EXPLORATORY ANALYSIS OF THE SPREAD OF CORONAVIRUS

INTESAR ABURAWI ALBKKOUSH

Master's Thesis

Graduate School

Izmir University of Economics

Izmir

2021

EXPLORATORY ANALYSIS OF THE SPREAD OF CORONAVIRUS

INTESAR ABURAWI ALBKKOUSH

A Thesis Submitted to

The Graduate School of Izmir University of Economics

Master's Program in Applied Statistics.

Izmir

2021

ABSTRACT

EXPLORATORY ANALYSIS OF THE SPREAD OF CORONAVIRUS

Albkkoush, Aburawi Intesar

Master's Program in Applied Statistics.

Advisor: Prof. Dr. Gözde Yazgı TÜTÜNCÜ

January, 2021

The thesis formulates comprehensive statistical models demonstrating the effects of the novel COVID-19 virus in various countries. The study is centred on analysing key variables responsible for the progression of the virus in several regions. Moreover, we sought to establish if any, a statistically significant relationship between influenza and pneumonia deaths in relation to COVID-19 and influenza related deaths. We formulated a regression model for the data set published by Kaggle (2020). Our research findings show that treatment in hospitalization is more effective than in home-isolation for infected patients. In addition, susceptibility to a high mortality rate for infected patients is prevalent among the elderly who are over the age of eighty with the exception of toddlers within the range of one to four year olds. Some of the most

notable factors are time, geographical factor, population density, and cultural factor. We use multiple linear regression models to determine the relationship between the number of COVID-19 deaths and the interpreted variables and also to determine the relationship between the number of COVID-19 cases and the interpreted variables. In both cases, we observe that there is a very strong direct relationship between the dependent variable and the independent variables. We use the fitted multiple linear regression equations to predict future values of COVID-19 deaths and cases. Our results indicate that we cannot forecast the number of COVID-19 deaths using age, symptoms, smoking, treatment and gender. We carry out Bayesian analysis and obtain a 95% credible interval (CI) which tells us that COVID-19 deaths are greatly affected by the independent variables.

Keywords: COVID-19, exploratory data analysis, SPSS Programming, Multiple Regression, Bayes Analysis, Influenza.

ÖZET

Koronavirüsün Yayılmasının Keşif Analizi

Albkkoush, Aburawi Intesar

Uygulamalı İstatistik Yüksek Lisans Programı

Tez Danışmanı: Prof. Dr. Gözde Yazgı TÜTÜNCÜ

OCAK, 2021

Bu tezde, yeni COVID-19 virüsünün çeşitli ülkelerdeki etkisini gösteren kapsamlı istatistiksel modeller önerilmiştir. Çalışma, virüsün çeşitli bölgelerde ilerlemesinden sorumlu temel değişkenleri analiz etmeye odaklanmıştır. Ayrıca, eğer varsa, influenza ve pnömoni ölümleri ile COVID-19 ve influenza ile ilişkili ölümler arasında istatistiksel olarak anlamlı bir ilişki kurulmaya çalışılmıştır Kaggle (2020). tarafından yayınlanan veri seti için bir regresyon modeli oluşturulmuş ve araştırma bulgularımız enfekte hastalar için hastanede tedavinin evde izolasyona kıyasla daha etkili olduğunu göstermektedir. Ek olarak, enfekte hastalarda seksen yaşın üzerindeki yaşlılarda yüksek ölüm oranı daha yaygındır ve, bir dört yaş aralığındaki çocuklar istisnadır. En dikkate değer faktörlerden bazıları zaman, coğrafya değişkeni, nüfus yoğunluğu ve kültür değişkenidir. COVID-19 ölümlerinin sayısı ile yorumlanan değişkenler

arasındaki ilişkiyi ve ayrıca COVID-19 vakalarının sayısı ile yorumlanan değişkenler arasındaki ilişkiyi belirlemek için çoklu doğrusal regresyon modeli kullanılmıştır. Her iki durumda da bağımlı değişken ve bağımsız değişkenler arasında güçlü bir ilişki olduğunu gözlemlenmiştir. COVID-19 ölümlerinin ve vakalarının gelecekteki değerlerini tahmin etmek için çoklu doğrusal regresyon denklemlerinden oluşturulan tahminler kullanılmıştır.Ancak sonuçlarımız, COVID-19 ölümlerinin sayısını yaş, semptomlar, sigara, tedavi ve cinsiyeti kullanarak çoklu doğrusal regresyon tahmin edemeyeceğimizi göstermiştir. Bu nedenle daha gelişmiş bir teknik olan Bayes Analiz yöntemleri uygulanmış ve COVID-19 ölümlerinin büyük ölçüde bağımsız değişkenlerden etkilendiğini gösteren % 95 güven aralığı elde edilmiştir.

Anahtar Kelimeler: COVID-19, keşifsel veri analizi, SPSS Programlama, Çoklu Regresyon, Bayes Analizi, İnfluenza.

To my dear father, my God protect him, who instilled in me the seriousness, diligence and the pursuit of knowledge. Remember him more and longer in his life to obey him…

To my tender mother, the symbol of love and giving. Remember her more and longer in her life to obey her…

To my dear husband, whom I had support in my life and which extended a helping hand to me in my scientific career…

To my dear children, may God protect them and be pleased with them and satisfy them. May God protect them…

To all those infected with the coronavirus, we wish them a speedy recovery and have mercy on the dead…

To my best teachers, and my foundation…

To my country, Libya, faraj God distress her soon…

To everyone whose goal was excellence, I dedicate the fruit of this humble effort…

ACKNOWLEDGEMENTS

Praise be to God, Lord of the Worlds, and prayers and peace be upon the most honourable of the prophets and messengers, our Prophet Muhammad, and upon all his family and companions.

After that, I thank God so much that he helped me to complete this thesis. Then I send the verses of thanks and gratitude to Supervisor" Prof. Dr. Gözde Yazgı TÜTÜNCÜ", the supervisor of the letter, who gave me a lot of her time, and the openness of her heart, the majesty of her creation and her distinctive style in following up the message had the greatest impact in helping to complete this work, and I asked God Almighty to reward her with the best reward He writes her work in the balance of her good deeds, and calls me the duty of loyalty and gratitude to thank everyone who supervised my teaching and extended a helping hand to me, even if with a kind word, I would also like to thank my friend Elif Duymaz, for her support and help , as well as all the trainers who facilitated my mission in order to reach success and last but not least, our prayers: Praise be to God.

PREFACE

The number of COVID-19 confirmed cases, released patients and the number of deaths are increasing rapidly in different countries. In all countries of the world, the number of diagnosed patients and the number of deaths for COVID-19 is gradually increasing (World Health Organisation, 2020). The national and local authorities are having a hard time to create a pattern, analyse and test hypotheses on the spread of COVID-19 in the world. The main aim of this thesis is to draw a statistical model for better understanding of the spread process of COVID-19 in different countries. An exploratory data analysis technique is implemented in order to study and analyse the reported COVID-19 data. We analyse the effects of COVID-19 in different countries and compare COVID-19 with other diseases such as influenza and pneumonia.

In this study, SPSS version 18 was used for analysis of the data. We compare between treated and confirmed, released, deceased patients and severity of the illness. The main research questions were as follows: Are there statistically significant differences between the average number of confirmed male cases and the average number of confirmed female cases; which country has more deceased and released and which country has less; are there differences in the average number of confirmed, released patients in different ages.

We also analyse the dynamics of the influenza. The main research questions in the case of influenza were as follows: which age group has the highest percentage of influenza deaths; which age group has the highest percentage of COVID-19 Deaths. Additionally, we also find answers to the question of whether there exists a statistically significant relationship between COVID-19 deaths and influenza deaths in different states or countries.

From our analysis of the COVID-19, we note that hospitalization results in the highest number of treated patients compared to home isolation. We find that in the severity of the illness, more than half of the infected patients recovered. However, we note that the percentage of deceased patients was more than those patients who were critically ill.

The findings from our research show that new confirmed cases in home isolation are more than in hospitalization. This means there is more attention to cleanliness and sterilization in hospitals than in home isolation. Furthermore, the cases of recovery for

home isolation are less than cases of recovery for patients who receive their treatment in the hospital. The reason for the low recovery rate is that the patients are not under medical supervision. In addition, there are more deaths in home isolation than admitted patients in hospitals, but however there is a small difference as shown by the percentages of 50.17% and 49.83% respectively.

From our analysis, we note that the highest number of deceased patients was in Colombia whilst the lowest was in China. Additionally, the highest mean number of released patients was in Colombia, whilst the lowest was in Canada and Malaysia.

Finally, in the data analysis of figure 16 we note that the highest value of influenza deaths was recorded in the age group of 65-74 years followed by the age group of 75-84 years with 24.62% and 24.43% respectively. On the other hand, the lowest value of deaths was in the age group 45-54 years and this could be because they have a lower percentage of being infected with the virus.

The figures 18, 19, 20, 21, 22 and 23 indicate that the highest number of COVID-19 deaths, Influenza deaths and Pneumonia deaths was recorded in New York and California. On the other hand, Idaho, Hawaii and Alabama have the lowest number of Influenza, Pneumonia and COVID-19 deaths respectively.

IZMIR

…/…/2021

INTESAR ABURAWI ALBKKOUSH

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In the beginning of 2020, the new corona virus (COVID-19) was escalating in China, and a large number of people had been infected (Livadiotis, 2020). At present, the outbreak in Turkey has been effectively controlled thanks to the strict curfew measures; however, the new corona virus is progressing rapidly in other countries. Currently, Europe has become the epicenter of the current outbreak of COVID-19. On March 11, the World Health Organization (WHO) declared the new pneumonia outbreak as a "global pandemic" (Stafford, 2020). According to Livadiotis (2020), the new corona virus has become a great threat to the health and safety of people all over the world due to its amazing spreading power and potential harm. Research on the progression, transmission and dynamics of the C0VID-19 pandemic is being widely conducted all over the world. Also, there are two types of diseases. An epidemic is a disease that affects many of people within a community, population, or region. A pandemic is an epidemic that's spread over multiple countries or continents (Stafford, 2020). We will touch on important advice on corona virus.

1. 1 - Research questions for COVID-19

In this research we will find answers to the following questions:

1.1.1- Which gender has the highest percentage of infections?

1.1.2- What are the treatment strategies available for infected patients and the different stages in the severity of the illness?

1.1.3 Which treatment strategy results in highest percentages of confirmed cases, recovered patients and deceased patients?

1.1.4- Which country has more deceased and recovered patients and which country has less?

1.1.5- Is there a link between severity of illness and gender?

1.1.6- Is there a link between severity of illness and treatments?

1.1.7- Is there a relationship between the smoking and severity of illness?

1.1.8 - Is there a statistically significant relationship between smoking cases and deceased, and released patients in different countries?

1.1.9-Are there statistically significant differences between the average number of

confirmed cases in males and the average number of confirmed cases in females?

1.1.10-Are there statistically significant differences between the average number of deaths in males and the average number of deaths in females?

1.1.11- Are there differences in the average age of patients in different countries

1.1.12-Are the numbers of deaths affected by age, date and smoking in different countries?

1.1.13-Are the numbers of cases affected by age, date and smoking in different countries?

1.1.14-Can we predict the number of deaths through age, symptoms, smoking, treatment and gender?

1.1.15-Can we predict the number of cases through age, symptoms, smoking, treatment and gender?

1.1.16-Are the numbers of deaths affected by smoking, severity of illness, infectious person, gender, treatment and symptoms in different countries?

1.2 - Research questions for Influenza

1.2.1- Which age group has the highest percentage of Influenza Deaths?

1.2.2- Which age group has the highest percentage of COVID-19 Deaths?

1.2.3- Which the country has more COVID-19 Deaths, Influenza Deaths and Pneumonia Deaths?

1.2.4 - Is there a statistically significant relationship between COVID-19 Deaths and Influenza Deaths in different countries?

1.2.5- Is there a statistically significant relationship between COVID-19 Deaths and Pneumonia Deaths in different countries?

1.2.6- Are there differences in the average numbers of COVID-19 Deaths and Influenza Deaths?

1.2.7- Are there differences in the average numbers of COVID-19 Deaths and Pneumonia Deaths?

1.3. The Importance of Research

The study entailed herein is very important because the Corona virus has led to huge losses in the most countries (Fonseca et al. 2020). This virus led to the closure of all

schools and public interests. This situation occurs for the first time in history all over the world. And this topic has become modern since the virus is spreading and researchers and scientists need to know which countries are most affected by this epidemic, and we will make some assumptions to obtain the required results.

The number of people infected with Coronavirus around the world has risen to more than 16 million and 883 thousand people, of whom more than 662 thousand have died, and more than 10 million and 450 thousand have recovered, according to the "World Meter" website, which specializes in counting virus victims (Chu et al. 2020).

In addition, most countries have lost a lot of their people, and people have lost their relatives and loved ones. Everyone has a relative, friend, or neighbour who is either infected or deceased. As for me, I lost many of my relatives and friends because of this deadly virus.

1.4 The aim

The aim of this study is to find a good model to in order to know which countries are most affected by this virus. Our objective is to study the relationship between the variables in a data set so that we would analyse the variables that are affecting the spread of the virus in different countries.

Another objective is to determine which countries are most affected by this dangerous virus. We also seek to find out if the prevalence of its commodities differs from one country to another, as well as to know which countries have controlled the spread of the virus.

CHAPTER 2: LITERATURE REVIEW

By the end of April 2020, death rates due to the novel COVID-19 had risen close to 1.2% in Germany (Müller et al. 2020), 8.6% in Netherlands, 11.9% in Italy (Furtuna et al., 2020). Chu et al. (2020) argues that Germany's low death rate could be attributed to quality controlled laboratories wide spread all over the country meant for the virus testing and early detection. Moreover, Stafford (2020) states that implementation of steep and rapid measures such as closure of businesses and schools including prohibition of gatherings might have alleviated the nation from a huge number of fatalities. Despite the low numbers, there is still a growing concern about future fatality rates (Mitić et al., 2020). Since, there might be a possibility of a reinfection among the patients (Yu, Chen and Chen, 2019).

Some other measures that have been effective are social-distancing (Smith, Adler, and Perelson 2010), shut down and quarantines (Giugliano, 2020), in eradicating the growth in infection rates. In addition, the National Academies of Sciences, Engineering and Medicine, (2020) highlight that there is some notable evidence in the reduced transmission efficiency of the virus due to environments with higher ambient temperature and humidity.

Black, Liu, and Mitchell (2020), argue that although, these measures outlined above are effective, the puzzle that remains unsolved is ascertaining factors which heavily influence the sharp growth rate of infected cases.

Deaths associated with influenza pandemics have been shown through a number of models (Osterholm, 2005; Meltzer, Cox, and Fukuda, 1999; Zhang, Meltzer, and Wortley, 2006). The Spanish flue pandemic that occurred during the 1918-20 is useful in predicting future pandemic deaths of related nature (Lemon et al., 2005), which resulted in close to 100 million deaths at most (Patterson, and Pyle, 1991; Johnson, and Mueller, 2002; Burnet, 1979). Most of these findings came from reliable sources such as historical documents and national commissions except (Noymer, and Garenne, 2000; Crosby, 2003; Collins, 1957).

Recently, an ever changing and varying estimates in range of fatality ratio for the novel corona virus disease 2019 (COVID-19) have been established. Verity et al. (2020) highlighted that primary estimates indicate that the fatality ratio for COVID-19 show a strong age gradient death risk.

Robust estimates, accounting for censoring and ascertainment biases show that mean duration of approximately 18 days from onset to death and a hospital discharge of about 25 days for survivors. Laboratory confirmed cases in mainland China show an estimated crude case fatality ration adjusted for censoring of about 3.7%. The case fatality ratio in China with adjustment for demography and under-ascertainment is 1.4%, ominously high ratios in advanced age groups ranging from 6.4% to 13.4% were obtained. Overall infection fatality ratio in China with an increasing profile in age is 0.7%. A maximum proportion of infected individuals' likelihood of being hospitalized is 18.4% and influenced by the advancement in age (Verity et al., 2020).
The factors connected with death of COVID-19 pneumonia patients, using univariate and multivariate logistic regression are found out to be, age of beyond 65 years, pre-existing concurrent cardiovascular or cerebrovascular diseases, $CD3^+$ $CD8^+$ T-cells and cardiac troponin I, the last two factors highly determine the mortality of COVID-19 pneumonia patients. There are no disparities in demographic and clinical presentation parameters among the dead and matched case-control survivors. Since numerous survivors are young people, the deceased patients compared to the young survivors had extremely increased concentrations of procalciton, cardiac troponin I, myoglobin and creatinine, reduced amount of $CD3^+CD8^+$ T-cells (Du et al. 2020).

The effect of the environmental temperature on the exponential growth rate in COVID-19 infected cases for USA - Italian region have a negative correlation relationship. A critical temperature of 86.1+4.3 F eradicates the COVID-19 exponential growth of the spread in the USA. This was shown by means of the analysis of regional data sets of infected cases through deriving the relationship between temperature and the exponential growth rate and the evaluation of its statistical confidence. Statistical analysis entailed fitting of linear statistical models with the inverse data set of the environmental temperature and the exponential growth rate. The reduced chi-square values and the p-value of extremes evaluated the statistical confidence of fitting.

Furthermore, comparison of the derived slopes was done by means combined testing measures and the Student's t-test (McHugh, 2013).

This goes to show how the Student's t-test has been of great importance in the study and analysis of the dynamics of the spread of the COVID-19. T-test is used on data following a normal distribution. Lee, In and Lee (2015) noted that samples drawn from normal distribution have a normal distribution influenced by number of samples. Lumley et al. (2002) stated that t-tests may still be applied even when the normality condition doesn't satisfactorily hold. Also, Box (1954) discovered that the equal variance condition has a little effect on the significance of level value.

The reproduction number which is the proportion to its logarithm is expected to result in a positive correlation with the exponential growth rate of COVID-19 spread. However, the rate in turn has a negative correlation with the incubation period which is an inverse proportional (McHugh, 2013).

Moreover, a forecast of a potential global mortality based on high-quality vital registration data in case a pandemic like the 1918-20 pandemic broke out and/ or reoccurred was considered. Countries with high-quality vital registration data for that pandemic contributed in excess mortality calculation. Ordinary least squares regression models were formulated and these linked excess mortalities to per-head income and income latitude. Forecasts for a 2004 mortality of a possible influenza pandemic were made using the formulated models. The findings show a discrepancy among population mortalities, a huge proportion of the discrepancy is clarified by the per-head income. Extrapolated mortality rates to the 2004 worldwide population forecasted an approximate of 62 million deaths associated with a related influenza pandemic, the mostly affected would have been third world countries (Murray et al., 2006).

The application of the ordinary least square regression models have had an undeniable contribution from the early 20th century and based on these previous models, impact of future pandemics can easily and readily be predicted. The primary reasons influencing inflated susceptibility to COVID-19 virus among the most vulnerable to the virus which are the elderly were investigated. These were found to be underlying conditions that hinder ability to recover and the changes to immune response with age.

Using R programming language to analyse the Kaggle published data set from John Hopkins University, the vulnerability and mortality rate of the elderly people because of the COVID-19 compared to the young and/ or the middle aged were assessed. The use of statistical software and data analysis have had a huge impact in analysing important data sets with great ease, especially in evaluating the effectiveness of some enhanced drug-developments. The ever-growing concerns of possible second waves of infections could be avoided through these, hence alleviating the burden on hospitals and health system capacities. It is inevitable for nations to open up their borders and start to function properly, but this could be done guided by the outcome of the analysis made by the R programming language (Müller et al., 2020).

Ogundokun et al. (2020) adopted the ordinary least squares estimator as a way to measure the influence of travelling history and contacts on the spread of COVID-19 in Nigeria. The regression equation was used to make predictions. Based on the diagnostic checks conducted, the fitted model was a best fit to the dataset and was free of any violation. Their results agree and support the decision made by the government made in restricting travelling because their analysis concluded that travelling history and contacts made resulted in an increase in the chances of people being infected with COVID-19 by 85% and 88% respectively.

Rath, Tripathy, and Tripathy (2020) applied Linear and Multiple Linear Regression techniques to a data set to envision the trend of the affected cases. Linear Regression and Multiple Linear Regression models are compared and the score of the model R2 tends to be 0.99 and 1.0 which suggests a strong prediction model to forecast the next coming day's active cases. From their study they concluded that these models gained remarkable accuracy in COVID-19 recognition.

Multiple regression and linear regression analyses were used replaceable by Ghosal et al. (2020) in order to discover a trend related to death counts expected at the 5th and 6th week of the COVID-19 in India. Their results indicate that the week 6 death count data was not significantly correlated with any of the chosen inputs and as a consequence they used an auto-regression technique in order to enhance the predictive ability of the regression model. They successfully predicted the average week 5 death count to be 211 with a 95% CI: 1.31–2.60) using a linear regression model. On the other hand, the auto-regression technique was used and also the week 5 death counts

as input, the linear regression model made a prediction of 467 for week 6 death count in India, while not overlooking the risk of over-estimation by most of the risk-based models.

Chaurasia, and Pal (2020) discovered and presented the estimated death rate by the ARIMA model and the regression model. The auto arima SARIMAX results, auto arima residual plots, ARIMA model results, and corresponding prediction plots on the training dataset were generated by the ARIMA model. The coefficients generated by the regression model were approximated, and the actual death cases and expected death cases were contrasted and analysed by the use of a regression model,

Choi, Kim, and Ahn (2019) collected the surveillance data of 164 countries by the use of the FluNet database, search queries from Google Trends, and temperature from 2010 to 2018. They collected data for influenza-like illness (ILI) in the U.S. from the Fluview database. A time lag between two time-series was identified and these were weekly surveillances for ILI, total influenza (Total INF), influenza A (INF A), and influenza B (INF B) viruses between two countries using cross-correlation analysis. Prediction models were expanded so as to forecast ILI, Total INF, INF A, and INF B of next season (after 26 weeks) in the U.S. and this was done using linear regression, auto regressive integrated moving average, and an artificial neural network (ANN). These models forecasted that the ILI for the U.S. in 2018–2019 would be later and lighter than those in 2017–2018.

Oviedo de la Fuente et al. (2016) used meteorological information in Galicia (Spain) to suggest a novel proposition to forecast the incidence of influenza. Their approach was an extension of the GLS methods in the multivariate framework to functional regression models with dependent errors. Results from a simulation study indicated that the GLS estimators provided better estimations of the parameters associated with the regression model and obtained extremely quality results. Thus, they improve the classical linear approach.

Darwish, Rahhal, and Jafar (2020) explored the performance of three distinct feature spaces in different models to predict the weekly influenza-like illness (ILI) rate in Syria using EWARS data from World Health Organization (WHO). Initially a time series feature space was employed. Seven models were applied and these are Naïve,

Average, Seasonal naïve, drift, dynamic harmonic regression (Dhr), seasonal and trend decomposition using loess (STL) and TBATS.

Uyanik, and Guler (2013) extracted data for multilinear regression analysis, from Sakarya University Education Faculty student`s lesson scores and their KPSS score. They examined the normality, linearity, no extreme values and missing value analysis assumptions of multilinear regression analysis. Analysis of the data that verified the assumptions was carried out by employing multiple regression and lessons measurement and evaluation, instructional techniques, counselling, program development and educational psychology estimated the KPSS respectively.

Bayesian Inference permits us to represent uncertainty as a probability. It is at the centre of the Bayesian approach. The Bayesian approach is compared with the Frequentist approach which bases probabilities on repeatable, random events and has null hypothesis testing at its heart in order to understand it. Bayesian Inference on the other hand does not test null hypotheses but includes prior knowledge and does not rely on duplication or necessarily uncertainty (IBM SPSS Forecasting, 2012).

SPSS Forecasting is fully combined with IBM SPSS Statistics, as a result all of its capabilities are allotted, and in addition features specifically designed to support forecasting (To, and Mandracchia, 2019). Forecasts have a major impact on profits since they assist you to develop and manage plans influencing a number of operational areas. SPSS Forecasting has the modern techniques you require without the disadvantages of traditional methods. Reliable forecasts are developed quickly, no matter how large the dataset or the number of variables involved, with the use of SPSS Forecasting. Besides, the forecasting error is reduced by automating the selection of the suitable models and their parameters.

However, with such wealth of statistical software and models, there remains a major concern over unverified myths on the dynamics and the spread of the COVID-19 especially on its effects according to gender and location. There remains a wide mystery concerning this side of things. This is an enormous motivation to the work I intend to cover, addressing such gaps and limitations to the studies that have been covered so far. The study we embark on seeks not only to cover the uncovered areas in research but it could widely be beneficial to the economics of various nations in

terms of how they incorporate insurance cover for their citizenry and on how they can gradually open up their borders for normal trade and for restocking additional health supplies among other issues of concern.

The previous literature covers a wide range of useful research findings and methodology regarding regression models dealing with pneumonia and influenza. Models that were formulated and made to predict future infection cases are still vital although they may need to be adjusted according to regions and other important factors that hadn't been evaluated in the past nor regarded as anything worth considering. Some of these factors include and are not only limited to the effects of the COVID on smoking and non-smoking group of people and the effects of gender, that is, the dynamics associated with each gender type in comparison to the other. Some handy methodology is still to be considered and integrated in modern models though appearing simple and basic but much more effective; these include the Student's t test and the Z-test for two samples, among others.

CHAPTER 3: METHODOLOGY

3.1 Statistical Hypothesis testing

What is a hypothesis test?

In general terms, a statistical hypothesis is a condition, statement or an assumption of one or more parameters of a population or a probability distribution (Yim et al., 2010). In hypothesis testing, we take an initial statement about the system in question to be a null hypothesis and we then test it against an opposite statement known as the alternative hypothesis based on a test statistic.

The general steps to be followed in a test of hypothesis are:

1) Identify the hypotheses: Stating the null hypothesis H_0, and the alternative hypothesis H_1. The H_1 always contradicts H_0.

2) Error Probability: It is usually the type I error known as the level of significance, α. It may also seldom include the type II error, β.

3) Rejection Criterion: State the test statistic (TS) and formulate the rejection region (RR), one that should be satisfied in order to reject H_0.

4) Calculations: Calculate the value of the test statistic based on the sample data in the problem.

5) Decision-making and conclusion: Make the decision whether to reject H_0 thus accepting H_1, or fail to reject H_0. It is noted that failing to reject H_0 does not affirm its acceptance. It may be only assumed accepted as per the given data sample. True acceptance of the null hypothesis can only be attained if it holds for all samples of the population. In general, to test hypothesis concerning mean, we use *z* and *t*-distributions and to test hypothesis concerning variances or standard deviations, we use the χ^2 (chi-squared) and F distributions.

The terms and notations for statistical hypothesis testing are very few. We define them:

- A unit is a single item or entity
- A population is a complete collection of units of a system
- A sample is a subset of a population

Table 1. Common notations for test of hypothesis.

Notation Description	Notation Description
H_0	Null hypothesis
H_1	Alternate hypothesis
μ	Population mean
$\bar{x}$	Sample mean
σ^2	Population variance
s^2	Sample variance
$\sigma/\sqrt{n}$	Standard error
$s/\sqrt{n}$	Estimated standard error
α	Type I error (level of significance)
β	Type II error
ν	Degree of freedom (nu)

Hypothesis Tests

Hypothesis testing is a formal process used by statistician to determine whether to reject a null hypothesis, based on sample data. According to Yim et al. (2010), hypothesis testing consists of these four steps:

1. State the hypotheses - This step involves stating both null and alternative hypotheses. The hypotheses should be stated in such a way that they are mutually exclusive. If one is true, then other must be false.

2. Formulate an analysis plan - The analysis plan is to describe how to use the sample data to evaluate the null hypothesis. The evaluation process focuses on a single test statistic.

3. Analyse sample data - Find the value of the test statistic (using properties like mean score, proportion, t statistic, z-score, etc.) stated in the analysis plan.

4. Interpret results - Apply the decisions stated in the analysis plan. If the value of the test statistic is very unlikely based on the null hypothesis, then reject the null hypothesis.

The formula for hypothesis test is:

Hypothesis test Formula: if $\bar{x}_1$ and $\bar{x}_2$ are the means of the two samples, Δ is the hypothesized difference between the population means (0 if testing for equal means), s_1 and s_2 are the standard deviations of the two samples, and n_1 and n_2 are the sizes of the two samples.

Hypothesis testing:

We have three cases:

- We want to test that the population mean, μ, is different from 50, i.e. $\mu_o = 50$.

H_0 : μ=50

H_1 : μ≠50

We want to test that the population mean, μ is greater than 50

H_0 : μ≤50

H_1 : μ > 50

- We want to test that the population mean, μ is less than 50.

H_0 : μ≥50

H_1 : μ < 50

3.1.1 - Chi-Square Test

Researchers must carry out a test of significance called the Chi-Square Test in order to determine whether the association between two qualitative variables is statistically significant. There are five steps to conduct this test.

The Chi-Square Statistic

The Chi square statistic denoted by x^2, is quite different from the other statistics which have been widely used in the hypotheses tests. In addition, it seems to bear little similarity to the theoretical chi square distribution. The chi square statistic is the same for the goodness of fit test and the test of independence. All the categories into which the data has been divided are used in both of these tests. The data acquired from the sample is known as the observed numbers of cases.

In the chi square tests, the null hypothesis makes a statement concerning how many cases are to be expected in each category if this hypothesis is correct. The chi square test is determined by the difference between the observed and the expected values for each category. The chi square statistic is defined as:

$$x^2=\Sigma \frac{(observed - Expected)^2}{\text{Expected}} \qquad (1)$$

where, O_i is the observed number of cases in category i, and E_i is the expected number of cases in category i.

This chi square test statistic is obtained by calculating the difference between the observed number of cases and the expected number of cases in each category. then we square the difference and divide it by the expected number of cases in that category. These resulting values are then summed up for all the categories, and the total is referred to as the chi square test statistic value.

Chi Square Calculation

Each entry in the summation can be referred to as "The observed minus the expected, squared, and divided by the expected." The chi square value for the test as a whole is "The sum of the observed minus the expected, squared, and divided by the expected."

The null hypothesis is a particular claim concerning how the data is distributed. The null and alternative hypotheses for each chi square test can be stated as follows:

$H_0 : O_i = E_i$

$H_1 : O_i \neq E_i$

If the claim made in the null hypothesis is true, the observed and the expected values are close to each other and O_i - E_i is small for each category. In the case where the observed data does not conform to what has been expected on the basis of the null hypothesis, the difference between the observed and expected values, O_i - E_i, is large. Thus, the chi square statistic is small when the null hypothesis is true and large when the null hypothesis is not true. A question arises on how large the χ^2 value must be in order to reject the null hypothesis.

We do not consider the general formula for determining the numbers degrees of freedom in our research, because this is different for the two types of chi square tests. the degrees of freedom is based on the number of categories which are used in the calculation of the statistic in each type of test.

The chi square statistic, along with the chi square distribution, permits the researcher to find out whether the data is distributed as claimed. If the chi square statistic is large enough to reject H_0, then the sample provides evidence that the distribution is not as claimed in H_0. If the chi square statistic is small or not very large, then the researcher may have insufficient evidence to reject the claim made in the null hypothesis (Uyanik, and Guler, 2013).

We carry out the chi square test as shown in Tables (4, 5 and 6).

3.1.2- Correlation test

The correlation coefficient r and the coefficient of determination r^2 measures summarize the strength of a linear relationship in samples only (McHugh, 2013). Different samples have different correlations, different r^2 values, and therefore potentially different conclusions. As always, the main objective is to draw conclusions about populations, not just samples. In order to do so, we either have to conduct a hypothesis test or calculate a confidence interval. We now outline how to conduct a hypothesis test for the population correlation coefficient, ρ.

Generally, a researcher should use the hypothesis test for the population correlation ρ to study about a linear association between two variables, when there is uncertainty on which variable should be regarded as the response.

If the Pearson's correlation is used to be, then one must carry out necessary checks to ensure that the Pearson's correlation is the appropriate statistic. This is done by ensuring the following four assumptions are passed:

1. The two variables must be measured at the interval or ratio scale.

2. There is a linear relationship between the two variables.

3. There should be no significant outliers. Outliers are single data points within your

data that do not follow the usual pattern.

4. The data should be approximately normally distributed.

Computing the Pearson's Correlation Coefficient, r

Given the bivariate set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the Pearson's Product Moment correlation Coefficient is calculated as follows:

$$r = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2}} \qquad (2)$$

Where, r is the Pearson's Product Moment Correlation Coefficient,

$i = 1,2 \ldots, n$

x_i is the number of the x-variable in a sample,

$\bar{x}$ is the mean of the x-values of the x-variable,

y_i is the number of the y-variable in a sample,

$\bar{y}$ is the mean of the y-values of the y-variable.

The correlation coefficient measures the intensity of the association between variables.

• r is a unit-less number.

• It cannot be used to extrapolate a change in y based on a change in x.

• If variables are highly correlated, then we may want to further investigate their association in order to determine if there is a causal mechanism operating.

Given the above relationship (Sharpe, 2015), we then see the results for our data set in tables (7, 19 and 20).

3.1.3 t-test

Let`s say you conduct an experiment in order to compare two groups and quantify the difference between them (Free Online courses, Wall Street mojo, 2021). For instance:

- Comparing the height of people in two countries.
- Comparing if the extent to which the brain is activated when watching happy or sad movies.

Analysis of the comparisons is done by performing different statistical analysis, such as t-test, which we describe below.

As a type of inferential statistic, a *t*-test is used to determine if there exists a statistical difference between two groups. Mathematically, the *t*-test demonstrates the problem by assuming that the means of the two distributions are equal (H_0: $\mu=\mu_0$). If the *t*-test rejects the null hypothesis (H_0: $\mu\neq\mu_0$), then it implies that the groups are not the same (that is different).

This test should be used when the groups have 20–30 sample sizes (Lee, In and Lee, 2015). Other tests more accurate than t-tests are the z-test, chi-square test or f-tests are used so as to examine more groups or larger sample sizes.

- The p-value and the critical values

McHugh (2013) defined the p-value and critical value as:

The p-value is the probability of getting test results which are as greatest possible as the results observed during the test, whilst making an assumption that the null hypothesis is correct.

The critical values of a statistical test are the borders of the acceptance region of the test (Box, 1954).

The p-value is the variable that enables us to reject the null hypothesis (H_0: $\mu=\mu_0$) or, in other words, to conclude that the two groups are distinct (Verity et al., 2020). However, since the p-value is just a value, there is a need to compare it with the critical value (α):

- If p-value > α (Critical value) then the decision is; fail to reject the null hypothesis of the statistical test.

- If p-value ≤ α (Critical value) then the decision is; reject the null hypothesis of the statistical test.

The critical value that mostly chosen by statisticians is α = 0.05. This 0.05 in other words implies that, if we carry out an experiment 100 times, 95% of the times we would be able to reject the null hypothesis and 5% we would fail to reject the null hypothesis.

Also, in some cases, statisticians choose the value of α = 0.01. The reduction of the critical value from 0.05 to 0.01 decreases the chance of a false positive (called a Type I error), but it also makes it more challenging to reject the null hypothesis. Therefore, if a critical value of 0.01 is used, the results are more trustworthy but however not easily obtained.

- A p-value that is greater than 0.1 means no sufficient evidence
- A p-value that is between 0.05 and 0.1 means weak evidence
- A p-value that is between 0.01 and 0.05 means there is evidence
- A p-value between 0.001 and 0.01 means there is strong evidence
- A p-value less than 0.001 means there is very strong evidence

When there is a difference between groups in a specific direction, the one-tailed test is appropriate (Du et al., 2020). It is infrequently used than the two-tailed test.

Performing a t-test

The t-test is useful in to approximating the difference between two group means by using the ratio of the difference in group means over the pooled standard error of both groups. It can be calculated manually using a formula or using statistical analysis software.

t-test formula

The formula for the two-sample t-test is shown as follows:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad (3)$$

In this formula, t is the t-value, $\bar{x}_1$ and $\bar{x}_2$ are the means of the two groups being compared, s^2 is the pooled standard error of the two groups, and n_1 and n_2 are the number of observations in each of the groups. Also, μ_1 and μ_2 are the population means.

A larger t-value shows that the difference between group means is greater than the pooled standard error, indicating a more significant difference between the groups. As we see it in tables (8 and 9).

3.1.4 Z- test

Z-Test Hypothesis Testing

1. Introduction.

The *Z*-test is a simple tool for hypothesis testing that can be used to identify whether a mean result (Free Online Courses, Wall Street mojo, 2021) when compared to a larger set is statistically significant.

2. Hypothesis Testing.

The steps of hypothesis testing using the z-test are as follows: identify our population, comparison distribution, hypothesis and assumptions. Choose an appropriate test.

3. One-Sample *Z*-Test Formula.

If $\bar{x}$ (*x*-bar) is the sample mean, Δ (delta) is the value you are comparing it with (the population mean), σ (sigma) is the population standard deviation.

A *Z*-test is any statistical test for which the distribution of the test statistic under the null hypothesis is approximately normally distributed. It is useful for testing the mean of a distribution. The *Z*-test has a single critical value (for instance, 1.96 for a two tailed test at 5%), for each level of significance in the confidence interval which

makes it more appropriate than the Student's *t*-test whose critical values are determined by the sample size (through the corresponding degrees of freedom).

Many test statistics are approximately normally distributed for large samples, because of the central limit theorem. If the sample size is large or the population variance is known, then many statistical tests can be appropriately carried out as approximate *Z*-tests. In the case where the population variance is unknown (and therefore has to be estimated from the sample itself), and the sample size is small (n < 30), the Student's *t*-test is more convenient.

When performing a Z-test given that *T* is a statistic that is approximately normally distributed under the null hypothesis, the following steps are used:

First, we approximate the expected value, μ of *T* under the null hypothesis, and get an estimate *s* of the standard deviation of *T*.

Secondly, we determine the properties of *T*: whether it is one tailed or two tailed.

For Null hypothesis H_0: $\mu \geq \mu_0$ vs alternative hypothesis H_1: $\mu<\mu_0$, it is upper/right-tailed (one tailed).

For Null hypothesis H_0: $\mu \leq \mu_0$ vs alternative hypothesis H_1: $\mu>\mu_0$, it is lower/left-tailed (one tailed).

For Null hypothesis H_0: $\mu = \mu_0$ vs alternative hypothesis H_1: $\mu \neq \mu_0$, it is two-tailed.

Third, calculate the standard score :

The Z-test works because the sum of normally distributed random variables is also normally distributed. We can perform Z-tests in instances where the underlying population is not normal. For n large and known population variance, then by the Central Limit Theorem, the distribution of the random variable Z is approximately the standard normal, and as a result we may apply the z-test here.

$$Z= \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \qquad (4)$$

A more difficult problem arises when the population variance is unknown. If n is small then there is a problem; however, if n is large (n ≥ 30 adequate for most distributions commonly encountered) the following approximation is quite good. We replace the unknown population variance with the sample variance, s^2.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad (5)$$

In this case x_i corresponds to each observation in the sample, and $\bar{x}$ the mean of the sample as always.

Everything else in the analysis does not change it is as before. Observe that the square root of the sample variance is the sample standard deviation s, and if given the sample standard deviation, one may use the following analogous formula;

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \qquad (6)$$

to calculate the test statistic. The point is that, for n large, $s^2$is a good approximation to the unknown population variance σ^2.

We use the Z-test as shown in tables (10, 21 and 22)

3.1.5 - Regression analysis

Regression analysis is performed in order to determine the correlations between two or more variables having cause-effect relations. Answers are sought in this research to questions such as:

- Are there any relations between dependent and independent variables?
- If there are any relations, what is the power of the relation?
- Is it possible to make future-oriented predictions regarding the dependent variable?
- If certain conditions are controlled, what influences does it have on a special variable?

The regression using one single independent variable is called univariate regression analysis while the analysis using more than one independent variable is called multivariate regression analysis (Oviedo de la Fuente et al., 2016).

Multivariate regression analysis model is formulated as in the following: group of variables have over another variable or variables

$$y = B_0 + B_1 x_1 + \ldots + B_n x_n + \ell \qquad (7)$$

where, y= dependent variable

x_1=independent variable

ℓ=error

B_0 , $B_1, \dots, B_n$ = constants (Ghosal et al., 2020).

We use regression analysis as shown it in the tables (11 and 12).

3.2-Time Series Modeler

The Time Series Modeler procedure estimates exponential smoothing, univariate Autoregressive Integrated Moving Average (ARIMA), and multivariate ARIMA (or transfer function models) models for time series, and produces forecasts. The procedure includes an Expert Modeler that attempts to automatically identify and estimate the best-fitting ARIMA or exponential smoothing model for one or more dependent variable series, thus eliminating the need to identify an appropriate model through trial and error. Alternatively, you can specify a custom ARIMA or exponential smoothing model (To, and Mandracchia, 2019).

Statistics. Goodness-of-fit measures: stationary *R*-square, *R*-square (R^2), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), maximum absolute error (MaxAE), maximum absolute percentage error (MaxAPE), normalized Bayesian information criterion (BIC). Residuals: autocorrelation function, partial autocorrelation function, Ljung-Box *Q*. For ARIMA models: ARIMA orders for dependent variables, transfer function orders for independent variables, and outlier estimates. Also, smoothing parameter estimates for exponential smoothing models.

Plots. Summary plots across all models: histograms of stationary *R*-square, *R*-square (R^2), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), maximum absolute error (MaxAE), maximum absolute percentage error (MaxAPE), normalized Bayesian information criterion (BIC); box plots of residual autocorrelations and partial autocorrelations. Results for individual models: forecast values, fit values, observed values, upper and lower confidence limits, residual autocorrelations and partial auto correlations (Chaurasia, and Pal, 2020).

Time Series Modeler Data Considerations

•Data.

The dependent variable and any independent variables should be numeric.

•Assumptions.

The dependent variable and any independent variables are treated as time series, meaning that each case represents a time point, with successive cases separated by a constant time interval.

•Stationarity

For custom ARIMA models, the time series to be modelled should be stationary. The most effective way to transform a non-stationary series into a stationary one is through a difference transformation--available from the dialog box.

•Forecasts.

For producing forecasts using models with independent (predictor) variables, the active dataset should contain values of these variables for all cases in the forecast period. Additionally, independent variables should not contain any missing values in the estimation period.

•Defining Dates

Although not required. This is done prior to using the Time Series Modeler and results in a set of variables that label the date associated with each case. It also sets an assumed periodicity of the data--for example, a periodicity of 12 if the time interval between successive cases is one month. This periodicity is required if you're interested in creating seasonal models. If you're not interested in seasonal models and don't require date labels on your output, you can skip the Define Dates dialog box. The label associated with each case is then simply the case number.

We show it in the table 13.

3.3 - Bayesian Inference

Bayesian Inference is at the core of the Bayesian approach, which is an approach that allows us to represent uncertainty as a probability. One way to understand the Bayesian approach is to contrast it with the Frequentist approach which bases probabilities on repeatable, random events and has null hypothesis testing at its heart. In contrast, Bayesian Inference does not test null hypotheses but incorporates prior knowledge and does not rely on repetition or necessarily randomness (IBM SPSS Forecasting, 2012).

Calculating Bayesian Inference

At the heart of Bayesian Inference is Bayes' Theorem, Equation 8 below:

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)} \quad (8)$$

where, *A* and *B* are events and:

P(*A*|*B*) = probability of *A* given *B* has occurred

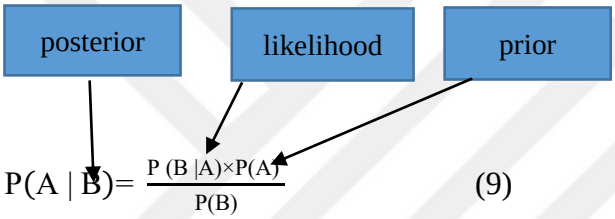
P(*B*|*A*) = probability of *B* given *A* has occurred

P(*A*) = probability of *A*

P(*B*) = probability of *B*

P(*A*|*B*) and P(*B*|*A*) are known as conditional probabilities, which is the probability of one event (*A* or *B*) occurring given another event (*A* or *B*) has already occurred.

.



$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)} \quad (9)$$

Where:

P(*A*|*B*) = posterior

P(*B*|*A*) = likelihood

P(*A*) = prior

Conducting Bayesian Analysis: Prior and Posterior Distributions

Bayesian analysis uses different terminology to Frequentist, so it is useful to review it alongside the key steps in a Bayesian approach.

Prior Distributions

The first step in a Bayesian analysis is to specify what is known as the Prior Distribution. As noted previously, one of the core differences in Bayesian Inference is that existing knowledge can be incorporated into the calculation of probabilities and the wider statistical model. This prior knowledge is known as Prior Distributions or Priors. In all Bayesian analysis, you have to specify Prior Distributions for all parameters in the model (e.g., means, regression coefficients, etc.). These Prior Distributions are based on our existing knowledge of the parameters before observing

our data; they may be based on previous studies and/or existing literature. Prior Distributions take the shape of different probability distributions, for example, a normal distribution. There are two types of Prior Distributions:

• Non-informative distributions. This type is used when we have no clear reason to expect one value over another and ranges from 0 to +/− infinity.

• Informative distributions.

This type is used when we want to take into account prior knowledge. Often these distributions will take the shape of a normal distribution and vary by mean and variance

The variance will vary by how certain you are that the parameter value will fall close to the estimate; low variance means high certainty and high variance means low certainty.

Observed Data

Once the Prior Distribution is established, you can then conduct your analysis on your observed data. Here, we would look at the observed evidence for the parameters (e.g., mean, variance) in the actual data. These parameters are calculated using a likelihood function, which tells us the most likely values for the unknown parameters given our data.

Posterior Distributions

The final step in a Bayesian analysis is to obtain what is known as the Posterior Distribution using Bayes' Theorem (see Equation 1). Our Prior Distribution (essentially our prior knowledge) is updated/modified by our observed data analysis, and from this, we can specify our Posterior Distribution (essentially our updated knowledge). The Posterior Distribution is usually obtained by Markov Chain Monte Carlo Methods via statistical software.

To summarise the relationship between the Prior distributions, observed data, and Posterior distribution in terms of updating or modifying our knowledge:

• If we had little or no knowledge to begin with (i.e., a non-informative Prior), whatever we learnt from our observed data would typically update our knowledge (i.e., our Posterior distribution) (IBM SPSS Forecasting, 2012).

• If we had some knowledge to begin with (i.e., an informative Prior) and the observed data confirmed this, then we would be more confident about our initial knowledge. In

a sense, the more knowledge we start with that is then confirmed by the data, then the greater our confidence about this knowledge.

Credible Intervals (CIs)

In Frequentist approaches, confidence intervals are used as one of a series of elements to assess our findings. Bayesian statistics does not use confidence intervals but something called credible intervals. The 95% CI is the central 95% of the Posterior Distribution, the range in which we think that it is 95% likely that the true figure lies, based on our Prior and observed data.

We have shown these in Tables (14, 15, 23 and 24).

Motivation

We used the Bayesian analysis since my forecasting and regression analyses were not very helpful in order to get good results. So, we study the linear dependencies or influences of predictor or explanatory variables on response variables using Bayesian analysis. We then predict or forecast future responses given future predictor data.

CHAPTER 4: DATA AND THE RESULTS

4.1 Coronavirus Dataset COVID-19

The 2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China. Earlier on, in the outbreak, many of the patients in Wuhan, China reportedly had some link to a large seafood and animal market, suggesting animal-to-person spread. However, a growing number of patients reported that they had not been to animal markets, indicating that person-to-person spread was occurring. At this time, it's unclear how easily or sustainably this virus is spreading between people.

Coronavirus Dataset COVID-19 is a collection of the COVID-19 data maintained by Our World in Data. It is updated daily and includes data on confirmed cases, deaths, and testing, as well as other variables of potential interest. More details about the data are in Google (2020).

The data I used was for the period stated from (27/01/2020to 15/08/2020) and it contains 34 columns and 1048576 observation

I chose this dataset because contains many important variables that are useful in the subject under study.

You can access the CSV file for the complete dataset on Kaggle (2020).

Table 2. Variables for Coronavirus Dataset COVID-19.

categorical data		value data	
name col	explain	name col	Value
severity illness		Id	
asymptomatic	4	age band	discrete
good	2	Age	discrete
critical	3	return_date_until_date_onset_symptoms	discrete
deceased	1	date_onset_symptoms_until_confirmed_date	discrete
cured	0	confirmed_date_until_released_date	discrete
gender		confirmed_date_until_deceased_date	discrete

male	0		
female	1	confirmed_date	discrete
transgender	2	confirmed_date_D	discrete
background_diseases_binary		deceased_date_D	discrete
TRUE	1	released_date_D	discrete
FALSE	0	return_date_D	discrete
treatment		date_onset_symptoms_D	discrete
hospital	2	len_people_infected_by_patient	discrete
clinic	1		
home isolation	0		
smoking			
TRUE	1		
FALSE	0		
severity_illness_infectious_person			
asymptomatic	4		
good	1		
critical	2		
deceased	3		
cured	0		
country			
background_diseases_name_of_disease			
TRUE	1		
FALSE	0		
symptoms_name_of_symptom			
TRUE	1		
FALSE	0		
severity_illness_over_time			
TRUE	1		
FALSE	0		

Table 3. Descriptive Statistics for Coronavirus Dataset COVID-19.

	Minimum	Maximum	Mean
age	0	119	42.12
confirmed	66	289	223.81
deceased	69	289	229.19
released	76	289	264.15

The results from table 3: shows the highest value of 76 in the released and the highest mean of 264.15 was also in the released.

4.1.1 Analysis of the Coronavirus Dataset COVID-19

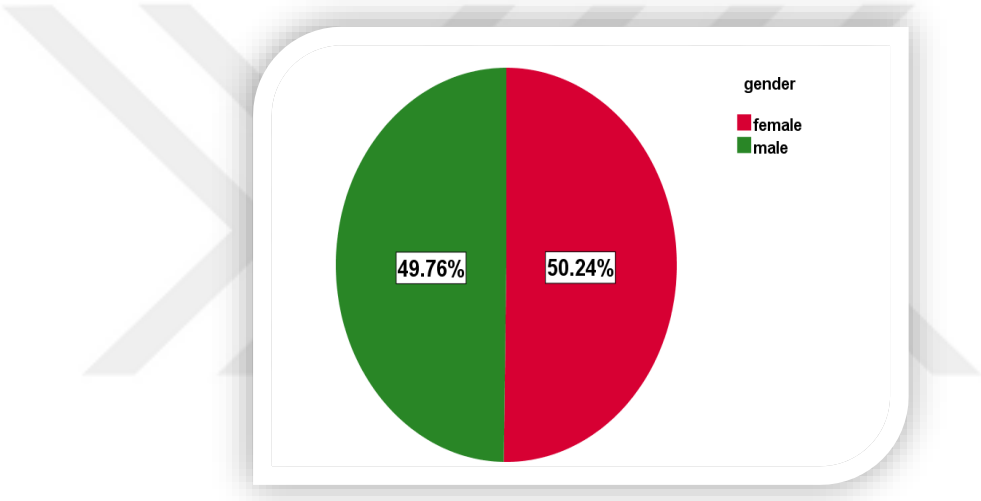


Figure 1. Pie chart for gender.

In Figure 1 we note that the rates of infection for male and female is approximately the same. The rate of infection for male is 49.76%, while for female is 50.24% which means that the number of female patients is slightly greater than the number of male patients. These results show that both male and female have an approximately equal chance of being infected with COVID-19.

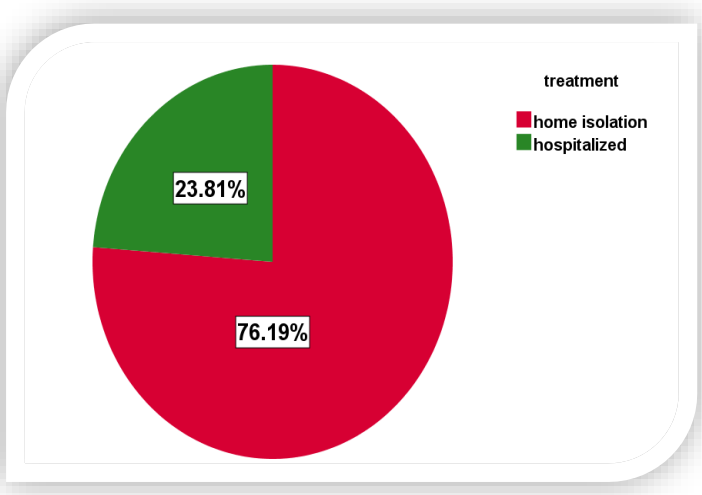


Figure 2. Pie chart for treatment.

In our study we consider two treatment strategies for COVID-19 which is hospitalization and home isolation. In our analysis process of figure 2, we can easily see that, the vast majority of treatment is done for hospitalized patients, it is equal to 76.19%, and however the treatment in home isolation is 23.81%. These results indicate that most of the patients are in a critical condition.

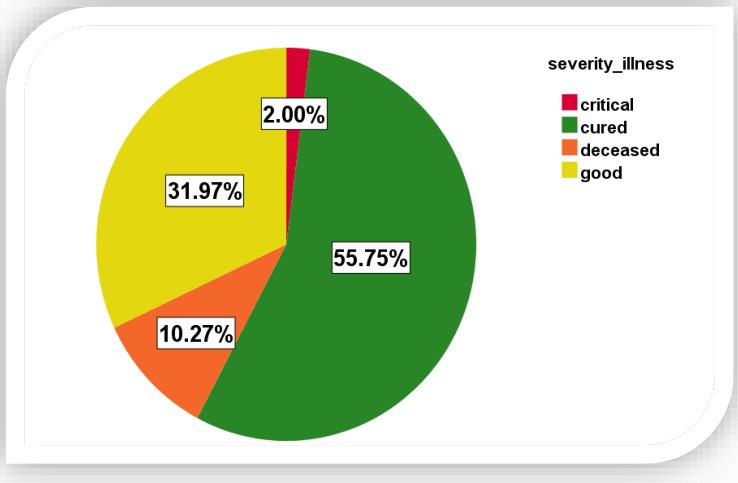


Figure 3. Pie chart for severity of illness.

For the severity of illness, we divided the patients into five groups which are: asymptomatic, critical, cured, deceased and good. Analysis of figure 3 shows that 55.75% of patients were cured, which is the greatest percentage, 31.97% were good, 10.27% were deceased, 2.80% were critical. These results mean that more than half of the patients recovered, however, we feel sorry because there were more deceased than critical patients.

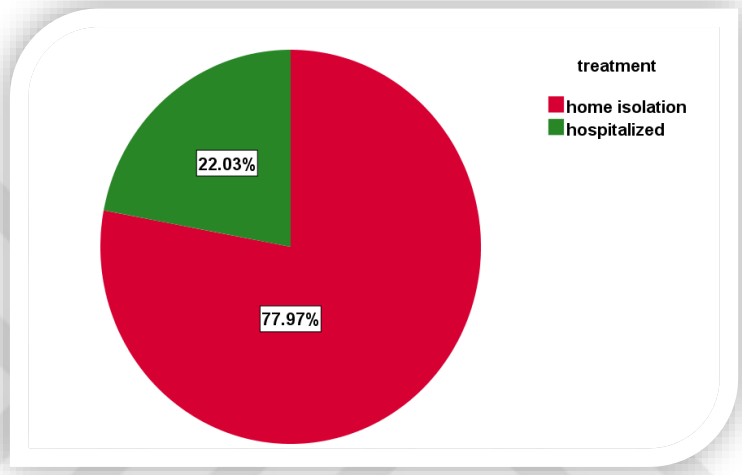


Figure 4. Pie chart for the mean confirmed patients and treatment.

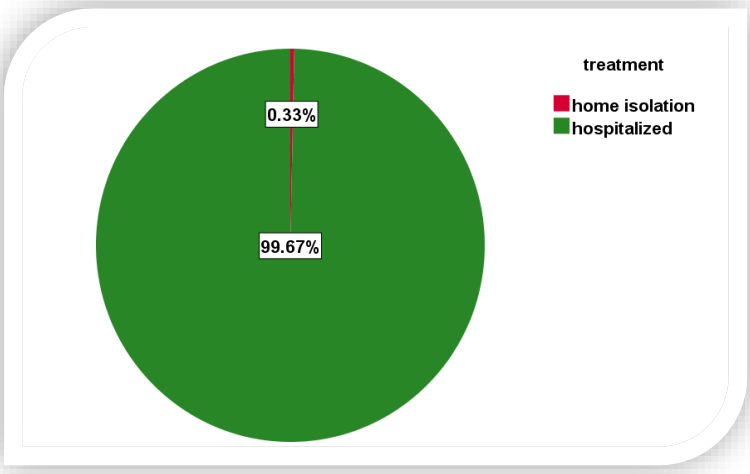


Figure 5. Pie chart for the mean released patients and treatment.

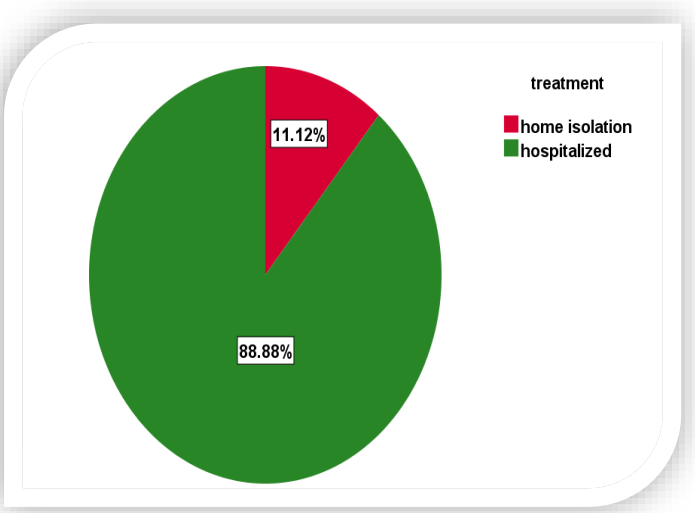


Figure 6. Pie chart for the mean deceased patients and treatment.

In our data analysis process, figure 4 shows that a higher number of new cases were confirmed in home isolation than hospitals, as indicated by the percentages of 77.97% and 22.03% respectively. This means that there is more attention to cleanliness and sterilization in hospitals than in home isolation. On the other hand, figure 5 shows that the cases of recovery for home isolation are 33% which is less than the cases of recovery for patients who receive their treatment in the hospital. The reason for low recovery in home isolation could be because the patients are not under medical supervision. In addition, figure 6 shows that there are more deaths in home isolation than religious patients who are in hospitals. However, there is a big difference in the mean number of deaths in home isolation and hospitalization as shown by the percentages of 11.12% and 88.88% respectively.

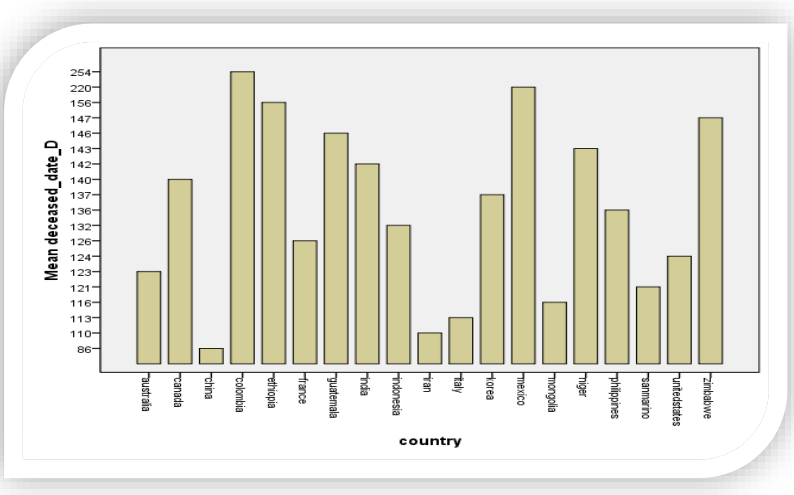


Figure 7. Bar chart for the mean deceased patients and country.

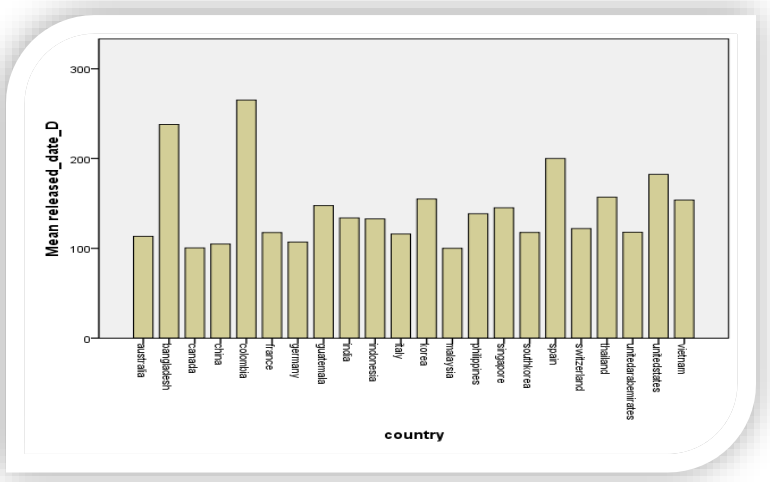


Figure 8. Bar chart for the mean released patients and country.

We carried out comparison of the mean number of deaths among countries or states. Analysis of figure 7 shows that the highest number of deceased patients was recorded in Colombia and the lowest was recorded in China. As well as in figure 8 the highest mean the number of released patients was recorded in Colombia, while the lowest was in recorded in Canada and Malaysia.

Chi-Square Tests

We investigate if there is a link or not between severity of illness and sex. We present our null and alternative hypotheses as follows.

$\boldsymbol{H_0}$ **:** There is no link between severity of illness and gender.

$\boldsymbol{H_1}$ **:** There is a link between severity of illness and gender.

Table 4. Chi-Square test results for severity illness and gender.

sex * severity illness	Value	Statistical significance
Pearson Chi-Square	607.288[a]	.000

From table 4, we have the p-value of Chi- Square test (0.000) <0.05. Thus we reject the null hypothesis and conclude that there is a statistically significant association between severity of illness and sex.

We also research on whether there is a link or not between severity of illness and treatments. The null and alternative hypotheses are as follows:

$\boldsymbol{H_0}$ **:** There is no link between severity of illness and treatments.

$\boldsymbol{H_1}$ **:** There is a link between severity of illness and treatments

Table 5. Chi-Square test results for severity illness and treatments.

Smoking severity illness	Value	Statistical significance
Pearson Chi-Square	607.288[a]	.000

The results from table 5 show a p-value of the Chi- Square test of (0.000) <0.05. Thus we reject the null hypothesis. We conclude that there is a statistically significant association between severity of illness and treatments.

We carried out an analysis on whether or not there is a relationship between the smoking and severity of illness. The null and alternative hypotheses are as follows:

$\boldsymbol{H_0}$ **:** There is no relationship between the smoking and severity of illness.

$\boldsymbol{H_1}$ **:** There is a relationship between the smoking and severity of illness.

Table 6. Chi-Square test results for severity illness and smoking.

Smoking severity illness	Value	Statistical significance
Pearson Chi-Square	66.807[a]	.000

The results from table 6 above of chi-square test show a p-value of the Chi- Square test (0.000) <0.05. The decision is to reject the null hypothesis. We then conclude that there is a statistically significant association between severity of illness and smoking.

Furthermore, we examine whether there exists a statistically significant relationship between smoking cases and deceased, confirmed and released patients in different countries. We present the null and alternative hypotheses as follows:

$\boldsymbol{H_0}$ **:** There is no statistically significant relationship between smoking cases and deceased and released patients in different countries

$\boldsymbol{H_1}$ **:** There is a statistically significant relationship between smoking cases and deceased and released patients in different countries

Table 7. Correlation coefficient for The Number of smoking *and* deceased, confirmed and released patients.

	Correlation coefficient	Statistical significance
Deceased patients	-.032**	.000
Confirmed patients	-.024**	.000
Released patients	-.001**	.000

In conclusion, table 7 shows that the correlation coefficients between smoking and deceased confirmed and released patients in different countries were all negative.
We note that the highest correlation coefficient was in released patients with a value of -0.001**, followed by confirmed patients with a value of -0.24** and the lowest correlation coefficient value was for the deceased patients a value of -0.32**. All the coefficients were statistically significant at the level of significance.

We researched on whether or not there are statistically significant differences between the average number of confirmed male and the average number of confirmed female. The null and alternative hypotheses are as follows:

$\boldsymbol{H_0}$ **:** There are no statistically significant differences between the average number of confirmed cases in males and the average number of confirmed cases in females.

$\boldsymbol{H_1}$ **:** There are statistically significant differences between the average number of confirmed cases in males and the average number of confirmed cases in females.

Table 8. Results for the *t*-test for the for confirmed cases in males and the average number of confirmed cases in females.

Gender	**Mean**	**Std.Deviation**	**t**	**Sig**
Male	223.85	25.903	14.758	.000
Female	221.43	26.807	14.332	.000

It is clear from the data of table 8 that the average number of confirmed cases in males reached a value of (223.85) with a standard deviation (25.903), which is higher than the average number of confirmed cases in females of (221.43) with a standard deviation (26.807). There is a probability characteristic of 1000 (smaller than the level

of significance) and accordingly it was decided that there are statistically significant differences at the 0.05 significance level between the average number of confirmed males and the average number of confirmed female commodities higher in the average.

Furthermore, we examine whether or not there are statistically significant differences between the average number of deceased male and the average number of deceased female. The null and alternative hypotheses are presented as follows:

H_0 : There are no statistically significant differences between the average number of deaths in males and the average number of deaths in females.

H_1 : There are statistically significant differences between the average number of deaths in males and the average number of deaths in females.

Table 9. Results t-test for the average number of deaths in males and the average number of deaths in females.

Gender	Mean	Std.Deviation	t	Sig
Male	220.72	25.839	6.495	.000
Female	217.73	26.811	6.298	.000

It is clear from the data of table 9 that the average number of deceased males reached a value of (220.72) with a standard deviation (25.839), which is higher than the average number of confirmed females of (221.43) with a standard deviation (26.811). There is a probability characteristic of 1000 (smaller than the level of significance) and accordingly it was decided that there are statistically significant differences at the 0.05 significance level between the average number of deceased males and the average number of deceased female commodities higher in the average.

We further examine whether or not there are differences in the average age of patients in different countries. The null and alternative hypotheses are as follows:

H_0 : There are no differences in the average age of patients in different countries and age.

H_1 : There are differences in the average age of patients in different countries and age.

Table 10. Z Test Two-Sample for the average age of patients in different countries.

	age	*deceased*	*age*	*confirmed*	*age*	*released*
Mean	42.11883	229.1867	42.11883	223.8075	42.11883	264.1527
Known Variance	317.599	936.7335	317.599	2489.549	317.599	653.0928
Observations	849866	58063	849866	1039724	849866	290162
Hypothesized Mean Difference	0		0		0	
z	-1456.02		-3453.3		-4334.08	
P(Z<=z) one-tail	0		0		0	
z Critical one-tail	1.644854		1.644854		1.644854	
P(Z<=z) two-tail	0		0		0	
z Critical two-tail	1.959964		1.959964		1.959964	

Table 10 shows the Z test Two-Sample for the number of new cases and male smokers and we observe that there are differences in the average numbers of confirmed patients and age with a Z value of -3459.407547 and a z critical one-tail of 1.959963985. All the coefficients were not statistically significant at the 0.05 level of significance.

We further examine whether or not the numbers of deaths are affected by age, date and smoking in different countries. The null and alternative hypotheses are as follows:

$\boldsymbol{H_0}$: The numbers of deaths are not affected by age, date and smoking in different countries

$\boldsymbol{H_1}$: The numbers of deaths are affected by age, date and smoking in different countries

Table 11. Regression analysis of age smoking and date.

COVID19 deaths	B	R	R Square	F	Sig.
(Constant)	152386.094	.960[a]	.921	162458.334	.000[b]
age	-.008				
smoking	.009				
date	1.105E-5				

In order to know the relationship between the number of COVID19 deaths and the interpreted variables, a multiple regression model was used. The results in table 11 show that the regression model was significant since the value of (F) reporting (0.960^a) in a statistical significance which is smaller than the level of significance (0.05). Also, there is a very strong direct relationship between the dependent variable and the independent variables.

We write the multiple regression equation as follows:

$Y(\text{COVID19 deaths}) = 152386.094 - 0.008(\text{age}) + 0.009(\text{smoking}) + 0.00001105(\text{date})$

Suppose we want to predict the number of deaths in the month of October for the age of 70 years and the patient who doesn`t smoke then, we obtain:

$Y(\text{COVID19 deaths}) = 152386.094 - 0.008(70) + 0.009(0) + 0.00001105(10) = 152375$

Furthermore, we examine whether or not the numbers of cases are affected by age, date and smoking in different countries. The null and alternative hypotheses are presented as follows:

$\boldsymbol{H_0}$ **:** The numbers of cases are not affected by age, date and smoking in different countries

$\boldsymbol{H_1}$ **:** The numbers of cases are affected by the age, date and smoking in different countries

Table 12. Regression analysis of age smoking and date

COVID19 cases	B	R	R Square	F	Sig.
(Constant)	159629.000	1.000[a]	1.000	379076056126440580.000	.000[b]
age	-4.822E-14				
smoking	1.568E-12				
date	1.157E-5				

In order to know the relationship between the number of COVID19 cases and the interpreted variables, a multiple regression model was used. Analysis of the results in Table 12 showed that the regression model was significant since the value of (F) reporting (1.000) in a statistical significance which is smaller than the level of significance (0.05). Also, there is a very strong direct relationship between the dependent variable and the independent variables.

We write the multiple regression equation as follows:

Y (COVID19 cases)= 159629.000-4.822E-14 (age)+ 1.568E-12 (smoking)+ 1.157E-5 (date) (12)

Suppose we want to predict the number of cases in the month of October, for the age of 70 years and the patient who doesn`t smoke then, we obtain:

Y(COVID19 cases)=159629.000-4.822E *70+ 1.568E *0+ 1.105E-5*10 =159955

We further examine whether or not we can predict the number of deaths through age, symptoms, smoking, treatment and gender. The forecasting testing is as follows:

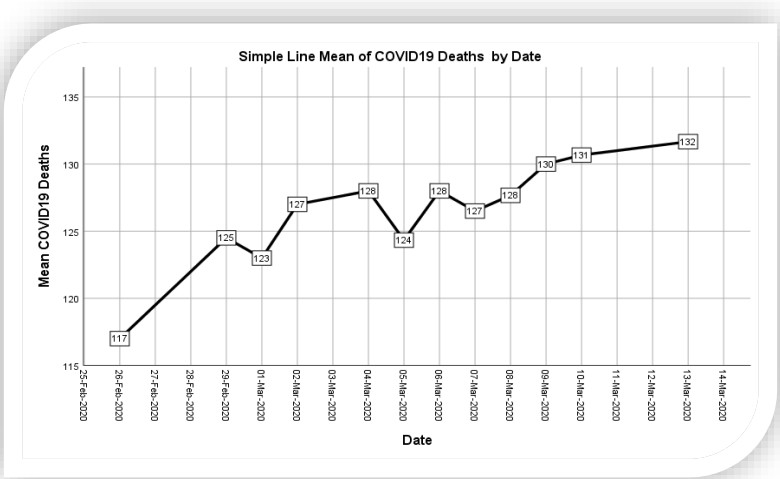


Figure 9. Line chart for the average number of deaths by date.

The previous figure shows that the number of deaths on 26 February was 117 people, and within three days, it increased to 125. At the beginning of the month of March, the number of deaths slightly decreased to 123 then rose again to 127, meaning there is a slight decrease in the number of deaths from time to time.

Table 13. Model Statistics of deaths through age, symptoms, smoking, treatment and gender.

Model	Number of Predictors	Stationary R-squared	Statistics	Sig.
COVID19 Deaths - Model_1	0	-1.522		.000..

From table 13, we note and conclude that we cannot predict the number of deaths using age, symptoms, smoking, treatment and gender.

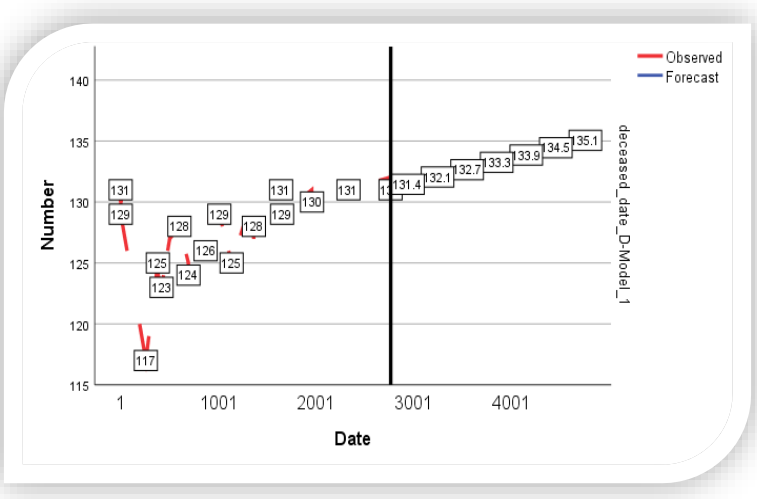


Figure 10. Line chart for forecast for death's apparel shows how you can automatically determine which model best fits your time series and independent variables.

Also, figure 10 shows us that we cannot predict the number of deaths because the line for forecasts does not appear. Only the observed data appears separately, but if we want to predict the number of deaths we need population, age, confirmed cases and other numerical variables.

We further examine whether or not the numbers of deaths are affected by smoking, severity of illness, infectious person, gender, treatment and symptoms in different countries. The Bayesian Inference is as follows:

$\boldsymbol{H_0}$ **:** The numbers of deaths are not affected by smoking, severity of illness, infectious person, gender, treatment and symptoms in different countries

$\boldsymbol{H_1}$ **:** The numbers of deaths are affected by smoking, severity of illness, infectious person, gender, treatment and symptoms in different countries

Table 14. Bayesian Inference for the numbers of deaths affected by smoking, severity of illness, infectious person, gender, treatment and symptoms.

Source	Sum of Squares	Mean Square	F	Sig	95% Credible Interval	
Regression					Lower	Upper
Residual	59920461.714	59920461.714	13.504	.000	4286454.874	4688031.891
Source	4437288.125	4437288.125				

From our analysis of the results, we note that p = 0.000, the mean is 4482810.973, and CIs is equal to [4286454.874, 4688031.891]. So in a Frequentist approach, we reject the null hypothesis of no difference in the mean. The probability of finding a difference of this or larger magnitude is 0%. The CIs tell us that 95% of the COVID-19 deaths are greatly affected by these independent variables.

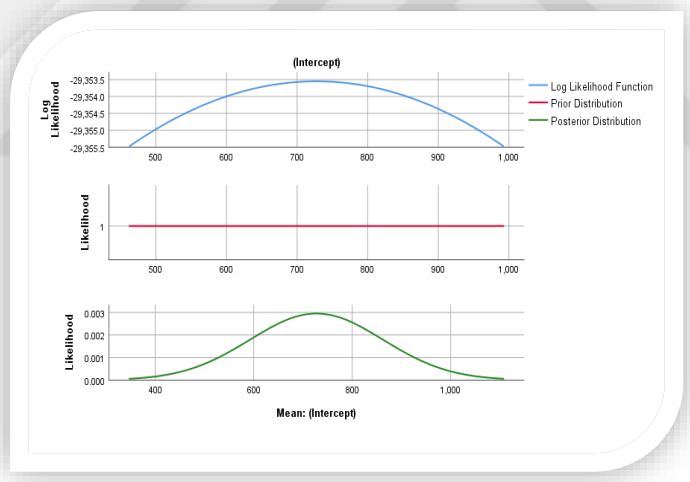


Figure 11. Non-Informative Prior Distribution, Distribution of Observed Data and Posterior Distribution for intercept.

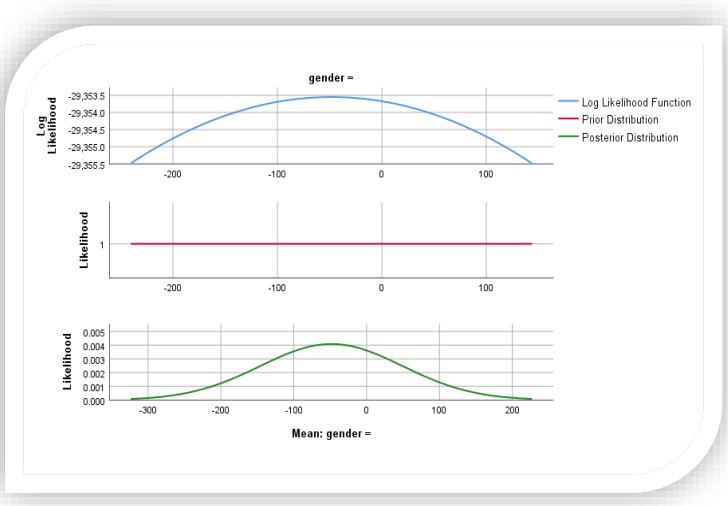


Figure 12. Non-Informative Prior Distribution, Distribution of Observed Data and Posterior Distribution for gender.

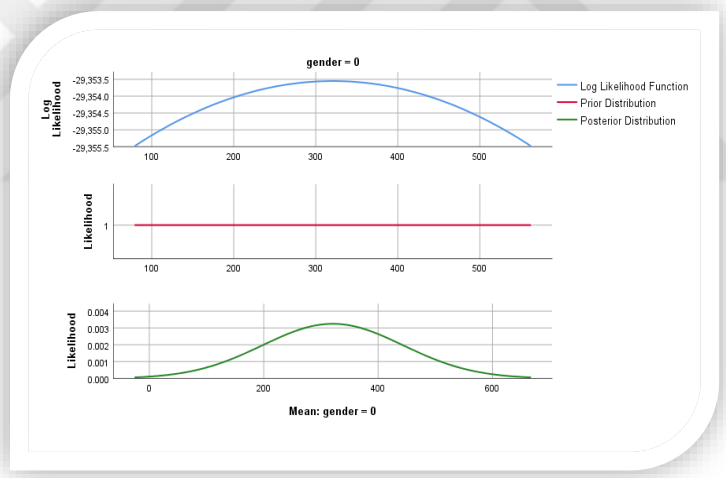


Figure 13. Non-Informative Prior Distribution, Distribution of Observed Data and Posterior Distribution for gender.

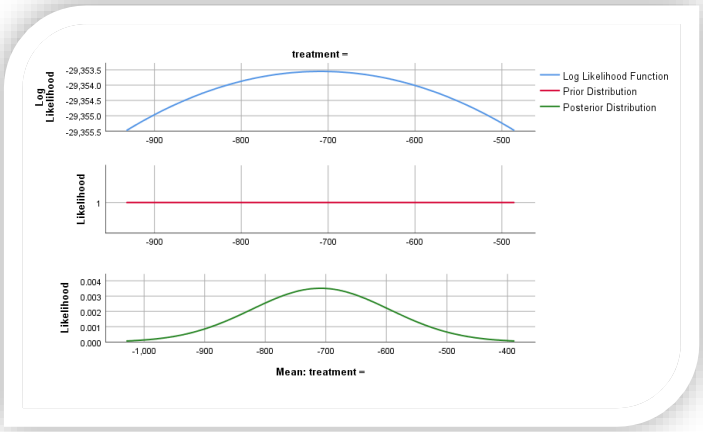


Figure 14. Non-Informative Prior Distribution, Distribution of Observed Data and Posterior Distribution for treatment.

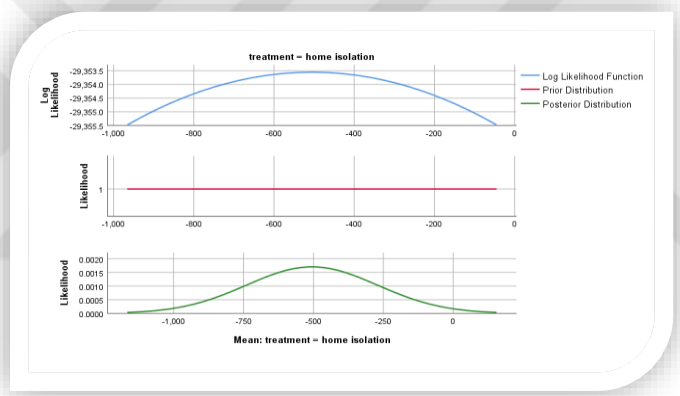


Figure 15. Non-Informative Prior Distribution, Distribution of Observed Data and Posterior Distribution for treatment.

Figures 11,12,13,14 and 15 illustrate the contrast between the three steps of Bayesian analysis if the Prior is a non-informative distribution. We can see that the Prior distribution is rectangular, the observed data distribution (the middle histogram) as well as the Posterior distribution (bottom histogram) is approximately normal.

And also the figures tell us that the highest number of deaths is among males and also in home isolation. Therefore, we recommend that treatment in hospitalized be taken because the females have more immunity.

Table 15. Bayesian Estimates of Coefficients[a,b,c.]

Parameter	Posterior			95% Credible Interval	
	Mode	Mean	Variance	Lower Bound	Upper Bound
(Intercept)	727.424	727.424	18268.640	462.498	992.351
smoking	.[d]	.[d]	.[d]	.[d]	.[d]
severity_illness_infectious_person	.[d]	.[d]	.[d]	.[d]	.[d]
gender =	-48.353	-48.353	9539.927	-239.798	143.092
gender = 0	320.698	320.698	15112.310	79.742	561.654
gender = 1	.[d]	.[d]	.[d]	.[d]	.[d]
treatment =	-708.890	-708.890	12921.196	-931.694	-486.085
treatment = home isolation	-505.348	-505.348	54894.805	-964.586	-46.111
treatment = hospitalized	.[d]	.[d]	.[d]	.[d]	.[d]
symptoms	.[d]	.[d]	.[d]	.[d]	.[d]

The table above shows that COVID-19 deaths are greatly affected by these independent variables. The number of deaths for males is higher than that of the females. Also, treatment in home isolation results in more deaths than other treatment methods. It is also clear from the previous table that the missing values indicate that they have less influence on the independent variable.

4.2 Influenza Dataset

The data I used it started from (1/02/2020 to 20/06/2020) and it contains 14 columns and 4762 observations.

I chose this dataset because it contains a lot of important variables that are useful in the subject under study and comparing it with the data of the Corona virus, given the similarities between them in several aspects.

You can access the CSV file for the complete dataset on Kaggle (2020).

Table 16. Explain variables for Influenza Dataset.

categorical data		value data	
name col	explain	name col	Value
Country		start week	Date
Sex			
Male			

Female			
Age group	4	end week	Date
Under 1 year	0	COVID-19 Deaths	discrete
1-4 years	1	Total Deaths	discrete
5-14 years	2	Pneumonia Deaths	discrete
15-24 years	3	confirmed_date_until_released_date	discrete
25-34 years	4	confirmed_date_until_deceased_date	discrete
35-44 years	5	Pneumonia and COVID-19 Deaths	discrete
45-54 years	6	Influenza Deaths	discrete
55-64 years	7	Pneumonia, Influenza, or COVID-19 Deaths	discrete
65-74 years	8		
75-84 years	9		
85 years and over	10		

4.2.1 Analysis of the Influenza Dataset

Table 17. Descriptive Statistics for Influenza Dataset.

	N	Maximum	Mean	Std. Deviation
COVID-19 Deaths	3839	4.4940E6	1282.533420	7.2560046E4
Total Deaths	3187	7.8850E8	2.489376E5	1.3967160E7
Pneumonia Deaths	3637	6.0798E6	1800.807294	1.0084189E5
Pneumonia and COVID-19 Deaths	3994	790733.3650	244.292358	1.2542792E4
Influenza Deaths	4013	16291.9053	10.045575	287.0135243

The Table 17 shows us that the highest value in mean deaths due to COVID – 19 is 1282.533420. Also, the number of COVID -19 deaths is ten times more than the number of influenza deaths.

For the case of influenza, we determine which age group has the highest percentage of Influenza deaths.

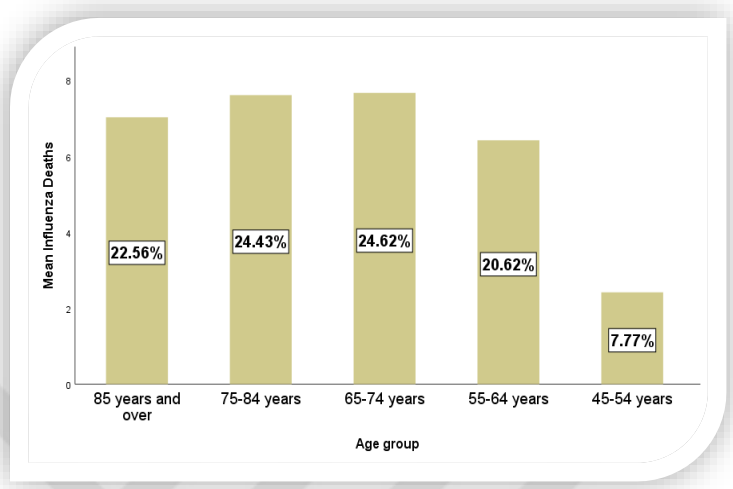


Figure 16. Bar chart for age group and Influenza Deaths.

In the data analysis of figure 16 we note that the highest value of influenza deaths was recorded in the age group of 65-74 years, followed by the age group of 75-84 years with percentages 24.62%, 24.43% respectively. On the other hand, the lowest value of influenza deaths was in the age group 45-54 years and this could be because they have a lower percentage of being infected with this virus.

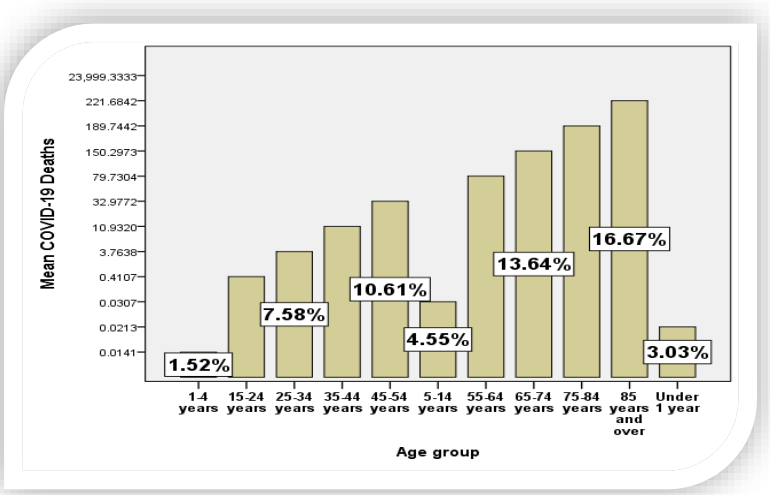


Figure 17. Bar chart for age group and COVID-19 Deaths.

In the analysis of figure 17, we also note that the highest mortality rate was recorded under the group of all ages, followed by the age group 85 years and above and next was the group 75-84 years. Moreover, the lowest one was in the age group of 1-4 years. Additionally, we note that the higher the age group, the higher the death rate in COVID-19, except for the age group of 1-4 years.

We determine which the country has more COVID-19 Deaths, Influenza Deaths and Pneumonia Deaths.

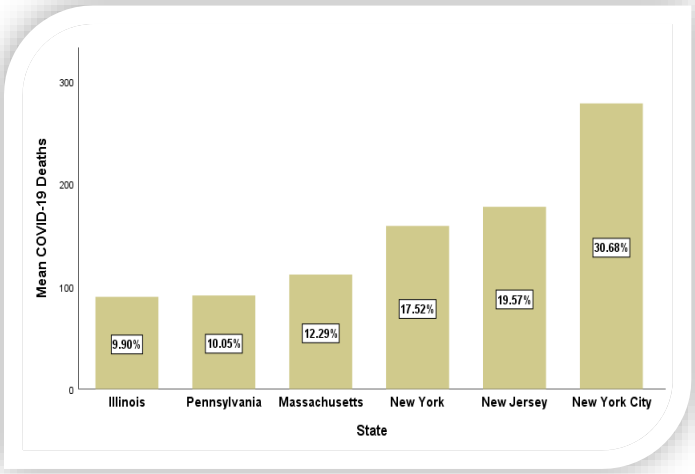


Figure 18. Bar chart for mean the COVID-19 Deaths by state.

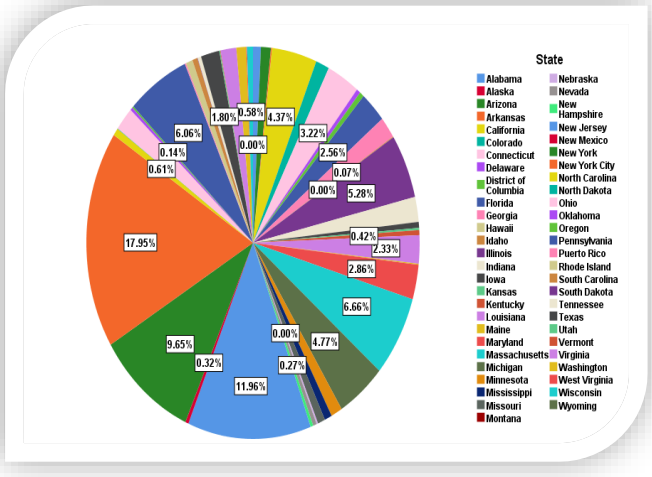


Figure 19. Pie chart for mean the COVID-19 Deaths by state.

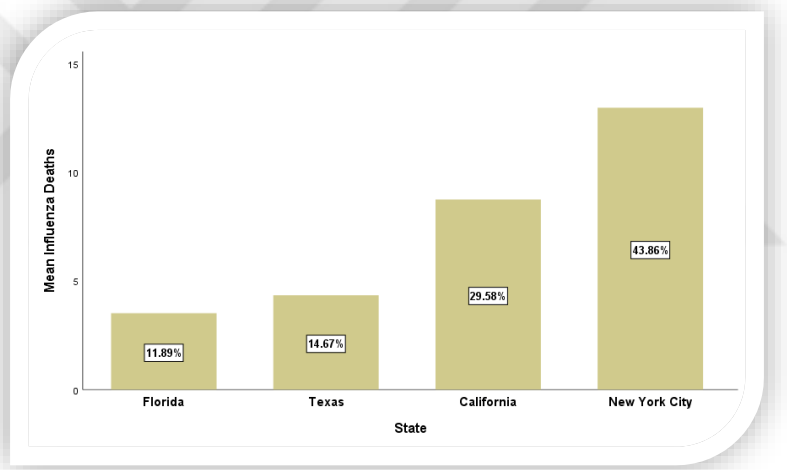


Figure 20. Bar chart for the mean Influenza Deaths by state.

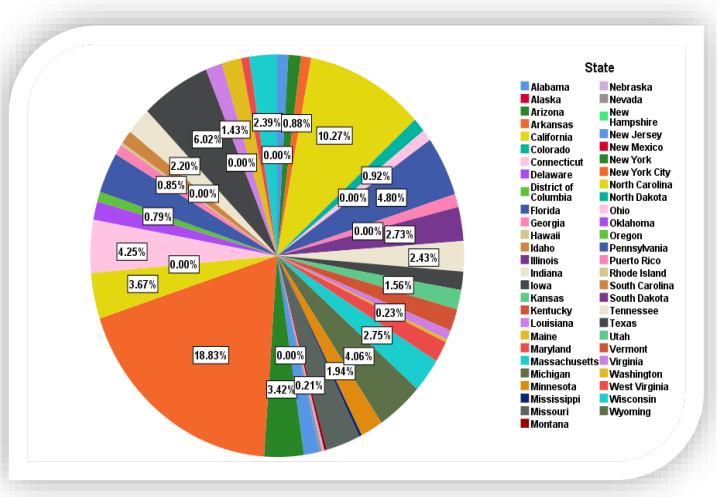


Figure 21. Pie chart for mean the Influenza Deaths by state.

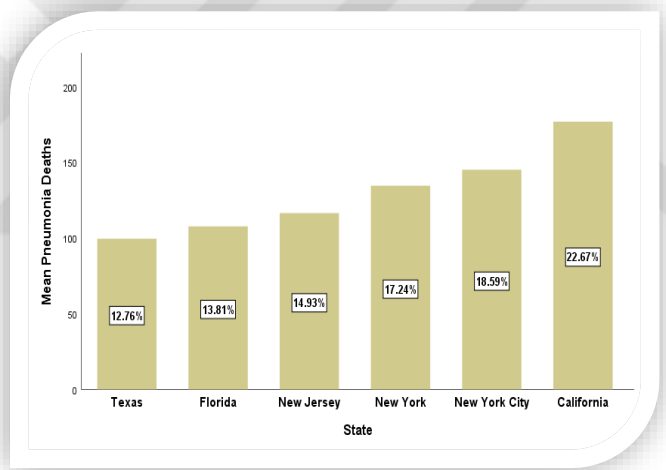


Figure 22. Bar chart for the mean Pneumonia Deaths by state.

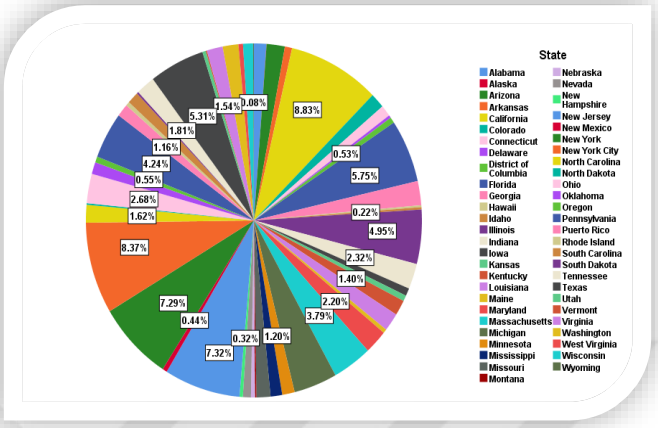


Figure 23. Pie chart for mean the Pneumonia Deaths by state.

The graphs shown in Figures above indicate that the New York City had the highest number of COVID-19 deaths and influenza deaths. California had the highest number of Pneumonia deaths. On the other hand, Idaho, Hawaii and Alabama had the least number of Influenza, Pneumonia and COVID-19 deaths respectively.

Table 18. One-Sample Kolmogorov-Smirnov Test.

	Probability value	Sig.
COVID-19 Deaths	.493	.000
Pneumonia Deaths	.494	.000
Total Deaths	.503	.000
Pneumonia and COVID-19 Deaths	.492	.000
Influenza Deaths	.486	.000
Pneumonia, Influenza, or COVID-19 Deaths	.496	.000

The previous table 18 shows the Kolmogorov-Smirnov Test for - COVID-19 Deaths, Pneumonia Deaths, Total Deaths, Pneumonia and COVID-19 Deaths and Pneumonia,

Influenza, or COVID-19 Deaths. Test distribution is not Normal.

We carry out an analysis to determine whether or not there is a statistically significant relationship between COVID-19 Deaths and Influenza Deaths in different states. The null and alternative hypotheses are as follows:

$\boldsymbol{H_0}$ **:** There is no statistically significant relationship between COVID-19 Deaths and Influenza Deaths in different countries.

$\boldsymbol{H_1}$ **:** There is a statistically significant relationship between COVID-19 Deaths and Influenza Deaths in different countries.

Table 19. Correlation coefficient for the Number of COVID-19 Deaths and Influenza Deaths.

	Correlation coefficient	Statistical significance
Influenza Deaths	.985**	.000

In conclusion**,** table 19 shows that there was a strong correlation between the number of COVID-19 and Influenza deaths, with a correlation coefficient of the number of Influenza Deaths of 0.985**. The coefficient was statistically significant at the level of significance.

We further research on whether or not there is a statistically significant relationship between COVID-19 Deaths and Pneumonia Deaths in different states. The null and alternative hypotheses are as follows:

$\boldsymbol{H_0}$ **:** There is no statistically significant relationship between COVID-19 Deaths and Pneumonia Deaths in different states

$\boldsymbol{H_1}$ **:** There is a statistically significant relationship between COVID-19 Deaths and Pneumonia Deaths in different states.

Table 20. Correlation coefficient for the Number of COVID-19 Deaths and Pneumonia Deaths

	Correlation coefficient	Statistical significance
Pneumonia Deaths	.992**	.000

In conclusion, table 20 shows the correlation coefficients between COVID-19 Deaths and Pneumonia Deaths in different states, where the correlation was strong directly and the correlation coefficient of the number of Pneumonia Deaths is 0.992**. The coefficient was statistically significant at the level of significance.

We further determine whether or not there are differences in the average numbers of COVID-19 Deaths and Influenza Deaths. The null and alternative hypotheses are as follows:

$\boldsymbol{H_0}$ **:** There is no difference in the average numbers of COVID-19 Deaths and Influenza Deaths

$\boldsymbol{H_1}$ **:** There is a difference in the average numbers of COVID-19 Deaths and Influenza Deaths.

Table 21. Correlation coefficient for The Number of COVID-19 Deaths and Influenza Deaths.

	COVID-19 Deaths	*Influenza Deaths*
Mean	111.9478759	5.987284966
Known Variance	4493989.882	16291.90532
Observations	3837	4011
Hypothesized Mean Difference	0	
Z	3.090811096	
P(Z<=z) one-tail	0.000998053	
z Critical one-tail	1.644853627	
P(Z<=z) two-tail	111.9478759	

From our analysis of Table 21, we conclude that there is a difference in the average numbers of COVID-19 Deaths and Influenza Deaths with a Z value of 3.090811096 and a z Critical two-tail of 1.959963985. The coefficients were statistically significant at the 0.05 level of significance.

We further examine whether there are differences in the average numbers of COVID-19 Deaths and Influenza Deaths. The null and alternative hypotheses are as follows:

$\boldsymbol{H_0}$ **:** There are no differences in the average numbers of COVID-19 Deaths and Pneumonia Deaths

$\boldsymbol{H_1}$ **:** There are differences in the average numbers of COVID-19 Deaths and Pneumonia Deaths

Table 22. Z-test Two-Sample for The Number of average number of COVID-19 Deaths and Pneumonia Deaths.

	COVID-19 Deaths	*Pneumonia Deaths*
Mean	111.9478759	129.1796424
Known Variance	4493989.882	6079838.985
Observations	3837	3635
Hypothesized Mean Difference	0	
Z	-0.323131783	
P(Z<=z) one-tail	0.373297721	
z Critical one-tail	1.644853627	
P(Z<=z) two-tail	0.746595441	

Analysis of table 22 indicates that there are differences in the average numbers of COVID-19 Deaths and Pneumonia Deaths with a Z value of -0.323131783 and a z Critical one-tail of 1.644853627. The coefficients were statistically significant at the 0.05 level of significance.

Finally, we determine whether or not the numbers of Influenza Deaths are affected by age group and gender. The null and alternative hypotheses are as follows:

$\boldsymbol{H_0}$: The average numbers of Influenza Deaths are not affected by age group and gender.

$\boldsymbol{H_1}$: The average numbers of Influenza Deaths are affected by age group and gender.

Table 23. Bayesian Inference for the average numbers of Influenza deaths affected by age group and gender.

Source	Sum of Squares	Mean Square	F	Sig	95% Credible Interval	
Regression					Lower	Upper
Residual	18435830.363	398805676.467	120.843	.000	11237.362	12267.470
Source	46894674.132	3153966.211				

From our analysis of the results, we note that p = 0.000, the mean is 1418140.797 and CIs is equal to [11237.362, 12267.470]. So in a Frequentist approach, we reject the null hypothesis of no difference in the mean. The probability of finding a difference of this or larger magnitude is 0%. The CIs tell us that 95% of the Influenza deaths are greatly affected by these independent variables.

Table 24. Bayesian Estimates of Coefficients[a, b,c].

Bayesian Estimates of Coefficients[a,b,c]					
	Posterior			95% Credible Interval	
Parameter	Mode	Mean	Variance	Lower Bound	Upper Bound
(Intercept)	8.039	8.039	44.157	-4.986	21.063
Age group = 1-4 years	-.160	-.160	57.383	-15.008	14.688
Age group = 15-24 years	-.465	-.465	57.819	-15.369	14.439
Age group = 25-34 years	.238	.238	61.262	-15.104	15.579
Age group = 35-44 years	.548	.548	62.630	-14.964	16.060
Age group = 45-54 years	1.989	1.989	62.576	-13.516	17.494
Age group = 5-14 years	-.029	-.029	57.744	-14.924	14.865
Age group = 55-64 years	6.444	6.444	63.562	-9.183	22.071
Age group = 65-74 years	7.508	7.508	63.781	-8.145	23.162
Age group = 75-84 years	7.365	7.365	62.035	-8.073	22.803

Age group = 85 years and over	6.921	6.921	61.734	-8.480	22.321
Age group = All Ages	1427.921	1427.921	1340.277	1356.163	1499.679
Age group = Under 1 year	.[d]	.[d]	.[d]	.[d]	.[d]
sex =	-8.731	-8.731	22.684	-18.067	.604
sex = female	-11.878	-11.878	37.661	-23.907	.150
sex = male	.[d]	.[d]	.[d]	.[d]	.[d]

The table above shows that Influenza deaths are greatly affected by these independent variables. The highest number of deaths was recorded for females. Also, the treatment of children under one year, results in less deaths than other age group methods.

It is also clear from the previous table that the missing values indicate that they have less influence on the independent variable.

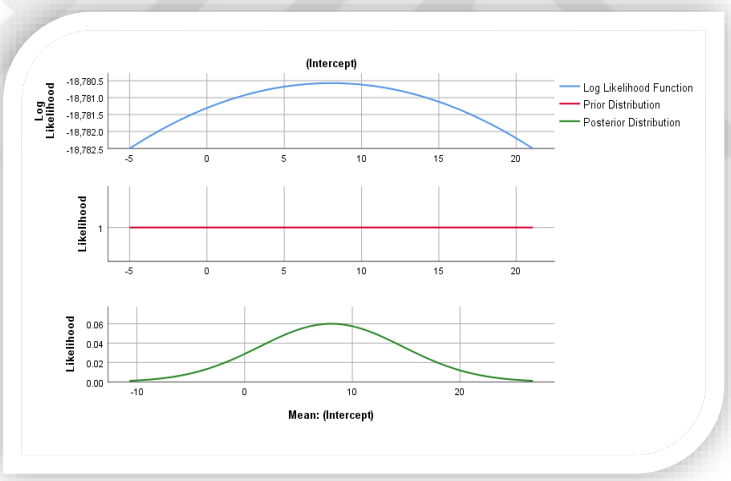


Figure 24. Non-Informative Prior Distribution, Distribution of Observed Data and Posterior Distribution for intercept.

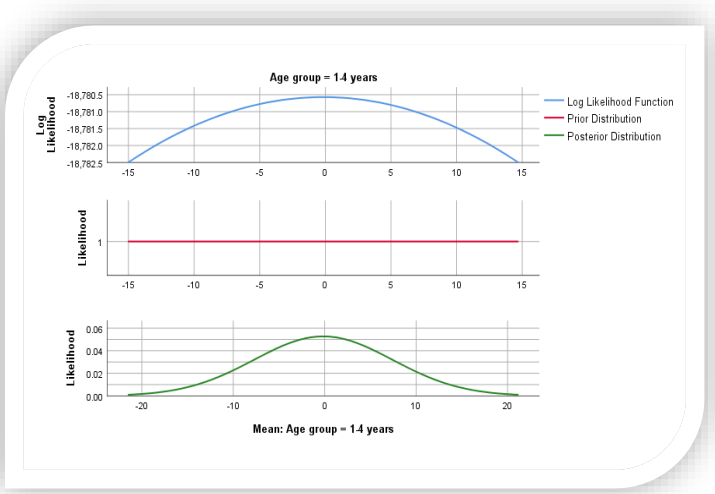


Figure 25. Non-Informative Prior Distribution, Distribution of Observed Data and Posterior Distribution for age group: 1-4 years.

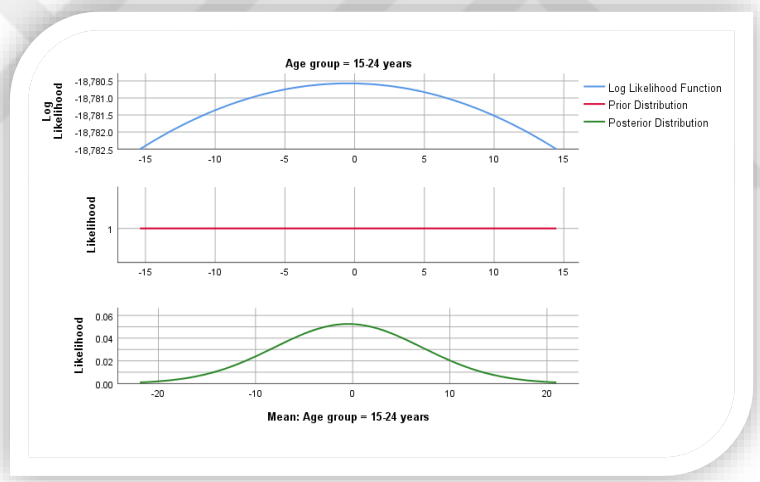


Figure 26. Non-Informative Prior Distribution, Distribution of Observed Data and Posterior Distribution for age group: 15-25 years.

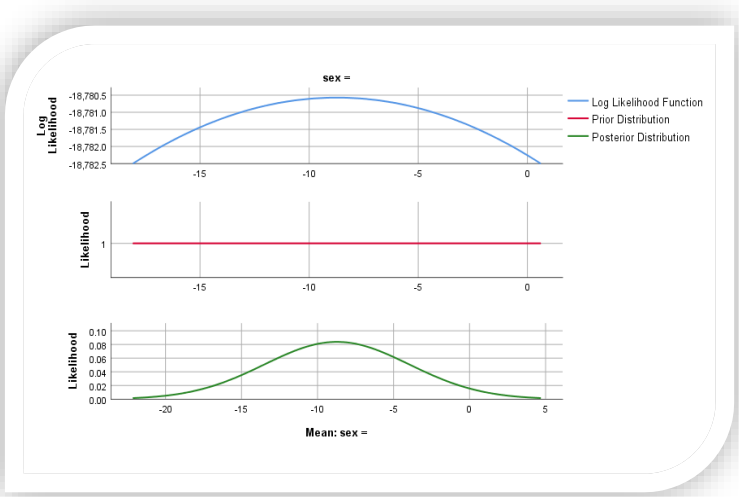


Figure 27. Non-Informative Prior Distribution, Distribution of Observed Data and Posterior Distribution for gender.

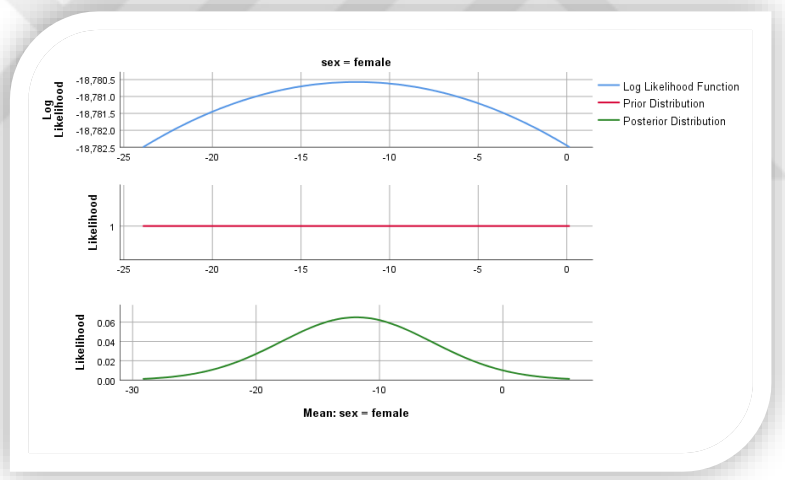


Figure 28. Non-Informative Prior Distribution, Distribution of Observed Data and Posterior Distribution for female.

Figures 22, 26, 27 and 28 illustrate the contrast between the three steps of Bayesian analysis if the Prior is a non-informative distribution. We can see that the Prior distribution is rectangular; the observed data distribution (the middle histogram) as well as the Posterior distribution (bottom histogram) is approximately normal.

And also the figures tell us which of them have the highest number of deaths. We observe that the females and children under one year of age have the least number of

deaths.

Therefore, we recommend that age group of above one year old be taken because the males have more immunity.

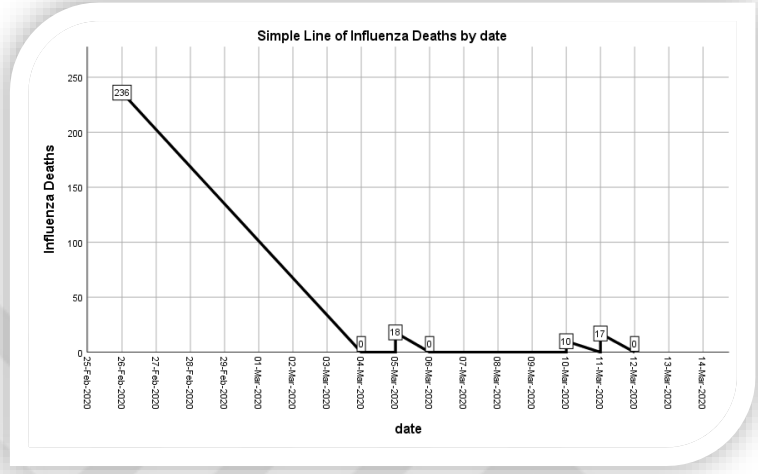


Figure 29. Line chart for the average number of influenza deaths by date

As shown in figure 29, the number of influenza deaths increased dramatically on January 25, then decreased significantly on 4 mar, from 250 to 0, after that it rose slightly, then stabilized from 6 to 10, and then began a small rise then a decrease.

CHAPTER 5: CONCLUSION

In this study we studied many models such as correlation, Z test and regression model, to estimate hypothesis test. We used the SPSS statistical program for analysing and answering some of the research questions. Then, we carried out data analysis for COVID-19. In this chapter we give a summary of our research findings.

Our results show that females have a higher chance of being infected with COVID-19 than males. The Figure 2 shows us that vast majority of treatment is for hospitalized patients, it is equal to 89%, and however the treatment for home isolation is 11%. From Figure 3 we see that 55.75% were cured, it is the greatest percentage, 31.97% were good (recovered), 10.27% were deceased, 2.80% respectively. These results mean that more than half patients recovered and we feel sorry because the patients who died were more than patients who were critically ill.

In addition, figures 4, 5 and 6 show us that confirmed cases in home isolation were more than in hospitals, as indicated by the percentages of 77.97%, 22.03% respectively. This means that more attention to cleanliness and sterilization is paid in hospitals than in home isolation. Also, the cases of recovery for home isolation are 0.33% which is less than the recovery cases of patients who receive their treatment in the hospitals, perhaps because they are not under medical supervision. Moreover, our analysis shows that there are more deaths in home isolation than religious patients who are in hospitals. However, there is a big difference as shown by the percentages of 11.12% and 88.88% respectively.
Also, the graph 7 shows the highest number of deceased was in the Colombia and the lowest one was in the China. Moreover, in figure 8 the highest mean number of released patients was in Colombia while the lowest was in Canada and Malaysia.

From our analysis we conclude that there is a statistically significant association between severity of illness and sex. Moreover, we conclude that there is a statistically significant association between severity of illness and treatments. We also conclude that there is a statistically significant association between severity of illness and smoking.

Table 7 shows that the correlation coefficients between smoking and deceased patients and also between confirmed and released patients in different countries were all negative. All the coefficients were statistically significant at the level of significance. We concluded that there are statistically significant differences at the 0.05 significance level between the average number of confirmed male cases and the average number of confirmed female cases higher in the average.

Our analysis indicates that there are differences in the average numbers of confirmed patients and age. The coefficients were not statistically significant at the 0.05 level of significance. Also we note that there are differences in the averages of released patients and age.

There is a very strong direct relationship between COVID-19 deaths and age, smoking and date in different countries. The relationship is illustrated by the multiple regression equation as follows:

Y (COVID19 deaths)= 152386.094-0.008(age)+ 0.009(smoking)+ 0.00001105(date)

The above relationship can be used to forecast the future values of COVID-19 deaths. There is a very strong direct relationship between COVID-19 cases and age, smoking and date. Thus, the number of cases is affected by age, smoking and date. We illustrated the relationship by the multiple regression equation as follows:

Y (COVID19 cases)= 159629.000-4.822E-14 (age)+ 1.568E-12 (smoking)+ 1.157E-5 (date) (12)

From table 14, we note and conclude that we cannot predict the number of deaths using age, symptoms, smoking, treatment and gender. We obtained the CIs which tell us that 95% of the COVID-19 deaths are greatly affected by smoking, severity of illness, infectious person, gender, treatment and symptoms in different countries. In addition, we also obtained the CIs which tell us that 95% of the Influenza deaths are greatly affected by age group and gender.

Figures 11,12,13,14 and 15 illustrate the contrast between the three steps of Bayesian analysis if the Prior is a non-informative distribution. Also, the figures tell us which of them have less number of deaths, for females and treatment in home isolation. Our Bayesian analysis of Table 15 also shows that COVID-19 deaths are greatly affected by these independent variables. The gender with the highest number of deaths is males.

Additionally, treatment in home isolation results in more deaths than treatment in hospitalization.

From our analysis, we also conclude that the number of COVID-19 deaths is ten times more than the number of influenza deaths. From our analysis of Figure 16 we note that the highest value of influenza deaths was recorded in age group of 65-74 years followed by the age group of 75-84 years with the percentages of 24.62% and 24.43% respectively. On the other hand, the lowest rate was in the age group 45-54 years and this might be, since they have a lower percentage of being infected with this virus.
On the other hand, we note that the highest mortality rate for COVID-19 was recorded under the group of all ages and the age groups of 75-84 years and 65-74 years have the same percentage of 0.52%. On the other hand, the lowest mortality rate was in children under 1 year and this could be because they have a lower percentage of being infected with this virus.

Furthermore, the graphs 18, 19, 20, 21, 22 and 23 shows that the highest numbers of COVID-19 Deaths and Influenza Deaths were recorded in New York City. The highest number of Pneumonia Deaths was recorded in California. On the other hand, Idaho, Hawaii and Alabama had the lowest numbers of Influenza, Pneumonia and COVID-19 deaths respectively.

Our results indicate that there is a slight decrease in the number of COVID-19 deaths from time to time whereas the number of influenza deaths increased dramatically on January 25, then decreased significantly on 4 March, from 250 to 0 and after that it rose again slightly, then stabilized from 6 to 10, and then began a small rise then a decrease. In both cases (COVID-19 and Influenza), we observe that we cannot predict the number of deaths by date only.

Our results indicate that there is a statistically significant relationship between COVID-19 Deaths and Influenza Deaths in different countries. We conclude that Influenza deaths are greatly affected by age group and gender**.** On the other hand, COVID-19 deaths are greatly affected by smoking, severity of illness, infectious person, gender, treatment and symptoms.

Also, we conclude that there is a statistically significant relationship between COVID-19 Deaths and Pneumonia Deaths in different states. We also conclude that there is a difference in the average numbers of COVID-19 Deaths and Influenza Deaths. Finally, our results indicate that there are differences in the average numbers of COVID-19 Deaths and Pneumonia Deaths.

Overall, our study managed to fulfil its aim and objectives and also it came up with answers for all the research questions.

REFERENCES

Arunachalam, R., Paulkumar, K., and Annadurai, G. (2012). *Phylogenetic analysis of pandemic influenza A/H1N1 virus*. Biologia, 67(1), pp. 14-31.

BBC (2020). *Symptoms of China Coronavirus chart.* [Online] Available at: https://images.app.goo.gl/mDrXZ9fmDGuYFBeZA . (Accessed 20/08/2020)

Black, A., Liu, D., and Mitchell, L. (2020). *How to flatten the curve of coronavirus, a mathematician explains.* [Online] Available at: https://theconversation.com/how-to-flatten-the-curve-of-coronavirus-a-mathematician-explains-133514. (Accessed 1/12/2020)

Box, G. E. (1954). *Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification.* The annals of mathematical statistics, 25(2), pp. 290-302.

Burnet, F. M. (1979). *Portraits of viruses: influenza virus A.* Intervirology, 11(4), pp. 201-214.

Byjus (2019). *Making a pie chart.* [Online] Available at: https://byjus.com/maths/pie-chart/#formula. (Accessed 13/11/2020)

Chaurasia, V. and Pal, S. (2020). *COVID‑19 Pandemic: ARIMA and Regression Model‑Based Worldwide Death Cases Predictions.* SN Computer Science 1:288.

Choi, S. B., Kim, J. and Ahn, I. (2019). *Forecasting type-specific seasonal influenza after 26 weeks in the United States using influenza activities in other countries.* PLoS One 14(11).

Chu, D. K., Pan, Y., Cheng, S. M., Hui, K. P., Krishnan, P., Liu, Y., Ng, D. Y. M., Wan, C. K.C., Yang, P., Wang, Q., Peiris, M. and Poon, L. L. M. (2020). *Molecular diagnosis of a novel coronavirus (2019-nCoV) causing an outbreak of pneumonia.* Clinical chemistry, 66(4), pp. 549-555.

Collins, S. D. (1957). *Influenza in the United States, 1887–1956.* Public Health Monograph, 48, pp. 51-73.

Crosby, A. W. (2003). *America's forgotten pandemic: the influenza of 1918*. 2nd edition, Cambridge University Press.

Darwish, A., Rahhal, Y. and Jafar, A. (2020). *A comparative study on predicting influenza outbreaks using different feature spaces: application of influenza-like*

illness data from Early Warning Alert and Response System in Syria. BMC Research Notes 13(33). doi: https://doi.org/10.1186/s13104-020-4889-5

Data.world (2020). *Provisional COVID-19 death counts by week.* [Online] Available at: https://data.world/us-hhs-gov/8b634eb1-fb4f-439d-99f0-bc392fac4f19 .(Accessed 8/09/2020)

Du, R. H., Liang, L. R., Yang, C. Q., Wang, W., Cao, T. Z., Li, M., Guo, G. Y., Du, J., Zheng, C., Zhu, Q., Hu, M., Li, X., Peng, P. and Shi, H. (2020). *Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study.* European Respiratory Journal, 55(5). doi: 10.1183/13993003.00524-2020

Fonseca, S., Rivas, I., Romaguera, D., Quijal, M., Czarlewski, W., Vidal, A., Fonseca, J. A., Ballester, J., Anto, J. M., Basagana, X., Cunha, L. M. and Bousquet, J. (2020). *Association between consumption of fermented vegetables and COVID-19 mortality at a country level in Europe.* MedRxiv. doi: https://doi.org/10.1101/2020.07.06.20147025.

Free online courses, Penn State Eberly College of Science. (2020). *Simple linear regression.* [Online] Available at: https://online.stat.psu.edu/stat501/lesson/1 . (Accessed 2/10/2020)

Free online courses, Wall Street mojo. (2020). *Differences between z-test and t-test.* [Online] Available at: www.wallstreetmojo.com/z-test-vs-t-test . (Accessed 27/11/2020)

Furtuna, N., Druc, A., Sajin, O., Spinu, C., Gutu, V. and Ceban, A. (2020). *Epidemiology of the initial period of novel coronavirus (COVID-19) pandemic in the Republic of Moldova.* One Health and Risk Management, 2(2). doi: https://doi.org/10.38045/ohrm.2020.1.11

Ghosal, S., Sengupta, S., Majumder, M. and Sinha, B. (2020). *Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020).* Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14, pp. 311-315.

Giugliano, F. (2020). *Greece shows how to handle the crisis.* Bloomberg Opinion.

Google (2020). *Details on the data set.* [Online] Available at: https://docs.google.com/spreadsheets/d/1awEY-04UK8wibkbZ1qfV6a-Q9YKScfP7qiAtWDsp9Jw/edit?usp=sharing .(Accessed 27/11/2020)

IBM SPSS Forecasting. (2020). *IBM Software Business Analytics*

Jargin, S. V. (2020). COVID-19: *Economic damage is a health risk.* American journal of preventive medicine, 6(3), pp. 62-64.

Johnson, N. P. and Mueller, J. (2002). *Updating the accounts: global mortality of the 1918-1920" Spanish" influenza pandemic.* Bulletin of the History of Medicine, pp. 105-115.

Kaggle (2020). *COVID-19 focus patients dataset.* [Online] Available at: https://www.kaggle.com/shirmani/characteristics-corona-patients. (Accessed 2/10/2020)

Kim, T. K. (2015). *T test as a parametric statistic.* Korean journal of anesthesiology, 68(6), pp. 540.

Lee, D. K., In, J. and Lee, S. (2015). *Standard deviation and standard error of the mean.* Korean journal of anesthesiology, 68(3), 220.

Lemon, S. M., Mahmoud, A., Mack, A. and Knobler, S. L. (Eds.). (2005). *The threat of pandemic influenza: are we ready? Workshop summary*. National Academies Press.

Livadiotis, G. (2020). *Statistical analysis of the impact of environmental temperature on the exponential growth rate of cases infected by COVID-19.* Plos one, 15(5), e0233875.

Lumley, T., Diehr, P., Emerson, S. and Chen, L. (2002). *The importance of the normality assumption in large public health data sets.* Annual review of public health, 23(1), pp. 151-169.

McHugh, M. L. (2013). *The chi-square test of independence.* Biochemia medica: Biochemia medica, 23(2), pp. 143-149.

Meltzer, M. I., Cox, N. J., and Fukuda, K. (1999). *The economic impact of pandemic influenza in the United States: priorities for intervention.* Emerging infectious diseases, 5(5), pp. 659.

Mitić, M., Janković, S., Mašković, P., Arsić, B., Mitić, J., and Ickovski, J. (2020). *Kinetic models of the extraction of vanillic acid from pumpkin seeds.* Open Chemistry, 18(1), pp. 22-30.

Mizumoto, K., and Chowell, G. (2020). *Estimating Risk for Death from Coronavirus Disease, China, January–February 2020.* Emerging infectious diseases, 26(6), pp. 1251.

Müller, O., Lu, G., Jahn, A. and Razum, O. (2020). *COVID-19 Control: Can Germany Learn From China?* International journal of health policy and management, 9(10), pp. 432-435.

Müller, S. A., Balmer, M., Charlton, W., Ewert, R., Neumann, A., Rakow, C., Schlenther, T. and Nagel, K. (2020). *A realistic agent-based simulation model for COVID-19 based on a traffic simulation and mobile phone data.* arXiv preprint arXiv:2011.11453.

Murray, C. J., Lopez, A. D., Chin, B., Feehan, D. and Hill, K. H. (2006). *Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918–20 pandemic: a quantitative analysis.* The Lancet, 368(9554), pp. 2211-2218.

Nastassopoulou, C., Russo, L., Tsakris, A. and Siettos, C. (2020). *Data-based analysis, modelling and forecasting of the COVID-19 outbreak.* PloS one, 15(3), e0230405.

National Academies of Sciences, Engineering, and Medicine. (2020). *Rapid expert consultation on social distancing for the COVID-19 pandemic*. Washington, DC: The National Academies Press. ss. 1-4. https://doi.org/10.17226/25753.

Noymer, A. and Garenne, M. (2000). *The 1918 influenza epidemic's effects on sex differentials in mortality in the United States.* Population and Development Review, 26(3), pp. 565-581.

Ogundokun, R. O., Lukman, A. F., Kibria, G. B. M., Awotunde, J. B. and Aladeitan, B. B. (2020). *Predictive modelling of COVID-19 confirmed cases in Nigeria.* Infectious Disease Modelling 5, pp. 543-548.

Osterholm, M. T. (2005). *Preparing for the next pandemic.* New England Journal of Medicine, 352(18), pp. 1839-1842.

Oviedo de la Fuente, M., Febrero-Bande, M., Munoz, M. P. and Dominguez, A. (2016). *Predicting seasonal influenza transmission using Regression Models with Temporal Dependence.* PLoS One 10(1371). doi: https://doi.org/10.1371/journal.pone.0194250

Patterson, K. D. and Pyle, G. F. (1991). *The geography and mortality of the 1918 influenza pandemic.* Bulletin of the History of Medicine, 65(1), pp. 4-21.

Rath, S., Tripathy, A. and Tripathy, A. R. (2020). *Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model.* Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14, pp. 1467-1474.

Sharpe, D. (2015). *Chi-Square Test is Statistically Significant: Now What?* Practical Assessment, Research, and Evaluation, 20(1), pp. 8.

Smith, A. M., Adler, F. R. and Perelson, A. S. (2010). *An accurate two-phase approximate solution to an acute viral infection model.* Journal of mathematical biology, 60(5), pp. 711-726.

Stafford, N. (2020). *Covid-19: Why Germany's case fatality rate seems so low*. Bmj, 369.

To, Y., and Mandracchia, J. T. (2019). *Learn About Hierarchical Linear Regression in SPSS With Data From Prison Inmates*. SAGE Publications, Limited.

Upton, G., and Cook, I. (1996). *Understanding statistics*. Oxford University Press.

Uyanik, G. K and Guler, N. (2013). *A study on multiple linear regression analysis.* Procedia- Social and Behavioral Sciences 106, pp. 234- 240.

Verity, R., Okell, L. C., Dorigatti, I., Winskill , P., Whittaker , C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P. G. T., Fu, H., Dighe, A., Griffin , J. T., Baguelin , M., Bhatia, S., Boonyasiri, A., Cori, A., Cucunubá, Z., FitzJohn, R., Gaythorpe, K., Green , W., Hamlet, A., Hinsley, W., Laydon, D., Nedjati-Gilani, G., Riley, S., Elsland, S., Volz, E., Wang, H., Wang, Y., Xi, X., Donnelly, C. A., Ghani, A. C. and Ferguson, N. M. (2020). *Estimates of the severity of coronavirus disease 2019: a model-based analysis.* The Lancet infectious diseases. 20(6), pp. 669-677.

Wallgren, A., Wallgren, B., Persson, R., Jorner, U., and Haaland, J. A. (1996). *Graphing statistics and data: Creating better charts*. Sage.

World Health Organisation (2020). *COVID-19 Epidemiology update.* [Online] Available at: https://www.aljazeera.net/news/2020/7/29/%D9%83%D9%88%D8%B1%D9%88%D9%86%D8%A7-23 . (Accessed 13/10/2020)

Yim, K. H., Nahm, F. S., Han, K. A., and Park, S. Y. (2010). *Analysis of statistical methods and errors in the articles published in the Korean journal of pain.* The Korean journal of pain, 23(1), pp. 35.

Yu, Y., Chen, K. C., and Chen, J. (2019). *Exclusive enteral nutrition versus corticosteroids for treatment of pediatric Crohn's disease: a meta-analysis.* World J. Pediatr. 15(1), pp. 26-36

Zhang, X., Meltzer, M. I., and Wortley, P. M. (2006). FluSurge—*a tool to estimate demand for hospital services during the next pandemic influenza.* Medical Decision Making, 26(6), pp. 617-623.