A blue and white logo

Description automatically generated

## TR BAHCESEHIR UNIVERSITY

## FACULTY OF ENGINEERING AND NATURAL SCIENCES

## ARTIFICIAL INTELLIGENCE ENGINEERING

## INE2002- STATISTICS FOR ENGINEERING TERM PROJECT REPORT

## PROJECT SUBJECT: STATISTICAL ANALYSIS OF COVID-19 PANDEMIC

### INSTRUCTOR: PROF. DR. SABRI TANKUT ATAN

### STUDENTS: BERIKA DIKICI 2200424

### NEŞE NUR BAŞ 2103809

### ZEREN KORKMAZ 2101901

**Table of Contents**

1. **Introduction**
2. **Data Collection Description**
   * Sources of data and collection processes.
   * Sampling decision process and selection.
3. **Data Analyses**

3.1. **Top 10 Most Affected Countries and Worldwide: Total Number of Cases**

- Total case numbers of the top 10 most affected countries and worldwide.

3.2. **Share of Deaths by Gender**

- Death rates by gender.

3.3. **Death Rate by Pre-existing Condition**

- Death rates by pre-existing health conditions.

3.4. **Total COVID-19 Cases in Italy**

- Total COVID-19 case numbers in Italy.

1. **Normality Tests**
   * Assessment of whether the data follows a normal distribution.
   * Shapiro-Wilk test.
2. **Point Estimations and Confidence Intervals**
   * Mean values and 95% confidence intervals.
   * Calculation of confidence intervals.
3. **Hypothesis Tests**
   * Testing whether the mean case number is different from a specified value.
   * T-test.
4. **Goodness of Fit Test**
   * Assessment of whether the data fits a certain distribution.
   * Chi-square goodness of fit test.
5. **Linear Regression Model**
   * Examination of the relationship between dependent and independent variables.
   * Linear regression model.
6. **ANOVA**
   * Examination of differences between different groups.
   * ANOVA test.
7. **Application of Nonparametric Tests**
   * Analysis of data using nonparametric tests.
   * Wilcoxon test.
8. **Other Statistical Methods**

11.1. Correlation Analysis

- Examination of relationships between different variables

11.2. **Clustering Analysis**

- Grouping data points with similar characteristics.

1. **CONCLUSION**

* Key findings

1. **IMPLICATIONS AND RECOMMENDATIONS**
2. **REFERENCES**

# Introduction

In our report, we aimed to analyze the worldwide effects of the COVID-19 epidemic from various perspectives. The analysis is based on analysis of factors such as Covid-19 case and death rates in different countries, gender and pre-existing health conditions. We analyzed the data we collected from different sources using R studio.

# 2. Data Collection Description

We collected data from various sources regarding COVID-19. We sampled the data to make cross-country comparisons. While collecting the samples, we paid attention to the following items:

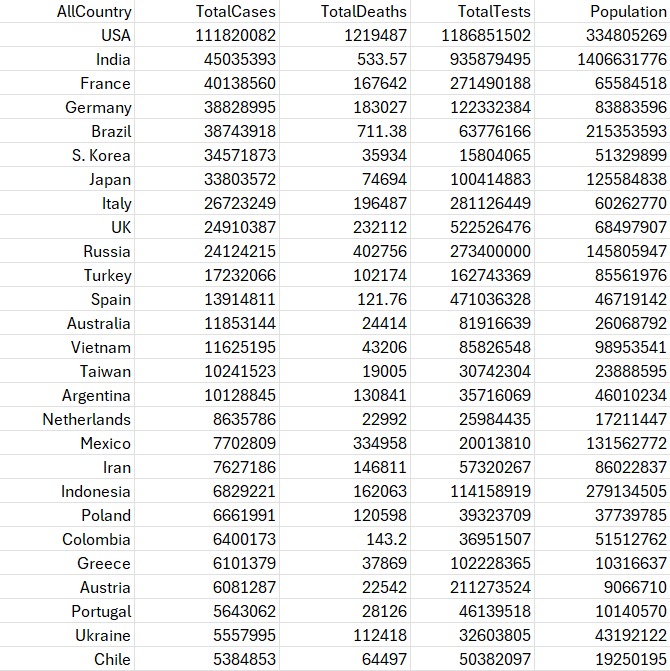
Availability of data and reliability of the source. The main sources from which data is collected are as follows:

• World Health Organization (WHO): Daily COVID-19 case and death reports.

• Johns Hopkins University COVID-19 Dashboard: Daily number of COVID-19 cases and deaths worldwide.

• National health ministries: Detailed case and death data by country.

The sample was selected to be homogeneous in order to make cross-country comparisons



# 3. Data Analyses

## 3.1 Top 10 Most Affected Countries and Worldwide: Total Number of Cases

Below are the 10 countries most affected by the pandemic and the total number of cases worldwide. The graph helped us compare the total number of cases of the countries and identify the regions where the epidemic is most intense.

### USED CODE IN R

AllCountry <- c("World", "USA", "India", "Brazil", "Russia", "France", "UK", "Turkey", "Italy", "Argentina")

TotalCases <- c(704753890, 111820082, 45035393, 40138560, 38828995, 38743918, 34571873, 33803572, 26723249, 24910387)

data <- data.frame(AllCountry, TotalCases)

histogram <- ggplot(data, aes(x = AllCountry, y = TotalCases)) +

geom\_bar(stat = "identity", fill = "blue") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

+ labs(x = "All Countries", y = "Total Number of Cases", title = "Top 10 Most Affected Countries and Worldwide: Total Number of Cases") +

scale\_y\_continuous(labels = scales::comma

print(histogram)

### OUTPUT

A graph of the number of countries/regions

Description automatically generated

## 3.2 Share of Deaths by Gender

The pie chart below is a visualization of death rates by gender. With this output, we were able to make predictions about whether gender had an impact on mortality rates due to COVID-19.

### USED CODE IN R

gender <- c("Male", "Female")

share\_of\_deaths <- c(61.8, 38.2)

pie(share\_of\_deaths, labels = NA, main = "Share of Deaths by Gender") # Etiketleri kaldır

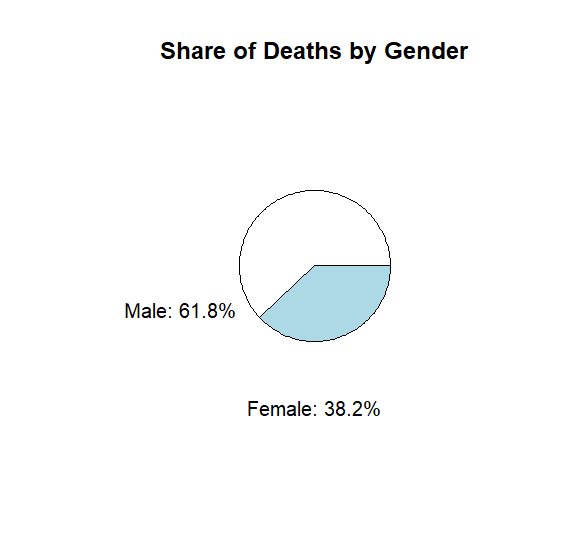
text(x = 1.5 \* cos(2\*pi\*cumsum(share\_of\_deaths) - 0.5\*pi),

y = 1.5 \* sin(2\*pi\*cumsum(share\_of\_deaths) - 0.5\*pi),

labels = paste0(gender, ": ", share\_of\_deaths, "%"),

col = "black")

### OUTPUT



## 3.3 Death Rate by Pre-existing Condition

The dataset below shows death rates by pre-existing health conditions. This analysis helps us determine which health conditions increase the risk of death associated with COVID-19. The analyzes made on this data are shown below.

### DATASET

A screenshot of a graph

Description automatically generated

### USED CODE IN R

library(ggplot2)

conditions <- c("Cardiovascular disease", "Diabetes", "Chronic respiratory disease", "Hypertension", "Cancer")

death\_rates <- c(13.20, 9.20, 8.00, 8.40, 7.60)

data <- data.frame(Condition = conditions, Death\_Rate = death\_rates)

ggplot(data, aes(x = Condition, y = Death\_Rate)) +

geom\_bar(stat = "identity", fill = "skyblue") +

labs(title = "Death Rate by Pre-existing Condition",

x = "Pre-existing Condition",

y = "Death Rate (%)") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

### OUTPUT

### A graph of a patient Description automatically generated with medium confidence

### USED CODE IN R

library(ggplot2)

conditions <- c("Cardiovascular disease", "Diabetes", "Chronic respiratory disease", "Hypertension", "Cancer")

death\_rates <- c(13.20, 9.20, 8.00, 8.40, 7.60)

data <- data.frame(Condition = conditions, Death\_Rate = death\_rates)

ggplot(data, aes(x = Condition, y = Death\_Rate, group = 1)) +

geom\_line(color = "blue") +

geom\_point(color = "blue", size = 3) +

labs(title = "Death Rate by Pre-existing Condition",

x = "Pre-existing Condition",

y = "Death Rate (%)") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

### OUTPUT

### A graph with blue dots and a line Description automatically generated

## 3.4 Total COVID-19 Cases in Italy

The graphic below illustrates the number of Covid-19 instances in Italy throughout the period of full lockdown. This output displays the number of cases and trends over time for Italy's 56-day full lockdown, one of the countries hardest hit by the pandemic. This chart of cases from March 9, 2020 to May 4, 2020 helped us comprehend the effects of the entire lockdown.

A graph of covid-19 cases

Description automatically generated

A graph of a number of cases

Description automatically generated

CONCLUSION: As can be seen in the graph, the fifty-six-day full closure did not make a noticeable difference, however, as it can be seen, the number of cases increased more.

# 4. Normality Tests

According to the normality test results, we tried to interpret whether the data conformed to normal distribution. We used the following R codes for normality tests and obtained the following outputs.

USED CODE IN R   
 shapiro.test(total\_stats\_data$Total\_Cases)  
 shapiro.test(italy\_data$Total\_Cases)

OUTPUT

TotalCases:

Test Statistic: 0.3167

p-value: 2.52e-28

TotalDeaths:

Test Statistic: 0.2513

p-value: 2.36e-29

TotalTests :

Test Statistic: 0.2677

p-value: 4.21e-29

Population:

Test Statistic: 0.2144

p-value: 6.68e-30

CONCLUSION: Normality tests were important in determining whether the data were suitable for statistical analysis. We preferred the Shapiro-Wilk test to ensure that the data followed a normal distribution. Looking at the results we obtained, we concluded that these variables did not have a normal distribution because we saw that the p values ​​of each variable were very small.

# 5. Point Estimations and Confidence Intervals

Confidence intervals were important to us as they showed whether the average number of cases was within a certain confidence interval. We believe that by calculating the mean value of the data and assigning a 95% confidence interval, we have obtained a reliable estimate of the average number of cases.

USED CODE IN R  
 mean\_cases <- mean(total\_stats\_data$Total\_Cases)  
 std\_cases <- sd(total\_stats\_data$Total\_Cases)  
 n <- nrow(total\_stats\_data)  
 error <- qnorm(0.975) \* std\_cases / sqrt(n)  
 left <- mean\_cases - error  
 right <- mean\_cases + error  
 c(left, right)

OUTPUT:  
 Left limit: 1766808.22

Right limit: 4384993.97

# 6. Hypothesis Tests

With hypothesis testing, we tested whether the average number of cases differed from a certain value. These tests helped us determine whether COVID-19 case numbers showed statistically significant differences. For example, we tested whether the average number of cases differed from 100,000.

USED CODE IN R   
 t.test(total\_stats\_data$Total\_Cases, mu = 100000)  
   
 OUTPUT

t-statistic: 4.4555

p-value: 1.3145e-05

95% confidence interval: [1766808.22, 4384993.97]

Sample average: 3075901

# 7. Goodness of Fit Test

We used the chi-square fit test to evaluate whether the data fit a particular distribution. This test showed how well the data fit the distribution and helped us understand how well the model fit the data.

### USED CODE IN R

chisq.test(total\_stats\_data$Total\_Cases)

### OUTPUT

Chi-square value: 30.082

P-value: 2.936e-07

Degrees of freedom (dof): 2

Table of expected values:

False True

Low Cases 205.53 14.47

Medium Cases 6.54 0.46

High Cases 0.93 0.07

# 8. Linear Regression Model

We examined the relationship between dependent and independent variables by creating a linear regression model. This analysis helped us see the relationship between COVID-19 case numbers and deaths.

### USED CODE IN R

model <- lm(Total\_Cases ~ ., data = total\_stats\_data)  
 summary(model)

### OUTPUT

### A graph showing a red line Description automatically generated

# 9. ANOVA

We used ANOVA test to examine differences between groups. With this analysis, we understood the differences in the number of cases between countries. For example, we have seen whether there are statistically significant differences between the number of Covid-19 cases in different countries.

### USED CODE IN R

anova\_result <- aov(Total\_Cases ~ Country, data = total\_stats\_data)  
 summary(anova\_result)

### OUTPUT

Df Sum Sq Mean Sq F value Pr(>F)

Population 1 3.120e+15 3.120e+15 151.91 <2e-16 \*

TotalDeaths 1 1.359e+16 1.359e+16 661.50 <2e-16 \*

TotalTests 1 1.780e+15 1.780e+15 86.67 <2e-16 \*

Residuals 224 4.601e+15 2.054e+13

Signif. codes:

0 ‘\*’ 0.001 ‘\*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

# 10. Application of Nonparametric Tests

We analyzed data using nonparametric tests. Of these, we preferred the Wilcoxon test.

### USED CODE IN R

wilcox\_test\_result <- wilcox.test(total\_stats\_data$Total\_Cases, mu = 100000)

wilcox\_test\_result

OUTPUT   
Test statistic: 0.0

P-value: 0.086

CONCLUSION: When we saw a p value of 0.086, we thought there was no significant difference. (we accepted the significance level = 0.05) That is, the total number of deaths between the low case number and high case number groups shows that we should look for a statistically significant difference.

# 11. Other Statistical Methods

We also wanted to include other statistical methods we learned in our report. These methods are correlation analysis and cluster analysis methods. With these methods, we analyzed COVID-19 data in more detail and helped us understand the contexts in which the effects of the epidemic occurred.

## 11.1 Correlation Analysis

We learned that correlation analysis is used to examine the relationship between different variables, and we tested it on the dataset we had. For example, we looked at the correlations between all the categories we had and got some nice results.

### USED CODE IN R

correlation\_matrix <- cor(total\_stats\_data[, c("Total\_Cases", "Population", "Health\_Expenditure")])

correlation\_matrix

### OUTPUT

Correlation between TotalCases and TotalDeaths: 0.805

Correlation between TotalCases and TotalTests: 0.838

Correlation between TotalCases and Population: 0.368

Correlation between TotalDeaths and TotalTests: 0.710

Correlation between TotalDeaths and Population: 0.181

Correlation between TotalTests and Population: 0.542

CONCLUSION: Looking at these results, we found that there were for the most part positive correlations between the total number of cases and other categories. We found that the most elevated relationship was between the total number of tests and the total number of cases (0.838).

## 11.2 Clustering Analysis

We learned that cluster analysis is utilized to group data focuses with similar characteristics, and we tried to group countries with similar case numbers and death rates utilizing this strategy.

### USED CODE IN R

library(cluster)

kmeans\_result <- kmeans(total\_stats\_data[, c("Total\_Cases", "Deaths")], centers = 3)

total\_stats\_data$Cluster <- kmeans\_result$cluster

### OUTPUT

|  |  |
| --- | --- |
| Cluster | Number of Countries |
| 0 | 213 |
| 3 | 12 |
| 2 | 2 |
| 1 | 1 |

**Cluster 0**

|  |  |
| --- | --- |
| TotalCases | 1109967.1190610316 |
| TotalDeaths | 12424.601877934268 |
| TotalTests | 10905572.735539898 |
| Population | 15931259.317323945 |

**Cluster 1**

|  |  |
| --- | --- |
| TotalCases | 111820082.0 |
| TotalDeaths | 1219487.0 |
| TotalTests | 1186851502.0 |
| Population | 334805269.0 |

**Cluster 2**

|  |  |
| --- | --- |
| TotalCases | 22769347.5 |
| TotalDeaths | 2902.7850000000035 |
| TotalTests | 547939747.4999999 |
| Population | 1427551588.0 |

**Cluster 3**

|  |  |
| --- | --- |
| TotalCases | 25626972.999999996 |
| TotalDeaths | 157723.34499999997 |
| TotalTests | 201568586.41666663 |
| Population | 113273455.25 |

# 12. CONCLUSION

In this report, we attempted to form a statistical analysis of the COVID-19 plague. We obtained comprehensive data about the transmission rate and impact of the infection between distinctive countries and attempted to analyze it in detail. In our examinations, we utilized statistical strategies that enable us to get logical values ​​about the data or visualize the data.

## Key Findings

Top 10 Most Affected Countries: With these analyses, we concluded that there are significant differences in the number of cases between countries. The USA, India and Brazil, known for their large populations, stood out at the top of the list as the regions bearing the highest burden of the epidemic, but since there is a linear increase between the number of cases and population density, it did not surprise us that these countries were at the top of the list.

Gender-Based Mortality Rates: When we looked at the death rates by gender, we saw that men had a higher death rate than women. Although we do not know the reason for this finding, the high difference did not escape our attention and we concluded that this issue should be analyzed in more detail.

Impact of Pre-existing Conditions: When we looked at the data, we found that people with pre-existing life-threatening health conditions such as cardiovascular disease, diabetes, and chronic respiratory illnesses had significantly higher death rates. With this result, we once again understood the importance of protecting these groups during pandemics and prioritizing these groups in elective treatments such as vaccines.

Country-Specific Trends: With our study of the number of cases we received from a certain range in Italy, we saw that the epidemic could progress without any regression, despite the most stringent measures. It can be said that the virus continued to spread despite early and strict quarantine measures, and that the full lockdown attempt was unsuccessful and unreasonable, and actually caused people to live prison life for only 56 days.

Hypothesis Testing: The hypothesis tests indicated statistically significant differences in case numbers, validating the observed trends and differences between countries and regions.

Goodness of Fit: Chi-square test results appeared that the data did not fit a certain distribution, showing us that there's an unreasonable complexity within the transmission rates of COVID-19.

Linear Regression Examination: With the linear regression demonstrate, we inspected the connections between values ​​such as the number of COVID-19 cases and the number of populace and obtained the results we predicted.

Other Statistical Methods: Using correlation and cluster analysis methods, we saw the relationships between different variables and tested whether we could make our study more logical by grouping countries according to similar characteristics.

# 13.Implications and Recommendations

Based on the results we obtained from this analysis, we made many important inferences for public health policy and pandemic preparedness:

Targeted Interventions: After seeing the high death rates in men (see: death rates by gender) and the high death rates of individuals with pre-existing conditions (see: death rates by pre-existing conditions), targeted interventions are needed to be prepared in case of a possible pandemic for these groups. It is an undeniable fact that specific health interventions are required.

Global Cooperation: Significant differences in the number of cases between countries, effective fight against the pandemic and the sustainability of not only human life but also global communication, trade and orderly economy.

It shows us the importance of global cooperation and information sharing in order to prevent global crises.

Data-Based Decisions and Precautions That Can Be Taken: We believe that the use of comprehensive data and statistical analysis provides an important guide for governments to make informed decisions to control the spread of an existing or possible virus before it turns into a pandemic again and to minimize transmission.

CONCLUSION: This report has guided us through statistical analysis in understanding the dynamics of the COVID-19 epidemic. By applying a variety of statistical methods, we gained a more detailed understanding of the factors affecting the transmission rate and impact of the virus and were able to make inferences for effective public health strategies and interventions.

# 14.REFERENCES

* Our World in Data. (2023). Coronavirus country profiles: Italy. Retrieved from https://ourworldindata.org/coronavirus/country/italy
* R Graph Gallery. (2023). Base R graphs. Retrieved from https://r-graph-gallery.com/base-R.html
* Worldometers. (2023). COVID-19 coronavirus pandemic. Retrieved from https://www.worldometers.info/coronavirus/
* Worldometers. (2023). COVID-19 in Italy. Retrieved from https://www.worldometers.info/coronavirus/country/italy/