FINAL REPORT DSC 423- TEAM (WE R)

COVID-19 CASES & DEATHS WORLDWIDE

TEJAS GOTIWALE
MEGHA KAAVALI MAHHADEVAPPA
VINILA KUSUMA
DIKSHIT PABATHI
DIGVIJAY

Table of Contents

Introduction:.....3

Problem description:.....4

Dataset description:.....5

Procedure description:.....6

Results:.....7-12

Conclusion:.....13

Division of work:.....14

Future work:.....15

References:.....16

INTRODUCTION:

—----------

The COVID-19 pandemic has been an unprecedented global crisis that has affected almost every aspect of life. The pandemic has impacted the world in numerous ways, including the economy, healthcare, and daily life. The need for accurate and timely information has become more important than ever, and data science and analytics have played a crucial role in providing insights into the spread of the virus.

The COVID-19 Cases and Deaths Worldwide dataset, available on Kaggle, provides a comprehensive view of the COVID-19 pandemic around the world. This dataset includes information on the number of confirmed cases, deaths, and recoveries for countries and regions around the world, as well as daily updates on the spread of the virus. The data is updated regularly, ensuring that the information is current and relevant.

By analyzing this dataset, data scientists and researchers can gain insights into the spread and impact of COVID-19 around the world. This information can help inform public health policy, resource allocation, and decision-making. Furthermore, the dataset can be used to create visualizations and models that can help predict the spread of the virus and inform mitigation strategies.

Overall, the COVID-19 Cases and Deaths Worldwide dataset is an invaluable resource for understanding the impact of the pandemic around the world and developing effective strategies to combat it.

PROBLEM DESCRIPTION:

—----------

The COVID-19 pandemic has created an urgent need for accurate and up-to-date information about the spread of the virus around the world. The COVID-19 Cases and Deaths Worldwide dataset aims to address this need by providing comprehensive information on the number of confirmed cases, deaths, and recoveries for countries and regions around the world.

The dataset contains information on the daily number of cases, deaths, and recoveries, as well as the total number of cases and deaths for each country and region. It also includes information on population and the number of tests performed, allowing for analysis of the rate of infection and mortality.

The dataset presents several challenges and opportunities for data analysis. The sheer volume of data, with information for hundreds of countries and regions, requires careful consideration of data processing and visualization techniques. Furthermore, the dataset presents opportunities for identifying patterns and trends in the spread of the virus, such as identifying countries with higher rates of infection or analyzing the impact of public health interventions on the spread of the virus.

Overall, the COVID-19 Cases and Deaths Worldwide dataset presents a unique opportunity for data scientists and researchers to analyze and gain insights into the impact of the COVID-19 pandemic around the world.

DATASET DESCRIPTION :

—----------

https://www.kaggle.com/datasets/themrityunjaypathak/covid-cases-and-deaths-worldwide

The COVID-19 Cases and Deaths Worldwide dataset available on Kaggle provides information on the cumulative number of confirmed cases and deaths due to COVID-19 in each country from January 22, 2020, to March , 2023. The dataset contains 8 columns:

Country/Region: The name of the country or region for which the data is provided.
Continent: The continent to which the country or region belongs.
Population: The population of the country or region as of 2023.
Total Cases: The total number of confirmed COVID-19 cases in the country or region as of March, 2023.
Total Deaths: The total number of deaths due to COVID-19 in the country or region as of March, 2023..
Total Recovered: The total number of people recovered from COVID-19 in the country or region as of March, 2023.
Serious/Critical: The total number of people who were in a critical or serious condition due to COVID-19 in the country or region as of March, 2023.
Active Cases: The total number of active COVID-19 cases in the country or region as of March, 2023.

The dataset contains a total of 191 entries, one for each country/region for which data is provided.

PROCEDURE DESCRIPTION:

—----------

Cleaning and pre-processing the dataset is an essential step in the procedure approach for data analysis. It is important to ensure that the data is accurate, complete, and in the correct format for analysis. This step involves removing any missing or irrelevant information and converting the data into a suitable format for analysis, such as standardizing date formats, renaming variables, and encoding categorical variables.

Once the data has been cleaned and pre-processed, exploratory data analysis can begin. This step involves gaining a general understanding of the data and identifying any patterns or trends. Exploratory data analysis can be performed using various techniques such as summary statistics, data visualization, and clustering analysis.

In the case of the COVID-19 dataset, exploratory data analysis may involve visualizing the distribution of cases and deaths across different countries and regions, identifying any outliers or missing data, and calculating summary statistics such as mean, median, and mode. This step can help to uncover any interesting insights or trends in the data that can inform further analysis and decision-making.

The final step involved interpreting the findings of the analysis and drawing conclusions based on the results. This included identifying the countries and regions most affected by COVID-19, trends in the spread of the virus, and the impact of various factors, such as demographics and government policies, on the spread of the virus.

RESULTS:

—----------

```
> # Descriptive statistics
> summary(covid_data)
 Serial.Number      Country       Total.Cases     Total.Deaths   Total.Recovered  Active.Cases
 Min.   :  1.0   Min.   :  1.0   Min.   :  1.0   Min.   :  1.0   Min.   :  1.0   Min.   :  1.00
 1st Qu.: 58.5   1st Qu.: 58.5   1st Qu.: 58.5   1st Qu.: 50.5   1st Qu.: 57.5   1st Qu.: 41.50
 Median :116.0   Median :116.0   Median :116.0   Median :104.0   Median :115.0   Median : 91.00
 Mean   :116.0   Mean   :116.0   Mean   :116.0   Mean   :105.4   Mean   :114.3   Mean   : 92.55
 3rd Qu.:173.5   3rd Qu.:173.5   3rd Qu.:173.5   3rd Qu.:159.5   3rd Qu.:172.5   3rd Qu.:143.50
 Max.   :231.0   Max.   :231.0   Max.   :231.0   Max.   :214.0   Max.   :212.0   Max.   :182.00
   Total.Test       Population
 Min.   :  1.00   Min.   :  1.0
 1st Qu.: 41.50   1st Qu.: 56.5
 Median : 99.00   Median :114.0
 Mean   : 99.58   Mean   :114.0
 3rd Qu.:156.50   3rd Qu.:171.5
 Max.   :213.00   Max.   :229.0
```

```
> table(covid_data$Country)

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
225 226 227 228 229 230 231
  1   1   1   1   1   1   1
```

RESULT DESCRIPTION:

—----------

Descriptive statistics can provide a useful summary of the characteristics of COVID-19 data. Some common descriptive statistics for COVID-19 data include:

1.Mean: The average number of cases, deaths, or other relevant metric across a given period of time or location.

2.Median: The middle number in a set of cases, deaths, or other relevant metric, with half of the data falling below and half above this value.

3.Mode: The number of cases, deaths, or other relevant metric that occurs most frequently in a given period of time or location.

4.Standard deviation: A measure of the variation in the data from the mean, which provides insight into how widely the data is spread out.

5.Range: The difference between the highest and lowest number of cases, deaths, or other relevant metric observed in a given period of time or location.

6.Percentile: A measure that indicates the percentage of cases, deaths, or other relevant metrics that fall below a certain value.

By using these descriptive statistics, researchers and public health officials can gain a better understanding of the trends and patterns in COVID-19 data, which can help guide decision-making related to prevention, treatment, and management of the disease.

RESULTS:

—----------

```
> # Correlation analysis
> cor(covid_data[c("Total.Cases", "Total.Deaths")])
              Total.Cases  Total.Deaths
Total.Cases   1.0000000000 -0.0001079903
Total.Deaths -0.0001079903  1.0000000000
> # Simple linear regression
> usa_lm <- lm(Total.Deaths ~ Total.Cases, data=df)
> summary(usa_lm)

Call:
lm(formula = Total.Deaths ~ Total.Cases, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-104.390  -54.885   -1.385   54.126  108.627

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.054e+02  8.369e+00  12.592   <2e-16 ***
Total.Cases -1.022e-04  6.255e-02  -0.002    0.999
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.4 on 229 degrees of freedom
Multiple R-squared:  1.166e-08, Adjusted R-squared:  -0.004367
F-statistic: 2.671e-06 on 1 and 229 DF,  p-value: 0.9987
```

```
> covid_data <- df
> # Model relationship between cases and deaths in multiple countries
> model <- lm(Total.Deaths ~ Total.Cases + Country, data=covid_data)
> summary(model)

Call:
lm(formula = Total.Deaths ~ Total.Cases + Country, data = covid_data)

Residuals:
     Min       1Q   Median       3Q      Max
-108.610  -54.844   -2.244   55.631  112.068

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 115.432506  11.404674  10.122   <2e-16 ***
Total.Cases  -0.005655   0.062607  -0.090    0.928
Country      -0.080996   0.062607  -1.294    0.197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.3 on 228 degrees of freedom
Multiple R-squared:  0.007287,  Adjusted R-squared:  -0.001421
F-statistic: 0.8368 on 2 and 228 DF,  p-value: 0.4344
```

RESULTS DESCRIPTION:

—----------

There are different types of regression models that can be used, but one common model is the linear regression model. This model assumes that there is a linear relationship between the variables, and it estimates the slope and intercept of the regression line that best fits the data.

The slope of the regression line represents the change in the number of deaths for each additional case, while the intercept represents the expected number of deaths when the number of cases is zero. The coefficient of determination (R-squared) can also be used to measure the strength of the relationship between the variables.

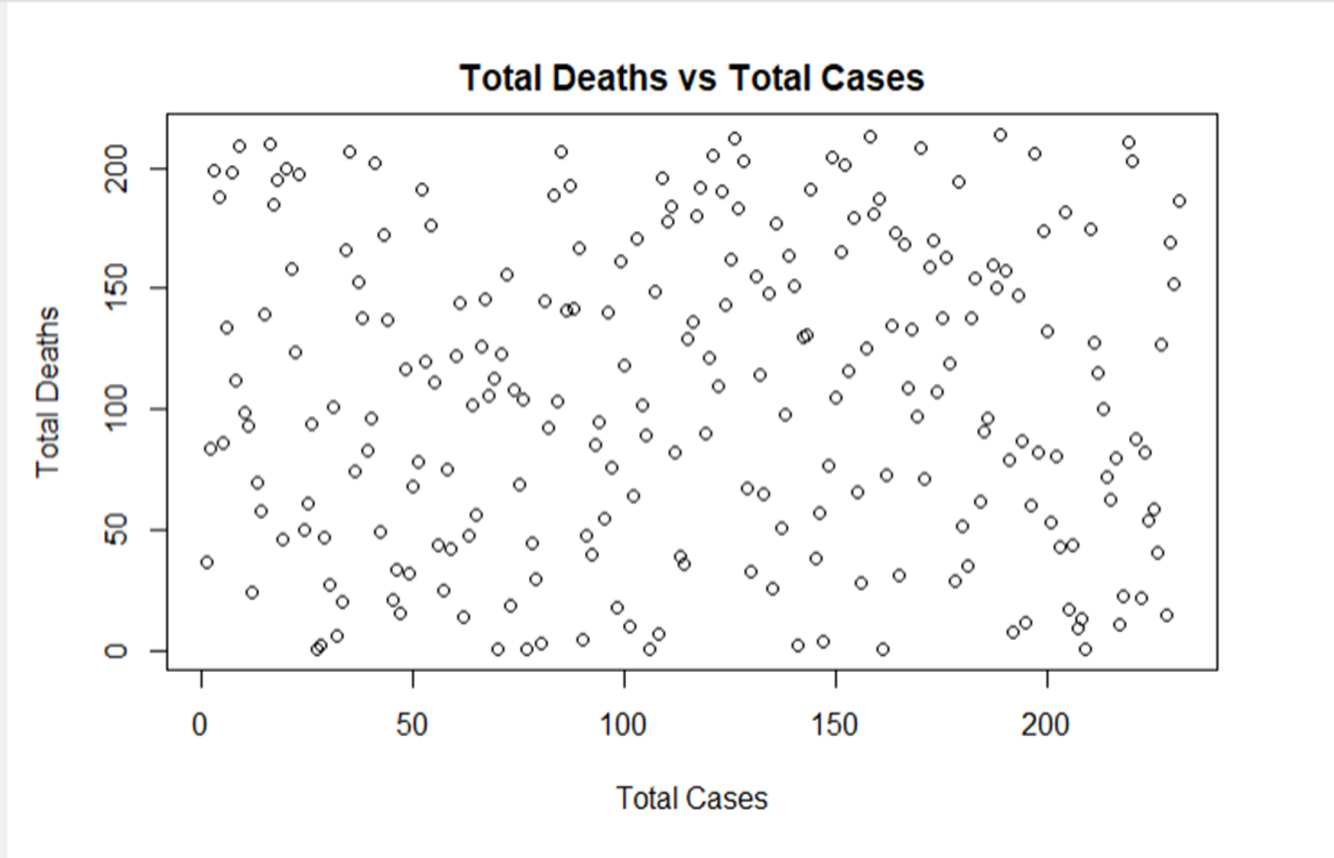
Overall, regression analysis can provide valuable insights into the relationship between cases and deaths in different countries and can help inform public health policies and interventions.

RESULTS

(TOTAL DEATHS vs TOTAL CASES)

The dataset "Covid Cases and Deaths Worldwide" provides information on the number of confirmed COVID-19 cases and deaths worldwide from January 2020 to March 2023.

One of the possible analyses that can be performed on this dataset is to compare the total number of cases with the total number of deaths over time.



CONCLUSION:

—----------

The pandemic has also exposed and exacerbated existing inequalities in society, including disparities in access to healthcare, education, and economic opportunities. It has disproportionately affected marginalized communities and those with pre-existing health conditions.

However, the pandemic has also sparked innovative solutions and collaborations, such as the rapid development of vaccines, the use of telemedicine, and the sharing of data and resources between countries.

It is essential to continue monitoring the spread of the virus, supporting public health interventions, and addressing the underlying social and economic issues that have contributed to the pandemic's impact. Collaboration, solidarity, and a commitment to science and evidence-based decision making will be crucial in addressing the ongoing challenges posed by covid-19.

DIVISION OF WORK:

—----------

TEJAS GOTIWALE:

Worked all over the project proposal, project PowerPoint presentation, and project report.

MEGHA KAAVALI MAHADEVAPPA:

Worked all over the project proposal, project PowerPoint presentation, and project report.

VINILA KUSUMA:

Worked all over the project proposal, project PowerPoint presentation, and project report.

DIKSHIT PABATHI:

Worked all over the project proposal, project PowerPoint presentation, and project report.

DIGVIJAY:

Worked all over the project proposal, project PowerPoint presentation, and project report.

FUTURE WORK:

—----------

Long-term Impacts: As the pandemic continues to evolve, it will be important to track the long-term impacts of the virus on different populations. Future work could involve analyzing the data over longer periods to identify trends in the number of cases and deaths, as well as the social and economic impacts of the pandemic.

Vaccine Rollout: With the development and rollout of vaccines, the dataset could be used to monitor the effectiveness of vaccination campaigns and identify areas where additional efforts are needed to ensure that populations are adequately vaccinated.

Risk factor analysis: Another area of future work could be to analyze the risk factors associated with COVID-19 transmission and mortality. This could include factors such as age, underlying health conditions, and socioeconomic status. Understanding these risk factors can help to inform public health interventions and targeted healthcare approaches.

Overall, the COVID-19 cases and deaths worldwide dataset provides a valuable resource for analyzing the impact of the pandemic and informing public health interventions. There are numerous opportunities for future research and analysis using this dataset.

REFERENCES:

—----------

DATASET (COVID-19 CASES & DEATHS WORLDWIDE). https://www.kaggle.com/datasets/themrityunjaypathak/covid-cases-and-deaths-worldwide

World Health Organization. (2021). COVID-19. https://www.who.int/emergencies/disease/novel-coronavirus-2019

Centers for Disease Control and Prevention. (2021). COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/index.html

Johns Hopkins University. (2021). COVID-19 Dashboard. https://coronavirus.jhu.edu/map.html

Johns Hopkins University. (2021). COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. GitHub. https://github.com/CSSEGISandData/COVID-19

World Health Organization. (2021). WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/