**COVID-19 Data Analysis**

**Data Source:** [**https://www.kaggle.com/datasets/sudalairajkumar/novel-corona-virus-2019-dataset**](https://www.kaggle.com/datasets/sudalairajkumar/novel-corona-virus-2019-dataset)

From above source following 3 files have been utilized:

* **covid\_19\_data.csv**
* **time\_series\_covid\_19\_confirmed\_US.csv**
* **time\_series\_covid\_19\_deaths\_US.csv**

**The purpose of this project I’ve covered the following:**

* calculating Confirmed, recovered, death cases from the dataset.
* Regression model on Observation Date vs Confirmed cases and plotting the graph for better visualization.
* Calculating Growth Rate of Confirmed and Death cases.
* Plotting US map to show the number of Confirmed cases (~State).
* Plotting Pivot table to gain more insights on Death cases using different types of graphs.

**#DATA STRUCTURE**

Text, letter

Description automatically generated

**# Total Numbers of Cases Worldwide.**

Total Confirmed Cases 🡪 **305766**

Total Recovered Cases 🡪 **18111**

Total Death Cases 🡪 **46932**

**# Plotting of Total Cases vs Observation Date**

Chart, histogram

Description automatically generated

**Comment:** According to these statistics, the number of confirmed cases climbed in February, rising from zero to 335,000 in March.

**# Plotting of Growth rate of Confirmed cases using Regression Method**

To assess the pace at which COVID-19 cases are rising or decreasing daily, the growth rate of cases is utilized to model the linear regression.

**Formula: Total no. of new cases / Total no. of new cases previous day \* 100**

Chart

Description automatically generated

**# Regression model of Growth Rate of Confirmed Cases**

Text, letter

Description automatically generated

**Comment:** The residuals of the observations are shown in the output above. The regression model's multiple R-squared is 0.5 percent, and the adjusted R-squared is -0.9 percent. The multiple R-squared coefficient denotes the multiple correlation between the responder and predictor variables. The model predictor variable's P-value is 0.5, which is excessively high.

**# Growth rate of Death Cases.**

Chart, histogram

Description automatically generated

**Comment:** The graph above depicts a surge in Death cases on two occasions, in the middle and late of February, when there were up to 5000 instances on consecutive days.

**# Growth Rate of Confirmed Cases in China**

Chart, histogram

Description automatically generated

**Comment:** The graph above illustrates that China saw a rise in covid cases from late January to late February, and that it steadily decreased with medical treatment.

**# Countries with high rate of Covid-19 Cases**

Chart, histogram

Description automatically generated

**Comment:** The following line graph depicts the growth of covid cases in March, with the United States being the most affected country, with cases reaching 20,000 by the end of the month.

**# Confirmed Cases in US**

* **Obtain using time\_series\_covid\_19\_confirmed\_US.csv**

![Map

Description automatically generated]()

**Comment:** The map depicts states with a high population density ratio and covid-19 instances. The state with the highest rate of covid-19 instances is Texas, followed by Georgia.

**P.S.** 🡪 The visualization looks more interactive when open in R.

**# Pivot Table of Death Cases in US**

Graphical user interface, chart, bar chart

Description automatically generated

**Comment:** The above bar chart depicts the death rate in the United States, with Missouri having the highest rate.

**# Code .R**

library(ggplot2)

library(caret)

library(dplyr)

library(readr)

library(highcharter)

library(rpivotTable)

**# Dataset**

covid <- read.csv(file.choose(), header = T)

summary(covid)

newcov <- covid[-c(10001:306429),]

str(newcov)

**# Converting character string to DATE**

newcov$ObservationDate <- as.Date(newcov$ObservationDate, format = "%m/ %d/ %Y")

**# Aggregating data**

confirmed\_cases <- aggregate(Confirmed ~ ObservationDate, newcov, sum)

death\_cases <- aggregate(Deaths ~ ObservationDate, newcov ,sum)

recovered\_cases <- aggregate(Recovered ~ ObservationDate, newcov ,sum)

totalConfirmed <- confirmed\_cases$Confirmed[length(confirmed\_cases$Confirmed)]

totalDeaths <- death\_cases$Deaths[length(death\_cases$Deaths)]

totalRecovered <- recovered\_cases$Recovered[length(recovered\_cases$Recovered)]

**# Scatter plot for all types of Cases**

ggplot() +

geom\_point(data = confirmed\_cases, aes(x = ObservationDate, y = Confirmed, color

= 'Confirmed Cases')) +

geom\_point(data = death\_cases, aes(x = ObservationDate, y = Deaths, color =

'Death Cases')) +

geom\_point(data = recovered\_cases, aes(x = ObservationDate, y = Recovered, color

= 'Recovered Cases')) +

xlab("Date") +

ylab("No. of Cases") +

# scale\_x\_discrete(guide = guide\_axis(check.overlap = TRUE)) +

scale\_y\_continuous(limits = c(0, 350000), breaks = seq(0, 350000, 50000)) +

labs( color = "Types of cases") +

ggtitle("COVID-19 cases")

**# Growth Rate of Covid cases**

C\_growrate = confirmed\_cases %>% arrange(ObservationDate) %>%

mutate(

date = ObservationDate,

grow = Confirmed - lag(Confirmed),

grow\_rate = grow/lag(grow) \* 100)

**# Eliminating Rows with NA**

C\_growrate$grow\_rate[is.na(C\_growrate$grow\_rate)] <- mean(C\_growrate$grow\_rate, na.rm = TRUE)

C\_growrate$grow[is.na(C\_growrate$grow)] <- mean(C\_growrate$grow, na.rm = TRUE)

str(C\_growrate)

**# Regression Model**

model <- lm(grow\_rate ~ ObservationDate, data = C\_growrate)

summary(model)

**# Plotting of Regression model**

ggplot(data = C\_growrate, aes(x = ObservationDate, y = grow\_rate, group = 1)) +

geom\_line(color = 'red') +

geom\_smooth(formula = y ~ x, method = "lm", se = T, col = 'black') +

#scale\_x\_discrete(guide = guide\_axis(check.overlap = TRUE)) +

xlab("Date") +

ylab ( "Growth Rate") +

ggtitle("Confirmed Cases Growth rate")

**### Similarly, we can do the plotting for death cases and recovered cases.**

**# Death Rate of Covid cases**

D\_growrate = death\_cases %>% arrange(ObservationDate) %>%

mutate(

date = ObservationDate,

grow = Deaths - lag(Deaths),

grow\_rate = grow/lag(grow) \* 100)

**# Plotting of Regression model**

ggplot(data = D\_growrate, aes(x = ObservationDate, y = grow\_rate, group = 1) ) +

geom\_line(color = 'blue') +

geom\_smooth(formula = y ~ x, method = "lm", se = T, col = 'black') +

#scale\_x\_discrete(guide = guide\_axis(check.overlap = TRUE)) +

xlab("Date") +

ylab ( "Growth Rate") +

ggtitle("Confirmed Cases Growth rate")

**# Cases in Countries**

Countries <- c("Mainland China","Italy","Spain","Germany","US","India")

C\_cases <- newcov %>% filter(Country.Region %in% Countries) %>%

group\_by(ObservationDate, Country.Region) %>% summarise(totalConfirmed = sum(Confirmed), .groups = 'drop')

China = C\_cases %>% filter(Country.Region == "Mainland China")

Italy = C\_cases %>% filter(Country.Region == "Italy")

Spain = C\_cases %>% filter(Country.Region == "Spain")

Germany = C\_cases %>% filter(Country.Region == "Germany")

US = C\_cases %>% filter(Country.Region == "US")

India = C\_cases %>% filter(Country.Region == "India")

China\_grow = China %>% group\_by(Country.Region) %>%

arrange(ObservationDate) %>%

mutate(

date = ObservationDate,

grow = totalConfirmed - lag(totalConfirmed))

China\_grow$ObservationDate <- as.Date(China\_grow$ObservationDate, format = "%m/ %d/ %Y")

Italy\_grow = Italy %>% group\_by(Country.Region) %>%

arrange(ObservationDate) %>%

mutate(

date = ObservationDate,

grow = totalConfirmed - lag(totalConfirmed))

Spain\_growth = Spain %>% group\_by(Country.Region) %>%

arrange(ObservationDate) %>%

mutate(

date = ObservationDate,

grow = totalConfirmed - lag(totalConfirmed))

Germany\_grow = Germany %>% group\_by(Country.Region) %>%

arrange(ObservationDate) %>%

mutate(

Diff\_date = ObservationDate,

grow = totalConfirmed - lag(totalConfirmed))

US\_grow = US %>% group\_by(Country.Region) %>%

arrange(ObservationDate) %>%

mutate(

Diff\_date = ObservationDate,

grow = totalConfirmed - lag(totalConfirmed))

India\_grow = India %>% group\_by(Country.Region) %>%

arrange(ObservationDate) %>%

mutate(

Diff\_date = ObservationDate,

grow = totalConfirmed - lag(totalConfirmed))

**# Covid-19 Cases growth in China**

ggplot(data = China\_grow, aes(x = ObservationDate , y = grow, group = 1)) +

geom\_line(color = "black") +

#scale\_x\_discrete(guide = guide\_axis(check.overlap = TRUE)) +

xlab("Date") +

ylab("Cases per day") +

ggtitle("COVID-19 cases per day trend of China")

**# Covid-19 cases growth in top Countries**

ggplot() +

geom\_line(data = Italy\_grow, aes(x = ObservationDate , y = grow, color = "Italy", group = 1), size = 1) +

geom\_line(data = Spain\_growth, aes(x = ObservationDate , y = grow, color = "Spain", group = 1), size = 1) +

geom\_line(data = Germany\_grow, aes(x = ObservationDate , y = grow, color = "Germany", group = 1), size = 1) +

geom\_line(data = US\_grow, aes(x = ObservationDate , y = grow, color = "US", group = 1), size = 1) +

geom\_line(data = India\_grow, aes(x = ObservationDate , y = grow, color = "India", group = 1), size = 1) +

#scale\_x\_discrete(guide = guide\_axis(check.overlap = TRUE)) +

xlab("Date") +

ylab("Cases per day")+

ggtitle("COVID-19 cases per day trend of Other Countries")

**# Mapping of Confirmed Cases in US State**

map <- read.csv(file.choose(), header = T)

summary(map)

US\_map <- map %>%

group\_by(Province\_State) %>%

summarise(count= n())

summary(US\_map)

highchart() %>%

hc\_title(text = "Confirmed Cases in US") %>%

hc\_subtitle(text = "time\_series\_covid\_19\_deaths\_US.csv") %>%

hc\_add\_series\_map(usgeojson, US\_map,

name = "Province\_State",

value = "count",

dataLabels = list(enabled = T, format = "{point.name}"),

joinBy = c("woename", "Province\_State")) %>%

hc\_mapNavigation(enabled = T)

**# Death cases in US Using Pivot Table**

heat <- read.csv(file.choose(), header = T)

rpivotTable(heat,

aggregatorName = "count",

cols = "count",

rows = "Province\_State",

rendererName = "Heatmap",

rainbow(50))