Statistics

Data Mining

NAME & SURNAME:..... NO:.....

```
# Load libraries
library(knitr)
library(tidyverse)
library(dplyr)
library(ggplot2)
library(tidyr)
```

The data used to generate this report contains 187 rows and 15 column. There are no missing values. This report presents an analysis of Confirmed , Deaths, Active and Recovered cases. In addition, the rates of deaths and recoveries both at the country levels and regional levels have been answered.

This report is guided by the following questions:

1. What is the minimum, average, maximum and standard deviations cases of Confirmed, deaths, and recovery?
2. What is the distribution of Confirmed cases?
3. Are there countries with extreme confirmed cases?

- What is the number of Confirmed, Deaths, Recovered, and Active cases in these countries?
- From which geographical region does these countries come from?

5. What is the relationship between the new cases?

- What is the rate of new deaths to new confirmed cases and new recovered cases?

6. Which countries are experiencing the highest mortality rates?

```
data <- read.csv("data.csv", stringsAsFactors = T)

# data dimensions
dim(data)
```

```
## [1] 187  15
```

```
# sum of missing values check for each column
missing <- sapply(data, function(x) sum(is.na(x)))
missing
```

```
##         Country.Region              Confirmed                 Deaths
##                      0                      0                      0
##              Recovered                 Active              New.cases
##                      0                      0                      0
##             New.deaths          New.recovered     Deaths...100.Cases
##                      0                      0                      0
##  Recovered...100.Cases Deaths...100.Recovered    Confirmed.last.week
##                      0                      0                      0
##         X1.week.change     X1.week...increase             WHO.Region
```

```
##                         0                         0                         0
```

```
# COUNT UNIQUE VALUES: Country/Region

data %>%
  distinct(Country.Region) %>%
  count()
```

```
##     n
## 1 187
```

There are 187 countries from 6 WHO recognized regions. 56,48, and 35 countries under investigation are from Europe, Africa and Americas respectively.

```
# COUNT UNIQUE VALUES: country/WHO.region

N <- data %>%
  group_by(WHO.Region) %>%
  summarise(N = n())
kable(N, caption = "Number of countries in per region")
```

Table 1: Number of countries in per region

WHO.Region	N
Africa	48
Americas	35
Eastern Mediterranean	22
Europe	56
South-East Asia	10
Western Pacific	16

What is the minimum, average, maximum and standard deviations cases of Confirmed, deaths, and recovery?

- **Summary Descriptive Statistics**

```
# EXAMINE VARIABALES: confirmed, deaths, recovered

Descr_stats <- data %>%
  select(Confirmed, Deaths, Recovered) %>%
  gather() %>%
  group_by(key) %>%
  summarise('min' = min(value), 'max' = max(value),
            'mean' = mean(value), 'median' = median(value),
            'sum' = sum(value), 'stdev' = sd(value))
kable(Descr_stats, caption = "Descriptive statistics")
```

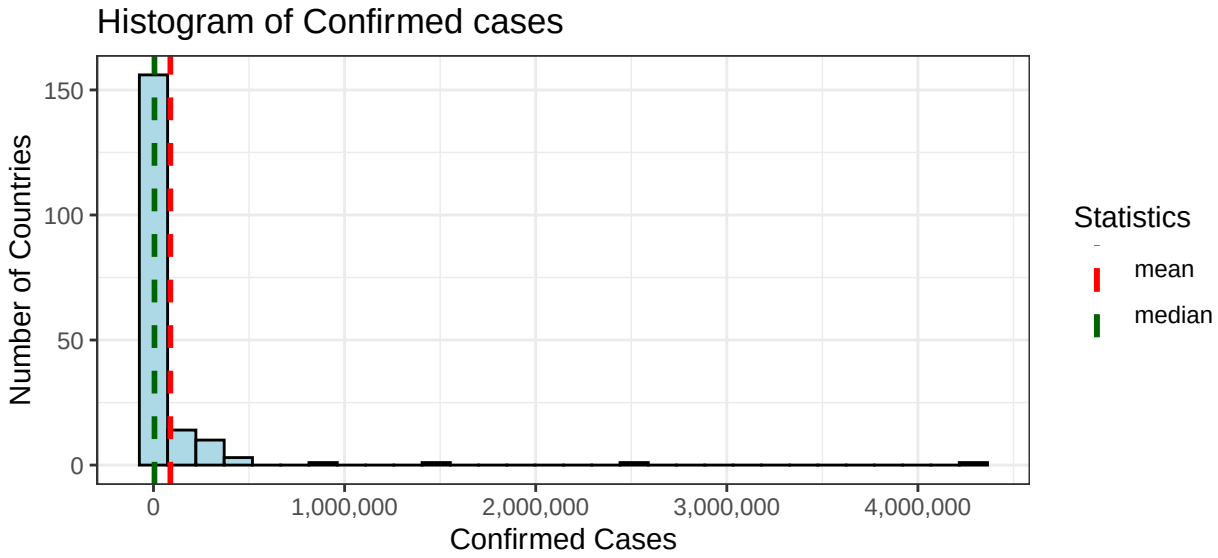
Table 2: Descriptive statistics

key	min	max	mean	median	sum	stdev
Confirmed	10	4290259	88130.936	5059	16480485	383318.7
Deaths	0	148011	3497.519	108	654036	14100.0
Recovered	0	1846641	50631.481	2815	9468087	190188.2

The minimum statistic is the least number of cases reported while the maximum value is the highest number of cases. Mean is the average cases and sum is the total number of cases. Standard deviation is the dispersion of these cases across all countries. According to the table above, the most affected country reported 4290259 confirmed cases while the least affected only had 10 cases. The sum of all cases was 16480485 where on average each country had 88131 cases.

Distribution of Confirmed cases.

```
# Create a histogram with summary statistics
data %>%
  ggplot(aes(Confirmed)) +
  geom_histogram(fill = "lightblue", color = "black") +
  geom_vline(aes(xintercept = mean(Confirmed),
                 color = "mean"), linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = median(Confirmed),
                   color = "median"), linetype = "dashed", size = 1) +
  scale_color_manual(name = "Statistics", values = c(mean = "red", median = "darkgreen")) +
  scale_x_continuous(labels = scales::comma) +
  ggtitle("Histogram of Confirmed cases") +
        xlab("Confirmed Cases") + ylab("Number of Countries") +
  theme_bw()
```



The histogram showws that confirmed cases are skewed to the right. There are countries with extreme cases which we consider to be outliers. These are the countries with more than 500,000 confirmed cases.

Countries with more than unusual confirmed cases

1. What is the number of Confirmed, Deaths, Recovered, and Active cases in these countries?
2. From which geographical region does these countries come from?

The table below shows that US, Brazil and India, and Russia are the countries with more than 500,000 confirmed cases. Both US and Brazil and in Americas Region. India from South-East Asia, and Russia from Europe also lies in this category. US, Brazil and India have more than 1 million cases. The number of Deaths, Recoveries, and active cases are directly proportional to the confirmed cases.

```
# country with more than 500,000 confirmed cases
extremes <- data %>%
```

```
  select(WHO.Region, Country.Region, Confirmed, Deaths, Recovered, Active) %>%
  filter(Confirmed >= 500000) %>%
  arrange(desc(Confirmed))
kable(extremes, caption = "Countries with cases more than 500000 by WHO Region")
```

Table 3: Countries with cases more than 500000 by WHO Region

WHO.Region	Country.Region	Confirmed	Deaths	Recovered	Active
Americas	US	4290259	148011	1325804	2816444
Americas	Brazil	2442375	87618	1846641	508116
South-East Asia	India	1480073	33408	951166	495499
Europe	Russia	816680	13334	602249	201097

Number of cases by Region

As can be seen from the table below, Europe has the highest number of cases followed by Americas. Europe is ranked the first with almost 2482843 confirmed cases where 893559 are active cases and 1391474 have already recovered. South-East Asia and Western Pacific are ranked lowest respectively with least cases.

```
# EXAMINE VARIABALES: confirmed, deaths, recovered, Active

data <- data %>%
  filter(Confirmed <= 500000)

Regional <- data %>%
  group_by(WHO.Region) %>%
  select(Confirmed, Deaths, Recovered, Active) %>%
  summarise(Confirmed = sum(Confirmed),
            Deaths = sum(Deaths),
            Recovered = sum(Recovered),
            Active = sum(Active)) %>%
  arrange(desc(Confirmed))
kable(Regional, caption = "Regionwise Cumulative cases")
```

Table 4: Regionwise Cumulative cases

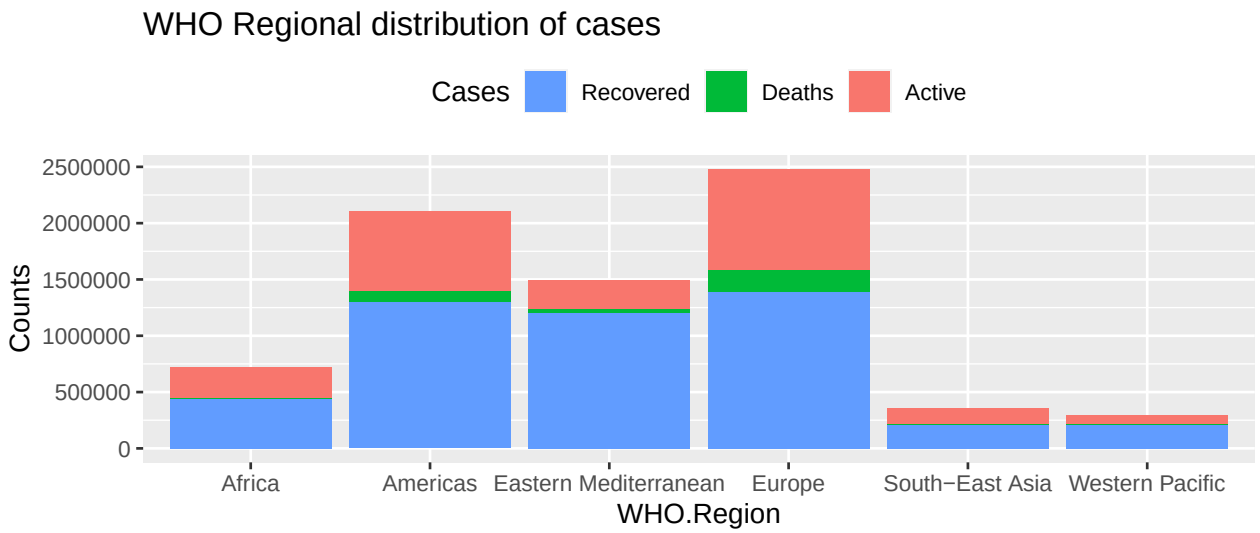
WHO.Region	Confirmed	Deaths	Recovered	Active
Europe	2482843	197810	1391474	893559
Americas	2106652	107103	1296171	703378
Eastern Mediterranean	1490744	38339	1201400	251005
Africa	723207	12223	440645	270339
South-East Asia	355224	7941	205767	141516
Western Pacific	292428	8249	206770	77409

The chart below shows WHO regional distribution of Recovered, Deaths and active cases, and total confirmed cases. Overall, more cases were recorded in the Europe, followed by Americas then Eastern Mediterranean. It is worth noting that these are the most populous countries in America.

```
long_Regional <- Regional %>%
  gather(Cases, Counts, 3:5)

# grouped bar plot
```

```
ggplot(long_Regional,
       aes(x = WHO.Region, y = Counts, fill = Cases)) +
  geom_col() +
  guides(fill = guide_legend(reverse = TRUE)) +
  labs(title = "WHO Regional distribution of cases") +
  theme(legend.position = "top")
```



Ratio Statistics of the New cases

- What is the rate of new deaths to new confirmed cases and new recovered cases?

In table 5 we have the summary of the ratio statistics for Confirmed and Recovered infections to the number of deaths. In western-Pacific there is one death for every 137 confirmed cases where 47 recovers. E#urope is nest with one death in every 76 confirmed cases. The ratio of deaths to confirmed is 1:27 and deaths to recovered is 1:18, implying that for every 27 cases 1 person dies and 18 recovers

```
# EXAMINE VARIABALES: New confirmed, New deaths, New recovered

New <- data %>%
  group_by(WHO.Region) %>%
  select(New.cases, New.deaths, New.recovered) %>%
  summarise(Confirmed = round(sum(New.cases)/sum(New.deaths),0),
            Deaths = sum(New.deaths)/sum(New.deaths),
            Recovered = round(sum(New.recovered)/sum(New.deaths),0)) %>%
  arrange(Confirmed)
kable(New, caption = "Ratio Statistics for new cases")
```

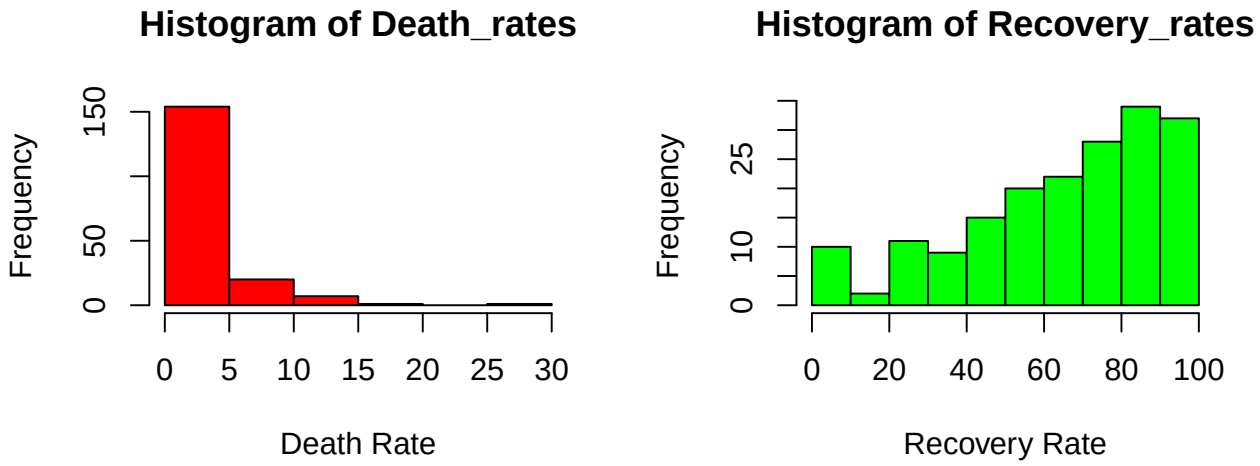
Table 5: Ratio Statistics for new cases

WHO.Region	Confirmed	Deaths	Recovered
Americas	27	1	18
Eastern Mediterranean	28	1	33
Africa	34	1	41
South-East Asia	47	1	41
Europe	76	1	40
Western Pacific	137	1	47

Comparing the distribution of the rates of Deaths and Recoveries

The rate of deaths is right skewed while the rate of recovery is left skewed. In most of the countries, the rates of death is between 0 and 5. On the other hand, the rates of recovery is more than 70% in many countries. We have some outlier with unusual death rates greater than 25%.

```
par(mfrow = c(1,2))
Death_rates <- data$Deaths...100.Cases
Recovery_rates <- data$Recovered...100.Cases
hist(Death_rates, xlab = "Death Rate", col = "red")
hist(Recovery_rates, xlab = "Recovery Rate", col = "green")
```



Which countries are experiencing the highest mortality rates?

In this case we shall use Cause Specific Death Rates to identify the countries with high CSDR. According the the histogram of death rates, any country with a rate greater than 25% is an outlier. From the table below, Yemen is experiencing usually very high rates of death cases. Other countries with high rates of deaths are UK, Belgium and Italy.

```
Rates <- data %>%
  select(Country.Region, Deaths...100.Cases) %>%
  arrange(desc(Deaths...100.Cases))
top <- head(Rates,5)
kable(top, caption = "Top 10 countries with highest death rates")
```

Table 6: Top 10 countries with highest death rates

Country.Region	Deaths...100.Cases
Yemen	28.56
United Kingdom	15.19
Belgium	14.79
Italy	14.26
France	13.71

Where do we have more recovery rates?

The rates of case recoveries(CRR) is used to identify the cpuntries with more probability of recovery. Dominica, Grenada and Holy Se has a 100% chance of recovery.

```
CRR <- data %>%
  select(Country.Region, Recovered...100.Cases) %>%
  arrange(desc(Recovered...100.Cases))
top <- head(CRR,5)
kable(top, caption = "Top 10 countries with highest death rates")
```

Table 7: Top 10 countries with highest death rates

Country.Region	Recovered...100.Cases
Dominica	100.00
Grenada	100.00
Holy See	100.00
Djibouti	98.38
Iceland	98.33

Which regions have the highest rates of recovery?

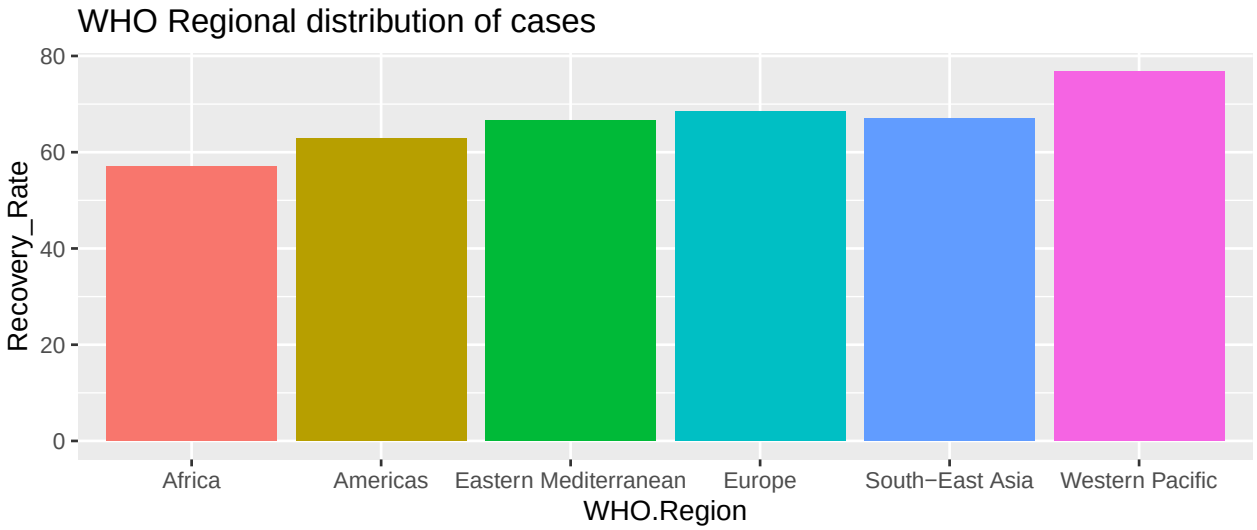
```
CSDR <- data %>%
  group_by(WHO.Region) %>%
  select(Recovered...100.Cases) %>%
  summarise(Recovery_Rate = mean(Recovered...100.Cases)) %>%
  arrange(Recovery_Rate)
kable(CSDR, caption = "Rates of recovery by WHO Region")
```

Table 8: Rates of recovery by WHO Region

WHO.Region	Recovery_Rate
Africa	57.01479
Americas	62.83909
Eastern Mediterranean	66.59318
South-East Asia	66.97556
Europe	68.54218
Western Pacific	76.80500

Africa has the least recovery rate while Western Pacific has the highest rate of recovery. However, there is no significant difference among these regions.

```
# grouped bar plot
ggplot(CSDR,
       aes(x = WHO.Region, y = Recovery_Rate, fill = WHO.Region)) +
  geom_col() +
  guides(fill = guide_legend(reverse = TRUE)) +
  labs(title = "WHO Regional distribution of cases") +
  theme(legend.position = "none")
```


WHO Regional distribution of cases
Recovery_Rate
80
60
40
20
0
Africa
Americas
Eastern Mediterranean
Europe
South−East Asia
Western Pacific
WHO.Region