R analysis on Covid-19 dataset

Lino Alis
Roberto Colangelo
Marco Cirelli
Davide Rossetti

In this part of the project, we work on a dataset that specifies covid-19 related cases and deaths for each day during the whole year of 2020, retrieved from the Github folder: https://github.com/minoue/Project_Python_R.git
The dataset appears to be composed of 61900 observations that are grouped under 12 variables. Because of the multitude of observations, we decided to perform an exploratory analysis on several combinations of variables grouped together, from which we will try to get interesting insights on the data. We therefore created several grouped data frames, from which we visualized the data in different ways.

We started by importing the libraries that are needed in order to perform the analysis, namely ggplot2, dplyr, reshape2 and tidyverse.
Then we proceeded by importing the csv file.

```
library(ggplot2)
library(dplyr)
library(reshape2)
library(tidyverse)
```

```
df <- read.csv('covid.csv', header=TRUE, sep=';')
head(df)
```

date <chr>	day <dbl>	month <dbl>	year <dbl>	cases <dbl>	deaths <dbl>	country <chr>	code <chr>	population <dbl>
1 31/12/2019	31	12	2019	0	0	Afghanistan	AF	38041757
2 01/01/2020	1	1	2020	0	0	Afghanistan	AF	38041757
3 02/01/2020	2	1	2020	0	0	Afghanistan	AF	38041757
4 03/01/2020	3	1	2020	0	0	Afghanistan	AF	38041757
5 04/01/2020	4	1	2020	0	0	Afghanistan	AF	38041757
6 05/01/2020	5	1	2020	0	0	Afghanistan	AF	38041757

6 rows | 1-10 of 13 columns

The first data frame created, df1, includes the variables "cases" and "deaths" grouped for every continent. As expected, there is a very strong positive correlation between the number of cases and deaths (around 0.966). Afterwards, we plotted in a scatterplot the observations of cases vs deaths for continent, and through the line plotted with them we observed that the number of cases in on average 40 times the number of deaths (as the slope of the line fitting the points is 40).

```
df1<-df %>%
  select(continent, deaths, cases) %>%
  group_by(continent)%>%
  summarise(deaths=sum(deaths),cases=sum(cases))
head(df1)
```

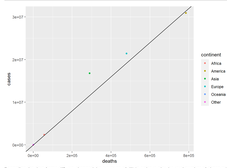
continent <chr>	deaths <dbl>	cases <dbl>
Africa	56334	2379827
America	765420	30887593
Asia	290129	16792046
Europe	479789	21400012
Oceania	1154	53440
Other	7	696

6 rows

```
cor(df1['cases'],df1['deaths'])
```

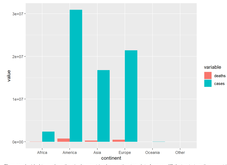
```
##           deaths
## cases 0.9859254
```

```
ggplot(df1, aes(x=deaths,y=cases, color=continent))+
  geom_point() + geom_abline(slope=40)
```



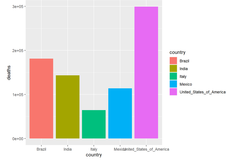
Regarding the data frame df2, we changed the structure of df1 in order to plot the total number of observations in the variables cases and deaths in the different continents, with two bar graphs.

```
df2<- melt(df1, id.vars='continent')
ggplot(df2,aes(x=continent,y=value,fill=variable)) +
  geom_bar(stat='identity',position='dodge')
```

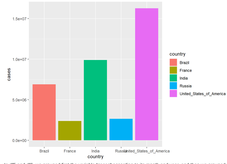


Then we decided to analyze the single countries by creating two data frames: df3 that puts together countries and deaths, and df4 that puts together countries and cases. We explicated the first five countries with the highest number of cases and then the ones with the highest number of deaths. We identified Brazil, India and USA as the countries with the highest number in both cases and deaths. In df3, among the countries with the highest number of deaths we can also identify Italy and Mexico and in df4 the two other countries with the highest number of cases are France and Russia.

```
df3<-df%>%
  select(country,deaths)%>%
  group_by(country)%>%
  summarise(deaths=sum(deaths))%>%
  arrange(desc(deaths))
df3<-df3[1:5,]
ggplot(df3,aes(x=country,y=deaths,fill=country))+
  geom_bar(stat='identity',position='dodge')
```




```
df4<-df%>%
  select(country,cases)%>%
  group_by(country)%>%
  summarise(cases=sum(cases))%>%
  arrange(desc(cases))
df4<-df4[1:5,]
ggplot(df4, aes(x=country,y=cases,fill=country))+
  geom_bar(stat='identity',position='dodge')
```



In df5 and df6, we grouped first the variable "cases" according to its month and year, and then we grouped the variable "deaths" according to its month and year. The graphs plot the total number of cases and deaths for each month of 2020. November was the month with highest number of cases and deaths.

```
df5 = df%>%
  select(month,cases,year)%>%
  group_by(year,month)%>%
  summarise(total_cases=sum(cases))%>%
  arrange(year)%>%
  mutate(Date=paste(as.character(month),as.character(year), sep="/"))%>%
  arrange(Date)
```

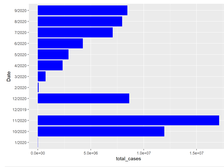
```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

```
head(df5)
```

year <dbl>	month <dbl>	total_cases <dbl>	Date <chr>
2020	1	9799	1/2020
2020	10	11949041	10/2020
2020	11	17134026	11/2020
2019	12	27	12/2019
2020	12	8642838	12/2020
2020	2	75422	2/2020

6 rows

```
ggplot(df5, aes(x=total_cases,y=Date)) +
  geom_bar(stat='identity',position='dodge',fill='blue')
```




```
df6 = df%>%
  select(month,deaths,year)%>%
  group_by(year,month)%>%
  summarise(total_deaths=sum(deaths))%>%
  arrange(year)%>%
  mutate(Date=paste(as.character(month),as.character(year), sep="/"))%>%
  arrange(Date)
```

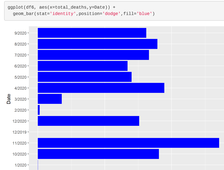
```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

```
head(df6)
```

year <dbl>	month <dbl>	total_deaths <dbl>	Date <chr>
2020	1	213	1/2020
2020	10	181054	10/2020
2020	11	271086	11/2020
2019	12	0	12/2019
2020	12	151565	12/2020
2020	2	2708	2/2020

6 rows

```
ggplot(df6, aes(x=total_deaths,y=Date)) +
  geom_bar(stat='identity',position='dodge',fill='blue')
```



We grouped again "deaths" and "cases" by date in df7, from which we printed the two dates of 2020 with the highest of these two values: the second of December was the day with more deaths, while the eleventh of December the day with more cases.

```
df7 = df%>%
  select(day,month,year,cases,deaths)%>%
  group_by(day,year,month)%>%
  summarise(cases=sum(cases),deaths=sum(deaths))
```

```
## `summarise()` has grouped output by 'day', 'year'. You can override using the `.groups` argument.
```

```
head(df7)
```

day <dbl>	year <dbl>	month <dbl>	cases <dbl>	deaths <dbl>
1	2020	1	0	0
1	2020	2	2121	46
1	2020	3	1843	58
1	2020	4	74647	4677
1	2020	5	85674	5650
1	2020	6	105626	3914

6 rows

```
df7[which.max(df7$deaths),]
```

day <dbl>	year <dbl>	month <dbl>	cases <dbl>	deaths <dbl>
2	2020	12	591801	13786

1 row

```
df7[which.max(df7$cases),]
```

day <dbl>	year <dbl>	month <dbl>	cases <dbl>	deaths <dbl>
11	2020	12	693352	12327

1 row