Top 200 common passwords by country 2021

Autor: Truman Tapia

December 2021

Dataset Description

The dataset used in this work is Top 200 common passwords by country 2021. Lets start by loading the data:

```
data <- read.csv("top_200_password_2020_by_country.csv")
dim(data)
```

```
## [1] 9800    8
```

We can see that the dataset has 9800 rows and 8 columns. The structure of the data is:

```
str(data)
```

```
## 'data.frame':    9800 obs. of  8 variables:
##  $ country_code           : chr  "au" "au" "au" "au" ...
##  $ country                : chr  "Australia" "Australia" "Australia" "Australia" ...
##  $ Rank                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Password               : chr  "123456" "password" "lizottes" "password1" ...
##  $ User_count             : int  308483 191880 98220 86884 75856 69344 68434 67130 37675 30844 ...
##  $ Time_to_crack          : chr  "< 1 second" "< 1 second" "3 Hours" "< 1 second" ...
##  $ Global_rank            : num  1 5 NA 16 2 3 15 4 6 NA ...
##  $ Time_to_crack_in_seconds: num  0 0 10800 0 0 0 0 0 0 120 ...
```

The columns of the dataset are:

- country_code: 2-digit country code (string).
- country: country name (string).
- Rank: ranking inside country (integer).
- Password: actual password (string).
- User_count: number of users that use the password inside a country (integer).
- Time_to_crack: time to crack the password (string).
- Global_rank: global ranking (numeric).
- Time_to_crack_in_seconds: number of seconds required to crack the password (numeric).

This dataset is important because it contains common passwords and also a measure of how weak a passwords is (*Time_to_crack_in_seconds*). Analyzing this data can give us insights about how vulnerable the users are to hack attacks. Also, we can compare the data among countries to detect the countries most vulnerable and resilient against possible hack attracts. Some of the insights that will be obtained from this analysis can be important in the technological society today.

Data cleaning

Given that the pairs of columns *country_code*-*country* and *Time_to_crack*-*Time_to_crack_in_seconds* contain similar information, I am going to get rid of one column in each pair:

```
data <- data[, colnames(data)[c(-1,-6)]]
```

Show the summary of data:

```
summary(data)
```

```
##    country               Rank          Password           User_count      
##  Length:9800        Min.   :  1.00   Length:9800        Min.   :     126  
##  Class :character   1st Qu.: 50.75   Class :character   1st Qu.:    1643  
##  Mode  :character   Median :100.50   Mode  :character   Median :    3948  
##                     Mean   :100.50                      Mean   :   34686  
##                     3rd Qu.:150.25                      3rd Qu.:   11518  
##                     Max.   :200.00                      Max.   :19000630  
##                                                                           
##   Global_rank     Time_to_crack_in_seconds
##  Min.   :  1.00   Min.   :0.000e+00       
##  1st Qu.: 22.00   1st Qu.:0.000e+00       
##  Median : 50.00   Median :2.000e+00       
##  Mean   : 65.34   Mean   :2.083e+06       
##  3rd Qu.:101.50   3rd Qu.:3.600e+02       
##  Max.   :200.00   Max.   :3.214e+09       
##  NA's   :6628                             
```

First, we see that only the column *Time_to_crack_in_seconds* have values of 0, but these values should be allowed because they correspond with extremely easy to crack passwords. Second, we have missing values in the variable *Global_rank* only. These correspond to passwords in the top 200 of a country, but these passwords are not popular enough to rank in the global top 200, then NA in this variable is a valid value. Third, lets check for duplicate rows:

```
library(dplyr)

length(unique(data)) == length(data)
```

```
## [1] TRUE
```

There are no duplicated rows. Fourth, we will check for outliers in the data.

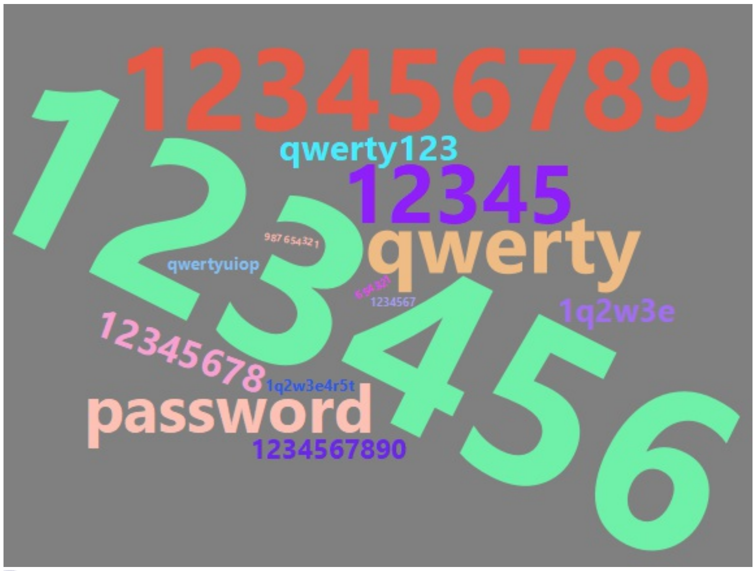
Let's check the column country by showing the countries that appear in the data:

```
table(data$country)
```

```
## 
##            Australia              Austria              Belgium 
##                  200                  200                  200 
##               Brazil               Canada                Chile 
##                  200                  200                  200 
##                China             Colombia       Czech Republic 
##                  200                  200                  200 
##              Denmark              Estonia              Finland 
##                  200                  200                  200 
##               France              Germany               Greece 
##                  200                  200                  200 
##              Hungary                India            Indonesia 
##                  200                  200                  200 
##              Ireland               Israel                Italy 
##                  200                  200                  200 
##                Japan                Korea               Latvia 
##                  200                  200                  200 
##            Lithuania             Malaysia               Mexico 
##                  200                  200                  200 
##          Netherlands          New Zealand              Nigeria
##                  200                  200                  200
##               Norway          Philippines               Poland
##                  200                  200                  200
##             Portugal              Romania               Russia
##                  200                  200                  200
##         Saudi Arabia      Slovak Republic         South Africa
##                  200                  200                  200
##                Spain               Sweden          Switzerland
##                  200                  200                  200
##             Thailand               Turkey              Ukraine
##                  200                  200                  200
## United Arab Emirates       United Kingdom        United States
##                  200                  200                  200
##              Vietnam
##                  200
```

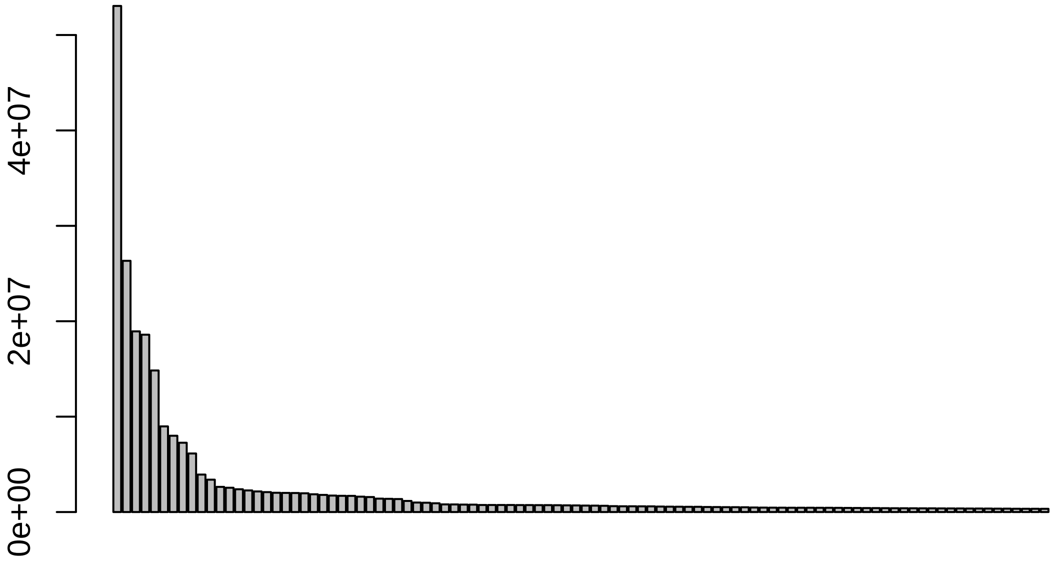
All countries are equally represented (200 times because of the 200 top). A word cloud can be useful to get an idea of the passwords in the column Password:

```
library(RColorBrewer)
library(wordcloud2)

password_frec <- data %>%
  group_by(Password) %>%
  summarise(frecuency=sum(User_count), time_to_crack=min(Time_to_crack_in_seconds)) %>%
  arrange(desc(frecuency))

wordcloud2(data=password_frec, size=1, color="random-light", backgroundColor="grey")
```



The most frequent passwords (the big ones) are simple sequence of numbers or words, like "123456" or "qwerty". Also, from the size of the passwords one sees that there is a peak of people using simple passwords. We can see this peak in the following distribution of the frequencies of users with the 100 most popular passwords:

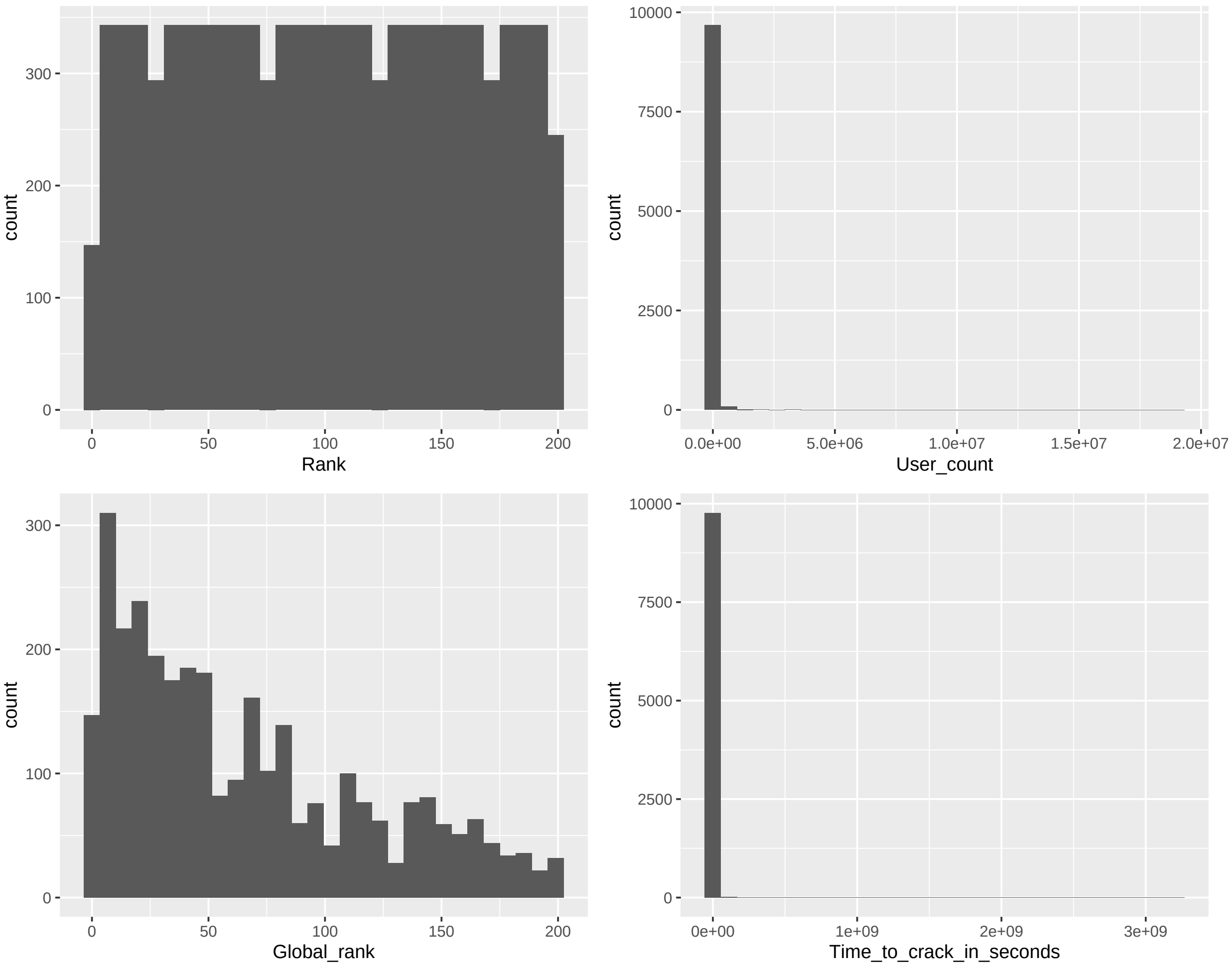
```
barplot(password_frec$frecuency[1:100])
```



It is interesting to realize that the popularity of passwords are highly packed in a few passwords (which can be beneficial to hackers).

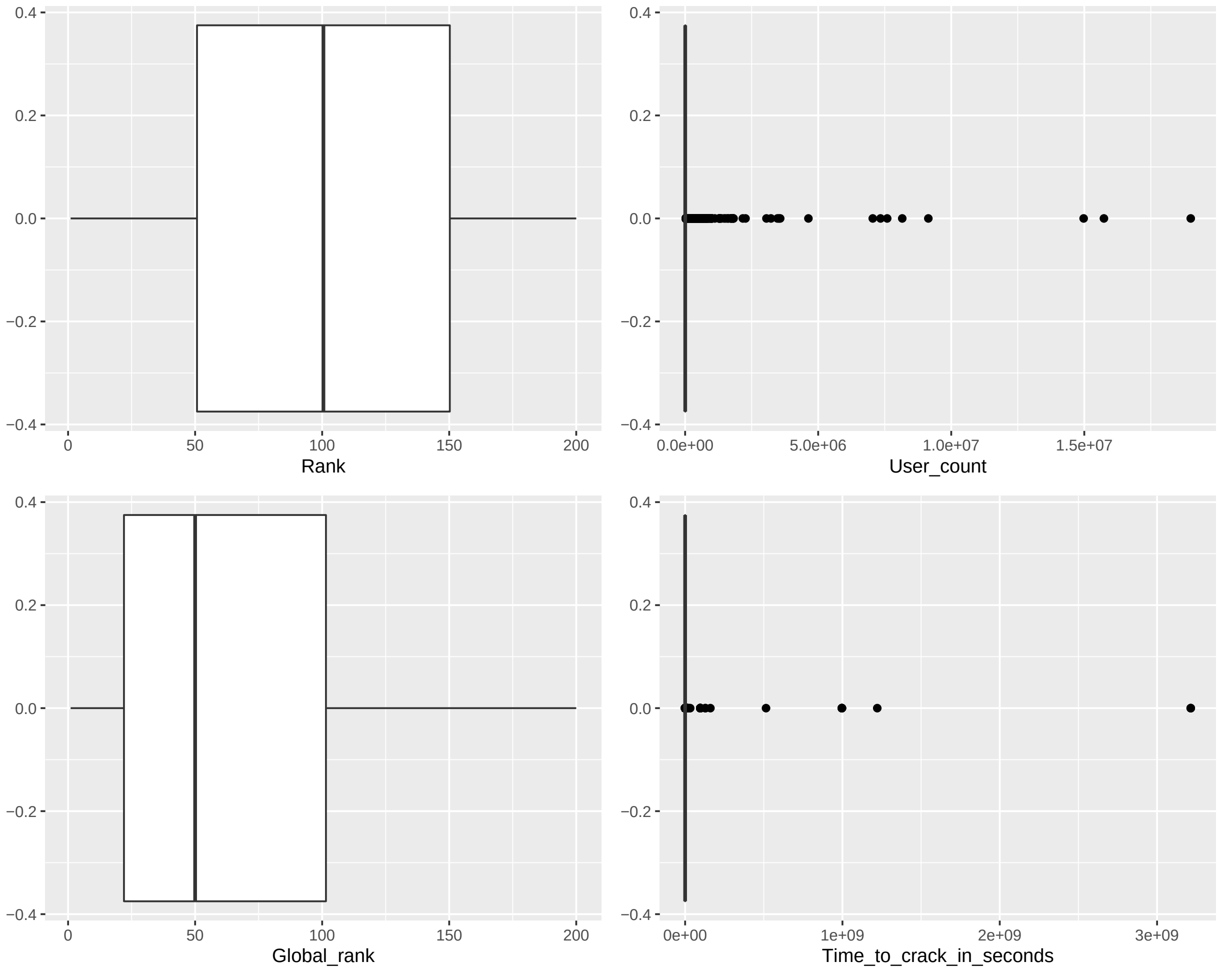
Analyzing the distributions of the numeric variables can be useful to identify outliers and inconsistencies in the data. Following the distribution of the numeric variables of the dataset:

```
library(ggplot2)
library(patchwork)

plot1 <- ggplot(data, aes(Rank)) + geom_histogram()
plot2 <- ggplot(data, aes(User_count)) + geom_histogram()
plot3 <- ggplot(data, aes(Global_rank)) + geom_histogram()
plot4 <- ggplot(data, aes(Time_to_crack_in_seconds)) + geom_histogram()

plot1+plot2+plot3+plot4+plot_layout(ncol=2)
```



The ranking by country (*Rank*) is normally distributed (as wee saw each country has 200 passwords). On the other hand, the global ranking shows that the top 50 global passwords are the most repeated passwords among several countries, as can be expected. The time to crack a password and the user count show similar distributions. This tell us that most passwords are easy to crack and most passwords are not very popular. Lets see the box plots of these variables:

```
plot1 <- ggplot(data, aes(Rank)) + geom_boxplot(outlier.colour="black", outlier.shape=16,
             outlier.size=2)
plot2 <- ggplot(data, aes(User_count)) + geom_boxplot(outlier.colour="black", outlier.shape=16,
             outlier.size=2)
plot3 <- ggplot(data, aes(Global_rank)) + geom_boxplot(outlier.colour="black", outlier.shape=16,
             outlier.size=2)
plot4 <- ggplot(data, aes(Time_to_crack_in_seconds)) + geom_boxplot(outlier.colour="black", outlier.shap
             outlier.size=2)

plot1+plot2+plot3+plot4+plot_layout(ncol=2)
```



In the variables *Rank* and *Global_rank* we do not have possible outliers. In the variable *User_count* and *Time_to_crack_in_seconds* we have very sharp distributions because the big majority of values are small, and some possible outliers have large scope. In reality these points are not outliers because they are real data entries. In the case of *User_count* most passwords are not highly popular, but those which are really popular seem outliers. And, in *Time_to_crack_in_seconds* most passwords are easy to crack, but few passwords are difficult to crack and these look like outliers.

Analysis

For this dataset we plan to answer 4 questions with the help of statistical tests. First, the easiest country where to crack passwords. Second, the hardest country where to crack a password. Third, is there a difference in the time to crack a password in the countries of Latin America? Fourth, the correlations between the variables.

What is the country where it is easiest to crack passwords? Is there a statistical difference between the easiest country and the second easiest?

To answer this question we will consider the mean of *Time_to_crack_in_seconds* by country as the key quantity to determine how difficult to crack a password is. Also, we will compare the sample data of the easiest and second easiest countries, to be sure that it is easier to crack a password in one country than in the other. First, we need to calculate the mean of the *Time_to_crack_in_seconds* for each country and get the countries with the lowest means:

```
means <- rep(0, length(unique(data$country)))
i <- 1

for (country in unique(data$country)){
  data.country <- data[data$country==country, ]
  means[i] <- sum(data.country$Time_to_crack_in_seconds*data.country$User_count)/sum(data.country$User_c
  i <- i+1
}

means <- data.frame(country=unique(data$country), mean_time_to_crack=means)
arrange(means, mean_time_to_crack)
```

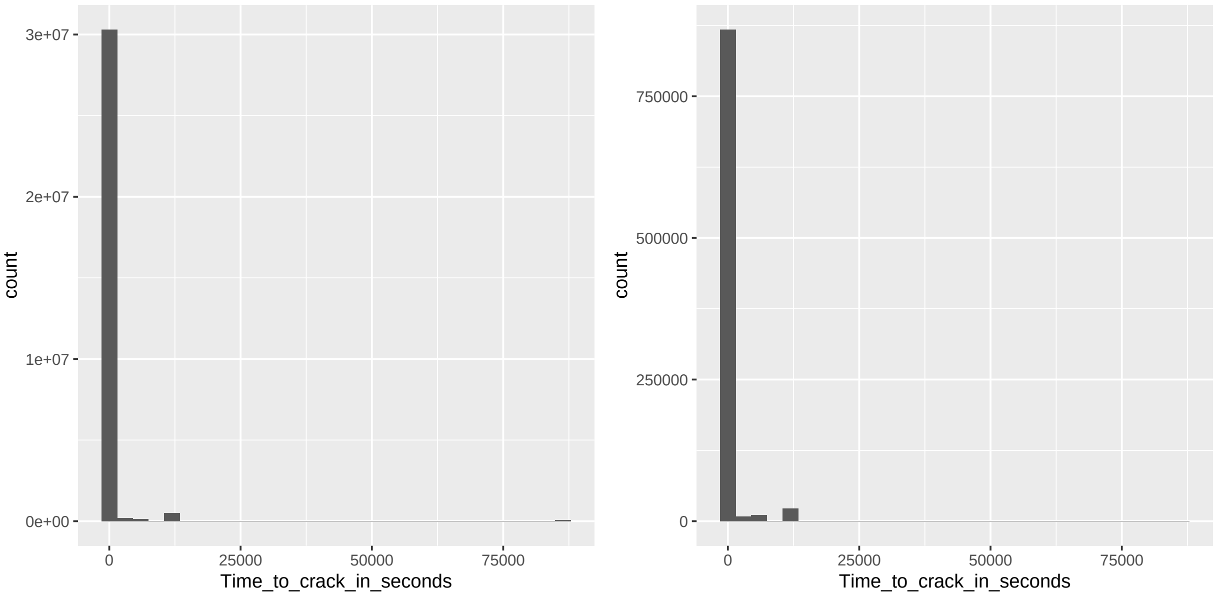
```
##                   country mean_time_to_crack
## 1           United States       4.397991e+02
## 2                   Korea       4.770809e+02
## 3    United Arab Emirates       7.352147e+02
## 4            South Africa       1.338933e+03
## 5                 Ireland       1.539977e+03
## 6          United Kingdom       1.600696e+03
## 7                 Hungary       2.031357e+03
## 8                  Israel       3.187626e+03
## 9                   Japan       3.386709e+03
## 10                 Sweden       3.702053e+03
## 11                 Norway       3.768317e+03
## 12        Slovak Republic       5.137800e+03
## 13                 Poland       6.096373e+03
## 14               Malaysia       6.861806e+03
## 15            Philippines       1.106513e+04
## 16                Ukraine       1.208829e+04
## 17                Finland       1.397032e+04
## 18                 Latvia       1.411447e+04
## 19                 France       1.575053e+04
## 20                  China       1.711886e+04
## 21                Romania       1.851171e+04
## 22                 Russia       2.026753e+04
## 23              Australia       2.455119e+04
## 24         Czech Republic       3.222404e+04
## 25            Switzerland       3.346555e+04
## 26                Belgium       3.368850e+04
## 27                Estonia       3.643986e+04
## 28                Nigeria       3.772692e+04
## 29           Saudi Arabia       4.020591e+04
## 30                  Italy       5.556758e+04
## 31            New Zealand       6.492299e+04
## 32               Thailand       7.176485e+04
## 33                Germany       1.340900e+05
## 34                 Mexico       1.349567e+05
## 35              Lithuania       1.364436e+05
## 36               Colombia       1.390695e+05
## 37                Denmark       1.879990e+05
## 38                  India       3.255156e+05
## 39            Netherlands       3.619271e+05
## 40                Austria       4.870151e+05
## 41                 Greece       7.006281e+05
## 42              Vietnam       7.433826e+05
## 43               Canada       1.093843e+06
## 44               Turkey       1.416007e+06
## 45                Spain       1.436433e+06
## 46             Portugal       1.733466e+06
## 47               Brazil       3.602670e+06
## 48                Chile       5.386567e+06
## 49            Indonesia       1.484593e+07
```

The two countries with the lowest means are US and Korea. Let's see if the difference in these two countries is statistically meaningful. Thus, we start by transforming the frequency data to samples:

```
library(vcdExtra)

data.us <- data[data$country=="United States", c("Time_to_crack_in_seconds", "User_count")]
data.us <- expand.dft(data.us, freq="User_count")

data.korea <- data[data$country=="Korea", c("Time_to_crack_in_seconds", "User_count")]
data.korea <- expand.dft(data.korea, freq="User_count")
```

To test if the difference in the data of these two countries is statistically meaningful, we must decide whether using a parametric or a non-parametric test. Parametric tests require that the data fulfill normality (follows a normal distribution), we test this by plotting histograms:

```
plot1 <- ggplot(data.us, aes(Time_to_crack_in_seconds)) + geom_histogram()
plot2 <- ggplot(data.korea, aes(Time_to_crack_in_seconds)) + geom_histogram()
plot1+plot2+plot_layout(ncol=2)
```



From the histograms we can see that the data of the two countries do not fulfill the normality condition. And, this can be checked by the Kolmogorov-Smirnov test:

```
ks.test(c(data.us$Time_to_crack_in_seconds, data.korea$Time_to_crack_in_seconds), pnorm, mean(c(data.us
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
```

```
## data:  c(data.us$Time_to_crack_in_seconds, data.korea$Time_to_crack_in_seconds)
## D = 0.46534, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

The null hypothesis (data follows a normal distribution) is rejected based on the p-value. Now, lets test homoscedasticity because it is another requirement for parametric tests:

```
fligner.test(list(data.korea$Time_to_crack_in_seconds, data.us$Time_to_crack_in_seconds))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(data.korea$Time_to_crack_in_seconds, data.us$Time_to_crack_in_seconds)
## Fligner-Killeen:med chi-squared = 1433.9, df = 1, p-value < 2.2e-16
```

The hypothesis of homogeneity of variances is rejected. Then, given that normality and homoscedasticity fails for the data of these two countries, I will use a non parametric test to test if the distributions of these two groups are different or not:

```
wilcox.test(data.korea$Time_to_crack_in_seconds, data.us$Time_to_crack_in_seconds)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data.korea$Time_to_crack_in_seconds and data.us$Time_to_crack_in_seconds
## W = 1.4518e+13, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

The Wilconxon test rejects the null hypothesis, then the distributions of *Time_to_crack_in_seconds* for the countries US and Korea are different, and surely US is the easiest country to crack a password.

What is the country where it is hardest to crack passwords? Is there a statistical difference between the hardest country and the second hardest?

We will follow a similar approach as in the last question. First, get the countries with the highest mean times to crack passwords:

```
arrange(means, desc(mean_time_to_crack))
```

```
##                country mean_time_to_crack
## 1            Indonesia       1.484593e+07
## 2                Chile       5.386567e+06
## 3               Brazil       3.602670e+06
## 4             Portugal       1.733466e+06
## 5                Spain       1.436433e+06
## 6               Turkey       1.416007e+06
## 7               Canada       1.093843e+06
## 8              Vietnam       7.433826e+05
## 9               Greece       7.006281e+05
## 10             Austria       4.870151e+05
## 11         Netherlands       3.619271e+05
## 12               India       3.255156e+05
## 13             Denmark       1.879990e+05
## 14            Colombia       1.390695e+05
## 15           Lithuania       1.364436e+05
## 16              Mexico       1.349567e+05
## 17             Germany       1.340900e+05
## 18            Thailand       7.176485e+04
## 19          New Zealand       6.492299e+04
## 20                Italy       5.556758e+04
## 21         Saudi Arabia       4.020591e+04
## 22              Nigeria       3.772692e+04
## 23              Estonia       3.643986e+04
## 24              Belgium       3.368850e+04
## 25          Switzerland       3.346555e+04
## 26       Czech Republic       3.222404e+04
## 27            Australia       2.455119e+04
## 28               Russia       2.026753e+04
## 29              Romania       1.851171e+04
## 30                China       1.711886e+04
## 31               France       1.575053e+04
## 32               Latvia       1.411447e+04
## 33              Finland       1.397032e+04
## 34              Ukraine       1.208829e+04
## 35          Philippines       1.106513e+04
## 36             Malaysia       6.861806e+03
## 37               Poland       6.096373e+03
## 38      Slovak Republic       5.137800e+03
## 39               Norway       3.768317e+03
## 40               Sweden       3.702053e+03
## 41                Japan       3.386709e+03
## 42               Israel       3.187626e+03
## 43              Hungary       2.031357e+03
## 44       United Kingdom       1.600696e+03
## 45              Ireland       1.539977e+03
## 46         South Africa       1.338933e+03
## 47 United Arab Emirates       7.352147e+02
## 48                Korea       4.770809e+02
## 49        United States       4.397991e+02
```

In this case the hardest countries are Indonesia and Chile. Get the sample data of these countries:

```
data.chile <- data[data$country=="Chile", c("Time_to_crack_in_seconds", "User_count")]
data.chile <- expand.dft(data.chile, freq="User_count")

data.indonesia <- data[data$country=="Indonesia", c("Time_to_crack_in_seconds", "User_count")]
data.indonesia <- expand.dft(data.indonesia, freq="User_count")
```

We make a Kolmogorov-Smirnov test to determine whether a parametric test applies or not to this subset of data:

```
ks.test(c(data.chile$Time_to_crack_in_seconds, data.indonesia$Time_to_crack_in_seconds),
        pnorm,
        mean(c(data.chile$Time_to_crack_in_seconds, data.indonesia$Time_to_crack_in_seconds)),
        sd(c(data.chile$Time_to_crack_in_seconds, data.indonesia$Time_to_crack_in_seconds)))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  c(data.chile$Time_to_crack_in_seconds, data.indonesia$Time_to_crack_in_seconds)
## D = 0.50966, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

The test rejects the null hypothesis as in the last section, then we will use a non-parametric test directly:

```
wilcox.test(data.chile$Time_to_crack_in_seconds, data.indonesia$Time_to_crack_in_seconds)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data.chile$Time_to_crack_in_seconds and data.indonesia$Time_to_crack_in_seconds
## W = 1.5733e+12, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

The distributions of the two countries are statistically different, then the hardest country to crack a password is Indonesia.

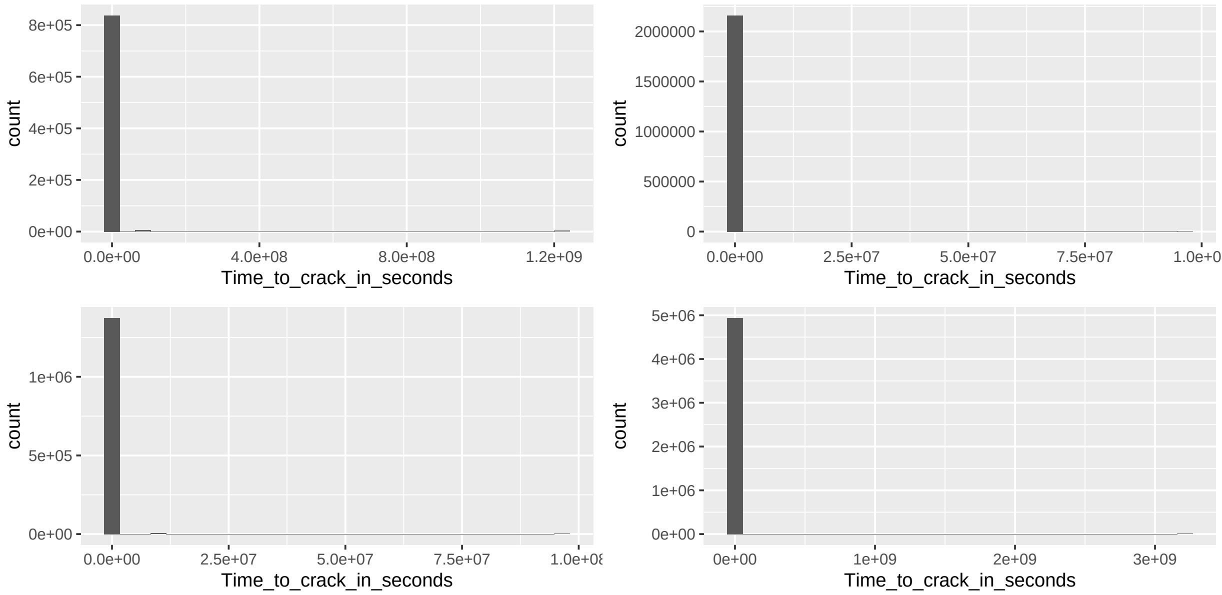
Is there any statistical difference in the time to crack a password among the countries in Latin America?

Answering this question can tell us if the people of countries of the same region follow similar patterns when dealing with their passwords. To answer this question we will compare the samples of the countries: Chile, Mexico, Colombia and Brazil. First, lets get the sample data of these countries (Chile was already obtained):

```
data.mexico <- data[data$country=="Mexico", c("Time_to_crack_in_seconds", "User_count")]
data.mexico <- expand.dft(data.mexico, freq="User_count")

data.colombia <- data[data$country=="Colombia", c("Time_to_crack_in_seconds", "User_count")]
data.colombia <- expand.dft(data.colombia, freq="User_count")

data.brazil <- data[data$country=="Brazil", c("Time_to_crack_in_seconds", "User_count")]
data.brazil <- expand.dft(data.brazil, freq="User_count")
```

Now, plot the histogram of these countries:

```
plot1 <- ggplot(data.chile, aes(Time_to_crack_in_seconds)) + geom_histogram()
plot2 <- ggplot(data.mexico, aes(Time_to_crack_in_seconds)) + geom_histogram()
plot3 <- ggplot(data.colombia, aes(Time_to_crack_in_seconds)) + geom_histogram()
plot4 <- ggplot(data.brazil, aes(Time_to_crack_in_seconds)) + geom_histogram()
plot1+plot2+plot3+plot4+plot_layout(ncol=2)
```



We see that the data do not follow a normal distribution, but let's corroborate this with a Kolmogorov-Smirnov test:

```
ks.test(c(data.chile$Time_to_crack_in_seconds, data.mexico$Time_to_crack_in_seconds, data.colombia$Time_
        pnorm,
        mean(c(data.chile$Time_to_crack_in_seconds, data.mexico$Time_to_crack_in_seconds, data.colombia$
        sd(c(data.chile$Time_to_crack_in_seconds, data.mexico$Time_to_crack_in_seconds, data.colombia$Ti
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  c(data.chile$Time_to_crack_in_seconds, data.mexico$Time_to_crack_in_seconds, data.colombia$Tim
## D = 0.50291, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Data does not follow a normal distribution, then a non-parametric tests will be used. To determine if there are statistical differences between the samples of these countries Kruskal-Wallis test will be used:

```
kruskal.test(list(data.chile$Time_to_crack_in_seconds, data.mexico$Time_to_crack_in_seconds, data.colom
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(data.chile$Time_to_crack_in_seconds, data.mexico$Time_to_crack_in_seconds, data.colombia$
## Kruskal-Wallis chi-squared = 96419, df = 3, p-value < 2.2e-16
```

The null hypothesis (countries follow the same distribution) is rejected, then the time to crack passwords in the countries of Latin American is different.

What is the correlation between the numerical variables of the dataset?

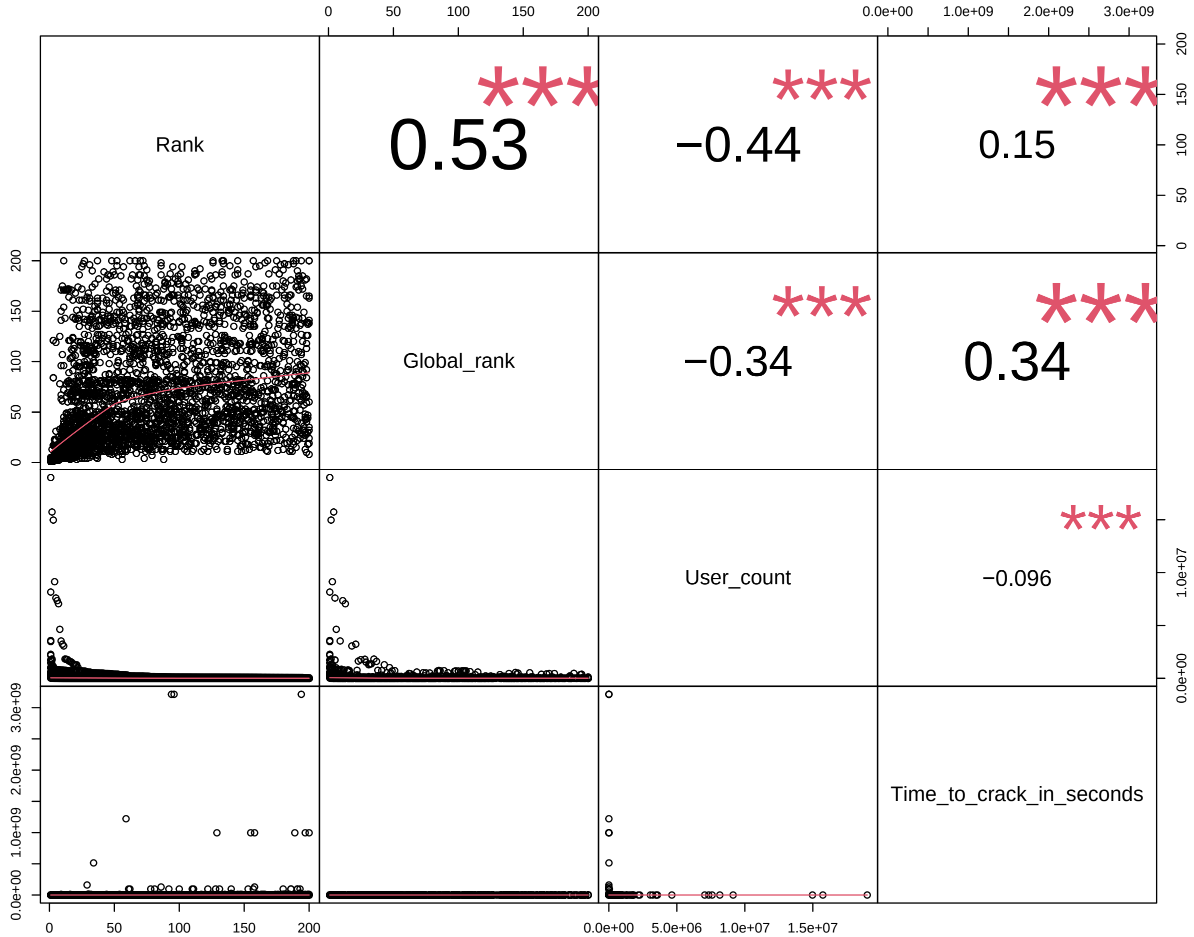
Following, the correlation between the numeric variables of the dataset along with the number of registers used and the p-values. Spearman correlation is used because we saw in the histograms that the variables do not follow a normal distribution:

```
library(Hmisc)

rcorr(as.matrix(data[,c("Rank", "Global_rank", "Time_to_crack_in_seconds", "User_count")]),type="spearma
```

```
##                            Rank Global_rank Time_to_crack_in_seconds User_count
## Rank                       1.00        0.53                     0.15      -0.44
## Global_rank                0.53        1.00                     0.34      -0.34
## Time_to_crack_in_seconds   0.15        0.34                     1.00      -0.10
## User_count                -0.44       -0.34                    -0.10       1.00
##
## n
##                          Rank Global_rank Time_to_crack_in_seconds User_count
## Rank                     9800        3172                     9800       9800
## Global_rank              3172        3172                     3172       3172
## Time_to_crack_in_seconds 9800        3172                     9800       9800
## User_count               9800        3172                     9800       9800
##
## P
##                          Rank Global_rank Time_to_crack_in_seconds User_count
## Rank                           0           0                        0
## Global_rank               0                0                        0
## Time_to_crack_in_seconds  0    0                                    0
## User_count                0    0           0
```

The variables that are more correlated are *User_count-Rank* and *Rank-Global_rank*. We can see in the p-values matrix (the last matrix) that all p-values are approximately 0, then correlation values are statistically significant. Given that the variables *Rank*, *Global_rank* and *User_count* depend on each other in an obvious way, we can focus our attention in the relationship of *Time_to_crack_in_seconds* and *User_count*. Because of the low correlation value that these two variables have, we can saw that they are not linearly related. We can get a better sense of the relationship between variables by plotting a scatter plot:

```
library("PerformanceAnalytics")

chart.Correlation(as.matrix(data[,c("Rank", "Global_rank", "User_count", "Time_to_crack_in_seconds")]),
```



The last graphic shows a scatter plot of the variables under the diagonal, and the correlation matrix above the diagonal where the asterisk show the statistical significant of the correlation (p-value information). In the scatter plot we notice that the variables are not correlated. About the scatter plot of *User_count* and *Time_to_crack_in_seconds* we have a few passwords with low User_count which are difficult to crack, but most passwords require little time to be cracked; there is no correlation between these variables.

Conclusions

In this work we have analysed some aspects of the dataset *Top 200 most common passwords*. At the beginning data preprocesing was done checking for 0 values, missings and outliers in the data. Also, variables were represented graphically through word cloud, histograms, bar plots and box plots; here it was noticed that there is a peak of people using the simplest passwords (number and/or letter sequences). Then, the analysis focused in answering the four questions proposed with the help of statistical tests. From the analysis, we found that the easiest country to crack a password is US, while the most difficult is Indonesia. In addition, the distributions of the time to crack a password among the countries of Latin America are not equivalent. Finally, the correlation between the variables of the dataset was studied with special attention in the variables *User_count* and *Time_to_crack_in_seconds*. No strong correlation between any of the variables was found, and most users have passwords with a low time to be cracked.