Assignment 1

Exploratory Data Analysis

Instructions

Using exploratory data analysis techniques, you must extract relevant information from a real database. The analyses, as well as the figures and conclusions, should be collected in a report with a **maximum length of 5 pages**. The source code should be included in a separated R script. The assignment can be submitted **individually or in pairs**. The evaluation criteria can be found at the end of this document.

Data

The file **titanic_train.RDATA** contains information about 668 passengers of the transatlantic ship Titanic. To read the data you must use the `load` function:

```r
load("titanic_train.RDATA")
head(titanic.train)
```

```
##    Survived Pclass    Sex Age SibSp Parch   Ticket    Fare Cabin Embarked
## 2         1      1 female  38     1     0 PC 17599 71.2833   C85        C
## 5         0      3   male  35     0     0   373450  8.0500              S
## 7         0      1   male  54     0     0    17463 51.8625   E46        S
## 10        1      2 female  14     1     0   237736 30.0708              C
## 11        1      3 female   4     1     1  PP 9549 16.7000    G6        S
## 14        0      3   male  39     1     5   347082 31.2750              S
```

The database contains 10 variables that can be described in the form:

Variable	Description
Survived	Survival (0 = No, 1 = Yes)
Pclass	Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
Sex	Sex (female, male)
Age	Age in years
SibSp	number of siblings or spouses aboard the Titanic
Parch	number of parents or children aboard the Titanic
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin number
Embarked	Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

The objective of this work is to carry out an exploratory analysis of the variables contained in the database to summarize their main characteristics and to summarize the relationships between them.

In a first approach, the main characteristics of each variable should be extracted, using the functions `table` and `prop.table` for the categorical variables and `summary` for the numerical variables. These analyses can be accompanied by graphs such as bar charts for the categorical variables, or histograms, density plots and boxplots for the numerical variables.

Then possible relationships between variables are proposed in forms of questions. This possible relations are verified or refuted with the graphical tools we have learned.

One variable of special attention is the **Survival** variable since this work constitutes a previous step to the implementation of different models of machine learning to predict the **Survival** variable based on the remaining ones.

Example:

One question we could ask is whether having any family members on board (whether siblings, spouses, children or parents) played any role in the chances of survival. We easily identified passengers who did not have any family members on board as follows

```
aux = titanic.train$Parch == 0 & titanic.train$SibSp == 0
sum(aux)
```

```
## [1] 392
```

and we could see that more than half of the passengers in this database were travelling alone. We can include a new variable **travels_alone** by attaching the following column

```
travels_alone = rep("No",length(aux))
travels_alone[aux] = "Yes"
titanic.train = cbind(titanic.train, travels_alone)
```

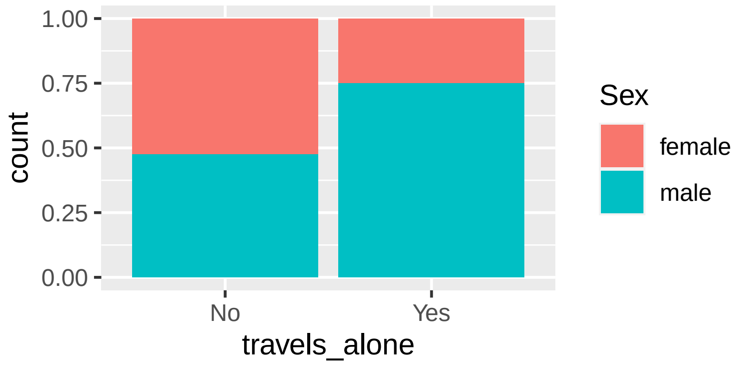
Using the following table we compared the survival ratio of the two groups (those who traveled alone versus those who did not)

```
prop.table(table(titanic.train$travels_alone, titanic.train$Survived),1)
```

```
##
##                 0         1
##    No   0.4927536 0.5072464
##    Yes 0.7040816 0.2959184
```

We observe that the survival ratio of those who travel alone is 30%, much lower than the ratio of those who did not travel alone, which is around 51%. Now, we wonder if being accompanied is a relevant fact, or if, for example, this higher mortality rate may be due to the fact that most of those travelling alone are men, whose survival rate was much lower than that of women as we have already seen. Let's make the following bar chart to graphically show the number of men and women in each group

```
library(ggplot2)
ggplot(titanic.train) + aes(x = travels_alone, fill = Sex ) +
  geom_bar(position = position_fill())
```



Where we confirm our theory that the higher mortality rate of the travelling group is due to the higher proportion of men.

Evaluation

For the evaluation of the work, the number of questions correctly formulated and answered in the report will be taken into account. Specifically, each question will be evaluated according to the following items:

1. Interest of the question.
2. Correct pre-processing of the data and quality of the figure.
3. Quality of the information provided by the figure(s) as well as the preliminary conclusions drawn from the figure(s).
4. Source code.
5. Originality will also be scored!

The maximum score can be obtained with 5 questions. However, it is recommended to do more to ensure a good evaluation. Remember that you can use any of the geometries seen in classes that includes frequency plots, histograms, density plots, scatter plots, boxplots and line plots. You can also use other graphics that you can find referenced in the R Graph Gallery.