R Final Project: The Titanic

Daniel Alonso & Ander Iturburu

October 28th, 2020

The Titanic disaster dataset

Introduction

In this project we will explore the Titanic Passenger's dataset. We intend to make a surface description of the datasets through plots, analysis of such plots and detailed descriptions of our observations relating them. We will also use caret to analyze the relationship of one of our categorical variables in terms of other variables (in our case, mortality of the titanic's passengers).

The project will be divided in two parts: The descriptive analysis and the use of caret to relate the variables. Each part will explore a few key points that we consider quite interesting. We want to create a general idea of the content of the dataset in order to later on effectively know what might have affected the mortality of passengers in the titanic. The teachings from such a tragic event must've taught us something right?

A few key points to explore will be:

- How are our most relevant variables distributed?
- How do these variables respond to grouping by categorical variables in the dataset?
- Which variables are correlated to which?
- What is really relevant when it comes to determining a possible increase in risk of death of the passengers?

Among other things, which hopefully shall paint an accurate full picture when it comes to the risk of death in this tragic event.

First off let's load the libraries we'll use during this project

```r
library(dplyr)
library(ggplot2)
library(fitdistrplus)
library(PerformanceAnalytics)
library(reshape2)
library(vcd)
library(EnvStats)
library(scales)
```

How does the data look?

Importing the data

```
data = read.csv('data/titanic.csv')
```

Dataset head

```
head(data)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                                                   Name
## 1 Braund, Mr. Owen Harris
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer)
## 3 Heikkinen, Miss. Laina
## 4 Futrelle, Mrs. Jacques Heath (Lily May Peel)
## 5 Allen, Mr. William Henry
## 6 Moran, Mr. James
##       Sex Age SibSp Parch              Ticket    Fare           Cabin Embarked
## 1 male     22     1     0 A/5 21171            7.2500                        S
## 2 female   38     1     0 PC 17599            71.2833 C85                    C
## 3 female   26     0     0 STON/O2. 3101282     7.9250                        S
## 4 female   35     1     0 113803              53.1000 C123                   S
## 5 male     35     0     0 373450               8.0500                        S
## 6 male     NA     0     0 330877               8.4583                        Q
```

Ommitting nans on a separate dataset

The purpose of this is to have a dataset with all the content which we will use for predictions, and a dataset with all the data, except nan values. The latter will be used exclusively for plotting and for descriptive analysis.

```
data_n <- na.omit(data)
```

Variables

ID:

- PassengerId: A unique identifying number for every passenger listed in an SQL-like fashion

Other identifying variables:

- Name: Name of the passenger
- Ticket: Ticket number/code for the passenger
- Cabin: Cabin the passenger boarded the cruise in

Continuous:

- Fare: Amount paid by the passenger to board the cruise
- Age: Age of the passenger during the cruise boarding in years

Discrete:

- SibSp: This is the total amount of siblings/step-siblings and spouses that a passenger has
- Sex: the passenger's sex
- Survived: a boolean variable that describes whether the passenger survived or not
- Pclass: a variable that describes the class in which the passenger sailed aboard the cruise ship (1st,2nd,3rd class)
- parch: This is the total amount of parents (mother, father) or children (son, daughter, step-son/daughter) that a given passenger has

Descriptive Analysis

Histogram for Age

```r
hist(data$Age)
```

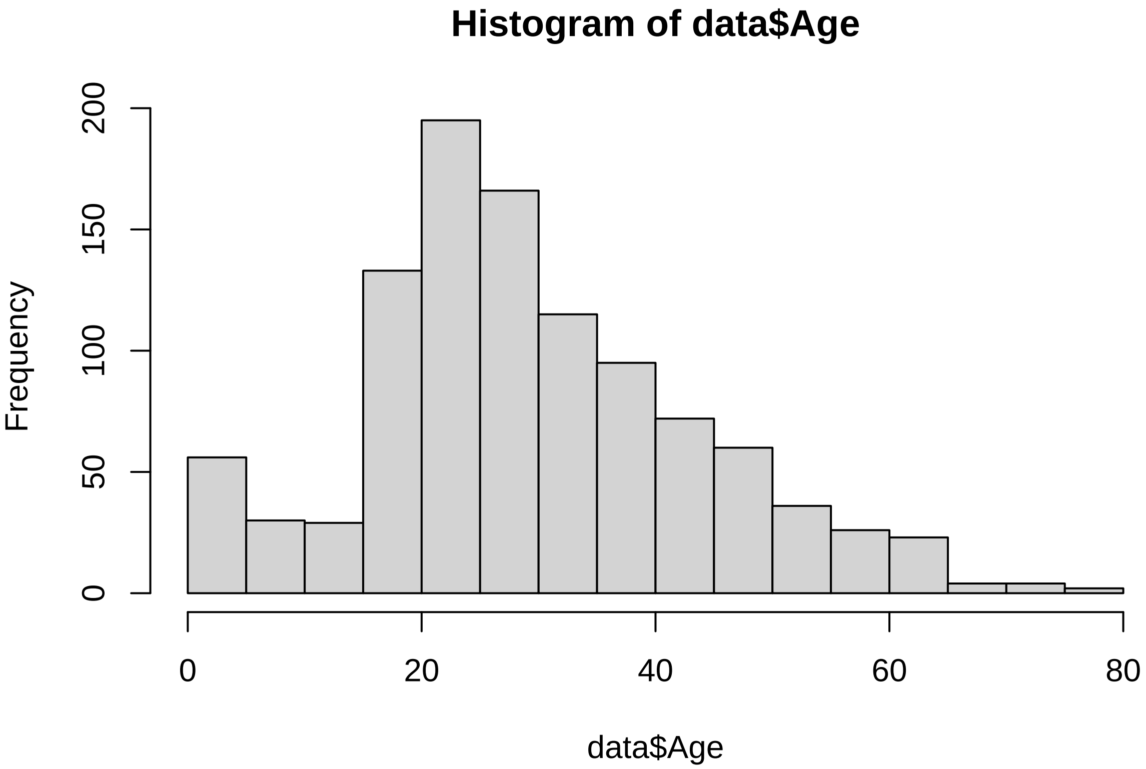

Figure 1

In Figure 1 we can see the age distribution shows the mean age between 25-35 years old. We can see there is a significantly larger right tail than left tail. With a large amount of young children, all, more than likely related to other passengers (their parents). The plot resembles a normal distribution in its concentration of individuals around the mean.

There are very few old people, which we could speculate could be due to the significantly lower life expectancy of back then (between 51 and 55 years old in 1910-1920 UK), or due to the fact that maybe a trip aboard the ship might have been aimed to families. Either way, there aren't many people above 70 (<5) and none above 80.

Histogram for Fare

```r
hist(data$Fare)
```

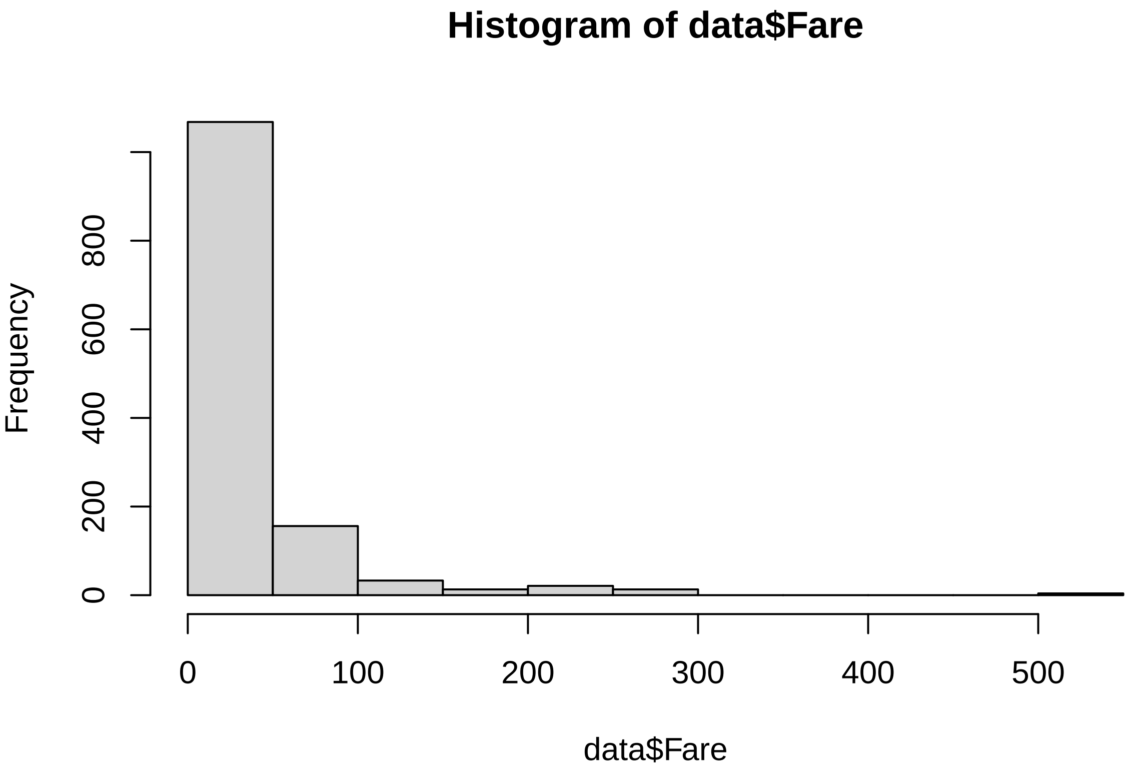

Figure 2

According to Figure 2 on our second continuous variable (Fare), we notice a significantly different situation Vs Age. We have a massively long left tail, as a result of some extremely highly priced tickets payed by a few passengers. Which shows a large and seemingly empty space between fares priced around 250-300 GBP and about 500+ GBP, with little to no tickets payed, showing evidence that some of the tickets, while expensive and very rare among all the individuals (250-300) are still far from the maximum Fare payed (500+). We can clearly see the data is extremely right-skewed.

Extremely expensive tickets, while interesting, are also quite rare. We see that almost all the data concentrates around the 0-100 GBP cost. None being actually free, but yes between 0-1 GBP. We could theorize such cheap tickets probably belonged to children, while most normal tickets were above 20 GBP.

For reference, the ticket prices to board the titanic were the following (in 1912 GBP):

- First Class (parlor suite) — £870
- First Class (berth)— £30
- Second Class — £12
- Third Class — £3 to £8

The conversion rate for £1 (in 1912) would be £115 (in 2019), which shows that even the cheapest tickets (while maybe relatively cheap to board a transatlantic cruise) were still relatively expensive by 1912 standards.

Fitting distributions to data: