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

```
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

Setting default plot sizes

```
options(repr.plot.width = 14, repr.plot.height = 8)
```

Histogram for Age

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

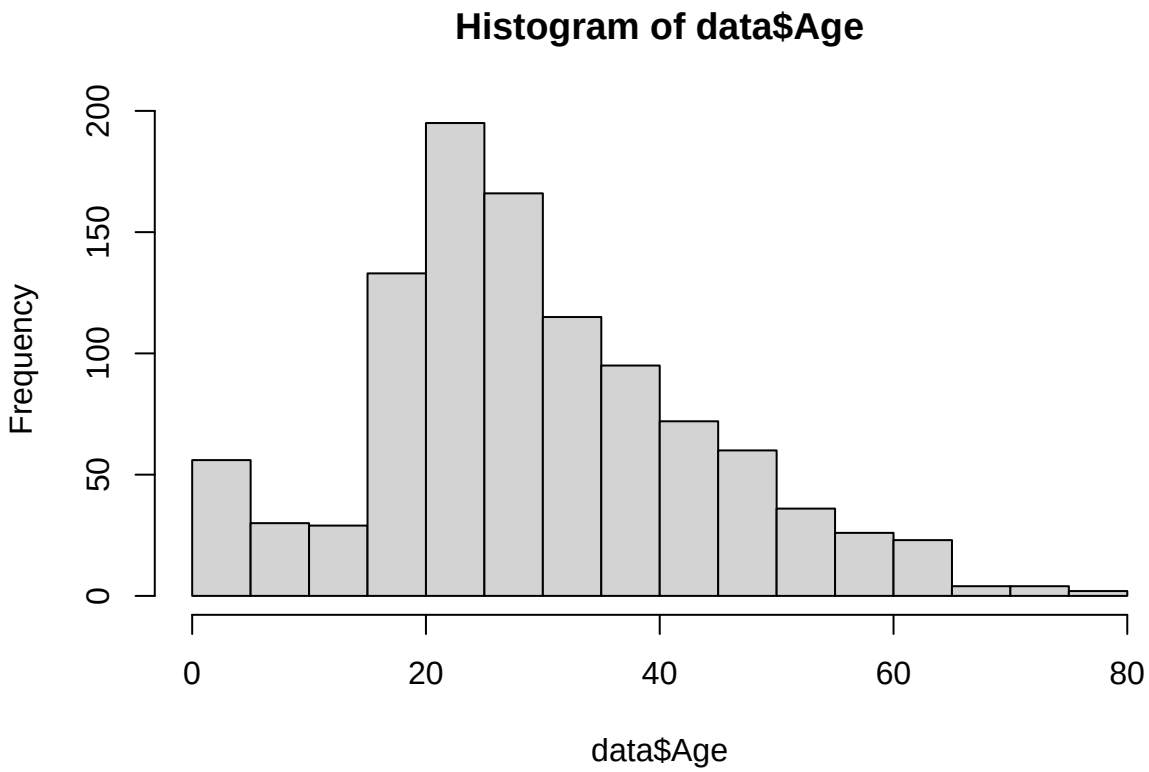


Figure 1

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

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

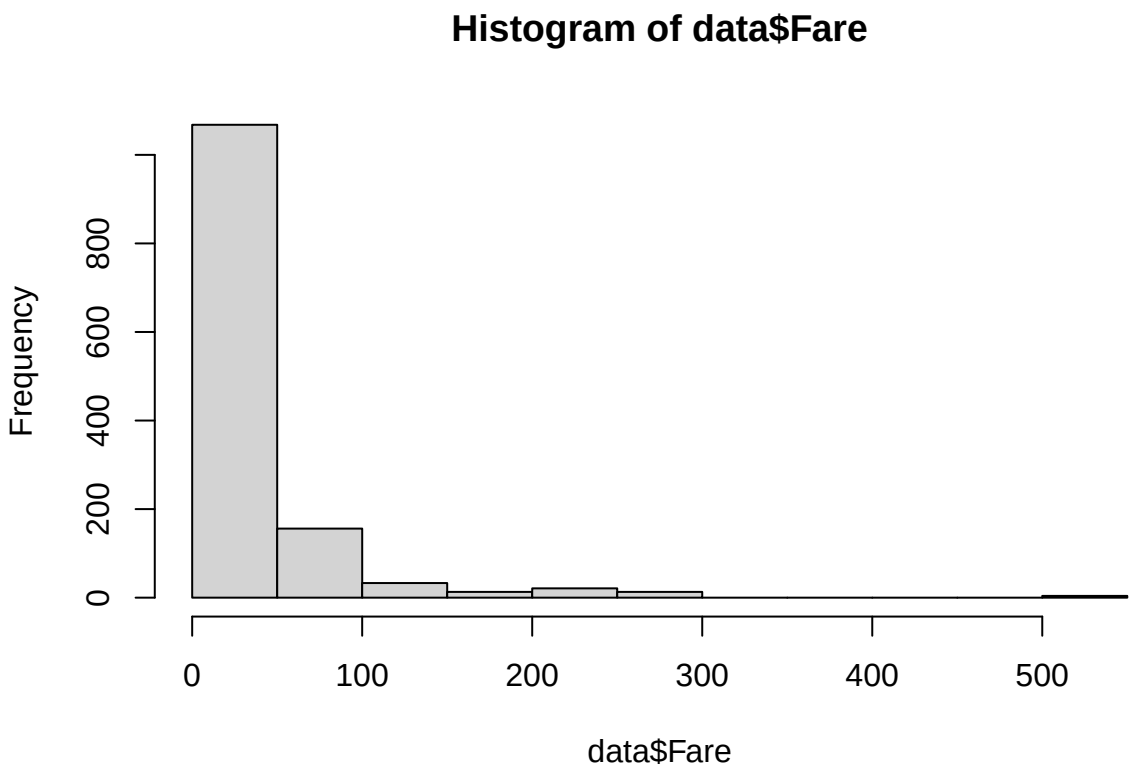


Figure 2

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 the normal distribution to Age

```
plot(summary(fitdist(data_n$Age,"norm")))
```

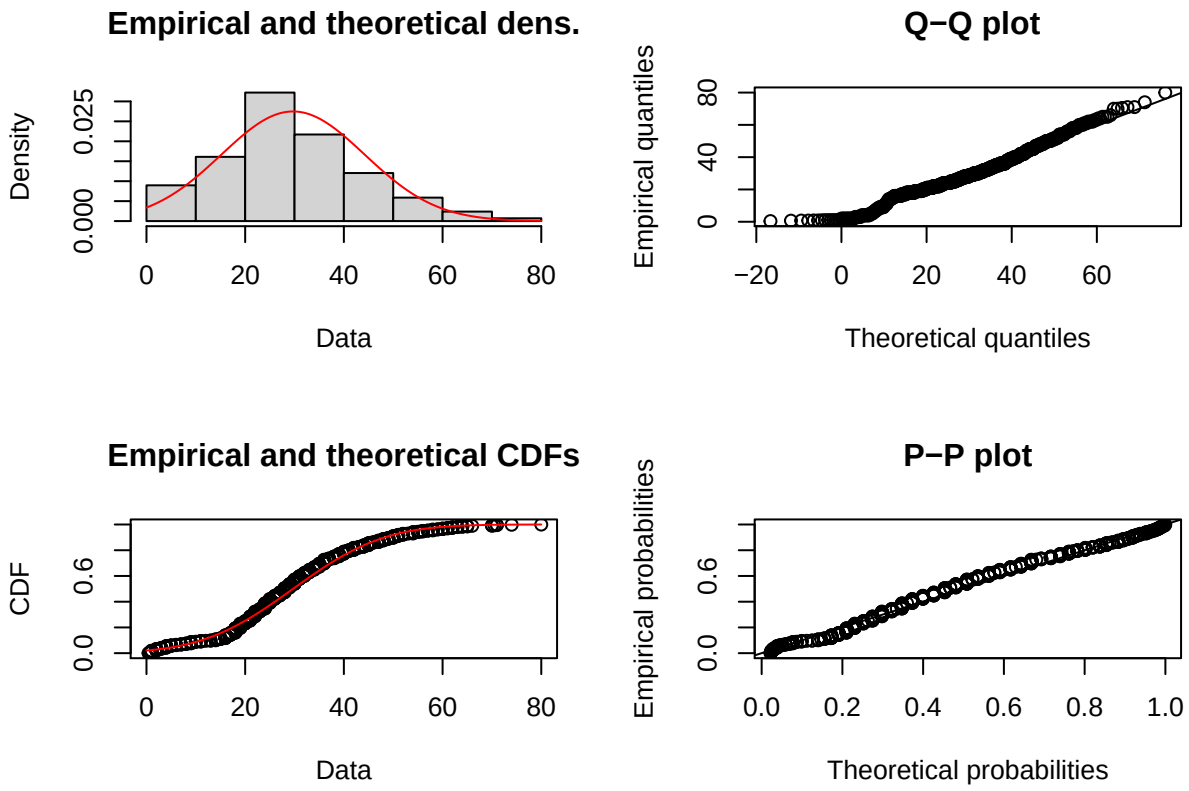


Figure 3

The age variable seems to show relative normality pretty much across the board with the exception of the slightly larger than usual right tail, as we saw in the histogram previously. This right tail can also be seen differring from a normal distribution in the QQ-plot, however, this overall doesn't hurt the relative normality the variable shows.

We can see the mean very very slightly to the right of a gaussian curve and a slightly longer left tail. Given the current size of the sample we could, perhaps, theorize that the set probably comes from a normal distribution.

An interesting thing to think about would be to, with proper statistical analysis, extrapolate the sample data to the population (everyone aboard the titanic) to see if it accurately describes the reality of the whole ship crew plus passengers. It would be out of the scope of this project, but quite interesting to think about.

Fitting a gamma distribution to Fare

```
plot(summary(fitdist(data_n$Fare+1,"gamma")))
```

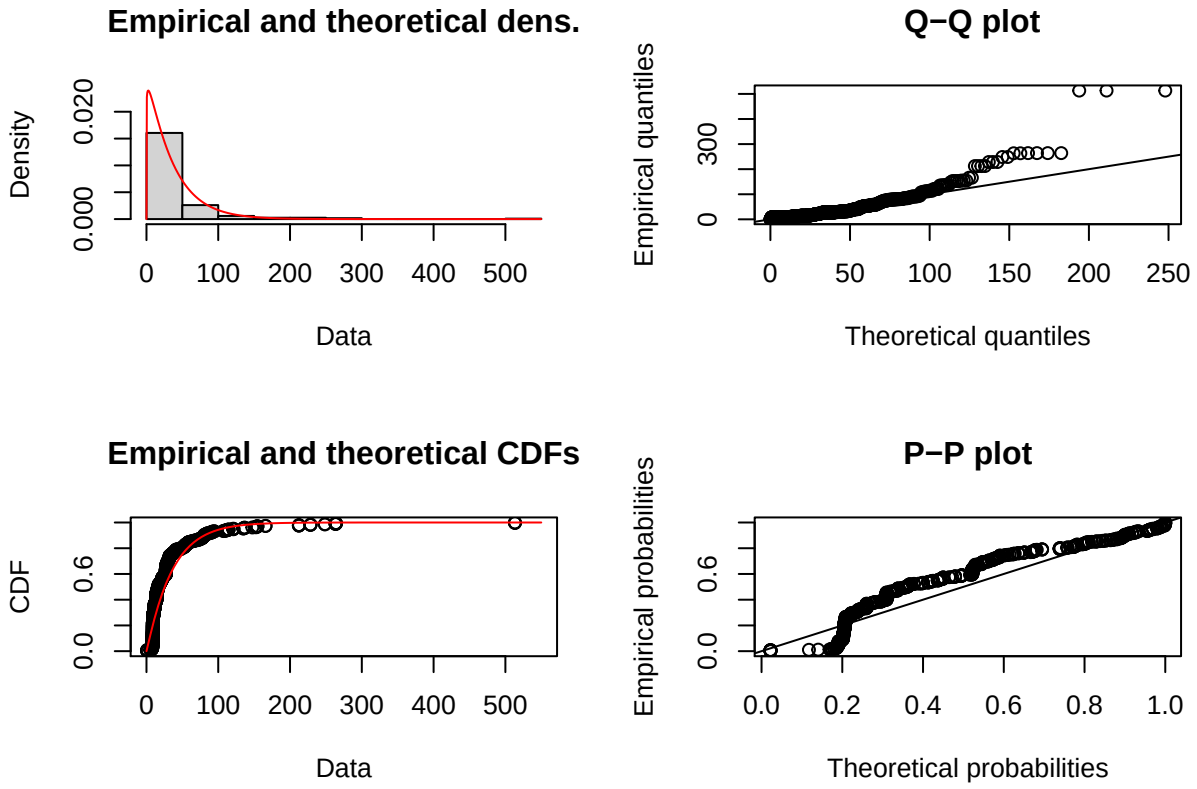


Figure 4

First of all, before plotting, we must sum a constant to Fare, we have chosen 1 to keep values relatively close to the original ones. As some fares are too low and by the nature of the gamma distribution these values would explode.

We can see that the data seems to follow a gamma distribution , with the exception of its incredibly long tail, which to an extent seems to fall above the theoretical quantiles in the QQ-plot, while the PP-plot seems to favour the left tail above the right tail.

We particularly like this variable as it is interesting to play with and seems to say a lot about the passengers aboard the boat and definitely does not seem to exhibit normality as most of the entries concentrate above relatively low prices, which matches very well the proportion of people in each class.

Boarding the ship wasn't exactly cheap, but we know for a fact that luxury tickets were significantly more expensive than third/second class tickets, therefore making the amount of people able to afford such cabins significantly lower.

Boxplot for Age

```
five <- data.frame(x=rep(1,5), five=fivenum(data_n$Age))

ggplot(data=data_n, aes(x=0, y=data_n$Age)) +
    geom_boxplot() +
    theme(text = element_text(size=18)) +
    ylab("Age") +
    scale_x_discrete(breaks = NULL) +
    geom_text(data=five, aes(x=0, y=five, label=five),nudge_x =0.5)
```

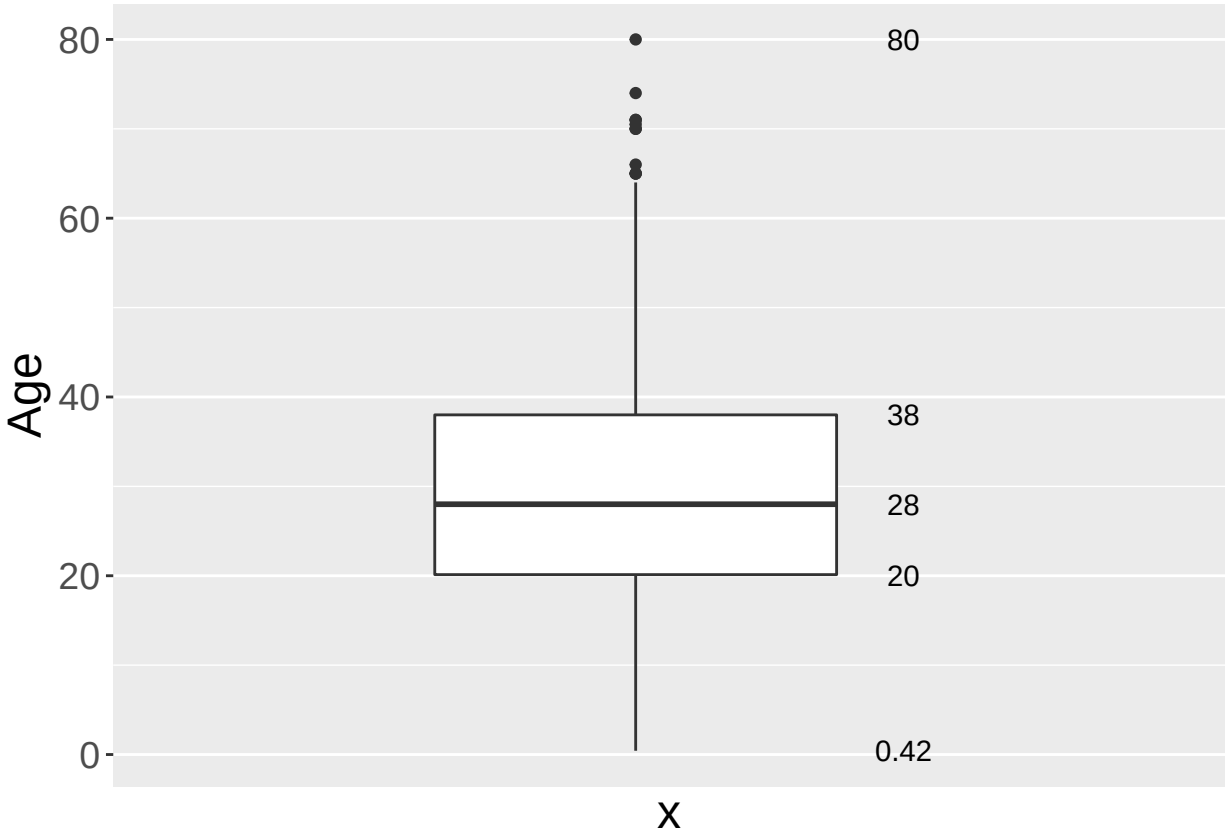


Figure 5

In the Figure 5 we see a boxplot for age with the median at 28, the 25th percentile at 20, the 75th percentile at 38, maximum value at 80 and minimum value at 0.42 (referring to a baby of less than 1 year old).

This shows the wide margin of ages within the ship very clearly. There were all sorts of people of all ages, predominantly men, but definitely all ages.

Older people are a bit more scarce for many reasons, but we could easily attribute such disparity to the life expectancy at the time, which we mentioned previously sat at around 51-55 years old in England.

The relative fragile state of old people during these times, the fact that they might have been significantly more frail than nowadays our old family members are could have discouraged them. Or maybe the cruise trip was simply marketed for younger people and families. However, all these are merely hypotheses.

Boxplot for Fare

```
five <- data.frame(x=rep(1,5), five=fivenum(data_n$Fare))

ggplot(data=data_n, aes(x=0, y=data_n$Fare)) +
    geom_boxplot() +
    theme(text = element_text(size=18)) +
    ylab("Fare") +
    scale_x_discrete(breaks = NULL) +
    geom_text(data=five, aes(x=0, y=five, label=five),nudge_x =0.5, size=1.7)
```

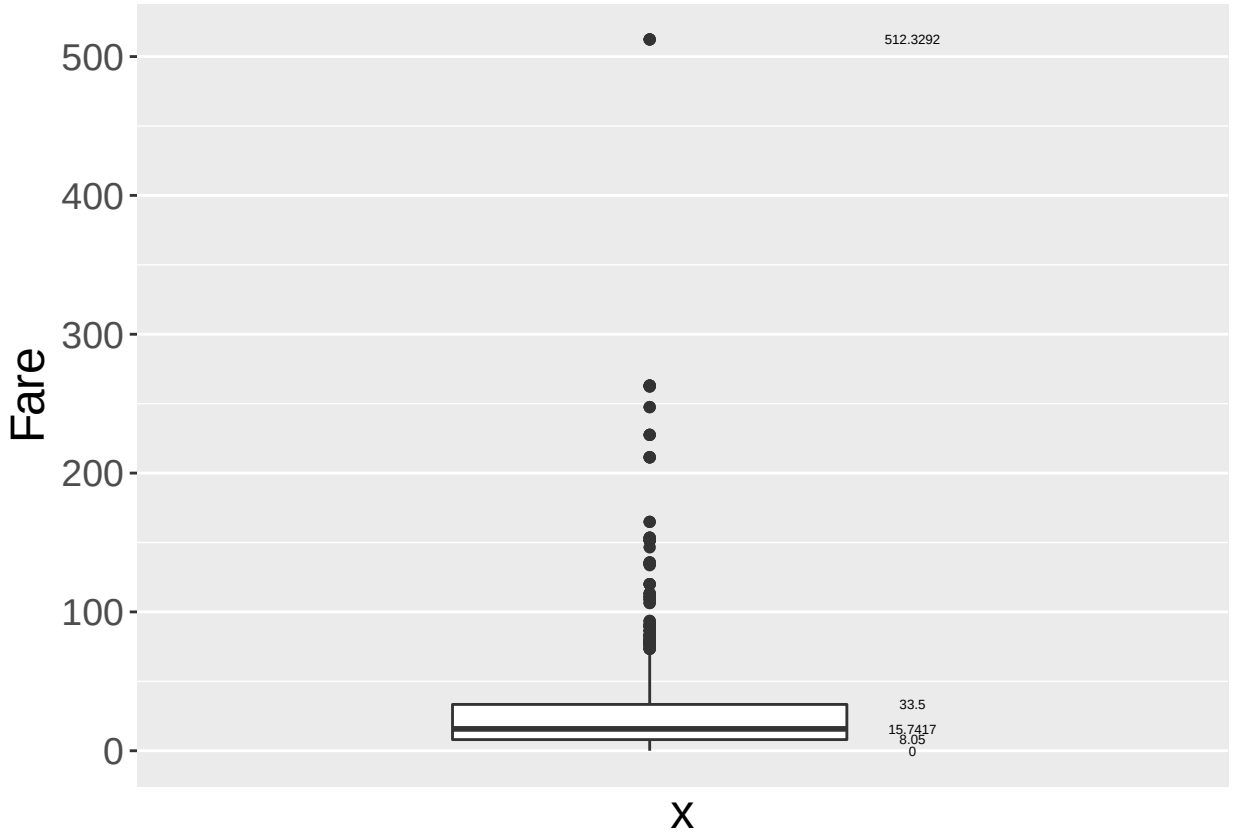


Figure 6

This boxplot for fare shows a really interesting pattern, similar to what the histogram painted. We can see the 25th and 75th percentile at 8.05 and 33.5 respectively, while the values in between represent most of the data and therefore most of the passengers.

All values above 100 are quite scarce, given that even at 100 GBP, the price was already significantly above the 30 GBP base fare for 1st class. These extreme values which peak at a very far maximum of about 512.33 (the maximum fare) represent a very small subset of very wealthy passengers above the ship.

Our minimum value is 0, which might represent children who probably didn't pay a fare, as maybe part of an offer for very very young children, which were, in fact, aboard the ship.

Quantiles for Age

```
plot(quantile(data_n$Age))
```

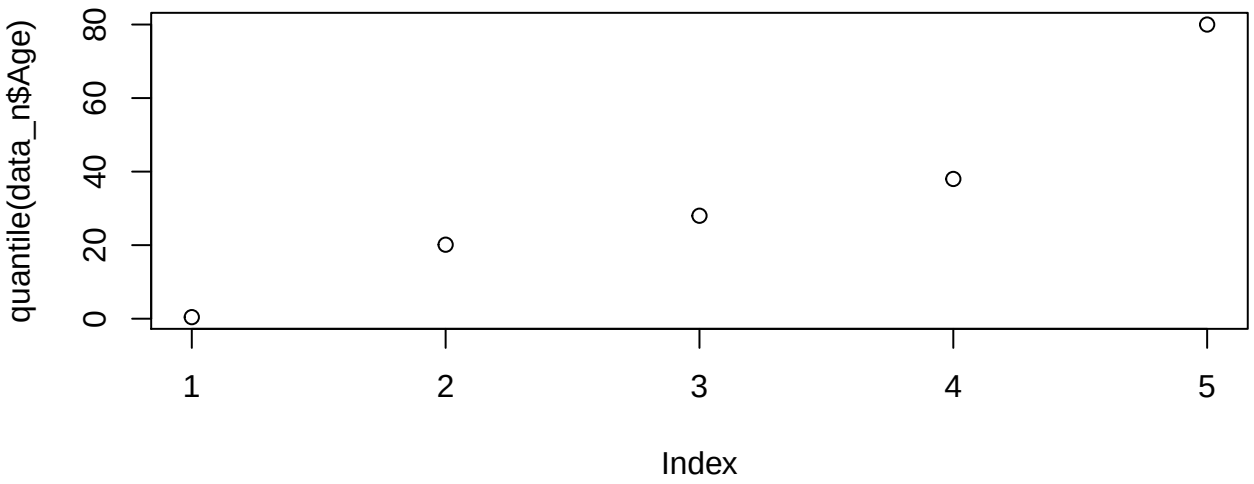


Figure 7

In the following plot we can see the quantiles more clearly for Age.

Quantiles for Fare

```
plot(quantile(data_n$Fare))
```

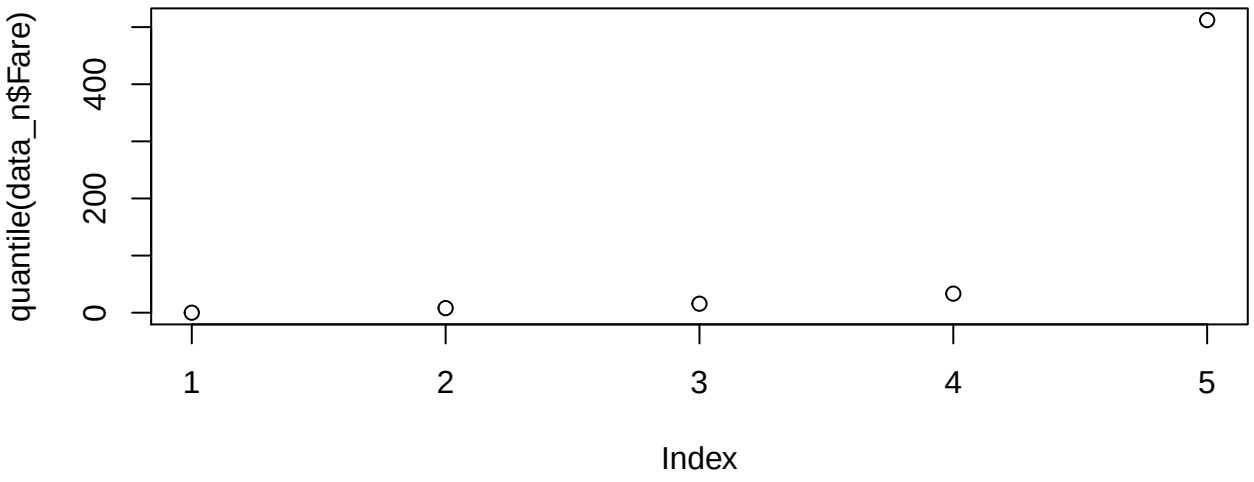


Figure 8

In the following plot we can see the quantiles more clearly for Fare

Both these plots clearly show the extreme nature of maximum values, especially for Fare.

Correlation between Age and Fare

```
log_data_n = data_n %>% mutate(logAge = log(Age), logFare = log(Fare+100))
ggplot(data=log_data_n, aes(x=logFare, y=logAge)) +
    geom_point() +
    theme(text = element_text(size=18))
```

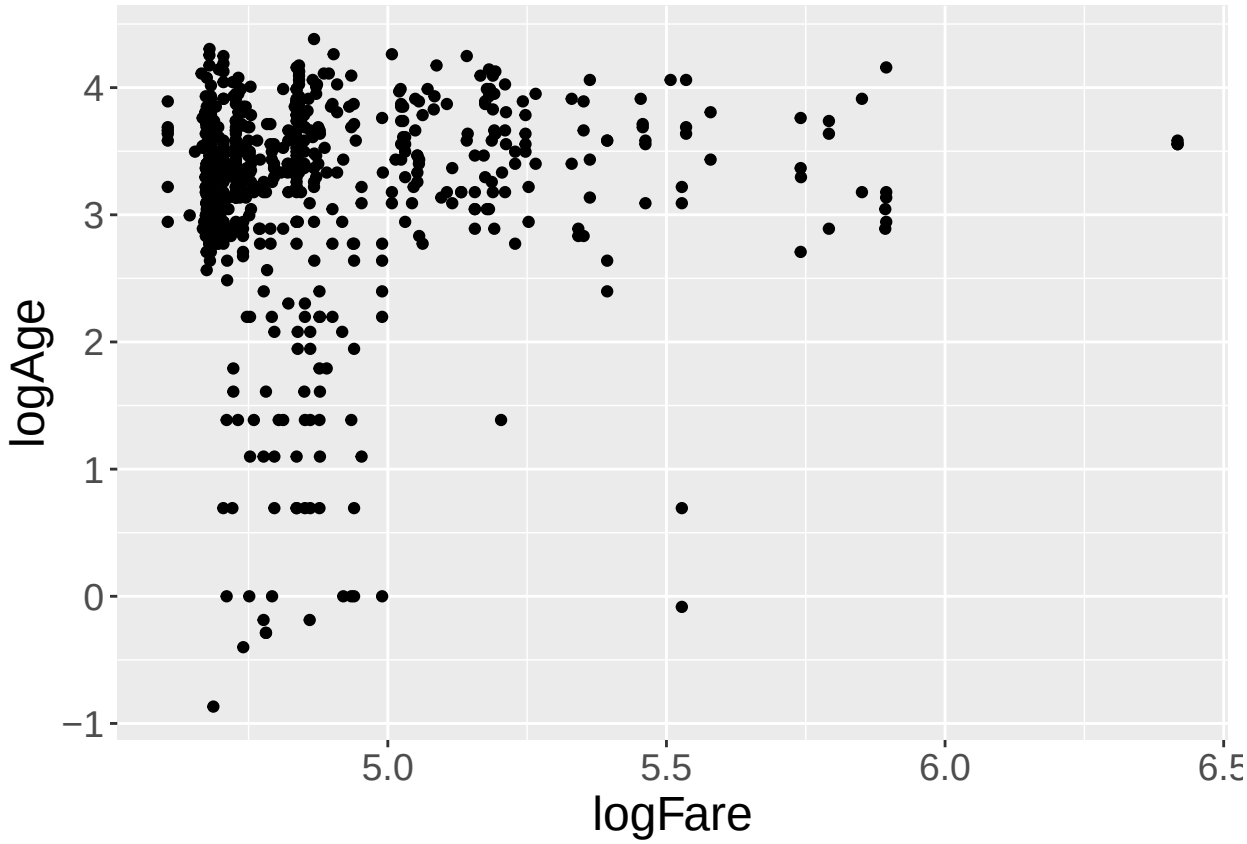


Figure 9

We decided to apply a log transformation for our plot in order to maybe try to make a plot that would better resemble one with a higher spearman correlation coefficient than what a pearson correlation coefficient would yield. However, such coefficient is not significantly larger, as the variables are not quite correlated.

The correlation is extremely low and even if we could draw a line to represent the shape both variables combined would take, we would really not show anything particularly relevant.

We decided to calculate the coefficient itself in order to be sure that it was as low as we expected.

```
cols <- data_n %>% dplyr::select(Age, Fare)
cor(cols, method="spearman")
```

```
##             Age      Fare
## Age  1.0000000 0.1350512
## Fare 0.1350512 1.0000000
```

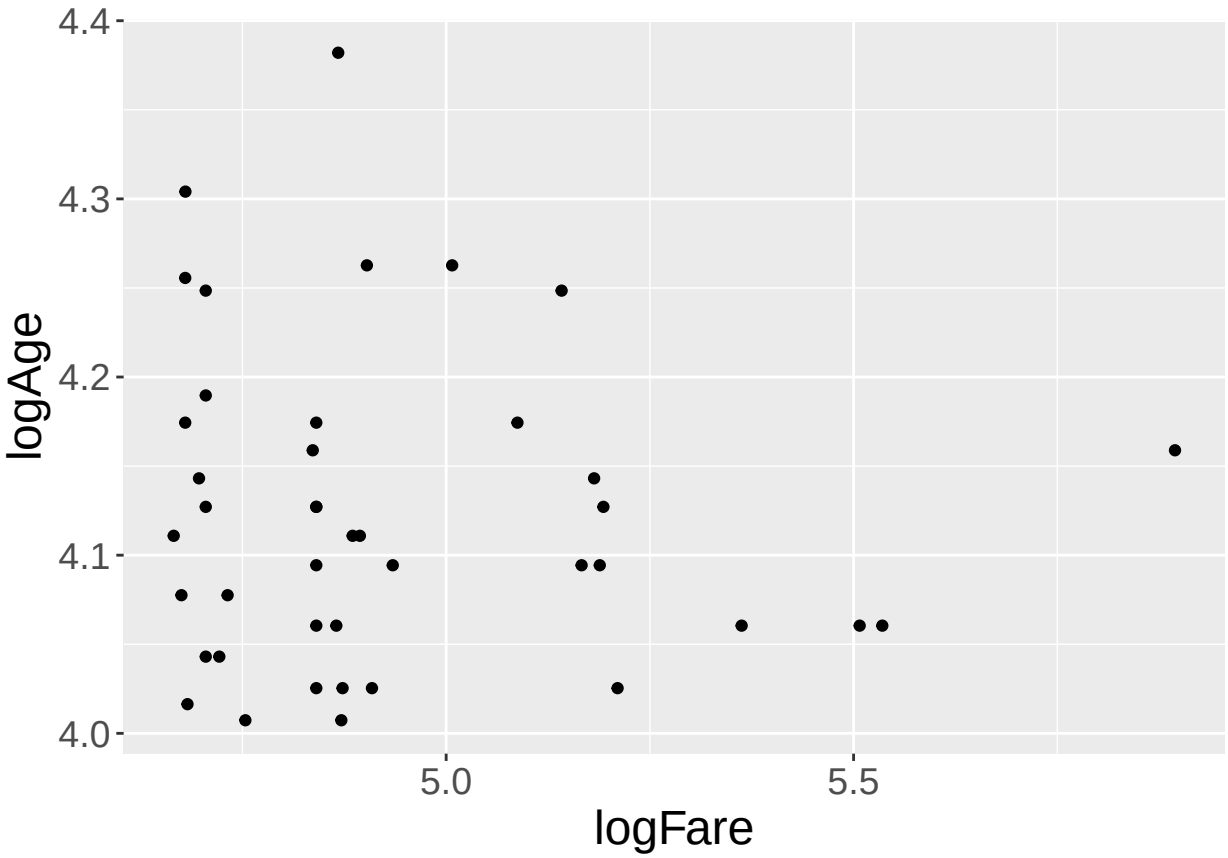
The result is about 0.135, which is, as expected, extremely low. We can faintly see the shape of somewhat of two typical spearman correlated variables. We can simply conclude that these two variables are not correlated, and thus age couldn't accurately predict fare nor the opposite.

What about older people?

Maybe if we segregated the age of the passengers before calculating the correlation, we could get a more interesting picture, let's try that out.

Correlation between Age and Fare for passengers older than 55 years of age

```
log_data_n = data_n %>%
              filter(Age >= 55) %>%
              mutate(logAge = log(Age), logFare = log(Fare+100))
ggplot(data=log_data_n, aes(x=logFare, y=logAge)) +
    geom_point() +
    theme(text = element_text(size=18))
```




```
cols <- log_data_n %>% dplyr::select(logAge, logFare)
cor(cols, method="spearman")
```

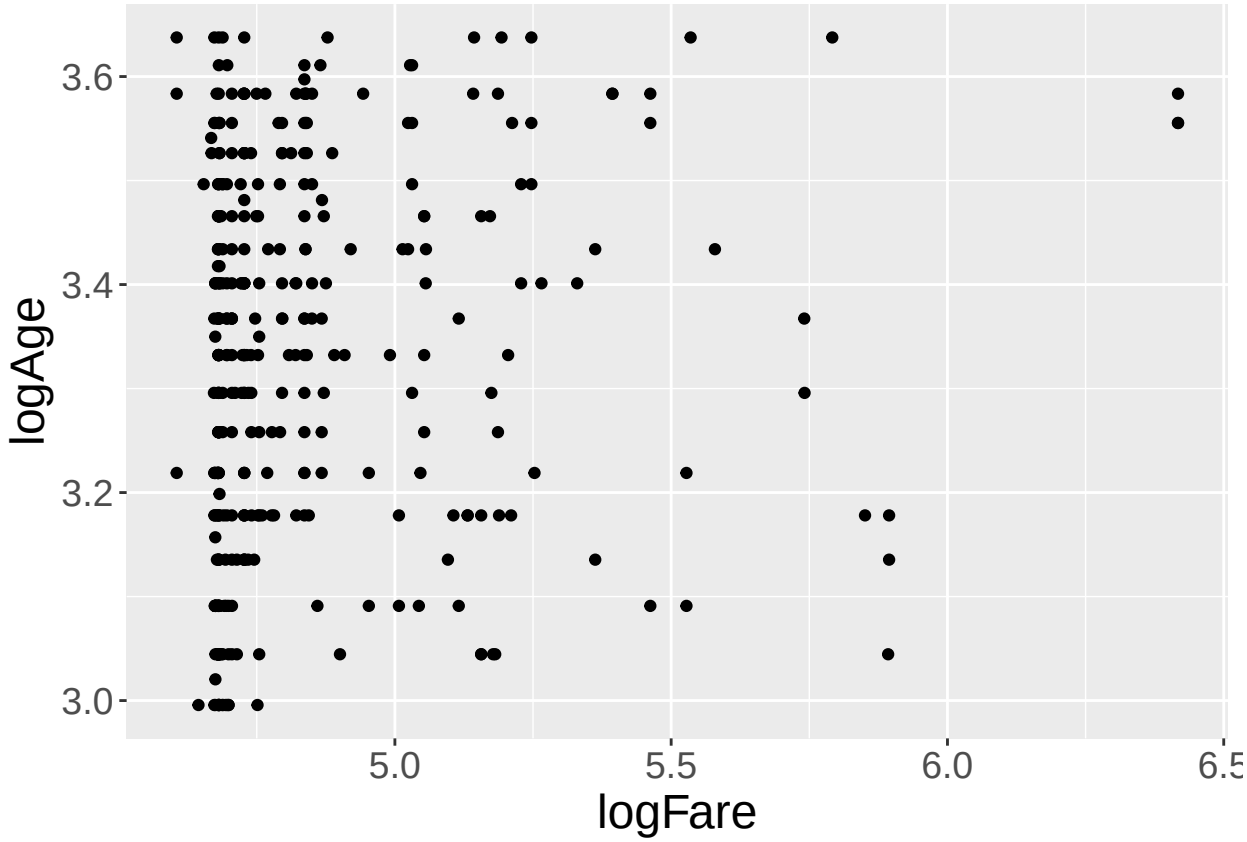
```
##             logAge    logFare
## logAge   1.0000000 -0.1050426
## logFare -0.1050426  1.0000000
```

We stick with our choice of using spearman, but this definitely did not work out for the older passengers.

What about ages between 25th and 75th percentiles?

Correlation between Age and Fare for passengers between 20 and 38 years of age

```
log_data_n = data_n %>%
              filter(Age >= 20 & Age <= 38 ) %>%
              mutate(logAge = log(Age), logFare = log(Fare+100))
ggplot(data=log_data_n, aes(x=logFare, y=logAge)) +
    geom_point() +
    theme(text = element_text(size=18))
```




```
cols <- log_data_n %>% dplyr::select(logAge, logFare)
cor(cols, method="spearman")
```

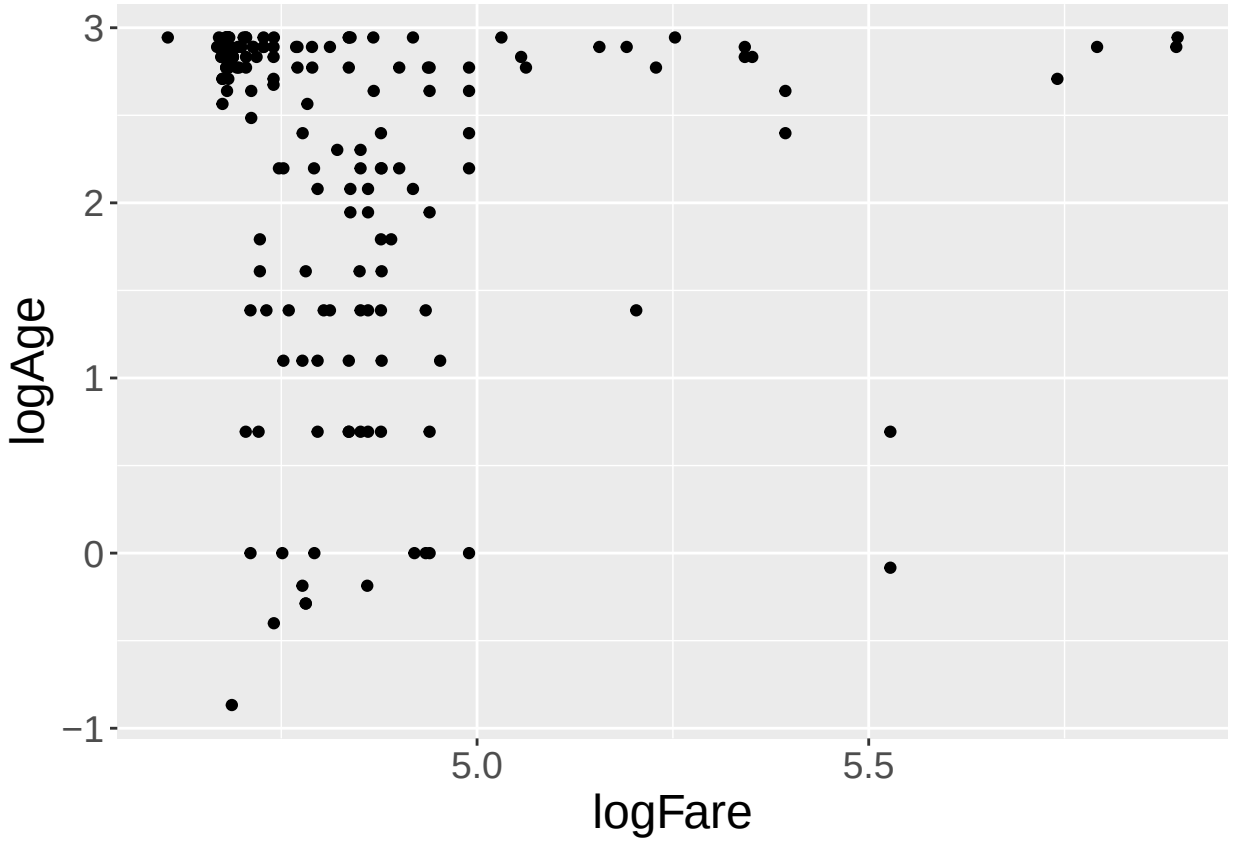
```
##            logAge   logFare
## logAge  1.0000000 0.2342904
## logFare 0.2342904 1.0000000
```

A tad bit better but still definitely not a stronger correlation.

What about ages below the 25th percentile?

Correlation between Age and Fare for passengers under 20 years of age

```
log_data_n = data_n %>%
              filter(Age < 20 ) %>%
              mutate(logAge = log(Age), logFare = log(Fare+100))
ggplot(data=log_data_n, aes(x=logFare, y=logAge)) +
    geom_point() +
    theme(text = element_text(size=18))
```




```
cols <- log_data_n %>% dplyr::select(logAge, logFare)
cor(cols, method="spearman")
```

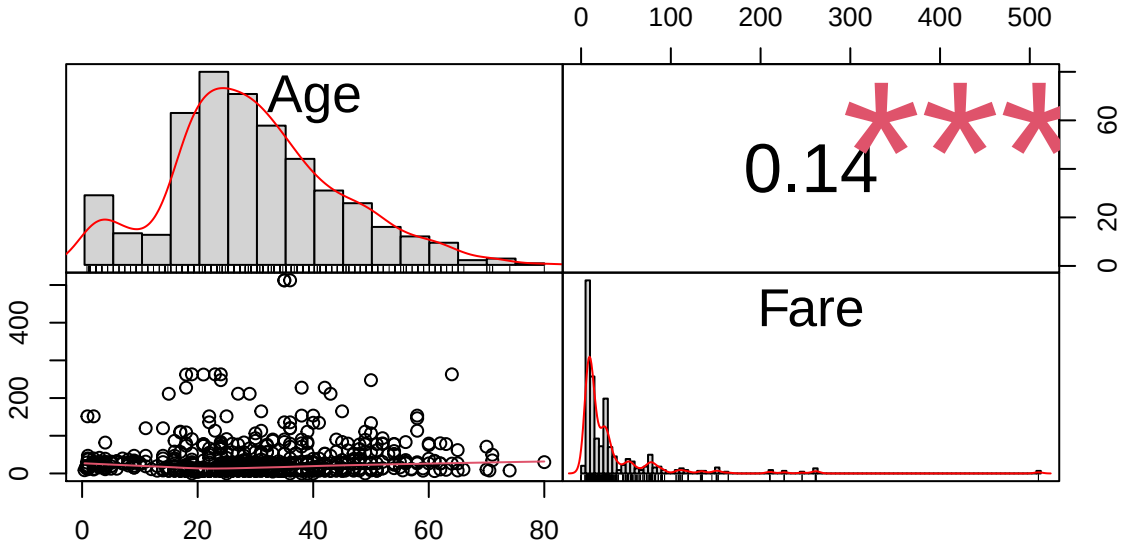
```
##             logAge    logFare
## logAge   1.0000000 -0.3443596
## logFare -0.3443596  1.0000000
```

Interesting!, so this is definitely stronger than our first example. We can say that most young people and kids are probably coming with their parents, therefore likely paying a cheaper ticket.

Performance Analytics plot

With a performance analytics plot we can see a histogram for each selected variable, a scatter plot with a plotted line through the points and the correlation number. Showing the power of custom R libraries!

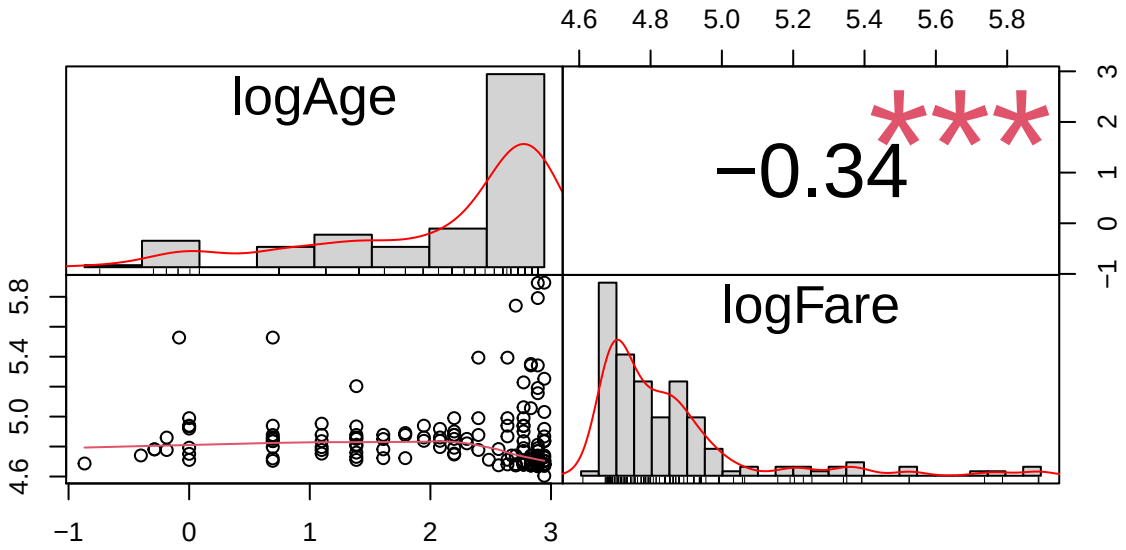
```
pa <- data_n %>% dplyr::select(Age, Fare)
chart.Correlation(pa, histogram=TRUE, pch=19, method="spearman")
```



Interesting!, so this is definitely stronger than our first example. We can say that most young people and kids are probably coming with their parents, therefore likely paying a cheaper ticket.

We can try using the dataset with the log transformations to view the scatter plot a bit less... messy

```
pa <- log_data_n %>% dplyr::select(logAge, logFare)
chart.Correlation(pa, histogram=TRUE, pch=19, method="spearman")
```



This didn't quite do much to the scatter plot but did give us a better correlation... suspicious.