EDA

Load Libraries

```
library(tidyverse)# data manipulation,create plots
library(DataExplorer) # correlation matrix
```

Load Data

```
df<-read.csv("houses_to_rent.csv")
```

```
head(df)
```

	X <int>	city <chr>	area <int>	rooms <int>	bathroom <int>	parking.spaces <int>	floor <chr>	animal <chr>	furniture <chr>
1	0	1	240	3	3	4	-	acept	furnished
2	1	0	64	2	1	1	10	acept	not furnished
3	2	1	443	5	5	4	3	acept	furnished
4	3	1	73	2	2	1	12	acept	not furnished
5	4	1	19	1	1	0	-	not acept	not furnished
6	5	1	13	1	1	0	2	acept	not furnished

6 rows | 1-10 of 15 columns

```
df<- df %>% select(-c(floor,animal,hoa,total,X,city,property.tax))
```

Type of Data

```
df %>% select(fire.insurance,rent.amount ) %>% summary()
```

```
##  fire.insurance     rent.amount     
##  Length:6080        Length:6080        
##  Class :character   Class :character   
##  Mode  :character   Mode  :character
```

We note that the price of fire insurance,floor and the price of rent. They are character type values, due to the fact that you have special characters, R standardizes and converts them to character type.

Convert variables to numeric

```
conv_to_numeric<-function(x){

  x<-gsub("[^\\d\\.]","", x)  # eliminate special charterers
  x<-as.numeric(x) # transform to numeric data

  return(x) # return data
}

attach(df)
```

```
fire_insurance<-sapply(fire.insurance,conv_to_numeric)
rent_amount<-sapply(rent.amount,conv_to_numeric)
```

```
df<-df %>%

  mutate(fire.insurance=fire_insurance,
         rent.amount=rent_amount )

# With the mutate function we make modifications to the data frame.
```

Do furnished houses have a higher rental price compared to those that are not?

```
df %>% group_by(furniture) %>% summarise(rent_amount_mean=mean(rent.amount))
```

furniture <chr>	rent_amount_mean <dbl>
furnished	5387.092
not furnished	4047.211

2 rows

The average price of furnished houses is higher than those that are not.

```
histogram=function(x,...){

  df %>%
    ggplot(aes(x=x,y=..density..)) +
    geom_histogram(color="black",fill="#FFFDCF") +
    geom_density(color="black",lwd=1) +
    geom_vline(aes(xintercept=mean(x),color="Mean")) +
    geom_vline(aes(xintercept=median(x),color="Median")) +
    labs(color="Statistics") +
    theme(legend.position = "top") +
    ...

}
```

```
rent_amount_histogram=histogram(df$rent.amount,labs(x="Rent Amount",title = "Rent Amount"))
area_histogram=histogram(df$area,labs(x="Area",title = "Area"))
fire_insurance_histogram=histogram(df$fire.insurance,labs(x="Fire Insurance",title = "Fire Insurance"))
rooms_histogram=histogram(df$rooms,labs(x="Rooms",title = "Rooms"))
bathroom_histogram=histogram(df$bathroom,labs(x="Bathroom",title = "Bathroom"))
parking_spces_histogram=histogram(df$parking.spaces,labs(x="Parking Spaces",title="Parking Spaces"))
```

```
scatter_plot<-function(x_feature,...){

  ggplot(data=df,aes(x=x_feature,y=rent.amount)) +
  geom_point(color="#7FDDFF",alpha=0.5) +
  theme_light() +
  geom_smooth(method = "lm",color="red") +
  ...
}
```

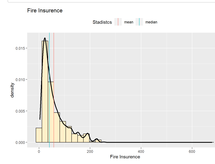
```
area_scatter<-scatter_plot(df$area,labs(x="Area",y="Rent Amount"))
fire_scatter<-scatter_plot(df$fire.insurance,labs(x="Fire Insurance",y="Rent Amount"))
rooms_scatter<-scatter_plot(df$rooms,labs(x="Rooms",y="Rent Amount"))
bathroom_scatter<-scatter_plot(df$bathroom,labs(x="Bathroom",y="Rent Amount"))
parking_scatter<-scatter_plot(df$parking.spaces,labs(x="Parking Spaces",y="Charges"))
```

```
library(gridExtra)
```

Histogram Plots

```
fire_insurance_histogram
```

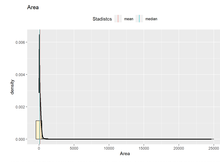
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



For the variable fire insurance there is a high amount of abnormal values. There are very few cases where they exceed R $250.

```
area_histogram
```

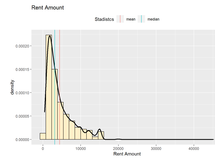
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There is an important bias for the variable area. At first glance, it can be seen that most of the departments have an area less than and equal to square meter.

```
rent_amount_histogram
```

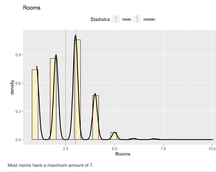
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



It can be considered that most of the rental houses are around a price below R $15,000.

```
rooms_histogram
```

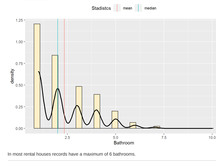
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Most rooms have a maximum amount of 7.

```
bathroom_histogram
```

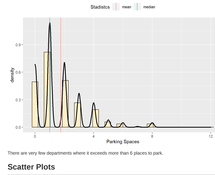
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In most rental houses records have a maximum of 6 bathrooms.

```
parking_spces_histogram
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

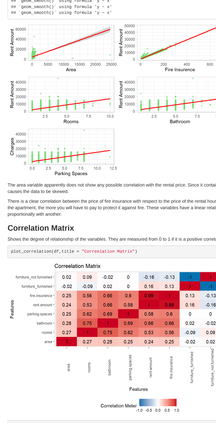


There are very few departments where it exceeds more than 6 places to park.

Scatter Plots

```
grid.arrange(area_scatter,
             fire_scatter,
             rooms_scatter,
             bathroom_scatter,parking_scatter)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



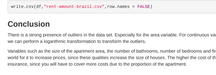
The area variable apparently does not show any possible correlation with the rental price. Since it contains a good number of outliers, which causes the data to be skewed.

There is a clear correlation between the price of fire insurance with respect to the price of the rental house, since the more expensive the price of the apartment, the more you will have to pay to protect it against fire. These variables have a linear relationship, that is, they increase proportionally, as much as

Correlation Matrix

Shows the degree of relationship of the variables. They are measured from 0 to 1 if it is a positive correlation, otherwise it is measured from 1 to -1.

```
plot_correlation(df,title = "Correlation Matrix")
```




```
### Save Dataset
```

```
write.csv(df,"rent-amount-brazil.csv",row.names = FALSE)
```

Conclusion

There is a strong presence of outliers in the data set. Especially for the area variable. For continuous variables, that is, those values with decimals, we can perform a logarithmic transformation to transform the outliers.

Variables such as the size of the apartment area, the number of bathrooms, number of bedrooms and fire insurance. It makes all the sense in the world for it to increase prices, since these qualities increase the size of houses. The higher the cost of the apartment, the higher the price of fire insurance, since you will have to cover more costs due to the proportion of the apartment.