EDA

Load Libraries

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

Load Data

```r
df<-read.csv("rent_amount.csv")
```

```r
head(df)
```

	city	area	rooms	bathroom	parking.spaces	floor	animal	furniture	rent.amount
	<chr>	<int>	<int>	<int>	<int>	<chr>	<chr>	<chr>	<chr>
1	yes	240	3	3	4	-	acept	furnished	R$8,000
2	no	64	2	1	1	10	acept	not furnished	R$820
3	yes	443	5	5	4	3	acept	furnished	R$7,000
4	yes	73	2	2	1	12	acept	not furnished	R$1,250
5	yes	19	1	1	0	-	not acept	not furnished	R$1,200
6	yes	13	1	1	0	2	acept	not furnished	R$2,200

6 rows | 1-10 of 11 columns

Type of Data

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

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

```r
conv_to_numeric<-function(x){


  x<-gsub("[//R$,]","",x)   # eliminate special charterers
  x<-as.numeric(x) # transform to numeric data

  return(x) # return data
}
```

```r
attach(df)
```

```
The following objects are masked from df (pos = 4):

    animal, area, bathroom, city, fire.insurance, floor, furniture, parking.spaces,
    rent.amount, rooms, total
```

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

```r
df<-df %>%

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

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

```r
unique(floor)
```

```
 [1] "-"  "10" "3"  "12" "2"  "16" "6"  "4"  "1"  "7"  "13" "9"  "14" "5"  "8"  "15" "11" "19" "20"
[20] "24" "23" "17" "18" "22" "27" "85" "28" "25" "29" "35" "21" "31" "99" "26" "68" "32" "51"
```

We note that for unique values of the floor variable an underscore appears. Which indicates that the rental house is independent, that is, it is not established in a building.

```r
floor_modify<-ifelse(floor=="-",0,floor)
# Where the hyphen appears, we transform it to 0 for the reasons explained above.
floor_modify<-as.numeric(floor_modify) # transform to numeric data


df<- df %>% mutate(floor=floor_modify)
```

```r
df<- df %>% select(-total)
```

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

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

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

2 rows

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

Accepting animals increases the price?

```r
df %>% group_by(animal) %>% summarise(animal_mean=mean(rent.amount))
```

animal	animal_mean
<chr>	<dbl>
acept	4585.849
not acept	3768.856

2 rows

The average price of apartments where animals are accepted is higher. Since they are allowing you the luxury of putting your pets in the apartment.

```r
 histogram<-function(x,...){

   df %>%
     ggplot(aes(x=x,y=..density..)) +
     geom_histogram(color="black",fill="#FFF0C9") +
     geom_density(color="black",lwd=1) +
     geom_vline(aes(xintercept=mean(x),color="mean")) +
     geom_vline(aes(xintercept=median(x),color="median")) +
     labs(col="Stadistcs") +
     theme(legend.position = "top") +
     ...


 }
```

```r
rent_amount_histogram=histogram(df$rent.amount,labs(x="Rent Amount",title = "Rent Amount"))
area_hsitogram=histogram(df$area,labs(x="Area",title = "Area"))
fire_insurence_histogram=histogram(df$fire.insurance,labs(x="Fire Insurence",title = "Fire Insurence"))
rooms_histogram=histogram(df$rooms,labs(x="Rooms",title = "Rooms"))
bathroom_histogram=histogram(df$bathroom,labs(x="Bathroom",title = "Bathroom"))
floor_histogram=histogram(df$floor,labs(x="Floor",y="Rent Amount",title = "Floor"))
```

```r
scattter_plot<-function(x_feature,...){

  ggplot(data=df,aes(x=x_feature,y=rent.amount)) +
  geom_point(color="#77dd77",alpha=0.5) +
  theme_light() +
  geom_smooth(method = "lm",color="red") +
  ...
}
```

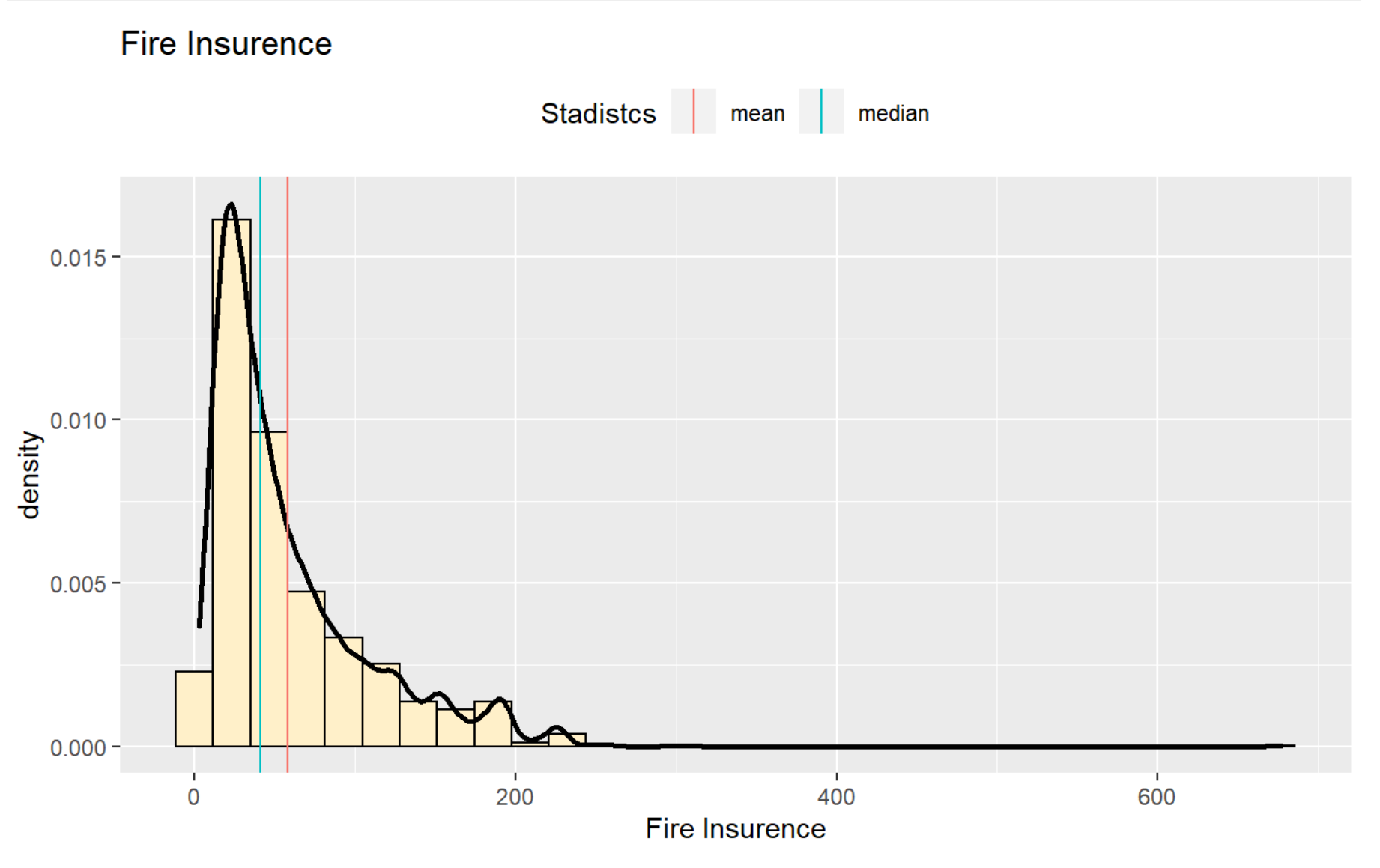
```r
area_scatter<-scattter_plot(df$area,labs(x="Area",y="Rent Amount"))
fire_scatter<-scattter_plot(df$fire.insurance,labs(x="Fire Insurence",y="Rent Amount"))
rooms_scatter<-scattter_plot(df$rooms,labs(x="Rooms",y="Rent Amount"))
bathroom_scatter<-scattter_plot(df$bathroom,labs(x="Bathroom",y="Rent Amount"))
floor_scatter<-scattter_plot(df$floor,labs(x="Floor",y="Rent Amount"))
```

```r
library(gridExtra)
```

Histogram Plots

```r
fire_insurence_histogram
```

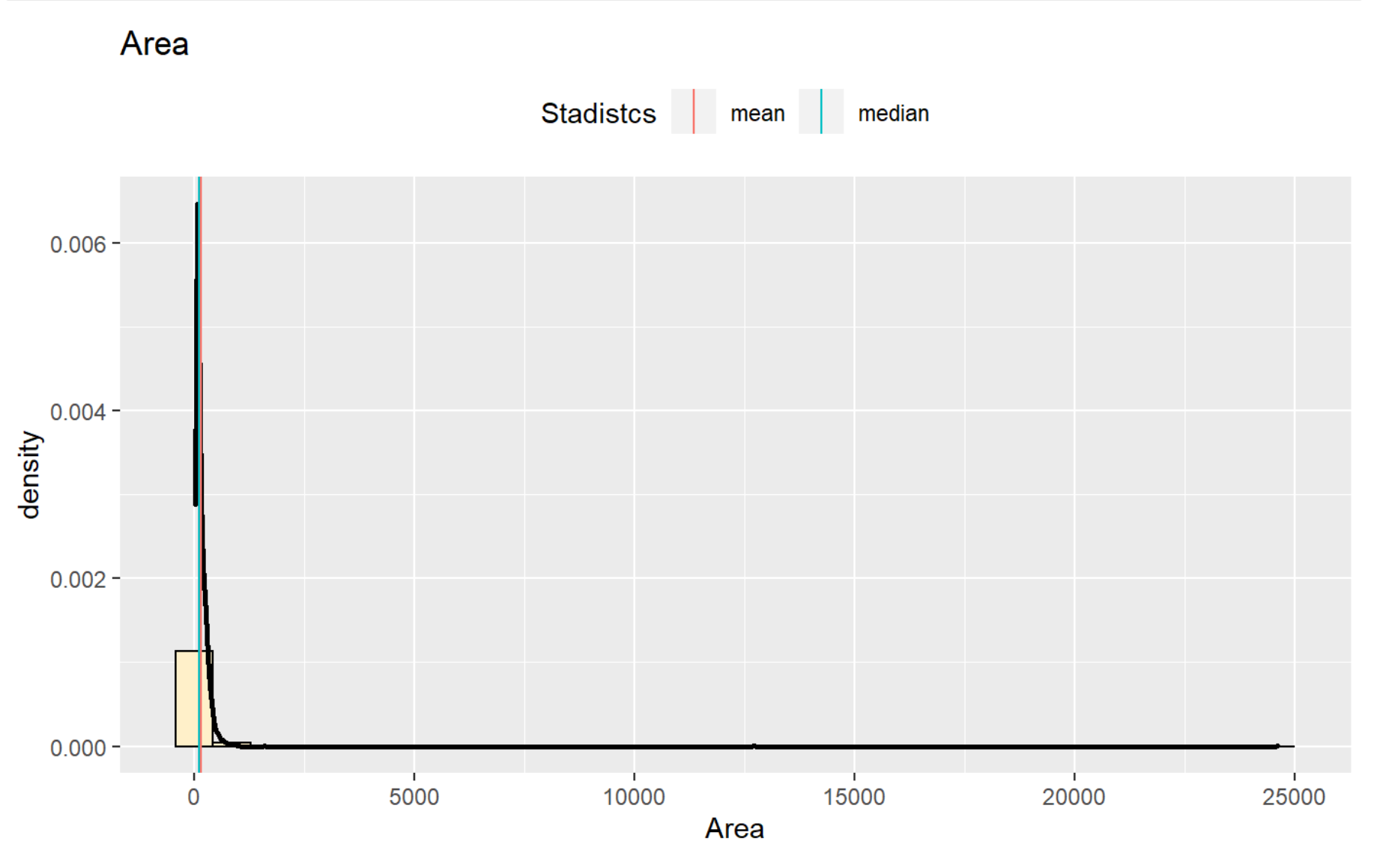
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```r
area_hsitogram
```

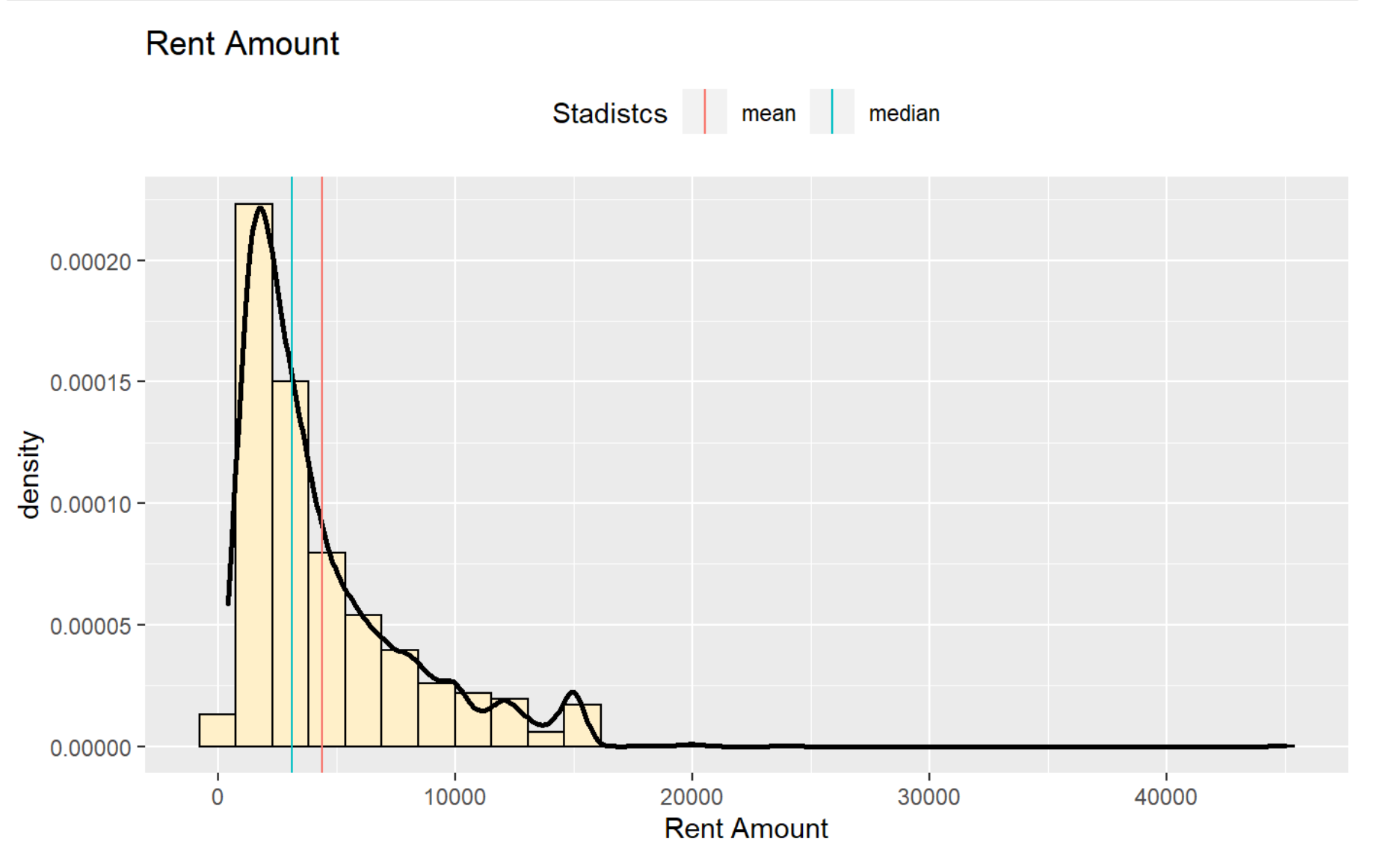
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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.

```r
rent_amount_histogram
```

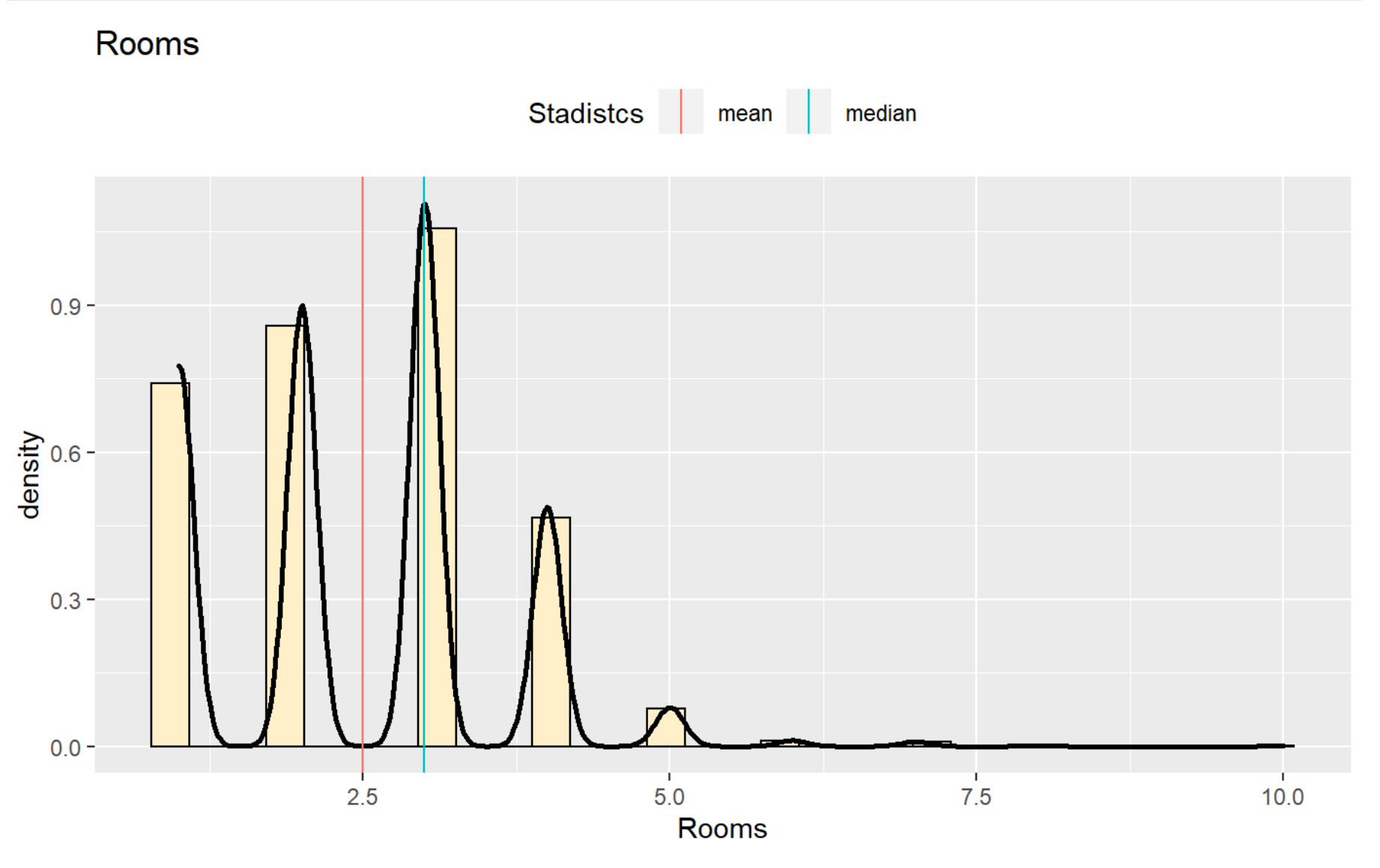
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```r
rooms_histogram
```

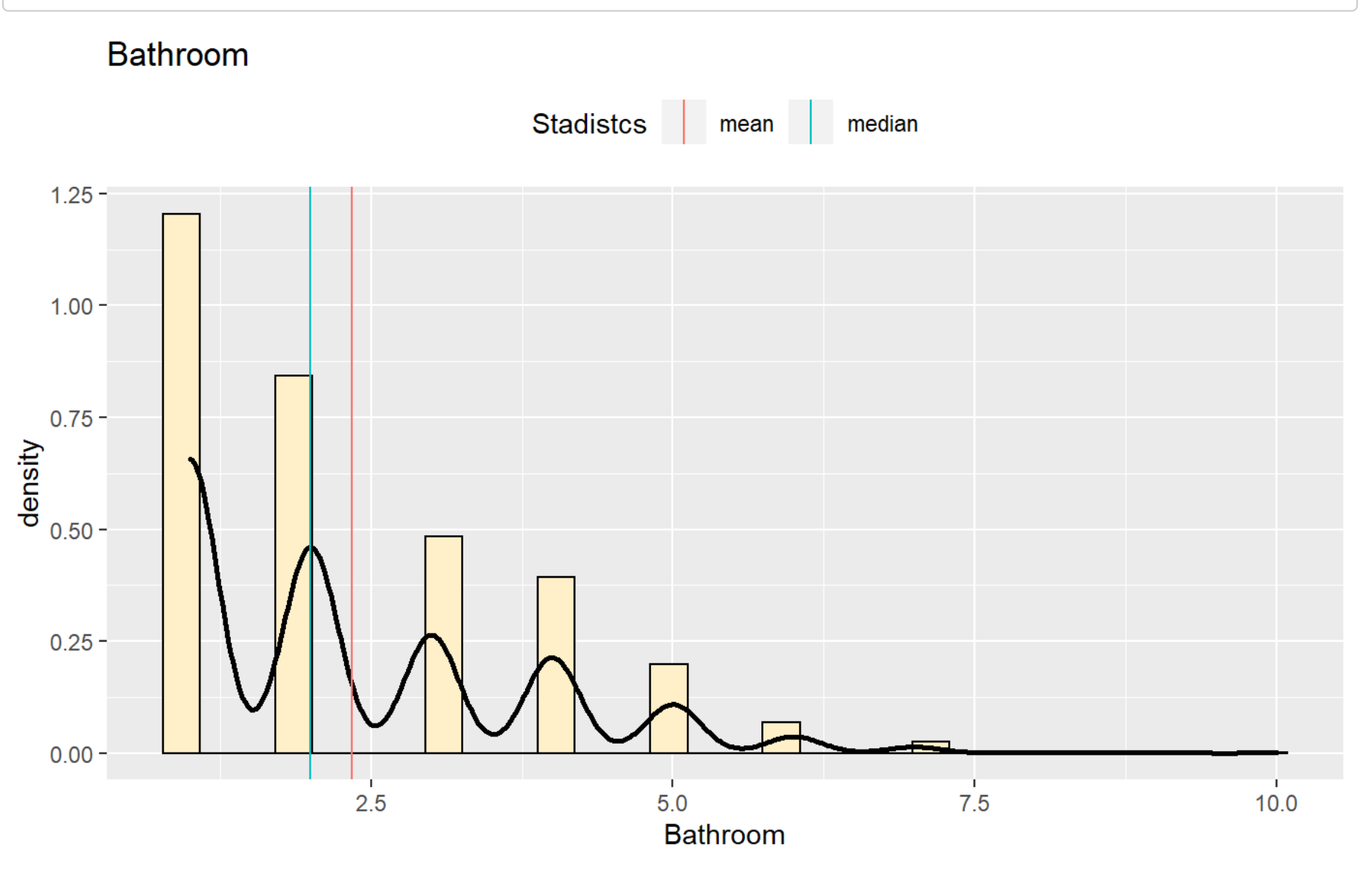
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Most rooms have a maximum amount of 7.

```r
bathroom_histogram
```

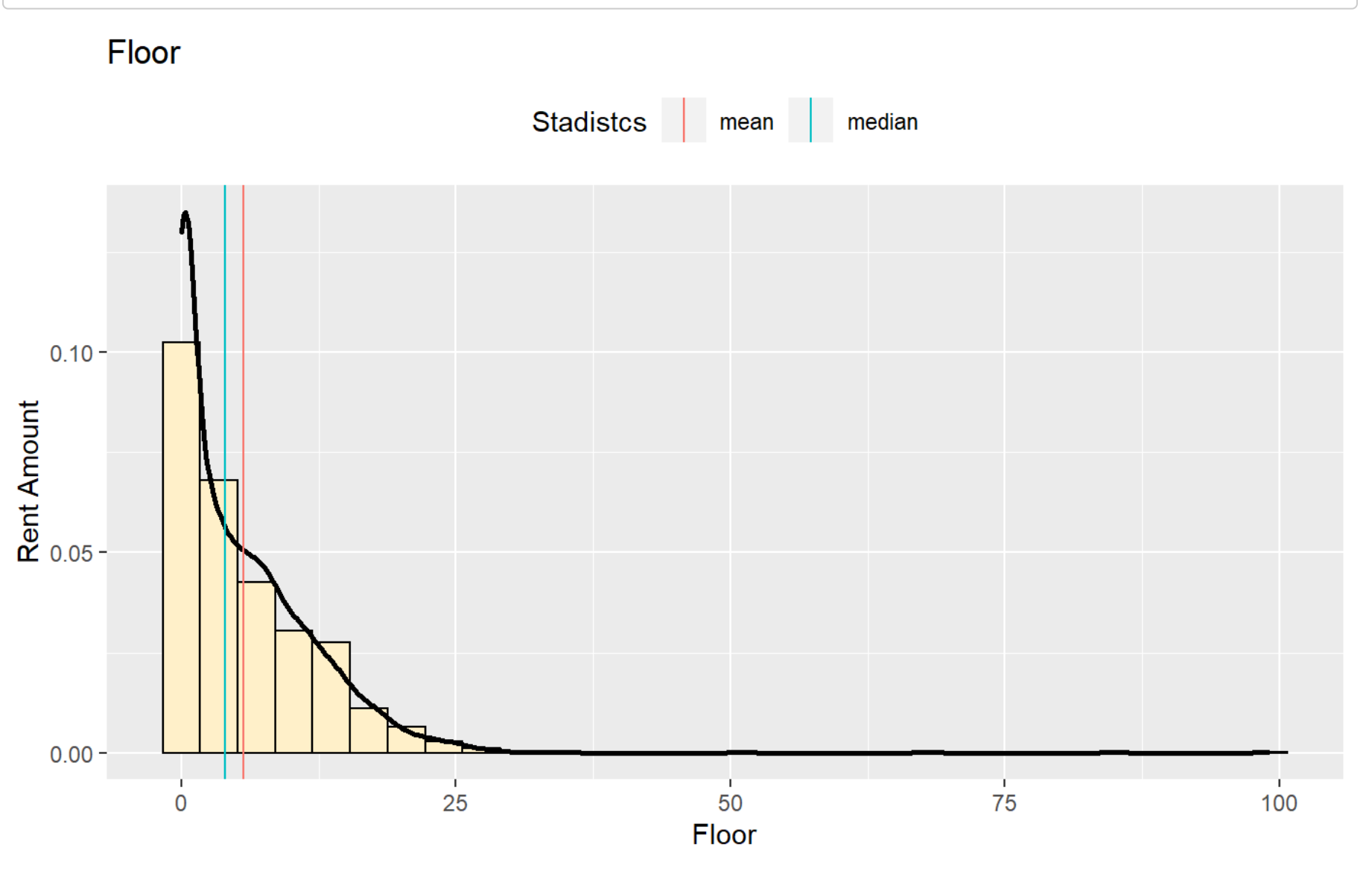
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```r
floor_histogram
```

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



Most of the buildings where the apartments are rented are less than 25 stories.

Scatter Plots

```r
grid.arrange(area_scatter,
             fire_scatter,
             rooms_scatter,
             bathroom_scatter,
             floor_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.

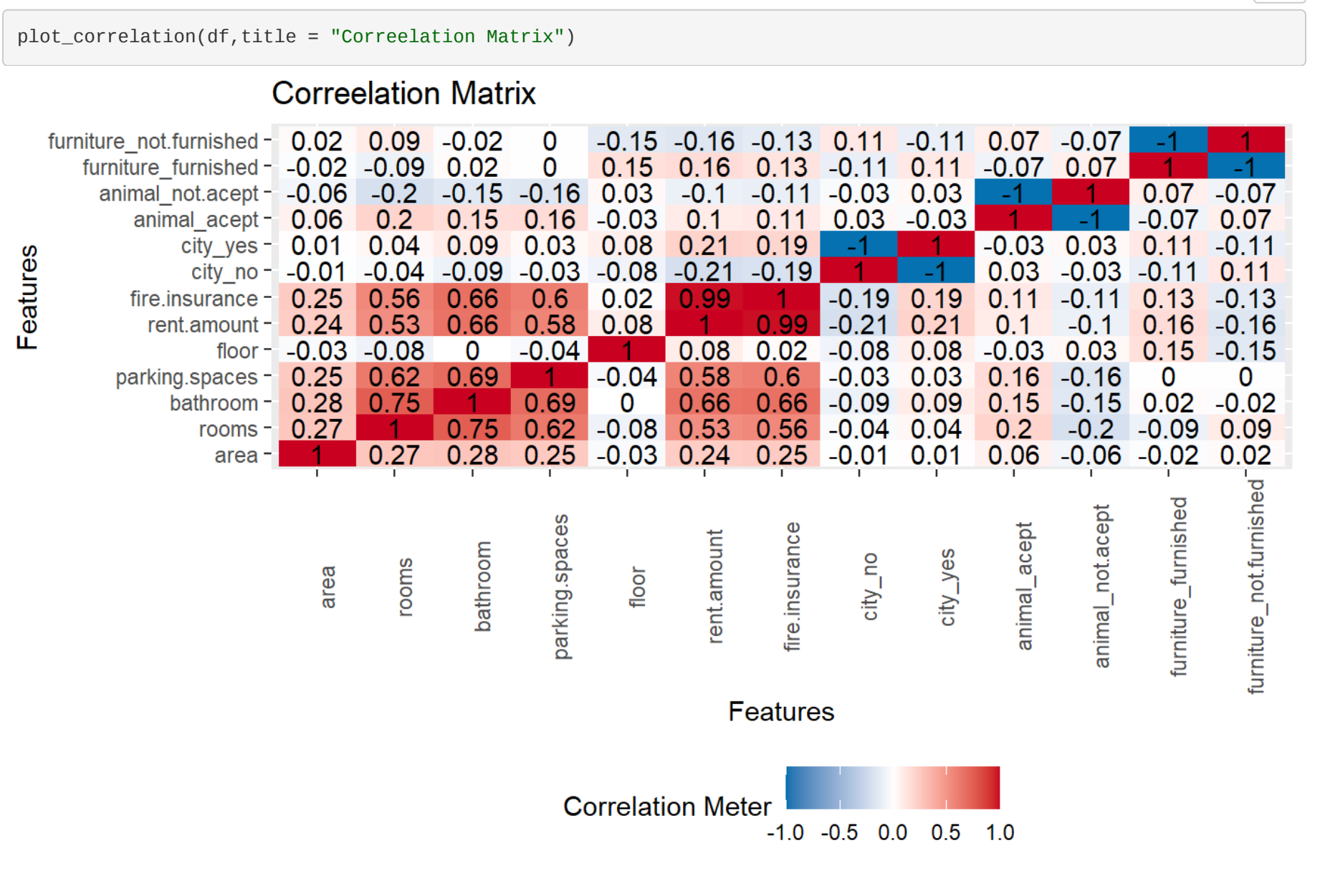
There is a clear correlation between the price of fire insurance with respect to the price of the rental house, since the more expensive the price of the apartment, the more you will have to pay to protect it against fire. These variables have a linear relationship, that is, they increase proportionally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

The variable floor, rooms and bathrooms have to raise the price of rent rent something that is logical.

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.

```r
plot_correlation(df,title = "Correelation Matrix")
```



Save Dataset

```r
write.csv(df,"rent-amount.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.