

QMB Exercise 1 - Exploring Housing Rents

Dino Nienhold

19 Feb, 2015

Introduction

The following report is based on the QMB Exercise 1 - Exploring Housing Rents. The task description pdf file is bis_ex1-HousingRents.pdf

Requirements

Please make sure that you the following packages loaded in your workspace.

```
library("dplyr")
```

```
##  
## Attaching package: 'dplyr'  
##  
## The following object is masked from 'package:stats':  
##  
##     filter  
##  
## The following objects are masked from 'package:base':  
##  
##     intersect, setdiff, setequal, union
```

```
library("ggplot2")
```

Data Set

Please make sure you have the file housingrents.csv in the subdirectoy Data in your workspace.

```
housingrents <- read.csv("../Data/housingrents.csv", sep=";")
```

Task 1

There are 152 observations and 7 variables in the dataset. The str command gives an overview of the variable types:

```
str(housingrents)
```

```
## 'data.frame': 152 obs. of 7 variables:
## $ id : int 1 2 3 4 6 7 8 10 11 13 ...
## $ rooms : int 1 1 1 1 1 1 1 1 1 1 ...
## $ area : int 34 35 50 45 35 40 43 45 37 60 ...
## $ rent : int 310 749 281 483 515 530 480 560 580 510 ...
## $ nre : int 0 0 0 0 0 0 0 0 0 0 ...
## $ econage: int 24 40 34 30 27 31 30 28 50 42 ...
## $ balcony: Factor w/ 2 levels "no","yes": 2 1 2 1 1 NA 1 1 1 1 ...
```

There are 14 NA values in the balcony variable.

```
summary(housingrents)
```

```
##      id      rooms      area      rent
## Min.   : 1.00   Min.   :1.000   Min.   : 18.00   Min.   : 250.0
## 1st Qu.: 38.75   1st Qu.:2.000   1st Qu.: 60.00   1st Qu.: 793.8
## Median : 76.50   Median :3.000   Median : 83.00   Median :1046.0
## Mean   : 76.50   Mean   :3.171   Mean   : 86.84   Mean   :1240.3
## 3rd Qu.:114.25   3rd Qu.:4.000   3rd Qu.:105.00   3rd Qu.:1552.8
## Max.   :152.00   Max.   :6.000   Max.   :250.00   Max.   :4725.0
##      nre      econage      balcony
## Min.   :0.0000   Min.   : 0.00   no  :61
## 1st Qu.:0.0000   1st Qu.:22.00   yes :77
## Median :0.0000   Median :31.00   NA's:14
## Mean   :0.3355   Mean   :30.18
## 3rd Qu.:1.0000   3rd Qu.:39.00
## Max.   :1.0000   Max.   :60.00
```

Task 2

Data Processing

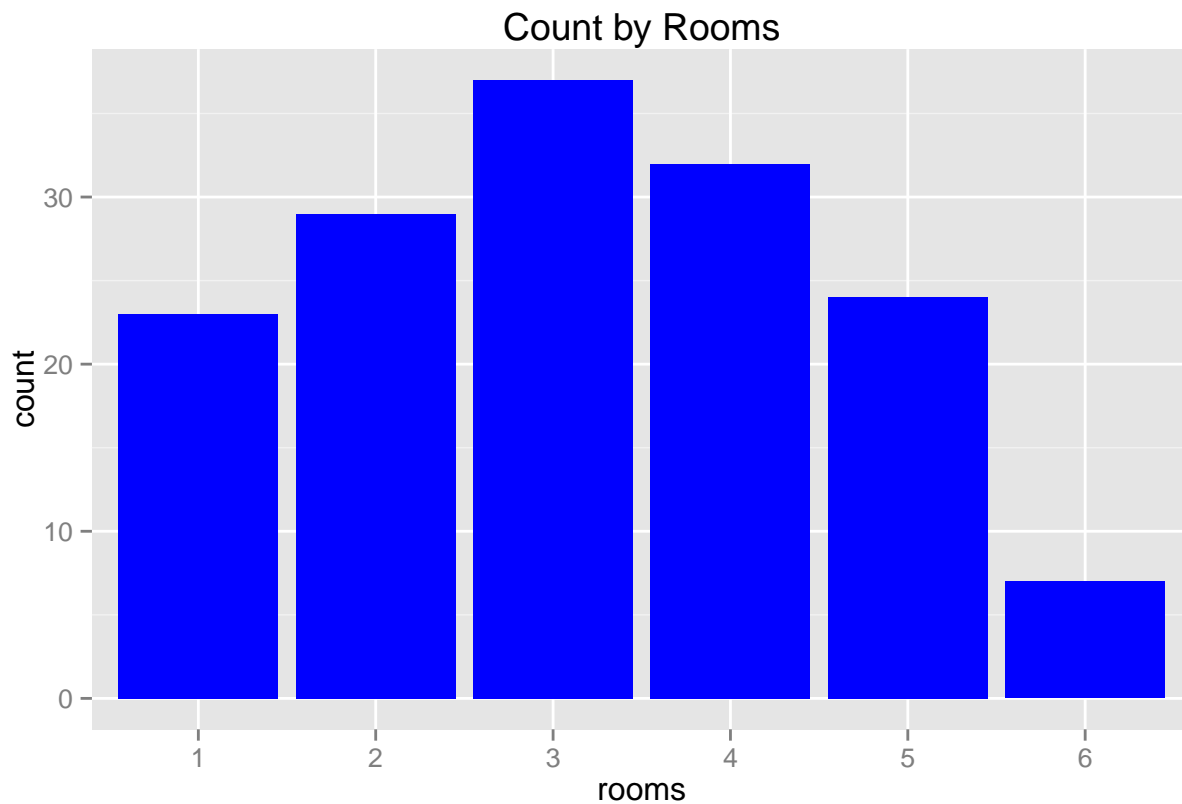
For analysis purposes it is necessary to convert the rooms and NRE variable to a factor.

```
housingrents <- mutate(housingrents, rooms = factor(rooms),  
  nre = factor(nre, levels=c(0,1), labels=c("no", "yes")))
```

Plotting

The following chart shows the frequency of appartments according to their numberof rooms.

```
ggplot(data=housingrents, aes(x=rooms, label=rooms)) +  
  geom_bar(fill="blue") +  
  ggtitle("Count by Rooms")
```



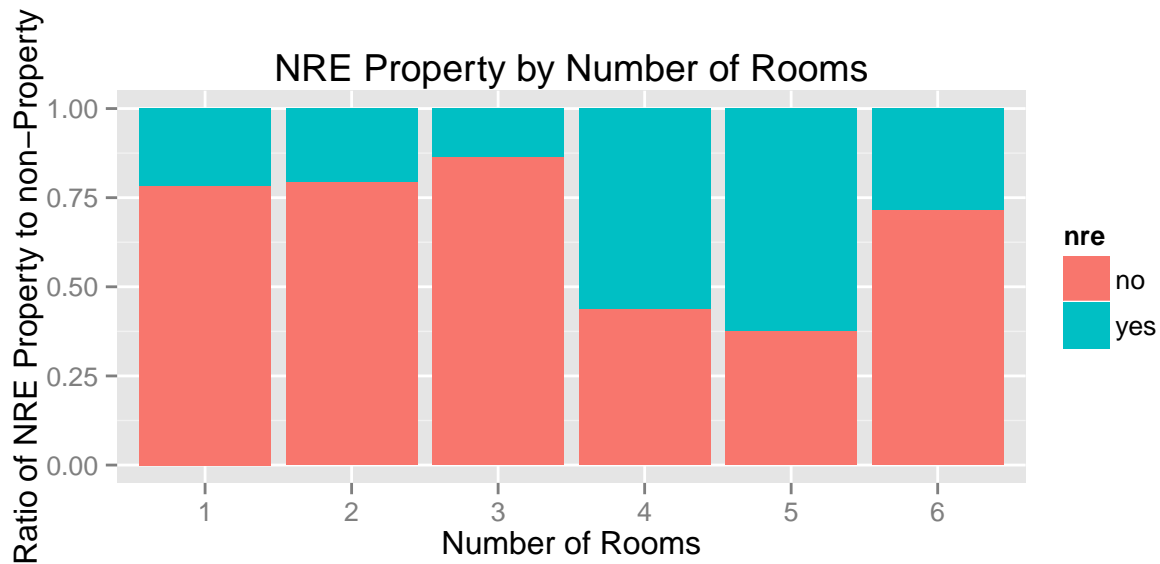
Conclusion

In the dataset 3 room apartments are the most frequent one. Compared to the other appartments there are only a few 6 room apartments.

Task 3

In this section the contingency table for rooms and NRE are calculated and plotted.

```
rooms2nre <- xtabs(~rooms+nre, data=housingrents)
rooms2nre <- prop.table(rooms2nre,1)
ggplot(data.frame(rooms2nre), aes(x=rooms, y=Freq, fill=nre)) +
  geom_bar(stat="identity") +
  xlab("Number of Rooms") +
  ylab("Ratio of NRE Property to non-Property") +
  ggtitle("NRE Property by Number of Rooms")
```



Contingency Table with row percentages

```
addmargins(prop.table(rooms2nre,1))
```

```
##      nre
## rooms   no    yes   Sum
##  1  0.7826087 0.2173913 1.0000000
##  2  0.7931034 0.2068966 1.0000000
##  3  0.8648649 0.1351351 1.0000000
##  4  0.4375000 0.5625000 1.0000000
##  5  0.3750000 0.6250000 1.0000000
##  6  0.7142857 0.2857143 1.0000000
## Sum 3.9673627 2.0326373 6.0000000
```

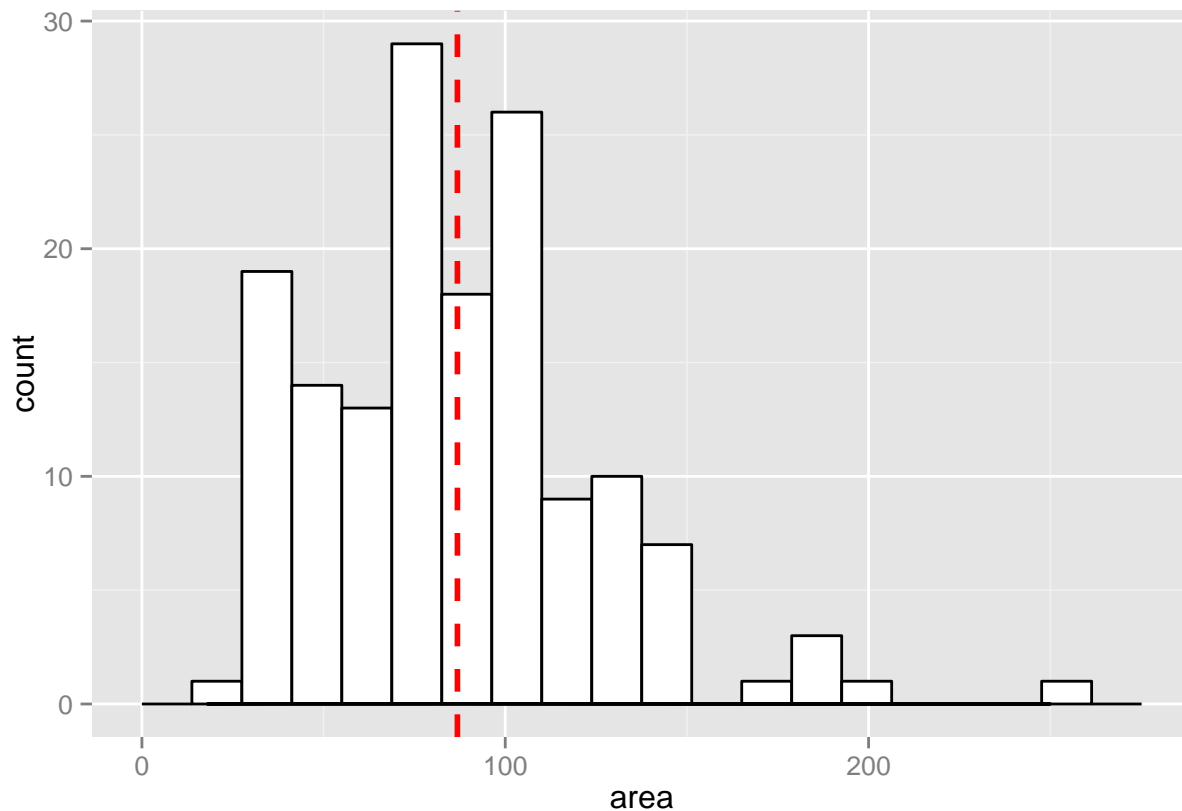
Conclusion

The apartments owned by NRE are mostly 4 or 5 room apartments.

Task 4

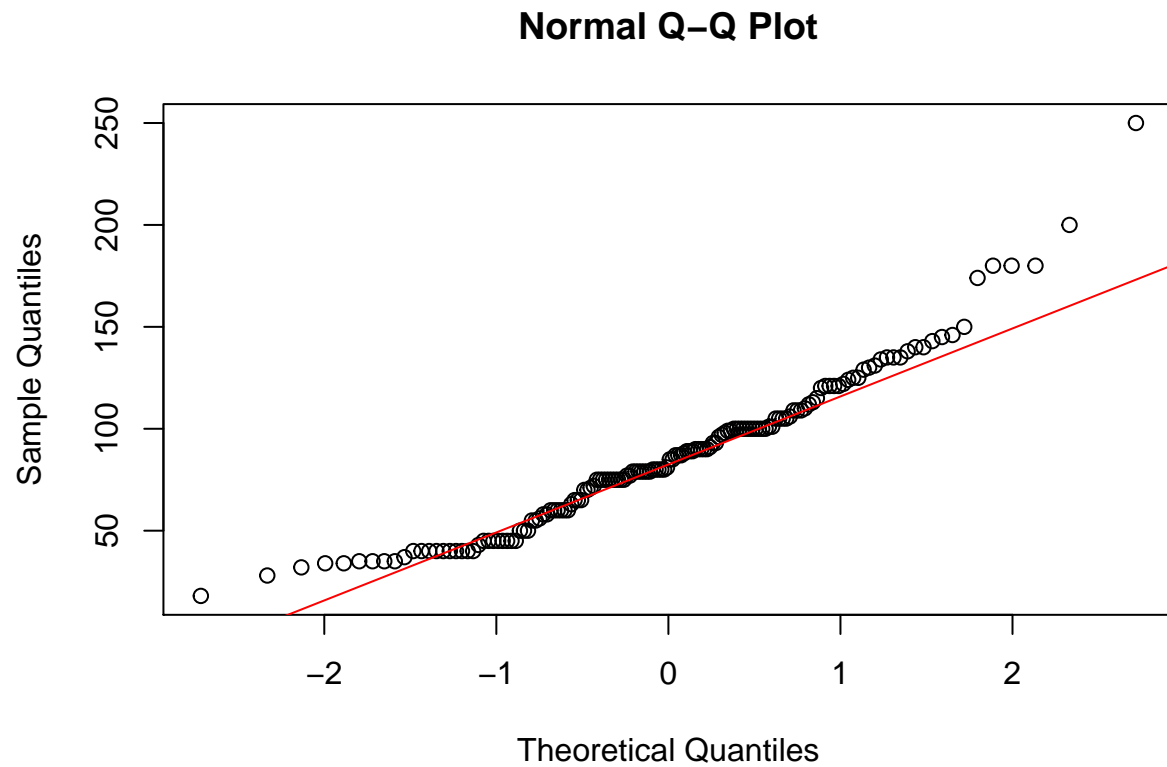
In this section the distribution of the variable area is analyzed to see if it is normally distributed.

```
#Calculate binwidth based on the Freedman-Diaconis rule
bw <- diff(range(housingrents$area)) / (2 * IQR(housingrents$area) /
  length(housingrents$area)^(1/3))
#Plot the histogram
ggplot(housingrents, aes(x = area)) +
  geom_histogram(binwidth = bw, colour="black", fill="white") +
  geom_density(alpha=.2) +
  geom_density(alpha=.5, fill="#FF6666") +
  geom_vline(aes(xintercept=mean(area, na.rm=T)),
    color="red", linetype="dashed", size=1)
```



The qqplot should help to get a better understanding whether the distribution of the variable is normally distributed.

```
qqnorm(housingrents$area)
qqline(housingrents$area, col = 2)
```



Conclusion

The area variable is not normally distributed.

Task 5

In this section the mean/median of the rent of NRE and non-NRE apartments are compared

```
ggplot(housingrents,aes(y=rent,x=nre, fill=nre)) +  
  geom_boxplot() +  
  xlab("Property owned by NRE") +  
  ylab("Amount of Rent") +  
  ggtitle("Rent Comparison of NRE and non-NRE apartments") +  
  stat_summary(fun.y=mean, geom="point", shape=5, size=4)
```



The following table shows the mean and median rent by nre and non-nre apartments:

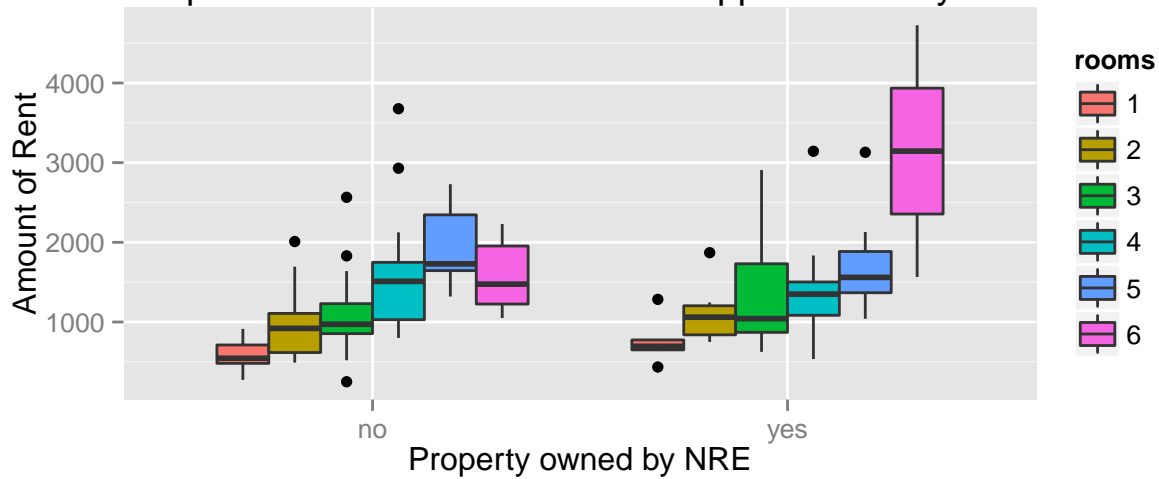
```
select(housingrents,rent,nre) %>%  
  group_by(nre) %>%  
  summarise(mean = mean(rent), median = median(rent))
```

```
## Source: local data frame [2 x 3]  
##  
##   nre    mean median  
## 1  no 1136.584   974  
## 2  yes 1445.745  1325
```

Both the plot and table show that NRE owned apartments have higher mean and median. The analysis is extended to see if the difference persists across all apartment categories.

```
ggplot(housingrents,aes(y=rent,x=nre, fill=rooms)) +  
  geom_boxplot() +  
  xlab("Property owned by NRE") +  
  ylab("Amount of Rent") +  
  ggtitle("Rent Comparison of NRE and non-NRE apartments by # of rooms")
```

Rent Comparison of NRE and non-NRE apartments by # of rooms



```
select(housingrents,rent,nre,rooms) %>%
  group_by(nre,rooms) %>%
  summarise(mean = mean(rent), median = median(rent))
```

```
## Source: local data frame [12 x 4]
## Groups: nre
##
##   nre rooms    mean median
## 1   no     1  581.2778   545
## 2   no     2  980.3043   920
## 3   no     3 1056.7188   972
## 4   no     4 1625.1429  1510
## 5   no     5 1920.4444  1730
## 6   no     6 1586.8000  1475
## 7  yes     1  767.8000   695
## 8  yes     2 1126.0000  1060
## 9  yes     3 1435.0000  1042
##10  yes     4 1360.5000  1350
##11  yes     5 1678.9333  1560
##12  yes     6 3145.0000  3145
```

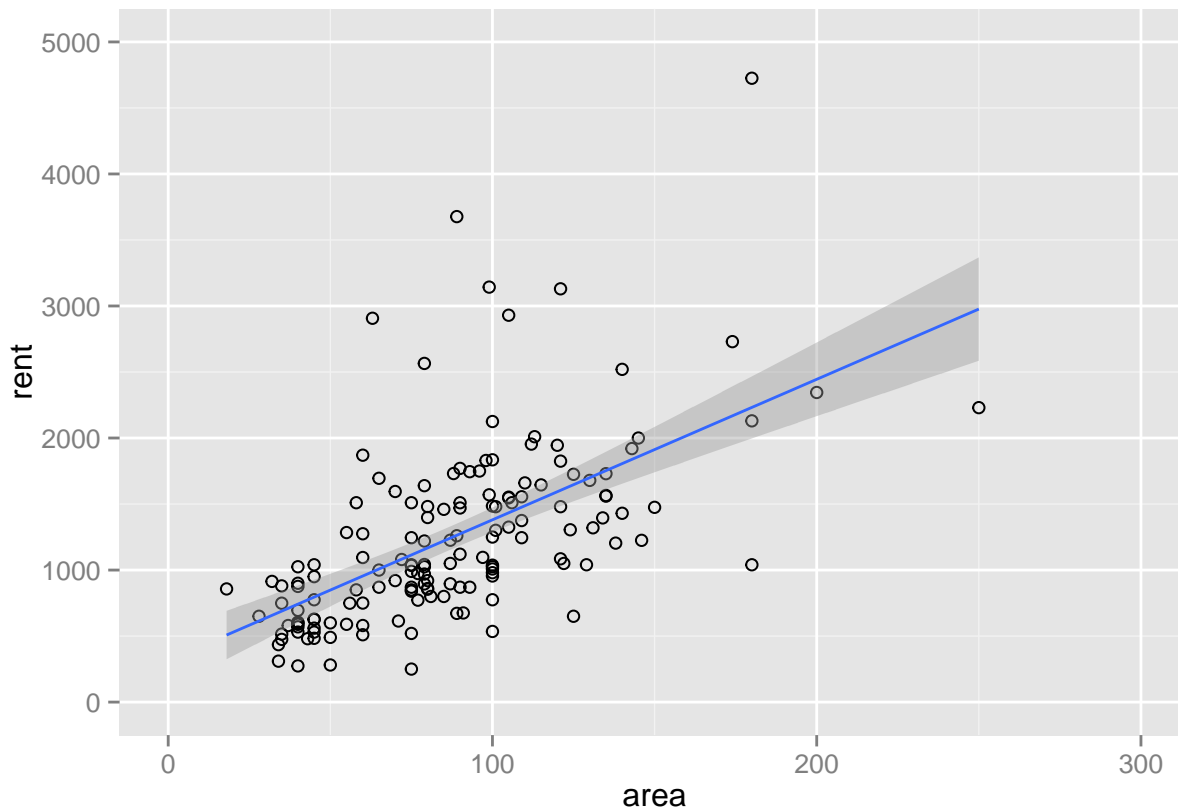
Conclusion

For 4 and 5 room appartments the median and mean rent amounts of the NRE appartments are lower.

Task 6

This sections shows that the variables rent and the area are associated (based on a linear model).

```
ggplot(housingrents, aes(x=area, y=rent)) +  
  geom_point(shape=1) + geom_smooth(method=lm) +  
  ylim(0, 5000) +  
  xlim(0,300)
```



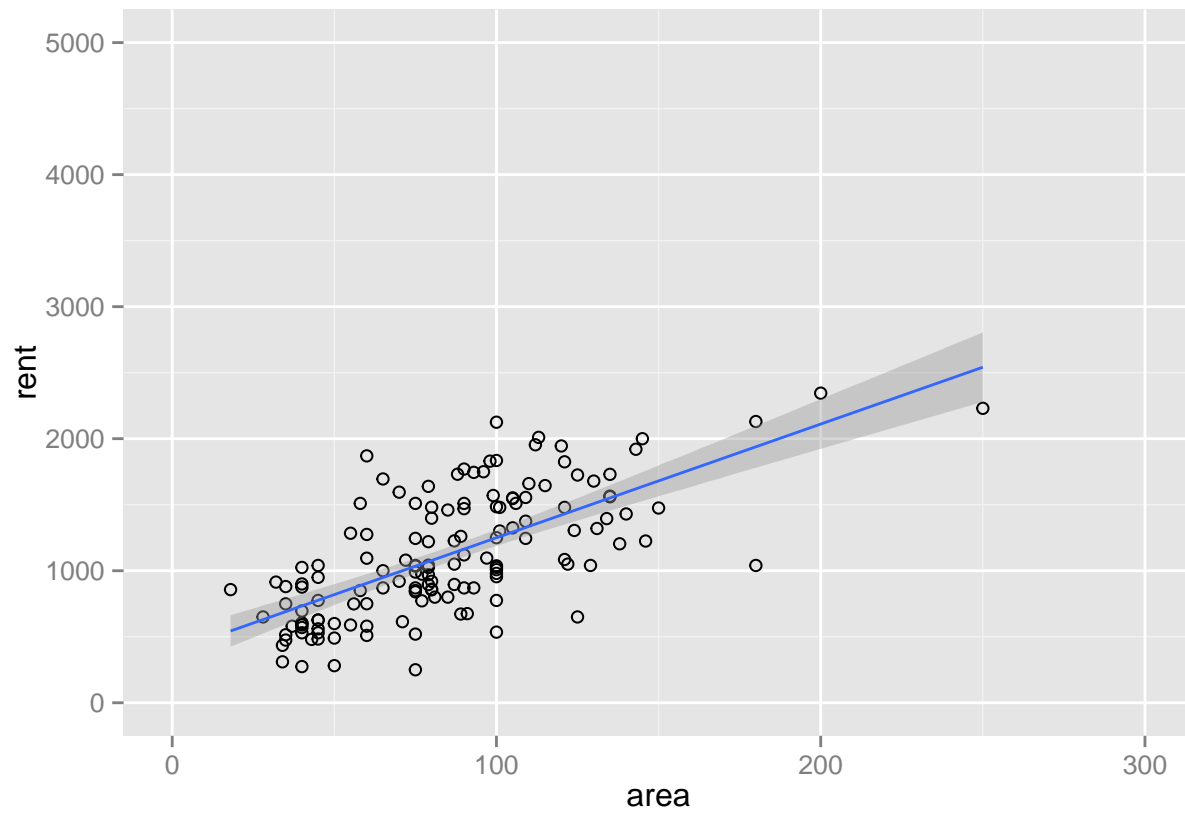
The outliers are appartments with a high rent >2500

```
filter(housingrents, rent > 2500)
```

##	id	rooms	area	rent	nre	econage	balcony
## 1	60	3	79	2565	no	43	no
## 2	64	3	63	2907	yes	13	no
## 3	98	4	89	3677	no	0	yes
## 4	102	4	99	3143	yes	19	<NA>
## 5	119	4	105	2930	no	12	yes
## 6	125	5	174	2730	no	4	yes
## 7	141	5	140	2520	no	10	yes
## 8	142	5	121	3130	yes	33	yes
## 9	151	6	180	4725	yes	24	no

Without the outliers the association (based on a linear model) between the rent and the area stays around the same.

```
ggplot(filter(housingrents, rent <= 2500 ), aes(x=area, y=rent)) +  
  geom_point(shape=1) + geom_smooth(method=lm) +  
  ylim(0, 5000) +  
  xlim(0,300)
```

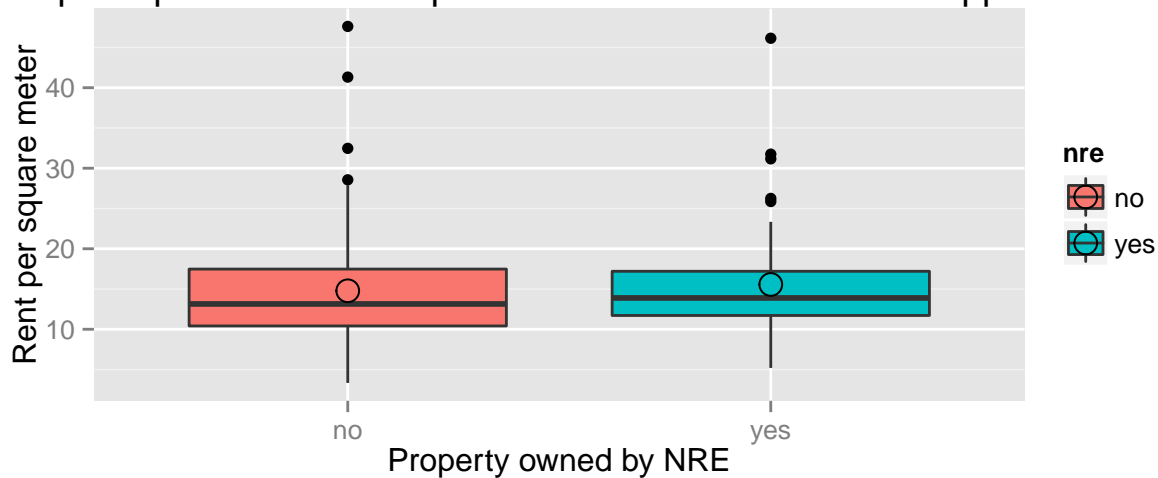


Task 7

In this section the rent per square meter by NRE and non-NRE apartments are analyzed.

```
#Create new variable rps (rent per square meter)
housingrents <- mutate(housingrents,rps = rent/area)
ggplot(housingrents,aes(y=rps,x=nre, fill=nre)) +
  geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", shape=1, size=4) +
  xlab("Property owned by NRE") +
  ylab("Rent per square meter") +
  ggtitle("Rent per square meter Comparison of NRE and non-NRE appartments")
```

Rent per square meter Comparison of NRE and non-NRE appartments



The following table shows the mean and median rent per square meter by NRE and non-NRE appartments:

```
select(housingrents,rps,nre) %>%
  group_by(nre) %>%
  summarise(mean = mean(rps), median = median(rps) )
```

```
## Source: local data frame [2 x 3]
##
##   nre    mean  median
## 1  no 14.77912 13.14286
## 2  yes 15.56434 13.88889
```

Conclusion

The table shows that NRE are only slightly more expensive when measured by rent per square meter.