STAT420 Final Project

Jocelyn Xu, Anahi Rodriguez

Due 11/4/2022

Final Project

```
library(ggplot2)
```

Data Set Up

```
# immo = read.csv("apartments.csv")
# head(immo)
# dim(immo)
```

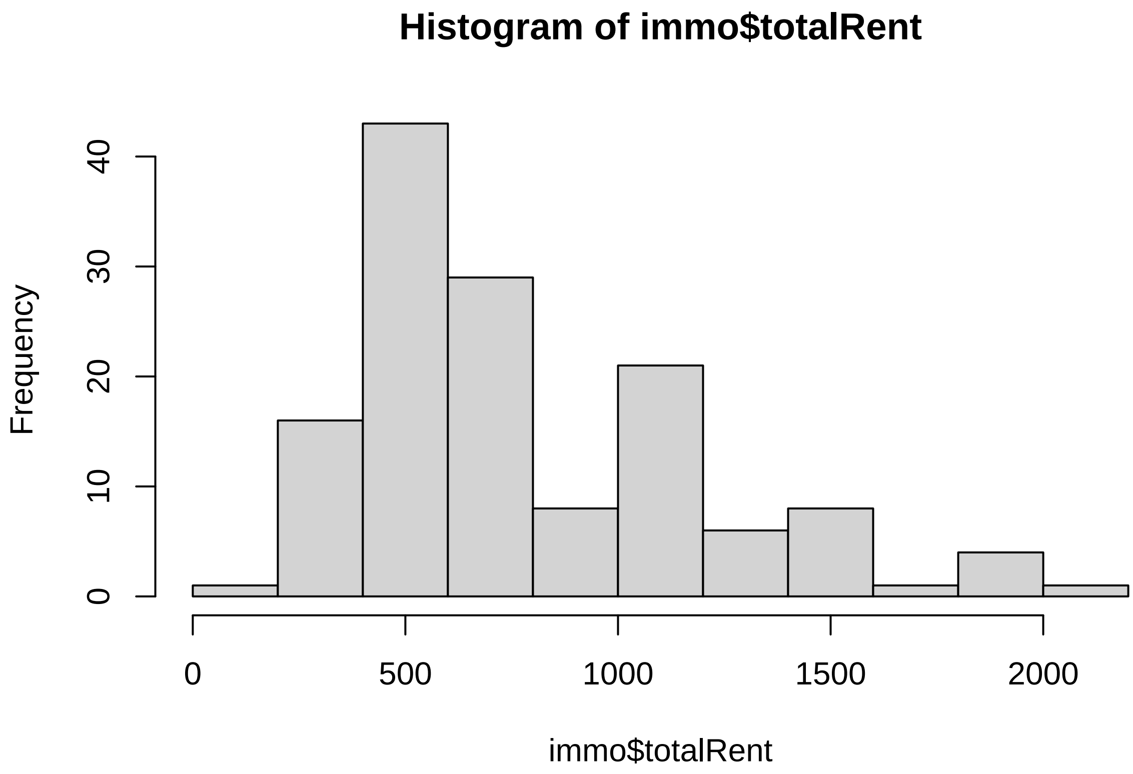
Histogram and summary on response variable

```
immo$totalRentNoHigh = with(immo, ifelse(immo$totalRent < 1000, immo$totalRent, NA))

# original summary and histogram for response variable before trimming
summary(immo$totalRent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   177.2   480.0   647.5   784.9  1075.0  2110.0
```

```
hist(immo$totalRent)
```



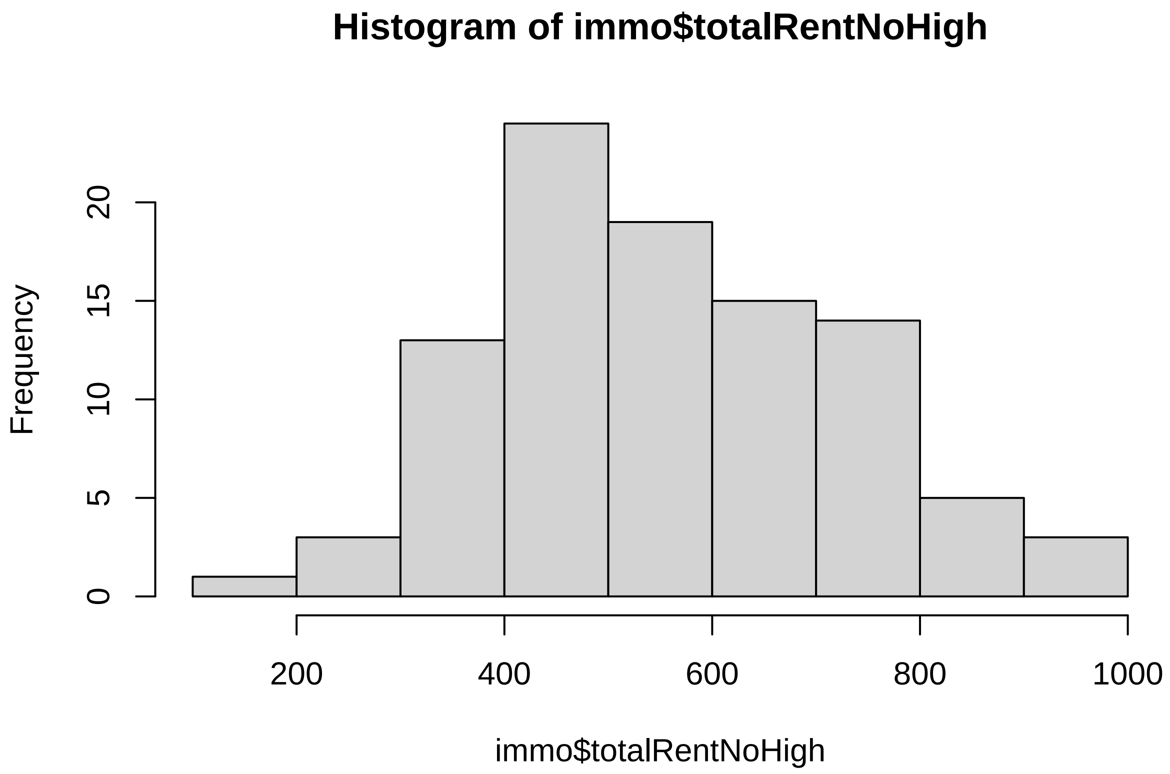
```
sd(immo$totalRent)
```

```
## [1] 405.4092
```

```
# original summary and histogram for response variable after trimming
summary(immo$totalRentNoHigh)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   177.2   440.0   540.0   560.7   680.0   975.0      41
```

```
hist(immo$totalRentNoHigh)
```



```
sd(immo$totalRentNoHigh, na.rm = T)
```

```
## [1] 170.2261
```

Answer:
Response variable of interest: total rent

Explanation on taking off the high end:
With the high end, It is very hard to tell the shape of the histogram because of the wide range and the lack of data on the high end. Although we do lose nearly 30% of the date doing this, we chose the threshold <1000 because the distribution is closer to a normal distribution and has a lower overall variance.

Interpretation on histogram:
The histogram shows the spread of the total rent in Germany. Overall, the histogram is right-skewed, with a center around mid 500s and the range from 0 to 975 with a cutting threshold or 0 to 2110 without a cutting threshold.

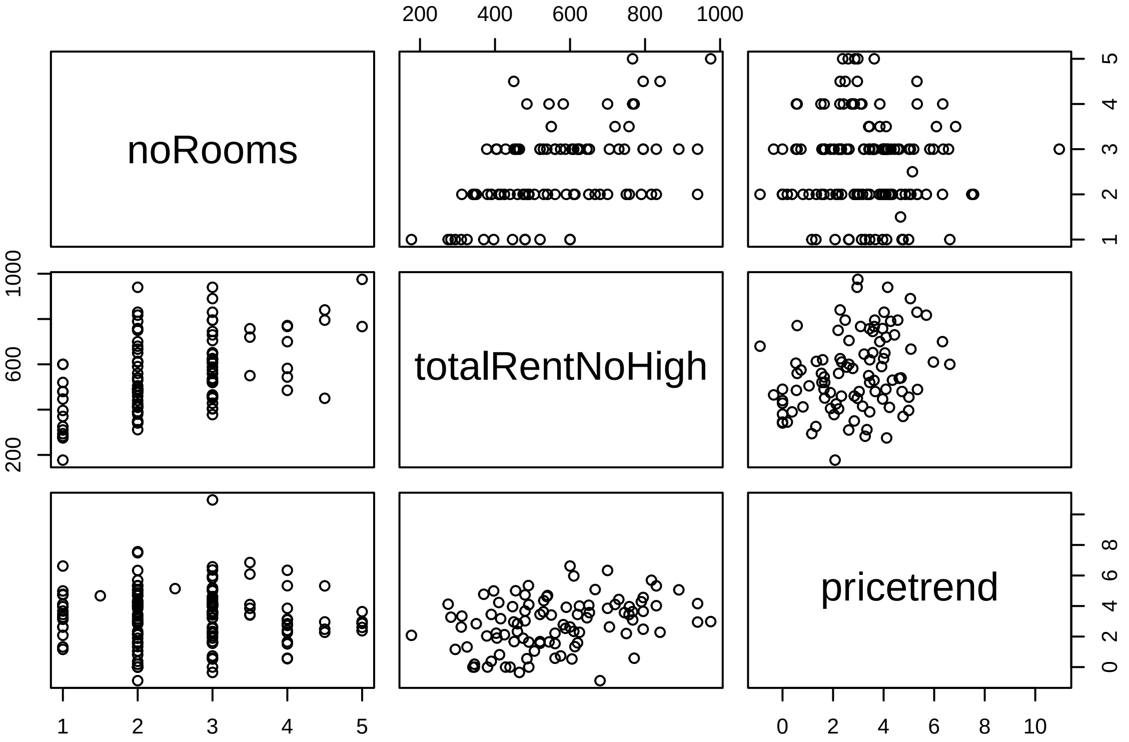
Scatterplots on quantitative predictors

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
for_matrices = immo %>%
  select(noRooms, totalRentNoHigh, pricetrend)

pairs(for_matrices)
```



```
# ggplot(immo, aes(x = noRooms, y = totalRentNoHigh)) +
#   geom_point() +
#   geom_smooth(method = 'lm', se = F, color = 'purple', formula = 'y ~ x') +
#   labs(x = "number of rooms", y = "total rent",
#     title = "Scatterplot of the relationship between the number of rooms and the total rent")
#
#
# ggplot(immo, aes(x = pricetrend, y = totalRentNoHigh)) +
#   geom_point() +
#   geom_smooth(method = 'lm', se = F, color = 'purple', formula = 'y ~ x') +
#   labs(x = "price trend", y = "total rent",
#     title = "Scatterplot of the relationship between the price trend and the total rent")
```

Answer:
Interpretation:
It seems that there is a large variability in the price trend variable when plotted against total rent, however, I still am able to recognize that this is a positive relationship. For the number of rooms plot, I see that there is a much greater variability for two and three bedroom apartments compared to one, four, and five bedroom apartments. Additionally, I see that there exists data for three and a half as well as four and a half bedroom apartments, possibly accounting for some of the missing variability at the higher end that we see around two and three bedrooms. Lastly, the relationship between our predictor variables is hard to identify but it seems that therre exists a weak positive relationship.

Fit a model

```
immo_model = lm(totalRentNoHigh ~ noRooms + pricetrend +balcony + typeOfFlat,
                data=immo)
summary(immo_model)
```

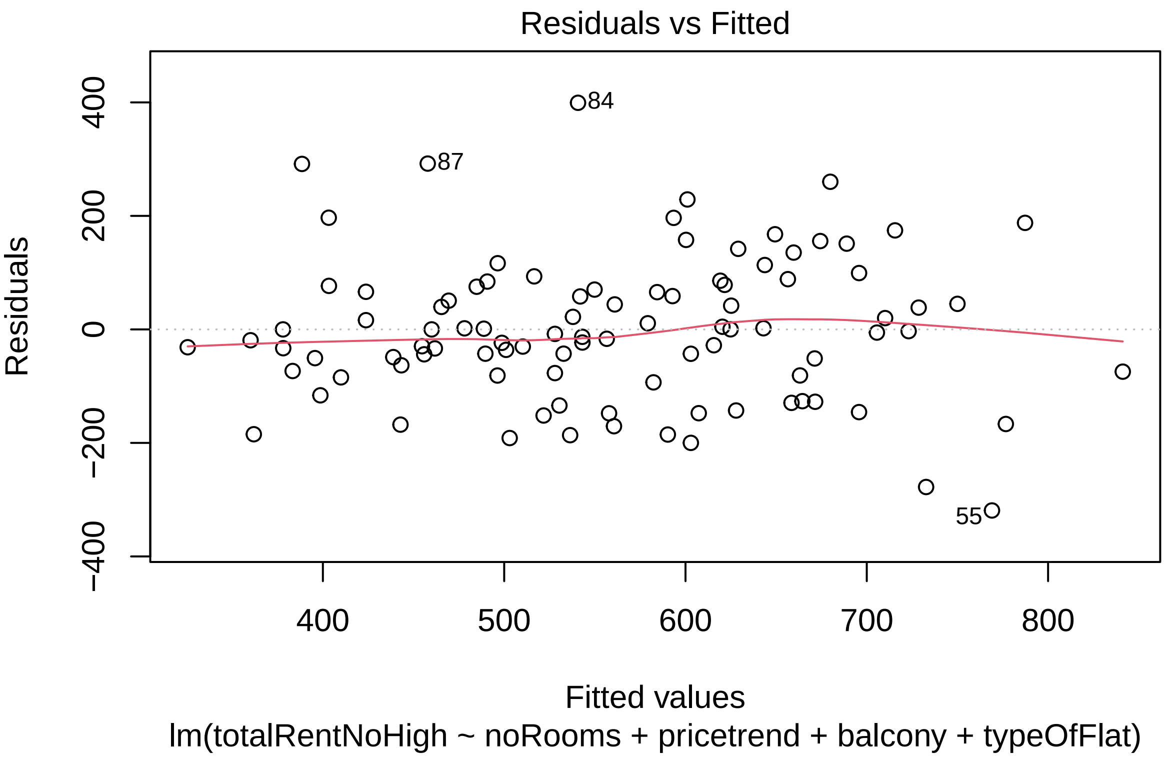
```
##
## Call:
## lm(formula = totalRentNoHigh ~ noRooms + pricetrend + balcony +
##     typeOfFlat, data = immo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -319.05  -76.98   -5.59   75.17  399.26
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   188.270     50.544   3.725 0.000344 ***
## noRooms                        91.164     14.475   6.298 1.15e-08 ***
## pricetrend                     39.653      8.319   4.767 7.35e-06 ***
## balconyTRUE                    53.167     29.023   1.832 0.070351 .
## typeOfFlatground_floor        -63.643     69.314  -0.918 0.361036
## typeOfFlathalf_basement       -76.378    134.522  -0.568 0.571632
## typeOfFlatother                25.036     57.283   0.437 0.663138
## typeOfFlatraised_ground_floor -217.820    134.450  -1.620 0.108793
## typeOfFlatroof_storey          19.521     40.371   0.484 0.629914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 133 on 88 degrees of freedom
##   (41 observations deleted due to missingness)
## Multiple R-squared:  0.4407, Adjusted R-squared:  0.3898
## F-statistic: 8.666 on 8 and 88 DF,  p-value: 1.195e-08
```

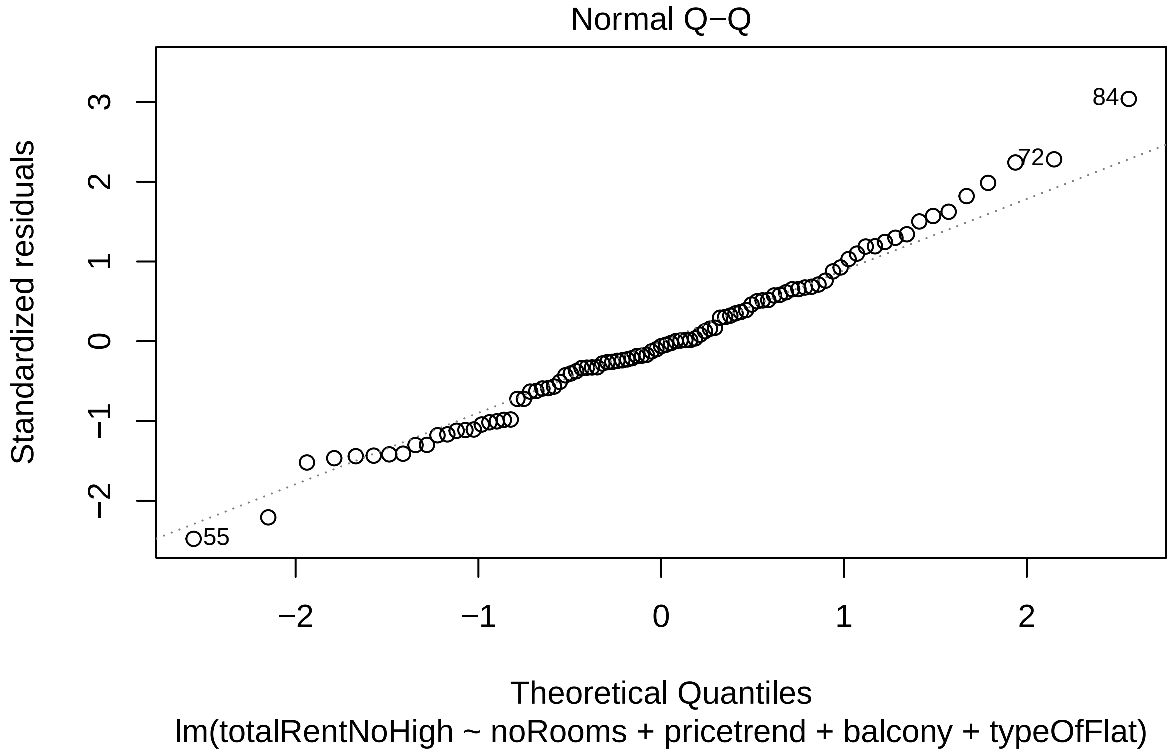
Answer:
fitted model: estimated total rent = 188.270 + (91.164 * number of rooms) + (39.653 * price trend) + (53.167 * balcony) + (-63.643 * ground floor flat) + (-76.378 * basemant flat) + (25.036 * other type of flat) + (-217.820 * raised ground floor flat) + (19.521 * roof flat) R_square = 0.4407

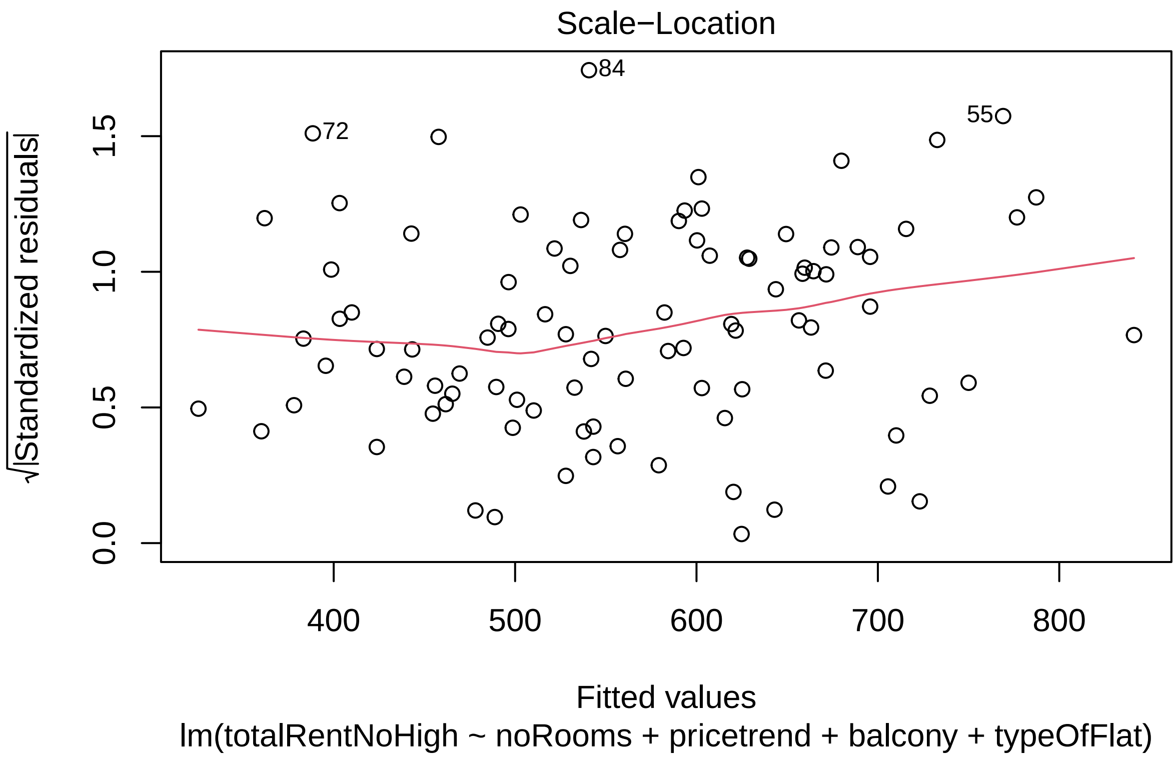
Plot the model

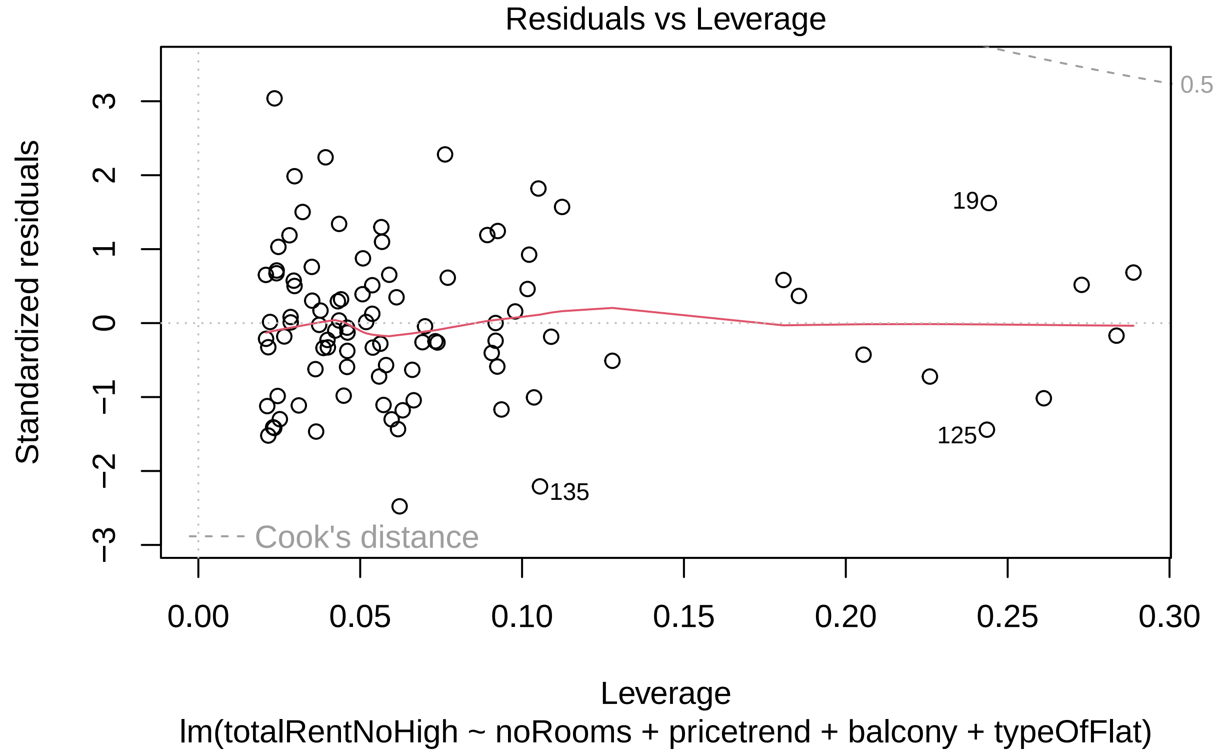
```
plot(immo_model)
```

```
## Warning: not plotting observations with leverage one:
##   41, 95
```


Residuals vs Fitted
Residuals
400
200
0
−200
−400
400
500
600
700
800
84
87
55
Fitted values
lm(totalRentNoHigh ~ noRooms + pricetrend + balcony + typeOfFlat)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−2
−1
0
1
2
84
72
55
Theoretical Quantiles
lm(totalRentNoHigh ~ noRooms + pricetrend + balcony + typeOfFlat)





Project Update Notes:

- We have removed out base rent predictor variable and replaced it with a different quantitative variable: price trend
- We have also trimmed our data