The Real Estate Market in Germany

Data have been converted to csv .

So the first step is to import the data into R studio.

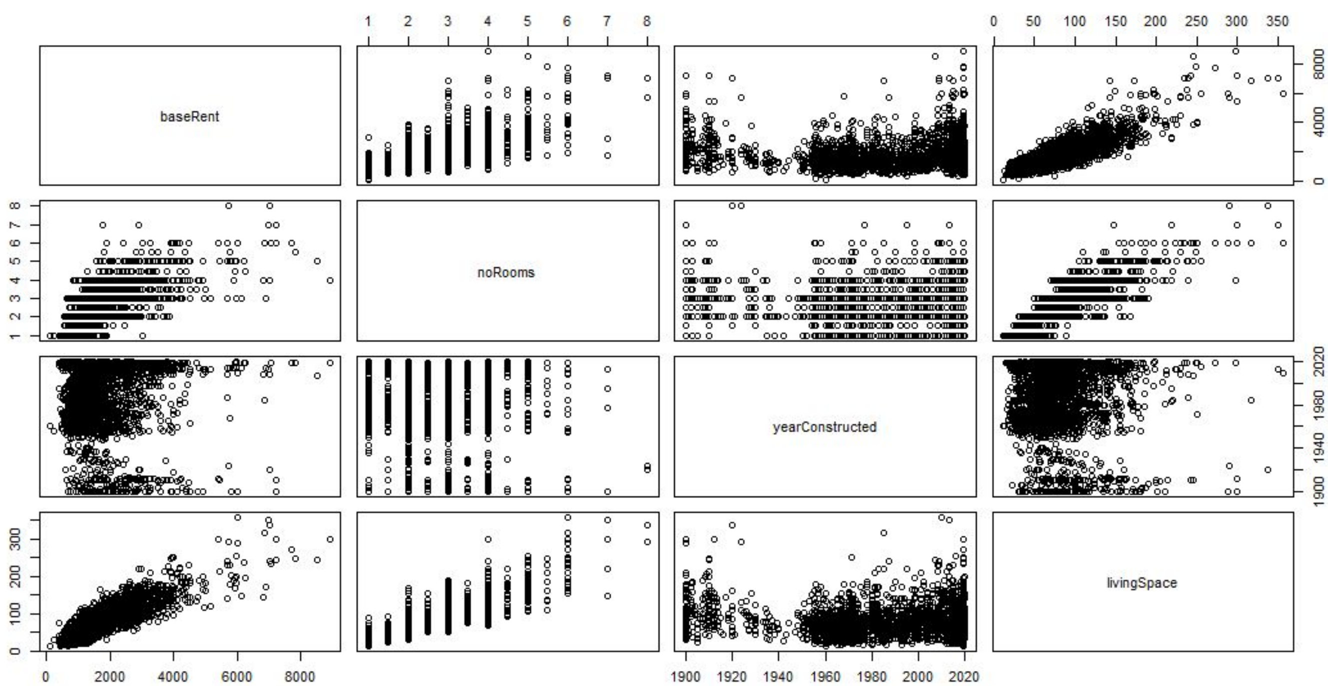
Line 1: `MUNICH_data<- read.csv("MUNICH_data.csv",stringsAsFactors = FALSE)`

Line 2 (determine the dimension of the data): `dim(MUNICH_data)`

Response: 3111 by 30. (from over 266,000 down to this)

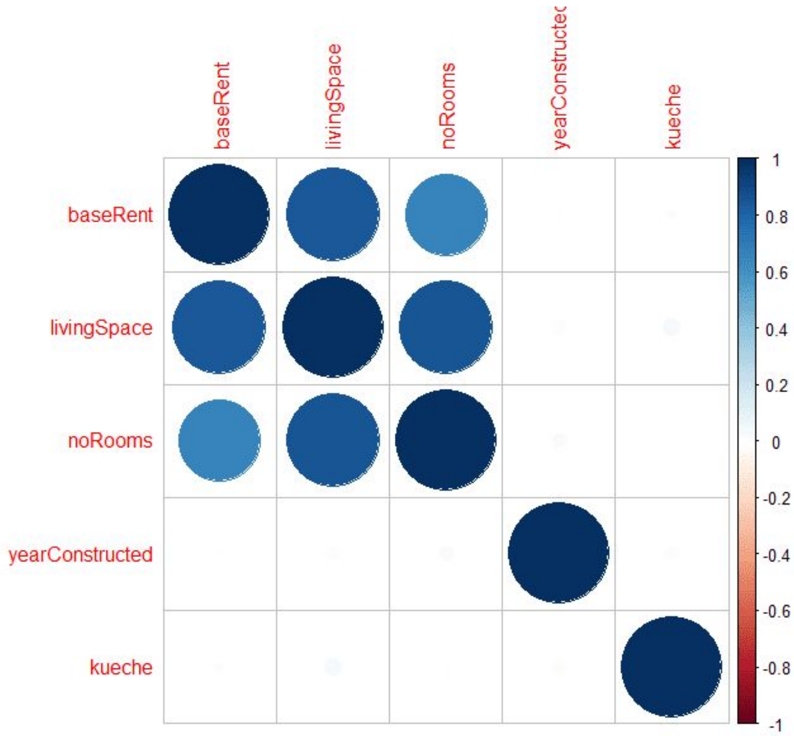
Line 3 (Box plot): `pairs(~ baseRent+noRooms+yearConstructed+livingSpace , MUNICH_data).`

Response:



Observation: Positive relationship between rent and living space, rooms.

Line 4 (correlation plot without causality): corrplot(corr, method = "circle")
Response:



Observation: Not surprising when you look at the box plot.

Now let's start predicting models. We start with the simple multiple linear regression.

Line 5 (lin, lin model): regression_2 <- lm(baseRent ~ livingSpace+noRooms+yearConstructed+balcony1+kueche,MUNICH_data)

Response:

```
> summary(regression_2)

Call:
lm(formula = baseRent ~ livingSpace + noRooms + yearConstructed +
    balcony1 + kueche, data = MUNICH_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1765.2  -317.8   -46.2   265.1  3843.5

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -738.1498   569.0987  -1.297   0.1947
livingSpace       23.4134     0.4344  53.899   <2e-16 ***
noRooms         -171.5369    16.1678 -10.610   <2e-16 ***
yearConstructed    0.5388     0.2872   1.876   0.0607 .
balcony1          -8.9558    23.7531  -0.377   0.7062
kueche           -40.8004    18.3689  -2.221   0.0264 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 500.1 on 3105 degrees of freedom
Multiple R-squared:  0.7108,    Adjusted R-squared:  0.7104
F-statistic:  1526 on 5 and 3105 DF,  p-value: < 2.2e-16
```

Observation: 71% of the data was captured by the model. For a data set of over 3000, this is very good. But we also see that the forecasting error is 500 Eur/sqm. This is quite a lot and the model needs to be worked on.

Line 6 (log log model): regression_3 <- lm(log(baseRent) ~ log(livingSpace)+noRooms+yearConstructed+balcony1+kueche,MUNICH_data)

Response:

```
> summary(regression_3)

Call:
lm(formula = log(baseRent) ~ log(livingSpace) + noRooms + yearConstructed +
    balcony1 + kueche, data = MUNICH_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.46375 -0.18503 -0.00816  0.19914  0.94873 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.2503191  0.3326714  12.776   <2e-16 ***
log(livingSpace)  0.7815750  0.0203918  38.328   <2e-16 ***
noRooms          -0.0108088  0.0095021  -1.138    0.255    
yearConstructed  -0.0001134  0.0001628  -0.697    0.486    
balcony1          0.0217575  0.0135501   1.606    0.108    
kueche           -0.0039313  0.0104026  -0.378    0.706    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2839 on 3105 degrees of freedom
Multiple R-squared:  0.6493,	Adjusted R-squared:  0.6487 
F-statistic:  1150 on 5 and 3105 DF,  p-value: < 2.2e-16
```

Observation: 65% of the data was captured by the model. But we also see that the forecasting error is 28%.

Conclusion:

It goes without saying that this is an extremely simple model. We can improve upon our predictions quite a lot with some heavier feature engineering. I encourage those who are interested to try playing around with this data set by including more variables and trying out different model prediction approaches.