German Rental Advanced Analysis and predction Models Part 1: Feature Analysis and Engeneering

Based on the Kaggle immo_data dataset. More informations can be found here: https://www.kaggle.com/datasets/corrieaar/apartment-rental-offers-in-germany

packages install

```
mirror = "http://cran.us.r-project.org"
install.packages("moments", repos=mirror)
```

```
## Installing package into 'C:/Users/Nutzer/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'moments' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Nutzer\AppData\Local\Temp\RtmpOKQna5\downloaded_packages
```

```
install.packages("ggplot2", repos=mirror)
```

```
## Installing package into 'C:/Users/Nutzer/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Nutzer\AppData\Local\Temp\RtmpOKQna5\downloaded_packages
```

```
install.packages("car", repos=mirror)
```

```
## Installing package into 'C:/Users/Nutzer/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'car' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Nutzer\AppData\Local\Temp\RtmpOKQna5\downloaded_packages
```

```
install.packages("MASS", repos=mirror)
```

```
## Installing package into 'C:/Users/Nutzer/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'MASS' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'MASS'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): Problem C:
## \Users\Nutzer\Documents\R\win-library\4.0\00LOCK\MASS\libs\x64\MASS.dll nach C:
## \Users\Nutzer\Documents\R\win-library\4.0\MASS\libs\x64\MASS.dll zu kopieren:
## Permission denied
```

```
## Warning: restored 'MASS'
```

```
##
## The downloaded binary packages are in
##  C:\Users\Nutzer\AppData\Local\Temp\RtmpOKQna5\downloaded_packages
```

```
install.packages("dummies", repos=mirror)
```

```
## Installing package into 'C:/Users/Nutzer/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'dummies' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Nutzer\AppData\Local\Temp\RtmpOKQna5\downloaded_packages
```

packages loads

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

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library("car")
```

```
## Warning: package 'car' was built under R version 4.0.5
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.0.5
```

```
library("MASS")
```

```
## Warning: package 'MASS' was built under R version 4.0.5
```

```
library("dummies")
```

```
## dummies-1.5.6 provided by Decision Patterns
```

Preparing the Data

```
data_raw = read.csv('immo_data_raw.csv')
head(data_raw)
```

```
##                regio1 serviceCharge                   heatingType
## 1 Nordrhein_Westfalen        245.00                central_heating
## 2     Rheinland_Pfalz        134.00 self_contained_central_heating
## 3             Sachsen        255.00                  floor_heating
## 4             Sachsen         58.15               district_heating
## 5              Bremen        138.00 self_contained_central_heating
## 6  Schleswig_Holstein        142.00                    gas_heating
##   telekomTvOffer telekomHybridUploadSpeed newlyConst balcony picturecount
## 1  ONE_YEAR_FREE                       NA      FALSE   FALSE            6
## 2  ONE_YEAR_FREE                       NA      FALSE    TRUE            8
## 3  ONE_YEAR_FREE                       10       TRUE    TRUE            8
## 4  ONE_YEAR_FREE                       NA      FALSE    TRUE            9
## 5           <NA>                       NA      FALSE    TRUE           19
## 6           NONE                       NA      FALSE    TRUE            5
##   pricetrend telekomUploadSpeed totalRent yearConstructed   scoutId
## 1       4.62               10.0       840            1965  96107057
## 2       3.47               10.0        NA            1871 111378734
```

```
## 3       2.72               2.4      1300            2019 113147523
## 4       1.53              40.0        NA            1964 108890903
## 5       2.46                NA       903            1950 114751222
## 6       4.48               2.4        NA            1999 115531145
##   noParkSpaces      firingTypes hasKitchen            geo_bln cellar
## 1            1              oil      FALSE Nordrhein_Westfalen   TRUE
## 2            2              gas      FALSE     Rheinland_Pfalz  FALSE
## 3            1             <NA>      FALSE             Sachsen   TRUE
## 4           NA district_heating      FALSE             Sachsen  FALSE
## 5           NA              gas      FALSE              Bremen  FALSE
## 6           NA              gas       TRUE  Schleswig_Holstein  FALSE
##   yearConstructedRange baseRent houseNumber livingSpace
## 1                    2    595.0         244       86.00
## 2                    1    800.0        <NA>       89.00
## 3                    9    965.0           4       83.80
## 4                    2    343.0          35       58.15
## 5                    1    765.0          10       84.97
## 6                    5    315.2          1e       53.43
##                    geo_krs      condition  interiorQual petsAllowed
## 1                  Dortmund      well_kept        normal        <NA>
## 2         Rhein_Pfalz_Kreis    refurbished        normal          no
## 3                   Dresden first_time_use sophisticated        <NA>
## 4       Mittelsachsen_Kreis           <NA>          <NA>        <NA>
## 5                    Bremen    refurbished          <NA>        <NA>
## 6 Schleswig_Flensburg_Kreis      well_kept          <NA>          no
##                         street                 streetPlain  lift baseRentRange
## 1  Sch&uuml;ruferstra&szlig;e           SchÃ¼ruferstraÃŸe FALSE             4
## 2              no_information                       <NA> FALSE             5
## 3                   Turnerweg                 Turnerweg  TRUE             6
## 4 Gl&uuml;ck-Auf-Stra&szlig;e          GlÃ¼ck-Auf-StraÃŸe FALSE             2
## 5 Hermann-Henrich-Meier-Allee Hermann-Henrich-Meier-Allee FALSE             5
## 6                 Hardeseiche                Hardeseiche FALSE             2
##     typeOfFlat geo_plz noRooms thermalChar floor numberOfFloors noRoomsRange
## 1 ground_floor   44269       4       181.4     1             3            4
## 2 ground_floor   67459       3          NA    NA            NA            3
## 3    apartment    1097       3          NA     3             4            3
## 4        other    9599       3        86.0     3            NA            3
## 5    apartment   28213       3       188.9     1            NA            3
## 6    apartment   24891       2       165.0    NA            NA            2
##   garden livingSpaceRange                  regio2
## 1   TRUE               4                 Dortmund
## 2  FALSE               4        Rhein_Pfalz_Kreis
## 3  FALSE               4                  Dresden
## 4  FALSE               2      Mittelsachsen_Kreis
## 5  FALSE               4                   Bremen
## 6  FALSE               2 Schleswig_Flensburg_Kreis
##                       regio3
## 1                     SchÃ¼ren
## 2                BÃ¶hl_Iggelheim
## 3 Ã„uÃŸere_Neustadt_Antonstadt
## 4                    Freiberg
## 5           Neu_Schwachhausen
## 6                    Struxdorf
##
```

```
## 1
## 2 Alles neu macht der Mai â\200" so kann es auch fÃ¼r Sie in 2019 sein! GenieÃŸen Sie das â\200žreine
## 3
## 4
## 5
## 6
##
## 1
## 2
## 3
## 4
## 5 Diese Wohnung wurde neu saniert und ist wie folgt ausgestattet:\n\n- 3 gerÃ¤umige Zimmer\n- Wohnzi
## 6
##   heatingCosts energyEfficiencyClass lastRefurbish electricityBasePrice
## 1           NA                  <NA>            NA                   NA
## 2           NA                  <NA>          2019                   NA
## 3           NA                  <NA>            NA                   NA
## 4        87.23                  <NA>            NA                   NA
## 5           NA                  <NA>            NA                   NA
## 6           NA                  <NA>            NA                   NA
##   electricityKwhPrice  date
## 1                  NA May19
## 2                  NA May19
## 3                  NA Oct19
## 4                  NA May19
## 5                  NA Feb20
## 6                  NA Feb20
```

Since I am interested in rental price prediction the first thing is to just drop all rows where this information is not given. One could impute these values one way or another but to keep the accuracy on the data as high as possible I am just going to removes these entries from the data.

```
rental_na = which(is.na(data_raw$totalRent))
removal_fraction = length(rental_na) / length(data_raw$totalRent)
removal_fraction
```

```
## [1] 0.1507049
```

```
data_raw = data_raw[-rental_na, ]
head(data_raw)
```

```
##                regio1 serviceCharge                   heatingType
## 1 Nordrhein_Westfalen           245                central_heating
## 3             Sachsen           255                  floor_heating
## 5              Bremen           138 self_contained_central_heating
## 7             Sachsen            70 self_contained_central_heating
## 8              Bremen            88                central_heating
## 9  Baden_WÃ¼rttemberg           110                    oil_heating
##   telekomTvOffer telekomHybridUploadSpeed newlyConst balcony picturecount
## 1  ONE_YEAR_FREE                       NA      FALSE   FALSE            6
## 3  ONE_YEAR_FREE                       10       TRUE    TRUE            8
## 5           <NA>                       NA      FALSE    TRUE           19
## 7  ONE_YEAR_FREE                       10      FALSE   FALSE            9
## 8  ONE_YEAR_FREE                       10      FALSE    TRUE            5
## 9  ONE_YEAR_FREE                       NA      FALSE   FALSE            5
##   pricetrend telekomUploadSpeed totalRent yearConstructed   scoutId
```

```
## 1       4.62              10.0    840.00            1965  96107057
## 3       2.72               2.4   1300.00            2019 113147523
## 5       2.46                NA    903.00            1950 114751222
## 7       1.01               2.4    380.00              NA 114391930
## 8       1.89               2.4    584.25            1959 115270775
## 9       3.77              40.0    690.00            1970 106416361
##   noParkSpaces     firingTypes hasKitchen            geo_bln cellar
## 1            1             oil      FALSE Nordrhein_Westfalen   TRUE
## 3            1            <NA>      FALSE            Sachsen   TRUE
## 5           NA             gas      FALSE             Bremen  FALSE
## 7           NA            <NA>      FALSE            Sachsen   TRUE
## 8           NA gas:electricity      FALSE             Bremen   TRUE
## 9            1             oil       TRUE  Baden_WÃ¼rttemberg   TRUE
##   yearConstructedRange baseRent houseNumber livingSpace            geo_krs
## 1                    2   595.00         244       86.00           Dortmund
## 3                    9   965.00           4       83.80            Dresden
## 5                    1   765.00          10       84.97             Bremen
## 7                   NA   310.00          14       62.00 Mittelsachsen_Kreis
## 8                    2   452.25          35       60.30             Bremen
## 9                    2   580.00        <NA>       53.00   Emmendingen_Kreis
##         condition  interiorQual petsAllowed                      street
## 1       well_kept        normal        <NA>  Sch&uuml;ruferstra&szlig;e
## 3  first_time_use sophisticated        <NA>                  Turnerweg
## 5     refurbished          <NA>        <NA> Hermann-Henrich-Meier-Allee
## 7 fully_renovated          <NA>        <NA>                  Am Bahnhof
## 8            <NA>          <NA>        <NA>            Lesumer Heerstr.
## 9       well_kept sophisticated          no              no_information
##                   streetPlain  lift baseRentRange   typeOfFlat geo_plz noRooms
## 1           SchÃ¼ruferstraÃŸe FALSE             4 ground_floor   44269       4
## 3                    Turnerweg  TRUE             6    apartment    1097       3
## 5 Hermann-Henrich-Meier-Allee FALSE             5    apartment   28213       3
## 7                  Am_Bahnhof FALSE             2         <NA>    9599       2
## 8            Lesumer_Heerstr. FALSE             3 ground_floor   28717       3
## 9                       <NA> FALSE             4  roof_storey   79211       2
##   thermalChar floor numberOfFloors noRoomsRange garden livingSpaceRange
## 1       181.4     1              3            4   TRUE               4
## 3          NA     3              4            3  FALSE               4
## 5       188.9     1             NA            3  FALSE               4
## 7          NA     1              4            2   TRUE               3
## 8        63.0    NA             NA            3  FALSE               2
## 9       138.0     2              2            2  FALSE               2
##               regio2                     regio3
## 1           Dortmund                   SchÃ¼ren
## 3            Dresden Ã„uÃŸere_Neustadt_Antonstadt
## 5             Bremen            Neu_Schwachhausen
## 7 Mittelsachsen_Kreis                   Freiberg
## 8             Bremen                  St._Magnus
## 9  Emmendingen_Kreis                  Denzlingen
##
## 1
## 3 Der Neubau entsteht im Herzen der Dresdner Neustadt.\nDas BaugrundstÃ¼ck befindet sich inmitten ei
## 5
## 7
## 8
```

```
## 9
##
## 1
## 3
## 5 Diese Wohnung wurde neu saniert und ist wie folgt ausgestattet:\n\n- 3 gerÃ¤umige Zimmer\n- Wohnzi
## 7
## 8
## 9
##   heatingCosts energyEfficiencyClass lastRefurbish electricityBasePrice
## 1           NA                  <NA>            NA                   NA
## 3           NA                  <NA>            NA                   NA
## 5           NA                  <NA>            NA                   NA
## 7           NA                  <NA>            NA                   NA
## 8           44                     B            NA                   NA
## 9           NA                     E            NA                   NA
##   electricityKwhPrice  date
## 1                  NA May19
## 3                  NA Oct19
## 5                  NA Feb20
## 7                  NA Feb20
## 8                  NA Feb20
## 9                  NA Feb20
```

facilities and description columns may contain useful information for humans and can be the base for a complex deep learning model but have no value for my regression model so they get droped as well.

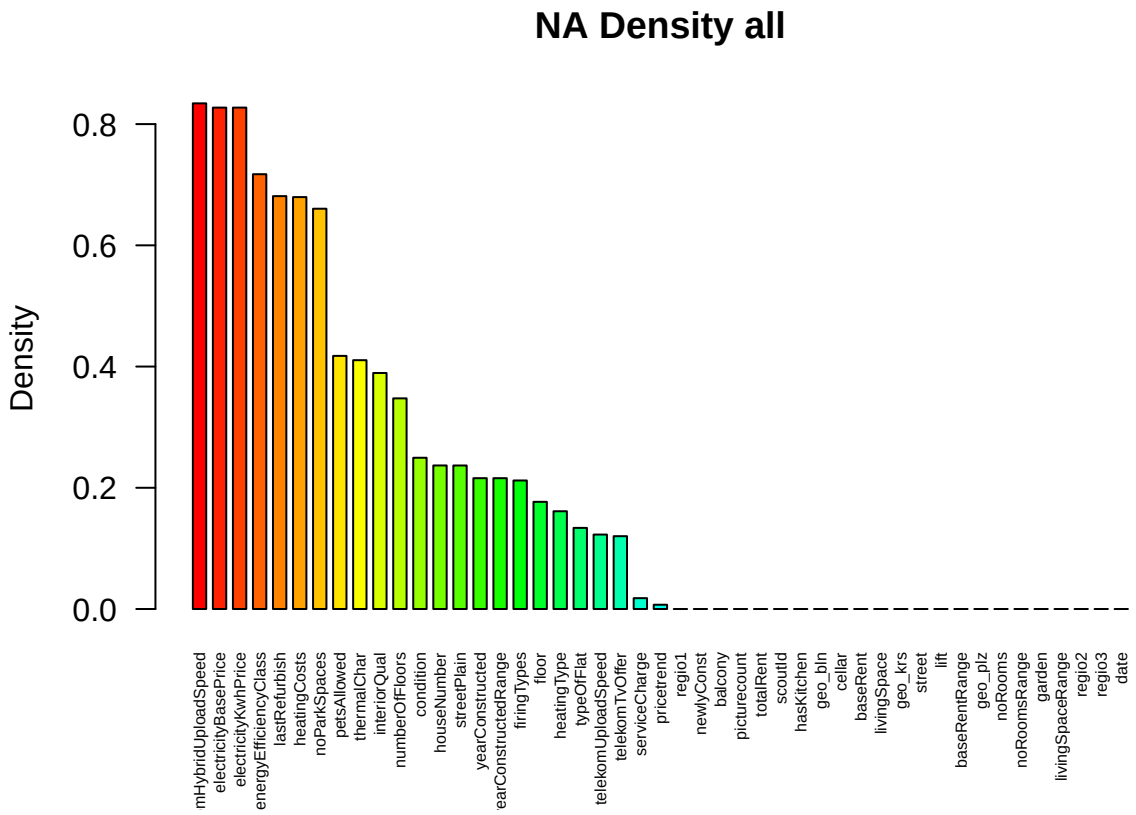
```
data_raw = subset(data_raw, select=-c(facilities, description))
head(data_raw)
```

```
##                regio1 serviceCharge                   heatingType
## 1 Nordrhein_Westfalen           245                central_heating
## 3             Sachsen           255                  floor_heating
## 5              Bremen           138 self_contained_central_heating
## 7             Sachsen            70 self_contained_central_heating
## 8              Bremen            88                central_heating
## 9  Baden_WÃ¼rttemberg           110                    oil_heating
##   telekomTvOffer telekomHybridUploadSpeed newlyConst balcony picturecount
## 1  ONE_YEAR_FREE                      NA      FALSE   FALSE            6
## 3  ONE_YEAR_FREE                      10       TRUE    TRUE            8
## 5           <NA>                      NA      FALSE    TRUE           19
## 7  ONE_YEAR_FREE                      10      FALSE   FALSE            9
## 8  ONE_YEAR_FREE                      10      FALSE    TRUE            5
## 9  ONE_YEAR_FREE                      NA      FALSE   FALSE            5
##   pricetrend telekomUploadSpeed totalRent yearConstructed   scoutId
## 1       4.62               10.0    840.00            1965  96107057
## 3       2.72                2.4   1300.00            2019 113147523
## 5       2.46                 NA    903.00            1950 114751222
## 7       1.01                2.4    380.00              NA 114391930
## 8       1.89                2.4    584.25            1959 115270775
## 9       3.77               40.0    690.00            1970 106416361
##   noParkSpaces     firingTypes hasKitchen             geo_bln cellar
## 1            1             oil      FALSE Nordrhein_Westfalen   TRUE
## 3            1            <NA>      FALSE             Sachsen   TRUE
## 5           NA             gas      FALSE              Bremen  FALSE
## 7           NA            <NA>      FALSE             Sachsen   TRUE
```

```
## 8           NA gas:electricity      FALSE              Bremen   TRUE
## 9            1             oil       TRUE  Baden_WÃ¼rttemberg   TRUE
##   yearConstructedRange baseRent houseNumber livingSpace             geo_krs
## 1                    2   595.00         244       86.00            Dortmund
## 3                    9   965.00           4       83.80             Dresden
## 5                    1   765.00          10       84.97              Bremen
## 7                   NA   310.00          14       62.00 Mittelsachsen_Kreis
## 8                    2   452.25          35       60.30              Bremen
## 9                    2   580.00        <NA>       53.00   Emmendingen_Kreis
##         condition  interiorQual petsAllowed                      street
## 1       well_kept        normal        <NA>  Sch&uuml;ruferstra&szlig;e
## 3  first_time_use sophisticated        <NA>                   Turnerweg
## 5     refurbished          <NA>        <NA> Hermann-Henrich-Meier-Allee
## 7 fully_renovated          <NA>        <NA>                  Am Bahnhof
## 8            <NA>          <NA>        <NA>            Lesumer Heerstr.
## 9       well_kept sophisticated          no              no_information
##                   streetPlain  lift baseRentRange   typeOfFlat geo_plz noRooms
## 1           SchÃ¼ruferstraÃŸe FALSE             4 ground_floor   44269       4
## 3                   Turnerweg  TRUE             6    apartment    1097       3
## 5 Hermann-Henrich-Meier-Allee FALSE             5    apartment   28213       3
## 7                  Am_Bahnhof FALSE             2         <NA>    9599       2
## 8            Lesumer_Heerstr. FALSE             3 ground_floor   28717       3
## 9                        <NA> FALSE             4  roof_storey   79211       2
##   thermalChar floor numberOfFloors noRoomsRange garden livingSpaceRange
## 1       181.4     1              3            4   TRUE               4
## 3          NA     3              4            3  FALSE               4
## 5       188.9     1             NA            3  FALSE               4
## 7          NA     1              4            2   TRUE               3
## 8        63.0    NA             NA            3  FALSE               2
## 9       138.0     2              2            2  FALSE               2
##                regio2                       regio3 heatingCosts
## 1            Dortmund                     SchÃ¼ren           NA
## 3             Dresden Ã„uÃŸere_Neustadt_Antonstadt           NA
## 5              Bremen            Neu_Schwachhausen           NA
## 7 Mittelsachsen_Kreis                      Freiberg           NA
## 8              Bremen                   St._Magnus           44
## 9   Emmendingen_Kreis                   Denzlingen           NA
##   energyEfficiencyClass lastRefurbish electricityBasePrice electricityKwhPrice
## 1                  <NA>            NA                   NA                  NA
## 3                  <NA>            NA                   NA                  NA
## 5                  <NA>            NA                   NA                  NA
## 7                  <NA>            NA                   NA                  NA
## 8                     B            NA                   NA                  NA
## 9                     E            NA                   NA                  NA
##    date
## 1 May19
## 3 Oct19
## 5 Feb20
## 7 Feb20
## 8 Feb20
## 9 Feb20
```

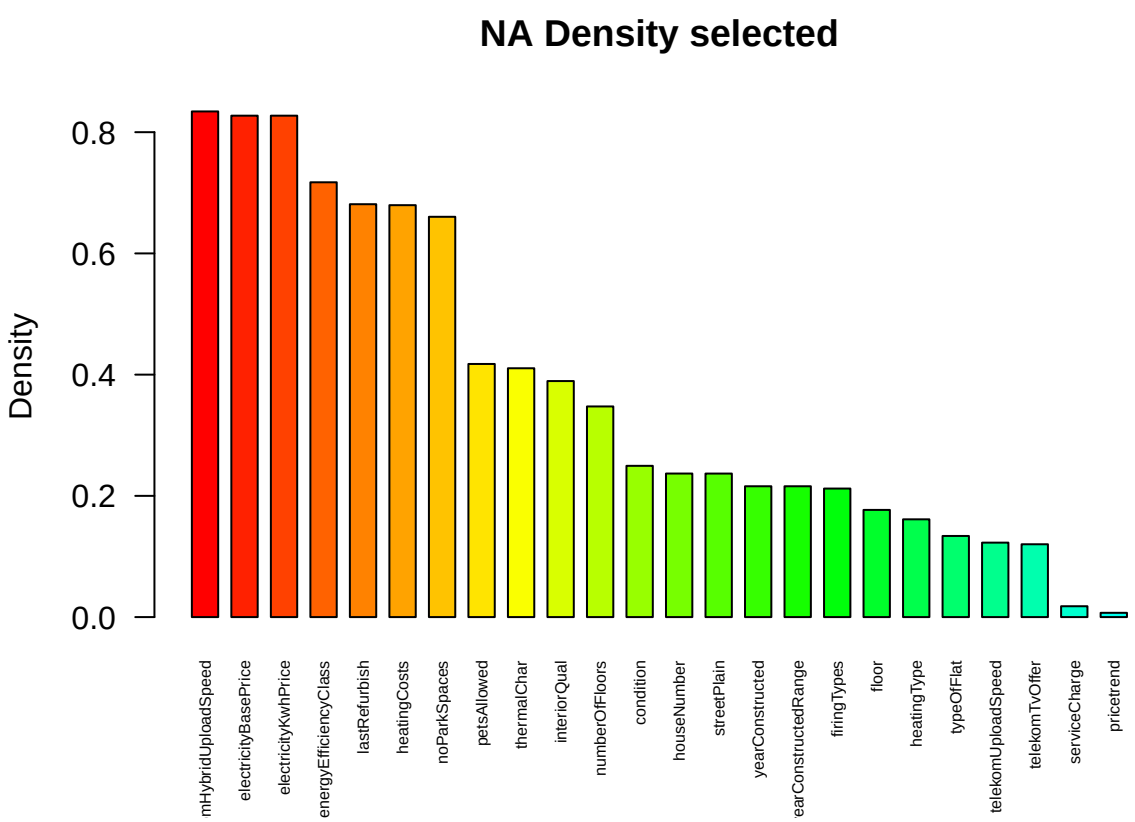
Lets have a look at the NA density of our features.

```
#no_nas = apply(data_raw, MARGIN=2, function(x) {sum(is.na(x))})
no_nas = colSums(is.na(data_raw))
na_density = no_nas / nrow(data_raw)
na_density = na_density[order(na_density, decreasing = TRUE)]
barplot(na_density, main="NA Density all", ylab="Density", space=0.5,
        col = rainbow(length(na_density)), cex.names = 0.5, las=2)
```



It looks like there are some features without any data missing. Lets take a closer look at the more relevant features with more NA values.

```
no_na = which(na_density == 0)
many_na = na_density[-no_na]
barplot(many_na, main="NA Density selected", ylab="Density", space=0.5,
        col = rainbow(length(na_density)), cex.names = 0.5, las=2)
```



This is not looking too great. Many feature have a rather high NA density. As a rule of thumb I will try to impute features with a maximum NA density of 0.3. There are of course methods to impute way higher percentages but I dont want to spend too much time with advanced feature imputation methods.

```
na_remove_cols = which(na_density >= 0.3)
length(na_remove_cols)
```

```
## [1] 11
```

```
na_density[na_remove_cols]
```

```
## telekomHybridUploadSpeed     electricityBasePrice      electricityKwhPrice
##                0.8342202                0.8272041                0.8272041
##    energyEfficiencyClass           lastRefurbish             heatingCosts
##                0.7173295                0.6811455                0.6795645
##             noParkSpaces             petsAllowed             thermalChar
##                0.6603688                0.4176006                0.4105101
##             interiorQual          numberOfFloors
##                0.3893743                0.3475144
```

There are 11 features in question. Before blindly removing them by the arbitrary cutoff of 0.3 lets take a quick look on the features and if we still might want to keep them due to importance.

- telekomHybridUploadSpeed: probably not too important
- electricityBasePrice: sounds important but hard to guess
- electricityKwhPrice: same as for the base price, we could probably find this information online but would involve too much searching
- energyEfficiencyClass: as for the eletricity
- lastRefurbish: this sounds important and should correlate to the price. We might want to keep it even

tho alot of data is missing.

- heatingCosts: important aswell and correlates to the EfficencyClass sadly too hard to guess and averaging would probably introduce too much of an bias
- noParkSpaces: prbably important, even more in non rural areas. We could assume that NA equals to no parkspaces and keep it.
- PetsAllowed: should not be too important especially since small pets are allowed by law anyways
- thermalCHar: directly relates to energyclass! could replace the information lost by droping the other columns and with 0.41 not remotely as bad as the other features.
- interiorQUal: seems important and also not too bad with ~0.39 missing values could just be imputed with "average condition"
- numberofFloors: seems to have an impact aswell so lets keep it.

To conclude it appears as if we should increase the "cutoff" to 0.4 to include most of the features discussed and include the ones above as exception.

```
data_raw = subset(data_raw, select=-c(telekomHybridUploadSpeed,
                                      electricityBasePrice,
                                      energyEfficiencyClass,
                                      heatingCosts, petsAllowed))
head(data_raw)
```

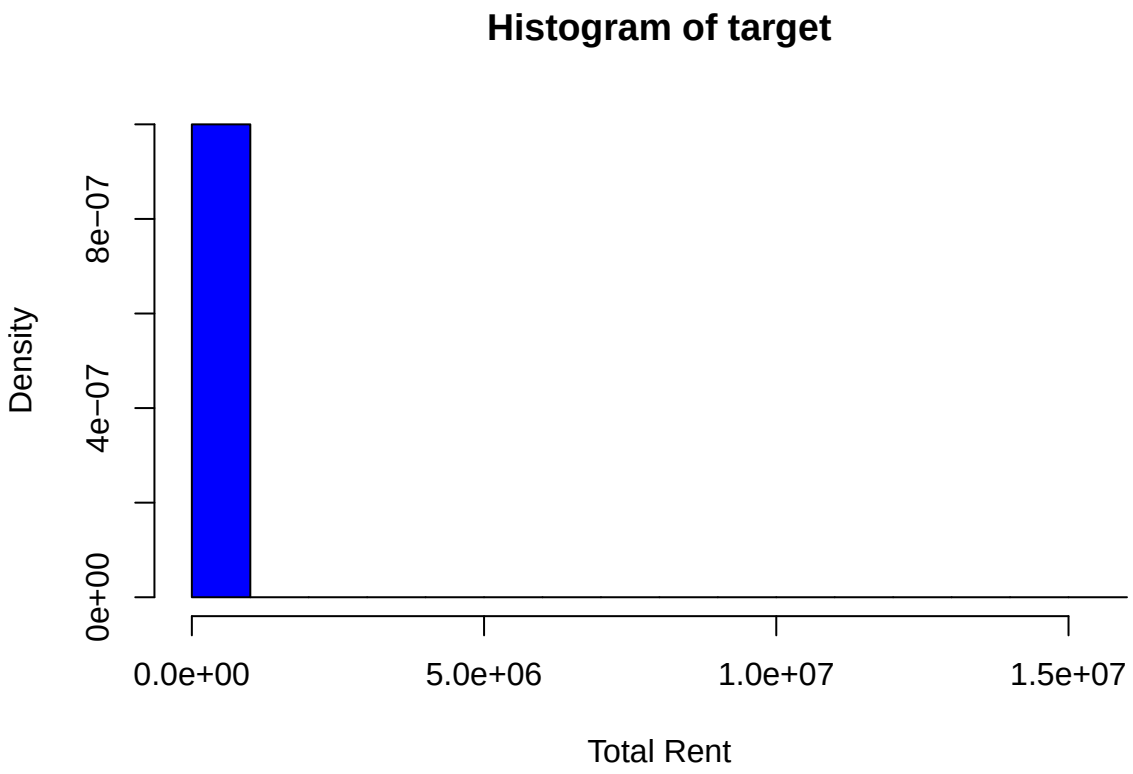
```
##                regio1 serviceCharge                   heatingType
## 1 Nordrhein_Westfalen           245                central_heating
## 3             Sachsen           255                  floor_heating
## 5              Bremen           138 self_contained_central_heating
## 7             Sachsen            70 self_contained_central_heating
## 8              Bremen            88                central_heating
## 9  Baden_WÃ¼rttemberg           110                    oil_heating
##   telekomTvOffer newlyConst balcony picturecount pricetrend telekomUploadSpeed
## 1  ONE_YEAR_FREE      FALSE   FALSE            6       4.62                10.0
## 3  ONE_YEAR_FREE       TRUE    TRUE            8       2.72                 2.4
## 5           <NA>      FALSE    TRUE           19       2.46                  NA
## 7  ONE_YEAR_FREE      FALSE   FALSE            9       1.01                 2.4
## 8  ONE_YEAR_FREE      FALSE    TRUE            5       1.89                 2.4
## 9  ONE_YEAR_FREE      FALSE   FALSE            5       3.77                40.0
##   totalRent yearConstructed   scoutId noParkSpaces     firingTypes hasKitchen
## 1    840.00            1965  96107057            1             oil      FALSE
## 3   1300.00            2019 113147523            1            <NA>      FALSE
## 5    903.00            1950 114751222           NA             gas      FALSE
## 7    380.00              NA 114391930           NA            <NA>      FALSE
## 8    584.25            1959 115270775           NA gas:electricity      FALSE
## 9    690.00            1970 106416361            1             oil       TRUE
##               geo_bln cellar yearConstructedRange baseRent houseNumber
## 1 Nordrhein_Westfalen   TRUE                    2   595.00         244
## 3             Sachsen   TRUE                    9   965.00           4
## 5              Bremen  FALSE                    1   765.00          10
## 7             Sachsen   TRUE                   NA   310.00          14
## 8              Bremen   TRUE                    2   452.25          35
## 9  Baden_WÃ¼rttemberg   TRUE                    2   580.00        <NA>
##   livingSpace             geo_krs       condition  interiorQual
## 1       86.00            Dortmund       well_kept        normal
## 3       83.80             Dresden  first_time_use sophisticated
## 5       84.97              Bremen     refurbished          <NA>
## 7       62.00 Mittelsachsen_Kreis fully_renovated          <NA>
## 8       60.30              Bremen            <NA>          <NA>
```

```
## 9       53.00   Emmendingen_Kreis       well_kept sophisticated
##                         street                streetPlain  lift baseRentRange
## 1  Sch&uuml;ruferstra&szlig;e           SchÃ¼ruferstraÃŸe FALSE             4
## 3                   Turnerweg                   Turnerweg  TRUE             6
## 5 Hermann-Henrich-Meier-Allee Hermann-Henrich-Meier-Allee FALSE             5
## 7                  Am Bahnhof                  Am_Bahnhof FALSE             2
## 8            Lesumer Heerstr.            Lesumer_Heerstr. FALSE             3
## 9              no_information                        <NA> FALSE             4
##     typeOfFlat geo_plz noRooms thermalChar floor numberOfFloors noRoomsRange
## 1 ground_floor   44269       4       181.4     1              3            4
## 3    apartment    1097       3          NA     3              4            3
## 5    apartment   28213       3       188.9     1             NA            3
## 7         <NA>    9599       2          NA     1              4            2
## 8 ground_floor   28717       3        63.0    NA             NA            3
## 9  roof_storey   79211       2       138.0     2              2            2
##   garden livingSpaceRange             regio2                      regio3
## 1   TRUE                4            Dortmund                    SchÃ¼ren
## 3  FALSE                4             Dresden Ã„uÃŸere_Neustadt_Antonstadt
## 5  FALSE                4              Bremen            Neu_Schwachhausen
## 7   TRUE                3 Mittelsachsen_Kreis                     Freiberg
## 8  FALSE                2              Bremen                   St._Magnus
## 9  FALSE                2   Emmendingen_Kreis                    Denzlingen
##   lastRefurbish electricityKwhPrice  date
## 1            NA                  NA May19
## 3            NA                  NA Oct19
## 5            NA                  NA Feb20
## 7            NA                  NA Feb20
## 8            NA                  NA Feb20
## 9            NA                  NA Feb20
```

Feature and Target Analysis

I now will go over all columns to gain more insight into the data. I will start with the target "totalRent".

```
target = data_raw$totalRent
hist(target, freq=FALSE, col="blue", xlab="Total Rent")
```



Apparently we have some extreme outliers lets check them out.

```
target = sort(target, decreasing=TRUE)
head(target, 10)
```

```
##  [1] 15751535  1234567  1150900  1000000   485350   108000    64651    63204
##  [9]    51570    37600
```

```
tail(target, 10)
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0
```

Lets note: we have some extreme high rents which seem like errors in the data (1234567) but also some really low values or even 0. In the next step I will find out with how many outliers we are dealing with.

```
high_rents = which(target > 10000)
low_rents = which(target < 200)
length(high_rents)
```

```
## [1] 34
```

```
length(low_rents)
```

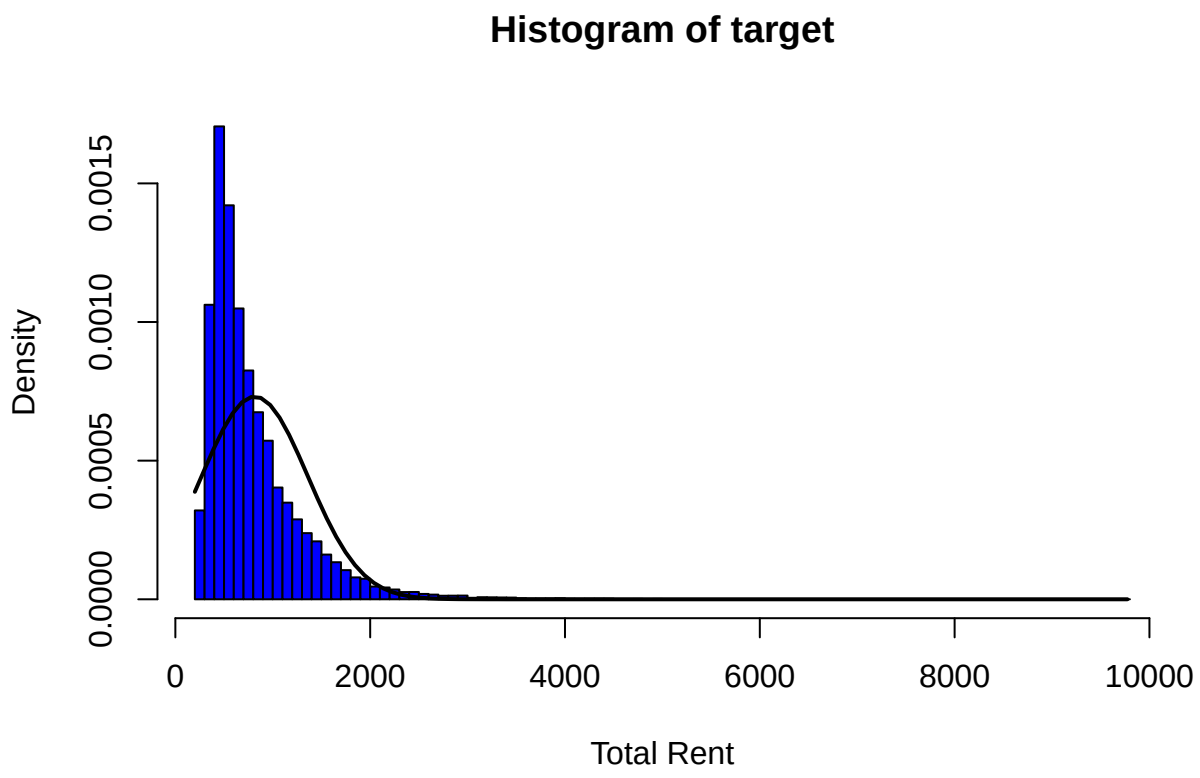
```
## [1] 670
```

Around 700. Not too bad we can just leave them out without reducing our data set by too much. I decided that monthly rents above 20.000e seem unreasonable and threat them as outliers.

```
remove_ind = c(high_rents, low_rents)
target = target[-remove_ind]
data_raw = data_raw[-remove_ind, ]
```

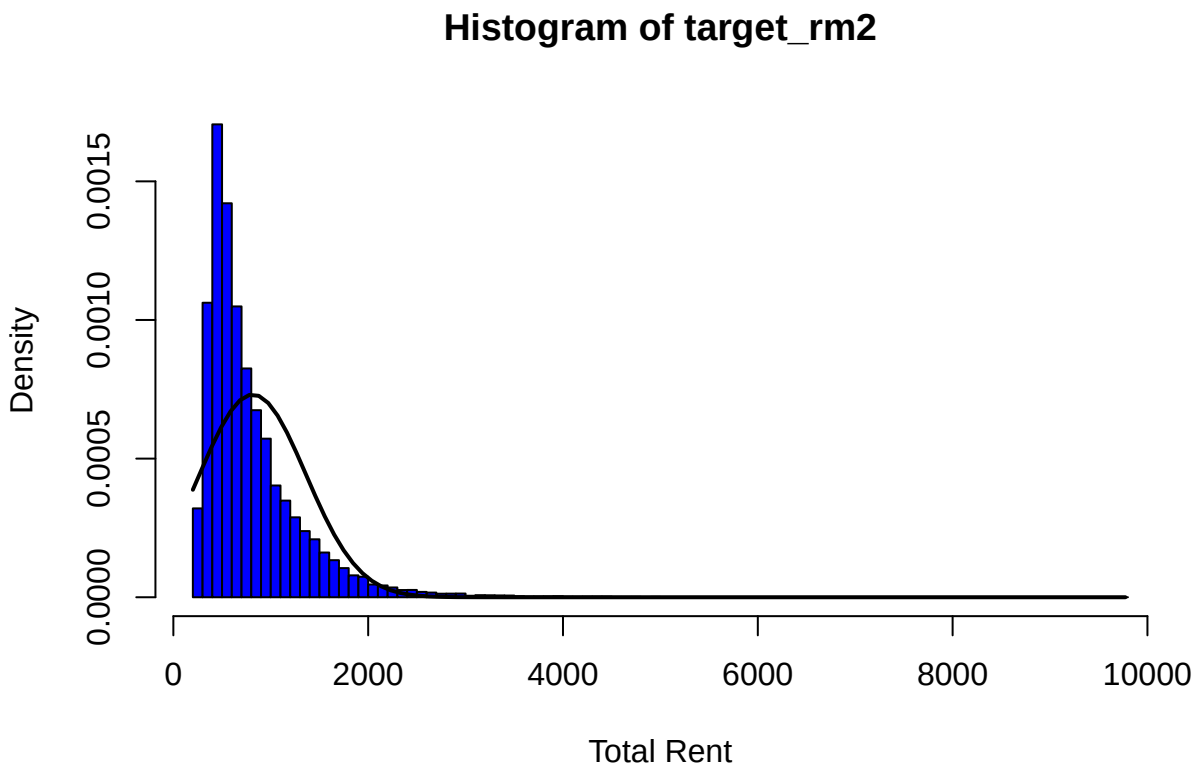
Lets check out the histogram again!

```
hist(target, freq=FALSE, col="blue", xlab="Total Rent", breaks=100)
xdist = seq(min(target), max(target), length=100)
ydist = dnorm(xdist, mean(target), sd(target))
lines(xdist, ydist, col="black", lwd=2)
```

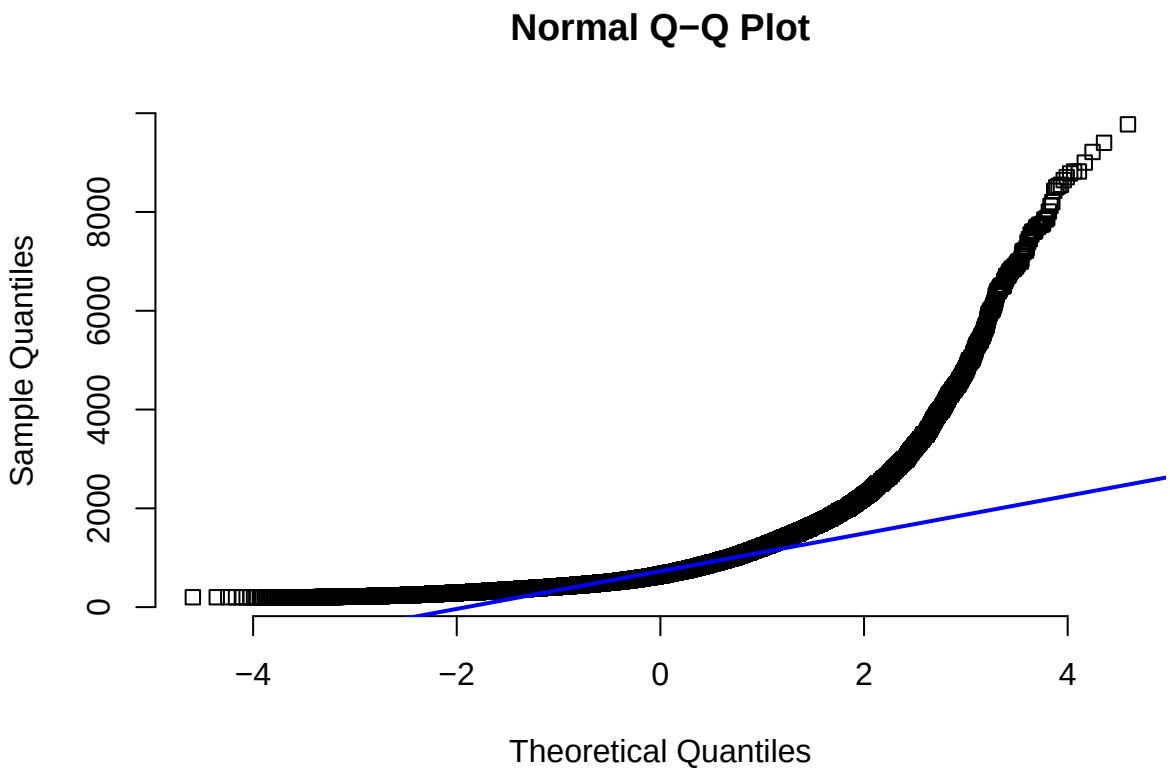


Still not good enough to get an idea of how our target distribution looks like. Lets cut this down further.

```
remove_ind_2 = which(target > 10000)
target_rm2 = target
hist(target_rm2, freq=FALSE, col="blue", xlab="Total Rent", breaks=100)
xdist = seq(min(target_rm2), max(target_rm2), length=100)
ydist = dnorm(xdist, mean(target_rm2), sd(target_rm2))
lines(xdist, ydist, col="black", lwd=2)
```




```
qqnorm(target_rm2, pch=0.05, frame=FALSE)
qqline(target_rm2, col="blue", lwd=2)
```



The distribution appears to be left skewed with a strong right tail. Lets get some statistics.

```
skew = skewness(target_rm2)
kur = kurtosis(target_rm2)
skew
```

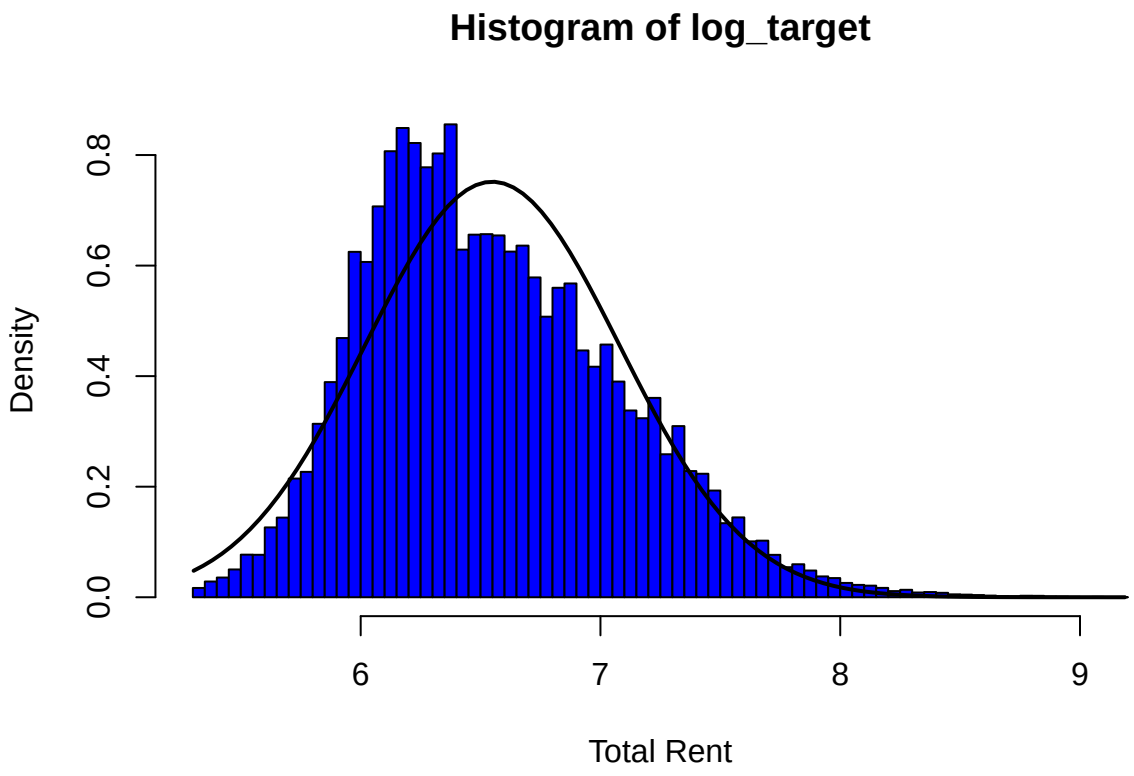
```
## [1] 3.159866
```

```
kur
```

```
## [1] 22.67893
```

Lets take a look on a logarithmic scale

```
log_target = log1p(target_rm2)
hist(log_target, freq=FALSE, col="blue", xlab="Total Rent", breaks=100)
xdist = seq(min(log_target), max(log_target), length=100)
ydist = dnorm(xdist, mean(log_target), sd(log_target))
lines(xdist, ydist, col="black", lwd=2)
```



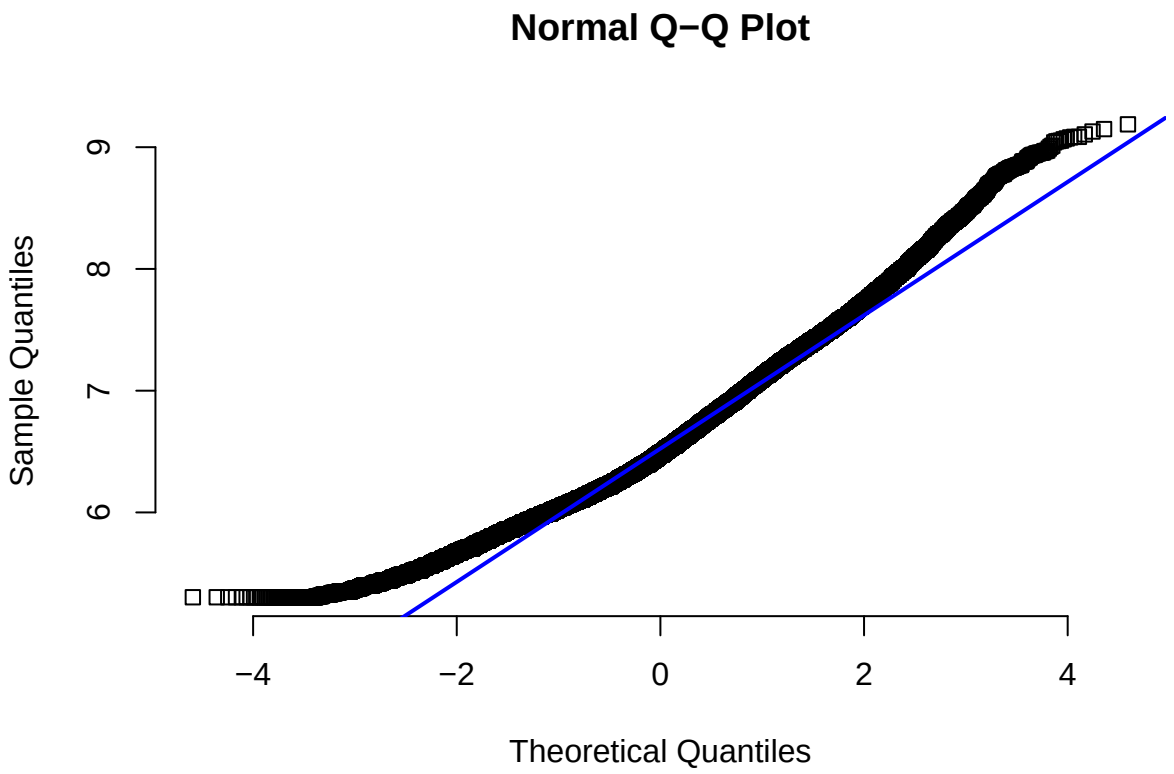

```
skewness(log_target)
```

```
## [1] 0.5620409
```

```
kurtosis(log_target)
```

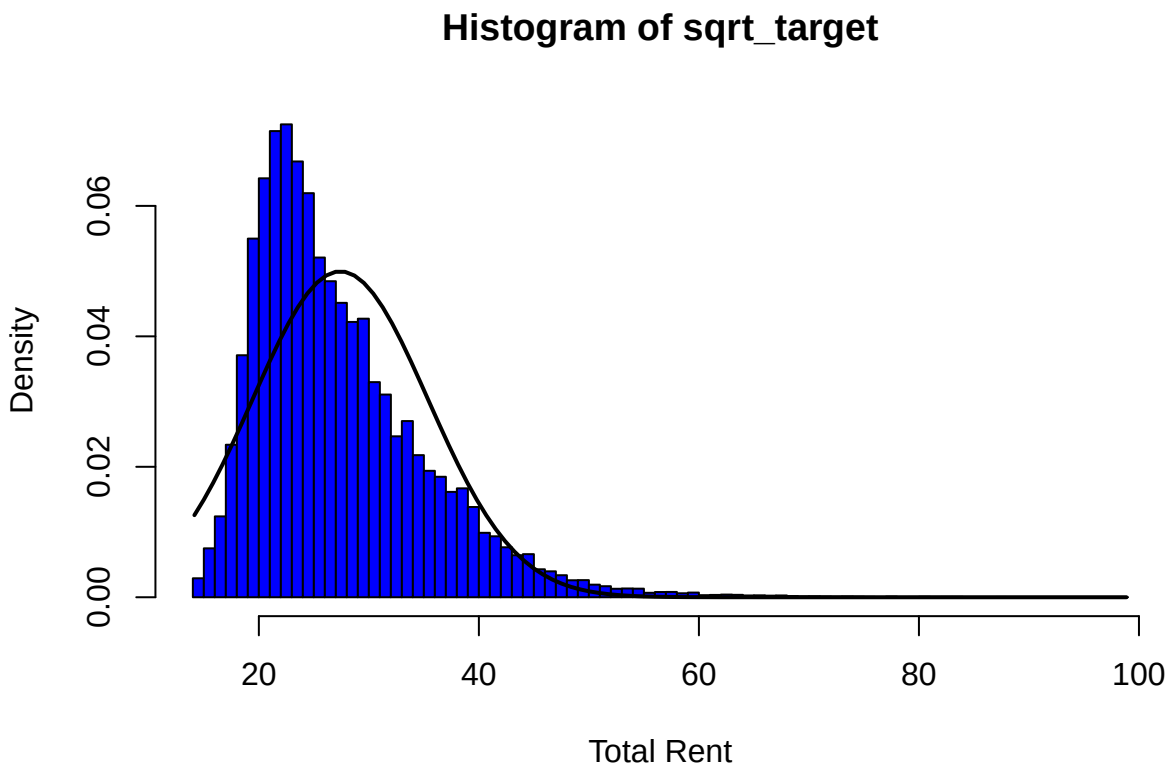
```
## [1] 3.180503
```

```
qqnorm(log_target, pch=0.05, frame=FALSE)
qqline(log_target, col="blue", lwd=2)
```



Still not perfect but looks way better! Lets see if sqrt transform does any better.

```
sqrt_target = sqrt(target_rm2)
hist(sqrt_target, freq=FALSE, col="blue", xlab="Total Rent", breaks=100)
xdist = seq(min(sqrt_target), max(sqrt_target), length=100)
ydist = dnorm(xdist, mean(sqrt_target), sd(sqrt_target))
lines(xdist, ydist, col="black", lwd=2)
```




```
skewness(sqrt_target)
```

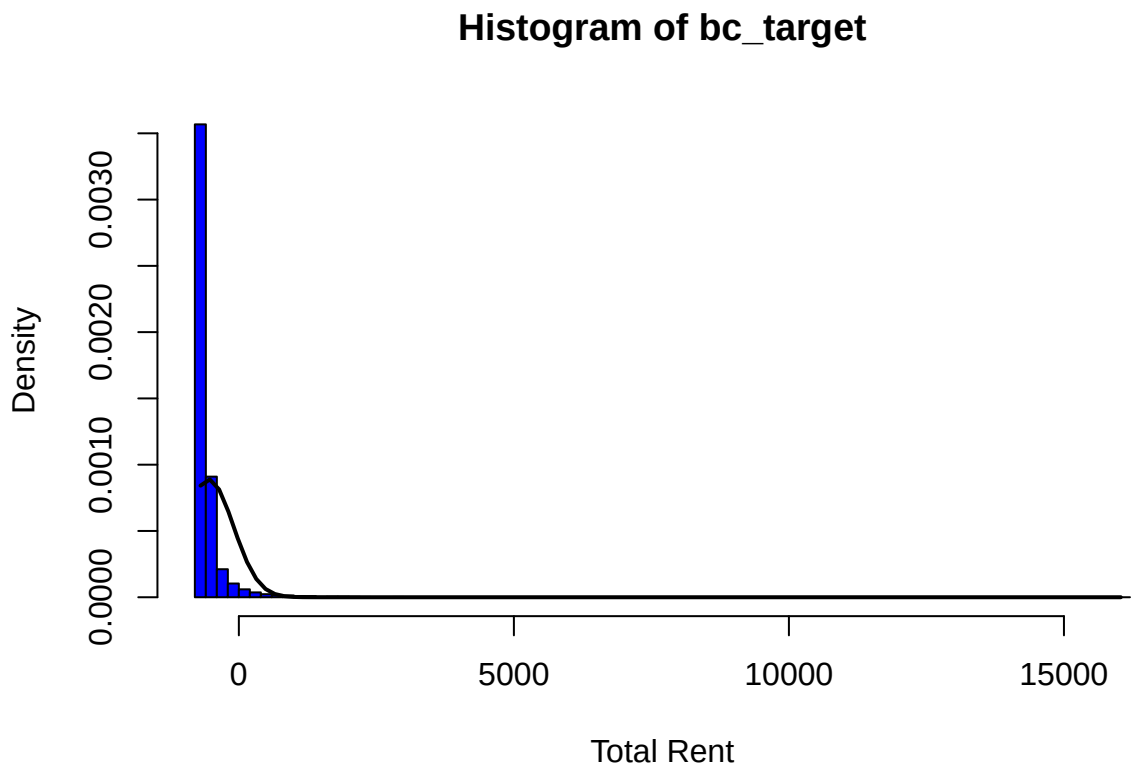
```
## [1] 1.479911
```

```
kurtosis(sqrt_target)
```

```
## [1] 6.812695
```

Nah this looks worse! Lets bring the big guns and use a Box-Cox transform

```
bc_target = boxCoxVariable(target_rm2)
hist(bc_target, freq=FALSE, col="blue", xlab="Total Rent", breaks=100)
xdist = seq(min(bc_target), max(bc_target), length=100)
ydist = dnorm(xdist, mean(bc_target), sd(bc_target))
lines(xdist, ydist, col="black", lwd=2)
```




```
skewness(bc_target)
```

```
## [1] 11.50415
```

```
kurtosis(bc_target)
```

```
## [1] 213.3626
```

this made it worse - how disappointing! So the best bet is the log1p transformation for now.

```
data_raw$totalRent = log1p(data_raw$totalRent)
data_raw = data_raw[-remove_ind, ]
```

Lets now get over the features one by one, transform possibly categorical variables and fill in the NA's

```
names(data_raw)
```

```
##  [1] "regio1"              "serviceCharge"       "heatingType"
##  [4] "telekomTvOffer"      "newlyConst"          "balcony"
##  [7] "picturecount"        "pricetrend"          "telekomUploadSpeed"
## [10] "totalRent"           "yearConstructed"     "scoutId"
## [13] "noParkSpaces"        "firingTypes"         "hasKitchen"
## [16] "geo_bln"             "cellar"              "yearConstructedRange"
## [19] "baseRent"            "houseNumber"         "livingSpace"
## [22] "geo_krs"             "condition"           "interiorQual"
## [25] "street"              "streetPlain"         "lift"
## [28] "baseRentRange"       "typeOfFlat"          "geo_plz"
## [31] "noRooms"             "thermalChar"         "floor"
## [34] "numberOfFloors"      "noRoomsRange"        "garden"
```

```
## [37] "livingSpaceRange"     "regio2"               "regio3"
## [40] "lastRefurbish"        "electricityKwhPrice"  "date"
```

```
no_nas_updated = colSums(is.na(data_raw))
row_nas = rowSums(is.na(data_raw))
```

regio1

```
temp = data_raw$regio1
head(temp, 10)
```

```
##  [1] "Nordrhein_Westfalen" "ThÃ¼ringen"          "Sachsen_Anhalt"
##  [4] "Baden_WÃ¼rttemberg"  "Berlin"              "Berlin"
##  [7] "Sachsen_Anhalt"      "Sachsen"             "Sachsen"
## [10] "Sachsen"
```

regio1 appears to give information about the "Bundesland" - We can simply encode these.

```
unique(temp)
```

```
##  [1] "Nordrhein_Westfalen"    "ThÃ¼ringen"             "Sachsen_Anhalt"
##  [4] "Baden_WÃ¼rttemberg"     "Berlin"                "Sachsen"
##  [7] "Mecklenburg_Vorpommern" "Bayern"                "Brandenburg"
## [10] "Hessen"                 "Niedersachsen"         "Rheinland_Pfalz"
## [13] "Hamburg"                "Schleswig_Holstein"    "Bremen"
## [16] "Saarland"
```

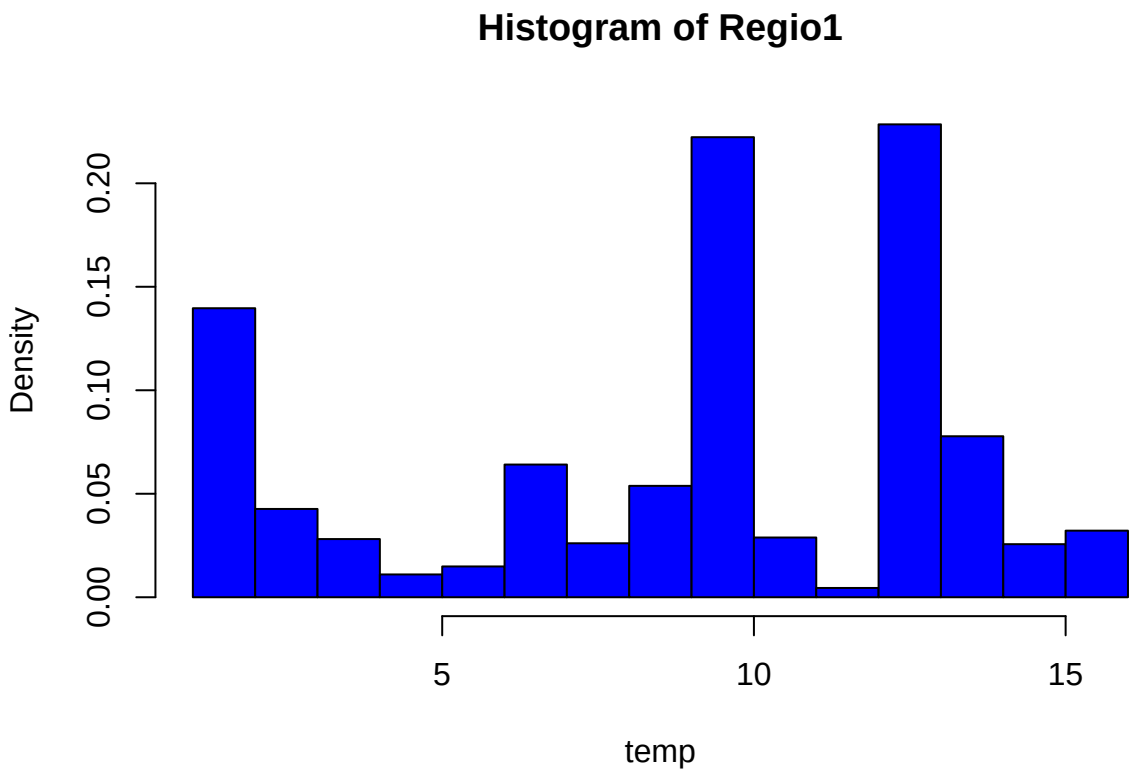
Good news! No NA's we have to deal with.

```
which(is.na(temp))
```

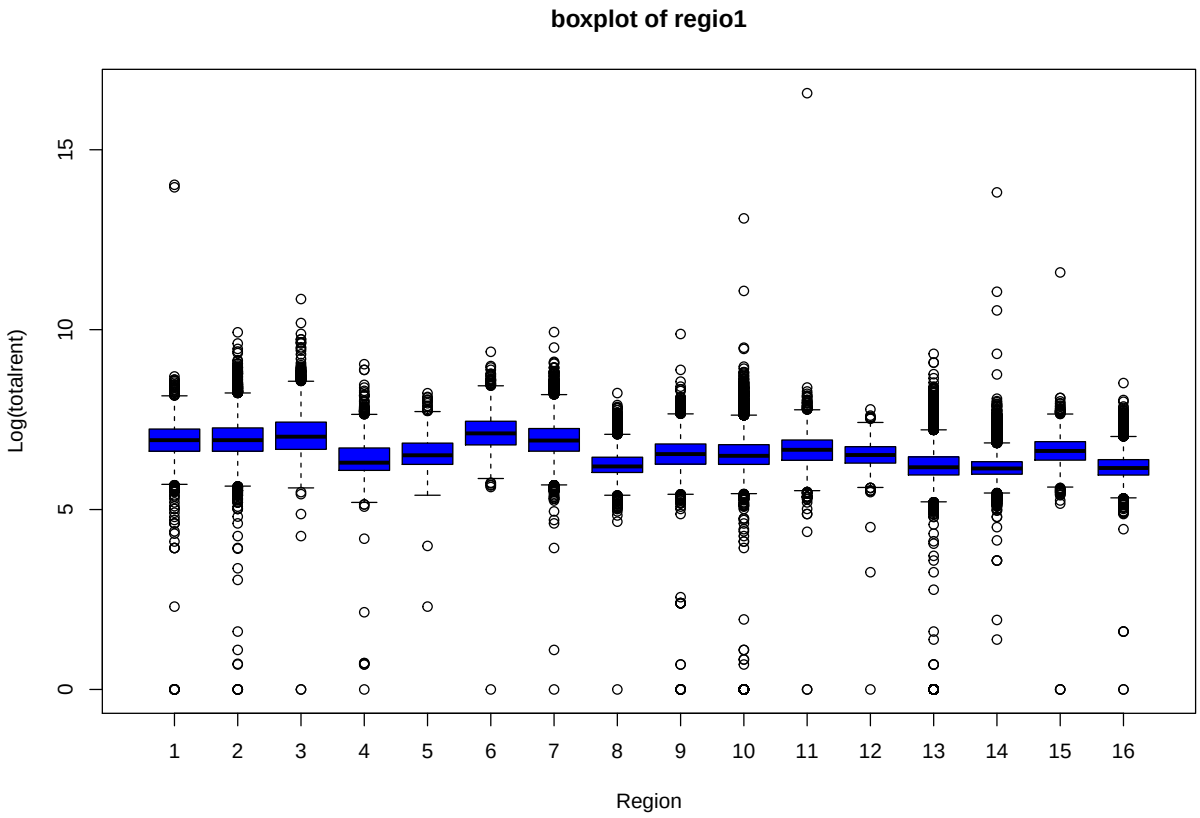
```
## integer(0)
```

lets convert to our dummy variable

```
temp = as.numeric(as.character(factor(temp, labels = 1:length(unique(temp)))))

hist(temp, col="blue", main="Histogram of Regio1", breaks=length(unique(temp))
     , freq = FALSE)
```




```
boxplot(data_raw$totalRent~temp, col="blue", stroke="red",
        main="boxplot of regio1", xlab = "Region", ylab="Log(totalrent)")
```




```
data_raw$regio1 = temp
```

regio2

```
temp = data_raw$regio2
head(temp, 10)
```

```
##  [1] "DÃ¼sseldorf"      "Gera"              "Magdeburg"         "BÃ¶blingen_Kreis"
##  [5] "Berlin"           "Berlin"            "Magdeburg"         "Dresden"
##  [9] "Chemnitz"         "Leipzig"
```

regio2 seems to narrow the geolocation further down to the city level.

```
length(unique(temp))
```

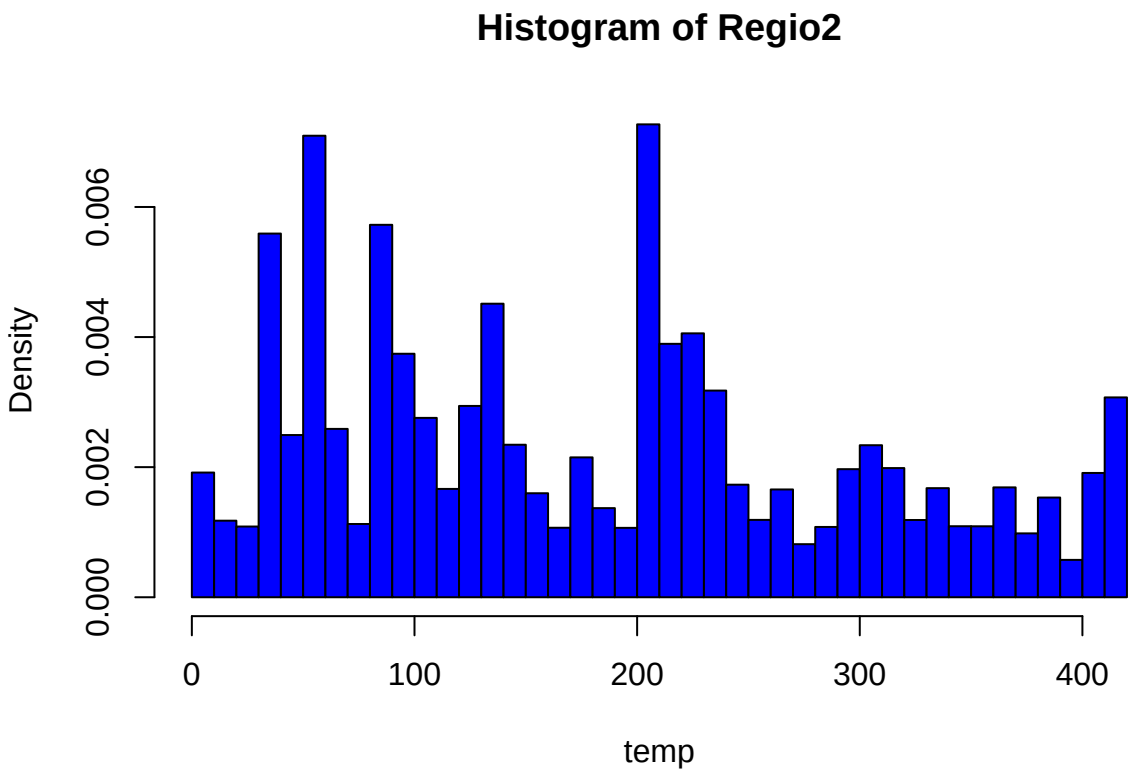
```
## [1] 419
```

```
which(is.na(temp))
```

```
## integer(0)
```

great no NA's here again.

```
temp = as.numeric(as.character(factor(temp, labels = 1:length(unique(temp)))))

hist(temp, col="blue", main="Histogram of Regio2", breaks=50, freq = FALSE)
```




```
data_raw$regio2 = temp
```

regio3

```
temp = data_raw$regio3
head(temp, 10)
```

```
##  [1] "Rath"                  "Stadtmitte"            "Leipziger_Str."
##  [4] "BÃ¶blingen"             "Kreuzberg_Kreuzberg"   "Tiergarten_Tiergarten"
##  [7] "Werder"                "Leubnitz_Neuostra"     "Zentrum"
## [10] "Reudnitz_Thonberg"
```

regio2 seems to narrow the geolocation further down to the precise location.

```
length(unique(temp))
```

```
## [1] 8334
```

```
which(is.na(temp))
```

```
## integer(0)
```

Due to the size of the data set I dont see that drilling down to this level will be beneficial since there are not too many points per location to go by. Thus we drop this column.

```
data_raw = subset(data_raw, selec=-c(regio3))
```

servicecharge

```
temp = data_raw$serviceCharge
head(temp, 10)
```

```
##  [1] 293.00 250.00 115.00 295.00 228.78 147.08 193.00 169.00 200.00  50.00
```

These seem to be the side-costs for maintenence and service.

```
length(which(is.na(temp)))
```

```
## [1] 4081
```

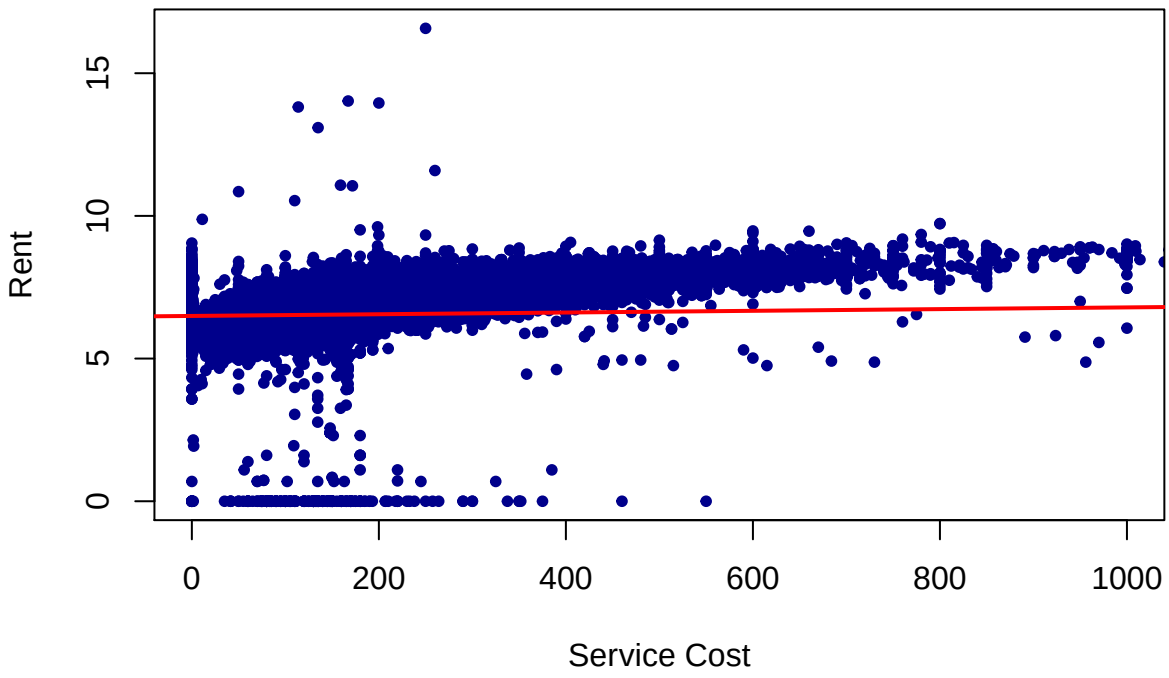
It appears as if we still have some NA's here. I will impute the average over each regio1 for the missing values.

```
averages = aggregate(temp, list(data_raw$regio1), mean, na.rm=TRUE)



for (i in which(is.na(temp))){
  for (j in 1:length(unique(data_raw$regio1))){
    if (data_raw$regio1[i] == j){
      temp[i] = averages[j, 2]
    }
  }
}

sum(is.na(temp))
```

```
## [1] 0
```

Now that we filled missing values lets plot them vs the rent

```
plot(temp, data_raw$totalRent, type = "p", xlim = c(0, 1000),
     pch=20, col="darkblue", xlab="Service Cost", ylab="Rent")
temp_model = lm(data_raw$totalRent~temp)
abline(temp_model, col="red", lwd=2)
```




```
summary(temp_model)
```

```
##
## Call:
## lm(formula = data_raw$totalRent ~ temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.511  -0.375  -0.060   0.345  10.004
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.494e+00  1.324e-03 4905.71   <2e-16 ***
## temp        2.990e-04  3.702e-06   80.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5719 on 227593 degrees of freedom
## Multiple R-squared:  0.02788,    Adjusted R-squared:  0.02787
## F-statistic:  6527 on 1 and 227593 DF,  p-value: < 2.2e-16
```

```
data_raw$serviceCharge = temp
```

heatingtype

```
temp = data_raw$heatingType
head(temp, 10)
```

```
##  [1] "floor_heating"    "central_heating"  "central_heating"  "floor_heating"
##  [5] "floor_heating"    "district_heating" "floor_heating"    "central_heating"
##  [9] "gas_heating"      "district_heating"
```

heatingType gives information about the heating -duh. Another Categorical Variable!

```
unique(temp)
```

```
##  [1] "floor_heating"                  "central_heating"
##  [3] "district_heating"               "gas_heating"
##  [5] NA                               "self_contained_central_heating"
##  [7] "combined_heat_and_power_plant"  "heat_pump"
##  [9] "oil_heating"                    "wood_pellet_heating"
## [11] "electric_heating"               "night_storage_heater"
## [13] "stove_heating"                  "solar_heating"
```

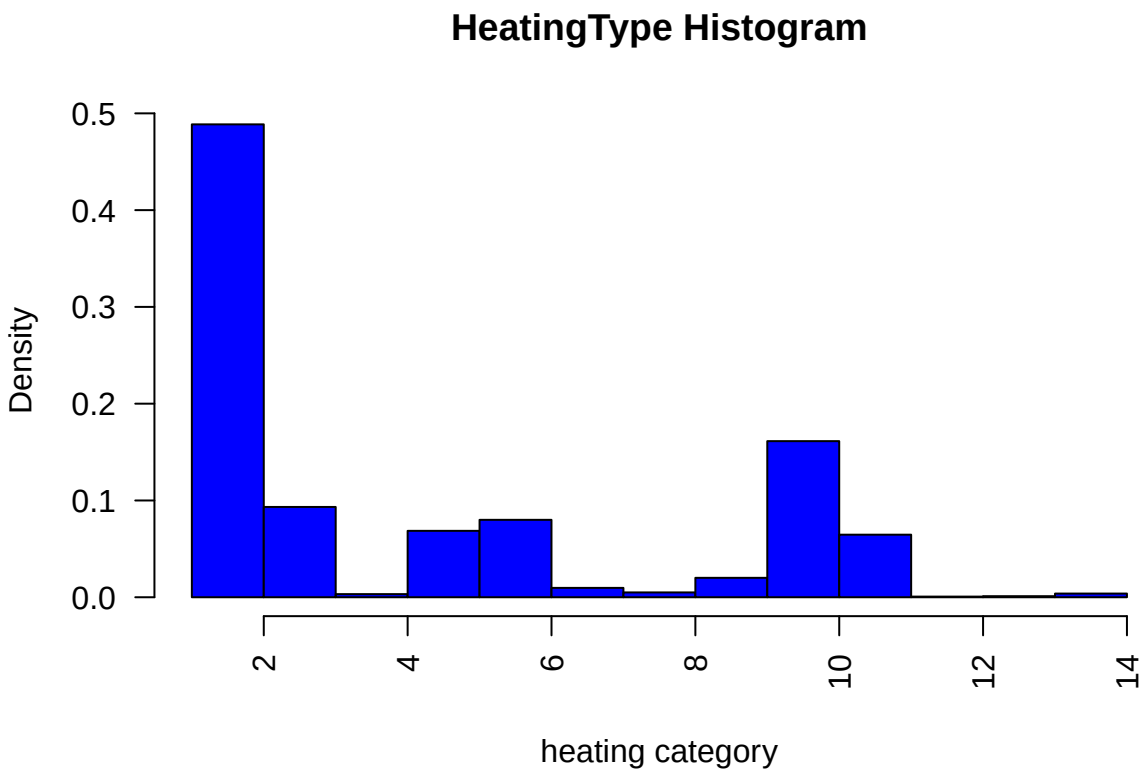
```
sum(is.na(temp))
```

```
## [1] 36714
```

Since we dont know the heating we will introduce a new category "other" for the NA's. Another possibility it to use the most common heating.

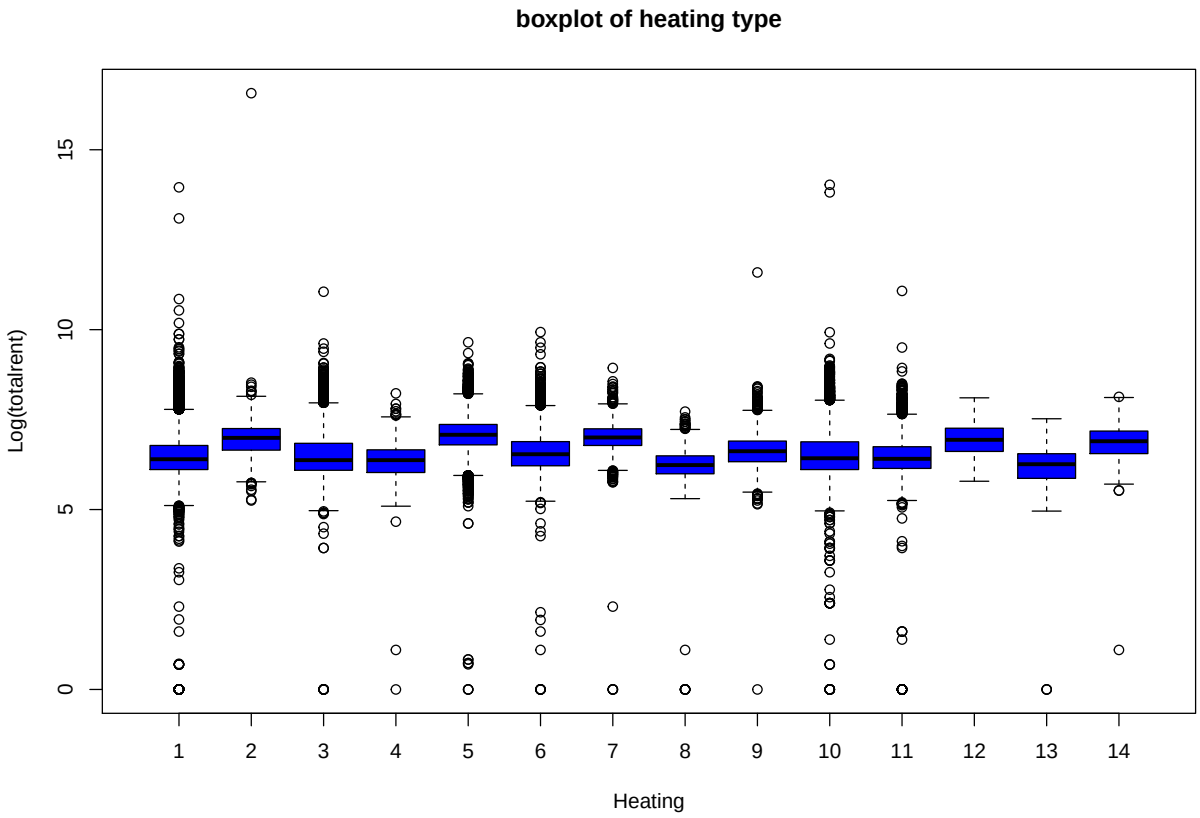
```
temp[is.na(temp)] = "other"
temp = as.numeric(as.character(factor(temp, labels = 1:length(unique(temp)))))
```

```
hist(temp, las=2, freq = FALSE, col="blue", xlab="heating category",
     main="HeatingType Histogram", breaks = length(unique(temp)))
```



We see that most buildings fall into the heating Category 1. Instead of introducing other we could have just assumed the NA's to be of the same type or randomly draw from the distribution. To improve performance we might come back to this later.

```
boxplot(data_raw$totalRent~temp, col="blue", stroke="red",
        main="boxplot of heating type", xlab = "Heating", ylab="Log(totalrent)")
```




```
data_raw$heatingType = temp
```

telekomTvOffer

```
temp = data_raw$telekomTvOffer
head(temp, 10)
```

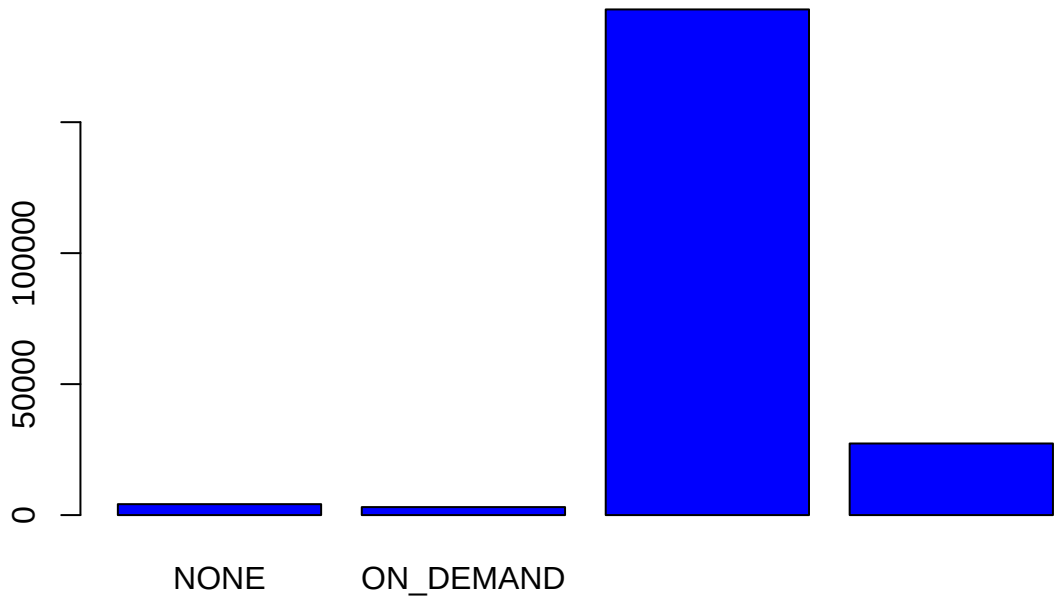
```
##  [1] NA               "ONE_YEAR_FREE" "ONE_YEAR_FREE" NA
##  [5] NA               "ONE_YEAR_FREE" "ONE_YEAR_FREE" "ONE_YEAR_FREE"
##  [9] "ONE_YEAR_FREE" "ONE_YEAR_FREE"
```

```
unique(temp)
```

```
## [1] NA               "ONE_YEAR_FREE" "NONE"          "ON_DEMAND"
```

This features tells us about the telekom TV offers for the flat. it apparently has 4 possible values including NA. Lets look at the histogram fist before deciding what to do with NA's.

```
plot(factor(temp, exclude=NULL), col="blue")
```



We see that almost all buildings get the one year free option. This feature will probably not be too interesting but lets add the NA's to this category for now and decide later if we are going to use it or not.

```
temp[is.na(temp)] = "ONE_YEAR_FREE"
temp = as.numeric(as.character(factor(temp, labels = 1:length(unique(temp)))))
data_raw$telekomTvOffer = temp
```

newlyConnst

```
temp = data_raw$newlyConst
head(temp, 10)
```

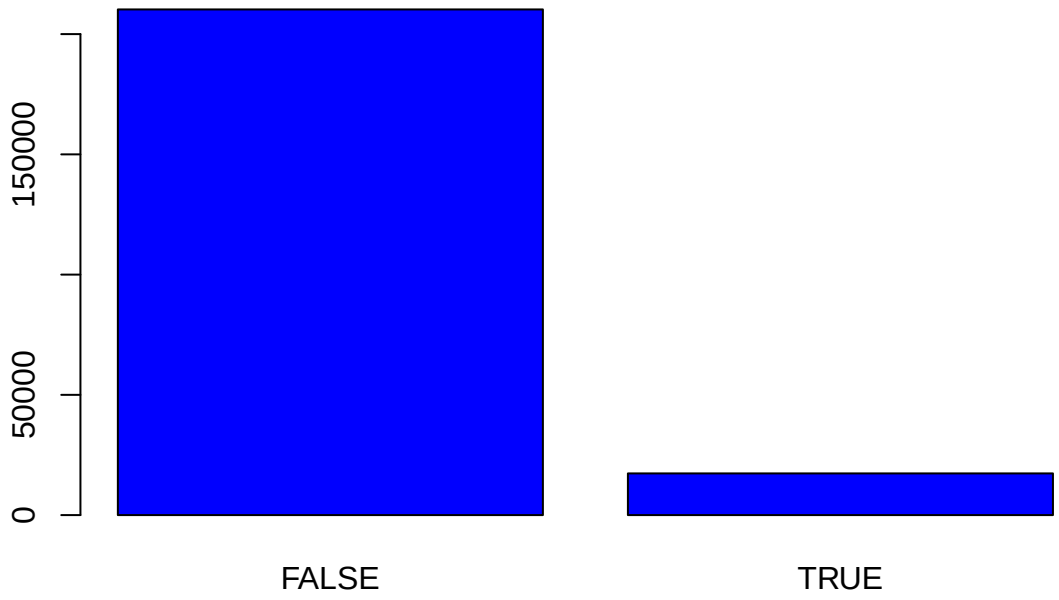
```
##  [1]  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
```

```
unique(temp)
```

```
## [1]  TRUE FALSE
```

This is great we do not have any NA's and this appears to be a bloolean Variable telling us if a building was newly constructed or not. Lets check the weight of the two categories.

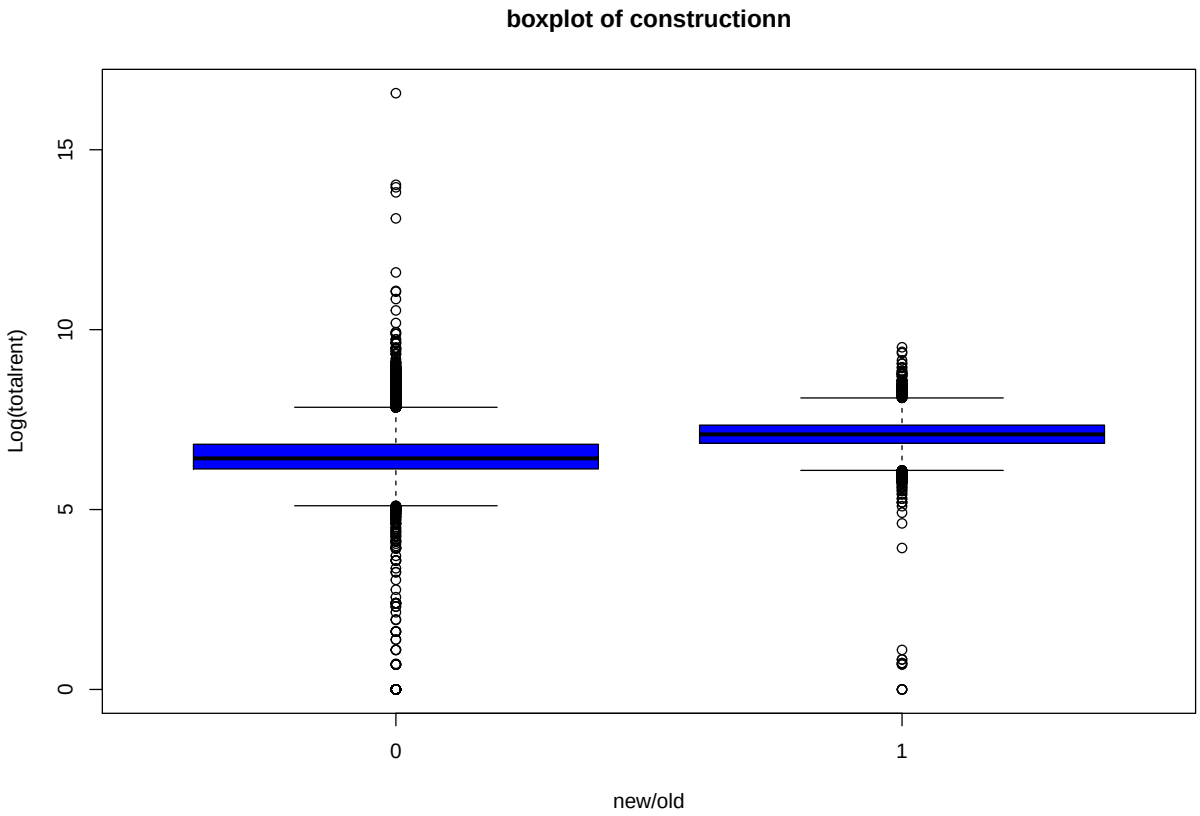
```
plot(as.factor(temp), col="blue")
```



Apparently most buildings are not new, as was to be expected. I am just going to enncode this as True = 1 False = 0

```
temp = ifelse(temp == FALSE, 0, 1)
```

```
boxplot(data_raw$totalRent~temp, col="blue", stroke="red",
        main="boxplot of constructionn", xlab = "new/old",
        ylab="Log(totalrent)")
```



The trend is that new flats rent for more then old ones. Not too surprising either.

```
data_raw$newlyConst = temp
```

balcony

```
temp = data_raw$balcony
head(temp, 10)
```

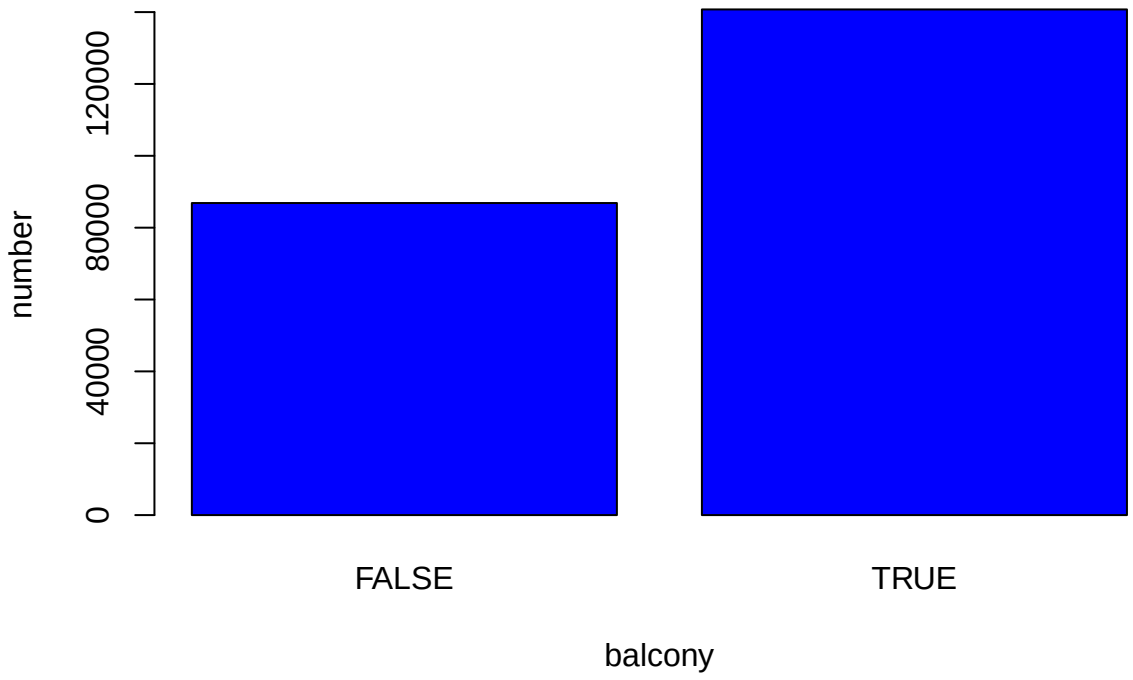
```
##  [1]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
```

```
unique(temp)
```

```
## [1]  TRUE FALSE
```

This feature is rather self explanatory. Boolean variable if a flat has a balcony or not.

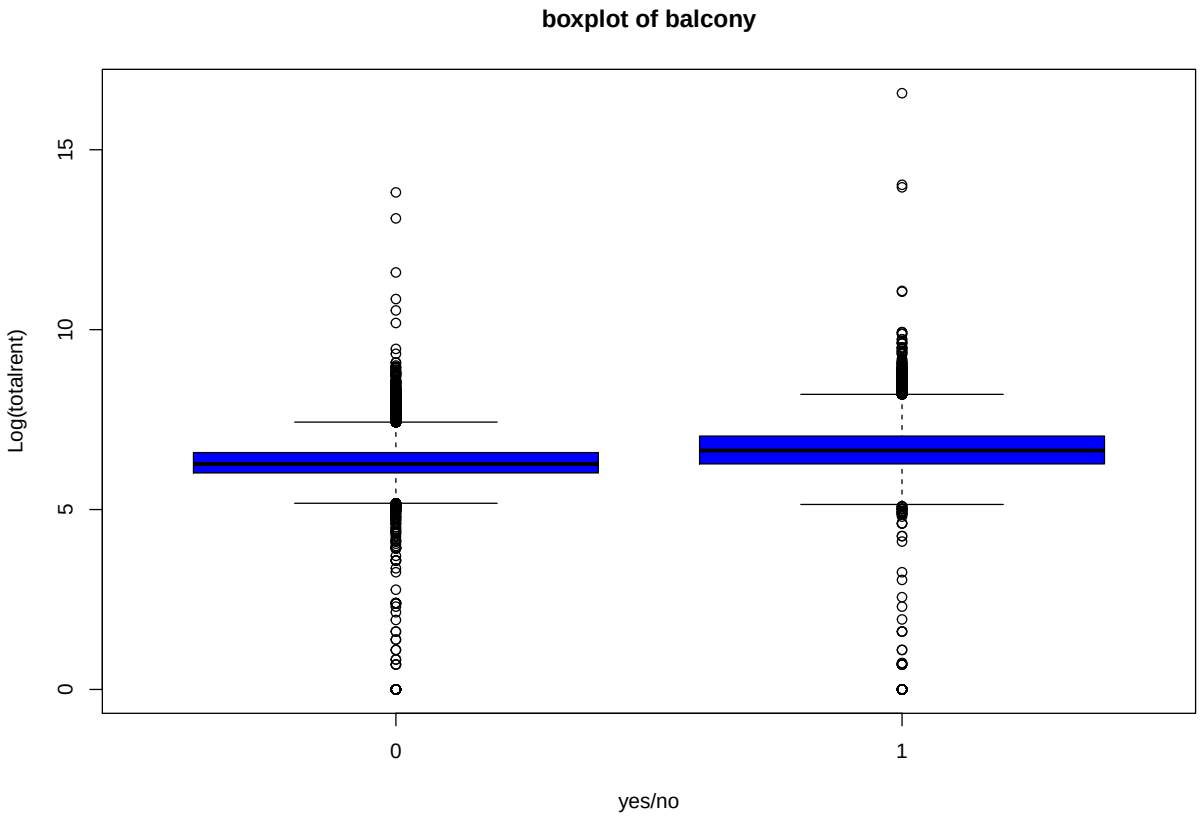
```
plot(as.factor(temp), col="blue", xlab="balcony", ylab="number")
```



This is surprising! Most flats actually do have a balcony. Lets encode in the same fashion.

```
temp = ifelse(temp == FALSE, 0, 1)
```

```
boxplot(data_raw$totalRent~temp, col="blue", stroke="red",
        main="boxplot of balcony", xlab = "yes/no", ylab="Log(totalrent)")
```



This looks promising. Price for flats with a balcony are slightly higher in the log presentation as flats without.

```
data_raw$balcony = temp
```