Austin

may use: price, availabilty, distance, number of reviews, minimum nights

DISTANCE: calculate distance in meter calculated to km, its measured from the downtown of Austin

Cleaning:

drop neighbourhood column due large number of missing values

drop prices that are over 100 per night (calculated by dividing price by minimum nights)

drop where the availability is lower than 1 day and more than 365

look after:

some cases, the minimum nights for rent more than 365 days -> left or price per nights price analysis

**why use these variables:**

distance:

avilability:

number of revies:

good neighbourhood:

dummys:

minimum nights:

classifying:

set availability to 200 beacuse on visualization there was a turning point there. also, the mean and median are near as well which indicates a regular availability of apartments

set distance over 15 because there is also a turning point in prices in visualization. however, the 3rd quatiles is appr. 9, which indicates that less than 25% of the variables will be 0.

set number of reviews is set to 100 because it starts to stag somewhere there

good neighbourhood defined by 7740 as datavisualization suggested, there is a significant decrease in prices. it can be also explained by the distance

For analyis:

t1 -> model table for cross validation and models -> saved in test\_table (in folder\_modellev)

RMSE:

> model3\_level\_holdout\_rmse

[1] 30

> model3\_level\_work\_rmse

[1] 28.3

>

> model4\_level\_holdout\_rmse

[1] 30

> model4\_level\_work\_rmse

[1] 28.2

>

Random forest:

foresttest.tex

foresttest2.tex

c("available\_morethan\_30", 'reviews\_morethan\_100', "distance\_morethan\_15", "good\_neighbourhood") as:

available\_morethan\_30 – available

should do:

use of variables:

ln\_price not suggested without long right take (high skewness). Also, do not want to overcomplicate the analysis (g3b)

for Random: