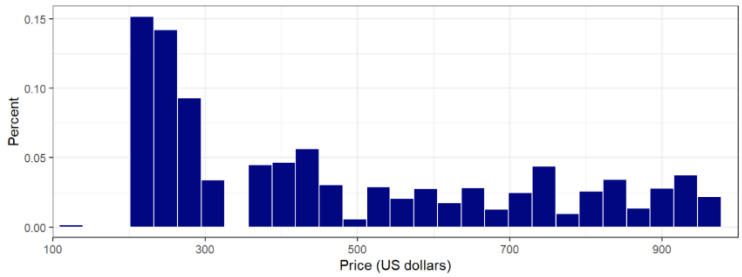
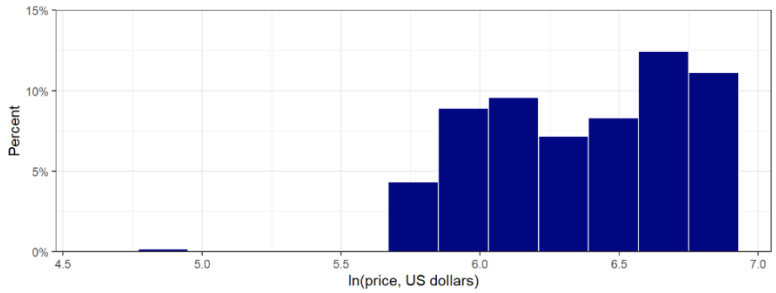
HW2 SalgaB Airbnb price analysis

The current analysis focuses on predicting Airbnb apartment prices for a company, which main activity is to rent apartments. The company has many apartments with places for 2 - 6 guests. They are interested in predicting the prices of their apartment, so they can price their apartments. The company's apartments are located in Sydney and its neighbourhood. The goal of the analysis is to predict the price that may be appropriate for their apartments. The analysis uses the Airbnb dataset from http://insideairbnb.com/get-the-data.html. The dataset has a single data table that includes more than 20 000 observations. The price variable is in US dollars. The first part of the analysis is to check the price variable and its log form. The following two histograms show the distribution of price and log price:

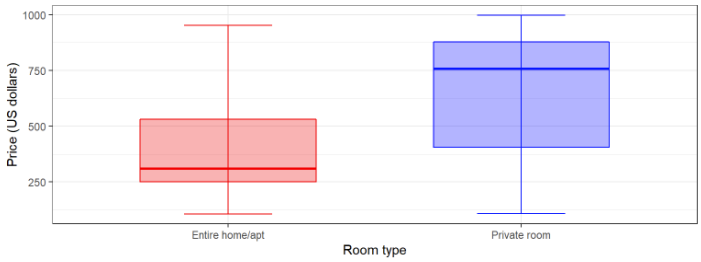
1. Histogram: Price



2. Histogram: Logarithm form of Price



Price data shows that apartment prices are strongly skewed with a long right tail. Log price is close to normally distributed between 5.5 and 7. To predict the analysis uses many important predictor variables: accommodate, number of rooms, number of beds. Also, some variables are related to reviews and not to the apartments directly: average review score, the number of reviews, the date of the first.

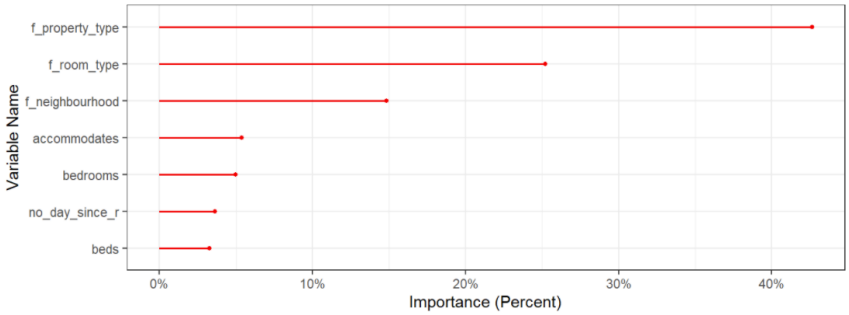


The 3rd histogram shows the average price by room. Entire homes and apartments are cheaper than Private rooms. Whole apartments are around 260 US dollars and Private rooms are around 750 US dollars.

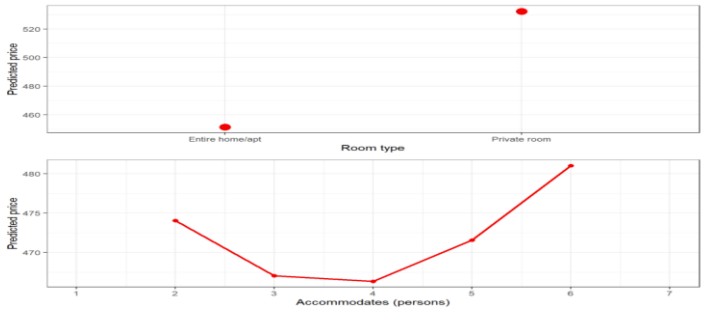
For data cleaning, many key decisions were made to clear all NA values, remove unnecessary variables and modify many others for easier analysis. During the data cleaning a huge number of apartments was dropped, so the final dataset has 12 419 observations.

For the analysis, 3 predictors were made for predicting price. They differ in the predictor variables included. The predictors are ordered by increasing complexity. The first predictor contains 14 basic variables about apartments. The second and third predictors also contain the same basic variables, but those contain variables

about reviews, reviews, and binary variables (the second predictor has 30 variables). The third predictor contains 39 variables since it has some specific variables for the LASSO model.
Before the prediction analysis from the 12 419 observation 20% holdout set (3724 observations) was selected. The remaining random 80% is the training set (8695 observations). In the analysis, these datasets were used for cross-validation.
First, the analysis started with estimating two random forest models. For model tuning, five-fold cross-validation was used, and eight variables were at the random forest for each split. All the models were optimized for Mean Squared Error.
The second random forest model has a slightly better result at RMSE as well. 1,54 dollars better RMSE at the second model (model_1 = 212.59 model_2 = 211.05). R2 shows a similar value as well: The second model fits better by 0.7%, so I used this second model for later analysis.



The diagnostics part of the analysis started with variable importance. The graph shows the variables grouped. The graph shows that the most important predictors are property and room type, after that neighbourhood is important as well.
The next step is to pick two important variables: the number of guests to accommodate and the room type.
The first graph shows the predicted price for different room types. By the room types, the predicted prices are much more volatile. The Private rooms (530 US dollars) are 80 US dollars more expensive than the Entire homes and apartments.



The second graph shows the predicted prices of apartments by the number of accommodates. The accommodates variable again shows interesting results: The predicted values for 3 and 4 people are the cheapest, an apartment for two people is more expensive than a one with a place for five people. Maybe this trend is because apartments for two people are aiming to be more luxurious.
For further investigation on the performance of the random forest models, first, the sample was divided into two groups based on the accommodate variable. Smaller apartments with three or fewer guests, and large apartments with 4 or more guests. For the meaningful comparison, RMSE was divided by the mean price. Large apartments have 13% better performance in terms of prediction. The neighbourhood was also compared. The predictive performances of them show no major difference. By property type, the difference is similar to the difference in room type performance. The results show that: small entire apartments are the hardest to predict.

```
##              Var.1     RMSE Mean.price RMSE.price
## 1   Apartment size       NA         NA         NA
## 2        large apt 182.5280   452.0926  0.4037403
## 3        small apt 230.1747   498.4272  0.4618020
## 4             Type       NA         NA         NA
## 5  Entire home/apt 185.9551   409.1640  0.4544758
## 6     Private room 263.8439   650.4420  0.4056379
## 7         District       NA         NA         NA
## 8           Auburn 177.6492   494.4717  0.3592708
## 9     Campbelltown 260.7803   623.3636  0.4183438
## 10 City Of Kogarah 172.4841   535.6818  0.3219899
## 11         Hornsby 208.8842   491.0426  0.4253892
## 12    North Sydney 177.3000   394.3179  0.4496372
## 13        Randwick 226.7637   506.2703  0.4479103
## 14          Sydney 207.6007   442.7341  0.4689061
## 15        Waverley 204.7022   459.6601  0.4453339
## 16       Woollahra 212.3092   472.9935  0.4488629
## 17             All 211.4507   478.9428  0.4414947
```

The analysis aims to find the best model and use that for helping the company in Sydney. For that, a comparison table was made. The models are the following: linear regression, LASSO, single regression tree with CART, two random forests, and GBM.

```
##                                  CV RMSE
## OLS                             216.1087
## LASSO (model w/ interactions)   213.3559
## CART                            212.1992
## Random forest 1: smaller model  212.5989
## Random forest 2: extended model 209.8692
## GBM                             208.5171
```

```
##                                 Holdout RMSE
## OLS                                 220.1260
## LASSO (model w/ interactions)       215.5275
## CART                                215.7493
## Random forest 1: smaller model      214.0365
## Random forest 2: extended model     211.4507
## GBM                                 210.2321
```

The GBM and the second Random Forest performed equally well. These models are followed by the CART and then the LASSO model. The last contender is OLS. The best model, estimate it on the work set and evaluate it on the holdout set. The best model is GBM and its holdout set RMSE is 210.2. The expectation is to make an error of 210 US dollars when using the model on the live data. The holdout set RMSE is quite close to the cross-validated RMSE.

So, the company should consider the RMSE, when they set their models to predict the prices of their apartments. The RMSE is quite huge and the R2 is also not so good. The analysis could be better if I could generate binary variables from the amenities variable.