# Final Project: a look into Brazilian Houses rents

2023-05-24

*Matteo Carucci, Alessandro Natoli, Tommaso Agudio and Lorenzo Ciampana.*

# 1. Understanding the dataset

In the first part of the project we will investigate some interesting trends and insights on the house rent market in Brazil. Let's get a broad idea of what the dataset looks like, checking any irregularities in the data.

*__Important Disclaimer: We understand the importance of providing clear and smart code, however we prioritized the importance of our findings. To check all our elaborations and techniques tried, please refer to the complete Rscript code - Some important chunks have been omitted from this report due to their size.__

## 1.1 Data cleaning, getting rid of nulls.

We now start with data preparation. Some duplicates were found and nulls in the `floor` column.

## 1.2 What about missing floor data?

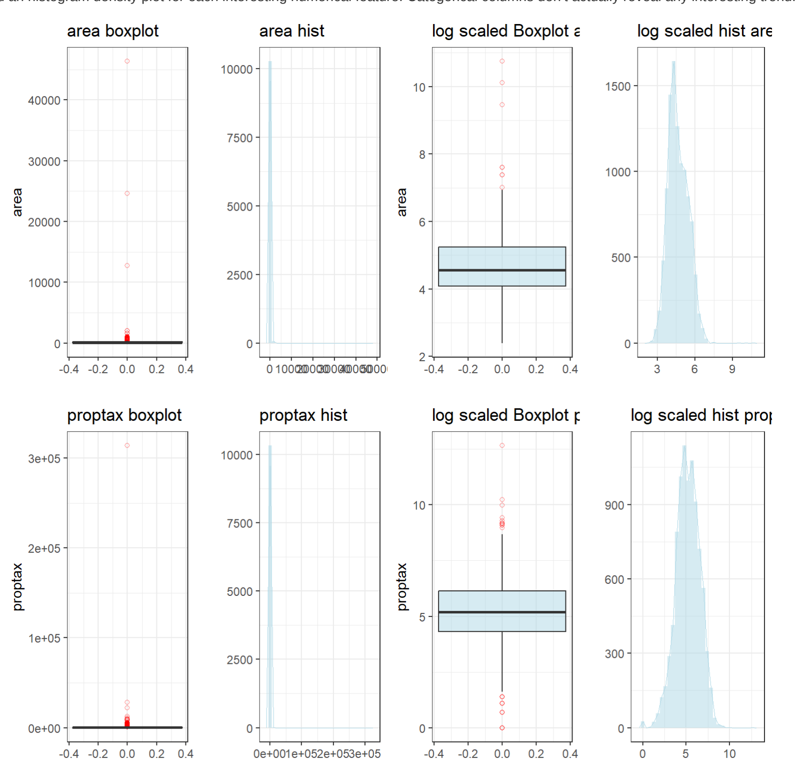
Looking at the cells above, We could either use mean or median imputation for replacing the nulls in the floor column. Anyways, the MICE package offers a set of very powerful methods to impute and estimate the missing values, so we will use it! In particular we used PMM Mice method which stands for predictive mean matching, a regression model which uses the other variables as predictors; trivially, it uses rows having similar predictors of the missing value row, and impute the missing value of the row by using the "similar" values as predictors.

***A summary of the new cleaned data.***

```
##              city          area             rooms          bathroom
##  Belo Horizonte:1209   Min.   :   11.0   Min.   : 1.00   Min.   :1.000
##  Campinas      : 823   1st Qu.:   59.0   1st Qu.: 2.00   1st Qu.:1.000
##  Porto Alegre  :1154   Median :   95.0   Median : 3.00   Median :1.000
##  Rio de Janeiro:1431   Mean   :  152.5   Mean   : 2.54   Mean   :1.283
##  São Paulo     :5712   3rd Qu.:  190.0   3rd Qu.: 3.00   3rd Qu.:1.000
##                        Max.   :46335.0   Max.   :13.00   Max.   :9.000
##  parking.spaces     floor               animal             furniture
##  Min.   :0.00   Min.   :  1.000   acept    :8073   furnished    :2515
##  1st Qu.:1.00   1st Qu.:  2.000   not acept:2256   not furnished:7814
##  Median :2.00   Median :  5.000
##  Mean   :1.33   Mean   :  6.482
##  3rd Qu.:2.00   3rd Qu.:  9.000
##  Max.   :8.00   Max.   :301.000
##       hoa              rent          proptax            fireins
##  Min.   :      0   Min.   :  450   Min.   :     0.0   Min.   :  3.00
##  1st Qu.:    180   1st Qu.: 1599   1st Qu.:    41.0   1st Qu.: 21.00
##  Median :    571   Median : 2750   Median :   130.0   Median : 37.00
##  Mean   :   1092   Mean   : 3967   Mean   :   377.1   Mean   : 54.28
##  3rd Qu.:   1289   3rd Qu.: 5000   3rd Qu.:   390.0   3rd Qu.: 70.00
##  Max.   :1117000   Max.   :45000   Max.   :313700.0   Max.   :677.00
```

# 2. EDA, is there anything affecting the rental prices?

After imputing the data we can now explore more in detail the distribution of each feature to identify eventual outliers and irregularities. From the summary, one can immediately see that there are some outliers in the features - The `area`, the `hoa` (Homeowners tax), fire and property taxes have some very suspicious maximum, suggesting that some houses may be completely irrelevant for our analysis. We'll see by looking at a boxplot and an histogram density plot for each interesting numerical feature. Categorical columns don't actually reveal any interesting trend.




***Above, you can see just a snapshot of a significant skewed distributions (area).***

As suspected most variables present some outliers. All the distributions are visibly right skewed, with most values being in the interquartile range and some houses with remarkably high features. Although it is true that we might log scale to make the distributions more "normal", that would underestimate the magnitude of the outliers.

## 2.1 Removing the *outliers*: some houses data just don't make sense

As we have an understanding of the dataset (we carefully analyzed it before taking this decision), we are now ready to get rid of the *outliers* as they might destabilize our results and the accuracy of our models. Obviously, we considered only numerical variables that make sense; for instance, it does not make sense to delete the house on the 300th floor (=, neither one with many rooms. In total, ***268 houses were disregarded***.

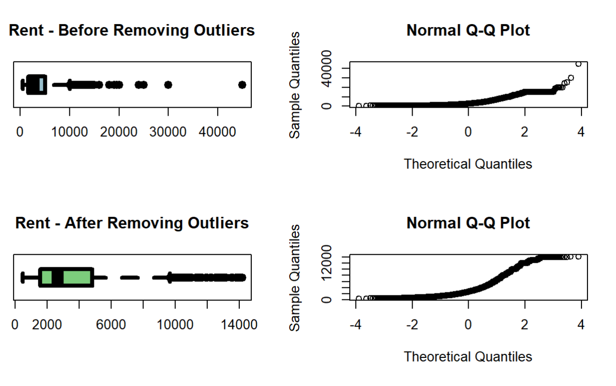
```
# Columns to consider for outlier detection
cols <- c("area", "hoa","rent", "proptax")
data_out <- data_imputed[, cols]

#using z-scores method
z_scores <- apply(data_out, 2, function(x) abs((x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE)))

# Identify rows with at least one z-score greater than 3 (considered outliers)
outlier_rows <- row.names(data_out)[apply(z_scores, 1, function(x) any(x > 3))]
# Print the number of outliers detected
cat("Number of outliers detected:", length(outlier_rows), "\n")

#new df without outliers
data3 <- data_imputed[!(rownames(data_imputed) %in% outlier_rows), ]

#231 out of 268 have rent == 15000, trivially some data entry errors
```




## 2.2 Thoughts on outliers, the importance of looking into them.

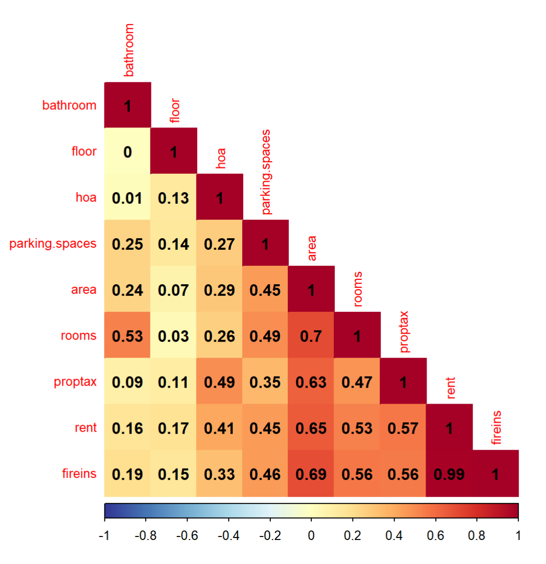
- For the outliers removal we decided to apply the *z-scores method*, setting the threshold to 3. Interesting things came up doing so - for instance most of the houses removed had the same rent price of 15000 - it is not difficult to notice that it may be a trivial data entry error! Also, some houses were incredibly big but had the same rooms of those which were way smaller. It is true that the **parking spaces** area could be included in the total `area`, but is it worth to consider these likely wrong typed data?
- By looking at the Q-Q plot it also came up that it is very unlikely that some of them were actually part of the dataset, some type errors might have been done given that data was crawled by rent ads data. Still some extremes are present in the new data but as already mentioned, it is very common to have house prices following a right skewed distribution in our opinion; Most of high ended houses are located in Sao Paulo, where most of the luxurious houses are at. This could suggest that it may be profitable for the agency to invest there.

## 2.3 Our previous try, the IQR method.

Before, We chose to apply the IQR method setting 1.5 as step for outliers in the lower quartile and a more relaxed 2 multiplier for extreme values - The approach with 1.5 as multiplier was too strict, deleting legitimate houses. The amount of deleted data was considerable ($\approx 12\%$), and it is highly unlikely that this amount of data contain all real outliers. Indeed, we assume that is compeletely normal to have very expensive and fancy houses in specific areas; once again most of the detected "false outliers" were located at Sao Paulo, where houses' rents are generally higher.

# 3. Correlation among variables: Can it affect our model?

We must assess the relevance of our predictors. We can check if there's any linear dependence among them by looking at the correlation matrix.

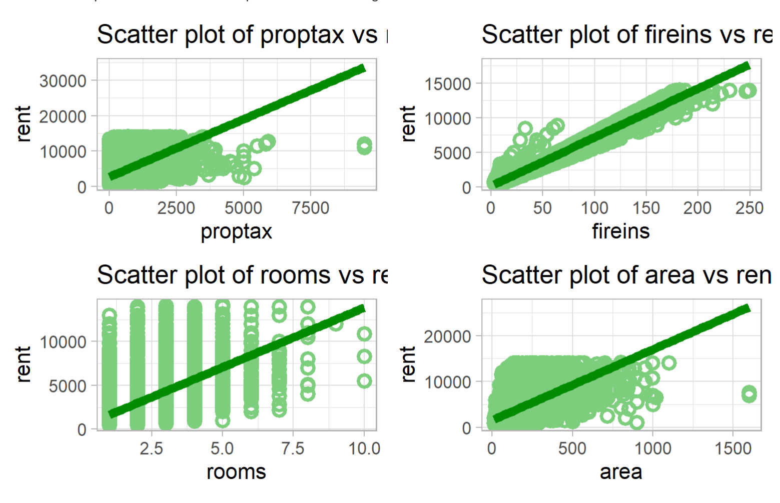



## 3.1 Is there a risk that linear models will be affected by Multicollinearity?

- Looking at the correlation heatmap, it is easy to see why rooms, area and the rent amounts are correlated - trivially, one would want to spend more if there are more rooms and space in a house and viceversa.
- On the other hand, it is remarkable the correlation of `rent` with the fire insurance `fireins`. It could make sense that the price of the insurance is correlated with the rent amount - this usually works for car insurance.
- No apparent risk of multicollinearity among predictors as the only 2 highly correlated variables are the fire insurance and our target. Houses' taxes are somehow correlated but not so much, we maybe can try to do a little feature engineering to sort this out.

## 3.2 Fitting with one predictor - is there linearity?

Since the end goal is to predict the rental prices, it would be interesting to see the single relationships between predictors and target (`rent`). We will show the variables having correlation higher than 0.4 with rent - the other coefficients are not that significant. The fitted regression lines below show the linear relationship between each chosen predictor and the target.



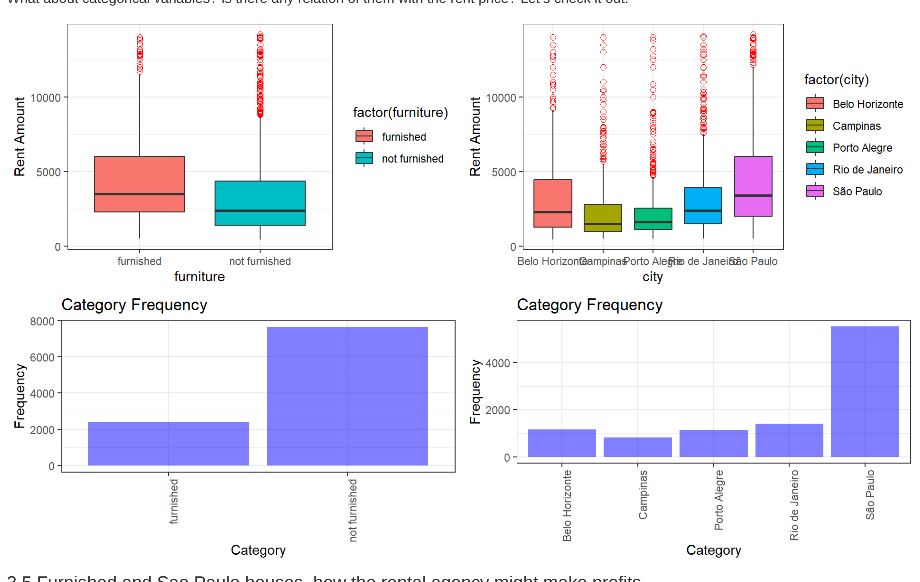


## 3.3 The importance of fire insurance.

Ironically, the scatterplots show that even taking the `fireins` as a single predictor could be a decent solution! (using a very simple linear regression, it can be seen a clear positive correlation). However we must investigate other relevant information that may come useful for our regression. Area also seems to have a positive relation with the rental prices as well as the property taxes. We will see if there is a risk of high variability in our models, since the bias should be quite low.

## 3.4 Categorical variables inspection

What about categorical variables? is there any relation of them with the rent price? Let's check it out!




## 3.5 Furnished and Sao Paulo houses, how the rental agency might make profits.

From the boxplots 3 important considerations can be taken away:

- Houses in Sao Paulo cost more than in other places as the interquartile and the median show. Also, the majority of the most expensive are located in this city.
- Whist animal allowance (`animal`) does not make a huge difference in rent prices, furniture obviously does. The lower quartile of the furnished home rental price corresponds to the median of not furnished ones, indicating a relevant difference in rental prices.
- It is interesting to see that most of the outliers we took out were "not furnished" homes but they had very high prices! Maybe they were really big properties like villas and entire buildings… given that some floor data was missing and we imputed it.

# 4. What do we really want to predict?

Although the primary objective was to identify the features that would make the real estate more profitable, that is pretty easy to see. We have already seen how rents are higher in the Sao Paulo area, and there's a trivial positive correlation between area and rent amount. Our main objective is to get a comprehensive view of rental prices, pointing out the major differences by their features. Some questions might include: Is there a significant difference in prices according to each `city`? or maybe: What is the amount of money that I'm likely to spend on monthly rents, given that I want x `rooms` and y `bathrooms`?) and first and foremost, what is a fair price I should pay for a rent in these areas? We will see whether the assumption we are doing on profitability really holds true.

## 4.1 Elastic net is the right balance between Lasso and Ridge regression.

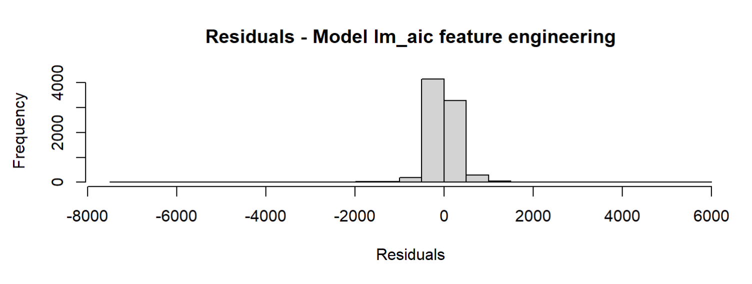
The first thing to do is to select the relevant features to include in the model. The first method that we will use to predict the rents amount is *Elastic net*. *Lasso* and *Ridge regression* won't be that useful (we tried them), as they usually don't handle well multicollinearity (In the case of lasso, it just picks one of the correlated and disregard the other basically). On the contrary as we saw in class, *Elastic Net* can balance well the L1 and L2 penalties by adjusting alpha!

## 4.2 Stepwise methods, great but less robust than regularization ones.

Important Disclaimer: Including `fireins` dramatically improves the model. We will show the best models and compare the performance with and without this variable; The positive correlation to the target is almost 1, making it the perfect feature, however evaluating the models without this feature could be beneficial to confirm the general goodness of the models. Before let's check out how a simple linear model with AIC and BIC criteria perform! Since AIC and BIC perform roughly the same, we will show the AIC as we don't have so many predictors to prefer a more penalized and strict approach. The AIC provides a good way to balance between complexity and generalization to new data; however they are less reliable, since variable selection could introduce some bias towards specific features - for instance, it may give high importance to `fireins` (we don't need to know how important that is) and disregard features which are actually influential because it wants to reduce the risk of overfitting and model complexity.

***Please Note that you can find the chunk code of the models in the Rscript. The chunk was very large and hence we did not include it in the report. Nonetheless, the 3 AIC models are listed below***:

1. AIC criterion (model with fireins and without it, back and forward elimination included!) lm_aic <- stepAIC(lm(rent ~ ., data = train_data), direction = "both", trace = FALSE)
2. AIC Linear Model without `fireins` lm_aic2 <- stepAIC(lm(rent ~ hoa + proptax + area + rooms + city + bathroom + floor + parking.spaces + furniture + animal, data = train_data), direction = "both", trace = FALSE)
3. model with some feature engineering to reduce predictors (combining correlated and disregarding less relevant) lm_aic3 <- stepAIC(lm(rent ~ hoa*proptax* + *area*rooms + city + bathroom + parking.spaces + fireins + furniture, data = train_data), direction = "both", trace = FALSE)




```
##                       Model             RMSE         R-Squared
## 1               Elastic Net 328.143644754368 0.987726156116444
## 2 Linear Model AIC complete 327.207350573513 0.987796098251815
## 3         Linear Model AIC2  1951.8139644678 0.565760384317353
## 4         Linear Model AIC3 322.091908057188 0.988174698344181
```

## 4.3 The importance of feature engineering.

It is interesting to see the results:

- The best performing model (AIC with featured engineering), which obviously includes `fireins` had explained almost 99% of the variability (R^2), its residuals distribution was not perfectly normal but a little right-skewed. This can be due to the fact that some non-linear relationship may be present between predictors (especially those having modest correlation) and our target. If we take a closer look at the fitted line with `fireins` as only predictor, we can clearly see some anomalies for low values of `fireins` (underestimated rents) that may be identified by more complex models; it is likely that this model highly relied on `fireins` and hence underestimated some rents.
- The model without `fireins` has performed quite poorly in comparison with the other 2. It is though visible that the model has no bias in its predictions (by looking at the distributions of residuals) but cannot understand the complexity of our data.
- The featured engineering has helped and shown how the animal feature and parking spaces were little helpful - they had low coefficients/completely removed both in elastic net and AIC first models. The combination of the correlated features might have reduced very little noise, contributing to a slightly lower RMSE and higher R-Squared. Some cities were also considered useless in affecting the price, though some had quite high coefficients, meaning that they have negative/positive correlation with the target! Also, rooms coefficients suggested a low impact on the target, reasonably because area had an higher influence.
- In all the models the `area` and the `city` play an important role: houses in Porto Alegre are generally deemed to have negative coefficients (negative correlation with rent, that is, the expected rent in Porto alegre is lower), suggesting as we have seen with the EDA, that it may be a good deal to rent a house there. From the Real Estate agency perspective instead, furnished homes and houses in Sao Paulo have a positive correlation with the rent prices, but this was expected. It is curious to see that a unit increase of area decreases the rent price in the some models, but why? Simple, on average we discovered that houses are way larger in Porto Alegre and Campinas, which have a negative correlation with prices, outweighing the effect of large houses in crowded cities like Rio and Sao Paulo that have on average smaller houses and higher rents.

# 5. Fancier methods: Random Forest and GAM Splines

As linear models work pretty well, we nonetheless think that some non-linear relationship can be captured. We tried numerous models, but these 2 work really well especially for these reasons:

## 5.1 Random forest, "handles it all".

- Random Forest solves most of the issues of other decision trees methods. By bagging, it trains the decision trees with random sampled data, avoiding correlation in predictions (this would increase variability, as the model would not be flexible and fail to fit well with new data!). As we did a gridsearch, the algorithm showed very similar RMSE with different tuning parameters, proving its robustness. Eventually we have chosen to have 500 trees and mtry set to 4 to reduce the risk of overfitting (the more trees, the less weight to each tree prediction, hence less dependence on a single one! - On the other hand, the less mtry, the less number of random predictors in each tree, reducing the complexity of the model). The bias was also very low as testified by the validation set RMSE and R-squared.

## 5.2 Generalized additive models caught non-linearity trends.

- On the other hand, GAMs are good for this situation since they should detect the non-linear relationships we mentioned before through splines, special functions that can detect non-linear relationships and introduce smooth curves that fit the data.

***Please Note that you can find the chunk code of the models in the Rscript. The chunk was very large and hence we did not include it in the report. A thorough description of the models is available on the Rmarkdown.***

```
##                          Model             RMSE         R-Squared
## 1                  Elastic Net 328.143644754368 0.987726156116444
## 2    Linear Model AIC complete 327.207350573513 0.987796098251815
## 3            Linear Model AIC2  1951.8139644678 0.565760384317353
## 4            Linear Model AIC3 322.091908057188 0.988174698344181
## 5       Random forest Complete 261.601085502017 0.992621002840413
## 6     Random forest no fireins 1726.59444650851 0.661352933411327
## 7                         Gam1  299.61462564809 0.989789052757959
## 8 Gam with feature engineering 290.874514881628 0.990359252815521
```

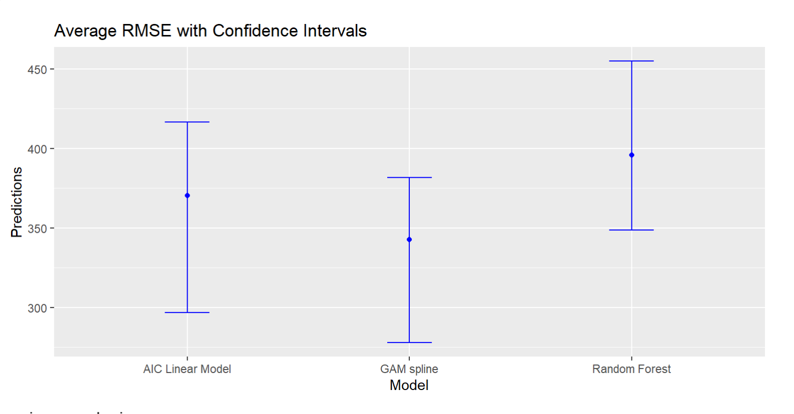
## 5.3 Validation set results.

Looking at the dataframe, there's little doubt which method to choose. Not only random forest is the best choice in terms of model perfomance indicators, but it also is surely the most flexible - we also had a significant improvement without the fireins feature, meaning that the model was able to explain further variability.

## 5.4 The variance-bias trade off, a personalized "bootstrapping".

On the one hand, the ability to adapt to different kind of houses is crucial for the real estate agency and also for potential families which want to invest money wisely. On the other, our results must be as accurate as possible. We chose to use a "personalized" bootstrapping to assess the model performance. We are basically taking 10 sampled training and test sets and allowing replacement, unlike cross-validation which does not train and test on already seen data (folds), we are basically sampling 80% of the data for training and 20 as validation set for 10 iterations. To enhance computational efficiency and capture enough variability, we chose to sample 30% of the dataset and then split 80-20; Not only is the size that provides most reliable and stable estimates (we tried sampling 50%, 63% and 80%) but also perfectly depicts the concept of the **bias-variance tradeoff**.

***Please Note that you can find the chunk code of the models in the Rscript. The chunk was very large and hence we did not include it in the report.***



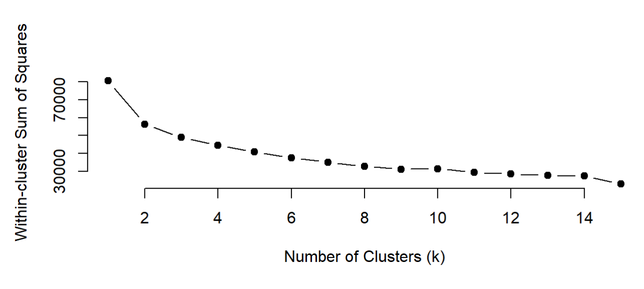

## 5.5 Regression conclusions.

- Our bootstrapping estimates confirms our initial belief, Random forest strikes **bias-variance tradeoff**. Even though GAM has lower average RMSE (and hence, less expected bias), the Confidence interval is larger, indicating that the model has higher variability and a little bias to underestimate rentals (We can notice looking at the upper CI). On the other hand, Random forest has no systematic bias as the confidence interval and its estimates are perfectly normally distributed, with both the upper and lower CI being roughly the same! The secret? Random forest is a bagging method, hence reduces overfitting and gerenalizes better! In this instance, the RMSE estimation is important but does not tell all - our random forest model is tuned with very conservative hyperparameters and increasing the amount of training data (as we did with the validation set results) shows the goodness of the model.
- Another interpretation of the confidence interval is the following: With CI we are estimating error that the models have. What if we could assume that CI could be interpreted as a threshold for good deals for the real estate agency and for affordable houses? Imagine one wants to be 95% sure that the home to rent has a fair price; he/she can look at the predictions of similar homes and estimates of the models error - the one can check whether the difference **rental offered - predicted rental** lies in the estimated error interval, that is, for the predicted rental price based on the same brazilian house market segment, is the price in line with those of similar homes?

# 6. Clustering: Can we identify the most profitable and convenient houses?

As we said, there are certain factors that make an house more profitable, first and foremost its location and area; the initial idea though was to get a comprehensive idea of the rental market and identify also the most convenient ones! Clustering will help us identify diverse rental groups, giving advice of where, how spacious an house should be to save some money. We will start by implementing **kmeans**, using the elbow method and silhouette score to assess the right number of clusters k.

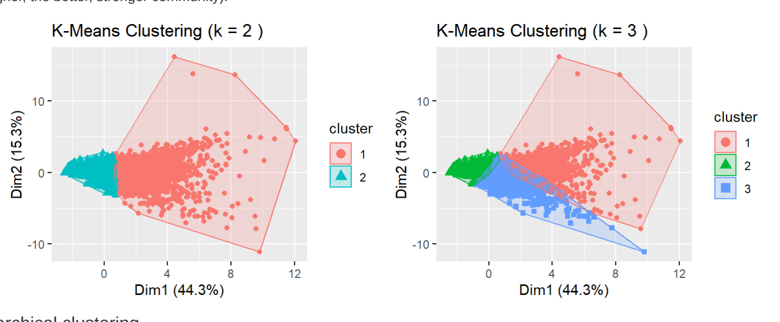
***Below, Within sum of squares shown by the elbow method.***




```
## NULL
```

## 6.1 The best k for Kmeans.

It is evident that both the **elbow method** and the **silhouette scores** suggest that the right number of clusters is k = 2. Indeed the silhouette score by k measures the quality of clustering, the higher it is, the better, as it indicates that they are more distinguishable and separate. On the other hand the within sum of squares indicated in the elbow plot indicates roughly "how close" data points are in the cluster, that is, how segregated they are (the higher, the better, stronger community).



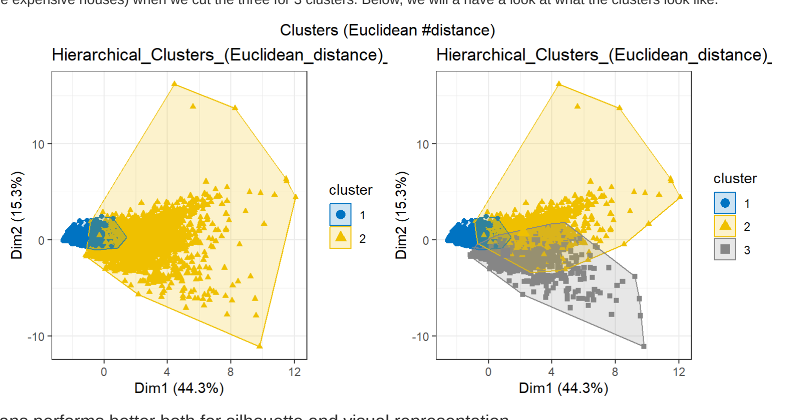

## 6.2 Hierarchical clustering.

Now let's get into **hierarchical clustering**! We want to effectively compare which method works best to identify the distinct home groups. We compared the euclidean and row correlation but the first behaves better overall, delineating clearer and more defined clusters. ***Note that the dendogram plot is not included - in the section below We will describe what it looks like.***

```
## Silhouette scores for euclidean distance: 0.3006866 0.3151515 0.3168659
```

## 6.3 Dendogram insights.

How do visualized clusters change according to their number k? For hierarchical clustering k = 2 and k = 3 look again the best parameters - apparently the algorithm finds 2 very distant clusters with k = 2, and 2 closer groups (presumably middle and low income houses) with a separate group (more expensive houses) when we cut the three for 3 clusters. Below, we will a have a look at what the clusters look like.

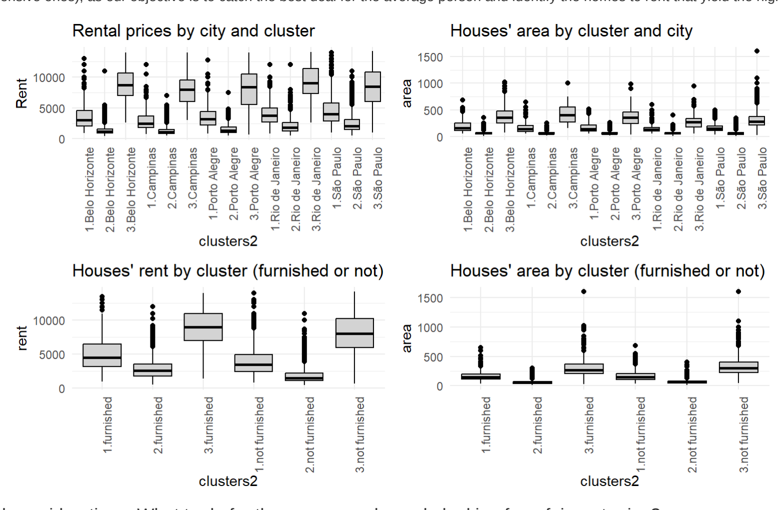



## 6.4 Kmeans performs better both for silhouette and visual representation.

Having compared **hierarchical** and **Kmeans** methods, the results are pretty clear. Although the first made a good job, **Kmeans** behaves better as clusters are more separate and segregated (both visually and in terms of silhouettes scores).

# 7. Do these clusters really tell what we want?

Our aim was to detect more profitable and more convenient houses in general - has **kmeans** detected different houses profiles? For responding to this, we clustered using both K = 2 and k = 3 number of clusters to look into them. We focused the investigation on prices and `area` as well as the `furniture` and the `city`. The 2 just mentioned were the most influential feature that changed the rental price; features like `floor`, `animal` and `parking spaces` don't really tell us much. To answer our questions we used k = 3 - in general houses are divided into 3 categories, small/low income ones, mid income, and luxury ones. It is important to understand that silhouette score and the elbow method are not enough to choose the correct k, because te context and the analysis objective come first. In this way, we carefully analyzed the cluster 2 (average houses) and cluster 3 (more expensive ones), as our objective is to catch the best deal for the average person and identify the homes to rent that yield the highest profit.




## 7.1 Final considerations: What to do for the agency and people looking for a fair rent price?

Looking at the 3 clusters boxplots, the results are even clearer than before. **From the real estate perspective it is clear that:**

- Kmeans has differentiated houses by area and rent price. Although some outliers are visible in the clusters, it is easy to see that the most profitable houses (given same area) are those in Sao Paulo and Rio de Janeiro. Small houses belonging to cluster 1 in particular are very overpriced when compared to other cities, suggesting that these may yield higher profits. If we value count the cluster cities, we notice that most of them are in Sao Paulo. Though, the cluster 3 which contains the most expensive and largest houses have predominantly houses in the city, with a stunning 78% of them; the other clusters are much more balanced, suggesting once again that Sao Paulo is the city to invest in for the real estate agency.
- Also, it is worth investing in furnished houses - Their area is comparable to not furnished ones and the difference in rental prices is negligible - buying non furnished homes to rent may not be a good deal for the agency, given that they need to furnish them to rent to clients (especially for most expensive ones).

**From the people looking for rents instead:**

- Houses in Campinas are much more worth the money. Although they may not be as appealing as the more luxurious and fancier in Sao Paulo, the cluster 1 revealed that mid class houses in Campinas are on average bigger and cheaper. Also, the median price for furnished mid-class homes is insignificantly higher than not furnished ones, so it may be worth renting them, with the median area being also very similar.