Data Analysis Final Project

Kevin Wang and Simone Angelo Meli

2024-06-19

Point 1 and 2: Data Description and Pre-Processing

The dataset chosen is the 'Brazilian houses' dataset and the tasks are to firstly analyze the driving factors that increase rent prices in Brazil and then clustering the houses for rental according to their characteristics. The dataframe consists of 10692 observations and 12 variables, 3 being categorical while the rest numerical. We will be taking a closer look at these variables and how they interact with eachother once we clean the data such as handling duplicates and missing values.

Handling duplicates:

```
## [1] "The number of duplicate rows deleted are 363"
```

Handling missing values

We noticed that there's no NA values in the data but there is are 2369 "-" values in 'floor' feature, which could either represent a missing value or it could suggest that '-' symbol represents properties with no floors such as a house not a part of a building or apartment. This could also indicate the ground floor as in brazil they do not follow the US naming conventions for floors where ground floor is labeled as level 1 but in Brazil ground floor is labeled like in europe simply by describing it as the ground floor or "T" for "Térreo".

We proceed by handling these '-' values first substituting them with 'NA' and then analyzing them to understand whether we can distinguish what they actually represent.

To understand if the "-" represents a house that is not a part of the condominium, one clear feature stands out which could help deduce this which is the 'hoa'. 'hoa' represents the Monthly Homeowners Association Tax which means if a house is not a part of a condominium the value 'hoa' will be 0. To identify this relationship we firstly identify rows with NA in 'floor' feature and then count zeroes in 'hoa' for the NA rows.

```
## [1] "Number of 0s in hoa for rows with NA in floor: 2015"
```

The results depict that 2015 values out of 2369, around 85%, which had '-' as their 'floor' also had 0 as thier 'hoa' value. From this we can safely assume that the '-' represents houses without a condomiunium or 0 floors as there are still a percentage of houses who could possibly be a part of a condominium but just at floor 0. Through this we decide to subsitute all NA values with 0.

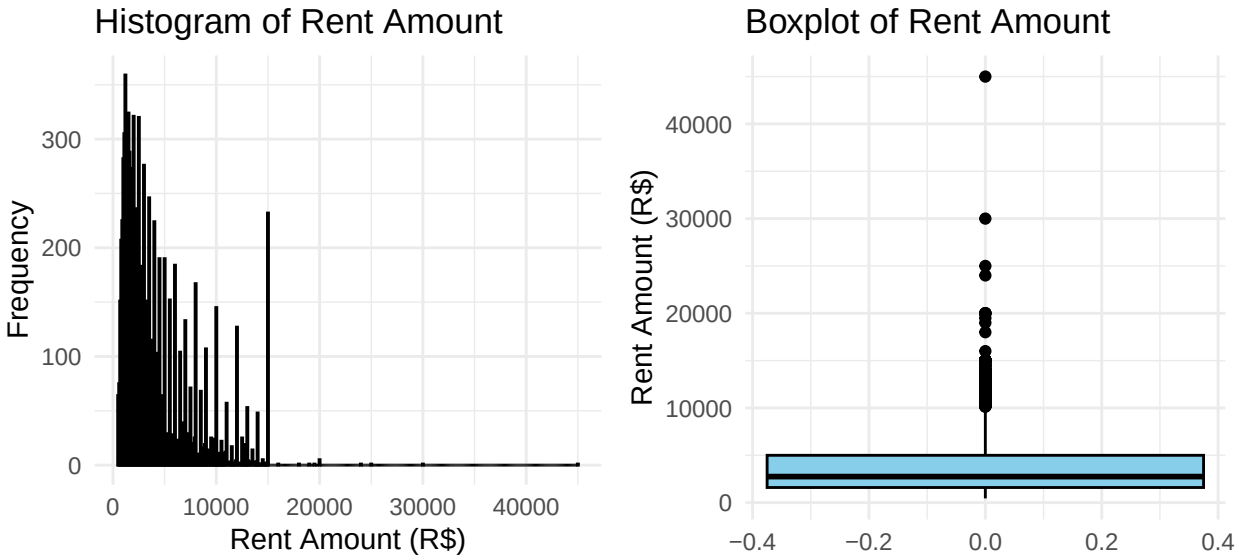
Encoding numerical and categorical variables

We decide to encode all variables as numeric except for 'city', 'animal', and 'furniture' which are categorical features and encoded as factors.

Summary statistics for numerical variables

The summary statistics reveal several points in the data set. The ones that stand out the most are the outliers in features such as hoa..R.. where the 3rd quantile is 1289 while the max is 1117000 or in the property.tax..R.. feature where the 3rd quantile is 390 but the max value is 313700. It clear that these maximum values can be seen as outliers in our dataset or interpreted as such. Another interesting point that canbe brought up later is the fact that the bathroom feature is mostly composed of houses with 1 bathroom as shown from median, min, 1st quantile, and 3rd quantile, but it seems like there are houses which reach maximum of 9 bathrooms, and so it seems like those could be outliers too and if thats the case bathroom could be considered a feature to drop in our analysis as it is mostly composed of 1. To visualize better the distribution of features and the outliers we create histograms and boxplots.

Histogram and Boxplot of numerical variables



From both the histogram and the boxplot many outliers can be visualized. From the histogram outliers are suggested by the extremely left skewed graphs while from the boxplots they can be visualized by the spread of the points and how far they distance from the mean. The outliers can be seen as very expensive houses or a possible errors or inconsistencies in the data as the data was gathered through a web-crawler. In order to obtain more accurate data we choose to remove some of these outliers.

Removing values four standard deviation away from their mean

We chose to remove the outliers with value that were four standard deviation away from their mean as we did not want to remove too many values and drastically affect the data, and we opted for four st deviations away from the mean which only removed 187 values out of 10329 which is 1.81% of the data.

Correlation matrix of numerical variables

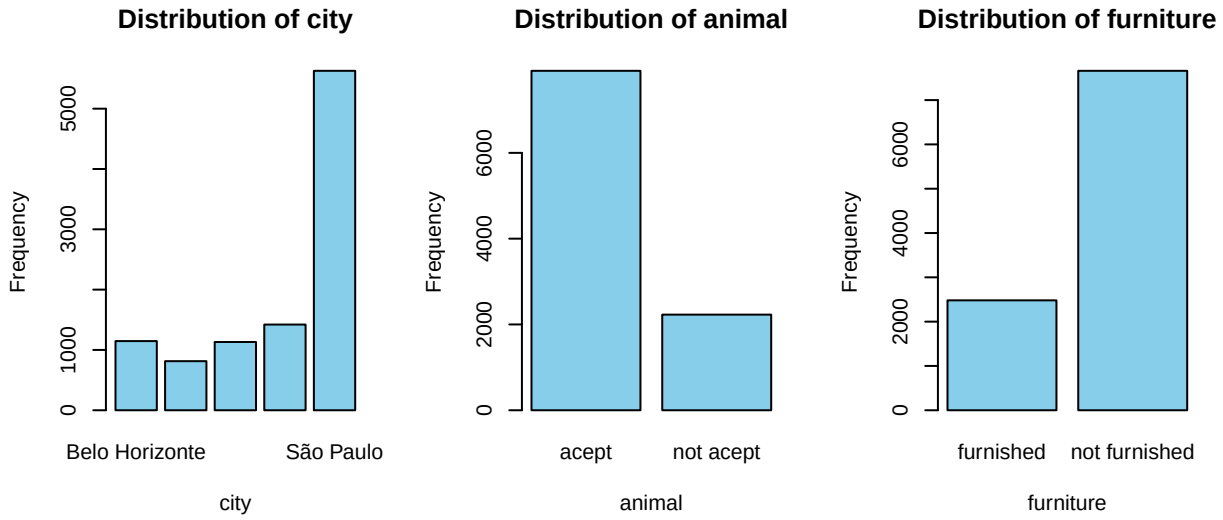
In the correlation matrix we can view the correlations of each feature with eachother, and understand key relationships. Seeing that the objective of our first task is to build a predictive model and find out the rent amount according to the house specifics, we mainly focus on the rent amount feature as our response variable and how it interacts with the other features.

Plotting only the correlation of rent amount with the rest of the features

```
##                area               rooms            bathroom      parking.spaces
##           0.65586035          0.53092465          0.08460863          0.44224173
##               floor            hoa..R..    rent.amount..R..   property.tax..R..
##           0.10676744          0.45719084          1.00000000          0.58043290
## fire.insurance..R..
##           0.98795414
```

From the table above we can see clear relationships between rent amount and area of the house as well rent amount with rooms and with property tax. The most distinct correlation we see is with rent amount and fire insurance with around a 0.988 correlation depicting a very strong linear relationship. We can also note a few features which have no correlation with any other feature or very poor correlation such as floor and bathroom which we could look to drop as features. Parking spaces and hoa, also do not reach a correlation of above 0.5 with rent amount and could be dropped during feature selection.

Categorical feature analysis



From the distribution of each categorical variable we are able to deduce various points. Firstly the city with the most houses registered in the data frame is São Paulo by a large margin which will be taken into account in future analysis as data is bias towards houses located in São Paulo. Furthermore, we are able to note that most houses do accept animals, around 8000, with around 2000 who don't. The same goes for the furniture distribution where their is clear bias for non furnished houses in the data frame over furnished ones.

Since rent amount is our response variable, we will visualize the relationship of rent amount with other categorical features to understand whether rent amount is affected by any of them.

Rent amount distribution by city

```
##             city rent.amount..R..
## 1 Belo Horizonte         3429.044
## 2       Campinas         2362.113
## 3   Porto Alegre         2213.986
```

```
## 4 Rio de Janeiro         3288.792
## 5      São Paulo         4640.364
```

From the data shown above, we are able to deduce multiple points. São Paulo has the highest mean rent, possibly due to its status as Brazil's largest city and financial hub, attracting significant business and professional demand. Other cities like Rio de Janeiro and Belo Horizonte also show high rents possibly due to economic activity, population density, and quality of life. Campinas and Porto Alegre have lower mean rents possibly due to their smaller economic scale, lower cost of living, and local economic conditions compared to larger cities like São Paulo and Rio de Janeiro.

Perform one-way ANOVA

We can verify whether the difference in mean is significant or not through a one-way anova test:

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  rent.amount..R.. and city
## F = 325.34, num df = 4.0, denom df = 2933.3, p-value < 2.2e-16
```

Based on the ANOVA results, there are significant differences in mean rent amounts between the cities as p-value is lower than 0.05. The same process is applied to rent amount distribution by animal and by furniture with t-test to confirm significant differences.

Rent amount distribution by animal

```
##      animal rent.amount..R..
## 1     acept         3965.432
## 2 not acept         3487.206
```

Seems like houses that do accept animals have a higher mean than houses that do not which could make sense as higher flexibility is awarded with higher mean rent. Based on the T-test results, there is a significant difference in mean rent amounts between whether a house accepts or does not accept animals as seen from p-value lower than 0.05.

Rent amount distribution by furniture

```
##       furniture rent.amount..R..
## 1     furnished         4905.130
## 2 not furnished         3521.971
```

Seems like houses that are furnished have a higher mean than houses that are not which could make sense as furniture can play a big part in a house's demand as well as appealability. Based on the T-test results, there is a significant difference in mean rent amounts between whether a house is furnished or not as seen from p-value lower than 0.05.

Point 3: Task 1

The objective of our first task is to build a predictive model to estimate the rent amount based on house specifics, with rent.amount..R.. as the response variable applied to regression models.

Practical Relevance

In practical terms, predicting the rent amount accurately is valuable for both renters and landlords. Renters can make informed decisions based on their budget and preferences, while landlords can set competitive prices to maximize occupancy and revenue. This model can also assist real estate agencies in providing data-driven advice to their clients.

Point 4: Lower Dimensional Models

Let's investigate the relationship between the response variable (rent.amount..R..) and a few selected variables (area, rooms, bathroom, furniture) to understand their importance in predicting rent. Multiple Linear Regression model and a Random Forest model are used to explore these relationships.

Multiple Linear Regression

```
##                           Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)             2030.0066 78.0784089  25.999589 2.514139e-144
## area                      12.9195  0.2559507  50.476503  0.000000e+00
## rooms                    687.4923 33.6643293  20.421980  7.081361e-91
## bathroom                -440.0825 52.4870596  -8.384591  5.766608e-17
## furniturenot furnished -1510.6067 54.7270950 -27.602538 8.997140e-162
```

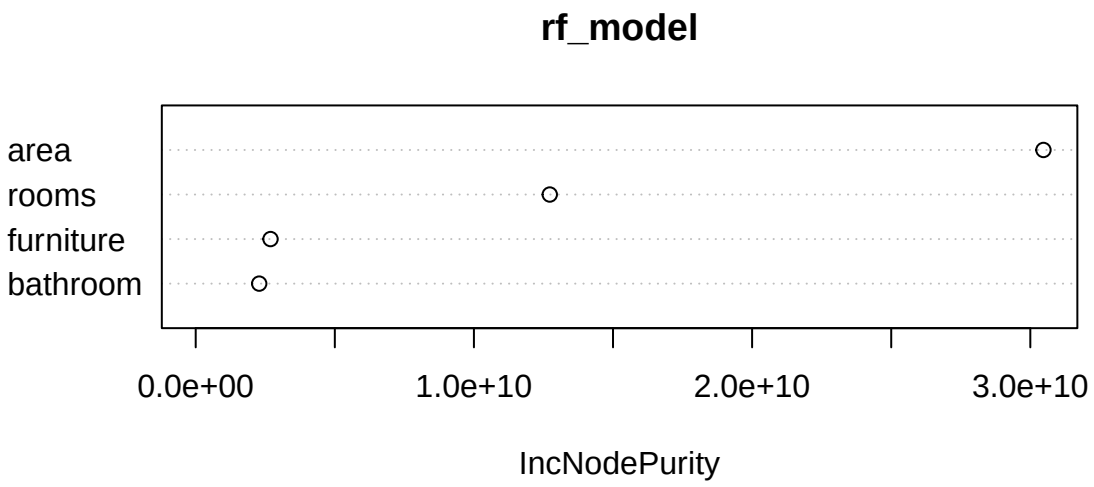
Results: The intercept represents the expected rent amount when all predictors (area, rooms, bathroom, and furniture) are zero. In this context, an intercept of 2030.0066 suggests that even with no area, rooms, bathrooms, or furnishings, the base rent amount starts at approximately 2030.01 (currency units, likely Brazilian Reais). The extremely low p-value indicates this estimate is highly significant.

Random Forest

```
##           IncNodePurity
## area         30472041533
## rooms        12726583364
## bathroom      2282222049
## furniture     2688485010
```

The results depict the total increase in node purity contributed by each variable, averaged across all trees in the forest. A higher IncNodePurity score indicates a greater contribution to reducing impurity, making the variable more important for predicting the target variable, which in our case is the rent amount.

1. area has the highest score, indicating it's the most important predictor.
2. rooms are also highly important but less so than area.
3. bathroom and furniture contribute less to the model's predictions compared to area and rooms.



Point 5: Model implementation

Linear Model with Stepwise Selection (AIC and BIC)

```
## AIC Stepwise Selection:

## MSE: 119581.3

## RMSE: 345.8053

## R-squared: 0.98685

## BIC Stepwise Selection:

## MSE: 119709.8

## RMSE: 345.9911

## R-squared: 0.98683
```

The comparison between AIC and BIC results shows minimal differences, with BIC yielding a marginally higher MSE and RMSE, and a very slightly lower R-squared value. This slight difference suggests that both criteria lead to similar models in terms of performance, but BIC, which penalizes model complexity more strongly, might result in a more parsimonious model. Overall, both stepwise selection methods provide robust models with high explanatory power and reasonable predictive accuracy for the rent amount.

Penalized Approache (Lasso)

```
## Lasso (lambda.min) MSE: 119376.7

## Lasso (lambda.min) RMSE: 345.5094
```

```
## Lasso (lambda.min) R-squared: 0.9885972
```

```
## Lasso (lambda.1se) MSE: 135379.4
```

For the Lasso regression using the lambda.min value, the Mean Squared Error (MSE) is 119376.7, the Root Mean Squared Error (RMSE) is 345.5094, and the R-squared value is 0.9885972. These results indicate that the Lasso model explains approximately 98.86% of the variance in the rent amount, showcasing a high level of accuracy and a good fit. When using the lambda.1se value for the Lasso regression, the MSE increases to 135379.4, suggesting that this model, which is simpler due to stronger regularization, does not perform as well as the model using lambda.min.

Non-linear Models (KNN, Splines, XGBoost)

```
## k-Nearest Neighbors:
```

```
## MSE: 2425025
```

```
## RMSE: 1557.249
```

```
## R-squared: 0.7679
```

These metrics suggest that the KNN model explains approximately 76.75% of the variance in the rent amount, which is relatively moderate. However, the high MSE and RMSE values indicate that the model's predictions have a substantial error margin, implying that KNN may not be the best model for this dataset.

Splines

```
## Splines:
```

```
## MSE: 114389.8
```

```
## RMSE: 338.2156
```

```
## R-squared: 0.98917
```

The Splines model explains approximately 98.92% of the variance in the rent amount, demonstrating a high level of accuracy. The lower MSE and RMSE compared to the KNN model suggest that the Splines model has a much smaller error margin, making it a more suitable choice for predicting rent amounts in this dataset.

XGBoost

```
## XGBoost:
```

```
## MSE: 46386.76
```

```
## RMSE: 215.3759
```

```
## R-squared: 0.99553
```

The XGBoost model explains approximately 99.56% of the variance in the rent amount, signifying a very strong fit. The relatively low MSE and RMSE values suggest that the model's predictions are highly accurate

Overall Models Comparison

```
##     Model        MSE      RMSE R_squared
## 1     AIC  119581.31  345.8053 0.9868543
## 2     BIC  119709.85  345.9911 0.9868302
## 3   Lasso  119376.74  345.5094 0.9885972
## 4     KNN 2425024.68 1557.2491 0.7678969
## 5 Splines  114389.82  338.2156 0.9891748
## 6 XGBoost   46386.76  215.3759 0.9955325
```

```
## Test Set Performance:
```

```
## MSE: 47348.9
```

```
## RMSE: 217.598
```

```
## R-squared: 0.9955557
```

These metrics indicate that the model maintains a high level of accuracy on the test set, explaining approximately 99.55% of the variance in the rent amount. The relatively low MSE and RMSE values suggest that the model's predictions are close to the actual values, confirming its robustness and generalizability to unseen data.

Calculate prediction errors and confidence intervals

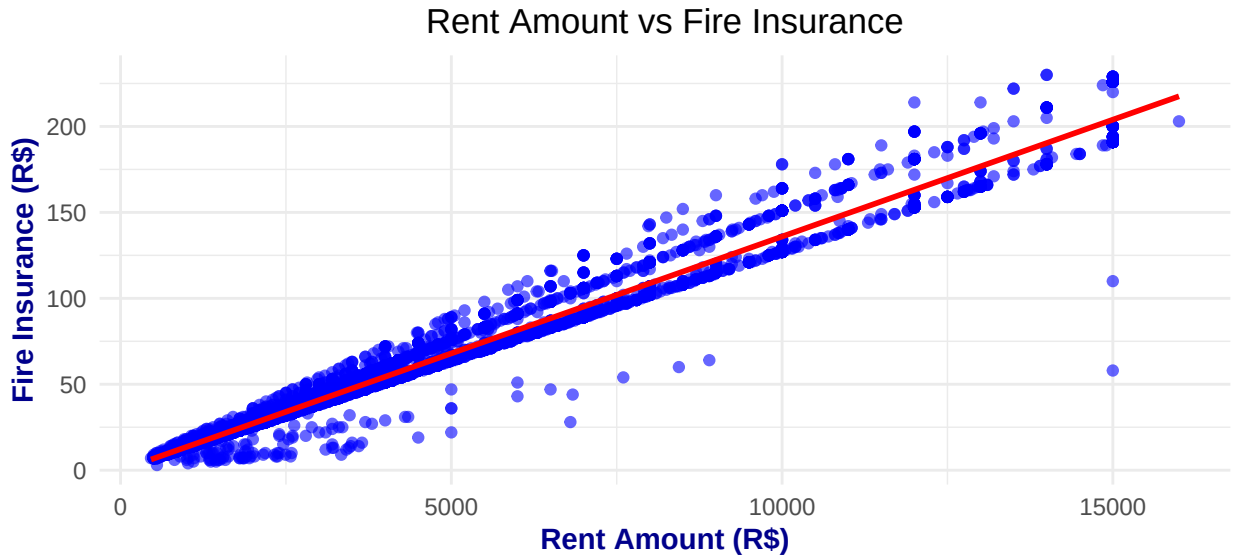
```
## Confidence Interval ( 95 %): [ -411.02 ,  440.24 ]
```

By subtracting the actual rent amounts in the test set from the predicted values, we obtain the prediction errors. The mean and standard deviation of these errors are then used to construct a 95% confidence interval for the error. This interval, which ranges from -412.57 to 442.02, indicates that the predicted rent prices are expected to deviate from the actual rent prices by this range on average.

Feature Importance Analysis

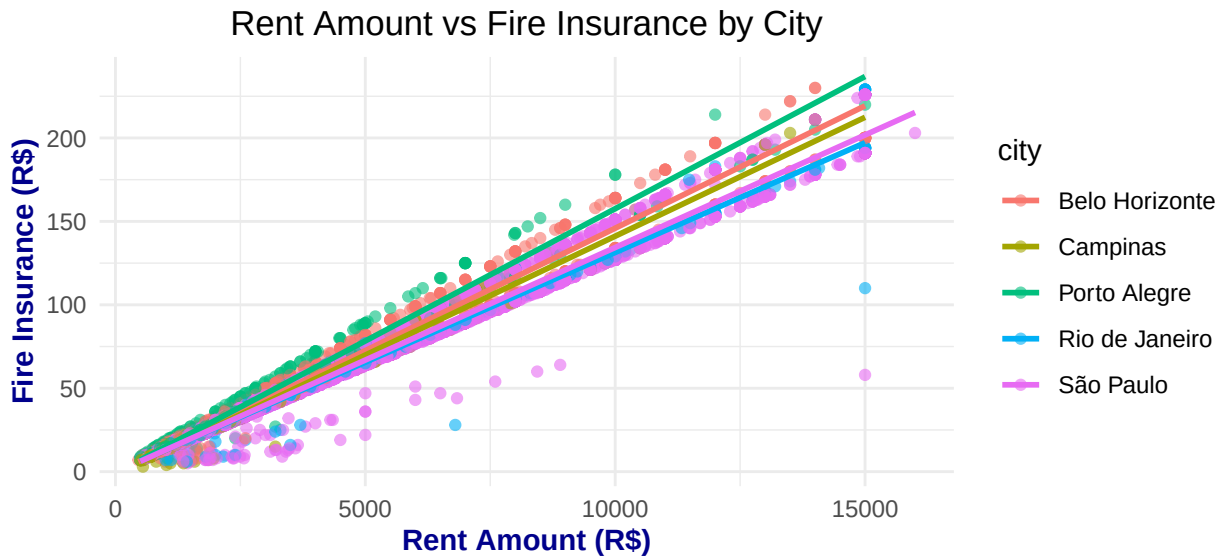
The most striking observation from the plot is that the feature fire.insurance..R.. is overwhelmingly the most influential factor Other features such as floor, city, hoa..R.. have negligible importance in comparison, with their importance scores being close to zero. This suggests that while these features contribute to the model, their impact is minimal relative to fire.insurance..R… The dominance of fire.insurance..R.. underscores its critical role in determining rental prices, possibly reflecting the higher replacement costs associated with more valuable properties, which in turn command higher rents.

Further Analysis on Fire Insurance



The scatter plot depicts the relationship between rent amount and fire insurance costs, showing a clear linear trend. The plot visually supports the conclusion that properties commanding higher rents are associated with higher fire insurance costs, likely due to their greater value and replacement costs.

Scatter plot with cities



The plot also highlights that, while the overall trend is similar, some cities have a wider spread of data points, suggesting greater variability in fire insurance costs for similar rent amounts. This analysis reinforces the critical role of location in determining both rent and associated insurance costs, and it underscores the importance of considering city-specific factors when predicting rental prices.

Point 6: Drawing Conclusion

In conclusion, our comprehensive analysis and modeling efforts have provided valuable insights into the rental market dynamics, highlighting the critical factors influencing rent prices. The XGBoost model, with its high R-squared value and low prediction errors, has proven to be an exceptionally accurate tool for predicting rental prices, effectively capturing the nuances of the data. Our findings underscore the significant impact of fire insurance costs on rental prices, as evidenced by the feature importance analysis. This strong correlation suggests that properties with higher fire insurance premiums generally command higher rents, reflecting their increased value and associated risks. Additionally, the variation in the relationship between rent and fire insurance costs across different cities indicates the importance of location-specific factors in rental pricing. The detailed analysis of rental prices in relation to various features, including the number of rooms, area, and presence of amenities such as parking spaces and furniture, provides a nuanced understanding of the market. For instance, while traditional factors like the number of rooms and area are important, our model suggests that fire insurance costs and city-specific characteristics play a more crucial role in determining rental prices.

Point 7: Task 2 Clustering

The objective of our second task is to cluster the houses for rental according to their characteristics using clustering models such as K-means and hierchical clustering.

Practical relevance

In practical terms, predicting clustering the houses for rental according to their characteristics can further help understand what dominant features lead to successful houses for rental in the market and this increases the amount of valuable knowledge that can be achieved by new companies who want to enter the real estate market.

Feature selection and data scaling

Before scaling our data, we firstly used one-hot encoding for all categorical variables, then we decide to remove a few features which we noticed had no correlation or very little correlation (< 50) with the whole data, as well as categorical features, city and animal which although affects rent amount, we believe that removing them and comparing the clusters later with these categorical features would be more effective. We decide to leave categorical feature 'furniture' in as it can be considered a more crucial component of a house.

After removing unnecessary features and scaling the remaining data, we are left with 7 features. Initially, we applied clustering models directly to these 7 features. However, we found that using PCA to further reduce the dimensionality to 3 principal components resulted in higher silhouette scores for both clustering models. Therefore, we decided to use PCA for dimensionality reduction before applying the clustering models.

Top 5 contributors for all principle components

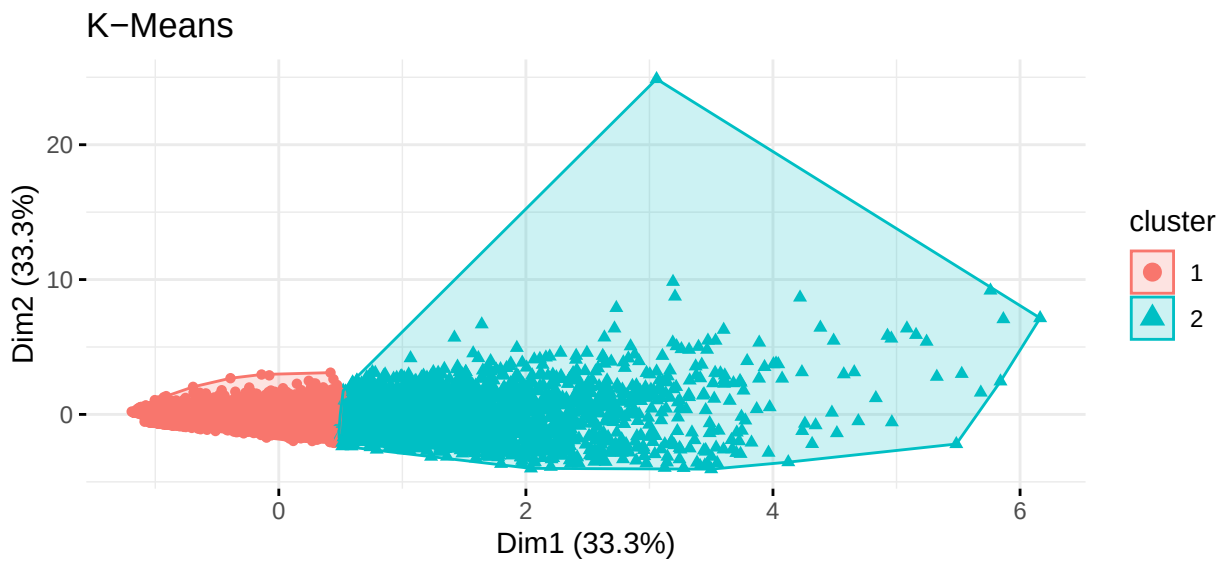
```
##      PC1                   PC2                       PC3
## [1,] "rent.amount..R.."    "furniturenot furnished" "hoa..R.."
## [2,] "fire.insurance..R.." "rooms"                  "property.tax..R.."
## [3,] "area"                "area"                   "fire.insurance..R.."
## [4,] "property.tax..R.."   "hoa..R.."               "rent.amount..R.."
## [5,] "rooms"               "rent.amount..R.."       "furniturenot furnished"
```

Clustering models

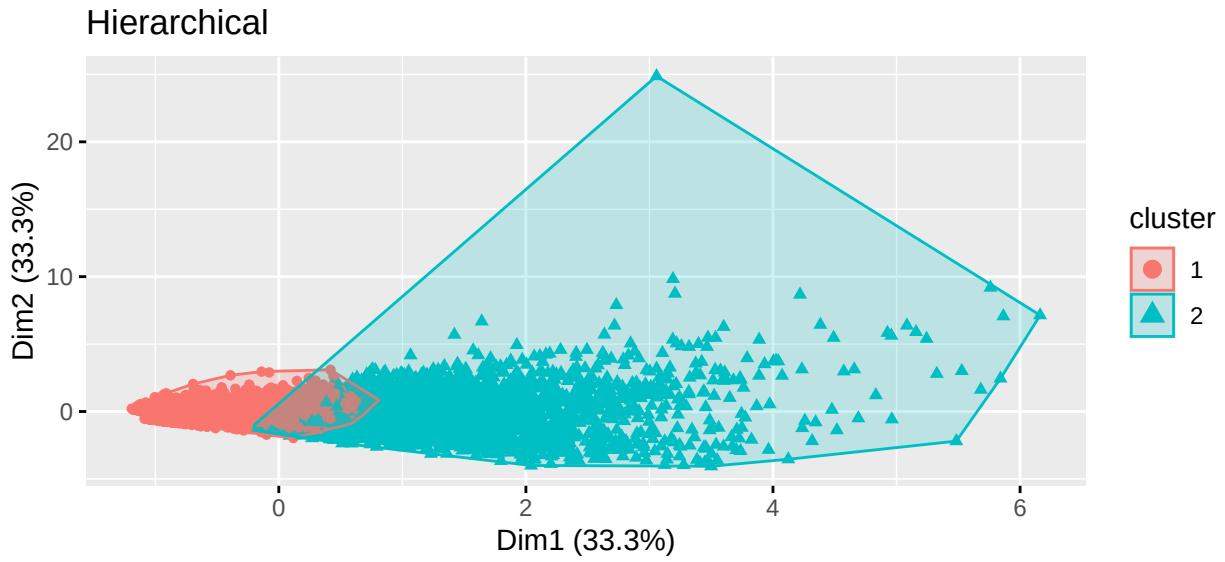
Using the silhuoette score we firstly find the optimum k number of clusters for each model, K-Means and Hierchical Clustering. We used the Euclidean distance for both clustering models due to its more intuitive geometric interpretation and widespread use in clustering analyses.

```
## [1] "Best k for K-Means: 2 , Best k for Hierarchical Clustering: 2"
```

Visualizing the partitions for K-means and Hierchical Clustering




```
## [1] "K-Means Average Silhouette width: 0.512229287473311"
```




```
## [1] "Hierarchical Clustering Average Silhouette width: 0.497794434086226"
```

We observe the PCs' mean for each cluster of each clustering model to understand more distinctively what group of houses each cluster represents.

PC statistics

```
##             Cluster    PC1    PC2    PC3
## 1 K-Means Cluster 1 -0.925 -0.008 -0.018
## 2 K-Means Cluster 2  2.879  0.026  0.055
```

```
##                   Cluster    PC1    PC2    PC3
## 1 Hierarchical Cluster 1 -0.940 -0.063 -0.049
## 2 Hierarchical Cluster 2  2.717  0.181  0.142
```

Both K-Means and Hierarchical Clustering display two clusters with similar partitioning and characteristics. In K-Means, the two clusters differ most in PC1, with cluster 1 having a negative value of -0.925 and cluster 2 a positive value of 2.879. Similarly, in Hierarchical Clustering, cluster 1 has a PC1 value of -0.940 and cluster 2 a value of 2.717. PC1's top contributing features—rent amount, fire insurance, and area—suggest that cluster 1 represents smaller, lower-value houses, while cluster 2 represents larger, higher-value houses.

Further examining PC2 and PC3, although the differences are smaller, cluster 1 has negative values and cluster 2 has positive values in both clustering models. PC2 is mainly influenced by furniture not furnished, rooms, and area, while PC3 is influenced by HOA, property tax, and fire insurance. Despite the unexpected negative contribution of 'furniture not furnished' in cluster 2, whilst having a higher mean rent in furnished houses, the rest of the features in PC2 and PC3 stay consistent with our assumption.

We opt for K-Means for further analysis due to its higher silhouette score of 0.512 as well as its more distinct partioning contrasted by the slight overlapping partitioning from hierchical clustering.

Cluster properties

Firstly, from the assumption we made we decide to call cluster 1 and 2, Low value and High value houses respectively. Now we take a closer look at house characteristics.

```
##             Cluster Average_Rent_Amount Average_Area
## 1  Low Value houses            2396.730       88.820
## 2 High Value Houses            8416.119      298.758
##   Average_Furniture_not_Furnished
## 1                            0.78
## 2                            0.68
```

From the results we can distinctly tell the two groups apart as Low value houses have a much lower average rent amount as well as area than high value houses. One further point to note, is that since we had unexpected result on furniture not furnished, we looked into it further and found out indeed that high value houses had a lower average of furniture not furnished than lower value houses, which aligns with our assumption. Furthermore we decide to compare our clusterization with available labels that were not included in the clustering process such as city and animal.

```
## [1] "Summary of houses in each city for Cluster 1"
##             City Count Percentage
## 1 Belo Horizonte   873      11.37
## 2       Campinas   718       9.35
## 3   Porto Alegre  1037      13.51
## 4 Rio de Janeiro  1203      15.67
## 5      São Paulo  3845      50.09
```

```
## [1] "Summary of houses in each city for Cluster 2"
##             City Count Percentage
## 1 Belo Horizonte   274      11.11
## 2       Campinas    96       3.89
## 3   Porto Alegre    95       3.85
## 4 Rio de Janeiro   218       8.84
## 5      São Paulo  1783      72.30
```

Task 2 Conclusions

In conclusion, our analysis reveals two primary clusters of houses: Low Value Houses and High Value Houses. As expected, Low Value Houses tend to have smaller living areas, fewer rooms, and lower overall costs. In contrast, High Value Houses feature larger living spaces, more rooms, and significantly higher prices.

We further compared our clusters with external labels that were not included in the clustering process and unearthed some insights on the distribution of houses across different cities. We firstly discovered that cluster 1 (Low value houses) comprises a much larger percentage of houses from the dataframe than cluster 2 (high value houses). This distribution is logical, as the demand for affordable housing is generally higher than that for expensive properties.

Although houses in both clusters are predominantly located in São Paulo, we recall from the distribution table that the city with the most houses registered in the data frame is São Paulo by a large margin therefore we account for that bias. Dominance is especially pronounced in Cluster 2, where 72% of the houses are situated in São Paulo. This likely reflects São Paulo's status as Brazil's largest city and financial hub. On the other hand, Cluster 1 shows a more even distribution of houses across various cities, indicating that affordable housing is more widely available throughout Brazil.

We also explored the relationship between our clusters and whether houses accepted animals, but this factor did not yield any significant insights.

From a strategic perspective, for a new company entering the real estate market, São Paulo emerges as a critical focal point due to the high concentration of high-value houses and the overall volume of properties available. However, given the competitive nature of São Paulo's market, new entrants might also consider expanding their operations to less competitive regions where they can acquire properties at lower costs and tap into the broader market for affordable housing.