R Rated Final Project - Brazilian Houses

Janhvi Goje, Ayham Alroumi, Davoud Danish

2024-05-07

Purpose

We pick the 'Brazilian Houses' dataset.

The goal of this project is to conduct a thorough data-driven analysis of the rent prices in different Brazilian cities in order to help a new company understand what kind of houses grant the larger (rent) revenue before investing themselves, in particular focusing on determining characteristics that correlate with higher rent prices. Additionally, we seek to explore whether rental markets can be segmented into distinct groups based on geographical location.

Description of the Dataset

This dataset encompasses 10,962 rental properties across various Brazilian cities and includes 12 distinct features, aimed at providing insights into the house rental market in some of Brazil's key cities.

Five of these variables are continuous: fire insurance, property tax, HOA fee, rent amount and area; while the remianing three are categorical: animal accepted,apartment furnished, and city (where property is located).

Data Cleaning

Let's start of by cleaning the dataset by looking at any NAs or duplicates present.

```
anyDuplicated(Data)
```

```
## [1] 245
```

```
anyNA(Data)
```

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

We notice that there are duplicates and no NAs, therefore we proceed by removing the duplicates.

```
Data <- unique(Data)
```

However, we do notice that in spite of no missing data in the form of 'NA' is present, there are '-' present in floor. Let's check how many of these values are present.

```
## [1] "Count for floor = '-': 2369"
```

We have 2369 observations with floor '-' in our dataset. This might suggest that the '-' symbol likely represents properties located on the ground floor or standalone houses that are not part of a multi-story condominium. If the later were to be true afterall, it makes sense for the 'HOA' variable (Monthly Homeowners Association Tax) to not exist since it is a measure of a fee paid by residents in order to cover the cost of maintaining common areas, building amenities, and utilities. Let's check this in order to verify if the assumption we're making is appropriate or not.

```
## [1] "Number of observations with floor = '-' and HOA = 0: 2015"
```

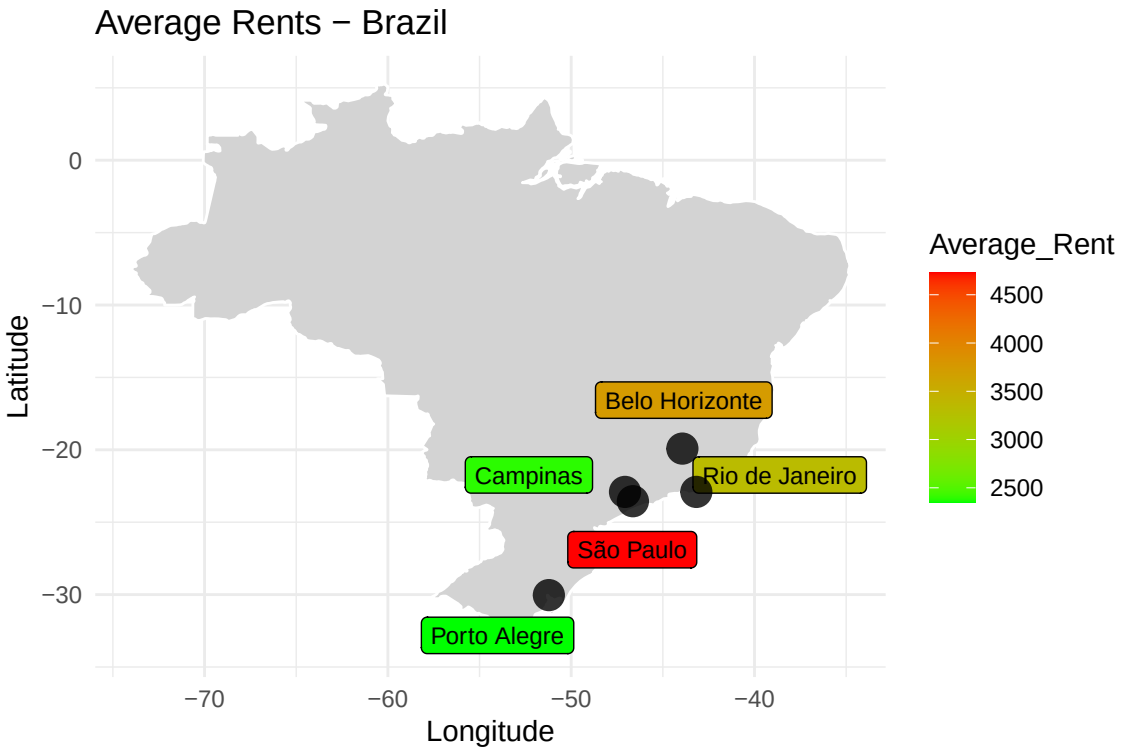
```
## [1] "Proportion of HOA = 0: 85.0569860700718 %"
```

Around 85% of the '-' observations have HOA as 0 which goes to show that our assumption is valid and therefore, it is appropriate to replace '-' with 0 in floor.

```
Data[Data == "-"] <- 0
```

Lets now categorise variables as numeric, for continous and factor, for categorical.

Since we have the information on cities, it might be helpful to visualise the data in the form of a map. We therefore calculate the average rent prices as per the cities and visualise them using 'map_plot' function of 'ggplplot2'.



São Paulo seems to have the most expensive average rent prices as we can see. Belo Horizonte is next, followed by Rio de Janeiro, having comparatively lower rental prices while Campinas and Porto Alegre have the lowest rental prices.

Further, let's look at the correlations between certain factors

```
cor(Data$fire.insurance..R..,Data$rent.amount..R..)
```

```
## [1] 0.9872013
```

We notice that fire insurance has the highest correlation with the rent prices, indicating that fire insurance could be a significant variable for predictive modeling.

Task 1

The objective of our first task is to *Build a predictive model and find out the rent amount according to the house specifics*, by using regression methods on our response variable: *rent.amount..R..*

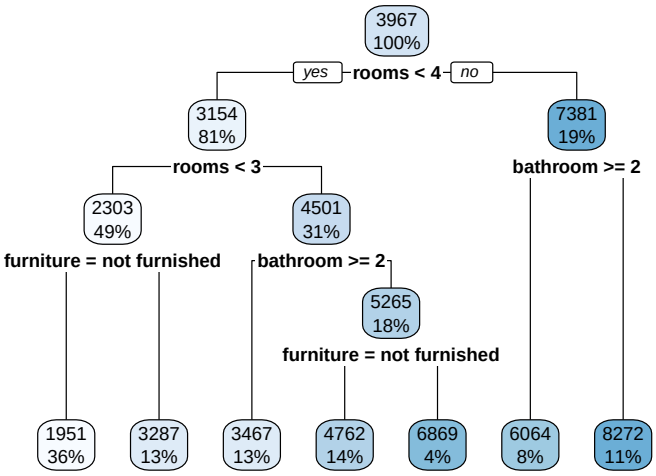
Lower Dimensional Models

Let's delve into the relationship of our response variable with other variables in order understand which variables are important for our case. First, we use a Log-Linear regression model to measure the percentage increase when there is a unit increase in area.

```
##                 Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) 7.9303056926 7.946043e-03 998.01942 0.000000e+00
## area        0.0002528586 1.401744e-05  18.03886 1.207089e-71
```

We see that there is a positive coefficient for area, suggesting a positive relationship between the area of a property and its rent price; as the area increases, the rent price also tends to increase. Given the logarithmic transformation of the rent, a coefficient of 0.0002528586 for the area implies a multiplicative effect on the original rent scale, i.e. approximately a 0.025% increase in rent for a one-unit increase in area.

Now, let's put ourselves in the perspective of the people wanting to rent a place in Brazil through the construction of a Decision Tree. This is integral in understanding the factors that influence a persons decision when selecting a house to rent.



This decision tree helps in understanding which factors significantly influence rental prices. Properties with more rooms and additional bathrooms generally command higher rents.The unfurnished status, particularly in properties with fewer rooms and bathrooms, often correlates with lower rent, suggesting that smaller, unfurnished properties are on the cheaper end of the rental market spectrum.

Following the right-most branches tends to associate with higher rent. This is depicted by properties having 4 or more rooms and possibly other features not fully visible in the cut-off image. Conversely, properties with less than 4 rooms and possibly fewer bathrooms or being not furnished are associated with cheaper rent, as seen in the left-most branches.

Data Pre-Processing

We begin by addressing outliers in our dataset, starting with the continuous variables, which we'll handle using the Interquartile Range (IQR) method. For discrete numerical variables, we will visually inspect boxplots to identify any outliers. This step allows us to manually scrutinize and decide on the removal of data points that appear significantly distant from the rest of the data distribution.

We proceed by encoding the categorical variables 'animal' and 'furniture' into binary numerical format. In the original dataset, these variables are represented as character strings, such as "accept" and "not accept" for 'animal'. To streamline data processing and facilitate statistical analysis, we convert these categories into 0s and 1s.

We divide the dataset into training, validation, and test subsets. We standardize these subsets using the mean and standard deviation calculated from the training set, ensuring that our model generalizes well on unseen data by using the same scale across all subsets. Additionally, we combine the training and validation sets into a primary training dataset, which we will later use for final model assessments and predictions on the test set.

Lastly, we encode categorical variables and keep an un-encoded version of the training and validation sets for later.

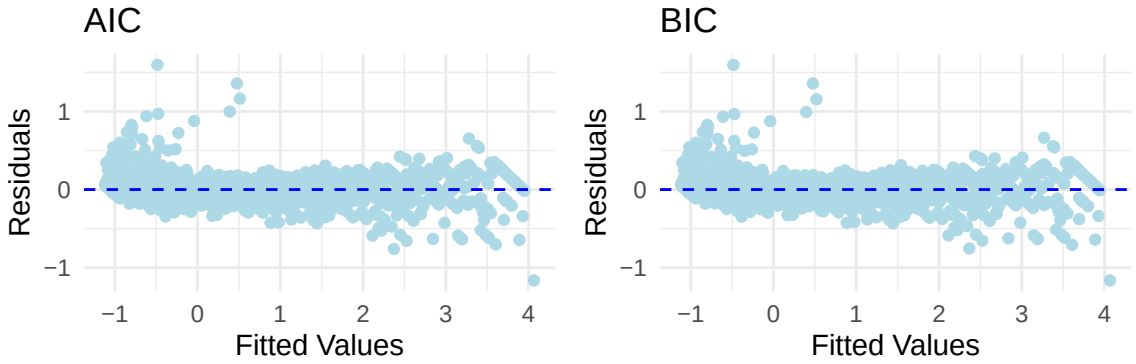
Constructing Models and Performance Assessment

AIC-BIC (Linear Model)

We implement AIC and BIC stepwise selection for multiple regression models and evaluate the performance on the validation set using **MSE**, **RMSE** and **R Squared**.

```
##   Model     MSE    RMSE R_squared
## 1   AIC 0.01437 0.11986   0.98842
## 2   BIC 0.01435 0.11978   0.98841
```

The two display similar results but BIC slightly better, so we're gonna prefer it over AIC. We can also plot the residuals against the fitted values:



The fact that we get a similar result for AIC and BIC is a good indication of balance between model complexity and goodness of fit.

Lasso and Group-Lasso (Penalised Approach)

We implement a lasso regression using glmnet and group-Lasso using gglasso.

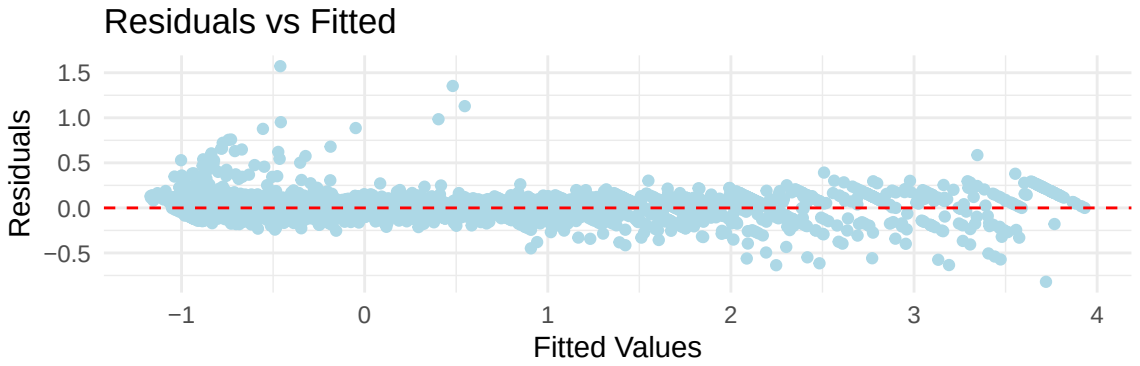
Let's compare the two models using minimum lambd since lambda.min values get a lower error on the validation set.

```
##     Model     MSE    RMSE R_squared
## 1   Lasso 0.01426 0.11942   0.98523
## 2 grLasso 0.01433 0.11969   0.98516
```

Once again, we notice that results are similar, however, Lasso has slightly lower errors and a higher R Squared value.

GAM (Non-linear Model)

We initially prepare the smoothing terms for both numerical and categorical variables in order to put up a GAM. In this stage, the training and validation sets that we previously conserved are used in their original, non-encoded forms. Next, we combine these smoothing terms to create the GAM formula. We then utilise this method to fit the model, which enables us to handle nonlinear relationships and variable interactions while efficiently analysing the data.



In this plot, the residuals predominantly cluster around the zero line, with most falling within the $[-0.5, 0.5]$ range, suggesting a generally strong fit of the model. Next, we will apply the model to the validation dataset and assess its performance using the designated metrics.

XGBoost

For XGBoost, we separate the dependent and independent variables. Following this, we convert the training and validation datasets into xgb.DMatrix format.

We tune the parameters `nrounds` (the number of boosting rounds), `max_depth` (the maximum tree depth) and `eta` (learning rate). We use 10-fold cross validation to do so.

Now, we use the optimal parameters to fit the finalised XGBoost model and make and assess predictions on the validation set.

Model Selection

```
##     Model     MSE    RMSE R_squared
## 1     BIC 0.01435 0.11978   0.98841
## 2   LASSO 0.01426 0.11942   0.98523
## 3     GAM 0.01308 0.11438   0.98644
## 4 XGBoost 0.00771 0.08781   0.99201
```

The above table helps us look at the performance of each model. We notice that *XGBoost* is best performing model and we hence, select it for our ultimate test set.

Model Implementation

We create the training set, by combining the initial training and validation set. We also scale and encode the categorial variables.

After doing so, we convert the training and test sets into xgb.Dmatrices to proceed with model fitting, and then predictions are made on the test set. It is important to evaluate these predictions (including for unscaled data for interpretability) and estimate a 95% confidence interval for the error, which will give an interval estimate of the possible deviation of the predicted rent prices from the true rent prices.

```
## [1] "MSE:  0.0128"
```

```
## [1] "RMSE:  0.08781 | RMSE for unscaled predictions 327.94212"
```

```
## [1] "R Squared:  0.99201"
```

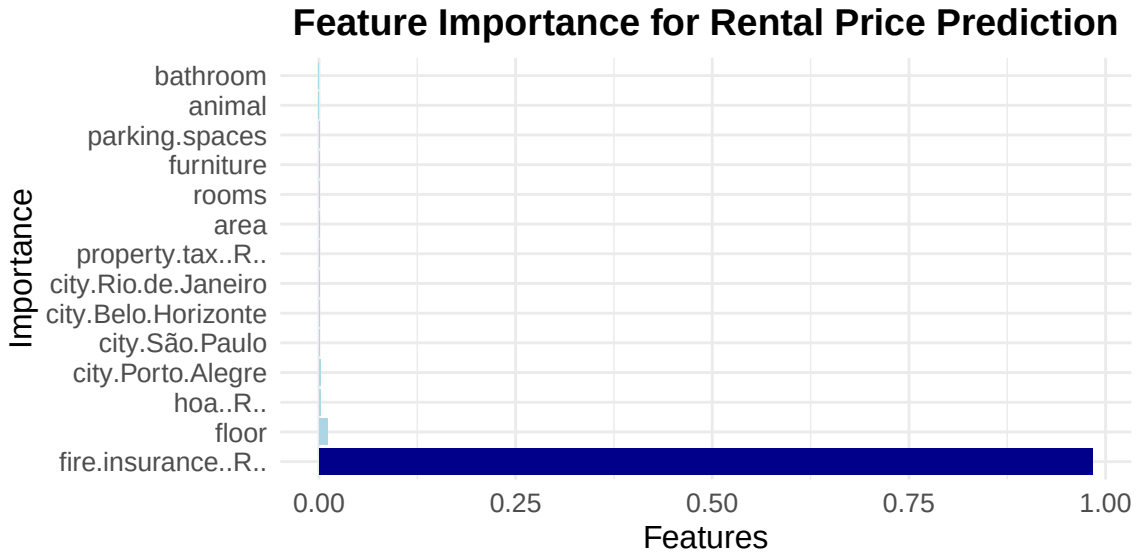
```
## Confidence Interval ( 95 %): [ -654.3607 ,  630.6373 ]
```

Our analysis demonstrates that the model consistently delivers robust performance on unseen datasets, as evidenced by the low error rates on the test set, which are comparable to those on the validation set. The confidence intervals calculated offer a prediction of the true errors associated with forecasting rent prices in Brazilian reais. Specifically, the predicted rent prices could deviate by up to **654.36 Real** below or **630.64**

Real above the actual rent prices, which is a favorable outcome considering the diversity of rental prices we are analyzing.

This level of precision in predicting rent prices would allow our company to make well-informed decisions about potential rental property investments in Brazil's major urban areas. By understanding the probable returns on these investments through accurate rent price predictions, we can strategically recommend investment in properties that promise higher returns, focusing on those with lower acquisition costs but high rental potential.

Moreover, the model's high R-squared value confirms that it effectively captures a significant portion of the variance in rental prices through the variables used. This indicates a robust relationship between how the chosen independent variables help predict rent prices, affirming the reliability of our model in predicting market behaviors based on these factors. Let's proceed with seeing what is the individual impact of these independent variables.

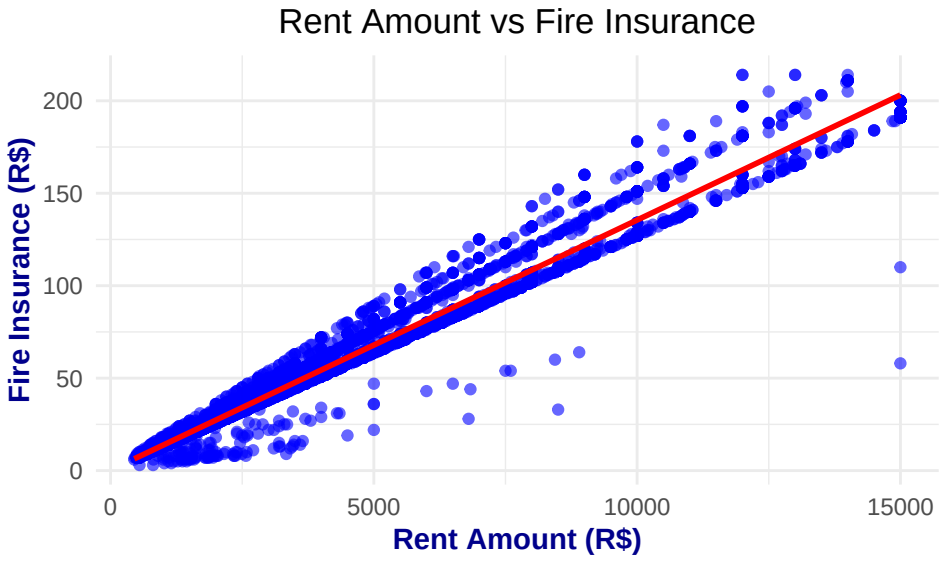


The feature importance plot clearly highlights that the variable `fire.insurance..R..` is the most influential in predicting rent prices, significantly outstripping other variables in its impact. This observation underscores the critical role of fire insurance costs in forecasting rental rates.

Additionally, the variable `city` also stands out for its importance, corroborating our earlier findings from a geospatial analysis. This analysis showed distinct variations in average rent prices across different cities, with Porto Alegre and São Paulo being particularly significant. Porto Alegre was noted for having the lowest average rents, whereas São Paulo exhibited the highest, thus underlining the substantial effect of location on rental costs.

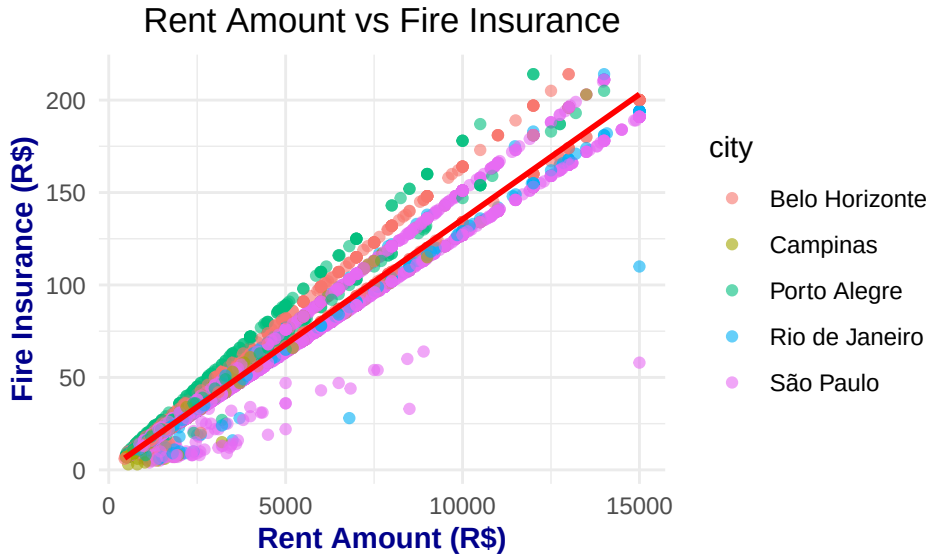
Other notable variables include `floor` and `hoa..R..`, indicating that these factors also contribute valuable insights into rent price predictions. Interestingly, factors such as the house's area and the number of rooms, which one might typically assume to be crucial, appear to have less influence than expected. This phenomenon suggests that the location of a property might overshadow its size or number of rooms in determining rental prices. For example, renting a larger house in Porto Alegre might be more affordable than renting a smaller apartment in central São Paulo.

Given the significant role highlighted for `fire.insurance..R..`, let's explore it further using a scatter plot.



The scatter plot reveals a pronounced linear relationship between fire insurance costs and rent amounts, which is not unexpected. Fire insurance premiums are frequently determined based on the property's value and related risk factors. Consequently, properties commanding higher rents are usually of greater value. These properties naturally have higher replacement costs, which in turn elevate the fire insurance premiums.

Additionally, it's important to note that the location of a property (City) can also impact insurance premiums. The presence of multiple distinct lines in the plot suggests that different cities may be represented by each line. Let's look at it.

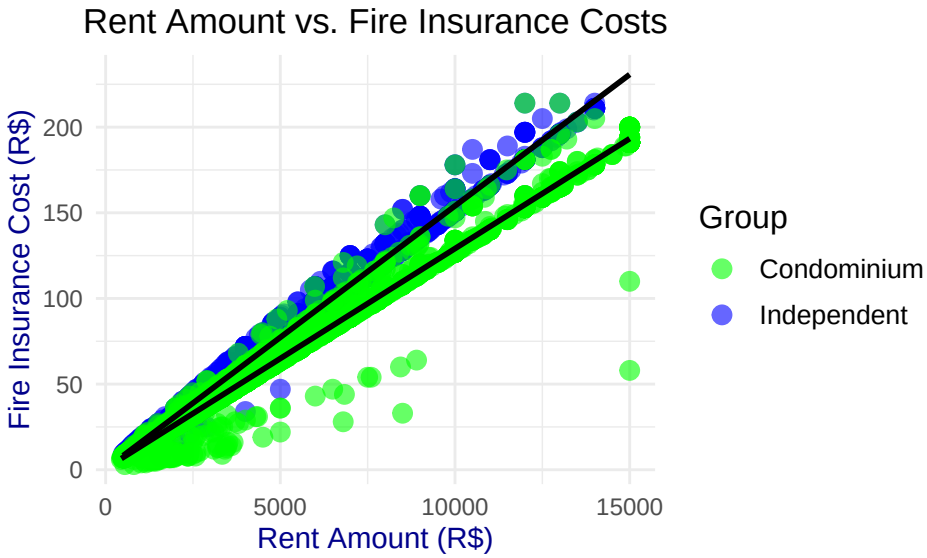


Upon analyzing the scatter plot, it was noted that cities with lower average rent values exhibited steeper slopes on the lines representing them. This suggests that, for properties of equal value, those in cities with lower rent averages would face higher fire insurance costs. This occurs because properties in these lower-rent cities tend to be larger in terms of area and room count, leading to higher replacement costs and, consequently, higher insurance premiums.

Moreover, the plot shows two distinct lines for each city, indicating the presence of another significant factor influencing both rent amounts and fire insurance costs. This factor appears to be related to the variables `floor` and `hoa..R..`, which were significant in our predictive model. A key determinant here is whether the property is located within a condominium.

For properties, a `floor` value of 0 could mean it is either on the ground floor of a condominium or it is a standalone house. The determining factor between these scenarios is the HOA fee. A zero HOA fee suggests the property is a standalone house, whereas a nonzero HOA fee indicates the property is part of a condominium.

To delve deeper into this relationship, we propose creating a focused plot that differentiates between houses located within condominiums and independent houses. This will allow us to better understand how these factors interact in specific urban contexts.



The scatter plot clearly supports our hypothesis, indicating that independent house renters incur higher fire insurance costs than apartment renters in condominiums, even at equivalent rent levels. This discrepancy may stem from the absence of shared fire safety features in independent houses that are typically present in condominiums.

Such features in apartments help mitigate fire risks and, consequently, lower insurance premiums. Furthermore, independent houses often require additional insurance coverage for separate structures like garages, contributing to the increased premium costs.

Drawing Conclusions - Task 1

To summarise the conclusions from our analysis, a strategic investment approach that encompasses both independent houses and condominium apartments could effectively cater to a broad spectrum of rental market demands. This strategy enables the company to meet the varying preferences and priorities of different renters. For instance, those who value privacy, ample space, and freedom may prefer renting independent houses, despite potentially higher rent and fire insurance costs and locations further from urban centers. Conversely, renters seeking affordability may be drawn to apartments within condominiums.

The company ought to conduct a thorough evaluation of potential investment returns, considering factors such as location, initial investment costs, associated risks, and available capital. Our predictive model, which reliably estimates rent prices, proves to be an indispensable tool in this analytical process, aiding the company in making informed investment decisions.

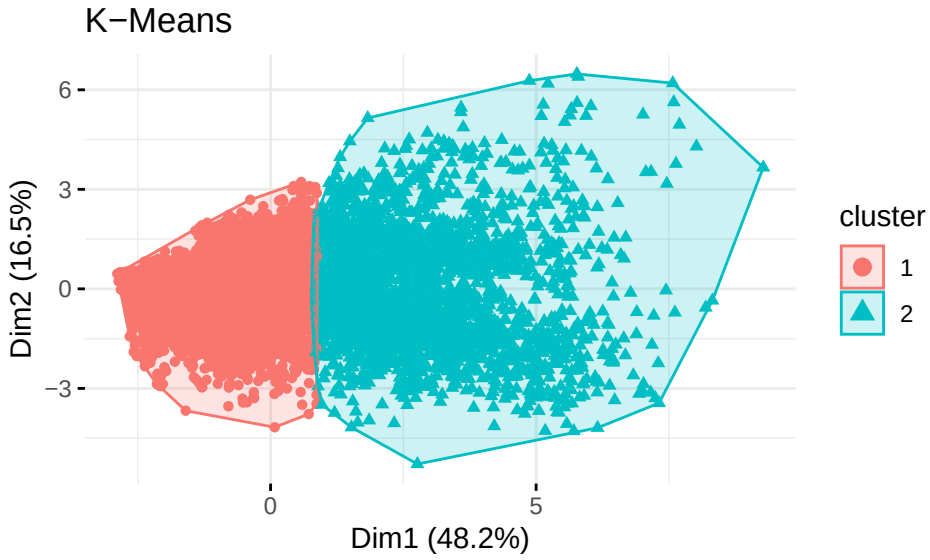
Task 2

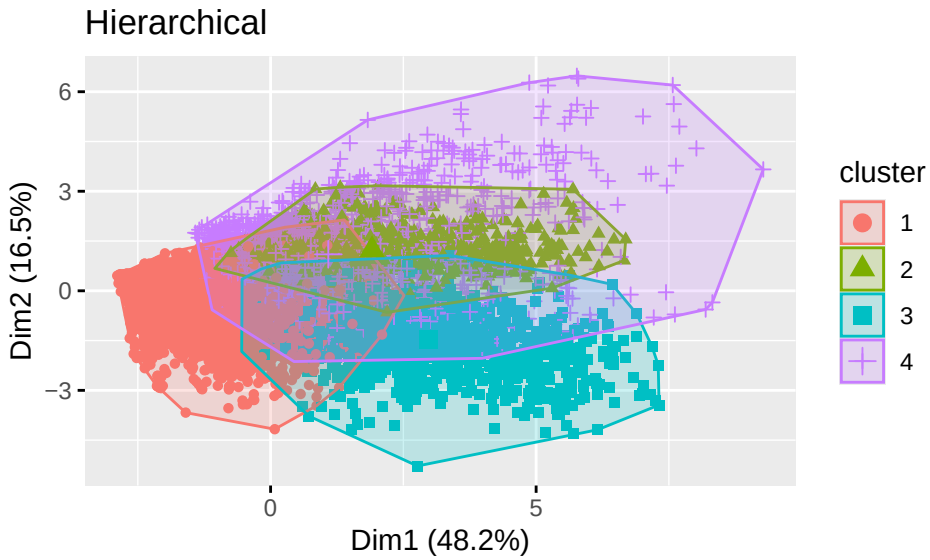
The objective of our second task is to *Cluster the houses for rental according to their characteristics*. We use K-Means and Hierarchical Clustering to do so and select the optimal number of k using the silhouette method.

```r
# removing categoral variables
Data_sc <- Data %>% select_if(is.numeric)
Data_sc <- as.data.frame(scale(Data_sc))

dist_p <- factoextra::get_dist(Data_sc,method = "pearson")
dist_e <- factoextra::get_dist(Data_sc,method = "euclidean")

set.seed(444)
```

```
## [1] "Best k for K-Means: 2 ,Best k for for Hierachial Clustering: 4"
```

We visualise the partitions for both methods: K-Means and Hierarchial Clustering.






```
## [1] "K-Means Average Silhouette width: 0.391492088470083"
```

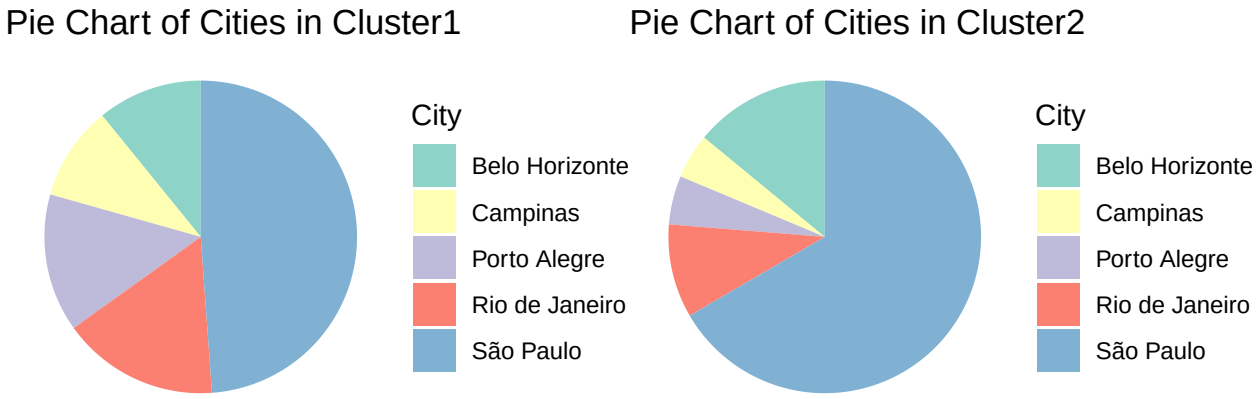
```
## [1] "Hierarchical Clustering Average Silhouette width: 0.308704751306791"
```

We notice that with hierarchial clustering we got some issues. As we can see above, the number of K is 4 which resulted in overlapping and further, it was also difficult to draw a direct conclusion on the basis of characteristics that determine what cluster a data point belongs to. Further, K-Means even has a higher

average silhouette width and for these reasons, we choose **K-Means as our choice of clustering** method and base our further analysis on the same. This clustering is valid as we can assume it to base clusters to determine two groups: affordable housing and expensive housing. Let's explore these clusters further.

```
##     Cluster Avg_Rent_Amount Avg_Property_Tax Avg_Rent_Per_Room
## 1 Cluster 1        2208.685          106.086          1277.183
## 2 Cluster 2        6968.054          583.458          2159.968
```

We make the very interesting observation that the average rent amount in cluster 2 is significantly higher than that of cluster 1 which as priorly mentioned, signifiess lower and higher end housing. We notice the same for average property tax and average rent per room where cluster 2 is prominently more 'expensive' than cluster 1. Let's lastly visualise if this aligns with our very first visualisation of certain cities being more expensive than others.



Drawing Conclusions - Task 2

Analyzing the data reveals distinct clusters representing different segments of the housing market in the region of Brazil under study, with a significant concentration of properties in Sao Paulo, reflecting both high and low-end housing options. The observed patterns indicate a correlation between property tax, rent per room, and overall rent prices within the identified clusters.

Considering business objectives, if the company seeks to expand its real estate holdings with a focus on *affordability*, targeting properties within *cluster 1* would be advantageous. These properties likely offer lower acquisition costs, enabling the company to scale its portfolio while maintaining competitive rental rates, appealing to budget-conscious renters.

Alternatively, if the company aims to emphasize premium properties with *higher rental potential*, prioritizing acquisitions within *cluster 2* is recommended. Despite potentially higher acquisition costs, these properties are positioned to yield greater rental income, aligning with a strategy geared towards luxury housing markets.

For portfolio diversification, leveraging insights from both clusters facilitates a balanced approach. By considering properties across a spectrum of cost ranges, the company can broaden its rental offerings, catering to diverse tenant preferences and optimizing investment returns.