ESTIMATION OF RENTAL PRICES IN MADRID

GROUP 3, SECTION 01: Alexandre Olivier, Anchal Jaiswal, Asier Sarasua Amundarain, Ignacio Rupérez Larrea, Jina Kim, Nicolas Clemens Linsenmaier, Rahul Singh

Executive Summary

The number of property sales transactions in Spain during 2016 is up significantly (14%) compared to 2015 with greatest increases being seen in the Madrid, Barcelona, Girona and Valencia. According to the latest price study of the leading Spanish property portal Idealista on prices for November 2017, the Madrid region has risen faster at + 1.5% than the rest of Spain where the average monthly increase was 1.2%. Two other provinces are still more expensive for housing: Guipúzcoa (2750 € / sqm) and Vizcaya (2,599 € / sqm), Catalunya (2,050 € / sqm). Moreover, foreign investors continue to be the key drivers of the Spanish prime property residential market representing 65% of sales during 2016 (Figure 1). These buyers cited ‘secondary residence’ as one of the main motivations for their purchases.

Approach: Our team decided to perform a regression analysis to predict real-estate rental prices and find good opportunities in the market for investments in flats to rent, which may be less than the theoretical prices. Our goal is to use Idealista data to provide an estimation of rental prices in Madrid city using certain attributes.

Dataset overview

The original dataset consisted of 2188 listings with 15 variables. Some address related variables were too specific to be used to fit a model, while additional variables were created to categorise listings better. These were the changes made to the dataset:

- Excluded variables: ‘Address’, ‘Number’, ‘Zone’. All these variables could be better represented by higher level variables such as ‘Area’
- New variables:
    - ‘Region’: higher-level division of area by the 4 directions and the center
    - ‘Inside_M30’: whether or not the listing is within the M-30 orbital motorway that surrounds the central Madrid area. Listings inside this motorway are considered in the most central area of Madrid, and thus prices may differ
    - ‘Studio’: 1 if ‘Bedrooms’ is 0
    - ‘basement_gf’: 1 if floor is -1 or 0, 1 for other floors. Highest floors are captured by the value ‘Penthouse’, but listings that are in the basement could equally have an effect on the rent
- Missing values
    - 181 values were missing, all from the variable ‘Floor’. 96 of them were cottages, which could be reasonably be thought as one story houses, and thus were imputed by value 1. All other 85 listings were dropped (5% of dataset)

The final analysis was finally conducted with 2007 listings and 16 variables.

Data Exploration

These were some of the important findings during the data exploration stage:

- Distribution of rent is positively skewed. The median price of rent is 1400, while there are a few outlier rent prices that pull the mean price to 1843.
- Most areas have rent prices lower than the average rent price, which makes sense due to the outliers pulling the mean price up. It is interesting to note that the areas with rent higher than the mean price have large variance in prices (Salamanca, Chamartin, Chamberi, Moncloa, Retiro), indicating that there are different factors that determine the prices in these areas. The ‘Region’ variable with higher-level area information show larger price variance for each region, indicating that it may not be the best variable to predict price.
- Most listings had 1, 2, or 3 bedrooms, with 2 bedrooms being the most common. Rent prices tend to increase by more number of bedrooms, but at the same time the variance of the price

increases too. This may be because absurd number of bedrooms (6+) may be for other uses (such as student rents) rather than for a single house owner.

- In general, a bigger house means higher rent price, and the pearson’s correlation r score is high with 0.83. However, there seems to be a lot more variation in prices in smaller houses.
- When it comes to rent prices by types of housing, there are too many outliers to safely conclude that different types of housing have a strong impact on the rent prices.
- In general, areas within the M-30 motorway have higher prices, and also have more outliers with the highest rent prices.
- Rent prices tend to be higher for listings that face towards the street (i.e., ‘outer’ direction), though with more variances.

Model Creation and Validation

Our final model that was derived after our analysis was composed of 5 independent explanatory variables:

$$\boxed{= -391.6766 + \beta_1 Sq..Mt. + \beta_2 Area + \beta_3 InsideM30 + \beta_4 Outer + \beta_5 Penthouse}$$

First, the dataset was divided into training and testing set in order to test and validate the model later, comparing the predicted and the actual rent prices. Next, the stepwise regression method was carried out to determine which variables were the most significant. We used the result of this analysis, together with the conclusion from the data exploration stage, to choose the most significant variables, which were: Sq..Mt., Area, Inside_M30, Outer, Penthouse, Cottage, Studio, Bedrooms and Duplex.

After Stepwise Regression, several models were created with these variables and ANOVA Tests were performed to identify the significant independent variables (with a 95% confidence interval). The non-significant variables were removed, and the variables which yielded the highest possible R squared were selected for further analysis. In the case of our final model, the adjusted R-squared equalled 0.7728 (see Figures 2 to 12 for detailed insight on different analysis performed).

Finally, we checked if our model accomplished the linear regression statistical assumptions:

- **Lack of correlation between independent variables**

According to the Variance Inflation Factor (VIF), none of the independent variables in our model has a VIF > 5, so we assumed that there is no collinearity among them.

- **Heteroscedasticity**

After the graphical analysis of the variance of residuals, we observed that the residuals did not follow any particular pattern and that the variation is evenly distributed. Additionally, according to the Global Linear Model Assumptions test, our model was not homoscedastic.

- **Normality of residuals**

Our graphical analysis showed that the residuals of our linear regression model followed a normal distribution. The slight variation from the normal distribution curve could be attributed to the presence of significant outliers.

- **Independence of residuals**

The Durbin-Watson test for our model was 1.77, which showed that there is a moderate positive correlation between the residuals of our model.

Description of variable coefficients:

The coefficients of the variables explain the prices of the listings as follows:

- An increase in 1 square meter size increases the price by 12.6 euros
- The listing being within the M-30 motorway increases the price by 401.2 euros
- The listing being located towards the street increases the price by 169.9 euros
- If the listing is a penthouse (located on the top floor), then the price increases by 243 euros
- When compared to the Arganzuela area, Centro, Chamartin, Chamberi, Moncloa, Retiro, and Salamanca had significantly higher rent prices. Other areas' rent prices were not significantly different according the model.

Conclusions & Recommendations

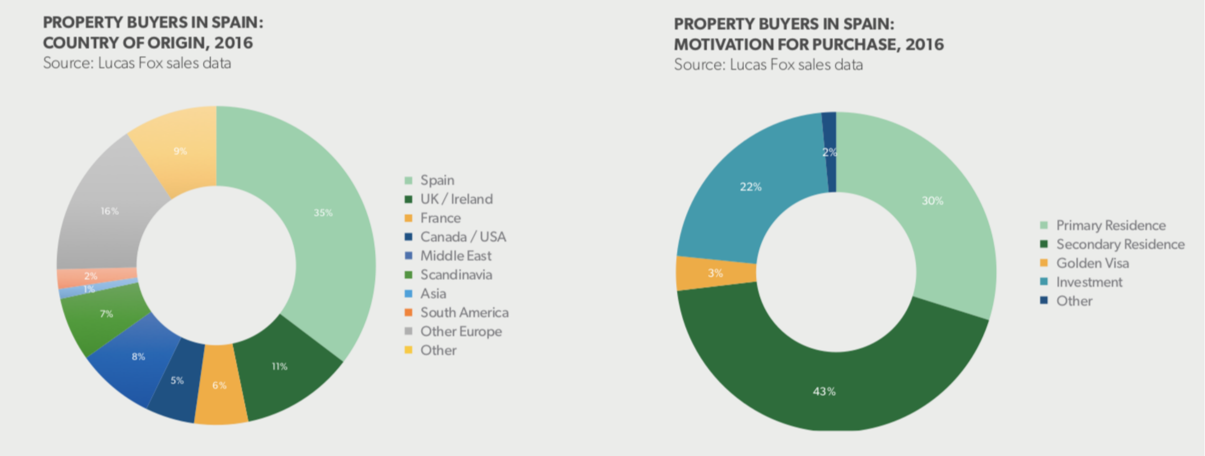
In conclusion, around 78% of rental prices in Madrid could be explained by observed variation in surface area, location (whether inside or outside the M30), orientation and listed category of the house (a penthouse, studio, etc.).

This model could be used by a real-estate agency, for example, to identify opportunities in the market. They would be able to choose whether or not to invest, rent or sublet a listing if the theoretical estimated price is higher than the actual market price. This would be an exciting opportunity for them as they could then rent it back under a 'normal' market price and earn profit out of it. For instance, they could find a flat in an expensive area with a large surface area and optimal orientation that does not follow the model and present an investment-opportunity for them.

Regarding improving the model, we recommend increasing the input dataset to add more observations and variables (such as inclusion of other variables may further explain the pricing too, such as age, condition, or the purpose of listing-student home or home-office space). The dataset used for analysis was relatively a small sample for a city like Madrid (from 2007) and should be updated to include recent years. To be able to create a more accurate model we recommend that the marketing team must collect more historical data as the real-estate market changes over time on factors such as micro- and macroeconomic elements for example.

Annexure

Figure 1: Property market overview in Spain



PROPERTY BUYERS IN SPAIN: COUNTRY OF ORIGIN, 2016
Source: Lucas Fox sales data

PROPERTY BUYERS IN SPAIN: MOTIVATION FOR PURCHASE, 2016
Source: Lucas Fox sales data

Table 1: list of variables in original dataset

Variable	Type	Meaning
id	Integer	Listing id
Area	Categorical	Name of the area
Address	Character	Street address
Number	Integer	Street number (not always completed)
Zone	Categorical	Sub category of Area
Rent	Integer	Monthly price of listing in euros
Bedroom	Integer	Number of rooms in the listing
Sq. Mt	Integer	Size of the listing in square meter
Floor	Integer	Floor number (not always completed or applicable)
Outer	Boolean	Takes value 1 if the listing is on the street side and 0 when on the inside of the building
Elevator	Boolean	Takes value 1 if there is an elevator and 0 if not
Penthouse	Boolean	Takes value 1 if the listing is a penthouse and 0 if not
Cottage	Boolean	Takes value 1 if the listing is a cottage and 0 if not
Duplex	Boolean	Takes value 1 if the listing is a duplex and 0 if not
Semi-detached	Boolean	Takes value 1 if the listing is semi-detached and 0 if not

Table 2: List of new variables included

Variable	**Type**	**Meaning**
Region	Categorical	Higher-level region information: North, South, East, West, Central
Inside_M30	Boolean	Takes value 1 if the listing is inside the M-30 motorway and 0 if not
basement_gf	Boolean	Takes value 1 if the listing is located on ground floor or below it and 0 if above ground floor

Figure 2: Distribution of rent price

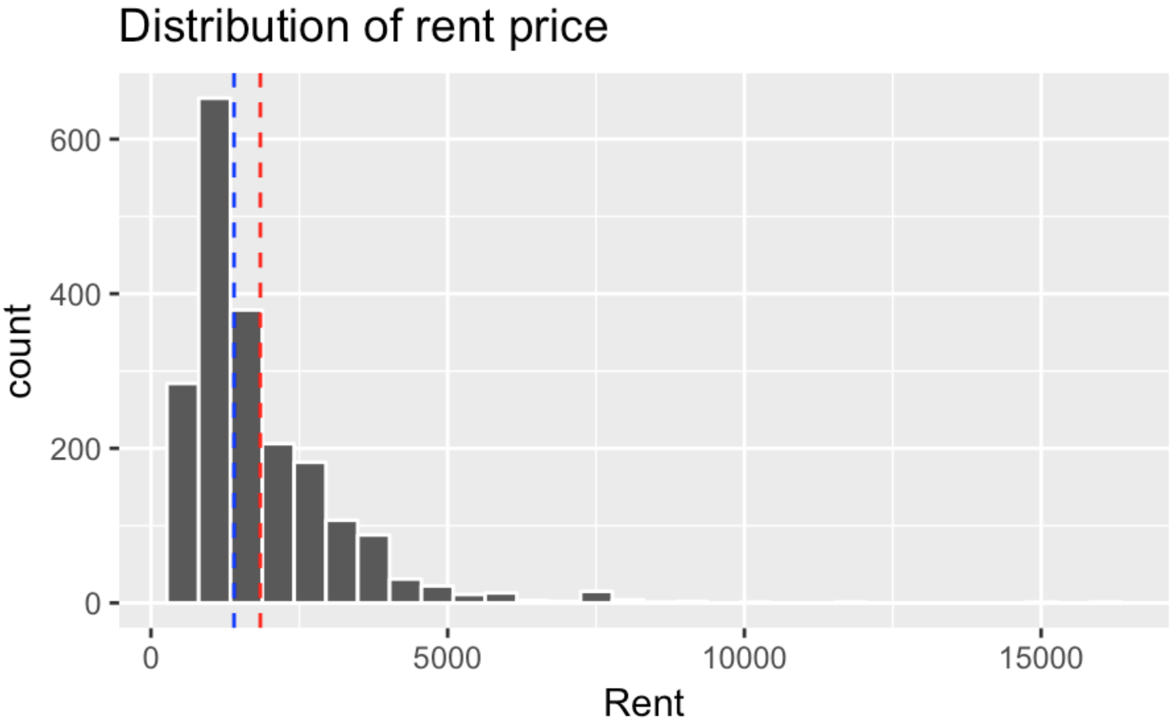


Figure 3: Distribution of rent price by area

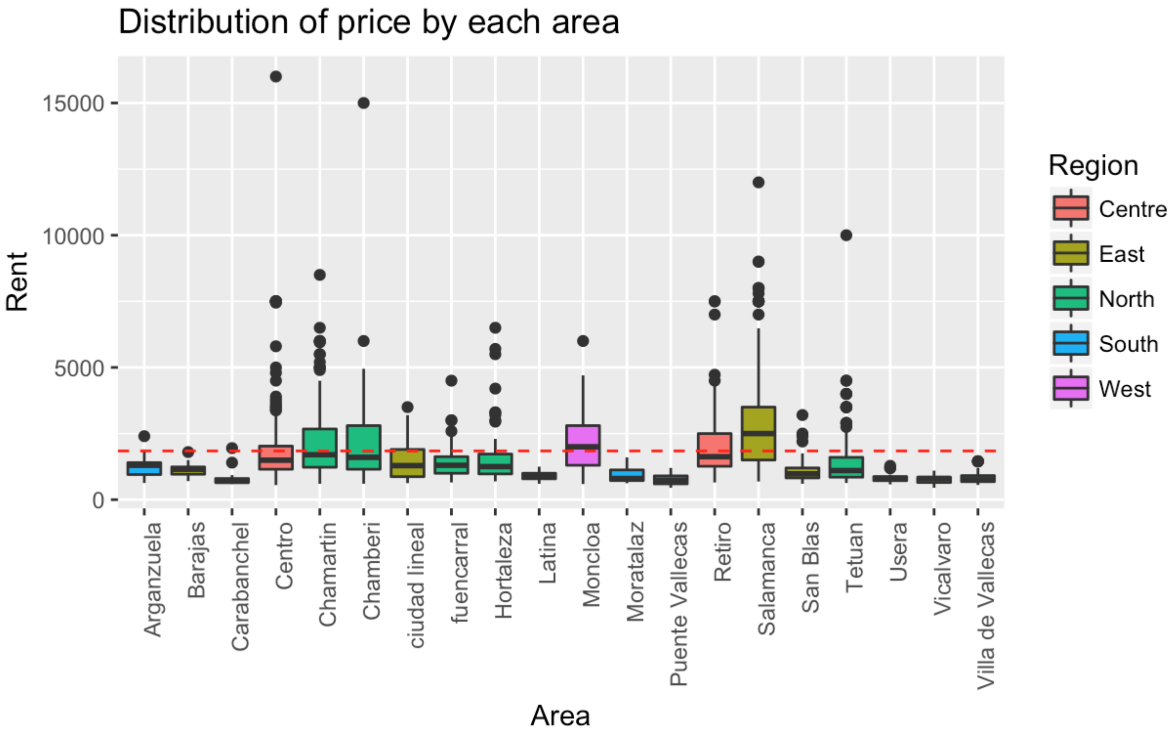


Figure 4: Distribution of rent price by region

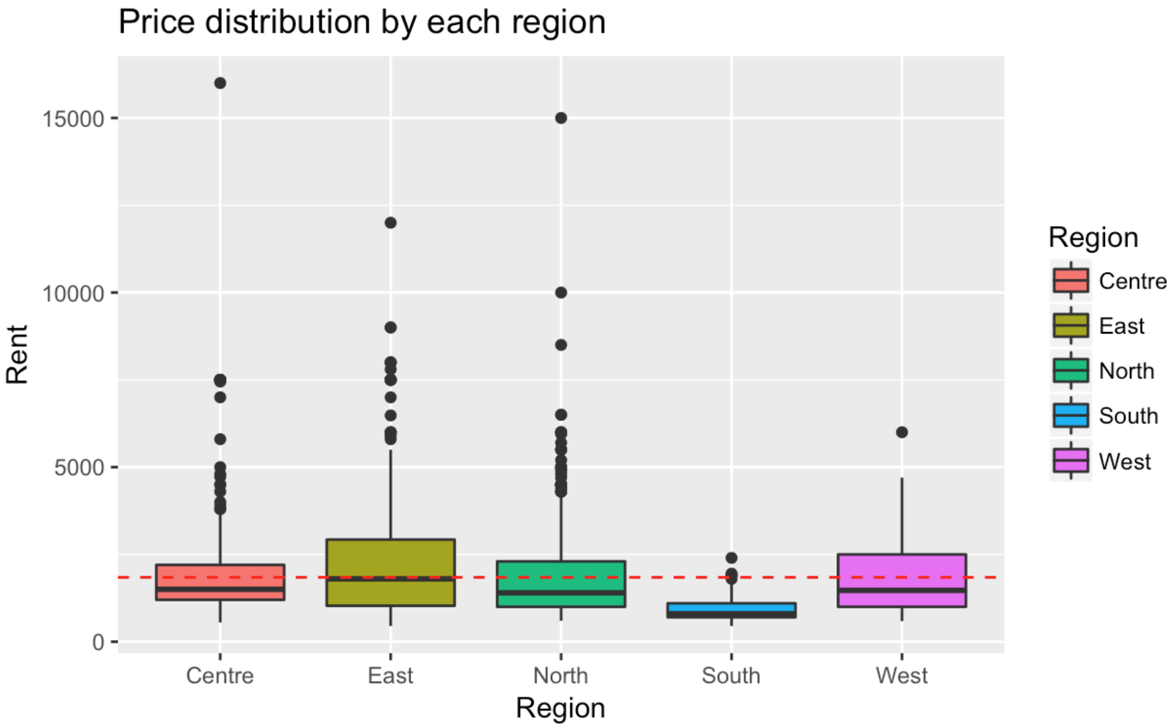


Figure 5: Number of bedrooms and rent distribution by each number

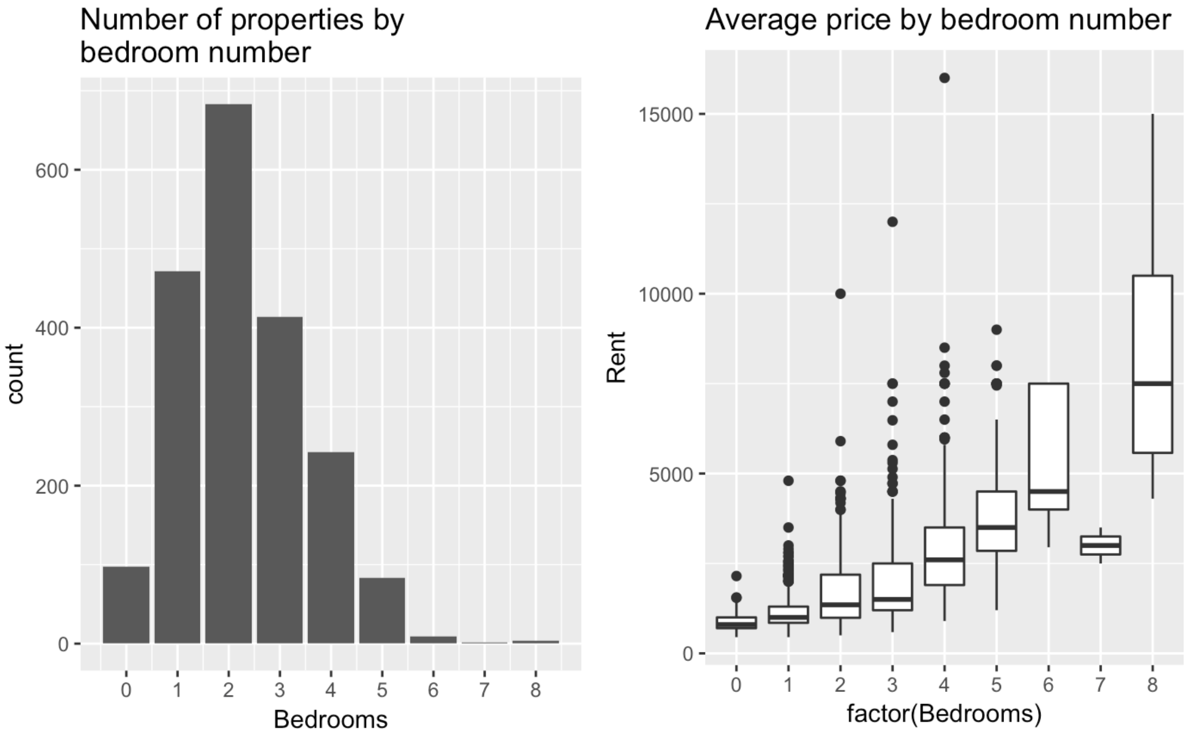


Figure 6: Rent price vs square meter size of listing

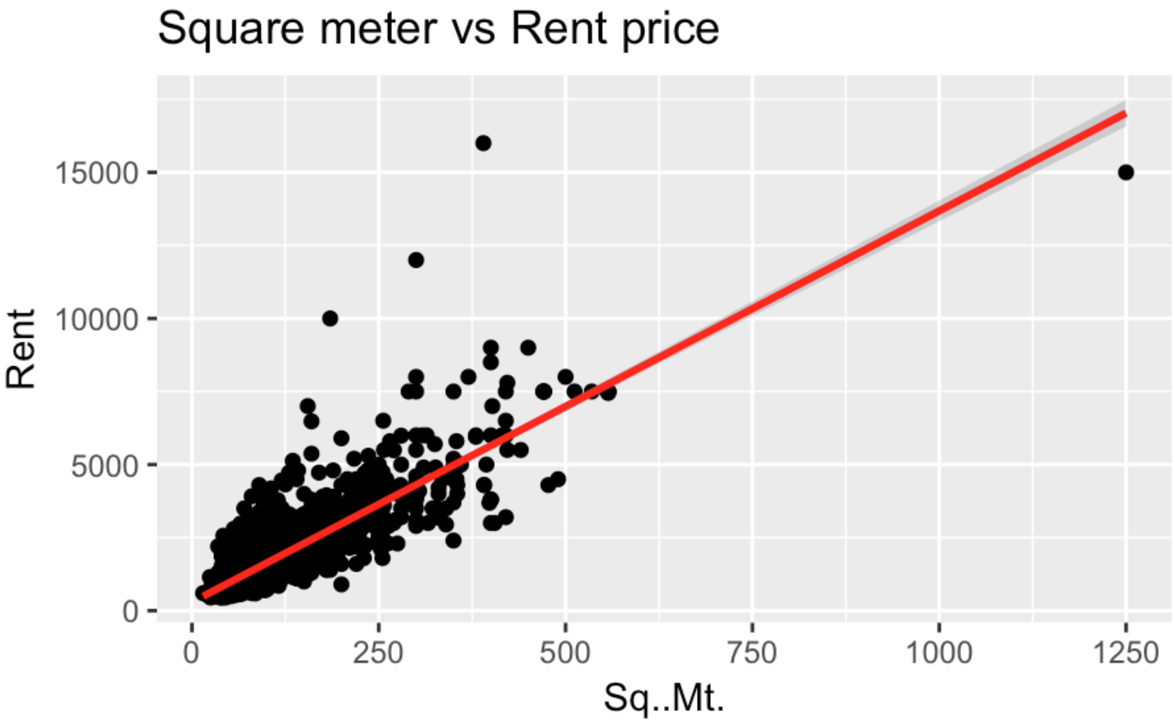
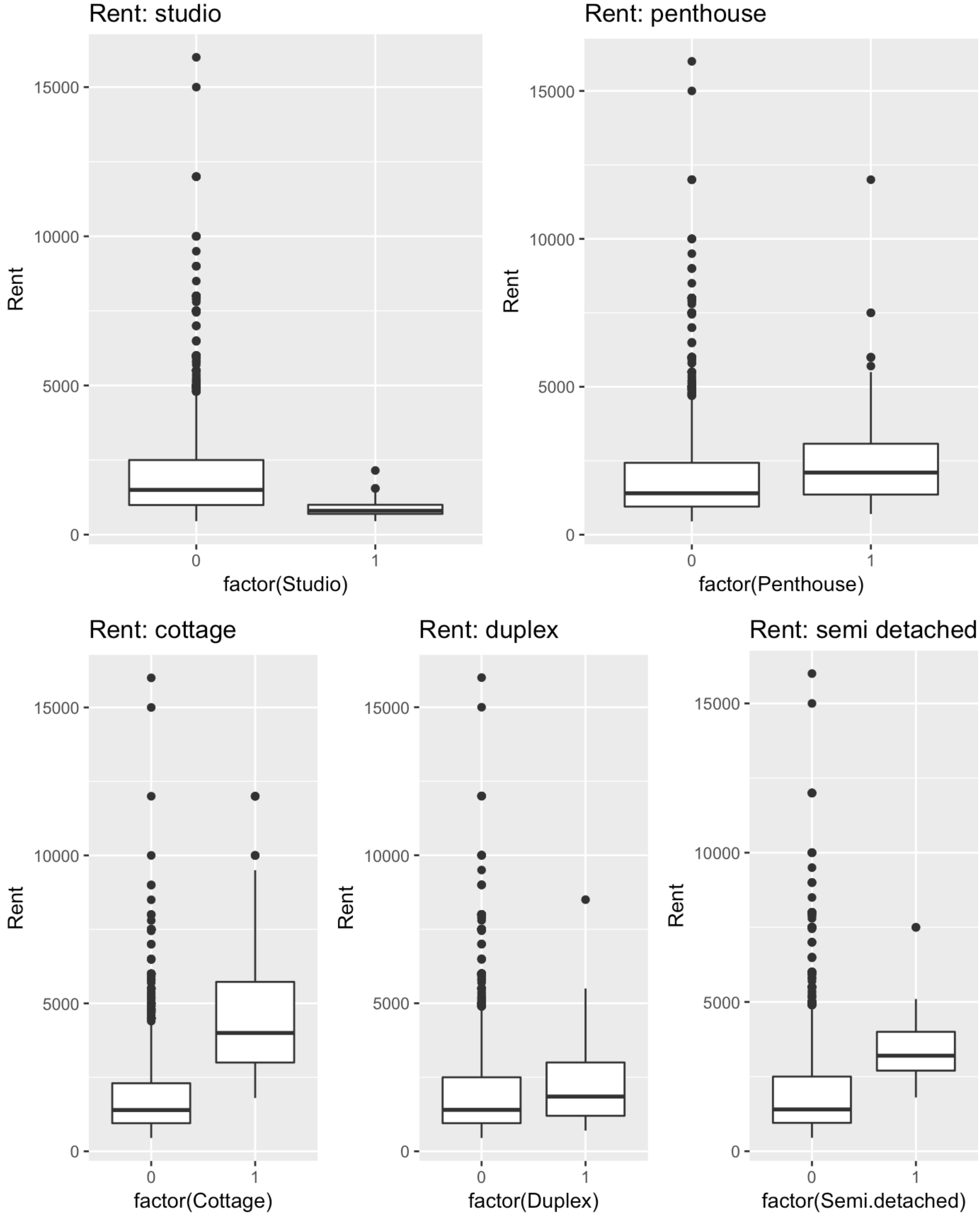


Figure 7: Distribution of price by each housing type



Rent: studio

Rent: penthouse

Rent: cottage

Rent: duplex

Rent: semi detached

Figure 8: Rent prices by area in and out of the M30 motorway

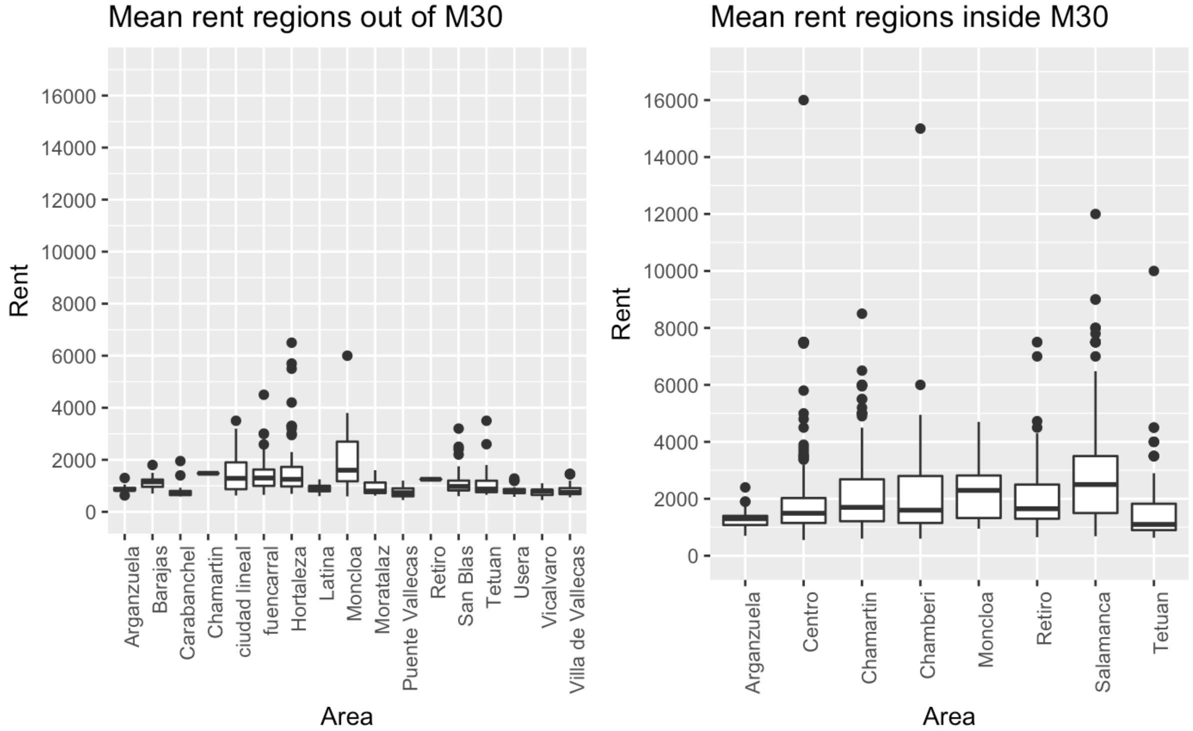


Figure 9: Rent prices for listings facing inwards vs outwards towards the street

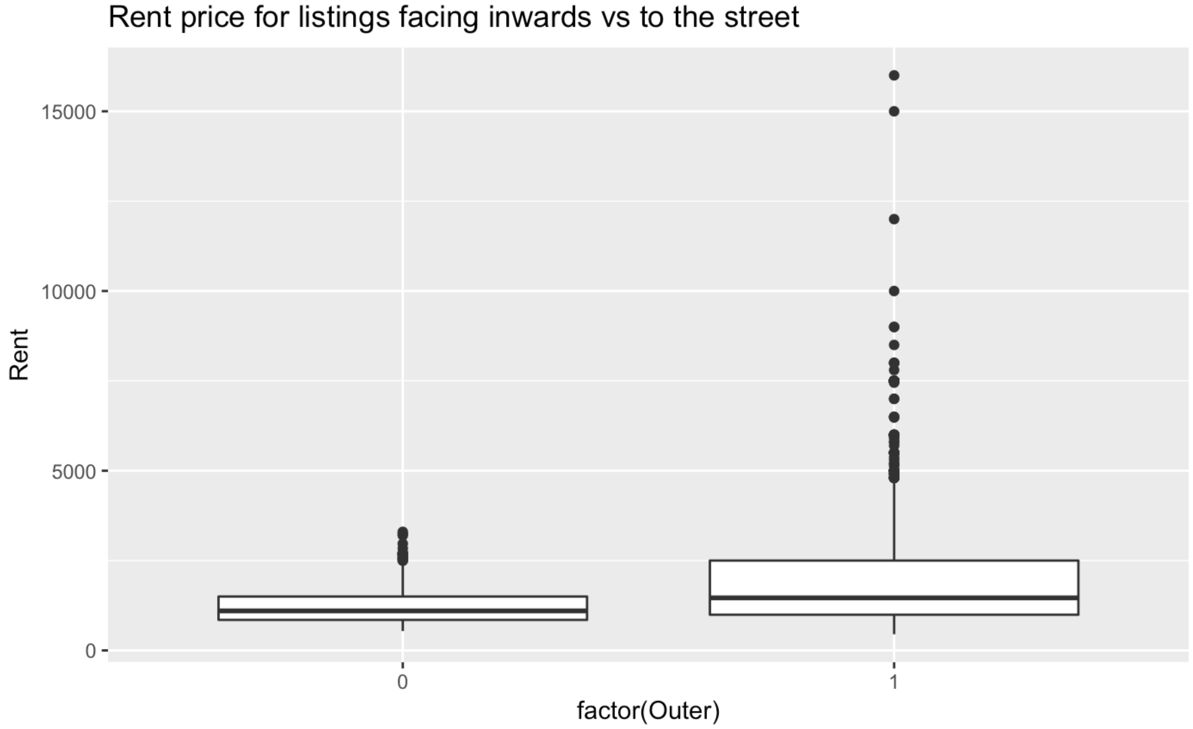
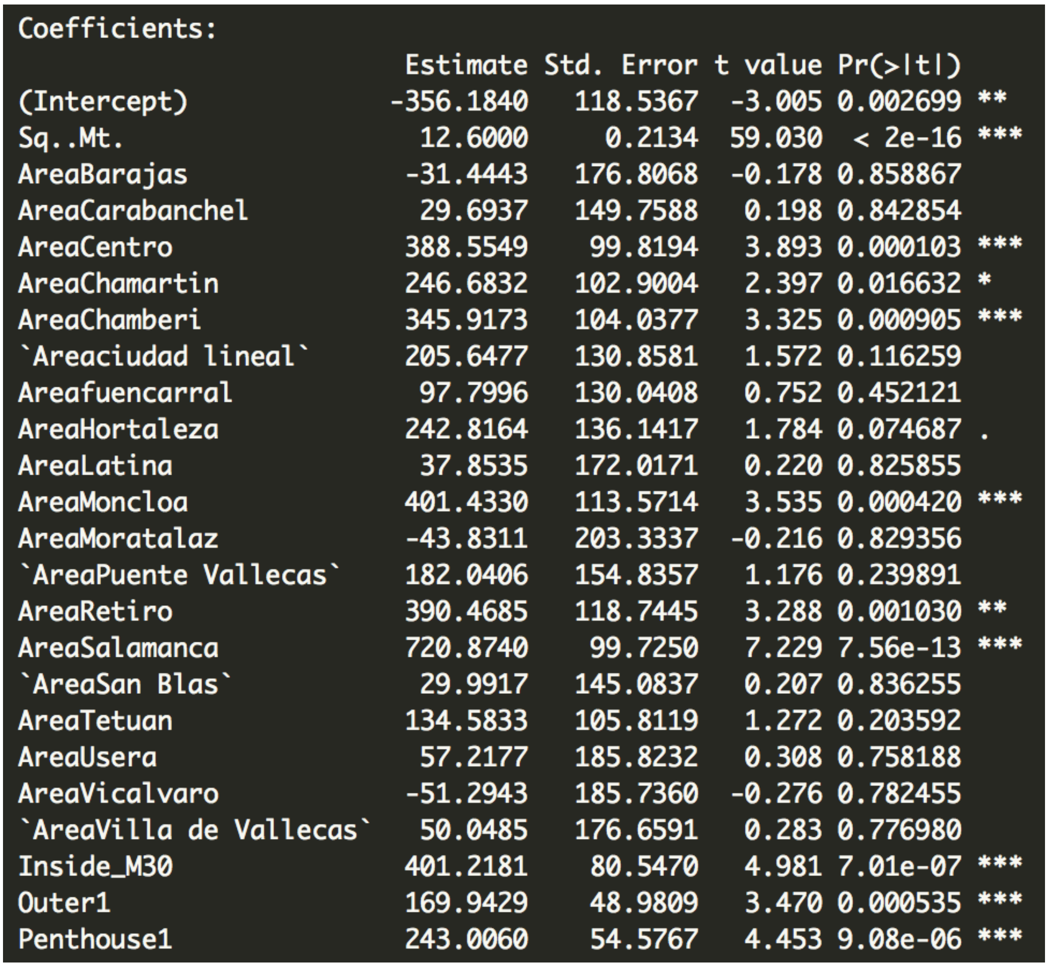


Table 3: Summary of variable coefficients



```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -356.1840   118.5367  -3.005 0.002699 ** 
Sq..Mt.                   12.6000     0.2134  59.030  < 2e-16 ***
AreaBarajas              -31.4443   176.8068  -0.178 0.858867    
AreaCarabanchel           29.6937   149.7588   0.198 0.842854    
AreaCentro               388.5549    99.8194   3.893 0.000103 ***
AreaChamartin            246.6832   102.9004   2.397 0.016632 *  
AreaChamberi             345.9173   104.0377   3.325 0.000905 ***
`Areaciudad lineal`      205.6477   130.8581   1.572 0.116259    
Areafuencarral            97.7996   130.0408   0.752 0.452121    
AreaHortaleza            242.8164   136.1417   1.784 0.074687 .  
AreaLatina                37.8535   172.0171   0.220 0.825855    
AreaMoncloa              401.4330   113.5714   3.535 0.000420 ***
AreaMoratalaz            -43.8311   203.3337  -0.216 0.829356    
`AreaPuente Vallecas`    182.0406   154.8357   1.176 0.239891    
AreaRetiro               390.4685   118.7445   3.288 0.001030 ** 
AreaSalamanca            720.8740    99.7250   7.229 7.56e-13 ***
`AreaSan Blas`            29.9917   145.0837   0.207 0.836255    
AreaTetuan               134.5833   105.8119   1.272 0.203592    
AreaUsera                 57.2177   185.8232   0.308 0.758188    
AreaVicalvaro            -51.2943   185.7360  -0.276 0.782455    
`AreaVilla de Vallecas`   50.0485   176.6591   0.283 0.776980    
Inside_M30               401.2181    80.5470   4.981 7.01e-07 ***
Outer1                   169.9429    48.9809   3.470 0.000535 ***
Penthouse1               243.0060    54.5767   4.453 9.08e-06 ***
```

Figure 10: Heteroscedasticity check

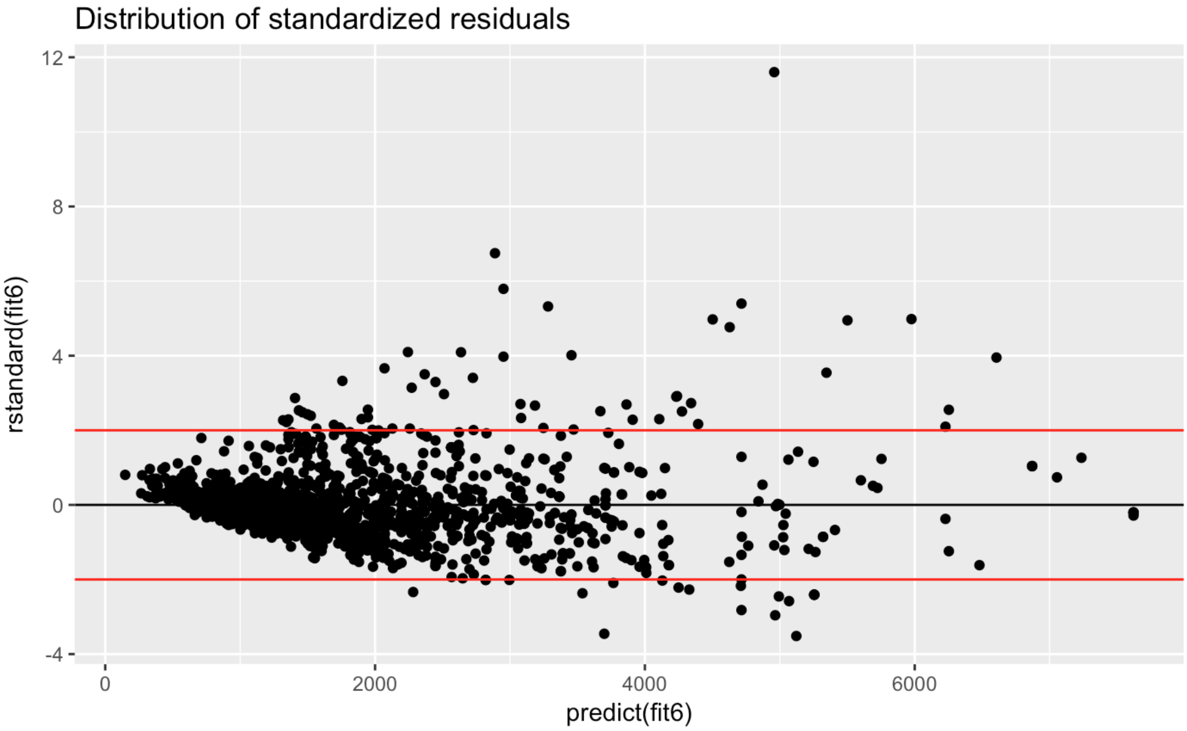


Figure 11: Residual normality test

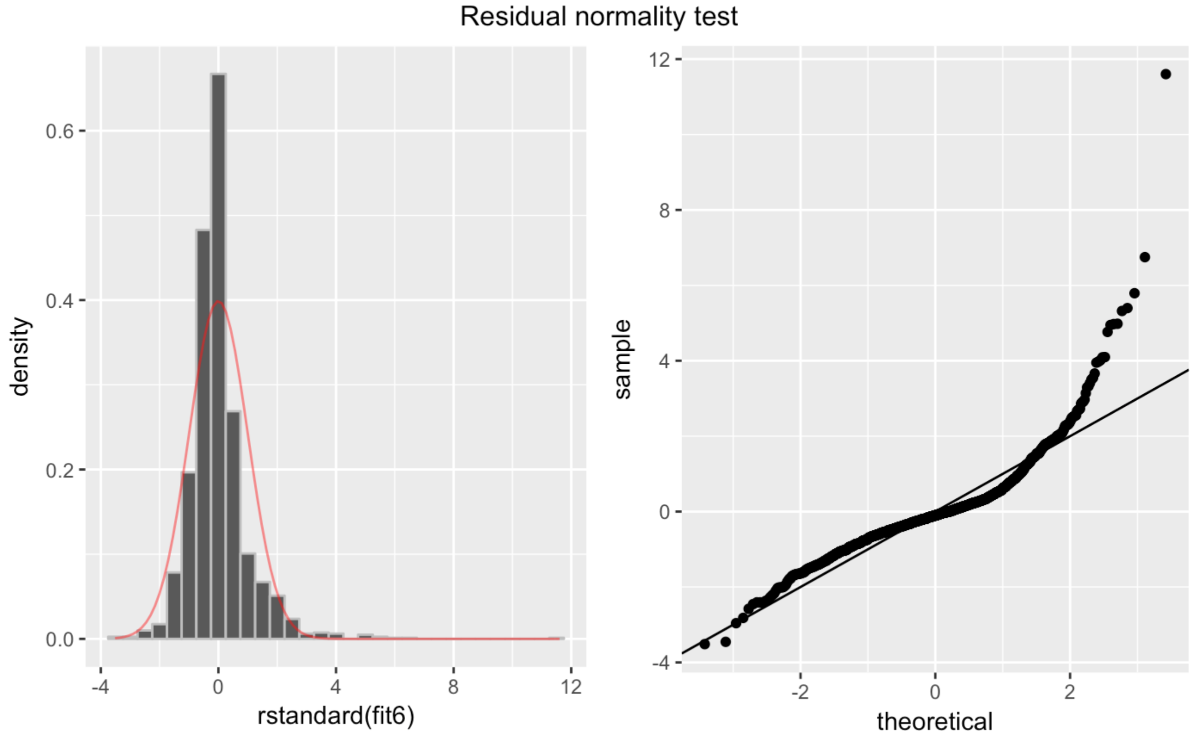
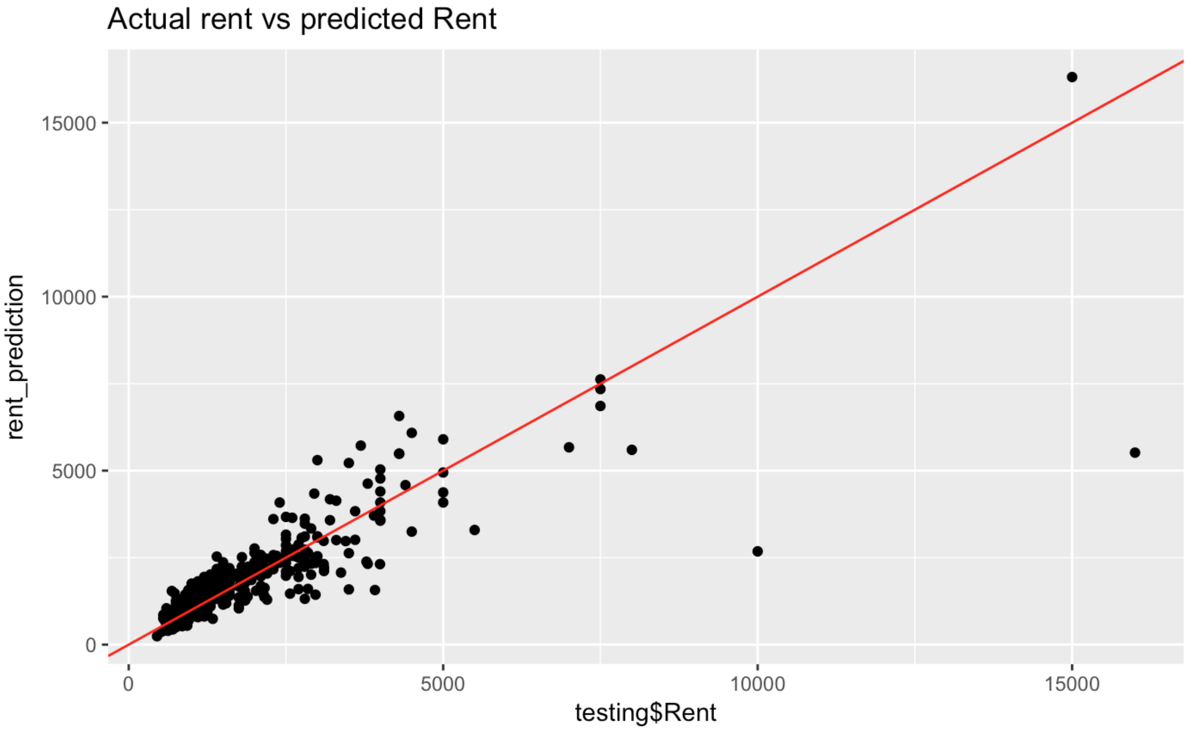


Figure 12: Actual rent price (testing set) vs predicted rent price


Actual rent vs predicted Rent
rent_prediction
testing$Rent