Apartments for rent classified

Elena Costagliola

25/08/2020

Contents

1 Introduction

The Apartments for Rent Classified original dataset was found on UCI - MLR but, as it was missing on the website, it has been personally requested to the Source. On UCI the description informs about two possible versions of the same dataset: one with 10,000 rows and the other with 100,000. The Source sent only the 10000 rows' file.

These data were cleaned to fill the missing values when possible, in other cases it was decided to delete the rows which showed unreliability of the original source. The dataset, in fact, results from the collection of data about apartments for rent classified in the USA. Unusual values would be more probably a mistype in the ad than a real information.

The goal of this report is to estimate the rental price of an apartment after identifying its most influential features. For this purpose, regression methods will be used.

The dataset consists of 6479 apartments for rent classified (from now on, *ARC*) and 37 features which characterize each one, including the corresponding price of rent. In this context it is possible to treat the ARCs as bundles of characteristics and prices can be attached to each of them. The statistical and economic method to deal with this kind of situation is known as *hedonic method*. In fact, it is a regression technique which allows to estimate how much of the final value of a good is determined by its characteristics. Moreover, this is associated to the economic concept of utility in which each amenity of the apartment increases the overall utility of the ARC itself.

For this reason, it will be analyzed how inserting amenities in the ads would affect the potential rental price, ceteris paribus.

Moreover, the relation between the geographic area and the rental price of the apartment will be the focus of the analysis, as well as the relation between the size of the apartment and the number of bedrooms.

2 Exploratory Data Analysis

2.1 The features

```
##     newcategory        class_price     bathrooms       bedrooms     bedrooms_f
##  apartment:5600   [100,1000) :2008   Min.   :1.00   Min.   :0.000   0 :  95   
##  studio   : 879   [1000,2000):3636   1st Qu.:1.00   1st Qu.:1.000   1 :3211   
##                   [2000,3000): 599   Median :1.00   Median :1.000   2 :2271   
##                   [3000,4000): 151   Mean   :1.31   Mean   :1.657   3 : 675   
##                   [4000,5000):  50   3rd Qu.:2.00   3rd Qu.:2.000   4 : 190   
##                   [5000,Inf) :  35   Max.   :5.00   Max.   :9.000   5+:  37   
##                                                                               
##  pets          price       square_meters                           SDiv     
##  0:1126   Min.   :  224   Min.   : 12.08   south_south_atlantic     :1406  
##  1:5353   1st Qu.:  931   1st Qu.: 60.20   south_westsouth_central  :1374  
##           Median : 1230   Median : 73.30   midwest_eastnorth_central: 981  
##           Mean   : 1401   Mean   : 81.66   west_pacific             : 868  
##           3rd Qu.: 1600   3rd Qu.: 96.80   midwest_westnorth_central: 549  
##           Max.   :19500   Max.   :529.54   northeast_middle_atlantic: 479  
##                                            (Other)                  : 822  
##         SReg      dishwasher elevator patio_deck pool     storage  refrigerator
##  midwest  :1530   0:3226     0:5839   0:4020     0:3230   0:4937   0:3366      
##  northeast: 786   1:3253     1: 640   1:2459     1:3249   1:1542   1:3113      
##  south    :2871                                                                
##  west     :1292                                                                
##                                                                                
##                                                                                
##                                                                                
##  AC       basketball cable_satellite gym      internet_access clubhouse
##  0:5592   0:6162     0:4814          0:5019   0:5050          0:5163   
##  1: 887   1: 317     1:1665          1:1460   1:1429          1:1316   
##                                                                        
##                                                                        
##                                                                        
##                                                                        
##                                                                        
##  parking  garbage_disposal fireplace washer_dryer playground gated    hot_tub 
##  0:2646   0:5270           0:5415    0:5381       0:5701     0:5992   0:6137  
##  1:3833   1:1209           1:1064    1:1098       1: 778     1: 487   1: 342  
##                                                                               
##                                                                               
##                                                                               
##                                                                               
##                                                                               
##  tennis   wood_floors view     alarm    TV       doorman  luxury   golf    
##  0:5995   0:5920      0:6289   0:6436   0:6275   0:6451   0:6382   0:6328  
##  1: 484   1: 559      1: 190   1:  43   1: 204   1:  28   1:  97   1: 151  
##                                                                            
##                                                                            
##                                                                            
##                                                                            
## 
```

2.1.1 *Categories of Apartments for Rent Classified*

As seen in the summary above, most of the ARCs are classified as “apartment”, while there is a small bunch of ARCs defined as studios. Comparing the numbers of bedrooms in the two categories (Figure 1), there is a small part of the apartments with no bedrooms, which seems to be unlinkely for an apartment but realistic for studios.

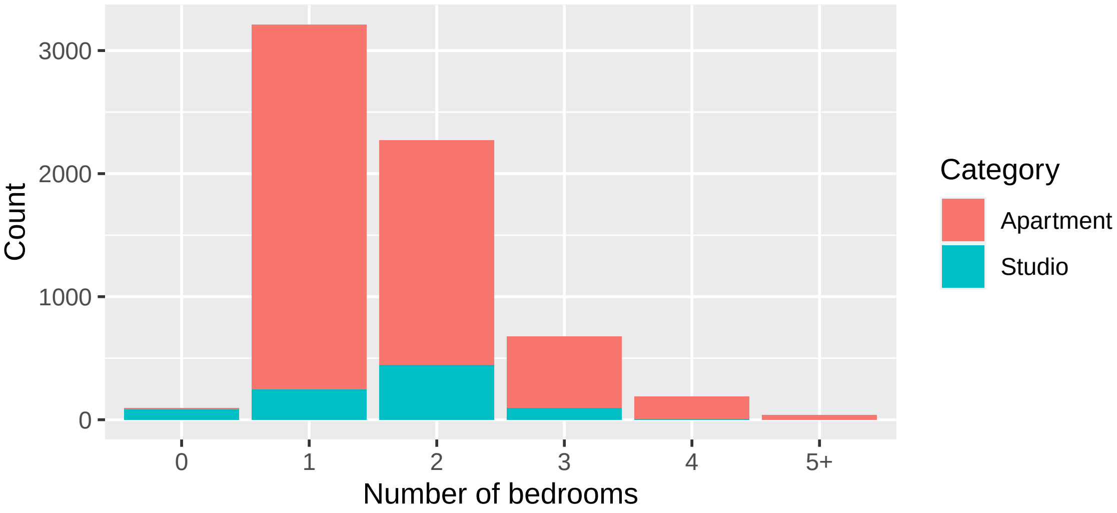

Figure 1: Number of bedrooms by category of Apartment for Rent Classified.

In particular, there are 11 apartments without bedrooms and they have been excluded from the present analysis, as they have been considered mistakes. It is worth noting, instead, that the ARCs defined as studios in some cases have more than one bedroom. It was decided to keep this information despite it seems unusual.

An insight on these data has been conducted including the price of the ARCs (Figure 2) and it shows that the prices of ARCs are not affected by the number of rooms. This point is also confirmed by a low correlation (0.2870922) between the number of bedrooms and prices.

This is probably due to the fact that having more bedrooms at the same square meters, on average, could reduce the value of the ARC rather than increasing it by resulting in smaller rooms.

2.1.2 *Class of price*

The feature’s summary shows that most of the ARCs are under 2000$ per month. Moreover, the Figure 3 gives an idea on the class of price of an ARC by its square meters and geographic position. Statistical divisions were preferred to Regions to keep more precision in the analysis, as will be seen below in this report.

However, the Figure 3 shows that there is a tendency for bigger ARCs to belong to higher classes of prices in all the Statistical Divisions, even if a bigger difference can be seen only in the classes from 3000$ on.

2.1.3 *Number of bathrooms*

This feature has decimal values, probably due to the presence of lobbies in the ARCs.

2.1.4 *Pets allowed in the apartment*

This is a boolean feature which takes the value 1 when pets are allowed, 0 otherwise. The summary shows that most of the ARCs allow them.

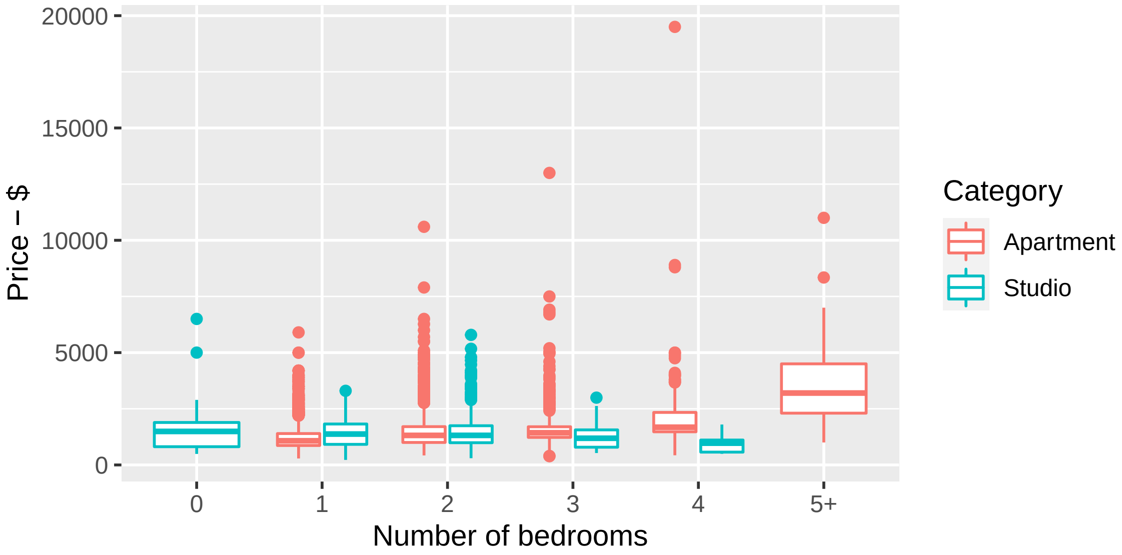

Figure 2: Number of Bedrooms and Prices ($) compared by category of Apartment for Rent Classified.

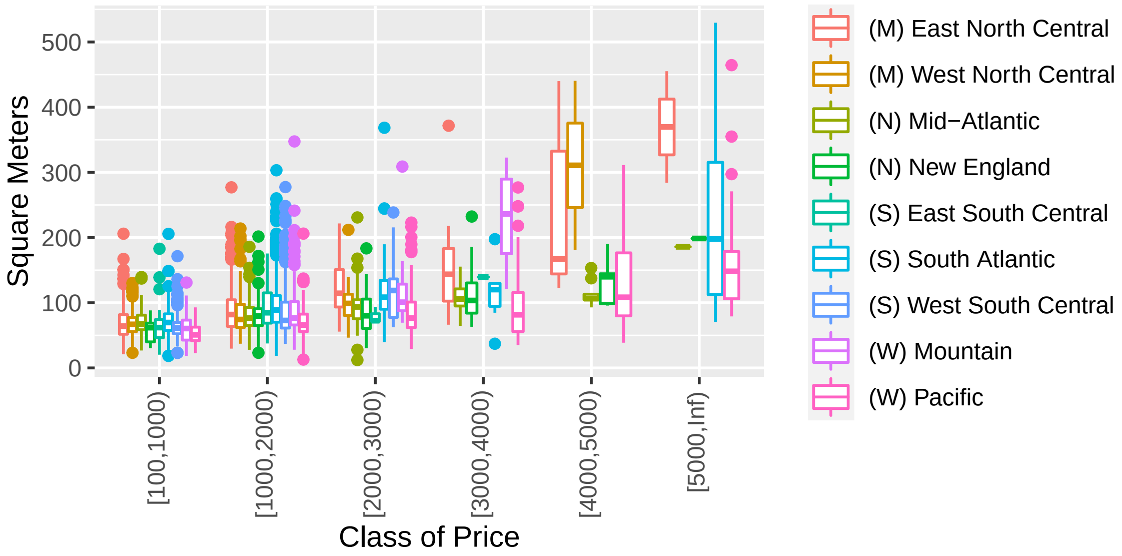

Figure 3: Square meters and Class of Prices ($) compared by Statistical Division.

2.1.5 *Price and Square meters of the Apartment for Rent Classifies*

These two features are here considered together since the high correlation expected between the two.

The Figure 4 shows that the distributions of square meters and price are very skewed, so they are logarithmically scaled to have normal distributions.

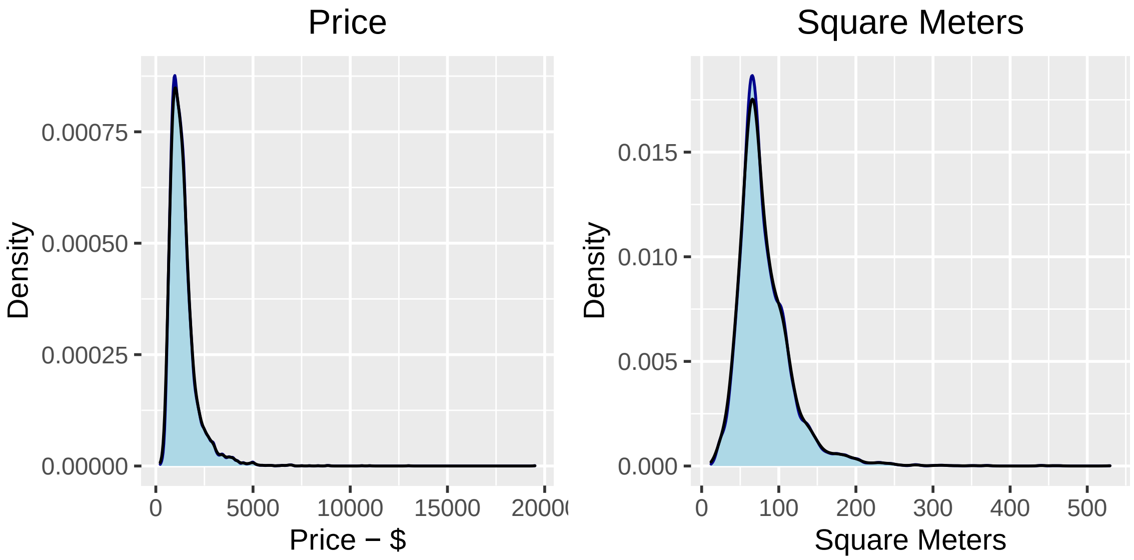

Figure 4: Density Plots of Price and Square Meters.

To better understand how prices and square meters are distributed within the Divisions, the Figure 5 fits a linear model using the logarithmically scaled price and square meters.

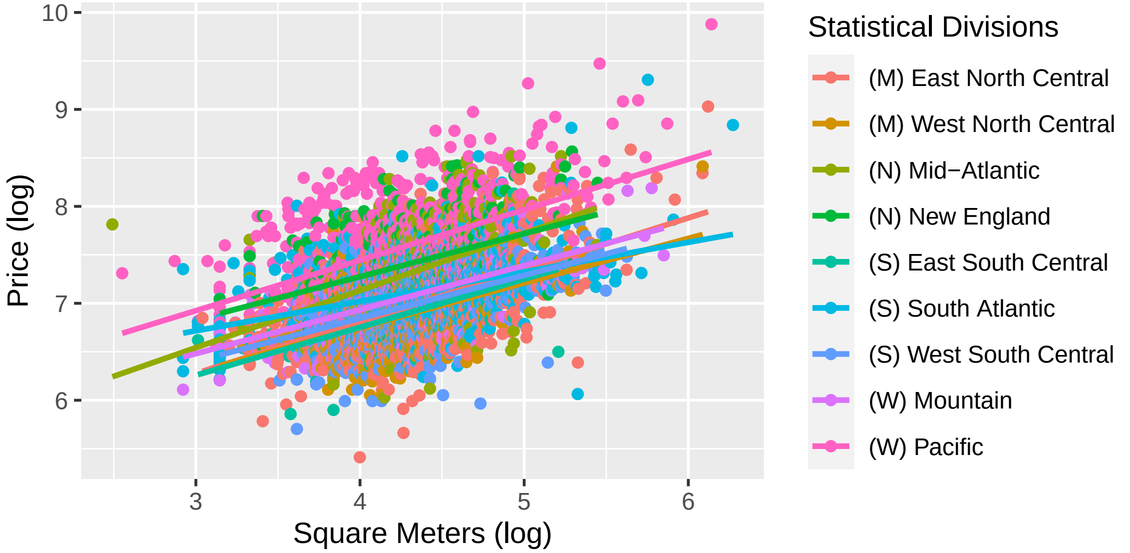

Figure 5: Normalized Price and Square meters by Statistical Divisions.

It shows that the most expensive ARCs are in the Pacific and, on the opposite, the East South Central and West North central have the less expensive ones. The comparison between the slopes of the lines shows that the prices in the Pacific and Mid-Atlantic increase faster then in the other divisions. On the opposite the

South Atlantic seems to be cheaper at the same square meters.

To improve the comprehension of the plot, in this context has been used the feature about the Regions to have an overall idea on the relation between price and square meters.

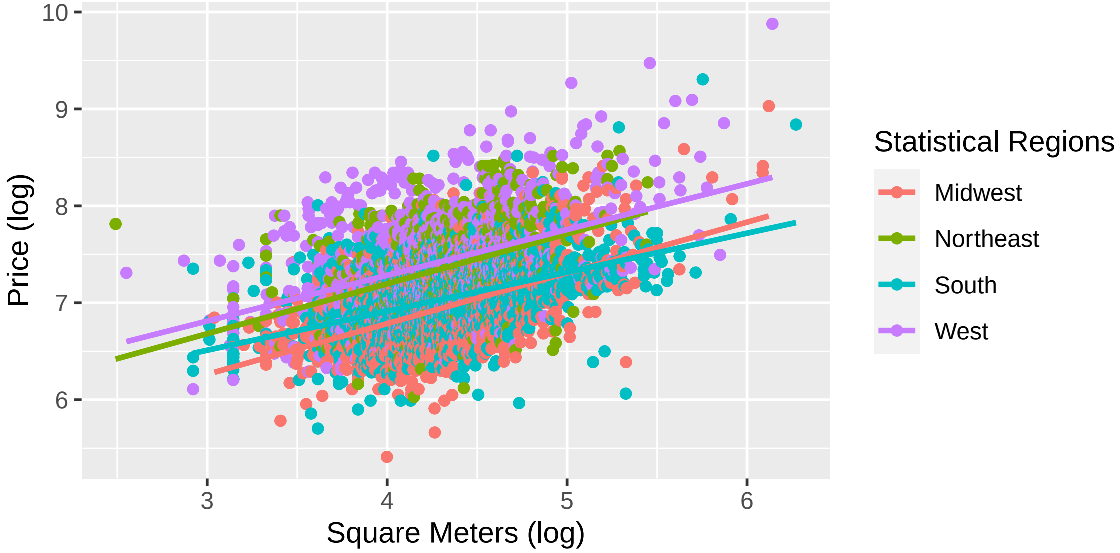

Figure 6: Normalized Price and Square meters by Statistical Regions.

Figure 6 confirms the observations made for Pacific and West North Central in Figure 5. It's evident here that the Northeast, together with the West especially for ARCs from 245 on, show the most expensive ARCs.

Also it is more evident than in the previously seen Figure 5 that, above the level of around 5 log(square_meters), the rental price of ARCs in the Midwest tend to be higher than those in the South, confirming that in the South the price increases more slowly than in the Midwest. This also means that an apartment around 148 square meters in the Midwest is more expensive than one with the same square meters in the South.

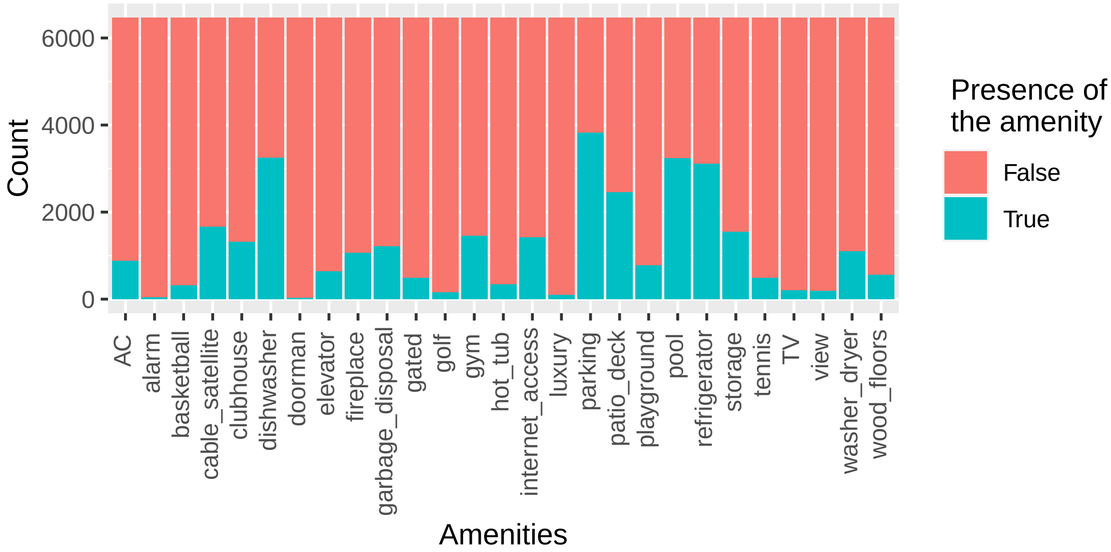

Figure 7: Presence of amenities in the Apartments for Rent Classified.

2.1.6 *Amenities*

The least 27 columns refer to the amenities offered by the ARC. As shown in the Figure 7, amenities are offered only by a small part of the ARCs. Roughly speaking, the only amenities offered by more than the 50% of the ARCs are dishwasher, parking, pool and refrigerator.

The impact of these variables on the ARC's final price will be on of the objects of this analysis.

3 Model Fitting and Features' Selection

The scope of this section is to select the most relevant features which affect the final price of the ARCs. To reach the goal two methods will be used: the K-Fold Cross Validation and LASSO (Least Absolute Shrinkage and Selection Operator).

Dataset has been splitted into two subsets: the training set consisting of 2/3 of data and the test set for the remaining 1/3, both randomly selected. The first one will be used to fit the model, the second one to predict the result and evaluate the accuracy.

3.1 K-Fold Cross Validation

To perform Cross Validation it has been set a K = 10, so the training set is randomly partitioned in 10 equal sizes subsets. Each of the folds is used as validation set at least one time, while the other partitions are used as training sets. This procedure is repeated 10 times.

```
##                   (Intercept)             newcategorystudio
##                    5.00015739                    0.12762752
##                     bathrooms                         pets1
##                    0.10072672                   -0.05733650
##                 square_meters SDivmidwest_westnorth_central
##                    0.42548211                   -0.07242012
## SDivnortheast_middle_atlantic      SDivnortheast_newengland
##                    0.35455906                    0.42530654
##   SDivsouth_eastsouth_central      SDivsouth_south_atlantic
##                   -0.07059308                    0.13268226
##             SDivwest_mountain              SDivwest_pacific
##                    0.10804309                    0.61028830
##                   dishwasher1                     elevator1
##                    0.04064254                    0.21538290
##                         pool1                 refrigerator1
##                    0.03469431                   -0.01186780
##                           AC1                   basketball1
##                   -0.02904302                   -0.10100311
##              cable_satellite1                          gym1
##                   -0.02606840                    0.03747566
##              internet_access1                      parking1
##                    0.02569766                    0.02662304
##             garbage_disposal1                    fireplace1
##                   -0.07842986                   -0.04233221
##                 washer_dryer1                   playground1
##                   -0.02896408                   -0.12725043
##                      hot_tub1                       tennis1
##                    0.04325074                   -0.03699208
##                  wood_floors1                         view1
##                    0.11162462                    0.08022332
##                        alarm1                      doorman1
##                     0.06245224                    0.14373745
##                        luxury1                         golf1
##                     0.09663671                    0.03862088
```

The model was fitted using both the backward and forward stepwise selection, but the second one was preferred since the model was less complex than that selected by the backward. In fact, with the Cross Validation performed by using forward stepwise algorithm the variables selected were `nbest_fwd` , while with the backward three more variables were included, with the same mean squared errors of `round(mse_fwd, 3)`

The resulting model seems to be quite coherent with what was discovered in the explorative analysis. The number of bedrooms, in fact is excluded from this model since it is not significant. Further, the ARC being a studio increases the overall price by 12.8%, while if pets are allowed the price is almost 6% lower.

It is worth analyzing the coefficient of the square meters. Since this feature and the price are both logarithmically scaled, this coefficient indicates that to an increase of the ARC's size by 1% corresponds an increase of the price by more than 0.42%. In economic terms, this coefficient can be interpreted as the elasticity of the price with relative to the size of the apartments. In fact, an elasticity < 1 implies that by increasing the quantity (in terms of square meters) the price increases more slowly.

As far as geographical features are concerned, the baseline used was East North Central (Midwest). In fact, an ARC located in the West North Central (Midwest) and East South Central (South) are underprized with respect to East North Central (Midwest) by 7.5%. In decreasing order the most expensive ARCs seem to be located in: - Pacific (+60.9%) - New England (+42.6%) - Mid-Atlantic (+35.4%) - South Atlantic (+13.2%) - Mountain (10.8%)

The only geographic area excluded from the model is West South Central.

Finally, let's focus on the most intriguing aspect: the value of amenities offered by the ARC. In the EDA it was evident that only a small number of the ARCs listed the amenities present, but the question was how much they impact on the final price of the ARC. The selected model confirms that on average ARCs with elevator, woodfloors, view, doorman and luxury tend to increase their final price. It's worth noting, instead, that amenities like pool, gym, parking and hot tub slightly increase the overall final price. On the opposite ARCs listing AC, garbage disposal, fireplace, washer dryer, playground, basketball and tennis court in the neighbourhood decrease the overall price. The non significant amenities excluded by this model are the TV, alarm, gated, internet access, clable satellite, refrigerator, storage, patio deck and the proximity to clubhouse and golf.

3.2 Least Absolute Shrinkage and Selection operator (LASSO)

As mentioned above, LASSO was here used to perform the feature selection. In particular, the model was fitted using the best tuning parameter (lambda) chosen with cross validation.

```
## [1] 37
```

```
## 42 x 1 sparse Matrix of class "dgCMatrix"
##                                           s0
## (Intercept)                    5.0445851049
## (Intercept)                    .
## newcategorystudio              0.1229238828
## bathrooms                      0.1007964715
## bedrooms                       .
## pets1                         -0.0498008411
## square_meters                  0.4182211108
## SDivmidwest_westnorth_central -0.0780910644
## SDivnortheast_middle_atlantic  0.3356356835
## SDivnortheast_newengland       0.4075069506
## SDivsouth_eastsouth_central   -0.0682757971
## SDivsouth_south_atlantic       0.1162002071
## SDivsouth_westsouth_central   -0.0091720529
## SDivwest_mountain              0.0898381185
## SDivwest_pacific               0.5945069541
## dishwasher1                    0.0301662091
## elevator1                      0.2138234876
## patio_deck1                   -0.0009304941
## pool1                          0.0332256887
## storage1                      -0.0029597543
## refrigerator1                 -0.0075014628
## AC1                           -0.0230872906
## basketball1                   -0.0946426531
## cable_satellite1              -0.0211019569
## gym1                           0.0335910084
## internet_access1               0.0188026090
## clubhouse1                    -0.0059812401
## parking1                       0.0211060292
## garbage_disposal1             -0.0722692491
## fireplace1                    -0.0367779456
## washer_dryer1                 -0.0216092790
## playground1                   -0.1235475818
## gated1                         .
## hot_tub1                       0.0394087625
## tennis1                       -0.0340432361
## wood_floors1                   0.1048080340
## view1                          0.0746930372
## alarm1                         0.0417750535
## TV1                            .
## doorman1                       0.1274439216
## luxury1                        0.0860156931
## golf1                          0.0268103749
```

Unfortunately this model includes more variables than the one fitted with the forward stepwise. However, it's worth of noting that also in this case the number of bedrooms is excluded from the analysis, confirming the initial assumption about the number of bedrooms considered less important than the size in square meters of the ARC.

In this model only two of the amenities are excluded, but all the amenities in common with the previous model present approximatively the same coefficients. On the opposite, the coefficients of the amenities excluded by the forward selection are here negative but with very small values.

The same can be said for the other coefficients. Therefore, the geographical areas of New England and Mountain give more value to the ARC by 2 percentage points over the previous model.

4 Conclusions

The price of ARCs, that is apartments for rent classified, has been estimated fitting two models, the first one selected the best features using the stepwise algorithm, the second one inseting in the model a tuning parameter. The final model chosen has been the one selected by stepwise to avoid complexity: the two models, in fact, present the same MSE and their coefficients are very similar.

The analysis confirms that the number of bedrooms does not affect the final price of the ARC, since it is not included in either models. On the opposite the square meters are.

It has been analized the relation between geographic area and rental prices, finding out that the most valuable areas are Pacific (which increases the overall price by 60.9%), New England (+42.6%) and Mid-Atlantic (+35.4%). Conversely, West North Central and East South Central seems to be the less valuable ones.

Particular attention has been given to the amenities offered by the ARCs, pointing out the ones having a greater impact on final price, in particular the presence of the elevator which increases the price by more than 21%, followed by the presence of a doorman (+14.4%), typical of buildings in the business or residential areas. On the opposite, seven of the amenities listed have a negative impact on the price, most of them being sport facilities. One possible explanation coud be the fact that the apartments are located in area outside of the center city. Distance from the center city could be an interesting feature for future analysis of the role of amenities.