**Capstone Project Report**

**Housing Rent Prediction**

Ujwal Kumar

10/28/2022

**Introduction**

As international or domestic students, many students arrive from various locations and countries, making it difficult for them to estimate the rent they must pay. Moreover, each student has unique requirements and options. As a result, their calculations for budgeting and spending could be challenging. Since housing is one of the most important basic requirements, we decided to do an analysis that will be useful for the current students as well as future cohorts in determining the proper rent price that is most suitable for them in order to attend the university. Each student has a unique background and specific preference. This is why we chose this dataset, which included characteristics such as the number of bedrooms, bathrooms, furnished options, laundry, parking options, pets allowed, smoking allowed or not, wheelchair accessibility, or even electric vehicle charging capabilities. We believe that these characteristics could be the factor contributing to the rental price.

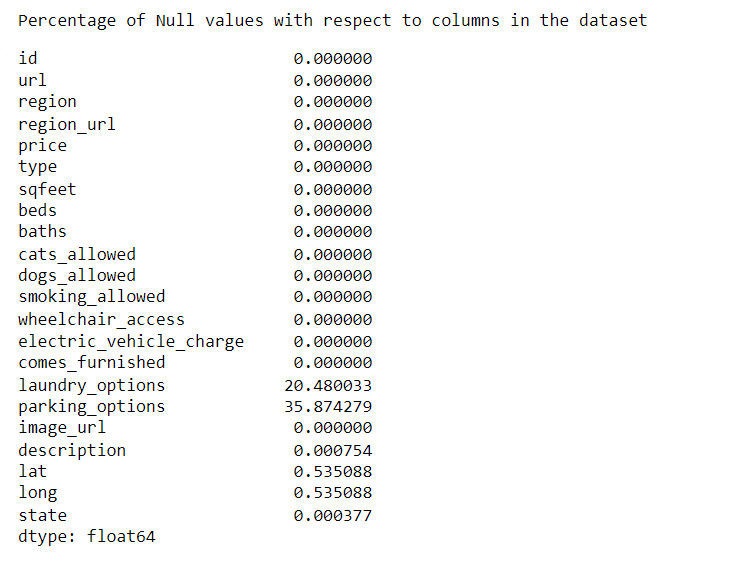
**Questions To Investigate:**

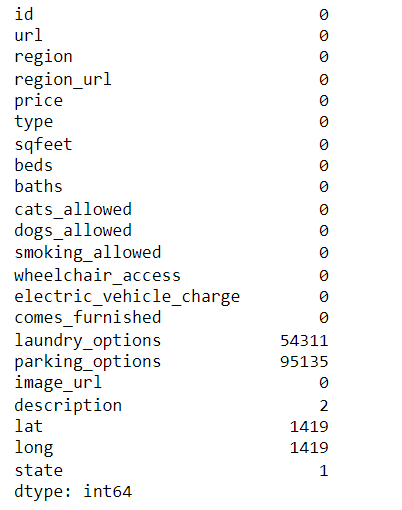
1. Which residential type is the most suitable to rent based on the individual requirements and amenities? (Logistic Regression)
2. According to the students' individual needs and amenities, which price is the most suitable? (Random Forest Regression and XgBoost Regression)
3. With respect to price and amenities, which region to expect? In addition, which features play an important role in prediction. (Random Forest Classifier)
4. What is the expected number of beds based on the region and amenities? (Random Forest Classifier)

**Exploratory Data Analysis**

**Data cleaning**

In our raw dataset we have numerous null values, which cause the difficulty of making any decision whether to drop these values or handle these null values. Therefore, we will use the percentage to decide the data we should remove.

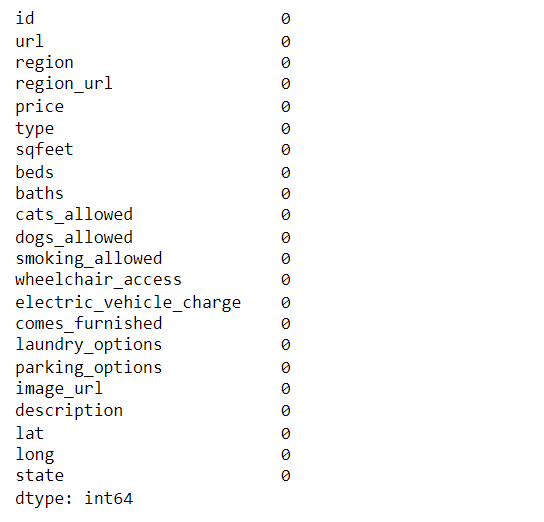




(

( Figure 1 : Original data summary) ( Figure 2 : The percentage of null values)

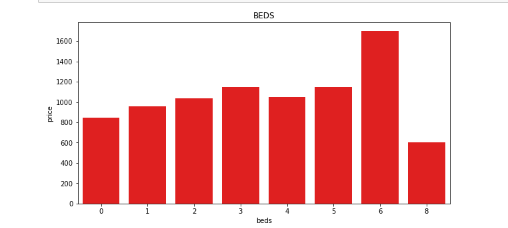
According to this method, we can see that laundry\_options and parking\_options have significant null values. But when it comes to lat, long, state and description they have very less percentage of null values. Hence we are dropping them directly. As for laundry\_options and parking\_options. We are replacing the null values with the mean so as to maintain the column distribution.Once we have replaced the null values and dropped the ones with less significant percentage. Now we have our dataset with 0 null values, as shown in the table below.



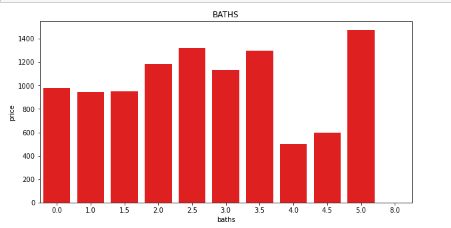
( Figure 3: Data summary after removing the null value)

In Feature engineering, we engineer/pre-process the House Rent Dataset's variables. We have removed missing-value variables, Temporal variables, and categorical variables to numerical variables. After that, we begin by examining the 'description' column, which is provided by the short listing description below.

**Treating Outlier**

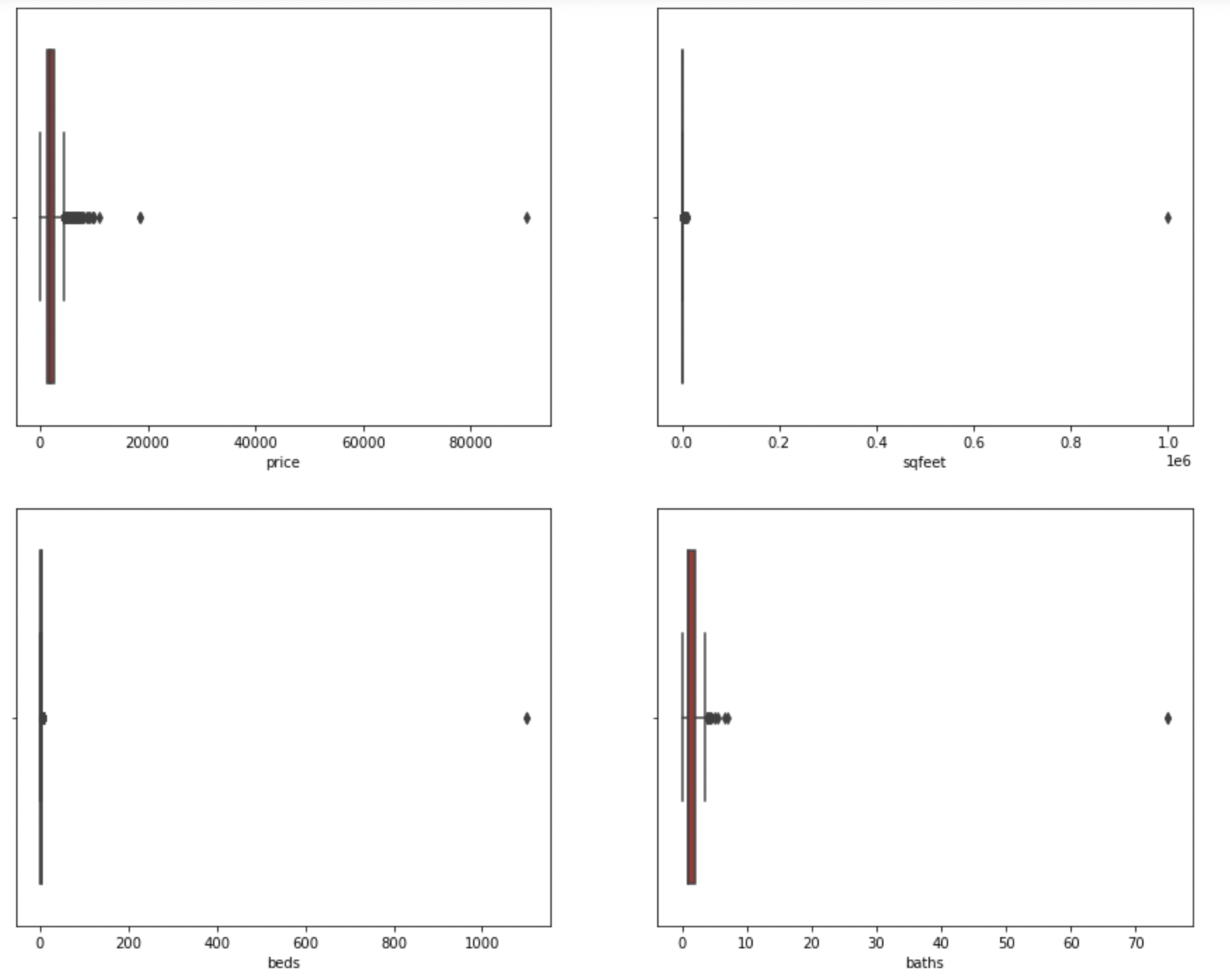
****

( Figure 4 : Number of bed vs Price )

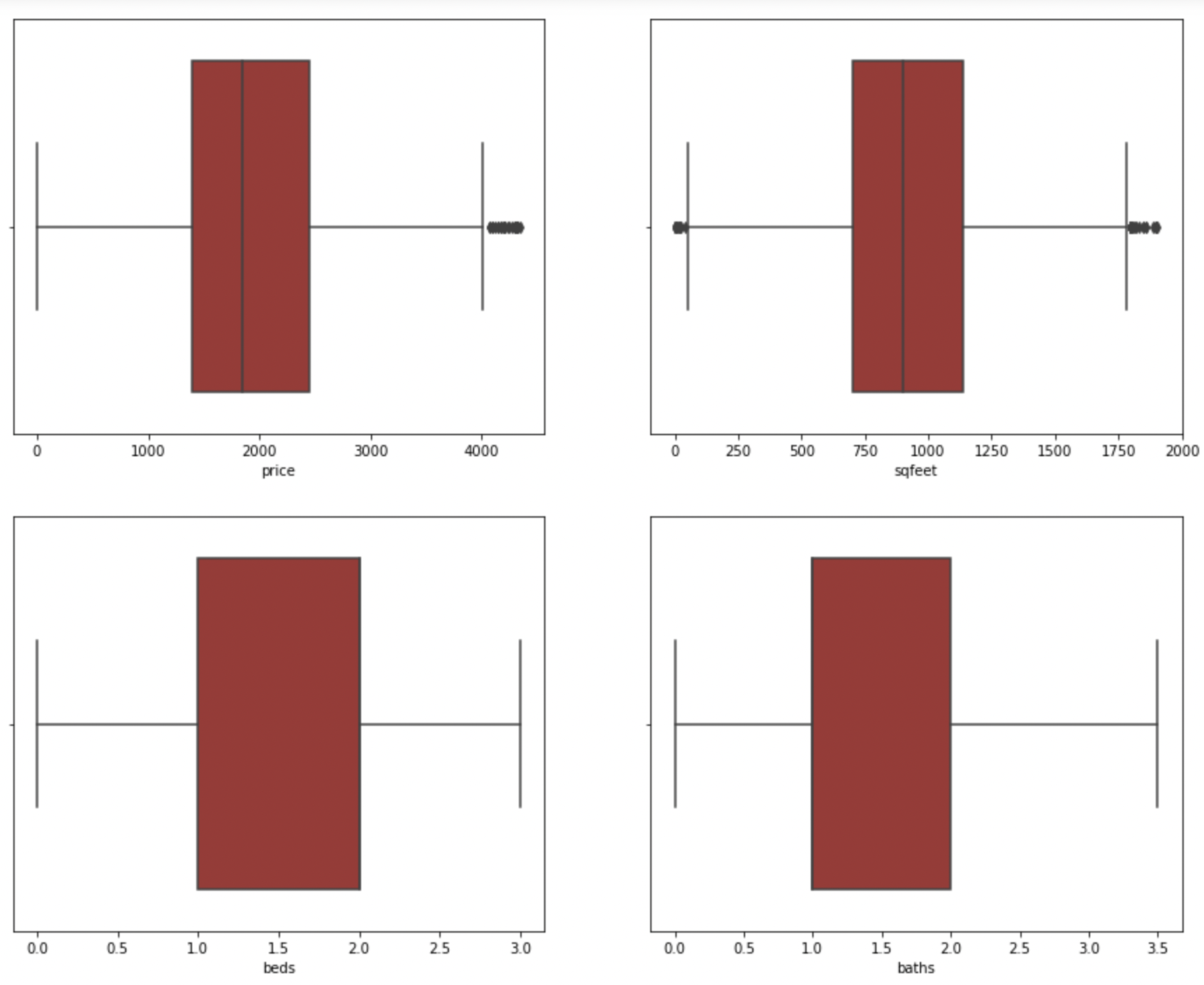
****

( Figure 5 : Number of bath vs Price )

This plot shows us Beds vs Price distribution for the dataset. From the chart, we can see the abnormal data on data that has the 6 and 8 beds. Moreover, the next plot for the number of baths vs Price distribution.We can see that there are some discrepancies in the relation of Price and Baths as well. Therefore, we intend to analyze the data to find the cause of this issue.

****

( Figure 6: Data with outlier)

****

( Figure 7: Without outlier)

Based on the findings in Figures 4 and 5, we decided to examine the outlier on continuous variables such as price, sqfeet, beds, and baths. In Figure 6, the outlier in the dataset is readily apparent, which could degrade our prediction performance. Therefore, we choose to eliminate them using the interquartile method.

**Adding New Features for Analysis**

According to the description we will analyze and create different features in our dataframe. For example, we have substring ''pool ” from “swimming pool” in order to create a new feature named ‘has\_pool’ and encode it to 1 and 0. Similarly we also created the following new feaures, including grill, fireplace, gymNearby, school/clgNearby, wifiFacilities, valetService, shoppingNearby, sportsPlayground and diningNearby. Below is the information of the description field example that we have in the original dataset.

**Description example:**

**“**

This is a sample description for our dataset, “*This pet friendly 1 bedroom upgraded apartment will be available for February. Rent is $980.00 with heat and hot water included. Apartment features include: \*Great living and closet space \*Connecting Dinette Area \*Generously sized bedroom \*Open-plan kitchen \*Excellent location \*Less than 10 minutes away from Naval Submarine Base! 2 pet limit per apartment and pets up to 75 pounds welcome! Please call for breed restrictions. Conveniently located near Waterford, Groton, Mystic, Old Lyme, East Lyme, and Niantic. Easy highway access. Within 10 minutes of major shopping plazas, great restaurants, beaches and more! Call today for more information! show contact info Equal Housing Opportunity*” .

**”**

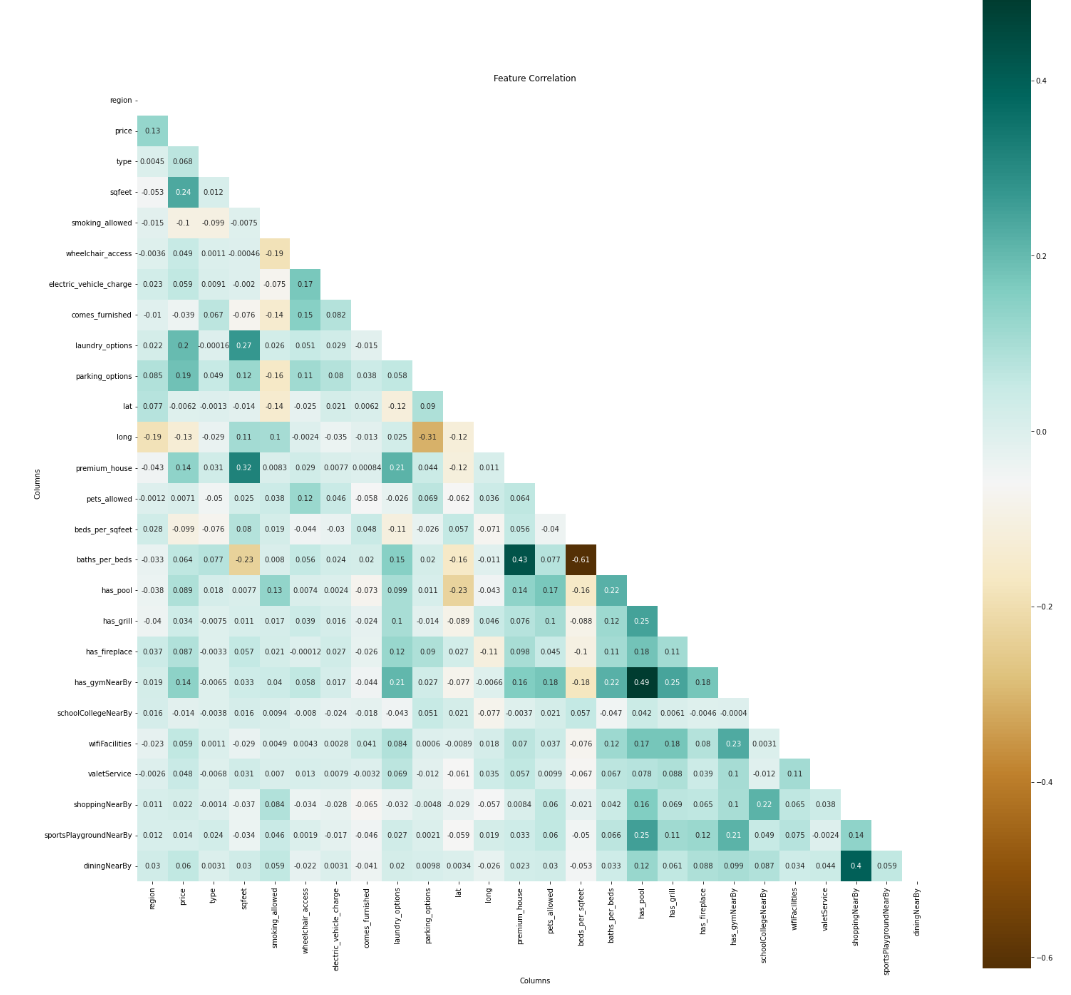
Therefore, we have added new features to our dataset like premium\_apartment by categorizing it into 1 by using the following criteria: when the number of bedrooms are greater than 1 and no of bathrooms are greater than number of bedrooms, we will consider it a premium apartment. Some other newly engineered features are beds\_per\_sqfeet and baths\_per\_beds.

**Data Preparation:**

Once the data was free of null values. We decided to check the correlation between the target variable and features on the original dataset. As you can see, there was some correlation between beds and baths and cats\_allowed and dogs\_allowed.

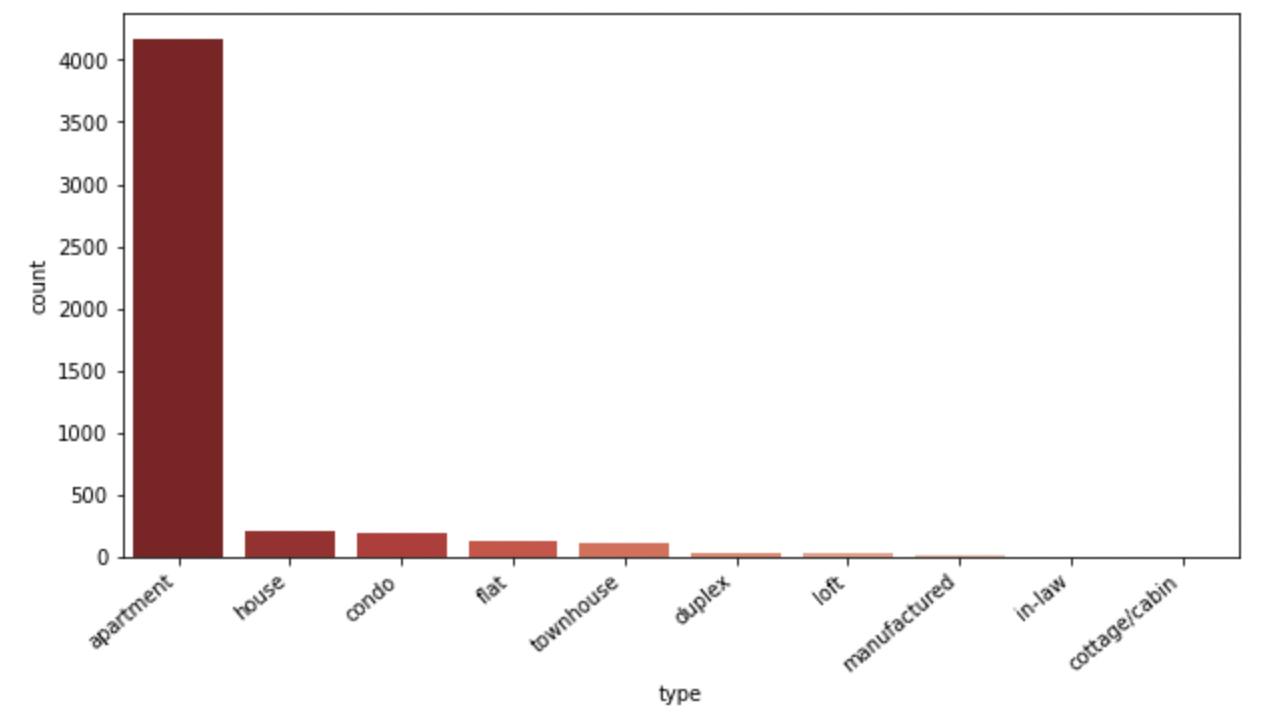
****

( Figure 11 : Correlation Between Features Before Adding New Features)



( Figure 10 : Correlation Between Features After Adding New Features)

Moreover, we also plot the correlation metric after we added the newly engineered features. We can see a correlation between gym\_nearBy and pool\_nearBy which are our newly added features. Similarly beds\_per\_bath and premium\_house also have a high correlation. All these new features will help us in the further modeling steps for the model to perform better.

****

( Figure 8 : The distribution of each apartment type)

Next, we investigate the data type in order to understand the data before we perform the logistic regression. When we look at the distribution of each residential category, two categories that have the most proportion are apartment and house. However, the number of apartments is slightly higher than other categories.

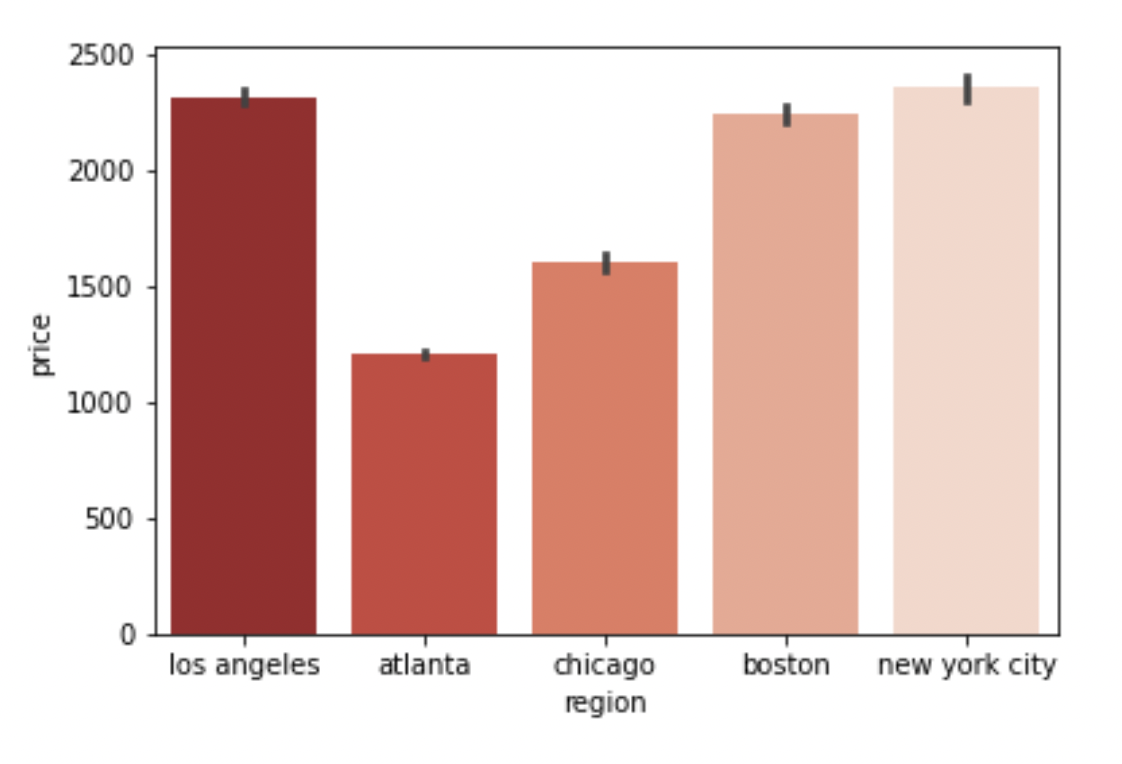
**Subset Data Preparation**

According to Bloomberg (2016, Florida), the top five cities with the largest college students are New York City, Los Angeles, Chicago, Boston, and Atlanta. From objective 2 until objective 4, we decided to extract the data which only consists of the regions mentioned above and go ahead with this subset of data to make the next predictions and classifications.

Moreover, we also added some extension predictors utilized from the description data field that we believed could be useful for our prediction, including pool, grill, fireplace, gym, school nearby, wifi, etc.

Next, let's visualize this subset of data and draw out some meaningful conclusions to use further

**Renting price in Top 5 Cities:**

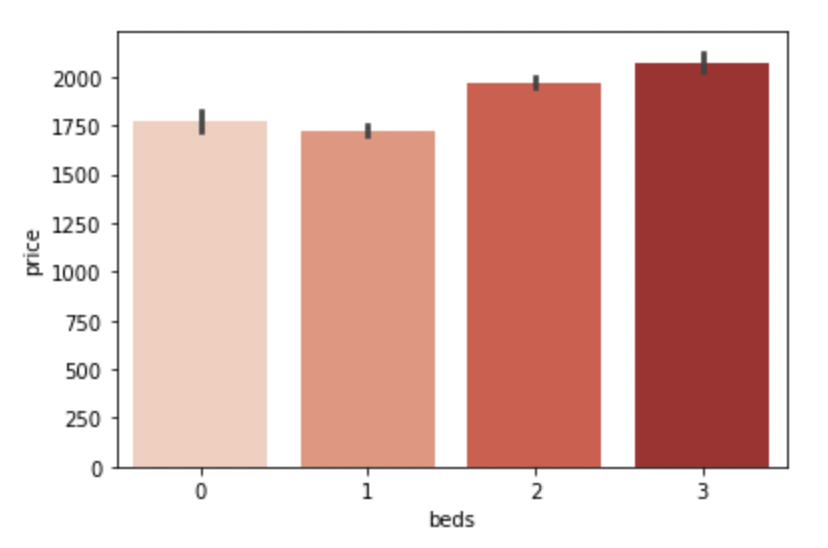
****

( Figure 12 : Renting price in top 5 states)

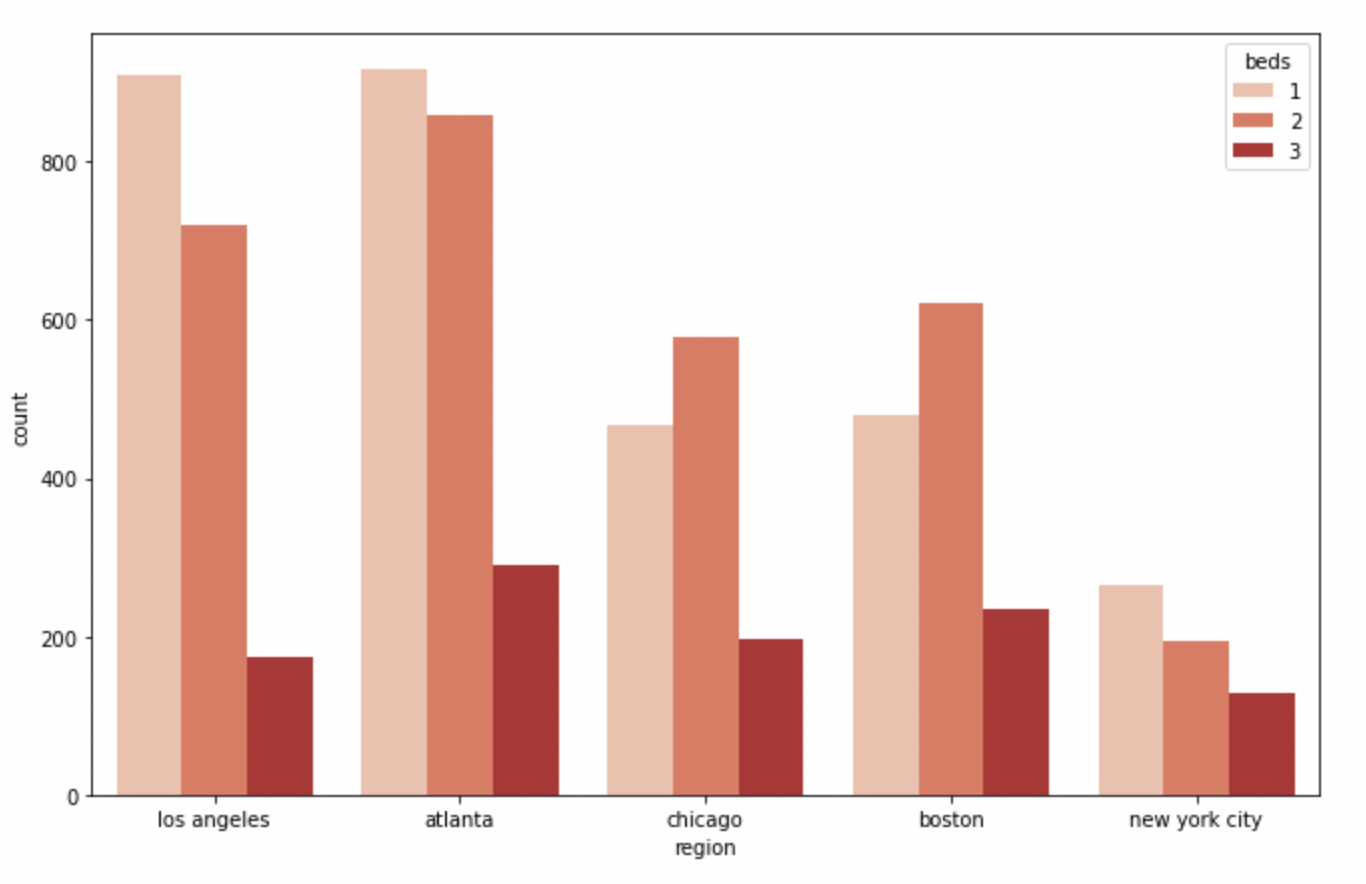
Based on the chart, we can see that the cities that have the highest average rent price are New York City, followed by Los Angeles, Boston, Chicago and Atlanta, respectively.

**Number of beds vs Price:**

The visualization below highlights that when the price increases as the number of beds in an apartment increases. According to this chart, we can get a better understanding of the trend when the number of beds is increased.

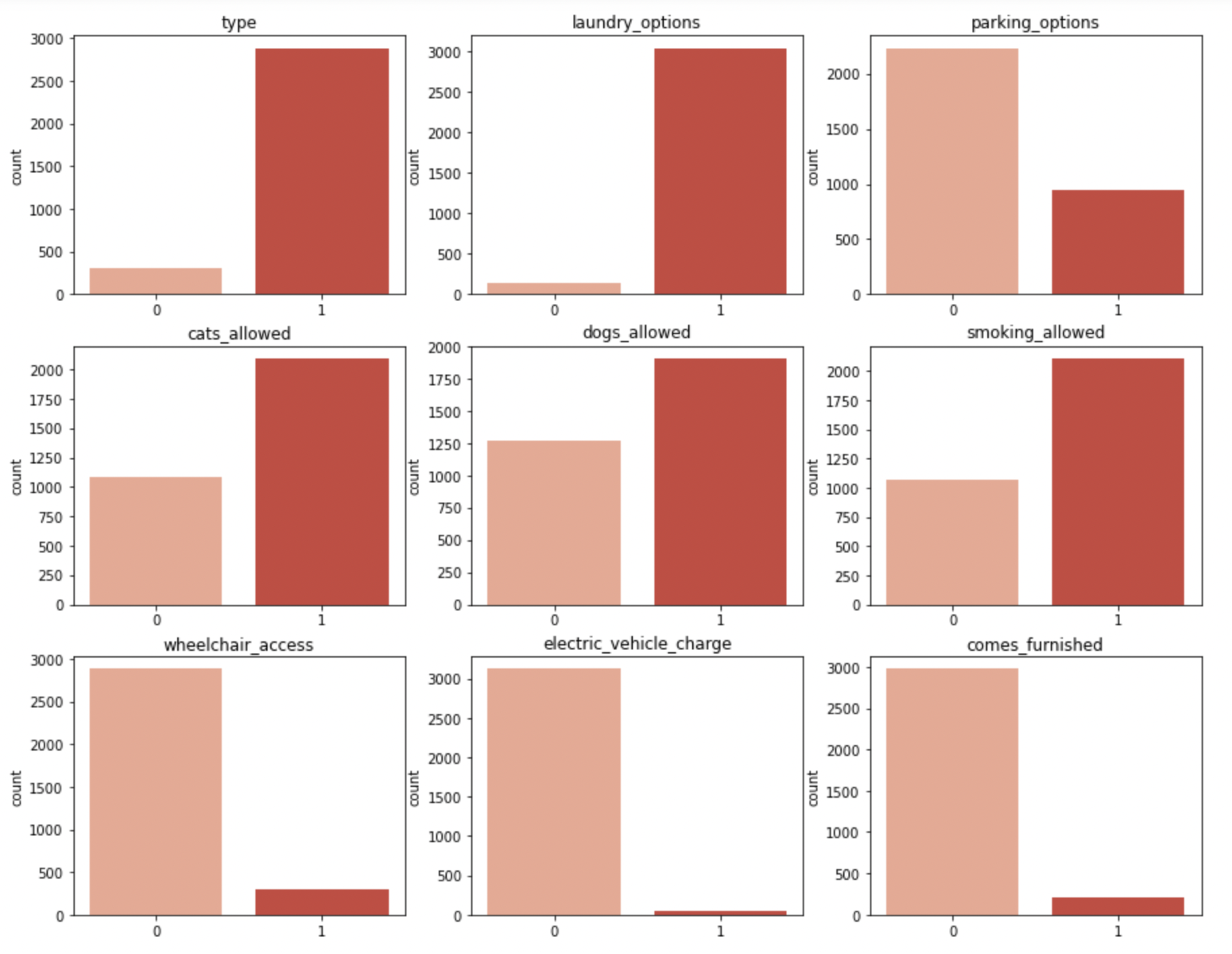
****( Figure 13 : Number of beds vs price)

Next we examined the number of beds available in the top five cities, which are Atlanta, Boston, Chicago, Los Angeles, and New York City in order to comprehend the data before we perform the classification.



( Figure 14 : Number of beds in five cities)

Based on the chart, we can see that the city that has the most one bedroom is Los Angeles, the most two bedroom and three bedroom is Atlanta. Moreover, the trend observed in Atlanta, Los Angeles, and New York City is that 1 bed, 2 beds, and then 3 beds are in decreasing order, whereas in Boston and Chicago, 2 beds, 1 bed, and 3 bed order is observed.

****

( Figure 15 : The distribution of all features for logistic regression)

In our subset of data which consists of top 5 Regions. We decided to evaluate how the boolean data is distributed. So here it is a plot of all the dummy variables from the dataset where 0 represents the house type and 1 represents the apartment type.

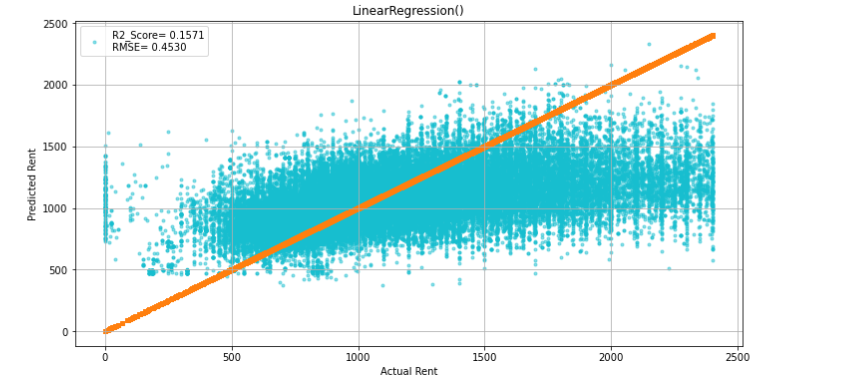
We learned from this that the majority of apartments had laundry facilities, but the home does not. Furthermore, apartments are more likely than houses to permit pets. Apartments, on the other hand, have very limited electric charging capabilities and wheelchair access.

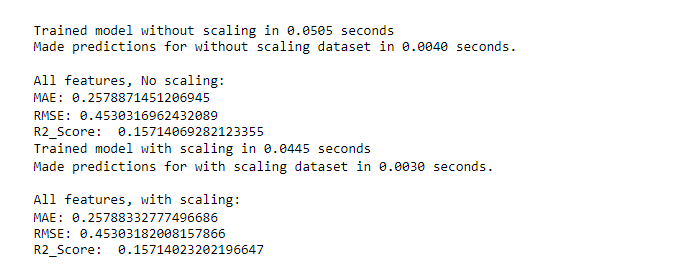
**Predictive Model**

***Objective 1:***

Predict the price of a given apartment with the respect to amenities (Features) provided.

**Model 1: Linear Regression**

****

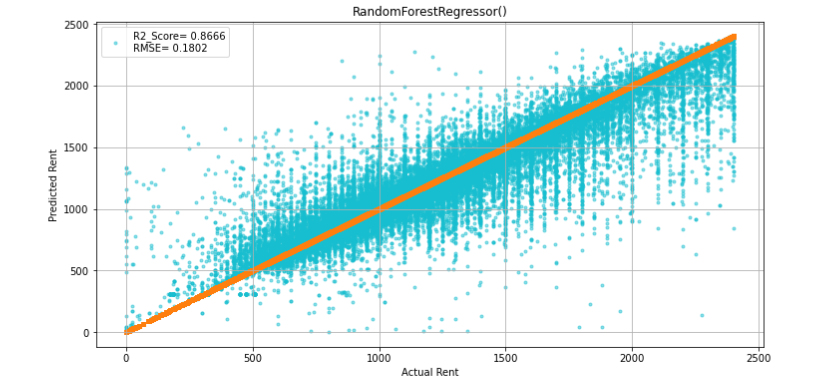
****

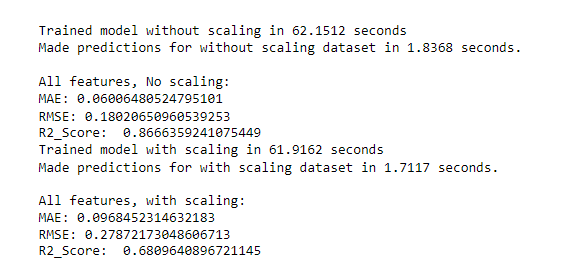
( Figure 16 : Linear Regression Result)

For the first objective, we apply linear regression to predict the price by using all existing data fields, excluding the price, and including the new data fields we had created. We split the data into testing and training datasets by 33% and 67%, respectively. After we performed the test, as expected, the performance was not good for Linear Regression, which we can see from the R2-score of only 0.15. Next, if we look at the orange line in the graph which represents the actual rents and the cyan circles plotted against actual rents, on the x-axis, are the predicted rents. The Linear Model worked terribly, which was expected as there were no linear relationships among the features. Feature scaling has a small positive effect on LR's prediction performance.

Therefore, we utilized other statistical models to determine the most accurate price prediction model.

**Model 2: Random Forest Regression**

****

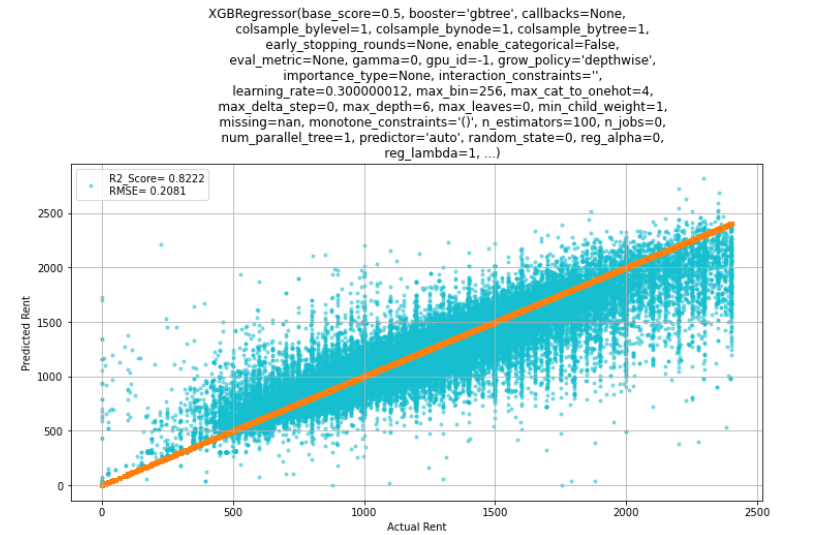
****

( Figure 17 : Random Forest Regression Result)

Moving to the Random Forest Regression. The performance is much better than Linear Regression as shown in the R2-score that has more than 0.85, which is more than the last model by over 6 times. Moreover, you can see the Orange line and the Cyan circles plotted in the graph as much more aligned and give a much better performance than the Linear Regression which shows that there is a strong relationship between actual rent and predicted rent.

**Model 3: XgBoost Regression**

Next, we tried the last model to evaluate the prediction by using the XgBoost regression to compare the prediction with the previous two models. As you can see, the performance of the XgBoost model has a pretty good R2-score at 0.8. However, it is still lower than the random forest regression.



( Figure 18 : XgBoost Regression Result)

**Interpretation**

Till now we used house rent dataset to build a monthly rent predictor using 3 different learning regressors (Linear Regression, Random Forest and XGB Regressor) were tested, and we have achieved the best prediction performance using Random Forest while Linear Regression, achieved the worst performance of the 3.

The best prediction performance was achieved using Random Forest regressor, using all features in the dataset, and resulted in the following metrics:

1.Mean Absolute Error (MAE): 0.0060

2.Root mean squared error (RMSE): 0.180

3.R-squared Score (R2\_Score): 0.8677

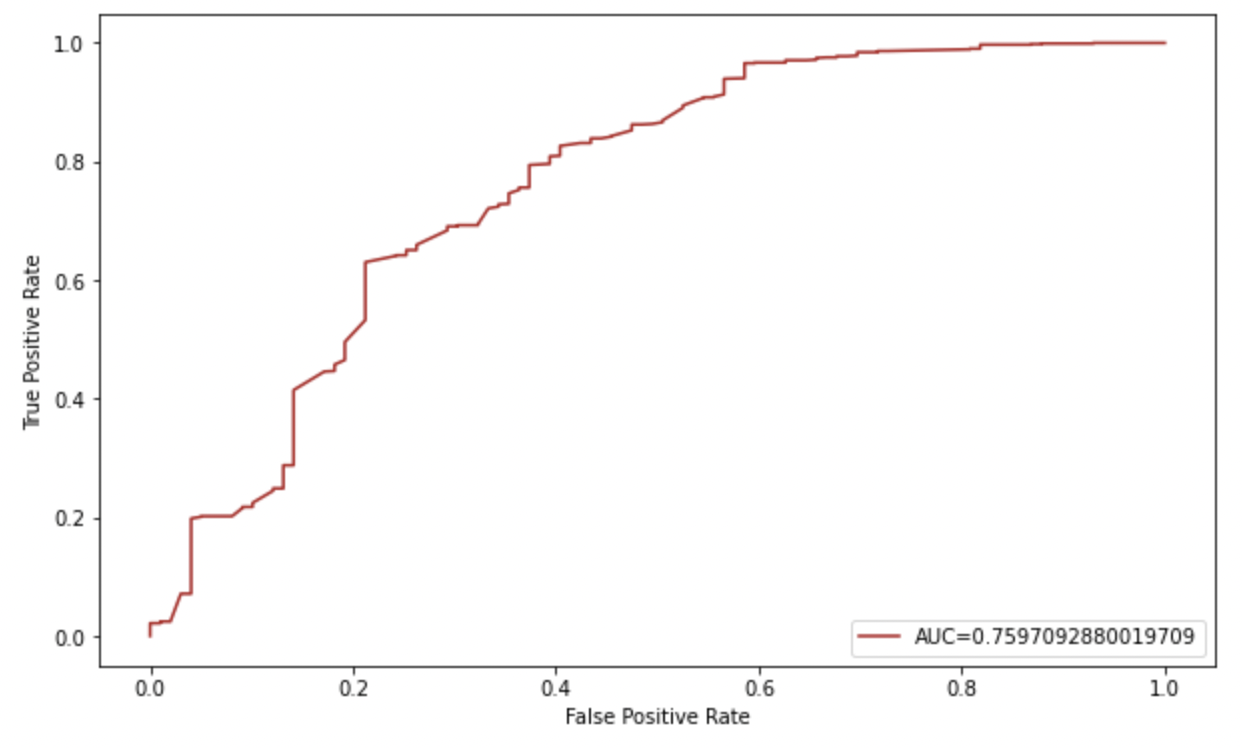
***Objective 2:***

Find the most suitable apartment with the given amenities requirement

**Model 4: Logistic Regression**

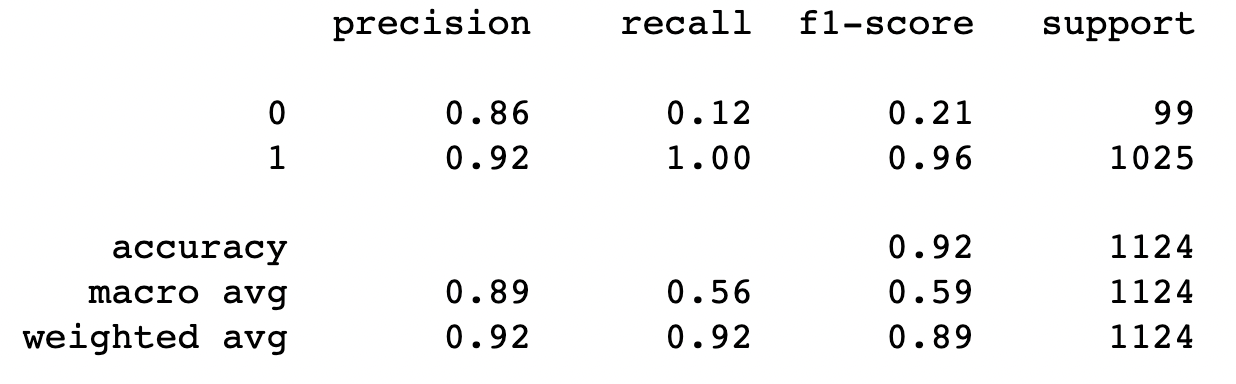
Our second objective is to predict the housing type through multiple categorical features that we have in order to give the recommendation for the student to choose the suitable apartment that matches their requirement amenities. According to the objective, we believe that logistic regression is the most suitable method, which is the appropriate regression analysis to conduct when the dependent variable is binary to explain the relationship between one dependent binary variable and another independent categorical variable.

In order to accomplish the task, we split the dataset into train and test data, 33% for the test size and 67% for the train. Next, we fit the model to minimize the sum of squares of the distance between the points and the regression line (squared in order to avoid negative differences).



( Figure 19 : ROC Curve and AUC )

After that, we will calculate the true positive rate and false positive rate and generate a ROC curve. The closer the curve approaches the upper left corner of the diagram, the more effective the model is at categorizing the data. We can see from the graph that our model performs exceptionally well based on the outcome. In addition, the AUC is used to determine the model's precision. The more closely AUC approaches 1, the better the model. In this instance, our AUC is 0.75, indicating that the model did a good discrimination.



( Figure 20 : Classification Report of Logistic Regression )

Lastly, we generate the classification report according to the result of our prediction model.

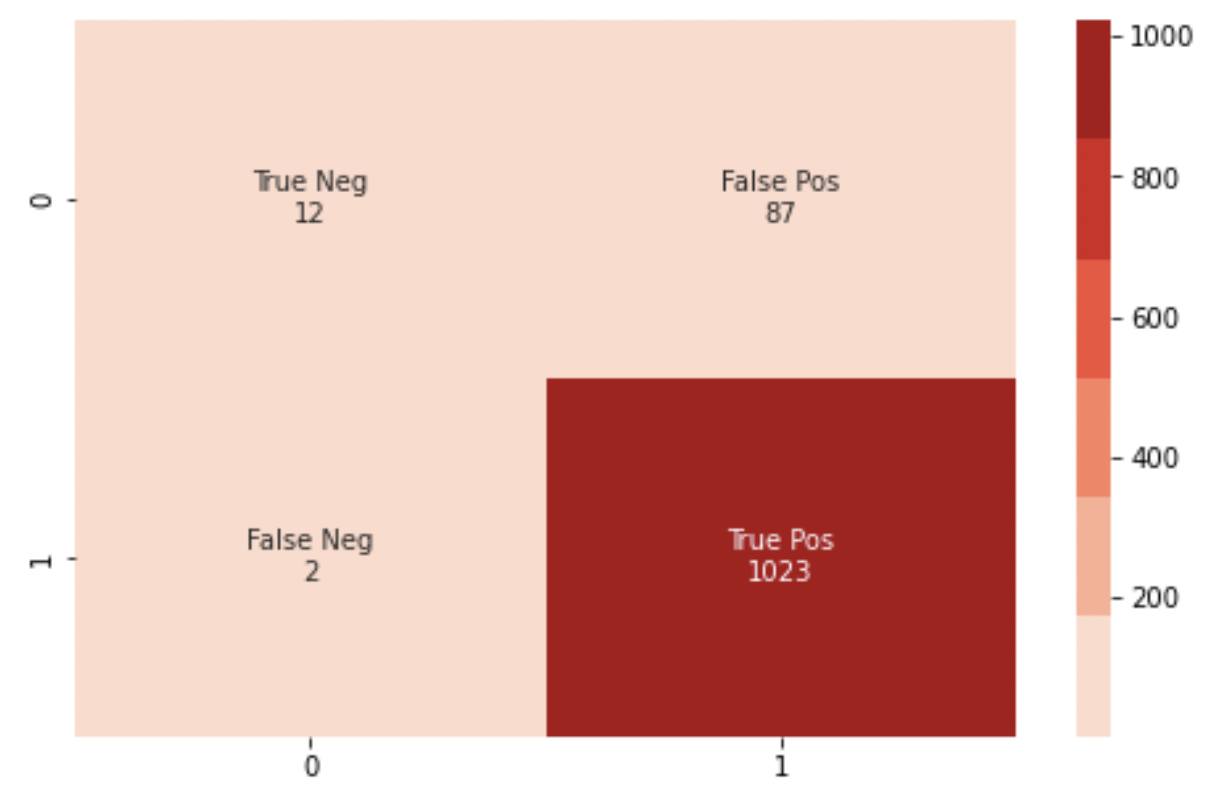
**Prediction:** Indicating the percentage of our correct prediction

**Recall:** The percentage of positive cases that we catch

**F1-score:** The percentage of positive prediction

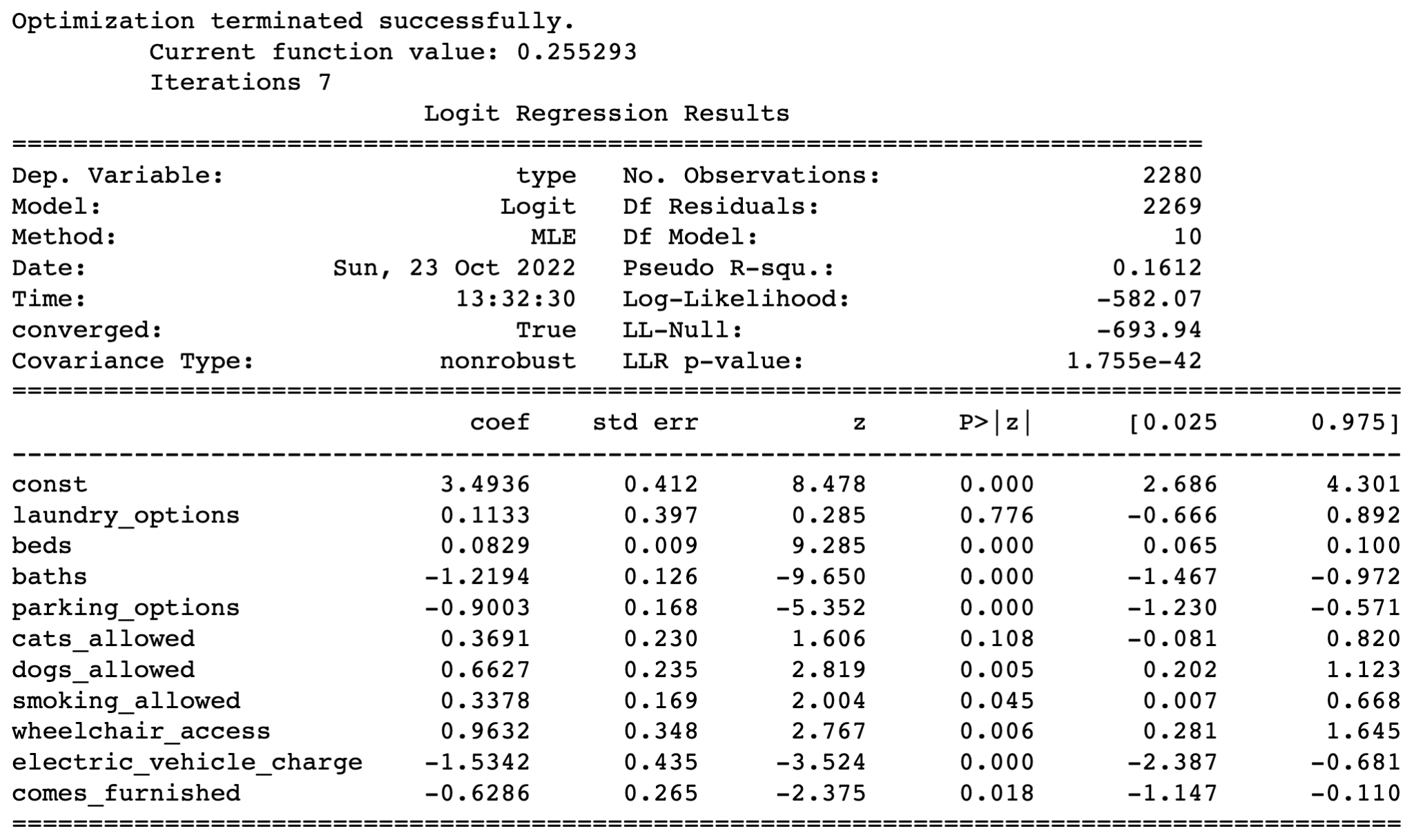
**Support:** The actual number of instances of the class within the given dataset

According to the results, our model performed exceptionally well with a precision of **0.89**, a recall of **0.56**, and a f1-score of **0.59**. For the whole output, the accuracy is **0.92**. Therefore, we can be certain that our model made a remarkable prediction. However, we obtained a low recall and precision; we must look into the possible causes for this result.

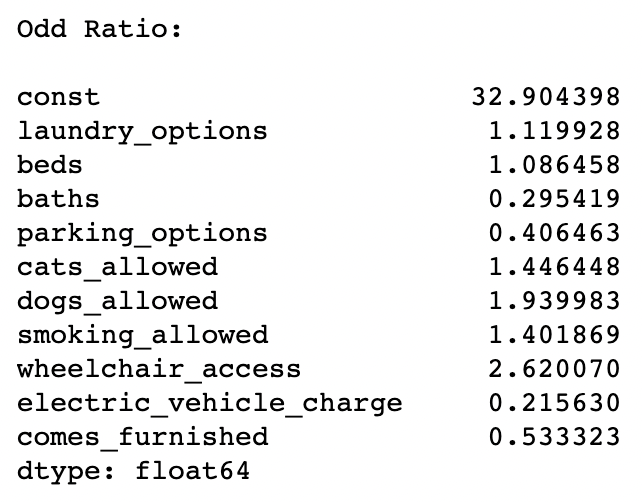


( Figure 21 : Confusion Matrix )

The confusion matrix result indicates that the model can correctly predict over 1023 apartments with the false negative prediction at only 87. Nevertheless, we can see that our prediction is slightly biased in this case since the number of apartments is more than houses, pretty much as we mentioned earlier in the EDA section, and also, this could be the possible reason that we got a low recall and precision outcome.



( Figure 22 : Logistic Regression Regression Coefficient )



( Figure 23 : Odd Ratio )

Lastly, we used the regression and odd ratio to determine how each predictor variable has an influence on the response variable. According to the regression coefficient, positive coefficients indicate that the event becomes more likely to happen as the predictor increases. Negative coefficients represent that the event becomes less likely to happen as the predictor increases.

After that, I convert the model coefficient to the odds ratio to better understand how each independent variable influences the response variable. In this case, we can imply that for each one-unit increase in the number of beds, the odds of apartment type rose by a factor of approximately 1.08, or one unit changing of the laundry option rose the odds of the apartment type by 1.11.

**Interpretation**

The overall accuracy that we have for the prediction is 0.92 which is excellent. However, the data we have is a little bit biased since we got the apartment data over house data by over ten times. Therefore, we can resolve this issue by resampling the data or clustering the abundant class instead of relying on random samples to cover the variety of the training samples.

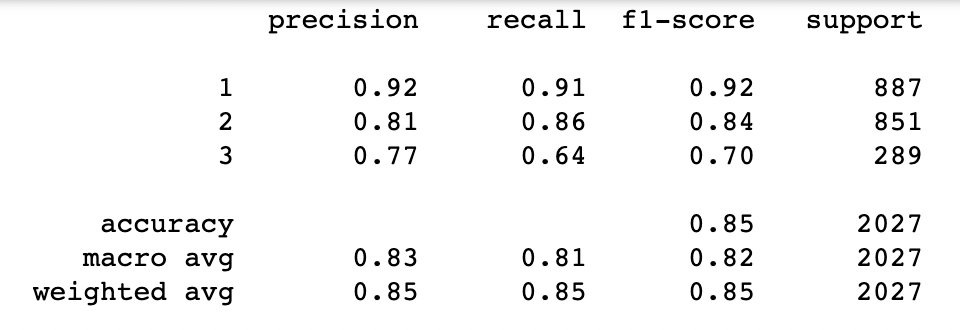
***Objective 3:***

Find the most suitable for renting regarding the amenities preferred.

**Model 4: Random Forest Classifier**

Random Forest Classifier model will be used to predict suitable regions and expected number of beds with the given amenities requirement. In addition, we will look into detail as to which features have the most influence on the prediction. As usual, we split the train data into 67% and 33% for the test data. Finally, we evaluate the algorithms.

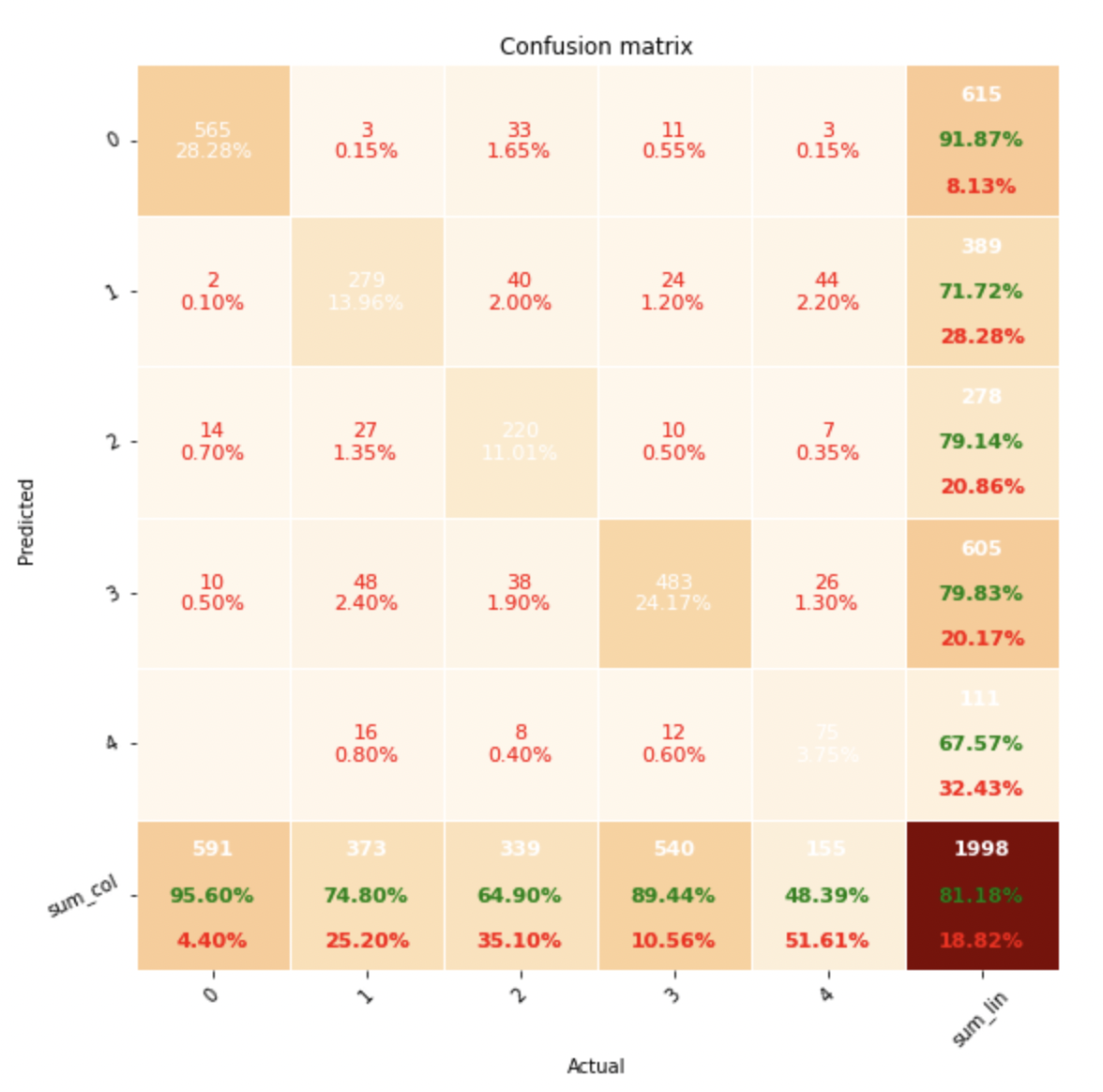
Results of region prediction is:



( Figure 24 : Classification Report )

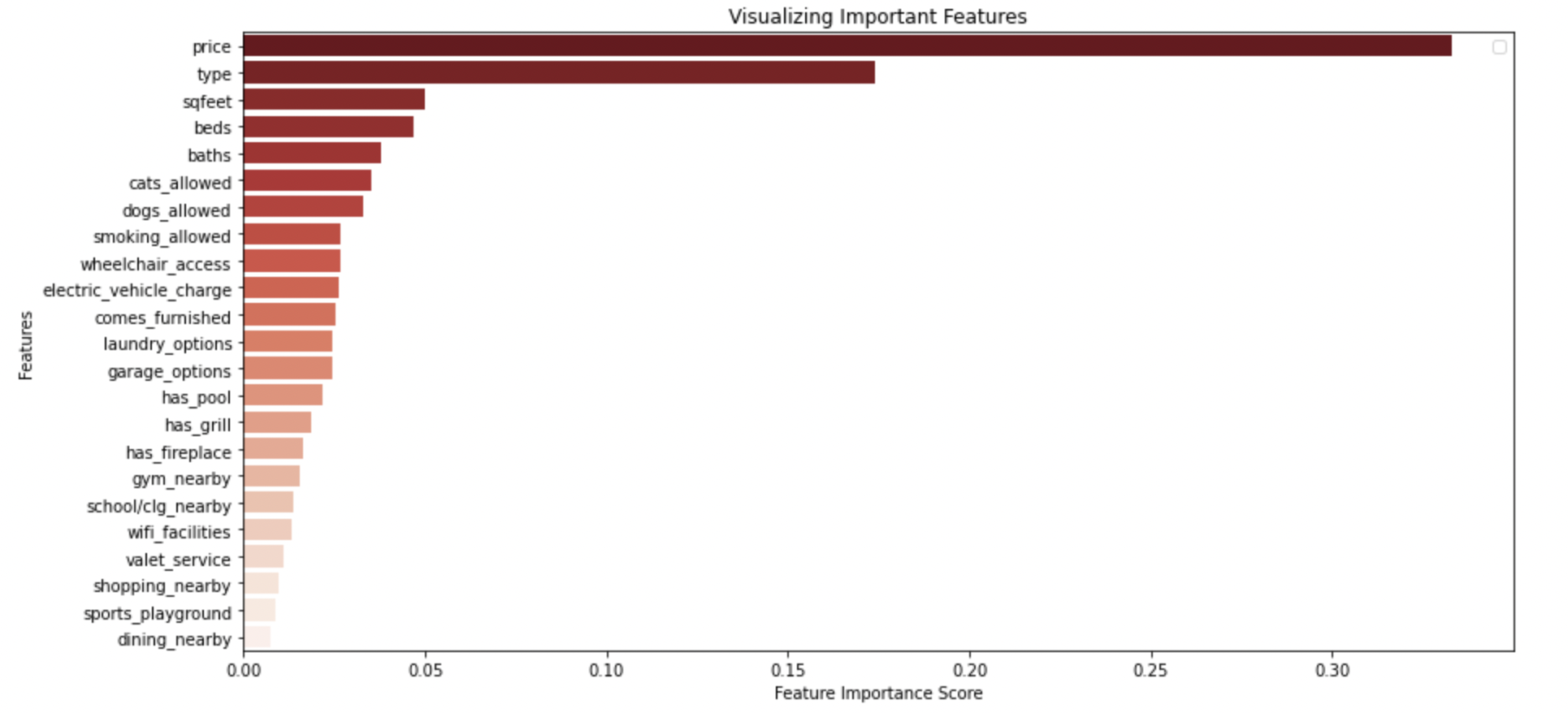
If we look at the classification report, we can see that our model did very well, with an overall accuracy of 0.78, which is considered good.

Next, we use the confusion matrix to determine the accuracy of each variable prediction. Most estimators have an accuracy of over 70%, and the overall accuracy is around 81%, implying that the model has a very good prediction accuracy.



( Figure 25 : Confusion matrix )

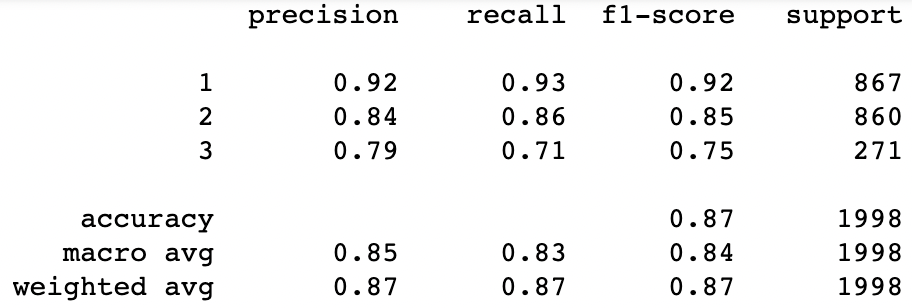
Moreover, we visualize the model features' importance score to make us interpret easier on what features have the most influence on the response variables. We can see that the price is the most important estimator in this model.

( Figure 26 : Feature Importance Score )

***Objective 4:***

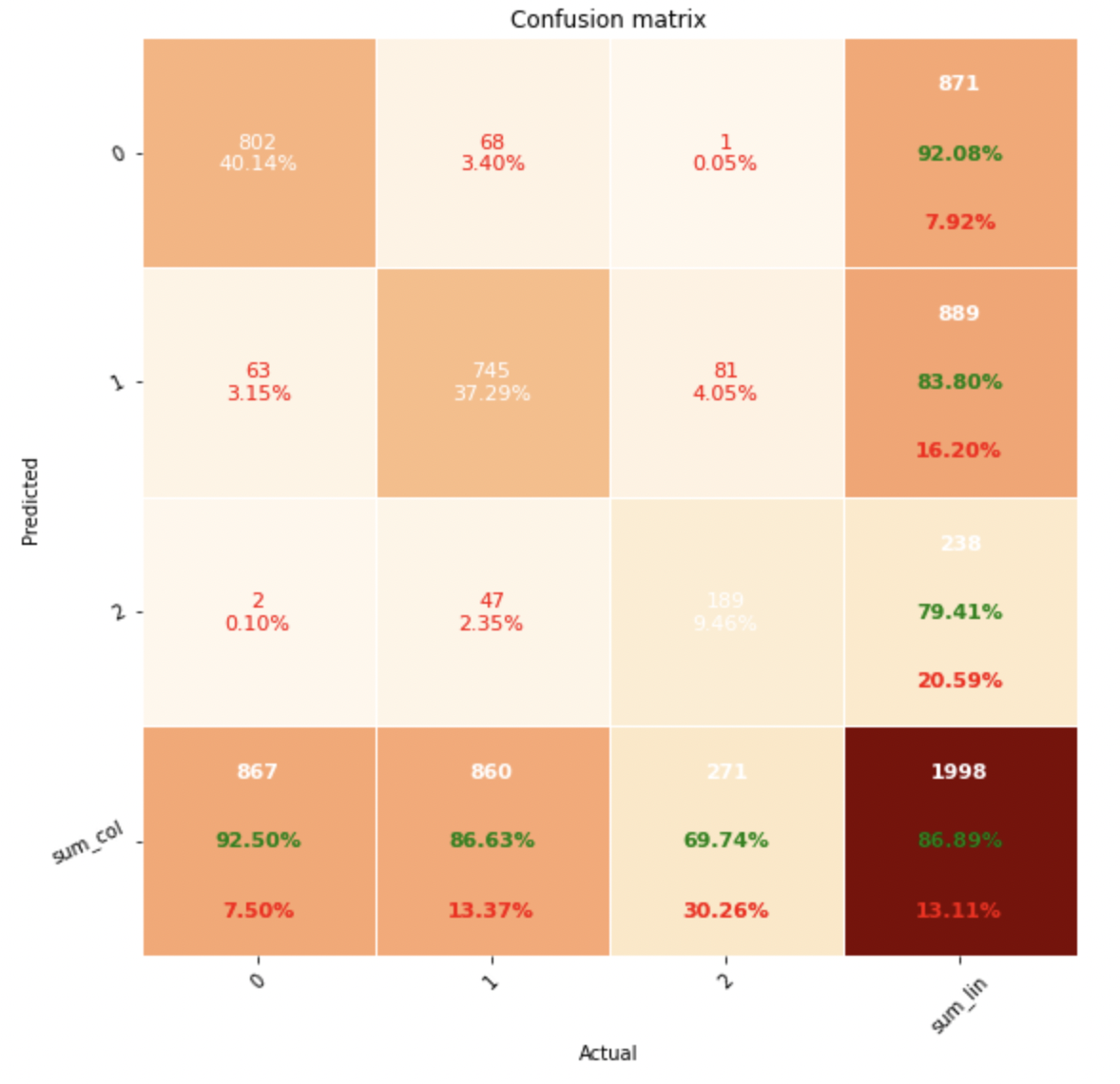
Find the expected number of beds based on the region and amenities

**Model 5: Random Forest Classifier**

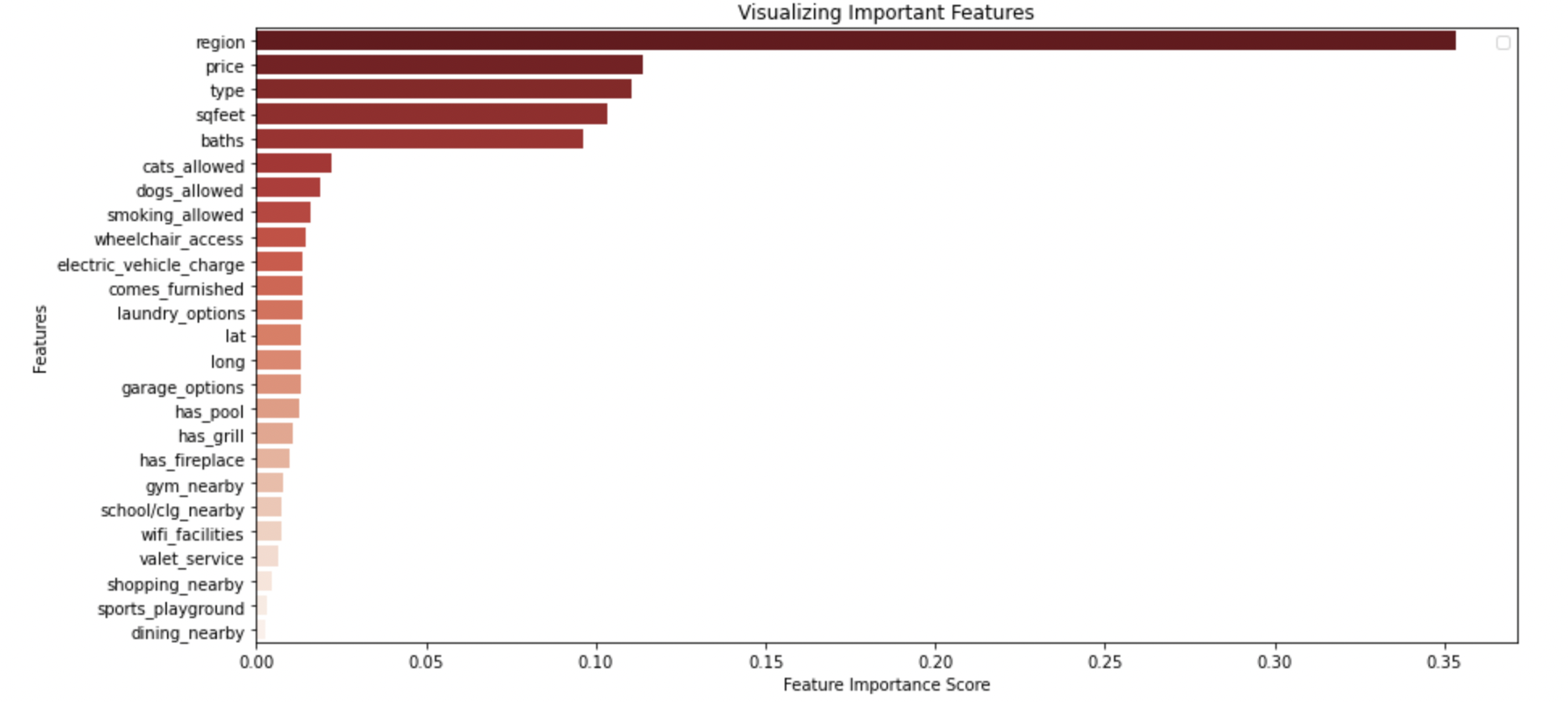
After that, we created another model to estimate the number of bedrooms. Based on the result, the model has an accuracy of over 0.87, which is an excellent performance. 

( Figure 27 : Classification Report )

Then, if we look at the confusion matrix, most variables have an accuracy of over 70%, which is excellent.



( Figure 28 : Confusion Matrix )



( Figure 29 : Feature Importance Score)

As in the previous model, we display the features' importances score to determine the critical variables in our prediction. We can see that region is the most important feature in this case.

**Conclusion**

After conducting our descriptive and predictive analyses in depth, The following are concluding and scope for improvement considerations:

* The best model for predicting the price amongst Linear Regression, Random Forest and XgBoost Regressor is random forest regression with an RMSE of 0.18, whereas linear regression is the worst model for the price prediction amongst the three.
* For the logistic regression, the factor that has the most influence in altering the residential type is wheelchair access, followed by cat and dog allowed. On the other hand, the number of bathrooms, parking spaces, electrical outlets, and furnished options are features that have the least effect on altering to apartment type.
* For the city prediction, the most important feature for our prediction is the price, and the least important feature is dining nearby, while the most important feature that affects the number of bedrooms is the region and the least is the dining nearby.

**Bibliography**

Ahmed, R. (2020, June 2). *House-rent-prediction-dataset*. Kaggle. Retrieved October 14, 2022, from <https://www.kaggle.com/datasets/rkb0023/houserentpredictiondataset>

Florida, R. (2016, September 8). America's biggest college towns. Bloomberg.com. Retrieved from https://www.bloomberg.com/news/articles/2016-09-08/america-s-biggest-college-towns

*supervised learning*. scikit. (n.d.). Retrieved October 14, 2022, from <https://scikit-learn.org/stable/supervised_learning.html#supervised-learning>