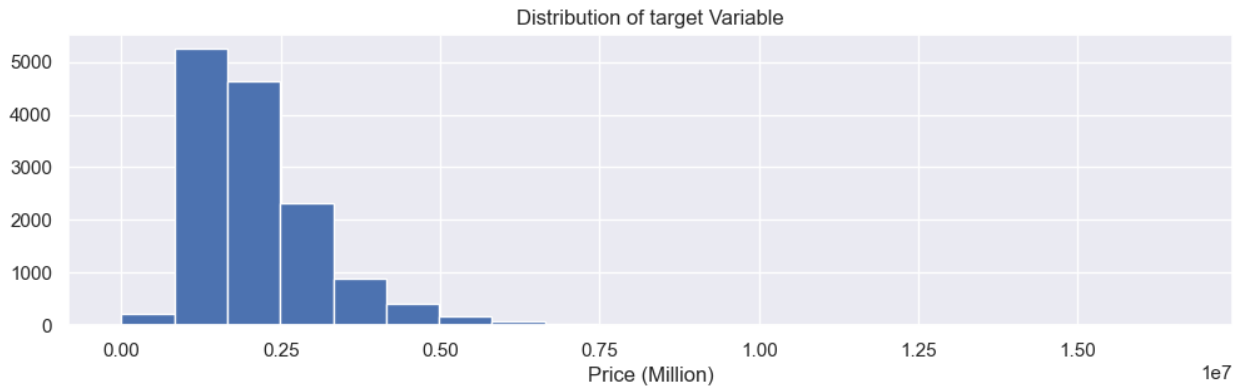
Oluwafemi's Executive Report On DSN AI Bootcamp Qualification 2023 Hackathon

Started my analysis by assessing the dataset provided, the train, test, and submission file. After observing I noticed I had 14,000 rows and 7 columns for my train dataset while the test dataset contains 6000 rows and 6 columns. Went on to assess the summary statistics for every feature (column in both my test and train dataset.

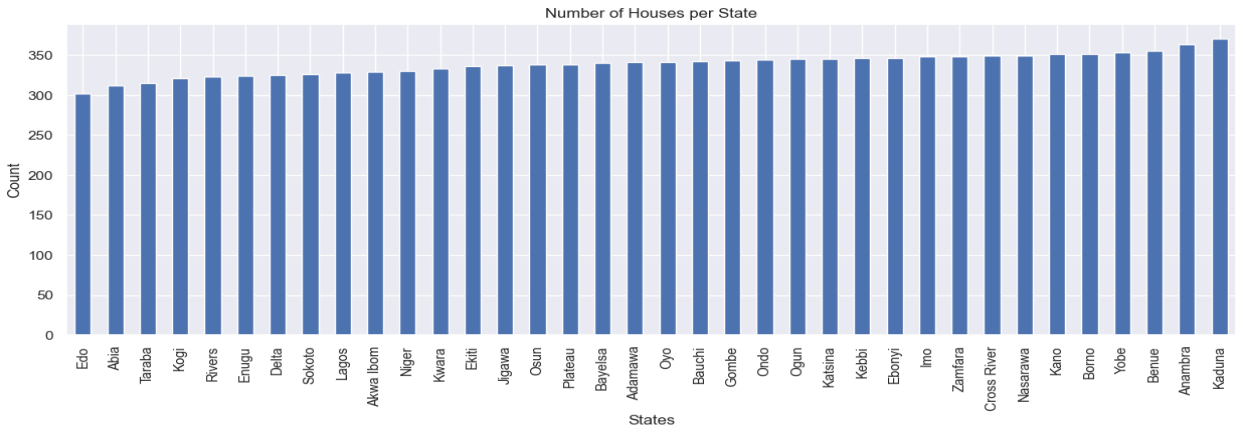
I plotted a histogram plot of my target variable where I noted the target variable which is the price was left skewed.

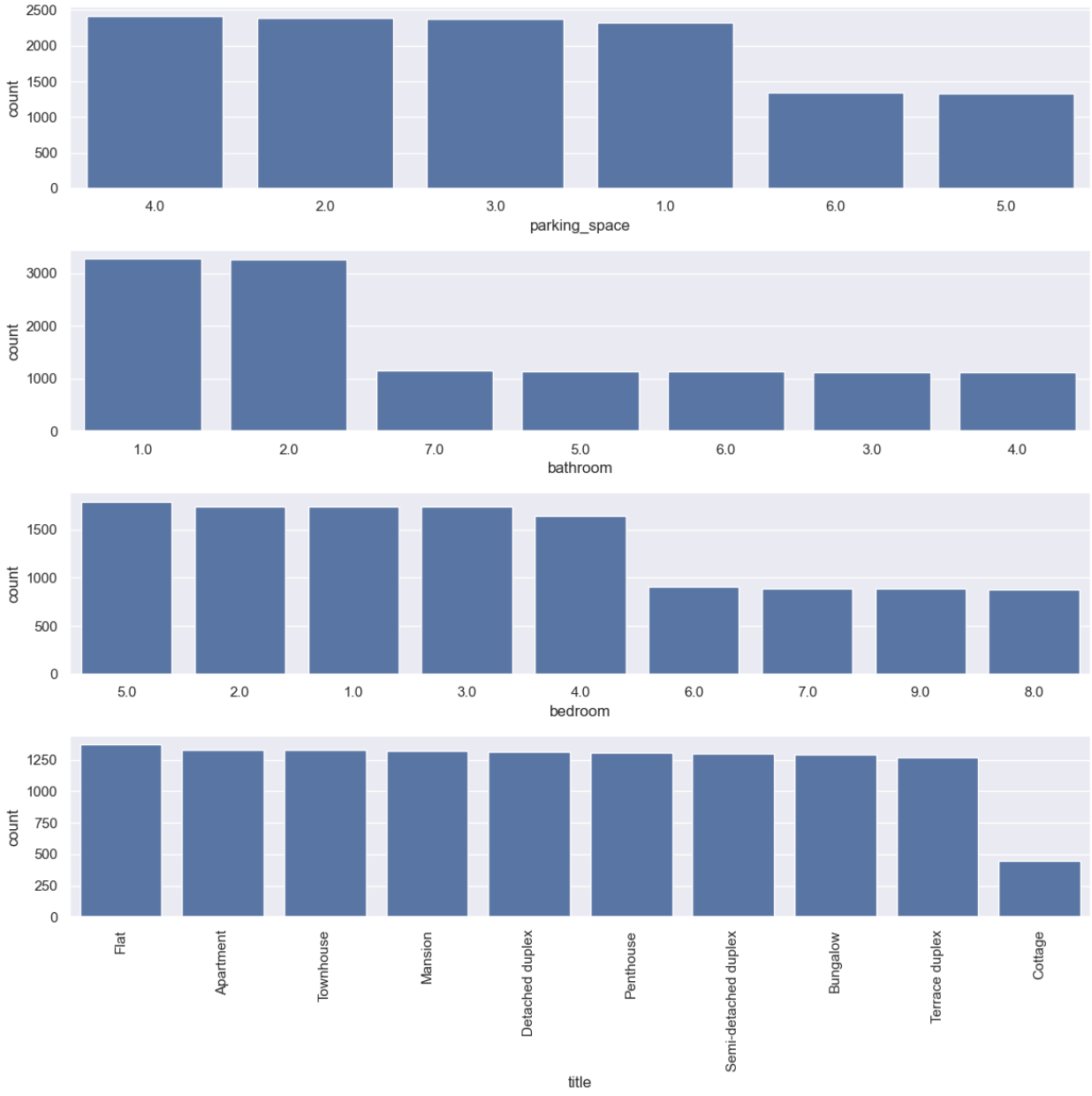


Proceeded to check for duplicates and the cardinality for each feature. The ID column particularly had a very high cardinality which could affect our model, concluded it was an identifier and not a good predictor.

Proceeded to check the percentage of missing values where I noted 5 variables (loc, parking space, bathroom, bedroom, and title) had about 13% missing values. I went further to visualize the distribution of the missing values using a heatmap.

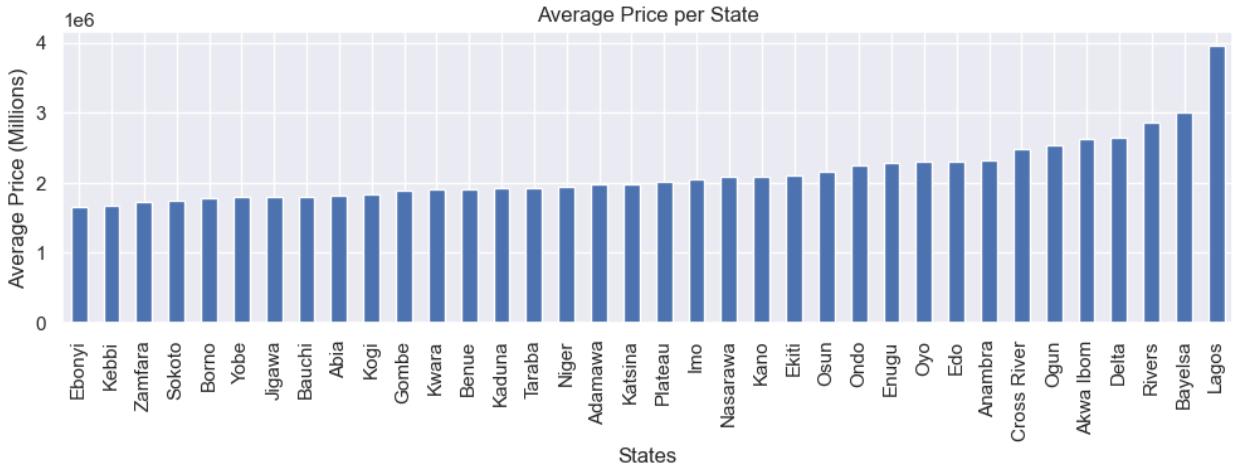
I performed univariate analysis using a count plot for the categorical variables as shown in the diagram below.



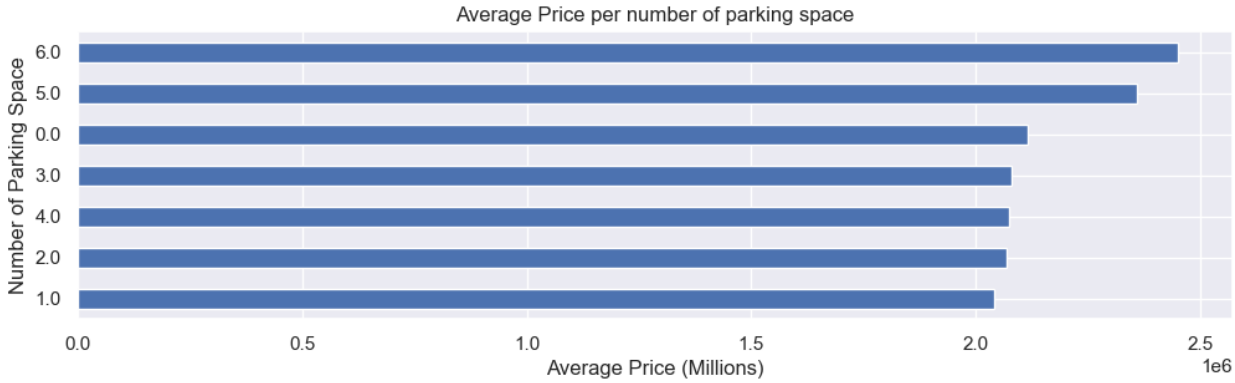


I filled the bedroom and bathroom column with standard deviation, and parking space with zero, then proceeded to dropped rows that had missing values for title and loc.

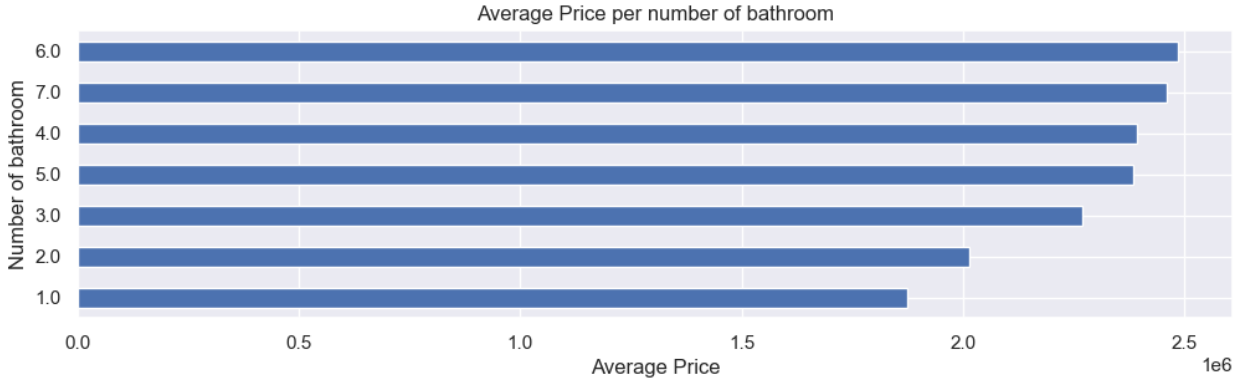
I proceeded with Bivariate analysis where I focused on the relationship of other variables to the average price as shown in the following diagram.



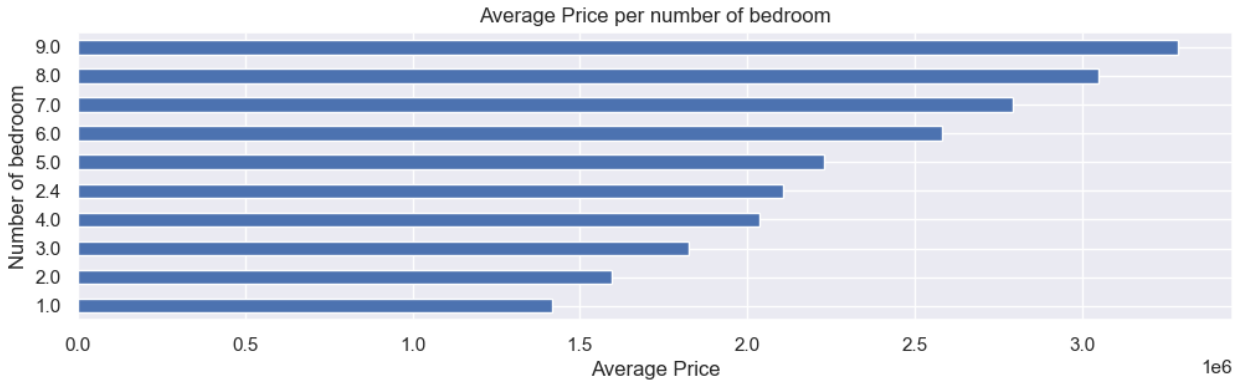
From diagram 2 it turns out that houses in the Lagos State had the highest average price



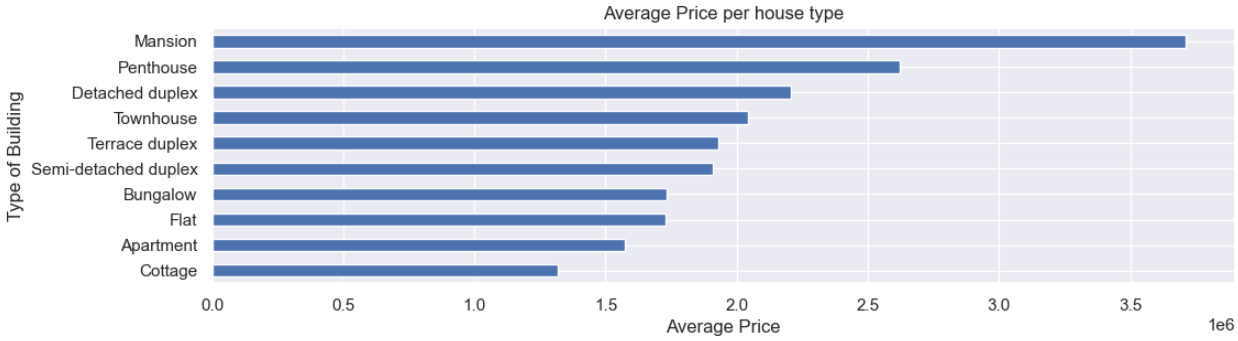
For parking, there was a direct relationship between the number of parking spaces and to average price except for houses with 3 parking spaces having more than houses with 4 parking spaces.



The same trend could be observed for the number of bathrooms to price except for houses with 4 bathrooms being more expensive than houses with 5 bathrooms and houses with 6 bathrooms being more expensive than houses with 7 bathrooms.

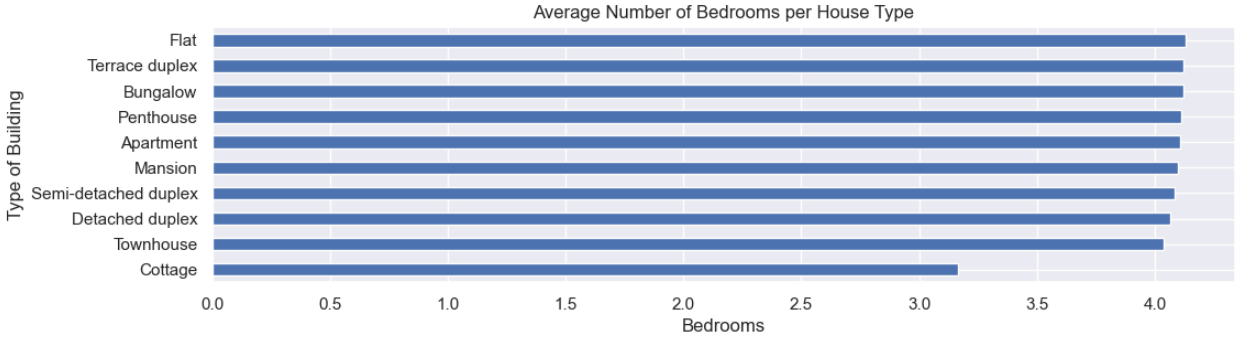


For the number of bedrooms, it turns out that there was a direct positive relationship with the average price.

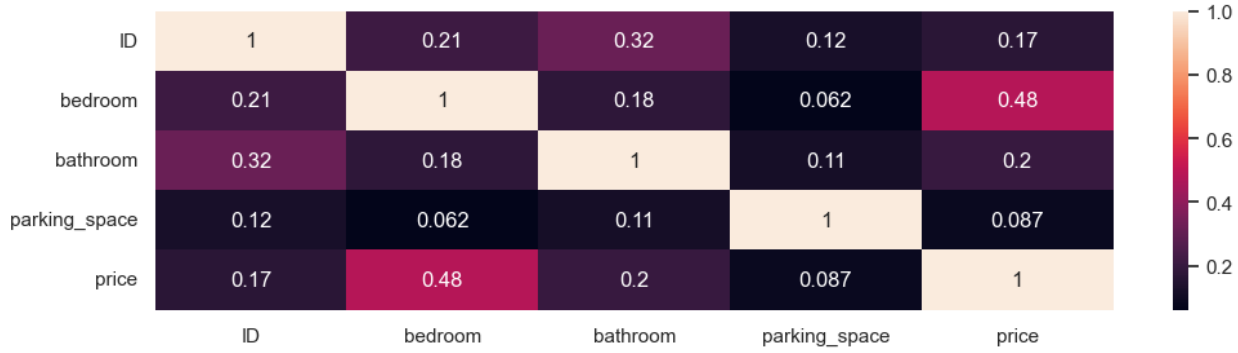


This shows the average price for each house type.

I also check the average number of bedrooms by house type.



I proceeded to check the relationship with all variables to check for correlation and multicollinearity as seen below.



I proceeded to perform some feature engineering based on the observation.

I started by manually encoding the house types by ranking them in order of the average price from the smallest to largest.

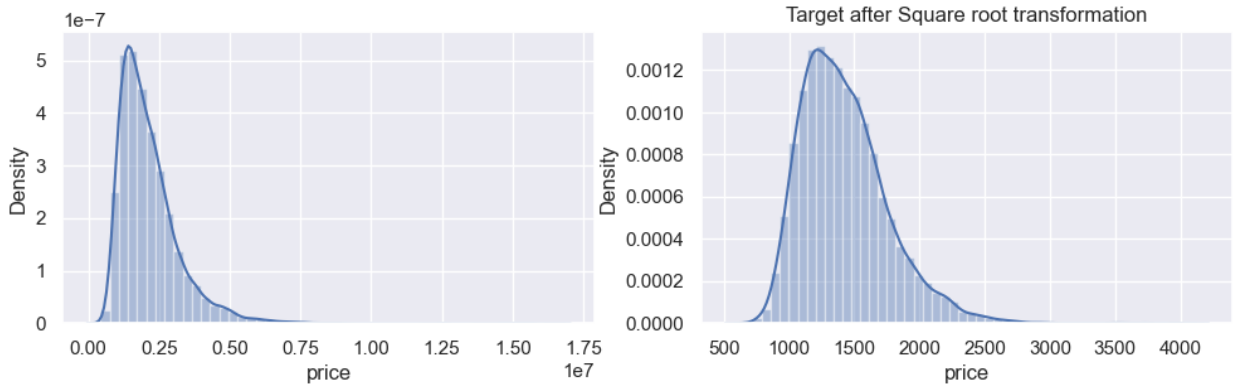
I continued to perform a one-hot encoding on the loc (state) and geopolitical zone column and later dropped the id, geopolitical and loc columns afterward.

Lastly, I created a new column labeled bedroom-to-bathroom ratio.

Model Training:

The next step was to train my model.

I performed a square root transformation on the target variable due to its skewness.



80% of my dataset was used to train while 20% to test.

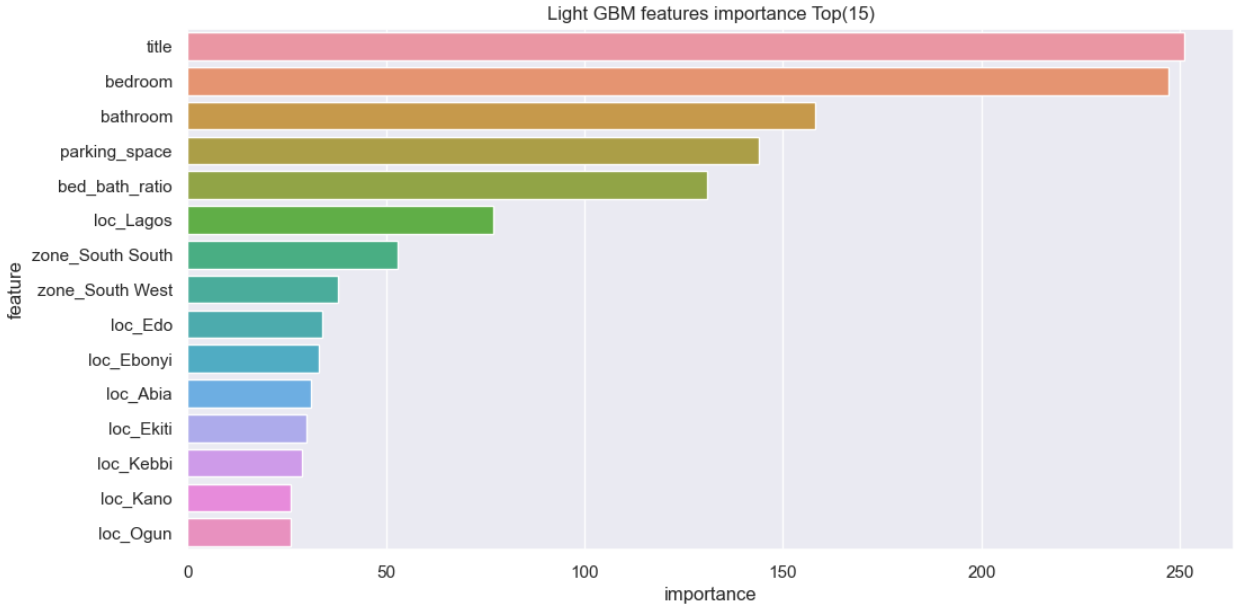
Baseline Model:

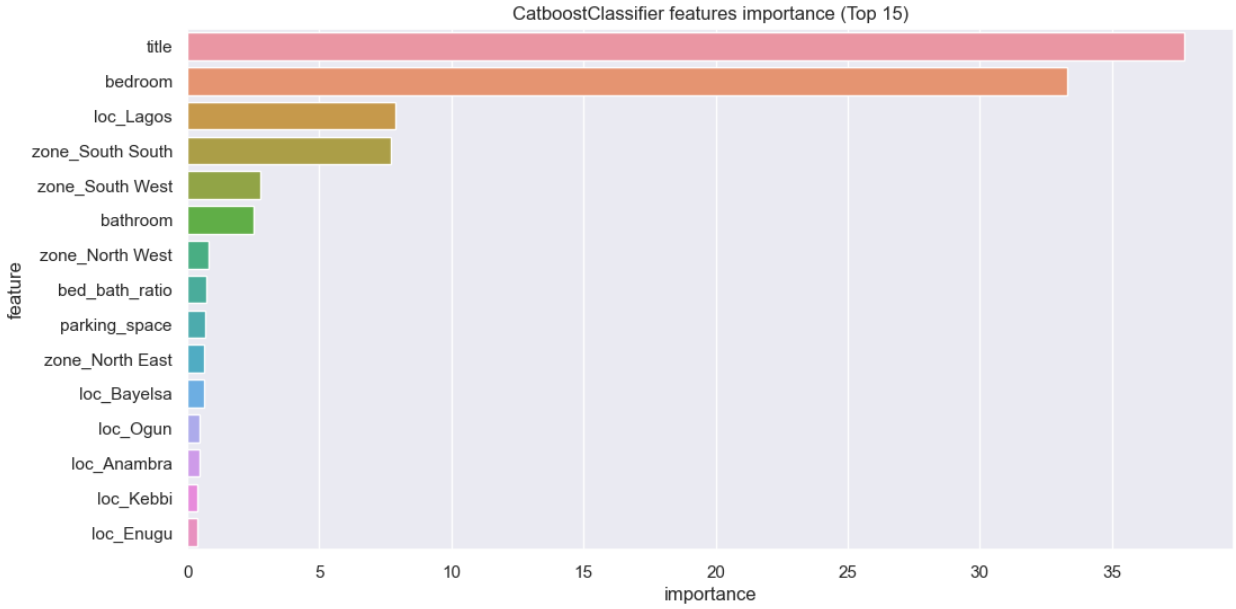
I used light GB to perform my baseline model with a rmse score of 610964.

To improve on the score, I tuned some hyperparameters and K-fold splitting techniques to avoid overfitt ing my model. In return, I got an average rmse score of 583677.

To improve the generalization of my model, I employed another model, the CatBoost Regressor and perf ormed some hyperparameter tunning and used the K-fold splitting method to avoid overfitting. In retur n, I got an average rmse score of 582572.

To interpret my result, I plotted a feature importance chart to show the top 15 variables that were stron g predictors of the house price with their respective coefficients as shown in the figure below.





Lastly, I blended the result of the two models with 70% of the catboost model and 30% of the light GB m odel to obtain my final submission.