WAZOBIA REAL ESTATE HOUSE PRICE PREDICTION EXECUTIVE REPORT

PROBLEM STATEMENT

Wazobia Real Estate Limited is a prominent real estate company operating in Nigeria. With a vast portfolio of properties, they strive to provide accurate and competitive pricing for houses. However, they have been facing challenges in accurately predicting the prices of houses in the current market. To overcome this hurdle, Wazobia Real Estate Limited is seeking the expertise of data scientists like you to develop a robust predictive model(Zindi, 2023).

The objective of this hackathon is to create a powerful and accurate predictive model that can estimate the prices of houses in Nigeria. By leveraging the provided dataset, you will analyse various factors that impact house prices, identify meaningful patterns, and build a model that can generate reliable price predictions. The ultimate goal is to provide Wazobia Real Estate Limited with an effective tool to make informed pricing decisions and enhance their competitiveness in the market(Zindi, 2023).

SOLUTION

In every data science project, there are four basic steps to follow. There are data collection, exploratory data analysis, data preprocessing, model building evaluation.

DATA COLLECTION

Wazobia real estate has provided a dataset to solve this challenge on zindi. The data contain 7000 entries for the training and 3000 entries for testing the performance of the model on real word data(test data).

EXPLORATORY DATA ANALYSIS

The train data contains 14,000 observations and 7 features while the test data contains 6,000 observations and 6 columns. The test data has 6 features as a result of absence of the target feature or column. From the 7 columns in the train data, 4 columns are numerical while 3 columns are categorical.

The dataset has no duplicate but some missing values and the names of all features present in the dataset are ID, loc, title, bedroom, bathroom, parking_space, price. The ID column is a unique number for all entries in the dataset, the loc feature stands for location of the house, title stands for title of the house like mansion, duplex etc, bedroom stands for number of bedroom in the house, bathroom stands for number of bathroom in the house, parking_space is the number of parking spaces in the house and price is the cost of the house which is the target variable.

Univariate Analysis

The building title with the highest number of houses is the Flat, followed by Apartment and the least is cottage as shown in the image below

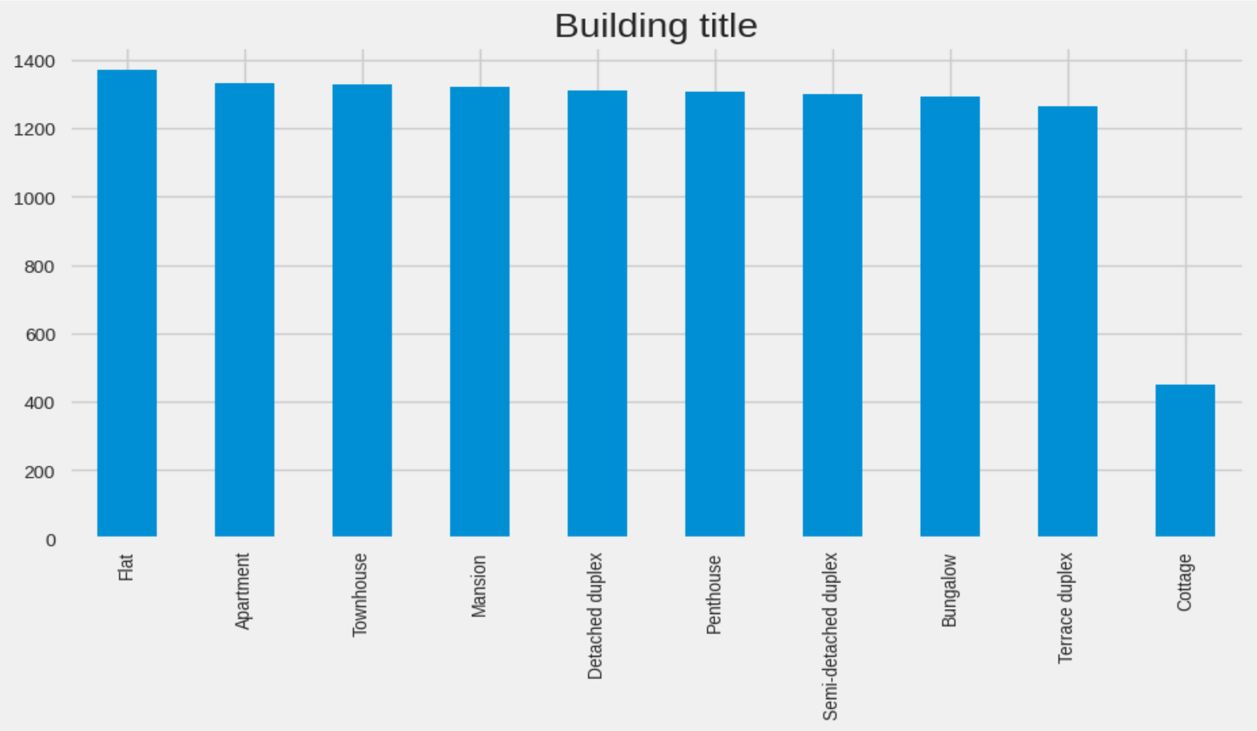


Fig 1: Distribution of the house title

Most of the houses are located in Kaduna, Anambra, Benue while Edo state has the fewest number of houses as shown in fig 2 below

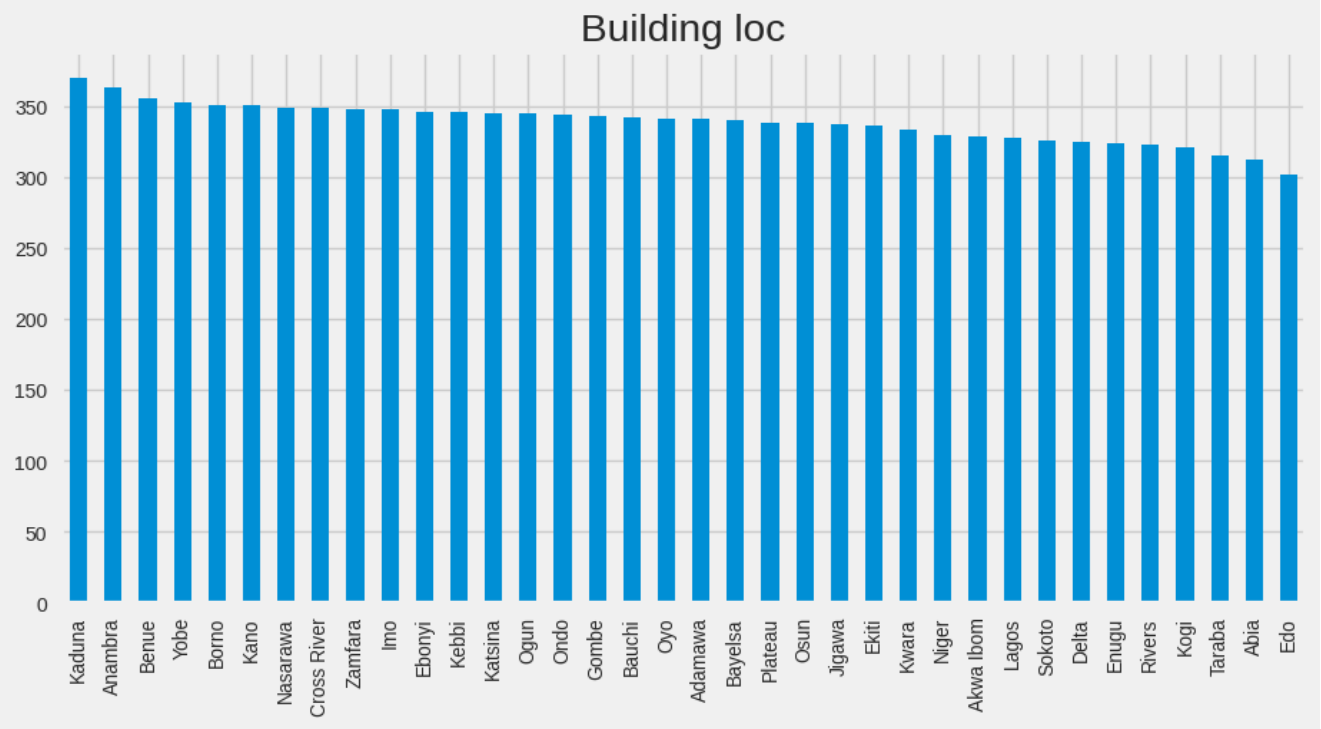


Fig 2: Distribution of house location

Most of the houses has number of parking spaces ranging from 1 - 4 with the highest been 4 and the least are 6 and 5 having almost the same values as in fig 3

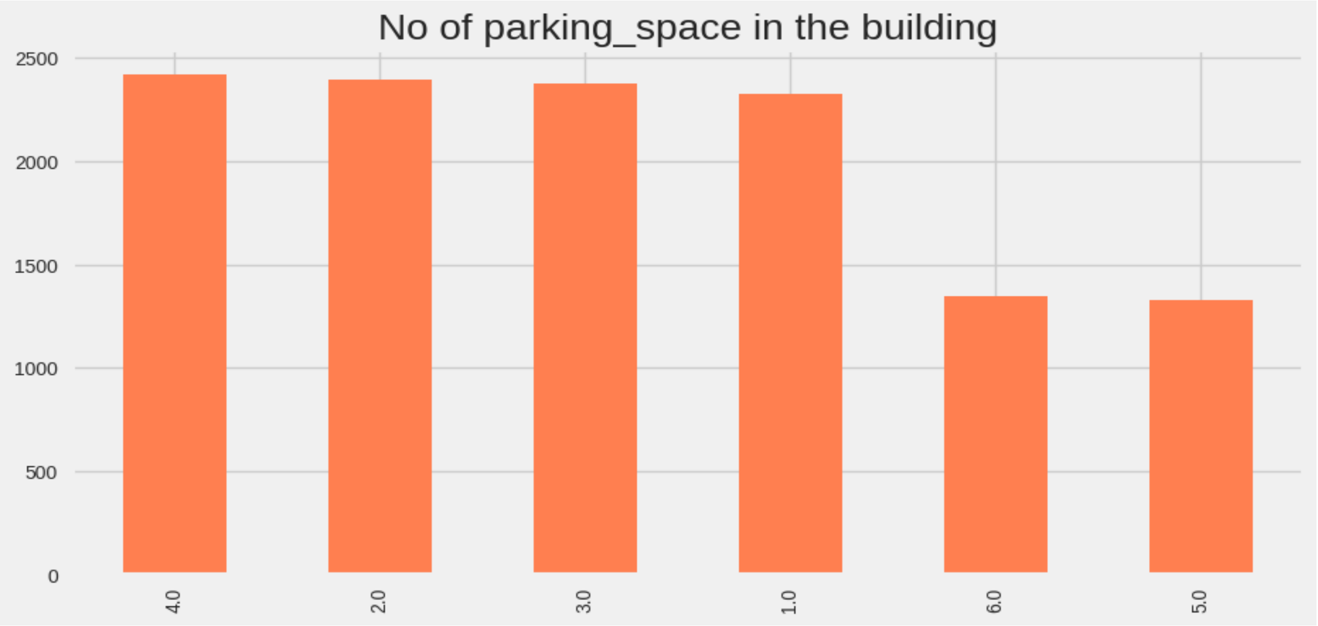


Fig 3: Distribution of number of parking spaces in the houses

Most of the houses has number of parking spaces ranging from 1 - 5 with the highest been 5 and the least ranging from 6 -9 with the lowest 8 as in fig 4

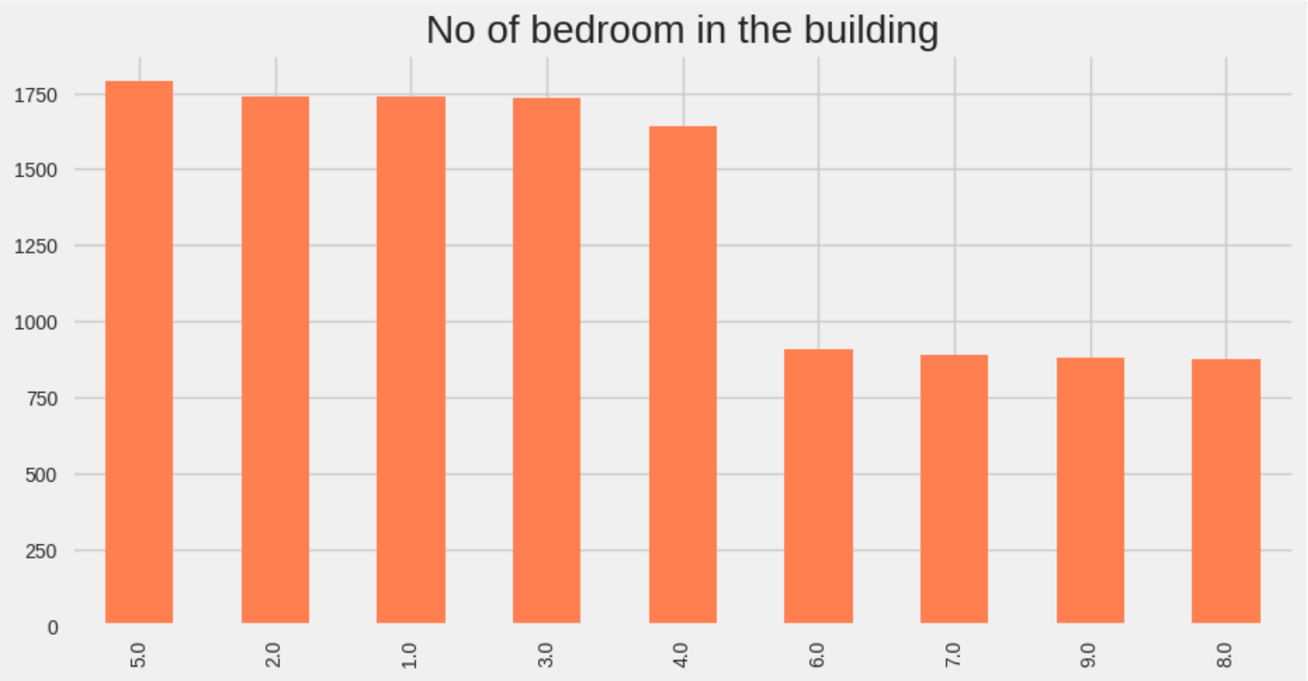


Fig 4: Distribution of the number of bedroom

Most of the houses has 1 or 2 number of bathroom with the highest been 1(Although, there is no significant difference between 1 and 2) and the least ranging from 3 -7 with the lowest 4(Although no significant difference between 3, 4, 5, 6, 7) as in fig 5

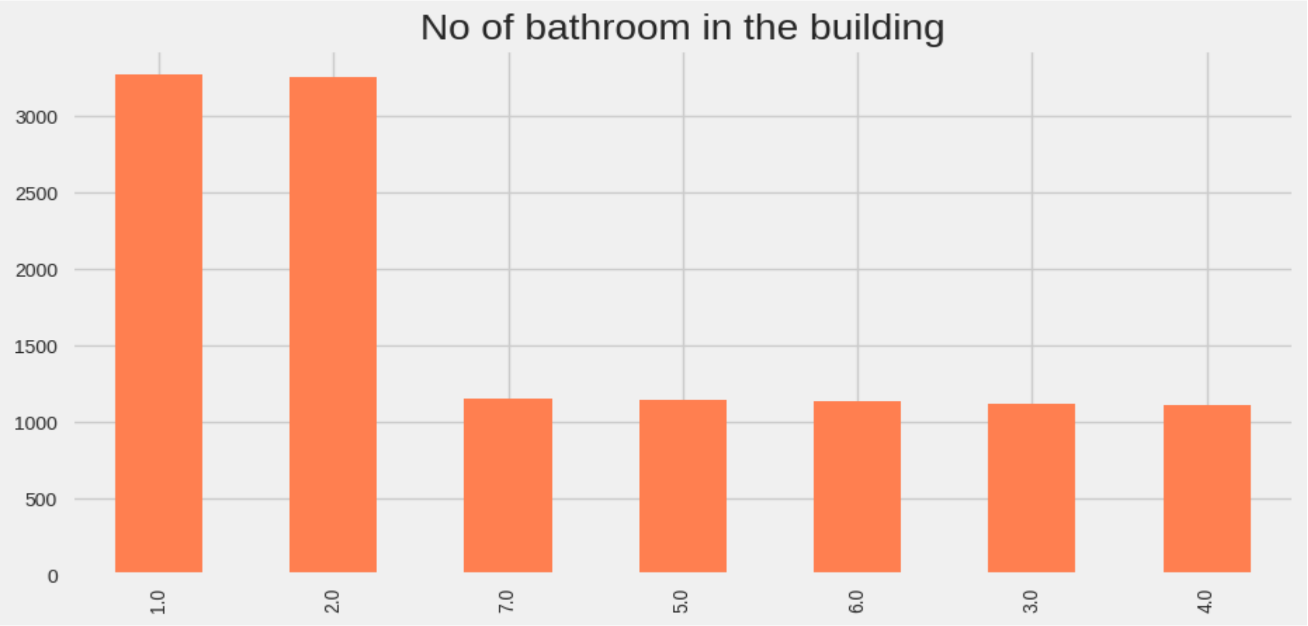


Fig 5: Distribution of the number of bathroom available in the houses

Bi-variate Analysis

- Most of the detached duplex and semi detached duplex house title have 5 bedrooms
- ,Most of the houses with title Apartment, Terrace, Townhouse, Flat and cottage have 1 bedroom
- Most of Mansion buildings has 3 bedrooms
- Most of bungalow and penthouse buildings has 2 bedrooms

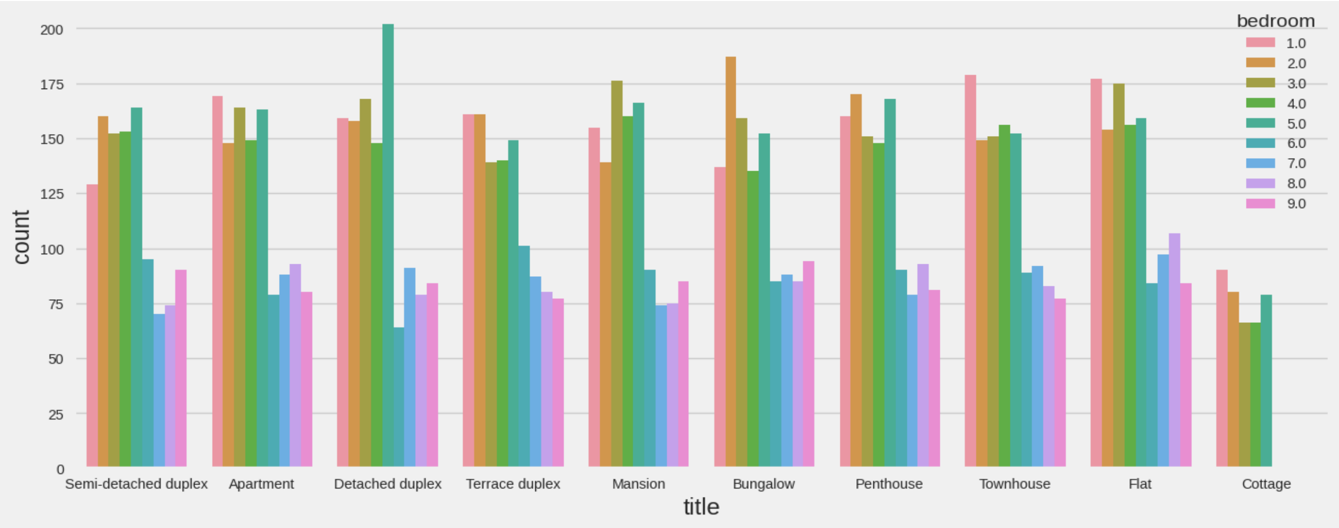


Fig 6: Relationship between title and number of bedrooms

- Most of the houses with title Apartment, Terrance, mansion, bungalow and townhouse have 1 bathroom
- Most of the houses with title Semi-detached, detached, penthouse, flat and cottage have 2 bathrooms

Shown in Fig 7 below

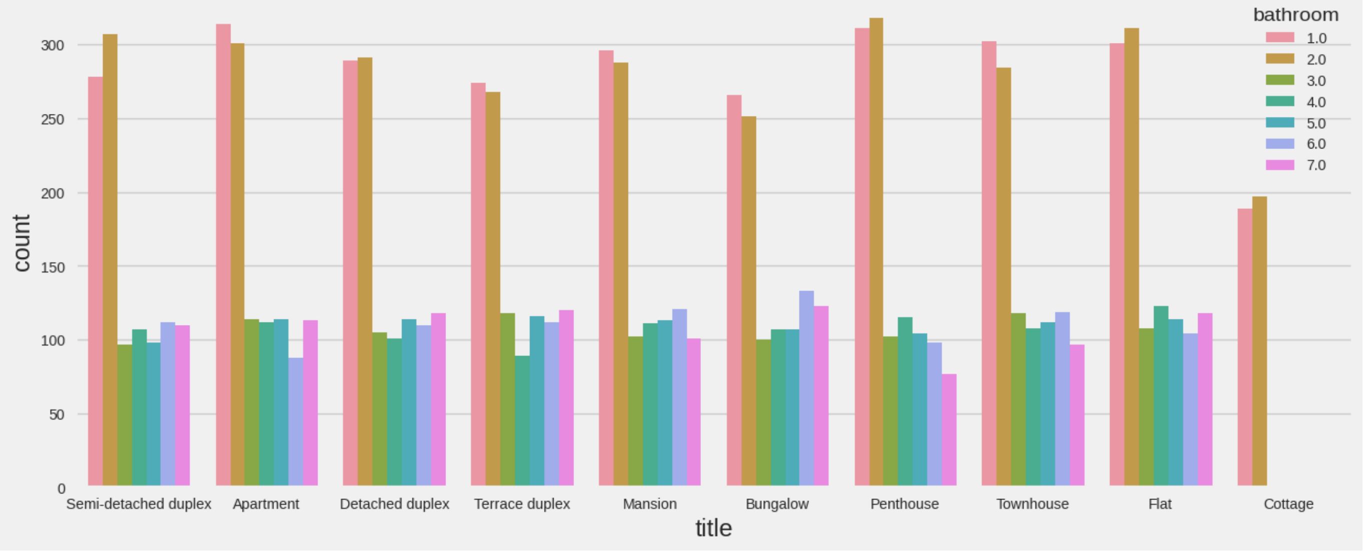


Fig 7: Relationship between title and number of bathrooms

Apartment buildings has 4 and highest number of parking spaces as shown in fig 8 below

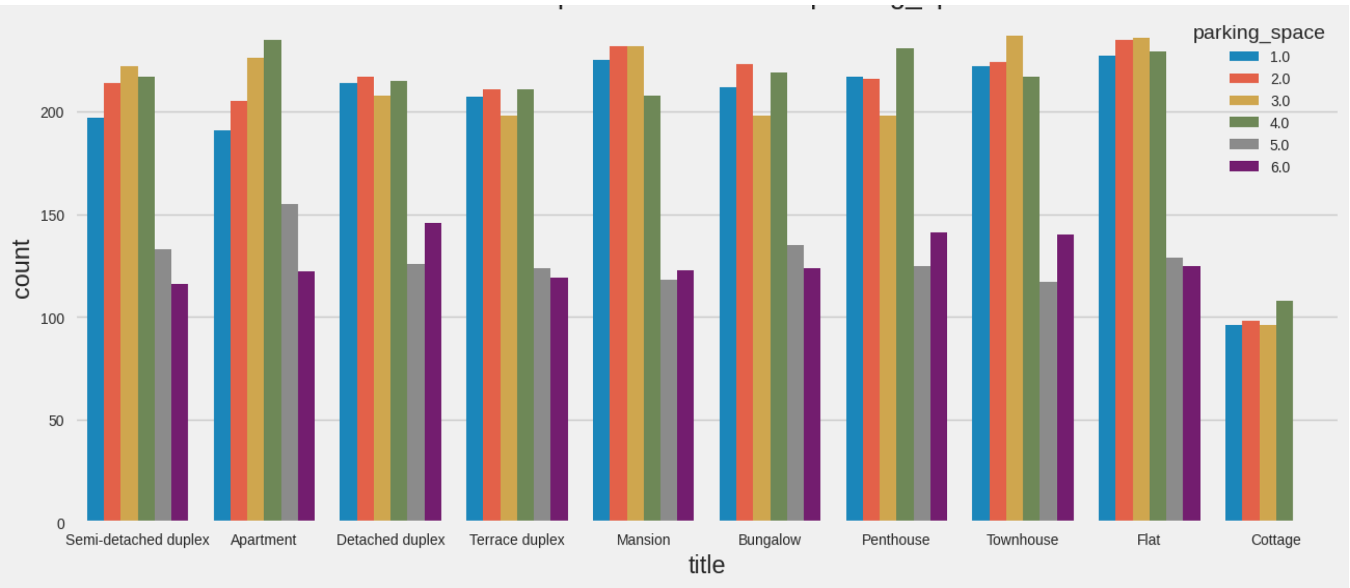


Fig 8: Relationship between title and parking space

Other notable findings are as follows:

- Highest house costs 16568486.16 and lowest 431967.29
- The Most expensive building is found in Lagos and the least expensive building is found in Gombe

- The Most expensive House title is Mansion and the least expensive House title is Cottage
- Houses with 2 bathrooms are least expensive
- Houses with 5 bedrooms are the most expensive and houses with 1 bedroom are the least expensive
- Houses with 3 parking spaces are the most expensive and houses with 1 parking space are the least expensive

None of the features were found to correlated as shown below in fig 9

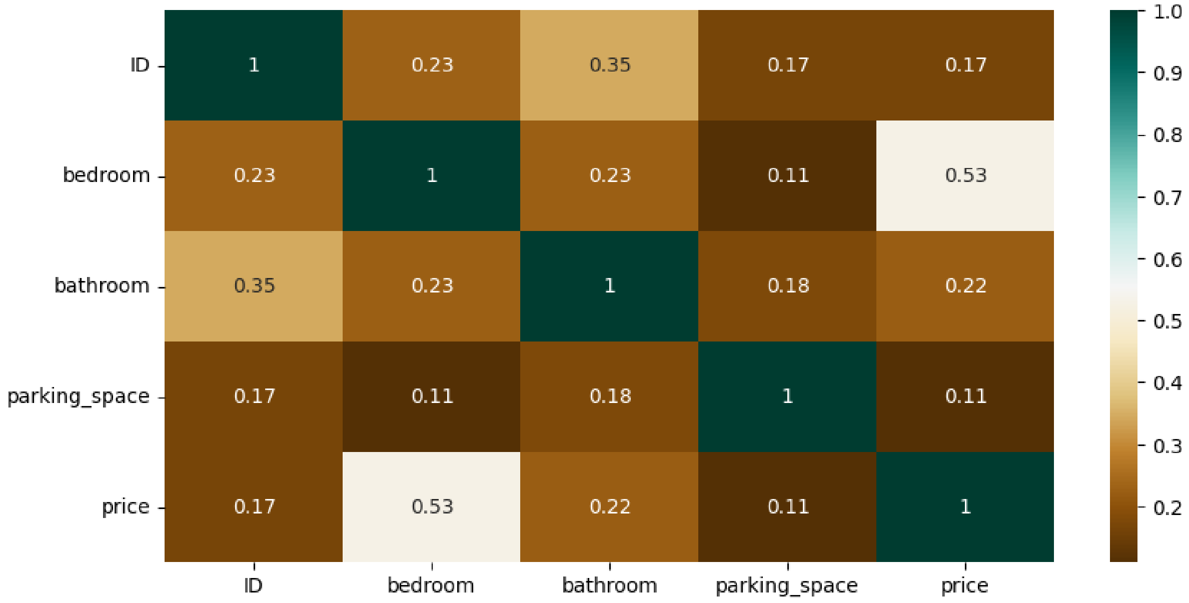


Fig 9: Correlation plot of the features

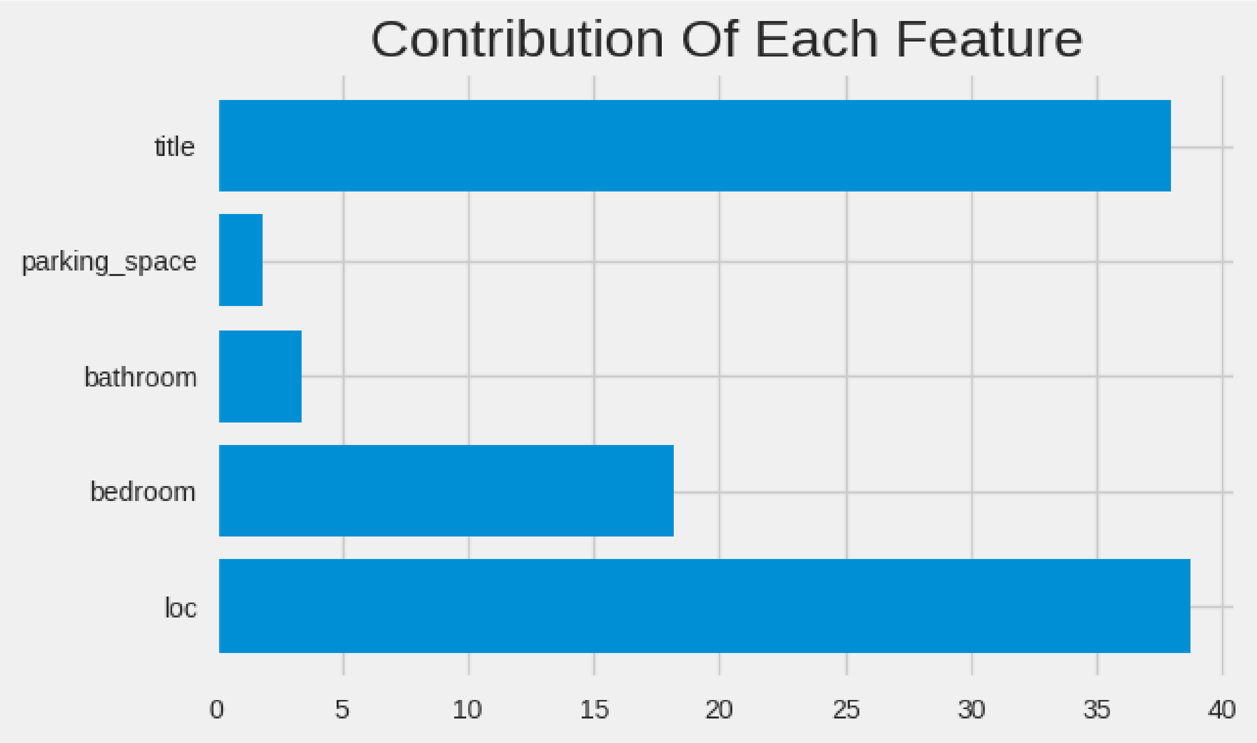
DATA PREPARATION

The categorical columns with missing values were filled with the mode in which the numericals were filled with the median and the categorical values were encoded using label encoded. The dataset was oversampled using a random oversampler to balance the class present in the title column. After the oversampling, the observation increased to 30940. The ID column was also dropped to avoid data leakage.

MODEL BUILDING

Before building the model, the dependent and target feature(price) was separated and then the dataset was further divided into training and test set in the ratio 9:1. 90% of the data was used to train the model and the remaining 10% was used to test the performance of the model before using the final real world test dataset.

The training set data was fitted to a lightgm, catboost and xgboost most and after the evaluation and submission, catboost performed best.



The title and location of the house are the major factors responsible for the price of houses.

Parameter tuning was used to improve the performance of the model but the model didn’t improve even after tuning the parameters with optuna.

To improve the model more, getting more training data can be more effective as all the features present are very important and none can be dropped.

Link to hackathon: https://zindi.africa/competitions/free-ai-classes-in-every-city-hackathon-2023