Project Description

Description:

The Wazobia Real Estate Limited Predictive Modeling Hackathon is a data science competition focused on accurately predicting house prices in Nigeria. Participants will work with a provided dataset containing property features and prices. The objective is to build a powerful predictive model that helps Wazobia Real Estate make informed pricing decisions and enhance their market position.

Goal:

Develop a robust predictive model capable of estimating house prices accurately in Nigeria. The model should leverage the dataset's features to generate reliable price predictions for Wazobia Real Estate's properties.

Objectives:

1. Data Analysis: Conduct exploratory data analysis to understand the dataset's characteristics, identify outliers, and address missing values.

2. Model Building: Build predictive models using machine learning algorithms, aiming for high accuracy in price predictions.

3. Model Evaluation: Evaluate model performance using appropriate metrics to select the best-performing model.

4. Interpretability: Gain insights into the factors influencing house prices by interpreting the model results.

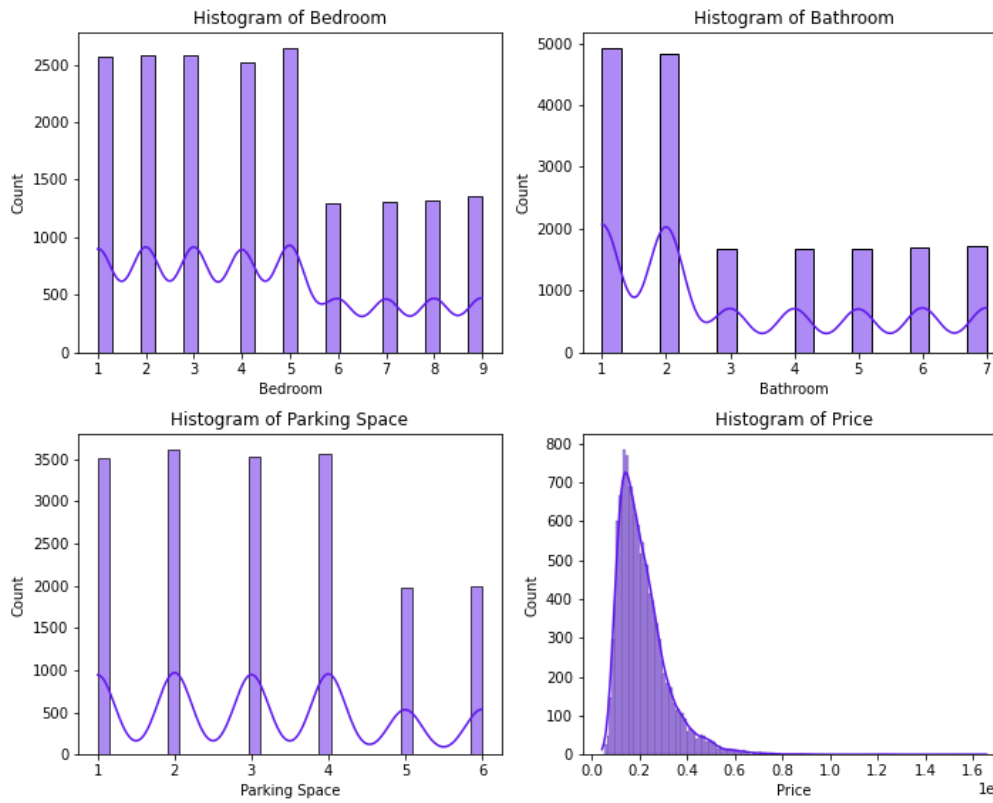
5. Business Impact: Provide actionable insights for Wazobia Real Estate to improve pricing strategies and market competitiveness.

By achieving these objectives, participants will help Wazobia Real Estate overcome pricing challenges, enhance decision-making, and deliver enhanced value to their customers.

Exploratory Data Analysis(EDA)

Summary Statistics

In this dataset, we have information on various features related to houses, such as the number of bathrooms, bedrooms, parking spaces, and their corresponding prices. Let's take a closer look at the summary statistics for each feature:



Bathroom:

- Count: 18,195
- Mean: 3.12
- Standard Deviation: 2.04
- Minimum: 1.00
- 25th Percentile: 1.00
- Median (50th Percentile): 2.00
- 75th Percentile: 5.00
- Maximum: 7.00

The distribution of the number of bathrooms appears to be positively skewed, with most houses having 1, 2, or 3 bathrooms. Some houses have up to 7 bathrooms, which might be considered outliers. The mean

(3.12) is less than the median (2.00), indicating a positive skewness. The distribution is right-skewed, with a longer right tail. The 75th percentile (5.00) is higher than the median, further confirming the presence of right skewness. This suggests that a few houses with a higher number of bathrooms contribute to the elongated right tail.

Bedroom:

- Count: 18,201
- Mean: 4.32
- Standard Deviation: 2.45
- Minimum: 1.00
- 25th Percentile: 2.00
- Median (50th Percentile): 4.00
- 75th Percentile: 6.00
- Maximum: 9.00

The distribution of the number of bedrooms seems to be approximately symmetrical although slightly positively skewed. Most houses have 2 to 6 bedrooms, with a few outliers having as few as 1 bedroom and as many as 9 bedrooms. The mean (4.32) is greater than the median (4.00), indicating a slight positive skewness. The distribution might be slightly right-skewed, with a slightly longer right tail. However, the difference between the mean and median is relatively small, suggesting a relatively symmetrical distribution.

Parking Space:

- Count: 18,189
- Mean: 3.16
- Standard Deviation: 1.60
- Minimum: 1.00
- 25th Percentile: 2.00
- Median (50th Percentile): 3.00
- 75th Percentile: 4.00
- Maximum: 6.00

The distribution of parking spaces also appears to be approximately symmetrical. Most houses have 2 to 4 parking spaces, with a few outliers having as few as 1 space and as many as 6 spaces. The mean (3.16) is greater than the median (3.00), indicating a slight positive skewness. The distribution might be slightly right-skewed, with a slightly longer right tail. Similar to the bedroom feature, the difference between the mean and median is relatively small, suggesting a relatively symmetrical distribution.

Price:

- Count: 14,000
- Mean: 2,138,082
- Standard Deviation: 1,083,057
- Minimum: 431,967.3
- 25th Percentile: 1,393,990
- Median (50th Percentile): 1,895,223
- 75th Percentile: 2,586,699

- Maximum: 16,568,490

The distribution of house prices is highly positively skewed, with most prices falling in the lower range. Some houses have significantly higher prices, leading to outliers. The mean (2,138,082) is greater than the median (1,895,223), indicating a positive skewness. The distribution is likely to be right-skewed, with a longer right tail. The 75th percentile (2,586,699) is also higher than the median, further confirming the presence of right skewness. This suggests that a few houses with higher prices contribute to the elongated right tail, pulling the mean higher than the median.

- Outliers are present in the bathroom, bedroom, parking space, and price features.
- These outliers may significantly impact the model's performance and predictions, and we may need to handle them carefully during the modeling process.
- Depending on the model chosen, we can consider various techniques such as data transformation, outlier removal, or robust modeling to address the effect of outliers on our predictive model.

Remarks on Missing Values:

loc: 1813 missing values (12.95%)

The "loc" feature has 1813 missing values, accounting for approximately 12.95% of the total data. These missing values represent the location of the houses in the dataset. It's essential to handle these missing values appropriately to avoid any biases in the analysis or predictive models.

title: 1722 missing values (12.30%)

The "title" feature has 1722 missing values, which accounts for approximately 12.30% of the total data. The "title" refers to the type of property or house (e.g., duplex, apartment, mansion). Addressing these missing values properly is crucial for any analysis or model that involves property types.

bedroom: 1799 missing values (12.85%)

The "bedroom" feature has 1799 missing values, representing approximately 12.85% of the total data. The number of bedrooms is an important factor affecting house prices. Handling these missing values appropriately is essential to avoid bias and ensure accurate price predictions.

bathroom: 1805 missing values (12.89%)

The "bathroom" feature has 1805 missing values, accounting for approximately 12.89% of the total data. Similar to the number of bedrooms, the number of bathrooms significantly impacts house prices. Properly handling these missing values is crucial for accurate price estimation.

parking_space: 1811 missing values (12.93%)

The "parking_space" feature has 1811 missing values, representing approximately 12.93% of the total data. Adequately dealing with these missing values is essential, as parking space availability can be a significant factor in determining house prices.

price: 0 missing values (0.00%)

Fortunately, the target variable "price" has no missing values, which is essential for building predictive models. A complete target variable ensures that we can use the available data to train and test the model effectively.

Addressing missing values appropriately is a critical step in the data preprocessing phase. The choice of the imputation method should be carefully considered based on the data distribution and the impact of missing values on the analysis or predictive model. Proper handling of missing values will lead to more robust and accurate predictions for house prices in Nigeria.

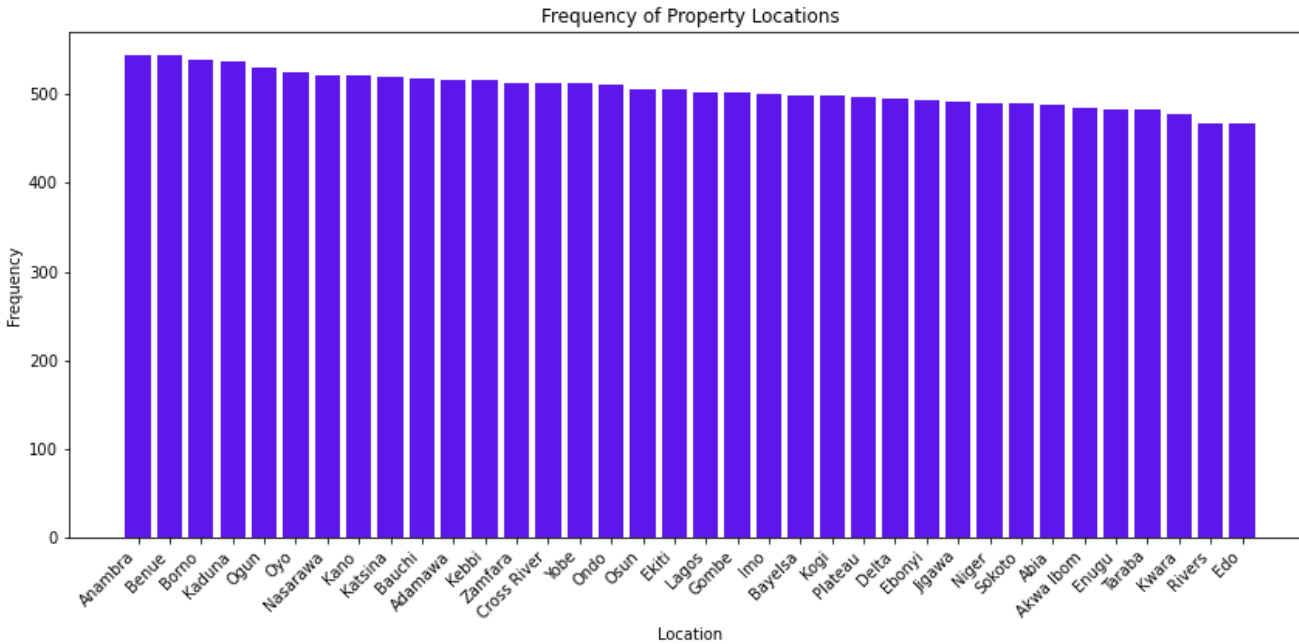
Based on the given data, we have two categorical features: 'loc' and 'title'. Let's analyze each feature and provide insights. Upon analyzing the frequency counts of property titles and locations, we can make the following observations:

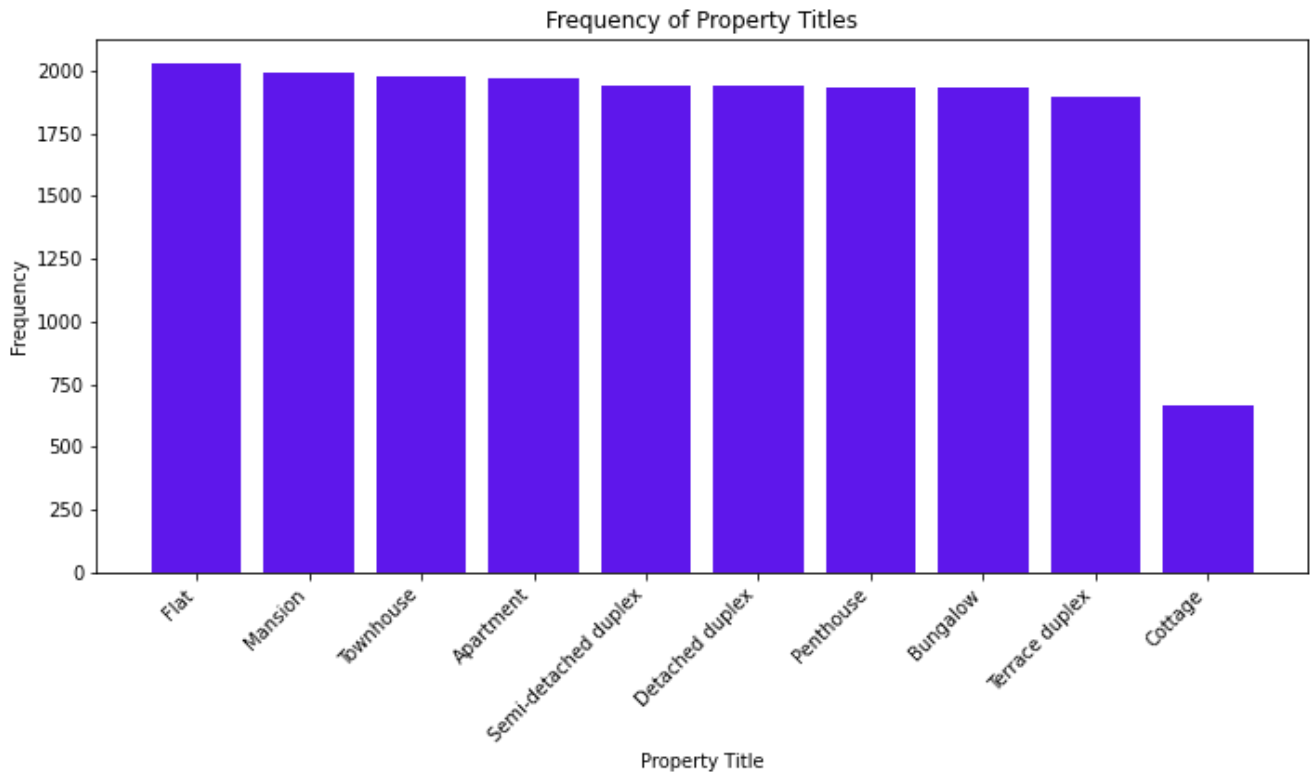
Property Titles:

- The most common property type is "Flat," with a frequency of 2026. Followed closely are "Mansion" and "Townhouse" with frequencies of 1995 and 1980, respectively.
- "Apartment," "Semi-detached duplex," and "Detached duplex" have similar frequencies, ranging from 1938 to 1973.
- "Cottage" has the lowest frequency among the property titles, with only 665 occurrences.

Property Locations:

- The states "Anambra" and "Benue" have the highest number of properties, with each having 543 occurrences in the dataset.
- Several states like "Borno," "Kaduna," "Ogun," and "Oyo" also have a high frequency of properties, ranging from 530 to 543.
- On the other hand, states like "Edo," "Rivers," "Kwara," and "Taraba" have the lowest number of properties, ranging from 467 to 477.


Frequency of Property Locations
Frequency
0
100
200
300
400
500
Anambra
Benue
Borno
Kaduna
Ogun
Oyo
Nasarawa
Kano
Katsina
Bauchi
Adamawa
Kebbi
Zamfara
Cross River
Yobe
Ondo
Osun
Ekiti
Lagos
Gombe
Imo
Bayelsa
Kogi
Plateau
Delta
Ebonyi
Jigawa
Niger
Sokoto
Abia
Akwa Ibom
Enugu
Taraba
Kwara
Rivers
Edo
Location


Frequency of Property Titles
Frequency
0
250
500
750
1000
1250
1500
1750
2000
Flat
Mansion
Townhouse
Apartment
Semi-detached duplex
Detached duplex
Penthouse
Bungalow
Terrace duplex
Cottage
Property Title

Missing Values:

- The columns "loc" and "title" have some missing values, with 1813 and 1722 missing entries, respectively. These missing values need to be handled appropriately during data preprocessing to avoid bias in the analysis and modeling.

To make better data-driven decisions, further exploration is necessary. Addressing missing values is crucial for building a reliable predictive model. We can consider strategies like imputation or removal of rows with missing values based on the extent of missing data and its potential impact on the analysis and predictions.

From the distribution of property titles and locations, we can make the following **recommendations** for Wazobia Real Estate Limited:

Property Type Diversification: The analysis shows that "Flat," "Mansion," and "Townhouse" are the most common property types in the dataset. Wazobia Real Estate Limited can use this insight to focus on developing and marketing more properties of these popular types. Additionally, they can explore opportunities to diversify their portfolio by offering properties in less common types, such as "Cottage" or "Penthouse," to cater to a broader range of customer preferences.

Location-Based Market Strategy: The dataset reveals that certain states, like "Anambra" and "Benue," have a higher concentration of properties compared to others. Wazobia can capitalize on this information to design location-specific marketing and pricing strategies. They can also consider expanding their presence in states with lower property frequencies to tap into emerging markets.

Pricing Insights: Wazobia can analyze how property prices vary across different property types and locations. By understanding the price distribution, they can set competitive pricing for their properties. They should pay attention to outliers in the price distribution, as these may indicate overpriced or underpriced properties that need adjustment.

Missing Value Handling: The analysis shows that there are missing values in the "loc" and "title" columns. Wazobia should carefully handle these missing values during data preprocessing. Imputation techniques can be used to fill in missing values or, if necessary, rows with significant missing data can be removed. Ensuring data completeness will contribute to the accuracy and reliability of the predictive model.

Customer Preference Analysis: To further understand customer preferences, Wazobia can conduct additional surveys or collect more data on customer requirements, such as preferred property features, amenities, and locations. By incorporating customer feedback, they can develop properties that align with the demands of the target market.

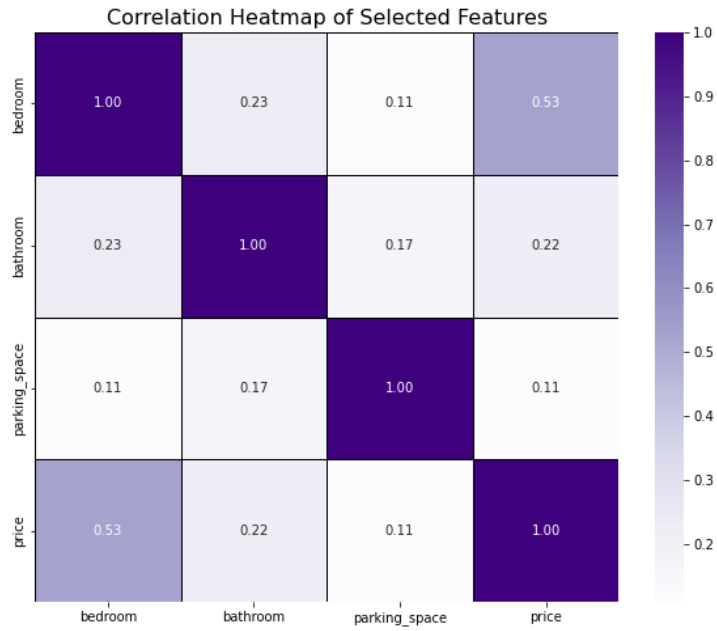
Predictive Modeling: Building a predictive model based on historical property prices and various property features can significantly benefit Wazobia. The model can be used to forecast property prices accurately, helping the company make data-driven pricing decisions and stay ahead in the competitive real estate market.

Market Trends Monitoring: It is essential for Wazobia to continuously monitor market trends, changes in customer preferences, and the competitive landscape. They can use this information to adapt their

strategies and offerings accordingly, ensuring they remain competitive and meet the evolving demands of the market.

Correlation Matrix Report

The correlation matrix reveals the relationship between different features and the target variable (price).



Bedroom and Price: A moderate positive correlation of 0.53 indicates that the number of bedrooms has a significant influence on house prices. As the number of bedrooms increases, the price tends to rise.

Bathroom and Price: The bathroom feature shows a weak positive correlation (0.22) with house prices. While it still contributes to the price prediction, its impact is not as strong as the number of bedrooms.

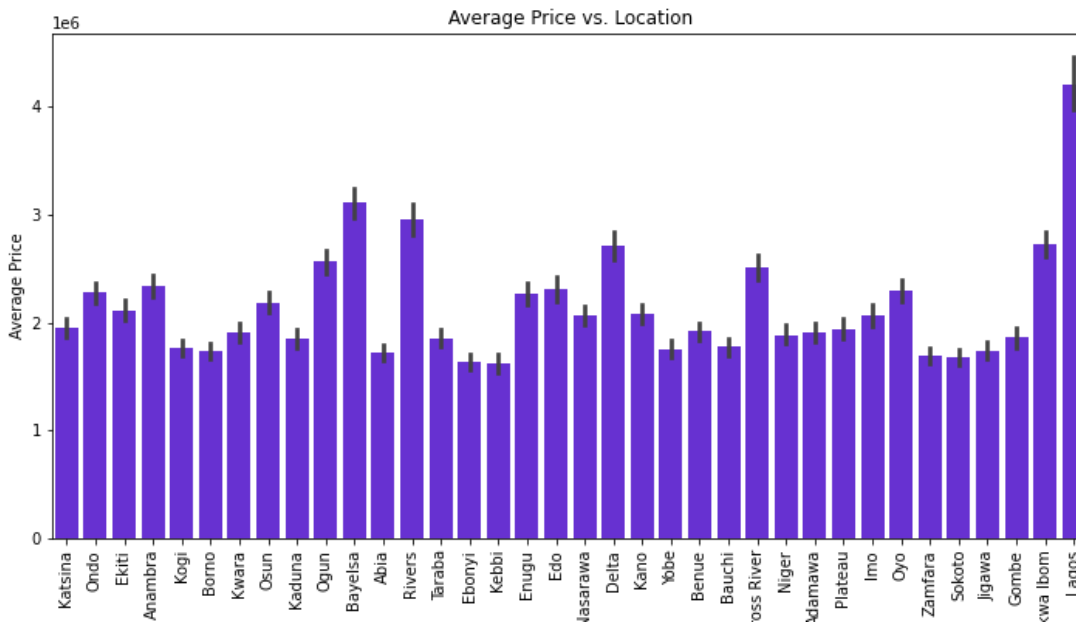
Parking Space and Price: The parking space feature exhibits a weak positive correlation (0.11) with house prices. It suggests that having more parking spaces can slightly affect the house price.

Overall, the correlation matrix highlights that the number of bedrooms has the most predictive power for estimating house prices, followed by bathrooms and parking spaces, although to a lesser extent.

The analysis of the average price by location and title provides valuable insights into the pricing trends of properties in different regions and types.

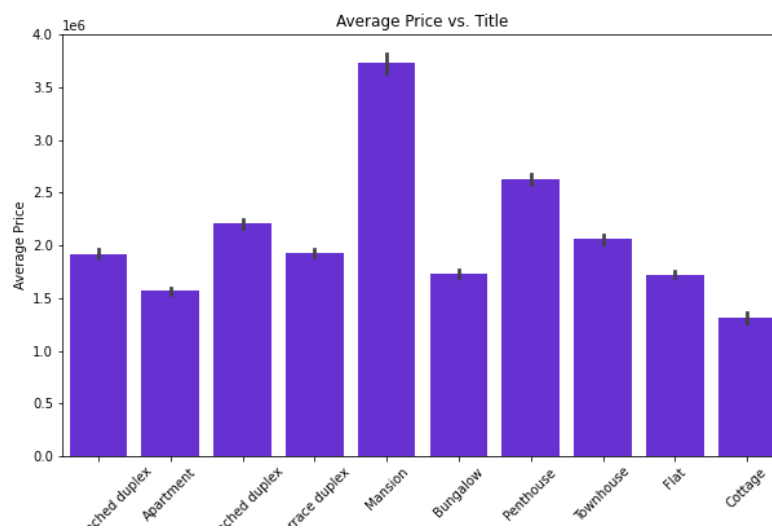
Location Analysis:

- The average prices of properties vary significantly across different states in Nigeria. Lagos stands out as the region with the highest average property price at approximately 4.21 million Naira, indicating a strong demand for real estate in the commercial hub.
- On the other hand, Ebonyi, Kebbi, and Zamfara have relatively lower average property prices, suggesting a more affordable real estate market in those regions.



Title Analysis:

- The type of property also plays a crucial role in determining its average price. Mansion properties command the highest average price of around 3.73 million Naira, making them the most expensive category.
- Cottage properties have the lowest average price at approximately 1.31 million Naira, indicating a more budget-friendly option for buyers.



Recommendations from Analysis:

- The real estate company, Wazobia Real Estate Limited, should focus on leveraging the high demand in Lagos to offer premium properties and maximize profits.
- The company could consider expanding its offerings in regions with a relatively lower average price, like Ebonyi, Kebbi, and Zamfara, to target budget-conscious customers.
- Since location and title both significantly impact property prices, it is essential to incorporate these features into the predictive model for accurate price estimation.

Conclusion from Analysis:

In this exploratory data analysis (EDA) of the real estate dataset provided by Wazobia Real Estate Limited, we gained valuable insights into the distribution and relationship of various features with the target variable 'price'. The dataset contained 14000 rows of data with several missing values in 'loc' and 'title' columns. We carefully analyzed the data and made decisions on how to handle the missing values.

1. Handling Missing Values:

As the 'bedroom', 'bathroom', and 'parking_space' features exhibited skewed distributions, we opted to fill their missing values with the median. Using the median is suitable in this case, as it is less sensitive to extreme values compared to the mean. For these features, advanced imputation methods could be affected by the skewed distribution, leading to potential inaccuracies in the imputed values.

However, for 'loc' and 'title', which are categorical features, it is challenging to impute missing values accurately due to the lack of relevant information. Since there are only 1813 and 1722 missing values in 'loc' and 'title', respectively, and considering the size of the dataset, it is recommended to drop rows with missing values in these columns. Imputing these categorical features may introduce bias and potentially distort the analysis or predictive modeling.

2. Correlation and Multicollinearity:

The correlation matrix showed that 'bedroom', 'bathroom', and 'parking_space' have moderate positive correlations with 'price'. These features can be considered important predictors of house prices, and their relationships with 'price' can be explored further in regression modeling.

3. Location and Property Type:

The bar plots revealed the relationship between 'loc' and 'price', as well as 'title' and 'price'. They showed the average price for different locations and property types. The 'loc' and 'title' features are crucial in predicting house prices, as they significantly impact the average price.

EDA has provided us with valuable insights into the dataset, guiding us on the appropriate handling of missing values and identifying key predictors of house prices. It is essential to use robust and suitable imputation methods based on the distribution of the features. Additionally, the 'loc' and 'title' columns

should be handled with care, and rows with missing values in these columns should be dropped to preserve data integrity and avoid biased analyses. The results obtained from this EDA will serve as a foundation for building predictive models that accurately estimate house prices for Wazobia Real Estate Limited.

Data Preprocessing

Data Cleaning

- **Handling Missing Values in 'loc' and 'title' Columns**

Process: We removed rows with missing values in the 'loc' and 'title' columns.

Reason: The number of missing values in the 'loc' and 'title' columns was relatively small compared to the total dataset. As a result, dropping these rows was a prudent decision to avoid introducing potential biases and uncertainties that could arise from imputing categorical data. By eliminating these incomplete records, we ensured the integrity and accuracy of the data used for analysis and modeling. This approach aligns with the goal of providing Wazobia Real Estate Limited with reliable and precise pricing information, enabling them to make informed decisions with confidence in the highly competitive real estate market.

- **Imputing Missing Values in 'bedroom', 'bathroom', and 'parking_space' Columns**

Process: We filled the missing values in the 'bedroom', 'bathroom', and 'parking_space' columns with their respective medians.

Reason: The 'bedroom', 'bathroom', and 'parking_space' columns are numeric features that have a right-skewed distribution, as evidenced by their mean being greater than their median. Filling the missing values with the median is a suitable approach in this context, as it is robust to outliers and ensures that the imputed values are representative of the central tendency of the data. This method provides a reasonable estimate for the missing values, minimizing potential distortions in the analysis and modeling process. By preserving the distribution characteristics of these features, we maintain the overall quality and reliability of the dataset, allowing for accurate predictions of house prices for Wazobia Real Estate Limited.

Features Engineering

- Creating Binary Indicator Variables for 'Lagos' and 'Mansion' Categories

Process: We added two binary indicator variables, 'is_lagos' and 'is_mansion', to the DataFrame. The 'is_lagos' variable takes the value of 1 if the location is 'Lagos', indicating that the property is in Lagos,

Nigeria; otherwise, it takes the value of 0. Similarly, the 'is_mansion' variable takes the value of 1 if the title of the property is 'Mansion', signifying that the property is classified as a mansion; otherwise, it takes the value of 0.

Reason: The decision to create binary indicators for 'Lagos' and 'Mansion' stems from our Exploratory Data Analysis (EDA), where we observed unusual patterns in both classes. Lagos properties displayed significantly higher prices compared to other locations, suggesting a unique price dynamic specific to the Lagos market. On the other hand, 'Mansion' properties exhibited distinct characteristics and significantly higher prices relative to other property types, indicating a premium segment in the housing market.

By isolating these classes using binary indicators, we can account for the specific effects of Lagos properties and mansions on house prices independently. This approach enhances the predictive modeling process, providing valuable insights to Wazobia Real Estate Limited for pricing decisions and market positioning.

- Deriving New Features for Comfort and Size

Process: We created three new features, 'comfort_ind', 'size', and 'comfort_by_size', to capture additional insights from the existing data. 'comfort_ind' represents the comfort index, calculated as the ratio of the number of bedrooms to the number of bathrooms. 'size' represents the total size of the property, calculated as the sum of the number of bedrooms, bathrooms, and parking spaces. Lastly, 'comfort_by_size' is an interaction feature obtained by multiplying the 'comfort_ind' with the 'size'.

Reason: These new features aim to provide a deeper understanding of the relationship between the number of bedrooms, bathrooms, and parking spaces in the context of property comfort and size. The 'comfort_ind' helps assess how the ratio of bedrooms to bathrooms influences the comfort level of a property, as a higher ratio may indicate more private living spaces for occupants. The 'size' feature provides a comprehensive measure of the property's overall space, accounting for the essential rooms like bedrooms, bathrooms, and parking spaces.

The 'comfort_by_size' interaction feature combines the information from both 'comfort_ind' and 'size', enabling us to identify properties that offer a high level of comfort while considering their overall size. By including these features in our predictive model, we can better capture the complexities of how property attributes impact prices, ultimately aiding Wazobia Real Estate Limited in more accurate pricing and decision-making.

- Categorizing Population Density Levels

Process: We created a new feature, 'population_density_level', to categorize the locations based on their population density. We classified the locations into ten population density levels, ranging from Level 1 to Level 10, with Level 1 representing the highest population density areas.

Reason: Categorizing the locations based on population density can provide valuable insights into how population density influences property prices. Areas with higher population densities tend to have higher

demand for housing, which can impact property prices significantly. By assigning each location to a specific population density level, our predictive model can better capture the variations in property prices due to differences in population density. This information will be instrumental in helping Wazobia Real Estate Limited accurately price properties and optimize their business strategy to target specific market segments.

- Encoding Location Information

Process: The 'loc' column in the combined_df DataFrame was encoded using TargetEncoder, where locations were ranked based on their corresponding house prices, and then the encoded values were smoothed by truncation.

Reason: By encoding location information, we transform categorical data into numerical values, making it easier for our predictive model to understand the relationship between locations and property prices. Smoothing the encoded 'loc' values helps prevent overfitting, a common problem where the model becomes too complex and performs poorly on new data. The process simplifies the data and makes it easier for the model to generalize, ensuring more accurate and reliable predictions. This way, Wazobia Real Estate Limited can make informed pricing decisions based on property locations, enhancing its competitiveness in the market.

Machine Learning Workload

Model Building

In the model building process, we utilized the CatBoostRegressor, a powerful gradient boosting algorithm, to predict house prices in the real estate market. The model was configured with the RMSE (Root Mean Squared Error) as the loss function, as our primary objective was to minimize prediction errors. To avoid overfitting, we applied 5-fold cross-validation with shuffling, using the KFold method. The evaluation metric used for assessing the model's performance was the RMSE, which measures the average deviation of predicted prices from the actual prices in the dataset.

Model Evaluation

An RMSE (Root Mean Squared Error) of 416,786 means that, on average, the model's predictions are off by approximately 416,786 units in terms of the target variable, which in this case is the house price. Since RMSE is a measure of the model's prediction accuracy, a lower RMSE indicates better performance and higher precision in predicting house prices. Therefore, the RMSE value of 416,786 suggests that our model's predictions have a considerable variance from the actual house prices in the dataset.

While the RMSE is a useful metric for assessing prediction accuracy, it's essential to compare it to the range of target variable values and understand the domain-specific implications. If the RMSE is

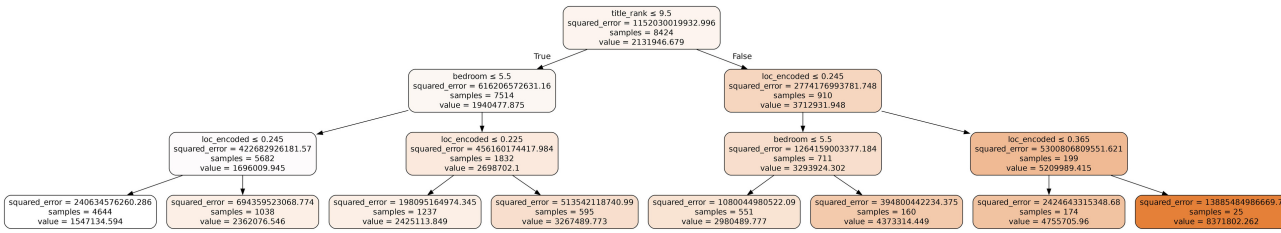
significantly smaller compared to the price range, it indicates that the model's predictions are relatively accurate. However, if the RMSE is close to or larger than the price range, it suggests that the model's predictions may not be reliable, and there is room for improvement.

The cross-validation RMSE score was calculated to evaluate the model's generalization ability. With a value of 416,786, this score indicates the average root mean squared error across the 5 folds, showing how well the model performed on unseen data during cross-validation. However, it's important to note that the performance on the leaderboard (LB) might differ from the cross-validation score due to variations in the test dataset, which can affect model predictions. The LB score, which is 310,010, is the model's performance on the competition's test dataset and provides a more realistic estimate of how the model will perform on new, unseen data.

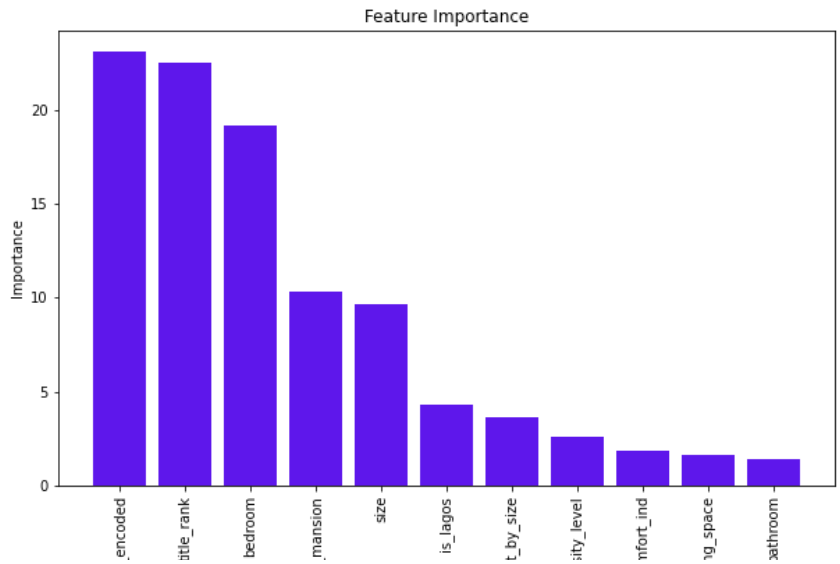
Overall, the RMSE is a crucial evaluation metric for our machine learning project. By minimizing the RMSE, we aim to build a model that can accurately predict house prices, thus providing valuable insights to Wazobia Real Estate Limited. The model's ability to generalize to unseen data is assessed through cross-validation, ensuring that it performs well on different subsets of the training data. The LB score, on the other hand, helps us gauge the model's true performance on unseen data, which is essential for real-world application. With the model built and evaluated, we can confidently provide Wazobia Real Estate Limited with a robust predictive tool to aid their pricing decisions and enhance their competitiveness in the Nigerian real estate market.

Machine Learning Explainability

Decision trees serve as a fundamental building block in the realm of explainable machine learning (ML) due to their inherent transparency and interpretability. Unlike complex black-box models, such as deep neural networks, decision trees offer a clear and human-understandable representation of the decision-making process. This characteristic makes decision trees an attractive choice when explainability and interpretability are crucial aspects of the machine learning task.



Features Importance



Based on the analysis, the top three most important features for predicting house prices are "loc_encoded," "title_rank," and "bedroom," with importance scores of 23.06%, 22.46%, and 19.12% respectively. These features have the highest impact on the model's performance.

Other significant features include "is_mansion" (10.31%), "size" (9.63%), "is_lagos" (4.28%), and "comfort_by_size" (3.64%). These features also contribute significantly to the model's predictive ability.

On the other hand, "comfort_ind," "parking_space," and "bathroom" have relatively lower importance scores, indicating that they have a lesser impact on the model's predictions compared to other features.

Conclusion

In conclusion, the data exploration and analysis have provided valuable insights into the factors influencing house prices in Nigeria. The predictive model, based on the Decision Tree and CatBoost Regressor, shows promise in accurately estimating property prices, with key features like location, property type, number of bedrooms, and property size playing significant roles. The model's interpretability is high, making it easier for stakeholders to understand the factors driving its predictions.

To improve its business, Wazobia Real Estate Limited can leverage the predictive model to make informed pricing decisions. By using the model to estimate property prices accurately, the company can competitively price its houses and optimize revenue generation. The model can also aid in identifying potentially undervalued or overvalued properties, helping the company make strategic investment decisions.

To maintain the model's performance, Wazobia should regularly update it with fresh data. As the real estate market dynamics change, the model's predictions may need adjustment. Regularly retraining the model with up-to-date data will ensure its continued accuracy and relevance.

From the analysis, it is evident that location and property type are crucial factors affecting prices. Therefore, Wazobia can focus on acquiring properties in high-demand locations and popular property types to optimize profits. Moreover, the company can explore adding more relevant features, such as amenities, proximity to essential services, and property condition, to further enhance the model's predictive capabilities.

Additionally, Wazobia should monitor potential outliers and further investigate instances of missing data to ensure data quality and avoid biased predictions. By conducting periodic audits and data quality checks, the company can maintain the model's robustness and reliability.

Implementing the predictive model and leveraging data-driven insights will empower Wazobia Real Estate Limited to make better pricing decisions, stay competitive in the market, and deliver enhanced value to its customers. With a robust and well-maintained predictive model, the company can achieve improved profitability and strengthen its position as a leading real estate player in Nigeria.