## Predictive Analytics for Real Estate: Price Prediction of Houses using Machine Learning

**PROJECT MANAGEMENT**

Ø **Choosing role**: - Project manager has assigned different roles to all the team members that are aligned with their interests and qualifications. These roles include researcher, Solution and design, Development, Testing- QA, Documentation, and Implementation/Deployment.

|  |  |  |
| --- | --- | --- |
| **Name** | **Student ID** | **Roles & Responsibilities** |
| Tajdar Unnisa Begum | 500201392 | Project Manager, Development, Documentation |
| Nanda Kishore Karicherla | 500197946 | Solution and design, Implementation/Deployment |
| Husanpreet Kaur | 500195671 | Researcher, Testing & QA |

Ø **Timeline**: For this week data set was provided and with the help of python, we were able to clean the data and filter out all the necessary data required to improve the quality of data. We researched on real estate industry and found that this industry is dependent on multiple factors such as prices, location, size, and locality.

Ø **Platform :** The project makes use of the Python programming language and few data analysis tools, including scikit-learn, pandas, and NumPy. Jupyter Notebook and Google collab is used to set up the development environment, and Git is used to maintain version control. We have used Microsoft Teams and outlook for sharing our researches, outcome, resources, files worked on and discussion.

Ø **Action plan:** The team worked together to specify the project's objectives, goals, and deliverables. The processes necessary to clean and transform the raw data, generate the prediction model, and assess its effectiveness were outlined in a plan of action.

Ø **MEETING DOCUMENTATION AND PROOF OF DISCUSSION**: Most of the discussions were in the class and shared documents on teams and outlook. Below is the screenshot attached.

**SOLUTION FLOW**

Ø **Requirement Finalization:** The team conducted thorough discussions to understand their specific requirements and expectations for the predictive model of Natty City data set.

Ø **Approach:** The team decided to follow a supervised machine learning approach for now, using a regression algorithm to predict house prices based on the provided features. The data will be divided into training and testing sets to evaluate the model's performance.

Ø **Data Cleaning and Transformation:** This week our team has performed data cleaning tasks, handling missing values, outliers, and inconsistent data. Feature engineering techniques will be employed to extract relevant information from the raw data and create meaningful features for the model.

Ø **Model Development:** we will work together to develop and fine-tune the predictive model using appropriate regression algorithms. The model will be trained on the training data and validate using the testing data.

Ø **Evaluation:** The model's performance will be evaluated using various metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared. The team will iteratively refine the model based on the evaluation results.

**PROJECT RESOURCES AND METHODOLOGY**

Ø **Resource Selection:** The team selected Python as the best programming language for data analysis and machine learning activities because of its large library and ecosystem support. The chosen libraries, including pandas and scikit-learn, offered effective tools for model creation and data manipulation.

Ø **Resource Quality Considerations:** The team will be using established documentation, user reviews to assure the credibility and quality of the resources it chose. To maintain the readability and quality of the code, best practices and coding standards will be adhered to.

Ø **Methods of Implementation:** The project will be using a collaborative and iterative process. To encourage a productive workplace, routine code reviews, pair programming, and knowledge exchange sessions will be held. Git version control will make it possible to integrate code quickly and effectively.

Ø **Analytics Techniques:** To obtain insights into the data, spot trends, and spot abnormalities, exploratory data analysis (EDA) methods will be used. The prediction model's significant characteristics will be derived using feature extraction and selection techniques.

**FINDING AND ANALYSIS**

Several features are frequently considered while developing a machine learning model for predicting home prices. These characteristics aid in identifying the crucial elements that affect a home's value. Here are some essential characteristics that are frequently included in models for predicting home price changes:

**DRAWBACKS OR FAILINGS:**

Ø One of the major challenges faced by the team was the presence of missing data in the raw dataset. Imputation techniques were applied to handle missing values, but it may have introduced some level of bias in the model.

**RESOURCES USED:**

[**https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e**](https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e)

[**https://www.kaggle.com/code/yasserh/housing-price-prediction-best-ml-algorithms**](https://www.kaggle.com/code/yasserh/housing-price-prediction-best-ml-algorithms)

Ø **Hours of Labor:** The team utilized AIP class hours, spread across researching on real estate, Exploring data and data cleaning.

Ø **Costs:** The project utilized existing resources on google, and no additional costs were incurred.

In conclusion, this project aims to develop a predictive model for Natty City to estimate house prices based on various features.

Introduction:

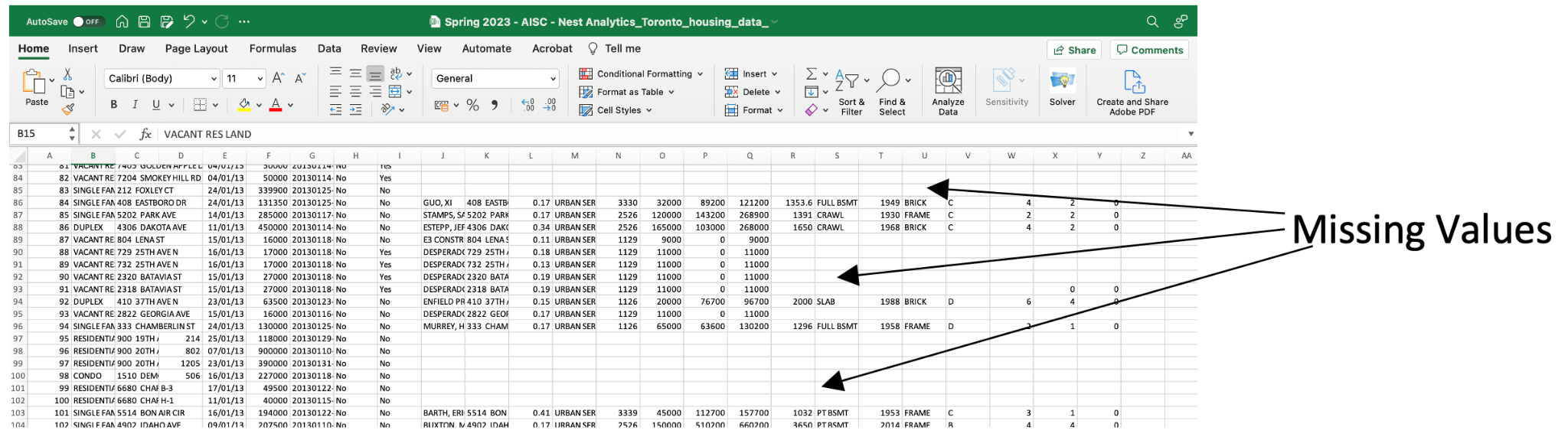
Price prediction is an essential task in the real estate industry, letting companies make good property pricing decisions. Various techniques and strategies must be employed to predict prices accurately. In this report, we will discuss the techniques required for price prediction in the real estate industry. We will cover handle missing values, outliers, categorical data, date data, address data, name data, feature selection techniques, feature ranking techniques, feature binning techniques, feature generation and combination techniques, kernel techniques, time series techniques such as ARMA, ARIMA, VARIMA, Holt-Winters and other sequence models, evaluation techniques for time series and regression models.

**Introduction:**

Price prediction is an essential task in the real estate industry, letting companies make good property pricing decisions. Various techniques and strategies must be employed to predict prices accurately.

In this report, we will discuss the techniques required for price prediction in the real estate industry.

1. Missing Data: Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset. Below is a sample of the missing data from the Housing data.



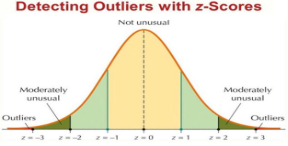
a. Handling Missing Values: Missing values are a common challenge in real estate data. Techniques to handle missing values include:

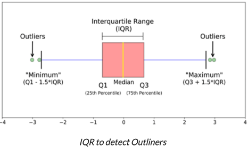
· Deleting: If missing values are few and randomly distributed, deleting the corresponding rows or columns can be an option.

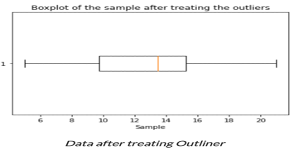
· Imputation: Techniques such as mean, median, mode imputation, or more advanced methods like regression imputation or multiple imputation can be used to estimate missing values.

· Advanced techniques: Use more sophisticated methods like k-nearest neighbors (KNN) imputation or matrix factorization.

2. Outlier: Outlier is an observation in a given dataset that lies far from the rest of the observations. That means an outlier is vastly larger or smaller than the remaining values in the set.







a. Handling Outliers:

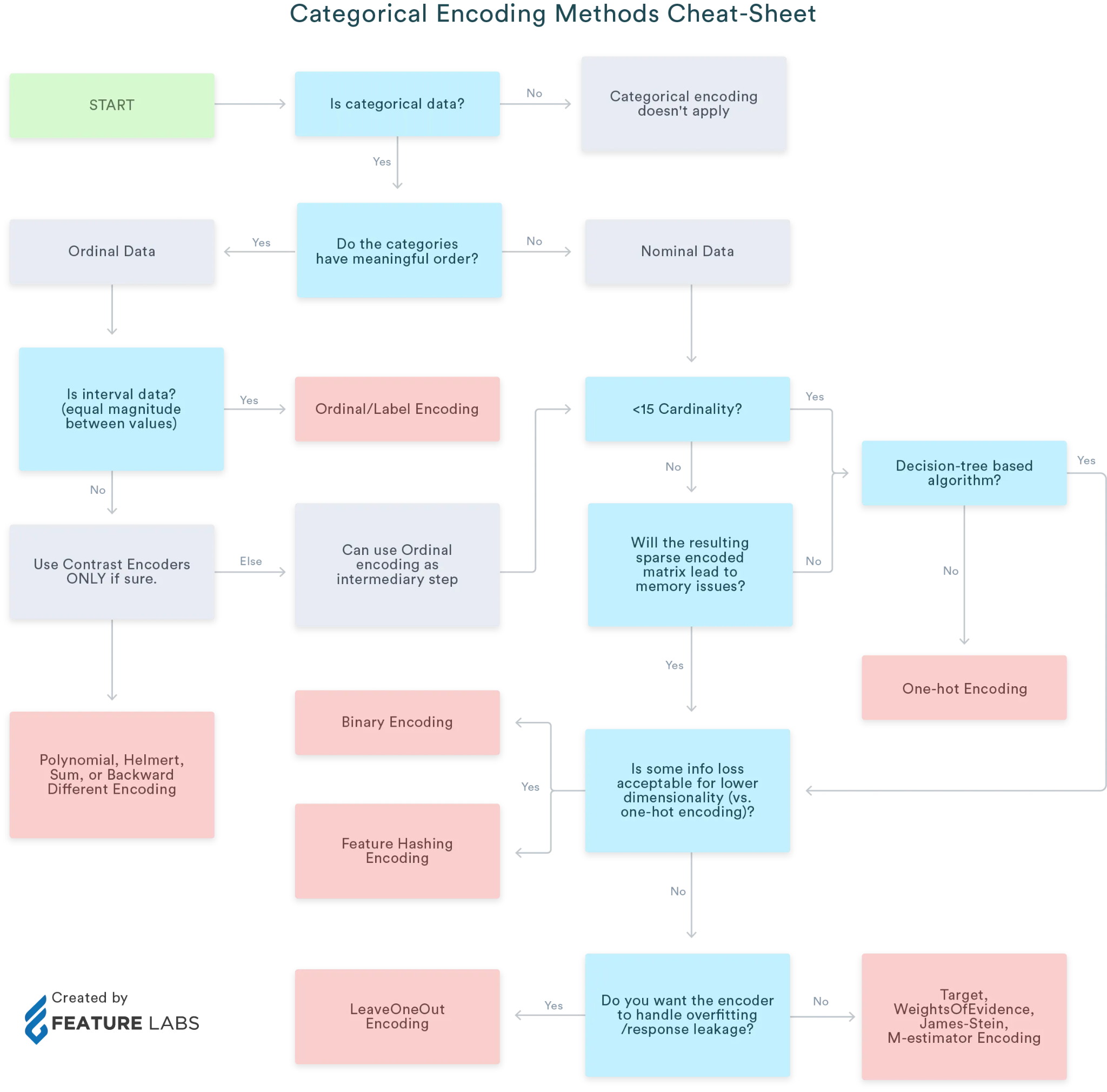
o Statistical methods: Use statistical techniques like z-score, interquartile range (IQR), or boxplots to identify and handle outliers.

o Winsorization: Replace extreme values with a predefined percentile value.

o Transformation: Apply data transformations like log transformation to reduce the impact of outliers.

3. Categorical Data: Categorical data is non-numerical information that is divided into groups. As its name suggests, categorical data describes categories or groups.

a. Handling Categorical Data:



4. Date, Address, and Name Data: Date, address, and name data require special treatment for price prediction

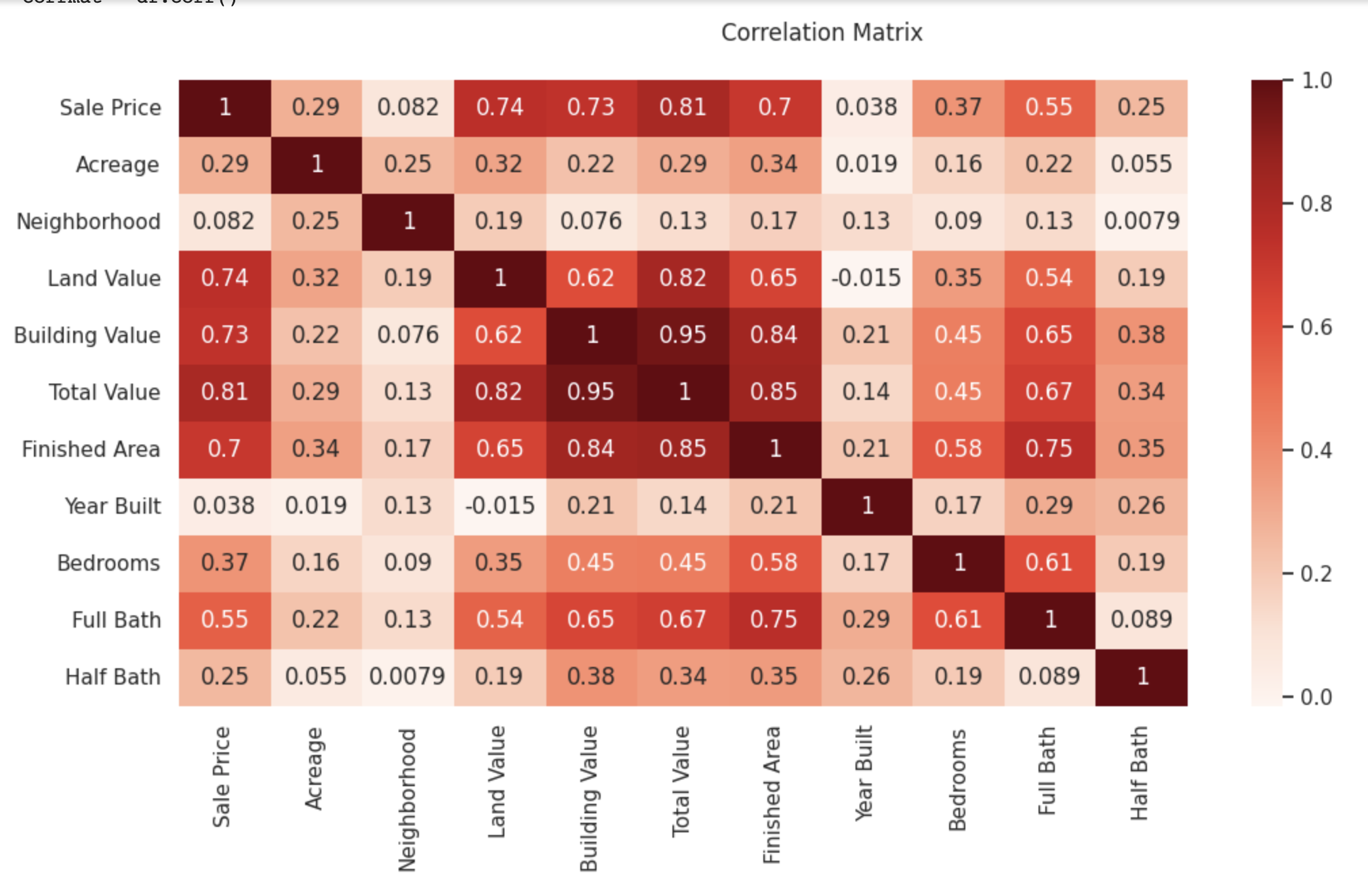
• Date Data: Extracting features like a day of the week, month, or year and incorporating time-related patterns.

• Address Data: Extracting location features like zip code or neighbourhood and incorporating spatial patterns.

• Name Data: Extracting relevant information from names, such as property owner demographics or property characteristics.

5. Feature Selection and Ranking Techniques:

• Correlation analysis: Identifying features with high correlation to the target variable.

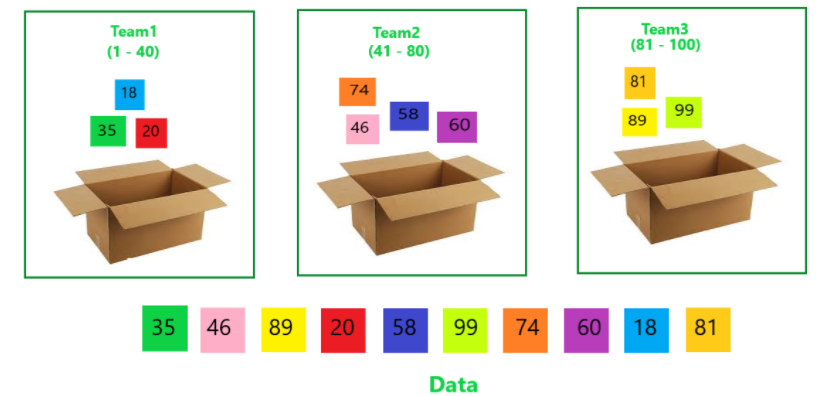


• Recursive Feature Elimination: Iteratively removing features with the most minor importance.

• L1 Regularization: Using regularization techniques like Lasso regression to shrink coefficients and perform feature selection.

6. Feature Binning, Generation, and Combination:

• Binning:



• Feature Generation: Creating new features by transforming existing variables or combining them through mathematical operations.

• Feature Combination: Creating interaction terms or cross-product features to capture synergistic effects.

7. Kernel Techniques: Kernel methods, such as Support Vector Regression (SVR), can capture complex non-linear relationships between features and target variables.

8. Time Series Techniques:

• Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), Vector Autoregressive Moving Average (VARIMA): Models to capture temporal dependencies and seasonality.

• Holt-Winters: Seasonal time series model for forecasting.

• Sequence Models: Advanced techniques like Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to capture sequential patterns.

9. Evaluation Techniques for Time Series and Regression:

• Mean Squared Error (MSE) or Root Mean Squared Error (RMSE): Measures the difference between predicted and actual values.

• Mean Absolute Percentage Error (MAPE): Measures the relative difference between predicted and actual values.

• R-squared (R2): Measures the proportion of variance in the target variable explained by the model.

**Conclusion:** Accurate price prediction in the real estate industry requires applying various techniques. This report covered techniques to handle missing values, outliers, categorical data, date data, address data, and name data, as well as feature selection, feature binning, feature generation and combination techniques, kernel techniques, time series models, and evaluation techniques for time series and regression. Real estate professionals can make more informed decisions and enhance their pricing strategies by employing these techniques.

**Project Plan:** [Trello Board](https://trello.com/b/BTo9AoVE/aip-13-project-plan)

**Code:** [Collab on git hub](https://github.com/nanda1296/Predictive-Analytics-for-Real-Estate/blob/main/Project_Group_A13.ipynb)

**References:** [A](https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/) , [B](https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02) , [C](https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/) , [D](https://www.analyticsvidhya.com/blog/2020/10/getting-started-with-feature-engineering/) , [E](https://www.tableau.com/learn/articles/time-series-analysis) , [F](https://towardsdatascience.com/what-is-an-arima-model-9e200f06f9eb) , [G](https://towardsdatascience.com/what-is-an-arima-model-9e200f06f9eb) , [H](https://www.kaggle.com/code/yasserh/housing-price-prediction-best-ml-algorithms) , [I](https://www.kaggle.com/code/chanakyavivekkapoor/house-price-prediction) , [J](https://www.geeksforgeeks.org/house-price-prediction-using-machine-learning-in-python/) , [K](https://stackoverflow.com/questions/58756515/onehotencoder-object-has-no-attribute-get-feature-names) , [L](https://towardsdatascience.com/support-vector-regression-svr-one-of-the-most-flexible-yet-robust-prediction-algorithms-4d25fbdaca60) , [M](https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e) , [N](https://www.kaggle.com/code/gaurav896/nashville-housing-initial-analysis) , [O](https://www.kaggle.com/code/gauthampughazh/house-sales-price-prediction-svr)

• <https://www.kaggle.com/code/yasserh/housing-price-prediction-best-ml-algorithms>

• <https://www.kaggle.com/code/chanakyavivekkapoor/house-price-prediction>

<https://www.geeksforgeeks.org/house-price-prediction-using-machine-learning-in-python/>

<https://stackoverflow.com/questions/58756515/onehotencoder-object-has-no-attribute-get-feature-names>

<https://towardsdatascience.com/support-vector-regression-svr-one-of-the-most-flexible-yet-robust-prediction-algorithms-4d25fbdaca60>

<https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>

• <https://www.kaggle.com/code/gaurav896/nashville-housing-initial-analysis>

• <https://www.kaggle.com/code/gauthampughazh/house-sales-price-prediction-svr>