**Project Report: Predicting Housing Prices**

**Introduction**

In this project, we aim to build predictive models for housing prices using a dataset obtained from Kaggle. The dataset contains various features that describe properties, including numerical and categorical data. The project encompasses data exploration, preprocessing, model building, and evaluation. The goal is to create models that can accurately predict housing prices based on the provided features.

**Data Exploration**

* **Importing and Previewing Data**

We began by importing the dataset and examining its structure. The dataset consists of various columns, and it's essential to understand its dimensions and column names to proceed effectively.

* **Numerical and Categorical Features**

We separated the dataset into numerical and categorical features to gain insights into their summary statistics. This step involved generating summary statistics for numerical features and categorical feature statistics, helping us understand the data's distribution and characteristics.

* **Missing Value Analysis**

Detecting missing values is crucial for data quality. We identified missing values in both numerical and categorical features, highlighting columns with null values and their respective percentages. This information helps us decide how to handle missing data during preprocessing.

* **Exploration of the Dependent Variable**

We analyzed the dependent variable, 'SalePrice,' to understand its distribution. We explored the following aspects:

* + Distribution: We examined whether the 'SalePrice' variable follows a normal distribution and if normalization is necessary.
  + Quantile-Quantile (Q-Q) Plot: We generated a Q-Q plot to assess how well 'SalePrice' conforms to a normal distribution.

**Key Questions and Analysis**

We posed several questions to extract insights from the dataset:

* **Distribution of Dwelling Types and Their Relation to Sale Prices**

We analyzed the distribution of building types and their impact on sale prices. The data showed that '1Fam' was the most common building type, and we found a clear association between building type distribution and average sale prices.

* **Does Zoning Impact Sale Prices**

We investigated whether zoning types have an influence on sale prices. The results indicated that certain zoning types, such as 'FV' and 'RL,' were associated with higher average sale prices.

* **Does Street and Alley Access Types Affect Sale Price**

We explored the impact of street and alley access types on sale prices. The analysis suggested that properties with 'Pave' access had higher sale prices.

* **Average Sale Price by Property Shape and Contour**

We examined the relationship between property shape, contour, and sale prices. The data revealed differences in sale prices based on property shape and contour.

* **Correlation between Property Age and Sale Price**

We calculated the correlation between property age and sale prices. The result indicated the strength of the relationship between these two variables.

* **Price Changes Over the Years**

We analyzed whether sale prices changed over the years, presenting the trends in sale prices over time.

**Data Preprocessing and Pipeline**

To prepare the data for modeling, we implemented data preprocessing steps and created a data pipeline. The pipeline included:

* Handling missing values
* Scaling numerical features
* Encoding categorical features
* Principal Component Analysis (PCA) for dimensionality reduction

**Model Building and Parameter Tuning**

We selected three regression models: Linear Regression, Random Forest, and XGBoost, to predict housing prices. For each model, we:

* Split the data into training and testing sets
* Defined hyperparameter grids for tuning
* Utilized GridSearchCV to find the best hyperparameters
* Conducted 3-fold cross-validation

The models were trained and fine-tuned, with the best hyperparameters and Root Mean Squared Error (RMSE) recorded for each model.

**Principal Component Analysis (PCA)**

We applied PCA to reduce dimensionality and retrained the models with the reduced dataset. PCA helped mitigate multicollinearity issues and improved the performance of the Linear Regression model.

**Model Evaluation**

We evaluated the models using the test dataset and calculated the RMSE for each model. The results indicated how well the models generalized to new data, where three models has almost the same result with XGBoost performing slightly better.