# Project Documentation: HDB Flat Prices Regression Analysis

## Overview

This project focuses on analyzing and predicting resale prices of Housing Development Board (HDB) flats in Singapore from 1990 to March 2021. The dataset used for analysis is sourced from Kaggle, containing over 800,000 rows of transaction data. The goal is to build regression models that can accurately predict resale prices based on various features.

## Data Sources

1. **ALL Prices 1990-2021 Mar.csv**: The primary dataset containing transaction details of HDB resale flats. It includes calculated and mapped columns, such as CPI index and lease percentage.
2. **Balas Table.csv**: Provides ratios of leasehold land value to freehold land value for each year of remaining lease, used in land valuations.
3. **MAS Core Inflation.csv**: Contains Singapore's core Consumer Price Index (CPI) values from January 1990 to February 2021.
4. **complete.csv**: Includes unique block addresses with geocoded full addresses and latitude-longitude coordinates.
5. **gni per capita.csv**: Provides Singapore's Gross National Income (GNI) per capita in nominal S$ for the years 1990-2020.
6. **HDB machine learning.xlsx**: An analysis file presenting a basic exploration of numerical variables influencing resale prices, using linear regression algorithms.

## Data Exploration

### 1. Importing Libraries

The project begins by importing necessary libraries such as pandas, numpy, and seaborn for data manipulation and visualization.

### 2. Reading Data

The main dataset, 'ALL Prices 1990-2021 mar.csv', is read into a Pandas DataFrame named 'data'.

### 3. Data Observation

Various visualizations are generated to observe trends in the data, including box plots, kernel density plots, and geographical distributions using latitude and longitude.

# Data Analysis

### 4. Data Analysis

Several questions are addressed, including the presence of null values, data types, mean and maximum values, and identification of unuseful features.

### 5. Data Preparation for Processing

Data encoding and scaling are performed to prepare the dataset for model building. Label encoding is applied to categorical variables, and scaling is carried out using MaxAbsScaler and StandardScaler.

# Model Building

### 6. Model Building

Linear regression models are constructed to predict resale prices. The following models are explored:

1. Simple Linear Model
2. Polynomial Regression
3. Ridge Regression
4. Lasso Regression
5. Elastic-Net Regression

Each model is evaluated and visualized, with an emphasis on exploring polynomial features and their effects on model performance.

# Model Comparison

### 7. Comparing Models

Various metrics, including Max Error, Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared, are utilized to compare the performance of different regression models. The results are presented for models with varying degrees of polynomial features.

# Conclusion

This documentation provides a comprehensive overview of the regression analysis conducted on HDB flat prices. The project involves data exploration, analysis, and model building to predict resale prices accurately. The comparison of different regression models allows for insights into their strengths and

weaknesses, aiding in the selection of the most suitable model for predicting HDB resale prices. The code and analysis serve as a valuable resource for anyone interested in understanding and predicting housing prices in Singapore.