CANADIAN HOUSE PRICE PREDICTION

Github link: https://github.com/vrushitamalani/machinelearningfinalproject

INTRODUCTION

Nowadays the prices of Canadian is booming for serval reason and plenty of factor affecting this such as market value of land because of the development and the power of purchasing the house among the people eventually the demand of the houses skyrocketing in Canada so I have take dataset of the province of Canada in order to demonstrate how many factor play crucial role in the price.

This project aims to build a machine learning model that estimated that house price base on key real estate attribute .the goal is provide data driven approaches to support home buyers and seller and investors to make flawless decision.

To utilize a dataset that includes variables for example house size ,number of bedrooms and bathrooms , neighborhood income levels and distance to the city center we explore the relationship between house price . the findings from the study can supported potential in order to making flawless decisions

This project aims to make machine learning model that estimate house price based on historical and simulated real estate data from Canada. The rectification include the full pipeline from data cleaning to model calculation

**Dataset Overview: Canadian House Prices**

This dataset is simulated version of house data from Canada . it is designed to resemble real world housing market data and include several key feature that affect home prices

**Explanation of Columns (Typical)**

|  |  |  |
| --- | --- | --- |
| Colum name | Data type | Desciption |
| **City** | Object | Name of the city where the property is located. |
| |  | | --- | | **Province** |  |  | | --- | |  | | |  | | --- | | Object | | Canadian province code (e.g., ON, MB). |
| **Size\_sqft** | Integer | Total area of the house in square feet. |
| **Bedrooms** | Integer | Number of bedrooms in the house. |
| **Bathrooms** | Integer | Number of bathrooms in the house. |
| **Year\_Built** | Integer | Year in which the house was built. |
| **Amenities\_Score** | Float | Composite score 1-10 present the quality and quantity of nearby amenities |
| **Crime\_Rate** | Float | Crime rate in the area, indicating safety (higher = worse). |
| **Median\_Income** | Integer | Median income of residents in the neighborhood (in CAD). |
| **Distance\_to\_City\_Center\_km** | Float | Distance from the house to the city center (in kilometers). |

Data Exploration Discussion

Prior to training any model it is vital to understand the structure , quality and patterns within dataset .

**1.DATASET OVERVIEW**

There are 11 dataset in this project which is show that residential properties in Canadian province and house\_price describe the actual market value

**2.DATA TYPE**

* Int and float : those are only for the numeric variables
* Object : for the categorical variable like city and province
* The data types are appropriate for regression analysis after encoding categorical variables.

**3. STATISTICAL SUMMARY**

Using df.descibe() , we will get,

* Size\_sqft : range from 600 to 4000 sqft
* Mediam\_income : spans from 40000 to 150000 CAD
* House\_price : has values ranging from 30000 to approx. 1500000 CAD
* Crime\_rate and amenities\_score are on a moralization scale from 1 to 10

4. **DATA DISTRIBUTION**

* House\_Price is **right-skewed**, as predict in real estate where expensive homes pull the average up.
* Size\_sqft, Median\_Income, and Distance\_to\_City\_Center\_km also show spread-out distributions.

**PREPROCESSING AND FEATURE ENGINEERING**

Preprocessing is vital step in machine learning that confirm that ensure the data is in a suitable format for modeling feature engineering further increscent the model by creating or transfer feature that make patterns in the data more apparent

1. **Data Cleaning and Handling Missing Values**

As observed in the data exploration phase there are no missing values in the dataset so there are no condition to removal of rows in the dataset

If the missing values existed, common method like mean number columns or categorial columns could be used

1. **FEATURE ENCODING**

As we have categorial columns like city and province these need to be encoded into numerical values so they can be used by the machine learning model

* **OneHotEncoding** : this method convert categorical values into the format that can be easily used by the machine learning algorithm it create a binary column for each category in the feature

City: Toronto, Vancouver, Montreal

City\_Toronto, City \_Vancouver, City\_Montreal with values 0 or 1.

1. **FEATURE SCALING**

Many machine learning models such as linear regression,KNN and SVM need feature scaling since they rely on the distance metrics and can perform poorly if feature are not scaled

In our dataset size\_sqft,median\_income,crime\_rate and distance to city center km are continuous numeric variable and scaling them is vital

StandardScaler is apply to ensure all feature have zero mean and unit variance . this is particulate important when feature have different units

1. **FEATURE ENGINEERING**

Feature engineering involves creating new feature or transforming existing feature to improve model performance   
  
Example:

**Price per Square Foot**

This is a useful feature in real estate as it control price according to house size.

**Formula:**

Price\_per\_sqft = house\_price/size\_sqft

This can be provided value in the property value per unit area and make the model more robust

df[‘price\_per\_sqft’] =df[‘house price’]/df[‘size\_sqft’]

1. **Data Transformation**

in some cases transformation like log transformation or power transformation can be used to make data distribution more normal

**Example**

if your target variable house\_price is heavily applied foe twice , applying log transformation might help

df ['house\_price ] = np.log(df[ house\_price ])

Model Training & Selection

Model training is the process of using historical data to teach a machine lerning algorithm how to make prediction in this project we are give the focus on estimating house prices based on the multiple feature such as houce size, number of bedrooms , neighborhood income,crime rate and more

1. **MODEL SELECTION**

Given that our problem is regression task (we will take the above variable house\_price) we will evalute multiple model to see which performs best

**Linear regression :** a simple model that assumes a liner relationship between the feature and the target variable this is a good baseline model for comparison

**Random forest regressor :** a more strong non liner model that uses multiple decision trees to predict the target it is great for capturing complex relationships in the data and less outperforms simpler model likewise linear regression

**2.TRAINING THE MODELS**

* Linear Regression :

This model will be train on all the feature to get ideia the linear realationship between them and target variable house \_price

It will assumes that the relationship between the predictiors and the target is liner meaning it tries to find the best fit line through the data

* **RANDOM FOREST REGRESSOR**

This model uses ensemble learning to combine the result of multiple decision trees each tree is traind on random subset of the data and the prediction from all the tree are averaged to get the final result

It is non liner meaning it can capture more hard patterns and interactions between feature

**3.MODEL EVALUATION**

* RMSE [root mean squared error] measures the average magnitude of the errors in pridictions with a focus on large errors
* MAE[mean absolute error] : similar to RMSE but less sensitive to outliners
* R²[coefficient of determination]:indicates how well the model explains the variable in the target variable a value closer to 1 indicates a better fit

**4.MODEL COMPARISON AND SELECTION**

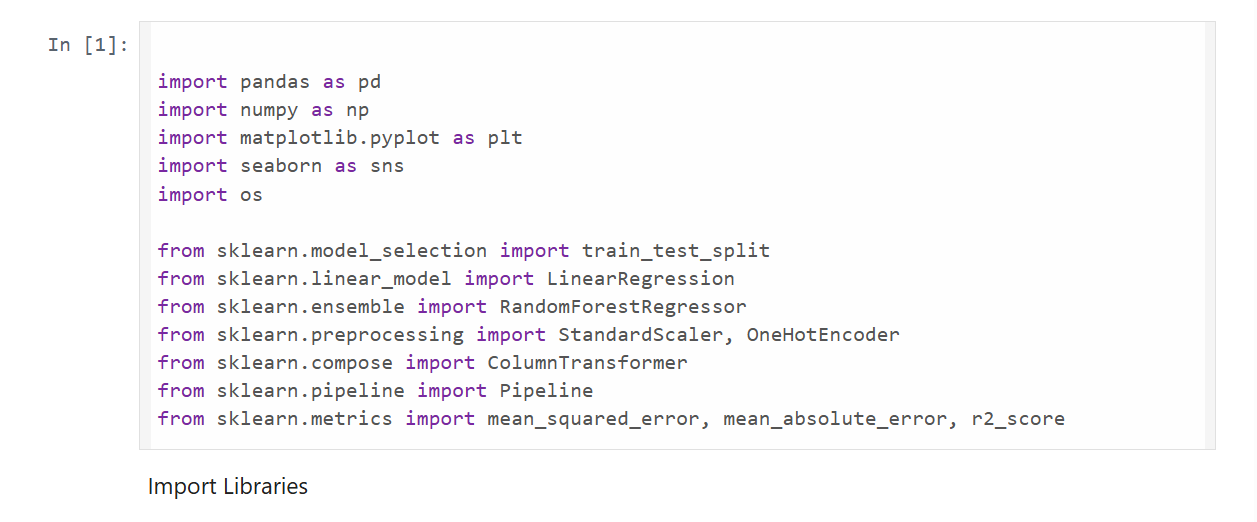
Once we evalute both models using the metrics above we can compare them

If random forest provides significantly better RMSE and R² it would be our model of choice

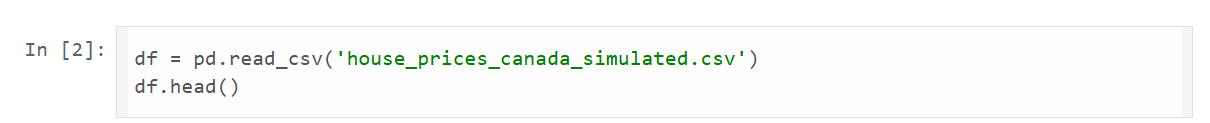
Linear regression can still be useful as a quick baseline model especially if the relationship between feature and price is mostly linear

EXPLANATION PROJECT

1. IMPORT LIBRARIES

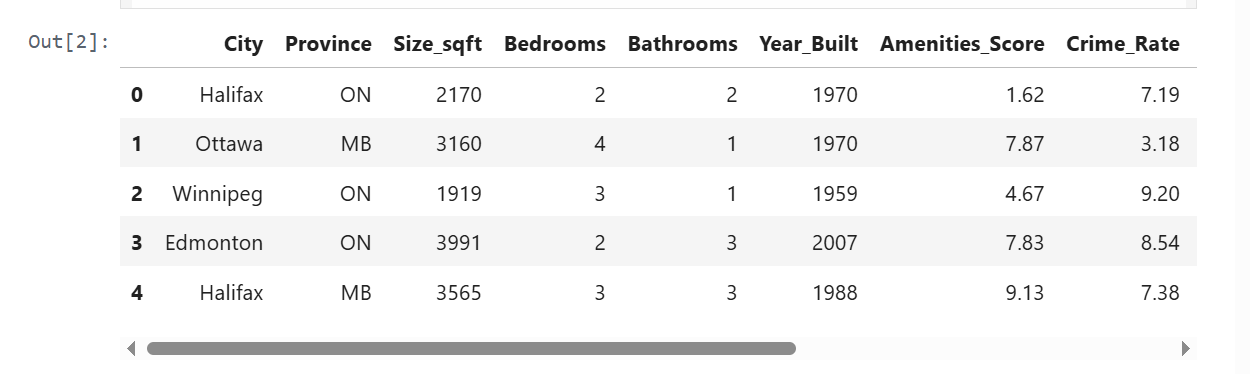


2.LOAD DATATSET

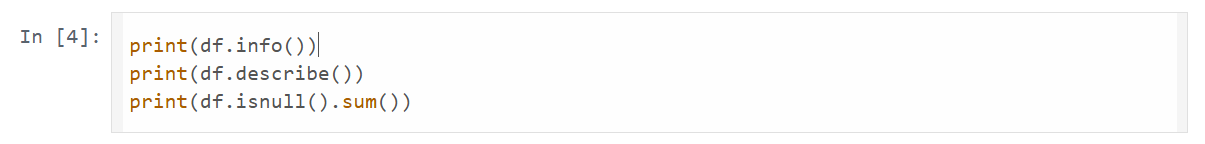


* pd.read\_csv() reads the csv file into the pandas dataframe
* df.head() displays the rows of the datafram

OUTPUT.

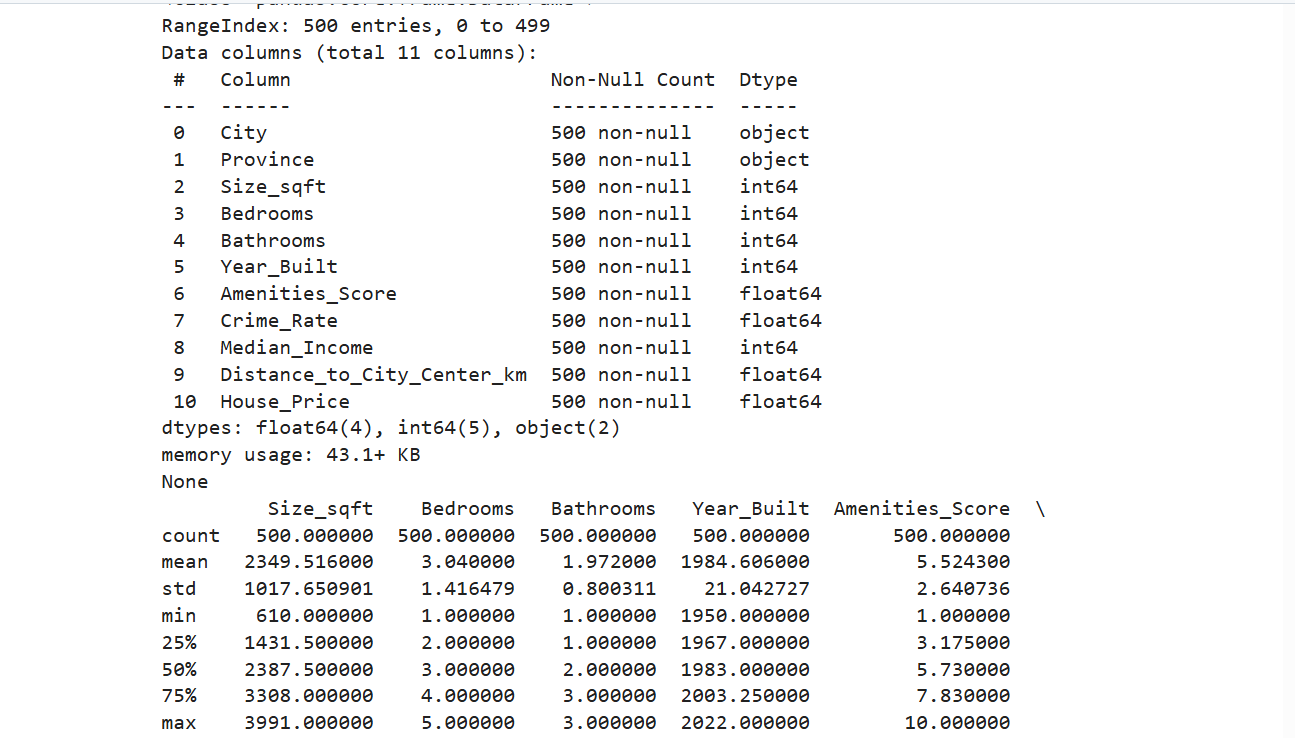


3.INVESTIGATE DATA

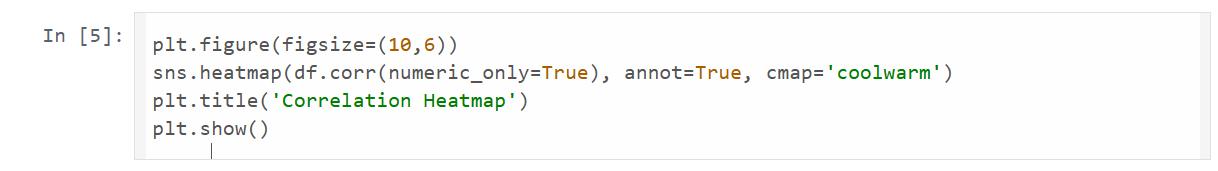


* df.info() displays the summary of the dataframe including the number of entries column names ,data type and non null values
* df.descibe() shows summary staticas of numerical colums such as mean , std , min , max
* df.isnull().sum() checks for missing values in each column and give the count of missing entries

OUTPUT

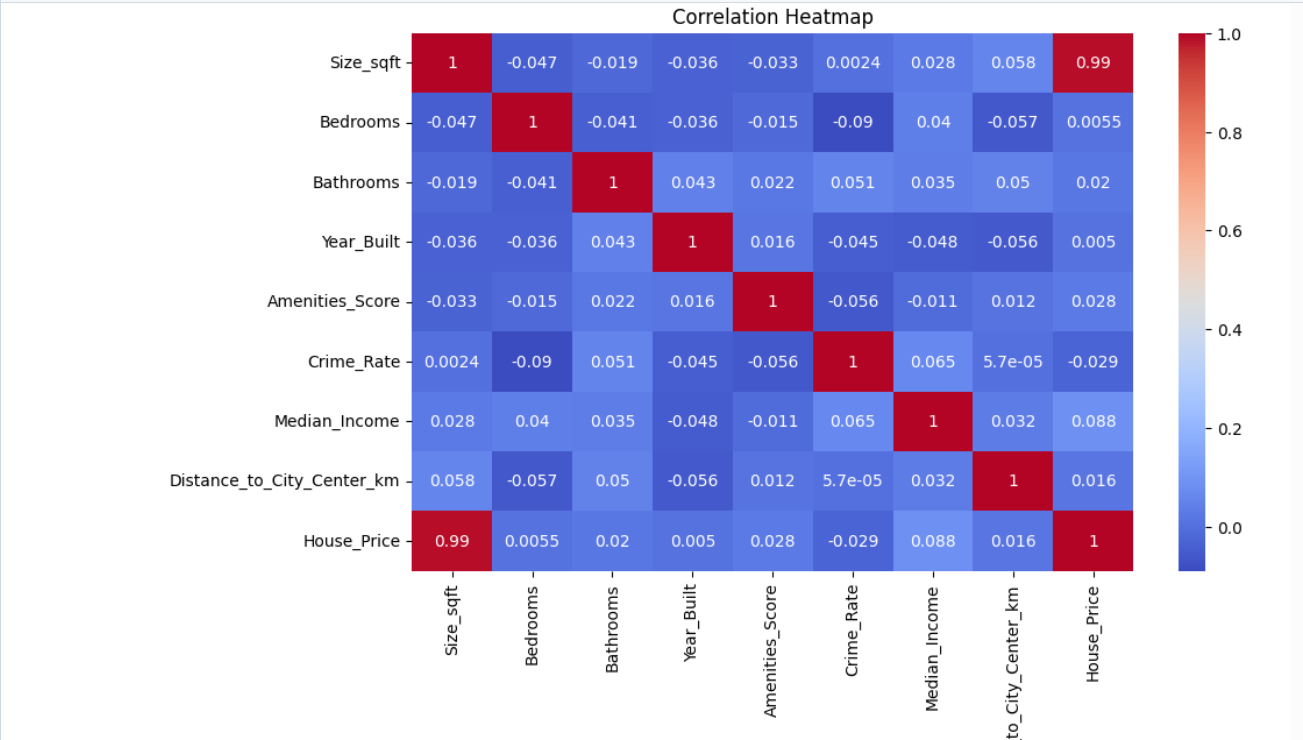


4. Visualize Correlation

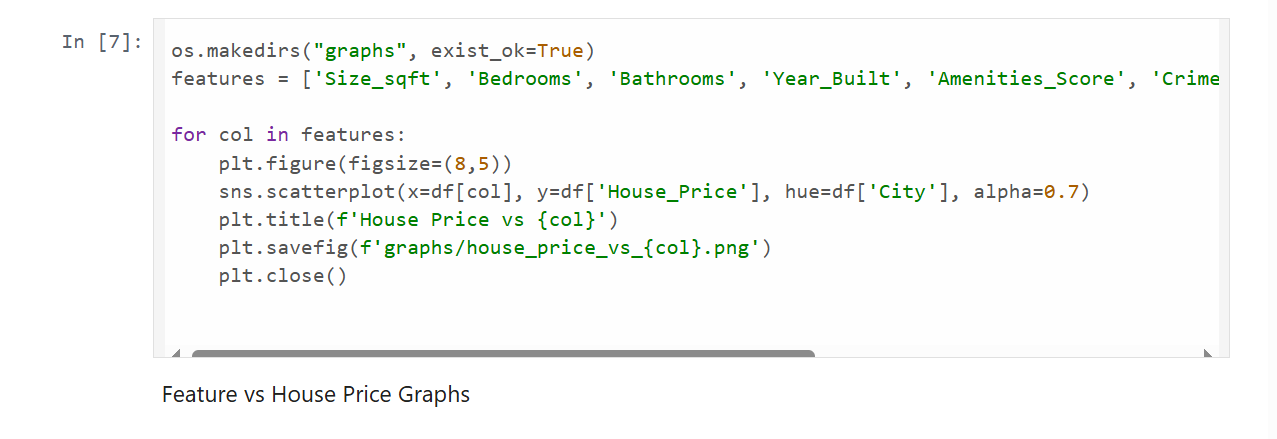


* Plt.figure(figsize=(10,6)) set the figure size to 10x6 for the plot
* df.corr() calculate the correlation between number feature
* annot=true shows the correlation values on the heatmap
* cmap=’coolwarm’ sets the color scheme for the heatmap

OUTPUT



5. Feature vs House Price Graphs

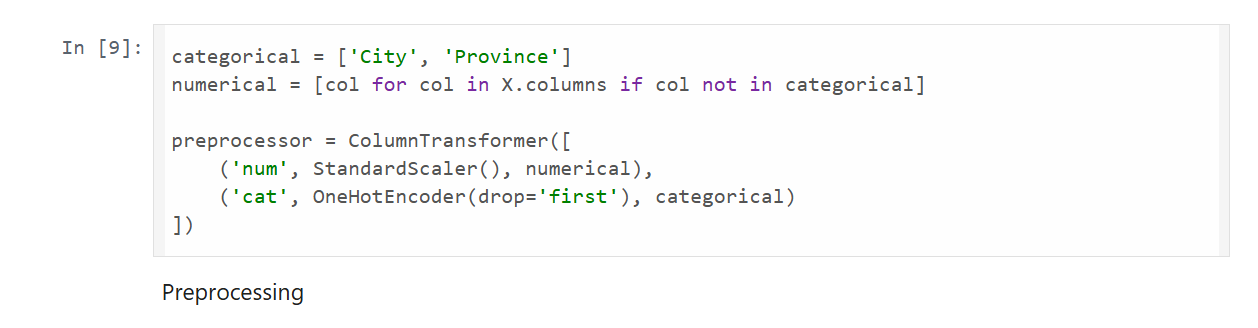


* os.makedirs(“graphs” , exits\_ok= true) create a folder name if it do not already exits to save the plot images

6.Prepare features and target

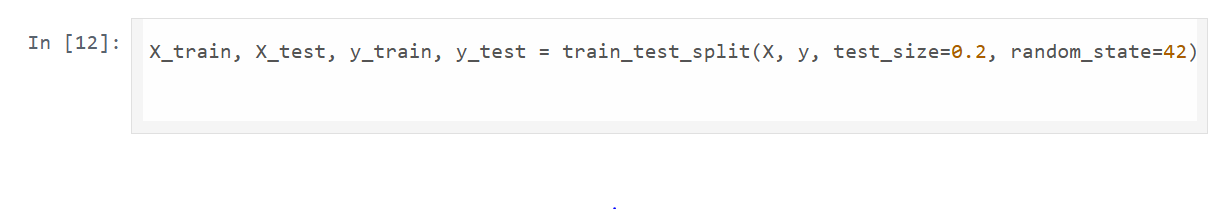


* X contains all the features so it drops the house price column from the datafram
* Y contains the target variable which we want to predict

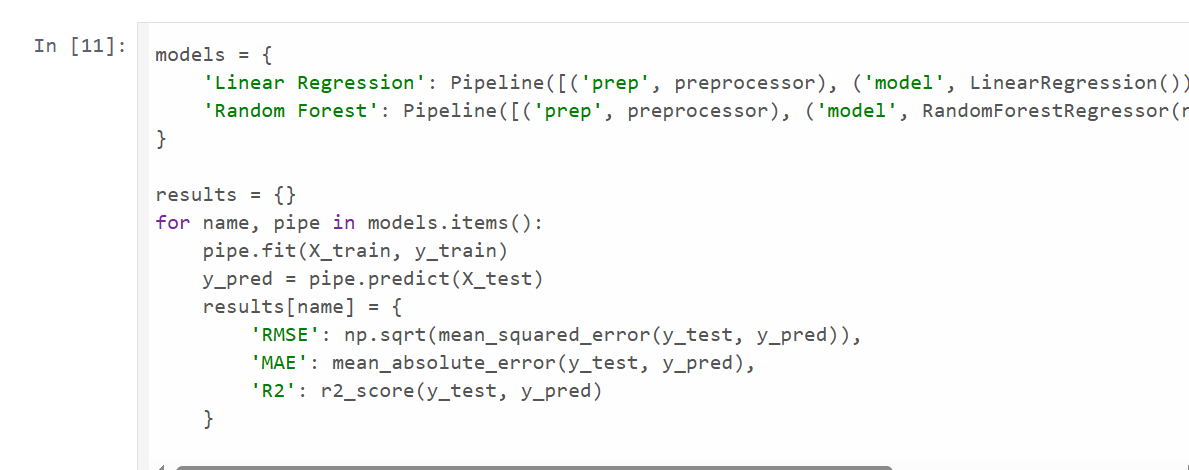
7.Preprocessing

Categorical = [‘city’,’province’] .list of categorical columns that need encodeing

Numerical=[] list of numerical column by the categorical colums

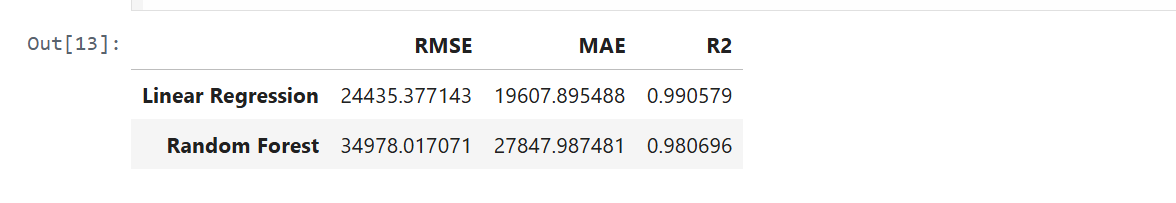
8.TRAIN AND SPILIT

9.BUILD AND TRAIN DATA

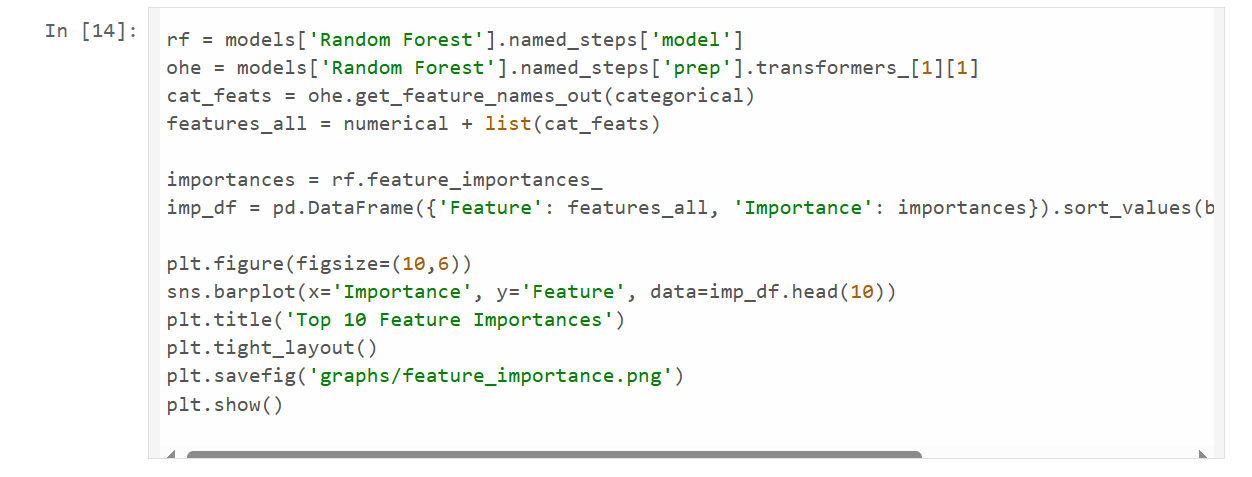


10.FRAME

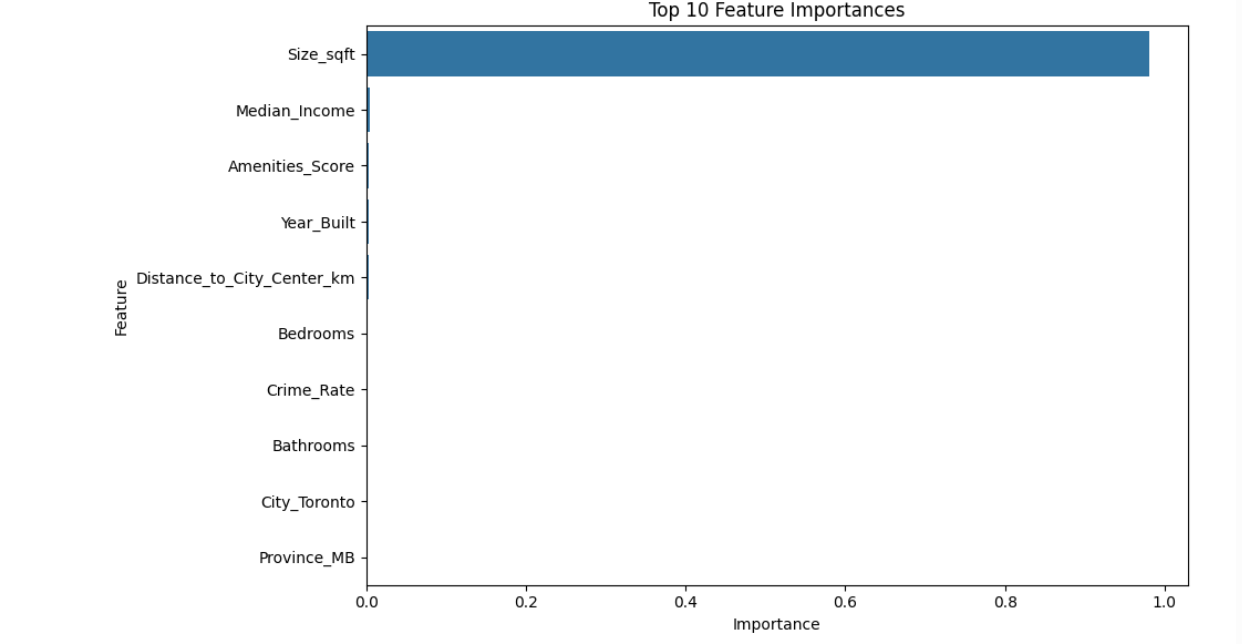
OUTPUT



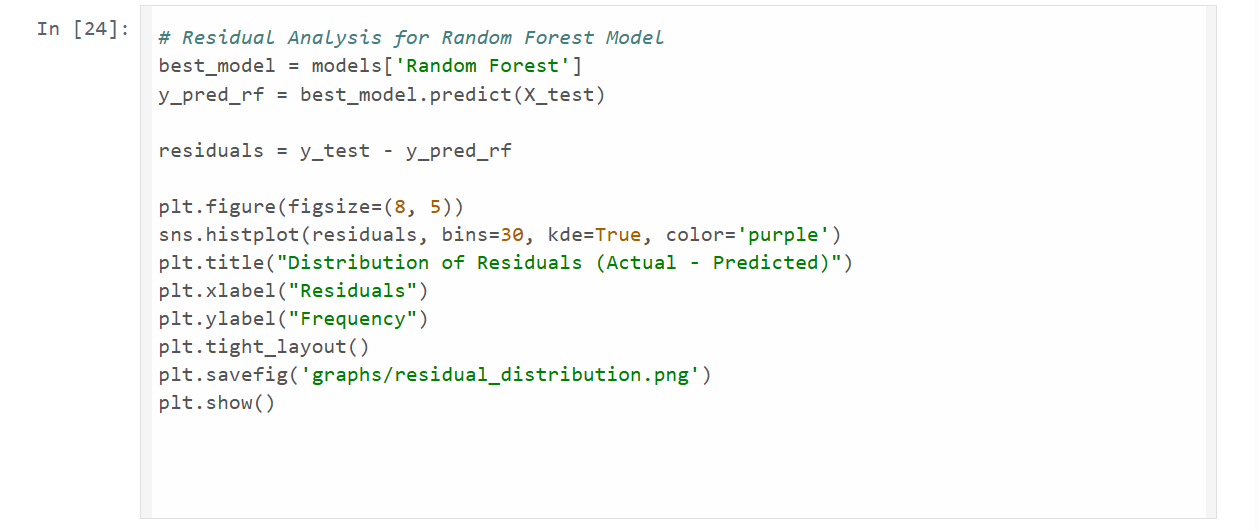
11.FEATURES IMPORTANCE RANDOM FOREST



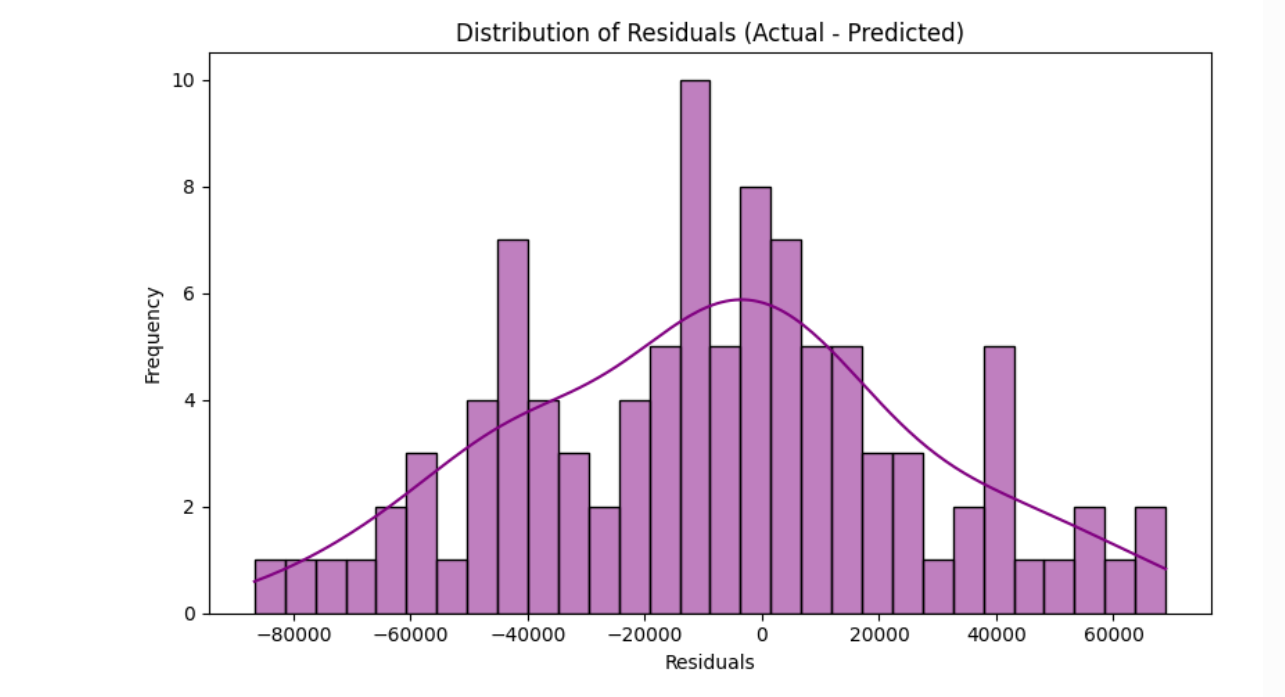
OUTPUT



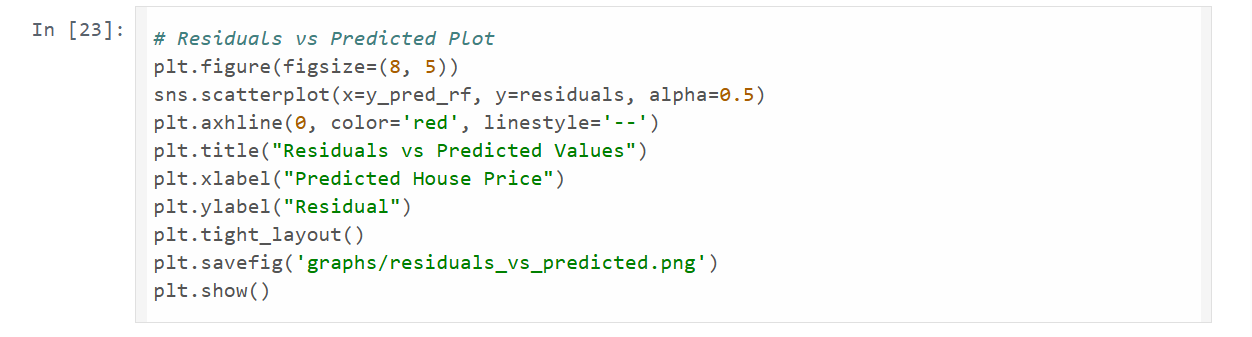
12. Residual Analysis for Random Forest Model



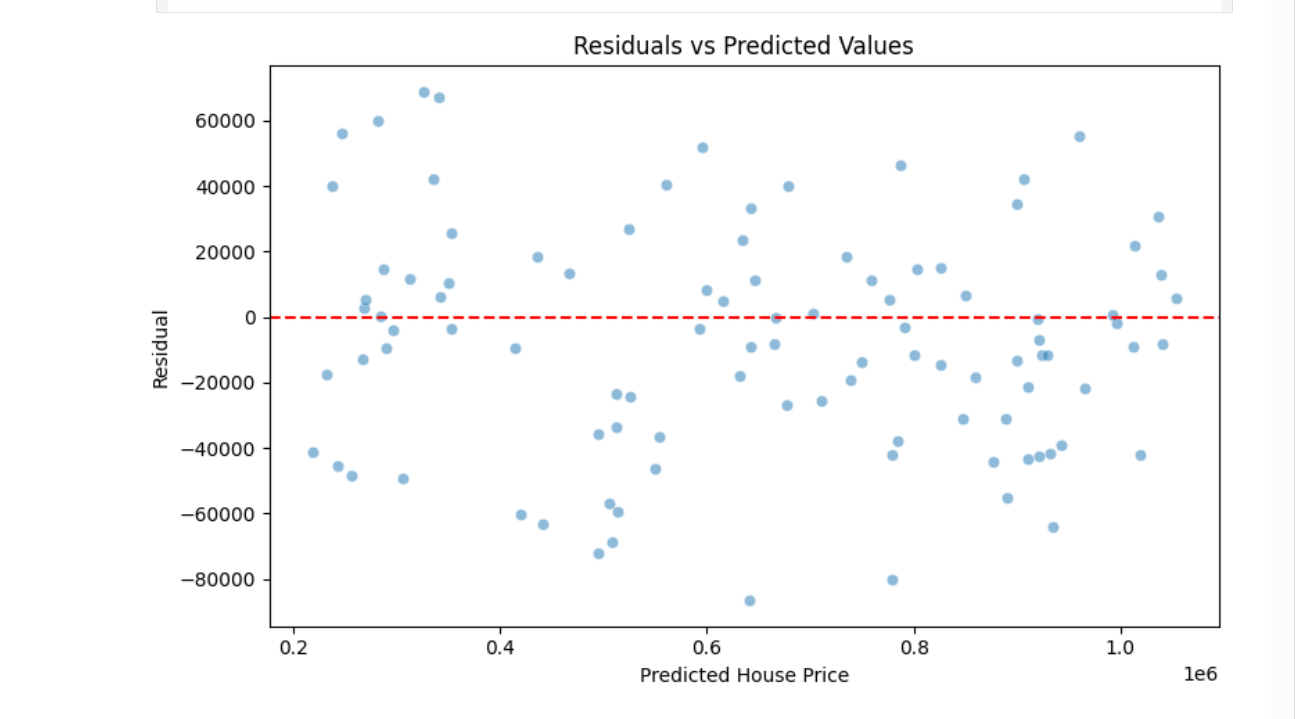
OUTPUT

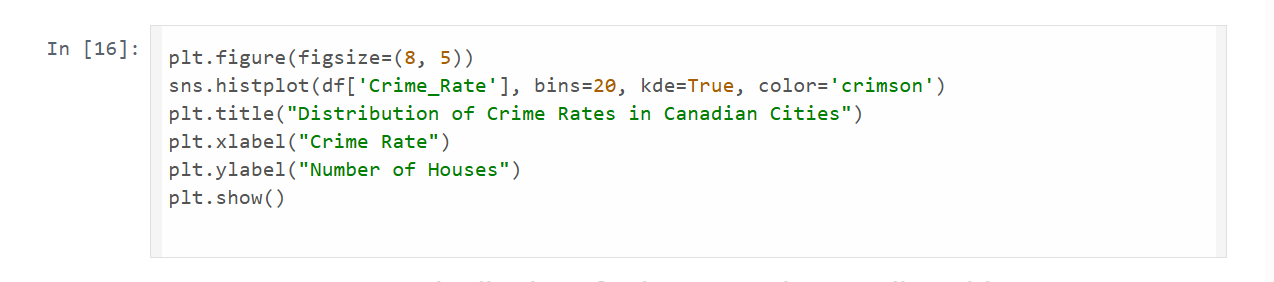


13. Residuals vs Predicted Plot

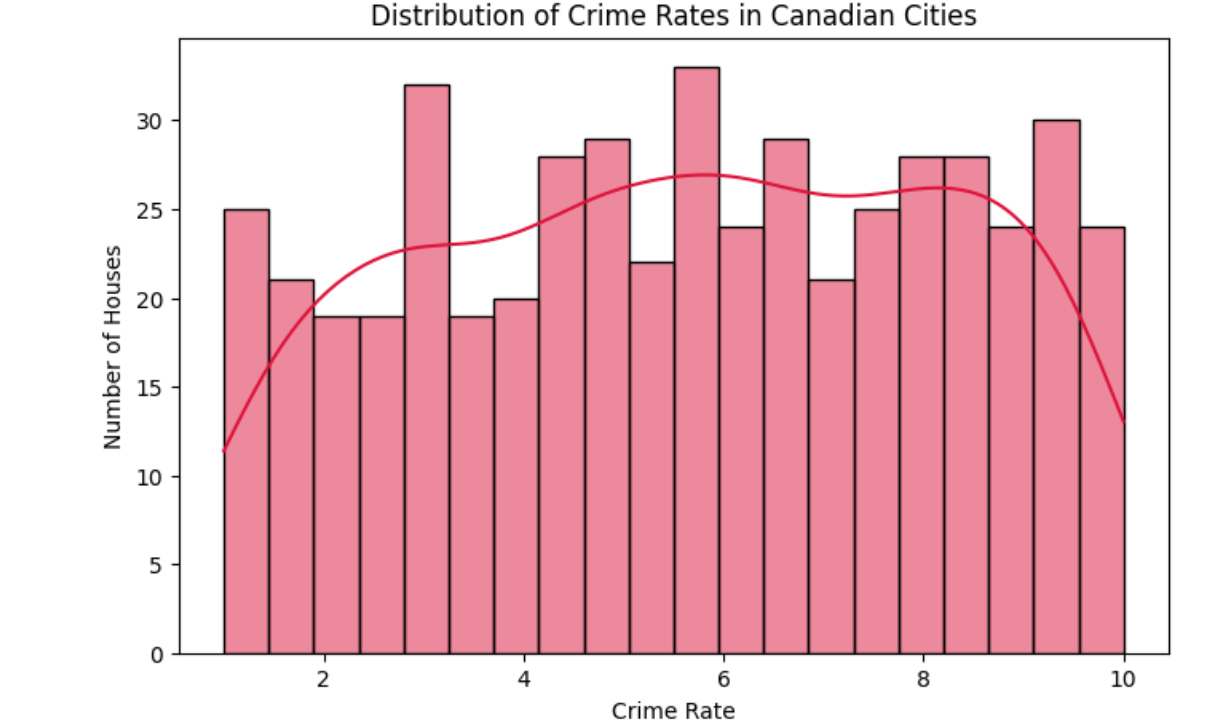


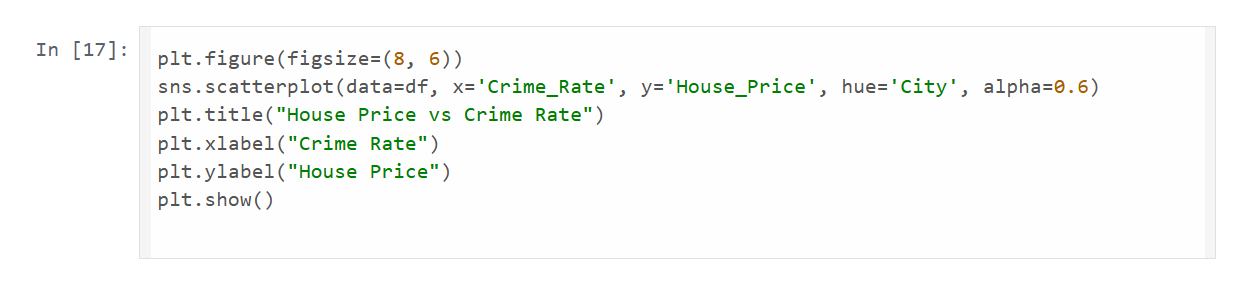
OUTPUT



14. Residual Analysis for Random Forest Model

OUTPUT

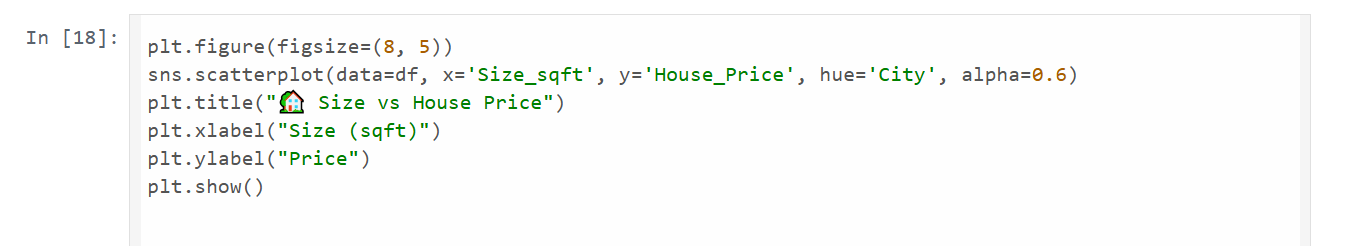


15. HOUSE PRICE VS CRIME RATE

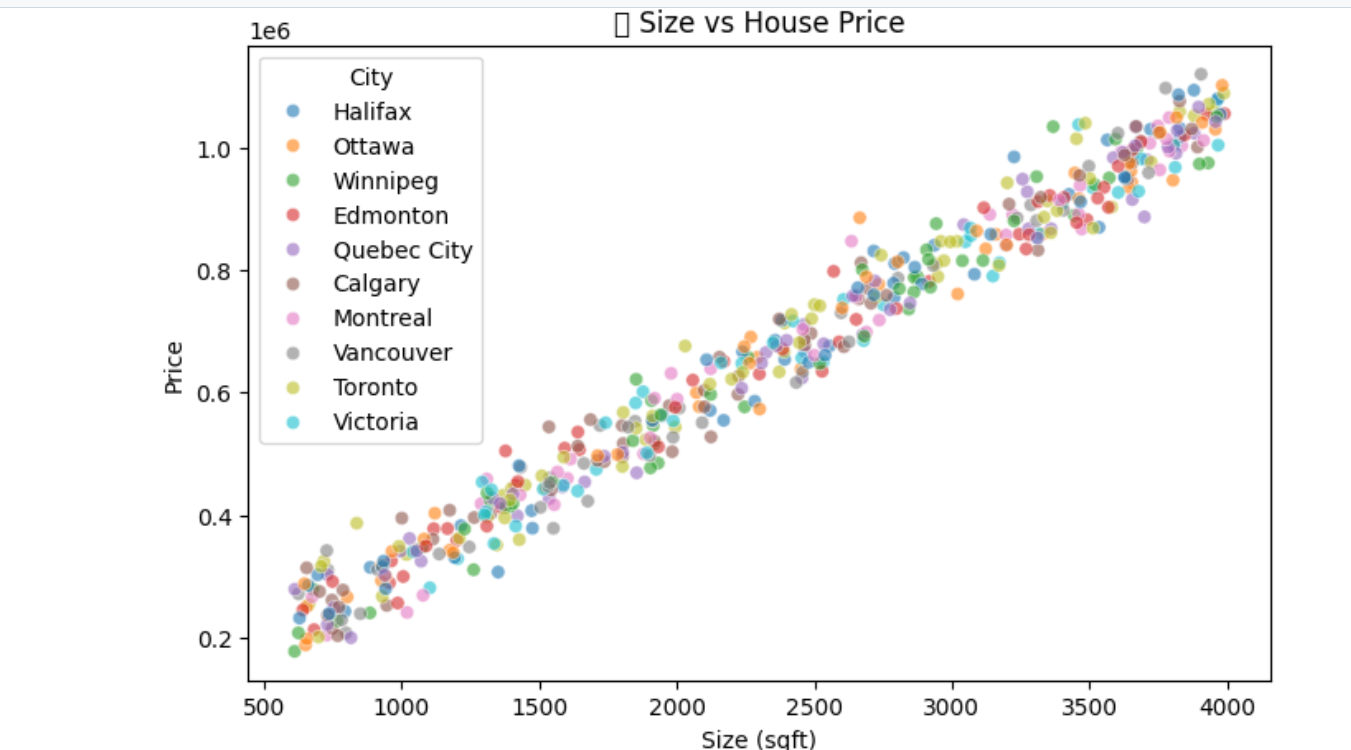
OUTPUT:



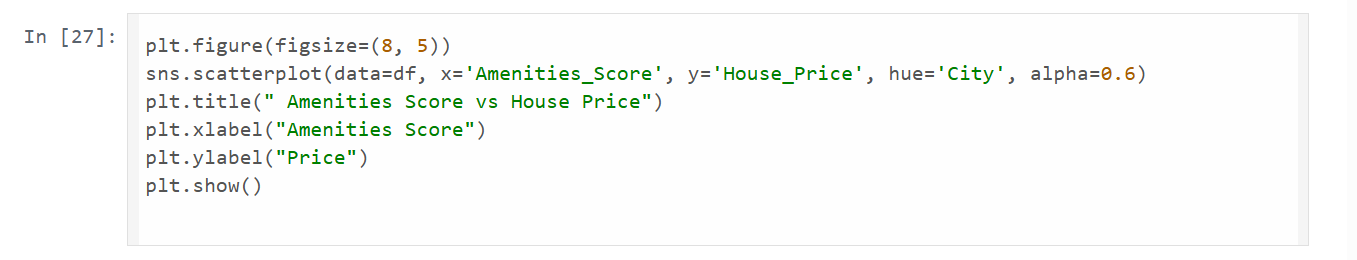
16. HOUSE SIZE VS HOUSE PRICE



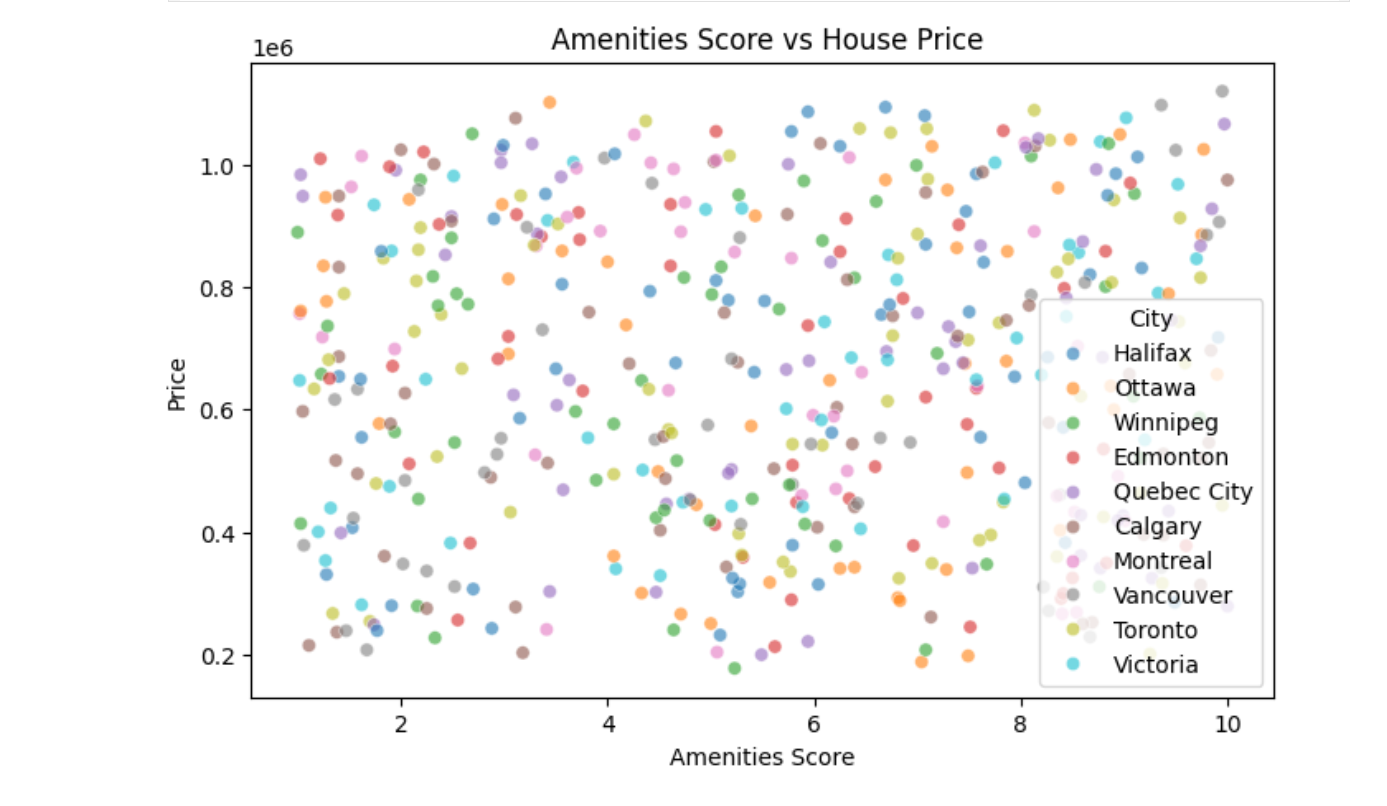
OUTPUT

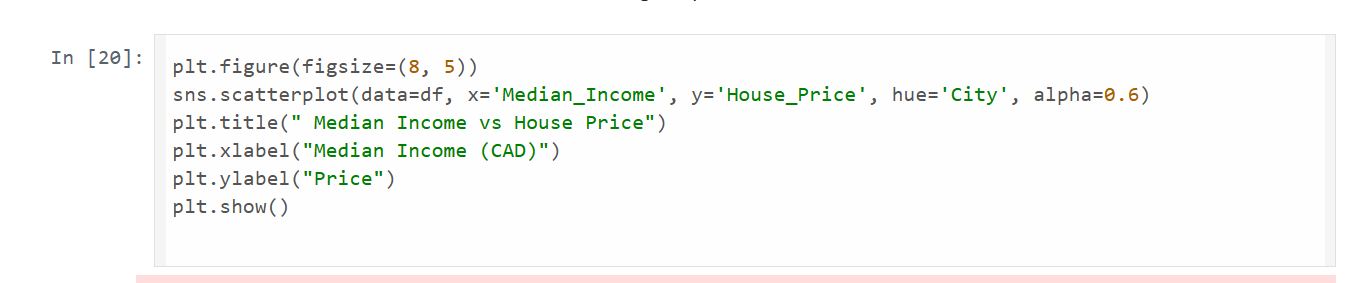


17. AMENITIES SCORE VS HOUSE PRICE

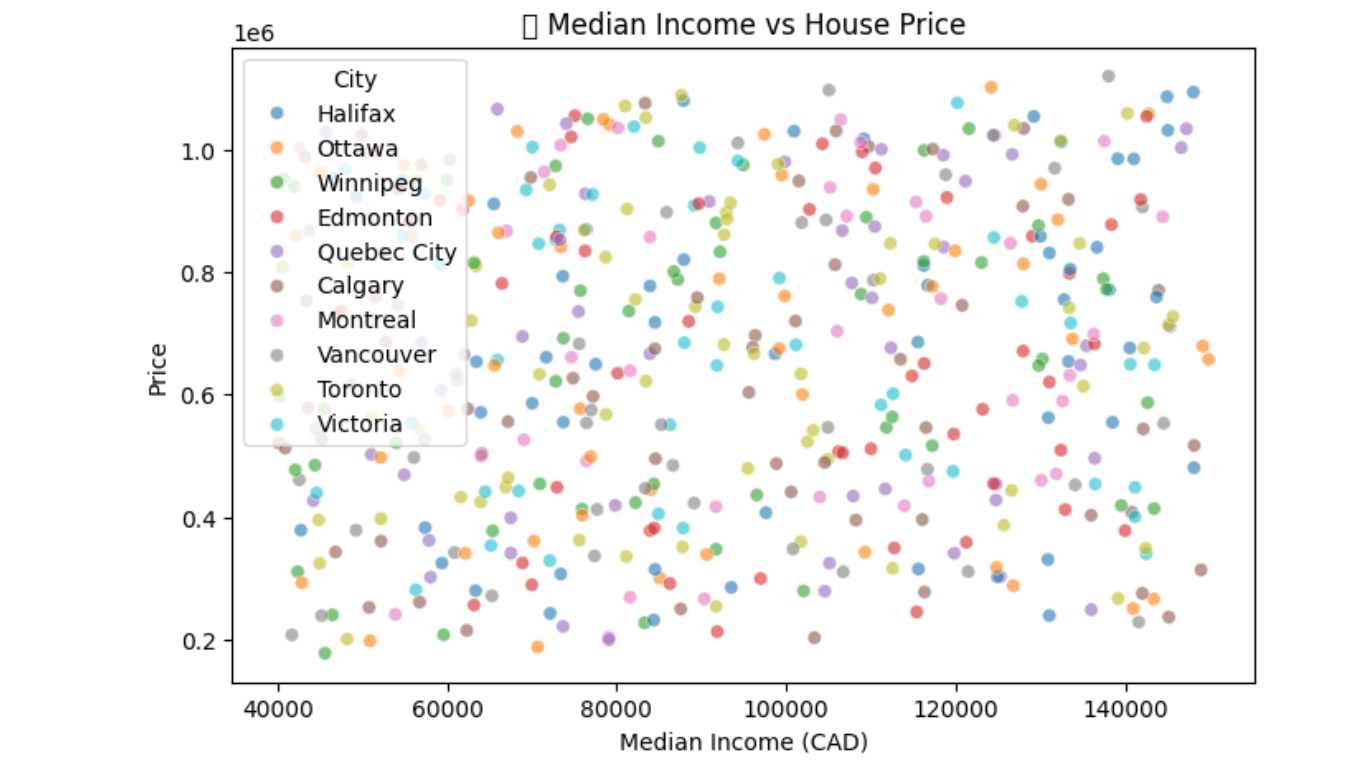


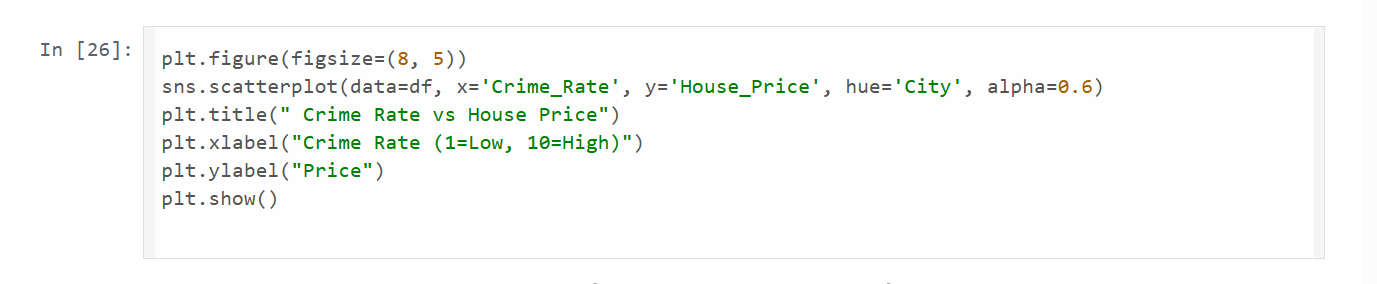
OUTPUT



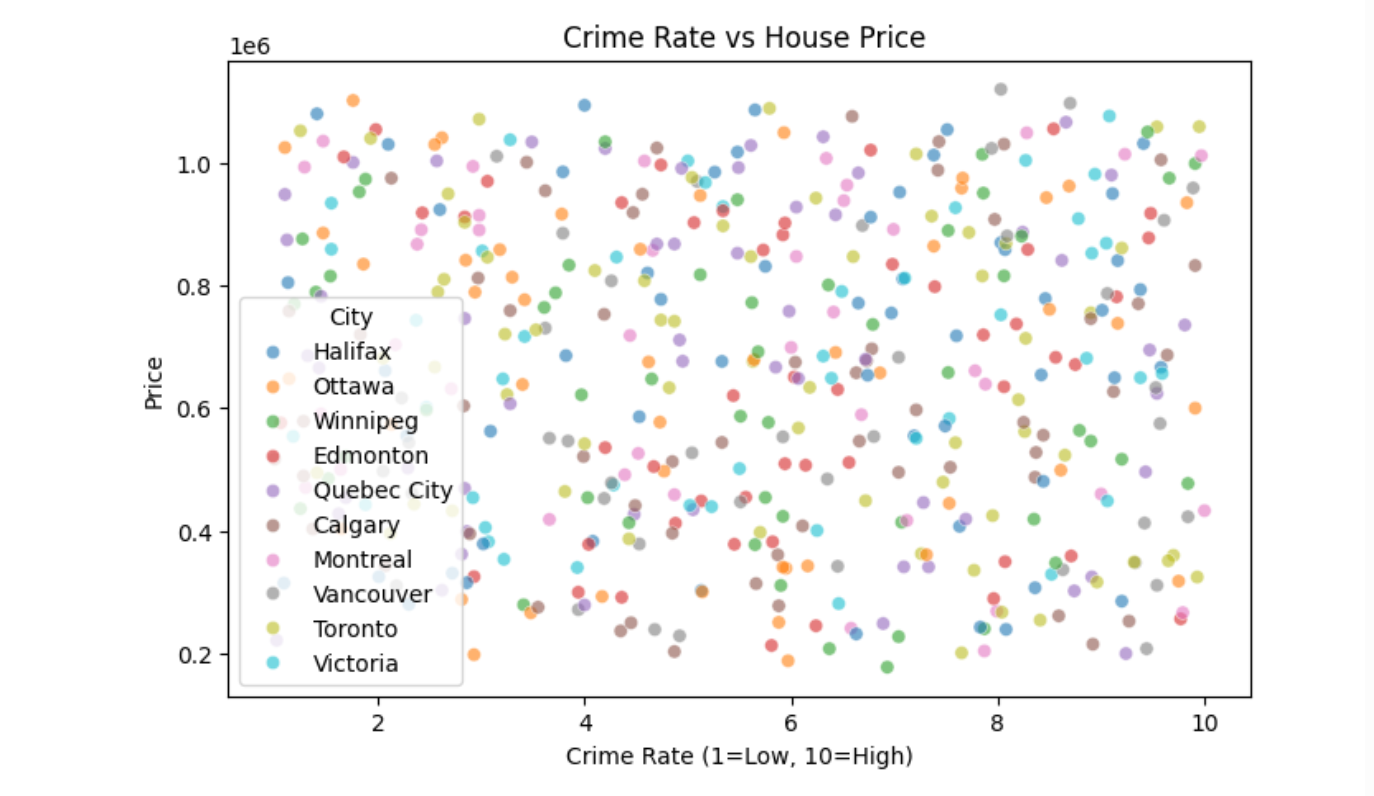
18. MEDIAN INCOME VS HOUSE PRICE

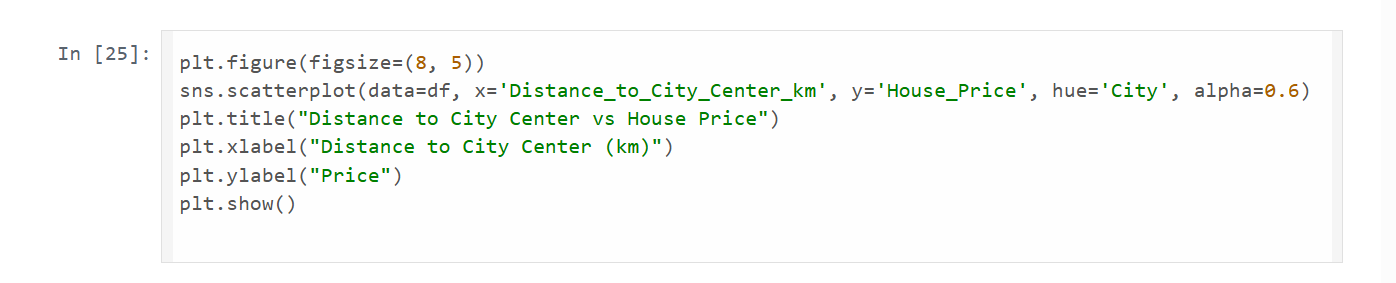
OUTPUT:



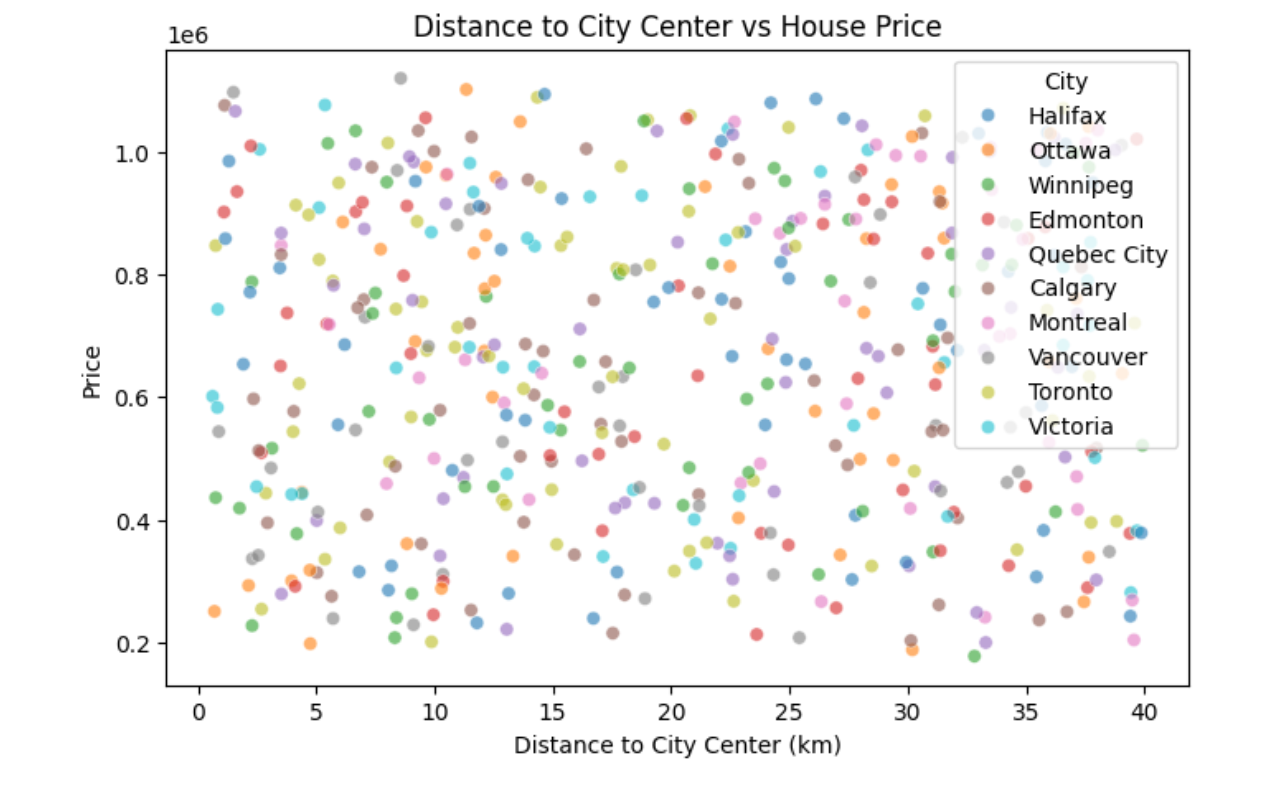
19. CRIME RATE VS HOUSE PRICE

OUTPUT

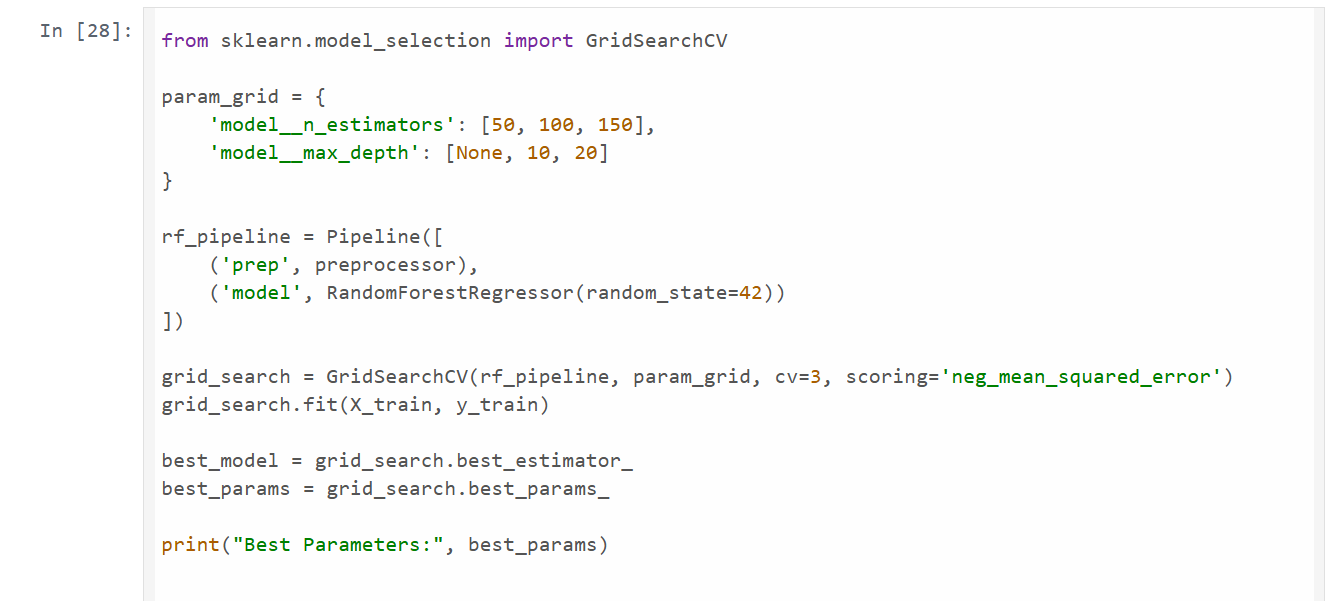


20. DISTANCE TO CITY CENTER VS HOUSE PRICE

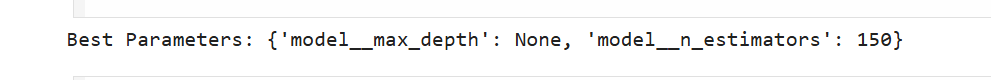
OUTPUT



21. BEST PARAMETERS



OUTPUT



Future Improvements for the Project

**1.hyperparameter tuning and advanced model**

To using with more advanced models l CatBoost , which are known to provide best performance over random forest in many regression tasks

1. **Feature engineering**

Interaction feature : create new feature by combining once such as price per square and distance to city center per bedrooms

Log transformation : to use log transformation on twice feature like houce price to make the distribution more normal, improving model prediction

1. **Model interpretability**

Use tool like SHAP or LIME provide best summary to understand which feature most usable for prediction

Feature impotance visualization foe tree based model will help stakeholder understand what drives house price