Predicting the price of houses

Soumyendra Shrivastava, Siddhant Sancheti, and Tanuja Reddy Maligareddy, *SJSU*

**Abstract**—Building a model that can precisely forecast house prices based on a collection of input data is the goal of a house price prediction project. This can assist developers, investors, and real estate agents in making educated decisions on the value of real estate, as well as house buyers and sellers.

**Index Terms**—NatWalkInd (National walkability index), D2B\_E8MIXA(mix of employment types in a block group),

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

————————————————

* *Soumyendra Shrivastava; E-mail: soumyendra.shrivastava@sjsu.edu.*
* *TanujaReddy Maligierddy; E-mail: tanujareddy.maligireddy@sjsu.edu*
* *Siddhant Sancheti; E-mail: siddhant.sancheti@sjsu.edu*
* *Code Repository: (See Appendix A)*

—————————— ◆ ——————————

# 1 Introduction

The real estate market is one of the most important and complex industries in the world. With the constantly changing market dynamics, predicting the prices of houses accurately has become increasingly challenging. This is where machine learning comes into play. In recent years, machine learning has become an increasingly popular tool for predicting housing prices due to its ability to handle large volumes of data and identify patterns that are not easily visible to the human eye.

In this project, we aim to use machine learning techniques to predict the prices of houses. We will use a dataset consisting of various features such as location, number of bedrooms, square footage, etc. to train our model. This project aims to build a model that can accurately predict the price of a house given its features.

We will begin by exploring the dataset and performing exploratory data analysis (EDA) to understand the relationship between the different features and the target variable (house price). We will then preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features. After that, we will train different machine learning models such as linear regression, decision trees, and random forests and compare their performance using metrics such as root mean squared error (RMSE) and R-squared.

The results of this project will not only help us understand the factors that affect the price of a house but also provide valuable insights to real estate agents and investors who are looking to make informed decisions. With accurate predictions, buyers and sellers can make better decisions and negotiate better deals, while investors can identify profitable opportunities in the market.

Overall, this project will demonstrate the power of machine learning in predicting housing prices and its potential to revolutionize the real estate industry.

# 2 Data Collection

The data for our house price prediction project was collected from two primary sources: Zillow and the Federal Housing Finance Agency (FHFA). Zillow is a well-known online real estate marketplace that provides a variety of data on housing markets, including property value estimates, rental prices, and historical sales data. We used Zillow's public API to access housing market data for our study.

In addition, we also utilized the House Price Index (HPI) data provided by the FHFA, which is a government-sponsored enterprise that regulates the housing finance market. The FHFA HPI is a comprehensive measure of changes in single-family house prices across the United States.

The combination of Zillow and FHFA data provided a robust and comprehensive dataset for our house price prediction project.

## 3 Exploratory Data Analysis and Visualization(EDAV)

Using lmplot() to visualize the relationship between two variables, i.e total\_sqft(on the x-axis) and price(on the y-axis) we observe a quite good correlation between the square footage and price

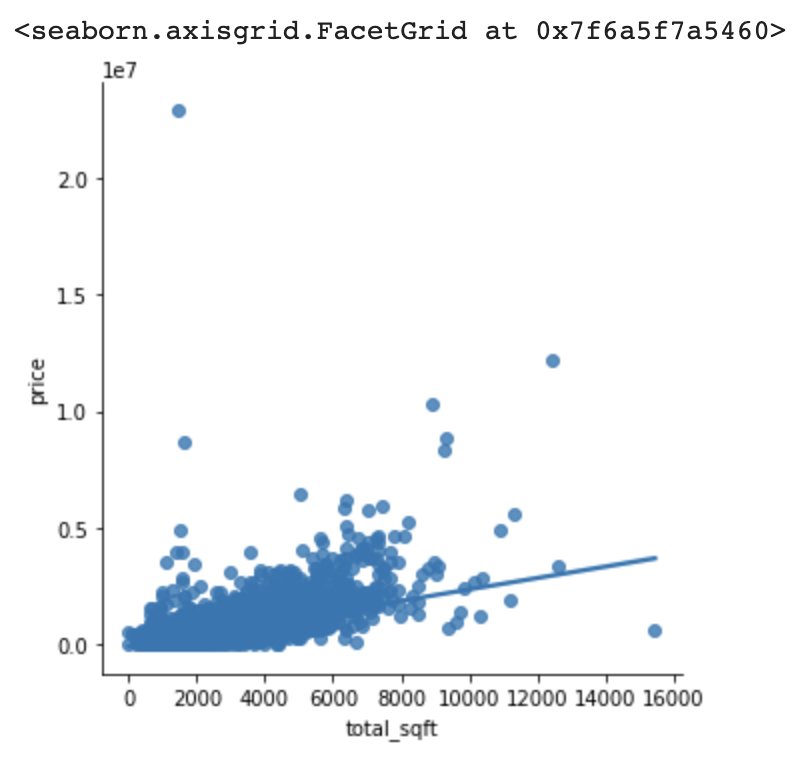


Fig 1:Correlation between total\_sqft and price

Visualization of bedrooms distribution

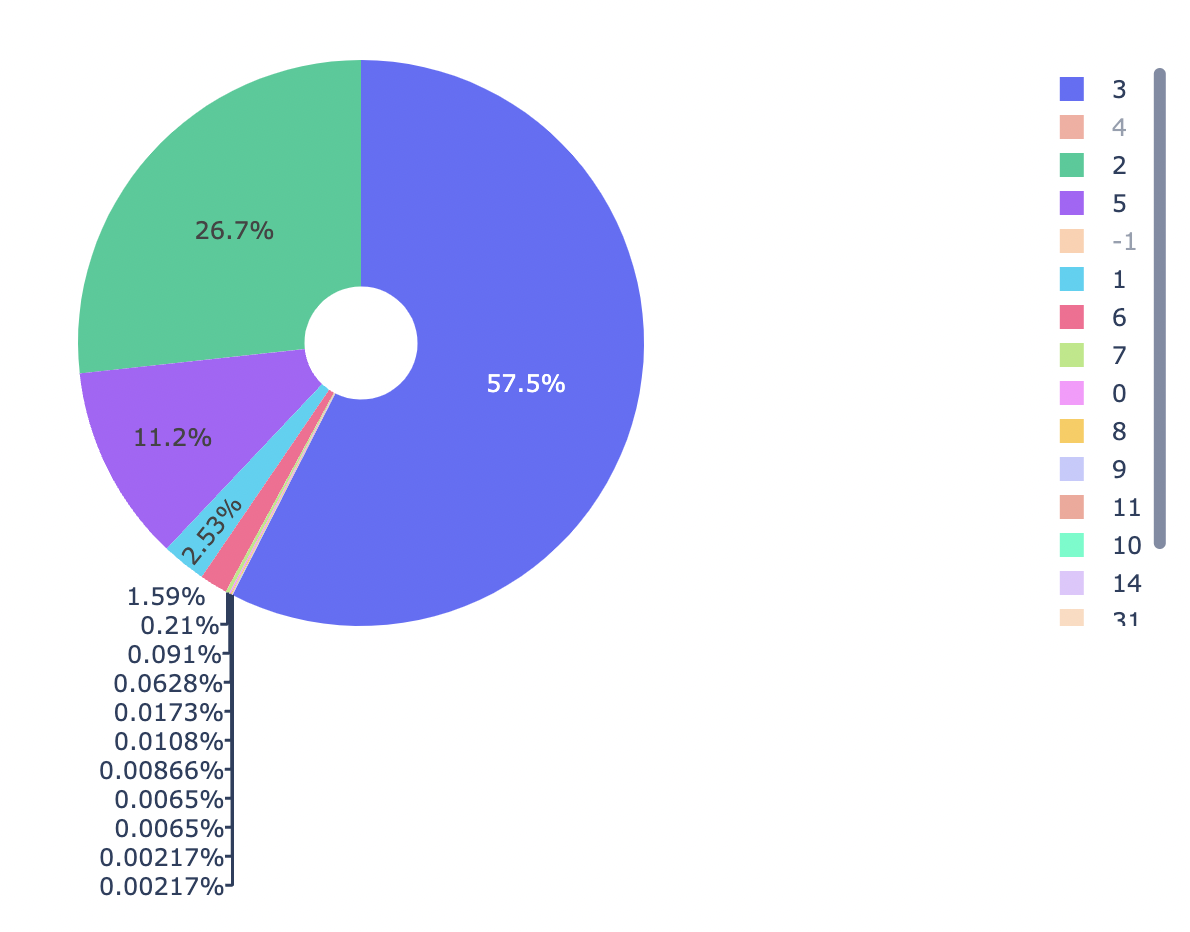


Fig 2: Bedroom Distribution

Using countplot() on the categories bedrooms,review\_score, and stories we found that the apartments with 3 bedrooms, with -1 review score, and 1 storied are the most popular.

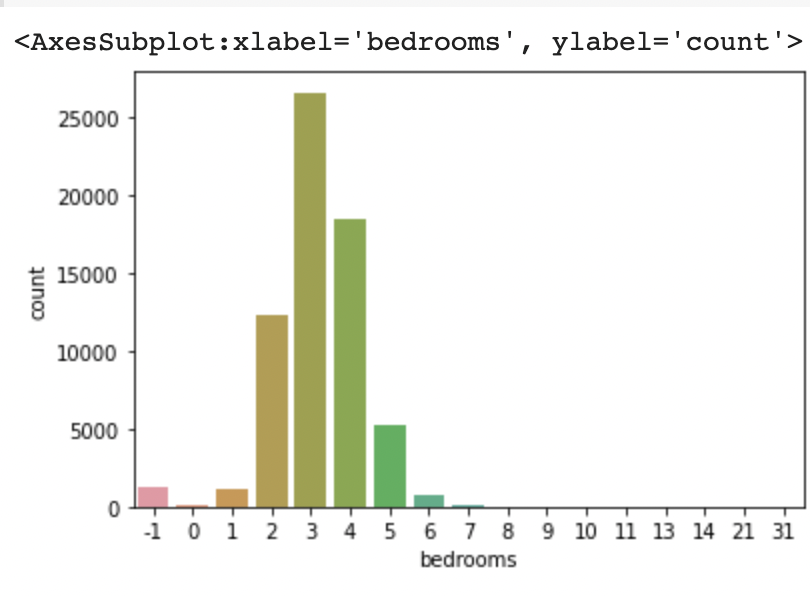


Fig 3: Countplot of bedrooms

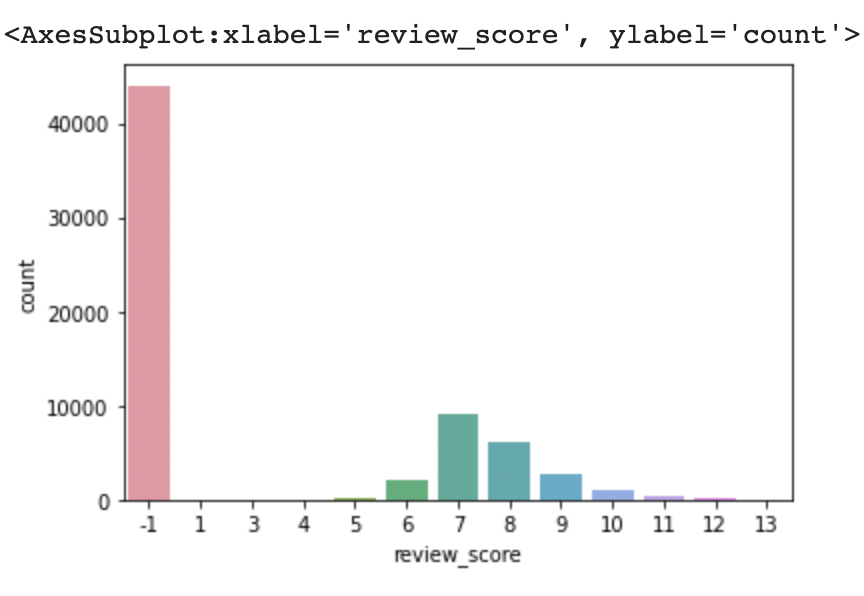


Fig 4: Countplot of review\_score

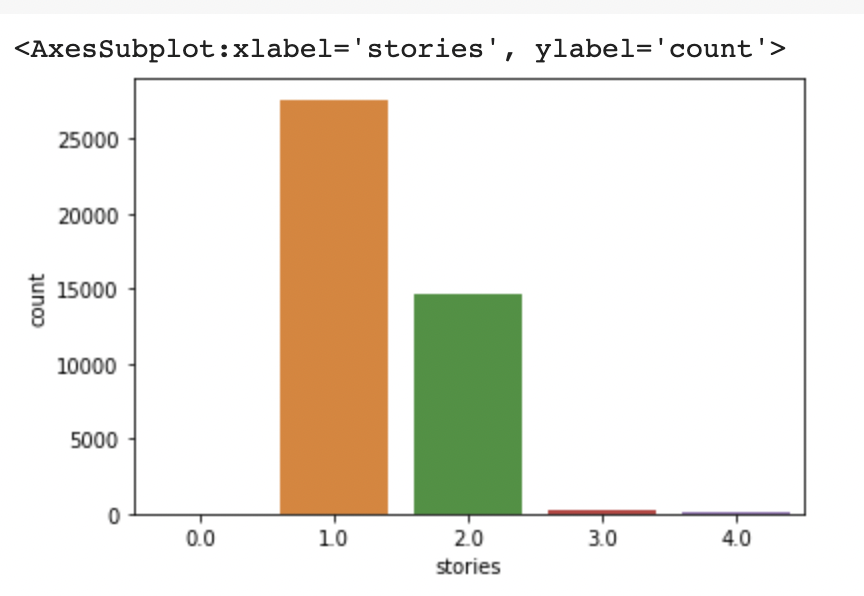


Fig 5: Countplot of stories

To understand the graphical representations of the distribution of the variables of the dataset i.e bedrooms, total\_sqft, price, review\_score we used the basic histplot() function and we found that the data is almost right-skewed, so we tried to fix this by using log transformation which is used to transform skewed data to approximately conform to normality.

Despite using the log transformation we observed some outliers in these columns which we can't be fixed using so we fixed them manually.

# 4. Feature Importance

# To identify the important features of the dataset we used the Decision tree Gini Scores and Random Forest classification. We plotted a decision tree using graphviz

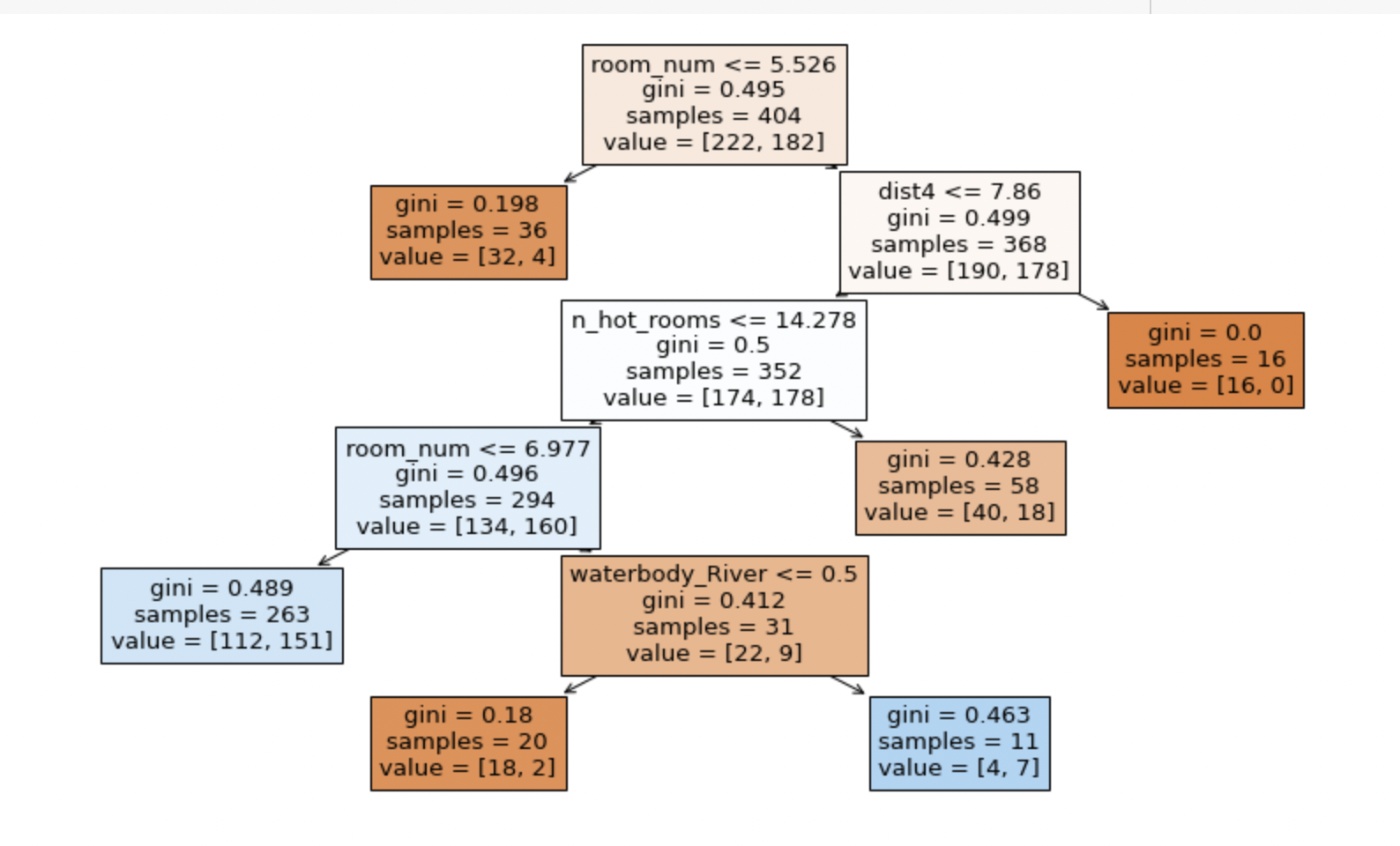


Fig 6: Decision tree plot using Gini Scores

Calculated the performance measures on the training dataset and based on the scores we got later subjected the dataset to hyper tuning to resolve overfit resulting in decent performance scores.

Defined a function FeatureImp() which takes the resultant data model as the input which resulted in the following feature importance graph

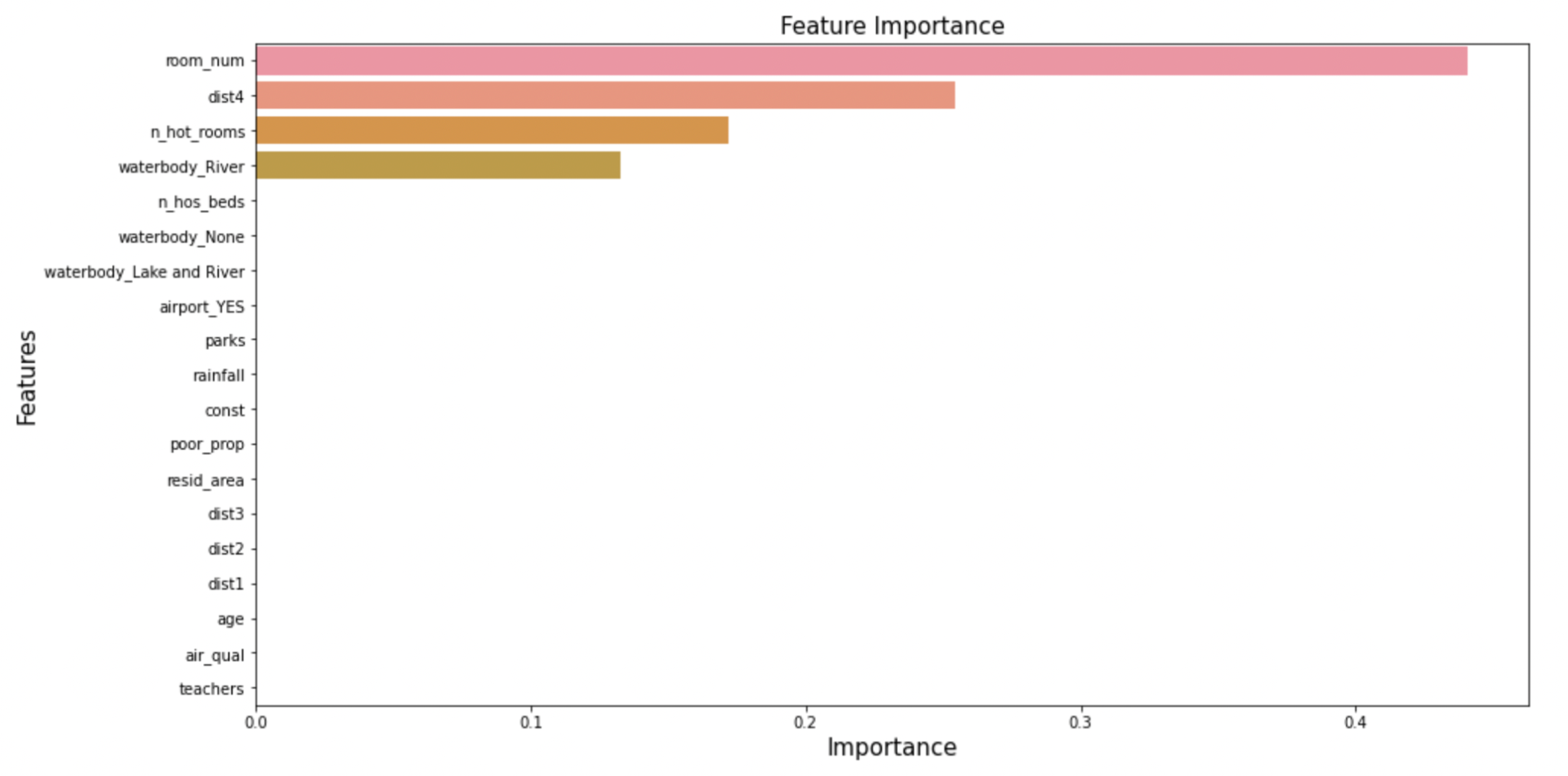


Fig 7: Feature Importance

## 5. Fractal Clustering

For Fractal Clustering we used dataset 3 in which we analyzed some variables like Zipcode, City, Metro, and County name. Later we did the EDA on zip code by plotting the average home value by month and year.

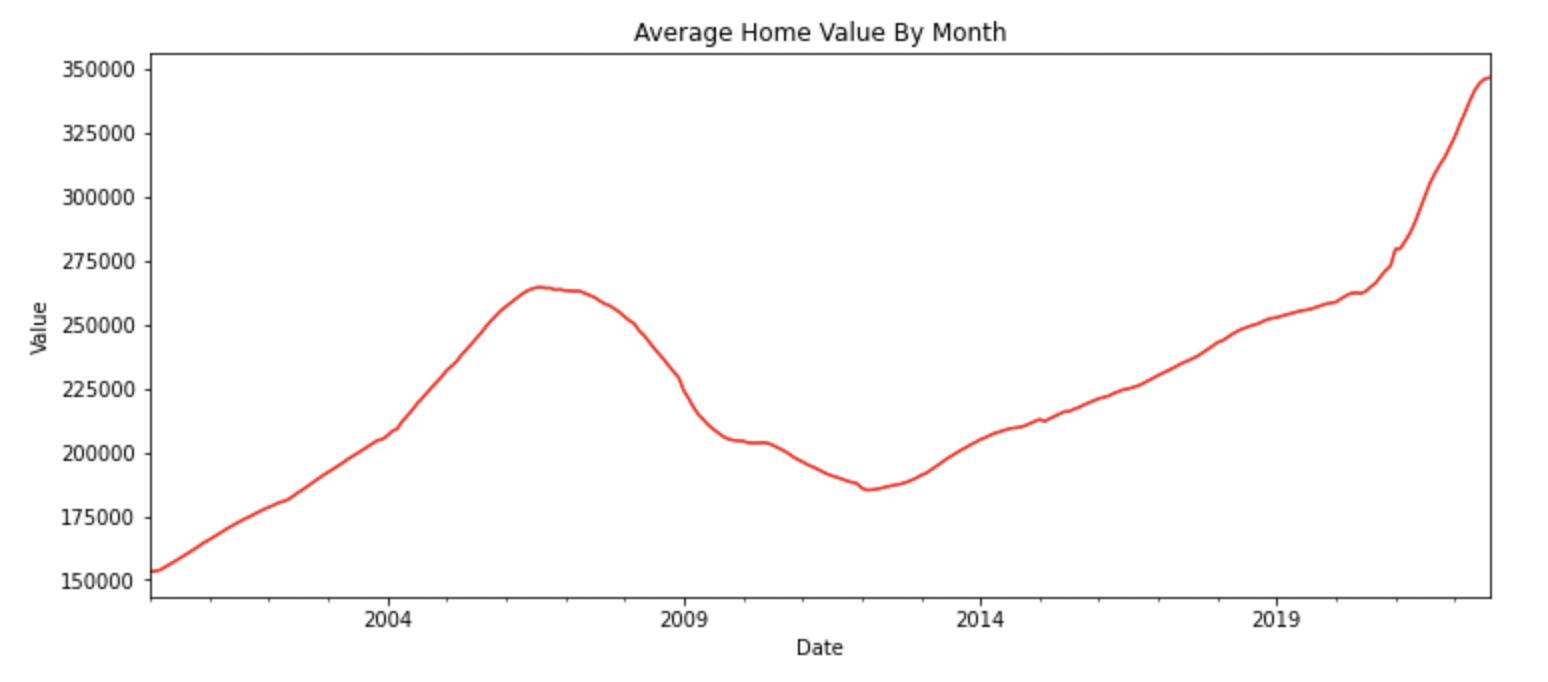


Fig 8:Average home value by month

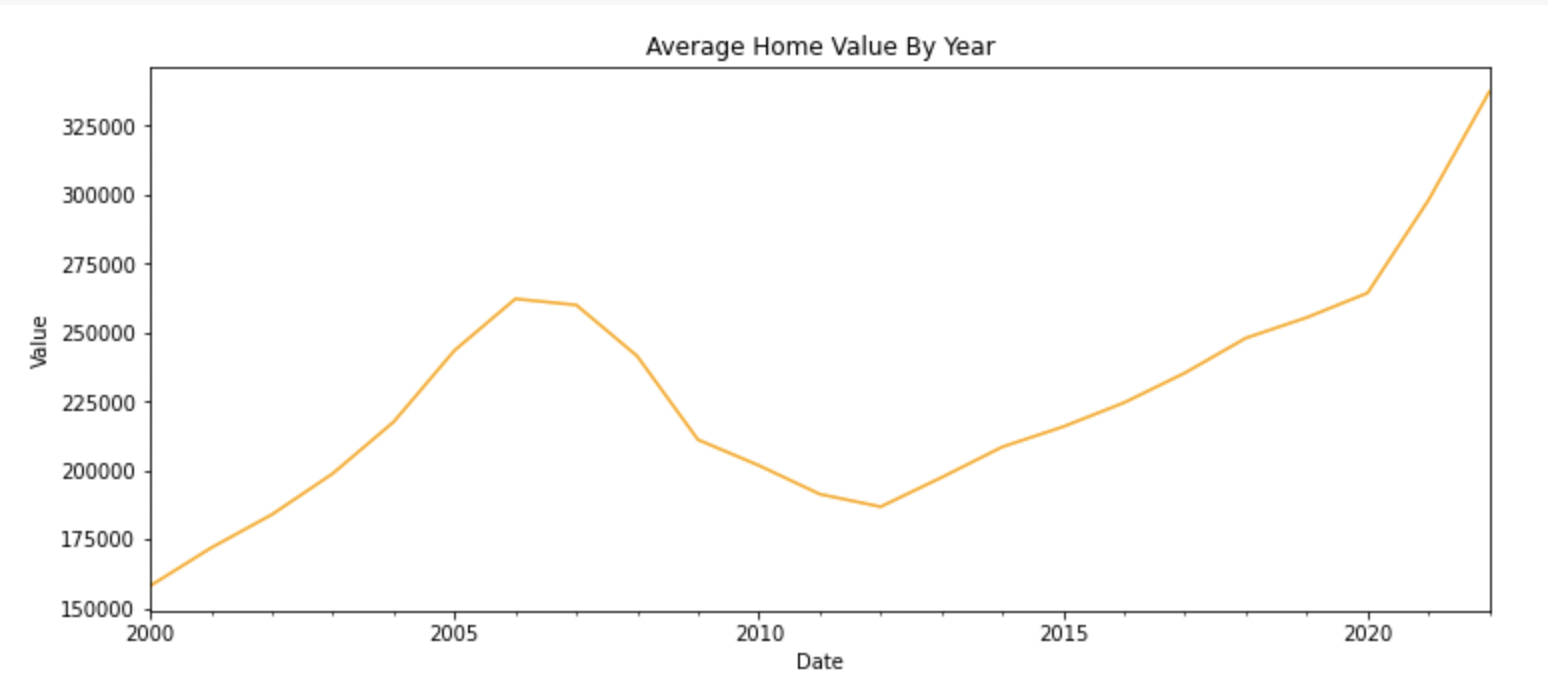


Fig 9: Average home value by year

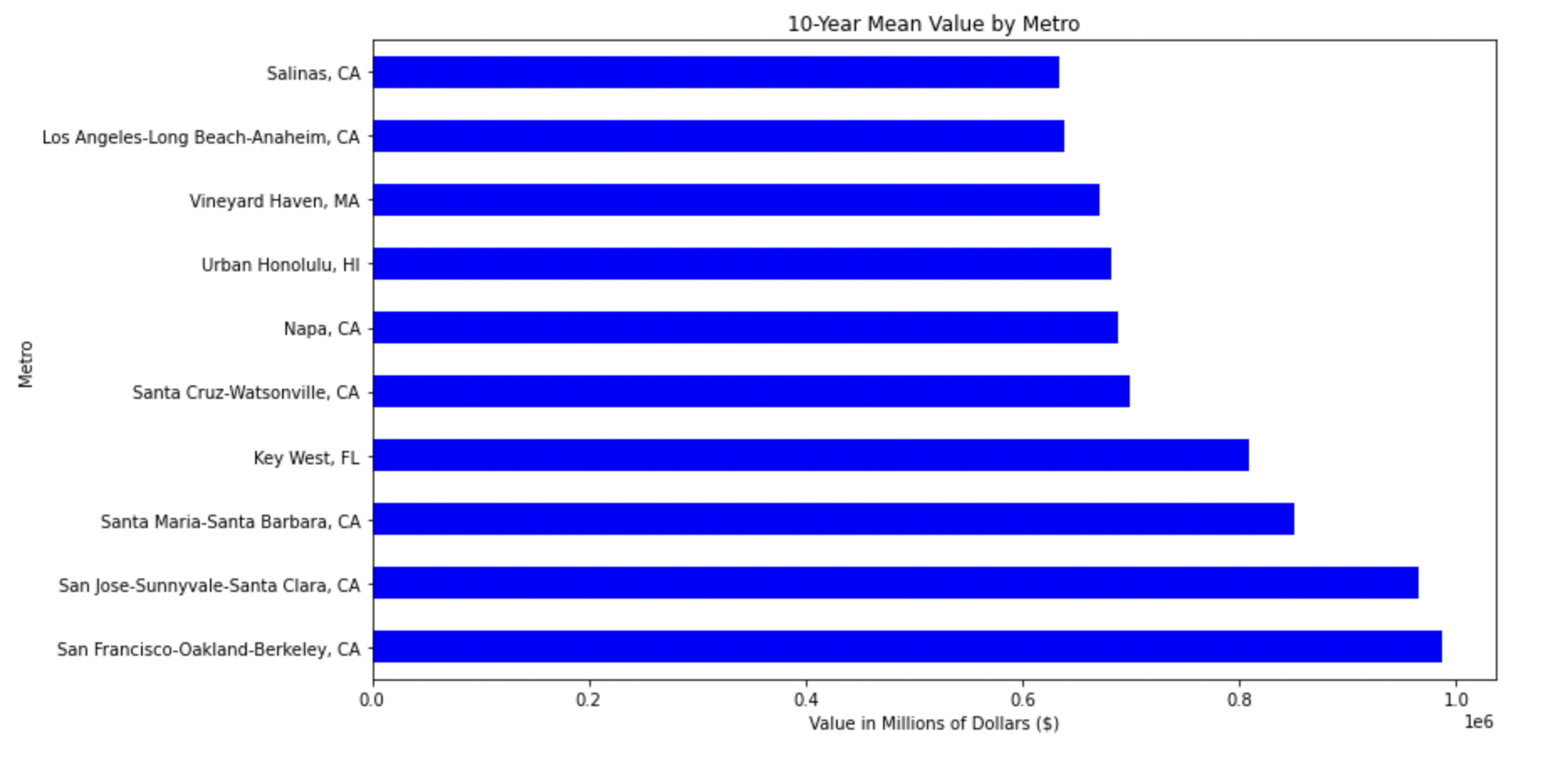


Fig 10:10-year mean home value by Metro

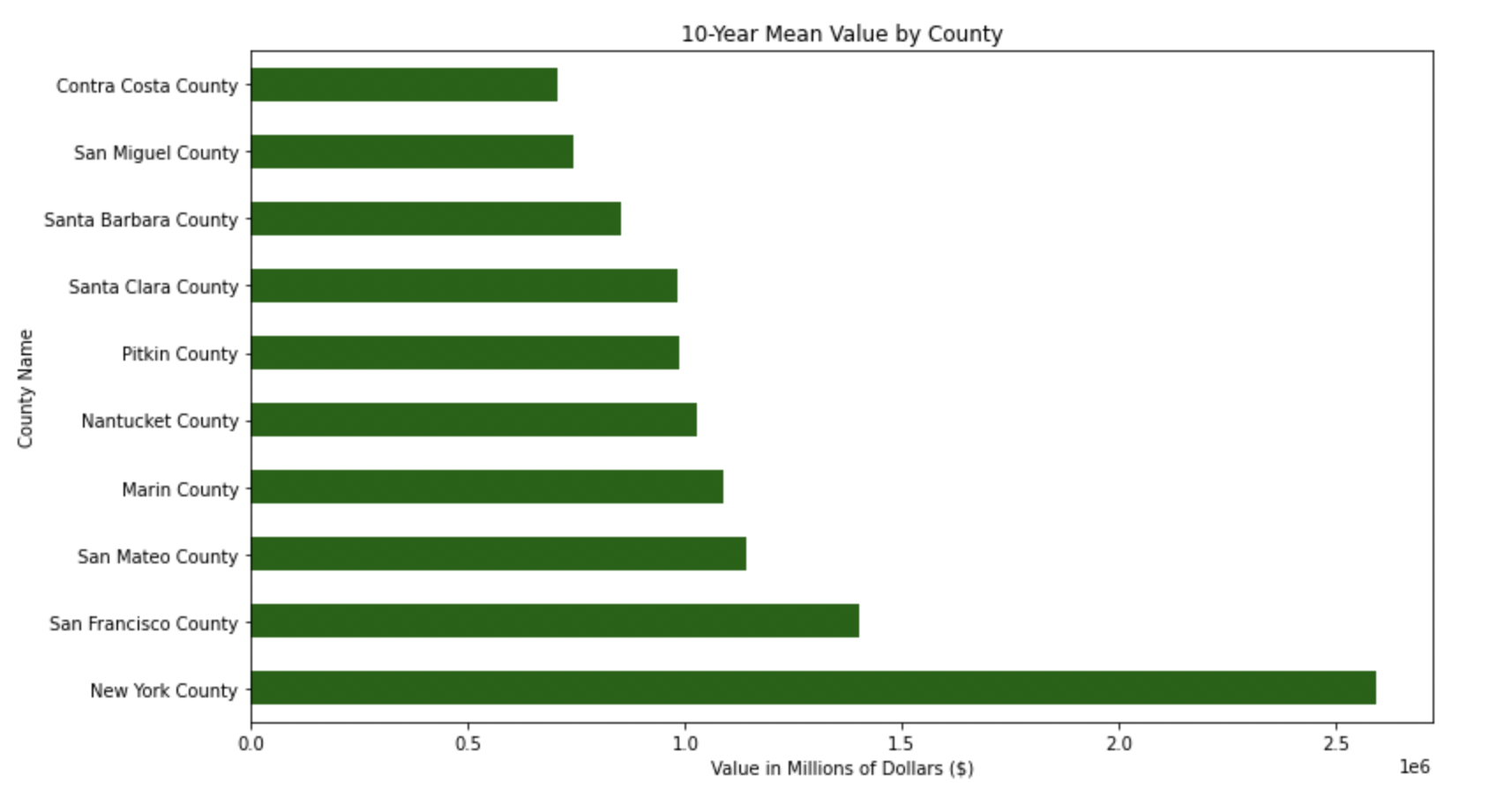


Fig 11:10-year mean home value by County

## 5.1 Golden Cluster

To find out the golden cluster we performed three iterations w.r.t the variable zipcode.

In the first trial, the clustering performance is silhouette score: 0.06 and sse within a cluster: 2594193.480335761 with a scatter plot

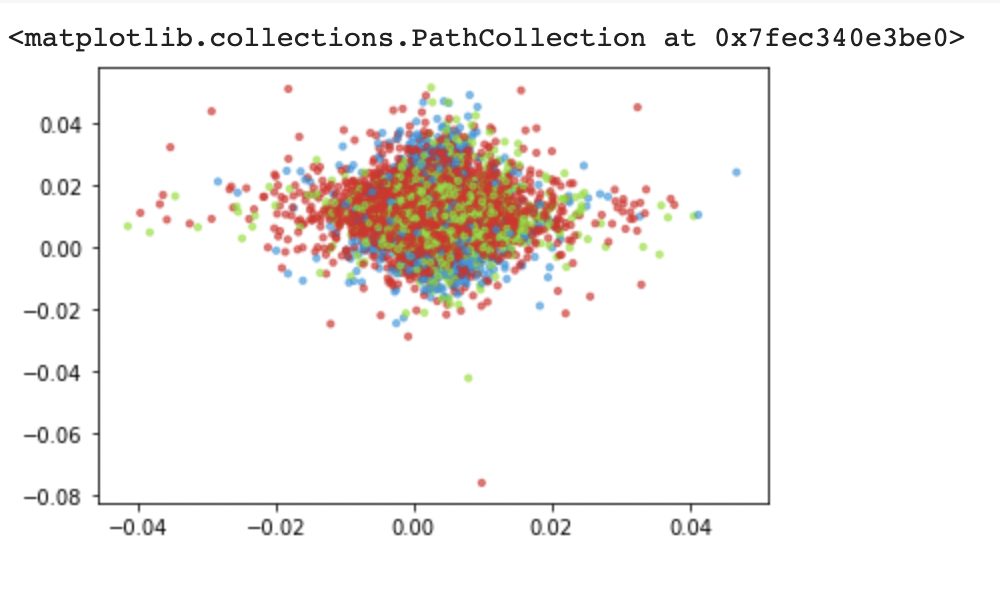
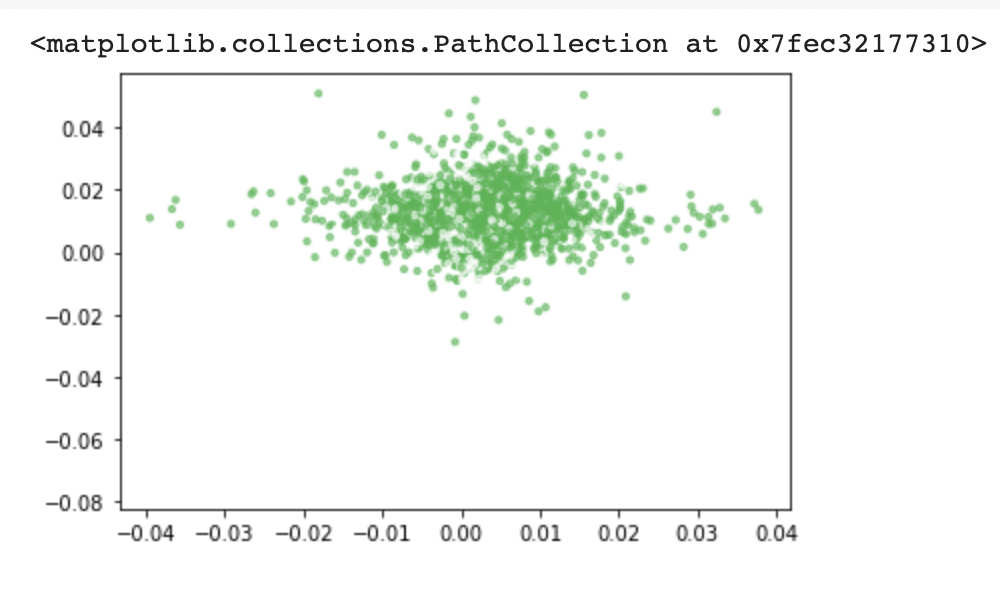


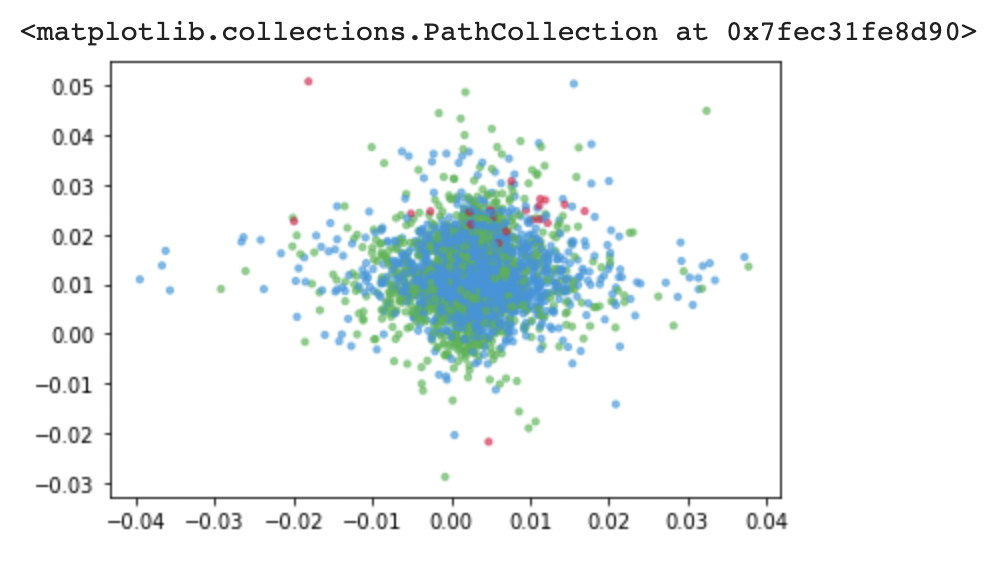
Fig 12 : First iteration

In the second trial, the clustering performance is silhouette score: 0.04 and sse within a cluster: 1926939.3814955656 with a scatter plot



## Fig 13: Second Iteration

In the third trial, the clustering performance is silhouette score: 0.02 and sse within the cluster: 1331352.0741808838 with a scatter plot



## Fig 14 : Third Iteration

On performing the third iteration we found that 2 clusters are almost shared and divided. Hence, one prominent cluster from iteration 2 itself will become our **Golden Cluster**.

## 6. Data Enrichment

## 6.1 Latent Variables and Manifolds

In a house price prediction project, latent features and manifolds can be used to represent and understand the complex relationships between the various features of a house and its price.

As mentioned in the assignment, after we profiled our 1st dataset which is a Zillow Home Value Index (ZHVI) time series dataset we added another dataset, EPA\_SmartLocationDatabase to add the latent variables and plot the results.

We have chosen our latent variables to be the National Walking Index and Mix of Employment types. We have done mapping the latent variables by defining a lv\_mapper() function followed by training and testing the data in 3 different ways i.e considering each latent variable separately and combining train and test data.

### **6.2 Muller Loop**

For the Muller loop, we defined a muller loop function with train and test datasets as the input parameters. We defined the muller loop function such that it considers different classifiers and calculates MSE, MAE, RSquared, and Test Accuracy which gives the best classifier with the best Test Accuracy score.

**6.2.1 Muller Loop Classification:**

Considering the metrics ​​MSE, MAE, RSquared, Test Accuracy, Recall, Precision, and iterating through different classifiers Nearest Neighbors, Linear SVM, RBF SVM, Decision Tree, Random Forest, Neural Net, AdaBoost, and Naive Bayes we found that **K-Nearest Neighbours algorithm** was the best performing algorithm with a Score (test, accuracy) of 57.84. So we use KNN to classify the data.

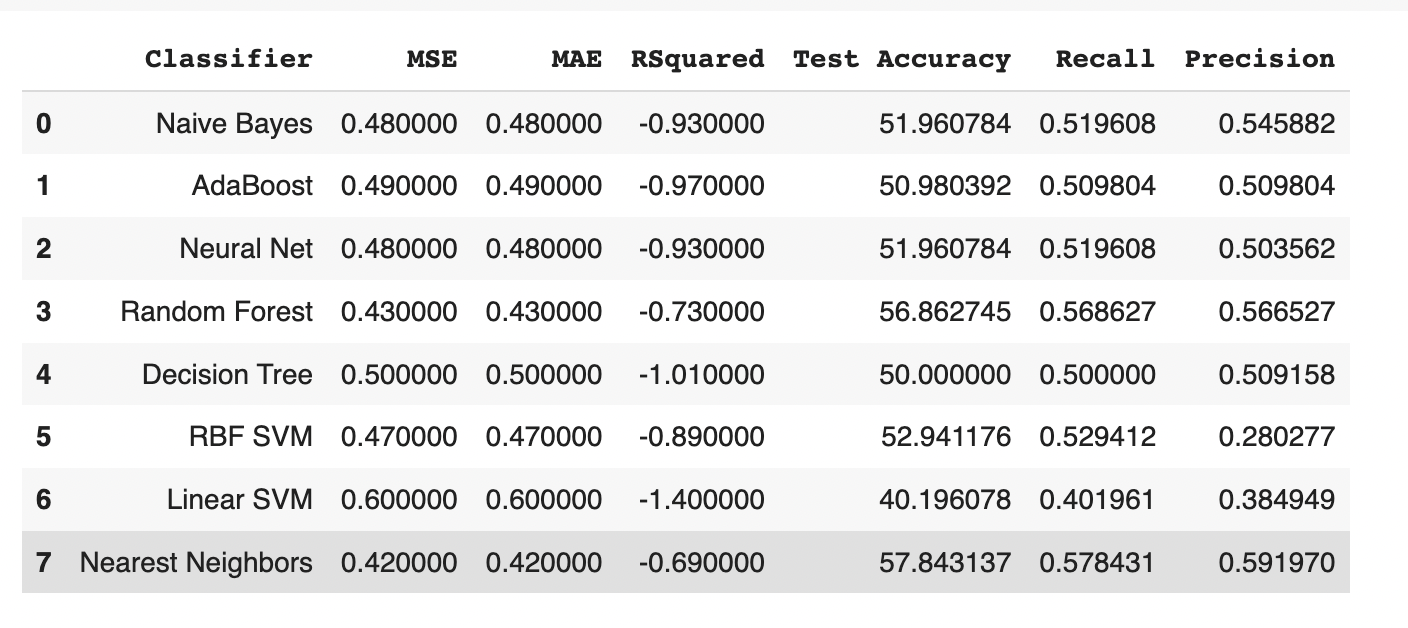


Fig. 15. Muller Loop Classification

**6.2.2 Muller Loop Regression :**

Similar to classification calculating the metrics MSE(MeanSquaredError), MAE(MeanAbsoluteError), RSquared, and Test Accuracy by iterating the train and test data through the regressors Linear Regression, MLP Regressor, RandomForest Regressor, Gradient Boosting Regressor, KNeighbors Regressor we found that RandomForest is the best regressor, with a Score (test, accuracy) of 3.77

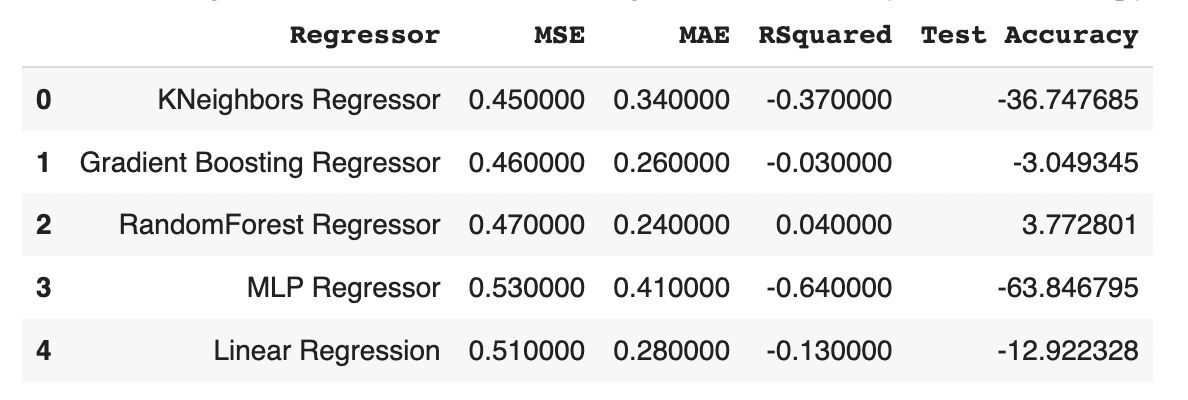


Fig. 16. Muller Loop Regression

**6.3 Confusion Matrix and Metrics**

Plotted a confusion matrix by defining a function confMatrix() taking y\_test and y\_pred as input parameters returning a confusion matrix of the above inputs which is further taken as an input to a plotting function plot\_cm() and gave the following matrix plot

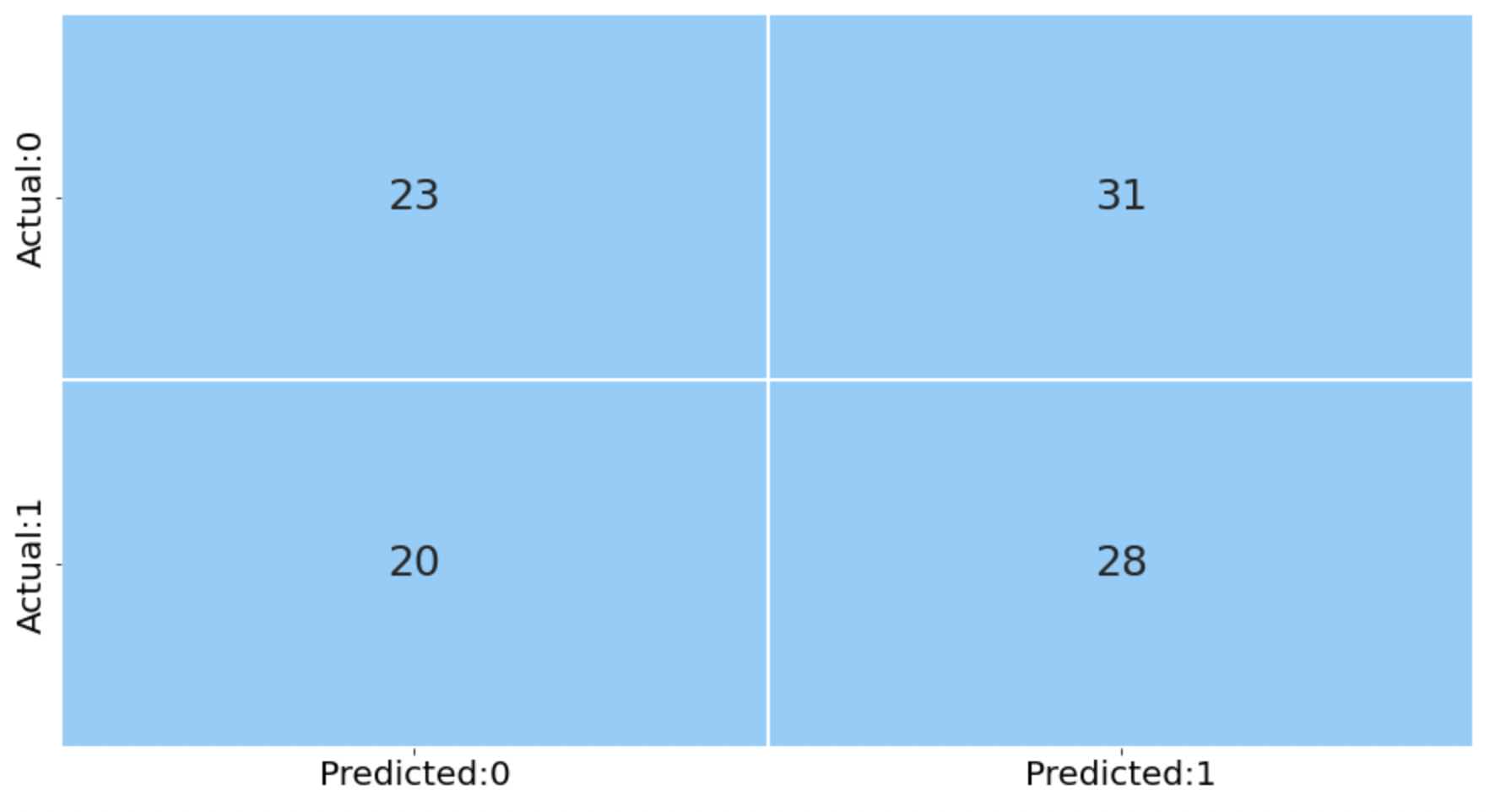


Fig 17 : Confusion matrix

Later defined a new function calculates metrics() taking confusion matrix,y\_test and y\_pred as the inputs to calculate the metrics given by

precision = TP / (TP+FP)

recall = TP / (TP+FN)

specificity = TN / (TN+FP)

f1\_score = 2\*((precision\*recall)/(precision+recall))

accuracy = (TN+TP) / (TN+FP+FN+TP)

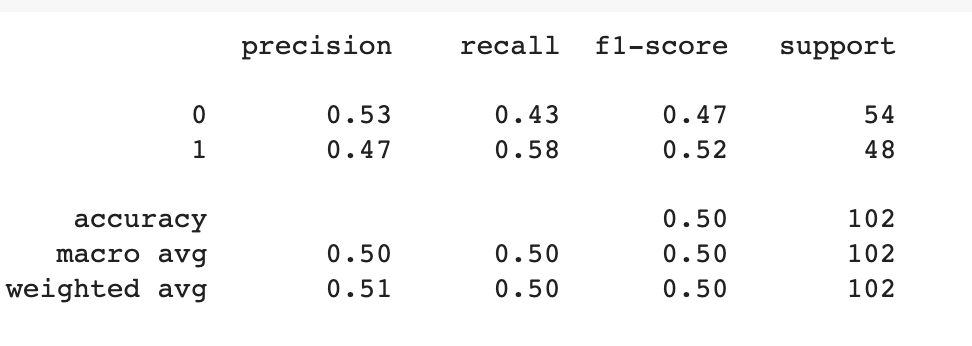
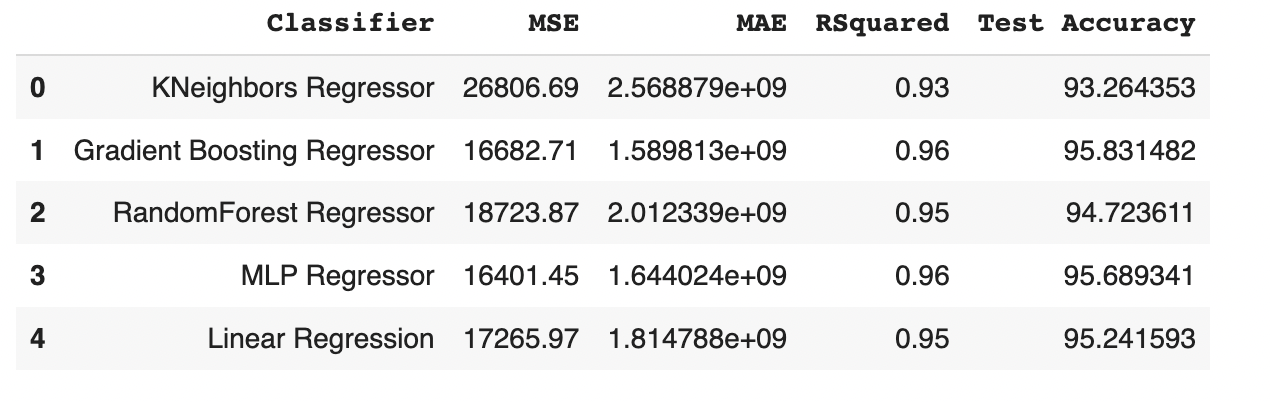


Fig 18 : Performance Measures

We applied the Muller loop function on the dataset considering the conditions without any latent variables, with each latent variable, and on the combination of the latent variables.

We found that the one without latent variables gives Gradient Boosting Regressor with Score (test, accuracy) of 95.83 as its best classifier.



The best classifier for the case With Latent Variables National Walking Index and Mix of Employment Types separately is MLP Regressor with scores (test, accuracy) of 95.92 and 95.94 respectively.

### **6.4 With Latent Variable 1: National Walking Index**

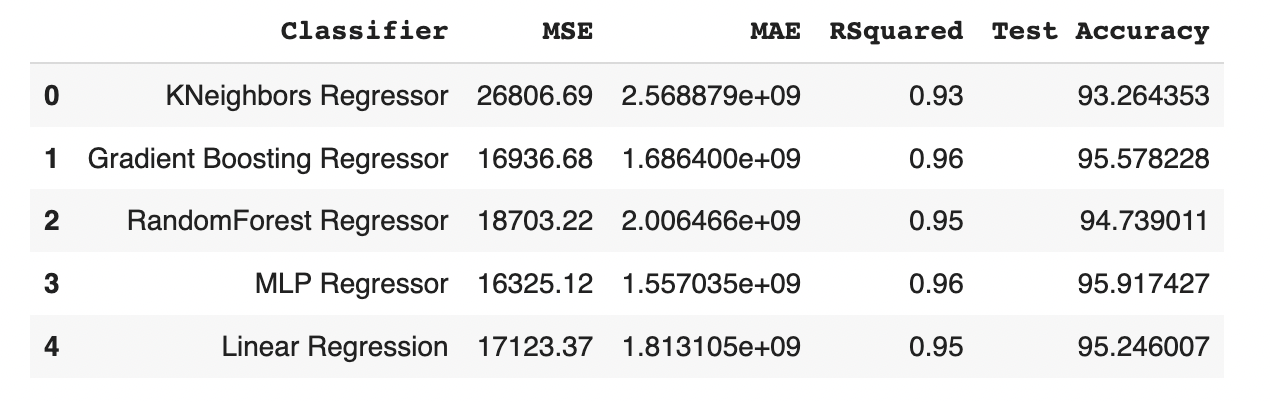


Fig 20: Muller Loop Regressor with National Walking Index

### **6.5 With Latent Variable 2: Mix of Employment Types**

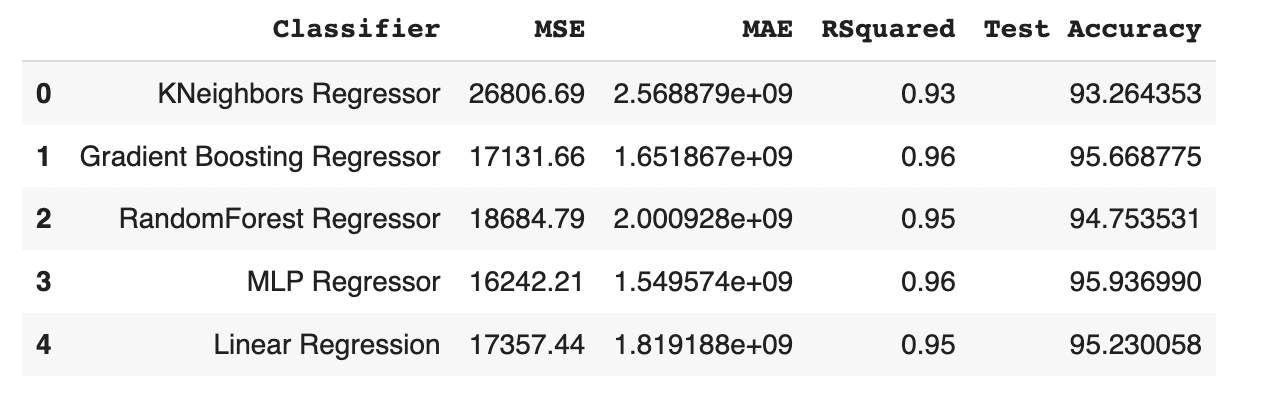


Fig 21: Muller Loop Regressor with Mix of Employment Types

**6.6 With Latent Variables 1 and 2: National Walking Index and Mix of Employment Types**

The best classifier for the case which is a combination of both the Latent Variables National Walking Index + Mix of Employment Types is the Gradient Boosting Regressor with a Score (test, accuracy) of 95.71

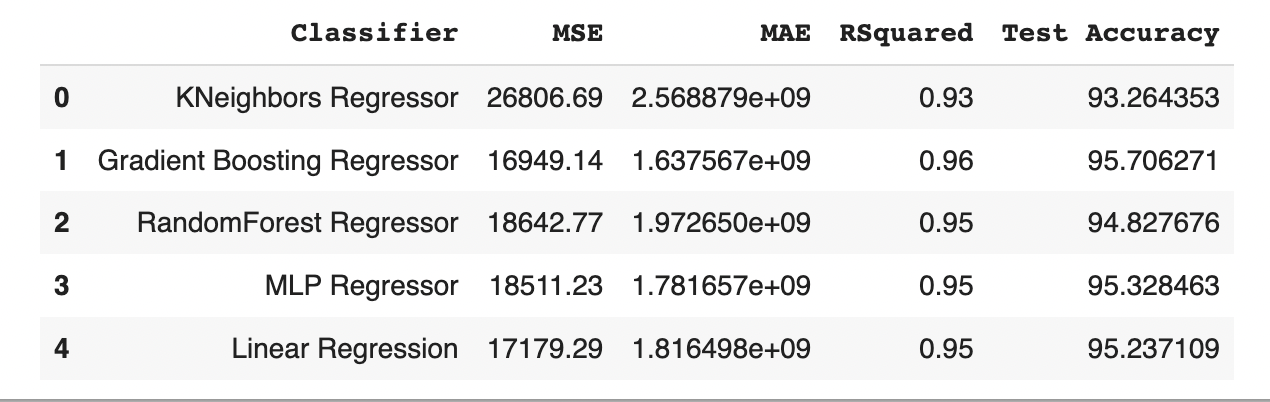
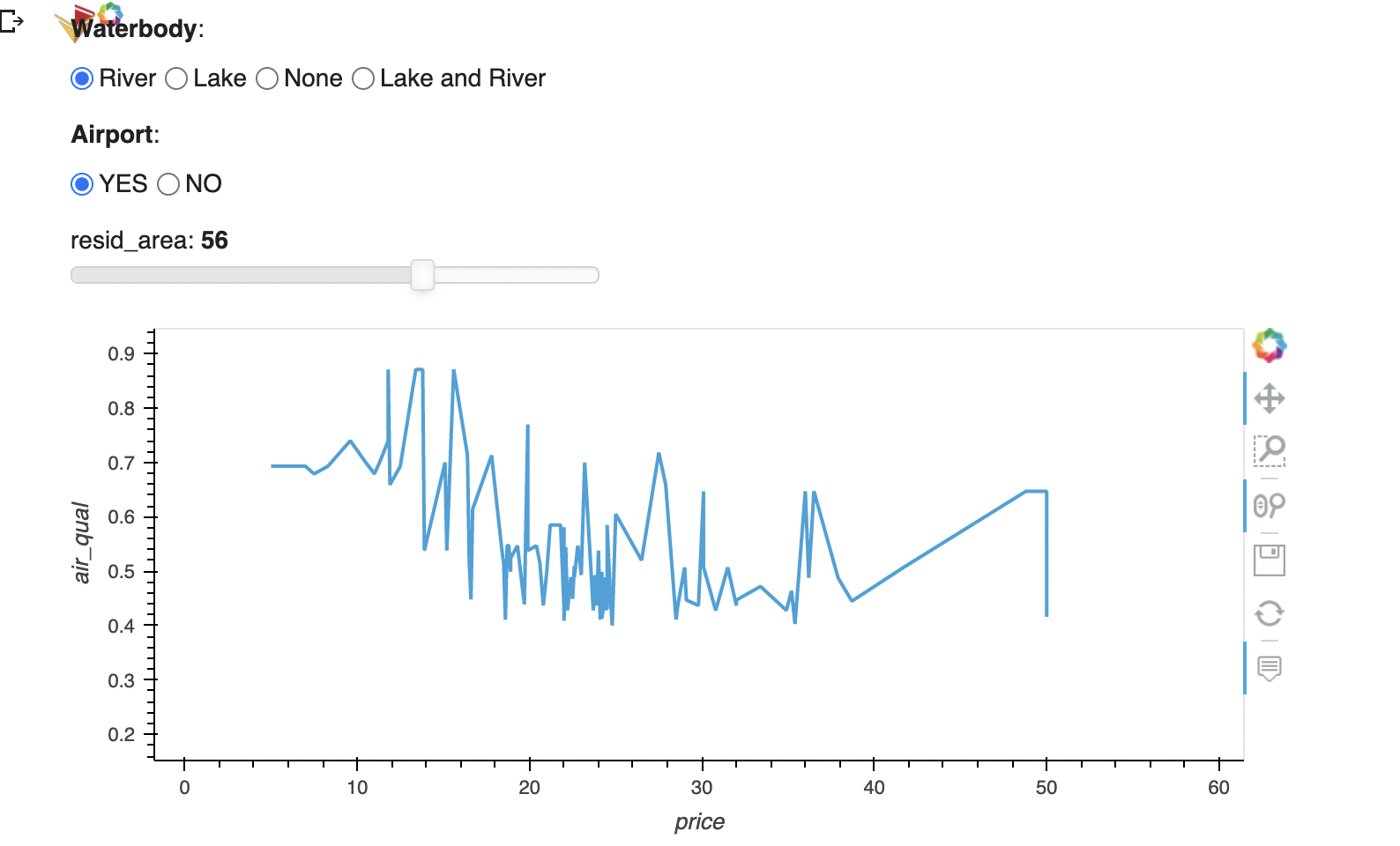


Fig 22: Muller Loop Regressor with National Walking Index and Mix of Employment Types

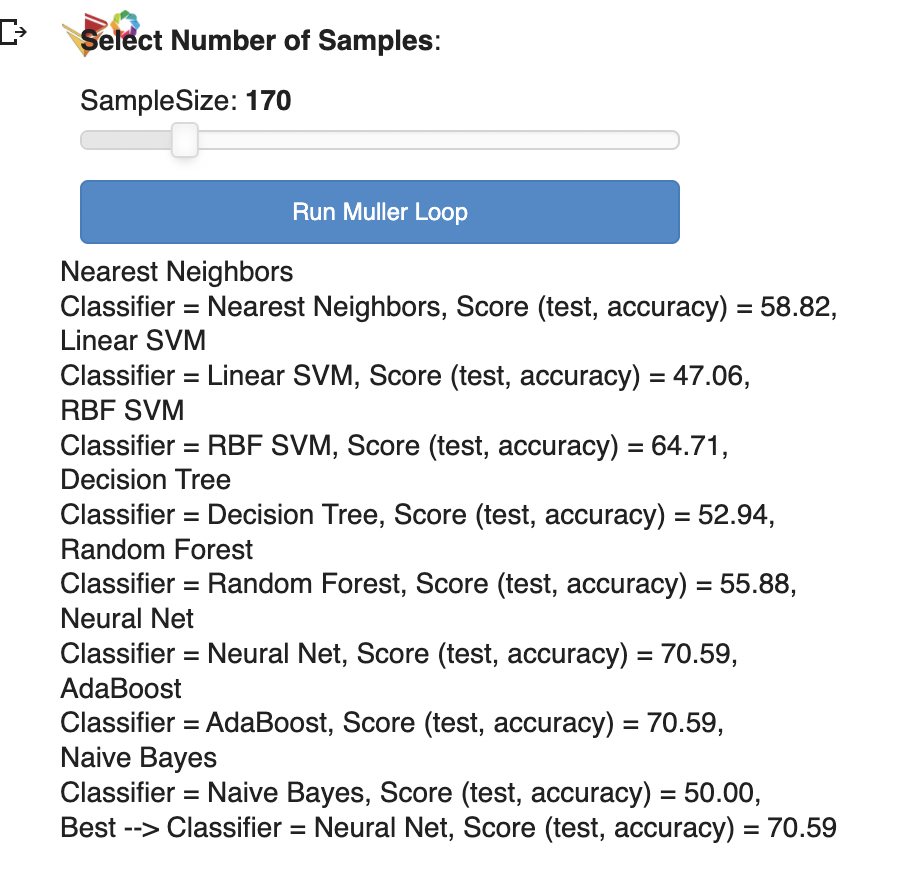
# Distributions of your Data

We distributed the data and visualized them using Holoviews importing the holoview extension ‘bokeh’.

Considering the hiver options to be Water body and Airport which had the options river, lake, none, river and lake and yes/no respectively. We also took the resid\_area constraint on a slider altogether plotting a graph between air quality and price.



Also did an interactive holoview plot for running a muller loop depending on the number of samples



# 7 Conclusions

After the third data enrichment the accuracy of prediction improves from 60% to 66% and to 94% but using different regression (Random Forest – First latent variable, MLP – Second Latent Variable).

We also did the other modeling using Muller Loop Regression for identifying the best model. We found that Random Forest is the best model regression for our dataset.

## Appendices

Appendix A

**Final Colab Link**

**https://colab.research.google.com/drive/1VXYi8FORM-hn7JTuS8DdT1Xsqj0Lbt7B#scrollTo=ye0VJYiW9FgV**

**Github Link**

**https://github.com/sidsanc/257\_MachineLearning/blob/main/Project/MIDTERM\_NUMERIC\_PROJECT.ipynb**

**Document Final Report**

[**https://drive.google.com/file/d/1QUtcbq75nYwTmu7vVsLkYQ6AKeHK2OX\_/view?usp=sharing**](https://drive.google.com/file/d/1QUtcbq75nYwTmu7vVsLkYQ6AKeHK2OX_/view?usp=sharing)

**Data Links:**

DataSet 1

House\_pricing.csv

Drive Link: [**https://drive.google.com/file/d/1lKFZEqth\_T9wFHN2XTxOhSRwFek1hi9Z/view**](https://drive.google.com/file/d/1lKFZEqth_T9wFHN2XTxOhSRwFek1hi9Z/view)

Dataset 2

Home\_sales-formatdate.csv

Drive Link:

[**https://drive.google.com/file/d/1KAvZblGU7rVfbhSM8S-RvD3jlsucIPyB/view**](https://drive.google.com/file/d/1KAvZblGU7rVfbhSM8S-RvD3jlsucIPyB/view)

Dataset 3

Drive Link:

[**https://drive.google.com/file/d/1HfTP590blvZIYS3MhlQyZK3or4lgIGaV/view**](https://drive.google.com/file/d/1HfTP590blvZIYS3MhlQyZK3or4lgIGaV/view)

Dataset 4

Drive Link: [**https://drive.google.com/file/d/1p4BNeW3kp7-Ruiukio3q25O\_SA2pICCY/view**](https://drive.google.com/file/d/1p4BNeW3kp7-Ruiukio3q25O_SA2pICCY/view)

Dataset 5

EPA\_SmartLocationDatabase\_V3\_Jan\_2021\_Final.csv

Drive Link: [**https://drive.google.com/file/d/1yCz6FYet7\_dwq4sHQxc67xcujvQbxTuU/view**](https://drive.google.com/file/d/1yCz6FYet7_dwq4sHQxc67xcujvQbxTuU/view)

Dataset 6

Drive link: [**https://drive.google.com/file/d/1paSdcAVmtMVaT4kmYVcuVUInfEISik\_W/view**](https://drive.google.com/file/d/1paSdcAVmtMVaT4kmYVcuVUInfEISik_W/view)

Colab Link:

[**https://colab.research.google.com/drive/1VXYi8FORM-hn7JTuS8DdT1Xsqj0Lbt7B#scrollTo=ye0VJYiW9FgV**](https://colab.research.google.com/drive/1VXYi8FORM-hn7JTuS8DdT1Xsqj0Lbt7B#scrollTo=ye0VJYiW9FgV)

## Acknowledgments

We would like to acknowledge our Professor Dr. Ali Arsanjani for guidance throughout this Spring-2023 semester.

## References

1. EPA USA

https://www.epa.gov/smartgrowth/smart-location-mapping

1. Inspection Support

https://www.inspectionsupport.com/resources/cities-with-the-worst-hoa-fees/