# Post Graduate Program - Data Science

## In Partnership With Purdue University

## Course Project - Data Science with Python

## California Housing Price Prediction

Submitted by:

Lavkush Singh

Submitted to:

Purdue University – Simplilearn

# Agenda

- Introduction
- Dataset Summary
- Exploratory Data Analysis – Column Variables
- Exploratory Data Analysis – Columns w.r.t New Feature
- Correlation Matrix
- Predictive Model Analysis
- Predictive Model Summary
- Appendix

# Introduction

- California Census Data published by US Census Bureau
- Dataset has 10 types of metrics such as the population, median income, median housing price, and so on for each block group in California.
- There are 20,640 districts in the project dataset.
- This is **Supervised – Regression** Problem
- This project aims at building a model of housing prices to predict median house values in California using the provided dataset.
- Linear Regression is used to report the results. However, Random Forest Regressor is also used for comparison purpose.

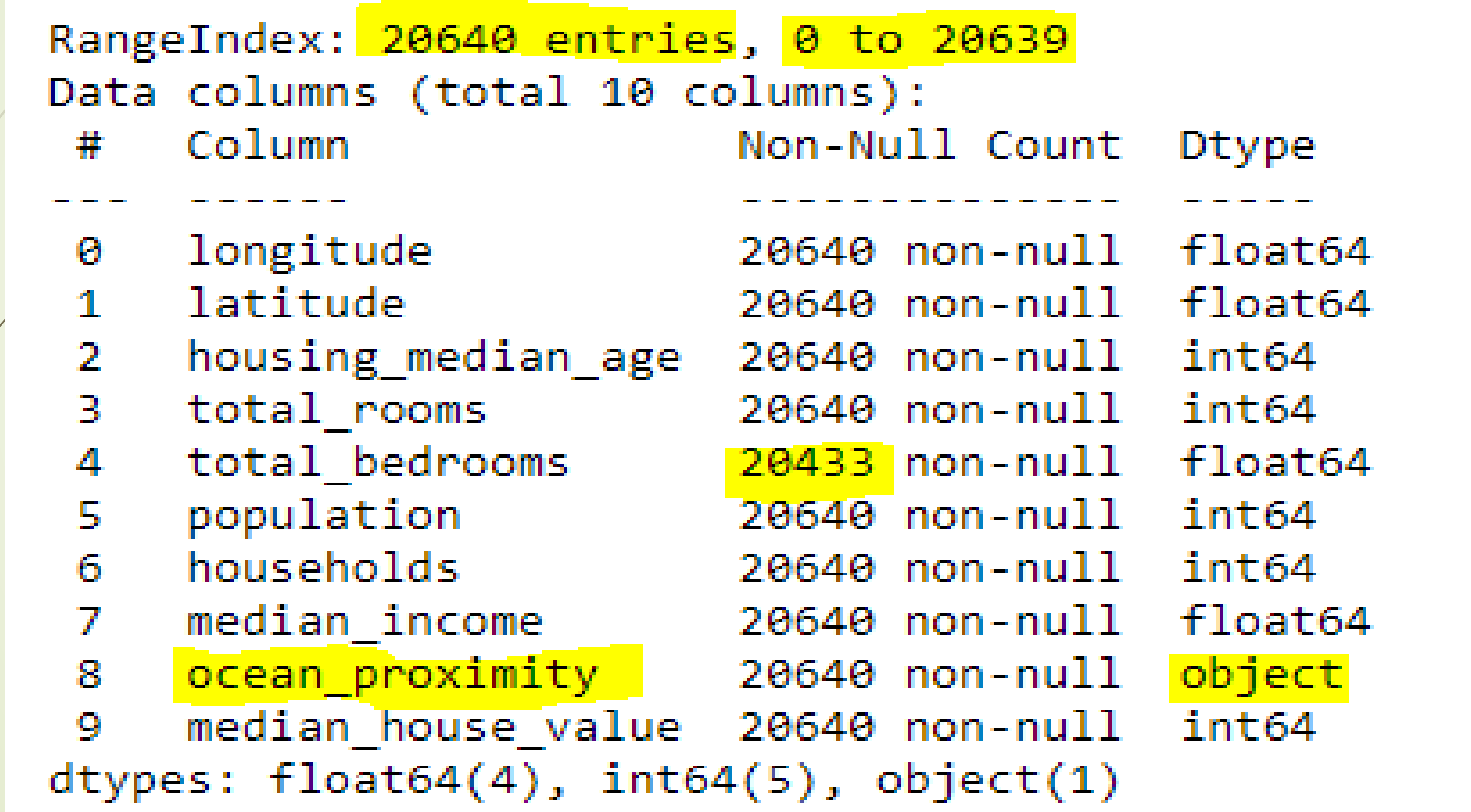
# Dataset Summary

- 20640 observations (rows) of  10 variables (columns)
- ‘total_bedrooms’ column had 207 missing values, median value is imputed
- All the columns were numerical except ‘ocean_proximity’ column which had 5 categories.
- One hot encoding was used to pre-process ‘ocean_proximity’ column
- Numerical column values varies in scale and range
- Data has skewed distributions. It is visually presented in the code file, however it is not accounted because the tasks given were pre-defined, and was not a part of ‘Analysis Tasks to be performed’

# Exploratory Data Analysis

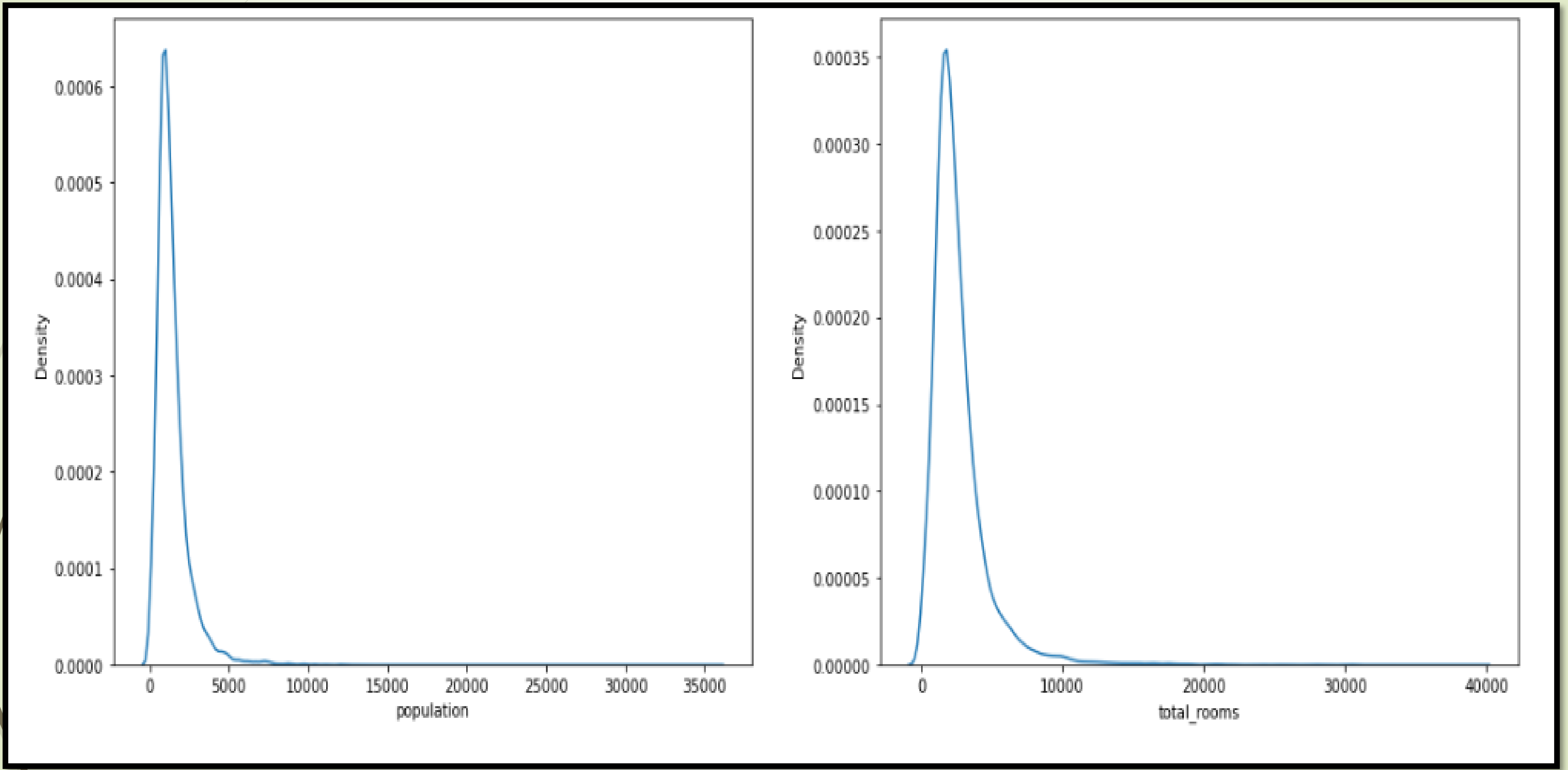
## Column Variables

# Understanding Data – Datatypes, Dimension, Null Values Summary

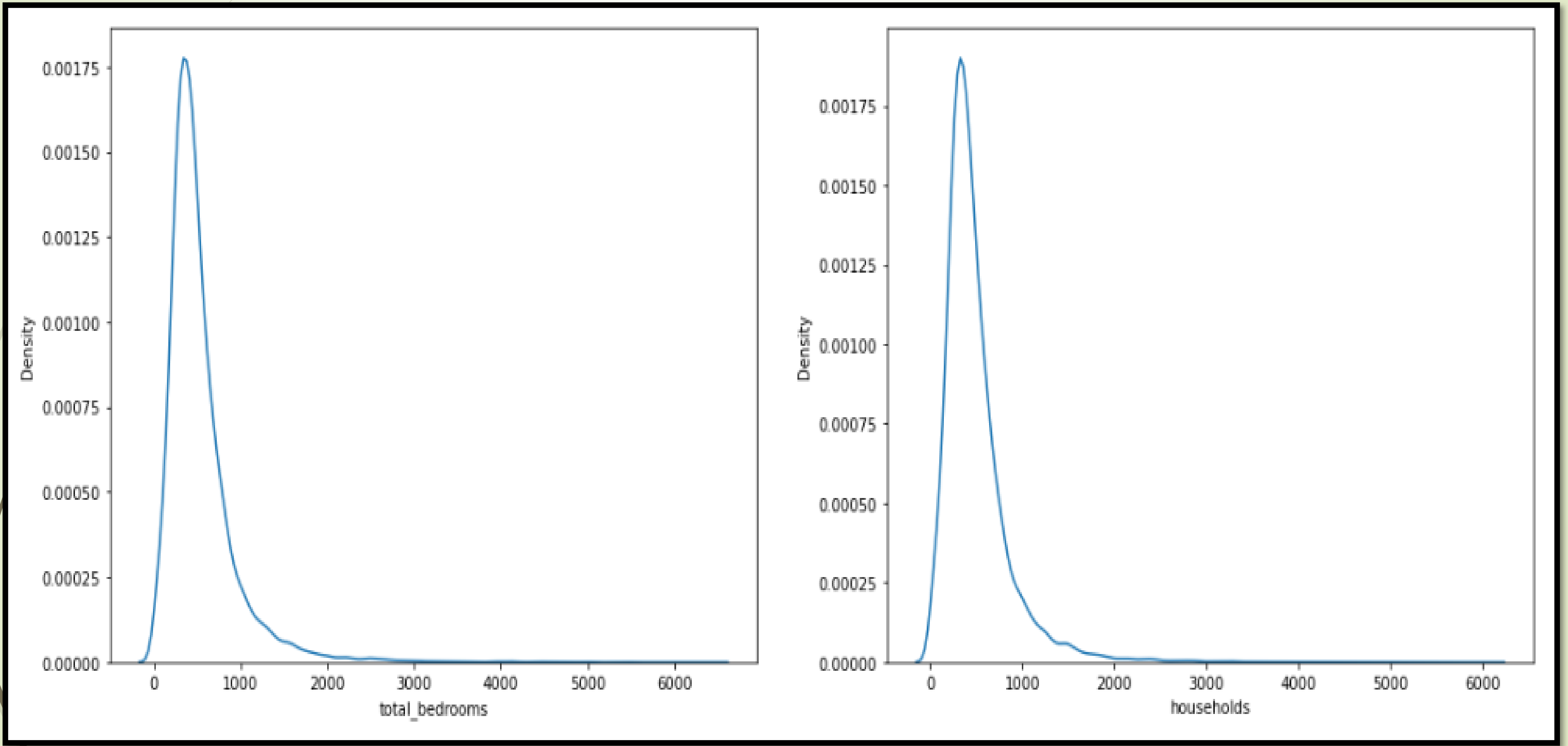


```
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  int64  
 3   total_rooms         20640 non-null  int64  
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  int64  
 6   households          20640 non-null  int64  
 7   median_income       20640 non-null  float64
 8   ocean_proximity     20640 non-null  object 
 9   median_house_value  20640 non-null  int64  
dtypes: float64(4), int64(5), object(1)
```

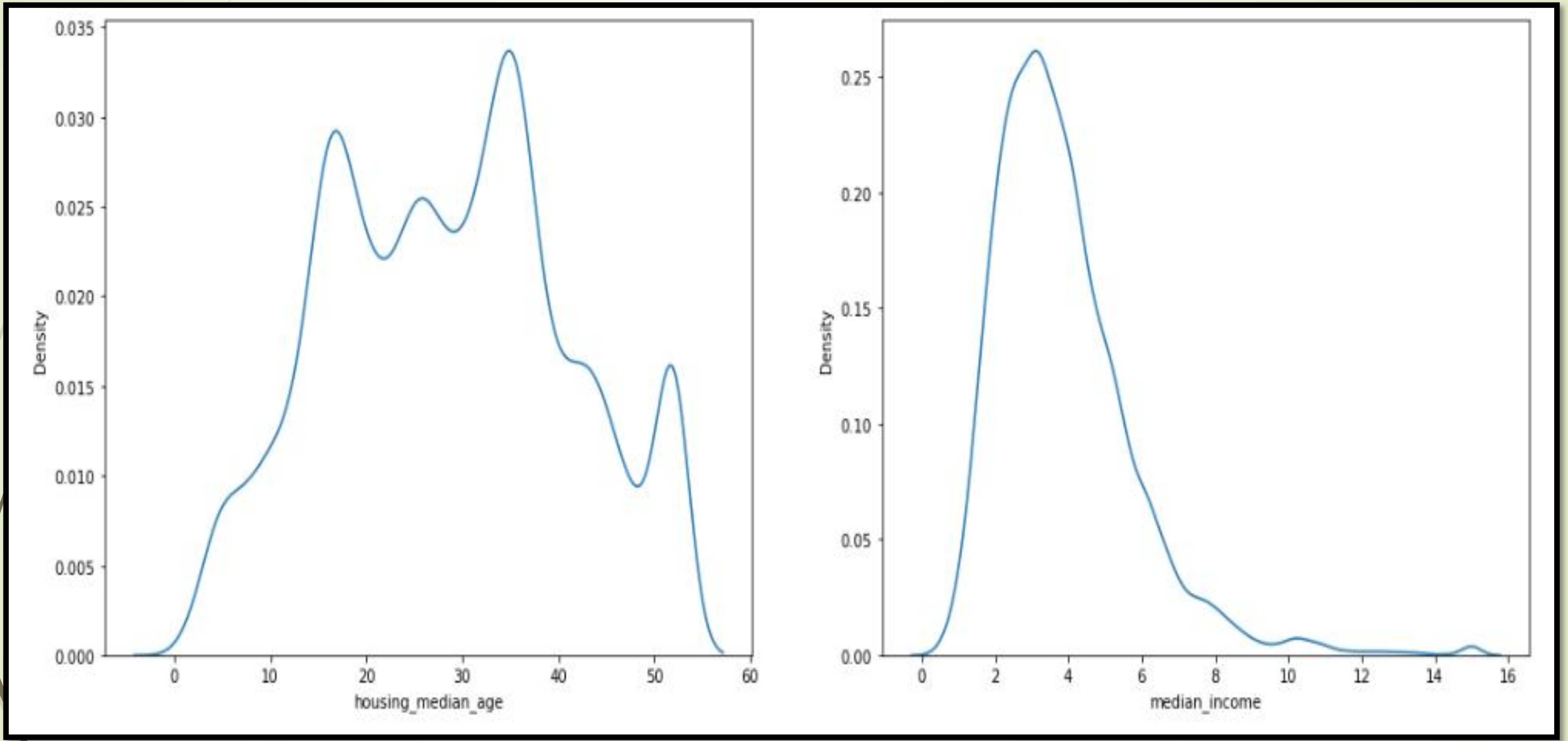
# Distribution – ‘population’, ‘total_rooms’



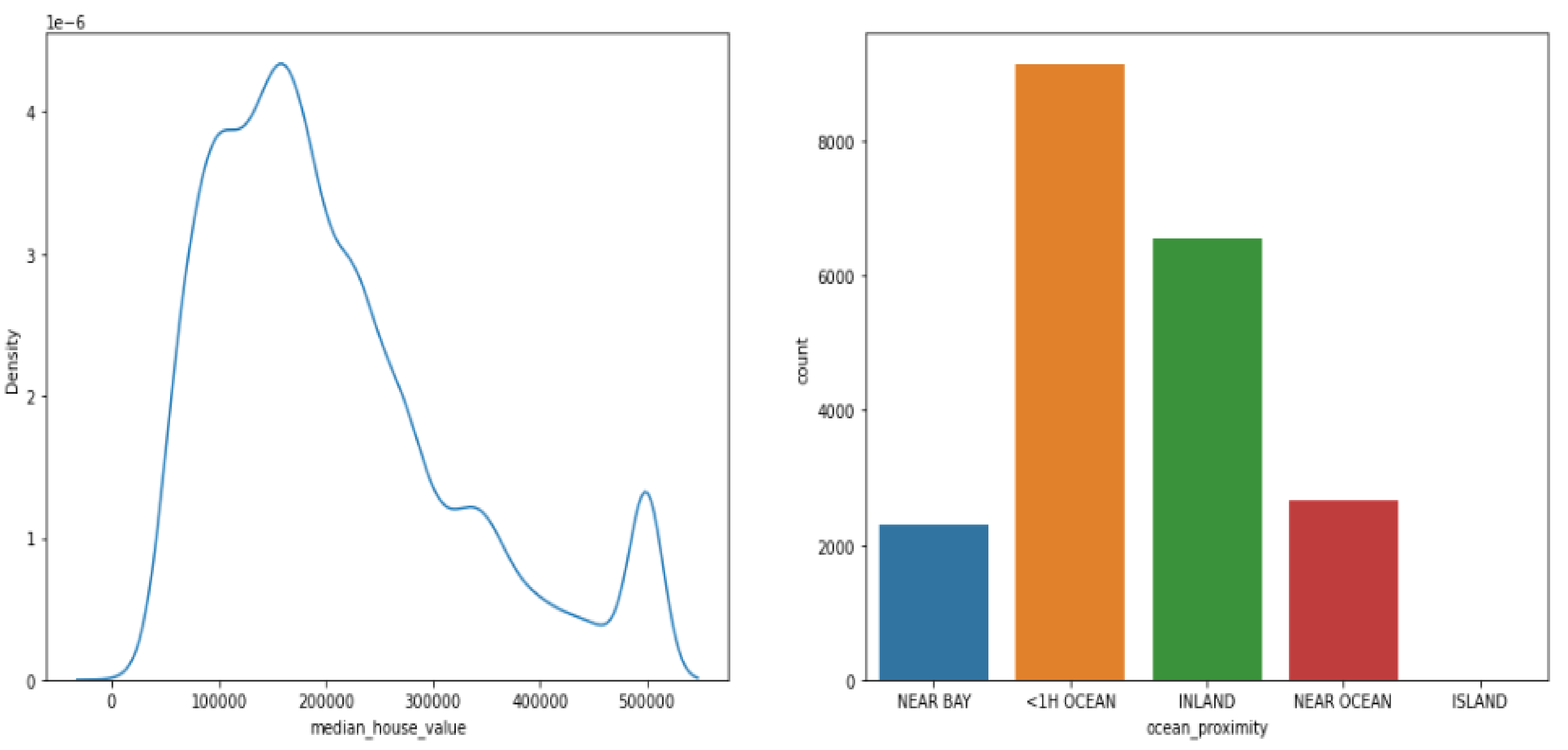
# Distribution – ‘total_bedrooms’, ‘households’



# Distribution – ‘housing_median_age’, ‘median_income’



# Distribution – ‘median_house_value’, ‘ocean_proximity’

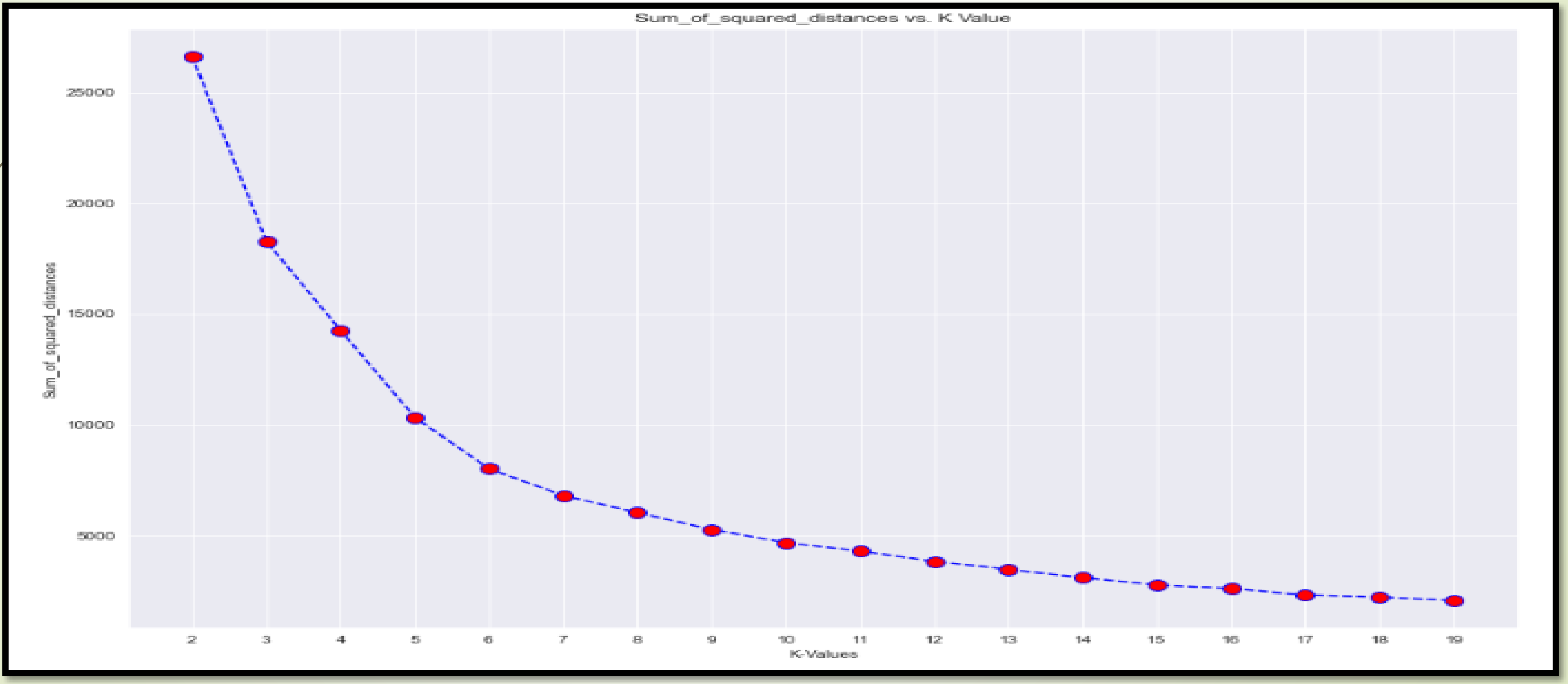


# Exploratory Analysis Summary

- We have 10 columns and 20640 entries in dataset
- ‘total_bedrooms’ column has total of 207 missing observations
- ‘latitude’ and ‘longitude’ are spatial data columns
- Columns – ['population', 'total_rooms', 'total_bedrooms', 'households’, ‘median_income’] has similar distribution – Right Tailed or Right Skewed Distribution
- Columns - [‘housing_median_age’, ‘median_house_value’] has uneven distribution.
- Column ‘ocean_proximity’ is categorical – of which houses near ‘<1H Ocean’ are highest and houses near ‘Island’ are lowest.

# Exploratory Analysis Summary

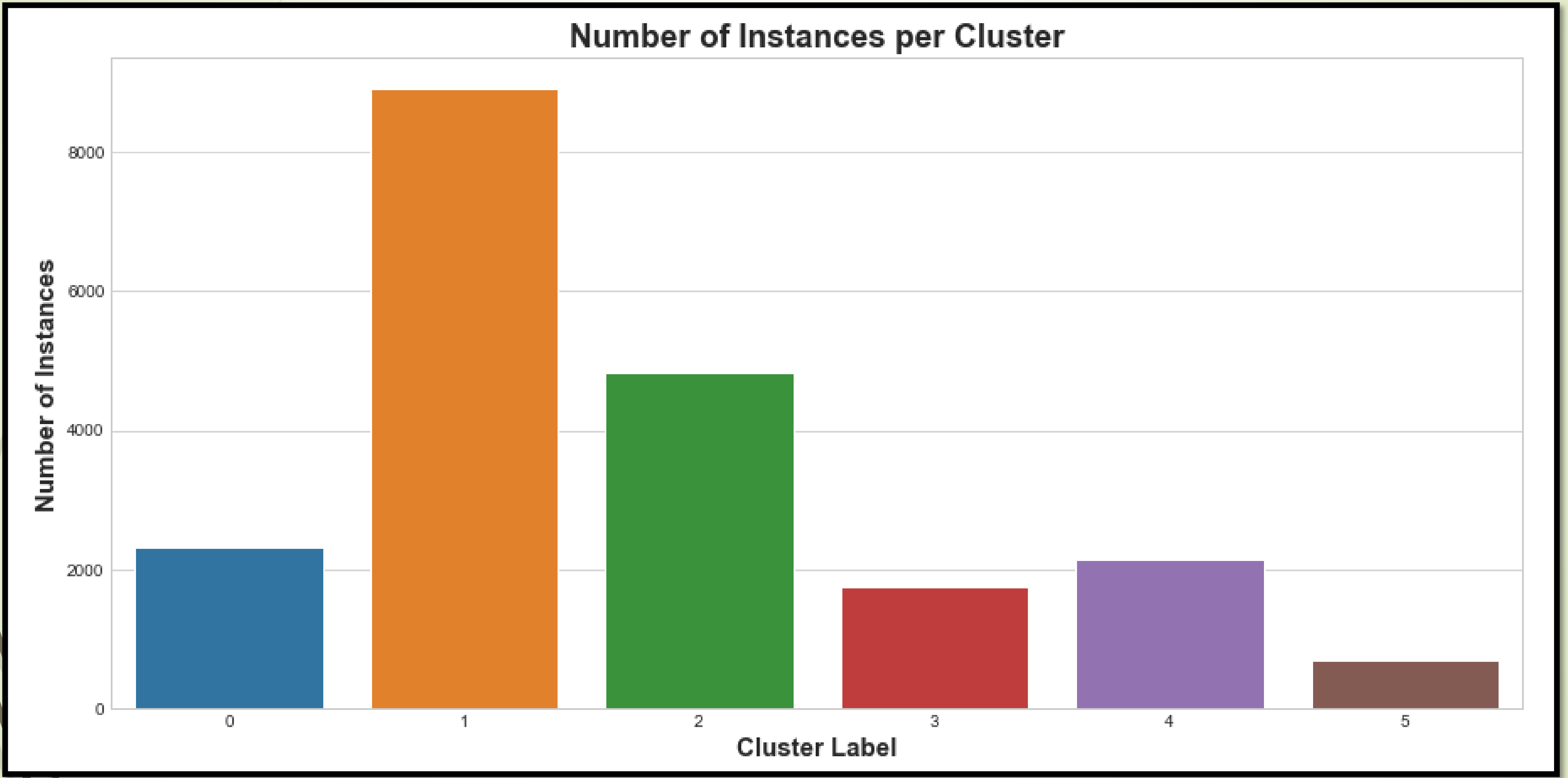
- ‘latitude’ and ‘longitude’ columns has been converted to categorical column using KMeans Clustering Algorithm with K = 6 based on the below elbow method plot. (Refer code file submitted)



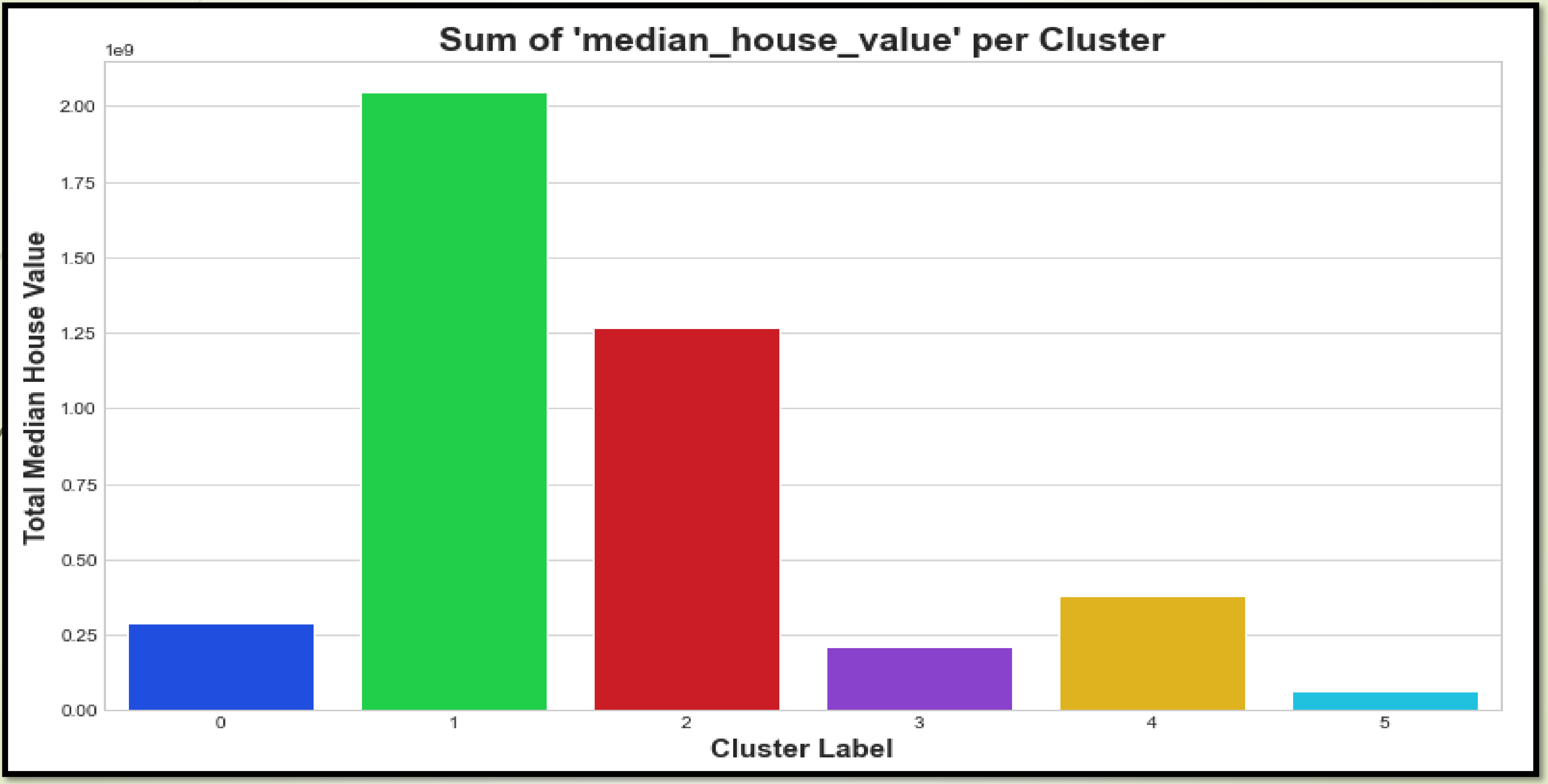
# Exploratory Data Analysis

## Features based on Latitude-Longitude Clusters

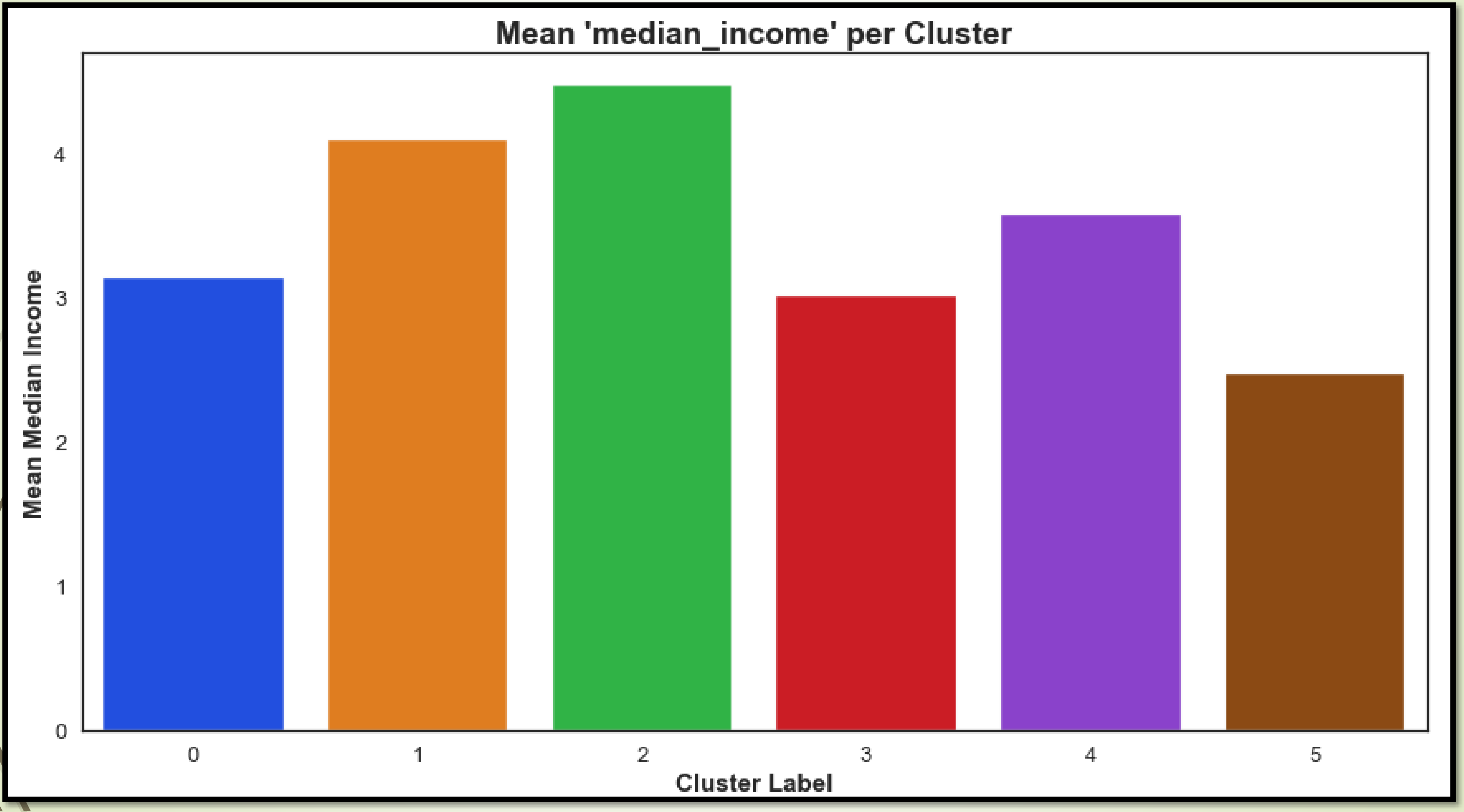
# Exploratory Data Analysis



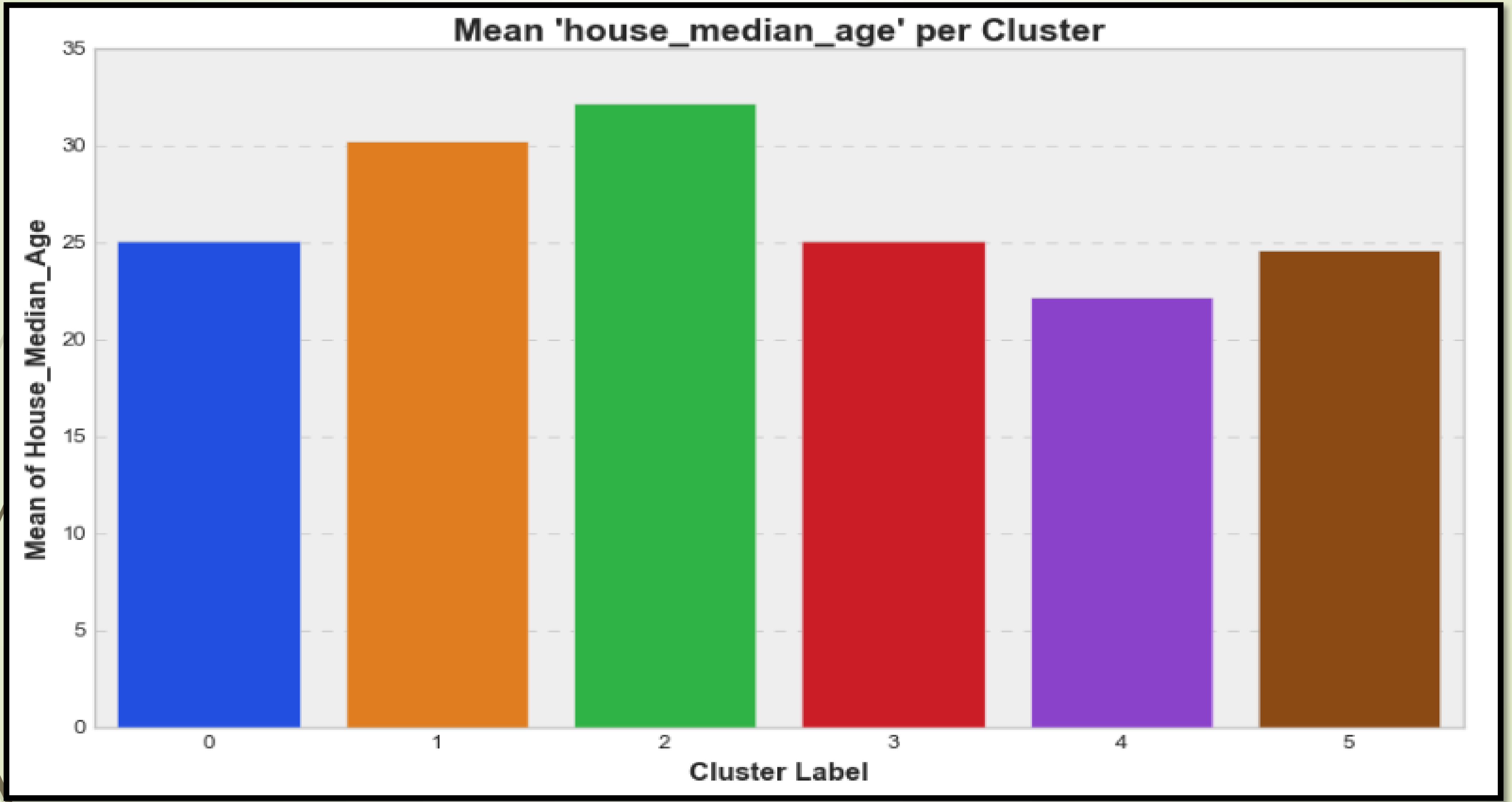
# Exploratory Data Analysis



# Exploratory Data Analysis



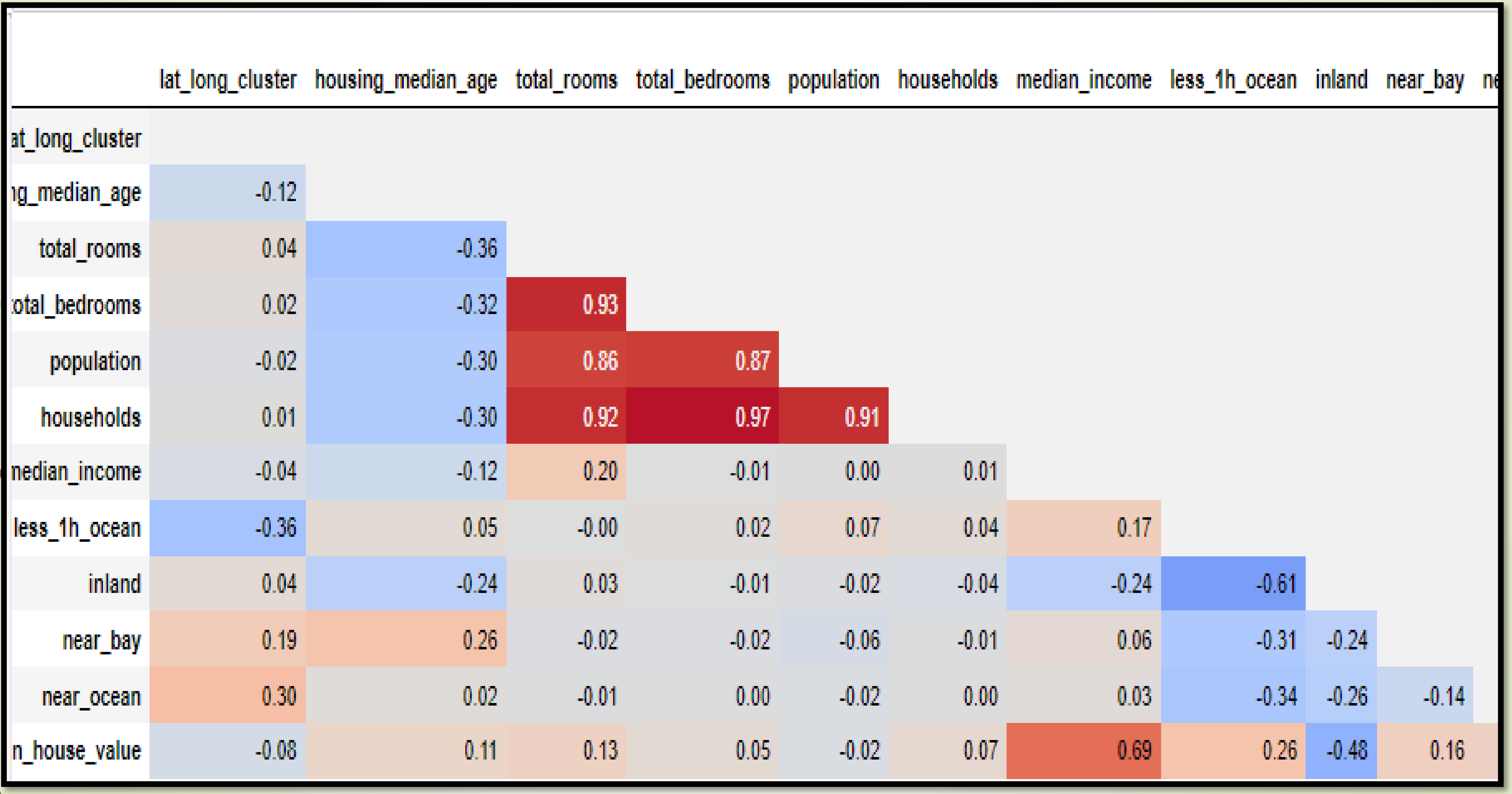
# Exploratory Data Analysis



# Exploratory Data Analysis Insights

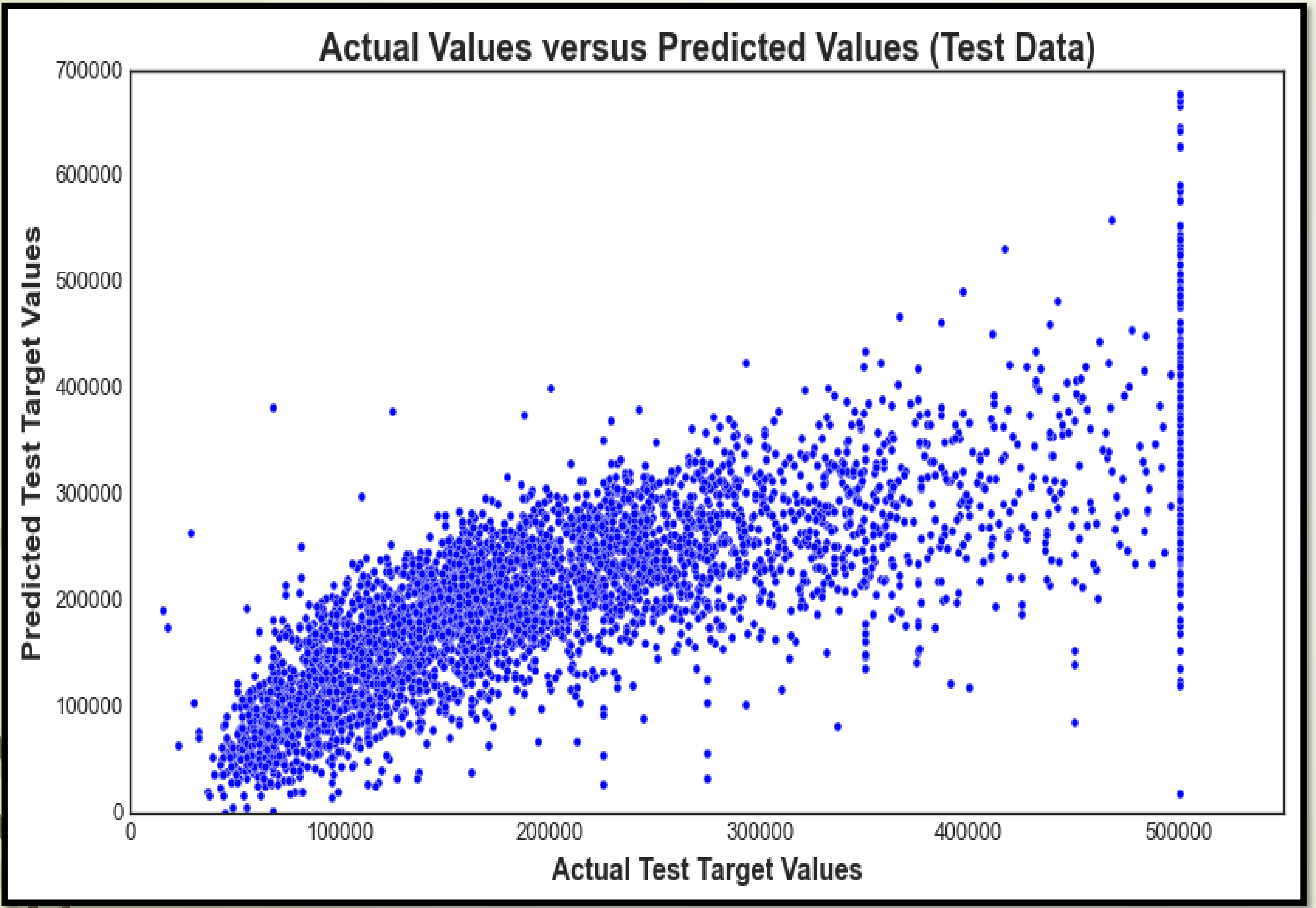
- Latitude and Longitude data has been converted to six clusters (0 to 5), based on KMeans Clustering Algorithm
- Cluster 1 has highest number of instances, where as cluster 5 has lowest
- Similar trend follows for sum of ‘median_house_value’. It means that the median house values (total) is more for cluster 1 and low for cluster 5
- Population living in cluster 2 are rich, meaning their mean income is highest, and lowest is for people living in cluster 5
- Oldest house of the city are located in cluster 1 and 2, where as newer houses have been constructed in rest of the clusters

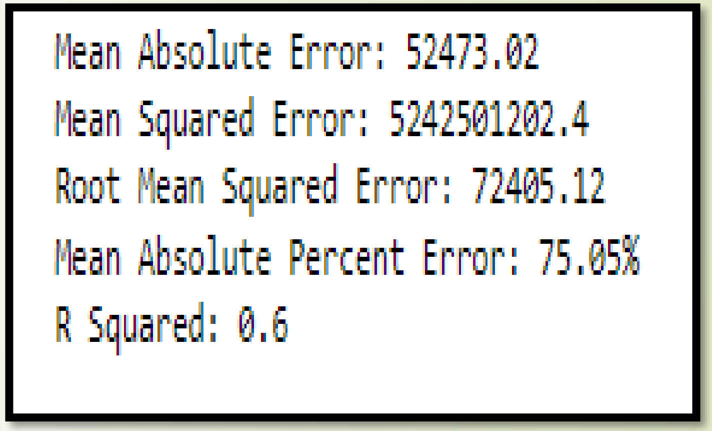
# Correlation Matrix



| | lat_long_cluster | housing_median_age | total_rooms | total_bedrooms | population | households | median_income | less_1h_ocean | inland | near_bay | ne |
|---|---|---|---|---|---|---|---|---|---|---|---|
| at_long_cluster | | | | | | | | | | | |
| g_median_age | -0.12 | | | | | | | | | | |
| total_rooms | 0.04 | -0.36 | | | | | | | | | |
| otal_bedrooms | 0.02 | -0.32 | 0.93 | | | | | | | | |
| population | -0.02 | -0.30 | 0.86 | 0.87 | | | | | | | |
| households | 0.01 | -0.30 | 0.92 | 0.97 | 0.91 | | | | | | |
| nedian_income | -0.04 | -0.12 | 0.20 | -0.01 | 0.00 | 0.01 | | | | | |
| less_1h_ocean | -0.36 | 0.05 | -0.00 | 0.02 | 0.07 | 0.04 | 0.17 | | | | |
| inland | 0.04 | -0.24 | 0.03 | -0.01 | -0.02 | -0.04 | -0.24 | -0.61 | | | |
| near_bay | 0.19 | 0.26 | -0.02 | -0.02 | -0.06 | -0.01 | 0.06 | -0.31 | -0.24 | | |
| near_ocean | 0.30 | 0.02 | -0.01 | 0.00 | -0.02 | 0.00 | 0.03 | -0.34 | -0.26 | -0.14 | |
| n_house_value | -0.08 | 0.11 | 0.13 | 0.05 | -0.02 | 0.07 | 0.69 | 0.26 | -0.48 | 0.16 | |

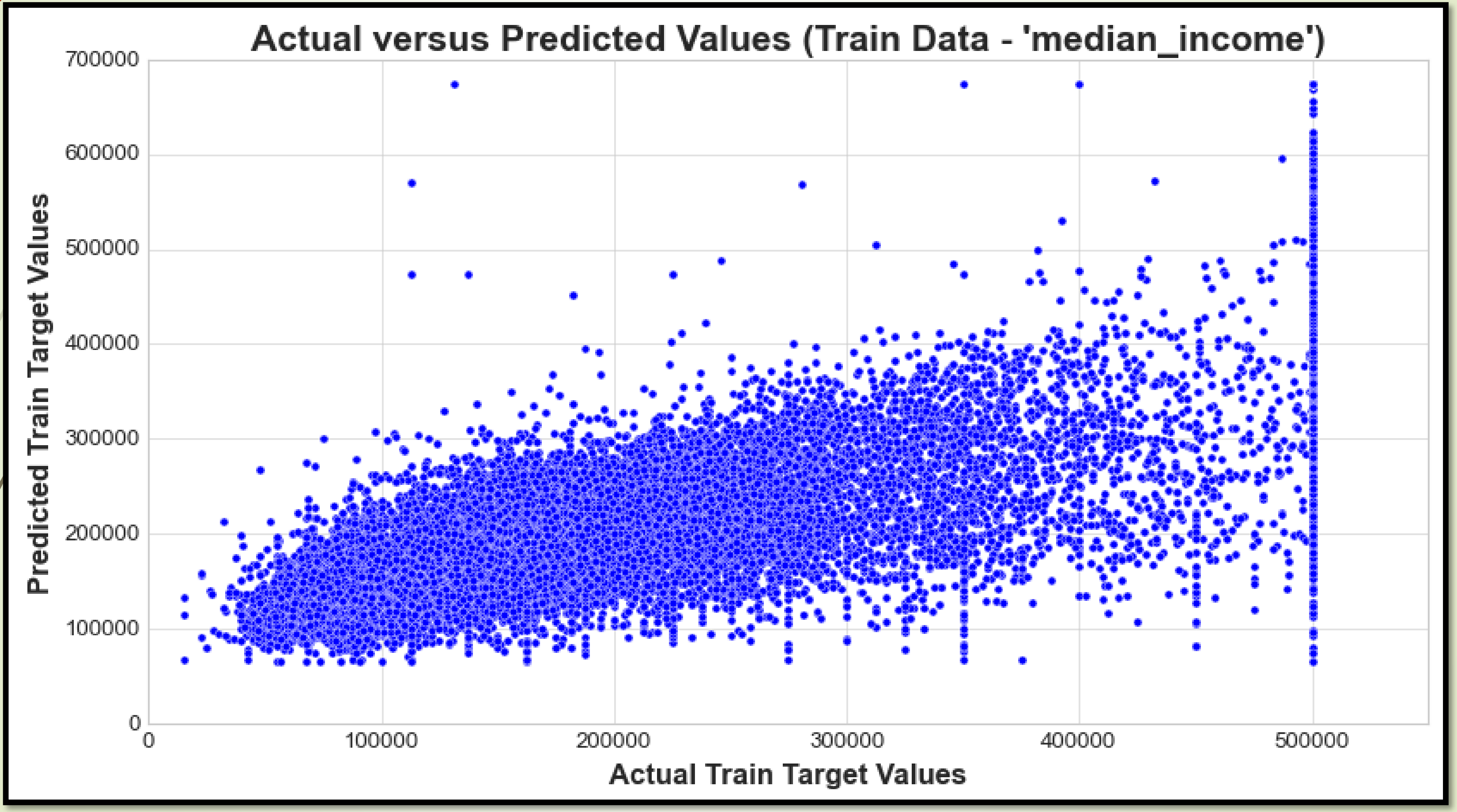
# Linear Regression Model Analysis



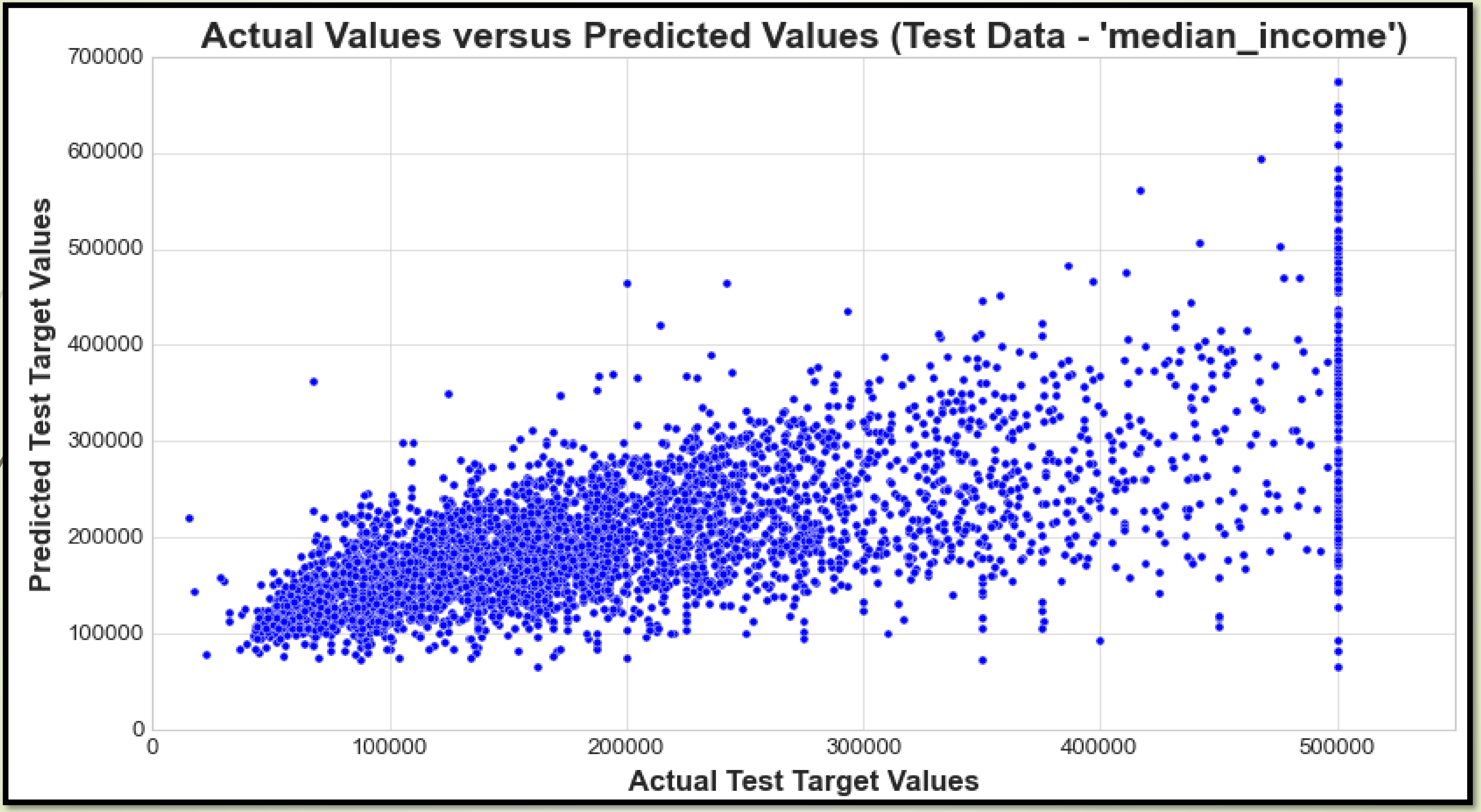


```
Mean Absolute Error: 52473.02
Mean Squared Error: 5242501202.4
Root Mean Squared Error: 72405.12
Mean Absolute Percent Error: 75.05%
R Squared: 0.6
```

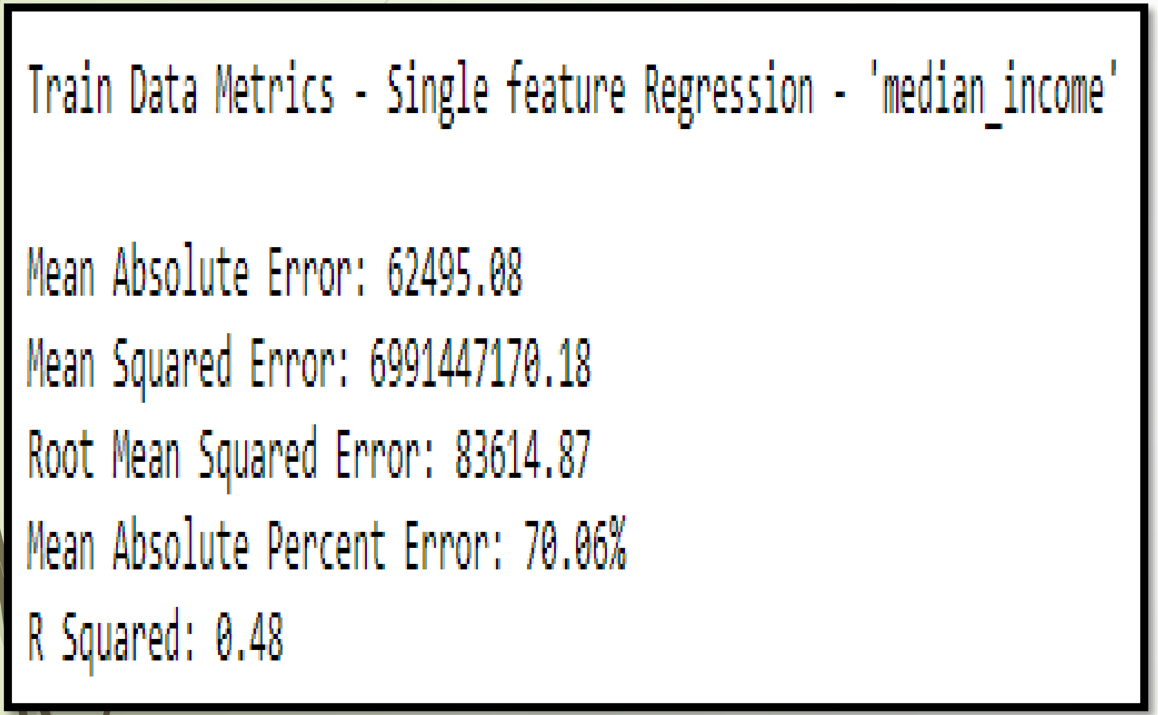
# Linear Regression Model Analysis



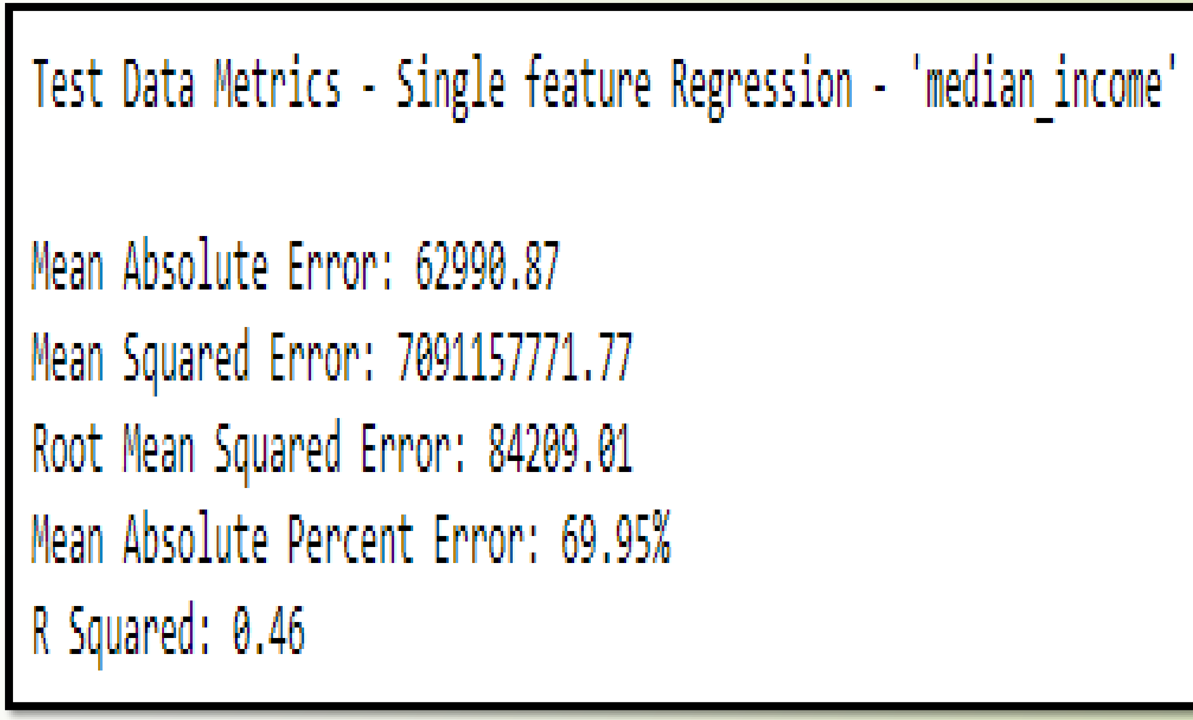
# Linear Regression Model Analysis



# Linear Regression Model Analysis

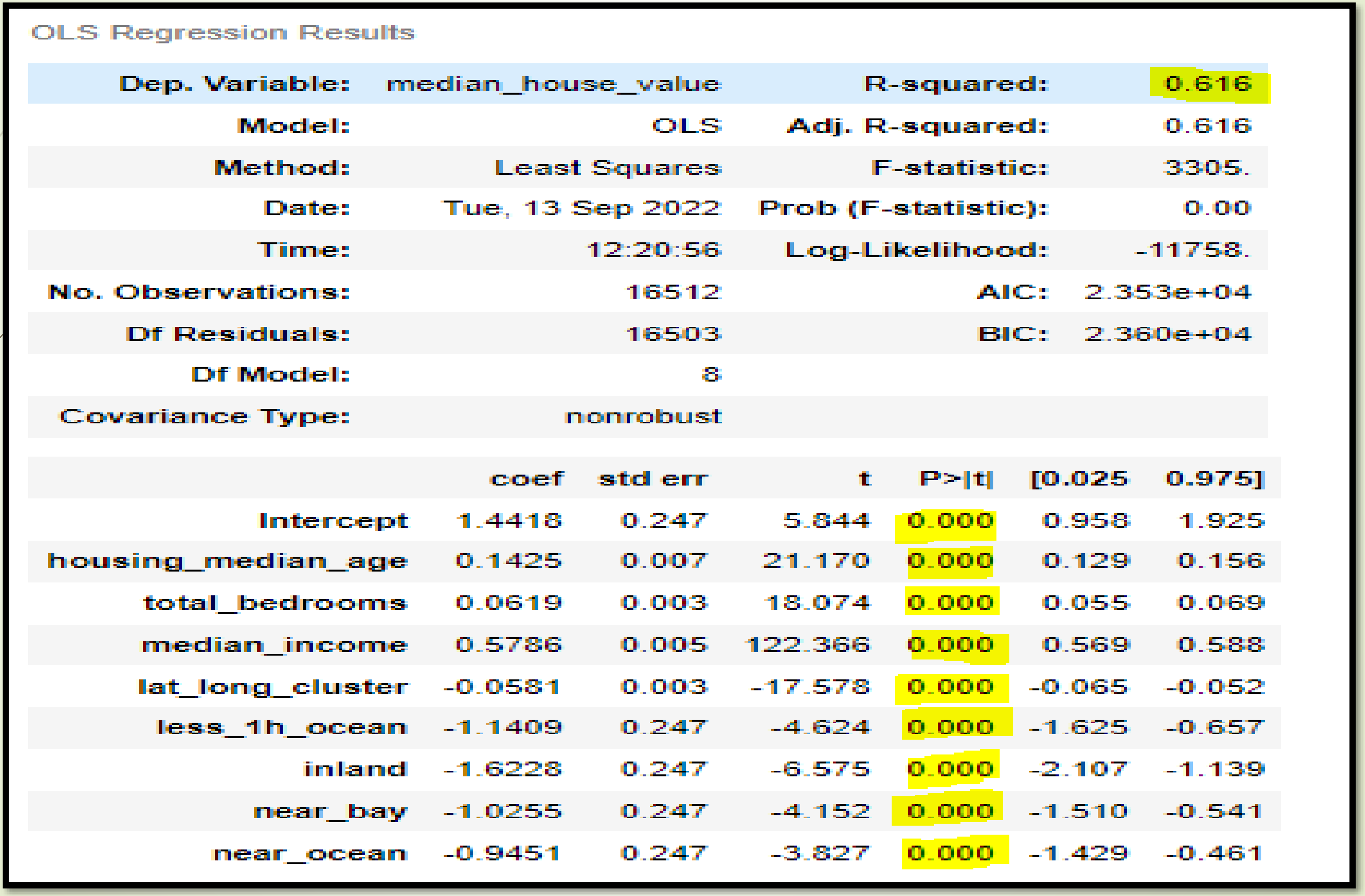


```
Train Data Metrics - Single feature Regression - 'median_income'


Mean Absolute Error: 62495.08
Mean Squared Error: 6991447170.18
Root Mean Squared Error: 83614.87
Mean Absolute Percent Error: 70.06%
R Squared: 0.48
```



```
Test Data Metrics - Single feature Regression - 'median_income'


Mean Absolute Error: 62990.87
Mean Squared Error: 7091157771.77
Root Mean Squared Error: 84209.01
Mean Absolute Percent Error: 69.95%
R Squared: 0.46
```

# Linear Regression Model Analysis

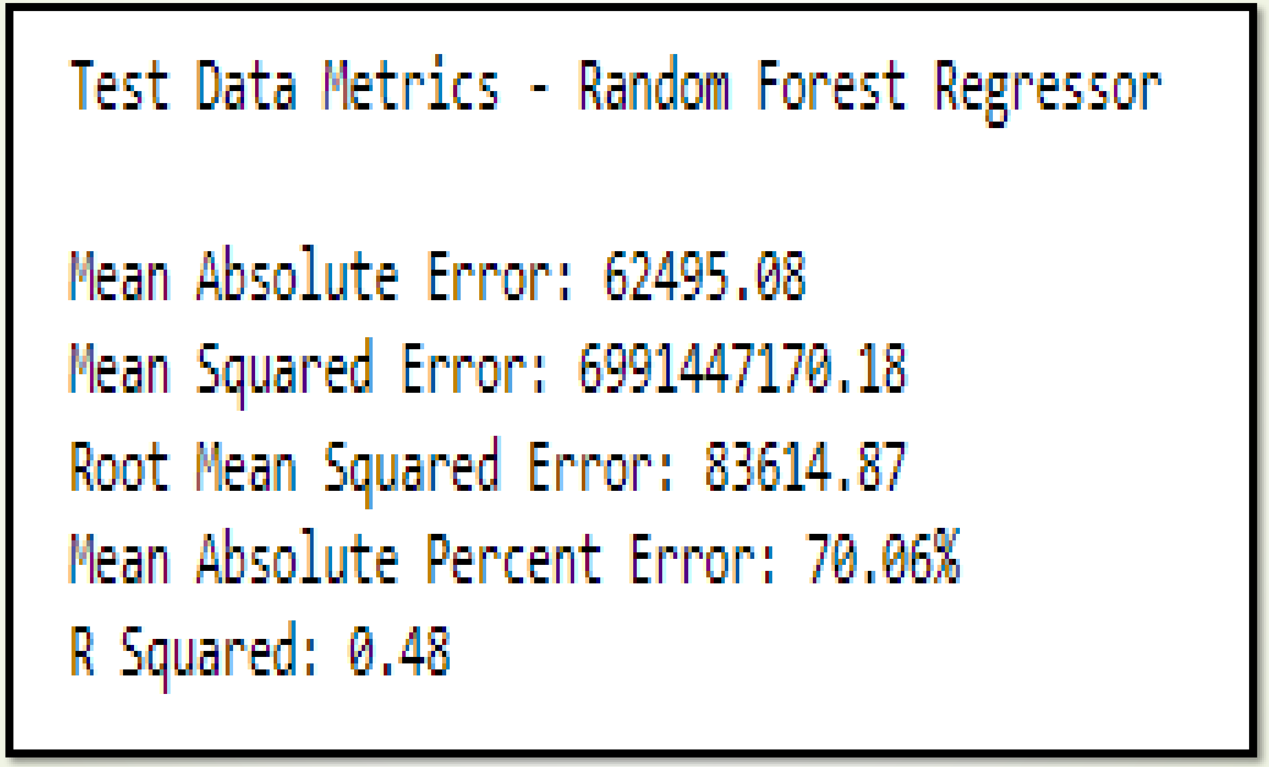


OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | median_house_value | R-squared: | 0.616 |
| Model: | OLS | Adj. R-squared: | 0.616 |
| Method: | Least Squares | F-statistic: | 3305. |
| Date: | Tue, 13 Sep 2022 | Prob (F-statistic): | 0.00 |
| Time: | 12:20:56 | Log-Likelihood: | -11758. |
| No. Observations: | 16512 | AIC: | 2.353e+04 |
| Df Residuals: | 16503 | BIC: | 2.360e+04 |
| Df Model: | 8 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 1.4418 | 0.247 | 5.844 | 0.000 | 0.958 | 1.925 |
| housing_median_age | 0.1425 | 0.007 | 21.170 | 0.000 | 0.129 | 0.156 |
| total_bedrooms | 0.0619 | 0.003 | 18.074 | 0.000 | 0.055 | 0.069 |
| median_income | 0.5786 | 0.005 | 122.366 | 0.000 | 0.569 | 0.588 |
| lat_long_cluster | -0.0581 | 0.003 | -17.578 | 0.000 | -0.065 | -0.052 |
| less_1h_ocean | -1.1409 | 0.247 | -4.624 | 0.000 | -1.625 | -0.657 |
| inland | -1.6228 | 0.247 | -6.575 | 0.000 | -2.107 | -1.139 |
| near_bay | -1.0255 | 0.247 | -4.152 | 0.000 | -1.510 | -0.541 |
| near_ocean | -0.9451 | 0.247 | -3.827 | 0.000 | -1.429 | -0.461 |

# Linear Regression Model Analysis



```
Test Data Metrics - Random Forest Regressor


Mean Absolute Error: 62495.08
Mean Squared Error: 6991447170.18
Root Mean Squared Error: 83614.87
Mean Absolute Percent Error: 70.06%
R Squared: 0.48
```

# Predictive (Linear) Model Summary

- Regression Model did not perform well, has RMSE of 72405.12, MAPE 75.05% with R Square value of 0.6
- Using only single feature (median_income), the model gives better result in terms of MAPE, but does error increases and the model explainability decreases. It has RMSE of 84209.01, MAPE 69.95% with R Square value of 0.46 for test data.
- Comparing the regression metrics for single feature (median_income), of train and test data, observed that the metrics are almost same, which implies that data does not have problem of overfitting and underfitting, therefore, feature engineering is done well.
- Random Forest Regressor, which is believed to be robust against outliers, was also build on the data, and it did not show much improvement. RMSE 83614.87, MAPE 70.06% and R Squared 0.48
- However, future scope of predictive modelling could be to test out other ensemble models, cross validation techniques, further feature engineering and to get more relevant data.

# Appendix

- Please refer 'California Housing Price Prediction-Lavkush.pdf' file, submitted along with this PPT
- Because the code was developed in jupyter notebook, it has source code along with the detailed analysis and report
- All the graphs included in this presentation can also be found in that project report
- This PPT is just a glimpse of the analysis done, for quick reference. Detailed work is present in the project report – "California Housing Price Prediction-Lavkush.pdf".

# Thank you!