# Real Estate Investment Opportunities in the State of Texas

| Project Report Presented to  Professor Vijay Eranti |
| --- |
| CMPE-255 Sec 48 - Data Mining |
| By |
| Anupama Kurudi, Gunjan Srivastava, Krishna Jha, Shivam Tomar  Team BingeCoders |
|  |

ABSTRACT

By Anupama Kurudi, Gunjan Srivastava, Krishna Jha, Shivam Tomar

Texas, the Lone Star State has a population of over 28 million making it the second most populated state in the US. This large population brings with it many real estate investment opportunities. So it’s natural for property investors from all across the US to show interest in the Texas real estate market and the State of Texas has no state property tax. All property is appraised at full market value, and taxes are assessed by local county assessors on 100% of appraised value.

The Texas real estate market is excellent when it comes to traditional investing. One of the reasons why this is the case is its economic success. With the 15th best state economy, 2nd largest state economy, and a GDP of $1.645 trillion last year, the state lives up to its “everything’s bigger in Texas” motto. The state’s job market is very diversified, with the largest industries including construction, schooling, restaurants and school services, software publishing, and energy sectors. The Texan economy has experienced great growth in the last year even with the covid pandemic.

The median household income of the state increased by 4.67% and its job market grew by 2.15% over the year, according to Data USA. Being a cool buyer’s market (whereas many large states are currently hot seller’s markets), the Texas real estate market provides plenty of opportunities for traditional long-term rentals and investment.

**Table of Contents**

[Chapter 1. Project Overview 1](#_heading=h.1t3h5sf)

[1.1 Introduction 1](#_heading=h.3rdcrjn)

[Chapter 2: Data Collection 2](#_heading=h.44sinio)

[Chapter 3: Data Understanding](#_heading=h.gdvkok3snud0) 3

[Chapter 4: Data Cleaning](#_heading=h.zii5653nphk) 5

Chapter 5: [Visualization](#_heading=h.osbbcl9mjk1z) 6

[Chapter 6. Methodologies](#_heading=h.i9dur7hw02cx) 8

6[.1 Feature Classification](#_heading=h.8hn0qwmmbnbp) 8

[6.2 PCA](#_heading=h.3fwokq0)9

[6.3 Normalization](#_heading=h.4buniycfp3nk) 10

6.[4 Regression](#_heading=h.2u6wntf) 12

[Chapter 7. Fractal Clustering 1](#_heading=h.puypxxp2aule)4

[Chapter 8. Training and Prediction](#_heading=h.4bvk7pj) 20

[Chapter 9. Testing and Analysis](#_heading=h.2r0uhxc) 21

[Chapter 10. Conclusion](#_heading=h.i3qxdggsozct) 21

[Chapter11. Deployment - MLOps](#_heading=h.umy4fks7j18z) 22

[Chapter 12. ACKNOWLEDGMENT](#_heading=h.kbc963voo8h) 22

[References](#_heading=h.1jlao46) 23

## Chapter 1. Project Overview

## 1.1 Introduction

The housing market is at an all time high with demand going steeply up while the supply remains fairly limited. Investors, first time home buyers and individual buyers all largely look for properties where they can maximize the square footage at the least possible price. Additionally, families with children also look for areas that have a good school ranking. This also ensures that the property value remains the same or goes higher in case they want to sell these properties in the future.

To help investors, agents and buyers find this set of properties that yield the best combination of house attributes including location, we build a system that outputs a Golden Cluster of properties. This cluster is a set of around a dozen properties that would be the best set of properties to be investing in. This system takes into account the parameters such as low price per square foot and high school ranking. It can also get the list of properties for specific zip codes.

## 1.2 Business Objective

## 

The business objective of the project are as follows:

* The objective of the project is to train the model to predict the house property price, square feet area and school popularity based on the inputs data in the state of Texas.
* Find the Golden cluster with the properties that have high school popularity and the high square feet area with the less price for a good business use case.
* Predicting the investment properties for an investor who wants to buy a house for a low price but with a higher square feet area and a good school rank.
* Finding the Golden cluster which signifies that it has the lowest costing house that has the highest school ranking. Relevant latent variables are taken into consideration.

The school ranking ranges from 1 to 100 where 1 means highest school ranking and 100 means least popular.

## Chapter 2: Data Collection

## 

Three datasets have been used from different sources including Kaggle, government websites, scraping from zillow to have large datasets to predict the house prices for Texas.

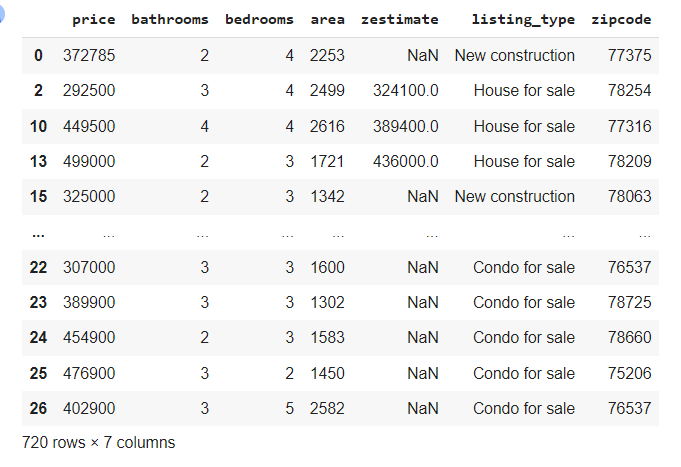
**Base Dataset:** Real Estate dataset has been downloaded from Kaggle and uploaded to Google Drive. This the base dataset that can be found [here](https://drive.google.com/file/d/1qBdrJYIzg8RqH5Y4KoqDIp4tIR4pPY_i/view?usp=sharing)

**School Dataset:** Downloaded school dataset from Texas Government website. This contains the features related to schools in Texas along with the zipcodes. The school dataset can be found [here](https://drive.google.com/file/d/10SiS52ScxegmuJK3s3LfgSoJYgyu4F16/view?usp=sharing)

**Scraped Dataset:** For the third dataset, we have scraped the real estate data from zillow.com using Beautiful soup library. Extracted data for different listing types of properties and later amalgamated into the base dataset in three iterations. Data for three types of properties are been extracted :

* House for sale
* Townhouse for sale
* Sqft-condo for sale

As the last step, all the three datasets have been merged together and provided to the next phase of the project.

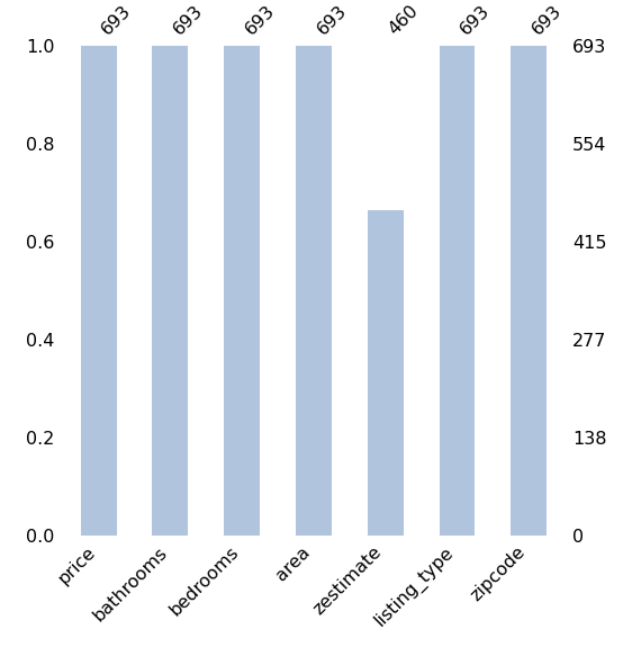
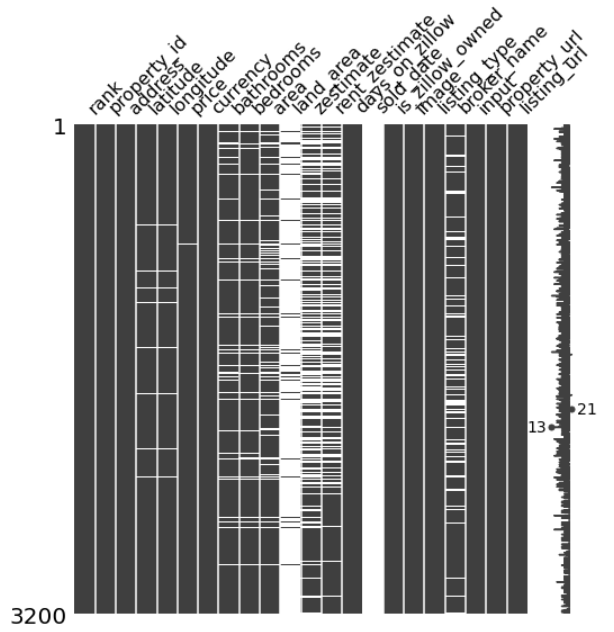


**Merged dataset**

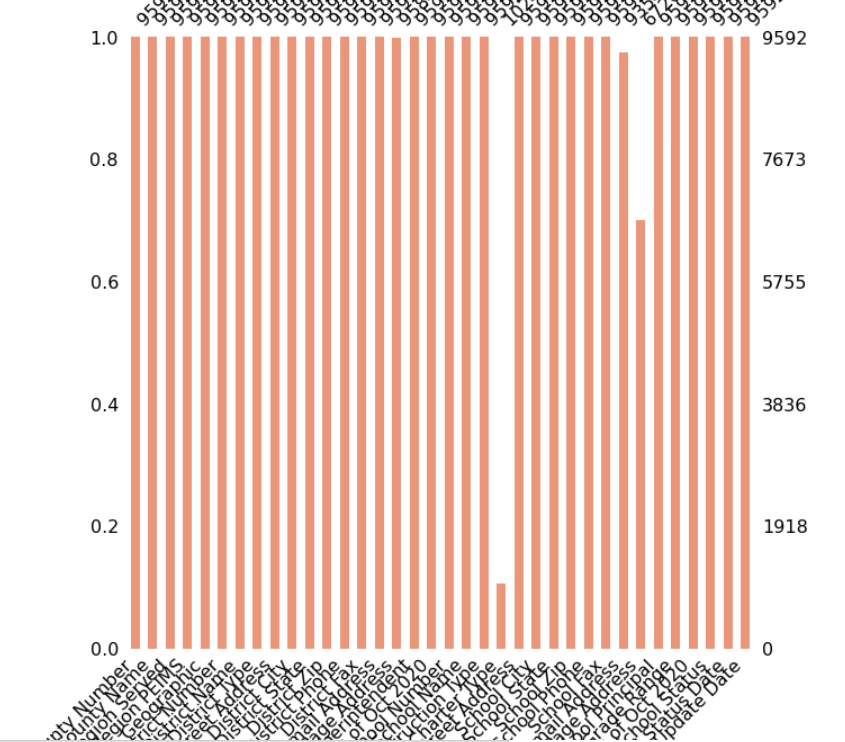
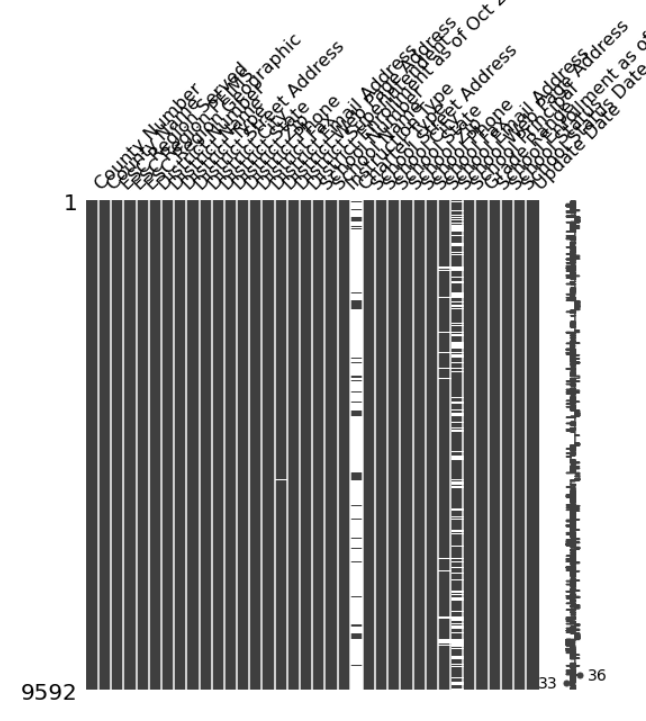
## Chapter 3. Data Understanding

## 

## The base dataset and school dataset have been visualized to have better understanding of the data. This gave us a clear picture of how sparse or dense the data is and how many null values we have in the datasets that would be handled in the data preparation and cleaning phase.



**Figure 1: Data understanding on Base Dataset**



**Figure 2: Data understanding on School Dataset**

## Chapter 4: Data Cleaning

Data cleaning has been performed as the next step where multiple steps have been taken which includes:

### Dropping the columns which are not required for our use cases.

### Extracting the zip code and State from the address and dropping the address once the state and zipcode have been extracted from it.

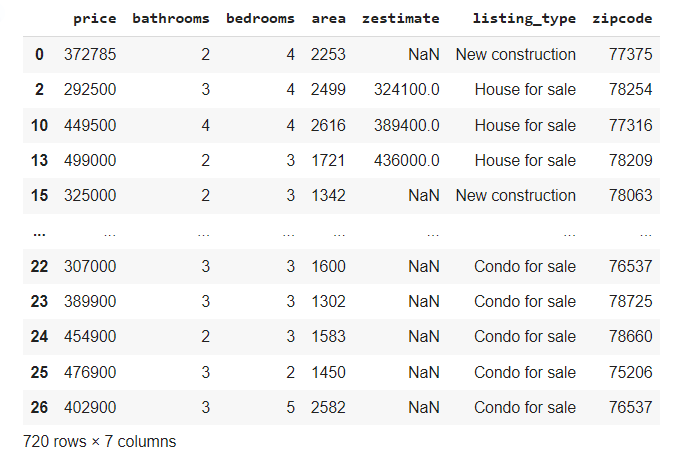
### Removing the sq ft unit from the area as it needs to be converted to integer value.

### Dropping the null values for a subset of the columns.

### Converting the data types for the columns for better understanding and visualization.

### Grouped the same zip code and calculating the aggregate sum of the corresponding values.

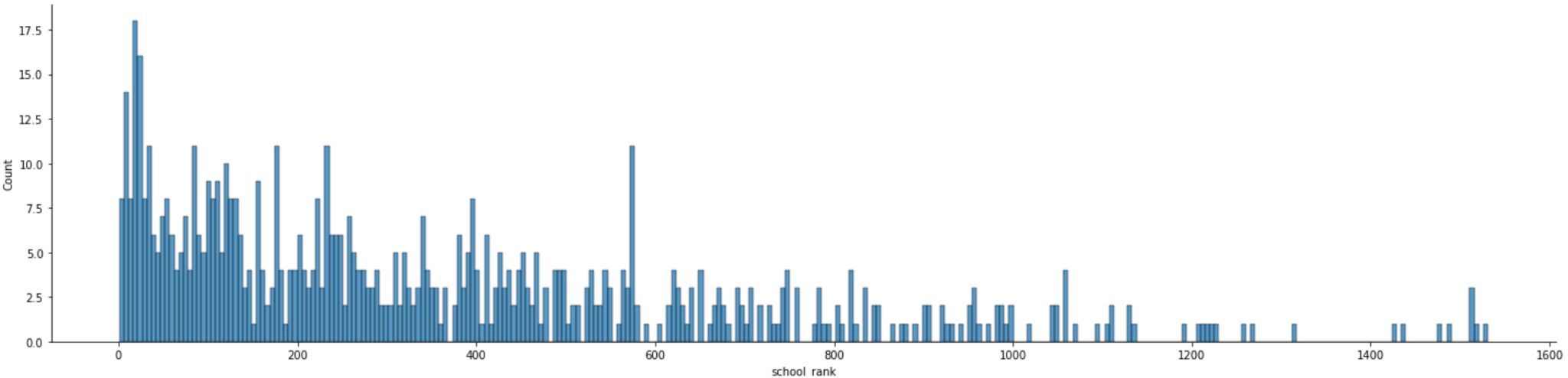
* Reindexing has been done on the school rank from high to low in the range of 1 to 100.



**Figure 3: Dataset after Data Cleaning Phase**

## Chapter 5. Visualization

To have a better understanding of the merged data, we plot graphs for each column value. This helps us visually understand the data distribution. We perform visualizations to see if the data is normally distributed. This way we can further pick the columns to work on and drop the ones that are not required.



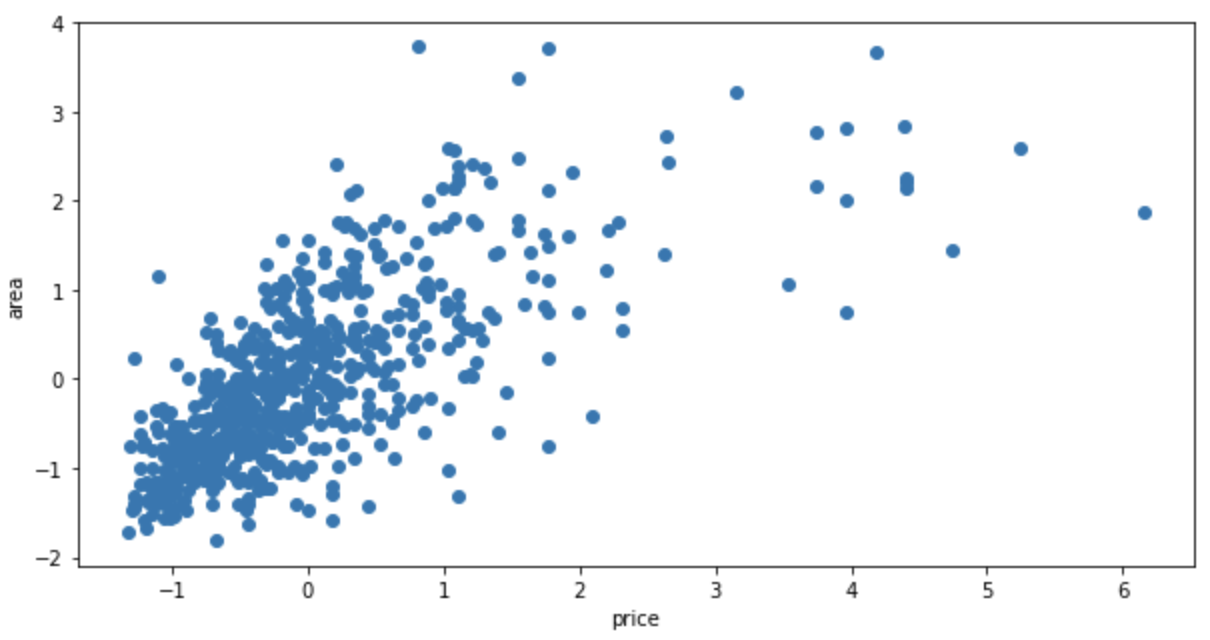
**Figure 4: School ranking data distribution**



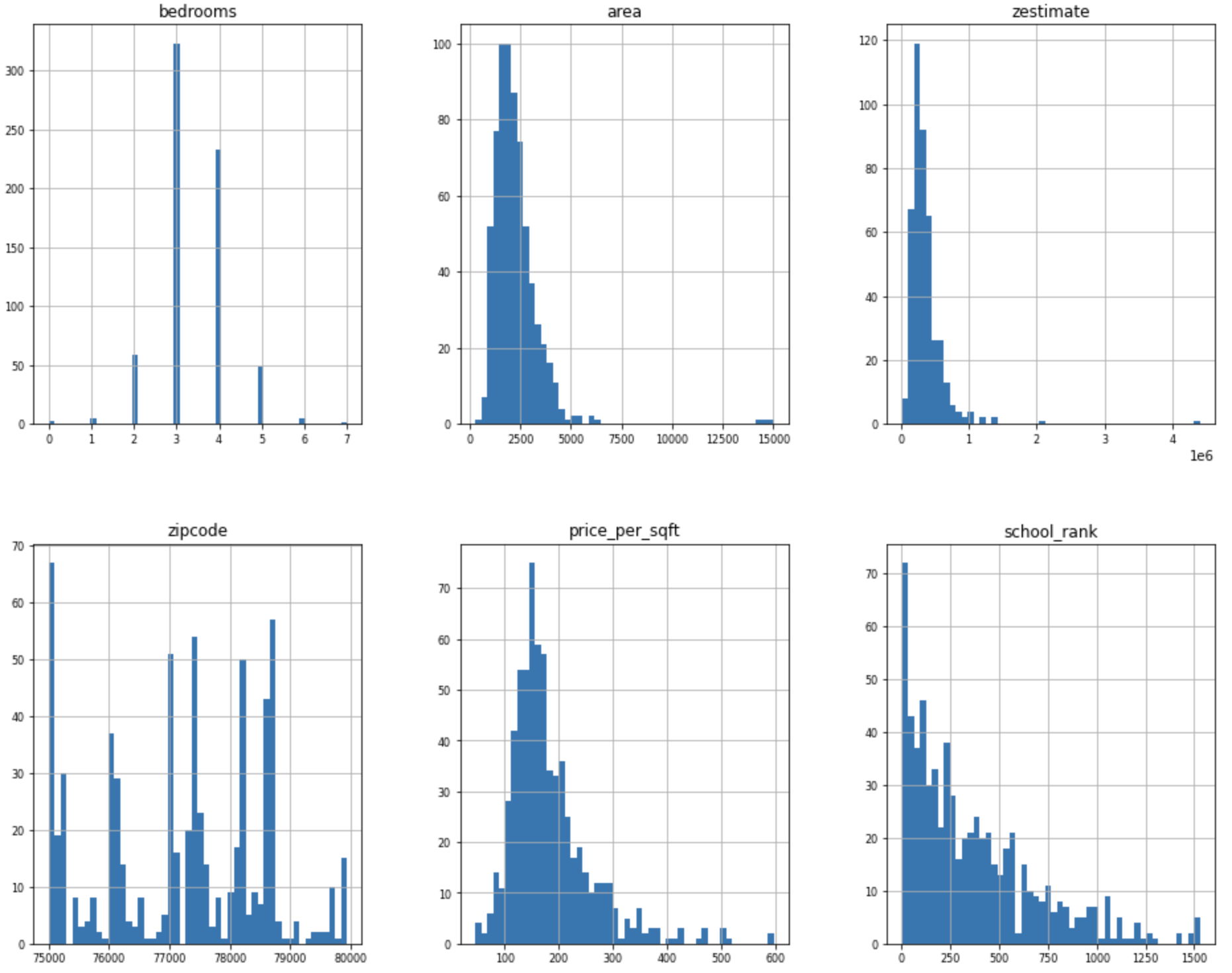
**Figure 5: Confusion Matrix on the merged dataset**



**Figure 6: Data distribution of latent variable - *price\_per\_sqft***



**Figure 7: Scatter plot of price vs area**



## 

## Chapter 6. Methodologies

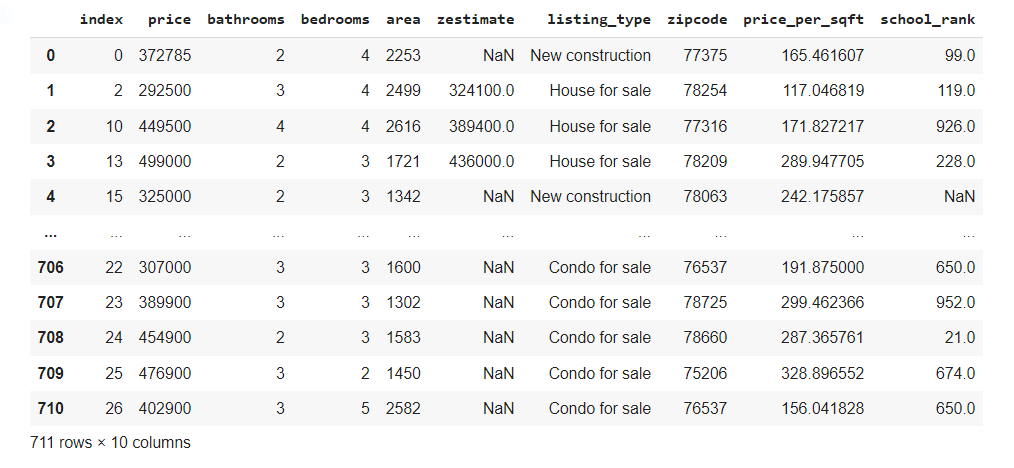
## 

### 6.1 Feature Classification

### 

Feature classification has been done by adding new features from the existing set of features, Classified "price per square feet area" based on the price and area columns.

.

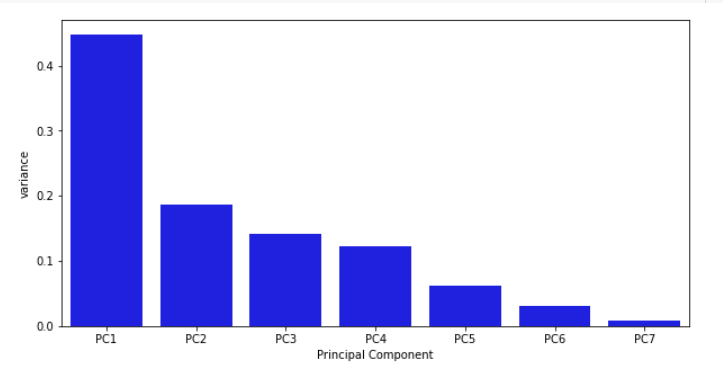


**Figure 8: Merged dataset with Classified column**

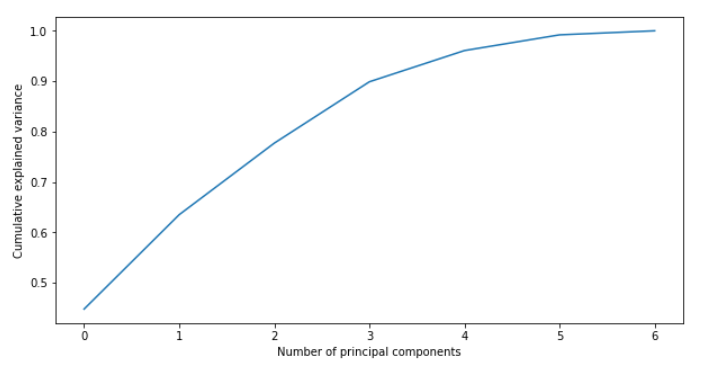
# 

### 6.2 PCA

Performed PCA to do dimensionality reduction on the merged dataset. Calculated the eigenvalues and eigenvectors. Principal components have been calculated too and plot them on the histogram.



**Figure 9 :Histogram Representation of the PCA**

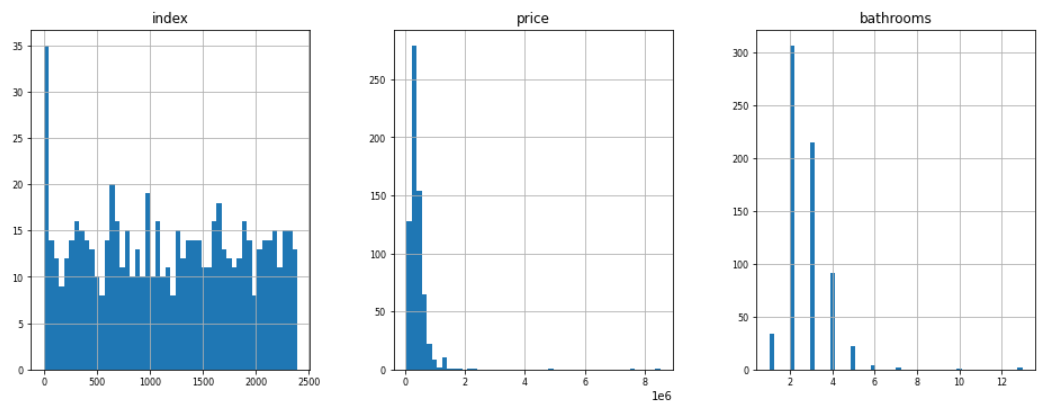


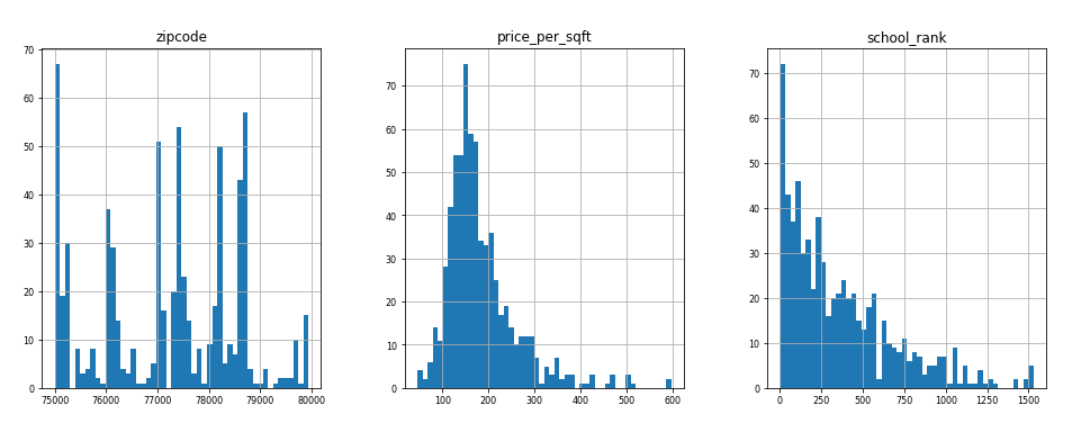
**Figure 10: PCA vs Cumulative explained variance**

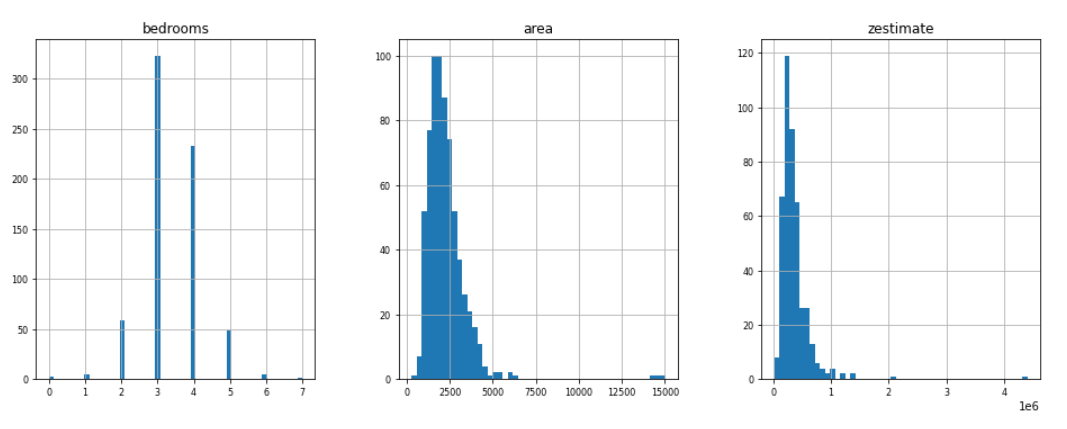
### 6.3 Normalization

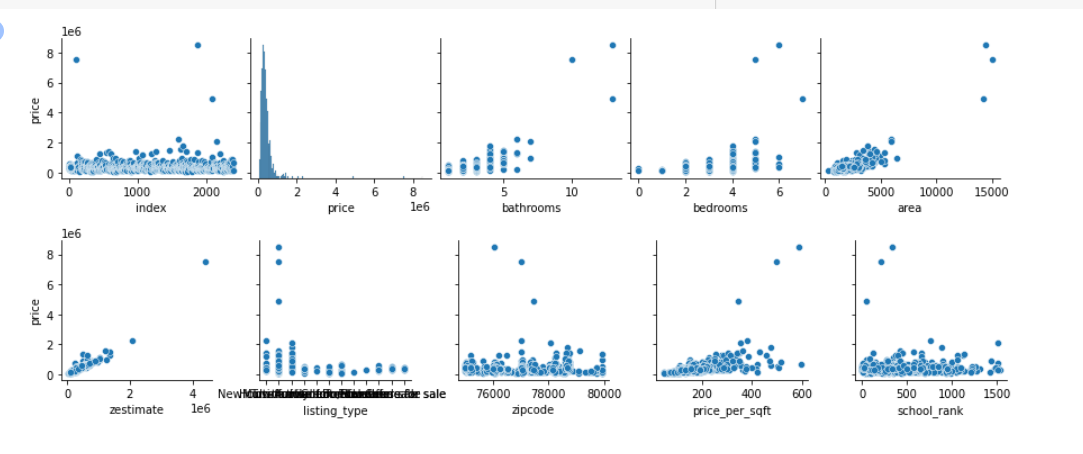
Normalizing the feature school rank between the range of 1 to 100 as the data was ranging from 100 to 1400. For the popularity range, 1 being the most popular and 100 being the least popular.

Normalization has been done to achieve better understanding and for more clear data variance.









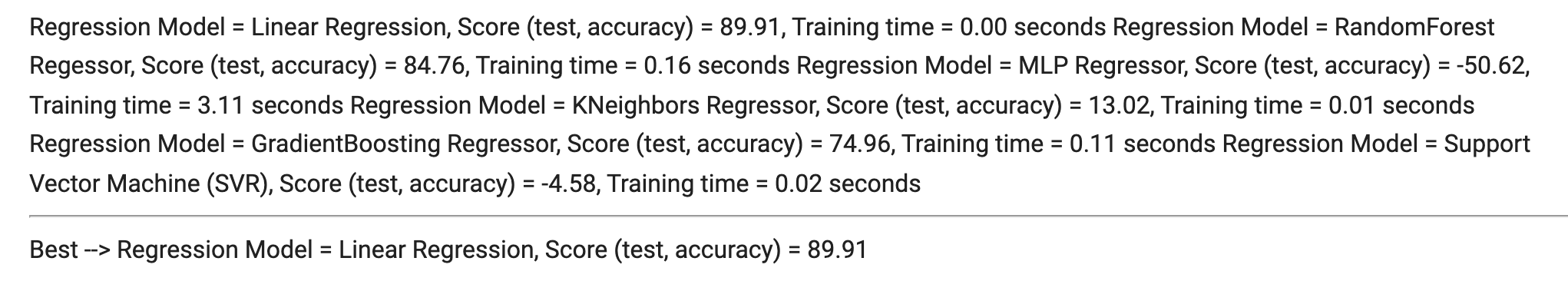
**Figure 11: Visualization of dataset after Normalization**

### 6.4 Regression

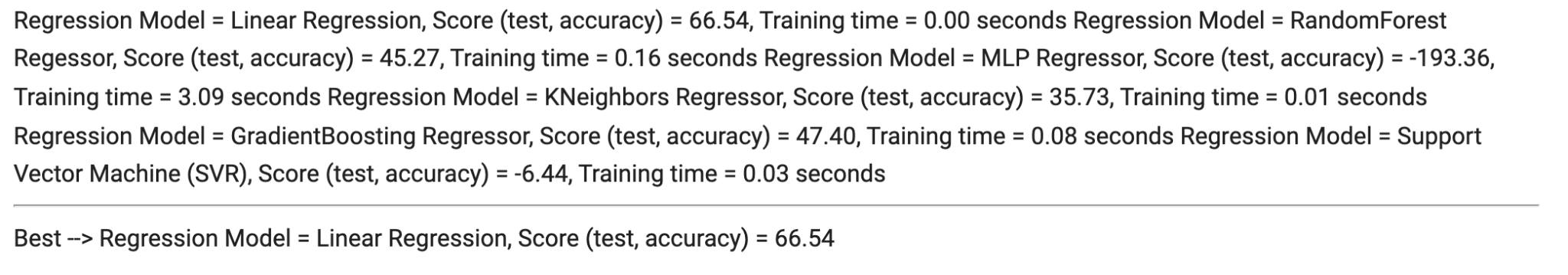
The following regression algorithms are applied to find the most suited regression algorithm for the given dataset:

* Linear Regression
* RandomForest Regressor
* MLP Regressor
* KNeighborsRegressor
* GradientBoostingRegressor
* Support Vector Machine (SVR)

We find that Linear Regression performs the best on the test dataset with a score of 89 and therefore pick this model to do further predictions. We also use Linear Regression models to test the performance of predictions with and without latent variables. The latent variable, price\_per\_sqft is added to the dataset. It is found that of the predictions without latent variables in the dataset, the model performs at a test and accuracy score of 66%, whereas the model performs better at a rate of 89% with the latent variable. Therefore, we choose to continue analysis with the latent variable. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).



**Figure 12: with Latent Variable**



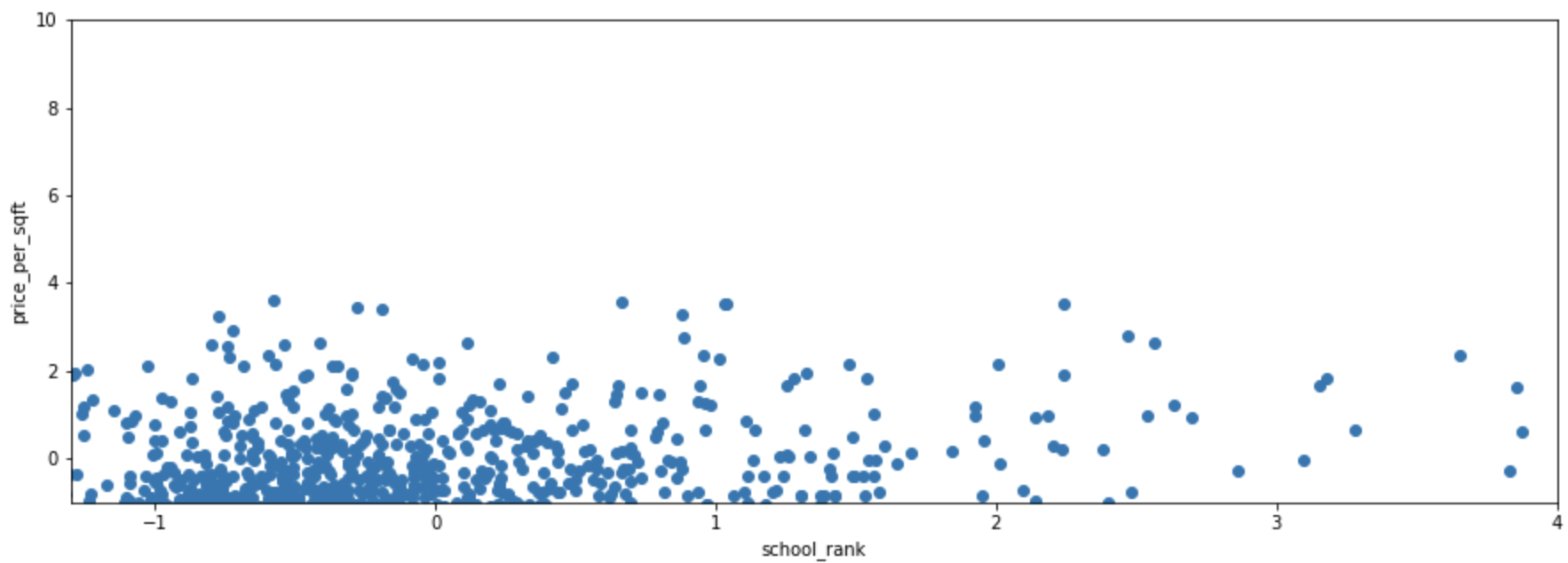
**Figure 13: without Latent Variable**

As per results, the best regression algorithm for the data set is Linear Regression (Score: 66.54). We will use the Linear Regression algorithm to train a model and use it to predict the house prices to help a buyer understand the market.

# 

## CHAPTER 7: Fractal Clustering

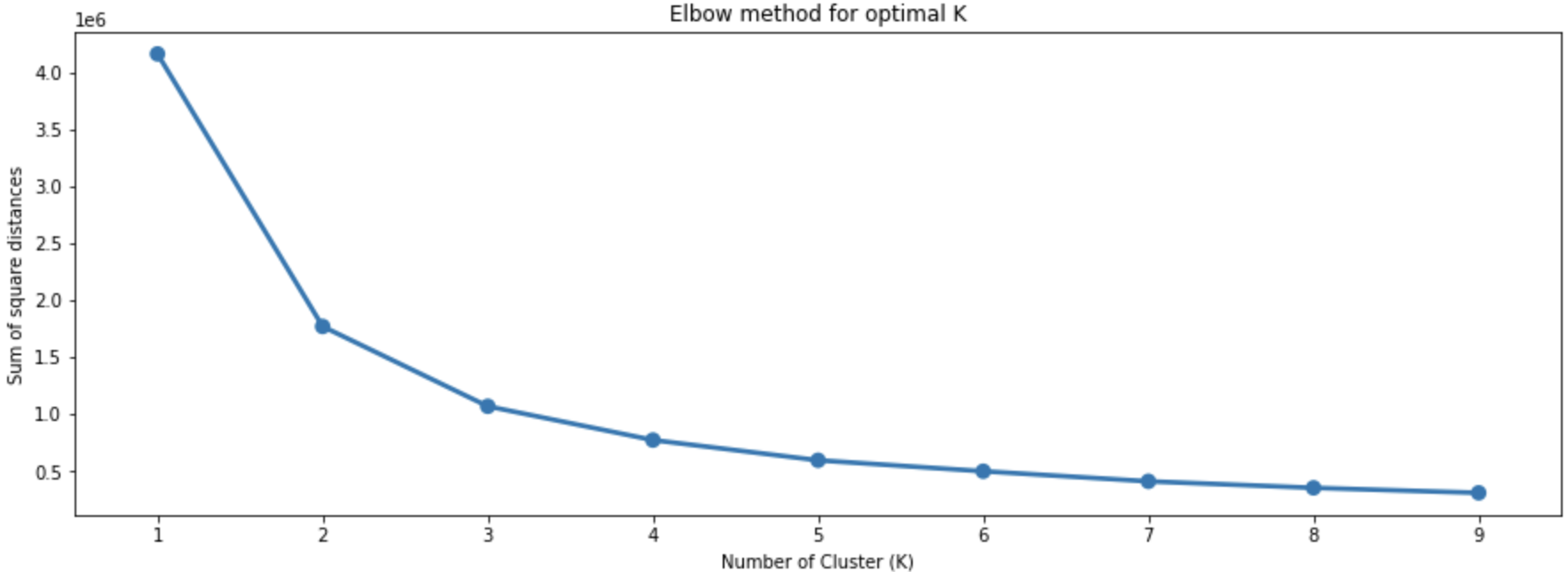
In the second part of the project, we find a golden cluster of properties which are a set of properties that meet a certain criteria. In our project, the criteria is one that is most popular in the housing market - low cost of price per square foot. Many buyers are looking to maximize the square footage they can buy at a price and often look for good school districts. This also ensures a good resale value in the future. Therefore we set out to find a cluster of houses that lie on the lower left corner of the graph, that represents low price per sqft and high School ranking.(figure) The y-axis represents the price per sqft and the x-axis represents the school ranking.



**Figure 14: Graph showing the distribution of data points**

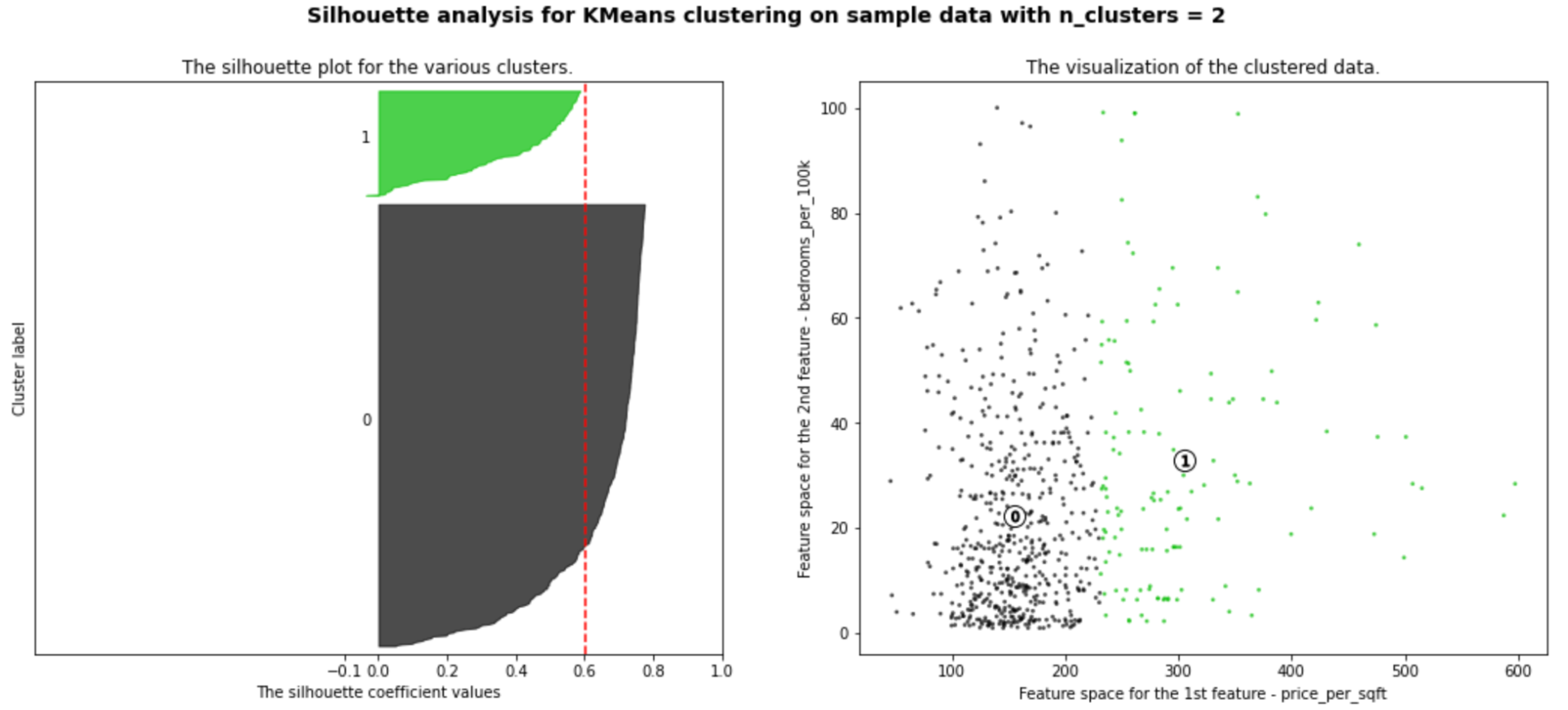
## Iteration 1

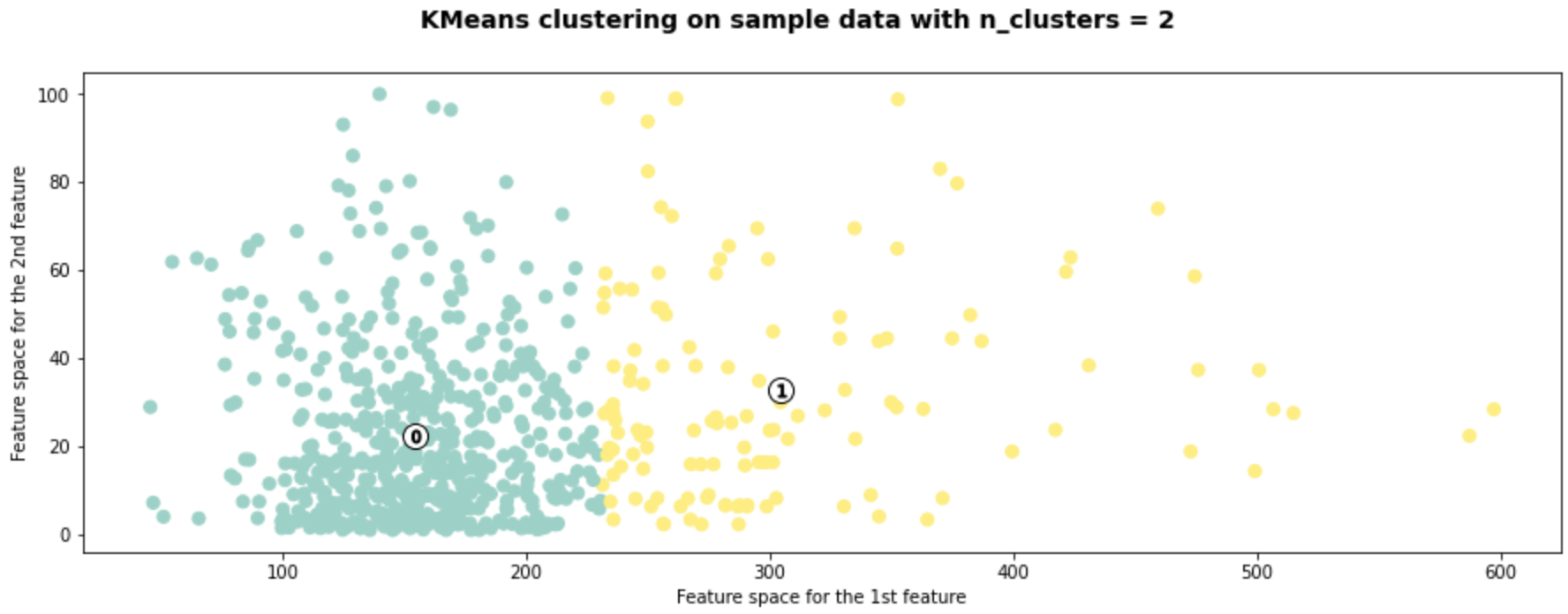
The clustering algorithm used here is K-Mean clustering. In our first iteration we perform the elbow method to determine the ideal number of clusters, the value of ‘k’.



We calculate the Silhouette score to see which value of ‘k’ gets the best score. Through this, we also understand how the data points are clustered for different values of ‘k’.

The Elbow Graph shows that there is a significant drop from point 2 to point 3 and then continues to drop from there. The silhouette analysis shows that for n\_clusters = 2, the average silhouette\_score is : 0.594619586166531. The highest score and the best choice in the number of clusters. Therefore the ideal number of clusters after this analysis is found to be 2.





After performing the first iteration of the KMeans clustering, we obtain Cluster 1 and Cluster 0.

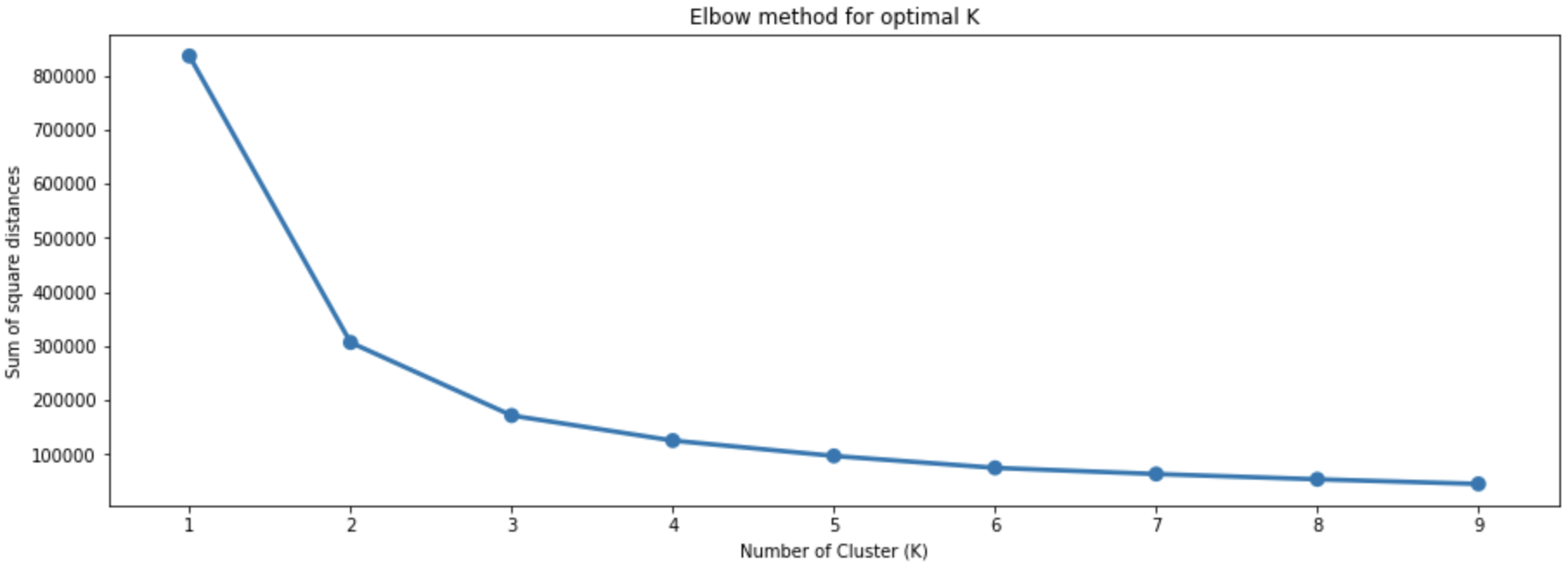
We need the golden cluster that is at the bottom left corner of the graph where we have a low price per square feet for the most popular school based on the ranking.

Let us extract Cluster 0 and perform further clustering on it.

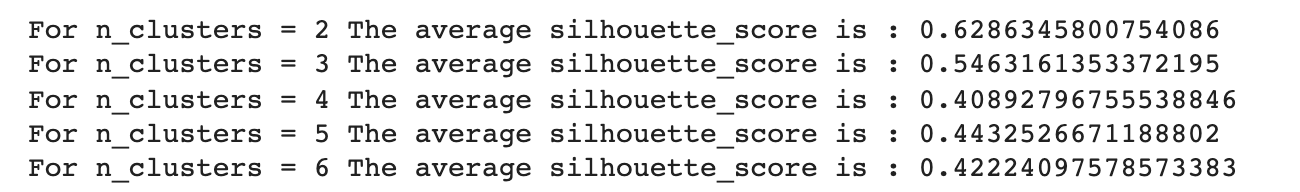
# 

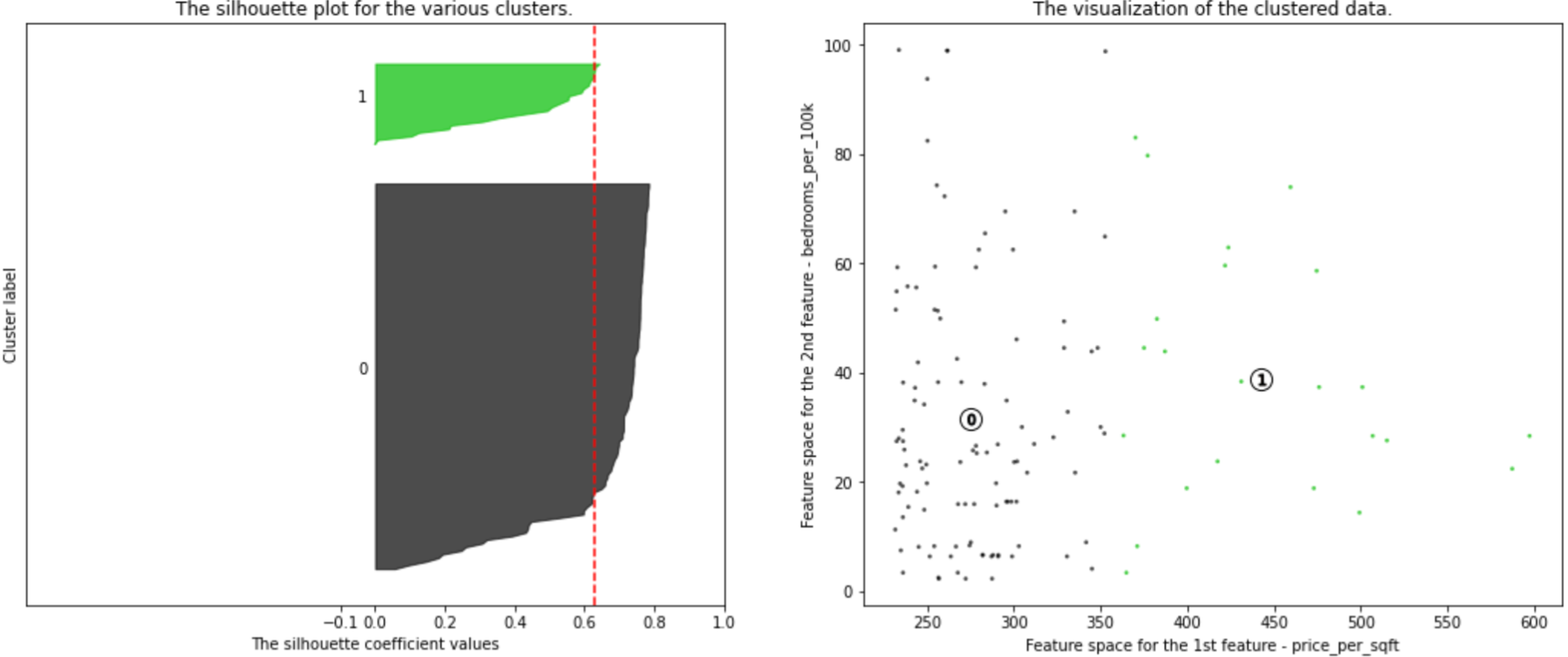
# 

## Iteration 2



From the Elbow graph we find that the ideal number of clusters is 2.



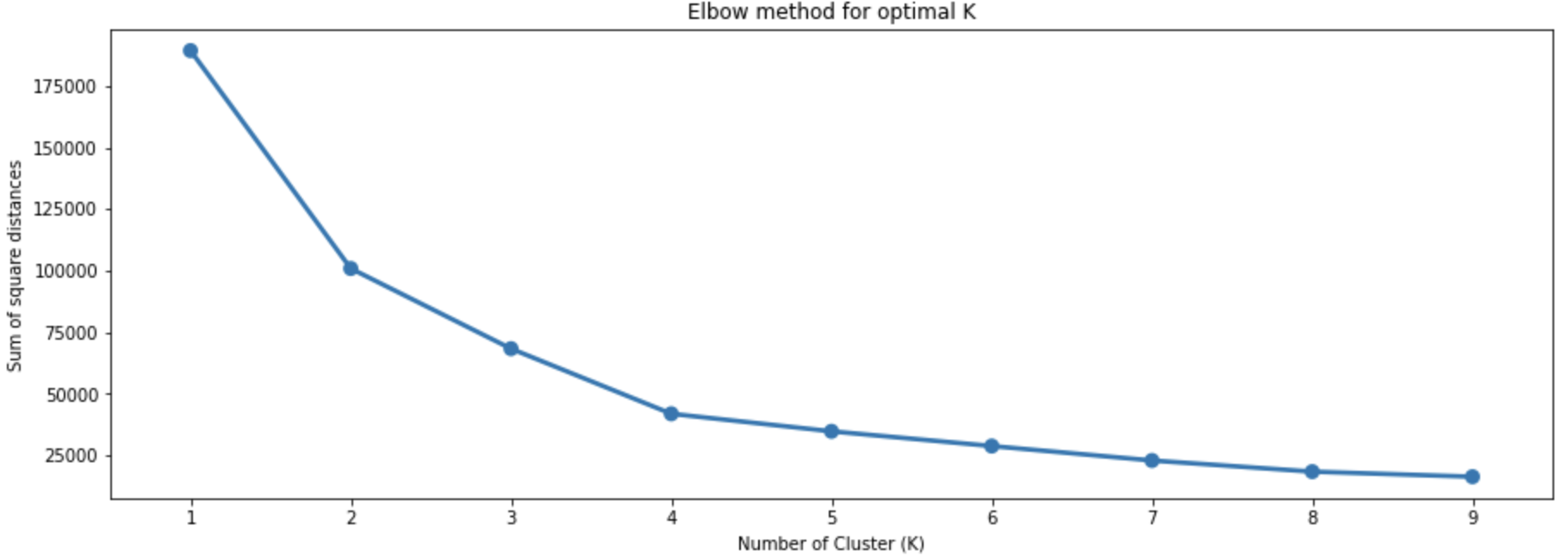


We find that the best Silhouette score = 0.4124687915434243 is for number of clusters = 2.

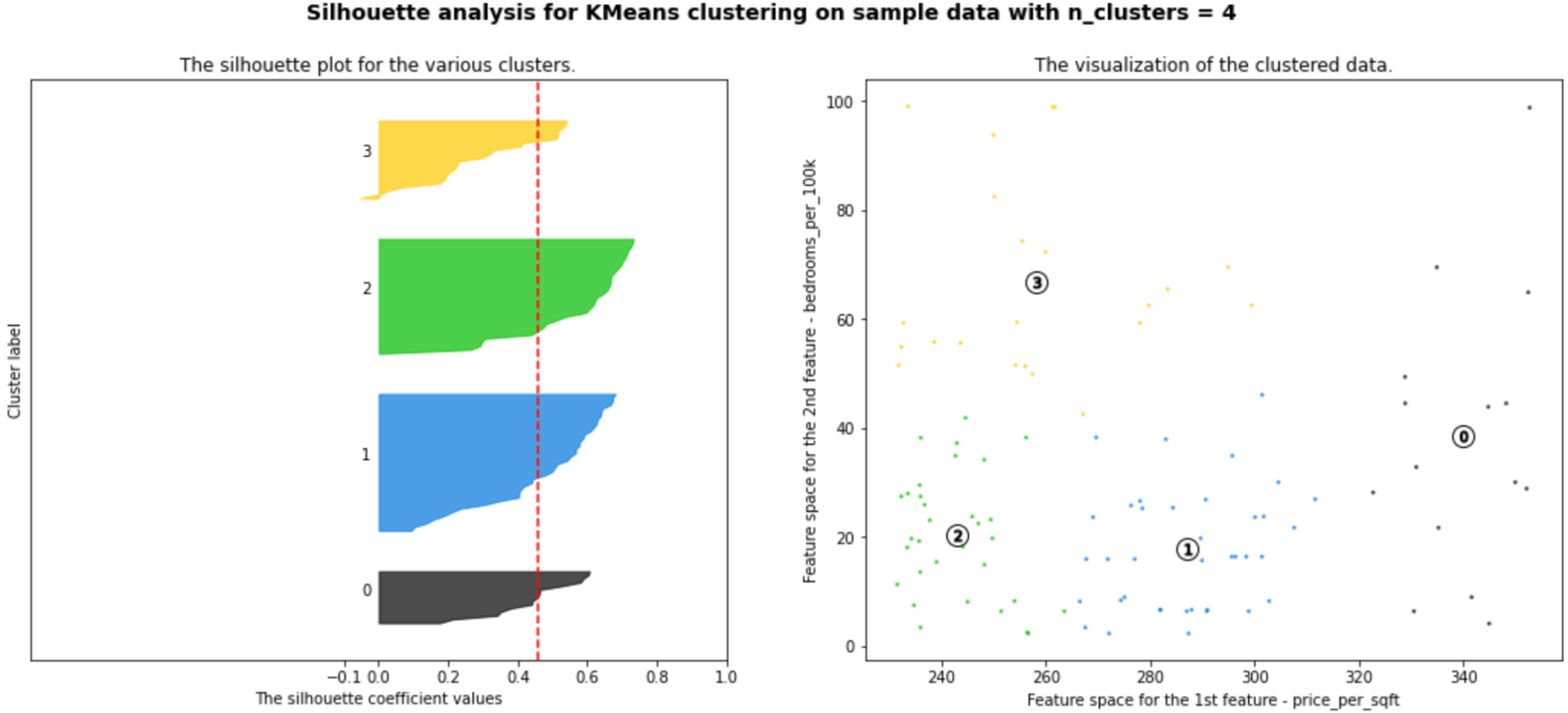
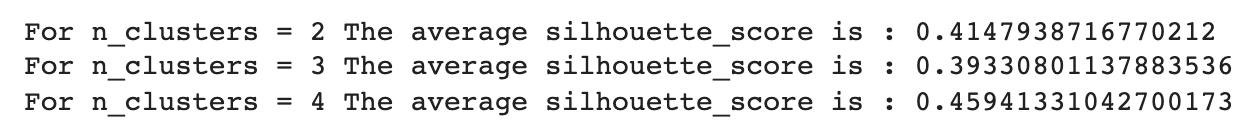


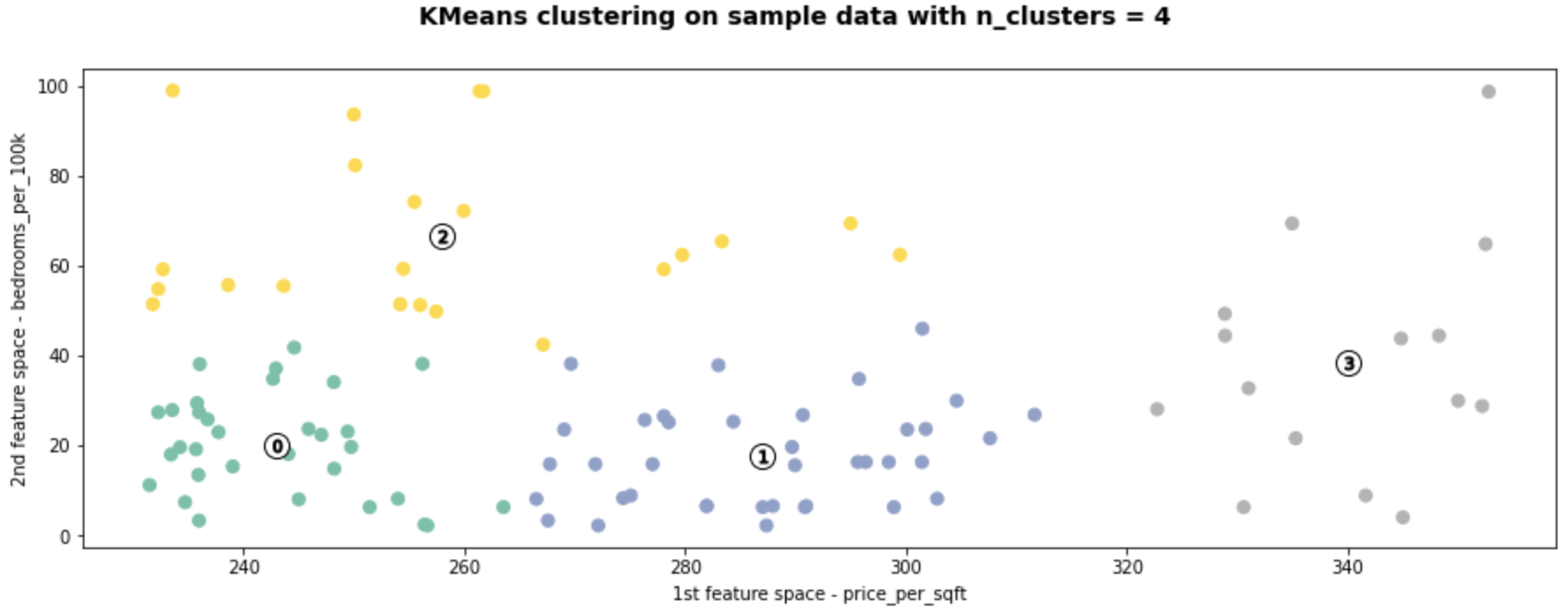
Let us extract Cluster 1 and perform further clustering on it.

## Iteration 3:

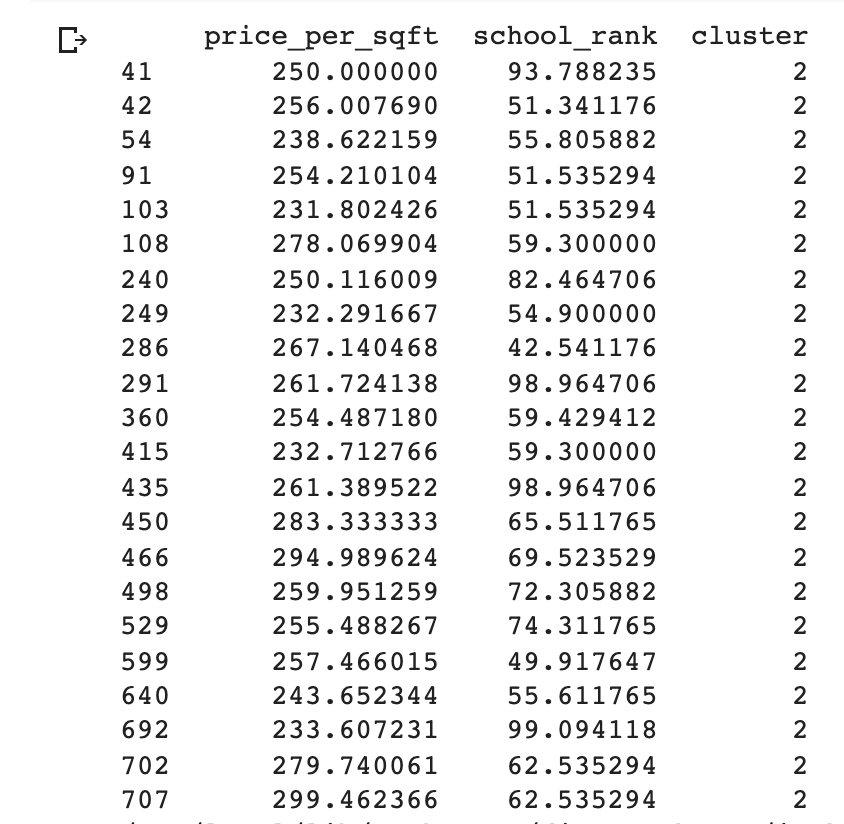


From the Elbow graph we find that the ideal number of clusters is 4.





Between cluster size 3 and 4, number of clusters = 4 provides the Golden Cluster which is Cluster no 2. Shown above in yellow marked 2.



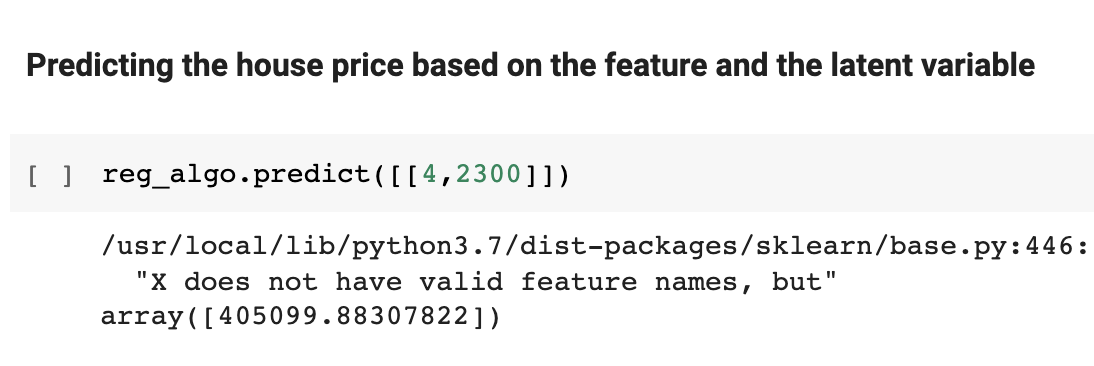
**These properties seen above are the Golden Cluster deemed to be the ideal set of properties for investing.** They are identifiable by their indices.

## Chapter 8. Training And Predictions

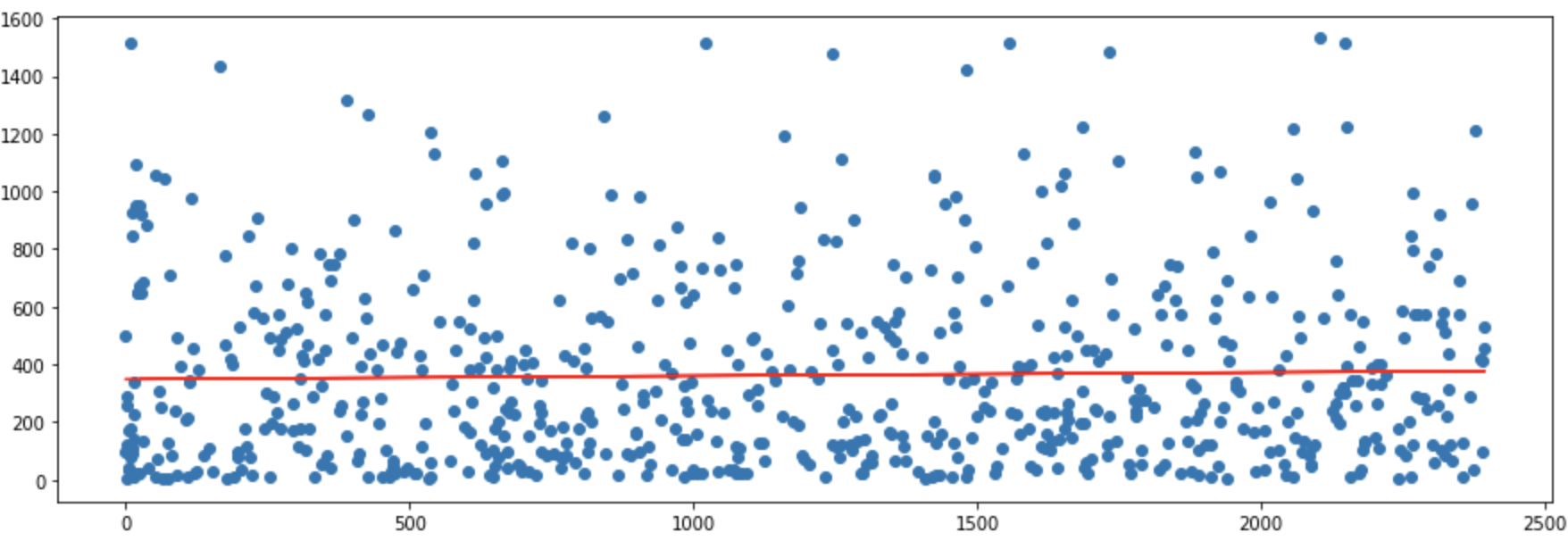
### 

### 

We input the school rank and square footage values into the predictor and get a result of the price. This way if a buyer or an investor looking for a house with a certain school rank and square footage, can input these values to know the budget they need to have for such a property.



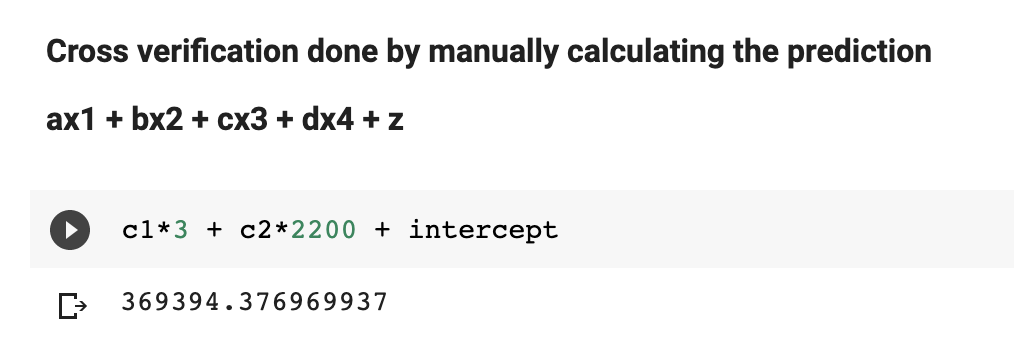
As we can see in the above prediction, to purchase a property whose district school rank is a minimum of 4 and square footage is 2300 sqft, a buyer needs to have a budget of about $ 405,099.



**Figure 15: Data distribution of linear regressor predictor.**

## Chapter 9. Testing And Analysis

We then perform calculations as seen below:

****

**We find that the predicted values are not far from the actual calculated values.**

The predicted value for the same was around $388,758.

## 

## Chapter 10. Conclusion

# In Texas, the price of the houses with square feet area ranging between 2100 to 2500 sq ft cost between 310k to 340k.

# The house price prediction becomes more accurate when we introduce the latent variables from a different dataset.

# The Regression score was better when we ran it with the latent variables. The variance of the data got improved.

# It can be concluded that the house prices go high when the school ranking is better, which means as the school popularity increases, the property prices increase.

# Through fractal clustering, we found that there are properties with number 15 that have low price per square feet with high school ranking that would be an ideal buy for any investor.

# Any investor who is interested in buying a property in Texas, can easily find a house ranging between 310 - 340k with a square feet area of around 2400 and with a *high* school ranking.

# 

## 

## Chapter 11. Deployment, Operation(MLOps)

The deployment of the trained model is done with the help of Vertex AI, which is Google cloud platform’s Machine Learning Pipeline solution. Deployment part of the project also maps with the 6th stage of the CRISP-DM process. The deployment pipeline is written in python and contains following 4 components:

* Load data
* Prepare data
* Train Model
* Deploy Model

Let’s see these steps individually in brief:

**Load Data**

This step load the CSV formatted data from remote storage and pass it to the next stage of pipeline.



***Fig 11.1 Load data component***

**Prepare Data**

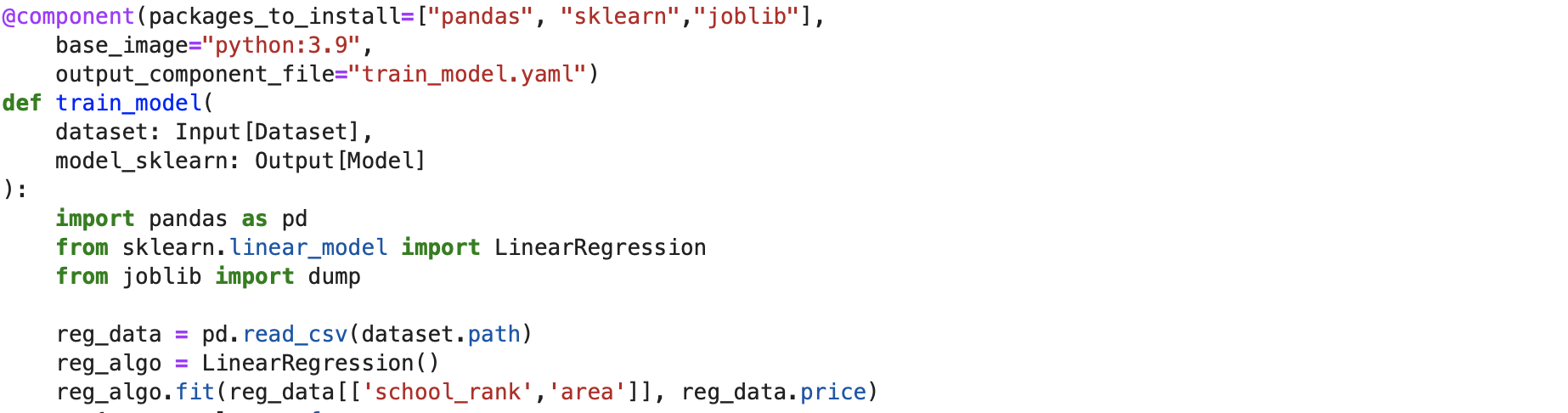
This step of the pipeline performs various preprocessing and cleaning steps on the data like data cleaning, feature engineering, amalgamation, normalization etc. The cleaned data is further passed to the train model step for training of the model.



***Fig 11.2 Prepare data component***

**Train Model:**

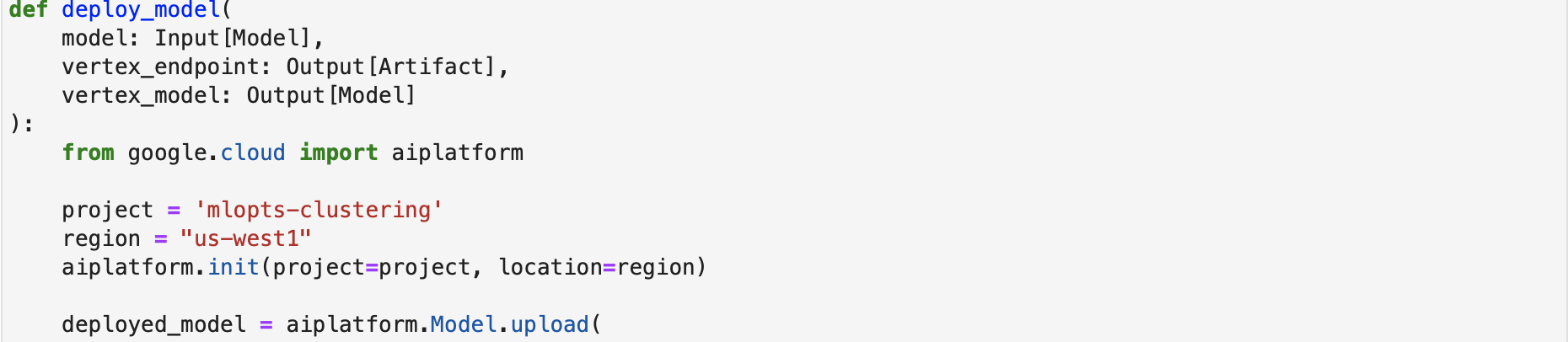
This is the step of the pipeline that trains the model and pass it to the deploy model stage.



***Fig 11.3 Deploy model component***

**Deploy Model:**

This step deploys the model and creates an endpoint for it.

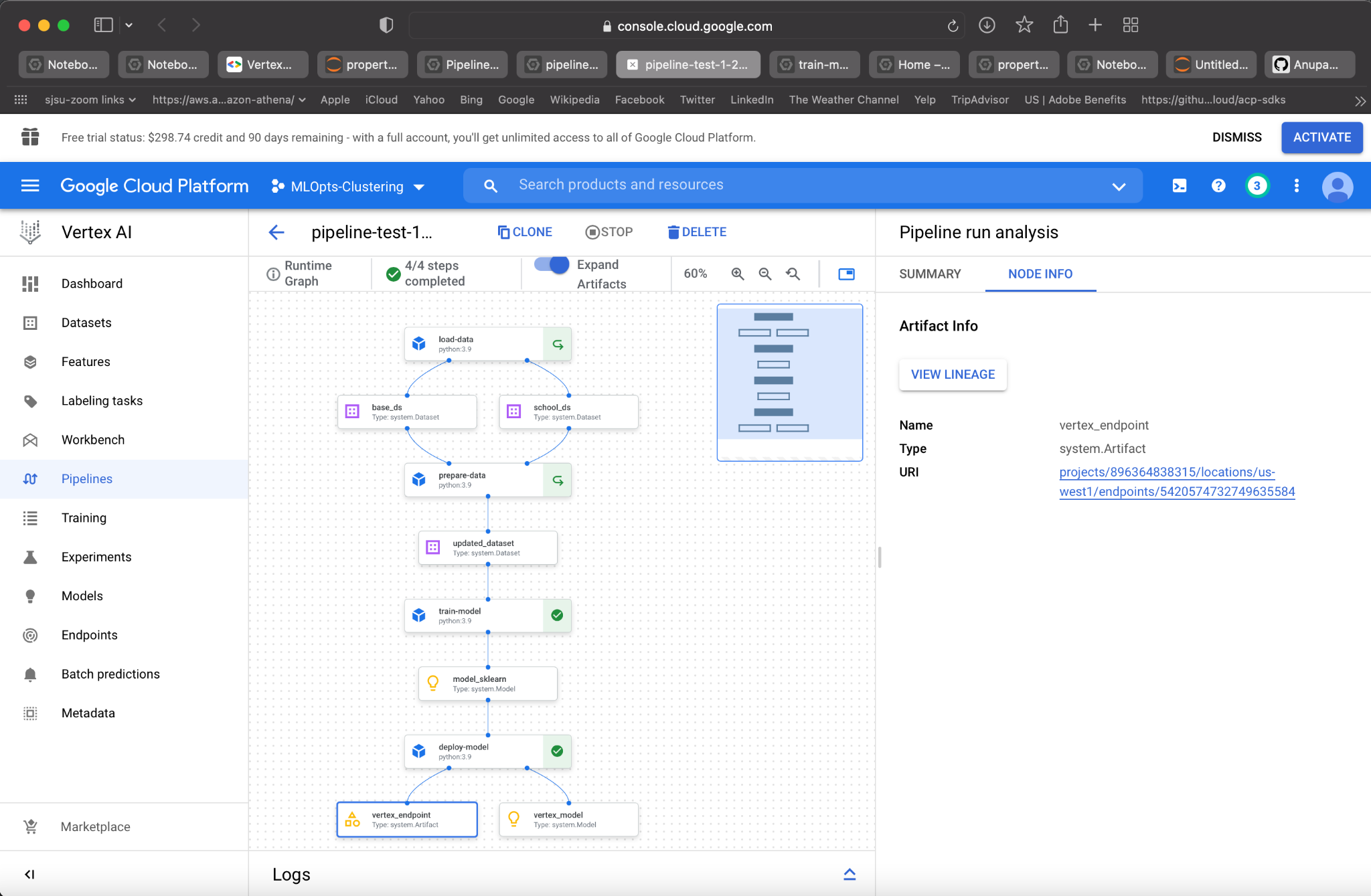


***Fig 11.4 Deploy model component***

**Pipeline:**

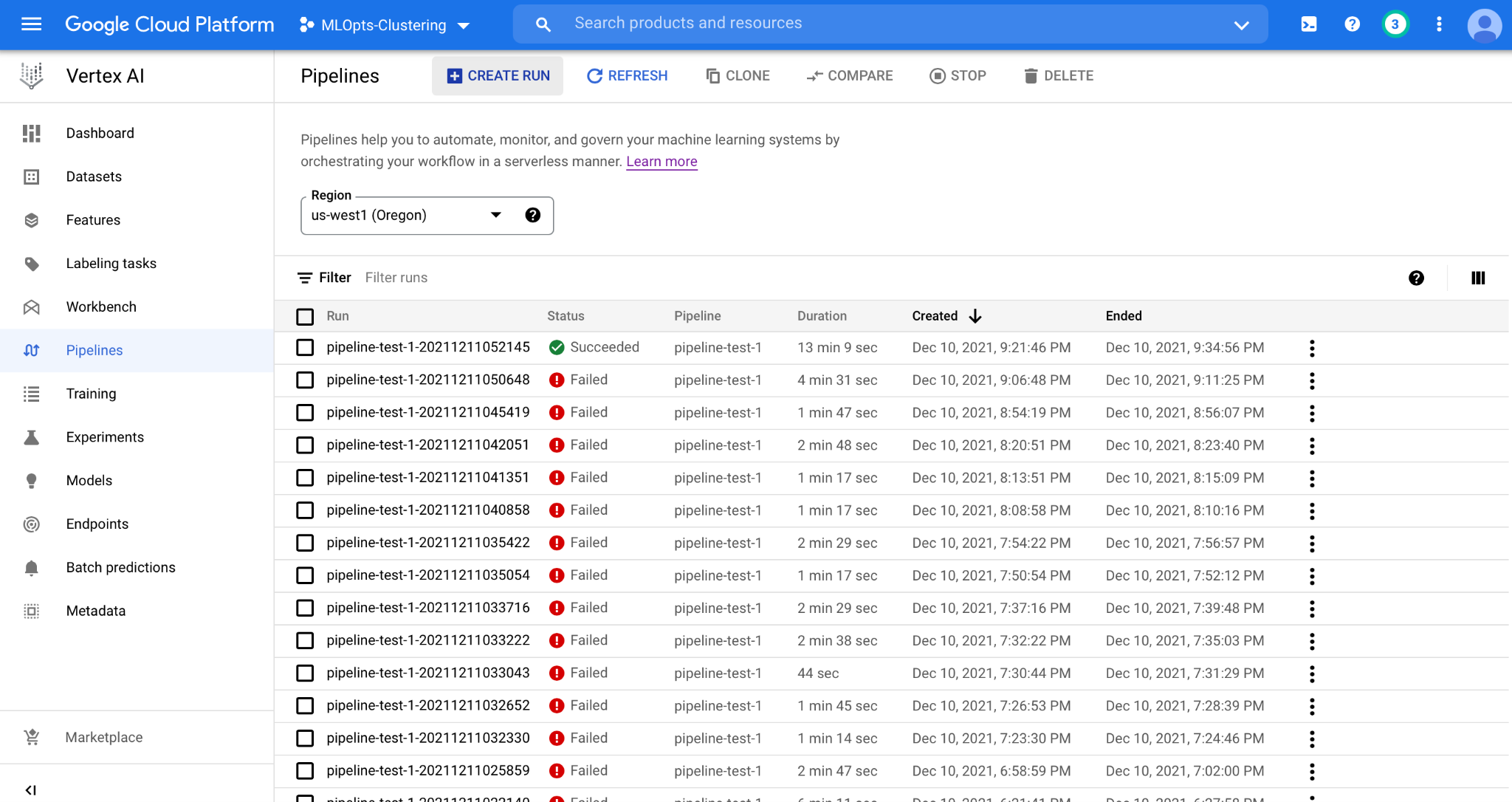
Finally pipeline combines all the four steps and executes them sequentially and deploy the model. It also creates an endpoint for the deployed mode.

Fig 11.5 shows successful pipeline deployment.



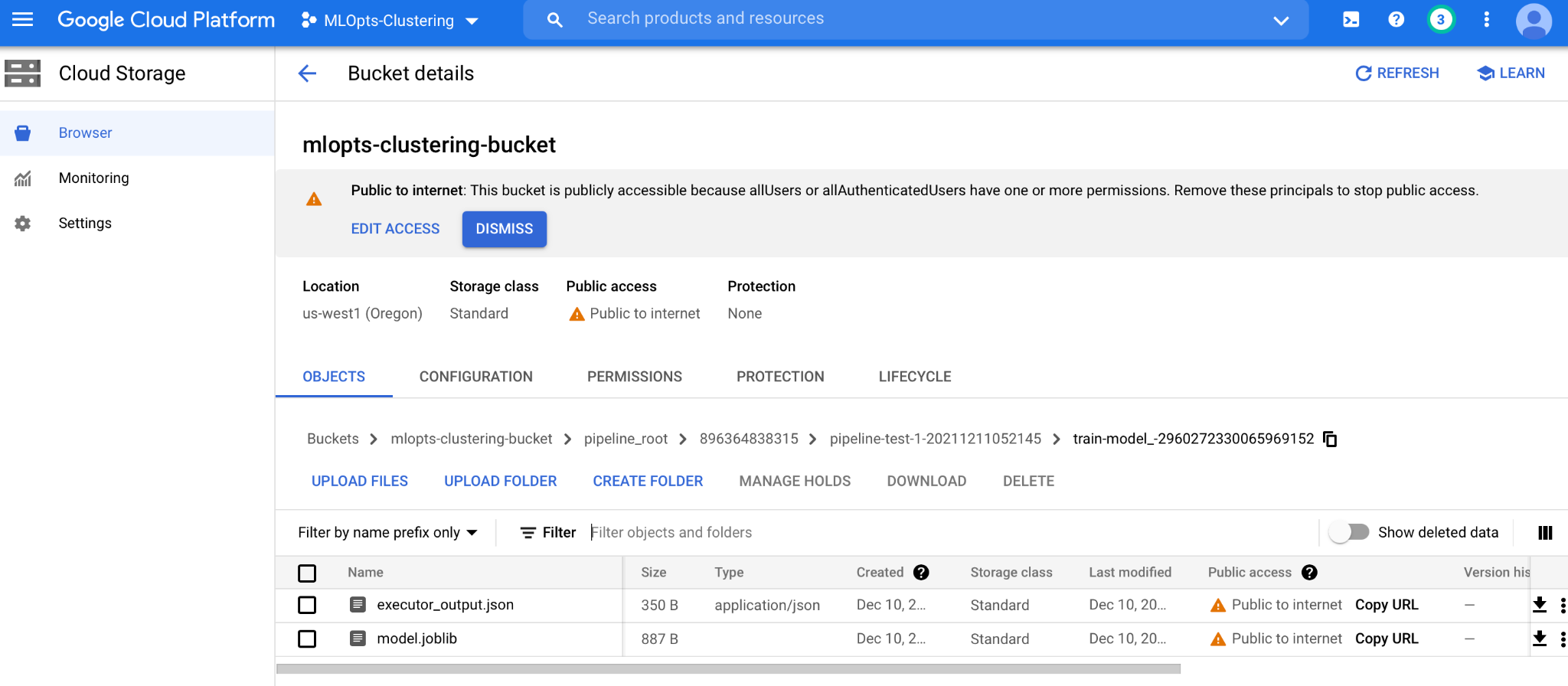
***Fig 11.5 End to end ML Pipeline***

Fig 11.6 shows the successful pipeline status.



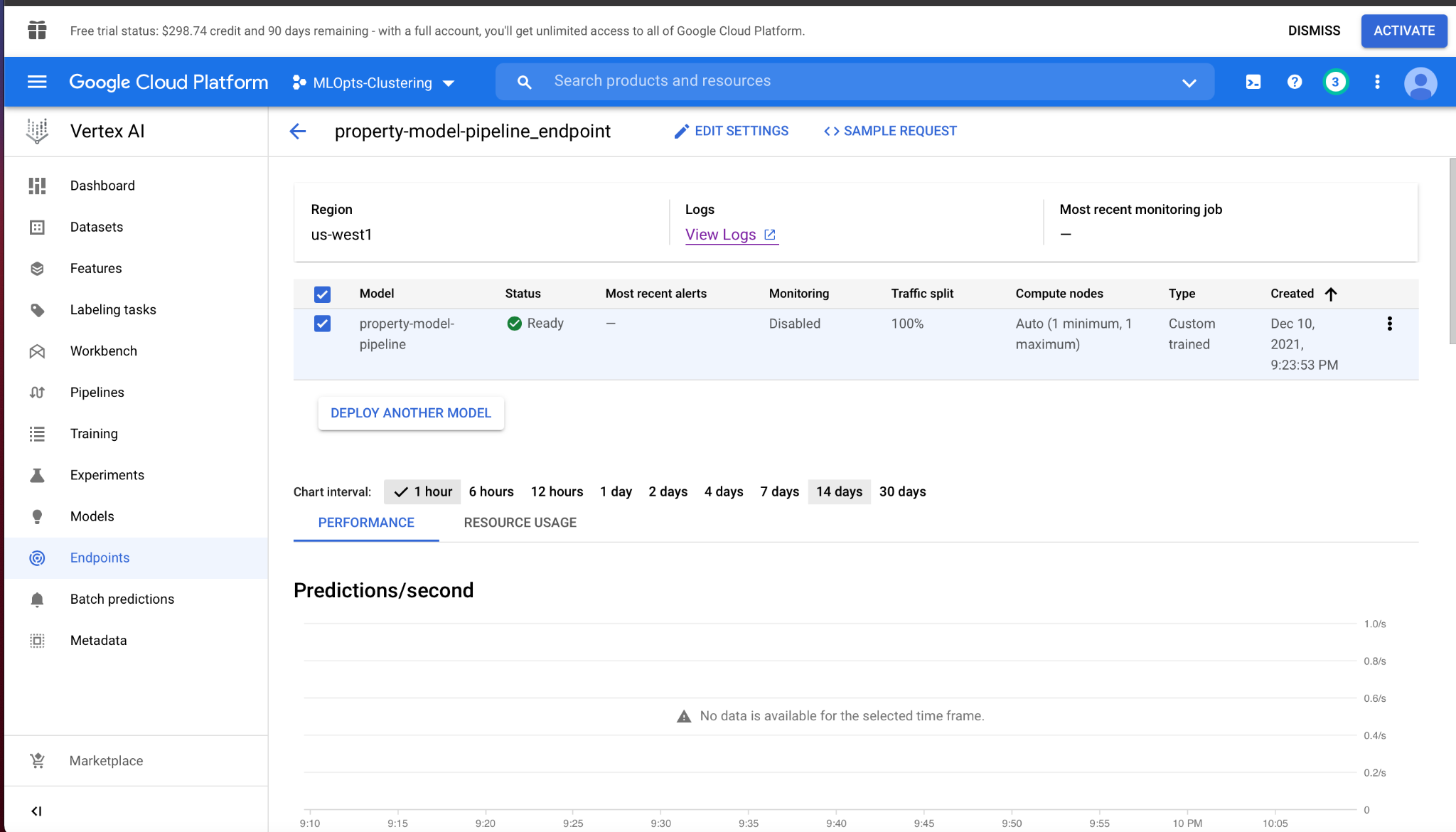
***Fig 11.6 shows the successful pipeline status.***

Fig 11.7 shows the storage bucket created for ML Pipeline and model added to it.



***Fig 11.7 Storage bucket created for ML Pipeline and model added to it.***

Fig 11.8 shows the endpoint created for the pipeline.



***Fig 11.8 Endpoint created for the model.***

## 

## 

## Chapter 12. ACKNOWLEDGMENT

This work is a part of Fall '21 Final project for CMPE-255 Sec 48 - Data Mining under Professor Vijay Eranti at San Jose State University. We thank Professor Vijay Eranti for his support and guidance throughout the course of working on this project.

# References

[1] <https://www.politieacademie.nl/kennisenonderzoek/kennis/mediatheek/PDF/86218.pdf>

[2] <https://arxiv.org/abs/1805.03620>

[3]<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3393434>