University of Strathclyde Glasgow

CS989 : BIG DATA FUNDAMENTALS COURSEWORK

REPORT – Perth House Price Prediction

Table of Contents

LIST OF FIGURES:

CHAPTER 1 : INTRODUCTION

In the UK, there is a long-standing issue with housing that has gotten worse in recent years. Politicians, analysts, academics, and the public there is an agreement that the UK is undergoing a housing crisis. Our housing issues have several essential facets, such as increased homelessness and an especially unstable private rental market. But what analysts and policymakers refer to as the housing crisis may be most notable for the drop in home ownership and record high property prices.

From few years Glasgow is also facing the same problem as the number of incoming students had increased after Covid-19. Many, students are homeless as facing the outrage of high rent due to the housing crisis. Many, were told by the universities to defer their education plan for the next year. Not only students but even other individuals are facing the issues as per data Glasgow's homeless population has increased as a result of the housing crisis as there were reportedly 1,500 homeless individuals living in the city in 2023. So, is only lack of houses is responsible for the crisis and if yes then the question arises how the government will architect the city housing plan to stop the long going crisis.

The housing crisis is complex with no easy solution. So, In this research I will be focussing on to determine how a community may plan for housing, we must first consider what characteristics make a house, what a tenant might need, and the overall pattern of homes in a region. In the pages that follow, we'll try to address these questions as well as many others.

CHAPTER 2 : DATASET AND ANALYSIS

2.1 DATASET

Since I couldn't locate a decent enough dataset for UK houses, I looked for one for Perth, Australia. Perth is one of the world's most remote towns and fourth most populated city of Australia, perched on the banks of the vast Swan River, between the Indian Ocean and the dunes of the Null arbour Desert. With lot of greenery and having the sunniest days in all the capital cities of Australia is perfect sunny town for people to enjoy.

But Everything is not perfect in the sunny town. In a metropolis of roughly 20 lac people, this year the number of homes available for rent are fewer than2500. The Shortage has surged the rent to a very high extent while increasing the cost of buying a house to 8 times the average Australian Median Income.

My aim is to understand the housing problem through the analysis of copious data and figures. We may contribute to evidence-based decision-making and policy development by spotting patterns, trends, and viable solutions. This will help to alleviate the housing crisis and advance the welfare of people and communities.

2.2 Source

The Perth Housing Dataset is used in this report which is publically available. The user report This data was scraped from following sources to gain information of 322 Perth suburbs:

a. house[dot]speakingsame[dot]com (used to get information on houses being sold in various regions according to their pincode)

b. bettereducation[dot]com (to get region wise school rankings)

c. data[dot]gov[dot]au (to get longitudinal and latitudinal data of regions in Australia

The Dataset is taken from Kaggle (https://www.kaggle.com/datasets/syuzai/perth-house-prices) it consists of several informative columns crucial for predicting the house prices, it consist of 33656 rows and 13 Columns which are :

1. Address: Street address of the house(Character variable)

2. Suburb: This is the name of the suburb (Character variable).

3.Price: Selling Price of the House (Continuous variable).

4. Bedrooms: Number of bedrooms in the house,(Integer).

5. Bathrooms: Number of bathrooms in the house(Integer).

6. Garage: Number of Garage in the house,(Integer).

7. Build_year: The year house was constructed,(Float).

8. Land_area: Size of land in square meters, (Integer)

9. Floor_area: Total Floor Area of the house(Float).

10.Cbd_dist: Distance of the house from central business district,(Integer)

11.Nearest_stn: Name of the nearest train station, (Character).

12.Nearest_stn_dist: Distance to the nearest train station,(Integer)

13.Date_sold: Date on which the property was sold, (Object).

14. Postcode – Postcode for the location of property(Integer)

14.Latitude: This is the latitudinal address of the house,(Float).

15.Longitude: This is the longitudinal address of the house,(Float).

16.Nearest_sch: Name of the nearest school ATAR (Australian Tertiary Admission Rank) approved school, (Character).

17.Nearest_sch_dist: Distance of the house to the nearest ATAR approved school in kms,(Float) 18.Nearest_sch_rank: ATAR approved rank of the school,(Float)

Chapter 3: DATASET ANALYSIS:-

3.1 Key Challenges and Problems :

Missing Values :

Missing values are common while dealing with the real-world datasets and this can become one of the key obstacles while doing the analysis. In this case there was null values in the 3 columns which were Garage , Build year , Nearest_sch_rank as shown in fig-3.1 . As, simply we can not drop the missing values as can effect our analysis so null values in the columns Garage and Build year were imputed by there mean values. As the percentage of missing values was 32.5% in Nearest_sch_rank column which was too high ,so dropped the rows consisting of null values as can effect the accuracy of the model if imputed.

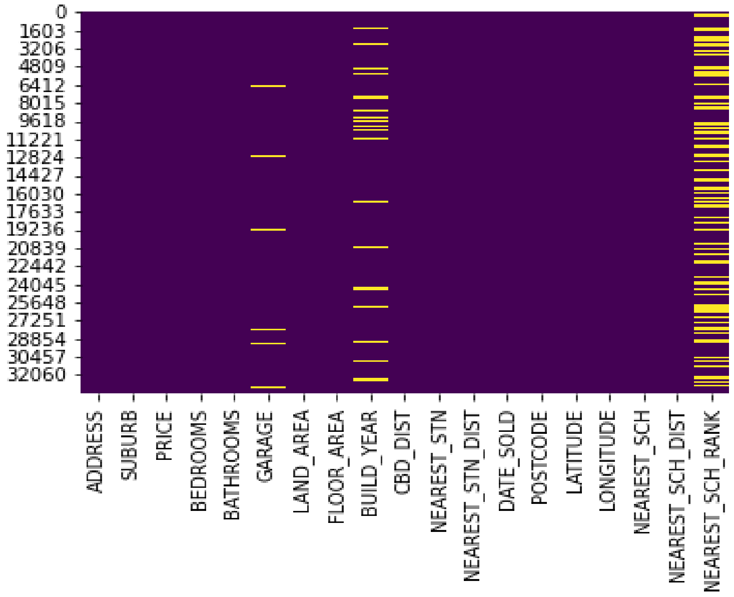

Figure 3.1 (Heatmap showing Null Values)

Date Time Format:

Because the Date column was in object format, the data were handled as strings. To fix this, I used the pandas pd.to_datetime function, which transformed the Date variable to datetime64 format.and finally the original Date column was dropped from dataset.

Columns With Categorical Data:

Columns like Address , Suburb, Nearest_stn , Nearest_Sch were dropped as they were mere text and with high cardinality so , was of not much use as we have Postcode for the location and knowing the distance only to school and station will be sufficient for our analysis.

Removing Outliers :

Few entries are very unfamiliar if we observe the maximum number of bathrooms are 16 moreover amount of garages exceeds to 99 which is impossible if a house has maximum of 10 bedrooms. So, remove all the rows having more than 5 number of garages and having more than bathrooms.

3.2 : Descriptive Analysis :

As the Price of houses depend on different factors which like geography, location, Near by Areas etc. We discovered that the median price of a house is about $545,000, and the average price is about $640,000.As average and mean are close it indicates the smooth and normal distribution which can be observed in figure 3.2.

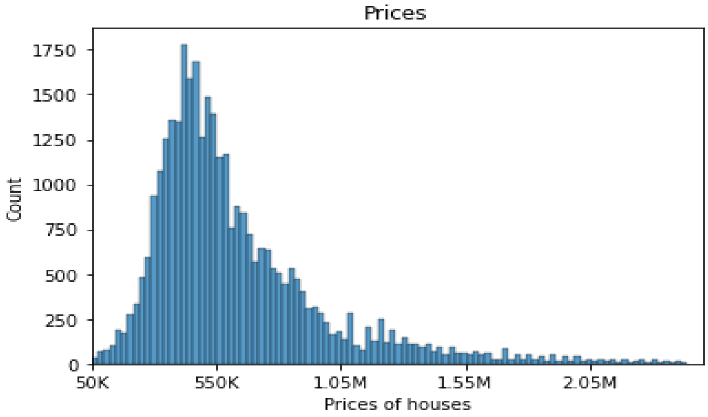

Figure 3.2

3.2.1 :Factor affecting Prices

We are interested in the Price variable and major factors affecting it. So, the basic factors in my views which can effect the selling price of the house are floor area , number of bedrooms, number of washroom's and number of garage's in the house. But if we observe (figure 3.2.1, figure 3.2.2, figure 3.2.3) it shows that price of houses was not proportional to the increase in above features but dispersion of price was quite varied.

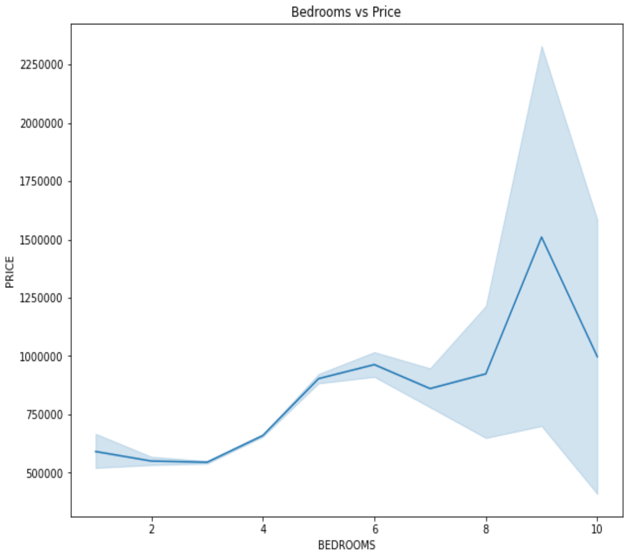

Figure 3.2.1

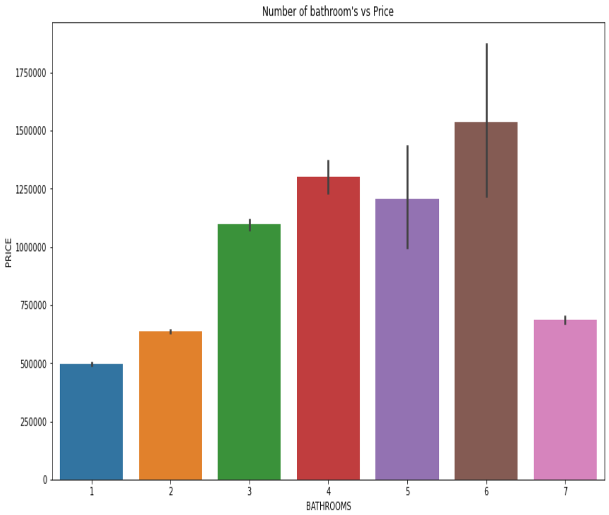

Figure 3.2.2

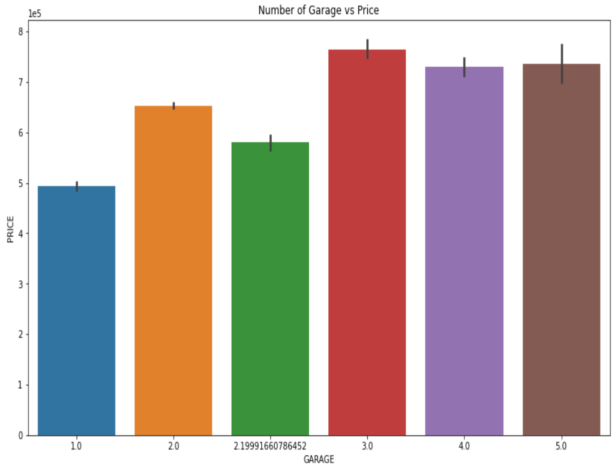

Figure 3.2.3

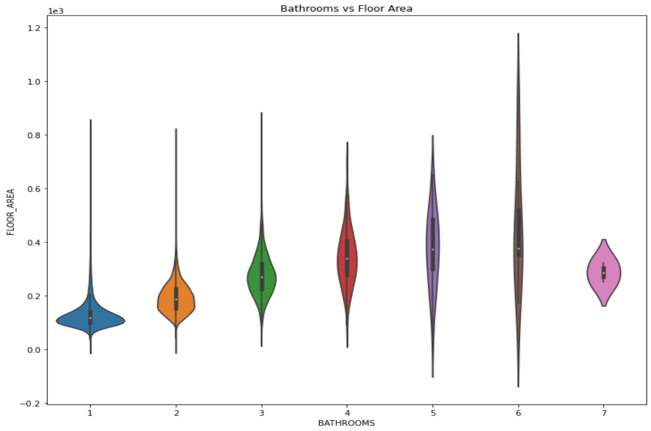

Figure 3.2.4

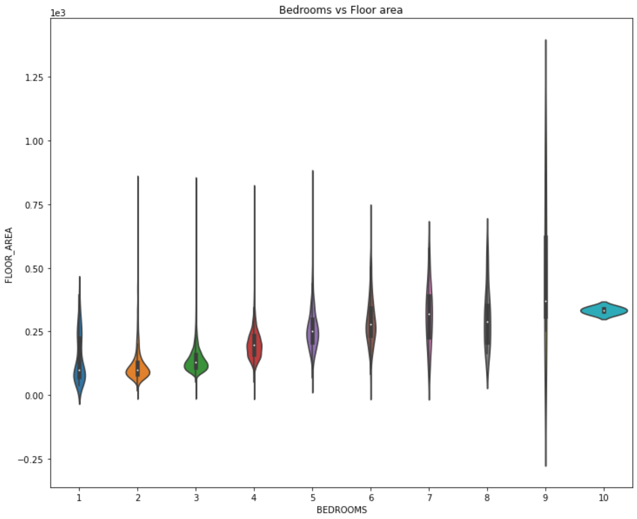

Figure 3.2.5

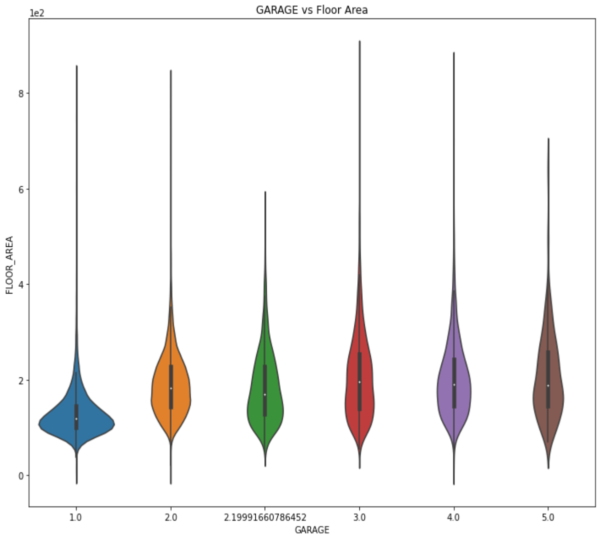

Figure 3.2.6

So, it might be a possibility that the size of a house's floor plan may affect its cost. For instance, if a house has more bedrooms, bathrooms, or a bigger garage than another with the same floor area, the price may vary or vice versa. Conversely, if two homes have the same number of bedrooms, bathrooms, and garages but different floor plans, they may be priced equally. However, this assumption also failed as we can see the violin plot in figure 3.2.4 and figure 3.2.5 as the number of bathrooms and bedrooms rise, the median of floor area also increases .

3.3 Geography :

I plotted the Latitudinal and Longitudinal coordinates of the houses to understand their geographical distribution. As, can be observed in the figure 3.3 the maximum houses are located at latitude 31 and longitude 116 and it demonstrates how geographically diverse our data is. This might imply that the information shows that different areas of Perth have varied characteristics and traits. On the other hand, less focus on specific locations may restrict our model's forecast accuracy.

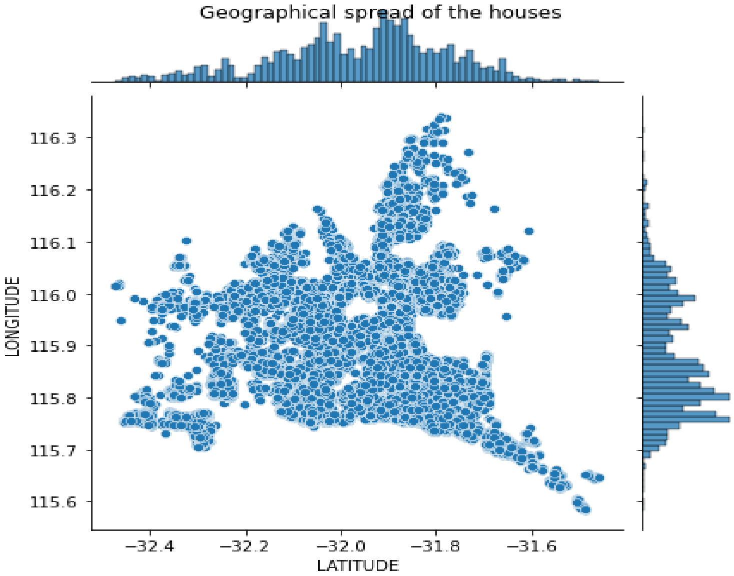

Figure 3.3

3.4: CORRELATION

In figure 3.4.1 there is heatmap showing correlation between the different features of the dataset. The floor area is correlated with the price (0.55) which is understandable as more the floor area more is the price but it is not highly correlated. Price is negatively correlated with Nearest School distance (-0.028) and is acceptable as far the school is the price of house would be lower.

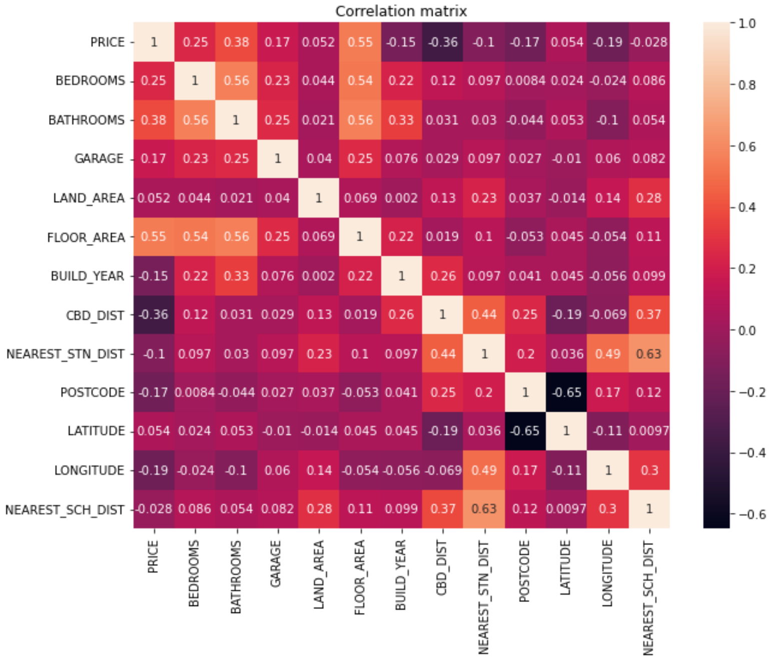

Figure 3.4.1

We should see the relationship between the floor and the price as they are most correlated with each other as it is intuitive larger the floor area more will be the price if we see it as a buyer. If we observe the figure 3.4.2 there is a linear relationship, but it doesn't seem that strong. Moreover, it has high slope which mean slight increase in floor area will increase the price at very high rate and vice versa.

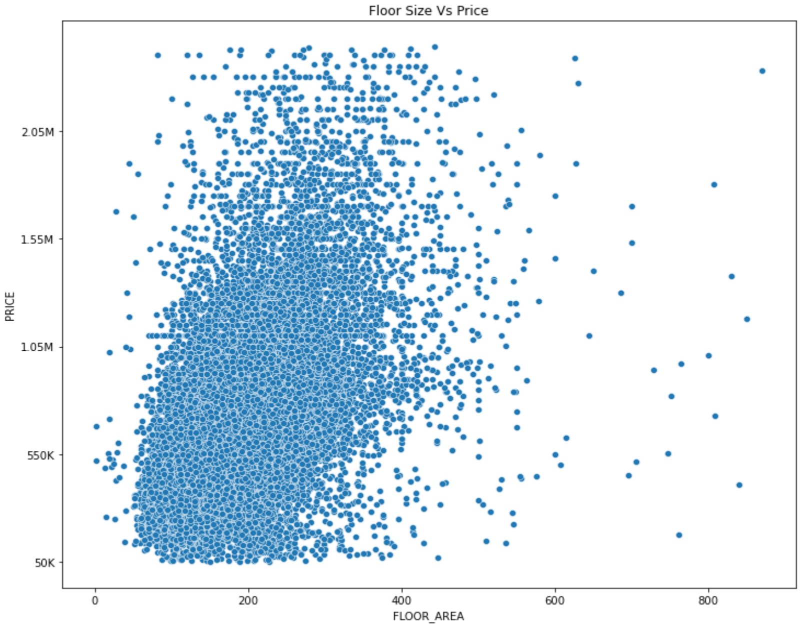

Figure 3.4.2

CHAPTER 4: Supervised Learning Model

A model is trained using a collection of labelled data in supervised learning, a type of machine learning. The output values that the model is attempting to predict are represented by the labels. In order to forecast the output labels for new input data, the model learns to link the input data with the output labels. In our case data means all the remaining quantities except the Price. The purpose of the analysis is to predict the Prices of houses.
We will train 67% of the data while 33% is our test data to know how our model performed.

4.1 Linear Regression

Linear regression is defined as an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events.
As I have split the data into test data and train data. After performing Linear regression
It was surprising that the number of bedrooms having negative impact on price while bathrooms, garages, and total floor area and land area have a positive impact on price. Another peculiarity is the distance to the nearest school and distance to nearest station both have a positive coefficient, despite the fact that, the greater the distance, the lower the price of a house should be, because it would make education possibilities less accessible, and similarly with train station as far it will be it will take more time to commute therefore the property less enticing and cheaper. The distance to common business district has negative impact as more the distance far will be the employment opportunity. The results were quite surprising which question the validity of our linear regression model. To check how our model performed we did few test :
1.R-squared value tells the proportion of variation in the dependent variable. Our R-squared value is 0.57(approx.) which is quite low

2. The values for Mean Absolute error(156050.607), Root mean squared error(231604.12611671677) which are also not up to the mark.

4.2 Random Forests Regressor

As our Linear Regression model did not perform well, I decided to go with random forests regressor for better results. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. I tried multiple number of trees(n_estimator) but after 50 estimator the model was perform similarly.

The performance of Random Forests Regressor was much better than the Linear Regression as:

1. R-squared value is 0.80 far better than of Linear Regression

2. The values for Mean Absolute error(93758.4625752683), Root mean squared error(157148.63448096954) lower than the Previous one’s.

The Random forest regressor model was good fit as compared to Linear Regression can also be seen in figure 4.2.1 and figure 4.2.2 the plot is more linear and points are fitted near the r egression line.

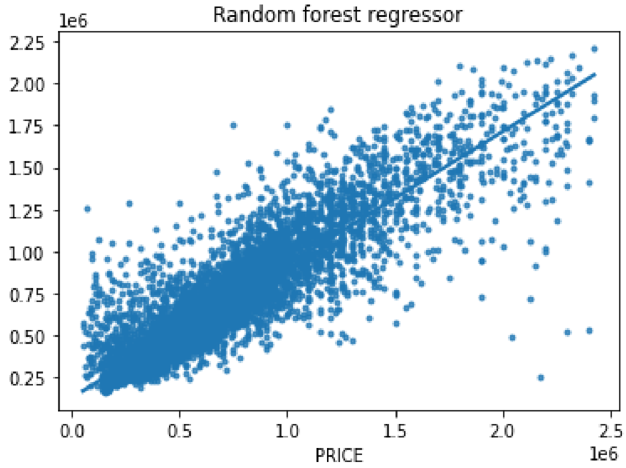

Figure 4.2.1

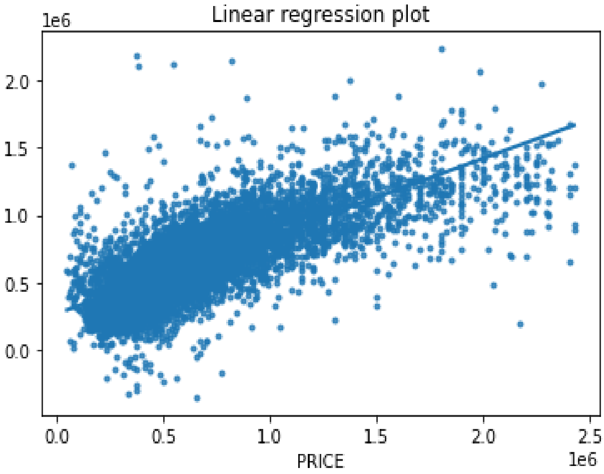

Figure 4.2.2

As the linear regression model did not performed and decision tree performed was good but in my views there are too many outliers which are creating noise in the dataset or the given features are not enough to predict the future prices of houses in Perth.

CHAPTER 5: UNSUPERVISED LEARNING MODEL

A type of machine learning called unsupervised learning involves training the model on a set of unlabelled data. Without assistance from the labels, the model develops its own ability to spot patterns in the data. As we can see, this dataset primarily was never meant to be used for unsupervised learning because it already includes the Price feature . However, to make use of this method we will drop the Price column from our dataset and treat our data as unlabelled. I will use k means clustering to identify the cluster's or various groupings of houses.

5.1 K Means Clustering

K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters, each cluster consisting of a prototype of the cluster. It minimizes within-cluster variances, but not regular Euclidean distances. A similar centroid exists for each cluster in K-Means.

As, most of the features were having different parameters so I scaled the features of the data by using Standard Scaler. Then it is very important to know the value for "k" which is the total quantity of clusters into which our data will be split. In order to get the optimal "k," I calculated the sum of squares for each successive iteration of the k-means model, with "k" values were ranging from 1 to 50.. The elbow curve was then obtained by plotting these error_ rate points against the number of clusters supplied in each model.

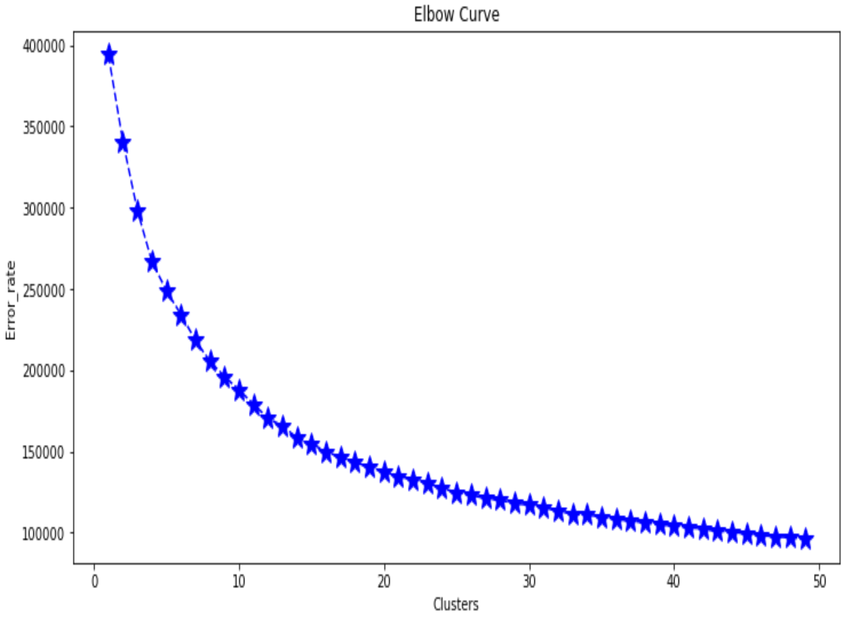

Figure 5.1.1

As, can be seen in figure 5.1.1 an elbow graph showing the elbow shape is created at 10 so appropriate numbers of clusters should be 10. Using k = 10 then checking the score if clustering is working or not.

Silhouette Score : The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It can be used to study the separation distance between the resulting clusters. I got the value of 0.27 which is not good.

Homogeneity score : It measures how much the sample in cluster are similar, it value ranges from 0 to 1. I scored 0.058 which is not good.

Completeness Score : The degree to which the data points in a cluster are comparable to one another is gauged by the completeness score in K-means clustering, the number goes from 0 to 1, with 1 indicating flawless labelling. I scored 0.17 which is again not good.

The scatter plots of the Ten clusters (figure 5.1.2, figure 5.1.3, figure 5.1.4) supplement the restrictions suggested by the robustness tests of Silhouette, Homogeneity, and Completeness score. We notice that several of the clusters are closely connected as well as there's plenty of overlap between them. There are additionally a few obvious outliers disrupting the clusters. Perhaps it is one severe constraint introduced by the data, making clustering problematic in the first place. Another cause for the clustering's failure is the indistinguishability of the characteristics, which prevents the clustering from guiding itself in the direction of pricing.

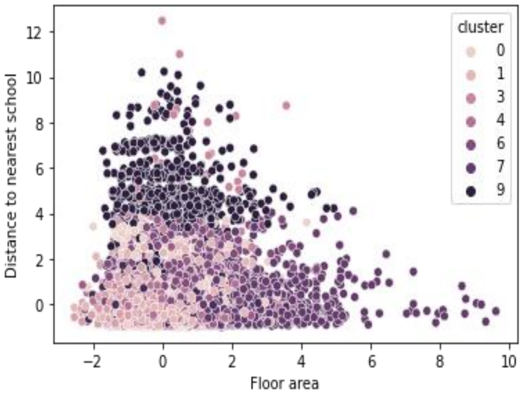

Figure 5.1.2

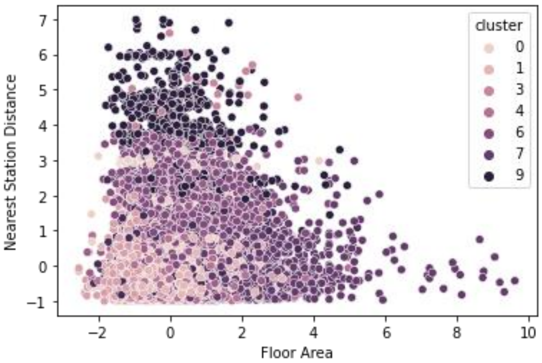

Figure 5.1.3

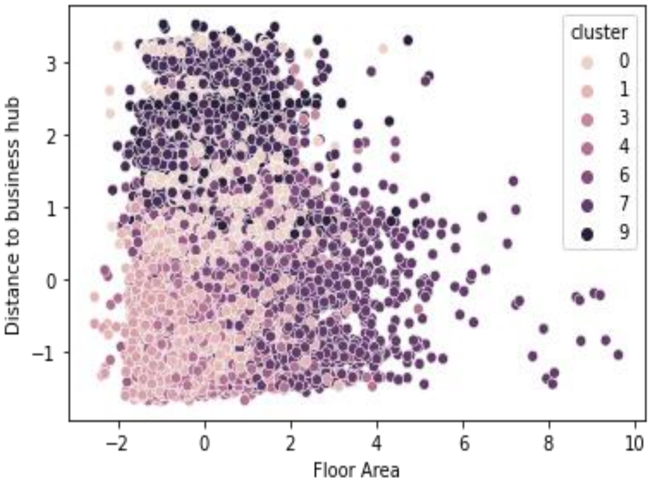

Figure 5.1.4

CHAPTER 6 : REFLECTION

The main goal of using this type of data was to create a model that could forecast the price of houses. I might have not chosen this dataset as there with data points with high noise and outliers which should be presumed by me as data was taken from public websites. However, there were certain aspects which I overlooked should have been dealt in earlier stages which could have given me more robust results even for the Linear Regression Model.

It was discovered when working with the unsupervised learning approach, K Means Clustering , that this sort of data is not suited for clustering analysis. Moreover , I should have selected only 2 or 3 features which were highly correlated to get few and well defined clusters.

If I had a second chance, I would have chosen a dataset that would have worked for both strategies.
Overall, it was a good learning experience through this exercise hence learnt about multiple data analytics tools in python.

APPENDIX:

Python Version: 3.9

Packages used:

- Numpy
- Pandas
- Matplotlib.pyplot
- Seaborn
- Sklearn
- Scipy

Bibliography

Analytics Vidhya. (2021). *Dealing With Missing Values in Python*. [online] Available at: https://www.analyticsvidhya.com/blog/2021/05/dealing-with-missing-values-in-python-a-complete-guide/.

GeeksforGeeks. (2019). *Silhouette Index – Cluster Validity index | Set 2*. [online] Available at: https://www.geeksforgeeks.org/silhouette-index-cluster-validity-index-set-2/.

Mulheirn, I. (2019). *Tackling the UK housing crisis: is supply the answer?* [online] Available at: https://housingevidence.ac.uk/wp-content/uploads/2019/08/20190820b-CaCHE-Housing-Supply-FINAL.pdf.

The Independent. (2022). *Scotland faces housing shortfall of up to 100,000 new homes, report finds*. [online] Available at: https://www.independent.co.uk/news/uk/scotland-homes-scottish-government-covid-b2032539.html.

Wikipedia Contributors (2019). *k-means clustering*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/K-means_clustering.