Regression on Real Estate Dataset

Objective

The main objective of the analysis is prediction of price per unit area variable given its features.

Introduction

Real Estate price prediction is a dataset originally compiled for regression analysis, linear regression, multiple regression, and predictive tasks. The dataset consists of purchase date, age of property, location, house price of unit area, and distance to nearest station.
The dataset is obtained from [Kaggle (https://www.kaggle.com/quantbruce/real-estate-price-prediction)](https://www.kaggle.com/quantbruce/real-estate-price-prediction) by **bruce.**

```
First five rows of dataset
```

	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
0	1	2012.917	32.0	84.87882	10	24.98298	121.54024	37.9
1	2	2012.917	19.5	306.59470	9	24.98034	121.53951	42.2
2	3	2013.583	13.3	561.98450	5	24.98746	121.54391	47.3
3	4	2013.500	13.3	561.98450	5	24.98746	121.54391	54.8
4	5	2012.833	5.0	390.56840	5	24.97937	121.54245	43.1

```
There are 414 rows and 8 columns in dataset.
```

Descriptive statistics

Descriptive statistics are those that summarizes the dataset and provide some quick insights.

```
Summary table
```

	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude
count	414.000000	414.000000	414.000000	414.000000	414.000000	414.000000	414.000000
mean	207.500000	2013.148971	17.712560	1083.885689	4.094203	24.969030	121.533361
std	119.655756	0.281967	11.392485	1262.109595	2.945562	0.012410	0.015347
min	1.000000	2012.667000	0.000000	23.382840	0.000000	24.932070	121.473530
25%	104.250000	2012.917000	9.025000	289.324800	1.000000	24.963000	121.528085
50%	207.500000	2013.167000	16.100000	492.231300	4.000000	24.971100	121.538630
75%	310.750000	2013.417000	28.150000	1454.279000	6.000000	24.977455	121.543305
max	414.000000	2013.583000	43.800000	6488.021000	10.000000	25.014590	121.566270

The mean price of unit area of house is $ 37.98K

Data Cleaning

Data Cleaning is the process of converting the raw data into the form which aids in the process of analysis. Data Cleaning methodology is as follows:-

1. Changing column names.
2. Handling missing values.
3. Feature Engineering
4. Dropping Irrelevent attributes.
5. Dropping Duplicate rows.

1. Changing column names

In dataset, with each column name some additional information is also associated signifying its role in prediction analytics. In this step of data cleaning column names are changed to appropriate ones.

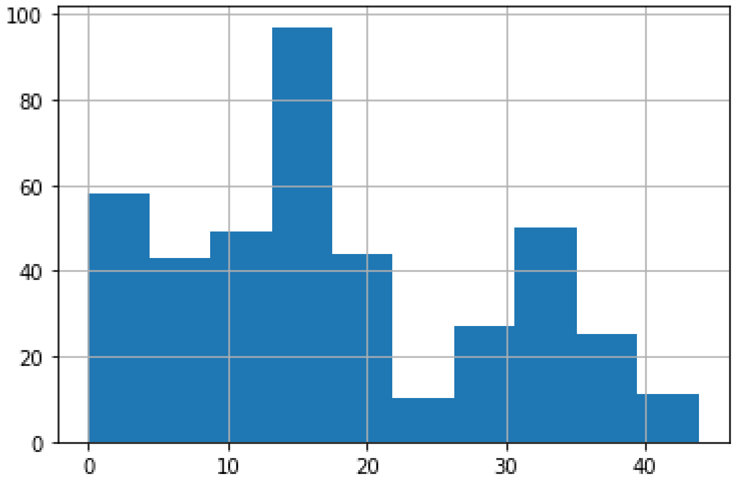
	No	Transaction date	Age	Distance from nearest MRT station	No. of convenience stores	Latitude	Longitude	Price/Area
0	1	2012.917	32.0	84.87882	10	24.98298	121.54024	37.9
1	2	2012.917	19.5	306.59470	9	24.98034	121.53951	42.2
2	3	2013.583	13.3	561.98450	5	24.98746	121.54391	47.3
3	4	2013.500	13.3	561.98450	5	24.98746	121.54391	54.8
4	5	2012.833	5.0	390.56840	5	24.97937	121.54245	43.1

2. Handling missing values

Looking for missing values in the dataset

```
No                                 0
Transaction date                   0
Age                                0
Distance from nearest MRT station  0
No. of convenience stores          0
Latitude                           0
Longitude                          0
Price/Area                         0
dtype: int64
```

```
<matplotlib.axes._subplots.AxesSubplot at 0xef4f230>
```



There are no missing values in dataset. But there are some observations in which Age is numbered as 0. Since Age of 0 doesn't make sense, we will remove those rows.

	No	Transaction date	Age	Distance from nearest MRT station	No. of convenience stores	Latitude	Longitude	Price/Area
0	1	2012.917	32.0	84.87882	10	24.98298	121.54024	37.9
1	2	2012.917	19.5	306.59470	9	24.98034	121.53951	42.2
2	3	2013.583	13.3	561.98450	5	24.98746	121.54391	47.3
3	4	2013.500	13.3	561.98450	5	24.98746	121.54391	54.8
4	5	2012.833	5.0	390.56840	5	24.97937	121.54245	43.1

3. Feature Engineering

Adding new feature Year in dataset

	No	Transaction date	Age	Distance from nearest MRT station	No. of convenience stores	Latitude	Longitude	Price/Area
0	1	2012.917	32.0	84.87882	10	24.98298	121.54024	37.9
1	2	2012.917	19.5	306.59470	9	24.98034	121.53951	42.2
2	3	2013.583	13.3	561.98450	5	24.98746	121.54391	47.3
3	4	2013.500	13.3	561.98450	5	24.98746	121.54391	54.8
4	5	2012.833	5.0	390.56840	5	24.97937	121.54245	43.1

4. Dropping irrelevent attributes

Dropping irrelevent features like 'No' and 'Transaction date' which do not help in analysis.

	Age	Distance from nearest MRT station	No. of convenience stores	Latitude	Longitude	Price/Area	Year
0	32.0	84.87882	10	24.98298	121.54024	37.9	2012
1	19.5	306.59470	9	24.98034	121.53951	42.2	2012
2	13.3	561.98450	5	24.98746	121.54391	47.3	2013
3	13.3	561.98450	5	24.98746	121.54391	54.8	2013
4	5.0	390.56840	5	24.97937	121.54245	43.1	2012

5. Dropping Duplicate rows

Removing duplicate rows from dataset which have same Age, Distance from nearest MRT station, No. of convenience stores and Year.

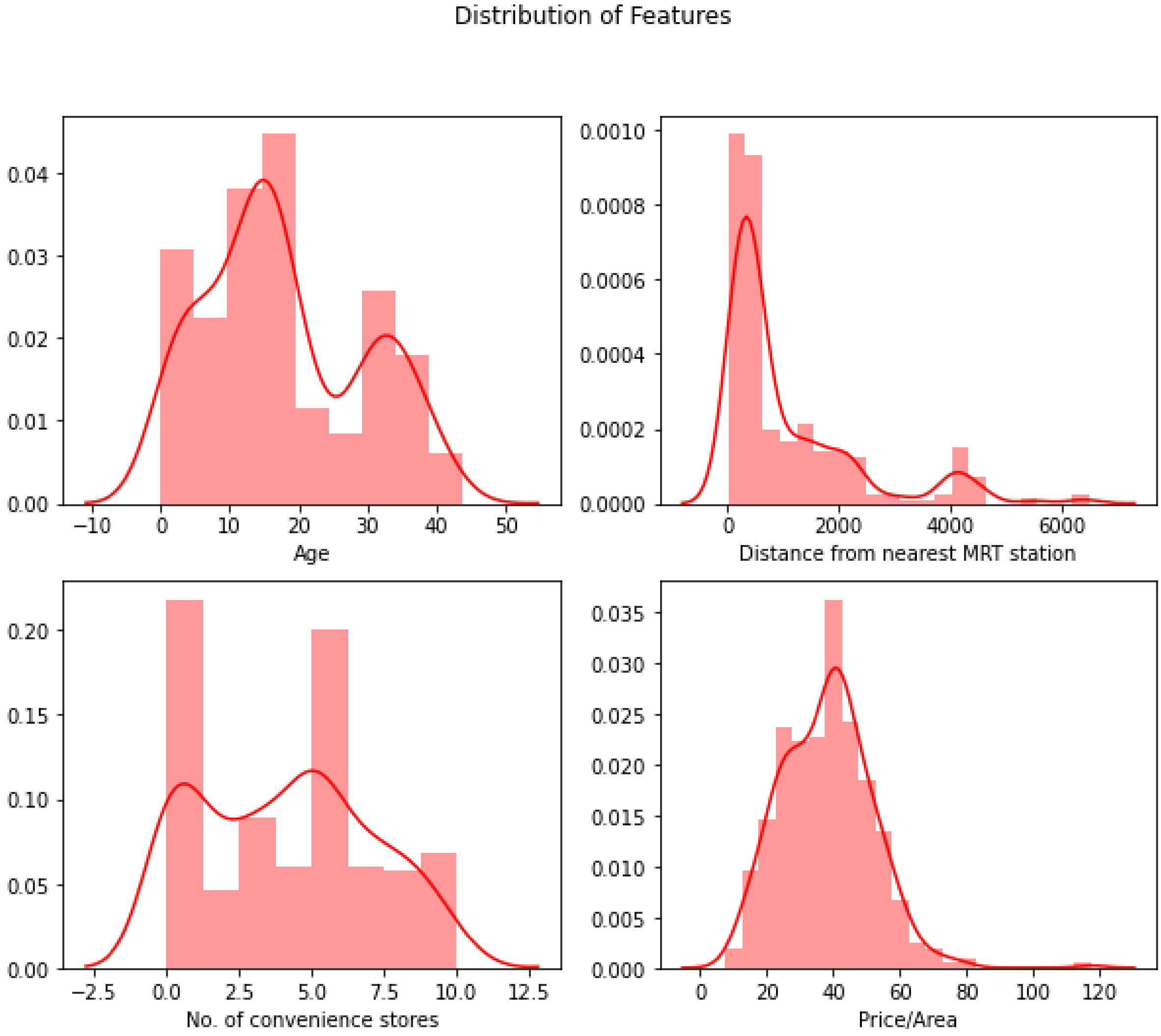
```
After removing duplicate rows there are 364 rows and 7 columns in dataset
```

Exploratory Data Analysis

```
First five rows of dataset
```

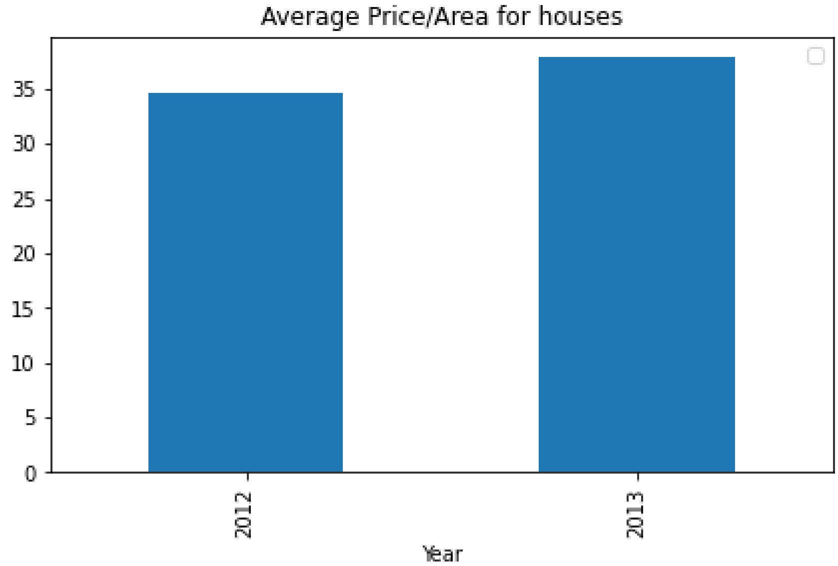
	Age	Distance from nearest MRT station	No. of convenience stores	Latitude	Longitude	Price/Area	Year
0	32.0	84.87882	10	24.98298	121.54024	37.9	2012
1	19.5	306.59470	9	24.98034	121.53951	42.2	2012
2	13.3	561.98450	5	24.98746	121.54391	47.3	2013
3	5.0	390.56840	5	24.97937	121.54245	43.1	2012
4	7.1	2175.03000	3	24.96305	121.51254	32.1	2012

```
Distribution of features
Text(0.5, 0.98, 'Distribution of Features')
```



The distribution of age shows that in dataset the observations having Age between 20 to 30 is quite less. The distribution of Distance from nearest MRT station is skewed. The distribution of No. of convenient stores and Price / Area is quite normal.

```
Average price of unit Area for houses
<matplotlib.legend.Legend at 0x145027f0>
```



The above plot shows that the average price of unit area of the property has got increased in year 2013 from year 2012.

```
Top 5 Positions having largest Price/Area in year 2012
   Latitude  Longitude  Year  Price/Area
0  24.95836  121.53756  2012        57.8
1  24.97528  121.54541  2012        57.4
2  24.98085  121.54391  2012        56.2
3  24.98419  121.54243  2012        55.1
4  24.98343  121.53762  2012        54.4

Top 5 Positions having largest Price/Area in year 2013
   Latitude  Longitude  Year  Price/Area
0  24.97460  121.53046  2013       117.5
1  24.97703  121.54265  2013        78.3
2  24.97071  121.54069  2013        78.0
3  24.96756  121.54230  2013        67.7
4  24.97345  121.54093  2013        63.3
```

Both years do not intersect at any Location. In 2012 Location having Latitude and Longitude (24.95836, 121.53756) is having highest Price/Area. In 2013 Location having Latitude and Longitude (24.97460, 121.53046) is having highest Price/Area.

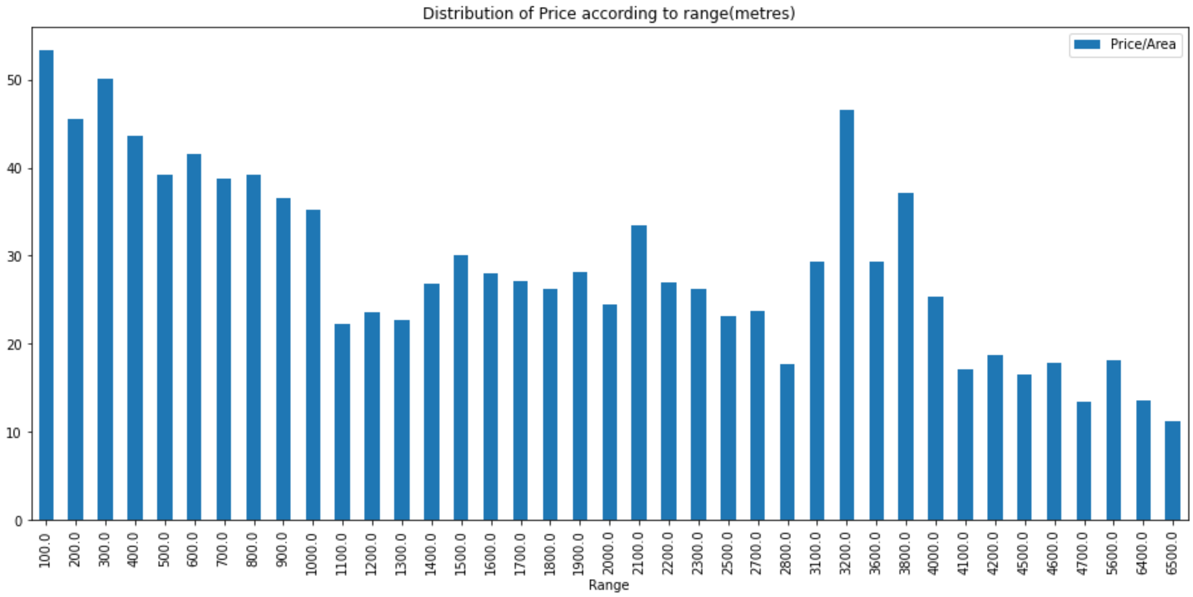
```
Bottom 5 Positions having smallest Price/Area in year 2012
   Latitude  Longitude  Year  Price/Area
0  24.94297  121.50342  2012        15.0
1  24.94235  121.50357  2012        14.7
2  24.94960  121.53018  2012        13.8
3  24.94968  121.53009  2012        13.7
4  24.94925  121.49542  2012        13.2

Bottom 5 Positions having smallest Price/Area in year 2013
   Latitude  Longitude  Year  Price/Area
0  24.93885  121.50383  2013        13.0
1  24.94935  121.53046  2013        12.8
2  24.94375  121.47883  2013        12.2
3  24.95719  121.47353  2013        11.2
4  24.96172  121.53812  2013         7.6
```

Both years do not intersect at any Location. In 2012 Location having Latitude and Longitude (24.94925, 121.49542) is having smallest Price/Area. In 2013 Location having Latitude and Longitude (24.96172, 121.53812) is having smallest Price/Area.

Relationship between Distance from nearest MRT station and Average Price/Area

Creating new feature Range that represents the distance of nearest MRT station in steps of 100.

```
Text(0.5, 1.0, 'Distribution of Price according to range(metres)')
```



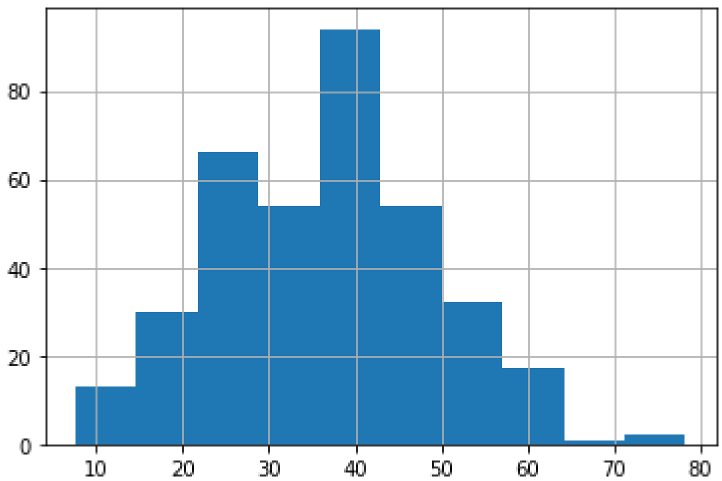
From the above plot it can be inferred that with every 100m increase in distance to nearest MRT station Price/Value decreases. But there are also some special cases where average Price/Value is also large for larger distance like for 3100m to 4100m.

Modelling

Determining Normality of Target Variable

1. Visually

```
<matplotlib.axes._subplots.AxesSubplot at 0x1077cab0>
```



It does look normal.

2. Statistical Test

```
NormaltestResult(statistic=2.5108032516775682, pvalue=0.2849613773865552)
```

The p-value is greater than 0.05 so we fail to reject the null hypothesis that it is normal. So, target variable is normal.

Testing Regression

Baseline: Linear Regression

```
R2 score for Baseline Linear Regression is  0.524339082629021
```

Polynomial Regression(Degree = 2)

```
R2 score for Polynomial Regression of degree 2 is 0.5902648469495535
```

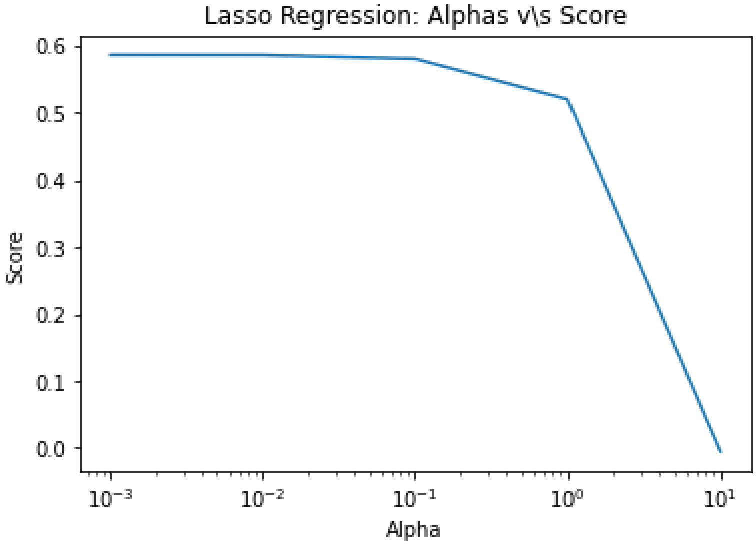
Polynomial Regression(Degree = 3)

```
R2 score for Polynomial Regression of degree 3 is 0.37798514715488524
```

Since R2 score of model with polynomial features with degree 2 is greater, we will tune it to achive higher accuracy.

Lasso Regression

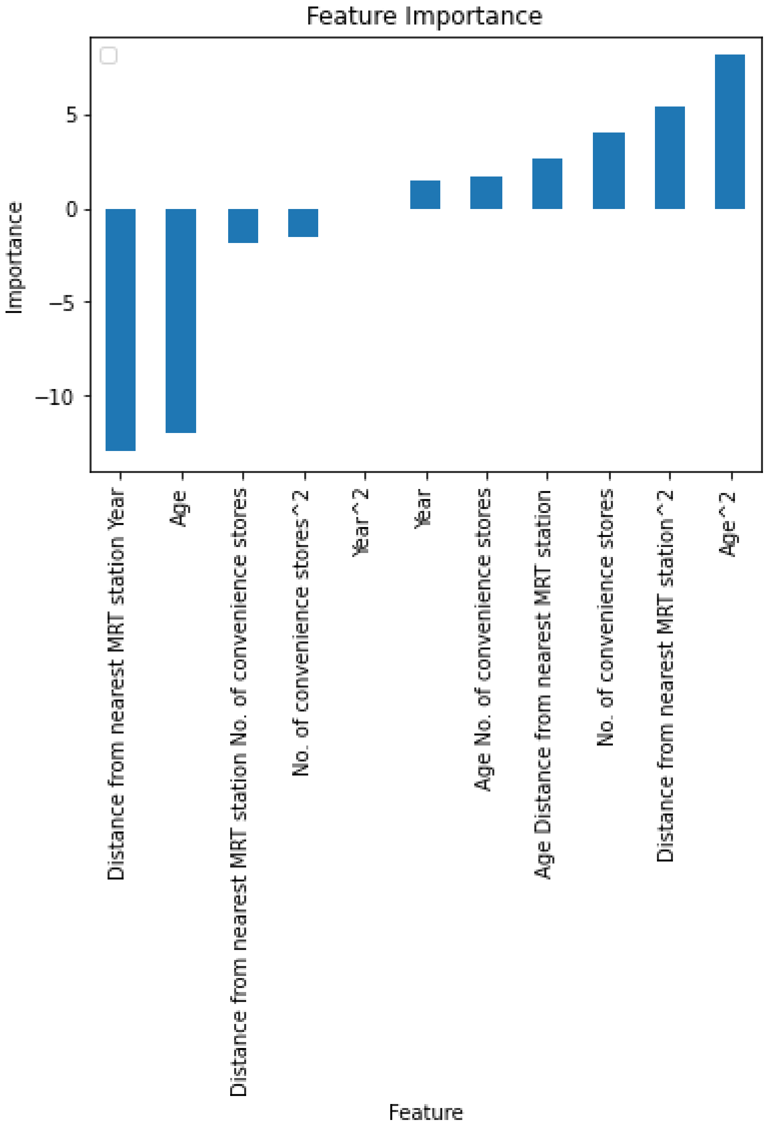
```
Text(0, 0.5, 'Score')
```




```
R2 score for Lasso Regression for alpha = 0.001 is  0.6059722423896299
```

The best model for prediction according to accuracy is Lasso Regression with hyperparameter alpha = 0.01. It gives a R2-score of about 0.60 i.e., 60% of the variability in outcome variable is explained by model.

```
<matplotlib.legend.Legend at 0x147dac10>
```



Conclusions

The features which reduces the price per unit area of property are:

1. Interaction of Year and Distance from nearest MRT station.
2. Age.
3. Interaction of No. of convenience stores and Distance from nearest MRT station.
4. Square of No. of convenience stores

The features that increases the price per unit area of property are:

1. Square of Age.
2. Square of Distance of Nearest MRT station
3. No. of convenience stores.
4. Interaction of Age and Distance from nearest MRT station.
5. Interaction of No. of convenience stores and Age

6. Year

Future Work

The analysis can be extended by analysing the price values at different latitudes and longitudes, through spatial analysis by collecting more data. Modelling can be extended by trying more linear regression models like Ridge Regression which prevents overfitting.