National University of Computer and Emerging Sciences

Plot Price Prediction based on Zameen.pk

Group Members

Saad Waseem	19L-1003
Muhammad Hasaan	19L-1011
Laiba Gohar	19L-2367

FAST School of Computing

National University of Computer and Emerging Sciences

Lahore, Pakistan

1. Dataset Overview

The dataset consists of different CSV files based on locations in Lahore that are present in Zameen. pk. The dataset consists of the following significant fields after cleaning.

- Type (Commercial or Residential)
- Price
- Location
- Area
- Creation Date
- Amenities
- Instalment Values

1.1 Data Gathering

The data was scraped using the Beautiful Soup Library of Python. The target was Lahore hence all the available areas gathered are from Lahore. Link to the plots of every location was saved and opened one by one to scrape the data. The whole process took about 4 days. Threading was used to speed up the process of data gathering.

1.2 Data Cleaning and Wrangling

This section explains the data cleaning and wrangling part.

- Unnecessary columns were dropped (Baths, Purpose and Bed)
- Amenities were sorted so that they could be categorically encoded
- The *Punjab, Lahore* part was removed from the location since this is common in all rows
- The price was converted to a standard Rs unit (Lakh and Crore were dropped)
- The area was converted to a standard marla unit
- The datatypes of the price and area were converted to numeric
- Advance payment, monthly instalments and remaining instalments were calculated and added as separate columns
- Instalment value was standardised (Lakh, Crore and Thousand were removed)
- The instalments were merged into a single column
- Amenities were categorically encoded
- Price, area and instalment value were all standardised

2. Exploratory Data Analysis

This section covers the exploratory data analysis of the dataset. This was done for deliverable 2. First, every feature will be looked at individually. Bivariate data analysis will follow that discussion. In the end, findings, observations and conclusions will be presented.

2.1 Individual Feature Exploration

This sub-section will cover the analysis of the features individually.

2.1.1 Price

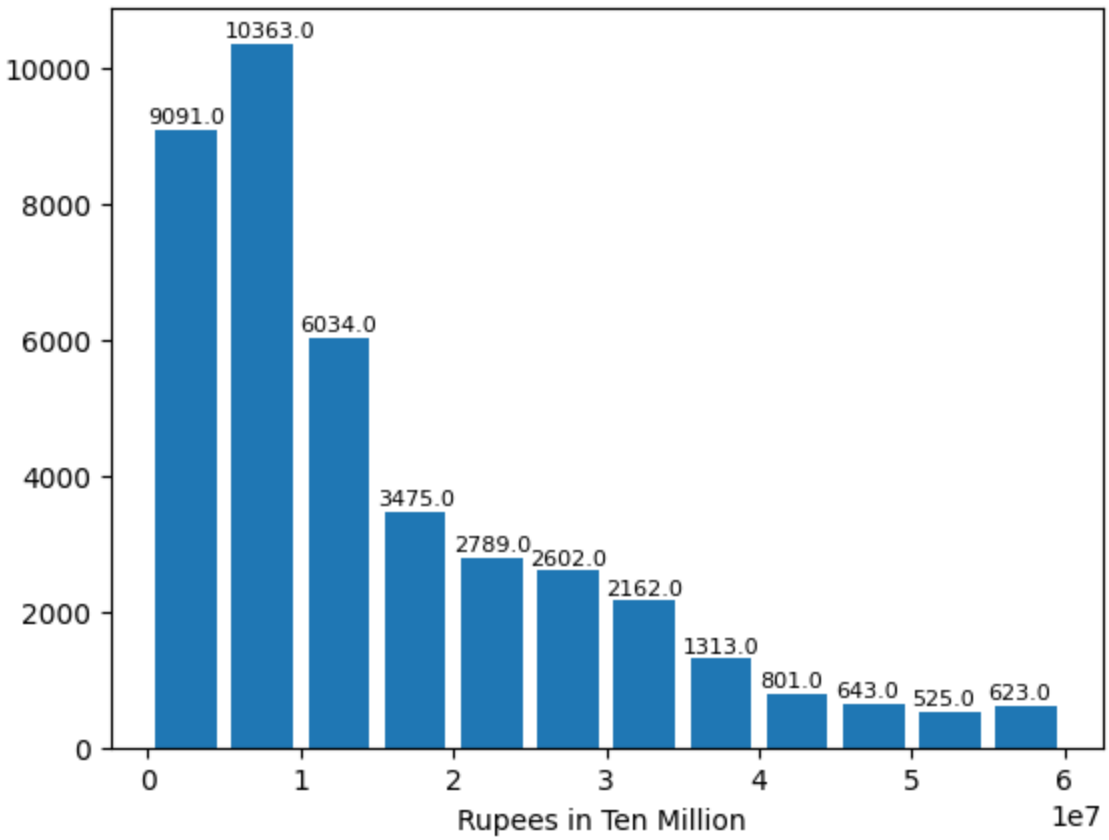
Price describes the amount required to make the purchase. This is what we will try to predict at the end of the project.

2.1.1.1 Summary Statistics

Property	Value
Mean	23268700
Min	100000
Max	990000000
25%	5500000
50%	11200000
75%	26000000
Standard Deviation	44330470

2.1.1.2 Histogram

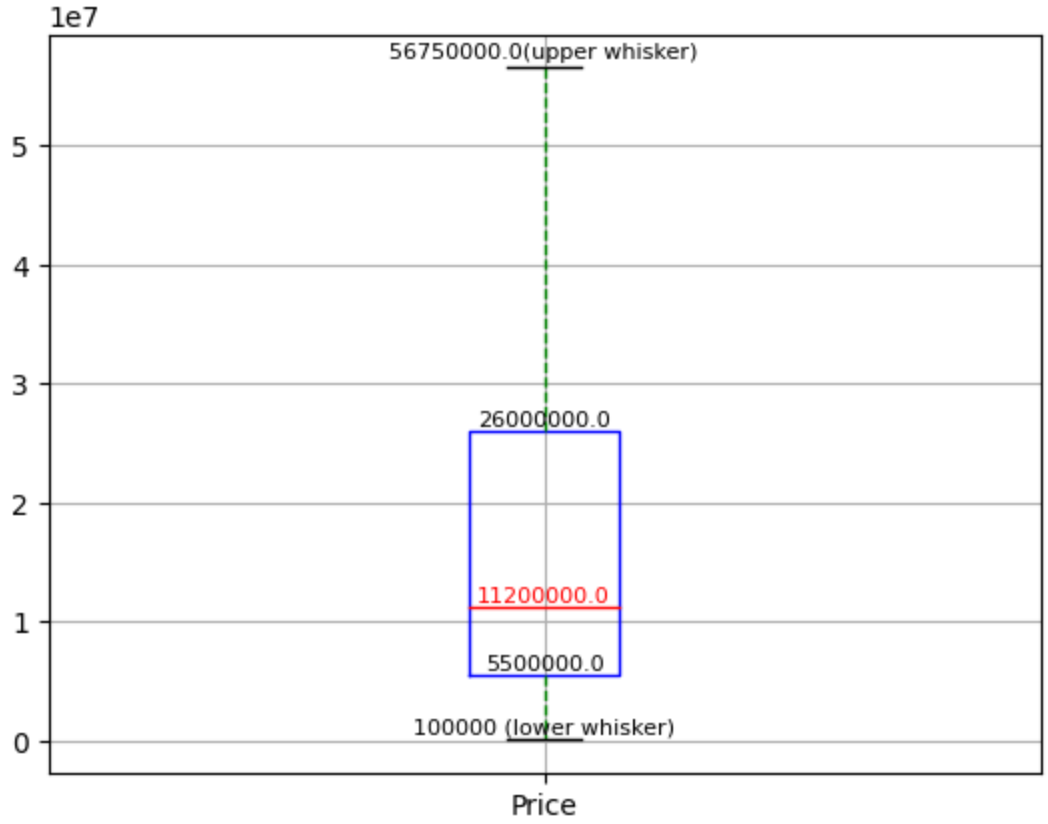
A histogram of the Price column was made using the matplotlib library. The bins are of size 5 million or 50 lakhs. The last bin represents every plot having a price > 60 million.



As it is evident, the graph is right-skewed. One more thing that is evident is that most plots are in the range of 50 lakhs - 100 lakhs or 5 - 10 million.

2.1.1.3 Boxplot

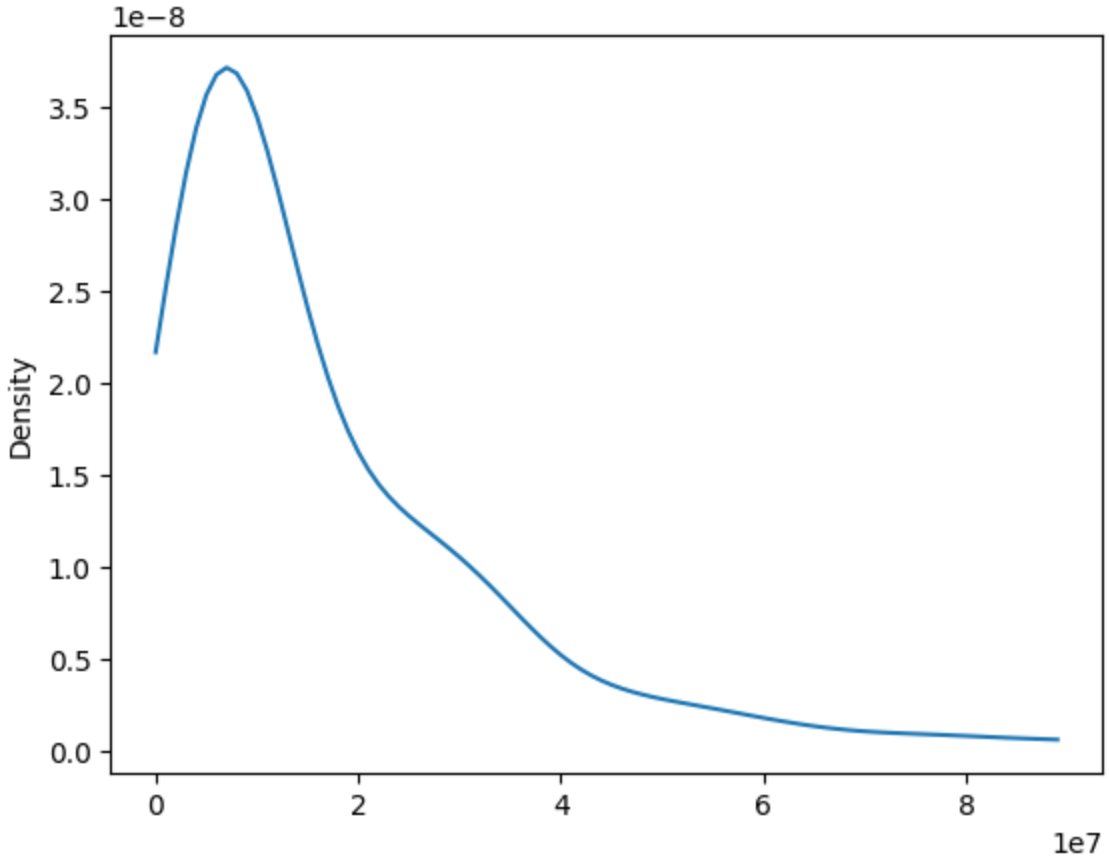
Another way to analyze the price column is to make a boxplot. Again, matplotlib was used to display a boxplot.



Here, we can see the values of the mean, Q3, Q1 and the whiskers. This boxplot again shows us that the data is skewed to the right. Both the boxplot and histogram for price suggest that the data is right-skewed, it is safe to assume that the outliers are affecting the statistics too much.

2.1.1.3 Density Graph Price

The density graph for the price is again conveying that the data is right-skewed. It also tells us that most plots are listed on Zameen. pk are under 20 million or 2 Crore.



2.1.2 Area

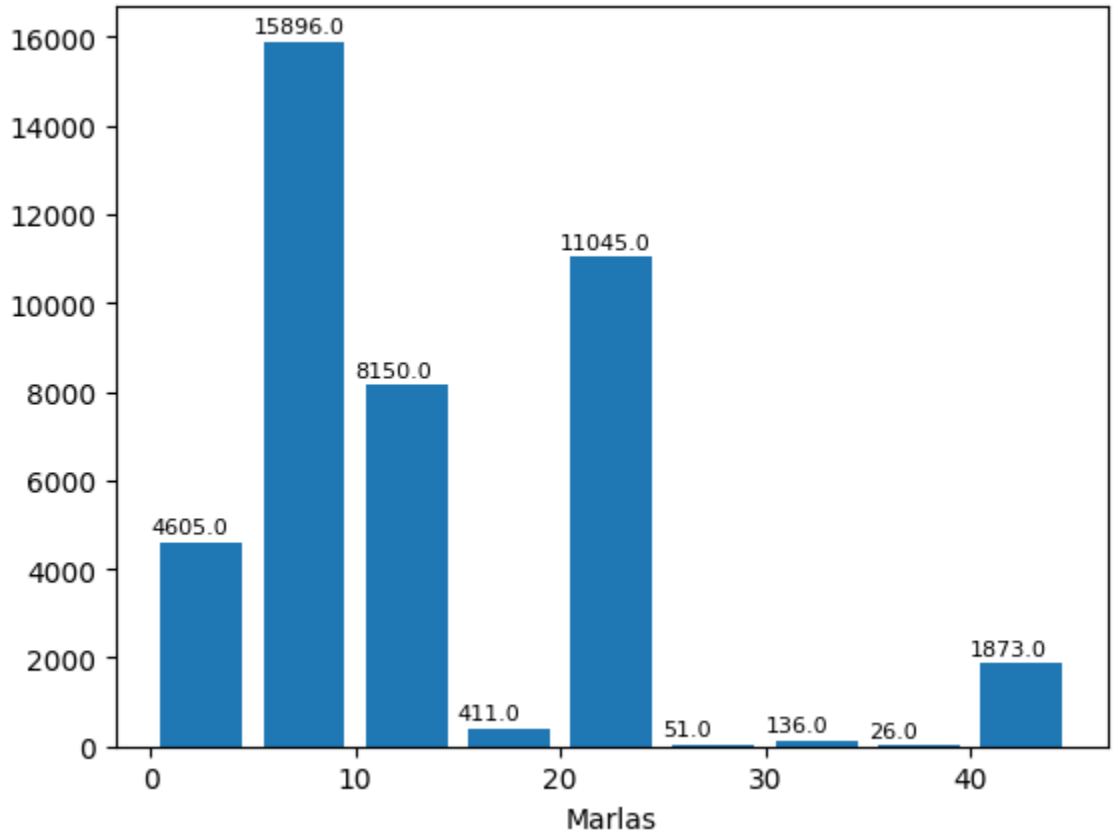
Area describes the land area of the plot. The unit of area used is Marlas.

2.1.2.1 Summary Statistics

Property	Value (In Marlas)
Mean	22.419965
Min	0
Max	19200
25%	5
50%	10
75%	20
Standard Deviation	241.82

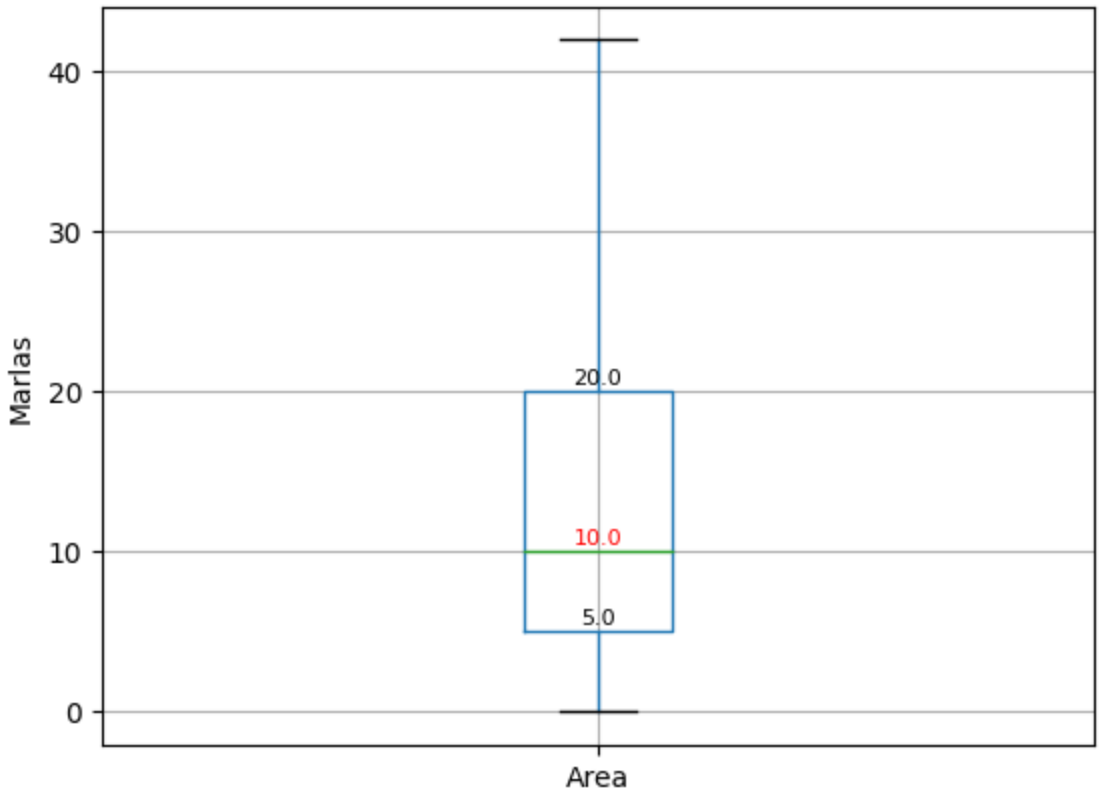
2.1.2.2 Histogram

The bins are of 5 marlas. It is evident that the most number of plots listed on the website are between 5-10 marlas. Spikes can be seen at 5 Marla, 10 Marla, 20 Marlas (1 Kanal) and 40 Marlas (2 Kanal). This is because, in Pakistan, these are the standard plot sizes. Other than that we can also see a spike in the first bin due to the increase of 3 Marla plots in recent years.



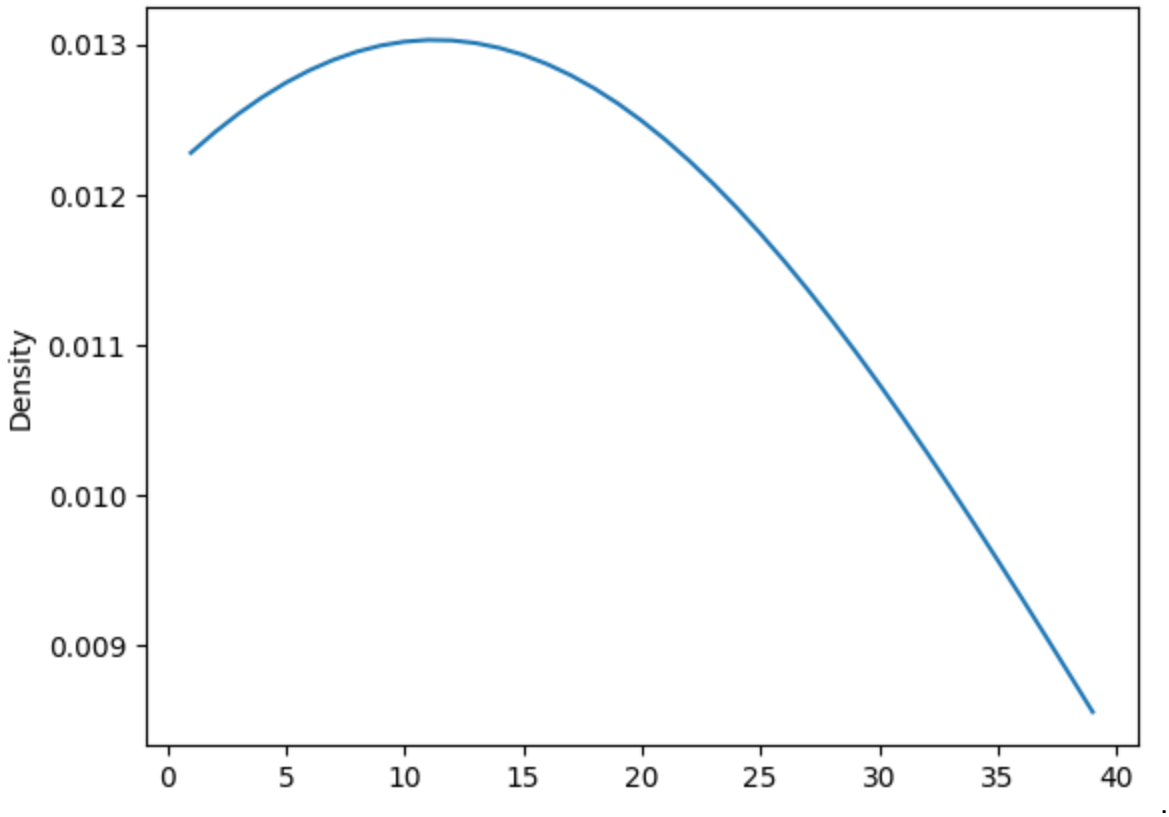
2.1.2.3 Boxplot

The boxplot gives us the Q1, Q2, Q3 and the whiskers. We can assume that any plot > 40 marlas (2 Kanal) are an outlier.



2.1.2.3 Density Graph Price

The density graphs tell that the most number of plots available are 10 marlas. The curve is almost elliptical in nature which tells us that 5 and 10 Marla plots are the most famous.



2.1.3 Instalment Value

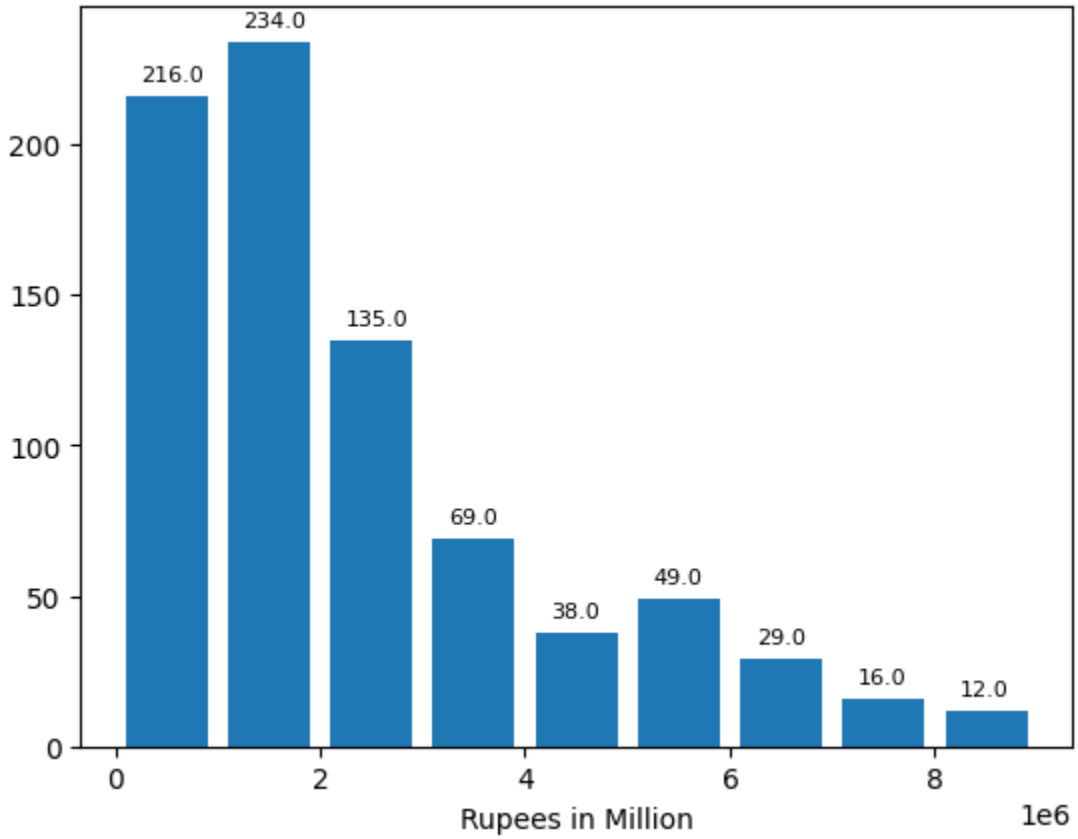
The instalment value gives the total amount due in instalments.

2.1.3 Summary Statistics

Property	Value (In Marlas)
Mean	4353525
Min	0
Max	100000000
25%	1024000
50%	1965000
75%	4378000
Standard Deviation	8229414

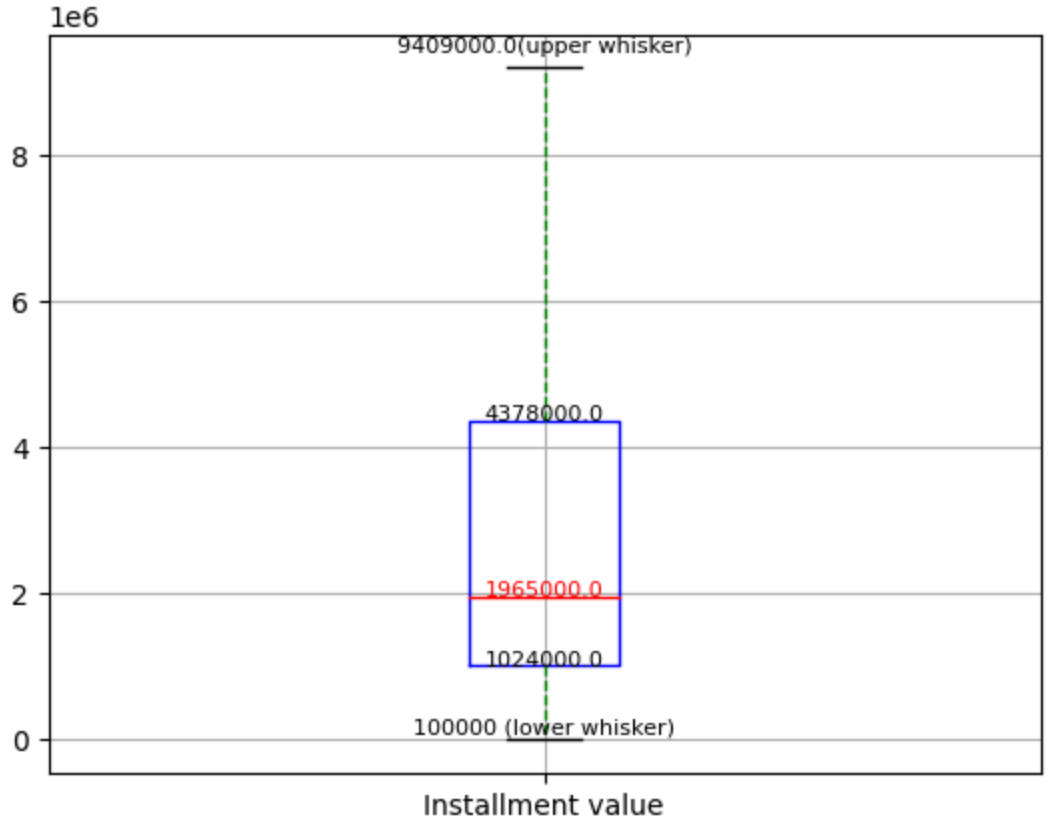
2.1.3.2 Histogram

According to this histogram, most of the plots have instalment value less than 2 Million (2 Lakhs). The graph is again right-skewed.



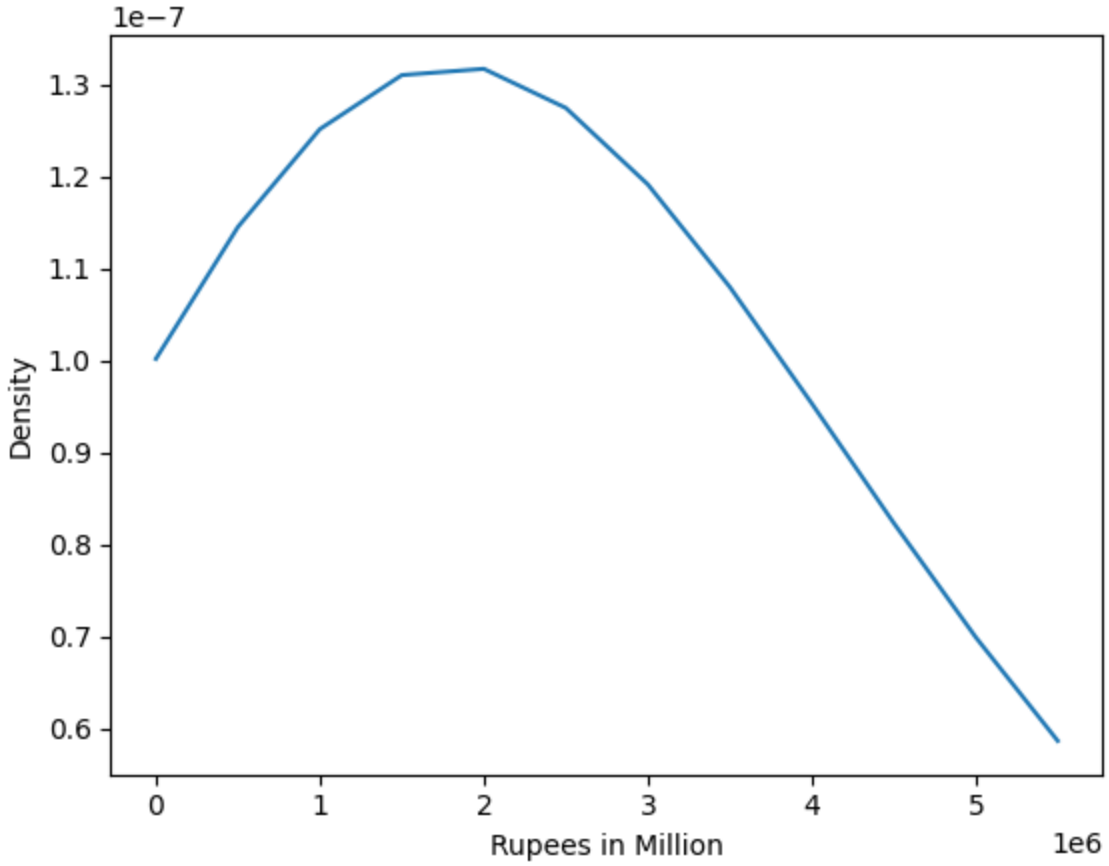
2.1.3.3 Boxplot

The boxplot indicates that the data is right-skewed. It also tells us that any value over 95000000 is an outlier.



2.1.3.3 Density Graph Price

The density graph displays that on average, the instalment value for a plot lies between 1-3 million. The density is low at the start and it falls off gradually as the value of the instalment increases.

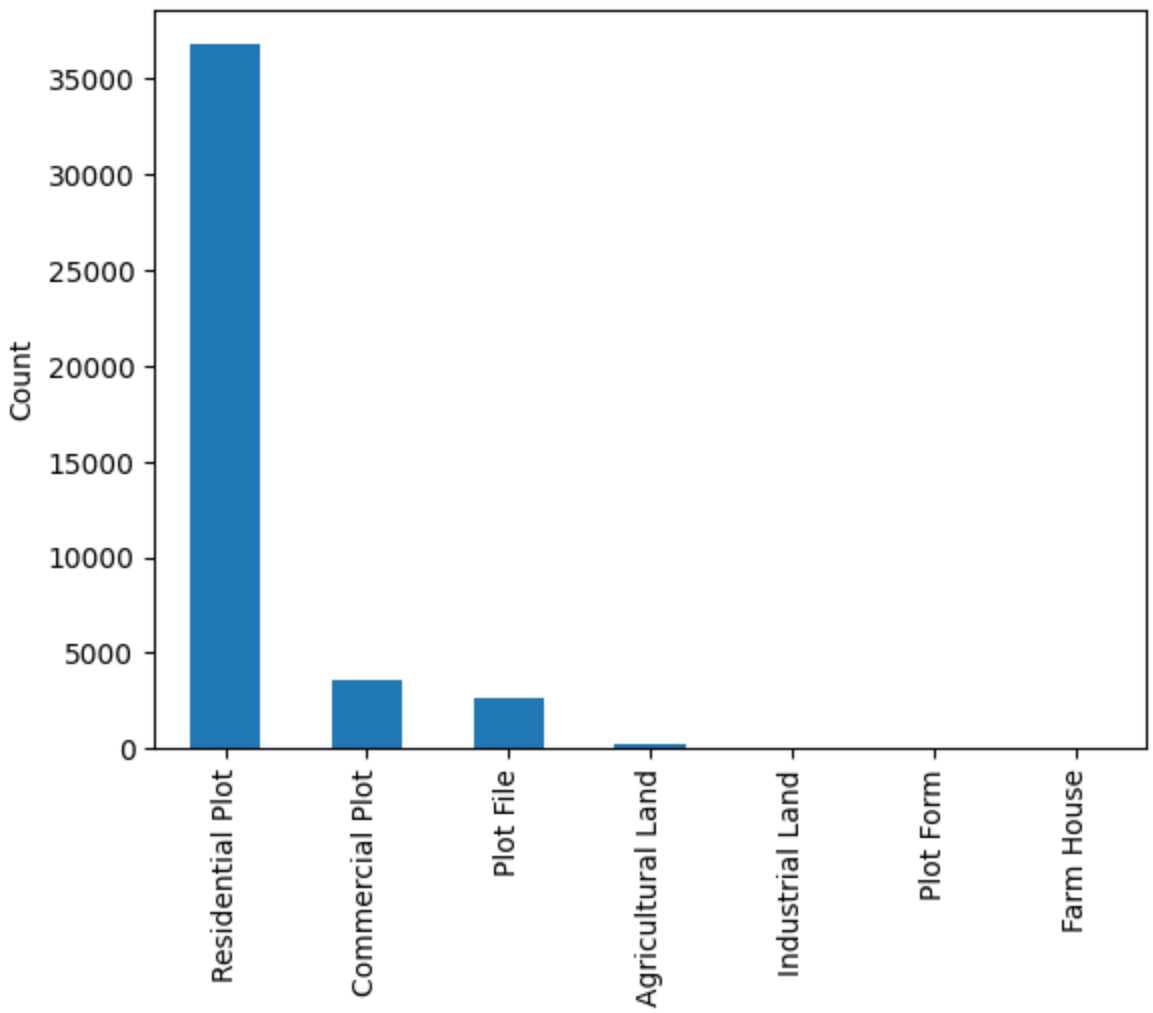


2.1.4 Type of Plot

Type indicates the type of the plot. We can see that a vast majority of the plots are Residential. It is almost 85% while Commercial Plots and Plot Files make up ~15% of total plots.

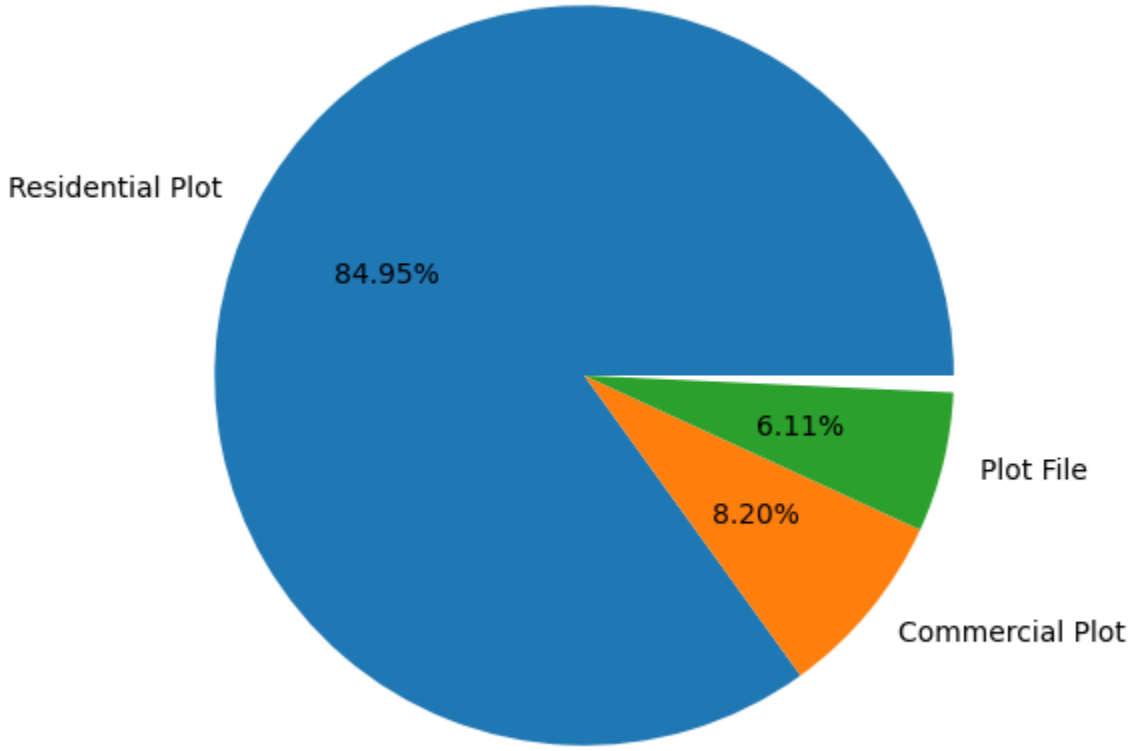
2.1.4.1 Bar Graph Displaying Counts

This graph displays the count of the types of plots. It can be seen that Plot Form, Industrial Plot and Farm House are negligible


Count
35000
30000
25000
20000
15000
10000
5000
0
Residential Plot
Commercial Plot
Plot File
Agricultural Land
Industrial Land
Plot Form
Farm House

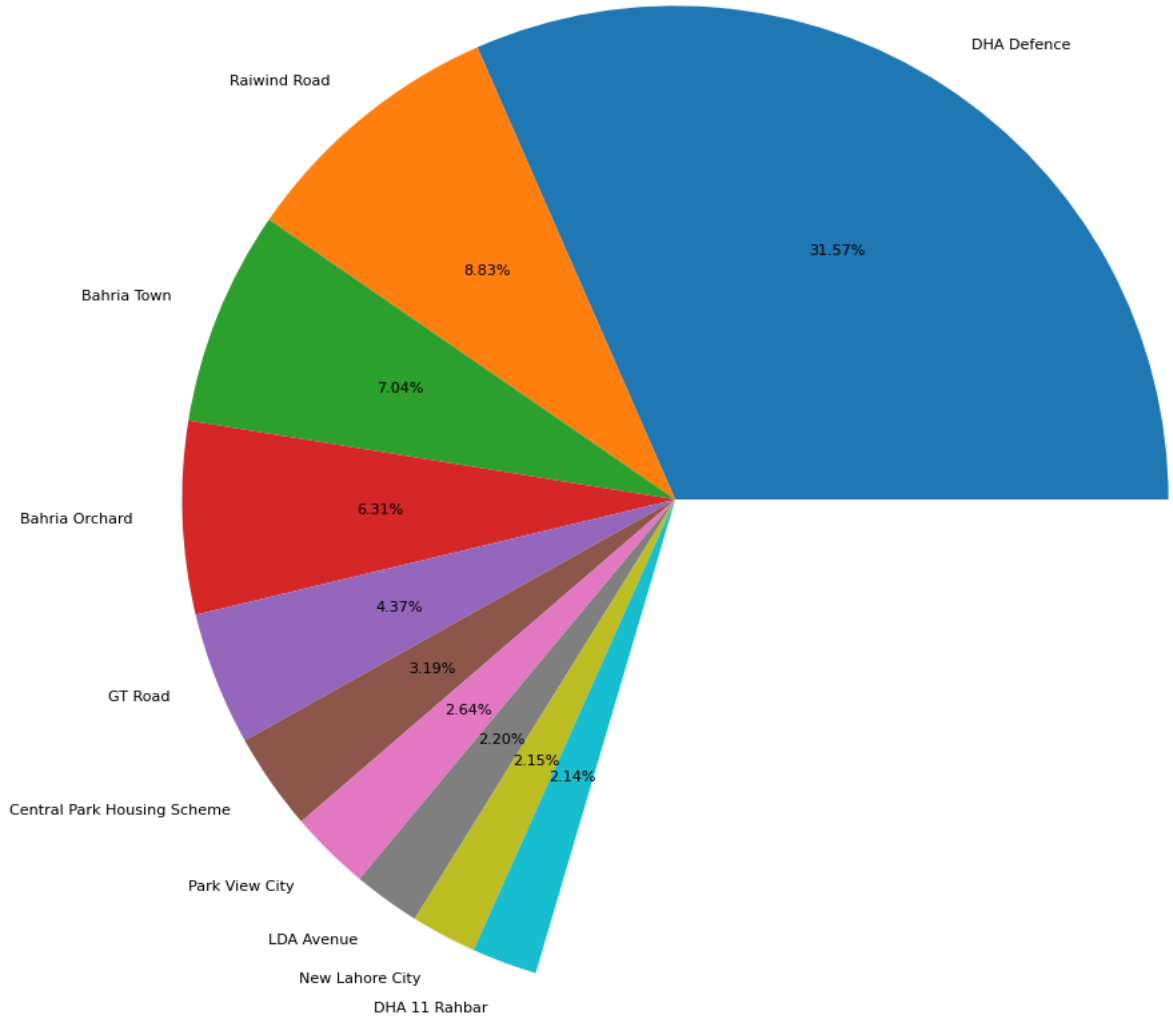
2.1.4.2 Pie Chart

This chart gives us the distribution of total plots. Residential plots account for almost 85% of total plots.



2.1.5 Location

Location gives us the location of the plot. As we can see, DHA Defense accounts for almost 32% of total plots. The top 10 locations by frequency can be seen in the graph.

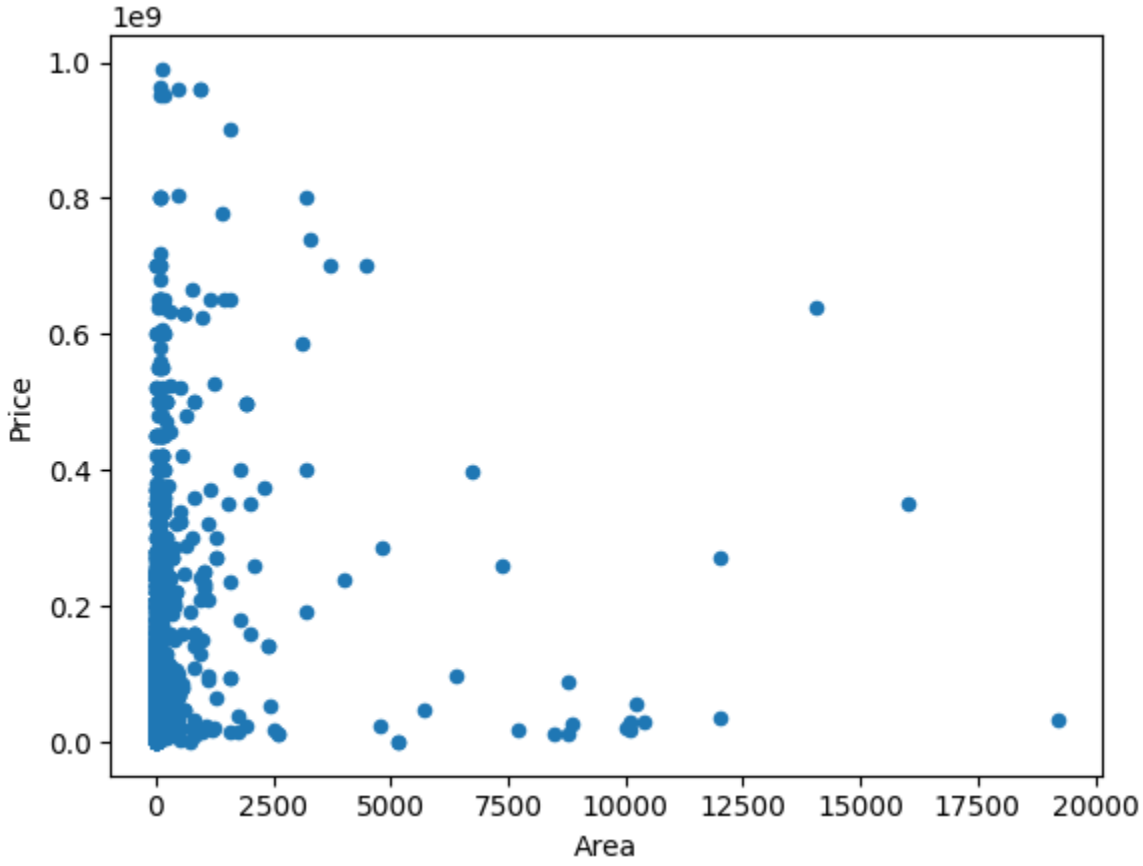


2.2 Combined Feature Exploration

2.2.1 Area - Price Scatter Plot

2.2.1.1 All Plots

This scatter plot gives us a plot between the Area and the Price of all plots. Since the data is varying a lot, no reasonable assumption can be made. We can see that due to a few outliers, the scale of the graph is massively shifted.

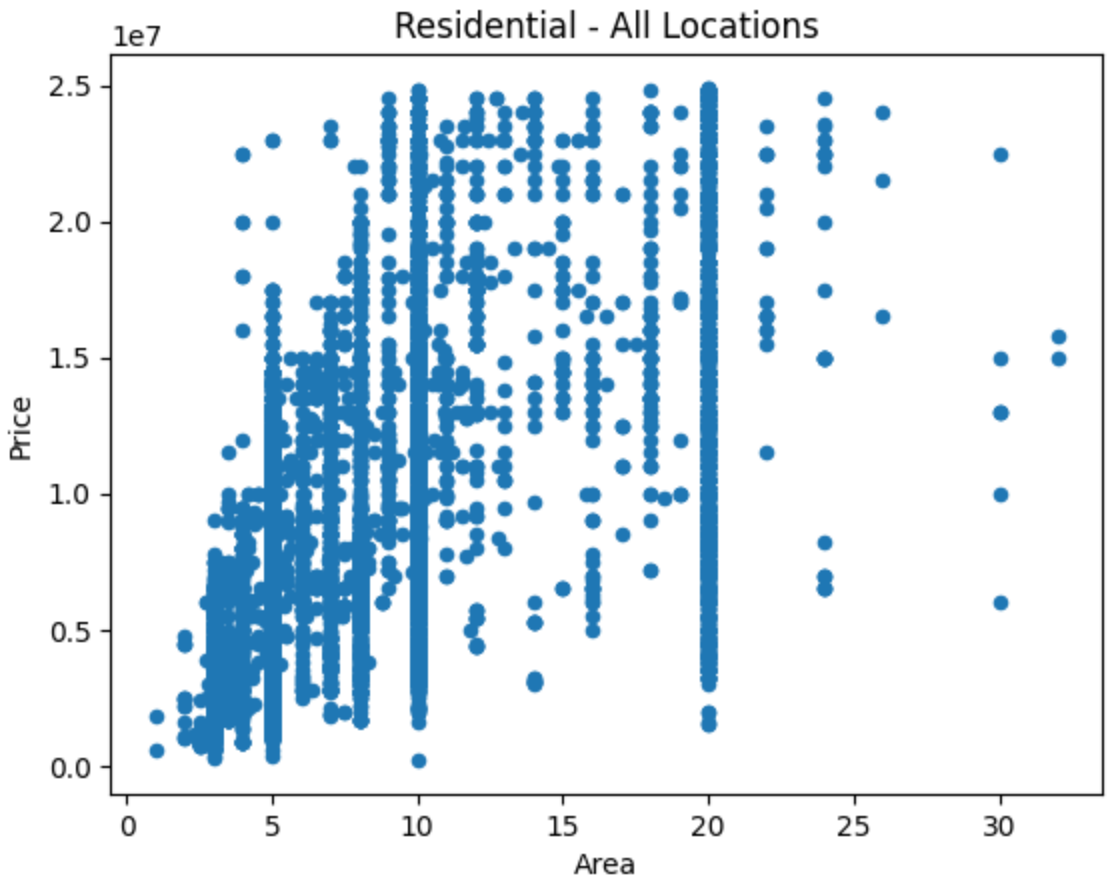


2.2.1.1.1 Residential Plots

This section describes the scatter plot for all residential locations.

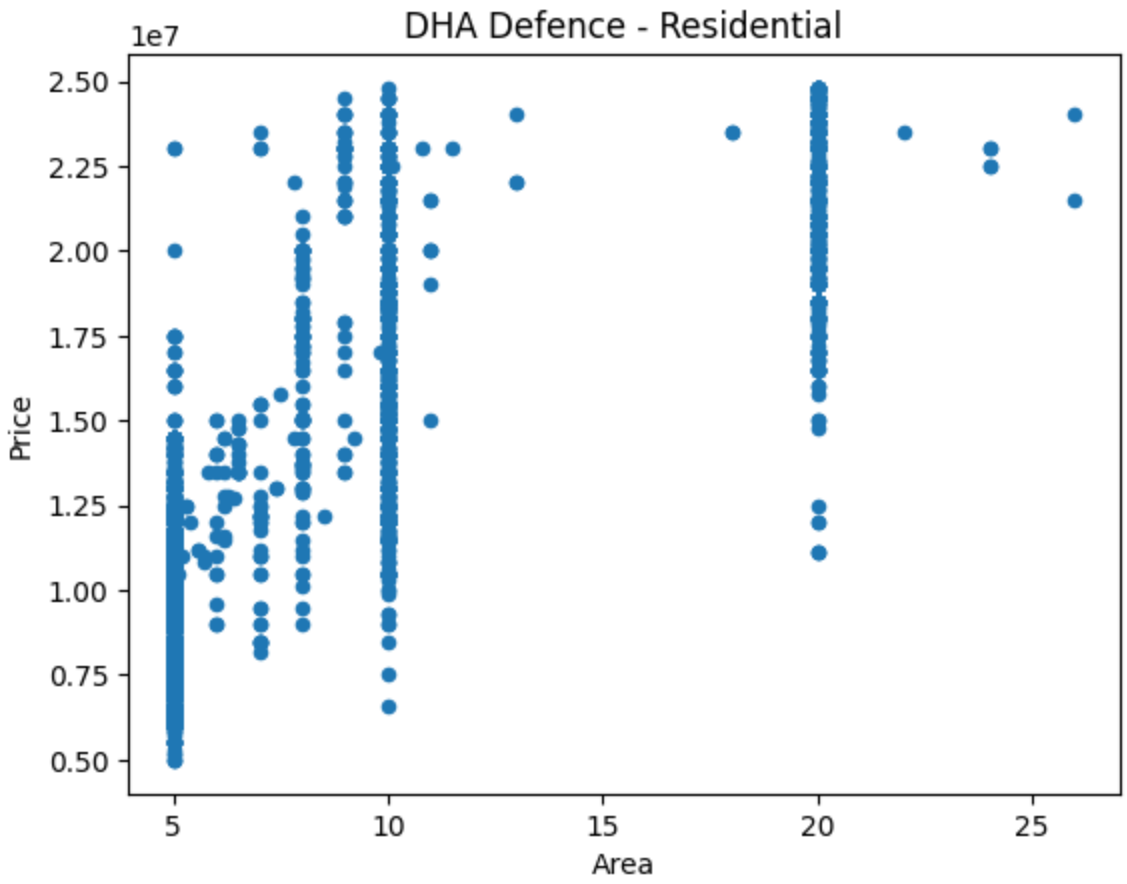
2.2.1.1.1.1 All Plots

As it is evident from the scatter plot, there is no obvious relationship between the Area and Price of all Residential plots. Location may be playing a part in this. Hence, the scatter plots were plotted against every location. This report will cover the top 4 scatter plots.



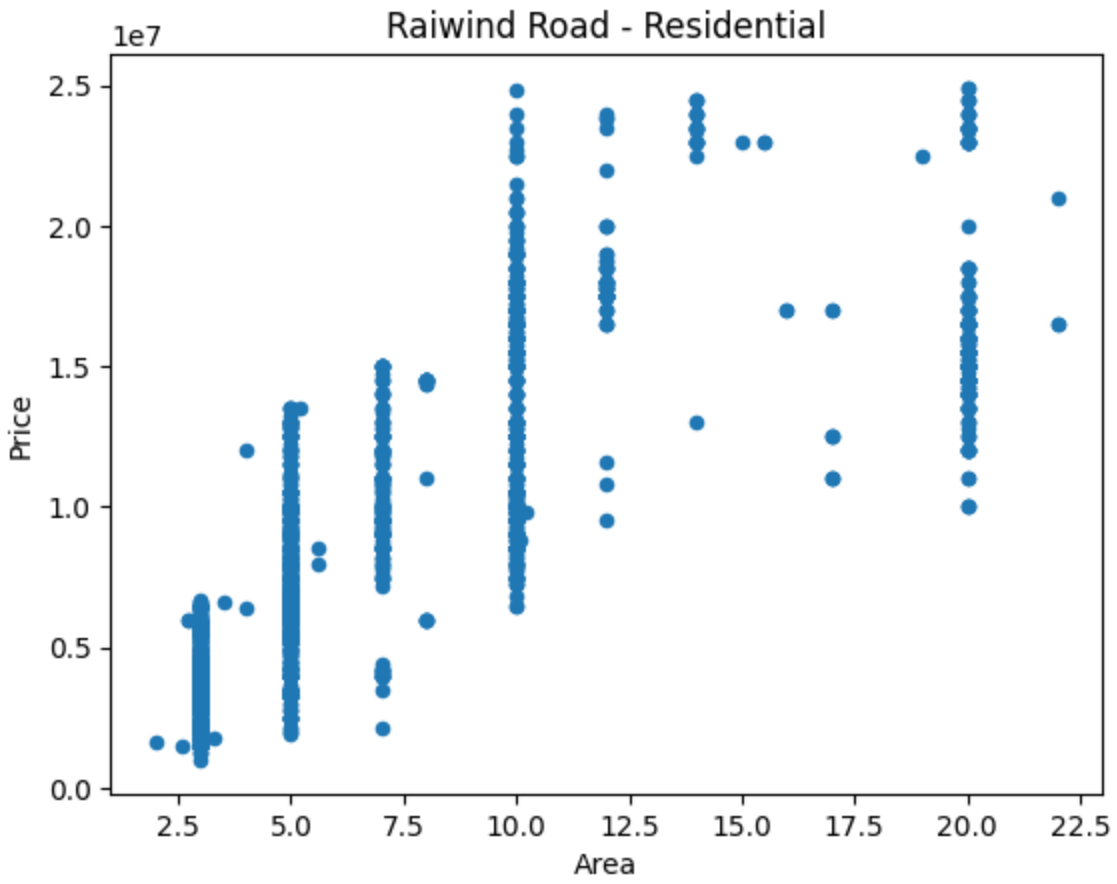
2.2.1.1.1.2 DHA Defense

DHA Defense has all the original phases of DHA along with the future planned phases. Phase 9 is included in this too. It is hard to find a relationship as different phases have different rates for plots.



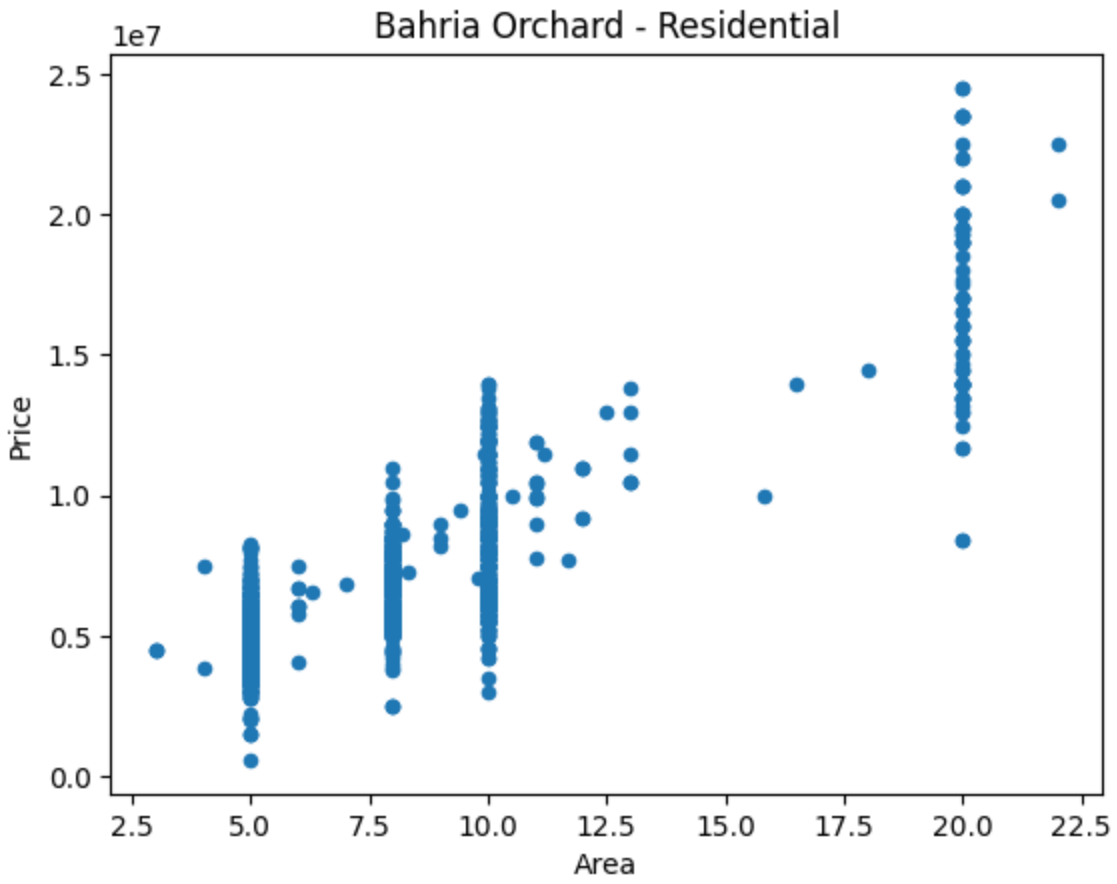
2.2.1.1.1.3 Raiwind Road

We can see that the average price increases as the area increases for this location.



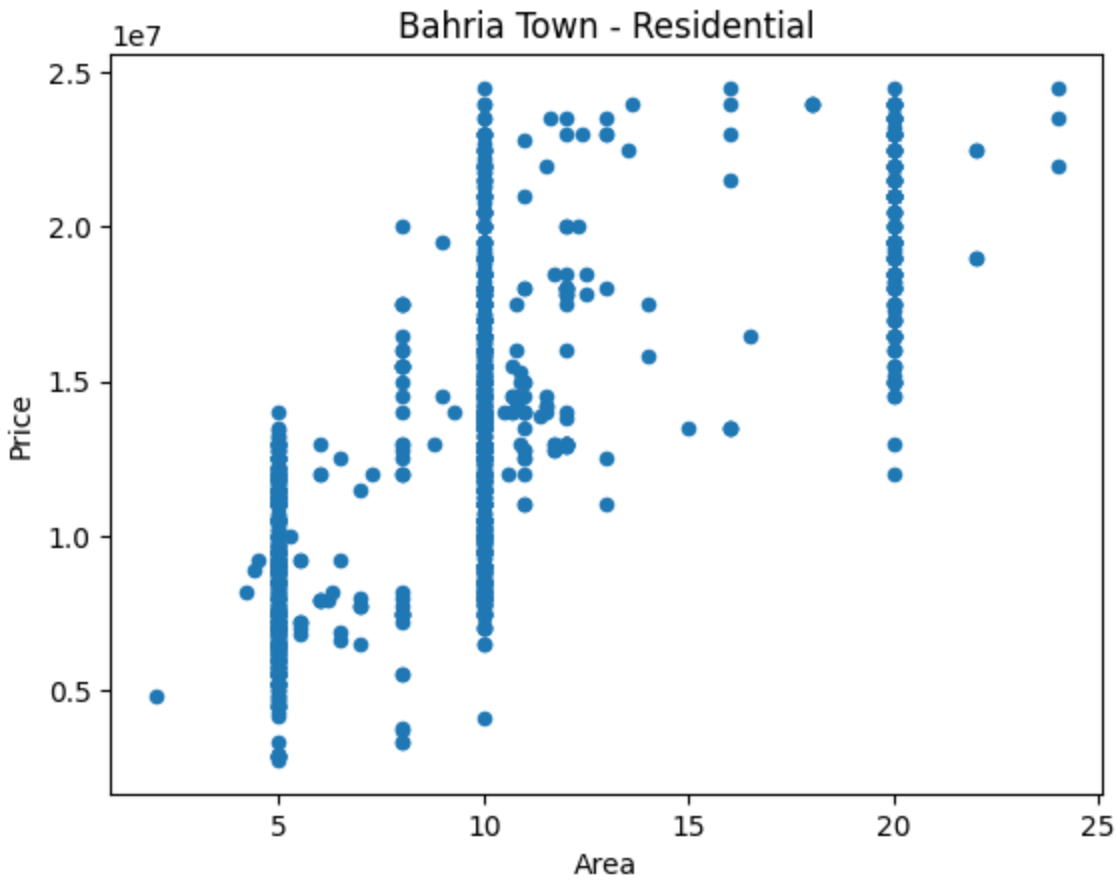
2.2.1.1.1.4 Bahria Orchard

As seen with Raiwind Road, the average price increases with the increase in area for Bahria Orchard too.



2.2.1.1.1.5 Bahria Town

As seen with Raiwind Road and Bahria Orchard, the average price increases with the increase in area for Bahria Town too.

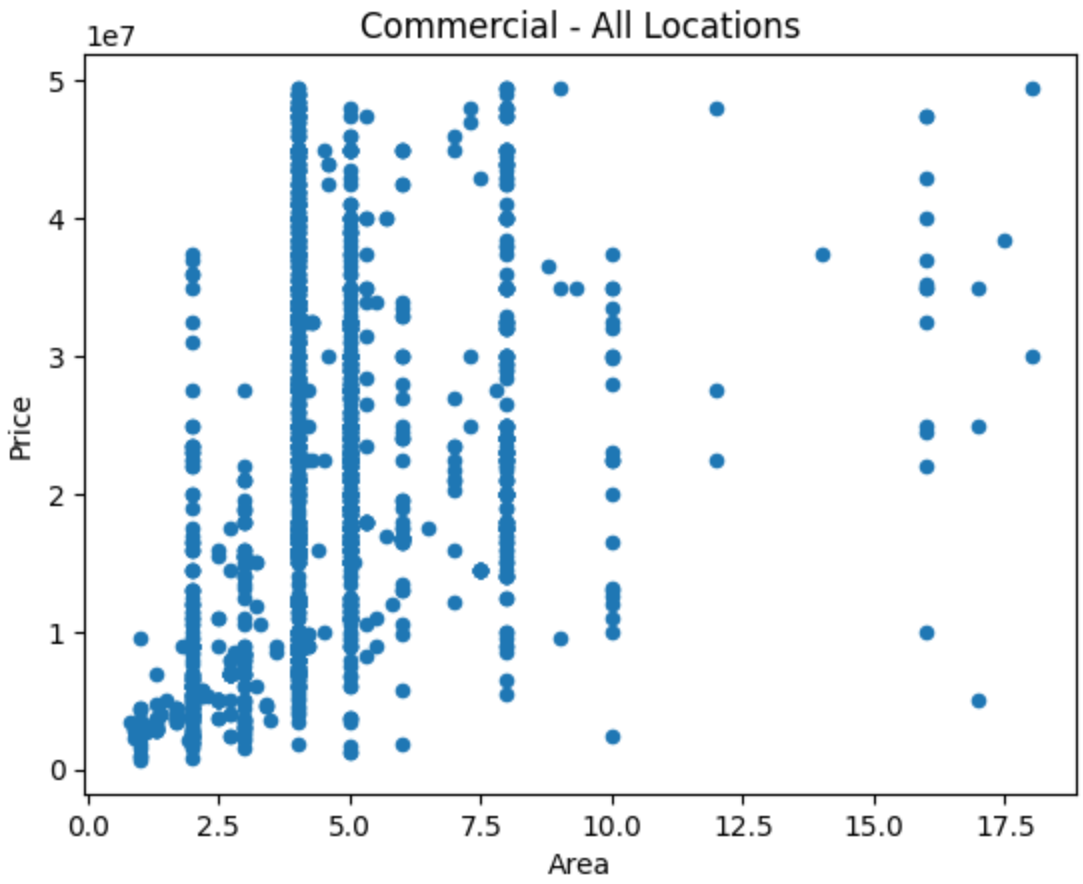


2.2.1.1.2 Commercial Plots

This section covers the commercial plots.

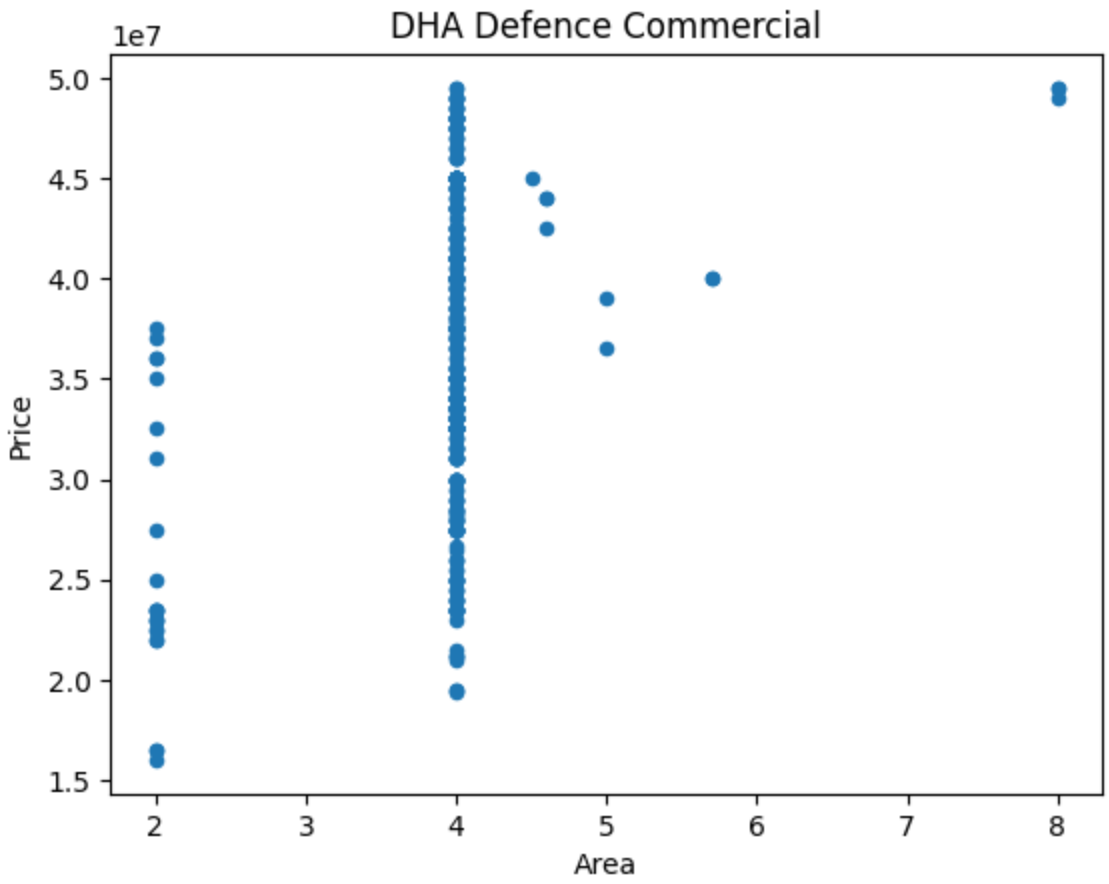
2.2.1.1.2.1 All Plots

As we can see, we cannot derive a relationship based on all locations. We have to filter by locations



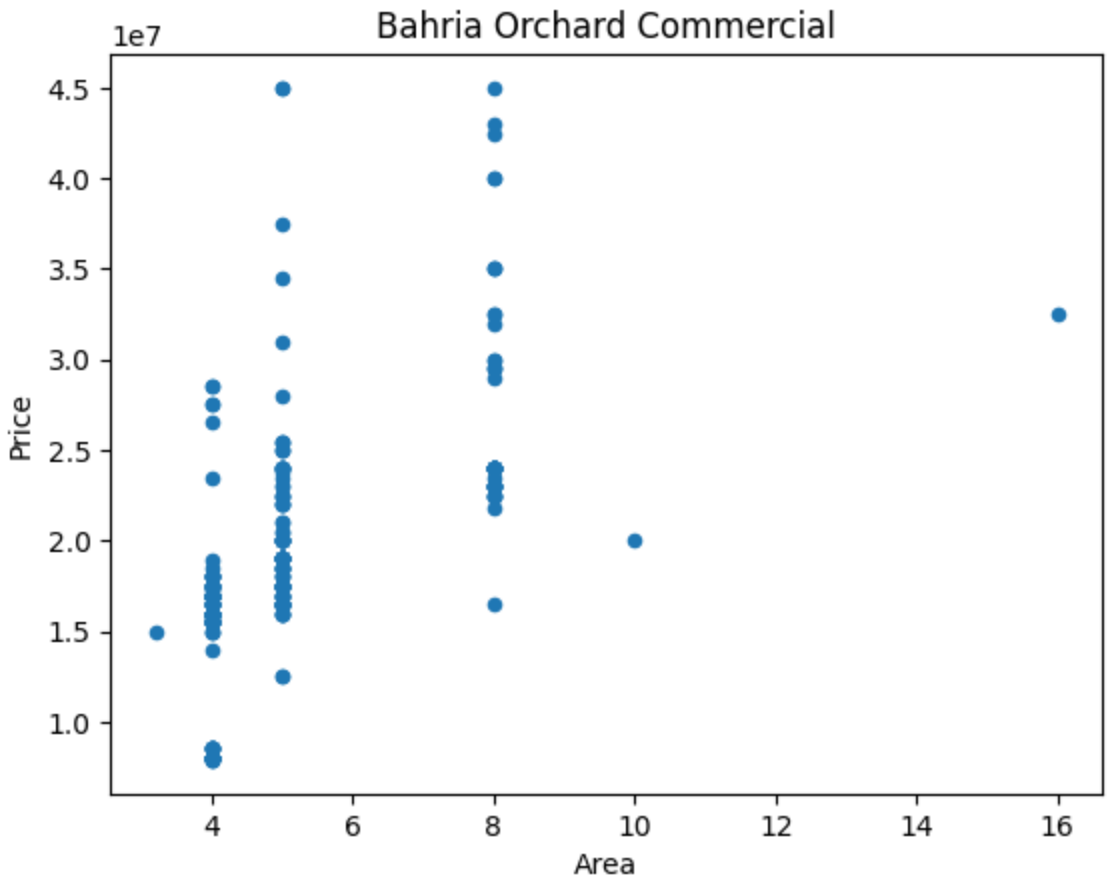
2.2.1.1.2.2 DHA Defense

This scatter plot shows somewhat of a positive relationship, however, as with the problem with DHA's residential plots, the commercial plots are a bit awkward too.



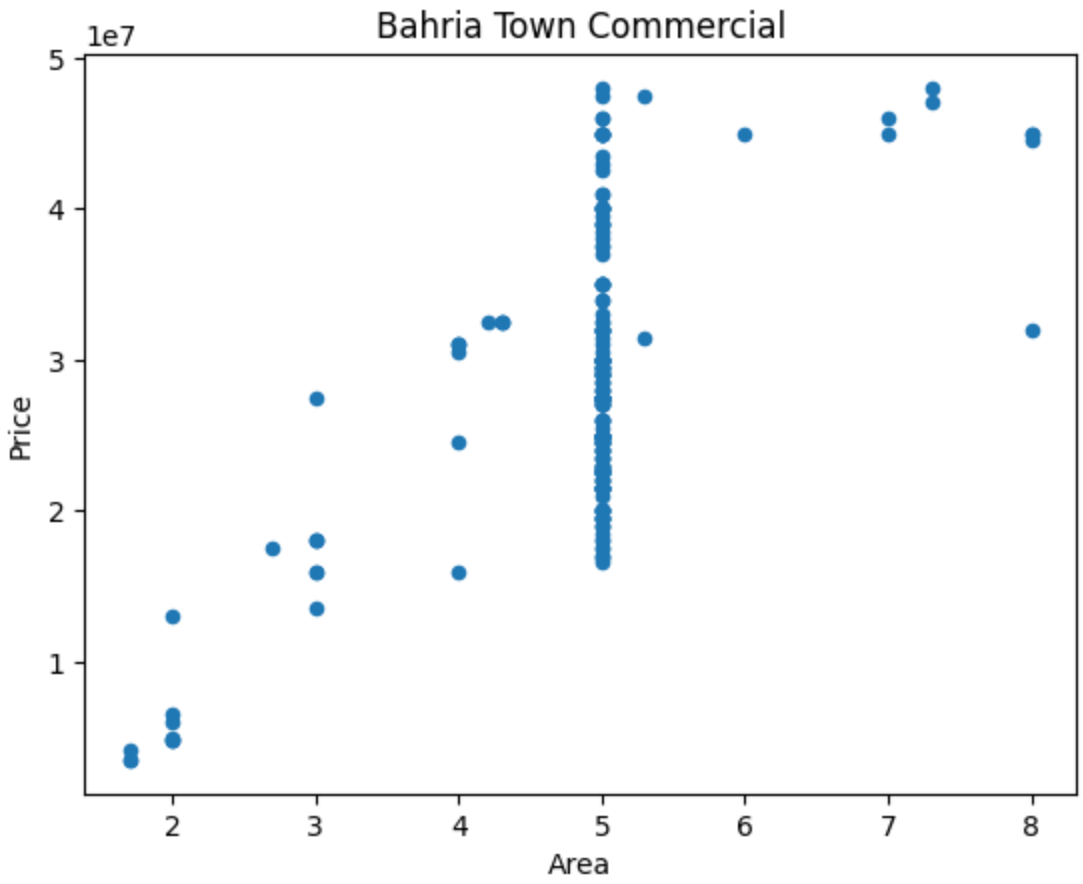
2.2.1.1.2.3. Bahria Orchard

The average price increases with the increase of area in Bahria Orchard.



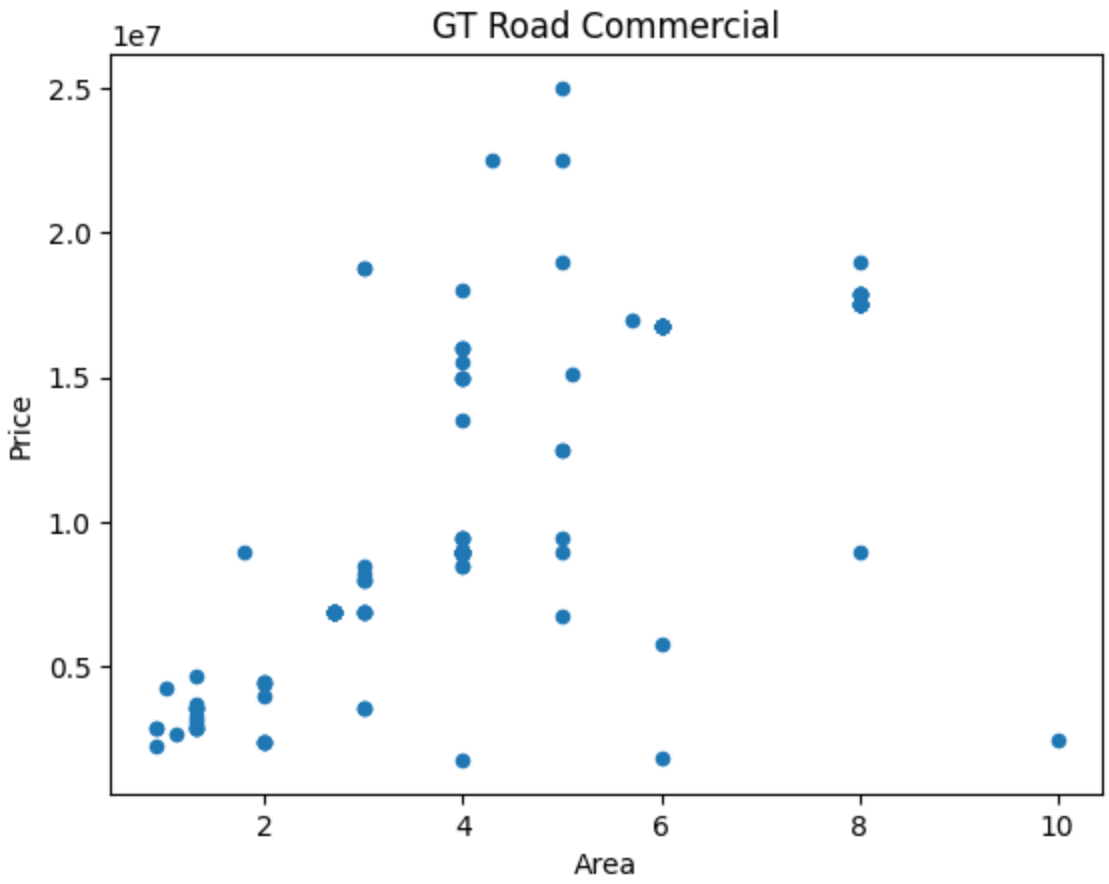
2.2.1.1.2.4 Bahria Town

As it is the case with Bahria Town Orchard, the average price increases with the increase in area.



2.2.1.1.2.5 GT Road

It is hard to gauge the relationship between price and area on GT Road. As it is the case with DHA, GT Road also covers a vast area. We believe that if GT road is further divided into different areas, it can give a better scatter plot.

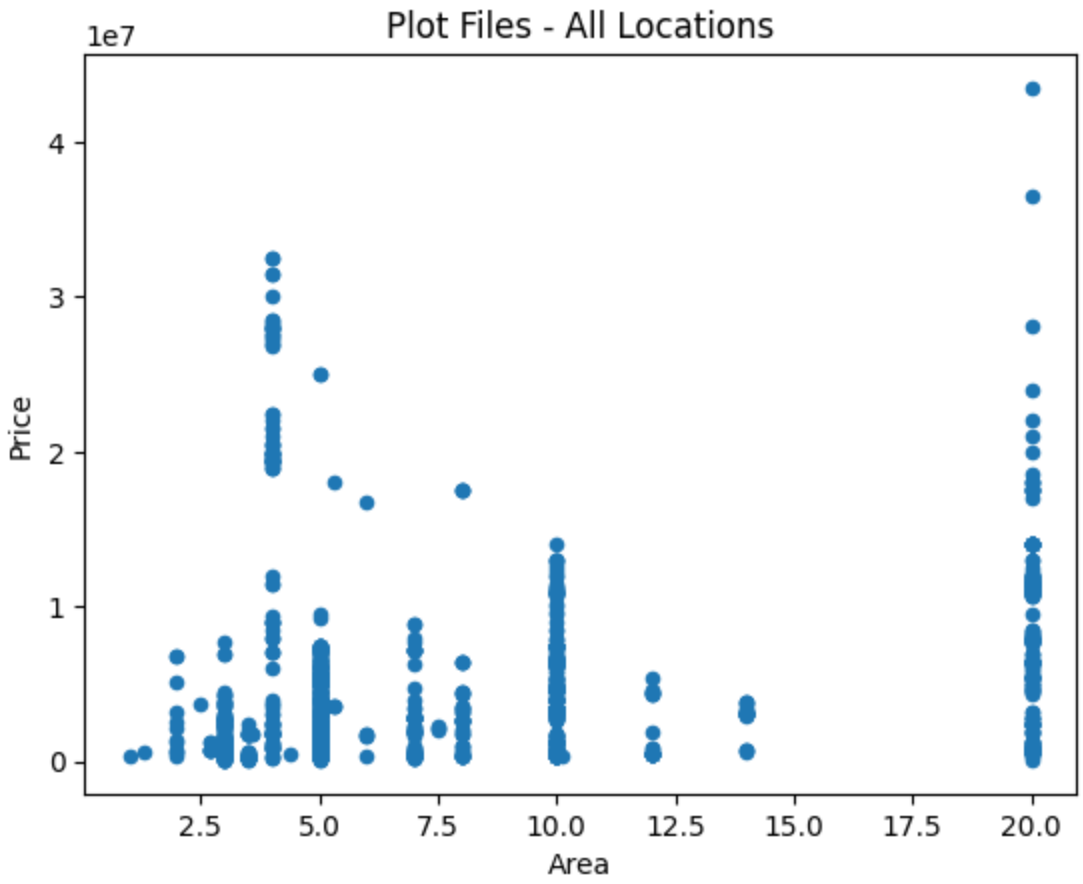


2.2.1.1.3 Plot Files

This section covers the scatter plots for the Plot Files.

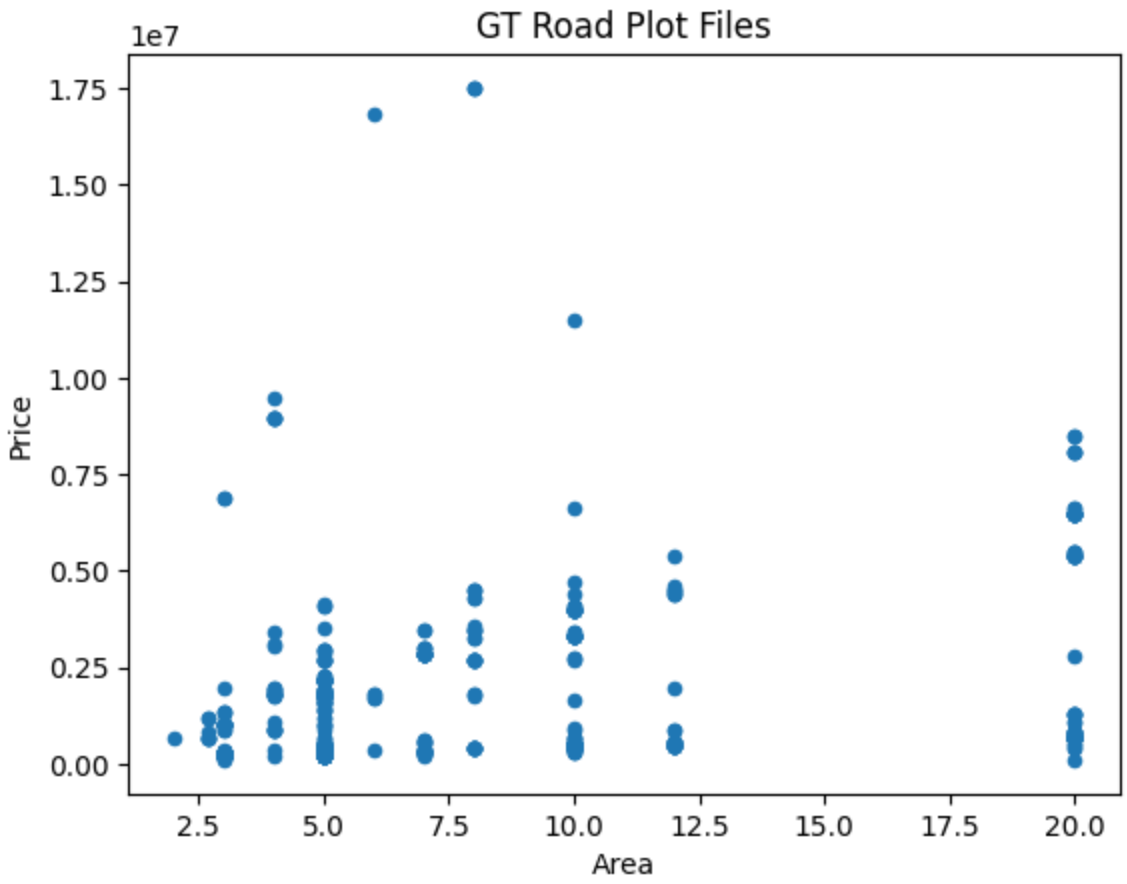
2.2.1.1.3.1 All Plots

As it was the case with commercial and residential plots, we cannot devise a suitable relationship if we consider all locations.



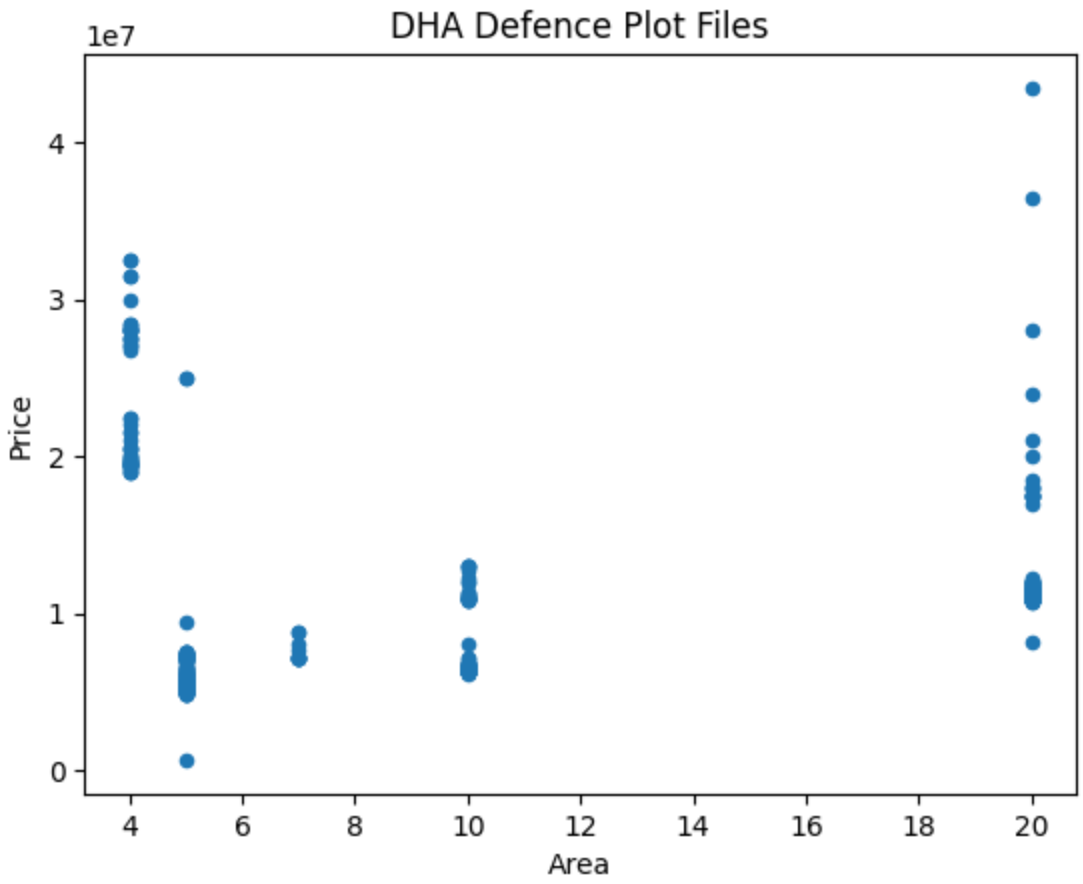
2.2.1.1.3.2 GT Road

If we ignore the outliers, the relationship is positive.



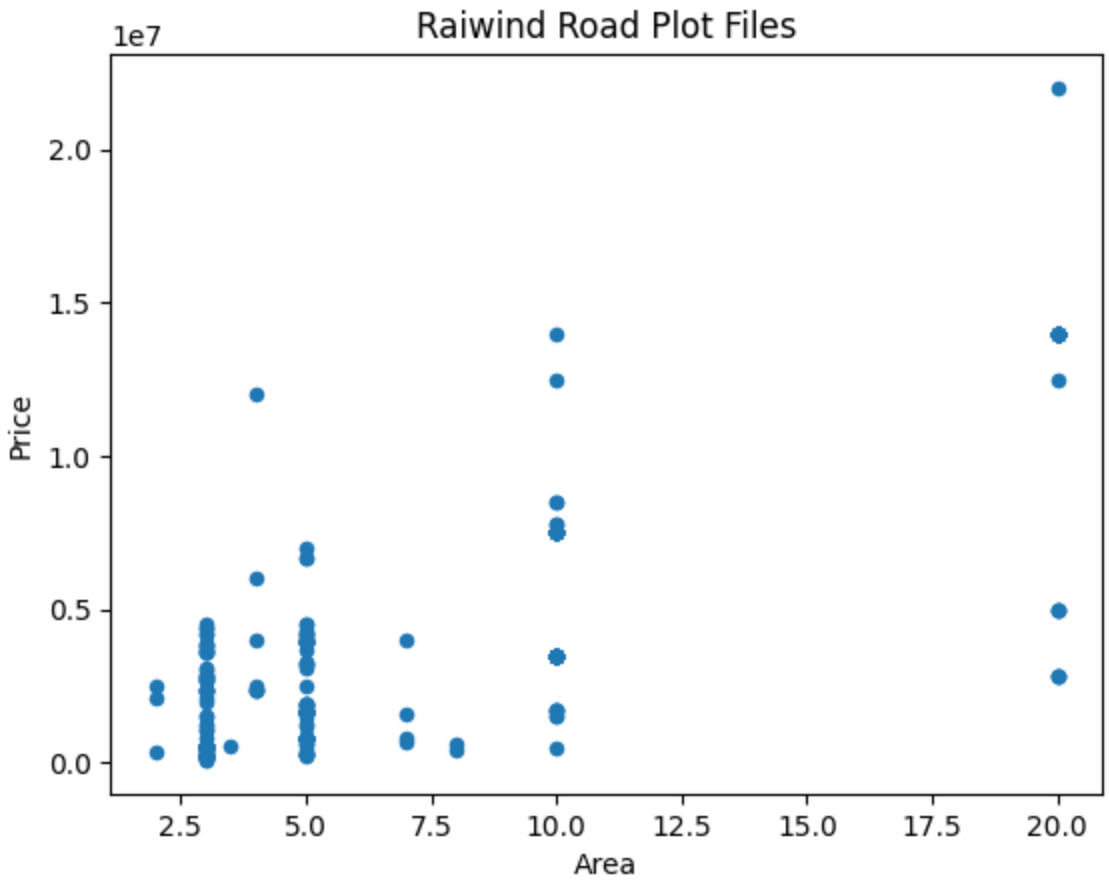
2.2.1.1.3.3. DHA Defense

Defense again proves awkward as the price of plot files for 3-4 Marla houses is a lot. This is again a variation in the new and old phases.



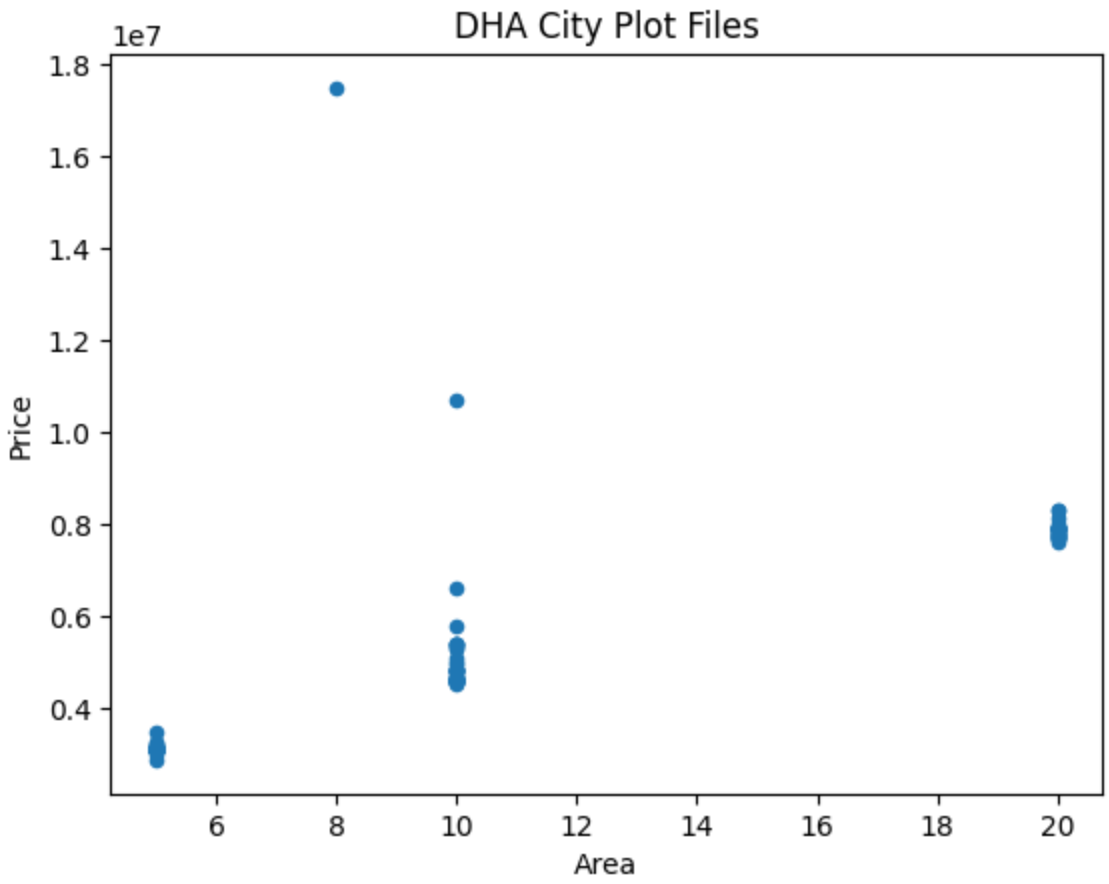
2.2.1.1.3.4 Raiwind Road

It can be observed that the plot files on Raiwind Road have a weakly positive relationship.



2.2.1.1.3.5 DHA City

There is a positive relationship between area and price for DHA City.

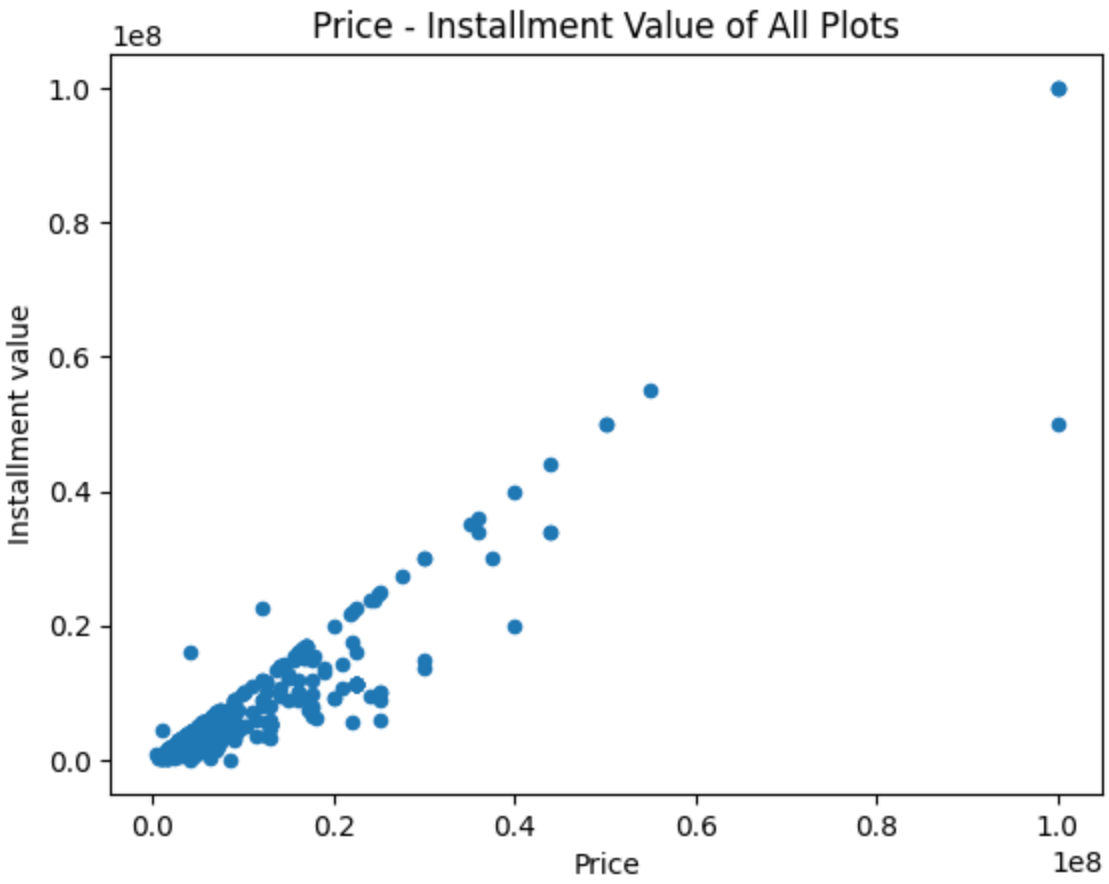


2.2.2 Price - Instalment Value Scatter Plot

This section covers the Price - Instalment value scatter plots. Instalment value is the value owed to the owner in order to gain full ownership of the plot. It is often used over 5-10 year plans. Price in this case means a downpayment and the Instalment value indicates the amount still owed.

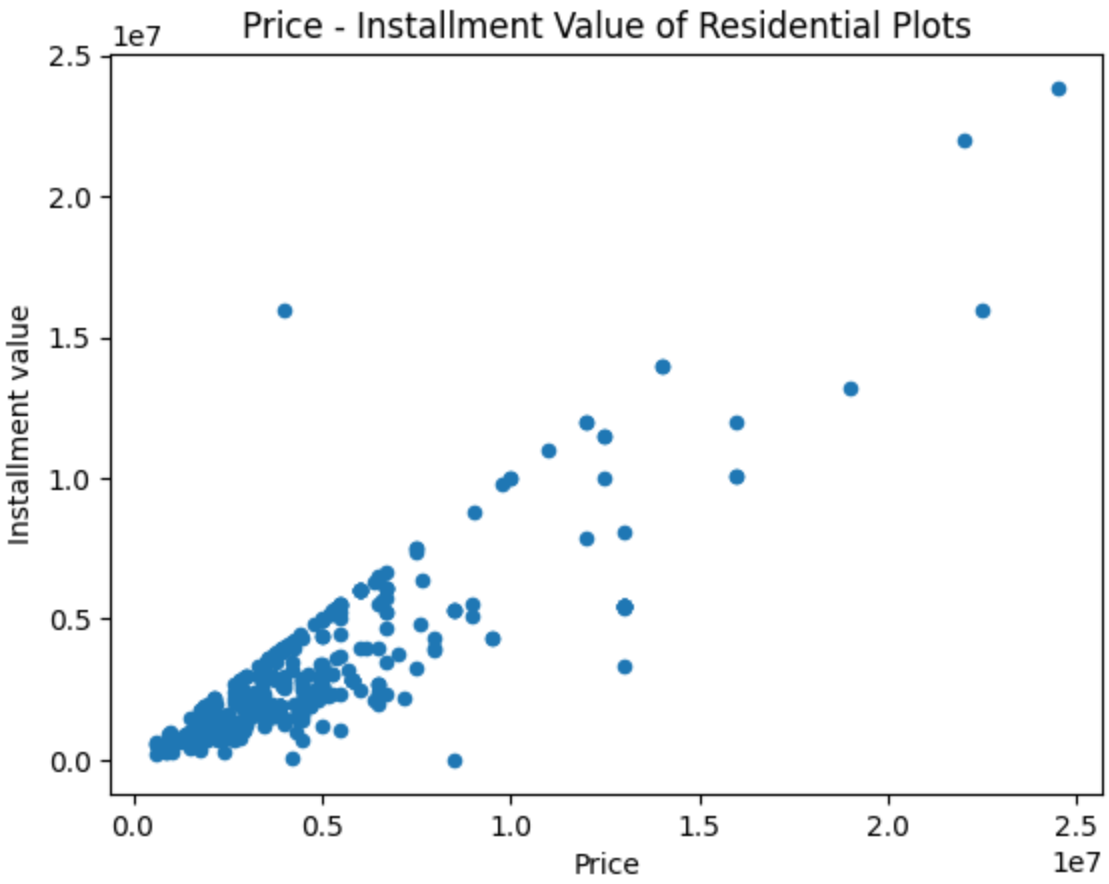
2.2.2.1 All Plots

As we can see from the scatter plot, as the price of the plot increases, it's instalment value increases too. Using all plots gives us a good indication, however, just to be sure, we divided the plots by their type too.



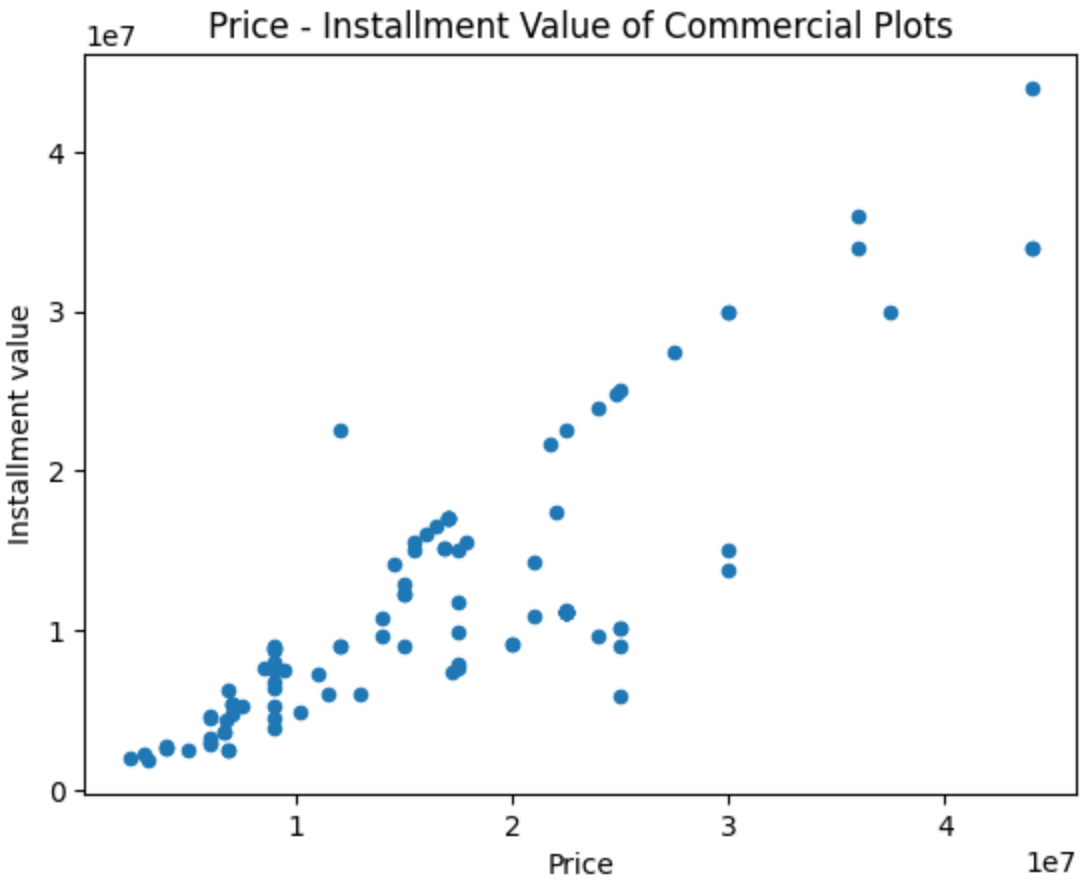
2.2.2.2 Residential

Residential plots also show a positive relationship between the price and the instalment value of the graph.



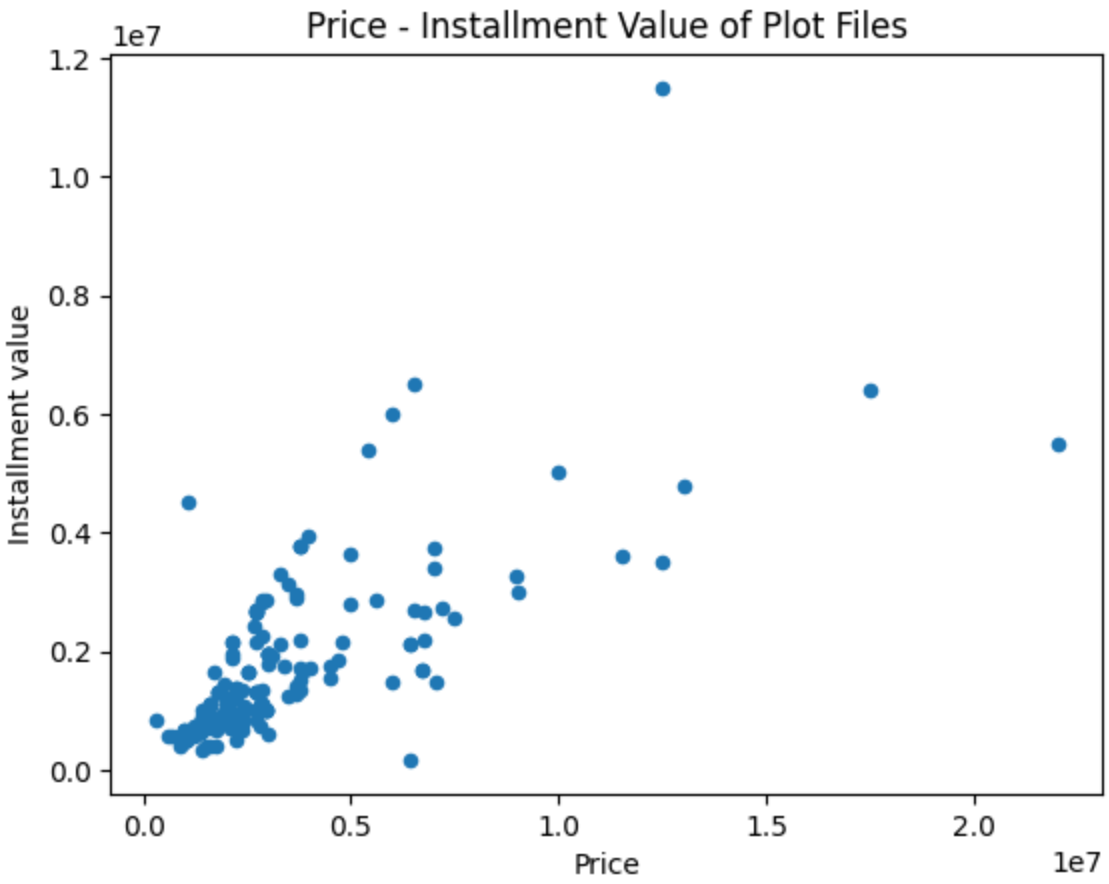
2.2.2.3 Commercial

Confirming the findings from the residential plots, commercial plots also show a positive relationship between price and instalment value



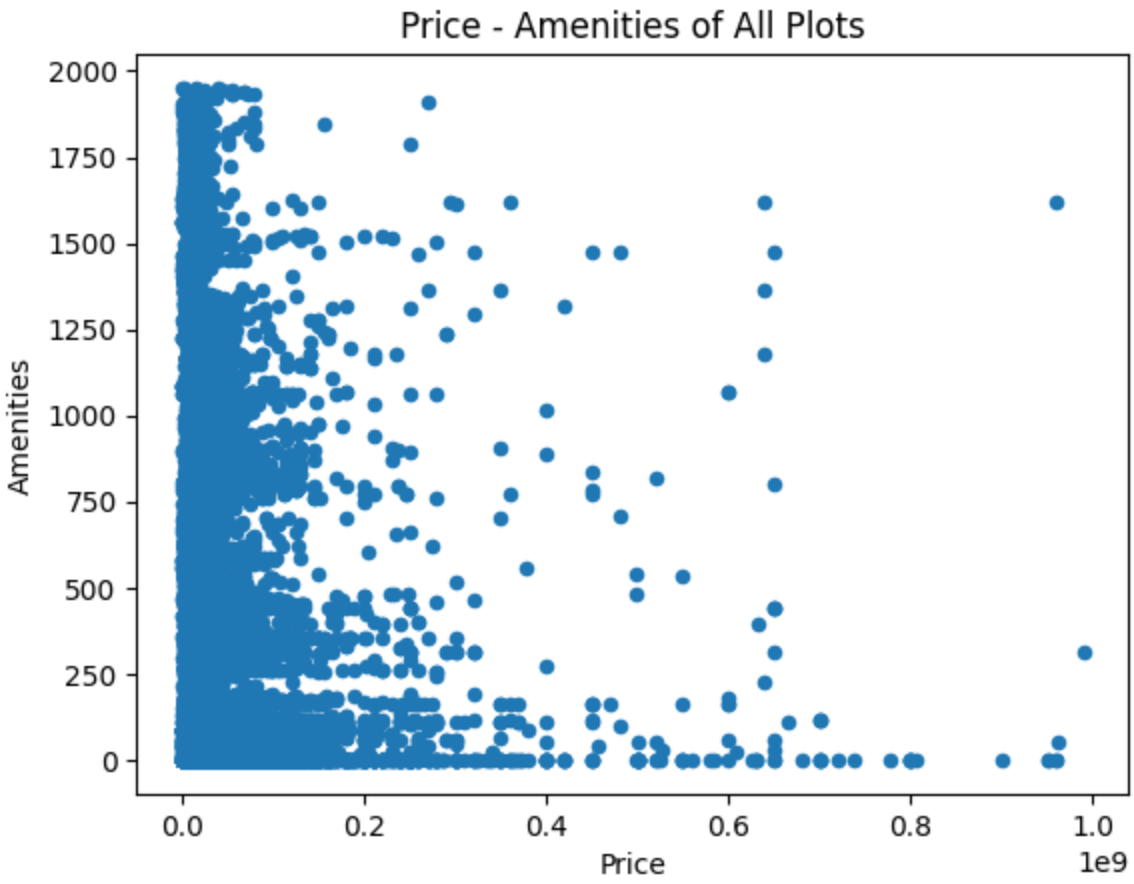
2.2.2.4 Plot Files

Plot files also indicate a positive relationship between the price and instalment value.



2.2.3 Price - Amenities Scatter Plot

This section covers the scatter plots between Price and Amenities. As with the previous relationships, the type of the plot is changed to find observations. In the case of all types of plots, the amenities appear random. Hence, we cannot say for sure if there is a relationship between the two.

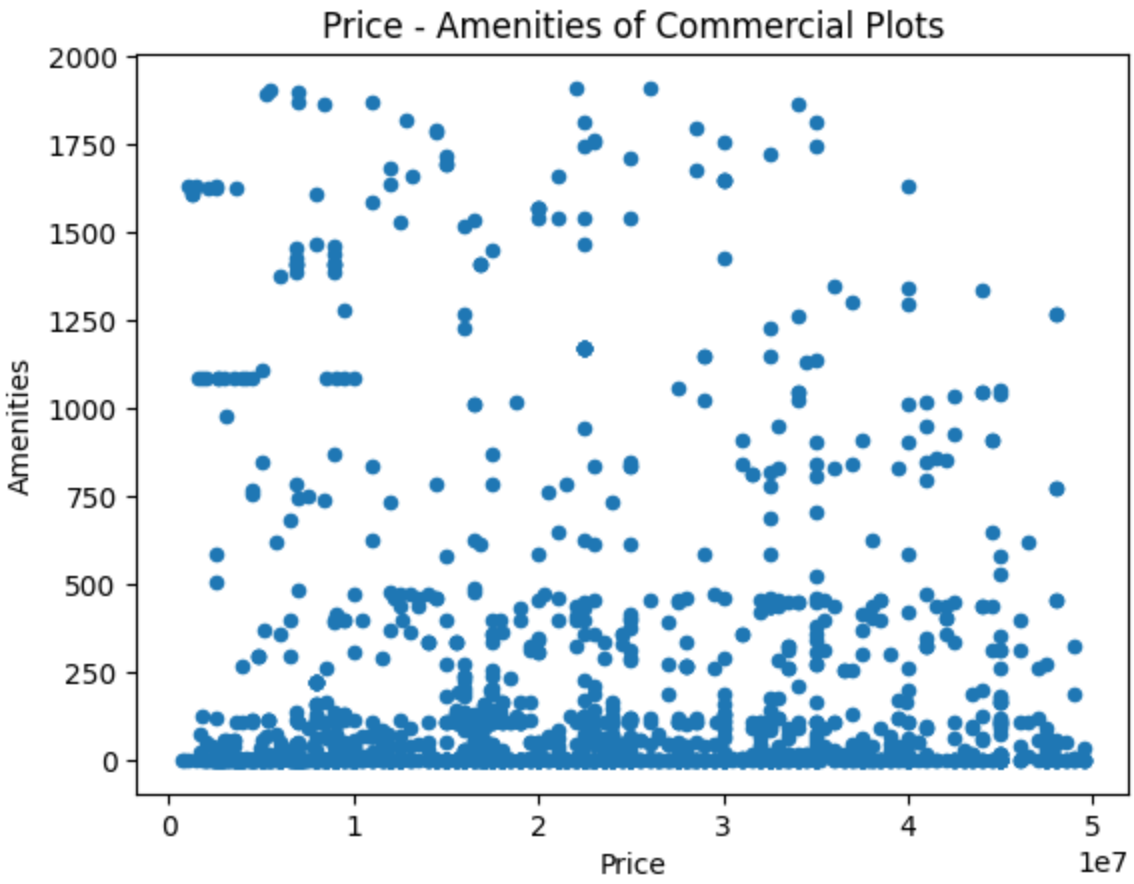


2.2.3.1 Residential

The scatter plot is shaped like a square, again making it hard to deduce any relationship between the 2 attributes.

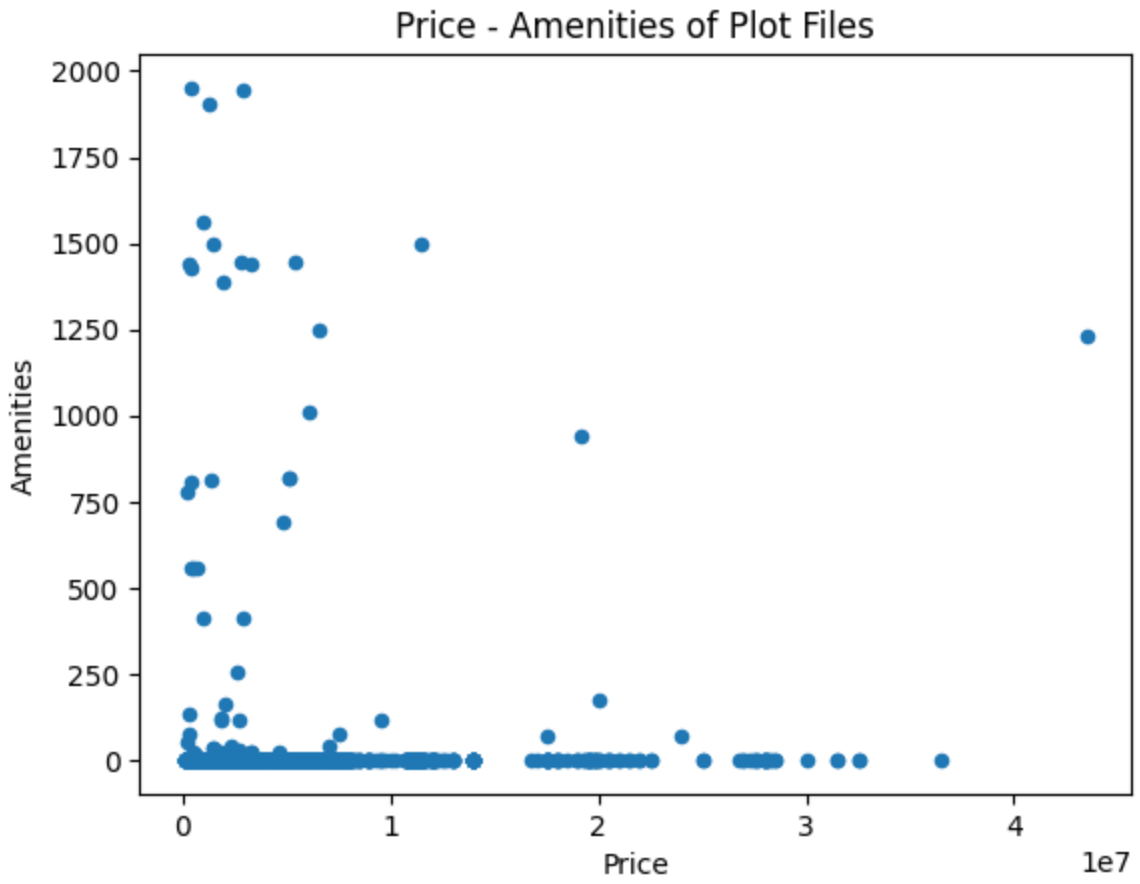
2.2.3.2 Commercial

Same is the case with commercial plots. No solid relationship can be found.



2.2.3.3 Plot Files

Plot Files also do not show any relationship between price and amenities.

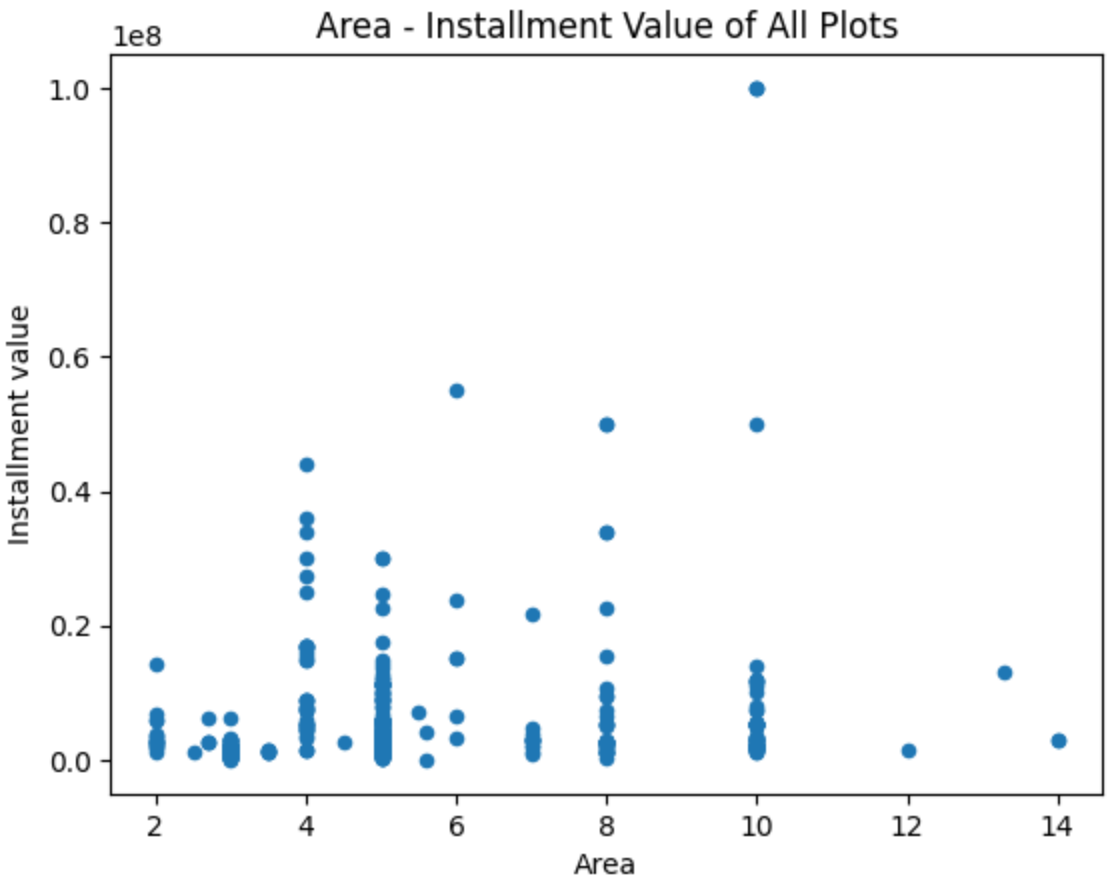


2.2.4 Area - Installment Value Scatter Plot

This section covers the relationship between area and instalment value. The intuition is that the instalment value will increase with the area.

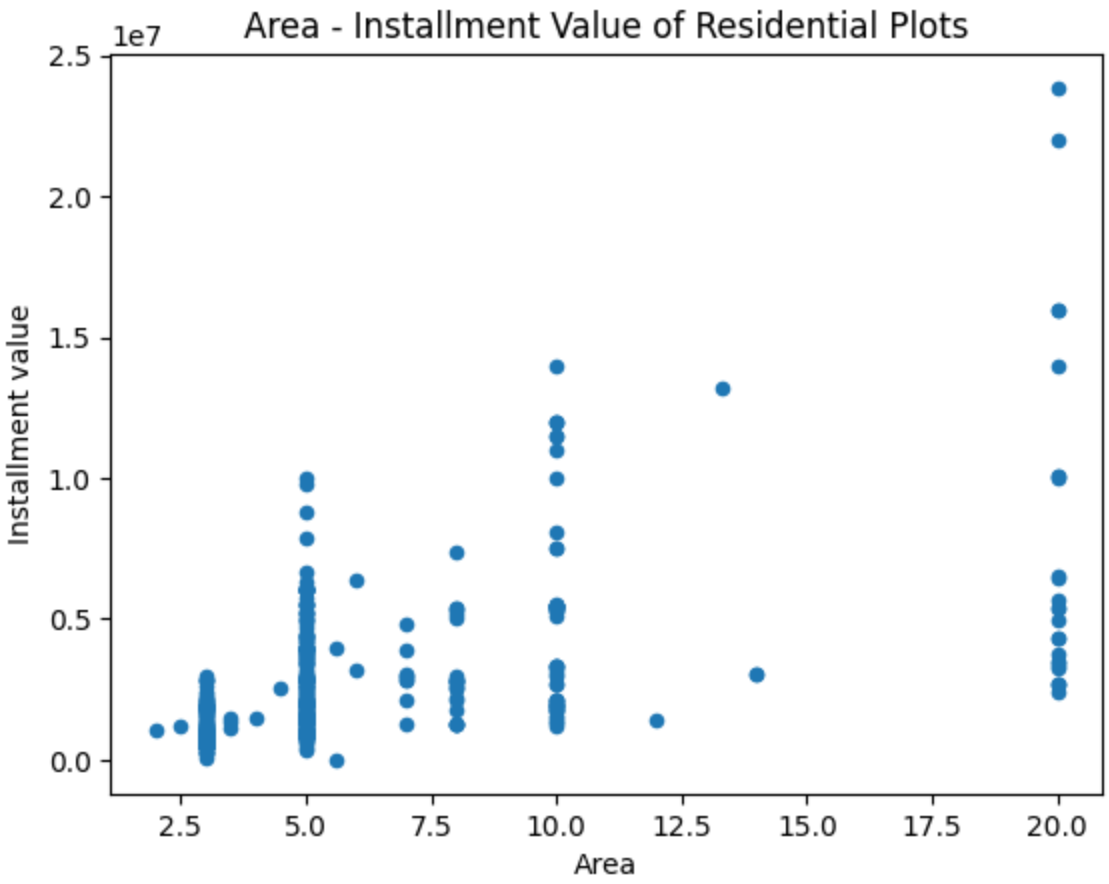
2.2.4.1 All Plots

No concrete relationship can be found if we consider all types of plots. Hence, we will consider different types of plots as done in the previous relationships.



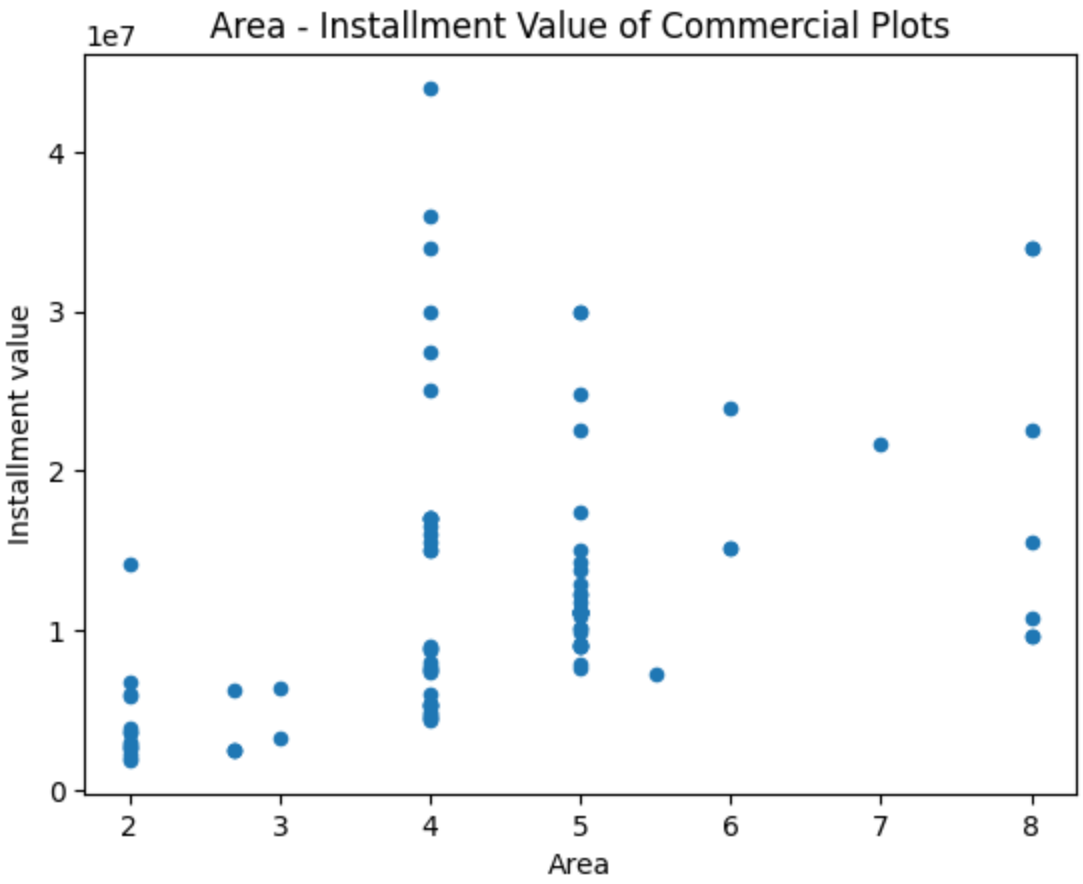
2.2.4.1 Residential

A weak positive relationship can be observed.



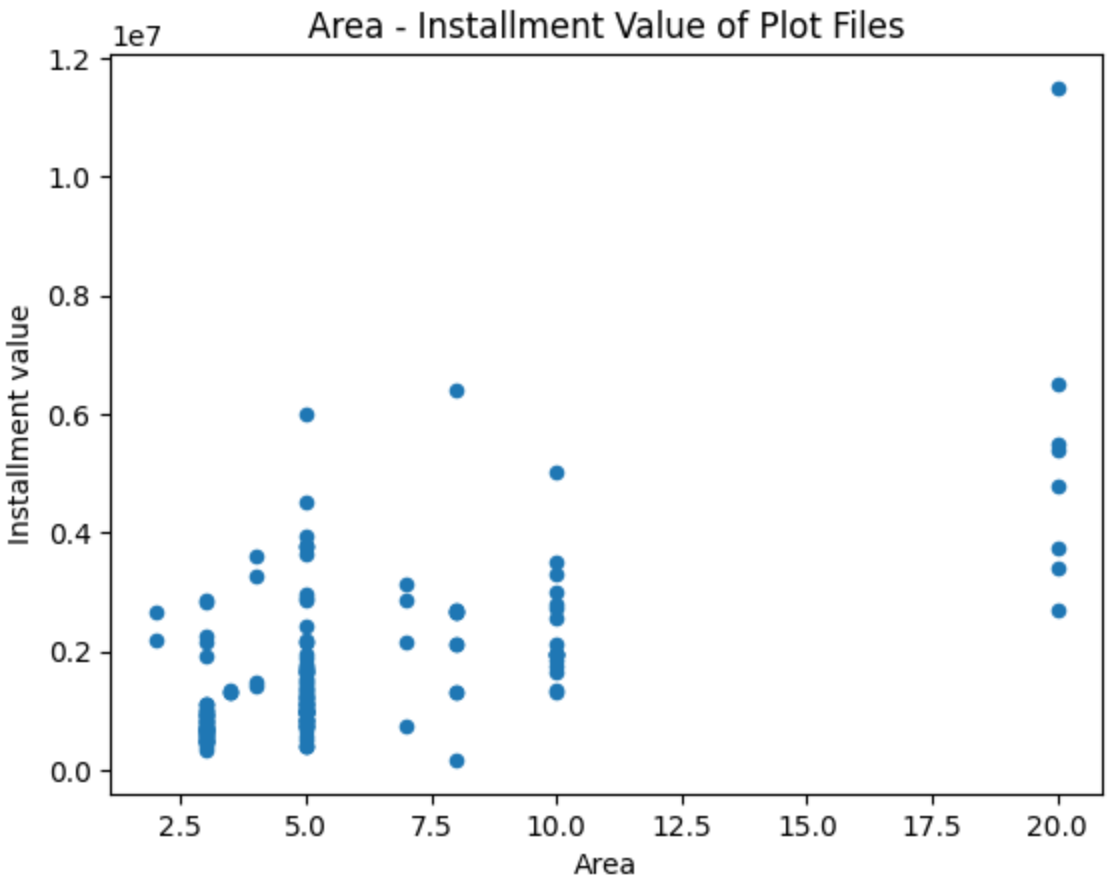
2.2.4.2 Commercial

4 Marla plots tend to deviate from the rest of the plot. One possible explanation is the abundance of new commercial buildings in Bahria Town and Bahria Orchard that are of 4 Marlas. This makes 4 Marla Instalments very high compared to the rest of the graph.



2.2.4.3 Plot Files

A positive relationship can be observed for plot files.

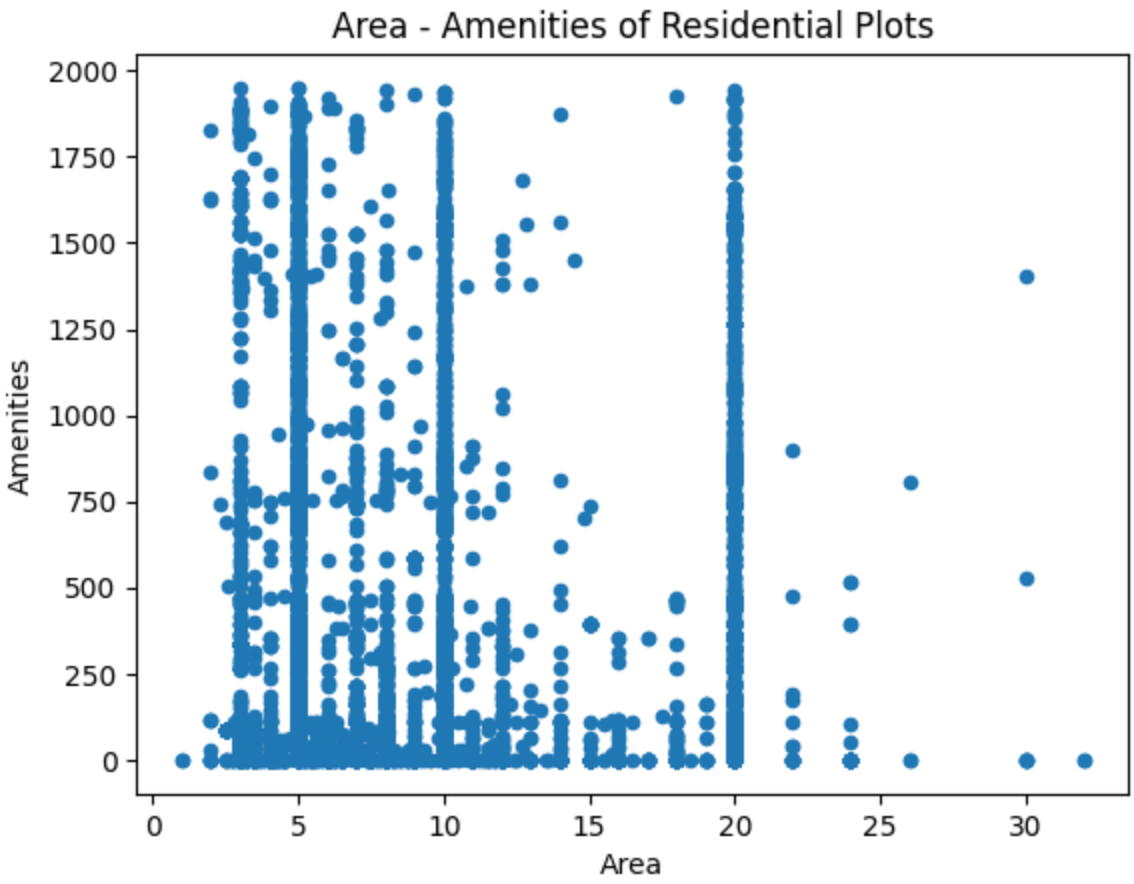


2.2.5 Area - Amenities Scatter Plot

This section covers the scatter plots for Area and Amenities.

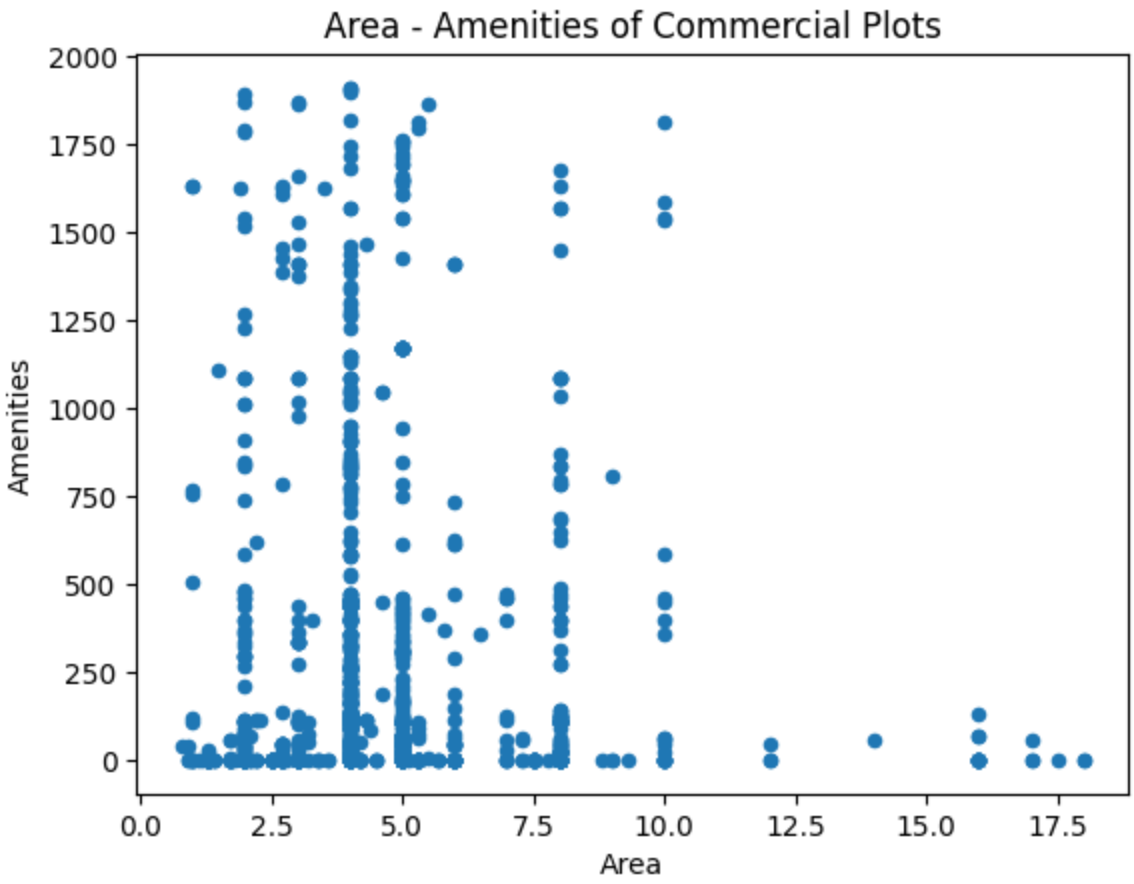
2.2.5.1 Residential

No relationship can be observed.



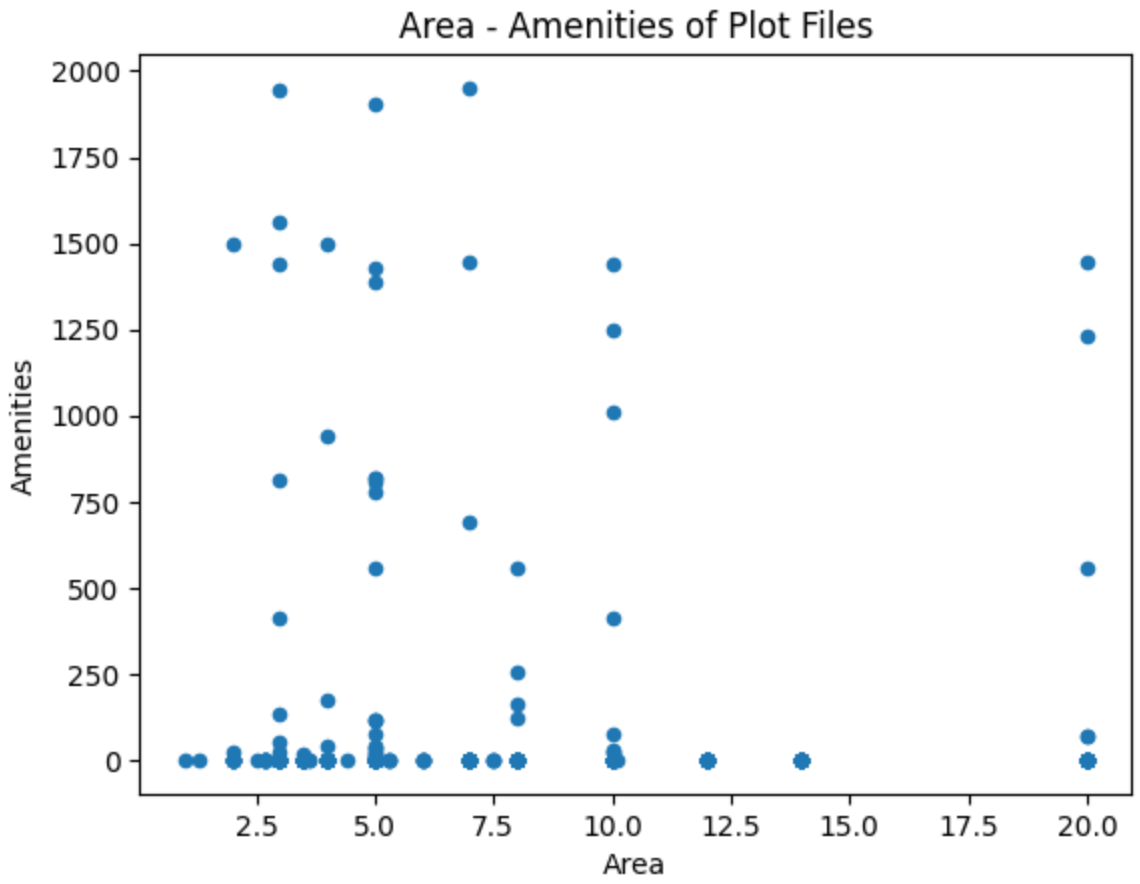
2.2.5.2 Commercial

No relationship can be observed.



2.2.5.3 Plot Files

No relationship can be observed.



2.3 Correlation Matrix

This section displays the correlation matrix of the data frames. As was the case with the scatter plots, the matrices have been segregated based on location to get a better relationship.

2.3.1 All Plots

	Price	Area	Amenities	Installment value
Price	1.000000	0.172782	0.048141	0.942146
Area	0.172782	1.000000	0.038810	0.170875
Amenities	0.048141	0.038810	1.000000	0.117228
Installment value	0.942146	0.170875	0.117228	1.000000

2.3.2 Residential Plots

	Price	Area	Amenities	Installment value
Price	1.000000	0.667615	0.036754	0.859967
Area	0.667615	1.000000	0.022749	0.564025
Amenities	0.036754	0.022749	1.000000	0.002705
Installment value	0.859967	0.564025	0.002705	1.000000

2.3.3 Commercial Plots

	Price	Area	Amenities	Installment value
Price	1.000000	0.329575	0.011261	0.837038
Area	0.329575	1.000000	-0.036881	0.448234
Amenities	0.011261	-0.036881	1.000000	0.127260
Installment value	0.837038	0.448234	0.127260	1.000000

2.3.4 Plot Files

	Price	Area	Amenities	Installment value
Price	1.000000	0.357646	0.027881	0.718948
Area	0.357646	1.000000	-0.011988	0.647612
Amenities	0.027881	-0.011988	1.000000	-0.013994
Installment value	0.718948	0.647612	-0.013994	1.000000

2.4 Covariance Matrix

This section provides the covariance matrix for all types of plots. It gives us the value by which an attribute changes and it is a good measure of the relationship between 2 attributes.

2.4.1 All Plots

	Price	Area	Amenities	Installment value
Price	1.965191e+15	1.852210e+09	7.540480e+08	7.234226e+13
Area	1.852210e+09	5.847615e+04	3.316005e+03	1.843426e+07
Amenities	7.540480e+08	3.316005e+03	1.248448e+05	5.145787e+08
Installment value	7.234226e+13	1.843426e+07	5.145787e+08	6.772325e+13

2.4.2 Residential

	Price	Area	Amenities	Installment value
Price	3.493014e+13	2.123906e+07	7.939390e+07	7.139453e+12
Area	2.123906e+07	2.897466e+01	4.475627e+01	5.574975e+06
Amenities	7.939390e+07	4.475627e+01	1.335870e+05	4.116451e+06
Installment value	7.139453e+12	5.574975e+06	4.116451e+06	6.963251e+12

2.4.3 Commercial

	Price	Area	Amenities	Installment value
Price	1.528650e+14	8.356314e+06	5.137674e+07	6.519184e+13
Area	8.356314e+06	4.205464e+00	-2.790886e+01	5.481422e+06
Amenities	5.137674e+07	-2.790886e+01	1.361649e+05	6.129608e+08
Installment value	6.519184e+13	5.481422e+06	6.129608e+08	7.088769e+13

2.4.4 Plot Files

	Price	Area	Amenities	Installment value
Price	2.242236e+13	8.882877e+06	1.633328e+07	2.869713e+12
Area	8.882877e+06	2.751177e+01	-7.779022e+00	3.360546e+06
Amenities	1.633328e+07	-7.779022e+00	1.530570e+04	-3.946177e+06
Installment value	2.869713e+12	3.360546e+06	-3.946177e+06	1.968864e+12

3. Observations, Deductions and Conclusion

- According to the histogram, boxplot and density graph, the price attribute is left-skewed. This can negatively affect the prediction values. A reasonable assumption would be to exclude the plots that have a price > 50000000. This will give us a better prediction model.
- Since the price is left-skewed, it affects the central statistics like the mean. Removing outliers will give us better central statistics.
- According to the histogram, boxplot and density graph, the area attribute is left-skewed. This can negatively affect the prediction values. A reasonable assumption would be to exclude the plots that have an area > 40. This will give us a better prediction model.
- Since the area is left-skewed, it affects the central statistics like the mean. Removing outliers will give us better central statistics.
- According to the histogram, boxplot and density graph, the instalment value attribute is left-skewed. This can negatively affect the prediction values. A reasonable assumption would be to exclude the plots that have an instalment value > 10000000. This will give us a better prediction model.
- Since the instalment is left-skewed, it affects the central statistics like the mean. Removing outliers will give us better central statistics.
- Including types of plots other than residential, commercial or plot files is not necessary as these types make up almost 99% of the total data. There is not enough data for other classes.
- DHA Defense consists of almost 32% of all plots at Zameen.
- Since DHA Defense covers a lot of areas (Phase 1 - Phase 10), it may have to be broken down further. Since there are fluctuations in data due to the different phases of Defence.
- Bahria Town and Bahria Orchard have almost similar scatter plots and relationships. So, it might be plausible to combine them without losing too much accuracy.
- No concrete conclusion can be made about the relationship between price and area if we look at all types of plots. The relationship can be better understood by narrowing down the types and then by location. It shows a positive relationship.
- No concrete conclusion can be made about the relationship between price and area if we look at all types of plots. The relationship can be better understood by narrowing down the types and then by location. It shows a positive relationship.
- The relationship between price and instalment value can be assessed easily using scatter plots. It can be said that they have a strong positive relationship.
- Amenities are independent of any attribute. This is evident in the scatter plots. One explanation may be that the data has not been scrapped properly as intuition tells us that amenities should increase with area and price. Another reason can be that the website does not list the amenities correctly.

- If we look at all plots, price and area are not as strongly related however, the relationship increases when we take into account the type of plot. It is most significant in the residential plots. It is weaker in residential plots and plot files.
- The correlation matrices confirm that amenities do not have a very strong relationship with the price. The small relationship can be dismissed because of the very small value of correlation across all matrices.
- Price is very strongly related (positively) to the Instalment value.
- Area has a positive relationship with price and Instalment value.

In conclusion, the dataset that was prepared in D1 was explored which gave us useful insights. These insights will greatly help us in predicting the price. Some adjustments to the dataset will also be made in order to ger a better system.