Assignment 2- Introduction To Data Science

22L-7503

BDS-3A

Part A. Preprocessing

1. In this step, you are required to apply the preprocessing steps that you've covered in the course. Specifically, for each of the input dimensions, fill in the following (add rows and complete the table for all input dimensions).

Iris:

Dim Name	Data Type	Total Instances	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	Std_ Dev
SepalLength	Float 64	150	0	0	4.30	7.90	5.0	5.84	5.80	0.672	0.82
Sepal Width	Float 64	150	0	1	2.0	4.40	3.0	3.05	3.00	0.185	0.43
PetalWidth	Float 64	150	0	0	0.10	2.50	0.2	1.12	1.30	0.578	0.76

Titanic:

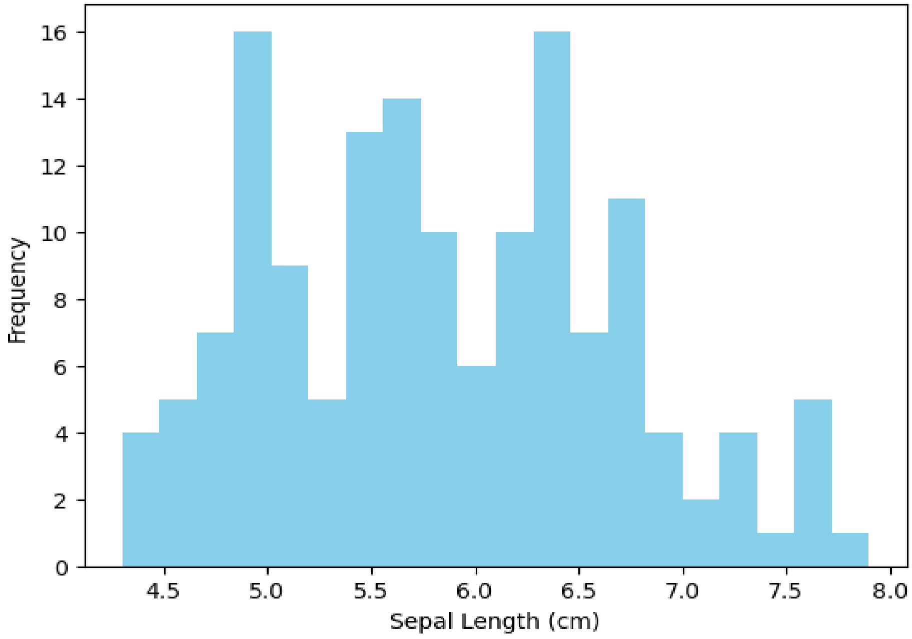
Dim Name	Data Type	Total Instances	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	Std_ Dev
Age	Float 64	714	177	0	0.42	80.0	24.0	29.7	28.0	210.25	14.5
SibSp	Int64 (Disc. Data)	891	0	30	0.00	8.00	0.00	0.52	0.00	1.21	1.10
Fare	Float 64	891	0	20	0.00	51.32	8.05	32.2	14.4	2470	49.7

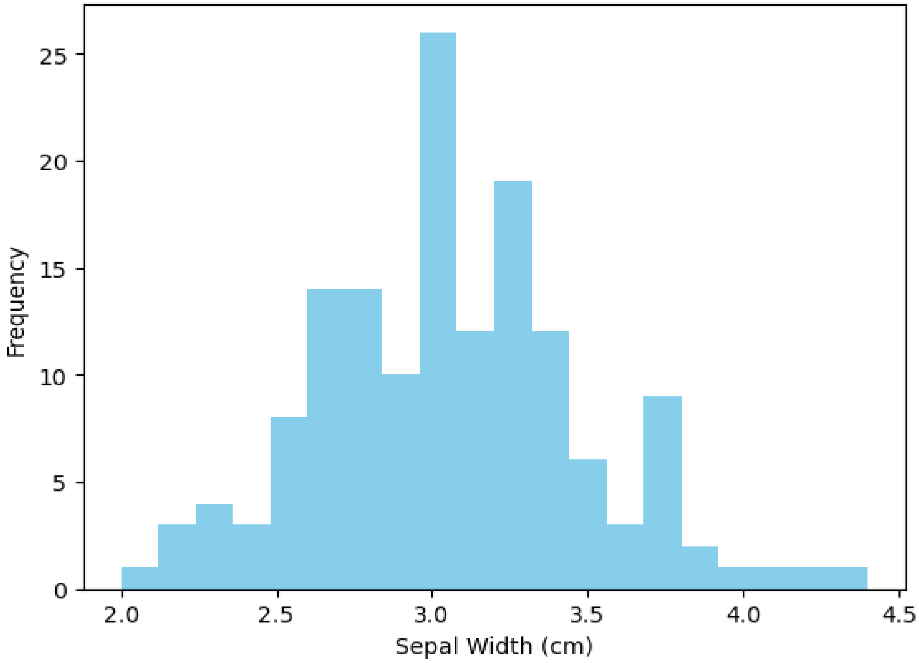
Housing Prices

Dim Name	Data Type	Total Instances	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	Std_ Dev
Area	int64	545	0	7	1650	16200	6000	5.1k	4600	4.7M	2.1k
Price	int64	545	0	6	1.75M	3.43M	3.5M	4.7M	4.3M	3.45B	4.7M
Bedrooms	int64	545	0	2	1	6	3	2.96	3	0.544	2.96

2. For each of the input dimension, plot histogram and comment the type of distribution the dimension exhibits. Further, visualize each dimension using a Box Plot. Specifically, for each of the input dimension, you're required to fill the following table (duplicate it for each of the 15 dimensions).

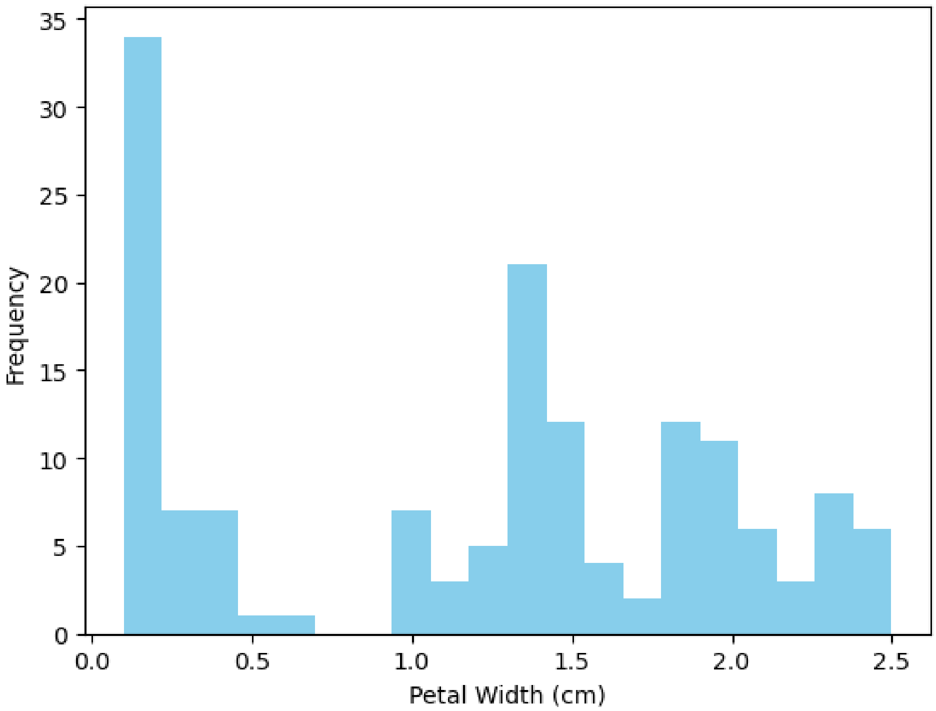
Iris:

SepalLength
Histogram

Comments: Values are clustered on the left side, this means that the values are positively skewed. Also mentioning that, most of the data lies in the lower quartiles.

SepalHeight
Histogram

Comments: The graph is almost and nearly symmetric, but it is Positively Skewed as the mean, median and mode are almost the same.

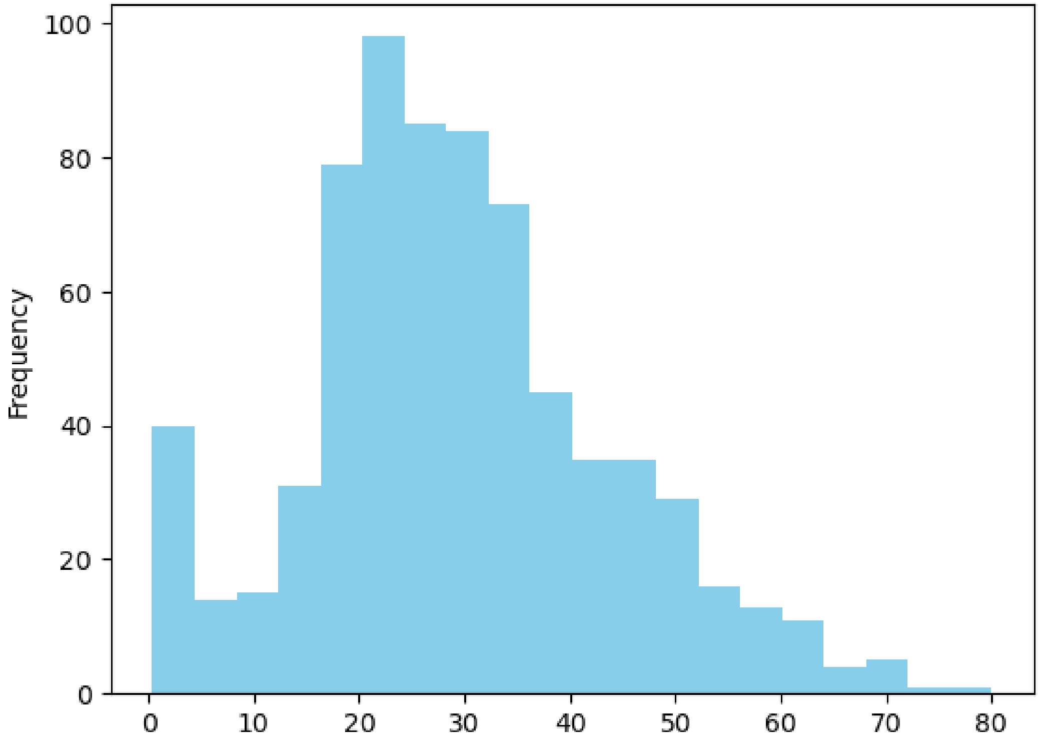
PetalWidth

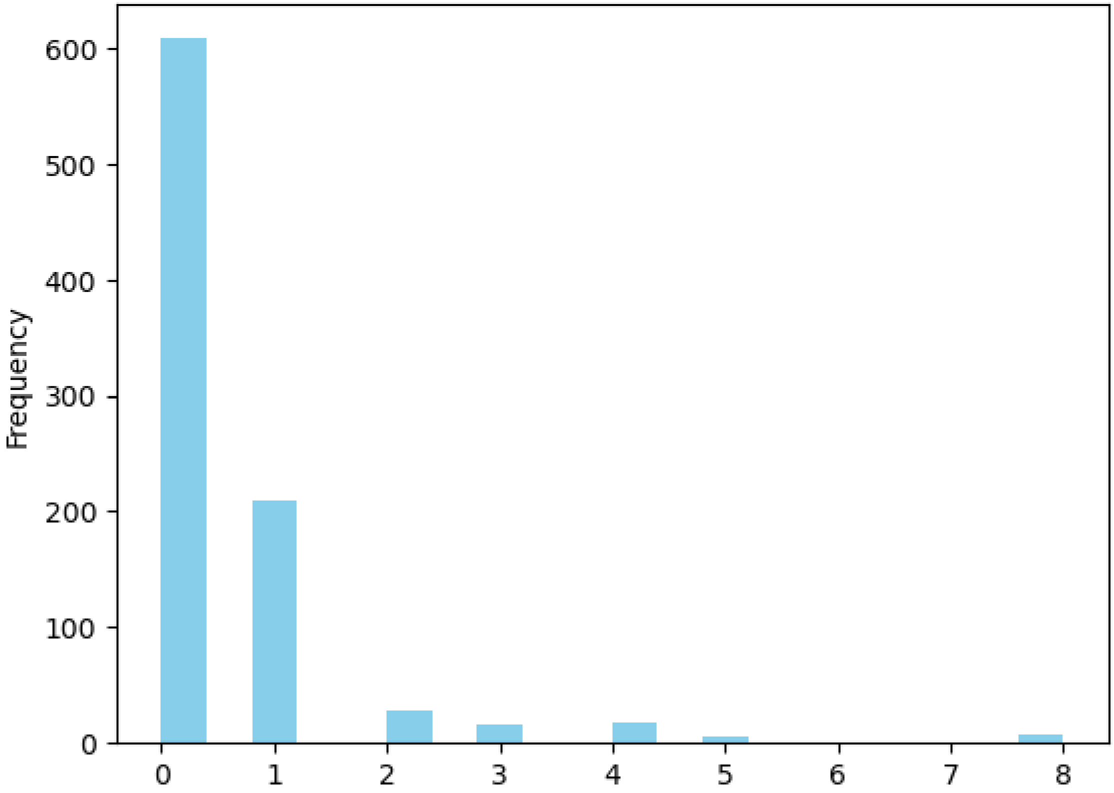
Histogram

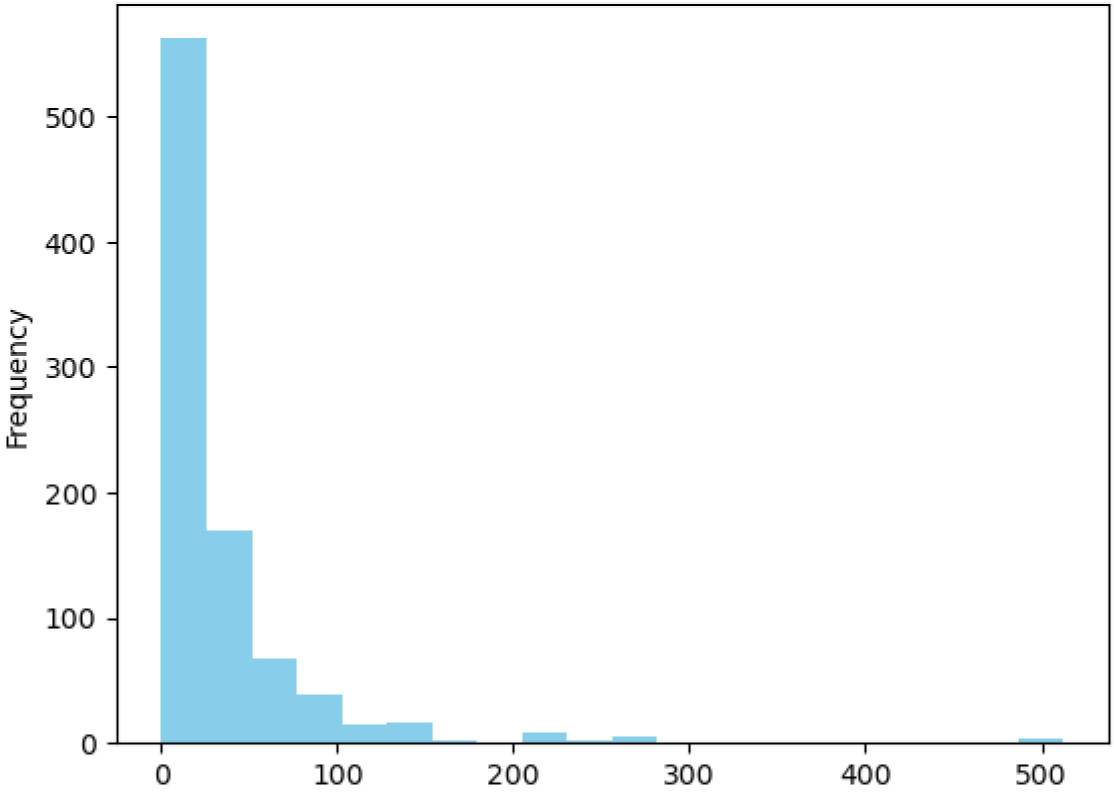


Comments: The Graph is Positively skewed, the most data lies in the lower quartile of the data set.

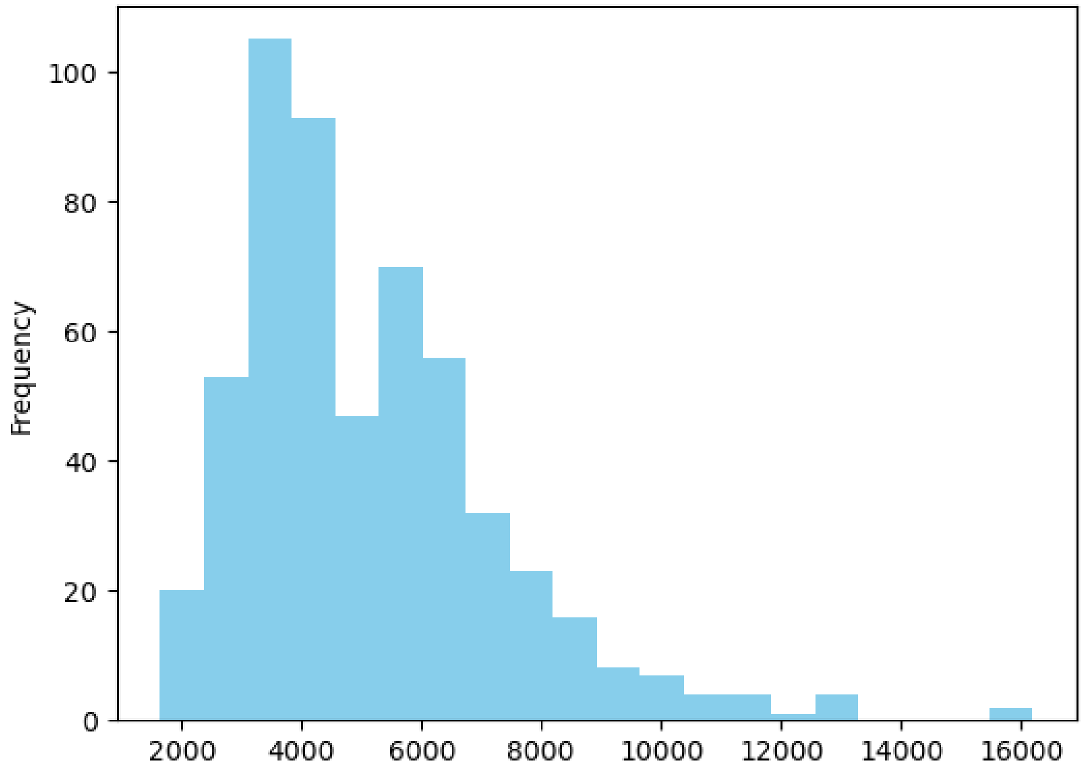
Titanic:

Age
Histogram

Comments:The Graph is Positively skewed, most the people in the ship were aged 20 to 40 as the data lies more in these percentiles.

SibSp
Histogram

Comments: The graph is Positively Skewed, stating that most of the people onboard were traveling alone.

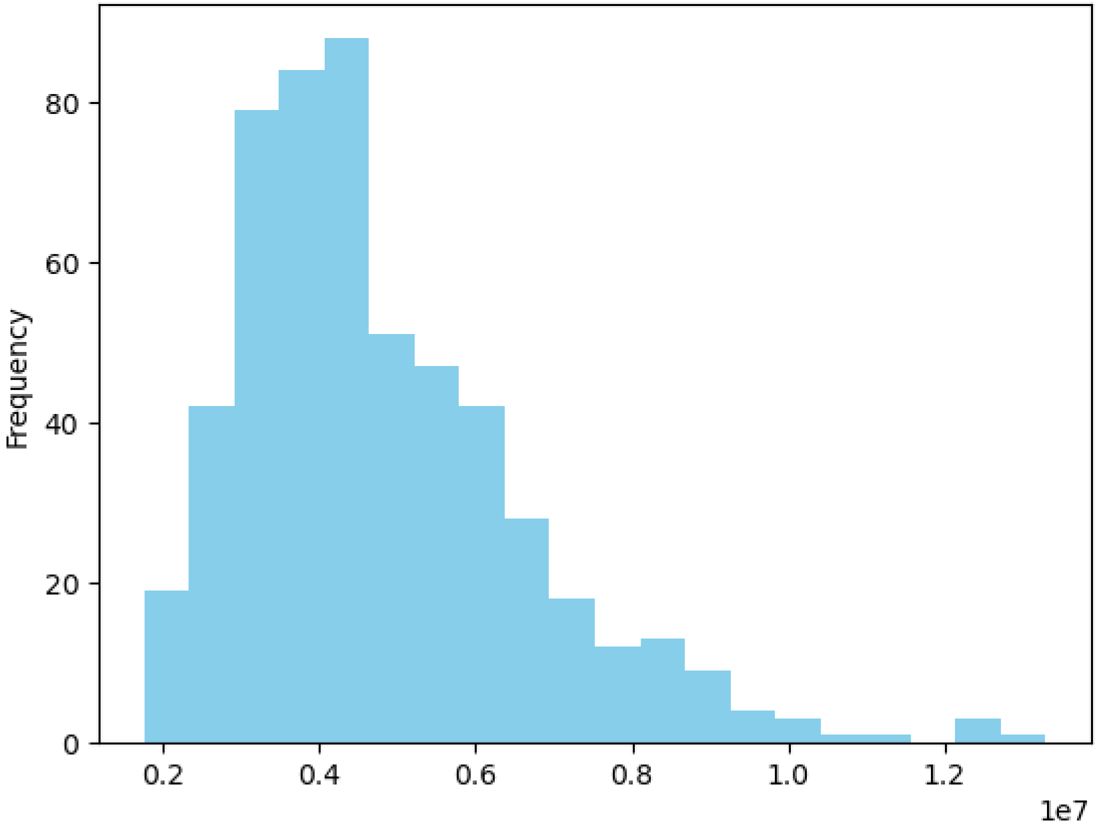
Fare
Histogram

Comments: The Graph is Positively Skewed.

Housing Prices:

Area
Histogram

Comments: Positively Skewed Graph, representing most data is based on small aread houses.

Price

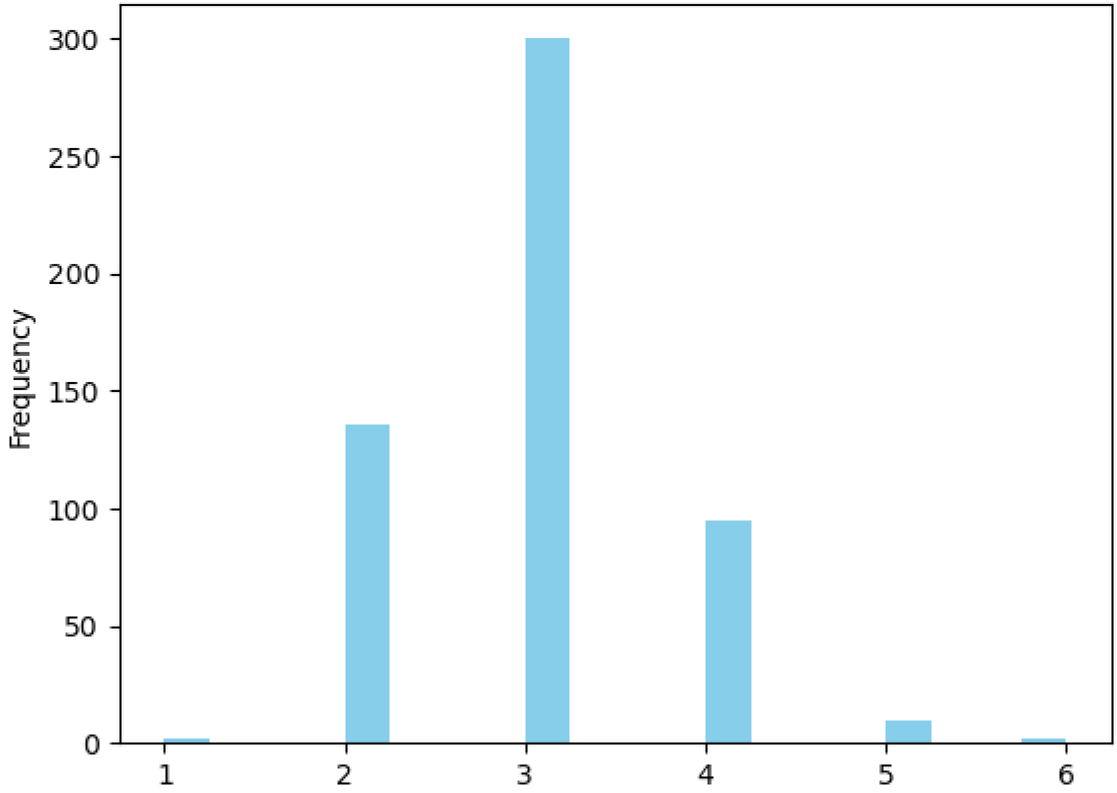
Histogram



Comments: Positively Skewed Graphs that represent that the prices above 2M are most common.

Bedrooms

Histogram



Comments: It is a discrete data, yet the data is positively skewed, with 3 as the most houses with these number of bedroom.

3. Find the missing values in each of the dimension (do this for both input and output dimensions), and fill these using an “appropriate” methodology that we’ve discussed in the class. You may also choose to drop a certain sample based on your analysis. Mention your approach and its justification.

Iris: .

Dim Name	Number of Missing Values	Filled using OR Dropped	Reason for selecting a certain approach
SepalLength	0	-	-
SepalWidth	0	-	-
PetalWidth	0	-	-

Titanic:

Dim Name	Number of Missing Values	Filled using OR Dropped	Reason for selecting a certain approach
Age	177	Filled using median	The median is less sensitive to outliers than mean, replacing it with mean will disturb the graph and median will also help in maintaining the age frequency.
SibSp	0	-	-
Fare	0	-	-

Housing Prices:

Dim Name	Number of Missing Values	Filled using OR Dropped	Reason for selecting a certain approach
Area	0	-	-
Price	0	-	-
Bedrooms	0	-	-

4. For each of the dimension, find out the outliers (noisy data) and handle these appropriately.

Iris:

Dim Name	Number of Outliers	Smooth using/ Dropped	Reason for selecting a certain approach
SepalLength	0	-	-
SepalWidth	1	Replacing the Value with Median	The median is less sensitive to outliers than mean, replacing it with mean will disturb the graph and median will also help in maintaining the age frequency. And we cannot afford dropping because of the quantity of data.
PetalWidth	0	-	-

Titanic:

(Post Removing Nulls)

Dim Name	Number of Outliers	Smooth using/ Dropped	Reason for selecting a certain approach
Age	7	Smooth using median	Maintaining the Graph
SibSp	30	Smooth using median	Discrete Data
Fare	20	Smooth using median	Dropping many values is not possible, it can affect the total result.

Housing Prices:

Dim Name	Number of Outliers	Smooth using/ Dropped	Reason for selecting a certain approach
Area	7	Dropped	Already have Data,and the Data depends on multiple factors, it can result to false and miscalculations
Price	6	Dropped	Already have Data,and the Data depends on multiple factors, it can result to false and miscalculations
Bedrooms	2	Dropped	Already have Data,and the Data depends on multiple factors, it can result to false and miscalculations