Assignment 2- Introduction To Data Science

Instructions:

- **Submit only one colab (.ipynb) file and one this report file (.pdf).**
- **Files should be named as yourrollnumber.ipynb (22L7521.ipynb, 22L7521.pdf)**
- **You are provided with three dataset files (Iris, Titanic, Housing) .csv files**
- **You have to provide code for all three datasets of the necessary steps described in the tables of each question.**
- **Only the mentioned columns/features mentioned for each dataset should be used.**
- **IN Q.2 you are only required to make the histograms and leave the BoxPlot part.**

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 dimension, 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											
Sepal Width											
SepalHeight											

Titanic:

Dim Name	Data Type	Total Instances	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	Std_ Dev

Age											
SibSp											
Fare											

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											
Price											
Bedrooms											

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	Box Plot
Comments:	Comments:

SepalHeight	
Histogram	Box Plot
Comments:	Comments:

SepalWidth	
Histogram	Box Plot
Comments:	Comments:

Titanic:

Age	
Histogram	Box Plot
Comments:	Comments:

SibSp	
Histogram	Box Plot

Comments:	Comments:

Fare	
Histogram	Box Plot
Comments:	Comments:

Housing Prices:

Area	
Histogram	Box Plot
Comments:	Comments:

Price	
Histogram	Box Plot
Comments:	Comments:

Bedrooms	
Histogram	Box Plot
Comments:	Comments:

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			
SepalWidth			
SepalHeight			

Titanic:

Dim Name	Number of Missing Values	Filled using OR Dropped	Reason for selecting a certain approach
Age			
SibSp			
Fare			

Housing Prices:

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

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			
SepalWidth			
SepalHeight			

Titanic:

Dim Name	Number of Missing Values	Filled using OR Dropped	Reason for selecting a certain approach
Age			

SibSp			
Fare			

Housing Prices:

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