**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 | Std Dev | Variance |
| Sepal Length | Float 64 | 150 | 0 | 0 | 4.3 | 7.9 | 5 .0 | 5.843 | 5.8 | 0.828066 | 0.685694 |
| Sepal Width | Float 64 | 150 | 0 | 4 | 2.0 | 4.4 | 3.0 | 3.054 | 3.0 | 0.433594 | 0.188004 |
| Petal Length | Float 64 | 150 | 0 | 0 | 1.0 | 6.9 | 1.5 | 3.758 | 4.35 | 1.76442 | 3.113179 |
| Petal Width | Float 64 | 150 | 0 | 0 | 0.1 | 2.5 | 0.2 | 1.198 | 1.3 | 0.763161 | 0.582414 |

**Titanic:**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dim Name | Data Type | Total Instances | Number of Nulls | Number of Outliers | Min. Value | Max Value | Mode | Mean | Median | Std | Variance |
| Age | Float 64 | 714 (removing nulls) | 177 | 11 | 0.42 | 80.0 | 24.0 | 29.69911 | 28.0 | 14.526497 | 211.019125 |
| SibSp | Int 64 | 891 | 0 | 46 | 0 | 8 | 0 | 0.523008 | 0.0 | 1.102743 | 1.216043 |
| Fare | Float 64 | 891 | 0 | 116 | 0.0 | 512.3292 | 8.05 | 32.204208 | 14.4542 | 49.693429 | 2469.436846 |

**Housing Prices**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dim Name | Data Type | Total Instances | Number of Nulls | Number of Outliers | Min. Value | Max Value | Mode | Mean | Median | Std | Variance |
| Area | Int 64 | 545 | 0 | 12 | 1650 | 16200 | 6000 | 5150.5412 | 4600 | 2170.141023 | 4709512.057576 |
| Price | Int 64 | 545 | 0 | 15 | 1750000 | 13300000 | 3500000,4200000 | 4766729.2477 | 4340000.0 | 1870439.615657 | 3498544355820.573242 |
| Bedrooms | Int 64 | 545 | 0 | 12 | 1 | 6 | 3 | 2.965138 | 3.0 | 0.738064 | 0.544738 |

**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: Normal distribution | Comments: |

|  |  |
| --- | --- |
| **SepalWidth** | |
| Histogram | Box Plot |
|  |  |
| Comments: Normal distribution | Comments: |

|  |  |
| --- | --- |
| **PetalLength** | |
| Histogram | Box Plot |
|  |  |
| Comments: Right skewed | Comments: |

**Titanic:**

|  |  |
| --- | --- |
| **Age** | |
| Histogram | Box Plot |
|  |  |
| Comments: Left skewed | Comments: |

|  |  |
| --- | --- |
| **SibSp** | |
| Histogram | Box Plot |
|  |  |
| Comments: Most people travel alone | Comments: |

|  |  |
| --- | --- |
| **Fare** | |
| Histogram | Box Plot |
|  |  |
| Comments: majority of fares are low | Comments: |

**Housing Prices:**

|  |  |
| --- | --- |
| **Area** | |
| Histogram | Box Plot |
|  |  |
| Comments: majority lies between 2500-7500 | Comments: |

|  |  |
| --- | --- |
| **Price** | |
| Histogram | Box Plot |
|  |  |
| Comments: Majority prices are medium | Comments: |

|  |  |
| --- | --- |
| **Bedrooms** | |
| Histogram | Box Plot |
|  |  |
| Comments: Majority of houses have 3 bedrooms or 2 bedrooms | 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 | 0 | 0 | 0 |
| SepalWidth | 0 | 0 | 0 |
| SepalHeight | 0 | 0 | 0 |

**Titanic:**

|  |  |  |  |
| --- | --- | --- | --- |
| Dim Name | Number of Missing Values | Filled using OR Dropped | Reason for selecting a certain approach |
| Age | 177 | mean | its about 20% of data so droping it could result in biasness of model to be trained |
| SibSp | 0 | 0 | 0 |
| Fare | 0 | 0 | 0 |

**Housing Prices:**

|  |  |  |  |
| --- | --- | --- | --- |
| Dim Name | Number of Missing Values | Filled using OR Dropped | Reason for selecting a certain approach |
| Area | 0 | 0 | 0 |
| Price | 0 | 0 | 0 |
| Bedrooms | 0 | 0 | 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 | 0 | 0 |
| SepalWidth | 4 | dropped | Only 2.7% of data wont make a difference |
| SepalHeight | 0 | 0 | 0 |

**Titanic:**

|  |  |  |  |
| --- | --- | --- | --- |
| Dim Name | Number of Missing Values | Filled using OR Dropped | Reason for selecting a certain approach |
| Age | 11 | drop | Dont need peoples age unless you need to train your model according to age |
| SibSp | 46 | Drop | Can be dropped because its only 5% of data |
| Fare | 116 | drop | 13% of values are outliers so it can be dropped |

**Housing Prices:**

|  |  |  |  |
| --- | --- | --- | --- |
| Dim Name | Number of Missing Values | Filled using OR Dropped | Reason for selecting a certain approach |
| Area | 12 | drop | Its only 2.2% of the data |
| Price | 15 | Drop | Its around 2.5% of the data |
| Bedrooms | 12 | drop | dropping 2.2% of data wont change anything |