

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

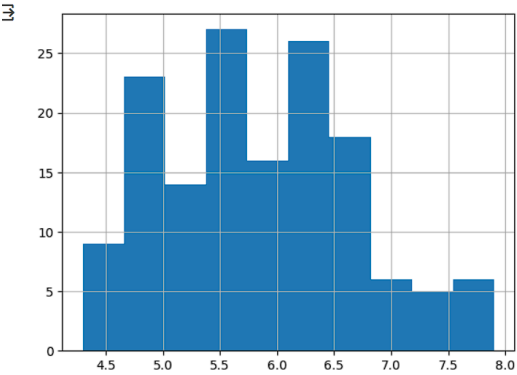
Housing Prices

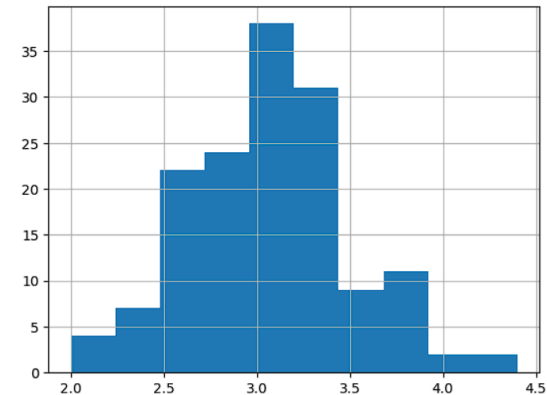
Dim Name	Data Type	Total Instances	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Std	Variance
Area	Int 64	545	0	12	1650	16200	6000	5150.5412	4600	2170.141023	4709512.057576
Price	Int 64	545	0	15	175000	1330000	350000,420000	4766729.2477	434000.0	1870439.615657	3498544355820.573242
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,

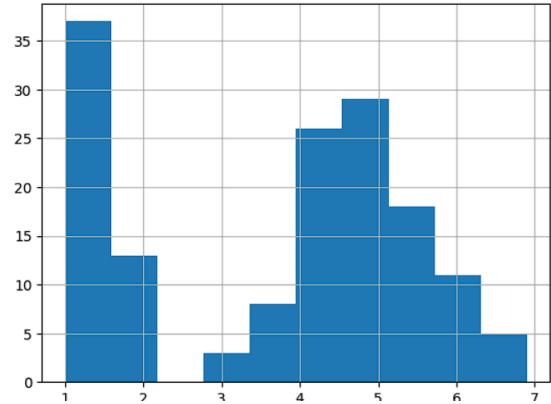
for each of the input dimension, you’re required to fill the following table (duplicate it for each of the 15 dimensions).

Iris:

SepalLength																							
Histogram	Box Plot																						
<pre>import matplotlib.pyplot as plt Iris_data['SepalLengthCm'].hist(bins=10) #SEPAL LENGTH plt.show()</pre>  <p>A histogram showing the frequency distribution of SepalLengthCm. The x-axis ranges from 4.5 to 8.0 with major ticks every 0.5 units. The y-axis ranges from 0 to 25 with major ticks every 5 units. The histogram consists of 10 blue bars. The distribution is roughly bell-shaped but slightly skewed to the right, with peaks around 5.5 and 6.2.</p> <table border="1"><caption>Histogram Data for SepalLengthCm</caption><thead><tr><th>Bin Range</th><th>Frequency</th></tr></thead><tbody><tr><td>4.5 - 5.0</td><td>9</td></tr><tr><td>5.0 - 5.5</td><td>23</td></tr><tr><td>5.5 - 6.0</td><td>14</td></tr><tr><td>6.0 - 6.5</td><td>27</td></tr><tr><td>6.5 - 7.0</td><td>16</td></tr><tr><td>7.0 - 7.5</td><td>26</td></tr><tr><td>7.5 - 8.0</td><td>18</td></tr><tr><td>8.0 - 8.5</td><td>6</td></tr><tr><td>8.5 - 9.0</td><td>5</td></tr><tr><td>9.0 - 9.5</td><td>6</td></tr></tbody></table>	Bin Range	Frequency	4.5 - 5.0	9	5.0 - 5.5	23	5.5 - 6.0	14	6.0 - 6.5	27	6.5 - 7.0	16	7.0 - 7.5	26	7.5 - 8.0	18	8.0 - 8.5	6	8.5 - 9.0	5	9.0 - 9.5	6	
Bin Range	Frequency																						
4.5 - 5.0	9																						
5.0 - 5.5	23																						
5.5 - 6.0	14																						
6.0 - 6.5	27																						
6.5 - 7.0	16																						
7.0 - 7.5	26																						
7.5 - 8.0	18																						
8.0 - 8.5	6																						
8.5 - 9.0	5																						
9.0 - 9.5	6																						
Comments: Normal distribution	Comments:																						

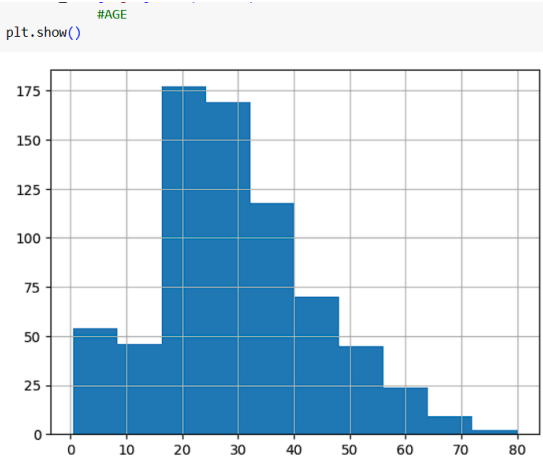
SepalWidth																							
Histogram	Box Plot																						
<pre>Iris_data['SepalWidthCm'].hist(bins=10) #SEPAL WIDTH plt.show()</pre>  <p>A histogram showing the frequency distribution of SepalWidthCm. The x-axis ranges from 2.0 to 4.5 with major ticks every 0.5 units. The y-axis ranges from 0 to 35 with major ticks every 5 units. The histogram consists of 10 blue bars. The distribution is roughly bell-shaped and centered around 3.0.</p> <table border="1"><caption>Histogram Data for SepalWidthCm</caption><thead><tr><th>Bin Range</th><th>Frequency</th></tr></thead><tbody><tr><td>2.0 - 2.5</td><td>4</td></tr><tr><td>2.5 - 3.0</td><td>7</td></tr><tr><td>3.0 - 3.5</td><td>22</td></tr><tr><td>3.5 - 4.0</td><td>24</td></tr><tr><td>4.0 - 4.5</td><td>38</td></tr><tr><td>4.5 - 5.0</td><td>31</td></tr><tr><td>5.0 - 5.5</td><td>9</td></tr><tr><td>5.5 - 6.0</td><td>11</td></tr><tr><td>6.0 - 6.5</td><td>2</td></tr><tr><td>6.5 - 7.0</td><td>2</td></tr></tbody></table>	Bin Range	Frequency	2.0 - 2.5	4	2.5 - 3.0	7	3.0 - 3.5	22	3.5 - 4.0	24	4.0 - 4.5	38	4.5 - 5.0	31	5.0 - 5.5	9	5.5 - 6.0	11	6.0 - 6.5	2	6.5 - 7.0	2	
Bin Range	Frequency																						
2.0 - 2.5	4																						
2.5 - 3.0	7																						
3.0 - 3.5	22																						
3.5 - 4.0	24																						
4.0 - 4.5	38																						
4.5 - 5.0	31																						
5.0 - 5.5	9																						
5.5 - 6.0	11																						
6.0 - 6.5	2																						
6.5 - 7.0	2																						

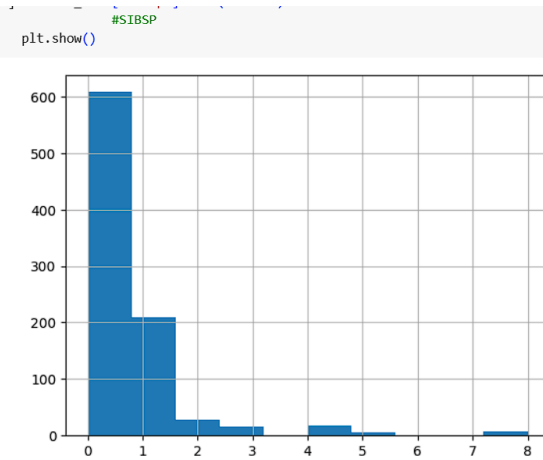
Comments: Normal distribution	Comments:
-------------------------------	-----------

PetalLength																											
Histogram	Box Plot																										
<pre>iris_data['Petal.Lengthcm'].hist(bins=10) #PETAL_LENGTH plt.show()</pre>  <table border="1"> <caption>Histogram Data (Approximate)</caption> <thead> <tr> <th>Petal.Lengthcm Bin</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1.0 - 1.5</td><td>37</td></tr> <tr><td>1.5 - 2.0</td><td>13</td></tr> <tr><td>2.0 - 2.5</td><td>13</td></tr> <tr><td>2.5 - 3.0</td><td>0</td></tr> <tr><td>3.0 - 3.5</td><td>3</td></tr> <tr><td>3.5 - 4.0</td><td>8</td></tr> <tr><td>4.0 - 4.5</td><td>26</td></tr> <tr><td>4.5 - 5.0</td><td>29</td></tr> <tr><td>5.0 - 5.5</td><td>18</td></tr> <tr><td>5.5 - 6.0</td><td>11</td></tr> <tr><td>6.0 - 6.5</td><td>11</td></tr> <tr><td>6.5 - 7.0</td><td>5</td></tr> </tbody> </table>	Petal.Lengthcm Bin	Frequency	1.0 - 1.5	37	1.5 - 2.0	13	2.0 - 2.5	13	2.5 - 3.0	0	3.0 - 3.5	3	3.5 - 4.0	8	4.0 - 4.5	26	4.5 - 5.0	29	5.0 - 5.5	18	5.5 - 6.0	11	6.0 - 6.5	11	6.5 - 7.0	5	
Petal.Lengthcm Bin	Frequency																										
1.0 - 1.5	37																										
1.5 - 2.0	13																										
2.0 - 2.5	13																										
2.5 - 3.0	0																										
3.0 - 3.5	3																										
3.5 - 4.0	8																										
4.0 - 4.5	26																										
4.5 - 5.0	29																										
5.0 - 5.5	18																										
5.5 - 6.0	11																										
6.0 - 6.5	11																										
6.5 - 7.0	5																										
Comments: Right skewed	Comments:																										

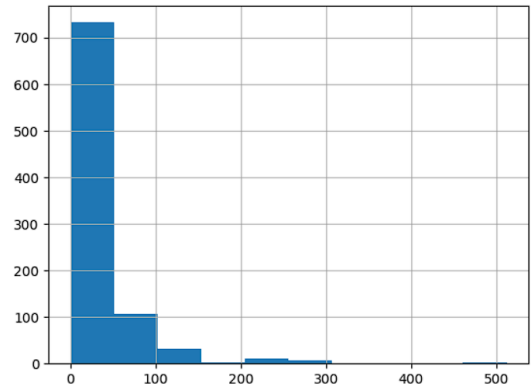
Titanic:

Age	
Histogram	Box Plot

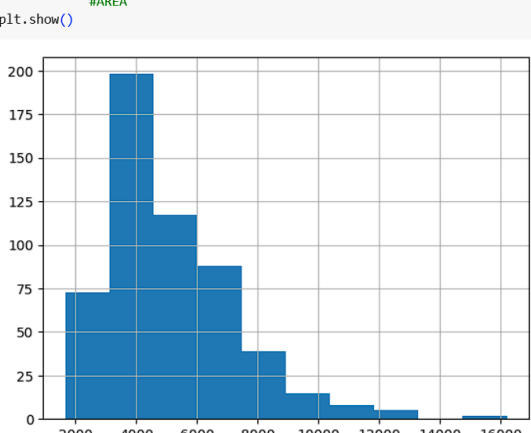
 <p>A histogram showing the distribution of AGE. The x-axis represents age from 0 to 80, and the y-axis represents frequency from 0 to 175. The distribution is left-skewed, with a peak frequency of approximately 175 for the age group 20-25.</p>	
Comments: Left skewed	Comments:

SibSp	
Histogram	Box Plot
 <p>A histogram showing the distribution of SibSp. The x-axis represents the number of siblings/spouses from 0 to 8, and the y-axis represents frequency from 0 to 600. The distribution is highly right-skewed, with a peak frequency of approximately 600 for the value 0.</p>	
Comments: Most people travel alone	Comments:

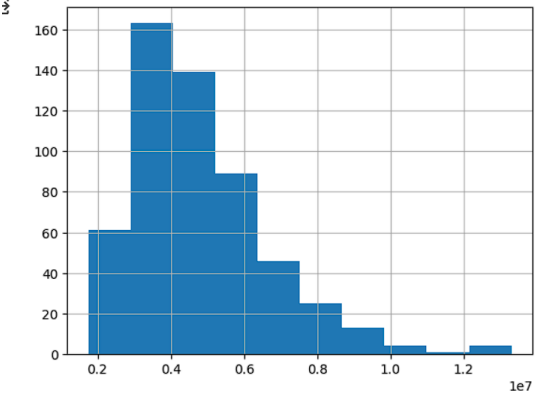
Fare	
Histogram	Box Plot

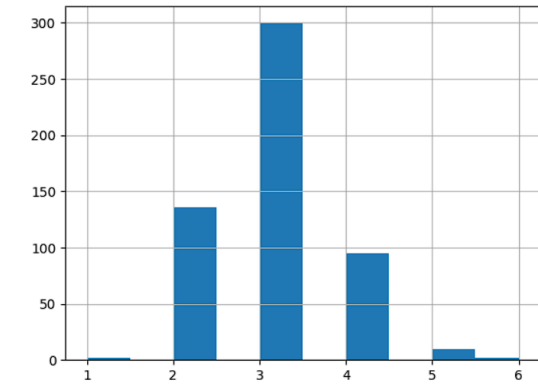
<pre>plt.show()</pre>  <p>A histogram showing the distribution of fares. The x-axis represents fare values from 0 to 500, and the y-axis represents the frequency from 0 to 700. The distribution is highly right-skewed, with a peak frequency of approximately 750 for fares between 0 and 50. The frequency drops sharply for higher fare values, with very few fares exceeding 300.</p>	
<p>Comments: majority of fares are low</p>	<p>Comments:</p>

Housing Prices:

Area	
Histogram	Box Plot
<pre>plt.show()</pre>  <p>A histogram showing the distribution of housing prices. The x-axis represents price values from 2000 to 16000, and the y-axis represents the frequency from 0 to 200. The distribution is right-skewed, with a peak frequency of approximately 200 for prices between 3000 and 4000. The frequency decreases as the price increases, with a long tail extending towards 16000.</p>	
<p>Comments: majority lies between 2500-7500</p>	<p>Comments:</p>

Price

Histogram	Box Plot
<pre>#PRICE</pre> <pre>plt.show()</pre> 	
Comments: Majority prices are medium	Comments:

Bedrooms	
Histogram	Box Plot
<pre>#BEDROOMS</pre> <pre>plt.show()</pre> 	
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 dropping 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