

Introduction to Data Science
Course Project
Report Document

<Muhammad Ahmad>

<21L-5617>

<Section 3A>

Instructions: Read These Carefully Before Starting

1. Due Date: Sunday 4th December 2022 – 11:59PM
2. Submission will be taken on Google Classroom
3. Submit only the following 2 files named like the following:
 - a. Code File (Jupyter Notebook): L210000_Code.ipynb
 - b. Report Document (This File): L210000_Report.pdf
4. Project will not be evaluated if:
 - a. You submit python (.py) files
 - b. You submit multiple .ipynb files
 - c. You submit compressed (.rar or .zip) files
 - d. You submit any files other than the required PDF and IPYNB
5. Upload data files directly to Google Colab - do not use Google Drive or GitHub linking method
6. All source files needed to complete this project are uploaded with it on Google Classroom.
7. Do not add the data file with your submission on Google Classroom.

Not following these instructions will lead to mark deduction.

Please try to use Microsoft Word instead of Google Docs to edit this document and to export it as a PDF file for final submission.

Happy Coding 🐱

TA Emails

Section A, C - Muhammad Maarij 1192347@lhr.nu.edu.pk

Section B, D - Hira Ijaz 1192377@lhr.nu.edu.pk

For this project you will be applying machine learning models (both regression and classification) to the dataset which contains information about various individuals, their clothing, and its properties along with other atmospheric elements such as temperature, pressure humidity etc. The users also provided feedback on if they feel cold or not. The feedback (through AMV and PMV) which is based on the following mapping:

The following table shows the mapping of sensations:

Value	Thermal Sensation
+3	hot
+2	warm
+1	slightly warm
0	neutral
-1	slightly cool
-2	cool
-3	cold

The dataset is given in an excel file named CollectedData.xlsx, see sheet 2 of excel file. The dimension names (column headers) are not mentioned in the given file. The table below describes the columns which will be of your interest.

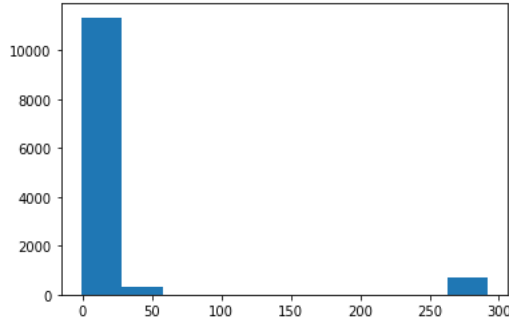
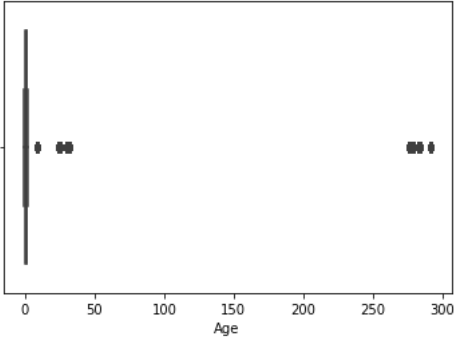
Column number	Feature Name	Feature Description
3	Age	Age
22	Clo	Clothing insulation
19	Met	Met Rate
26	Dewpt	Dewpt
27	PlaneRadTemp	plane radiant temperature
37	Ta	Average air temperature
38	Tmrt	Average mean radiant temperature
40	Vel	Air Velocity
42	AirTurb	Air Turbulance
43	Pa	Vapor Pressure
44	Rh	Humidity
74	TaOutdoor	Outdoor Air Temperature
77	RhOutdoor	Outdoor Humidity
8	AMV	Classification response variable
49	PMV	Regression response variable

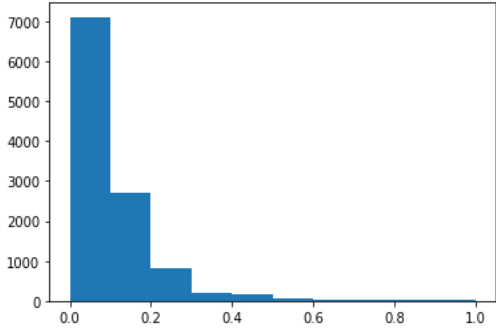
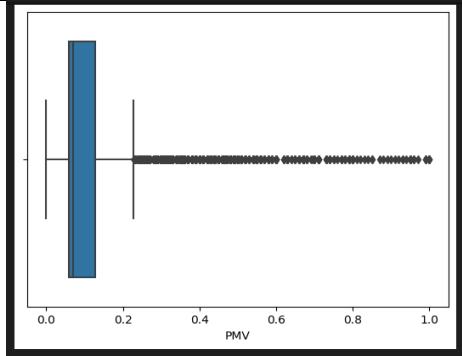
Part A. Preprocessing

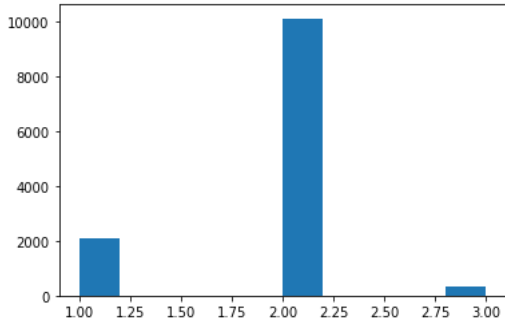
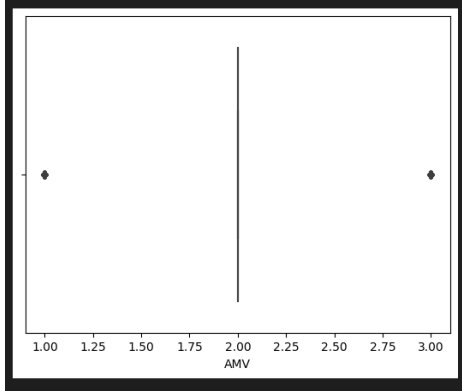
1. In this step, you are required to apply the preprocessing steps that you've covered in the course. Specifically, for each of the input dimension, fill in the following (add rows and complete the table for all input dimensions).

Dim Name	Data Type	Total Instances	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	STD
Age	Float64	18566	2916	4	0.00	1996	24.0	308.6	35.0	462556.5	680.1
Clo	Float64	18566	1406	71	0.150	2.13	0.77	0.778	0.751	0.049	0.22
Met	Float64	18566	1887	55	0.100	4.50	1.0	1.065	1.10	0.183	0.428
Dewpt	Float64	18566	3552	0	-1.95	26.89	17.4	13.621	14.10	34.845	5.903
PlaneRadTemp	Float64	18566	7022	245	-7.42	11.70	0.3	0.217	0.20	1.084	1.041
Ta	Float64	18566	20	138	15.96	31.00	23.2	23.179	23.13	2.053	1.432
Tmrt	Float64	18566	3701	277	16.61	37.44	22.5	23.450	23.35	2.257	1.502
Vel	Float64	18566	3700	161	0.00	1.880	0.1	0.1124	0.100	0.006	0.079
AirTurb	Float64	18566	5601	2	0.00	102.4	0.5	18.265	0.500	627.05	25.04
Pa	Float64	18566	4656	39	0.00	27.07	2.1	5.123	1.55	66.522	8.156
Rh	Float64	18566	35	0	7.400	79.30	64.0	42.528	43.27	226.84	15.06
TaOutdoor	Float64	18566	3568	10	-24.9	32.35	27.55	17.175	18.20	113.75	10.665
RhOutdoor	Float64	18566	19	1	0.00	100.3	0.0	61.098	68.79	610.30	24.704
AMV	Float64	18566	55	0	-3.00	3.00	0.0	0.100	0.00	1.214	1.102
PMV	Float64	18566	696	114	-4.17	2.500	0.1	-0.07	-0.03	0.289471	0.5388

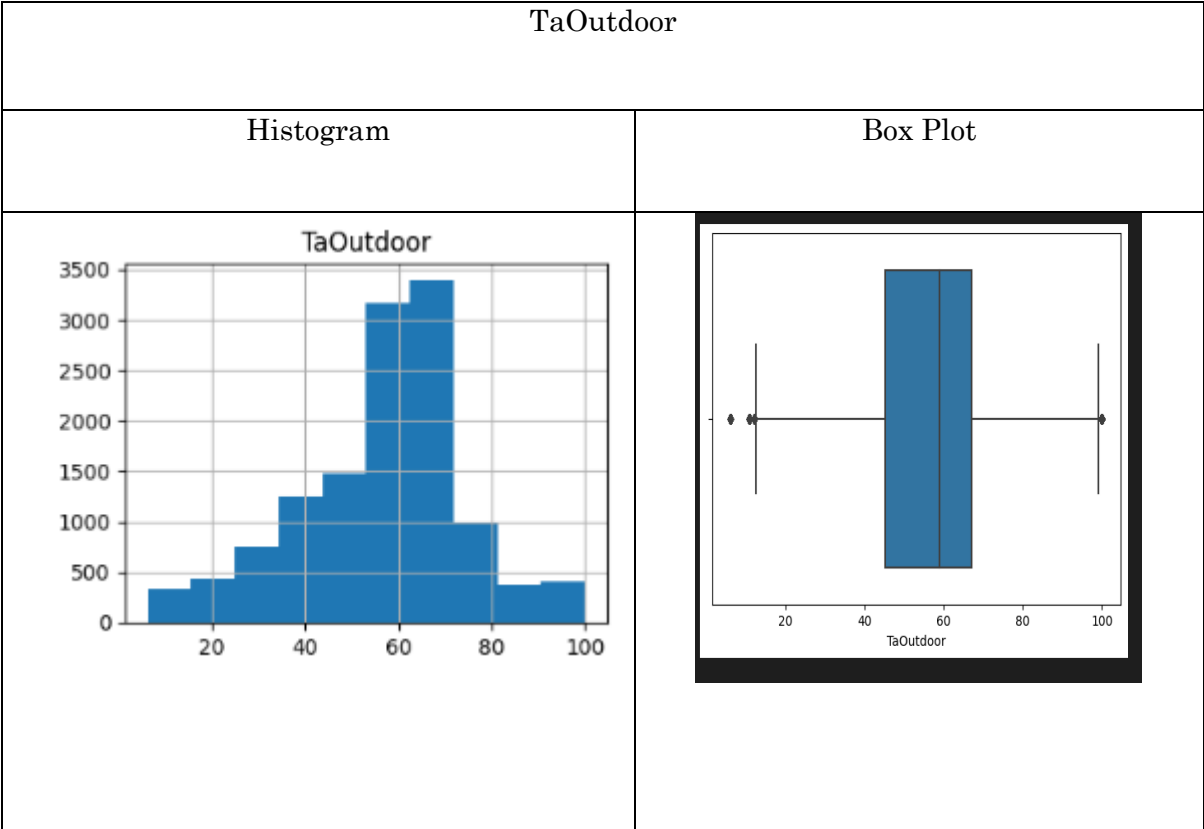
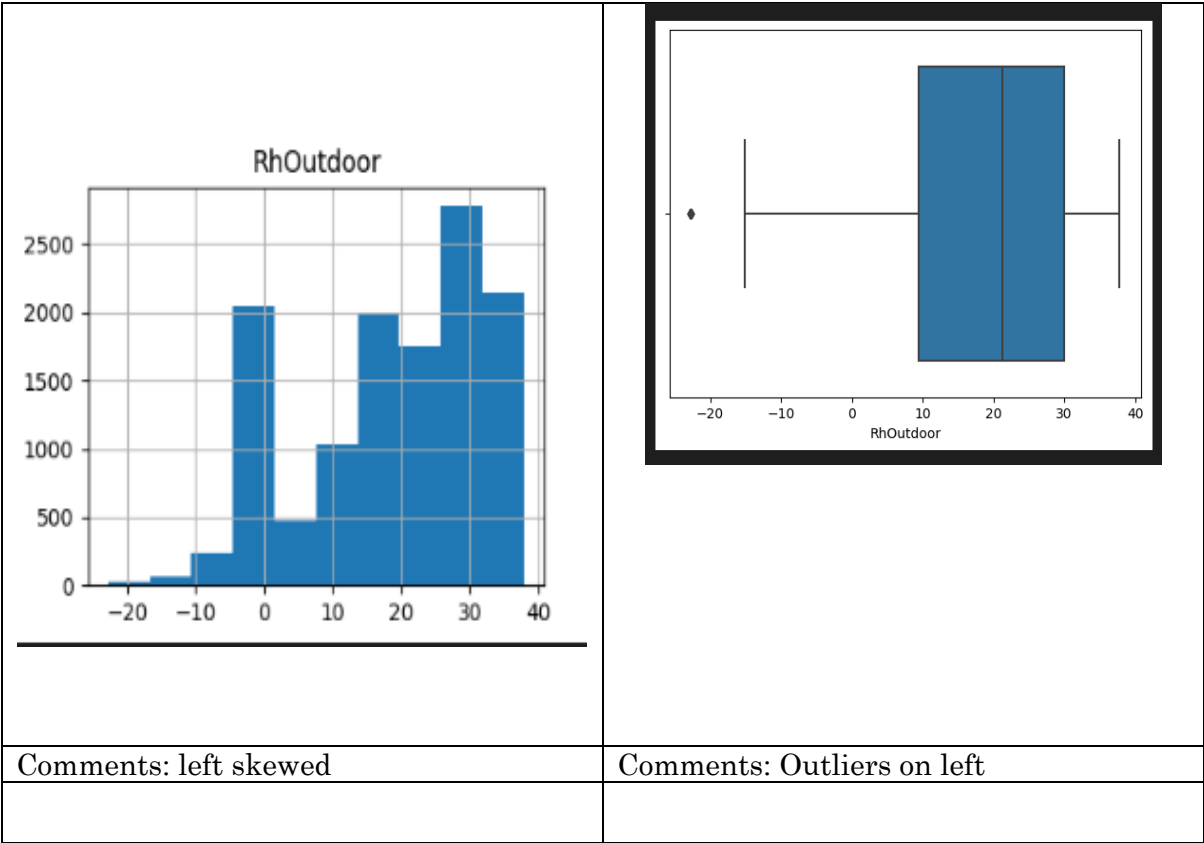
2. For each of the input dimension, plot histogram and comment the type of distribution the dimension exhibits. Further, visualize each dimension using a Box Plot. Specifically, for each of the input dimension, you're required to fill the following table (duplicate it for each of the 15 dimensions).

Age	
Histogram	Box Plot
	
Comments: right Skewwd	Comments: Outliers on righth

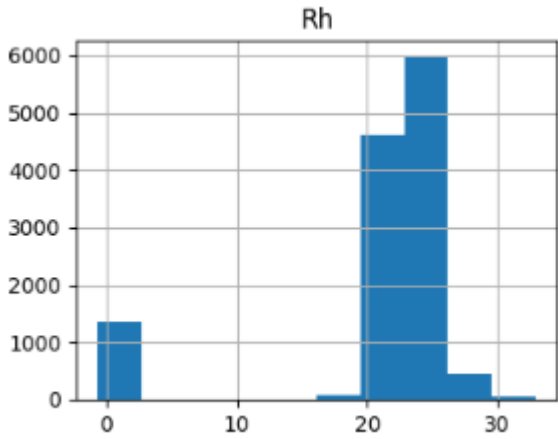
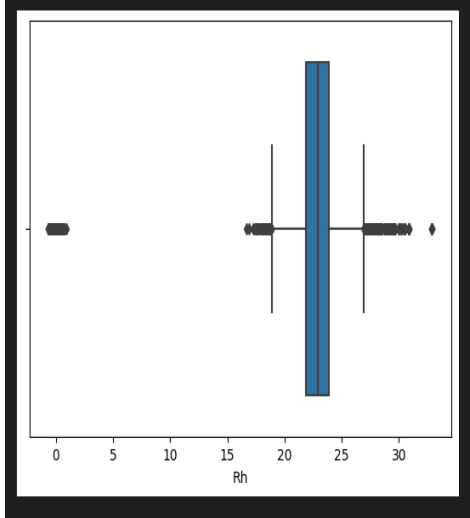
PMV	
Histogram	Box Plot
	
Comments: right Skewwd	Comments: outlies on rigth

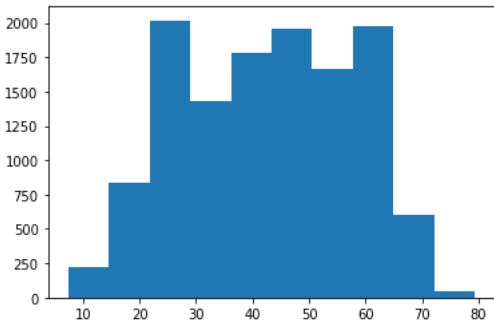
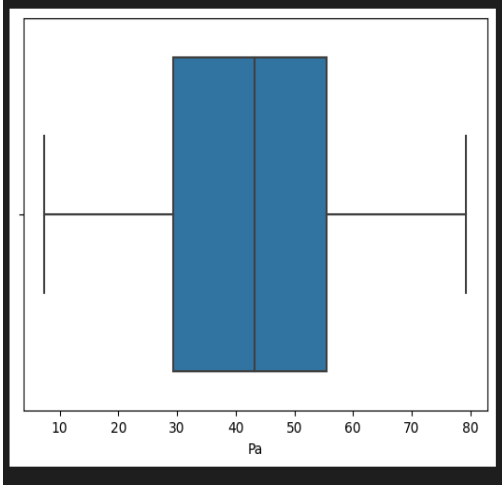
AMV	
Histogram	Box Plot
	
Comments:Normal	Comments: outliers on both sides

RhOutdoor	
Histogram	Box Plot

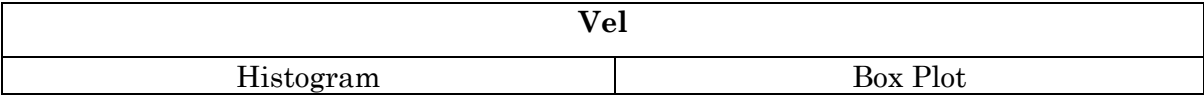
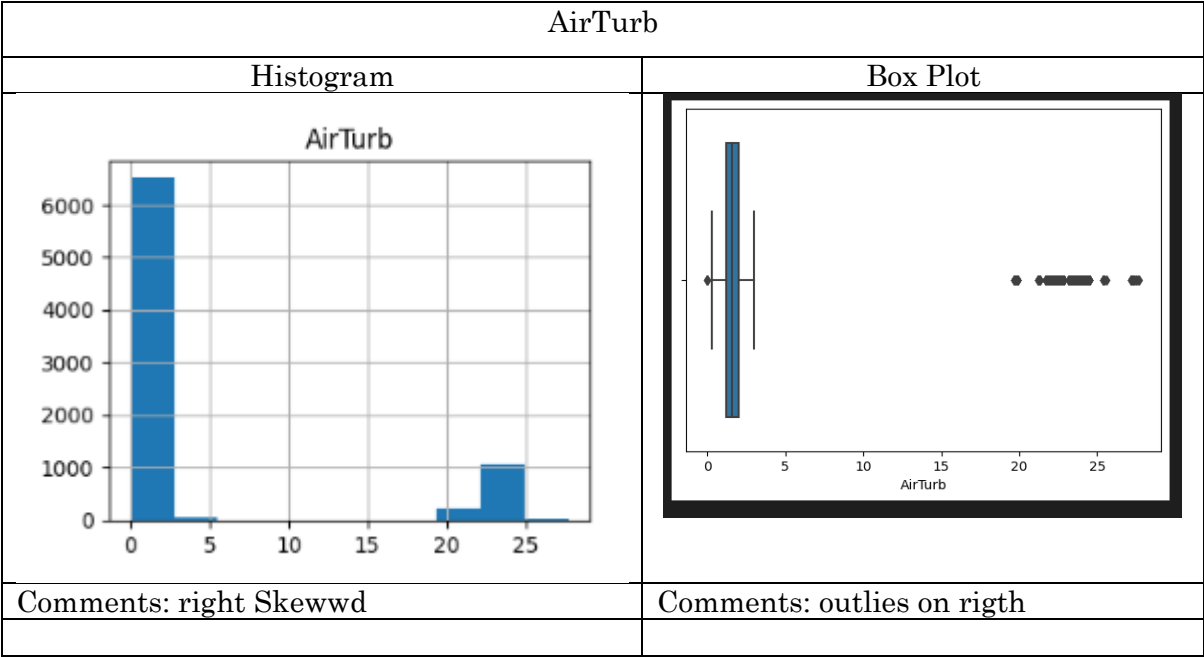
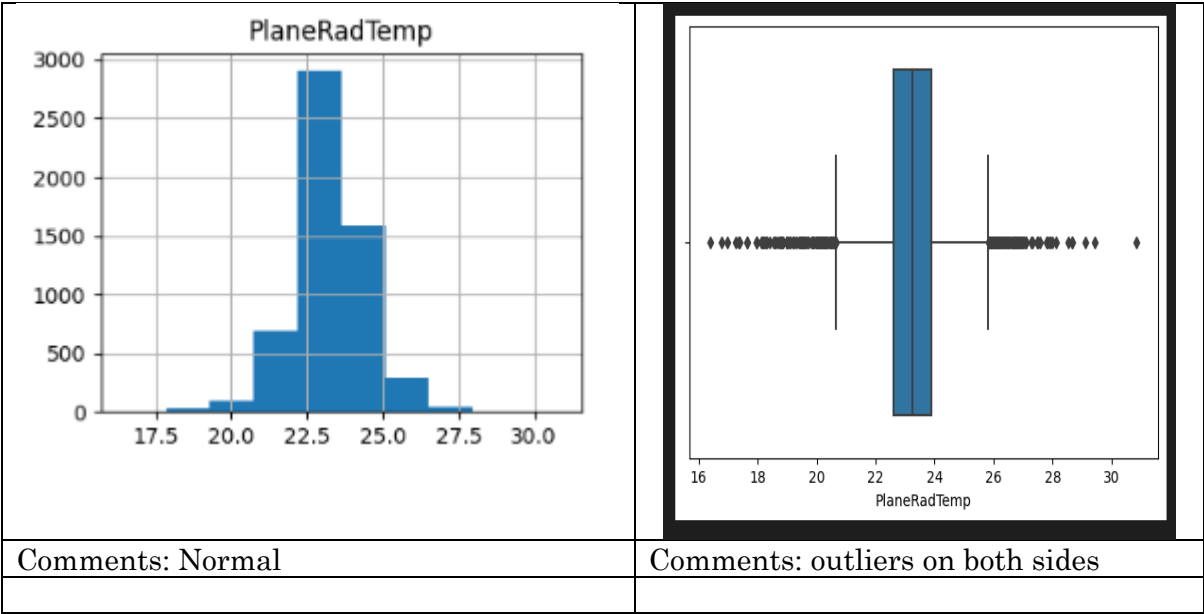


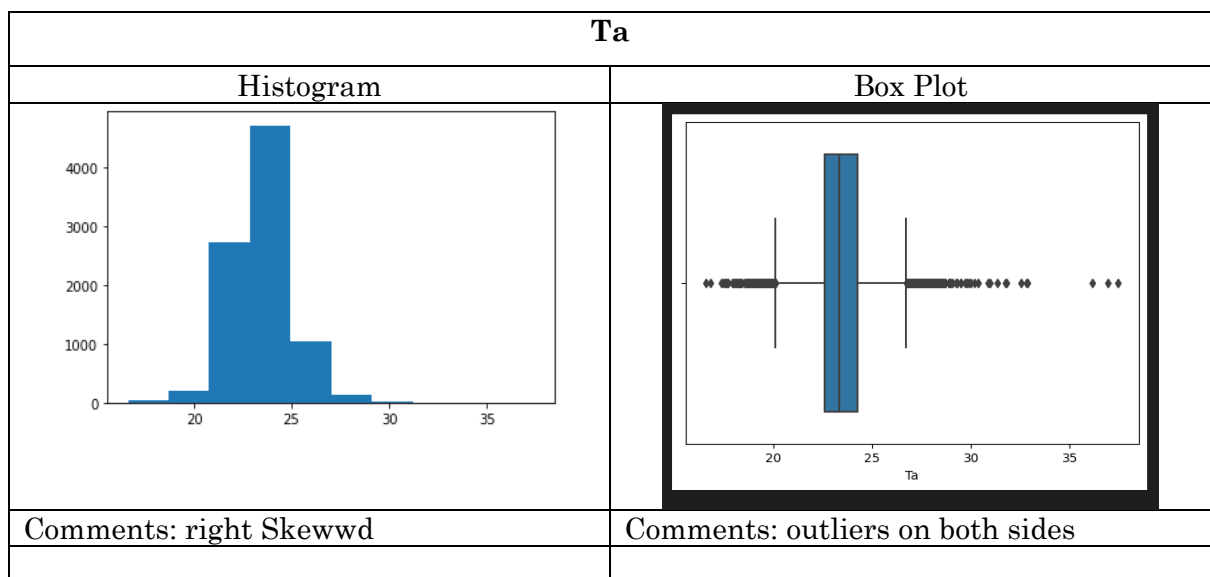
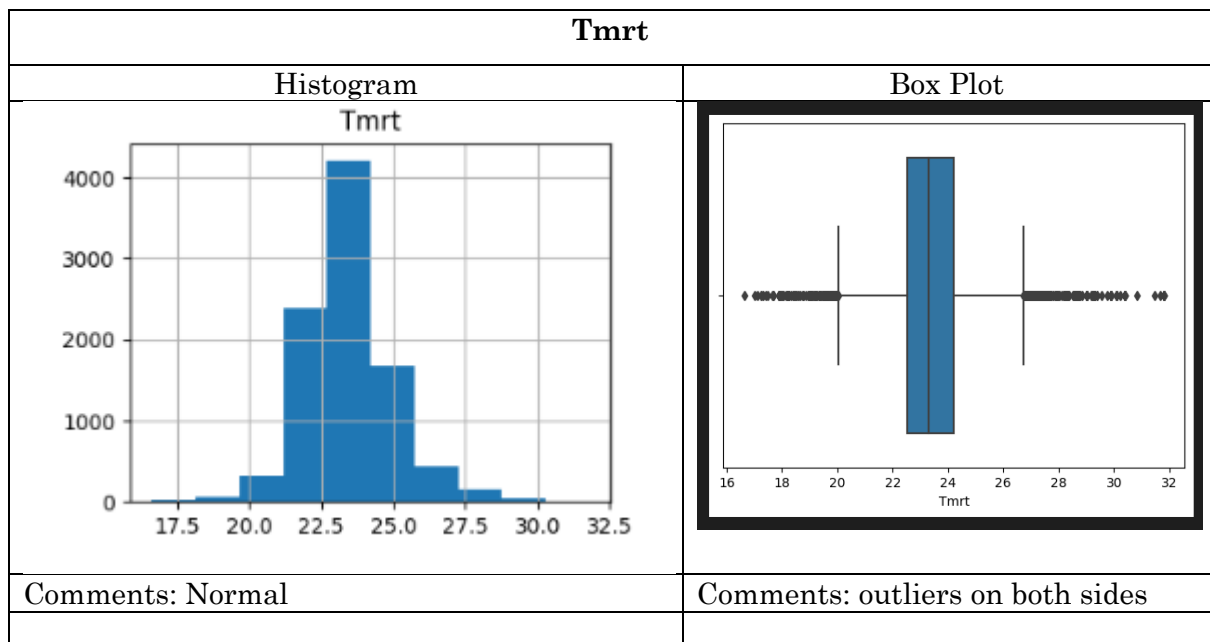
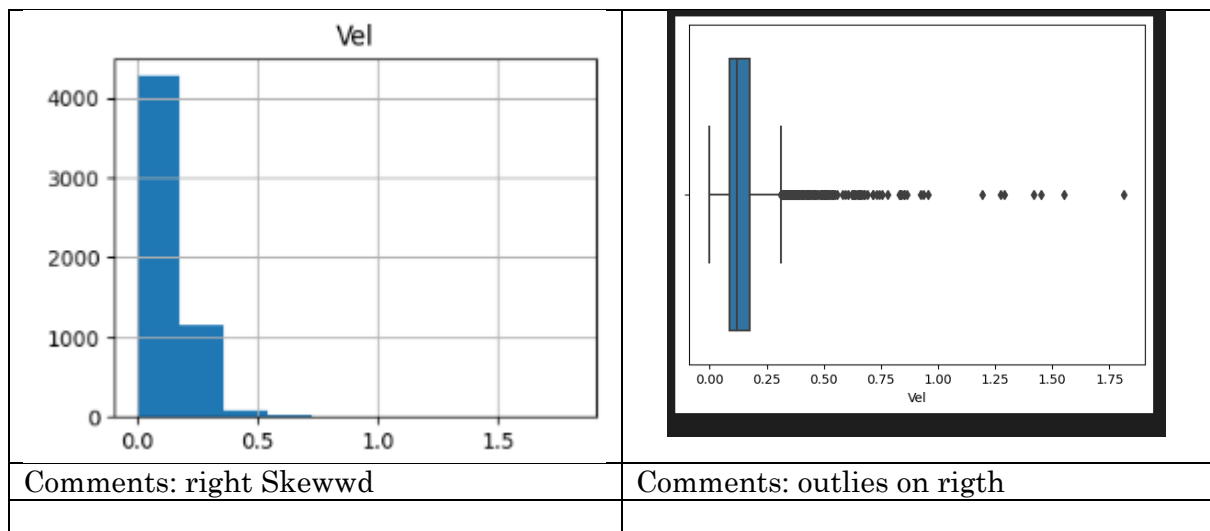
Comments: left skewed	Comments: outliers on both sides

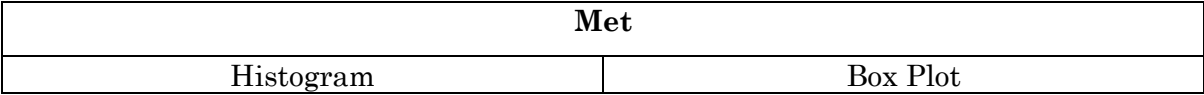
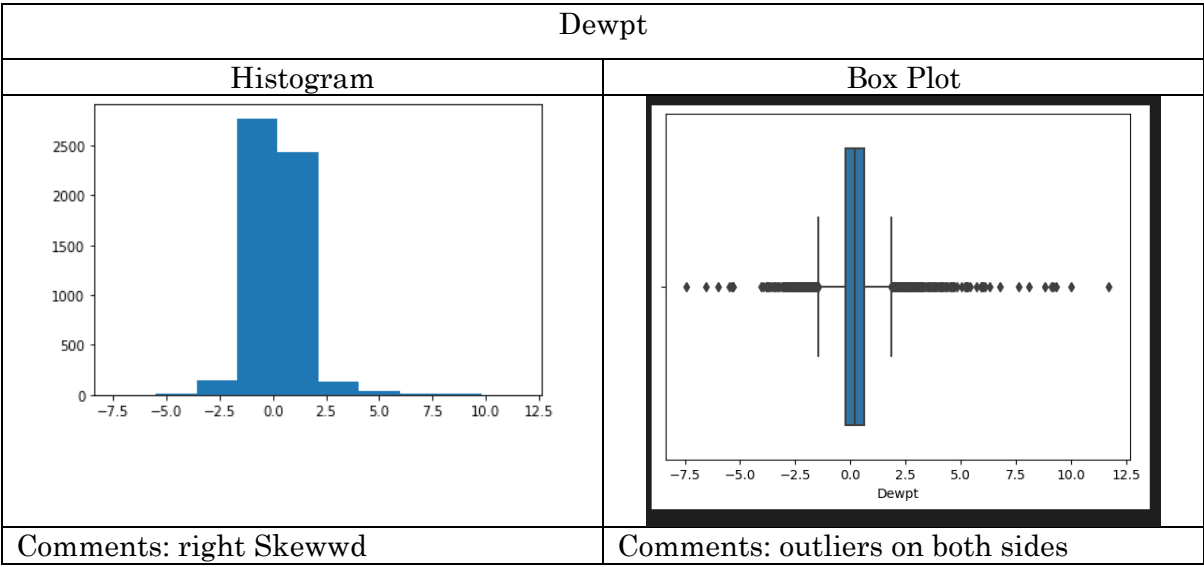
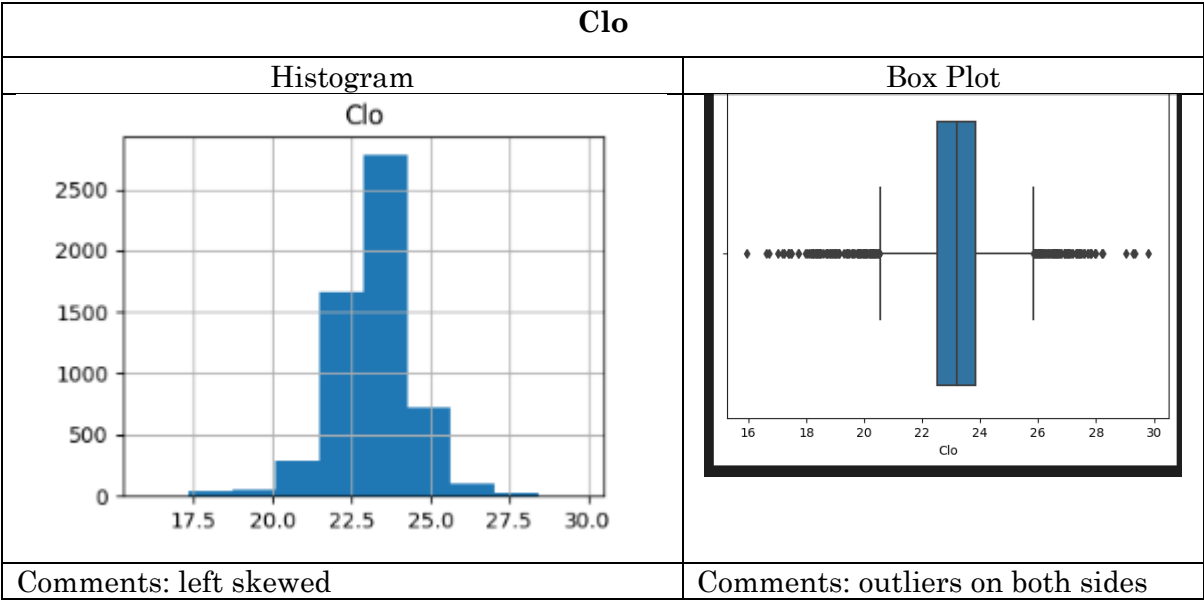
Rh	
Histogram	Box Plot
	
Comments: left skewed	Comments: outliers on both sides

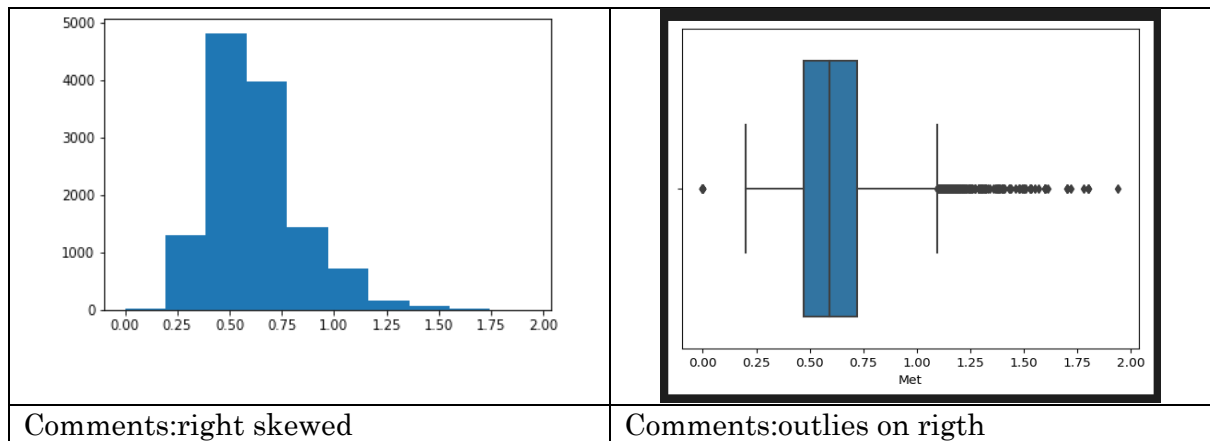
Pa	
Histogram	Box Plot
	
Comments: Normal	Comments: No outliers

PlaneRadTemp	
Histogram	Box Plot









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.

Dim Name	Number of Missing	Filled Using OR Dropped	Reason for selecting a certain Approach
Age	2916	Median	Data is Skewed
Clo	1406	Median	Data is Skewed
Met	1887	Mean	Data is not Skewed
Dewpt	3552	Median	Data is Skewed
PlaneRadTemp	7022	Median	Data is Skewed
Ta	20	Median	Data is Skewed
Tmrt	3701	Median	Data is Skewed
Vel	3700	Median	Data is Skewed
AirTurb	5601	Median	Data is Skewed
Pa	4656	Mean	Data is not Skewed
Rh	35	Median	Data is Skewed
TaOutdoor	3568	Mean	Data is not Skewed
RhOutdoor	19	Median	Data is Skewed
AMV	55	Median	Data is Skewed
PMV	696	Median	Data is Skewed

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

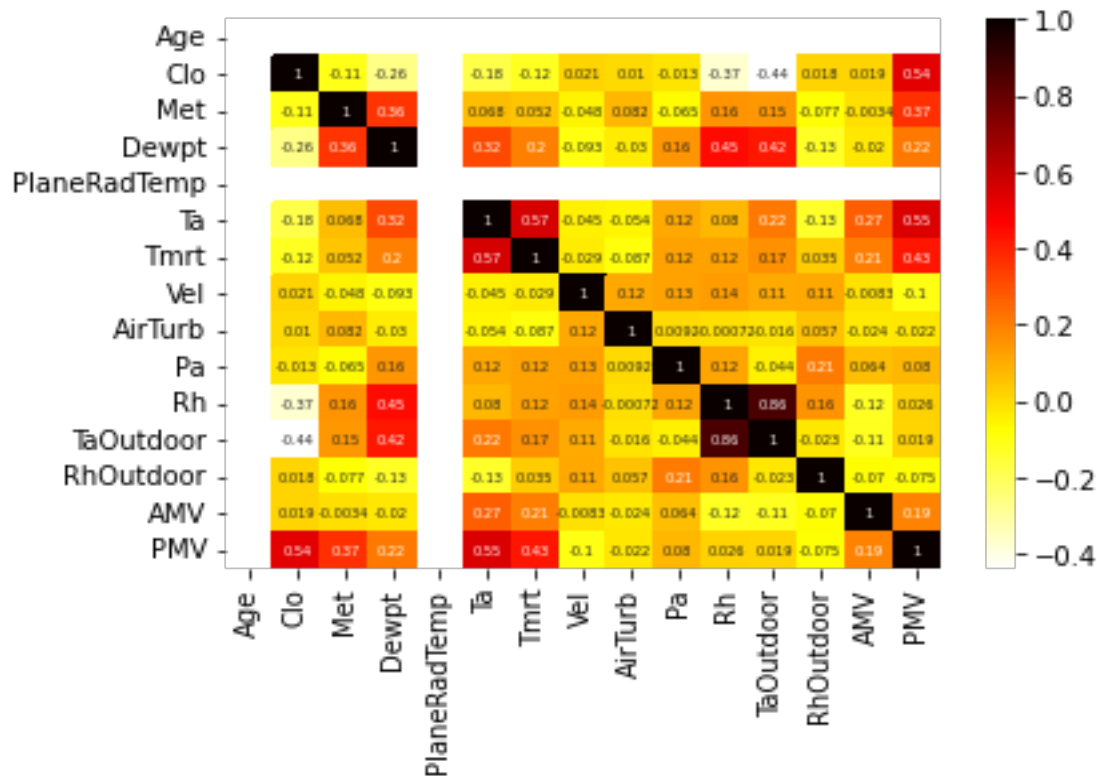
Dim Name	Number of Outliers	Smooth using/ Dropped	Reason for selecting a certain Approach
Age	4	Z-Score	Removal of outliers
Clo	71	Z-Score	Removal of outliers
Met	55	Z-Score	Removal of outliers
Dewpt	0	Z-Score	Removal of outliers
PlaneRadTemp	245	Z-Score	Removal of outliers
Ta	138	Z-Score	Removal of outliers
Tmrt	277	Z-Score	Removal of outliers
Vel	161	Z-Score	Removal of outliers
AirTurb	2	Z-Score	Removal of outliers
Pa	39	Z-Score	Removal of outliers
Rh	0	Z-Score	Removal of outliers
TaOutdoor	10	Z-Score	Removal of outliers
RhOutdoor	1	Z-Score	Removal of outliers
AMV	0	Z-Score	Removal of outliers
PMV	114	Z-Score	Removal of outliers

5. Using the variance that you've calculated above, for each dimension, comment whether you'll select the input dimension or no. (don't drop a dimension at this point)

Dim Name	Variance	Apply filter or no, reason
Age	462556.5	Low Variance Filter to remove low variance to become useful data
Clo	0.049	Low Variance Filter to remove low variance to become useful data
Met	0.183	Low Variance Filter to remove low variance to become useful data
Dewpt	34.845	Low Variance Filter to remove low variance to become useful data
PlaneRadTemp	1.084	Low Variance Filter to remove low variance to become useful data
Ta	2.053	Low Variance Filter to remove low variance to become useful data
Tmrt	2.257	Low Variance Filter to remove low variance to become useful data
Vel	0.006	Low Variance Filter to remove low variance to become useful data
AirTurb	627.05	Low Variance Filter to remove low variance to become useful data
Pa	66.522	Low Variance Filter to remove low variance to become useful data
Rh	226.84	Low Variance Filter to remove low variance to become useful data

TaOutdoor	113.75	Low Variance Filter to remove low variance to become useful data
RhOutdoor	610.30	Low Variance Filter to remove low variance to become useful data
AMV	1.214	Low Variance Filter to remove low variance to become useful data
PMV	0.289471	Low Variance Filter to remove low variance to become useful data

6A. Create a correlation matrix (Heat Map) for all the dimensions (input and output).



[Add correlation matrix here]

6B. Using the above correlation matrix, comment what are the most informative dimensions, and which are the least. Note that, be careful since we have two response variables in the dataset (i.e., PMV and AMV regression and classification respectively)

The least informative dimension is “PlaneRadTemp, Age”

The most informative dimension is “PMV”

7. Apply entropy followed by information gain on the selected columns. Specify your selection criteria.

Dim Name	Entropy	Info Gain	Reason
Age	0.0		
Clo	4.78		
Met	1.99		
Dewpt	1.86		
PlaneRadTemp	0.0		
Ta	3.47		
Tmrt	4.2		
Vel	1.911		
AirTurb	0.05		
Pa	2.928		
Rh	6.497		
TaOutdoor	5.155		
RhOutdoor	5.044		
AMV	2.09		
PMV	4.90		

Part B. Applying Algorithms

1. For this part, split the data randomly into 80/20 percent. Where 80% represents the training data. Also normalize the dataset as you see fit.

```
scaler = MinMaxScaler()
normalized_data = scaler.fit_transform(Selecting_the_data)
print("Normalized Dataset In The Form Of Array :- \n")
print(normalized_data)
print("\n")
print("Normalized data in the form of Data Set : ")
normalized_data = pd.DataFrame(normalized_data , columns =
Selecting_the_data.columns)
print(normalized_data)
```

2A. Apply forward selection, considering PMV as response variable and Multilinear regression as machine learning model. Create a table, that mentions dimensions, and performance achieved. Which is the optimal feature set, and why.

Feature Vector	Performance achieved
[1,2,3,5,6,7,8,9,10,11,12,13]	Accuracy of the model is 94.92

2B. Apply backward selection, considering PMV as response variable and Multilinear regression as machine learning model. Create a table, that mentions dimensions, and performance achieved. Which is the optimal feature set, and why.

Feature Vector	Performance achieved
[1,2,3,5,6,7,8,9,10,11,12,13]	Accuracy of the model is 94.92

3A. Apply forward selection, considering AMV as response variable and Logistic regression as machine learning model. Create a table, that mentions dimensions, and performance achieved. Which is the optimal feature set, and why.

Feature Vector	Performance achieved
[3,4,10,11]	Accuracy of the model is 100

3B. Apply backward selection, considering AMV as response variable and Logistic regression as machine learning model. Create a table, that mentions dimensions, and performance achieved. Which is the optimal feature set, and why.

Feature Vector	Performance achieved
[0,1,3,5,6,9,10,11,13]	Accuracy of the model is 100

4. Using the optimal feature vector that you've figured out from your analysis above, apply 3-fold cross validation for both regression and classification problems (PMV and AMV respectively). Write down the optimal parameters values for each of the model. Further, plot confusion matrix for the classification part.

#3-Fold-Cross-Validaton For Regression Problem i.e PMV

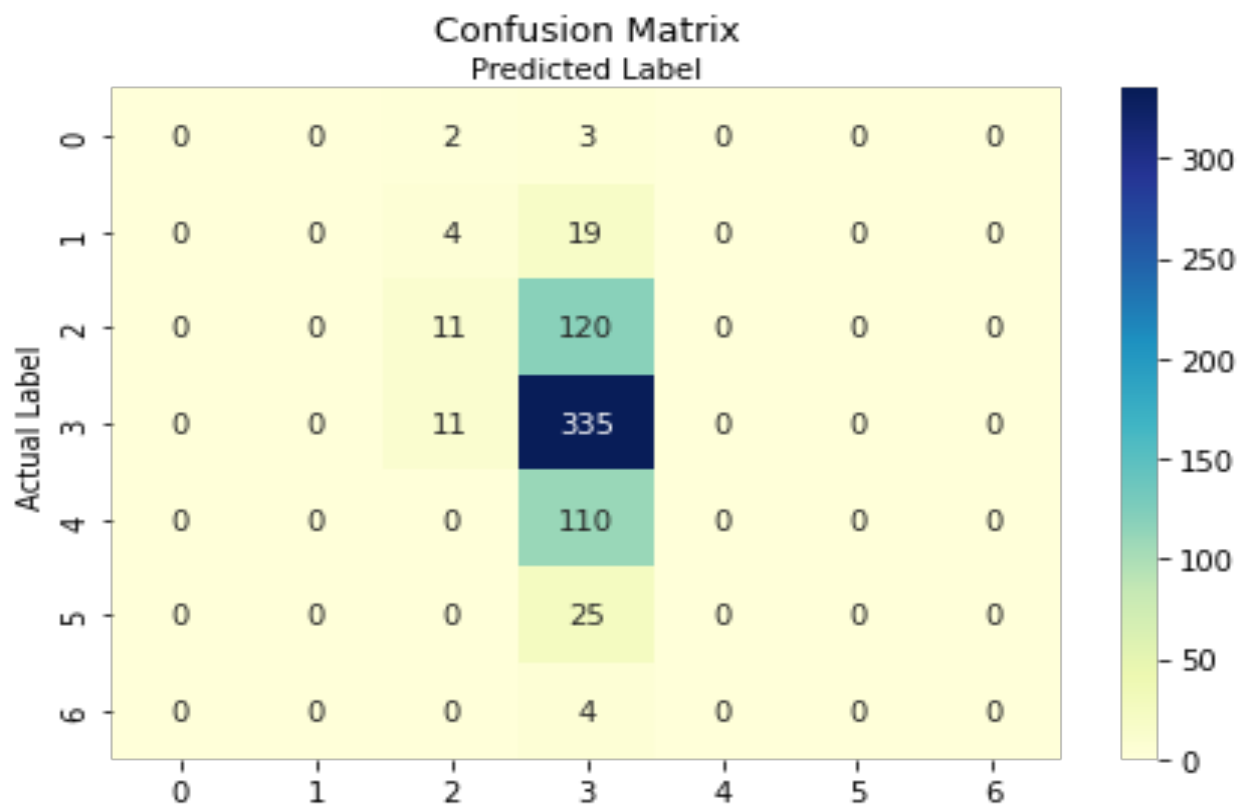
[94.93012999 94.11054873 94.02743803]

0.9713703666411915

#3-Fold-Cross-Validaton For Regression Problem i.e AMV

[7.09891347 4.56131035 6.55049852]

0.24637858635538992



Accuracy : 53.72670807453416

Precision : 13.381261595547308