Introduction to Data Science

Course Project

Report Document

<Anna Rai>

<21L-5696>

<Section 3B>

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 l192347@lhr.nu.edu.pk

Section B, D - Hira Ijaz l192377@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.

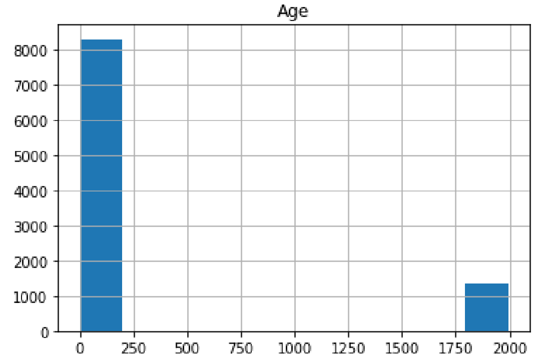
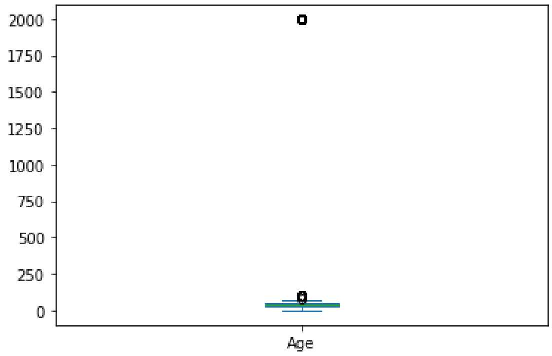
Column number	Feature Name	Feature Description
3	Age	Age
22	Clo	Clothing insulation
19	Met	Met Rate
26	Dewpt	Dewpt
27	PlaneRadTemp	plane radiant temperature
37	Ta	Average air temperature
38	Tmrt	Average mean radiant temperature
40	Vel	Air Velocity
42	AirTurb	Air Turbulance
43	Pa	Vapor Pressure
44	Rh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

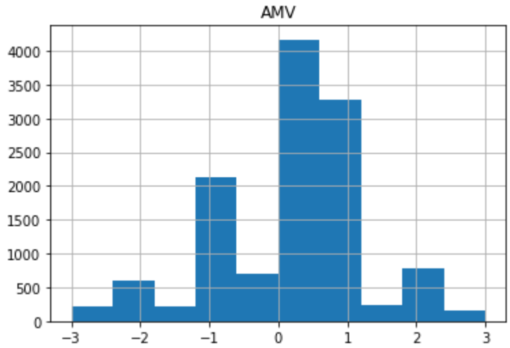
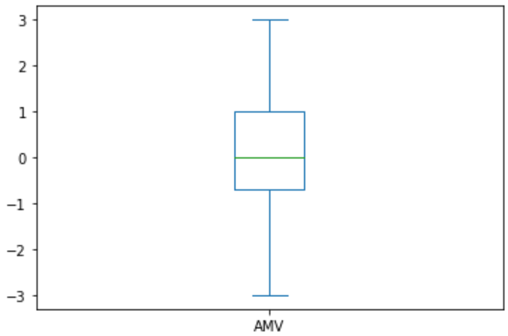
Part A. Preprocessing

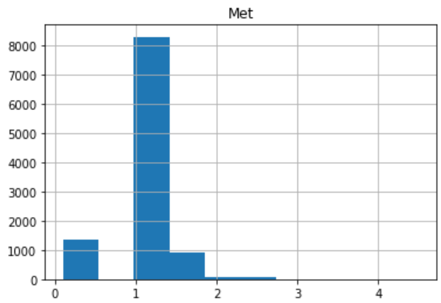
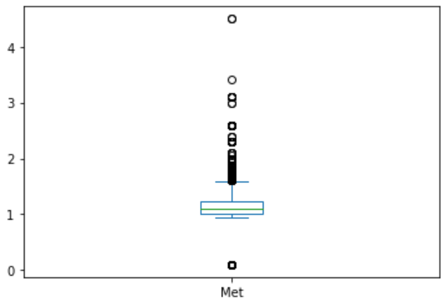
1. In this step, you are required to apply the preprocessing steps that you've covered in the course. Specifically, for each of the input dimension, fill in the following (add rows and complete the table for all input dimensions).

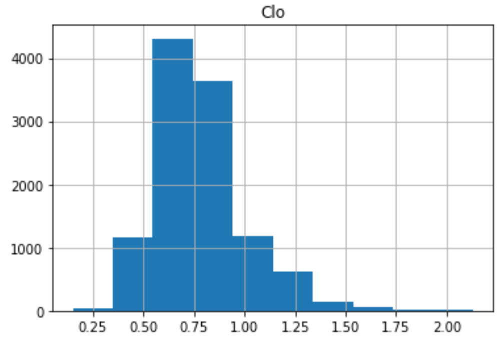
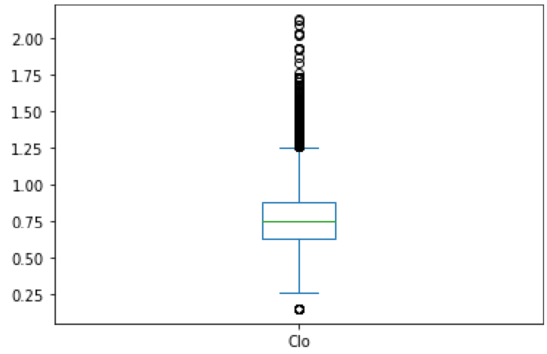
Dim Name	Data Type	Total Instances	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	STD
Age	Float64	9650	2916	1359	0.000	1996.00000	24.0	308.637202	35.000000	462556.556104	680.115105
AMV	Float64	12511	55	0	-3.000	3.00000	0.0	0.100735	0.000000	1.214621	1.102099
Met	Float64	10679	1887	1732	0.100	4.50000	1.0	1.066003	1.100000	0.184022	0.428978
Clo	Float64	11160	1406	373	0.150	2.13000	0.77	0.778492	0.751700	0.049281	0.221992
Dewpt	Float64	9014	3552	0	-1.953	26.89675	17.4	13.621447	14.100000	34.845928	5.903044
PlaneRadTemp	Float64	5544	7022	452	-7.420	11.70000	0.3	0.217785	0.200000	1.084022	1.041164
Ta	Float64	12546	20	540	15.960	31.00000	23.2	23.178861	23.136667	2.054606	1.433390
Tmrt	Float64	8865	3701	344	16.610	37.44500	22.5	23.450261	23.358438	2.258867	1.502953
Vel	Float64	8866	3700	309	0.000	1.88000	0.1	0.112439	0.100000	0.006248	0.079041
AirTurb	Float64	6965	5601	2	0.000	102.45000	0.5	18.265870	0.500000	627.057129	25.041109
Pa	Float64	7910	4656	1352	0.000	27.70000	2.1	5.123996	1.550667	66.522562	8.156136
Rh	Float64	12531	35	0	7.400	79.30000	64.0	42.529203	43.280000	226.835983	15.061075
PMV	Float64	11870	696	259	-4.170	2.50000	0.1	-0.073676	-0.030000	0.289461	0.538016
TaOutdoor	Float64	11198	1368	124	-24.900	32.35000	27.55556	17.174585	18.200000	113.743733	10.665071
RhOutdoor	Float64	12547	19	1349	0.000	100.35000	0.0	61.100365	68.795799	610.282477	24.703896

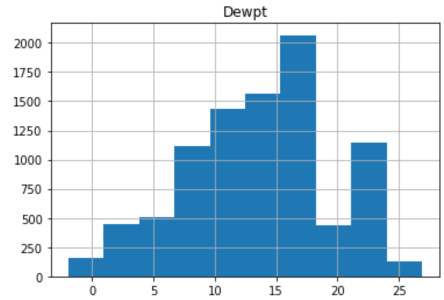
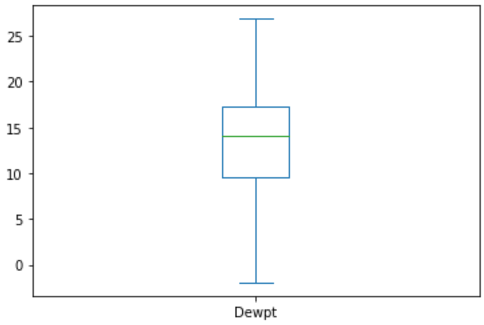
2. For each of the input dimension, plot histogram and comment the type of distribution the dimension exhibits. Further, visualize each dimension using a Box Plot. Specifically, for each of the input dimension, you're required to fill the following table (duplicate it for each of the 15 dimensions).

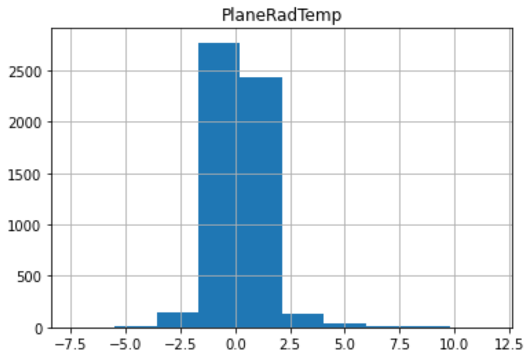
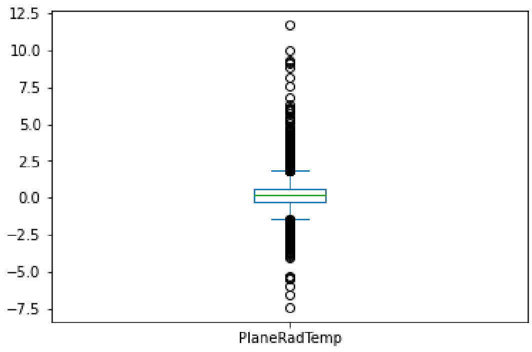
Age	
Histogram	Box Plot
	
Comments: Data is normally distributed	Comments: The outliers are out of range and they are not normal

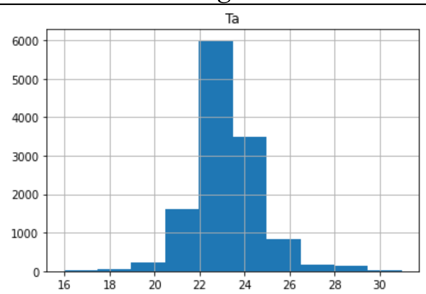
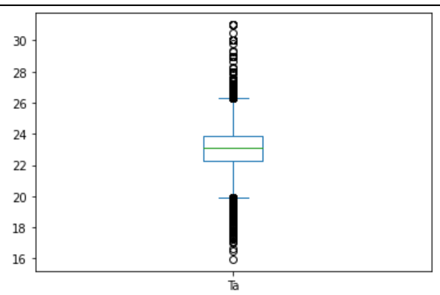
AMV	
Histogram	Box Plot
	
Comments: Data is normally distributed	Comments: Data has no large outliers

Met	
Histogram	Box Plot
	
Comments: Data is positively skewed	Comments: There are large no. of outliers

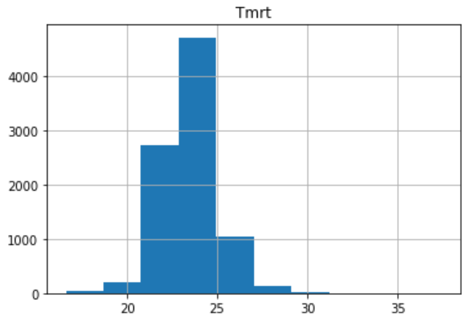
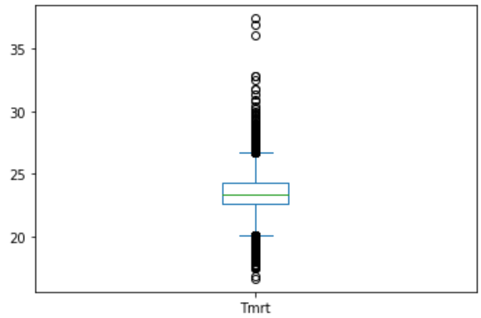
Clo	
Histogram	Box Plot
	
Comments: Data is normally distributed	Comments: There are large no. of outliers

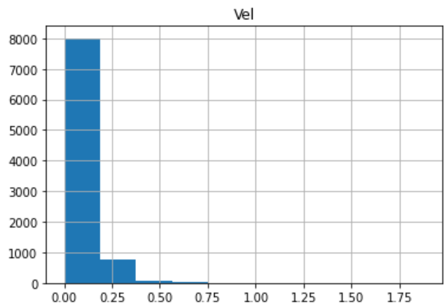
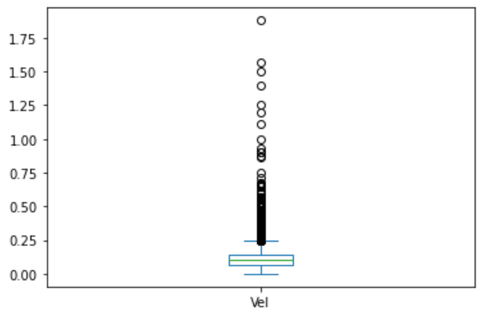
Dewpt	
Histogram	Box Plot
	
Comments: Data is normally distributed	Comments: Data has no large outliers

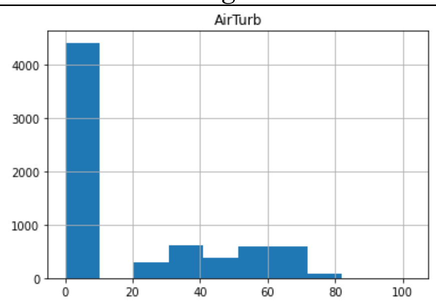
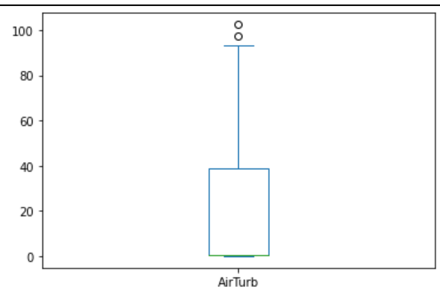
PlaneRadTemp	
Histogram	Box Plot
	
Comments: Data is normally distributed	Comments: There are large no. of outliers

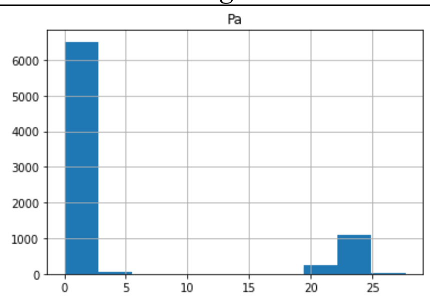
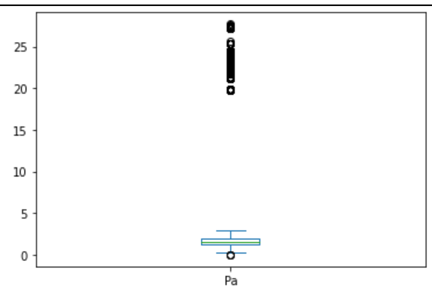
Ta	
Histogram	Box Plot
	

Comments: Data is normally distributed	Comments: There are large no. of outliers

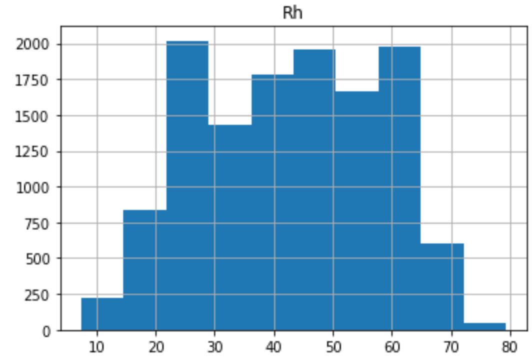
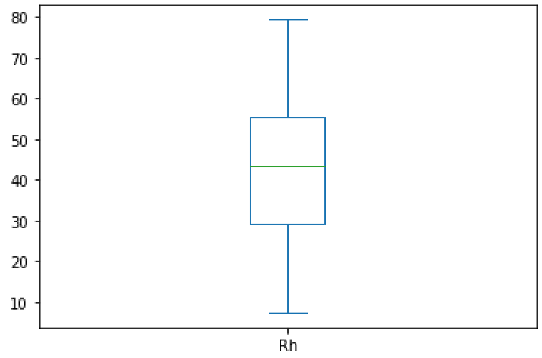
<table>
<tr><th colspan="2">Tmrt</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments: Data is normally distributed</td><td>Comments: There are large no. of outliers</td></tr>
</table>

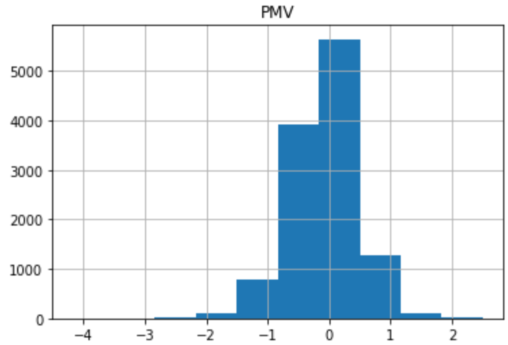
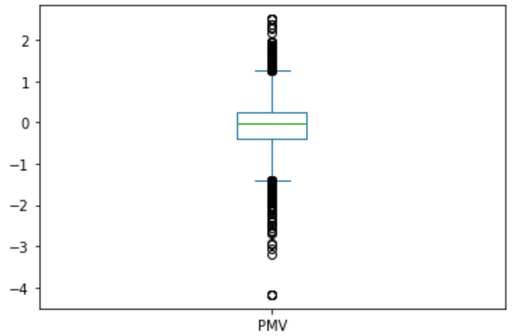
<table>
<tr><th colspan="2">Vel</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments: Data is positively skewed</td><td>Comments: There are large no. of outliers</td></tr>
</table>

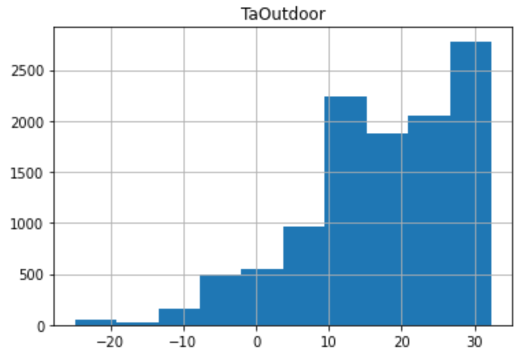
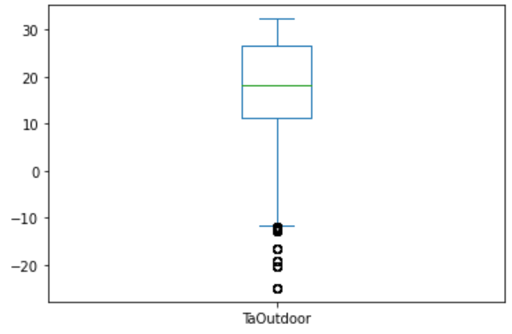
<table>
<tr><th colspan="2">AirTurb</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments: Data is positively skewed</td><td>Comments: Data is positively skewed so it is not ready to use</td></tr>
</table>

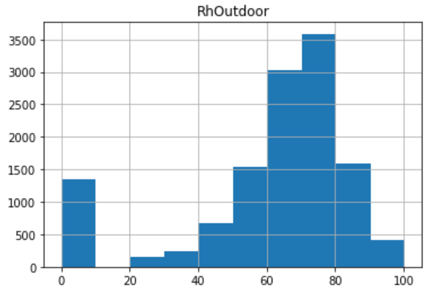
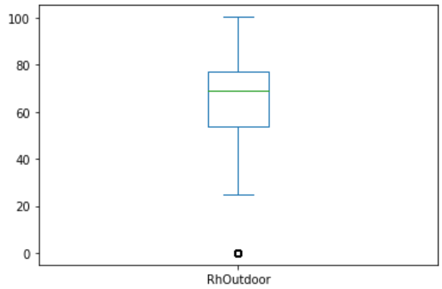
<table>
<tr><th colspan="2">Pa</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
</table>

Comments: Data is normally distributed	Comments: There are large no. of outliers

Rh	
Histogram	Box Plot
	
Comments: Data is normally distributed	Comments: Data has no outliers

PMV	
Histogram	Box Plot
	
Comments: Data is normally distributed	Comments: There are large no. of outliers

TaOutdoor	
Histogram	Box Plot
	
Comments: Data is negatively skewed	Comments: There are few no. of outliers

RhOutdoor	
Histogram	Box Plot
	
Comments: Data is normally distributed	Comments: There are few no. of outliers

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 Values	Filled using OR Dropped	Reason for selecting a certain approach
Age	2916	Median	Median is 35 which is normal while mean is 300+
AMV	55	Mean	Because mean is close to mode
Met	1887	Mean	There is large no. of outliers
Clo	1406	Mean	Because mean is close to mode
Dewpt	3552	Mean	Because mean is close to mode
PlaneRadTemp	7022	Drop	Large no. null values
Ta	20	Mean	Because mean is close to mode
Tmrt	3701	Mean	Because mean is close to mode
Vel	3700	Mean	Because mean is close to mode
AirTurb	5601	Drop	Large no. null values
Pa	4656	Drop	Large no. null values
Rh	35	Mean	Because mean is close to mode
PMV	696	Mean	Because mean is close to mode
TaOutdoor	1368	Mean	Because mean is close to mode
RhOutdoor	19	Mean	There is large no. of outliers

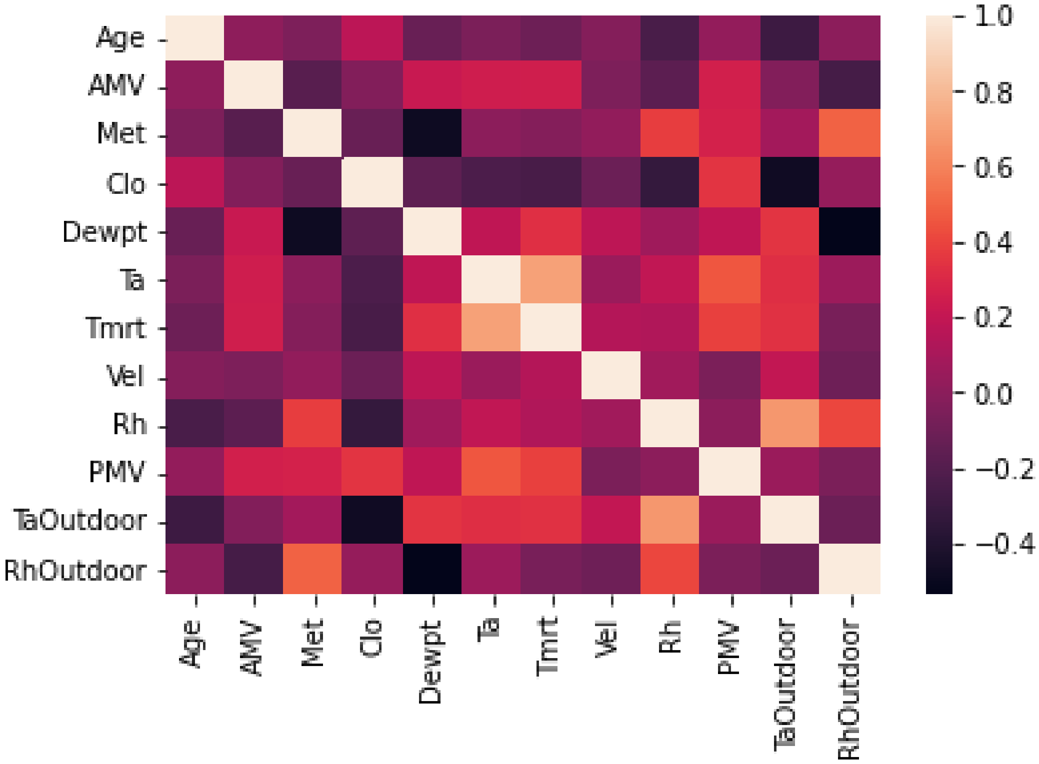
4. For each of the dimension, find out the outliers (noisy data) and handle these appropriately.

Dim Name	Number of Outliers	Smooth using/ Dropped	Reason for selecting a certain approach
Age	1359	Median	There is large no. of outliers
AMV	0	None	No outliers
Met	1732	IQR	There is large no. of outliers
Clo	373	IQR	There is large no. of outliers
Dewpt	0	None	No outliers
PlaneRadTemp	452	IQR	There is large no. of outliers
Ta	540	IQR	There is large no. of outliers
Tmrt	344	IQR	There is large no. of outliers
Vel	309	IQR	There is large no. of outliers
AirTurb	2	IQR	There is few no. of outliers
Pa	1352	IQR	There is large no. of outliers
Rh	0	None	No outliers
PMV	259	IQR	There is large no. of outliers
TaOutdoor	124	IQR	There is large no. of outliers
RhOutdoor	1349	IQR	There is large no. 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	46.584339	No
AMV	1.209305	Yes
Met	0.041571	Yes
Clo	0.033370	Yes
Dewpt	22.698750	No
Ta	1.667084	Yes
Tmrt	0.873679	Yes
Vel	0.001498	Yes
Rh	226.204128	No
PMV	0.239404	Yes
TaOutdoor	94.824547	No
RhOutdoor	394.684046	No

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



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)

For PMV:

Most informative dimensions:

Age, Vel, Dewpt, Rh, TaOutdoor, RhOutdoor

Least informative dimensions:

AMV, Met, Clo, Dewpt, PlaneRadtemp, Ta, Tmrt

For AMV:

Most informative dimensions:

Age, Met, Clo, Vel, AirTurb, Pa, Rh, TaOutdoor, RhOutdoor

Least informative dimensions:

Dewpt, PlaneRadTemp, Ta, Tmrt, PMV

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

Dim name	Entropy	Info Gain	Reason
Age	3.421	0.068	
AMV	3.507		
Met	4.836	0.299	
Clo	7.101	0.359	
Dewpt	7.451	0.738	
Ta	8.095	0.577	
Tmrt	7.119	0.552	
Vel	4.98	0.255	
Rh	10.879	1.186	
PMV	7.409	0.259	
TaOutdoor	7.579	0.418	
RhOutdoor	7.204	0.389	

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.

```
[[ 0.32956402  0.01883223  0.01412417 ...  0.00320148  0.10875613
   0.74387306]
 [ 0.3294843   0.01882767  0.01412076 ...  0.00480106  0.10872982
   0.74369314]
 [ 0.32855702  0.00938734  0.01408102 ...  0.00750987  0.10842382
   0.74160013]
 ...
 [ 0.55715325  0.01591866  0.01114306 ... -0.00117282  0.27339645
   0.3024521 ]
 [ 0.56068871  0.01601968  0.01121377 ... -0.00118026  0.27513131
   0.30437134]
 [ 0.56593498  0.01616957  0.0113187  ... -0.0011913   0.27770566
   0.30721929]]
```

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
Age, Clo, Dewpt, Ta, Tmrt	100%

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
Age, Clo, Dewpt, Ta, Tmrt	100%

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
Clo, Met, Ta, Tmrt, Vel	53.50%

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
Clo, Met, Ta, Tmrt, Vel	53.38%

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.

For Regression:

0.486

array([0.48644885, 0.44547353, 0.45150303, 0.46566876, 0.48301733])

For Classification:

0.536

array([0.51019393, 0.53654898, 0.50049751, 0.52139303, 0.51840796])

Confusion Matrix:

```
array([[   0,    0,    0,   19,    1,    0,    0],
       [   0,    0,    0,  121,    5,    0,    0],
       [   0,    0,    0,  439,   10,    0,    0],
       [   0,    0,    0, 1051,   27,    0,    0],
       [   0,    0,    0,  370,  291,    0,    0],
       [   0,    0,    0,  134,   25,    0,    0],
       [   0,    0,    0,   19,    2,    0,    0]])
```