Introduction to Data Science

Course Project

Report Document

<Abdullah Umar>

<21L-5604>

<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.

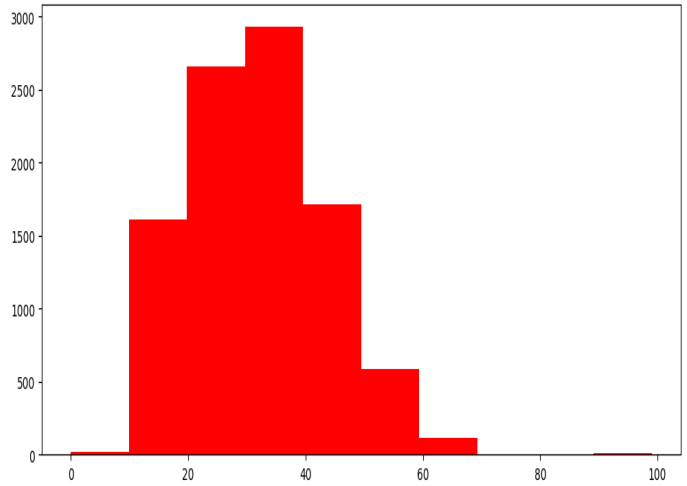
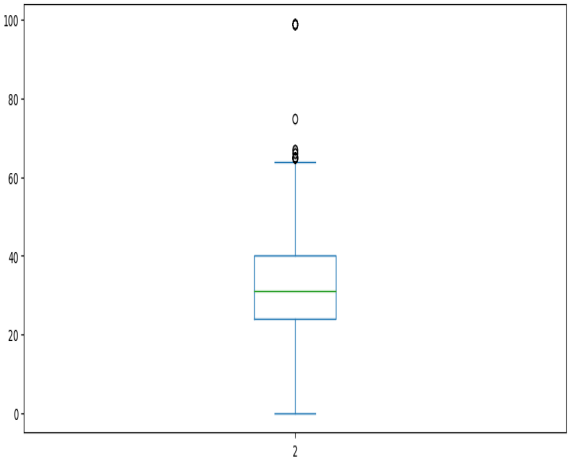
Column number	Feature Name	Feature Description
3	Age	Age
22	Clo	Clothing insulation
19	Met	Met Rate
26	Dewpt	Dewpt
27	PlaneRadTemp	plane radiant temperature
37	Ta	Average air temperature
38	Tmrt	Average mean radiant temperature
40	Vel	Air Velocity
42	AirTurb	Air Turbulance
43	Pa	Vapor Pressure
44	Rh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

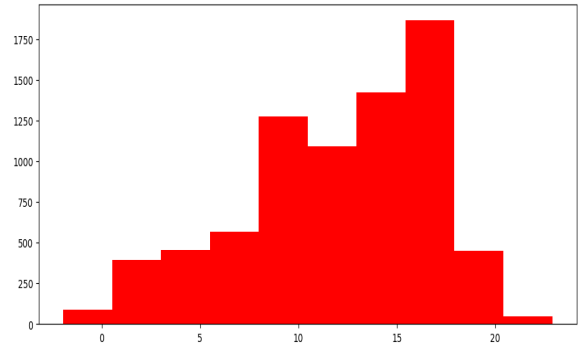
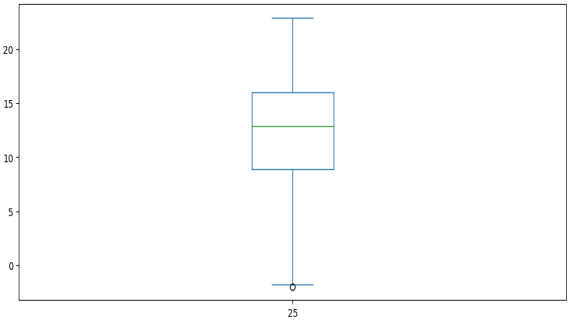
Part A. Preprocessing

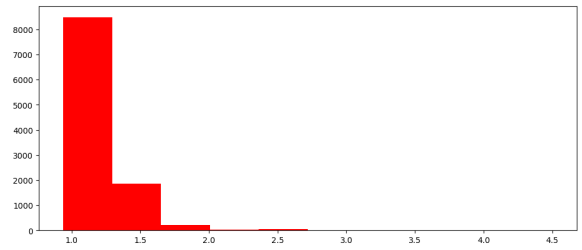
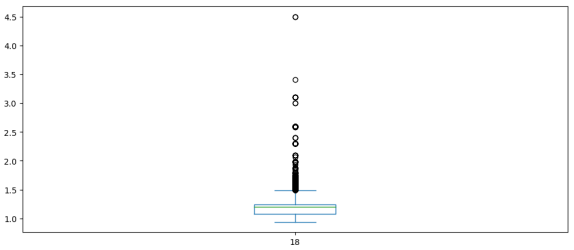
1. In this step, you are required to apply the preprocessing steps that you've covered in the course. Specifically, for each of the input dimensions, fill in the following (add rows and complete the table for all input dimensions).

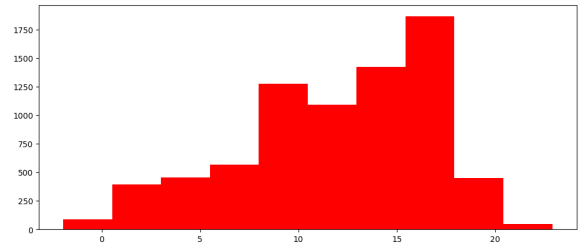
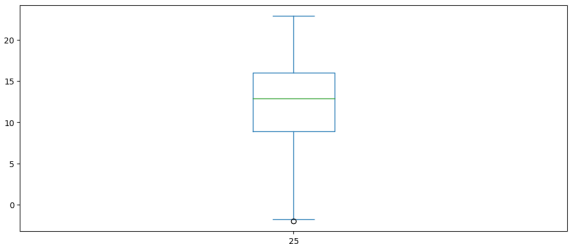
Dim Name	Data Type	Total Instances (without nulls)	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	STD
Age	Float64	9649	2917	37	0.0	99.0	24.0	31.98	31.0	133.48	11.55
Clo	Float64	12509	57	356	0.15	2.13	0.77	0.75	0.72	0.05	0.22
Met	Float64	**10679**	1887	838	0.93	4.5	1.2	1.2	1.2	0.04	0.22
Dewpt	Float64	7665	4901	1	-1.95	22.9	17.4	12.01	12.87	23.42	4.84
PlaneRadTemp	Float64	5544	7022	452	-7.42	11.7	0.3	0.21	0.2	1.08	1.04
Ta	Float64	11197	1369	425	15.96	31.0	23.2	23.20	23.13	2.15	1.46
Tmrt	Float64	8865	3701	344	16.61	37.44	22.5	23.45	23.35	2.25	1.50
Vel	Float64	8866	3700	309	0.0	1.88	0.1	0.11	0.1	0.006	0.079
AirTurb	Float64	5616	6950	1216	0.0	102.45	0.5	8.15	0.4	235.65	15.35
Pa	Float64	6561	6005	158	0.0	3.0	2.1	1.43	1.45	0.19	0.44
Rh	Float64	12531	35	0	7.4	79.3	64.0	46.5	47.88	209.03	14.45
TaOutdoor	Float64	12547	19	147	-24.9	32.35	27.55	18.27	20.7	112.63	10.61
RhOutdoor	Float64	12547	19	162	24.97	100.35	81.55	68.48	69.5	170.13	13.04
AMV	Float64	12511	55	0	-3.0	3.0	0.0	-0.11	0.0	1.30	1.14
PMV	Float64	12523	43	231	-4.17	2.5	-0.01	-0.13	-0.12	0.31	0.5

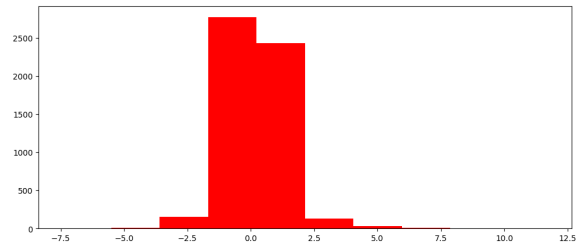
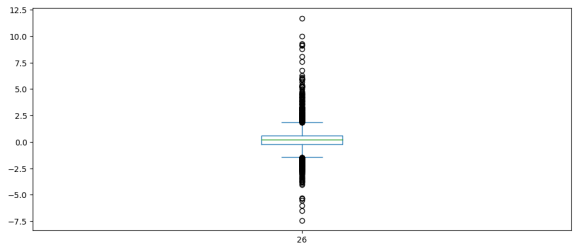
2. For each of the input dimensions, plot a histogram and comment on the type of distribution the dimension exhibits. Further, visualize each dimension using a Box Plot. Specifically, for each of the input dimensions, you're required to fill the following table (duplicate it for each of the 15 dimensions).

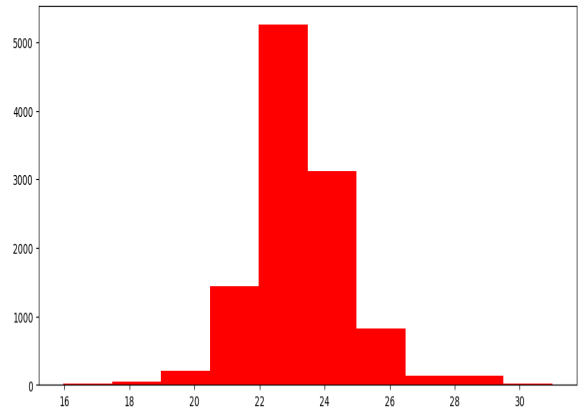
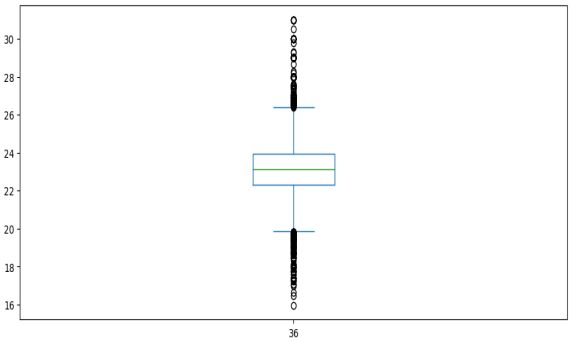
<table>
<tr><th colspan="2">Age</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>Comments:
The data is right skewed</td><td>Comments:
there are so 37 outliers</td></tr>
</table>

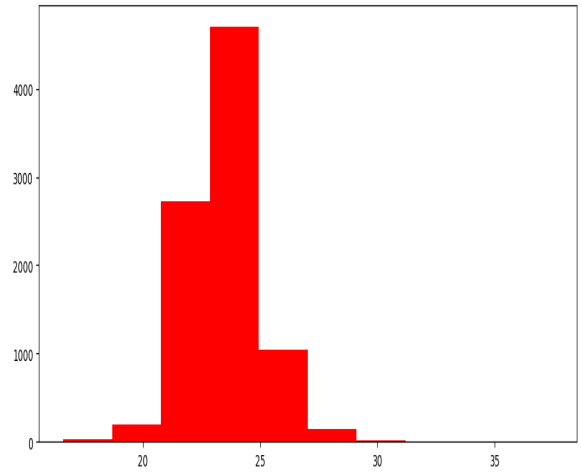
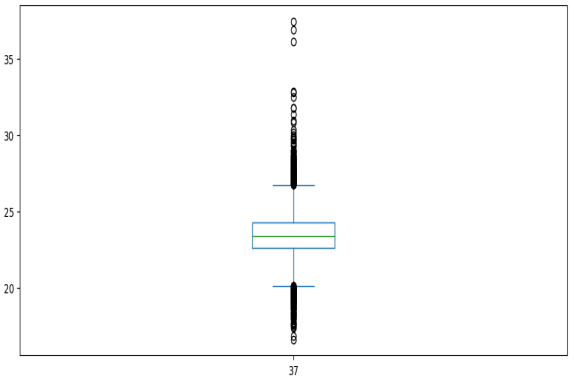
Clo	
Histogram	Box Plot
	
Comments: The data is LEFT skew	Comments: The outliers are 56

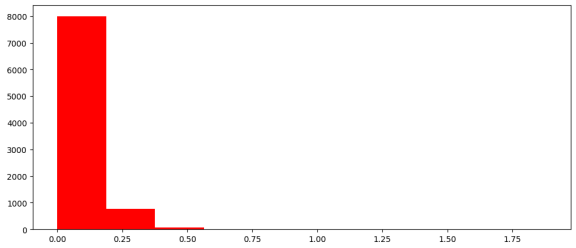
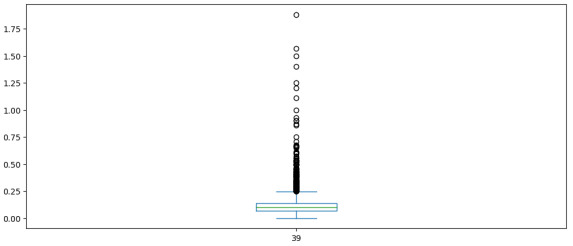
Met	
Histogram	Box Plot
	
Comments: The data is right skewed	Comments: the outliers are 838

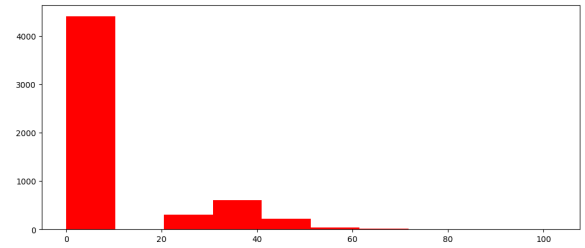
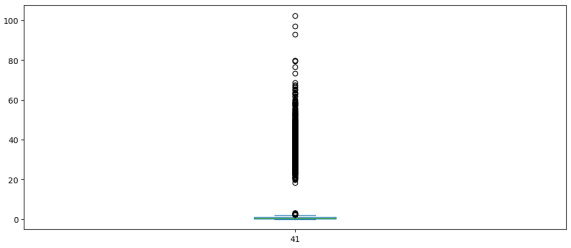
Dewpt	
Histogram	Box Plot
	
Comments: the data is RIGHT skewed	Comments: the outliers are 1

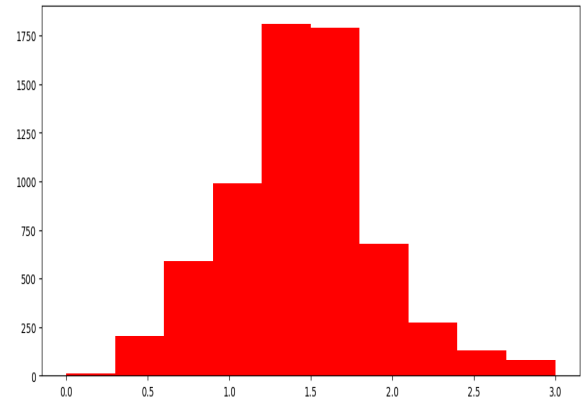
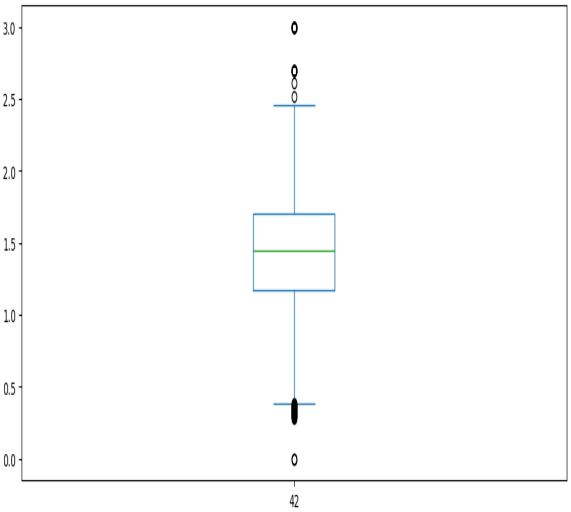
PlaneRadTemp	
Histogram	Box Plot
	
Comments: the data is left skewed	Comments: the outliers are 452

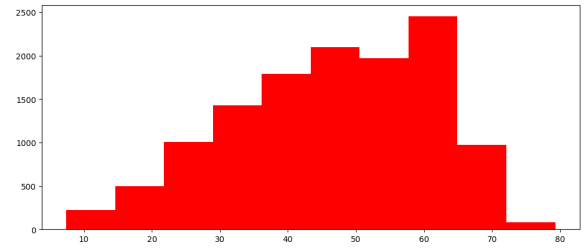
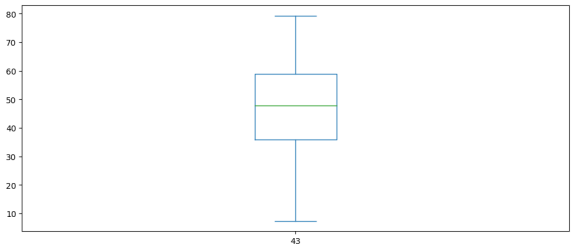
Ta	
Histogram	Box Plot
	
Comments: the data is right skewed	Comments: the outliers are 425

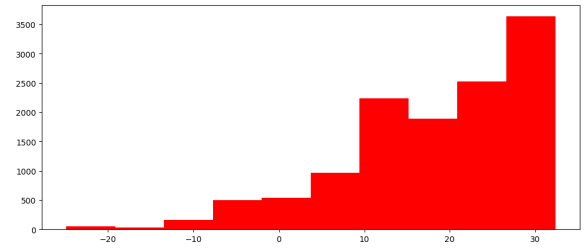
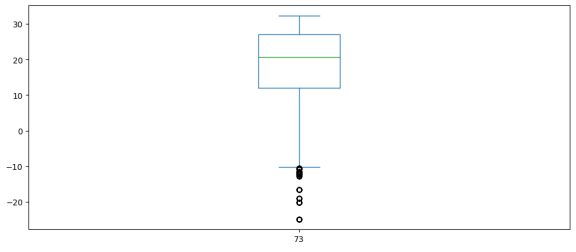
Tmrt	
Histogram	Box Plot
	
Comments: the data is right skewed	Comments: the outliers are 344

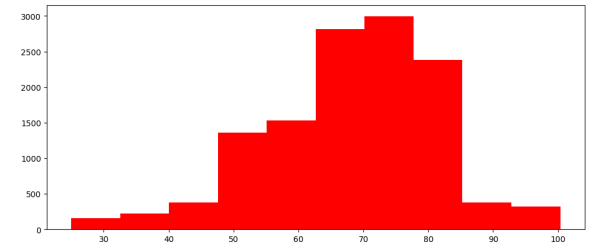
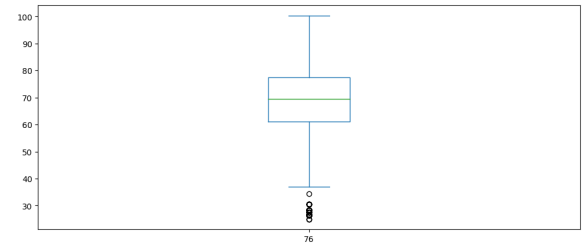
Vel	
Histogram	Box Plot
	
Comments: the data is left skewed	Comments: the outliers are 309

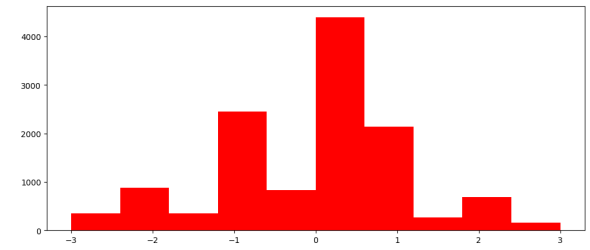
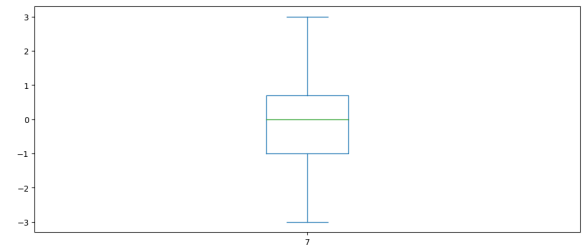
AirTurb	
Histogram	Box Plot
	
Comments: the data is right skewed	Comments: the outliers are 1216

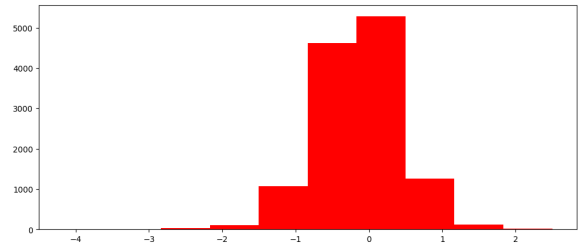
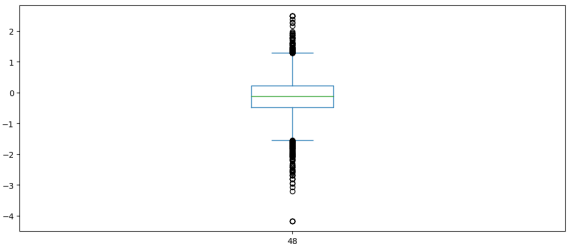
Pa	
Histogram	Box Plot
	
Comments: The data is normalized	Comments: the outliers are 158

Rh	
Histogram	Box Plot
	
Comments: the data is left skewed	Comments: the outliers are 0

TaOutdoor	
Histogram	Box Plot
	
Comments: the data is right skewed	Comments: the outliers are 147

RhOutdoor	
Histogram	Box Plot
	
Comments: the data is normilized	Comments: the outliers are 162

AMV	
Histogram	Box Plot
	
Comments: the table is normal	Comments: the outliers are 0

PMV	
Histogram	Box Plot
	
Comments: the data is simmitrical	Comments: the outliers are 231

3. Find the missing values in each of the dimensions (do this for both input and output dimensions), and fill these using an "appropriate" methodology that we've discussed in 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	2917	Filled using mean	the percentages of outliers to total numbers of entries is less than 2 percent, I assume the threshold 2 percent
Clo	57	Filled using median	the percentages of outliers to total numbers of entries is **greater** than 2 percent, I assume the threshold 2 percent
Met	1887	Filled using median	the percentages of outliers to total numbers of entries is **greater** than 2 percent, I assume the threshold 2 percent
Dewpt	4901	Filled using mean	the percentages of outliers to total numbers of entries is less than 2 percent, I assume the threshold 2 percent
PlaneRadTemp	7022	Filled using median	the percentages of outliers to total numbers of entries is **greater** than 2 percent, I assume the threshold 2 percent
Ta	1369	Filled using median	the percentages of outliers to total numbers of entries is **greater** than 2 percent, I assume the threshold 2 percent

Tmrt 3701	Filled using median	the percentages of outliers to total numbers of entries is **greater** than 2 percent, I assume the threshold 2 percent
Vel 3700	Filled using median	the percentages of outliers to total numbers of entries is **greater** than 2 percent, I assume the threshold 2 percent
AirTurb 6950	Filled using median	the percentages of outliers to total numbers of entries is **greater** than 2 percent, I assume the threshold 2 percent
Pa 6005	Filled using median	the percentages of outliers to total numbers of entries is **greater** than 2 percent, I assume the threshold 2 percent
Rh 35	Filled using mean	the percentages of outliers to total numbers of entries is less than 2 percent, I assume the threshold 2 percent
TaOutdoor 19	Filled using mean	the percentages of outliers to total numbers of entries is less than 2 percent, I assume the threshold 2 percent
RhOutdoor 19	Filled using mean	the percentages of outliers to total numbers of entries is less than 2 percent, I assume the threshold 2 percent
AMV 55	Filled using mean	the percentages of outliers to total numbers of entries

		is less than 2 percent, I assume the threshold 2 percent
PMV 43	Filled using mean	the percentages of outliers to total numbers of entries is less than 2 percent, I assume the threshold 2 percent

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

Dim Name Number of Nulls	Smooth using/ Dropped	Reason for selecting a certain approach
Age 2917	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measure of how to spread out the values
Clo 57	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to spread out the values are.
Met 1887	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures of how

		to spread out the values are.
Dewpt 4901	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to spread out the values are.
PlaneRadTemp 7022	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to spread out the values are.
Ta 1369	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to spread out the values are.
Tmrt 3701	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to spread out the values are.
Vel 3700	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to

		spread out the values are.
AirTurb 6950	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to spread out the values are.
Pa 6005	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to spread out the values are.
Rh 35	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to spread out the values are.
TaOutdoor 19	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to spread out the values are.
RhOutdoor 19	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to

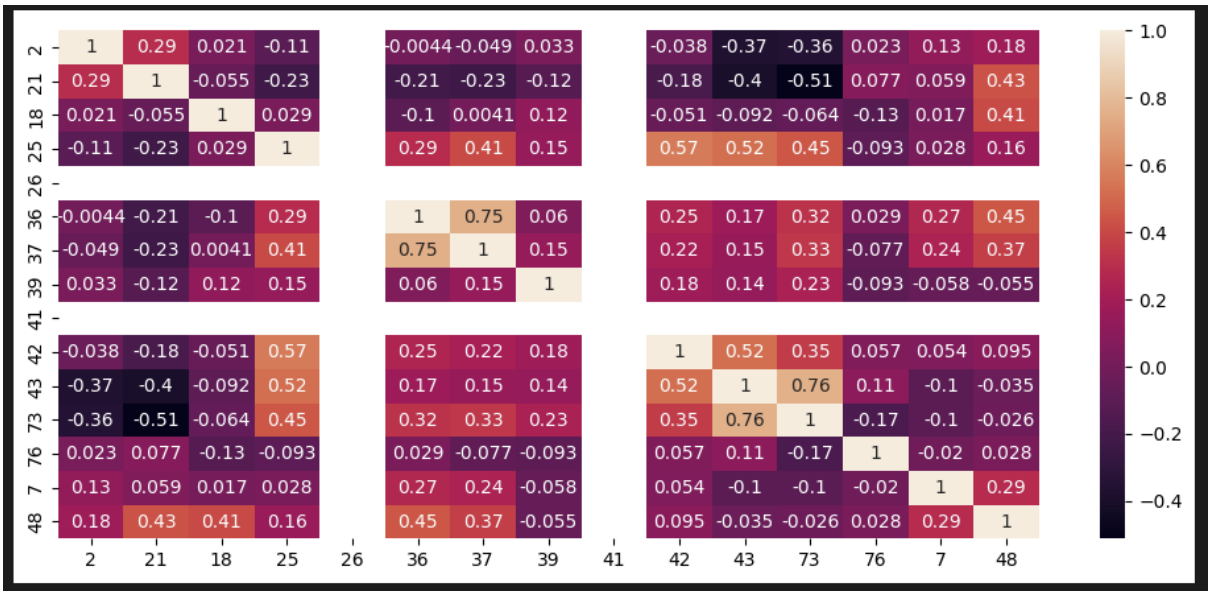
		spread out the values are.
AMV 55	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to spread out the values are.
PMV 43	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to spread out the values are.

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	133.48	I will not apply any filter as the data is so much diverse
Clo	0.05	I will apply the filter as the data has no variance in it so this data will not help us in training
Met	0.04	I will apply the filter as the data has no variance in it so this data will not help us in training
Dewpt	23.42	the data us slightly variant so may be I well apply any filter on it
PlaneRadTemp	1.08	I will apply the filter as the data has no variance in it so this data will not help us in training
Ta	2.15	I will apply the filter as the data has no variance in it so this data will not help us in training
Twrt	2.25	I will apply the filter as the data has no variance in it so this data will not help us in training
Vel	0.006	I will apply the filter as the data has no variance in it so this data will not help us in training
AirTurb	235.65	I will not apply any filter as the data is so much diverse
Pa	0.19	I will apply the filter as the data has no variance in it so this data will not help us in training
Rh	209.03	I will not apply any filter as the data is so much diverse
TaOutdoor	112.63	I will not apply any filter as the data is so much diverse
RhOutdoor	170.13	I will not apply any filter as the data is so much diverse

AMV	1.30	I will apply the filter as the data has no variance in it so this data will not help us in training
PMV	0.31	I will apply the filter as the data has no variance in it so this data will not help us in training

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)

73 and 21 are the **most informative** dimension as their correlation is weakest

43 and 73 are the **least informative** dimension as their correlation is strongest

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

Dim name	Entropy	Info Gain	Reason
Age	4.908375408580655		
Clo	4.908375408580655		
Met	4.908375408580655		
Dewpt	4.908375408580655		
PlaneRadTemp	4.908375408580655		
Ta	4.908375408580655		
Tmrt	4.908375408580655		
Vel	4.908375408580655		
AirTurb	4.908375408580655		
Pa	4.908375408580655		
Rh	4.908375408580655		
TaOutdoor	4.908375408580655		
RhOutdoor	4.908375408580655		
AMV	4.908375408580655		
PMV	4.908375408580655		

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.

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

Feature Vector	Performance achieved
5	0.198862681406491
1, 5	0.48472248715305477
1, 2, 5	0.7351465773947905
1, 2, 5, 10	0.7612276970564111
1, 2, 5, 7, 10	0.7716416877563868
1, 2, 5, 6, 7, 10	0.7751849899603429
0, 1, 2, 5, 6, 7, 10	0.778670715414786
0, 1, 2, 5, 6, 7, 10, 11	0.7811010098399922
0, 1, 2, 5, 6, 7, 10, 11, 12	0.781593593935969
0, 1, 2, 3, 5, 6, 7, 10, 11, 12	0.7817689376430383
0, 1, 2, 3, 5, 6, 7, 9, 10, 11, 12	0.7817698927274701
0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12	0.7817698927274701
0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	0.7817698927274701

The optimal feature set is (0, 1, 2, 5, 6, 7, 10, 11) as avg score is **0.781 as more feature set may have more accuracy but they have so much baggage**

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

Feature Vector	Performance achieved
0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	0.7867319436599411
0, 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12	0.7867319436599411
0, 1, 2, 3, 5, 6, 7, 10, 11, 12	0.7867319436599411
0, 1, 2, 5, 6, 7, 10, 11, 12	0.7867319120933685
0, 1, 2, 5, 6, 7, 10, 11	0.7864796539290779
0, 1, 2, 5, 6, 7, 10	0.7860286937573029
1, 2, 5, 6, 7, 10	0.783729046520725
1, 2, 5, 7, 10	0.7801472917337273
1, 2, 5, 10	0.7766023023627011
1, 2, 5	0.7664463652135877
1, 5	0.7421956337292546
5	0.4946979359140804
36	0.2041950927710311

The optimal feature set is (0, 1, 2, 5, 6, 7, 10) as avg score is 0.786 **as more feature set may have more accuracy but they have so much baggage**

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 (indexes)	Performance achieved
6	37.8%
0,6	39.2%
0,6,11	39.2%
0, 6, 9, 11	39.4%
0, 5, 6, 9, 11	39.5%
0, 5, 6, 8, 9, 11	39.7%
0, 5, 6, 8, 9, 11,12	39.7%
0, 3, 5, 6, 8, 9, 11, 12	0, 3, 5, 6, 8, 9, 10, 11, 12
0, 3, 5, 6, 8, 9, 10, 11, 12	39.8%
0, 1, 3, 5, 6, 8, 9, 10, 11, 12	39.8%
0, 1, 2, 3, 5, 6, 8, 9, 10, 11, 12	39.9%
0, 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12	39.9%
0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	40.0%

The optimal feature set is **(0, 3, 5, 6, 8, 9, 10, 11, 12**) as avg score **39.7%** is `as more feature set may have more accuracy but they have so much baggage`

3B. Apply backward selection, considering AMV as the response variable and Logistic regression as a machine learning model. Create a table, that mentions

the dimensions, and performance achieved. Which is the optimal feature set, and why.

Feature Vector (indexes)	**Performance achieved**
Index: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	**39.7%**
Index: 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12	**39.8%**
Index: 0, 1, 2, 4, 5, 6, 7, 8, 10, 11, 12	**39.9%**
Index: 0, 1, 2, 5, 6, 7, 8, 10, 11, 12	**39.9%**
Index: 0, 1, 2, 5, 6, 7, 8, 11, 12	**40.0%**
Index: 0, 1, 2, 5, 7, 8, 11, 12	**39.9%**
Index: 0, 1, 2, 5, 8, 11, 12	**40.0%**
Index: 0, 1, 2, 5, 11, 12	**0, 1, 2, 5, 11, 12**
Index: 0, 1, 5, 11, 12	**39.8%**
Index: 0, 5, 11, 12	**39.5%**
Index: 0, 5, 11	**39.0%**
Index: 5,11	**38.3%**
Index: 5	**37.9%**

The optimal feature set is (**0, 1, 2, 5, 11, 12**) as avg score **40.0%** is **`as more feature set may have more accuracy but they have so much baggage`**

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 parameter values for each of the models. Further, plot the confusion matrix for the classification part

linear regression 3 fold cross validation optimized is

```
1, 2, 3, 5 accuracy is

'cv_scores': array([0.75058379, 0.74754941, 0.73361997]),

  'avg_score': 0.7439177236228995,
```

logistic regression 3 fold cross validation optimized is

```
{'feature_idx': (4,),

  'cv_scores': array([0.38197553, 0.38436288, 0.38119403]),

  'avg_score': 0.3825108120988611,
```

```
confusion matrix

[[  0   0  16  63   0   0   0]

 [  0   0  16 217   0   0   0]

 [  0   0  28 565   1   0   0]

 [  0   0  32 894  12   0   0]

 [  0   0  11 449  17   0   0]

 [  0   0   3 154   8   0   0]

 [  0   0   1  27   0   0   0]]

accuracy  0.373508353221957
```