

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 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.

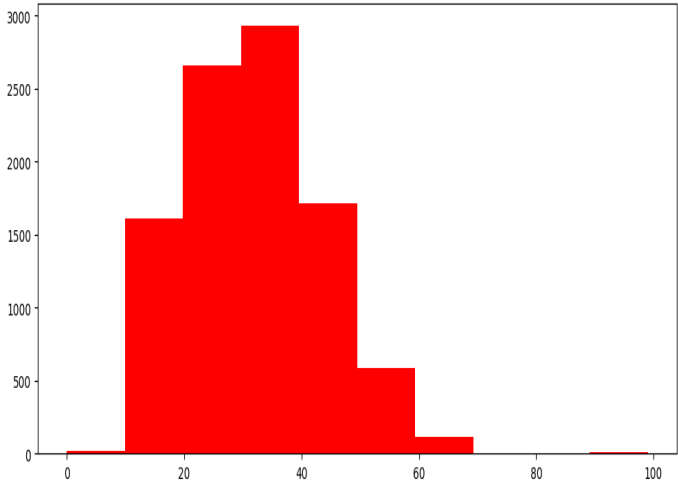
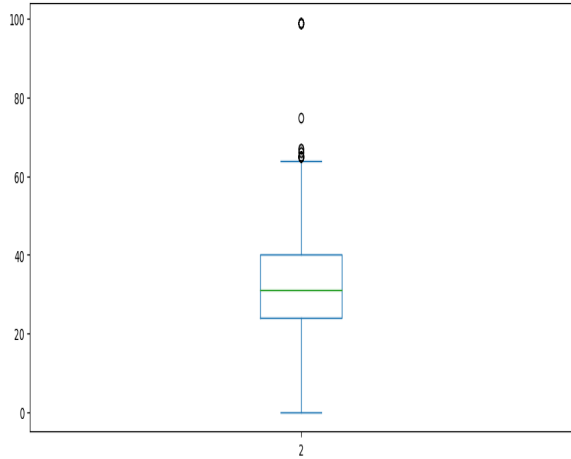
Column number	Feature Name	Feature Description
3	Age	Age
22	Clo	Clothing insulation
19	Met	Met Rate
26	Dewpt	Dewpt
27	PlaneRadTemp	plane radiant temperature
37	Ta	Average air temperature
38	Tmrt	Average mean radiant temperature
40	Vel	Air Velocity
42	AirTurb	Air Turbulance
43	Pa	Vapor Pressure
44	Rh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

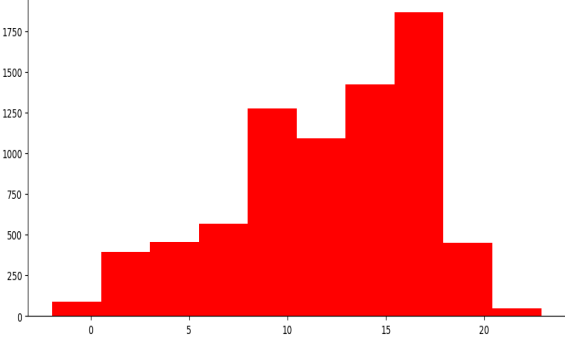
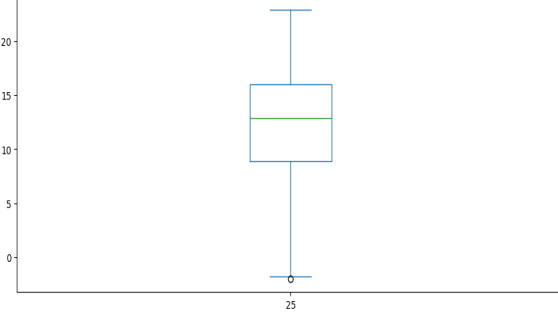
Part A. Preprocessing

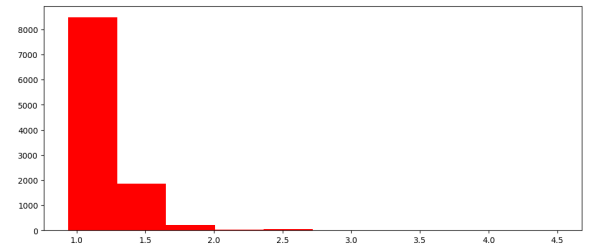
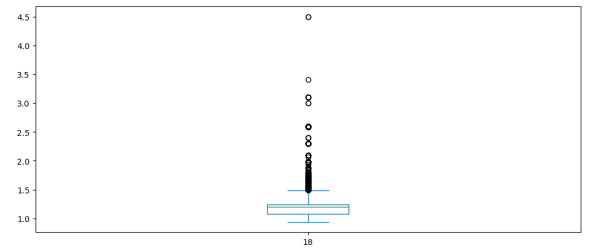
1. In this step, you are required to apply the preprocessing steps that you've covered in the course. Specifically, for each of the input dimensions, fill in the following (add rows and complete the table for all input dimensions).

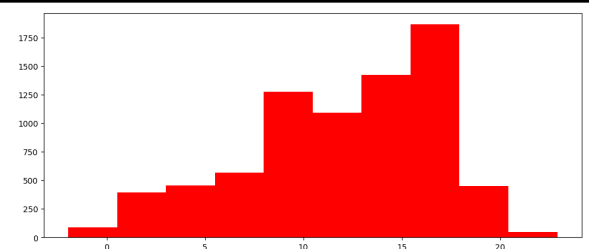
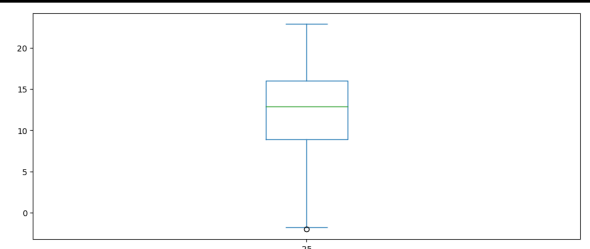
Dim Name	Data Type	Total Instances (without nulls)	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	STD
Age	Float64	9649	2917	37	0.0	99.0	24.0	31.98	31.0	133.48	11.55
Clo	Float64	12509	57	356	0.15	2.13	0.77	0.75	0.72	0.05	0.22
Met	Float64	10679	1887	838	0.93	4.5	1.2	1.2	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

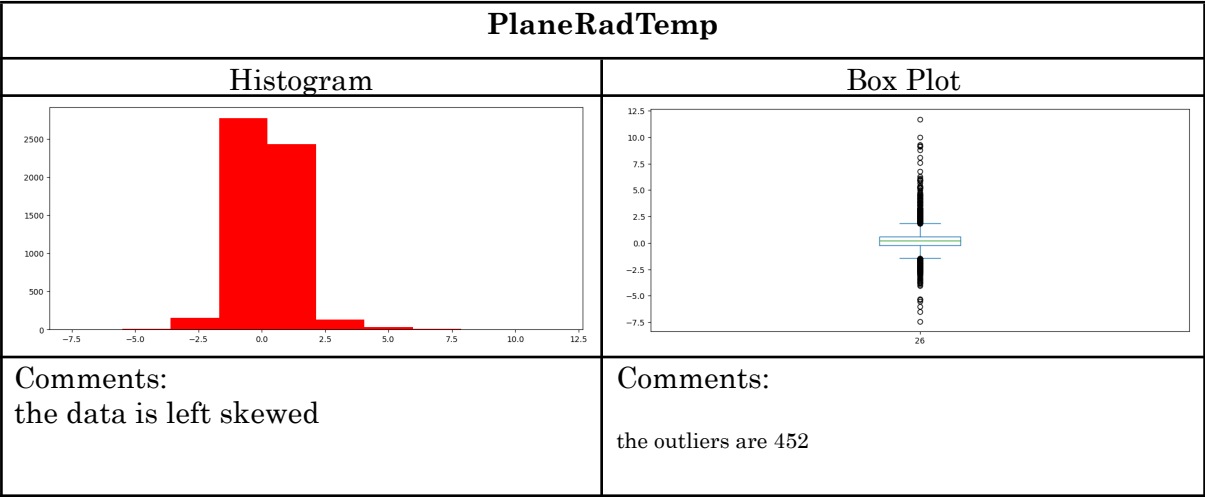
2. For each of the input dimensions, plot a histogram and comment on the type of distribution the dimension exhibits. Further, visualize each dimension using a Box Plot. Specifically, for each of the input dimensions, you're required to fill the following table (duplicate it for each of the 15 dimensions).

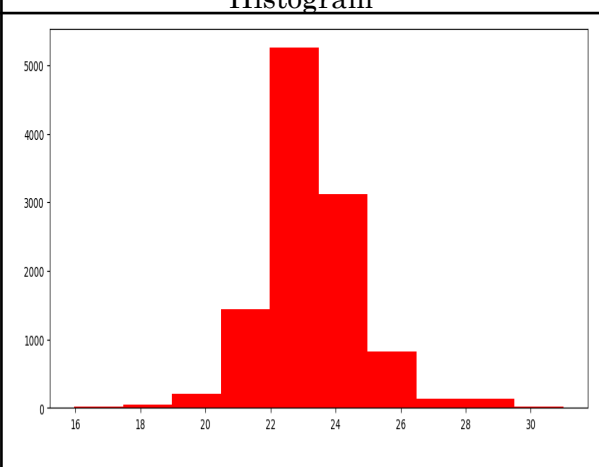
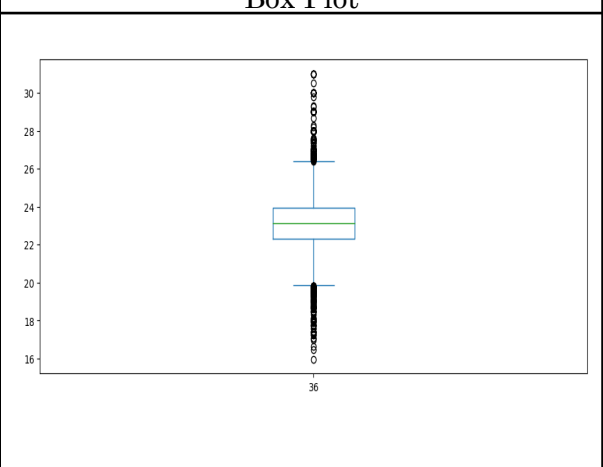
Age	
Histogram	Box Plot
 <p>The histogram displays the frequency of ages. The x-axis represents age from 0 to 100, and the y-axis represents frequency from 0 to 3000. The distribution is unimodal and right-skewed, with the highest frequency occurring in the 35-40 age range.</p>	 <p>The box plot shows the distribution of ages. The median is approximately 32, the interquartile range (IQR) is from about 25 to 40, and the whiskers extend from 0 to 65. There are numerous outliers, with the highest value reaching 100.</p>
Comments: The data is right skewed	Comments: there are so 37 outliers

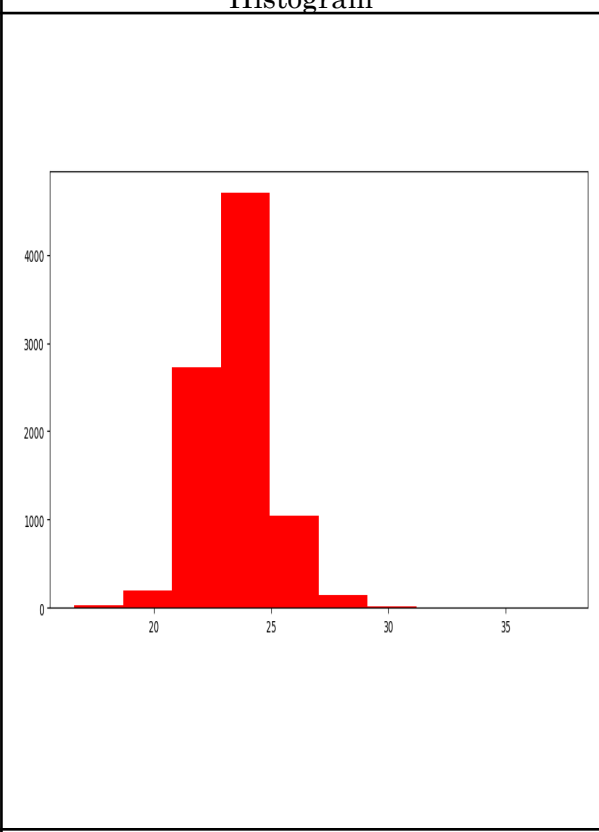
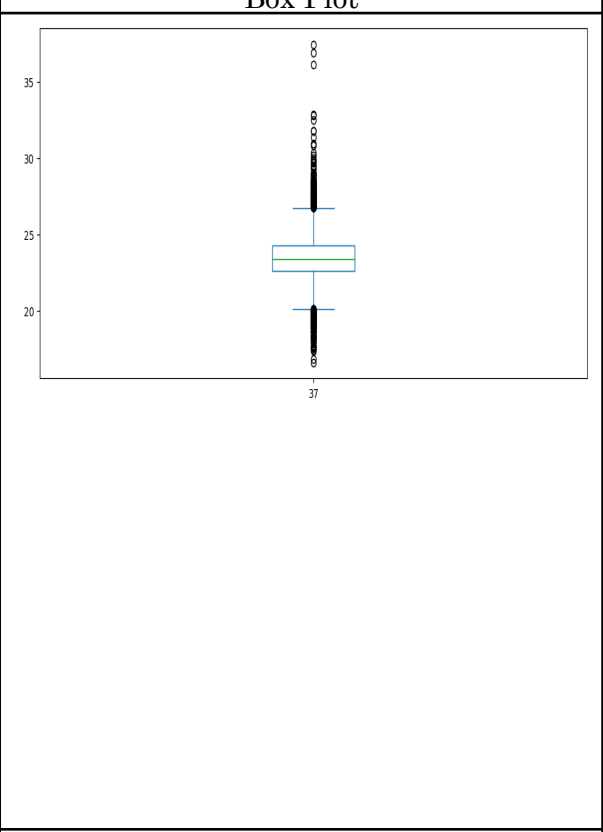
Clo																																																																	
Histogram	Box Plot																																																																
 <table border="1"><caption>Histogram Data</caption><thead><tr><th>Bin Range</th><th>Frequency</th></tr></thead><tbody><tr><td>0-1</td><td>100</td></tr><tr><td>1-2</td><td>400</td></tr><tr><td>2-3</td><td>450</td></tr><tr><td>3-4</td><td>550</td></tr><tr><td>4-5</td><td>600</td></tr><tr><td>5-6</td><td>750</td></tr><tr><td>6-7</td><td>1250</td></tr><tr><td>7-8</td><td>1100</td></tr><tr><td>8-9</td><td>1100</td></tr><tr><td>9-10</td><td>1400</td></tr><tr><td>10-11</td><td>1400</td></tr><tr><td>11-12</td><td>1800</td></tr><tr><td>12-13</td><td>1800</td></tr><tr><td>13-14</td><td>1800</td></tr><tr><td>14-15</td><td>1800</td></tr><tr><td>15-16</td><td>1800</td></tr><tr><td>16-17</td><td>1800</td></tr><tr><td>17-18</td><td>1800</td></tr><tr><td>18-19</td><td>450</td></tr><tr><td>19-20</td><td>450</td></tr><tr><td>20-21</td><td>50</td></tr><tr><td>21-22</td><td>50</td></tr><tr><td>22-23</td><td>50</td></tr><tr><td>23-24</td><td>50</td></tr><tr><td>24-25</td><td>50</td></tr></tbody></table>	Bin Range	Frequency	0-1	100	1-2	400	2-3	450	3-4	550	4-5	600	5-6	750	6-7	1250	7-8	1100	8-9	1100	9-10	1400	10-11	1400	11-12	1800	12-13	1800	13-14	1800	14-15	1800	15-16	1800	16-17	1800	17-18	1800	18-19	450	19-20	450	20-21	50	21-22	50	22-23	50	23-24	50	24-25	50	 <table border="1"><caption>Box Plot Statistics</caption><thead><tr><th>Statistic</th><th>Value (approx.)</th></tr></thead><tbody><tr><td>Minimum</td><td>0</td></tr><tr><td>First Quartile (Q1)</td><td>9</td></tr><tr><td>Median</td><td>13</td></tr><tr><td>Third Quartile (Q3)</td><td>16</td></tr><tr><td>Maximum</td><td>22</td></tr></tbody></table>	Statistic	Value (approx.)	Minimum	0	First Quartile (Q1)	9	Median	13	Third Quartile (Q3)	16	Maximum	22
Bin Range	Frequency																																																																
0-1	100																																																																
1-2	400																																																																
2-3	450																																																																
3-4	550																																																																
4-5	600																																																																
5-6	750																																																																
6-7	1250																																																																
7-8	1100																																																																
8-9	1100																																																																
9-10	1400																																																																
10-11	1400																																																																
11-12	1800																																																																
12-13	1800																																																																
13-14	1800																																																																
14-15	1800																																																																
15-16	1800																																																																
16-17	1800																																																																
17-18	1800																																																																
18-19	450																																																																
19-20	450																																																																
20-21	50																																																																
21-22	50																																																																
22-23	50																																																																
23-24	50																																																																
24-25	50																																																																
Statistic	Value (approx.)																																																																
Minimum	0																																																																
First Quartile (Q1)	9																																																																
Median	13																																																																
Third Quartile (Q3)	16																																																																
Maximum	22																																																																
<p>Comments:</p> <p>The data is LEFT skew</p>	<p>Comments:</p> <p>The outliers are 56</p>																																																																

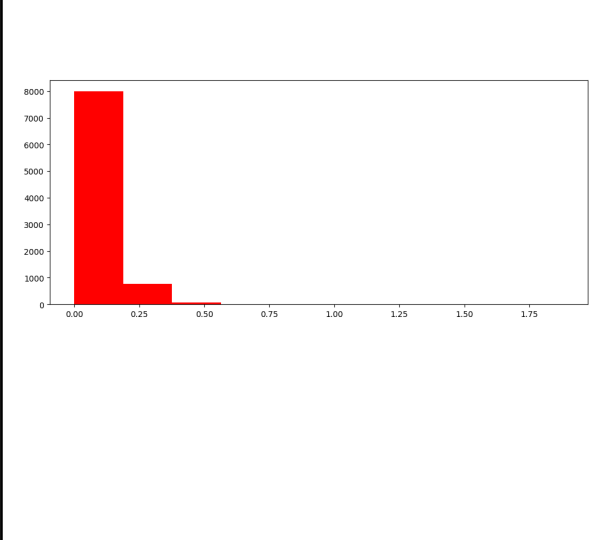
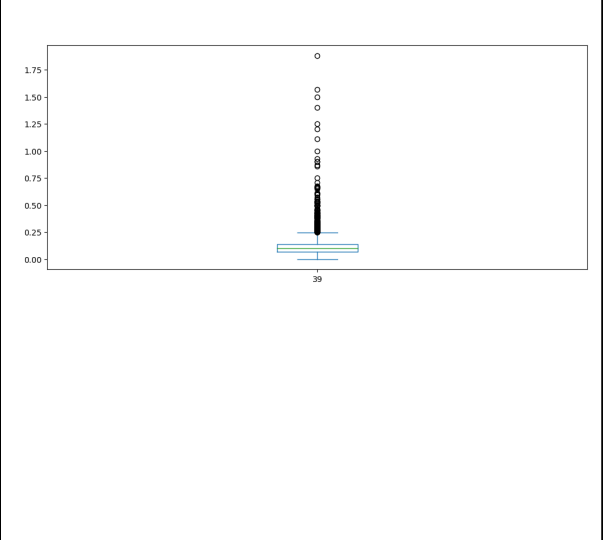
Met	
Histogram	Box Plot
	
Comments: The data is right skewed	Comments: the outliers are 838

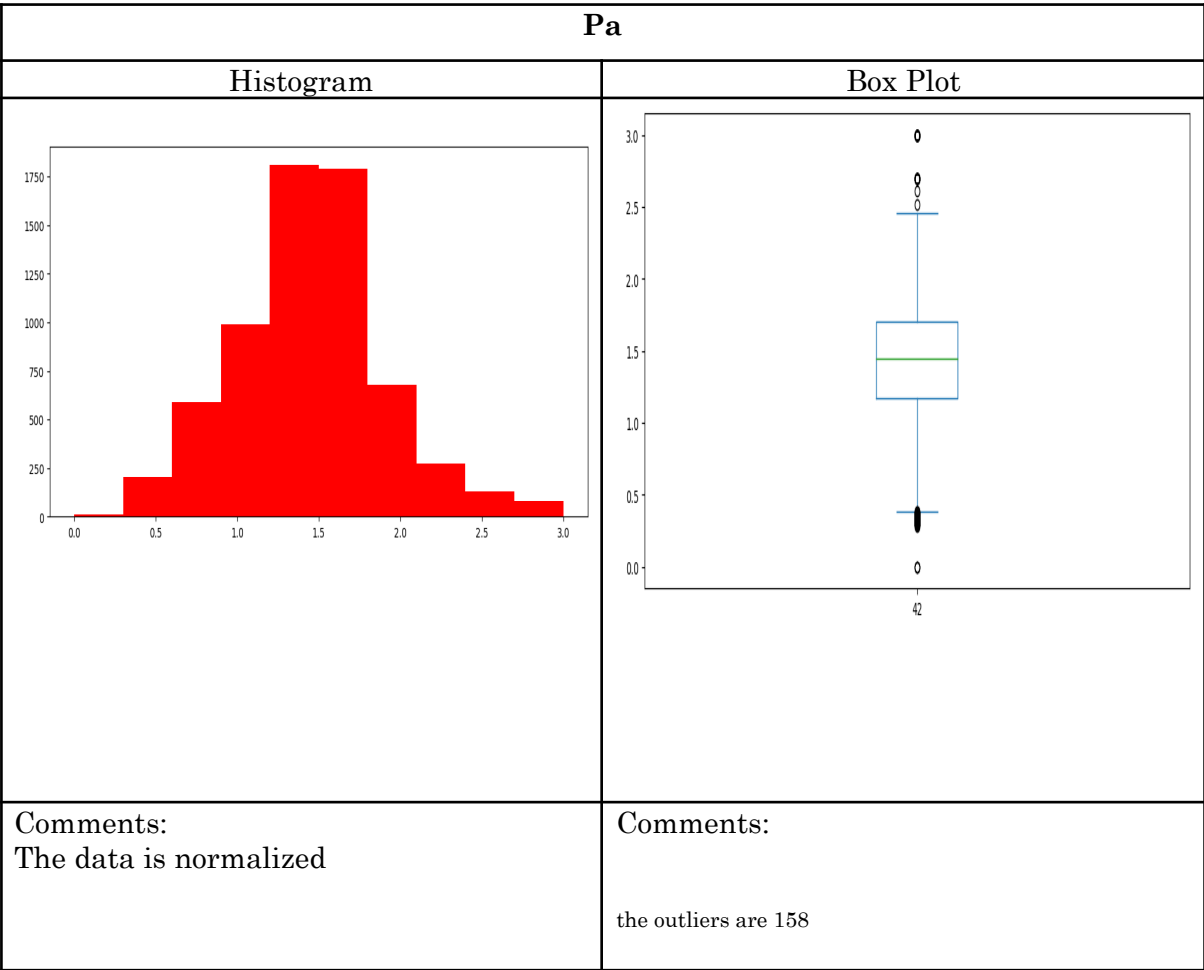
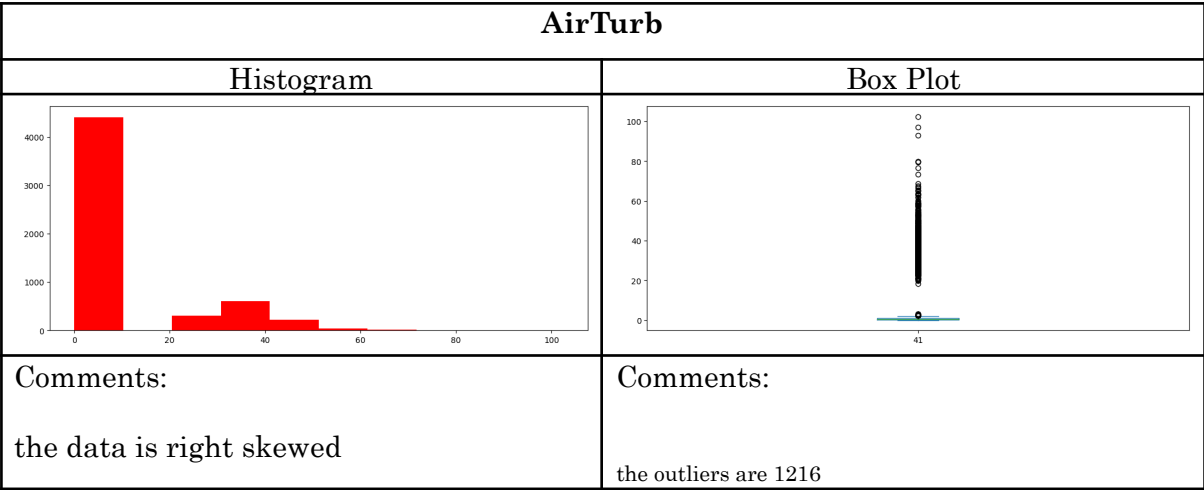
Dewpt	
Histogram	Box Plot
	
Comments: the data is RIGHT skewed	Comments: the outliers are 1

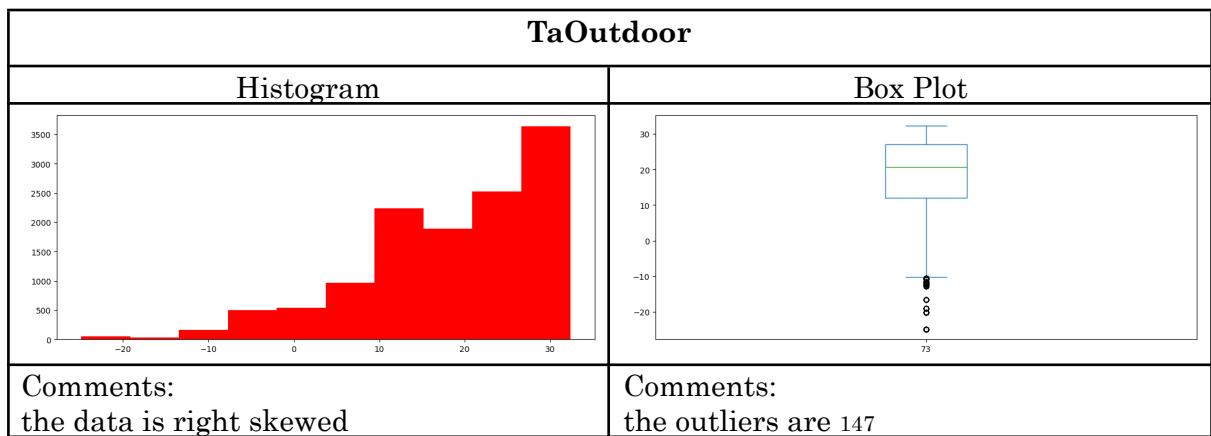
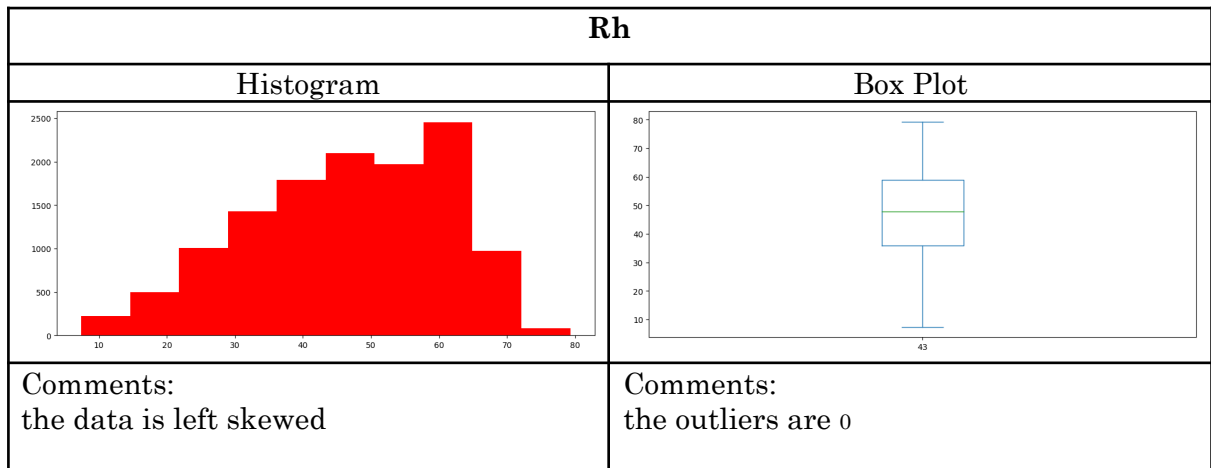


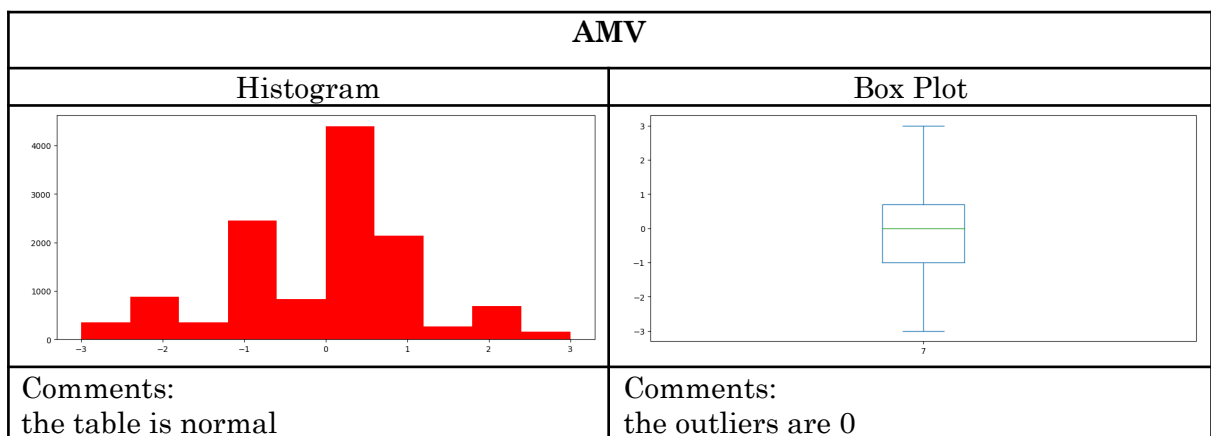
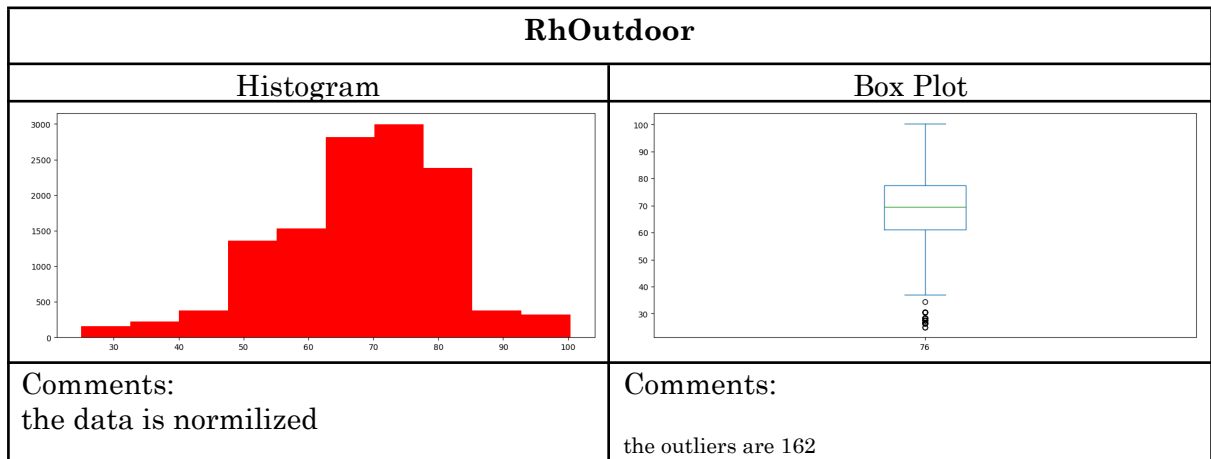
Ta																																													
Histogram	Box Plot																																												
 <p>A histogram showing the frequency distribution of Ta. The x-axis ranges from 16 to 30 with major ticks every 2 units. The y-axis ranges from 0 to 5000 with major ticks every 1000 units. The distribution is right-skewed, with a peak frequency of approximately 5200 at Ta = 22. The data is represented by red bars.</p> <table border="1"><thead><tr><th>Ta</th><th>Frequency</th></tr></thead><tbody><tr><td>16</td><td>0</td></tr><tr><td>17</td><td>0</td></tr><tr><td>18</td><td>0</td></tr><tr><td>19</td><td>0</td></tr><tr><td>20</td><td>200</td></tr><tr><td>21</td><td>1500</td></tr><tr><td>22</td><td>5200</td></tr><tr><td>23</td><td>3100</td></tr><tr><td>24</td><td>800</td></tr><tr><td>25</td><td>800</td></tr><tr><td>26</td><td>200</td></tr><tr><td>27</td><td>100</td></tr><tr><td>28</td><td>50</td></tr><tr><td>29</td><td>20</td></tr><tr><td>30</td><td>0</td></tr></tbody></table>	Ta	Frequency	16	0	17	0	18	0	19	0	20	200	21	1500	22	5200	23	3100	24	800	25	800	26	200	27	100	28	50	29	20	30	0	 <p>A box plot showing the distribution of Ta. The y-axis ranges from 16 to 30 with major ticks every 2 units. The box plot is blue with a green median line at approximately 23. The whiskers extend from approximately 19.5 to 26.5. There are numerous outliers represented by open circles, ranging from approximately 16 to 30. The data is right-skewed.</p> <table border="1"><thead><tr><th>Statistic</th><th>Value</th></tr></thead><tbody><tr><td>Minimum</td><td>16</td></tr><tr><td>Q1</td><td>22.5</td></tr><tr><td>Median</td><td>23</td></tr><tr><td>Q3</td><td>24</td></tr><tr><td>Maximum</td><td>26.5</td></tr></tbody></table>	Statistic	Value	Minimum	16	Q1	22.5	Median	23	Q3	24	Maximum	26.5
Ta	Frequency																																												
16	0																																												
17	0																																												
18	0																																												
19	0																																												
20	200																																												
21	1500																																												
22	5200																																												
23	3100																																												
24	800																																												
25	800																																												
26	200																																												
27	100																																												
28	50																																												
29	20																																												
30	0																																												
Statistic	Value																																												
Minimum	16																																												
Q1	22.5																																												
Median	23																																												
Q3	24																																												
Maximum	26.5																																												
<p>Comments:</p> <p>the data is right skewed</p>	<p>Comments:</p> <p>the outliers are 425</p>																																												

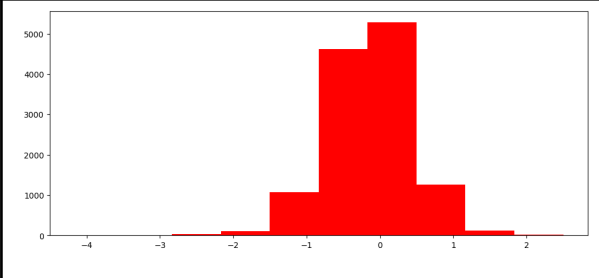
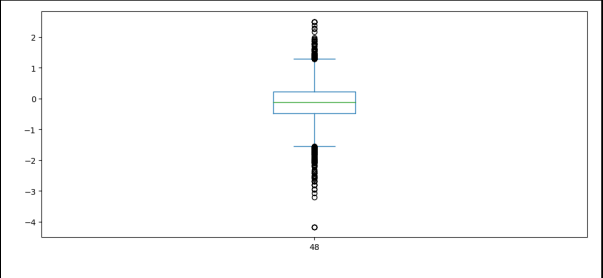
Tmrt																																																							
Histogram	Box Plot																																																						
 <p>A histogram showing the frequency distribution of Tmrt. The x-axis ranges from 15 to 35 with major ticks at 20, 25, 30, and 35. The y-axis ranges from 0 to 4000 with major ticks at 0, 1000, 2000, 3000, and 4000. The distribution is right-skewed, with a peak frequency of approximately 4500 at a Tmrt value of 24. The data is concentrated between 18 and 28.</p> <table border="1"><thead><tr><th>Tmrt Range</th><th>Frequency</th></tr></thead><tbody><tr><td>15-16</td><td>10</td></tr><tr><td>16-17</td><td>20</td></tr><tr><td>17-18</td><td>50</td></tr><tr><td>18-19</td><td>100</td></tr><tr><td>19-20</td><td>200</td></tr><tr><td>20-21</td><td>500</td></tr><tr><td>21-22</td><td>1000</td></tr><tr><td>22-23</td><td>2500</td></tr><tr><td>23-24</td><td>4500</td></tr><tr><td>24-25</td><td>3500</td></tr><tr><td>25-26</td><td>1500</td></tr><tr><td>26-27</td><td>1000</td></tr><tr><td>27-28</td><td>500</td></tr><tr><td>28-29</td><td>200</td></tr><tr><td>29-30</td><td>100</td></tr><tr><td>30-31</td><td>50</td></tr><tr><td>31-32</td><td>20</td></tr><tr><td>32-33</td><td>10</td></tr><tr><td>33-34</td><td>5</td></tr><tr><td>34-35</td><td>2</td></tr></tbody></table>	Tmrt Range	Frequency	15-16	10	16-17	20	17-18	50	18-19	100	19-20	200	20-21	500	21-22	1000	22-23	2500	23-24	4500	24-25	3500	25-26	1500	26-27	1000	27-28	500	28-29	200	29-30	100	30-31	50	31-32	20	32-33	10	33-34	5	34-35	2	 <p>A box plot showing the distribution of Tmrt. The y-axis ranges from 15 to 35 with major ticks at 20, 25, 30, and 35. The median is approximately 23.5. The interquartile range (IQR) is from about 22.5 to 24.5. Whiskers extend from 20 to 27. There are numerous outliers, with the highest values around 34 and 35.</p> <table border="1"><thead><tr><th>Statistic</th><th>Value</th></tr></thead><tbody><tr><td>Minimum</td><td>20</td></tr><tr><td>First Quartile (Q1)</td><td>22.5</td></tr><tr><td>Median</td><td>23.5</td></tr><tr><td>Third Quartile (Q3)</td><td>24.5</td></tr><tr><td>Maximum</td><td>27</td></tr></tbody></table>	Statistic	Value	Minimum	20	First Quartile (Q1)	22.5	Median	23.5	Third Quartile (Q3)	24.5	Maximum	27
Tmrt Range	Frequency																																																						
15-16	10																																																						
16-17	20																																																						
17-18	50																																																						
18-19	100																																																						
19-20	200																																																						
20-21	500																																																						
21-22	1000																																																						
22-23	2500																																																						
23-24	4500																																																						
24-25	3500																																																						
25-26	1500																																																						
26-27	1000																																																						
27-28	500																																																						
28-29	200																																																						
29-30	100																																																						
30-31	50																																																						
31-32	20																																																						
32-33	10																																																						
33-34	5																																																						
34-35	2																																																						
Statistic	Value																																																						
Minimum	20																																																						
First Quartile (Q1)	22.5																																																						
Median	23.5																																																						
Third Quartile (Q3)	24.5																																																						
Maximum	27																																																						
Comments: the data is right skewed	Comments: the outliers are 344																																																						

Vel	
Histogram	Box Plot
 <p>A histogram showing the frequency distribution of 'Vel' data. The x-axis ranges from 0.00 to 1.75 with increments of 0.25. The y-axis ranges from 0 to 8000 with increments of 1000. The distribution is highly left-skewed, with a very high frequency (approximately 8000) for values between 0.00 and 0.125, and a rapid decline in frequency for higher values, with a small bar around 0.375 and a negligible bar around 0.500.</p>	 <p>A box plot of the 'Vel' data. The y-axis ranges from 0.00 to 1.75 with increments of 0.25. The median is approximately 0.05. The interquartile range (IQR) is very narrow, spanning from about 0.02 to 0.08. Whiskers extend from approximately 0.00 to 0.15. Numerous outliers are plotted as open circles, starting from about 0.15 and extending up to 1.75. A label '39' is positioned below the plot area.</p>
Comments: the data is left skewed	Comments: the outliers are 309







PMV																											
Histogram	Box Plot																										
 <p>A histogram showing the frequency distribution of PMV values. The x-axis ranges from -4 to 2 with major ticks every 1 unit. The y-axis ranges from 0 to 5000 with major ticks every 1000 units. The bars are red. The distribution is roughly symmetric and bell-shaped, centered around 0. The highest frequency is in the bin from 0 to 1, reaching approximately 5200.</p> <table border="1"><thead><tr><th>PMV Bin</th><th>Frequency</th></tr></thead><tbody><tr><td>-3 to -2</td><td>100</td></tr><tr><td>-2 to -1</td><td>200</td></tr><tr><td>-1 to 0</td><td>1000</td></tr><tr><td>0 to 1</td><td>5200</td></tr><tr><td>1 to 2</td><td>1200</td></tr><tr><td>2 to 3</td><td>100</td></tr></tbody></table>	PMV Bin	Frequency	-3 to -2	100	-2 to -1	200	-1 to 0	1000	0 to 1	5200	1 to 2	1200	2 to 3	100	 <p>A box plot showing the distribution of PMV values. The y-axis ranges from -4 to 2 with major ticks every 1 unit. The box is light blue with a white median line at approximately 0. The whiskers extend from approximately -1.5 to 1.5. There are several outliers represented by small circles, with one outlier at approximately -4. The plot is labeled '48' at the bottom.</p> <table border="1"><thead><tr><th>Statistic</th><th>Value</th></tr></thead><tbody><tr><td>Minimum</td><td>-1.5</td></tr><tr><td>Q1</td><td>-0.5</td></tr><tr><td>Median</td><td>0.0</td></tr><tr><td>Q3</td><td>0.5</td></tr><tr><td>Maximum</td><td>1.5</td></tr></tbody></table>	Statistic	Value	Minimum	-1.5	Q1	-0.5	Median	0.0	Q3	0.5	Maximum	1.5
PMV Bin	Frequency																										
-3 to -2	100																										
-2 to -1	200																										
-1 to 0	1000																										
0 to 1	5200																										
1 to 2	1200																										
2 to 3	100																										
Statistic	Value																										
Minimum	-1.5																										
Q1	-0.5																										
Median	0.0																										
Q3	0.5																										
Maximum	1.5																										
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 median using	the percentages of outliers to total numbers of entries is greater than 2 percent, I assume the threshold 2 percent
Met	1887	Filled median using	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 median using	the percentages of outliers to total numbers of entries is greater than 2 percent, I assume the threshold 2 percent
Ta	1369	Filled median using	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

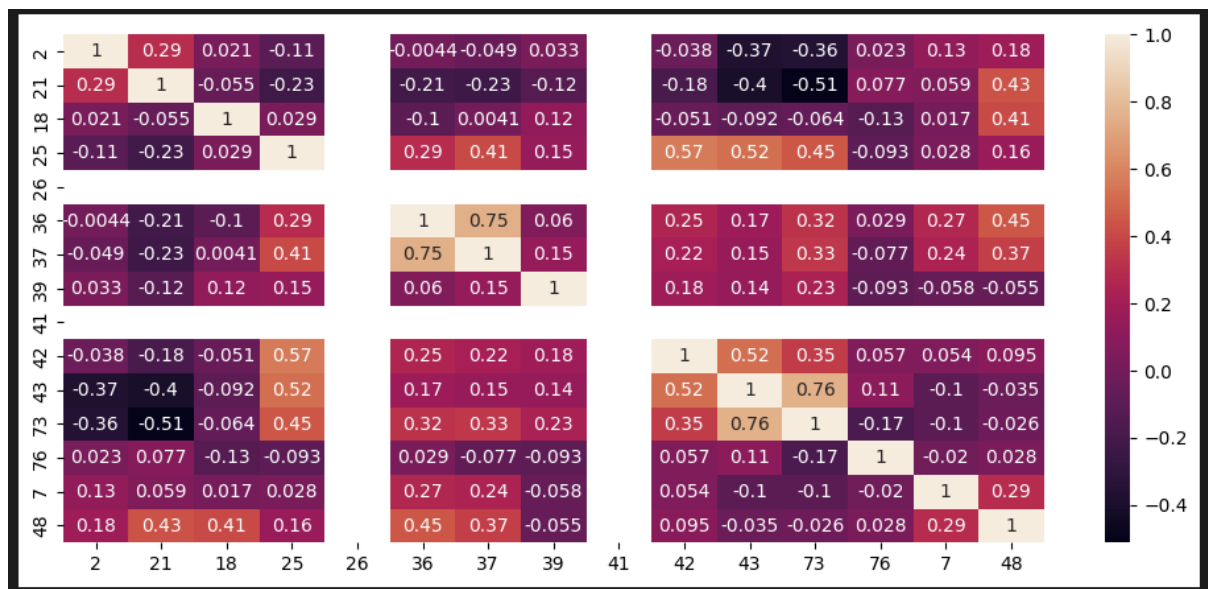
		spread out the values are.
AMV 55	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON is it measures how to spread out the values are.
PMV 43	IQR	We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3 the REASON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