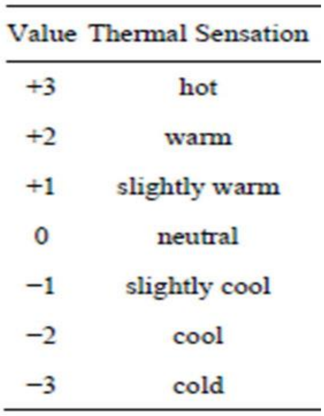
<MOAZZAM UMER>

<ANALYSIS REPORT>

For this project we 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.

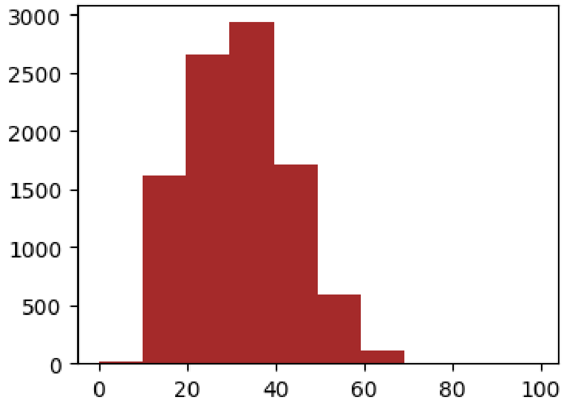
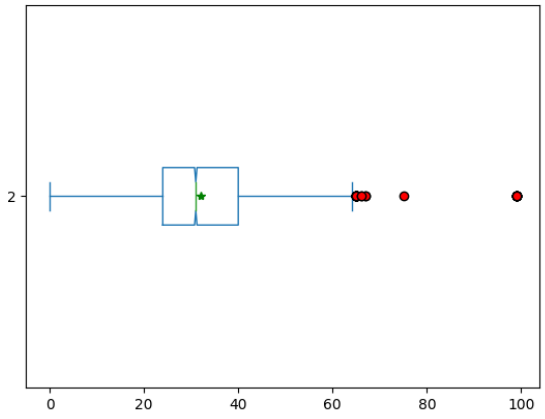
Column number	Feature Name	Feature Description
3	Age	Age
22	Clo	Clothing insulation
19	Met	Met Rate
26	Dewpt	Dewpt
27	PlaneRadTemp	plane radiant temperature
37	Ta	Average air temperature
38	Tmrt	Average mean radiant temperature
40	Vel	Air Velocity
42	AirTurb	Air Turbulance
43	Pa	Vapor Pressure
44	Rh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

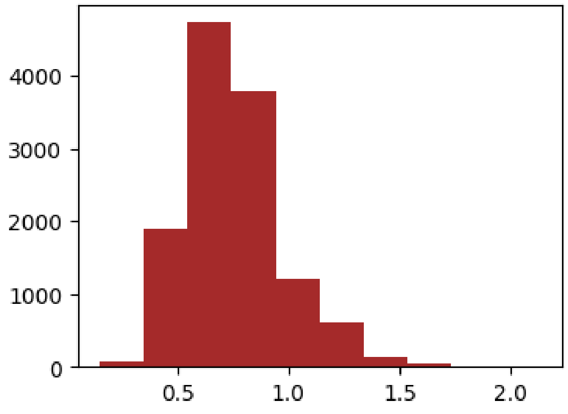
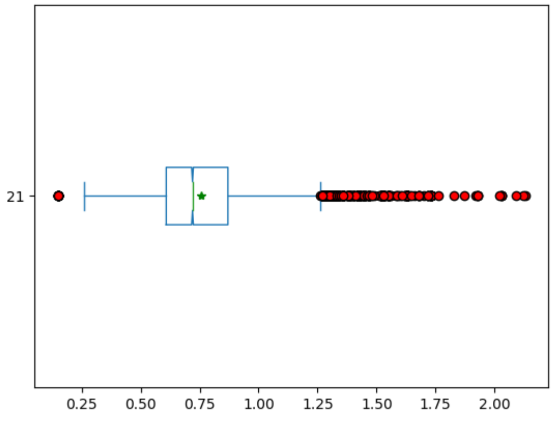
Part A. Preprocessing

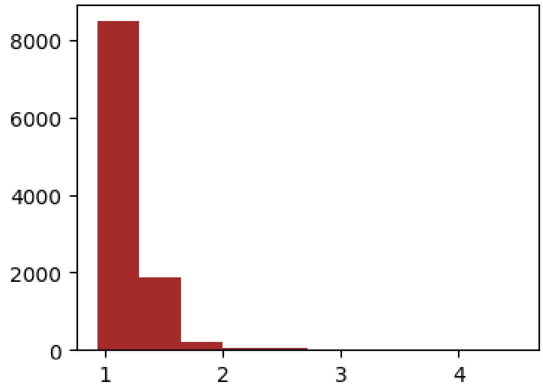
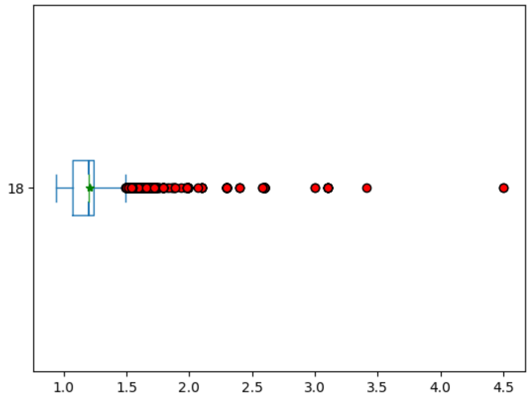
1. In this step, you are required to apply the preprocessing steps that you've covered in the course. Specifically, for each of the input dimension, fill in the following (add rows and complete the table for all input dimensions).

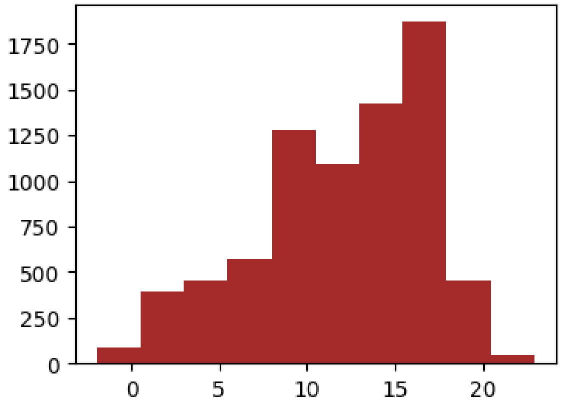
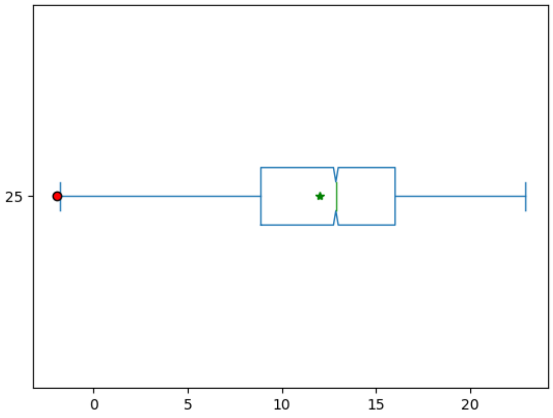
Dim Name	Data Type	Total Instances (without nulls)	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	STD
Age	Float64	9649	2917	37	0.0	99.0	24.0	31.98	31.0	133.48	11.55
Clo	Float64	12509	57	356	0.15	2.13	0.77	0.75	0.72	0.05	0.22
Met	Float64	10679	1887	838	0.93	4.5	1.2	1.2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

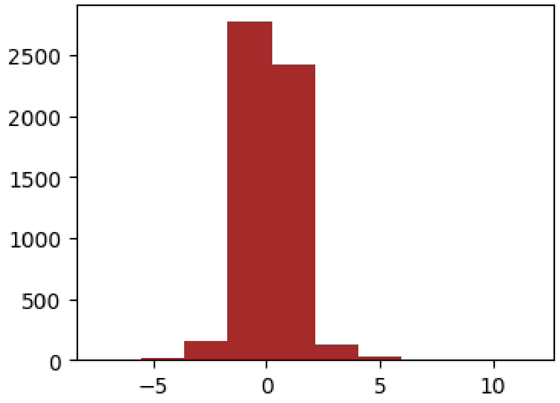
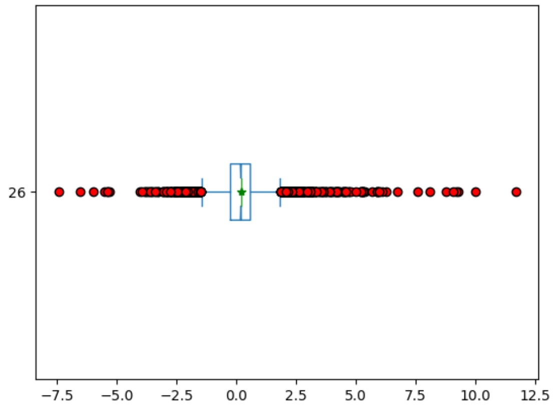
2. For each of the input dimension, plot histogram and comment the type of distribution the dimension exhibits. Further, visualize each dimension using a Box Plot. Specifically, for each of the input dimension, you're required to fill the following table (duplicate it for each of the 15 dimensions).

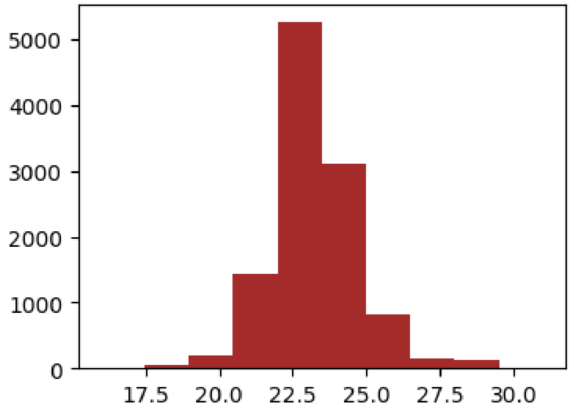
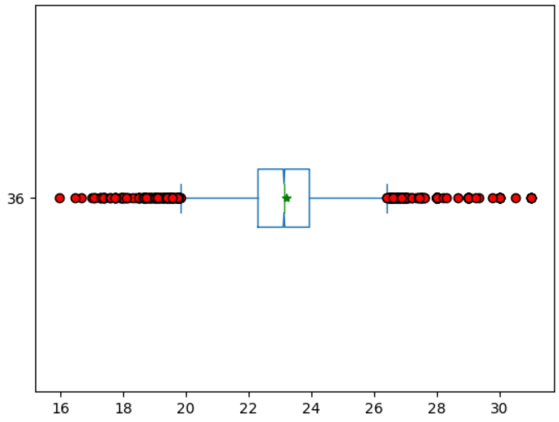
Age	
Histogram	Box Plot
	
Comments: Data distribution is right skewed. Most of the data lie between 20 to 50.	Comments: The standard limit of data is between 0 to 60. The notch in the box represents median and the marker represents mean of the data. Data points out of range are outliers

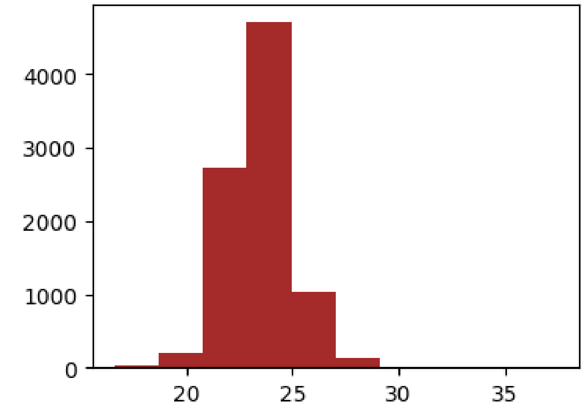
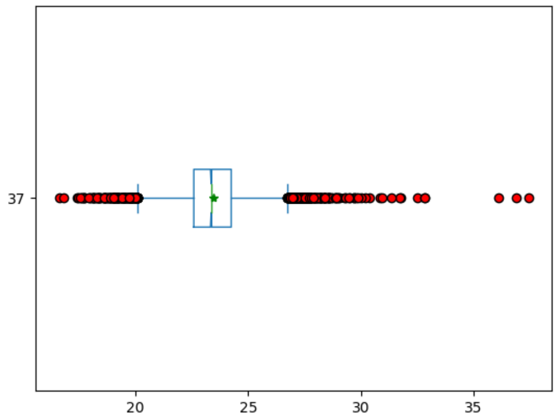
Clo	
Histogram	Box Plot
	
Comments: Data distribution is right skewed. Most of the data lie between 0.5 to 1.0.	Comments: The standard limit of data is between 0.25 to 1.25. The notch in the box represents median and the marker represents mean of the data. Data points out of range are outliers

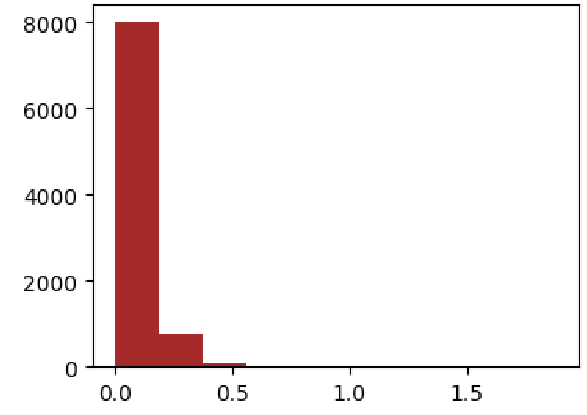
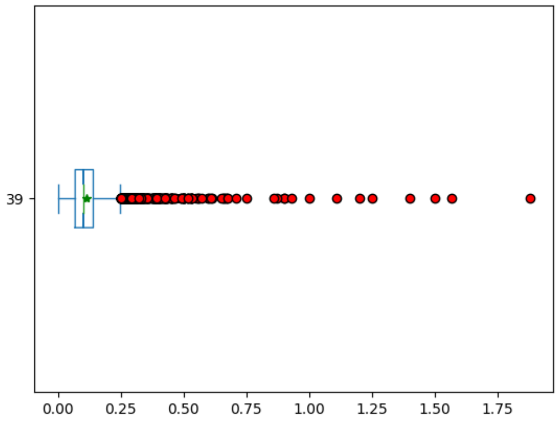
Met	
Histogram	Box Plot
	
Comments: Data distribution is right skewed. Most of the data lie between 1 to 1.5.	Comments: 50% of the data lies in the range of box. The notch in the box represents median and the marker represents mean of the data. Data points out of data range are outliers

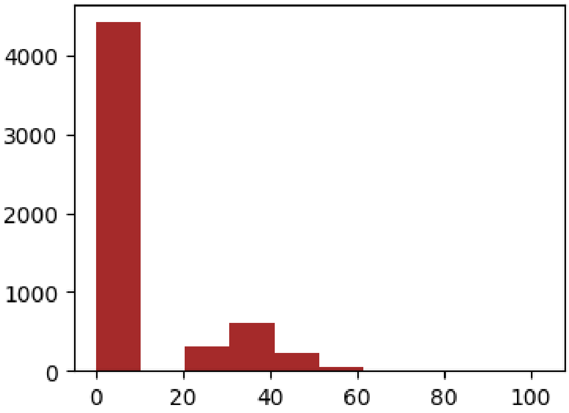
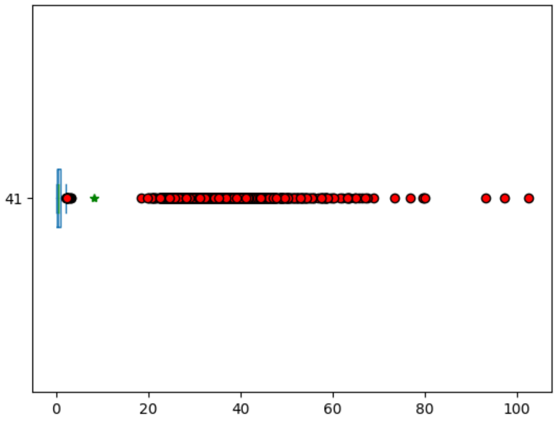
Dewpt	
Histogram	Box Plot
	
Comments: Data distribution is left skewed. Most of the data lie between 8 to 18.	Comments: 50% of the data lies in the range of box. The notch in the box represents median and the marker represents mean of the data. Data points out of data range are outliers

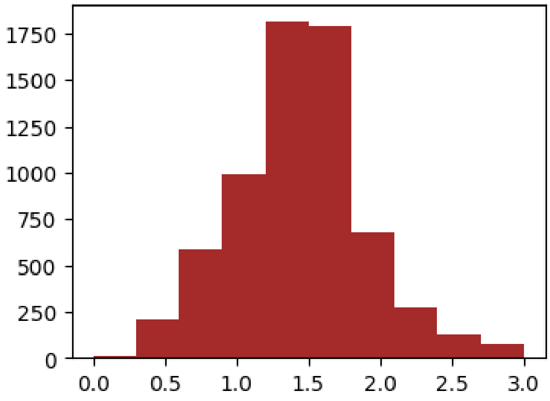
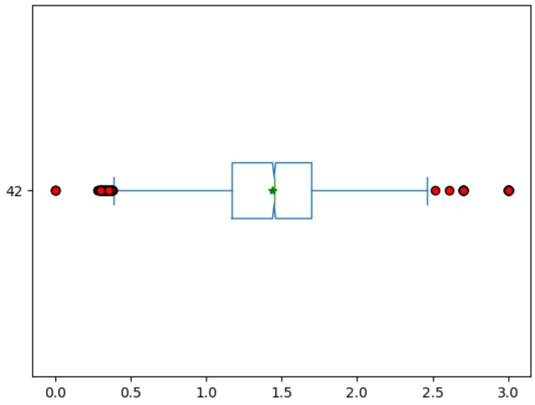
PlaneRadTemp	
Histogram	Box Plot
 	
Comments: Data distribution is symmetrical or unimodal. Most of the data lie at the center.	Comments: The standard range of data is between -2.5 to 2.5. The notch in the box represents median and the marker represents mean of the data. Data points out of data range are outliers

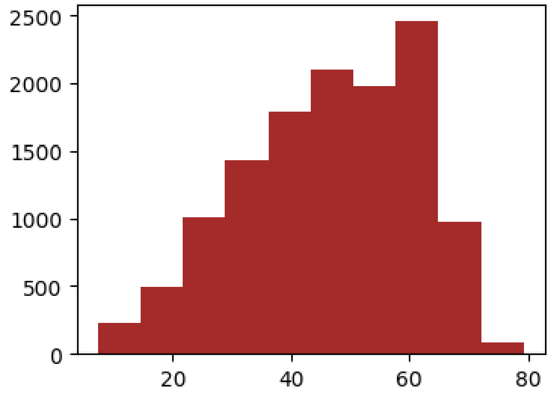
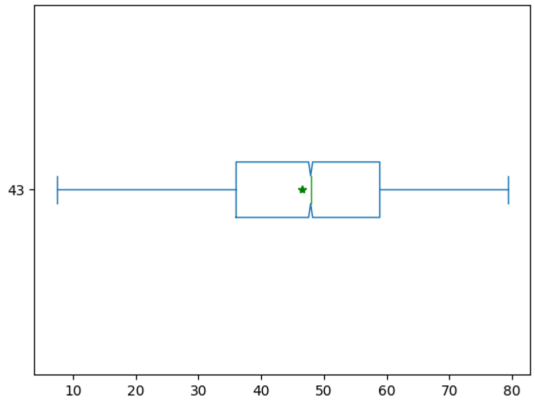
Ta	
Histogram	Box Plot
 	
Comments: Data distribution is symmetrical or unimodal. Most of the data lie at the center.	Comments: The standard range of data is between 20 to 26. The notch in the box represents median and the marker represents mean of the data. Data points out of data range are outliers

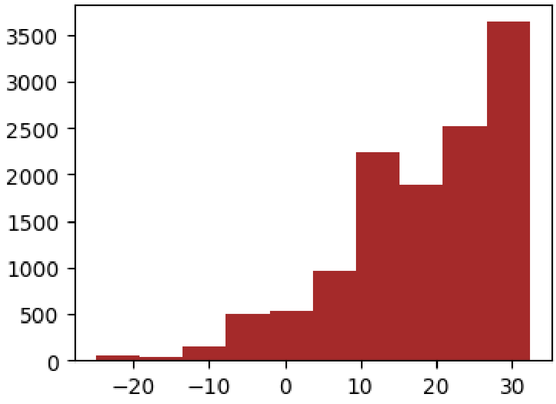
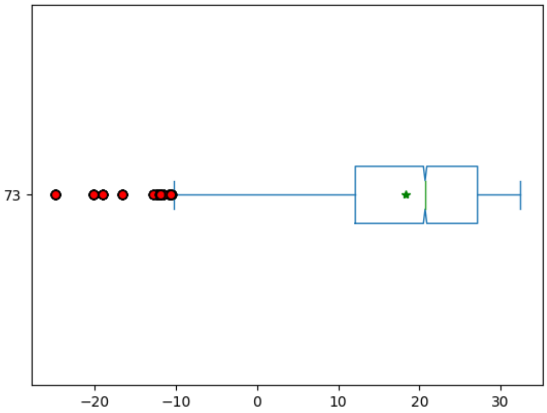
<table>
<tr><th colspan="2"><b>Tmrt</b></th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments: Data distribution is symmetrical or unimodal. Most of the data lie at the center.</td><td>Comments: The standard range of data is between 20 to 26. The notch in the box represents median and the marker represents mean of the data. Data points out of data range are outliers</td></tr>
</table>

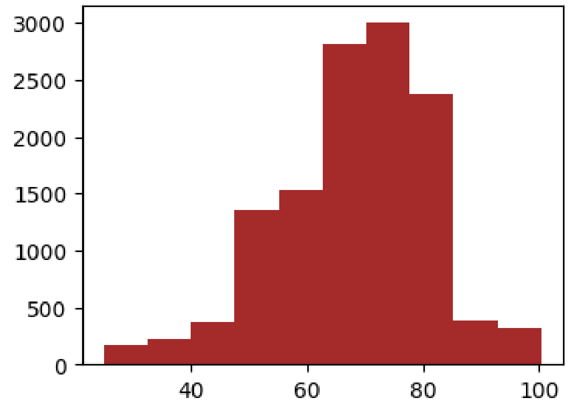
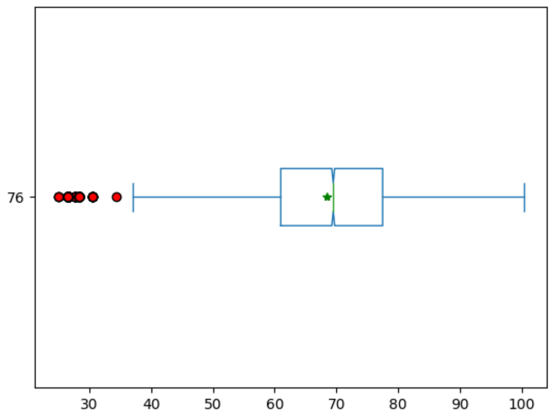
<table>
<tr><th colspan="2"><b>Vel</b></th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments: Data distribution is right skewed. Most of the data lie between 0.0 to 0.2.</td><td>Comments: The standard range of data is between 0.00 to 0.25. The notch in the box represents median and the marker represents mean of the data. Data points out of data range are outliers</td></tr>
</table>

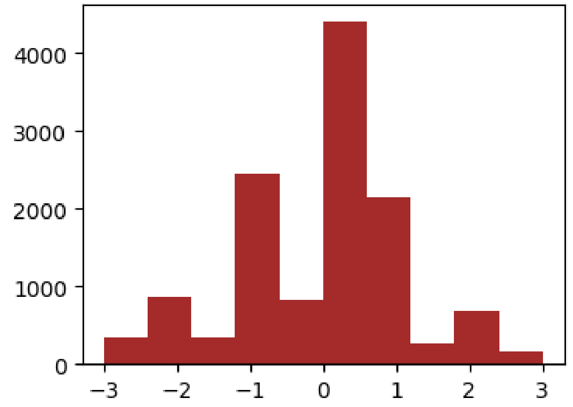
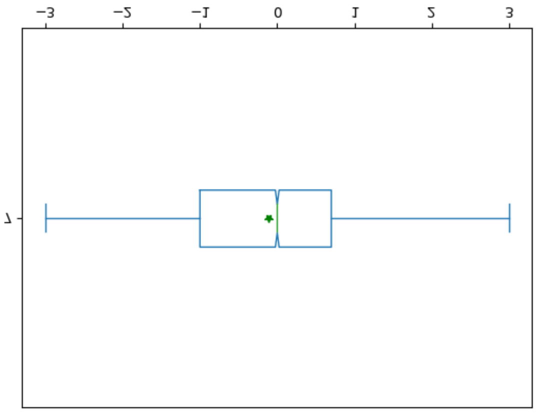
AirTurb	
Histogram	Box Plot
	
Comments: Data distribution multimodal.	Comments: Data distribution is multimodal. The notch in the box represents median and the marker represents mean of the data. Data points out of data range are outliers

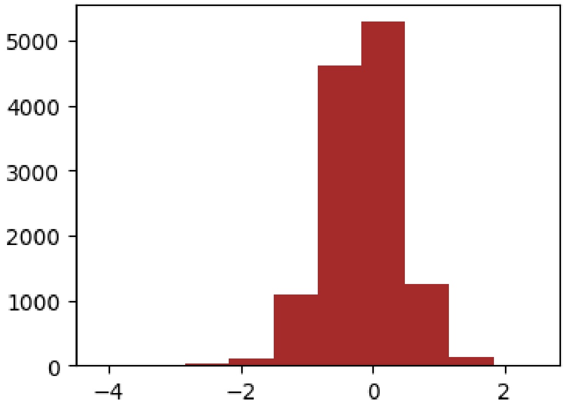
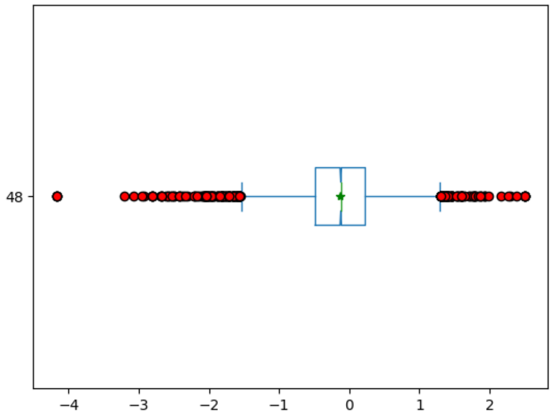
Pa	
Histogram	Box Plot
	
Comments: Data distribution is symmetrical or unimodal. Most of the data lie at the center.	Comments: The standard range of data is between 0.5 to 2.5. The notch in the box represents median and the marker represents mean of the data. Data points out of data range are outliers

<table>
<tr><th colspan="2">Rh</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td> </td><td> </td></tr>
<tr><td>Comments: Data distribution left skewed. Most of the data lie between 30 to 60.</td><td>Comments: The notch in the box represents median and the marker represents mean of the data. Data points out of data range are outliers</td></tr>
</table>

<table>
<tr><th colspan="2">TaOutdoor</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td> </td><td> </td></tr>
<tr><td>Comments: Data distribution left skewed. Most of the data lie between 10 to 30.</td><td>Comments: Data distribution is left skewed. The notch in the box represents median and the marker represents mean of the data. Data points out of data range are outliers</td></tr>
</table>

RhOutdoor	
Histogram	Box Plot
	
Comments: Data distribution is left skewed. Most of the data lies between 60 to 80.	Comments: 50% of the data lies in the range of box. The notch in the box represents median and the marker represents mean of the data. Data points out of data range are outliers

AMV	
Histogram	Box Plot
	
Comments: Data distribution is symmetrical or unimodal. Most of the data lie at the center.	Comments: The notch in the box represents median and the marker represents mean of the data. Data contains no outliers.

<table>
<tr><th colspan="2">PMV</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td> </td><td> </td></tr>
<tr><td>Comments: Data distribution is symmetrical or unimodal. Most of the data lie at the center.</td><td>Comments: The standard range of data is between -2 to 1. The notch in the box represents median and the marker represents mean of the data. Data points out of data range are outliers</td></tr>
</table>

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	2917	Filled using mean	There are very few or no outliers that may not affect mean so missing values are filled using mean
Clo	57	Filled using median	Number of outliers in this dimension can affect the mean so it is filled with median which is central value
Met	1887	Filled using median	Number of outliers in this dimension can affect the mean so it is filled with median which is central value
Dewpt	4901	Filled using mean	There are very few or no outliers that may not

			affect mean so missing values are filled using mean
PlaneRadTemp	7022	Filled using median	Number of outliers in this dimension can affect the mean so it is filled with median which is central value
Ta	1369	Filled using median	Number of outliers in this dimension can affect the mean so it is filled with median which is central value
Tmrt	3701	Filled using median	Number of outliers in this dimension can affect the mean so it is filled with median which is central value
Vel	3700	Filled using median	Number of outliers in this dimension can affect the mean so it is filled with median which is central value
AirTurb	6950	Filled using median	Number of outliers in this dimension can affect the mean so it is filled with median which is central value
Pa	6005	Filled using median	Number of outliers in this dimension can affect the mean so it is filled with median which is central value
Rh	35	Filled using mean	There are very few or no outliers that may not affect mean so missing values are filled using mean
TaOutdoor	19	Filled using mean	There are very few or no outliers that may not affect mean so missing values are filled using mean
RhOutdoor	19	Filled using mean	There are very few or no outliers that may not affect mean so missing values are filled using mean
AMV	55	Filled using mean	There are very few or no outliers that may not affect mean so missing

			values are filled using mean
PMV	43	Filled using mean	There are very few or no outliers that may not affect mean so missing values are filled using mean

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	37	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
Clo	356	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
Met	838	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
Dewpt	1	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
PlaneRadTemp	452	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
Ta	425	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
Tmrt	344	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
Vel	309	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR

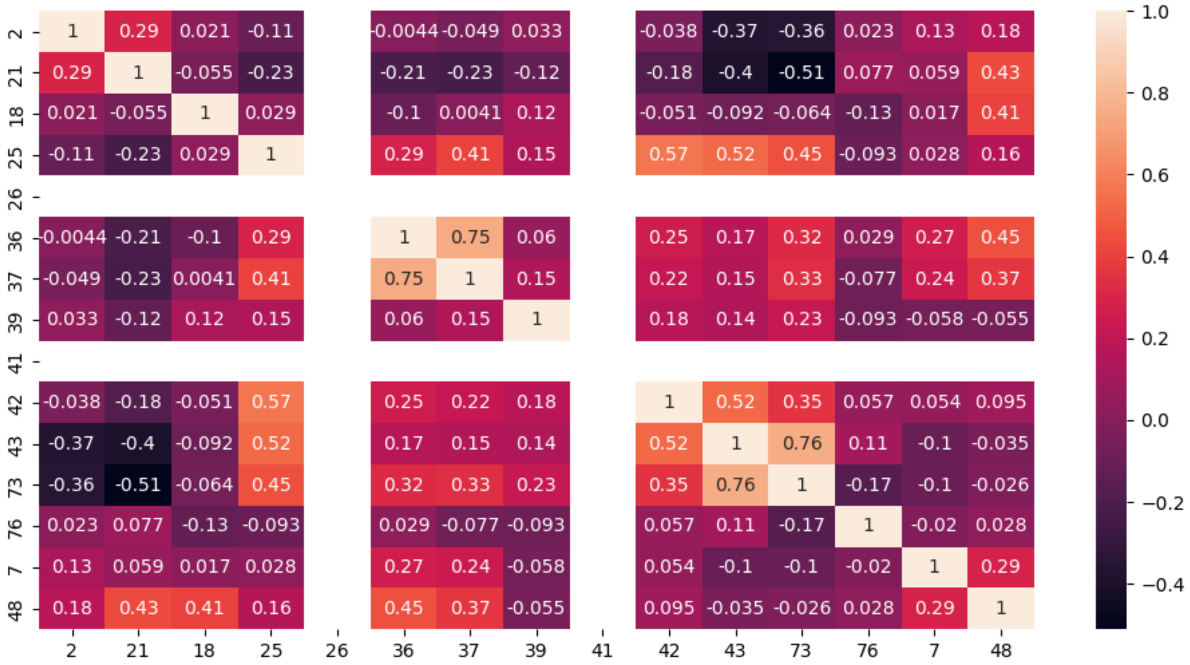
AirTurb	1216	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
Pa	158	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
Rh	0	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
TaOutdoor	147	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
RhOutdoor	162	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
AMV	0	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR
PMV	231	Handled using IQR method	To smooth data in a specific range between upper and lower limit of IQR

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	No, because high variance tells that data points are very spread from mean which means a good diversity in data that means more information for model training.
Clo	0.05	Yes, very low or zero variance indicates less diversity in data hence not good information for model. So this dimension can be dropped.
Met	0.04	Yes, very low or zero variance indicates less diversity in data hence not good information for model. So this dimension can be dropped.
Dewpt	23.42	The variance is comparatively higher so this dimension may not be dropped because it contain some diverse data.
PlaneRadTemp	1.08	Yes, very low or zero variance indicates less diversity in data hence not good information for model. So this dimension can be dropped.
Ta	23.13	The variance is comparatively higher so this dimension may not be dropped because it contain some diverse data.
Tmrt	2.25	Yes, very low or zero variance indicates less diversity in data hence not good information for model. So this dimension can be dropped.
Vel	0.006	Yes, very low or zero variance indicates less diversity in data hence not good information for model.
AirTurb	235.64	No, because high variance tells that data points are

		very spread from mean which means a good diversity in data that means more information for model training.
Pa	0.19	Yes, very low or zero variance indicates less diversity in data hence not good information for model. So this dimension can be dropped.
Rh	209.03	No, because high variance tells that data points are very spread from mean which means a good diversity in data that means more information for model training.
TaOutdoor	112.63	No, because high variance tells that data points are very spread from mean which means a good diversity in data that means more information for model training.
RhOutdoor	170.13	No, because high variance tells that data points are very spread from mean which means a good diversity in data that means more information for model training.

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



	2	21	18	25	26	36	37	39	41	42	43	73	76	7	48
2	1	0.29	0.021	-0.11		-0.0044	-0.049	0.033		-0.038	-0.37	-0.36	0.023	0.13	0.18
21	0.29	1	-0.055	-0.23		-0.21	-0.23	-0.12		-0.18	-0.4	-0.51	0.077	0.059	0.43
18	0.021	-0.055	1	0.029		-0.1	0.0041	0.12		-0.051	-0.092	-0.064	-0.13	0.017	0.41
25	-0.11	-0.23	0.029	1		0.29	0.41	0.15		0.57	0.52	0.45	-0.093	0.028	0.16
26															
36	-0.0044	-0.21	-0.1	0.29		1	0.75	0.06		0.25	0.17	0.32	0.029	0.27	0.45
37	-0.049	-0.23	0.0041	0.41		0.75	1	0.15		0.22	0.15	0.33	-0.077	0.24	0.37
39	0.033	-0.12	0.12	0.15		0.06	0.15	1		0.18	0.14	0.23	-0.093	-0.058	-0.055
41															
42	-0.038	-0.18	-0.051	0.57		0.25	0.22	0.18		1	0.52	0.35	0.057	0.054	0.095
43	-0.37	-0.4	-0.092	0.52		0.17	0.15	0.14		0.52	1	0.76	0.11	-0.1	-0.035
73	-0.36	-0.51	-0.064	0.45		0.32	0.33	0.23		0.35	0.76	1	-0.17	-0.1	-0.026
76	0.023	0.077	-0.13	-0.093		0.029	-0.077	-0.093		0.057	0.11	-0.17	1	-0.02	0.028
7	0.13	0.059	0.017	0.028		0.27	0.24	-0.058		0.054	-0.1	-0.1	-0.02	1	0.29
48	0.18	0.43	0.41	0.16		0.45	0.37	-0.055		0.095	-0.035	-0.026	0.028	0.29	1

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 dimensions are [AirTurb] (41) and [PlaneRadTemp] (26) because of zero variance and no correlation. These dimensions have same data with no diversity so therefore they have least information. The highest correlation is between the dimensions [TaOutdoor] (73) and [Rh] (43) that is 0.76. It means these dimensions are 76% related so these are less informative due to their correlation.

The smallest correlation is between the dimensions [Tmrt] (37) and [Met] (18) that is 0.004. It means these dimensions have very small or no similar data so both can provide diverse data and most information for model.

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

Dim name	Entropy	Info Gain	Reason
Age	4.9083		
Clo	7.6384		
Met	3.828		
Dewpt	6.01		
PlaneRadTemp	-0.0		
Ta	7.27		
Tmrt	7.11		
Vel	4.72		
AirTurb	-0.0		
Pa	2.3931		
Rh	11.033		
TaOutdoor	8.04		
RhOutdoor	7.42		

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 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.

Dataset=[Age,Clo,Met,Dewpt,PlaneRadTemp,Ta,Tmrt,Vel,AirTurb,Pa,Rh,TaOutdoor,RhOutdoor]

Feature Vector (indexes of dataset)	Performance achieved
Index: 5	20.2%
Index: 1,5	48.8%
Index: 1,2,5	73.7%
Index: 1,2,5,10	76.2%
Index: 1,2,5,7,10	77.3%
Index: 1,2,5,6,7,10	77.6%
Index: 0,1, 2, 5, 6, 7, 10	77.9%
Index: 0, 1, 2, 5, 6, 7, 10,11	78.1%
Index: 0, 1, 2, 5, 6, 7, 10, 11,12	78.2%
Index: 0, 1, 2, 3, 5, 6, 7, 10, 11, 12	78.2%
Index: 0, 1, 2, 3, 5, 6, 7, 9, 10, 11,12	78.2%
Index: 0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12	78.2%
Index: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	78.2%

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 (indexes pf dataset)	Performance achieved
Index: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	78.5%
Index: 0, 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12	78.5%
Index: 0, 1, 2, 3, 5, 6, 7, 9, 10, 11, 12	78.5%
Index: 0, 1, 2, 3, 5, 6, 7, 10, 11, 12	78.5%
Index: 0, 1, 2, 5, 6, 7, 10, 11, 12	78.5%
Index: 0, 1, 2, 5, 6, 7, 10, 11	78.5%
Index: 0, 1, 2, 5, 6, 7, 10	78.2%
Index: 1, 2, 5, 6, 7, 10	77.9%
Index: 1,2,5,7,10	77.6%
Index: 1,2,5,10	76.5%
Index: 1,2,5	74.0%
Index: 1,5	48.9%
Index: 5	20.2%

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.

Dataset=[Age,Clo,Met,Dewpt,PlaneRadTemp,Ta,Tmrt,Vel,AirTurb,Pa,Rh,TaOutdoor,RhOutdoor]

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

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 (indexes of dataset)	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	40.0%
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%

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.

PMV:

Optimal feature vector of Linear Regression: [Age,Clo,Met,Ta,Tmrt,Vel,Rh,TaOutdoor]

Optimal Parameter values:

Slope: [0.17127608, 1.63038218, 2.53695307, 1.29855889, 0.0554797, -1.6402265, 0.55752028]

Intercept: -6.510734207975688

Results after performing linear regression with 3-fold cross validation on optimal feature vector:

Feature Vector (indexes)	Performance Achieved
Indexes: 3	20.3%
Indexes: 1,3	49.2%
Indexes: 1,2,3	73.7%
Indexes: 1,2,3,6	76.1%
Indexes: 1,2,3,5,6	77.1%
Indexes: 1,2,3,4,5,6	77.4%
Indexes: 0,1,2,3,4,5,6	77.7%

AMV:

Optimal feature vector of Logistic Regression: [Age,Clo,Met,Ta,TaOutdoor,RhOutdoor]

Optimal Parameter values:

slope: [[-1.69995062, 0.66851565, -0.14374576, -3.4630399, 1.85646311]

[-1.07621328, -0.4939433, -0.29700983, -2.41723852, 1.61133338]

[0.42527379, -0.2868378, -0.49834243, -1.41727443, 0.63468891]

[1.08934985, -0.12896863, -1.01080899, 0.11030714, -0.43004845]

[0.52707207, 0.47852433, 0.30834143, 1.89281459, -0.55848888]

[0.03480724, 0.29351821, 1.38389517, 2.80767081, -1.50801068]

[0.69966096, -0.53080847, 0.25767041, 2.48676031, -1.60593739]]

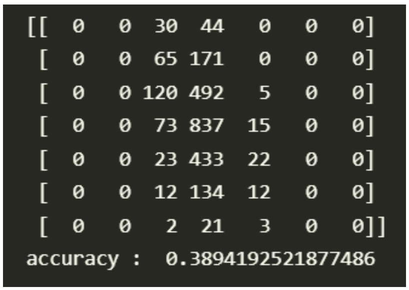
intercept: [-0.26059431, 1.42567285, 2.03965162, 2.63330672, -0.67733874, -2.59261472, -2.56808343]

Results after performing logistic regression with 3-fold cross validation on optimal feature vector:

Feature Vector (indexes)	Performance Achieved
Indexes: 1	38.0%

Indexes: 0,1	38.6%
Indexes: 0,1,3	38.7%
Indexes: 0,1,2,3	39.4%
Indexes: 0,1,2,3,4	39.7%

Confusion Matrix:



```
[[  0   0  30  44   0   0   0]
 [  0   0  65 171   0   0   0]
 [  0   0 120 492   5   0   0]
 [  0   0  73 837  15   0   0]
 [  0   0  23 433  22   0   0]
 [  0   0  12 134  12   0   0]
 [  0   0   2  21   3   0   0]]
accuracy :  0.3894192521877486
```