Introduction to Data Science

Course Project

Report Document

<Sophia Razzaq>

<21L-5607>

<BDS-3C>

Instructions: Read These Carefully Before Starting

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

Not following these instructions will lead to mark deduction.

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

Happy Coding 😺

TA Emails

Section A, C - Muhammad Maarij l192347@lhr.nu.edu.pk

Section B, D - Hira Ijaz l192377@lhr.nu.edu.pk

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

The following table shows the mapping of sensations:

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

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

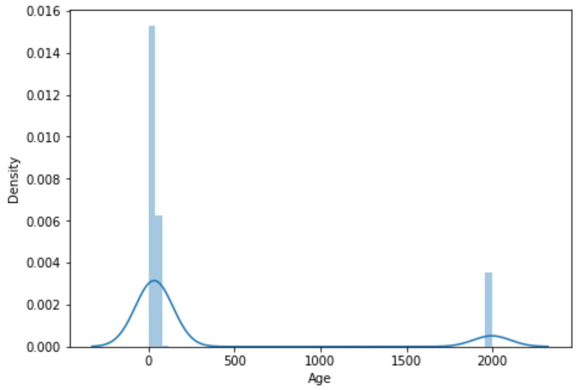
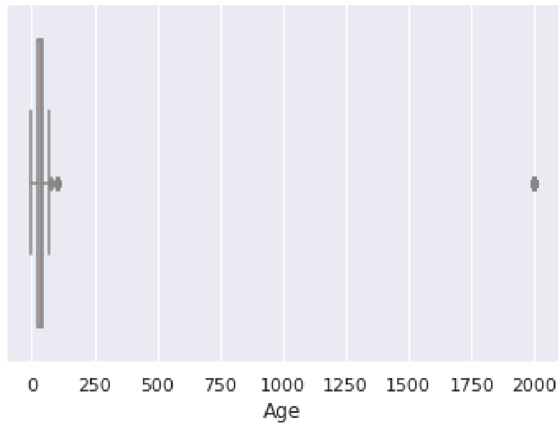
Column number	Feature Name	Feature Description
3	Age	Age
22	Clo	Clothing insulation
19	Met	Met Rate
26	Dewpt	Dewpt
27	PlaneRadTemp	plane radiant temperature
37	Ta	Average air temperature
38	Tmrt	Average mean radiant temperature
40	Vel	Air Velocity
42	AirTurb	Air Turbulance
43	Pa	Vapor Pressure
44	Rh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

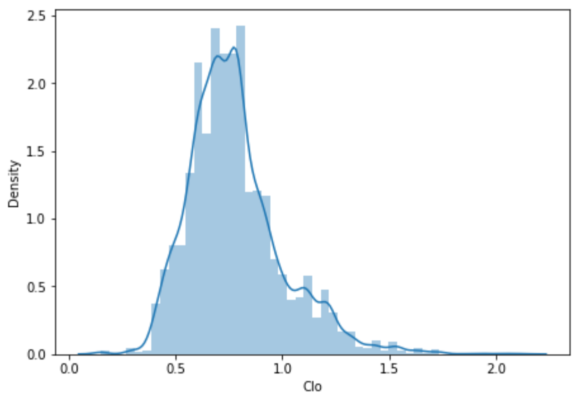
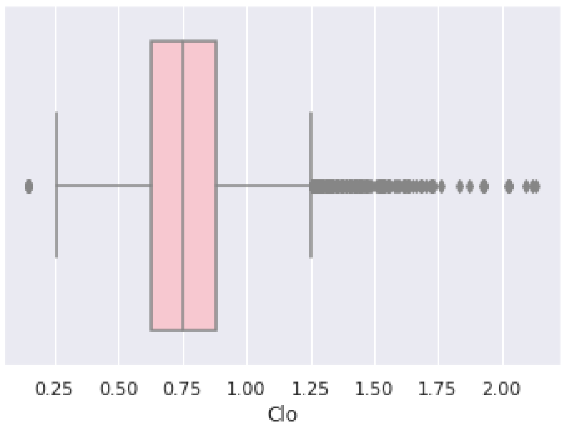
Part A. Preprocessing

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

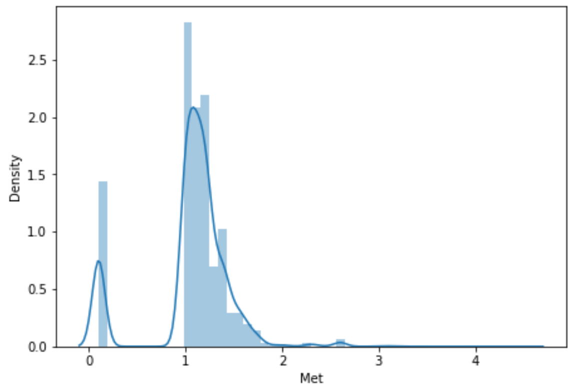
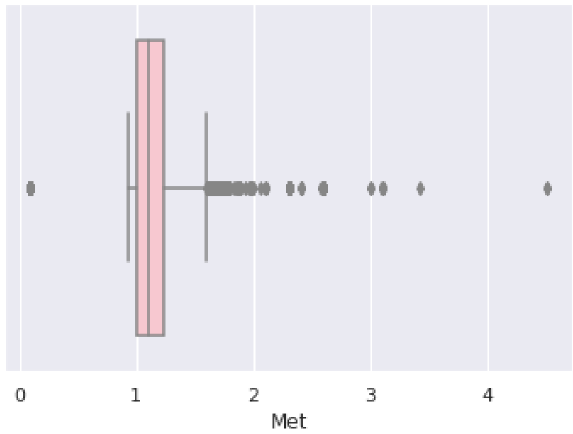
Dim Name	Data Type	Total Instances	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	STD
Age	float	9650	2915	1359	0	1996	24.0	308.63	35	462556.5	680.11
Clo	float	11159	1406	373	0.15	2.130	0.77	0.77	0.7	0.04	0.22199
Met	float	10678	1887	1732	0.10	4.50	1.0	1.06	1.1	0.18	0.4288
Dewpt	float	9014	3551	0	-1.9	26.89	17.4	13.61	14.1	34.84	5.903044
PlaneRedTemp	float	5544	7021	452	-7.42	11.7	0.3	0.21	0.2	1.0	1.0411
Ta	float	12545	20	539	15.96	31.0	23.2	23.17	23.12	2.0	1.432
Tmert	float	8864	3701	344	16.6	37.44	22.5	23.45	23.3	2.25	1.502
Vel	float	8865	3700	309	0.0	1.88	0.1	0.11	0.1	0.00	0.079044
AirTurb	float	6965	5600	2	0.0	102.45	0.5	18.2	0.5	627.05	25.0411
Pa	float	7910	4655	1352	0.0	27.7	2.1	5.1	1.55	66.5	8.156
Rh	float	12530	35	0	7.4	79.30	64.0	42.5	43.27	226.84	15.06
TaOutdoor	float	11197	1368	124	-24.9	32.35	27.56	17.1	18.2	113.75	10.665
RhOutdoor	float	12546	19	1349	0.0	100.35	0.0	61.0	68.79	610.30	24.704
AMV	float	12510	55	0	-3.0	3.00	0.0	0.1	0.0	1.2	1.10
PMV	float	11869	696	259	-4.17	2.5	0.1	-0.-7	-0.03	0.2	0.538

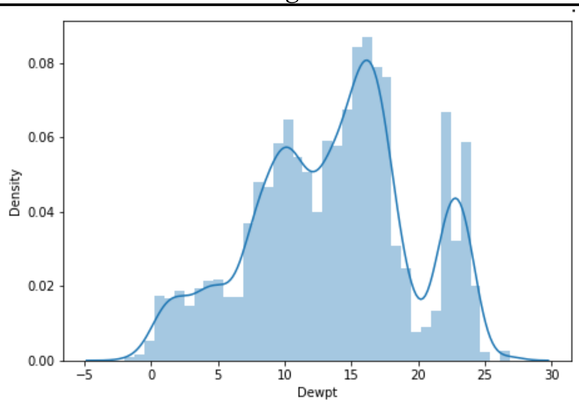
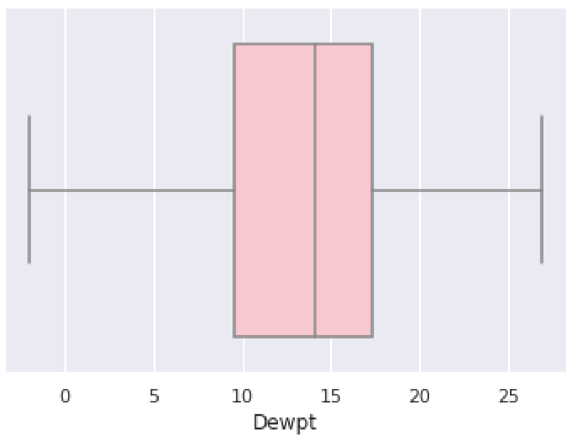
2. For each of the input dimension, plot histogram and comment the type of distribution the dimension exhibits. Further, visualize each dimension using a Box Plot. Specifically, for each of the input dimension, you're required to fill the following table (duplicate it for each of the 15 dimensions).

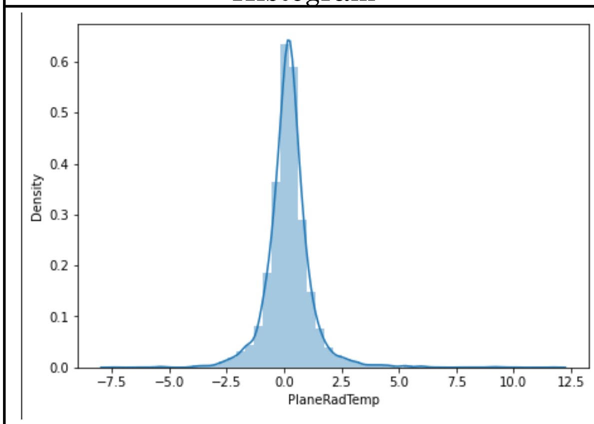
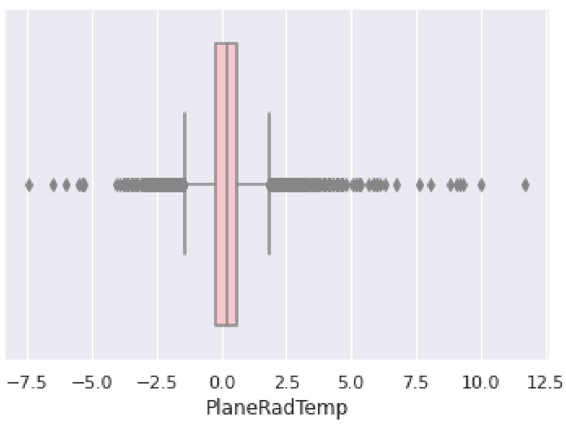
<table>
<tr><th colspan="2">Age</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments: positively skewed</td><td>Comments:Errors in Age column as values are greater than they should be</td></tr>
</table>

<table>
<tr><th colspan="2">Clo</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments:normally distributed but effected by outliers</td><td>Comments:many outliers</td></tr>
</table>

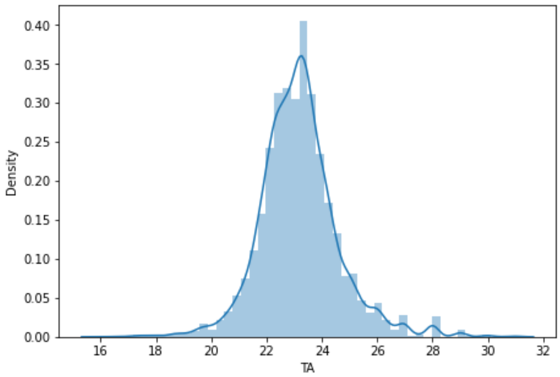
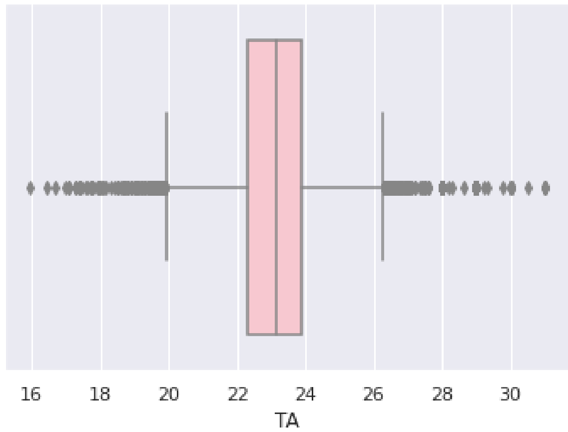
<table>
<tr><th colspan="2">Met</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
</table>

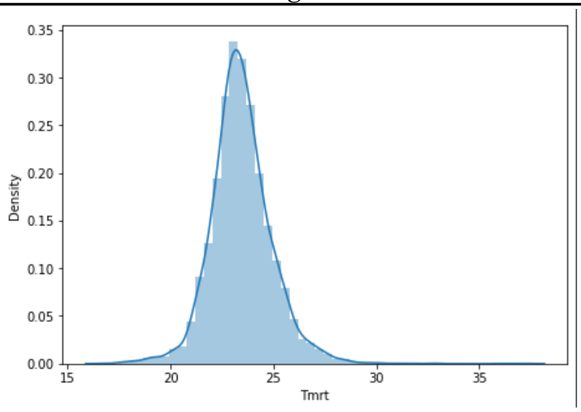
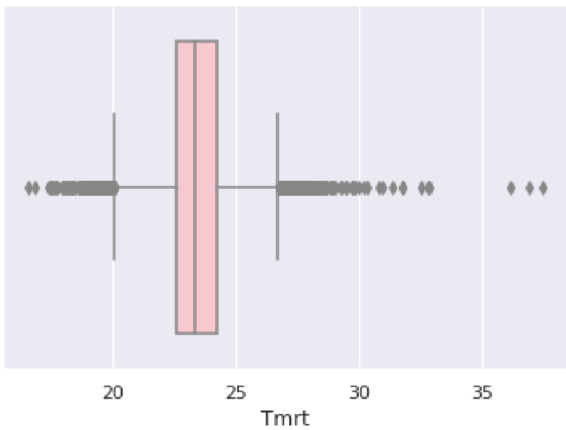
	
Comments:positively skewed	Comments:no outliers

Dewpth	
Histogram	Box Plot
	
Comments:normally distributed	Comments:no outliers,perfect

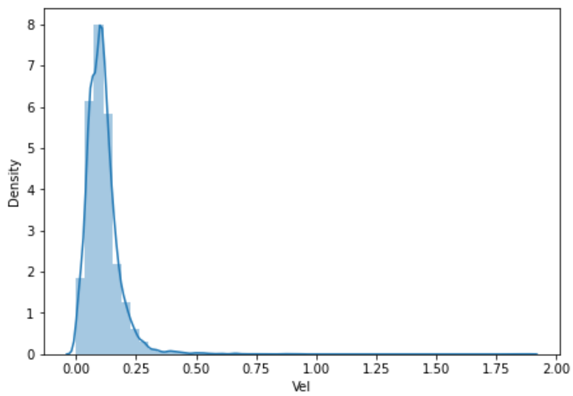
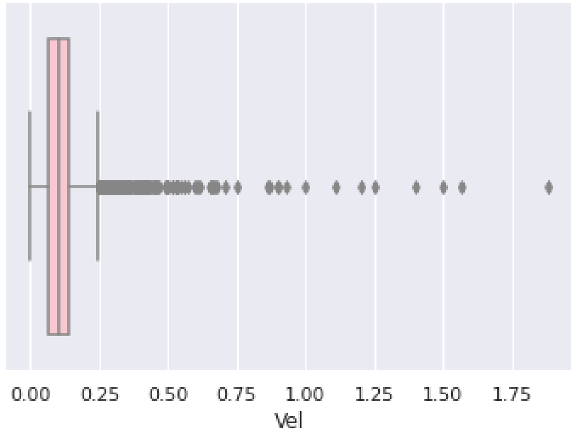
PlaneRadTemp	
Histogram	Box Plot
	

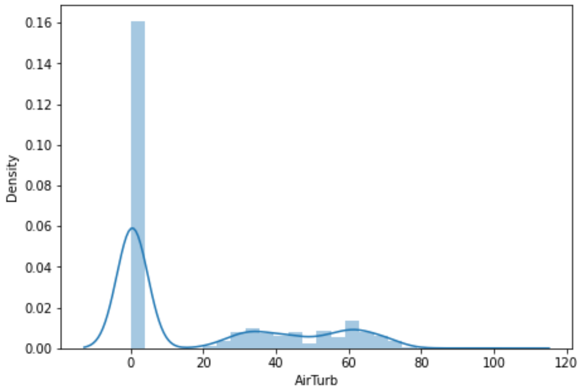
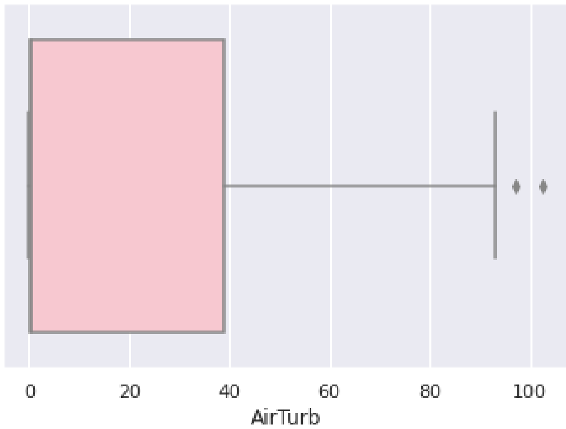
Comments:normal distribution	Comments:many outliers

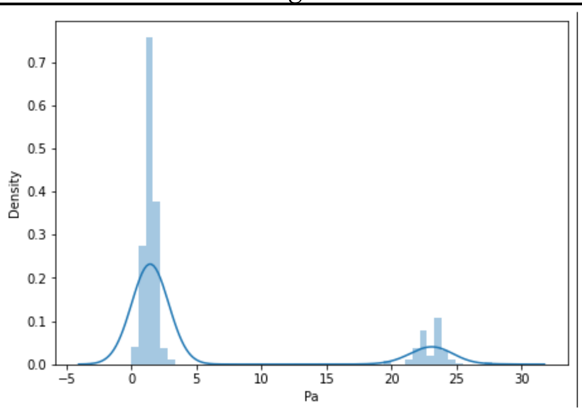
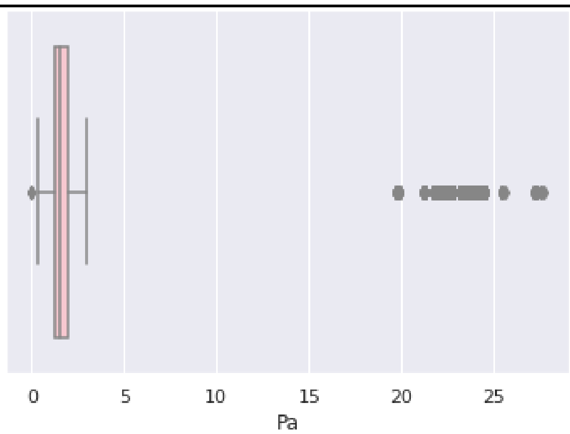
<table>
<tr><th colspan="2">TA</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments:normal distribution</td><td>Comments: many outliers</td></tr>
</table>

<table>
<tr><th colspan="2">Tmrt</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments:normal distribution</td><td>Comments:many outliers</td></tr>
</table>

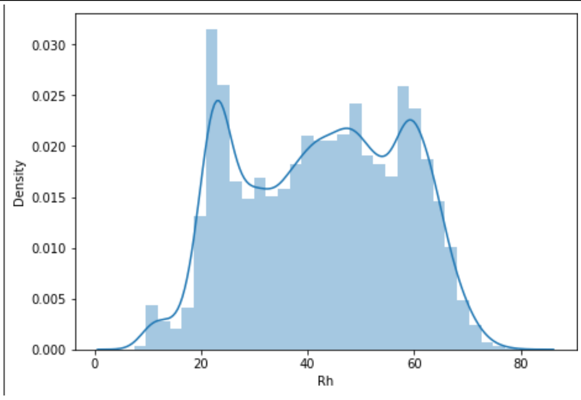
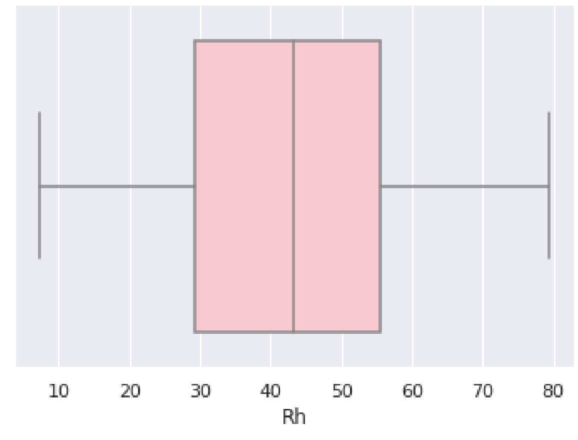
<table>
<tr><th colspan="2">Vel</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
</table>

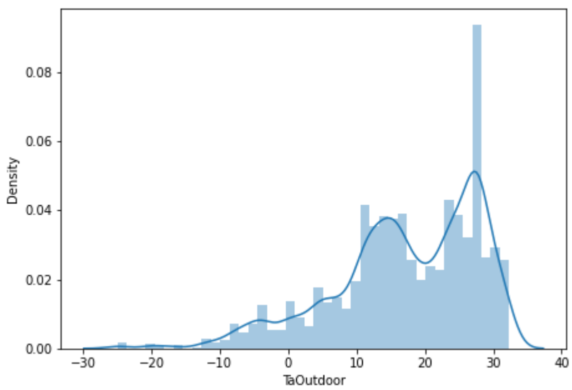
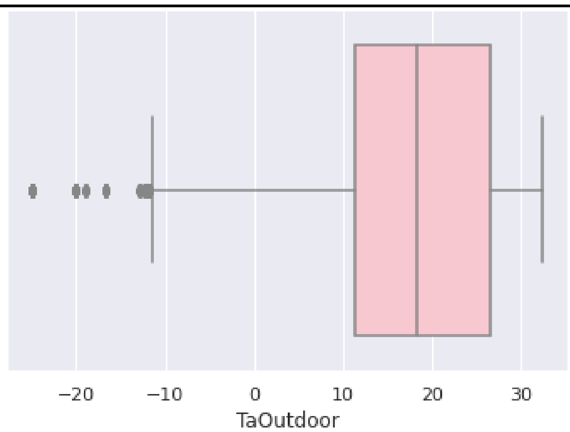
	
Comments:positively skewed	Comments:many outliers

AirTurb	
Histogram	Box Plot
	
Comments:positively skewed	Comments:very few outliers

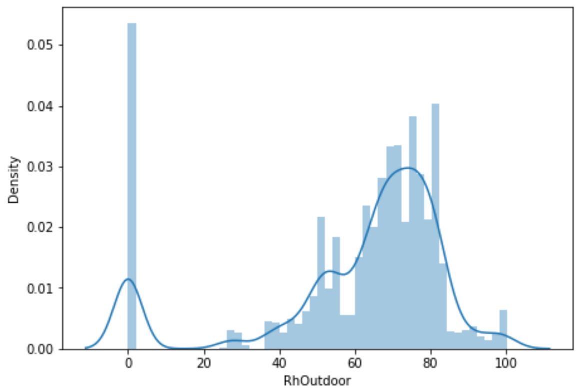
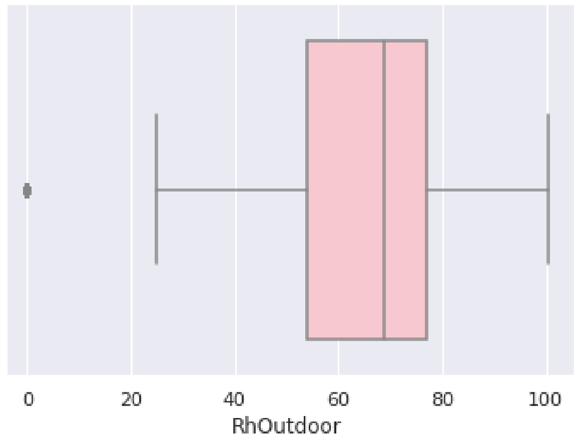
Pa	
Histogram	Box Plot
	

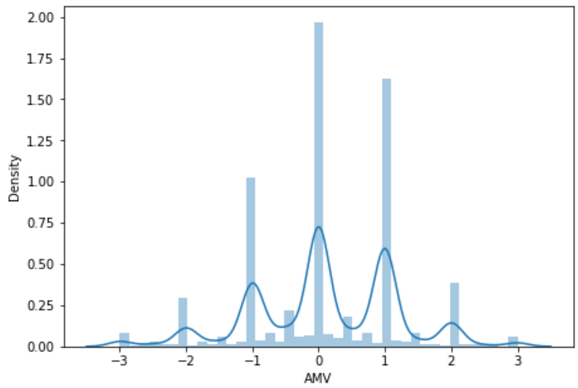
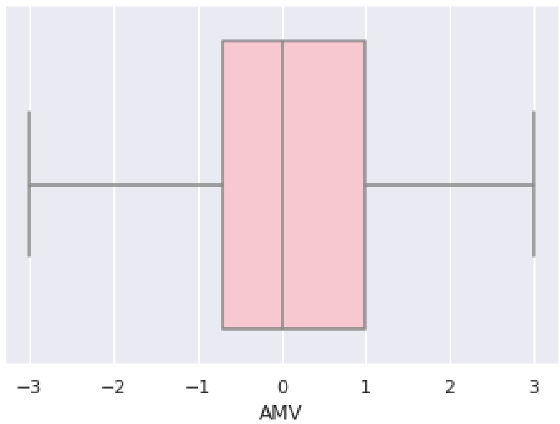
Comments:normal distribution	Comments:many outliers

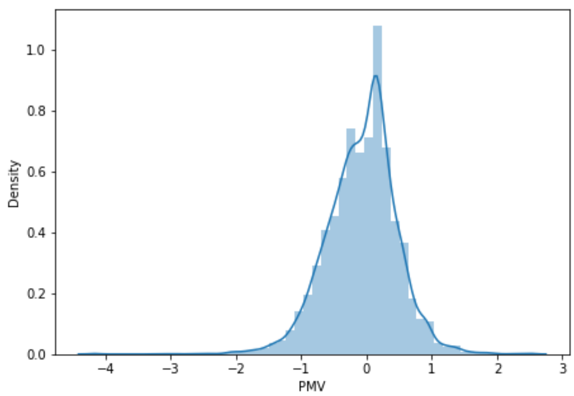
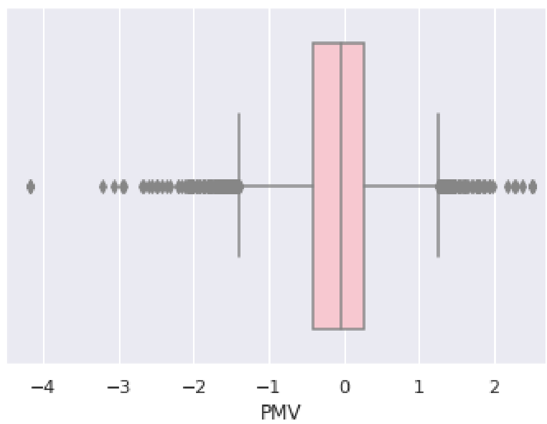
<table>
<tr><th colspan="2">Rh</th></tr>
<tr><td>Histogram</td><td>Box Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Comments:normal distribution</td><td>Comments:no outliers</td></tr>
</table>

<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:negatively skewed</td><td>Comments: few outliers</td></tr>
</table>

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

	
Comments:normal distribution	Comments:very few outliers

AMV	
Histogram	Box Plot
	
Comments:normal distribution	Comments:no outliers

PMV	
Histogram	Box Plot
	
Comments:normal distribution	Comments:many outliers

3. Find the missing values in each of the dimension (do this for both input and output dimensions), and fill these using an "appropriate" methodology that we've discussed in the class. You may also choose to drop a certain sample based on your analysis. Mention your approach and its justification.

Dim Name	Number of Missing Values	Filled using OR Dropped	Reason for selecting a certain approach
Age	2915	dropped	Too many missing values
Clo	1406	dropped	Mean can effect our original data as missing values ARE TOO MANY SO we dropped them
Met	1887	dropped	
Dewpt	3551	dropped	
PlanRedTemp	7021	dropped	
TA	20	FILLED using MEAN	missing values were very few
Tmrt	3701	dropped	
Vel	3700	dropped	
AirTurb	5600	dropped	
Pa	4655	dropped	
Rh	35	FILLED using MEAN	missing values were very few
TaOutdoor	1368	dropped	
RhOutdoor	19	FILLED using MEAN	missing values were very few
AMV	55	FILLED using MEAN	missing values were very few
PMV	696	dropped	

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	1379	smooth thru median	large num of outliers
COL	373	smooth thru median	large num of outliers
MET	1731	smooth thru median	large num of outliers
DEWPT	0	dropped	no outliers
PLANRADTEMP	452	smooth thru median	large num of outliers
TA	539	smooth thru median	large num of outliers
TMRT	342	smooth thru median	large num of outliers
VEL	309	smooth thru median	large num of outliers
AIRTURB	2	dropped	no outliers
PA	1352	smooth thru median	large num of outliers
RH	0	dropped	no outliers
TAOUTDOOR	124	smooth thru median	large num of outliers
RHOUTDOOR	1349	smooth thru median	large num of outliers
AMV	0	dropped	no outliers
PMV	259	smooth thru median	large num of outliers

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

METHODS

filter: information gain,corr with target,pairwise corr,var threshold

embedded

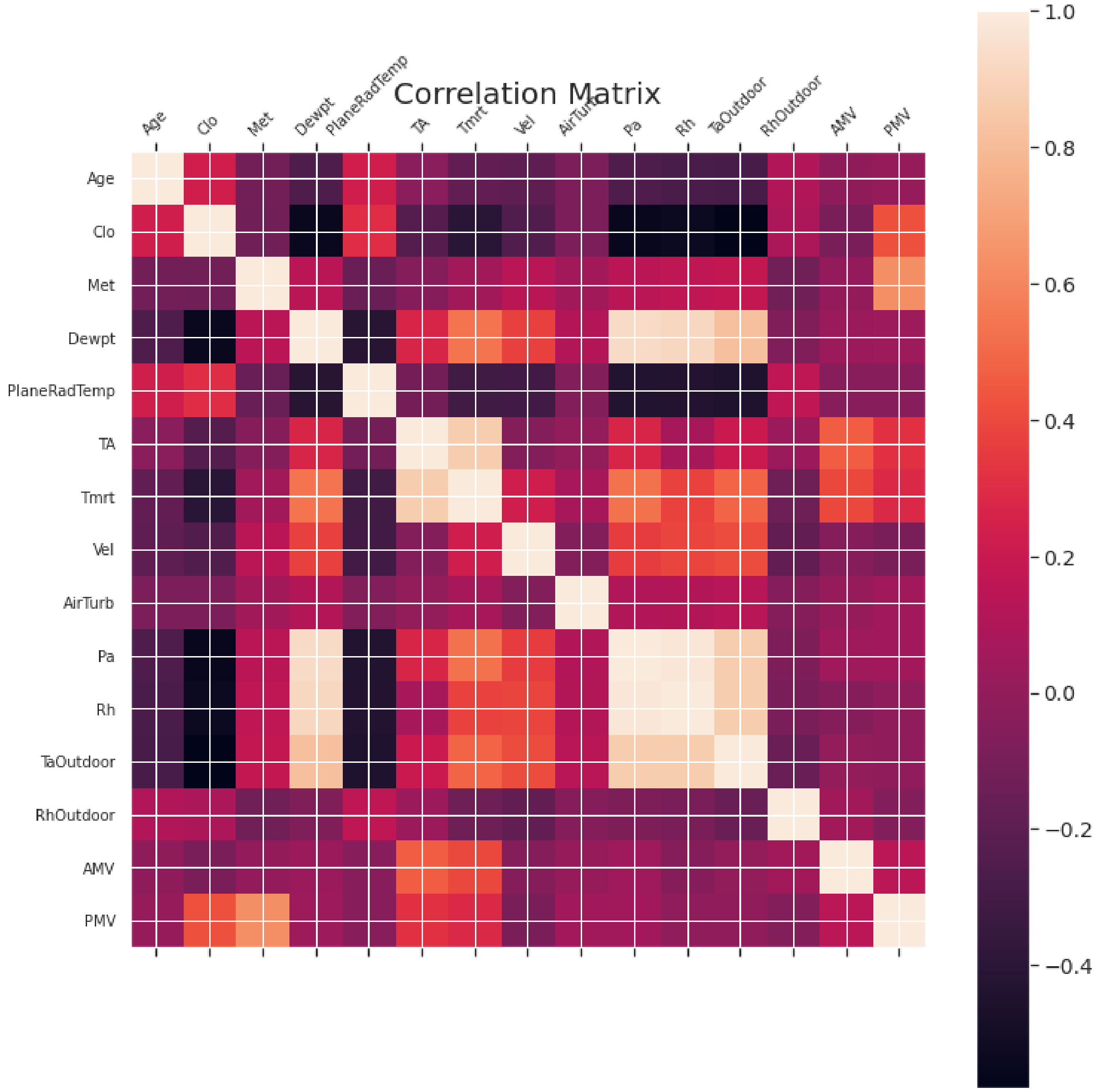
wrapper

Dim Name	Variance	Apply filter or no, reason
Age	105.46	no filter has been applied due to variety of data
Clo	0.08	yes, because value of threshold was zero
MET	0.45	yes, because value of threshold was zero
DEWPT	20.54	no filter has been applied due to variety of data
PLANREDTEMP	1.27	no filter has been applied due to variety of data
TA	1.09	no filter has been applied due to variety of data
TMRT	1.36	no filter has been applied due to variety of data
VEL	0.0015	yes, because value of threshold was zero
AIRTURB	0.51	yes, because value of threshold was zero
PA	0.186	yes, because value of threshold was zero
RH	216.4016	no filter has been applied due to variety of data
TAOUTDOOR	161.95	no filter has been applied due to variety of data
RHOUTDOOR	126.007	no filter has been applied due to variety of data
AMV	1.34	no filter has been applied due to variety of data
PMV	0.2	yes, because value of threshold was zero

Data was so vast and away from mean, no column had all same values thas why no filter hs been applied. Even if we apply we get back he original number of columns.

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

	Age	Clo	Met	Dewpt	PlaneRadTemp	TA	Tmrt	Vel	AirTurb	Pa	Rh	TaOutdoor	RhOutdoor	AMV	PMV
Age	1.000000	0.186989	-0.925265	0.699352	0.173246	-0.098748	-0.088926	0.000370	0.801354	0.997275	-0.497468	-0.273651	-0.871291	0.368119	0.125964
Clo	0.186989	1.000000	-0.079698	-0.330773	0.128719	-0.217531	-0.271392	-0.125144	-0.059634	-0.340309	-0.356797	-0.472870	0.081462	-0.040341	0.356949
Met	-0.925265	-0.079698	1.000000	-0.586927	0.010647	0.002913	-0.070052	0.041481	-0.714648	-0.924737	0.476259	-0.068409	0.749087	-0.231223	0.227606
Dewpt	0.699352	-0.330773	-0.586927	1.000000	-0.150887	0.236582	0.438170	0.161741	0.657528	0.684008	0.116914	0.706257	-0.630064	0.253157	0.245247
PlaneRadTemp	0.173246	0.128719	0.010647	-0.150887	1.000000	-0.013066	-0.066628	0.068490	0.178139	-0.213913	-0.195779	-0.186061	-0.038506	0.007446	0.023844
TA	-0.098748	-0.217531	0.002913	0.236582	-0.013066	1.000000	0.856613	0.080374	-0.076881	-0.104930	0.195278	0.325977	0.067023	0.240560	0.464282
Tmrt	-0.088926	-0.271392	-0.070052	0.438170	-0.066628	0.856613	1.000000	0.146586	-0.009348	0.392226	0.182434	0.417864	-0.117619	0.292196	0.432233
Vel	0.000370	-0.125144	0.041481	0.161741	0.068490	0.080374	0.146586	1.000000	0.352486	0.204115	0.136047	0.222993	-0.140477	-0.064838	-0.108440
AirTurb	0.801354	-0.059634	-0.714648	0.657528	0.178139	-0.076881	-0.009348	0.352486	1.000000	0.986307	-0.494304	0.183096	-0.846524	0.384551	0.149072
Pa	0.997275	-0.340309	-0.924737	0.684008	-0.213913	-0.104930	0.392226	0.204115	0.986307	1.000000	-0.600409	0.717430	-0.928318	0.389508	0.147360
Rh	-0.497468	-0.356797	0.476259	0.116914	-0.195779	0.195278	0.182434	0.136047	-0.494304	-0.600409	1.000000	0.752640	0.436306	-0.169236	0.006704
TaOutdoor	-0.273651	-0.472870	-0.068409	0.706257	-0.186061	0.325977	0.417864	0.222993	0.183096	0.717430	0.752640	1.000000	-0.168836	-0.022527	0.048520
RhOutdoor	-0.871291	0.081462	0.749087	-0.630064	-0.038506	0.067023	-0.117619	-0.140477	-0.846524	-0.928318	0.436306	-0.168836	1.000000	-0.279895	-0.073309
AMV	0.368119	-0.040341	-0.231223	0.253157	0.007446	0.240560	0.292196	-0.064838	0.384551	0.389508	-0.169236	-0.022527	-0.279895	1.000000	0.263758
PMV	0.125964	0.356949	0.227606	0.245247	0.023844	0.464282	0.432233	-0.108440	0.149072	0.147360	0.006704	0.048520	-0.073309	0.263758	1.000000



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)

<u>For PMV:</u>

Most informative dimensions: Age, Vel, AirTurb, Pa, Rh, TaOutdoor, RhOutdoor

Least informative dimensions: AMV, Met, Clo, Dewpt, PlaneRadtemp, Ta, Tmrt

<u>For AMV:</u>

Most informative dimensions: Age, Met, Clo, Vel, AirTurb, Pa, Rh, TaOutdoor, RhOutdoor

Least informative dimensions: Dewpt, PlaneRadTemp, Ta, Tmrt, PMV

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

Dim name	Entropy	Info Gain	Reason
Age	3.69	0.075	selected, info gain>
Clo	4.25	10.283	selected, info gain>
Met	3.18	0.75	selected, info gain>
dEWPT	6.28	3.59	selected, info gain>
PlaneRadTemp	5.97	3.408	selected, info gain>
TA	5.42	2.606	selected, info gain>
Tmrt	5.45	2.63	selected, info gain>
Vel	2.56	0.21	selected, info gain>
AirTurb	5.9	0.3	selected, info gain>
Pa	6.51	0.4	selected, info gain>
Rh	5.02	1.2	selected, info gain>
TaOutdoor	4.2	0.367	selected, info gain>
RhOutdoor	3.9	0.392	selected, info gain>
AMV	2.49	0.578	selected, info gain>
PMV	5.26	0.225	selected, info gain>

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.

Splitting the data randomly by using train_test_split function using sklearn library and normalizing:

```
# splitting the data randomly into 80/20 percent. Where 80% represents the training data. Also normalizing the dataset as you see fit.
from sklearn.model_selection import train_test_split
X = df.drop(['PMV'], axis=1)
y = df['PMV']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

```

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

Feature Vector	Performance achieved
Age, Pa, Rh, TaOutdoor, RhOutdoor	93%

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

Feature Vector	Performance achieved
['Age', 'Clo', 'TA', 'Tmrt', 'Vel', 'Pa', 'TaOutdoor	93%

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

Feature Vector	Performance achieved
Age, Vel, Rh, TaOutdoor, RhOutdoor	79%

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

Feature Vector	Performance achieved
Met, AirTurb, Rh, TaOutdoor, RhOutdoor	79%

['Age', 'Vel', 'Rh', 'TaOutdoor', 'RhOutdoor'] is the optimal feature vector as it is giving best accuracy than other features

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

3-fold validation

```
# define the model
model = LinearRegression()

# define the evaluation procedure
cv = KFold(n_splits=3, random_state=1, shuffle=True)

# evaluate the model
scores = cross_val_score(model, X_train, y_train, scoring='r2', cv=cv, n_jobs=-1)

# report performance
print('R2: %.3f (%.3f)' % ((scores).mean(), scores.std()))
```

confusion matrix

```
array([[  0,   0,   0,  26,   4,   0,   0],
       [  0,   0,   0, 117,  18,   0,   0],
       [  0,   0,   0, 394,  55,   0,   0],
       [  0,   0,   0, 960,  91,   0,   0],
       [  0,   0,   0, 356, 289,   0,   0],
       [  0,   0,   0, 118,  39,   0,   0],
       [  0,   0,   0,  29,   1,   0,   0]], dtype=int64)
```

Optimal parameter value for Classification:

0.0210

Optimal parameter value for Regression

0.41