Introduction to Data Science

Course Project

Report Document

**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:
   1. Code File (Jupyter Notebook): L210000\_Code.ipynb
   2. Report Document (This File): L210000\_Report.pdf
4. Project will not be evaluated if:
   1. You submit python (.py) files
   2. You submit multiple .ipynb files
   3. You submit compressed (.rar or .zip) files
   4. 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 😺

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

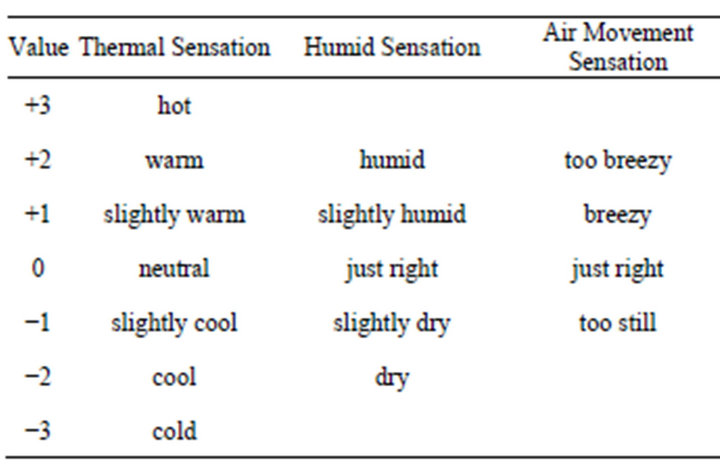
*TA Emails*

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:



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

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

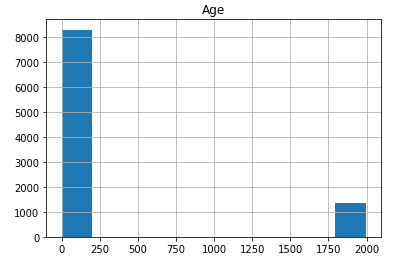
***Part A. Preprocessing***

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

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dim Name | Data Type | Total Instances | Number of Nulls | Number of Outliers | Min. Value | Max Value | Mode | Mean | Median | Variance | STD |
| Age | Float64 | 12566 | 2916 | 1359 | 0.0 | 1996.0 | 24.0 | 308.637 | 35.00 | 462556.556 | 680.115 |
| Clo | Float64 | 12566 | 1406 | 373 | 0.15 | 2.130 | 0.77 | 0.778 | 0.7517 | 0.04928 | 0.22199 |
| Met | Float64 | 12566 | 1887 | 1732 | 0.10 | 4.50 | 1.0 | 1.066 | 1.100 | 0.184022 | 0.42897 |
| Dewpt | Float64 | 12566 | 3552 | 0 | - 1.95 | 26.896 | 17.4 | 13.621 | 14.100 | 34.8459 | 5.90304 |
| PlaneRadTemp | Float64 | 12566 | 7022 | 452 | -7.42 | 11.70 | 0.3 | 0.217 | 0.200 | 1.08402 | 1.04116 |
| Ta | Float64 | 12566 | 20 | 540 | 15.96 | 31.0 | 23.2 | 23.178 | 23.136 | 2.054606 | 1.43339 |
| Tmrt | Float64 | 12566 | 3701 | 44 | 16.61 | 37.4450 | 22.5 | 23.450 | 23.358 | 2.258867 | 1.502953 |
| Vel | Float64 | 12566 | 3700 | 309 | 0.0 | 1.880 | 0.1 | 0.1124 | 0.100 | 0.006248 | 0.079041 |
| AirTurb | Float64 | 12566 | 5601 | 2 | 0.0 | 102.450 | 0.5 | 18.265 | 0.500 | 627.057129 | 25.041109 |
| Pa | Float64 | 12566 | 4656 | 1352 | 0.0 | 27.70 | 2.1 | 5.123 | 1.5506 | 66.522562 | 8.156136 |
| Rh | Float64 | 12566 | 35 | 0 | 7.4 | 79.30 | 64.0 | 42.529 | 43.2800 | 226.835983 | 15.06107 |
| TaOutdoor | Float64 | 12566 | 1368 | 124 | -24.9 | 32.350 | 27.55 | 17.174 | 18.200 | 113.7437 | 10.665071 |
| RhOutdoor | Float64 | 12566 | 19 | 1349 | 0.0 | 100.350 | 0.0 | 61.1003 | 68.7957 | 610.282477 | 24.703896 |
| AMV | Float64 | 12566 | 55 | 0 | -3.0 | 3.00 | 0.0 | 0.100735 | 0.00 | 1.214621 | 1.102099 |
| PMV | Float64 | 12566 | 696 | 259 | -4.17 | 2.500 | 0.1 | -0.0736 | -0.0300 | 0.289461 | 0.538016 |

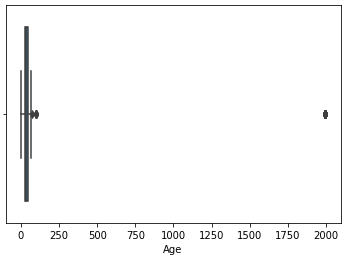
**2. For each of the input dimension, plot histogram and comment the type of distribution the dimension exhibits. Further, visualize each dimension using a Box Plot. Specifically, for each of the input dimension, you’re required to fill the following table (duplicate it for each of the 15 dimensions).**

**Age histogram:**



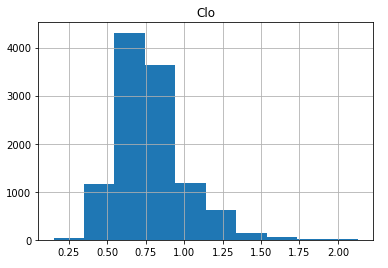
**The above histogram show distribution with an outlier**. While most of the circumference values fall between 0 and 200, some value falls within 1760 to 1795. The value that falls far away from all the other data in the data set is an outlier. Without considering the outliers it is symmetric distribution.

**Age Box Plot:**



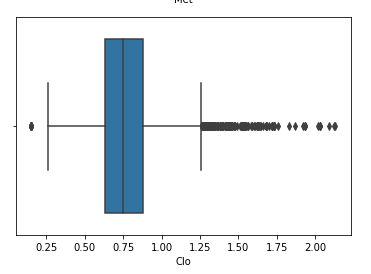
The box plot shape is symmetric as median is in middle of box and whiskers are about same on both sides of box and there exist outliers above the range of upper whisker.

**Clo histogram:**



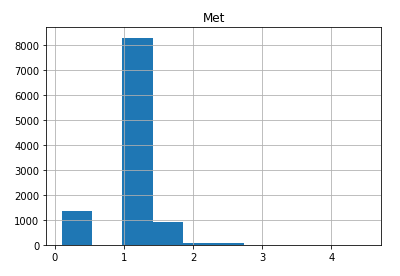
Here we have a unimodal distribution which is skewed right – the right tail of the distribution is longer than the left. Low values are more common in a skewed right distribution.

**Clo Box Plot:**



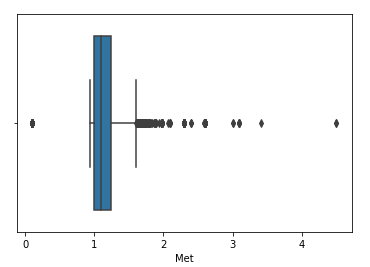
The box plot shape is symmetric as median is in middle of box and whiskers are about same on both sides of box and there exist outliers above the range of upper whisker and also below the range of lower whisker.

**Met histogram:**



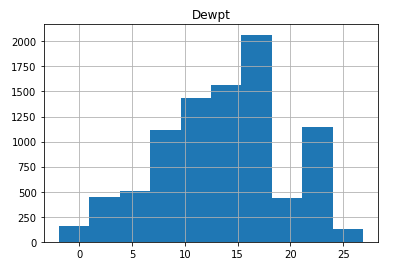
Above histogram is right skewed - the right tail of the distribution is longer than the left. Low values are more common in a skewed right distribution. **This histogram also display outlier**. While most of the circumference values fall between 1 and 2.7, some value falls within almost 0.1 to 0.4. The value that falls far away from all the other data in the data set is an outlier.

**Met Box Plot:**



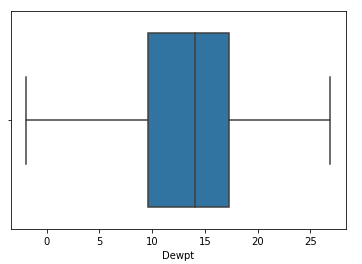
Above Box plot shows positively skewed distribution (skewed right) as the median is closer to the left of the box, and the whisker is shorter on the left end of the box. Outliers exist above the upper whisker and below the lower whisker.

**Dewpt histogram:**



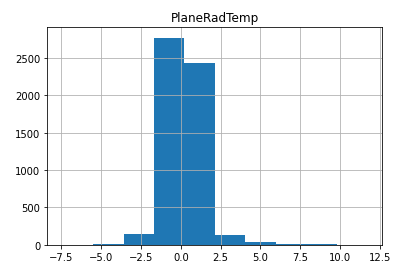
Here we have a distribution which is skewed left – the left tail of the distribution is longer than the right. High values are more common in a skewed left distribution.

**Dewpt Box Plot:**



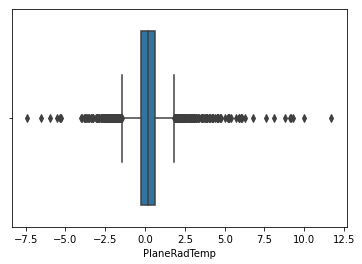
The distribution is negatively skewed (left skewed) in the box plot shown above as the median is closer to the right of the box and the whisker is shorter on the right end of the box. There are no outliers in this feature.

**PlaneRadTemp histogram:**



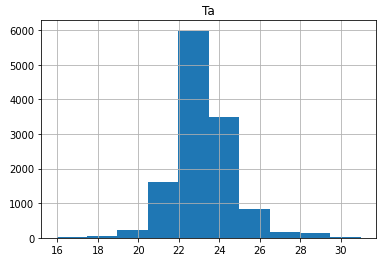
Histogram shows positively skewed distribution - the right tail of the distribution is longer than the left. Low values are more common in a skewed right distribution.

**PlaneRadTemp Box Plot:**



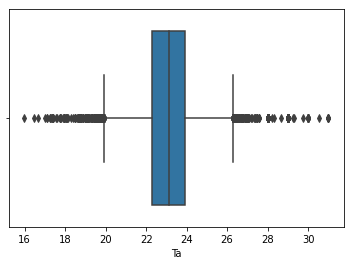
The box plot shape is symmetric as median is in middle of box and whiskers are about same on both sides of box and there exist outliers above the range of upper whisker and also below the range of lower whisker.

**Ta histogram:**



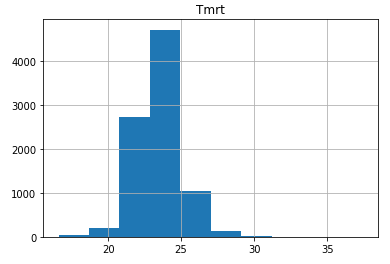
Here we have right skewed distribution - the right tail of the distribution is longer than the left. Low values are more common in a skewed right distribution.

**Ta Box Plot:**



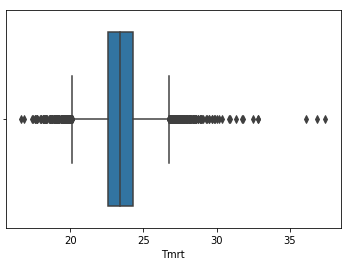
The box plot shape is symmetric as median is in middle of box and whiskers are about same on both sides of box and there exist outliers above the range of upper whisker and also below the range of lower whisker.

**Tmrt histogram:**



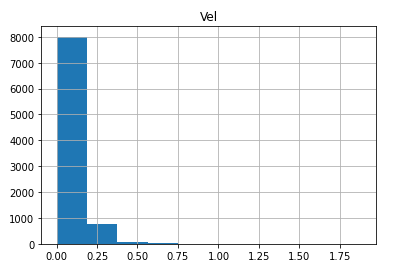
The above histogram shows that majority of the values are situated close to mean. If we draw a line curve around the values, it would look like a bell. We call this a bell curve for normal probability distribution.

**Tmrt Box Plot:**



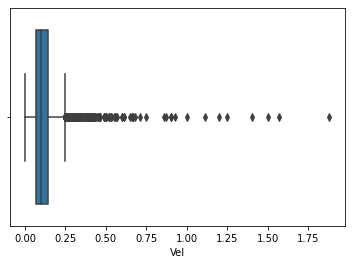
The box plot shape is symmetric as median is in middle of box and whiskers are about same on both sides of box and there exist outliers above the range of upper whisker and also below the range of lower whisker.

**Vel histogram:**



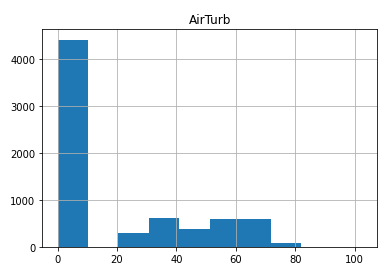
Above histogram shows right skewed nature in distribution as the right tail of the distribution is longer than the left. Low values are more common in a skewed right distribution.

**Vel Box Plot:**



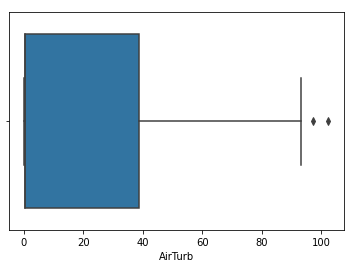
Above Box plot shows positively skewed distribution (skewed right) as the median is closer to the lower whisker of the box, and the whisker is shorter on the left of the box. Outliers exist above the upper whisker and below the lower whisker.

**AirTurb histogram:**



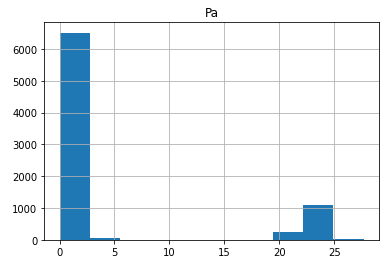
Positively skewed distribution of histogram the right tail of the distribution is longer than the left. Low values are more common in a skewed right distribution.

**AirTurb Box Plot:**



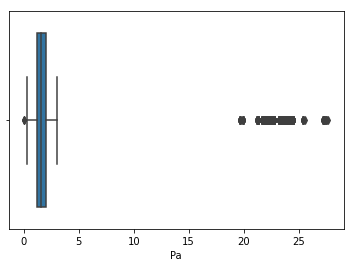
The Box plot shows positively skewed distribution (skewed right) as the median is very close to the lower whisker of the box, and the whisker is too much short on the left of the box. Outliers exist above the upper whisker and below the lower whisker.

**Pa histogram:**



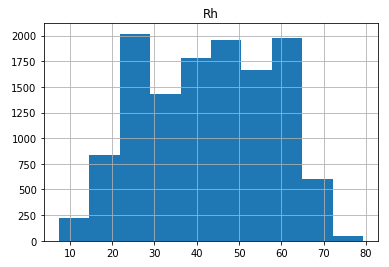
**The above histogram show right skewed distribution with an outlier.** While most of the circumference values fall between 0 and 6, some value falls within 19 to 28. The value that falls far away from all the other data in the data set is an outlier.

**Pa Box Plot:**



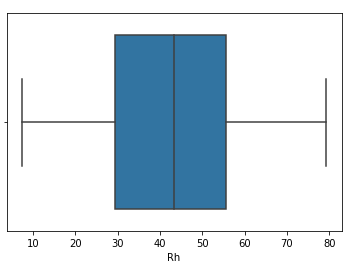
The box plot shape is symmetric as median is in middle of box and whiskers are about same on both sides of box and there exist outliers above the range of upper whisker and also below the range of lower whisker.

**Rh histogram:**



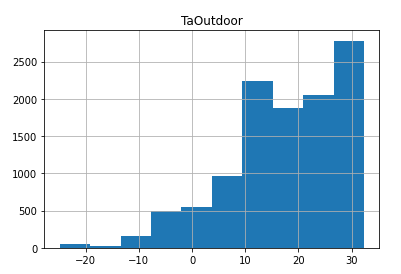
Random distribution is shown through above histogram as it lacks an apparent pattern and has several peaks.

**Rh Box Plot:**



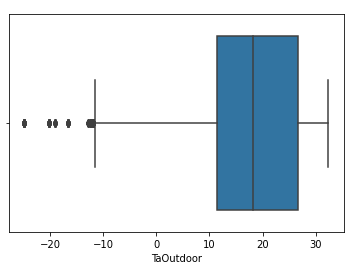
The distribution is negatively skewed (left skewed) in the box plot shown above as the median is closer to the right of the box and the whisker is shorter on the right end of the box. There are no outliers in this feature.

**TaOutdoor histogram:**



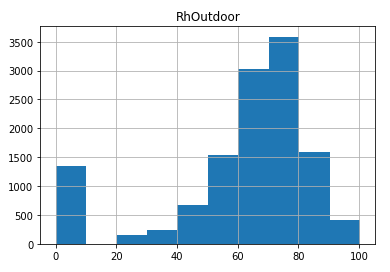
Here we have a distribution which is skewed left – the left tail of the distribution is longer than the right. High values are more common in a skewed left distribution.

**TaOutdoor Box Plot:**



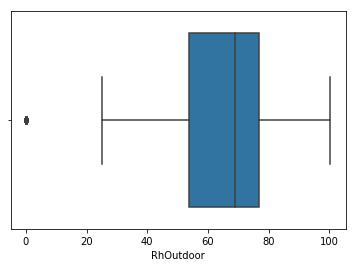
The distribution is negatively skewed (left skewed) in the box plot shown above as the whisker is shorter on the right end of the box. There are outliers below the range of lower whisker.

**RhOutdoor histogram:**



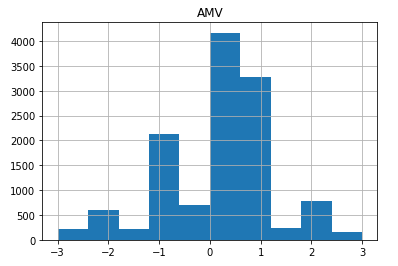
Left skewed distribution histogram - the left tail of the distribution is longer than the right. High values are more common in a skewed left distribution. **This histogram also show distribution with an outlier.**While most of the circumference values fall between 20 and 100, some value falls within 0 to 9. The value that falls far away from all the other data in the data set is an outlier. Without considering the outlier it is symmetric distribution.

**RhOutdoor BoxPlot:**



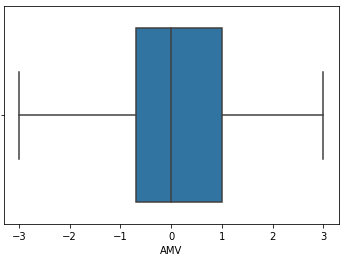
The distribution is negatively skewed (left skewed) in the box plot shown above as the median is closer to the right of the box and the whisker is shorter on the right end of the box. There is outlier below the range of lower quartile.

**AMV histogram:**



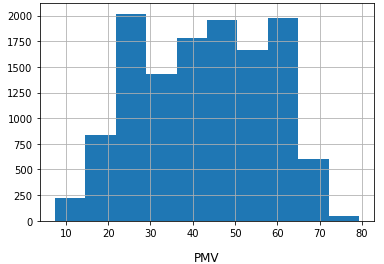
Normal distribution Histogram – If we draw a line curve around the values, it would look like a bell. We call this a bell curve for normal probability distribution.

**AMV Box Plot:**



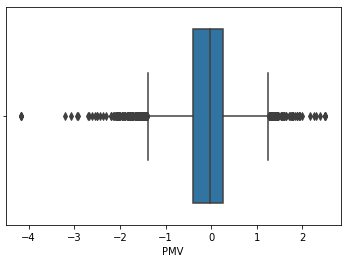
The box plot shape is symmetric as median is in middle of box and whiskers are about same on both sides of box and there exist outliers above the range of upper whisker

**PMV histogram:**



Random distribution is shown through above histogram as it lacks an apparent pattern and has several peaks.

**PMV Box Plot:**



The distribution is negatively skewed (left skewed) in the box plot shown above as the median is closer to the right of the box and the whisker is shorter on the right end of the box. There are outliers above and below the range of upper whisker and lower whisker respectively.

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

|  |  |  |  |
| --- | --- | --- | --- |
| Dim Name | Number of Missing Values | Filled using OR Dropped | Reason for selecting a certain approach |
| Age | 2916 | Filled using median | Median is not affected heavily by outliers. |
| Clo | 1406 | Filled using median | Median is not affected heavily by outliers. |
| Met | 1887 | Filled using median | Median is not affected heavily by outliers. |
| Dewpt | 3552 | Filled using mean | As there are no outliers we prefer to use Mean. |
| PlaneRadTemp | 7022 | Filled using median | Median is not affected heavily by outliers. |
| Ta | 20 | Filled using median | Median is not affected heavily by outliers. |
| Tmrt | 3701 | Filled using median | Median is not affected heavily by outliers. |
| Vel | 3700 | Filled using median | Median is not affected heavily by outliers. |
| AirTurb | 5601 | Filled using median | Median is not affected heavily by outliers. |
| Pa | 4656 | Filled using median | Median is not affected heavily by outliers. |
| Rh | 35 | Filled using mean | As there are no outliers we prefer to use Mean. |
| TaOutdoor | 1368 | Filled using median | Median is not affected heavily by outliers. |
| RhOutdoor | 19 | Filled using median | Median is not affected heavily by outliers. |
| AMV | 55 | Filled using mean | As there are no outliers we prefer to use Mean. |
| PMV | 696 | Filled using median | Median is not affected heavily by outliers. |

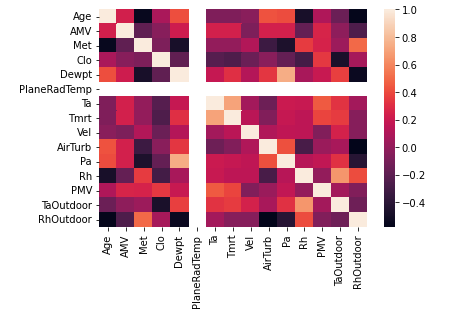
**4. For each of the dimension, find out the outliers (noisy data) and handle these appropriately.**

|  |  |  |  |
| --- | --- | --- | --- |
| Dim Name | Number of Outliers | Smooth using/ Dropped | Reason for selecting a certain approach |
| Age | 1359 | Smooth using IQR | Data is skewed and have outliers so IQR is used and is not affected by extreme outliers. |
| Clo | 373 | Smooth using IQR | Data is skewed and have outliers so IQR is used and is not affected by extreme outliers. |
| Met | 1732 | Smooth using IQR | Data is skewed and have outliers so IQR is used and is not affected by extreme outliers. |
| Dewpt | 0 | Smooth using IQR | As there is no outliers then all data points lies in range of upper limit and lower limit. |
| PlaneRadTemp | 452 | Smooth using IQR | Data is skewed and have outliers so IQR is used and is not affected by extreme outliers. |
| Ta | 540 | Smooth using IQR | Data is skewed and have outliers so IQR is used and is not affected by extreme outliers. |
| Tmrt | 44 | Smooth using IQR | Data is skewed and have outliers so IQR is used and is not affected by extreme outliers. |
| Vel | 309 | Smooth using IQR | Data is skewed and have outliers so IQR is used and is not affected by extreme outliers. |
| AirTurb | 2 | Smooth using IQR | Data is skewed and have outliers so IQR is used and is not affected by extreme outliers. |
| Pa | 1352 | Smooth using IQR | Data is skewed and have outliers so IQR is used and is not affected by extreme outliers. |
| Rh | 0 | Smooth using IQR | As there is no outliers then all data points lies in range of upper limit and lower limit. |
| TaOutdoor | 124 | Smooth using IQR | Data is skewed and have outliers so IQR is used and is not affected by extreme outliers. |
| RhOutdoor | 1349 | Smooth using IQR | Data is skewed and have outliers so IQR is used and is not affected by extreme outliers. |
| AMV | 0 | Smooth using IQR | As there is no outliers then all data points lies in range of upper limit and lower limit. |
| PMV | 259 | Smooth using IQR | Data is skewed and have outliers so IQR is used and is not affected by extreme outliers. |

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

|  |  |  |
| --- | --- | --- |
| Dim Name | Variance | Apply filter or no, reason |
| Age | 462556.556 | No, Variance is too large. |
| Clo | 0.04928 | Yes, Variance is too small |
| Met | 0.184022 | Yes, Variance is too small |
| Dewpt | 34.8459 | No, Acceptable Variance |
| PlaneRadTemp | 1.08402 | Yes, Variance is too small as compare to remaining features. |
| Ta | 2.054606 | No, Acceptable Variance |
| Tmrt | 2.258867 | No, Acceptable Variance |
| Vel | 0.006248 | Yes, Variance is too small |
| AirTurb | 627.057129 | No, Variance is too large. |
| Pa | 66.522562 | No, Variance is too large. |
| Rh | 226.835983 | No, Variance is too large. |
| TaOutdoor | 113.7437 | No, Variance is too large. |
| RhOutdoor | 610.282477 | No, Variance is too large. |

**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)**

Age, Clo, Met, Dewpt, Ta, Rh, Tmrt, Vel, TaOutdoor, RhOutdoor are the most informative features.

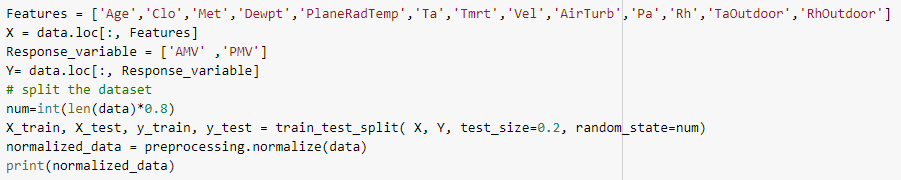
PlaneRadTemp, Pa, AirTurb are least informative features.

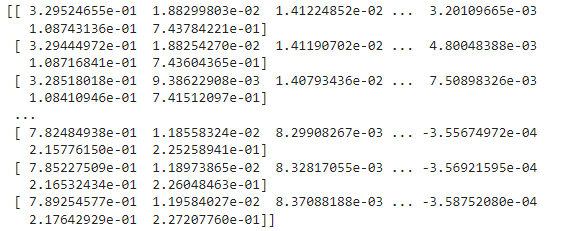
**7. Apply entropy followed by information gain on the selected columns. Specify your selection criteria.**

|  |  |  |  |
| --- | --- | --- | --- |
| Dim Name | Entropy | Info Gain | Reason |
| Age | 2.888 | -1.8888 | Not Selected |
| Clo | 4.9106 | -3.9106 | Not Selected |
| Met | 3.2142 | -2.2142 | Not Selected |
| Dewpt | 5.1644 | -4.164 | Not Selected |
| PlaneRadTemp | 0.0 | 1.0 | Select, Info Gain is higher than other i.e. best gain to define threshold. |
| Ta | 5.6105 | -4.6105 | Not Selected |
| Tmrt | 4.9337 | -3.9337 | Not Selected |
| Vel | 3.2738 | -2.2738 | Not Selected |
| AirTurb | 2.8905 | -1.8905 | Not Selected |
| Pa | 3.94523 | -2.9452 | Not Selected |
| Rh | 7.5408 | -6.54086 | Not Selected |
| TaOutdoor | 5.2277 | -4.2277 | Not Selected |
| RhOutdoor | 4.9908 | -3.99087 | Not Selected |
| AMV | 2.4305 | -1.4305 | Not Selected |
| PMV | 5.1143 | -4.11439 | Not Selected |

***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.**

|  |  |
| --- | --- |
| Feature Vector | Performance achieved |
| 'feature\_names': ('Ta',) | 'avg\_score': 0.2014807624630024 |
| 'feature\_names': ('Clo', 'Ta') | 'avg\_score': 0.4089589811346319 |
| 'feature\_names': ('Clo', 'Met', 'Ta') | 'avg\_score': 0.4950189073667701 |
| 'feature\_names': ('Clo', 'Met', 'Dewpt', 'Ta') | 'avg\_score': 0.6519397705470433 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta') | 'avg\_score': 0.7036034316233191 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Vel') | 'avg\_score': 0.7194322673181945 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Vel', 'RhOutdoor') | 'avg\_score': 0.7355021623276563 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Vel', 'Pa', 'RhOutdoor') | 'avg\_score': 0.7448780777852659 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Tmrt', 'Vel', 'Pa', 'RhOutdoor') | 'avg\_score': 0.7539356641661359 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'RhOutdoor') | 'avg\_score': 0.755292607681261 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'Rh', 'RhOutdoor' | 'avg\_score': 0.7563000768688709 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'Rh', 'TaOutdoor', 'RhOutdoor') | 'avg\_score': 0.7622755085426615 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'Rh', 'TaOutdoor', 'RhOutdoor') | 'avg\_score': 0.7622755085426614 |

As the avg\_score tells the accuracy of combinations of feature vector. Avg\_score will become same after combination of 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'Rh', 'TaOutdoor', 'RhOutdoor'), so we conclude it as best combination.

**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 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'Rh', 'TaOutdoor', 'RhOutdoor') | 'avg\_score': 0.7622755085426614 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'Rh', 'TaOutdoor', 'RhOutdoor') | 'avg\_score': 0.7622755085426615 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Tmrt', 'Vel', 'Pa', 'Rh', 'TaOutdoor', 'RhOutdoor') | 'avg\_score': 0.7617963984431413 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Tmrt', 'Vel', 'Pa', 'Rh', 'TaOutdoor') | 'avg\_score': 0.7588267819698067 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Vel', 'Pa', 'Rh', 'TaOutdoor') | 'avg\_score': 0.7523027038028517 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Vel', 'Rh', 'TaOutdoor') | 'avg\_score': 0.7376569884772619 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Vel', 'Rh')}, 6: {'feature\_idx': (0, 1, 2, 3, 5, 7) | 'avg\_score': 0.7231957717735322 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Vel')}, | 'avg\_score': 0.7194322673181945 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta') | 'avg\_score': 0.7036034316233191 |
| 'feature\_names': ('Clo', 'Met', 'Dewpt', 'Ta') | 'avg\_score': 0.6519397705470433 |
| 'feature\_names': ('Clo', 'Met', 'Ta') | 'avg\_score': 0.4950189073667701 |
| 'feature\_names': ('Clo', 'Ta') | 'avg\_score': 0.4089589811346319 |
| 'feature\_names': ('Ta',) | 'avg\_score': 0.2014807624630024 |

As the avg\_score tells the accuracy of combinations of feature vector. Avg\_score will become same after combination of 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'Rh', 'TaOutdoor', 'RhOutdoor'), so we conclude it as best combination.

**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 |
| 'feature\_names': ('Ta',) | 'avg\_score': 0.024898574098242254 |
| 'feature\_names': ('Met', 'Ta') | 'avg\_score': 0.03851598987811011 |
| 'feature\_names': ('Age', 'Met', 'Ta') | 'avg\_score': 0.0438685001268444 |
| 'feature\_names': ('Age', 'Met', 'Dewpt', 'Ta') | 'avg\_score': 0.04630861509317907 |
| 'feature\_names': ('Age', 'Met', 'Dewpt', 'Ta', 'Vel') | 'avg\_score': 0.04371107335482283 |
| 'feature\_names': ('Age', 'Met', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Vel') | 'avg\_score': 0.04725317572530863 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Vel') | 'avg\_score': 0.043632359968811985 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Vel', 'TaOutdoor') | 'avg\_score': 0.04418335367088755 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Vel', 'AirTurb', 'TaOutdoor') | 'avg\_score': 0.03654815522784027 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'TaOutdoor') | 'avg\_score': 0.04851258990148144 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'TaOutdoor') | 'avg\_score': 0.049142296989567846 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'Rh', 'TaOutdoor' | 'avg\_score': 0.04898487021754627 |
| 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'Rh', 'TaOutdoor', 'RhOutdoor') | 'avg\_score': 0.03646944184182943 |

As the avg\_score of combination 'feature\_names': ('Age', 'Clo', 'Met', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'TaOutdoor'), is higher than others so we consider it as best combination.

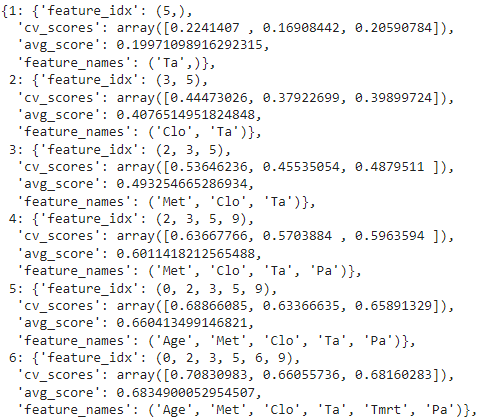
**3B. Apply backward selection, considering AMV as response variable and Logistic regression as machine learning model. Create a table, that mentions dimensions, and performance achieved. Which is the optimal feature set, and why.**

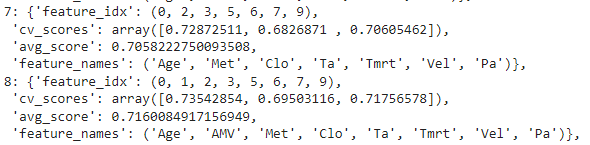
|  |  |
| --- | --- |
| Feature Vector | Performance achieved |
| ‘feature\_names: ('Age', 'Clo', 'Met', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'Rh', 'TaOutdoor', 'RhOutdoor') | 'avg\_score': 0.03646944184182943 |
| ‘feature\_names: ('Age', 'Clo', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'AirTurb', 'Pa', 'Rh', 'TaOutdoor', 'RhOutdoor') | ‘avg\_score': 0.049535863919621836 |
| 'feature\_names': ('Age', 'Clo', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'Pa', 'Rh', 'TaOutdoor', 'RhOutdoor') | 'avg\_score': 0.05158241195590252 |
| 'feature\_names': ('Age', 'Clo', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'Pa', 'Rh', 'RhOutdoor') | 'avg\_score': 0.051818552113934935 |
| 'feature\_names': ('Age', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'Pa', 'Rh', 'RhOutdoor') | 'avg\_score': 0.04764674265536262 |
| 'feature\_names': ('Age', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Pa', 'Rh', 'RhOutdoor') | 'avg\_score': 0.04308136626673642 |
| 'feature\_names': ('Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Pa', 'Rh', 'RhOutdoor') | 'avg\_score': 0.05150369856989179 |
| 'feature\_names': ('Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Pa', 'Rh') | 'avg\_score': 0.04748931588334104 |
| 'feature\_names': ('Dewpt', 'Ta', 'Tmrt', 'Pa', 'Rh') | 'avg\_score': 0.044813060758973955 |
| 'feature\_names': ('Dewpt', 'Tmrt', 'Pa', 'Rh') | 'avg\_score': 0.04040511114236933 |
| 'feature\_names': ('Dewpt', 'Tmrt', 'Rh') | 'avg\_score': 0.03717786231592657 |
| 'feature\_names': ('Dewpt', 'Rh') | 'avg\_score': 0.03402932687549476 |
| 'feature\_names': ('Dewpt',) | 'avg\_score': -0.03067307642538064 |

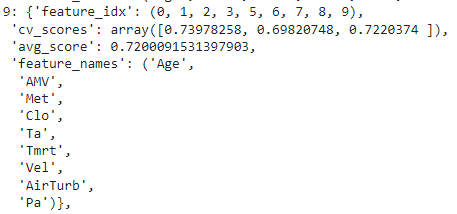
As the avg\_score of feature\_names: ('Age', 'Clo', 'Dewpt', 'PlaneRadTemp', 'Ta', 'Tmrt', 'Vel', 'Pa', 'Rh', 'TaOutdoor', 'RhOutdoor') is maximum so we consider it as best.

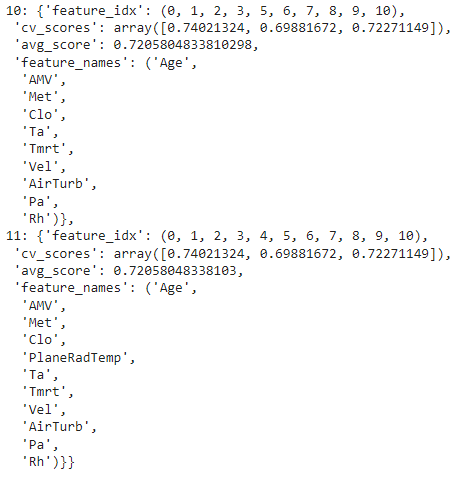
**4. Using the optimal feature vector that you’ve figured out from your analysis above, apply 3-fold cross validation for both regression and classification problems (PMV and AMV respectively). Write down the optimal parameters values for each of the model. Further, plot confusion matrix for the classification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

**LinearRegression:**

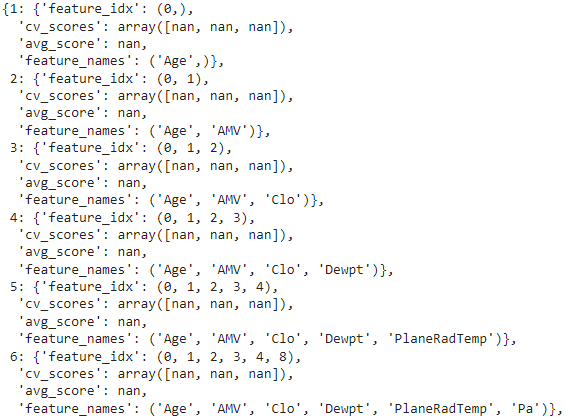


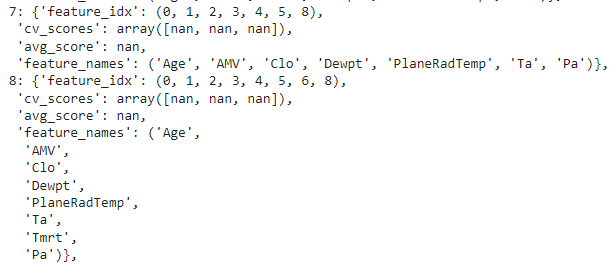


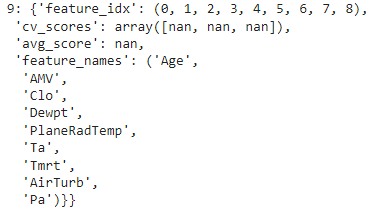




**Logistic Regression:**







**Confusion Matrix:**

