Descriptive Analysis – AQI Data

Table of Contents

S.no	Title	Page no.
1.	Data Screening/Cleaning	2
2.	Descriptive analysis – box plot	9
3.	Descriptive analysis – scatter plot	13
4.	Descriptive analysis – Histogram	17
5.	Systematic sampling method	20
6.	AQI Category comparison	22
7.	AQI Correlation	23
8.	Interpretation and Discussion	24
9.	Pros and Cons of descriptive techniques	25

Descriptive Analysis – AQI Data

1) Data Screening and Cleaning tasks:

a) Dataset Description:

Six datasets, each representing a year from 2016 to 2021, were used for performing the various tasks as part of this assignment. These datasets represent the hourly captured PM2.5 concentration in the city of Addis Ababa central. The datasets were first uploaded onto the SAS Platform and then imported into a custom-made library named 'custom' using the PROC Import statement. The details of each dataset after importing are as follows:

S.no	Dataset Name	no. of Rows	no. of columns
1.	aac_2016	8784	14
2.	aac_2017	8760	14
3.	aac_2018	8111	14
4.	aac_2019	6920	14
5.	aac_2020	8369	14
6.	aac_2021	5347	14

b) Variable description:

- Site: This variable represents the location where the profiling of PM2.5 concentration has been done. It is a categorical variable as the response is nonnumeric and nominal because countries can't be ranked in a particular order.
- Parameter: The Parameter variable helps in identify the type of pollutant the dataset is statistically speaking about. It is a categorical variable as the response is non-numeric and nominal because the parameters can't be ordered.
- DateLT: This variable gives the exact timestamp at which the pollutant level has been measured. It is a date time variable.
- Year: This variable tells us about the exact year during which the levels of pollution were recorded. It is a numeric variable of ordinal type as the numeric responses can be ranked in an order.

- Month: The Month variable tells us about the exact month during which the levels of pollution were recorded. It is a numeric variable of ordinal type as they can be ranked in an order.
- Day: This variable lists out the day on which the pollution levels are recorded. It is a numeric variable of ordinal as they can be logically ordered.
- Hour: This variable specifies the exact hour during which the recording has been made. It is a numeric variable of ordinal type as they can be ordered meaningfully.
- NowCast Concentration: This variable provides the concentration of NowCast that is measured at a given time. It is a numeric variable as the response is numeric in nature and can be further classified as ratio as it can be zero but never less that it.
- AQI: This variable gives the reading of AQI levels measured at a given time. It is a numeric variable as the response is numeric and can be further classified as ratio as it can be zero but never below that.
- AQI Category: This variable maps every AQI level to a particular category and hence it is a categorical variable. It can be further classified as ordinal as the categories can be ranked in a particular order.
- Raw Concentration: It gives the raw concentration levels measured at a given time. It is a numeric variable as the response is numeric and of ratio type as it can be zero but never below it.
- Concentration Unit: This variable gives the units in which the concentration is measured. It is a categorical variable of nominal type as there can be no specific order to rank the units.
- Duration: This variable gives the time that is taken to record one reading of PM2.5 by the device. It is a numeric ordinal variable as the response is numeric and can be ranked.
- QC_Name: This variable provides the status, such as valid or missing, of the captured recordings by the device. It is categorical ordinal as the response is non-numeric which can be ordered.

c) Merging datasets:

All the six datasets that were imported are merged into a master dataset, named as

'aac_master', before performing various screening and cleaning tasks. Four columns were removed from the master dataset which are believed to not compromise the findings of our present study and make the master table smaller and precise. These columns are Site, Parameter, Concentration unit and duration. A Data step with set and drop commands has been used to merge and drop the columns. The total rows and columns after merging are 46,291 and 10 respectively.

```
NOTE: There were 8784 observations read from the data set CUSTOM.AAC_2016.
NOTE: There were 8760 observations read from the data set CUSTOM.AAC_2017.
NOTE: There were 8111 observations read from the data set CUSTOM.AAC_2018.
NOTE: There were 6920 observations read from the data set CUSTOM.AAC_2019.
NOTE: There were 8369 observations read from the data set CUSTOM.AAC_2020.
NOTE: There were 5347 observations read from the data set CUSTOM.AAC_2021.
NOTE: The data set CUSTOM.AAC_MASTER has 46291 observations and 10 variables.
NOTE: DATA statement used (Total process time):
      real time           0.09 seconds
      user cpu time       0.01 seconds
      system cpu time     0.03 seconds
      memory              10668.50k
      OS Memory           40768.00k
      Timestamp           10/07/2021 02:14:03 AM
      Step Count                        40  Switch Count  4
      Page Faults                       0
      Page Reclaims                     2135
      Page Swaps                        0
      Voluntary Context Switches        144
      Involuntary Context Switches      5
      Block Input Operations            11264
      Block Output Operations           9480
```

d) Checking for Missing Data:

The master dataset is completely checked for missing values such as a character space ( ) or a period (.). The SAS tool provides us with a utility using which we can identify if there is any missing data in the dataset. The code returns the frequency of every variable in the dataset which helps in identifying the missing data. Upon running the code, it is identified that there is no missing data as such in the master dataset.

Missing Data Frequencies
Legend: ., A, B, etc = Missing

DateLT	Frequency	Percent
Non-missing	46291	100.00

Year	Frequency	Percent
Non-missing	46291	100.00

Month	Frequency	Percent
Non-missing	46291	100.00

Day	Frequency	Percent
Non-missing	46291	100.00

Hour	Frequency	Percent
Non-missing	46291	100.00

NowCastConc	Frequency	Percent
Non-missing	46291	100.00

AQI	Frequency	Percent
Non-missing	46291	100.00

AQI_Category	Frequency	Percent
Non-missing	46291	100.00

RawConc	Frequency	Percent
Non-missing	46291	100.00

QC_Name	Frequency	Percent
Non-missing	46291	100.00

Although there is no such cell which is physically blank or with a period, the master dataset is checked for yet another kind of missing data which is -999. The -999 are default values filled in the empty cells and it is a random missing data. A code has

been written using PROC SQL method and count command which returns the count of number of records with -999 values. Upon running the code, it is found that there are a total of 6553 missing values in the master dataset.

missing_values
6553

Since 6553 records constitute to approximately 14 percent of the total 46,291 records, this missing data cannot be handled by replacing with mean as it may have a great impact on the resulting trend. Hence the missing data is handled by removing them from the dataset. This is performed by implementing delete command using PROC SQL method.

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         /*code to fix missing values*/
70         PROC SQL;
71         delete from custom.aac_master
72         WHERE NowCastConc eq -999 OR RawConc eq -999 OR AQI eq -999;
NOTE: 6553 rows were deleted from CUSTOM.AAC_MASTER.

73
74         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
84
```

The master dataset ‘aac_master’ has a total of 39,738 rows after deleting the 6553 records.

Note:

Before deleting the observations, the data set has been checked for observations where Nowcast Conc. variable value can be used to calculate and correct the AQI values. But upon checking it is found that no observation holds for such a pair. Nevertheless, a code has been included to calculate the AQI using Nowcast Conc. in the code file.

e) Identifying Out of range values:

The master dataset ‘aac_master’ is checked for accuracy using the PROC Mean method. The PROC Mean method analyzes the complete dataset and provides with statistical results such as mean, standard deviation, minimum and maximum.

Accuracy check for master Data set

Variable	Mean	Std Dev	Minimum	Maximum	N
Year	2018.64	1.5147181	2016.00	2021.00	39738
Month	6.4489154	3.4204965	1.0000000	12.0000000	39738
Day	15.7080125	8.8200435	1.0000000	31.0000000	39738
Hour	11.5066184	6.9332626	0	23.0000000	39738
NowCastConc	24.2300569	14.5781602	0.8000000	266.8000000	39738
AQI	75.5689516	30.0553614	3.0000000	317.0000000	39738
RawConc	24.3282500	18.3705887	-15.0000000	985.0000000	39738

After observing the generated results, it is noted that RawConc. variable has negative values which is practically not possible. Hence a sql query with count command has been implemented using the PROC SQL method which returns the total number of observations whose value is less than 0.

negative_values
7

Upon running the query it is identified that 7 observations are in negative range. These observations have been removed from the master dataset by implementing the delete command using the PROC SQL method.

```
NOTE: 7 rows were deleted from CUSTOM.AAC_MASTER.

73         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time               0.03 seconds
      user cpu time           0.01 seconds
      system cpu time         0.01 seconds
      memory                  6921.78k
      OS Memory               30632.00k
      Timestamp               10/07/2021 05:00:34 AM
      Step Count                        35  Switch Count  3
      Page Faults                       0
      Page Reclaims                     568
      Page Swaps                        0
      Voluntary Context Switches        34
      Involuntary Context Switches      0
      Block Input Operations            9504
      Block Output Operations           2064
```

After deleting the 7 observations from ‘aac_master’, the updated row count is 39,731.

f) Accuracy check for AQI Category:

As part of the accuracy check the ‘aac_master’ dataset is checked for any observation where AQI levels are incorrectly mapped to the AQI category. To perform this a series of sql queries have been written using PROC SQL method where the code returns the observation, if found any, for which the category is labelled falsely. Upon running the code, no observation was selected which proves that all the AQI levels are correctly mapped to categories.

```
NOTE: No rows were selected.
103        QUIT;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.06 seconds
      user cpu time       0.03 seconds
      system cpu time     0.02 seconds
      memory              7557.56k
      OS Memory           31656.00k
      Timestamp           10/07/2021 05:14:51 AM
      Step Count                        47  Switch Count  0
      Page Faults                       0
      Page Reclaims                     1965
      Page Swaps                        0
      Voluntary Context Switches        45
      Involuntary Context Switches      0
      Block Input Operations            6336
      Block Output Operations           8
```

g) Accuracy check for duplicate records:

The master dataset ‘aac_master’ is also checked for duplication of records if any. Since the DateLT variable is a unique variable which cannot have repetitions, the dataset is checked if there are any.

To perform this a query has been written using count command implemented by PROC SQL method which returns the duplicate observations present in the dataset and the count for number of times repeated.

DateLT	Count
12MAR17:03:00:00	2
11MAR18:03:00:00	2
10MAR19:03:00:00	2

Upon running the query it is identified that there are 3 records in total which have duplicate records. Since timestamp is unique and duplication of it creates ambiguity, the records have been deleted using the delete command implemented by PROC SQL method. It is not a randomly missing data as the hour is same for all 3 records.

```
NOTE: 2 rows were deleted from CUSTOM.AAC_MASTER.

73
74          delete from custom.aac_master
75          where Year = 2018 AND Month = 3 AND Day = 11 AND Hour = 3;
NOTE: 2 rows were deleted from CUSTOM.AAC_MASTER.

76
77          delete from custom.aac_master
78          where Year = 2019 AND Month = 3 AND Day = 10 AND Hour = 3;
NOTE: 2 rows were deleted from CUSTOM.AAC_MASTER.

79          quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time               0.07 seconds
      user cpu time           0.04 seconds
      system cpu time         0.01 seconds
      memory                  7026.84k
      OS Memory               35756.00k
      Timestamp               10/07/2021 05:39:24 AM
```

After deleting the 6 observations from 'aac_master' dataset, the updated row count is 39,725.

h) Select values for every 6 hours:

The objective of this step is to select the values for every 6 hours from the master dataset 'aac_master' and form a new modified dataset from it. To perform this, Data step has been used along with Set and Where commands. The where command mentions the logic using which records for every 6 hours are selected.

```
NOTE: There were 6619 observations read from the data set CUSTOM.AAC_MASTER.
      WHERE Hour in (0, 6, 12, 18);
NOTE: The data set CUSTOM.AAC_MASTER has 6619 observations and 10 variables.
NOTE: DATA statement used (Total process time):
      real time               0.05 seconds
      user cpu time           0.00 seconds
      system cpu time         0.01 seconds
      memory                  2881.90k
      OS Memory               31916.00k
      Timestamp               10/07/2021 05:58:36 AM
      Step Count                        146  Switch Count  7
      Page Faults                       0
      Page Reclaims                     510
      Page Swaps                        0
      Voluntary Context Switches        96
      Involuntary Context Switches      0
      Block Input Operations            6336
      Block Output Operations           1544
```

The new modified dataset 'aac_master' has a total of 6619 rows.

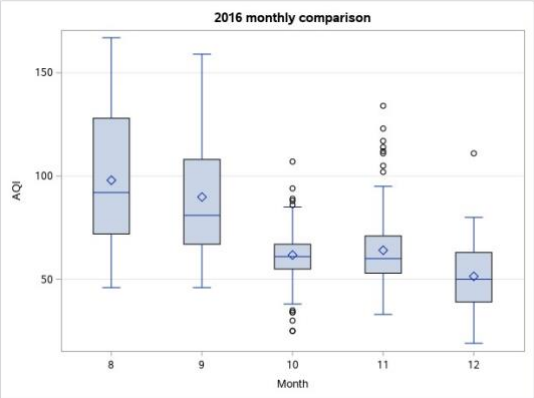
2) Descriptive analysis:

2.1 Monthly box plots to compare air pollution on yearly basis:

The below box plots have been plotted using the PROC SGPLOT method. The element of analysis is AQI, and it is categorized by month. The 'where' condition is used to retrieve the data belonging to the year of interest.

2.1.1 Box plot for year 2016:

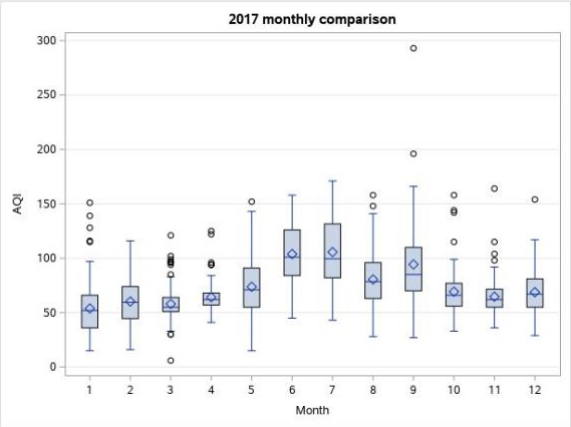
Upon analyzing the graph, it is inferred that majority of the data for the year 2016 is missing as we have boxplots for only 5 months from August 2016 to December 2016. One another observation from the graph is that the mean has been consistently decreasing from August to December. The AQI levels have dropped by approximately 50 percent during these 5 months. There are also a huge number of outliers identified in October and November in relative comparison to the other months. The highest AQI value is observed in the month of august standing around 170 and lowest in the month of December standing around 10.



2.1.2 Box plot for year 2017:

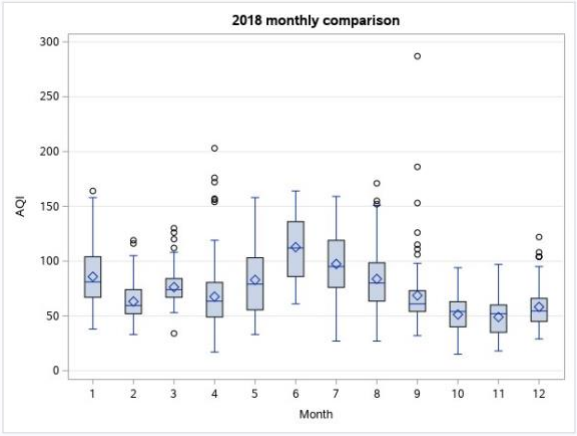
Upon observing the graph, it is identified that the first quadrimester has observed a low record of AQI levels. The average Mean stood around 55 approximately. Although the year started with low levels of AQI, the next 5 months starting from May to September has recorded a significant rise in the levels of AQI when compared to previous months. The average mean during this interval stood around 85. The last quarter of the year has again shown a sharp decline in the AQI levels, with average mean falling as low as to around 65. The highest AQI level is recorded in the month of July with a value of 175 and the lowest in the month of

January with a value of 20. The outliers are detected all the year round and has gone to extremes in the month of march and September.



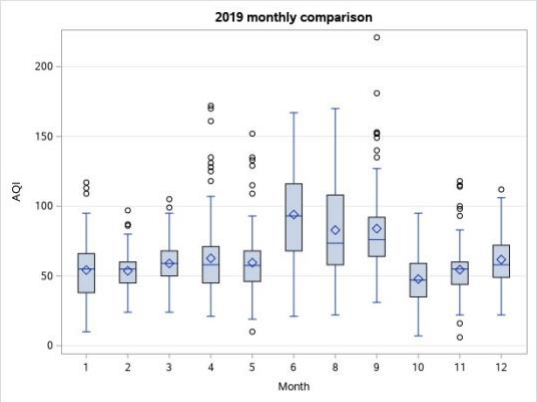
2.1.3 Box plot for year 2018:

On analyzing the graph, it is inferred that the AQI levels maintained a consistent average for the first five months, with the average standing at around 75. The average AQI observed a sudden spike in the month of June with the mean AQI rising by approx. 60% taking the value to 120. The mean again started to drop gradually from June to December with the mean AQI standing at approx. 55 by the end of the year. The highest AQI level is observed in the month of June standing at around 165 and the lowest is recorded in the month of April with AQI at approx. 20. Although April and September recorded a low mean AQI, many outliers with high AQI levels were observed in these months.



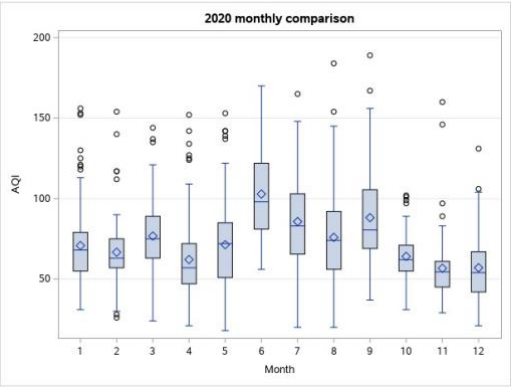
2.1.4 Box plot for year 2019:

The box plots for the year 2019 have maintained a very steady rate for most of the months, except for June to September, with average mean standing at around 55. Upon observing it is identified that the month of July has been missing from the dataset. A significant rise has been observed in the AQI levels from June and the trend continued for 3 moths until September with average mean for AQI standing at around 85. Outliers have been detected round the year with September having noted extreme outlier values. The highest AQI level is observed in the month of August with a value of approx. 175 and lowest in the month of October with value standing around 15.



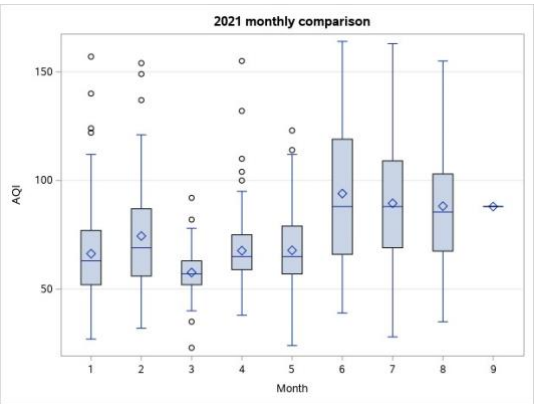
2.1.5 Box plot for year 2020:

The year of 2020 has observed a very random behavior in its distribution of AQI levels from month to month. There is no pattern in the recorded AQI levels. The outliers are distributed all along the scale. The highest AQI level is noted in the month of June with value standing at around 170. The lowest AQI levels are recorded multiple time in this year with the value going to as low as 10. Although the entire distribution is random, the month of June stood out for its abnormally extreme low and high values.

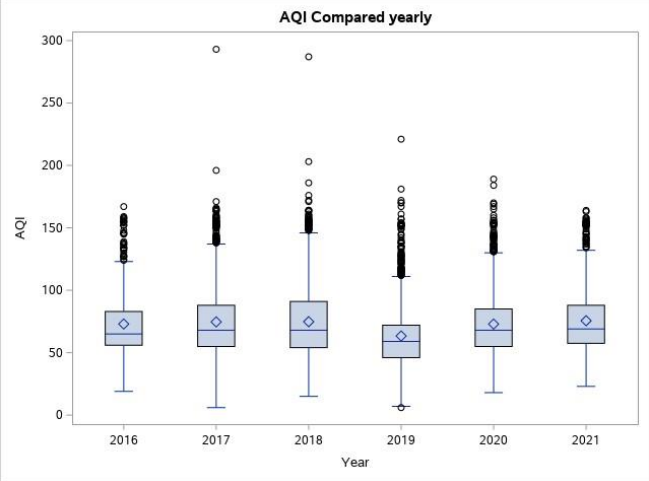


2.1.6 Box plot for year 2021:

The year 2021 has observations for a total of 8 months starting from January to August. The levels of AQI in 2021 started on a mid-note and maintained the same until May. The average AQI levels for the first 5 months stood around 65 but the values have gone up to as much as 120. Starting from June the AQI levels leaped very sharply and maintained these high values until August. The AQI levels have noted record high values in this year consecutively in the months of June and July with the value marked at around 165. The lowest AQI level has been registered in the month of May with an approx. value of 10. Outliers have been detected in the first 5 months but nothing from June as per records.



2.1.7 Box plot comparing years from 2016 to 2021:



The above box plot has been plotted using the PROC SGPLOT method. The element of analysis is AQI, and it is categorized by the variable Year.

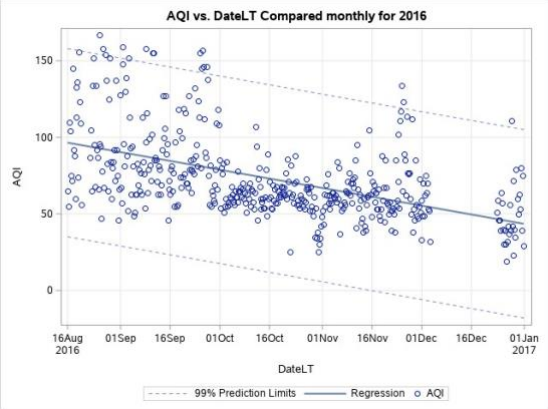
Upon analyzing and comparing the box plots on yearly level, it is very evident that the AQI levels have not been coming down but are maintaining a consistent average from 2016 to 2021. The average mean is standing at around 75 for all the 6 years combined.

The lowest AQI of all the 6 years has been registered in the year 2017 with an approx. value of 10. All the years have observed a high AQI level of approx. 165. The years 2017, 2018 and 2019 have noted 5 extreme outliers as per the records.

2.2 Monthly Scatter plots to compare Air pollution on yearly basis:

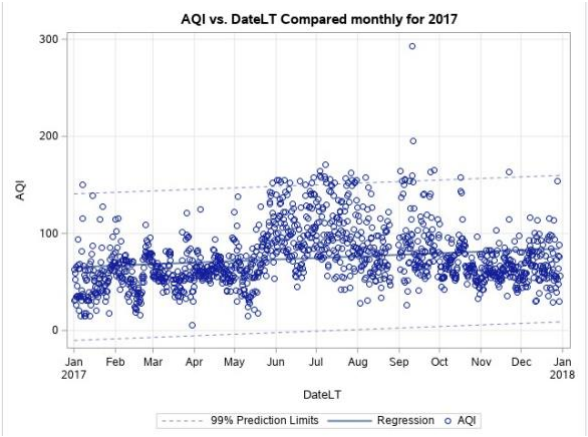
The below Scatter plots have been plotted using the PROC SGPLOT method. The graph has been plotted for AQI vs DateLT, for each year. The 'where' condition is used to retrieve the data belonging to the year of interest.

2.2.1 Scatter plot for year 2016:



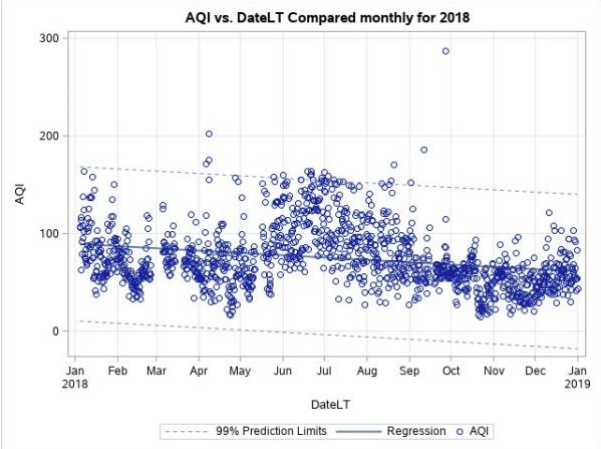
The above scatter plot displays an unpredictable weak trend because of the inconsistent data. The information inferred from this individual graph is not reliable owing to the inconsistency. But a vague negative relation can be observed as the AQI has been declining from August to December. High AQI values can be observed during the month of August and September.

2.2.2 Scatter plot for year 2017:



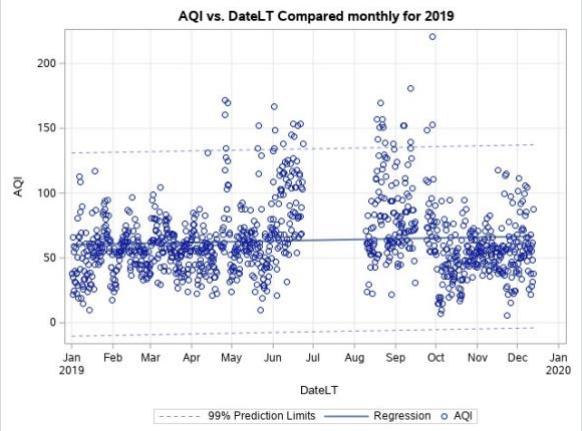
From the graph it can be inferred that there is a slight increase in the average AQI levels for the year 2017 from January to December. The highest AQI levels are observed during the months of June to September. 2 extreme outliers can be detected in the month of September.

2.2.3 Scatter plot for year 2018:



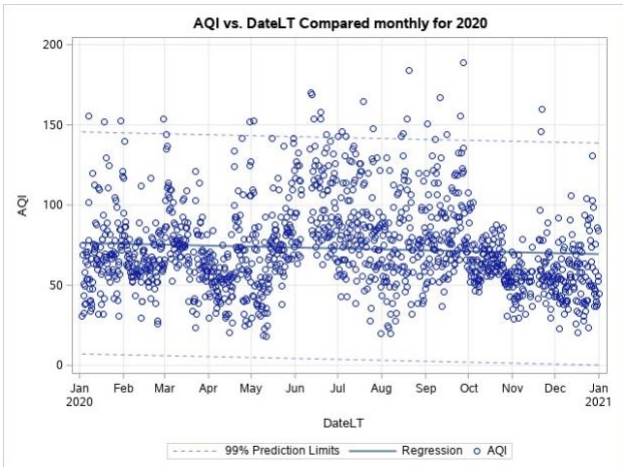
The graph of 2018 observes a surge in the AQI levels during the months of June and July, values going up to around 175. The overall average AQI has been noted to slightly decline showing a weak negative trend from January to December as the prediction limits show. A total of 3 outliers have been detected, 1 in the month of April and 2 in the month of September.

2.2.4 Scatter plot for year 2019:



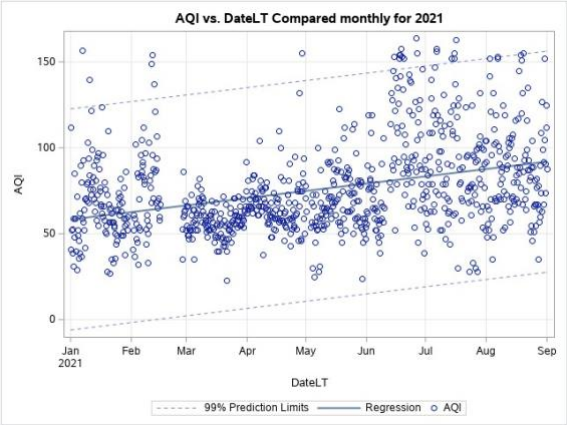
From the scatter plot it can be observed that the data is missing for the month of June. The AQI levels have been consistent throughout the year maintaining the same average from the beginning to the end. A surge can be observed in the months of June and August recording values as high as 155. Around 5 outliers are identified where 3 come from the month of April and 2 in the month of September.

2.2.5 Scatter plot for year 2020:



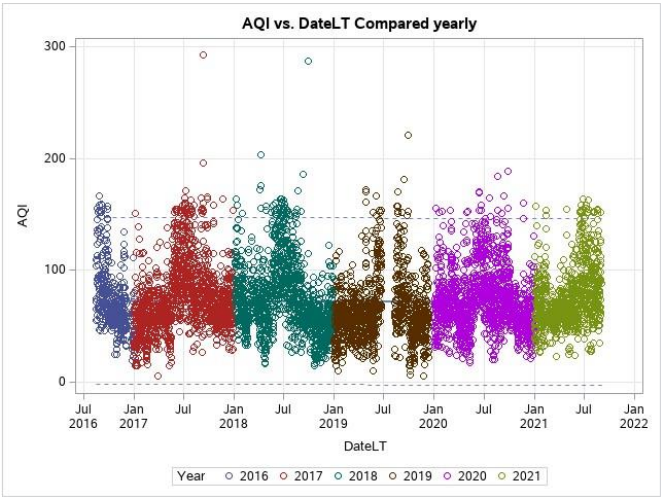
The mean AQI level has been consistent throughout the year. The graph is distributed on a large scale showing a great variation in AQI levels. The highest AQI level has been record in the month of June where the AQI level has gone up to as much as 170. Around 4 outliers have been detected in the graph, with 1 in the month of August,1 in the month of September and 2 in the month of November.

2.2.6 Scatter plot for the year 2021:



The graph represents data for the months from January to September. The overall trend shows a significant rise in the average mean by approx. 45%. The AQI levels recorded shows a wide range of values in the months of January and February. But the values further proceed to concentrate from March to May. From June the AQI levels are again widely scattered on the scale. High AQI levels are registered during the months of January, February, June and July which go up to as much as around 160. Around 6 outliers have been detected. 2 in the month of January, 3 in the month of February and 1 in April.

2.2.7 Scatter plot comparing years from 2016 to 2021:



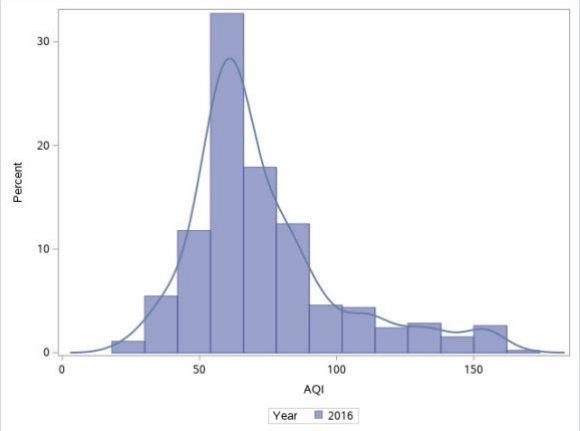
The graph above depicts the AQI levels against the DateLT scale. It was plotted using the PROC SGPLOT method. Upon analyzing the graph, it can be inferred that the mean AQI value has been consistent for the past 6 years with average standing at around 75. All the years have their upper bound touched to almost 175.

There are 5 outliers identified in the graph. 2 outliers in the year 2017, 2 in the year 2018 and 1 in the year 2019. The lowest AQI level has gone to as low as 10 during the year 2017 and 2019.

2.3 Monthly Histograms to compare Air pollution on yearly basis:

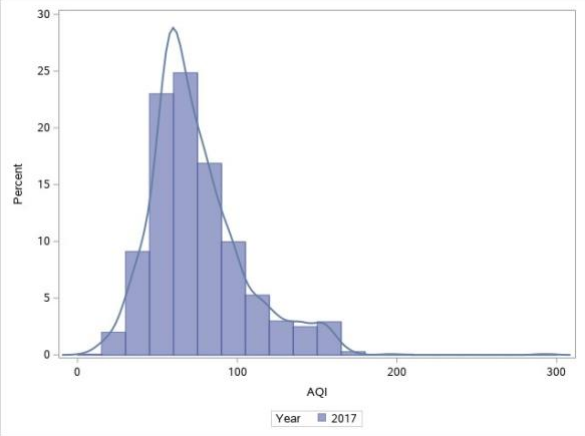
The below histograms have been plotted using the PROC SGPLOT method. The variable of interest is AQI and it is compared monthly for its levels each year. 'where' command has been used to retrieve yearly data from the master dataset.

2.3.1 Histogram for the year 2016:



The graph represents a right-skewed histogram as it has a peak at the left of the centre and the right gradually decreasing from the peak. From the graph we can infer that the AQI level of 60-70 contribute to the majority share, approx. 33%, of all the captured values.

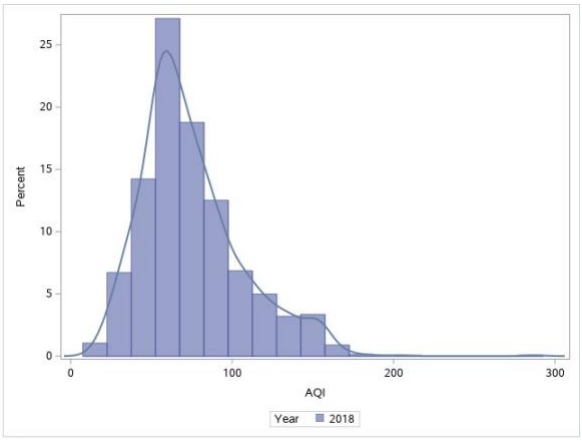
2.3.2 Histogram for the year 2017:



The graph represents the AQI distribution for the year 2017. The structure represents a right-skewed histogram as the peak is at the left of the centre and the

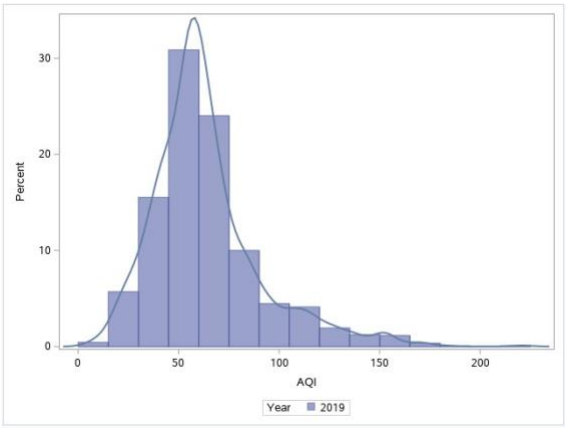
right of the peak gradually decreasing to the end. The 46% of the recorded AQI values during 2017 ranged from 40 to 60.

2.3.3 Histogram for the year 2018:



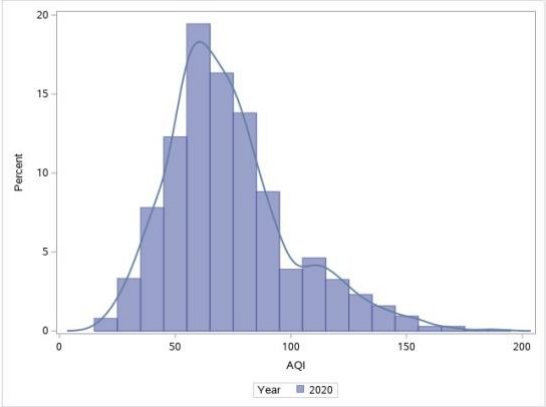
The above histogram is plotted for the year 2018 showing the AQI levels. The structure of the graph defines a right-skewed histogram as the peak is to the left of the centre and the right of the peak is shown to gradually drop to the extreme. Approx. 30% of the values of AQI ranged from 50-70 in year 2018.

2.3.4 Histogram for the year 2019:



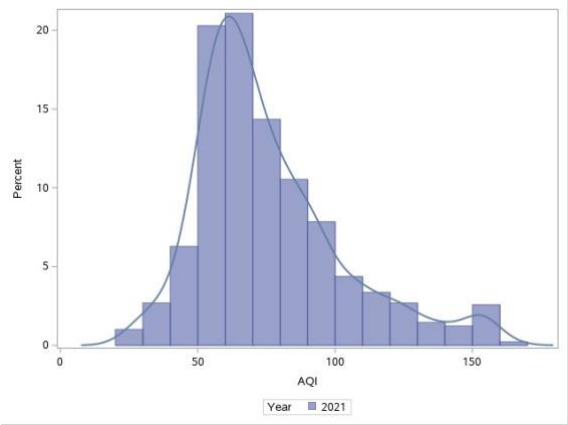
The above histogram is plotted for the year 2019 analyzing the variable AQI. The structure depicts a right-skewed model as the peak is at the centre of the left and the right declining from the centre to the right end. A significant share of around 30% constitutes of AQI values from 50-60.

2.3.5 Histogram for the year 2020:



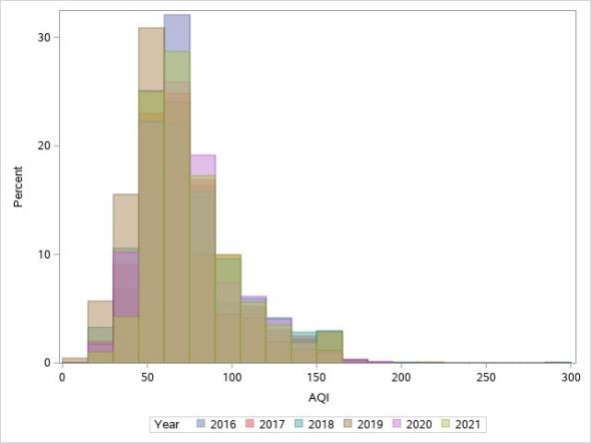
The graph above analyzes the AQI variable for the year 2020. The structure of the graph symbolizes a right-skewed model as the peak is at the centre of the left and the right gradually decreasing until its dead at the end of the right. The graph shows a wide range of AQI levels actively recorded in the year 2020 as approx. 59% of the values ranged from 50-80.

2.3.6 Histogram for the year 2021:



The above histogram shows the analysis of AQI variable for the year 2021. The structure resembles a right skewed model as the peak is at the center of the left and right side gradually dropping to the end. 40% of all the values recorded in the year ranged from 50-70. Approximately 9% of the values are below 50 and the rest 50% is above the 70.

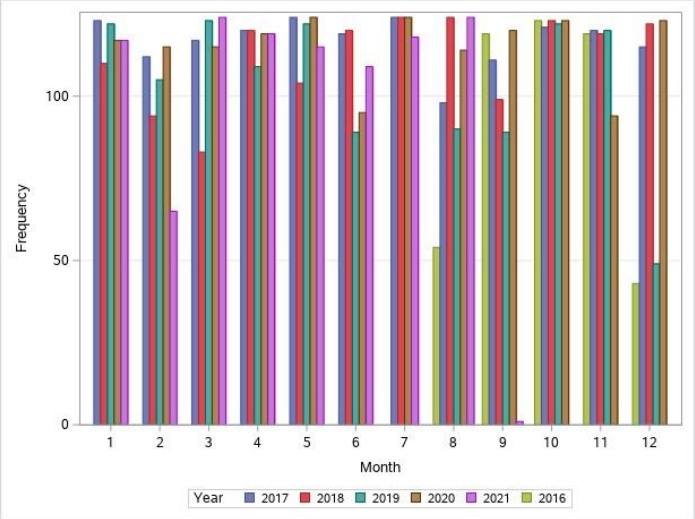
2.3.7 Histogram comparing years from 2016 to 2021:



The above graph has been plotted using the PROC SGPLOT method. The master dataset has been used to analyze the AQI variable which is grouped by year using the group condition. Upon comparing the graphs, it is clearly understood that all years have a very similar structure to their histograms which is right skewed as the peaks are at the center of the left and the right gradually decreasing to the end. As per the plotted graph it is clearly understood that the most common range of AQI levels are varying between 50-70 for all the 6 years. So, it is safe to assume that most of the days the AQI levels in the city is moderate and this has been consistently maintained for almost 6 years. The similarity in trends and lack of major abnormalities allows this dataset to be reliable to safely predict the future trends.

2.4 Systematic Sampling method:

To perform systematic sampling PROC SURVEYSELECT method has been implemented. The sampling is expected to pick 4 records a month from all the years of the dataset. Since we have 6 years as part of the master dataset ‘aac_master’, we know there are a total of 6*12 = 72 months. But it’s evident from the above analysis that the data is consistent and has many missing values. To help better understand the frequency of each month every year, a bar chart has been plotted for months grouping them by year using the PROC SGPLOT method.



Upon analyzing the above graph, it is evident that the data is inconsistent and It is observed that year 2016 has only 5 months, 2019 has 11 months and 2021 has only 8 months. That leaves behind a total of 12 months. So, the total months has come down to 72 – 12 = 60. Since we want to pick 4 records a month that brings the total sample space to 60*4 = 240.

To attain this, each year is sampled separately from the master dataset 'aac_master' using where condition. The strata has been set to month and sample size to 4 which gives 4 records from each month in that year. The yearly sampled datasets were named aac_sample16, aac_sample17, aac_sample18, aac_sample19, aac_sample20 and aac_sample21 each representing the years 2016, 2017, 2018, 2019, 2020 and 2021 respectively.

Selection Method	Systematic Random Sampling
Strata Variable	Month

Input Data Set	AAC_MASTER
Random Number Seed	103
Stratum Sample Size	4
Number of Strata	5
Total Sample Size	20
Output Data Set	AAC_SAMPLE16

Selection Method	Systematic Random Sampling
Strata Variable	Month

Input Data Set	AAC_MASTER
Random Number Seed	104
Stratum Sample Size	4
Number of Strata	12
Total Sample Size	48
Output Data Set	AAC_SAMPLE17

Selection Method	Systematic Random Sampling
Strata Variable	Month

Input Data Set	AAC_MASTER
Random Number Seed	105
Stratum Sample Size	4
Number of Strata	12
Total Sample Size	48
Output Data Set	AAC_SAMPLE18

Selection Method	Systematic Random Sampling
Strata Variable	Month

Input Data Set	AAC_MASTER
Random Number Seed	106
Stratum Sample Size	4
Number of Strata	11
Total Sample Size	44
Output Data Set	AAC_SAMPLE19

Selection Method	Systematic Random Sampling
Strata Variable	Month

Input Data Set	AAC_MASTER
Random Number Seed	107
Stratum Sample Size	4
Number of Strata	12
Total Sample Size	48
Output Data Set	AAC_SAMPLE20

Selection Method	Systematic Random Sampling
Strata Variable	Month

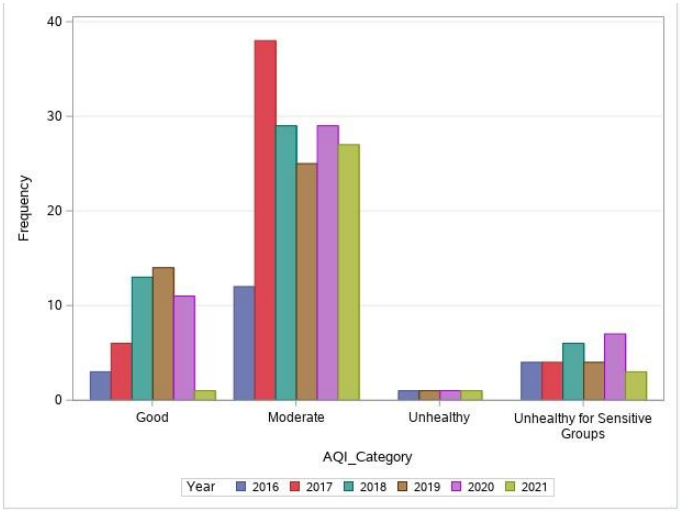
Input Data Set	AAC_MASTER
Random Number Seed	108
Stratum Sample Size	4
Number of Strata	8
Total Sample Size	32
Output Data Set	AAC_SAMPLE21

All the sampled datasets are then merged into one master sample dataset named as 'aac_sample' which has a total of 240 records as expected.

```
NOTE: There were 20 observations read from the data set CUSTOM.AAC_SAMPLE16.
NOTE: There were 48 observations read from the data set CUSTOM.AAC_SAMPLE17.
NOTE: There were 48 observations read from the data set CUSTOM.AAC_SAMPLE18.
NOTE: There were 44 observations read from the data set CUSTOM.AAC_SAMPLE19.
NOTE: There were 48 observations read from the data set CUSTOM.AAC_SAMPLE20.
NOTE: There were 32 observations read from the data set CUSTOM.AAC_SAMPLE21.
NOTE: The data set CUSTOM.AAC_SAMPLE has 240 observations and 12 variables.
NOTE: DATA statement used (Total process time):
      real time           0.03 seconds
      user cpu time       0.00 seconds
      system cpu time     0.01 seconds
      memory              2526.68k
      OS Memory           32960.00k
      Timestamp           10/08/2021 09:53:52 PM
      Step Count                        197  Switch Count  2
      Page Faults                       0
      Page Reclaims                     370
      Page Swaps                        0
      Voluntary Context Switches        95
      Involuntary Context Switches      0
      Block Input Operations            1440
      Block Output Operations           264
```

2.5 AQI Category comparison:

The below histogram has been plotted to analyze the AQI Category variable over the years from 2016 to 2021.

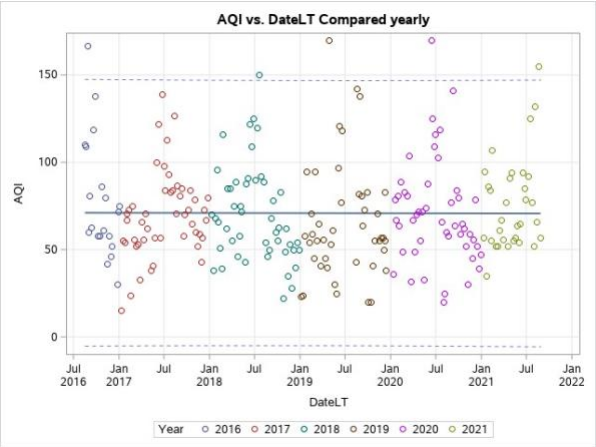


The PROC SGPLOT method has been implemented to obtain the histogram and has been grouped by year to compare effectively. The 'aac_sample' dataset has been used to plot the histogram. Upon observing it is identified that the most common AQI Category the city is experiencing is at 'Moderate' level and relatively 2017 has seen the most number of days with 'Moderate' AQI level. The next common AQI category is 'Good' and 2019 has observed the greatest number of days in this category. Followed by 'Good' we have 'Unhealthy for Sensitive groups' as the next common category. Days with this category is found very less and the year 2020 has seen the most days with this category among the 6 year lot. The 'unhealthy' category is least observed of all the categories and every year is recorded with almost same number of days with this category.

From this graph we can conclude that the city is maintaining a safe environment as approx. 70% of the time the pollutant levels are falling under Good to Moderate levels.

2.6 AQI Correlation among all years:

2.6.1 Scatter plot for AQI vs DateLT:



The above scatter plot has been plotted using the PROC SGPLOT method. The variables plotted are AQI and DateLT and the result is categorized by years. Upon analyzing the graph it is clearly understood that the plot is Non-Linear and is having no relation as the data is very widely distributed with no clear pattern. Hence from the plot we can conclude that they are weakly related. On comparing all years, we can safely assume to have 5 outliers detected from the plot each falling in the year 2016, 2018, 2019, 2020 and 2021.

2.6.2 AQI Correlation among the years:

1 With Variables:	Year
1 Variables:	AQI

Pearson Correlation Coefficients, N = 240	
	AQI
Year	0.00036

After using Scatter Plot the AQI and Year were checked for Correlation by implementing the PROC CORR method. The correlation coefficient (r) value has been calculated to be 0.00036 for the selected variables. This tell us that there is essentially no linear relation between the AQI levels and the Year. The levels are randomly distributed throughout the year and they do not depend on each other.

3) Interpretation and Discussion:

The purpose of this project is to analyze the six-year data of PM2.5 concentration levels observed in the city of Addis Ababa central and provide statistical interpretations in accordance with the U.S. Environmental Protection Guidelines. This data has been obtained from the official website of AirNow, which holds a record of all the PM2.5 concentration levels measured hourly by various fixed air quality monitoring instruments in the U.S. embassy. PM2.5 is regarded as the pollutant with the most adverse health impact and hence a subject of interest.

The master dataset has been obtained after performing various cleaning/screening techniques on the initial data and is analyzed using various analytical techniques such as scatter plots, histograms, boxplots, and other techniques.

On the initial cleaning of the data, it has been noticed that the data has duplicate values for three Timestamps. One strange pattern observed here is that all the three Timestamps belong to the same hour, midnight 3. Although they belong to different years, having the same hour might not be a random fault, so it needs to be addressed. It is also observed that the collected data is not consistent as there is so much data missing. The year 2016 has 7 months missing, 2019 has 1 month, and 2021 has 4 months missing.

After analyzing the data using box plots and scatter plots, it has been observed that the AQI levels are staying relatively low during the first few months and gaining value by approx. 50% during the mid of the year in June, July, and August. High AQI values have been observed very commonly during these months and tend to drop by the end of the year. On overall yearly comparison, it has been observed that the mean AQI levels have been very consistent, standing at around 75. The average upper bound of the AQI levels throughout the six years stands at approx. 165 whereas the average lower bound is at 17.8 approximately. On analyzing the data using histograms, it is understood that the AQI levels ranging from 50 – 70 are the most common levels, with an average share of 25% of all the levels recorded. 2019 has experienced the longest time with almost 30% share of AQI level in the 50 – 70 range. After performing systematic sampling on the datasets, the AQI Category has been compared over the years. It is found that the ‘moderate’ levels, which fall in the range of 51 – 100, constitute almost 60% of all the categories recorded in the 6 years. This is followed by the ‘Good’ category with almost 30% share. 2018 has observed a little spike of approx. 35% in recording the category of ‘Unhealthy for Sensitive Groups’, although this category shares as low as 17.5% on the overall

comparison. On comparing the datasets yearly, 5 outliers were identified. 2 outliers in the year 2017, 2 in the year 2018 and 1 in the year 2019.

On analyzing the correlation between year and AQI using the scatter plot, it is identified that they share a non-liner plot having no relation between them. The correlation coefficient r has been calculated for the two variables and is found to be 0.00036. This implies that there is no relation between the two.

In conclusion, the average AQI levels are very consistent and is predicted to follow the same course. This signifies that there are no measures being taken to get the concentration level under control or the measures in action proved not to be effective in bringing the average AQI level down.

Pros and Cons of descriptive analysis techniques used:

a) Box plots:

Pros: The box plots are very effective when the data is very large as huge data can be accommodated and effectively compared on a single graph in very less time. Data can be easily interpreted using the mean, min, max and outliers on the graph.

Cons: The major drawback of the boxplot is it hides the entire distribution. This limits the understanding of the variable in interest. Box plot only gives a high-level overview, for in-depth research other techniques are to be coupled with this.

b) Scatter plots:

Pros: Scatter plots are highly effective in identifying the correlation between two variables. The presence of every data point in the graph and its clear distribution helps in understanding the graphs in depth. Outliers, the minimum and the maximum are very clearly plotted in the graph.

Cons: Scatter plots only work for continuous variables. For discrete variable the graph shows odd results. Also, not more than two variables can be used for analysis. Scatter plots alone do not give the exact correlation for any two variable and hence requires correlation analysis to understand better.

c) Histograms:

Pros: Histograms are very easy to interpret as the data is divided into equal intervals and displayed using bins. It is very useful when comparing date with large range.

Frequency of occurrences can be easily understood at glance.
Cons: Histograms cannot accommodate large amounts of data. It does not show the exact values of the data as they are categorized into intervals. It is very difficult to analyze more datasets. It can be used only for continuous data.