Project Report

TEAM - **BEYOND DATA**

March 28, 2021

Table of Contents

Introduction

This report provides a detailed explanation of data analysis of Air Pollution data collected at Ho Chi Minh City for year 2019. An external data set consisting of Weather parameters is also used to find any relationship between weather conditions and Air pollution in Ho Chi Minh City. This report is intended for:

- Exploring the Datasets.
- Screening/Cleaning the data
- Identifying Patterns in the Dataset
- Gathering and Exploring External Weather Dataset
- Performing Inferential Tests to determine relationship between weather and air conditions
- Answering research questions based on analysis and test results

Dataset Description

❖ Variables

	VARIABLES	VARIABLE TYPE	DESCRIPTION
1	AQI	NUMERICAL, Discrete	AQI is Air quality Index and can range from 0-500.
2	AQI_CATEGORY	CATEGORICAL, Nominal	Tells whether the AQI level is Good, Moderate, Unhealthy for Sensitive groups, Unhealthy, Very Unhealthy, and Hazardous
3	Parameter	CATEGORICAL, Nominal	It denotes the parameter being used as a measure to represent the air pollution Has Fixed value – 'PM2.5 - Principal'
4	Raw Conc.	NUMERICAL, Continuous	Raw concentration is the hourly PM 2.5 concentrations in micrograms per cubic meter
5	Conc Unit.	CATEGORICAL, Nominal	Unit of PM 2.5 concentrations i.e., UG/M3 (micrograms per cubic meter)
6	NowCast Conc.	NUMERICAL, Continuous	Weighted average of hourly monitored concentrations.
7	Date (LT)	NUMERICAL, Date - Time	Timestamp information of the data points i.e., date and time when the raw concentration was recorded.
8	Day	NUMERICAL, Discrete	Discrete values can range from 1-31

9	Month	CATEGORICAL	Has values ranging from 1- 12 which are just labelling for Jan, Feb, March etc.
10	Year	CATEGORICAL, Ordinal	As it has fixed values – 2016,2017,2018,2019,2020,2021
11	Hour	NUMERICAL, Discrete	Discrete values from 0-23
12	Duration	CATEGORICAL, Nominal	Represents that the data is generated on hourly basis. Has Fixed Value – '1 Hr'
13	QC_Name	CATEGORICAL, Nominal	Identifier used for specifying whether the data is Valid, Invalid or Missing
14	Site	CATEGORICAL	Describes from which city is the data collected
15	24-hr. Midpoint Avg. Conc.	NUMERICAL, Continuous	Average concentration of previous 12 hours, given hour and next 11 hours

❖ Size

The number of observations and variables in each dataset is tabulated below:

DATASET	NUMBER OF OBSERVATIONS	NUMBER OF COLUMNS
TP. HCMN_2016_12	720	14
TP. HCMN_2017_12	744	14
TP. HCMN_2018_12	744	14
TP. HCMN_2019_12	744	14
TP. HCMN_2020_12	744	14
TP. HCMN_2021_02	528	14

External Source: Weather Dataset

The external data set is fetched from a weather forecasting website 'worldweatheronline.com' via API. An API client is written in Python programming language to grab JSON structured data from API response. The data is then transformed using pandas data frames and few of the prominent weather parameters are extracted in a CSV file.
Source: https://www.worldweatheronline.com/developer/api/docs/historical-weather-api.aspx

❖ Variables

The below tables list the variables of the weather dataset. There are **4392** observations and **7** variables in the Weather Data frame.

	VARIABLES	VARIABLE TYPE	DESCRIPTION
1	day	DATE	Date information of the data points i.e., date when the weather parameters were recorded.
2	time	NUMERICAL, Discrete	Discrete values of hours from 0-23
3	tempC	NUMERICAL, Discrete	Temperature in degree Celsius
4	windspeedKmph	NUMERICAL, Discrete	Wind speed in Kilometres per hour
5	weatherDesc	CATEGORICAL, Nominal	Weather Condition Description like Sunny, Clear, Cloudy etc.
6	humidity	NUMERICAL, Discrete	Humidity in percentage (%)
7	visibility	NUMERICAL, Discrete	Visibility in Kilometres

```
'data.frame':   4392 obs. of  7 variables:
 $ day          : Date, format: "2016-12-01" "2016-12-01" ...
 $ time         : int  0 1 2 3 4 5 6 7 8 9 ...
 $ tempC        : int  26 26 25 25 25 25 25 26 28 29 ...
 $ windspeedKmph: int  5 6 7 8 8 9 9 10 12 13 ...
 $ weatherDesc  : chr  "Clear" "Clear" "Clear" "Clear" ...
 $ humidity     : int  89 89 90 90 90 89 89 83 78 72 ...
 $ visibility   : int  10 10 10 10 10 10 10 10 10 10 ...
```

Data Pre-Processing

The datasets of Ho Chi Minh city air quality parameters are analyzed to fix any accuracy or missing value issues thereby creating a cleaned dataset for further analysis

❖ Handling Ho Chi Minh 2016 Dataset

From the Categorical and Numerical issues, we can say that data cleaning must be performed for Ho Chi Minh PM.2.5 2016_12 dataset to make it compatible with all the other datasets for the statistical analysis.

Frequencies for Categorical Variables HCMC 2016

Site	Frequency	Percent
Ho Chi Minh City	720	100.00

QC_Name	Frequency	Percent
Invalid	2	0.28
Valid	718	99.72

AQI_Category	Frequency	Percent
2	234	32.50
3	401	55.69
4	85	11.81

Numerical value	CATEGORY NAME
1	Good
2	Moderate
3	Unhealthy for Sensitive groups
4	Unhealthy
5	Very Unhealthy
6	Hazardous

Descriptive Statistics for Numeric Variables HCMC 2016

Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
Raw_Conc_	720	0	9.0000000	43.9736111	37.0000000	985.0000000	53.7419334
_24_hr__Midpoint_Avg__Conc_	720	0	20.7000000	41.2834722	40.8000000	77.0000000	11.0576598
AQI	720	0	69.0000000	114.3361111	114.0000000	162.0000000	23.5916424

Fixing Issues

- The AQI_Category is defined as number rather than Categories. This is addressed by replacing numeric values with US EPA Provided Category Labels
- The 24 hr Midpoint Avg Conc doesn't exist for other datasets, so this column is eliminated
- The NowCast Conc is calculated using the Raw Conc by applying SQL queries to aggregate past 12 hours data points and utilize formals provided by US EPA to compute the values.

❖ Merging Datasets

By now Ho Chi Minh 2016 dataset looks same as rest of the datasets. So, it is time to merge all the datasets into one dataset and call it as **Master Dataset.** The master dataset contains 4117 rows and 14 columns.

Analyzing Master Dataset

Since we merged all the years data together from Ho Chi Minh 2016 to Ho Chi Minh 2021, we will again analyze the variables for evaluating the issues.

Descriptive Statistics for Numeric Variables HCMC 2016-2021

Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
Raw_Conc_	4117	0	-999.0000000	18.9504494	30.0000000	161.0000000	126.1672478
AQI	4117	0	-999.0000000	81.7029390	91.0000000	176.0000000	134.9651034
NowCast_Conc_	4117	0	-999.0000000	19.1533641	30.5000000	129.6000000	124.8888206

Frequencies for Categorical Variables HCMC 2016-2021

Site	Frequency	Percent
Ho Chi Minh City	4117	100.00

QC_Name	Frequency	Percent
Missing	61	1.48
Valid	4056	98.52

AQI_Category	Frequency	Percent
Good	93	2.26
Moderate	2355	57.20
N/A	60	1.46
Unhealthy	387	9.40
Unhealthy for Sensitive Groups	1222	29.68

- We see that from table that there are 61 missing values in QC_Name
- We observe that there are 60 values mentioned as Not Applicable(N/A) for AQI_Category
- From descriptive statistics below, there are negative values (-999) found in Raw_Conc_, AQI and NowCast_Conc_ variables

Handling Master Dataset

Let us now begin by cleaning the master dataset for NowCast_Conc_, AQI and Raw_Conc variables in a step-by-step process as follows,

Removing Invalid and Negative values

- We will remove all the negative values from NowCast_Conc_, AQI and Raw_Conc variables
- Remove the "Invalid" value from QC_Name variable
- Replace all the removed values with a dot (.)

Backfilling method

- The missing/invalid values replaced by dot(.) are to be handled using Backfill Technique.
- For this, we have created a SAS macro to be used for all the three variables i.e., NowCast_Conc_, AQI and Raw_Conc

What this macro does:

This macro retains the values of previous non-missing data while traversing through the dataset row-by-row. At any row, if it checks a missing value (dot) for a given variable, it replaces it with the retained value from last non-missing row. This is thus known as backfilling technique.

Fixing Categorical variables

- Missing QC_name value is replaced with "Valid" for positive value of Raw_Conc_.
- The AQI_Category values are categorized based on the AQI value.

Descriptive Analysis

Descriptive statistics are basically used to describe features of the data.

❖ Summary Statistics

- Summary statistics helps to know the range of the data that a variable possess, such as Minimum, Maximum, Mean, Median, and quartile details
- Below table shows the summary statistics of all the variables in our study

```
                Site                    Parameter              Date__LT_        Year          Month
Ho Chi Minh City:4117   PM2.5 - Principal:4117   01DEC16:01:00:00:   1   Min.   :2016   Min.   : 1.00
                                                 01DEC16:02:00:00:   1   1st Qu.:2017   1st Qu.:12.00
                                                 01DEC16:03:00:00:   1   Median :2018   Median :12.00
                                                 01DEC16:04:00:00:   1   Mean   :2018   Mean   :10.71
                                                 01DEC16:05:00:00:   1   3rd Qu.:2020   3rd Qu.:12.00
                                                 01DEC16:06:00:00:   1   Max.   :2021   Max.   :12.00
                                                 (Other)          :4111
      Day             Hour            AQI          Raw_Conc_       Conc__Unit    Duration     QC_Name
Min.   : 1.0   Min.   : 0.00   Min.   : 21.00   Min.   :  0.00   UG/M3:4117   1 Hr:4117   Valid:4117
1st Qu.: 8.0   1st Qu.: 6.00   1st Qu.: 75.00   1st Qu.: 22.00
Median :15.0   Median :12.00   Median : 91.00   Median : 30.00
Mean   :15.3   Mean   :11.52   Mean   : 97.44   Mean   : 34.14
3rd Qu.:23.0   3rd Qu.:18.00   3rd Qu.:118.00   3rd Qu.: 43.00
Max.   :31.0   Max.   :23.00   Max.   :176.00   Max.   :161.00

                        AQI_Category  NowCast_Conc_
Good                          :  93   Min.   :  5.1
Moderate                      :2413   1st Qu.: 22.8
Unhealthy                     : 387   Median : 30.5
Unhealthy for Sensitive Groups:1224   Mean   : 34.1
                                      3rd Qu.: 42.1
                                      Max.   :129.6
```

❖ Year and Month wise Distribution

Year <int>	Month <int>	Total_Count <int>
2016	12	718
2017	12	741
2018	12	651
2019	12	737
2020	12	739
2021	2	527

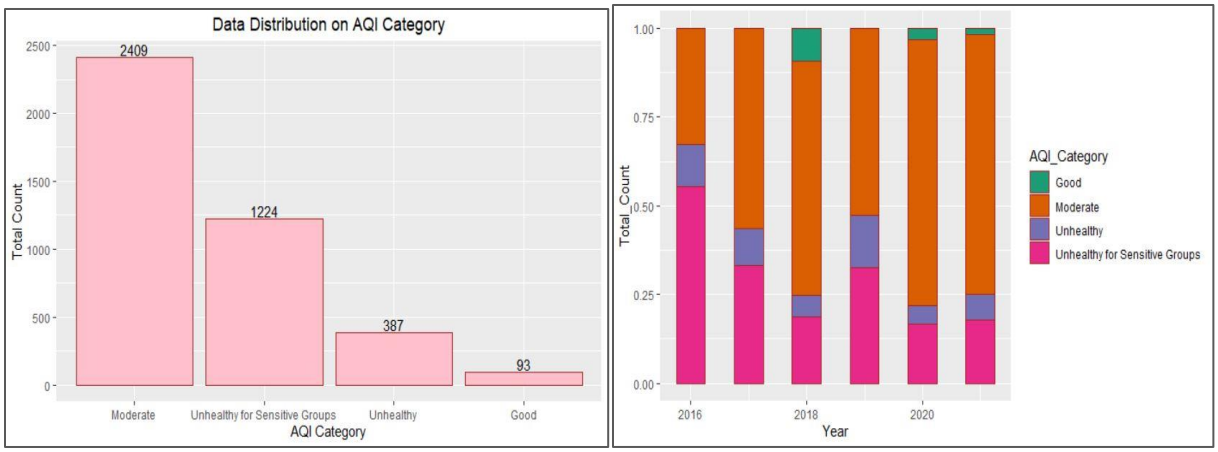
6 rows

Interpretations

- Above table depicts year and month wise distribution after removing the January month data
- All the years have December data except for 2021 which has data for February

❖ AQI Category

Proportion Year Wise AQI_Category Distribution



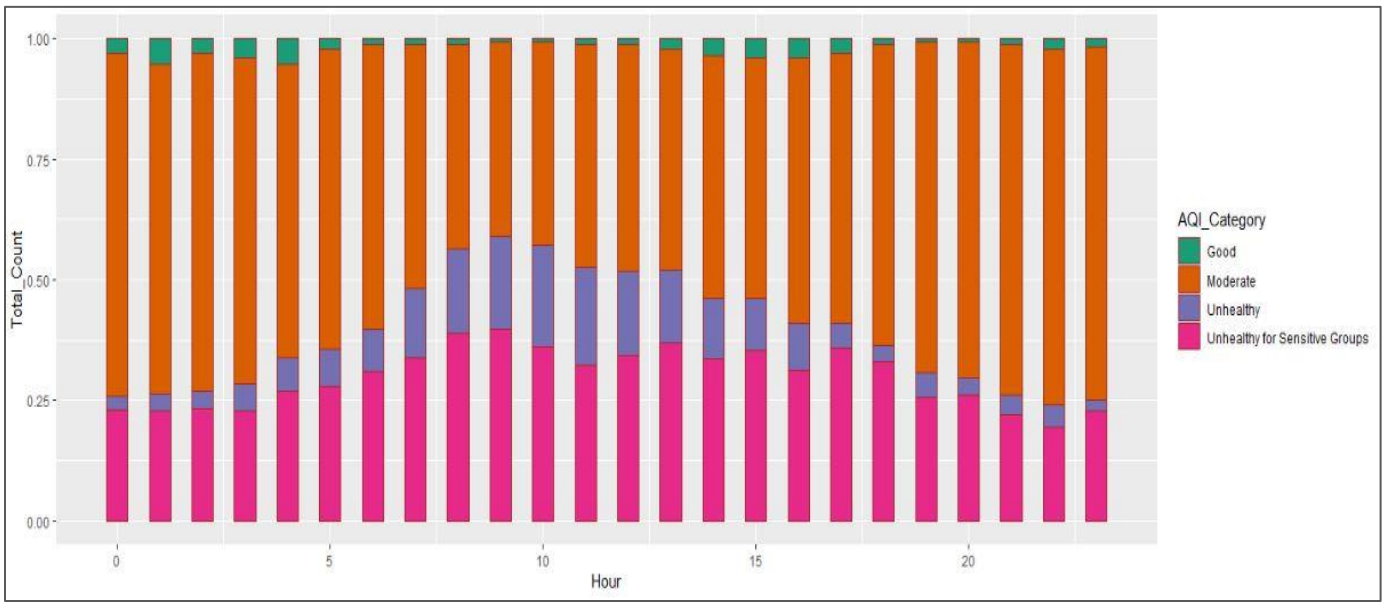
Checking the AQI distribution on a yearly basis to see if there are any conclusions that we can make for the Air Quality

Interpretations

- More than half of the observations falls into Moderate category i.e., 59% and 2% of the data is classified as Good AQI
- Year 2020 has only ~20% of the data with AQI categorized as Unhealthy which is lowest amongst all years and the pattern continues in the following year i.e., 2021

AQI_Category with Hour

Checking data on hourly basis to see if there are any patterns we can observe:

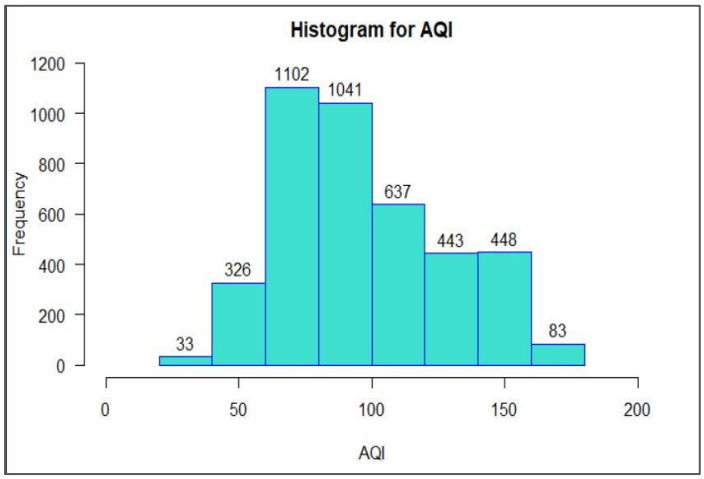


Interpretations

- Above figure shows the peak hours where AQI Category is Unhealthy for more than 50% of the data.
- The peak in Unhealthy Hours is ranging from 8:00 to 13:00 hour on a 24-hour scale
- For late night and early morning, 75% of the data has AQI in the range of Good to Moderate (Hours = 21,22,23,0,1,2)

❖ AQI

```
$Summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   75.00   91.00   97.46  118.00  176.00

$`Standard Deviation`
[1] 30.71327

$Range
[1]  21 176

$`Inter-Quartile Range(IQR)`
[1] 43
```



Interpretation

- AQI value ranging from 60 – 100 (approximately) has high number of observations

  Shape: Slightly Right Skewed | ***Spread***: Min = 21, Max = 176 | ***Centre:*** 75

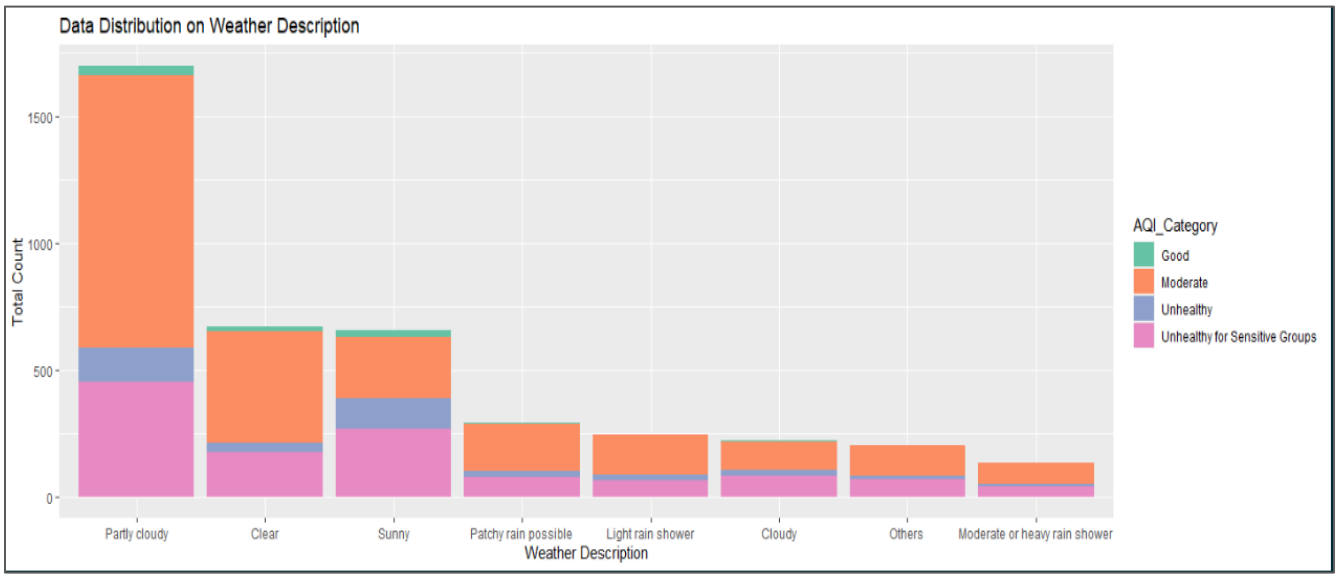
❖ WeatherDesc

weatherDesc <fctr>	Total_Count <int>	Perc_Contri <dbl>
Partly cloudy	1697	41.26
Clear	670	16.29
Sunny	655	15.93
Patchy rain possible	291	7.08
Light rain shower	245	5.96
Cloudy	219	5.32
Moderate or heavy rain shower	132	3.21
Light rain	42	1.02
Overcast	42	1.02
Moderate rain at times	24	0.58

1-10 of 19 rows

Weather Description has 19 distinct values. This can be reduced to a smaller number of high-level weather description categories.

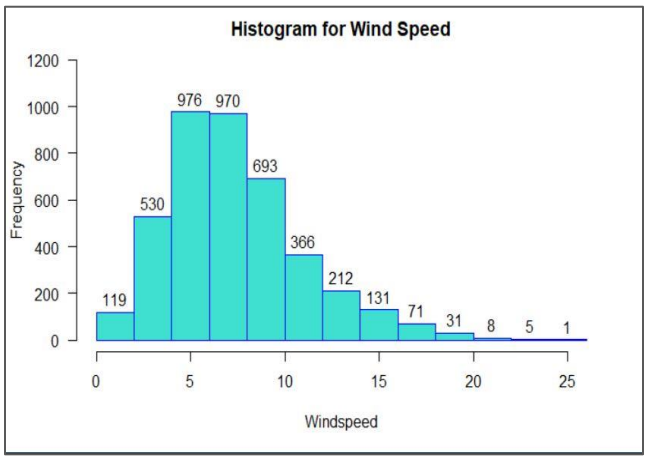
WeatherDesc and AQI_Category

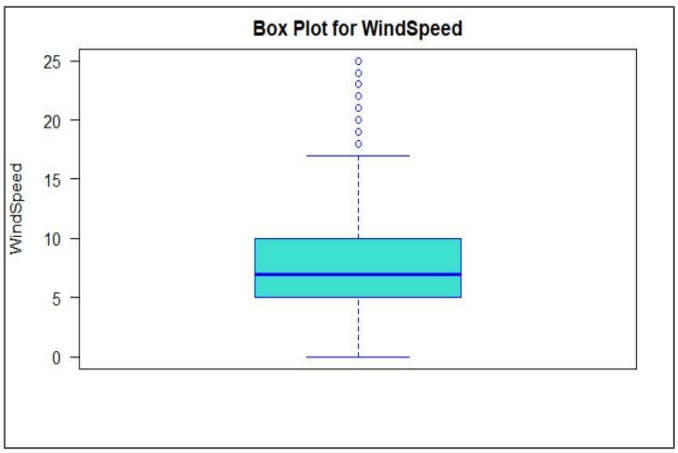


Interpretations

- The distributions of weather description on AQI Category for total absolute volume
- From the above chart, no significant relation can be found between Weather description and AQI Categories

❖ Wind Speed





Interpretations

- Data is skewed towards right
- Most of the data is concentrated towards lower values of the Windspeed, implying large number of observations has a lower windspeed
- Most cases the wind speed is in between 4 to 8 Kmph (approximately)
- For the data spread outside the upper whisker can be termed as outliers

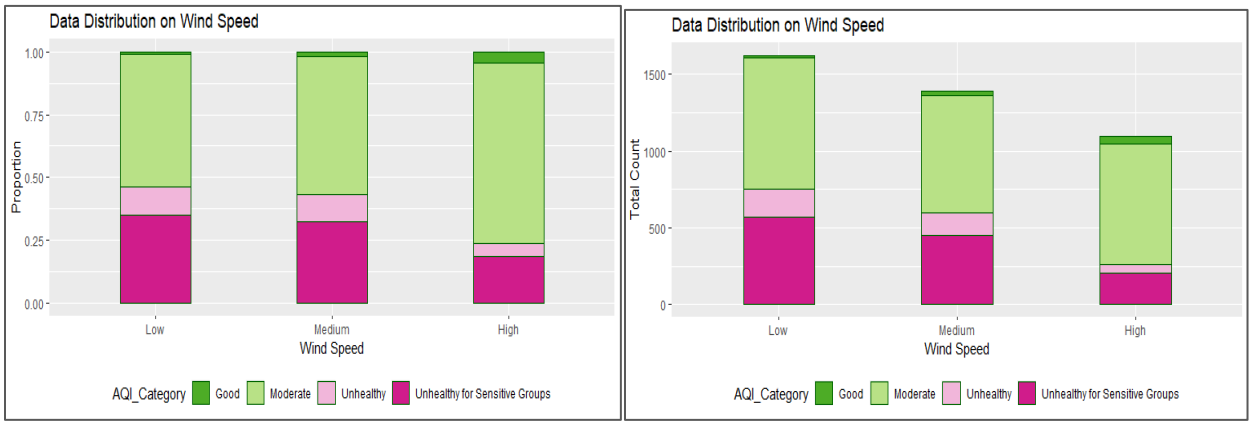
❖ Windspeed Category

Binning numerical windspeed data into 3 high level categories: Low, Medium and High

windspeed_Cat <fctr>	Total_Count <int>	Perc_Contri <dbl>
Low	1625	39.51
Medium	1389	33.77
High	1099	26.72

3 rows

Windspeed category with AQI category



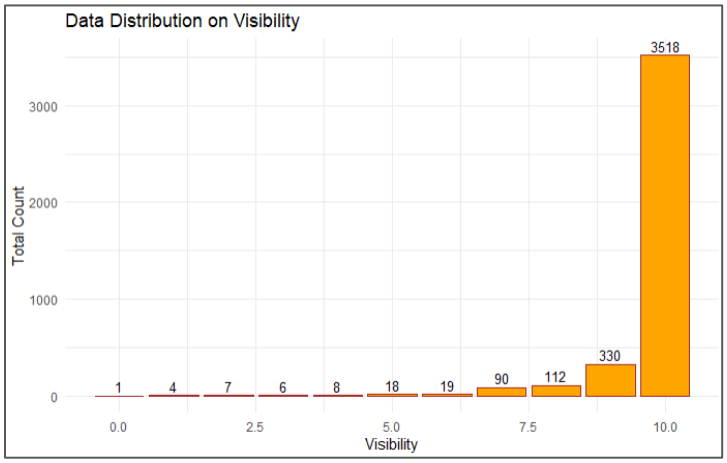
Data distribution of wind speed category on AQI category while taking total count into account

Data distribution of wind speed category on AQI category while taking total count into account for windspeed and proportions for AQI category.

Interpretations

- As wind speed increases AQI improves and shifts to Moderate – Good range. 76% of the data has AQI as Good and Moderate for wind speed categorized under "High"

❖ Visibility

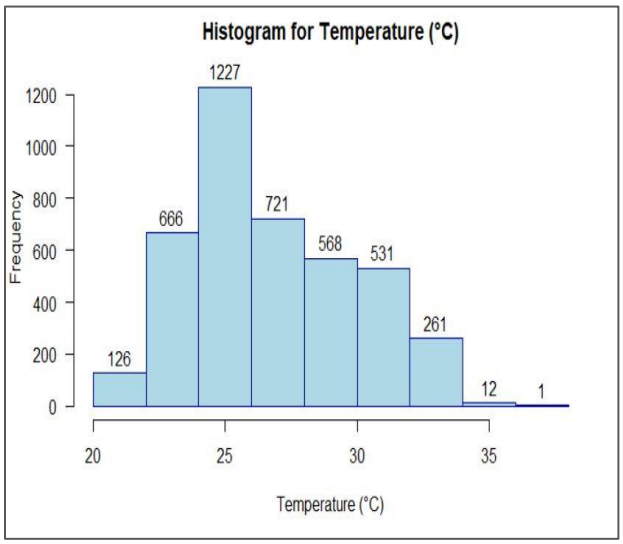


Interpretations

- More than 90% of the data is tagged against visibility = 10

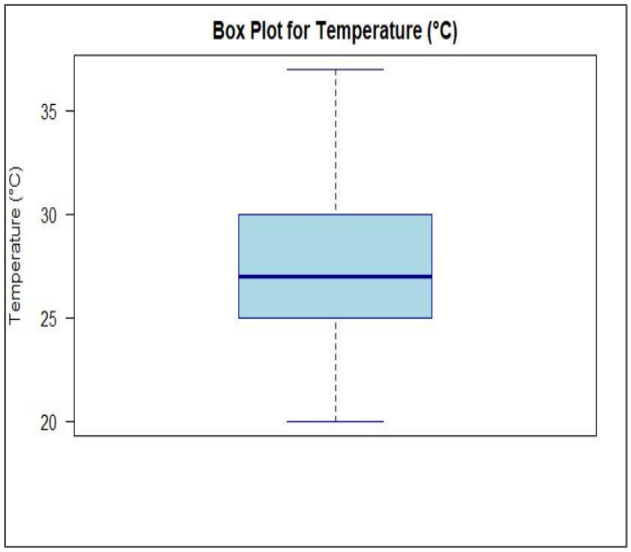
- Hence this variable cannot be used for any significant analysis as no inference could be derived using Visibility

❖ Temperature



Histogram

Shape: Slightly Right Skewed
Spread: Min = 20, Max = 37
Centre: 25



Box Plot

Min: 20, Max: 37
Mean: 27.5, Median: 27
Right Skewed, Mean > Median

Interpretation

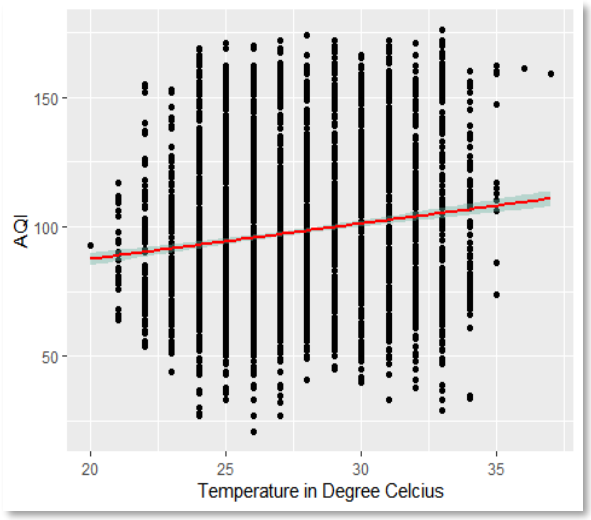
- Above histogram shows the frequency distribution of temperature
- Temperature for majority of the observations is concentrated around 25°C
- Box Plot the concentration of temperature values
- The values are concentrated from the range of 24 to 30 degree Celsius (approximately)
- Since the temperature scale is in the moderate range it would not influence AQI

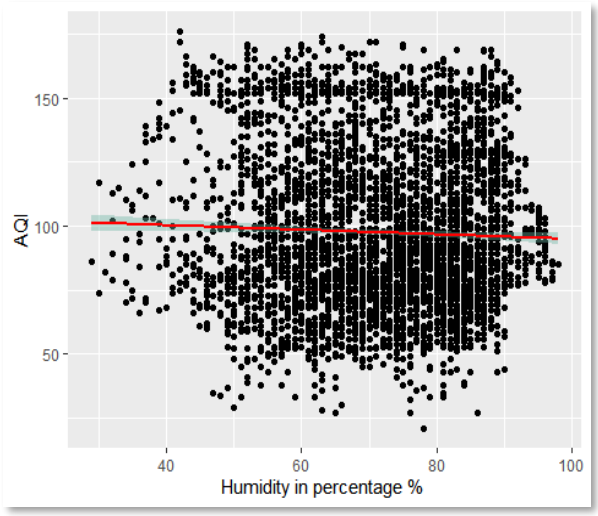
Inferential Analysis

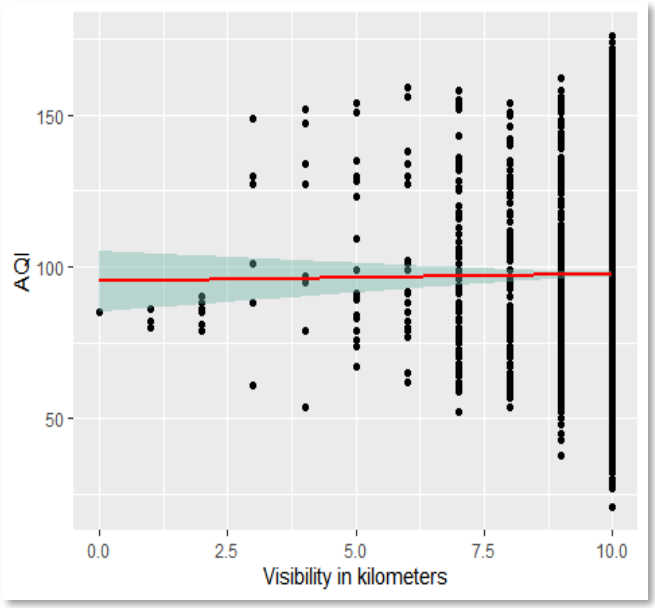
❖ Pearson Correlation Test

The Pearson correlation tests helps in analysing the relationship between two numerical variables. Here, the AQI of Ho Chi Minh City is being analysed to check for any relationship that may exist with the weather parameters namely Temperature, Humidity, Visibility and Windspeed.

AQI vs Temperature/Humidity/Visibility

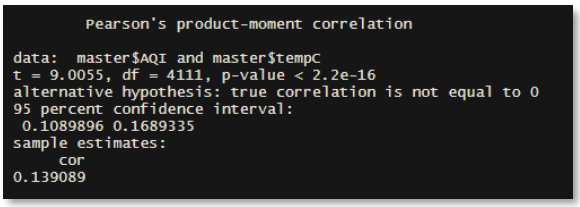


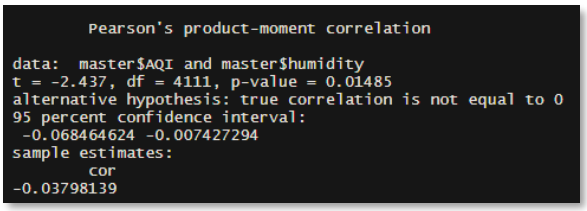


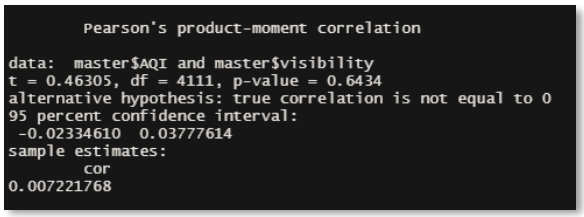


Interpretations

1. Temperature, Humidity and Visibility are explanatory variable, and AQI is response variable.
2. Form: The scatter plot shows the none of the variables have linear relationship with AQI as the data points are scattered in no clear pattern over the canvas.
3. Strength: It is evident that the strength of the correlation between the weather variables and AQI is quiet poor.







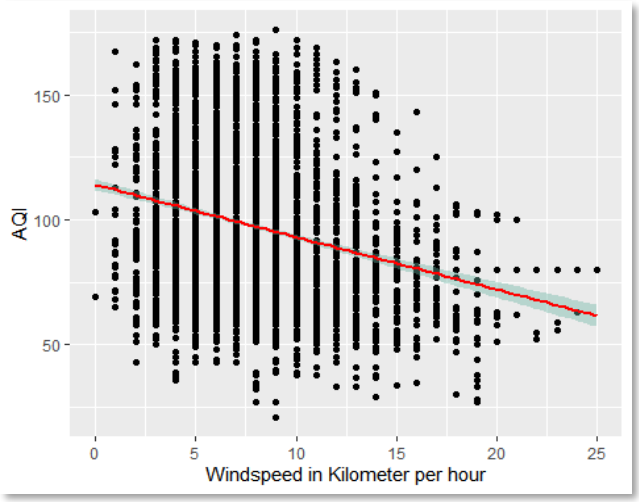
4. From the Pearson Correlation Test, it can be inferred that the relationship between AQI and Temperature/Humidity/Visibility is significant since the P value is very smaller than 0.05.
5. However, the value of correlation coefficients is very close to 0. Hence, we cannot reject the null hypothesis that true correlation is equal to 0. Therefore, it can be inferred that there is no correlation between AQI and Temperature/Humidity/Visibility.

Reasoning

The range of temperature as observed from scatterplot is very small and lies between 20 and 35. Similarly, the range of humidity as observed from scatterplot is moderate and lies between 50 to 85. Also, almost 85% of the values of visibility are fixed to 10 as shown in the bar graph below Moreover, the data available is just for the month of December that results in minimal and minor variations in weather parameters which is not enough to analyse the effect of variations in temperature/visibility/humidity on Air Quality. Thus, due to limited data and discrete nature of weather parameters, it cannot be inferred that AQI, and Temperature/Visibility/humidity are related. That is why, more data corresponding to other months with significant variations in values is required to make any conclusive statement.

AQI vs Windspeed

The below scatter plot describes the relationship between Air Quality Index and Windspeed in kilometres per hours of Ho Chi Minh City recorded hourly. Following the plot, is the Pearson correlation Test to gather insights about the strength and significance of the relationship.




```
        Pearson's product-moment correlation

data:  master$AQI and master$windspeedKmph
t = -16.399, df = 4111, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2762550 -0.2188792
sample estimates:
       cor
-0.2477844
```

Interpretations

1. **Form**: A slightly negatively decreasing linear form can be observed for Wind Speed.
2. **Strength**: It is evident that the strength of the correlation between two variables is quiet poor since the data points are spread instead of being gathered.
3. From the Pearson Correlation Test, it can be inferred **P value** is **significant** as it very less than 0.05.
4. The value of correlation coefficient is **-0.247** which is very weak. Thus, there is very weak and negative correlation between AQI and Windspeed.

Reasoning:

It can be inferred from the analysis and test results that there exists a linear relationship between the AQI and Windspeed. But since, AQI is not solely dependent on the speed of winds in the atmosphere, the strength of the relationship is very weak. The AQI is affected by a lot of variables which are out of the scope of this report. Thus, for a bivariate analysis between AQI and windspeed, existence of dependence can be observed but amount of effect of windspeed on AQI is low.

❖ Chi Square Analysis

The relationship between the categorical variables is based on frequency rather than values. In this analysis, the following methods are used for examining the relationships between the variables:

- Graphical: Stacked bar chart
- Descriptive statistics: Cross Tables or 2x2 Contingency Tables
- Hypotheses testing: Chi-square tests to test if two categorical variables are independent.
- Metric to measure the strength of the relation: Kendall's Tau and Spearman Correlation

In this analysis, the chi-square test to determine the association between the categorical variables is used. The first step is to define the Null and Alternative Hypothesis.

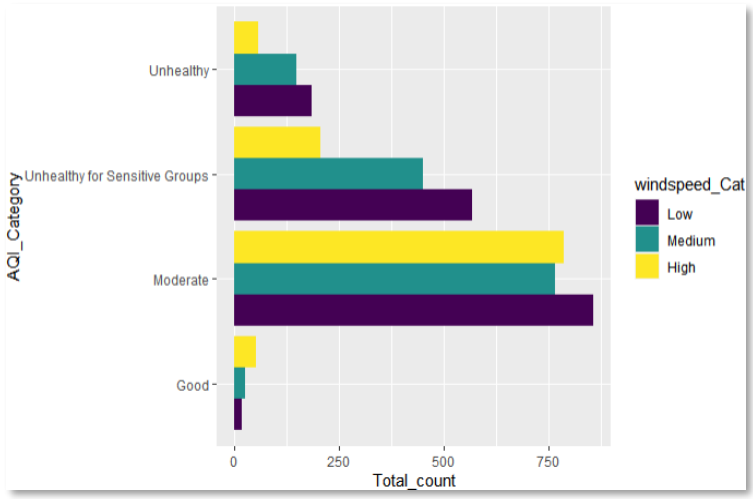
Null Hypothesis H_0: The two variables are independent of each other among all subjects in population.
Alternate Hypothesis H_a: The two variables are related to each other.
Interpretation:

AQI vs Windspeed

```
                              Low Medium High
  Good                         18     24   51
  Moderate                    856    766  787
  Unhealthy for Sensitive Groups 567    451  206
  Unhealthy                   184    148   55

        Pearson's Chi-squared test

data:  t1
X-squared = 177.05, df = 6, p-value < 2.2e-16


        Kendall's rank correlation tau

data:  as.numeric(master$windspeed_Cat) and as.numeric(master$AQI_Category)
z = -11.512, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.1626442
```

Interpretations

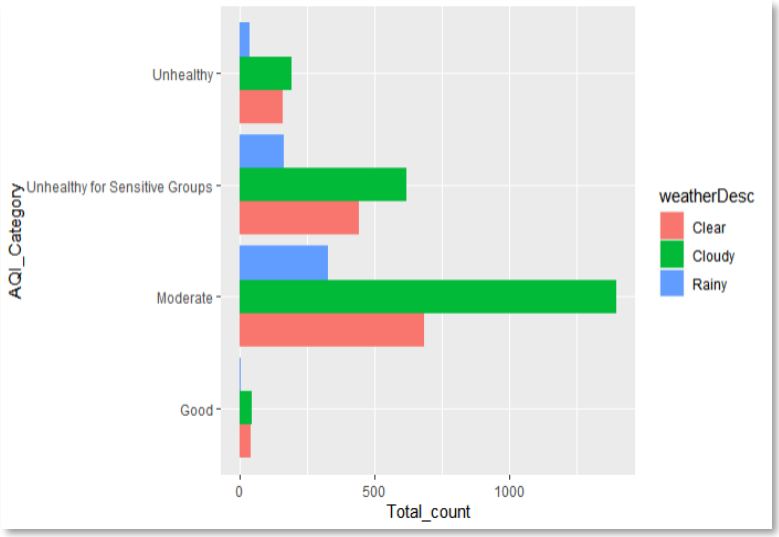
1. The side-by-side bar chart between the AQI and Windspeed categories shows that the groups of bars do not look alike. Thus, it can be inferred that the two variables are not independent to each other.
2. In the Contingency Table, for High values of Windspeed, a relationship exists with Good, Unhealthy for Sensitive Groups and Unhealthy AQI since the deviations from expected values of 31, 408 and 129 are quite high while for Moderate AQI, the distribution is almost similar.
3. The Chi-Squared value is significantly higher than the decision point/ critical value (for DF of 6 from decision point table). Also, the p-value for Chi-Square value is very smaller than 0.05. Thus, it can be inferenced that based on Chi-Square Test, the relationship between AQI and Windspeed is statistically Significant.



4. However, the strength of the relationship is too weak to be considered as significant. The Kendall's rank correlation tau values is -0.162 which is very close to zero, thus we cannot reject the null hypothesis and infer that though there is a significant relationship between the variables, the strength is very weak.

AQI vs Weather Description

```
                                Clear Cloudy Rainy
  Good                             42     46     5
  Moderate                        683   1398   328
  Unhealthy for Sensitive Groups  442    617   165
  Unhealthy                       158    194    35

        Pearson's Chi-squared test

data:  t1
X-squared = 51.355, df = 6, p-value = 2.513e-09


        Kendall's rank correlation tau

data:  as.numeric(master$weatherDesc) and as.numeric(master$AQI_Category)
z = -4.364, p-value = 1.277e-05
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.06265903
```



Interpretations

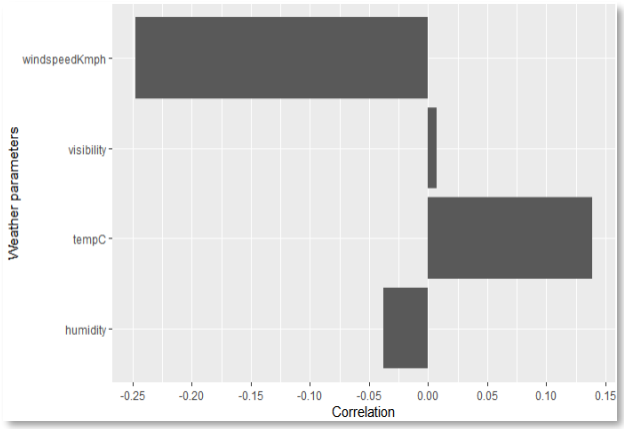
1. The side-by-side bar chart between the AQI and Weather Description categories shows that the groups of bars look alike though the proportion of frequency is highest for Moderate and lowest for Good AQI. Thus, it can be inferred that the two variables are independent to each other.
2. In the Contingency Table, a similar trend can be seen as the proportional contribution of Clear, Cloudy and Rainy weather is similar across various AQI Categories. However, the deviations for Rainy weather from expected values of 31, 803, 408 and 129 are quite high while for others it is not.
3. The Chi-Squared value is slightly higher than the decision point/ critical value (for DF of 6 from decision point table). Also, the p-value for Chi-Square value is very smaller than 0.05.

This is due to large sample size that results in higher probability of rejecting the null hypothesis. For practical implication, the strength analysis of correlation is needed.

4. The strength of the relationship clearly shows that no relationship exists between the two. The Kendall's rank correlation tau value is -0.062 which is very close to zero, thus we cannot reject the null hypothesis and infer that there is a no linear correlation between the variables.

❖ Conclusions

- The Pearson correlation test indicates that there is no correlation between the AQI and weather parameters except for Wind Speed. The Scatterplot between AQI and Windspeed displays negative linear relationship for higher values of windspeed. However, the strength of relationship is quite low to be used as a predictor of AQI.
- The side-by-side bar graph between AQI and weather parameters look almost alike except for Wind Speed. For windspeed, the graph is different for Moderate and Good AQI. This shows that a relationship does exist between AQI and Windspeed.
- The Contingency tables also show a similar result wherein the proportions of frequencies between AQI Categories and Weather Parameter categories are almost same except for higher range of values of weather parameters that have greater deviations from expected values.
- The Chi-Square statistics also shows a highly significant value above the decision point for AQI vs Windspeed with a very low p-value indicating significant relationship. However, for other weather parameters, the chi-squared value is slightly above the decision point with a low p-value indicating a slight relationship which must be confirmed with a strength analysis.
- Kendall's rank correlation quantifies the strength of relationships between AQI and Weather parameters is very low to be considered for any kind of regression. The highest tau value is 0.162 (for Windspeed) which is too low and close to zero.

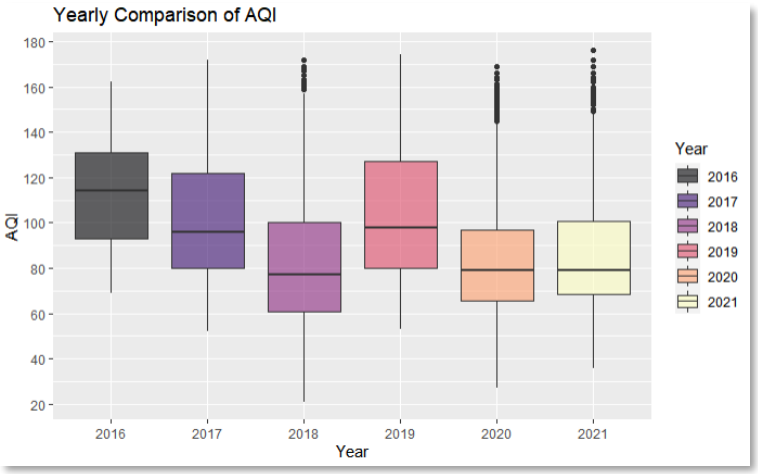


The graph proves that there is no linear correlation between AQI & humidity, AQI & Temperature, and AQI & Visibility. The correlation between Windspeed and AQI is very weak and negative. That is, the air quality of the city slightly improves when the speed of the wind is higher.

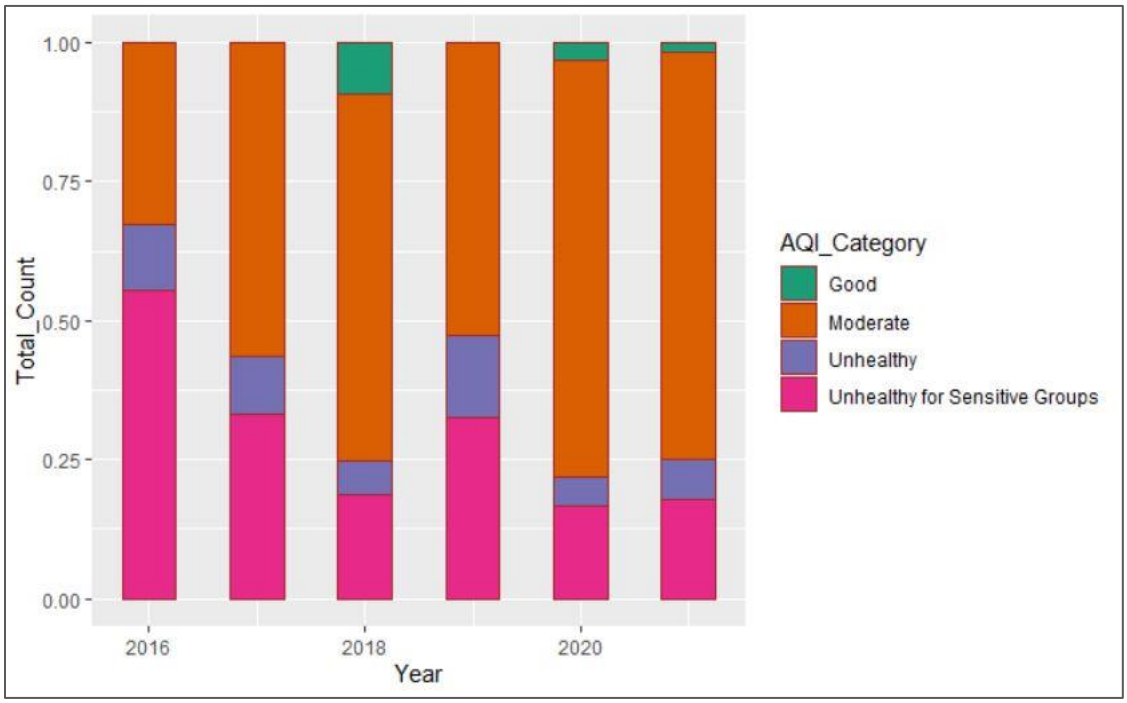
Research Questions

By the virtue of Descriptive and Inferential Analysis performed above, the following research questions are answered.

❖ Did the Air Pollution change over the period of last 6 years?



As we can see from the figure, AQI values tends to shift to the lower side post year 2020. It is evident from the Box plot, that the median AQI was highest in the year 2016. Hence 2016 was the most polluted year in the past five years based on the given data set. The value of Air quality gradually decreased over the years until 2018 but witnessed sudden rise in the year 2019. Post Covid, the air quality of the Ho Chi Minh City improved and hence AQI values were lower in the years 2020 and 2021. Checking the proportion distribution of the overall data on year basis split on AQI Category:



2016 – 70 % data categorized as Unhealthy AQI

2017 – 40% data categorized as Unhealthy AQI

2018 – 25% data categorized as Unhealthy AQI

2019 – 47% data categorized as Unhealthy AQI

2020 – 20% data categorized as Unhealthy AQI, 80% data categorized under Healthy group

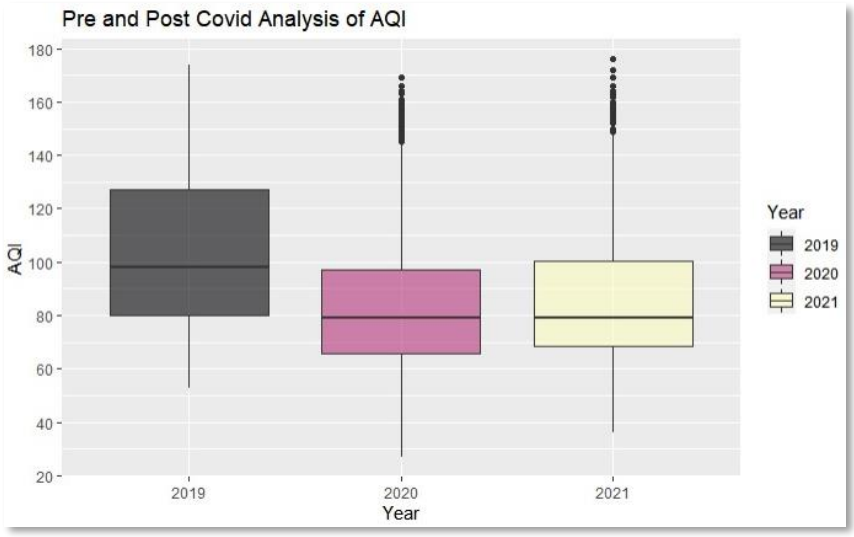
2021 – 25% data categorized as Unhealthy AQI, 75% data categorized under Healthy group

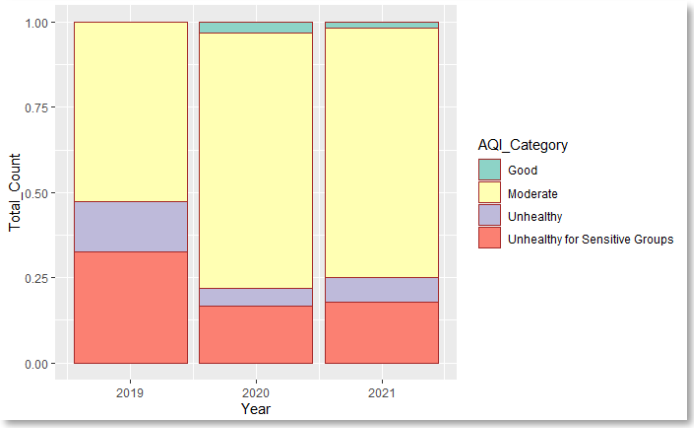
This clearly depicts, Air Pollution has seen reduction, i.e., only close to 20% of the data has AQI categorized as Unhealthy. Hence, we can infer from the above figures that AQI changes over the period of time.

❖ Did Lockdown in Ho Chi Minh City due to COVID-19 impact the Air Quality of the city?

To check the impact of lockdown in the city we will check the AQI for 3 specific years, year in which lockdown was imposed i.e., year 2020, a year before the lockdown i.e. 2019 and a year post lockdown i.e. 2021.

It is evident from the boxplot given below that the median AQI for years 2020 and 2021 is lesser than that of the year 2019. Hence it can be concluded that the AQI levels which were higher in the year 2019 got reduced after the lockdown. Thus, it can be clearly inferred that lockdown resulted in reduction of air pollution and improvement in Air Quality. This might be a result of shutdown of pollutants releasing industries and reduced number of vehicles on road.





The figure shows, improvement in Air Quality post lockdown:

- ***2019***: 52% of the data has AQI categorized as Moderate
- ***2020***: 80% of the data has AQI categorized as Moderate to Good
- ***2021***: 75% of the data has AQI categorized as Moderate to Good

```
        Paired t-test

data:  cdf$AQI.x and cdf$AQI.y
t = 13.627, df = 733, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 16.40159 21.92267
sample estimates:
mean of the differences
               19.16213
```

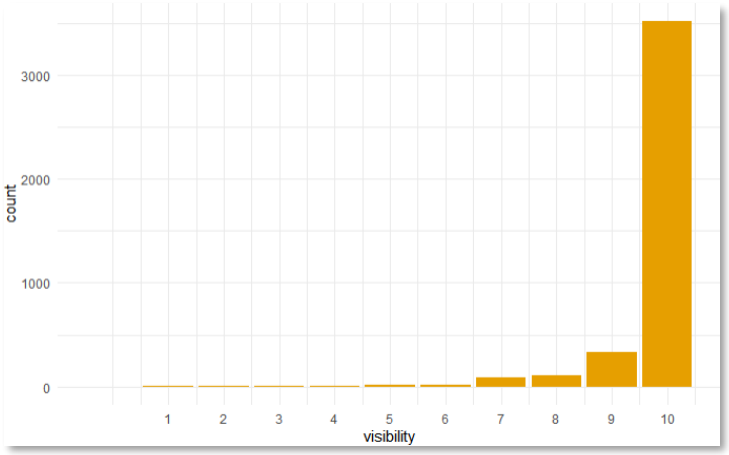
The Paired T-Test is performed using the AQI level measurements taken from corresponding hour of each day in the month of December for years 2019 and 2020 to perform Pre-Post Analysis. The subset of datasets for years 2019 and 2020 consisting of Day, Hour and AQI variables are merged(intersection, common values only) based on Day, and Hour(refer to Inferential Analysis in Appendix for R code) since, the month is December for both datasets. Each observation in the merged dataset consists of hourly AQI levels corresponding to each day of December month of 2019 and 2020.

The results from the paired t-test between AQI levels of year 2019 and 2020 clearly shows that the p-value is significant resulting in rejection of null hypothesis which states that *'True difference in means is equal to zero'*. Thus, the mean AQI for years 2019 and 2020 are statistically different with a t-value of 13.627. Also, the mean value of paired differences is approximately 19 which is significant enough to infer a reduction in Air Pollution between these years.

Hence, it can be said that the pollution decreased after the COVID-19 lockdown in city.

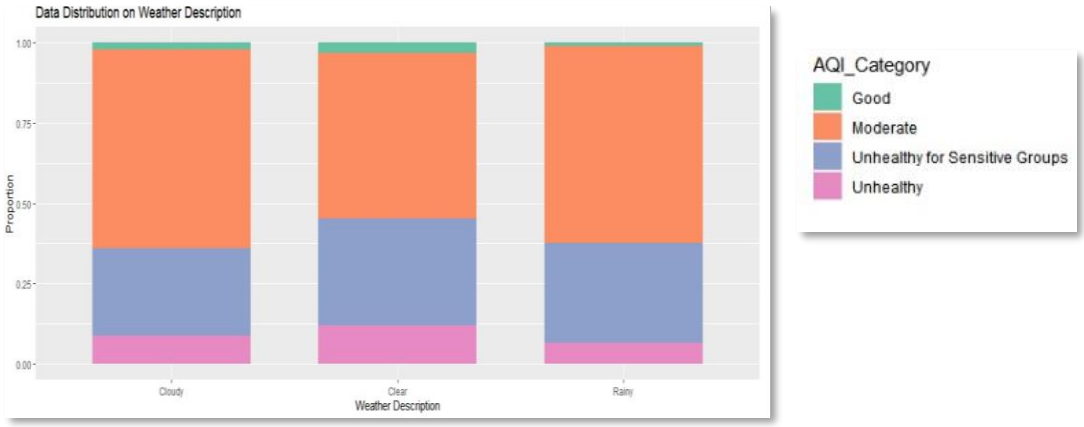
❖ What is the amount of reduction in visibility due to rise in air pollution?

It can be inferred based on the inferential analysis that the visibility is not affected by Air Quality Index variations. This is due to the fact that almost 85% of the values of visibility are fixed to 10 as shown in the bar graph below. Since, the data available is only for December month of the years 2016-2020, the variations in visibility are minimal. Thus, there is a need of data for different months to explore further about the relationship between the AQI and visibility in practical scenarios.



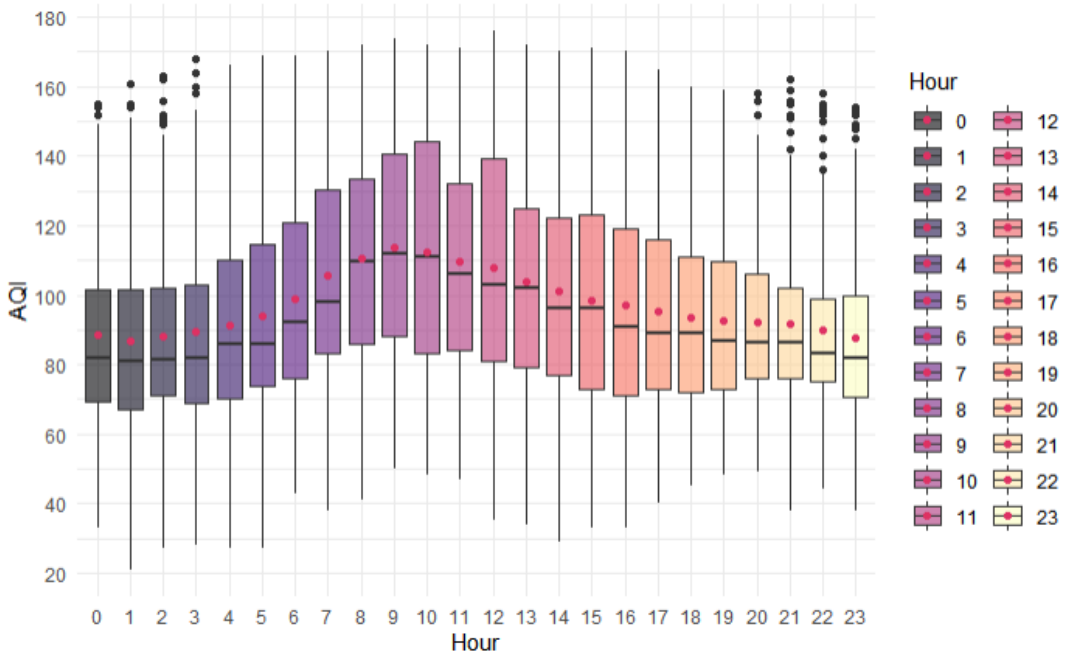
❖ Which weather condition (Sunny, Cloudy etc.) leads to major rise in Air Pollution?

In the month of December, the weather is mostly cloudy. The distribution of AQI categories for different weather conditions is similar. Thus, it cannot be inferred that any specific weather condition leads to major rise in Air Pollution.



❖ What are the peak air pollution hours in a day over the 5 years period?

As can be inferred from the Boxplot, 8:00am to 10:00 am are the peak pollution hours during a day with 09:00 am being the highest. This can be attributed to peak traffic and industrial working hours. Moreover, there is a gradual increase in air pollution from 4:00am to 09:00am and afterwards it decreases gradually till 11:00pm in the night.



❖ What is the effect of variations in Wind Speed on Air Pollution?

It is evident from the Scatterplot, the side-by-side bar plots, the Chi-Square and Kendall's Tests that Windspeed has a statistically significant effect on Air Pollution. As the windspeed increases above 10km per hour, there is an improvement in the Air Quality since the AQI levels drop by a value of 20. The lower strength of the relationship as clear from Pearson and Kendall's coefficient results from the fact that there are many other factors that cause the Air Pollution and rise in AQI levels. Thus, Windspeed plays a minor role in lowering the AQI levels.

Summary & Suggestions

- Ho Chi Minh City AQI data provided for the month of December for 5 consecutive years starting from **2016 to 2020**. For year **2021**, February month data is provided
- Gathered external data for consisting of **weather** parameters such as **Visibility, Humidity, Temperature, Wind Speed and Weather Description** via **API** written in **Python** to carry out the analysis if any weather metric has influence on Air Pollution
- Treated the missing and incorrect values in the data after descriptive analysis using **backfill** methods and appropriate mean/median technique
- AQI distribution in the data is such that **60%** of the data is categorized between **Good - Moderate AQI** and **40%** of the data is categorized as **Unhealthy AQI**
- **Peak hours** where pollution is at the **unhealthy** end are the morning rush hours ranging from morning **8:00 AM to 10:00 AM**
- Weather parameter such as **Visibility, Humidity, Temperature and Weather Description** do **not** have any **influence** on the Air Pollution. Major reasons being data availability only for the month of December, because of which most of the data points are between the **same range**. There is **not much variability** in the data to make any inferences.
- **Wind Speed** is identified as one variable which seem to have **impact** on the **Air Quality**. With increase in Windspeed, AQI shifts from **Unhealthy to Moderate-Good**
- The **Pearson correlation** test indicates that there is **no correlation** between the AQI and weather parameters **except for Wind Speed**.
- The **Scatterplot** between AQI and Windspeed displays **negative linear relationship** for **higher values** of windspeed. However, the **strength** of relationship is **quite low** to be used as a predictor of AQI.
- The **side-by-side bar** graph between AQI and weather parameters look **almost alike except for Wind Speed** which shows that a relationship does exist between AQI and Windspeed.
- The **Contingency tables** also show a similar result wherein the **proportions of frequencies** between AQI Categories and Weather Parameter categories are **almost same** except for higher range of values of weather parameters that have greater deviations from expected values.
- The **Chi-Square** statistics also shows a **highly significant** value above the decision point for **AQI vs Windspeed** with a very **low p-value** indicating **significant relationship**. However, for other weather parameters, relationship must be confirmed with a **strength analysis**.
- **Kendall's rank** correlation quantifies the **strength** of relationships between **AQI and Weather** parameters is **very low** to be considered for any kind of regression. **The highest tau value is 0.162 (for Windspeed)** which is too low and close to zero.

Appendix

❖ Sas Code

DATA DESCRIPTION

```
libname TP "/folders/myfolders/SASLAB/Group Project/Data";

/*_____________________________________________________________________*/
/*                              IMPORT                                 */
/*_____________________________________________________________________*/

PROC IMPORT DATAFILE="/folders/myfolders/SASLAB/Group
Project/Data/HoChiMinhCity_PM2.5_2016_12_MTD.csv"
                    OUT=TP.HCMN_2016_12
                    DBMS=CSV
                    REPLACE;
RUN;


PROC IMPORT DATAFILE="/folders/myfolders/SASLAB/Group
Project/Data/HoChiMinhCity_PM2.5_2017_12_MTD.csv"
                    OUT=TP.HCMN_2017_12
                    DBMS=CSV
                    REPLACE;
RUN;



PROC IMPORT DATAFILE="/folders/myfolders/SASLAB/Group Project/Data/
HoChiMinhCity_PM2.5_2018_12_MTD.csv"
                    OUT=TP.HCMN_2018_12
                    DBMS=CSV
                    REPLACE;
RUN;

PROC IMPORT DATAFILE="/folders/myfolders/SASLAB/Group
Project/Data/HoChiMinhCity_PM2.5_2019_12_MTD.csv"
                    OUT=TP.HCMN_2019_12
                    DBMS=CSV
                    REPLACE;
RUN;
```

```
PROC IMPORT DATAFILE="/folders/myfolders/SASLAB/Group
Project/Data/HoChiMinhCity_PM2.5_2020_12_MTD.csv"
                        OUT=TP.HCMN_2020_12
                        DBMS=CSV
                        REPLACE;
RUN;


PROC IMPORT DATAFILE="/folders/myfolders/SASLAB/Group Project/Data/
HoChiMinhCity_PM2.5_2021_02_MTD.csv"
                        OUT=TP.HCMN_2021_02
                        DBMS=CSV
                        REPLACE;
RUN;


/*_____________________________________________________________*/
/*                    DATA DESCRIPTION                         */
/*_____________________________________________________________*/

PROC CONTENTS
        data=tp.hcmn_2016_12;
RUN;


PROC CONTENTS
        data=tp.hcmn_2017_12;
RUN;


PROC CONTENTS
        data=tp.hcmn_2018_12;
RUN;


PROC CONTENTS
        data=tp.hcmn_2019_12;
RUN;


PROC CONTENTS
        data=tp.hcmn_2020_12;
RUN;
```

```
PROC CONTENTS
        data=tp.hcmn_2021_02;

RUN;

/*_____________________________________________________________*/
/*                        TOP 5 ROWS                           */
/*_____________________________________________________________*/

PROC PRINT
        data=tp.hcmn_2016_12
        (OBS=5);
RUN;

PROC PRINT
        data=tp.hcmn_2017_12
        (OBS=5);
RUN;

PROC PRINT
        data=tp.hcmn_2018_12
        (OBS=5)
RUN;

PROC PRINT
        data=tp.hcmn_2019_12
        (OBS=5);
RUN;

PROC PRINT
        data=tp.hcmn_2020_12
        (OBS=5);
RUN;

PROC PRINT
        data=tp.hcmn_2021_02
        (OBS=5);
RUN;


/*_____________________________________________________________*/
/*                          END                                */
/*_____________________________________________________________*/
```

DATA PREPROCESSING

```
*-------------------------------------------------------------------------------------------------------------*;

**********Creating a New Library for Hochiminh Pollution Data**********;

*-------------------------------------------------------------------------------------------------------------*;

libname GP "/folders/myfolders/Demo/DANA Team Project/GP_Lib";

options datestyle=dmy;

*-------------------------------------------------------------------------------------------------------------*;

FILENAME HCMC2016 "/folders/myfolders/Demo/DANA Team
Project/HoChiMinhCity_PM2.5_2016_12_MTD.csv";

FILENAME HCMC2017 "/folders/myfolders/Demo/DANA Team
Project/HoChiMinhCity_PM2.5_2017_12_MTD.csv";

FILENAME HCMC2018 "/folders/myfolders/Demo/DANA Team
Project/HoChiMinhCity_PM2.5_2018_12_MTD.csv";

FILENAME HCMC2019 "/folders/myfolders/Demo/DANA Team
Project/HoChiMinhCity_PM2.5_2019_12_MTD.csv";

FILENAME HCMC2020 "/folders/myfolders/Demo/DANA Team
Project/HoChiMinhCity_PM2.5_2020_12_MTD.csv";

FILENAME HCMC2021 "/folders/myfolders/Demo/DANA Team
Project/HoChiMinhCity_PM2.5_2021_02_MTD.csv";


*-------------------------------------------------------------------------------------------------------------*;

**********IMPORTING DATA FOR ALL YEARS FROM 2016 to 2021**********;

*-------------------------------------------------------------------------------------------------------------*;

proc import datafile=HCMC2016
            dbms=csv
            out=GP.HoChiMinh2016
            replace;
```

```
        guessingrows=MAX;
run;
proc import datafile=HCMC2017
                dbms=csv
                out=GP.HoChiMinh2017
                replace;
        guessingrows=MAX;
run;
proc import datafile=HCMC2018
                dbms=csv
                out=GP.HoChiMinh2018
                replace;
        guessingrows=MAX;
run;
proc import datafile=HCMC2019
                dbms=csv
                out=GP.HoChiMinh2019
                replace;
        guessingrows=MAX;
run;
proc import datafile=HCMC2020
                dbms=csv
                out=GP.HoChiMinh2020
                replace;
        guessingrows=MAX;
```

```
run;
proc import datafile=HCMC2021
                    dbms=csv
                    out=GP.HoChiMinh2021
                    replace;
          guessingrows=MAX;
run;
*-------------------------------------------------------------------------------------------------------*;
************************* ANALYSING DATASETS ************************;
*-------------------------------------------------------------------------------------------------------*;
*-------------------------------Analysing categorical variables----------------------------------*;
proc freq data=GP.HoChiMinh2016;
          tables Site QC_Name AQI_Category Parameter Conc__Unit Duration / nocum;
          title "Frequencies for Categorical Variables HCMC 2016";
run;
proc freq data=GP.HoChiMinh2017;
          tables Site QC_Name AQI_Category Parameter Conc__Unit Duration / nocum;
          title "Frequencies for Categorical Variables HCMC 2017";
run;
proc freq data=GP.HoChiMinh2018;
          tables Site QC_Name AQI_Category Parameter Conc__Unit Duration / nocum;
          title "Frequencies for Categorical Variables HCMC 2018";
run;
proc freq data=GP.HoChiMinh2019;
          tables Site QC_Name AQI_Category Parameter Conc__Unit Duration / nocum;
```

```
        title "Frequencies for Categorical Variables HCMC 2019";
run;
proc freq data=GP.HoChiMinh2020;
        tables Site QC_Name AQI_Category Parameter Conc__Unit Duration / nocum;
        title "Frequencies for Categorical Variables HCMC 2020";
run;
proc freq data=GP.HoChiMinh2021;
        tables Site QC_Name AQI_Category Parameter Conc__Unit Duration / nocum ;
        title "Frequencies for Categorical Variables HCMC 2021";
run;

*----------------------------------Analysing numerical variables---------------------------------;
proc means data=GP.HoChiMinh2016 n nmiss min mean median max std;
        var Raw_Conc_ _24_hr__Midpoint_Avg__Conc_ AQI;
        title "Descriptive Statistics for Numeric Variables HCMC 2016";
run;
proc means data=GP.HoChiMinh2017 n nmiss min mean median max std;
        var Raw_Conc_ NowCast_Conc_ AQI;
        title "Descriptive Statistics for Numeric Variables HCMC 2017";
run;
proc means data=GP.HoChiMinh2018 n nmiss min mean median max std;
        var Raw_Conc_ NowCast_Conc_ AQI;
        title "Descriptive Statistics for Numeric Variables HCMC 2018";
run;
proc means data=GP.HoChiMinh2019 n nmiss min mean median max std;
```

```
        var Raw_Conc_ NowCast_Conc_ AQI;
        title "Descriptive Statistics for Numeric Variables HCMC 2019";
run;
proc means data=GP.HoChiMinh2020 n nmiss min mean median max std;
        var Raw_Conc_ NowCast_Conc_ AQI;
        title "Descriptive Statistics for Numeric Variables HCMC 2020";
run;
proc means data=GP.HoChiMinh2021 n nmiss min mean median max std;
        var Raw_Conc_ NowCast_Conc_ AQI;
        title "Descriptive Statistics for Numeric Variables HCMC 2021";
run;
*-------------------------------------------------------------------------------------------------------*;
**************************CLEANING DATASETS**************************;
*-------------------------------------------------------------------------------------------------------*;
*---------------------------------------Handling 2016 Dataset---------------------------------------*;
*-----------------------------Removing Invalid and Negative Values-----------------------------*;
data work.HoChiMinh2016_intermediate;
        set GP.HoChiMinh2016(where=(QC_Name ^= 'Invalid'));
        drop _24_hr__Midpoint_Avg__Conc_;
run;
*-----------------------------------------Fixing Categories-----------------------------------------*;
data work.HoChiMinh2016_aqi_category_fixed;
        set work.HoChiMinh2016_intermediate;
        format AQI_Category_Name $40.;
        if AQI_Category=1 then
```

```
            AQI_Category_Name="Good";
        if AQI_Category=2 then
            AQI_Category_Name="Moderate";
        if AQI_Category=3 then
            AQI_Category_Name="Unhealthy for Sensitive Groups";
        if AQI_Category=4 then
            AQI_Category_Name="Unhealthy";
        if AQI_Category=5 then
            AQI_Category_Name="Very Unhealthy";
        if AQI_Category=6 then
            AQI_Category_Name="Hazardous";
        drop AQI_Category;
        rename AQI_Category_Name=AQI_Category;
run;
*--------------------------------------Calculating NowCast 2016----------------------------------*;
proc sql;
        create table work.calc_min_max_conc as
        select Date__LT_,
            (select max(Raw_Conc_)
             from work.HoChiMinh2016_aqi_category_fixed as b
             where intnx('hour', a.Date__LT_, -11, 'b') le b.Date__LT_ le a.Date__LT_) as
Max_Conc,
                (select min(Raw_Conc_)
                 from work.HoChiMinh2016_aqi_category_fixed as c
                 where intnx('hour', a.Date__LT_, -11, 'b') le c.Date__LT_ le a.Date__LT_) as
Min_Conc
```

```
        from work.HoChiMinh2016_aqi_category_fixed as a;
create table work.calc_min_max_weight as
        select *, case
                when Min_Conc/Max_Conc < 0.5 then 0.5
                else Min_Conc/Max_Conc
            end as weight
        from work.calc_min_max_conc;
        create table work.HoChiMinh2016_fixed as
        select a.*,
            (select ROUND(SUM(case
                        when intck('hour', b.Date__LT_, a.Date__LT_)=0
                        then b.Raw_Conc_
                        when intck('hour', b.Date__LT_, a.Date__LT_)=1
                        then b.Raw_Conc_ * (c.weight** 1)
                        when intck('hour', b.Date__LT_, a.Date__LT_)=2
                        then b.Raw_Conc_ * (c.weight** 2)
                                when intck('hour', b.Date__LT_, a.Date__LT_)=3
                                then b.Raw_Conc_ * (c.weight** 3)
                                when intck('hour', b.Date__LT_, a.Date__LT_)=4
                                then b.Raw_Conc_ * (c.weight** 4)
                                when intck('hour', b.Date__LT_, a.Date__LT_)=5
                                then b.Raw_Conc_ * (c.weight** 5)
                                when intck('hour', b.Date__LT_, a.Date__LT_)=6
                                then b.Raw_Conc_ * (c.weight** 6)
                                when intck('hour', b.Date__LT_, a.Date__LT_)=7
```

```
            then b.Raw_Conc_ * (c.weight** 7)
            when intck('hour', b.Date__LT_, a.Date__LT_)=8
            then b.Raw_Conc_ * (c.weight** 8)
            when intck('hour', b.Date__LT_, a.Date__LT_)=9
            then b.Raw_Conc_ * (c.weight** 9)
            when intck('hour', b.Date__LT_, a.Date__LT_)=10
            then b.Raw_Conc_ * (c.weight**10)
            when intck('hour', b.Date__LT_, a.Date__LT_)=11
            then b.Raw_Conc_ * (c.weight**11)
            else b.Raw_Conc_
        end)/SUM(case
            when intck('hour', b.Date__LT_, a.Date__LT_)=0
            then 1
            when intck('hour', b.Date__LT_, a.Date__LT_)=1
            then c.weight** 1
            when intck('hour', b.Date__LT_, a.Date__LT_)=2
            then c.weight** 2
            when intck('hour', b.Date__LT_, a.Date__LT_)=3
            then c.weight** 3
            when intck('hour', b.Date__LT_, a.Date__LT_)=4
            then c.weight** 4
            when intck('hour', b.Date__LT_, a.Date__LT_)=5
            then c.weight** 5
            when intck('hour', b.Date__LT_, a.Date__LT_)=6
            then c.weight** 6
```

```
                when intck('hour', b.Date__LT_, a.Date__LT_)=7
                then c.weight** 7
                when intck('hour', b.Date__LT_, a.Date__LT_)=8
                then c.weight** 8
                when intck('hour', b.Date__LT_, a.Date__LT_)=9
                 then c.weight** 9
                when intck('hour', b.Date__LT_, a.Date__LT_)=10
                then c.weight**10
                when intck('hour', b.Date__LT_, a.Date__LT_)=11
                then c.weight**11
                else 1
            end),
        0.1)
    from work.HoChiMinh2016_aqi_category_fixed as b
    where b.Date__LT_ between intnx('hour', a.Date__LT_, -11, 'b') and
    a.Date__LT_) as NowCast_Conc_
from
work.HoChiMinh2016_aqi_category_fixed as a
inner join work.calc_min_max_weight as c on a.Date__LT_=c.Date__LT_;
quit;
*-------------------------------------------------------------------------------------------------------*;
**********Merging Datasets to Create Master Dataset**********;
*-------------------------------------------------------------------------------------------------------*;

data GP.HoCHiMinh2016_2021_Master;
```

```
        set work.HoChiMinh2016_fixed GP.hochiminh2017 GP.hochiminh2018
GP.hochiminh2019 GP.hochiminh2020 GP.hochiminh2021;

        if QC_Name = 'Invalid' then delete;

run;

*----------------------------------------Analyze categorical variables-------------------------------;

proc freq data=GP.HoCHiMinh2016_2021_Master;

        tables Site QC_Name AQI_Category Parameter Conc__Unit Duration / nocum ;

        title "Frequencies for Categorical Variables HCMC 2016-2021";

run;

*----------------------------------------Analyze numerical variables---------------------------------;

proc means data=GP.HoCHiMinh2016_2021_Master n nmiss min mean median max std;

        var Raw_Conc_ AQI NowCast_Conc_;

        title "Descriptive Statistics for Numeric Variables HCMC 2016-2021";

run;

*---------------------------------Removing Invalid and Negative Values-------------------------;

data work.HoChiMinh_master_intermediate;

        set GP.HoCHiMinh2016_2021_Master;

        if Raw_Conc_ < 0 then Raw_Conc_ = '.';

        if AQI < 0 then AQI = '.';

        if NowCast_Conc_ < 0 then NowCast_Conc_='.';

run;

*----------------------------------------Back Filling Macro-----------------------------------------*;

%macro backfill(variable);

retain help_&variable;

if not missing(&variable)
```

```
then help_&variable = &variable;

else &variable = help_&variable;

drop help_&variable;

%mend;

data work.hochiminh2016_2021_missing_fixed;

set work.HoChiMinh_master_intermediate;

%backfill(Raw_Conc_);

%backfill(AQI);

%backfill(NowCast_Conc_);

run;

data GP.hochiminh2016_2021_cleaned;

set work.hochiminh2016_2021_missing_fixed;

if QC_Name='Missing' and Raw_Conc_^='.' and Raw_Conc_ >= 0 then QC_Name='Valid';

if AQI >= 0   and AQI <= 50  and AQI_Category='N/A' then AQI_Category = 'Good';

if AQI >= 51  and AQI <= 100 and AQI_Category='N/A' then AQI_Category = 'Moderate';

if AQI >= 101 and AQI <= 150 and AQI_Category='N/A' then AQI_Category = 'Unhealthy for
Sensitive Groups';

if AQI >= 151 and AQI <= 200 and AQI_Category='N/A' then AQI_Category = 'Unhealthy';

if AQI >= 201 and AQI <= 300 and AQI_Category='N/A' then AQI_Category = 'Very
Unhealthy';

if AQI >= 301 and AQI <= 500 and AQI_Category='N/A' then AQI_Category = 'Hazardous';

run;

*-------------------------------------------------------------------------------------------------------*;

**********Analysing Cleaned Master Dataset**********;

*-------------------------------------------------------------------------------------------------------*;

*---------------------------------Analysing Categorical Variables---------------------------------*;
```

```
title "Frequencies for Categorical Variables of Cleaned Dataset";
proc freq data=GP.hochiminh2016_2021_cleaned;
        tables QC_Name AQI_Category Year/ nocum plots=(freqplot);
run;
title "Frequencies for Categorical Variables of Cleaned Dataset";
proc freq data=GP.hochiminh2016_2021_cleaned;
        tables  Site Parameter Duration Conc__Unit/ nocum;
run;
*----------------------------------------Analyze numerical variables-------------------------------;
title "Descriptive Statistics for Numeric Variables of Cleaned Dataset";
proc means data=GP.hochiminh2016_2021_cleaned n nmiss min mean median max std;
        var NowCast_Conc_ AQI Raw_Conc_;
run;
title;
*---------------------------------------- Analysing Issues ---------------------------------------;
*-------------------------------------------------999 Values-----------------------------------------;
proc print data=GP.hochiminh2016_2021_cleaned label n;
where Raw_Conc_=-999 or AQI=-999 or NowCast_Conc_=-999;
run;
*-------------------------------------Missing QC_Name | Conc Values---------------------------;
proc print data=GP.hochiminh2016_2021_cleaned(where=(QC_Name=" or
QC_Name='Missing'));
var  Date__LT_ QC_Name Raw_Conc_ AQI  NowCast_Conc_;
run;
*-------------------------------------------------------------------------------------------------------*;
```

```
**************************Export Master Dataset**************************;

*-------------------------------------------------------------------------------------------------------*;

proc export data=GP.hochiminh2016_2021_cleaned

outfile="/folders/myfolders/DANA/Group_Project/HoChiMinhCity_PM2.5_2017_2021_Master
_MTD.csv" dbms=csv replace;

run;
```

❖ R Code

Descriptive Analysis

```
Install libraries for data wrangling , binning and visualization
install.packages("ggplot2")
install.packages("dplyr")
install.packages("OneR")

loading libraries for data wrangling , binning and visualization
library(dplyr)
library(ggplot2)
library(OneR)

reading the master dataset for Ho Chi Minh City combined for all years, path D:/Semester 1/DANA/Team
Project
df <- read.csv("HoChiMinhCity_PM2.5_2017_2021_Master_MTD.csv")
summary(df)

Check if there are any null values in the dataframe
any(is.na((df)))

year and month wise distribution
df %>%
  group_by(Year,Month) %>%
  summarise(Total_Count  = n())

Remove Jan month value from the dataset as there is only one value from the month of Jan from
2018,2019,2020 and 2021 which is not significant for the analysis
```

```
df= subset(df,Month != 1)
df %>%
   group_by(Year,Month) %>%
   summarise(Total_Count  = n())

list("Summary" = summary(df$AQI),
    "Standard Deviation" = sd(df$AQI, na.rm=TRUE),
    "Range" = range(df$AQI, na.rm=TRUE),
    "Inter-Quartile Range(IQR)" = IQR(df$AQI, na.rm=TRUE))
```

Check AQI distribution in the data

```
df %>%
   group_by(AQI_Category) %>%
   summarise(Total_Count  = n()) %>% mutate(Perc_Contri=
round(Total_Count/sum(Total_Count)*100,2)) %>%arrange(desc(Total_Count))
```

AQI distribution for the overall data

```
hist(df$AQI,
    main = "Histogram for AQI",
    xlab = "AQI",
    freq = TRUE,
   # probability = TRUE,
    breaks = 10,
    border = "Blue",
    col = "Turquoise",
    labels = TRUE,
    las=1   ,
    ylim = c(0,1200),
    xlim = c(0,200))
```

YOY AQI distribution ,coord_flip()

```
df$Year_New = factor(df$Year, order = TRUE, labels = c("2016","2017", "2018", "2019", "2020", "2021"))
ggplot(df, aes(x = Year_New, y = AQI, fill= Year_New) )+
   geom_boxplot() +
    stat_summary(fun.y = mean,
      geom = "point",
      size = 2,
      color = "red")+
   theme_classic()
```

Bar plot for AQI Category

```
df %>%
```

```
   group_by(AQI_Category) %>%
   summarise(Total_Count  = n()) %>% mutate(Perc_Contri=
round(Total_Count/sum(Total_Count)*100,2))  %>%
   ggplot(aes(x = reorder(AQI_Category, -Total_Count), y = Total_Count))+ geom_col(fill="pink",color =
"brown")+
   geom_text(aes(label = Total_Count), vjust = -0.2,size = 4)+
    labs(x = "AQI Category",
         y = "Total Count",
         title = "Data Distribution on AQI Category")+
  theme(axis.text.x = element_text(angle = 0),plot.title = element_text(hjust = 0.5))
```

Proportion distribution of AQI Category on yearly basis

```
df %>%
   group_by(Year,AQI_Category) %>%
   summarise(Total_Count  = n())  %>%
 ggplot(aes(x = Year, y = Total_Count,fill= AQI_Category))+ geom_col(color = "brown", width
=0.5,position = "fill")+
  scale_fill_brewer(palette = "Dark2")+
   theme(axis.text.x = element_text(angle = 0),plot.title = element_text(hjust = 0.5))
```

Check AQI distribution for Pre and Post COVID period

Year 2020 has only ~20% of the data with AQI categorized as Unhealthy which is lowest amongst all years

```
df %>% subset(Year %in% c(2019,2020,2021)) %>%
   group_by(Year,AQI_Category) %>%
   summarise(Total_Count  = n())  %>%
 ggplot(aes(x = Year, y = Total_Count,fill= AQI_Category))+ geom_col(color = "brown", position = "fill")+
  scale_fill_brewer(palette = "Set3"
```

Check hourly distribution of AQI Catgeory in the data

```
df %>%
   group_by(Hour,AQI_Category) %>%
   summarise(Total_Count  = n()) %>%
  ggplot(aes(x = Hour, y = Total_Count,fill= AQI_Category))+ geom_col(color = "brown", width
=0.5,position = "fill")+
  scale_fill_brewer(palette = "Dark2")
```

```

Above chart shows the peak hours where AQI Category = Unhealthy for more than 50% of the data

Hours = 8,9,10,11,12,13

Late night and early morning 75% of the data has AQI in the range of Good to Moderate (Hours = 21,22,23,0,1,2)
```

Merging the given data with external data gathered for weather paramters

```
df_weather <- read.csv("D:/Semester 1/DANA/Team Project/weather.csv")
View(df_weather)
new_df <- merge(df,df_weather,by = c("Day","Month","Year","Hour"))
```

Checking distribution on Weather Description

```
new_df %>%
  group_by(weatherDesc) %>%
  summarise(Total_Count  = n()) %>% mutate(Perc_Contri=
round(Total_Count/sum(Total_Count)*100,2)) %>%arrange(desc(Total_Count))
```

Clubbing related weather description for further analysis

```
new_df$weatherDescNew <- ifelse(new_df$weatherDesc %in% c("Partly
cloudy","Clear","Sunny","Patchy rain possible","Light rain shower","Cloudy","Moderate or heavy rain
shower"),new_df$weatherDesc , "Others")
new_df %>%
  group_by(weatherDescNew) %>%
  summarise(Total_Count  = n()) %>% mutate(Perc_Contri=
round(Total_Count/sum(Total_Count)*100,2)) %>%arrange(desc(Total_Count))
```

Relation between AQI Category and Weather Description

```
new_df %>%
  group_by(weatherDescNew,AQI_Category) %>%
  summarise(Total_Count  = n()) %>% mutate(Perc_Contri=
round(Total_Count/sum(Total_Count)*100,2)) %>%
 ggplot(aes(reorder(x = weatherDescNew, -Total_Count), y = Total_Count,fill= AQI_Category))+
 geom_bar( stat="identity")+
    labs(x = "Weather Description",
      y = "Total Count",
      title = "Data Distribution on Weather Description")+
  scale_fill_brewer(palette = "Set2")
```

AQI Category proportion distribution on Weather Description

```
new_df %>%
  group_by(weatherDescNew,AQI_Category) %>%
  summarise(Total_Count  = n()) %>% mutate(Perc_Contri=
round(Total_Count/sum(Total_Count)*100,2)) %>%
 ggplot(aes(reorder(x = weatherDescNew, -Total_Count), y = Total_Count,fill= AQI_Category))+
 geom_bar( position = "fill",stat="identity",width =0.7)+
    labs(x = "Weather Description",
      y = "Proportion",
      title = "Data Distribution on Weather Description")+
```

```
  scale_fill_brewer(palette = "Set2")
```

No significant relation between AQI and Weather Description

Check effect of windspeed on AQI, Histogram for WindSpeed

```
hist(new_df$windspeedKmph,
    main = "Histogram for Wind Speed",
    xlab = "Windspeed",
    freq = TRUE,
    #probability = TRUE,
    breaks = 10,
    border = "Blue",
    col = "Turquoise",
    labels = TRUE,
    las=1   ,
    ylim = c(0,1200))
```

Box Plot for WindSpeed

```
boxplot(new_df$windspeedKmph,
      las=1,
      main = "Box Plot for WindSpeed",
      #xlab = "Number of quantitues sold",
      ylab = "WindSpeed",
      col = "Turquoise",
      border = "blue",
      horizontal = F,
      outline = T)
```

Binning WindSpeed data points into relevant categories using content method which gives intervals of equal content via quantiles ,

```
new_df$windspeed_Cat <- bin( new_df$windspeedKmph, nbins = 3,method = "content",label =
c("Low","Medium","High"))
new_df %>% group_by(windspeed_Cat) %>%
  summarise(Total_Count  = n()) %>% mutate(Perc_Contri=
round(Total_Count/sum(Total_Count)*100,2))
```

AQI CAtegory distribution over WindSpeed

```
new_df %>% group_by(windspeed_Cat,AQI_Category) %>%
  summarise(Total_Count  = n()) %>% mutate(Perc_Contri=
round(Total_Count/sum(Total_Count)*100,2)) %>%
 ggplot(aes(x = windspeed_Cat, y = Total_Count,fill= AQI_Category))+
 geom_bar(stat="identity",width =0.4, color = "Dark Green")+
     labs(x = "Wind Speed",
       y = "Total Count",
```

```
        title = "Data Distribution on Wind Speed")+
  scale_fill_brewer(palette = "PiYG", direction = -1)
```

WindSpeed proportion distribution over AQI ACtegory

```
new_df %>% group_by(windspeed_Cat,AQI_Category) %>%
  summarise(Total_Count  = n()) %>% mutate(Perc_Contri=
round(Total_Count/sum(Total_Count)*100,2)) %>%
 ggplot(aes(x = windspeed_Cat, y = Total_Count,fill= AQI_Category))+
 geom_bar(position = "fill",stat="identity",width =0.4, color = "Dark Green")+
     labs(x = "Wind Speed",
       y = "Proportion",
       title = "Data Distribution on Wind Speed")+
  scale_fill_brewer(palette = "PiYG", direction = -1)
```

Above figure shows that as the wind speed increases AQI improves. For wind speed categorized under "High", 76% of the #data has AQI as Low and Medium

Data Distribution for Visibility Variable

```
new_df %>%
   group_by(visibility) %>%
   summarise(Total_Count  = n()) %>% mutate(Perc_Contri=
round(Total_Count/sum(Total_Count)*100,2)) %>%
 ggplot(aes(x = visibility, y = Total_Count))+
 geom_bar( stat="identity", fill = "Light Yellow", color = "Brown")+
  geom_text(aes(label=Total_Count), vjust=-0.3, size=3.5)+
     labs(x = "Visibility",
       y = "Total Count",
       title = "Data Distribution on Visibility") +
 theme_minimal()
```

Check effect of humidity on Air Quality, Histogram for Humidity

```
hist(new_df$humidity,
    main = "Histogram for Humidity",
    xlab = "Humidity",
    freq = TRUE,
    #probability = TRUE,
    breaks = 10,
    border = "Turquoise",
    col = "Pink",
    labels = TRUE,
    las=1   ,
    ylim = c(0,800),
```

```
    xlim = c(20,100))
```

Binning Humidity Variable for analysis

labels = c("Very Low", "Low", "Medium", "High", "Very High")

```
 new_df$humidity_cat <- cut( new_df$humidity,  c(-Inf, 30, 60, Inf), labels = c("Dry","Normal","Humid"))
new_df %>% group_by(humidity_cat) %>%
  summarise(Total_Count  = n()) %>% mutate(Perc_Contri=
round(Total_Count/sum(Total_Count)*100,2))
```

Proportion Distribution for Humidity VAribable for AQI Distribution

```
new_df %>% group_by(humidity_cat,AQI_Category) %>%
  summarise(Total_Count  = n()) %>% mutate(Perc_Contri=
round(Total_Count/sum(Total_Count)*100,2)) %>%
 ggplot(aes(x = humidity_cat, y = Total_Count,fill= AQI_Category))+
 geom_bar(position = "fill",stat="identity",width =0.4, color = "Dark Green")+
     labs(x = "Humidity",
       y = "Proportion",
       title = "Data Distribution on Humidity")+
  scale_fill_brewer(palette = "Set1", direction = -1)
```

No impact of humidity on Air Quality

Check effect of temperature on Air Quality

```
hist(new_df$tempC,
    main = "Histogram for Temperature (°C)",
    xlab = "Temperature (°C)",
    freq = TRUE,
    #probability = TRUE,
    breaks = 10,
    border = "Dark Blue",
    col = "light blue",
    labels = TRUE,
    las=1   ,
    ylim = c(0,1300))
```

temperature scale is in the normal moderate range, hence would not influence AQI

Box Plot for Temperature

```
boxplot(new_df$tempC,
      las=1,
      main = "Box Plot for Temperature (°C)",
      #xlab = "Number of quantitues sold",
      ylab = "Temperature (°C)",
      col = "light blue",
```

```
    border = "dark blue",
    horizontal = F,
    outline = T)
```

Exporting the final dataset

```
setwd("D:/Semester 1/DANA/Team Project")
write.csv(new_df,"Final_Merged_Dataset.csv",row.names = FALSE)
```

INFERENTIAL ANALYSIS

Libraries

```
library(readxl)
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.
3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflict
s() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(viridisLite)
library(viridis)
```

Import Merged Data Set

```
master<- read.csv('Final_Merged_Dataset.csv')
```

Import External Data Set

```
weather<- read.csv('weather.csv')
weather$day<- as.Date(weather$day)
```

External Data Set Description

```
str(weather)

size<-dim(weather)

print("Dimensions")
print("--------------------")
```

```
print(paste("Rows = ",size[1]))
print(paste("Variables = ",size[2]))
print("--------------------")

head(weather)
```

INFERENTIAL ANALYSIS

AQI vs Temperature

```
ggplot(master, aes(x=tempC, y=AQI)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) +
  xlab("Temperature in Degree Celcius")+
  scale_x_continuous()

## `geom_smooth()` using formula 'y ~ x'

cor.test(master$AQI, master$tempC)
```

AQI vs Humidity

```
ggplot(master, aes(x=humidity, y=AQI)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) +
  xlab("Humidity in percentage % ")+
  scale_x_continuous()

## `geom_smooth()` using formula 'y ~ x'

cor.test(master$AQI, master$humidity)
```

AQI vs visibility

```
ggplot(master, aes(x=visibility, y=AQI)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) +
  xlab("Visibility in kilometers")+
  scale_x_continuous()

## `geom_smooth()` using formula 'y ~ x'

cor.test(master$AQI, master$visibility)
```

AQI vs Windspeed

```
ggplot(master, aes(x=windspeedKmph, y=AQI)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) +
  xlab("Windspeed in Kilometer per hour")+
  scale_x_continuous()

## `geom_smooth()` using formula 'y ~ x'

cor.test(master$AQI, master$windspeedKmph)
```

CATegorical Analysis

```
master$AQI_Category <- factor(master$AQI_Category, ordered = TRUE, levels= c(
'Good', 'Moderate', 'Unhealthy for Sensitive Groups', 'Unhealthy'))

master$windspeed_Cat <- factor(master$windspeed_Cat, ordered = TRUE, levels=
c('Low', 'Medium', 'High'))

master$humidity_cat <- factor(master$humidity_cat, ordered = TRUE, levels= c(
'Dry', 'Normal', 'Humid'))

master$weatherDesc<- as.factor(ifelse(master$weatherDesc %in% c('Clear', 'Sun
ny'), 'Clear', ifelse(master$weatherDesc %in% c('Partly cloudy', 'Cloudy', 'P
atchy rain possible', 'Fog', 'Mist', 'Overcast'), 'Cloudy', 'Rainy')))
```

AQI vs Windspeed

```
t1 <- table(master$AQI_Category,master$windspeed_Cat)
t1
chisq.test(t1)
cor.test(y=as.numeric(master$AQI_Category), x=as.numeric(master$windspeed_Cat
),method = "kendall")

master %>%
  group_by(AQI_Category,windspeed_Cat) %>%
  summarise(Total_count = n()) %>%
  ggplot(aes(x = AQI_Category,y=Total_count,fill = windspeed_Cat))+
  geom_bar(position = 'dodge',stat = 'identity') +
  coord_flip()

## `summarise()` has grouped output by 'AQI_Category'. You can override using
the `.groups` argument.
```

AQI vs Humidity

```
master_temp <- droplevels(subset(master, humidity_cat != 'Dry'))
t1 <- master_temp %>% with(table(AQI_Category,humidity_cat))
t1
chisq.test(t1)
cor.test(y=as.numeric(master_temp$AQI_Category), x=as.numeric(master_temp$hum
idity_cat),method = "kendall")

master_temp %>%
  group_by(AQI_Category,humidity_cat) %>%
  summarise(Total_count = n()) %>%
  ggplot(aes(x = AQI_Category,y=Total_count,fill = humidity_cat))+
  geom_bar(position = 'dodge',stat = 'identity') +
  coord_flip()

## `summarise()` has grouped output by 'AQI_Category'. You can override using
the `.groups` argument.
```

AQI vs Weather Description

```
t1 <- table(master$AQI_Category,master$weatherDesc)
t1
chisq.test(t1)
cor.test(y=as.numeric(master$AQI_Category), x=as.numeric(master$weatherDesc),
method = "kendall")

master %>%
  group_by(AQI_Category,weatherDesc) %>%
  summarise(Total_count = n()) %>%
  ggplot(aes(x = weatherDesc,y=Total_count,fill = AQI_Category))+
  geom_bar(position = 'dodge',stat = 'identity') +
  coord_flip()
```

```
## `summarise()` has grouped output by 'AQI_Category'. You can override using
the `.groups` argument.
```

PRE AND POST COVID ANALYSIS

```
year_2019 <- subset(master, Year == 2019,select = c(Day,Hour,AQI))
year_2019$DT<- paste(as.character(year_2019$Day), as.character(year_2019$Hour
),sep = ':')
year_2020 <- subset(master, Year == 2020,select = c(Day,Hour,AQI))
year_2020$DT<- paste(as.character(year_2020$Day), as.character(year_2020$Hour
),sep = ':')

cdf<- merge(year_2019,year_2020,by="DT")
```

T test for AQI of Years 2019 and 2020

```
t.test(cdf$AQI.x,cdf$AQI.y ,paired = TRUE, alternative = "two.sided")
```

Yearly Analysis

```
master %>%
  ggplot( aes(x=as.factor(Year), y=AQI, fill=as.factor(Year))) +
    geom_boxplot() +
    scale_fill_viridis(discrete = TRUE, alpha=0.6, option="A") +
    labs(
    x = "Year",
    y = "AQI",
    fill = "Year",
    title = "Yearly Comparison of AQI"
        ) +
  scale_y_continuous(breaks = seq(0,200,20))
```

Box Plot Pre and Post Covid

```
subset(master, Year %in% c(2019,2020,2021)) %>%
  ggplot( aes(x=as.factor(Year), y=AQI, fill=as.factor(Year))) +
    geom_boxplot() +
```

```
    scale_fill_viridis(discrete = TRUE, alpha=0.6, option="A") +
    labs(
    x = "Year",
    y = "AQI",
    fill = "Year",
    title = "Pre and Post Covid Analysis of AQI"

    )+
  scale_y_continuous(breaks = seq(0,200,20))
```

Peak hours

```
master %>%
  ggplot( aes(x=as.factor(Hour), y=AQI, fill=as.factor(Hour))) +
    geom_boxplot() +
    scale_fill_viridis(discrete = TRUE, alpha=0.6, option="A") +
    labs(
    x = "Hour",
    y = "AQI",
    fill = "Hour"
    )+
  scale_y_continuous(breaks = seq(0,200,20)) +
  stat_summary(fun.y=mean, geom="point",color = '#DE3163')
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

Peak days

```
master %>%
  ggplot( aes(x=as.factor(Day), y=AQI, fill=as.factor(Day))) +
    geom_boxplot() +
    scale_fill_viridis(discrete = TRUE, alpha=0.6, option="A") +
    labs(
    x = "Day",
    y = "AQI",
    fill = "Day"
    )+
  scale_y_continuous(breaks = seq(0,200,20)) +
  stat_summary(fun.y=mean, geom="point",color = '#DE3163')
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

AQI and Weather Description

```
master %>%
    group_by(weatherDesc,AQI_Category) %>%
    summarise(Total_Count  = n()) %>% mutate(Perc_Contri= round(Total_Count/s
um(Total_Count)*100,2)) %>%
  ggplot(aes(reorder(x = weatherDesc, -Total_Count), y = Total_Count,fill= AQ
I_Category))+
  geom_bar( position = "fill",stat="identity",width =0.7)+
       labs(x = "Weather Description",
```

```
          y = "Proportion",
          title = "Data Distribution on Weather Description")+
   scale_fill_brewer(palette = "Set2")
```

```
## `summarise()` has grouped output by 'weatherDesc'. You can override using
the `.groups` argument.
```

```
## No significant relation between AQI and Weather Description
```

Correlation Summary

```
ss <- subset(master,select = c(AQI,tempC,humidity,visibility,windspeedKmph))
corr<-cor(x= ss[c(-1)],y=ss$AQI)
cdf <- data.frame(corr)
ggplot(cdf, aes(x= corr, y= row.names(cdf))) +
  geom_bar(stat = 'identity')+
  xlab("Correlation") +
  ylab("Weather parameters") +
  scale_x_continuous(breaks = seq(-0.5,0.5,0.05))
```

Visibility frequency bar plot

```
ggplot(master, aes(visibility)) +
  geom_bar(fill = "#E69F00")+
  scale_x_continuous(breaks = seq(1,10,1))
```