Analyzing Air Quality of Ho Chi Minh City

Objectives

- 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 conditions and air quality
- Answering research questions based on analysis and test results

Data Description

Ho Chi Minh Data Sets

- ✓ Hourly Concentrations of PM2.5 in Ho Chi Minh
- ✓ 6 datasets from year 2016 to 2021
- ✓ December month for 2016-2020 and February month for 2021
- ✓ Year 2016 has 24-hour midpoint average instead of Now Cast concentration
- ✓ 14 variables in each data set
- ✓ Variables of Interest :
  - AQI
  - AQI_Category
  - Now Cast Concentration
  - Raw Concentration
  - Data and Time

2016

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
8	AQI	Num	8	BEST12.	BEST32.
9	AQI_Category	Num	8	BEST12.	BEST32.
12	Conc__Unit	Char	5	$5.	$5.
3	Date__LT_	Num	8	DATETIME.	ANYDTDTM40.
6	Day	Num	8	BEST12.	BEST32.
13	Duration	Char	4	$4.	$4.
7	Hour	Num	8	BEST12.	BEST32.
5	Month	Num	8	BEST12.	BEST32.
2	Parameter	Char	17	$17.	$17.
14	QC_Name	Char	5	$5.	$5.
11	Raw_Conc_	Num	8	BEST12.	BEST32.
1	Site	Char	16	$16.	$16.
4	Year	Num	8	BEST12.	BEST32.
10	_24_hr__Midpoint_Avg__Conc_	Num	8	BEST12.	BEST32.

Others

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
9	AQI	Num	8	BEST12.	BEST32.
10	AQI_Category	Char	30	$30.	$30.
12	Conc__Unit	Char	5	$5.	$5.
3	Date__LT_	Num	8	DATETIME.	ANYDTDTM40.
6	Day	Num	8	BEST12.	BEST32.
13	Duration	Char	4	$4.	$4.
7	Hour	Num	8	BEST12.	BEST32.
5	Month	Num	8	BEST12.	BEST32.
8	NowCast_Conc_	Num	8	BEST12.	BEST32.
2	Parameter	Char	17	$17.	$17.
14	QC_Name	Char	5	$5.	$5.
11	Raw_Conc_	Num	8	BEST12.	BEST32.
1	Site	Char	16	$16.	$16.
4	Year	Num	8	BEST12.	BEST32.

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

Weather Dataset

- ✓ Data fetched from ‘worldweatheronline.com’ API using Python Client
- ✓ There are 4392 observations and 7 variables
- ✓ Hourly Data

```
'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 ...
```

day <date>	time <int>	tempC <int>	windspeedKmph <int>	weatherDesc <chr>	humidity <int>	visibility <int>
2016-12-01	0	26	5	Clear	89	10
2016-12-01	1	26	6	Clear	89	10
2016-12-01	2	25	7	Clear	90	10
2016-12-01	3	25	8	Clear	90	10
2016-12-01	4	25	8	Clear	90	10
2016-12-01	5	25	9	Clear	89	10

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

Data Preprocessing

Issues with HCMC 2016

- ✓ Dataset for Year 2016 is handled separately to align its structure with others
- ✓ AQI Categories are converted from numerical to Character Categories as per the definitions provided by US EPA.
- ✓ Visible outlier (985) corresponds to invalid Data based on QC_Name. There are 2 such invalid values which are eliminated.
- ✓ Values of Now Cast Concentration Value and AQI are evaluated using Raw concentration Values as per the formula.
- ✓ No Null values in Numerical Data Variables

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

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

$$w^* = 1 - \frac{c_{max} - c_{min}}{c_{max}} = \frac{c_{min}}{c_{max}}$$

and let

$$w = \begin{cases} w^* & \text{if } w^* > \frac{1}{2} \\ \frac{1}{2} & \text{if } w^* \leq \frac{1}{2} \end{cases}$$

With these definitions the PM NowCast[2] is given by:

$$NowCast = \frac{\sum_{i=1}^{12} w^{i-1} c_i}{\sum_{i=1}^{12} w^{i-1}}.$$

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

Parameter	Frequency	Percent
PM2.5 - Principal	720	100.00

Conc__Unit	Frequency	Percent
UG/M3	720	100.00

Duration	Frequency	Percent
1 Hr	720	100.00

Handling Master Data Set

- ✓ All datasets are merged to create a Master Dataset
- ✓ Negative Values of Raw Conc., AQI and NowCast Conc.(-999) are replaced with (.)
- ✓ Calculated NowCast, AQI and Raw_Conc using Backfilling Method.
- ✓ Evaluated AQI Categories Based on AQI Value to replace 'N/A'
- ✓ After handling missing data, QC_Name was marked as 'Valid'
- ✓ Eliminated 4 observations belonging to January month.

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

```
      AQI            Raw_Conc_       NowCast_Conc_
 Min.   : 21.00   Min.   :  0.00   Min.   :  5.10
 1st Qu.: 75.00   1st Qu.: 22.00   1st Qu.: 22.80
 Median : 91.00   Median : 30.00   Median : 30.50
 Mean   : 97.46   Mean   : 34.15   Mean   : 34.11
 3rd Qu.:118.00   3rd Qu.: 43.00   3rd Qu.: 42.10
 Max.   :176.00   Max.   :161.00   Max.   :129.60
'data.frame':   4113 obs. of  14 variables:
```

Research Questions

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

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

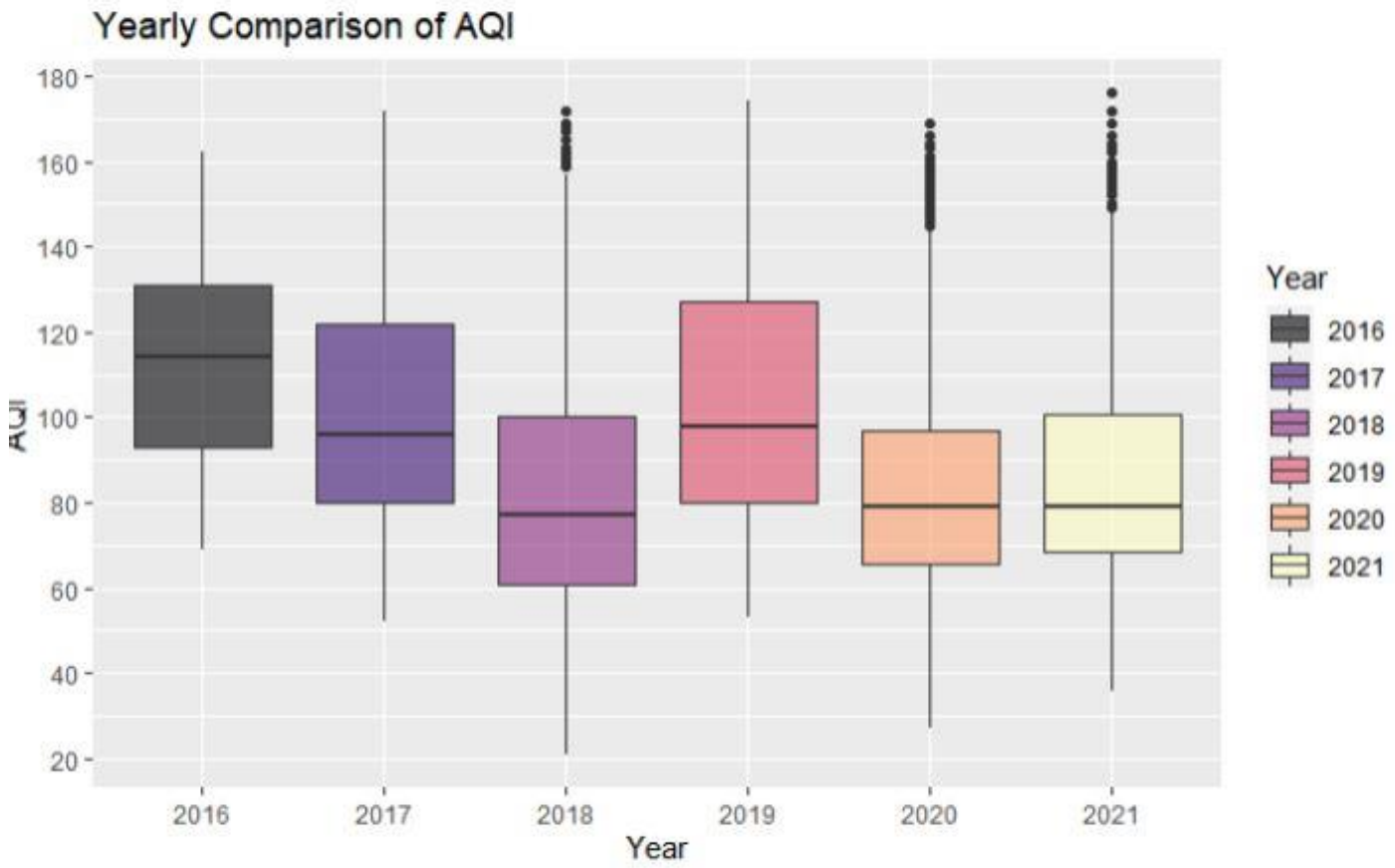
What are the peak air pollution hours in a day over the 6 years period?

What is the effect of variations in Humidity, Visibility, Temperature and Weather Description on Air quality ?

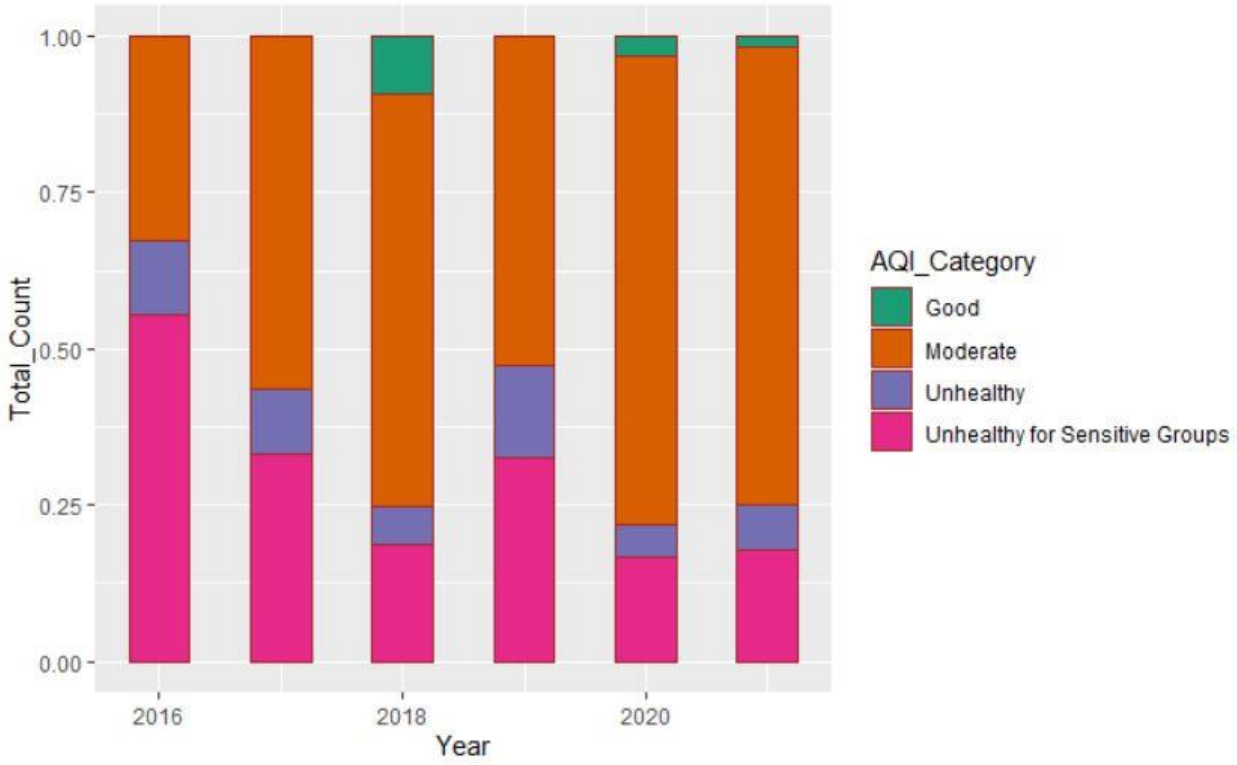
What is the effect of variations in Wind Speed on Air Pollution?

Change in Air Quality over the period of 6 years

- ✓ Highest AQI value in 2016
- ✓ Gradually reduced till 2018
- ✓ Rise in 2019
- ✓ Again, reduced for the years 2020 and 2021



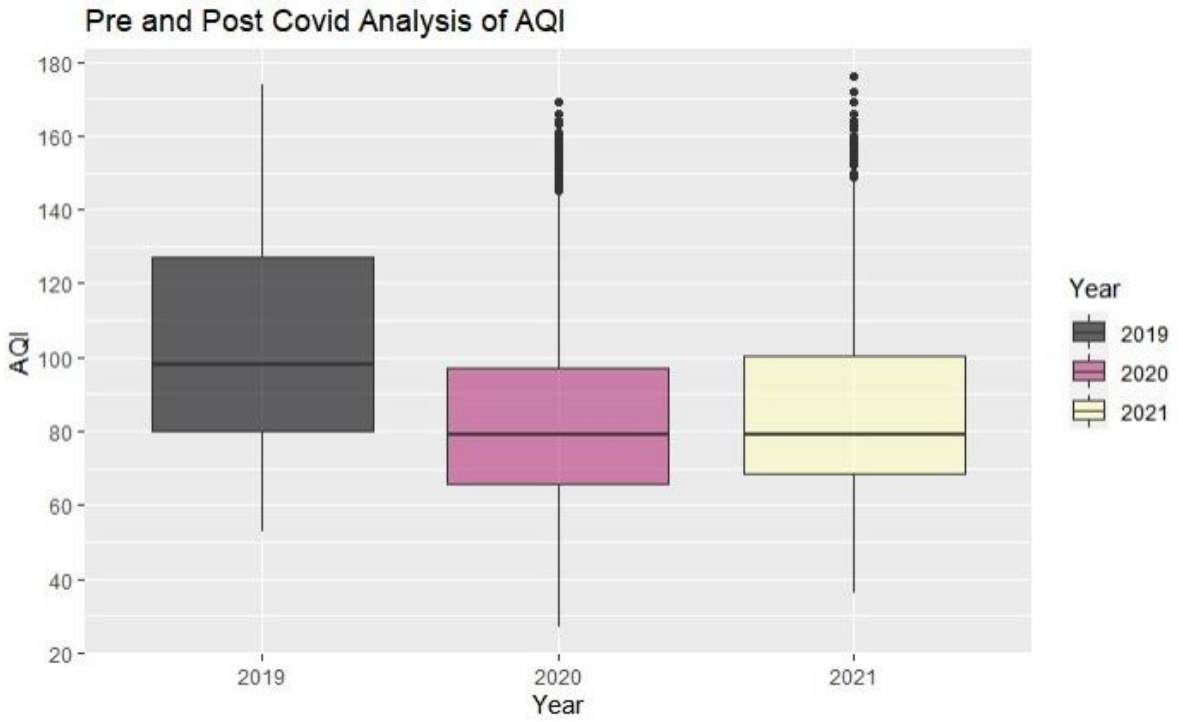
Proportion Distribution of AQI Categories

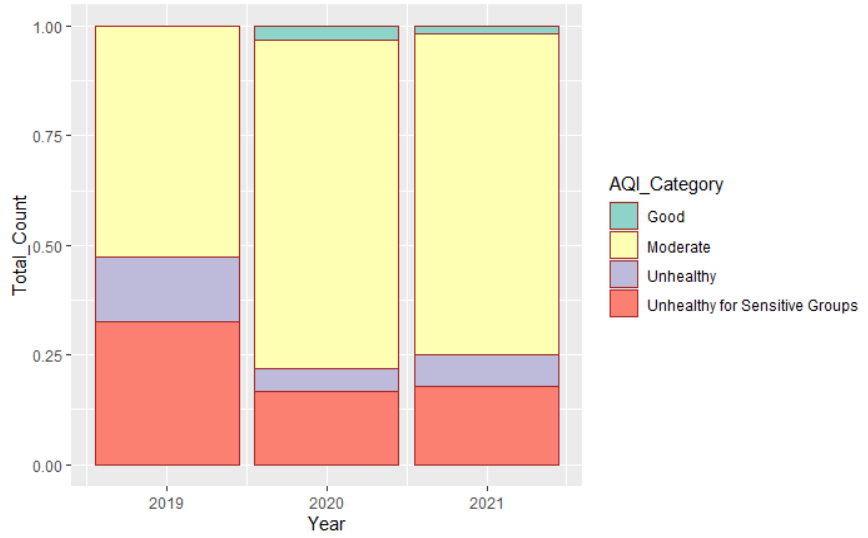


Year	Unhealthy AQI %
2016	70%
2017	40%
2018	25%
2019	47%
2020	20%
2021	25%

Pre – Post Covid Air Quality Analysis

- ✓ Box Plot Analysis illustrates reduction in AQI levels from the Year 2019 to 2021.
- ✓ This might be a result of lesser traffic due to imposed restrictions.
- ✓ Mosaic plot evidently displays decrease in proportions of unhealthy AQI categories over the given period.





- ✓ Paired T-test between AQI levels of years 2019 and 2020
- ✓ Significant P value resulting in rejection of Null Hypothesis
- ✓ Mean AQI for December month of the years 2019 and 2020 are significantly different with a T value of 13.627.
- ✓ It can be inferred that pollution decreased post Covid – 19 Lockdown.

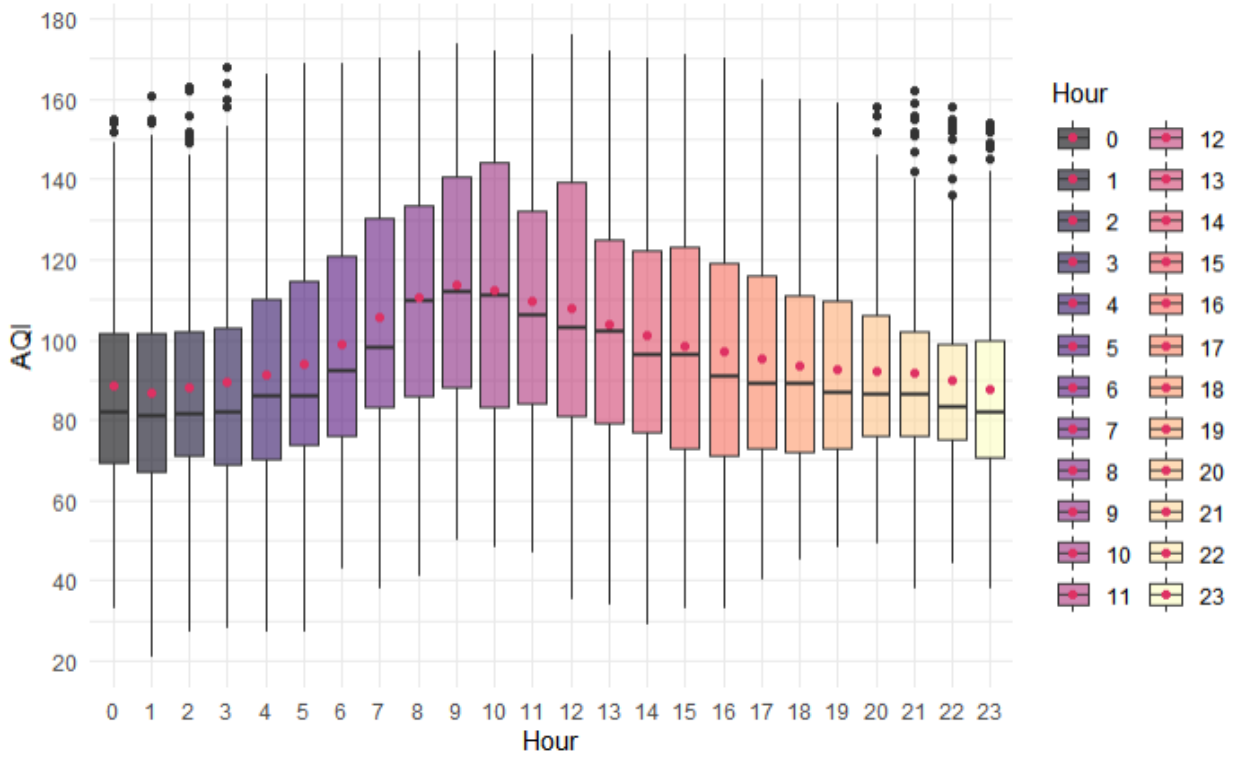
```
        Paired t-test

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
```

H_0	True difference in means is equal to 0
H_A	True difference in means is not equal to 0

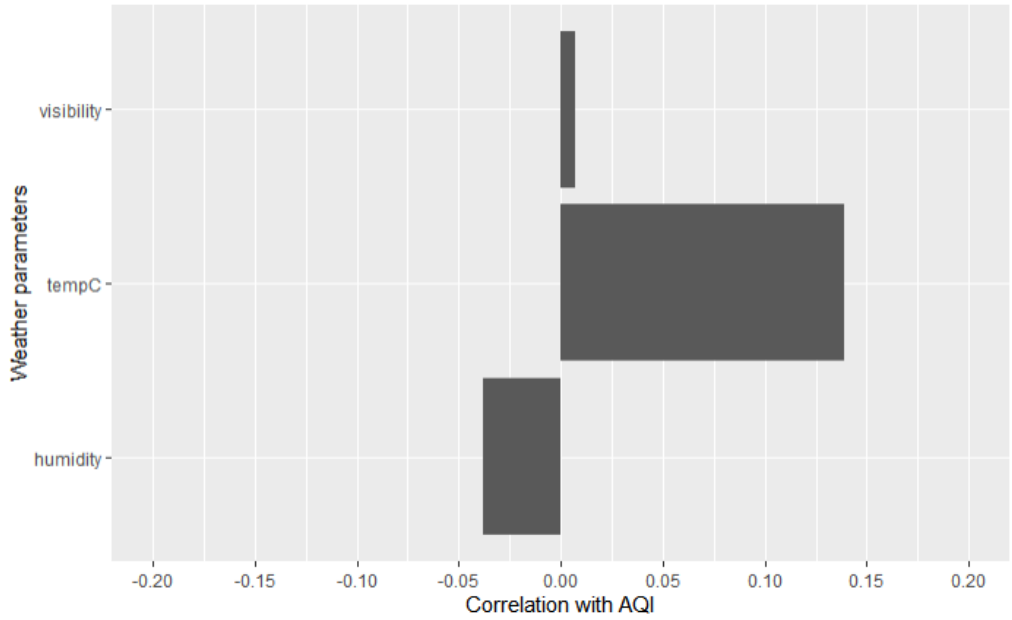
Peak Pollution Hours in a Day

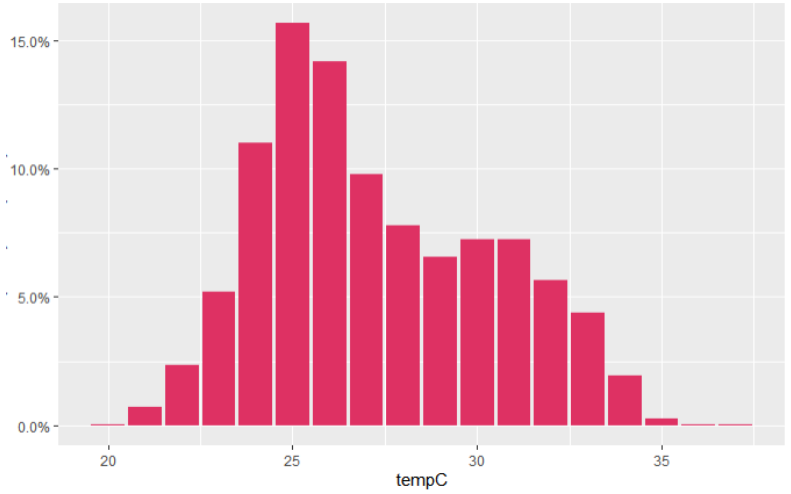
- ✓ Peak Hours of pollution for December Month is in the morning between 8:00 AM and 10: 00 AM.
- ✓ Air Quality gradually increases as the day wears in.
- ✓ Air Quality is best between 10:00 PM to 3:00 AM
- ✓ Air Quality decreases during the rush hours in the morning due to heavy vehicular traffic, increase of domestic activities and local emission sources.
- ✓ Weather experts say that pollution levels are the highest early in the day because of low wind movement, and thus PM 2.5 particles get trapped in the air.

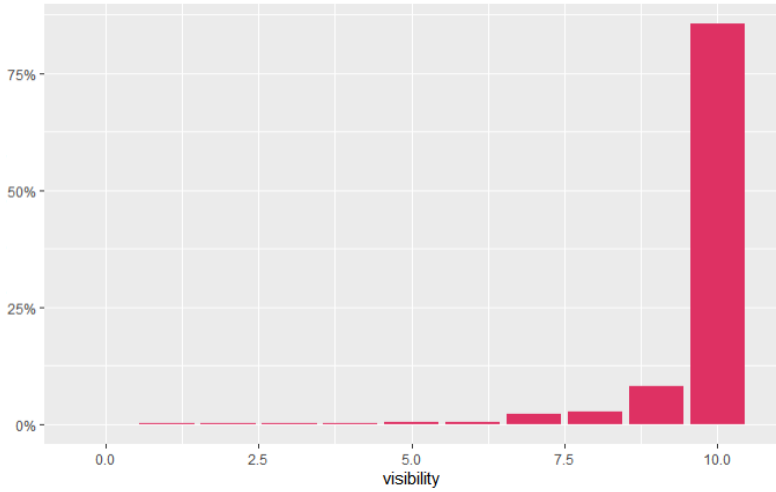


Effect of Weather Conditions on Air Quality

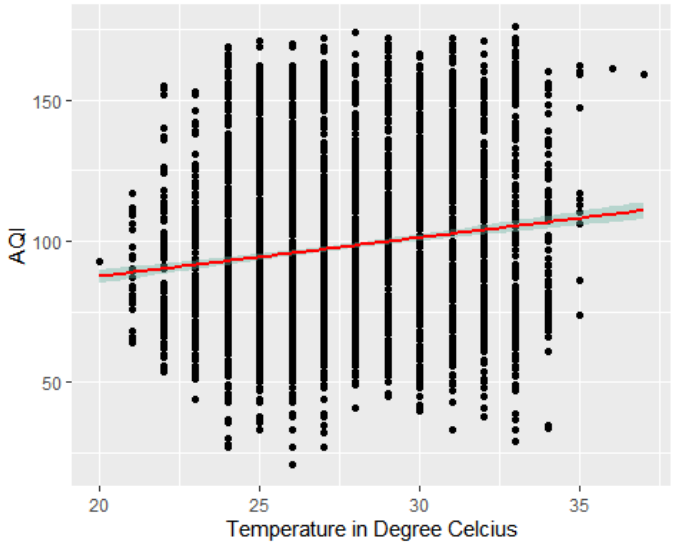
- ✓ No Linear correlations between AQI and Humidity/Temperature/Visibility
- ✓ Based on given Data, Humidity, Temperature and Visibility do not have much impact on Air quality.
- ✓ Approximately 85% of the values of visibility are fixed to 10 Kms
- ✓ Approximately 50 % of Temperature values are between 24 and 27 Degree Celsius
- ✓ Since data is available for only month of December, most of the data points lie in the same range.
- ✓ There is not much variability in data to make any inferences.

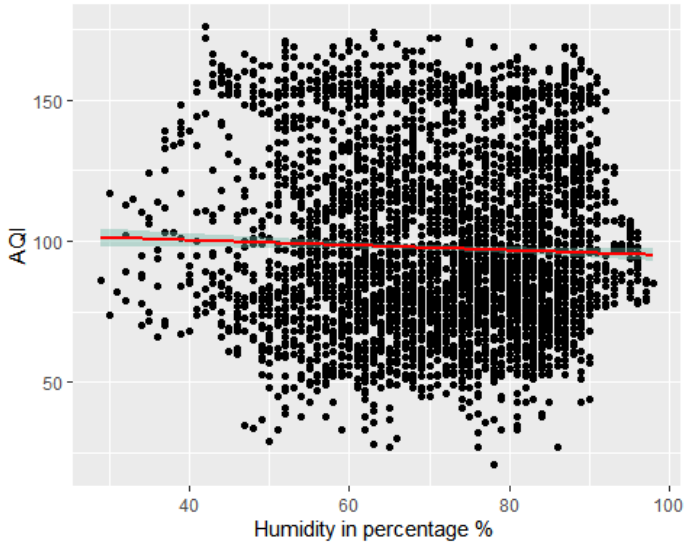


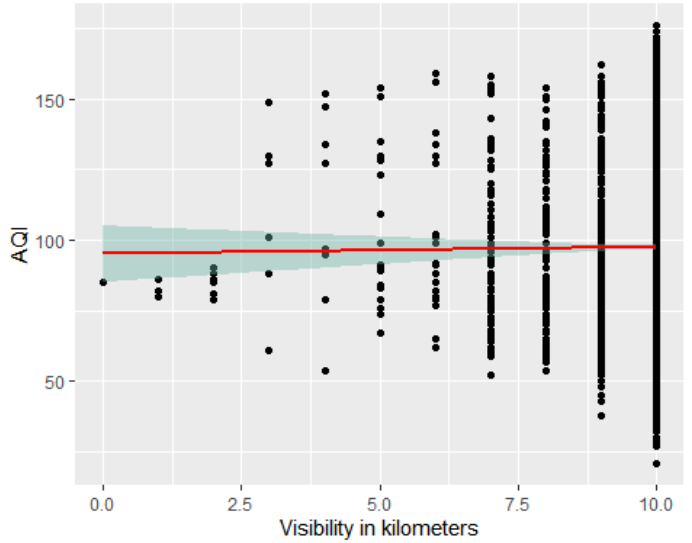




Scatter Plots

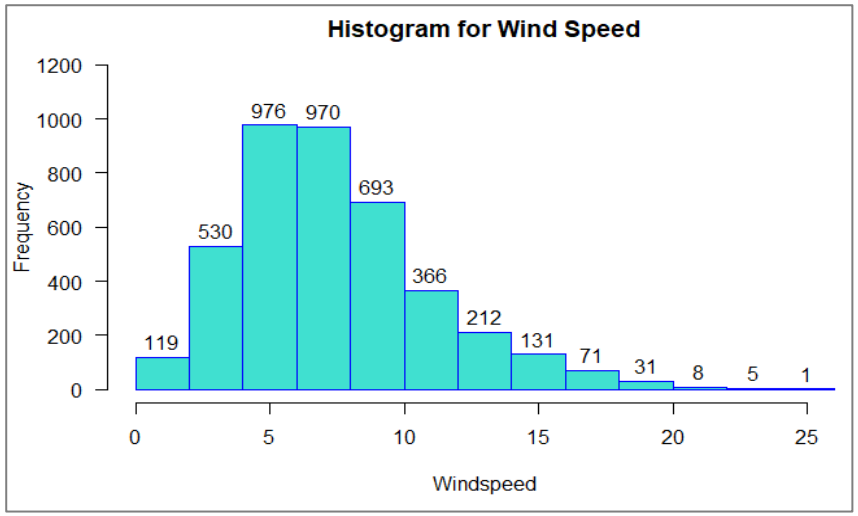




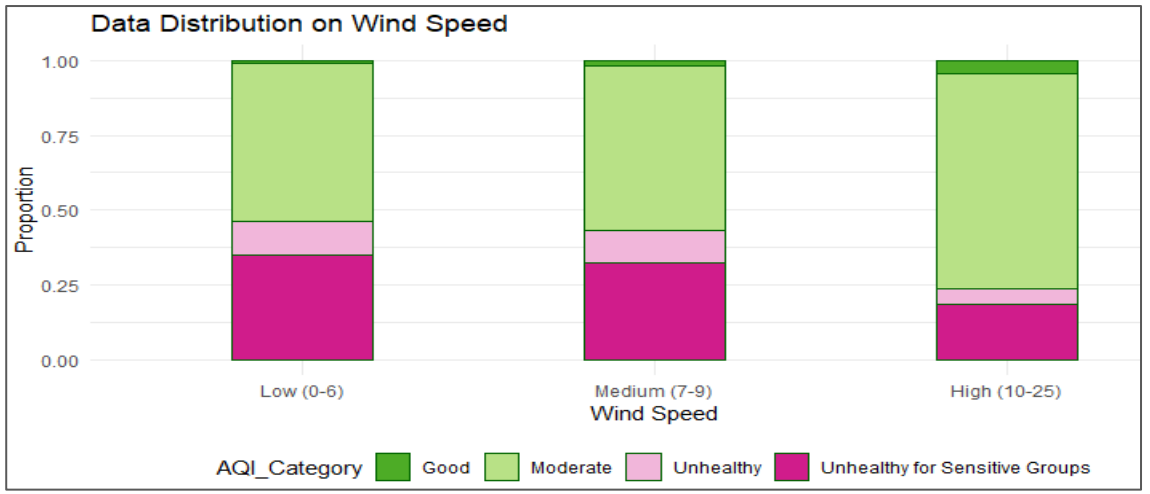


Effect of Variations in Wind Speed on Air Quality

- ✓ In general, higher the wind speed, the more contaminants are dispersed and lower the concentrations, which means ‘clearer the air’
- ✓ Binned wind speed into 3 high level categories basis the Modern Scale and earlier research performed
- ✓ Nearly 30% - 40% data distributed in each bin
- ✓ From the bar chart we see that as the wind speed increases, Air Quality improves. 76% of the data tagged as ‘Good – Moderate’ for High Wind Speed category as opposed to 56% for Medium and 52% for Low Wind Speed categories



Shape : Right Skewed, Unimodal
Spread : Min = 0 , Max = 25

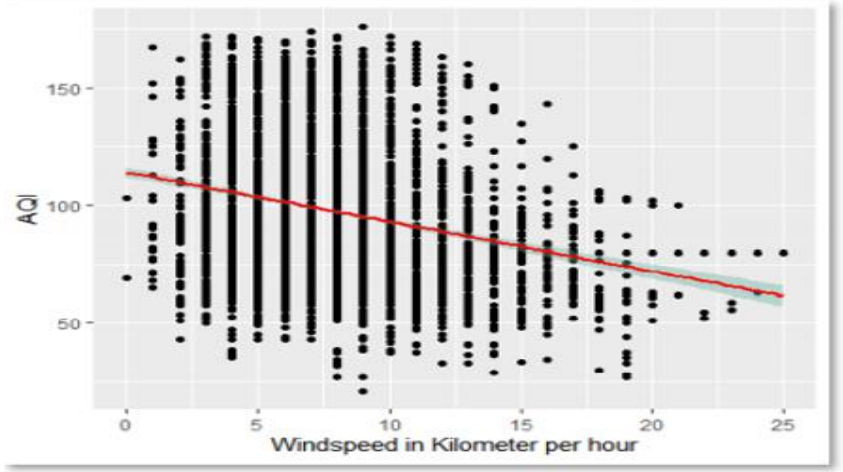


Pearson's Correlation

$

Chi Square Test

Null Hypothesis (H_0)	Wind Speed and Air Quality are independent of each other among all subjects in population
Alternate Hypothesis (H_A)	Wind Speed and Air Quality are related to each other




```
        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
```

Wind Speed is statistically significant in predicting the AQI, and has a weak negative correlation

```
                           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
```

p-value for Chi-Square value is smaller than 0.05. Thus, it can be inferenced that based on Chi-Square Test, the relationship between AQI and Windspeed is statistically Significant.

Conclusions

- ✓ Air Quality Index is the worst from morning 8:00 AM to 10:00 AM but it tends to improve as the day wears in.
- ✓ Post COVID – 19, Air Quality for the Ho Chi Minh city has shown improvement
- ✓ Weather parameters analyzed do not have any significant impact on the Air Quality since the data is collected only for the month of December and there's not much variation in these variables.
- ✓ Wind Speed shows weak negative correlation with AQI.
- ✓ Visibility, Temperature and Humidity variables are spread on almost a fixed range or either concentrated towards specific data point which do not help in interpreting any significant results
- ✓ The results may be different from actual conditions since the analysis is based on limited data.

Suggestions

- ✓ Reduce air pollution exposure in the Ho Chi Minh City by making use of Air Purifiers and Air Quality Monitors
- ✓ During peak pollution hours (8-10am), sensitive groups should limit/avoid outdoor activities.
- ✓ Residents are encouraged to wear good quality masks when outside in order to keep themselves protected from pollution effects
- ✓ It is advisable and safer to go out on windy days
- ✓ Promoting use of EVs (Electric Vehicles) to be an answer to pollution in city centers

Thank You