REPORT OF HCMC AIR POLLUTION

HOCHIMINH CITY AIR POLLUTION 2016 - 2021

April 2021
Team project

TEAM INTRODUCTION

Ayushi Singh

3 years of work experience in SQL programming and support and automation

Kevin Moroso

17 years experience in public policy and business planning.

Lien Pham

12 years of professional experience within data analytics strategy consulting

Sarath Ravikumar

3.5 years of experience as a Machine Learning data Associate

THE REPORT CONSISTS OF 5 PARTS

1 INTRODUCTION
- Introduction of HCMC
- Overview of HCMC's causes of pollution
- Methodology

2 DATA CLEANING
- Process of data cleaning

3 DESCRIPTIVE ANALYSIS
- Analysis of AQI and Raw conc categories and levels
  - 2016 – 2020
  - 2020- 2021

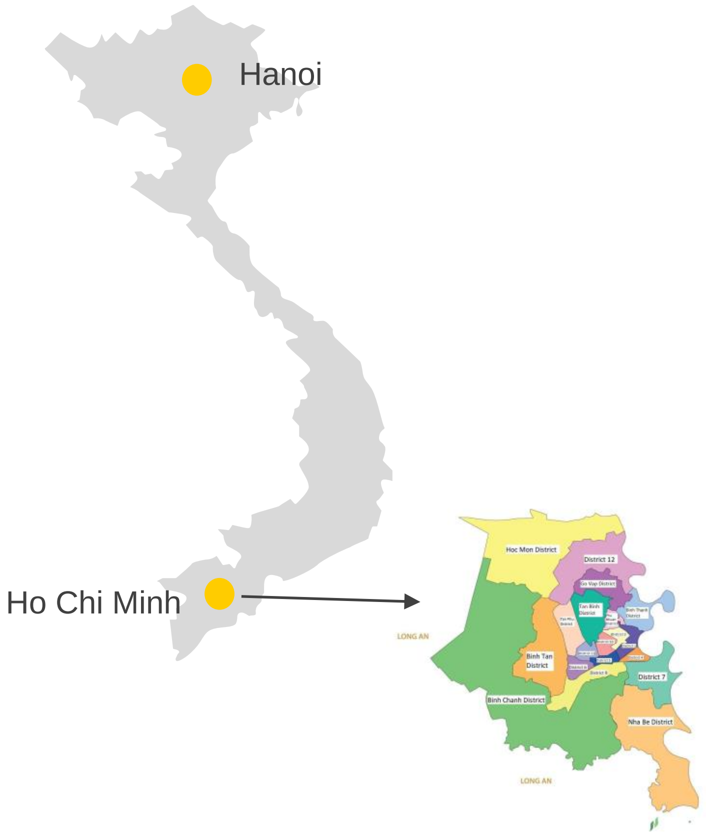
4 INFERENTIAL ANALYSIS
- Correlation methods
- Chi – square
- Linear regression

5 CONCLUSION
- Conclusions of all the analysis results
- Recommendations

BACKGROUND OF HCMC

THE LARGEST CITY IN VIETNAM & FAST IN INDUSTRIALIZATION

MAP OF VIETNAM

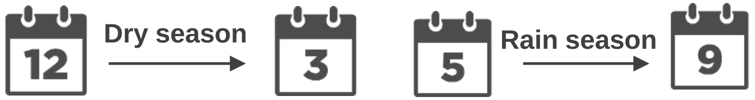


GENERAL & ECONOMICS

- Population of over **8 million** people
- Total area of over **2,095 km2**
- Comprises **19 districts**, District 1, along the Saigon River, where downtown Saigon is located
- Contribute **40%** of the **country's GDP**
- Its main industries includes textiles and garments, footwear, plastics, food processing, electricity, automobiles, electronics, computers, rubber tires and mechanical products
- **Tourism** also plays a very important role, convenient access from other countries by air, by road and by sea

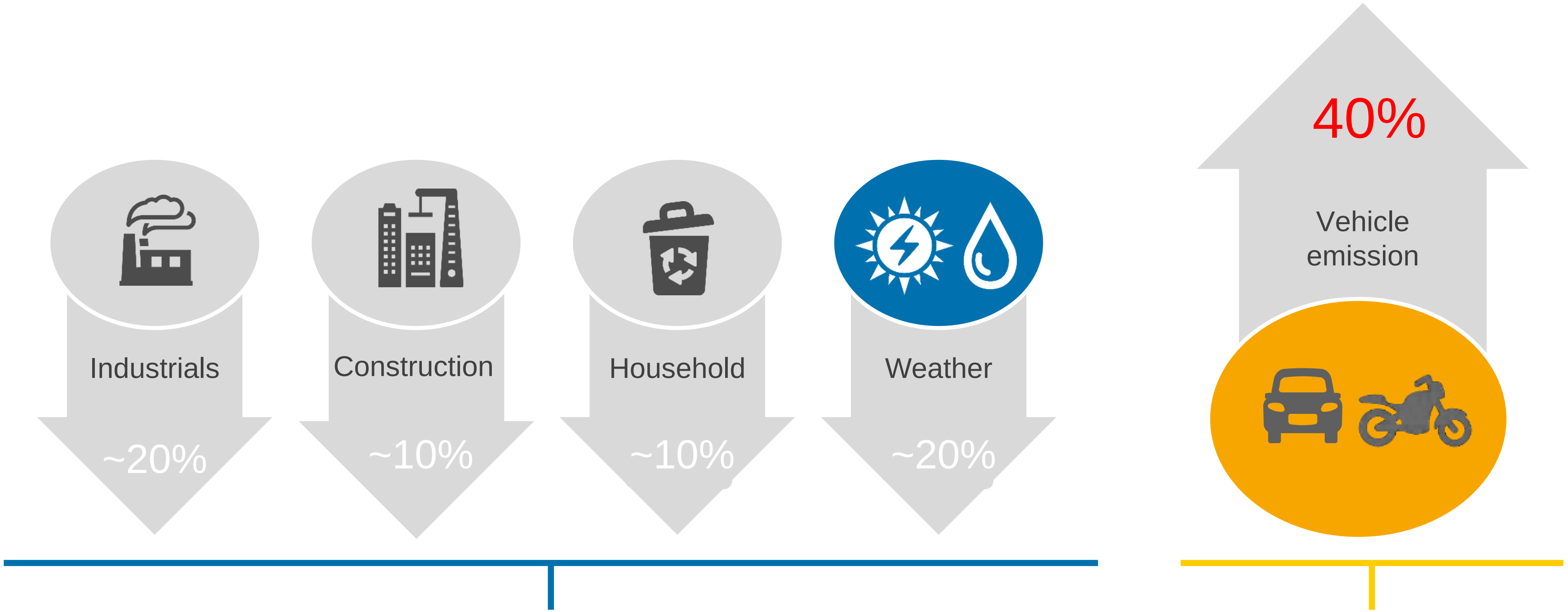
GEOGRAPY AND WEATHER

- There **are 2 distinct seasons:**
  - **Dry season** from **December to March** where Temperature ranges between 21C and 34C
  - **Rainy season** from May to September with monthly rainfall levels of 200mm to 300mm
- The northeast monsoon months from November to April
- The rainy southwest monsoon months of May through October. Humidity levels average 75% throughout the year but are higher during the rainy season



FACTORS THAT CAUSE POLLUTION IN HCMC

POLLUTION IS ATTRIBUTED BY MAINLY VEHICLE EMISSION



Industrial emissions, coal combustion and increasing number of motor vehicles using fossil fuels, dust dispersion and weather conditions are main causes of loss of air quality in urban and industrial areas

Main factor that contributes to pollution

METHODOLOGY

STATISTIC METHOD

Correlation

- The correlation methods for numeric variables and categorical variables will be performed to detect the relationships or associations of the variables

Linear regression

- LM are used to examine the association or relationship. Particularly, we will examine the LMs with **3 days** and **7 days lag**
- Examine the **interactive impact** of two continuous pairs on the dependent variable by multiply these two variables with each other

Chi square

- Chi – square test is used to measure the relationship among categorical variables

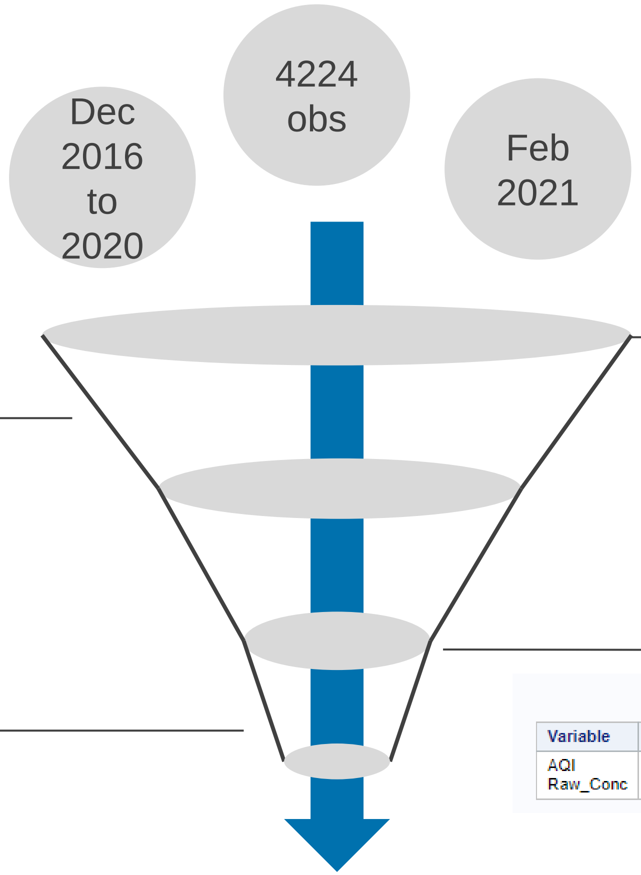
QUESTIONS TO ASK?

Question 1 & 2:
Is air quality getting worse, improving?
How does air quality in HCMC change at different times of the year?

Question 3: Is there a relationship between weather and air quality in Ho Chi Minh City?

Question 4: Are there ways to mitigate the impact of air pollution in Ho Chi Minh City?

DATA CLEANING PROCESS



- The year 2016 did not have AQI category in compliance with the rest of the dataset.
  - They were replaced with appropriate descriptive AQI category

- And also the variable 24 hr. Midpoint was present for 2016 while the other datasets had the variable Now Cast Conc
  - Both the variables were dropped before merging Description

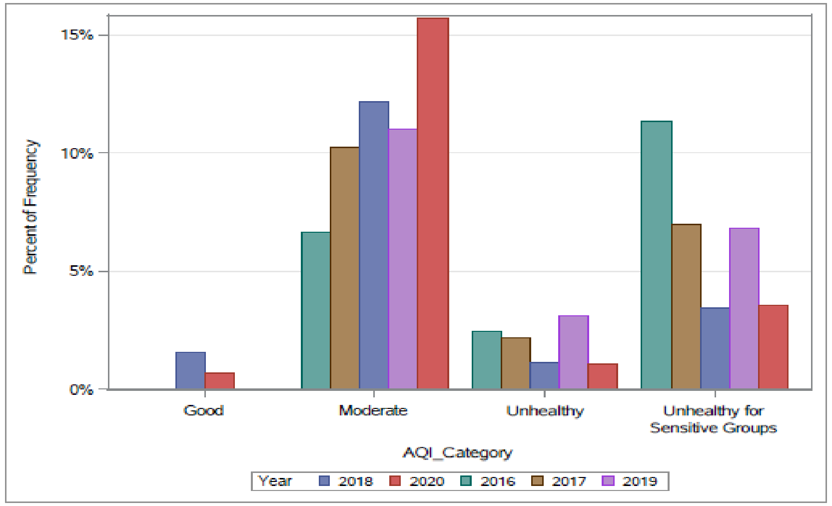
Descriptive Statistics for Numeric Variables

Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
AQI	4224	0	-999.0000000	59.5525568	90.0000000	176.0000000	203.2982446
Raw_Conc	4224	0	-999.0000000	42.2859848	30.0000000	985.0000000	193.6796759

4049 observations

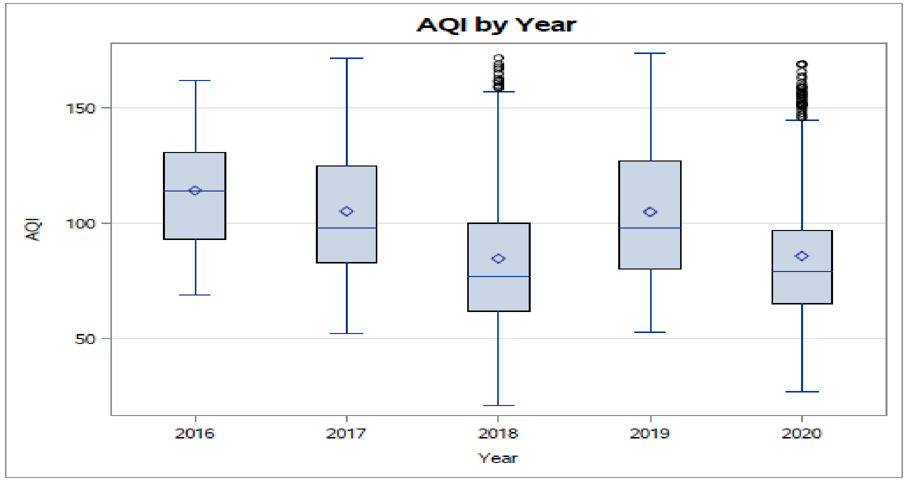
ANNUAL CHANGES IN AQI CATEGORIES AND AQI LEVELS

AQI LEVEL CATEGORIES



- Time periods that are unhealthy and unhealthy for sensitive groups displays a downward trend; time periods with air pollution levels that are good or moderate have increased.
- Trend is not linear and has displayed significant variability in each month.

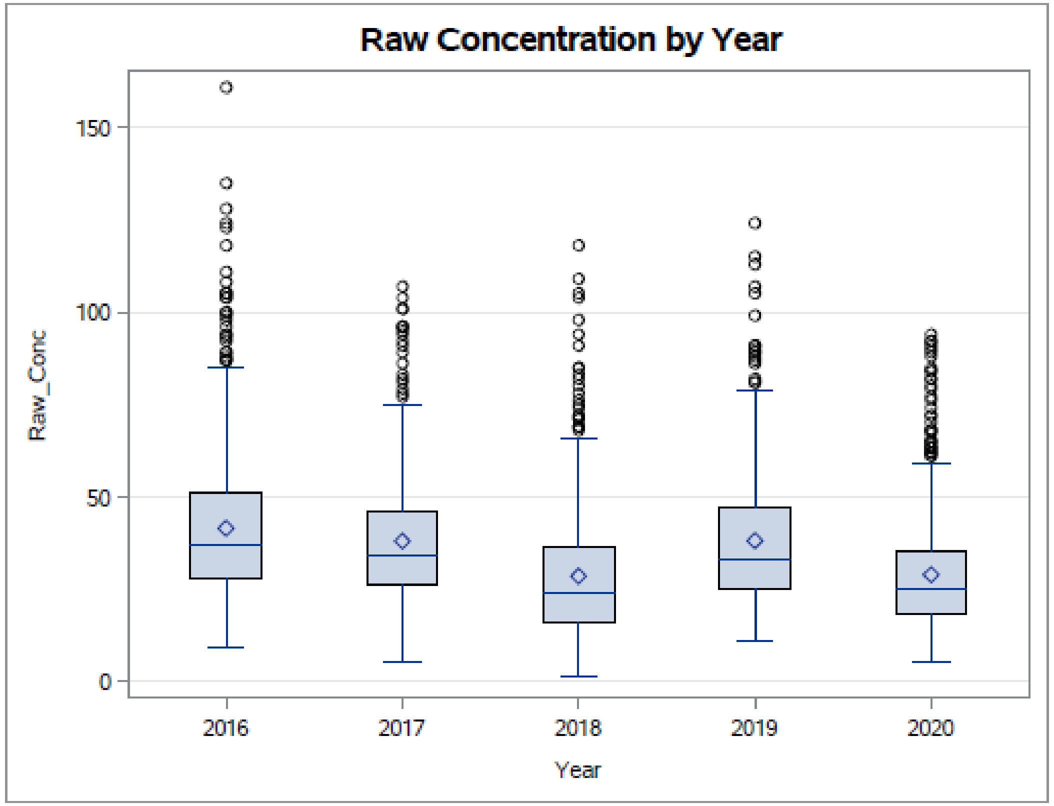
AQI LEVEL FROM 2016 TO 2020



- Mean and median air pollution levels have declined from a high in 2016, reaching their lowest levels in 2018, rising again somewhat in 2019, before declining again in 2010.
- Air pollution levels in most years are skewed to the right due to outliers of higher pollution levels.
- In years of lower average pollution, there are dramatic upswings in air pollution that are as high as years with higher average AQI levels.
- Greater variability in air pollution levels; while air pollution may have declined, it is more erratic.

ANNUAL CHANGES IN RAW CONCENTRATION LEVELS

RAW CONC FROM 2016 TO 2020



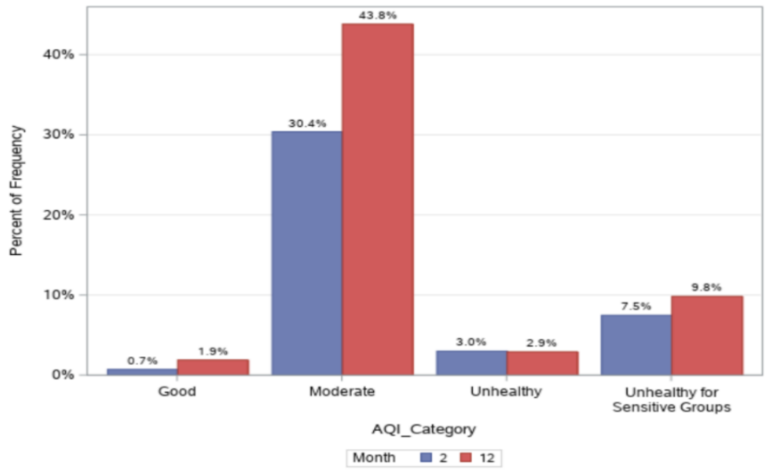
COMMENTS

- Mean Raw Concentration levels are trending downwards, though this is not a linear trend as they increased again in 2019.
- Mean is higher than the median in every year: most times show moderate air pollution levels but there are a large number of outliers with higher levels of air pollution
- Minimum levels of air pollution have not significantly changed, whereas maximum levels of air pollution have declined significantly. Correspondingly, the range has declined.

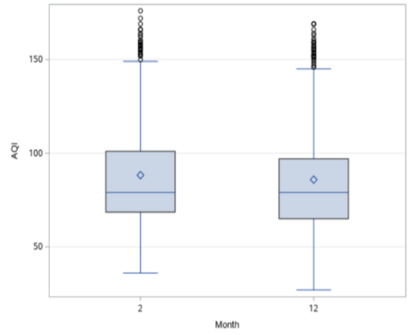
	2016 - December	2017 - December	2018 - December	2019 - December	2020 - December
Mean	41.35	37.85	28.32	37.97	28.74
Median	37	34	24	33	25
Minimum	9	5	1	11	5
Maximum	161	107	118	124	94
Standard Deviation	20.48	16.62	17.69	17.62	15.50
Range	152	102	117	113	89
Interquartile Range	23	20	20	22	17
Coefficient of Variation	49.52	43.91	62.45	46.40	53.93

MONTHLY CHANGES IN AQI CATEGORIES AND AQI LEVELS

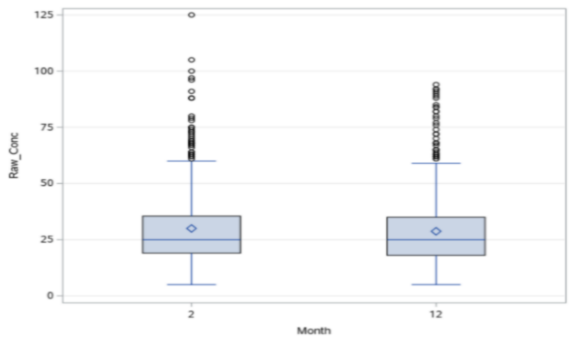
AQI LEVEL CATEGORIES FROM DEC 2020 TO FEB 2021



AQI



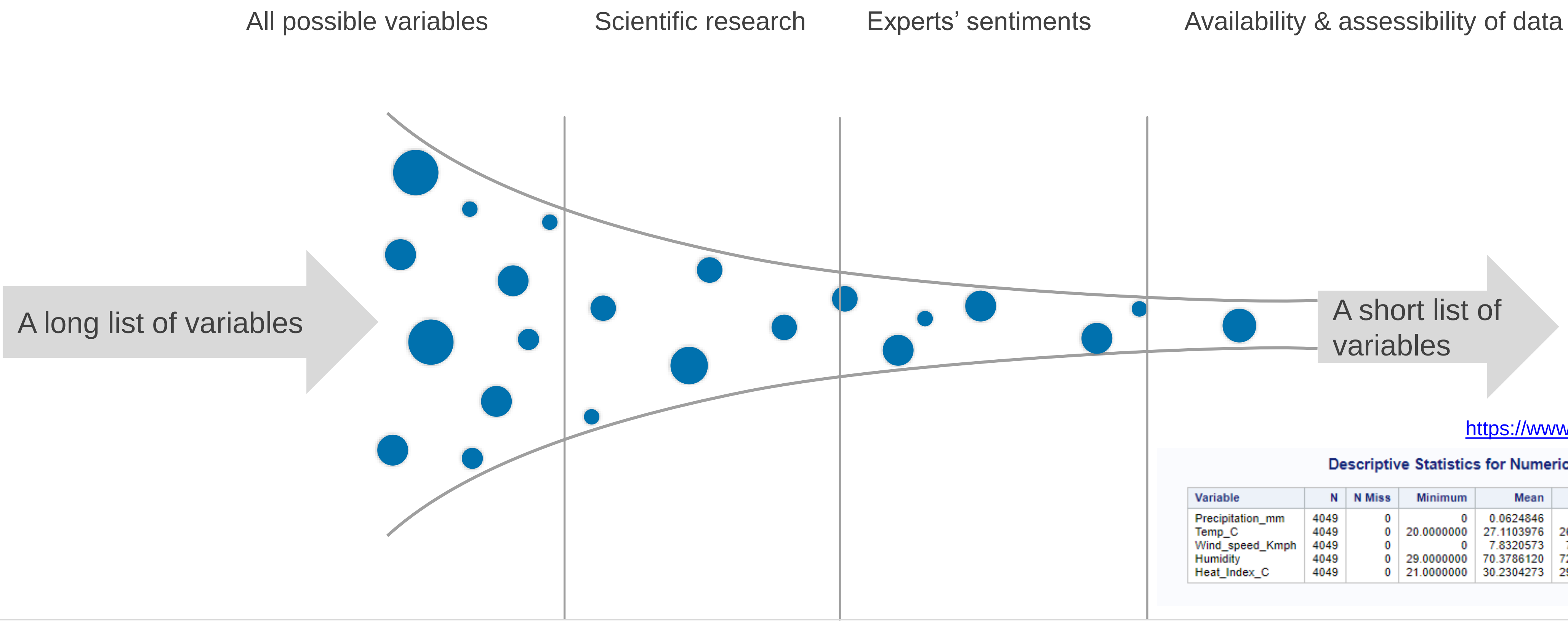
Raw Conc



COMMENTS

- December has more time periods of good air quality (3.23%) than in February (1.70%) but a similar number of time periods that are moderate, with air quality being moderate 74.97% of the time in December, and 73.11% of the time in February
- February had more time periods of Unhealthy (7.20%) or Unhealthy for Sensitive Groups (17.99%), than December (4.98% and 16.82% respectively)
- February 2021 had only slightly higher levels of air pollution than in December 2020.
- One might expect to see significantly higher levels in February due to:
  - Before and during Tet festival, city usually experiences high levels of travel, tourism, and the use of fireworks and firecrackers.
  - However, due to the COVID-19 pandemic, these activities were largely curtailed in 2021, reducing the potential impact they have on air pollution levels.

PROCESS OF SELECTING VARIABLES (EXTERNAL)



- Temperature
- Humidity
- Precipitation
- Windspeed
- Weather description
- Heat index

https://www.worldweatheronline.com

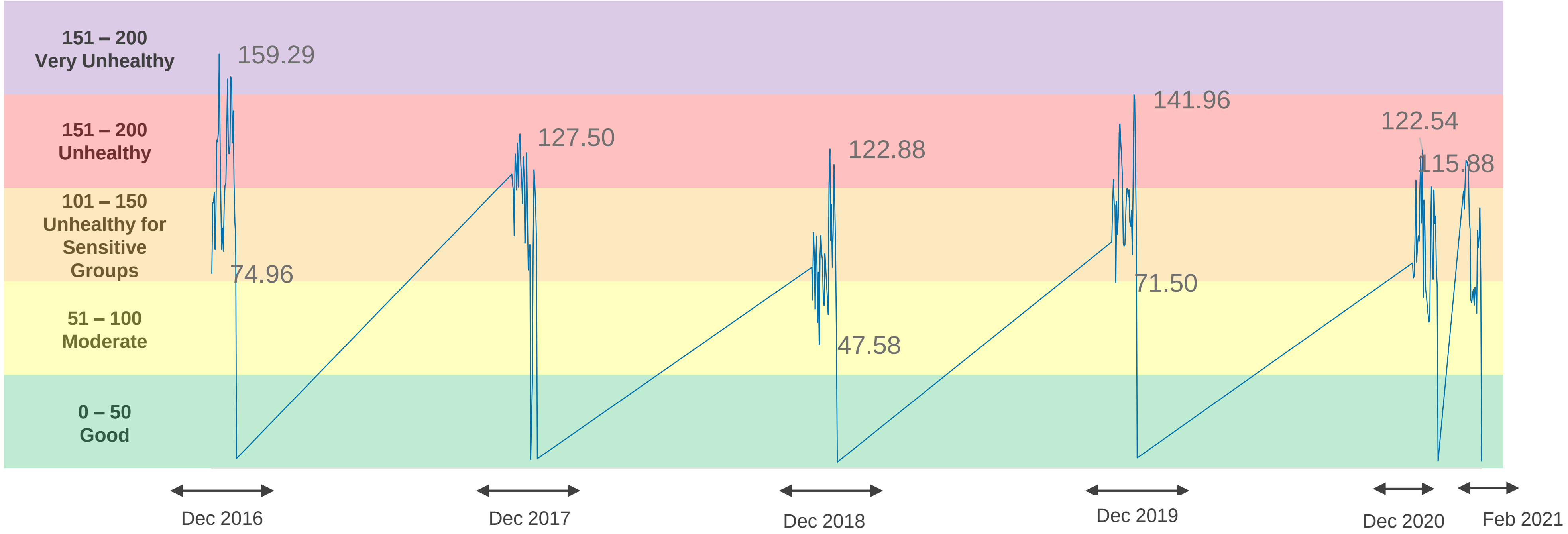
Descriptive Statistics for Numeric Variables

Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
Precipitation_mm	4049	0	0	0.0624846	0	5.6000000	0.2677915
Temp_C	4049	0	20.0000000	27.1103976	26.0000000	36.0000000	3.1344088
Wind_speed_Kmph	4049	0	0	7.8320573	7.0000000	24.0000000	3.5479424
Humidity	4049	0	29.0000000	70.3786120	72.0000000	98.0000000	13.2055858
Heat_Index_C	4049	0	21.0000000	30.2304273	29.0000000	41.0000000	3.6668317

- The **municipal environment department** says “A combination of **tropical convergence and cold air** in the atmosphere produced cloudy sky in HCMC and **high moisture levels**, which caused **air pollutants to condense into smog**.
- As there was **not enough sunlight** to heat up the ground, **temperature inversion** kicked in and prevented the smog from being dispersed into the upper atmosphere, confining it close to the ground and making them thicker and longer-lasting

AIR QUALITY IN IS CONSISTENTLY RANKED AS UNHEALTH THROUGHOUT THE YEARS

AIR QUALITY INDEX – 2016 TO 2021 DAILY AVERAGE

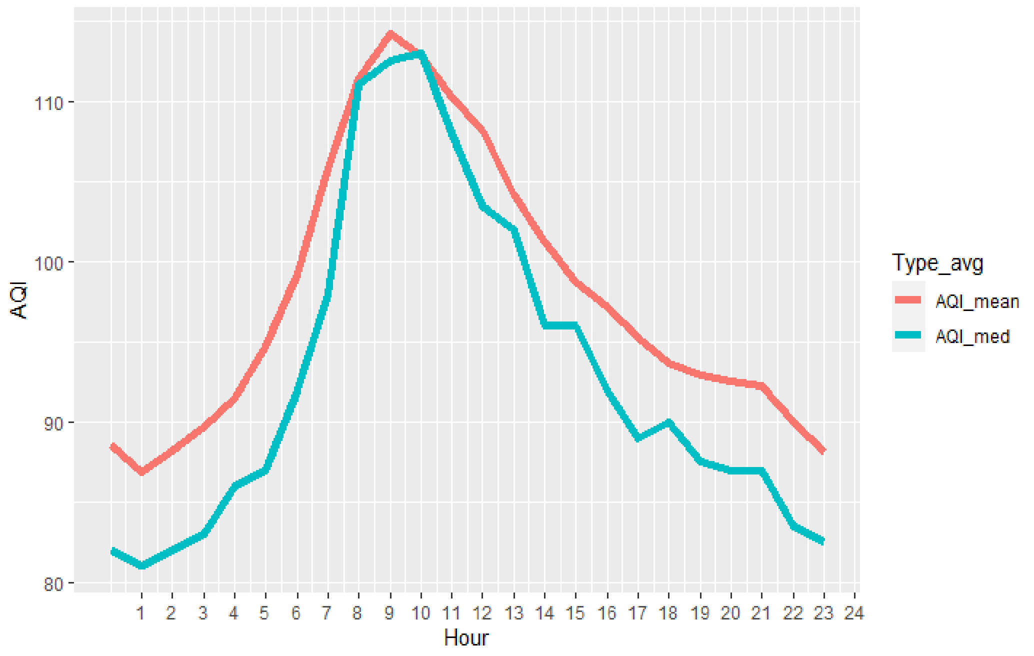


COMMENT

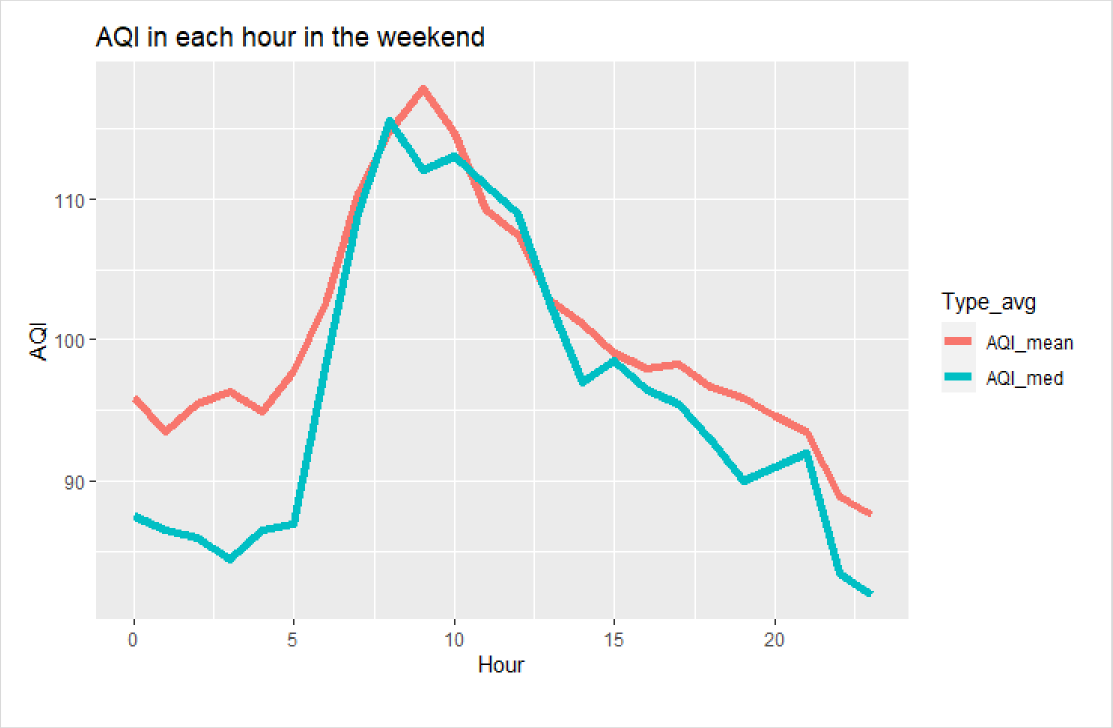
- From above graphs, we also see that there are some days in 2017 and 2020 that have extremely high AQI levels. Interestingly, some of the days are weekend:
  - 10 December 2016 (Total AQI for 24 hours: 3823 UG/m3 or average hourly AQI: 159 UG/m3)
  - 22 December 2019 (Total AQI for 24 hours: 2571 UG/m3 or average hourly AQI: 108 UG/m3)

HOURLY AQI LEVELS FOR ALL DAYS THROUGHOUT THE YEARS

HOURLY AQI LEVELS FOR ALL DAY FROM 2016-2021



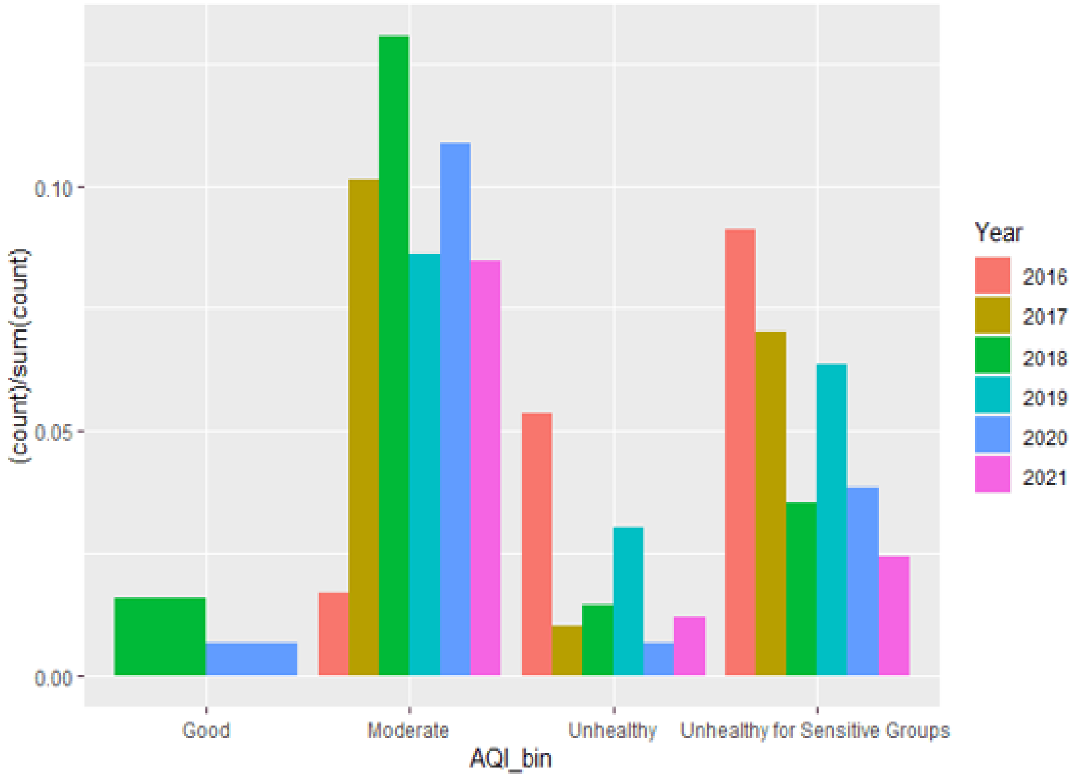
HOURLY AQI LEVELS FOR WEEKEND FROM 2016-2021



- The AQI levels are **highest** within from **7am to 10 am** for both weekdays and weekends
- The lowest AQI levels are **lowest** from **22h to 3 am** for both weekdays and weekends

PLOT AQI LEVELS FOR WEEKEND DAYS THROUGHOUT THE YEARS

AQI LEVELS FOR WEEKEND FROM 2016-2021

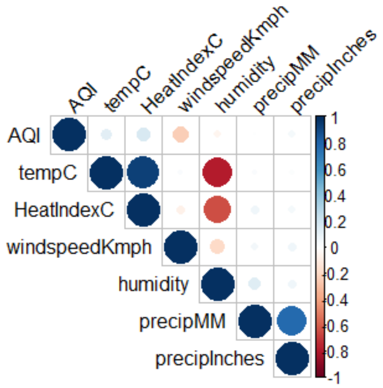


COMMENTS

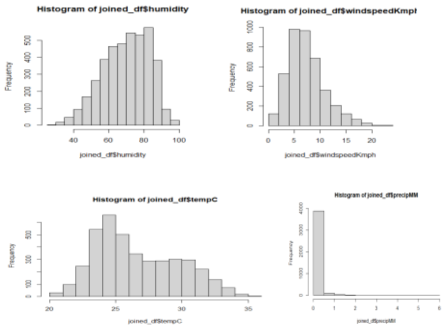
- **A total of 1195 weekend days** (for all years) and **573 days** of those are with AQI levels of 'Unhealthy' or 'Unhealthy for sensitive group', which accounts for **49%**
- From the above bar graph, we can see that years 2016, 2017, 2019, 2020 and 2021 almost do not have good air quality at all for weekend days
  - For December 2018, weekend days, the majority AQI level is moderate
  - December 2016, weekend days, has highest "unhealthy for sensitive group" among all years
  - December 2019, weekend days, has highest "unhealthy" AQI level among all years
  - December 2021's AQI "unhealthy" level is even higher than December for all weekend days
  - Year 2018 and 2020 have highest "moderate" AQI level for all weekend days

CORRELATION OF NUMERIC AND CATEGORICAL VARIABLES

CORRELATION MATRIX



DISTRIBUTION OF VARIABLES



COMPARISON OF NUMERIC AND CATEGORICAL MATRIX

Pairs	CramerV (Categorical variables)	Pairs	Correlation (Numeric variables)
AQI & Windspeed	0.1165437	**AQI & Wind speed**	-0.24
AQI & Weather	0.1334249	**AQI & HeatIndexC**	0.14
AQI & preciMM	0.03355126	**AQI & preciMM**	0.0078
AQI & tempC	0.1015877	**AQI & tempC**	0.12
AQI & humidity	0.09740825	**AQI & humidity**	-0.76

COMMENTS

- AQI & Windspeed and AQI & humidity are two pairs that have opposite correlations with 2 methods
- According to CramerV method, one variable (AQI) increases while another (Windspeed) decrease
- According to correlation method, two variables (AQI & Windspeed) increase relatively with each other
- AQI & Weather and AQI & tempC pairs have the same results with both methods, in other word, two variables move with the same direction, one increases then another will slightly increase accordingly

CHI-SQUARE TEST

	df	X-squared	p-value	Comments
AQI & weather	45	211.67	< 2.2e-16	Reject Ho, accept H1
AQI & humidity	12	115.26	< 2.2e-16	Reject Ho, accept H1
AQI & temperature	6	83.572	6.522e-16	Reject Ho, accept H1
AQI & preciMM	6	2.427	0.8765	Can't reject Ho
AQI & windspeed	12	164.99	< 2.2e-16	Reject Ho, accept H1

COMMENTS

- Hypothesis establishment
  - H0: There are NO such association between variables of each pair
  - H1: Ho is not true or there are associations between the variables
- From the chi-square test, we can conclude that the below pairs are associated with each other:
  - AQI & weather
  - AQI & humidity
  - AQI & temperature
  - AQI & windspeed

LINEAR REGRESSION MODEL

MODEL 1

```
lm(formula = AQI ~ tempC + humidity + precipMM + windspeedKmph,
    data = joined_df)

 Residuals:
    Min      1Q  Median      3Q     Max
-71.322 -22.257  -5.562  20.467  77.324

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    84.1424    11.1285   7.561 4.91e-14 ***
tempC           1.1901     0.2553   4.662 3.23e-06 ***
humidity       -0.0258     0.0624  -0.413    0.679
precipMM        2.2593     1.7456   1.294    0.196
windspeedKmph  -2.1733     0.1381 -15.738  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.66 on 4072 degrees of freedom
Multiple R-squared:  0.07656,       Adjusted R-squared:  0.07565
F-statistic:  84.4 on 4 and 4072 DF,  p-value: < 2.2e-16
```

COMMENTS

- R-square is 7.6%, which is quite low. That means the model can explain only 7.6% the variability of the dependent variable (AQI)
  - Temperature (tempC), windspeedKmP and humidity are statistically significant with the confidence interval of approximately 99%
  - TempC has a coefficient of 1.19, that means, if temperature increase 1 unit (degree), the AQI will increase 1.19 unit
  - Windspeed has a coefficient of -2.17 , that means, if windspeed increase 1 unit, the AQI will decrease -2.17 units

LINEAR REGRESSION FOR A SINGLE VARIABLE

MODEL 2

```
Call:
lm(formula = AQI ~ tempC, data = joined_df)

Residuals:
   Min     1Q Median     3Q    Max
-75.93 -23.22  -6.01  21.56  76.78

Coefficients:
            Estimate Std. Error t value
Pr(>|t|)
(Intercept)  64.8932     4.1845  15.508  <
2e-16 ***
tempC         1.2132     0.1533   7.912
3.23e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

Residual standard error: 30.58 on 4047
degrees of freedom
Multiple R-squared:  0.01523,    Adjusted
R-squared:  0.01499
F-statistic:  62.6 on 1 and 4047 DF,
p-value: 3.235e-15
```

MODEL 3

```
Call:
lm(formula = AQI ~ humidity, data =
joined_df)

Residuals:
    Min      1Q  Median      3Q     Max
-75.842 -23.212  -6.224  21.282  75.675

Coefficients:
             Estimate Std. Error t value
Pr(>|t|)
(Intercept) 106.48300    2.62242  40.605  <
2e-16 ***
humidity     -0.12361    0.03662  -3.375
0.000745 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

Residual standard error: 30.77 on 4047
degrees of freedom
Multiple R-squared:  0.002807,   Adjusted
R-squared:  0.002561
F-statistic: 11.39 on 1 and 4047 DF,  p-
value: 0.0007446
```

MODEL 4

```
Call:
lm(formula = AQI ~ windspeedKmph, data =
joined_df)

Residuals:
    Min      1Q  Median      3Q     Max
-74.326 -22.639  -6.326  20.256  80.674

Coefficients:
              Estimate Std. Error t value
Pr(>|t|)
(Intercept)   114.2672     1.1387  100.35
<2e-16 ***
windspeedKmph  -2.1046     0.1324  -15.89
<2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

Residual standard error: 29.89 on 4047
degrees of freedom
Multiple R-squared:  0.05874,    Adjusted
R-squared:  0.05851
F-statistic: 252.6 on 1 and 4047 DF,
p-value: < 2.2e-16
```

LINEAR REGRESSION MODELS FOR LAG – TIME DATASET

MODEL 5: 3 DAYS LAG

Call:
lm(formula = AQI ~ tempC + humidity + precipMM + windspeedKmph,
data = df)

```
Residuals:
    Min      1Q  Median      3Q     Max
-70.301 -21.099  -4.696  19.414  88.255

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  136.181081  10.890905  12.504   <2e-16 ***
tempC         -0.598544   0.249720  -2.397   0.0166 *
humidity       0.005723   0.061085   0.094   0.9254
precipMM       3.142195   1.705996   1.842   0.0656 .
windspeedKmph -2.909917   0.135499 -21.476   <2e-16
***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.98 on 4041 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.1169, Adjusted R-squared:
0.1161
F-statistic: 133.8 on 4 and 4041 DF,  p-value: < 2.2e-16
```

MODEL 6: 7 DAYS LAG

Call:
lm(formula = AQI ~ tempC + humidity + precipMM + windspeedKmph,
data = df)

```
Residuals:
    Min      1Q  Median      3Q     Max
-68.234 -20.622  -2.635  18.197  87.731

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  152.30746   10.48916  14.520  < 2e-16 ***
tempC         -1.61908    0.24044  -6.734 1.89e-11 ***
humidity       0.18047    0.05881   3.069  0.00216 **
precipMM       1.22642    1.64152   0.747  0.45503
windspeedKmph -2.99200    0.13087 -22.862  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.87 on 4037 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.1835, Adjusted R-squared:  0.1827
F-statistic: 226.8 on 4 and 4037 DF,  p-value: < 2.2e-16
```

COMMENTS

- If we performed LM for dataset with 3 days and 7 days lag, we achieved better R-square results, **12% and 18 % respectively**, that means this model can explain 12% the variability of AQI (independent variable) for the dataset with 3 days lag and 18% for 7 days lag
- The adjusted R-squared also increased from **11%** for the dataset of **3 days** lag to **18%** for the dataset of **7 days lag**
- TempC , windspeed and humidity are statistically significant with the confidence interval of approximately 99%
- TempC has a coefficient of -1.6, that means, if temperature increase 1 unit , the average of AQI will decrease -1.6 unit
- Humidity has a coefficient of 0.18, that means, if humidity increases 1 unit, the average of AQI will also increase 0.18 unit
- Windspeed has a coefficient of -2.99, that means, if windspeed increase 1 unit, the average of AQI will decrease -2.99 units

LINEAR REGRESSION MODELS FOR LAG – TIME DATASET & INTERACTIVE IMPACT OF EACH PAIRS

MODEL 7: 7 DAYS LAG & INTERACTIVE IMPACT OF TEMPERATURE

lm(formula = AQI ~ tempC + humidity + precipMM + windspeedKmph + tempC * windspeedKmph + tempC * humidity + tempC * precipMM, data = df)

```
Residuals:
    Min      1Q  Median      3Q     Max
-70.672 -20.215  -3.091  18.141  89.160

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        161.71786   26.13985   6.187 6.76e-10 ***
tempC               -1.82083    0.87661  -2.077 0.037852 *
humidity             1.07419    0.31216   3.441 0.000585 ***
precipMM           -35.68322   18.11143  -1.970 0.048883 *
windspeedKmph      -10.12353    1.16270  -8.707  < 2e-16 ***
tempC:windspeedKmph  0.26310    0.04219   6.236 4.94e-10 ***
tempC:humidity      -0.03573    0.01118  -3.197 0.001400 **
tempC:precipMM       1.39842    0.64976   2.152 0.031439 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.68 on 4034 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.1952,    Adjusted R-squared:  0.1938
F-statistic: 139.8 on 7 and 4034 DF,  p-value: < 2.2e-16
```

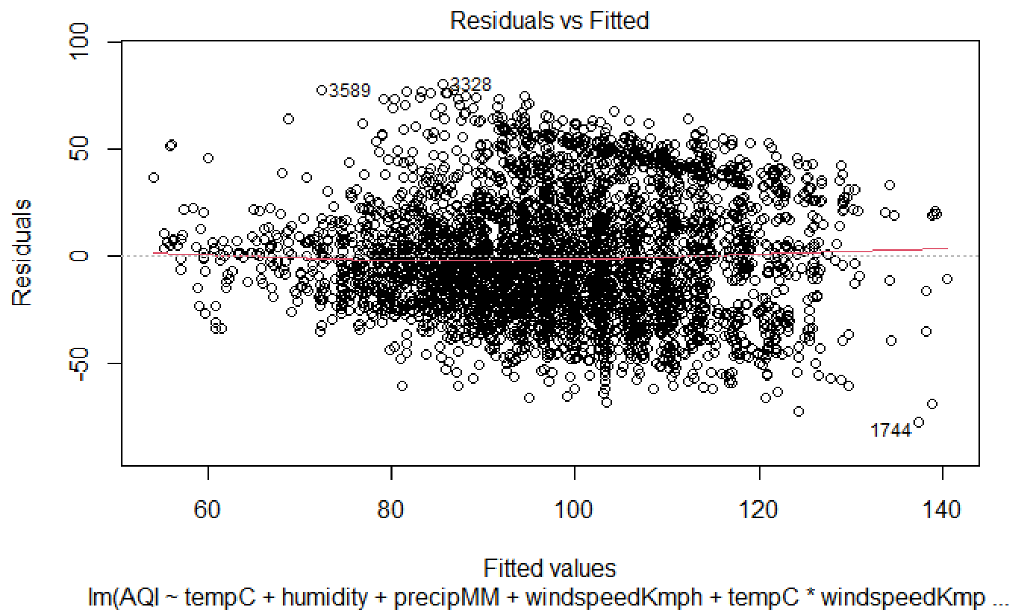
MODEL 8: 7 DAYS LAG & INTERACTIVE IMPACT OF EACH PAIRS

lm(formula = AQI ~ tempC + humidity + precipMM + windspeedKmph + tempC * windspeedKmph + tempC * humidity + tempC * precipMM + humidity * precipMM + humidity * windspeedKmph + precipMM * windspeedKmph, data = df)

```
Residuals:
    Min      1Q  Median      3Q     Max
-77.447 -20.224  -3.228  18.208  80.431
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.828e+02  3.578e+01   5.110 3.37e-07 ***
tempC                 -2.324e+00  1.029e+00  -2.259  0.02395 *
humidity               8.919e-01  3.551e-01   2.512  0.01206 *
precipMM              -2.364e+02  8.943e+01  -2.644  0.00823 **
windspeedKmph         -1.152e+01  2.667e+00  -4.319 1.60e-05 ***
tempC:windspeedKmph    2.863e-01  6.457e-02   4.434 9.48e-06 ***
tempC:humidity        -3.183e-02  1.120e-02  -2.841  0.00452 **
tempC:precipMM         5.903e+00  1.914e+00   3.083  0.00206 **
humidity:precipMM      6.026e-01  4.765e-01   1.265  0.20609
humidity:windspeedKmph 7.845e-03  1.553e-02   0.505  0.61344
precipMM:windspeedKmph 3.176e+00  5.318e-01   5.972 2.54e-09 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 27.55 on 4031 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.2036, Adjusted R-squared:  0.2016
F-statistic:   103 on 10 and 4031 DF,  p-value: < 2.2e-16
```

RESIDUAL PLOT FOR MODEL 8

RESIDUAL PLOT



COMMENTS

- The residual plot shows that the residuals are not very equally distributed across the regression line above and below the regression line.
- There is no pattern between residuals and fitted AQI, therefore there is no non- linear relationship between AQI and dependent variables

SUMMARY OF MODEL 7 & 8

GENERAL COMMENTS

- As mentioned in the methodology, I multiplied 2 continuous independent variables with each other to find out the interactive impact of the independent pair on the dependent variable, AQI
- **Model 7**: we examine the interactive impact of pairs: TempC*windspeedKmph , tempC*humidity , tempC*precipMM on the dependent variable, we achieve **R-square** of **19.5%,** that means this model can explain 19.5% the variability of AQI (independent variable)
- The **adjusted R-squared** also slightly increased from **19% to 20%** from from **model 7 to model 8**
- Windspeed, temp*windspeedKmP and precipMM*windspeedKmph are statistically significant with the confidence interval of approximately 99.9%

COMMENTS ON IMPACT OF EACH PAIR ON AQI LEVELS

- tempC:humidity, tempC:precipMM, precipMM are statistically significant with the confidence interval of approximately 99%
- **TempC and humidity** are statistically significant with the confidence interval of approximately 99%
- **Windspeed** has a coefficient of **-1.15**, that means, if windspeed increase 1 unit (km/h), the AQI will decrease -1.15 units (UG/m3)
- **tempC*windspeedKmph** has a coefficient of **2.863**, this means that the higher tempC make the negative impact of windspeed on AQI to be lower
- **precipMM*windspeedKmph** has a coefficient of **3.17**, we will have the same explaination
- **tempC* humidity** has a coefficient of **-3.18**, that means, when tempC increases, the **impact of humidity on AQI is negative**. In other words, the higher tempC, the more negative impact of humidity on AQI

EXECUTIVE SUMMARY & RECOMMENDATIONS

Question 1&2&4

POLLUTION TREND

- **Trend has reversed** over the **last 5 years** and there is less variability in air pollution levels, with fewer time periods of extreme and high levels
- Air pollution levels that are **unhealthy or unhealthy for sensitive groups** around **one-fifth** of the time - **levels are still quite high** and will continue to affect people's health
- Air pollution levels in HCMC have significantly increased over the decades since the rapid industrialization of Vietnam
- The government should encourage people to use public transport, reduce the reliance of the usage of fossil fuel in factories and households

Question 3

RELATIONSHIP ANALYSIS

- **CramerV method**, AQI levels increase when wind speed decreases, whereas the correlation method found that AQI levels increase when wind speed increases. However, **AQI levels** were found to increase as **temperature increased** using **both methods**.
- **Chi-square tests** showed that there is a relationship between **AQI levels and four weather-related variables**: weather, humidity, temperature, and wind speed
- The model that has the **highest R-square, 20%** in which we multiplied each two independent variables with each other to see the interactive impact of each pair on AQI and the dataset being used is **7-day time lag**

Example

For example, today is not crowded, but some days before, it was crowded with cold air like a blanket, accumulating fine dust that cannot spread causing the pollution index to rise

- The models with **the time lag datasets** explain the expert's sentiments or scientific evidence
- **Pollution** is caused by a combination of **factors**: vehicle emission, industrial operation, temperature, wind, dust dispersion and others
- **Weather** can explain the variability in air pollution levels to a certain level
- The **government** should give warning to its people about pollution levels by **predicting pollution levels** based on the combination of factors, especially forecast the pollution levels by **analyzing the patterns** of temperature from 3 to **a week before**

THANK YOU FOR YOUR ATTENTION!