Trends in Air Pollution in Ho Chi Minh City and the Associated Impact of Weather Conditions.

Written by:

Kevin Moroso – 098302599

Lien Pham – 100361334

Sarath Ravikumar – 100368894

Ayushi Singh – 100359100

March 28, 2021

Introduction

Beginning in the 1980s, Vietnam has undergone tremendous economic growth, surpassing 5% growth annually in every year but one since 1990. This has been driven by industrialization, and Vietnam has become an export-oriented economy, becoming the 23rd largest economy in the world. This has also led to large increases in air pollution in Ho Chi Minh City, which contributes around 40% of Vietnam's GDP, as a result of industrial and construction activities, as well as increases in vehicle ownership. The main drivers of air pollution in Ho Chi Minh City are transportation, especially due to older vehicles that do not comply with emissions standards, and construction, where the demolition of older buildings creating dust and the dispersal of cement powder when constructing new buildings.

The smog smothering Ho Chi Minh City is made of condensed air pollutants caused by high moisture and temperature inversion, which the municipal environment department says is "A combination of tropical convergence and cold air in the atmosphere [producing] cloudy sky in Ho Chi Minh City and high moisture levels, which [causes] 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."[1]

Air pollution levels can be affected by a number of factors, both anthropogenic (e.g. vehicle emissions) and natural (e.g. direction and strength of wind), and can change throughout the year and over time. This paper focuses on four questions:

1. Is air quality in Ho Chi Minh City getting worse, improving, or staying the same?
2. How does air quality in Ho Chi Minh City change at different times of the year?
3. Is there a relationship between weather and air quality in Ho Chi Minh City?
4. Given the information gathered from questions 1 to 3, are there ways to mitigate the impact of air pollution in Ho Chi Minh City?

Methodology

[1] In September 2019, the municipal environment department denied that air pollution was caused by forest fires in Indonesia as reported by some online media outlets, adding that similar pollution levels had been recorded in previous years.

To answer these 4 questions, a number of analyses were performed. In order to determine whether air quality in Ho Chi Minh City is getting worse, improving, or staying the same, air pollution data is compared for the month of December in each year from 2016 to 2020. In order to determine how air quality in Ho Chi Minh City changes at different times of the year, air pollution data is compared for the month of December 2020 to the month of February 2021. In order to determine whether there is a relationship between weather and air quality in Ho Chi Minh City, a variety of statistical tests were performed between air quality and humidity, temperature, precipitation, and wind speed.

The dataset created combines Raw Concentration pollution data and Air Quality Index ("AQI") levels with humidity, temperature, precipitation, and wind speed for the month of December from 2016-2020, and 23 days in the month of February 2021.[2] Appendix A includes a description of the variables in the datasets.

The original dataset had outliers that were out of the expected range ("985" and negative values) and missing values (which had been replaced in the original dataset with "-999") for Raw Concentration and AQI. This was cleaned by removing observations with missing air pollution data. Outliers beyond the expected range of a particular variable were also removed. Data was not replaced with medians or another number to ensure it didn't affect relationship tests between air pollution and precipitation. The original dataset had 4224 observations before cleaning, 174 observations were removed during cleaning (4%), and 4050 observations remained after cleaning. As less than 5% of observations were removed, it should not significantly affect the analysis. Appendix B includes SAS and R codes for the combining and cleaning of the dataset.

<u>Section 1: Changes in Air Quality in Ho Chi Minh City (2016-2020)</u>

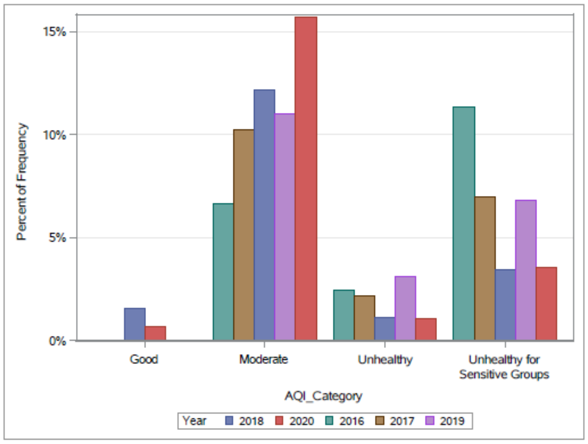
Air quality has changed in Ho Chi Minh City during the month of December over the 5-year period of 2016 to 2020. The analysis focuses on three different types of measurements: Raw Concentration levels, AQI levels, and AQI Categories. Raw concentration is the air pollution level of particulates 2.5 microns in diameter or smaller per cubic metre over a one hour period. NowCast is a weighted average of these hourly raw concentration observations from the most recent 12 hours but weighted more towards the more recent data and the AQI level is then

[2] Please see attached file for the cleaned dataset: *HCMC_master.csv*.

derived from this. AQI categories are then based on the AQI levels. Formulas and tables that explain these variables further can be found in Appendix C. Appendix D includes the SAS codes for the analysis in this section.

AQI Category Frequencies

There are six AQI categories: Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous. During the time period for the observations, there were no time periods that experienced air pollution levels of Very Unhealthy or Hazardous. This table and chart shows the frequency for each category within each month of December from 2016 to 2020.

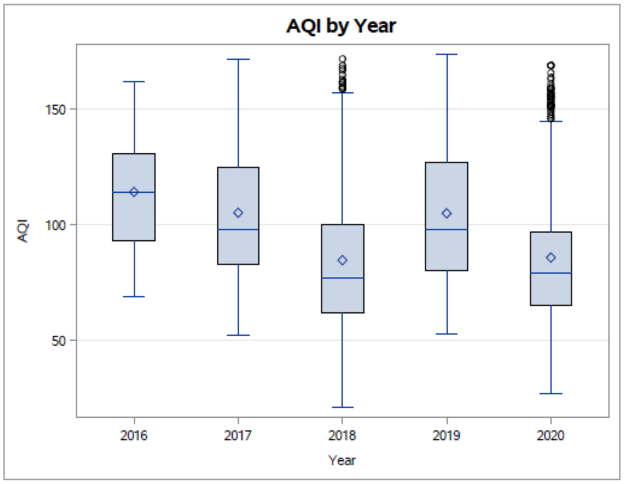


AQI Category	2016 - December	2017 - December	2018 - December	2019 - December	2020 - December
Good	0.00%	0.00%	8.55%	0.00%	3.26%
Moderate	32.59%	52.71%	66.56%	52.51%	74.90%
Good and Moderate combined	***32.59%***	***52.71%***	***75.11%***	***52.51%***	***78.16%***
Unhealthy for Sensitive Groups	55.57%	36.02%	18.66%	32.56%	16.82%
Unhealthy	11.84%	11.27%	6.22%	14.93%	5.02%
Unhealthy and Unhealthy for Sensitive Groups combined	***67.41%***	***47.29%***	***24.88%***	***47.49%***	***21.84%***

The proportion of time periods with air pollution that is unhealthy and unhealthy for sensitive groups displays a downward trend, with a high of 67.41% in 2016 (for both categories combined) to 21.84% in 2020, though this trend is not linear and has displayed significant variability in each month. Conversely, the proportion of time periods with air pollution levels that are good or moderate has increased. Even in the best year of December, 2020, air pollution levels are still unhealthy but this is a marked improvement from December 2016.

AQI Changes

AQI levels have also experienced change over this time period. These box plots show AQI levels data for the month of December in each year:



This table shows some summary statistics on AQI levels data:

	2016 - December	2017 - December	2018 - December	2019 - December	2020 - December
Mean	114.31	105.28	84.71	105.05	85.84
Median	114	98	77	98	79
Minimum	69	52	21	53	27
Maximum	162	172	172	174	169
Standard Deviation	23.61	28.14	31.67	30.20	28.19
Range	93	120	151	121	142
Interquartile Range	38	42	38	47	32

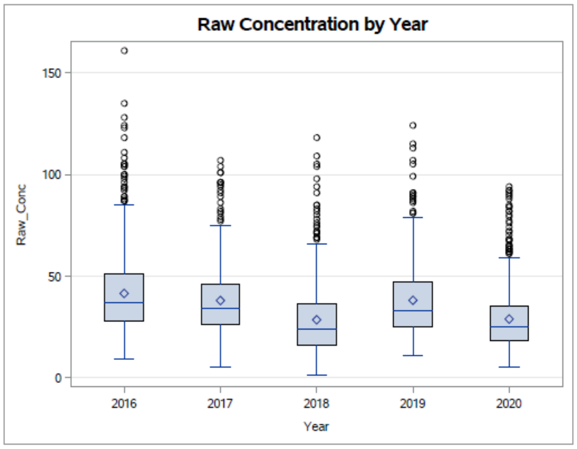
Mean and median air pollution levels have declined from a high in 2016, reaching their lowest levels in 2018, rising again somewhat in 2019, before declining again in 2010. 2016 displays a more symmetric distribution, with the mean and median almost the same; 2017-2020, air pollution levels are skewed to the right (the mean is higher than the median) due to outliers of higher pollution levels. The box plots show a significant number of extreme outliers in 2018 and 2020, the two years with the lowest average AQI levels, showing that in years of lower average pollution, there are dramatic upswings in air pollution that are as high as years with higher average AQI levels.

The range in air pollution levels is increasing, with slight increases in the maximum range, but a significant decrease in the minimum levels of air pollution. There are more days of moderate and good levels of air pollution. Standard deviation has increased, from a low of 23.61 in 2016, peaking in 2018 at 31.67. This indicates that there is greater variability in air pollution levels; while air pollution may have declined, it is more erratic. However, the interquartile range displays no clear trend – it is the outliers that are causing greater variability; when these are removed, the range in air quality levels is more consistent over time. It should be noted that AQI levels are determined by the NowCast Concentration measure, and the NowCast Concentration measure is designed to smooth out some of the variability in the Raw Concentration measure.

Data from 2016 shows no discernible pattern in distribution, broadly spread out among categories of moderate and unhealthy for sensitive groups; data from 2017-2020 is skewed to the right, indicating most time periods have moderate air pollution, with a smaller number of days of higher, unhealthy levels of air pollution. Histograms displaying the frequencies of AQI levels can be found in Appendix E.

Raw Concentration Changes

Raw Concentration data provides real time pollution levels without any smoothing out of the variability that the AQI level calculations use. These box plots show Raw Concentration data for the month of December in each year:



This table shows some summary statistics on AQI levels data:

	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

The mean Raw Concentration levels are trending downwards, decreasing from 41.35 in 2016 down to 28.74 in 2020, though this is not a linear trend as they increased again in 2019. Median Raw Concentration levels show a similar trend, going from 37 in 2016 to 25 in 2020. The mean is higher than the median in every year, indicating air pollution levels are positively skewed – most times show moderate air pollution levels but there are a large number of outliers with higher levels of air pollution. There are more outliers of higher Raw Concentration levels than shown in the AQI levels, as AQI levels are designed to smooth out some of the variability.

Minimum levels of air pollution have not significantly changed, whereas maximum levels of air pollution have declined significantly – from a peak of 161 in 2016 to a maximum level of only 94 in 2020, the lowest levels in the 5 year period. Correspondingly, the range has therefore declined, from 152 in 2016 to 89 in 2020. The standard deviation has also shown a moderate decline, from a high of 20.48 in 2016, to a low of 15.50 in 2020 – confirming less variability in air pollution levels. The interquartile range has also declined from a peak of 23 in 2016 to 17 in 2020; so even when outliers are excluded, there is less variability in air pollution levels.

The distribution of Raw Concentration data is similar for all years, being positively skewed to the right. This means that the outliers are for high levels of pollution and Raw Concentration levels are more concentrated at lower levels. Histograms displaying the frequencies of Raw Concentration levels can be found in Appendix E.

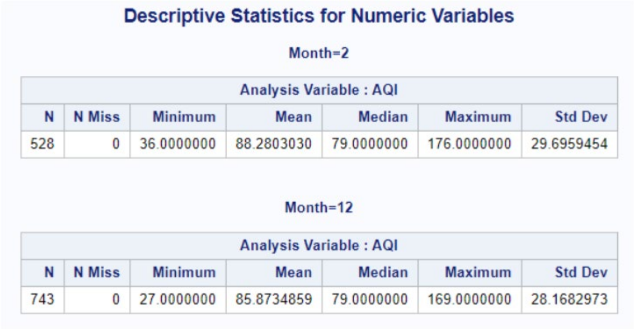
Conclusion

While air pollution levels in Ho Chi Minh City have significantly increased over the decades since the rapid industrialization of Vietnam, the trend seems to have reversed over the last 5 years, from a high in 2016 to a low in 2020. There is also less variability in air pollution levels, with fewer time periods of extreme and high levels. This is likely due to concerted action on the part of local authorities to reduce reliance on private vehicles, encouraging people to use public transport, and reducing congestion by varying working and school hours.[3] However, with 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.

Section 2: Difference in Air Quality in Ho Chi Minh City (February vs December)

The next analysis focuses on examining the difference in air pollution levels between December 2020 and February 2021[4], focusing again on AQI and Raw Concentration levels. SAS codes for this analysis can be found in Appendix F.

AQI Changes



Descriptive Statistics for Numeric Variables

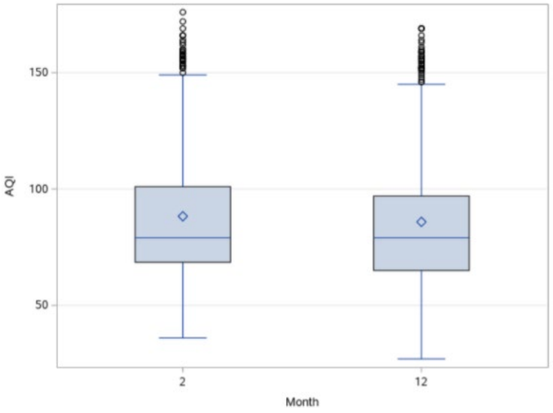
Month=2

Analysis Variable : AQI						
N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
528	0	36.0000000	88.2803030	79.0000000	176.0000000	29.6959454

Month=12

Analysis Variable : AQI						
N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
743	0	27.0000000	85.8734859	79.0000000	169.0000000	28.1682973

The table above displays summary statistics for AQI levels in February 2021 and then December 2020. Average AQI levels are somewhat higher in the month of February (88.28) than in December (85.87), and both the minimum and maximum levels of air pollution are higher in February. However, this is due to outliers in February as the median AQI levels are the same in both months at 79. This box plot shows the similarities in pollution levels in both months:

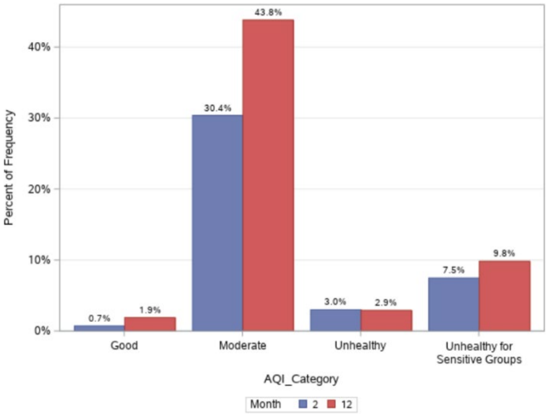


[3] HCMC continues to fight air pollution for the 2020-2030 period | Ho Chi Minh City | SGGP English Edition (sggpnews.org.vn)

[4] Data from February 1 to February 23, 2021 was available at the time of this analysis.

AQI Category Frequencies

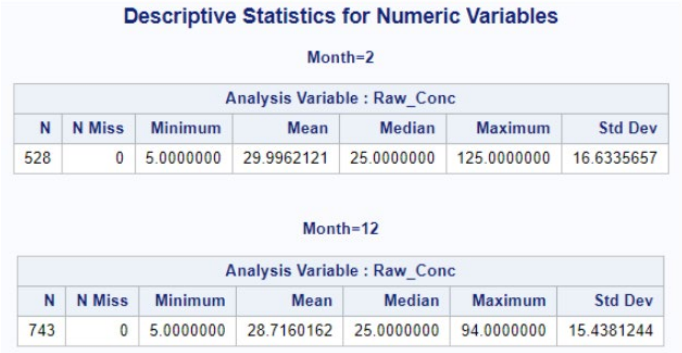
Focusing on the AQI categories for the two months, there are only some slight differences in air quality:



Neither February nor December had any time periods with air pollution levels that were Very Unhealthy or Hazardous. 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).

Raw Concentration Changes

Raw Concentration levels were also analyzed as AQI levels are designed to smooth out some of the variability in air pollution levels.



Descriptive Statistics for Numeric Variables

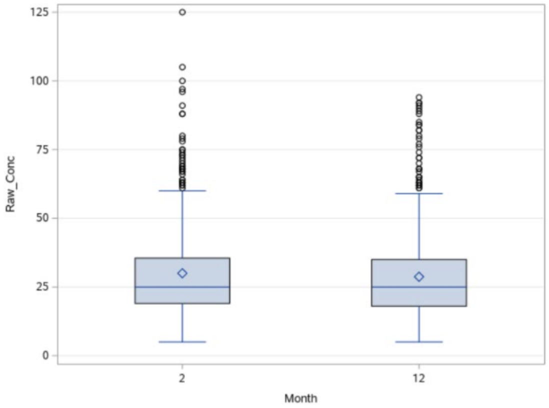
Month=2

Analysis Variable : Raw_Conc						
N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
528	0	5.0000000	29.9962121	25.0000000	125.0000000	16.6335657

Month=12

Analysis Variable : Raw_Conc						
N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
743	0	5.0000000	28.7160162	25.0000000	94.0000000	15.4381244

Both February and December have similar minimum levels of air pollution (5.0) but maximum levels in February are far higher at 125.0 when compared to December at 94.0 – February can have more dramatic swings upwards in air pollution levels. Despite these higher levels in February, median levels of air pollution are the same for each month (25.0) and average levels in February are only moderately higher at 30.0 when compared to December at 28.7. The boxplot below displays these similarities in air pollution levels in both months, with the exception of the outliers of higher air pollution levels in February:



Conclusion

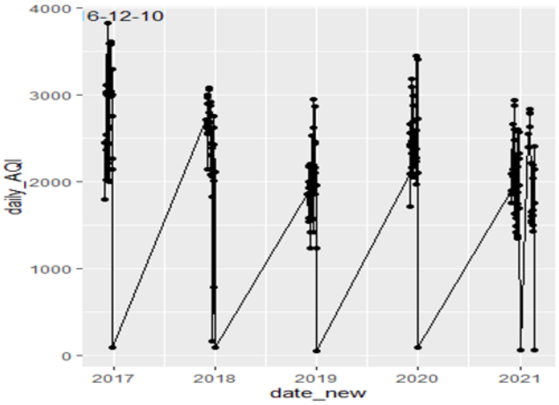
February 2021 had only slightly higher levels of air pollution than in December 2020. One might expect to see significantly higher levels in February, due to the Tet festival during that time period which usually experiences high levels of travel, tourism, and the use of fireworks and firecrackers. However, due to the COVID-19 pandemic, these activities were largely curtailed in 2021, reducing the potential impact they have on air pollution levels.

<u>Section 3: Relationship between Air Quality and Weather in Ho Chi Minh City</u>

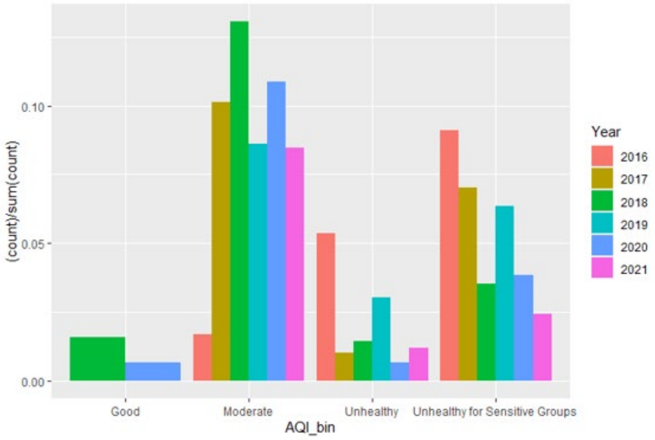
Weather can have a big impact on air quality, either suppressing or elevating pollution levels caused by man-made factors such as transportation and construction. In this section, we focus on the months included in previous sections, the months of December in 2016 through 2020, and the month of February 2021. Correlation methods were used for numeric and categorical variables to examine any relationship. A linear regression model was developed, using weather data with a 3- and 7-day time lag from air pollution data as it takes time for weather to impact air quality. In the linear regression model, we also examine the interactive impact of two continuous variables/pairs on the dependent variable by multiplying these two variables with each other. A chi-square test was used to measure the relationship among categorical variables. R Codes for this section can be found in Appendix G.

Preliminary analysis

Before examining the relationship between the different variables, differences in air pollution between weekdays and weekends were examined. The primary drivers of air pollution, such as construction and transport, are higher on weekdays. However, pollution levels would therefore be expected to accumulate throughout the weekdays and carry over into the weekend. This scatterplot displays AQI levels for weekends:



And this bar chart displays the frequency of AQI categories for weekends:



As the graphs above show, there are high levels of air pollution on weekends:

- For December 2018, the majority of time periods have an AQI of Moderate;
- December 2016 has the highest frequency of time periods where the AQI is Unhealthy for Sensitive Groups;
- December 2019 has the highest frequency of time periods where the AQI is Unhealthy;
- The frequency of time periods in December 2020 where the AQI is Unhealthy is even higher than December for all weekend days;

- Years 2018 and 2020 have the highest frequency of time periods where the AQI is Moderate.

There are a total of 1195 observations for weekend days in all years; 573 of these observations have an AQI that is Unhealthy or Unhealthy for Sensitive Groups, 49% of the total. This is a larger proportion than found in Sections 1 and 2. Therefore, we can conclude that the pollution is worse on weekends, despite the fact that construction and transportation activities are reduced on those days, and supporting the approach taken to examine time lags when examining the relationship between weather and air pollution.

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 during the weekend:

- 10 December 2016 (Total AQI over 24 hours: 3823; average hourly AQI: 159)
- 22 December 2019 (Total AQI over 24 hours: 2571; average hourly AQI: 108)

Correlation

Correlations were performed among all of the numerical variables. The following table shows the variables that had the highest correlations:

Variables	**R number**
Humidity & Temperature (Celsius)	-0.76
Humidity & Heat Index (Celsius)	-0.65
AQI & Wind speed	-0.24
Humidity and Wind speed	-0.19
Heat Index (Celsius) and AQI	0.14
Temperature (Celsius) & AQI	0.12

Using this method, the correlations were weak for each pair of variables. Therefore, we transformed some of the numerical variables above into categorical variables. Wind speed, precipitation, and humidity display skewed distributions, which means that a binning technique will not work as too many observations will be in low or high groups. Therefore, the data distribution itself decides bin ranges with no manual intervention as the bins will contain a uniform number of data points. The variables were divided by quantile to produce equal numbers for each group: Low, Medium, High. This method helps in partitioning the continuous valued distribution of a specific numeric field into discrete contiguous bins or intervals. Histograms displaying the distribution of the variables, along with an example of the binning technique, can be found in Appendix H.

Pairs	CramerV (Categorical variables)	Pairs	Correlation (Numeric variables)
AQI & Wind speed	0.1165437	AQI & Wind speed	-0.24
AQI & Weather [condition]	0.1334249	AQI & Heat Index	0.14
AQI & Precipitation	0.03355126	AQI & Precipitation	0.0078
AQI & Temperature	0.1015877	AQI & Temperature	0.12
AQI & Humidity	0.09740825	AQI & Humidity	-0.76

Two pairs of variables, AQI & Wind speed and AQI & Humidity, are negatively correlated using both the CramerV method and the correlation method. The CramerV method shows AQI increasing when wind speed decreases; low wind speed is related to higher levels of air pollution. Using the correlation method, two variables (AQI & Wind Speed) increase relatively with each other. AQI & Weather [Condition] and AQI & Temperature show the same results with both methods; when one increases then another will slightly increase.

Linear regression

Eight different linear regression models were developed to examine, with AQI as the dependent variable in each of them.

Model 1: AQI as dependent variable; Temperature, Humidity, Wind speed and Precipitation as predictors or independent variables

```
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
```

The R-square for Model 1 is quite low at 7.6%, meaning the model can only explain 7.6% the variability of AQI. Temperature, wind speed, and humidity are statistically significant with a confidence interval of approximately 99%. Temperature has a coefficient of 1.19 which means if temperature increases by 1 degree, AQI levels will increase by 1.19. Wind speed has a coefficient of -2.17 which means if wind speed increases by 1 km/h, AQI levels will decrease by -2.17.

Model 2: AQI as dependent variable; Temperature alone as an independent variable

```
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: AQI as dependent variable; Humidity alone is an independent variable

```
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: AQI as dependent variable; Humidity alone is an independent variable

```
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
```

Models 2, 3, and 4 each use a single predictor for AQI levels, Temperature, Humidity, and Wind speed. Each has an even lower R-square than Model 1 when we examined all the predictors/ independent variables together:

- Model 2 (AQI ~ Temperature): R-square: 1.5%
- Model 3 (AQI ~ Humidity): R-square: 0.28%
- Model 4(AQI ~ Wind speed): R-square: 5.8%

None of these predictor variables display a significant relationship to AQI levels. The next two models used a time lag to examine the relationship between weather and AQI levels.

Model 5: AQI as dependent variable; Temperature, Humidity, Wind speed and Precipitation as predictors or independent variables. 3-day time lag

```
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: AQI as dependent variable; Temperature, Humidity, Wind speed and Precipitation as predictors or independent variables. 7-day time lag

```
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
```

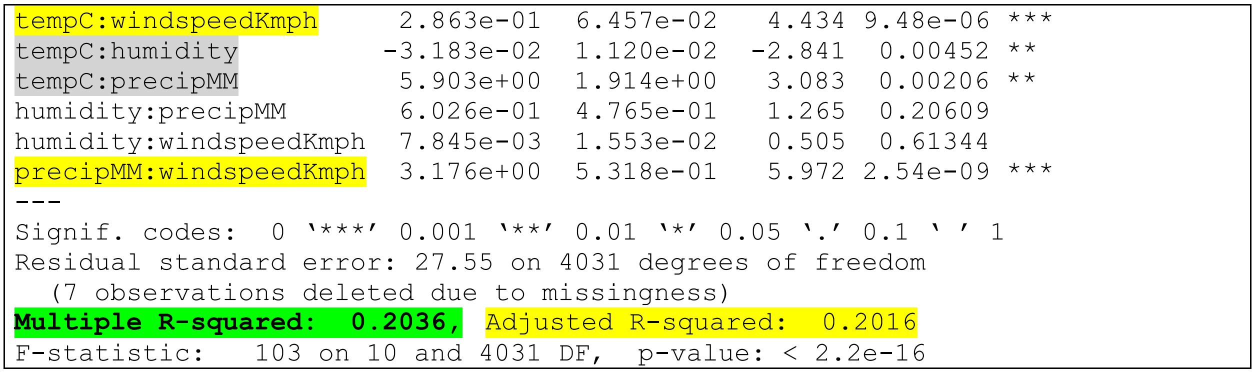
Models 5 and 6 provided better R-square results. Using a 3-day time lag for the predictor variables explained 12% of AQI levels; using a 7-day time lag for the predictor variables explained 18% of AQI levels. The adjusted R-squared also increased from 11% for Model 5 to 18% for Model 6. Temperature, Wind speed, and Humidity are statistically significant with a confidence interval of approximately 99%. Temperature has a coefficient of -1.6, which means that if temperature increases 1 degree Celsius, the AQI levels will decrease by 1.6. Humidity has a coefficient of 0.18, which means that if humidity increases by 1 unit, AQI levels will increase by 0.18. Wind speed has a coefficient of -2.99, which means that if wind speed increases by 1km/hr, AQI levels will decrease by 2.99.

*Model 7: AQI as dependent variable; Temperature, Humidity, Wind speed and Precipitation as predictors or independent variables; TempC*windspeedKmph , tempC*humidity , tempC*precipMM are also independent variables*

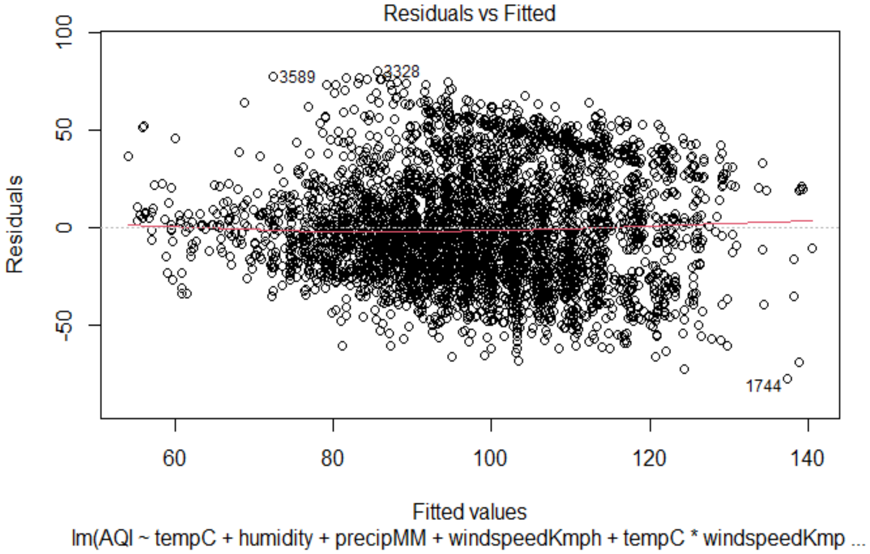
```
Residuals:
    Min      1Q  Median      3Q     Max 
-70.672 -20.215  -3.091  18.141  89.160 
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         161.71786   26.13985   6.187 6.76e-10 ***
tempC                -1.82083    0.87661  -2.077 0.037852 *  
humidity              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 : AQI as dependent variable; Temperature, Humidity, Wind speed and Precipitation as predictors or independent variables; tempC*windspeedKmph , tempC*humidity , tempC*precipMM, humidity * precipMM, humidity * windspeedKmph,precipMM * windspeedKmph are also independent variables*

```
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
```



For Models 7 and 8, two continuous independent variables were multiplied with each other to find out the interactive impact of the independent pair on the dependent variable, AQI. Model 7 examines the interactive impact of pairs: Temperature & Wind speed, Temperature & Humidity, and Temperature & Precipitation on the dependent variable. Model 7 has an R-square of 19.5%, which means this model can explain 19.5% of the variability of AQI. The adjusted R-squared has increased slightly from 19% in Model 6, to 20% in Model 7. All of the variables are statistically significant with a confidence interval of approximately 99%.

Wind speed has a coefficient of -1.15, which means that if wind speed increases by 1 km/h, AQI levels will decrease by 1.15, temperature with wind speed has a coefficient of 2.863 which means, the higher the temperature, the more negative impact of wind speed on AQI. Precipitation with wind speed has a coefficient of 3.17, indicating that the more precipitation, the more negative the impact of wind speed on AQI. And finally, temperature with humidity has a coefficient of -3.18, which means that when temperature increases, the impact of humidity on AQI levels is negative. In other words, the higher the temperature, the more negative the impact of humidity on AQI

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.

Chi-square tests

Chi-square tests were performed as an additional method to determine the relationship between weather and AQI levels, with H_0 indicating that there is NO relationship between the variables and H_1 indicating that there is a relationship between the variables:

Variables	df	X-squared	p-value	Comments
AQI & weather	45	211.67	< 2.2e-16	Reject H_0, accept H_1
AQI & humidity	12	115.26	< 2.2e-16	Reject H_0, accept H_1
AQI & temperature	6	83.572	6.522e-16	Reject H_0, accept H_1
AQI & precipitation	6	2.427	0.8765	Can't reject H_0
AQI & wind speed	12	164.99	< 2.2e-16	Reject H_0, accept H_1

From the chi-square test, we can conclude that the below pairs show a relationship between each other: AQI & weather; AQI & humidity; AQI & temperature; and AQI & wind speed.

Conclusion

The linear model performed with the dataset using a 3-day time lag (Model 5) and a 7-day time lag (Model 6) achieved better R-square results, with Model 5 explaining 12% of the variability of AQI levels and Model 6 explaining 18% of the variability. Using the CramerV method and correlation r for numeric variables, some of the results were contradictory:

- AQI & Wind speed and AQI & humidity are two pairs that have opposite correlations with both methods;
- Using the CramerV method, AQI increases while Wind speed decreases; using the correlation method, both AQI and Wind speed increase relative to each other;
- The two pairs, AQI & Weather and AQI & Temperature, have the same results using both methods (each variable moves in the same direction)

Using chi-square tests, we can conclude that there is a relationship between AQI and the following variables: weather, humidity, temperature, and wind speed.

Recommendations

Air pollution levels can have a significant impact on the health of people living in Ho Chi Minh City. Providing advance notice to people about worsening air quality would enable them to modify their behavior during those times, such as avoiding strenuous activities. It also enables the government to put in temporary measures to reduce activities that generate pollution, such as demolitions of old buildings or reducing traffic congestion. While it is generally difficult to directly predict air pollution levels for a given day, it is much easier to predict weather patterns, including precipitation, wind speed, humidity, and temperature. By using weather forecasts, governments can provide advisories to people, especially those more susceptible to the health impacts of air pollution, to avoid strenuous activities on those days, such as avoiding exercise, cycling, or walking long distances. Government can also take temporary measures to suspend construction-related work, requiring work places stagger start and end times, or prohibiting certain types of traffic at specific times of the day.

Conclusions

Industrialization and increasing wealth have gone hand in hand with pollution in every country over the last three centuries. Unsurprisingly, air pollution levels in Ho Chi Minh City have significantly increased over the decades since the rapid industrialization of Vietnam. Over the last 5 years, it appears that these have been decoupled, with economic growth continuing but a plateauing and even a reversal in the growth in air pollution levels. This is observed by looking at air pollution levels in the month of December in each of the last 5 years, from a high in 2016 to a low in 2020. There is also less variability in air pollution levels, with fewer time periods of extreme and high levels. This is likely due to concerted action on the part of local authorities to reduce reliance on private vehicles, encouraging people to use public transport, and reducing congestion by varying working and school hours. Work is currently underway to invest in better public transport systems, including bus rapid transit in Ho Chi Minh City. However, with 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.

When looking at two different times of year, February 2021 and December 2020, February did experience higher levels of air pollution but the difference was quite small. However, with the COVID-19 pandemic restricting many of the pollution-causing activities that usually take place in February around the Tet festival, pollution levels in February 2021 may have been significantly lower than usual.

Weather can explain some of the variability in air pollution levels. Some methods to explain the variability in air pollution levels were not successful, providing contradictory results. Using the CramerV method, it was found that 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 the CramerV and correlation methods.

Chi-square tests showed that there is a relationship between AQI levels and four weather-related variables: weather, humidity, temperature, and wind speed. Various linear regression models were tested in order to find the best predictors for air quality. Model 5 and 6, which combine Temperature, Humidity, Wind speed and Precipitation as predictors for air quality, and which uses a 3-day time lag and 7-day time lag, respectively with air quality being measured 7 days later, achieved the adjusted R-squared increased from 11% for Model 5 to 18% for Model 6. The model that has the highest R-square, 20%, is model 8, 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.

The model with the time lag datasets explains the expert's sentiments or scientific evidence, presented in the background. Science is evidence-based that can't be observed with naked eye. For example, today it may be a normal sunny day with average speeds of wind, but a few days ago, the city could have been covered with cold air like a blanket, accumulating fine dust that cannot spread causing the pollution index to rise. It is a combination of many factors: temperature, wind, dust dispersion. This does not mean that more vehicles will cause pollution to increase.

This model can be used to predict air quality 7 days into the future, providing a useful tool for the government to take action in advance to reduce the health impacts of air pollution.

Appendix A: Description of Dataset Variables

Name of Variable	Description of Variable	Type of Variable	Source
Site	Subject location of the observations. All observations are from Ho Chi Minh City, Vietnam.	Categorical variable	U.S. Consulate General Ho Chi Minh City
Parameter	describes what is being measured by the data. Data is for pollution particles that are 2.5 microns in diameter or smaller.	Categorical variable	U.S. Consulate General Ho Chi Minh City
Date_LT	Provides the date and time the air pollution information was collected at, taken at 1 hour intervals.	Numerical variable	U.S. Consulate General Ho Chi Minh City
Year	The year in which the information was collected, all information was collected in 2019.	Numerical variable	U.S. Consulate General Ho Chi Minh City
Month	Month in which the information was collected.	Numerical variable	U.S. Consulate General Ho Chi Minh City
Day	Day of which the information was collected	Numerical variable	U.S. Consulate General Ho Chi Minh City
Hour	The time of day at which the information was collected.	Numerical variable	U.S. Consulate General Ho Chi Minh City
AQI	The AQI (air quality index) level is calculated based on the NowCast Conc. data.	Numerical variable	U.S. Consulate General Ho Chi Minh City
AQI_Category	The AQI level is converted into one of 6 AQI Categories: Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous.	Categorical variable	U.S. Consulate General Ho Chi Minh City
Raw_Conc	Raw concentration is the air pollution levels of particulates 2.5 microns in diameter or smaller and a specific point in time, measured in UG/M^3.	Numerical variable	U.S. Consulate General Ho Chi Minh City
Conc_Unit	Unit of measurement to describe the data in the previous column.	Categorical variable	U.S. Consulate General Ho Chi Minh City
Duration	Duration states how long the PM2.5 was measured for and is the standard duration for collecting air pollution data	Numerical variable	U.S. Consulate General Ho Chi Minh City

QC_Name	Quality control status of the observation. There are three variables: Valid, Invalid, and Missing.	Categorical variable	U.S. Consulate General Ho Chi Minh City
Weather	Weather condition description	Categorical variable	Historical Forecast Weather: Historical Weather Forecasts | World Weather Online
Precipitation_mm	Total precipitation amount in mm	Numerical variable	Historical Forecast Weather: Historical Weather Forecasts | World Weather Online
TempC	Temperature in degrees Celsius.	Numerical variable	Historical Forecast Weather: Historical Weather Forecasts | World Weather Online
WindspeedKmph	Wind speed in kilometers per hour	Numerical variable	Historical Forecast Weather: Historical Weather Forecasts | World Weather Online
Humidity	Humidity in percentage (%)	Numerical variable	Historical Forecast Weather: Historical Weather Forecasts | World Weather Online

Appendix B: SAS codes for combining and cleaning dataset

```
Data HCMC2016_12;

infile '/folders/myfolders/Team
project/AQI_december/HoChiMinhCity_PM2.5_2016_12_MTD.csv' dlm=',' firstobs=2 missover
dsd;

    format Site $5.

        Parameter $17.

        Date_LT_ DATETIME13.

        Year 4.

        Month 2.

        Day 2.

        Hour 2.

        AQI 4.

        AQI_Category $30.

        _24_hr_MidPt_Avg 4.1

        Raw_Conc 3.

        Conc_Unit $5.

        Duration $4.

        QC_Name $10.

        Weather_Description $29.

        Precipitation_mm 3.1

        Precipitation_inches 3.1

        Temp_C        2.
```

```
          Wind_speed_Kmph 2.

          Humidity 2.

          Heat_Index_C 2.  ;

     input Site $ Parameter $ Date_LT_:ANYDTDTM40. Year Month Day Hour AQI
AQI_Category $ _24_hr_MidPt_Avg Raw_Conc Conc_Unit $ Duration $ QC_Name $
Weather_Description $ Precipitation_mm Precipitation_inches Temp_C  Wind_speed_Kmph
Humidity Heat_Index_C ;

     options datestyle=dmy;

*cleaning dataset ;

*Changing AQI category according to the value of AQI;

if AQI_Category = 2 then AQI_Category='Moderate';

else if AQI_Category = 3 then AQI_Category='Unhealthy for Sensitive Groups';

else if AQI_Category = 4 then AQI_Category='Unhealthy';

*Removing outliers ;

if Raw_Conc=985 then delete;

*dropping 24 hr mid pt. avg column for merging purpose;

drop _24_hr_MidPt_Avg;

run;

**************************************************************************************;

Data HCMC2017_12;

infile '/folders/myfolders/Team
project/AQI_december/HoChiMinhCity_PM2.5_2017_12_MTD.csv' dlm=',' firstobs=2 missover
dsd;

    format Site $5.
```

```
        Parameter $17.

        Date_LT_ DATETIME13.

        Year 4.

        Month 2.

        Day 2.

        Hour 2.

        NowCast_Conc 5.1

        AQI 4.

        AQI_Category $30.

        Raw_Conc 5.1

        Conc_Unit $5.

        Duration $4.

        QC_Name $10.

        Weather_Description $29.

        Precipitation_mm 3.1

        Precipitation_inches 3.1

        Temp_C      2.

        Wind_speed_Kmph 2.

        Humidity 2.

        Heat_Index_C 2. ;

    input Site $ Parameter $ Date_LT_:ANYDTDTM40. Year Month Day Hour NowCast_Conc
AQI AQI_Category $ Raw_Conc Conc_Unit $ Duration $ QC_Name $ Weather_Description $
Precipitation_mm Precipitation_inches Temp_C Wind_speed_Kmph Humidity Heat_Index_C ;
```

```
    options datestyle=dmy;
*Removing missing values;
if Raw_Conc=-999 or Raw_Conc=985 then delete;
if NowCast_Conc=-999 then delete;
if AQI=-999 then delete;
*dropping NowCast_Conc column for merging purpose;
drop NowCast_Conc;
run;
*******************************************************************;
Data HCMC2018_12;
infile '/folders/myfolders/Team project/AQI_december/HoChiMinhCity_PM2.5_2018_12_MTD.csv' dlm=',' firstobs=2 missover dsd;
    format Site $5.
         Parameter $17.
         Date_LT_ DATETIME13.
         Year 4.
         Month 2.
         Day 2.
         Hour 2.
         NowCast_Conc 5.1
         AQI 4.
         AQI_Category $30.
```

```
          Raw_Conc 5.1
          Conc_Unit $5.
          Duration $4.
          QC_Name $10.
          Weather_Description $29.
          Precipitation_mm 3.1
          Precipitation_inches 3.1
          Temp_C         2.
          Wind_speed_Kmph 2.
          Humidity 2.
          Heat_Index_C 2. ;
     input Site $ Parameter $ Date_LT_:ANYDTDTM40. Year Month Day Hour NowCast_Conc
AQI AQI_Category $ Raw_Conc Conc_Unit $ Duration $ QC_Name $ Weather_Description $
Precipitation_mm Precipitation_inches Temp_C Wind_speed_Kmph Humidity Heat_Index_C  ;
     options datestyle=dmy;
  *data cleaning;
*removing missing data;
if NowCast_Conc= -999 then delete;
if AQI= -999 then delete;
*there are missing data of -999, outliers 985 and 0 as values for raw conc. removing them;
if Raw_Conc <1 or Raw_Conc=985 then delete;
 *dropping NowCast_Conc column for merging purpose;
drop NowCast_Conc;
```

```
run;
**********************************************************************;
Data HCMC2019_12;
infile '/folders/myfolders/Team project/AQI_december/HoChiMinhCity_PM2.5_2019_12_MTD.csv' dlm=',' firstobs=2 missover dsd;
    format Site $5.
         Parameter $17.
         Date_LT_ DATETIME13.
         Year 4.
         Month 2.
         Day 2.
         Hour 2.
         NowCast_Conc 5.1
         AQI 4.
         AQI_Category $30.
         Raw_Conc 5.1
         Conc_Unit $5.
         Duration $4.
         QC_Name $10.
         Weather_Description $29.
         Precipitation_mm 3.1
```

```
            Precipitation_inches 3.1
            Temp_C        2.
            Wind_speed_Kmph 2.
            Humidity 2.
            Heat_Index_C 2. ;
      input Site $ Parameter $ Date_LT_:ANYDTDTM40. Year Month Day Hour NowCast_Conc AQI AQI_Category $ Raw_Conc Conc_Unit $ Duration $ QC_Name $ Weather_Description $ Precipitation_mm Precipitation_inches Temp_C Wind_speed_Kmph Humidity Heat_Index_C  ;
      options datestyle=dmy;
*data cleaning;
*removing missing data;
if NowCast_Conc= -999 then delete;
if AQI= -999 then delete;
*removing outliers;
if Raw_Conc=985 then delete;
 *dropping NowCast_Conc column for merging purpose;
drop NowCast_Conc;
run;
*********************************************************************;
Data HCMC2020_12;
infile '/folders/myfolders/Team project/AQI_december/HoChiMinhCity_PM2.5_2020_12_MTD.csv' dlm=',' firstobs=2 missover dsd;
     format Site $5.
```

```
          Parameter $17.
          Date_LT_ DATETIME13.
          Year 4.
          Month 2.
          Day 2.
          Hour 2.
          NowCast_Conc 5.1
          AQI 4.
          AQI_Category $30.
          Raw_Conc 5.1
          Conc_Unit $5.
          Duration $4.
          QC_Name $10.
          Weather_Description $29.
          Precipitation_mm 3.1
          Precipitation_inches 3.1
          Temp_C      2.
          Wind_speed_Kmph 2.
          Humidity 2.
          Heat_Index_C 2.;
     input Site $ Parameter $ Date_LT_:ANYDTDTM40. Year Month Day Hour NowCast_Conc
AQI AQI_Category $ Raw_Conc Conc_Unit $ Duration $ QC_Name $ Weather_Description $
Precipitation_mm Precipitation_inches Temp_C Wind_speed_Kmph Humidity Heat_Index_C ;
```

```
    options datestyle=dmy;

*data cleaning;

*removing missing data ;

if NowCast_Conc= -999 then delete;

if AQI= -999 then delete;

*there are missing data of -999 and outliers 985 as values for raw conc. - Removing them;

if Raw_Conc <1 or Raw_Conc=985 then delete;

*dropping NowCast_Conc column for merging purpose;

drop NowCast_Conc;
run;

***********************************************************************;

Data HCMC2021_02;

infile '/folders/myfolders/Team
project/AQI_december/HoChiMinhCity_PM2.5_2021_02_MTD.csv' dlm=',' firstobs=2 missover
dsd;

    format Site $5.

        Parameter $17.

        Date_LT_ DATETIME13.

        Year 4.

        Month 2.

        Day 2.

        Hour 2.

        NowCast_Conc 5.1
```

```
          AQI 4.
          AQI_Category $30.
          Raw_Conc 5.1
          Conc_Unit $5.
          Duration $4.
          QC_Name $10.
          Weather_Description $29.
          Precipitation_mm 3.1
          Precipitation_inches 3.1
          Temp_C        2.
          Wind_speed_Kmph 2.
          Humidity 2.
          Heat_Index_C 2.;
     input Site $ Parameter $ Date_LT_:ANYDTDTM40. Year Month Day Hour NowCast_Conc
AQI AQI_Category $ Raw_Conc Conc_Unit $ Duration $ QC_Name $ Weather_Description
$  Precipitation_mm Precipitation_inches Temp_C Wind_speed_Kmph Humidity Heat_Index_C
;
     options datestyle=dmy;
*dropping NowCast_Conc column for merging purpose;
drop NowCast_Conc;
run;
*********************************************************************;
*********************************Merging****************************;
```

```
Data HCMC_master;

*Merging all the datasets;

set  HCMC2016_12  HCMC2017_12  HCMC2018_12  HCMC2019_12  HCMC2020_12
HCMC2021_02 ;

run;

**Lockdown Analysis**;

*************************************************************************************;

Data HCMC2020_12_lockdown;

infile '/folders/myfolders/Team project/lockdown
analysis/HoChiMinhCity_PM2.5_2020_12_lockdown.csv' dlm=',' firstobs=2 missover dsd;

    format

         Site $5.

         Parameter $17.

         Date_LT_ DATETIME13.

         Year 4.

         Month 2.

         Day 2.

         Hour 2.

         NowCast_Conc 5.1

         AQI 4.

         AQI_Category $30.

         Raw_Conc 5.1

         Conc_Unit $5.
```

```
          Duration $4.

          QC_Name $10.;

     input Site $ Parameter $ Date_LT_:ANYDTDTM40. Year Month Day Hour NowCast_Conc
AQI AQI_Category $ Raw_Conc Conc_Unit $ Duration $ QC_Name $ ;

     options datestyle=dmy;

*data cleaning;

*deleting  missing data;

if NowCast_Conc= -999 then delete;

if AQI= -999 then delete;

*there are missing data of -999 and outliers 985 as values for raw conc. - deleting them;

if Raw_Conc <1 or Raw_Conc=985 then delete;

*Changing AQI category according to AQI value;

If AQI >=51 and AQI <=100 then AQI_Category ='Moderate';

run;

*************************************************************************;

Data HCMC2021_02_lockdown;

infile '/folders/myfolders/Team project/lockdown
analysis/HoChiMinhCity_PM2.5_2021_02_lockdown.csv' dlm=',' firstobs=2 missover dsd;

     format Site $5.

          Parameter $17.

          Date_LT_ DATETIME13.

          Year 4.

          Month 2.
```

```
        Day 2.
        Hour 2.
        NowCast_Conc 5.1
        AQI 4.
        AQI_Category $30.
        Raw_Conc 5.1
        Conc_Unit $5.
        Duration $4.
        QC_Name $10.;

    input Site $ Parameter $ Date_LT_:ANYDTDTM40. Year Month Day Hour NowCast_Conc AQI AQI_Category $ Raw_Conc Conc_Unit $ Duration $ QC_Name $ ;
    options datestyle=dmy;
*data Cleaning;
*deleting outliers;
if Raw_Conc <1 or Raw_Conc=985 then delete;
run;
****************************Merging******************************;
***merging 2 datasets for post and pre lockdown analysis***;
data lockdown_analysis;
set HCMC2020_12_lockdown HCMC2021_02_lockdown ;
run;
```

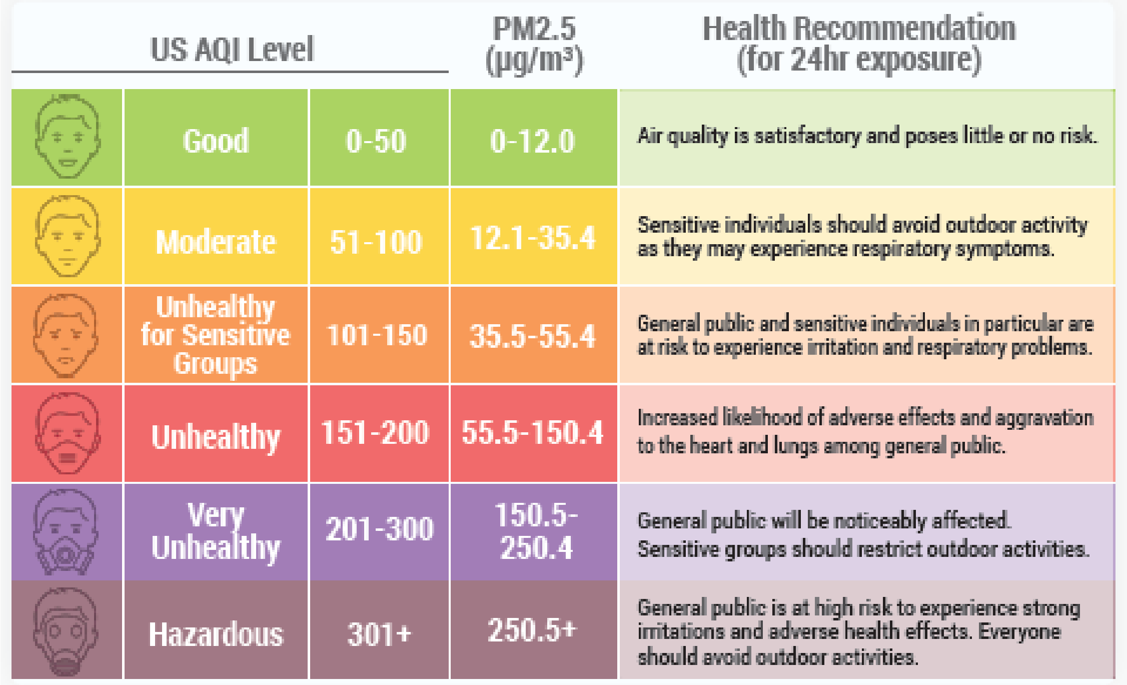
Appendix C: NowCast, AQI levels, and AQI categories

NowCast using the following formula to take the Raw Concentration data from the previous 12 hours, weighting recent observations more heavily:

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

C_1, C_2…. C_{12} are the Raw Concentration levels and W^i is the weighting.

The following chart shows how the NowCast data is converted into the AQI levels and the AQI categories:



US AQI Level			PM2.5 (µg/m³)	Health Recommendation (for 24hr exposure)
	Good	0-50	0-12.0	Air quality is satisfactory and poses little or no risk.
	Moderate	51-100	12.1-35.4	Sensitive individuals should avoid outdoor activity as they may experience respiratory symptoms.
	Unhealthy for Sensitive Groups	101-150	35.5-55.4	General public and sensitive individuals in particular are at risk to experience irritation and respiratory problems.
	Unhealthy	151-200	55.5-150.4	Increased likelihood of adverse effects and aggravation to the heart and lungs among general public.
	Very Unhealthy	201-300	150.5-250.4	General public will be noticeably affected. Sensitive groups should restrict outdoor activities.
	Hazardous	301+	250.5+	General public is at high risk to experience strong irritations and adverse health effects. Everyone should avoid outdoor activities.

Appendix D: SAS codes for changes in air quality 2016-2017

AQI Category Frequencies

```
proc sort data=PROJECT.QUERY out=Work.SortTempTableSorted;
    by Year;
run;
proc freq data=Work.SortTempTableSorted;
    tables AQI_Category / nocum plots=(freqplot);
    by Year;
run;
proc delete data=Work.SortTempTableSorted;
run;
```

AQI and Raw Concentration box plots

```
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=PROJECT.QUERY;
    title height=14pt "AQI by Year";
    vbox AQI / category=Year;
    yaxis grid;
run;
ods graphics / reset;
title;
graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=PROJECT.QUERY;
    title height=14pt "Raw Concentration by Year";
    vbox Raw_Conc / category=Year;
    yaxis grid;
run;
ods graphics / reset;
title;
```

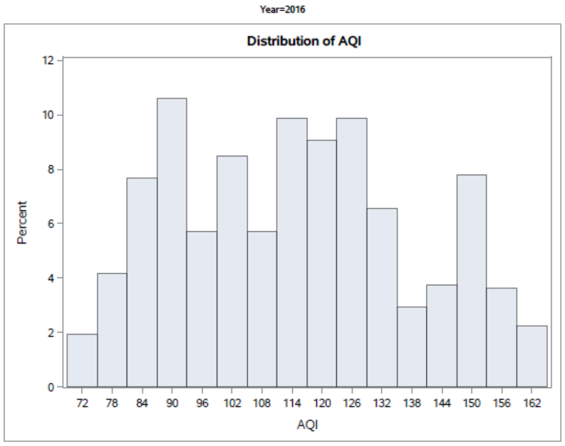
AQI and Raw Concentration summary statistics

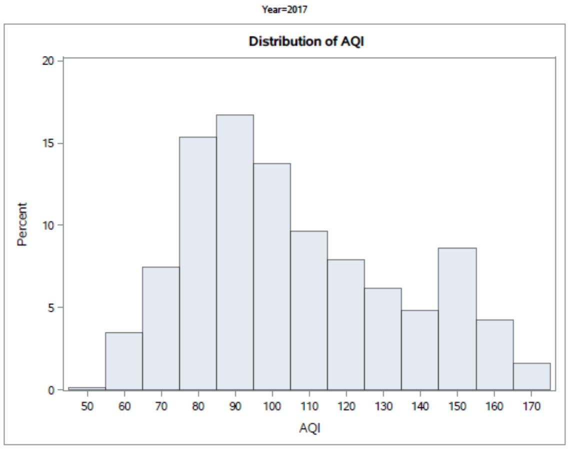
```
ods noproctitle;
ods graphics / imagemap=on;
proc sort data=PROJECT.QUERY out=WORK.TempSorted2236;
    by Year;
run;
proc means data=WORK.TempSorted2236 chartype mean std min max median stderr var
        mode range vardef=df cv q1 q3 qrange qmethod=os;
    var AQI Raw_Conc;
    by Year;
run;
```

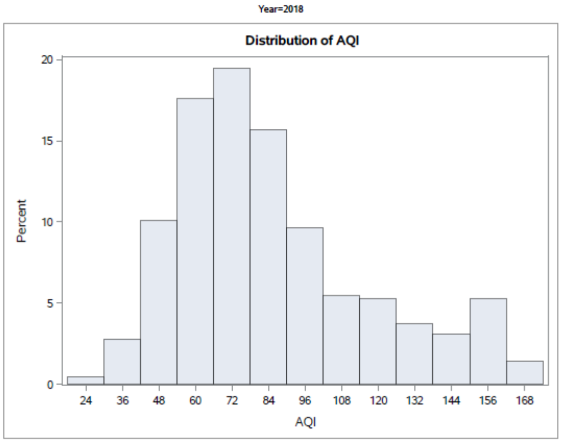
```
proc univariate data=WORK.TempSorted2236 vardef=df noprint;
	var AQI Raw_Conc;
	histogram AQI Raw_Conc;
	by Year;
run;
proc datasets library=WORK noprint;
	delete TempSorted2236;
	run;
```

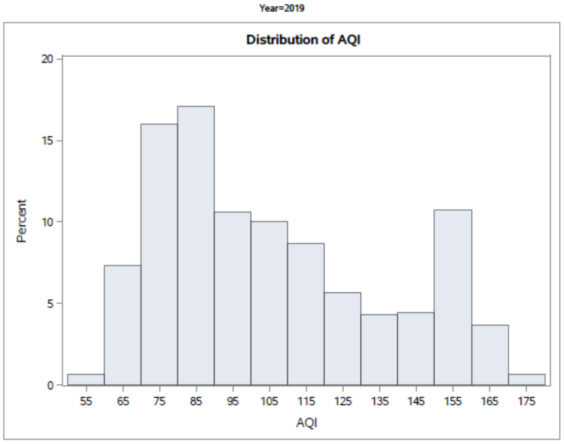
Appendix E: Distribution of AQI and Raw Concentration Levels

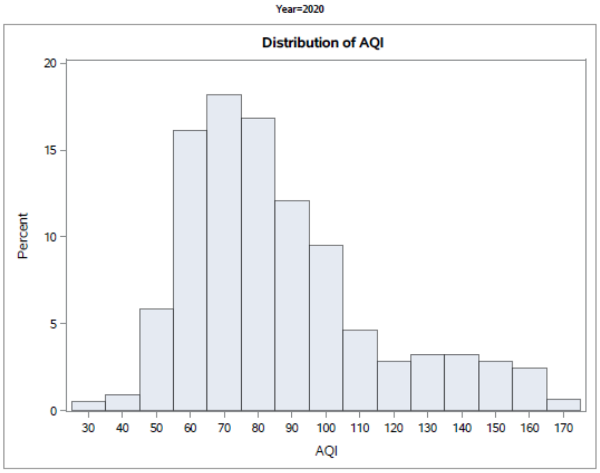
The graphs below show the frequency of AQI levels:



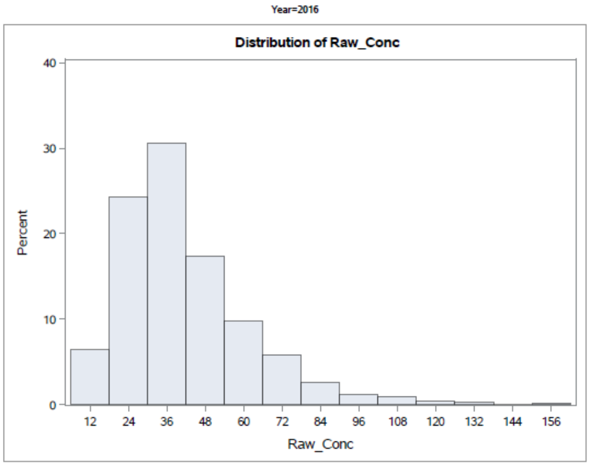


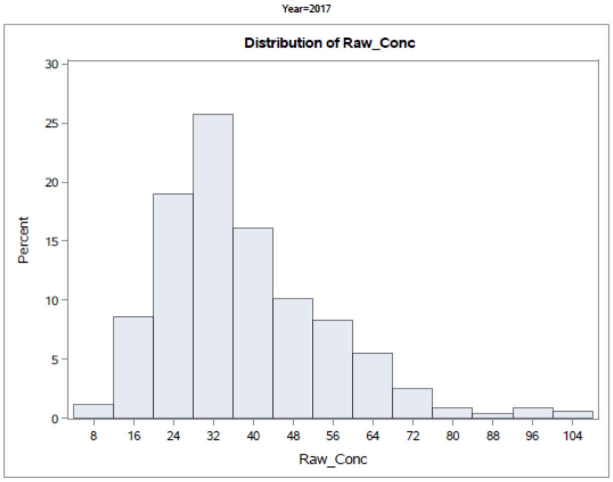


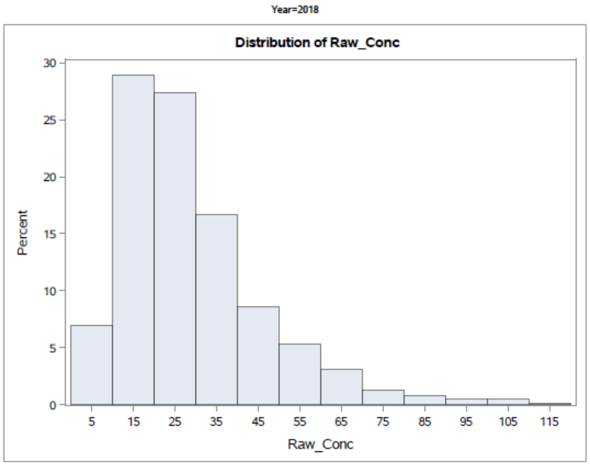


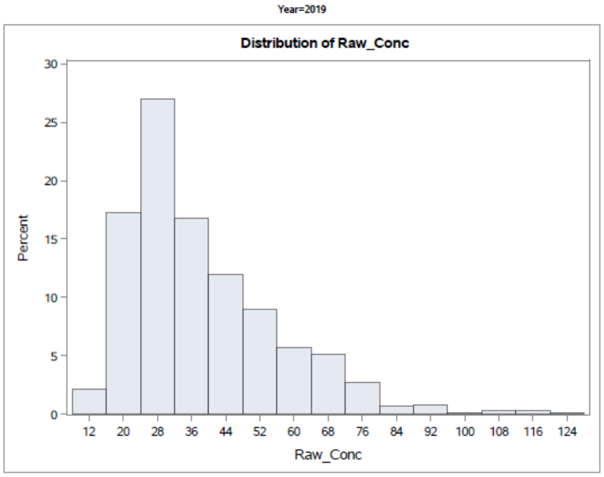


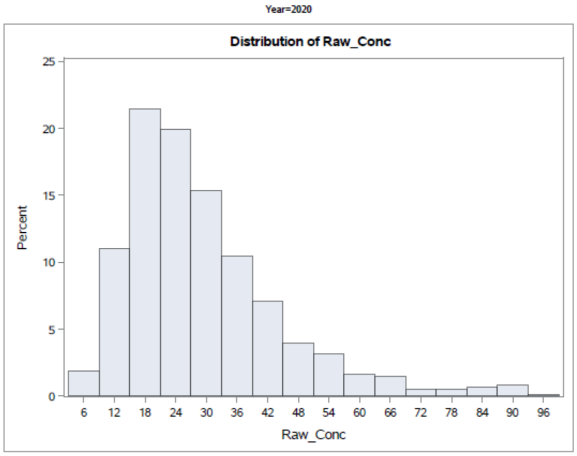
The graphs below show the frequency of Raw Concentration levels:











Appendix F: SAS Codes for Changes in Air Quality (February 2021 vs December 2020)

```
/**Deleting Data for month 1**/
proc sql;
delete from DECVSFEB_ANALYSIS
where month = 1;
run;


/**Descriptive statistics of AQI**/
ods noproctitle;
proc sort data=WORK.DECVSFEB_ANALYSIS out=_chardata_sorted;
        by Month;
run;
title "Descriptive Statistics for Numeric Variables";
proc means data=_chardata_sorted n nmiss min mean median max std;
        by Month;
        var AQI;
run;
title;
proc univariate data=_chardata_sorted noprint;
        by Month;
        histogram AQI;
run;
proc delete data=_chardata_sorted;
run;


/**BoxPlot AQI by month**/
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=WORK.DECVSFEB_ANALYSIS;
```

```
        vbox AQI / category=Month;
        yaxis grid;
run;
ods graphics / reset;

/**Descriptive statistics of AQI_Category**/
proc sort data=DECVSFEB_ANALYSIS out=Work.SortTempTableSorted;
    by Month;
run;
proc freq data=Work.SortTempTableSorted;
    tables AQI_Category / nocum plots=(freqplot);
    by Month;
run;
proc delete data=Work.SortTempTableSorted;
run;

/**side-by-side bar chart of AQI_Category feb vs dec**/
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=WORK.DECVSFEB_ANALYSIS;
        vbar AQI_Category / group=Month groupdisplay=cluster stat=percent datalabel;
        yaxis grid;
run;
ods graphics / reset;

/**Descriptive statistics of Raw_Conc**/
ods noproctitle;
proc sort data=WORK.DECVSFEB_ANALYSIS out=_chardata_sorted;
        by Month;
```

```
run;
title "Descriptive Statistics for Numeric Variables";
proc means data=_chardata_sorted n nmiss min mean median max std;
        by Month;
        var Raw_Conc;
run;
title;
proc univariate data=_chardata_sorted noprint;
        by Month;
        histogram Raw_Conc;
run;
proc delete data=_chardata_sorted;
run;

/**BoxPlot Raw_Conc by month**/
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=WORK.DECVSFEB_ANALYSIS;
        vbox Raw_Conc / category=Month;
        yaxis grid;
run;
ods graphics / reset;
```

Appendix G: R- Codes for Relationship between Weather and Air Quality

```
setwd('C:/Users/LIEN PHAM/Desktop/Langara/Data analysis and statistics infer/Project/Team
Project attached files')
getwd
library("readxl")
library(tidyverse)
library(lubridate)

link = dir("C:/Users/LIEN PHAM/Desktop/Langara/Data analysis and statistics
infer/Project/Team Project attached files",
                    full.names = FALSE)
str(link)

link = link[str_detect(link, "csv")]

df_17_21 <- data.frame()
for (i in (1:length(link))){

  # Import file

  df <- read.csv(file = paste0("C:/Users/LIEN PHAM/Desktop/Langara/Data analysis and
statistics infer/Project/Team Project attached files/",
                   link[i]))
  # Merge file
  df_17_21%>%
    rbind(df) -> df_17_21
}

df_17_21= df_17_21 %>%
  select(- NowCast.Conc.)

# $ Date..LT.   : chr  "12/1/2017 1:00" "12/1/2017 2:00" "12/1/2017 3:00" "12/1/2017 4:00" ...

df_2016 <- read.csv("C:/Users/LIEN PHAM/Desktop/Langara/Data analysis and statistics
infer/Project/HoChiMinhCity_PM2.5_2016_12_MTD.csv")

df_2016 = df_2016 %>%
  select(- X24.hr..Midpoint.Avg..Conc.)

# Merge all the raw files (2016 - 2021) for checking missing data purpose
df_master = rbind(df_2016,df_17_21)
View(df_master)

#4224 obs. of  13 variables
# checking NA/missing values for all variables (NO missing values for all variables)
```

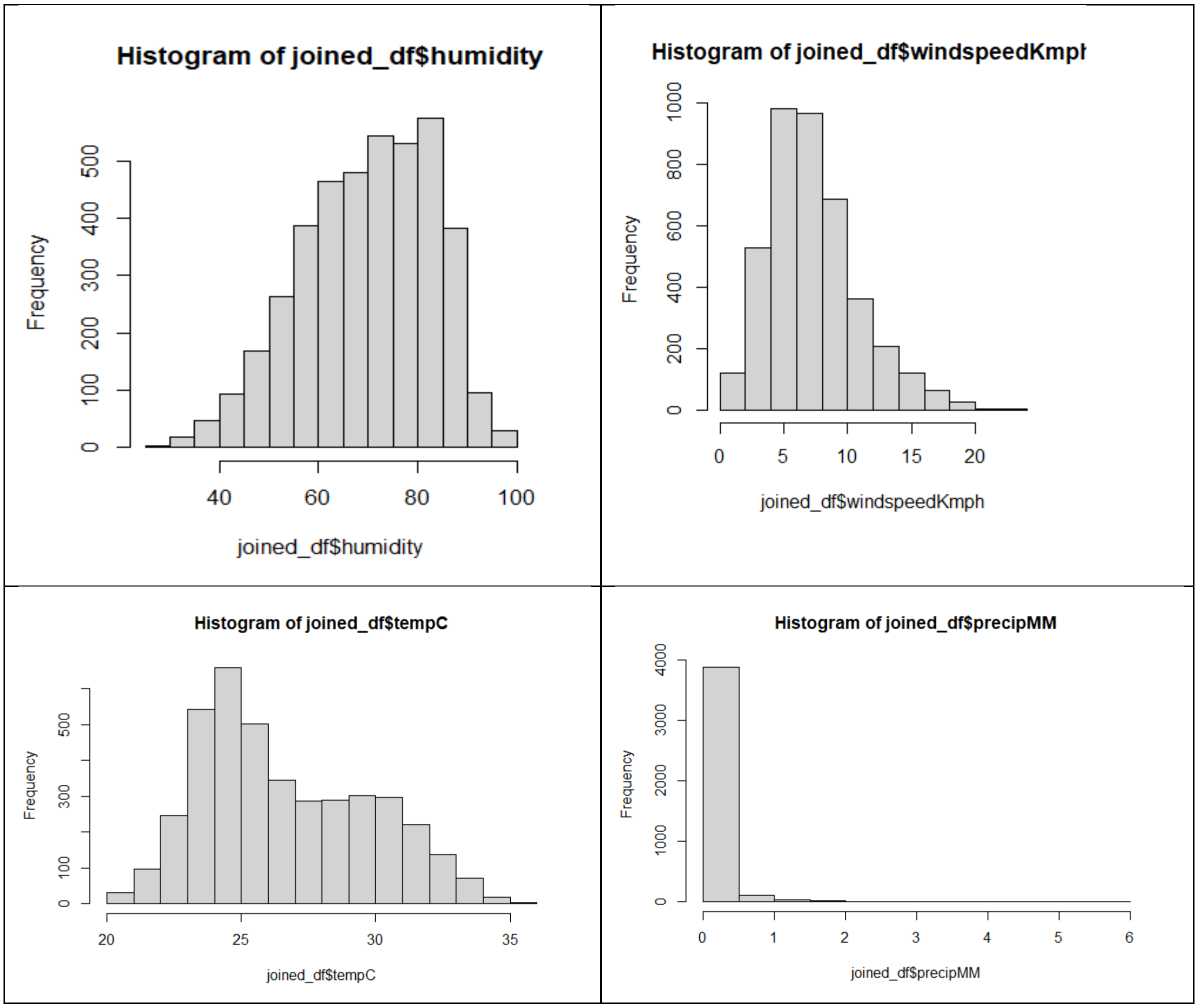
```
df_master %>%
  select(everything()) %>%
  summarise_all(funs(sum(is.na(.)))) -> missing_values

# checking outliners (147 obs with outliners)
df_master %>%
  filter(!AQI>=0 |!AQI<=176) -> df_master_check_outliners
View(df_master_check_outliners)

df_master %>%
  filter(!Raw.Conc. >0 |! Raw.Conc. < 985) -> df_master_check_outliners_Rawconc
View(df_master_check_outliners_Rawconc) # 28 obs removed

df_master = df_master %>%
  filter( AQI>=0 & AQI<=176,Raw.Conc. >0 & Raw.Conc. < 985 )

View(df_master) #(4049 obs including all years: 2016,2017,2018,2019,2020,2021)

external_data <- read.csv("C:/Users/LIEN PHAM/Desktop/Langara/Data analysis and statistics
infer/Project/weather_data_1hr.csv")

external_data = external_data %>%

select(date,time,weatherDesc,precipMM,precipInches,tempC,windspeedKmph,humidity,HeatInd
exC)
external_data = external_data %>%
  mutate(Time = time/100)
external_data = external_data %>%
  select(-time)

# change format of Date for external data
external_data$date_new = as.Date(external_data$date,format = c("%m/%d/%Y"))

# change format of Date for master data for merging
as.Date ("2018-12-01 01:00 AM", format = "%Y-%m-%d")
# "2018-12-01"

as.Date ("12/1/2017 1:00", format = "%m-%d-%Y")

df_master$date_new = as.Date(df_master$Date..LT.,format = c("%m/%d/%Y"))

df_master[is.na(df_master$date_new), "date_new"] =
as.Date(df_master[is.na(df_master$date_new), "Date..LT."], format = "%Y-%m-%d")

df_master[is.na(df_master$date_new), c("date_new", "Date..LT.")]
```

```
# combine master file with external dataset
library(lubridate)

joined_df <- left_join(df_master,external_data, by = c("date_new" = "date_new", "Hour" =
"Time"))

View(joined_df)

joined_df[2977, c("date_new", "Hour", "Date..LT.")]

df_master[is.na(df_master$date_new),]

join_df_16_20 = joined_df %>%
  filter(date_new >= "2016-12-01" & date_new <= "2020-12-31")

max(joined_df$date_new, na.rm = T) # check
unique(joined_df$date_new)

join_df_16_20 = joined_df %>%
  filter(date_new >= "2016-12-01" & date_new <= "2020-12-31")

unique(joined_df$AQI.Category)

# fix the AQI categories for the consistency
joined_df = joined_df %>%
  filter(AQI > 0) %>%
  mutate(AQI_bin = case_when(AQI.Category == "2" ~ "Moderate",
                          AQI.Category == "3" ~ "Unhealthy for Sensitive Groups",
                          AQI.Category == "4" ~ "Unhealthy",
                          TRUE ~ AQI.Category))

joined_df$AQI_bin = as.factor(joined_df$AQI_bin)

unique(joined_df_fix$weatherDesc)

joined_df %>% group_by(weatherDesc) %>% count()

write_xlsx(joined_df,'C:/Users/LIEN PHAM/Desktop/Langara/Data analysis and statistics
infer/Project/joined_df.xlsx')

####################################################################################

df_daily_all_variables = joined_df %>%
  group_by(date_new) %>%
  summarise(daily_AQI = sum(AQI),
        daily_wind = sum(windspeedKmph),
        daily_tempC = sum(tempC),
        daily_preci = sum(precipMM),
        daily_hum = sum(humidity))

library("writexl")
write_xlsx(df_daily_all_variables,'C:/Users/LIEN PHAM/Desktop/Langara/Data analysis and
statistics infer/Project/df_daily_all_variables.xlsx')


df_daily_AQI = joined_df %>%
  group_by(date_new) %>%
  summarise(daily_AQI = sum(AQI))

df_daily_AQI$date_new = as.Date(df_daily_AQI$date_new)

# plot daily AQI to find the dates with max AQI (3823 for 10/12/2016)
library(dplyr)
ggplot(data = df_daily_AQI) +
  geom_line(aes(date_new,daily_AQI))+
  geom_point(aes(date_new,daily_AQI)) +
  geom_text(data = df_daily_AQI %>% dplyr::filter(daily_AQI ==
max(df_daily_AQI$daily_AQI)),
        aes(date_new,daily_AQI, label = date_new), vjust = -0.3)

# Using yearly data, generate: a boxplot and a bar plot (choose one of variables)

joined_df$Year = as.factor(joined_df$Year)

joined_df %>% ggplot() + geom_boxplot(aes(x = Year, y = AQI, fill = Year))

joined_df %>% ggplot() + geom_col(aes(x = Year, y = AQI, fill = Year))

# The distributions of AQI, EACH month of HCMC
joined_df %>% ggplot() + geom_histogram(aes(AQI), fill = "lightblue", binwidth = 15) +
  facet_wrap(~ Year)

###############################################
# caterogize the wind_speed
library(lsr)
install.packages("DescTools")
library(DescTools)

joined_df %>%
  arrange(Wind_speed_Kmph) %>%
  View()
```

```
hist(joined_df$windspeedKmph)

quantile_windspeed <- quantile(joined_df$windspeedKmph,probs = seq(0, 1, 1/5))
group_windspeed <- cut(joined_df$windspeedKmph,c(0,5,6,8,10,24), labels =
c("very_low","low","medium","high","very_high"),
                    right = FALSE, include.lowest = TRUE)
table(group_windspeed, joined_df$AQI_bin)

tab <- table(group_windspeed, joined_df$AQI_bin)
CramerV(tab, bias.correct = FALSE) # r = 0.1165437 (Windspeed & AQI)
Phi(tab)
ContCoef(tab)
TschuprowT(tab)

tab4 <- table(joined_df$weatherDesc, joined_df$AQI_bin)
CramerV(tab4, bias.correct = FALSE) # r = 0.1334249 (weather & AQI)

# caterogize the precipMM
joined_df %>%
  arrange(precipMM) %>%
  View()
hist(joined_df$precipMM)
quantile_preciMM <- quantile(joined_df$precipMM, probs = seq(0, 1, 1/5))
group_preciMM  <- cut(joined_df$windspeedKmph,c(0,0.5,2,5.6), labels =
c("low","medium","high"),
                   right = FALSE, include.lowest = TRUE)
tab1 <- table(group_preciMM,joined_df$AQI_bin)
CramerV(tab1, bias.correct = FALSE)  # 0.03355126 (AQI & preciMM)


# caterogize the tempC
joined_df %>%
  arrange(tempC) %>%
  View()
hist(joined_df$tempC)
quantile_tempC <- quantile(joined_df$tempC,probs = seq(0, 1, 1/3))
group_tempC <- cut(joined_df$tempC,c( 20,25,28,36),labels = c("low","medium","high"),
                   right = FALSE, include.lowest = TRUE)
tab2<- table(group_tempC,joined_df$AQI_bin)
CramerV(tab2, bias.correct = FALSE)  # 0.1015877 (AQI & tempC)

joined_df %>% group_by(tempC) %>% count()

# caterogize the humidity
joined_df %>%
  arrange(humidity) %>%
  View()
hist(joined_df$humidity)
quantile_humidity <- quantile(joined_df$humidity,probs = seq(0, 1, 1/5))
group_humidity <- cut(joined_df$humidity,c(29,59,68,75,83,98), labels =
c("very_low","low","medium","high","very_high"),
                 right = FALSE, include.lowest = TRUE)
tab3 <- table(group_humidity,joined_df$AQI_bin)
CramerV(tab3, bias.correct = FALSE) # 0.09740825 (AQI & humidity)

#############################################
#correlation relationship (for numeric variables)
summary <- summary(joined_df[,c(8,10, 17:22)])

joined_df_re_col<- joined_df[,-c(1:7,9:16,23)] # remove columns (chr)
all_correlation <-cor(joined_df_re_col, use = "complete.obs")

install.packages("Hmisc")
library("Hmisc")
plot(AQI ~ tempC + windspeedKmph + tempC*windspeedKmph, data = joined_df_re_col, main
= "Plot")
pairs(~tempC + windspeedKmph + tempC*windspeedKmph,data = joined_df_re_col,
     main="Simple Scatterplot Matrix")

#############################################
# Linear model
lag_data <- read_excel("df_daily_all_variables.xlsx",sheet = "Sheet2")
LM_lag_data <- lm (AQI ~ tempC + humidity + precipMM + windspeedKmph, data = joined_df
)
summary(LM)

LM_mul <- lm(AQI ~ tempC + windspeedKmph + tempC*windspeedKmph, data = joined_df)
summary(LM_mul)

############################################
LM <- lm (AQI ~ tempC + humidity + precipMM + windspeedKmph, data = joined_df)
summary(LM)

lm_1<- lm (AQI ~ tempC , data = joined_df)
summary(lm_1)
abline(lm_1)

lm_2<- lm (AQI ~ humidity , data = joined_df)
summary(lm_2)
abline(lm_2)

lm_3<- lm (AQI ~ precipMM , data = joined_df)
```

```
summary(lm_3)
abline(lm_3)

lm_4<- lm (AQI ~ windspeedKmph, data = joined_df)
summary(lm_4)
abline(lm_4)

#################################################################
# Perform LM with lag-time dataset

setwd('C:/Users/LIEN PHAM/Desktop/Langara/Data analysis and statistics infer/Project')
getwd
library("readxl")
df_3 = read_excel("joined_df _lag_3.xlsx")
df_7 = read_excel("joined_df _lag_7.xlsx")

lm_6 <- lm(AQI ~ tempC + humidity + precipMM + windspeedKmph, data = df_3)
summary(lm_6)

LM_mul_lag7 <- lm(AQI ~ tempC + windspeedKmph + tempC*windspeedKmph, data = df_7)
summary(LM_mul_lag7)

LM_mul_lag7_tempC <- lm(AQI ~ tempC + humidity + precipMM + windspeedKmph +
tempC*windspeedKmph + tempC*humidity + tempC*precipMM, data = df)
summary(LM_mul_lag7_tempC)

LM_mul_lag7_all <- lm(AQI ~ tempC + humidity + precipMM + windspeedKmph +
tempC*windspeedKmph
                  + tempC*humidity + tempC*precipMM + humidity*precipMM +
humidity*windspeedKmph
                  + precipMM*windspeedKmph, data = df_7)

MD8 <- summary(LM_mul_lag7_all)
plot(LM_mul_lag7_all, which = 1)

###############################################################
#  find dates

date_day <- weekdays(as.Date(joined_df$date_new))
joined_df$date_day = date_day
View(joined_df)

# total 1195 weekend days and 573 days with Unhealthy or Unhealthy for sensitive groups
(accounts for 49%)
weekend = joined_df %>%
  filter(date_day%in% c('Saturday','Sunday') &
        AQI_bin %in% c('Unhealthy for Sensitive Groups','Unhealthy'))

as.factor(weekend$date_day)
weekend %>% ggplot() + geom_bar(mapping = aes(x = AQI))
#######################################################################
# Perform chi-square test

chisq.test(joined_df$AQI_bin,joined_df$weatherDesc)
chisq.test(joined_df$AQI_bin,group_humidity)
chisq.test(joined_df$AQI_bin,group_tempC)
chisq.test(joined_df$AQI_bin,group_preciMM)
chisq.test(joined_df$AQI_bin,group_windspeed)
```

Appendix H: Distribution of variables and example of binning techniques



Example of binning technique for ‘windspeed’ variable

```
quantile_windspeed
  0%  20%  40%  60%  80% 100%
   0    5    6    8   10   24
group_windspeed
 very_low       low    medium      high very_high
     1082       550       964       687       794

> t <- table(group_windspeed, joined_df$AQI_bin)
> t
group_windspeed Good Moderate Unhealthy Unhealthy for Sensitive Groups
      very_low    10      353        68                            217
      low          5      224        49                            152
      medium      12      532       125                            360
      high        13      494        90                            288
      very_high   48      748        55                            206
```