AIR POLLUTION REPORT OF SHANGHAI

Explanatory Data Analysis in SAS

BY EZGI POLAT

Introduction

In this Air Pollution Report of Shanghai, we used Air Now real-time data to analyze the air quality levels in Shanghai city of China.

Shanghai is located in China on the southern estuary of the Yangtze River. In 2019 the population was estimated to be around 25 million. This figure reflects the people who are registered as living there but does not take into account the transient workers. It is the largest populous urban area in China, surpassing the capital, Beijing. It is a very busy city because of its location and the Port of Shanghai is the busiest container port in the world. In 2018 the port handled 42 million 20-foot-long containers, 259 cruise vessels and 1.89 million passengers. (Shanghai Air Quality Index (AQI) and China Air Pollution | AirVisual, 2022)

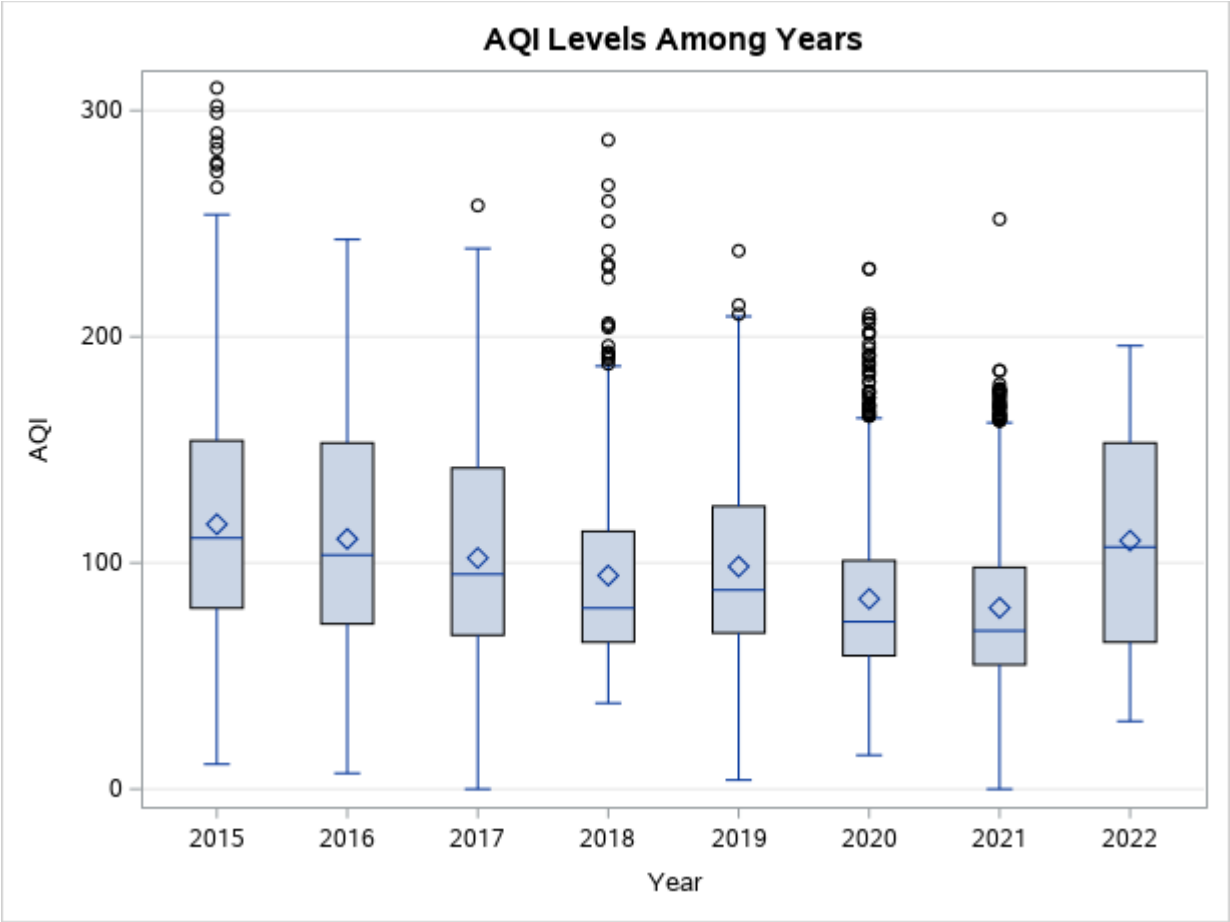
Summary

Data cleaning techniques to clean Air Now data by removing missing values. In the end, we created a 6-hourly dataset to use in our analysis. This report talks about the improvement in air quality levels in Shanghai over the years.

Our analysis showed us that there is a clear improvement in air pollution levels in Shanghai. Air pollution decreases over the years. These can be related to the Air Quality Plans of the Chinese Government.

Table of Contents

1. Comparison of Air Pollution Levels on Yearly Basis



We used box plot to compare air pollution levels. Box plots are helpful to see the distribution, skewness, median and interquartile range of the data. We will analyze the box plots in these 4 steps:

1) Comparison of the Interquartile Ranges and Whiskers

2) Comparison of the Medians

3) Identifying Outliers

4) The Skewness

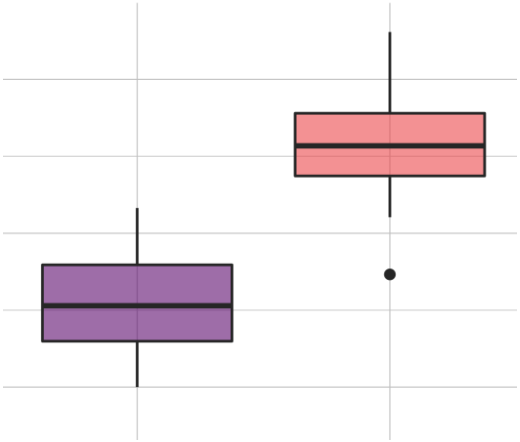
Comparison of the Interquartile Ranges and Whiskers

The lengths of the boxes representing the interquartile range differ in each year. To understand the actual size of the ranges, we can look at the Q1-Q3 values demonstrated below.

Year	Q1	Q3	Change	IQR
2015	80	154	-	74
2016	73	153	⇦⇨	80
2017	68	142	⇨|⇦	74
2018	65	114	⇨|⇦	49
2019	69	125	⇦⇨	56
2020	59	101	⇨|⇦	42
2021	55	98	⇨|⇦	43
2022	65	153	⇦⇨	88

In 2015, the interquartile range is 80 – 154. And in 2016, the length of the box got longer with the 73-153 values. The length of the box tells us how dispersed data is. Therefore, we can conclude that the 2016 data is **more dispersed** compared to the 2015 data because the length of the box got **longer**. To compare the change between all years, we can look at the 'Change' section of the table above. If the length of the box got longer compared to the previous year, the change is represented by '⇦⇨' icon. And if the length got shorter, the change is represented by '⇨|⇦' icon.

From the data, we can see that the longest box belongs to the year 2016 (excluding 2022) with an interquartile range of 80. And 2020 is the most concentrated distribution because it has the smallest IQR with 42. Because the year 2020 has a way shorter box (or smaller IQR), we can say that the more of its data is accumulated around the median. On the other hand, the year 2016 has a bigger IQR compared to all other years. When the IQR is noticeably bigger, we can say that the data has more variety.



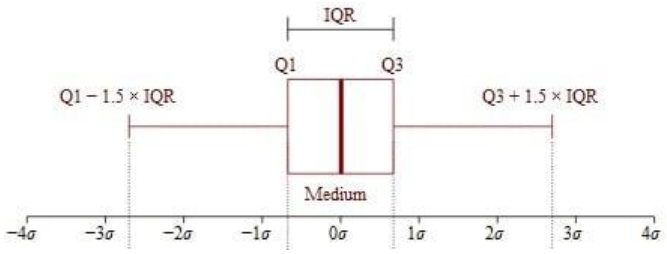
We can also look at the alignments of the boxes to see if they **overlap** with each other or not. If the boxes don't overlap with each other at all (like in the image below), then the 2 data are very different than each other. However, in our plot we can see that every year overlaps with its previous and next year's box somehow. Meaning that at least one of the lower and upper quartile values of that year aligns with the next or previous year's box.

Another way of understanding the distribution of the data is to evaluate the **whiskers**. The whiskers show us

how wide the distribution of the data is. The lower adjacent represents the value ‘min’ and the upper adjacent represents the value ‘max’ (excluding the outliers). The formula for the min and max values is demonstrated below.

Min = 1 - 1.5 * IQR

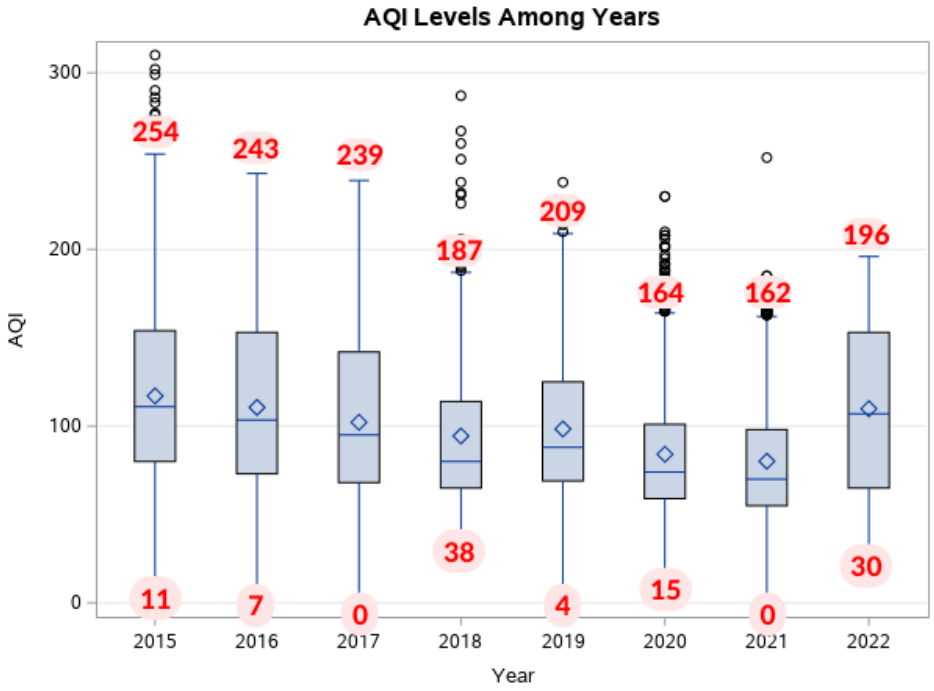
Max = Q3 + 1.5 * IQR



Source ‘https://www.statisticshowto.com/probability-and-statistics/interquartile-range/’

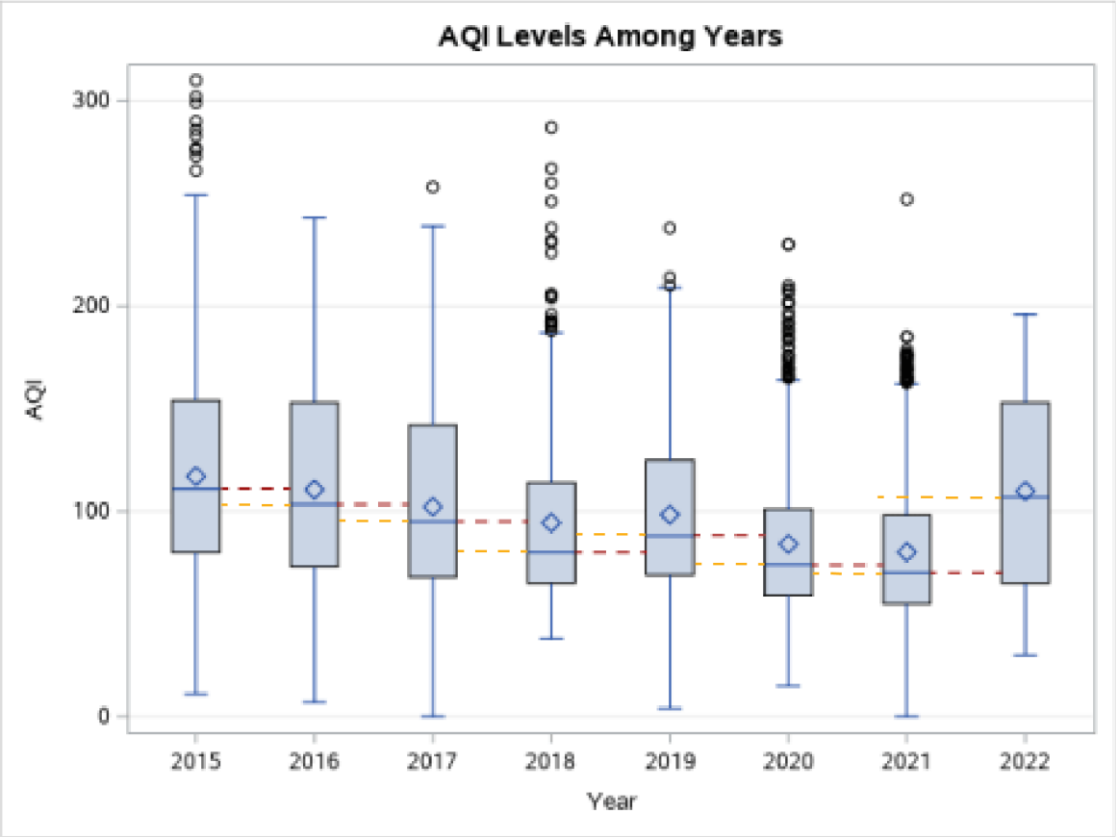
Anything that falls outside of the whiskers, which are equal to the size of 1.5 times of IQR, is considered as **outliers**. We will be talking about outliers in step 3.

Going back to our data, we can see the min and max values in the image below. The widest range belongs to 2015 with the ‘11-254’ values. In 2016, the range gets smaller which means that the data gets less scattered. As you we mentioned earlier, the range of the box gets wider in 2016, compared to 2015, but the whiskers range gets smaller. The smallest range belongs to the years of both 2018 and 2020.



Comparison of the Medians

We wanted to see if the line of medians aligns **inside** or **outside** of the next & previous year's boxes. And as we can see in the image below, all of the medians align with the inside of the boxes of the next & previous years. (Except the median of 2022 but we don't have enough data to talk about 2022. If we had, we could say that the data of 2022 is likely to be different than the 2021 data because the median line aligns outside of the box of 2021)



Identifying Outliers

We already mentioned that any value falling outside of the whiskers are considered as **outliers**. In our data set, we have a total of **110 outliers.** We can see the most outlier values in 2020 with 41 values. And the years 2016 & 2022 are the only years with 0 outlier values. Here are the outlier values of all years.

11	310	OUTLIER	2015	169	2015
12	302	OUTLIER	2015	168	2015
13	299	OUTLIER	2015	168	2015
14	290	OUTLIER	2015	221	2015
15	286	OUTLIER	2015	177	2015
16	283	OUTLIER	2015	185	2015
17	277	OUTLIER	2015	153	2015
18	276	OUTLIER	2015	150	2015
19	273	OUTLIER	2015	166	2015
20	266	OUTLIER	2015	197	2015

2015: 10 Outlier Values Detected
266 to 310

41	258	OUTLIER	2017	266	2015

2017: 1 Outlier Value Detected
258

52	287	FAROUTLIER	2018	74	2015
53	267	FAROUTLIER	2018	53	2015
54	260	OUTLIER	2018	81	2015
55	251	OUTLIER	2018	92	2015
56	238	OUTLIER	2018	104	2015
57	232	OUTLIER	2018	154	2015
58	231	OUTLIER	2018	180	2015
59	226	OUTLIER	2018	155	2015
60	206	OUTLIER	2018	176	2015
61	205	OUTLIER	2018	170	2015
62	204	OUTLIER	2018	171	2015
63	196	OUTLIER	2018	187	2015
64	193	OUTLIER	2018	224	2015
65	193	OUTLIER	2018	190	2015
66	192	OUTLIER	2018	154	2015
67	191	OUTLIER	2018	103	2015
68	190	OUTLIER	2018	79	2015
69	188	OUTLIER	2018	94	2015

2018: 18 Outlier Values Detected
188 to 287
2 Far Outlier Detected

80	238	OUTLIER	2019
81	214	OUTLIER	2019
82	210	OUTLIER	2019

2019: 3 Outlier Values Detected
210 to 238

93	230	FAROUTLIER	2020
94	230	FAROUTLIER	2020
95	210	OUTLIER	2020
96	208	OUTLIER	2020
97	208	OUTLIER	2020
98	206	OUTLIER	2020
99	202	OUTLIER	2020
100	202	OUTLIER	2020
101	201	OUTLIER	2020
102	197	OUTLIER	2020
103	195	OUTLIER	2020
104	192	OUTLIER	2020
105	192	OUTLIER	2020
106	191	OUTLIER	2020
107	189	OUTLIER	2020
108	188	OUTLIER	2020
109	187	OUTLIER	2020
110	185	OUTLIER	2020
111	184	OUTLIER	2020
112	183	OUTLIER	2020
113	180	OUTLIER	2020
114	180	OUTLIER	2020
115	176	OUTLIER	2020
116	176	OUTLIER	2020
117	175	OUTLIER	2020
118	175	OUTLIER	2020
119	173	OUTLIER	2020
120	173	OUTLIER	2020
121	170	OUTLIER	2020
122	170	OUTLIER	2020
123	170	OUTLIER	2020
124	169	OUTLIER	2020
125	169	OUTLIER	2020
126	168	OUTLIER	2020
127	167	OUTLIER	2020
128	167	OUTLIER	2020
129	166	OUTLIER	2020
130	166	OUTLIER	2020
131	165	OUTLIER	2020
132	165	OUTLIER	2020
133	165	OUTLIER	2020

2020: 41 Outlier Values Detected
165 to 230
2 Far Outliers Detected

144	252	FAROUTLIER	2021
145	185	OUTLIER	2021
146	185	OUTLIER	2021
147	179	OUTLIER	2021
148	177	OUTLIER	2021
149	177	OUTLIER	2021
150	176	OUTLIER	2021
151	176	OUTLIER	2021
152	176	OUTLIER	2021
153	176	OUTLIER	2021
154	175	OUTLIER	2021
155	174	OUTLIER	2021
156	174	OUTLIER	2021
157	173	OUTLIER	2021
158	172	OUTLIER	2021
159	171	OUTLIER	2021
160	171	OUTLIER	2021
161	170	OUTLIER	2021
162	170	OUTLIER	2021
163	169	OUTLIER	2021
164	169	OUTLIER	2021
165	169	OUTLIER	2021
166	169	OUTLIER	2021
167	168	OUTLIER	2021
168	167	OUTLIER	2021
169	166	OUTLIER	2021
170	166	OUTLIER	2021
171	165	OUTLIER	2021
172	165	OUTLIER	2021
173	165	OUTLIER	2021
174	165	OUTLIER	2021
175	164	OUTLIER	2021
176	164	OUTLIER	2021
177	163	OUTLIER	2021
178	163	OUTLIER	2021
179	163	OUTLIER	2021
180	163	OUTLIER	2021

2021: 37 Outlier Values Detected
163 to 252
1 Far Outlier Detected

We detected 5 **'Far Outlier'** values in the dataset. 2 far outliers in 2018, 2 far outliers in 2020 and 1 far outlier in 2021. Far outliers are the very extreme values that fall outside more than 3.0 times the IQR below the Q1 or above the Q3.

And if they fall between 1.5 times and 3.0 times the IQR below the Q1 or above the Q3, they are called **mild outliers**.

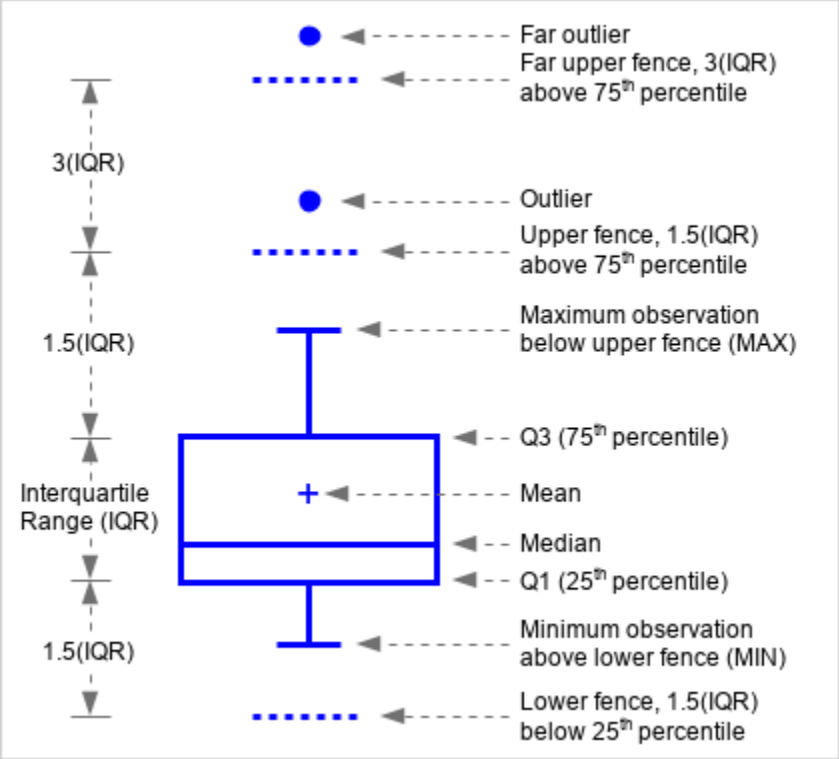
Therefore, we can say that there are 2 types of outliers: Far Outliers and Mild Outliers. The mathematical representation for these 2 types is demonstrated below.

Mild Outliers

Q1 - 3 * IQR <= X < Q1 - 1.5 * IQR
Q1 + 1.5 * IQR < X <= Q3 + 3 * IQR

Far Outliers

X < Q1 - 3 * IQR
X > Q3 + 3 * IQR



Source: 'https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/grstatgraph/p0vuh82v39fsasn1vqhzmhdl8y16.htm'

Now, let's go back to our dataset and crosscheck the far outlier values with the formula. For example, the mathematical process for the year 2018 would be:

49 * 3 = 147

65 – 147 = -47 Far Lower Fence

114 + 147 = 261 Far Upper Fence

The far outlier values are 267 and 287 meaning that both are bigger than the Far Upper Fence value of 261.

267 > 261 ✔

287 > 261 ✔

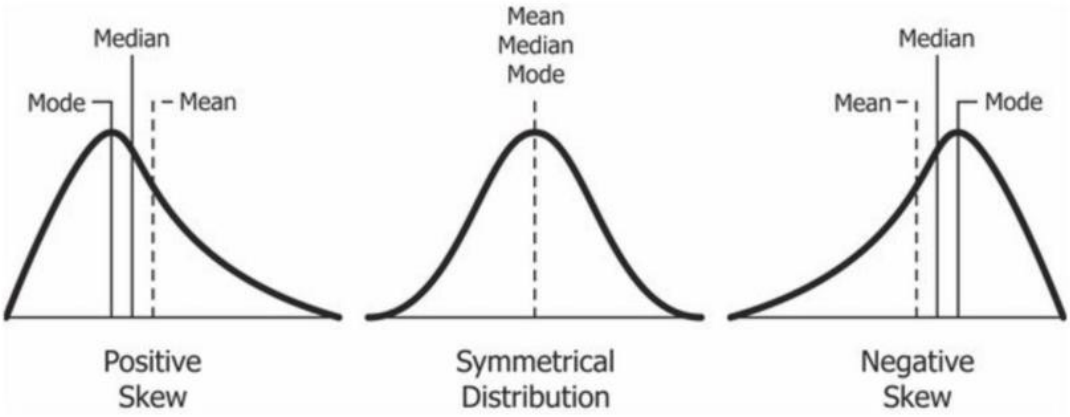
Year	Q1	Q3	IQR	3*IQR	Far Lower Fence	Far Upper Fence	Far Outlier Values
2018	65	114	49	147	-82	261	267, 287 ✔
2020	59	101	42	126	-67	227	230, 230 ✔
2021	55	98	43	129	-74	227	252 ✔

And we can see the values needed for the formula in the table above. As it is indicated, all Far Outlier values are bigger than the Far Upper Fence. Therefore, we can confirm that they are far outliers.

The Skewness

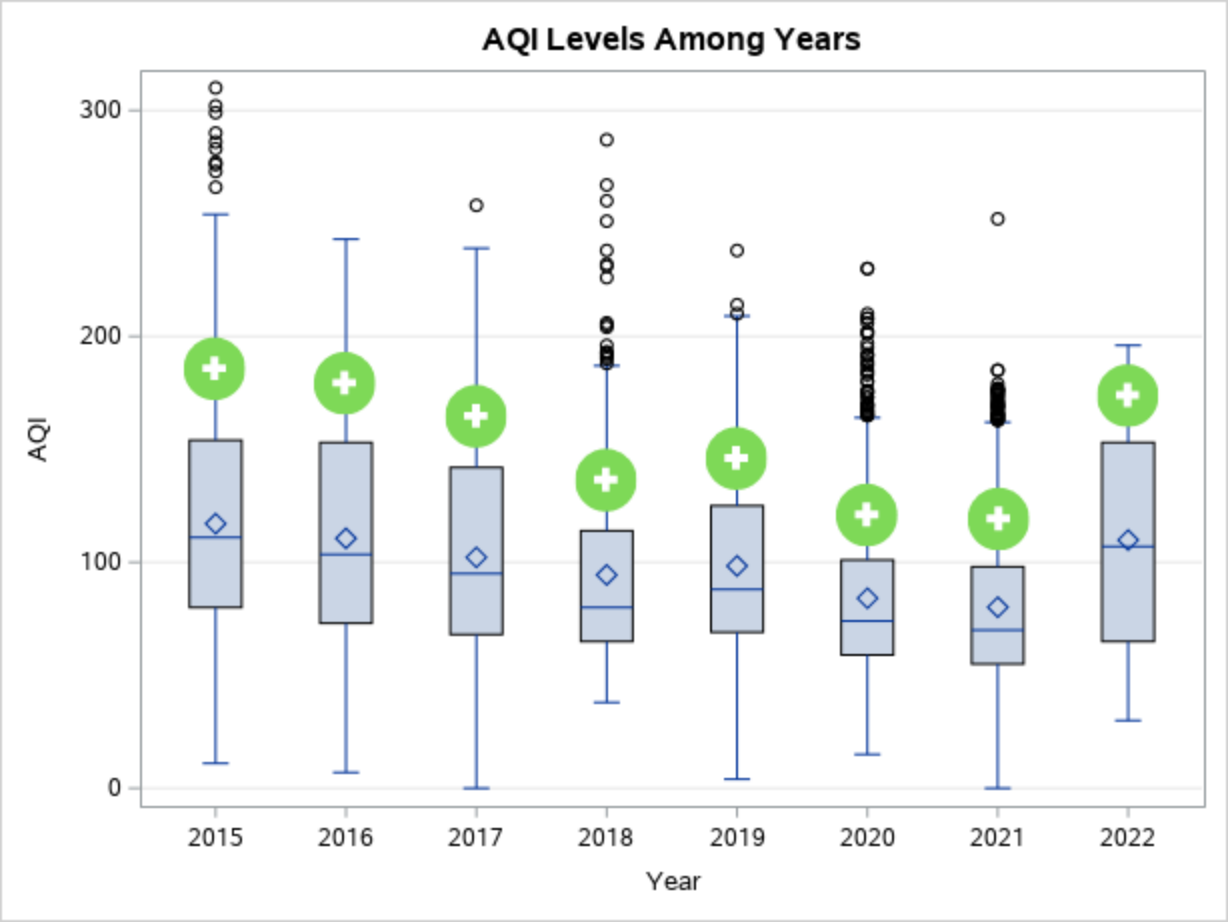
A normal/symmetrical distribution has 0 skewness with all median, mean and mode having the same values. And the Q1 and Q3 values has the same exact distance to the Q2 value. Meaning that "**Q2 – Q1 = Q3 – Q2**" in a normal distribution. To understand the skewness of our box plots, we need to look where the median is located. Is it closer to the Q1 value or the Q3 value? Or is it exactly in the middle? Answering these questions will help us to investigate the skewness of the box plots.

Now, as it is indicated in the image below, if the median line is located closer to Q1 value, then it means our box plot has a **positive skewness**. And if it is located closer to the Q3 value, then it means our box plot has a **negative skewness**.



Source: "https://stats.stackexchange.com/questions/481409/figuring-out-skewness-from-a-boxplot"

We can also see that in a positive skew, the mean is greater than the median; and in a negative skew, the mean is smaller than the median. So, let's look at our box plots to see their skewness.



As we can see in the plot, all our mean values are smaller the median values and lie under them. We can also see that the median values in all the years are closer to the Q1 value. A way to

crosscheck is to use the "**Q2 – Q1 = Q3 – Q2**" formula to find the middle value (The Q2 value if the box plot had a normal distribution) and see if the median is smaller or greater than the middle value.

From the formula, we can calculate the middle value as **(Q1 + Q3) / 2** and if the median is greater than the MV, the median is closer to Q3; meaning that it has a negative skewness. But if it is smaller than the MV, than it is closer to Q1; meaning that it has a positive skewness. And if it is the same value as MV, then it means the box plot has a normal distribution.

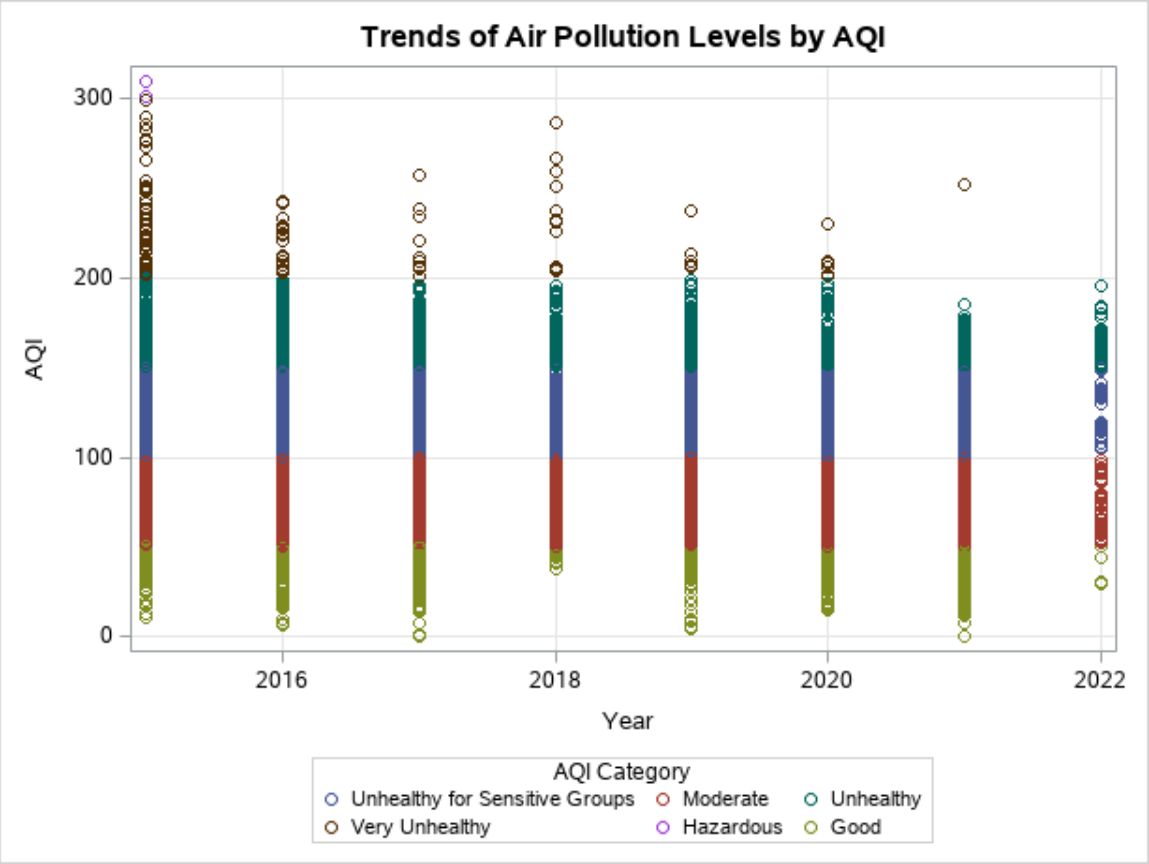
The median and middle values are indicated in the table below for all years. The median values are smaller than the middle values in all years.

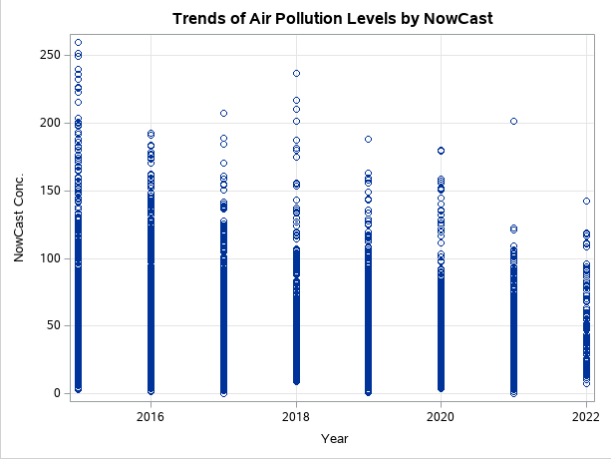
Year	Q1	Q3	Median	Middle Value
2015	80	154	111	117
2016	73	153	103.5	113
2017	68	142	95	105
2018	65	114	80	89.5
2019	69	125	88	97
2020	59	101	74	80
2021	55	98	70	76.5
2022	65	153	107	109

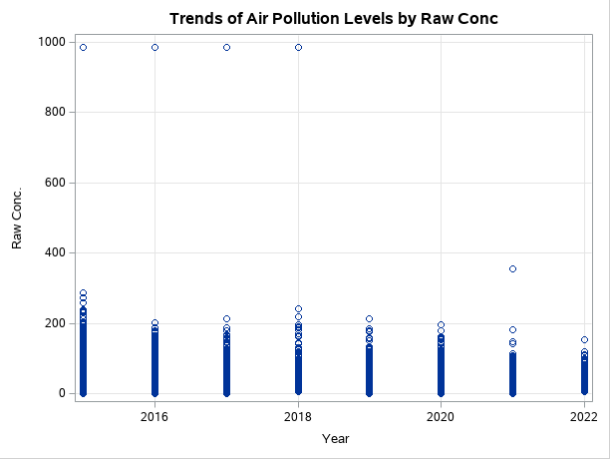
To sum up, the box plots of all years have a **positive skewness**. The median values are closer to Q1, the means are greater than the medians and lastly, the lower whiskers are shorter than the upper whiskers.

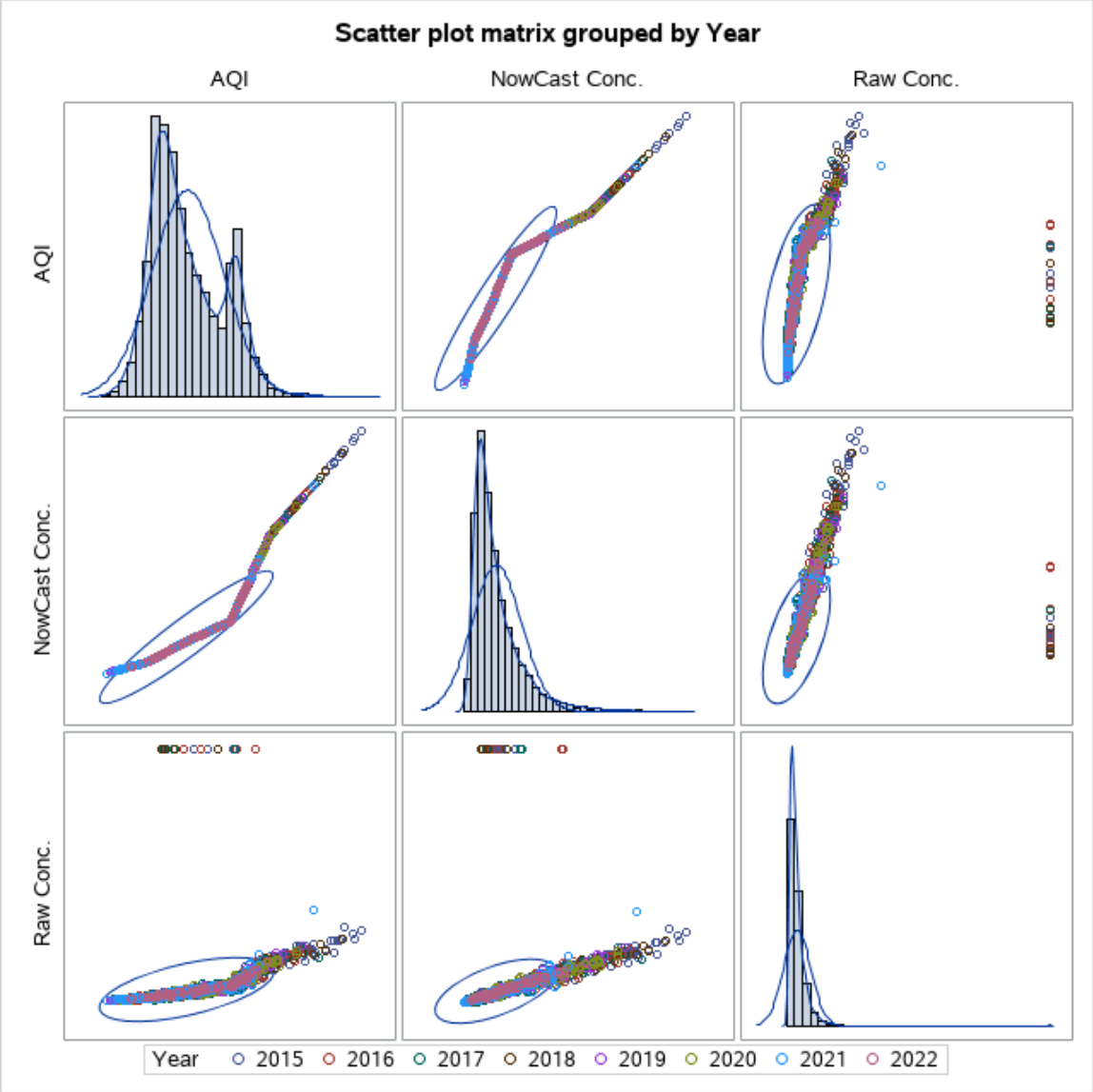
2. The Trends of Air Pollution Levels

We can investigate the air pollution trends among years in the scatter plot below. In 2015 data, we see a big range with obvious air pollutions levels from 'Good' to 'Hazardous'. In 2016, the air pollution data seems to have less values for the 'Hazardous' and 'Very Unhealthy' category. Which we can consider as an improvement. However, in order to comment about improvement of air pollution levels among the years, we can not just consider the range of data. We also need to judge the distribution of the data which we will be doing in the next section with Histograms.









This scatter plot matrix shows us the relationship between AQI, NoxCast Concentration and Raw Concentration values. We can see that there is a correlation between AQI vs. NowCast Concentration, NowCast Concentration vs. Raw Concentration and AQI vs. Raw Concentration. This is because we use Raw Concentration data to calculate AQI which represents the 24-hourly air quality and NowCast Concentration which represents the current air quality with a 12-hourly calculation.

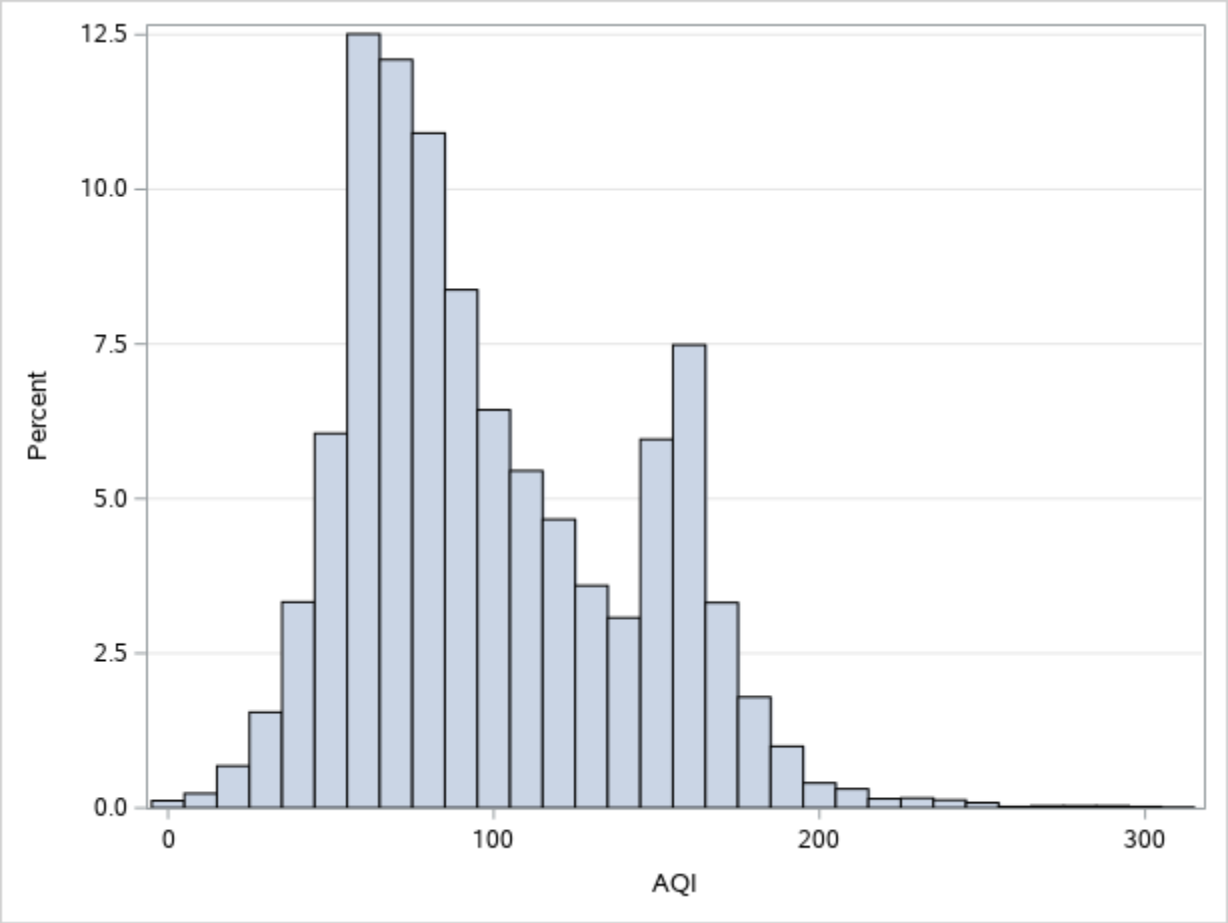
$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low}) + I_{low}$$

The Representation of AQI

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

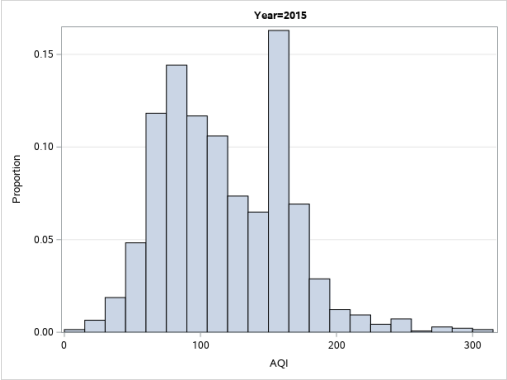
The Representation of NowCast

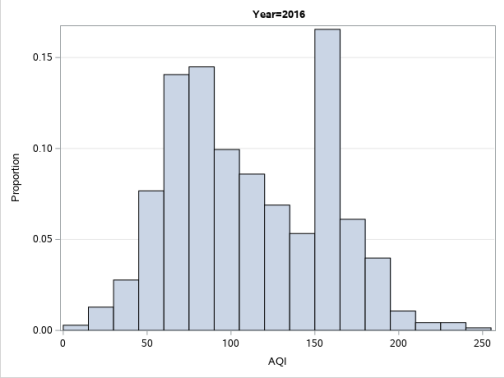
3. The Distribution of Air Pollution Levels

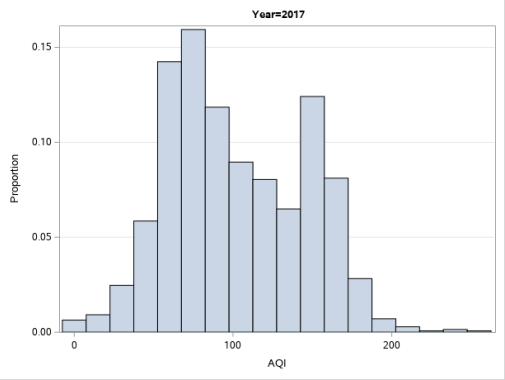


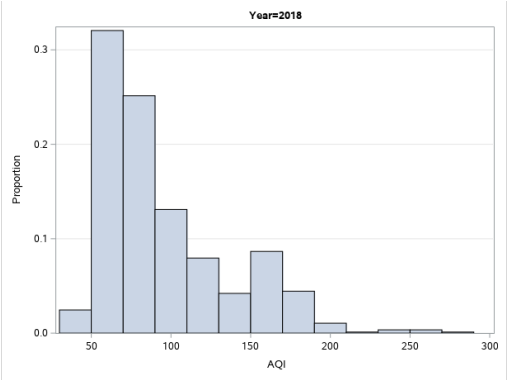
To investigate the distribution of air pollution levels in our dataset, we used histogram of AQI. As you can see, we have a **bimodal** histogram for air pollution levels meaning that the histogram has 2 data peaks: the first peak representing the 60-70 levels and the second peak representing the 160-170 levels.

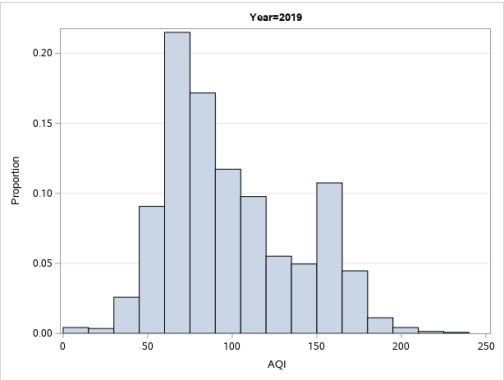
As we can see from the histogram, a big portion of the data is gathered around the 'Moderate' air quality levels. However, this histogram represents the all 7 years of data. Now, let's take a look at the distribution of air quality levels on a yearly basis so we can see the improvement or deterioration.

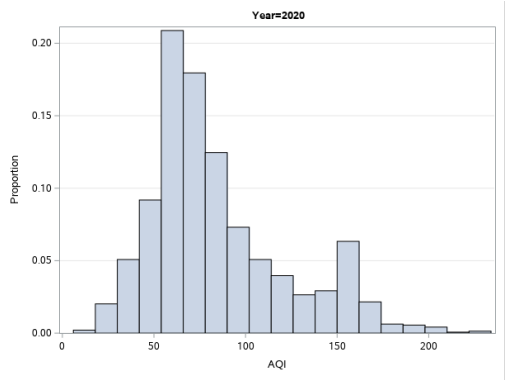


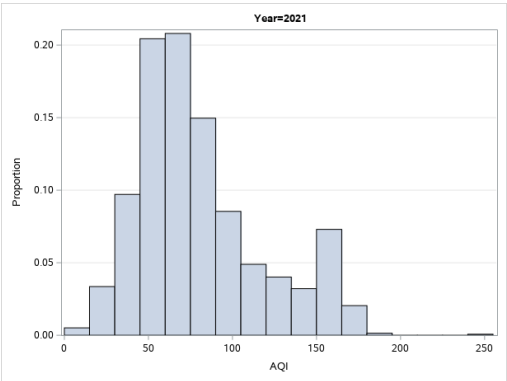


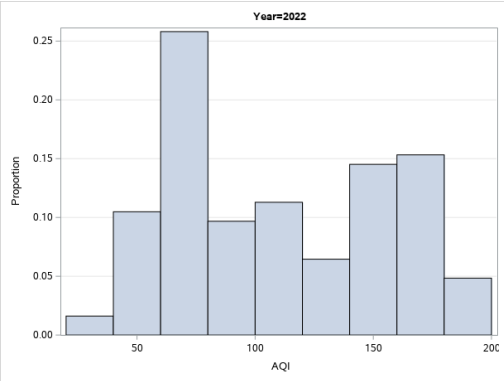




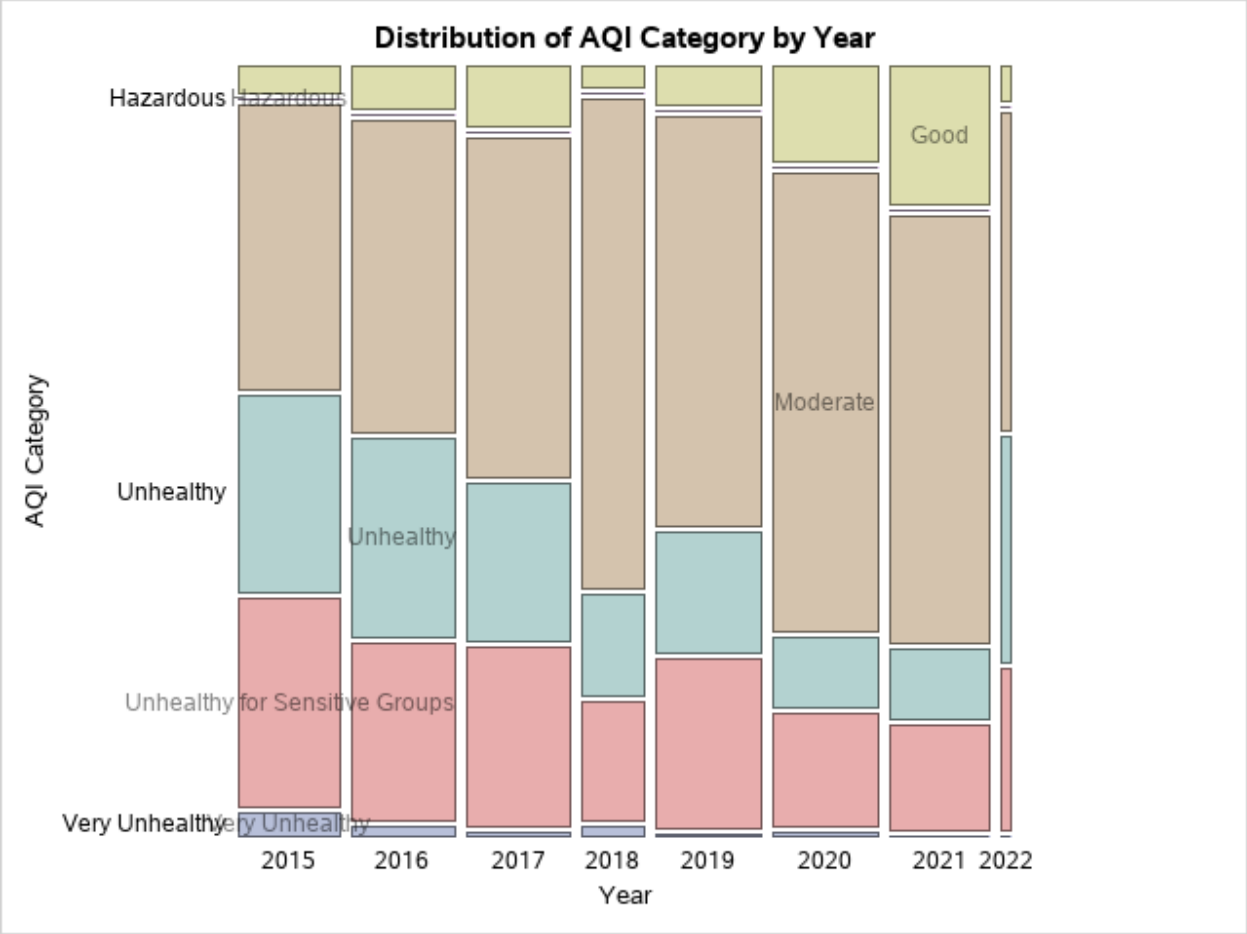








Here are the yearly histogram plots of the AQI data. We will be showing these yearly distributions in the graph below and talk about the distributions more.



We used this graph to see the distributions of air quality data on a yearly basis. It is a different representation of the yearly histograms that we can see in the previous page. And it allows us to compare the years.

The graph shows us a clear improvement behaviour between 2015-2017 years in air quality. In 2016 and 2017, we can see that the air quality levels get better compared to the previous year.

However, the year 2018 has an interesting data. We can see that the 'Very Unhealthy' levels has a slight increase but on the other hand, the categories that can be considered good such as 'Good' and 'Moderate' also has an increase compared to the previous year. And of course, 'Unhealthy' and 'Unhealthy for Sensitive Groups' levels has a big decrease. Therefore, we can still talk about an improvement.

And the following years also shows some level of improvement consecutively (excluding the 2022 due to lack of data).

Overall, we see a clear improvement in Shanghai's air quality over the years considering the histogram, scatter plot and box plot data.

4. Systematic Sampling of the Dataset

Systematic sampling is a type of probability sampling method in which sample members from a larger population are selected according to a random starting point but with a fixed, periodic interval. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size. Despite the sample population being selected in advance, systematic sampling is still thought of as being random if the periodic interval is determined beforehand and the starting point is random. (How Does Systematic Sampling Work?, 2022)

And in our dataset, we selected 4 records per month to apply systematic sampling in our original clean dataset with 9432 observations. The new dataset we created has a total of 332 observations and named as SampleSYS1.

The SURVEYSELECT Procedure

Selection Method	Systematic Random Sampling
Strata Variables	Year
	Month

Input Data Set	TABLESORTED
Random Number Seed	728954629
Stratum Sample Size	4
Number of Strata	83
Total Sample Size	332
Output Data Set	SAMPLESYS1

5. AQI Category Comparison Among All Years

One of the most common things that is used to investigate the relationship between categorical variables is a **Chi Square Test**. Therefore, we tried running a chi square test with AQI Category and Year variables to understand the association between them. And here we can see the results below.

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of AQI Category by Year

AQI Category	Year								
	2015	2016	2017	2018	2019	2020	2021	2022	Total
Good	3 0.90 8.82 6.25	3 0.90 8.82 6.25	4 1.20 11.76 8.33	2 0.60 5.88 5.00	5 1.51 14.71 10.42	7 2.11 20.59 14.58	10 3.01 29.41 20.83	0 0.00 0.00 0.00	34 10.24
Moderate	21 6.33 12.21 43.75	19 5.72 11.05 39.58	22 6.63 12.79 45.83	24 7.23 13.95 60.00	25 7.53 14.53 52.08	30 9.04 17.44 62.50	29 8.73 16.86 60.42	2 0.60 1.16 50.00	172 51.81
Unhealthy	16 4.82 30.77 33.33	12 3.61 23.08 25.00	13 3.92 25.00 27.08	2 0.60 3.85 5.00	5 1.51 9.62 10.42	3 0.90 5.77 6.25	1 0.30 1.92 2.08	0 0.00 0.00 0.00	52 15.66
Unhealthy for Sensitive Groups	8 2.41 11.43 16.67	13 3.92 18.57 27.08	9 2.71 12.86 18.75	10 3.01 14.29 25.00	12 3.61 17.14 25.00	8 2.41 11.43 16.67	8 2.41 11.43 16.67	2 0.60 2.86 50.00	70 21.08
Very Unhealthy	0 0.00 0.00 0.00	1 0.30 25.00 2.08	0 0.00 0.00 0.00	2 0.60 50.00 5.00	1 0.30 25.00 2.08	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 1.20
Total	48 14.46	48 14.46	48 14.46	40 12.05	48 14.46	48 14.46	48 14.46	4 1.20	332 100.00

Statistics for Table of AQI Category by Year

Statistic	DF	Value	Prob
Chi-Square	28	55.1132	0.0016
Likelihood Ratio Chi-Square	28	56.8895	0.0010
Mantel-Haenszel Chi-Square	1	9.0683	0.0026
Phi Coefficient		0.4074	
Contingency Coefficient		0.3773	
Cramer's V		0.2037	

WARNING: 48% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 332

In the first table, we can see the following values: Frequency, Percent, Row Pct and Col Pc.

As it can be seen, we highlighted the numbers to differentiate the values.

The yellow numbers represent the frequency values of the AQI categories on a yearly basis.

The pink numbers represent the overall percentage of the relative AQI category observations in each specific year.

The green numbers represent the row percentage of the number of observations in year over the total number in AQI Category.

The blue numbers represent the column percentage of the number of observations in year over the total number in AQI Category.

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of AQI Category by Year

AQI Category	Year								
	2015	2016	2017	2018	2019	2020	2021	2022	Total
Good	3	3	4	2	5	7	10	0	34
	0.90	0.90	1.20	0.60	1.51	2.11	3.01	0.00	10.24
	8.82	8.82	11.76	5.88	14.71	20.59	29.41	0.00	
	6.25	6.25	8.33	5.00	10.42	14.58	20.83	0.00	
Moderate	21	19	22	24	25	30	29	2	172
	6.33	5.72	6.63	7.23	7.53	9.04	8.73	0.60	51.81
	12.21	11.05	12.79	13.95	14.53	17.44	16.86	1.16	
	43.75	39.58	45.83	60.00	52.08	62.50	60.42	50.00	
Unhealthy	16	12	13	2	5	3	1	0	52
	4.82	3.61	3.92	0.60	1.51	0.90	0.30	0.00	15.66
	30.77	23.08	25.00	3.85	9.62	5.77	1.92	0.00	
	33.33	25.00	27.08	5.00	10.42	6.25	2.08	0.00	
Unhealthy for Sensitive Groups	8	13	9	10	12	8	8	2	70
	2.41	3.92	2.71	3.01	3.61	2.41	2.41	0.60	21.08
	11.43	18.57	12.86	14.29	17.14	11.43	11.43	2.86	
	16.67	27.08	18.75	25.00	25.00	16.67	16.67	50.00	
Very Unhealthy	0	1	0	2	1	0	0	0	4
	0.00	0.30	0.00	0.60	0.30	0.00	0.00	0.00	1.20
	0.00	25.00	0.00	50.00	25.00	0.00	0.00	0.00	
	0.00	2.08	0.00	5.00	2.08	0.00	0.00	0.00	
Total	48	48	48	40	48	48	48	4	332
	14.46	14.46	14.46	12.05	14.46	14.46	14.46	1.20	100.00

Statistics for Table of AQI Category by Year

Statistic	DF	Value	Prob
Chi-Square	28	55.1132	0.0016
Likelihood Ratio Chi-Square	28	56.8895	0.0010
Mantel-Haenszel Chi-Square	1	9.0683	0.0026
Phi Coefficient		0.4074	
Contingency Coefficient		0.3773	
Cramer's V		0.2037	

WARNING: 48% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 332

And in the table 2, we can see the **Chi-square test statistic** and **P-value** highlighted with purple and red colors.

Our Chi-square test value is **55.1132** and P-Value is **0.0016**. We can clearly see that the P-Value is less than the **significance level** of 0.05. Therefore, we reject the null hypothesis and conclude that there is a statistically significant association between the variables AQI Category and Year.

Statistics for Table of AQI Category by Year

Statistic	DF	Value	Prob
Chi-Square	28	55.1132	0.0016
Likelihood Ratio Chi-Square	28	56.8895	0.0010
Mantel-Haenszel Chi-Square	1	9.0683	0.0026
Phi Coefficient		0.4074	
Contingency Coefficient		0.3773	
Cramer's V		0.2037	
WARNING: 48% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Sample Size = 332

All these information aside, we can see that SAS gave us a WARNING. "Because 48% of the cells have expected counts less than 5. Chi-Square may not be a valid test."

We need to consider this while making a judgement about our Chi-Square Test.

6. AQI Correlation Among All Years

We wanted to see the strength and direction of the linear relationship between AQI and Year variables. In order to do that, we used Correlation Analysis and Correlation Matrix. As it can be seen in the results below; The Pearson Correlation Coefficient of these 2 variables are **-0.26739**.

Correlation Coefficients shows a value between **+1** and **-1**. When there is a **positive perfect correlation** between 2 variables, the correlation coefficient shows +1 as value. When there is a **negative perfect correlation** between 2 variables, the correlation coefficient shows -1 as value. And when there is no correlation between 2 variables, , the correlation coefficient would be 0.

Our correlation coefficient is between 0 and -1. It is closer to 0 than it is closer to -1. Therefore, we can say that the correlation between AQI and Year is weak, it doesn't have a strong correlation. But it has a negative correlation, which can be interpreted as there is an improvement in air pollution levels among the years.

The CORR Procedure

2 Variables:	AQI Year

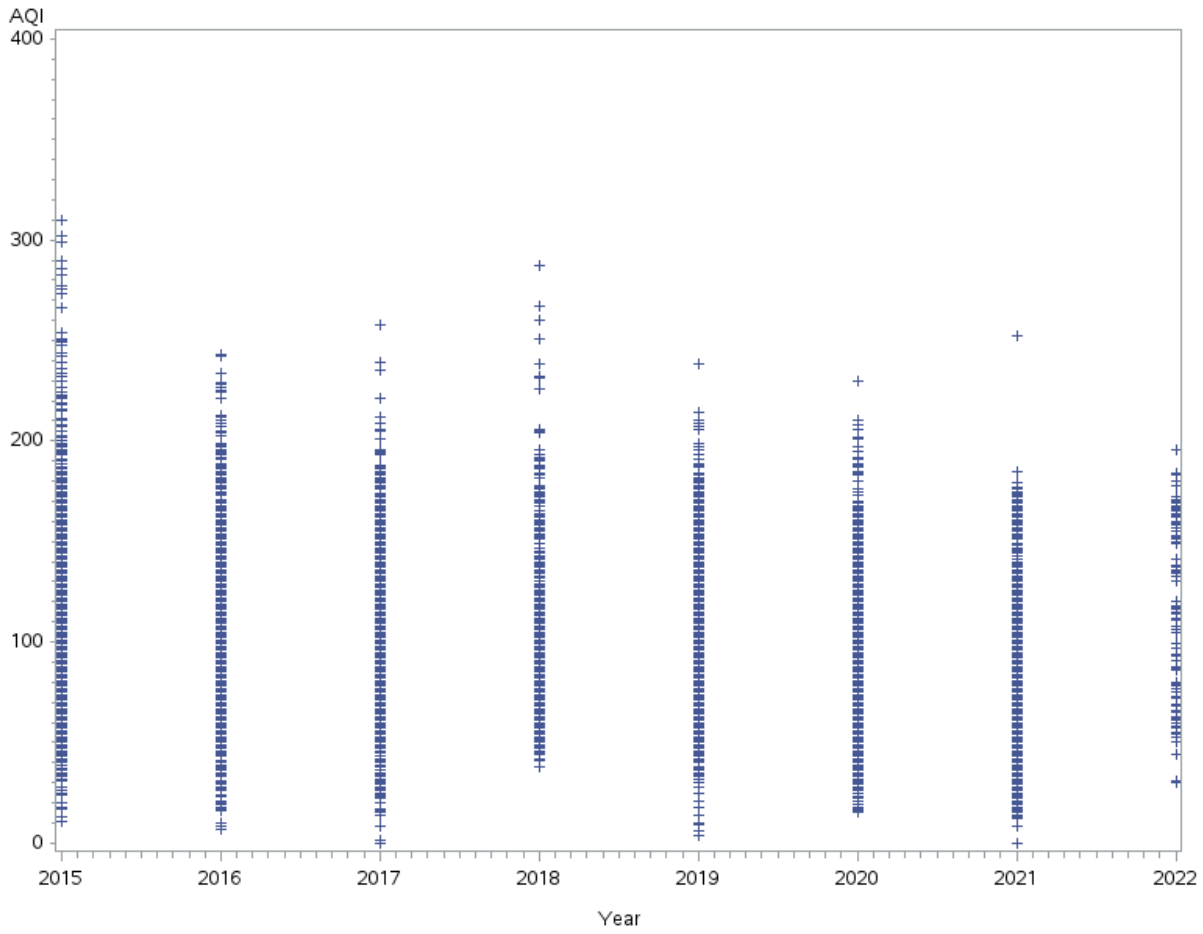
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
AQI	9432	98.49958	43.35218	929048	0	310.00000
Year	9432	2018	2.08502	19034294	2015	2022

Pearson Correlation Coefficients, N = 9432
Prob > |r| under H0: Rho=0

	AQI	Year
AQI	1.00000	-0.26739 <.0001
Year	-0.26739 <.0001	1.00000

The visual representation of Correlation Matrix of our Correlation Analysis:



References

Iqair.com. 2022. Shanghai Air Quality Index (AQI) and China Air Pollution | AirVisual. [online] Available at: <https://www.iqair.com/ca/china/shanghai> [Accessed 12 February 2022].

Investopedia. 2022. *How Does Systematic Sampling Work?*. [online] Available at:<https://www.investopedia.com/terms/s/systematic-sampling.asp#:~:text=Systematic%20sampling%20is%20a%20type,by%20the%20desired%20sample%20size.> [Accessed 11 February 2022].