Team Project Report

August 2021 (Term I)

Submitted by:

TEAM 6

Hemanshi Marwaha (100361239)

Jill Lau (100371189)

Christina Mathew (100374202)

Juan Figini (100373584)

TABLE OF CONTENTS

ACKNOWLEDGEMENT

Foremost, we would like to extend our sincere gratitude to our mentor Monica Nguyen for giving us this opportunity to complete this project and her able guidance, support, and encouragement throughout the course which shaped the present work as is shown in this report. We would also like to express our sincere thanks to the U.S. Embassies for establishing the air quality monitoring systems which enabled us to receive the data required to perform the analysis.

This project cannot be completed without the efforts and co-operation from our group members — *Hemanshi Marwaha, Christina Mathew, Jill Lau, and Juan Figini*. This project would not have been possible without the stimulating discussions and insightful comments.

INTRODUCTION

According to the Air Quality Life Index (AQLI), the air quality in Indonesia fails to meet the World Health Organization (WHO) guideline for concentrations of fine particulate matter (PM2.5). At the current pollution levels, an average Indonesian is expected to lose 1.2 years of life expectancy. ln parts of the country with particularly high particulate pollution like Indonesia's capital city Jakarta, the health effects are much larger. If the 2016 pollution levels are sustained over the lifetime, losing 2.3 years of life expectancy can be expected. The AQLI data show that air quality was not a pressing problem in Indonesia just two decades ago, but that air quality declined substantially in recent decades—with the steepest decline since 2013. From 1998 to 2016, the particulate air pollution concentrations have increased by 171 percent and the country went from being one of the cleaner countries in the world to one of the twenty most polluted. Since 2010, electricity generation from coal-fired power plants and gasoline and diesel consumption which are both contributors to PM2.5 air pollution has seen a sharp increase. However, the greatest spike has happened over just the last few years. Pollution more than doubled from 2013 to 2016 alone, with at least some of the increase due to intense fires.[1]

In 2017, the government required that all gasoline-fueled vehicles adopt Euro-4 fuel standards by September 2018 which demands the use of high-quality, cleaner fuels with a sulfur content no higher than 50 parts per million (ppm). This standard is ten times more stringent than the sulfur limit in the Euro-2 fuel that Indonesia previously used.

To combat air pollution from peat and forest fires and after the health and economic damages caused by the 2015 Southeast Asian Haze, the government stepped up its efforts and the Indonesian President enacted a moratorium on new peatland development and established the Peatland Restoration Agency (BRG). The country has experienced fewer fires since then and one of the possible reasons for that could be the BRG's efforts to rewet degraded peatlands. In 2018, the land area that experienced fires was 7 percent of the size of 2015's fires. However, it remains unclear whether the decrease in fires is due to the

[1] Michael Greenstone and Qing (Claire) Fan. 2019. *"Indonesia's Worsening Air Quality and its Impact on Life Expectancy*." AQLI. https://aqli.epic.uchicago.edu/wp-content/uploads/2019/03/Indonesia-Report.pdf.

government's efforts or simply milder weather conditions than in 2015 as some of the areas that did catch fire in 2018 are part of lands supposedly prioritized for peat restoration or protected from drainage.

Ethiopia is urbanizing rapidly. According to UNDESA, the population of Addis Ababa is estimated to have grown by 102% between 2000 and 2020. Urbanization is occurring at the expense of outdoor air quality. Controlling air pollution has always been a challenge in low and middle-income countries and Ethiopia is no different. Air pollution in Ethiopia is caused mainly by vehicles, followed closely by industry, then by domestic emissions.

Air pollution in these countries is characterized by the burning of biomass in household stoves and the use of adulterated liquid fuels. The rapid increase in vehicle numbers in major cities, driven by increased urban population, economic development and urbanization, old vehicles, poor maintenance, and inefficient public transport are factors driving up emissions in the transport sector.

Most of Ethiopia's industry is classed as being agro-based as it contributes to well over 50 percent of the country's GDP. Due to its location, it experiences favorable weather conditions which are ideal to support its agricultural businesses. Even though there are lots of factories on the industrial estates of Ethiopia, many of them use clean energy as their fuel source so do not contribute to air pollution in any significant amount.

Addis Ababa has a limited air quality monitoring network and lacks the long-term data required to ascertain variations of air quality temporally and spatially. Monitored PM2.5 data indicates that air quality in Addis is typically at levels considered 'moderate' to 'unhealthy for sensitive groups' according to the United States Environment Protection Agency's Air Quality Index.[2]

[2] asap-eastafrica.com. 2019. "*Air Quality Briefing Note: Addis Ababa (Ethiopia).*" ASAP East Africa. https://assets.publishing.service.gov.uk/media/5eb16f10e90e0723bd470fdf/ASAP_-_East_Africa_-_Air_Quailty_Briefing_Note_-_Addis_Abada.pdf.

The report uses four cleaned data sets consisting of the measurements of PM2.5 concentrations in Indonesia and Ethiopia across four locations — Jakarta Central, Jakarta South, Addis Ababa Central, and Addis Ababa School.

Air pollution is a significant health issue in Jakarta, causing more than 5 million illnesses a year. This stems mostly from the coal-fired power plants, factories, and vehicles emission. Ethiopia is urbanizing rapidly and it is estimated that the population of Addis Ababa has grown by 102% between 2000 and 2020. Due to a limited air quality monitoring network, Addis Ababa lacks the long-term data required to ascertain variations of air quality temporally and spatially.

The U.S. Embassy Jakarta has two air quality meters, one in Central and one in South Jakarta. The AirNow website indicates that there are two air quality stations —Addis Ababa Central and Addis Ababa School, installed by the American embassy in Ethiopia.

The two master datasets for Indonesia and Ethiopia have **18,233 rows**, 14 columns, and **10,941 rows**, 14 columns respectively.

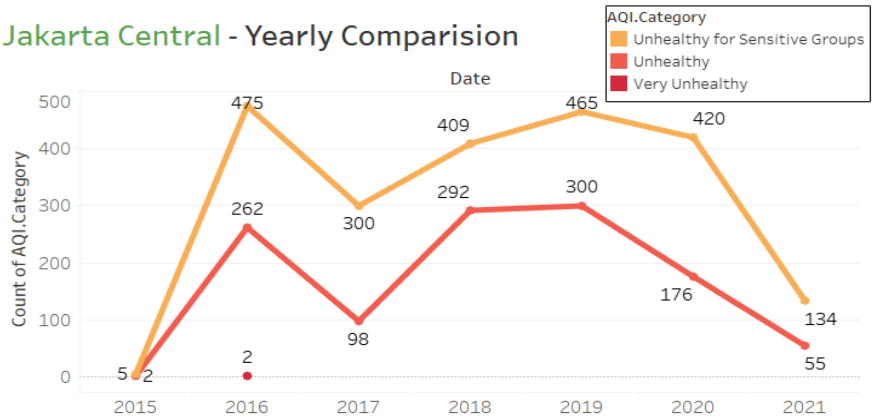
DESCRIPTIVE ANALYSIS

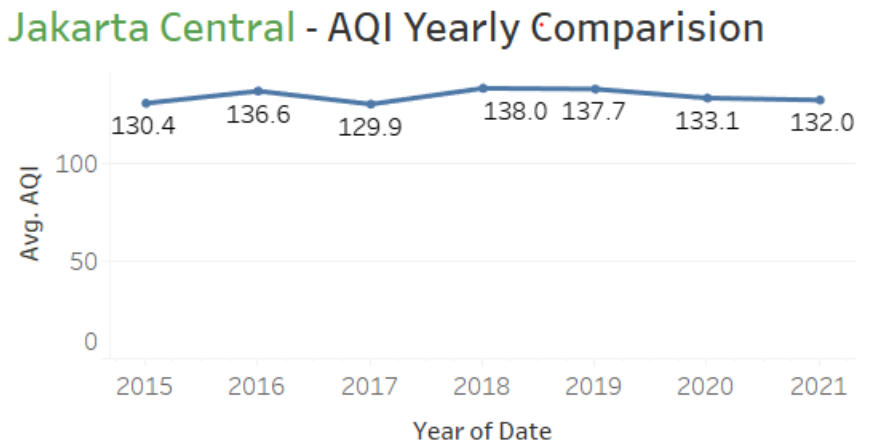
City-wise analysis for AQI categories and AQI

Air quality isn't a new problem for Indonesia. Jakarta's air has been heavily polluted for years. In 2017, the US Embassy's air monitoring stations – located on the rooftops of the US Embassy in Central Jakarta and the Ambassador's residence in South Jakarta – recorded just 26 days where the air quality could be deemed "good". Most of these were during the wet season when heavy monsoonal rain clears the skies and tamps down dust.[3]

Jakarta is the capital city of Indonesia, with some 10.7 million inhabitants as of 2020. It is well known as the economic, political, and cultural hub of the country. There are several contributing factors regarding the high levels of pollution in Jakarta. Jakarta's poor air quality is the result of the perfect storm of pollutants: numerous nearby coal-fired power plants, transport emissions, manufacturing, household emissions, construction, road dust, and open waste burning.

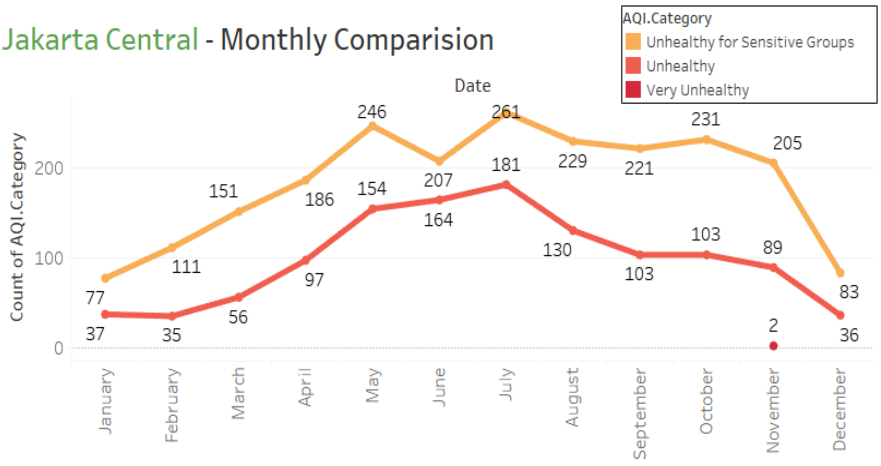
1) Jakarta Central

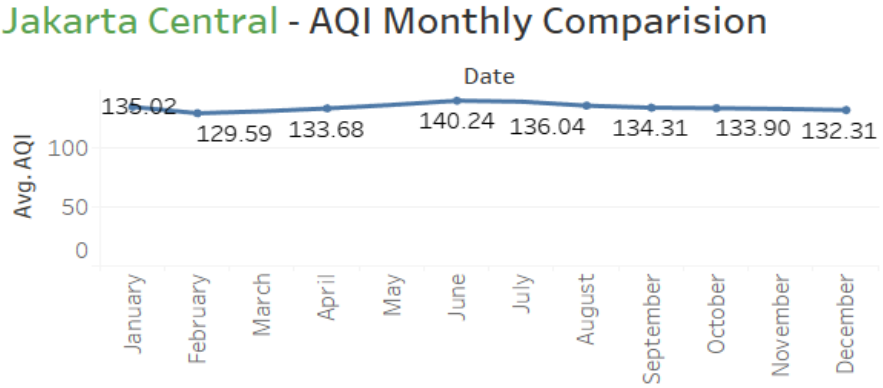




[3] Walton, Kate. 2019. "*Jakarta's air quality kills its residents – and it's getting worse*." theinterpreter. https://www.lowyinstitute.org/the-interpreter/jakarta-s-air-quality-kills-its-residents-and-it-s-getting-worse.

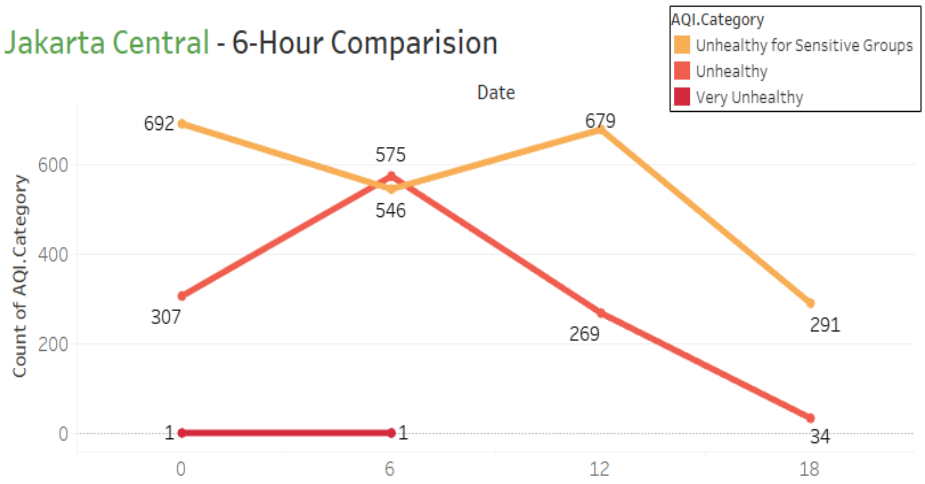
- The levels of air quality have somewhat started to improve after 2019. However, no significant change can be observed.
- The year 2016 followed by 2019 has the highest frequency of readings falling in the three unhealthy categories. Also, 2016 is the only year for which 'very unhealthy' air quality is also observed. However, a look at the average AQI indicates that the years 2018 and 2019 are the worst affected in terms of air quality.
- To improve the air quality, preventative measures such as wearing high-quality particle filtering masks, avoiding outdoor activities, and exercise when pollution levels are particularly high, were taken after the year 2019.
- The average amount of hazardous PM 2.5 particulates in the air rose more than 50% between 2017 and 2018. [4]

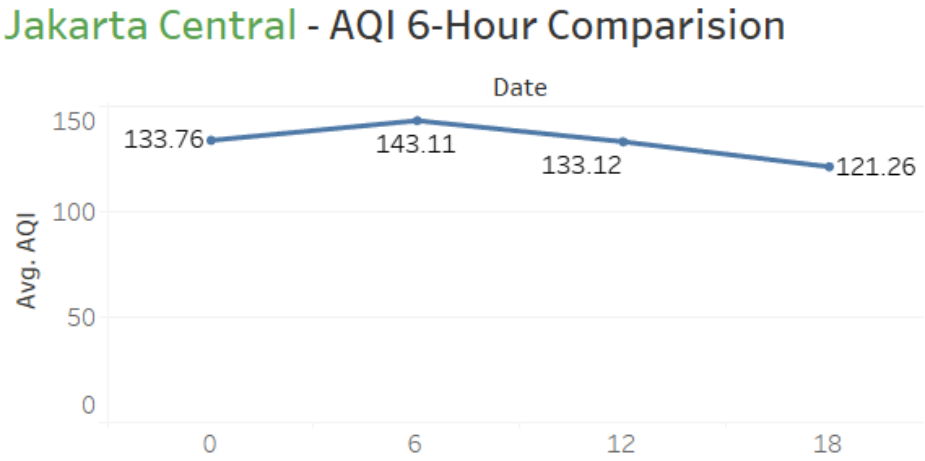




[4] Ben Westcott, and Tia Asmara. 2019. "*Angry citizens sue Indonesian government over growing air pollution*." edition.cnn.com. https://edition.cnn.com/2019/07/02/health/jakarta-pollution-law-suit-intl-hnk/index.html.

- Comparing the month-wise data for the given years, Jakarta Central experienced its worst spells of pollution from June to September, with particularly bad periods of pollution being recorded in June. This could be a result of forest fires during the dry season as the air quality always worsened during the dry season in Jakarta, between May and September.[5]
- February recorded an average AQI value of 129.59 making it comparatively better than the other months.

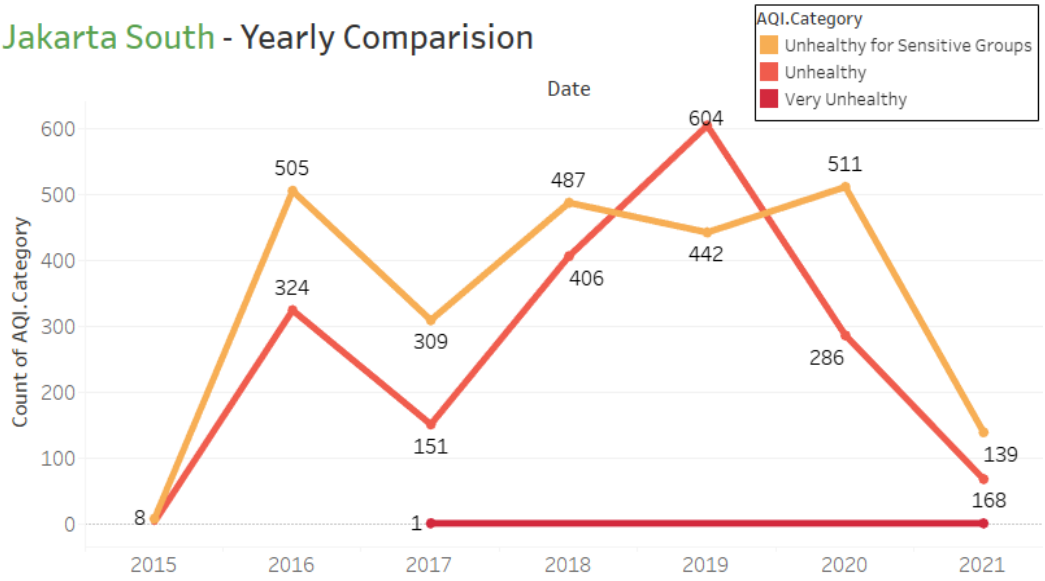


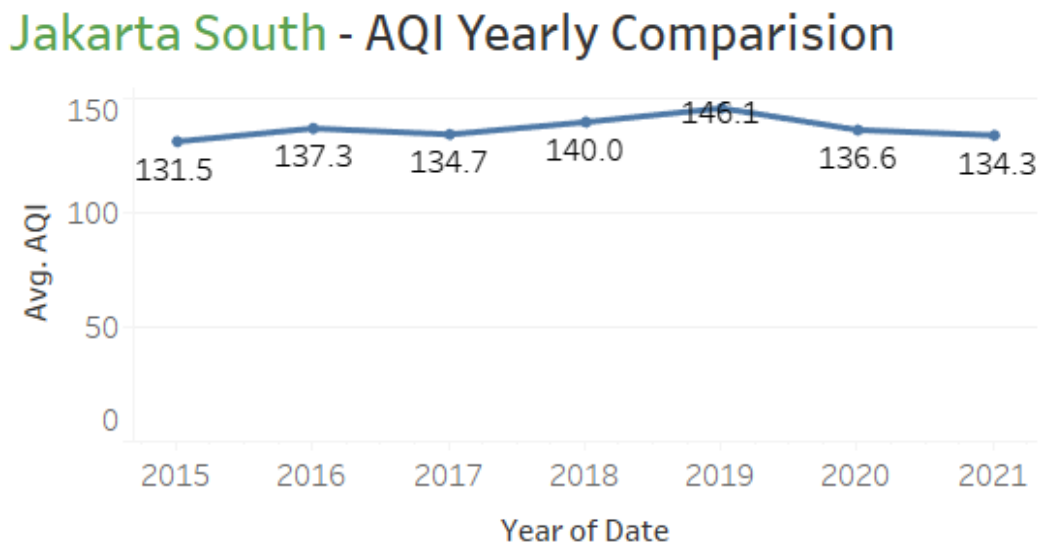


- The 18-hour recorded the least frequency of readings falling under the three unhealthy categories with an average AQI value of 121.26. This could be due to a comparatively lesser number of vehicles on the road and as a result lower emissions towards the end of the day.
- The maximum value of average AQI corresponds to the 6-hour, which also recorded the highest number of values falling in the '*Unhealthy*' category.

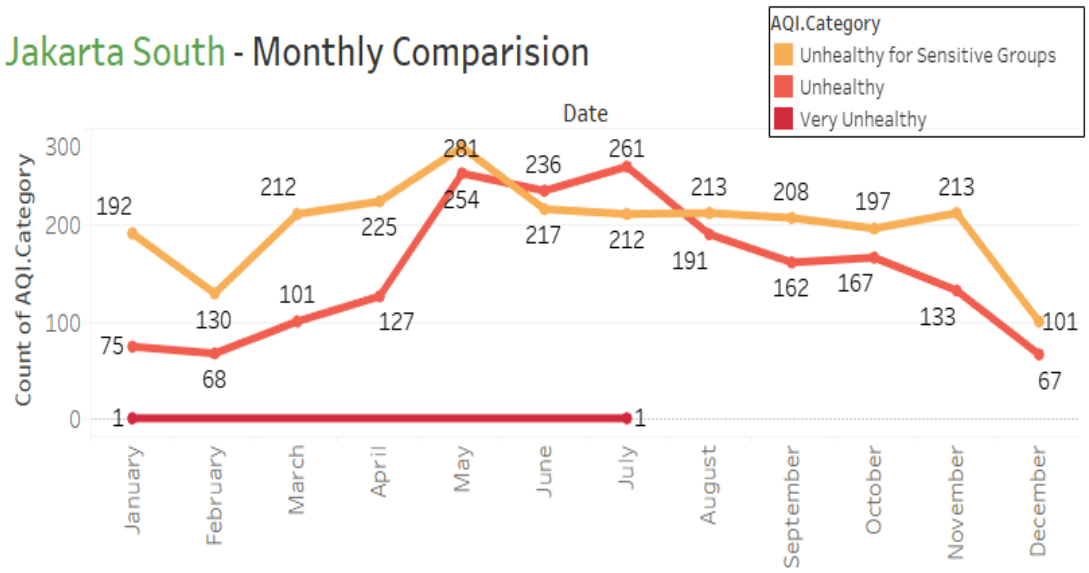
[5] Ben Westcott, and Tia Asmara. 2019. "*Angry citizens sue Indonesian government over growing air pollution*." edition.cnn.com. https://edition.cnn.com/2019/07/02/health/jakarta-pollution-law-suit-intl-hnk/index.html.

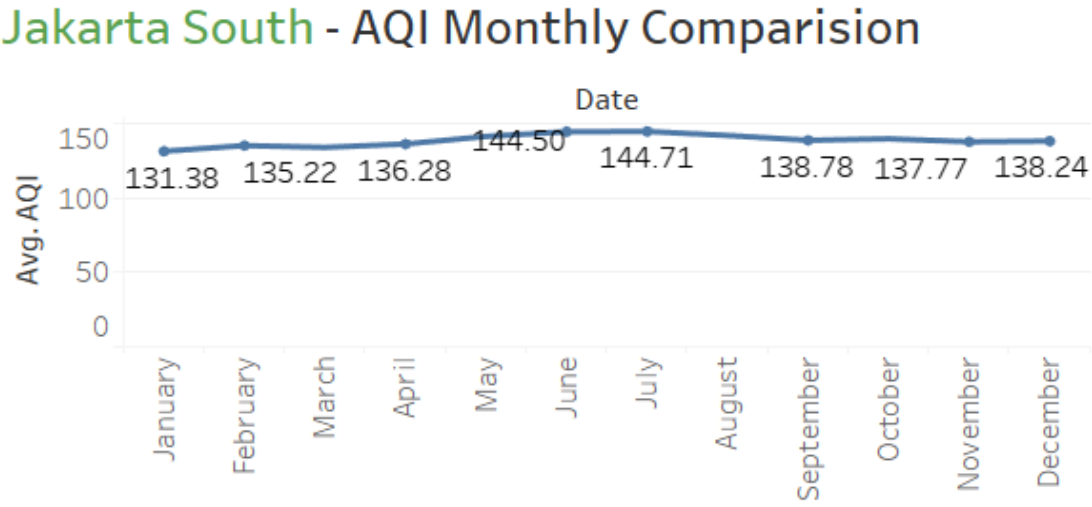
2) Jakarta South



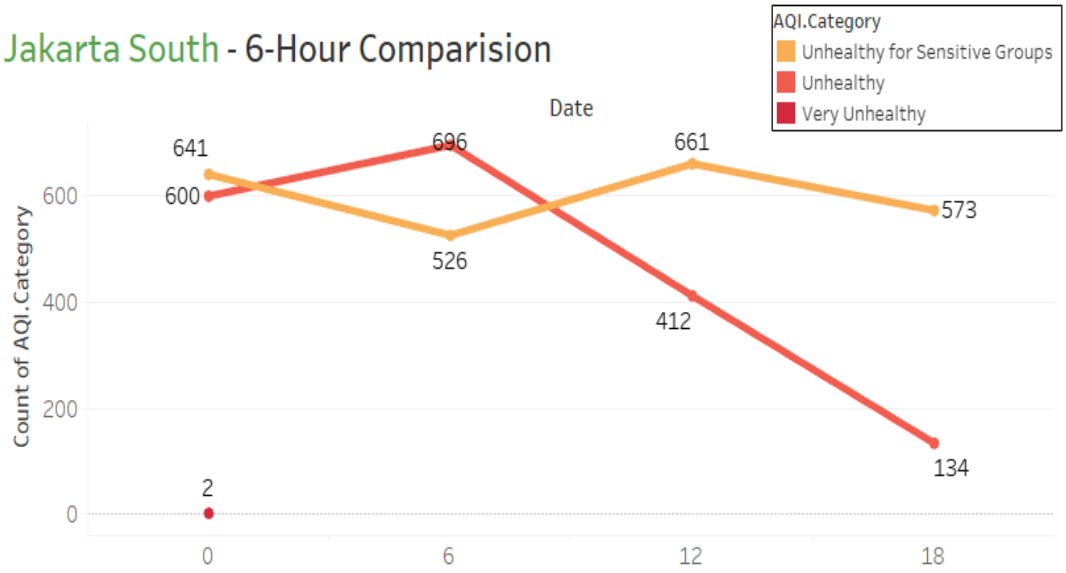


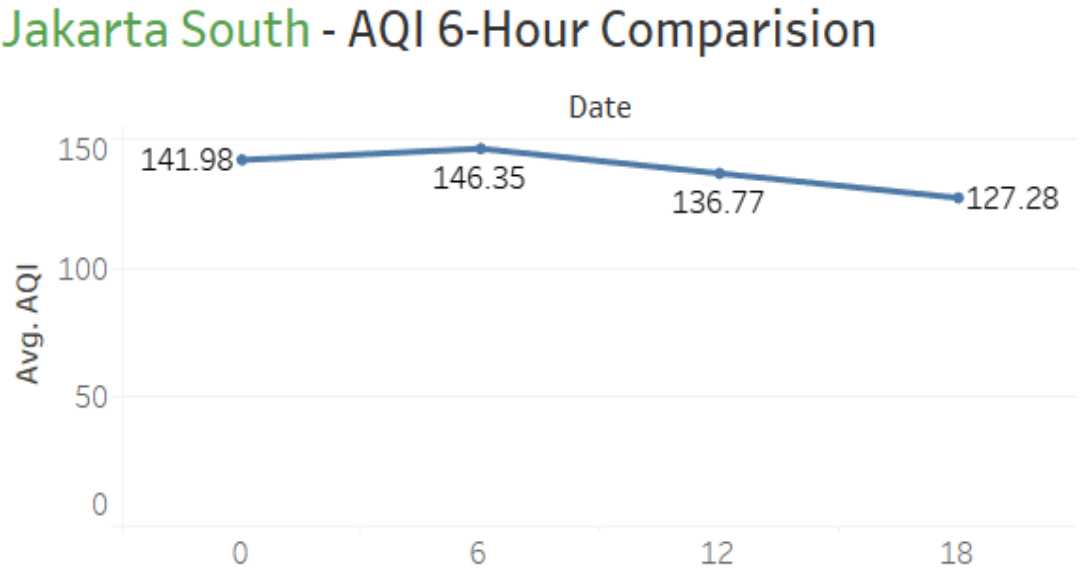
- Taking 2015 as the base year, the air quality for Jakarta South deteriorated significantly from 2015 to 2016.
- The year 2019 recorded the highest frequency of values falling in the '*Unhealthy*' category and is the worst year in terms of air quality with an average AQI value of 146.1
- While the average AQI per month didn't seem to fluctuate much, a downward trend in the average AQI can be observed after the year 2019 implying that the quality of air is improving.





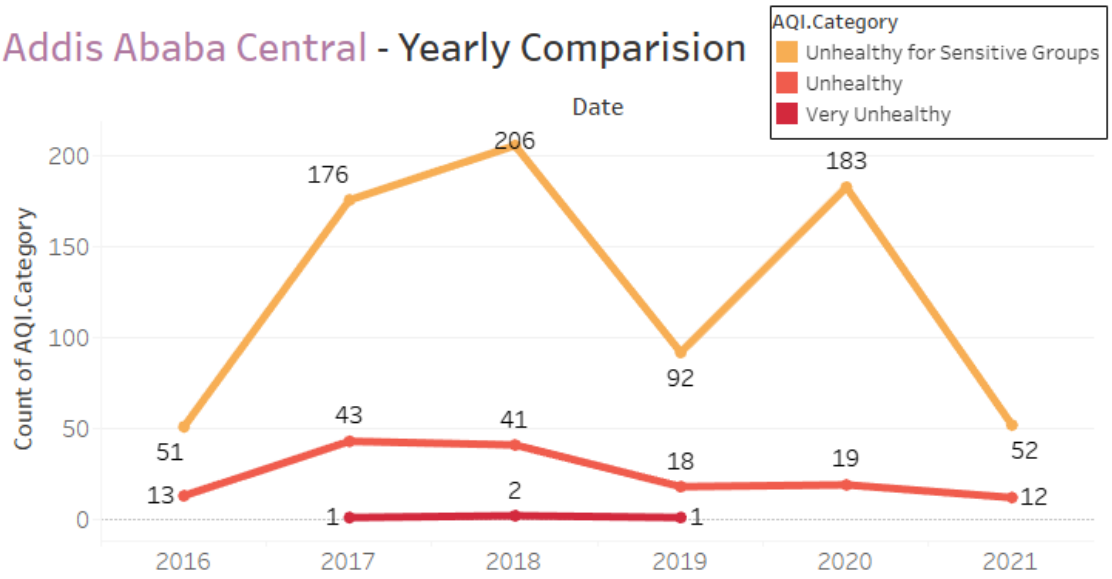
- Comparing the month-wise data for the given years, Jakarta South experienced its worst spells of pollution from May to September, with particularly bad periods of pollution being recorded in May and July. This could be a result of forest fires during the dry season.
- January has the least value for the average AQI throughout the years and is comparatively better than the other months in terms of air quality.

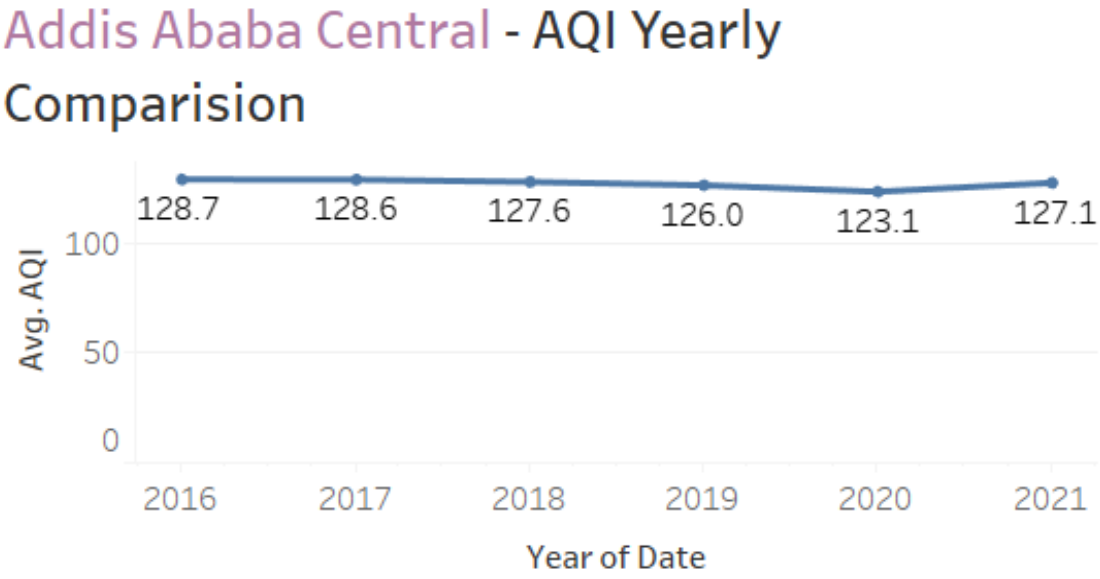




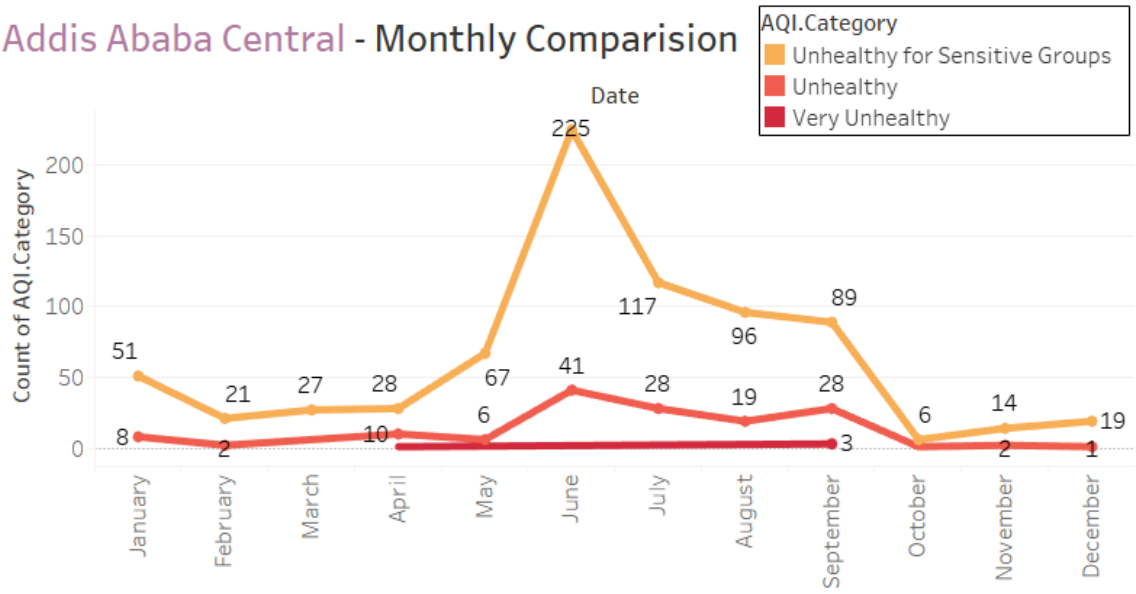
- The quality of air seems to remain almost the same for all the hours in the day except for the 18-hour for which there are only 134 readings falling in the '*Unhealthy*' category.
- It can also be seen that the average AQI seems to be comparatively higher at the 6-hour of the day and continues to decrease after that, dropping the average AQI to 127.28. This could be due to a comparatively lesser number of vehicles on the road and as a result lower emissions towards the end of the day.

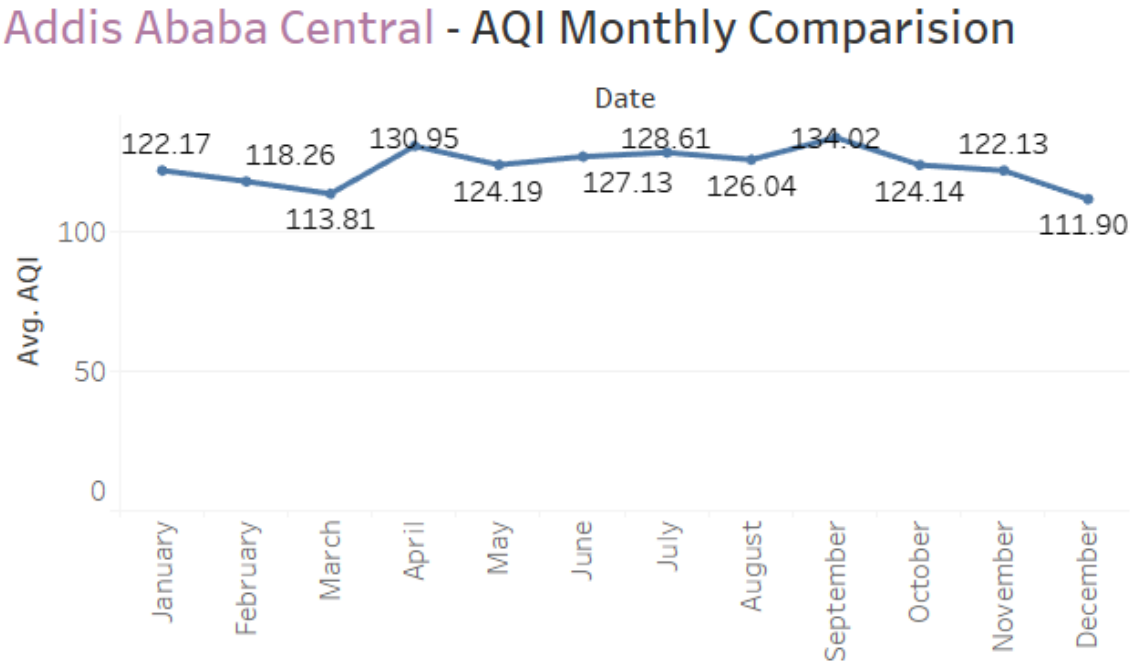
2) **Addis Ababa Central**



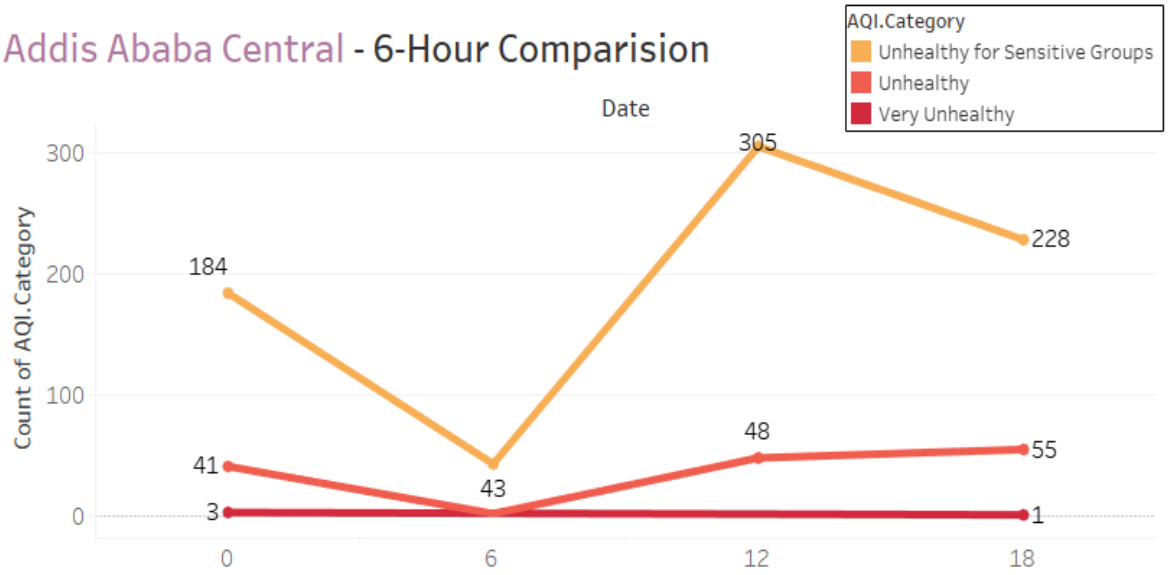


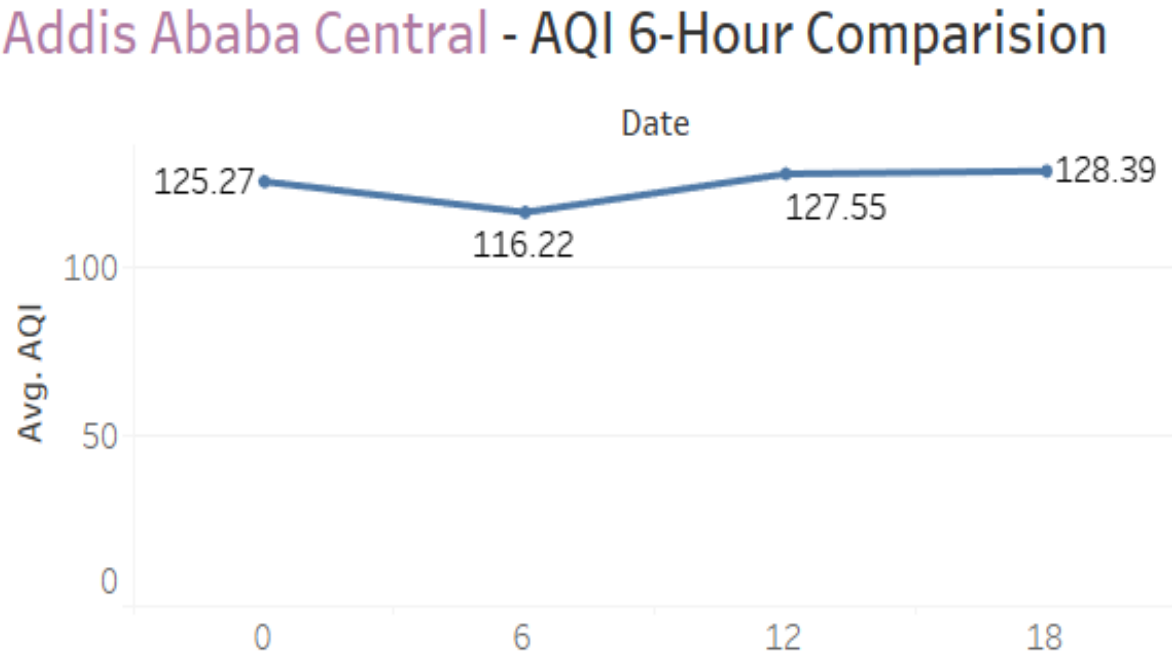
- The year 2018 records the highest frequency of readings falling in the '*Unhealthy for sensitive groups*' category and 2017 for the '*Unhealthy*' category
- While 2020 records a high number of readings under the unhealthy categories, the average AQI for the year is the lowest at 123.1 implying that the quality of air has somewhat improved.





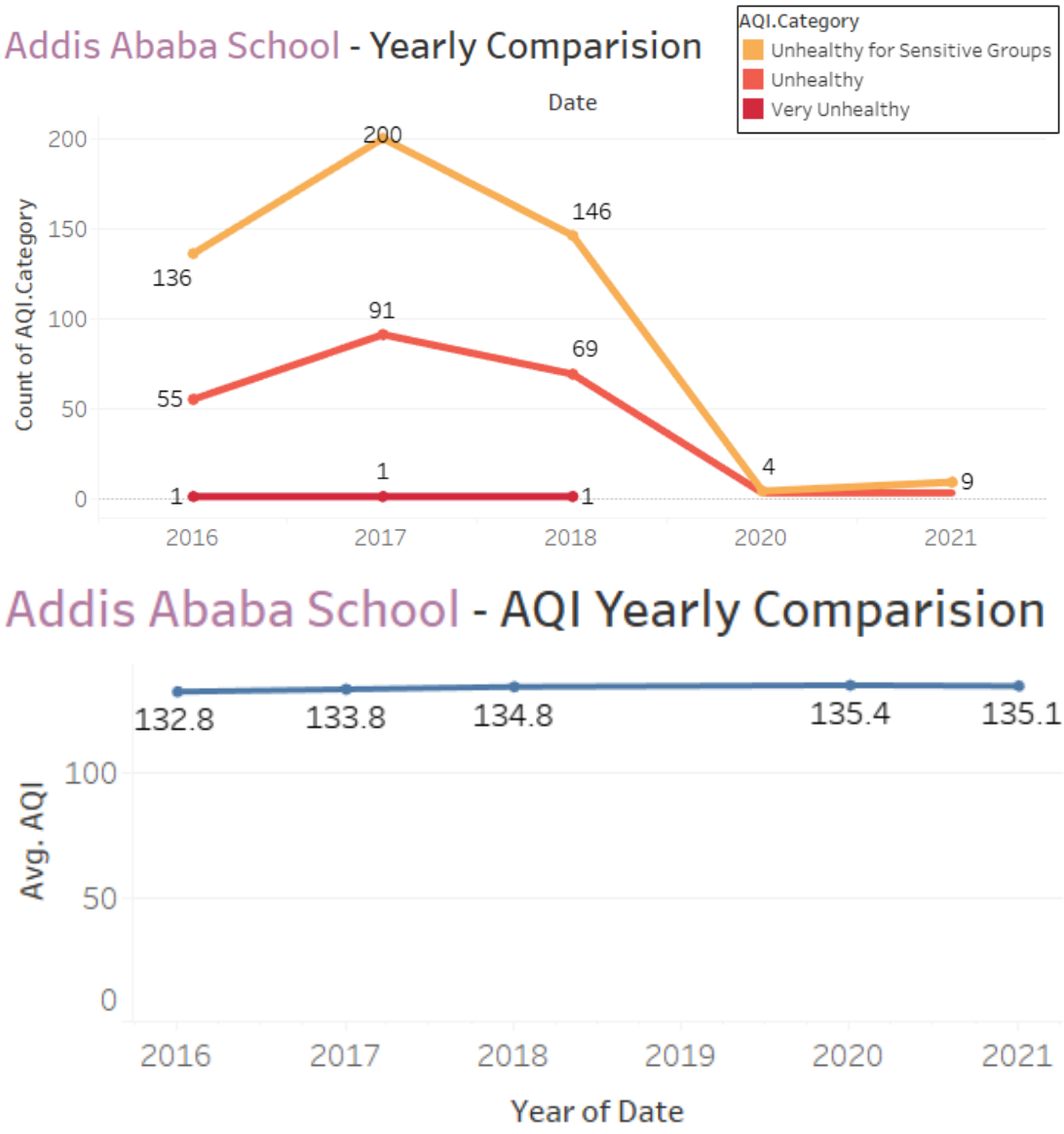
- Throughout the years, the month of June records the highest number of readings falling in the '*Unhealthy*' and '*Unhealthy for Sensitive Groups*' categories.
- September has the highest value of average AQI at 134.01 indicating that it is the worst month in terms of air quality, throughout the years.



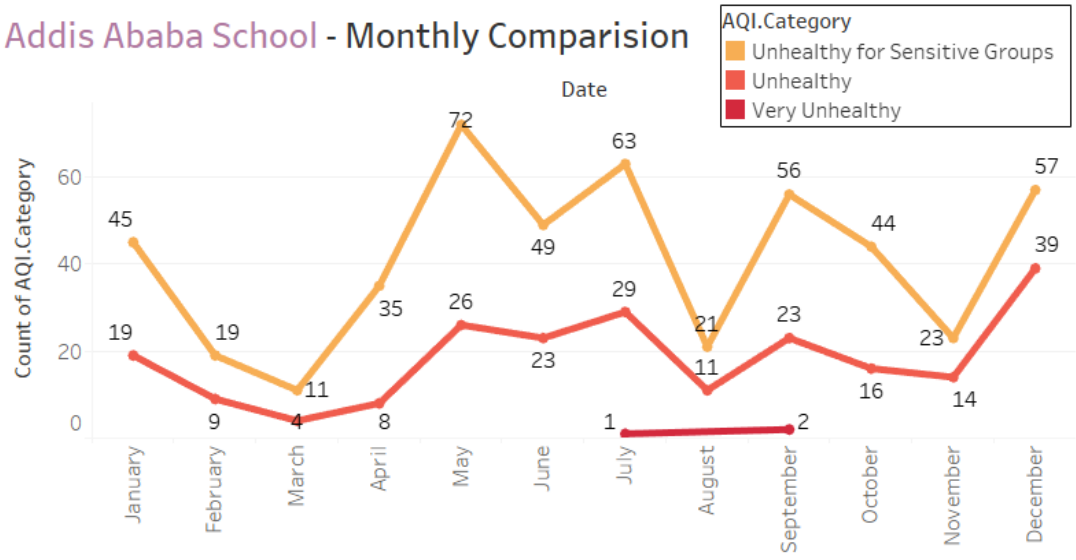


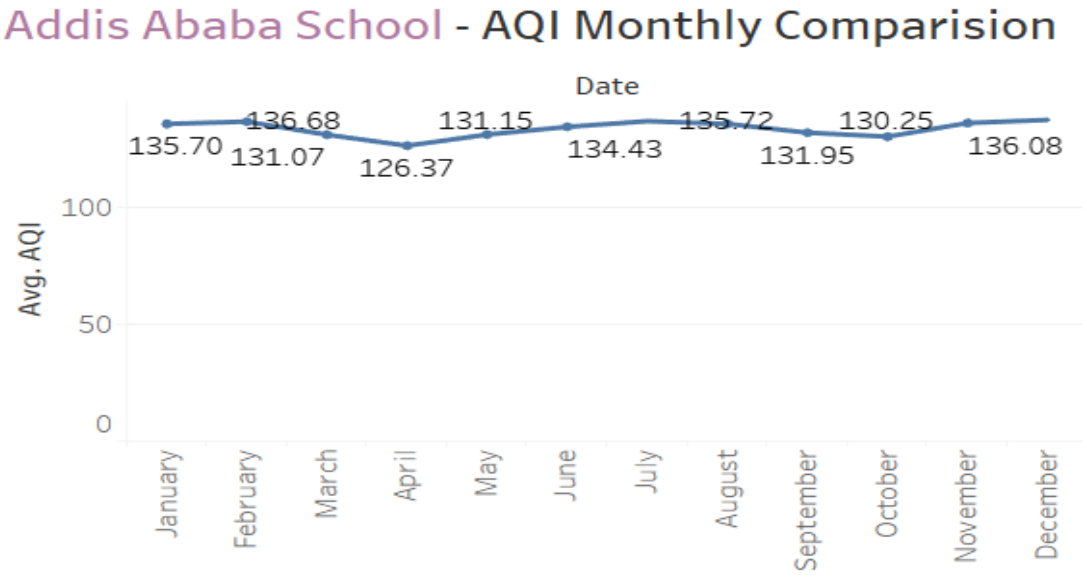
- The device records the lowest number of observations falling in the unhealthy categories at the 6-hour of the day and the quality of air seems to get comparatively better at that time.
- For the sensitive category, the air quality seems to be the worst at the 12-hour, while for the '*Unhealthy*' category, the quality continues to deteriorate after the 6-hour.
- The pollution rate is higher during the peak hours (beginning and end of the day) as the device is located on one of the busiest roads in the city.

3) Addis Ababa School

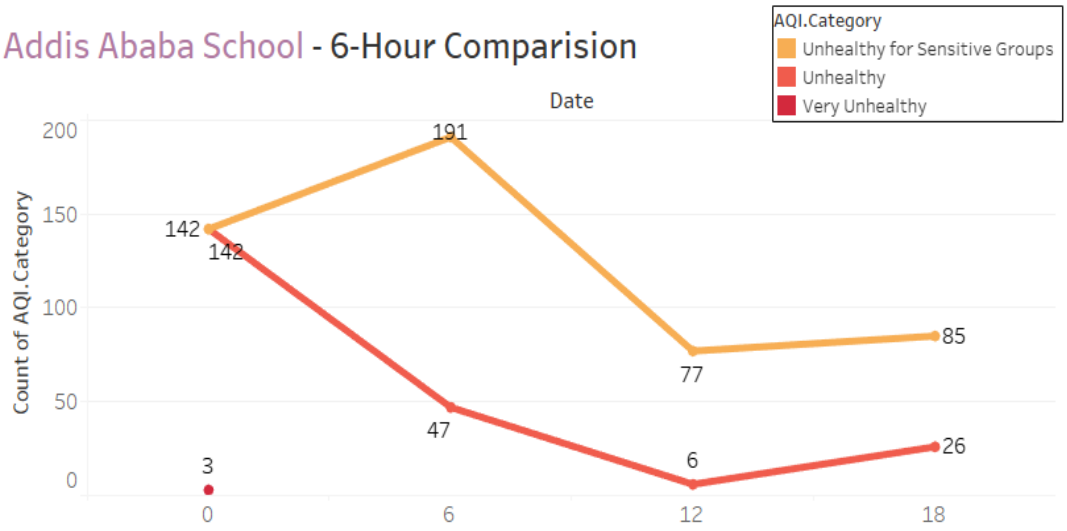


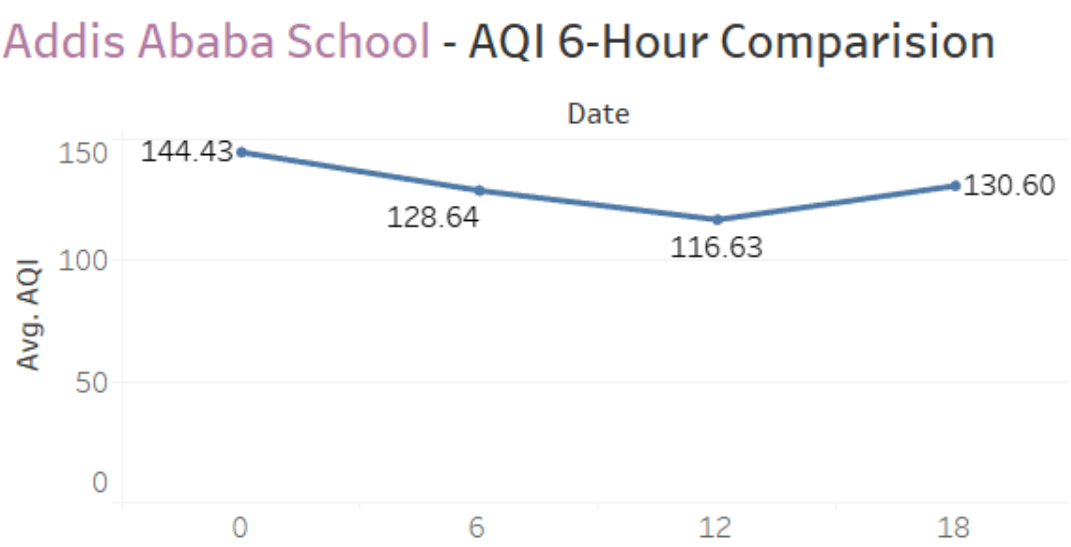
- The year 2017 recorded the maximum observations in the given unhealthy categories with 200 readings falling in the 'Unhealthy for sensitive groups' category and 91 in the 'Unhealthy' category.
- Since data for 2021 is incomplete, available readings for 2020 are very few and the data for 2019 is not available, we cannot make any conclusions about the trend of air quality. However, for the three years for which the data is available (2016-2018), it can be observed that 2016 has the least frequency of observations falling in the three unhealthy categories combined.
- There is not much difference in the values of average AQI throughout the years. It remains almost the same for all the years. However, 2016 has a comparatively lower value (132.8) for average AQI among all the years of analysis.





- May records the highest frequency of observations falling in the category of '*Unhealthy for sensitive groups*' and July for the '*Unhealthy*' category.
- The month of April seems to be comparatively better among the others in terms of air quality with an average AQI value of 126.37





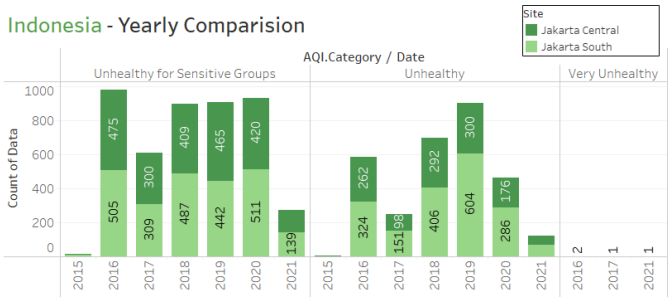
- The quality of air seems to be the best at the 12-hour during the day.
- The average AQI with a starting value of around 144 continued to drop during the day until 12-hour, after which it shows an increasing trend.
- The 6-hour seems to be affecting the sensitive category the most with 191 observations recorded in that category alone.
- The beginning of the day (0-hour) seems to be the worst during the day with 142 observations being recorded in both the '*Unhealthy for sensitive groups*' and '*Unhealthy*' categories each. This hour also records a few values falling under the '*Very unhealthy*' category.

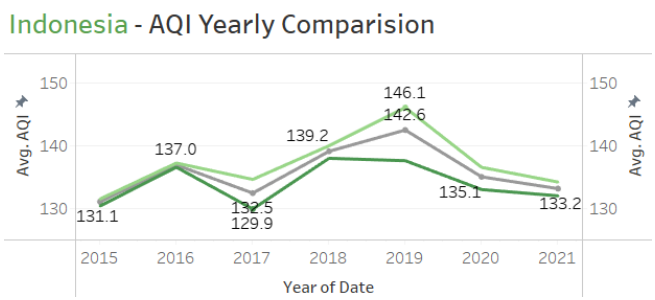
Country-wise analysis for AQI categories

1) Indonesia

The AQLI data show that air quality was not a pressing problem in Indonesia just two decades ago, but that air quality declined substantially in recent decades—with the steepest decline since 2013. From 1998 to 2016, the particulate air pollution concentrations have increased by 171 percent and the country went from being one of the cleaner countries in the world to one of the twenty most polluted.

Comparing the data by Year





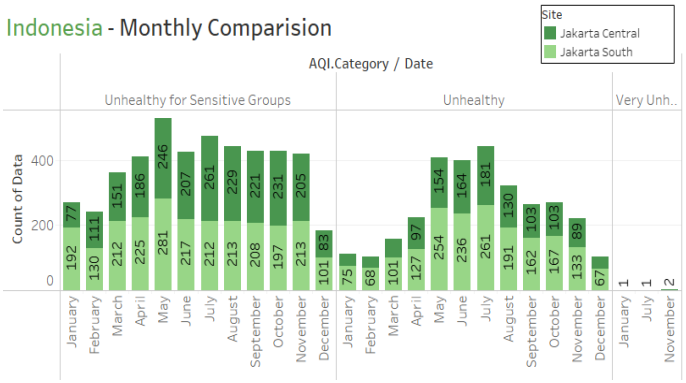
Indonesia - Yearly Comparision

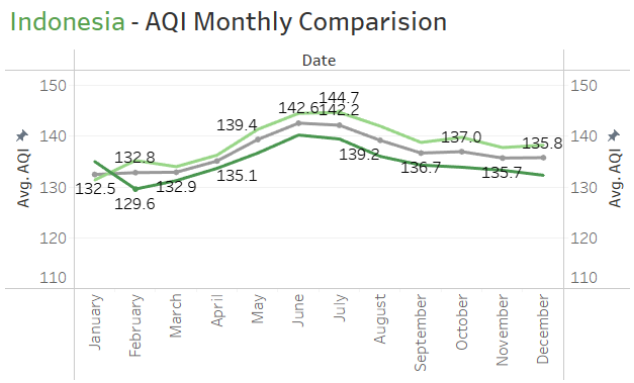
AQI.Category	Site	Year 2015	2016	2017	2018	2019	2020	2021	Grand ..
Unhealthy for Sensitive Groups	Jakarta Central	5	475	300	409	465	420	134	2,208
	Jakarta South	8	505	309	487	442	511	139	2,401
	Total	13	980	609	896	907	931	273	4,609
Unhealthy	Jakarta Central	2	262	98	292	300	176	55	1,185
	Jakarta South	3	324	151	406	604	286	68	1,842
	Total	5	586	249	698	904	462	123	3,027
Very Unhealthy	Jakarta Central		2						2
	Jakarta South			1				1	2
	Total		2	1				1	4
Grand Total		18	1,568	859	1,594	1,811	1,393	397	7,640

The above graphs for yearly comparison between the two cities of Indonesia suggest the following:

- The air quality of Jakarta South is even worse than Jakarta Central on most days with the majority of the readings falling in the '*Unhealthy for sensitive groups*' bracket.
- 2019 seems to be the worst year for Indonesia in terms of air quality with an average AQI of 142.6
- After the year 2017, the quality of air in Indonesia deteriorated continuously and significantly until 2019, after which the average AQI indicated a downward trend.
- To improve the air quality, preventative measures such as wearing high-quality particle filtering masks, avoiding outdoor activities, and exercise when pollution levels are particularly high, were taken after the year 2019.

Comparing the data by Month





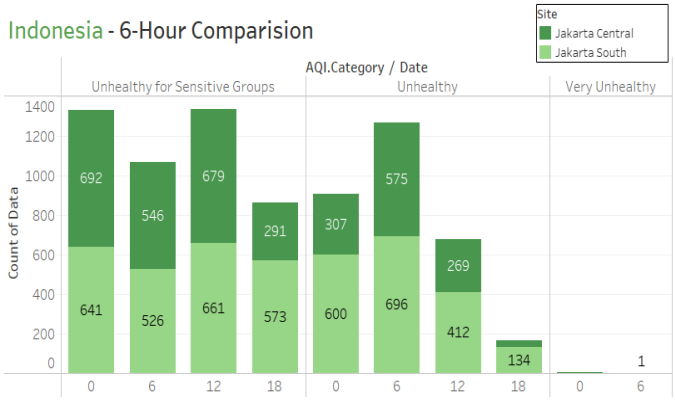
Indonesia - Monthly Comparision

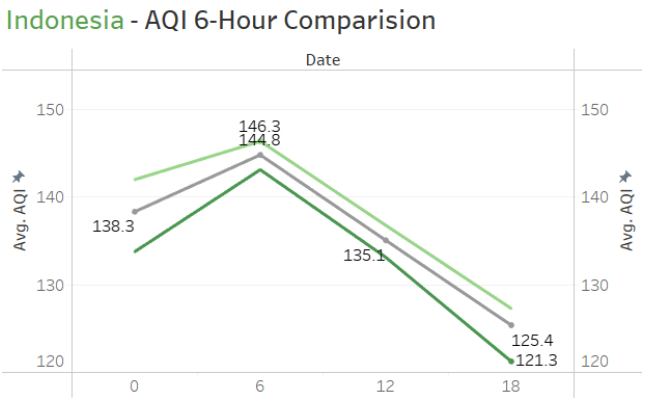
		Date												
AQI.Cat..	Site	1	2	3	4	5	6	7	8	9	10	11	12	Grand..
Unhealthy for Sensitive Groups	Jakarta Central	77	111	151	186	246	207	261	229	221	231	205	83	2,208
	Jakarta South	192	130	212	225	281	217	212	213	208	197	213	101	2,401
	Total	269	241	363	411	527	424	473	442	429	428	418	184	4,609
Unhealthy	Jakarta Central	37	35	56	97	154	164	181	130	103	103	89	36	1,185
	Jakarta South	75	68	101	127	254	236	261	191	162	167	133	67	1,842
	Total	112	103	157	224	408	400	442	321	265	270	222	103	3,027
Very Unhealthy	Jakarta Central											2		2
	Jakarta South	1						1						2
	Total	1						1				2		4
Grand Total		382	344	520	635	935	824	916	763	694	698	642	287	7,640

The above graphs for monthly comparison between the two cities of Indonesia suggest the following:

- For all the years of our analysis, the air quality in Jakarta South is worse than that of Jakarta Central during all the months.
- Throughout the years, the majority of the recorded values were found in the '*Unhealthy for sensitive groups*' category while there are only 4 recorded values for the '*Very unhealthy*' category.
- May is the most unhealthy month, particularly for the sensitive population. Overall, the months of June and July are unhealthy for everyone, throughout the year.

Comparing the data in a 6-hour interval





Indonesia - 6-Hour Comparision

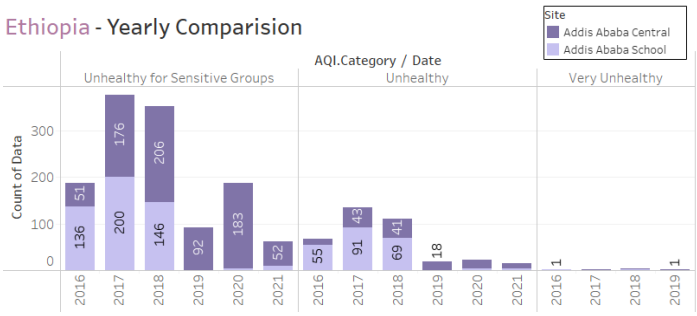
AQI.Cat..	Site	Date 0	6	12	18	Grand..
Unhealthy for Sensitive Groups	Jakarta Central	692	546	679	291	2,208
	Jakarta South	641	526	661	573	2,401
	Total	1,333	1,072	1,340	864	4,609
Unhealthy	Jakarta Central	307	575	269	34	1,185
	Jakarta South	600	696	412	134	1,842
	Total	907	1,271	681	168	3,027
Very Unhealthy	Jakarta Central	1	1			2
	Jakarta South	2				2
	Total	3	1			4
Grand Total		2,243	2,344	2,021	1,032	7,640

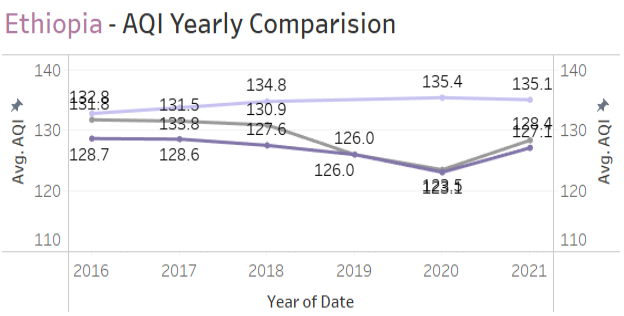
The above graphs for a 6-hourly comparison between the two cities of Indonesia suggest the following:

- The mean AQI is highest on the 6th hour of the day (144.8) and declines continuously after that, with a value of 125.4 for the 18-hour during the day. This could be due to the vehicle and industry emissions which start from the beginning of the day.
- Over the years, Indonesia recorded only very few readings which fall into the '*Very unhealthy*' bracket.

2) Ethiopia

Comparing the data by Year





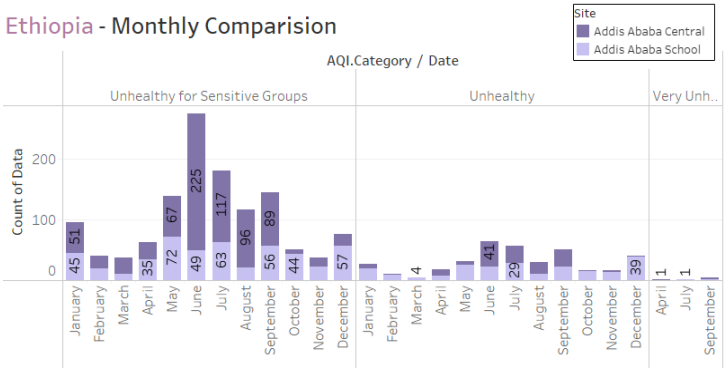
Ethiopia - Yearly Comparision

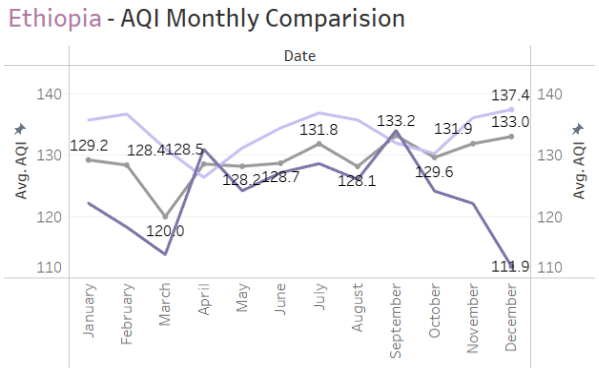
AQI.Category	Site	Year 2016	2017	2018	2019	2020	2021	Grand ..
Unhealthy for Sensitive Groups	Addis Ababa Central	51	176	206	92	183	52	760
	Addis Ababa School	136	200	146		4	9	495
	Total	187	376	352	92	187	61	1,255
Unhealthy	Addis Ababa Central	13	43	41	18	19	12	146
	Addis Ababa School	55	91	69		3	3	221
	Total	68	134	110	18	22	15	367
Very Unhealthy	Addis Ababa Central		1	2	1			4
	Addis Ababa School	1	1	1				3
	Total	1	2	3	1			7
Grand Total		256	512	465	111	209	76	1,629

The above graphs for yearly comparison between the two cities of Ethiopia suggest the following:

- The air quality in Ethiopia mostly lies in the 'Unhealthy for sensitive groups' bracket throughout the years.
- Although the average AQI observed is the least for the year 2020, it cannot be concluded that 2020 is the best year in terms of air quality due to the unavailability of complete data for 2020 at the Addis Ababa school.
- Also, the entire data corresponding to the site '*Addis Ababa school*' for the year 2019 is unavailable and so is half of the overall data for 2021, it is hard to predict the accurate trend. However, among the three years (2016-2018), 2016 seems to be the best year for Ethiopia in terms of air quality.

Comparing the data by Month





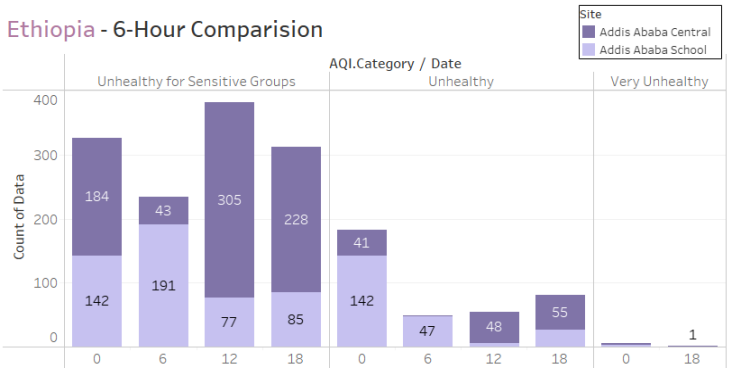
Ethiopia - Monthly Comprision

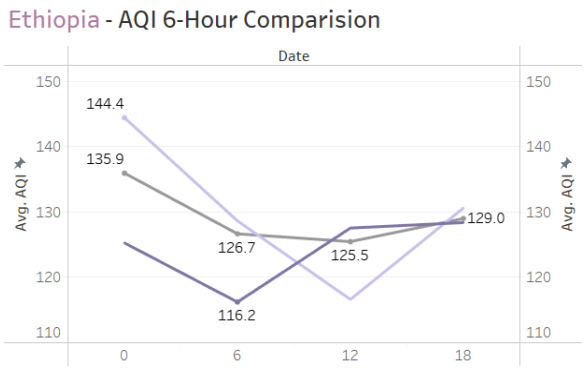
		Date												
AQI.Cat..	Site	1	2	3	4	5	6	7	8	9	10	11	12	Grand..
Unhealthy for Sensitive Groups	Addis Ababa Central	51	21	27	28	67	225	117	96	89	6	14	19	760
	Addis Ababa School	45	19	11	35	72	49	63	21	56	44	23	57	495
	Total	96	40	38	63	139	274	180	117	145	50	37	76	1,255
Unhealthy	Addis Ababa Central	8	2		10	6	41	28	19	28	1	2	1	146
	Addis Ababa School	19	9	4	8	26	23	29	11	23	16	14	39	221
	Total	27	11	4	18	32	64	57	30	51	17	16	40	367
Very Unhealthy	Addis Ababa Central				1					3				4
	Addis Ababa School							1		2				3
	Total				1			1		5				7
Grand Total		123	51	42	82	171	338	238	147	201	67	53	116	1,629

The above graphs for monthly comparison between the two cities of Ethiopia suggest the following:

- Most of the recorded observations fall in the '*Unhealthy for sensitive groups*' category. For Addis Ababa Central, June recorded a significantly higher number of readings under this category as compared to other months.
- The average AQI is the least for March.

Comparing the data in a 6-hour interval





Ethiopia - 6-Hour Comparision

AQI.Cat..	Site	Date 0	6	12	18	Grand..
Unhealthy for Sensitive Groups	Addis Ababa Central	184	43	305	228	760
	Addis Ababa School	142	191	77	85	495
	Total	326	234	382	313	1,255
Unhealthy	Addis Ababa Central	41	2	48	55	146
	Addis Ababa School	142	47	6	26	221
	Total	183	49	54	81	367
Very Unhealthy	Addis Ababa Central	3			1	4
	Addis Ababa School	3				3
	Total	6			1	7
Grand Total		515	283	436	395	1,629

- The average AQI for the country (and both cities) starting from 135.9 kept dropping until the 12th hour, after which it began to rise again, indicating the hours during the day when the air quality is comparatively better.
- The pollution rate is higher during the peak hours (beginning and end of the day) as the device is located on one of the busiest roads in the city. This could be due to more vehicles on the road during the peak hours and as a result more emissions.

Frequency of AQI Categories

Indonesia

Indonesia - Heatmap by Hour

CNT(AQI)
1 52

Month	Hour	Unhealthy for Sensitive Groups								Unhealthy								Very Unhealthy				Grand ..
		2015	2016	2017	2018	2019	2020	2021	Total	2015	2016	2017	2018	2019	2020	2021	Total	2016	2017	2021	Total	
1	0		24	6	16	12	20	11	89		12			7	12	7	38			1	1	128
	6		14	4	20	14	13	7	72		31	1		6	15	9	62					134
	12		26	3	16	6	14	1	66		5			2	4		11					77
	18		10	3	16		12	1	42						1		1					43
	Total		74	16	68	32	59	20	269		48	1		15	32	16	112			1	1	382
2	0		20	10	16	17	20	9	92		13	3	4	8	12	3	43					135
	6		13	11	18	16	12	7	77		13	3	5	19	10	4	54					131
	12		11	10	8	10	12		51		2	1		2			5					56
	18		7	1	4	2	7		21				1				1					22
	Total		51	32	46	45	51	16	241		28	7	10	29	22	7	103					344
3	0		25	15	23	9	21	28	121		12	7	11	10	7	5	52					173
	6		19	14	19	8	21	18	99		21	10	14	5	14	10	74					173
	12		23	13	20	6	21	18	101		8		2		8	5	23					124
	18		8	3	7	1	13	10	42		5		1			2	8					50
	Total		75	45	69	24	76	74	363		46	17	28	15	29	22	157					520
4	0		28	10	18	11	26	16	109		13	8	17	15	10	14	77					186
	6		26	16	14	21	26	20	123		19	8	26	13	20	21	107					230
	12		22	9	21	21	27	25	125		2	3	9	8	5	7	34					159
	18		5	5	14	14	14	2	54			2			2	2	6					60
	Total		81	40	67	67	93	63	411		34	21	52	36	37	44	224					635
5	0		22	26	17	19	30	30	144		28	15	35	25	11	5	119					263
	6		13	20	10	16	25	28	112		41	27	45	34	22	22	191					303
	12		35	21	24	33	29	31	173		13	17	26	14	8	7	85					258
	18		16	11	26	17	19	9	98		1	3	3	4	2		13					111
	Total		86	78	77	85	103	98	527		83	62	109	77	43	34	408					935
6	0		20	15	31	16	21	2	105		34	5	18	40	32		129					234
	6		9	22	20	12	20		83		41	12	28	42	32		155					238
	12		31	20	21	20	32		124		10	1	20	36	18		85					209
	18		23	5	25	33	26		112		4		4	13	10		31					143
	Total		83	62	97	81	99	2	424		89	18	70	131	92		400					824
7	0		25	17	21	24	33		120		21	9	38	37	15		120		1		1	241
	6		27	16	4	10	24		81		25	12	48	52	24		161					242
	12		40	27	14	22	22		125		7	8	45	40	23		123					248
	18		27	18	30	41	31		147		1	2	20	11	4		38					185
	Total		119	78	69	97	110		473		54	31	151	140	66		442		1		1	916
8	0		21	18	24	30	29		122		26	5	15	16	8		70					192
	6		27	18	9	26	16		96		22	17	25	31	36		131					227
	12		44	24	13	21	29		131		9	10	29	40	18		106					237
	18		21	9	19	29	15		93		1	2	4	7			14					107
	Total		113	69	65	106	89		442		58	34	73	94	62		321					763
9	0		26	19	30	22	28		125		17	9	15	23	8		72					197
	6		18	23	15	14	21		91		21	15	22	36	11		105					196
	12		30	35	32	11	33		141		6	8	13	40	5		72					213
	18		10	7	20	24	11		72		2	2	1	9	2		16					88
	Total		84	84	97	71	93		429		46	34	51	108	26		265					694
10	0		25	19	33	26	28		131		19	5	20	28	7		79					210
	6		19	14	20	18	13		84		20	4	26	40	10		100					184
	12		29	17	30	22	29		127		6		27	39	2		74					201
	18		22	5	17	29	13		86		2		6	8	1		17					103
	Total		95	55	100	95	83		428		47	9	79	115	20		270					698
11	0		30	13	28	23	31		125		12	4	20	27	9		72	1			1	198
	6	1	30	13	21	23	15		103		16	1	19	26	22		84	1			1	188
	12		26	13	35	26	19		119	1	13	4	4	22	2		46					165
	18		15	4	21	23	8		71		6	1	2	11			20					91
	Total	1	101	43	105	95	73		418	1	47	10	45	86	33		222	2			2	642
12	0		7	2	9	31	1		50		3	1	15	17			36					86
	6	5	4		14	27	1		51	3	3	4	7	30			47					98
	12	7	3	4	9	34			57	1			6	10			17					74
	18		4	1	4	17			26				2	1			3					29
	Total	12	18	7	36	109	2		184	4	6	5	30	58			103					287
Grand Total		13	980	609	896	907	931	273	4,609	5	586	249	698	904	462	123	3,027	2	1	1	4	7,640

The above heatmap of Indonesia shows the relative frequency by assigning each value a color representation. Heatmaps are used to show relationships between the two variables, one plotted on each axis. By observing how cell colors change across each axis, it can be observed if there are any patterns in value for one or both variables.

Observations:

- There is a high concentration of '*Unhealthy*' data from May to July for the years 2018 and 2019. During that period, the 18th-hour data was comparatively healthier than other hours of the day.
- The year 2017 seems to have the least frequency total, indicating that the year records the least number of readings falling in the category '*Unhealthy for Sensitive Groups*'.
- Overall, the year 2015 records the least frequency of readings falling in any of the three unhealthy categories.

Ethiopia

Ethiopia - Heatmap by Hour

CNT(AQI) 1 30

		Unhealthy for Sensitive Groups							Unhealthy							Very Unhealthy					
Month	Hour	2016	2017	2018	2019	2020	2021	Total	2016	2017	2018	2019	2020	2021	Total	2016	2017	2018	2019	Total	Grand ..
1	0		7	12	2	7	3	31		7	5		2	1	15						46
	6		19	10		2	6	37		3	5			2	10						47
	12		2	16	1	3	3	25			1		1		2						27
	18			3				3													3
	Total		28	41	3	12	12	96		10	11		3	3	27						123
2	0		7	5		3	5	20		1	1		1	1	4						24
	6		7	3			1	11		2	4				6						17
	12		1	2		1	2	6						1	1						7
	18			1			2	3													3
	Total		15	11		4	10	40		3	5		1	2	11						51
3	0		3			3		6			1				1						7
	6		9	1		4		14		1	2				3						17
	12			3	1	9		13													13
	18			2		3		5													5
	Total		12	6	1	19		38		1	3				4						42
4	0		5	8		3	1	17			5	2		1	8			1		1	26
	6		16	4	3		1	24			2				2						26
	12			7	3	4	1	15													15
	18		1	2	2	2		7			6	1	1		8						15
	Total		22	21	8	9	3	63			13	3	1	1	18			1		1	82
5	0		15	13		1	1	30		8	12	1			21						51
	6		10	17				27		1					1						28
	12		11	21	2	9	3	46		1	1				2						48
	18		9	19	3	4	1	36		3	4		1		8						44
	Total		45	70	5	14	5	139		13	17	1	1		32						171
6	0	1	16	18	5	4	7	51	2	8	8		1	3	22						73
	6	2	15	10	1	2	1	31													31
	12	2	26	25	16	20	15	104		3	6	3	1	2	15						119
	18	3	23	23	15	16	8	88	3	6	10	1	3	4	27						115
	Total	8	80	76	37	42	31	274	5	17	24	4	5	9	64						338
7	0	10	9	18		5		42	15	1	9				25			1		1	68
	6	2	2	1		1		6													6
	12	7	14	24		15		60		11	2		1		14						74
	18	13	19	30		10		72	3	5	10				18						90
	Total	32	44	73		31		180	18	17	21		1		57			1		1	238
8	0	8	2	6	2	4		22	7		5	2	1		15						37
	6	2		3	1			6													6
	12	12	9	10	6	11		48	3		1	3			7						55
	18	9	8	11	10	3		41	4	1	1	1	1		8						49
	Total	31	19	30	19	18		117	14	1	7	6	2		30						147
9	0	15	7	3	2	4		31	14	13	1	1	2		31	1	1	1	1	4	66
	6	13	1			1		15		2					2						17
	12	18	15	1	4	12		50	2	4		2	1		9						59
	18	18	13	2	6	10		49	3	3	1	1	1		9		1			1	59
	Total	64	36	6	12	27		145	19	22	2	4	4		51	1	2	1	1	5	201
10	0	6	14	6		1		27	1	6	5				12						39
	6		11	4				15		1					1						16
	12		3					3			1				1						4
	18	2	2			1		5		2	1				3						8
	Total	8	30	10		2		50	1	9	7				17						67
11	0	5	6	2	2	1		16		7					7						23
	6	1	11	1				13		6					6						19
	12	3	1		1			5		2			1		3						8
	18	1	2					3													3
	Total	10	20	3	3	1		37		15			1		16						53
12	0	12	11	3	4	3		33	6	13			3		22						55
	6	21	10			4		35	5	13					18						53
	12		4	2		1		7													7
	18	1						1													1
	Total	34	25	5	4	8		76	11	26			3		40						116
Grand Total		187	376	352	92	187	61	1,255	68	134	110	18	22	15	367	1	2	3	1	7	1,629

From the above heatmap of Ethiopia, the following observations can be made:

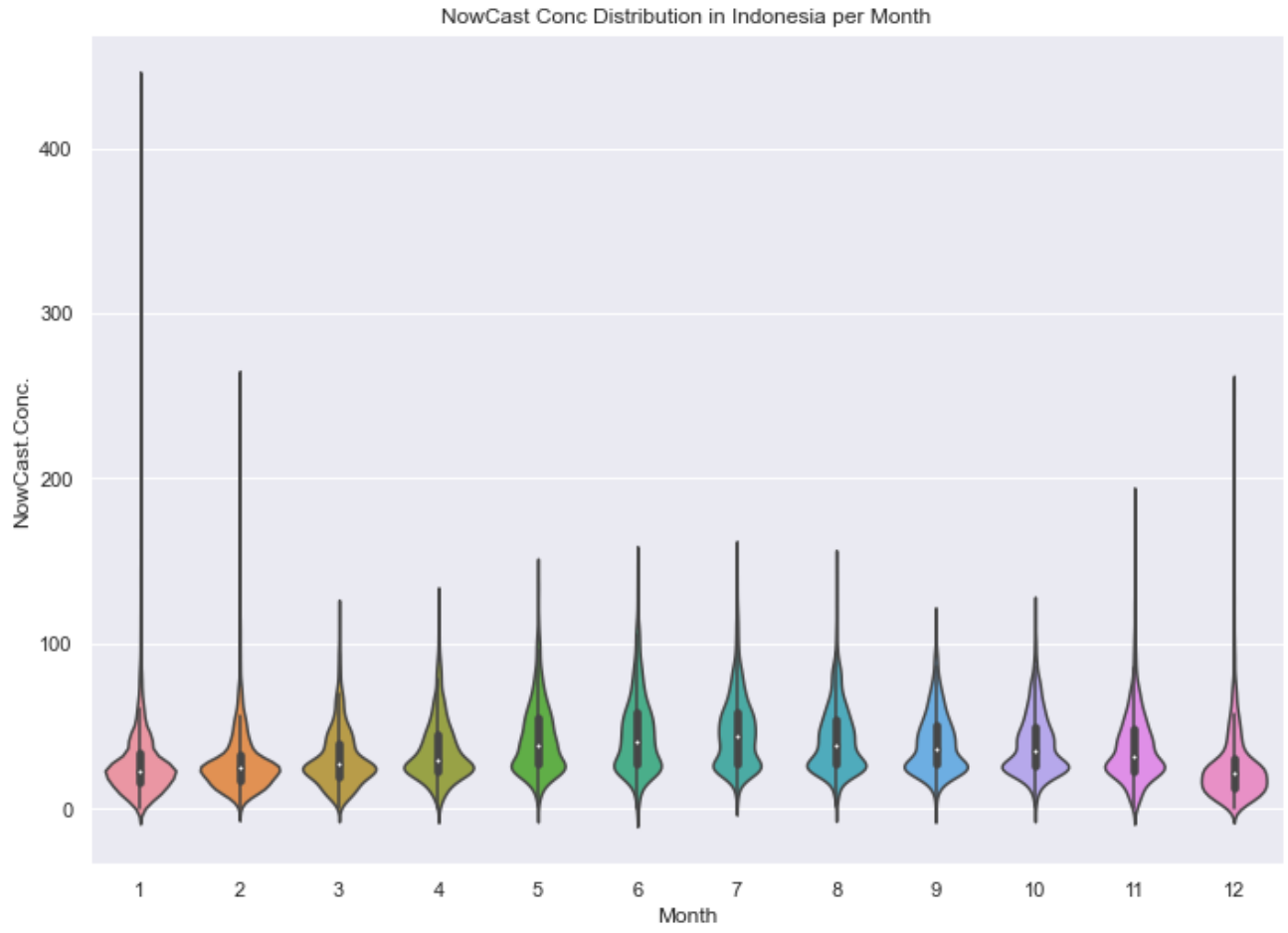
- There is a high concentration of data falling in the category '*Unhealthy for sensitive groups*' from May to July for the years 2017 and 2018.
- June records the highest frequency of values in all the three unhealthy categories combined,
- Overall, the year 2017 records the highest frequency of data falling in the three unhealthy categories.
- The entire data corresponding to the site '*Addis Ababa school*' for the year 2019 is unavailable and so is half of the overall data for 2021, therefore, it is hard to predict the accurate trend. However, among the three years (2016-2018), 2016 seems to have recorded the least number of observations under the unhealthy categories.

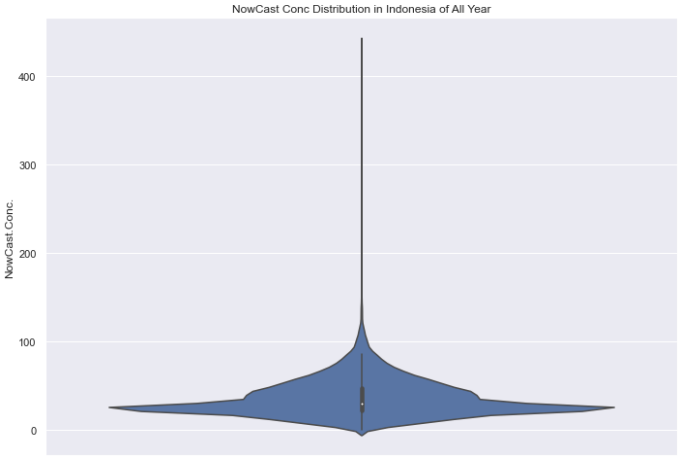
Distribution of variables using Violin Plots

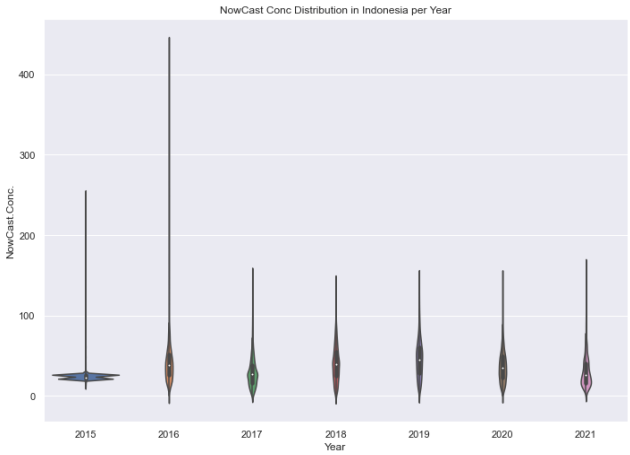
After comparing the AQI categories, the data distribution and density of the numerical data (NowCast, AQI, and Raw Concentration) are checked using the violin graphs. Violin plots are similar to box plots, except that they also show the probability density of the data at different values.

For Indonesia

Distribution of NowCast Concentration



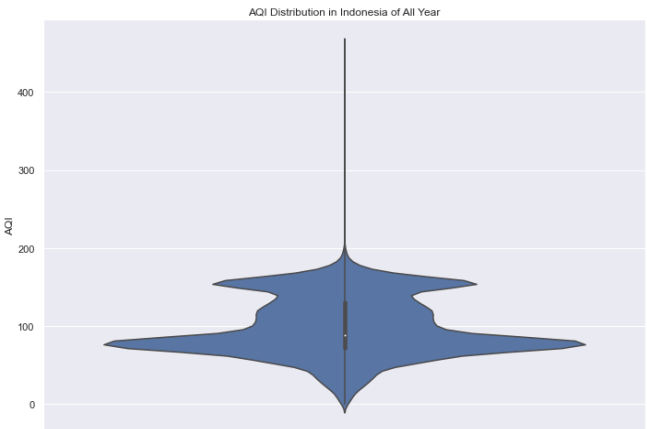


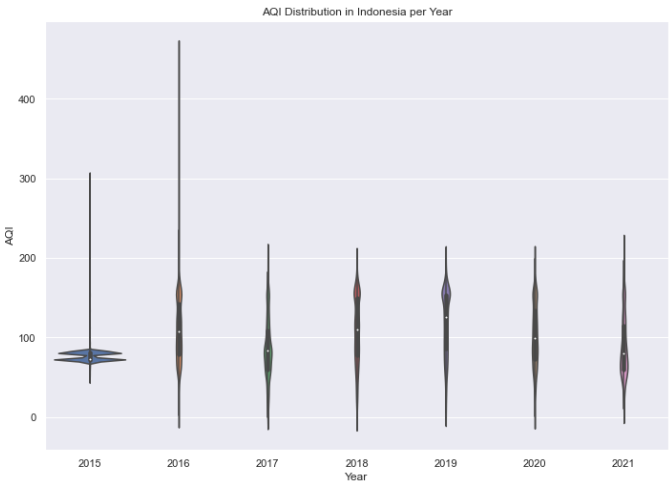


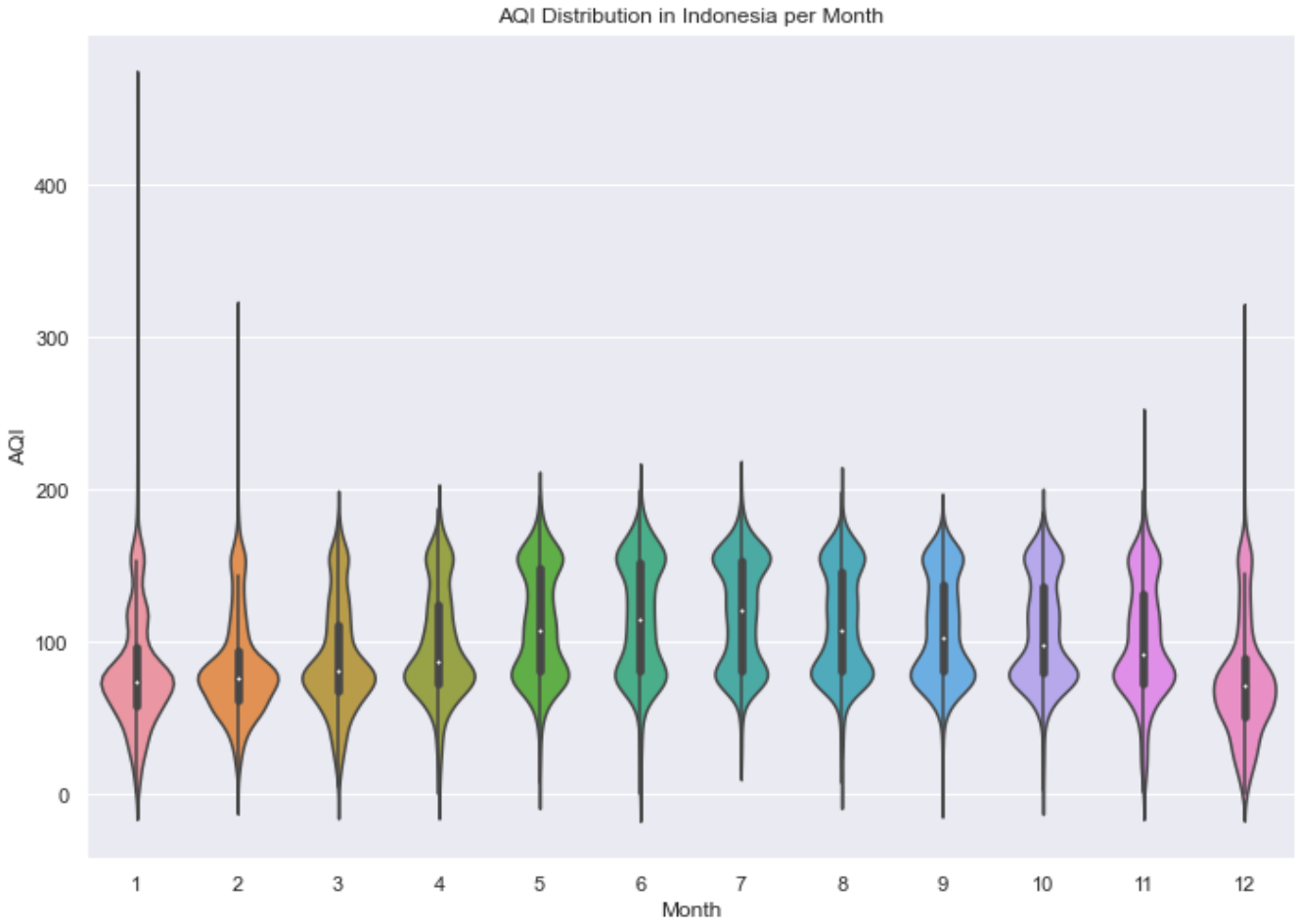
From the above Violin Plots for NowCast Concentration of Indonesia, the following observations can be made:

- The overall NowCast Data of Indonesia is concentrated around 25.
- The data corresponding to the year 2015 is more concentrated in the graphs and is different from other years. The plot for the other years is flatter indicating an even distribution.
- The median of May to November is higher in comparison and the shape is narrower, which means the data is evenly distributed.
- The median value for December to April is lower among all the months and the shape of the distribution (extremely skinny on each end and wide in the middle) indicates that the values of NowCast Concentration are highly concentrated around the median.

Distribution of AQI



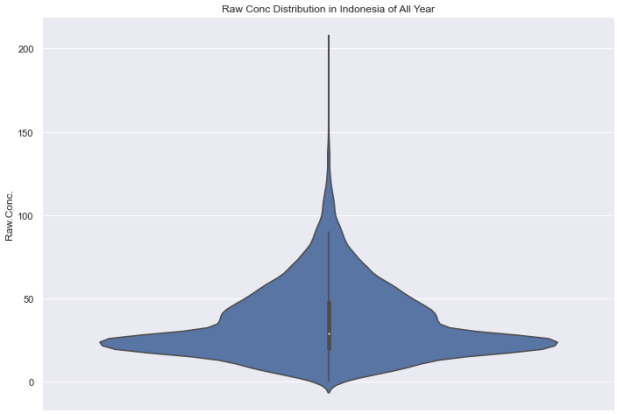


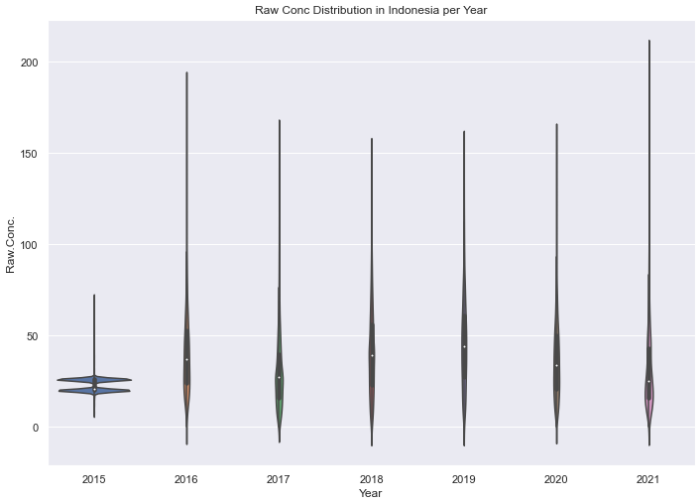


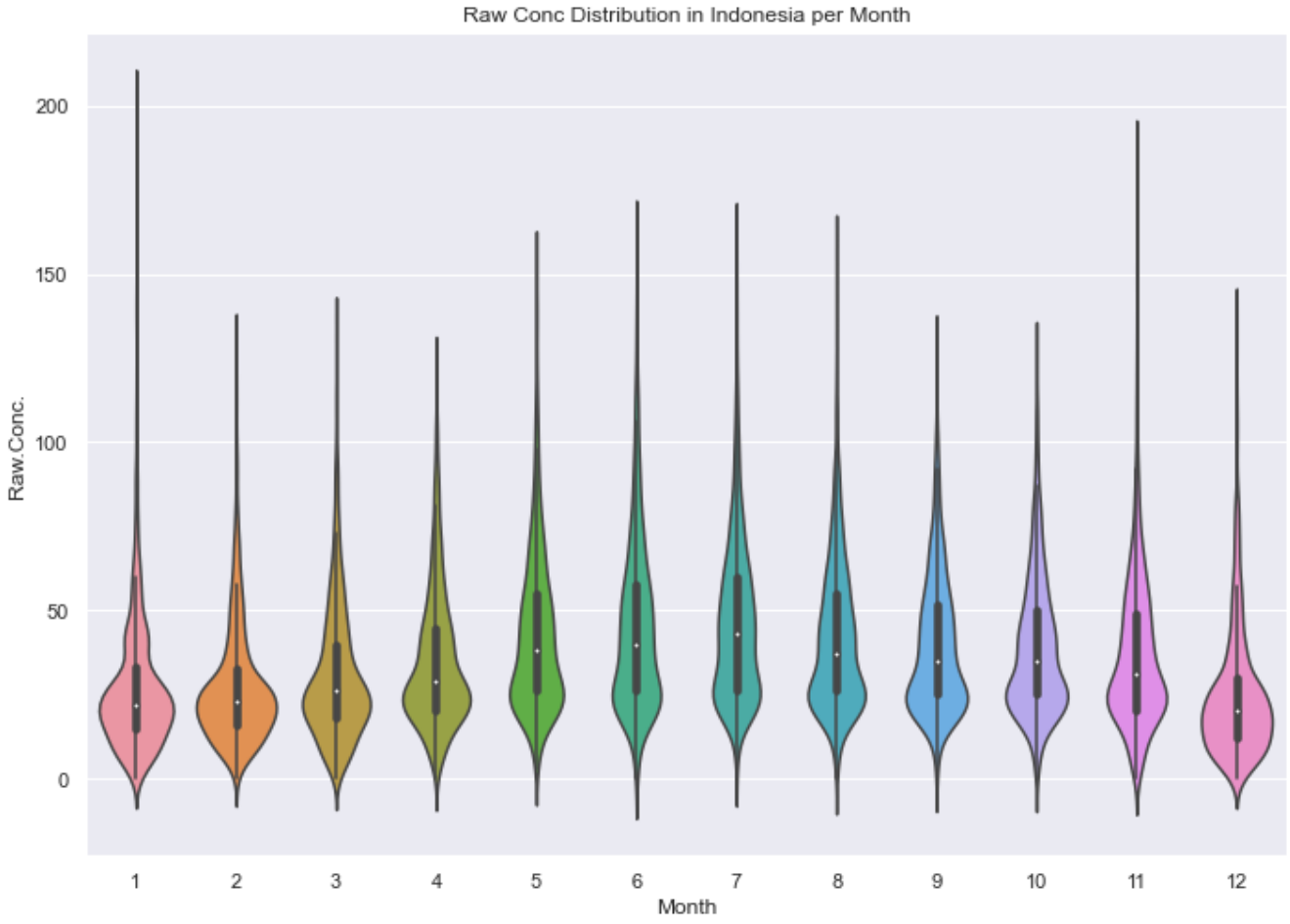
From the above Violin Plots for AQI of Indonesia, the following observations can be made:

- The overall AQI Data of Indonesia is concentrated at around 90 and 170.
- The data corresponding to the year 2015 is more concentrated in the graphs and is totally different from other years. The plot for the other years is flatter indicating an even distribution.
- The median of May to November is higher in comparison and the shape is narrower, which means the data is evenly distributed.
- The median value for December to April is comparatively lower among all the months and the shape of the distribution (extremely skinny on each end and wide in the middle) indicates that the values of AQI are highly concentrated around the median.

Distribution of Raw Concentration





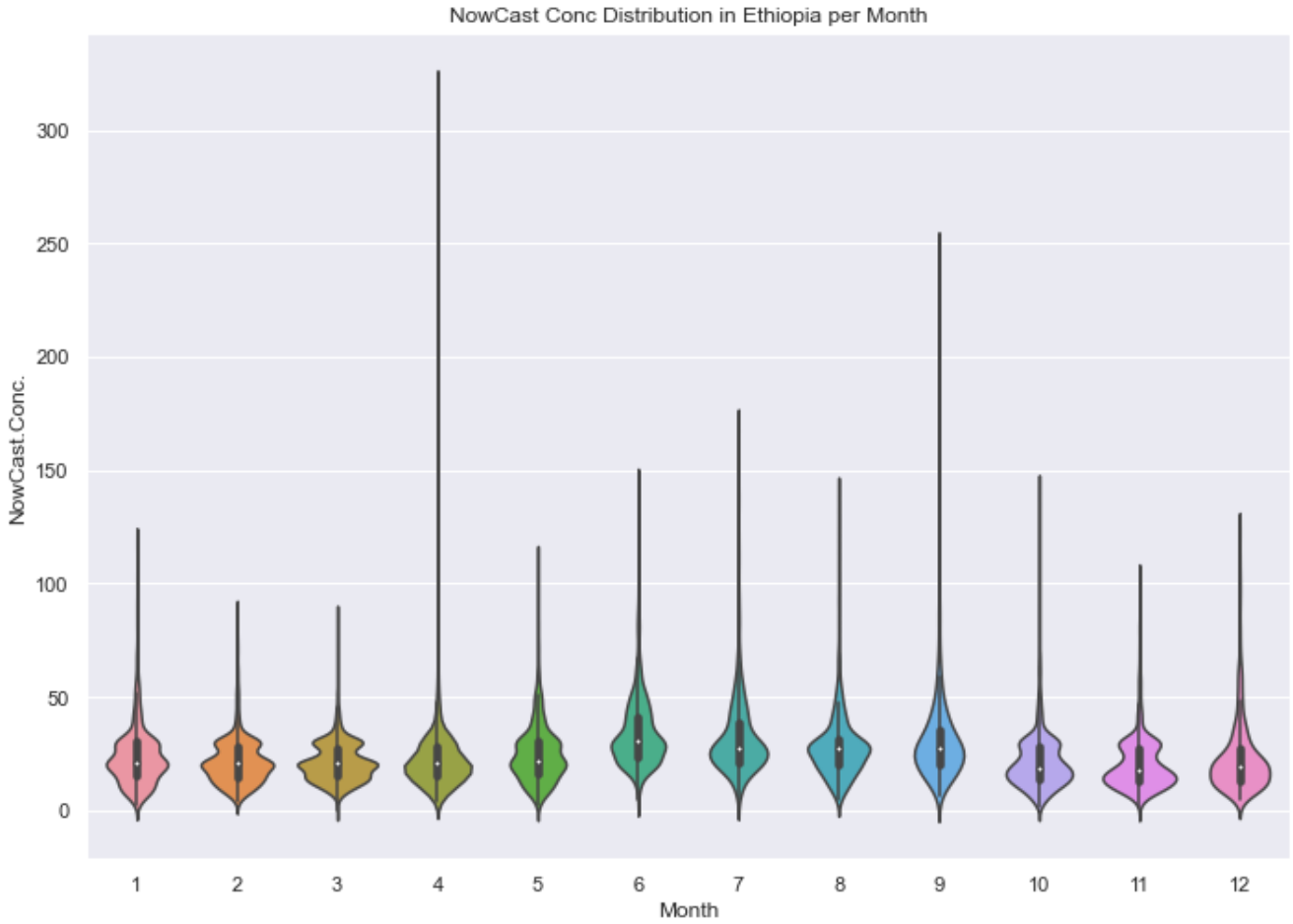


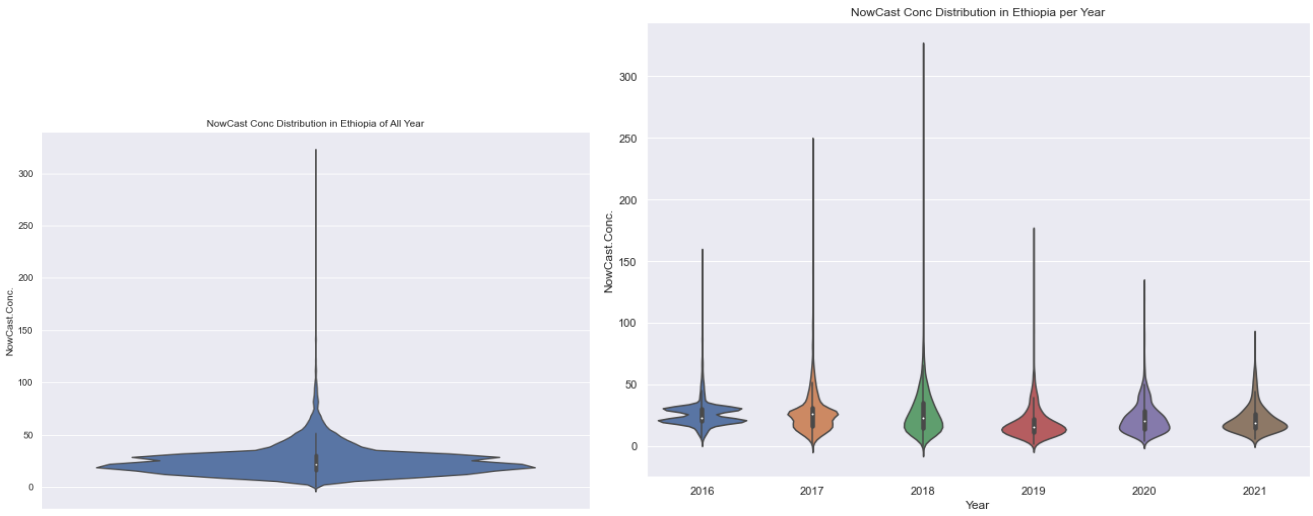
From the above Violin Plots for Raw Concentration of Indonesia, the following observations can be made:

- The overall Raw Concentration Data of Indonesia is concentrated at around 25.
- The data corresponding to the year 2015 is more concentrated in the graphs and is different from other years. The plot for the other years is flatter indicating an even distribution.
- The median of May to November is higher in comparison and the shape is narrower, which means the data is evenly distributed.
- The median value for December to April is comparatively lower among all the months and the shape of the distribution (extremely skinny on each end and wide in the middle) indicates that the values of Raw Concentration are highly concentrated around the median.

For Ethiopia

Distribution of NowCast Concentration

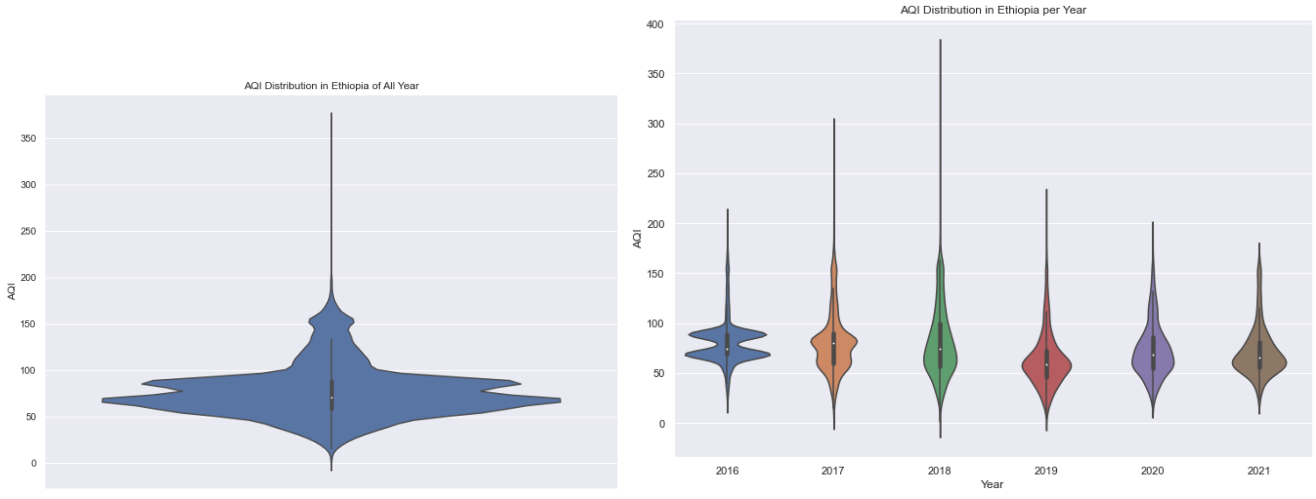


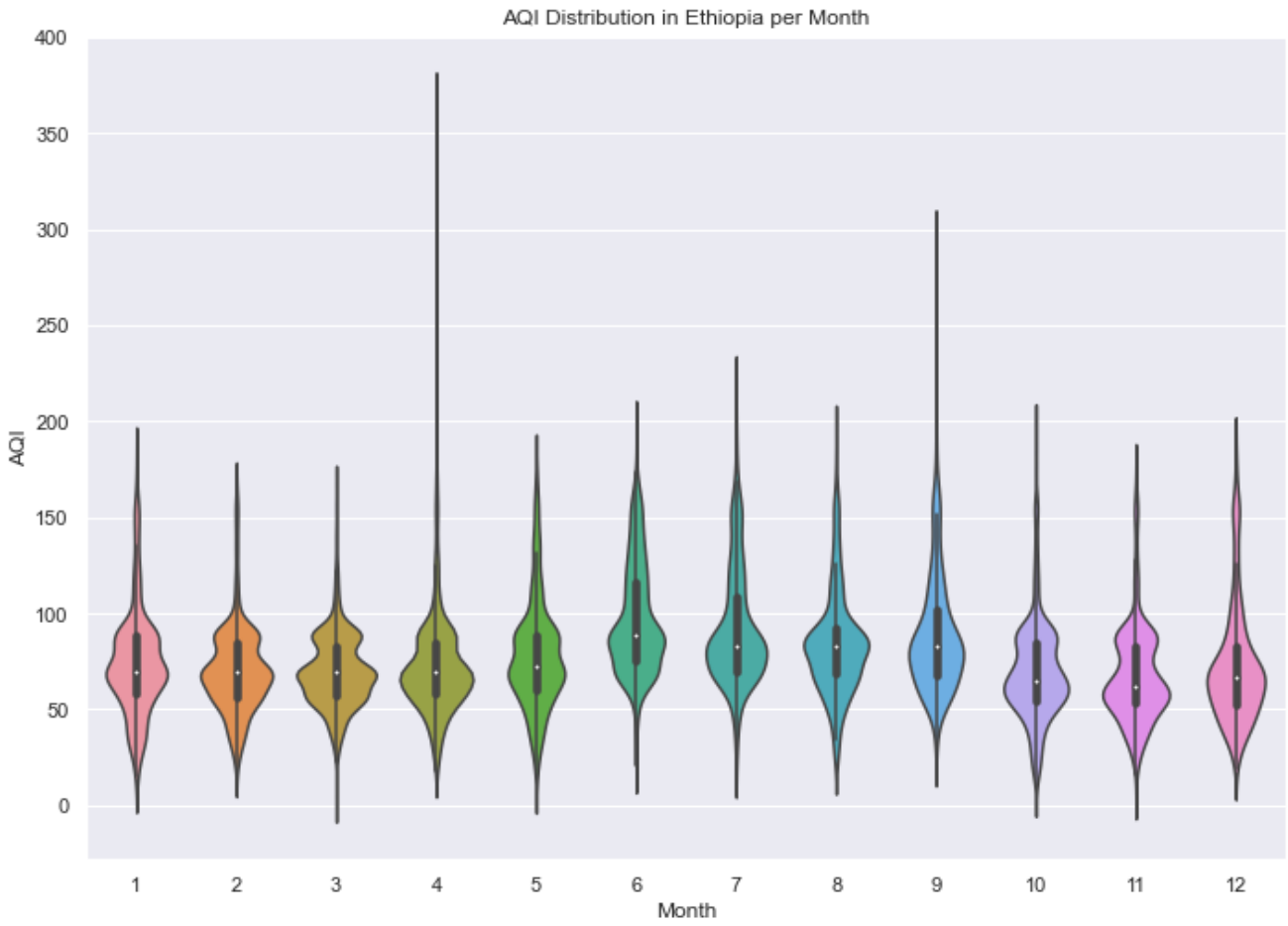


From the above Violin Plots for NowCast Concentration of Ethiopia, the following observations can be made:

- The overall NowCast concentration Data of Ethiopia is concentrated below 25.
- The median value is nearly the same for all the months.
- There are two peaks found in 2016 data, which is quite different from all the other years. Most of the years have a distribution with one peak.
- The median of June to September is higher in comparison and the shape is narrower, which means the data is evenly distributed.
- The median value for October to May is comparatively lower among all the months and the shape of the distribution (extremely skinny on each end and wide in the middle) indicates that the values of Nowcast Concentration are highly concentrated around the median.

Distribution of AQI

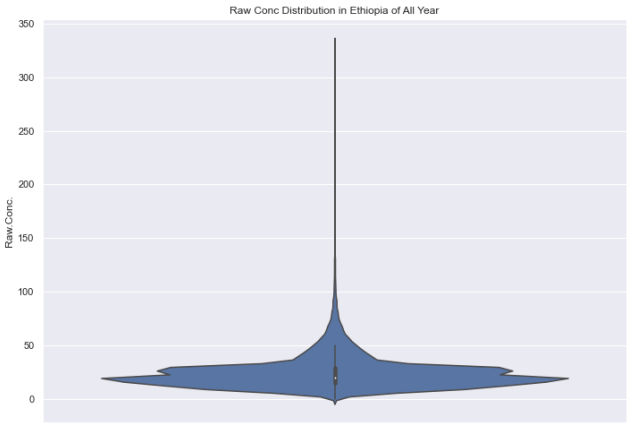


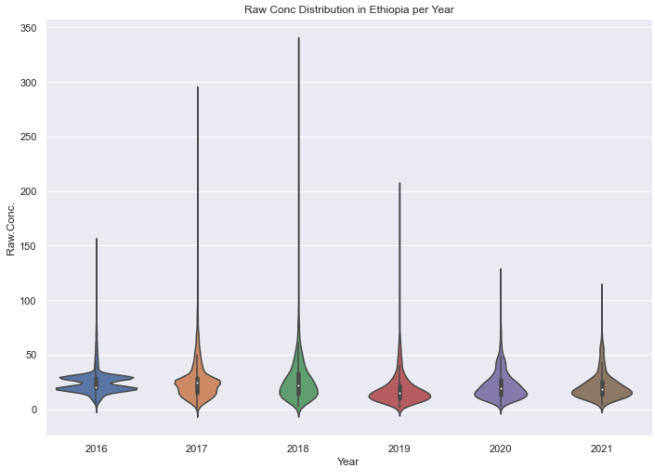


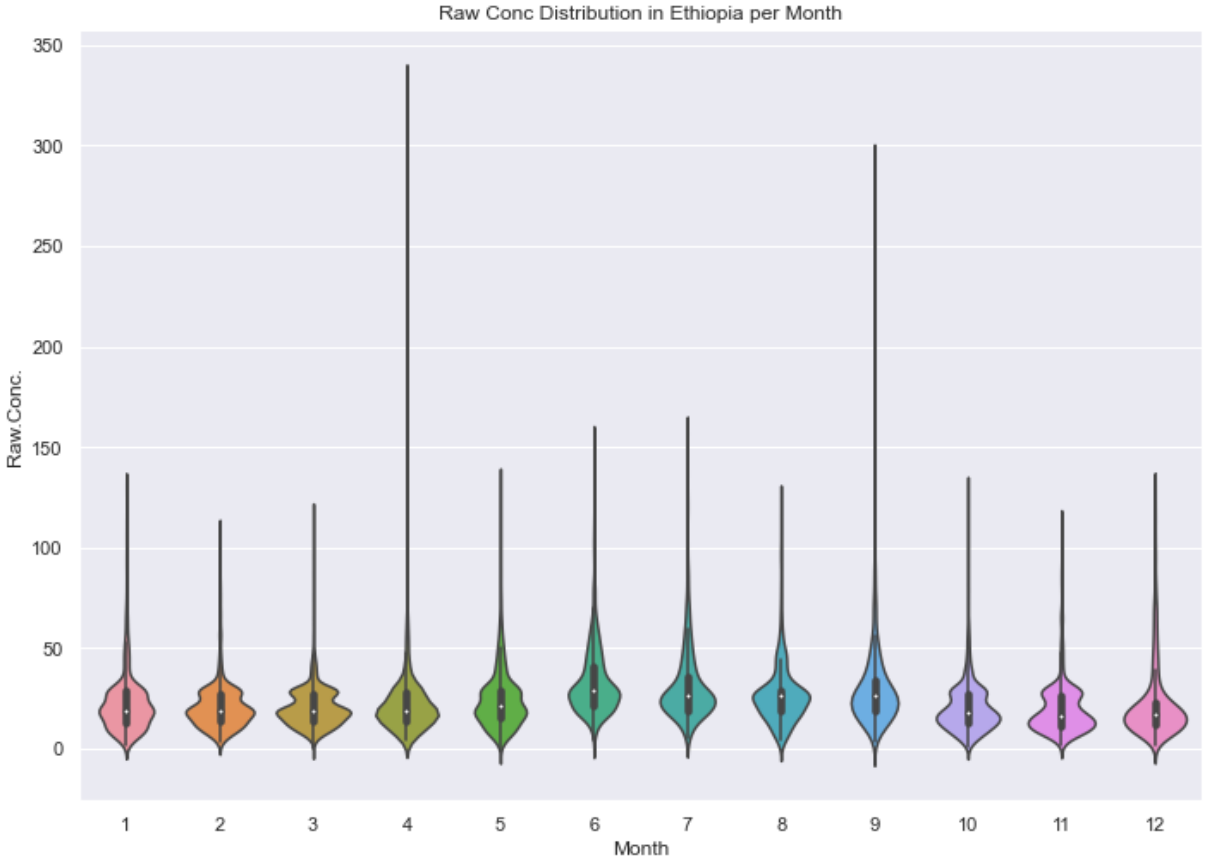
From the above Violin Plots for AQI of Ethiopia, the following observations can be made:

- The overall AQI Data of Ethiopia is concentrated at around 70 and 90.
- There are two peaks found in 2016 data, which is quite different from all the other years. Most of the years have a distribution with one peak.
- The median of June to September is higher in comparison and the shape is narrower, which means the data is evenly distributed.
- The median value for October to May is comparatively lower among all the months and the shape of the distribution (extremely skinny on each end and wide in the middle) indicates that the values of AQI are highly concentrated around the median.

Distribution of Raw Concentration







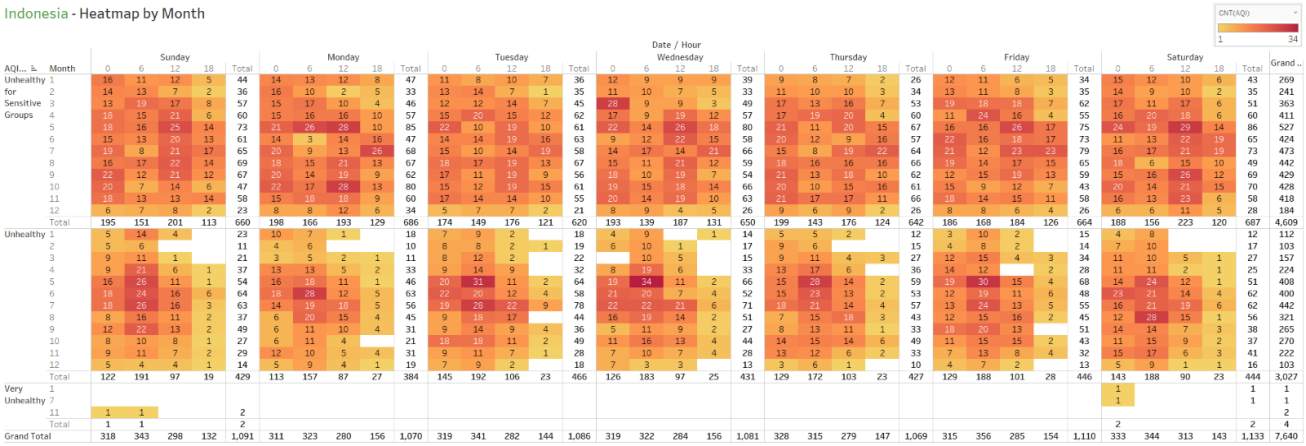
From the above Violin Plots for Raw Concentration of Ethiopia, the following observations can be made:

- The overall raw concentration Data of Ethiopia is concentrated below 25.
- There are two peaks found in 2016 data, which is quite different from all the other years. Most of the years have a distribution with one peak.

- The median of June to September is higher in comparison and the shape is narrower, which means the data is evenly distributed.
- The median value for October to May is comparatively lower among all the months and the shape of the distribution (extremely skinny on each end and wide in the middle) indicates that the values of AQI are highly concentrated around the median.

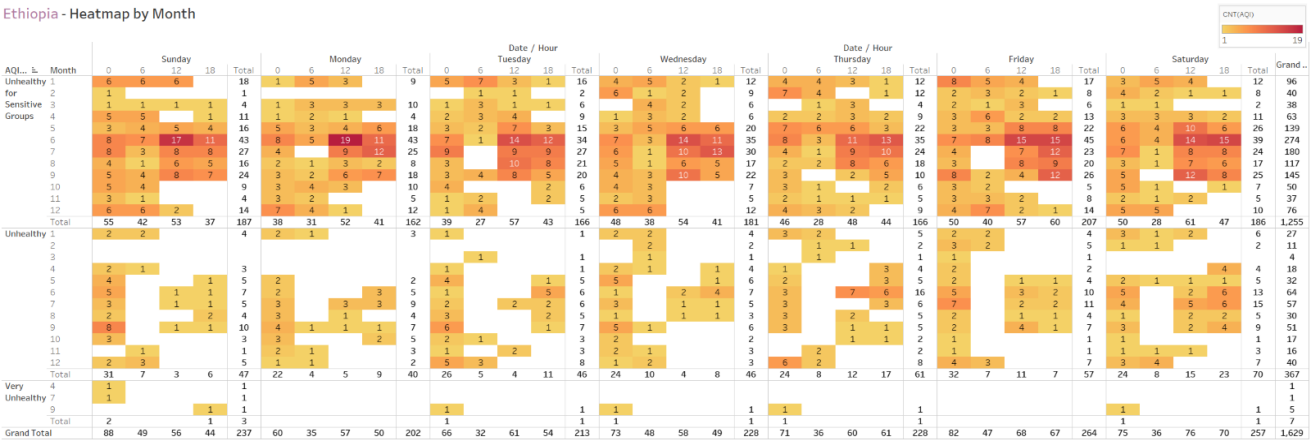
Overall six-hourly AQI Categories trend

For Indonesia



- The 12-hour for Saturday and Sunday has the highest frequency of readings falling under the '*Unhealthy for Sensitive Groups*' category, throughout the years.
- The 6-hour of everyday records the highest frequency of readings falling under the '*Unhealthy*' category, throughout the years.

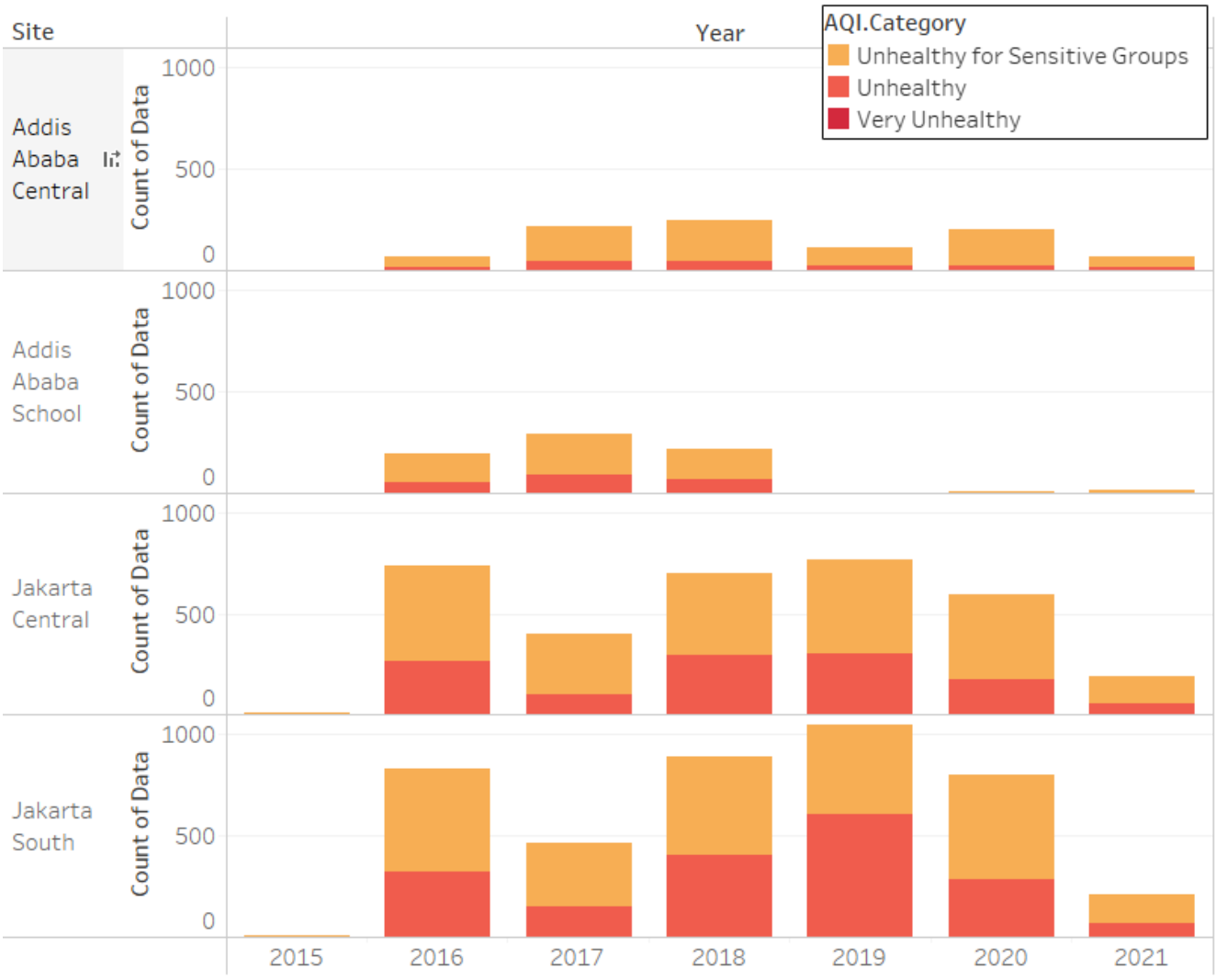
For Ethiopia



- The majority of the data falling under the "*Unhealthy for sensitive groups*" category is concentrated from July to September.
- The 12-hour of almost everyday records the highest frequency of readings falling under the '*Unhealthy for sensitive groups*' category, throughout the years.

Comparison of the AQI Categories between cities

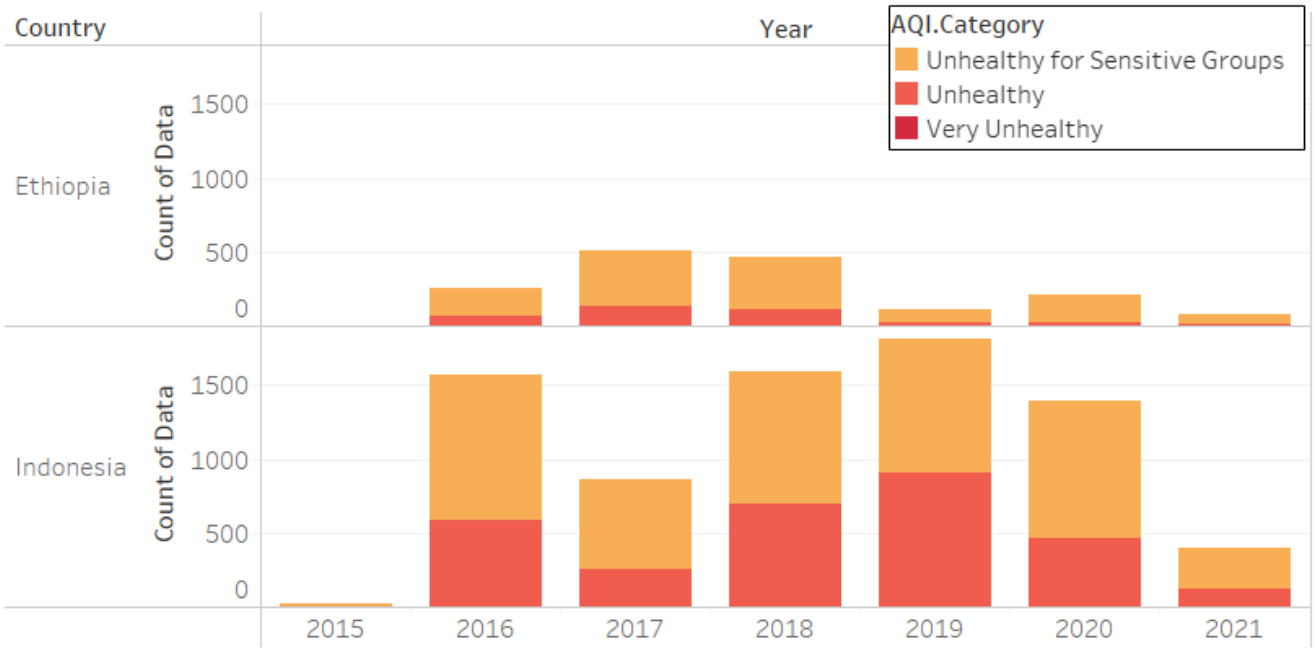
City-wise analysis



Following observations can be made from the above bar charts for the AQI Categories per year:

- The frequency distributions of the two cities within the countries look alike.
- The frequencies of records falling under the three unhealthy categories are comparatively lower in Jakarta Central than Jakarta South. It can be concluded that Jakarta South has worse air quality in comparison to Jakarta Central.
- Since the data for the 'Addis Ababa School' site is not complete, it is difficult to compare the two sites of Ethiopia.
- Jakarta South is the most polluted city among the four, where the frequency of all categories is the highest throughout the years of analysis.

Country-wise analysis



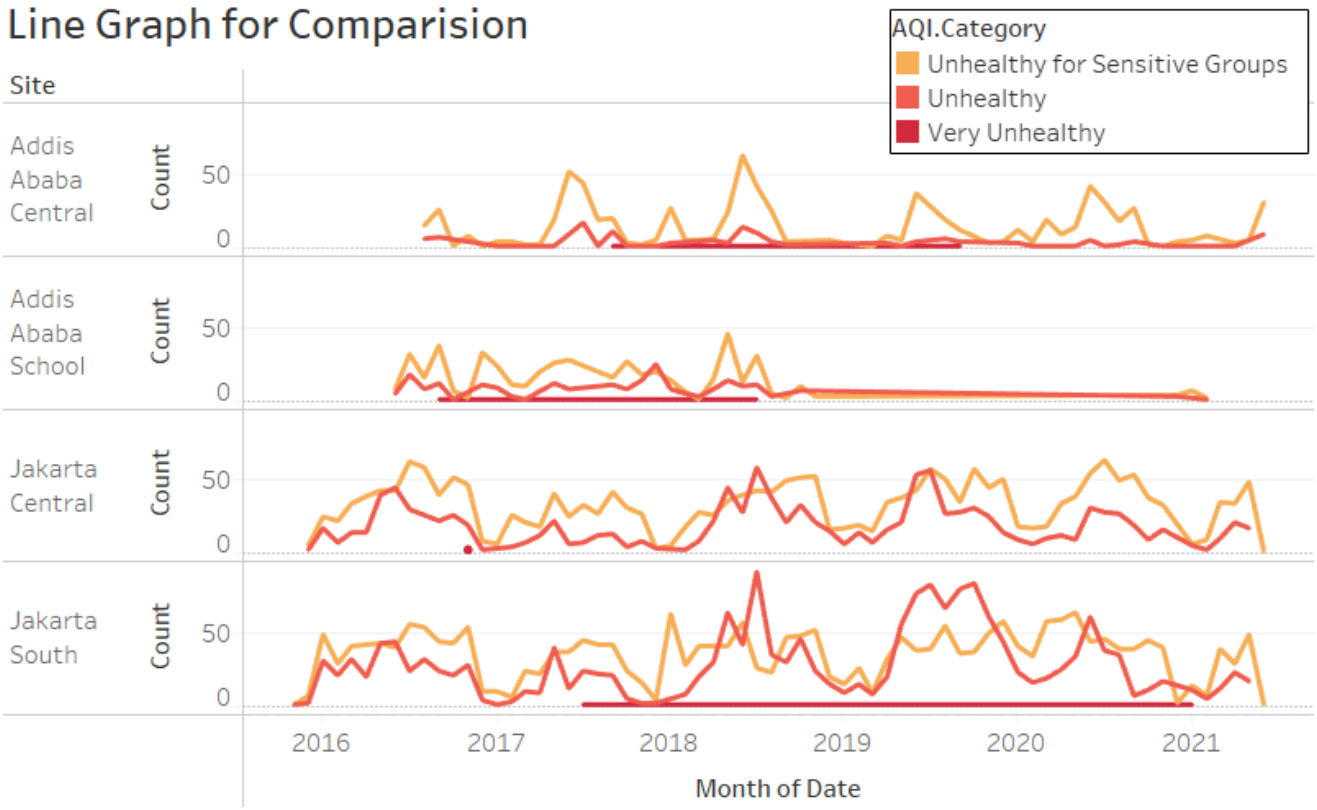
Following observations can be made from the above bar charts for the AQI Categories per year:

- The frequencies of records falling under the three unhealthy categories are much lower in Ethiopia. As a result, it can be concluded that the air quality of Indonesia is significantly worse in comparison to Ethiopia, throughout the years of our analysis.
- No significant improvement can be seen in the air quality of Indonesia in spite of a lockdown during the pandemic.
- Indonesia has become an outlier in Southeast Asia due to its plans to build new coal plants.[6]

[6] Nicholas, Hans. 2020. "*Lockdown should have cleared up Jakarta's air. Coal plants kept it dirty*." Mongabay. https://news.mongabay.com/2020/08/jakarta-air-pollution-coal-power-plant-covid-lockdown-crea-study/.

Line graph for the AQI Categories

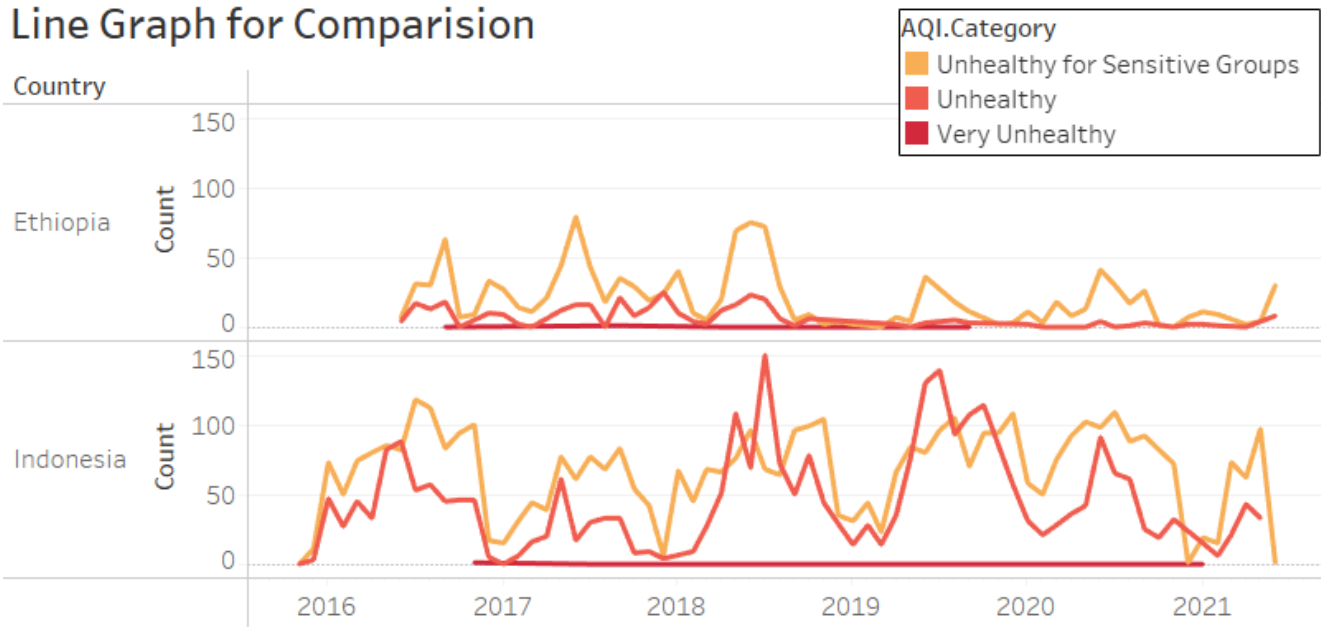
City-wise analysis



Following observations can be made from the above line graphs for the AQI Categories per year:

- The trend of the distributions of the two cities within the countries look alike.
- The '*Unhealthy*' and '*Very Unhealthy*' lines are flatter in Ethiopia, which indicates comparatively better air quality
- For all the cities, the peaks can be observed during the middle of the year, followed by a drop towards the beginning and the end.
- Jakarta South is the most polluted city among the four and has the highest frequency of unhealthy readings, throughout the years.

Country-wise analysis



From the above line charts, the following observations can be made:

- The trend indicates much lower frequencies of records falling under the three unhealthy categories in Ethiopia. As a result, it can be concluded that the air quality of Indonesia is significantly worse in comparison to Ethiopia, throughout the years of our analysis.
- No significant overall improvement can be seen in the air quality of Indonesia in spite of a lockdown during the pandemic.

INFERENTIAL ANALYSIS

The inferential analysis is done on the Air Quality dataset as well as for Covid Cases & Deaths for the cities in Indonesia and Ethiopia. The dataset for Covid cases/deaths contained cumulative frequencies, hence the data was cleaned to get accurate frequency per day.

Correlation Analysis

Aim:

Correlation is done between AQI, Nowcast. Conc. & Raw Conc. with Covid Cases & Death to check whether air quality affects the number of cases and/deaths. The month is also used for analysis to check whether a particular time in the year could explain the covid cases – like climate, any significant event in the country, festivals, etc.

Approach:

The Correlation analysis is performed country-wise, as the Covid dataset contains data for each country rather than for each city. The hypothesis is:

i. Null- Hypothesis (H_0): X-variable (AQI, Nowcast Conc. & Raw Conc.) and Y-variable (Covid Cases & Covid Deaths) do not have a linear correlation among all subjects in the population.
ii. Alternative Hypothesis (H_a): X-variable (AQI, Nowcast Conc. & Raw Conc.) and Y-variable (Covid Cases & Covid Deaths) have a linear correlation among all subjects in the population.

As the AQI dataset contains hourly data whereas the Covid dataset contains daily data, two approaches will be used in the correlation analysis.

1. Correlating Mean of values per Day with Covid cases & deaths

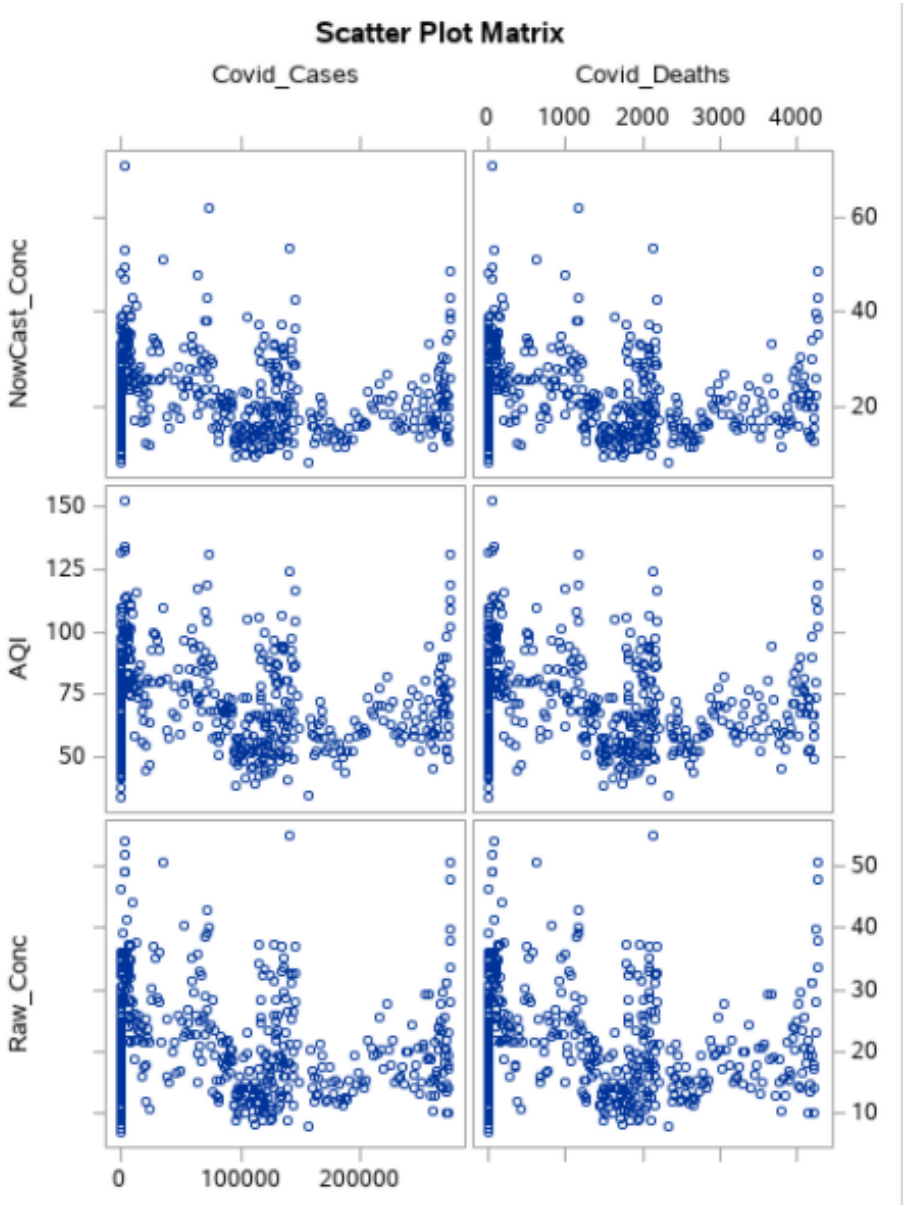
On taking the mean of AQI per day and merging with the Covid dataset, the following results are obtained.

Ethiopia:

Variable	Covid Cases		Covid Deaths	
	Pearson Value	P-Value	Pearson Value	P-Value
AQI	-0.17273	<.0001	-0.16629	0.0002
Nowcast. Conc.	-0.18334	<.0001	-0.17643	<.0001
Raw Conc.	-0.17273	<.0001	-0.17672	<.0001

On conducting Pearson correlation analysis on the AQI, Nowcast. Conc. & Raw Conc. of the cities in Ethiopia with the number of confirmed Covid cases & deaths, the following are observed:

- For Covid cases & Covid deaths, all three variables - AQI, Nowcast. Conc. & Raw Conc., have a p-value less than 0.05, which shows that it is probable to reject the null hypothesis.
- The correlation coefficient (R) is slightly greater than the decision point taken from the statistics table, for samples with n greater than 1000.
- The Correlation coefficient is negative for all cases, which signifies a negative linear relationship, as evident in the scatterplot graph shown below.
- To conclude, all variables are slightly negatively correlated with the Covid Cases and Covid Deaths.

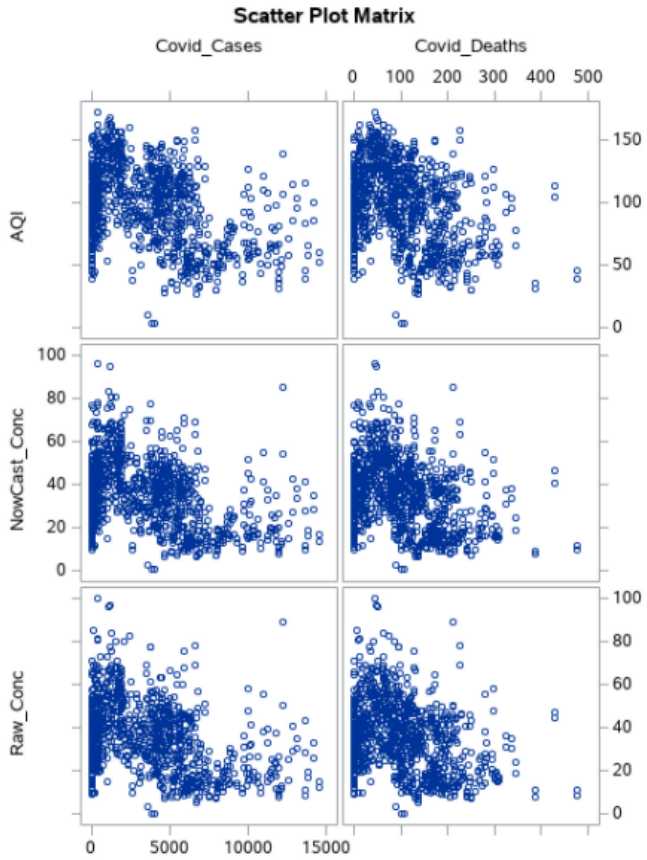


Indonesia:

Variable	Covid Cases		Covid Deaths	
	Pearson Value	P-Value	Pearson Value	P-Value
AQI	-0.41811	<.0001	-0.35000	0.0002
Nowcast. Conc.	-0.39208	<.0001	-0.32937	<.0001
Raw Conc.	-0.38347	<.0001	-0.32318	<.0001

On conducting Pearson correlation analysis on the AQI, Nowcast. Conc. & Raw Conc. of the cities in Indonesia with the number of confirmed Covid cases & deaths, the following are observed:

- For both Covid cases & Covid deaths, all three variables - AQI, Nowcast. Conc. & Raw Conc., have a p-value less than 0.05, which shows that it is probable to reject the null hypothesis.
- The correlation coefficient (R) is greater than the decision point taken from the statistics table, for samples of n greater than 1000.
- The Correlation coefficient is also negative for all cases, which signifies a negative linear relationship, as shown in the scatterplot graph below.
- To conclude, all variables are slightly negatively correlated with the Covid Cases and Covid Deaths.



2. Correlating by Using Last hour per day with Covid cases & Deaths

On taking the AQI of the last hour per day and merging with the Covid dataset, the following results are obtained.

Ethiopia:

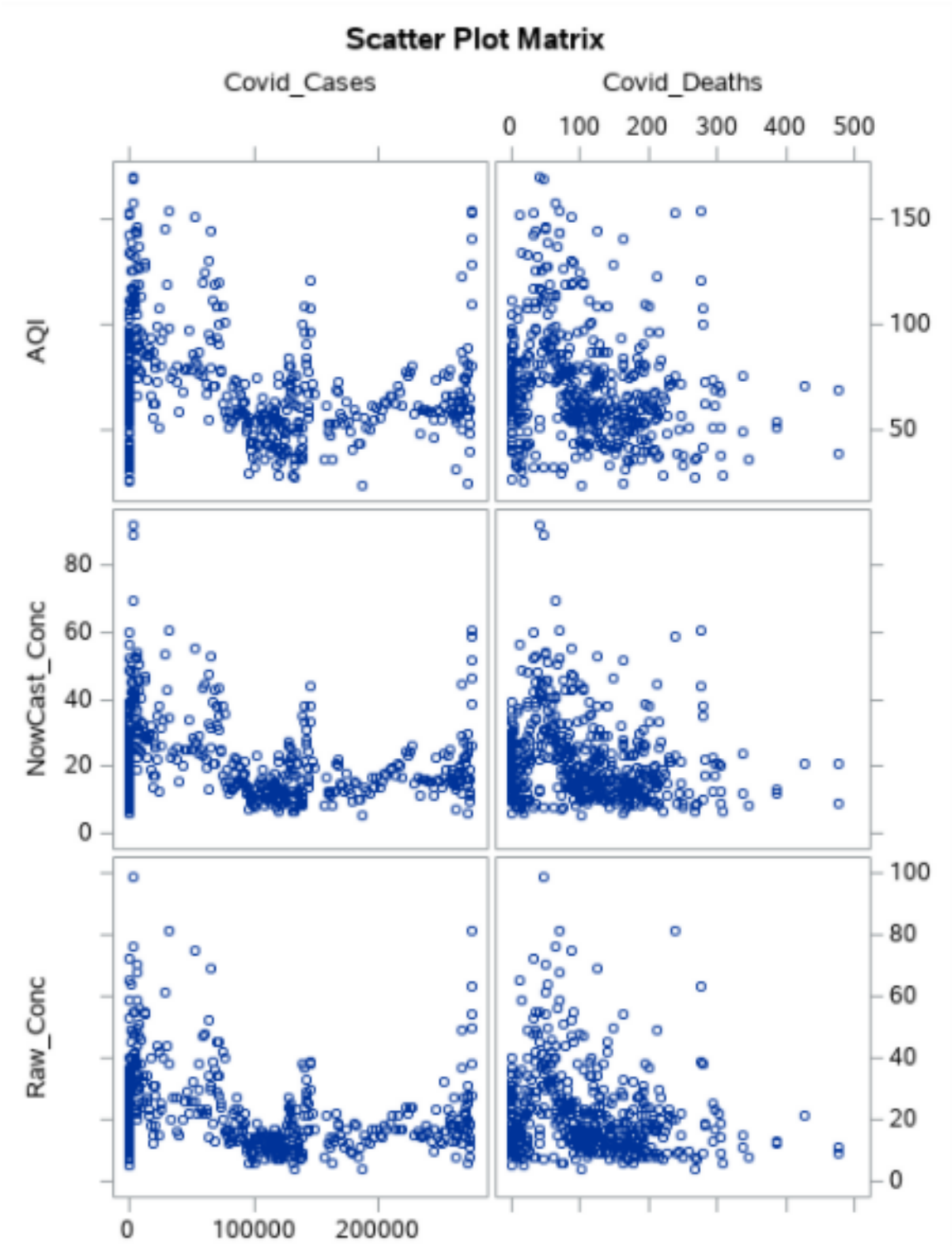
Variable	Covid Cases		Covid Deaths	
	Pearson Value	**P-Value**	**Pearson Value**	**P-Value**
AQI	-0.25134	<.0001	-0.21276	<.0001
Nowcast. Conc.	-0.26121	<.0001	-0.21119	<.0001
Raw Conc.	-0.24184	<.0001	-0.20901	<.0001
	Spearman Value	**P-Value**	**Spearman Value**	**P-Value**
Month	0.05489	0.2137	0.05128	0.2453

On conducting Pearson correlation analysis on the AQI, Nowcast. Conc. & Raw Conc. of the cities in Ethiopia with the number of confirmed Covid cases & deaths, the following are observed:

- For covid cases and deaths, all three variables AQI, Nowcast. Conc. & Raw Conc., have a p-value less than 0.05, which shows that it is probable to reject the null hypothesis.
- The correlation coefficient (R) is greater than the decision point taken from the statistics table, for samples with n greater than 1000.
- The Correlation coefficient is negative for all cases, which signifies a negative linear relationship, as evident in the scatterplot graph shown below.
- To conclude, AQI, Nowcast. Conc. & Raw Conc. are slightly negatively correlated with Covid Cases and deaths.

For the Spearman analysis, the following are observed:

- Month and Covid Cases have a p-value of more than 0.05, hence there is no significant correlation between them.
- Month and Covid Death also have a p-value of more than 0.05, hence there's no significant linear relationship between them.
- To conclude, Month has no linear correlation with Covid Cases or Deaths.


Scatter Plot Matrix
Covid_Cases
Covid_Deaths
0
100
200
300
400
500
AQI
150
100
50
NowCast_Conc
80
60
40
20
0
Raw_Conc
100
80
60
40
20
0
0
100000
200000

Indonesia:

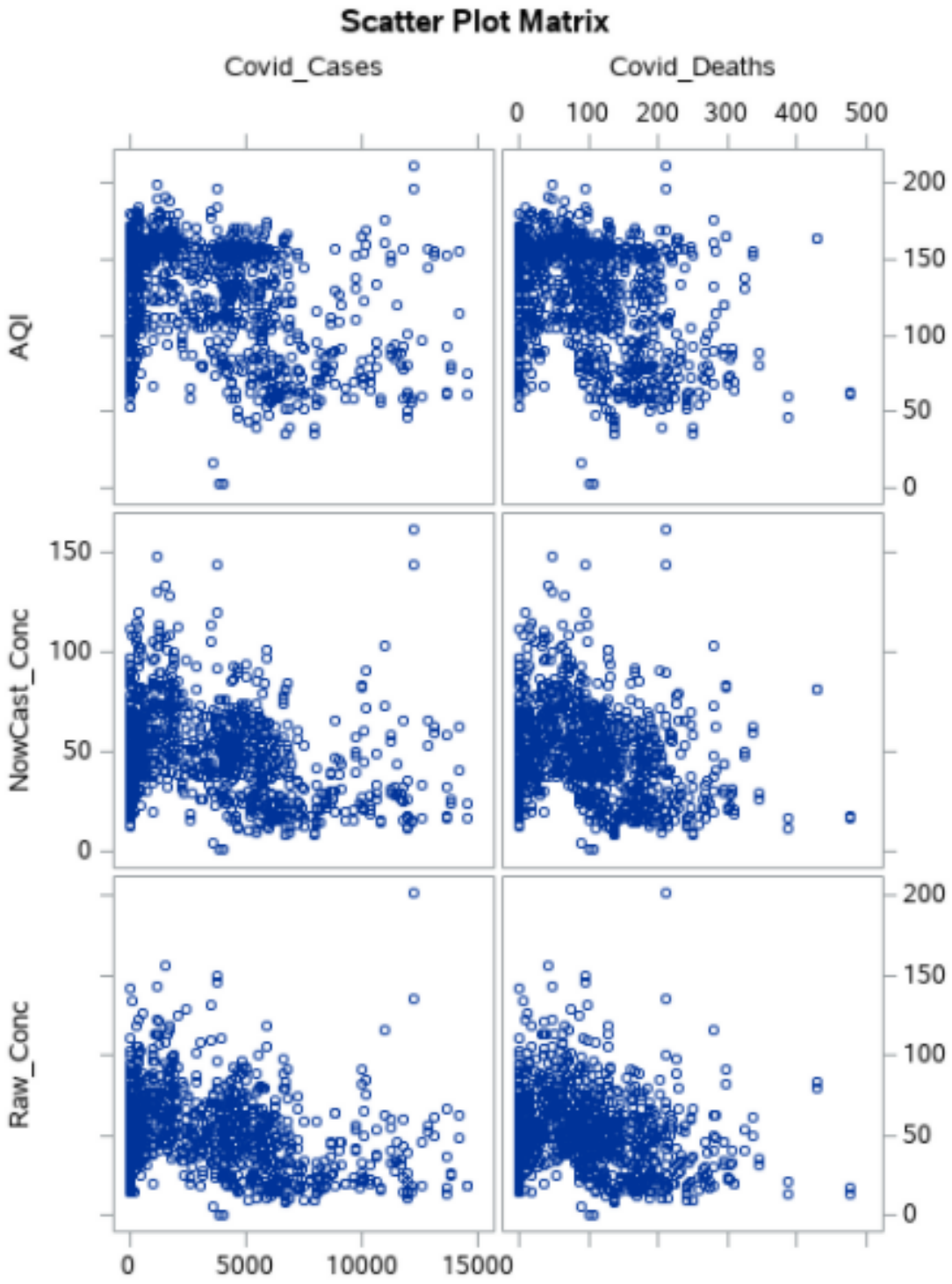
Variable	Covid Cases		Covid Deaths	
	Pearson Value	**P-Value**	**Pearson Value**	**P-Value**
AQI	-0.35431	<.0001	-0.30282	<.0001
Nowcast. Conc.	-0.30990	<.0001	-0.26854	<.0001
Raw Conc.	-0.29904	<.0001	-0.26554	<.0001
	Spearman Value	**P-Value**	**Spearman Value**	**P-Value**
Month	0.06363	0.0460	0.05566	0.0809

On conducting Pearson correlation analysis on the AQI, Nowcast. Conc. & Raw Conc. of the cities in Indonesia with the number of confirmed Covid cases & deaths, the following are observed:

- For both Covid cases & Covid deaths, all three variables AQI, Nowcast. Conc. & Raw Conc., have a p-value less than 0.05, which shows that it is probable to reject the null hypothesis.
- The correlation coefficient (R) is greater than the decision point taken from the statistics table for linear correlation, for samples with n greater than 1000.
- The Correlation coefficient is also negative for all cases, which signifies a negative linear relationship, as shown in the scatterplot graph below.
- To conclude, all three variables are slightly negatively correlated with Covid Cases and Deaths.

For the Spearman analysis, the following are observed: -

- Month and Covid Cases have a p-value of less than 0.05, hence there is a significant correlation between them.
- The Correlation coefficient is positive, which signifies a positive linear relationship.
- Month and Covid Death have a p-value of more than 0.05, hence there is no significant linear relationship between them.
- To conclude, Month is slightly positively correlated with Covid Cases but not with Covid Death.


Scatter Plot Matrix
Covid_Cases
Covid_Deaths
0
100
200
300
400
500
AQI
NowCast_Conc
Raw_Conc
0
50
100
150
200
0
5000
10000
15000

Linear Regression

Aim:

The analysis aims to test and see if by using AQI, NowCast_Conc, or Month, the number of Covid Cases & Covid Deaths can be predicted.

Approach:

For the Linear Regression Analysis, "AQI" & "Month" will be used as Predictor, and "Covid Cases" & "Covid Deaths" will be used as Response. The analysis will be separate for each country as the covid cases/deaths are different for each country.

i. **Null- Hypothesis (H_0)**: X-variable (AQI, Nowcast Conc. & Month) does not have an impact on prediction of Y-variable (Covid Cases & Covid Deaths).
ii. **Alternative Hypothesis (H_a)**: X-variable (AQI, Nowcast Conc. & Month) has an impact on prediction of Y-variable (Covid Cases & Covid Deaths).

Ethiopia:

The result of the Linear regression is shown below.

Predictor (Independent)	Response (Dependent)	Parameter Estimate	Standard Error	P-value	Variance P-Value	R-Square
AQI	Covid Cases	-824.33035	140.43076	<.0001	<.0001	0.0632
AQI	Covid Death	-0.66595	0.13530	<.0001	<.0001	0.0453
NowCast_Conc	Covid Cases	-1895.33858	309.84046	<.0001	<.0001	0.0682
NowCast_Conc	Covid Death	-1.46245	0.29943	<.0001	<.0001	0.0446
Month	Covid Cases	-782.43546	1163.08590	0.5014	0.5014	0.0009
Month	Covid Death	-0.84739	1.10986	0.4455	0.4455	0.0011

From the results, it is observed that:

- Only for AQI by Covid Death & NowCast_Conc by Covid Death is the model having a p-value of less than 0.05 for a variance, which means the variance can be explained by the models. The other models have values greater than 0.05, hence the model may not be able to explain the variance.
- The R-Square value for all the models is quite small, less than 1%, which may make the regression model invalid for some cases.

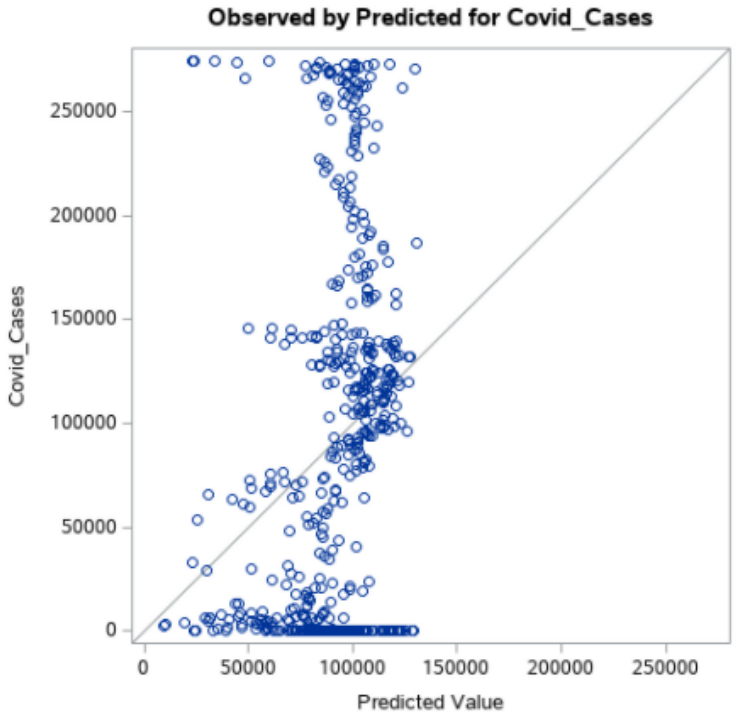
The following model further explains the analysis for AQI by Covid Cases and AQI by Covid Death models.

AQI by Covid Cases Model:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.575943E11	2.575943E11	34.46	<.0001
Error	511	3.820138E12	7475807516		
Corrected Total	512	4.077732E12			

Root MSE	86463	R-Square	0.0632
Dependent Mean	91454	Adj R-Sq	0.0613
Coeff Var	94.54283		

The p-value is quite small i.e. <0.05, hence the model does a good job of explaining the variability in the data.
R-Square value is 0.0632 which means that AQI can account for 0.06% of the variability in Covid cases, which is quite small, hence the regression may not be valid in some cases.



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	149640	10622	14.09	<.0001
AQI	1	-824.33035	140.43076	-5.87	<.0001

From the table above, it is observed that the p-value is less than 0.05, hence the Parameter Estimate is statistically significant. The parameter estimate for AQI is given as approximately -824, which means that an increase in AQI results in Covid cases decreased by 824.

From the graph of the Observed Value by Predicted value, it is observed that most data points do not lie on the line. Hence the model is not valid i.e., AQI may not be a valid predictor for Covid Cases.

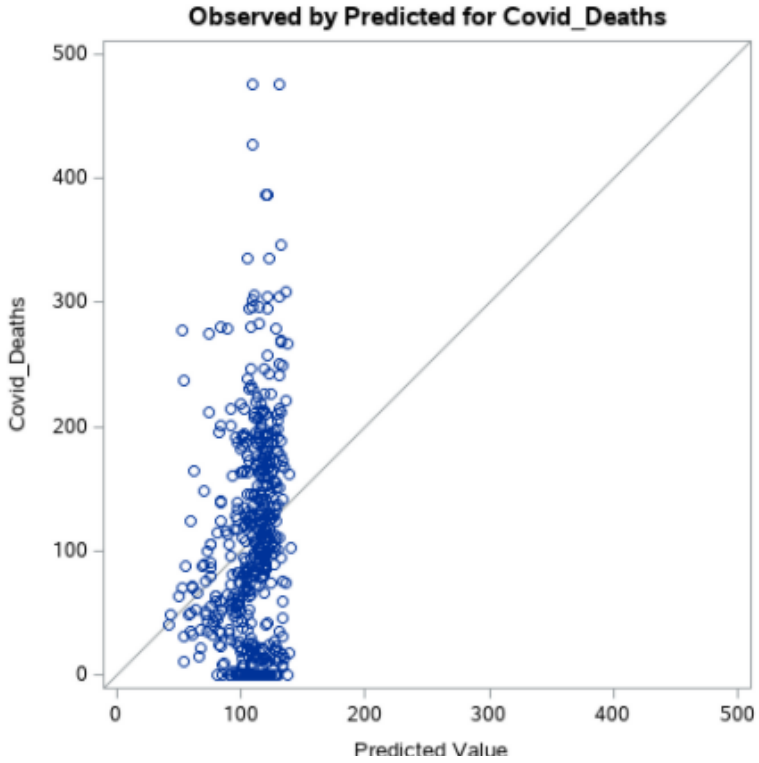
AQI by Covid Deaths Model:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	168118	168118	24.23	<.0001
Error	511	3545887	6939.11428		
Corrected Total	512	3714005			

Root MSE	83.30135	R-Square	0.0453
Dependent Mean	108.54191	Adj R-Sq	0.0434
Coeff Var	76.74579		

The p-value is greater than 0.05, hence the model may not be significant in explaining the variability in the data.
R-Square value is 0.0453 which means that AQI can account for 0.04% of the variability in Covid cases, which is quite small, hence the regression may not be valid in some cases.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	155.54900	10.23382	15.20	<.0001
AQI	1	-0.66595	0.13530	-4.92	<.0001



From the graph of the Observed Value by Predicted value, it is observed that the majority of data points do not lie on the line. Hence the model is not valid i.e., AQI may not be a valid predictor for Covid Death.

Indonesia:

Predictor	Response	Parameter Estimate	Standard Error	P-Value	Variance P-Value	R-Square
AQI	Covid Cases	-31.79963	2.68791	<.0001	<.0001	0.1255
AQI	Covid Death	-0.67749	0.06829	<.0001	<.0001	0.0917
NowCast_Conc	Covid Cases	-43.67689	4.29137	<.0001	<.0001	0.0960
NowCast_Conc	Covid Death	-0.94343	0.10838	<.0001	<.0001	0.0721
Month	Covid Cases	-48.78332	31.60496	0.1230	0.1230	0.0024
Month	Covid Death	-0.52474	0.7886	0.5060	0.5060	0.0005

From the results, it is observed that:

- The model AQI by Covid Cases and AQI by Covid Deaths & NowCast_Conc by Covid Cases and NowCast_Conc by Covid Deaths are having a p-value of less than 0.05 for variance, which means the model may be able to explain the variance. Whereas the other models have values greater than 0.05, hence the model may not be able to explain the variance.
- The R-Square values for all the models are quite small, less than 1%, which may make the regression model invalid for some cases.
- For Month by Covid Death, the R-Square value is 0, which does not explain any variance at all.

The following further explains the analysis for AQI by Covid Cases and AQI by Covid Death model.

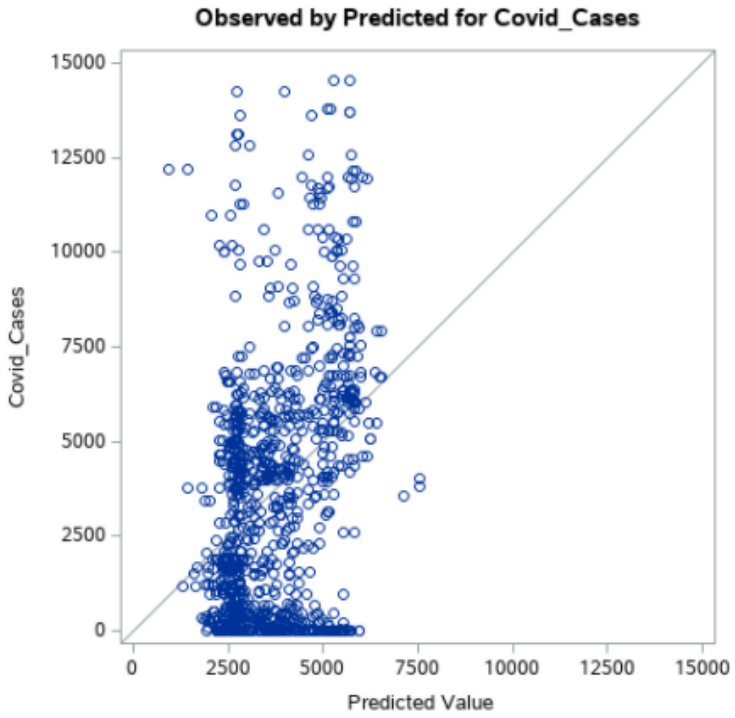
AQI by Covid Cases Model:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1333593262	1333593262	139.96	<.0001
Error	975	9289923430	9528127		
Corrected Total	976	10623516692			

Root MSE	3086.76637	R-Square	0.1255
Dependent Mean	3688.44831	Adj R-Sq	0.1246
Coeff Var	83.68740		

The p-value is less than 0.05, hence the model is significant in explaining the variability in the data. R-Square value is 0.1255 which means that AQI can account for 0.12% of the variability in Covid cases, which is quite small, hence the regression may not be valid in some cases.



From the graph of the Observed Value by Predicted value, it is observed that most data points do not lie on the line. Hence the model is not valid i.e., AQI may not be a valid predictor for Covid Cases.

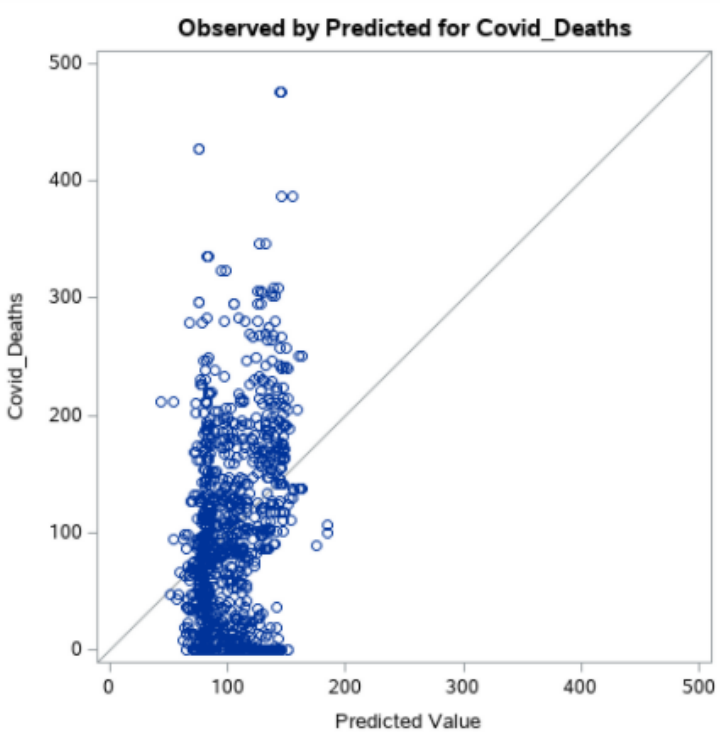
AQI by Covid Deaths Model:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	605320	605320	98.43	<.0001
Error	975	5995981	6149.72404		
Corrected Total	976	6601301			

Root MSE	78.42018	R-Square	0.0917
Dependent Mean	102.31013	Adj R-Sq	0.0908
Coeff Var	76.64947		

The p-value is quite small i.e. <0.05, hence the model does a good job of explaining the variability in the data.

R-Square value is 0.0917 which means that AQI can account for 0.09% of the variability in Covid deaths, which is quite small, hence the regression may not be valid in some cases.



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	186.79391	8.87737	21.04	<.0001
AQI	1	-0.67749	0.06829	-9.92	<.0001

From the table above, it is observed that the p-value is less than 0.05, hence the Parameter Estimate is statistically significant. The parameter estimate for AQI is given as -0.677, which means that an increase in AQI results in Covid deaths decreased by 0.677.

From the graph of the Observed Value by Predicted value, it is observed that most data points do not lie on the line. Hence the model is not valid i.e., AQI may not be a valid predictor for Covid Death.

T-Test Analysis

Aim:

According to AirNow.gov, an AQI value of below 100 is considered satisfactory. The test aims to see if the mean AQI of each city is greater than 90.[7]

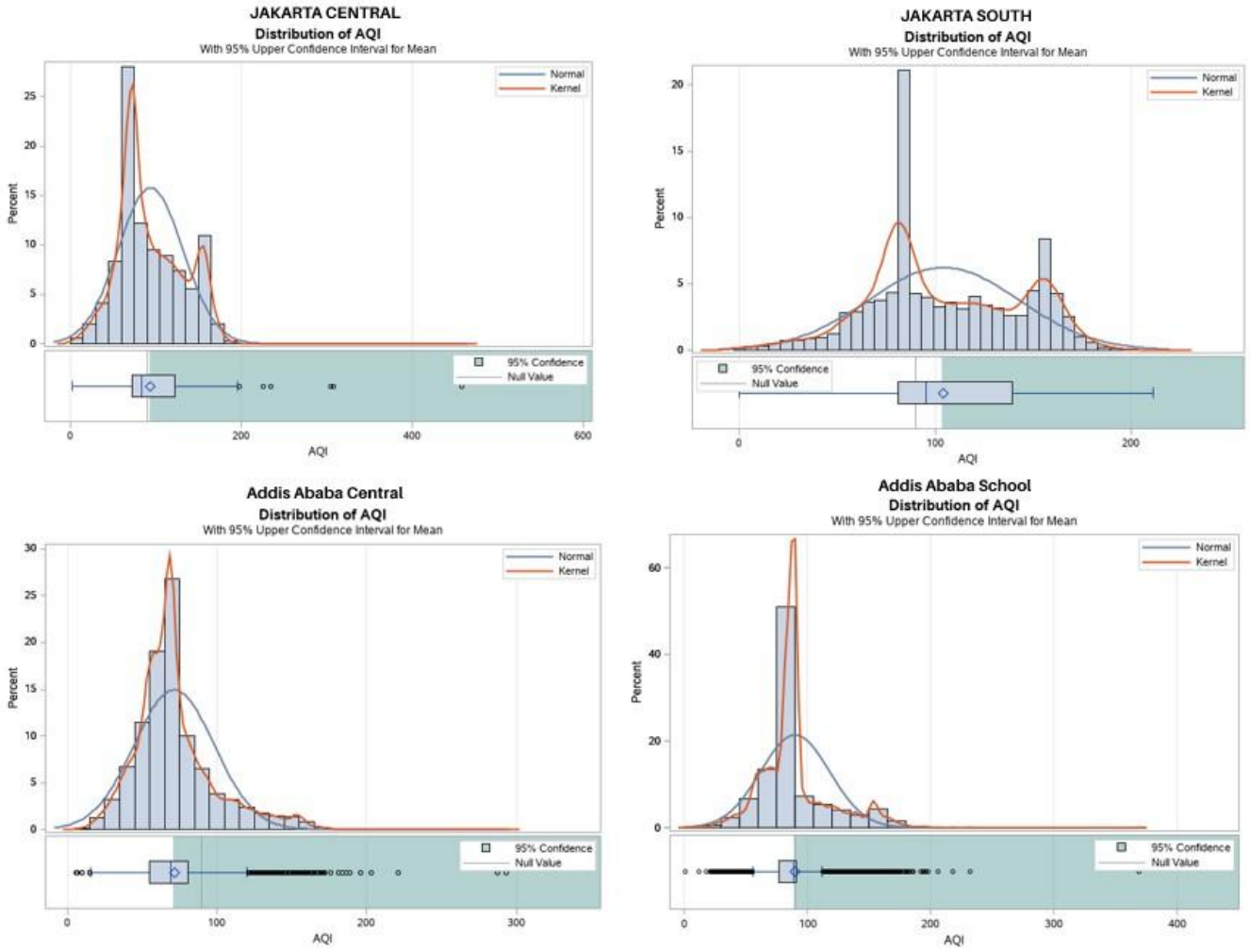
Approach:

T-Test will be used for testing the hypothesis. The following parameters will be used:

i. **Null- Hypothesis (H_0)**: The mean AQI of each city is equal to 90

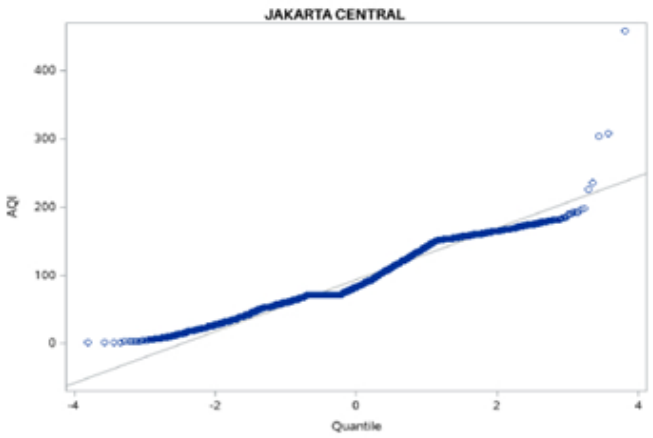
ii. **Alternative Hypothesis (H_a)**: The mean AQI of each city is greater than 90

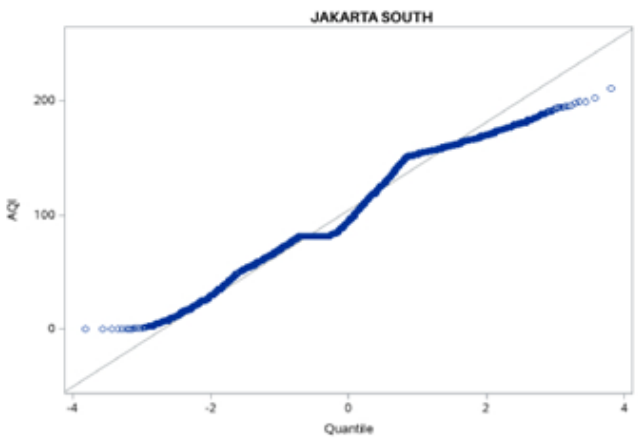
Before conducting the T-Test analysis, a test of normality is performed to ensure that the data is normally distributed.

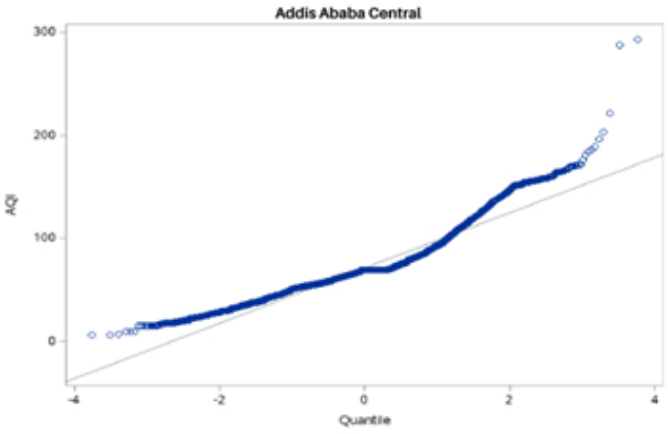


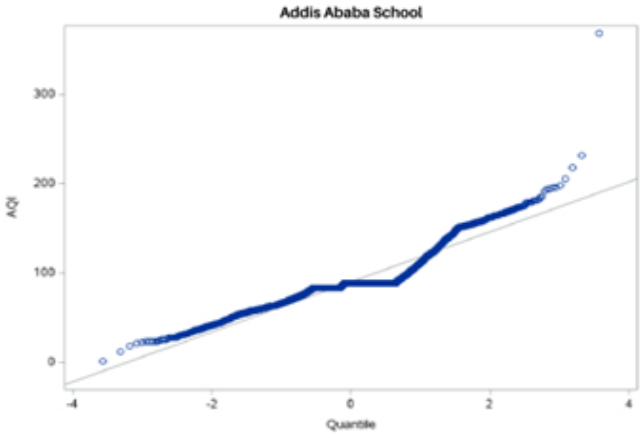
[7] airnow.gov. n.d. "*Air Quality Index (AQI) Basics*." AirNow. Accessed August 2, 2021. https://www.airnow.gov/aqi/aqi-basics/.

The graphs above show a “Bell-shaped” curve, which signifies normal distribution for the data for each of the cities. The points for the Q-Q plot also mostly lie on the straight line to show normal distribution as shown below.









JAKARTA CENTRAL

Variable: AQI

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.131192	Pr > D	<0.0100
Cramer-von Mises	W-Sq	34.32404	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	183.4443	Pr > A-Sq	<0.0050

JAKARTA SOUTH

Variable: AQI

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.129285	Pr > D	<0.0100
Cramer-von Mises	W-Sq	30.9085	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	169.5627	Pr > A-Sq	<0.0050

ABABA CENTRAL

Variable: AQI

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.162369	Pr > D	<0.0100
Cramer-von Mises	W-Sq	38.15114	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	203.1452	Pr > A-Sq	<0.0050

ABABA SCHOOL

Variable: AQI

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.255169	Pr > D	<0.0100
Cramer-von Mises	W-Sq	34.82001	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	168.0899	Pr > A-Sq	<0.0050

Further verification can be done by observing the results of tests for normality. Since the p-value of all the normality tests is less than 0.05, it can be concluded that the data for each city has a normal distribution.

The result and conclusion of the t-test analysis is shown as below:

City	DF	t-value	P-value	Conclusion
Jakarta Central	9090	9.41	<.0001	H_0 rejected; Mean of AQI is greater than 90
Jakarta South	9131	35.24	<.0001	H_0 rejected; Mean of AQI is greater than 90
Addis Ababa Central	7378	-59.77	1.0000	Not enough evidence to accept or reject H_0
Addis Ababa School	3561	0.05	0.4807	H_0 accepted; Mean of AQI is equal to 90

Chi-Square Analysis

AQI Category by Covid Cases

Aim:

The aim of the test is to check whether AQI Category & Covid Cases are associated with each country.

Approach:

Chi-Square test will be conducted for the hypothesis. The following parameters are used for the test: -

i. **Null Hypothesis (H_0):** AQI Category & Covid Cases are independent (i.e. they are not associated)
ii. **Alternative Hypothesis (H_a):** AQI Category & Covid Cases are dependent (i.e. they are associated)

Since the "Covid Cases" variable is numerical, it will be converted to a categorical variable by binning it into three equal bins consisting of "low", "medium" & "high" values.

Ethiopia:

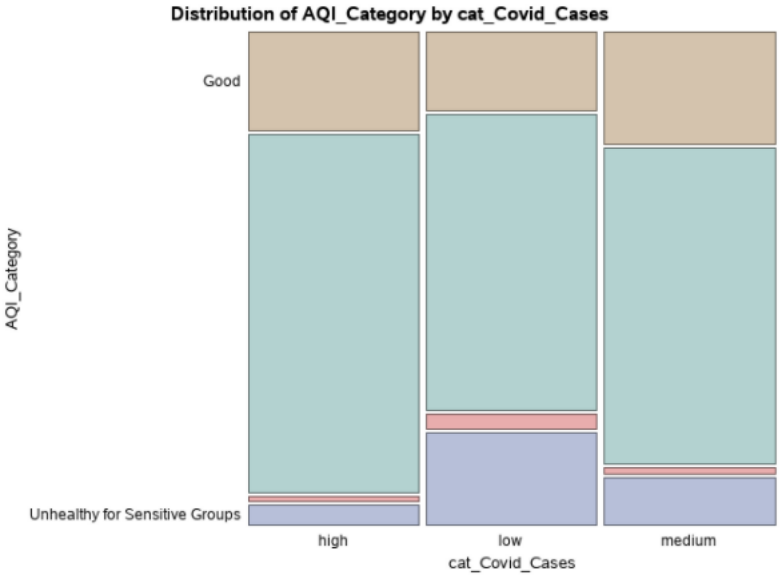


Table of AQI_Category by cat_Covid_Cases

AQI_Category	cat_Covid_Cases			
	high	low	medium	Total
Good	35	28	40	103
Moderate	126	105	113	344
Unhealthy	2	5	2	9
Unhealthy for Sensitive Groups	7	33	17	57
Total	170	171	172	513

The above table shows the Contingency table of AQI Category vs. Cat_Covid_Cases. Some cell values have frequencies of less than 5, which may affect the validity of the Chi-Square test.

Statistics for Table of AQI_Category by cat_Covid_Cases

Statistic	DF	Value	Prob
Chi-Square	6	24.2138	0.0005
Likelihood Ratio Chi-Square	6	24.6246	0.0004
Mantel-Haenszel Chi-Square	1	0.9827	0.3215
Phi Coefficient		0.2173	
Contingency Coefficient		0.2123	
Cramer's V		0.1536	

WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 513

df	dec.pt.	df	dec. pt.	df	dec. pt.
1	3.84	14	23.68	26	38.89
2	5.99	15	25.00	27	40.11
3	7.81	16	26.30	28	41.34
4	9.49	17	27.59	29	42.56
5	11.07	18	28.87	30	43.77
6	12.59	19	30.14	40	55.80
7	14.07	20	31.41	50	67.50
8	15.51	21	32.67	60	79.10
9	16.92	22	33.92	70	90.50
10	18.30	23	35.17	80	102.00
11	19.68	24	36.42	90	113.00
12	21.03	25	37.65	100	124.30
13	22.36				

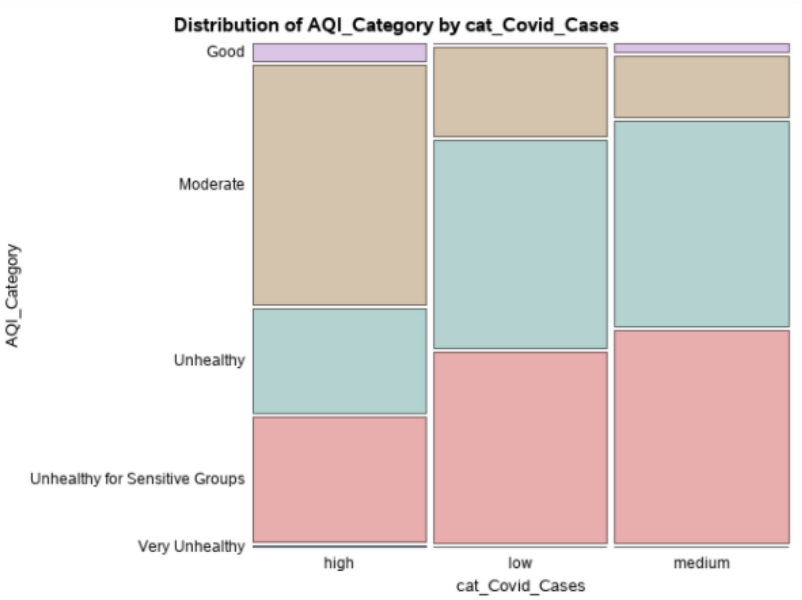
The results of the test are as follows:

- Taking df = 6 from the above table, the decision point obtained is equal to 12.59. Finally, it is verified that 12.59 is minor than the X^2 value of 24.21.
- The p-value is < 0.0005, which is less than 0.05. Hence the Null hypothesis could be rejected.
- Having p-value < 0.05 and X^2 > dp-value, it could be concluded that AQI Category & Covid Cases are dependent on the country of Ethiopia.
- However, since there are cells containing frequencies of less than 5, the test may not be valid and hence no conclusions can be drawn.

Indonesia:

Table of AQI_Category by cat_Covid_Cases

AQI_Category	cat_Covid_Cases			
	high	low	medium	Total
Good	12	0	6	18
Moderate	160	59	41	260
Unhealthy	70	138	137	345
Unhealthy for Sensitive Groups	83	127	143	353
Very Unhealthy	1	0	0	1
Total	326	324	327	977



The above table shows the Contingency table of AQI Category vs. Cat_Covid_Cases. The cell values for the "Very Unhealthy" & "Good" row contain cells with a frequency below 5, which may make the Chi-Square test invalid.

Statistics for Table of AQI_Category by cat_Covid_Cases

Statistic	DF	Value	Prob
Chi-Square	8	151.6243	<.0001
Likelihood Ratio Chi-Square	8	154.2617	<.0001
Mantel-Haenszel Chi-Square	1	77.8423	<.0001
Phi Coefficient		0.3939	
Contingency Coefficient		0.3665	
Cramer's V		0.2786	

Sample Size = 977

df	dec.pt.	df	dec. pt.	df	dec. pt.
1	3.84	14	23.68	26	38.89
2	5.99	15	25.00	27	40.11
3	7.81	16	26.30	28	41.34
4	9.49	17	27.59	29	42.56
5	11.07	18	28.87	30	43.77
6	12.59	19	30.14	40	55.80
7	14.07	20	31.41	50	67.50
8	15.51	21	32.67	60	79.10
9	16.92	22	33.92	70	90.50
10	18.30	23	35.17	80	102.00
11	19.68	24	36.42	90	113.00
12	21.03	25	37.65	100	124.30
13	22.36				

The results of the test are as follows:

- Taking df = 8 from the above table, the decision point obtained is 15.51. Finally, it is verified that 15.51 is smaller than X^2 (151.6).
- The p-value is < 0.0001, which is less than 0.05. Hence the Null hypothesis could be rejected.
- Having p-value < 0.05 and X^2 > dp-value, it could be concluded that AQI Category & Covid Cases are dependent on the country of Indonesia, and hence there is a significant association between the two variables.

AQI Category by Site

Aim:

The test aims to check whether AQI Category & Site (City) are associated by testing the hypothesis on a merged dataset comprising all 4 cities. This analysis will help in investigating whether the location of a place determines the AQI Category.

Approach:

Chi-Square test will be conducted for the hypothesis. The following parameters are used for the test: -

i. **Null Hypothesis (H_0):** AQI Category & Site are independent (i.e. they are not associated)
ii. **Alternative Hypothesis (H_a):** AQI Category & Site are dependent (i.e. they are associated)

Table of AQI_Category by Site

AQI_Category	Site				
	Addis Ababa Central	Addis Ababa School	Jakarta Central	Jakarta South	Total
Good	1183	151	767	520	2621
Moderate	5286	2691	4926	4367	17270
Unhealthy for Sensitive Groups	760	495	2208	2401	5864
Unhealthy	146	221	1185	1842	3394
Very Unhealthy	4	3	2	2	11
Hazardous	0	1	3	0	4
Total	7379	3562	9091	9132	29164

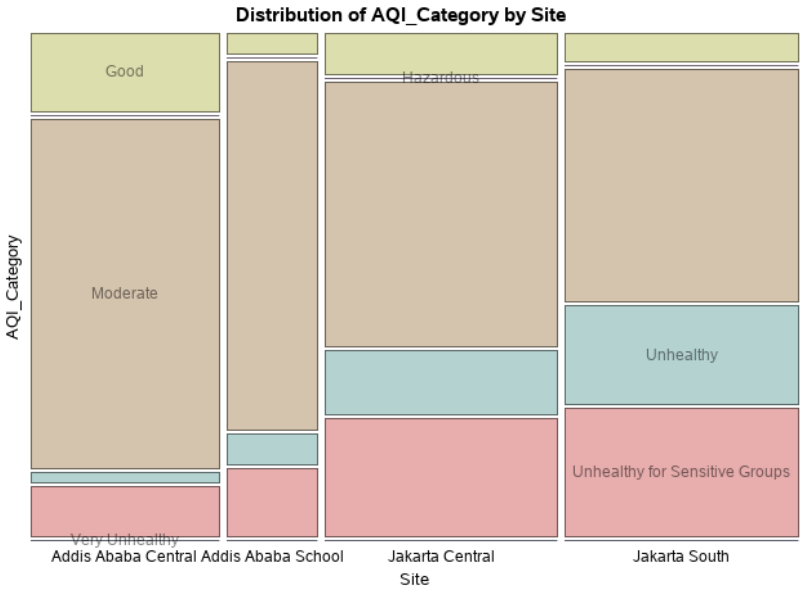
The above table shows the Contingency table of AQI Category vs. Site. The cell values for "Very Unhealthy" & "Hazardous" rows are below 5, which might make the Chi-Square test invalid.

Statistics for Table of AQI_Category by Site

Statistic	DF	Value	Prob
Chi-Square	15	3153.5032	<.0001
Likelihood Ratio Chi-Square	15	3402.1118	<.0001
Mantel-Haenszel Chi-Square	1	1839.0796	<.0001
Phi Coefficient		0.3288	
Contingency Coefficient		0.3124	
Cramer's V		0.1899	

WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 29164



The results of the test are as follows:

- The p-value of the Chi-Square test is <0.0001, which is less than 0.05, hence the Null hypothesis can be rejected.
- Therefore, it can be concluded that AQI Category & Site are dependent, and hence there is a significant association between the two variables.
- However, as observed earlier, there are 33% cell values with expected counts of less than 5, which invalidates the Chi-Square Test.
- Hence no valid conclusions can be obtained.

Conclusion

For Jakarta, 2019 was the most polluted year as the AQI Category with values above the acceptable ranges were observed the most in 2019. June-September were the most affected months, which could be explained due to the dry season and forest fire. The air quality in Indonesia also depicted an upward trend from the year 2017-19, which could be due to an increase in using coal for electricity generation, as well as gasoline and diesel consumption from the year 2016 onwards.[8]

There was also a lack of PM2.5 standards imposed until the year 2017, when the government mandated the Euro-4 fuel standards to be adopted by gasoline-fueled vehicles by September 2018. The regulations on coal plant emissions are also less strict compared to other countries, due to which even with the onset of Covid-19 in 2020, the average AQI continues to be above satisfactory levels.[9] With COVID-19 having brought large portions of the city (and world) to a standstill, one would expect the pollution levels to drop, but instead, they have been rising consistently despite lesser instances of international and domestic tourism. This is blamed largely on the previously mentioned coal-based power plants and factories.

For Addis Ababa Central, 2018 was the most polluted year with AQI categories in the unhealthy range having the highest frequencies. The month of June was the worst in terms of air quality for the sensitive group. Addis Ababa School had limited data with no data available for the year 2019, and very little data available from 2020 onwards. Hence the analysis could only be performed for 2016-18, where 2017 was the worst year with the highest frequencies of category in Unhealthy category. Ethiopia's air quality mostly lies in the Unhealthy for sensitive group category.

The higher air quality level could be explained as a large percentage of Ethiopia's vehicles are more than 15 years old and hence are not as efficient at reducing harmful emissions as compared to the newer vehicles.[10] There was also missing data for the year 2019, which is due to the lack of sufficient real-time permanent monitoring stations needed to get accurate results.[11]

Inferential analysis was conducted on the dataset using various tests. As the Air Quality dataset contained data on an hourly basis and the Covid19 dataset contained data on daily basis, it was difficult to merge the two datasets accurately.

---

[8] Michael Greenstone and Qing (Claire) Fan. 2019. "*Indonesia's Worsening Air Quality and its Impact on Life Expectancy*." AQLI. https://aqli.epic.uchicago.edu/wp-content/uploads/2019/03/Indonesia-Report.pdf.

[9] iqair.com. n.d. "*Air quality in Jakarta*." IQAir. Accessed August 2, 2021. https://www.iqair.com/indonesia/jakarta.

[10] iqair.com. n.d. "*Air quality in Ethiopia.*" IQAir. Accessed August 2, 2021. https://www.iqair.com/us/ethiopia.

[11] iqair.com. n.d. "*Air quality in Addis Ababa*." IQAir. Accessed August 2, 2021. https://www.iqair.com/ethiopia/addis-ababa.

As covid cases and deaths are mostly calculated at the end of the day, the datasets were merged by considering only the last hour of the day for the Air Quality dataset. The findings of the tests are as follows:

1. A negative correlation was observed between AQI and the number of Covid cases/deaths. In Indonesia, this could be explained due to the less stringent regulations on coal emissions for power consumption, hence why the AQI was still high even though the Covid cases/deaths decreased due to lockdown.
   In Ethiopia, a negative correlation was also observed for covid cases, but not for covid deaths. However, the dataset was insufficient with missing values for 2019 and incomplete values for 2020, hence we cannot make any valid conclusions here.

2. The tests for linear regressions were also not valid, as Air quality may not be the only factor affecting covid cases/deaths, and hence the models could not be used to give accurate predictions.

3. Using T-test analysis, the mean AQI for each city could be accurately tested against a benchmark value of 90 where Jakarta Central & Jakarta South had mean AQI over the benchmark value, Addis Ababa School was equal to benchmark and no conclusion could be obtained for Addis Ababa Central.

4. Dependency between AQI Category and covid cases could not be concluded due to insufficient data for Ethiopia, whereas for Indonesia it was found that there is a dependency between Covid cases and AQI category.

Due to differences in the way the dataset was presented (hourly vs daily dataset), there could be some biases in the test results. Some limitations observed with the tests:

- Correlation assumes a linear relationship between the variables, which can limit the interpretation. Its interpretation could not be accounted for causality as well. As observed, Air quality and covid cases/deaths showed a negative correlation, which meant that as air quality improved the covid cases went up, which ideally doesn't make sense. But after more in-depth research, it was concluded that other factors contributed to such a result.

- Similarly with linear regression, as it assumes a linear relationship it couldn't be applied accurately to the dataset.

Recommendations

Indonesia:

- Using fossil fuels for power generation leads to higher air pollution In Indonesia, hence reducing dependency on fossil fuels would help.
- Shifting to cleaner sources of energy could also help in improving air quality.
- As forest fires are also one of the main causes, the Government can implement policies to reduce deforestation and implement alternative methods of farming.

Ethiopia:

- Biomass fuel, wood, and charcoal are widely used as fuel in Ethiopia.[12] Switching to cleaner fuels would help in improving the air quality.
- Steps could be taken to reduce harmful emissions from vehicles or to switch to vehicles using cleaner fuels as well.
- More air monitoring systems could be established to help better understand the air quality of a region, and high levels of air pollution can be mitigated.

[12] iqair.com. n.d. "*Air quality in Ethiopia.*" IQAir. Accessed August 2, 2021. https://www.iqair.com/us/ethiopia.

Bibliography

airnow.gov. n.d. “Air Quality Index (AQI) Basics.” AirNow. Accessed August 2, 2021. https://www.airnow.gov/aqi/aqi-basics/.

ASAP East Africa. 2019. “Air Quality Briefing Note: Addis Ababa (Ethiopia).” asap-eastafrica.com. https://assets.publishing.service.gov.uk/media/5eb16f10e90e0723bd470fdf/ASAP_-_East_Africa_-_Air_Quailty_Briefing_Note_-_Addis_Abada.pdf.

asap-eastafrica.com. 2019. “Air Quality Briefing Note: Addis Ababa (Ethiopia).” ASAP East Africa. https://assets.publishing.service.gov.uk/media/5eb16f10e90e0723bd470fdf/ASAP_-_East_Africa_-_Air_Quailty_Briefing_Note_-_Addis_Abada.pdf.

“Average Weather in Addis Ababa.” n.d. weatherspark.com. https://weatherspark.com/y/100668/Average-Weather-in-Addis-Ababa-Ethiopia-Year-Round.

Greenstone, Michael, and Qing (Claire) Fan. 2019. “Indonesia’s Worsening Air Quality and its Impact on Life Expectancy.” AQLI. https://aqli.epic.uchicago.edu/wp-content/uploads/2019/03/Indonesia-Report.pdf.

iamat.org. 2020. “Indonesia General Health Risks: Air Pollution.” IAMAT. https://www.iamat.org/country/indonesia/risk/air-pollution#.

iqair.com. n.d. “Air quality in Addis Ababa.” IQAir. Accessed August 2, 2021. https://www.iqair.com/ethiopia/addis-ababa.

iqair.com. n.d. “Air quality in Ethiopia.” IQAir. Accessed August 2, 2021. https://www.iqair.com/us/ethiopia.

iqair.com. n.d. “Air quality in Jakarta.” IQAir. Accessed August 2, 2021. https://www.iqair.com/indonesia/jakarta.

News Desk (The Jakarta Post). 2020. “Jakarta air pollution causes 5.5 million illnesses yearly: Anies Jakarta Governor Anies Baswedan has said that air pollution is a significant health issue in the capital, causing more than 5 million illnesses a year. (Tribunews/handout) Share this article.” The Jakarta Post. https://www.thejakartapost.com/news/2020/09/23/jakarta-air-pollution-causes-5-5-million-illnesses-yearly-anies.html.

Nicholas, Hans. 2020. “Lockdown should have cleared up Jakarta’s air. Coal plants kept it dirty.” Mongabay. https://news.mongabay.com/2020/08/jakarta-air-pollution-coal-power-plant-covid-lockdown-crea-study/.

Syakriah, Ardila. 2020. “Pollution kills more than 230,000 Indonesians per year.” The Jakarta Post. https://www.thejakartapost.com/news/2019/12/30/pollution-kills-more-than-230000-indonesians-per-year-report.html.

The University of Birmingham. n.d. “How can we take the learnings of COVID-19 lockdown and turn them into a brighter environmental future for millions across the Global South?” www.birmingham.ac.uk. Accessed July 31, 2021. https://www.birmingham.ac.uk/research/quest/sustainable-environments/covid-19-and-air-pollution.aspx.

The U.S. Embassy and Consulates in Indonesia. n.d. “U.S Embassy Jakarta Air Quality Monitor.” id.usembassy.gov. https://id.usembassy.gov/embassy-consulates/airqualitymonitor/.

Walton, Kate. 2019. “Jakarta’s air quality kills its residents – and it’s getting worse.” theinterpreter. https://www.lowyinstitute.org/the-interpreter/jakarta-s-air-quality-kills-its-residents-and-it-s-getting-worse.

Westcott, Ben, and Tia Asmara. 2019. “Angry citizens sue Indonesian government over growing air pollution.” edition.cnn.com. https://edition.cnn.com/2019/07/02/health/jakarta-pollution-law-suit-intl-hnk/index.html.

APPENDIX

```
/**concatonate all countries **/


FILENAME JCFinal "/home/u58581363/MyPractice/Data/JAKARTACENTRAL.csv" TERMSTR=CRLF;

options validvarname=v7;
proc import datafile=JCFinal
                                    out=work.JakartaCentral
                                    dbms=CSV
                                    REPLACE;
                                    guessingrows=max;
run;


FILENAME REFFILE '/home/u58581363/MyPractice/Data/JSMAIN.csv';
options validvarname=v7;
PROC IMPORT DATAFILE=REFFILE
            DBMS=CSV
            OUT=WORK.JakartaSouth;
            GETNAMES=YES;
            guessingrows=max;
RUN;

FILENAME central "/home/u58581363/MyPractice/Data/ABABACENTRAL.csv" TERMSTR=CRLF;

options validvarname=v7;
proc import datafile=central
                                    out=work.AddisAbabaCentral
                                    dbms=CSV
                                    REPLACE;
                                    guessingrows=max;
run;

FILENAME  Abschool "/home/u58581363/MyPractice/Data/ABABASCHOOL.csv" TERMSTR=CRLF;

options validvarname=v7;
proc import datafile=Abschool
                                    out=work.AddisAbabaSchool
                                    dbms=CSV
                                    REPLACE;
                                    guessingrows=max;
run;


data ALL_MAIN;
            retain  Site Parameter Date__LT_ Year Month Day Hour NowCast_Conc_ AQI AQI_Category Raw_Conc_
Conc__unit
Duration QC_Name;
            length AQI_Category $30;
            format AQi_Category $30.;
            length Site $20;
            format Site $20.;
            set  JakartaCentral JakartaSouth  AddisAbabaSchool AddisAbabaCentral;

run;

/*select only the last two year*/
proc sql;
create table AllCovTemp as
```

```
select Site ,Parameter, Date__LT_, Year, Month, Day, Hour, NowCast_Conc_ ,AQI, AQI_Category ,Raw_Conc_, Conc__unit,
Duration ,QC_Name
from ALL_MAIN
where year in (2020,2021);
quit;
```

```
FILENAME Covcases "/home/u58581363/MyPractice/Data/time_series_covid19_confirmed_global_new.csv"
TERMSTR=CRLF;
FILENAME Covd "/home/u58581363/MyPractice/Data/time_series_covid19_deaths_global_new.csv" TERMSTR=CRLF;

options validvarname=v7;
proc import datafile=Covcases
                                    out=work.Cov_cases_temp
                                    dbms=CSV
                                    REPLACE;
                                    guessingrows=max;
run;

options validvarname=v7;
proc import datafile=Covd
                                    out=work.Cov_deaths_temp
                                    dbms=CSV
                                    REPLACE;
                                    guessingrows=max;
run;

/******** modify cov_cases dataset by transposing and selecting only the required country ***********/
/**covid cases indonesia***/

data Cov_cases_indo;
set Cov_cases_temp;
where  (country_region='Indonesia') ;
run;


proc transpose data=cov_cases_indo
out=work.Cov_cases_transpose_temp
name=dates
prefix=covid_cases;
run;

data Cov_cases_transpose;
set Cov_cases_transpose_temp;
if _n_ < 3 then delete;
run;

/**split date into year month and day**/
data cov_cases_split;
set cov_cases_transpose;
year=scan(dates,3,'_');
month=scan(dates,1,'_');
day=scan(dates,2,'_');
run;
```

```
data cov_cases_final;
set cov_cases_split;
ch_year=input(year, 4.);
ch_month=input(month, 4.);
ch_day=input(day, 4.);
drop year month day;
rename ch_year=Year ch_month=Month ch_day=Day;
run;
/**covid cases ethopia***/

data Cov_cases_etho;
set Cov_cases_temp;
where  (country_region='Ethiopia') ;
run;


proc transpose data=cov_cases_etho
out=work.Cov_cases_transpose_ethotemp
name=dates
prefix=covid_cases;
run;

data Cov_cases_ethotranspose;
set Cov_cases_transpose_ethotemp;
if _n_ < 3 then delete;
run;

/**split date into year month and day**/
data cov_cases_ethosplit;
set cov_cases_ethotranspose;
year=scan(dates,3,'_');
month=scan(dates,1,'_');
day=scan(dates,2,'_');
run;


/** convert to numeric values**/
data cov_cases_ethofinal;
set cov_cases_ethosplit;
ch_year=input(year, 4.);
ch_month=input(month, 4.);
ch_day=input(day, 4.);
drop year month day;
rename ch_year=Year ch_month=Month ch_day=Day;
run;

/**** covid deaths ***/
data Cov_deaths_indo;
set Cov_deaths_temp;
where  (country_region='Indonesia') ;
run;


proc transpose data=cov_deaths_indo
out=work.Cov_deaths_transpose_temp
name=dates
prefix=covid_deaths;
run;

data Cov_deaths_transpose;
set Cov_deaths_transpose_temp;
```

```
if _n_ < 3 then delete;
run;

/**split date into year month and day**/
data cov_deaths_split;
set cov_deaths_transpose;
year=scan(dates,3,'_');
month=scan(dates,1,'_');
day=scan(dates,2,'_');
run;

/** convert to numeric values**/
data cov_deaths_final;
set cov_deaths_split;
ch_year=input(year, 4.);
ch_month=input(month, 4.);
ch_day=input(day, 4.);
drop year month day;
rename ch_year=Year ch_month=Month ch_day=Day;
run;

/**** covid deaths ethopia***/
data Cov_deaths_etho;
set Cov_deaths_temp;
where  (country_region='Ethiopia') ;
run;


proc transpose data=cov_deaths_etho
out=work.Cov_deaths_transpose_ethotemp
name=dates
prefix=covid_deaths;
run;

data Cov_deaths_ethotranspose;
set Cov_deaths_transpose_ethotemp;
if _n_ < 3 then delete;
run;

/**split date into year month and day**/
data cov_deaths_ethosplit;
set cov_deaths_ethotranspose;
year=scan(dates,3,'_');
month=scan(dates,1,'_');
day=scan(dates,2,'_');
run;

/** convert to numeric values**/
data cov_deaths_ethofinal;
set cov_deaths_ethosplit;
ch_year=input(year, 4.);
ch_month=input(month, 4.);
ch_day=input(day, 4.);
drop year month day;
rename ch_year=Year ch_month=Month ch_day=Day;
run;



/*******************************************/
```

```
/** create new dataset taking average aqi of all days **/
proc sql;
create table AllCov_mean as
select Site, parameter, year, month, day, hour, avg(NowCast_Conc_) as NowCast_Conc, avg(aqi) as AQI,
AQI_Category, avg(Raw_Conc_) as Raw_Conc, Conc__unit, Duration, QC_Name
from AllCovTemp/*used the main dataset*/
group by Site, year, month, day;
quit;

/** re-evaluate AQI categories **/
data AllCov_mean_cat;
set AllCov_mean;
        if 0<aqi<51 then AQI_Category='Good';
        if 51<aqi<101 then AQI_Category='Moderate';
        if 101<aqi<151 then AQI_Category='Unhealthy for Sensitive Groups';
        if 151<aqi<201 then AQI_Category='Unhealthy';
        if 201<aqi<301 then AQI_Category='Very Unhealthy';
        if 301<aqi then AQI_Category='Hazardous';
run;


proc sql;
create table AllCov_mean_final as
select distinct year, month, day, Site, parameter, aqi, NowCast_Conc,AQI_Category, Raw_Conc, Conc__unit, Duration,
QC_Name
from AllCov_mean_cat;
quit;



/*** join datasets for mean aqi **/
proc sql;
create table AllCov_mean_main as
select All.Site, All.parameter,All.year, All.month,  All.day, All. NowCast_Conc, All.aqi, All.AQI_Category,
All.Raw_Conc, All. Conc__unit,All. Duration, All.QC_Name, cov.covid_cases1  as Covid_Cases,
covdeaths.covid_deaths1 as Covid_Deaths
from AllCov_mean_final All
inner join cov_cases_final cov on
cov.year=All.year and
cov.month=All.month and
cov.day=All.day
inner join cov_deaths_final covdeaths on
covdeaths.year=All.year and
covdeaths.month=All.month and
covdeaths.day=All.day
where all.Site="Jakarta Central" or Site="Jakarta South"
union
select All.Site, All.parameter,All.year, All.month,  All.day, All. NowCast_Conc, All.aqi, All.AQI_Category,
All.Raw_Conc, All. Conc__unit,All.Duration, All.QC_Name, covetho.covid_cases1  as Covid_Cases,
covdeathsetho.covid_deaths1 as Covid_Deaths
from AllCov_mean_final All
inner join cov_cases_ethofinal covetho on
covetho.year=All.year and
covetho.month=All.month and
covetho.day=All.day
inner join cov_deaths_ethofinal covdeathsetho on
covdeathsetho.year=All.year and
covdeathsetho.month=All.month and
covdeathsetho.day=All.day
```

```
where all.Site="Addis Ababa Central" or Site="Addis Ababa School"
;
quit;

/*** country wise **/
proc sql;
create table AllCov_mean_main_Indonesia as
select All.Site, All.parameter,All.year, All.month,  All.day, All. NowCast_Conc, All.aqi, All.AQI_Category,
All.Raw_Conc, All. Conc__unit,All. Duration, All.QC_Name, cov.covid_cases1  as Covid_Cases,
covdeaths.covid_deaths1 as Covid_Deaths
from AllCov_mean_final All
inner join cov_cases_final cov on
cov.year=All.year and
cov.month=All.month and
cov.day=All.day
inner join cov_deaths_final covdeaths on
covdeaths.year=All.year and
covdeaths.month=All.month and
covdeaths.day=All.day
where all.Site="Jakarta Central" or Site="Jakarta South";
quit;
proc sql;
create table AllCov_mean_main_Ethiopia as
select All.Site, All.parameter,All.year, All.month,  All.day, All. NowCast_Conc, All.aqi, All.AQI_Category,
All.Raw_Conc, All. Conc__unit,All.Duration, All.QC_Name, covetho.covid_cases1  as Covid_Cases,
covdeathsetho.covid_deaths1 as Covid_Deaths
from AllCov_mean_final All
inner join cov_cases_ethofinal covetho on
covetho.year=All.year and
covetho.month=All.month and
covetho.day=All.day
inner join cov_deaths_ethofinal covdeathsetho on
covdeathsetho.year=All.year and
covdeathsetho.month=All.month and
covdeathsetho.day=All.day
where all.Site="Addis Ababa Central" or Site="Addis Ababa School"
;
quit;


/****************************** import dataset with latest hours **************/
OPTIONS VALIDVARNAME=V7;
FILENAME outfile1 '/home/u58581363/MyPractice/Data/AbabaCentralDaily.csv';
FILENAME outfile2 '/home/u58581363/MyPractice/Data/AbabaSchoolDaily.csv';
FILENAME outfile3 '/home/u58581363/MyPractice/Data/JakartaCentralDaily.csv';
FILENAME outfile4 '/home/u58581363/MyPractice/Data/JakartaSouthDaily.csv';

options validvarname=v7;
proc import datafile=outfile1
                                out=work.AbabaCentralDaily
                                dbms=CSV
                                REPLACE;
                                guessingrows=max;
run;

options validvarname=v7;
proc import datafile=outfile2
                                out=work.AbabaSchoolDaily
                                dbms=CSV
                                REPLACE;
                                guessingrows=max;
```

```
run;

options validvarname=v7;
proc import datafile=outfile3
                                    out=work.JakartaCentralDaily
                                    dbms=CSV
                                    REPLACE;
                                    guessingrows=max;
run;

options validvarname=v7;
proc import datafile=outfile4
                                    out=work.JakartaSouthDaily
                                    dbms=CSV
                                    REPLACE;
                                    guessingrows=max;
run;

data ALL_HOUR_MAIN;
            retain  Site Parameter Date__LT_ Year Month Day Hour NowCast_Conc_ AQI AQI_Category Raw_Conc_
Conc__unit
Duration QC_Name;
            length AQI_Category $30;
            format AQi_Category $30.;
            length Site $20;
            format Site $20.;
            set  AbabaCentralDaily AbabaSchoolDaily JakartaCentralDaily  JakartaSouthDaily;

run;

/** re-evaluate AQI categories **/
data ALL_HOUR_MAIN;
set ALL_HOUR_MAIN;
            if 0<aqi<51 then AQI_Category='Good';
            if 51<aqi<101 then AQI_Category='Moderate';
            if 101<aqi<151 then AQI_Category='Unhealthy for Sensitive Groups';
            if 151<aqi<201 then AQI_Category='Unhealthy';
            if 201<aqi<301 then AQI_Category='Very Unhealthy';
            if 301<aqi then AQI_Category='Hazardous';
run;

/*** join datasets for hourly aqi**/
proc sql;
create table AllCov_hour_main as
select All.Site, All.parameter, All.Date__LT_, All.year,  All.month, All.day, All.NowCast_Conc_ as NowCast_Conc,  All.aqi,
All.AQI_Category,
All.Raw_Conc_ as Raw_Conc, All.Conc__unit,All.Duration,  All.QC_Name, cov.covid_cases1 as Covid_Cases,
covdeaths.covid_deaths1 as Covid_Deaths
from ALL_HOUR_MAIN All
inner join cov_cases_final cov on
cov.year=All.year and
cov.month=All.month and
cov.day=All.day
inner join cov_deaths_final covdeaths on
covdeaths.year=All.year and
covdeaths.month=All.month and
covdeaths.day=All.day
where Site="Jakarta Central" or Site="Jakarta South"
union
select All.Site, All.parameter, All.Date__LT_, All.year,  All.month, All.day, All.NowCast_Conc_ as NowCast_Conc,  All.aqi,
All.AQI_Category,
```

```
All.Raw_Conc_ as Raw_Conc, All. Conc__unit,All. Duration, All.QC_Name, covetho.covid_cases1 as Covid_Cases,
covdeathsetho.covid_deaths1 as Covid_Deaths
from ALL_HOUR_MAIN All
inner join cov_cases_ethofinal covetho on
covetho.year=All.year and
covetho.month=All.month and
covetho.day=All.day
inner join cov_deaths_final covdeathsetho on
covdeathsetho.year=All.year and
covdeathsetho.month=All.month and
covdeathsetho.day=All.day
where Site="Addis Ababa Central" or Site="Addis Ababa School"
;
quit;


/** country wise **/
proc sql;
create table AllCov_hour_main_Indonesia as
select All.Site, All.parameter, All.Date__LT_, All.year,  All.month, All.day, All.NowCast_Conc_ as NowCast_Conc,  All.aqi,
All.AQI_Category,
All.Raw_Conc_ as Raw_Conc, All.Conc__unit,All.Duration,  All.QC_Name, cov.covid_cases1 as Covid_Cases,
covdeaths.covid_deaths1 as Covid_Deaths
from ALL_HOUR_MAIN All
inner join cov_cases_final cov on
cov.year=All.year and
cov.month=All.month and
cov.day=All.day
inner join cov_deaths_final covdeaths on
covdeaths.year=All.year and
covdeaths.month=All.month and
covdeaths.day=All.day
where Site="Jakarta Central" or Site="Jakarta South";
quit;

proc sql;
create table AllCov_hour_main_Ethiopia as
select All.Site, All.parameter, All.Date__LT_, All.year,  All.month, All.day, All.NowCast_Conc_ as NowCast_Conc,  All.aqi,
All.AQI_Category,
All.Raw_Conc_ as Raw_Conc, All. Conc__unit,All. Duration,  All.QC_Name, covetho.covid_cases1 as Covid_Cases,
covdeathsetho.covid_deaths1 as Covid_Deaths
from ALL_HOUR_MAIN All
inner join cov_cases_ethofinal covetho on
covetho.year=All.year and
covetho.month=All.month and
covetho.day=All.day
inner join cov_deaths_final covdeathsetho on
covdeathsetho.year=All.year and
covdeathsetho.month=All.month and
covdeathsetho.day=All.day
where Site="Addis Ababa Central" or Site="Addis Ababa School"
;
quit;


/** correlaton of aqi & covid FOR MEAN DATA**/
title 'CORRELATION ANALYSIS FOR MEAN DATA OF ALL CITIES';
ods noproctitle;
ods graphics / imagemap=on;
proc corr data=WORK.ALLCOV_MEAN_MAIN_ETHIOPIA pearson nosimple plots=matrix;
        var Covid_Cases Covid_Deaths;
```

```
        with  AQI NowCast_Conc Raw_Conc;
run;

proc corr data=WORK.ALLCOV_MEAN_MAIN_INDONESIA pearson nosimple
                plots(maxpoints=10000)=matrix;
        var Covid_Cases Covid_Deaths;
        with AQI NowCast_Conc Raw_Conc;
run;



/** correlaton of aqi & covid FOR ORIGINAL DATA**/
title 'CORRELATION ANALYSIS FOR ETHIOPIA';
ods noproctitle;
ods graphics / imagemap=on;

proc corr data=WORK.ALLCOV_HOUR_MAIN_ETHIOPIA pearson nosimple plots=matrix;
        var Covid_Cases Covid_Deaths;
        with  AQI NowCast_Conc Raw_Conc;
run;
title 'CORRELATION ANALYSIS FOR INDONESIA';
proc corr data=WORK.ALLCOV_HOUR_MAIN_INDONESIA pearson nosimple
plots(maxpoints=10000)=matrix;
        var Covid_Cases Covid_Deaths;
        with  AQI NowCast_Conc Raw_Conc;
run;




/** spearmen correlaton of month with covid**/
title 'CORRELATION ANALYSIS FOR AQI vs MONTH OF ALL  CITIES';
ods noproctitle;
ods graphics / imagemap=on;

ods noproctitle;
ods graphics / imagemap=on;

proc corr data=WORK.ALLCOV_MEAN_MAIN_ETHIOPIA pearson spearman nosimple
                plots=matrix;
        var Covid_Cases Covid_Deaths;
        with Month;
run;


proc corr data=WORK.ALLCOV_MEAN_MAIN_INDONESIA pearson spearman nosimple
                plots=matrix;
        var Covid_Cases Covid_Deaths;
        with Month;
run;




/****** Linear Regression Analysis **********/
```

```
title 'REGRESSION ANALYSIS FOR  AQI & COVID CASES';
ods noproctitle;
ods graphics / imagemap=on;

title 'ETHIOPIA';
proc reg data=Work.AllCov_hour_main_Ethiopia alpha=0.05 plots(only)=(diagnostics residuals
                        fitplot observedbypredicted);
            model Covid_Cases=AQI /;
            run;
quit;

title 'INDONESIA';
proc reg data=Work.AllCov_hour_main_Indonesia alpha=0.05 plots(only)=(diagnostics residuals
                        fitplot observedbypredicted);
            model Covid_Cases=AQI /;
            run;
quit;

title 'REGRESSION ANALYSIS AQI & COVID DEATHS';
title 'ETHIOPIA';
proc reg data=Work.AllCov_hour_main_Ethiopia alpha=0.05 plots(only)=(diagnostics residuals
                        fitplot observedbypredicted);
            model Covid_Deaths=AQI /;
            run;
quit;

title 'INDONESIA';
proc reg data=Work.AllCov_hour_main_Indonesia alpha=0.05 plots(only)=(diagnostics residuals
                        fitplot observedbypredicted);
            model Covid_Deaths=AQI /;
            run;
quit;

/*no plots*/
title 'REGRESSION ANALYSIS FOR  MONTH & COVID CASES';
ods noproctitle;
ods graphics / imagemap=on;

title 'ETHIOPIA';
proc reg data=Work.AllCov_hour_main_Ethiopia alpha=0.05 ;
            model Covid_Cases= Month/;
            run;
quit;

title 'INDONESIA';
proc reg data=Work.AllCov_hour_main_Indonesia alpha=0.05 ;
            model Covid_Cases=  Month/;
            run;
quit;

title 'REGRESSION ANALYSIS Month & COVID DEATHS';
title 'ETHIOPIA';
proc reg data=Work.AllCov_hour_main_Ethiopia alpha=0.05 ;
            model Covid_Deaths=Month/;
            run;
quit;

title 'INDONESIA';
proc reg data=Work.AllCov_hour_main_Indonesia alpha=0.05 ;
            model Covid_Deaths= Month /;
            run;
```

```
quit;



/**/
title 'REGRESSION ANALYSIS FOR Nowcast & COVID CASES';
ods noproctitle;
ods graphics / imagemap=on;

title 'ETHIOPIA';
proc reg data=Work.AllCov_hour_main_Ethiopia alpha=0.05 ;
        model Covid_Cases= NowCast_Conc/;
        run;
quit;

title 'INDONESIA';
proc reg data=Work.AllCov_hour_main_Indonesia alpha=0.05 ;
        model Covid_Cases=  NowCast_Conc/;
        run;
quit;

title 'REGRESSION ANALYSIS NowCast_Conc & COVID DEATHS';
title 'ETHIOPIA';
proc reg data=Work.AllCov_hour_main_Ethiopia alpha=0.05 ;
        model Covid_Deaths=NowCast_Conc/;
        run;
quit;

title 'INDONESIA';
proc reg data=Work.AllCov_hour_main_Indonesia alpha=0.05 ;
        model Covid_Deaths= NowCast_Conc /;
        run;
quit;



/** AQI CATEGORY vs Month CHI SQUARE TEST **/
title 'CHI SQUARE ANALYSIS';

title 'ADDIS ABABA CENTRAL';
proc freq data=WORK.central;
        tables  (AQI_Category) *(Month) / chisq nopercent norow nocol nocum
                plots(only)=(freqplot mosaicplot);
run;

title 'ADDIS ABABA SCHOOL';
proc freq data=WORK.abschool;
        tables  (AQI_Category) *(Month) / chisq nopercent norow nocol nocum
                plots(only)=(freqplot mosaicplot);
run;

title 'JAKARTA CENTRAL';
proc freq data=WORK.JcFINAL;
        tables  (AQI_Category) *(Month) / chisq nopercent norow nocol nocum
                plots(only)=(freqplot mosaicplot);
run;

title 'JAKARTA SOUTH';
proc freq data=WORK.JSFINAL;
        tables  (AQI_Category) *(Month) / chisq nopercent norow nocol nocum
                plots(only)=(freqplot mosaicplot);
```

```
run;


/************* T-Test analysis ******************/
ods noproctitle;
ods graphics / imagemap=on;

/**JAKARTA CENTRAL**/
/* Test for normality */
title 'JAKARTA CENTRAL';
proc univariate data=WORK.JakartaCentral normal mu0=90;
		ods select TestsForNormality;
		var AQI;

run;
/* t test */
proc ttest data=WORK.JakartaCentral sides=U h0=90;
		var AQI;

run;

/**JAKARTA SOUTH**/
/* Test for normality */
title 'JAKARTA SOUTH';
proc univariate data=WORK.JakartaSouth normal mu0=90;
		ods select TestsForNormality;
		var AQI;

run;
/* t test */
proc ttest data=WORK.JakartaSouth sides=U h0=90 plots(showh0);
		var AQI;

run;

/**ABABA CENTRAL**/
/* Test for normality */
title 'ADDIS ABABA CENTRAL';
proc univariate data=WORK.AddisAbabaCentral normal mu0=90;
		ods select TestsForNormality;
		var AQI;

run;
/* t test */
proc ttest data=WORK.AddisAbabaCentral sides=U h0=90 ;
		var AQI;

run;

/**ABABA SCHOOL**/
/* Test for normality */
title 'ADDIS ABABA SCHOOL';
proc univariate data=WORK.AddisAbabaSchool normal mu0=90;
		ods select TestsForNormality;
		var AQI;

run;
/* t test */
proc ttest data=WORK.AddisAbabaSchool sides=U h0=90 ;
		var AQI;
```

```
run;


/********** AQI CATEGORY vs Covid Cases CHI SQUARE  TEST ***************/

title 'ETHIOPIA';
proc hpbin data=Work.AllCov_hour_main_Ethiopia computehist computequantile pseudo_quantile;
  input Covid_Cases / numbin=3;

data AllCov_hour_main_Ethiopia_cat;
set AllCov_hour_main_Ethiopia;
format cat_Covid_Cases $8.;
run;

data AllCov_hour_main_Ethiopia_cat;
set AllCov_hour_main_Ethiopia_cat;
if Covid_Cases <  8490 then cat_Covid_Cases='low';
if  8491 <=Covid_Cases and Covid_Cases< 124281 then cat_Covid_Cases='medium';
if            124280 <= Covid_Cases then cat_Covid_Cases='high';
run;

ods noproctitle;

proc freq data=WORK.ALLCOV_HOUR_MAIN_ETHIOPIA_CAT;
            tables  (AQI_Category) *(cat_Covid_Cases) / chisq nopercent norow nocol nocum
                        plots(only)=(freqplot mosaicplot);
run;

title 'INDONESIA';
proc hpbin data=Work.AllCov_hour_main_Indonesia computehist computequantile pseudo_quantile;
  input Covid_Cases / numbin=3;

data AllCov_hour_main_Indonesia_cat;
set AllCov_hour_main_Indonesia;
format cat_Covid_Cases $8.;
run;

data AllCov_hour_main_Indonesia_cat;
set AllCov_hour_main_Indonesia_cat;
if Covid_Cases <  1331 then cat_Covid_Cases='low';
if  1331 <=Covid_Cases and Covid_Cases< 4850 then cat_Covid_Cases='medium';
if            4850 <= Covid_Cases then cat_Covid_Cases='high';
run;

ods noproctitle;

proc freq data=WORK.ALLCOV_HOUR_MAIN_INDONESIA_CAT;
            tables  (AQI_Category) *(cat_Covid_Cases) / chisq nopercent norow nocol nocum
                        plots(only)=(freqplot mosaicplot);
run;
```