Statistics Project Report

AIR QUALITY ANALYSIS on the basis of PM10 parameter.

DF-2009 Batch
Aegis School of Data Science

Submitted To:
Dr. Vinay Kulkarni

Submitted By:
Jayshri Gupta
Pravesh Raikwar

INTRODUCTION

We decided to analyze the air quality data of Mandideep, Sector D, M.P. to find some underlying principles or patterns which might give us an insight into how severe the problem is.

How Is AQI Measured?

Eight pollutants namely particulate matter (PM) 10, PM2.5, Ozone (O3), Sulphur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), lead (Pb) and ammonia (NH3) act as major parameters in deriving the AQI of an area. Short-term (upto 24-hours and 8-hours for CO and O3) National Ambient Air Quality Standards are prescribed for these pollutants.

Particulate Matter
Using a nationwide network of monitoring sites, EPA has developed ambient air quality trends for particle pollution, also called Particulate Matter (PM). PM_{10} describes inhalable particles, with diameters that are generally 10 micrometers and smaller. Under the Clean Air Act, EPA sets and reviews national air quality standards for PM. Air quality monitors measure concentrations of PM throughout the country. EPA, state, tribal and local agencies use that data to ensure that PM in the air is at levels that protect public health and the environment. Nationally, average PM10 concentrations have decreased over the years.

Exploratory Data Analysis(EDA)

DATA

The data of Air Quality Index has been derived for Sector-D Industrial Area, Mandideep - MPPCB from the state of Madhya Pradesh for the year 2020(01/01/2020 to 31/12/2020)

Mandideep is a town with municipality in Goharganj sub-district of Raisen district in the Indian state of Madhya Pradesh. Mandideep is 23 km from Bhopal and is basically an Industrial township which came into existence in late 1970s.

The data obtained had following columns:

```
In [4]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 65536 entries, 0 to 65535
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   location     65536 non-null  object
 1   city         65536 non-null  object
 2   country      65536 non-null  object
 3   utc          65536 non-null  object
 4   local        65536 non-null  object
 5   parameter    65536 non-null  object
 6   value        65536 non-null  float64
 7   unit         65536 non-null  object
 8   latitude     65536 non-null  float64
 9   longitude    65536 non-null  float64
 10  attribution  65536 non-null  object
dtypes: float64(3), object(8)
memory usage: 5.5+ MB
```

Description of columns:

- Location : Describes the area under the observation
- City: Mandideep(the city taken into consideration)
- Country : INDIA
- Utc: The Universal time stamp for the observations
- Local: The local timestamp for the observations.
- Parameter: Several parameters that affect the Air Quality Index value (PM25, PM10, SO_2, CO, NO_2, O_3)
- Value: The observed value of the parameter at a particular time of the day.
- Unit: $\mu g/m^3$

- Latitude: The geographical position of the place.(Here, 23.10844)
- Longitude: The geographical position of the place.(Here, 77.511428)
- Attribution: The URLs for various APIs.

location	city	country	utc	local	parameter	value	unit	latitude	longitude	attribution
Sector-D Industrial Area, Mandideep - MPPCB	Mandideep	IN	2020-12-19T17:30:00Z	2020-12-19T23:00:00+05:30	so2	23.4	Âµg/mÂ³	23.10844	77.511428	[{"url":"https://app.cpcbccr.c dashboard-all/caaqm-landing' Pollution Control Board"}]
Sector-D Industrial Area, Mandideep - MPPCB	Mandideep	IN	2020-12-19T17:30:00Z	2020-12-19T23:00:00+05:30	co	1100	Âµg/mÂ³	23.10844	77.511428	[{"url":"https://app.cpcbccr.c dashboard-all/caaqm-landing' Pollution Control Board"}]
Sector-D Industrial Area, Mandideep - MPPCB	Mandideep	IN	2020-12-19T17:30:00Z	2020-12-19T23:00:00+05:30	no2	61.9	Âµg/mÂ³	23.10844	77.511428	[{"url":"https://app.cpcbccr.c dashboard-all/caaqm-landing' Pollution Control Board"}]
Sector-D Industrial Area, Mandideep - MPPCB	Mandideep	IN	2020-12-19T17:30:00Z	2020-12-19T23:00:00+05:30	o3	6.4	Âµg/mÂ³	23.10844	77.511428	[{"url":"https://app.cpcbccr.c dashboard-all/caaqm-landing' Pollution Control Board"}]
Sector-D Industrial Area, Mandideep - MPPCB	Mandideep	IN	2020-12-19T17:30:00Z	2020-12-19T23:00:00+05:30	pm25	63	Âµg/mÂ³	23.10844	77.511428	[{"url":"https://app.cpcbccr.c dashboard-all/caaqm-landing' Pollution Control Board"}]

Redundant Columns:

- Location
- City
- Country
- Units
- Longitude
- Latitude
- Attribution

DATA INTO CONSIDERATION:

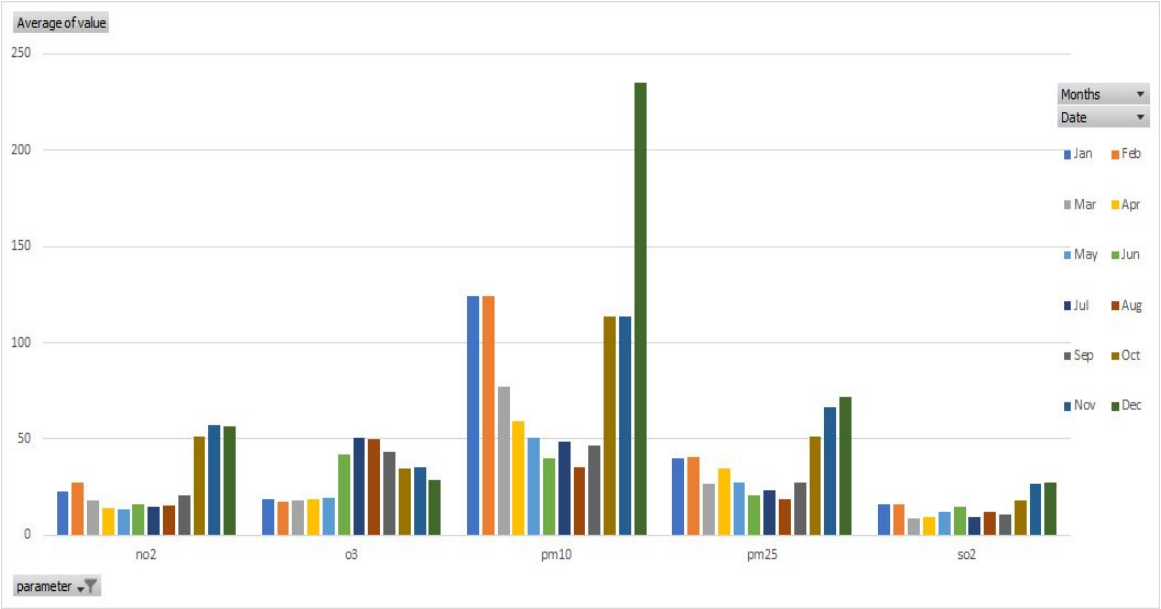
Total rows= 65537, Columns = 3

- Parameter
- Value
- Local(Trimmed for date)

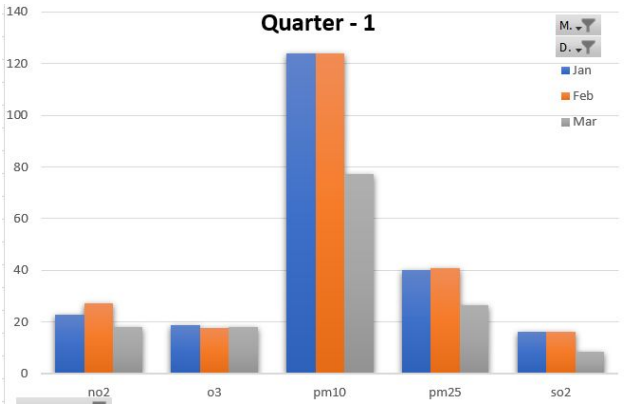
	A	B	C
1	Date	Parameter	Value
2	19-12-2020	so2	23.4
3	19-12-2020	co	1100
4	19-12-2020	no2	61.9
5	19-12-2020	o3	6.4
6	19-12-2020	pm25	63
7	19-12-2020	pm10	128
8	19-12-2020	pm25	24
9	19-12-2020	pm10	95
10	19-12-2020	o3	37.3
11	19-12-2020	so2	14.7
12	19-12-2020	no2	103.2
13	19-12-2020	co	800
14	19-12-2020	no2	28.5
15	19-12-2020	co	700
16	19-12-2020	so2	10.3
17	19-12-2020	o3	5.3
18	19-12-2020	pm25	57

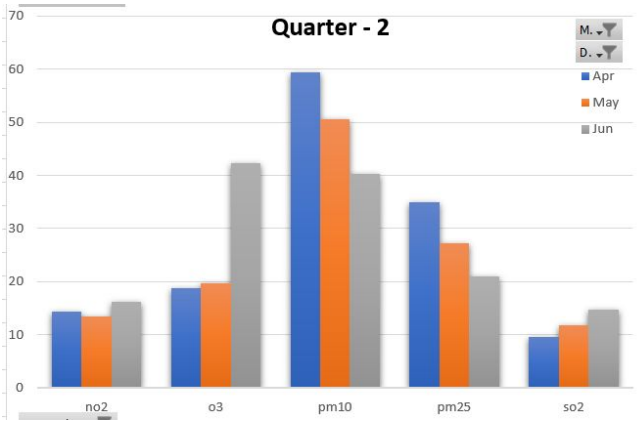
VISUAL ANALYSIS:

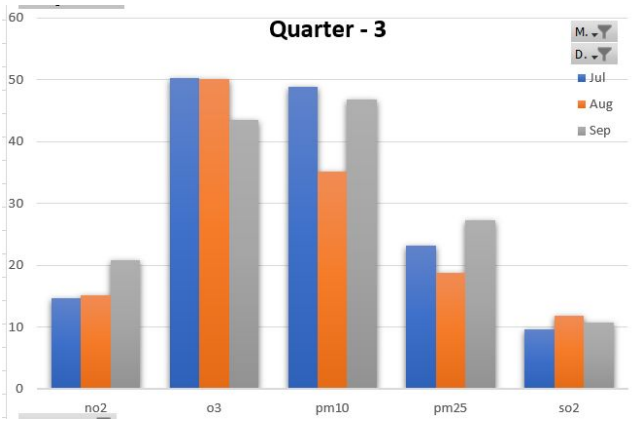
Plotting of various parameter values in different months.

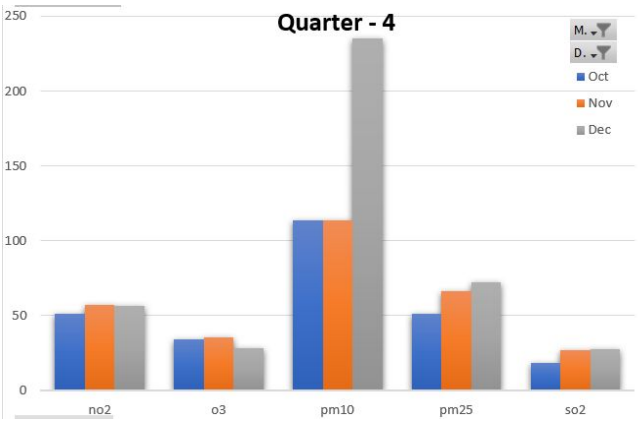


Quarterly Plots

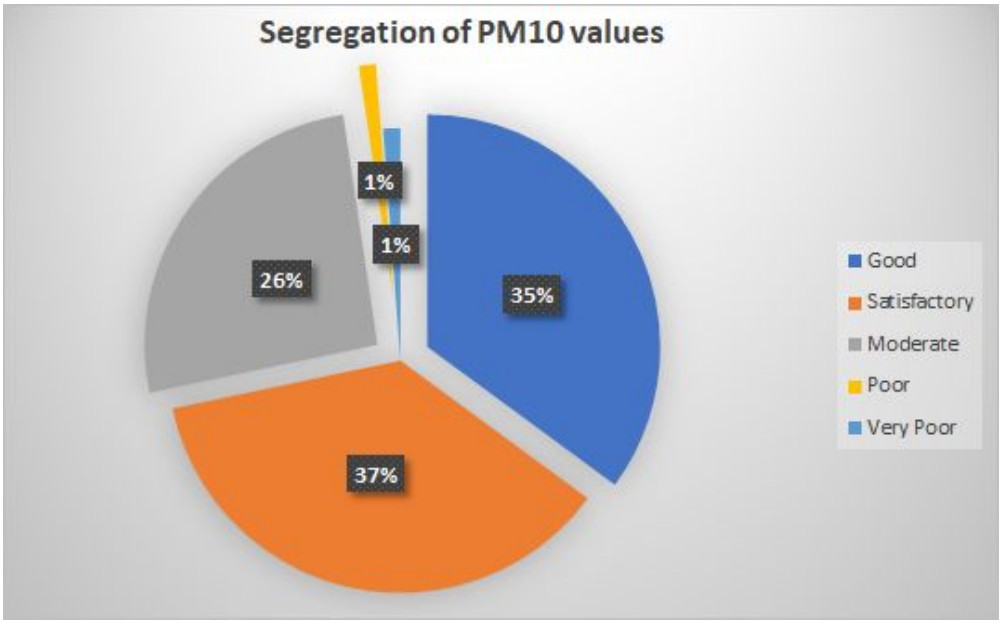




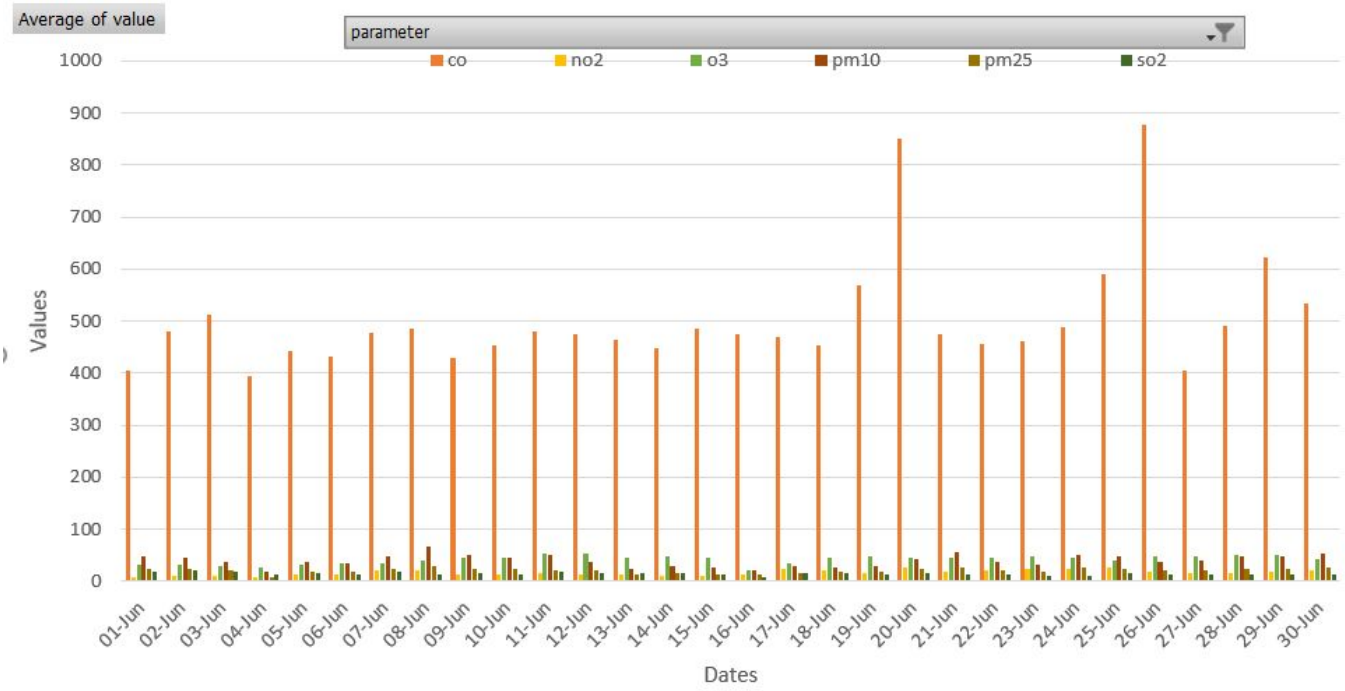




Pie chart showing what percentage of PM10 values lie in Good, Satisfactory, Moderate, Poor, Very Poor in accordance to the standard PM10 values in a day.



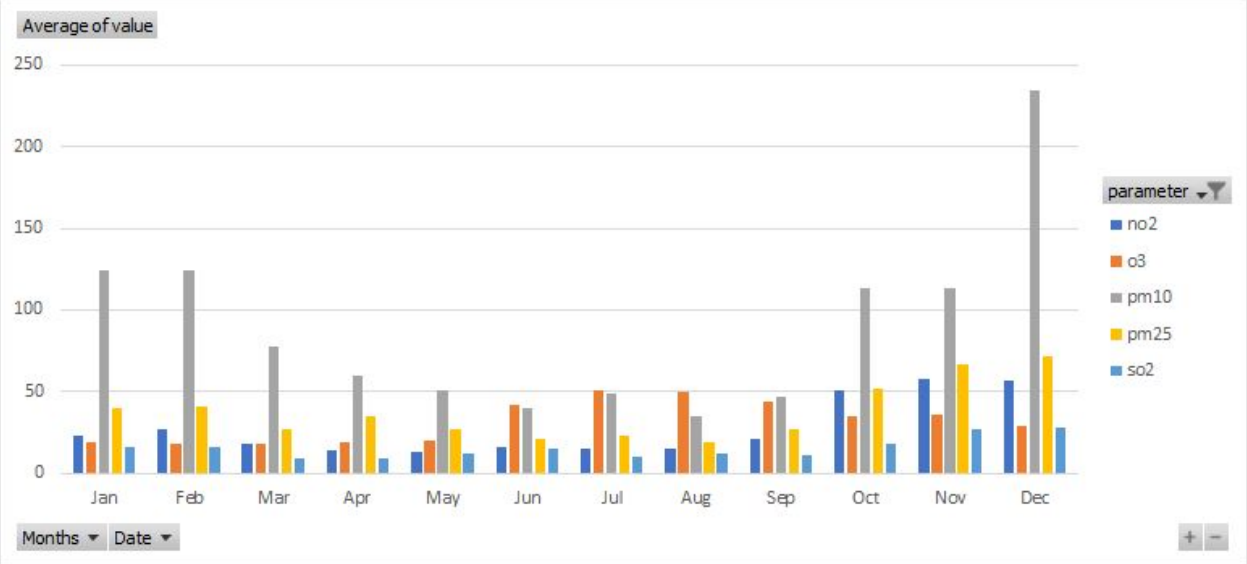
Bar plot of June data



PLOTS USED:

Using Pivot charts and tables in Excel, we analyse the data over the span of the entire year for different parameters to choose the most significant ones for our analysis.

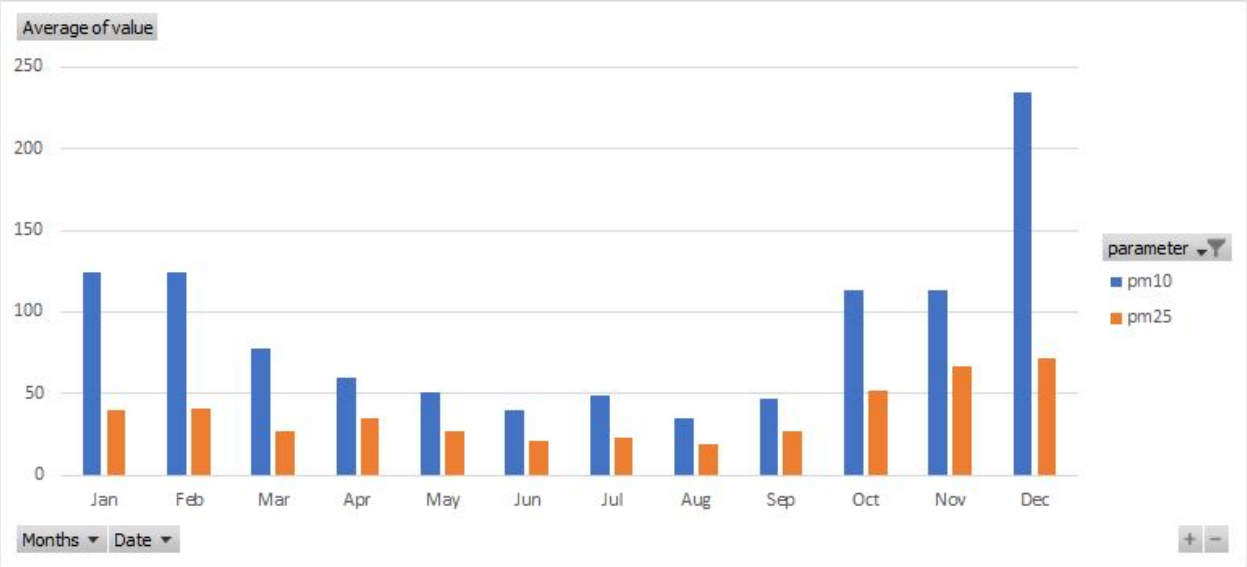
Average of value	Column Labels					
Row Labels	no2	o3	pm10	pm25	so2	C
⊞ Jan	22.67465181	18.81488331	124.1065644	40.12532079	16.09458599	
⊞ Feb	27.19189189	17.64075738	123.9447236	40.94931164	16.18787475	
⊞ Mar	17.99440613	17.88909513	77.44207317	26.44207317	8.605801773	
⊞ Apr	14.34466886	18.81731765	59.46256817	34.96661898	9.526347009	
⊞ May	13.40608247	19.68169597	50.51599587	27.28202479	11.875	
⊞ Jun	16.28644252	42.27878077	40.22791712	20.98251366	14.66946039	
⊞ Jul	14.605	50.31857977	48.93784787	23.20871985	9.678277886	
⊞ Aug	15.13533835	50.19545455	35.18181818	18.71969697	11.8744186	
⊞ Sep	20.78011696	43.56432749	46.86549708	27.30994152	10.81024096	
⊞ Oct	51.11743119	34.47965616	113.6749311	51.54317549	18.27885196	
⊞ Nov	57.30199336	35.38422819	113.410828	66.56507937	26.7550173	
⊞ Dec	56.5472973	28.43661972	234.9113924	72.08860759	27.46111111	



Inferences from the above graph:

1. The most significant values are of PM10 and second of the PM 25 values. Hence, we consider these parameters for our study further.
2. There has been a major decrement and increment in the two halves of the years respectively- Air Quality Index(Parameters) have declined in the first half of the year and have significantly increased in the second half after the month of June.
3. The values of SO_2 have been very low in comparison to those of other parameters hence, we can drop it from our study point of view for this project.

Further, we plot the data that we require for our study(According to our consideration)



Note: We have used PM10 and PM25 data as it is supposed to be the most significant factor affecting the Air Quality Index and hence for the reasons stated below:

Many studies have demonstrated a close relationship between particulate matter (PM10) pollution and deterioration in human health. The key properties of airborne particles are generally considered to be the size of aerosols and the associated capacity for penetration into the human respiratory system. This is supported by epidemiological evidence. As such, the concentration of PM10 has been monitored extensively in urban areas in many western countries followed by many Asian countries.

AQI Category, Pollutants and Health Breakpoints

AQI Category (Range)	PM_{10} (24hr)	$PM_{2.5}$ (24hr)	I
Good (0–50)	0–50	0–30	(
Satisfactory (51–100)	51–100	31–60	4
Moderately polluted (101–200)	101–250	61–90	8
Poor (201–300)	251–350	91–120	1
Very poor (301–400)	351–430	121–250	2
Severe (401–500)	430+	250+	4

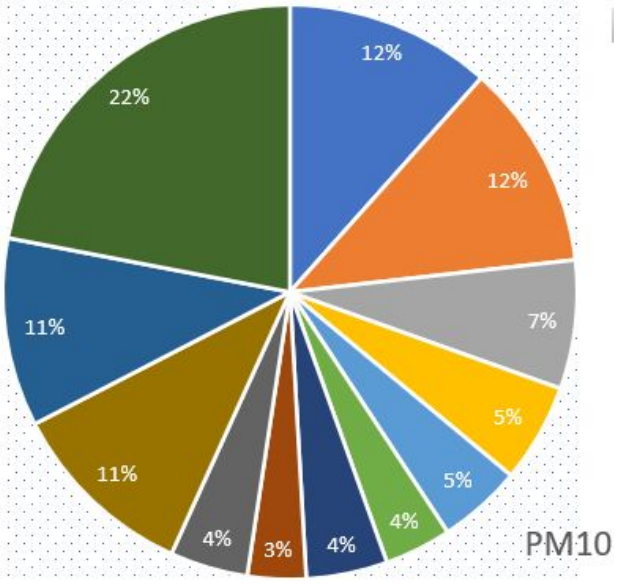
The Standard level of PM25 and Pm 10 can be found in the table.

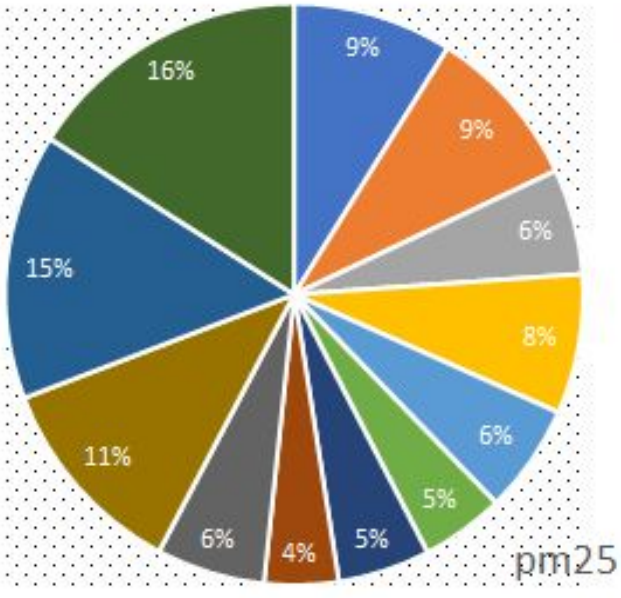
DESCRIPTIVE STATISTICS

Descriptive Analysis (PM10)	
Mean	84.73862
Standard Error	0.63679
Median	66
Mode	53
Standard Deviation	67.66172
Sample Variance	4578.108
Kurtosis	17.18176
Skewness	3.068847
Range	897
Minimum	0
Maximum	897
Sum	956699
Count	11290
Confidence Level(95.0%)	1.248219

Descriptive Analysis(PM25)	
Mean	34.34272
Standard Error	0.203475
Median	28
Mode	27
Standard Deviation	21.68891
Sample Variance	470.4089
Kurtosis	13.19974
Skewness	2.724764
Range	289
Minimum	5
Maximum	294
Sum	390202
Count	11362
Confidence Level(95.0%)	0.398846

Plotting other charts to understand Data Trends:





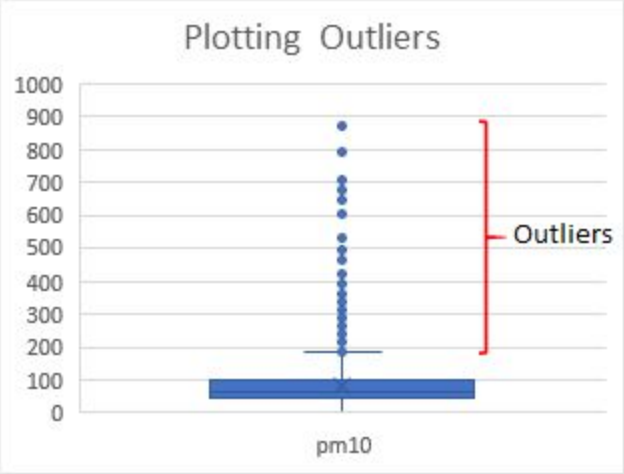
DATA CLEANING:

Data:
PM10 values for each day in 2020

Outlier identification method:
IQR

Outlier replacement:
Mean of the population

Identifying Outliers	
Q1	44
Q2	66
Q3	101
IQR	57
Maximum	186.5
Minimum	-41.5
Number of Outliers	791
Total Data Points	11290
%Outliers in the data	7.0062002



HYPOTHESIS TESTING:

Hypothesis: The average PM10 contribution in AQI between the first quarter of the year and the second quarter of the year is the same.

Ho: μ1=μ2 μ1-μ2=0

Ha: μ1≠μ2 μ1-μ2≠0

Where:

μ1 is the mean of the PM10 data for the first quarter in 2020.
μ2 is the mean of the PM10 data for the second quarter in 2020.

Proportions into consideration:

Proportion 1 : The entire cleansed data of the first three months(January,2020 to March, 2020) of PM10 having **5251** data points.

Descriptive Analysis for quarter 1	
Mean	90.29058007
Standard Error	0.512422941
Median	84.73861825
Mode	84.73861825
Standard Deviation	37.12856949
Sample Variance	1378.530673
Kurtosis	-0.154125346
Skewness	0.478155943
Range	175
Minimum	11
Maximum	186
Sum	474025.5453
Count	5250
Confidence Level(95.0%)	1.004562151

Proportion 2 : The entire cleansed data of the first three months(January,2020 to March, 2020) of PM10 having **3904** data points.

Descriptive Analysis for quarter 2	
Mean	52.28814952
Standard Error	0.383670031
Median	48
Mode	44
Standard Deviation	23.96939942
Sample Variance	574.5321085
Kurtosis	3.082724442
Skewness	1.312259423
Range	172
Minimum	3
Maximum	175
Sum	204080.6476
Count	3903
Confidence Level(95.0%)	0.75221277

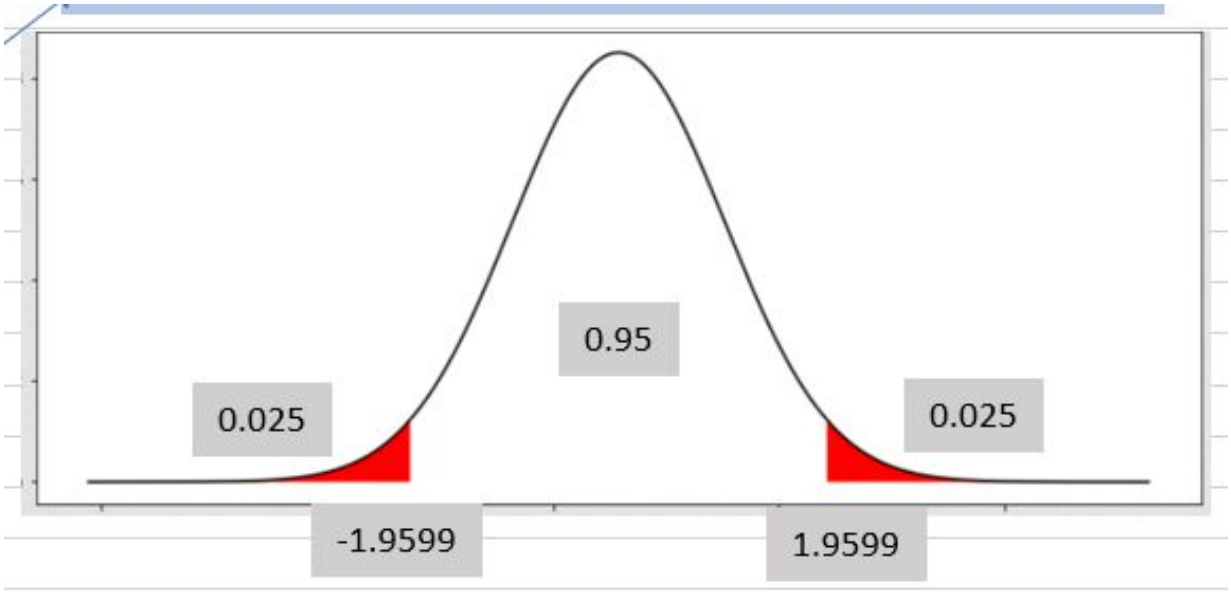
Test used = z-Test: Two Sample for Means
Two-tailed test

Reasons:

- Normally distributed data
- Independent proportion
- Large sample size

z-Test: Two Sample for Means

	Quarter 1	*Quarter 2*
Mean	90.29058007	52.28814952
Known Variance	1378.53	574.53
Observations	5250	3903
Hypothesized Mean Difference	0	
z	59.365806	
P(Z<=z) one-tail	0	
z Critical one-tail	1.644853627	
P(Z<=z) two-tail	0	
z Critical two-tail	1.959963985	



Conclusion:

(z value) **59.366 >> 1.959** (z critical)

z value significantly greater than z critical two-tail, hence we neglect null hypothesis and suggest alternative hypothesis.

ANOVA:

Comparison of PM10 data between different weeks of June, 2020. The data used is displayed below.

Week 1	Week 2	Week 3	Week 4
48.58139535	65.71794872	25.17647059	37.2
46.42424242	50.29268293	21.85	30.75609756
36.68965517	46.36666667	29.18421053	50.88888889
17.30434783	49.51351351	26.28125	47.07142857
36.35483871	37.89285714	28.87179487	37.36363636
34.05882353	23	43.05714286	39.97297297
47.5	30.2	57.30434783	48.13888889

Carrying out **Anova: Single Factor test** on the above data to analyse if there is a difference between the week 1, week 2, week 3, and week 4 of June, 2020.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Week 1	7	266.913303	38.13047186	120.2218463
Week 2	7	302.983669	43.28338128	202.0682491
Week 3	7	231.7252167	33.10360238	159.0077254
Week 4	7	291.3919132	41.62741618	52.75266795

F<Fcritical implies that there is a significant difference between the PM10 values between the weeks of June.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	425.3830566	3	141.7943522	1.062029566	0.383638	3.00878657
Within Groups	3204.302932	24	133.5126222			
Total	3629.685989	27				

Since,

$$F < F_{Critical}$$

We conclude that there is a significant difference between the PM10 values between the weeks of June.

An ANOVA test can tell you if your results are significant overall, but it won't tell you exactly where those differences lie. After you have run an ANOVA and found significant results, then you can run Tukey's HSD to find out which specific groups' means (compared with each other) are different. The test compares all possible pairs of means. Now, we carry out Tukey's test to find out which two weeks have a statistically significant difference.

TUKEY'S TEST:

To determine exactly which group means are different, we can perform a Tukey-Kramer post hoc test.
We use the Q-value method to analyse the results of the significance difference. We find the Q critical value using the following formula:

$$\textbf{Q critical value} = Q^* \sqrt{(s^2 \text{ pooled} / n)}$$

- Q = Value from Studentized Range Q Table
- s^2 pooled = Pooled variance across all groups
- n = Sample size for a given group

$Q_{critical}$ = 3.9047
Pooled Variance = 133.513

Tukey-Kramer Post Hoc Test

Qvalue used(n:28, k:4) =	3.9047	Pooled Variance=	133.5126222
	Abs Mean Difference	**Q critical**	**Signficance**
Week 1 vs week2	5.152909422	8.526480821	FALSE
Week 2 vs week3	10.1797789	8.526480821	TRUE
week 3 vs week 4	8.523813797	8.526480821	FALSE
week 4 vs week 1	3.496944319	8.526480821	FALSE

Significance is calculated by comparing **|Mean Difference|** and **$Q_{critical.}$**

INFERENCE:

- The difference in means between week2 and week3 is statistically significant.