Final Report

Title: Air Quality Analysis

Mini Project 3: Python for Data Science

Student: Ghefua Yembu

XULA ID: 900285543

Course: CPSC 3603

Term: Fall 2023

Abstract

Poor air quality results in several respiratory track infections and diseases such as asthma, lung cancer and other complications. In an effort to understand and hopefully reduce the effect of poor air quality on human health, environmental scientist started monitoring air quality for analysis and categorisation based on the Environmental protection Agency(EPA) standards to identify low and high risk areas. This project focuses analysing and deriving meaningful insights from air quality data that can create awareness to scientist and others affected by air quality in an effort to increase the quality of health and life. The data used for this project was collected by the EcoStem research team at Xavier University of Louisiana.

Keywords:
Air Quality Index- A measurement of air pollution;
Particulate Matter- microscopic liquid droplets or solids that can be inhaled and cause serious health problems.

Contents

Introduction

Particulate Matter are microscopic liquid droplets and solid particles in the atmosphere that cause air pollution and are hazardous when inhaled. The particulate matter that posses the greatest health risk is PM 2.5(micrometers) and it's effect on air pollution is measured through air quality index value (AQI2.5). This project is focused on different ways through which insights could be obtained from the raw PM data collected using Arduino sensors. Some of the questions these project answers concerning air quality are listed below.

Research questions/hypotheses

Hypothesis: Analysing PM2.5 data could generate actionable insights that can be useful to stake holders such as researchers and the general public.
1. What does the originally collected data look like?
2. How can particulate matter data collected from sensor be cleaned and created into a format that can be more easily analyzed.
3. How can air quality be classified into different categories?
4. What is the general distribution of the AQI2.5 values in the Art Center
5. What are the highest values of AQI2.5 for each day.
6. What is the dominant category AQI2.5 values in the Art Center fall in?
7. In what ways can the distribution of AQI2.5 values in the different categories be visualized?
8. In what day of the week, time of the day and month of the year did the highest AQI2.5 value occur ?
8. How can air quality be differentiated into different months?
9. What day of the week had the highest AQI2.5 value?
10. What day of the week had the highest AQI2.5 value?
11. What month of the year had the highest AQI2.5 value?
12.How are air quality values for each month distributed across the different categories?

Dataset

This data was collected for the purpose of Air Quality Monitoring in real time from the Art Center, a location on Xavier University of Louisiana's campus in the year 2022 by the EcoStem research team at Xavier. The data was collected at 3 minute intervals, from a particulate matter(PM) sensor recording the concentration of particulate matter(small solid and liquid particles that result in air pollution) in the atmosphere. Different sizes of PM are collected including PM10micrometers and PM 2.5 micrometer.
The raw data comes in 23 columns and 65535 rows. After the deletion of null values and the use of feature engineering to create two additional categorical columns a data frame containing 25 rows and 65515 columns is generated. Initially, the data contains one column with data type of 'datetime64[ns]' while the rest of the columns are of

data type 'object'. Most of the analysis and research questions concerning this data set will be related to Air Quality index 2.5 (AQI2.5) which is the derived air quality from particulate matter of size 2.5 micrometers (PM2.5) because this is the most hazardous to human health, thus the data type of this column(AQI2.5) will be converted to a more suitable data type that can be easily analysed.

Data Analysis

Data was converted from a an excel file to a pandas data frame.

All rows containing missing values were dropped to improve the quality of the data and output. The data type of AQI 2.5 was changed from object to integer to facilitate analysis.

A column was created to reflect the 3 categories each of the data points fell in(good,moderate, bad). Another column was created to show the name of the months when the data was collected.

After viewing the shape of the dimensions of the data, a histogram was plotted to view the overall distribution of AQI2.5 data in the Art Center. The histogram was right skewed with most of the data points falling within the EPA standards for good air quality(¡50).

The highest AQI2.5 value for each day was isolated as well as the overall highest value With the use of the 'loc' function.

To know the dominant category air quality values in the art center fall, a bar chart and pie chart were generated based on the number of AQI2.5 values in each category.

The highest AQI2.5 value was used to extract the specific day of the week, month and time of the day when this value occurred.

An area plot was created of maximum AQI2.5 values for each day to visualize when the highest AQI2.5 value of the year was recorded.

A bar graph was created of the AQI2.5 values in each month, day of the week and month of the year.

The groupby function was also used to see how the different categories of air quality in each month were distributed.

Data Visualization

visualization goes here

```
df.head(-5)
```

	Date	Day of the week	Time of the day	yr	mon	day	hr	min	sec	day_frac	...	PM2.5	PM10	AQI_5	AQI10	p0.3	p0.5	p1.0	p2.5	p5	p10
0	2022-05-01	Sunday	Night	2022	5	1	0	2	53	1.002	...	15.3	16.4	57.0	15	2136.57	576.29	91.09	6.14	1.55	1.09
1	2022-05-01	Sunday	Night	2022	5	1	0	5	47	1.00402	...	14.6	15.3	56.0	14	2086.27	562.72	86.46	3.99	0.94	0.64
2	2022-05-01	Sunday	Night	2022	5	1	0	8	42	1.00604	...	14.6	15	56.0	14	2059.74	548.98	86.48	2.93	0.65	0.35
3	2022-05-01	Sunday	Night	2022	5	1	0	11	36	1.00806	...	14.9	15.7	56.0	15	2067.21	562.92	80.99	5.32	1.04	0.43
4	2022-05-01	Sunday	Night	2022	5	1	0	14	30	1.01007	...	14.9	16	56.0	15	2134.63	581.41	88.14	3.21	1.79	0.6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
65525	2022-06-25	Saturday	Night	2022	6	25	5	49	48	25.24292	...	20.4	21.3	68.0	20	2488.33	688.37	126.55	10.98	1.5	0.1
65526	2022-06-25	Saturday	Night	2022	6	25	5	52	41	25.24492	...	19	20.9	65.0	19	2544.48	689.24	118.11	7.03	2.96	0.8
65527	2022-06-25	Saturday	Night	2022	6	25	5	55	36	25.24695	...	17.6	20.1	62.0	19	2329.74	635.2	108.41	12.1	3.82	0.96
65528	2022-06-25	Saturday	Night	2022	6	25	5	58	30	25.24896	...	18	20	63.0	19	2369.2	644.62	111.1	8.29	3.08	0.9
65529	2022-06-25	Saturday	Day	2022	6	25	6	1	25	25.25098	...	19.4	20.5	66.0	19	2456.28	675.95	124.03	8.15	1.71	0.37

65510 rows × 23 columns

Figure 1: Raw data

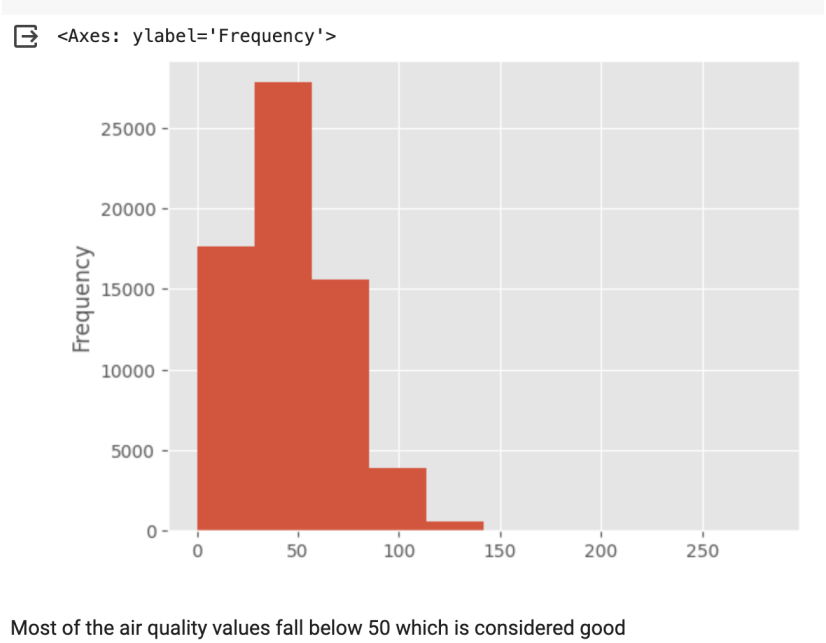

Figure 2: Right skewed histogram

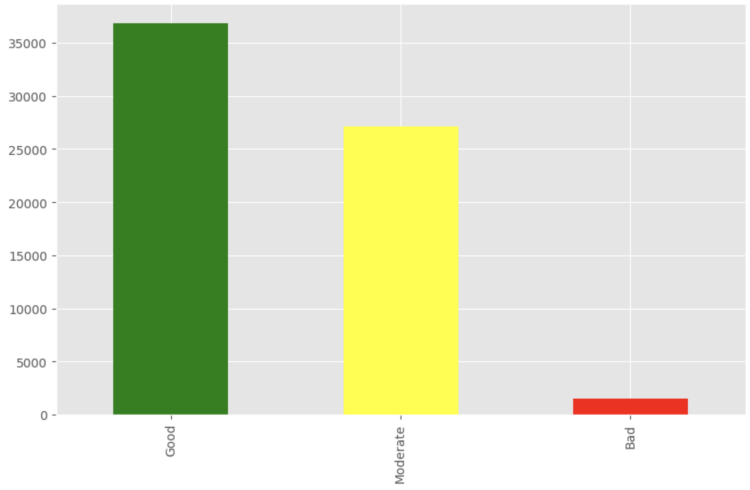

Figure 3: Bar chart showing air quality category distribution

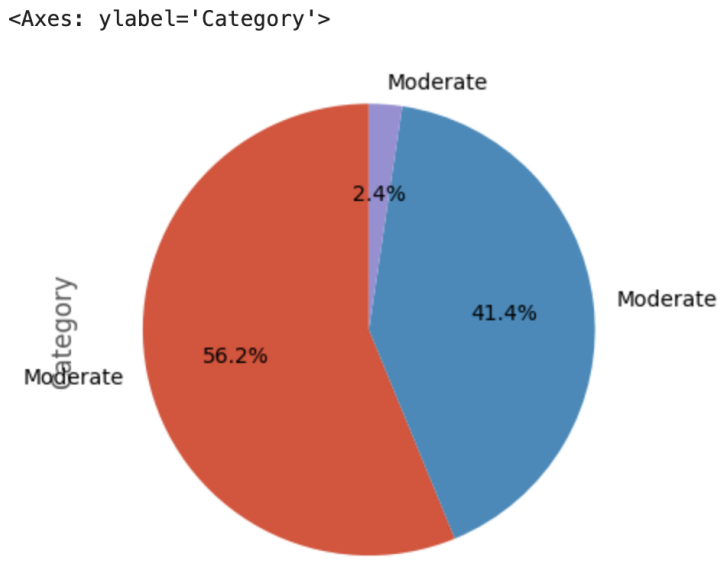

Figure 4: Pie chart showing air quality category distribution

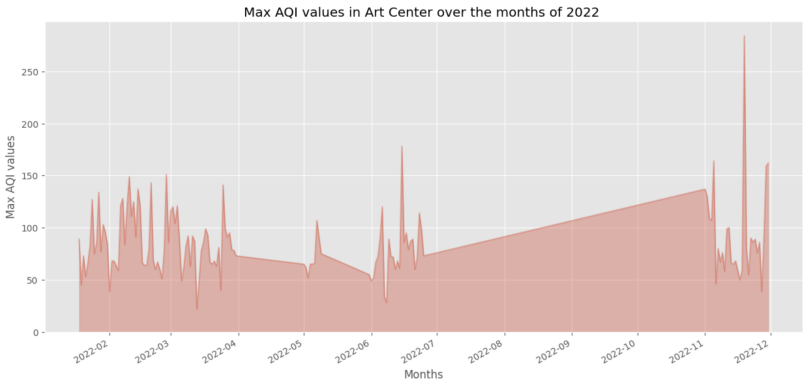

Figure 5: Area plot showing the highest air quality values through out the year

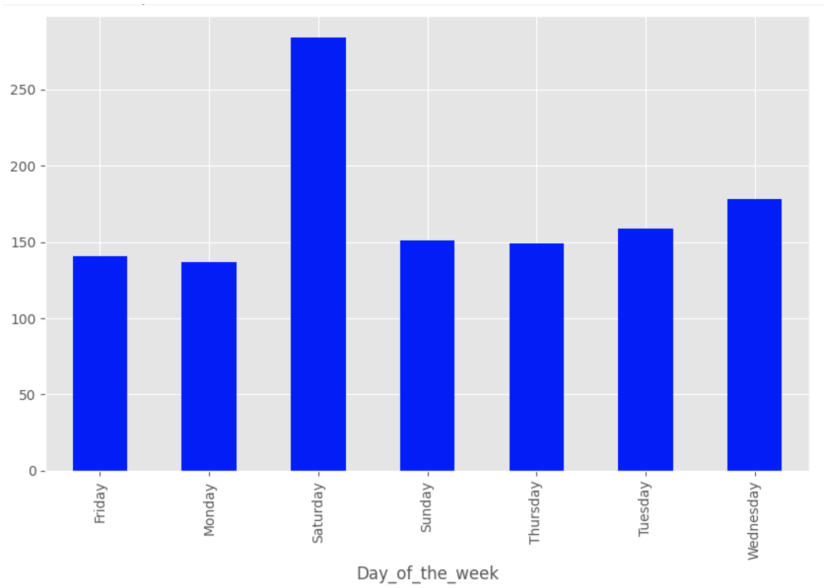

Figure 6: Bar graph showing how air quality varied weekly

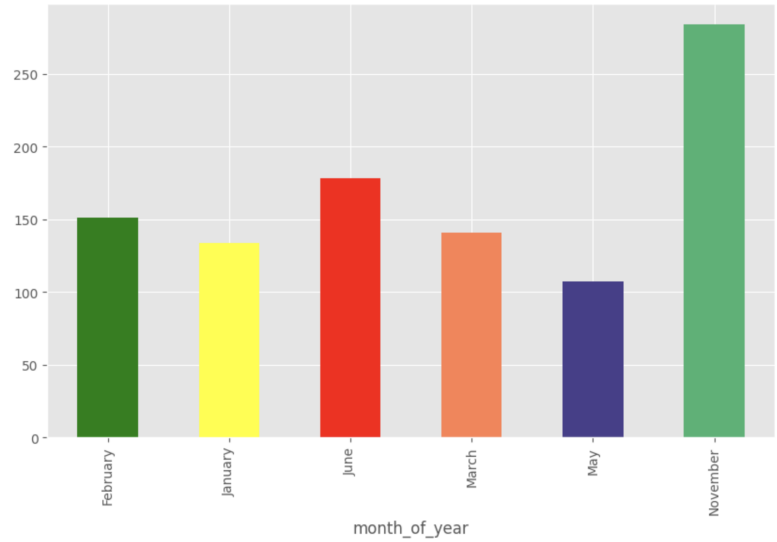

Figure 7: Bar graph showing how air quality varied monthly

	Date	Day_of_the_week	Time_of_the_day	month_of_year
47586	2022-11-19	Saturday	Night	November

Figure 8: Extract of the day of the week, time of the day and month when the highest PM2.5 value occurred

Conclusion

After analysing air quality data from the Art center for the year of 2022 the following conclusions could be drawn. Air quality around the art center can be classified as good because this is the dominant category most of the data points fall in. The month of November had the poorest air quality with a maximum AQI2.5 value of 284 occurring on Saturday 11/19/2022 at night. All graphs plotted to visualize air quality supported that the highest AQI2.5 values occurred on a Saturday in November 2022 which makes the analysis reliable.

Visualising and analysing air quality data like this or in many other ways not explored in this project helps inform scientist and others who are affected by air quality on insights that could drive positive change in the health and wellbeing of humans.

[1]

Bibliography

[1] U EPA. Air quality index: a guide to air quality and your health. *Washington, DC, US Environmental Protection Agency*, 2003.