CL-205 Project: Comparing Air Quality Levels in Two Cities

Shaswat Kumar (210020126)

Contents

1 Introduction

Problem Statement

The objective of this study is to compare air quality levels between two major cities by analyzing Air Quality Index (AQI) data. The goal is to determine whether one city consistently maintains better air quality than the other. By examining AQI values, I aim to assess the average pollution levels, evaluate their variability, and establish if the observed differences are statistically significant.

Quantifiable Analysis

To determine whether one city has better air quality than the other, I performed the following steps:

- Calculated and compared the mean AQI for each city.
- Analyzed the variance and distribution of AQI data in both cities.
- Constructed a 95% confidence interval for the difference in mean AQI levels to determine statistical significance.
- Draw a conclusion on whether one city has significantly better air quality or if no meaningful difference exists.

Availability of Data

The AQI data were sourced from publicly available datasets from Kaggle. The dataset contains daily AQI values, timestamps, and city identifiers. I focused on comparing two cities, New Delhi and Kolkata, over a defined period of 2 years.

Methodology

1. Data Collection
    - Obtained AQI data for both cities (New Delhi and Kolkata) from the selected dataset.
    - Cleaned the dataset by removing any missing or incomplete data entries, ensuring data quality.
2. Data Preprocessing
    - Filtered the AQI data to include only relevant data points.
    - Organised the data into two separate groups—one for each city.
3. Exploratory Data Analysis
    - Plotted histograms of the AQI values for each city to visualise the distribution of AQI levels.
    - Calculated the sample sizes (n1, n2) for both cities, i.e., the number of AQI observations.
    - Computed sample averages and sample variances for the AQI data from both cities.

Statistical Inference

- Using the sample statistics, calculated the difference in mean AQI between the two cities.
- Assuming normal distribution or large sample sizes, computed the 95% confidence interval for the difference between the two population means.

Conclusion

- Examined whether zero is included in the confidence interval. If the interval does not contain zero, I can conclude that there is a statistically significant difference in the AQI levels between the two cities.
- If zero falls within the confidence interval, we conclude that any observed differences in AQI are not statistically significant.

2 Data Analysis

Source

After downloading data from the specified source, I obtained two CSV files containing AQI data:

- **Daily AQI Data:** Includes daily AQI values for multiple cities.
- **Hourly AQI Data:** Includes hourly AQI values for multiple cities.

Overview

Below are sample entries from the two files used in this study:

Daily AQI Data

Shows sample entries from the daily data file:

City	Datetime	PM2.5	NO	CO	SO2	O3	AQI	AQI Bucket
Kolkata	2018-11-06	65.23	52.23	41.32	2.14	28.32	112.0	Moderate
Ahmedabad	2015-01-29	83.13	28.71	6.93	49.52	59.76	209.0	Poor
Delhi	2015-01-21	159.54	12.27	9.01	7.05	18.56	338.0	Very Poor
Patna	2016-01-14	412.87	42.07	4.23	13.69	13.99	0.66	Severe
Bhopal	2020-06-16	14.33	3.65	12.29	17.27	0.49	90.0	Satisfactory

Table 1: Sample data points from the Daily AQI Data file.

Hourly AQI Data

Shows sample entries from the hourly data file:

City	Date & time	PM2.5	NO	CO	SO2	AQI	AQI Bucket
Ahmedabad	2015-02-02 22:00:00	186.53	48.7	1.02	5.12	1000.0	Severe
Ahmedabad	2015-02-02 23:00:00	224.6	41.39	2.83	2.13	1000.0	Severe
Ahmedabad	2015-02-03 00:00:00	244.33	34.64	2.30	2.77	1000.0	Severe
Ahmedabad	2015-02-03 01:00:00	260.6	37.68	3.53	4.03	1000.0	Severe
Ahmedabad	2015-02-03 02:00:00	248.77	35.65	1.88	0.98	1000.0	Severe

Table 2: Sample data points from the Hourly AQI Data file.

Preprocessing

Presenting the data cleaning methods used (with python codes).

- **Data Cleaning:**
  Removed missing values and unnecessary columns

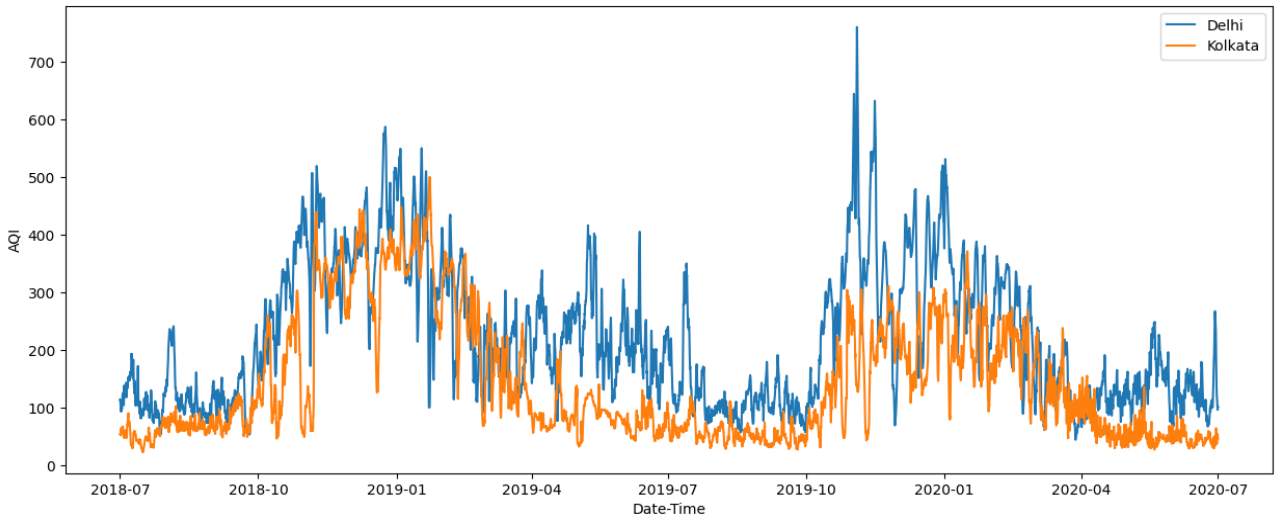
```
# Removing rows with missing AQI values and unnecessary columns
data = data.dropna(subset=['AQI'])
data = data[['City', 'Date', 'AQI']]
```

- **Data Transformation:**
  Selected two cities (Delhi and Kolkata) for comparison within a shared timeframe

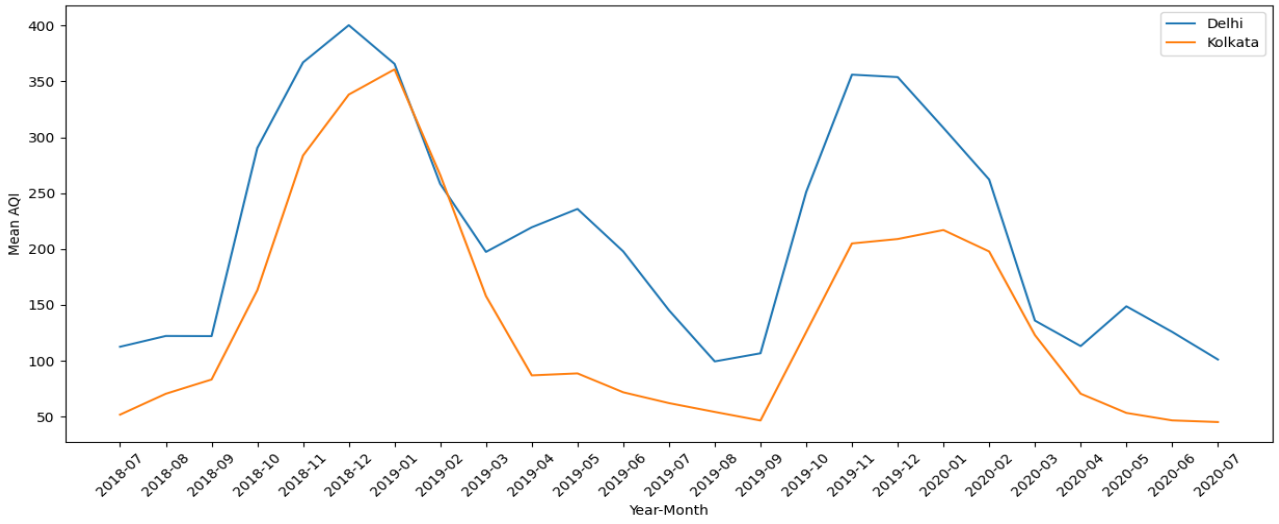
```
# Filter for cities and set a common timeframe
cities = ['Delhi', 'Kolkata']
data = data[data['City'].isin(cities)]
data['Date'] = pd.to_datetime(data['Date'])

# Separate data for each city
Delhi_data = data[data['City'] == 'Delhi']
Kolkata_data = data[data['City'] == 'Kolkata']

# Limit Delhi data to the desired timeframe
Delhi_data=Delhi_data[Delhi_data['Date']>='2018-05-07 00:00:00']
```

Plots

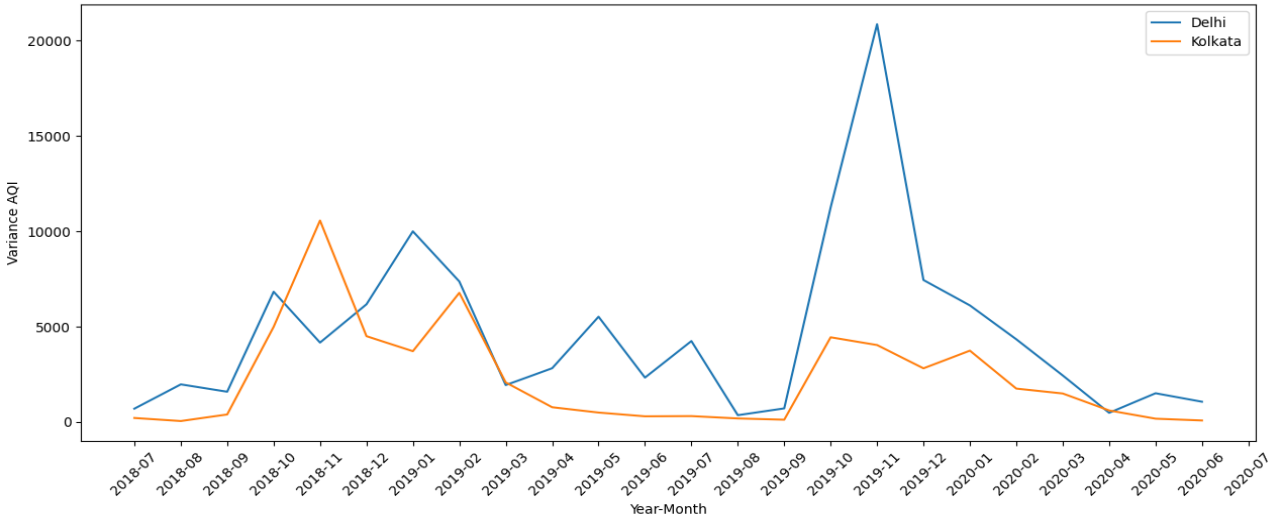
- Plot of daily AQI values for both cities



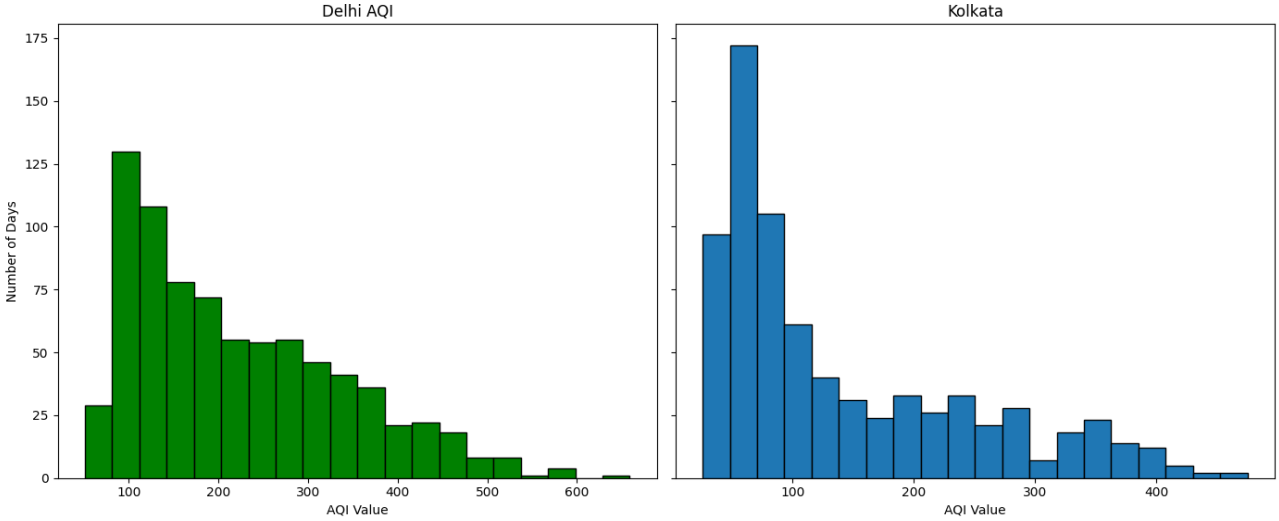
- Plot of monthly mean AQI values for both cities



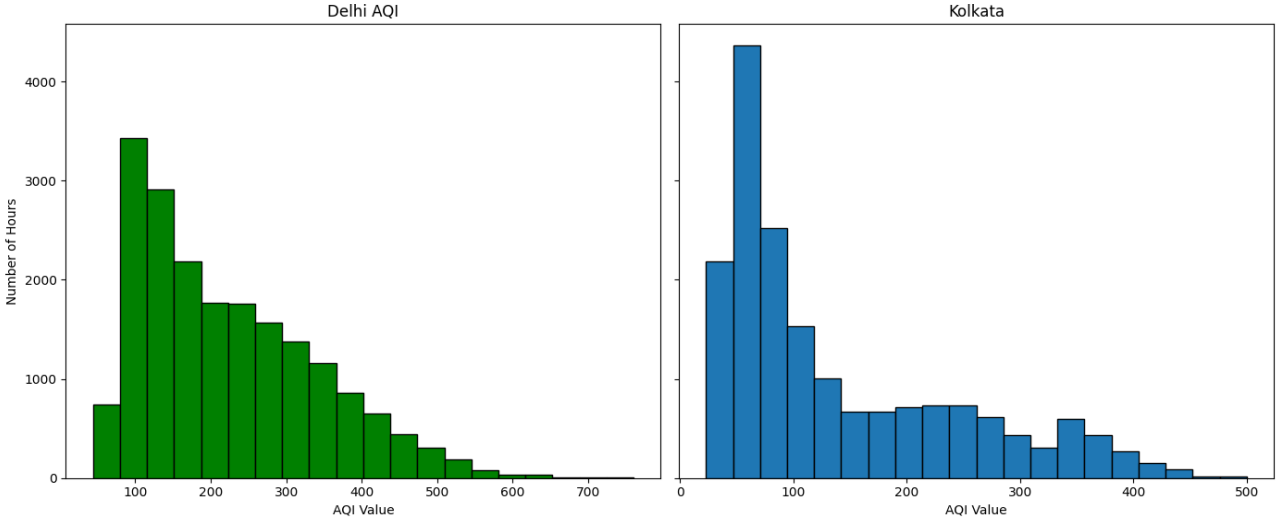
- Plot of monthly variance AQI values for both cities



- Plot of number of days vs observed AQI values



- Plot of number of hours vs observed AQI values



3 Approach

Available Approaches

With two datasets available—one containing daily AQI values and the other with hourly values—I evaluated which would better suit this study. I plotted AQI against the frequency of observations (days and hours) for Delhi over one month to assess distribution and data density.

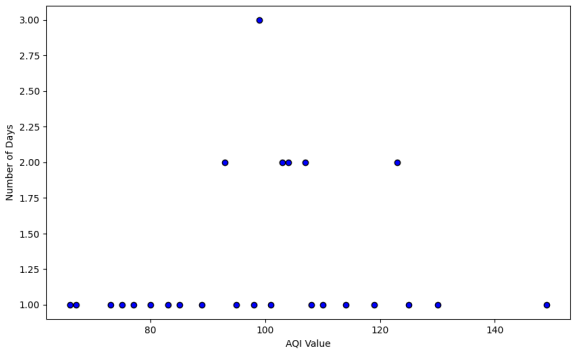


Figure 1: Daily AQI Values for Delhi

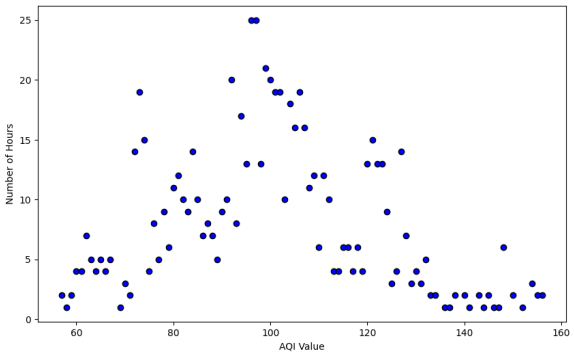


Figure 2: Hourly AQI Values for Delhi

Conclusion

The hourly AQI data exhibited a clearer Gaussian distribution and provided a higher data density than the daily dataset, making it more suitable for robust sampling. Consequently, I selected the hourly AQI dataset for further analysis. I planned to analyze hourly AQI values over a two-year span (July 1, 2018, to July 1, 2020), dividing data into 12 samples of two-month intervals.

Methodology

I began by calculating the mean and variance of each sample for both cities, creating a table with these values.

Table 3: Mean of hourly AQI values for Delhi and Kolkata

City	Mean											
Delhi	117.25	207.67	384.09	314.60	208.27	217.04	122.24	180.13	355.05	286.00	124.62	144.00
Kolkata	61.00	123.64	311.39	315.70	122.84	80.21	57.98	86.73	206.92	207.63	97.02	69.0

Table 4: Variance of hourly AQI values for Delhi and Kolkata

City	Variance											
Delhi	1406.15	11446.14	5903.83	12152.18	2725.03	4852.26	2892.99	11333.01	14436.58	6067.91	1722.40	1444.38
Kolkata	246.30	4423.74	8147.41	7503.88	2813.29	559.73	299.19	3959.34	3573.60	3148.87	1981.91	1243.83

Calculation of sample mean and variance

To calculate the sample mean and variance for each city, I used the following formulas:

$$\overline{X}_{\text{City}} = \frac{1}{12}\sum_{i=1}^{12}\overline{X}_i$$

$$s^2_{\text{City}} = \frac{\sum_{i=1}^{12}(n-1)s_i^2}{(12-1)\cdot n} \quad (n = 30)$$

Using these calculations, I obtained:

$$\overline{X}_{\text{Delhi}} = 221.75, \quad s^2_{\text{Delhi}} = 22490083.44$$
$$\overline{X}_{\text{Kolkata}} = 145.01, \quad s^2_{\text{Kolkata}} = 6796891.94$$

Differences in mean and standard error of difference

Next, I calculated the difference in means and the standard error of the difference using:

$$D = \overline{X}_{\text{Delhi}} - \overline{X}_{\text{Kolkata}}$$

$$SE = \sqrt{\frac{s^2_{\text{Delhi}}}{N_1} + \frac{s^2_{\text{Kolkata}}}{N_2}} \quad (N_1 = N_2 = 12 * 30 = 360)$$

The results were:

$$D = 76.74 \quad SE = 285.22$$

Creating 95% Confidence Interval

Finally, to construct a 95% Confidence Interval for the mean difference, I used:

$$CI = D \pm Z_{\alpha/2} \cdot SE$$

Using $Z_{\alpha/2} \approx 1.96$, the confidence interval is:

$$[-482.28, 635.76]$$

4 Conclusion

Based on the 95% confidence interval for the mean difference in AQI between Delhi and Kolkata, we observe that the interval includes zero. This implies there is **no statistically significant difference** in the average AQI levels between the two cities at the 95% confidence level. Thus, we cannot conclusively state that one city has consistently better or worse air quality than the other over the analyzed period.
Datasets used, all plots and python code is available here.