Course Project Report

STA 2101: Statistics & Probability

Project Title : Dhaka AQ

Student Name : Md. Shahriar Alam
Student ID : 242014180
University of Liberal Arts Bangladesh (ULAB)

October 14, 2025

Abstract

This project analyzes the link between weather and air quality in Dhaka. It uses the "Dhaka Daily Air Quality and Weather" dataset. The study identifies key weather factors that impact the Air Quality Index (AQI). The analysis applies the statistical and probability concepts of STA 2101.

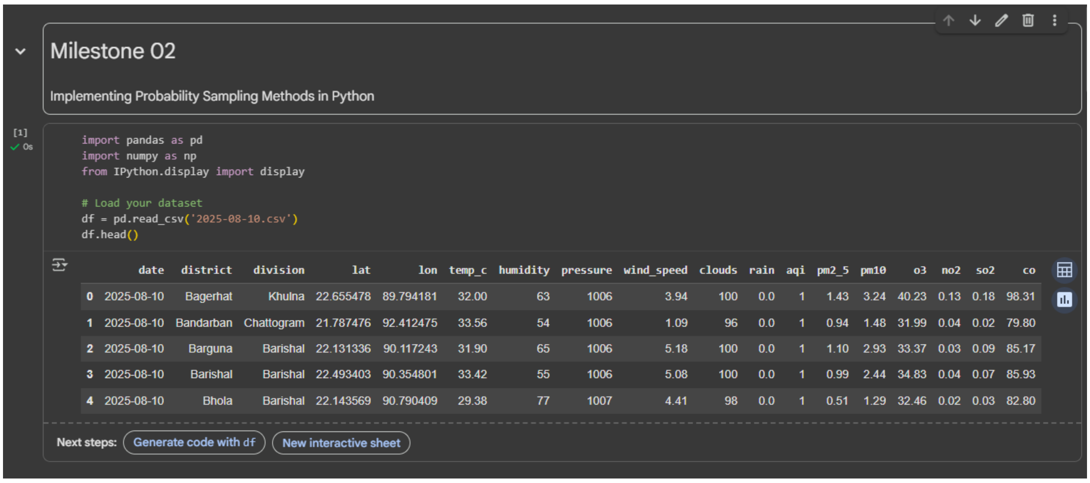
Contents

1 Milestone 1: Dataset Selection

- **Dataset Name :** Dhaka Daily Air Quality and Weather Dataset
- **Dataset URL :** `https://www.kaggle.com/datasets/albab12/dhaka-daily-air-quality-and-weather-dataset`
- **Description :** This project uses the "Dhaka Daily Air Quality & Weather" dataset. The dataset provides daily records for Dhaka, Bangladesh. It contains two main types of information: air quality and weather. The air quality data includes the Air Quality Index (AQI). It also measures several pollutants. These pollutants include PM2.5, PM10, nitrogen dioxide, ozone, carbon monoxide, and sulfur dioxide. The weather data includes daily temperature. It also has information on humidity, barometric pressure, and wind speed.

  This dataset was chosen because it provides comprehensive variables for both air quality and weather. This makes it ideal for studying the relationship between these factors in Dhaka using Statistics & Probability concept.

2 Milestone 02: Probability Sampling Methods



Milestone 02

Implementing Probability Sampling Methods in Python

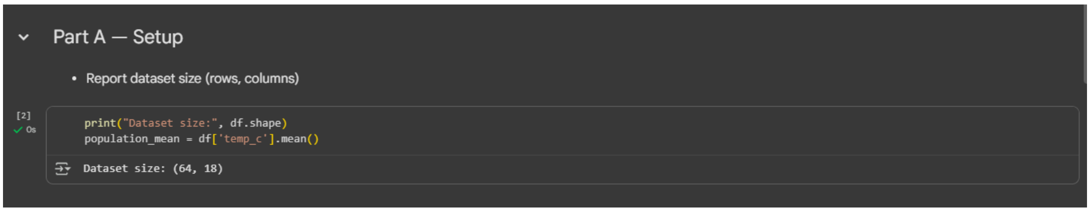
[1]

```
import pandas as pd
import numpy as np
from IPython.display import display

# Load your dataset
df = pd.read_csv('2025-08-10.csv')
df.head()
```

	date	district	division	lat	lon	temp_c	humidity	pressure	wind_speed	clouds	rain	aqi	pm2_5	pm10	o3	no2	so2	co
0	2025-08-10	Bagerhat	Khulna	22.655478	89.794181	32.00	63	1006	3.94	100	0.0	1	1.43	3.24	40.23	0.13	0.18	98.31
1	2025-08-10	Bandarban	Chattogram	21.787476	92.412475	33.56	54	1006	1.09	96	0.0	1	0.94	1.48	31.99	0.04	0.02	79.80
2	2025-08-10	Barguna	Barishal	22.131336	90.117243	31.90	65	1006	5.18	100	0.0	1	1.10	2.93	33.37	0.03	0.09	85.17
3	2025-08-10	Barishal	Barishal	22.493403	90.354801	33.42	55	1006	5.08	100	0.0	1	0.99	2.44	34.83	0.04	0.07	85.93
4	2025-08-10	Bhola	Barishal	22.143569	90.790409	29.38	77	1007	4.41	98	0.0	1	0.51	1.29	32.46	0.02	0.03	82.80

Figure 1: Overview

Part A — Setup



Part A — Setup

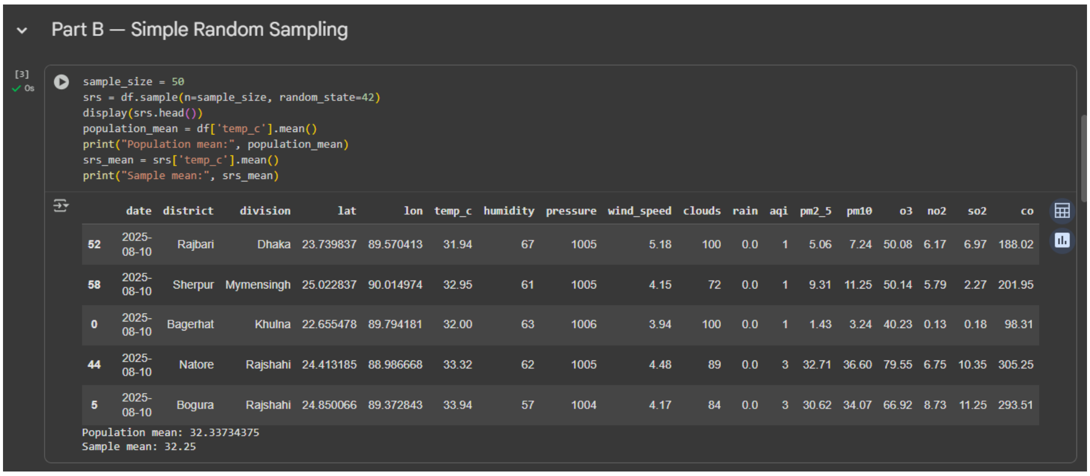
- Report dataset size (rows, columns)

```
print("Dataset size:", df.shape)
population_mean = df['temp_c'].mean()
```

```
Dataset size: (64, 18)
```

Figure 2: Setup

Part B — Simple Random Sampling



Part B — Simple Random Sampling

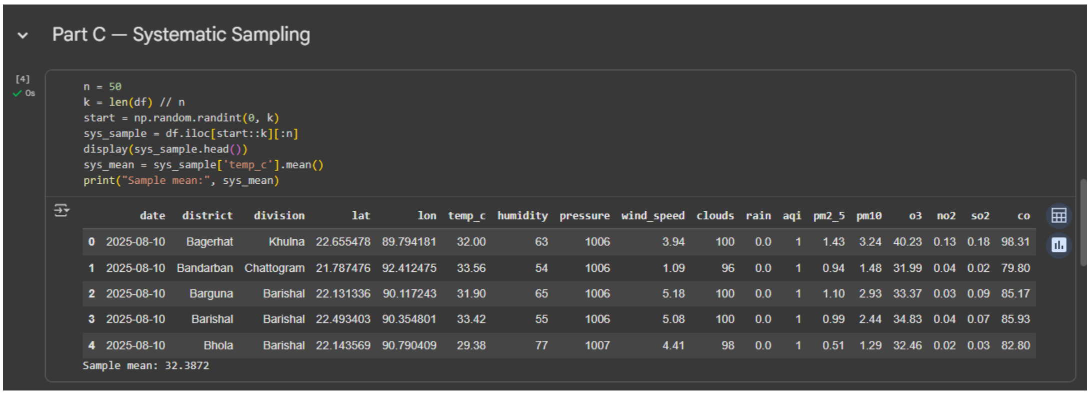
```
sample_size = 50
srs = df.sample(n=sample_size, random_state=42)
display(srs.head())
population_mean = df['temp_c'].mean()
print("Population mean:", population_mean)
srs_mean = srs['temp_c'].mean()
print("Sample mean:", srs_mean)
```

	date	district	division	lat	lon	temp_c	humidity	pressure	wind_speed	clouds	rain	aqi	pm2_5	pm10	o3	no2	so2	co
52	2025-08-10	Rajbari	Dhaka	23.739837	89.570413	31.94	67	1005	5.18	100	0.0	1	5.06	7.24	50.08	6.17	6.97	188.02
58	2025-08-10	Sherpur	Mymensingh	25.022837	90.014974	32.95	61	1005	4.15	72	0.0	1	9.31	11.25	50.14	5.79	2.27	201.95
0	2025-08-10	Bagerhat	Khulna	22.655478	89.794181	32.00	63	1006	3.94	100	0.0	1	1.43	3.24	40.23	0.13	0.18	98.31
44	2025-08-10	Natore	Rajshahi	24.413185	88.986668	33.32	62	1005	4.48	89	0.0	3	32.71	36.60	79.55	6.75	10.35	305.25
5	2025-08-10	Bogura	Rajshahi	24.850066	89.372843	33.94	57	1004	4.17	84	0.0	3	30.62	34.07	66.92	8.73	11.25	293.51

```
Population mean: 32.33734375
Sample mean: 32.25
```

Figure 3: Simple Random Sampling

Part C — Systematic Sampling



Part C — Systematic Sampling

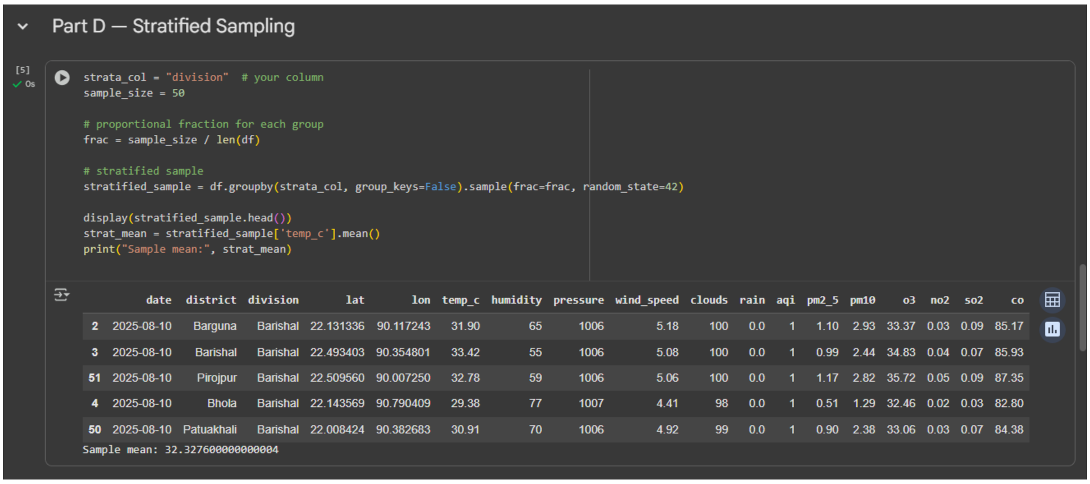
```
n = 50
k = len(df) // n
start = np.random.randint(0, k)
sys_sample = df.iloc[start::k][:n]
display(sys_sample.head())
sys_mean = sys_sample['temp_c'].mean()
print("Sample mean:", sys_mean)
```

	date	district	division	lat	lon	temp_c	humidity	pressure	wind_speed	clouds	rain	aqi	pm2_5	pm10	o3	no2	so2	co
0	2025-08-10	Bagerhat	Khulna	22.655478	89.794181	32.00	63	1006	3.94	100	0.0	1	1.43	3.24	40.23	0.13	0.18	98.31
1	2025-08-10	Bandarban	Chattogram	21.787476	92.412475	33.56	54	1006	1.09	96	0.0	1	0.94	1.48	31.99	0.04	0.02	79.80
2	2025-08-10	Barguna	Barishal	22.131336	90.117243	31.90	65	1006	5.18	100	0.0	1	1.10	2.93	33.37	0.03	0.09	85.17
3	2025-08-10	Barishal	Barishal	22.493403	90.354801	33.42	55	1006	5.08	100	0.0	1	0.99	2.44	34.83	0.04	0.07	85.93
4	2025-08-10	Bhola	Barishal	22.143569	90.790409	29.38	77	1007	4.41	98	0.0	1	0.51	1.29	32.46	0.02	0.03	82.80

```
Sample mean: 32.3872
```

Figure 4: Systematic Sampling

Part D — Stratified Sampling



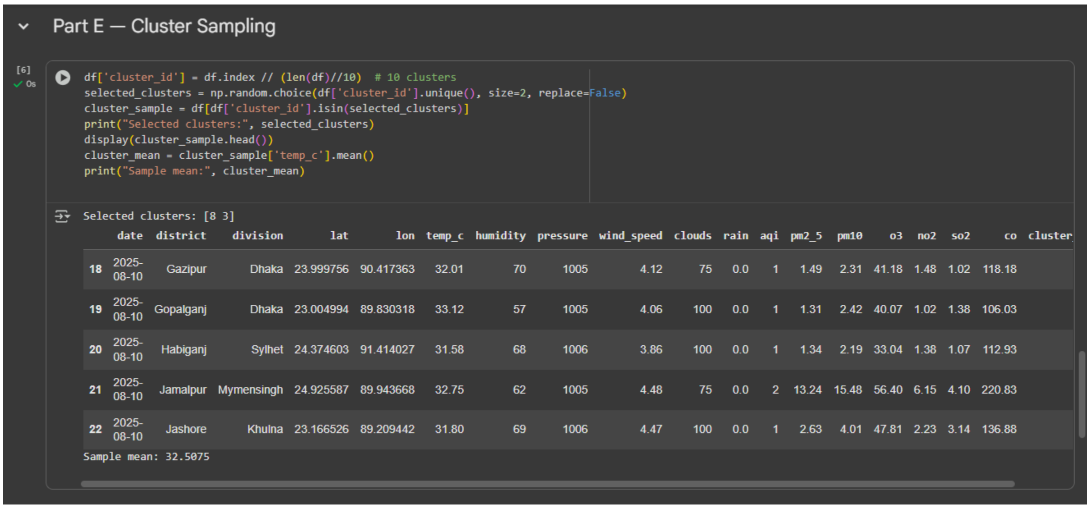
```python
strata_col = "division"  # your column
sample_size = 50

# proportional fraction for each group
frac = sample_size / len(df)

# stratified sample
stratified_sample = df.groupby(strata_col, group_keys=False).sample(frac=frac, random_state=42)

display(stratified_sample.head())
strat_mean = stratified_sample['temp_c'].mean()
print("Sample mean:", strat_mean)
```

	date	district	division	lat	lon	temp_c	humidity	pressure	wind_speed	clouds	rain	aqi	pm2_5	pm10	o3	no2	so2	co
2	2025-08-10	Barguna	Barishal	22.131336	90.117243	31.90	65	1006	5.18	100	0.0	1	1.10	2.93	33.37	0.03	0.09	85.17
3	2025-08-10	Barishal	Barishal	22.493403	90.354801	33.42	55	1006	5.08	100	0.0	1	0.99	2.44	34.83	0.04	0.07	85.93
51	2025-08-10	Pirojpur	Barishal	22.509560	90.007250	32.78	59	1006	5.06	100	0.0	1	1.17	2.82	35.72	0.05	0.09	87.35
4	2025-08-10	Bhola	Barishal	22.143569	90.790409	29.38	77	1007	4.41	98	0.0	1	0.51	1.29	32.46	0.02	0.03	82.80
50	2025-08-10	Patuakhali	Barishal	22.008424	90.382683	30.91	70	1006	4.92	99	0.0	1	0.90	2.38	33.06	0.03	0.07	84.38

```
Sample mean: 32.327600000000004
```

Figure 5: Stratified Sampling

Part E — Cluster Sampling



```python
df['cluster_id'] = df.index // (len(df)//10)  # 10 clusters
selected_clusters = np.random.choice(df['cluster_id'].unique(), size=2, replace=False)
cluster_sample = df[df['cluster_id'].isin(selected_clusters)]
print("Selected clusters:", selected_clusters)
display(cluster_sample.head())
cluster_mean = cluster_sample['temp_c'].mean()
print("Sample mean:", cluster_mean)
```

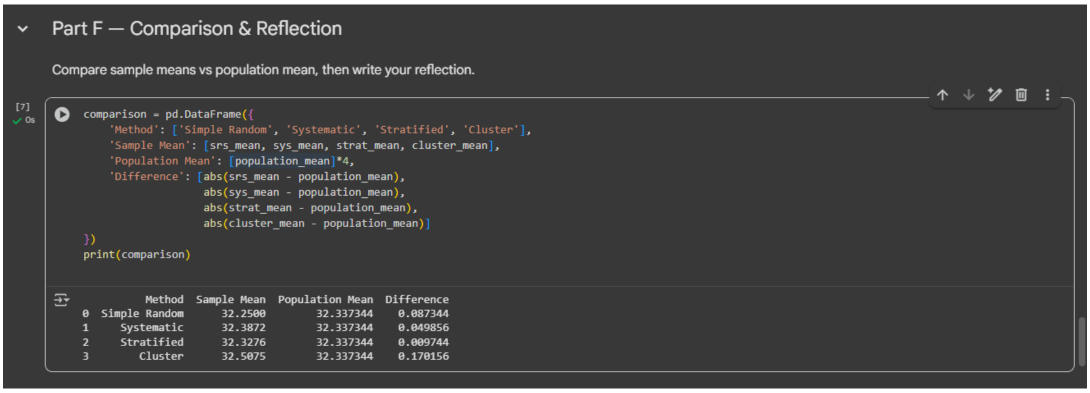
```
Selected clusters: [8 3]
```

	date	district	division	lat	lon	temp_c	humidity	pressure	wind_speed	clouds	rain	aqi	pm2_5	pm10	o3	no2	so2	co	cluster
18	2025-08-10	Gazipur	Dhaka	23.999756	90.417363	32.01	70	1005	4.12	75	0.0	1	1.49	2.31	41.18	1.48	1.02	118.18	
19	2025-08-10	Gopalganj	Dhaka	23.004994	89.830318	33.12	57	1005	4.06	100	0.0	1	1.31	2.42	40.07	1.02	1.38	106.03	
20	2025-08-10	Habiganj	Sylhet	24.374603	91.414027	31.58	68	1006	3.86	100	0.0	1	1.34	2.19	33.04	1.38	1.07	112.93	
21	2025-08-10	Jamalpur	Mymensingh	24.925587	89.943668	32.75	62	1005	4.48	75	0.0	2	13.24	15.48	56.40	6.15	4.10	220.83	
22	2025-08-10	Jashore	Khulna	23.166526	89.209442	31.80	69	1006	4.47	100	0.0	1	2.63	4.01	47.81	2.23	3.14	136.88	

```
Sample mean: 32.5075
```

Figure 6: Cluster Sampling

Part F — Comparison & Reflection



Part F — Comparison & Reflection

Compare sample means vs population mean, then write your reflection.

[7]
0s

```
comparison = pd.DataFrame({
    'Method': ['Simple Random', 'Systematic', 'Stratified', 'Cluster'],
    'Sample Mean': [srs_mean, sys_mean, strat_mean, cluster_mean],
    'Population Mean': [population_mean]*4,
    'Difference': [abs(srs_mean - population_mean),
                   abs(sys_mean - population_mean),
                   abs(strat_mean - population_mean),
                   abs(cluster_mean - population_mean)]
})
print(comparison)
```

```
          Method  Sample Mean  Population Mean  Difference
0  Simple Random      32.2500        32.337344    0.087344
1     Systematic      32.3872        32.337344    0.049856
2     Stratified      32.3276        32.337344    0.009744
3        Cluster      32.5075        32.337344    0.170156
```

Figure 7: Comparison and Reflection

In this milestone, I used four probability sampling methods. The population mean was 32.337344. All methods gave sample means close to it but not the same. Stratified sampling gave the closest mean 32.3276. The difference was very small. This happened because it kept the same group ratio as the dataset. Simple random sampling gave 32.25 which is a bit lower than the population mean. Systematic sampling gave 32.3872 which is slightly higher. Cluster sampling gave 32.5075 which is the farthest from the population mean.

Simple random sampling was the easiest. I wrote only one line of code and got the result. It did not need any group or pattern. Systematic sampling was also easy. I just needed k and a start point. Stratified sampling was harder. I had to use a division column and take samples from each group by proportion. Cluster sampling was easy to code but tricky to pick clusters.

Each method works for different goals. Simple random sampling is good for small or mixed data. Systematic sampling is good when data has no clear order. Stratified sampling is best for datasets with clear groups. Cluster sampling is useful for large data that is grouped by place or type. From this, I saw stratified sampling gave the most accurate result. Simple random sampling was the easiest to use.

3 Milestone 3: Data Visualization

Add graphs and figures using LaTeX. Example:

4 Milestone 4: Probability Distributions

Identify probability distributions in your dataset. Perform fitting, plots, and discuss results.

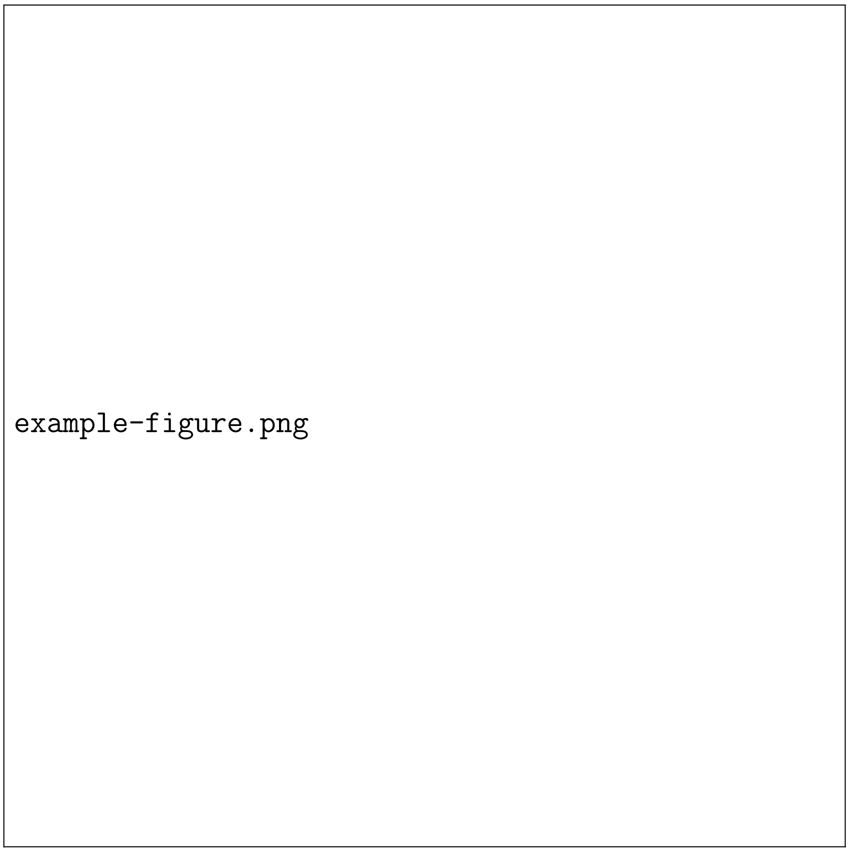


Figure 8: Sample dataset visualization (replace with your figure)

5 Milestone 5: Hypothesis Testing

State hypotheses, perform tests, and report conclusions.

6 Milestone 6: Regression Analysis

Fit regression models, explain coefficients, and evaluate model fit.

7 Milestone 7–12: Further Analysis

Continue documenting each milestone here as instructed in class.

8 Final Conclusion

Summarize the overall findings of your project. Mention challenges, learning outcomes, and possible future work.

References

List your references here in proper citation format. If you prefer, you may use BibTeX.