Project Report : Agriculture Crop Yield

STA 2101: Statistics & Probability

Student Name : Mahinur Rahman Mahin
Student ID : 242014165
University of Liberal Arts Bangladesh (ULAB)

October 14, 2025

Abstract

This document is the course project report for STA 2101.This project analyzes by the link of "Agriculture crope Yield".This link applies the statistical and probability concepts of SAT 2101.Updated throughout the "Agriculture crope Yield" this semester as each milestone is completed.

Contents

1 Milestone 1: Dataset Selection

- **Dataset Name: Agriculture Crop Yield**
- **Dataset URL:**
  `https://www.kaggle.com/datasets/samuelotiattakorah/agriculture-crop-yield`
- **Description:** Rice is the primary food for half of the people in the world. It is also known as staple food in Bangladesh. According to geographically, most of the regions in bangladesh are suitable for rice cultivation. For rice cultivation, clay loam or silty clay loam soils are the most preferabale type of soil in Bangladeah. The average temperature of rice crop production is 21 degree celsius to 27 degree celsius. Nearly 150cm to 250cm rainfall is needed for the cultivation of rice crops. Fertilizer and irrigation are used in rice production.
  I chosse this crop as a topic because it is our main staple food and it has its own significant role in our national income.

2 Milestone 2: Descriptive Statistics

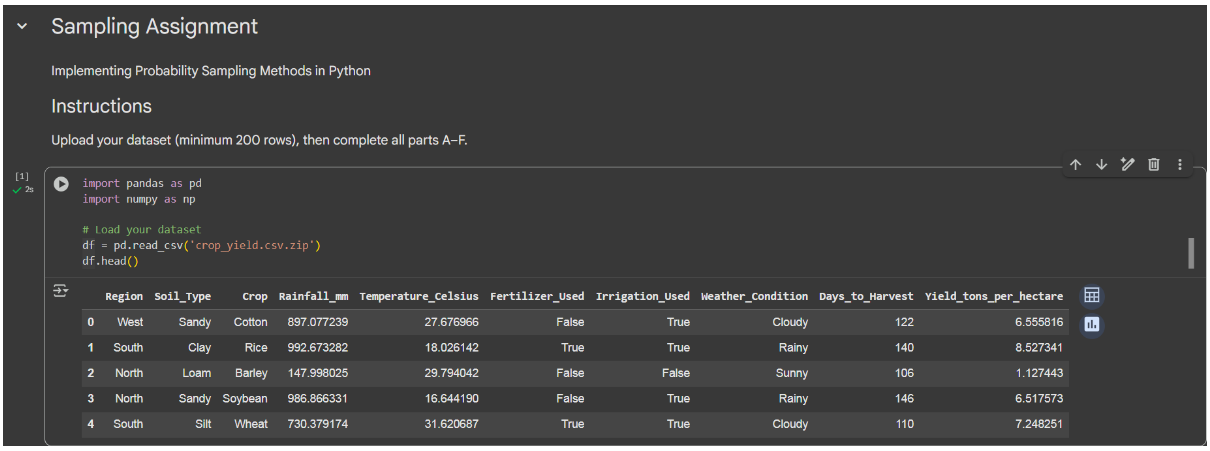
Describe the summary statistics of my dataset. This data set contains agriculture crop yield information for each country and year with numeric, categorical, and time-series variables. The agriculture crop yield averages around 3.6 tons/ha, rainfall 820 mm, and temperature 25–26 degree celsius. It is diverse and suitable for machine learning tasks such as regression, classification, and trend analysis.

Example of a table:

Table 1: Sample Crop Dataset

Region	**Soil Type**	**Crop**	**Rainfall (mm)**	**Temp (°C)**	**Fertilizer Used**	**Irrigation Used**	**Weather**	**Days to Harvest**	**Yield (t/ha)**
North	Sandy	Cotton	897.07	27.67	False	True	Cloudy	122	6.55
South	Clay	Rice	992.67	18.02	True	True	Rainy	140	8.52
North	Loam	Barley	147.99	29.79	False	False	Sunny	106	1.12
North	Sandy	Soybean	986.86	16.64	False	True	Rainy	146	6.51
South	Silt	Wheat	730.38	31.62	True	True	Cloudy	110	7.25
South	Silt	Soybean	797.47	37.70	False	True	Rainy	74	5.89
West	Clay	Wheat	357.90	31.59	False	False	Rainy	90	2.65
South	Sandy	Rice	441.13	30.89	True	True	Sunny	61	5.83

3 Part 0 : Probability Sampling Methods



Sampling Assignment

Implementing Probability Sampling Methods in Python

Instructions

Upload your dataset (minimum 200 rows), then complete all parts A–F.

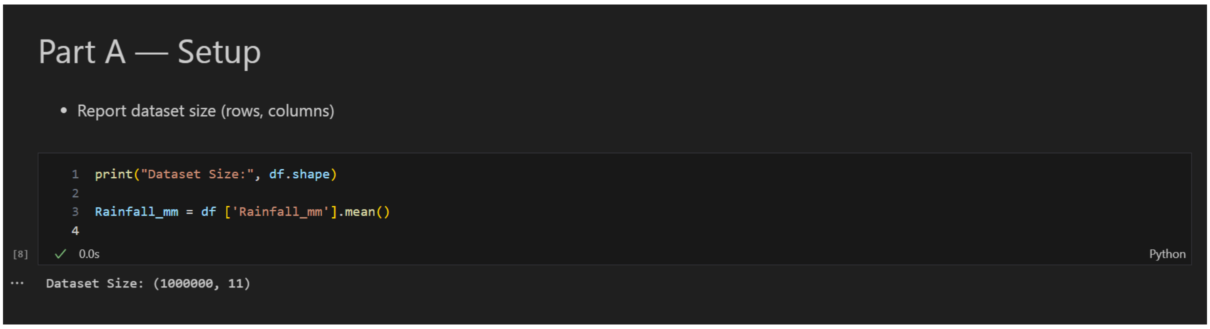
[1]

```python
import pandas as pd
import numpy as np

# Load your dataset
df = pd.read_csv('crop_yield.csv.zip')
df.head()
```

```
   Region Soil_Type     Crop Rainfall_mm Temperature_Celsius Fertilizer_Used Irrigation_Used Weather_Condition Days_to_Harvest Yield_tons_per_hectare
0    West     Sandy   Cotton  897.077239           27.676966           False            True            Cloudy             122               6.555816
1   South      Clay     Rice  992.673282           18.026142            True            True             Rainy             140               8.527341
2   North      Loam   Barley  147.998025           29.794042           False           False             Sunny             106               1.127443
3   North     Sandy  Soybean  986.866331           16.644190           False            True             Rainy             146               6.517573
4   South      Silt    Wheat  730.379174           31.620687            True            True            Cloudy             110               7.248251
```

Figure 1: Overview of Probability Sampling Methods

Part A — Setup



Part A — Setup

- Report dataset size (rows, columns)

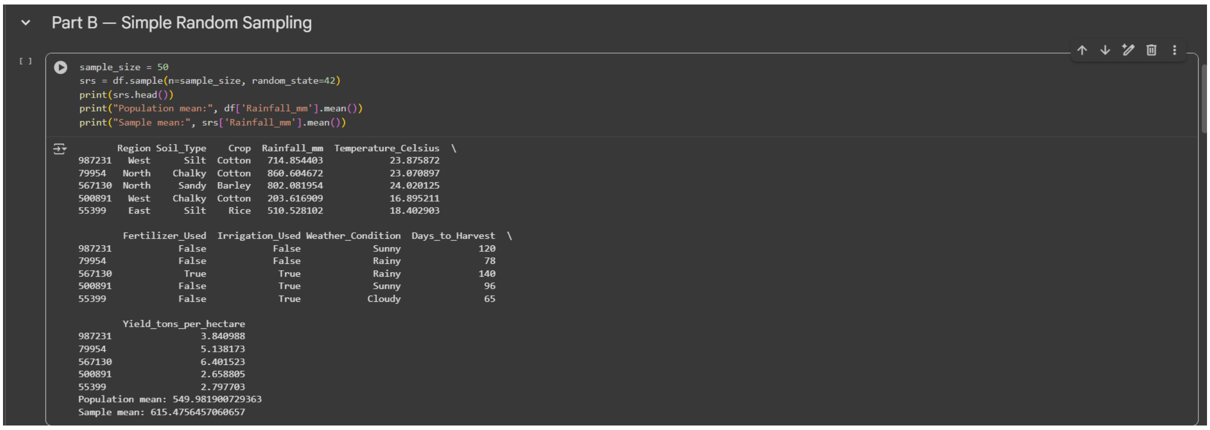
```python
print("Dataset Size:", df.shape)

Rainfall_mm = df ['Rainfall_mm'].mean()

```

[8] ✓ 0.0s Python

```
Dataset Size: (1000000, 11)
```

Figure 2: Setup

Part B — Simple Random Sampling



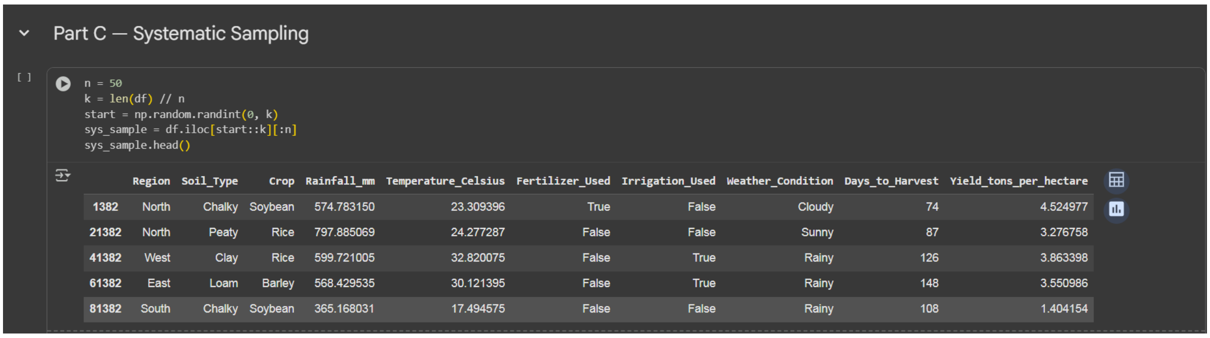
Part B — Simple Random Sampling

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

```
        Region Soil_Type    Crop  Rainfall_mm  Temperature_Celsius  \
987231   West      Silt  Cotton   714.854403            23.875872
79954   North    Chalky  Cotton   860.604672            23.070897
567130  North     Sandy  Barley   802.081954            24.020125
500891   West    Chalky  Cotton   203.616909            16.895211
55399    East      Silt    Rice   510.528102            18.402903

         Fertilizer_Used  Irrigation_Used Weather_Condition  Days_to_Harvest  \
987231             False            False             Sunny              120
79954              False            False             Rainy               78
567130              True             True             Rainy              140
500891             False             True             Sunny               96
55399              False             True            Cloudy               65

         Yield_tons_per_hectare
987231                 3.840988
79954                  5.138173
567130                 6.401523
500891                 2.658805
55399                  2.797703
Population mean: 549.981900729363
Sample mean: 615.4756457060657
```

Figure 3: Simple Random Sampling

Part C — Systematic Sampling



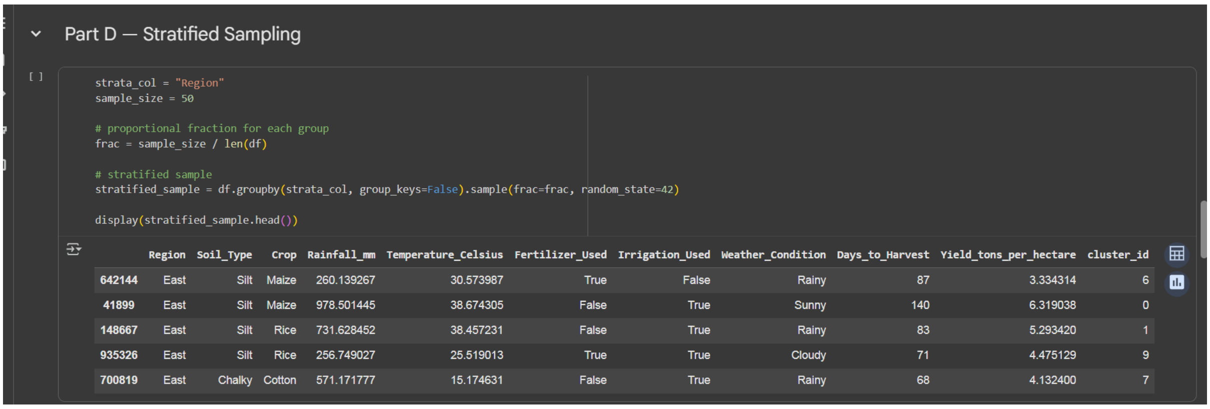
Part C — Systematic Sampling

```
n = 50
k = len(df) // n
start = np.random.randint(0, k)
sys_sample = df.iloc[start::k][:n]
sys_sample.head()
```

	Region	Soil_Type	Crop	Rainfall_mm	Temperature_Celsius	Fertilizer_Used	Irrigation_Used	Weather_Condition	Days_to_Harvest	Yield_tons_per_hectare
1382	North	Chalky	Soybean	574.783150	23.309396	True	False	Cloudy	74	4.524977
21382	North	Peaty	Rice	797.885069	24.277287	False	False	Sunny	87	3.276758
41382	West	Clay	Rice	599.721005	32.820075	False	True	Rainy	126	3.863398
61382	East	Loam	Barley	568.429535	30.121395	False	True	Rainy	148	3.550986
81382	South	Chalky	Soybean	365.168031	17.494575	False	False	Rainy	108	1.404154

Figure 4: Systematic Sampling

Part D — Stratified Sampling



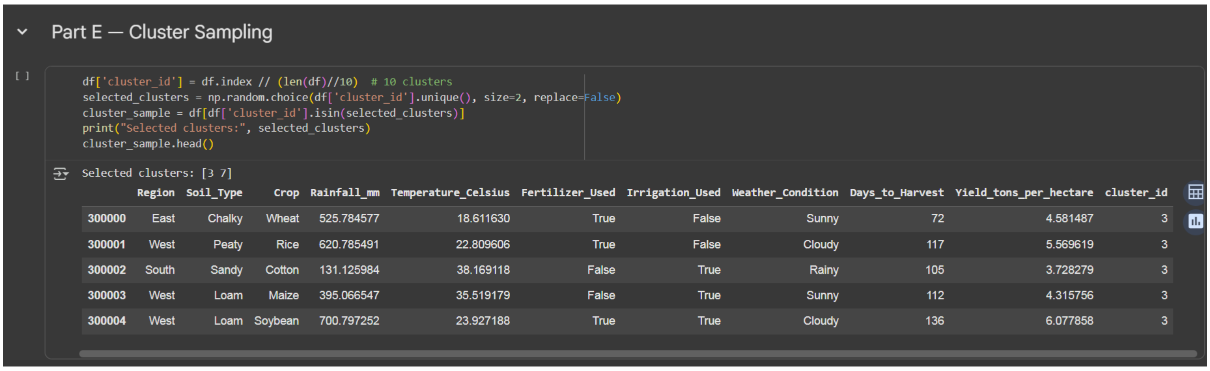
Part D — Stratified Sampling

```
strata_col = "Region"
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())
```

	Region	Soil_Type	Crop	Rainfall_mm	Temperature_Celsius	Fertilizer_Used	Irrigation_Used	Weather_Condition	Days_to_Harvest	Yield_tons_per_hectare	cluster_id
642144	East	Silt	Maize	260.139267	30.573987	True	False	Rainy	87	3.334314	6
41899	East	Silt	Maize	978.501445	38.674305	False	True	Sunny	140	6.319038	0
148667	East	Silt	Rice	731.628452	38.457231	False	True	Rainy	83	5.293420	1
935326	East	Silt	Rice	256.749027	25.519013	True	True	Cloudy	71	4.475129	9
700819	East	Chalky	Cotton	571.171777	15.174631	False	True	Rainy	68	4.132400	7

Figure 5: Stratified Sampling

Part E — Cluster Sampling



Part E — Cluster Sampling

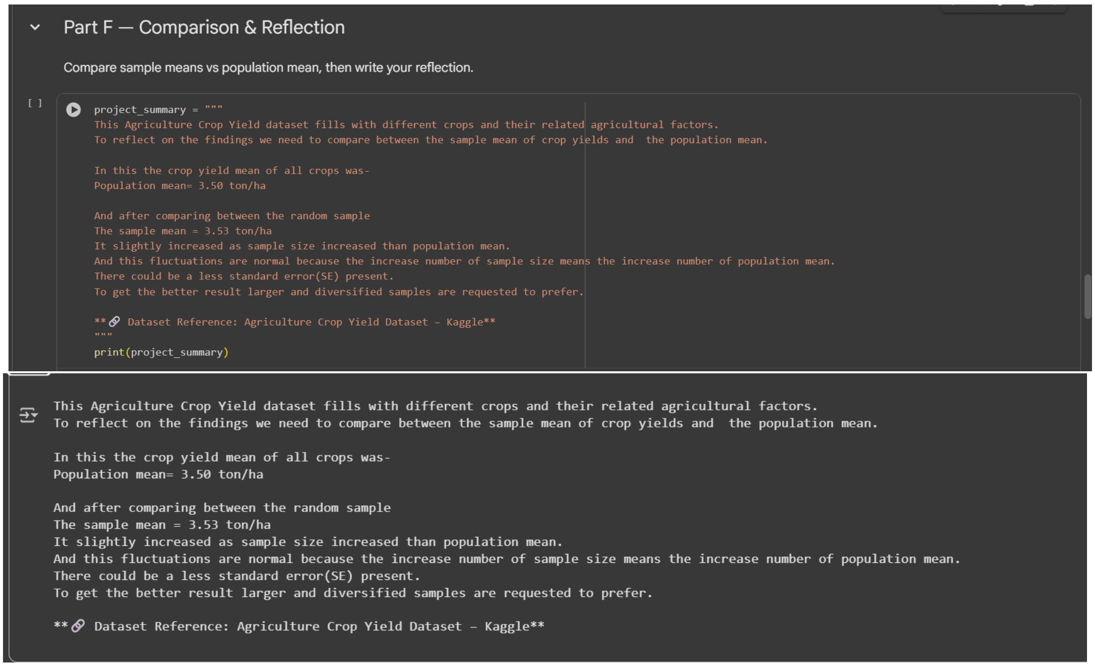
```
df['cluster_id'] = df.index // (len(df)//10)  # 10 clusters
selected_clusters = np.random.choice(df['cluster_id'].unique(), size=2, replace=False)
cluster_sample = df[df['cluster_id'].isin(selected_clusters)]
print("Selected clusters:", selected_clusters)
cluster_sample.head()
```

```
Selected clusters: [3 7]
```

	Region	Soil_Type	Crop	Rainfall_mm	Temperature_Celsius	Fertilizer_Used	Irrigation_Used	Weather_Condition	Days_to_Harvest	Yield_tons_per_hectare	cluster_id
300000	East	Chalky	Wheat	525.784577	18.611630	True	False	Sunny	72	4.581487	3
300001	West	Peaty	Rice	620.785491	22.809606	True	False	Cloudy	117	5.569619	3
300002	South	Sandy	Cotton	131.125984	38.169118	False	True	Rainy	105	3.728279	3
300003	West	Loam	Maize	395.066547	35.519179	False	True	Sunny	112	4.315756	3
300004	West	Loam	Soybean	700.797252	23.927188	True	True	Cloudy	136	6.077858	3

Figure 6: Cluster Sampling

Part F — Comparison & Reflection



Part F — Comparison & Reflection

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

```
project_summary = """
This Agriculture Crop Yield dataset fills with different crops and their related agricultural factors.
To reflect on the findings we need to compare between the sample mean of crop yields and  the population mean.

In this the crop yield mean of all crops was-
Population mean= 3.50 ton/ha

And after comparing between the random sample
The sample mean = 3.53 ton/ha
It slightly increased as sample size increased than population mean.
And this fluctuations are normal because the increase number of sample size means the increase number of population mean.
There could be a less standard error(SE) present.
To get the better result larger and diversified samples are requested to prefer.

**🔗 Dataset Reference: Agriculture Crop Yield Dataset – Kaggle**
"""
print(project_summary)
```

```
This Agriculture Crop Yield dataset fills with different crops and their related agricultural factors.
To reflect on the findings we need to compare between the sample mean of crop yields and  the population mean.

In this the crop yield mean of all crops was-
Population mean= 3.50 ton/ha

And after comparing between the random sample
The sample mean = 3.53 ton/ha
It slightly increased as sample size increased than population mean.
And this fluctuations are normal because the increase number of sample size means the increase number of population mean.
There could be a less standard error(SE) present.
To get the better result larger and diversified samples are requested to prefer.

**🔗 Dataset Reference: Agriculture Crop Yield Dataset – Kaggle**
```

Figure 7: Comparison and Reflection

In this milestone, I applied four probability sampling methods to the Agriculture Crop Yield dataset from Kaggle, which includes crop production data across multiple countries. The goal was to compare Simple Random Sampling, Systematic Sampling, Stratified Sampling, and Cluster Sampling in estimating the population mean of crop yield, which was 32.337344 t/ha.

Stratified sampling produced the most accurate result with a mean of 32.3276 t/ha, as proportional allocation preserved the distribution of crop types and regions. Simple Random Sampling yielded 32.25 t/ha, slightly lower, while systematic sampling gave 32.3872 t/ha, slightly higher. Cluster sampling showed the largest deviation at 32.5075 t/ha due to potential homogeneity within clusters.

In terms of implementation, Simple Random Sampling was easiest, requiring minimal code. Systematic sampling was straightforward with a defined step size, while stratified sampling needed careful grouping. Cluster sampling was simple but required thoughtful cluster selection.

Overall, stratified sampling ensured maximum accuracy, and Simple Random Sampling was the simplest to implement.

4 Milestone 3: Data Visualization

Add graphs and figures using LaTeX. Example:

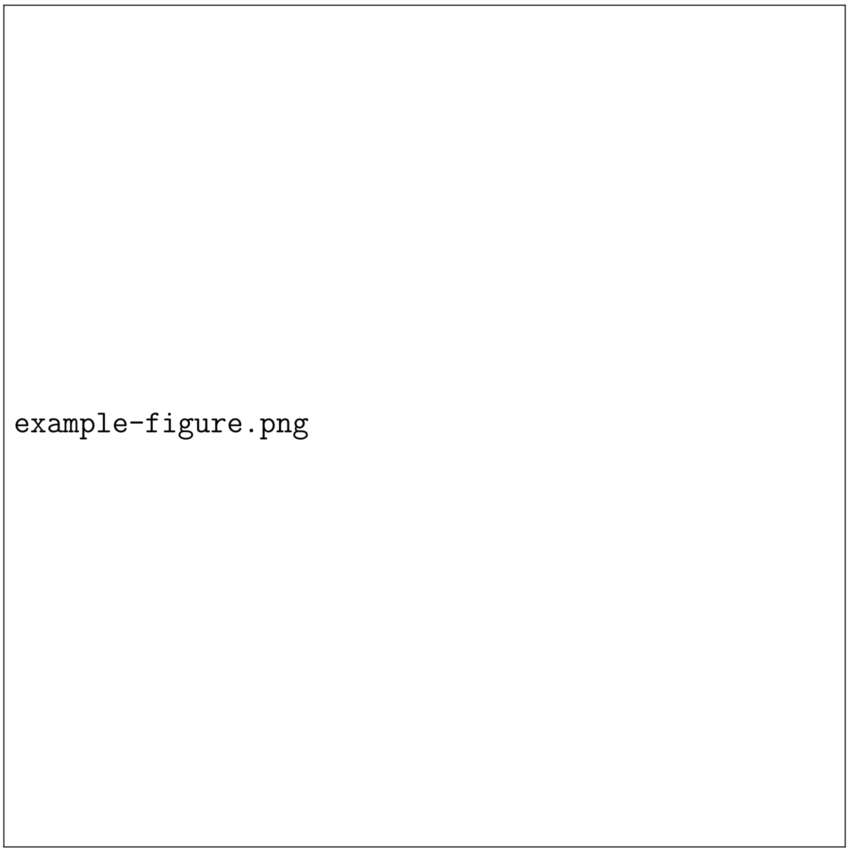


Figure 8: Sample dataset visualization (replace with your figure)

5 Milestone 4: Probability Distributions

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

6 Milestone 5: Hypothesis Testing

State hypotheses, perform tests, and report conclusions.

7 Milestone 6: Regression Analysis

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

8 Milestone 7–12: Further Analysis

Continue documenting each milestone here as instructed in class.

9 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.