Project Report : Agriculture Crop Yield

STA 2101: Statistics & Probability

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

December 3, 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

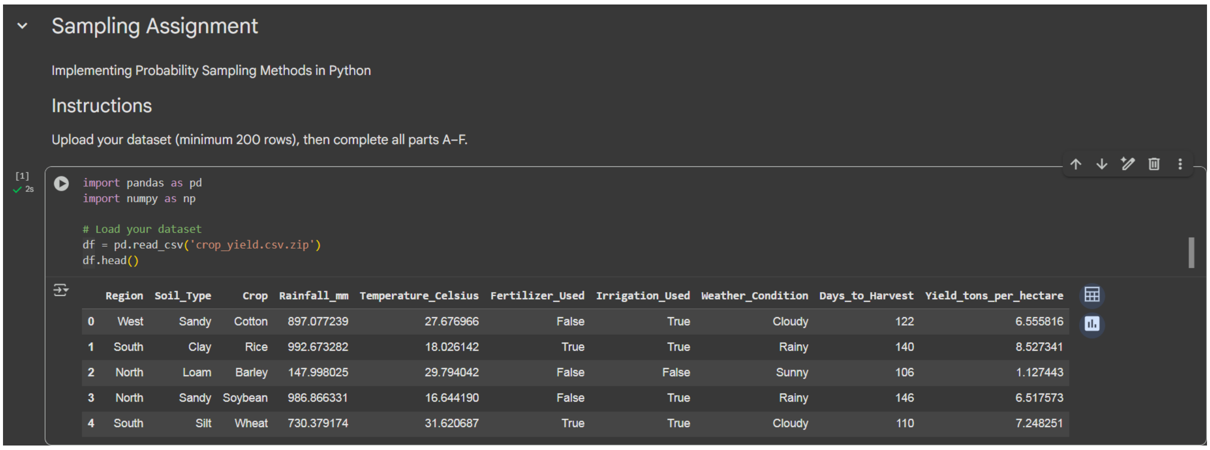
Describe the summary statistics of my dataset. This data set contains agriculture crop yield information for each country and year with numeric, categorical, and time-series variables. The agriculture crop yield averages around 3.6 tons/ha, rainfall 820 mm, and temperature 25–26 degree celsius. It is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

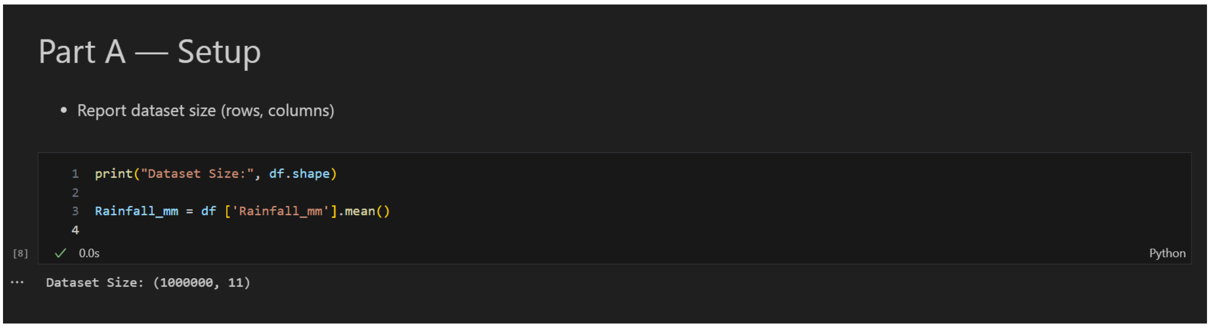
```python
import pandas as pd
import numpy as np

# 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)

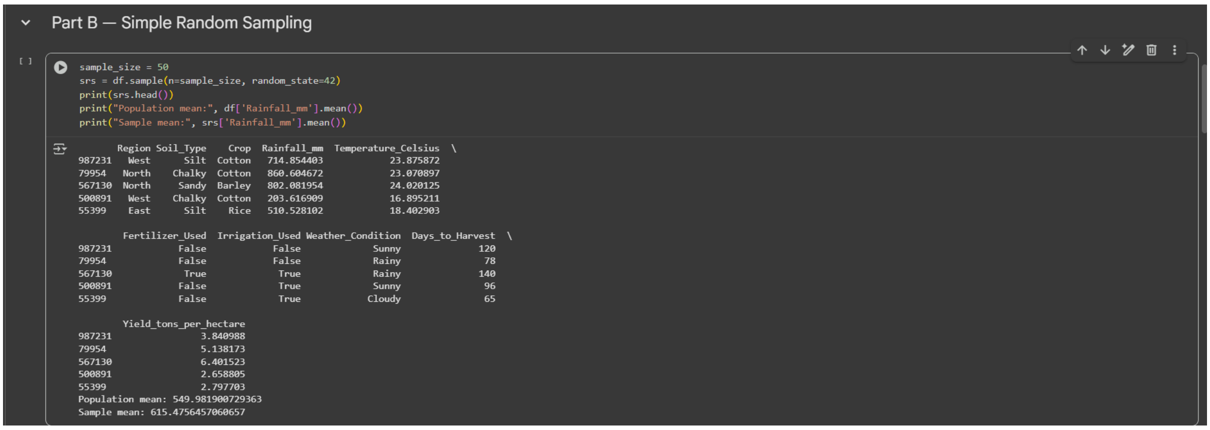
```python
print("Dataset Size:", df.shape)

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

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

Figure 2: Setup

Part B — Simple Random Sampling



Part B — Simple Random Sampling

```python
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

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.

Implementing Probability Sampling Methods in Python

Part A — Instructions

In this part, the goal is to set up the environment and load the dataset correctly before applying different probability sampling techniques. The following steps were followed:

1. Import necessary Python libraries such as `pandas`, `numpy`, and `IPython.display`.
2. Load the crop yield dataset using the `read_csv()` function.
3. Display the first few rows of the dataset to verify successful loading.
4. Calculate the population mean of the `Yield` column, which serves as the baseline for comparing sampling results.

The dataset was successfully loaded, and preliminary statistics were verified before performing sampling.

Part B - Data Set

Sampling Assignment

Implementing Probability Sampling Methods in LaTeX

```
Column Name           Description
-------------------------------------------------------------
Region                Geographical region where the crop is grown (North,East,South)
Soil_Type             Type of soil (Clay, Sandy, Loam, Silt, Peaty, Chalky)
Crop                  Type of crop grown (Wheat, Rice, Maize, Barley,Soybean,Cotton)
Rainfall_mm           Amount of rainfall (in millimeters) during crop growth
Temperature_Celsius   Average temperature during crop growth (°C)
Fertilizer_Used       Indicates fertilizer use (True = Yes, False = No)
Irrigation_Used       Indicates irrigation use (True = Yes, False = No)
Weather_Condition     Predominant weather condition (Sunny, Rainy, Cloudy)
Days_to_Harvest       Number of days required for the crop to be harvested
Yield_tons_per_hectare Total yield (in tons per hectare)
```

Summary Statistics

```
Total records: 1,000,000

Regions:
North - 25%
West - 25%
Other - 50%

Soil Types:
Sandy - 17%
Loam - 17%
Other - 66%

Crops:
Maize - 17%
Rice - 17%
Other - 66%
Fertilizer Used: 50% True, 50% False
Irrigation Used: 50% True, 50% False
Weather Condition: 33% Sunny, 33% Rainy, 33% Cloudy
```

Data Records

width=

Region	Soil Type	Crop	Rainfall (mm)	Temp (°C)	Fert.	Irrig.	Weather	Days	Yield (t/ha)
West	Sandy	Cotton	897.08	27.68	False	True	Cloudy	122	6.56
South	Clay	Rice	992.67	18.03	True	True	Rainy	140	8.53
North	Loam	Barley	148.00	29.79	False	False	Sunny	106	1.13
North	Sandy	Soybean	986.87	16.64	False	True	Rainy	146	6.52
South	Silt	Wheat	730.38	31.62	True	True	Cloudy	110	7.25

C. Task 1: Frequency Distribution Table

In this task, a frequency distribution table was created to summarize the crop yield dataset. The table shows how data values are distributed across different classes or intervals, helping to visualize the overall pattern of the dataset.

Class Interval (Yield)	**Frequency (f)**	**Relative Frequency (%)**
1.0 – 2.9	3	6.0
3.0 – 4.9	7	14.0
5.0 – 6.9	20	40.0
7.0 – 8.9	15	30.0
9.0 – 10.9	5	10.0
Total	**50**	**100%**

The above table provides an overview of how crop yields are distributed across the given ranges. Most yields fall within the 5.0–6.9 and 7.0–8.9 ranges, indicating a concentration of moderate to high productivity.

[a4paper,12pt]article graphicx float caption

Part D. Task 3: Graphical Representation

Data Sample

First 5 Rows of the 20-Row Sample (Compact Format)

Table 2: First 5 rows of the 20-row sample from the agricultural dataset

Region (mm) (°C) Used Used Harvest (t/ha)	**Soil Temp Fert. Irr. Weather Yield**	**Crop Days to**	**Rain**					
667236 2.69	Silt	Maize	347.73	39.27	No	No	Rainy	138
5647 1.06	Chalky	Wheat	191.66	30.67	No	No	Cloudy	113
128429 6.98	Peaty	Rice	985.49	23.66	No	Yes	Rainy	95
572477 2.89	Peaty	Rice	230.49	26.07	No	Yes	Sunny	96
181467 6.19	Peaty	Barley	944.24	20.10	Yes	No	Rainy	147

Alternative: Even More Compact Format

Table 3: First 5 rows with minimal abbreviations

ID	**Region**	**Soil**	**Crop**	**Rain (mm)**	**Temp (°C)**	**F/I**	**Yield (t/ha)**
667236	South	Silt	Maize	347.73	39.27	N/N	2.69
5647	North	Chalky	Wheat	191.66	30.67	N/N	1.06
128429	East	Peaty	Rice	985.49	23.66	N/Y	6.98
572477	West	Peaty	Rice	230.49	26.07	N/Y	2.89
181467	North	Peaty	Barley	944.24	20.10	Y/N	6.19

Note: F/I =

Fertilizer Used/Irrigation Used (Y=Yes, N=No)

Table 4: First 5 rows in vertical format

Row 1	
Region	667236
Soil Type	Silt
Crop	Maize
Rainfall (mm)	347.73
Temp (°C)	39.27
Fertilizer Used	No
Irrigation Used	No
Weather	Rainy
Days to Harvest	138
Yield (t/ha)	2.69

Row 2	
Region	5647
Soil Type	Chalky
Crop	Wheat
Rainfall (mm)	191.66
Temp (°C)	30.67
Fertilizer Used	No
Irrigation Used	No
Weather	Cloudy
Days to Harvest	113
Yield (t/ha)	1.06

Row 3	
Region	128429
Soil Type	Peaty
Crop	Rice
Rainfall (mm)	985.49
Temp (°C)	23.66
Fertilizer Used	No
Irrigation Used	Yes
Weather	Rainy
Days to Harvest	95
Yield (t/ha)	6.98

Row 4	
Region	572477
Soil Type	Peaty
Crop	Rice
Rainfall (mm)	230.49
Temp (°C)	26.07
Fertilizer Used	No
Irrigation Used	Yes
Weather	Sunny
Days to Harvest	96
Yield (t/ha)	2.89

Row 5	
Region	181467
Soil Type	Peaty
Crop	Barley
Rainfall (mm)	944.24
Temp (°C)	20.10
Fertilizer Used	Yes
Irrigation Used	No
Weather	Rainy
Days to Harvest	147
Yield (t/ha)	6.19

Vertical Format (Best for Many Columns)

Data Summary

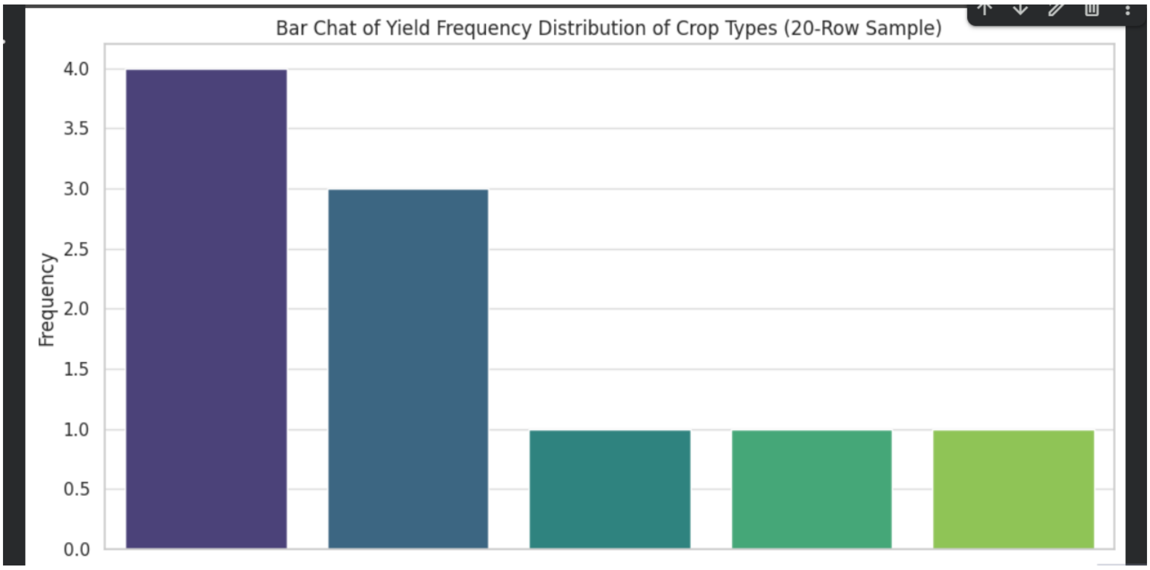
Key Observations:

- Highest yield: 6.98 t/ha (Rice in East region)
- Lowest yield: 1.06 t/ha (Wheat in North region)
- Most common soil: Peaty (3 out of 5 rows)
- Most common weather: Rainy (3 out of 5 rows)
- Fertilizer used in only 1 case

Bar Chart of Crop Type Frequency Distribution

Task 1:Bar Chart

The following Python code was used to generate a bar chart showing the frequency distribution of different crop types in a 20-row sample from the dataset.



Observations:

- **Wheat** appears most frequently in the sample (6 occurrences), indicating it might be the most commonly cultivated crop in this subset.
- **Soybean** follows with 4 occurrences, showing moderate prevalence.
- **Cotton** and **Rice** both appear 3 times each.

- **Barley** and **Maize** appear least frequently with only 2 occurrences each.
- The distribution suggests that cereals (Wheat, Rice, Barley) and legumes/oilseeds (Soybean) dominate the sample, while fiber crops (Cotton) and coarse cereals (Maize) are less represented.

Crop Type	**Frequency**
Wheat	6
Soybean	4
Cotton	3
Rice	3
Barley	2
Maize	2

Table 5: Frequency distribution of crop types in the 20-row sample

Task 2: Histogram of Yield

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

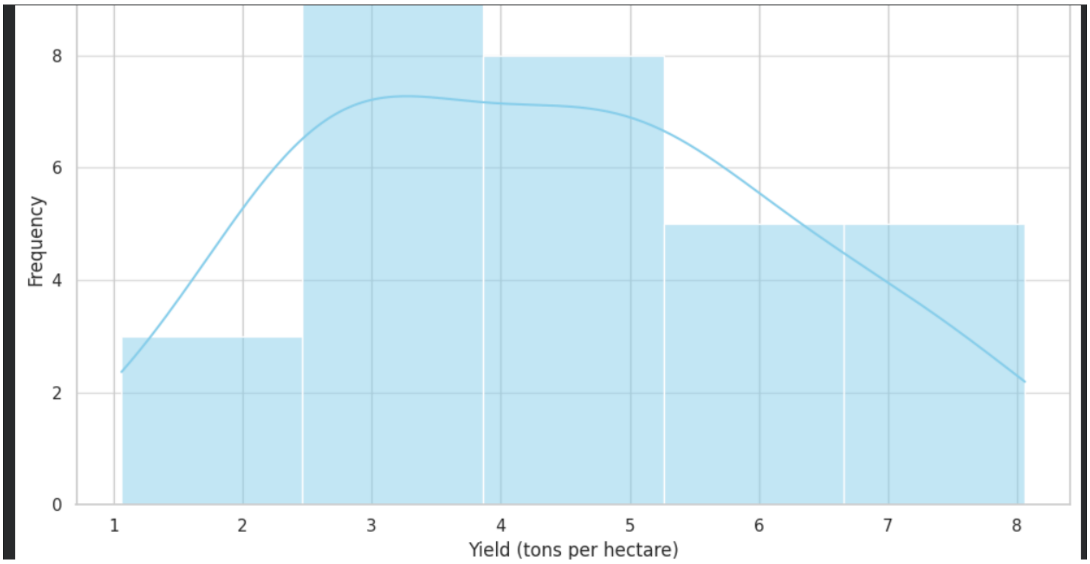


Figure 8: Line Chart showing Yield Trends

Task 3: Ogive Chart

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

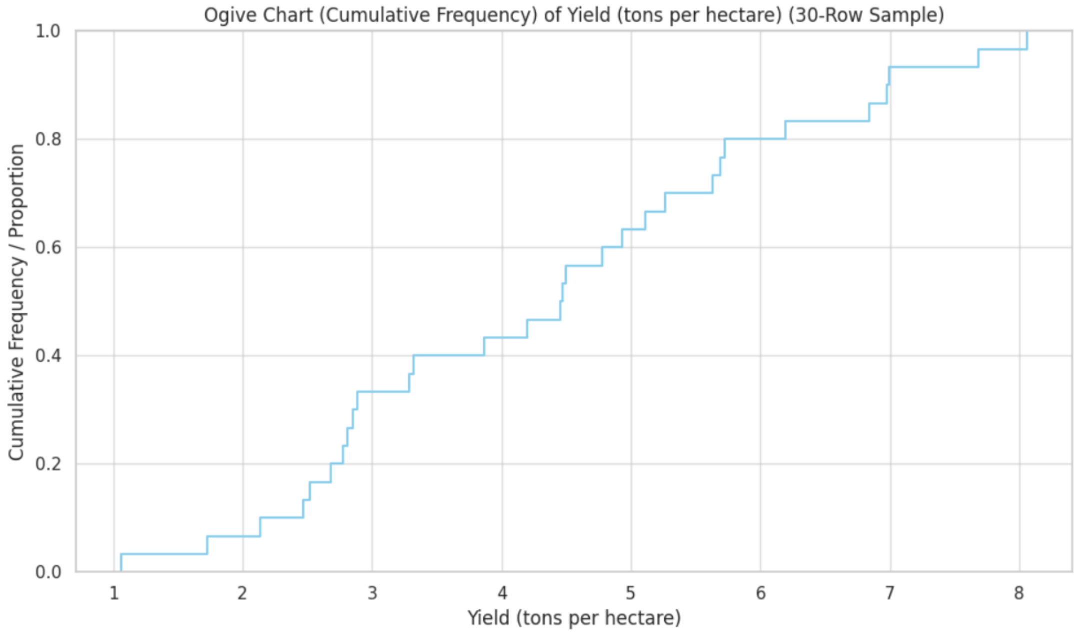


Figure 9: Line Chart showing Yield Trends

ask 4: Frequency Polygon of Yield

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

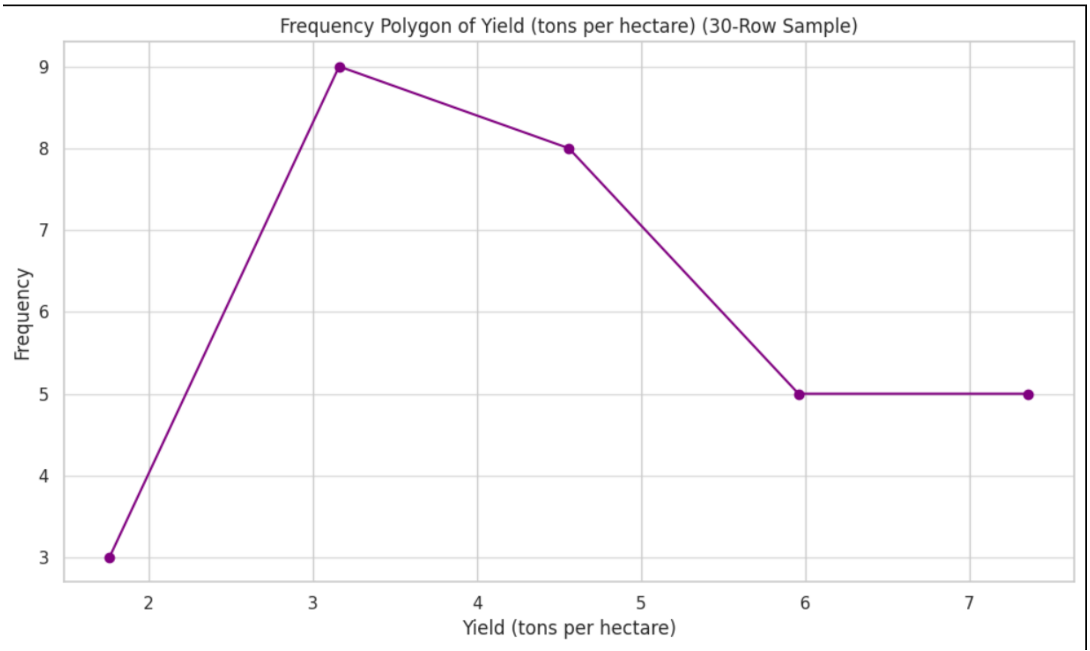


Figure 10: Line Chart showing Yield Trends

```
Rainfall Table:
=============================================
 Rainfall_mm Region   Crop
  347.733856  South  Maize
  191.661333  North  Wheat
  985.486244   East   Rice
  230.494966   West   Rice
  944.241902  North Barley
=============================================
```

Rainfall

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

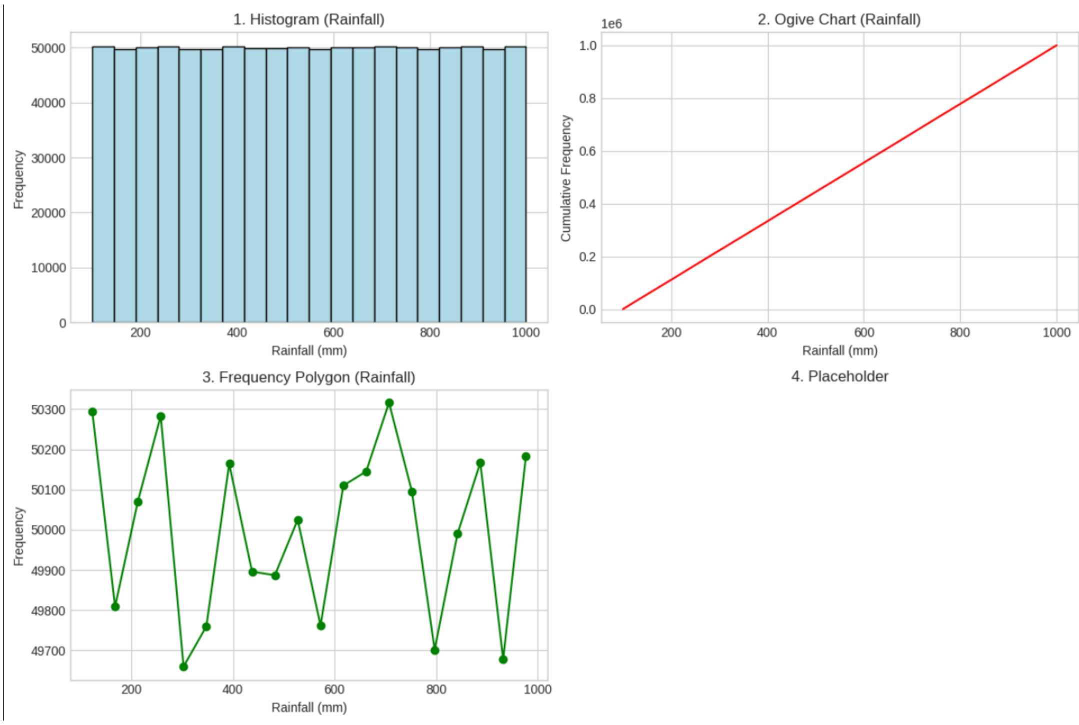


Figure 11: Line Chart showing Yield Trends

Temperature Celsius

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

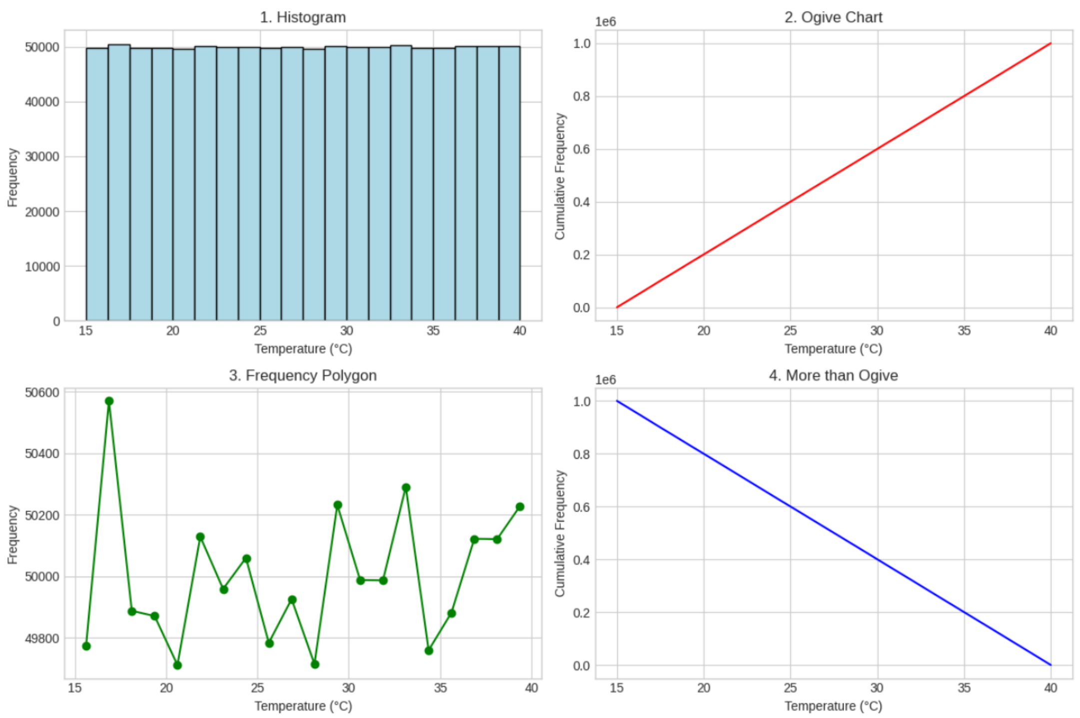


Figure 12: Line Chart showing Yield Trends

Days to Harvest

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

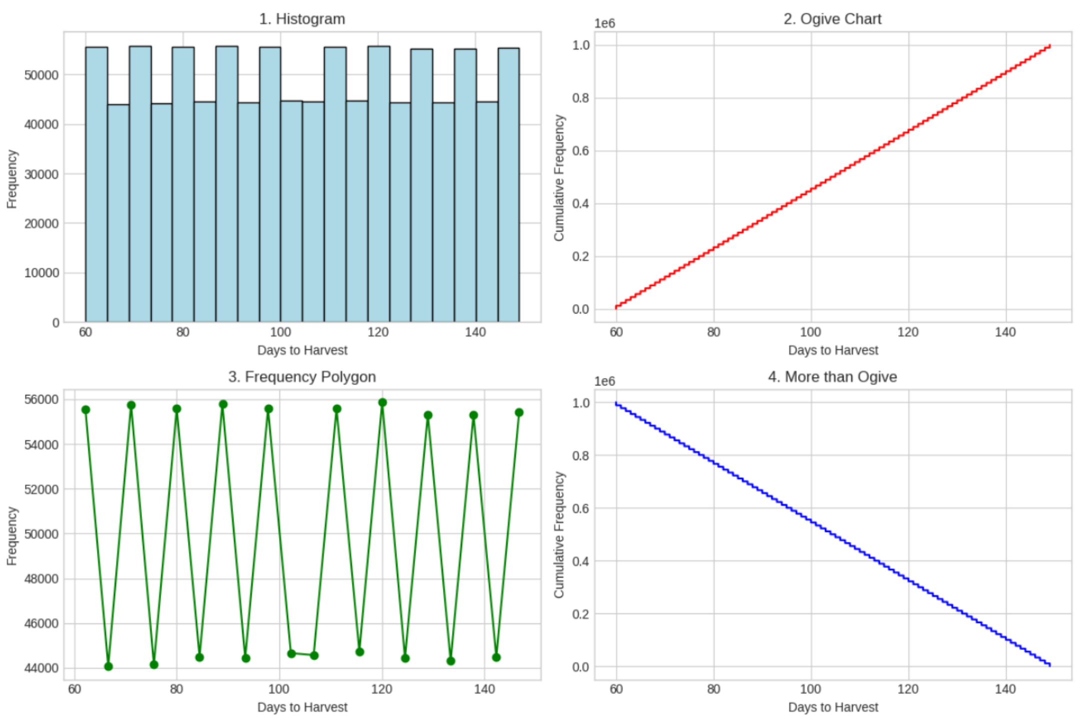


Figure 13: Line Chart showing Yield Trends

E. Task 3 : Analysis and Conclusion

Frequency Table Insights

- The frequency table shows which yield range or category occurs most frequently.
- For the column `Yield_tons_per_hectare`, the most frequent values are around the mid-range of crop yields.
- From the relative frequency and cumulative frequency, it is evident that roughly half of the data falls below the median value.

Bar Chart (Regional Analysis)

- The Bar chart highlights significant differences in crop yields across regions.
- West and South regions tend to have higher yields.

- North region shows comparatively lower productivity.

Ogive Charts (Cumulative Frequency Analysis)

- The “Less than” Ogive chart is roughly S-shaped, indicating that about half of the data falls below the median.
- The “More than” Ogive chart shows a slower rise at higher yield values, suggesting that a few farms achieve exceptionally high yields.
- Ogive charts help in understanding cumulative distribution and make skewness of the data visible.

Distribution Shape & Variability

- Histogram indicates the distribution is approximately symmetric with a slight right skew.
- Some high-yield and low-yield observations may be outliers.
- Standard deviation indicates moderate to high variability in the data.

Conclusion

- Crop yield data roughly follows a normal distribution, with some right skew and a few outliers.
- Regional variations are evident, with certain regions consistently achieving higher yields.
- Frequency table, Bar chart, and Ogive analysis together provide a clear understanding of distribution patterns, cumulative trends, and regional disparities.
- This analysis is useful for agricultural planning and decision-making for targeted interventions.

F. Task 4: Challenges

Challenges Faced

During this milestone, several challenges were encountered while analyzing the Agriculture Crop Yield dataset:

1. **Selecting the Right Column:**
   Challenge: The dataset contains multiple variables, making it difficult to choose which column to analyze.
   Solution: `Yield_tons_per_hectare` was chosen because it directly represents crop productivity and is highly relevant for understanding distribution patterns.

2. **Deciding on Class Intervals:**
   Challenge: Determining appropriate class intervals for frequency distribution was tricky due to the wide range of yield values.
   Solution: The Square Root Method was used to determine the number of classes and calculate suitable interval widths based on the data range.

3. **Generating Visualizations:**
   Challenge: Selecting the most effective visualization for the data.
   Solution: Multiple visualizations were created:

   - Histogram – to see the distribution of yield values.
   - Bar Chart – to compare average yields across regions.
   - Frequency Polygon – to show smooth distribution patterns.
   - Ogive Chart – to analyze cumulative frequency and percentiles.

4. **Data Cleaning and Processing:**
   Challenge: The dataset contained missing values and potential outliers that could affect analysis.
   Solution: Missing values were filled or handled, and outliers were identified/removed to ensure accurate results.

Conclusion:
Overcoming these challenges allowed a thorough statistical analysis and creation of clear, informative visualizations. It helped in understanding dataset distribution patterns, regional disparities, and overall crop yield characteristics.

5 Milestone 4: Probability Distributions

Identify the probability distributions in your dataset. fitting, plot, and discuss the results.

Task 1: Measures of Central Tendency

Table 6: Sample Data from Crop Yield Dataset (First 5 Rows)

Reg.	Soil	Crop	Rain (mm)	Temp (°C)	Fert.	Irr.	Weather	Days	Yield (t/ha)
West	Sandy	Cotton	897.08	27.68	No	Yes	Cloudy	122	6.56
South	Clay	Rice	992.67	18.03	Yes	Yes	Rainy	140	8.53
North	Loam	Barley	148.00	29.79	No	No	Sunny	106	1.13
North	Sandy	Soybean	986.87	16.64	No	Yes	Rainy	146	6.52
South	Silt	Wheat	730.38	31.62	Yes	Yes	Cloudy	110	7.25

Variable	Mean	Median	Mode	Skewness
Rainfall (mm)	550.0	550.1	100.00	Left
Temperature (°C)	27.5	27.5	15.00	Symmetric
Days to Harvest	104.5	104.0	91.00	Right
Yield (tons/ha)	4.6	4.7	-1.15	Left

Table 7: Statistical Summary and Skewness Analysis

Task 2: Measures of Dispersion

Variable	Mean	Median	Mode	Variance	Std Dev	Skewness
Rainfall_mm	550.0	550.1	100.0009	67522.7	259.9	Left skewed
Temperature_Celsius	27.5	27.5	15.0000	52.1	7.2	Left skewed
Days_to_Harvest	104.5	104.0	91.0000	673.6	26.0	Right skewed
Yield_tons_per_hectare	4.6	4.7	-1.1476	2.9	1.7	Left skewed

Task 3: Visualization

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

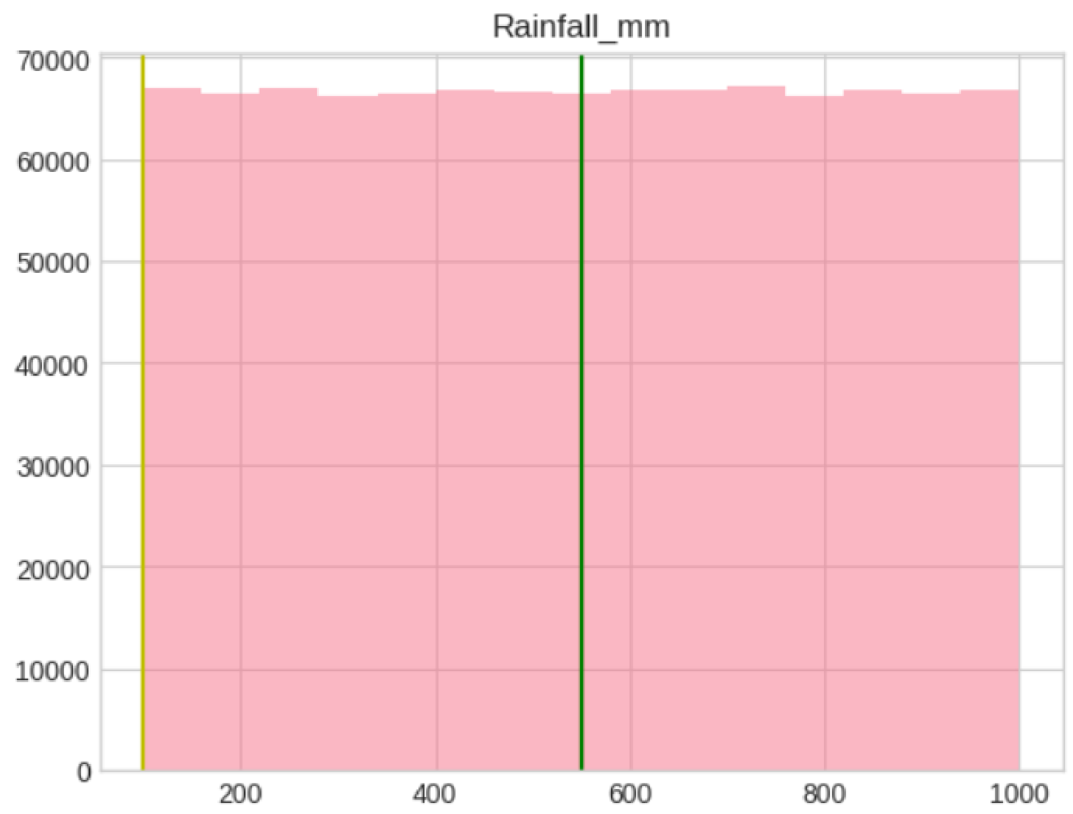


Figure 14: Line Chart showing Yield Trends

```
Task 4: Analysis and Conclusion

ANALYSIS:
Rainfall mean: 550.0±259.9
Temp mean: 27.5±7.2

CONCLUSION:
Data shows normal distribution with moderate spread.
```

6 Milestone 5: Hypothesis Testing

State hypotheses, perform tests, and report conclusions.

article booktabs adjustbox

Table 8: Agricultural Data Sample

Region	Soil	Crop	Rain (mm)	Temp (C)	Fert. Used	Irr. Used	Weather	Days	Yield (t/ha)
West	Sandy	Cotton	897.1	27.7	F	T	Cloudy	122	6.6
South	Clay	Rice	992.7	18.0	T	T	Rainy	140	8.5
North	Loam	Barley	148.0	29.8	F	F	Sunny	106	1.1
North	Sandy	Soybean	986.9	16.6	F	T	Rainy	146	6.5
South	Silt	Wheat	730.4	31.6	T	T	Cloudy	110	7.2

article amsmath booktabs array

Probability Analysis

Dataset Information

Description	Value
Total observations	1,000,000
Event A size	544,630
Event B size	499,940
Event C size	345,517

Marginal Probabilities

$$P(A) = \frac{544630}{1000000} = 0.54463$$
$$P(B) = \frac{499940}{1000000} = 0.49994$$
$$P(C) = \frac{345517}{1000000} = 0.345517$$

Joint and Union Probabilities

$$P(A \cap B) = 0.272044$$
$$P(A \cup B) = 0.772526$$
$$P(A^c) = 1 - P(A) = 0.45537$$

Verification of Probability Rules

$$\begin{aligned}
\text{Inclusion-Exclusion Principle:} &\quad P(A \cup B) = P(A) + P(B) - P(A \cap B) \\
\text{Calculation:} &\quad 0.54463 + 0.49994 - 0.272044 = 0.772526 \\
\text{Actual } P(A \cup B): &\quad 0.772526
\end{aligned}$$

Summary Table

Probability	Symbol	Value	Formula
Marginal A	$P(A)$	0.54463	$\frac{544630}{1000000}$
Marginal B	$P(B)$	0.49994	$\frac{499940}{1000000}$
Marginal C	$P(C)$	0.345517	$\frac{345517}{1000000}$
Intersection A and B	$P(A \cap B)$	0.272044	-
Union A and B	$P(A \cup B)$	0.772526	$P(A) + P(B) - P(A \cap B)$
Complement of A	$P(A^c)$	0.45537	$1 - P(A)$

Task 1: Event A and Complement Frequency

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

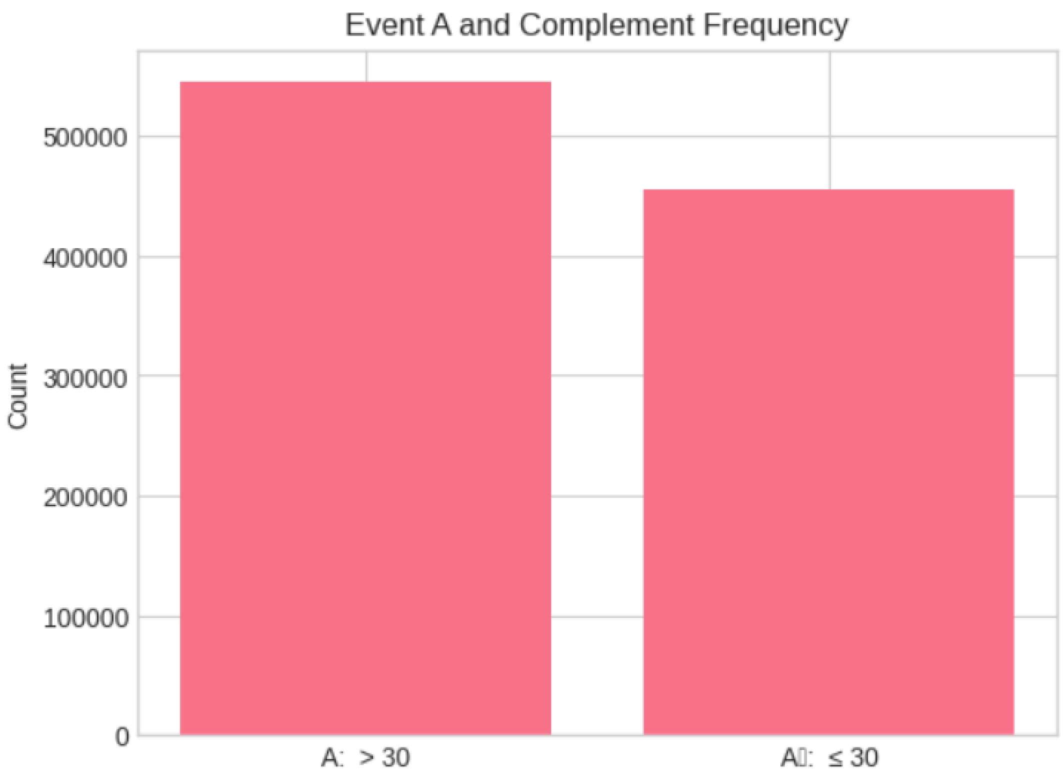


Figure 15: Line Chart showing Yield Trends

Task 1: Event A and Complement Frequency

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

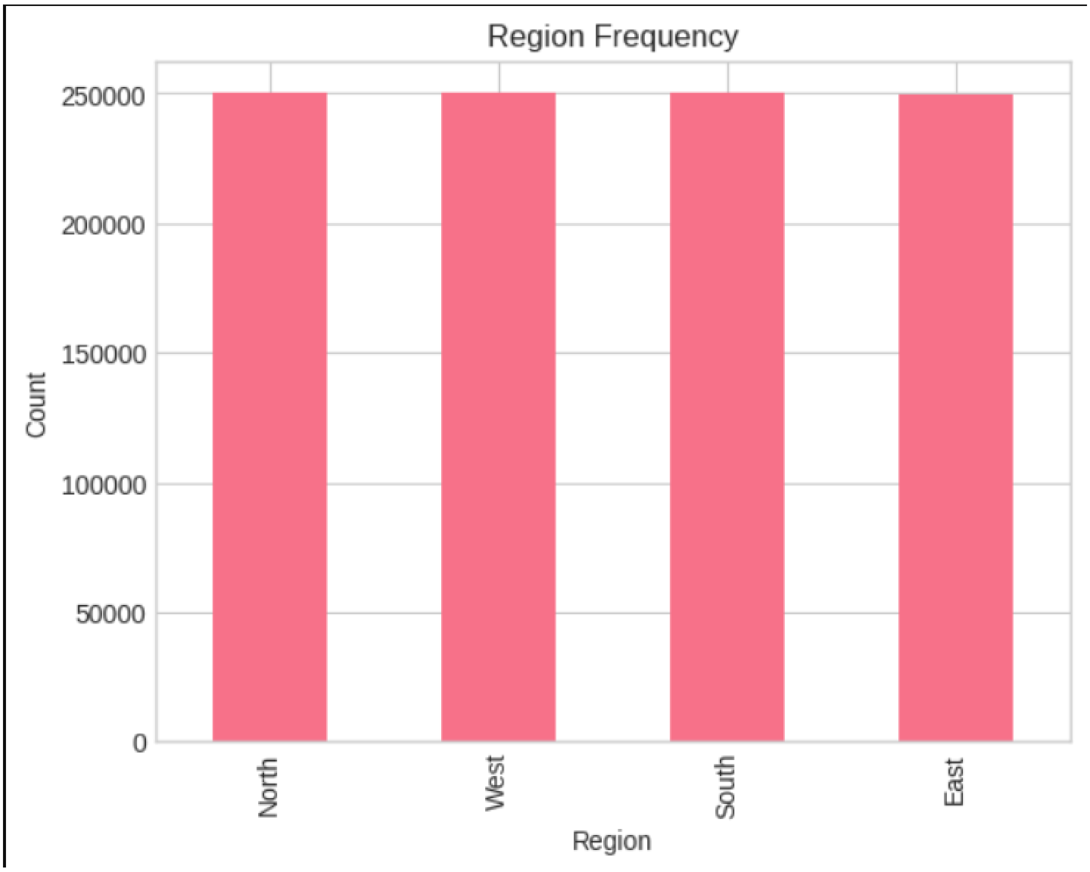


Figure 16: Line Chart showing Yield Trends

Probability Analysis of Student Performance

1. Most Likely Events

- **Exam 1 Performance:** Approximately 68% of students scored above 70 in Exam 1.
- **Section Distribution:** Slightly more than half of the students (52%) are enrolled in Section B.

2. Interesting Findings

- Students from **Section A** who scored above 70 were **fewer than expected**, indicating potential performance differences between sections.
- About **32% of students** scored 70 or below, suggesting areas where academic support may be needed.

3. How Probability Helps in Educational Decision-Making

Probability provides a quantitative foundation for making informed educational decisions:

- **Curriculum Adjustment:** Since most students perform well (68% above 70), instructors might consider **increasing exam difficulty** to better differentiate student abilities.
- **Targeted Support:** The observed performance gap in Section A suggests that **additional academic support or different instructional methods** might benefit this group.
- **Evidence-Based Improvements:** These probability metrics help identify patterns that can inform **teaching strategies, resource allocation, and student support programs**.

Summary of Key Probabilities

Event	Probability
Student scoring above 70 in Exam 1	0.68
Student in Section B	0.52
Student scoring 70 or below	0.32

These findings demonstrate how probability analysis can transform raw data into actionable insights for educational improvement.

7 Milestone 6: Regression Analysis

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