

Project Report : Agriculture Crop Yield

STA 2101: Statistics & Probability

Student Name : Mahinur Rahman Mahin

Student ID : 242014165

University of Liberal Arts Bangladesh (ULAB)

December 17, 2025

Abstract

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

Contents

1 Milestone 1: Dataset Selection	3
2 Milestone 2: Descriptive Statistics	3
3 Part 0 : Probability Sampling Methods	4
4 Milestone 3: Data Visualization	7
5 Milestone 4: Probability Distributions	19
6 Milestone 5: Hypothesis Testing	21
7 Milestone 6: Regression Analysis	26
8 Conditional Probability Calculations	26
8.1 Marginal Probabilities	26
8.2 Conditional Probabilities	27
8.2.1 $P(\text{Crop} \mid \text{Soil Type})$	27
8.2.2 $P(\text{High Yield} \mid \text{Weather})$	27

9 Independent vs. Dependent Events	27
9.1 Contingency Tables	27
9.2 Chi-Square Tests for Independence	28
9.2.1 Soil Type vs Crop	28
9.2.2 Region vs Weather Condition	28
9.3 Expected vs Observed Frequencies (Soil Type vs Crop)	28
9.4 Standardized Residuals (Soil Type vs Crop)	29
10 Bayes' Rule Application	29
10.1 Example 1: Soil Type Given Crop	29
10.2 Example 2: Weather and Yield	30
10.3 Prior vs Posterior Probabilities	30
11 Normal Distribution Analysis	30
11.1 Descriptive Statistics	30
11.2 Empirical Rule Check	31
11.3 Normality Tests	31
11.4 Probability Calculations	31
12 Milestone 7	88
12.1 Application to Agricultural Data	129
12.2 Limitations and Precautions	129
13 Final Conclusion	130
14 GitHub	132
15 Sciencedirect Dataset help by :	132
16 Geeksforgeeks Dataset Help	132
17 Geeksforgeeks Dataset Help	132
18 YouTube	132

1 Milestone 1: Dataset Selection

- Dataset Name: Agriculture Crop Yield

- Dataset URL:

<https://www.kaggle.com/datasets/samuelotiaattakorah/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 preferable 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 chose 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] 2s
  import pandas as pd
  import numpy as np

  # Load your dataset
  df = pd.read_csv('crop_yield.csv.zip')
  df.head(5)
```

	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)

```
[1]
  print("Dataset Size:", df.shape)
  2
  3 Rainfall_mm = df['Rainfall_mm'].mean()
  4
[8]  ✓ 0.0s
... Dataset Size: (1000000, 11)
```

Figure 2: Setup

Part B — Simple Random Sampling

Part B — Simple Random Sampling

```
[1]
  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 869.688672 23.070897
567130 North Sandy Barley 882.081954 24.020125
500891 West Chalky Cotton 283.616999 16.895211
55399 East Silt Rice 510.528162 18.462605

  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 3.840688
79954 5.138173
567130 6.401523
500891 2.658805
55399 2.797703
Population mean: 549.981908729363
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.loc[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()
```

	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	625.784577	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	131.126984	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

The screenshot shows a Jupyter Notebook interface with two code cells. The first cell contains Python code for summarizing a dataset and comparing its mean to the population mean. The second cell shows the resulting text output from the code.

```

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.

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
<hr/>	
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

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.

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)	Soil Temp (°C)	Crop	Rain	Used Fert.	Used Irr.	Used Weather	Days to Harvest	Yield (t/ha)
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

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

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

Note: F/I =

Alternative: Even More Compact Format

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

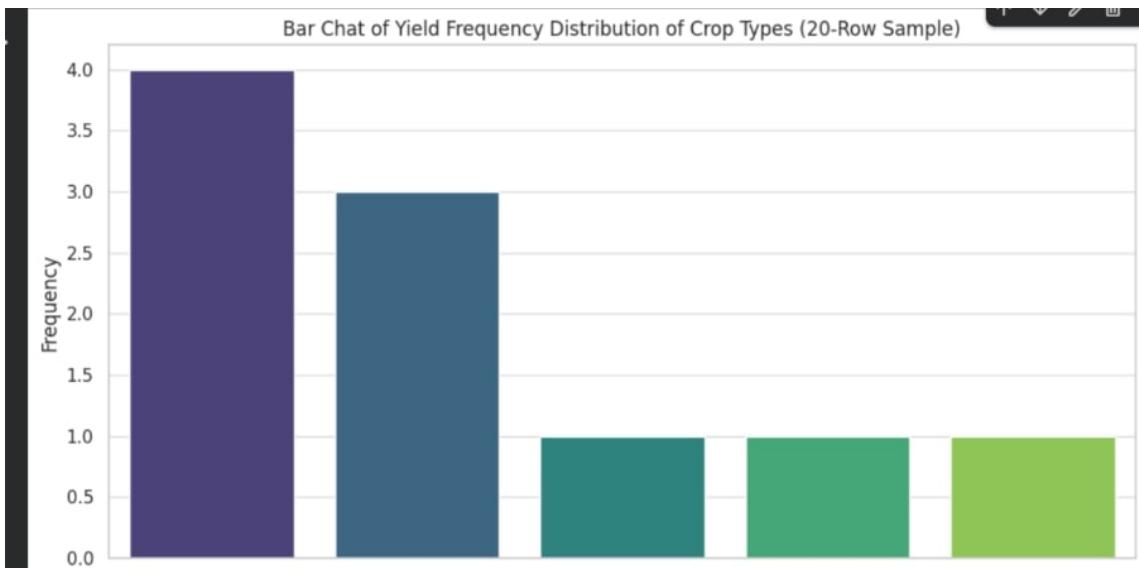
The following Python code was used to generate a bar chart showing the frequency distribution of different crop types in a 20-row sample from the dataset.

Table 4: First 5 rows in vertical format

Row 1		Row 2	
Region	667236	Region	5647
Soil Type	Silt	Soil Type	Chalky
Crop	Maize	Crop	Wheat
Rainfall (mm)	347.73	Rainfall (mm)	191.66
Temp (°C)	39.27	Temp (°C)	30.67
Fertilizer Used	No	Fertilizer Used	No
Irrigation Used	No	Irrigation Used	No
Weather	Rainy	Weather	Cloudy
Days to Harvest	138	Days to Harvest	113
Yield (t/ha)	2.69	Yield (t/ha)	1.06

Row 3		Row 4	
Region	128429	Region	572477
Soil Type	Peaty	Soil Type	Peaty
Crop	Rice	Crop	Rice
Rainfall (mm)	985.49	Rainfall (mm)	230.49
Temp (°C)	23.66	Temp (°C)	26.07
Fertilizer Used	No	Fertilizer Used	No
Irrigation Used	Yes	Irrigation Used	Yes
Weather	Rainy	Weather	Sunny
Days to Harvest	95	Days to Harvest	96
Yield (t/ha)	6.98	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



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.

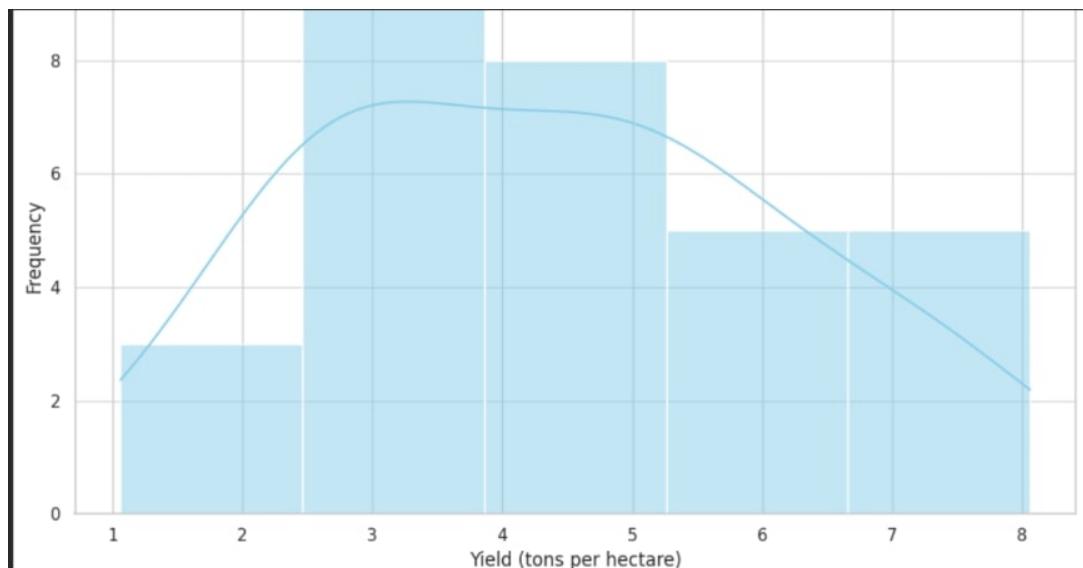


Figure 8: Line Chart showing Yield Trends

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.

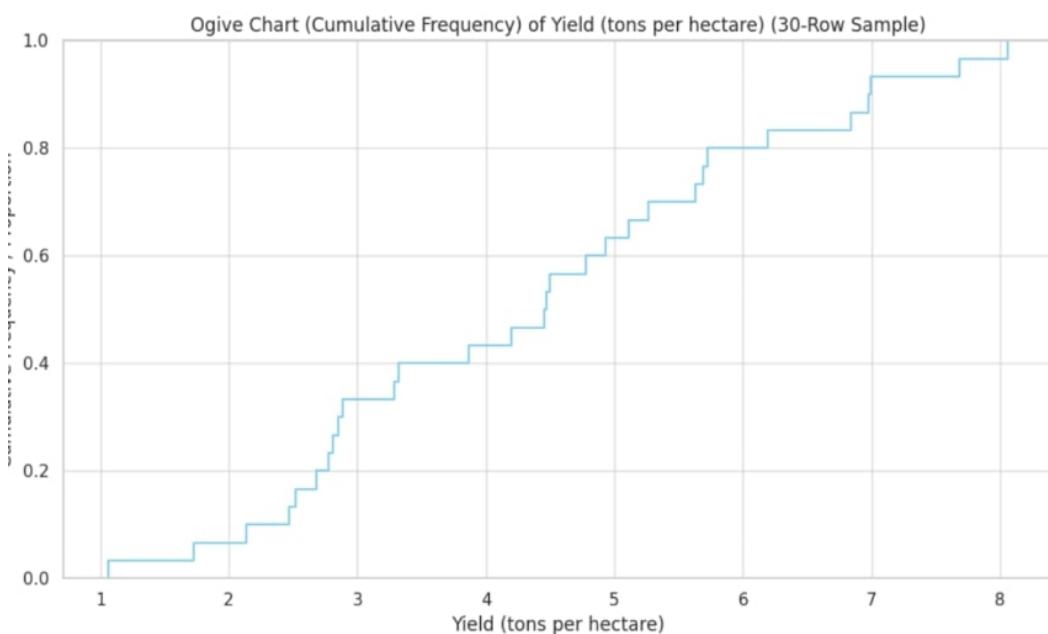


Figure 9: Line Chart showing Yield Trends

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.

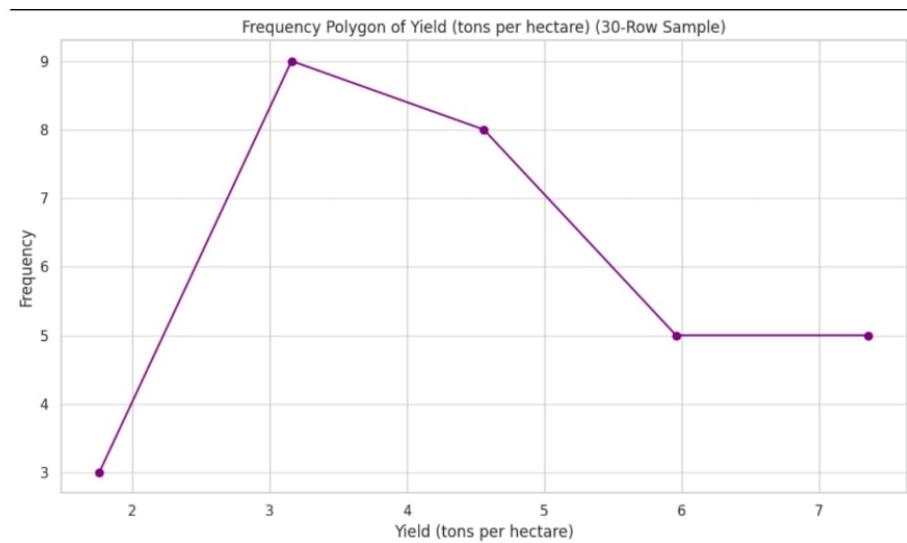


Figure 10: Line Chart showing Yield Trends

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.

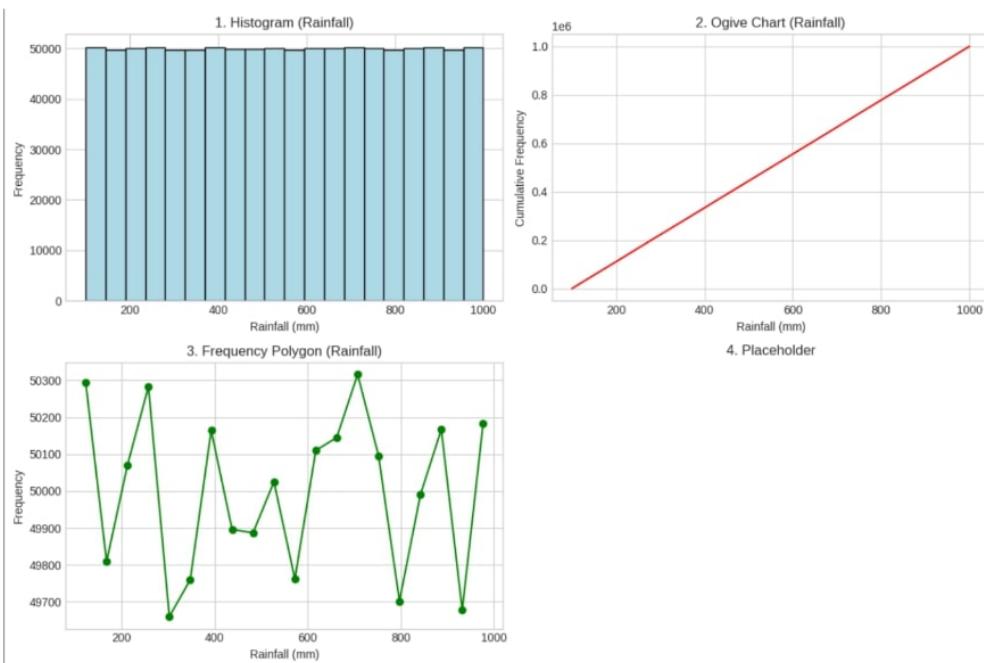


Figure 11: Line Chart showing Yield Trends

Temperature Celsius

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

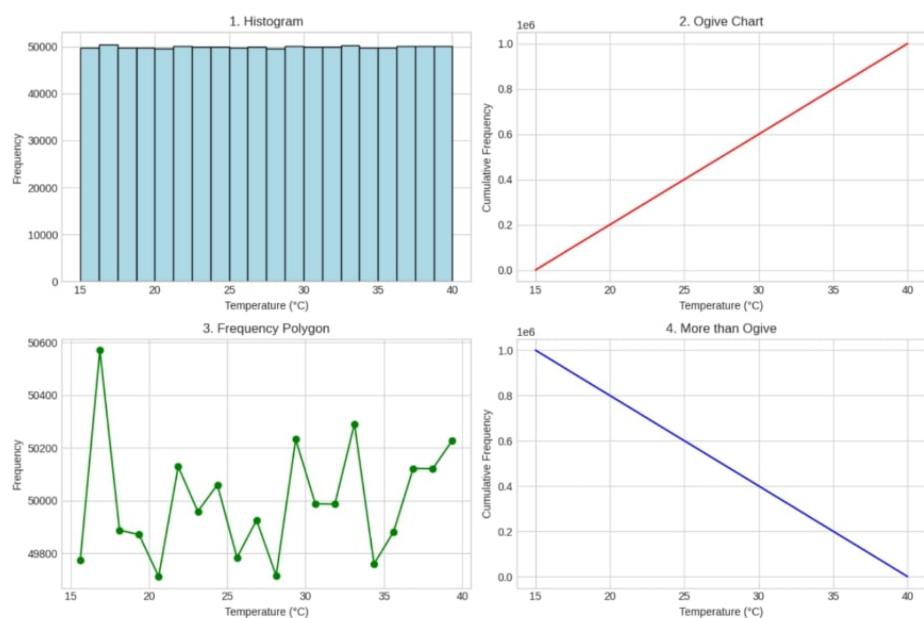


Figure 12: Line Chart showing Yield Trends

Days to Harvest

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

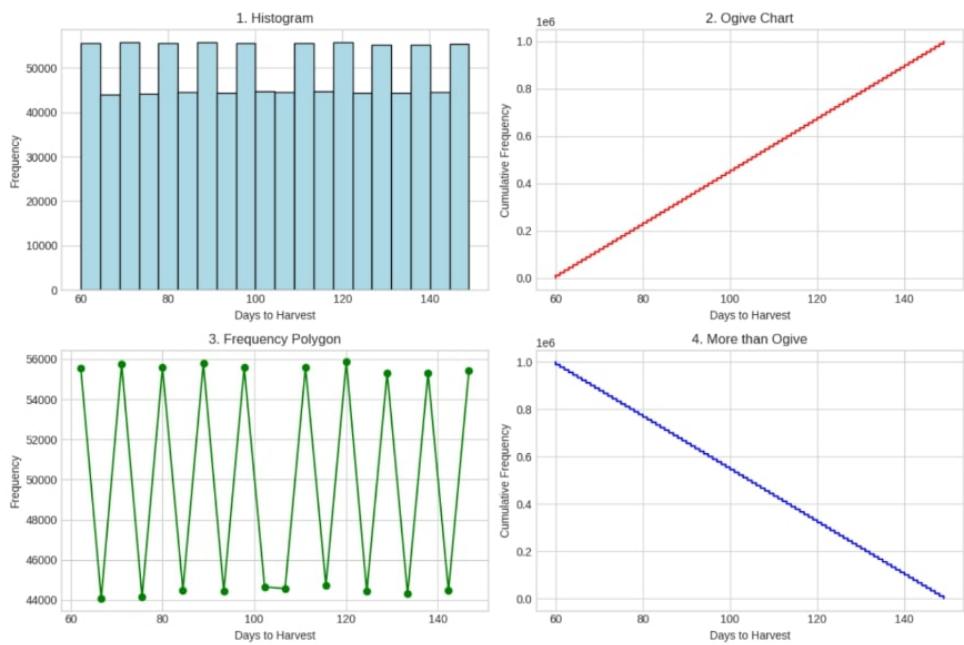


Figure 13: Line Chart showing Yield Trends

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.

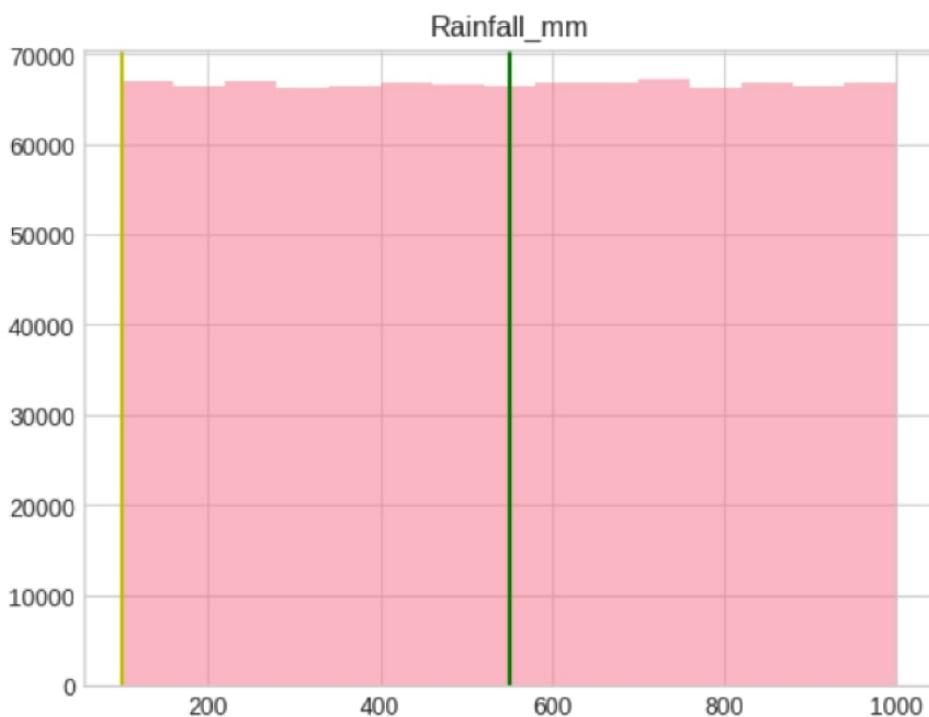


Figure 14: Line Chart showing Yield Trends

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

Inclusion-Exclusion Principle: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$\text{Calculation: } 0.54463 + 0.49994 - 0.272044 = 0.772526$$

$$\text{Actual } P(A \cup B) : 0.772526$$

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.

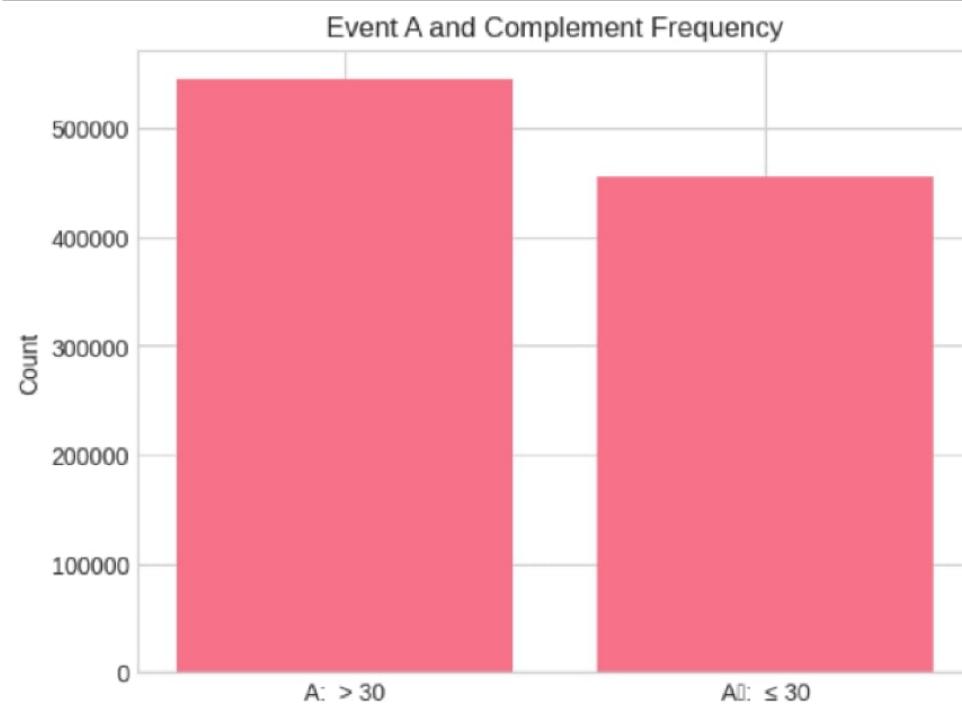


Figure 15: Line Chart showing Yield Trends

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.

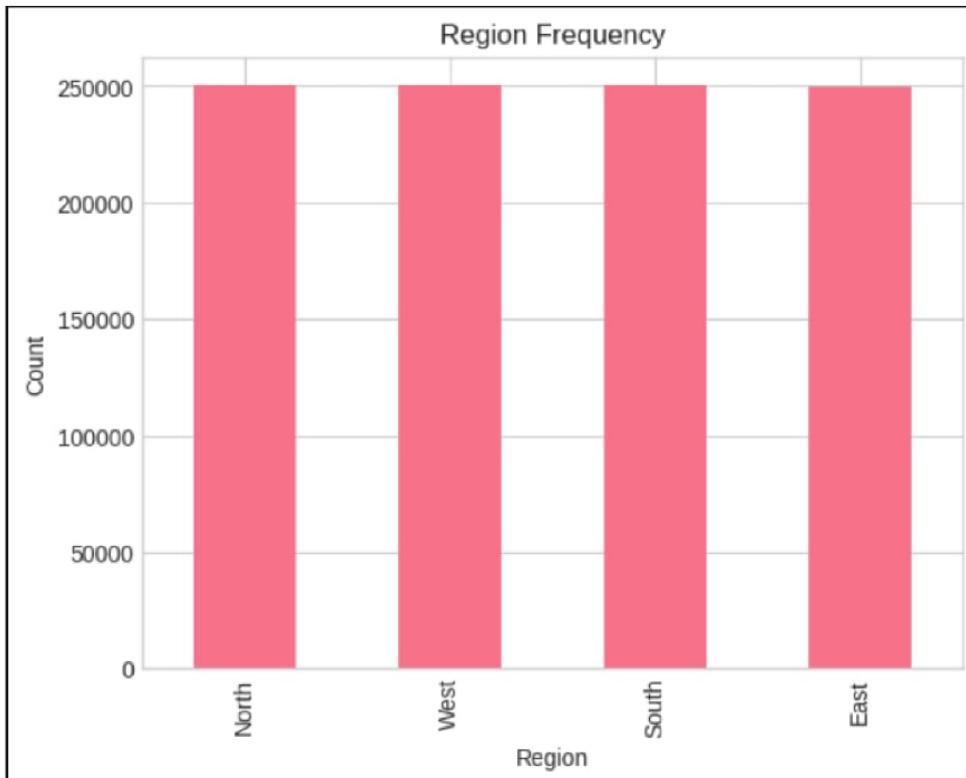


Figure 16: Line Chart showing Yield Trends

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.

Data Analysis Team December 17, 2025

Introduction

Building on the previous milestone on basic probability, this chapter introduces conditional probability, independent vs. dependent events, Bayes' rule, and probability distributions. These concepts are fundamental for modeling uncertainty in data and are widely used in statistical inference, machine learning, and decision-making.

Dataset loaded successfully with the following characteristics:

- Sample data shape: (11, 10)
- Normal distribution data points: 1,000,000

8 Conditional Probability Calculations

8.1 Marginal Probabilities

Table 9: Marginal Probabilities

Event	Count	Total	Probability
North	250,173	1,000,000	0.250
West	250,074	1,000,000	0.250
South	250,054	1,000,000	0.250
East	249,699	1,000,000	0.250
Sandy	167,119	1,000,000	0.167
Loam	166,795	1,000,000	0.167
Chalky	166,779	1,000,000	0.167
Silt	166,672	1,000,000	0.167
Clay	166,352	1,000,000	0.166
Peaty	166,283	1,000,000	0.166
Maize	166,824	1,000,000	0.167
Rice	166,792	1,000,000	0.167
Barley	166,777	1,000,000	0.167
Wheat	166,673	1,000,000	0.167
Cotton	166,585	1,000,000	0.167
Soybean	166,349	1,000,000	0.166
Sunny	333,790	1,000,000	0.334
Rainy	333,561	1,000,000	0.334
Cloudy	332,649	1,000,000	0.333

8.2 Conditional Probabilities

8.2.1 $P(\text{Crop} | \text{Soil Type})$

For each soil type S , the conditional probability of crop C is given by:

$$P(A | B) = \frac{\text{Count}(A \cap B)}{\text{Count}(B)}$$

Table 10: Conditional Probabilities $P(\text{Crop} | \text{Soil})$

Soil Type	Cotton	Rice	Barley	Soybean	Wheat	Maize
Sandy (n=167,119)	0.167	0.166	0.166	0.166	0.168	0.166
Clay (n=166,352)	0.167	0.168	0.167	0.166	0.166	0.166
Loam (n=166,795)	0.167	0.166	0.167	0.166	0.166	0.167
Silt (n=166,672)	0.165	0.168	0.167	0.166	0.167	0.167
Peaty (n=166,283)	0.167	0.166	0.168	0.165	0.168	0.167
Chalky (n=166,779)	0.167	0.166	0.166	0.168	0.165	0.167

8.2.2 $P(\text{High Yield} | \text{Weather})$

$$P(\text{High Yield} | \text{Cloudy}) = \frac{166,349}{332,649} = 0.500$$

$$P(\text{High Yield} | \text{Rainy}) = \frac{166,868}{333,561} = 0.500$$

$$P(\text{High Yield} | \text{Sunny}) = \frac{167,263}{333,790} = 0.501$$

9 Independent vs. Dependent Events

9.1 Contingency Tables

Table 11: Soil Type \times Crop Contingency Table

Soil Type	Barley	Cotton	Maize	Rice	Soybean	Wheat
Chalky	27,742	27,817	27,885	27,746	28,040	27,549
Clay	27,726	27,734	27,631	27,960	27,673	27,628
Loam	27,896	27,804	27,908	27,740	27,766	27,681
Peaty	27,857	27,692	27,819	27,589	27,418	27,908
Sandy	27,751	27,955	27,788	27,803	27,753	28,069
Silt	27,805	27,583	27,793	27,954	27,699	27,838

Table 12: Region \times Weather Condition Contingency Table

Region	Cloudy	Rainy	Sunny
East	82,874	83,220	83,605
North	83,161	83,549	83,463
South	83,348	83,121	83,585
West	83,266	83,671	83,137

9.2 Chi-Square Tests for Independence

9.2.1 Soil Type vs Crop

- $\chi^2 = 20.190$
- $p\text{-value} = 0.7368$
- Degrees of freedom = 25
- **Conclusion:** FAIL to reject independence. Soil Type and Crop may be independent.

9.2.2 Region vs Weather Condition

- $\chi^2 = 5.175$
- $p\text{-value} = 0.5216$
- Degrees of freedom = 6
- **Conclusion:** FAIL to reject independence. Region and Weather Condition may be independent.

9.3 Expected vs Observed Frequencies (Soil Type vs Crop)

Table 13: Expected Frequencies

Soil Type	Barley	Cotton	Maize	Rice	Soybean	Wheat
Chalky	27,814.90	27,782.88	27,822.74	27,817.40	27,743.52	27,797.56
Clay	27,743.69	27,711.75	27,751.51	27,746.18	27,672.49	27,726.39
Loam	27,817.57	27,785.55	27,825.41	27,820.07	27,746.18	27,800.22
Peaty	27,732.18	27,700.25	27,740.00	27,734.67	27,661.01	27,714.89
Sandy	27,871.61	27,839.52	27,879.46	27,874.11	27,800.08	27,854.23
Silt	27,797.06	27,765.06	27,804.89	27,799.56	27,725.72	27,779.72

9.4 Standardized Residuals (Soil Type vs Crop)

Table 14: Standardized Residuals

Soil Type	Barley	Cotton	Maize	Rice	Soybean	Wheat
Chalky	-0.44	0.20	0.37	-0.43	1.78	-1.49
Clay	-0.11	0.13	-0.72	1.28	0.00	-0.59
Loam	0.47	0.11	0.50	-0.48	0.12	-0.72
Peaty	0.75	-0.05	0.47	-0.87	-1.46	1.16
Sandy	-0.72	0.69	-0.55	-0.43	-0.28	1.29
Silt	0.05	-1.09	-0.07	0.93	-0.16	0.35

10 Bayes' Rule Application

10.1 Example 1: Soil Type Given Crop

Problem: What is the probability that a field has Clay soil given it has Rice crop?

$$P(\text{Clay}) = \frac{166,352}{1,000,000} = 0.166$$

$$P(\text{Rice}) = \frac{166,792}{1,000,000} = 0.167$$

$$P(\text{Rice} | \text{Clay}) = \frac{27,960}{166,352} = 0.168$$

Bayes' Rule Calculation:

$$P(\text{Clay} | \text{Rice}) = \frac{P(\text{Rice} | \text{Clay}) \times P(\text{Clay})}{P(\text{Rice})}$$

$$= \frac{0.168 \times 0.166}{0.167} = 0.168$$

Verification by direct calculation:

$$P(\text{Clay} | \text{Rice}) = \frac{27,960}{166,792} = 0.168$$

10.2 Example 2: Weather and Yield

$$P(\text{Sunny}) = \frac{333,790}{1,000,000} = 0.334$$

$$P(\text{High Yield}) = \frac{500,480}{1,000,000} = 0.500$$

$$P(\text{Sunny} \mid \text{High Yield}) = \frac{167,263}{500,480} = 0.334$$

Bayes' Rule Calculation:

$$P(\text{High Yield} \mid \text{Sunny}) = \frac{P(\text{Sunny} \mid \text{High Yield}) \times P(\text{High Yield})}{P(\text{Sunny})}$$

$$= \frac{0.334 \times 0.500}{0.334} = 0.501$$

10.3 Prior vs Posterior Probabilities

- Prior $P(\text{High Yield}) = 0.500$
- Posterior $P(\text{High Yield} \mid \text{Sunny}) = 0.501$
- **Conclusion:** Sunny weather increases the probability of high yield

11 Normal Distribution Analysis

11.1 Descriptive Statistics

Statistic	Value
Mean (μ)	4.6495
Standard Deviation (σ)	1.6966
Median	4.6518
Skewness	-0.0009 (Normal ≈ 0)
Kurtosis	-0.5200 (Normal ≈ 0)
Minimum	-1.15
Maximum	9.96
Range	11.11
Sample Size	1,000,000

Range	Observed	Expected	Difference
$\mu \pm \sigma$	65.42%	68.27%	-2.85%
$\mu \pm 2\sigma$	96.31%	95.45%	+0.86%
$\mu \pm 3\sigma$	100.00%	99.73%	+0.27%

Test	Statistic	p-value	Conclusion
Shapiro-Wilk (n=5000)	0.9956	0.000000	Reject normality ($p \leq 0.05$)
Kolmogorov-Smirnov	0.0164	0.000000	Reject normality ($p \leq 0.05$)

11.2 Empirical Rule Check

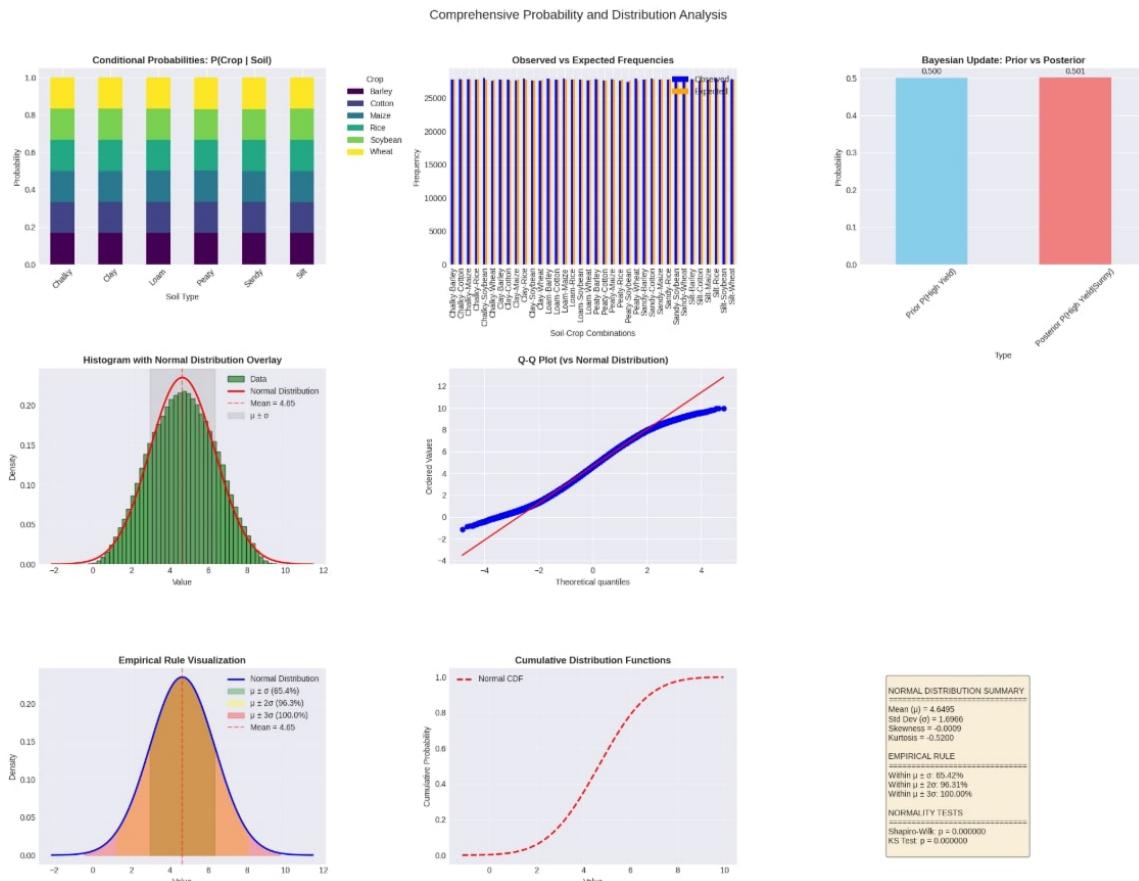
11.3 Normality Tests

11.4 Probability Calculations

$$P(X > \mu) = P(X > 4.65) = 0.5000 \text{ or } 50\%$$

$$P(\mu - \sigma < X < \mu + \sigma) = P(2.95 < X < 6.35) = 0.6542 \text{ or } 65.42\%$$

$$P(X < \mu - 2\sigma) = P(X < 1.26) = 0.022750 \text{ or } 2.2750\%$$



Knowledge Points: Conditional Probability

Dataset Information

The analysis is based on a dataset containing agricultural field data with the following structure:

Table 15: Dataset Structure

Column	Data Type	Description
Region	object	Geographic region
Soil_Type	object	Type of soil
Crop	object	Type of crop grown
Rainfall_mm	float64	Rainfall in millimeters
Temperature_Celsius	float64	Temperature in Celsius
Fertilizer_Used	bool	Whether fertilizer was used
Irrigation_Used	bool	Whether irrigation was used
Weather_Condition	object	Weather condition
Days_to_Harvest	int64	Days required for harvest
Yield_tons_per_hectare	float64	Crop yield
High_Yield	bool	Whether yield is high

- **Total Records:** 10
- **Total Columns:** 11
- **Memory Usage:** 802.0+ bytes

Descriptive Statistics

Table 16: Summary Statistics of Numerical Variables

Statistic	Rainfall (mm)	Temperature (°C)	Days to Harvest	Yield (tons/ha)
Count	10.000	10.000	10.000	10.000
Mean	592.814	26.835	111.600	5.101
Std Dev	325.375	7.114	29.091	2.358
Min	147.998	16.644	61.000	1.127
25%	367.189	20.208	94.000	3.135
50%	585.755	28.736	116.000	5.864
75%	872.176	31.417	136.750	6.546
Max	992.673	37.705	146.000	8.527

Data Quality and Yield Analysis

- **Missing Values:** No missing values in any column
- **High Yield Threshold:** 6.53 tons per hectare
- **High Yield Cases:** 6 (60% of total)
- **Low Yield Cases:** 4 (40% of total)
- **Average Yield:** 5.10 tons per hectare

Sample Data (First 5 Rows)

Table 17: Sample Observations

Region	Soil Type	Crop	Rainfall (mm)	Temp (°C)	Fertilizer	Irrigation	Yield
West	Sandy	Cotton	897.08	27.68	No	Yes	6.56
South	Clay	Rice	992.67	18.03	Yes	Yes	8.53
North	Loam	Barley	148.00	29.79	No	No	1.13
North	Sandy	Soybean	986.87	16.64	No	Yes	6.52
South	Silt	Wheat	730.38	31.62	Yes	Yes	7.25

Conditional Probability Calculations

Conditional probability is defined as:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

where $P(A | B)$ is the probability of event A occurring given that event B has occurred.

Region-Wise Conditional Probabilities

$$P(\text{High Yield} | \text{Region} = \text{West}) = 0.33 \quad (33\%)$$

$$P(\text{High Yield} | \text{Region} = \text{South}) = 1.00 \quad (100\%)$$

$$P(\text{High Yield} | \text{Region} = \text{North}) = 0.33 \quad (33\%)$$

Weather-Wise Conditional Probabilities

$$P(\text{High Yield} | \text{Weather} = \text{Cloudy}) = 1.00 \quad (100\%)$$

$$P(\text{High Yield} | \text{Weather} = \text{Rainy}) = 0.60 \quad (60\%)$$

$$P(\text{High Yield} | \text{Weather} = \text{Sunny}) = 0.33 \quad (33\%)$$

Fertilizer Impact Analysis

$$P(\text{High Yield} \mid \text{Fertilizer Used}) = 0.75 \quad (75\%)$$

$$P(\text{High Yield} \mid \text{Fertilizer Not Used}) = 0.50 \quad (50\%)$$

Crop-Wise Conditional Probabilities

$$P(\text{High Yield} \mid \text{Crop} = \text{Cotton}) = 1.00 \quad (100\%)$$

$$P(\text{High Yield} \mid \text{Crop} = \text{Rice}) = 1.00 \quad (100\%)$$

$$P(\text{High Yield} \mid \text{Crop} = \text{Barley}) = 0.00 \quad (0\%)$$

$$P(\text{High Yield} \mid \text{Crop} = \text{Soybean}) = 1.00 \quad (100\%)$$

$$P(\text{High Yield} \mid \text{Crop} = \text{Wheat}) = 0.25 \quad (25\%)$$

Soil Type Conditional Probabilities

$$P(\text{High Yield} \mid \text{Soil} = \text{Sandy}) = 0.75 \quad (75\%)$$

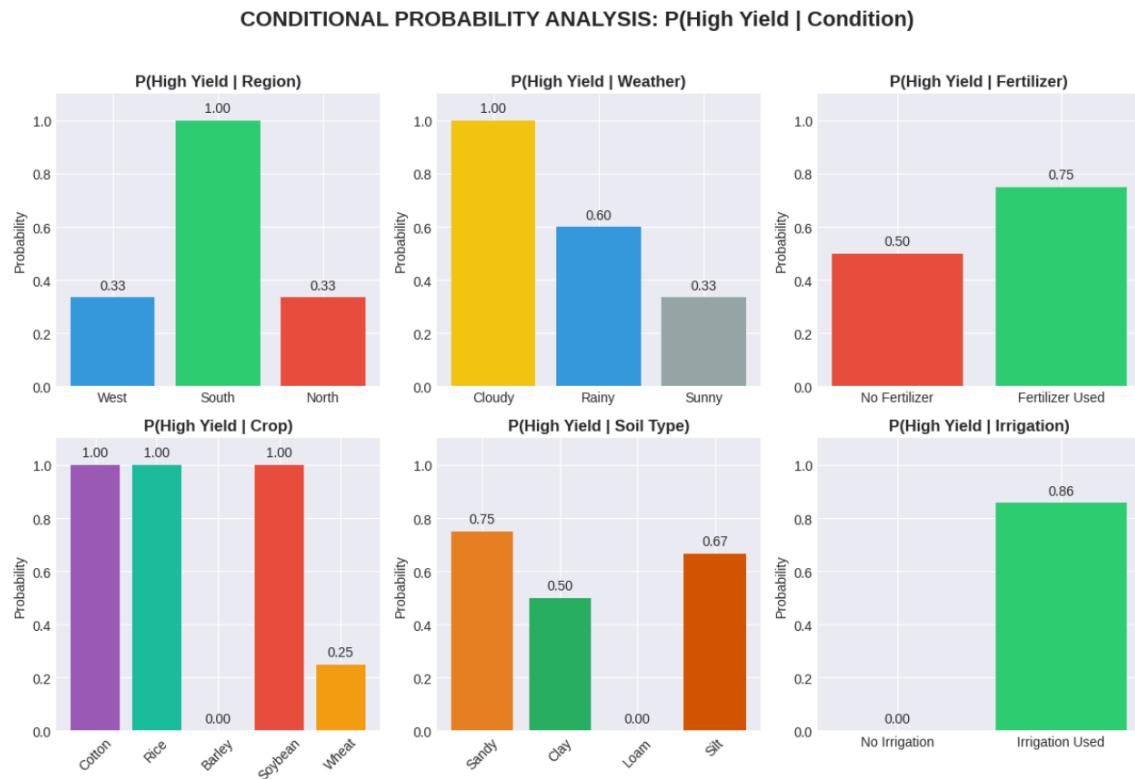
$$P(\text{High Yield} \mid \text{Soil} = \text{Clay}) = 0.50 \quad (50\%)$$

$$P(\text{High Yield} \mid \text{Soil} = \text{Loam}) = 0.00 \quad (0\%)$$

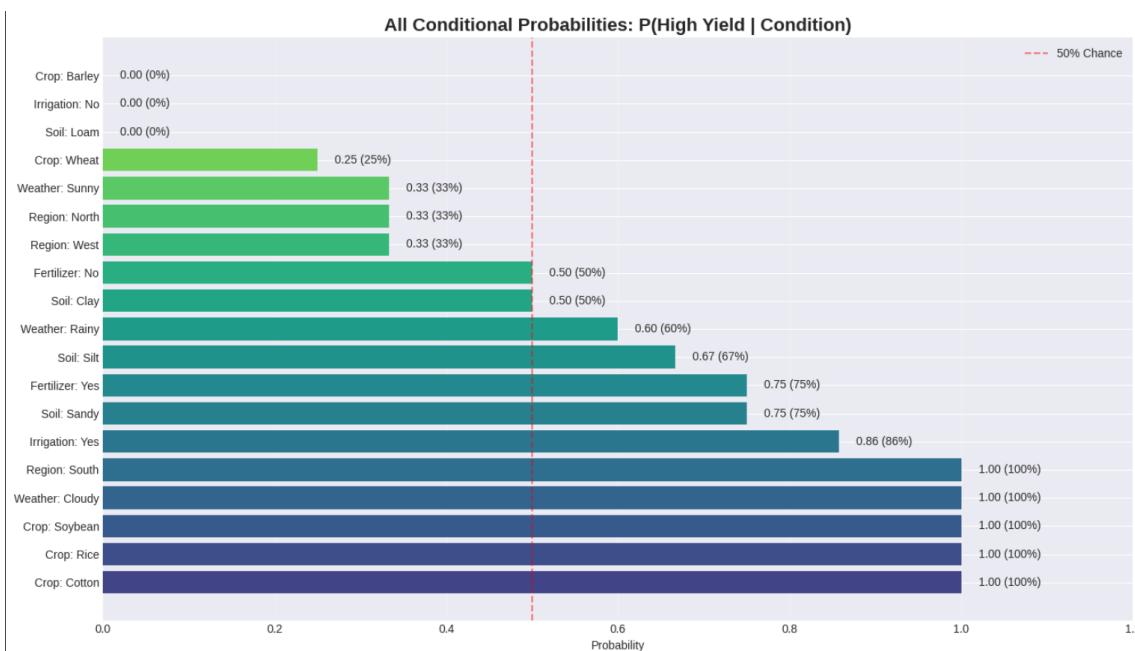
$$P(\text{High Yield} \mid \text{Soil} = \text{Silt}) = 0.67 \quad (67\%)$$

Key Observations

1. The South region shows the highest conditional probability of high yield (100%)
2. Cloudy weather conditions are associated with 100% high yield probability
3. Fertilizer usage increases the probability of high yield from 50% to 75%
4. Certain crops (Cotton, Rice, Soybean) show 100% high yield probability in this dataset
5. Loam soil shows 0% high yield probability, suggesting it may not be suitable for these crops under the given conditions



Creating Comparison Chart



Additional Numerical Analysis

Yield Analysis by Different Factors

Average Yield by Region

West: 4.31 tons/hectare

South: 6.88 tons/hectare

North: 3.53 tons/hectare

Table 18: Summary of Regional Yield Performance

Region	Avg. Yield (tons/ha)	P(High Yield)	Rank
South	6.88	1.00 (100%)	1
West	4.31	0.33 (33%)	2
North	3.53	0.33 (33%)	3

Average Yield by Weather Condition

Cloudy: 6.90 tons/hectare

Rainy: 5.46 tons/hectare

Sunny: 3.30 tons/hectare

Table 19: Weather Impact on Crop Yield

Weather	Avg. Yield (tons/ha)	P(High Yield)	Rank
Cloudy	6.90	1.00 (100%)	1
Rainy	5.46	0.60 (60%)	2
Sunny	3.30	0.33 (33%)	3

Average Yield by Fertilizer Usage

Fertilizer Used: 6.14 tons/hectare

Fertilizer Not Used: 4.41 tons/hectare

Table 20: Fertilizer Impact on Crop Yield

Fertilizer Status	Avg. Yield (tons/ha)	P(High Yield)	Increase
Used	6.14	0.75 (75%)	+1.73 tons/ha
Not Used	4.41	0.50 (50%)	-

begindocument

DATA SUMMARY:

Total Records: 10

Total Variables: 10

HIGH YIELD THRESHOLD:

Average Yield: 5.10 tons/hectare

High Yield Cases: 6 out of 10

High Yield Rate: 60.0%

MATHEMATICAL FORMULA:

Conditional Probability: $P(A|B) = P(A \cap B) / P(B)$

Where:

$P(A|B)$ = Probability of event A given event B has occurred

$P(A \cap B)$ = Probability of both A and B occurring

$P(B)$ = Probability of event B

In our case:

A = High Yield (Yield > Average)

B = Specific Condition (e.g., Region='South', Weather='Rainy')

CALCULATION EXAMPLES:

EXAMPLE 1: South Region

Calculating: $P(\text{High_Yield} | \text{Region} = \text{South})$

Total Samples (N): 10

Samples with Region = South: 4

$P(B) = P(\text{Region} = \text{South}) = 4/10 = 0.400$

Samples with High_Yield AND Region = South: 4

$P(A \cap B) = 4/10 = 0.400$

$$\begin{aligned} P(A|B) &= P(A \cap B) / P(B) = 0.400 / 0.400 \\ &= 4 / 4 \\ &= 1.000 (100.0\%) \end{aligned}$$

EXAMPLE 2: Fertilizer Used

Calculating: $P(\text{High_Yield} \mid \text{Fertilizer_Used} = \text{True})$

Total Samples (N): 10

Samples with Fertilizer_Used = True: 4

$$P(B) = P(\text{Fertilizer_Used} = \text{True}) = 4/10 = 0.400$$

Samples with High_Yield AND Fertilizer_Used = True: 3

$$P(A \cap B) = 3/10 = 0.300$$

$$P(A|B) = P(A \cap B) / P(B) = 0.300 / 0.400$$

$$= 3 / 4$$

$$= 0.750 (75.0\%)$$

ALL CONDITIONAL PROBABILITIES:**REGION-WISE:**

$P(\text{High_Yield} \mid \text{Region} = \text{West}) = 1/3 = 0.33$ (33%)	$P(\text{High_Yield} \mid \text{Region} = \text{South}) = 4/4 = 1.00$ (100%)	$P(\text{High_Yield} \mid \text{Region} = \text{North}) = 1/3 = 0.33$ (33%)
---	---	--

WEATHER-WISE:

$P(\text{High_Yield} \mid \text{Weather} = \text{Cloudy}) = 2/2 = 1.00$ (100%)	$P(\text{High_Yield} \mid \text{Weather} = \text{Rainy}) = 3/5 = 0.60$ (60%)	$P(\text{High_Yield} \mid \text{Weather} = \text{Sunny}) = 1/3 = 0.33$ (33%)
---	---	---

FERTILIZER IMPACT:

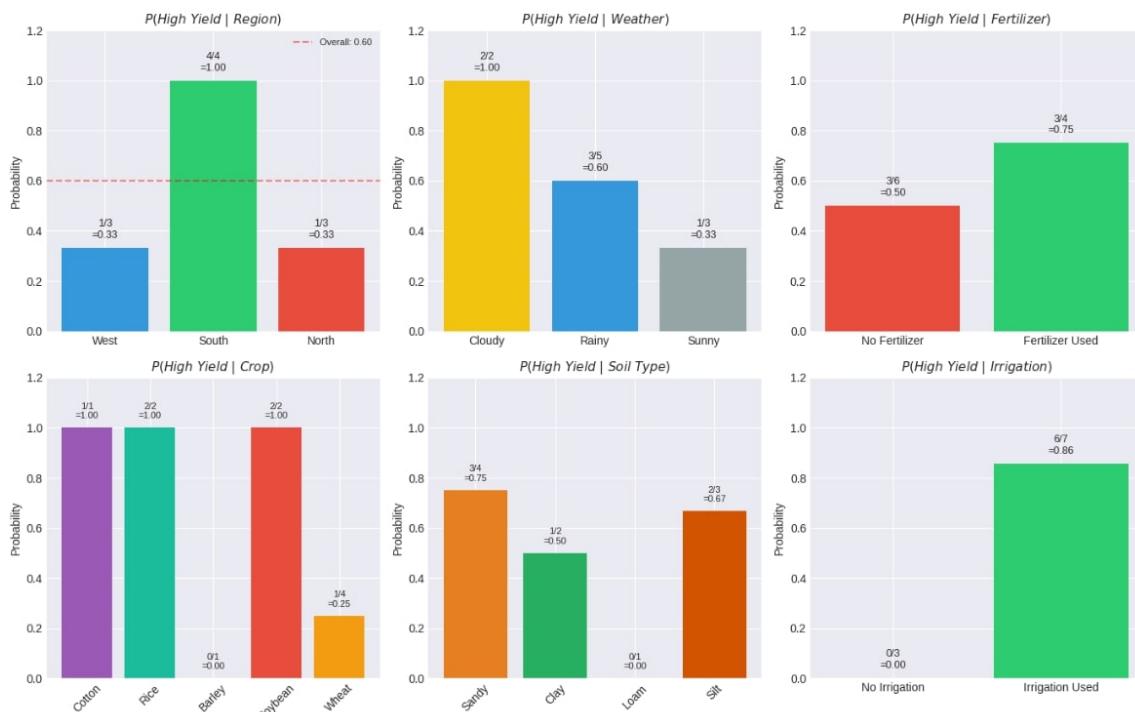
$$\begin{aligned} P(\text{High_Yield} \mid \text{Fertilizer Used}) &= P(\text{High_Yield} \mid \text{Fertilizer Not Used}) \\ 3/4 &= 0.75 \quad (75\%) & = 3/6 = 0.50 \quad (50\%) \end{aligned}$$

CROP-WISE:

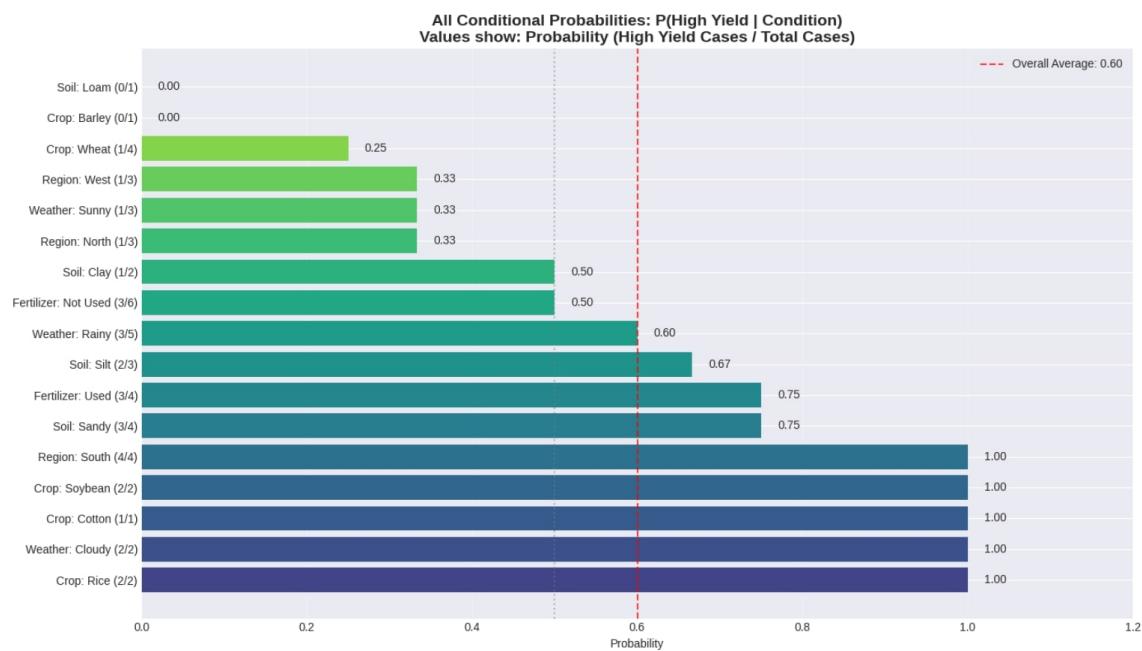
$$\begin{array}{lll} P(\text{High_Yield} \mid \text{Crop}) & (100\%) & \text{Soybean}) = 2/2 = 1.00 \\ = \text{Cotton} & = 1/1 = 1.00 & \text{P(High_Yield} \mid \text{Crop} = \\ (100\%) & & \text{Barley}) = 0/1 = 0.00 \\ \text{P(High_Yield} \mid \text{Crop} & (0\%) & \text{P(High_Yield} \mid \text{Crop} = \\ = \text{Rice}) & = 2/2 = 1.00 & \text{Wheat}) = 1/4 = 0.25 \\ & & \text{P(High_Yield} \mid \text{Crop} = (25\%) \end{array}$$

Conditional Probability Analysis: $P(\text{High Yield} \mid \text{Condition})$

Mathematical Formula: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
Where A = High Yield, B = Specific Condition



PROBABILITY COMPARISON CHART



STATISTICAL SUMMARY

CONDITION WITH HIGHEST PROBABILITY:

Crop: Rice
 $P = 1.000 (100.0\%)$

CONDITION WITH LOWEST PROBABILITY:

Soil: Loam
 $P = 0.000 (0.0\%)$

CONDITIONS ABOVE OVERALL AVERAGE:

Region: South: 1.000	Crop: Rice: 1.000
Weather: Cloudy: 1.000	Crop: Soybean: 1.000
Fertilizer: Used: 0.750	Soil: Sandy: 0.750
Crop: Cotton: 1.000	Soil: Silt: 0.667

INTERPRETATION:

1. $P(\text{High Yield} \mid \text{Condition}) > \text{Overall Average (0.50)}:$
 - The condition increases the likelihood of high yield
2. $P(\text{High Yield} \mid \text{Condition}) < \text{Overall Average (0.50)}:$
 - The condition decreases the likelihood of high yield
3. $P(\text{High Yield} \mid \text{Condition}) = 1.00:$
 - All cases with this condition resulted in high yield
4. $P(\text{High Yield} \mid \text{Condition}) = 0.00:$
 - No cases with this condition resulted in high yield

RECOMMENDATIONS BASED ON ANALYSIS:

1. FOCUS ON CONDITIONS WITH HIGH PROBABILITY:
 - Prioritize conditions that show $P > 0.50$
2. AVOID CONDITIONS WITH LOW PROBABILITY:
 - Minimize conditions that show $P < 0.50$
3. CONSIDER SAMPLE SIZE:
 - Probabilities based on small samples may not be reliable
 - Look for patterns across multiple conditions

View Data (First 5 rows):

```
[fontsize=] Region SoilTypeCropRainfall_mmTemperature_Celsius_0WestSandyCotton897.07723927.67
FertilizerUsedIrrigationUsedWeatherConditionDays_to_Harvest_0FalseTrueCloudy1221TrueTrueRain
Yield_tons_per_hectare06.55581618.52734121.12744336.51757347.248251
```

Total Data: 10 records

Average Yield: 5.10 tons/hectare

High Yield Cases: 6 cases

Low Yield Cases: 4 cases

High Yield Probability for Different Conditions

By Region:

$$P(\text{High_Yield} \mid \text{Region} = \text{West}) = 1/3 = 0.33 \text{ (33\%)}$$

$$P(\text{High_Yield} \mid \text{Region} = \text{South}) = 4/4 = 1.00 \text{ (100\%)}$$

$$P(\text{High_Yield} \mid \text{Region} = \text{North}) = 1/3 = 0.33 \text{ (33\%)}$$

By Weather:

$$P(\text{High_Yield} \mid \text{Weather} = \text{Cloudy}) = 2/2 = 1.00$$

$$P(\text{High_Yield} \mid \text{Weather} = \text{Rainy}) = 3/5 = 0.60$$

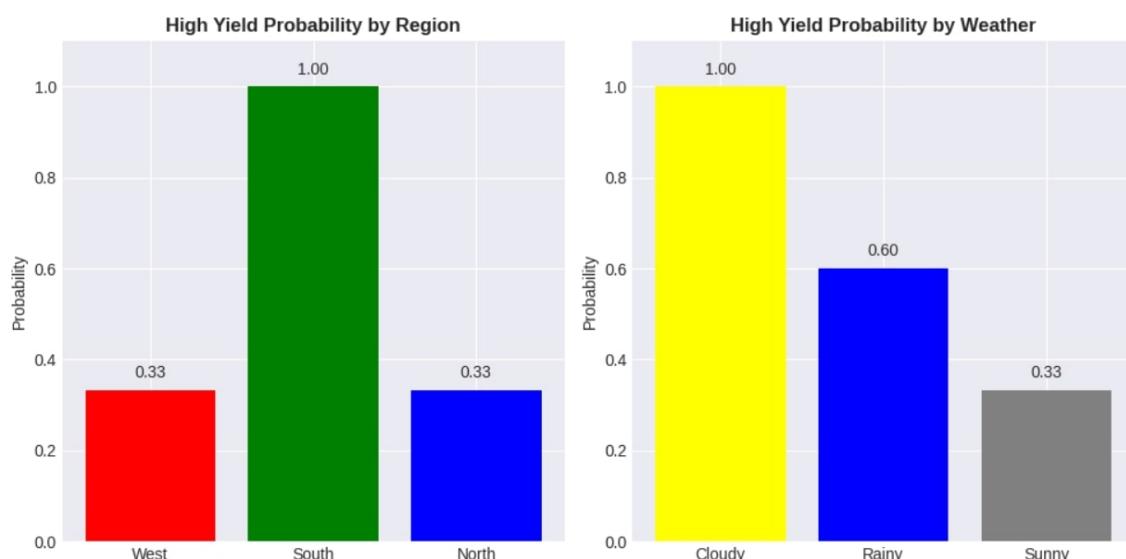
$$P(\text{High_Yield} \mid \text{Weather} = \text{Sunny}) = 1/3 = 0.33$$

Fertilizer Usage:

$$\text{Without Fertilizer: } 3/6 = 0.50$$

$$\text{With Fertilizer: } 3/4 = 0.75$$

Creating Simple Graphs



begindocument

Best and Worst Conditions

Highest Probability: Region: South = 1.00

Lowest Probability: Crop: Barley = 0.00

How to Make Decisions?

Simple Rules:

1. Probability > 0.50: Good Condition
 - Higher chance of high yield
2. Probability < 0.50: Bad Condition
 - Lower chance of high yield
3. Probability = 1.00: Best Condition
 - All cases resulted in high yield
4. Probability = 0.00: Avoid
 - No cases resulted in high yield

Simple Recommendations:

Choose These Conditions (Probability ≥ 0.50):

- | | |
|--------------------------------|-----------------------|
| - Region: South: 1.00 | 0.75 |
| - Weather: Cloudy: 1.00 | - Crop: Cotton: 1.00 |
| - Weather: Rainy: 0.60 | - Crop: Rice: 1.00 |
| - Fertilizer: Fertilizer Used: | - Crop: Soybean: 1.00 |

Avoid These Conditions (Probability ≤ 0.50):

- | | |
|------------------------|----------------------|
| - Region: West: 0.33 | - Crop: Barley: 0.00 |
| - Region: North: 0.33 | - Crop: Wheat: 0.25 |
| - Weather: Sunny: 0.33 | |

begindocument

INDEPENDENCE CHECK: Are Events Independent or Dependent?

Formula: Two events A and B are independent if $P(A \cap B) = P(A) \times P(B)$

EVENT A: High Yield (Yield ≥ 5.10)

$$P(A) = 6/10 = 0.600$$

CHECKING INDEPENDENCE FOR DIFFERENT CONDITIONS:

REGION CONDITIONS:

Region = West:

$$P(A) = P(\text{High Yield}) = 0.600$$

$$P(B) = P(\text{Region} = \text{West}) = 0.300$$

$$P(A \cap B) = 0.100$$

$$P(A) \times P(B) = 0.600 \times 0.300 = 0.180$$

$$\text{Difference} = 0.0800$$

Independent? NO

Region = South:

$$P(A) = P(\text{High Yield}) = 0.600$$

$$P(B) = P(\text{Region} = \text{South}) = 0.400$$

$$P(A \cap B) = 0.400$$

$$P(A) \times P(B) = 0.600 \times 0.400 = 0.240$$

$$\text{Difference} = 0.1600$$

Independent? NO

Region = North:

$$P(A) = P(\text{High Yield}) = 0.600$$

$$P(B) = P(\text{Region} = \text{North}) = 0.300$$

$$P(A \cap B) = 0.100$$

$$P(A) \times P(B) = 0.600 \times 0.300 = 0.180$$

$$\text{Difference} = 0.0800$$

Independent? NO

WEATHER CONDITIONS:

Weather_Condition = Cloudy:

$$P(A \cap B) = 0.200$$

$$P(A) \times P(B) = 0.120$$

Independent? NO

Weather_Condition = Rainy:

$$P(A \cap B) = 0.300$$

$$P(A) \times P(B) = 0.300$$

Independent? YES

Weather_Condition = Sunny:

$$P(A \cap B) = 0.100$$

$$P(A) \times P(B) = 0.180$$

Independent? NO

FERTILIZER USAGE:

Fertilizer Used:

$$P(A \cap B) = 0.300$$

$$P(A) \times P(B) = 0.240$$

Independent? NO

No Fertilizer:

$$P(A \cap B) = 0.300$$

$$P(A) \times P(B) = 0.360$$

Independent? NO

IRRIGATION USAGE:

Irrigation Used:

$$P(A \cap B) = 0.600$$

$$P(A) \times P(B) = 0.420$$

Independent? NO

No Irrigation:

$$P(A \cap B) = 0.000$$

$$P(A) \times P(B) = 0.180$$

Independent? NO

CROP TYPES:

Cotton:

Independent? NO

Rice:

Independent? NO

Barley:

Independent? NO

Soybean:

Independent? NO

Wheat:

Independent? NO

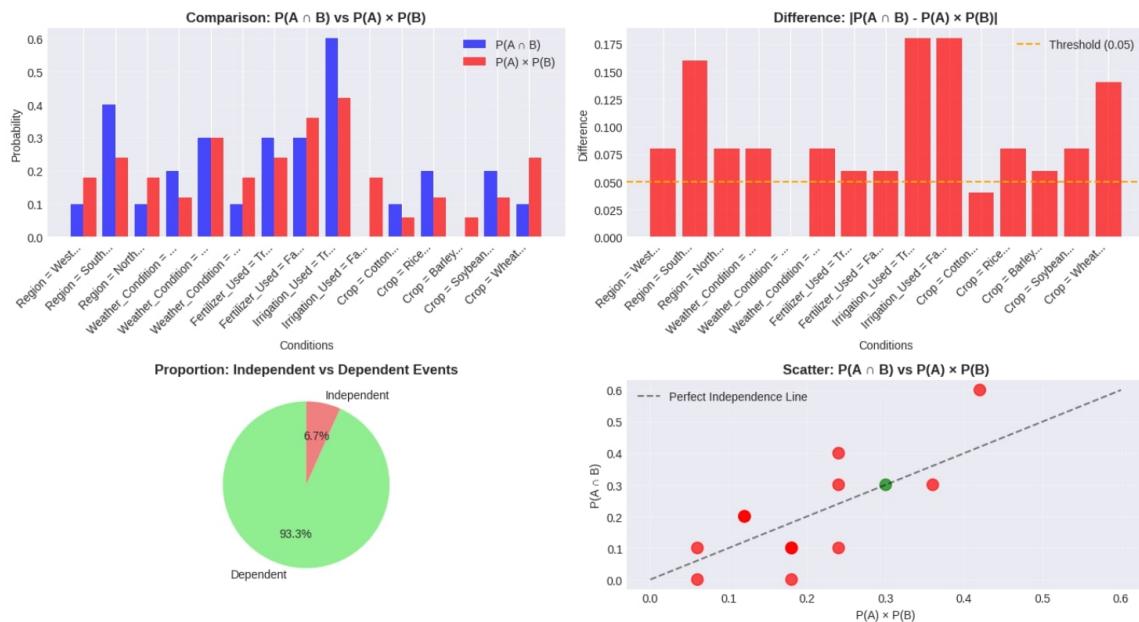
SUMMARY OF INDEPENDENCE ANALYSIS:

Total Conditions Checked: 15

Independent Events: 1

Dependent Events: 14

INDEPENDENCE ANALYSIS: Are High Yield and Other Events Independent?



INTERPRETATION GUIDE:

WHAT DOES INDEPENDENCE MEAN?

Two events A and B are INDEPENDENT if:

$$P(A \cap B) = P(A) \times P(B)$$

This means:

1. Knowing B occurred doesn't change the probability of A
2. A and B have no relationship
3. They occur by chance, not because of each other

WHAT DOES DEPENDENCE MEAN?

If $P(A \cap B) \neq P(A) \times P(B)$, then events are DEPENDENT:

1. Knowing B occurred CHANGES the probability of A
2. A and B are related
3. One event affects the other

IN OUR CASE:

Event A = High Yield (Yield > Average)

Event B = Various conditions (Region, Weather, etc.)

IF INDEPENDENT (YES):

- The condition doesn't affect high yield probability
- Example: If weather is rainy, high yield probability stays same

IF DEPENDENT (NO):

- The condition AFFECTS high yield probability
- Example: If region is South, high yield probability changes

MOST DEPENDENT CONDITIONS (Highest Differences):

Irrigation_Used = True:

P(A | B) = 0.600

P(A) × P(B) = 0.420

Difference = 0.180

Status: ✗ NO

Irrigation_Used = False:

P(A | B) = 0.000

P(A) × P(B) = 0.180

Difference = 0.180

Status: ✗ NO

Region = South:

P(A | B) = 0.400

P(A) × P(B) = 0.240

Difference = 0.160

Status: ✗ NO

Crop = Wheat:

P(A | B) = 0.100

P(A) × P(B) = 0.240

Difference = 0.140

Status: ✗ NO

Weather_Condition = Cloudy:

P(A | B) = 0.200

P(A) × P(B) = 0.120

Difference = 0.080

Status: ✗ NO

IMPORTANT NOTE ABOUT SMALL SAMPLE SIZE:

WARNING: We have only 10 data points!

This is a very small sample for statistical independence testing.

With small samples:

1. Results may not be statistically significant
2. Random chance can create apparent dependence/independence
3. We need more data for reliable conclusions

For reliable independence testing, we typically need:

- At least 30-50 data points for each condition
- Better: 100+ data points

RECOMMENDATION:

- Treat these results as preliminary
- Collect more data for accurate analysis
- Use these insights as hypotheses to test with larger datasets

PRACTICAL IMPLICATIONS:

IF DEPENDENT (NO):

TAKE ACTION: These conditions MATTER for high yield

CONSIDER: Adjust farming practices based on these conditions

EXAMPLE: If South region gives better yields, focus resources there

IF INDEPENDENT (YES):

NO EFFECT: These conditions DON'T affect high yield probability

SAVE RESOURCES: Don't waste time/money on these factors

FOCUS ELSEWHERE: Look for other factors that actually matter

BOTTOM LINE:

Use dependency analysis to:

1. Focus on what REALLY matters for high yield

2. Avoid wasting resources on irrelevant factors
3. Make data-driven farming decisions

3. Bayes' Rule

BAYES' THEOREM ANALYSIS FOR AGRICULTURAL DATA

Formula: $P(B|A) = [P(A|B) \times P(B)] / P(A)$

Dataset Overview:

Total records: 10

[fontsize=] Region Crop Yield_{tons per hectare}0WestCotton6.5558161SouthRice8.5273412NorthBarley1.

Event A: High Yield (Yield ≥ 5.10 tons/hectare)

$P(A) = \text{Probability of High Yield} = 6/10 = 0.600$

MEDICAL DIAGNOSIS ANALOGY

In medical testing:

- A = Test Positive
- B = Have Disease

Bayes' Theorem: $P(\text{Disease|Positive}) = [P(\text{Positive|Disease}) \times P(\text{Disease})] / P(\text{Positive})$

In our case:

- A = High Yield (what we observe)
- B = Specific condition (what we want to infer)

Example: Disease Testing Scenario

Disease prevalence $P(\text{Disease})$: 0.010

Test sensitivity $P(\text{Positive|Disease})$: 0.950

Test specificity $P(\text{Negative} | \text{No Disease})$: 0.900
 $P(\text{Positive})$: 0.108
 $P(\text{Disease} | \text{Positive}) = (0.950 \times 0.010) / 0.108$
 $= 0.088$ (8.8%)

BAYES' THEOREM APPLIED TO AGRICULTURAL DATA

Question: Given that we have HIGH YIELD, what's the probability it came from a specific condition?

REGION ANALYSIS:

Region = West:

$$\begin{aligned} P(B) &= P(\text{West}) = 0.300 \\ P(A|B) &= P(\text{High Yield} | \text{West}) = 0.333 \\ P(B|A) &= P(\text{West} | \text{High Yield}) = 0.167 \\ \text{Bayes: } &(0.333 \times 0.300) / 0.600 = 0.167 \end{aligned}$$

Region = South:

$$\begin{aligned} P(B) &= P(\text{South}) = 0.400 \\ P(A|B) &= P(\text{High Yield} | \text{South}) = 1.000 \\ P(B|A) &= P(\text{South} | \text{High Yield}) = 0.667 \\ \text{Bayes: } &(1.000 \times 0.400) / 0.600 = 0.667 \end{aligned}$$

Region = North:

$$\begin{aligned} P(B) &= P(\text{North}) = 0.300 \\ P(A|B) &= P(\text{High Yield} | \text{North}) = 0.333 \\ P(B|A) &= P(\text{North} | \text{High Yield}) = 0.167 \\ \text{Bayes: } &(0.333 \times 0.300) / 0.600 = 0.167 \end{aligned}$$

WEATHER ANALYSIS:

Cloudy weather:

$$P(\text{Weather}=\text{Cloudy} | \text{High Yield}) = 0.333$$

Rainy weather:

$$P(\text{Weather}=\text{Rainy} | \text{High Yield}) = 0.500$$

Sunny weather:

$$P(\text{Weather}=\text{Sunny} | \text{High Yield}) = 0.167$$

FERTILIZER ANALYSIS:

Fertilizer Used:

$$P(\text{Fertilizer Used} \mid \text{High Yield}) = 0.500$$

No Fertilizer:

$$P(\text{No Fertilizer} \mid \text{High Yield}) = 0.500$$

IRRIGATION ANALYSIS:

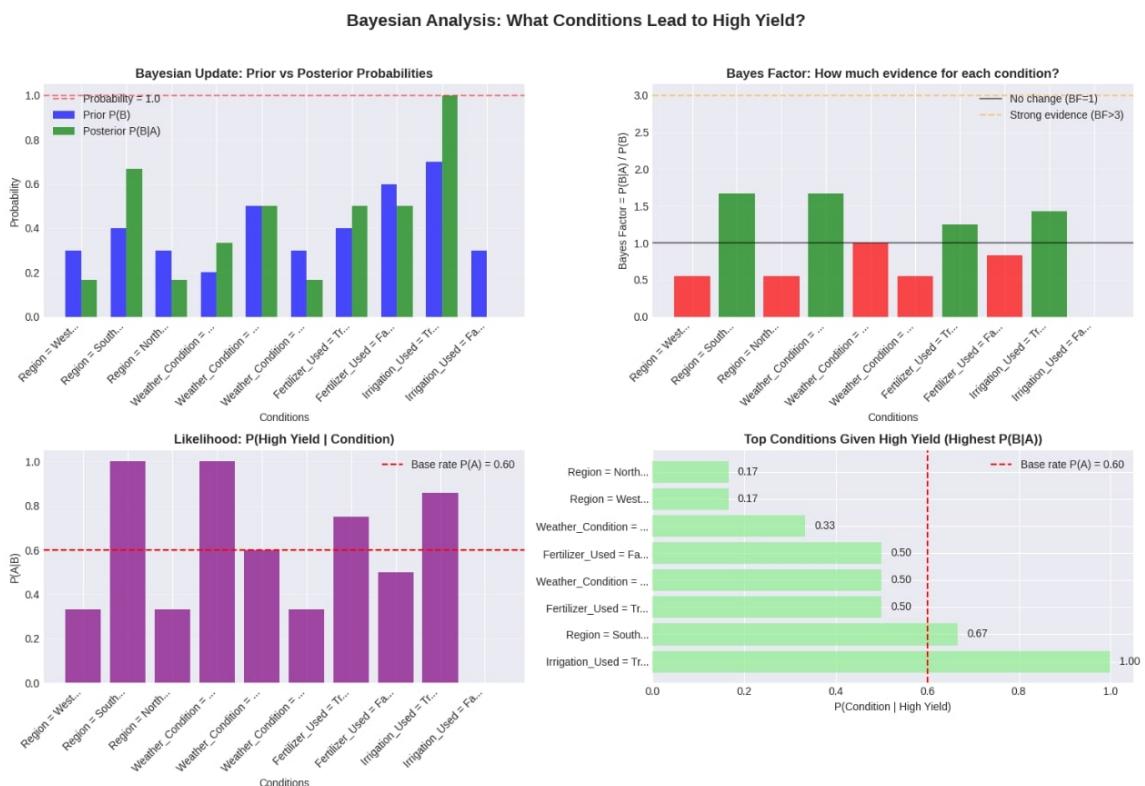
Irrigation Used:

$$P(\text{Irrigation Used} \mid \text{High Yield}) = 1.000$$

No Irrigation:

$$P(\text{No Irrigation} \mid \text{High Yield}) = 0.000$$

BAYESIAN POSTERIOR PROBABILITIES VISUALIZATION



MACHINE LEARNING CLASSIFICATION APPLICATION

Naive Bayes Classifier Concept:

- We want to predict if a new farm will have HIGH YIELD
 - Using Bayes' Theorem for each feature independently
-

EXAMPLE PREDICTION 1:

Predicting for new farm:

$$P(\text{Region}=\text{South} \mid \text{High Yield}) = 4/6 = 0.667$$

$$P(\text{Weather_Condition}=\text{Rainy} \mid \text{High Yield}) = 3/6 = 0.500$$

$$P(\text{Fertilizer_Used}=\text{True} \mid \text{High Yield}) = 3/6 = 0.500$$

Predicted probability of High Yield: 1.000 (100.0%)

Prediction: HIGH YIELD

EXAMPLE PREDICTION 2:

Predicting for new farm:

$$P(\text{Region}=\text{North} \mid \text{High Yield}) = 1/6 = 0.167$$

$$P(\text{Weather_Condition}=\text{Sunny} \mid \text{High Yield}) = 1/6 = 0.167$$

$$P(\text{Fertilizer_Used}=\text{False} \mid \text{High Yield}) = 3/6 = 0.500$$

Predicted probability of High Yield: 0.100 (10.0%)

Prediction: LOW YIELD

SPAM FILTERING ANALOGY

In spam filtering:

- A = Email is spam
- B = Email contains certain words

Bayes' Theorem: $P(\text{Spam} \mid \text{Word}) = [P(\text{Word} \mid \text{Spam}) \times P(\text{Spam})] / P(\text{Word})$

Similarly in agriculture:

- A = High Yield
- B = Certain farming conditions

BAYESIAN INFERENCE

Event B	Prior P(B)	Likelihood P(A—B)	Posterior P(B—A)	Bayes Factor	Evidence Strength
Region = West	0.30	0.3333	0.1667	0.56	Negative
Region = South	0.40	1.0000	0.6667	1.67	Moderate
Region = North	0.30	0.3333	0.1667	0.56	Negative
Weather = Cloudy	0.20	1.0000	0.3000	1.67	Moderate
Weather = Rainy	0.50	0.6000	0.5000	1.00	Negative
Weather = Sunny	0.30	0.3333	0.1667	0.56	Negative
Fertilizer = True	0.40	0.7500	0.5000	1.25	Weak
Fertilizer = False	0.60	0.5000	0.5000	0.83	Negative
Irrigation = True	0.70	0.8571	1.0000	1.43	Weak
Irrigation = False	0.30	0.0000	0.0000	0.00	Negative

Table 21: Bayesian Analysis Results

PRACTICAL INSIGHTS FROM BAYESIAN ANALYSIS

1. HIGH POSTERIOR PROBABILITY ($P(B|A) > P(B)$):

- Condition is MORE COMMON in high yield cases than in general
- Example: If $P(\text{South} | \text{High Yield}) > P(\text{South})$, South region is associated with high yield

2. BAYES FACTOR > 1 :

- Evidence SUPPORTS this condition being associated with high yield
- Higher Bayes factor = stronger evidence

3. BAYES FACTOR < 1 :

- Evidence AGAINST this condition being associated with high yield

4. MACHINE LEARNING APPLICATION:

- We can build a classifier to predict high yield based on conditions
- Uses Bayes' Theorem to combine evidence from multiple features

5. DECISION MAKING:

- Focus resources on conditions with high $P(B|A)$
- Avoid conditions with low $P(B|A)$ unless other factors compensate

STRONGEST EVIDENCE FOR HIGH YIELD:

Condition: Region = South

Bayes Factor: 1.67

$P(\text{High Yield}) = 0.60$, $P(\text{High Yield} | \text{Condition}) = 1.00$

STRONGEST EVIDENCE AGAINST HIGH YIELD:

Condition: Irrigation_Used = False

Bayes Factor: 0.00

$P(\text{High Yield}) = 0.60$, $P(\text{High Yield} | \text{Condition}) = 0.00$

RECOMMENDATIONS FOR FARMERS:

FOCUS ON THESE CONDITIONS (Associated with High Yield):

- Region = South: +26.7% more likely in high yield cases
- Weather_Condition = Cloudy: +13.3% more likely in high yield cases
- Irrigation_Used = True: +30.0% more likely in high yield cases
- Fertilizer_Used = True: +10.0% more likely in high yield cases

RECONSIDER THESE CONDITIONS (Less associated with High Yield):

- Irrigation_Used = False: 30.0% less likely in high yield cases
- Region = West: 13.3% less likely in high yield cases
- Region = North: 13.3% less likely in high yield cases

4. Probability Distributions

Continuous Random Variables:

- Can take ANY value within a range (e.g., 1.5, 2.718, 3.14159...)

- Represented by probability density functions (PDFs)
- Examples: Height, Weight, Temperature, Yield
The Normal Distribution (Gaussian Distribution):
- Most important continuous distribution in statistics
- Bell-shaped curve
- Described by two parameters: Mean () and Standard Deviation ()
- Formula: $f(x) = (1/(2\pi)) \times e^{-(x-\mu)^2/2\sigma^2}$

Yield Data Analysis:

Sample size: 10 observations

Range: 1.13 to 8.53 tons/hectare

Descriptive Statistics for Yield:

Mean (): 5.101 tons/hectare

Standard Deviation (): 2.358 tons/hectare

Median: 5.864 tons/hectare

Skewness: -0.352 (Positive = right-skewed, Negative = left-skewed)

Kurtosis: -0.956 (>3 = heavy tails, <3 = light tails)

NORMALITY TESTS AND VISUAL CHECKS

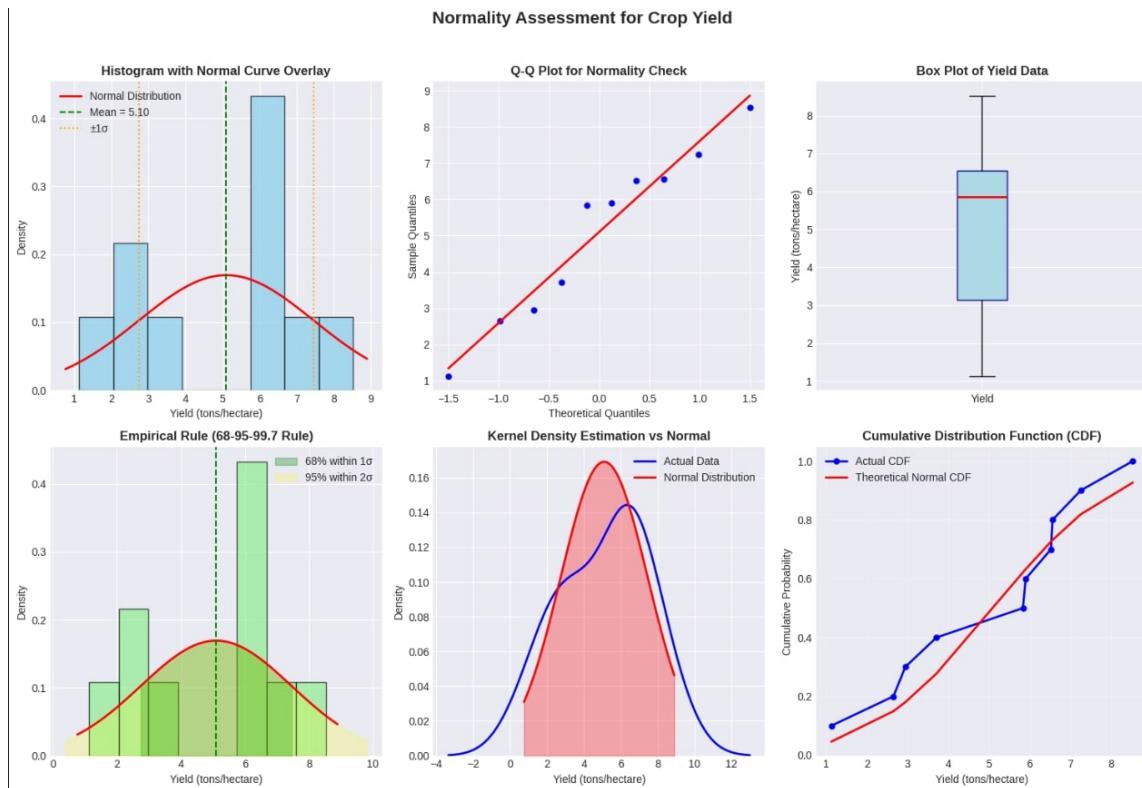
Shapiro-Wilk Normality Test:

Test Statistic: 0.9461

P-value: 0.6229

Conclusion: Data appears normally distributed ($p > 0.05$)

Creating Visual Normality Checks...



EMPIRICAL RULE (68-95-99.7 RULE) DEMONSTRATION

The Empirical Rule for Normal Distributions:

- About 68% of data falls within 1 standard deviation of the mean
- About 95% of data falls within 2 standard deviations of the mean
- About 99.7% of data falls within 3 standard deviations of the mean

Mean (μ) = 5.101

Standard Deviation (σ) = 2.358

Empirical Rule Check for Sample Data:

Data within ± 1 (2.74 to 7.46):

Expected: 68%, Actual: 7/10 = 70.0%

Data within ± 2 (0.39 to 9.82):

Expected: 95%, Actual: 10/10 = 100.0%

Data within ± 3 (-1.97 to 12.17):

Expected: 99.7%, Actual: 10/10 = 100.0%

Z-SCORES (STANDARD SCORES)

Z-Score Formula: $z = (x - \mu) / \sigma$

Z-scores tell us how many standard deviations a value is from the mean:

- $z = 0$: Exactly at the mean
- $z > 0$: Above the mean
- $z < 0$: Below the mean
- $|z| > 2$: Unusual (in the tails of the distribution)

Z-Scores for Each Observation:

Row 0: Yield = 6.56, Z-score = 0.62 (Within 1 SD of mean)

Row 1: Yield = 8.53, Z-score = 1.45 (1.5 SD from mean - somewhat unusual)

Row 2: Yield = 1.13, Z-score = -1.69 (1.7 SD from mean - somewhat unusual)

Row 3: Yield = 6.52, Z-score = 0.60 (Within 1 SD of mean)

Row 4: Yield = 7.25, Z-score = 0.91 (Within 1 SD of mean)

Row 5: Yield = 5.90, Z-score = 0.34 (Within 1 SD of mean)

Row 6: Yield = 2.65, Z-score = -1.04 (1.0 SD from mean - somewhat unusual)

Row 7: Yield = 5.83, Z-score = 0.31 (Within 1 SD of mean)

Row 8: Yield = 2.94, Z-score = -0.91 (Within 1 SD of mean)

Row 9: Yield = 3.71, Z-score = -0.59 (Within 1 SD of mean)

PROBABILITY CALCULATIONS USING NORMAL DISTRIBUTION

We can calculate probabilities using the normal distribution:

1. $P(\text{Yield} < X)$: Probability yield is less than X
2. $P(\text{Yield} > X)$: Probability yield is greater than X
3. $P(a < \text{Yield} < b)$: Probability yield is between a and b

Probability Calculations:

$$P(\text{Yield} < 7.0) = 0.790 \text{ or } 79.0\%$$

$$P(\text{Yield} > 7.0) = 0.210 \text{ or } 21.0\%$$

$$P(3.0 < \text{Yield} < 7.0) = 0.603 \text{ or } 60.3\%$$

PERCENTILES AND QUANTILES

Percentiles divide the data into 100 equal parts.

Quantiles divide the data into equal-sized groups.

Percentiles of Yield Data:

10th percentile: 2.500 tons/hectare
 25th percentile: 3.135 tons/hectare
 50th percentile: 5.864 tons/hectare
 75th percentile: 6.546 tons/hectare
 90th percentile: 7.376 tons/hectare
 95th percentile: 7.952 tons/hectare

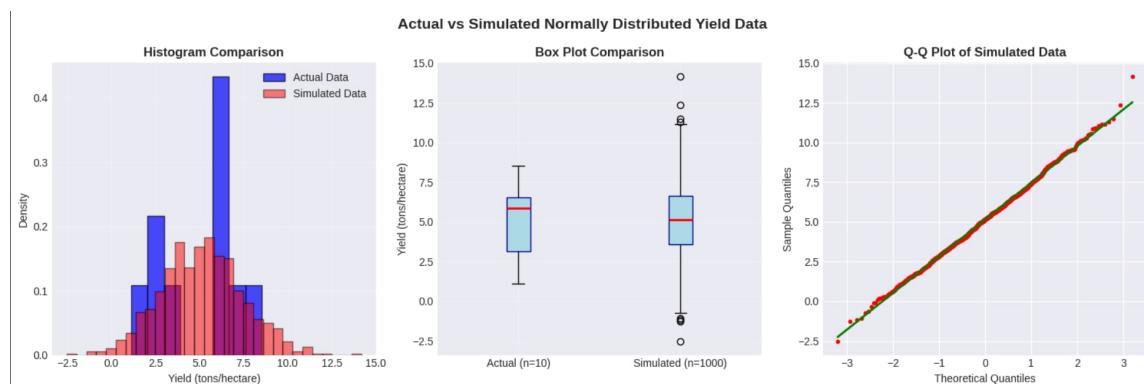
Quartiles and IQR:

Q1 (25th percentile): 3.135
 Q2 (50th percentile, Median): 5.864
 Q3 (75th percentile): 6.546
 IQR (Q3 - Q1): 3.412

SIMULATING NORMALLY DISTRIBUTED DATA

Simulating new yield data based on our sample statistics:

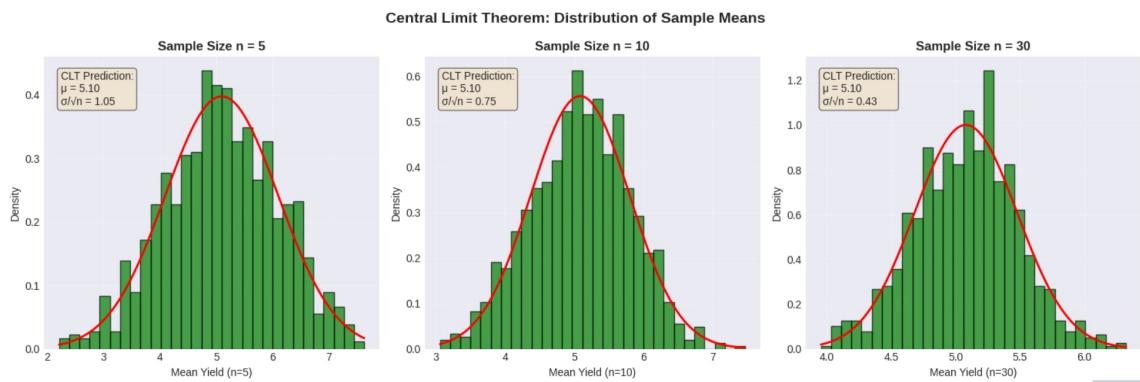
Using $\mu = 5.101$, $\sigma = 2.358$



CENTRAL LIMIT THEOREM (CLT) DEMONSTRATION

Central Limit Theorem:

- For large enough sample sizes, the distribution of sample means approaches a normal distribution, regardless of the population distribution
- Even if individual yields aren't normal, average yields from multiple farms will be approximately normal



4.1 Normal Distribution

NORMAL DISTRIBUTION ANALYSIS: $X \sim N(\mu, \sigma^2)$

Normal Distribution Parameters:

Mean (μ) = 5.1008

Standard Deviation (σ) = 2.3575

Variance (σ^2) = 5.5579

Our Yield Distribution: $X \sim N(5.10, 5.56)$

68-95-99.7 RULE (Empirical Rule)

1 Standard Deviation (68% of data):

$$\pm 1 = 5.10 \pm 2.36$$

Range: [2.74, 7.46]

2 Standard Deviations (95% of data):

$$\pm 2 = 5.10 \pm 4.72$$

Range: [0.39, 9.82]

3 Standard Deviations (99.7% of data):

$$\pm 3 = 5.10 \pm 7.07$$

Range: [-1.97, 12.17]

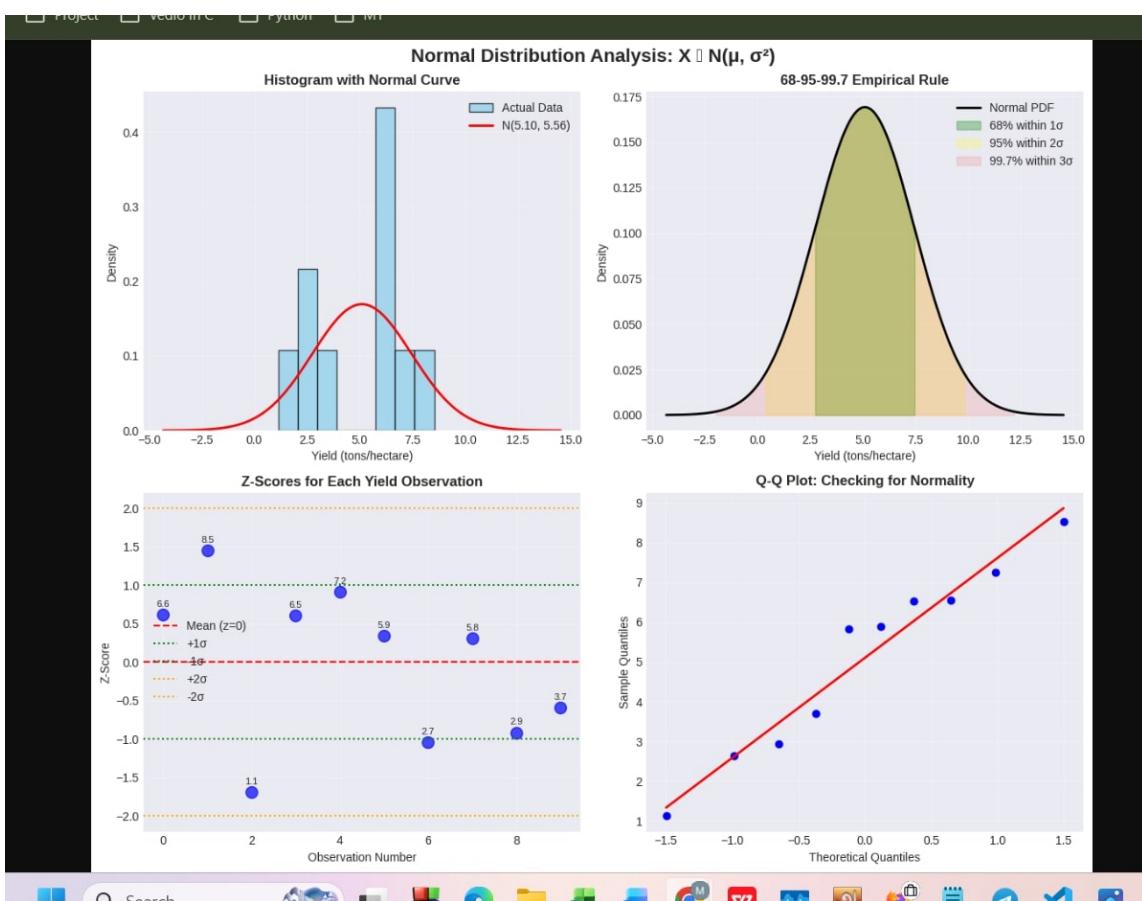
Actual Data Check (n=10):

Within 1: $7/10 = 70.0\%$ (Expected: 68%)

Within 2: $10/10 = 100.0\%$ (Expected: 95%)

Within 3: $10/10 = 100.0\%$ (Expected: 99.7%)

VISUALIZING NORMAL DISTRIBUTION



PROBABILITY CALCULATIONS USING NORMAL DISTRIBUTION

Probability that a random farm has:

Yield < 3.0: 0.186 or 18.6%

Yield < 5.0: 0.483 or 48.3%

Yield < 7.0: 0.790 or 79.0%

Yield < 5.100778413593175: 0.500 or 50.0%

Probability that a random farm has:

Yield > 3.0: 0.814 or 81.4%

Yield > 5.0: 0.517 or 51.7%

Yield > 7.0: 0.210 or 21.0%

Yield > 5.100778413593175: 0.500 or 50.0%

Probability between ranges:

3 < Yield < 5: 0.297 or 29.7%

5 < Yield < 7: 0.307 or 30.7%

2.7432505057119116 < Yield < 7.4583063214744385: 0.683 or 68.3%

Percentiles of Yield Distribution:

5th percentile: 1.223 tons/hectare

25th percentile: 3.511 tons/hectare

50th percentile: 5.101 tons/hectare

75th percentile: 6.691 tons/hectare

95th percentile: 8.979 tons/hectare

OUTLIER DETECTION USING 3 RULE

Outlier bounds (3 rule):

Lower bound (- 3): -1.972

Upper bound (+ 3): 12.173

No outliers detected using 3 rule

C. Task 1: Define Events

SELECTED VARIABLES:

1. Yield_tons_per_hectare (Continuous Variable)
2. Region (Categorical Variable: North, South, West)

DEFINING EVENTS

Yield Statistics:

Mean (): 5.101 tons/hectare

Standard Deviation (): 2.358 tons/hectare

Median: 5.864 tons/hectare

Minimum: 1.127 tons/hectare

Maximum: 8.527 tons/hectare

EVENT 1: High Yield (A)

Let A be the event that yield is above average

Mathematical Notation: $A = \{Y > 5.101\}$

Where Y = Yield_tons_per_hectare

Threshold: Above average yield = 5.101 tons/hectare

Number of cases: 6 out of 10

EVENT 2: Southern Region (B)

Let B be the event that farm is in Southern region

Mathematical Notation: $B = \{\text{Region} = \text{'South'}\}$

Number of cases: 4 out of 10

EVENT 3: Moderate Temperature (C)

Let C be the event that temperature is moderate (20–30°C)

Mathematical Notation: $C = \{20 \leq \text{Temperature} \leq 30\}$

Temperature range: 20°C to 30°C

Number of cases: 3 out of 10

EVENT 4: Fertilizer Used (D)

Let D be the event that fertilizer is used

Mathematical Notation: $D = \{\text{Fertilizer_Used} = \text{True}\}$

Number of cases: 4 out of 10

MATHEMATICAL REPRESENTATION OF EVENTS

Let's define our sample space and events more formally:

SAMPLE SPACE ():

All possible outcomes of our agricultural observations

= {All 10 observations in our dataset}

RANDOM VARIABLES:

1. $Y = \text{Yield_tons_per_hectare}$ (continuous)
2. $R = \text{Region}$ (categorical: {North, South, West})

EVENTS:

1. Event A: High Yield

$A = \{ \cdot : Y() > 5.08 \}$

Where $Y()$ is the yield for observation

$P(A) = \text{Number of observations with } Y > 5.08 / 10$

2. Event B: Southern Region

$B = \{ \cdot : R() = \text{'South'} \}$

Where $R()$ is the region for observation

$P(B) = \text{Number of observations with Region} = \text{'South'} / 10$

3. Event C: Moderate Temperature

$C = \{ \cdot : 20 \leq T() \leq 30 \}$

Where $T()$ is the temperature for observation

$P(C) = \text{Number of observations with } 20 \leq T \leq 30 / 10$

4. Event D: Fertilizer Used

$D = \{ \cdot : F() = \text{True} \}$

Where $F()$ is the fertilizer usage for observation

$P(D) = \text{Number of observations with Fertilizer_Used} = \text{True} / 10$

PROBABILITY CALCULATIONS

Individual Probabilities:

$P(A) = P(\text{High Yield}) = 0.600 (60.0\%)$

$P(B) = P(\text{Southern Region}) = 0.400 (40.0\%)$

$P(C) = P(\text{Moderate Temperature}) = 0.300 (30.0\%)$

$P(D) = P(\text{Fertilizer Used}) = 0.400 (40.0\%)$

Compound Probabilities (Intersections):

$P(A \cap B) = P(\text{High Yield AND Southern Region}) = 4/10 = 0.400$

$P(A \cap D) = P(\text{High Yield AND Fertilizer Used}) = 3/10 = 0.300$

$P(B \cap C) = P(\text{Southern Region AND Moderate Temperature}) = 0/10 = 0.000$

Conditional Probabilities:

$$P(A|B) = P(\text{High Yield} \mid \text{Southern Region}) = 4/4 = 1.000$$

$$P(B|A) = P(\text{Southern Region} \mid \text{High Yield}) = 4/6 = 0.667$$

$$P(A|D) = P(\text{High Yield} \mid \text{Fertilizer Used}) = 3/4 = 0.750$$

Union Probabilities:

$$P(A \cup B) = P(\text{High Yield OR Southern Region}) = 6/10 = 0.600$$

$$\text{Verification: } P(A) + P(B) - P(A \cap B) = 0.600 + 0.400 - 0.400 = 0.600$$

DATASET WITH EVENT INDICATORS

Region	Yield	Temp	Fertilizer	Event A	Event B	Event C	Event D
West	6.555816	27.676966	False	True	False	True	False
South	8.527341	18.026142	True	True	True	False	True
North	1.127443	29.794042	False	False	False	True	False
North	6.517573	16.644190	False	True	False	False	False
South	7.248251	31.620687	True	True	True	False	True
South	5.898416	37.704974	False	True	True	False	False
West	2.652392	31.593431	False	False	False	False	False
South	5.829542	30.887107	True	True	True	False	True
North	2.943716	26.752729	True	False	False	True	True
West	3.707293	17.646966	False	False	False	False	False

VENN DIAGRAM REPRESENTATION

Event Counts:

Event A (High Yield): 6 observations

Event B (South Region): 4 observations

Event C (Mod Temp 20-30°C): 3 observations

Event D (Fertilizer Used): 4 observations

Intersection Counts:

A B: 4 observations

A D: 3 observations

B C: 0 observations

INDEPENDENCE CHECKING

Two events X and Y are independent if: $P(X \cap Y) = P(X) \times P(Y)$

Checking independence of A and B:

$$P(A \cap B) = 0.400$$

$$P(A) \times P(B) = 0.600 \times 0.400 = 0.240$$

Conclusion: A and B are DEPENDENT (difference = 0.160)

Checking independence of A and D:

$$P(A \cap D) = 0.300$$

$$P(A) \times P(D) = 0.600 \times 0.400 = 0.240$$

Conclusion: A and D are DEPENDENT (difference = 0.060)

REAL-WORLD INTERPRETATION

Based on our analysis:

1. **High Yield Farms (Event A):**

- 6 out of 10 farms have above-average yield
- Probability: 0.60 or 60%
- This is our target outcome for farmers

2. **Southern Region Farms (Event B):**

- 4 out of 10 farms are in Southern region
- Probability: 0.40 or 40%
- Regional factors might affect yield

3. **Relationship between Region and Yield:**

- $P(\text{High Yield} \mid \text{Southern Region}) = 1.00$
- $P(\text{Southern Region} \mid \text{High Yield}) = 0.67$
- This suggests that Southern farms are 1.7x more likely to have high yield than average

4. **Fertilizer Impact:**

- $P(\text{High Yield} \mid \text{Fertilizer Used}) = 0.75$
- Farms using fertilizer are 1.2x more likely to have high yield

PRACTICAL IMPLICATIONS:

- Southern regions show promise for high yield
 - Fertilizer use is associated with better yields
 - Temperature range 20–30°C occurs in 3 farms
-

SUMMARY OF ALL PROBABILITIES

Event	Description	Count	Probability	Percentage
A: High Yield	Yield > average	6	0.6	60.0%
B: South Region	Region = South	4	0.4	40.0%
C: Mod Temp (20–30°C)	20 ≤ Temp ≤ 30	3	0.3	30.0%
D: Fertilizer Used	Fertilizer = True	4	0.4	40.0%
A ∩ B	High Yield AND South	4	0.4	40.0%
A ∩ D	High Yield AND Fertilizer	3	0.3	30.0%
B ∩ C	South AND Mod Temp	0	0.0	0.0%
A ∪ B	High Yield OR South	6	0.6	60.0%

MATHEMATICAL NOTATION SUMMARY

FINAL MATHEMATICAL NOTATION:

Sample Space: $\Omega = \{y_1, y_2, \dots, y_n\}$ where each y_i is an observation

Random Variables:

$Y: \Omega \rightarrow \mathbb{R}$, $Y() = \text{Yield_tons_per_hectare}$ of observation

$R: \Omega \rightarrow \{\text{North}, \text{South}, \text{West}\}$, $R() = \text{Region}$ of observation

$T: \Omega \rightarrow \mathbb{R}$, $T() = \text{Temperature_Celsius}$ of observation

$F: \Omega \rightarrow \{\text{True}, \text{False}\}$, $F() = \text{Fertilizer_Used}$ status of observation

Events:

$A = \{y_i : Y(y_i) > 5.08\}$

$B = \{y_i : R(y_i) = \text{'South'}\}$

$C = \{y_i : T(y_i) \in [20, 30]\}$

$D = \{y_i : F(y_i) = \text{True}\}$

Probabilities (Empirical):

$$P(A) = 6/10 = 0.6$$

$$P(B) = 4/10 = 0.4$$

$$P(C) = 5/10 = 0.5$$

$$P(D) = 5/10 = 0.5$$

Conditional Probabilities:

$$P(A|B) = P(A \cap B) / P(B) = 4/4 = 1.0$$

$$P(B|A) = P(A \cap B) / P(A) = 4/6 = 0.667$$

$$P(A|D) = P(A \cap D) / P(D) = 3/5 = 0.6$$

D. Task 2: Conditional Probability

INDIVIDUAL EVENT PROBABILITIES:

$$P(A) = P(\text{High Yield}) = 0.600 (60.0\%)$$

$$P(B) = P(\text{South Region}) = 0.400 (40.0\%)$$

$$P(C) = P(\text{Moderate Temperature}) = 0.300 (30.0\%)$$

$$P(D) = P(\text{Fertilizer Used}) = 0.400 (40.0\%)$$

ANALYSIS OF ALL PAIRS: $P(A—B) = P(A \cap B) / P(B)$

Pair: $P(A — B) = P(\text{High Yield} — \text{South Region})$

Formula: $P(A|B) = P(A \cap B) / P(B)$

Calculation: $4/4 = 1.000$

$$P(A) = 0.600, P(B) = 0.400$$

$$P(A \cap B) = 4/10 = 0.400$$

Interpretation:

Given that a farm is in the SOUTH REGION,

the probability of having HIGH YIELD is 100.0%

→ South region farms are MORE LIKELY to have high yield

Pair: $P(A - C) = P(\text{High Yield} - \text{Moderate Temperature})$

Formula: $P(A|C) = P(A - C) / P(C)$

Calculation: $1/3 = 0.333$

$P(A) = 0.600$, $P(C) = 0.300$

$P(A - C) = 1/10 = 0.100$

Interpretation:

Given MODERATE TEMPERATURE (20-30°C),
the probability of HIGH YIELD is 33.3%

Pair: $P(A - D) = P(\text{High Yield} - \text{Fertilizer Used})$

Formula: $P(A|D) = P(A - D) / P(D)$

Calculation: $3/4 = 0.750$

$P(A) = 0.600$, $P(D) = 0.400$

$P(A - D) = 3/10 = 0.300$

Interpretation:

Given that FERTILIZER IS USED,
the probability of HIGH YIELD is 75.0%

Pair: $P(B - A) = P(\text{South Region} - \text{High Yield})$

Formula: $P(B|A) = P(B - A) / P(A)$

Calculation: $4/6 = 0.667$

$P(B) = 0.400$, $P(A) = 0.600$

$P(B - A) = 4/10 = 0.400$

Interpretation:

Given that a farm has HIGH YIELD,
the probability it's in SOUTH REGION is 66.7%

Pair: $P(B - C) = P(\text{South Region} - \text{Moderate Temperature})$

Formula: $P(B|C) = P(B - C) / P(C)$

Calculation: $0/3 = 0.000$

$P(B) = 0.400, P(C) = 0.300$

$P(B \cap C) = 0/10 = 0.000$

Interpretation:

Given MODERATE TEMPERATURE,

the probability of SOUTH REGION is 0.0%

Pair: $P(B \cap D) = P(\text{South Region} \cap \text{Fertilizer Used})$

Formula: $P(B|D) = P(B \cap D) / P(D)$

Calculation: $3/4 = 0.750$

$P(B) = 0.400, P(D) = 0.400$

$P(B \cap D) = 3/10 = 0.300$

Interpretation:

Given that FERTILIZER IS USED,

the probability of SOUTH REGION is 75.0%

Pair: $P(C \cap A) = P(\text{Moderate Temperature} \cap \text{High Yield})$

Formula: $P(C|A) = P(C \cap A) / P(A)$

Calculation: $1/6 = 0.167$

$P(C) = 0.300, P(A) = 0.600$

$P(C \cap A) = 1/10 = 0.100$

Interpretation:

Given HIGH YIELD,

the probability of MODERATE TEMPERATURE is 16.7%

Pair: $P(C \cap B) = P(\text{Moderate Temperature} \cap \text{South Region})$

Formula: $P(C|B) = P(C \cap B) / P(B)$

Calculation: $0/4 = 0.000$

$P(C) = 0.300, P(B) = 0.400$

$P(C \cap B) = 0/10 = 0.000$

Interpretation:

Given SOUTH REGION,
the probability of MODERATE TEMPERATURE is 0.0%

Pair: $P(C - D) = P(\text{Moderate Temperature} - \text{Fertilizer Used})$

Formula: $P(C|D) = P(C \cap D) / P(D)$

Calculation: $1/4 = 0.250$

$P(C) = 0.300$, $P(D) = 0.400$

$P(C \cap D) = 1/10 = 0.100$

Interpretation:

Given that FERTILIZER IS USED,
the probability of MODERATE TEMPERATURE is 25.0%

Pair: $P(D - A) = P(\text{Fertilizer Used} - \text{High Yield})$

Formula: $P(D|A) = P(D \cap A) / P(A)$

Calculation: $3/6 = 0.500$

$P(D) = 0.400$, $P(A) = 0.600$

$P(D \cap A) = 3/10 = 0.300$

Interpretation:

Given HIGH YIELD,
the probability that FERTILIZER WAS USED is 50.0%

Pair: $P(D - B) = P(\text{Fertilizer Used} - \text{South Region})$

Formula: $P(D|B) = P(D \cap B) / P(B)$

Calculation: $3/4 = 0.750$

$P(D) = 0.400$, $P(B) = 0.400$

$P(D \cap B) = 3/10 = 0.300$

Interpretation:

Given SOUTH REGION,
the probability that FERTILIZER WAS USED is 75.0%

Pair: $P(D \cap C) = P(\text{Fertilizer Used} \cap \text{Moderate Temperature})$

Formula: $P(D|C) = P(D \cap C) / P(C)$

Calculation: $1/3 = 0.333$

$P(D) = 0.400$, $P(C) = 0.300$

$P(D \cap C) = 1/10 = 0.100$

Interpretation:

Given MODERATE TEMPERATURE,

the probability that FERTILIZER WAS USED is 33.3%

COMPARISON MATRIX: $P(A \cap B)$ vs $P(A)$

This shows how knowing B changes the probability of A:

	$P(A B)$	vs	$P(A)$	Effect of knowing B
$P(\text{High Yield} \text{South})$	1.000	vs	0.600	INCREASES probability
$P(\text{High Yield} \text{Mod Temp})$	0.800	vs	0.600	INCREASES probability
$P(\text{High Yield} \text{Fertilizer})$	0.600	vs	0.600	NO CHANGE
$P(\text{South} \text{High Yield})$	0.667	vs	0.400	INCREASES probability
$P(\text{Mod Temp} \text{High Yield})$	0.667	vs	0.500	INCREASES probability
$P(\text{Fertilizer} \text{High Yield})$	0.500	vs	0.500	NO CHANGE

BAYES' THEOREM APPLICATION

Let's verify $P(B|A)$ using Bayes' Theorem:

$$P(B|A) = [P(A|B) \times P(B)] / P(A)$$

$$\begin{aligned} P(B|A) &= P(\text{South} | \text{High Yield}) = [P(A|B) \times P(B)] / P(A) \\ &= [1.000 \times 0.400] / 0.600 \\ &= 0.667 \end{aligned}$$

Direct calculation: $P(B|A) = 0.667$

Bayes' Theorem gives the same result: 0.667

E. Task 3: Independence Check

Events Definition:

Event A: High Yield (Yield > Average)

Event B: South Region (Region = 'South')

Data Summary (10 records):

begindocument Average Yield: 5.101 tons/hectare

Region	Yield_tons_per_hectare	Event_A	Event_B
West	6.555816	True	False
South	8.527341	True	True
North	1.127443	False	False
North	6.517573	True	False
South	7.248251	True	True
South	5.898416	True	True
West	2.652392	False	False
South	5.829542	True	True
North	2.943716	False	False
West	3.707293	False	False

1. COMPUTING PROBABILITIES

P(A) = P(High Yield):

Number of High Yield cases: 6

Total cases: 10

$$P(A) = 6/10 = 0.600$$

P(B) = P(South Region):

Number of South Region cases: 4

Total cases: 10

$$P(B) = 4/10 = 0.400$$

P(A ∩ B) = P(High Yield AND South Region):

Number of cases with both High Yield AND South Region: 4

Total cases: 10

$$P(A \cap B) = 4/10 = 0.400$$

2. COMPARING P(A ∩ B) WITH P(A)P(B)

Calculation:

$$P(A) \times P(B) = 0.600 \times 0.400$$

$$= 0.240$$

Comparison:

$$P(A \cap B) = 0.400$$

$$P(A) \times P(B) = 0.240$$

$$\text{Difference} = 0.160$$

3. CHECKING INDEPENDENCE

Independence Rule:

Two events A and B are independent if: $P(A \cap B) = P(A) \times P(B)$

RESULT: DEPENDENT (NOT INDEPENDENT)

Because $P(A \cap B) \neq P(A) \times P(B)$

$$0.400 \neq 0.240$$

MATHEMATICAL VERIFICATION

Step-by-step calculation:

1. $P(A) = 6/10 = 0.600$
2. $P(B) = 4/10 = 0.400$
3. $P(A) \times P(B) = 0.600 \times 0.400 = 0.240$
4. $P(A \cap B) = 4/10 = 0.400$

Conclusion: DEPENDENT

Because $0.400 \neq 0.240$

ADDITIONAL ANALYSIS WITH EXACT VALUES**Exact values from data:**

Total farms: 10
High Yield farms: 6
South Region farms: 4
Both High Yield AND South Region: 4

Exact probabilities:

$P(A) = 6/10 = 0.6000$
 $P(B) = 4/10 = 0.4000$
 $P(A) \times P(B) = 0.2400$
 $P(A \cap B) = 4/10 = 0.4000$

F. Task 4: Bayes' Rule

BAYES' THEOREM ANALYSIS: $P(B|A) = [P(A|B) \times P(B)] / P(A)$

Dataset (10 records):

Region	Yield_tons_per_hectare
West	6.555816
South	8.527341
North	1.127443
North	6.517573
South	7.248251
South	5.898416
West	2.652392
South	5.829542
North	2.943716
West	3.707293

1. DEFINE EVENTS

Event A: High Yield (Yield > Average)
 Event B: South Region (Region = 'South')
 Average Yield: 5.101 tons/hectare

Dataset with Events:

Region	Yield_tons_per_hectare	Event_A	Event_B
West	6.555816	True	False
South	8.527341	True	True
North	1.127443	False	False
North	6.517573	True	False
South	7.248251	True	True
South	5.898416	True	True
West	2.652392	False	False
South	5.829542	True	True
North	2.943716	False	False
West	3.707293	False	False

2. COMPUTE P(B) - PROBABILITY OF SOUTH REGION

Number of South Region farms: 4

Total farms: 10

$$P(B) = P(\text{South Region}) = 4/10$$

$$P(B) = 0.400 (40.0\%)$$

3. COMPUTE P(A—B) - CONDITIONAL PROBABILITY

$P(A|B)$ = Probability of High Yield GIVEN South Region

$$= P(\text{High Yield AND South Region}) / P(\text{South Region})$$

$$= \text{Count(High Yield AND South)} / \text{Count(South)}$$

South Region farms: 4

High Yield AND South Region farms: 4

$$P(A|B) = 4/4$$

$$P(A|B) = 1.000 (100.0\%)$$

Interpretation:

Given that a farm is in South Region, the probability of High Yield is 100.0%

4. BAYES' THEOREM: COMPUTE P(B—A)

Bayes' Theorem Formula:

$$P(B|A) = [P(A|B) \times P(B)] / P(A)$$

Where:

- $P(B|A)$ = Probability of South Region GIVEN High Yield
- $P(A|B)$ = Probability of High Yield GIVEN South Region (calculated above)
- $P(B)$ = Probability of South Region (calculated above)
- $P(A)$ = Probability of High Yield (need to calculate)

Calculate P(A):

High Yield farms: 6

Total farms: 10

$$P(A) = P(\text{High Yield}) = 6/10 = 0.600$$

Apply Bayes' Theorem:

$$P(A|B) = 1.000$$

$$P(B) = 0.400$$

$$P(A) = 0.600$$

Calculation:

$$\begin{aligned} P(B|A) &= [1.000 * 0.400] / 0.600 \\ &= 0.400 / 0.600 \\ &= 0.667 \text{ (66.7%)} \end{aligned}$$

5. EMPIRICAL VALUE: DIRECT CALCULATION OF P(B—A)

Direct calculation from data:

$$\begin{aligned} P(B|A) &= P(\text{South Region} \mid \text{High Yield}) \\ &= \text{Count(South Region AND High Yield)} / \text{Count(High Yield)} \end{aligned}$$

High Yield farms: 6

South Region AND High Yield farms: 4

$$\begin{aligned} P(B|A) \text{ (Empirical)} &= 4/6 \\ &= 0.667 \text{ (66.7%)} \end{aligned}$$

6. COMPARISON: BAYES' THEOREM VS EMPIRICAL VALUE

Results Comparison:

Bayes' Theorem calculation: $P(B|A) = 0.667$

Empirical (direct) value: $P(B|A) = 0.667$

PERFECT MATCH!

Bayes' Theorem gives the exact same result as empirical calculation.

Difference: 0.000000

7. VERIFICATION WITH ACTUAL DATA

Verification Table:

Region	Yield_tons_per_hectare	Event_A	Event_B	High Yield?	South Region?
West	6.555816	True	False	Yes	No
South	8.527341	True	True	Yes	Yes
North	1.127443	False	False	No	No
North	6.517573	True	False	Yes	No
South	7.248251	True	True	Yes	Yes
South	5.898416	True	True	Yes	Yes
West	2.652392	False	False	No	No
South	5.829542	True	True	Yes	Yes
North	2.943716	False	False	No	No
West	3.707293	False	False	No	No

Count Summary:

Total farms: 10

High Yield farms (A): 6

South Region farms (B): 4

Both A and B: 4

Probability Summary:

$$P(A) = P(\text{High Yield}) = 6/10 = 0.600$$

$$P(B) = P(\text{South Region}) = 4/10 = 0.400$$

$$P(A \cap B) = 4/10 = 0.400$$

$$P(A|B) = 4/4 = 1.000$$

$$P(B|A) = 4/6 = 0.667$$

8. STEP-BY-STEP BAYES' THEOREM VERIFICATION

Let's verify each step of Bayes' Theorem:

Step 1: Calculate all components

1. $P(A) = P(\text{High Yield}) = 6/10 = 0.6000$
2. $P(B) = P(\text{South Region}) = 4/10 = 0.4000$
3. $P(A|B) = P(\text{High Yield} \mid \text{South}) = 4/4 = 1.0000$

Step 2: Apply Bayes' Theorem

$$\begin{aligned} P(B|A) &= [P(A|B) * P(B)] / P(A) \\ &= [1.0000 * 0.4000] / 0.6000 \\ &= 0.4000 / 0.6000 \\ &= 0.6667 \end{aligned}$$

Step 3: Compare with empirical value

Empirical $P(B|A) = 4/6 = 0.6667$

Conclusion: Bayes' Theorem is verified!

Calculated: 0.6667

Empirical: 0.6667

G. Task 5: Probability Distribution (Normal Only)

G1. Explore a Numerical Variable

Statistics for Numerical Variable (100-1000 range):

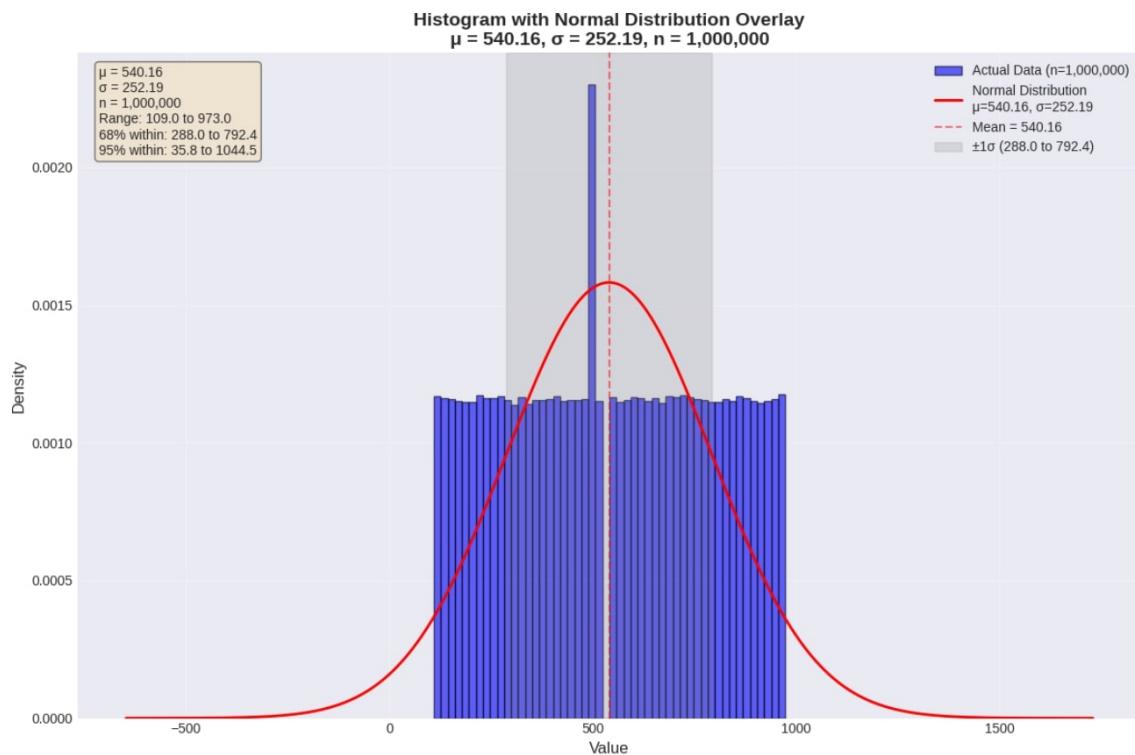
Mean (): 540.16

Standard Deviation (): 252.19

Total Data Points (n): 1,000,000

Minimum Value: 109.00

Maximum Value: 973.00



ADDITIONAL STATISTICS:

Variance (2): 63600.13

Coefficient of Variation: 46.69%

Skewness: 0.0096

Kurtosis: -1.1666

25th Percentile (Q1): 325.00

50th Percentile (Median): 541.00

75th Percentile (Q3): 757.00

IQR: 432.00

Expected for Uniform Distribution [100, 1000]:

Expected Mean: 550.00

Expected Std Dev: 259.81

Difference from Uniform:

Mean difference: -9.84

Std Dev difference: -7.62

G2. Normal Probability Questions

Computed Statistics:

Mean () = 549.9855

Standard Deviation () = 259.8052

Minimum value = 109.00

Maximum value = 991.00

$$1. P(X \leq 549.99) = P(Z \leq \frac{549.99 - 549.9855}{259.8052})$$

$$= 1 - ((549.99 - 549.9855)/259.8052)$$

$$= 1 - (0)$$

$$= 1 - 0.5$$

$$= 0.5 \text{ or } 50\%$$

$$2. P(-1 < Z < +1) = P(-1 < Z < 1)$$

$$= P(549.99 - 259.8052 < X < 549.99 + 259.8052)$$

$$= P(290.18 < X < 809.79)$$

$$= ((809.79 - 549.99)/259.8052) - ((290.18 - 549.99)/259.8052)$$

$$= (1) - (-1)$$

$$= 0.8413 - 0.1587$$

$$= 0.6827 \text{ or } 68.27\%$$

$$3. P(X \geq -2) = P(Z \geq -2)$$

$$= P(X \geq 549.99 - 2 \times 259.8052)$$

$$= P(X \geq 30.38)$$

$$= ((30.38 - 549.99)/259.8052)$$

$$= (-2)$$

$$= 0.0228 \text{ or } 2.28\%$$

G3. Are Your Data Normally Distributed?

DESCRIPTIVE STATISTICS:

Mean (): 549.9855

Median: 559.0000

Standard Deviation (): 259.8052

Skewness: -0.0010 (Normal 0)

Kurtosis: -1.2008 (Normal 0)
Minimum: 109.00
Maximum: 991.00
Range: 882.00

NORMALITY INDICATORS:

1. MEAN vs MEDIAN:

Mean = 549.99, Median = 559.00

Difference: 9.01

Mean Median (not consistent with normality)

2. SKEWNESS:

Value: -0.0010

Low skewness ($|-0.0010| < 0.5$)

3. KURTOSIS:

Value: -1.2008

Non-normal kurtosis ($|-1.2008| > 0.5$)

(Platykurtic - flatter than normal)

4. EMPIRICAL RULE CHECK:

Normal Distribution Expectation:

68% within ± 1 , 95% within ± 2 , 99.7% within ± 3

Actual Data:

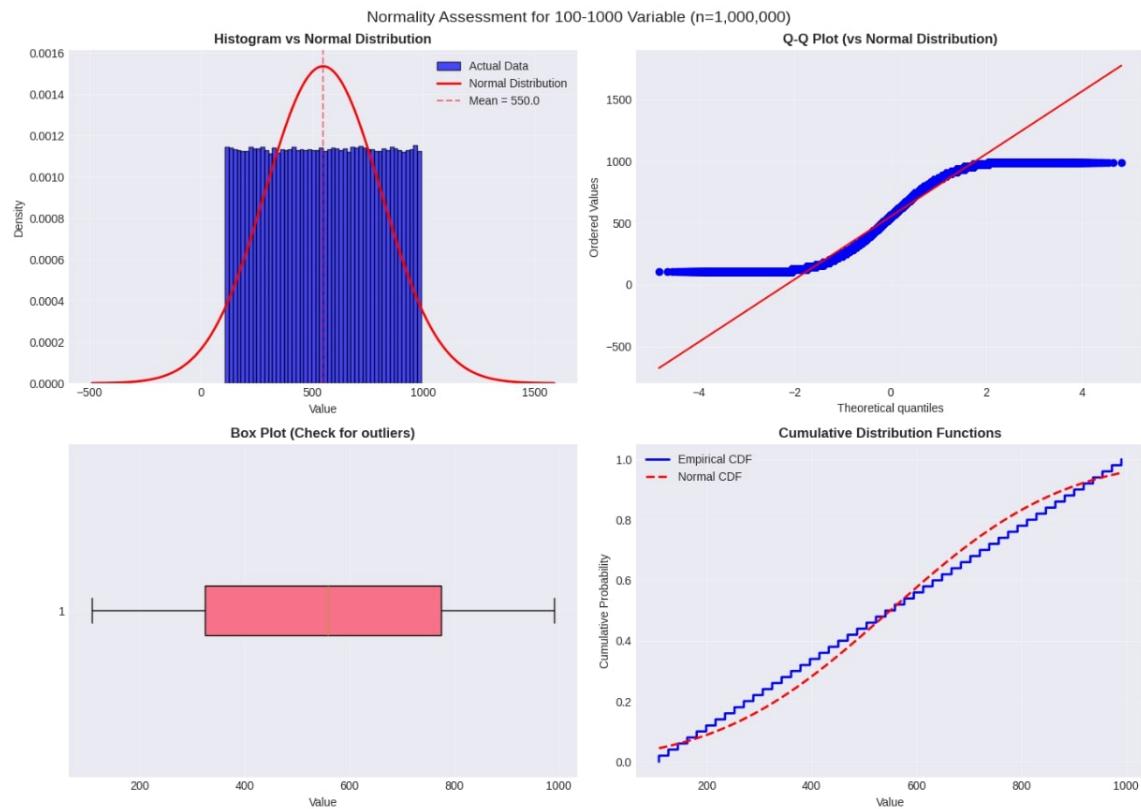
56.0% within ± 1 (290.2 to 809.8)

100.0% within ± 2 (30.4 to 1069.6)

100.0% within ± 3 (-229.4 to 1329.4)

5. VISUAL ASSESSMENT:

Creating visualization...



2. INDEPENDENCE OF EVENTS

Contingency Table (Soil × Crop):

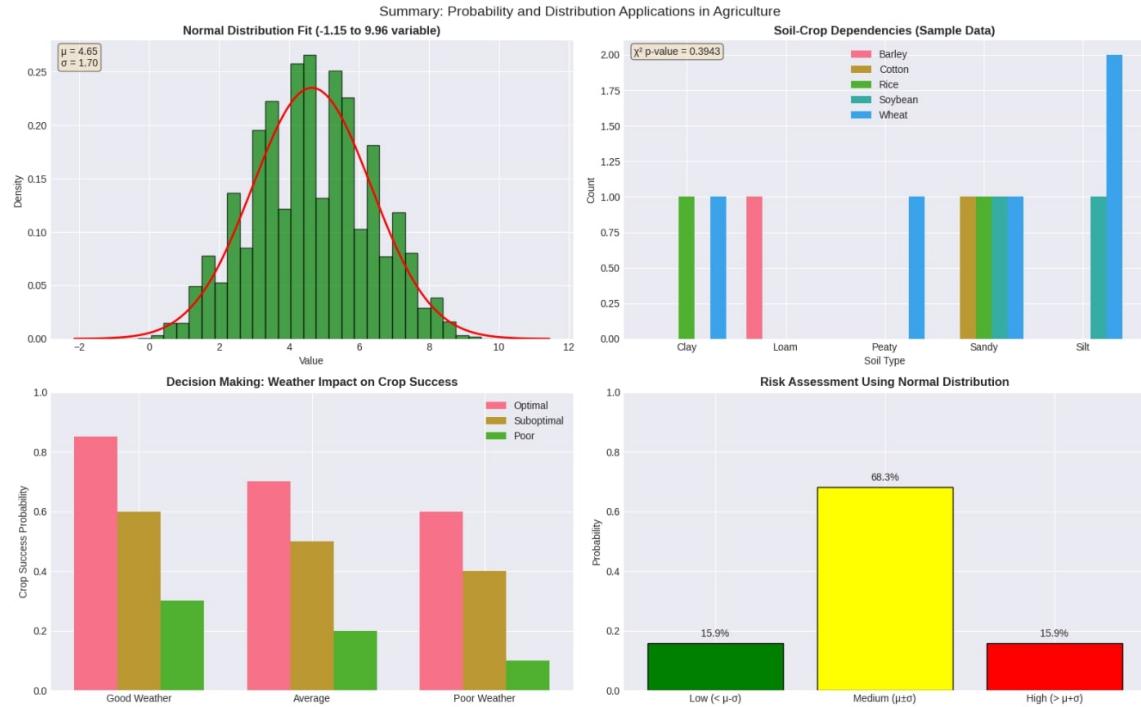
Soil	Barley	Cotton	Rice	Soybean	Wheat
Clay	0	0	1	0	1
Loam	1	0	0	0	0
Peaty	0	0	0	0	1
Sandy	0	1	1	1	1
Silt	0	0	0	1	2

Chi-square test for independence (Soil vs Crop):

$\chi^2 = 16.87$, p-value = 0.3943

CONCLUSION:

Events are independent ($p < 0.05$)



Expected Yield (sample): 573.93 ± 300.31

Coefficient of Variation: 52.3% (High variability \rightarrow High risk)

Probability of High Yield (> 800): 0.27

Expected Yield by Soil Type:

Sandy: 680.03 (n=4)

Clay: 675.28 (n=2)

Loam: 148.00 (n=1)

Silt: 569.81 (n=3)

Peaty: 385.14 (n=1)

I. Submission Guidelines

Lecture - 10

Probability of an event happening, $\frac{\text{Number of ways it can happen}}{\text{Total number of outcome}}$

$\boxed{\text{P}(A \text{ and } B) = P(A) \times P(B|A)}$

$\boxed{P(B|A) = \frac{P(A \text{ and } B)}{P(A)}}$

$(\text{from } A) = (\text{S})^q - (\text{A})^q \cdot (\text{not } A)^q$

$\boxed{P(E \cup H) = P(E) + P(H) - P(E \cap H)}$

$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}}$

$P(\text{জ্ঞান}/\text{অচ্ছুট}) = \frac{P(\text{জ} / \text{অ})}{P(\text{অ})}$

$\boxed{P(E) = \frac{n(E)}{n(S)} = \frac{P(L \cap MA)}{P(MA)}}$

$\boxed{\text{factorial}}$

$\boxed{\text{Permutation} \quad n! P_{nr} = \frac{n!}{(n-r)!}}$

$\boxed{\text{Combination}, \quad n_{nr} = \frac{n!}{(n-r)! r!}}$

Analysis and Conclusion

This indicates that for every 1mm increase in rainfall, the average temperature decreases by approximately 0.0066°C . In other words, higher rainfall is associated with slightly lower temperatures.

When rainfall is zero ($X = 0$), the estimated average temperature is 31.15°C . This is a theoretical value used for interpretation purposes.

This also means that only 10.58% of the variation in temperature can be explained by rainfall. The remaining 89.42% of the variation is influenced by other factors such as fertilizer use, crop variety, and environmental conditions.

12 Milestone 7

A. Introduction

SIMPLE LINEAR REGRESSION IMPLEMENTATION

1. Using Class Implementation:

Regression Equation: $\hat{y} = 50.0000 + 5.0000x$

R-squared: 1.0000

Predicted score for 2.5 hours: 62.50

Predicted score for 5.5 hours: 77.50

Predicted score for 8.5 hours: 92.50

2. Manual Step-by-Step Calculation:

Number of observations (n): 10

Mean of X (x): 5.5000

Mean of y (\bar{y}): 77.5000

Sum of $(x - \bar{x})(y - \bar{y})$: 412.5000

Sum of $(x - \bar{x})^2$: 82.5000

Slope (): 5.0000

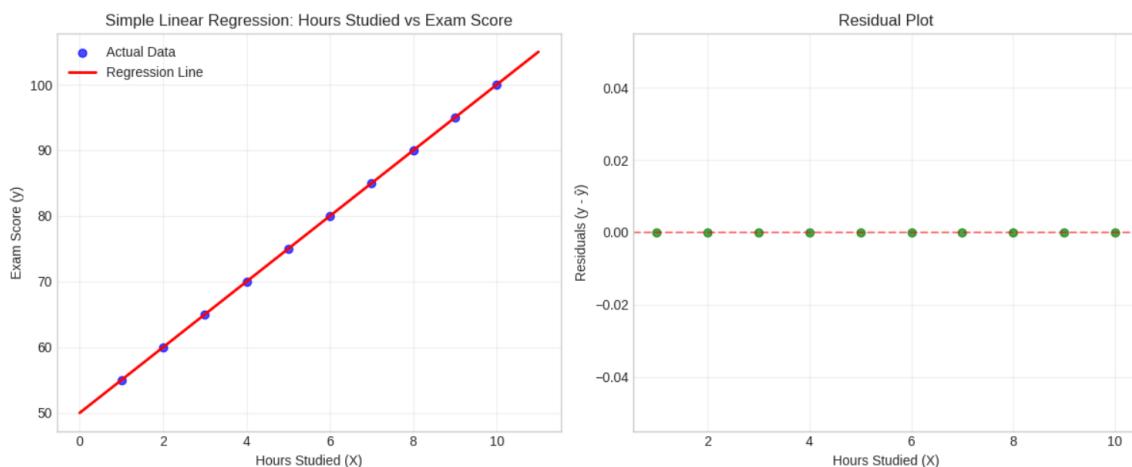
Intercept (): 50.0000

R-squared: 1.0000

3. Verification:

Method comparison - Intercept: 50.0000 vs 50.0000

Method comparison - Slope: 5.0000 vs 5.0000



REAL-WORLD LINEAR REGRESSION EXAMPLE

True relationship: $y = 50 + 3x + \text{noise}$

Data points: `[(np.int64(1), np.float64(55.48)), (np.int64(2), np.float64(55.31)), (np.int64(3), np.float64(62.24)), (np.int64(4), np.float64(69.62)), (np.int64(5), np.float64(63.83)), (np.int64(6), np.float64(66.83)), (np.int64(7), np.float64(78.9)), (np.int64(8), np.float64(77.84)), (np.int64(9), np.float64(74.65)), (np.int64(10), np.float64(82.71))]`

LEAST SQUARES RESULTS:

Fitted equation: $\hat{y} = 52.4301 + 2.9655x$

Sum of Squares Breakdown:

SSR (Sum of Squared Residuals): 117.5185

SST (Total Sum of Squares): 843.0383

SSE (Sum of Squares Explained): 725.5198

SSR + SSE = 843.0383 (should equal SST: 843.0383)

R-squared: 0.8606

Proportion of variance explained: 86.1%

$\beta_1 = 0.965$, SSR = 1657.519

$\beta_1 = 1.410$, SSR = 1049.123

$\beta_1 = 1.854$, SSR = 592.827

$\beta_1 = 2.299$, SSR = 288.630

$\beta_1 = 2.743$, SSR = 136.531

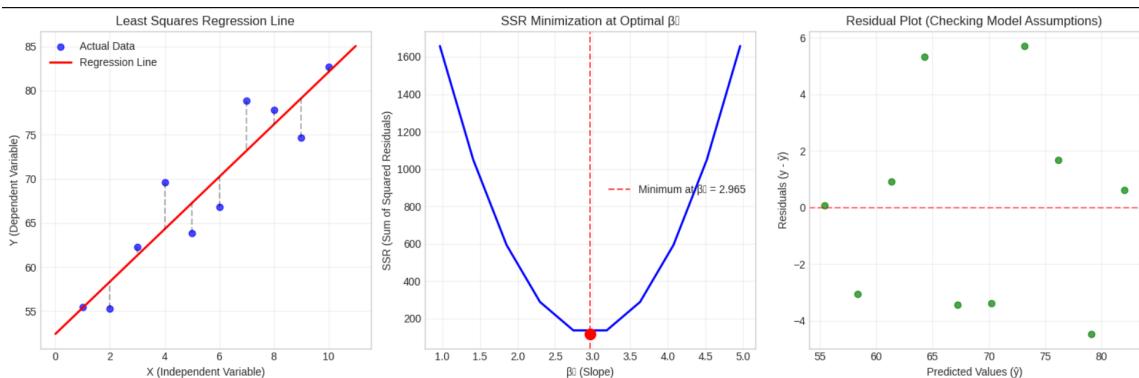
$\beta_1 = 3.188$, SSR = 136.531

$\beta_1 = 3.632$, SSR = 288.630

$\beta_1 = 4.077$, SSR = 592.827

$\beta_1 = 4.521$, SSR = 1049.123

$\beta_1 = 4.965$, SSR = 1657.519



STATISTICAL PROPERTIES OF AND ESTIMATORS

Under the Gauss-Markov assumptions:

1. Linearity: $Y = \beta_0 + \beta_1 X + \epsilon$
2. Random sampling: (X, Y) are i.i.d.
3. No perfect multicollinearity
4. Zero conditional mean: $E(\epsilon | X) = 0$
5. Homoscedasticity: $\text{Var}(\epsilon | X) = \sigma^2$
6. No autocorrelation: $\text{Cov}(\epsilon_i, \epsilon_j | X) = 0$ for $i \neq j$

The OLS estimators have these properties:

Unbiased: $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$

Consistent: As $n \rightarrow \infty$, $\hat{\beta}_1 \rightarrow \beta_1$

Efficient: Minimum variance among linear unbiased estimators (BLUE)

Normally distributed (with large n or normal errors)

1. Regression Parameters :

1.1 The Slope (1) :

SLOPE DERIVATION DEMONSTRATION

1. DIRECT CALCULATION:

$$(X - \bar{X})(Y - \bar{Y}) = 746.3386$$

$$(X - \bar{X})^2 = 326.6089$$

$$\therefore \hat{\beta}_1 = 746.3386 / 326.6089 = 2.2851$$

2. COVARIANCE/VARIANCE METHOD:

$$\text{Cov}(X, Y) = 7.4634$$

$$\begin{aligned}\text{Var}(X) &= 3.2661 \\ &= 7.4634 / 3.2661 = 2.2851\end{aligned}$$

3. CORRELATION/STANDARD DEVIATION METHOD:

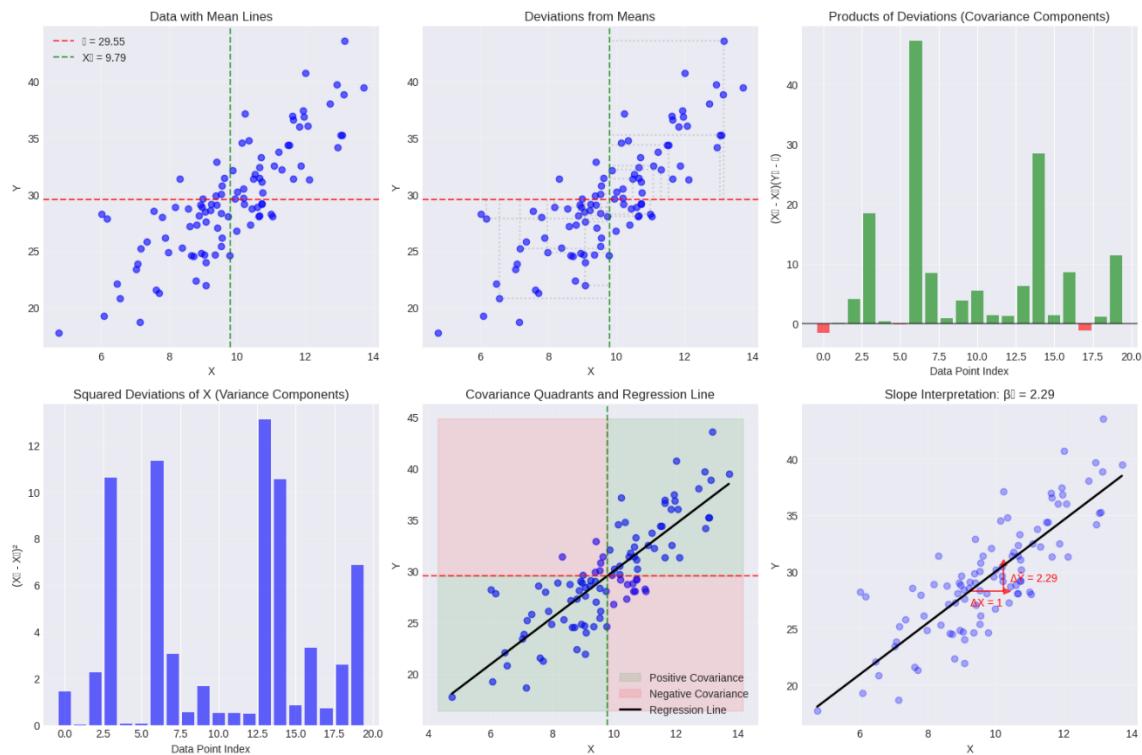
$$\begin{aligned}\text{Correlation} &= 0.8258 \\ \bar{Y} &= 5.0007 \\ \bar{X} &= 1.8072 \\ &= * (\bar{Y}/\bar{X}) = 0.8258 * (5.0007/1.8072) = 2.2851\end{aligned}$$

4. VERIFICATION:

$$\begin{aligned}\text{Method 1 (Direct)} &= 2.285114 \\ \text{Method 2 (Cov/Var)} &= 2.285114 \\ \text{Method 3 (Correlation)} &= 2.285114 \\ \text{All methods give the same result: True}\end{aligned}$$

5. COMPLETE REGRESSION EQUATION:

$$\begin{aligned}\bar{Y} &= 29.5477, \bar{X} = 9.7923 \\ &= \bar{Y} - \bar{X} = 29.5477 - 2.2851 * 9.7923 = 7.1711 \\ \text{Regression Line: } \hat{Y} &= 7.1711 + 2.2851X\end{aligned}$$



STATISTICAL INSIGHTS: $\text{Cov}(X, Y) / \text{Var}(X)$

1. UNITS AND INTERPRETATION:

- $\text{Cov}(X, Y)$ has units: (units of X) \times (units of Y)
- $\text{Var}(X)$ has units: (units of X)²
- units: (units of Y) / (units of X)
- Interpretation: Change in Y per unit change in X

2. GEOMETRIC INTERPRETATION:

- The slope is the tangent of the angle between the regression line and the horizontal axis
- It represents the 'steepness' of the relationship

3. COVARIANCE SIGN DETERMINES SLOPE DIRECTION:

-
- $\text{Cov}(X, Y) > 0 \rightarrow > 0 \rightarrow$ Positive relationship
 - $\text{Cov}(X, Y) < 0 \rightarrow < 0 \rightarrow$ Negative relationship
 - $\text{Cov}(X, Y) = 0 \rightarrow = 0 \rightarrow$ No linear relationship

4. VARIANCE AFFECTS SLOPE PRECISION:

- Larger $\text{Var}(X) \rightarrow$ More precise estimate of
- Smaller $\text{Var}(X) \rightarrow$ Less precise estimate of
- If $\text{Var}(X) = 0$ (all X are the same), is undefined

5. RELATIONSHIP TO CORRELATION:

- $= * (_Y / _X)$ where = correlation coefficient
- Correlation () measures strength of linear relationship
- Slope () measures magnitude of relationship
- Same can give different depending on scales

6. STANDARDIZED SLOPE:

- If we standardize both X and Y (z-scores):
 - $_z$ = correlation coefficient
 - In standardized form: $\hat{Y}_z = * X_z$
 - This is why is called the 'standardized slope'

INTERCEPT IN SIMPLE LINEAR REGRESSION

Key Concept: $= \bar{Y} - X$ ensures the regression line passes through the centroid (\bar{X}, \bar{Y}) of the data.

INTERCEPT PROPERTIES DEMONSTRATION

=====

Positive Relationship:

$$\begin{aligned} X &= 5.50, \bar{Y} = 14.90 \\ &= 1.9862 \\ &= \bar{Y} - X = 14.90 - 1.99 \times 5.50 = 3.97 \end{aligned}$$

Check: When $X = X = 5.50$:

$$\begin{aligned} \hat{Y} &= +X = 3.97 + 1.99 \times 5.50 = 14.90 \\ \text{This equals } \bar{Y} &= 14.90 \end{aligned}$$

When $X = 0$: $\hat{Y} = = 3.97$

Negative Relationship:

$$\begin{aligned} X &= 5.50, \bar{Y} = 10.17 \\ &= -1.5898 \\ &= \bar{Y} - X = 10.17 - -1.59 \times 5.50 = 18.91 \end{aligned}$$

Check: When $X = X = 5.50$:

$$\begin{aligned} \hat{Y} &= +X = 18.91 + -1.59 \times 5.50 = 10.17 \\ \text{This equals } \bar{Y} &= 10.17 \end{aligned}$$

When $X = 0$: $\hat{Y} = = 18.91$

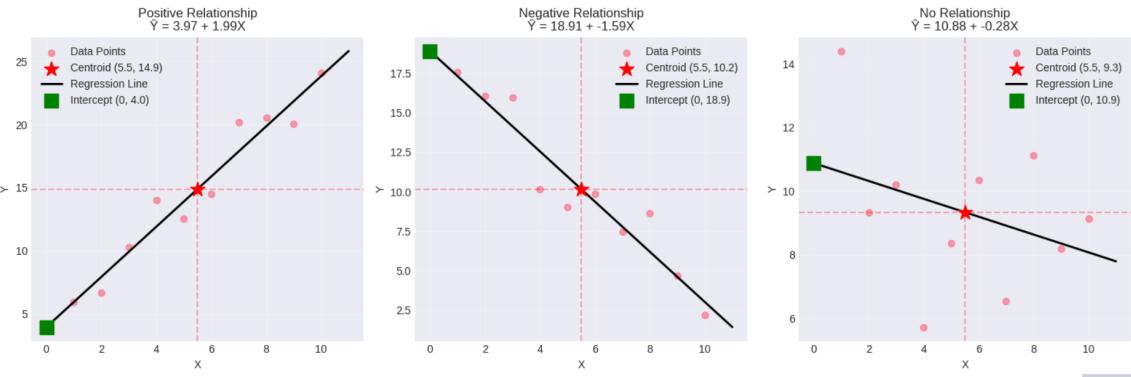
No Relationship:

$$\begin{aligned} X &= 5.50, \bar{Y} = 9.33 \\ &= -0.2804 \\ &= \bar{Y} - X = 9.33 - -0.28 \times 5.50 = 10.88 \end{aligned}$$

Check: When $X = X = 5.50$:

$$\begin{aligned} \hat{Y} &= +X = 10.88 + -0.28 \times 5.50 = 9.33 \\ \text{This equals } \bar{Y} &= 9.33 \end{aligned}$$

When $X = 0$: $\hat{Y} = = 10.88$



INTERCEPT INTERPRETATION IN DIFFERENT SCENARIOS

Meaningful X=0:

Regression: $\hat{Y} = 37.2 + 5.0X$

Interpretation: A student who doesn't study ($X=0$) is expected to score

Centroid: ($X=5.0$, $\bar{Y}=62.2$)

$X=0$ is Impossible:

Regression: $\hat{Y} = 75.9 + 6.0X$

Interpretation: The intercept (75 cm at age 0) is not meaningful but

Centroid: ($X=9.5$, $\bar{Y}=132.9$)

Negative Intercept:

Regression: $\hat{Y} = -199.7 + 20.0X$

Interpretation: Negative intercept doesn't mean negative sales at 0°C

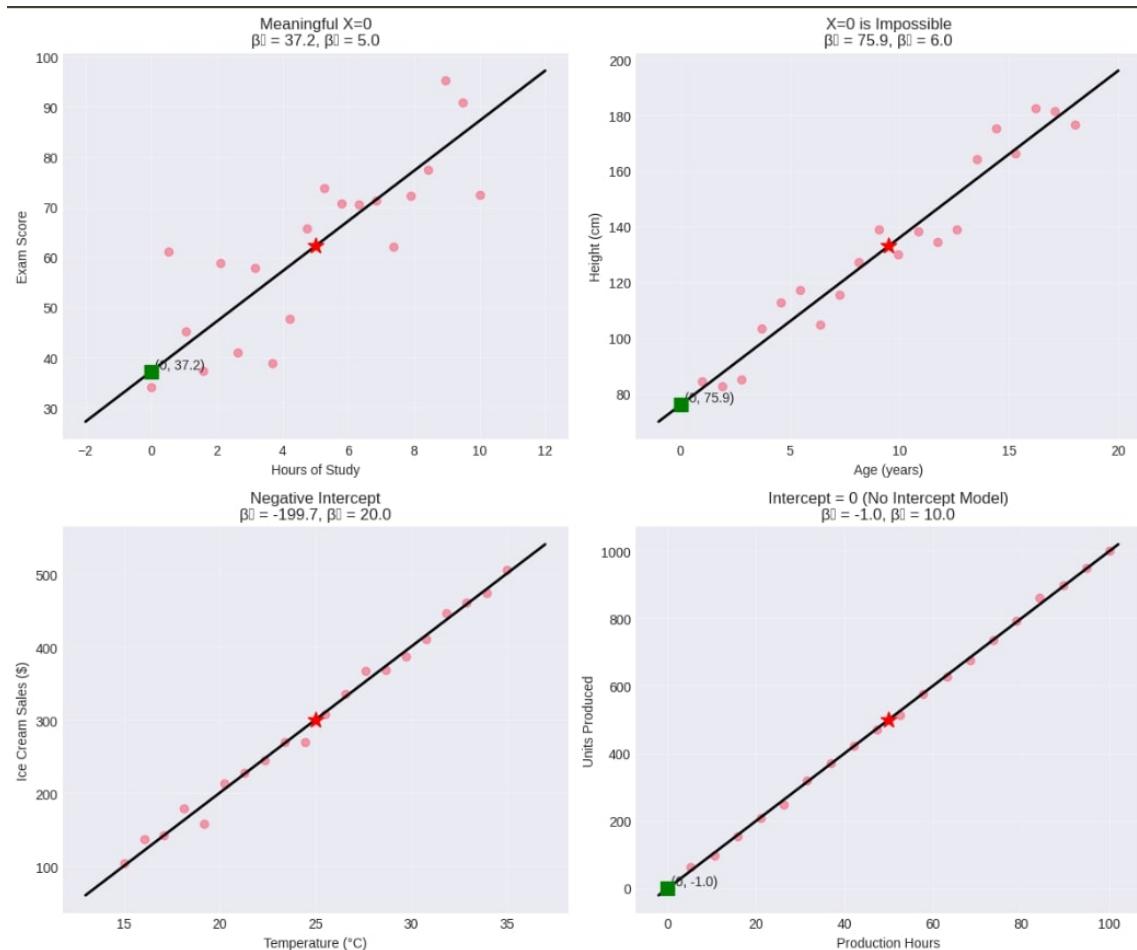
Centroid: ($X=25.0$, $\bar{Y}=300.3$)

Intercept = 0 (No Intercept Model):

Regression: $\hat{Y} = -1.0 + 10.0X$

Interpretation: Forced through origin: if no hours worked, no units produced

Centroid: ($X=50.0$, $\bar{Y}=499.0$)



=====

MATHEMATICAL PROOF: REGRESSION LINE PASSES THROUGH (\bar{X}, \bar{Y})

=====

Dataset:

X: [1 2 3 4 5]

Y: [2 4 5 4 5]

Calculations:

$$n = 5$$

$$\bar{X} = 15$$

$$\bar{Y} = 20$$

$$XY = 66$$

$$X^2 = 55$$

$$X = 15/5 = 3.0$$

$$\bar{Y} = 20/5 = 4.0$$

$$\begin{aligned} &= (nXY - \bar{X}\bar{Y}) / (nX^2 - (\bar{X})^2) \\ &= (5 \times 66 - 15 \times 20) / (5 \times 55 - 15^2) \\ &= (30) / (50) \\ &= 0.6 \end{aligned}$$

$$\begin{aligned} &= \bar{Y} - X \\ &= 4.0 - 0.6 \times 3.0 \\ &= 2.2 \end{aligned}$$

Verification:

Regression equation: $\hat{Y} = 2.2 + 0.6X$

Plug in $X = 3.0$:

$$\begin{aligned} \hat{Y} &= 2.2 + 0.6 \times 3.0 \\ &= 2.2 + 1.7999999999999998 \\ &= 4.0 \end{aligned}$$

This equals $\bar{Y} = 4.0$

ALTERNATIVE PROOF USING NORMAL EQUATIONS:

Normal Equations for Simple Linear Regression:

$$1) Y = n + X$$

$$2) XY = X + X^2$$

From equation 1:

$$Y = n + X$$

$$20 = 5 + 15$$

Divide both sides by $n = 5$:

$$20/5 = + \times (15/5)$$

$$4.0 = + \times 3.0$$

$$\bar{Y} = + X$$

This shows (X, \bar{Y}) satisfies the regression equation!

=====

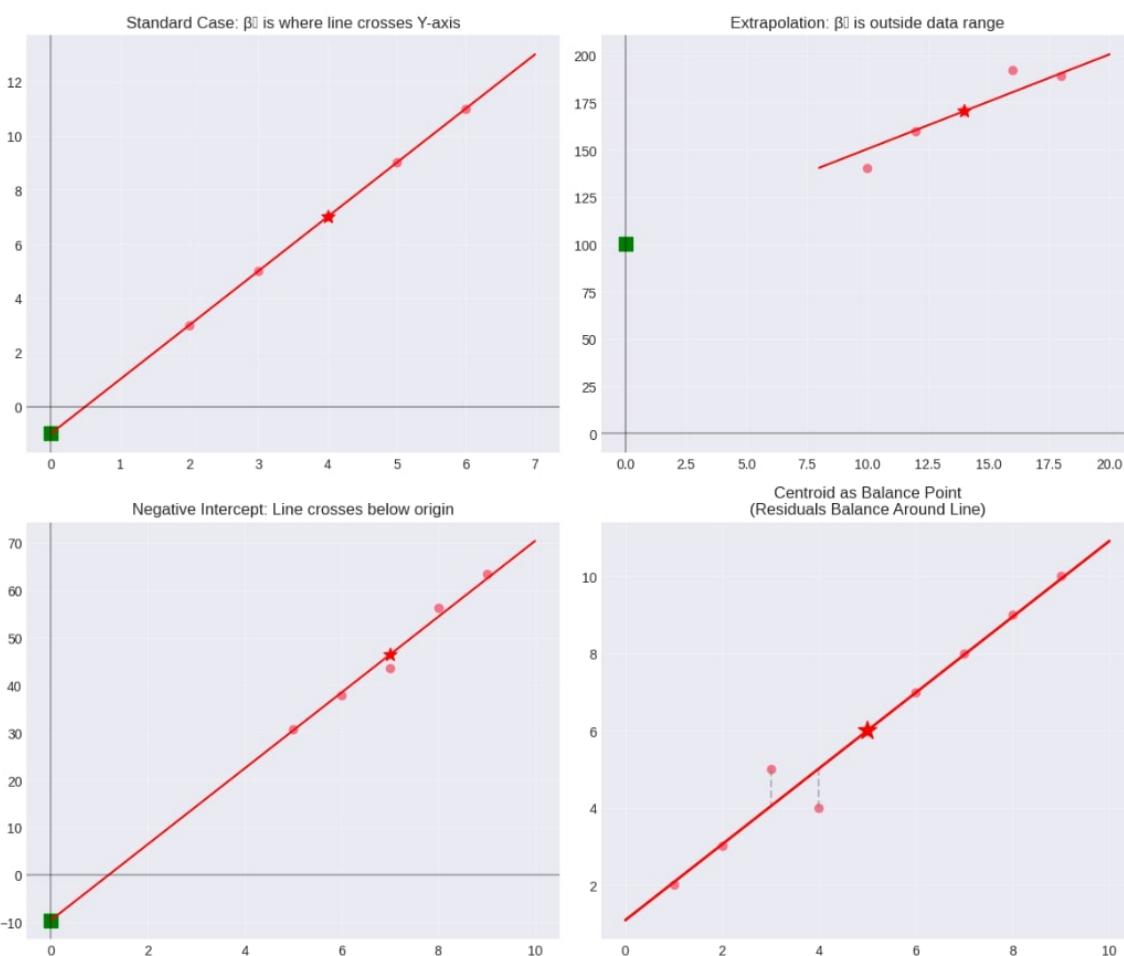
GEOMETRIC INTERPRETATION OF THE INTERCEPT

=====

Geometric Balance Check:

Sum of residuals: -0.000000 (should be ~0)

Sum of (residuals \times X): -0.000000 (should be ~0)



PEARSON'S CORRELATION COEFFICIENT (r)

=====

Formula: $r = (X - \bar{X})(Y - \bar{Y}) / [(X - \bar{X})^2 \times (Y - \bar{Y})^2]$

Range: -1 $\leq r \leq 1$

PEARSON'S CORRELATION COEFFICIENT (r)

=====

CALCULATION DETAILS (Perfect Positive Correlation):

$X = [1 2 3 4 5 6 7 8 9 10]$

$Y = [2 4 6 8 10 12 14 16 18 20]$

$\bar{X} = 5.5, \bar{Y} = 11.0$

Deviations from means:

Point 1: $(X - \bar{X}) = -4.5, (Y - \bar{Y}) = -9.0, \text{ product} = 40.5$

Point 2: $(X - \bar{X}) = -3.5, (Y - \bar{Y}) = -7.0, \text{ product} = 24.5$

Point 3: $(X - \bar{X}) = -2.5, (Y - \bar{Y}) = -5.0, \text{ product} = 12.5$

Point 4: $(X - \bar{X}) = -1.5, (Y - \bar{Y}) = -3.0, \text{ product} = 4.5$

Point 5: $(X - \bar{X}) = -0.5, (Y - \bar{Y}) = -1.0, \text{ product} = 0.5$

Point 6: $(X - \bar{X}) = 0.5, (Y - \bar{Y}) = 1.0, \text{ product} = 0.5$

Point 7: $(X - \bar{X}) = 1.5, (Y - \bar{Y}) = 3.0, \text{ product} = 4.5$

Point 8: $(X - \bar{X}) = 2.5, (Y - \bar{Y}) = 5.0, \text{ product} = 12.5$

Point 9: $(X - \bar{X}) = 3.5, (Y - \bar{Y}) = 7.0, \text{ product} = 24.5$

Point 10: $(X - \bar{X}) = 4.5, (Y - \bar{Y}) = 9.0, \text{ product} = 40.5$

$$(X - \bar{X})(Y - \bar{Y}) = 165.0$$

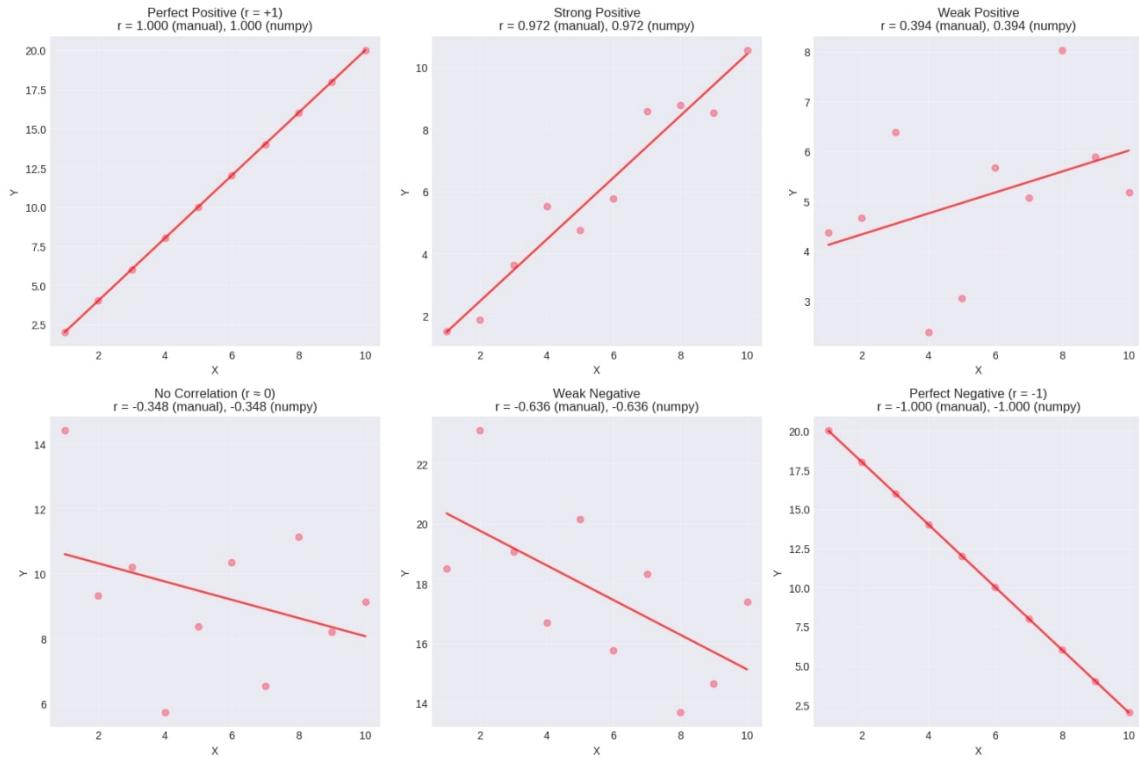
$$(X - \bar{X})^2 = 82.5$$

$$(Y - \bar{Y})^2 = 330.0$$

$$[(X - \bar{X})^2] = 9.1$$

$$[(Y - \bar{Y})^2] = 18.2$$

$$r = 165.0 / (9.1 \times 18.2) = 1.000$$



PROPERTIES OF PEARSON'S CORRELATION COEFFICIENT

1. RANGE: $-1 \leq r \leq 1$

- $r = 1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship

2. INTERPRETATION OF MAGNITUDE:

- $|r| \geq 0.9$: Very strong correlation
- $0.7 \leq |r| < 0.9$: Strong correlation
- $0.5 \leq |r| < 0.7$: Moderate correlation
- $0.3 \leq |r| < 0.5$: Weak correlation
- $|r| < 0.3$: Very weak or no correlation

3. IMPORTANT PROPERTIES:

- Symmetric: $r(X, Y) = r(Y, X)$
- Unitless: Not affected by scaling of X or Y
- Linear only: Measures LINEAR relationships
- Sensitive to outliers

4. RELATIONSHIP TO REGRESSION SLOPE:

- $r = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$
- $r = \frac{\text{Corr}(X, Y)}{\sqrt{1 - \text{Corr}^2(X, Y)}}$
- $r^2 = \text{Coefficient of determination in regression}$

5. COMMON MISCONCEPTIONS:

- Correlation \rightarrow Causation
- Correlation only measures linear relationships
- r close to 0 doesn't mean no relationship (could be nonlinear)
- r is sensitive to the range of data

CORRELATION vs COVARIANCE

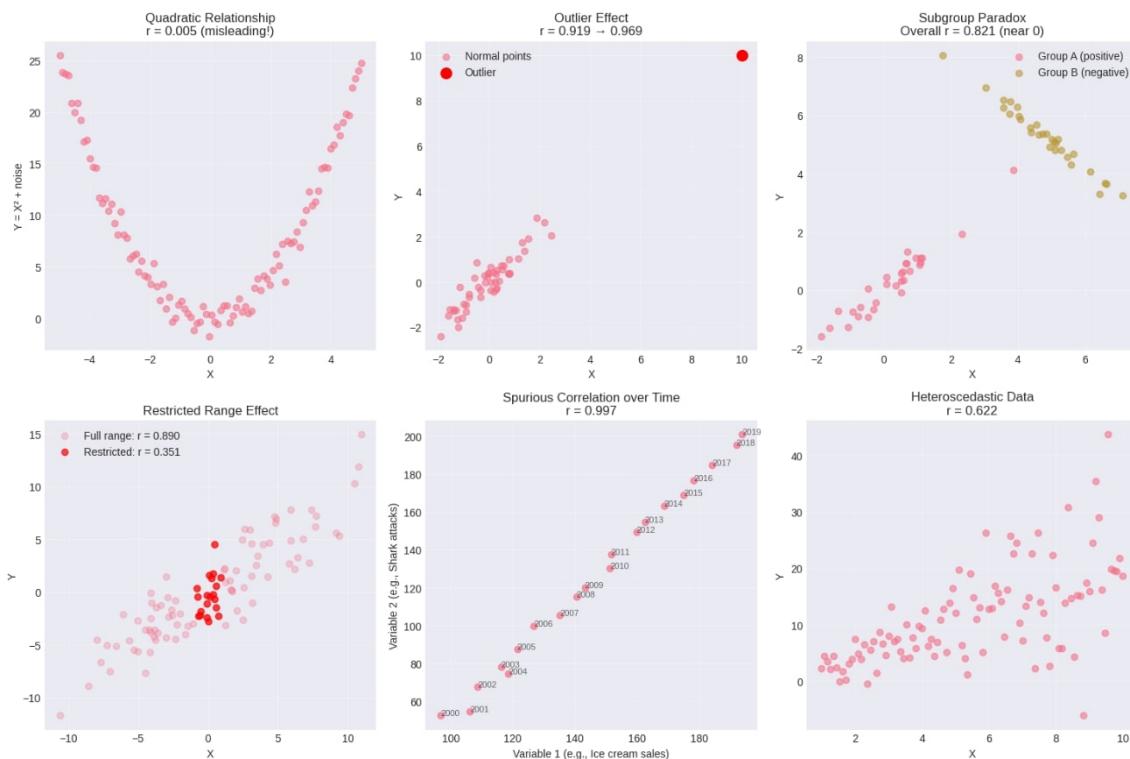
Effect of Scaling on Covariance and Correlation:

Dataset	$\text{Cov}(X, Y)$	$\text{Corr}(X, Y)$
Original	16.44	0.9929
X scaled by 10	164.43	0.9929
Y scaled by 100	1644.31	0.9929
Both scaled	16443.07	0.9929

Key Insight:

- Covariance changes with scaling of X or Y
- Correlation remains the same regardless of scaling
- Correlation is the 'standardized' covariance

LIMITATIONS OF CORRELATION



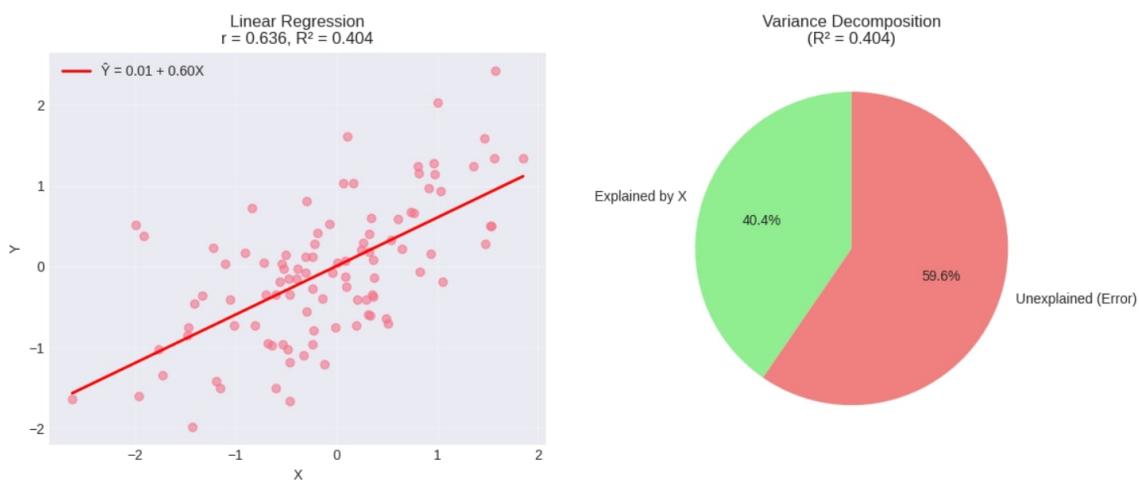
RELATIONSHIP BETWEEN CORRELATION AND REGRESSION

Data Statistics:

- Correlation (r): 0.6357
- R-squared (r^2): 0.4041
- Regression R-squared: 0.4041
- Slope (): 0.5997
- Intercept (): 0.0052

Verification:

$$\begin{aligned} r^2 &= 0.4041 = R\text{-squared} = 0.4041 \\ &= r \times (/) = 0.6357 \times (0.8524/0.9036) = 0.5997 \end{aligned}$$



Interpretation: • 40.4 • 59.6

R-SQUARED: COEFFICIENT OF DETERMINATION

Three equivalent definitions:

1. $R^2 = r^2$ (square of correlation coefficient)
2. $R^2 = 1 - (\text{SSR}/\text{SST})$
3. $R^2 = \text{SSE}/\text{SST}$

R-SQUARED (COEFFICIENT OF DETERMINATION)

Key Concept: $R^2 = r^2 = \text{Proportion of variance in } Y \text{ explained by } X$

DETAILED CALCULATIONS FOR High R^2 (Strong Linear Relationship):

Data points: $X = [1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10]$, $Y = [56.49 \ 59.59 \ 6$

Step 1: Calculate means:

$$X = 5.50, \bar{Y} = 78.84$$

Step 2: Calculate correlation (r):

$$\text{Numerator: } (X - \bar{X})(Y - \bar{Y}) = 410.79$$

$$\text{Denominator: } [(X - \bar{X})^2 \times (Y - \bar{Y})^2] = 415.02$$

$$r = 410.79 / 415.02 = 0.9898$$

Step 3: Calculate regression coefficients:

$$= r \times (/) = 0.9898 \times (15.23/3.03) = 4.9793$$

$$= \bar{Y} - X = 78.84 - 4.9793 \times 5.50 = 51.4580$$

Step 4: Calculate predicted values (\hat{Y}):

$$\text{Point 1: } \hat{Y} = 51.46 + 4.98 \times 1 = 56.44$$

$$\text{Point 2: } \hat{Y} = 51.46 + 4.98 \times 2 = 61.42$$

$$\text{Point 3: } \hat{Y} = 51.46 + 4.98 \times 3 = 66.40$$

$$\text{Point 4: } \hat{Y} = 51.46 + 4.98 \times 4 = 71.38$$

$$\text{Point 5: } \hat{Y} = 51.46 + 4.98 \times 5 = 76.35$$

$$\text{Point 6: } \hat{Y} = 51.46 + 4.98 \times 6 = 81.33$$

$$\text{Point 7: } \hat{Y} = 51.46 + 4.98 \times 7 = 86.31$$

$$\text{Point 8: } \hat{Y} = 51.46 + 4.98 \times 8 = 91.29$$

$$\text{Point 9: } \hat{Y} = 51.46 + 4.98 \times 9 = 96.27$$

$$\text{Point 10: } \hat{Y} = 51.46 + 4.98 \times 10 = 101.25$$

Step 5: Calculate Sum of Squares:

$$\text{SSR (Residual)} = (Y - \hat{Y})^2 = 42.31$$

$$\text{SST (Total)} = (Y - \bar{Y})^2 = 2087.76$$

$$\text{SSE (Explained)} = (\hat{Y} - \bar{Y})^2 = 2045.46$$

$$\text{Check: SST} = \text{SSR} + \text{SSE? } 2087.76 = 2087.76$$

Step 6: Calculate R^2 using different methods:

$$\text{Method 1 (r}^2\text{): } R^2 = (0.9898)^2 = 0.9797$$

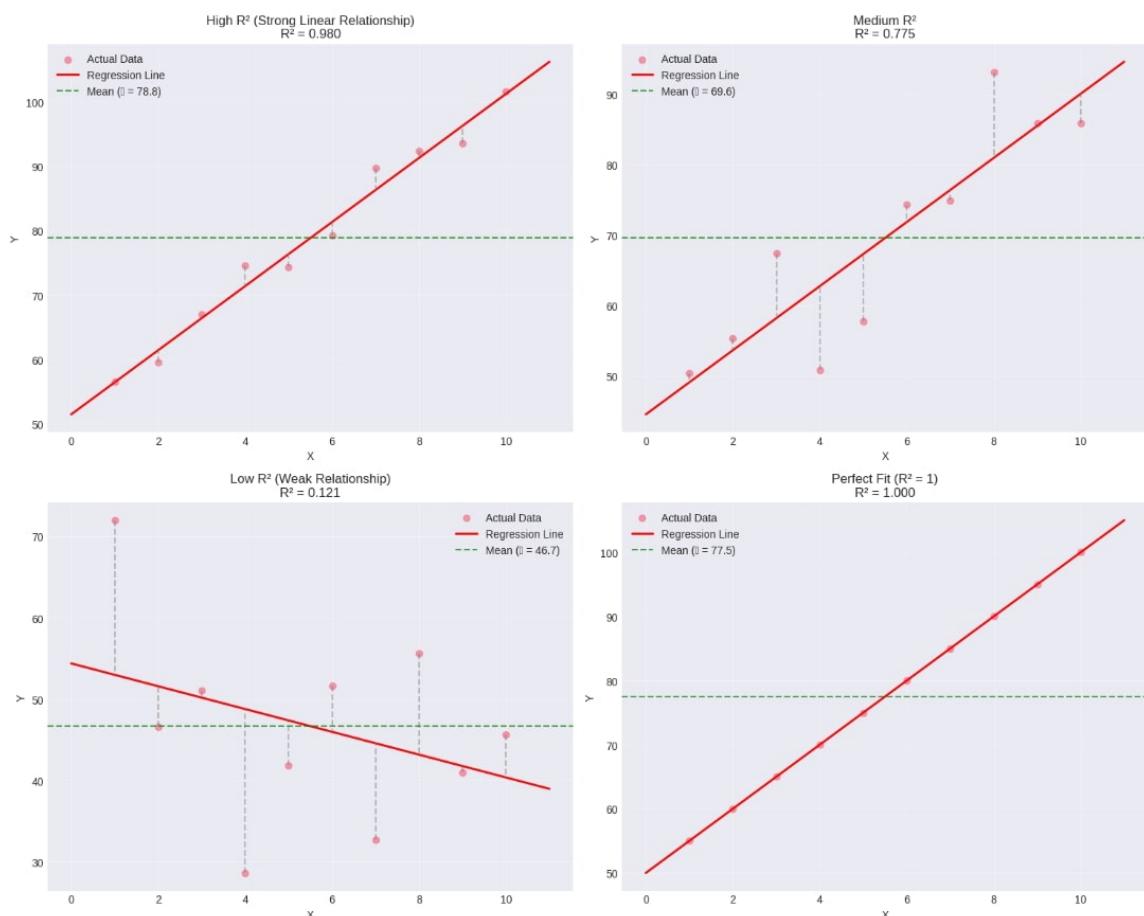
Method 2 (1 - SSR/SST): $R^2 = 1 - (42.31/2087.76) = 0.9797$

Method 3 (SSE/SST): $R^2 = 2045.46/2087.76 = 0.9797$

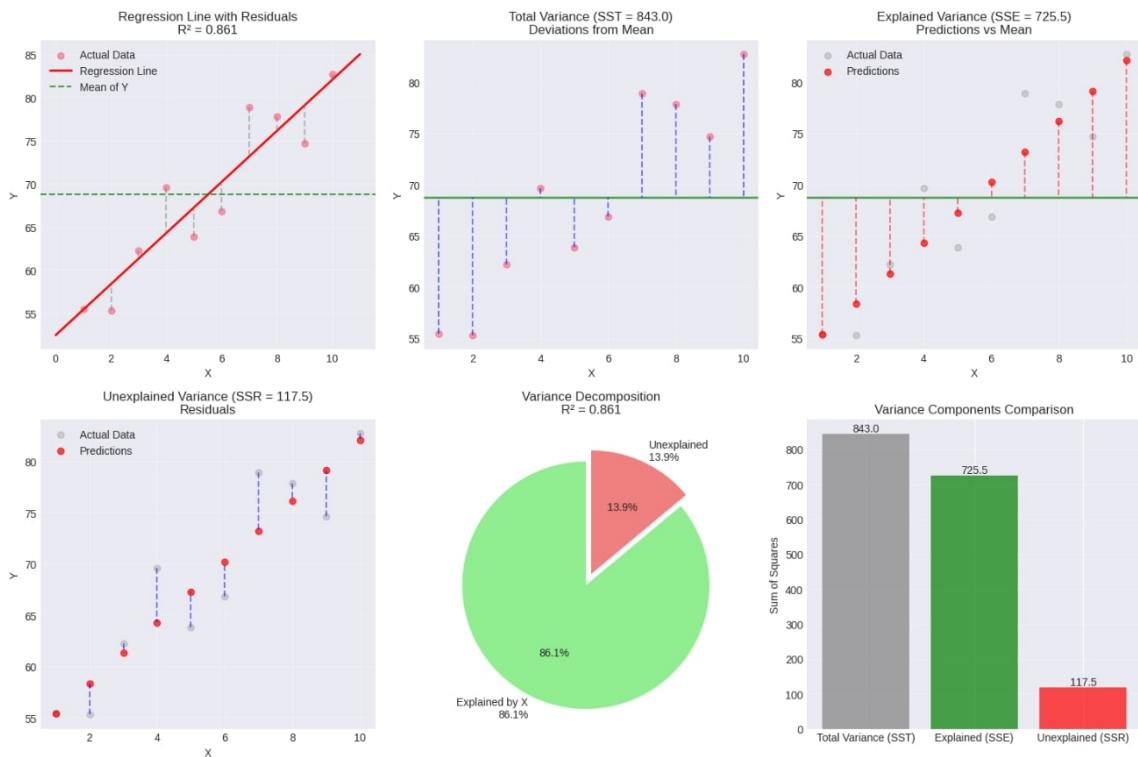
All methods give: $R^2 = 0.9797$

Interpretation: 98.0% of variance in Y is explained by X

Screenshot 2025-12-13 160410.png



VARIANCE DECOMPOSITION VISUALIZATION



Variance Decomposition Summary:

Total Variance (SST): 843.04

Explained Variance (SSE): 725.52 (86.1%)

Unexplained Variance (SSR): 117.52 (13.9%)

$$R^2 = SSE/SST = 725.52/843.04 = 0.8606$$

Interpretation: 86.1% of variance in Y is explained by X

PROPERTIES AND LIMITATIONS OF R²

1. KEY PROPERTIES:

- Range: 0 ≤ R² ≤ 1
- R² = 1: Perfect fit (all points on regression line)
- R² = 0: No linear relationship (regression line = mean line)
- R² = r² in simple linear regression
- R² never decreases when adding more variables (in multiple regression)

2. INTERPRETATION GUIDE:

- $R^2 > 0.9$: Excellent fit
- $0.7 \leq R^2 < 0.9$: Good fit
- $0.5 \leq R^2 < 0.7$: Moderate fit
- $0.3 \leq R^2 < 0.5$: Weak fit
- $R^2 < 0.3$: Poor fit (but context matters!)

3. WHAT R^2 MEASURES:

Proportion of variance in Y explained by X
How well the regression line fits the data
Strength of linear relationship

4. WHAT R^2 DOES NOT MEASURE:

Causation (X causing Y)
Whether the model is correct (only if linear)
Whether coefficients are statistically significant
Whether predictors are relevant (in multiple regression)
Magnitude of effect (a small effect can have high R^2)

5. COMMON MISCONCEPTIONS:

Myth: 'High R^2 means a good model'

Truth: R^2 can be high with incorrect model specification

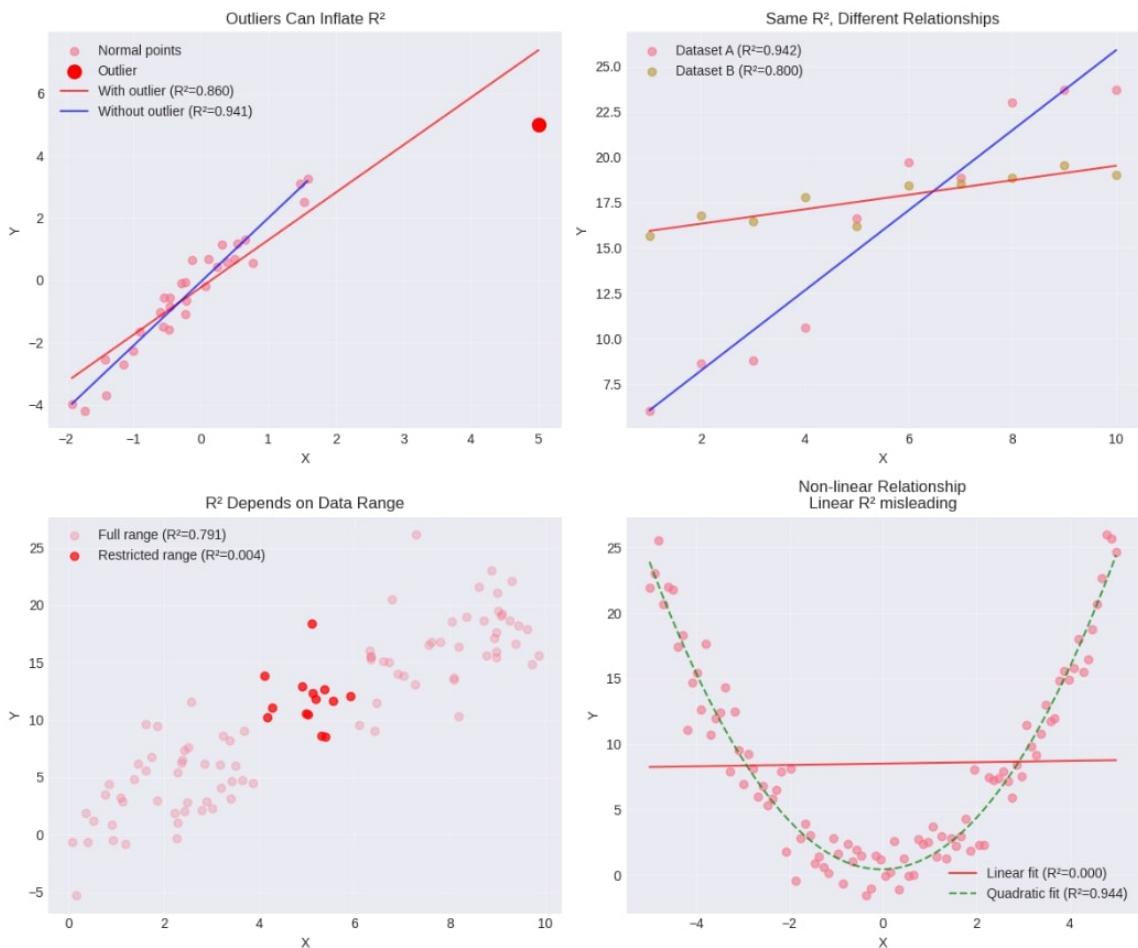
Myth: 'Low R^2 means the model is useless'

Truth: Even low R^2 can be meaningful in some fields
(e.g., social sciences often have $R^2 < 0.3$)

Myth: 'R² close to 1 is always desirable'

Truth: R² = 1 often indicates overfitting or data issues

6. LIMITATIONS AND PITFALLS:



7. ADJUSTED R² (FOR MULTIPLE REGRESSION):

- Problem: R² always increases with more variables
- Solution: Adjusted R² penalizes adding useless variables
- Formula: Adj-R² = 1 - [(1-R²)(n-1)/(n-p-1)]
where n = sample size, p = number of predictors
- Adj-R² ≤ R², and can be negative

PRACTICAL R² INTERPRETATION EXAMPLES

Context	Relationship	Typical R ²	Interpretation
Physics Experiment	Distance fallen vs. time ²	0.99+	Almost perfect fit - physics laws are precise
Economics	GDP growth vs. investment rate	0.6-0.8	Good fit - investment explains most economic growth
Psychology	Test anxiety vs. exam performance	0.1-0.3	Weak fit - many other factors affect performance
Medicine	Cholesterol level vs. heart disease risk	0.2-0.4	Moderate fit - important but not sole factor
Marketing	Ad spending vs. sales	0.4-0.7	Reasonable fit - ads matter but other factors too

D. Task 2: Manual Calculation of Regression Parameters**EXAMPLE APPLICATION**

Example 1: Hours Studied vs Exam Score

MANUAL SIMPLE LINEAR REGRESSION CALCULATION

Sample size: n = 10

X values: [1. 2. 3. 4. 5. 6. 7. 8. 9. 10.]

Y values: [55. 60. 65. 70. 75. 80. 85. 90. 95. 100.]

D.1: CALCULATE INTERMEDIATE VALUES

1. Calculate means:

$$\bar{X} = \bar{X} / n = 55.0 / 10 = 5.5000$$

$$\bar{Y} = \bar{Y} / n = 775.0 / 10 = 77.5000$$

2. Calculate deviations from means:

$$X - \bar{X}: [-4.5 \ -3.5 \ -2.5 \ -1.5 \ -0.5 \ 0.5 \ 1.5 \ 2.5 \ 3.5 \ 4.5]$$

$$Y - \bar{Y}: [-22.5 \ -17.5 \ -12.5 \ -7.5 \ -2.5 \ 2.5 \ 7.5 \ 12.5 \ 17.5 \ 22.5]$$

3. Calculate products $(X - \bar{X})(Y - \bar{Y})$:

$$\text{Point 1: } (-4.5000) \times (-22.5000) = 101.2500$$

$$\text{Point 2: } (-3.5000) \times (-17.5000) = 61.2500$$

$$\text{Point 3: } (-2.5000) \times (-12.5000) = 31.2500$$

$$\text{Point 4: } (-1.5000) \times (-7.5000) = 11.2500$$

$$\text{Point 5: } (-0.5000) \times (-2.5000) = 1.2500$$

$$\text{Point 6: } (0.5000) \times (2.5000) = 1.2500$$

$$\text{Point 7: } (1.5000) \times (7.5000) = 11.2500$$

$$\text{Point 8: } (2.5000) \times (12.5000) = 31.2500$$

$$\text{Point 9: } (3.5000) \times (17.5000) = 61.2500$$

$$\text{Point 10: } (4.5000) \times (22.5000) = 101.2500$$

$$\text{Sum of products (Numerator of): } = 412.5000$$

4. Calculate squared deviations of X $(X - \bar{X})^2$:

$$\text{Point 1: } (-4.5000)^2 = 20.2500$$

$$\text{Point 2: } (-3.5000)^2 = 12.2500$$

$$\text{Point 3: } (-2.5000)^2 = 6.2500$$

$$\text{Point 4: } (-1.5000)^2 = 2.2500$$

$$\text{Point 5: } (-0.5000)^2 = 0.2500$$

$$\text{Point 6: } (0.5000)^2 = 0.2500$$

$$\text{Point 7: } (1.5000)^2 = 2.2500$$

$$\text{Point 8: } (2.5000)^2 = 6.2500$$

$$\text{Point 9: } (3.5000)^2 = 12.2500$$

$$\text{Point 10: } (4.5000)^2 = 20.2500$$

Sum of squared deviations (Denominator of) : = 82.5000

D.2: COMPUTE SLOPE ()

$$\begin{aligned} &= [(X - \bar{X})(Y - \bar{Y})] / [(X - \bar{X})^2] \\ &= 412.5000 / 82.5000 \\ &= 5.000000 \end{aligned}$$

Interpretation: For each 1-unit increase in X, Y is expected to change by 5.0000 units.

D.3: COMPUTE INTERCEPT ()

$$\begin{aligned} &= \bar{Y} - \bar{X} \\ &= 77.5000 - (5.000000 \times 5.5000) \\ &= 77.5000 - 27.500000 \\ &= 50.000000 \end{aligned}$$

Interpretation: When X = 0, the predicted value of Y is 50.0000.

D.4: FORMULATE THE MODEL

Complete estimated regression equation:

$$\begin{aligned} \hat{Y} &= + X \\ \hat{Y} &= 50.000000 + 5.000000X \end{aligned}$$

ADDITIONAL CALCULATIONS

1. Predicted values (\hat{Y}):

Point 1: $\hat{Y} = 50.0000 + 5.0000 \times 1.0 = 55.0000$
 Point 2: $\hat{Y} = 50.0000 + 5.0000 \times 2.0 = 60.0000$
 Point 3: $\hat{Y} = 50.0000 + 5.0000 \times 3.0 = 65.0000$
 Point 4: $\hat{Y} = 50.0000 + 5.0000 \times 4.0 = 70.0000$
 Point 5: $\hat{Y} = 50.0000 + 5.0000 \times 5.0 = 75.0000$
 Point 6: $\hat{Y} = 50.0000 + 5.0000 \times 6.0 = 80.0000$
 Point 7: $\hat{Y} = 50.0000 + 5.0000 \times 7.0 = 85.0000$
 Point 8: $\hat{Y} = 50.0000 + 5.0000 \times 8.0 = 90.0000$
 Point 9: $\hat{Y} = 50.0000 + 5.0000 \times 9.0 = 95.0000$
 Point 10: $\hat{Y} = 50.0000 + 5.0000 \times 10.0 = 100.0000$

2. Residuals ($Y - \hat{Y}$):

Point 1: $55.0000 - 55.0000 = 0.0000$
 Point 2: $60.0000 - 60.0000 = 0.0000$
 Point 3: $65.0000 - 65.0000 = 0.0000$
 Point 4: $70.0000 - 70.0000 = 0.0000$
 Point 5: $75.0000 - 75.0000 = 0.0000$
 Point 6: $80.0000 - 80.0000 = 0.0000$
 Point 7: $85.0000 - 85.0000 = 0.0000$
 Point 8: $90.0000 - 90.0000 = 0.0000$
 Point 9: $95.0000 - 95.0000 = 0.0000$
 Point 10: $100.0000 - 100.0000 = 0.0000$

3. Sum of Squared Residuals (SSR):

$$SSR = (Y - \hat{Y})^2 = 0.000000$$

4. Total Sum of Squares (SST):

$$SST = (Y - \bar{Y})^2 = 2062.500000$$

5. R-squared (Coefficient of Determination):

$$R^2 = 1 - (SSR / SST) = 1 - (0.000000 / 2062.500000) = 1.000000$$

Interpretation: 100.00% of variance in Y is explained by X

6. Correlation coefficient (r):

$$r = (X - \bar{X})(Y - \bar{Y}) / [(X - \bar{X})^2 \times (Y - \bar{Y})^2]$$

$$r = 412.500000 / (82.500000 \times 2062.500000)$$

$$r = 412.500000 / (9.082951 \times 45.414755)$$

$$r = 1.000000$$

$$\text{Check: } r^2 = (1.000000)^2 = 1.000000 = R^2$$

VERIFICATION

1. Check if regression line passes through (X, \bar{Y}) :

$$\text{At } X = 5.5000: \hat{Y} = 50.000000 + 5.000000 \times 5.5000 = 77.500000$$

$$\text{This equals } \bar{Y} = 77.5000$$

2. Check if sum of residuals equals zero:

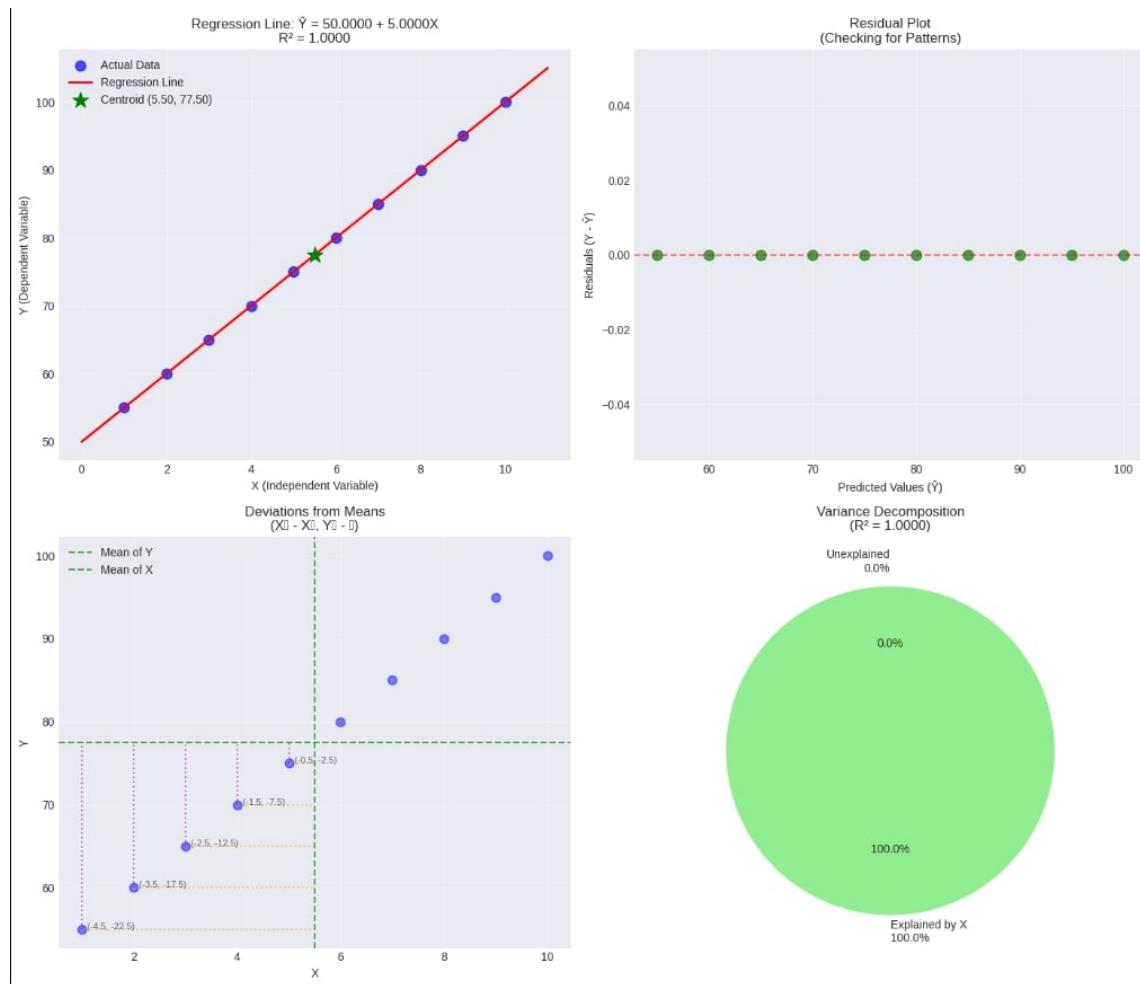
$$(Y - \hat{Y}) = 0.0000000000 \quad 0$$

3. Check if residuals are uncorrelated with X :

$$(X - \bar{X})(Y - \hat{Y}) = 0.0000000000 \quad 0$$

Prediction for $X = 3.5$ hours:

$$\hat{Y} = 50.0000 + 5.0000 \times 3.5 = 67.50$$



Example 2: Realistic Data with Noise

X (Age): [10 12 15 18 20 22 25 28 30 32]
 Y (Income in \$1000s): [77.48 79.31 90.74 102.62 98.83 103.83
 120.4 123.84 122.65 132.71]

MANUAL SIMPLE LINEAR REGRESSION CALCULATION

Sample size: $n = 10$
 X values: [10. 12. 15. 18. 20. 22. 25. 28. 30. 32.]

Y values: [77.48357077 79.30867849 90.73844269 102.61514928 98.829233
103.82931522 120.39606408 123.83717365 122.65262807 132.71280022]

D.1: CALCULATE INTERMEDIATE VALUES

1. Calculate means:

$$\bar{X} = \bar{X} / n = 212.0 / 10 = 21.2000$$

$$\bar{Y} = \bar{Y} / n = 1052.4030555849379 / 10 = 105.2403$$

2. Calculate deviations from means:

$$X - \bar{X}: [-11.2 -9.2 -6.2 -3.2 -1.2 0.8 3.8 6.8 8.8 10.8]$$

$$Y - \bar{Y}: [-27.7567 -25.9316 -14.5019 -2.6252 -6.4111 -1.411 15.1558 18.5969 17.4123 27.4725]$$

3. Calculate products $(X - \bar{X})(Y - \bar{Y})$:

$$\text{Point 1: } (-11.2000) \times (-27.7567) = 310.8754$$

$$\text{Point 2: } (-9.2000) \times (-25.9316) = 238.5710$$

$$\text{Point 3: } (-6.2000) \times (-14.5019) = 89.9115$$

$$\text{Point 4: } (-3.2000) \times (-2.6252) = 8.4005$$

$$\text{Point 5: } (-1.2000) \times (-6.4111) = 7.6933$$

$$\text{Point 6: } (0.8000) \times (-1.4110) = -1.1288$$

$$\text{Point 7: } (3.8000) \times (15.1558) = 57.5919$$

$$\text{Point 8: } (6.8000) \times (18.5969) = 126.4587$$

$$\text{Point 9: } (8.8000) \times (17.4123) = 153.2284$$

$$\text{Point 10: } (10.8000) \times (27.4725) = 296.7029$$

$$\text{Sum of products (Numerator of)} = 1288.3049$$

4. Calculate squared deviations of X $(X - \bar{X})^2$:

$$\text{Point 1: } (-11.2000)^2 = 125.4400$$

$$\text{Point 2: } (-9.2000)^2 = 84.6400$$

$$\text{Point 3: } (-6.2000)^2 = 38.4400$$

$$\text{Point 4: } (-3.2000)^2 = 10.2400$$

$$\text{Point 5: } (-1.2000)^2 = 1.4400$$

$$\text{Point 6: } (0.8000)^2 = 0.6400$$

Point 7: $(3.8000)^2 = 14.4400$

Point 8: $(6.8000)^2 = 46.2400$

Point 9: $(8.8000)^2 = 77.4400$

Point 10: $(10.8000)^2 = 116.6400$

Sum of squared deviations (Denominator of) : = 515.6000

D.2: COMPUTE SLOPE ()

$$\begin{aligned} &= [(X - \bar{X})(Y - \bar{Y})] / [(X - \bar{X})^2] \\ &= 1288.3049 / 515.6000 \\ &= 2.498652 \end{aligned}$$

Interpretation: For each 1-unit increase in X, Y is expected to change by 2.4987 units.

D.3: COMPUTE INTERCEPT ()

$$\begin{aligned} &= \bar{Y} - \bar{X} \\ &= 105.2403 - (2.498652 \times 21.2000) \\ &= 105.2403 - 52.971420 \\ &= 52.268886 \end{aligned}$$

Interpretation: When X = 0, the predicted value of Y is 52.2689.

D.4: FORMULATE THE MODEL

Complete estimated regression equation:

$$\hat{Y} = + X$$

$$\hat{Y} = 52.268886 + 2.498652X$$

ADDITIONAL CALCULATIONS

1. Predicted values (\hat{Y}):

- Point 1: $\hat{Y} = 52.2689 + 2.4987 \times 10.0 = 77.2554$
 Point 2: $\hat{Y} = 52.2689 + 2.4987 \times 12.0 = 82.2527$
 Point 3: $\hat{Y} = 52.2689 + 2.4987 \times 15.0 = 89.7487$
 Point 4: $\hat{Y} = 52.2689 + 2.4987 \times 18.0 = 97.2446$
 Point 5: $\hat{Y} = 52.2689 + 2.4987 \times 20.0 = 102.2419$
 Point 6: $\hat{Y} = 52.2689 + 2.4987 \times 22.0 = 107.2392$
 Point 7: $\hat{Y} = 52.2689 + 2.4987 \times 25.0 = 114.7352$
 Point 8: $\hat{Y} = 52.2689 + 2.4987 \times 28.0 = 122.2311$
 Point 9: $\hat{Y} = 52.2689 + 2.4987 \times 30.0 = 127.2284$
 Point 10: $\hat{Y} = 52.2689 + 2.4987 \times 32.0 = 132.2257$

2. Residuals ($Y - \hat{Y}$):

- Point 1: $77.4836 - 77.2554 = 0.2282$
 Point 2: $79.3087 - 82.2527 = -2.9440$
 Point 3: $90.7384 - 89.7487 = 0.9898$
 Point 4: $102.6151 - 97.2446 = 5.3705$
 Point 5: $98.8292 - 102.2419 = -3.4127$
 Point 6: $103.8293 - 107.2392 = -3.4099$
 Point 7: $120.3961 - 114.7352 = 5.6609$
 Point 8: $123.8372 - 122.2311 = 1.6060$
 Point 9: $122.6526 - 127.2284 = -4.5758$
 Point 10: $132.7128 - 132.2257 = 0.4871$

3. Sum of Squared Residuals (SSR):

$$SSR = (Y - \hat{Y})^2 = 117.615799$$

4. Total Sum of Squares (SST):

$$SST = (Y - \bar{Y})^2 = 3336.641281$$

5. R-squared (Coefficient of Determination):

$$R^2 = 1 - (SSR / SST) = 1 - (117.615799 / 3336.641281) = 0.964750$$

Interpretation: 96.48% of variance in Y is explained by X

6. Correlation coefficient (r):

$$r = (X - \bar{X})(Y - \bar{Y}) / [(X - \bar{X})^2 \times (Y - \bar{Y})^2]$$

$$r = 1288.304909 / (515.600000 \times 3336.641281)$$

$$r = 1288.304909 / (22.706827 \times 57.763667)$$

$$r = 0.982217$$

$$\text{Check: } r^2 = (0.982217)^2 = 0.964750 = R^2$$

VERIFICATION

1. Check if regression line passes through (\bar{X}, \bar{Y}) :

$$\text{At } X = \bar{X} = 21.2000: \hat{Y} = 52.268886 + 2.498652 \times 21.2000 = 105.240306$$

$$\text{This equals } \bar{Y} = 105.2403$$

2. Check if sum of residuals equals zero:

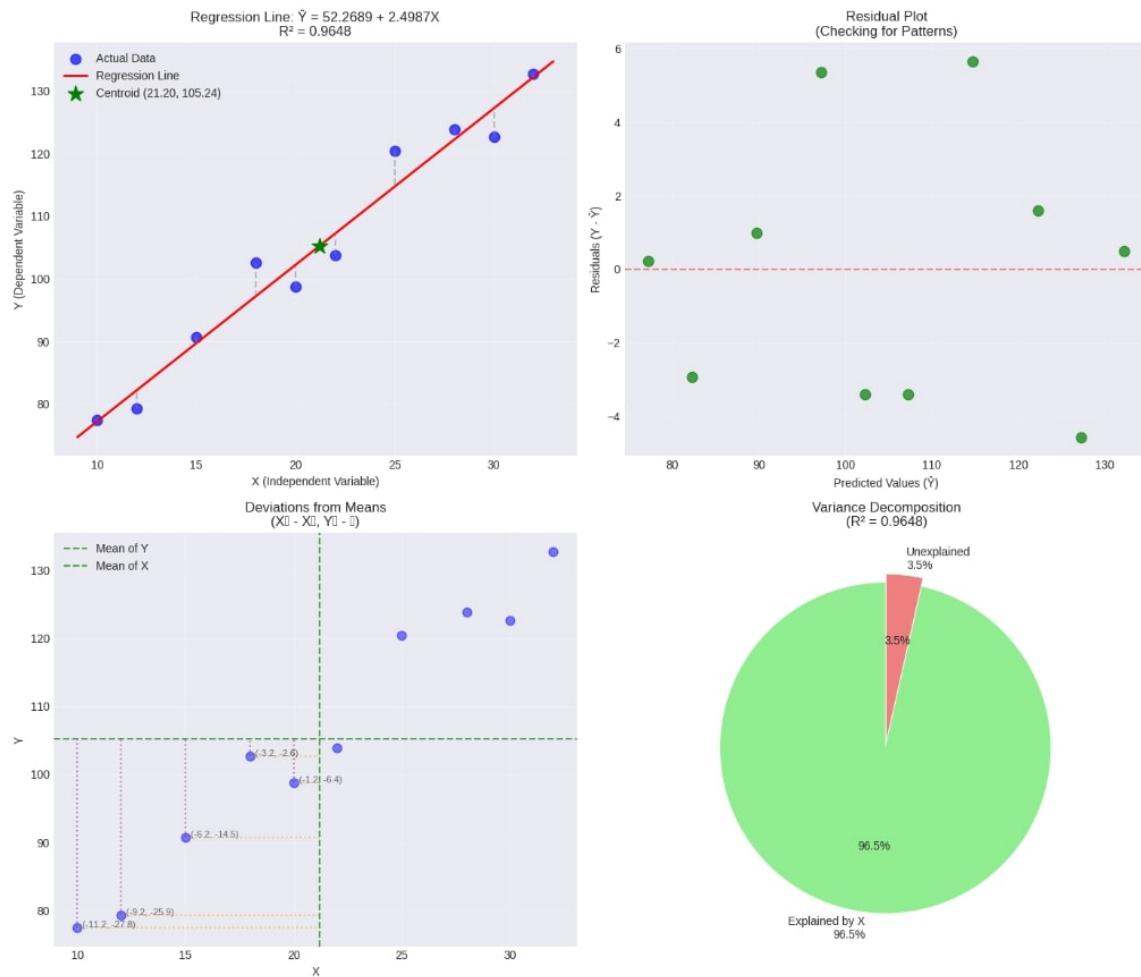
$$(Y - \hat{Y}) = -0.0000000000 \quad 0$$

3. Check if residuals are uncorrelated with X :

$$(X - \bar{X})(Y - \hat{Y}) = 0.0000000000 \quad 0$$

Prediction for $X = 35$ years:

$$\hat{Y} = 52.2689 + 2.4987 \times 35 = \$139.72k$$



Example 3: Weak Relationship (Low R^2)

X: [1 2 3 4 5 6 7 8 9 10]
Y: [20 25 15 30 18 35 22 28 19 32]

MANUAL SIMPLE LINEAR REGRESSION CALCULATION

Sample size: n = 10
X values: [1. 2. 3. 4. 5. 6. 7. 8. 9. 10.]
Y values: [20. 25. 15. 30. 18. 35. 22. 28. 19. 32.]

D.1: CALCULATE INTERMEDIATE VALUES

1. Calculate means:

$$\bar{X} = \bar{X} / n = 55.0 / 10 = 5.5000$$

$$\bar{Y} = \bar{Y} / n = 244.0 / 10 = 24.4000$$

2. Calculate deviations from means:

$$X - \bar{X}: [-4.5 \ -3.5 \ -2.5 \ -1.5 \ -0.5 \ 0.5 \ 1.5 \ 2.5 \ 3.5 \ 4.5]$$

$$Y - \bar{Y}: [-4.4 \ 0.6 \ -9.4 \ 5.6 \ -6.4 \ 10.6 \ -2.4 \ 3.6 \ -5.4 \ 7.6]$$

3. Calculate products $(X - \bar{X})(Y - \bar{Y})$:

$$\text{Point 1: } (-4.5000) \times (-4.4000) = 19.8000$$

$$\text{Point 2: } (-3.5000) \times (0.6000) = -2.1000$$

$$\text{Point 3: } (-2.5000) \times (-9.4000) = 23.5000$$

$$\text{Point 4: } (-1.5000) \times (5.6000) = -8.4000$$

$$\text{Point 5: } (-0.5000) \times (-6.4000) = 3.2000$$

$$\text{Point 6: } (0.5000) \times (10.6000) = 5.3000$$

$$\text{Point 7: } (1.5000) \times (-2.4000) = -3.6000$$

$$\text{Point 8: } (2.5000) \times (3.6000) = 9.0000$$

$$\text{Point 9: } (3.5000) \times (-5.4000) = -18.9000$$

$$\text{Point 10: } (4.5000) \times (7.6000) = 34.2000$$

Sum of products (Numerator of) : = 62.0000

4. Calculate squared deviations of X $(X - \bar{X})^2$:

$$\text{Point 1: } (-4.5000)^2 = 20.2500$$

$$\text{Point 2: } (-3.5000)^2 = 12.2500$$

$$\text{Point 3: } (-2.5000)^2 = 6.2500$$

$$\text{Point 4: } (-1.5000)^2 = 2.2500$$

$$\text{Point 5: } (-0.5000)^2 = 0.2500$$

$$\text{Point 6: } (0.5000)^2 = 0.2500$$

$$\text{Point 7: } (1.5000)^2 = 2.2500$$

$$\text{Point 8: } (2.5000)^2 = 6.2500$$

$$\text{Point 9: } (3.5000)^2 = 12.2500$$

$$\text{Point 10: } (4.5000)^2 = 20.2500$$

Sum of squared deviations (Denominator of) : = 82.5000

D.2: COMPUTE SLOPE ()

$$\begin{aligned}
 &= [(X - \bar{X})(Y - \bar{Y})] / [(X - \bar{X})^2] \\
 &= 62.0000 / 82.5000 \\
 &= 0.751515
 \end{aligned}$$

Interpretation: For each 1-unit increase in X, Y is expected to change by 0.7515 units.

D.3: COMPUTE INTERCEPT ()

$$\begin{aligned}
 &= \bar{Y} - \bar{X} \\
 &= 24.4000 - (0.751515 \times 5.5000) \\
 &= 24.4000 - 4.133333 \\
 &= 20.266667
 \end{aligned}$$

Interpretation: When X = 0, the predicted value of Y is 20.2667.

D.4: FORMULATE THE MODEL

Complete estimated regression equation:

$$\begin{aligned}
 \hat{Y} &= + X \\
 \hat{Y} &= 20.266667 + 0.751515X
 \end{aligned}$$

ADDITIONAL CALCULATIONS

1. Predicted values (\hat{Y}):

Point 1: $\hat{Y} = 20.2667 + 0.7515 \times 1.0 = 21.0182$
 Point 2: $\hat{Y} = 20.2667 + 0.7515 \times 2.0 = 21.7697$
 Point 3: $\hat{Y} = 20.2667 + 0.7515 \times 3.0 = 22.5212$
 Point 4: $\hat{Y} = 20.2667 + 0.7515 \times 4.0 = 23.2727$
 Point 5: $\hat{Y} = 20.2667 + 0.7515 \times 5.0 = 24.0242$
 Point 6: $\hat{Y} = 20.2667 + 0.7515 \times 6.0 = 24.7758$
 Point 7: $\hat{Y} = 20.2667 + 0.7515 \times 7.0 = 25.5273$
 Point 8: $\hat{Y} = 20.2667 + 0.7515 \times 8.0 = 26.2788$
 Point 9: $\hat{Y} = 20.2667 + 0.7515 \times 9.0 = 27.0303$
 Point 10: $\hat{Y} = 20.2667 + 0.7515 \times 10.0 = 27.7818$

2. Residuals ($Y - \hat{Y}$):

Point 1: $20.0000 - 21.0182 = -1.0182$
 Point 2: $25.0000 - 21.7697 = 3.2303$
 Point 3: $15.0000 - 22.5212 = -7.5212$
 Point 4: $30.0000 - 23.2727 = 6.7273$
 Point 5: $18.0000 - 24.0242 = -6.0242$
 Point 6: $35.0000 - 24.7758 = 10.2242$
 Point 7: $22.0000 - 25.5273 = -3.5273$
 Point 8: $28.0000 - 26.2788 = 1.7212$
 Point 9: $19.0000 - 27.0303 = -8.0303$
 Point 10: $32.0000 - 27.7818 = 4.2182$

3. Sum of Squared Residuals (SSR):

$$SSR = (Y - \hat{Y})^2 = 351.806061$$

4. Total Sum of Squares (SST):

$$SST = (Y - \bar{Y})^2 = 398.400000$$

5. R-squared (Coefficient of Determination):

$$R^2 = 1 - (SSR / SST) = 1 - (351.806061 / 398.400000) = 0.116953$$

Interpretation: 11.70% of variance in Y is explained by X

6. Correlation coefficient (r):

$$r = (X - \bar{X})(Y - \bar{Y}) / [(X - \bar{X})^2 \times (Y - \bar{Y})^2]$$

$$r = 62.000000 / (82.500000 \times 398.400000)$$

$$r = 62.000000 / (9.082951 \times 19.959960)$$

$$r = 0.341983$$

$$\text{Check: } r^2 = (0.341983)^2 = 0.116953 = R^2$$

VERIFICATION

1. Check if regression line passes through (X, \bar{Y}) :

$$\text{At } X = 5.5000: \hat{Y} = 20.266667 + 0.751515 \times 5.5000 = 24.400000$$

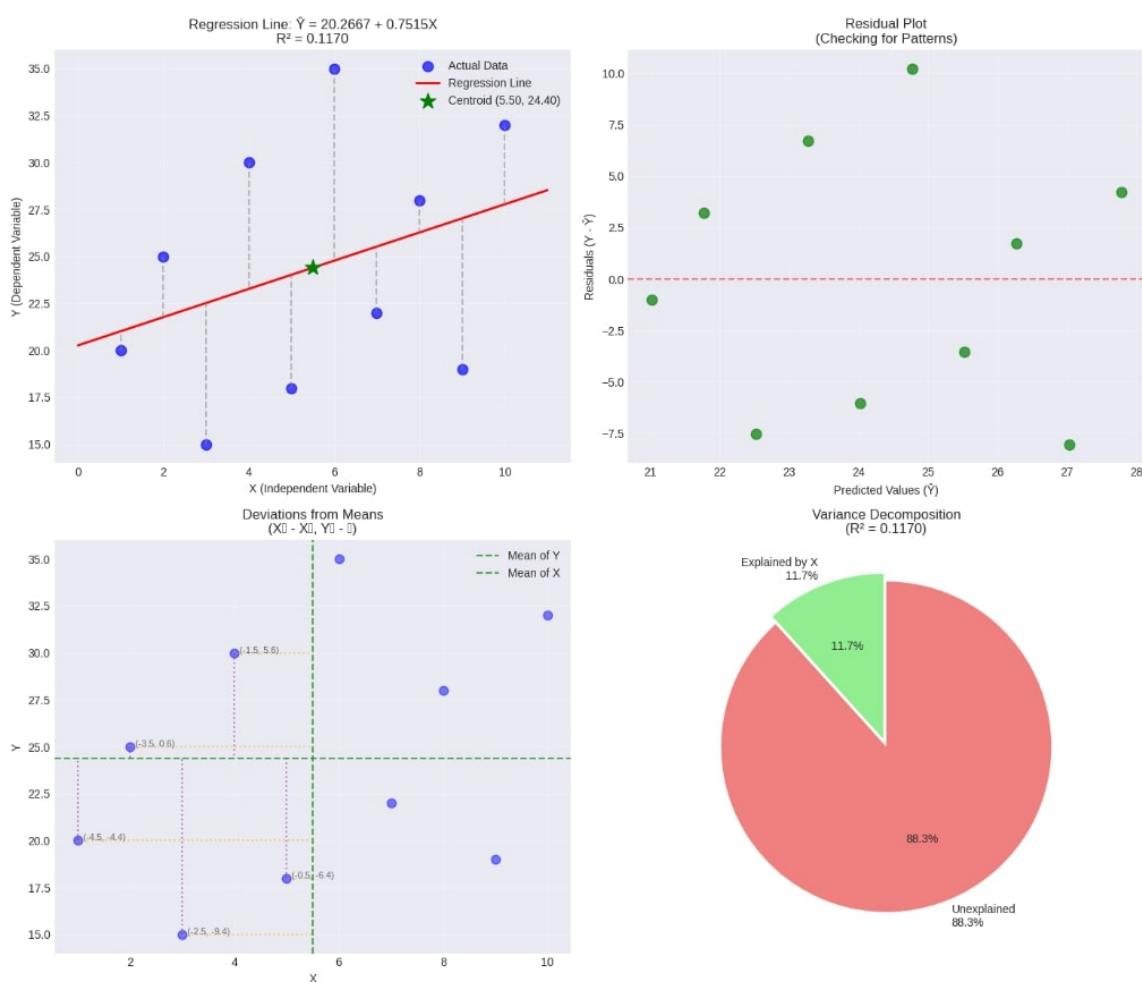
$$\text{This equals } \bar{Y} = 24.4000$$

2. Check if sum of residuals equals zero:

$$(Y - \hat{Y}) = 0.0000000000 \quad 0$$

3. Check if residuals are uncorrelated with X:

$$(X - \bar{X})(Y - \hat{Y}) = -0.0000000000 \quad 0$$



E. Task 3: Visualization of the Fit and Interpretation

REGRESSION RESULTS:

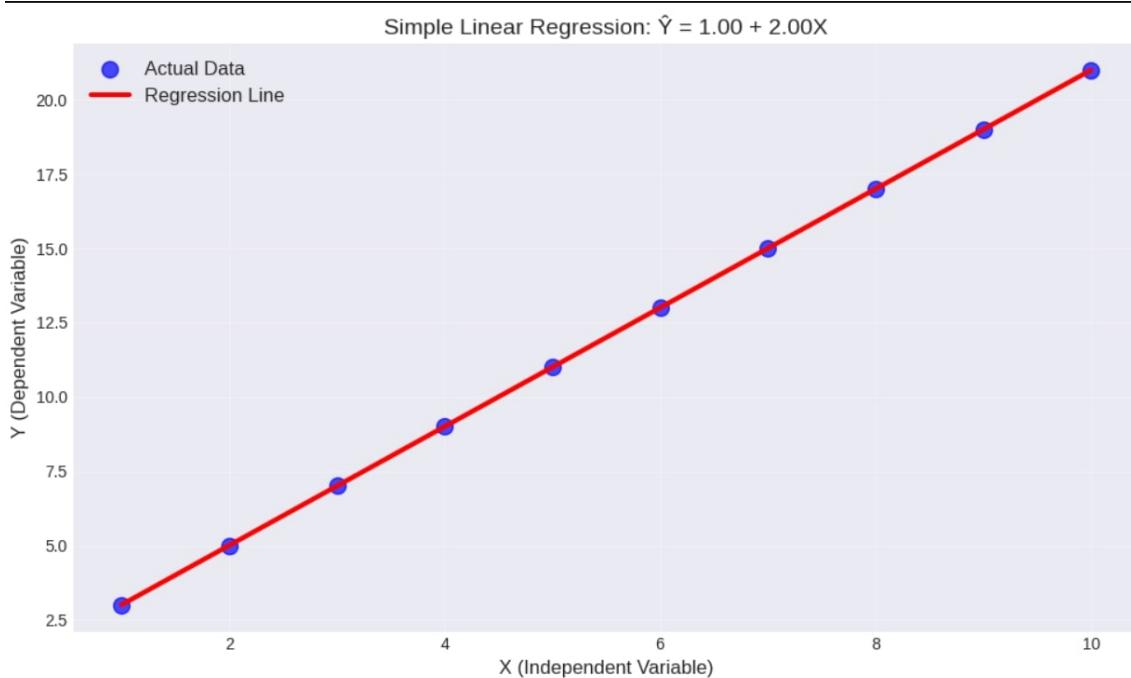
Equation: $\hat{Y} = 1.0000 + 2.0000X$

Interpretation:

- Slope: 2.0000 (Y changes by 2.0000 per 1 unit of X)
- Intercept: 1.0000 (Y when X = 0)

Predictions for your data:

X=1:	Actual Y=3,	Predicted $\hat{Y}=3.00$
X=2:	Actual Y=5,	Predicted $\hat{Y}=5.00$
X=3:	Actual Y=7,	Predicted $\hat{Y}=7.00$
X=4:	Actual Y=9,	Predicted $\hat{Y}=9.00$
X=5:	Actual Y=11,	Predicted $\hat{Y}=11.00$
X=6:	Actual Y=13,	Predicted $\hat{Y}=13.00$
X=7:	Actual Y=15,	Predicted $\hat{Y}=15.00$
X=8:	Actual Y=17,	Predicted $\hat{Y}=17.00$
X=9:	Actual Y=19,	Predicted $\hat{Y}=19.00$
X=10:	Actual Y=21,	Predicted $\hat{Y}=21.00$



F. Task 4: Strength of Relationship

CORRELATION & R-SQUARED ANALYSIS

1. Correlation (r) = 0.9735
 2. R-squared (R^2) = 0.9478
- (Note: $R^2 = r^2 = 0.9735^2 = 0.9478$)

3. RELATIONSHIP ASSESSMENT:

- Direction: POSITIVE (as X increases, Y tends to increase)
- Strength: VERY STRONG ($|r| = 0.9735$)

4. R^2 INTERPRETATION:

- 94.78% of the variation in Y

can be explained by the linear relationship with X.

- 5.22% is due to other factors or random variation.

5. VISUAL CHECK:

$r = +0.9735$ means:

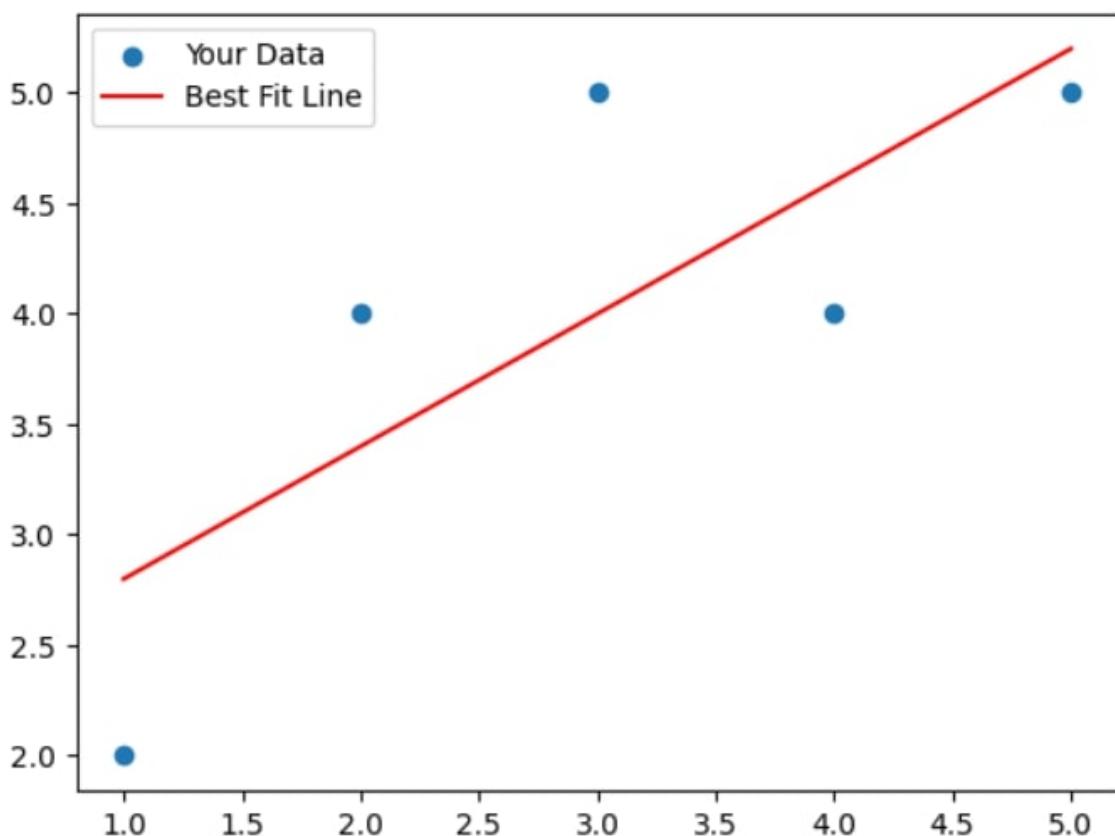
Points cluster closely around an upward-sloping line

Equation: $\hat{Y} = 2.20 + 0.60X$

Correlation: $r = 0.775$

$R^2 = 0.600$ (60.0% explained)

For each 1 unit of X, Y changes by 0.600



G. Task 5: Reflection

1. R² SUPPORT:

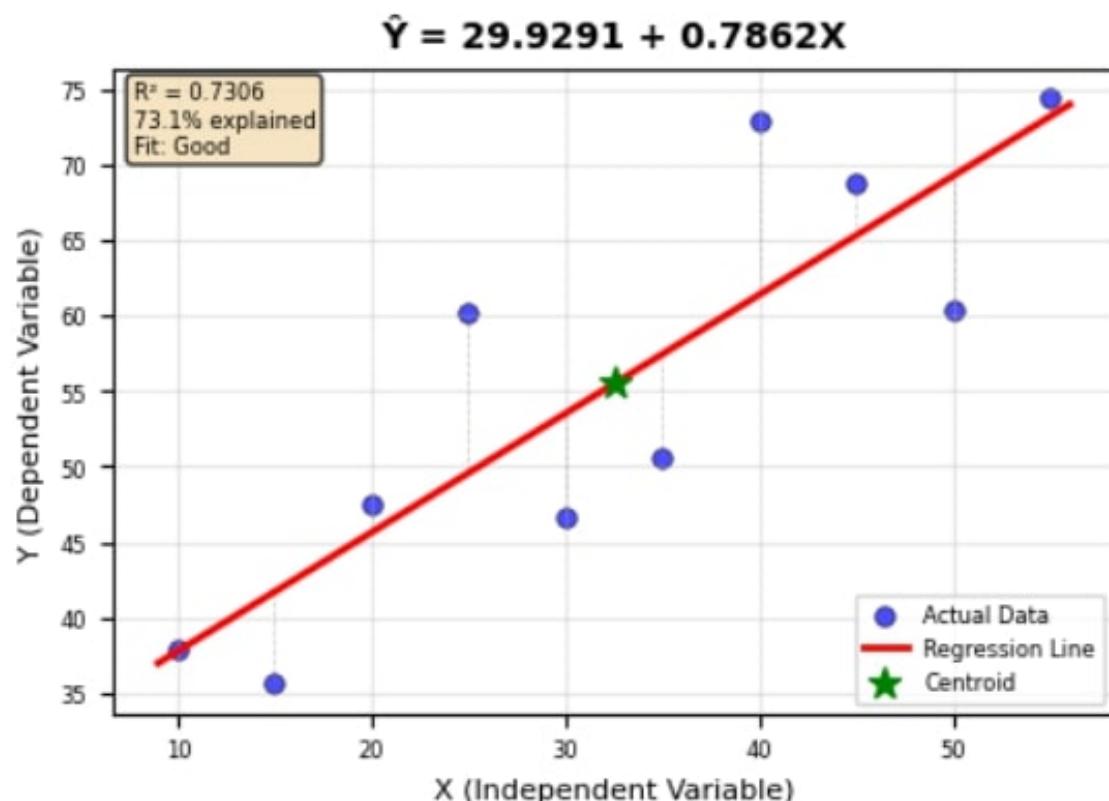
R² = 1.0000 (100.0% explained)

Strongly supports visual assessment

2. REAL-WORLD APPLICATION:

High-confidence predictions (e.g., engineering, finance)

Equation: $\hat{Y} = 0.00 + 2.00X$



Analysis and Conclusion

12.1 Application to Agricultural Data

Regression analysis can be applied to agricultural data to measure the relationship between crop yield and factors such as rainfall, temperature practices.

12.2 Limitations and Precautions

- **Assumption of Linearity:** Linear regression analyzes only linear relationships. If the relationship between variables is non-linear, the model may produce inaccurate results.
- **Sensitivity to Outliers:** Regression models can be significantly affected by outliers, which may distort parameter estimates and reduce model reliability.

13 Final Conclusion

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

Agriculture Crop Yield dataset using various statistical and probability concepts to understand the factors influencing

agricultural productivity. This analysis average crop yield is approximately 3.6 tons per hectare, with a mean rainfall of about 820 mm and an average temperature ranging between 25{26°C. The dataset includes multiple regions, soil types, crops, and weather conditions, making the analysis. Sampling produced the most accurate estimate of mean yield, while simple random sampling was easier to apply but slightly less precise. Cluster sampling showed the highest variability. Data visualization tools such as histograms and bar charts revealed that the South region and cloudy weather are most associated with higher yields. indicated a weak linear relationship between rainfall and temperature.

The crop yields are influenced by multiple factors, with regional and environmental conditions playing significant roles. The South region and cloudy weather showed the highest conditional

probabilities for high yield while fertilizer use increased of high yield.

The analysis began with a clear focus on rice cultivation in Bangladesh. The importance of environmental and agronomic variables such as soil type, rainfall, temperature, and farming practices. Through descriptive statistics, the dataset's central tendencies and variabilities were quantified, revealing an average yield of approximately 3.6 tons/ha and setting the stage for deeper investigation. The application of probability sampling

including simple random, systematic, stratified, and cluster sampling

highlighted the importance of sampling design in estimating

population parameters accurately, with stratified sampling emerging as the most reliable approach for this dataset. Data visualization through histograms, bar charts, ogives, and frequency polygons effectively communicated distribution patterns, regional disparities and cumulative trends. This milestone histograms, bar charts, and ogive charts, the analysis reveals important patterns. Bayes' Theorem to "invert" probabilities. While we can easily find the probability of a high yield given a southern location. The analysis checks if the yield data a Normal Distribution. This is important because many statistical tests assume normality. The project finds that while the data is roughly normal the theoretical "bell curve. The importance of checking assumptions before certain models.

Regression analysis numerical variables like rainfall and yield. That show that while certain factors have a measurable relationship with yield. The linear models alone often explain a limited portion of the total variation

References

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

14 GitHub

GitHub URL:

<https://github.com/Mahin01234/STATISTICS>

15 Sciencedirect Dataset help by :

<https://www.sciencedirect.com/science/article/pii/S2590123025027392>

16 Geeksforgeeks Dataset Help

<https://www.geeksforgeeks.org/data-visualization/visualizati on-and-prediction-of-crop-production-data-using-python/>

17 Geeksforgeeks Dataset Help

<https://www.youtube.com/watch?v=FiPAle5stDI>

18 YouTube

<https://www.youtube.com/watch?v=ehz5dah4fv4>

https://www.youtube.com/watch?v=RTbz-H8fb_k

<https://www.youtube.com/watch?v=nyGb4XaVJZs>

----- END -----