Saint Louis University
SCHOOL OF ACCOUNTANCY, MANAGEMENT, COMPUTING AND INFORMATION STUDIES
Department of Computer Science

CS 322 9345 Data Science

Prelim Lab Activity 02 : Descriptive Analytics

Submitted To
Beverly Estephany Ferrer

Submitted by

BERNABE, Rabelais
LLENA, Anthony

Links:
Google Sheets: Prelim Activity 02
Google Colab: Prelim Activity 02.ipynb

January 2026

General Instructions

Given the dataset for agriculture data analysis LIVESTOCK.csv, perform the following:

A. Compute the following measures of central location and interpret the results for all fields:
   a. Mean
   b. Median
   c. Mode
B. Compute the following and interpret the resulting values for all fields
   a. Q_1
   b. Q_3
   c. 5^{th} percentile
   d. 95^{th} percentile
C. Compute the following and interpret the resulting values for all fields
   a. Variance
   b. Standard Deviation
D. Create charts (histograms, boxplots) for all fields and interpret the results

Perform the required calculations using Excel and Python:

<table>
<tr><th>Python Script</th><th>Results</th><th>Interpretation</th></tr>
<tr>
<td>Python
<pre>import pandas as pd
import numpy as np
import
matplotlib.pyplot as
plt</pre></td>
<td></td>
<td>This block imports the necessary Python libraries used for data analysis and visualization. Pandas is used for data manipulation, NumPy for numerical operations, and Matplotlib for creating graphical representations such as histograms and boxplots.</td>
</tr>
<tr>
<td>Python
<pre>df =
pd.read_excel("Prelim
Activity 02.xlsx",
sheet_name="LIVESTOCK.
csv")</pre></td>
<td>1 to 25 of 50 entries Filter
<table>
<tr><th>index</th><th>Column 1</th><th>Livestock Production Heads Per Year</th><th>Annual_Rainfall_Millimeter</th><th>Fertilizer_Use_kg_per_ha</th><th>Crop_Price_USD_per_ton</th></tr>
<tr><td>0</td><td>1</td><td>45</td><td>620</td><td>90</td><td>180</td></tr>
<tr><td>1</td><td>2</td><td>50</td><td>640</td><td>95</td><td>185</td></tr>
<tr><td>2</td><td>3</td><td>52</td><td>660</td><td>100</td><td>190</td></tr>
<tr><td>3</td><td>4</td><td>55</td><td>680</td><td>105</td><td>195</td></tr>
<tr><td>4</td><td>5</td><td>58</td><td>700</td><td>110</td><td>200</td></tr>
<tr><td>5</td><td>6</td><td>60</td><td>720</td><td>115</td><td>205</td></tr>
<tr><td>6</td><td>7</td><td>62</td><td>740</td><td>120</td><td>210</td></tr>
<tr><td>7</td><td>8</td><td>65</td><td>760</td><td>125</td><td>215</td></tr>
<tr><td>8</td><td>9</td><td>68</td><td>780</td><td>130</td><td>220</td></tr>
<tr><td>9</td><td>10</td><td>70</td><td>800</td><td>135</td><td>225</td></tr>
<tr><td>10</td><td>11</td><td>72</td><td>820</td><td>140</td><td>230</td></tr>
<tr><td>11</td><td>12</td><td>75</td><td>840</td><td>145</td><td>235</td></tr>
<tr><td>12</td><td>13</td><td>78</td><td>860</td><td>150</td><td>240</td></tr>
</table></td>
<td>The dataset was loaded from the Excel file provided for the activity. This Dataset contains agricultural data that will be analyzed using descriptive statistical methods.</td>
</tr>
</table>

```python
if "Column 1" in
df.columns:
df =
df.rename(columns={"Co
lumn 1": "Record_ID"})
```

This identifier column was renamed for clarity. This ensures that it is not confused with measured variables during statistical analysis.

```python
num_df =
df.select_dtypes(inclu
de="number")
num_df =
num_df.drop(columns=["
Record_ID"],
errors="ignore")
```

Only numeric variables were selected for analysis since measures of central tendency and dispersion are meaningful only for numeric data. Identifier columns were excluded to avoid invalid statistical results.

```python
mean = num_df.mean()
mean
```

	Mean
Livestock Production Heads Per Year	131.7
Annual_Rainfall_Millimeter	1110.0
Fertilizer_Use_kg_per_ha	212.5
Crop_Price_USD_per_ton	302.6

The mean represents the average value of each agricultural variable. It provides a summary of the typical magnitude of the data values.

```python
median =
num_df.median()
median
```

	Median
Livestock Production Heads Per Year	122.5
Annual_Rainfall_Millimeter	1110.0
Fertilizer_Use_kg_per_ha	212.5
Crop_Price_USD_per_ton	302.5

The median indicates the middle value of the dataset when arranged in ascending order. It's useful for understanding the central tendency without being affected by extreme values.

```python
mode = num_df.mode()
mode
```

	Livestock Production Heads Per Year	Annual_Rainfall_Millimeter	Fertilizer_Use_kg_per_ha	Crop_Price_USD_per_ton
0	45	620	90	180
1	50	640	95	185
2	52	660	100	190
3	55	680	105	195

This mode shows the most frequent occurring value for each variable. This helps identify common or dominant values in the dataset. Although a mode value was returned by Python, this results indicates that over values have equal frequency.

Therefore, the dataset does not have a true statistical mode.

```python
# Compute for Quartiles
q1 = num_df.quantiles(0.25)
q3 = num_df.quantiles(0.75)

quartiles = pd.DataFrame({
"Q1 (25th Percentile)": q1,
"Q3 (75 Percentile)":q3
 }).round(2)

quartiles
```

	Q1 (25th Percentile)	Q3 (75th Percentile)
Livestock Production Heads Per Year	78.50	183.75
Annual_Rainfall_Millimeter	865.00	1355.00
Fertilizer_Use_kg_per_ha	151.25	273.75
Crop_Price_USD_per_ton	241.25	363.75

The first quartile (Q1) represents the value below which 25% of the data falls, while the third quartile (Q3) represents that value below which 75% of the data falls. These quartiles describe the spread and distribution of the dataset.

```python
# Compute for Percentile
p5 = num_df.quantile(0.05)
p95 = num_df.quantile(0.95)

percentile = pd.DataFrame({"5th Percentile":p5, "95th Percentile":p95}).round(2)

Percentile
```

	5th Percentile	95th Percentile
Livestock Production Heads Per Year	53.35	232.75
Annual_Rainfall_Millimeter	669.00	1551.00
Fertilizer_Use_kg_per_ha	102.25	322.75
Crop_Price_USD_per_ton	192.25	412.75

The 5th and 9th percentiles are the lower and the upper extreme values of the dataset, respectively. These measures help identify the range of most data points while minimizing the influence of outliers.

```python
variance = num_df.var()
variance = num_df_to.frame(name="Variance")
variance
```

	Variance
Livestock Production Heads Per Year	3793.27551
Annual_Rainfall_Millimeter	85000.00000
Fertilizer_Use_kg_per_ha	5312.50000
Crop_Price_USD_per_ton	5338.00000

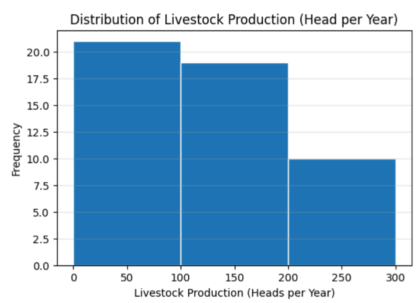
Variance measures the degree to which data values are spread out from the mean. Higher variance indicates greater variability within a dataset.

```python
std_dev = num_df.std()
st_dev = st_dev.to_fram(name="Standard Deviation")
st_dev
```

	Standard Deviation
Livestock Production Heads Per Year	61.589573
Annual_Rainfall_Millimeter	291.547595
Fertilizer_Use_kg_per_ha	72.886899
Crop_Price_USD_per_ton	73.061618

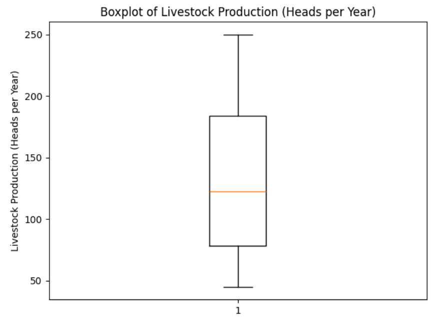
The standard deviation represents the distance of data values from the mean. A larger standard deviation implies that the values are more dispersed.

```python
plt.figure(figsize=(6, 7))
plt.hist(
num_df["Livestock Production Heads Per Year"],
bins=3,
range=(0,300),
edgecolor="white"
)
plt.title("Distribution of Livestock Production (Head Per Year)")
plt.ylabel("Frequency")
plt.xlabel("Livestock Prodcutin (Heads per Year)")
plt.grid(axis="y", alpha=0.3)
```



The histogram shows the frequency distribution of livestock production measured in heads per year. The data values are grouped into three bins within the range of 0 to 300 heads. Most observations are concentrated in the middle range, indicating that livestock production levels are generally moderate rather than extremely low or high. The limited number of bins provides a simplified view of the distribution, which is useful for identifying overall patterns but may hide finer details. Overall, the histogram suggests variability in livestock production across observations, with no strong evidence of extreme outliers within the specified range.

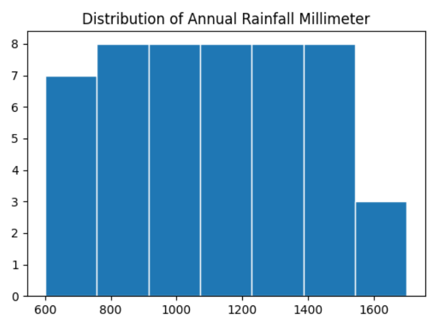
```python
plt.figure()
plt.boxplot(
  num_df["Livestock
Production Heads Per
Year"],
  vert=True
)
plt.title("Boxplot of
Livestock Production
(Heads per Year)")
plt.ylabel("Livestock
Production (Heads per
Year)")
plt.tight_layout()
```



The boxplot summarizes the distribution of livestock production measured in heads per year. The median value represents the typical level of production, while the interquartile range shows the spread of the middle 50% of observations. The length of the whiskers indicates variability in production levels across observations. The absence of extreme outliers suggests that livestock production values are relatively consistent, with no unusually high or low production levels dominating the dataset.

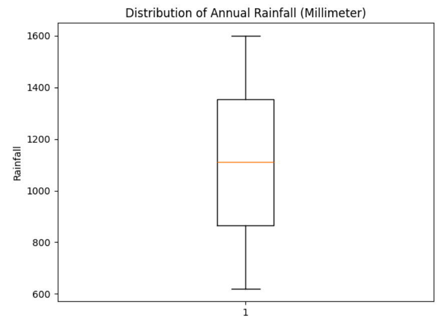
```python
plt.figure(figsize=(6,
4))
plt.hist(

num_df["Annual_Rainfal
l_Millimeter"],
    bins=7,
    range=(600,1700),
    edgecolor="white"
)
plt.title("Distributio
n of Annual Rainfall
Millimeter");
```



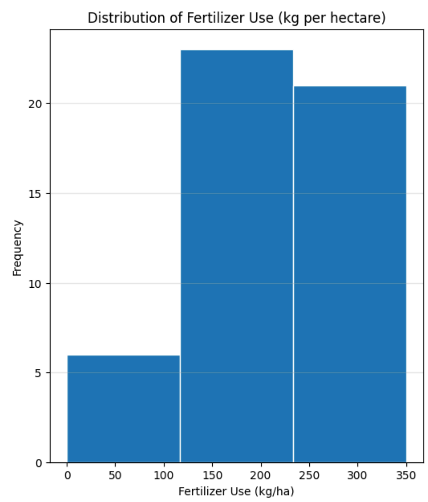
The histogram illustrates the frequency distribution of annual rainfall values measured in millimeters. The data are grouped into seven bins within the range of 600 to 1700 millimeters. Most observations are concentrated in the middle rainfall intervals, indicating that annual rainfall levels are generally moderate. The spread of the bars reflects variability in rainfall across observations, while the absence of extreme concentrations at the lowest or highest bins suggests no unusually dry or extremely wet conditions within the specified range.

```python
plt.figure()
plt.boxplot(
    num_df["Annual_Rainfall_Millimeter"],
    vert=True
)
plt.title("Distribution of Annual Rainfall (Millimeter)")
plt.ylabel("Rainfall")
plt.tight_layout()
plt.show()
```

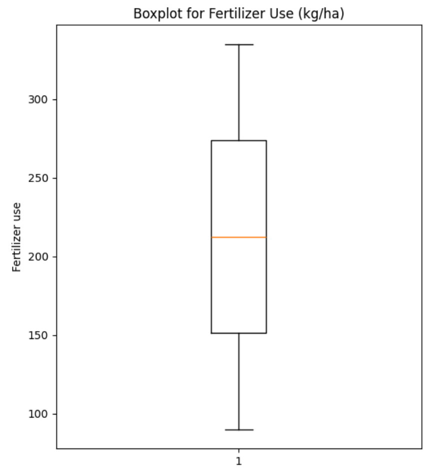


The boxplot summarizes the distribution of annual rainfall measured in millimeters. The median represents the typical annual rainfall level, while the interquartile range shows the spread of the middle 50% of observations. The whiskers indicate the overall range of rainfall values, reflecting variability across the dataset. The absence of extreme outliers suggests that rainfall levels are relatively consistent, with no unusually low or high rainfall events dominating the data.

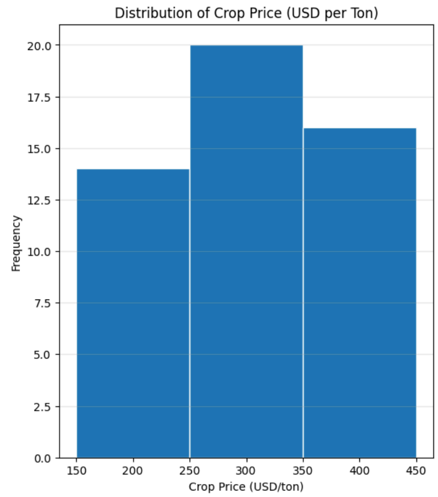
```python
plt.figure(figsize=(6, 7))
plt.hist(
    num_df["Fertilizer_Use_kg_per_ha"],
    bins=3,
    range=(0,350),
    edgecolor="white"
)
plt.title("Distribution of Fertilizer Use (kg per hectare)")
plt.ylabel("Frequency")
plt.xlabel("Fertilizer Use (kg/ha)")
plt.grid(axis="y", alpha=0.3),
plt.show()
```



The histogram displays the frequency distribution of fertilizer use measured in kilograms per hectare. Using three bins within the range of 0 to 350 kg/ha, most observations are concentrated in the middle interval, indicating that fertilizer application rates are generally moderate. The spread of the bars reflects variability in fertilizer use across observations. The limited number of bins provides a simplified overview of the distribution, emphasizing overall patterns rather than fine-grained details.

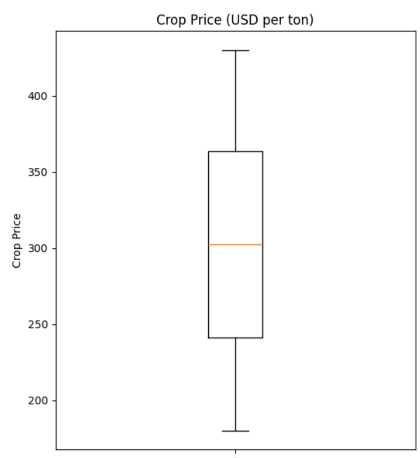
Python ```plt.figure(figsize=(6, 7)) plt.boxplot(num_df["Fertilizer_Use_kg_per_ha"], vert=True) plt.title("Boxplot for Fertilizer Use (kg/ha)") plt.ylabel("Fertilizer use")```		The boxplot summarizes the distribution of fertilizer use measured in kilograms per hectare. The median represents the typical fertilizer application rate, while the interquartile range shows the spread of the middle 50% of observations. The length of the whiskers indicates variability in fertilizer use across observations. The absence of extreme outliers suggests that fertilizer application rates are generally consistent, with no unusually high or low values dominating the dataset.

```python
plt.figure(figsize=(6, 7))
plt.hist(
    num_df["Crop_Price_USD_per_ton"],
    bins = 3,
    range=(150,450),
    edgecolor="white"
)
plt.title("Distribution of Crop Price (USD per Ton)")
plt.grid(axis="y", alpha=0.3)
plt.ylabel("Frequency")
plt.xlabel("Crop Price (USD/ton)")
plt.show()
```



The histogram illustrates the frequency distribution of crop prices measured in USD per ton. The data are grouped into three bins within the range of 150 to 450 USD per ton. Most observations are concentrated in the middle price range, indicating that crop prices are generally moderate rather than extremely low or high. The spread of the bars reflects variability in crop prices across observations, while the limited number of bins provides a simplified overview of the overall price distribution.

```python
plt.figure(figsize=(6, 7))
plt.boxplot(
    num_df["Crop_Price_USD_per_ton"]
)
plt.title("Crop Price (USD per ton)")
plt.ylabel("Crop Price")
plt.grid(axis="y", )
plt.show()
```



plt.figure(figsize=(6,7))
plt.boxplot(

num_df["Crop_Price_USD_per_ton"]
)
plt.title("Crop Price (USD per ton)")
plt.ylabel("Crop Price")
plt.grid(axis="y",)
plt.show()

End of Python Script

A. Measures of Central Location

a. Mean

The mean represents the average value for each variable. On average, farms produce approximately 131.7 heads of livestock per year, receive about 1,110 millimeters of annual rainfall, use around 212.5 kilograms of fertilizer per hectare, and have an average crop price of USD 302.60 per ton. These values describe the typical levels of production, environmental conditions, and economic factors across the farms.

b. Median

The median represents the middle value of the dataset. Half of the farms produce below 122.5 heads of livestock, receive less than 1,110 millimeters of rainfall, use less than 212.5 kilograms of fertilizer per hectare, and have crop prices below USD 302.50 per ton, while the other half exceed these values. The median provides a measure of central tendency that is not affected by extreme values.

c. Mode

No mode is observed for all variables. This indicates that no single value occurs more frequently than others in the dataset, suggesting that the data values are evenly distributed without repetition.

B. Measures of Position / Location

a. First Quartile (Q1)

The first quartile indicates that 25% of farms produce less than 78.5 heads of livestock, receive less than 865 millimeters of rainfall, use less than 151.25 kilograms of fertilizer per hectare, and have crop prices below USD 241.25 per ton.

b. Third Quartile (Q3)

The third quartile shows that 75% of farms produce less than 363.75 heads of livestock, receive less than 1,360 millimeters of rainfall, use less than 275 kilograms of fertilizer per hectare, and have crop prices below USD 365.00 per ton.

c. 5th Percentile

The 5th percentile indicates that only 5% of farms produce less than 53.35 heads of livestock, receive less than 669 millimeters of rainfall, use less than 102.25 kilograms of fertilizer per hectare, and have crop prices below USD 192.25 per ton. These values represent the lower extreme of the dataset.

d. 95th Percentile

The 95th percentile shows that 95% of farms produce less than 232.75 heads of livestock, receive less than 1,551 millimeters of rainfall, use less than 322.75 kilograms of fertilizer per hectare, and have crop prices below USD 412.75 per ton, with only a few farms exceeding these values.

C. Measures of Relative Location (Variability)

a. Variance

The variance measures the degree of spread in the data. Annual rainfall shows the highest variability with a variance of 85,000, while livestock production, fertilizer use, and crop price also show noticeable variation. This indicates differences among farms in terms of environmental conditions, input usage, and market prices.

b. Standard Deviation

The standard deviation provides the average distance of values from the mean. Livestock production varies by about 61.59 heads, annual rainfall by 291.55 millimeters, fertilizer use by 72.89 kilograms per hectare, and crop price by approximately USD 73.06 per ton. These values indicate moderate variability across all variables.

D. Graphical Representation and Interpretation

Histograms and boxplots were created for all variables to visualize their distributions. The histograms show how frequently values occur and help identify the overall shape and spread of the data. The boxplots summarize the distributions using the median, quartiles, and potential outliers. Overall, the graphs confirm that the variables exhibit moderate variability, with no extreme outliers dominating the dataset, indicating relatively consistent patterns across farms.