**Topics: Descriptive Statistics and Probability**

1. Look at the data given below. Plot the data, find the outliers and find out

|  |  |
| --- | --- |
| **Name of company** | **Measure X** |
| Allied Signal | 24.23% |
| Bankers Trust | 25.53% |
| General Mills | 25.41% |
| ITT Industries | 24.14% |
| J.P.Morgan & Co. | 29.62% |
| Lehman Brothers | 28.25% |
| Marriott | 25.81% |
| MCI | 24.39% |
| Merrill Lynch | 40.26% |
| Microsoft | 32.95% |
| Morgan Stanley | 91.36% |
| Sun Microsystems | 25.99% |
| Travelers | 39.42% |
| US Airways | 26.71% |
| Warner-Lambert | 35.00% |

Answers:

import matplotlib.pyplot as plt

import pandas as pd

# Given data (percentage values)

data = {

'Name of company': [

'Allied Signal', 'Bankers Trust', 'General Mills', 'ITT Industries',

'J.P.Morgan & Co.', 'Lehman Brothers', 'Marriott', 'MCI',

'Merrill Lynch', 'Microsoft', 'Morgan Stanley', 'Sun Microsystems',

'Travelers', 'US Airways', 'Warner-Lambert'

],

'Measure X': [

'24.23%', '25.53%', '25.41%', '24.14%', '29.62%', '28.25%', '25.81%',

'24.39%', '40.26%', '32.95%', '91.36%', '25.99%', '39.42%', '26.71%', '35.00%'

]

}

# Convert percentage values to float numbers

data['Measure X'] = [float(value.strip('%')) for value in data['Measure X']]

# Create a DataFrame

df = pd.DataFrame(data)

# Step 3: Plot the data

plt.figure(figsize=(8, 5))

plt.boxplot(df['Measure X'])

plt.title('Boxplot of Measure X')

plt.ylabel('Percentage')

plt.show()

# Step 4: Identify outliers using IQR method

Q1 = df['Measure X'].quantile(0.25)

Q3 = df['Measure X'].quantile(0.75)

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

outliers = df[(df['Measure X'] < lower\_bound) | (df['Measure X'] > upper\_bound)]

print("Outliers:")

print(outliers)

# Step 5: Calculate mean, standard deviation, and variance

mean = df['Measure X'].mean()

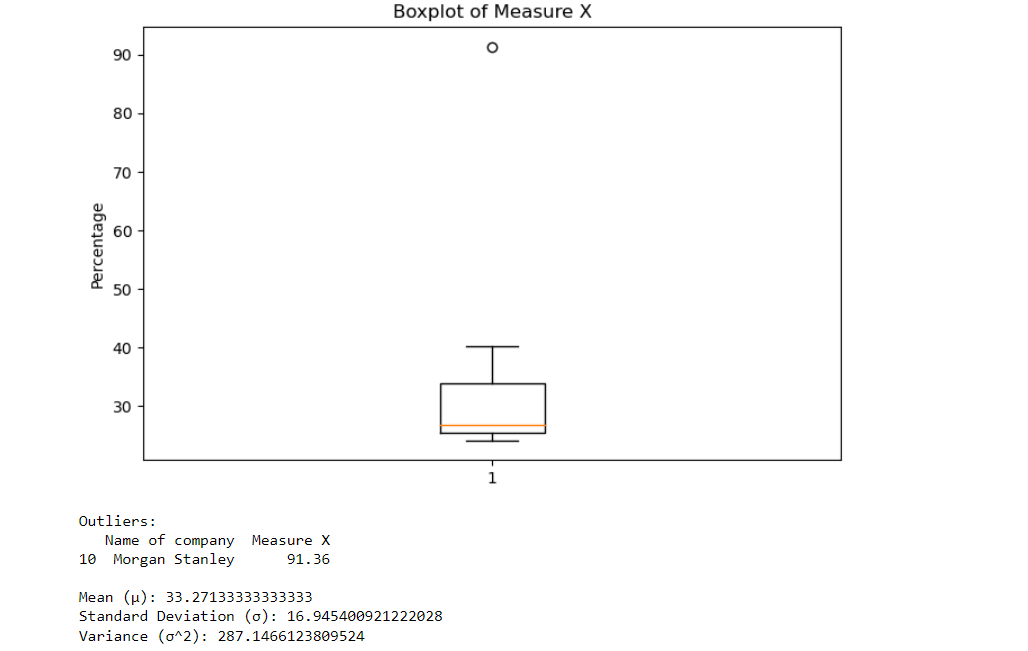
std\_dev = df['Measure X'].std()

variance = df['Measure X'].var()

print("\nMean (μ):", mean)

print("Standard Deviation (σ):", std\_dev)

print("Variance (σ^2):", variance)



Please find below output :

Outlier from the graph is

|  |  |
| --- | --- |
| Morgan Stanley | 91.36% |

Mean: 33:27

Standard Deviation: 16.94

Variance:287.14



Answer the following three questions based on the box-plot above.

1. What is inter-quartile range of this dataset? (please approximate the numbers) In one line, explain what this value implies.

Answer:

IQR = Q3-Q1

= 12-5

= 7

1. What can we say about the skewness of this dataset?

Answer: Right Skewed Data

1. If it was found that the data point with the value 25 is actually 2.5, how would the new box-plot be affected?

Answer: Q1 value will start from 2.5 rather than 5, and also there will be changes in IQR and also no outlier in the minimum of 19



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

Answer: Mode of the data set lie between 4 to 8

1. Comment on the skewness of the dataset.

Answer: Right Skewed Histogram

1. Suppose that the above histogram and the box-plot in question 2 are plotted for the same dataset. Explain how these graphs complement each other in providing information about any dataset.

Answer: Histograms and box plots are very similar in that they both help to visualize and describe numeric data. Although histograms are better in determining the underlying distribution of the data, box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space. And if we plot above data set in Box plot it there will be one Outlier. Histogram is negative curved graph.

1. AT&T was running commercials in 1990 aimed at luring back customers who had switched to one of the other long-distance phone service providers. One such commercial shows a businessman trying to reach Phoenix and mistakenly getting Fiji, where a half-naked native on a beach responds incomprehensibly in Polynesian. When asked about this advertisement, AT&T admitted that the portrayed incident did not actually take place but added that this was an enactment of something that “could happen.” Suppose that one in 200 long-distance telephone calls is misdirected. What is the probability that at least one in five attempted telephone calls reaches the wrong number? (Assume independence of attempts.)

Answer:

We can use binomial distribution to calculate the probability of least one in five attempted telephone calls reaches the wrong number.

Probability of success (misdirected calls) = p = 1/200

Number of rails is n = 5

The probability that a single call is correctly directed is (1 - p) = (199/200).

The probability that all five calls are correctly directed is (199/200)^5 = (0.995)^5 = 0.9752

The probability that at least one call is misdirected is 1 – 0.9752 =0.0248 \* 100 = 2.48%

The probability that at least one in five attempted telephone calls reaches the wrong number is about 2.48%.

1. Returns on a certain business venture, to the nearest $1,000, are known to follow the following probability distribution

|  |  |
| --- | --- |
| x | P(x) |
| -2,000 | 0.1 |
| -1,000 | 0.1 |
| 0 | 0.2 |
| 1000 | 0.2 |
| 2000 | 0.3 |
| 3000 | 0.1 |

1. What is the most likely monetary outcome of the business venture?

Answer: The most likely monetary outcome of the business venture is the value with the highest probability mass function , which is 2000 dollars since it has a PMF of 0.3.

1. Is the venture likely to be successful? Explain

Answer: The venture is likely to be successful because there is a high probability of 0.6 for earning a positive return. That is the outcome of $1000,$2000,$3000 have positive probability (0.2, 0.3, 0.1) respectively.

1. What is the long-term average earning of business ventures of this kind? Explain

Answer:

The long-term average earning of business ventures of this kind can be calculated as:

E(X) = (-2000)(0.1) + (-1000)(0.1) + (0)(0.2) + (1000)(0.2) + (2000)(0.3) + (3000)(0.1) = 800

Therefore, the long-term average earning of business ventures is 800 dollars.

1. What is the good measure of the risk involved in a venture of this kind? Compute this measure

Answer:

A good measure of the risk involved in a venture of this kind is the standard deviation, which can be computed as:

σ = sqrt[(-2000 - 800)^2\*(0.1) + (-1000 - 800)^2\*(0.1) + (0 - 800)^2\*(0.2) + (1000 - 800)^2\*(0.2) + (2000 - 800)^2\*(0.3) + (3000 - 800)^2\*(0.1)] = 1854.15

Therefore, the standard deviation of the probability distribution is 1854.15 dollars, which means high level of risk.