**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% |

**CODE FOR THE GIVEN QUESTION**

**# Data given**

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%']}

**# Create a Data Frame from the data**

Import pandas as pd

df = pd.DataFrame(data)

**# Convert 'Measure X' to numeric values (remove the % sign and convert to float data type)**

df['Measure X'] = df['Measure X'].str.rstrip('%').astype(float)

**# Plot the 'Measure X' values**

Import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))

plt.bar(df['Name of company'], df['Measure X'])

plt.xticks(rotation=90)

plt.xlabel('Name of Company')

plt.ylabel('Measure X')

plt.title('Measure X for Companies')

plt.tight\_layout()

plt.show()

**# Calculate the mean, standard deviation, and variance**

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

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

var\_x = df['Measure X'].var()

print(f"Mean (μ): {mean\_x:.2f}")

print(f"Standard Deviation (σ): {std\_x:.2f}")

print(f"Variance (σ^2): {var\_x:.2f}")

**OUTPUT:**

**Mean (μ): 33.27**

**Standard Deviation (σ): 16.95**

**Variance (σ^2): 287.15**

**# Calculate the first quartile (Q1) and third quartile (Q3)**

q1 = df['Measure X'].quantile(0.25)

q3 = df['Measure X'].quantile(0.75)

**OUTPUT:**

**Q1 = 25.47**

**Q3 = 33.975**

**# Calculate the interquartile range (IQR)**

IQR = q3 - q1

IQR

**OUTPUT:**

**IQR = 8.505000000000003**

**# Define the lower whisker and upper whisker for outliers**

lower\_whisker = q1 - 1.5 \* IQR

upper\_whisker = q3 + 1.5 \* IQR

**OUPUT:**

**lower\_whisker =12.712499999999995**

**upper\_whisker =46.7325**

**# Find outliers**

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

outliers

**OUTPUT:**

**Name of company Measure X**

**Morgan Stanley 91.36**



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.

**Solution:**

**We can draw the following inferences from the above box plot**

**IQR(inter quartile range) = Q3-Q1**

**Q1 = 5**

**Q3 = 12**

**Therefore IQR = Q3-Q1**

**= 12-5**

**= 7**

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

**Answer:**

**By visualizing the above box plot we can say that the box plot is skewed towards the**

**right side and the median is on the left side.**

**Thus, we can say that the box plot is not following the normal distribution**

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:**

**The data point 25 is an outlier in the above box plot. If it was found that the value was 2.5 instead of 25 then the skewness of the box plot would be reduced and there will be high chances that the data would be normally distributed as there will be no outliers.**



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

**Answer:**

**The mode of the dataset lies between 4 and 10 approximately.**

1. Comment on the skewness of the dataset.

**Answer:**

**By visualizing the above histogram, we can say that the skewness of the data set is on the right side and thus have tail on the right side.**

**Hence, we can say that the histogram is positively skewed.**

**Therefore Mean>Median>Mode**

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: Both the box plot and the histogram are positively skewed thus both have the outliers on the right side. The outlier data point on box plot and histogram is same that is 25. By visualizing them we can say that there is a possibility that both box plot and histogram represents the same dataset.**

**Through box plot we can identify the median by the IQR and in the histogram it is easy to identify mode as we can see where the data points are highly repeated.**

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.)

**Solution:**

**According to the situation stated above in the question. From 200 long distance**

**Calls 1 call is getting connected to the wrong person.**

**Probability of call getting misdirected = 1**

**200**

**Probability of call correctly connected = 1-1 = 199**

**200 200**

**Findig the probability that at least one of the five attempted calls connects to the wrong number**

**Number of calls (n) = 5**

**P = 1**

**200**

**Q = 199**

**200**

**P(x) = probability of at least one of the five attempted calls connects to the wrong number**

**Findig the probability of at least one of the five attempted calls connects to the wrong number by using below formula :**

**# nCr = n! \* (n-r)!**

**r!**

**P(x) = ⁿCₓ pˣ qⁿ⁻ˣ**

**P(1)** = **(nCx) (p^x) (p^n-x)**

**P(1) = (5C1) (1/200) ^1(199/200) ^5-1**

**P(1) = 0.0245037**

**From the above calculations we can say that the probability that at least one of the five attempted calls connects to the wrong numbers is 0.0245037.**

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 |

**Solution: E(X) =Sum X\*P(X) | E(X^2) = X^2\*P(X)**

|  |  |
| --- | --- |
| **-200** | **400000** |
| **-100** | **100000** |
| **0** | **0** |
| **200** | **200000** |
| **600** | **1200000** |
| **300** | **900000** |
| **Total : 800 2800000** | |

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

**Answer: The most likely monetary outcome of the business venture is 2000$ as we can see that the probability for 2000$ is the highest, that is 0.3.**

1. Is the venture likely to be successful? Explain

**Answer: we can say whether the venture will be successful by performing the calculations below:**

**p(x>0)+p(x>1000)+p(x>2000)+p(x>3000) = 0.2+0.2+0.3+0.1**

**= 0.8**

**From the above calculations we can say that there are good chances for the venture to be successful as we can see that the probability of earning 0 or more than 0 is 0.8 = 80% and the chances of making loss is 20%.**

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

**Answer: The long-term average earning of business venture is expected to be**

**= Sum (X\*P(X))**

**= 800$**

**Therefore, we can say that the average earnings will be 800$ .**

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

**Answer: The good measure of the risk involved in a venture of this kind depends on the variability in the distribution.**

**Higher variance leads to more chances of risk**

**Var (X) = E(X^2) – (EX))^2**

**= 2800000 – 800^2**

**= 2160000**