**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% |

Solution:

# import module

from tabulate import tabulate

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sn

data\_1 = {'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 }

Company\_name = list(data\_1.keys())

Measure\_X = list(data\_1.values())

fig = plt.figure(figsize = (25,10))

# creating the bar plot

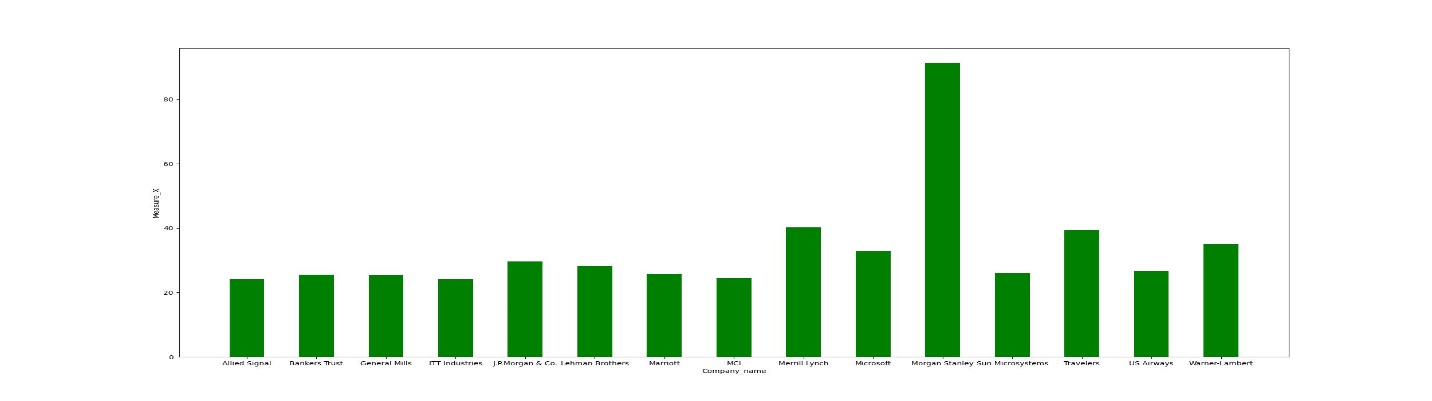
plt.bar(Company\_name, Measure\_X, color ='green',width = 0.5)

plt.xlabel('Company\_name')

plt.ylabel('Measure\_X')

# Saving the figure.

plt.savefig("output.jpg")



import statistics

data\_2= [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]

x=statistics.mean(data\_2)

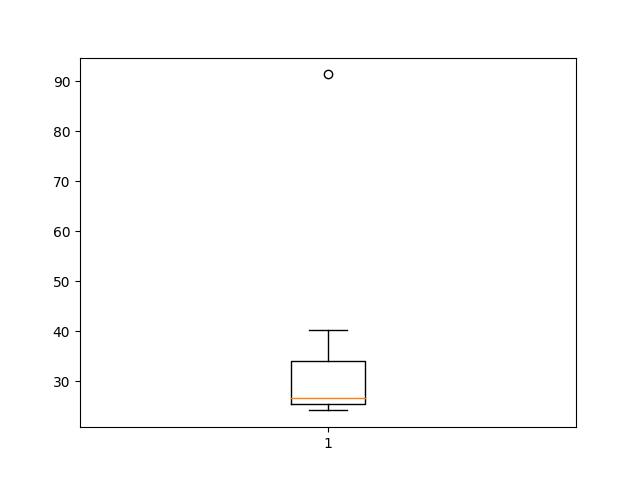
y=statistics.stdev(data\_2)

z=statistics.variance(data\_2)

print('Mean : ',x,'Standard Dev : ',y,'Variance : ',z)

plt.boxplot(data\_2,vert = True)

**Mean** : 33.27133333333333 **Standard Dev** : 16.945400921222028 **Variance** : 287.1466123809524



From the boxplot it is observed that the outlier is 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.

Sol:

Assume the dataset= {1,2,3,5,5,6,6,8,9,12,13,25}

Median(Q2)=(6+8)/2=7

Q1: Consider {1,2,3,5,5,6,6}

Median for Q1 is 5

Q2:Consider {8,9,12,13,25}

Median for Q3 is 12

Q3=12, Q2=7, Q1=5

inter-quartile range (IQR ): It tells about spreading of data. More the IQR high will be the dispersion, vice-versa. It is calculated as : Q3-Q1= 12-5=7

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

Sol: Hence, the median is closer to the bottom of the box and the whisker is shorter on the lower end of the box, the distribution is right-skewed

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?

Sol: If we change 25 to 2.5, the values of Q1,Q2,Q3,IQR gets change and it is observed that there are no outliers.



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

Sol: two modes are there, it is a bi modal

1. Comment on the skewness of the dataset.

Sol: right-skewed or positive skewed

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.

Sol: both the plots conveys the same information like the outliers, skewness and the dispersion. For to draw inferences quickly from large data set boxplot is preferred in finding outliers.

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.)

Sol:

Here there are two possible outcomes, one is call directed properly, other is misdirected.

Hence there are two possible outcomes, consider binomial distribution:

P(x:n,p) = nCx px (1-p)n-x

n = 5

p = 1/200

Calculate the probability that at least one in five attempted telephone calls reaches the wrong number

P(1:5,1/200) = at least one in five attempted telephone calls reaches the wrong number

= (5C1) (1/200)1 (199/200)5-1

= 0.024

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?

Sol: Most likely monetary outcome of the business venture is $ 2000 as it has maximum Probability  0.3

1. Is the venture likely to be successful? Explain

Sol: If P(X>0) > P(X<=0) , then the venture is successful

P(X=1000)+ P(X=2000)+ P(X=3000) > P(X=0)+ P(X=-1000)+ P(X=-2000)

0.2+0.3+0.1 > 0.1+0.1+0.2

0.6>0.4

Hence the venture is successful

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

Sol: = $800

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

Sol: find the standard dev and compare with average, if the deviation is more from the average value, then the risk will be more

SD==

= 2800000

= 800

Var (X) = E(X^2) –(E(X))^2

= 2800000 – 800^2

= 2160000

SD==1469.6 which is much high than Avg=800 hence there is a high risk