**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% |

import pandas as pd

import numpy as np

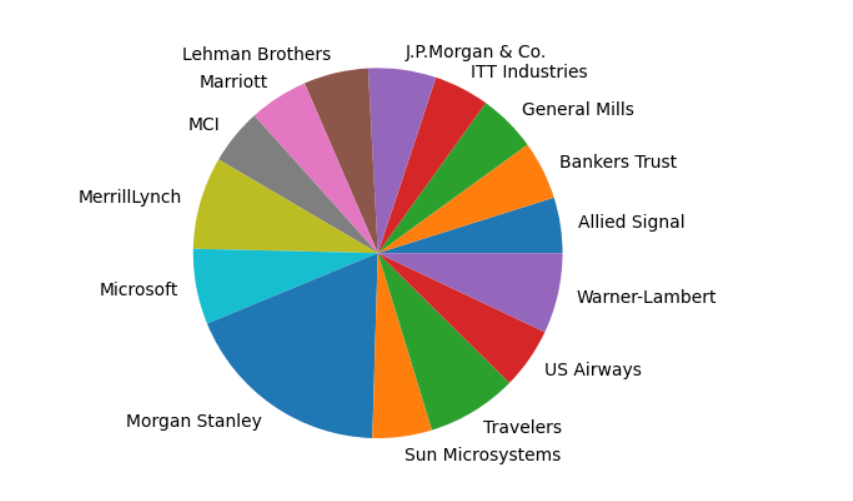
import matplotlib.pyplot as plt

data=pd.Series([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])

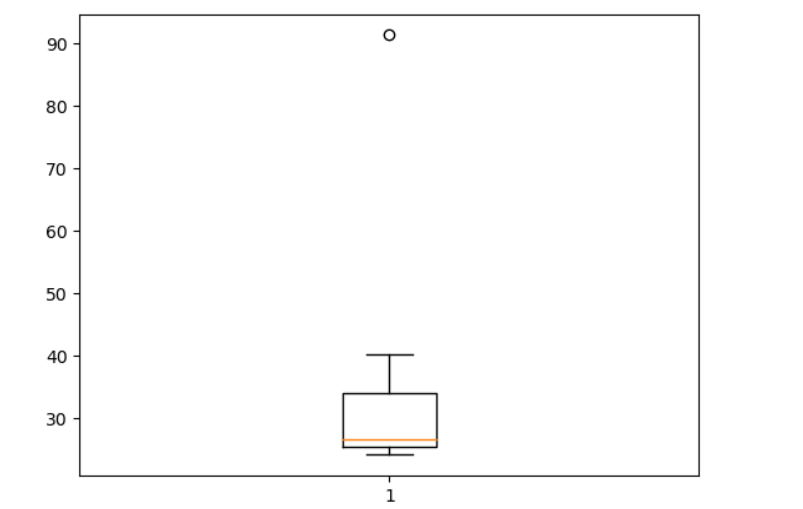
names=['Allied Signal','Bankers Trust','General Mills','ITT Industries','J.P.Morgan & Co.','Lehman Brothers','Marriott','MCI','MerrillLynch','Microsoft','Morgan Stanley',

'Sun Microsystems','Travelers','US Airways','Warner-Lambert']

plt.pie(data,labels=names)



plt.boxplot(data)



The boxplot is used to identify the Outliers in the data. The above boxplot indicates the outlier which is the value represented in the form of small dot.

The outlier in this data is the extreme value which is 91.36%

Mean of the population

data.mean()

33.271333333333

Variance of the population data

data.var()

287.1466123809524

Standard Deviation of population

data.std()

16.945400921222028



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.

Inter Quartile Range (IQR) = Q3 – Q1 = 12 – 5 = 7

IQR implies to the range between the upper quartile which is the 75% level and the lower quartile which is the 25% level. IQR tells us the range of the middle 50% of data.

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

The above dataset is positively skewed as it has the longer tale on the right side of the median. The dataset is said to be Positively skewed if it has the longer tail on the right side that is skewness > 0

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?

The datapoint 25 in the above boxplot indicates the outlier. Outliers are generally the extreme values in the dataset which affect or influence the mean of the dataset.

In the above boxplot the outlier 25 is on the right side and so it has the longer tail on the right side and the median, mean would get affected if it was 2.5 instead of 25.

And there would not be any outlier in the dataset, it would also not be a skewed data and it might be a normal distribution of the probabilities.



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

Ans) Mode of the above dataset would be the frequently occurring values of X with respect to Y which is between 5 to 8 i.e., Mode = 20

1. Comment on the skewness of the dataset.

Ans) The dataset is positively skewed as it has the longer tail on the right side of the plot curve and hence if it is positively skewed it means Skewness > 0

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.

Ans) Both the plot in the question 2 and 3 are positively skewed and have the longer tail of curve towards the right side.

They are called skewed data which means both of them are not Normally distributed and the skewness is > 0. This is how both the graphs complement each other.

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.)

Probability of a call getting misdirected = 1/200

Probability of a call not getting misdirected = 1 – (1/200) = 199/200

Total attempted telephone calls = 5

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

= 1 – (199/200)5 = 1 – 0.975 = 0.025

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?

The most likely monetary outcome is at the $2000 in the business venture because it has the highest probability when compared to the other values.

1. Is the venture likely to be successful? Explain

Yes, this venture is likely to be successful because probabilities for the dollar amount greater than 0 i.e., P(x>0) are adding up to 0.2+0.3+0.1 = 0.6 which is the 60% of the amount comes under the profit.

Hence, a business venture with almost 60%nof profit is likely to be successful.

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

The total average expected earning value of business ventures

= x\*p(x)

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

=-200-100+200+600+300 = 800

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

The good measure of risk is defined by the probabilities of x value less than 0 which is the negative amount values

P(x<0) =P (-2000) +P (-1000) = 0.1+0.1 = 0.2 which is the 20% of the amount goes under losses.