**Topics: Descriptive Statistics and Probability**

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

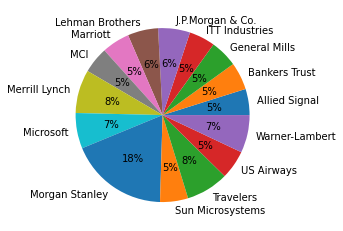
|  |  |
| --- | --- |
| **Name of company** | **Measure X** |
| Allied Signal | 24.23% |
| Bankers Trust | 25.53% |
| General Mills | 25.41% |
| ITT Industries | 24.14% |
| J.P.Morgan & Co. | 29.62% |
| Lehman Brothers | 28.25% |
| Marriott | 25.81% |
| MCI | 24.39% |
| Merrill Lynch | 40.26% |
| Microsoft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

Ans :

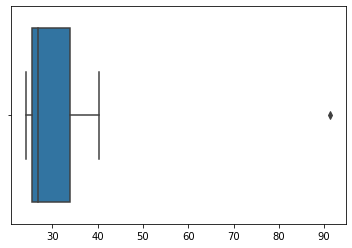
x=pd.Series([24.23,25.53,25.41,24.14,29.62,28.25,25.81,24.39,40.26,32.95,91.36,25.99,39.42,26.71,35.00])

name=['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']

plt.pie(x,labels=name,autopct='%1.0f%%')



sns.boxplot(x)



From the above boxplot, we can say that there is one outlier which measures more than 90% which is Morgan Stanley i.e. 91.36%.

Also the data is right skewed because mean is greater than median.

Mean:

= 33.27133333333333

Median= 26.71

Standard Deviation: = 16.945400921222028

Variance:  *=*287.1466123809524



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.
2. What can we say about the skewness of this dataset?
3. If it was found that the data point with the value 25 is actually 2.5, how would the new box-plot be affected?

Ans : i) First Quantile Range(Q1) = 5

Third Quantile Range(Q3) =12

Inter-Quantile Range (IQR) = Q3- Q1

= 12-5

= 7

This value implies that median value is 7.

ii) From the above boxplot, we can say that dataset is positively or right skewed.

iii) From the above boxplot, we can conclude that data point with value 25 is an outlier. If its actual value is 2.5 then there will be no outliers in dataset and positive skewness will get reduced and we can see normal distribution for the above data set.



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?
2. Comment on the skewness of the dataset.
3. 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 :

1. From the above histogram, we can say that the mode of this dataset will lie between 4-10 approximately.
2. The dataset is positively or right skewed since mean > median.
3. If both histogram and box-plot are plotted together than both will give information about outliers & both will be positively skewed. Also from histogram we can visualize mode and from boxplot we can visualize median.
4. 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.)

Ans :

Probability of call misdirecting (a) = 1/200

Probability of call not misdirecting (b) = 1-1/200

= 199/200

P(x) = ⁿCₓ aˣ bⁿ⁻ˣ

n = 5

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

= 1 - no call reaches wrong number

= 1 – P (0)

= 1 -  ⁵C₀(1/200)⁰(199/200)⁵⁻⁰

= 1 – (199/200)5

= 0.02475

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?
2. Is the venture likely to be successful? Explain
3. What is the long-term average earning of business ventures of this kind? Explain
4. What is the good measure of the risk involved in a venture of this kind? Compute this measure

Ans :

1. The most likely monetary outcome of the business venture is 2000 as it has highest probability value.
2. The probability that the venture will make profit i.e., greater than 0

= P (1000) + P (2000) + P (3000)

= 0.2 + 0.3 + 0.1

= 0.6

So there are 60% chances of the venture likely to be successful.

1. Expected Value = ∑ P(X).E(X)

= (-2000 \* 0.1) + (-1000 \* 0.1) + (0 \* 0.2) + (1000 \* 0.2) +

(2000 \* 0.3) + (3000 \* 0.1)

= -200 + -100 + 0 + 200 + 600 + 300

= 800

The long-term average earning of business ventures of this kind will be 800.

1. The good measure of the risk involved in a venture of this kind depends on the Variability in distribution. Higher Variance means more chances of risk

Var (X) = E(X^2) –(E(X))^2

= 2800000 – 800^2

= 2160000