**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:-** R- Code-

install.packages("readxl")

library(readxl)

BS2<-read\_excel("C://Users//Lenovo//Desktop//ExcelR//BS2.xlsx")

View(BS2)

mean(BS2$`Measure X`)

sd(BS2$`Measure X`)

var(BS2$`Measure X`)

hist(BS2$`Measure X`)

boxplot(BS2$`Measure X`)

OutVals=boxplot(BS2$`Measure X`)$out

OutVals

Output-

>mean(BS2$`Measure X`)

[1] 0.3327133

>sd(BS2$`Measure X`)

[1] 0.169454

>var(BS2$`Measure X`)

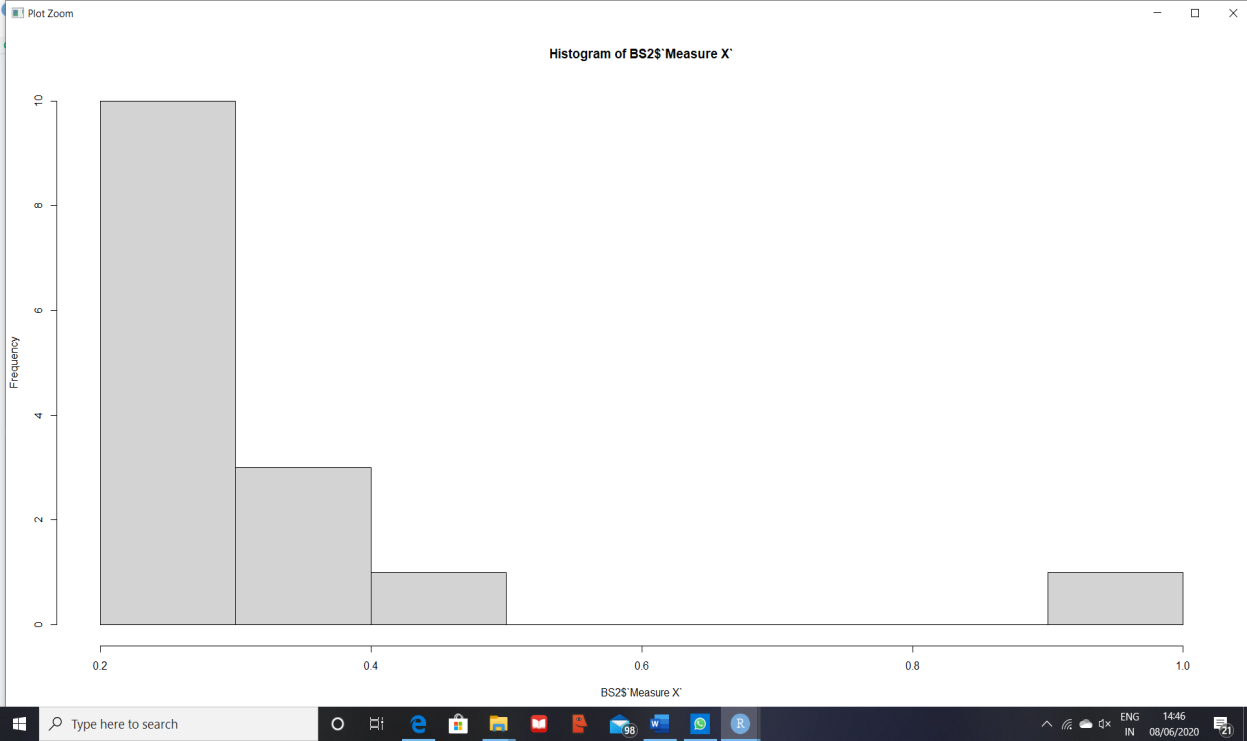
[1] 0.02871466

>OutVals

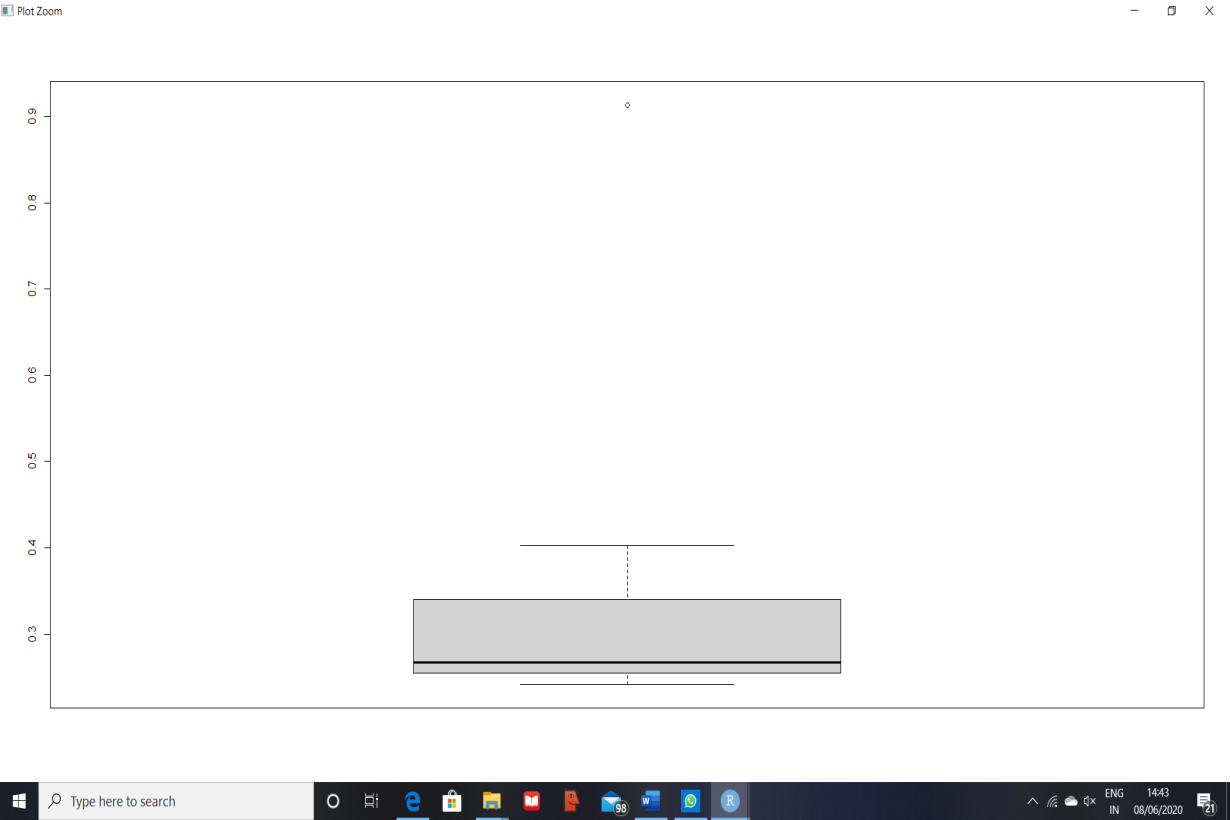
[1] 0.9136

Plot-

1. Histogram Plot



1. Boxplot





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) Inter-quartile Range = I3-I1 = 12-5= ~**7**

In [descriptive statistics](https://en.wikipedia.org/wiki/Descriptive_statistics), the **interquartile range** (**IQR**), also called the **midspread**, **middle 50%**, or **H‑spread**, is a measure of [statistical dispersion](https://en.wikipedia.org/wiki/Statistical_dispersion), being equal to the difference between 75th and 25th [percentiles](https://en.wikipedia.org/wiki/Percentiles), or between upper and lower [quartiles](https://en.wikipedia.org/wiki/Quartile).

Here, as per observation, median is also approximately (~)7. Hence, IQR is nearly equal to median value.

(ii) As the longer part of the box is to the right of the median, the dataset is **positively skewed** or say **skewed Right.**

1. Outlier (25) is at the Right side of the boxplot. If its treated as 2.5 then there will be no change in the plot.



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:-** (i) As per observation, data has Bi modal value of **21 (~Approx)** for the **interval 4**

**to 8 (~approx)**. Hence, mode of the given dataset lies somewhere between 4 to 8 (~Approx) of Y value.

(ii) Data is **Right or Positively skewed.**

(iii) A histogram is highly useful when wide or very little variances exist among the observed frequencies for a particular data set. As seen in the graph, the histogram shows that there are two peaks within the data, indicating it is Bi-modal (two commonly recurring groups of numbers).While in the box plot these values would average one another out and will distribute the data evenly. Median can be easily predicted or any dataset using boxplot while we can easily check the modal values.

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.)

**Ans:-** Lets define an event E = The call is misdirected

Probability of Event E is P(E) = 1/200

So, P(not E) = 1 – P(E) = 1 – (1/200) = 199/200

Probability that at least one in 5 attempted call reaches the wrong number is

= 1 - Probability that no attempted call reaches the wrong number

= 1 – [(199/200) \* (199/200) \* (199/200) \* (199/200) \* (199/200)]

= 1 - (199/200)^5

= 7920399001/200^5

= 0.025

**Therefore, Probability that at least one in 5 attempted call reaches the wrong number is 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?
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:-** (i) P(𝑋) = ∑ 𝑋𝑖∗𝑃(𝑋𝑖) where 𝑖= 1 to 6

=(−$2000 ∗ 0.10)–( $1000 ∗ 0.10) + 0 + ($1000 ∗ 0.20) + ($2000 ∗ 0.30)+

($3000 ∗ 0.10)

= $800

(ii) First Calculate the probability the business return is a non-negative number

(𝑋 ≥ 0) = 𝑃(0) + 𝑃($1000) + 𝑃($2000) + 𝑃($3000) = 0.20 + 0.20 + 0.30 + 0.10

= 0.80

If the venture can maintain for long term business then eventually it will be successful since the probability of non-negative return is higher than 0.50 and the expected value for return is a positive number ($800).

(iii) From the above question requirement, we have to consider similar business ventures of this type whose distribution of the returns is similar to this venture. In that case we say that the expected value of returns to this particular venture is the required average.

(-2000\*0.1)+(-1000\*0.1)+0+(1000\*0.2)+(2000\*0.3)+(3000\*0.1)=800

Therefore, the long-term average earning for these type of ventures would be around $800.

(iv) Risk stems from the possible variability in the expected returns. Therefore a good measure to evaluate the risk for a venture of this kind would be variance or standard deviation of the variable X.

> sd(ex$x)

[1] 1870.829

> var(ex$x)

[1] 3500000

The large value of standard deviation of $1870 is considered along with the average returns of $800 indicates that this venture is highly risky.