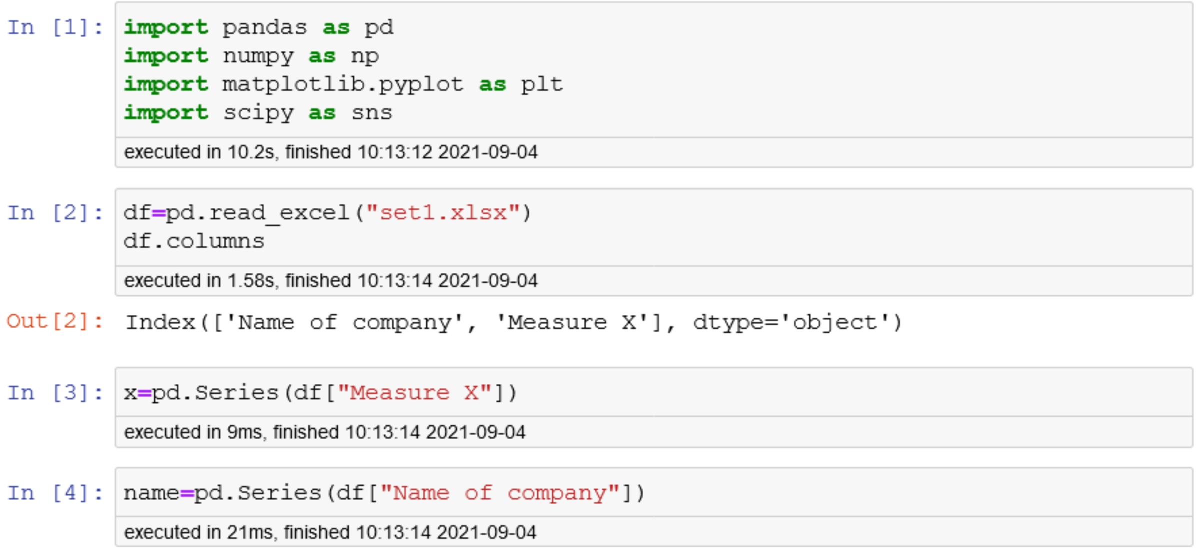
Topics: Descriptive Statistics and Probability

1. Look at the data given below. Plot the data, find the outliers and find out μ, σ, σ^2

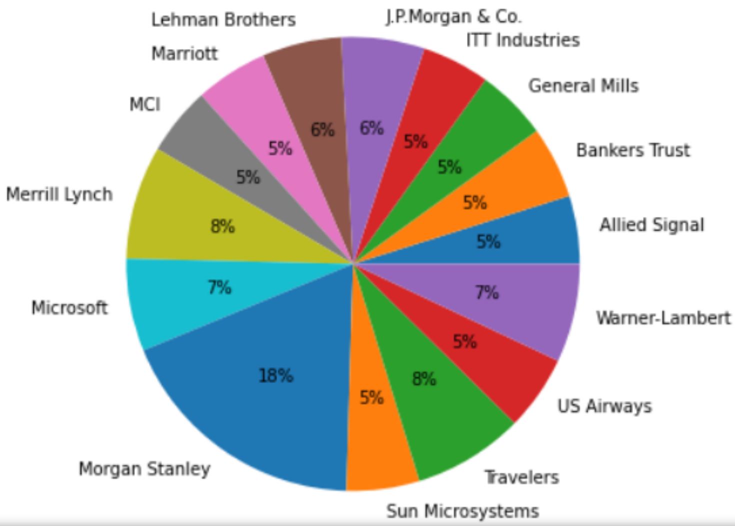
Name of company	Measure X
Allied Signal	24.23%
Bankers Trust	25.53%
General Mills	25.41%
ITT Industries	24.14%
J.P.Morgan & Co.	29.62%
Lehman Brothers	28.25%
Marriott	25.81%
MCI	24.39%
Merrill Lynch	40.26%
Microsoft	32.95%
Morgan Stanley	91.36%
Sun Microsystems	25.99%
Travelers	39.42%
US Airways	26.71%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



```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import scipy as sns
        executed in 10.2s, finished 10:13:12 2021-09-04

In [2]: df=pd.read_excel("set1.xlsx")
        df.columns
        executed in 1.58s, finished 10:13:14 2021-09-04

Out[2]: Index(['Name of company', 'Measure X'], dtype='object')

In [3]: x=pd.Series(df["Measure X"])
        executed in 9ms, finished 10:13:14 2021-09-04

In [4]: name=pd.Series(df["Name of company"])
        executed in 21ms, finished 10:13:14 2021-09-04
```

```
In [5]: # Pie Plot
        plt.figure(figsize=(6,8))
        plt.pie(x,labels=name,autopct='%1.0f%%')
        plt.show()
```

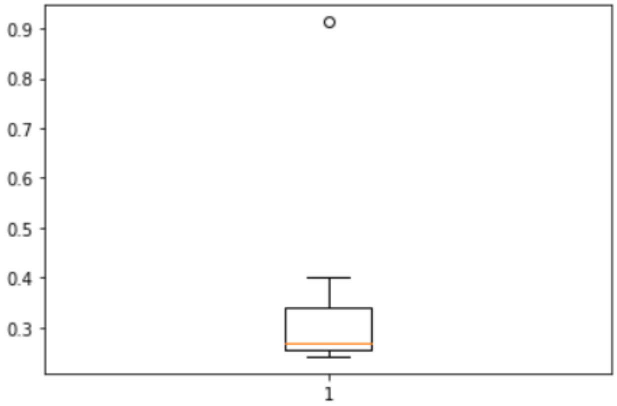
executed in 546ms, finished 10:24:18 2021-09-04



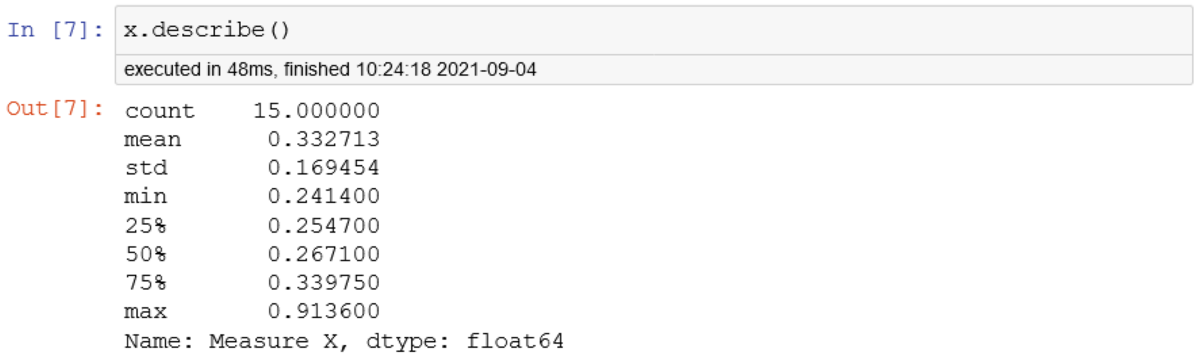
```
In [6]: plt.boxplot(x)
```

executed in 301ms, finished 10:24:18 2021-09-04

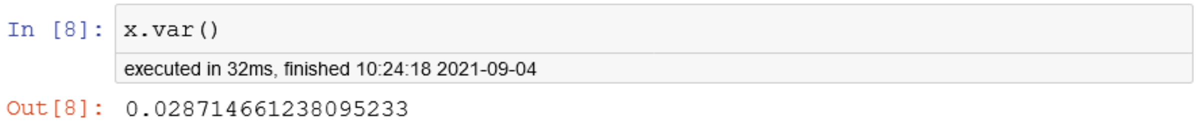
```
Out[6]: {'whiskers': [<matplotlib.lines.Line2D at 0x18d5c2d0250>,
          <matplotlib.lines.Line2D at 0x18d5c2d0e50>],
         'caps': [<matplotlib.lines.Line2D at 0x18d5c2de340>,
          <matplotlib.lines.Line2D at 0x18d5c2de790>],
         'boxes': [<matplotlib.lines.Line2D at 0x18d5c2d0c70>],
         'medians': [<matplotlib.lines.Line2D at 0x18d5c2def10>],
         'fliers': [<matplotlib.lines.Line2D at 0x18d5c2dec40>],
         'means': []}
```



Inference: There is one outliar: Morgan Stanley at 91.36%

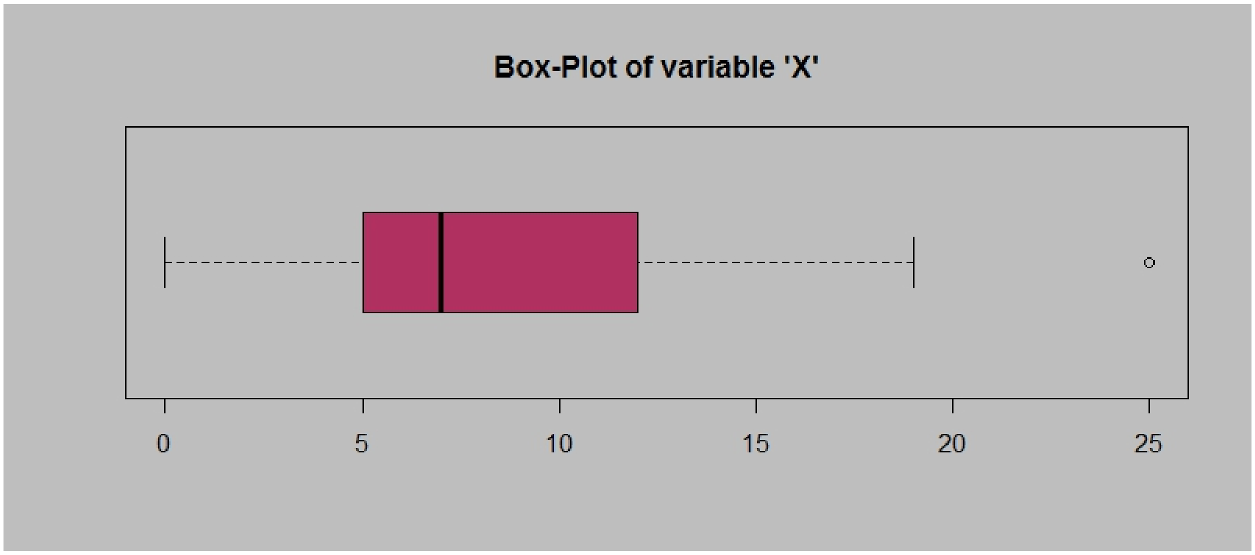


```
In [7]:  x.describe()
         executed in 48ms, finished 10:24:18 2021-09-04

Out[7]:  count    15.000000
         mean      0.332713
         std       0.169454
         min       0.241400
         25%       0.254700
         50%       0.267100
         75%       0.339750
         max       0.913600
         Name: Measure X, dtype: float64
```



```
In [8]:  x.var()
         executed in 32ms, finished 10:24:18 2021-09-04

Out[8]:  0.028714661238095233
```

2.



Answer the following three questions based on the box-plot above.

(i) What is inter-quartile range of this dataset? (Please approximate the numbers) In one line, explain what this value implies.

ANS:

INTER QUARTILE RANGE(IQR)=Q3-Q1

IQR=12-5=7, This represents the range which contains 50% of the data points.

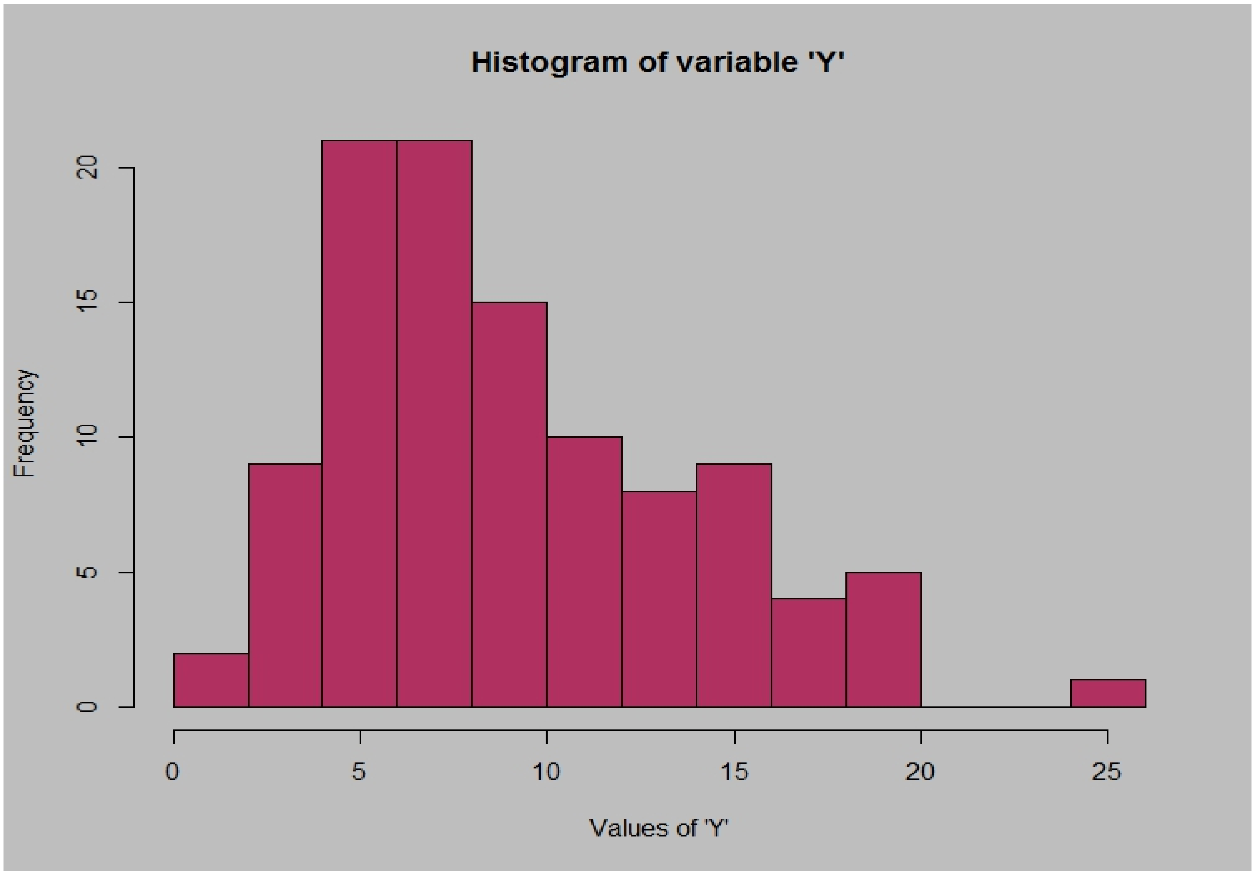
(ii) What can we say about the skewness of this dataset?

ANS: Right Skewed.

(iii) 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: 2.5 will be not considered as outlier. The boxplot will start from 0 to 20 in representation.

3.



Answer the following three questions based on the histogram above.

(i) Where would the mode of this dataset lie?

ANS: Mode lies between 4 & 8.

(ii) Comment on the skewness of the dataset.

ANS: Datasets is Right Skewed.

(iii) 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:

- Median in boxplot and Mode in histogram
- Histogram provides the frequency distribution so we can see how many times each data point is occurring however boxplot provides the quantile distribution i.e., 50% data lies between 5 and 12.
- Boxplot provides whisker length to identify outliers, no information from histogram. We can only guess looking at the gap that 25 may be an outlier.

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 calls misdirected, p (μ) = 1/200

Probability of calls not misdirected, p (μ) = 1 -1/200 = 199/200

We can use the formula for binomial distribution

No. of calls done, n = 5

Probability that one of the cells misdirected = 1-P (0)

$= 1-{}^{n}c_{0}\, p^{x}\, q^{1-x} = 1-{}^{5}c_{0}\, (1/200)^{0}(199/200)^{5}$

$1-(199/200)^{5} = 0.02475 \approx$ **2.45%**

5. 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

(i) What is the most likely monetary outcome of the business venture?

ANS: Max. P = 0.3 for P (2000). So most likely outcomes are 2000

(ii) Is the venture likely to be successful? Explain

ANS: P(x>0) = 0.6, implies there is a 60% chance that the venture would yield profits or greater than expected returns. P (Incurring losses) is only 0.2. So, the venture is likely to be successful.

(iii) What is the long-term average earning of business ventures of this kind? Explain

ANS: Weighted average = x8P(x) = 800. This means the average expected earnings over a long period of time would be 800(including all losses and gains over the period of time)

(iv) What is the good measure of the risk involved in a venture of this kind? Compute this measure

ANS: P(loss) = P (x = -2000) +P (x =-1000) = 0.2.

So, the risk associated with this venture is 20%.