**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% |

Answer: The outlier in the boxplot: Morgan Stanley 91.36%

* Mean = 33.271333
* Standard deviation = 16.945401
* Variance = 287.1466123809524

**Code for visualizing the above data:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

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']

x.mean()

x.var()

x.std()

fig,ax=plt.subplots(2,2,figsize=(10,10))

ax[0,0].plot(name,x)

ax[0,0].set\_title('line plot')

ax[1,0].scatter(name,x)

ax[1,0].set\_title('scatter plot')

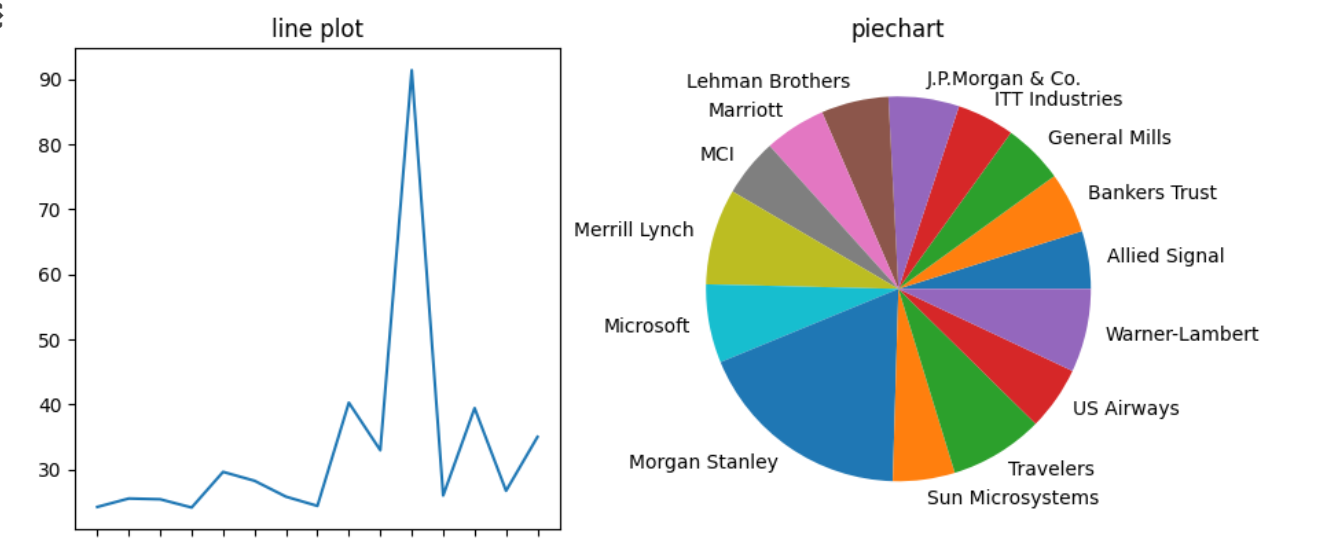
ax[0,1].pie(x,labels=name)

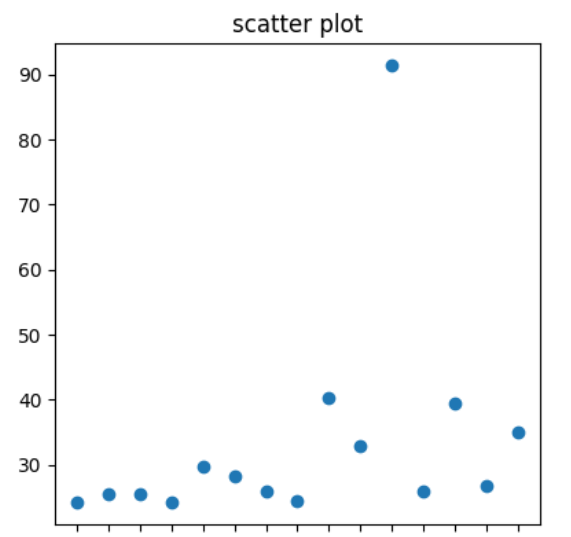
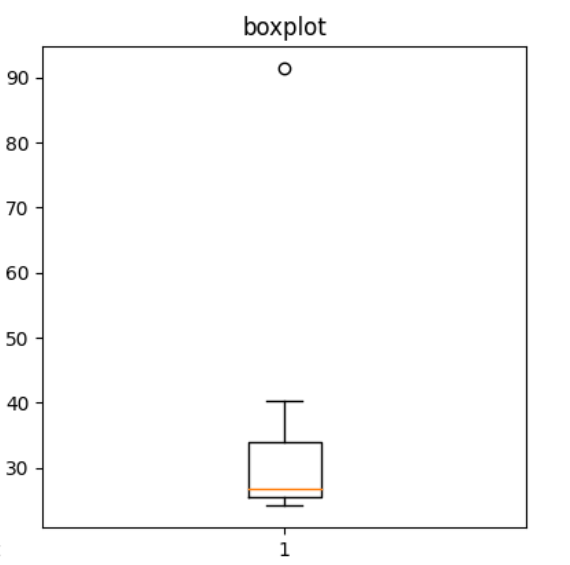
ax[0,1].set\_title('piechart')

ax[1,1].boxplot(x)

ax[1,1].set\_title('boxplot')

plt.show()





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?

**Answer:** (i) Lower quartile (LQ) = 5

Upper quartile (UQ) = 12

Inter-quartile range(IQR) = UQ – LQ = 12 – 5 = 7

* IQR is equal to median.

(ii) The right-skewed median is towards the left side it is not normal distribution.

(iii) In that case there would be no Outliers on the given dataset because of the outlier the data had positive skewness it will reduce and the data will normal distributed



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.

**Answer:** (i) The mode of this dataset will lie between 4 to 8.

(ii) This dataset is right skewed and positive in nature.

(iii) The Median in boxplot is 7 and the Mode in histogram is 4.

The 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% of data lies between 5 and 13.

Boxplot provides whisker length to identify outliers, no information from a histogram. We can only guess looking at the gap that 25 may be an outlier.

They both are right-skewed and both have outliers the median can be easily visualized in box plot as in histogram mode is more visible.

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.)

**Answer:** Number of Calls(n) = 5, p = 1/200, q = 199/200

P(x) = at least one in five attempted telephone calls reaches the wrong number

P(x) = ⁿCₓ pˣ qⁿ⁻ ˣ

P(x) = (nCx) (p^x) (q^n-x)

P(x) = (5C1) (1/200)^1 (199/200)^5-1

P(x) = 0.0245037

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

**Answer:** i) Most likely monetary outcome of the business venture is 2000

Its maximum probability is 0.3

(ii) P(x>0) = 0.6, implies there is a 60% chance that the venture would

yield profits or greater than expected returns. So the venture is likely to

be successful.

(iii) Weighted average = x\*P(x) = 80%. This means the average expected

earnings over a long period would be 800(including all losses and gains

over the period of time).

(iv) P(Incurring loss) = P(x= -2000)+P(x=-1000)=0.2. So the risk

associated with this venture is 20%.