**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% |

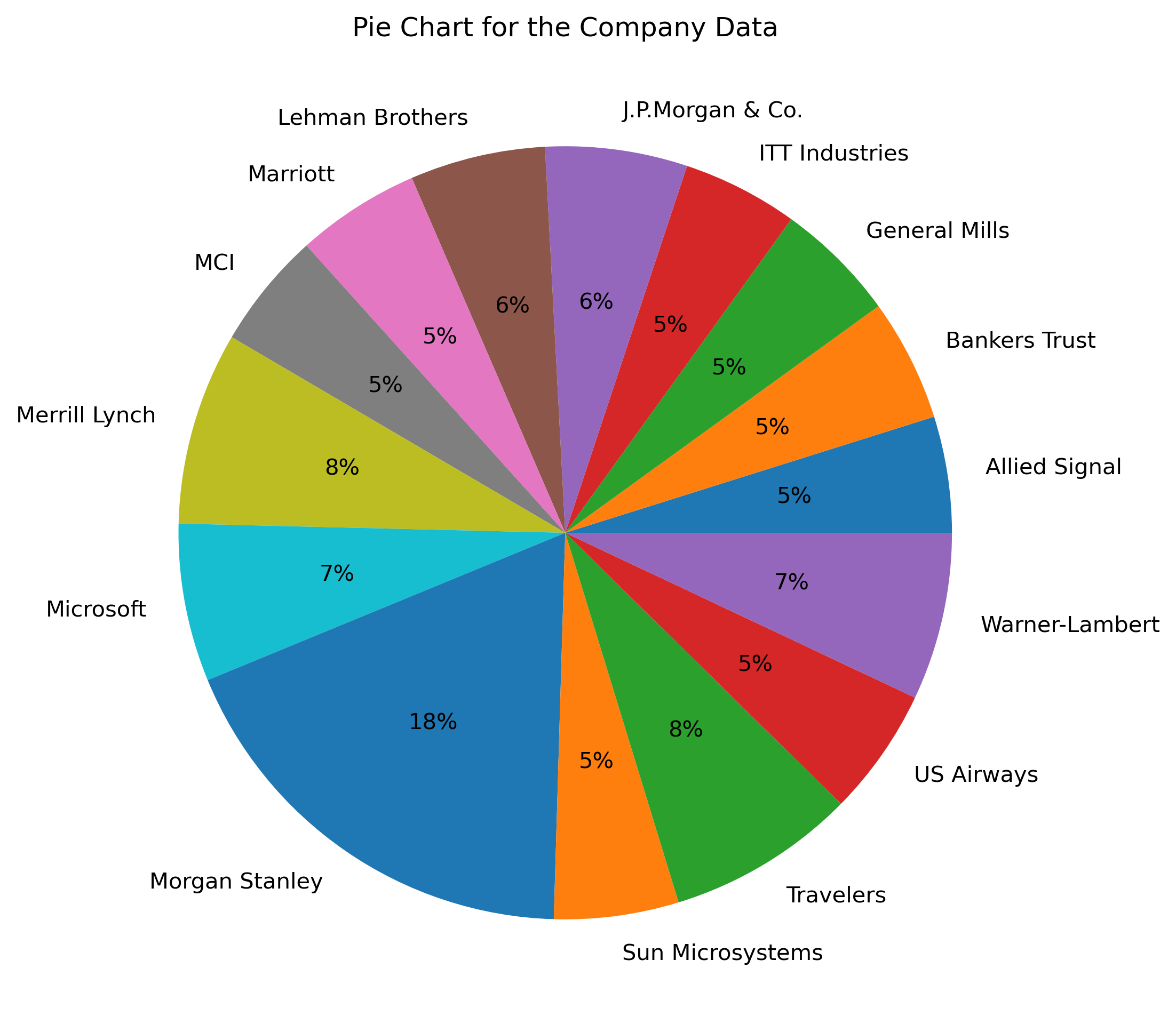
**Ans:**

**import** numpy **as** np  
import pandas **as** pd   
import matplotlib.pyplot **as** plt  
import seaborn **as** sns  
%**matplotlib** inline

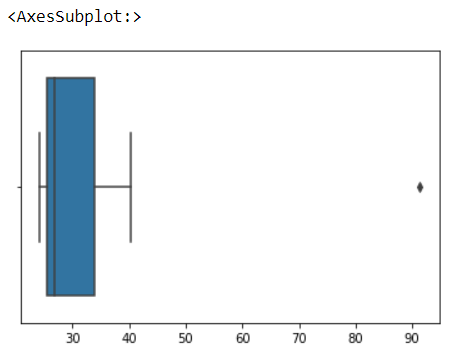
Measure = pd**.**Series([24.23,25.53,25.41,24.14,29.62,28.25,25.81,24.39,40.26,32.95,91.36,25.99,39.42,26.71,35.00])

name\_company**=**['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']

*#pie plot*  
*fig* **=** plt**.**figure(figsize**=**(15,8))  
plt**.**pie(measure, labels **=** name\_company, autopct **=** '%1.0f%%')  
plt**.**title('Pie Chart for the Company Data')  
plt**.**show  
fig**.**savefig('Pie Chart for the Company Data', dpi**=**300, bbox\_inches**=**'tight')



#Outliers

sns.boxplot(measure)  


From the above graph we can conclude “Morgan Stanley is an outlier of 91.36”.

#Mean(μ)

measure.mean()

**33.27133333333333**

#Standard Deviation(σ)

measure.std()

**16.945400921222028**

#variance(σ^2)

measure.var()

**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.

**Ans:** Inter quartile is Q3-Q1 =12-5 =7. this represents the range which contains 50% of the data points.

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

**Ans:** Positively skewed or right skewed.

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?

**Ans:** 2.5 will be not considered an outliner.The box plot will start from 0 and end at 20 in representation.



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

**Ans:** Between 4 and 8

1. Comment on the skewness of the dataset.

**Ans:** Positively Skewed or Right Skewed.

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:** We cannot differentiate mode in box plot but we can do that in histogram. – 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.

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:** One wrong number out of 200 Probability of wrong number: P(WN) = 1/200 = 0.005 Probability of not wrong number: 1 - P(WN) =(1- 1/200) = 0.995

Number of Calls = 5

P(x) = ⁿCₓpˣqⁿ⁻ˣ

n = 5

p = 1/200

q = 199/200

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

= 1 - none of the call reaches the wrong number

= 1 - P(0)

= 1 - ⁵C₀(1/200)⁰(199/200)⁵⁻⁰

= 1 - (199/200)⁵

= 0.0247

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
5. It can be seen from the above table that for x = 2000, the value of P(X) most. Hence, the most likely monetary outcome of the business venture is x = 2000.
6. 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.
7. Weighted average = x\*P(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)
8. P(loss) = P(x= -2000)+P(x=-1000)=0.2. So the risk associated with this venture is 20%.