**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% |

Mean : 33.27133333333333

Variance : 268.00350488888887

Standard Deviation : 16.370812590976932

Outliers Find

Outliers = []

def detect\_outliers(data):

q1 = data.quantile(.25)

q3 = data.quantile(.75)

iqr = q3 – q3

upper\_taill = q3 + 1.5 \* iqr

lower\_tail = q1 - 1.5 \* iqr

for out in data:

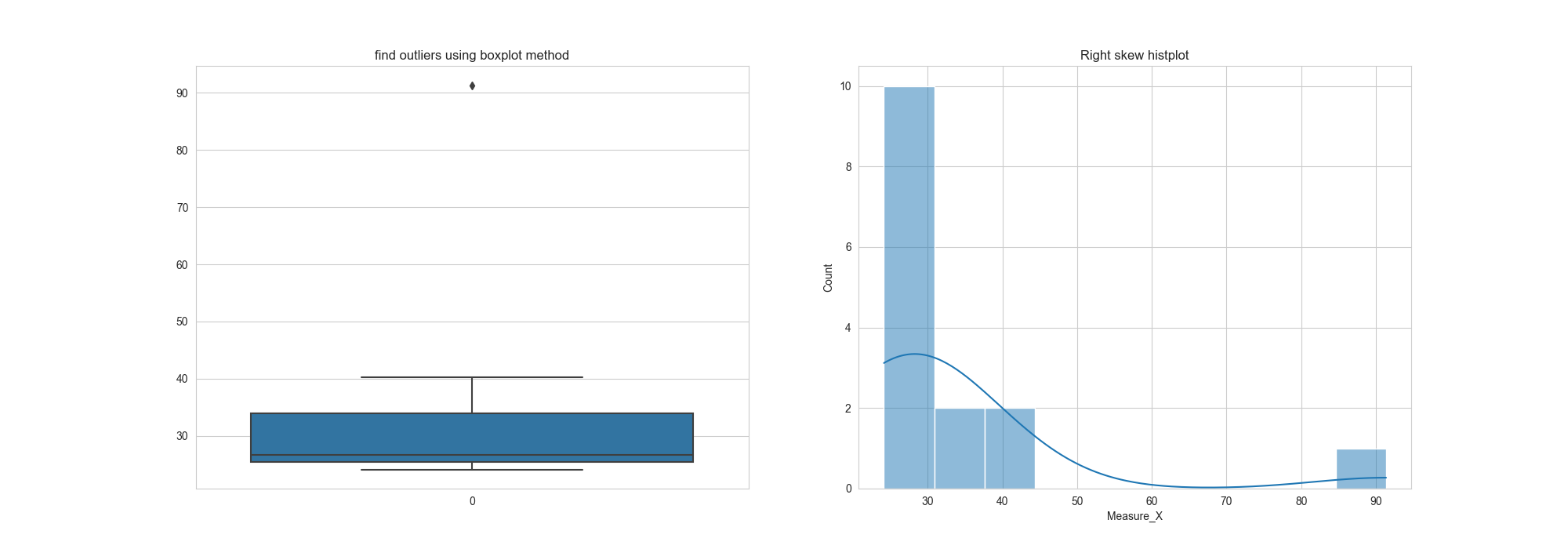
if out > upper\_taill or out < lower\_tail:

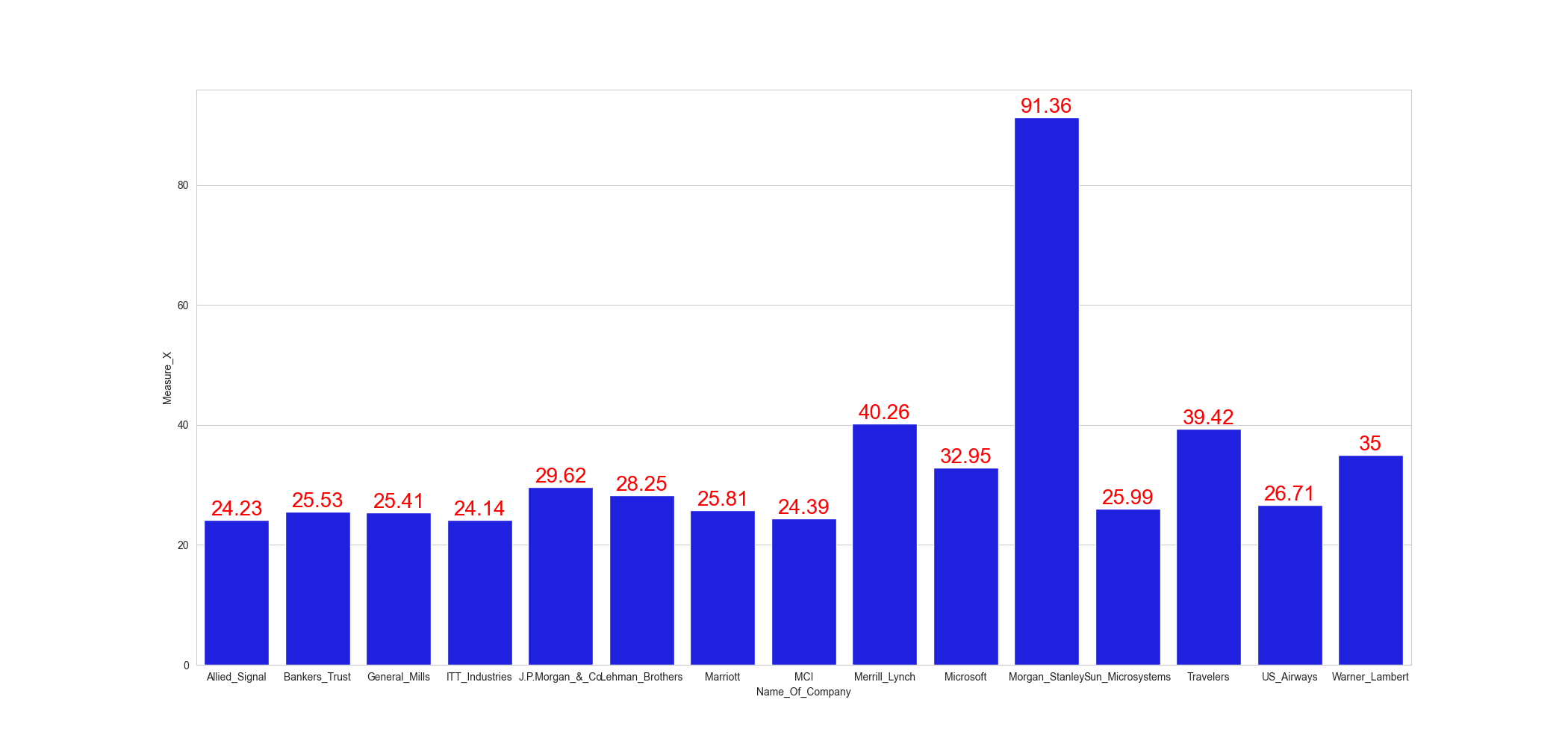
Outliers.append(out)

print(Outliers)

detect\_outliers(df['Measure\_X'])

Ans = [91.36]





2.



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 = IQR = Q3 – Q1 => 12-5=7 (approx.)

IQR is like the mid-spread of the data, which lies at 7, in the above case.

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

Ans = From the above boxplot we can say that the distribution of x is positively 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 = the box plot will still get skewed a little more to the left since most of the observation are towards the left before the median.

3.



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

Ans = We need to have actual data to get exact value of the mode.

1. Comment on the skewness of the dataset.

Ans = the mode can lie between 4 and 10 because there are many values in this range but this is just an assumption. The 2 bars of the same height don’t indicate more every time. By observing we can say that mode lies in between 4 to 8.

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 = Histogram helps in finding out mode, skewness, kurtosis. Boxplot is used to find IQR and outliers. Both have an outlier at 25.Moreover, if question1’s boxplot and question 2’s histogram is for the data set, then both together represent that the data is positively / left skewed. And also, the middle line in the box plot denotes the median, for which we can find where the median lies for our dataset by extending the same line onto the histogram. Also, both the histogram and the boxplot complement each other. The whiskers in boxplot, to the right are longer than that of left indicating that the number of occurrences of the observation is comparatively lower. Also, the IQR in the boxplot can be visualized in the histogram showing the range of the data.

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 =

Probability of an event ,(calls getting misdirected) E = 1 /200 = 0.005

Probability of call not misdirecting = 1 - (1/200) => 199/200 = 0.995

Probability of at least one of five is wrong number

1– Probability of at least one of five is not wrong number

=> 1-(1-0.005)^5

=> 0.024

(or)

P(x) = nC.pxqn-x

At least one in 5 calls 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?

Ans = Most likely outcome of business venture is 2000 since probability is highest with 0.3

1. Is the venture likely to be successful? Explain

Ans = For the venture to be successful, yes, P(x=1000) + P(x=2000) + P(x=3000) = 0.2+0.3+0.1 = 0.6

1. What is the long-term average earning of business ventures of this kind? Explain

Ans = Long-term earning for such venture :

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

1. What is the good measure of the risk involved in a venture of this kind? Compute this measure

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

Var (X) = E(X^2) –(E(X))^2

= 2800000 – 800^2

= 2160000