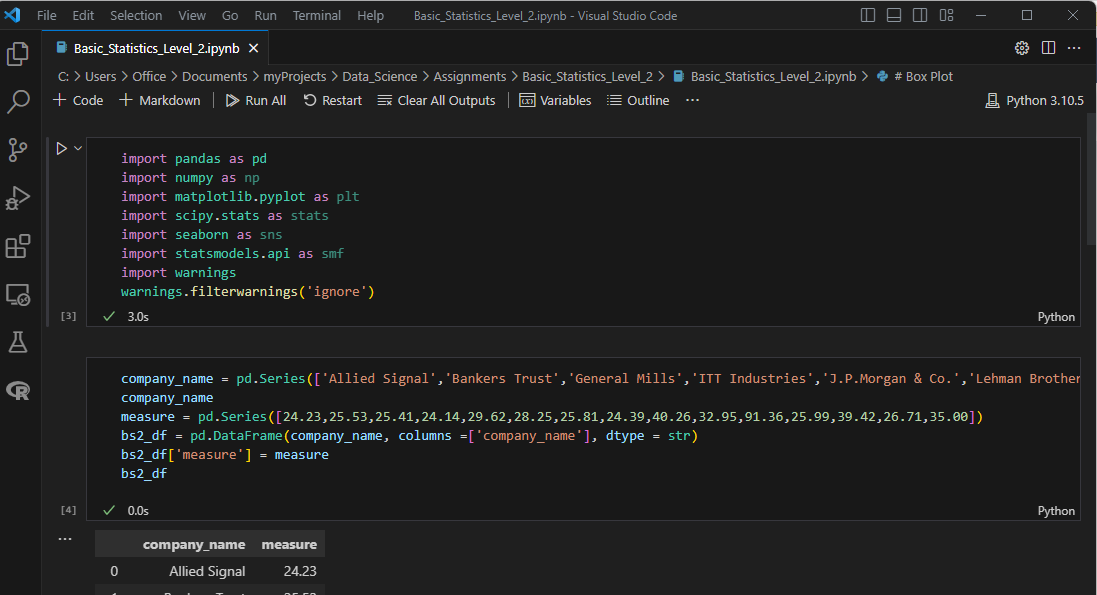
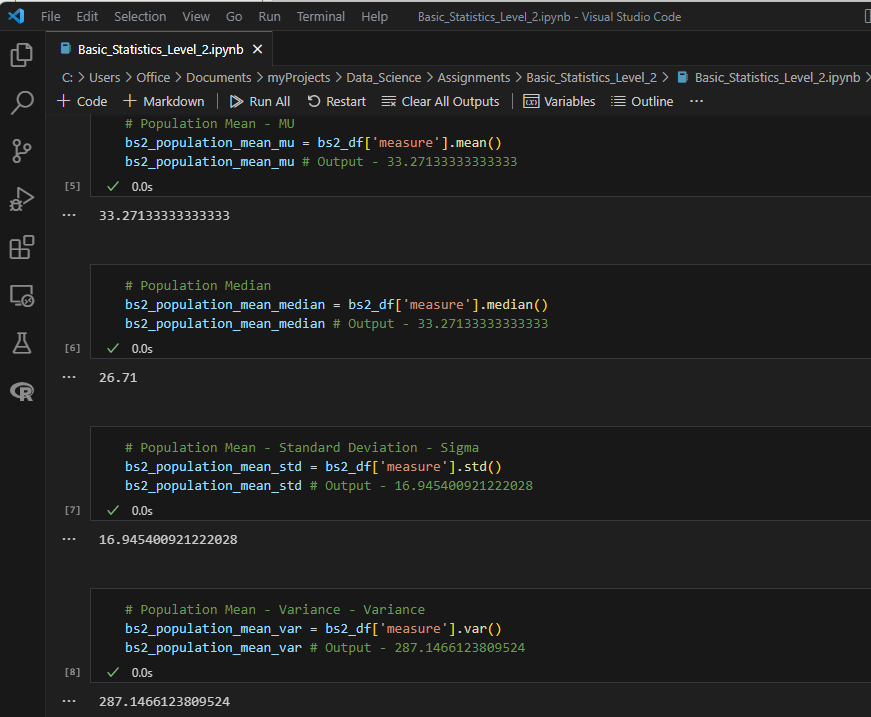
**Topics: Descriptive Statistics and Probability**

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% |

**Extracting Population Mean (), Standard Deviation () and Variance () using Python**

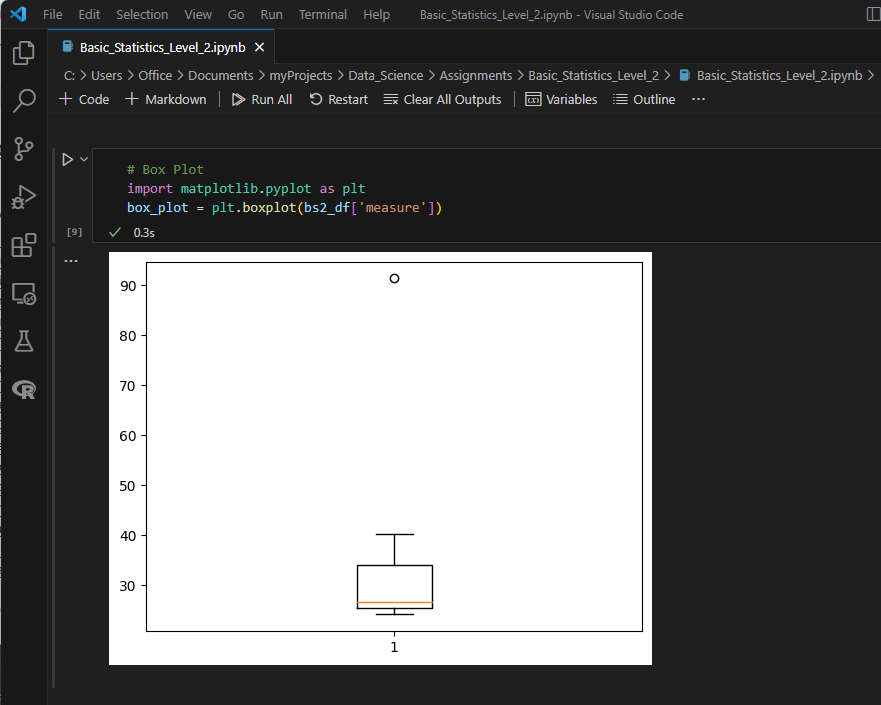


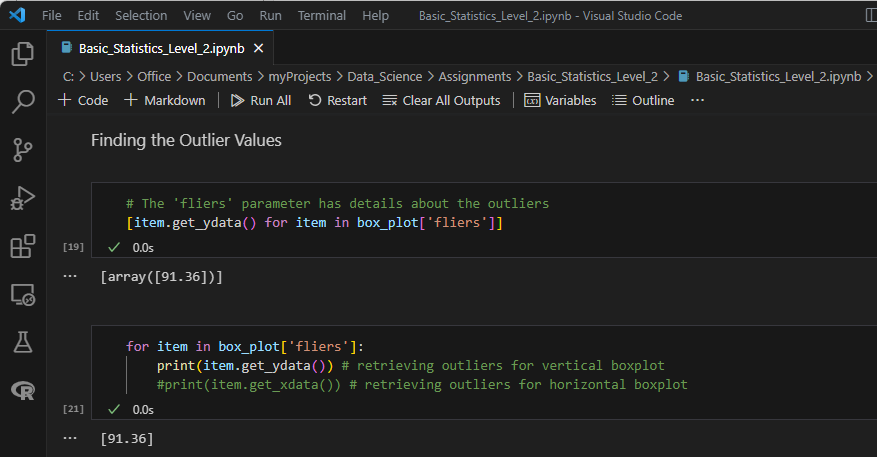


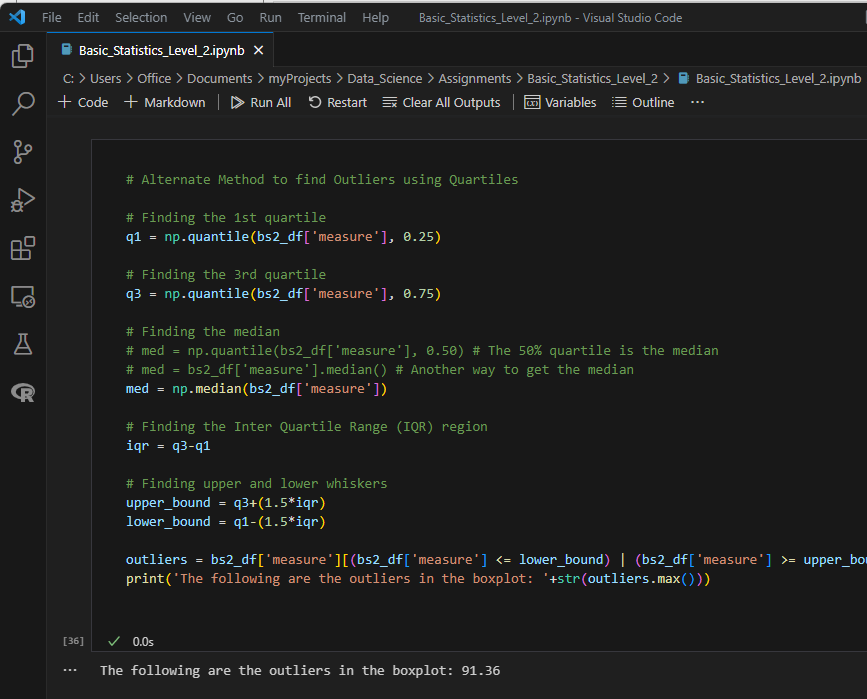
**Population Mean (): 33.27133333333333**

**Standard Deviation (): 16.945400921222028**

**Variance (): 287.1466123809524**







**Outlier is 91.36, which is the 11th Data Point in Measure.**



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.

**Solution:**

The Upper Quartile (UQ) – 12 (approximate value based on image)

The Lower Quartile (LQ) – 5 (approximate value based on image)

The Inter Quartile Range (IQR) for the dataset – (UQ-LQ) = 12-5 = 7 (approx.)

The Inter Quartile Range tells us the range of the middle half of the data.

The median is seen as 7 (approx.) based on the chart.

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

**Solution:**

Based on the chart, we can say that the data is skewed to the right i.e., Positively Skewed. There is a long tail on the right side due to an outlier value at 25.

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?

**Solution:**

The outlier value 25 was the only outlier and was causing the long tail to the right and the median shifting to the right. By correction of the data point with the value 25 to 2.5, there would be no outliers and the mean and median will get changed with both get shifted towards the left-hand side (as the new value is lesser than the median) and the skewness that was earlier towards the right will get reduced. It all depends on the other data points and their locations; if all are near the median and mean and this correction causes the mean and median to be the same, the result will be a curve that is symmetrical or normally distributed. If more points are towards the right of the median, then it may still remain skewed towards the right but the skewness will definitely get reduced, becoming closer to a normal distribution.



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

**Solution:**

The mode will be **between 4 and 8 (as these are the most frequent and are equal i.e., 22 approx.)**

1. Comment on the skewness of the dataset.

**Solution:**

The data is right skewed (positively skewed) with possibly one odd outlier value around the upper extreme i.e., data point 25, and therefore a long tail towards the upper extreme. The mode < median < mean. The median would be between 6 and 10.

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.

**Solution:**

If we assume that both the Box Plot in Question 2 and the Histogram for Question 3 are plotted for the same data, we can see that both graphs provide us with the same information mentioned below and complement each other –

1. Both graphs indicate the skewness towards the right i.e., the data is Positively skewed. The comparatively large right whisker in the box-plot and the spread of more frequencies towards the right in the histogram shows the skewness towards the right.
2. Both graphs show that there is one outlier points on 25.
3. Both graphs give us an indication of where the median lies, with the box plot showing the value of median more accurately but with histogram we can get a possible estimation.
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.)

**Solution:**

If 1 in 200 long-distance telephone calls are getting misdirected.

probability of call misdirecting (P(m)) = 1/200

Probability of call not Misdirecting or Correct Calls (P(c)) = 1-1/200 = 199/200

The probability of at least one in five attempted telephone calls reaches the wrong number i.e., (P(mc))

Total calls made (tc) = 5

Total expected misdirected (mc) = 1

Total expected not-misdirected (cc) = 5-1 = 4

P(mc) = tcCmc \* P(m)mc \* P(c)cc

P(mc) = (5C1) \* (1/200)1 \* (199/200)4

where -

P(m) = 1/200

P(c) = 199/200

tc = mc + cc

tc = 5

mc = 1

cc = 4

(Short Way) -

P(mc) = 1 – (5C5 \* (199/200)5)

**= 0.02475124687812502**

OR (Alternate Way)

P(mc) = (5C1) \* (1/200)1 \* (199/200)4

= 5 \* 1/200 \* (199/200)4

= 5 /200 \* (0.995)4

= 0.025 \* (0.995)4

= 0.025 \* 0.98015

= **0.024503737515625**

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?

**Solution:**

**The most likely monetary outcome of the business venture is 2000**, which is the one with the highest percentage probability, when compared to others i.e., 0.3 or 30%

1. Is the venture likely to be successful? Explain

**Solution:**

The probability of the venture being successful is when the profit is > 0 i.e.,

P(success) = P(x = 1000) + P(x = 2000) + P(x = 3000)

= 0.2 + 0.3 + 0.1

= 0.6 or 60%

So, there is a 60% chance of the venture making a profit, which is a good percentage

When we calculate the total values for all probabilities, it shows as –

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

= -200 + -100 + 0 + 200 + 600 + 300

= 800

The total earning of the venture is 800 and the highest probability of earning is 2000, so it is profitable.

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

**Solution:**

The long-term average of the business venture is the sum of all probabilities and their respective values i.e.,

|  |  |  |
| --- | --- | --- |
| **x** | **P(x)** | **x \* P(x)** |
| -2,000 | 0.1 | -200 |
| -1,000 | 0.1 | -100 |
| 0 | 0.2 | 0 |
| 1000 | 0.2 | 200 |
| 2000 | 0.3 | 600 |
| 3000 | 0.1 | 300 |
| **Total** | | **800** |

P(x = -2000) \* -2000 + P(x = -1000) \* -1000 + P(x = 0) \* 0 + P(x = 1000) \* 1000 + P(x = 2000) \* 2000 + P(x = 3000) \* 3000

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

= -200 + -100 + 0 + 200 + 600 + 300

= 800

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

**Solution:**

The good measure of the risk involved in a venture of this kind depends on the variance and standard deviation of the distribution i.e., the higher the variance and standard deviation, the more the chances of risk.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **x** | **P(x)** | **x \* P(x)** | **x^2** | **(x^2) \* P(x)** |
| -2,000 | 0.1 | -200 | 40,00,000 | 400000 |
| -1,000 | 0.1 | -100 | 10,00,000 | 100000 |
| 0 | 0.2 | 0 | 0 | 0 |
| 1000 | 0.2 | 200 | 10,00,000 | 200000 |
| 2000 | 0.3 | 600 | 40,00,000 | 1200000 |
| 3000 | 0.1 | 300 | 90,00,000 | 900000 |
|  |  | **800** |  | **2800000** |

V(X) = E(X^2) –(E(X)) ^ 2 = 2800000 – 800 ^ 2 = 2160000

SD(X) = √2160000 = 1469.694