|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Integer (Discrete) |
| Results of rolling a dice | Integer (Discrete) |
| Weight of a person | Float (Continuous) |
| Weight of Gold | Float (Continuous) |
| Distance between two places | Float (Continuous) |
| Length of a leaf | Float (Continuous) |
| Dog's weight | Float (Continuous) |
| Blue Color | Boolean (Discrete) |
| Number of kids | Integer (Discrete) |
| Number of tickets in Indian railways | Integer (Discrete) |
| Number of times married | Integer (Discrete) |
| Gender (Male or Female) | Boolean (Discrete) |

Basic Statistics Level 1

**Q1) Identify the Data type for the Following:**

**Q2) Identify the Data types, which were among the following**

Nominal, Ordinal, Interval, Ratio.

|  |  |
| --- | --- |
| Data | Data Type |
| Gender | Nominal |
| High School Class Ranking | Ordinal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ordinal |

**Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?**

Possible outcomes are: HHH, HHT, HTH, HTT, THH, THT, TTH, TTT

No. of events when two heads and one tail are obtained: 3 (HHT, HTH, THH)

Therefore,

P(two heads and one tail) = 3/8

**Q4) Two Dice are rolled, find the probability that sum is**

1. **Equal to 1**
2. **Less than or equal to 4**
3. **Sum is divisible by 2 and 3**

Total possible outcomes are = 6^2 = 36

Possible outcomes: 2,3,4,5,6,7,8,9,10,11,12

1. Sum equals to 1 is not a possibility

Therefore, P (sum equal to 1) = 0/36 = 0

1. P (Sum less than or equal to 4) = P (sum equal to 2) +P (sum equal to 3) +

P (sum equal to 4)

= 1/36 + 2/36 + 3/36

= 6/36 = 1/6

1. P (sum is divisible by 2 and 3) = P (sum is divisible by 6)

= P (sum is 6 or 12)

= 5/36 + 1/36 = 6/36 = 1/6

**Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?**

So, there are a total of 7 balls in the bag. We have to draw 2 balls at random.

Number of ways to draw 2 balls from the bag =

Number of ways to draw 2 non blue balls from the bag =

P (drawing 2 non blue balls) = Number of ways to draw 2 non blue balls from the bag / Number of ways to draw 2 balls from the bag

= 10 / 21

**Q6) Calculate the Expected number of candies for a randomly selected child**

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

(I did not understand the question)

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

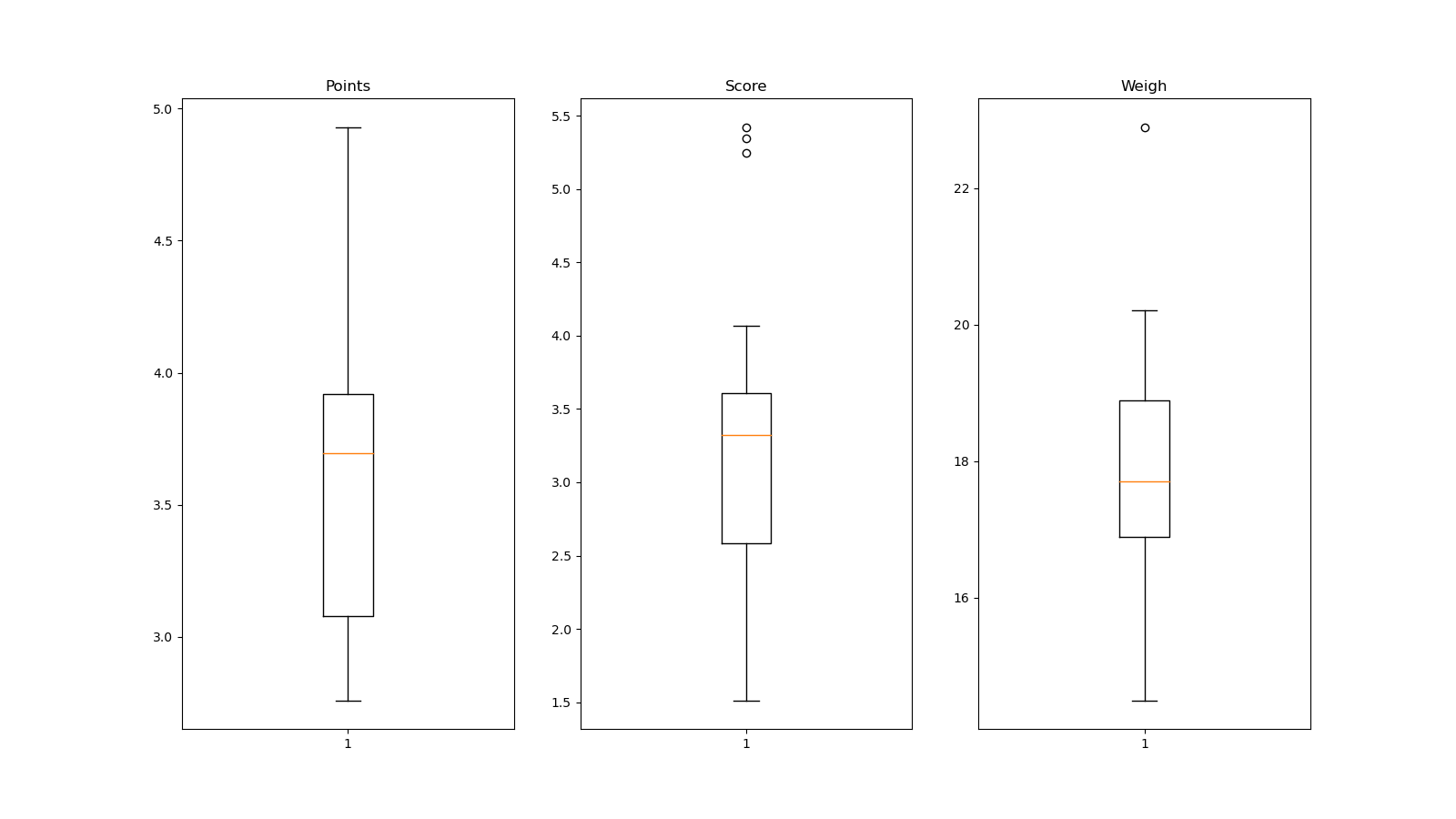
* For Points,Score,Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

Answer:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.596563 | 3.217250 | 17.848750 |
| Median | 3.695 | 3.325 | 17.710 |
| Mode | 3.07 | 3.44 | 17.02 |
| Variance | 0.28588 | 0.957378 | 3.1931 |
| Standard Deviation | 0.5346 | 0.97845 | 1.7869 |
| Range | 2.17 | 3.911 | 8.399 |



Q8) Calculate Expected Value for the problem below

1. The weights (X) of patients at a clinic (in pounds), are

108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

Answer:

Expected value is = sum of all observations / total number of observations

= 138.625

Therefore, if one patients is chosen at random then the expected weight is 138.625 kgs

**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Use Q9\_a.csv**

**Answer:**

|  |  |  |
| --- | --- | --- |
|  | Speed | Distance |
| Skewness | -0.11395477012828319 | 0.7824835173114966 |
| Kurtosis | -0.5771474239437371 | 0.24801865717051808 |

Here's how to interpret skewness values:

**Negative Skewness (Left Skewness):** If the skewness is less than zero, it indicates that the left tail of the distribution is longer or fatter than the right tail. In other words, the distribution is skewed to the left. The mean is typically less than the median in a negatively skewed distribution.

Mean < median

**Positive Skewness (Right Skewness):** If the skewness is greater than zero, it indicates that the right tail of the distribution is longer or fatter than the left tail. The distribution is skewed to the right, and the mean is typically greater than the median in a positively skewed distribution.

Mean > median

Zero Skewness: A skewness value of zero indicates a perfectly symmetrical distribution.

Here's a general interpretation of Kurtosis value:

**Positive Kurtosis:** A distribution with a positive kurtosis has heavier tails and a more pronounced peak than a normal distribution. It may indicate the presence of outliers or extreme values.

**Negative Kurtosis:** A distribution with negative kurtosis has lighter tails and a flatter peak than a normal distribution. It may suggest that the distribution has fewer outliers or is less prone to extreme values.

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | ST | WT |
| Skewness | 1.5814536794423764 | - 0.6033099322115126 |
| Kurtosis | 0.8194658792266849 | 0.8194658792266849 |

(Please check the attached. ipynb for detail)

**Q10) Draw inferences about the following boxplot & histogram**



Here, histogram peak has skew right and tail has a right skew. Here mean > median. And we can expect to have outliers on the maximum side of the data.



Here, we can see that, outliers does exists on the maximum side of the data.

**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

Answer: We will calculate confidence interval using stats.t.interval function.

Here, 94% confidence interval = (198.7376089443071, 201.2623910556929)

98% confidence interval = (198.4381860483216, 201.5618139516784)

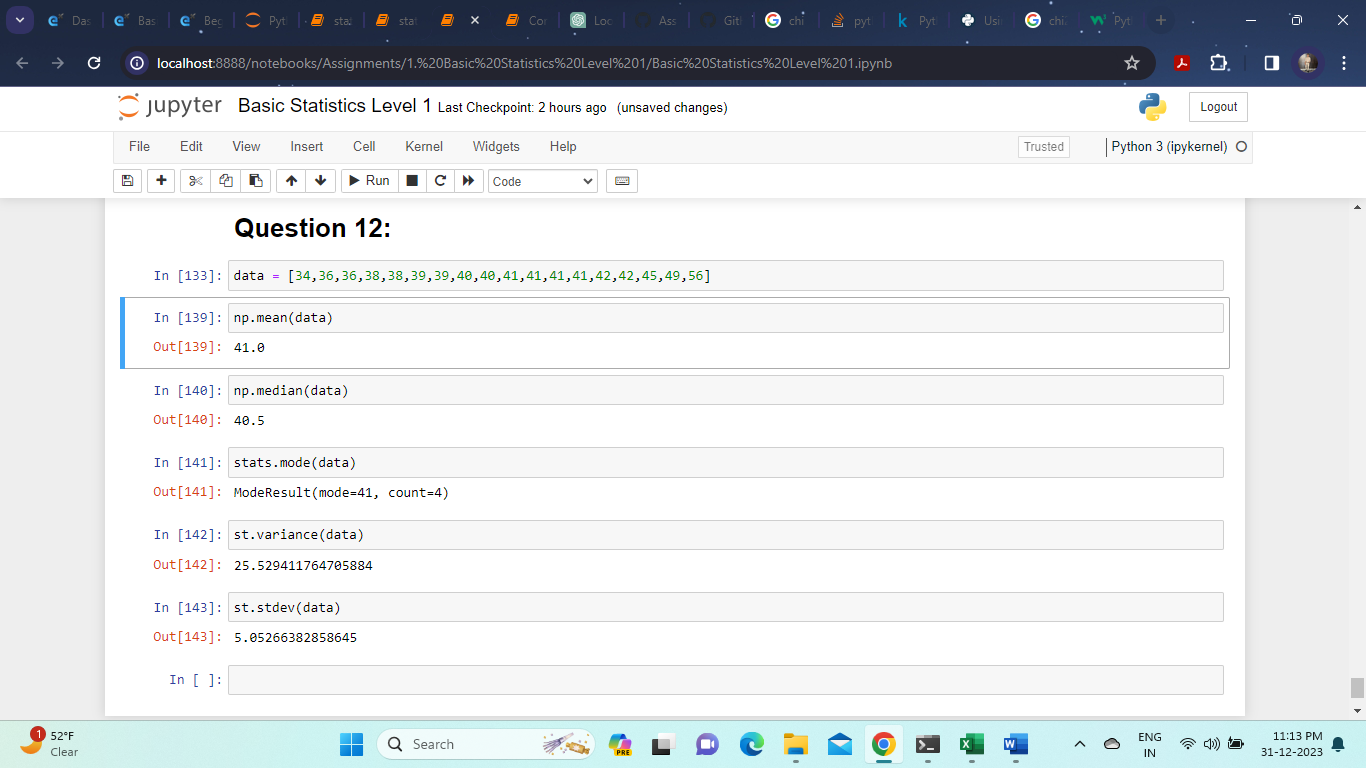
96% confidence interval = (198.6214037429732, 201.3785962570268)

**Q12)** Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean, median, variance, standard deviation.
2. What can we say about the student marks?

Answer:



Here, mean of the data is 41. And standard deviation is approx. 5.  
So 68% of the marks lies within (41-5, 41+5) = (36,46)  
and 95% of the marks lies within (41 – 2\*5, 41 + 2\*5) = (41 – 10, 41 + 10)   
 = (31, 51)  
and 99.7% of the marks lies within (41 – 3\*5, 41 +3\*5) = (41 – 15, 41 + 15)   
 = (26, 56)

Q13) What is the nature of skewness when mean, median of data are equal?

Ans: Skewness is 0. The data has symmetrical distribution.

Q14) What is the nature of skewness when mean > median ?

Ans: When mean > median, the data is right skewed. Or it has positive skewness.

Q15) What is the nature of skewness when median > mean?

Ans: When median > mean, the data is left skewed. Or it has negative skewness.

Q16) What does positive kurtosis value indicates for a data ?

Ans: distribution with a positive kurtosis has heavier tails and a more pronounced peak than a normal distribution. It may indicate the presence of outliers or extreme values.

Q17) What does negative kurtosis value indicates for a data?

Ans: A distribution with negative kurtosis has lighter tails and a flatter peak than a normal distribution. It may suggest that the distribution has fewer outliers or is less prone to extreme values.

Q18) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

Ans: The boxplot does not represent a normal distribution. Here the data is skewed. The median of the data is closer to the max of the data.

What is nature of skewness of the data?

Ans: The data is negatively skewed being skewed toward left.

What will be the IQR of the data (approximately)?

Ans: IQR = Q3 upper quantile – Q1 lower quantile = 18 – 10 = 8

Q19) Comment on the below Boxplot visualizations?



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

Ans: The median of both the datasets is equal, but the range of the data represented by boxplot 1 is quite low as compared to that of data represented by boxplot 2.

(Both the datasets seem to have no outliers, and has almost zero skewness).

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

Ans:

P (MPG > 38) = 0.348  
P (MPG < 40) = 0.729  
P (20 < MPG < 40) = 0.013

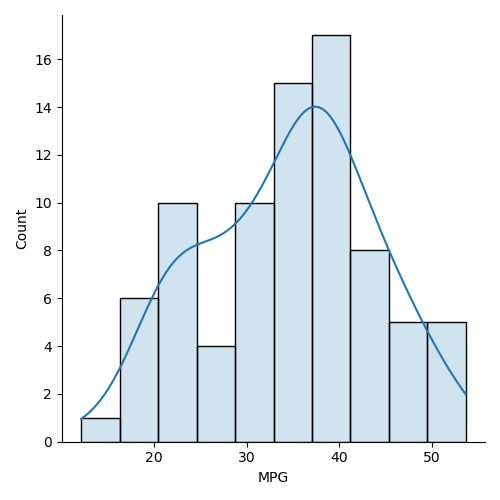
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

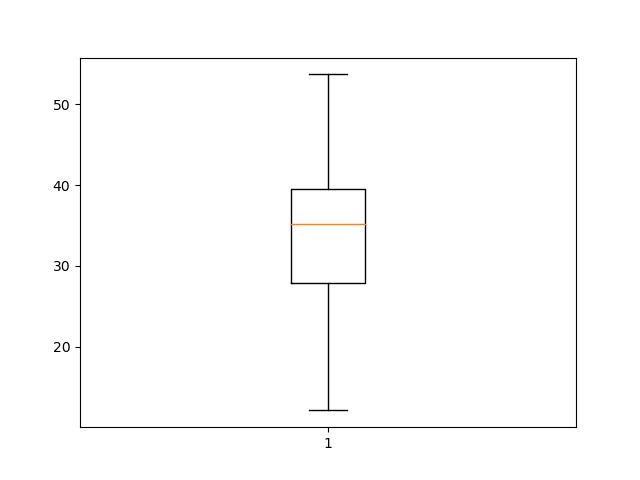
Dataset: Cars.csv

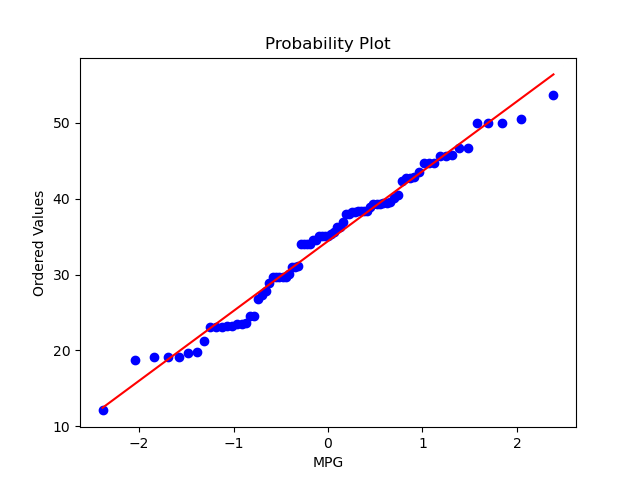
We can check for the normal distribution of the data using two ways

1. Visual: histogram (bell- shaped) and QQ-plot (straight line)
2. Descriptive statistics: skewness (=0) and kurtosis (=0)



This is not perfectly bell shaped, but it is quite close.





QQ -plot seems to align in a straight line. Quite close.

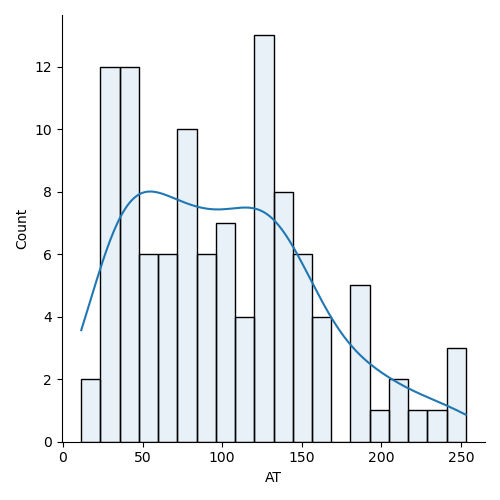
Skewness of the data: -0.17463433818755686

Kurtosis of the data: -0.6477383960245926

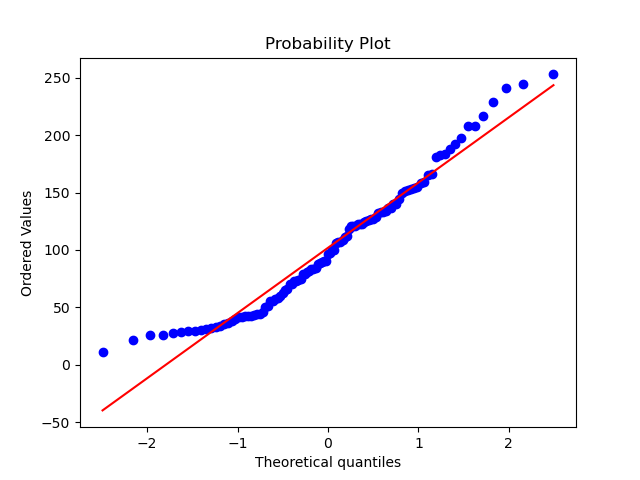
1. Check Whether the Adipose Tissue (AT) and Waist circumference (Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

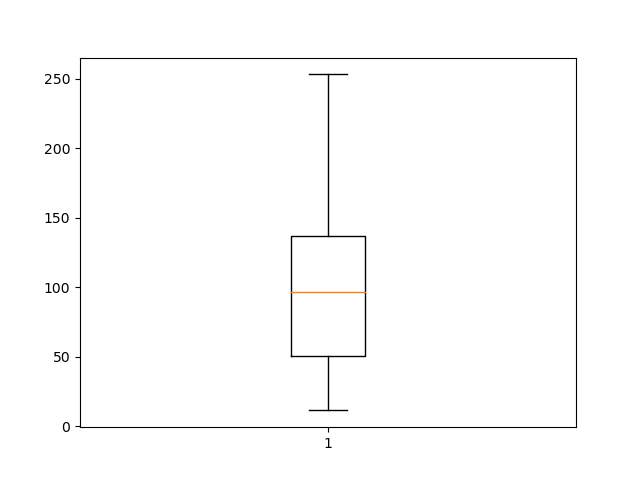
For Adipose Tissue:



The data is clearly positively skewed.



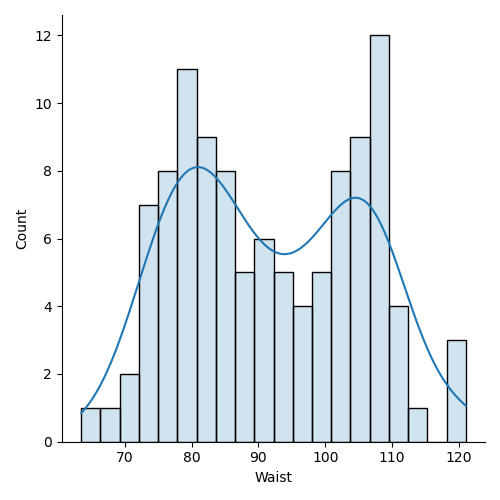
The QQ-plot doesn’t seem to align in the straight line.



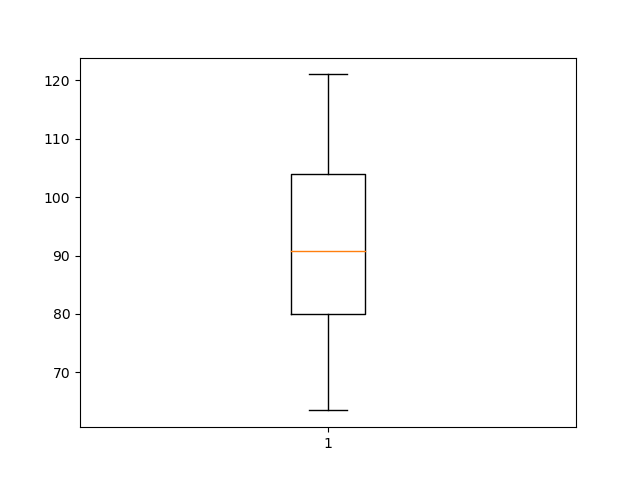
Skewness of the data: 0.5767896975987847

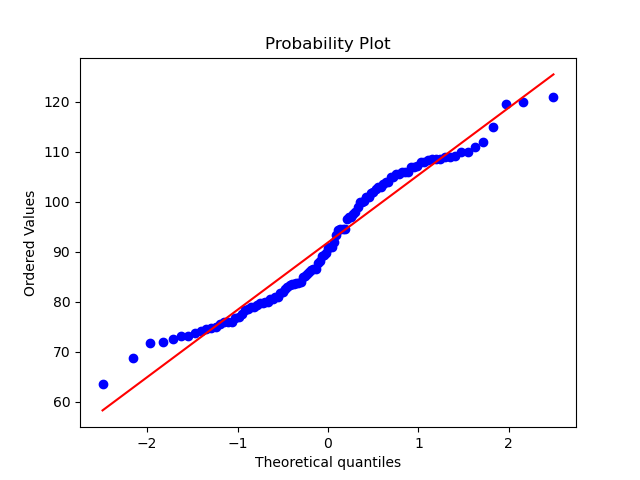
Kurtosis of the data: -0.3271884938021854

Waist circumference:



Clearly, this is not at all bell shaped.





Skewness of the data: 0.1322041763592883

Kurtosis of the data: -1.1072764806858817

So, here we can clearly say that the data is not normally distributed.

Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval.

Answer:

We calculate the z-score using the stats.norm.ppf function. The results are:  
90% confidence interval: 1.64  
94% confidence interval: 1.88  
60% confidence interval: 0.84

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25.

Answer:

We calculated the t-score using the stats.norm.ppf function. So, here are the results of the calculations:

T-score for a 95% confidence interval with 25 sample size: 2.064  
T-score for a 96% confidence interval with 25 sample size: 2.172  
T-score for a 99% confidence interval with 25 sample size: 2.797

Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

So, we used stats.t.cdf(t\_score, df) and we got 0.32. Which means that there is a probability of 32%.