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

**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 | Nominal |
| 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?**

ANS- Three coins are tossed,

Total number of Sample possible combinations = 23 = 8

Sample= {HHH, HHT, HTH, THH, TTH, THT, HTT, TTT}

Number of combinations that have two heads and one tail = 3, (HHT, HTH, TTH)

The probability of two heads and one tail when three coins are tossed simultaneously are

P (Two heads and One tail) = Two heads and One tail /Number of desired outcomes

= ⅜ or 0.375

**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

ANS- for Two Dice total number of outcome -36

a) Equal to 1: There is only one way to get a sum of 1: (1, 1).

1. Probability = (Equal to 1) / (Total Number of Outcomes)

= 1 / 36

1. Less than or equal to 4:

The possible combinations for sums less than or equal to 4 are: (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1).

Probability = (Less than or equal to 4) / (Total Number of Outcomes)

= 6 / 36 = 1 / 6

1. P (Sum is divisible by 2 and 3) = N (Event (Sum is divisible by 2 and 3)) / N(Event (Two dice rolled)) = 6 / 36 = 1/6 = 0.16 = 16.66%

**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?**

ANS- Total number of balls = (2 + 3 + 2) = 7  
Let S be the sample space.  
Then, n(S) = Number of drawing 2 balls out of 7  
=7C2​  
= (2×1)(7×6)​  
=21  
Let E = drawing 2 balls, none of which is blue.  
∴n(E)= Number of drawing 2 balls out of (2 + 3) balls.  
=5C2​  
= (2×1)(5×4)​  
=10  
P(E)=n(S)n(E)​=2110​

**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

ANS = Expected Value = Σ (Value \* Probability)

Expected Value = (1 \* 0.015) + (4 \* 0.20) + (3 \* 0.65) + (5 \* 0.005) + (6 \* 0.01) + (2 \* 0.120)

= 0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.24

= 3.10

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

**For Points, Score, Weight**

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

Use Q7.csv file

ANS

Mean for Points = 3.59, Score = 3.21 and Weigh = 17.84

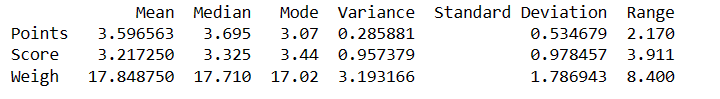
Median for Points = 3.69, Score = 3.32 and Weigh = 17.71

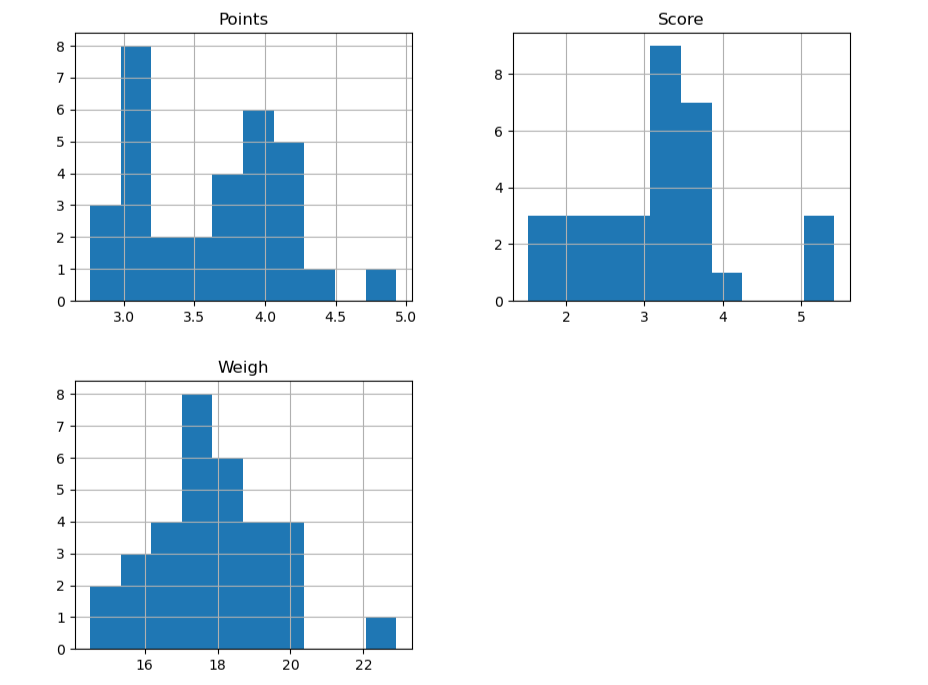
Mode for Points = 3.07, Score = 3.44 and Weigh = 17.02

Variance for Points = 0.28, Score = 0.95, Weigh = 3.19

Standard Deviation for Points = 0.53, Score = 0.97, Weigh = 1.78

Range [Min-Max] for Points [3.59 – 4.93], Score [3.21 – 5.42] and Weigh [17.84 – 22.9]





**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**

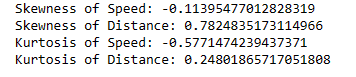
**Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?**

ANS: weights of patients =108, 110, 123, 134, 135, 145, 167, 187, 199

Expected value = Sum (X \* Probability of X) = (1/9)(108)+ (1/9)(110)+ (1/9)(123)+ (1/9)(134)+ (1/9)(145)+ (1/9)(167)+ (1/9)(187)+ (1/9)(199) = 145.33

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

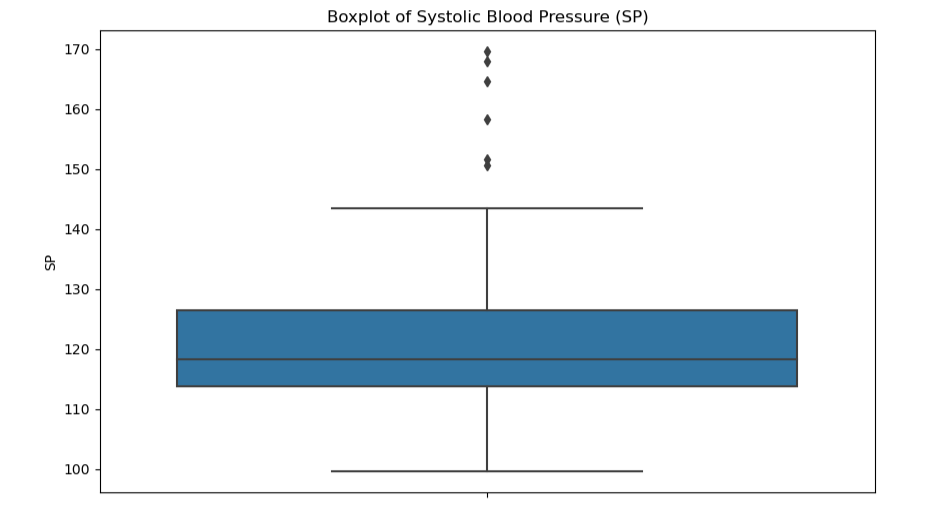
**Cars speed and distance Use Q9\_a.csv**

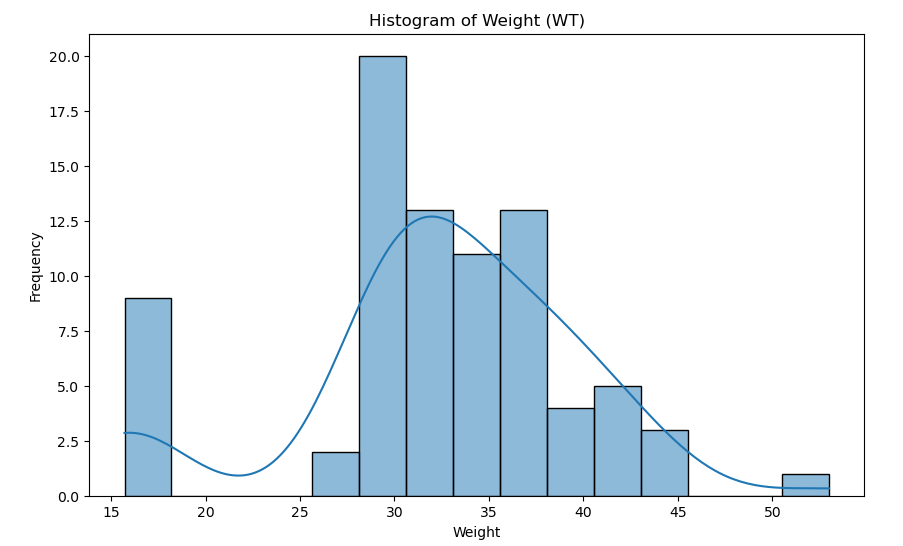


**SP and Weight(WT)**

**Use Q9\_b.csv**

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





**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?**

conf\_94 =stats.t.interval(alpha = 0.94, df=1999, loc=200, scale=30/np.sqrt(2000))

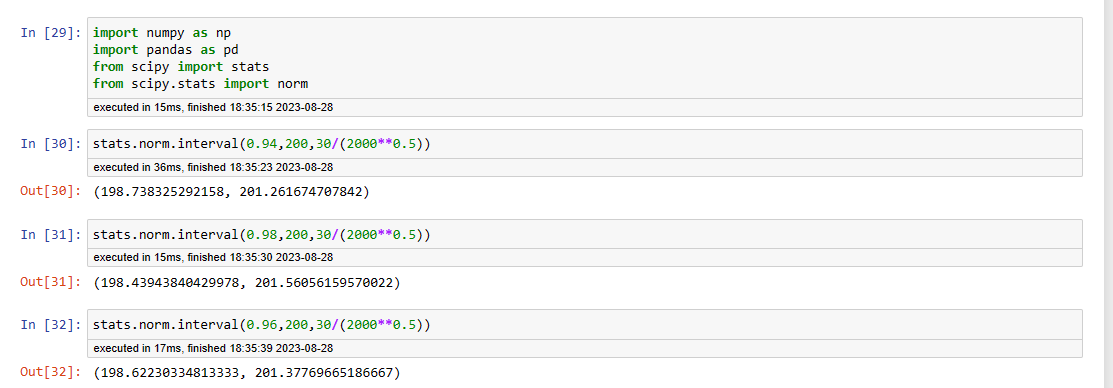
print(np.round(conf\_94,0))

print(conf\_94)

For 94% confidence interval Range is [ 198.73 – 201.26]

For 98% confidence interval range is [198.43 – 201.56]

For 96% confidence interval range is [198.62 – 201.37]



**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.**

Mean =41, Median =40.5, Variance =25.52 and Standard Deviation =5.05

1. **What can we say about the student marks?**

Don’t have outliers and the data is slightly skewed towards right because mean is greater than median



The average score is around 41.0.

The middle value is 40.5, which is close to the mean.

Variance and Standard Deviation: The variance and standard deviation are high.

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

ANS : When the mean and median of a dataset are equal, it indicates that the data has a symmetric distribution. The mean is at the center of the distribution. The median is also at the center of the distribution. The values on one side of the center are similar to the values on the other side.

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

ANS : When the mean is greater than the median, the distribution of the data is right-skewed or positively skewed.

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

ANS : When the median is greater than the mean, the distribution is said to be negatively skewed. This means that the tail of the distribution is stretched out to the left.

**Q16) what does positive kurtosis value indicates for a data?**

ANS : Data with positive kurtosis have more outliers than a normal distribution.

**Q17) what does negative kurtosis value indicates for a data?**

ANS : Negative kurtosis value indicates that wider peak and thinner tails.

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



What can we say about the distribution of the data?

ANS -Not normally distributed

What is nature of skewness of the data?

ANS- -Negative skewness

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

ANS- -10-18

**Q19) Comment on the below Boxplot visualizations?**



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

ANS. Both Boxplot 1 and Boxplot 2 have a similar interquartile range Boxplot 1 have a narrow range between the minimum and maximum values compared to Boxplot 2. & middle values of the data are similar in both plots

**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)
  4. ANS. P(MPG>38)

Ans: a. P(MPG>38)= np.round(1 - stats.norm.cdf(38, loc= q20.MPG.mean(), scale= q20.MPG.std()),3)

print('P(MPG>38)=',Prob\_MPG\_greater\_than\_38)

P(MPG>38)= 0.348

1. P(MPG<40)= np.round(stats.norm.cdf(40, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('P(MPG<40)=',prob\_MPG\_less\_than\_40)

P(MPG<40)= 0.729

1. P (20<MPG<50)

P (20<MPG )= np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('p(MPG>20)=',(prob\_MPG\_greater\_than\_20))

p(MPG>20)= 0.943

prob\_MPG\_less\_than\_50 = np.round(stats.norm.cdf(50, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

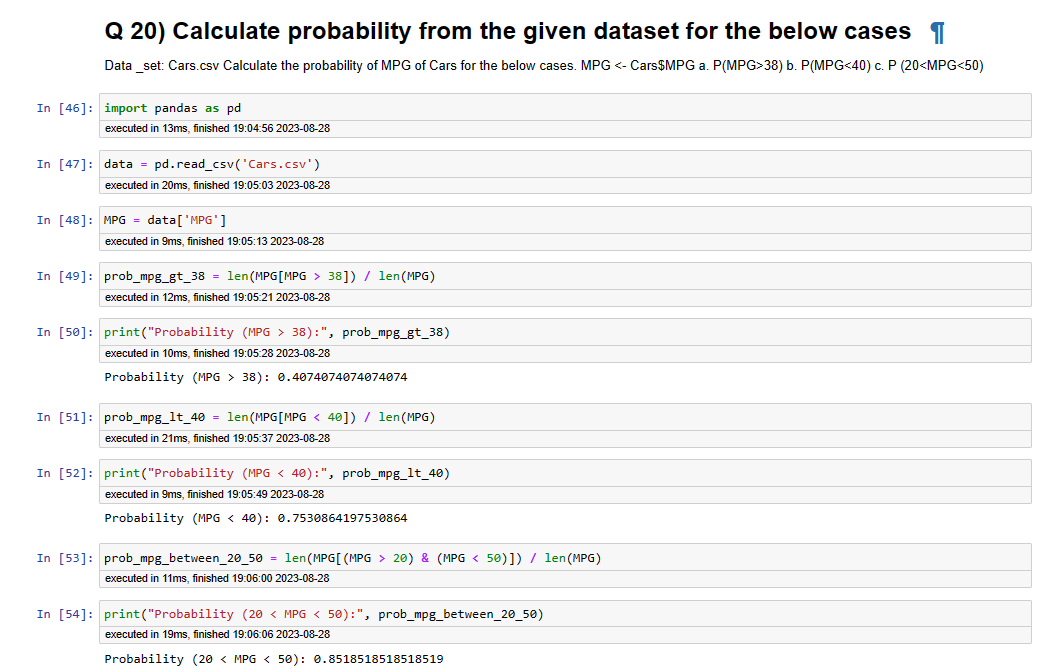
print('P(MPG<50)=',(prob\_MPG\_less\_than\_50))

P(MPG<50)= 0.956

P (20<MPG<50)= (prob\_MPG\_less\_than\_50) - (prob\_MPG\_greater\_than\_20)

print('P(20<MPG<50)=',(prob\_MPG\_greaterthan20\_and\_lessthan50))

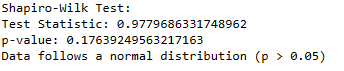
P(20<MPG<50)= 0.013000000000000012



**Q 21) Check whether the data follows normal distribution**

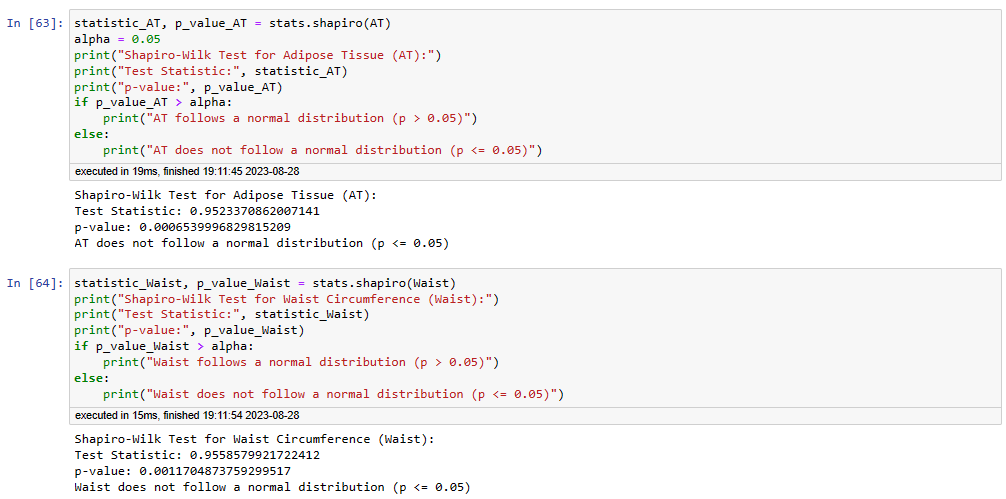
1. **Check whether the MPG of Cars follows Normal Distribution**

**Dataset: Cars.csv**

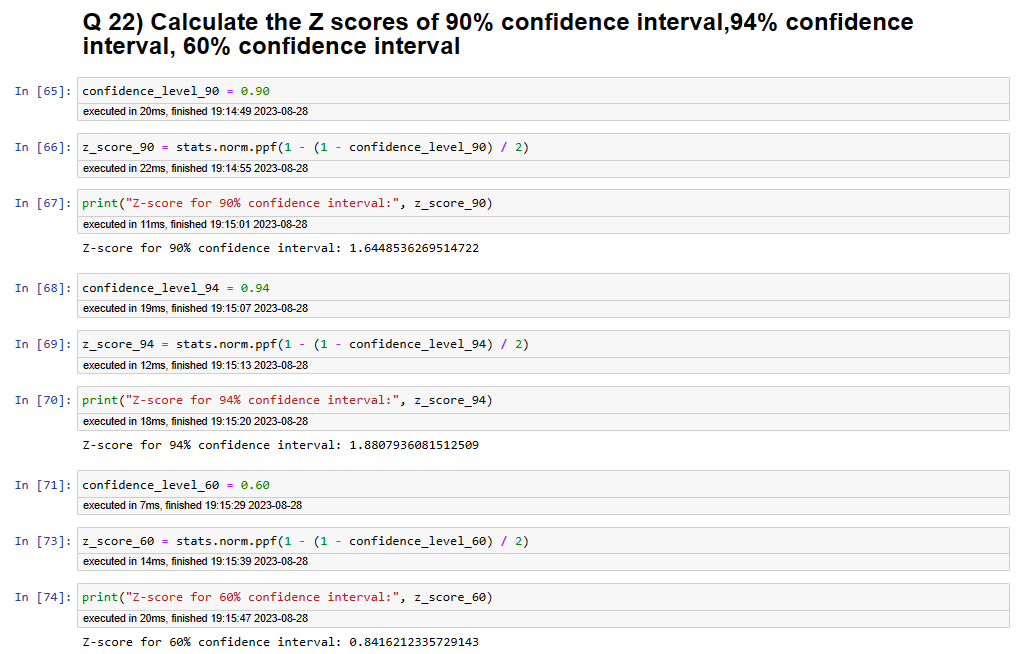


1. **Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution**

**Dataset: wc-at.csv**



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



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



**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**

ANS

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

Ans: import numpy as np

Import scipy as stats

t\_score = (x - pop mean) / (sample standard daviation / square root of sample size)

(260-270)/90/np.sqrt(18))

t\_score = -0.471

stats.t.cdf(t\_score, df = 17)

0.32 = 32%