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

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 | ratio |
| Weight | ratio |
| Hair Color | ratio |
| Socioeconomic Status | Nominal |
| Fahrenheit Temperature | ratio |
| Height | ratio |
| Type of living accommodation | ordinal |
| Level of Agreement | ordinal |
| IQ(Intelligence Scale) | ordinal |
| Sales Figures | interval |
| Blood Group | Nominal |
| Time Of Day | interval |
| Time on a Clock with Hands | interval |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | ratio |
| SAT Scores | ordinal |
| Years of Education | interval |

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

--- **23 =8**

**Possible combinations are [HHH, HHT, HTH, THH, TTH, THT, HTT, TTT]**

**The number of combinations which have two heads and one tail are:**

**HHT, HTH, TTH which makes them 3 in number.**

**P(Two heads and One Tail) = 3 =3/8= 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

------- **total possible outcome =62=36**

1. **Probability is equal to 1 is**

**Sum will always exceed 1 so**

**P= 0/36 = 0**

1. **Probability Less than or equal to 4**

**(1,3)(3,1)(2,2)(1,2)(2,1)(1,1)**

**P = 6/36 = 0.1666**

1. **Probability Sum is divisible by 2 and 3**

**(6,6)(3,3)**

**P = 2/36 =0.05555**

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?

---------------------------------------------------------------

**Total ball is 2 yellow +3 green +2 blue= 7 ball.  
It can be written as 7C2,  
So None of the ball is blue that is 7- 2 (blue ball) is = 5 ball.Or  5C2  
So 5C2/7C2 = (5!/3!\*2!) // (7!/5!\*2!) = 10/21**

**P =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

-----------------

**Expected number of candies = (1\**0.015) + (4\**0.20) + (3\**0.65) + (5\**0.005) + (6\**0.01) + (2\**0.120) = 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24 = 3.085**

**E(x) = 3085**

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**

**------------------------------------------------------------------------------**

**---For points**

**Mean – 3.596**

**Median— 3.695**

**Mode—3.92**

**Var—0.285**

**Standard var-- 0.5346787**

**---- For score**

**Mean –3.217**

**Median—3.325**

**Mode—3.44**

**Var –0.9573**

**Standard var-- 0.9784574**

**---For weight**

**Mean –17.848**

**Median----17.71**

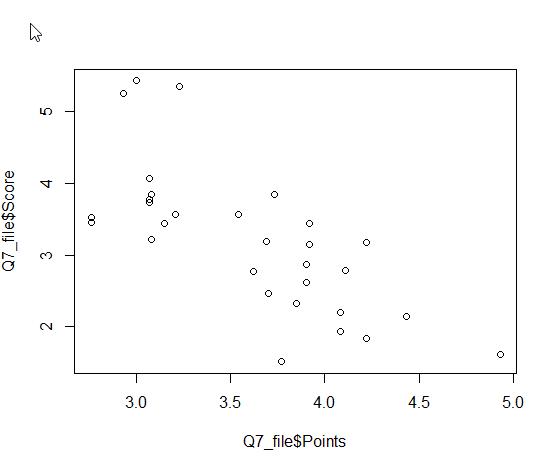
**Mode---17.02**

**Var---3.193**

**Standard var ----1.786943**

**Score increasing as point are decreasing as per observation from graph**

**Both are somewhat inversely proportional to each other**

****

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?

-----------------------------------------------------------------------------

**Expected Value  =  ∑ ( probability  \* Value )**

**∑ P(x)\*E(x)**

**there are 9 patients**

**Probability of selecting each patient = 1/9**

**Ex = 108, 110, 123, 134, 135, 145, 167, 187, 199**

**P(x) = 1/9  1/9   1/9  1/9   1/9   1/9   1/9   1/9  1/9**

**Expected Value  =  (1/9)(108) + (1/9)110  + (1/9)123 + (1/9)134 + (1/9)135 + (1/9)145 + (1/9(167) + (1/9)187 + (1/9)199**

**= (1/9) ( 108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)**

**= (1/9) \* (  1308)**

**= 145.33**

**Expected Value of the Weight of that patient = 145.33**

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

Cars speed and distance

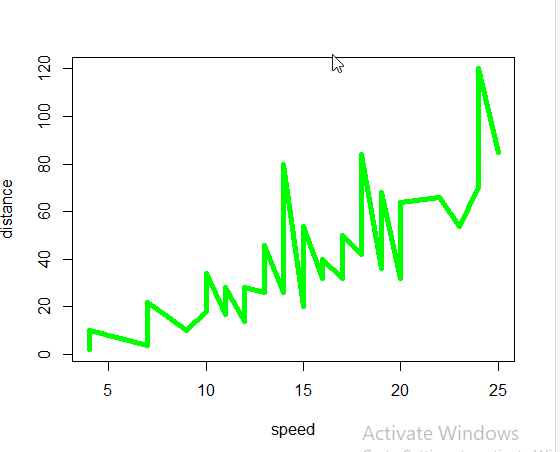
Use Q9\_a.csv

**Skewness for Speed = -0.113**

**Skewness for distance = 0.7824**

**Kurtois for Speed = 2.422**

**Kurtosis for distance = 3.248**

****

**SP and Weight(WT)**

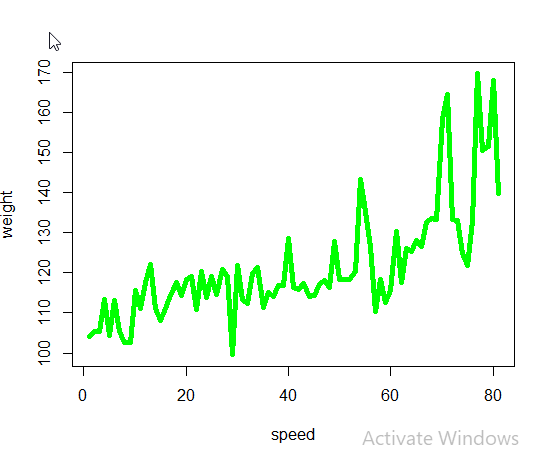
**Use Q9\_b.csv**

**Skewness for SP = 1.581**

**Skewness for WT=-0.6033**

**Kurtois for SP = 5.723**

**Kurtosis for WT= 3.819**

****

Q10) Draw inferences about the following boxplot & histogram



----------------------------------------

**Histogram is right skewed also known as positive skewed means The mean is greater than the median**

**100 chickweight has hightest frequency and 25 is lowest.**



**Median is not centered. And Upper quertile (3rd) is bigger than 1st Quertile**

**There are few ouliers at top**

**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?

-----------

**From T distribution table For a 94% confidence interval, the critical value is 1.88, for a 98% confidence interval, the critical value is 2.33, and for a 96% confidence interval, the critical value is 1.96. The degrees of freedom in this case are 1,999 (n-1).**

**Using these values, we can calculate the confidence intervals as follows:**

**94% confidence interval:**

**Margin of error = 1.88 \* (30 / sqrt(2000)) = 1.26**

**Confidence interval = 200 ± 1.26 = [198.74, 201.26]**

**98% confidence interval:**

**Margin of error = 2.33 \* (30 / sqrt(2000)) = 1.56**

**Confidence interval = 200 ± 1.56 = [198.44, 201.56]**

**96% confidence interval:**

**Margin of error = 1.96 \* (30 / sqrt(2000)) = 1.31**

**Confidence interval = 200 ± 1.31 = [198.69, 201.31]**

**Therefore, we can say with 94% confidence that the average weight of an adult male in Mexico is between 198.74 and 201.26 pounds, with 98% confidence that it is between 198.44 and 201.56 pounds, and with 96% confidence that it is between 198.44 and 201.56 pounds.**

**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?

---------------------------

**Mean = 34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56/18**

**= 41**

**Meadian = 40+41/2 = 40.5**

**Variance = Variance (σ2) is the average of squared difference from mean**.

**= (34-41)2+(36-41)2+(36-41)2+(38-41)2+(38-41)2+(39-41)2+(39-41)2+(40-41)2+(40-41)2+(41-41)2+(41-41)2+(41-41)2+(41-41)2+(42-41)2+(42-41)2+(45-41)2+(49-41)2+(56-41)2/18**

**=49+25+25+9+9+4+4+1+1+0+0+0+0+1+1+16+64+225/18**

**=25.52**

**SD = square root of 25.52 =5.05**

**We can say about student mark is range is almost common (i.e 2-3 most of the time)**

**And there one outlier which is 55.**

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

---------**Normal distribution**

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

------------------ **positive skewness**

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

---------------**negavtive skewness**

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

-------------- **Positive values of kurtosis indicate that distribution is peaked and possesses thick tails**

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

-------- **A distribution with a negative kurtosis value indicates that the distribution has lighter tails than the normal distribution.**

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



What can we say about the distribution of the data?

* **There is more data in first Quertile(lower Quertile) . And median is not centered**

What is nature of skewness of the data?

* **Left skewed**

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

* **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.

---------------------

**Box plot 2 have more data or we can say widely spread data than boxplot 1**

**Boxplot 2 have centered median. So we can say its normal distribution**

**Boxplot 1 median is not centered. And its 3rd quartile is bigger.**

**Both of them don’t have any outliers**

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

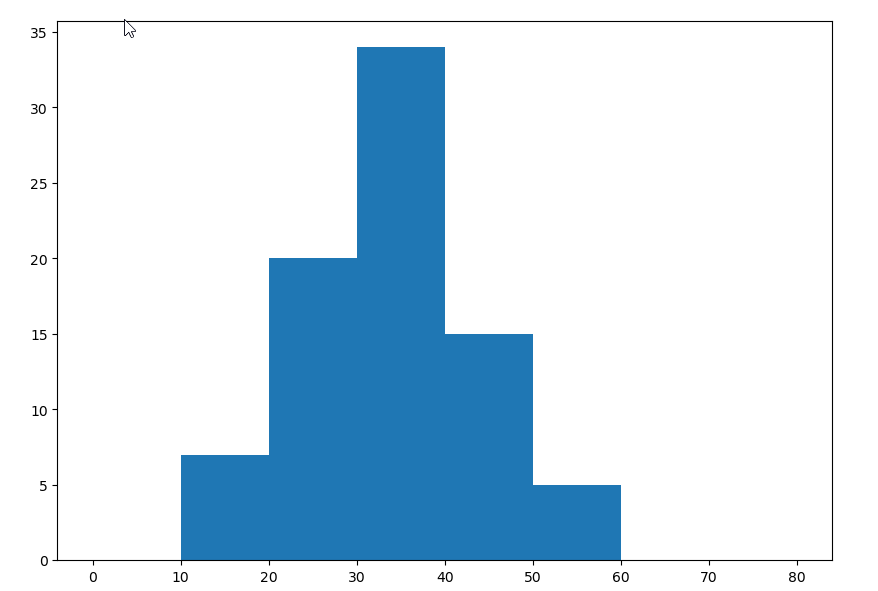
* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

---------- **Q 20) .ipynb**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

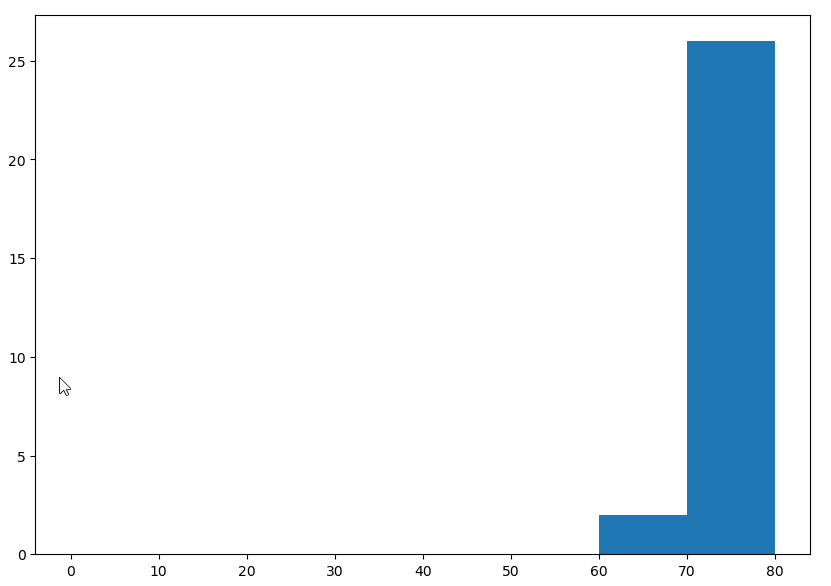
Dataset: Cars.csv



**Yes, its close to the normal distribution**

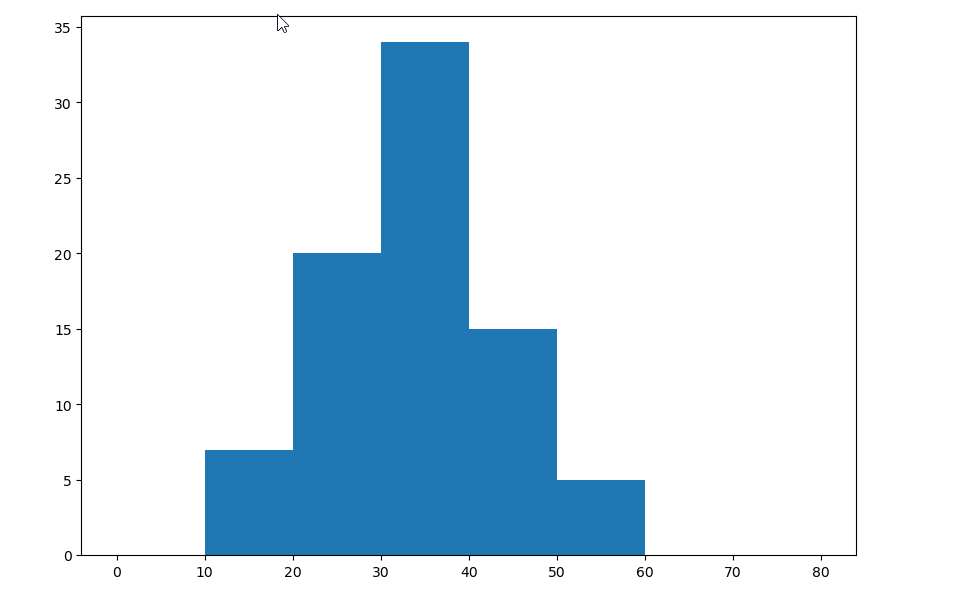
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv



**Waist Circumference(Waist) histogram**

**It is left skewed (Negative). Its not following normal distribution**



**Adipose Tissue (AT) histogram:**

**Its normal distribution**

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

----------------

**90% confidence interval: Z = 1.645**

**94% confidence interval: Z = 1.880**

**60% confidence interval: Z = 0.253**

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

--------------------

**Df = 25- 1 = 24**

**For a 95% confidence interval and df=24, the t score is 1.71.**

**For a 96% confidence interval and df=24, the t score is 1.82.**

**For a 99% confidence interval and df=24, the t score is 2.49.**

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

------- Ans. **Q24 (Basic Statistics Level-1).ipynb**

n = 18

SD = 90

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom