|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | **Discrete** |
| Results of rolling a dice | **Discrete** |
| Weight of a person | **Continues** |
| Weight of Gold | **Continues** |
| Distance between two places | **Continues** |
| Length of a leaf | **Continues** |
| Dog's weight | **Continues** |
| 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 | **Nominal** |
| Level of Agreement | **Ordinal** |
| IQ(Intelligence Scale) | **Interval** |
| Sales Figures | **Ratio** |
| Blood Group | **Nominal** |
| Time Of Day | **Interval** |
| Time on a Clock with Hands | **Interval** |
| Number of Children | **Ratio** |
| Religious Preference | **Nominal** |
| Barometer Pressure | **Interval** |
| SAT Scores | **Interval** |
| Years of Education | **Nominal** |

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

**The probability of getting two heads and one tail on tossing three coins at once is equal to 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

**Sol:**

1. If two dices were rolled, then total possible cases =36

Total Favourable cases (Having sum =1) = 0

As minimum sum is 2 for outcome (1,1)

Hence, probability is 0

1. Equal to 4 =(1, 3)(2,2)(3,1) = 3/36 = 1/12
2. divisible by 2 is 18/36 + divisible by 3 is 12/36 = 5/6

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?

**Sol: 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

Sol: **Expected number of candies for a randomly selected child**

**=  1 \* 0.015  + 4\*0.20  + 3 \*0.65  + 5\*0.005  + 6 \*0.01  + 2 \* 0.12**

**= 0.015 + 0.8  + 1.95 + 0.025 + 0.06 + 0.24**

**=  3.09**

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**

**Sol:** By using **df.describe() in Python using pandas**  
 Mean Of Point = 115.09/32 = 3.6

Mean Of Score = 102.952/32 = 3.21

Mean Of Weigh = 571.16/32 = 17.84

**By using : print("\n----------- Calculate Median -----------\n")**

**print(df.median())**

Points 3.695

Score 3.325

Weigh 17.710

dtype: float64

**By using df.std()**

Points 0.534679

Score 0.978457

Weigh 1.786943

dtype: float64

**By using df.var()**

Points 0.285881

Score 0.957379

Weigh 3.193166

dtype: float64

**Min Range Max Range**

**Points : 2.76 4.93**

**Score : 1.51 5.42**

**Weigh : 14.5 22.9**

1. **Mean and median are approx. equal, so there is no outliers on this data set.**
2. **The dataset is normally distributed because there is no outliers.**
3. **Skewness is zero because dataset is normally distributed.**

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?

**Sol: 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**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Sol:** By using df.skew() By using df.kurtosis()

speed -0.117510 speed -0.508994

dist 0.806895 dist 0.405053

Inferences:

1. Varible “speed” is negative skewed data because the concentration of the data is left side.
2. Variable “distance” is positive skewed data because the concentration of the data is right side.
3. Variable “speed” is negative kurtosis of the data because the peakness of the data distribution is wide.
4. Variable “distance” is positive kurtosis of the data because the peakness of the curve is short.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Sol:** By using df.skew() By using df.kurtosis()

SP 1.611450 SP 2.977329

WT -0.61475 WT 0.950291

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



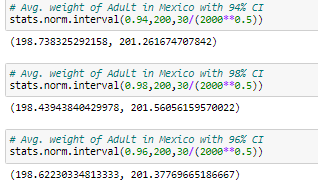
**Solution :**

**Inferences:**

1. **Histogram has positive outliers because there is positive skewness.**
2. **Histogram are only used for continuous data.**
3. **Box contains the first quartile median and the third quartile and the whiskers are the lines extending to the minimum and maximum.**
4. **Boxplot has outliers because there is positive skewness.**
5. **Boxplot is also know as whiskers plot.**

**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?

**Solution:**



**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?

**Solution:**

**Mean of Student Marks = 41.0**

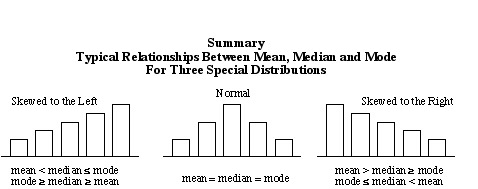
**Median = 40.5**

**Standard Deviation = 5.052664**

**Variance = 25.529411764705884**

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

Sol: **Zero Skewness**



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

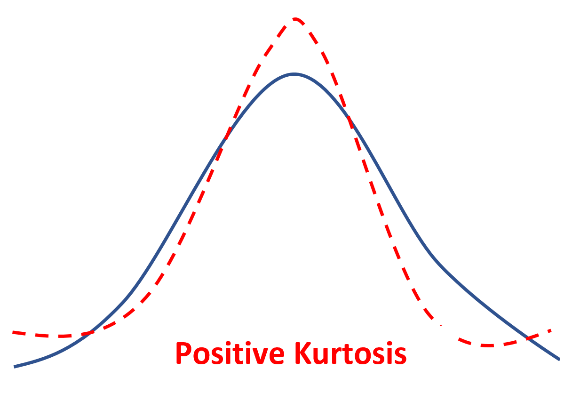
**Sol:** If the **mean** is greater than the **median**, the distribution is positively **skewed** means **Positive Skewness.**

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

**Sol:** If the **mean** is less than the **median**, the distribution is negatively **skewed** means **Negative Skewness**.

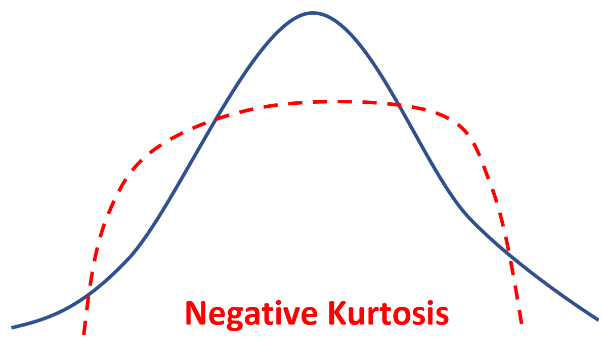
Q16) What does positive kurtosis value indicates for a data ?

**Sol:** Positive values of kurtosis indicate that a distribution is peaked and possess thick tails.



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

**Sol:** Negative values of kurtosis indicate that a distribution is flat and has thin tails.



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



What can we say about the distribution of the data?

**Solution : Distribution Is Not Normal According to the Figure.**

What is nature of skewness of the data?

**Solution : Distribution Is Positively Skewed According To The Figure.**

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

**Solution : IQR according to this data is located from Q3=18 and Q1=10.**

**So IQR = Q3 - Q1 = 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.

**Solution : a) Both Boxplot 1 & boxplot 2 are normally distributed.**

1. **Both Boxplot 1 & boxplot 2 are symmetric in nature.**
2. **Boxplot 1 having data from 230 to 285 and boxplot 2 having data from 100 to 310.**

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)

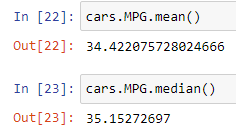
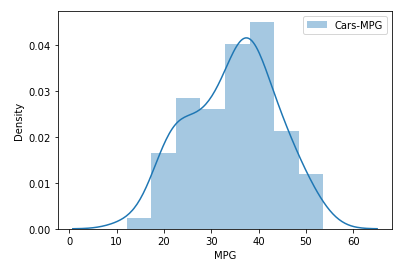
c. P (20<MPG<50)

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Sol :**

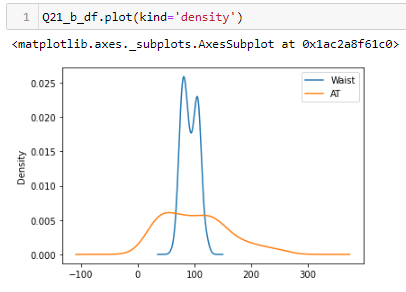
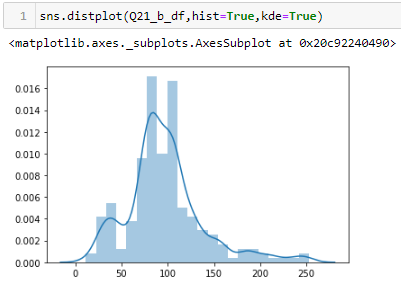


**Inference:** MPG of Cars does follow normal distribution approximately (as mean and median are approx. same)

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

Dataset: wc-at.csv

**Solution : By The Observed Plot We Can See That Data Set Is Not Following Normal Distribution.**

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

Sol: **formula using – AL=(1+CI)/2 & By Z-table**

**90% confidence interval, Z score = 1.65**

**94% confidence interval, Z score = 1.89**

**60% confidence interval, Z score = 0.84**

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

**Solution : For 95%=stats.t.ppf(0.975,24)= 1.9739**

**For 96%=stats.t.ppf(0.98,24)= 2.1715**

**For 99%=stats.t.ppf(0.995,24)=2.7969**

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

**Solution:**

**t - statistics for the data is given as follows:**

**https://tex.z-dn.net/?f=t%3D%5Cdfrac%7Bx-%5Cmu%7D%7B%5Cfrac%7Bs%7D%7B%5Csqrt%20n%7D%7D**

**x = mean of the sample of bulbs =  260**

**μ = population mean = 270**

**s = standard deviation of the sample = 90**

**n = number of items in the sample = 18**

**https://tex.z-dn.net/?f=t%3D%5Cdfrac%7B260-270%7D%7B%5Cfrac%7B90%7D%7B%5Csqrt%2018%7D%7D**

**https://tex.z-dn.net/?f=t%20%3D%20%5Cdfrac%7B-10%7D%7B%5Cfrac%7B90%7D%7B3%5Csqrt%202%7D%7D**

**https://tex.z-dn.net/?f=t%20%3D%20%5Cdfrac%7B-10%7D%7B%5Cfrac%7B30%7D%7B%5Csqrt%202%7D%7D**

**https://tex.z-dn.net/?f=t%20%3D%20%5Cdfrac%7B-1%20%5Ctimes%20%5Csqrt%202%7D%7B3%7D**

**t = - 0.471**

**degrees of freedom is n – 1 = 18-1= 17**

**probability that t < - 0.471 with 17 degrees of freedom assuming the population mean is true**

**t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of 0.3218 assuming the mean life of the bulbs is 300 days.**