|  |  |
| --- | --- |
| 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 | Categorical |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Categorical |
|  |  |

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) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ordinal |
| Years of Education | Ratio |

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

ANS:- When 3 coins are tossed there are total 8 outcomes.

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

Two heads and one tail are = {HHT, THH, HTH}

Probability = Number of favourable outcome

Total number of favourable outcome

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

ANS:- Total possible outcomes = 30

Outcomes = (1, 1) (1, 2) (1, 3) (1, 4) (1, 5) (1, 6)

(2, 1) (2, 2) (2, 3) (2, 4) (2, 5) (2, 6)

(3, 1) (3, 2) (3, 3) (3, 4) (3, 5) (3, 6)

(4, 1) (4, 2) (4, 3) (4, 4) (4, 5) (4, 6)

(5, 1) (5, 2) (5, 3) (5, 4) (5, 5) (5, 6)

(6, 1) (6, 2) (6, 3) (6, 4) (6, 5) (6, 6)

1. Equal to 1 = 0

Reason:-Since dice as the least number as 1 so the least sum is 2

1. Less than or equal to 4 = 6/36 = 1/6

Reason:-Number of outcomes are [(1,1),(1,2), (1,3),(2,1),(2,2),(3,1)]

1. Sum is divisible by 2and 3 = 6/36 = 1/6

Reason:- If sum is 6 or 12 then they will be divisible by 2 and 3.

Number of outcomes are [(1,5),(2,4), (3,3),(4,2),(5,1),(6,6)]

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?

ANS:- Total number of balls 2+3+2=7

Two balls are drawn at random so No of outcomes= 7C2 = 7\*6/(2\*1) = 21

No blue balls means only 5 balls left in that have to pick 2 so No of outcomes = 5C2 = 5\*4/(2\*1) =10

Probability of No balls are blue = 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

ANS:- p = (1\*0.015)+(4\*0.20)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.120) = 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**

**ANS:-**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **MEAN** | **MEDIAN** | **MODE** | **VERIENCE** | **S.D** | **RANGE** |
| **POINTS** | 3.597 | 3.695 | 3.07 ,  3.92 | 0.286 | 0.535 | 2.17 |
| **SCORE** | 3.217 | 3.325 | 3.44 | 0.957 | 0.978 | 3.911 |
| **WEIGH** | 17.849 | 17.710 | 17.02 ,  18.90 | 3.193 | 1.787 | 8.399 |

* All the calculations are done by using python pandas library.
* In ‘Points’ and ‘Weigh’ data sets there are 2 mode values because these sets have two most repeating values.
* In all three data sets mean ≠ median ≠ mode.
* Points and Score are negatively skewed because Median>mean.
* Weigh is positively skewed because mean>median.

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:-**

There are totally 9 patients so probability of getting each patient is = 1/9

Expected value of the patient = (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)

Expected value of the patient = 145.33 Pounds.

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

**Cars speed and distance**

**Use Q9\_a.csv**

**ANS:-**

|  |  |  |
| --- | --- | --- |
|  | **SKEWNESS** | **KURTOSIS** |
| **SPEED** | -0.118 | -0.509 |
| **DISTANCE** | 0.807 | 0.405 |

* Here Speed is **negatively skewed** and Distance is **positively skewed.**
* Since Speed skewness is between -0.5 to +0.5 we can say data are **fairly symmetrical.**
* Since Distance skewness is between +0.5 to 1 we can say data are **moderately skewed.**
* Speed and Distance is **Platykrutic** since kurtosis <3.

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | **SKEWNESS** | **KURTOSIS** |
| **SP** | 1.611 | 2.977 |
| **WEIGHT** | -0.615 | 0.950 |

* Here Weight is **negatively skewed** and SP is **positively skewed.**
* Since SP skewness is greater than +1 we can say data are **highly skewed.**
* Since Weight skewness is between -0.5 to -1 we can say data are **moderately skewed.**
* SP and Weight is **Platykrutic** since kurtosis <3.

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



ANS:-

* This is positively skewed because tail is longer on the right side of the distribution.
* Maximum number of chick weight is in range 50-100.
* Minimum number of chick weight is in range 350-400.
* In this case mean>median.



ANS:-

* This box plot has outliers.
* Outliers are extreme values.
* Outliers are above the max value line so it’s a positively skewed distribution.
* Here mean>median.

**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?

ANS:- Confidence interval = x̄ ±z (s/)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **94%** | **98%** | **96%** |
| **Lower** | 198.74 | 198.43 | 198.62 |
| **Upper** | 201.26 | 201.56 | 201.38 |

**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.

ANS:- Mean=41

Median=40.5

Variance=24.11

Standard deviation=4.91

1. What can we say about the student marks?

ANS:-

* Students marks has a variance so marks are not normally distributed.
* From Box Plot we can confirm that most the values lies between 34 to 45.
* We can also say that there are two outlayers in the data which are 49 and 56.
* From distplot we can say that the data is **positively skewed**

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

ANS:- There is no skewness the data is normally distributed.

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

ANS:- This is positive skewness. In this the tail of the distribution is longer on the

right side.

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

ANS:- This is negative skewness. In this the tail of the distribution is longer on the

left side.

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

ANS:- Positive kurtosis is called as Leptokurtic and in this distribution the peak is higher and sharper and tail is flatter than normal distribution**.**

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

ANS:- Negative kurtosis is called as Platykurtic and in this distribution the peak is

lower and broaderand tail isthinner than normal distribution**.**

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



What can we say about the distribution of the data?

ANS:- The data is not normally distributed .

What is nature of skewness of the data?

ANS:- This data is negatively skewed data.

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

ANS:- 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.

ANS:-

* Both box plots are normally distributed.
* No outliers in both the box plots.
* Both plots have the same Inter Quartile Range.
* First plot data are distributed in smaller range as compare to the second plot.

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)

Ans:- 0.347

* 1. P(MPG<40)

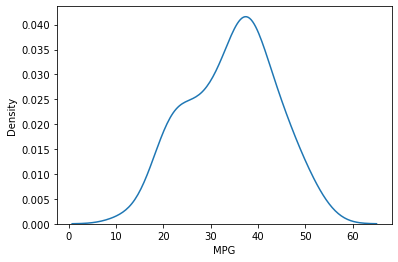
Ans:- 0.729

c. P (20<MPG<50)

Ans:- 0.898

Q 21) Check whether the data follows normal distribution

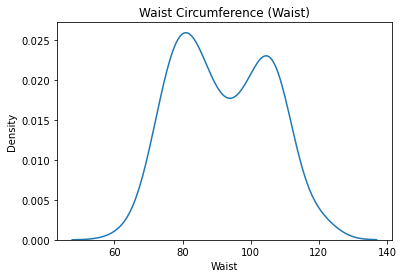
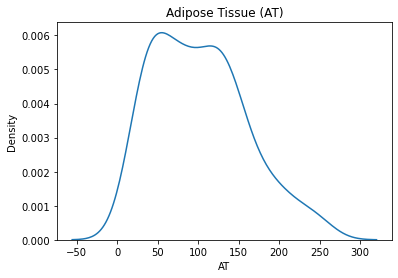
1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

MPG follows 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



Both Adipose Tissue (AT) and Waist Circumference(Waist) doesn’t fallow normal distribution.

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

Ans:- Z Score of 90% confidence interval = 1.645

Z Score of 94% confidence interval = 1.880

Z Score of 60% confidence interval = 0.842

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

Ans:- T Score of 95% confidence interval = 2.064

T Score of 96% confidence interval = 2.172

T Score of 99% confidence interval = 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

ANS:-

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

t= (260-270)/(90/(√18))

t=-0.471

Degrees of freedom = 18-1= 17

P = 0.32