|  |  |
| --- | --- |
| 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/Binary |

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/Binary |
| 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 | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

-3/8

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

1. Equal to 1

* 0

1. Less than or equal to 4

* 6/36=1/6

1. Sum is divisible by 2 and 3

* 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?

* 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

* 3.09

1 \* 0.015 + 4\*0.20 + 3 \*0.65 + 5\*0.005 + 6 \*0.01 + 2 \* 0.12=3.09

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

Mean for all the variables:

Points - 3.596563; Score - 3.217250; Weigh - 17.848750

Median for all the variables:

Points - 3.695; Score - 3.325; Weigh - 17.710

Mode for all the variables:

Points - 3.07; Score - 3.44; Weigh - 17.02

Standard Deviation and Variance for all the variables:

Points - 0.534679 & 0.285881; Score - 0.978457 & 0.957379; Weigh - 1.786943 & 3.193166

Range for all the variables:

Points - [2.76 - 4.93]; Score - [1.513 - 5.424]; Weigh - [14.5 - 22.9]

From Standard Deviation for all three variables, it can be observed that the variation for the variable Weigh is greater than that of the variables Points and score i.e. the deviation from the mean for the Variable Weigh is much greater than other two variables.



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?

* 163.5

1/8[108+110+123+134+135+145+167+187+ 199]=163.5

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

**Cars speed and distance**



**SP and Weight(WT)**



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

Skewness and Kurtosis for the variables Speed and Distance:

speed = -0.117510 & -0.508994; dist = 0.806895 & 0.405053

Skewness and Kurtosis for the variables SP and Weight:

SP = 1.611450 & 2.977329; Weight = -0.614753 & 0.950291;

Interpretation:

Speed-

Since Skewness value of Speed is negative then the values in Speed Variable follow Negative(Left) Skewed Distribution

Since Kurtosis value of Speed is negative then the values in Speed Variable follow Negative(Platykurtic) Kurtosis Distribution

Distance-

Since Skewness value of Distance is positive then the values in Distance Variable follow Positively(Right) Skewed Distribution

Since Kurtosis value of Distance is positive then the values in Distance Variable follow Positive(Leptokurtic) Kurtosis Distribution

Note:

1.If skewness is less than -1 or greater than 1, the distribution is highly skewed. If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed. If skewness is between -0.5 and 0.5, the distribution is approximately symmetric

2.Any distribution with kurtosis == 0 is called mesokurtic. A distribution with kurtosis <0 is called platykurtic. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader. The Critical value for kurtosis are considered using Fisher’s definition of kurtosis where kurtosis of normal == 0.0

SP-

Since Skewness value of SP is positive then the values in SP Variable follow Positively(Right) Skewed Distribution

Since Kurtosis value of SP is positive then the values in SP Variable follow Positive(Leptokurtic) Kurtosis Distribution

Weight-

Since Skewness value of Weight is negative then the values in Weight Variable follow Negative(Left) Skewed Distribution

Since Kurtosis value of Weight is positive then the values in Weight Variable follow Positive(Leptokurtic) Kurtosis Distribution

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



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

Most of the data points are concentrated in the range 50-150 with average frequency 150 and the frequency sees a downfall as the weight increases.

So, the expected value for the above distribution is 75.

The above distribution has a long tail towards right so it can be concluded that it is heavily Right Skewed



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

Area covered by upper part of the Box is greater than the lower part which means that the Distribution follows Right(Positively) Skewed Distribution and it is noticed that there are outliers on the upper side of box plot and there is less data points between Q1 and bottom point

**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 ?

* The range of 94% confidence interval = [198.739,201.2611]
* The range of 98% confidence interval = [198.437,201.563]
* The range of 96% confidence interval = [198.6248,201.3752]

**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 (Average) - 41; Median - 40.5; Range - 22; Mode - 41; Standard Deviation - 4.9103066; Variance - 24.11111

1. What can we say about the student marks?

* Most of the Datapoints lie close to the mean value.

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

* If the mean, median, and mode are approximately equal to each other, the distribution can be assumed to be approximately symmetrical i.e. the distribution has Skewness of zero

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

* Positively Skewed (Right skewed)

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

* Negatively Skewed (Left skewed)

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

* The distribution has heavier tails and a sharper peak than the normal distribution

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

* The distribution has lighter tails and a flatter peak than the normal distribution

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



What can we say about the distribution of the data?

* The spread is larger i.e. variation in the data is more and Median is greater than Mean

What is nature of skewness of the data?

* Left(Negative) Skewed Distribution, median is greater than mean

What will be the IQR of the data (approximately)?   
- 8

Q19) Comment on the below Boxplot visualizations?



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

* IQR is different for both Boxplots
* The spread (interquartile distance) is smaller for Boxplot1 rather than Boxplot2
* From the Boxplots, it can be noticed that whisker level is high in Boxplot2
* No outliers in both the boxplots
* Mean and Median are equal hence distribution is symmetrical

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.

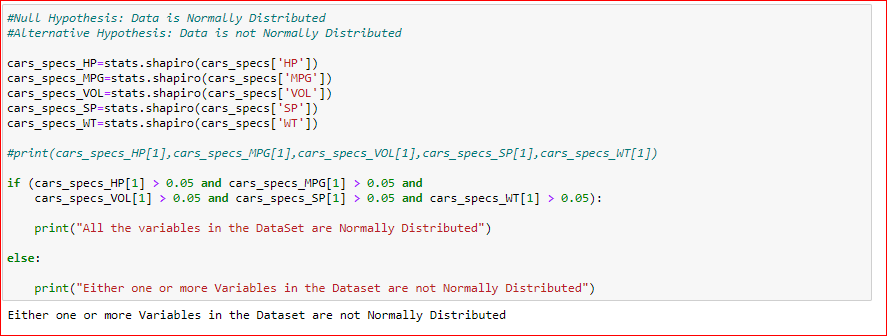
MPG <- Cars$MPG

* 1. P(MPG>38)
  + P(MPG>38) = 0.34827~28 records which is close to the Actual Number of records with MPG>38 i.e. 33. The Difference in Actual and Theoretical value is that the Theoretical Value has been calculated assuming the data is Normally Distributed about the mean perfectly
  1. P(MPG<40)
  + P(MPG<40) = 0.72907 ~ 59 records which is close to the Actual Number of records with MPG<40 i.e 61. The Difference in Actual and Theoretical value is that the Theoretical Value has been calculated assuming the data is Normally Distributed about the mean perfectly

c. P (20<MPG<50)

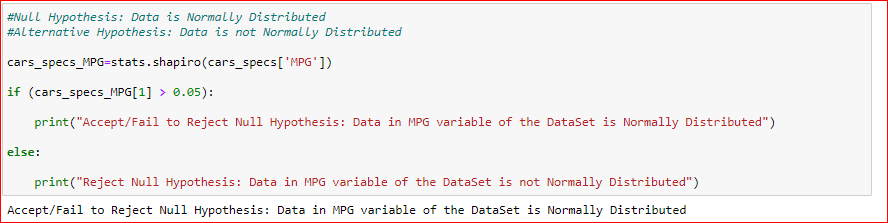
- P(20<MPG<50) = 0.89932 ~ 73 records

Q 21) Check whether the data follows normal distribution



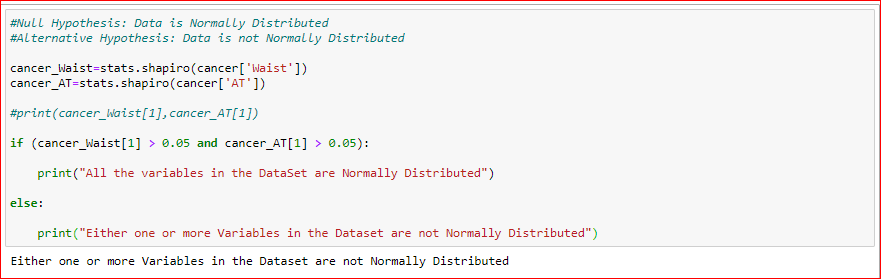
1. Check whether the MPG of Cars follows Normal Distribution

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

* Z score of 90% confidence interval is 1.64
* Z score of 94% confidence interval is 1.88
* Z score of 60% confidence interval is 0.84

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

* t-score of 95% confidence interval is 2.064
* t-score of 96% confidence interval is 2.172
* t-score of 99% confidence interval is 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

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

t = - 0.471;

With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average P(X < 260) is **0.3218** assuming the mean life of the bulbs is 300 days