|  |  |
| --- | --- |
| 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 | 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 | 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 | ratio |
| SAT Scores | Interval |
| Years of Education | ratio |

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

2/3 &1/3

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

1. Equal to 1

ANS: unusual event

1. Less than or equal to 4

ANS: 5/36

1. Sum is divisible by 2 and 3

ANS: 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?

ANS: 5/7

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 : 3.09

Expected value = Σx.p(x)

Here no. of. candies = x

Respective probabilities = p(x)

Expected value = 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean | 3.5965625 | 3.21725 | 17.84875 |  |
| Median | 3.695 | 3.325 | 17.71 |  |
| Mode | 3.92 | 3.44 | 17.02 |  |
| Std | 0.534678736 | 0.978457443 | 1.786943236 |  |
| Variance | 0.285881351 | 0.957378968 | 3.193166129 |  |
| max | 4.93 | 5.424 | 22.9 |  |
| min | 2.76 | 1.513 | 14.5 |  |
| RANGE | 2.17 | 3.911 | 8.4 |  |

**Use Q7.csv file**

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: 145.33

Expected value = Σx.p(x)

Here x = 108, 110, 123, 134, 135, 145, 167, 187, 199

P(x) = 1/9

E(x) = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**ANS: df.speed.skew() =>** -0.11750986144663393

If skewness value is negative (-) then the data is stated as negatively skewed or left skewed. Since our data is negatively skewed here, it means that it has a lower number of data points having low values.

And if value is lied between -0.5 to 0.5 then it is saying as fairly symmetrical

So, speed in data set Q9\_a.csv is fairly symmetrical.

df.dist.skew()=> 0.8068949601674215

Since our data is positively skewed here, it means that it has a higher number of data points having low values

If skewness value is positive (+) then the data is stated as positively skewed or right skewed

So, distance in data set Q9\_a.csv is right skewed or positive skewed.

df.speed.kurt() => -0.5089944204057617

Negative values of kurtosis indicate that a distribution is flat and has thin tails. Platykurtic distributions have negative kurtosis values.

A platykurtic distribution is flatter (less peaked) when compared with the normal distribution, with fewer values in its shorter (i.e. lighter and thinner) tails.

df.dist.kurt()=>0.4050525816795765

Positive values of kurtosis indicate that a distribution is peaked and possess thick tails. Leptokurtic distributions have positive kurtosis values.

A leptokurtic distribution has a higher peak and taller (i.e. fatter and heavy) tails than a normal distribution.

If kurtosis = 0 then it is equal to normal distribution.

**SP and Weight(WT)**

**Use Q9\_b.csv**

df.SP.skew() => 1.6114501961773555

If skewness value is positive (+) then the data is stated as positively skewed or right skewed. Since our data is positively skewed here, it means that it has a higher number of data points having low values

So, distance in data set Q9\_a.csv is right skewed or positive skewed.

df.WT.skew() => -0.6147533255357768

A negatively skewed distribution is the distribution with the tail on its left side. The value of skewness for a negatively skewed distribution is less than zero.

Since our data is negatively skewed here, it means that it has a lower number of data points having low values.

**df.SP.kurt() =>** 2.9773289437871764

Positive values of kurtosis indicate that a distribution is peaked and possess thick tails. Leptokurtic distributions have positive kurtosis values. In a data the obtained kurtosis is positive then middle values in that data frame are so high compared to the left and right side values.

df.WT.kurt() => 0.9502914910300326

Positive values of kurtosis indicate that a distribution is peaked and possess thick tails. Leptokurtic distributions have positive kurtosis values. In a data the obtained kurtosis is positive then middle values in that data frame are so high compared to the left and right side values.

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



In the above histogram most of the data values lies on the left side of the histogram. It shows that the is positively skewed.

A positively skewed distribution is the distribution with the tail on its right side. The value of skewness for a positively skewed distribution is greater than zero. As you might have already understood by looking at the figure, the value of mean is the greatest one followed by median and then by mode.



In above box plot, there so many outliers at the top of the upper quartile and when compared to the lower quartile range the upper quartile is so bigger.

And the data used to plot the above box plot is slightgly positive skewed.

**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: At 94% confidence interval is (143.57619175546247, 256.42380824453755)

At 96% confidence interval is (138.38753268104531, 261.61246731895466)

At 98% confidence interval is (130.2095637787748, 269.7904362212252)

**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

Variance= 25.52941176

Standard deviation= 5.052663829

Median= 40.5

1. What can we say about the student marks?

ANS: The given data is normally distributed it follows very negligibly deviated bell curve almost all mean = median = mode= 41.

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

ANS: Normal distributed, skewness=0

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

ANS: Negatively skewed

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

ANS: Positively skewed

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

ANS: Positive values of kurtosis indicate that a distribution is peaked and possess thick tails. Leptokurtic distributions have positive kurtosis values. In a data the obtained kurtosis is positive then middle values in that data frame are so high compared to the left and right-side values.

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

ANS: Negative values of kurtosis indicate that a distribution is flat and has thin tails. Platykurtic distributions have negative kurtosis values.

A platykurtic distribution is flatter (less peaked) when compared with the normal distribution, with fewer values in its shorter (i.e., lighter and thinner) tails.

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



What can we say about the distribution of the data?

ANS: In the above box plot, we observe that median= 15, first quartile Q1=10 & third quartile Q3= 18, from this values we can find Inter quartile range (IQR) = 18-10= 8

And from above box plot we observe that median is slightly deviated to right side it shows that taken is negatively skewed and from formula maximum= Q3+IQR\*1.5 and minimum= Q1+IQR\*1.5. We get maximum=30 and minimum= 22.

What is nature of skewness of the data?

ANS: Negative skewness

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: In the both the box plots 1 & 2 median value is same and both the boxplots doesn’t have outliers .

Box plot 1 slightly negative skewed and plot 1 data frame located in smaller region than box plot 2. Box plot 2 is normally distributed and its skewness is zero.

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) = 0.3475939404145300
2. P(MPG<40) = 0.72934986041579

c. P (20<MPG<50) = 0.8988689076273199

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

ANS: In dataset: cars.csv the MPG of cars doesn’t follow normal distribution, because in normal distribution mean=median=mode but in cars.csv the mean, median & mode of MPG of cars are not equal.

Mean= 34.422

Median= 35.152

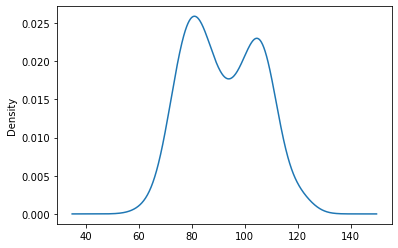
Mode= 29.629

And line plot of MPG is not in bell shape.

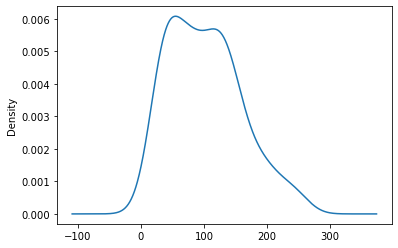
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

ANS: In Dataset: wc-at.csv the Adipose Tissue (AT) and Waist Circumference (Waist) are doesn’t follow the normal distribution



Waist line plot (mean!=median!=mode)



AT lineplot(mean! =median! =mode)

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

ANS: 90% = 1.282

94% = 1.555

60% = 0.253

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

ANS: 95%= 1.7108820799694275

96%= 1.8280511719596342

99%= 2.492159473157762

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 - statistics for the data is given as follows:



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 = -10/(90/3√2)

t = -10/(30/√2)

t = (-1\*√2) / 3

t = - 0.471

For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.

The probability that **t < - 0.471 with 17 degrees of freedom** assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a 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.