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

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 | Interval |
| 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 | Ordinal |
| 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?

* S = { HHH ,HHT,HTH,HTT,TTT,TTH,THT,THH}
* A = {HHT,HTH,THH}
* P(A) =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

* A) 0
* B) 1/6
* C) 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?



Let S be the sample space.  
Then, n(S) = Number of ways of drawing 2 balls out of 7  
=7C2​  
=(2×1)(7×6)​  
=21  
Let E = Event of drawing 2 balls, none of which is blue.  
∴n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls.  
=5C2​  
=(2×1)(5×4)​  
=10  
∴P(E)=n(S)n(E)​=21/10​

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

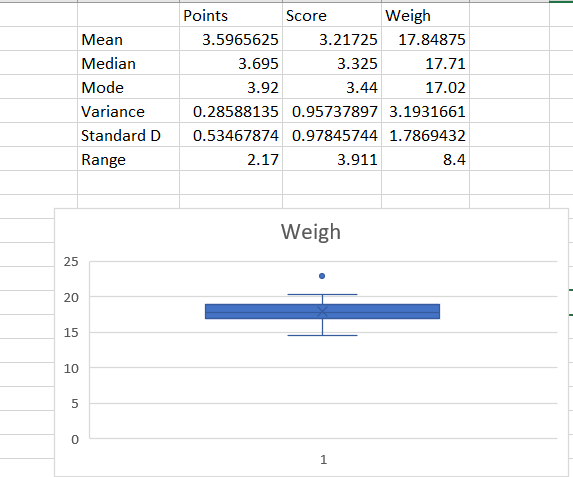
* 1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.120
* 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24
* 3.09

*So the expected value of candies = 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.



* Points and Score column has normal graph it don’t contains any outlier.
* Weigh column has right skewed graph it contain a outlier that is 22.9

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?

* P(x) = 108,110,123,134,135,145,167,187,199

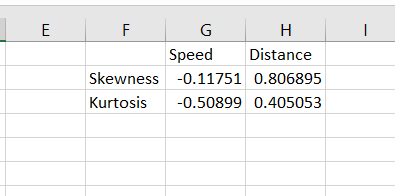
|  |  |  |
| --- | --- | --- |
| X | P(x) | P(x)\*X |
| 1 | 108 | 108 |
| 2 | 110 | 220 |
| 3 | 123 | 369 |
| 4 | 134 | 536 |
| 5 | 135 | 675 |
| 6 | 145 | 870 |
| 7 | 167 | 1169 |
| 8 | 187 | 1496 |
| 9 | 199 | 1791 |

E(X) = = 7234

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

**Cars speed and distance**

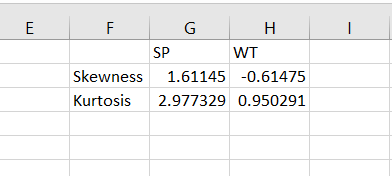




* For speed it is negative skew & for distance it is positive skew.
* So it contains outliers.
* For speed kurtosis is plat & for distance kurtosis is peaked.

**SP and Weight(WT)**





* For SP it is positive skew & for WT it is negative skew.
* So it contains outliers.
* For SP kurtosis is more peaked & for WT kurtosis is less peak than SP.

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



* It is a right skewed graph, so it contains outliers which would between 300-400.
* 50% of data is between 50 - 150



* The given data is right skewed.
* Above the upper extreme there exist few dots which means data con

**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?

* CI = X̄ ± Z×(s/√n)
* X̄ = 200
* n = 2000
* Std.S = 30
* Z = 1.8808, For 94% [198.738 – 201.262]
* Z = 2.3263, For 98% [198.439 – 201.561]
* Z = 2.0537, For 96% [198.622 – 201.378]

**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 = 41
* Median = 40.5
* Var = 25.52941
* Std = 5.052664

1. What can we say about the student marks?

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

* The skewness will be 0, symmetrical so data will be normally distributed.

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

* The nature of skewness is positive. Most of the data will be lying on the left side of the plot.

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

* The nature of skewness is negative. Most of the data will be lying on the right side of the plot.

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

* It indicates the distribution is peaked with thick tail.

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

* It indicates that distribution is flat with thin tail.

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



What can we say about the distribution of the data?

* The data is not equally distributed across the plane.
* The median of the data is approximately 15.2
* 25 percent of the data lies between 0-10
* 50 percent of the data lies between 10-18
* 25 percent of the data lies after 18-20

What is nature of skewness of the data?

* The nature will be left skewed for the given data.

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

* Q1 = 10
* Q2 = 15.2
* Q3 = 18
* IQR = Q3 – Q1
* IQR = 8

Q19) Comment on the below Boxplot visualizations?



* 1) Slightly less distributed, it is spread in range 250-290 approximately.
* 2) Highly distributed across the plane, it is spread in range 200-350 approximately.

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

* Here when we compare box plot 1 with box plot 2, we can say that the data in box plot 1 is widely spread. Here the main inference is that since the data range varies high in box plot 2 it is hard to make a prediction in box plot 2. The median in the 2 box plots is equal. And the data spread in both of them are 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)

= mean (MPG)

=34.42208

= sd(MPG)=9.131445

= 1 – pnorm(38, mean(MPG),sd(MPG))

= 0.330 = 33%

* 1. P(MPG<40)

=pnorm(40, mean(MPG),sd(MPG))

=0.7293499

=72.3%

* 1. P (20<MPG<50)

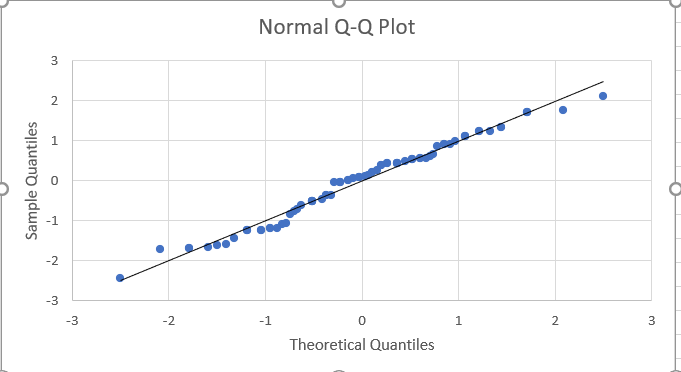
=pnorm(50, mean(MPG),sd(MPG)) – pnorm(20,mean(MPG),sd(MPG)) =0.955 -0.057

=0.8988689

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

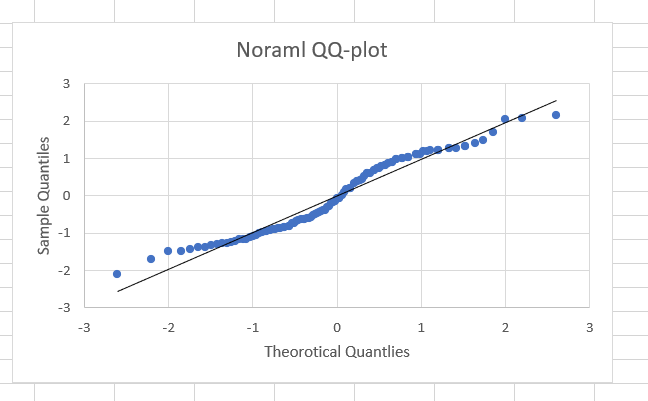
Dataset: Cars.csv



* After plotting QQ-plot we can see the majority of data lies near the line so our data is normal distributed.

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

Dataset: wc-at.csv



* After looking at the QQ-plot, we can say that the points are not near the line. Data contains some outilers.

Hence the distribution is abnormal.

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

* 90%
* 95+2.5
* 97.5
* qnorm(0.975) =1.96
* 94%
* 94+4
* 97
* qnorm(0.97) =1.88
* 60%
* 60 + 20
* 80
* qnorm(0.80) = 0.841

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

* T-Score Calculation
* T((1,alpha),(n-1))
* Here n = 25
* n-1 = 24
* 95%
* qt(0.975,24)
* 2.063899
* 96%
* qt(0.98,24)
* 2.171545

* 99%
* qt(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

* Sample size = n = 18
* Sample mean = x = 260 days
* Sample standard deviation = s = 90 days
* 260 – 270/90/SQRT(18)
* -10/9.487
* -1.054