|  |  |
| --- | --- |
| 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 | 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 or Ratio |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval or Ratio |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Nominal |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Ratio |
| Years of Education | Ratio |

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

**Ans :-** 3/8 or 37%

There are 3 favorable outcomes and each time if we have the condition like

Three coins are tossed and from those two heads and one tail then we can

Say that (H,H,T), (H,T,H), (T,H,H)

So from those 2 possible outcomes every time the coin is tossed then we can

Say that 2 power 3 that gives 2\*2\*2 = 8

Probability = (Number of Favorable outcomes) / (Total Number of

outcomes)

Probability = 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 :-**

1. 1/36

Two dice are rolled and each dice has 6 sides so 6 \* 6 = 36

And question asked is equal to 1 so from those 36 outcomes only one has asked so 1/36

1. 1/6

Less than or equal to 4 are 2, 3, 4

Sum of 2 will be the combination of (1,1)

Sum of 3 will be the combination of (1,2), (2,1)

Sum of 4 will be the combination of (1,3), (2,2), (3,1)

So, there are total 6 favorable outcomes and overall are 36

6/36 can be cancelled by 6 and it will become 1/6

c) 1/36

P (Sum is divisible by 2 and 3) = N (Event (Sum is divisible by 2 and 3)) / N(Event (Two dice rolled)) = 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?**

**Ans :-**

1st Draw

Probability of drawing a red ball on the first draw = 2 / 7

Probability of drawing a green ball on the first draw = 3 / 7

Sum of the Red and Green is

2 / 7 + 3 / 7 = 5 / 7

2nd Draw

Since the first draw has done, now the total number of balls present are 6.

Probability of drawing a red ball on the second draw = 2 / 6

Probability of drawing a green ball on the second draw 4 / 6

Multiplication Rule

(Probability of first draw being non-blue) \* (Probability of second draw being non-blue) = (5/7) \* (2/6) = 10/42 = 5/21

So, the probability that none of the balls drawn is blue is **5/21**.

**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 :-**

(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**

**Ans :-**

**Points** : Mean 3.596563

Median 3.695

Mode 0 3.07 1 3.92

Variance 0.2858813508064516

Standard Deviation 0.534679

Range 2.760 - 4.930

The points column says that the minimum starts from 2.760 and maximum is 4.930, and it clearly says that there are no outliers in the Points column. The Mean value is 3.59 and Median is 3.695 which clearly says that it’s a right skewed data and if we see the boxplot there are more data in between Q1 and Q2 as compare to between Q2 and Q3.

**Score** : Mean 3.217250

Median 3.325

Mode 0 3.44

Variance 0.9573789677419353

Standard Deviation 0.978457

Range 1.513 - 5.424

The Score column is near normal distribution, where as in between 4 and 5 there are no data points so that we have a bum in between. How ever we have more data points are in between the 3 and 4 that’s the reason we can say that it has the peakiness i.e. kurtosis. Yes there are outliers and we have only at one side and those are lies in between the 5 and 5.5 so that we can separate those points can see the normal data how it looks.

**Weigh** : Mean 17.848750

Median 17.710

Mode 0 17.02 1 18.90

Variance 3.1931661290322575

Standard Deviation 1.786943

Range 14.500 - 22.900

The Weigh column is also a near normal distribution where as there are no data points on the point 21 so that we can see that we have a bum in the distplot and if we see there are only some outliers not much as compare to the Score column and if we want the better accuracy we can remove the outliers into separate dataset and work on that separately and these are the some of the comments and the inferences on the Q7 dataset.

**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 :-**

For this we have to sum all the values and divide it by the total no. of values

i.e. 108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199

sum is 1198

and the number of items are 9

so 1198/9 that’s value is approximately 133.11

Answer is 133

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans :-**

As we can see that we have only 2 columns that we can use for the inferences i.e. speed and dist columns so we can see there is a linear relationship between those two columns and also speed column is a normally distributed where as dist column is a right skewed data and there are no outliers in the speed data and where as dist data contains the outliers and only one side that means we can separate the outliers or we can even remove those data points and we can see the data how it looks without outliers data.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans :-**

In this dataset we have only 2 columns and if we see that in the SP column we have many outliers and there are many chances that if we try to predict with that dataset without changing the dataset from outlier and non-outlier dataset we can get the bad results so it’s better to divide the outlier data to separate dataset and non-outlier data into the separate dataset and draw inferences and if we see that it’s a clearly right skewed data if we see the data points only until the around 90 to 140 it would be near normal distribution and where as we have many data points from 140 even so that it’s tail has extended from 140 to until 180 so that it has become the right skewed data.

If we see the WT column even here also we have outliers and especially we can see that both sides we have outliers so that we can say that we can not just delete or remove the outliers and if we want to delete it not at all possible either we need to do good EDA or the best another thing we can do is separate the outliers data into one dataset and we have to see the how the results would look and these are some of the comments and inferences we can draw from the given dataset.

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



**Ans :-**

From the barplot we can say that Weight’s are right skewed data and from the boxplot we can say that there are outliers. Maximum of the weights are in the range of 50 and 100 and further it has gradually decreased and typically highest value would be near 400 and minimum value is 0. The frequency of the data is varied through out the data and highest frequency is around 200.

**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:-**

# Avg. weight of Adult in Mexico with 94% CI

stats.norm.interval(0.94,200,30/(2000\*\*0.5))

* (198.738325292158, 201.261674707842)

# Avg. weight of Adult in Mexico with 98% CI

stats.norm.interval(0.98,200,30/(2000\*\*0.5))

* (198.43943840429978, 201.56056159570022)

# Avg. weight of Adult in Mexico with 96% CI

stats.norm.interval(0.96,200,30/(2000\*\*0.5))

* (198.62230334813333, 201.37769665186667)

**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 41.0

Median 40.5

Variance 25.52

Standard Deviation 5.05

1. **What can we say about the student marks?**

**Ans :-**

From the marks list we can say that it’s a right skewed data where as minimum value is 34 and maximum value is 56. If we consider the values from 34 to 45 then we can see that It looks near normal distribution. Where as we can consider the 49 and 56 are the outliers, in that case we can remove the outliers by just removing the outliers because these are very less in count.

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

**Ans :-**

If Mean and Median of the data are equal then we can say that the distribution is exactly normal. We can say that it will be like 1st Standard Deviation has the 68% of the data, 2nd Standard Deviation has the 95% of the data and 3rd Standard Deviation has the 99% of the data.

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

**Ans :-**

If Mean is greater than Median then the data is called the right skewed data.

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

**Ans :-**

If Median is greater than Mean then the data is called the left skewed data.

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

**Ans :-**

If kurtosis value is positive then we can say that the data will be looked like the peaked will be very high.

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

**Ans :-**

If kurtosis value is negative then we can say the data will be like dispersed around the entire distribution. There will be no any peakness in the whole data.

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



**What can we say about the distribution of the data?**

**Ans :-**

From the Boxplot we can say that most of the data lies in the left side of the data and clearly we can say the it’s a left skewed data that means Mean is less than Median. We can also say that there could be outliers towards the left side of the data that means less number of values are more as compare to the large number of values. Minimum value is around 1 and maximum value could be 19.

**What is nature of skewness of the data?**

**Ans :-**

There are two types of the skewness those are positive and the negative skewness. If the skewness is right skewed data then it is a positive skewness that also tells that mean is greater than the median value of the data. If the skewness is left skewed data then it is a negative skewness that also tells that mean is less than median value of the data.

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

**Ans :-**

In the boxplot we have 4 whiskers and those are partitioned by this 5 points and those are called the 5 Point Summary.

We have **Minimum value, Q1, Q2, Q3, Maximum value.**

If the distribution is normally distributed then we can say the each whisker has the 25% of the data and combination of 4 whiskers is equals to 100%.

**IQR = Q3-Q1**

And we can also say that any value less than **Q1-1.5(IQR)** and any value greater than **Q3+1.5(IQR)** are the outliers statistically.

**Q19) Comment on the below Boxplot visualizations?**



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

**Ans :-**

From the diagram we can say that there are no any outliers in the data. If we compare both boxplot 1 and 2 we can say that median value and mean values are almost same for both and both are Normally Distribution data.

Clearly we can say that Boxplot 1 has the very less number of values in the dataset and whereas in the boxplot 2 has the many values as compare to the 1 and these are some of the inferences.

**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)**
  3. **P (20<MPG<50)**

**Ans:-**

1. # P(MPG>38)

1-stats.norm.cdf(38,df.MPG.mean(),df.MPG.std()) => 0.3475

1. # P(MPG>40)

stats.norm.cdf(40,df.MPG.mean(),df.MPG.std()) => 0.7293

1. # P (20<MPG<50)

stats.norm.cdf(0.50,df.MPG.mean(),df.MPG.std())-stats.norm.cdf(0.20,df.MPG.mean(),df.MPG.std()) => 1.2430

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans :-**

No actually but we can say that it is partially normal distribution or near normal distribution, we can see from the distplot that in between 20 and 30 we got a bump that made the distribution some what different from the normal distribution.

Conclusion is NO, MPG does not follow 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

**Ans :-**

AT : No it’s not a normal distributed data, from the distplot we can say that there are 2 peaks in the data and since we can say that it’s not a normally distributed data.

Waist : No even it’s not a normal distributed data, from the distplot we can say that it is a right skewed data and there is no peaked ness in the data even and

Both AT and Waist are not normally distributed data.

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

**Ans :-**

z score at 90 % confidence interval is 1.645, at 94 % confidence interval is 1.555 and at 60 % confidence interval is 0.253.

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

**Ans :-**

Here we have confidence interval for sample size of 25

And Degrees of freedom would be n – 1 i.e. 25-1=24

stats.t.ppf(0.95,24) = 1.7108

stats.t.ppf(0.96,24) = 1.8280

stats.t.ppf(0.99,24) = 2.4921

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 :-**

Ho = Average life of Bulb >= 260 days

Ha = Average life of Bulb < 260 days

t=(sample\_mean-Population\_mean)/(sample\_SD/sqrt(n))

sample mean = 260

Population mean = 270

Sample standard deviation = 90

n = 18

degrees of freedom = n -1 =18 -1 = 17 => df=17

t=(260-270)/(90/18\*\*0.5)

t = -0.4714

p\_value = 1-stats.t.cdf(abs(-0.4714), df=17)

p\_value = 0.3216