**Q1) Identify the Data type for the Following:**

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

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

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

Possibilities of results = HHH, HTH, HHT, TTT, THT, TTH, THH, HTT. Therefore, Probability of getting two heads and one tail = 3/8 = 37.5%

**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**

P (Sum is equal to 1) = 0/36 = 0%

P (Sum is less than or equal to 4) = 6/36 = 1/6 = 16.66%

P (Sum is divisible by 2 and 3) = 6/36 = 1/6 = 16.66%

**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?**

N(2 balls randomly drawn) = 7! /2! \*5!

= (7\*6\*5\*4\*3\*2) / (2\*1) \* (5\*4\*3\*2)

= (7\*6) / (2\*1)

= 21

N(None of the balls drawn is blue) = 5! / 2! \* 3!

= (5\*4\*3\*2) / (2\*1) \* (3\*2)

= 5\*2 = 10

P (None of the balls drawn is blue) = N (None of the balls drawn is blue) /

N (Event (2 balls are drawn randomly from

bag)

= 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

Expected number of candies for a randomly selected child

= 0.015\*1 + 0.20\*4 + 0.65\*3 + 0.005\*5 + 0.01\*6 + 0.120\*2

= 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**

**Mean**

* for Points = 3.59
* for Score = 3.21
* for Weigh = 17.84

**Median**

* for Points = 3.695
* for Score = 3.325
* for Weigh = 17.710

**Mode**

* for Points = 3.07
* for Score = 3.44
* for Weigh = 17.02

**Variance**

* for Points = 0.28
* for Score = 0.95
* for Weigh = 3.19

**Standard Deviation**

* for Points = 0.53
* for Score = 0.97
* for Weigh = 1.78

**Range**

* for Points = 2.17
* for Score = 3.911
* for Weigh = 8.39

****

**Inferences**





We can conclude from the analysis that, there are some little amount outliers in Scores & Weigh. The data is close to normal distribution as there is not much difference between the mean and median values.

**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?**

Expected Value = (Probability of x \* x)

= (1/9\*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)

= 12 + 12.22 + 13.66 + 14.88 + 15 + 16.11 + 18.55 + 20.77 + 22.11

= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

*print(df['speed'].skew())* = -0.1175

*print(df['dist'].skew())* = 0.806

*print(df['speed'].kurtosis())* = -0.5089

*print(df['dist'].kurtosis())* = 0.405

From the above analysis we can conclude that the data of car speed is fairly symmetrical and car distance is moderately skewed

**SP and Weight(WT)**

**Use Q9\_b.csv**

*print(df['SP'].skew())* = 1.611

*print(df['WT'].skew())* = -0.614

*print(df['SP'].kurtosis())* = 2.977

*print(df['WT'].kurt())* = 0.950

From the above analysis we can conclude that the data of SP is highly skewed and WT is moderately skewed

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



The Histogram is right skewed as the tail is at the right. Therefore Mean>Median. We might have outliers on the higher side of the data.



The box plot has outliers at the upper extreme and no outliers are the lower extreme

**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?**

*print(stats.norm.interval(0.94,200,30/* *np.sqrt(2000)))* = (198.73, 201.26)

*print(stats.norm.interval(0.96,200,30/* *np.sqrt(2000)))* = (198.62, 201.37)

*print(stats.norm.interval(0.98,200,30/* *np.sqrt(2000)))* = (198.43, 201.56)

**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, Variance = 25.52, Standard Deviation = 5.05

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

The data is right skewed as the value of mean is greater than median

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

The data is normally skewed and a symmetry lies within the distribution.

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

The data is right skewed if the mean is greater than median which indicates a long-right hand side tail.

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

The data is left skewed if the mean is greater than median which indicates a long-left hand side tail.

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

The data has heavier tails and is more peaked.

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

The data has thinner tails and is less peaked.

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



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

The distribution are not normally distributed as the median is at the higher side

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

The data is left skewed as the median is greater than mean.

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

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.**

There are no outliers in both the boxplots. The median of both the boxplots is also similar between the range of 250-275 and there seems to be no skewness so both the boxplots are therefore normally distributed.

**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)**

*print(df['MPG'].mean())* = 34.422

*print(df['MPG'].std())* = 9.131

*P\_MPG>38 = 1 - stats.norm.cdf(38,34.42,9.13)* = 0.347

1. **P(MPG<40)**

*P\_(MPG<40) = stats.norm.cdf(40,34.42,9.13)*

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

*P (20<MPG) = 1 - stats.norm.cdf(20,34.42,9.13)* = 0.942

*P (MPG<50) = stats.norm.cdf(50,34.42,9.13)* = 0.956

*P (20<MPG<50) = P (MPG<50) - P (20<MPG)*

*= 0.956* – 0.942

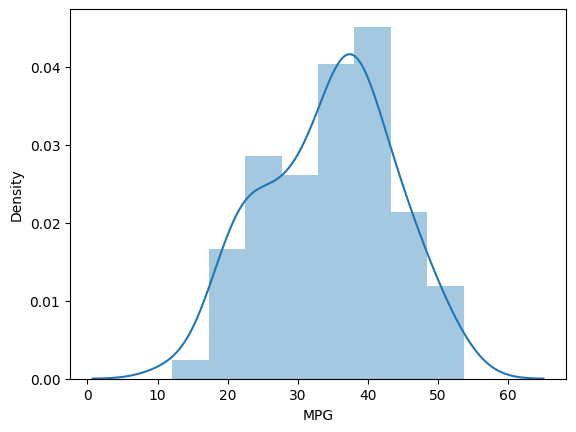
*= 0*.131

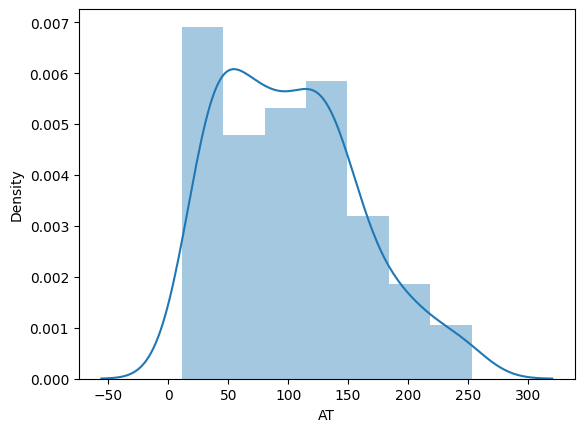
**Q 21) Check whether the data follows normal distribution**

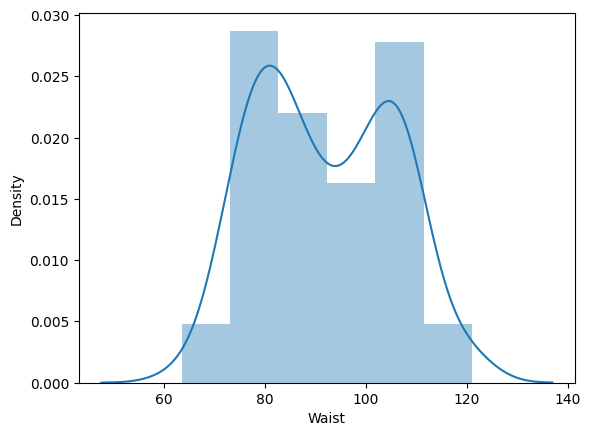
1. **Check whether the MPG of Cars follows Normal Distribution**

**Dataset: Cars.csv**

The MPG of Cars 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**



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

*print(stats.norm.ppf(0.05)) =* - 1.644

*print(stats.norm.ppf(0.03))* = -1.880

*print(stats.norm.ppf(0.2))* = -0.841

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

*print(stats.t.ppf(0.025,df=24)) =* -2.06

*print(stats.t.ppf(0.03,df=24)) =* -1.97

*print(stats.t.ppf(0.005,df=24)) =* -2.76

**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**

t\_score = (sample mean - population mean) / (sample standard deviation / square root of sample size)

= (260-270)/(90/np.sqrt(18))

= -0.471

*stats.t.cdf(-0.471,df=17)*

= 0.32 = 32%