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

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

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

**ans: 3/8**

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

* Equal to 1 = **0**
* Less than or equal to 4 = 6/36= **1/6**
* 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?

**Ans: 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

**ANS= 4.090**

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** | **score** | **weigh** |
| **Mean** | **3.5965** | **3.2172** | **17.8487** |
| **Median** | **3.695** | **3.325** | **17.710** |
| **mode** | **3.07, 3.92** | **3.44** | **17.02, 18.90** |
| **Range** | **2.17** | **3.911** | **8.400** |
| **variance** | **0.2858** | **0.9573** | **3.1931** |
| **standard deviation** | **0.5346** | **0.9784** | **1.7869** |

**- Mode: Points has 2 modes ,Score has one and Weigh has 2 modes.**

Q8) Calculate Expected Value for the problem below

* 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: 145.334**

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

**ans: 1. Cars speed and distance (Use Q9\_a.csv)**

**Skenwss: Speed= -0.11395477 , Distance:= 0.78248352**

**Kurtosis: Speed= -0.57714742,Dist= 0.24801866**

**2. SP and Weight(WT)(Use Q9\_b.csv)**

**Skenwss:SP=1.58145368, WT= -0.60330993**

**Kurtosis: SP= 2.72352149, WT= 0.81946588**

Q10) Draw inferences about the following boxplot & histogram



ans: **Range is 0 to 400 , where max frequency lies between 50 to 100 weight. the distribution follows a right sided skewness and most of the data is concentrated in left hand side.(mean>median>mode)**



**ans: Many Outliers are present way outside the upper 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?

given: n=2000; s=3000000; u=200; sd=30

z=x-u/s.d

**ans**:

**94%= 246.65**

**98%= 264.45**

**96%= 255.55**

**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**

* Find

**mean = 41.0,**

**median,= 40.5**

**variance = 24.111,**

**standard deviation= 4.91030**

* What can we say about the student marks?

**as the mean is greater than the median, the data is right sided skewed i.e. most data is concentrated before the mean and it has some outliers on the right side!**

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

**the data will be noramlly distributed with equal curves on either sides (Bell shaped where most values are concentrated arounf the mean and median value !)**

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

**The data is right skewed and most data lies on the left side of the mean. and has smaller values as contents of the data**

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

**The data is left skewed and most data lies on the right side of the mean. and has larger values as contents of the data**

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

**data has highter peaks, thin bell thickness and thick tails than normal distribution**

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

**data has Lower peaks, thick bell thickness and small tails than normal distribution**

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



What can we say about the distribution of the data?

**1.data is highly distributed on right side and thus, more values lie in the box region (50% values) i.e. in 10 to 18. and outer values can be called OUTLIERS.**

**2. Mean is around 15.**

**3. upper extreme and lower extremes cannot be seen here.**

What is nature of skewness of the data?

**Skewness is left sided, most values concentrsting on right region. (mean>median)**

What will be the IQR of the data (approximately)?   
  
**IQR: 16.25-12.50 = 3.75**

Q19) Comment on the below Boxplot visualizations?



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

ans:

**Its a Normal Distribution both Boxplots.**

**Boxplot 1 has upper extreme around 285 and lower extreme lower than 250. and Q1=250, Q3=275(lower and upper bound respectively)**

**Boxplot 1 and 2 has same mean=262.5**

**boxplot2 has lower and upper bound 225 and 300 respectively.**

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

* **P(MPG>38): 33/81=0.40740**
* **P(MPG<40): 61/81=0.7530**

**c. P (20<MPG<50): 69/81=0.85185**

Q 21) Check whether the data follows normal distribution

* Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**ans: It does not follow Normal distribution(as there are 2 peaks):**



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

Dataset: wc-at.csv

ans: **Adipose Tissue(AT) :**

**though the curve seems partially bell shaped, it still has 2 peaks , which means the data is not normally distributed.**

**Waist Circumference(WC):**

**the data has 2 peaks , which means the data is not normally distributed.**

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

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

**ans: z scores are 1.29, 1.56, 0.24 respectively.**

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

**ans:**

**scipy.stats.t.ppf(q, df)**

**where:**

**q: The significance level to use**

**df: The degrees of freedom= (sample size-1)**

**t scores of 95% = 1.7108**

**t scores of 96% =1.8280**

**t scores of 99%= 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:

**sample set of n=18**

**X bar= mean of sample= 260**

**meu= mean of population=270**

**s= standard deviation of sample=90**

**t=(X bar- meu)/(s/sqrt(n))**

**t score= -0.471404**

**ANS: Probability that 18 randomly selected bulbs would have an average life of no more than 260 days**

**= 0.3216727 (One tailed hypothesis)**