|  |  |
| --- | --- |
| 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 | Interval |
| Celsius Temperature | Interval |
| Weight | Interval |
| 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 | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| 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?

**Ans) Total number of events= {HHH, HHT, HTT, TTT, TTH, THH, HTH, THT} =8**

**Required events=3**

**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 2and 3

**Ans) Total number of outcomes when two dice are rolled=6\*6=36.**

**(1, 1)(1, 2)(1, 3)(1, 4)(1, 5)(1, 6)**

**(2, 1)(2, 2)(2, 3)(2, 4)(2, 5)(2, 6)**

**(3, 1)(3, 2)(3, 3)(3, 4)(3, 5)(3, 6)**

**(4, 1)(4, 2)(4, 3)(4, 4)(4, 5)(4, 6)**

**(5, 1)(5, 2)(5, 3)(5, 4)(5, 5)(5, 6)**

**(6, 1)(6, 2)(6, 3)(6, 4)(6, 5)(6, 6)**

**a) Equal to 1 = 0% probability**

**b) Less than or equal to 4= 6/36 = 1/6**

**c) sum is divisible by 2 and 3**

**{ 2 3 4 5 6 7**

**3 4 5 6 7 8**

**4 5 6 7 8 9**

**5 6 7 8 9 10**

**6 7 8 9 10 11**

**7 8 9 10 11 12}**

**Probability=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) total number of events= === 21**

**Interested events === 10**

**Probability that none of the balls is blue =10/21=0.47**

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)**

**Expected number = E(x) ==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.

**Use Q7.csv file**

**Using python**

**Jupyter notebook**

1. **Input - import pandas as pd**

**Output –**

1. **Input - df = pd.read\_csv("C:/Users/HP/Downloads/Q7.csv")**

**Output –**

1. **Input – df**

**Output –**

| **Unnamed: 0** | **Points** | **Score** | **Weigh** |
| --- | --- | --- | --- |
| **0** | Mazda RX4 | 3.90 | 2.620 | 16.46 |
| **1** | Mazda RX4 Wag | 3.90 | 2.875 | 17.02 |
| **2** | Datsun 710 | 3.85 | 2.320 | 18.61 |
| **3** | Hornet 4 Drive | 3.08 | 3.215 | 19.44 |
| **4** | Hornet Sportabout | 3.15 | 3.440 | 17.02 |
| **5** | Valiant | 2.76 | 3.460 | 20.22 |
| **6** | Duster 360 | 3.21 | 3.570 | 15.84 |
| **7** | Merc 240D | 3.69 | 3.190 | 20.00 |
| **8** | Merc 230 | 3.92 | 3.150 | 22.90 |
| **9** | Merc 280 | 3.92 | 3.440 | 18.30 |
| **10** | Merc 280C | 3.92 | 3.440 | 18.90 |
| **11** | Merc 450SE | 3.07 | 4.070 | 17.40 |
| **12** | Merc 450SL | 3.07 | 3.730 | 17.60 |
| **13** | Merc 450SLC | 3.07 | 3.780 | 18.00 |
| **14** | Cadillac Fleetwood | 2.93 | 5.250 | 17.98 |
| **15** | Lincoln Continental | 3.00 | 5.424 | 17.82 |
| **16** | Chrysler Imperial | 3.23 | 5.345 | 17.42 |
| **17** | Fiat 128 | 4.08 | 2.200 | 19.47 |
| **18** | Honda Civic | 4.93 | 1.615 | 18.52 |
| **19** | Toyota Corolla | 4.22 | 1.835 | 19.90 |
| **20** | Toyota Corona | 3.70 | 2.465 | 20.01 |
| **21** | Dodge Challenger | 2.76 | 3.520 | 16.87 |
| **22** | AMC Javelin | 3.15 | 3.435 | 17.30 |
| **23** | Camaro Z28 | 3.73 | 3.840 | 15.41 |
| **24** | Pontiac Firebird | 3.08 | 3.845 | 17.05 |
| **25** | Fiat X1-9 | 4.08 | 1.935 | 18.90 |
| **26** | Porsche 914-2 | 4.43 | 2.140 | 16.70 |
| **27** | Lotus Europa | 3.77 | 1.513 | 16.90 |
| **28** | Ford Pantera L | 4.22 | 3.170 | 14.50 |
| **29** | Ferrari Dino | 3.62 | 2.770 | 15.50 |
| **30** | Maserati Bora | 3.54 | 3.570 | 14.60 |
| **31** | Volvo 142E | 4.11 | 2.780 | 18.60 |

1. **Input - df.shape**

**Output -** (32, 4)

1. **Input - df.columns**

**Output -** Index(['Unnamed: 0', 'Points', 'Score', 'Weigh'], dtype='object')

1. **Input – df.dtypes**

**Output –**

6]:

Unnamed: 0 object

Points float64

Score float64

Weigh float64

dtype: object

1. **Input - df.info()**

**Output -** <class 'pandas.core.frame.DataFrame'>

RangeIndex: 32 entries, 0 to 31

Data columns (total 4 columns):

# Column Non-Null Count Dtype

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

0 Unnamed: 0 32 non-null object

1 Points 32 non-null float64

2 Score 32 non-null float64

3 Weigh 32 non-null float64

dtypes: float64(3), object(1)

memory usage: 1.1+ KB

1. **Input – df.describe()**

**Output -**

| **Points** | **Score** | **Weigh** |
| --- | --- | --- |
| **count** | 32.000000 | 32.000000 | 32.000000 |
| **mean** | 3.596563 | 3.217250 | 17.848750 |
| **std** | 0.534679 | 0.978457 | 1.786943 |
| **min** | 2.760000 | 1.513000 | 14.500000 |
| **25%** | 3.080000 | 2.581250 | 16.892500 |
| **50%** | 3.695000 | 3.325000 | 17.710000 |
| **75%** | 3.920000 | 3.610000 | 18.900000 |
| **max** | 4.930000 | 5.424000 | 22.900000 |

1. **Input – df.mean()**

**Output -** Points 3.596563

Score 3.217250

Weigh 17.848750

dtype: float64

1. **Input -df.std()**

**Output –**

Points 0.534679 Score 0.978457 Weigh 1.786943 dtype: float64

1. **Input - df.median()**

**Output -** Points 3.695 Score 3.325 Weigh 17.710 dtype: float64

1. **Input - df.min()**

**Output -**

Unnamed: 0 AMC Javelin

Points 2.76

Score 1.513

Weigh 14.5

dtype: object

1. **Input - df.max()**

**Output –**

Unnamed: 0 Volvo 142E

Points 4.93

Score 5.424

Weigh 22.9

dtype: object

1. **Input – df.mode()**

**Output -**

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: EV=Σx/n ==145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Answers –**

**Ans) #using e1071 package**

**> skewness(ex2\_csv$speed)**

**[1] -0.7983898**

**> kurtosis(ex2\_csv$speed)**

**[1] -0.2260851**

**> skewness(ex2\_csv$dist)**

**[1] 1.150886**

**> kurtosis(ex2\_csv$dist)**

**[1] 1.466731**

**#Using moments package**

**> skewness(ex2\_csv$speed)**

**[1] -0.8448909**

**> kurtosis(ex2\_csv$speed)**

**[1] 2.991396**

**> skewness(ex2\_csv$dist)**

**[1] 1.217917**

**> kurtosis(ex2\_csv$dist)**

**[1] 4.816933**

**Inferences: as you can see from the above data, there is a huge difference in the kurtosis values when e1071 and moments package are compared with each other. This is due to different equations used by the packages to find kurtosis.**

**SP and Weight(WT)**

**Ans)**

**#using e1071 package**

**> skewness(ex3\_csv$SP)**

**[1] -0.3898407**

**> skewness(ex3\_csv$WT)**

**[1] -1.230919**

**> kurtosis(ex3\_csv$SP)**

**[1] -1.034207**

**> kurtosis(ex3\_csv$WT)**

**[1] 0.5979244**

|  |
| --- |
| **#using moments package**  **skewness(ex3\_csv$SP)**  **[1] -0.4076944**  **> skewness(ex3\_csv$WT)**  **[1] -1.287292**  **> kurtosis(ex3\_csv$SP)**  **[1] 2.086738**  **> kurtosis(ex3\_csv$WT)**  **[1] 3.819284** |

**Use Q9\_b.csv**

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



**Ans: The above boxplot suggests that the distribution has lots of outliers towards 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?.

Ans: n=2000

= 200

s= 30

Confidence Interval Estimate= Z => 200 Z

94% Confidence: norms.stats(0.97)

[1] 1.880794=Z

200 1.88\* =**198.74 – 201.26**

98% Confidence: norms.stats(0.99)

[1] 2.326348=Z

200 2.33\* =**198.44-201.56**

96% Confidence: norms.stats(0.99)

[1] 2.053749

200 2.05\* = **198.62-201.38**

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

Ans: 1) > mean(ex$scores)

[1] 41

> median(ex$scores)

[1] 40.5

> var(ex$scores)

[1] 25.52941

> sd(ex$scores)

[1] 5.052664

**2)  Mean > Median, This implies that the distribution is slightly skewed towards right. No outliers are present.**

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

Ans) no skewness, symmetric

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

Ans) Right skewed(tail on the right side).

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

Ans) Left Skewed(tail on the left side).

Q16) What does positive kurtosis value indicates for adata ?

Ans) peakness (sharp peak) and less variation.

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

Ans) less peakness (Broad peak) and more variation.

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



What can we say about the distribution of the data?

Ans) it is not a Normal Distribution

What is nature of skewness of the data?

Ans) It is left skewed.

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

Ans) Inter Quartile Range =Upper Quartile- Lower Quartile => 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) 1) The median of the two boxplots are same approximately 260.

2) The boxplots are not skewed in +ve or –ve direction.

3) Outliers doesn’t exist in both of the boxplots.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG ofCars for the below cases.

MPG<- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

1. Ans) > mean(Cars$MPG)
2. [1] 34.42208
3. P(MPG>38):
4. > sd(Cars$MPG)
5. [1] 9.131445
6. > pnorm(38,34.42,9.13)
7. [1] 0.652513
8. P(MPG>38)=1-P(MPG<38)(PS: Z-table gives you only less than probabilities)
9. >1 - 0.65
10. [1] 0.35
11. P(MPG<40):
12. pnorm(40,34.42,9.13)
13. [1] 0.7294571
15. P (20<MPG<50):
16. > pnorm(50,34.42,9.13)-pnorm(20,34.42,9.13)
17. [1] 0.8989178

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans) Follows Normal distribution as indicated by qq-plot.



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

Dataset: wc-at.csv

**Ans) waist follows Normal Distribution from the below QQ-plot**

**> qqnorm(wc\_at$Waist)**

**> qqline(wc\_at$Waist)**



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

Ans) 90% 🡪> qnorm(0.95)

[1] 1.644854

94% 🡪> qnorm(0.97)

[1] 1.880794

60% 🡪> qnorm(0.8)

[1] 0.8416212

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

Ans) 95% 🡪> qt(0.975,24)

[1] 2.063899

96% 🡪> qt(0.98,24)

[1] 2.171545

99% 🡪 qt(0.995,24)

[1] 2.79694

Q 24**)**A Government companyclaims 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) µ=270, =260, SD=90, n=18, df=n-1=18-1= 17

tscore= = = -10/21.23= -0.47

> pt(-0.47,17)

[1] 0.3221639

Required probability = 0.32=32%