Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | discreate |
| 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 |

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 data |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Groupd | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | interval |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | Inerval |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

Ans-

Possible outcomes for Three coin tossed ={HHH,HHT,HTT,TTT,TTH,THH,THT,HTH}

Two heads and on tail are obtained={HHT,THH,HTH}

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-

Possible no of outcomes after rolling two dice are=36

1. Sum Equal to 1-When we roll a dice the smallest number we get is 1 if we roll two we get two 1 and its sum is more than two so there is no way we can get sum 1,so the probability of getting sum 1=0
2. Less than or equal to 4- Possible outcomes-{(1,1)(1,2)(1,3)(2,1)(2,2)(3,1)}=6

Probability=6/36=1/6

c)Sum is divisible by 2 and 3-The number which divisible by 2 and 3 are -6&12

Possible outcomes are-{(1,5)(2,4)(3,3)(6,6)}

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-

Randomly selected child -1/6  
for the candies E(x) =1/6(1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.012+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

Ans-

The variance of the point is very less so we can say that values in points are not fluctuating too much.

Variance of weigh is very high the values in weigh are fluctuating a lot.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.597 | 3.217 | 17.85 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |
| Variance | 0.29 | 0.96 | 3.19 |
| Standard Deviation | 0.53 | 0.98 | 1.79 |
| Range | 2.17 | 3.911 | 8.4 |

df.describe()

(#will describe the data sets by providing values like mean, std, max and min, etc)

df.mode() (#will show the mode fo each column)

df.var() (#will provide the variance for given dataset for columns with numerical 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?

Ans- EV=Σx/n

One of the patients selected randomly -1/9(108+110+123+134+135+145+167+187+199)

Expected weight - 1/9\*1308

145.33

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

Cars speed and distance Use Q9\_a.csv, Use Q9\_b.csv

Ans-

*Using Q9\_a.csv*

*Speed and Distance*

*The Skewness for Speed is Negative ,i.e. most of the data points are in right side of plot(Right Skewed data)*

*The Kurtosis is negative ,the curve over the plot is wide .*

Summary

|  |  |
| --- | --- |
|  | Speed |
| Skewness | -0.1175 |
| Kurtosis | -0.5774 |

Shape

Description automatically generated

Skewness and Kurtosis for Distance Q9\_a.csv-

The Skewness for distance is Positive, All the data accumulated at right side of the graph

The Kurtosis for distance is Positive

Summary

|  |  |
| --- | --- |
|  | Distance |
| Skewness | 0.8068 |
| Kurtosis | 0.2480 |

A picture containing chart

Description automatically generated

Using Q9\_b.csv

|  |  |
| --- | --- |
|  | Speed(SP) |
| Skewness | 1.5814 |
| Kurtosis | 2.7235 |

Text

Description automatically generated

|  |  |
| --- | --- |
|  | Weight(WT) |
| Skewness | -0.6033 |
| Kurtosis | 0.8194 |

Text

Description automatically generated

Q10) Draw inferences about the following boxplot & histogram

Shape, rectangle

Description automatically generated



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-

The information we have

Sample size N=2000

Average weigh=200

Sample standard deviation=30

Formula=(m - t\*(s/sqrt(N), m + t\*(s/sqrt(N))

Confidence interval for

94%- 143.54 , 256.45

96%- 138.34 , 261.65

98%- 130.15 , 269.84

: AVG\_WGT1 = stats.norm.interval(0.97, loc = 200, scale = 30)

print('Average weight of adult in Mexico at 94% confidence interval', np.round(AVG\_WGT1, 3))

Average weight of adult in Mexico at 94% confidence interval [134.897 265.103]

AVG\_WGT2 = stats.norm.interval(0.99, loc = 200, scale = 30)

print('Average weight of adult in Mexico at 98% confidence interval', np.round(AVG\_WGT2, 3))

Average weight of adult in Mexico at 98% confidence interval [122.725 277.275]

AVG\_WGT3 = stats.norm.interval(0.98, loc = 200, scale = 30)

print('Average weight of adult in Mexico at 96% confidence interval', np.round(AVG\_WGT3, 3))

Average weight of adult in Mexico at 96% confidence interval [130.21 269.79]

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 | 41.00 |
| median | 40.50 |
| variance | 25.53 |
| standarddev | 5.05 |

df = pd.DataFrame(df)

df.mean()

0 41.0

dtype: float64

df.median()

0 40.5

dtype: float64

df.std()

0 5.052664

dtype: float64

df.var()

0 25.529412

dtype: float64

2. mean>median so we can say that the data is positively skeed i.e. many kids got marks above 40,No outlier is present

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

-Symmetric Distribution of data, no skewness

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

- Right skewed(tail on the right side).

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

- Left Skewed(tail on the left side).

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

- peakness (sharp peak) and less variation.

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

- 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?

it is not a Normal Distribution

What is nature of skewness of the data?

It is left skewed.

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

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.



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 of Cars for the below cases

MPG <- Cars$MPG

* 1. P(MPG>38) = 34.66%
  2. P(MPG<40) = 73.06%
  3. P (20<MPG<50) = 94.45%

print("Probabilty that 'MPG' > 38 = ", np.round(1-stats.norm.cdf(38, loc = df4.MPG.mean(), scale = df4.MPG.std()), 3))

Probabilty that 'MPG' > 38 = 0.348

print("Probabilty that 'MPG' < 40 = ", np.round(stats.norm.cdf(40, loc = df4.MPG.mean(), scale = df4.MPG.std(. )), 3))

Probabilty that 'MPG' < 40 = 0.729

print("Probabilty that 20 <'MPG' < 40 = ", np.round((1-stats.norm.cdf

(20, loc = df4.MPG.mean(), scale = df4.MPG.std())) -

(stats.norm.cdf(40, df4.MPG.mean(), scale = df4.MPG.std())) , 3))

Probabilty that 20 <'MPG' < 40 = 0.214

print("Probabilty that 'MPG' < 40 = ", np.round(stats.norm.cdf(70, loc = 60, scale = 10), 5))

Probabilty that 'MPG' < 40 = 0.84134

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Follows Normal distribution as indicated by qq-plot.

Chart, line chart, scatter chart

Description automatically generated

import statsmodels.api as smf

import pylab as py

smf.qqplot(df["MPG"],line='45')

py.show()

b)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

Chart, histogram

Description automatically generated

> import statsmodels.api as smf

import pylab as py

smf.qqplot(df5["Waist"],line='45')

py.show()

Chart

Description automatically generated

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

Z score for 90%=1.281

94%=1.554

60%=0.253

print('Z scores at 90% confidence interval is', np.round(stats.norm.ppf(.95), 2))

print('Z scores at 94% confidence interval is', np.round(stats.norm.ppf(.97), 2))

print('Z scores at 60% confidence interval is', np.round(stats.norm.ppf(.80), 2))

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

Confidence interval For

95%= -2.063, 2.063

96%= -2.171, 2.171

99%= -2.796, 2.796

print(' t scores at 95% confidence interval is', np.round(stats.t.ppf(0.975, df = 24), 2))

print(' t scores at 96% confidence interval is', np.round(stats.t.ppf(0.98, df = 24), 2))

print(' t scores at 99% confidence interval is', np.round(stats.t.ppf(0.995, df = 24), 2))

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

print('critical value = ', np.round(t\_value, 2))

print('probabilty for average life of no more than 260 days is', np.round(stats.t.cdf(t\_value, df=17), 2))

**critical value = -0.47**

**probabilty for average life of no more than 260 days is 0.32**