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

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

**A:** The Probability of getting two heads and one tails in the toss of three coins simultaneously is 3/8 or 0.375.

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

**A:**

1. Equal to 1 **=** 0,

sum of 2 dice numbers will not be equal to 1 assuming if we get ‘1’ from 1st dice and ‘1’ from 2nd dice , sum will always be >=2 then probability is 0

1. Less than or equal to 4 = 6/36= 0.1667,

there are 6 possible outcomes for sum to be 4 divided by total possible outcomes that’s 36 then probability is 0.1667

1. Sum is divisible by 2 and 3 = 5/36= 0.138 ,

The sum divisible by both 2 & 3 are 5 possible numbers divided by total possible outcomes that is 36 then probability is 0.138

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?

**A:** = (5/7)\*(4/6) = (10/21) = 0.47,

Total balls= 7

2 balls are drawn in which none should be blue =7-2=5

Possible outcomes for none balls be blue= 5! / 2! \* 3! = (5\*4) / (2\*1) = 10

2 balls are drawn randomly from bag = 7! / 2! \* 5!

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

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

Total possible outcomes =21, so probability of none being blue balls= 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

**A:** Expected no. of Candies for a randomly selected Child

= (1\*0.015)+(4\*0.20)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.12)

=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**

**A:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column1** | **Column2** | **Column3** | **Column4** |
| Mean | 3.5965625 | 3.21725 | 17.84875 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Column1** | **Column2** | **Column3** | **Column4** |
| Variance | 0.29 | 0.96 | 3.19 |
| STD DEV | 0.53 | 0.98 | 1.79 |
| Range | 2.17 | 3.91 | 8.4 |

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?

**A:** Expected value = Sum (X \* Probability of X)

(1/9)\*(108+110 +123 +134 +135 +145 +167 +187 +199)

= 1308/9 = 145.33

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

**Cars speed and distance**

**A:** Speed has a negative kurtosis, meaning that the distribution is light in the tails compared to the ideal normal distribution. The skewness of the distance is 0.81, which means that the distribution is not symmetrical compared to the ideal normal distribution.

|  |  |  |
| --- | --- | --- |
| **Column1** | **Speed** | **Distance** |
| Skewness | -0.11751 | 0.806895 |
| Kurtosis | -0.50899 | 0.4050526 |

**SP and Weight(WT)**

**Use Q9\_b.csv**

**A:** The SP skewness is 1.61, which is > 1, so the data is skewed to the left and not normally distributed. Even the kurtosis is 2.98, which is > 1, so the data is highgly peaked compared to the normal distribution.

|  |  |  |
| --- | --- | --- |
| **Column1** | **SP** | **Weight** |
| Skewness | 1.6114502 | -0.6147533 |
| Kurtosis | 2.9773289 | 0.9502915 |

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



**A:** The highly positive skewed and peaking kurtosis shown by this histogram



**A:** This boxplot shows six to seven outliers that are dragging the data in their vicinity and affecting our machine learning model.

**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?

**A: import numpy as np**

**import scipy.stats as stats**

**For94 =stats.t.interval(alpha = 0.94, df=1999, loc=200, scale=30/np.sqrt(2000))**

**print(np.round(For94,2)) = [198.74 201.26]**

**For98 =stats.t.interval(alpha = 0.98, df=1999, loc=200, scale=30/np.sqrt(2000))**

**print(np.round(For98,2)) = [198.44 201.56]**

**For96 =stats.t.interval(alpha = 0.96, df=1999, loc=200, scale=30/np.sqrt(2000))**

**print(np.round(For96,2)) = [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.

**A:**

|  |  |
| --- | --- |
| **Column1** | **Column2** |
| mean | 41 |
| median | 40.35 |
| variance | 25.53 |
| standard deviation | 5.05 |

1. What can we say about the student marks?

**A:** Distribution of the Data is Positive Skewed and asymmetrical, with most of the student's scores falling between 34 and 42.

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

**A:**It will be perfectly symmetrical and skewness will be 0.

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

**A:**Skewness will be positive and distribution will be rightly skewed.

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

**A:** Skewness will be negative and distribution will be left skewed.

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

**A:** A positive kurtosis represents a peaked distribution in the center and short tails.

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

**A:**A negative kurtosis represents data distributed more at tails and has thick tails.

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



What can we say about the distribution of the data?

**A:**Upper Quartile Q3 is 18 , Lower Quartile is Q1 is 10 , Median is 15 approx.

What is nature of skewness of the data?

**A:**The data distribution is negatively skewed.

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

**A:** IQR = Q3- Q1= 8

Q19) Comment on the below Boxplot visualizations?



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

**A:** Boxplot 1 shows that the data are tightly distributed between 255 and 275, with few positive peaks. Boxplot 2 shows that the data are normally distributed and that the medians for both boxplots are equal around 265. And finally there are no outliers in the dataset.

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)

**A:** import scipy.stats as stats

cars=pd.read\_csv(“cars.csv”)

np.round(1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std()),3**)** = **0.348**

* 1. P(MPG<40)

**A:** np.round(stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std()),3) **= 0.729**

c. P (20<MPG<50)

**A:** np.round(stats.norm.cdf(50,cars.MPG.mean(),cars.MPG.std()) - stats.norm.cdf(20,cars.MPG.mean(),cars.MPG.std()),3) = **0.899**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**A:**

|  |  |  |
| --- | --- | --- |
|  | Median | Mean |
| HP | 100 | 117.4691 |
| MPG | 35.15273 | 34.42208 |
| VOL | 101 | 98.76543 |
| SP | 118.2087 | 121.5403 |
| WT | 32.73452 | 32.41258 |
| dtype: | float64 | float64 |

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

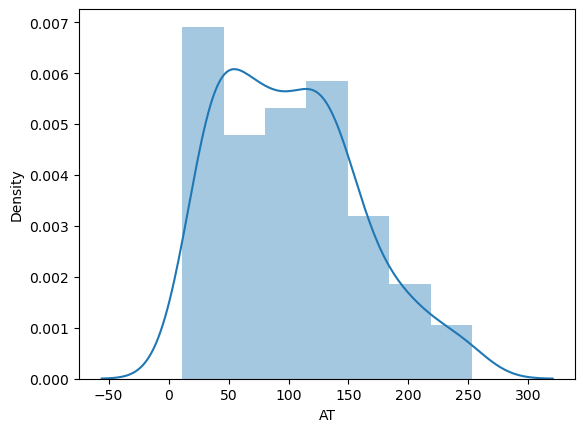
Dataset: wc-at.csv

**A:** import pandas as pd

import seaborn as sns

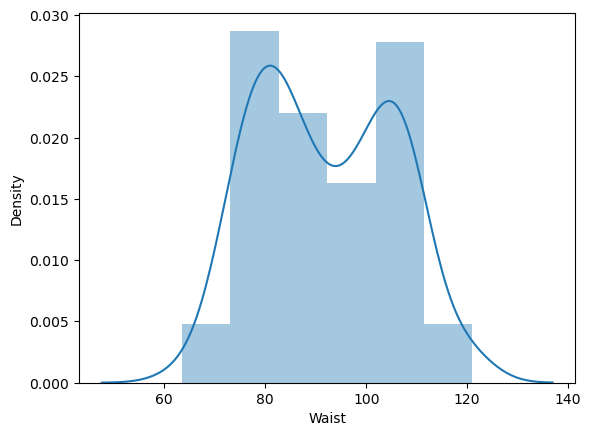
wc\_at=pd.read\_csv(‘wc-at.csv’)

sns.distplot(wc\_at.AT)

****

Adipose Tissue (AT) data is left skewed which is little negative skewness as we can see mean(101.89) is far away from median(96.54).

**sns.distplot(wc\_at.Waist)**

****

Waist Circumference distribution is normally distributed as we can see mean(91.90) and median(90.8) are near to each other

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

**A:** Z-score for 90% confidence interval

np.round(stats.norm.ppf(0.95),2)

= **1.64**

Z-score for 94% confidence interval

np.round(stats.norm.ppf(0.97),2)

= **1.88**

Z-score for 60% confidence interval

np.round(stats.norm.ppf(0.8),2)

= **0.84**

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

**A:** t scores for 95% confidence interval for sample size of 25

stats.t.ppf(0.975,24) # df = n-1 = 24

= **2.06**

t scores for 96% confidence interval for sample size of 25

stats.t.ppf(0.98,24)

= **2.171**

t scores for 99% confidence interval for sample size of 25

stats.t.ppf(0.995,24)

= **2.79**

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

**A:** LetsAssume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days

and Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days

find t-scores at x=260; t=(s\_mean-p\_mean)/(std/sqrt(n))

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

t= -0.47

Find P(X>=260) for null hypothesis

p\_value=1-stats.t.cdf(abs(t\_scores),df=n-1)

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

p\_value = 0.321

probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 0.321