|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Countable |
| Results of rolling a dice | Countable |
| 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 | Nominal |
| Number of kids | Countable |
| Number of tickets in Indian railways | Countable |
| Number of times married | Countable |
| Gender (Male or Female) | Nominal |

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

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

Ans:

The probability of two heads and one tail are 3/8🡺0.375

The events that occur are {TTT, HHH, HHT, HTH, HTT, THH, TTH, THT}

The possible outcomes are three (3) out of eight events (8)

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

1. Equal to 1

Ans: Probability =0 (since minimum sum of two dice is equal to 2 for non-possible event the probability is zero)

1. Less than or equal to 4

Ans: Probability (x<=4)

Possible outcomes for x<=4 are {1,1;1,2;1,3;2,1;2,2;3,1}

So, probability is 6/36 🡺0.16

1. Sum is divisible by 2 and 3

The only numbers that are divisible by both 2 & 3 are 6 &12

Possible outcomes (x/2&3)🡺{3,3;2,4;5,1;6,6;4,2;1,5}

So, probability is 6/36🡺0.16

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: Let the total number of balls = (2+3+2)🡺7

Let S be the sample space

Then, n(S)=Number of Ways of Drawing 2 balls out of 7

=7C2

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

=21

Let E= event of drawing 2 balls none of then is blue

n(E) = Number of ways of Drawing 2 balls out of (5 balls)

=5C2

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

=10

P(E) =n(E)/n(S) =10/21

=10/21

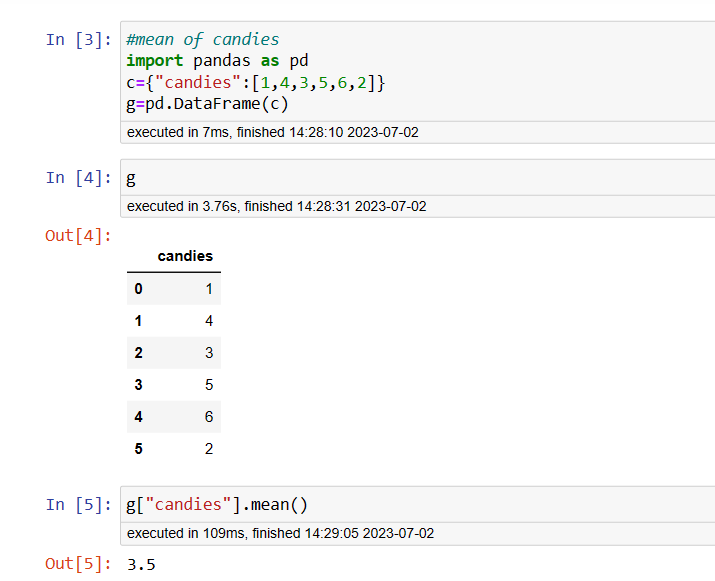
Q6) Calculate the Expected number of candies for a randomly selected child

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



Inference:

The expected number of candies for a randomly selected child is the mean of the candies count is 3.5

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”**

import pandas as pd

df = pd. read\_csv("Q7.csv")

df["Points"]. describe ()

output:

count 32.000000

mean 3.596563

std 0.534679

min 2.760000

25% 3.080000

50% 3.695000

75% 3.920000

max 4.930000

Name: Points, dtype: float64

P1 = df["Points"]. median ()

P2 = df["Points"]. mode ()

P3 = df["Points"]. var ()

P4 = (df["Points"].max())-(df["Points"]. min ())

print(P1)

print(P2)

print(P3)

print(P4)

output:

3.5965625000000006

3.6950000000000003

0 3.07

1 3.92

Name: Points, dtype: float64

0.28588135080645166

“Score”

import pandas as pd

“Score”

df = pd.read\_csv("Q7.csv")

df["Score"]. describe ()

output:

count 32.000000

mean 3.217250

std 0.978457

min 1.513000

25% 2.581250

50% 3.325000

75% 3.610000

max 5.424000

Name: Score, dtype: float64

S1 = df["Score"]. median ()

S2 = df["Score"]. mode ()

S3 = df["Score"]. var ()

print(S1)

print(S2)

print(S3)

output:

3.325

0 3.44

Name: Score, dtype: float64

0.9573789677419356

In [16]:

[ “Weigh”]

import pandas as pd

df = pd. read\_csv("Q7.csv")

df["Weigh"]. describe ()

count 32.000000

mean 17.848750

std 1.786943

min 14.500000

25% 16.892500

50% 17.710000

75% 18.900000

max 22.900000

Name: Weigh, dtype: float64

W1 = df["Weigh"]. median ()

W2 = df["Weigh"]. mode ()

W3 = df["Weigh"]. var ()

print(W1)

print(W2)

print(W3)

output:

17.71

0 17.02

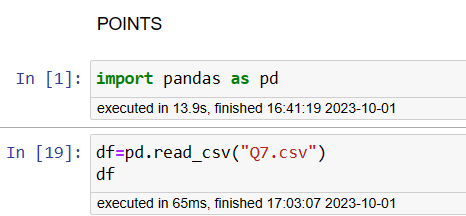
1 18.90

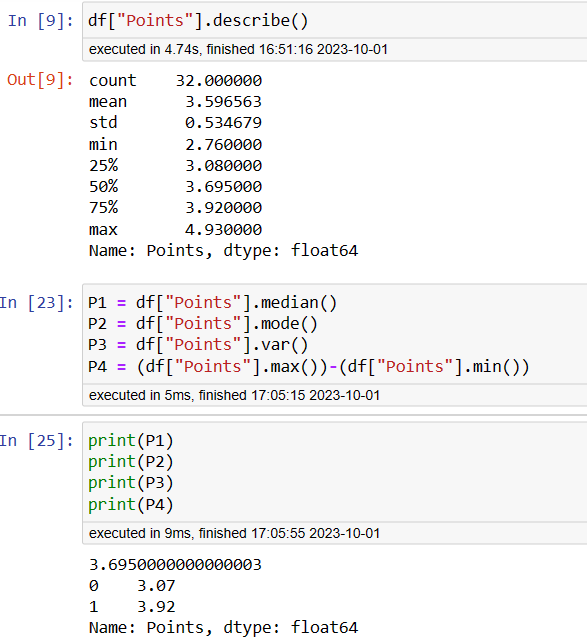
Name: Weigh, dtype: float64

3.193166129032258

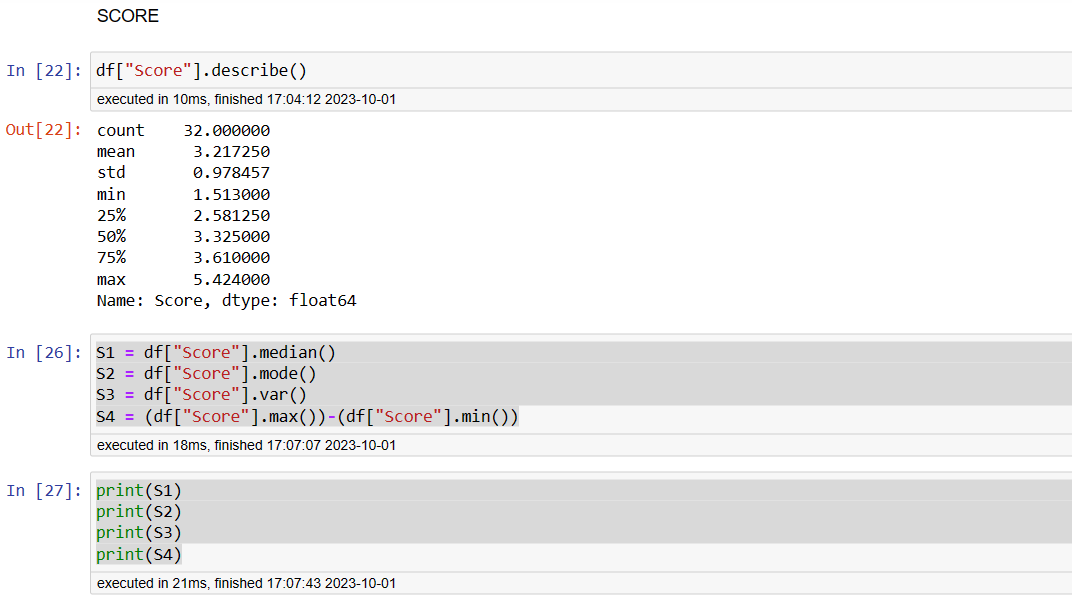
**INFERENCES.**

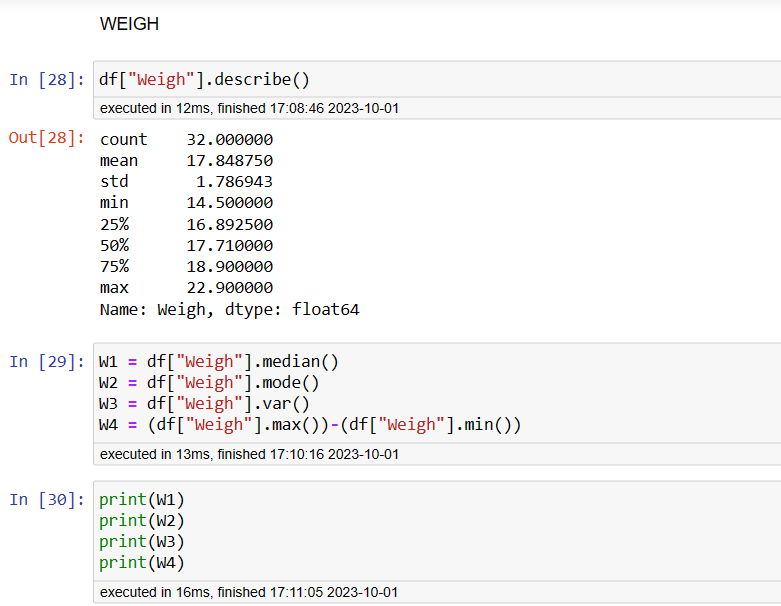
The values of Mean, Median, Mode, Variance, Standard Deviation, and Range is

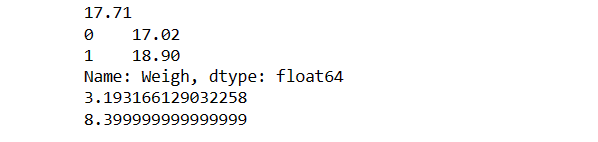
calculated for the variables (Points, Score and Weigh) and values are obtained****

****

****

****





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: import pandas as pd

Import numpy as np

df= {“Weigh”. [108,110,123,134,135,145,167,187,199]}

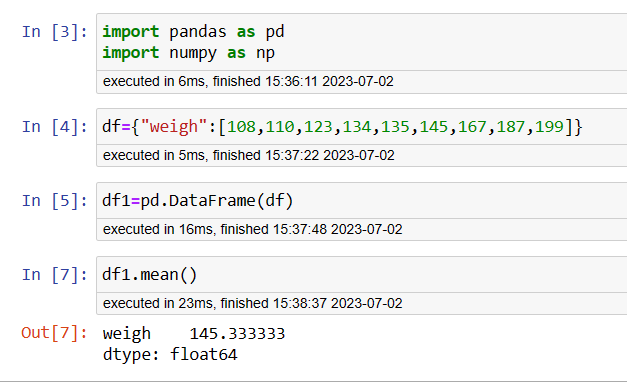
df1=pd.DataFrame(df)

df1[“Weigh”. Mean ()]

output:

Weigh :145.3333

dtype: float64



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

**Cars speed and distance**

**Use Q9\_a.csv**

import pandas as pd

df = pd. read\_csv("D:/assingment/Q9\_a.csv")

“speed”

(df["speed"]. kurt ()). round (2)

(df["speed"]. skew ()). round (2)

Output:

-0.51

-0.12

“Distance”

df = pd. read\_csv("D:/assingment/Q9\_a.csv")

(df["dist"]. kurt ()). round (2)

(df["dist”]. skew ()). round (2)

Output:

0.41

0.81

**INFERENCES:**

The values for kurtosis and skewness of distance and speed is obtained and observed.

**SP and Weight(WT)**

**Use Q9\_b.csv**

“SP”

import pandas as pd

df = pd.read\_csv("D:/assingment/Q9\_b.csv")

df

(df["SP"].kurt()).round(2)

(df["SP"].skew()).round(2)

Output:

2.98

1.61

INFERENCES:

The values for kurtosis and skewness of SP is obtained and observed.

“WT”

import pandas as pd

df = pd.read\_csv("D:/assingment/Q9\_b.csv")

df

(df["WT"].kurt()).round(2)

(df["WT"].skew()).round(2)

Output:

0.95

-0.61

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



Answer:

Histogram: - Chick weight data is right skewed or positively skewed.---- Yes

- More than 50% Chick Weight is between 50 to 150. ---- Yes

- Most of the chick weight is between 50 to 100. --- Yes



* The data is right side skewed
* The outliers are moved upper side.

**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?

Answer:

from scipy import stats

a= stats.norm.interval(0.94, loc=200, scale=30)

print(a)

b = stats.norm.interval(0.96, loc = 200, scale=30)

print(b)

c = stats. norm. Interval(0.92, loc = 200, scale =30)

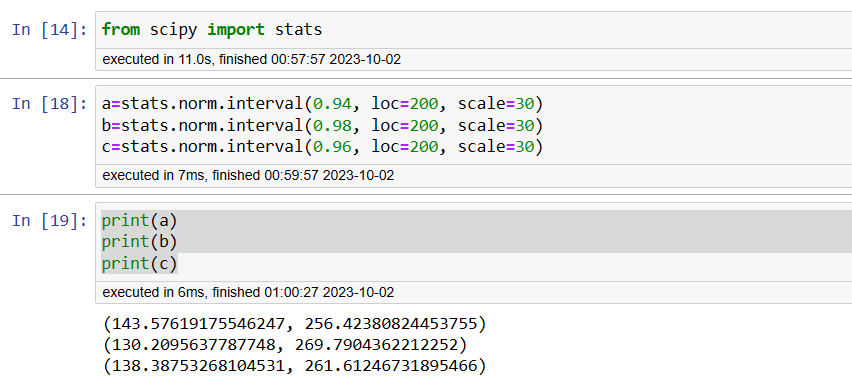
Print(c)

Output:

(143.57619175546247, 256.42380824453755)

(138.38753268104531, 261.61246731895466)

(130.2095637787748, 269.7904362212252)



**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:

import numpy as np

import pandas as pd

df = {“Score”: [ 34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]}

df1 = pd. Data Frame(df)

print(df1["Score"])

print((df1[“Score”]). mean ())

print((df1[“Score”]). median ())

print((df1[“Score”]). var ())

print((df1[“Score”]). std ())

output:

41.0

40.5

25.529411764705884

5.05266382858645

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

Ans: It is symmetrically distributed

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

Ans: Nature of skewness is positively skewed

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

Ans: Nature of skewness is negatively skewed

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

Ans: It indicates the peakedness of the graph ( it has high peak value)it is above

Symmetrical

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

Ans: It indicates the low peakedness of the graph .it is below symmetrical, peak value is low

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



What can we say about the distribution of the data?

Ans: The data is distributed between 10-18

What is nature of skewness of the data?

Ans: The nature of the skewness is negatively skewed because most of the mean is lied at the right side of the box plot

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

Ans: IQR=(Q3-Q1)

Q3=18, Q1=10

IQR= 18-10🡺8

Q19) Comment on the below Boxplot visualizations?



Ans:

There are no outliers in both the Box plot. Both the box plots share the same median that is approximately in the range of 250 to 275 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range

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

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)
  2. P(MPG<40)

c. P (20<MPG<50)

A solution: P(MPG>38)

From scipy. stats import norm

df = pd. read\_csv ("D:/assingment/Cars (1).csv")

a = df["MPG"]. mean ()

b = df["MPG"]. std ()

c = norm(a,b)

e = 1-c.cdf (38)

print(e)

Output:

0.3475939251582705

B solution: P(MPG<40)

df = pd. read\_csv ("D:/assingment/Cars (1).csv")

a = df["MPG"]. mean ()

b = df["MPG"]. std ()

c = norm(a,b)

e = 1-c.cdf (40)

print(e)

output:

0.27065012378483844

C solution: P (20<MPG<50)

df = pd. read\_csv ("D:/assingment/Cars (1).csv")

a = df["MPG"]. mean ()

b = df["MPG"]. std ()

c = norm(a,b)

e = 1-c.cdf (50)

f = 1-c.cdf (20)

g = e-f

print(g)

output:

0.8988689169682046

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans: the MPG of the car follows the 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

Ans: Adipose tissue (AT) and waist circumference (Waist)does not follow the normal distribution

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

Answer :

from scipy. stats import norm

from scipy import stats

# z score for 90% confidence interval

Stats.norm.ppf (0.95)

Output:

1.6448536269514722

# z score for 94% confidence interval

Stats.norm.ppf (0.97)

Output:

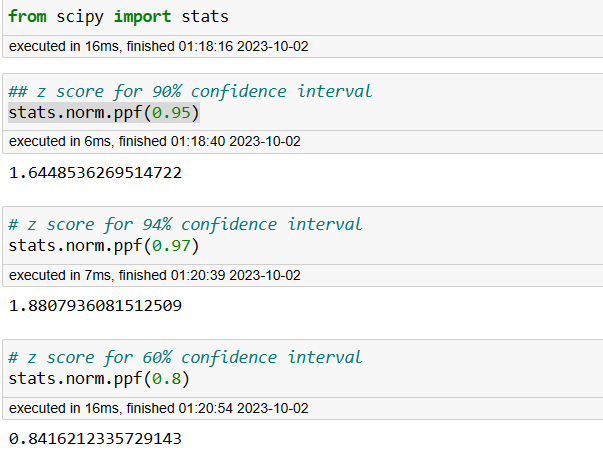
1.8807936081512509

# z score for 60% confidence interval

Stats.norm.ppf (0.8)

Output:

0.8416212335729143



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

Ans:

From scipy import stats

From scipy. stats import norm

# t scores of 95% confidence interval for sample size of 25

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

output:

2.0638985616280205

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

Stats.t.ppf (0.98,24)

Output:

2.1715446760080677

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

Stats.t.ppf (0.995,24)

Output:

2.796939504772804

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:

From scipy import stats

From scipy.stats import norm

# find t-scores at x=260; t=(s\_mean-P\_mean)/(s\_SD/sqrt(n))

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

t

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

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

p\_value

# OR p\_value=stats.t.sf(abs(t\_score),df=n-1)... Using sf function

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

p\_value

# Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days

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

Output:

-0.4714045207910317

0.32167411684460556

0.32167411684460556