Q1) Identify the Data type for the Following:

Activity	Data Type
Number of beatings from Wife	Discrete data type
Results of rolling a dice	Discrete data type
Weight of a person	Continuous data type
Weight of Gold	Continuous data type
Distance between two places	Continuous data type
Length of a leaf	Continuous data type
Dog's weight	Continuous data type
Blue Color	Discrete data type
Number of kids	Discrete data type
Number of tickets in Indian railways	Discrete data type
Number of times married	Discrete data type
Gender (Male or Female)	Discrete data type

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	Nominal
Fahrenheit Temperature	Ratio
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	Interval
Number of Children	Interval

Religious Preference	Nominal
Barometer Pressure	Ratio
SAT Scores	Interval
Years of Education	Nominal

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

Soln:- The total possible outcome 2^3 = 8

HHH,HHT,HTT,THT,TTH,HTH,THH,TTT

Number of favorable outcomes =3
P(two head and one tail)=3/8 = 0.375

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

a) Equal to 1
b) Less than or equal to 4
c) Sum is divisible by 2 and 3

Soln:-

Two dice are thrown here therefore, the set of possible outcomes when we roll a die are {1, 2, 3, 4, 5, 6}

The total possible outcome 2^6 = 36

a) the sum is equal to 1 is zero because they starts with (1,1) i.e there are zero possible outcomes of getting 1 as a sum.

P(sum equal to 1) = 0

b) When we roll two dice, the possibility of getting number 4 is (1, 3), (2, 2), and (3, 1).

P(sum less than 4) = 3 / 36 = 1 / 12

c) When we roll two dice, the possibility of getting number sum divisible by 2 and 3 is (1 , 5) , (3 , 3) , (4 , 2) , (5 , 1) , (6 , 6)

P(sum is divisible by 2 and 3) = 5 / 36

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?
Soln:-

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
n(S) = 7C2

n(S) = (7×6) / (2×1)

n(S) = 21

Let E = Event of 2 balls, none of which is blue
∴ n(E) = Number of ways of drawing 2 balls out of (2 + 3) balls
n(E) = 5C2

n(E) = (5×4) / (2×1)

n(E) = 10

∴P(E) = n(E) / n(S) = 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

Soln:-

Expected number of candies for randomly selected child = 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

Soln:-

Points: Mean = 3.596563, Median= 3.695, Mode = "numeric", Variance= 0.2858814, Standard deviation= 0.5346787, Range = 2.17

Score: Mean = 3.21725, Median = 3.325, Mode = "numeric", Variance= 0.957379, Standard deviation = 0.9784574, Range = 3.91

Weight: Mean = 17.84875, Median= 17.71, Mode= "numeric", Variance= 3.193166, Standard deviation= 1.786943, Range = 8.39

Comments / Inferences:-

- Mean value are closer for both 'Point' and 'Score'.
- Points do not contain any outliers.
- Score contains 3 outliers.
- Weight contain 1 outlier.

Q8) Calculate Expected Value for the problem below

a) 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?

Soln:-

Expected Value = $\sum$ (probability * Value)

$\sum P(x).E(x)$

There are 9 patients

Probability of selecting each patient = 1/9

Ex = 108, 110, 123, 134, 135, 145, 167, 187, 199

P(x) = 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9

Expected Value = (1/9)(108) + (1/9)110 + (1/9)123 + (1/9)134 + (1/9)135 + (1/9)145 + (1/9(167) + (1/9)187 + (1/9)199

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

= (1/9) (1308)

= 145.33

Expected Value of the Weight of that patient = 145.33

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

Cars speed and distance

Use Q9_a.csv

Soln:-

```
In [4]: df.skew()
Out[4]: Index    0.000000
        speed   -0.117510
        dist     0.806895
        dtype: float64

In [5]: df.kurtosis()
Out[5]: Index   -1.200000
        speed   -0.508994
        dist     0.405053
        dtype: float64
```

For speed:

Skewness = -0.117510, skewness value is negative so it is left skewed.

Kurtosis = -0.508994, negative kurtosis

For distance:

Skewness = 0.8068, skewness value is positive so it is right skewed.

Kurtosis = 0.405053, positive kurtosis

SP and Weight(WT)

Use Q9_b.csv

```
In [4]: df.skew()
Out[4]: Unnamed: 0     0.000000
        SP             1.611450
        WT            -0.614753
        dtype: float64

In [5]: df.kurtosis()
Out[5]: Unnamed: 0    -1.200000
        SP             2.977329
        WT             0.950291
        dtype: float64
```

For SP:

Skewness = 1.61145, skewness value is positive so it is right skewed.

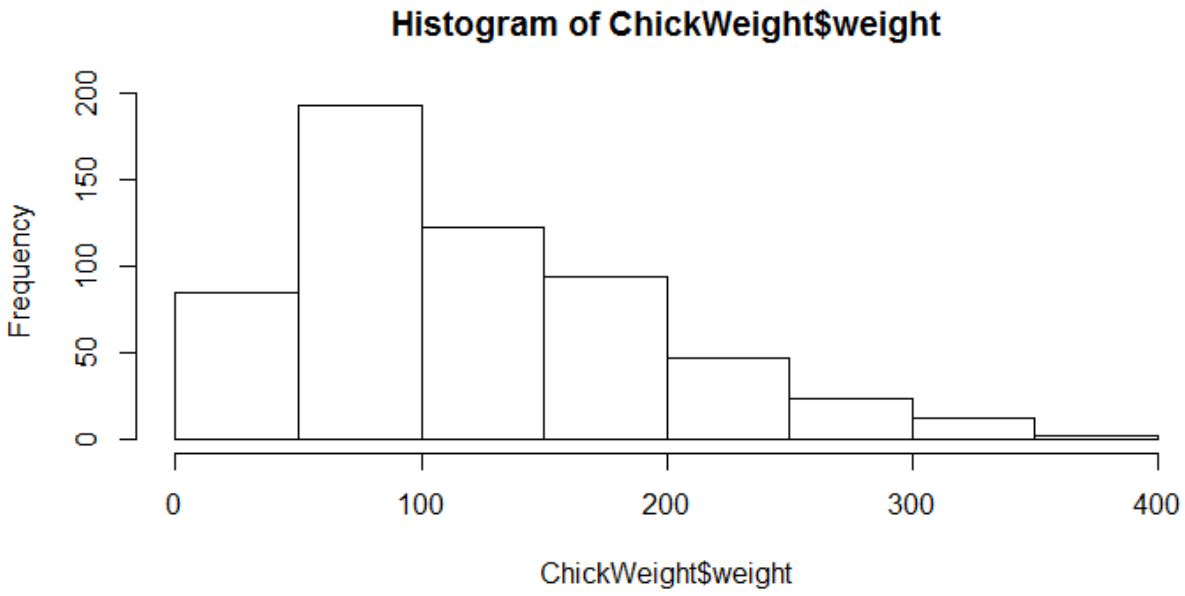
Kurtosis = 2.977329, positive kurtosis

For Weight:

Skewness = -0.61475, skewness value is negative so it is left skewed.

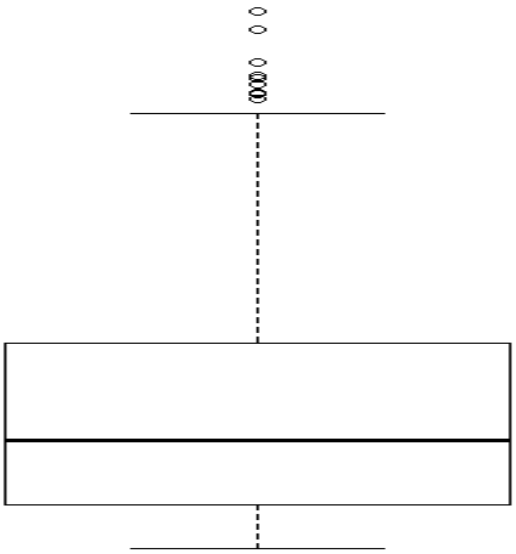
Kurtosis = 0.950291, positive kurtosis

Q10) Draw inferences about the following boxplot & histogram



Soln:

- The most of the data points are concerated in the range 50-100 with frequency 200.
- And least range of weight is 400 somewere around 0-10. So the expected value the above distribution is 75.
- Skewness- We can notice a long tail towards right so it is heavily right skewed.



Median is less than mean right skewed and we have outlier on the upper side of box plot and there is less data points between Q1 and bottom point.

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?

Soln:

```
In [1]:
import numpy as np
import pandas as pd
from scipy import stats
from scipy.stats import norm

In [13]:
# Avg. weight of Adult in Mexico with 94% CI
stats.norm.interval(0.94,200,30/(2000**0.5))

Out[13]: (198.738325292158, 201.261674707842)

In [14]:
# Avg. weight of Adult in Mexico with 98% CI
stats.norm.interval(0.98,200,30/(2000**0.5))

Out[14]: (198.43943840429978, 201.56056159570022)

In [15]:
# Avg. weight of Adult in Mexico with 96% CI
stats.norm.interval(0.96,200,30/(2000**0.5))

Out[15]: (198.62230334813333, 201.37769665186667)

In [ ]:
```

- Sample mean of $\overline{x} = 200$.
- Sample standard deviation of $s = 30$.
- Sample size of $n = 2000$.

The interval is:

$\overline{x} \pm t\frac{s}{\sqrt{n}}$

- In which t is the critical value for the two-tailed confidence interval.

Considering a 94% confidence level, using a calculator, with 200 - 1 = 199 dof, the critical value is t = 1.8916, hence:

$\overline{x} - t\frac{s}{\sqrt{n}} = 200 - 1.8916\frac{30}{\sqrt{2000}} = 198.73$

$\overline{x} + t\frac{s}{\sqrt{n}} = 200 + 1.8916\frac{30}{\sqrt{2000}} = 201.27$

The 94% confidence interval is (198.73, 201.27).

Considering a 96% confidence level, using a calculator, with 200 - 1 = 199 dof, the critical value is t = 2.0673, hence:

$\overline{x} - t\frac{s}{\sqrt{n}} = 200 - 2.0673\frac{30}{\sqrt{2000}} = 198.61$

$\overline{x} + t\frac{s}{\sqrt{n}} = 200 + 2.0673\frac{30}{\sqrt{2000}} = 201.39$

The 96% confidence interval is (198.61, 201.39).

Considering a 98% confidence level, using a calculator, with 200 - 1 = 199 dof, the critical value is t = 2.3452, hence:

$$\overline{x} - t\frac{s}{\sqrt{n}} = 200 - 2.3452\frac{30}{\sqrt{2000}} = 198.43$$

$$\overline{x} + t\frac{s}{\sqrt{n}} = 200 + 2.3452\frac{30}{\sqrt{2000}} = 201.57$$

The 98% confidence interval is (198.43, 201.57).

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

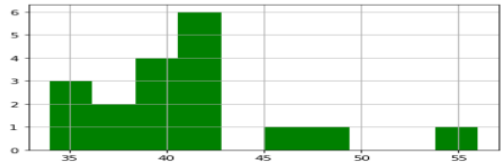
1)Find mean, median, variance, standard deviation.

```
In [45]: import numpy as np
         import pandas as pd
         from matplotlib import pyplot as plt
         import statistics

In [46]: marks=[34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

In [47]: statistics.mean(marks)

Out[47]: 41

In [48]: statistics.median(marks)

Out[48]: 40.5

In [49]: statistics.mode(marks)

Out[49]: 41

In [50]: statistics.stdev(marks)

Out[50]: 5.05266382858645

In [51]: statistics.variance(marks)

Out[51]: 25.529411764705884
```

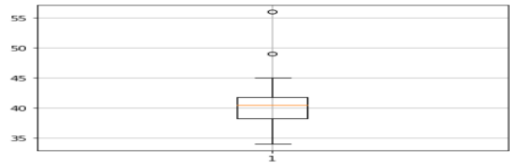
Mean= 41, Median= 40.5, variance= 24.111, Standard deviation= 5.05

2)What can we say about the student marks?

```
In [63]: plt.hist(marks, color='g')
         plt.grid()
         plt.show()
```




```
In [66]: plt.boxplot(marks)
         plt.grid()
         plt.show()
```



From above plot we can say that mean of marks of student is 41 which is slightly greater than median.

- Most of the students got marks in between 41-42.
- There are two outlier 49,56.

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

Soln: Symmetrical

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

Soln: Right skewed

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

Soln: Left skewed

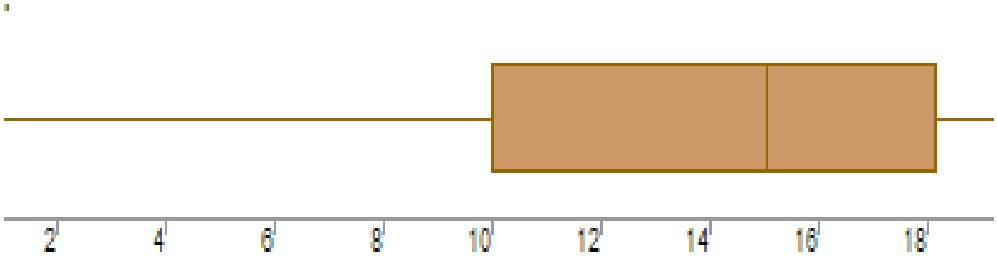
Q16) What does positive kurtosis value indicates for a data ?

Soln: The data is normally distributed and kurtosis value is 0.

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

Soln: The distribution of the data has lighter tails and a flatter peaks than the normal distribution.

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



What can we say about the distribution of the data?

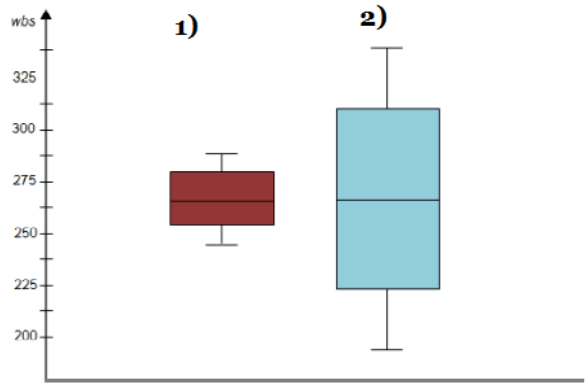
- Let's assume above box plot is about age's of the students in a school.
- 50% of the people are above 10 yrs old and remainig are less.
- And students who's age is above 15 are approx 40%.

What is nature of skewness of the data?

Left skewed, median is greater than mean.

What will be the IQR of the data (approximately)?
Approximately= -8

Q19) Comment on the below Boxplot visualizations?



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

Soln:

- By observing both the plots whisker's level is high in boxplot 2.
- Mean and median are equal hence distribution is symmetrical.

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

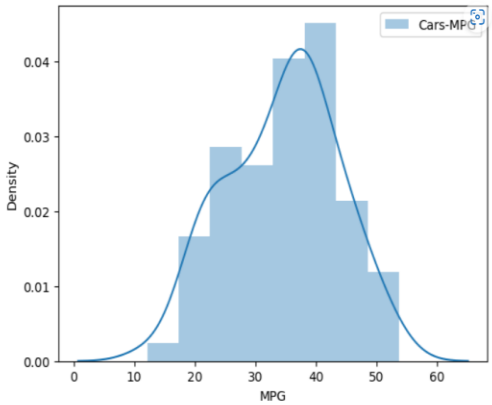
a. P(MPG>38)
b. P(MPG<40)
c. P (20<MPG<50)

```
[47]
# P(MPG>38)
1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std())

0.3475939251582705

[45]
# P(MPG<40)
stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std())

0.7293498762151616

[46]
# P (20<MPG<50)
stats.norm.cdf(0.50,cars.MPG.mean(),cars.MPG.std())-stats.norm.cdf(0.20,cars.MPG.mean(),cars.MPG.std())

1.2430968797327613e-05
```

Q 21) Check whether the data follows normal distribution

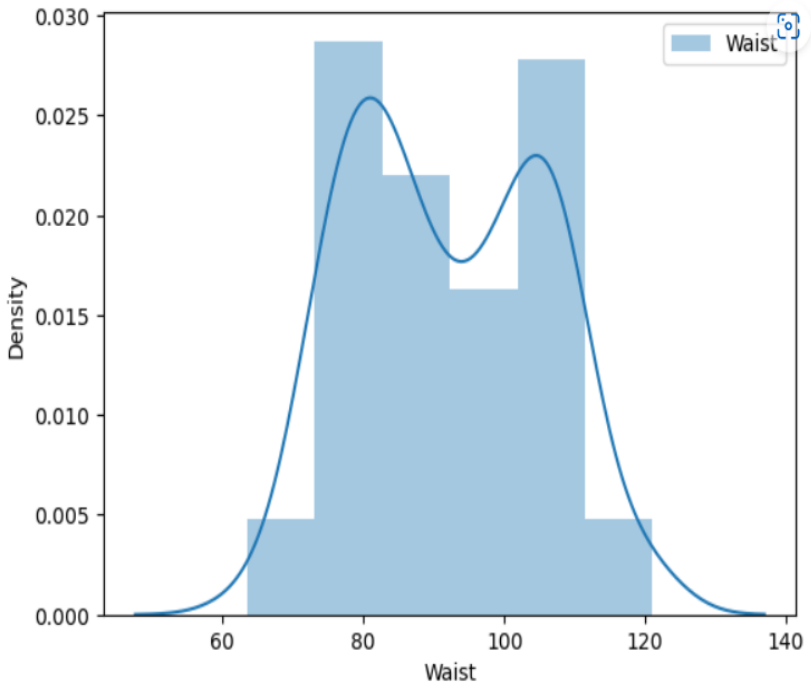
a) Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv




```
In [11]: df.MPG.mean()
Out[11]: 34.422075728024666

In [12]: df.MPG.median()
Out[12]: 35.15272697
```

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



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

Soln:

For 90% confidence interval:

We have the significance level at 5 % (as it is a two tailed test)

that is:

$\alpha = 5\ \% = 0.05$

z at $\alpha = 0.05$ from the z table will be:

$z = 1.645$.

For 94 % confidence interval, we get:

We have the significance level at 3 % (as it is a two tailed test)

that is:

$\alpha = 3\ \% = 0.03$

z at $\alpha = 0.03$ from the z table will be:

$z = 1.555$.

For 60 % confidence interval, we get:

We have the significance level at 20 % (as it is a two tailed test)

that is:

$\alpha = 20\ \% = 0.2$

z at $\alpha = 0.2$ from the z table will be:

$z = 0.253$

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

Soln:

```
In [12]: # t scores of 95% confidence interval for sample size of 25
         stats.t.ppf(0.975,24)  # df = n-1 = 24

Out[12]: 2.0638985616280205

In [13]: # t scores of 96% confidence interval for sample size of 25
         stats.t.ppf(0.98,24)

Out[13]: 2.1715446760080677

In [14]: # t scores of 99% confidence interval for sample size of 25
         stats.t.ppf(0.995,24)

Out[14]: 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

Soln:

In [7]:
```
from scipy import stats
from scipy.stats import norm
```

In [8]:
```
# Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days
# Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days
```

In [9]:
```
# find t-scores at x=260; t=(s_mean-P_mean)/(s_SD/sqrt(n))
t=(260-270)/(90/18**0.5)
t
```

Out[9]:
```
-0.4714045207910317
```

In [10]:
```
# Find P(X>=260) for null hypothesis
```

In [11]:
```
# 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
```

Out[11]:
```
0.32167411684460556
```