

Activity	Data Type
Number of beatings from Wife	Discrete / Count
Results of rolling a dice	Discrete / Count
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	Categorical
Number of kids	Discrete / Count
Number of tickets in Indian railways	Discrete / Count
Number of times married	Discrete / Count
Gender (Male or Female)	Binary

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	Nominal
Celsius Temperature	Interval
Weight	Ordinal
Hair Color	Nominal
Socioeconomic Status	Ordinal
Fahrenheit Temperature	Interval
Height	Ordinal
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	Ratio

Number of Children	Nominal
Religious Preference	Nominal
Barometer Pressure	Ratio
SAT Scores	Ordinal
Years of Education	Ordinal

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

probability of one head= $\frac{1}{2}$ , Probability of two heads= $\frac{1}{2}+\frac{1}{2}$

Probability of one tail= $\frac{1}{2}$

= $\frac{1}{2}+\frac{1}{2}+\frac{1}{2}=\frac{3}{8}$

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

a) The sum is equal to 1 is zero

b) Less than or equal to 4

(1,3),(2,2),(3,1) therefore  $n(b) = 3/36 = 1/12$

c) When 2 dice are rolled find the probability of getting a sum divisible by 3 - 12.

d) When 2 dice are rolled find the probability of getting a sum divisible by 2 - 0.5

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?

Probability of two balls drawn are blue  $P(b)=\frac{2}{5}$

none of the balls drawn are blue= $1-p(b)=1-\frac{2}{5}=0.6$

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

$$= 1 * 0.015 + 4 * 0.20 + 3 * 0.65 + 5 * 0.005 + 6 * 0.01 + 2 * 0.12$$

$$= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24$$

$$= 3.090$$

$$= 3.09$$

Expected number of candies for a randomly selected child = 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**

Ferrari 302	3.54	3.57	14.6
Maserati E	3.54	3.57	14.6
Volvo 142	4.11	2.78	18.6
Mean	3.596563	3.21725	17.84875
Median	3.695	3.325	17.71
Range	2.17	3.911	8.4
Mode	3.92	3.44	17.02
Variance	0.285881	0.957379	3.193166
Standard d	0.523431	0.972369	1.768883

### Comments:

- **There were outliers in the Weights as there variance and Standard deviation**

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?

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

Answer: Here i mentioned the sample screen shots and i have given skewness

Index	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10
7	10	18
8	10	26

44	22	66
45	23	54
46	24	70
47	24	92
48	24	93
49	24	120
50	25	85
-3.77627E-17	-0.11751	0.806895

- Inferences is that distance has positive skewness and Speed and index has negative skewness

### Kurtosis

- Index -1.200000
- speed -0.508994
- dist 0.405053
- Speed distribution is negative kurtosis i.e. flatter than normal distribution)
- Distance distribution is positive kurtosis i.e. peaked than normal distribution)

### SP and Weight(WT)

Use Q9\_b.csv

A	B	C	D
	SP	WT	
1	104.1854	28.76206	
2	105.4613	30.46683	
3	105.4613	30.1936	
4	113.4613	30.63211	
5	104.4613	29.88915	
6	113.1854	29.59177	
7	105.4613	30.30848	
8	102.5985	15.84776	
9	102.5985	16.35948	
10	115.6452	30.92015	
11	111.1854	29.36334	
12	117.5985	15.75353	

70	158.3007	37.14173
71	164.5985	15.82306
72	133.416	44.01314
73	133.1401	43.35312
74	124.7152	52.99775
75	121.8642	42.6187
76	132.8642	42.77822
77	169.5985	16.13295
78	150.5766	37.92311
79	151.5985	15.76963
80	167.9445	39.4231
81	139.8408	34.94861
Skewness	1.61145	-0.61475
Kurtosis	2.977329	0.950291

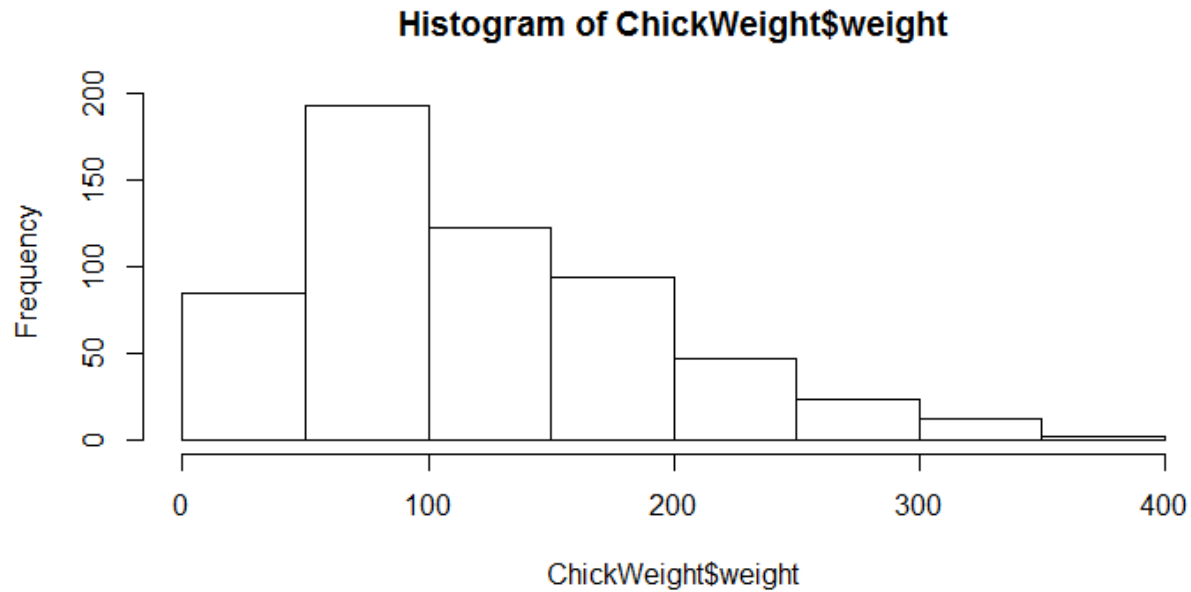
- Inferences is that Speed has positive skewness and Distance has negative skewness

#### Comments:

- Weight distribution is Positive kurtosis i.e. peaked than normal distribution)
- Speed distribution is positive kurtosis i.e. peaked than normal distribution)

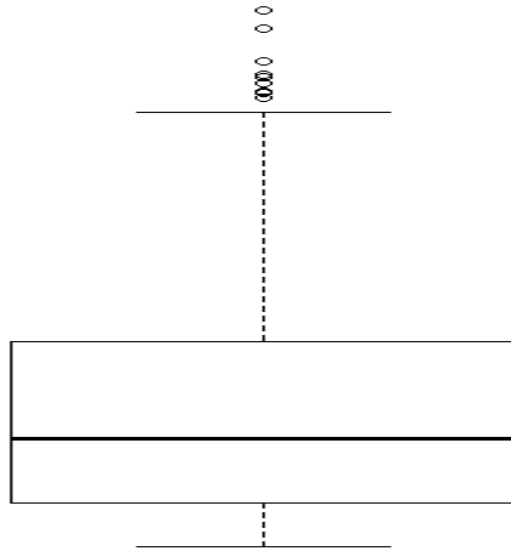
**Q10) Draw inferences about the following boxplot & histogram**

---



Inferences:

- when the chickWeight\$weight is in the range of 50-100 then the frequency is in peak 200
- when ChickWeight\$weight is in the range of 350-400 then the frequency is in zero 0
- when ChickWeight\$weight is in the range of 0-50 then the frequency is in 80
- when ChickWeight\$weight is in the range of 100-150 then the frequency is in zero 130
- The curve looks like it has positive skewness or right side skewed



Inferences:

- There are outliers after the upper extreme or upper whisker

**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?

*# Avg. weight of Adult in Mexico with 94% CI*

```
stats.norm.interval(0.94,200,30/(2000**0.5))
```

```
(198.738325292158, 201.261674707842)
```

*# Avg. weight of Adult in Mexico with 98% CI*

```
stats.norm.interval(0.98,200,30/(2000**0.5))
```

```
(198.43943840429978, 201.56056159570022)
```

*# Avg. weight of Adult in Mexico with 96% CI*

```
stats.norm.interval(0.96,200,30/(2000**0.5))
```

```
(198.62230334813333, 201.37769665186667)
```



**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?

34	36	36	38	38	39	39	40	40	41	41	41	41	42	42	45	49	56	Median
																		40.5
																		Mean
																		41
																		Variance
																		25.5294
																		Standard
																		5.05266

Mean:41

Median:40.3

Variance: 25.5204

Standard Deviation:5.05266

2. The average student marks are 41

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

If the distribution is symmetric, then the mean is equal to the median, and the distribution has zero skewness

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

If the mean is greater than the median, the distribution is positively skewed.

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

If the median is greater than the mean, the distribution is negatively skewed.

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

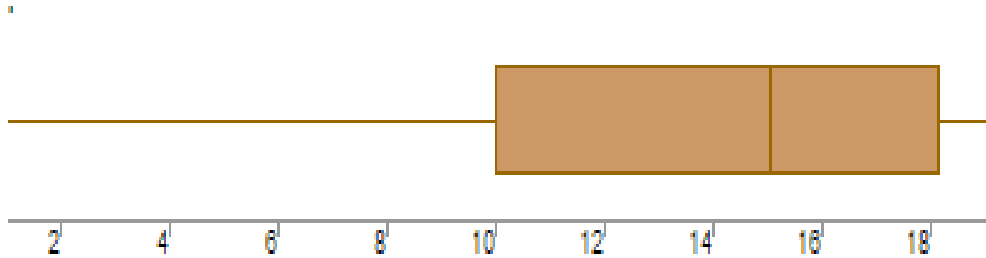
Positive excess values of kurtosis (>3) indicate that a distribution is peaked and possess thick tails.A leptokurtic distribution has a higher peak (thin bell) and taller (i.e. fatter and heavy) tails than a normal distribution.

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

Negative excess values of kurtosis (<3) indicate that a distribution is flat and has thin tails. A platykurtic distribution is flatter (less peaked) when compared with

the normal distribution, with fewer values in its shorter (i.e. lighter and thinner) tails.

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



What can we say about the distribution of the data?

Lower Quartile was 10

Median was 15

Upper Quartile was 18

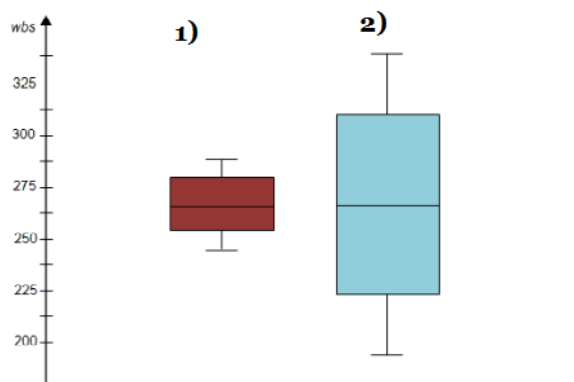
What is nature of skewness of the data?

It is negative skewness

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

IQR:  $Q3 - Q1: 18 - 10 = 8$

Q19) Comment on the below Boxplot visualizations?



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

Inferences:

Boxplot 1 and Boxplot 2 has a median of 262.5

Lower Extreme for Boxplot 1 is 250 and Boxplot 2 is 200

upper extreme for Boxplot 1 is 300 and Boxplot 2 is 350

upper Quartile for Boxplot 1 is 275 and Boxplot 2 is 312.5

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.  $P(\text{MPG} > 38)$
- b.  $P(\text{MPG} < 40)$
- c.  $P(20 < \text{MPG} < 50)$

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

0.3475939251582705
```

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

0.7293498762151616
```

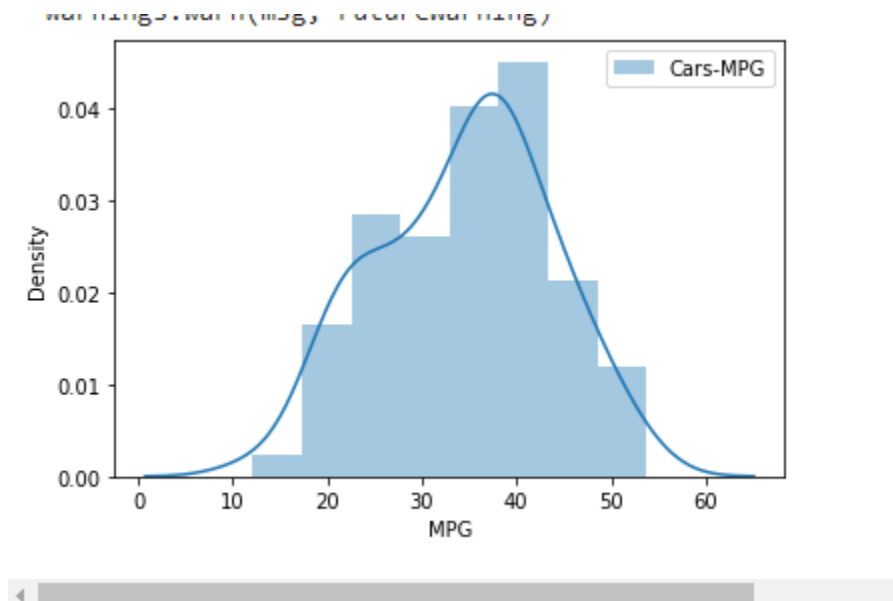
```
[9] # P (20<MPG<50)
stats.norm.cdf(0.50,cars.MPG.mean(),cars.MPG.std())-stats.norm.cdf(0.20,cars.MPG.mean(),cars.MPG.std())

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

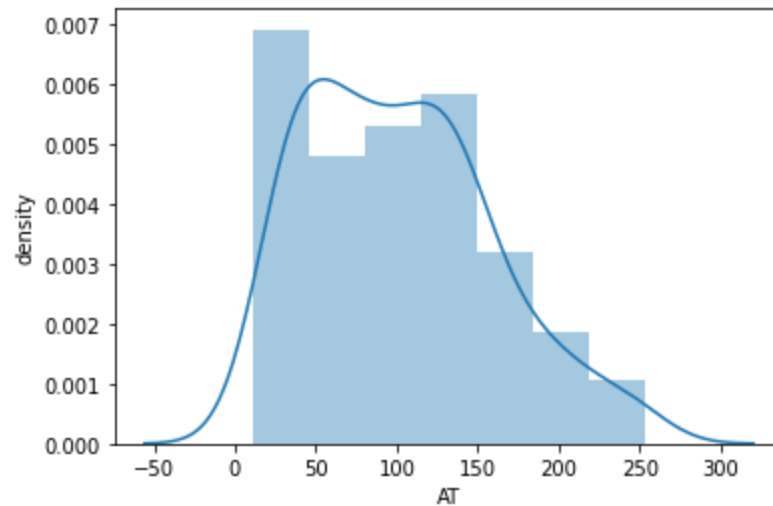


b) Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

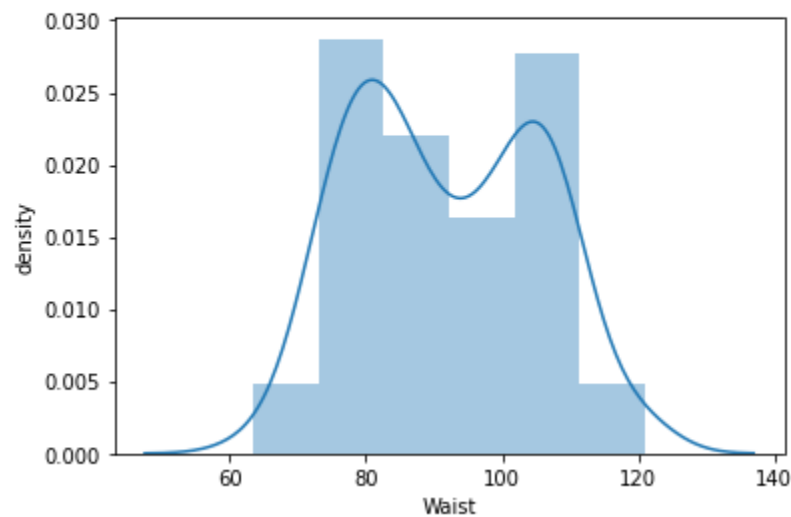
```
[15] # plotting distribution for Adipose Tissue (AT)
sns.distplot(wcat.AT)
plt.ylabel('density');
```

 /usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning



```
[14] # plotting distribution for Waist Circumference (Waist)
sns.distplot(wcat.Waist)
plt.ylabel('density');
```

/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning



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

```
[18] #Q22
      from scipy import stats
      from scipy.stats import norm
```

```
[19] # Z-score of 90% confidence interval
      stats.norm.ppf(0.95)
```

1.6448536269514722

```
[20] # Z-score of 94% confidence interval
      stats.norm.ppf(0.97)
```

1.8807936081512509

```
[21] # Z-score of 60% confidence interval
      stats.norm.ppf(0.8)
```

0.8416212335729143

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

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

2.0638985616280205

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

2.1715446760080677

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

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

x = mean of the sample of bulbs = 260

$\mu$  = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

```
[26] # find t-scores at x=260; t=(s_mean-P_mean)/(s_SD/sqrt(n))
      t=(260-270)/(90/18**0.5)
      t
      -0.4714045207910317
```

For probability calculations, the number of degrees of freedom is  $n - 1$ .

The probability that  $t < -0.471$  with 17 degrees of freedom assuming the population mean is true, the t-value is less than the t-value obtained. With 17 degrees of freedom and a t score of  $-0.471$ , the probability of the bulbs lasting less than 260 days on average of 0.3216 assuming the mean life of the bulbs is 300 days.

```
[29] # 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
```

```
0.32167411684460556
```

```
[30] # 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
```

```
0.32167411684460556
```