DATA SCIENCE - Assignment
Basic Statistics _Level 1

Name: GUVVALA SOMASEKHAR REDDY mail: showmove5697@gmail.com

Q1) Identify the Data type for the Following:

ACTIVITY	DATA TYPE
Number of beatings from Wife	Numerical(Discrete)
Results of rolling a dice	Numerical (Discrete)
Weight of a person	Numerical (Continuous)
Weight of Gold	Numerical (Continuous)
Distance between two places	Numerical (Continuous)
Length of a leaf	Numerical (Continuous)
Dog's weight	Numerical (Continuous)
Blue Color	Categorical(Nominal)
Number of kids	Numerical (Discrete)
Number of tickets in Indian railways	Numerical (Discrete)
Number of times married	Numerical (Discrete)
Gender (Male or Female)	Categorical(Nominal)

Q2) Identify the Data types, which were among the 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	Ordinal
Time on a Clock with Hands	Interval
Number of Children	Nominal
Religious Preference	Nominal
Barometer Pressure	Interval
SAT Scores	Interval
Years of Education	Ratio

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

ANS: The probability that two heads and one tail are obtained is **0.375**

Given 3 coins are tossed.

Probability = Number of Favorable Outcomes / Total Number of Outcomes

Number of Possible outcomes from each coin are 2, head – 'H' and tail – 'T'.

Total observations: 8

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

Number of favorable cases: {HHT, HTH, THH} = 3

- The probability that two heads and one tail are obtained **3/8=0.375**

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

a) Equal to 1 ANS: **0**

b) Less than or equal to 4 and ANS: **1/6**

c) Sum is divisible by 2 and 3 ANS: **1/6**

When Two Dice are rolled
Total observations: $6c_1 * 6c_1 = 36$

(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6),
(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6),
(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6),
(4, 1), (4, 2), (4, 3), (4, 4), (4, 5) (4, 6),
(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6),
(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6).

Probability = Number of Favorable Outcomes / Total Number of Outcomes
a. Is equal to 1
Number of favorable case to get those sum 1 is 0
The probability that sum 1 is = 0/36 = **0**

b. Less than or equal to 4
Number of favorable cases to get those sum Less than or equal to 4 = 6
{ (1,1),(1,2),(2,1),(1,3),(3,1),(2,2)}
The probability that cases to get those sum is Less than or equal to 4 = 6/36=**1/6**

c. Sum is divisible by 2 and 3
Number of favorable cases to get those sum is divisible by 2 and 3= 6
{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1), (6, 6)}
The probability that case to get those sum is divisible by 2 and 3= 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: **10/21**

Probability = Number of Favorable Outcomes / Total Number of Outcomes

Total number of balls
= (red 2 + green 3 + blue 2) = 7
Let S be the sample space
Then, n(S) = Number of ways of drawing 2 balls out of 7
$n(S)=7C2=(7\times6)(2\times1)=21$
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

$n(E)=5C2=(5\times4)(2\times1)=10$

∴ The probability that none of the balls drawn is blue $P(E)=n(E)/n(S)=10/21$

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
ANS: **3.09**

Expected number of candies for a randomly selected child
$= \sum P(x).E(x)$

=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

ANS:

	Mean	Median	Mode	Variance	STD	Range
Points	3.596563	3.695000	3.92	0.285881	0.534679	1.63
Score	3.217250	3.325000	3.44	0.957379	0.978457	3.911
Weigh	17.848750	17.710000	17.02	3.193166	1.786943	8.4

- These mean, median and mode are approximately same in each individual.
- The data points are likely to frame normal distribution.

Code:

```
Q7[['Points','Score','Weigh']].agg(['mean','median','var','std','min','max'])
Q7[['Points','Score','Weigh']].agg(['mode'])
```

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?

ANS: **145. 33**

Expected Value = ∑ (probability* Value)

∑ P(x).E(x)

There are 9 patients

Probability of selecting each patient = 1/9

E(x)	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)

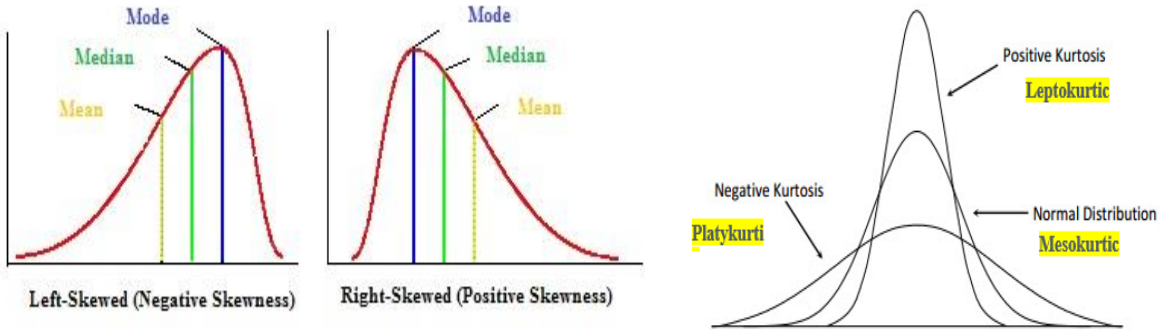
= 145.33

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

ANS:



	Skewness	Kurtosis
Speed	-0.1175	-0.509
distance	0.8069	0.4051

- Speed is negatively skewed and the distribution has its tail on the left side of the distribution
- Distance is positively skewed and it has tail on the on the right side of the distribution
- Kurtosis of speed is negative then kurtosis less than normal distribution and it has lower tail
- Kurtosis of distance is positive then kurtosis more than normal distribution and it has upper skinny tails

Code:

```
print('speed skewness: ',round(Q9_a['speed'].skew(),4))
print('distance skewness: ',round(Q9_a['dist'].skew(),4))
print('speed kurtosis: ',round(Q9_a['speed'].kurt(),4))
print('distance kurtosis: ',round(Q9_a['dist'].kurt(),4))
```

SP and Weight (WT) Use Q9_b.csv

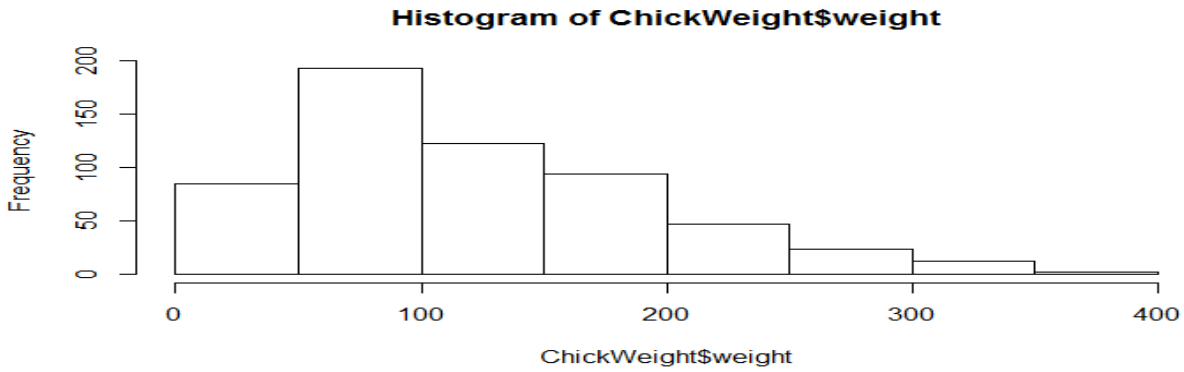
	Skewness	Kurtosis
SP	1.6115	2.9773
WT	-0.6148	0.9503

- WT is negatively skewed and the distribution has its tail on the left side of the distribution
- SP is positively skewed and it has tail on the on the right side of the distribution
- Both SP and WT have positive kurtosis implies, so the kurtosis more than normal distribution and they have long and skinny tails

Code:

```
print('sp skewness: ',round(Q9_b.SP.skew(),4))
print('wt skewness:',round(Q9_b.WT.skew(),4))
print('sp kurtosis: ',round(Q9_b.SP.kurt(),4))
print('wt kurtosis: ',round(Q9_b.WT.kurt(),4))
```

Q10) Draw inferences about the following boxplot & histogram

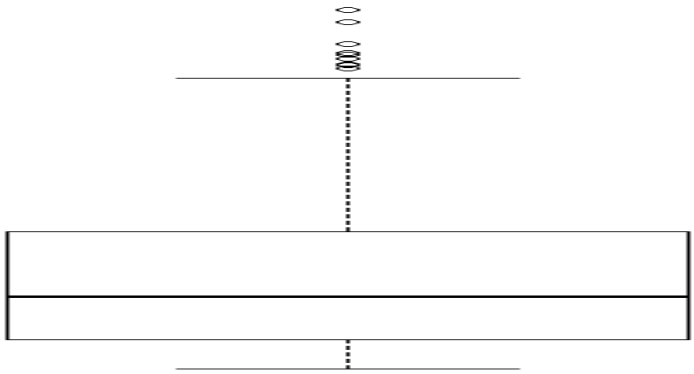


ANS:

Histogram:

- This is an asymmetric distribution.
- It is a right tailed and positively skewed.
- Outliers can find in the right end.

Boxplot:



- This is an asymmetric distribution.
- Outliers are on the right extreme 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?

ANS:

The confidence interval for 94% is: [143.5762, 256.4238]
The confidence interval for 98% is: [130.2096, 269.7904]
The confidence interval for 96% is: [138.3875, 261.6125]

Code:

```
print(" The confidence interval for 94% is :  ",np.round(stats.norm.interval(0.94,200,30),4))
print(" The confidence interval for 98% is :  ",np.round(stats.norm.interval(0.98,200,30),4))
print(" The confidence interval for 96% is :  ",np.round(stats.norm.interval(0.96,200,30),4))
```

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.

ANS:

Mean : 41
Median : 40.5
Variance : 24.11
Standard Deviation : 4.91

Code:

```
a = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]
print (" mean               : ",st.mean(a))
print (" median             : ",st.median(a))
print (" varience           : ",round(np.var(a),2))
print (" standard deviation : ",round(np.std(a),2))
```

2) What can we say about the student marks?

ANS:

- The average marks of students are 41.
- Min and max marks are 34 and 56 respectively.
- Maximum students have scored marks between 38 and 42.

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

ANS: It is following normal distribution without skewness.

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

ANS: Then the distribution will be negatively skewed.

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

ANS: Then the distribution will be positively skewed.

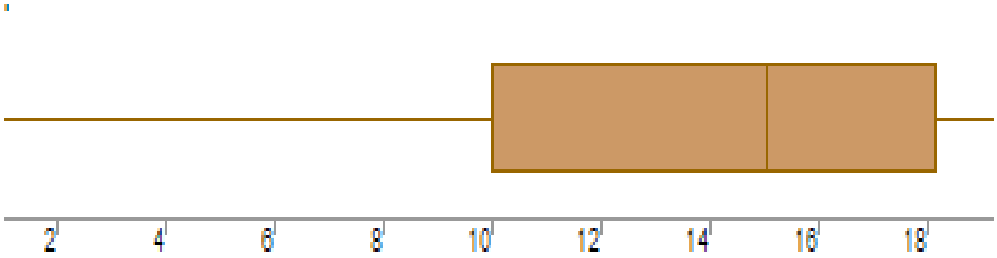
Q16) what does positive kurtosis value indicates for a data?

ANS: Plot of data will have sharp thin peak and data will be denser in short range.

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

ANS: Plot of data will have broader peak and data will be spread over wide range.

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



What can we say about the distribution of the data?

ANS:

- This is negatively skewed and with outliers on the left side of median.
- The median is present around 15.

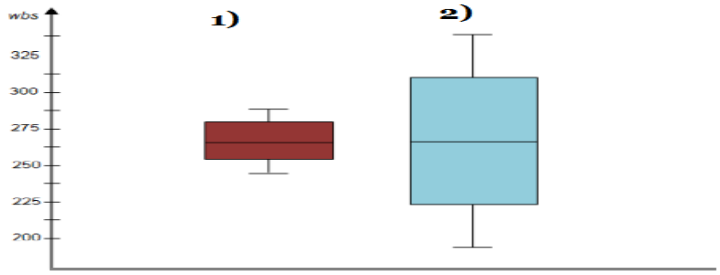
What is nature of skewness of the data?

ANS: Negatively Skewed

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

ANS: 8

Q19) Comment on the below Boxplot visualizations?



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

ANS:

- These two plots are normal distributed with having their median around 260
- Both of the plots do not have outliers and these are not skewed.
- Plot 2 is covering more area (or range) than Plot 1.

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) ANS: 0.3476

b)P(MPG<40) ANS: 0.7293

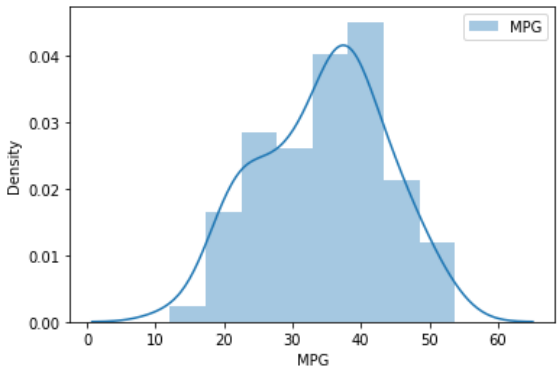
c)P (20<MPG<50) ANS: 0.8989

Code:

```
m=cars['MPG'].mean()
s=cars['MPG'].std()
print('a. P(MPG>38)      :', 1- round(stats.norm.cdf(38,m,s),4))
print('b. P(MPG<40)      :', round(stats.norm.cdf(40,m,s),4))
print('c. P(20<MPG<50)   :', round(stats.norm.cdf(50,m,s)-stats.norm.cdf(20,m,s),4))
```

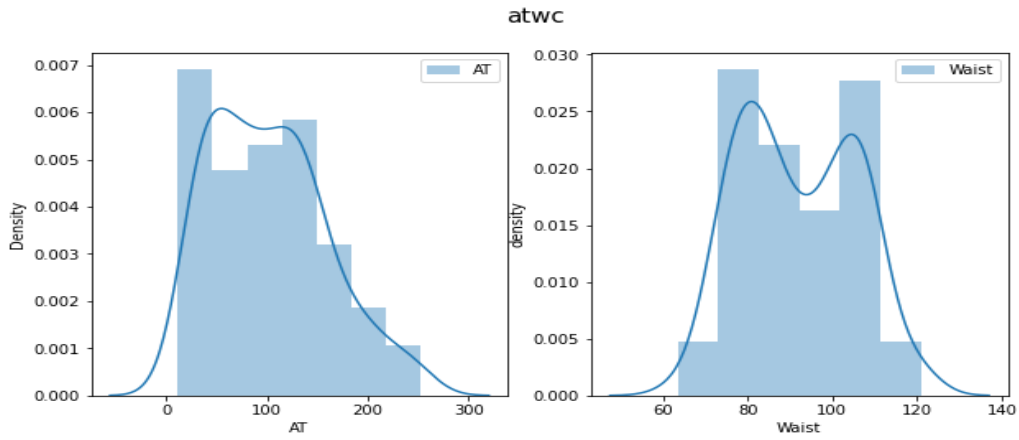
Q 21) Check whether the data follows normal distribution
A) Check whether the MPG of Cars follows Normal Distribution
Dataset: Cars.csv

```
import seaborn as sn

sn.distplot(cars.MPG,label='MPG');
plt.xlabel('MPG');
plt.ylabel('Density');
plt.legend();

#This is not follwing normal distribution
```



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

```
# both are not follwing normal disribution
```



Code:

```
plt.figure(figsize=(10,5));
plt.subplot(1,2,1);
plt.suptitle('atwc',fontsize=15);

sn.distplot(atwc.AT, label = 'AT');

plt.xlabel('AT');
plt.ylabel('Density');
plt.legend();

plt.subplot(1,2,2);
sn.distplot(atwc.Waist, label = 'Waist');

plt.xlabel('Waist');
plt.ylabel('density');
plt.legend();
```

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

ANS:

The Z score of 90% confidence interval: 1.2816
The Z score of 94% confidence interval: 1.5548
The Z score of 60% confidence interval: 0.2533

Code:

```
print('The Z score of 90% confidence interval : ',round(stats.norm.ppf(0.90),4))
print('The Z score of 94% confidence interval : ',round(stats.norm.ppf(0.94),4))
print('The Z score of 60% confidence interval : ',round(stats.norm.ppf(0.60),4))
```

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

ANS:

The t score of 95% confidence interval for sample size of 25: 1.7109
The t score of 96% confidence interval for sample size of 25: 1.8281
The t score of 99% confidence interval for sample size of 25: 2.4922

Code:

```
print('The t score of 95% confidence interval for sample size of 25: ',round(stats.t.ppf(0.95,25-1),4))
print('The t score of 96% confidence interval for sample size of 25: ',round(stats.t.ppf(0.96,25-1),4))
print('The t score of 99% confidence interval for sample size of 25: ',round(stats.t.ppf(0.99,25-1),4))
```

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- Score: -0.4714
The probability that 18 randomly selected bulbs would have

an average life of no more than 260days : 0.3217

Code:

```
print('T score : ',round((260-270)/(90/np.sqrt(18)),4))

print('The probability that 18 randomly selected bulbs would have an average life of no more than 260 days :',
      round(stats.t.cdf(-0.4714,18-1),4))
```