Q1) Identify the Data type for the Following:

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

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	Ratio
Fahrenheit Temperature	Interval
Height	Ratio
Type of living accommodation	Ordinal
Level of Agreement	Ordinal
IQ(Intelligence Scale)	Ratio
Sales Figures	Ratio
Blood Group	Ordinal
Time Of Day	Interval
Time on a Clock with Hands	Ordinal
Number of Children	Nominal
Religious Preference	Ratio

Barometer Pressure	Interval
SAT Scores	Ordinal
Years of Education	Ratio

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

Sol :

(HHH, HHT, HTH, THH, TTH, THT, HTT, TTT – 8 outcomes) – & (Two head and one tail are HHT, HTH, TTH so 3 Probability) so there are **3/8 - 0.375**

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

a) Equal to 1 **- 0**
b) Less than or equal to 4 - : **6/36 = 1/6**
c) 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?

Sol:

Total Number of balls = **2+3+2 = 7**

Number of ways of drawing 2 balls out of 7= $^{7}\mathbf{C}_{2}$ **= (7 x 6) / (2 x 1) = 42/2 = 21**

Number of balls other than blue = **5**

Number of ways of drawing 2 balls out of 5 = $^{5}\mathbf{C}_{2}$ **= (5 x 4) / (2 x 1)**

=20/2 = 10

∴ Required Probability = **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

Sol :

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

= 1*0.015 + 4*0.20 + 3*0.65 + 5*0.005 + 6*0.01 + 2*0.120 **= 3.090**

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


Q7.csv


Q7 .ipynb

Sol:

In [8]:
```
import os
import pandas as pd
import numpy as np
```

In [9]:
```
data = pd.read_csv("Q7.csv")
```

In [10]:
```
summary =  data.describe()
var = data.var()
Mode = data.mode()
```

```
C:\Users\RAJMAH~2\AppData\Local\Temp/ipykernel_22564/3942730837.py:2: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError.  Select only valid columns before calling the reduction.
  var = data.var()
```

In [11]:
```
print(summary)
```

```
            Points      Score      Weigh
count  32.000000  32.000000  32.000000
mean    3.596563   3.217250  17.848750
std     0.534679   0.978457   1.786943
min     2.760000   1.513000  14.500000
25%     3.080000   2.581250  16.892500
50%     3.695000   3.325000  17.710000
75%     3.920000   3.610000  18.900000
max     4.930000   5.424000  22.900000
```

In [12]:
```
print(var)
```

```
Points    0.285881
Score     0.957379
Weigh     3.193166
dtype: float64
```

In [13]:
```
print(Mode)
```

```
             Unnamed: 0  Points  Score  Weigh
0             AMC Javelin    3.07   3.44  17.02
1    Cadillac Fleetwood    3.92    NaN  18.90
2           Camaro Z28     NaN    NaN    NaN
3     Chrysler Imperial     NaN    NaN    NaN
4           Datsun 710     NaN    NaN    NaN
5      Dodge Challenger     NaN    NaN    NaN
6           Duster 360     NaN    NaN    NaN
7          Ferrari Dino     NaN    NaN    NaN
8              Fiat 128     NaN    NaN    NaN
9            Fiat X1-9     NaN    NaN    NaN
10       Ford Pantera L     NaN    NaN    NaN
11          Honda Civic     NaN    NaN    NaN
12       Hornet 4 Drive     NaN    NaN    NaN
13    Hornet Sportabout     NaN    NaN    NaN
14  Lincoln Continental     NaN    NaN    NaN
15         Lotus Europa     NaN    NaN    NaN
16        Maserati Bora     NaN    NaN    NaN
17            Mazda RX4     NaN    NaN    NaN
18        Mazda RX4 Wag     NaN    NaN    NaN
19             Merc 230     NaN    NaN    NaN
20            Merc 240D     NaN    NaN    NaN
21             Merc 280     NaN    NaN    NaN
22            Merc 280C     NaN    NaN    NaN
23           Merc 450SE     NaN    NaN    NaN
24           Merc 450SL     NaN    NaN    NaN
25          Merc 450SLC     NaN    NaN    NaN
26     Pontiac Firebird     NaN    NaN    NaN
27        Porsche 914-2     NaN    NaN    NaN
28       Toyota Corolla     NaN    NaN    NaN
29        Toyota Corona     NaN    NaN    NaN
30              Valiant     NaN    NaN    NaN
31           Volvo 142E     NaN    NaN    NaN
```

In [14]:
```
print("Inferences : mean mode and median are not equal so we can say that data is skewed")
```

```
Inferences : mean mode and median are not equal so we can say that data is skewed
```

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?

Sol :

Q8.ipynb

```
In [4]: import numpy as np
        import pandas as pd

In [5]: weight = [108, 110, 123, 134, 135, 145, 167, 187, 199]
        df = pd.DataFrame(weight)

In [6]: weight_of_paitenet =  df.mean()
        print("The Expected Value of the Weight of that patient is ",weight_of_paitenet)

        The Expected Value of the Weight of that patient is  0    145.333333
        dtype: float64
```

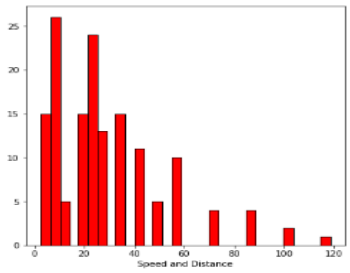
Q9) Calculate Skewness, Kurtosis & draw inferences on the following dat

Cars speed and distance Use Q9_a.csv , SP and Weight(WT) Use Q9_b.csv

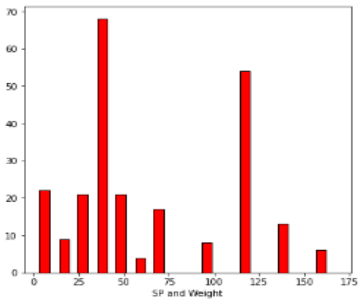
Q9.ipynb Q9_b.csv Q9_a.csv

Sol :

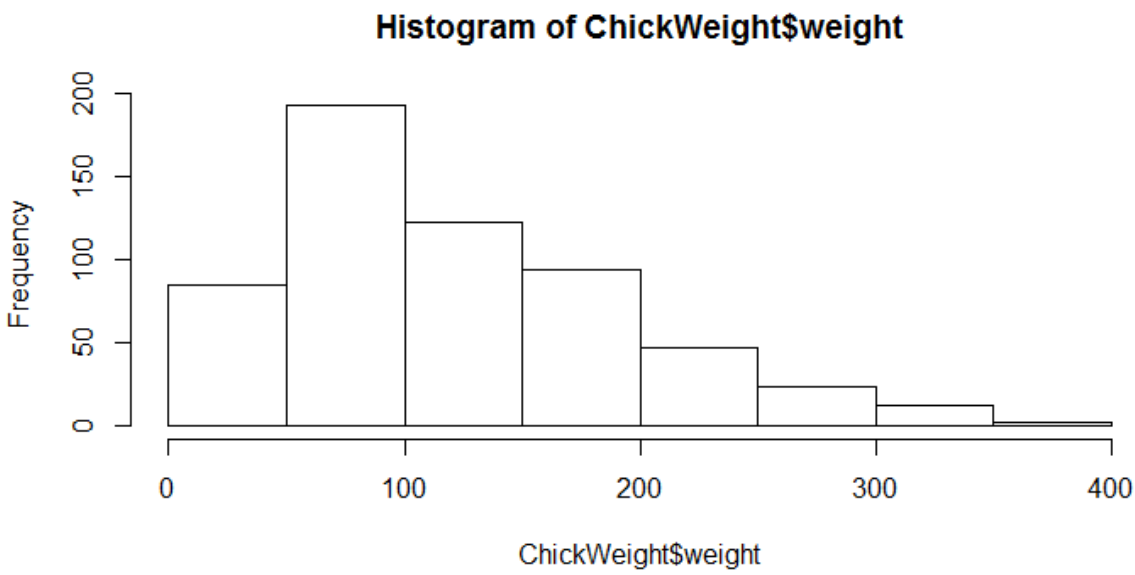
```
In [1]: import os
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt

In [2]: data = pd.read_csv("Q9_a.csv")

In [3]: data.skew()
Out[3]: Index    0.000000
        speed   -0.117510
        dist     0.806895
        dtype: float64

In [4]: data.kurtosis()
Out[4]: Index   -1.200000
        speed   -0.508994
        dist     0.405053
        dtype: float64

In [5]: plt.figure(figsize=(6,6),facecolor ="white")
        plt.hist(data,facecolor ="red",edgecolor ="black",bins =8)
        #creates histogram with 5bins and colours filled init.
        plt.xlabel("Speed and Distance")
Out[5]: Text(0.5, 0, 'Speed and Distance')
```



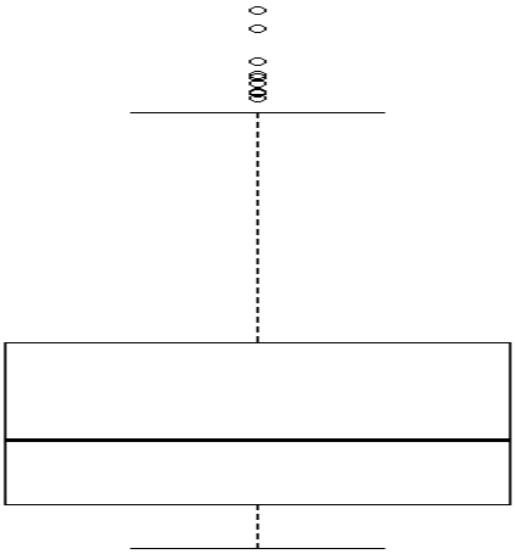

```
In [6]: data = pd.read_csv("Q9_b.csv")

In [7]: data.skew()
Out[7]: Unnamed: 0    0.000000
        SP            1.611450
        WT           -0.614753
        dtype: float64

In [8]: data.kurtosis()
Out[8]: Unnamed: 0   -1.200000
        SP            2.977329
        WT            0.950291
        dtype: float64

In [10]: plt.figure(figsize=(6,6),facecolor ="white")
         plt.hist(data,facecolor ="red",edgecolor ="black",bins =8)
         #creates histogram with 5bins and colours filled init.
         plt.xlabel("SP and Weight")
Out[10]: Text(0.5, 0, 'SP and Weight')
```



Q10) Draw inferences about the following boxplot & histogram



inference : The distribution is right skew('+ve'), Mean >Median



Inference : The distribution has lots of outliers towards upper extreme

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?

Sol :

Q11.ipynb

```
In [1]: import os
        import numpy as np
        from scipy import stats

In [2]: stats.norm.interval(0.94, loc=200 , scale=30/np.sqrt(2000))  # Lower to uper Limit
Out[2]: (198.738325292158, 201.261674707842)

In [3]: stats.norm.interval(0.98, loc=200 , scale=30/np.sqrt(2000)) # Lower to uper Limit
Out[3]: (198.43943840429978, 201.56056159570022)

In [4]: stats.norm.interval(0.96, loc=200 , scale=30/np.sqrt(2000)) # Lower to uper Limit
Out[4]: (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?

Sol :

Q12.ipynb

In [1]:
```
import pandas as pd
```

In [2]:
```
data = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]
df = pd.DataFrame(data)
```

In [3]:
```
df.describe()
```

Out[3]:

	0
count	18.000000
mean	41.000000
std	5.052664
min	34.000000
25%	38.250000
50%	40.500000
75%	41.750000
max	56.000000

In [4]:
```
df.mode()
```

Out[4]:

	0
0	41

In [5]:
```
df.var()
```

Out[5]:
```
0    25.529412
dtype: float64
```

In [6]:
```
print("we can say that avg mark student marks is : 41")
```

```
we can say that avg mark student marks is : 41
```

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

Sol : skewness=0, Symmetric

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

Sol : Right skewed(tail on the right side)

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

Sol : Left skewed(tail on the left side)

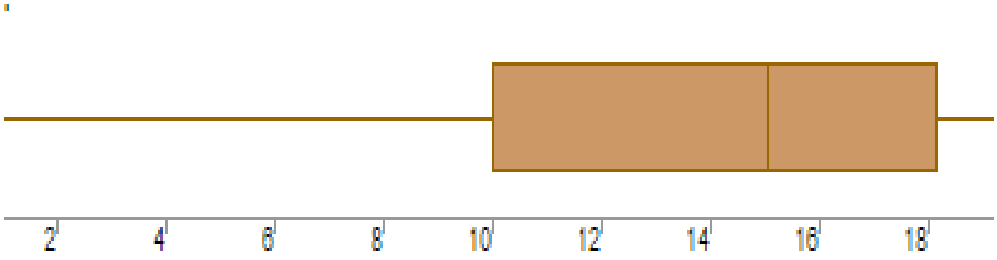
Q16) What does positive kurtosis value indicates for a data ?

Sol : Sharp Peak, Thick Tails

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

Sol : Broad Peak, Wide Tails

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



What can we say about the distribution of the data?

Sol : It is Not a Normal Distribution

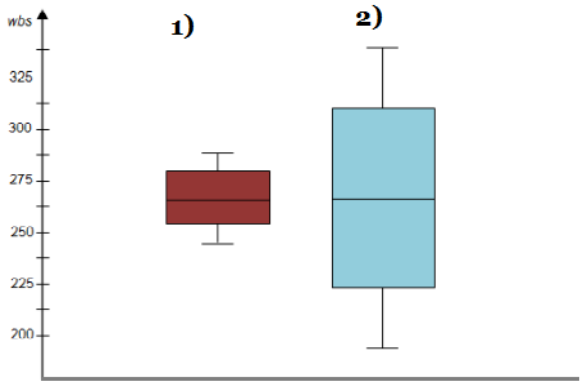
What is nature of skewness of the data?

Sol : Left Skewed

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

Sol : Inter Quartile Range=Upper Quartile-Lower Quartile=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.

Sol :

1) The median of the two boxplots are same
2) Both are Normally Distributed

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)

Sol :

Q20.ipynb

In [5]:
```
import pandas as pd
from scipy import stats
```

In [6]:
```
data = pd.read_csv("Cars.csv")
data
```

Out[6]:

	HP	MPG	VOL	SP	WT
0	49	53.700681	89	104.185353	28.762059
1	55	50.013401	92	105.461264	30.466833
2	55	50.013401	92	105.461264	30.193597
3	70	45.696322	92	113.461264	30.632114
4	53	50.504232	92	104.461264	29.889149
...	...	...	...	...	...
76	322	36.900000	50	169.598513	16.132947
77	238	19.197888	115	150.576579	37.923113
78	263	34.000000	50	151.598513	15.769625
79	295	19.833733	119	167.944460	39.423099
80	236	12.101263	107	139.840817	34.948615

81 rows × 5 columns

In [7]:
```
Avg = data['MPG'].mean()
std = data['MPG'].std()
```

In [10]:
```
Out_38 = (1 - stats.norm.cdf(38,loc=Avg,scale=std)) #P(MPG>38)
Out_38
```

Out[10]:
```
0.3475939251582705
```

In [11]:
```
Out_40 = stats.norm.cdf(40,loc=Avg,scale=std) #P(MPG<40)
Out_40
```

Out[11]:
```
0.7293498762151616
```

In [12]:
```
Out_40_20 = stats.norm.cdf(50,loc=Avg,scale=std)- stats.norm.cdf(20,loc=Avg,scale=std) #P(20<MPG<50)
Out_40_20
```

Out[12]:
```
0.8988689169682046
```

In [ ]:

Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Sol :

Cars.csv Q21.ipynb

In [11]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [12]:
```
data = pd.read_csv("Cars.csv")
data
```

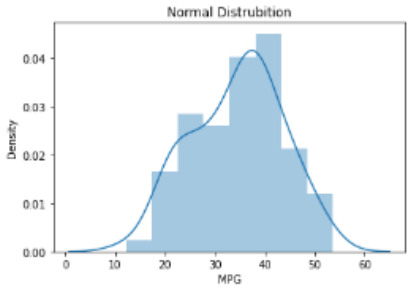
Out[12]:

	HP	MPG	VOL	SP	WT
0	49	53.700681	89	104.185353	28.762059
1	55	50.013401	92	105.461264	30.466833
2	55	50.013401	92	105.461264	30.193597
3	70	45.696322	92	113.461264	30.632114
4	53	50.504232	92	104.461264	29.889149
...	...	...	...	...	...
76	322	36.900000	50	169.598513	16.132947
77	238	19.197888	115	150.576579	37.923113
78	263	34.000000	50	151.598513	15.769625
79	295	19.833733	119	167.944460	39.423099
80	236	12.101263	107	139.840817	34.948615

81 rows × 5 columns

In [19]:
```
sns.distplot(data['MPG'])
plt.title("Normal Distrubition");
plt.show()
```

```
C:\Users\rajmah60018\Anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```



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

Sol :

In [22]:
```
data = pd.read_csv("wc-at.csv")
data
```

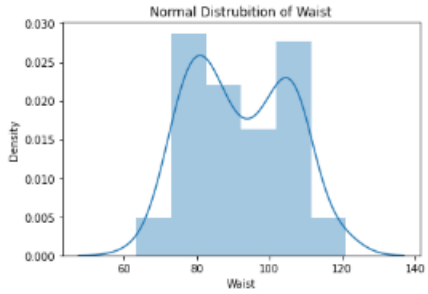
Out[22]:

	Waist	AT
0	74.75	25.72
1	72.60	25.89
2	81.80	42.60
3	83.95	42.80
4	74.65	29.84
...	...	...
104	100.10	124.00
105	93.30	62.20
106	101.80	133.00
107	107.90	208.00
108	108.50	208.00

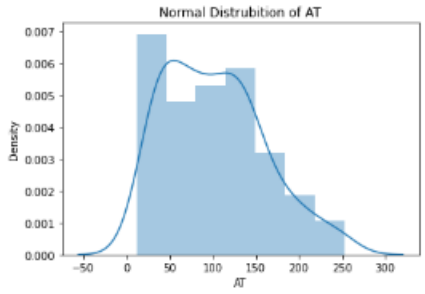
109 rows × 2 columns

In [24]:
```
sns.distplot(data['Waist'])
plt.title("Normal Distrubition of Waist");
plt.show()
sns.distplot(data['AT'])
plt.title("Normal Distrubition of AT");
plt.show()
```

```
C:\Users\rajmah60018\Anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated funct
ion and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with simil
ar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```




```
C:\Users\rajmah60018\Anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated funct
ion and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with simil
ar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```



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

Sol :

Q22.ipynb

```
In [6]: from scipy import stats

In [13]: Z_90 = stats.norm.ppf(0.90)
         Z_90

Out[13]: 1.2815515655446004

In [14]: Z_94 = stats.norm.ppf(0.94)
         Z_94

Out[14]: 1.5547735945968535

In [15]: Z_80 = stats.norm.ppf(0.80)
         Z_80

Out[15]: 0.8416212335729143
```

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

Sol :

Q23.ipynb

```
In [1]: from scipy import stats

In [3]: Z_95 = stats.t.ppf(0.90,df=25)
        Z_95

Out[3]: 1.3163450765592588

In [4]: Z_96 = stats.t.ppf(0.96,df=25)
        Z_96

Out[4]: 1.8248284689556018

In [5]: Z_99 = stats.t.ppf(0.99,df=25)
        Z_99

Out[5]: 2.4851071754106413
```

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

Sol :

Q24.ipynb

```
In [1]: from scipy import stats

In [2]: S_mean = 260
        P_mean = 270
        std = 90
        Tscore = (260 -270)/(90/(18**0.5))

In [3]: Tscore

Out[3]: -0.4714045207910317

In [4]: stats.t.cdf((Tscore),17)

Out[4]: 0.32167253567098364
```