Q1) Identify the Data type for the Following:

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

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	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	Nominal
Time on a Clock with Hands	Ratio
Number of Children	Ratio
Religious Preference	Ordinal

Barometer Pressure	Interval
SAT Scores	Ratio
Years of Education	Ratio

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

Total scenarios:

8

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

Possible scenarios : 3

$\therefore$ probability = $3/8 = 0.375$

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

Total possible outcomes when 2 dices are rolled

$= 6 \times 6 = 36$

a) Equal to 1 = 0 (Not possible, as min. sum when 2 dices are rolled are 1+1 = 2)

b) ≤ 4

possible outcomes = (1,1) (1,2) (1,3) (2,1) (2,2) (3,1)

$\therefore P(\leq 4) = \frac{6}{36} = \frac{1}{6}$

c) 2 and 3

Sum is divisible by _____

Total scenarios possible where sum is divisible by

2 & 3 = 24

∴ P(24/36) = 2/3

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?

Total no. of balls = 2R + 2B + 3G = 7

Total no. of ways to draw 2 balls out of 7 balls

$= {}^{7}C_2 = \frac{7!}{2! \times (7-2)!} = \frac{7 \times 6}{2 \times 1} = 21$ ways

Total no. of ways to draw 2 balls out of 2R and 3G.

$= {}^{5}C_2 = \frac{5 \times 4}{2} = 10$

∴ P(No blue balls drawn) = $\frac{10}{21}$

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

Expected prob. of children getting candies at random = $1 \times 0.015 + 4 \times 0.20 + 3 \times 0.65 + 5 \times 0.005 + 6 \times 0.01 + 2 \times 0.120$

$= 3.09$

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

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

Out[3]:

	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 [8]: df.head(5)
```

Out[8]:

	Unnamed: 0	Points	Score	Weigh
0	Mazda RX4	3.90	2.620	16.46
1	Mazda RX4 Wag	3.90	2.875	17.02
2	Datsun 710	3.85	2.320	18.61
3	Hornet 4 Drive	3.08	3.215	19.44
4	Hornet Sportabout	3.15	3.440	17.02

```
In [10]: df.median()
```

```
Out[10]: Points     3.695
         Score      3.325
         Weigh     17.710
         dtype: float64
```

```
In [15]: df[['Points','Score','Weigh']].mode()
```

Out[15]:

	Points	Score	Weigh
0	3.07	3.44	17.02
1	3.92	NaN	18.90

for code check jupyter file

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?

Expected value = $\frac{(108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)}{9}$

here
no. of patients = 9

$= 145.33$

$\therefore$ Expected value = 145.33

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

Cars speed and distance

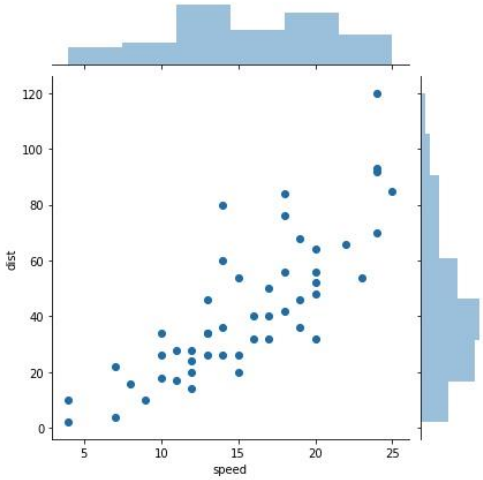
Use Q9_a.csv

SP and Weight(WT)

In [33]:
```
from scipy.stats import skew
from scipy.stats import kurtosis
print("skewness : ",skew(df_q9))
print("kurtosis : ",kurtosis(df_q9))
```

```
skewness :  [ 0.         -0.11395477  0.78248352]
kurtosis :  [-1.20096038 -0.57714742  0.24801866]
```

In [34]:
```
sns.jointplot(x= 'speed', y= 'dist', data=df_q9)
plt.show()
```



Inferences on Car Speed vs Distance

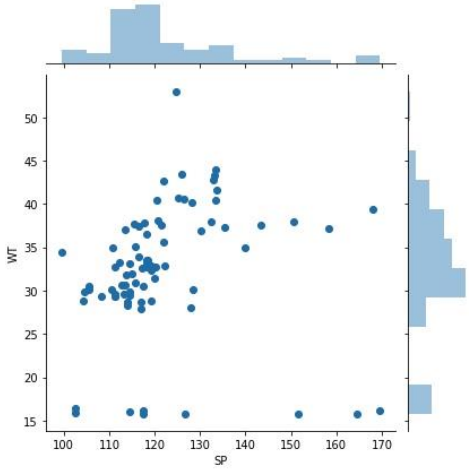
From the above plot we can say that Speed and Distance have a positive correlation , and once Speed increases distance increases too. There are a few outliers like when speed is 4 , dist is 10. So ignoring these outliers we can say that +ve correlation holds true for most of the data.

Details about steps are mentioned in attached notebook.

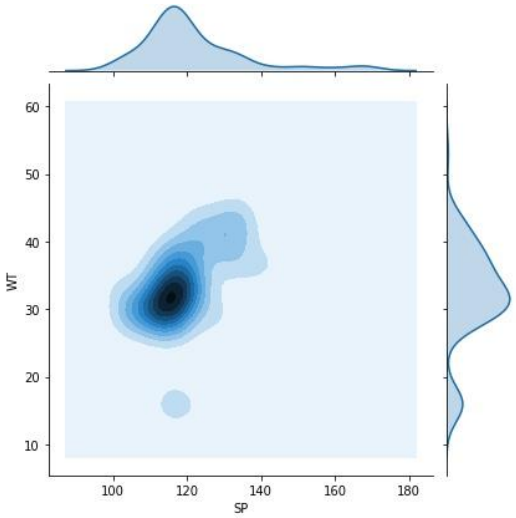
Use Q9_b.csv

```
skewness :  [ 0.          1.58145368 -0.60330993]
kurtosis :  [-1.20036585  2.72352149  0.81946588]
```

```
]: sns.jointplot(x= 'SP', y= 'WT', data=df_q9b)
   plt.show()
```




```
]: sns.jointplot(x='SP', y= 'WT', data=df_q9b,kind='kde')
   plt.show()
```




```
In [10]: df_q9b.describe()
```

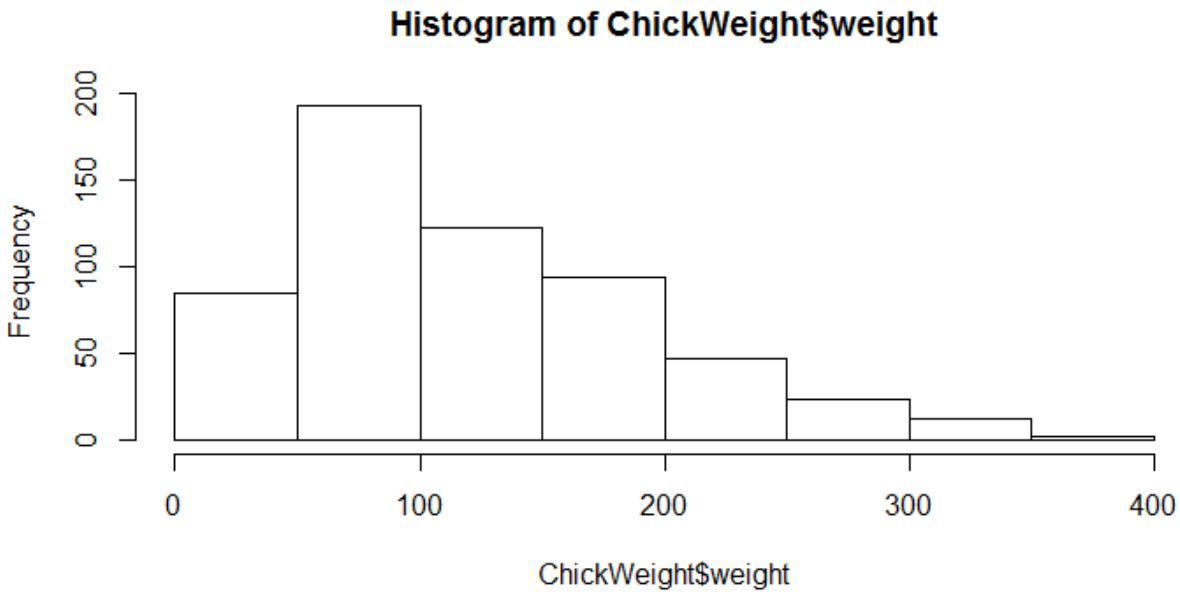
Out[10]:

	Unnamed: 0	SP	WT
count	81.000000	81.000000	81.000000
mean	41.000000	121.540272	32.412577
std	23.526581	14.181432	7.492813
min	1.000000	99.564907	15.712859
25%	21.000000	113.829145	29.591768
50%	41.000000	118.208698	32.734518
75%	61.000000	126.404312	37.392524
max	81.000000	169.598513	52.997752

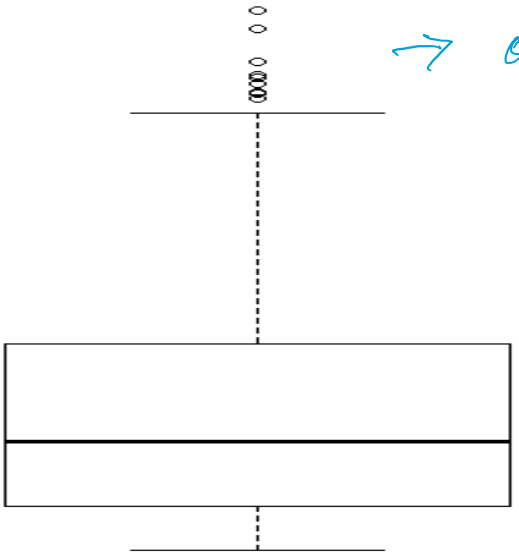
Inferences regarding speed and weight

From the density plot and the summary , we can make out that most of the cars weigh close to 30 and their speed averages at 120.

Q10) Draw inferences about the following boxplot & histogram



Boxplot obs:



→ outliers

1) First 25% of data have higher values compared to rest of the data

2) The median lies on the lower side as majority of data is at lower 75%

Histogram obs:

1) Maximum of the data falls in the range from 50-100, so the mode of the data lies in this range

⇒ The range of 'weight' is from 0–400.

⇒ The distribution is positively skewed as the distribution is biased towards right.

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?

$n = 2000$, $\bar{x} = 200$ $S = 30$

Confidence interval formulae = $\bar{x} \pm Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$

94%

$\alpha = 1 - \frac{94}{100} = \frac{6}{100} = 0.06$

$\alpha/2 = 0.03$

critical probability = $94 + 0.03 = 97\%$

∴ t score for 97% and degrees of freedom = $n - 1$
$= 1999$

from scipy import stats

stats.t.ppf(0.97,df = 1999) = gives value of 1.88

Substituting in formulae of C.I =

$$200 \pm 1.88 \cdot \frac{30}{\sqrt{2000}}$$

$$= 200 \pm 1.88 \cdot 0.67$$

$$= 200 \pm 1.26$$

∴ CI for 94% = (198.74, 201.26)

Similarly for 98% CI

$\alpha/2 \quad 1 - \frac{98}{100} = 0.02$

$\alpha/2 = 0.01$

$\therefore$ Critical probability $= 98 + 0.01$
$= 99\%$

$\therefore$ t table prob. $= 2.32$

$\therefore \quad 200 \pm 2.32 \cdot 0.67$

$= 200 \pm 1.55$

C.I for 98% = (196.45, 201.55)

C.I for 96%

$\alpha/2 \quad 1 - \frac{96}{100} = 0.04$

$\alpha/2 = 0.02$

critical prob. = 98%

$\therefore$ t table prob. for 98% = 2.055

$\therefore$ CI $= 200 \pm 2.055 \cdot 0.67$

$= 200 \pm 1.376$

$= (198.63, 201.376)$

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 [6]: students = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

In [12]: np.median(students)
Out[12]: 40.5

In [13]: np.mean(students)
Out[13]: 41.0

In [14]: np.std(students)
Out[14]: 4.910306620885412

In [15]: from scipy.stats import skew
         from scipy.stats import kurtosis
         print("skewness : ",skew(students))
         print("kurtosis : ",kurtosis(students))

         skewness :  1.5428846814037365
         kurtosis :  2.6216313788782957
```

median

mean

standard deviation

Skewness

2) What can we say about the student marks?

→ Range → 34 – 56

outliers → 49, 56 (as seen from distplot)

average score = 41

Skewness = 1.5 so the distribution is rightly skewed.

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

Skewness = 0, the distribution is symetrical

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

positively skewed

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

negatively skewed data

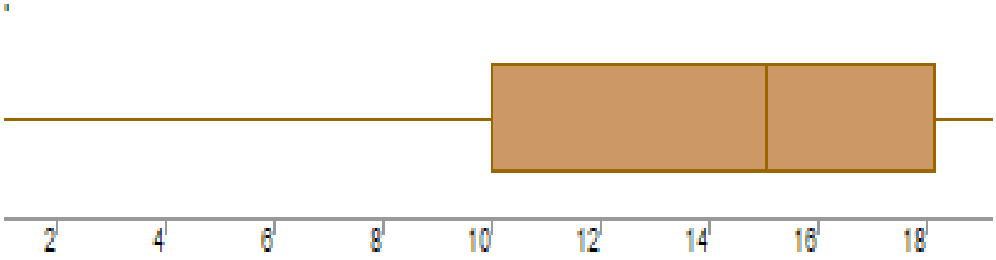
Q16) What does positive kurtosis value indicates for a data ?

positive kurtosis indicates the majority of the data lies close to the tails/end and the spread is large

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

most of the data lies close to the center and spread is less

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



What can we say about the distribution of the data?

spread = 1 – 20 IQR = 10 – 18

median = 15.5

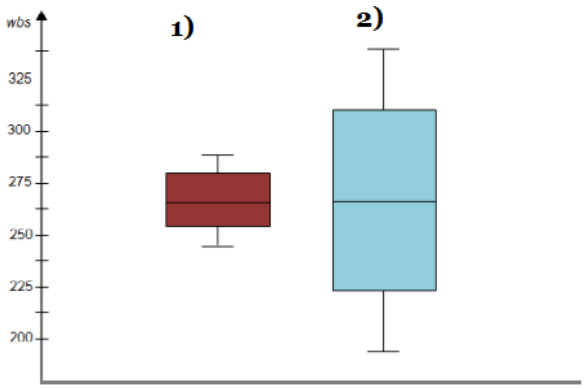
What is nature of skewness of the data?

median is greater than mean so the distribution is negatively skewed.

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

IQR = 10 – 18

Q19) Comment on the below Boxplot visualizations?



1) Spread = 237.5 – 287.5

IQR = 255 – 280

median = 262.5

2) median = 262.5

spread = 190 – 327.5

IQR = 225 – 312.5

Boxplot 2 data is more widely spread than the data from boxplot 1. The median is same for both boxplots.

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

a. P(MPG>38) → 33/81
b. P(MPG<40) → 61/81
c. P (20<MPG<50) → 69/81

Details are present in the attached jupyter notebook.

Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution
Dataset: Cars.csv

check jupyter notebook

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

check jupyter notebook.

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

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

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

check jupyter notebook

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

check jupyter notebook