Locating Extreme Outliers: Z-Score

To compute the Z-**score (Standard score)** of a data value, **subtract the mean** and **divide by the standard deviation.**

The Z-score is the number of standard deviations a data value is from the mean.

A data value is considered an extreme outlier if its Z-score is **less than -3.0 or greater than +3.0.**

The **larger** the absolute value of the Z-score, the **farther** the data value is from the mean.

$$Z = \frac{X - \overline{X}}{S}$$

where X represents the data value

$\overline{X}$ is the sample mean

S is the sample standard deviation

Suppose the **mean** math SAT score is 490, with a standard deviation of 100.

Compute the Z-score for a test score of 620.

$$Z = \frac{X - \overline{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1.3$$

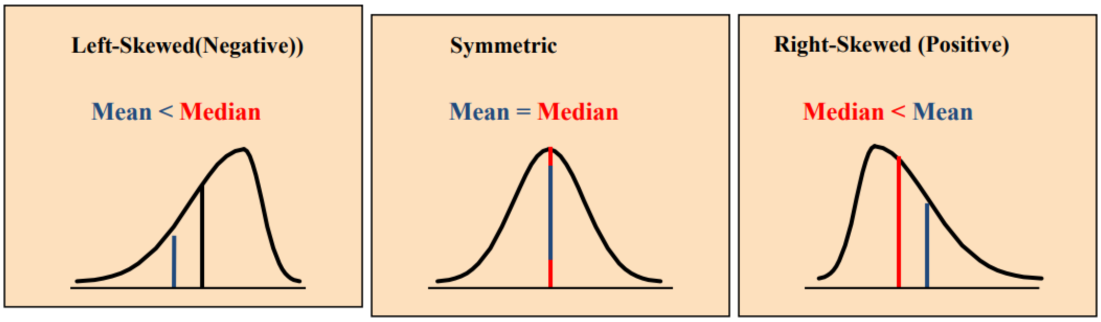
A score of 620 is 1.3 standard deviations above the mean and would **not be considered an outlier**.

Shape of a Distribution

Describes how data are distributed

Measures of shape

- Symmetric or skewed



Numerical Descriptive Measures for a Population

Descriptive statistics discussed previously described a ***sample*, not the *population*.**

Summary measures describing a population, called **parameters**, are denoted with Greek letters.

Important population parameters are the population **mean, variance, and standard** deviation.

Numerical Descriptive Measures for a Population: The mean μ

- The population mean is the sum of the values in the population divided by the population size, N

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

Where

μ = population mean

N = population size

X_i = i^{th} value of the variable X

Numerical Descriptive Measures For A Population: The Variance σ^2

Average of squared deviations of values from the mean

– Population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Where μ = population mean

N = population size

X_i = i^{th} value of the variable X

Numerical Descriptive Measures For A Population: The Standard Deviation σ

- Most commonly used measure of variation
- **Shows variation about the mean**
- Is the square root of the population variance
- Has the same units as the original data

Population standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

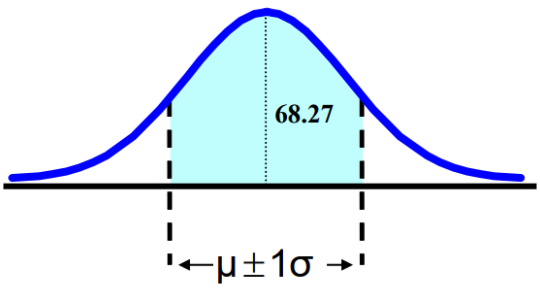
Sample statistics versus population parameters

Measure	Population Parameter	Sample Statistic
Mean	μ	$\overline{X}$
Variance	σ^2	S^2
Standard Deviation	σ	S

Numerical Descriptive Measures: The Empirical Rule for distribution of data

The empirical rule approximates the variation of data in **a bell-shaped** distribution

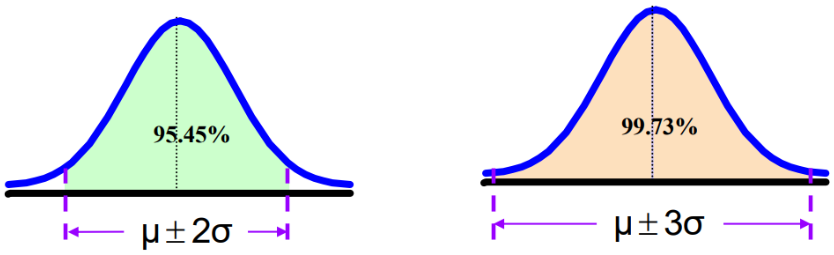
Approximately 68. 27% of the data in a bell shaped distribution is within 1 standard deviation of the mean or $\mu \pm 1\sigma$



The Empirical Rule

Approximately 95.45% of the data in a bell-shaped distribution lies within two standard deviations of the mean, or $\mu \pm 2\sigma$

Approximately 99.73% of the data in a bell-shaped distribution lies within three standard deviations of the mean, or $\mu \pm 3\sigma$



Using the Empirical Rule

Suppose that the variable Math SAT scores is bell-shaped with a **mean of 500 and a standard deviation of 90**. Then,

- 68.27% of all test takers scored between 410 and 590 ???
- 95.45% of all test takers scored between 320 and 680 ???
- 99.73% of all test takers scored between 230 and 770 ???

Numerical Descriptive Chebyshev Rule

Regardless of how the **data are distributed**, at least $(1 - 1/k^2) \times 100\%$ of the values will fall within k standard deviations of the mean (for $k > 1$)

– Examples:

At least

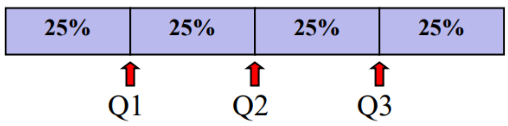
$(1 - 1/2^2) \times 100\% = 75\%$ .... $k=2$ $(\mu \pm 2\sigma)$

$(1 - 1/3^2) \times 100\% = 89\%$ .... $k=3$ $(\mu \pm 3\sigma)$

Another way of describing numerical data is through an <u>**exploratory data analysis**</u> that includes:
Quartile,
Five number summary,
and the Box plot.

Quartiles

Quartiles split the ranked **data into 4 segments with** an equal number of values per segment



- The first quartile, Q_1, is the value for which **25% of the observations are smaller** and 75% are larger
- Q_2 is the same as the median (50% of the observations are smaller and 50% are larger)
- Only 25% of the observations are **greater than the third quartile**

Locating Quartiles

Find a quartile by determining the value in the appropriate **position** in the ranked data, where

First quartile position:	Q1 = (n+1)/4 ranked value
Second quartile position:	Q2 = (n+1)/2 ranked value
Third quartile position:	Q3 = 3(n+1)/4 ranked value

- where n is the number of observed values

Calculation Rules

When calculating the ranked position use the following rules

- If the result is a **whole** number then it is the **ranked** position to use

- If the result is a **fractional half** (e.g. 2.5, 7.5, 8.5, etc.) then **average** the two corresponding data values.

- If the result is **not a whole number or a fractional half then round the result to the nearest integer** to find the ranked position.

Locating Quartiles

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

(n = 9)

Q_1 is in the $\boxed{(9+1)/4 = 2.5 \text{ position}}$ of the ranked data

so use the value half way between the 2nd and 3rd values,

so $\boxed{\mathbf{Q_1 = 12.5}}$

Q_1 and Q_3 are measures of **non-central location**
Q_2 = median, is a measure of **central tendency**

Quartile Example

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

(n = 9)

Q_1 is in the (9+1)/4 = 2.5 position of the ranked data,

so **Q_1 = (12+13)/2 = 12.5**

Q_2 is in the (9+1)/2 = 5th position of the ranked data,

so **Q_2 = median = 16**

Q_3 is in the 3(9+1)/4 = 7.5 position of the ranked data,

so **Q_3 = (18+21)/2 = 19.5**

Quartile Measures: The Interquartile Range (IQR)

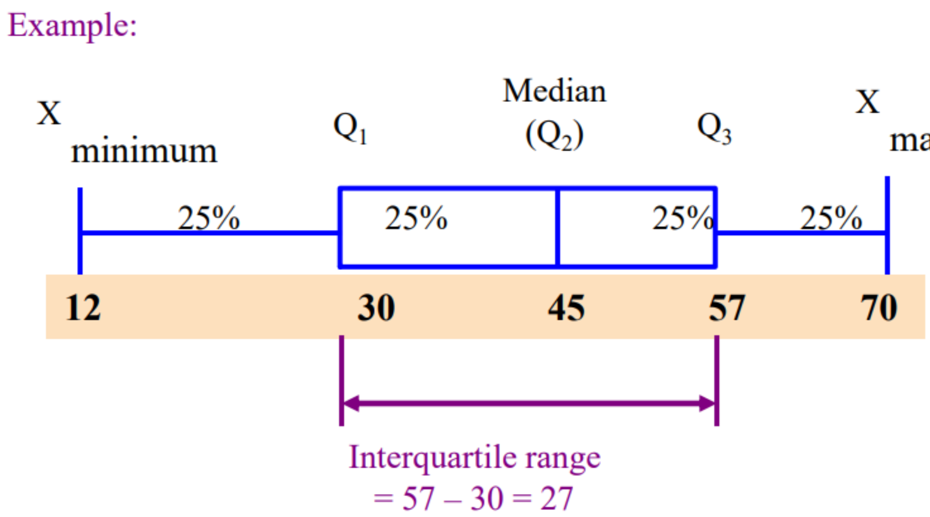
The IQR is $Q_3 – Q_1$ and measures the spread in the **middle 50% of** the data

The IQR is also called the **midspread** because it covers the middle 50% of the data

The IQR is ***a measure of variability that is not influenced by*** <u>**outliers**</u> or extreme values

Measures like Q_1, Q_3, and IQR that are ***not influenced by outliers are called*** <u>***resistant measures***</u>

Calculating The Interquartile Range



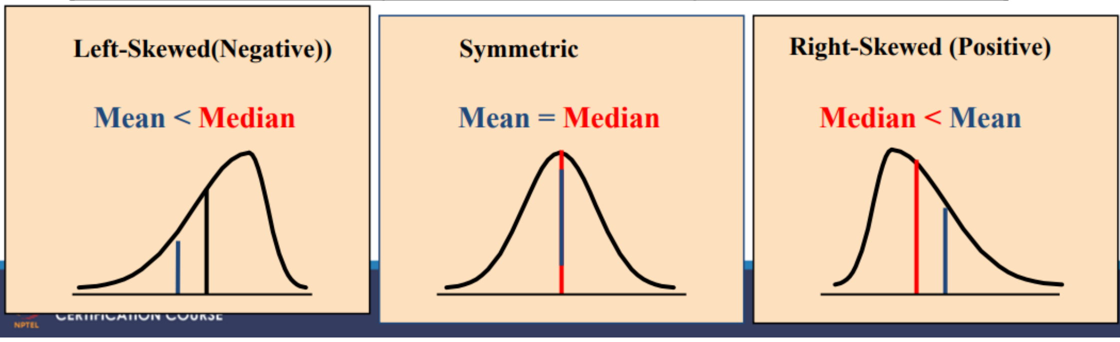
The Five Number Summary

The five numbers that help describe the **center, spread and shape of data are:**

- $X_{smallest}$
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- $X_{largest}$

Relationships among the five-number summary and distribution shape

Left-Skewed	Symmetric	Right-Skewed
Median – $X_{smallest}$ > $X_{largest}$ – Median	Median – $X_{smallest}$ ≈ $X_{largest}$ – Median	Median – $X_{smallest}$ < $X_{largest}$ – Median
Q_1 – $X_{smallest}$ > $X_{largest}$ – Q_3	Q_1 – $X_{smallest}$ ≈ $X_{largest}$ – Q_3	Q_1 – $X_{smallest}$ < $X_{largest}$ – Q_3
Median – Q_1 > Q_3 – Median	Median – Q_1 ≈ Q_3 – Median	Median – Q_1 < Q_3 – Median

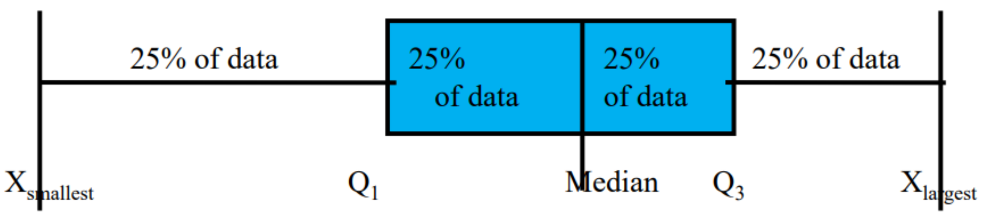


Five Number Summary and The Boxplot

- The Boxplot: A Graphical display of the data based on the five-number summary:

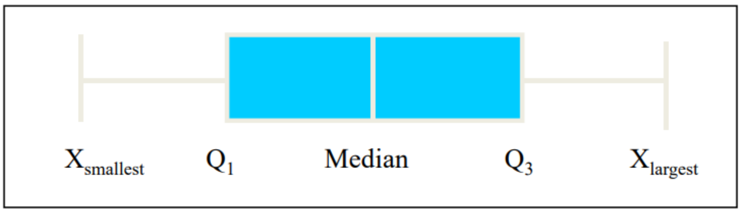
$X_{smallest}$ -- Q_1 -- Median -- Q_3 -- $X_{largest}$

Example:



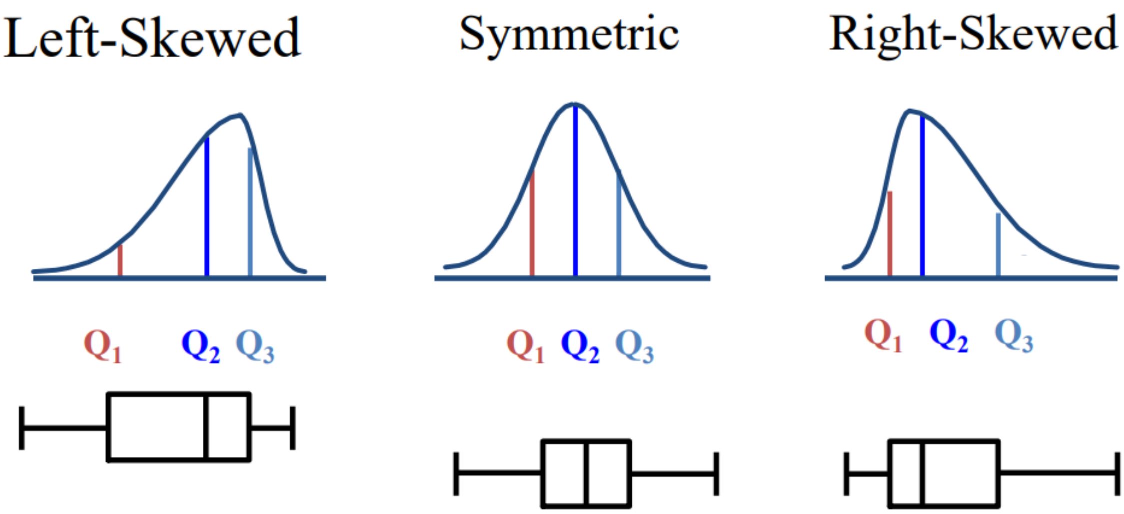
Five Number Summary: Shape of Boxplots

- If data **are symmetric around** the median then the box and central line are **centered between the endpoints**



- A Boxplot can be shown in either a **vertical or horizontal** orientation

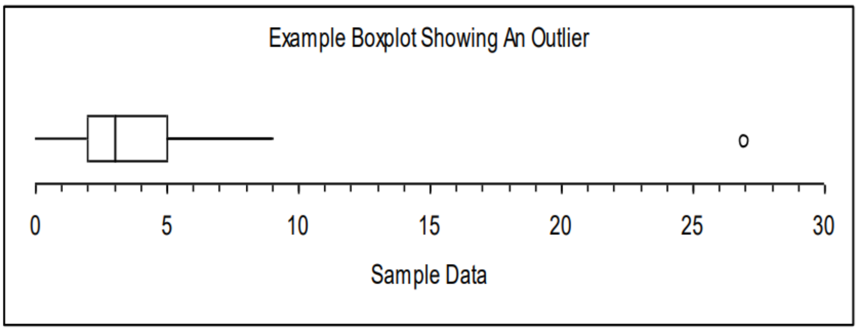
Distribution Shape and The Boxplot



Boxplot example showing an outlier

•The boxplot below of the same data shows the outlier **value of 27 plotted separately**

•A value is considered an outlier if it is **more than 1.5 times** the interquartile range **below Q_1** or **above Q_3**



Find outliers?

850	875	4700	4900	5300	5700	6700	7300	7700	8100
8300	8400	8700	8700	8900	9300	9500	9500	9700	10000
10300	10500	10700	10800	11000	11300	11300	11800	12700	12900
13100	13500	13800	14900	16300	17200	18500	20300	21310	21315

Q1= 8100
Q3=12900
IQR= 12900-8100=4800
1.5*IQR = 7200
Outliers= Q1-7200 =8100-7200=900
Outliers=Q3+7200=12900+7200=20100
Any data point below 900 and above 20100 are outliers.

Z score can also be used to know outliers

xi	xi-$\bar{x}$	(xi-x)/SD
240	-140	-1.237437797
260	-120	-1.060660969
350	-30	-0.265165242
350	-30	-0.265165242
420	40	0.353553656
510	130	1.149049383
530	150	1.325826211
Mean 380		

SD= 113

Is there any outlier???

Relationship between <u>two numerical </u>variable

1 Covariance

2 Coefficient of correlation

- The **covariance** measures the **strength of the linear** relationship between **two numerical variables** (X & Y)

- The sample covariance:

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

- Only concerned with the ***<u>strength of the relationship</u>***
- **No causal effect** is implied

Interpreting Covariance

- **Covariance** between two variables:

cov(X,Y) > 0 ⟶ X and Y tend to move in the same direction

cov(X,Y) < 0 ⟶ X and Y tend to move in opposite directions

cov(X,Y) = 0 ⟶ X and Y are independent

- The covariance has a major flaw:
  - It is not possible to determine the ***relative strength of the relationship*** from the **size** of the covariance

Find covariance??

Sr no	City	Hamburger (x)	Movie Tickets (y)
1	Tokyo	5.99	32.66
2	London	7.62	28.41
3	New York	5.75	20.00
4	Sydney	4.45	20.71
5	Chicago	4.99	18.00
6	San Francisco	5.29	19.50
7	Boston	4.39	18.00
8	Atlanta	3.7	16.00
9	Toronto	4.62	18.05
10	Rio de Janeiro	2.99	9.90
Avg		4.98	20.12

Sr no	City	Hamburger (x)	Movie Tickets (y)	(x-x bar)*(y-ybar)
1	Tokyo	5.99	32.66	12.6654
2	London	7.62	28.41	21.8856
3	New York	5.75	20.00	-0.0924
4	Sydney	4.45	20.71	-0.3127
5	Chicago	4.99	18.00	-0.0212
6	San Francisco	5.29	19.50	-0.1922
7	Boston	4.39	18.00	1.2508
8	Atlanta	3.7	16.00	5.2736
9	Toronto	4.62	18.05	0.7452
10	Rio de Janeiro	2.99	9.90	20.3378
Avg		4.98	20.12	Sum= 61.53

Covariance = 61.53/9=6.83, we can't tell whether this value is an indictor of strong or weak relationship.

Coefficient of Correlation

- Measures the relative strength of the linear relationship between two numerical variables
- Sample coefficient of correlation:

$$\boxed{r = \frac{\operatorname{cov}(X,Y)}{S_X S_Y}}$$

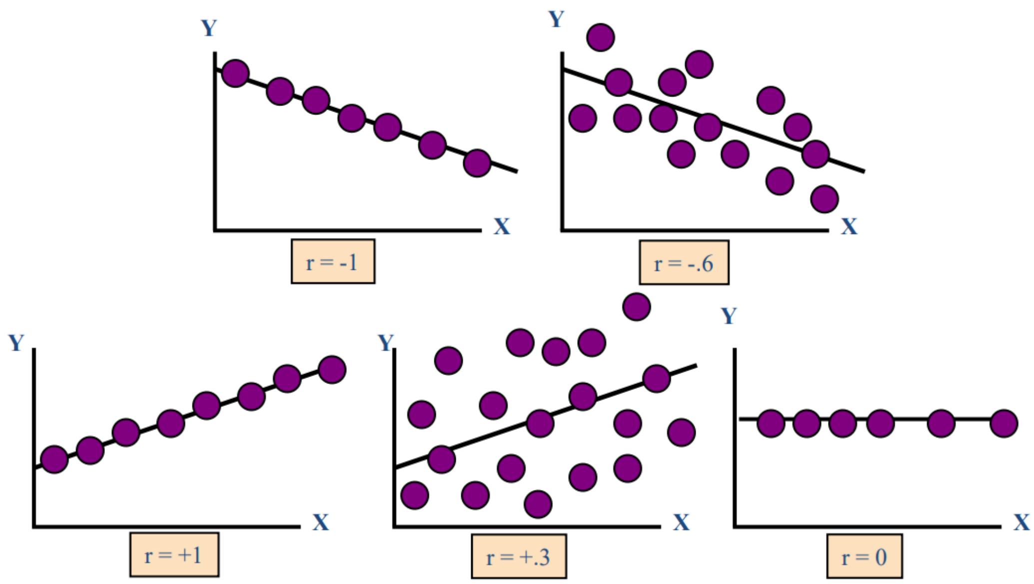
where

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \qquad S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \qquad S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

Features of the Coefficient of Correlation

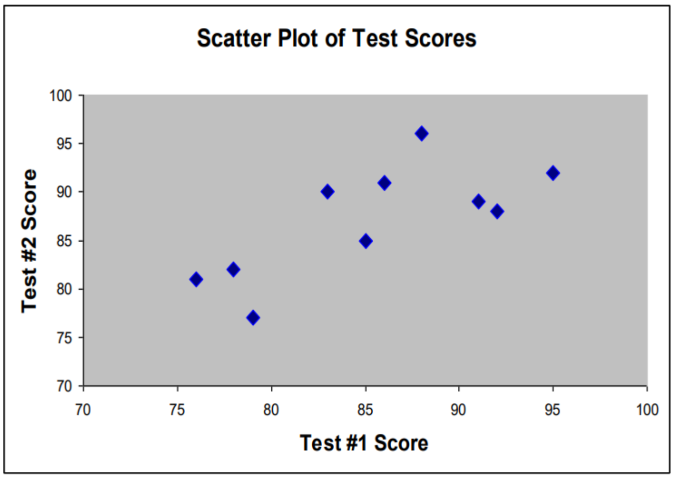
- The **population** coefficient of correlation is referred as **ρ**.
- The **sample** coefficient of correlation is referred to as **r**.
- Either ρ or r have the following **features**:
  - Unit free
  - Ranges between **–1 and 1**
  - The closer to –1, the stronger the **negative linear** relationship
  - The closer to 1, the stronger the **positive linear** relationship
  - The closer to 0, the **weaker** the **linear** relationship

Scatter Plots of Sample Data with Various Coefficients of Correlation



Interpreting the Coefficient of Correlation Using Microsoft Excel

- r = .733
- There is a relatively strong positive linear relationship between test score #1 and test score #2.
- Students who scored high on the first test tended to score high on second test.



Product	Calories	Fat
Dunkin' Donuts Iced Mocha Swirl latte (whole milk)	240	8
Starbucks Coffee Frappuccino blended coffee	260	3.5
Dunkin' Donuts Coffee Coolatta (cream)	350	22
Starbucks Iced Coffee Mocha Expresso (whole milk and whipped cream	350	20
Starbucks Mocha Frappuccino blended coffee (whipped cream)	420	16
Starbucks Chocolate Brownie Frappuccino blended coffee (whipped cream)	510	22
Starbucks Chocolate Frappuccino Blended Crème (whipped cream)	530	19

a) Compute covariance
b) Compute coefficient of correlation
c) Which is valuable in expressing relationship
d) What conclusion can you reach about relationship

a) Compute covariance : 591.66
b) Compute coefficient of correlation: r = 0.71
c) Which is valuable in expressing relationship: **correlation**
d) What conclusion can you reach about relationship: strong positive relationship