Class Task + Lab Report: 2

Fall 2021

CSE422L Data Analytics Lab

Submitted by: **Ayaz Mehmood**

Registration No.:**18PWCSE1652**

Section: **A**

“On my honor, as student of University of Engineering and Technology, I have neither given nor received unauthorized assistance on this academic work.”

Student Signature: _______________

Submitted to:

Engr. Mian Ibad Ali Shah

Last date of Submission:

Sunday, 14 November 2021

Department of Computer Systems Engineering

University of Engineering and Technology, Peshawar

OBJECTIVE:

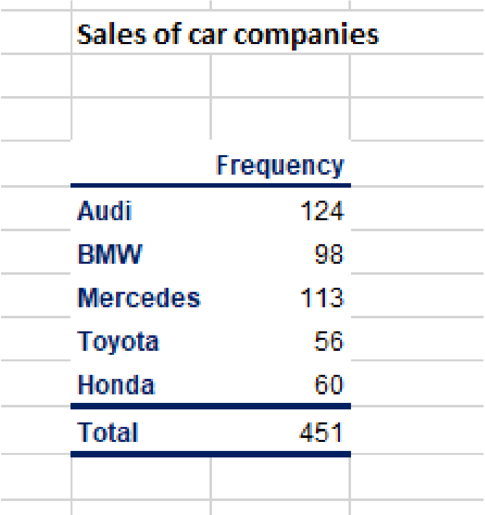
The basic Objective of this lab is:

- To know about the types of data
- To know different level of measurement
- To know and use different visualization Technique
- To know and use different statistic measure.

TASKS

Categorical Variable:

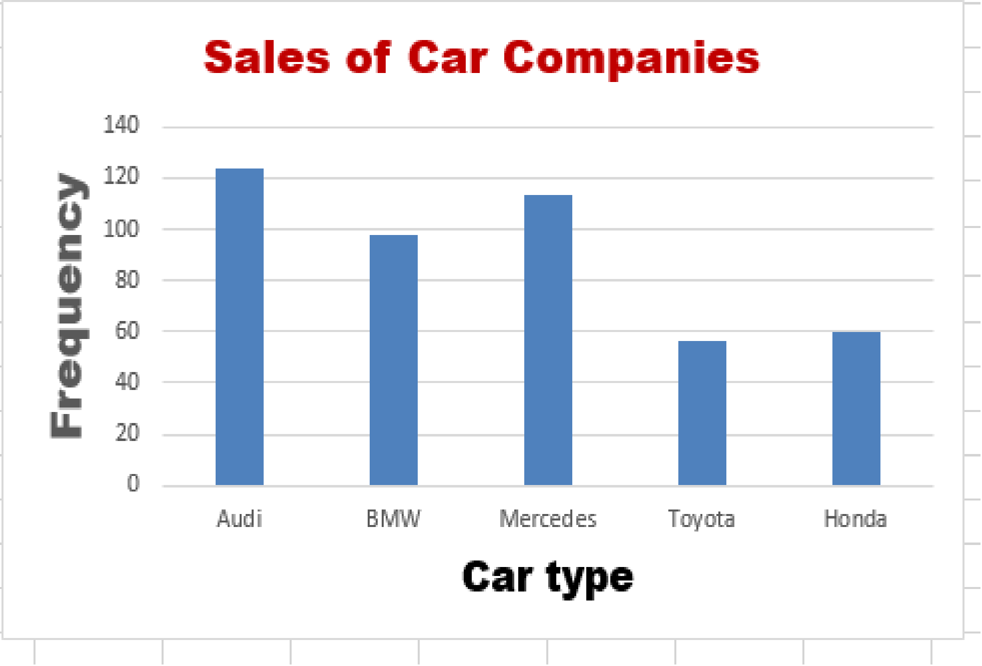
A variable that can take on one of a limited, and usually fixed number of possible values. For example: Yes/No, Male/Female, Car brands etc.



Sales of car companies

	Frequency
Audi	124
BMW	98
Mercedes	113
Toyota	56
Honda	60
Total	451

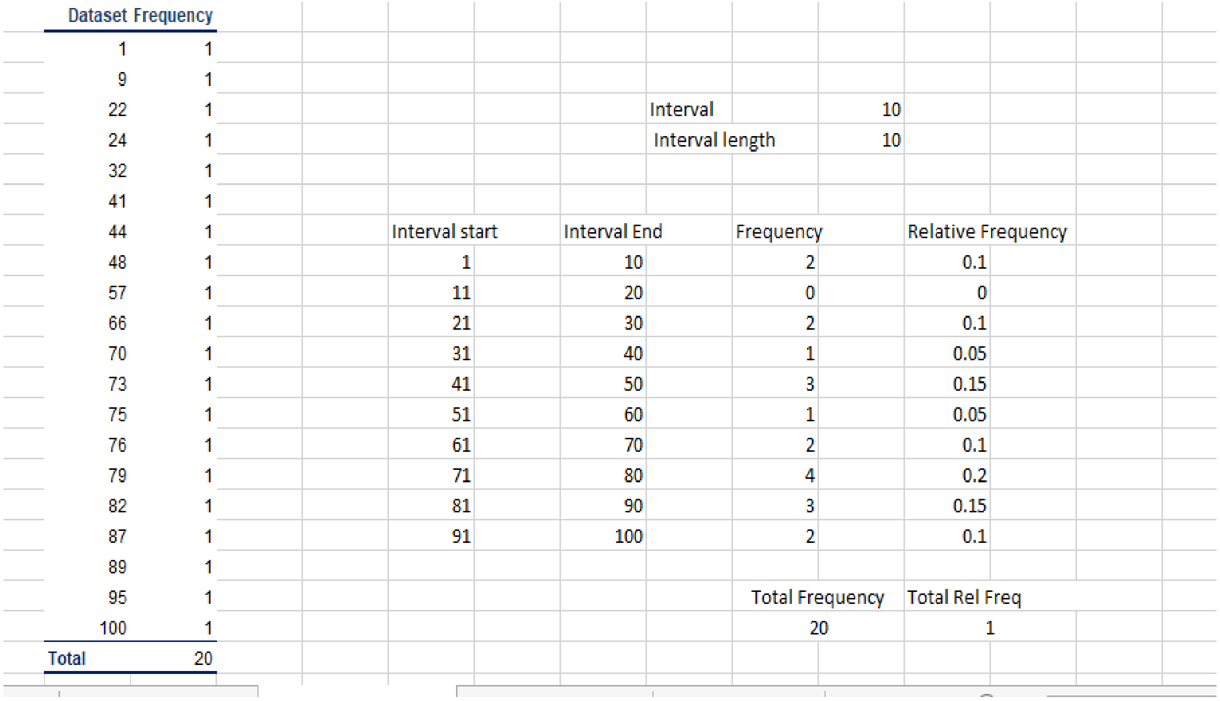
We have to Plot the frequency values



Numeric Variable:

Numeric variable have values that describe a measureable quantity as a number, like ‘how many’, ‘how much’.

For example: Human age, height of a person etc.



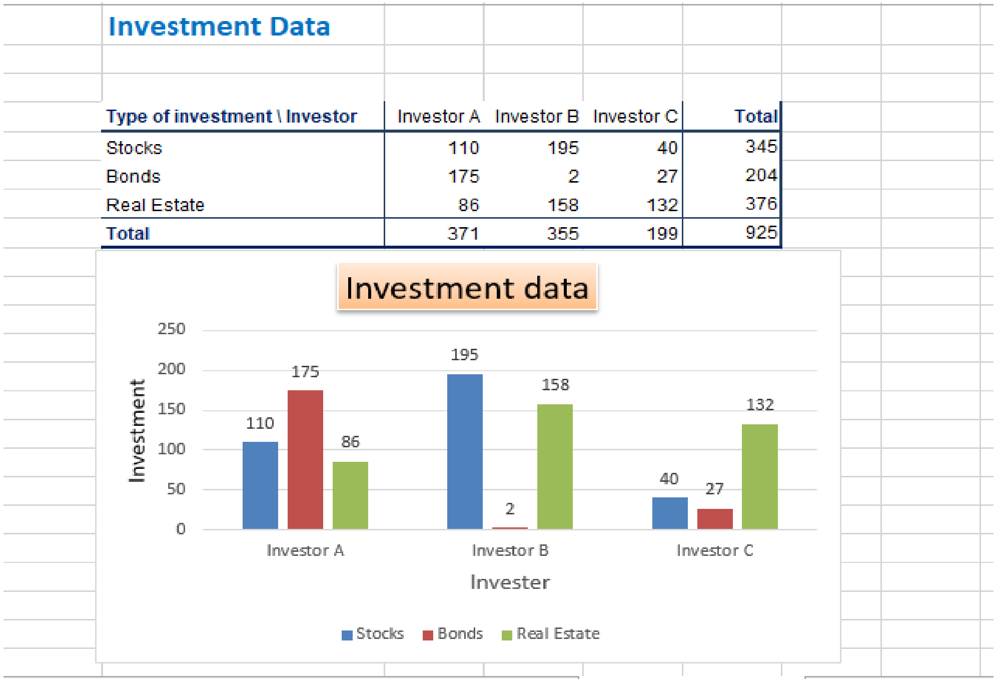
Dataset	Frequency
1	1
9	1
22	1
24	1
32	1
41	1
44	1
48	1
57	1
66	1
70	1
73	1
75	1
76	1
79	1
82	1
87	1
89	1
95	1
100	1
Total	20

Interval	10
Interval length	10

Interval start	Interval End	Frequency	Relative Frequency
1	10	2	0.1
11	20	0	0
21	30	2	0.1
31	40	1	0.05
41	50	3	0.15
51	60	1	0.05
61	70	2	0.1
71	80	4	0.2
81	90	3	0.15
91	100	2	0.1

Total Frequency	Total Rel Freq
20	1

Cross Table Dataset:

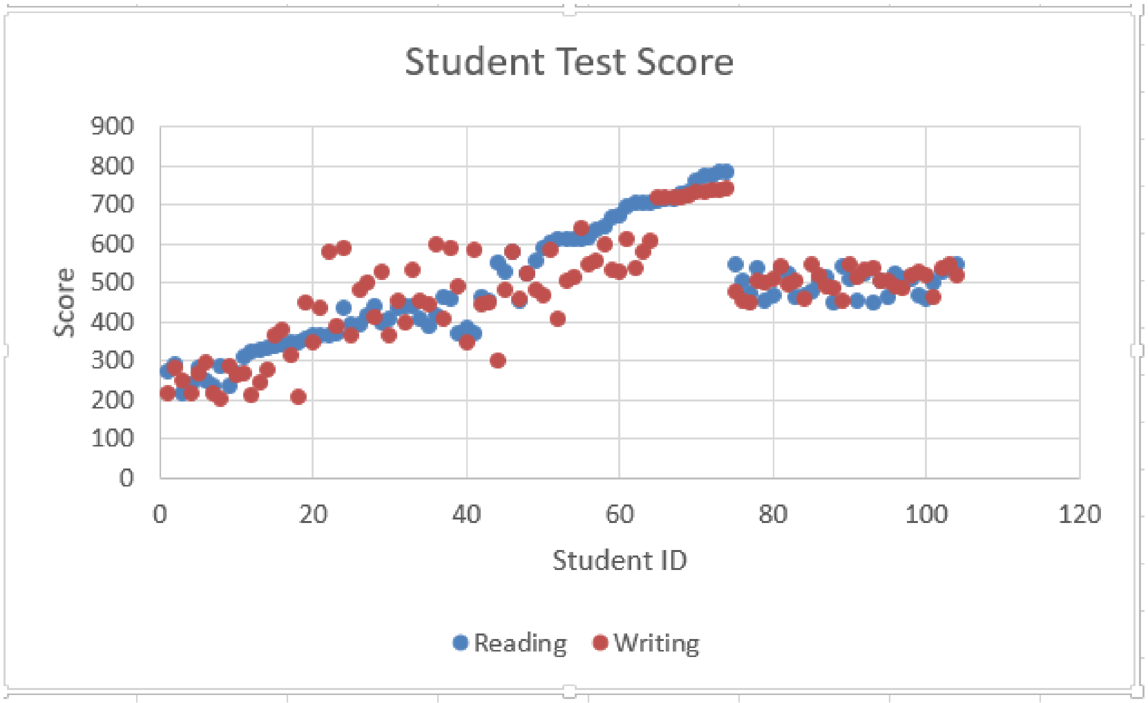


Investment Data

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	110	195	40	345
Bonds	175	2	27	204
Real Estate	86	158	132	376
Total	371	355	199	925

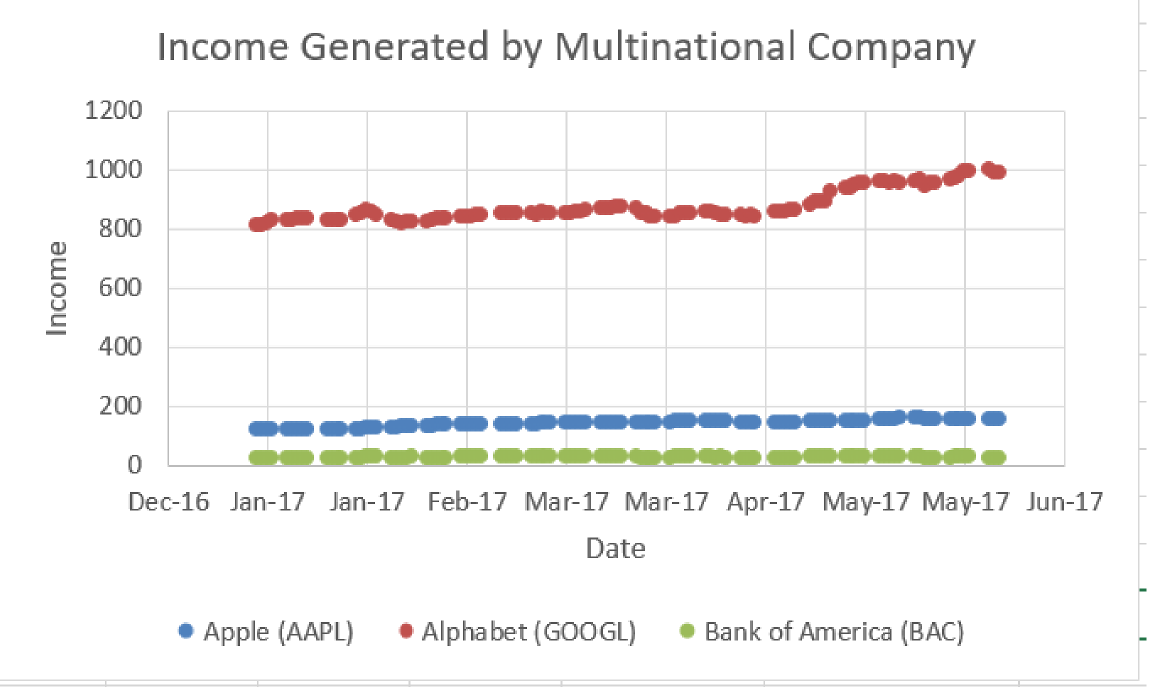
In the graph each inverter is represented in X axis, each investor Invest in stocks, Bonds and Real Estate. The Y axis has the investment

Scatter Plot Data:
Part 1 Plot:



Blue points represent reading score of a student and red represent Writing score

Part 2 Plot:



Blue points represent AAPL income, red represent GOOGL and green represent BAC

Measure of Central Tendency:

Measure of central tendency help us to find the middle or the average of a dataset.

The most common measures of Central tendency are the **mean**, **median** and **mode**

Mean:

Mean is equal to the sum of all the values in the data set divided by the number of values in the data set.

The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value

Here in the pizza price in Pakistan there is an outlier in Islamabad Pizza Price which is the amount 6600

Median:

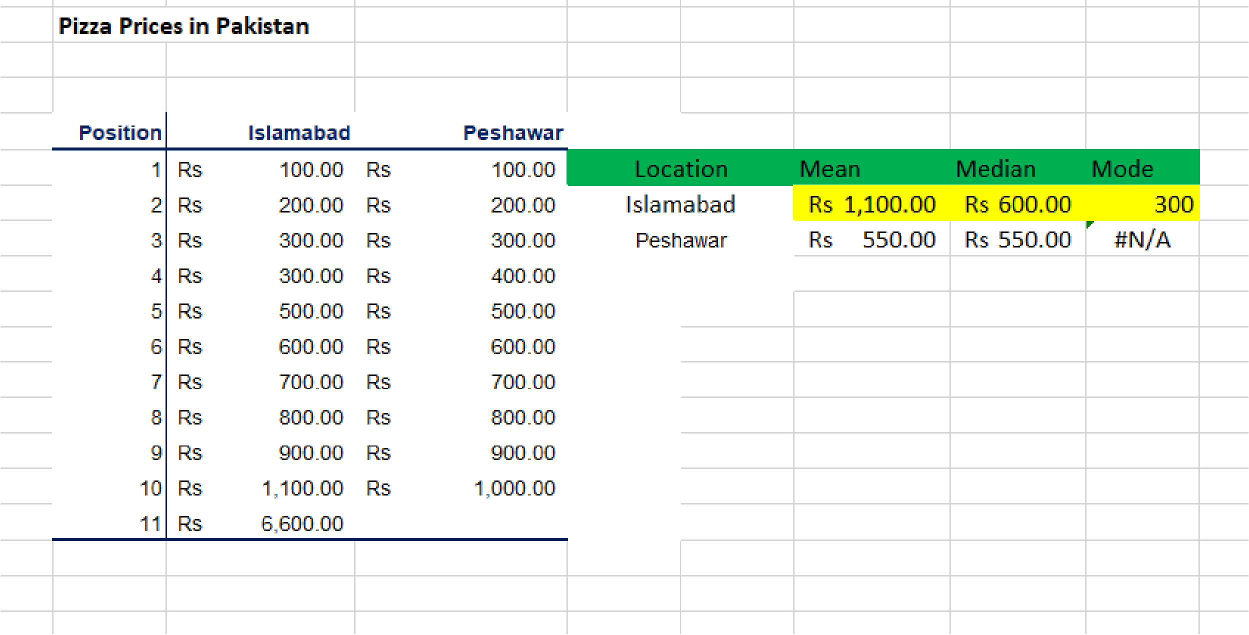
The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data.

Mode:

The mode is the most frequent score in our data set.

Normally Mode is used to categorical data where we wish to know the most common category.

In the below dataset we have given pizza price in Islamabad and Peshawar



Pizza Prices in Pakistan

Position	Islamabad	Peshawar
1	Rs 100.00	Rs 100.00
2	Rs 200.00	Rs 200.00
3	Rs 300.00	Rs 300.00
4	Rs 300.00	Rs 400.00
5	Rs 500.00	Rs 500.00
6	Rs 600.00	Rs 600.00
7	Rs 700.00	Rs 700.00
8	Rs 800.00	Rs 800.00
9	Rs 900.00	Rs 900.00
10	Rs 1,100.00	Rs 1,000.00
11	Rs 6,600.00	

Location	Mean	Median	Mode
Islamabad	Rs 1,100.00	Rs 600.00	300
Peshawar	Rs 550.00	Rs 550.00	#N/A

Variance:

Variance measures how far each number in the set is from the mean and thus from every other number in the set.

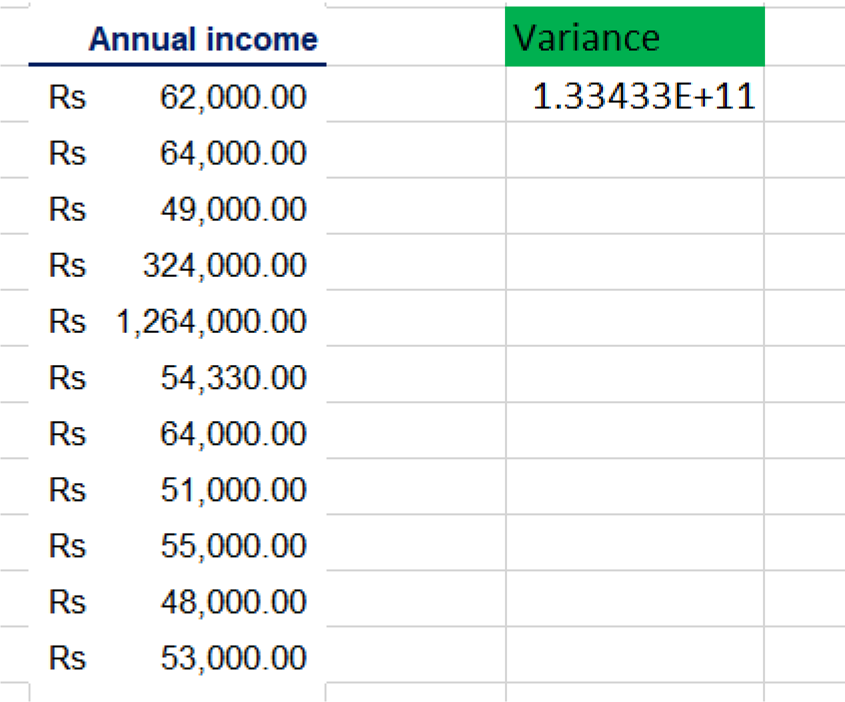
It is calculated by taking the differences between each number in the data set and the mean, then squaring the differences to make them positive, and finally dividing the sum of the squares by the number of values in the data set.

-Population Variance: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$

-Sample Variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

What does this no tell you?

The no tell us the spread of data from the mean value. This shows the variability in the dataset. Greater the no shows the no are far away from the mean value.



Annual income			Variance
Rs	62,000.00		1.33433E+11
Rs	64,000.00		
Rs	49,000.00		
Rs	324,000.00		
Rs	1,264,000.00		
Rs	54,330.00		
Rs	64,000.00		
Rs	51,000.00		
Rs	55,000.00		
Rs	48,000.00		
Rs	53,000.00		

Standard Deviation and Coefficient of Variation:

Standard Deviation

Standard deviation is the degree of dispersion or the scatter of the data points relative to its mean

STD is like Variance but under root in order to deal with large values of variance

- Population SD: $\sigma = \sqrt{\sigma^2}$

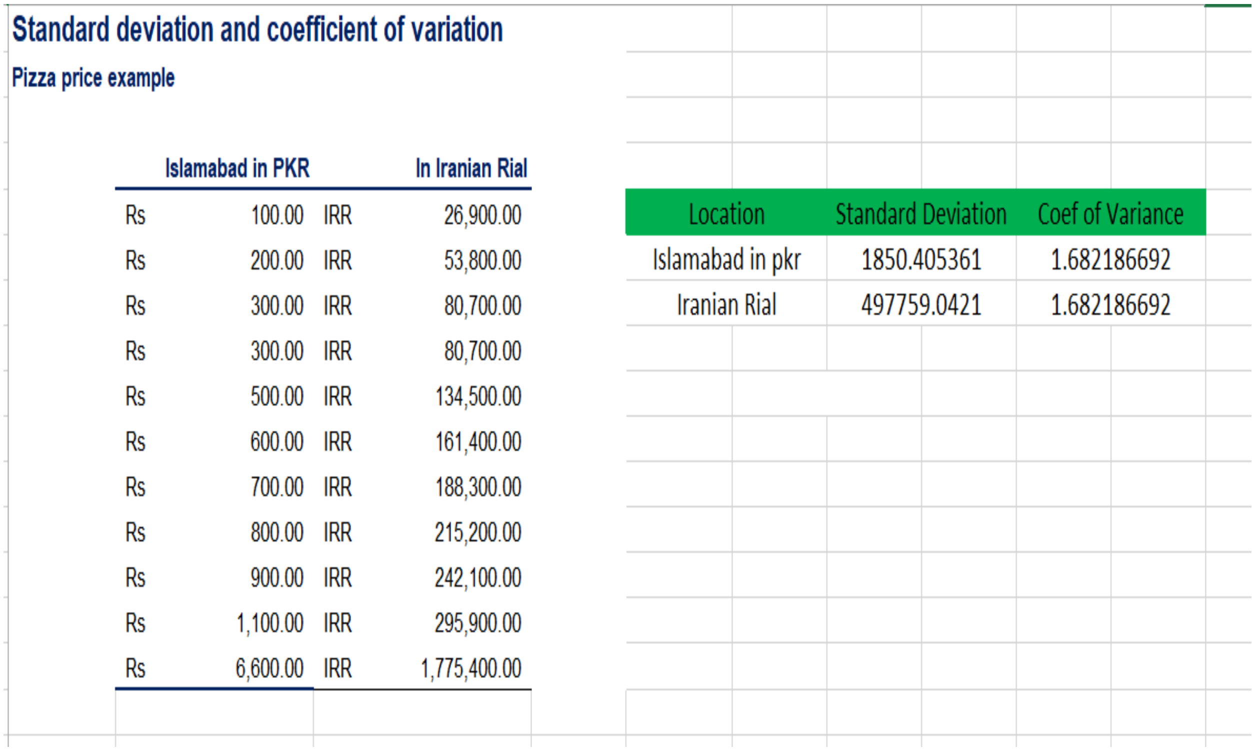
-Sample SD: $s = \sqrt{s^2}$

Coefficient of Variation:

The coefficient of variation (CV) is a measure of relative variability. It is the ratio of the standard deviation to the mean (average)

The coefficient of Variation come into play because variance and standard deviation are scale dependent.

Here in the below example we have the same thing but in different scale/Currency. That's why the STD result is different while the coefficient of variation are the same.



Standard deviation and coefficient of variation

Pizza price example

	Islamabad in PKR		In Iranian Rial
Rs	100.00	IRR	26,900.00
Rs	200.00	IRR	53,800.00
Rs	300.00	IRR	80,700.00
Rs	300.00	IRR	80,700.00
Rs	500.00	IRR	134,500.00
Rs	600.00	IRR	161,400.00
Rs	700.00	IRR	188,300.00
Rs	800.00	IRR	215,200.00
Rs	900.00	IRR	242,100.00
Rs	1,100.00	IRR	295,900.00
Rs	6,600.00	IRR	1,775,400.00

Location	Standard Deviation	Coef of Variance
Islamabad in pkr	1850.405361	1.682186692
Iranian Rial	497759.0421	1.682186692

Correlation and Covariance:

Covariance:

Covariance is a measure of how much two random variables vary together. It does not measure the variation amount

-Population covariance: $\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x) * (y_i - \mu_y)}{N}$

-Sample covariance: $s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$

Correlation:

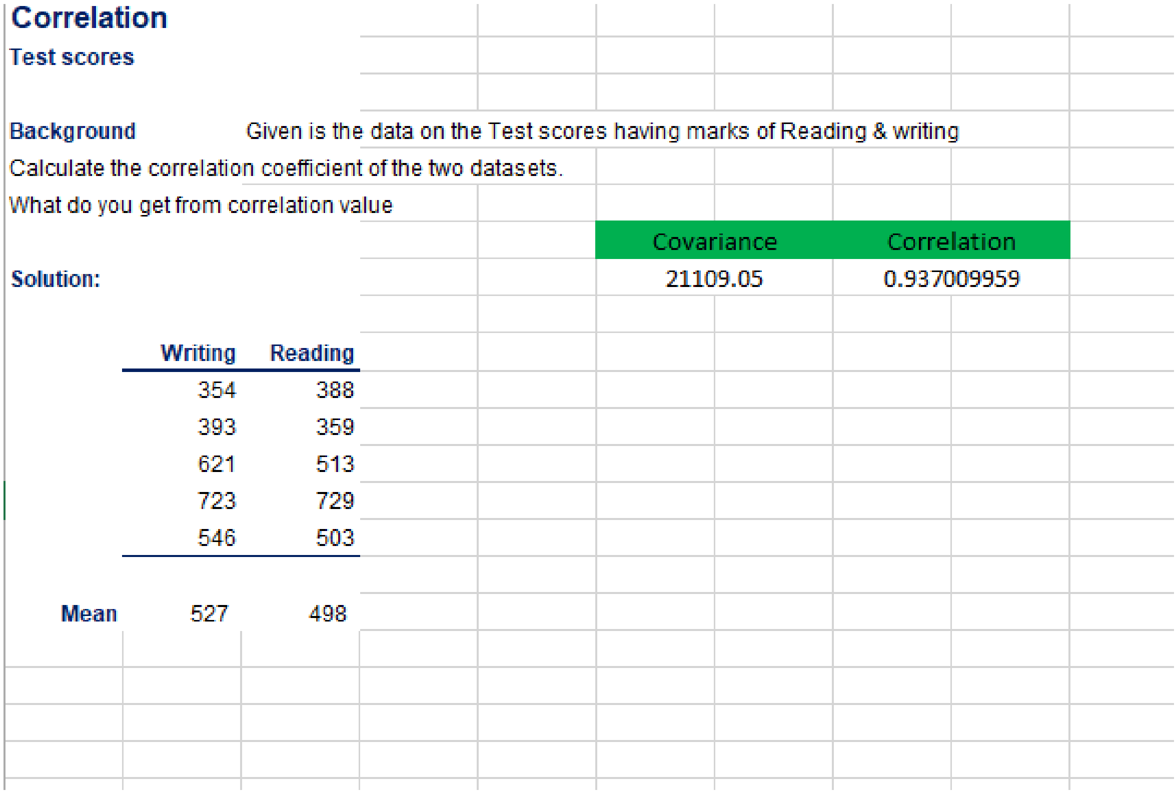
Correlation means association. It is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.

A **positive correlation** is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases

A **negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other

A **zero correlation** exists when there is no relationship between two variables.

Correlation= $\frac{CoV(x,y)}{SD(x) * SD(y)}$



Correlation

Test scores

Background Given is the data on the Test scores having marks of Reading & writing

Calculate the correlation coefficient of the two datasets.

What do you get from correlation value

Covariance	Correlation
21109.05	0.937009959

Solution:

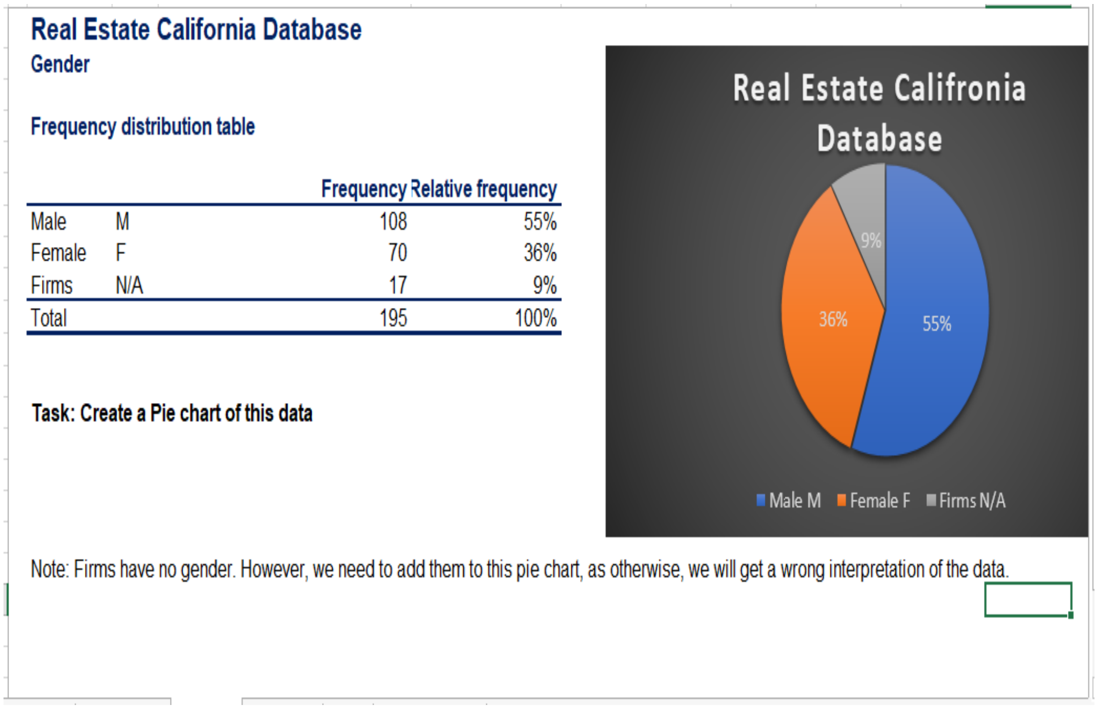
	Writing	Reading
	354	388
	393	359
	621	513
	723	729
	546	503
Mean	527	498

What do you get from Correlation values?

It shows strong positive correlation. The students who have higher scores in writing also have higher scores in reading.

HOME TASK

Gender:



Real Estate California Database

Gender

Frequency distribution table

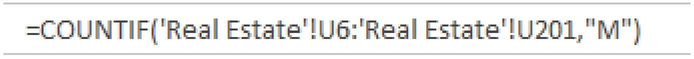
		Frequency	Relative frequency
Male	M	108	55%
Female	F	70	36%
Firms	N/A	17	9%
Total		195	100%

Task: Create a Pie chart of this data

Real Estate Califronia Database
9%
36%
55%
■ Male M ■ Female F ■ Firms N/A

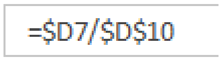
Note: Firms have no gender. However, we need to add them to this pie chart, as otherwise, we will get a wrong interpretation of the data.

Formula of Frequency



=COUNTIF('Real Estate'!U6:'Real Estate'!U201,"M")

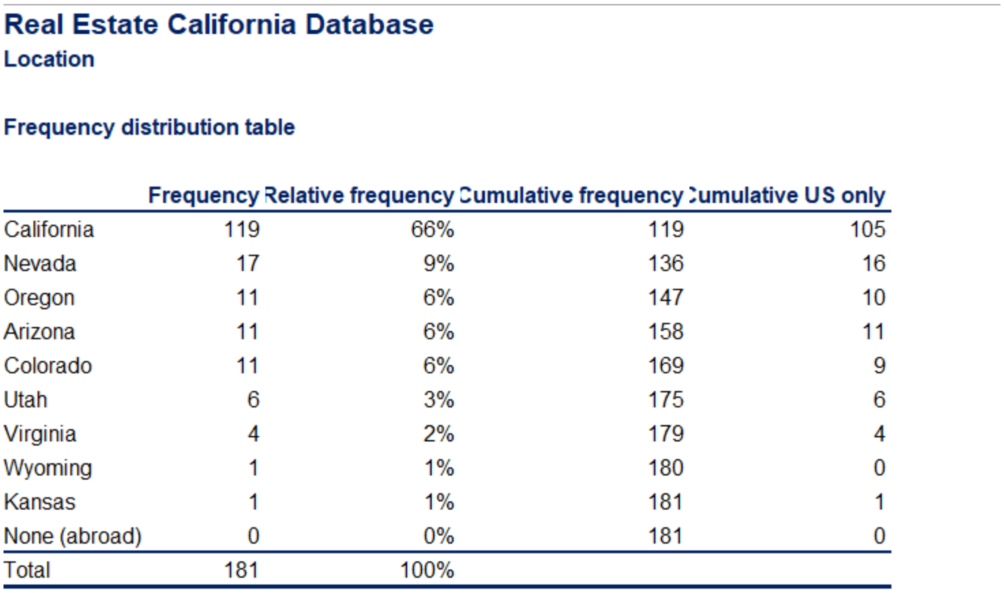
Formula of Relative Frequency



=$D7/$D$10

D7 is the frequency component of Male and D10 is the total frequency

Location:



Real Estate California Database

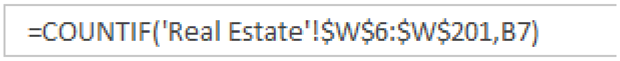
Location

Frequency distribution table

	Frequency	Relative frequency	Cumulative frequency	Cumulative US only
California	119	66%	119	105
Nevada	17	9%	136	16
Oregon	11	6%	147	10
Arizona	11	6%	158	11
Colorado	11	6%	169	9
Utah	6	3%	175	6
Virginia	4	2%	179	4
Wyoming	1	1%	180	0
Kansas	1	1%	181	1
None (abroad)	0	0%	181	0
Total	181	100%		

Total Frequency=181

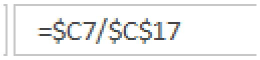
Formula of Frequency



=COUNTIF('Real Estate'!$W$6:$W$201,B7)

Sum of Relative frequency should be 100%.

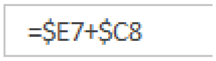
Formula of Relative Frequency



=$C7/$C$17

Last class Cumulative frequency should be equal to total frequency

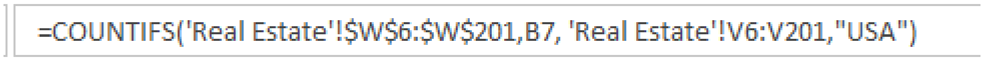
Formula of Cumulative Frequency



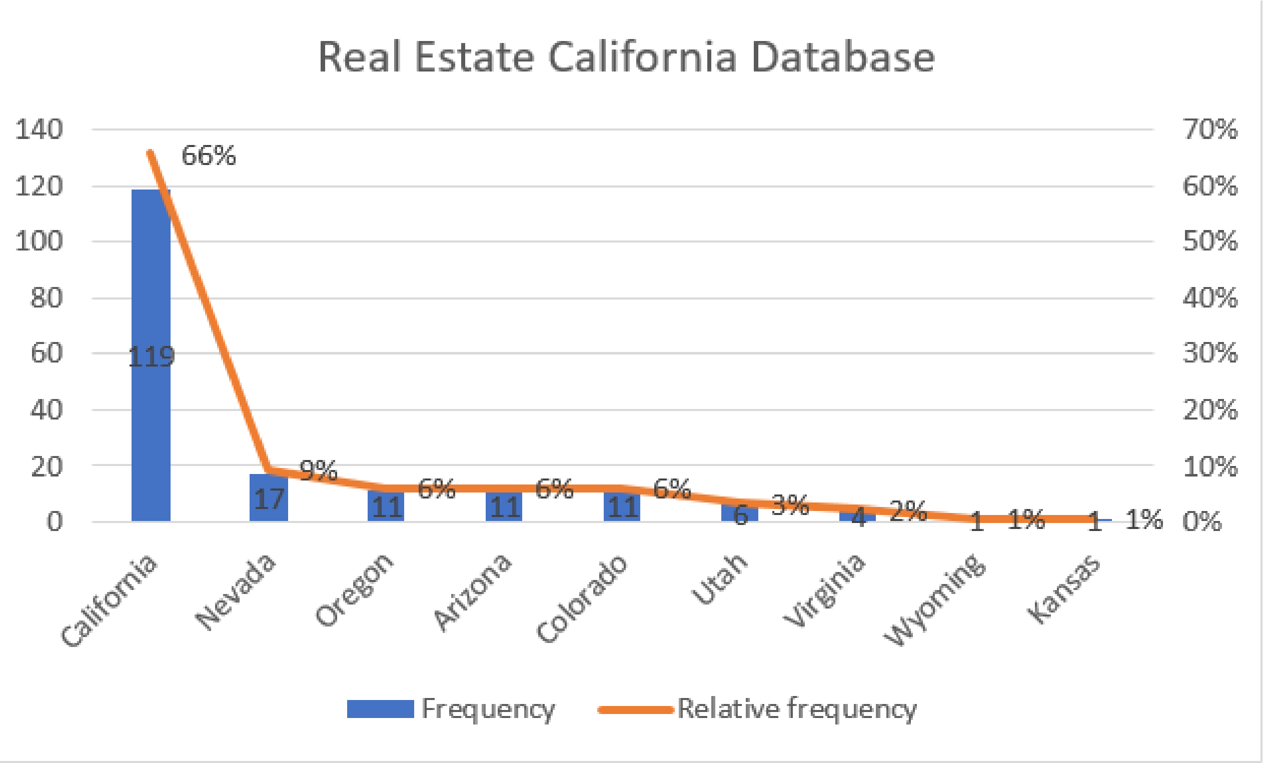
=$E7+$C8

Formula of Cumulative Frequency for USA only

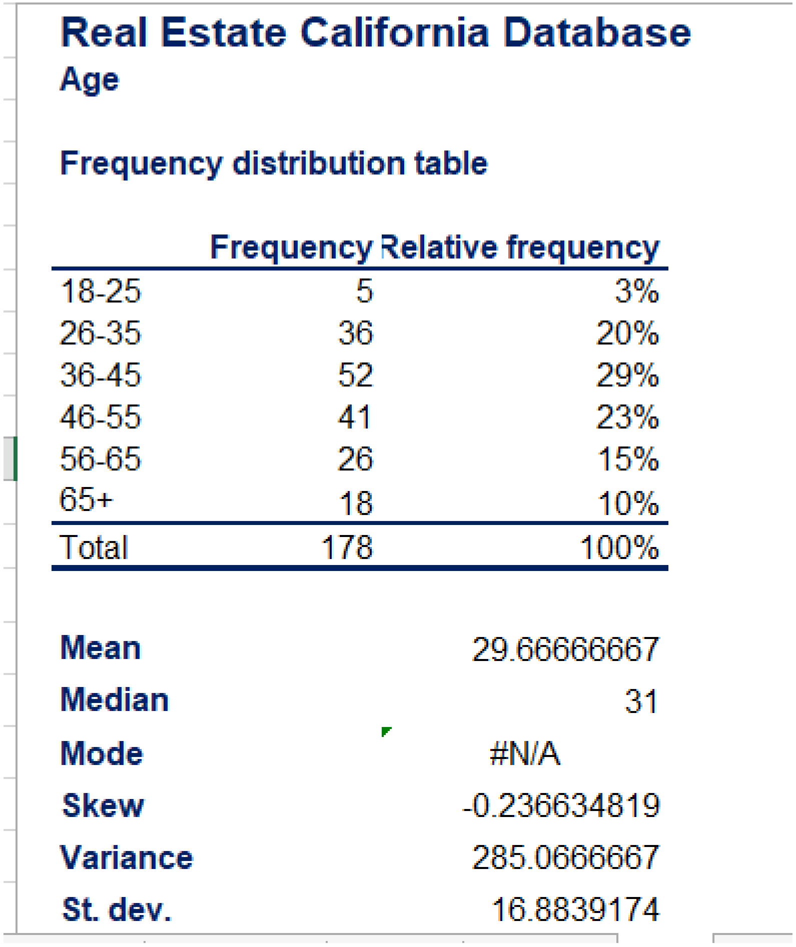
Find out the frequency of places of USA only



=COUNTIFS('Real Estate'!$W$6:$W$201,B7, 'Real Estate'!V6:V201,"USA")



Age:



Real Estate California Database

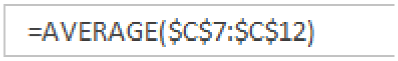
Age

Frequency distribution table

	Frequency	Relative frequency
18-25	5	3%
26-35	36	20%
36-45	52	29%
46-55	41	23%
56-65	26	15%
65+	18	10%
Total	178	100%

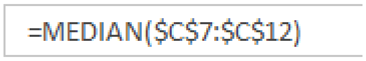
Mean	29.66666667
Median	31
Mode	#N/A
Skew	-0.236634819
Variance	285.0666667
St. dev.	16.8839174

Mean formula:



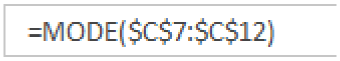
=AVERAGE($C$7:$C$12)

Median Formula:



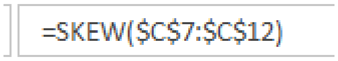
=MEDIAN($C$7:$C$12)

Mode Formula:



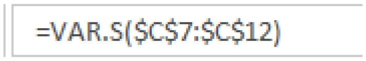
=MODE($C$7:$C$12)

Skewness Formula:



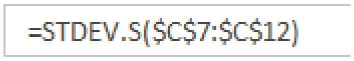
=SKEW($C$7:$C$12)

Variance Formula:

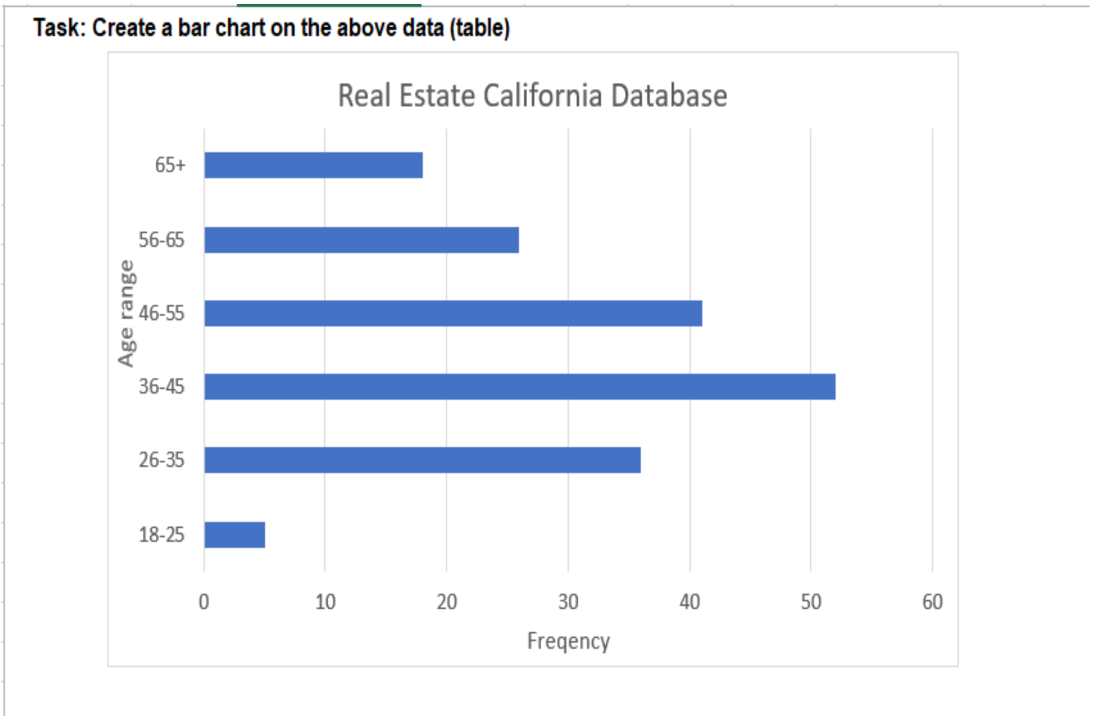


=VAR.S($C$7:$C$12)

St.dev Formula:

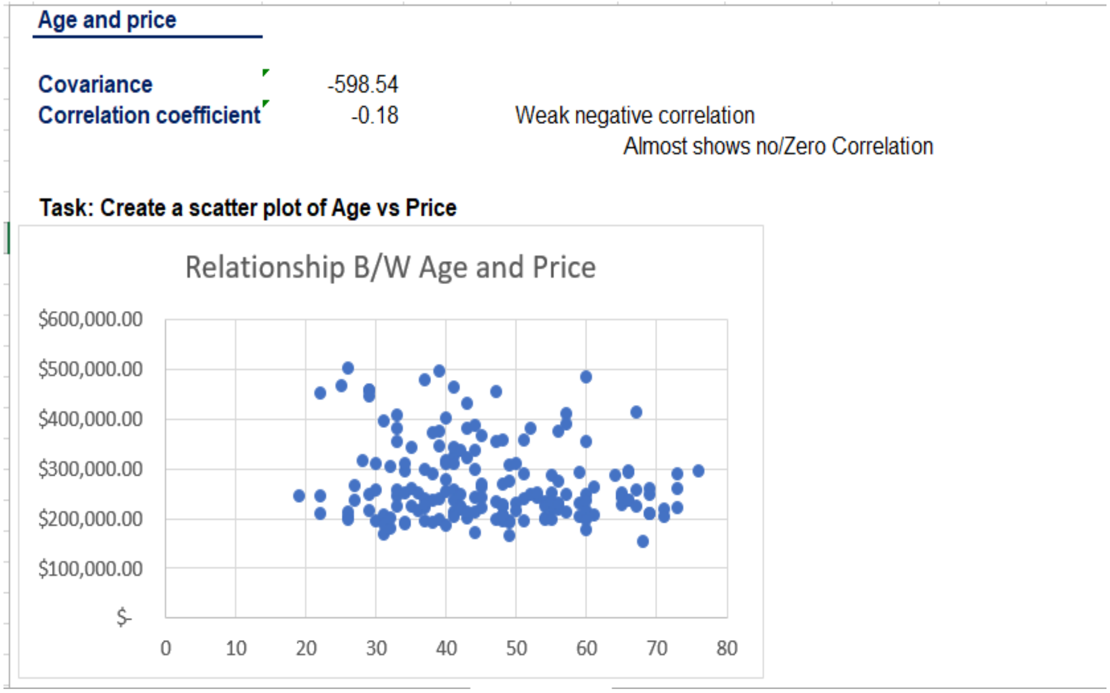


=STDEV.S($C$7:$C$12)



Task: Create a bar chart on the above data (table)

Age and Price Relationship:

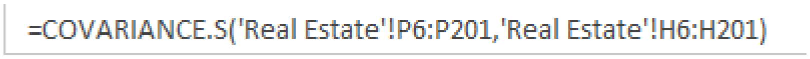


Age and price

Covariance	-598.54	
Correlation coefficient	-0.18	Weak negative correlation
		Almost shows no/Zero Correlation

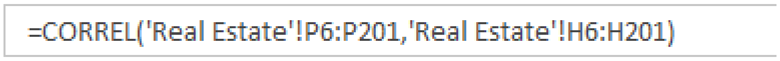
Task: Create a scatter plot of Age vs Price

Covariance Formula:



=COVARIANCE.S('Real Estate'!P6:P201,'Real Estate'!H6:H201)

Correlation Formula:



=CORREL('Real Estate'!P6:P201,'Real Estate'!H6:H201)