Correlation Analysis

COMP 1810 –Data and Web Analytics.

Correlation Analysis

Correlation Analysis explains relationship between variables. It is concerned majorly with Correlation and Covariance.

Covariance and Correlation practically measure the same relationship between two random variables, but with different approach.

While Covariance is a measure of the relationship between two random variables and from infinite -∞ to +∞ values, correlation measure from –1 to + 1. Therefore, correlation values are standardized, hence make more sense.

Why Correlation is used instead of Covariance.

In comparing two variables Covariance does not provide enough and accurate information as regard to the strength of the relationship between two variables, imagine a calculated covariance between two sets of variables A and B or C and D could be 0.45 or 67 or 20000 or even from -∞ to +∞. How do we compare these values? Therefore, we used correlation because the values are standardized eg from –1 to +1. Correlation is the normalized version of covariance. It shows both relationship and strength of the relationship.

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$cov_{x,y}$ = covariance between variable x and y

x_i = data value of x

y_i = data value of y

$\bar{x}$ = mean of x

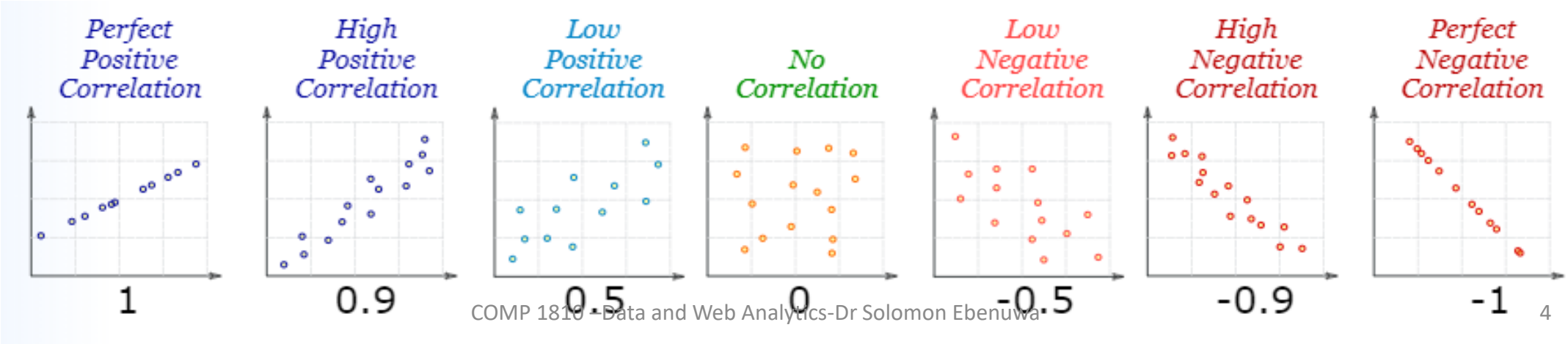
$\bar{y}$ = mean of y

N = number of data values

Correlation

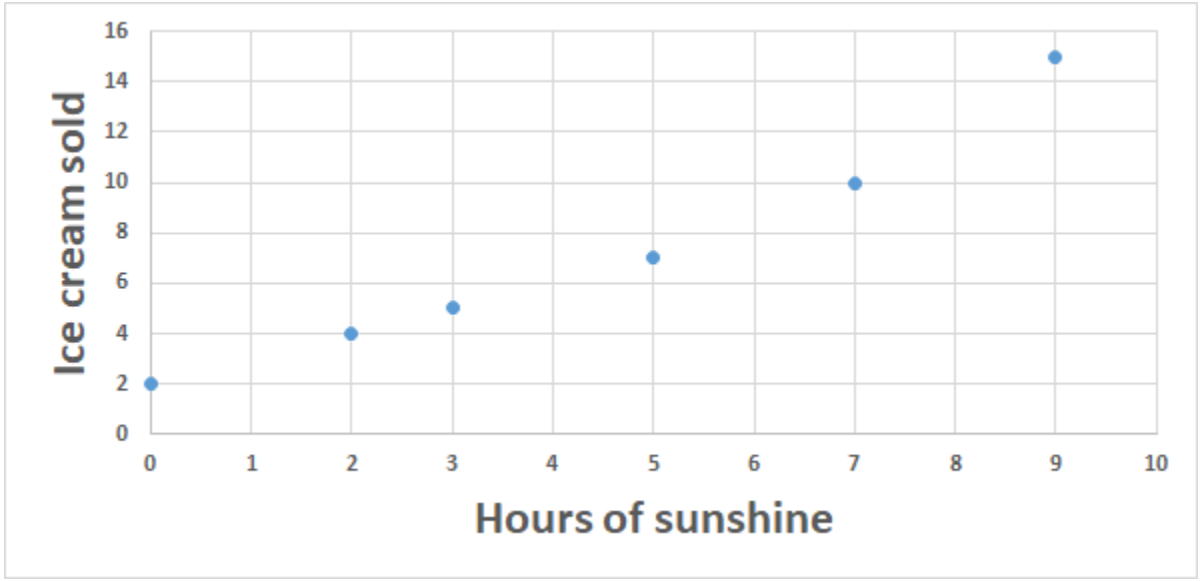
Correlation: Is the linear relation between two variables. An example is a correlation between temperature of the day and the amount of ice cream sold.

When the temperature of the day is high more ice cream is sold. This is **positive correlation** since both are moving in the same direction (High Temperature, More ice cream). Two variables may also **correlate negatively**. Both moving in opposite direction. An example of negative correlation is the amount of money you spend for heating vs the temperature of the day. That is heat amount get higher as temperature get lower. Correlation is measured between +1 to –1. Scattered point graph like and 10 are used to show correlation between two variables



Correlation

"x" Hours of Sunshine	"y" Ice Creams Sold
0	2
2	4
3	5
5	7
7	10
9	15

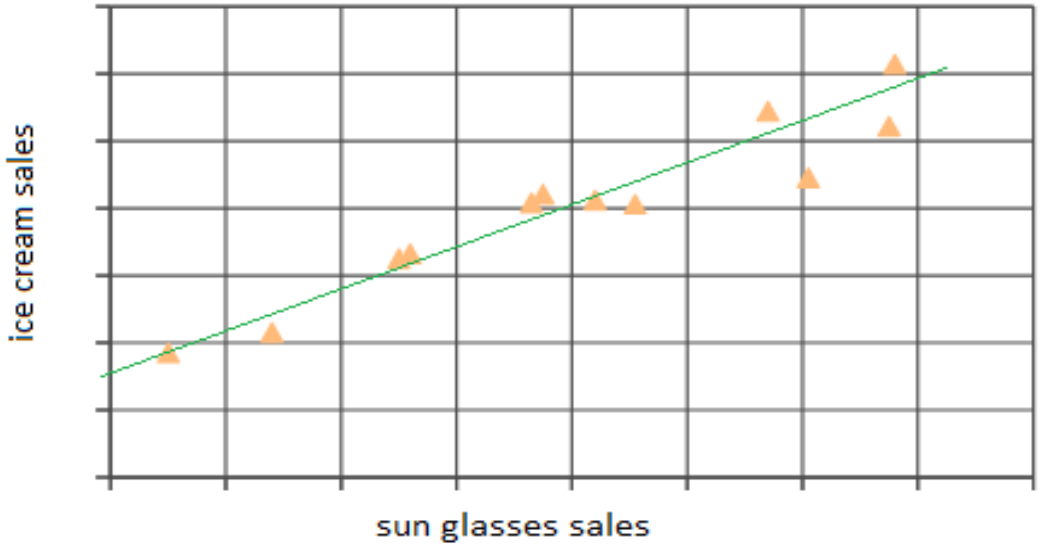


The line of best fit (line that passes majority of points) may be inserted to have a reference point to assess the level of the error among the plotted points

Correlation is not Causation.

If two variables are correlated does not mean that each of them causes the other. They could be a third factor that made each to correlate. There are lots of elaborate examples. If people carry umbrella because it is raining. That does not mean that umbrella caused rain. Investigate some other examples of Correlation is not Causation.

There is correlation between ice cream sales and sun glasses sold. But does that mean that sun glasses made people want to buy ice cream?



Calculating Correlations

Calculating Correlations: Two major types of correlation are calculated; these are **Pearson's** product correlation coefficient and **Spearman's** rank correlation coefficient. In this course we would concentrate on **Pearson's** product correlation coefficient as is the most popular, we will concentrate on **Pearson's**

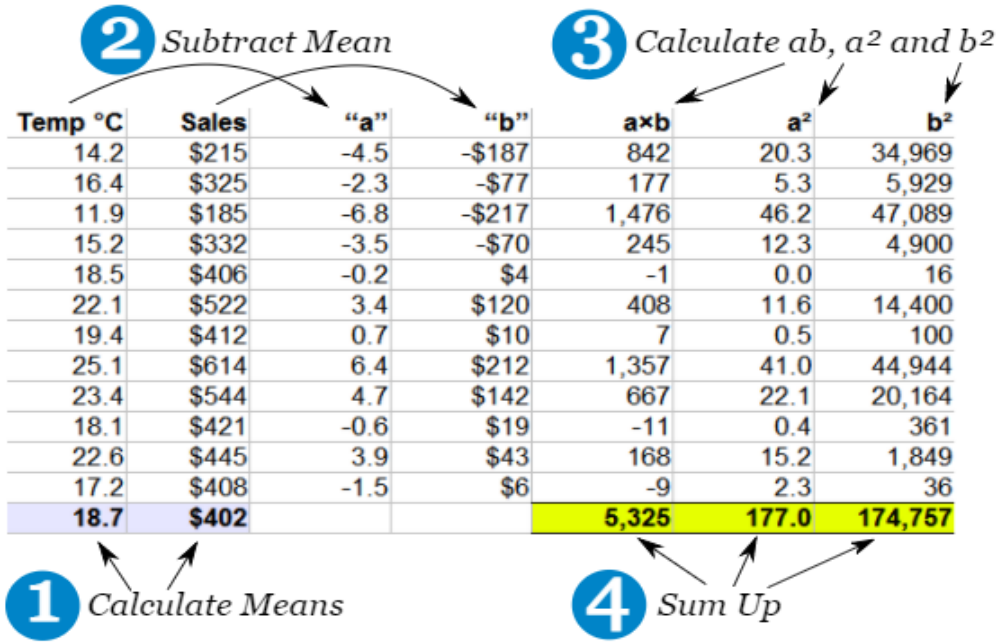
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Calculating Correlations.

Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

- Step 1: Find the mean of **x**, and the mean of **y**
- Step 2: Subtract the mean of x from every x value (call them "**a**"), and subtract the mean of y from every y value (call them "**b**")
- Step 3: Calculate: **ab**, $\mathbf{a^2}$ and $\mathbf{b^2}$ for every value
- Step 4: Sum up **ab**, sum up $\mathbf{a^2}$ and sum up $\mathbf{b^2}$
- Step 5: Divide the sum of ab by the square root of [(sum of a^2) × (sum of b^2)]

Here is how I calculated the first Ice Cream example (values rounded to 1 or 0 decimal places):



Temp °C	Sales	"a"	"b"	a×b	a²	b²
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

5 $\frac{5{,}325}{\sqrt{177.0 \times 174{,}757}} = 0.9575$

(https://www.mathsisfun.com/data/correlation.html)

importance of Correlation

Correlation is one of the techniques for attributes selections in machine learning. Most datasets are made up of many variables, these variables are correlated with each other and the target labels. Below is a section of wine quality dataset. With eleven variables (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, density, pH, sulphates, alcohol), these variables (also called attributes or features), have relationship with one another and also has relationship with the quality of wine. Therefore, they are used to predict the quality of a wine which are measured in a scale of 1 to 10 (only 5 and 6 shown here)

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

importance of Correlation

In all the eleven variables some are more relevant than others in predicting the quality of wine, some may not even be necessary at all rather may add noise into the process of determining the quality of wine. The extents of this relevance determine the significance of the variable in predicting the quality of wine.

End.