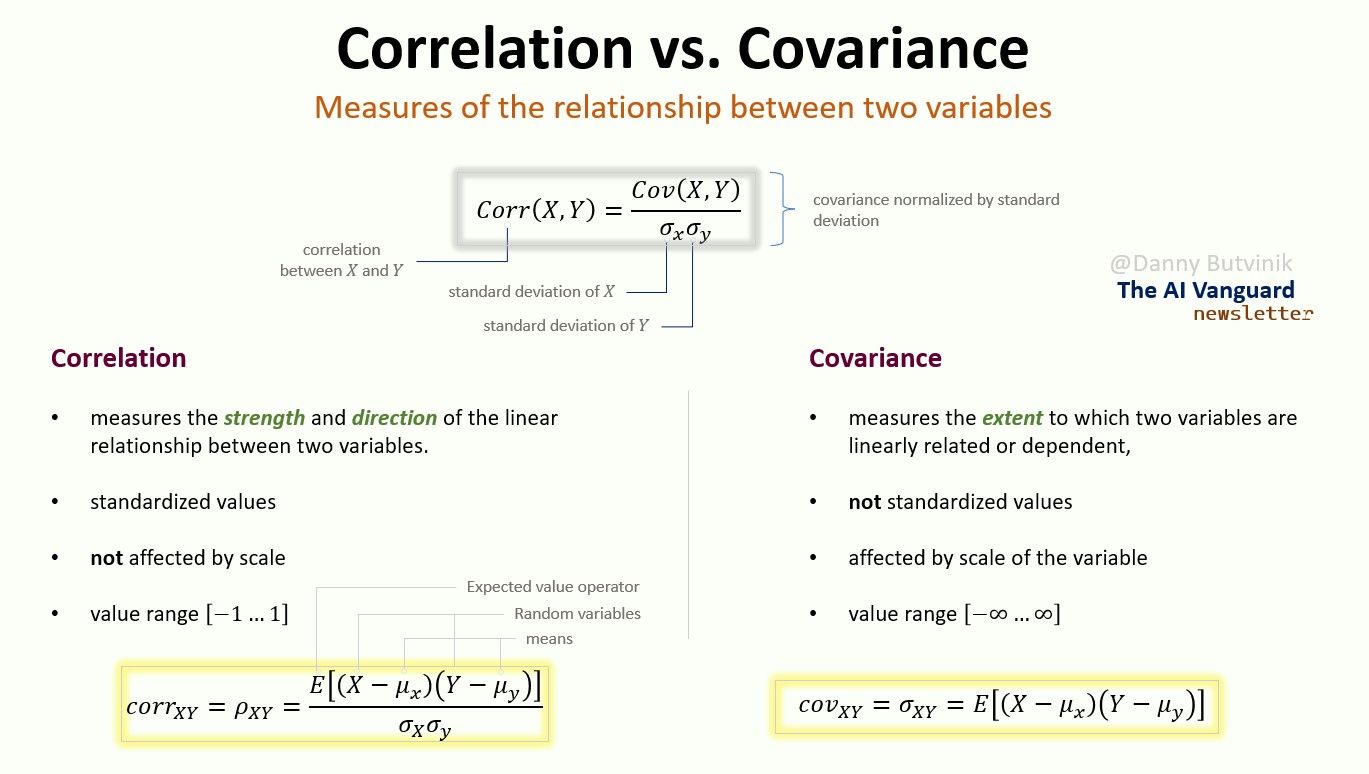
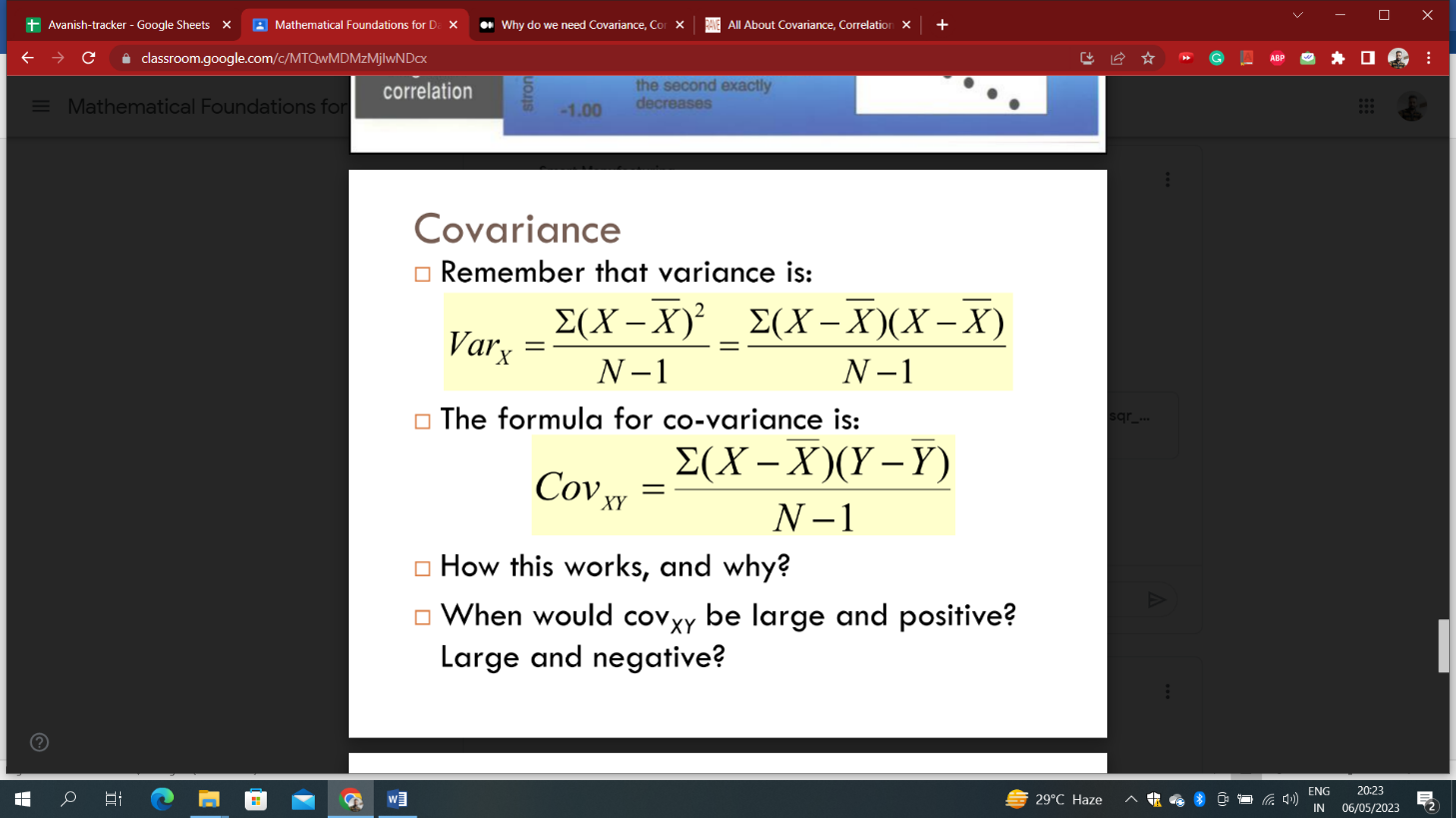
**Correlation, causation and Covariance**

Correlation and covariance are essential tools in various fields, such as statistics, data science, machine learning, and data analysis. They serve as useful measures for determining the relationship between two variables.  
These concepts are particularly significant in artificial intelligence and machine learning, as they are frequently employed in linear regression and neural networks to model and predict the relationship between variables.  
   
However, they have different properties and may be used in different contexts depending on the research question and the data being analyzed.  
   
Understanding the relationship between two variables is a critical aspect of data analysis. Two commonly used measures of relationships are correlation and covariance.  
While both provide useful insights into the relationship between two variables, they have distinct properties and may be used in different contexts.

Credit: The AI Vanguard Newsletter | Danny Butvinik



Note:

1. **Pearson Correlation Coefficient is represented by ρ** for a population and by **r** for a sample.
2. The expected value(E) is the arithmetic mean of a large number of independently selected outcomes of a random variable.

**Types of Correlation:**

**1. Positive correlation**

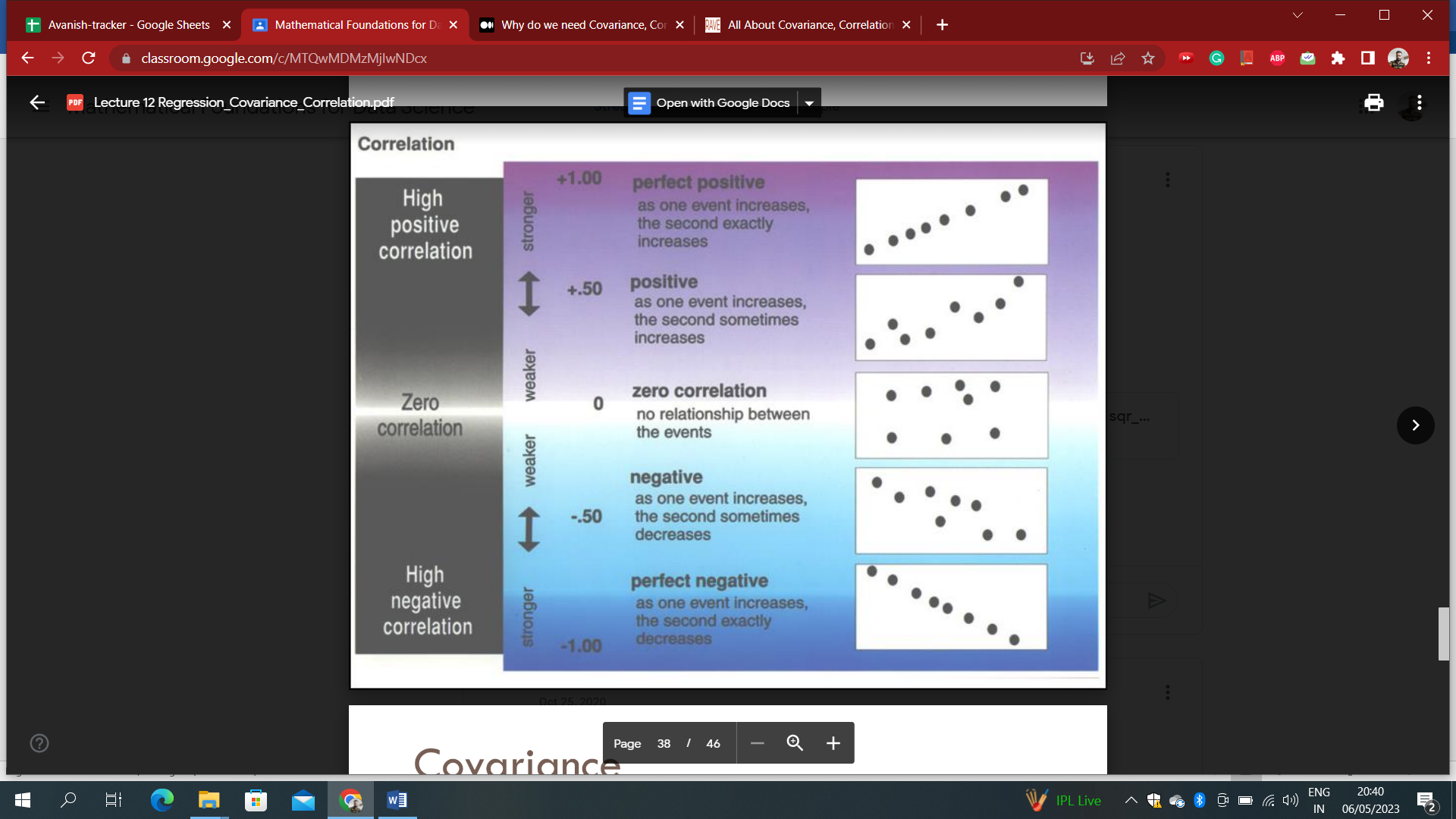
* **High values of X tend to be associated with high values of Y.**
* **As X increases, Y increases**

**2. Negative correlation**

* **High values of X tend to be associated with low values of Y.**
* **As X increases, Y decreases**

**3. No correlation**

* **No consistent tendency for values on Y to increase or decrease as X increases**

****

**Calculation in pandas:**

import pandas as pd  
import numpy as np  
# Setting a seed so the example is reproducible   
np.random.seed(123)  
df = pd.DataFrame(np.random.randint(low= 0, high= 20, size= (5, 2)), columns= ['X', 'Y'])  
df[['X', 'Y']].cov()

In the code below, we have determined the correlation between literacy rate and sex ratio from data of 640  Indian districts.

from scipy.stats.stats import pearsonr

pearsonr(data['literacy\_percent-dist'],data['sex\_ratio'])[0]

**Note:**

1. Correlation and covariance are sensitive to outliers, so checking for outliers is important before calculating these measures.  
2. While correlation measures the linear relationship between two variables, it may not capture the full extent of the relationship if it is not linear. In cases where the relationship is not linear, other measures, such as nonparametric correlation coefficients or nonlinear regression, may be more appropriate.  
3. Sometimes, a high correlation coefficient may not necessarily imply causality between the two variables. The correlation only measures the association between two variables, and other factors may affect the relationship between the two variables.

4. With small samples, correlations can be unreliable. The smaller the sample size, the more likely we are to observe a correlation that is further from 0, even if the true correlation (obtained if we had data for the entire population) was 0. The standard way of quantifying this is to use **p-values**. **In academic research, a common rule of thumb is that when p is greater than 0.05, the correlation should not be trusted.**  
5. Covariance and correlation can be calculated using different methods, such as raw data, deviations from the mean, or data ranks. The choice of method can affect the resulting correlation or covariance coefficient.  
   
**Example**  
Suppose we have two variables, X and Y, and we want to measure their relationship. We calculate the covariance and correlation coefficients and obtain the following results:  
Covariance: 500  
Correlation: 0.8  
At first glance, X and Y appear to have a strong, positive relationship. However, upon further inspection, we find one outlier in the data driving the results. After removing the outlier, we recalculate the covariance and correlation coefficients and obtain the following results:  
Covariance: 200  
Correlation: 0.6  
We can see that the correlation coefficient decreased, indicating a weaker relationship between X and Y, and the covariance decreased even more. This illustrates the importance of checking for outliers and the potential for spurious correlations when working with real-world data.  
It's crucial to carefully examine the data and ensure that no outliers or other elements that might skew the results are responsible for the findings.

**Causation:**

**It** indicates that one event results from the occurrence of the other event; i.e., there is a causal relationship between the two events. This is also referred to as cause and effect.

For Example: After I exercise, I feel physically exhausted. This is cause-and-effect because I’m purposefully pushing my body to physical exhaustion when doing exercise. The muscles I used to exercise are exhausted (effect) after I exercise (cause). This cause-and-effect is confirmed.

**All causations are correlations, but not all correlations are causations.**

Examples of correlation, NOT causation: “On days where I go running, I notice more cars on the road. “ I, personally, am not CAUSING more cars to drive outside on the road when I go running. It’s just that because I go running outside, I see more cars than when I stay at home. This relationship is not cause-and-effect because neither the cars nor I are impacting each other.

**Conclusion:**

While covariance identifies how two variables vary simultaneously, correlation determines how change in one variable affects the change in another variable.

The two variables are correlated with each other and there is also a causal link between them. A correlation doesn't imply causation, but **causation always implies correlation**.