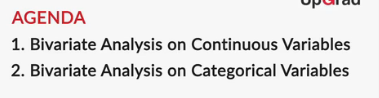
**Bivariate Analysis on Continuous Variables**



Continuous Variables are those we can add up.

Categorical Variables are those we can group by.

To summarise, **correlation** is a number between **-1 and 1** which quantifies the extent to which two variables ‘correlate’ with each other.

* If **one increases** as the **other increases**, the correlation is **positive**
* If **one decreases** as the **other increases**, the correlation is **negative**
* If one **stays constant** as the **other varies**, the correlation is **zero**

In general, a positive correlation means that two variables will increase together and decrease together, e.g. an increase in rain is accompanied by an increase in humidity. A negative correlation means that if one variable increases the other decreases, e.g. in some cases, as the price of a commodity decreases its demand increases.

A perfect positive correlation means that the correlation coefficient is exactly 1. This implies that as one variable moves, either up or down, the other one moves in the same direction. A perfect negative correlation means that two variables move in opposite directions, while a zero correlation implies no relationship at all.

There are **two things** which you would have noticed from the lecture:

1. The **correlation matrix of stock prices** of different countries gives a real sense of the relationship between many variables.
2. The correlated variables are **grouped by similarities**, and correlation can also be calculated for ‘groups of variables. This is called **‘clustering’** which you will study in detail later in the machine learning course, but the idea is to form a hierarchy of similar groups of variables.

If you are confused about the correlation matrix used above and its interpretation, you can visit this [link](https://gramener.com/retail/clothing-sales), where a similar case study has been explained. Basically, the top-right half represents the correlation coefficient and the left bottom half has the scatter plot between the 2 variables.

# Bivariate Analysis on categorical variables

The categorical bivariate analysis is essentially an extension of the **segmented univariate** analysis to another categorical variable. In **segmented** **univariate analysis**, you compare metrics such as ‘**mean of X**’ across various segments of a categorical variable, e.g. mean marks of a student are higher for ‘degree and above’ than other levels of the mother’s education; or the median income of educated parents is higher than that of uneducated ones, etc.

In the **categorical bivariate analysis**, you extend this comparison to other categorical variables and ask — is this true for all categories of another variable, say, men and women? Take another categorical variable, such as state, and ask — is the median income of educated parents higher than that of uneducated ones **in all states**?

Thus, you are drilling down into another categorical variable and getting closer to the true patterns in the data. In fact, you may also go to the next level and ask — is the median income of educated parents higher than that of uneducated ones (variable 1)  in all states (variable 2) for all age groups (variable 3)? This is what you may call ‘trivariate analysis’, and though it gives you a more granular version of the truth, it gets a bit complex to make sense of and explain to others (and hence it is not usually done in EDA).  
  
Thus, remember that doing only conducting segmented univariate analysis may deceive you into thinking that a certain phenomenon is true without asking the question — is it true for all sub-populations or is it true only when you aggregate information across the entire population?

So in general, there are two fundamental aspects of analysing categorical variables:

1. To see the distribution of two categorical variables. For example, if you want to compare the number of boys and girls who play games, you can make a ‘cross table’ as given below:

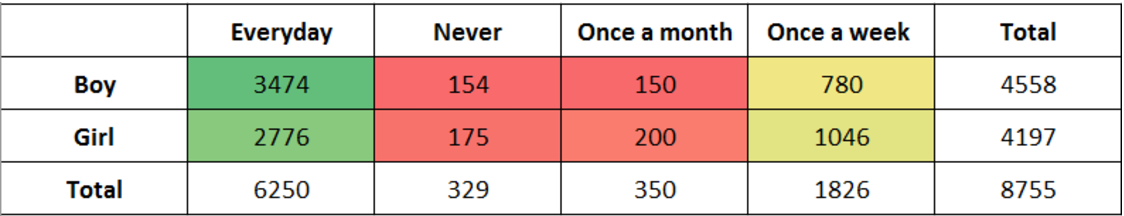


Fig-1: Cross Table

From this table, firstly, you can compare boys and girls across a fixed level of ‘play games’, e.g. a higher number of boys play games every day than girls, a higher number of girls never play games than boys, etc. And secondly, you can compare the levels of ‘play games’ across a fixed value of gender, e.g. most boys play every day and very few play once a month or never.

1. To see the **distribution of two categorical variables** with one continuous variable. For example, you saw how a student’s percentage in science is distributed based on the father’s occupation (categorical variable 1) and the poverty level (categorical variable 2).

Usually, you do not analyze more than two variables at a time, though there are ways to do that. Machine learning models are essentially a way to do that, some of which you will learn in the next course.