BLOG DATA SCIENCE

Check the first module to find out about data cleaning( the step to clean data is more direct and I should focus on the practical learning for less irrelevant details

**Mission 3.1: First, some theory**

A key part of the exploratory analysis is using Descriptive Statistics.

Descriptive statistics

**Descriptive statistics** are a set of techniques used to summarize and describe the main features of a dataset.

The three main types of descriptive statistics are:

**Frequency Distribution**: A table or graph that illustrates the number of occurrences of each value or group of values in a dataset.

**Measures of Central Tendency**: Estimates of the middle or average values in the dataset. These include the mean, median, and mode.

**Measures of Variability**: Describes how spread out the data is. Range, standard deviation, and variance are all measures of variability.

**Pivot Tables**

Pivot tables are a data tool to summarize and organize large amounts of data in a meaningful way. It allows you to quickly look up the information you need.

We’ll be looking at answering the first five questions from Mission 1.3 Defining the problem.

**Questions from Mission 1.3**

1. What are the total numbers of subscribers in each subject?
2. How does the average content duration/price/number of students vary across different subjects?
3. How many courses are free and paid for each subject?
4. What is the average price of web development courses at different levels?
5. What are the 20 most popular courses? Also, include the following information:  
     
   *- Their level  
   - Whether they are free or paid  
   - Whether any are free beginner courses  
   - Duration of the course*

**Our analysis will consist of three parts.**

In this mission we will:

* Calculate some of the descriptive statistics such as the mean, median, minimum, and maximum values for content duration/price of the course/and number of subscribers.

And in the next mission, we will:

* Create pivot tables to answer questions 1 to 4.
* Learn to use the IF function, VLOOKUP and Sorting to answer question 5.

**Let’s get started**

**Q: Calculate descriptive statistics for “Content Duration”, “Price”, and “Subscribers”**

For these variables, we’ll calculate the mean and median values (measures of central tendency), as well as the minimum and maximum values.

We’ll be using the following formulas:

=AVERAGE

=MEDIAN

=MIN

=MAX

PROCEDURE WITH GOOGLE SHEET

we noticed that the minimum value was zero for the content duration and subscribers and we deleted those entries as they were outliers and can give us misleading results.

Generally, an analyst would decide what values to keep and remove. For example, it could be that they will perform analysis on all courses with more than 10 subscribers. In that case, we would need to delete the lower value.

Now, looking at our calculations, try answering the following questions:

1. What are the total numbers of subscribers in each subject?
2. How does the average content duration/price/number of students vary across different subjects?
3. How many courses are free and paid for each subject?
4. What is the average cost of web development courses at different levels?



We can’t, right? Our insights have given us a good starting point, but they are not enough to answer the questions we asked at the beginning of the course.

We have established the basics; now let's use another powerful tool to answer those key questions: **the pivot table.**