**Problem Statement I - Data Preparation**

**Business problem**

Global Mart is an online supergiant store that has worldwide operations. This store takes orders and delivers across the globe and deals with all the major product categories — consumer, corporate and home office.

As a sales manager for this store, you have to forecast the sales of the products for the next 6 months, so that you have a proper estimate and can plan your inventory and business processes accordingly.

The store dataset has the following 5 attributes and their data description is as given below:

| **Attributes** | **Description** |
| --- | --- |
| Order-Date | The date on which the order was placed |
| Segment | The segment to which the product belongs |
| Market | The market to which the customer belongs |
| Sales | Total sales value of the transaction |
| Profit | Profit made on the transaction |

If you check the entries in the dataset, you will see that the store caters to 7 different geographical market segments and 3 major customer segments, i.e. consumer, corporate and home as can be seen in the table below.

| **Market** | **Segment** |
| --- | --- |
| Africa | Consumer |
| APAC (Asia Pacific) | Corporate |
| Canada | Home Office |
| EMEA(Middle East) |  |
| EU (European Union) |  |
| LATAM (Latin America) |  |
| US (United States) |  |

Based on these, there are **21 unique "Market-Segments"** for which the sales forecasts can be made. That is the **dataset** needs to be **prepared** such that you get the **Order-Date**, **Sales** and **Profit** for the **21 market segments**.

Now, due to certain unpredictable circumstances in the market, as a company, you are prioritizing only the best and most consistent market segment in terms of profitability. You want to see which market segment is the most consistently profitable. And then, you want to **forecast the sales** for that **most consistently profitable market-segment only.** This way you know that the market region your company is investing in will be beneficial for the company as the forecasts will be reliable. As of now, you do not want to focus on other market segments that might have not been very consistent and profitable to your company.

**So, not all of these 21 market segments are important from the store’s point of view. You need to find out the most consistently profitable market-segment from the above and forecast the sales and demand for that single market-segment only and not for all.**

Now the question is how do you find that most profitable market segment from the 21 market segments ?

**How to find the most profitable market segment ?**

By now, it’s clear that you only need to work on one market-segment which is the most consistently profitable. To find the most consistently profitable market-segment you will be using a measure called **"Coefficient of Variation (CoV)"**. The coefficient of variation or CoV is nothing but the ratio of the standard deviation to mean for the data that it is being calculated for.

But why consider the coefficient of variation here and not the standard deviation for measuring how much variation is present in the data for each of the 21 market segments? and how will the coefficient of variation lead you to the most profitable market segment for which eventually you will want to forecast the sales?

The **coefficient of variation is a ratio of the standard deviation to mean**. Once you have prepared the data such that you have the Order-Date, Sales and Profit against each of the 21 market segments, and not in the manner as it was in the initial dataset, you can check the standard deviation and the mean calculated on profit for all the 21 market segments and compare. You will find that these values vary a lot and hence it is meaningless to compare the 21 market segment's profits based on the standard deviation and their mean.

Actually, standard deviations are meaningless to compare different datasets as you would see for these 21 market-segments as well. As a better metric to compare the variance between the segments you use the **coefficient of variation** which will **normalise the standard deviation with the mean** and give you a comparative figure on the basis of which you can identify the most profitable market segment.

As a part of this assignment, you will find the coefficient of variations for all these 21 market segments and list them in a table and compare them.

Now, **it is for you to reason out** whether the **most profitable market segment** should have the **least value of CoV or the highest value of CoV**. However, please note that as a Sales manager you want to **forecast the sales** where the market segment is reliable or in other words, there is **less variation in the profits.**

You can read more about coefficient of variation [here](https://365datascience.com/coefficient-variation-variance-standard-deviation/).

Data understanding and preparation

Now let's begin the actual work. You have seen that the data had 5 attributes. These are Order-Date, Market, Segment, Sales, and Profit for each transaction. The complete data dictionary has been given below again for your reference. The “Market” attribute has 7-factor levels representing the geographical market sector that the customer belongs to. The “Segment” attribute tells which of the 3 segments that customer belongs to.

For doing the data preparation —

1. Find the 21 unique Market Segments by combining the respective 7 geographical markets for each of 3 segments such as Home office, Consumer and Corporate.
2. Your dataset after the above step should have Order-Date, Sales, Profit against each market segment such as APAC-Consumer, APAC-Home Office and so on.
3. Once you have understood the dataset of the global data store, you need to get the order date in the required month-year format to make it a monthly aggregated transaction data. For this, **convert the order-date into a date-time format** for getting it into the **Month-year**format; you will get the data for 48 months now.
4. After the above step, perform the **train-test split** such that you take the **42 months as the train data**and the **6 months as the test data**.
5. Calculate the **CoV on the profit** for each of the 21 market segments on the train data.
6. Find the most profitable market segment by **comparing the 21 CoV values**.

Once you have understood the dataset of the global data store, you need to get the order date in the required month-year format to make it a monthly aggregated transaction data. Then, you need to concatenate the market and segment columns such that you get the time series data consisting of order, sales, profit, market segment. The perform the train-test split for the such that you take the 42 months as train data and the 6 months as the test data.

For each of the 21 unique market segments, then calculate the coefficient of variation on the respective profits and compare the CoVs. Conclude the most profitable market segment based on the CoV value. It is expected that you do some research on understanding CoV and how to find it out for all the given market segments in Python.

**NOTE**: Till now, you have done only the data preparation part, in the next segment, you will understand the other part of the problem statement, that is to forecast the SALES for the most profitable market segment.