Case Study 3 – Customer Behavior (Patterns and Trends)

The aim of this Case Study is to search and find patterns and trends about how a customer behaves. This helps is gaining a clear knowledge on what the customer likes, and dislikes and we can come up with better business decisions.

Using Data Driven The dataset **OnlineRetail.xlsx** was obtained from Kaggle and it consists of 8 attributes and 541909 observations. I used **R** Programming Language and Packages in R (*tidyr, dplyr, ggplot2, maps, tidyverse*) for processing the data and plotting the graphs.

Modifying the data types and changing the data type into factors was the first step for processing the data. It was an efficient way to plot graphs with factors as your datatypes in R.

Instead of working on the whole data at once (which made the graphs unreadable), I decided to split the data and take the data of specific countries as examples for plotting. Initially, it worked well for a few countries (such as USA, Japan, Canada) but countries with a lot of entries and multiple customers and invoices (such as Germany, France), graphs were not as good as expected (mostly unreadable/ packed with lot of points) and this made the data unusable. So, while plotting, I didn't include these countries and worked on the smaller chunks of the data to plot the graphs.

Including the World Map was not the idea I had in the beginning, but it felt challenging and interesting to work on the maps. I even thought of using Tableau instead of maps package and change the technology used, but it was a good learning experience for me work on maps package and include the World Map in my code. World Map can be improved a little more by adding more indices and elements in the graph, it was new for me and I'm trying to find ways to improve it

There were two major things that I thought of working but was unable to do so.

- Firstly, I was unable to find the correlation among the attributes as the attributes were limited and it was not efficient to work on it.
- Secondly, even if I have a lot of data (541909 obv.) but I only used 1500-2000 observations to plot graphs as I was working country wise.

It was a setback that I had 2 more attributes – Description and StockCode which I was unable to use while plotting. With further improvements in the generated graphs, we can use *Description* and we can get the product details (that would be after the plotting, just to know the product which has customer's interest). There would be a way to generate a graph such a way that when we hover cursor at the point, it would give the Description of the product and the attribute would be used. This is one way of using the attribute, but I couldn't come up with a way to use *StockCode*. Maybe I could have used it which plotting the graphs (like *InvoiceNo*) but that would not help us in drawing any conclusions.

I wanted to draw out the whole data and use it as effectively as possible, but with the restrictions like available attributes type, missing values and unused data, the finding and the conclusions drawn are limited. This may lead a decision which would not be helpful for business.