**Explanation and Theory Part**

1. **Data Understanding**

In the jupyter notebook named Data Understand.ipynb I have imported the dataframe in the sqlite database. I have used functions which accept 3 arguments viz. Connection, table name and the column name to query out the required columns from the dataframe. All the steps of preprocessing techniques mentioned in the Task 1 of the Data Understanding have been implemented in the above highlighted .ipynb notebook.

1. **Perform RMF Segmentation**

In the jupyter notebook named rfm-analysis-online-retail.ipynb. The notebook is self exploratory since the code block present in the .ipynb notebook has been explained in depth. Moreover I have segmented the customer into 10 categories which is :

1. Hibernating
2. At Risk
3. Can’t Lose
4. About to Sleep
5. Need Attention
6. Loyal Customer
7. Promising
8. New Customer
9. Potential Loyalist
10. Champion

The customer segmentation has been shown with the help of the tree map.

1. **Customer Segmentation with k-means**

In the jupyter notebook named K--Means.ipynb. The notebook is self exploratory since the code block present in the .ipynb notebook has been explained in depth. The dataset consists of approximately 59k rows and 8 columns from the UK-based and registered non-store online retail. The company sells all occasion gifts. The dataset was calculated between 01/12/2010 and 09/12/2011.

Before applying K-means algorithm we applied various preprocessing techniques which are as follows:

1. Null value treatment.
2. Outlier Detection (IQR Strategy to remove outliers).
3. Standard Scaling of the dataset.
4. Analyzing the customer based on:

4.1. Recency

4.2. Frequency

4.3. Monetary

1. Analyzing the Correlation between dependent variables
2. Pair plot of Recency, Frequency, Monetary
3. RMF Score generation

**3.1. Justification of K-value**

After preprocessing the dataset, we applied K-means clustering. In K-mean

clustering we need to decide the number of the cluster manually. So, I tried with the number of clusters inthe range of [2,8] inclusive to generate the loss and applied silhouette Analysis to find the optimum number of clusters for the dataset.

**Silhouette Analysis:**

silhouette score=(p-q)/max(p,q)

Where,

p= mean distance to the points in the nearest cluster that the data point is not a part.

q= mean intra-cluster distance to all the points in its own cluster.

The value of the silhouette lies in the range of [-1,1] inclusively. A score closer to 1 indicates that the data point is very similar to other data points in the cluster and a score closer to -1 indicates that the data point is not similar to the data points in its cluster. By experimenting with the number of the cluster while training the model, we found that k-means model loss stops decreasing after the model with 3 clusters i.e it displayed the elbow pattern shape in the graph with 3 clusters. The score is shown below for reference:

For n\_clusters=2, the silhouette score is 0.5415858652525395

For n\_clusters=3, the silhouette score is 0.5084896296141937

For n\_clusters=4, the silhouette score is 0.4814862659467589

For n\_clusters=5, the silhouette score is 0.46627005641897035

For n\_clusters=6, the silhouette score is 0.41723982911099605

For n\_clusters=7, the silhouette score is 0.41705890301596343

For n\_clusters=8, the silhouette score is 0.39439948764516547

1. **Review of the result**

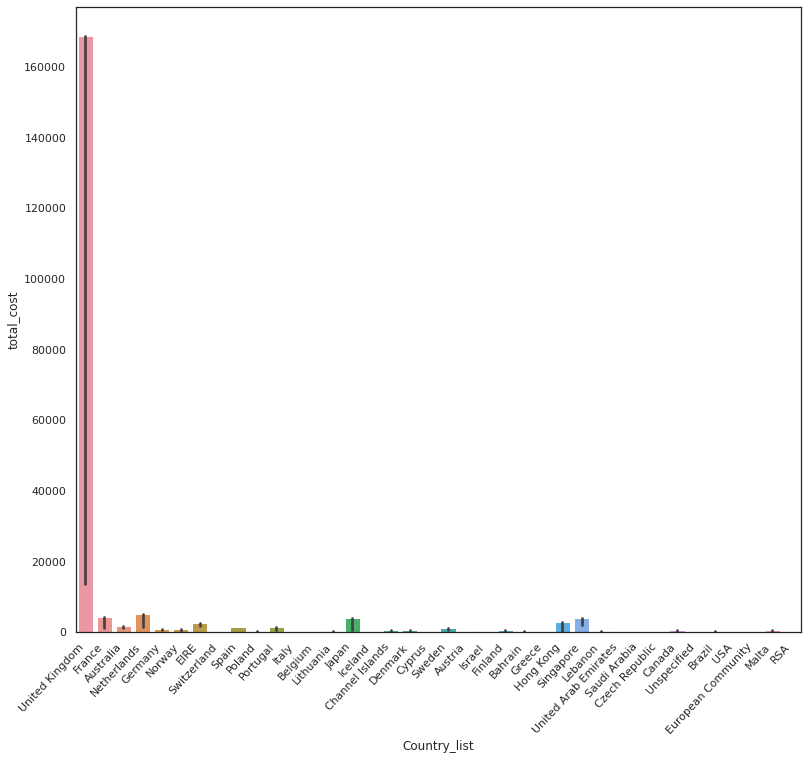
**4.1. Identification of the business value**

The below images depict the items customers bought from the store. The size of the text is directly proportional to the number of the items bought from the customer. As we can see items such as metal object, water bottles lunch bag, object of the red colors and many more are most buyed item by the customers from the shop whereas some items such as vintage leaf, wall arts, girl design, art making and many other items which are sold less frequently as compared to the other items.

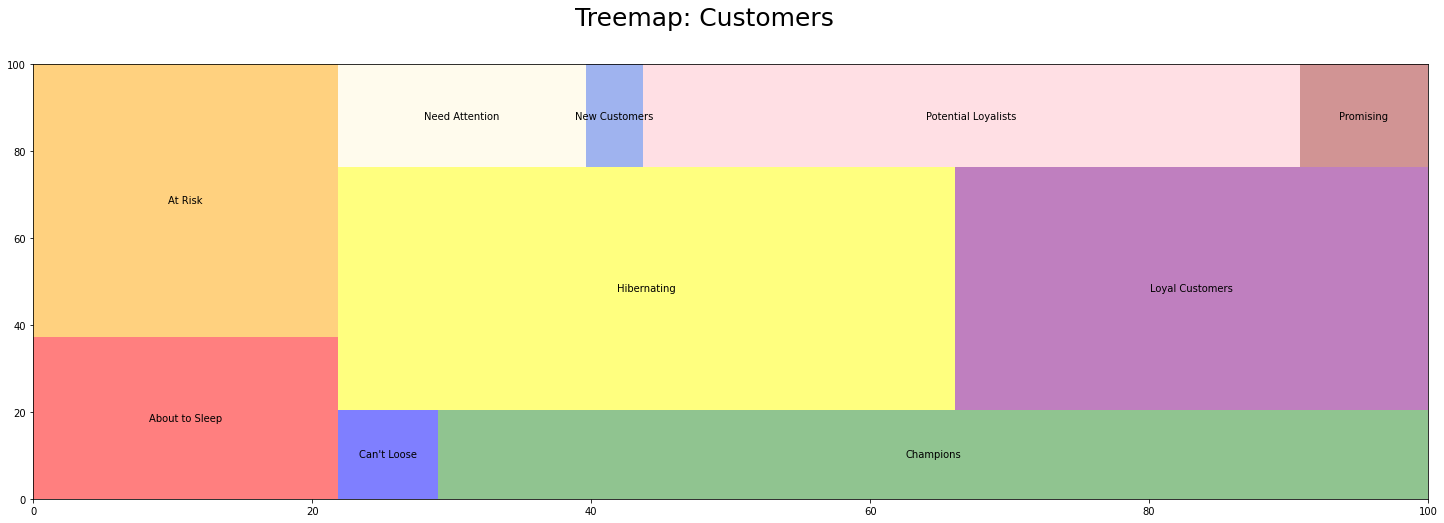


**Fig 1. Product sold**

Since the retail shop is present in 40 countries where the UK secures the top most place by generating more than 60% of the total revenue whereas in other countries the revenue is very negligible. Since the retail shop is based in UK so it has idea of the customer requirements especially for the people in the UK but we can see from the below graph that sales from the other counties is almost negligible as compared to the UK so the retail shop need to focus on the customer need by considering cultural value, geological as well as topographical needs as well as the online market price as to attract the customers.



The below image shows the different segments of the customers from the online sales dataset which are segmented based on the recency, frequency and monetary.



As we can see from the above graph the number of customers who are loyal and the number of customers who are hibernating is almost equal so we need to transition the hibernating customer into revenue generating customer by offering attractive schemes or some other means so they start generating revenue for the retail stores. One beneficial thing for the shop is that less than 30% of the customers are at risk and about to sleep. If we deal with the customers who are at risk and about to sleep we need to find the exact data of their needs, choice and their expectation and why they are not buying from the retail store and buying from some other shop and implement the necessary steps to convert them into other segments. Now, we have almost an equal number of customers who are promising and can't lose so we need to transition them by giving special offers into loyal customer segments or champion segments. If we analyze the tree map. We are attracting approximately 10% of the new customers. We need to focus on the transitioning of customers into champions as well as loyal customers as well as we need to attract new customers and transition them into loyal customers so we can expand the business of the retail store as well to generate revenue globally.

1. **Data Mart Design**

Based on the findings after analyzing the data, RMF model and K-means the main metric for designing the data mart with justification is mentioned below :

**Customer Needs/Requirements:** Since the retail store is spread over 47+countries, there is a need to analyze the customer requirements based on the geography, culture and topography of the location. We have seen that the UK has the maximum number of the revenue because the company is based in the UK . It understands the requirements of the local peoples and hence it is directly reflected in the data whereas other countries show very less revenue.

**Recent Trends:** Nowadays social media has a high impact on individuals' lives, so people try to follow what they see. We need to design the inventory in such a way that it is a combination of the recent trends as well as the people's needs.

**Geographical/Topological region:** People's needs/requirements are directly related to their culture and culture is related to the geographical location. SInce the was originally based in the UK it is performing well but if we will see other countries revenue we can easily analyze that they are lagging behind because the inventory is designed based on the UK region and for other countries it is irrelevant.

**Inventory/Product list:**  As we have seen in Fig.1 that some of the products demand is huge whereas some of the product demand is negligible. So we need to redesign the inventory in many countries in order to increase the revenue. The inventory should be based on the geographical location of the shop, recent trends and customer requirements.

**Marketing:** Marketing is the most important part of the business. If the online retail shop has reached its customers virtually it can attract a huge number of customers. Along with the campaign based on social media and internet we need to attract customers by offering them schemes based on their cluster or segmentation so that they must be transitioned into loyal customers or champions.