Problem Statement

Customer Segmentation: Explore and identify different segments present in the customer transaction data.

Dataset : The dataset contains transactions on an e-commerce website between the period **Feb 2018** to **Feb 2019** from customers across different countries.

Solution:

I use **K - means** clustering algorithms for user segmentation. For implementation I use **Scikit-Learn** python library. **NLTK** (Stamming, Tokenization) python library is used for text processing.

All steps of Segmentation are as follow:

- Firstly, I do some data cleaning like NULL value and duplicate entry removal. After it I get proper **401569** entries because above 50% of entries are duplicate and others are NULL with user id.
- Then do basic cleaning like time type transformation from object to datetime.

At the end of basic cleaning, I analyze all transactions that are in between **2018-02-12** and **2019-02-20**.

- Removal of negative quantity and check it with hypothesis testing.
- Some variable transformation and new variable generation.

Total numbers of unique products, transactions and customers.

	Products	Transactions	Customers
Quantity	3200	18632	4339

- I show basic **data analysis (daily, monthly, country wise)** charts and insights in Notebook.
- **Clustering on description** of each items and convert it into **5 category**
  (Both Transaction wise and user wise)
- For that I use only simple **one hot encoding**.
- I **ignore the country** variable because above **96%** of transactions are from the **United Kingdom**. So, they are not useful for clustering.
- Create new features from that clustering process and apply **standardization** on final dataframe
- For clustering validation, I use **silhouette score** and choose the number of clusters according to it.
- **Segment all user ids** into **10 cluster** and number of users in each cluster as follow:

	0	1	2	3	4	5	6	7	8	9
Customers	310	1460	1	1250	2	8	739	235	62	272

- At the end I do **Principal Component Analysis** for visualization of clusters.