1. **Cleaning Data:**
    a. Userids and Itemcodes with -1 were removed. These on inspection, seemed to correspond to banking and shipping charges, etc.
    b. There were -ve values for NumberOfProductsPurchased column. This was assumed to correspond to returned orders. There were duplicate entries for these transactions.
    c. There were zeros for CostPerItem column. This was assumed to correspond to cancelled orders. There were 2 copies of each of these transactions.
    d. Some transaction dates had 2028 recorded. These were replaced with 2018.
    e. ‘Country’ of some users were ‘Unspecified’. These were replace with United Kingdom, from where more than 89% of transactions happened. Some users had made transactions from multiple countries. The country of maximum purchases per user was assigned to them.
    f. Some itemcodes had different descriptions. These descriptions are closely related - different sizes, colors, etc of the same basic product. The first occurring description is kept for all items.

2. **Feature Engineering (per user)**
    a. Average transaction hour (in 24 hour format).
    b. Avg, minimum, maximum amounts spent on valid (not returned/cancelled) transactions.
    c. Average items returned and total number of orders cancelled.
    d. Continent of (maximum) purchases (one hot encoding).
    e. Clusters of purchase history:
        i. Purchases were represented as the number of days from first purchase (per user).
        ii. Clusters are separated by 15 (a threshold number of) days of inactivity.
        iii. Features: number of clusters, avg. number of (unique) days of purchase across clusters, and total number of unique days of purchase.
    f. Days since last purchase (current day is chosen as latest date in dataset).
    g. Amount spent on different product categories:
        i. Words are filtered for English stopwords, colors, sizes, specific punctuations, etc.
        ii. The stem of each filtered word in the descriptions was extracted (nltk stemmers).
        iii. CountVectorizer (counts occurrences) is used to extract the top 120 words.
        iv. Each description is converted into a vector representation of 120 binary variables.
        v. KMeans clustering is performed and K = 9 was chosen based on silhouette score.
        vi. Total amount spent by each user on each product category is recorded.

3. **Feature selection (using KMeans)**
    a. Two features found to be poor performers - continent information and days since last purchase.
    b. Removing either of the two gives better overall silhouette scores, but the best score was obtained by removing both these features.
    c. Number of optimal clusters is found to be 2, by silhouette analysis.

4. **Future Work**
    a. Other algorithms like Gaussian Mixture Models or DBSCAN could be employed.