Online Retail Sales Analysis

DATA PREPROCESSING:

Data Wrangling :

Cleaned the data by removing NaN/ non-finite values and removed outliers using Standard Deviation methods.

RFM model:

RFM (recency, frequency, monetary) analysis is a marketing technique used to determine quantitatively which customers are the best ones by examining how recently a customer has purchased (recency), how often they purchase (frequency), and how much the customer spends (monetary). RFM analysis is based on the marketing axiom that "80% of your business comes from 20% of your customers."

Recency (R): Recency is the most important predictor of who is more likely to respond to an offer. Customers who have purchased recently are more likely to purchase again when compared to those who did not purchase recently.

Frequency (F): The second most important factor is how frequently these customers purchase. The higher the frequency, the higher is the chances of these responding to the offers.

Monetary Value (M): The third factor is the amount of money these customers have spent on purchases. Customers who have spent higher contribute more value to the business as compared to those who have spent less.

CLUSTERING:

k-means algorithm is used to cluster the sample data. k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori.

Let $X = \{x1,x2,x3,\ldots\ldots..,x_n\}$ be the set of data points and $V = \{v1,v2,\ldots\ldots,v_c\}$ be the set of centers.

Algorithm:

1) Randomly select 'c' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Recalculate the new cluster center.
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3.

We fetch the elbow point and perform clustering using 3 attributes, monetary ('Amount'), recency value ('rec_val') and frequency ('freq_val'). We perform 6 clusters.