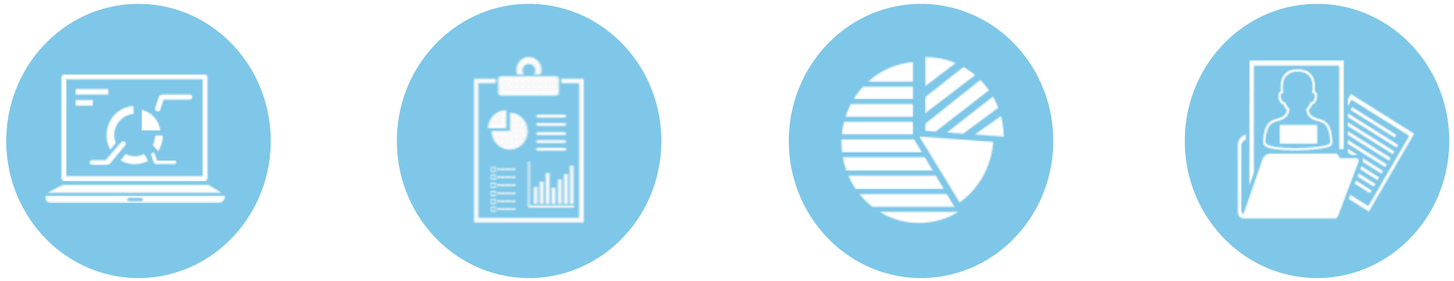
Data Science Capstone Project

Project 3: Retail



Problem Statement

It is a critical requirement for business to understand the value derived from a customer. RFM is a method used for analyzing customer value.

Customer segmentation is the practice of segregating the customer base into groups of individuals based on some common characteristics such as age, gender, interests, and spending habits

Perform customer segmentation using RFM analysis. The resulting segments can be ordered from most valuable (highest recency, frequency, and value) to least valuable (lowest recency, frequency, and value).

Dataset Description

This is a transnational data set which contains all the transactions that occurred between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique and all-occasion gifts.

Variables	Description
InvoiceNo	Invoice number. Nominal, a six digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation
StockCode	Product (item) code. Nominal, a five digit integral number uniquely assigned to each distinct product
Description	Product (item) name. Nominal
Quantity	The quantities of each product (item) per transaction. Numeric
InvoiceDate	Invoice Date and time. Numeric, the day and time when each transaction was generated
UnitPrice	Unit price. Numeric, product price per unit in sterling
CustomerID	Customer number. Nominal, a six digit integral number uniquely assigned to each customer
Country	Country name. Nominal, the name of the country where each customer resides

Project Task: Week 1

Data Cleaning:

1. Perform a preliminary data inspection and data cleaning.
   a. Check for missing data and formulate an apt strategy to treat them.
   b. Remove duplicate data records.
   c. Perform descriptive analytics on the given data.

Project Task: Week 1

Data Transformation:

2. Perform cohort analysis (a cohort is a group of subjects that share a defining characteristic). Observe how a cohort behaves across time and compare it to other cohorts.
   a. Create month cohorts and analyze active customers for each cohort.
   b. Analyze the retention rate of customers.

Project Task: Week 2

Data Modeling :

1. Build a RFM (Recency Frequency Monetary) model. *Recency* means the number of days since a customer made the last purchase. *Frequency* is the number of purchase in a given period. It could be 3 months, 6 months or 1 year. *Monetary* is the total amount of money a customer spent in that given period. Therefore, big spenders will be differentiated among other customers such as MVP (Minimum Viable Product) or VIP.

Project Task: Week 2

Data Modeling :

2. Calculate RFM metrics.
3. Build RFM Segments. Give recency, frequency, and monetary scores individually by dividing them into quartiles.
    b1. Combine three ratings to get a RFM segment (as strings).
    b2. Get the RFM score by adding up the three ratings.
    b3. Analyze the RFM segments by summarizing them and comment on the findings.

Note: Rate “recency" for customer who has been active more recently higher than the less recent customer, because each company wants its customers to be recent.

Note: Rate “frequency" and “monetary" higher, because the company wants the customer to visit more often and spend more money.

Project Task: Week 3

Data Modeling :

1. Create clusters using k-means clustering algorithm.
   a. Prepare the data for the algorithm. If the data is asymmetrically distributed, manage the skewness with appropriate transformation. Standardize the data.
   b. Decide the optimum number of clusters to be formed.
   c. Analyze these clusters and comment on the results.

Project Task: Week 4

Data Reporting:

1. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:
   a. Country-wise analysis to demonstrate average spend. Use a bar chart to show the monthly figures
   b. Bar graph of top 15 products which are mostly ordered by the users to show the number of products sold
   c. Bar graph to show the count of orders vs. hours throughout the day
   d. Plot the distribution of RFM values using histogram and frequency charts
   e. Plot error (cost) vs. number of clusters selected
   f. Visualize to compare the RFM values of the clusters using heatmap

Thank You