SCTP Capstone Project

# **Data Understanding and Solution Identification**:

## Data Understanding:

1. InvoiceNo – Invoice number
2. StockCode – Product Identifier
3. Description – Product Description
4. Quantity – Transactional Quantity
5. InvoiceDate – Date and Time of purchase
6. UnitPrice – Per unit price for the item
7. CustomerID – Customer Identifier
8. Country – Country of Customer

## Solution Identification/Business Intelligence

1. Customer Segmentation
2. Sales Performance
3. Inventory Management
4. Geographical analysis

As there are many multiple possible potential business intelligence options available, I have streamed the data line for customer segmentation which allows for customer identification from as early as their first purchase . This business intelligence is concluded after the inconclusive nature of sales trend due to the scope of data.

### Role and Scenario

The date set includes detailed sales data made on an E-commerce platform over the period of a year, along with customer demographic information, transaction histories, and product descriptions. The business is aiming to understand purchasing patterns, identify key customer segments, and uncover sales trends to improve both product offerings and targeted marketing strategies. As a data analyst, my role is to analyze sales trends and customer behavior to provide actionable insights that can help the e-commerce business optimize its marketing, sales strategies, and customer engagement.

# Data Preparation:

## Data Cleaning

### Data Cleaning

Customer ID noted to have missing values. As such dropped due to the purposed BI  
Before:  
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

After:  
A screenshot of a computer

Description automatically generated

Using:  
A screen shot of a computer program

Description automatically generated

### Date Normalization

Amend raw CSV to only use date, and convert invoice date into year, month day:  
A screen shot of a computer program

Description automatically generated  
A screenshot of a computer

Description automatically generated

Converting customer ID to integers due recurring float value:  
A screenshot of a computer

Description automatically generated

# Exploratory Data Analysis (EDA)

## Data Shape and Size

A screenshot of a computer

Description automatically generated

## Descriptive Statistics

A screenshot of a computer

Description automatically generated

## Outliers in Data set

### Shorthand view of Order region clustering:

A screenshot of a computer

Description automatically generated

Unspecified Country within the scope (outliers)

A screenshot of a phone

Description automatically generated

The 8 counts of “Unspecified” are then removed from the scope due to the inability to move forward with the lack of specificity on such data

# Pipeline Development

## Heat Map Visualization of the sales

A screenshot of a computer screen

Description automatically generated

## Identification of most prominent Customers

### Grouping of Invoices to CustomerID

A screenshot of a computer

Description automatically generated

### Total sales based on CustomerID

A screenshot of a computer

Description automatically generated

### Visualization of Top 10 Sales of customer

A screenshot of a computer

Description automatically generated

Due to the nature of large samples, we can’t properly illustrate the customers of which are widely spread out in a large range. As such I had adopted a broken axis methodology to better understand the CustomerID

### Broken Axis Plot

A screenshot of a graph

Description automatically generated

As the broken axis plot was found to allow for better visualization where all the CustomerID are visible, I had found that the information obtained from the charts as I found the visualization had not been able to provide any additional information and analyzation I had not move forward in further broken axis.

## Monthly Sales

A screenshot of a computer

Description automatically generated

I was able to identify the monthly sales pattern across the noting that 2011 is a weaker month due to the data not entirely encompassing the full month.

### Sales trend prediction Model

I had at this point attempted to create an simple sales trend analysis.

#### Linear Regression

A screenshot of a computer program

Description automatically generated

It is found that Iinear regression has returned a negative R-squared values suggesting a poor fit as such we will attempt and polynomial regression as an alternative

#### Polynomial Regression

A screenshot of a computer program

Description automatically generated

Polynomial regression has also been found to provide a negative R-squared value. Thus, based on the above analysis, I will be pivoting into a categorical approach using product descriptors.

## Natural Language Toolkit, Keywords

Due to the large hand nature of the differing product descriptors, there is a need to implement an word cloud to best identify the category names to such as “Furniture”, “Jewelry” etc.

A screenshot of a computer program

Description automatically generated

We had identified 1483 unique descriptions after tokenizing the descriptions and stemming the keywords.

### Visualization of the Keywords

Snapshot of keywords occurrence

A graph of words

Description automatically generated

Due the large amount of keywords within over 14000 and the most common keyword (heart) occurring over 200 times there is a need to revaluate the keywords. As such I had dropped the certain keywords that are less useful in describing the products such as “pink” and any low occurrence keywords

### Dropping of color and low occurrence keywords

A computer screen shot of a code

Description automatically generated

### Conversion of keywords for use

A screenshot of a computer

Description automatically generated

In the above we created a matrix of which all preserved keywords were assigned a column for which we would assign 1 if said keyword is present in the records this is done for clustering

### Creating Clusters and evaluating silhouette scores

A screen shot of a computer program

Description automatically generated

To better visualize the silhouette scores to best select the number of clusters of which we would recommend.

A screen shot of a computer screen

Description automatically generated

Based on the inherent Silhouette scores it is recommended that I use 8 clusters. However, in practice I had found that 8 clusters would return multiple clusters with little entries as such the cluster values of 5 had been chosen.

8 Clusters:

A screenshot of a computer program

Description automatically generated

5 Clusters:

A screenshot of a computer program

Description automatically generated

### Word Cloud

A screenshot of a computer

Description automatically generated

As the word cloud does not yield any obvious keywords, we are unable to conclusively categorize the products based on their description

### Principle Component Analysis

A graph with a line

Description automatically generated

PCA is performed to better understand the number of components required to explain the data. It is noted that we require over 100 components to explain 90% of the variance in data. Thus, the number of components is limited.

### Visualization of clusters

A group of colored blots

Description automatically generated

## Categorizing Customers

### Categorizing Product

We have products based on their unit price

A screenshot of a computer screen

Description automatically generated

### Total Sales of Product per Category

A screenshot of a computer screen

Description automatically generated

### Evaluation of customer

A screenshot of a computer

Description automatically generated

Noted the number of customers with single purchases

A screenshot of a computer

Description automatically generated

### Principle Component Analysis

A graph with blue and green squares

Description automatically generated

### Identifying K values

A screen shot of a graph

Description automatically generated

A screen shot of a computer screen

Description automatically generated

It is noteworthy that at this point our optimal K cluster has deviated from the expected K value. As per our the source content where I had adopted the customer segmentation direction from [source](https://www.kaggle.com/code/fabiendaniel/customer-segmentation#Customer-segmentation). We have deviated from the expected no. of clusters.

Due to the nature of this capstone project and the intended time frame, I had taken a pause on the further evaluation of the classification, prediction and testing on of the afore mentioned customer segmentation. As such I will be providing an intermediary pause for the conclusion of the project while continuing to work on this project on my own time separated from the project evaluation.

## Documentation and Recommendations

It is understood that the scope of which I had initially intended deviated from the original purpose of sales trend analysis. However, we can summarize the findings based on current conjecture to the following,

* There is no inherent sales trend based on the existing sales data
* The bulk of the sales are obtained from UK
* There is no clear umbrella for the categorical data regarding the product descriptors

It is recommended that further analysis is to be carried out such as performing K-mode analysis for clustering and taking into account holiday patterns for the product sales eg. Christmas decorations during the November and December period.