

MLG382: CYO Project Report

Ann-Mari De Beer

Jason Romario Maracha Bond

Tiaan Kemp

Jeané Christine Thomas - 577864

**Thursday, 24 April 2025**

Dino Giovannoni • Hellen Machingauta • Francois Venter

Table of Contents

[Customer Segmentation Title 3](#_Toc196003088)

[Introduction 3](#_Toc196003089)

[Problem Statement 3](#_Toc196003090)

[Understanding the Data 3](#_Toc196003091)

[Numeric Clean Summary 3](#_Toc196003092)

[Exploratory Data Analysis (EDA) 3](#_Toc196003093)

[1. Univariate Analysis 3](#_Toc196003094)

[ML Model 8](#_Toc196003095)

[Model Comparison 9](#_Toc196003096)

# Customer Segmentation Title

*GitHub link:  
Render link:*

## Introduction

## Problem Statement

Group customers into segments to better understand behaviour and improve marketing strategies.

## Understanding the Data

### Numeric Clean Summary

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Quantity** | **UnitPrice** | **CustomerID** |
| **count** | 401604.00 | 401604.00 | 401604.00 |
| **mean** | 12.183273 | 3.474064 | 15281.160818 |
| **std** | 250.283037 | 69.764035 | 1714.006089 |
| **min** | -80995.00 | 0.00 | 12346.00 |
| **25%** | 2.00 | 1.25 | 13939.00 |
| **50%** | 2.00 | 1.95 | 15145.00 |
| **75%** | 12.00 | 3.75 | 16784.00 |
| **max** | 80995.00 | 38970.00 | 18287.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Quantity** | **UnitPrice** | **CustomerID** | **InvoiceMonth** | **Invoice Day** |
| **Quantity** | 1.000000 | -0.001243 | -0.003457 | -0.002354 | -0.000237 |
| **UnitPrice** | -0.001243 | 1.000000 | -0.004524 | -0.003140 | -0.001911 |
| **CustomerID** | -0.003457 | -0.004524 | 1.000000 | 0.029855 | -0.001930 |
| **InvoiceMonth** | -0.002354 | -0.003140 | 0.029855 | 1.000000 | -0.117818 |
| **InvoiceDay** | -0.000237 | -0.001911 | -0.001930 | -0.117818 | 1.000000 |

## Exploratory Data Analysis (EDA)

### 1. Univariate Analysis

#### 1.1 Correlation heatmap

A screenshot of a graph

AI-generated content may be incorrect.

Figure 1

#### 1.2 Frequently Purchased From Countries

A graph with different colored bars

AI-generated content may be incorrect.

Figure 2: Frequently Purchased From Countries

**Observation:** As seen in Figure 2, the United Kingdom is the country where customers purchase items the most form.

#### 1.3 Orders Per Month

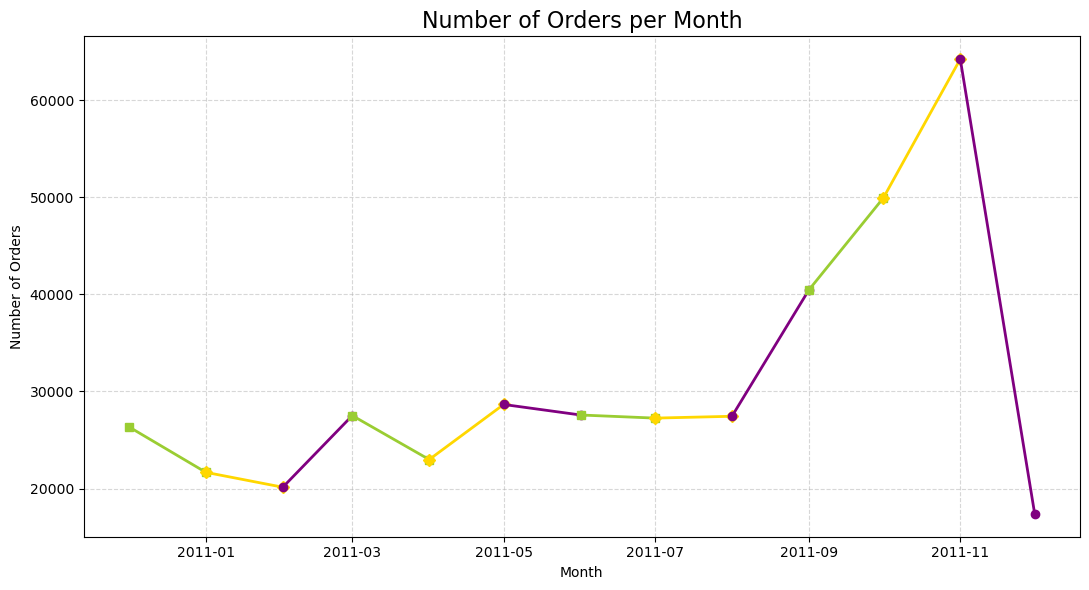


Figure 3: Number of orders per Month

**Observation:** Figure 3 shows that November is

#### 1.4 Orders Per Day

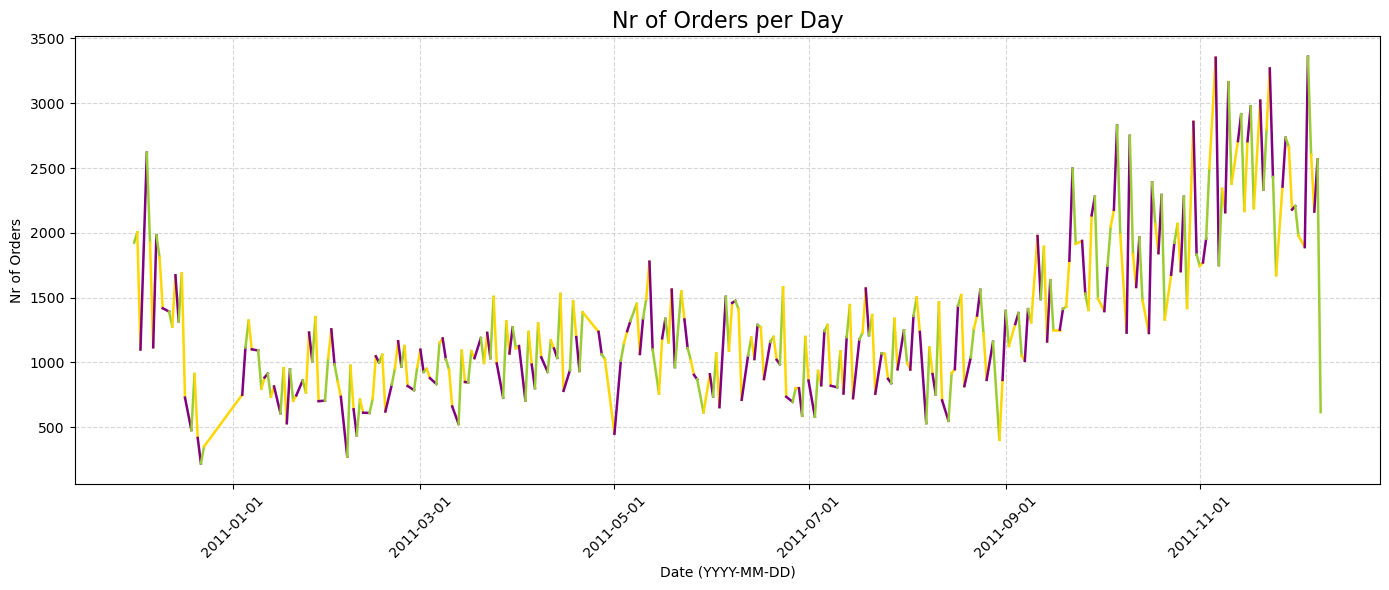


Figure 4: Number of orders per

**Observation:** Figure 3 shows that November is

#### 1.5

A collage of different colored squares

AI-generated content may be incorrect.

Figure 5

A group of graphs with different colors

AI-generated content may be incorrect.

Figure 6

**Observation:**

## ML Model

A graph with a line

AI-generated content may be incorrect.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **Recency** | **Frequency** | **Monetary** | **Num\_Customers** |
| 0 | 10.787565 | 28.575130 | 12190.961710 | 193 |
| 1 | 248.927577 | 1.805942 | 453.425572 | 1077 |
| 2 | 5.090909 | 109.909091 | 124150.432727 | 11 |
| 3 | 42.768360 | 4.374313 | 1317.266164 | 3091 |

#### K-distance graph for DBSCAN

A graph with numbers and lines

AI-generated content may be incorrect.

#### Dimensionality Reduction and Visualization

A graph of a diagram

AI-generated content may be incorrect.

### Model Comparison

A graph with colored dots

AI-generated content may be incorrect.

**Observation:**

**KMeans:** Clusters are fairly compact and balanced. KMeans assumes spherical clusters, which looks decent here.

**Agglomerative:** Similar results to KMeans, with some variation. Can perform better on non-globular shapes.

**DBSCAN:**    Very different — likely identified outliers or noise (see sparse points). Best if you want to ignore anomalies.