**Technical Report**

Alba Francisco

Project-UK-High-Value-Customers-Identification

[fralba806@gmail.com](mailto:fralba98@gmail.com)

**2. DATA INFORMATION**

**2.1 Data source**

The data is available for all public in Kaggle:

Data link: <https://www.kaggle.com/vik2012kvs/high-value-customers-identification>

**2.2 Data description**

|  |  |
| --- | --- |
| Ecommerce | |
| Number of attributes: | 8 |
| Number of observations: | 541904 |
| Missing values: | 136534 of 4335264 (3.15%) |
| Duplicated rows: | 5269 (0.97%) |
| File size: | 40.7 MB |
| Attributes types | |
| Number: | 3 |
| Category: | 0 |
| Binary: | 0 |
| Date: | 1 |
| URL: | 0 |
| Geographic: | 1 |
| Text: | 3 |

**2.3 Data exploration**

|  |  |
| --- | --- |
| InvoiceNo | |
| Type: | Text |
| Missing values: | 0 |
| Description: | Unique value which is assigned to every transaction |

|  |  |
| --- | --- |
| StockCode | |
| Type: | Text |
| Missing values: | 0 |
| Description: | Code assigned to every product |

|  |  |
| --- | --- |
| Description | |
| Type: | Text |
| Missing values: | 1454 |
| Description: | Name of each product |

|  |  |
| --- | --- |
| Quantity | |
| Type: | Number |
| Missing values: | 0 |
| Minimum: | -80995 |
| Q1: | 1 |
| Q2: | 3 |
| Q3: | 10 |
| Maximum: | 80995 |
| Mean: | 9.55 |
| Range: | 161950 |
| Interquartile range: | 9 |
| Standard deviation: | 218.081 |
| Variance: | 47559.39 |
| Description: | Quantity of the item purchased in the transaction |

|  |  |
| --- | --- |
| InvoiceDate | |
| Type: | Date |
| Missing values: | 0 |
| Description: | Date when was made the transaction |

|  |  |
| --- | --- |
| UnitPrice | |
| Type: | Number |
| Missing values: | 0 |
| Minimum: | -11062.06 |
| Q1: | 1.25 |
| Q2: | 2.08 |
| Q3: | 4.13 |
| Maximum: | 38970 |
| Mean: | 4.611 |
| Range: | 161950 |
| Interquartile range: | 2.88 |
| Standard deviation: | 96.76 |
| Variance: | 9362.47 |
| Description: | Price per unit of product |

|  |  |
| --- | --- |
| CustomerID | |
| Type: | Number |
| Missing values: | 135080 |
| Description: | ID that represents a single customer |

|  |  |
| --- | --- |
| Country | |
| Type: | Geographic |
| Missing values: | 0 |
| Description: | Country where resides the customer |

**2.4 Quality assurance**

|  |  |
| --- | --- |
| Problem: | Missing values, duplicate values, anomalies |
| Involved attributes: | InvoiceNo, StockCode, Descripcion, Quantity, InvoiceDate, UnitPrice, CustomerID, Country |
| Possible course of action: | Remove missing values, remove duplicate rows, remove anomalies, replace anomalies, replace missing values with mean. |

**3. DATA PREPARATION**

**3.1 Data selection**