College Dublin
Computing • IT • Business

Online Retail

Natalia Gomes Fernandes
James Garza, Strategic Thinking
Higher Diploma in Science in Data Analytics for Business

28th July 2024

Introduction

- E-Commerce / Retail Competitive Market
- UCI ‘Online Retail’ Dataset
- 500,000 + transactions analysed

Project Goals

- Hypothesis: Effective segmentation through buying habits.
  - Create distinct customer segments
  - Predict segment for new customer

Technologies Used

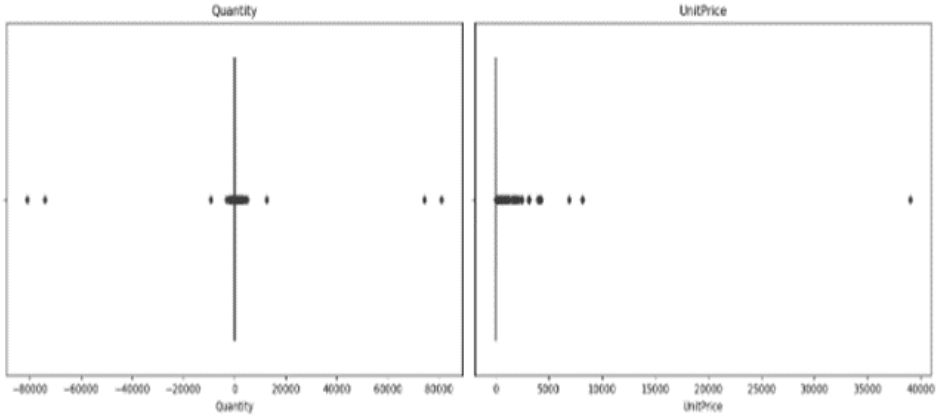
- Unsupervised: K-Means Clustering
- Supervised: Decision Trees and Random Forest
- Python Libraries: NumPy, Pandas, Scikit-learn and more

Data Overview

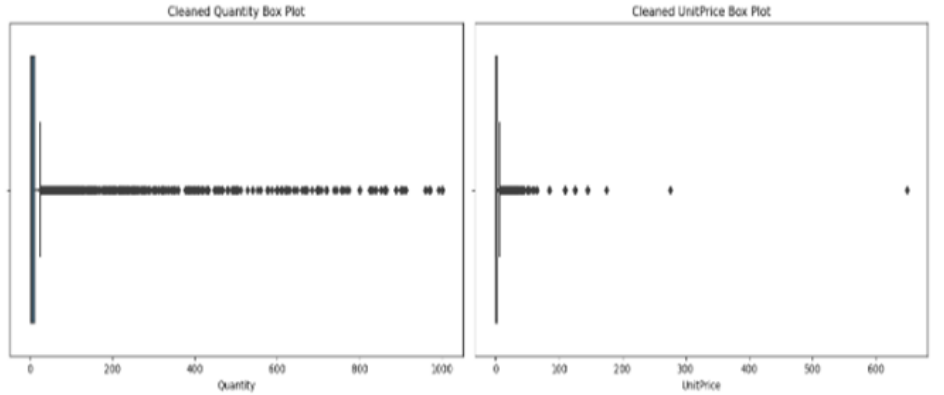
- 8 features – 541,909 observations
- Key features: InvoiceNo, StockCode, Quantity, Price and CustomerID

Data Preparation

- Handling missing values
- Removing duplicates
- Outlier detection and removal

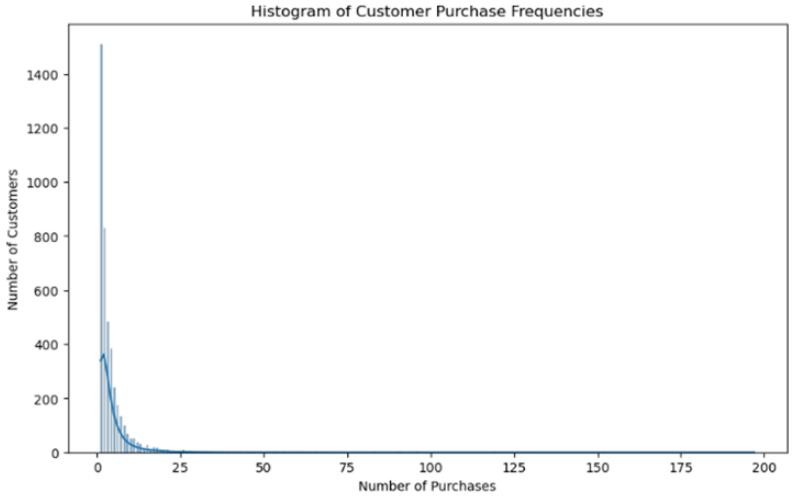


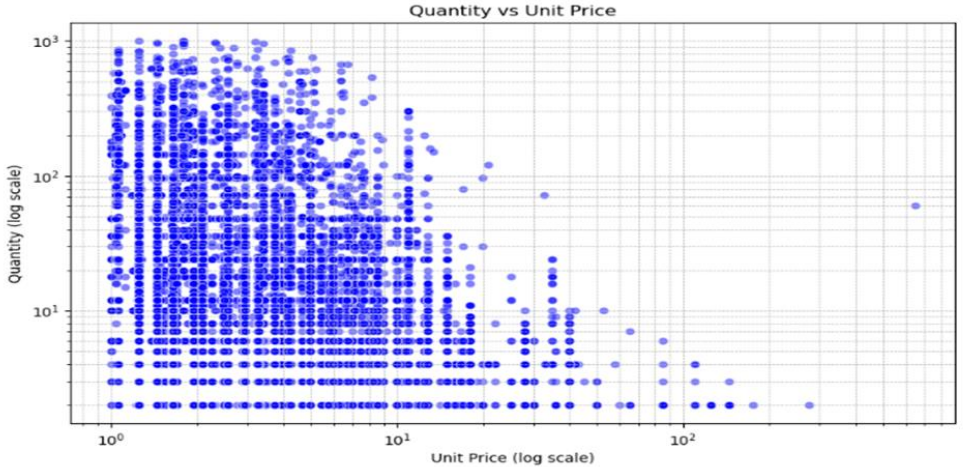
	Quantity	InvoiceDate	UnitPrice	CustomerID
count	401604.000000	401604	401604.000000	401604.000000
mean	12.183273	2011-07-10 12:08:23.848567552	3.474064	15281.160818
min	-80995.000000	2010-12-01 08:26:00	0.000000	12346.000000
25%	2.000000	2011-04-06 15:02:00	1.250000	13939.000000
50%	5.000000	2011-07-29 15:40:00	1.950000	15145.000000
75%	12.000000	2011-10-20 11:58:30	3.750000	16784.000000
max	80995.000000	2011-12-09 12:50:00	38970.000000	18287.000000
std	250.283037	NaN	69.764035	1714.006089



Exploratory Data Analysis

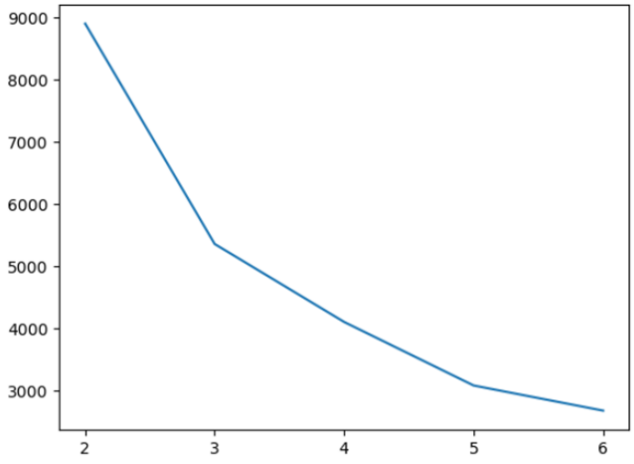
- Customer purchase frequencies
- Total spend vs Number of orders
- Quantity vs Unit Price relationship

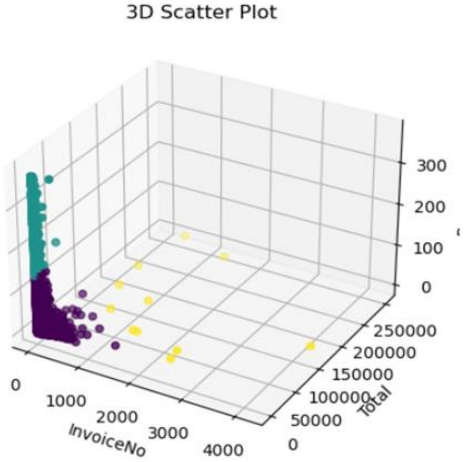




K-Means Clustering

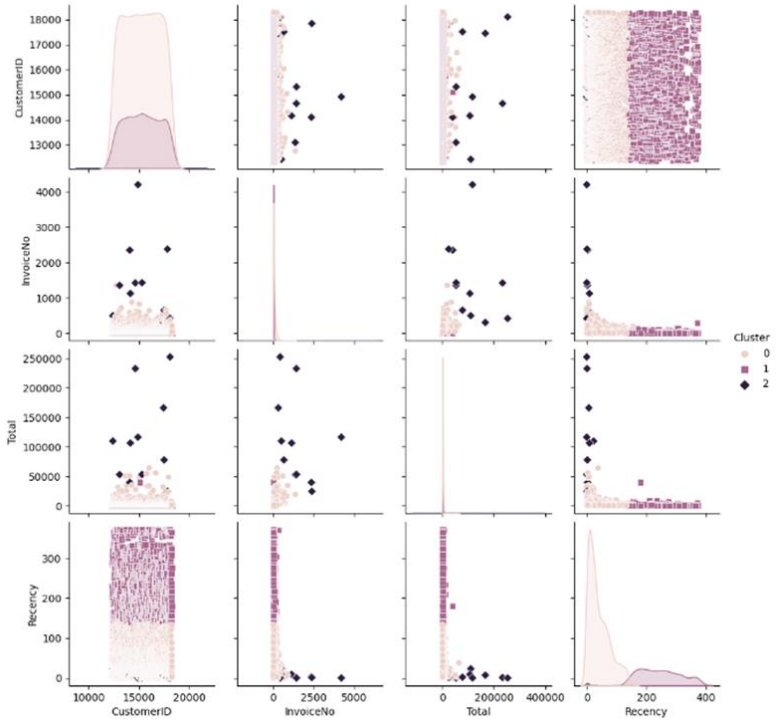
- 3 distinct customer segments identified
- Elbow Method and silhouette score
- 3D visualisation of clusters





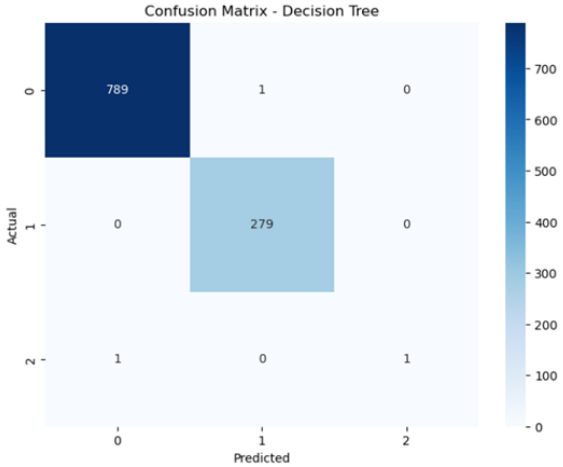
Cluster Overview

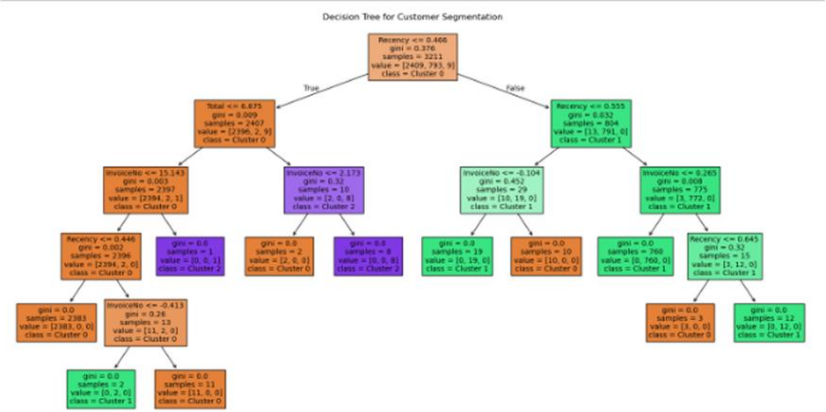
- Cluster 0: Moderate recency
- Cluster 1: Least recent purchases
- Cluster 2: Most active customers



Decision Tree Model

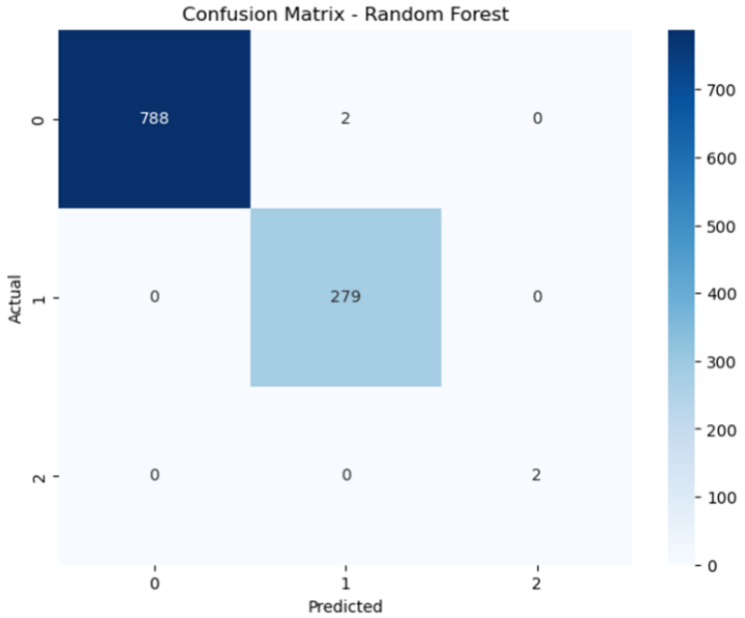
- Predicts cluster for new customer
- 99,81 accuracy achieved
- Confusion Matrix results





Random Forest Model

- Matched Decision Tree accuracy
- More consistent performance
- 100% accuracy for Cluster 1 and 2



Model Comparison

- Both models: High accuracy (99,81%)
- Random Forest: More consistent in cross-validation

Well-separated clusters contribute to high accuracy

```
Cross-validation scores - Decision Tree: [0.99416569 0.99883314 0.99649533 0.9953271  0.99182243]
```

```
Cross-validation scores: [0.99416569 0.99883314 0.99766355 0.9953271  0.99299065]
```

Business Implication

- Tailored marketing strategies
- Optimized inventory management
- Focused customer retention
- Efficient resource allocation

Conclusion

- Powerful insights into customer behaviour
- Predictive capabilities for new customers
- Foundation for data-driven decision making

- Thank you!!