Online Retail II Data Set Analysis

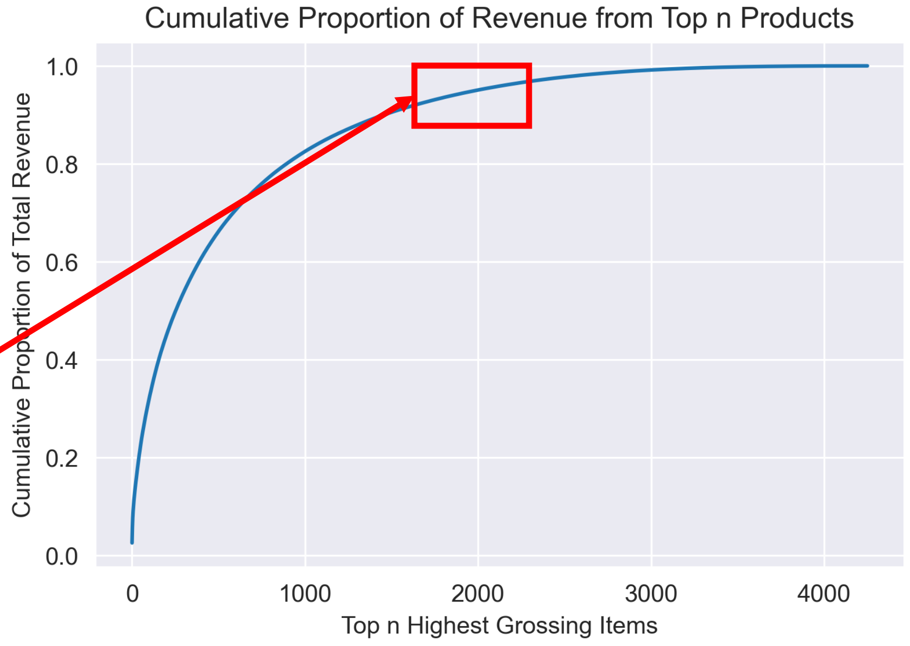
Evan Glas

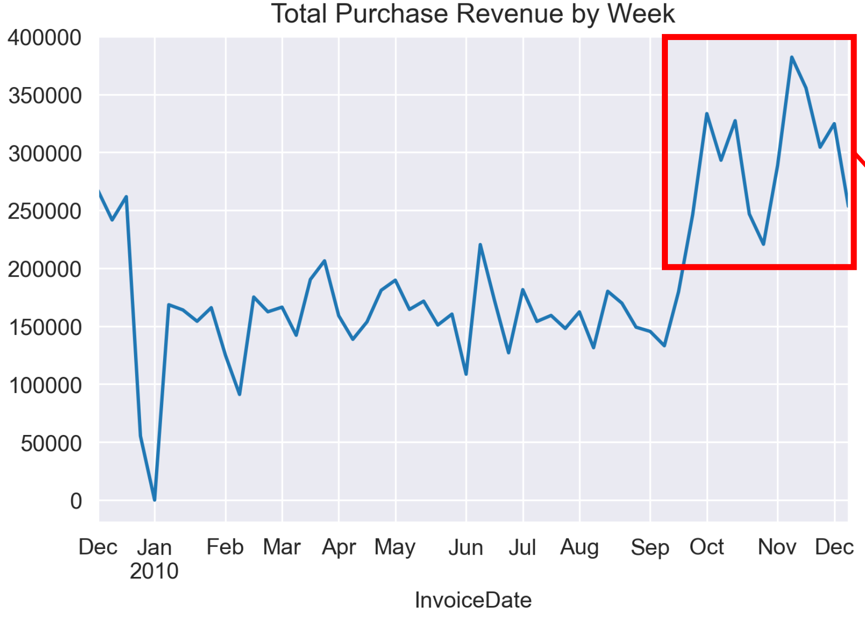
Goals

- Build understanding of 2009-2010 customer dataset
  - Visualize aggregate purchasing patterns
  - Determine aggregate sales information
- Quantify individual customer purchasing patterns
  - Analyze recency of purchases, frequency of purchases, and total monetary value of customers
  - Determine which features correspond to higher lifetime customer value
- Determine which factors predict retention rate, repeat customers
  - Will a customer make an additional purchase?

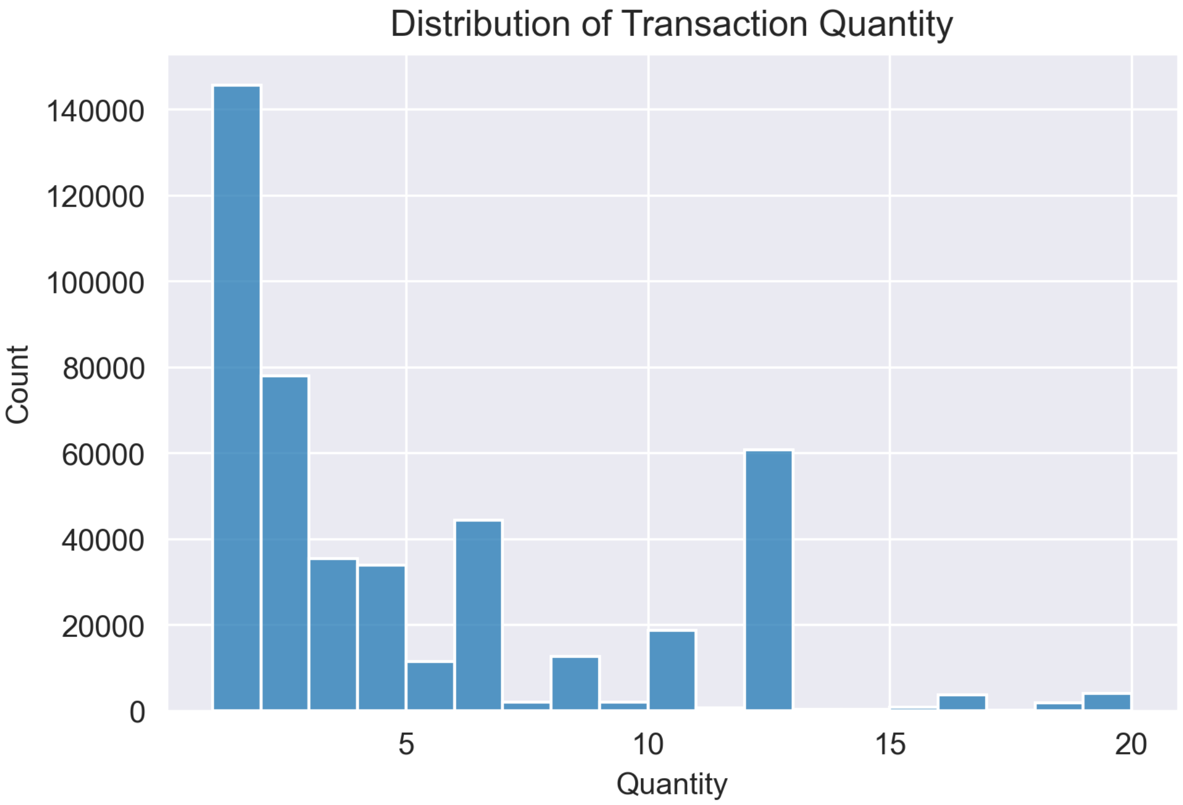
2009 EDA

- **£ 10,306,265** Gross Purchase Revenue
- **4,300** unique customers
- **4,251** unique products sold
  - Top **2000** products account for **95%** of store revenue
- Sold products to customers in **40 countries**
  - Of all purchases,
    - **92.5% UK**
    - 1.8% Ireland
    - 1.5% Germany
    - 1% France
    - 3.2% Other

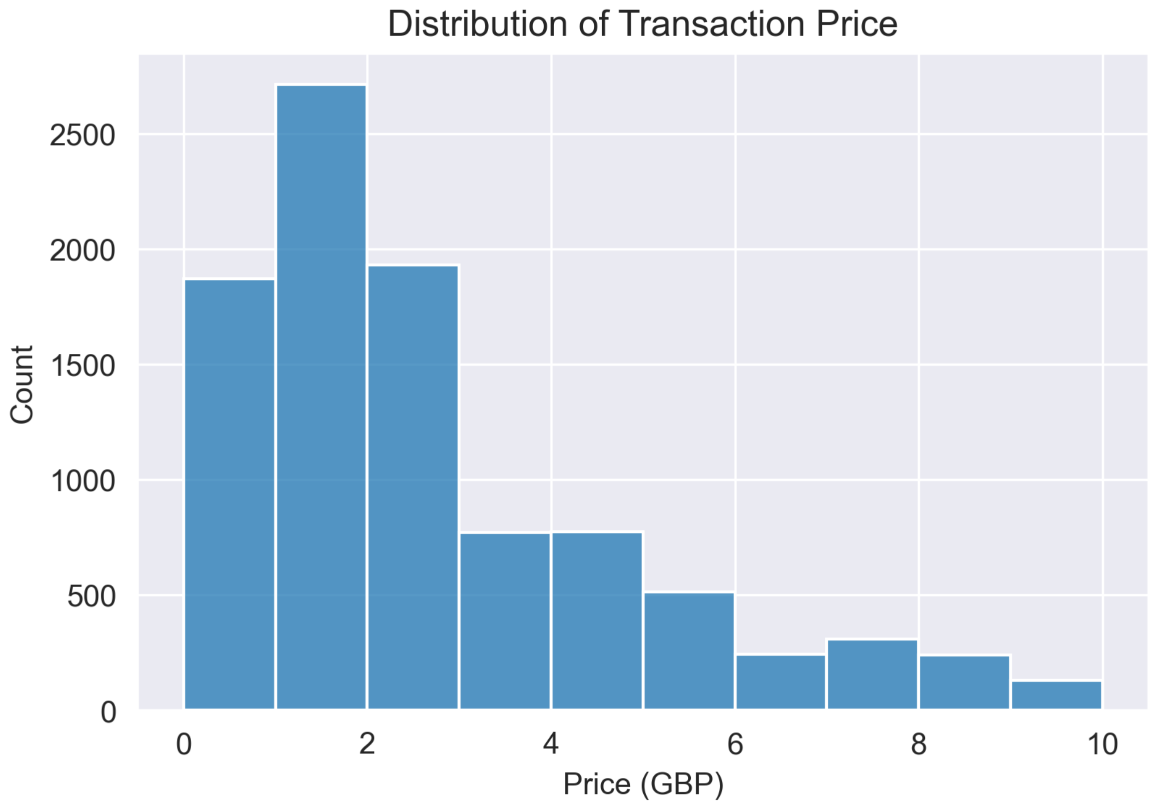




Sales remain relatively stable through most of year, peak in holiday season



Distribution of Transaction Quantity



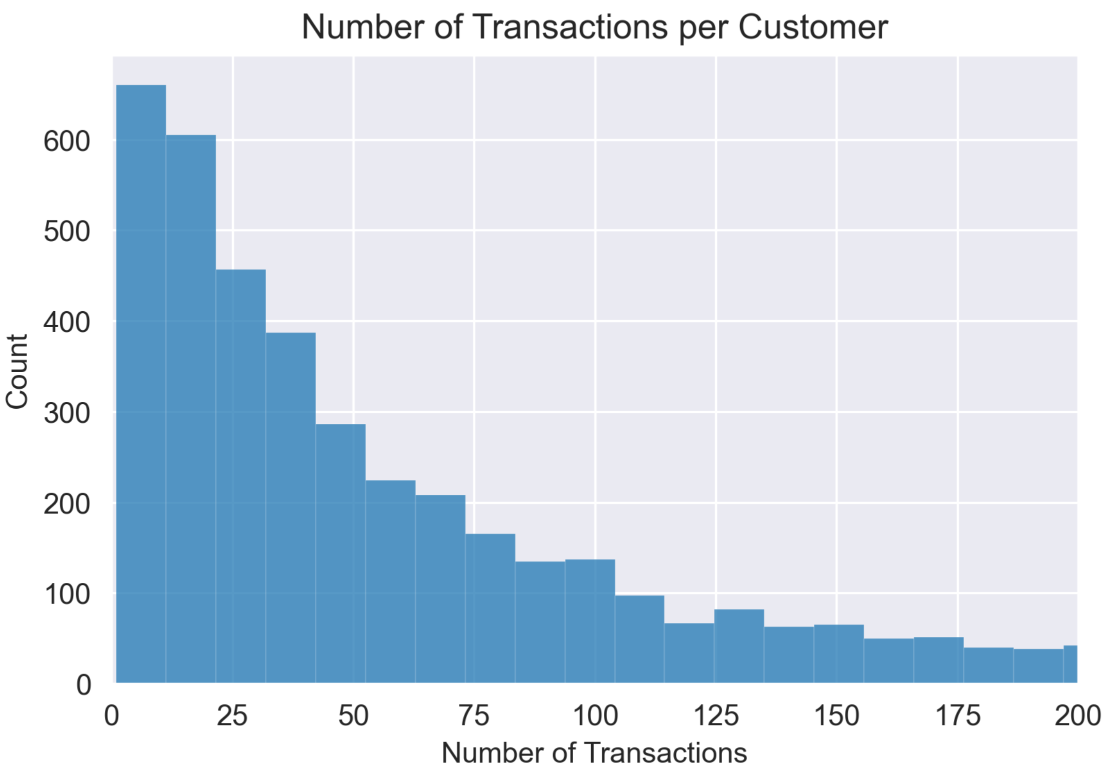
Distribution of Transaction Price

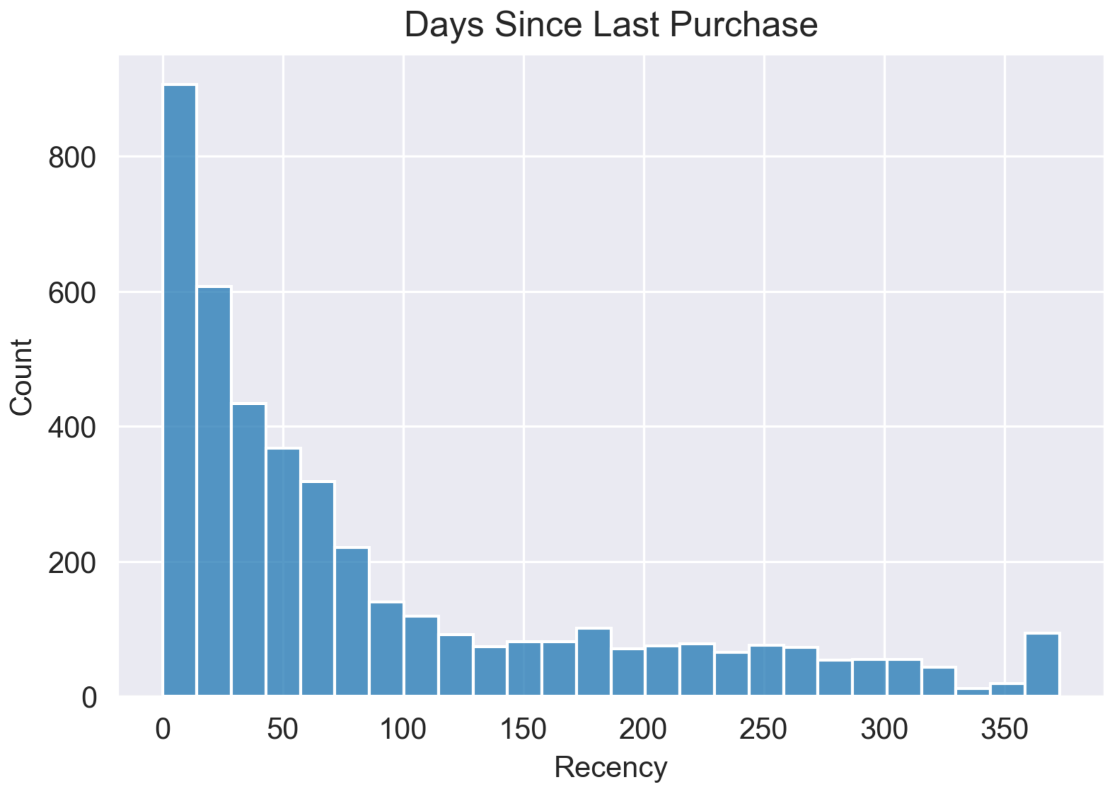
Descriptive Statistics (Quantity):

Average: 11.4
25%: 1
50%: 3
75%: 10
Max: 19152

Descriptive Statistics (Price):

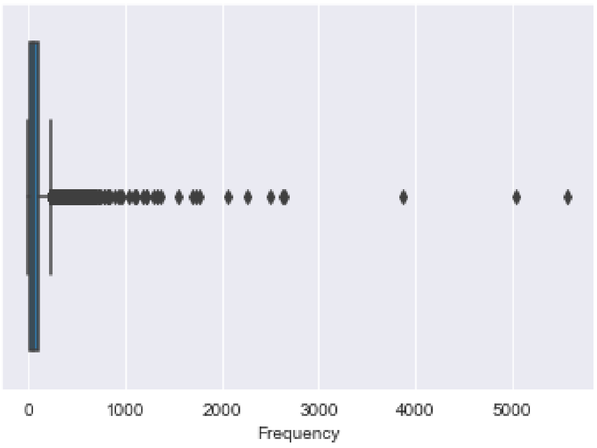
Average: 4.25
25%: 1.25
50%: 2.1
75%: 4.21
Max: 25111





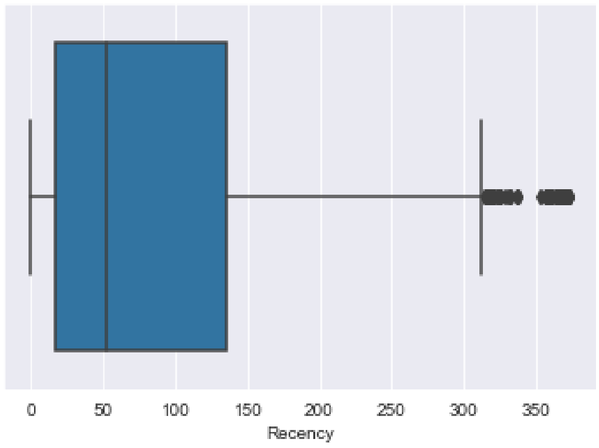
Descriptive Statistics (Frequency):

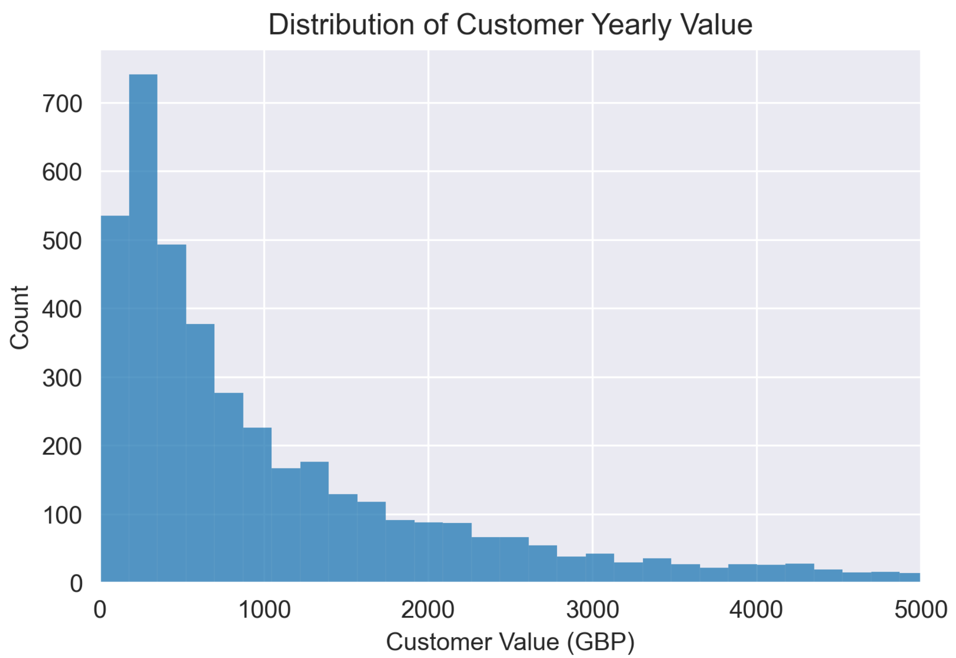
Average: 95
25%: 18
50%: 44
75%: 102
Max: 5570

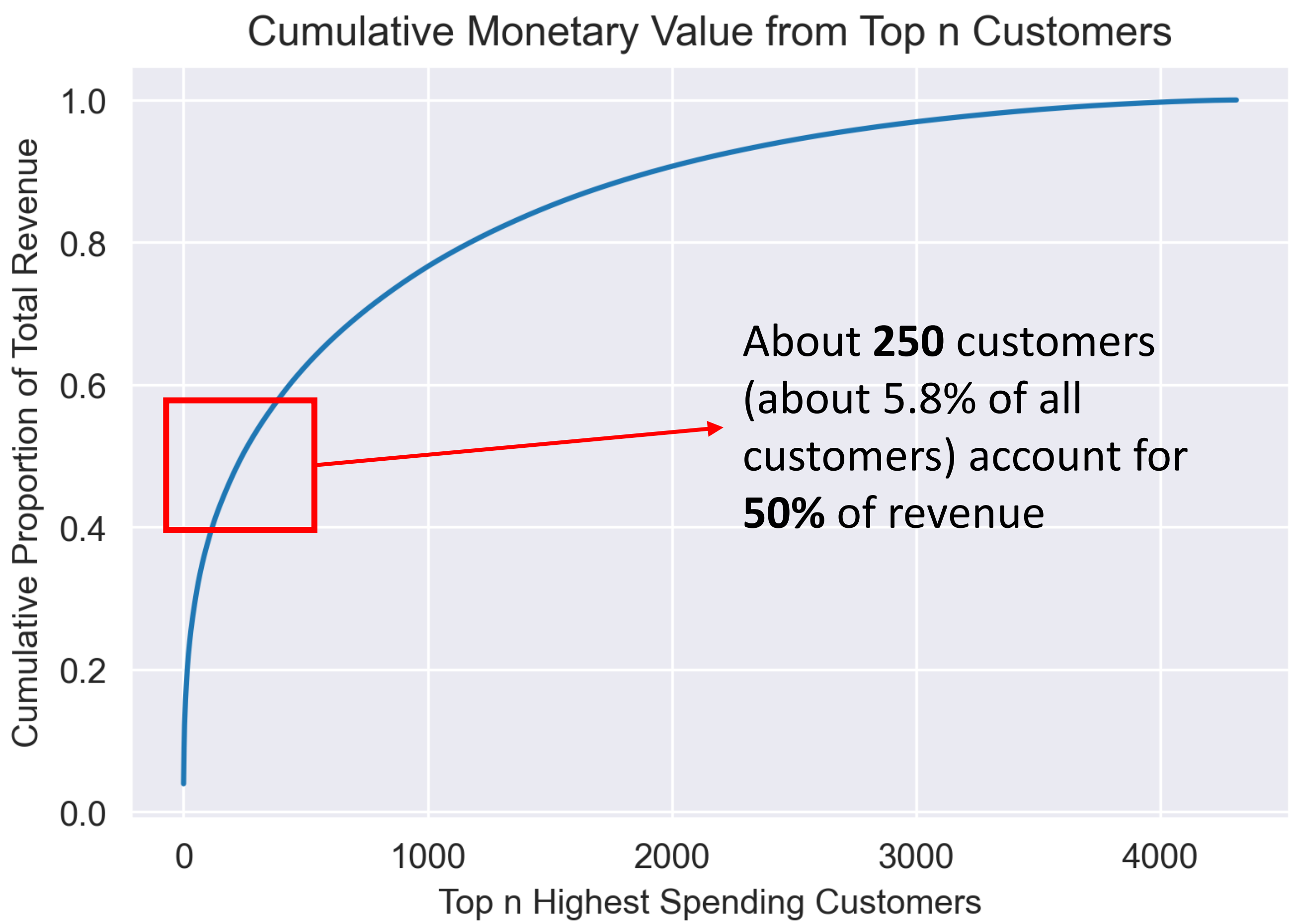


Descriptive Statistics (Recency):

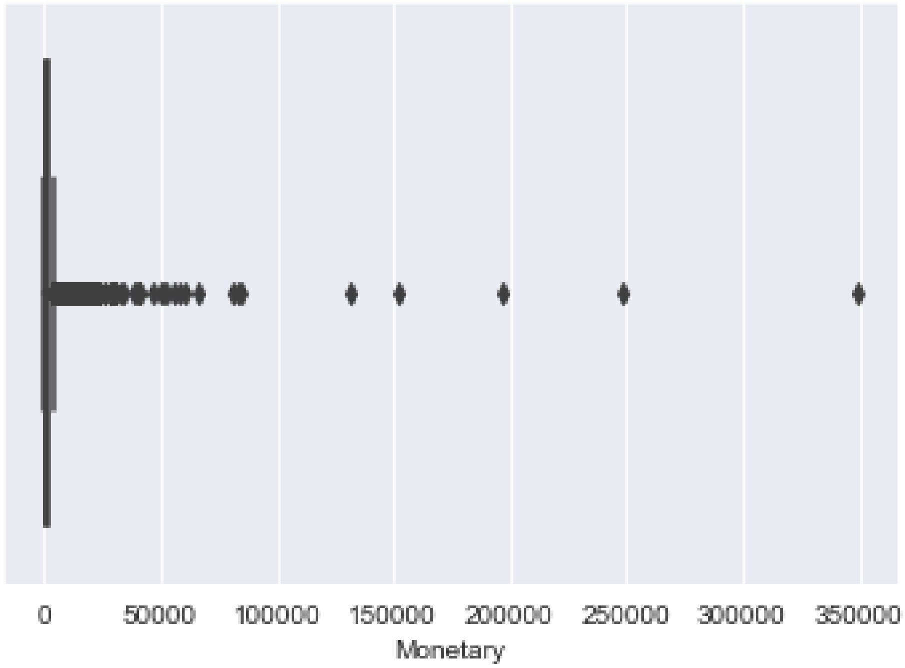
Average: 90 days
25%: 17 days
50%: 52 days
75%: 135 days







About **250** customers (about 5.8% of all customers) account for **50%** of revenue



Descriptive Statistics (Monetary Value):

Average: 2048 GBP

25%: 308

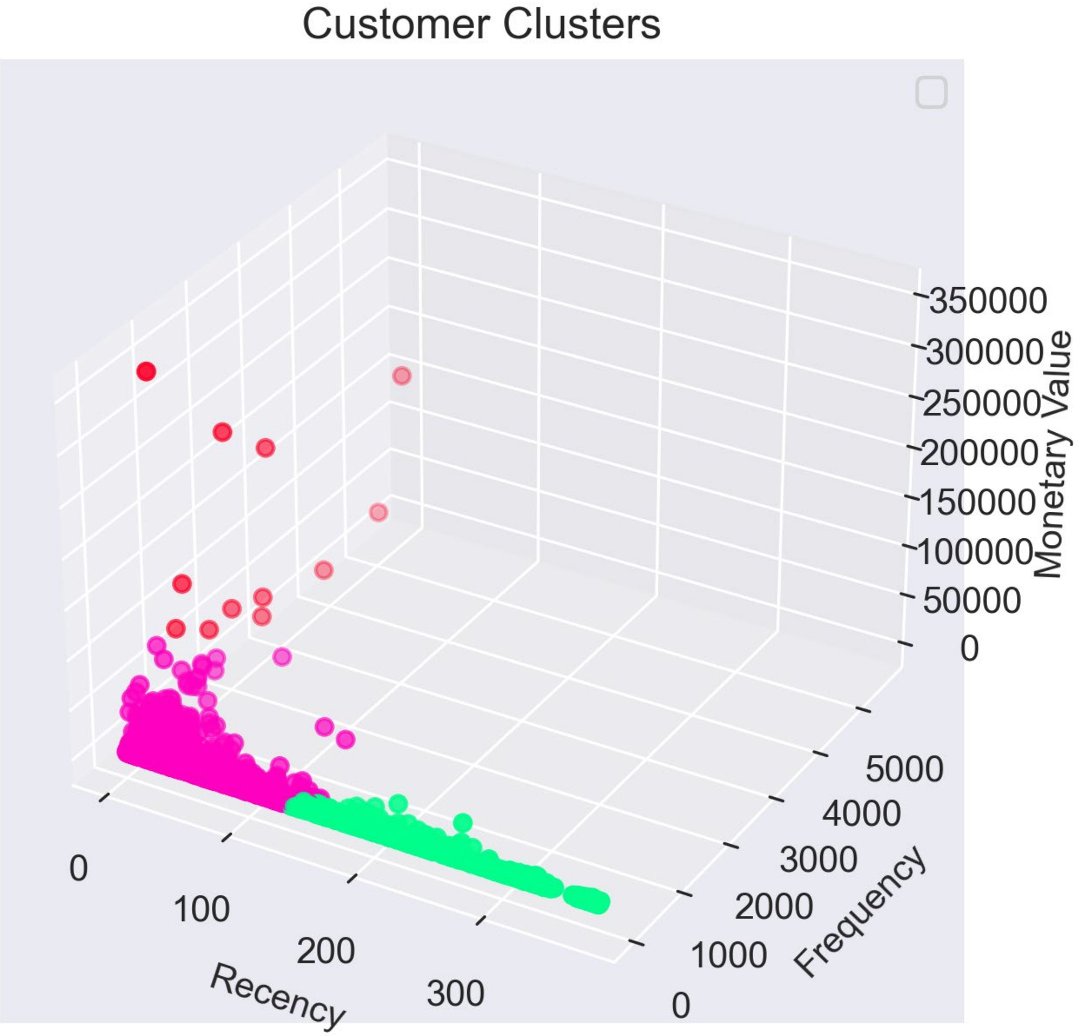
50%: 706

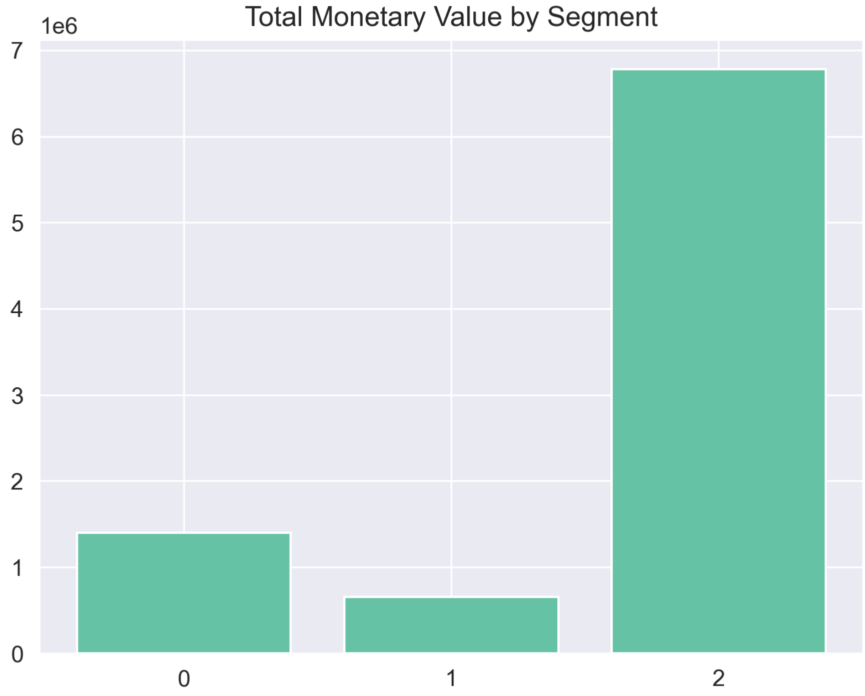
75%: 1723

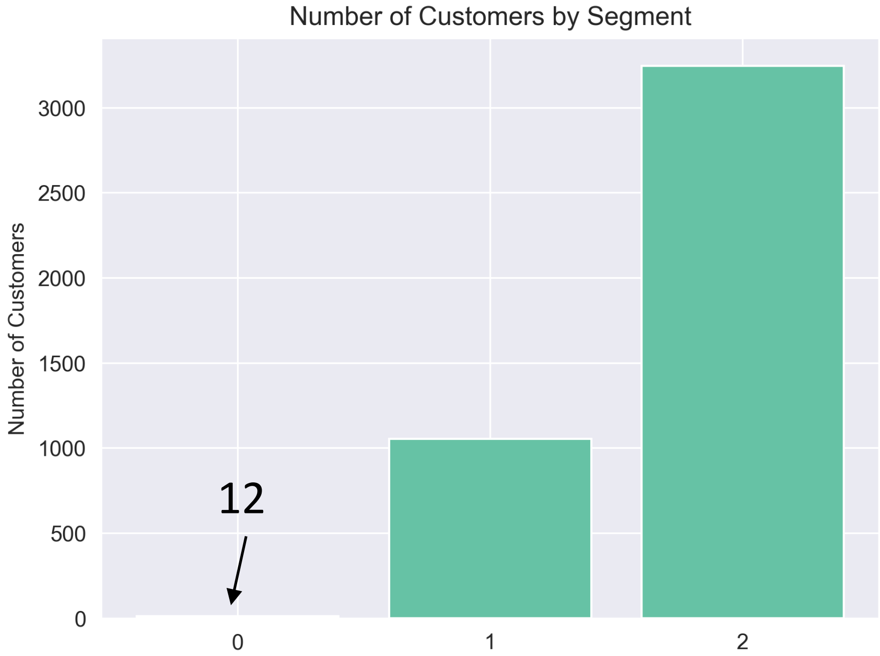
Max: 350,000

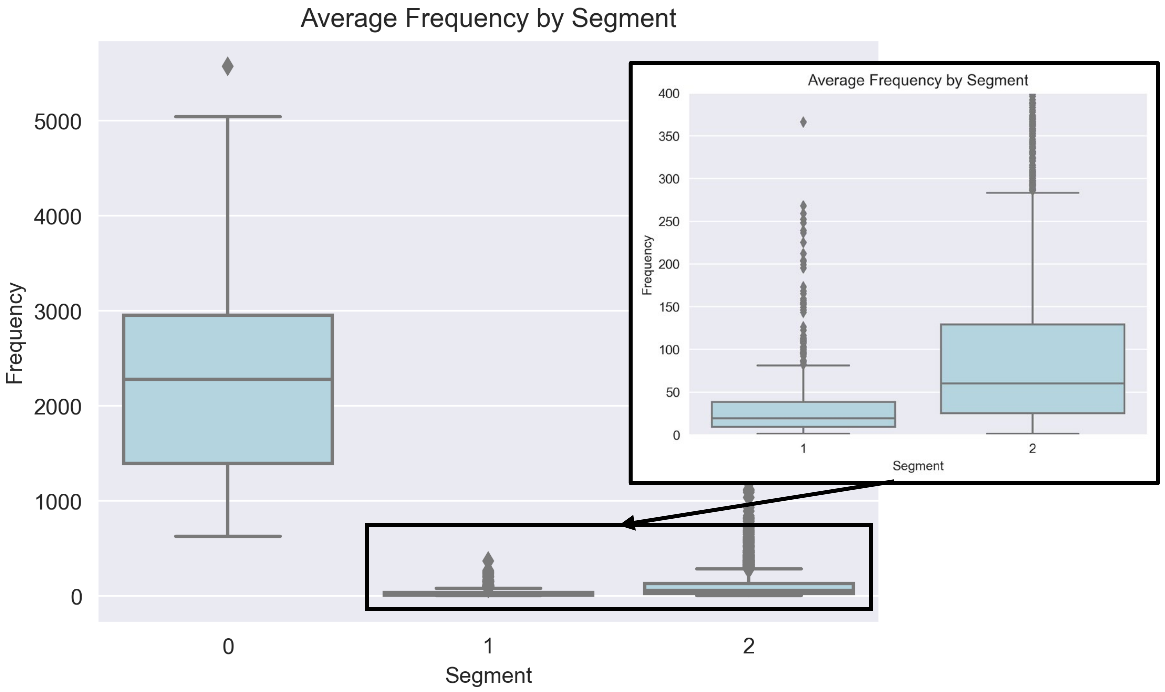
K-Means Customer Segmentation

Customer Clusters

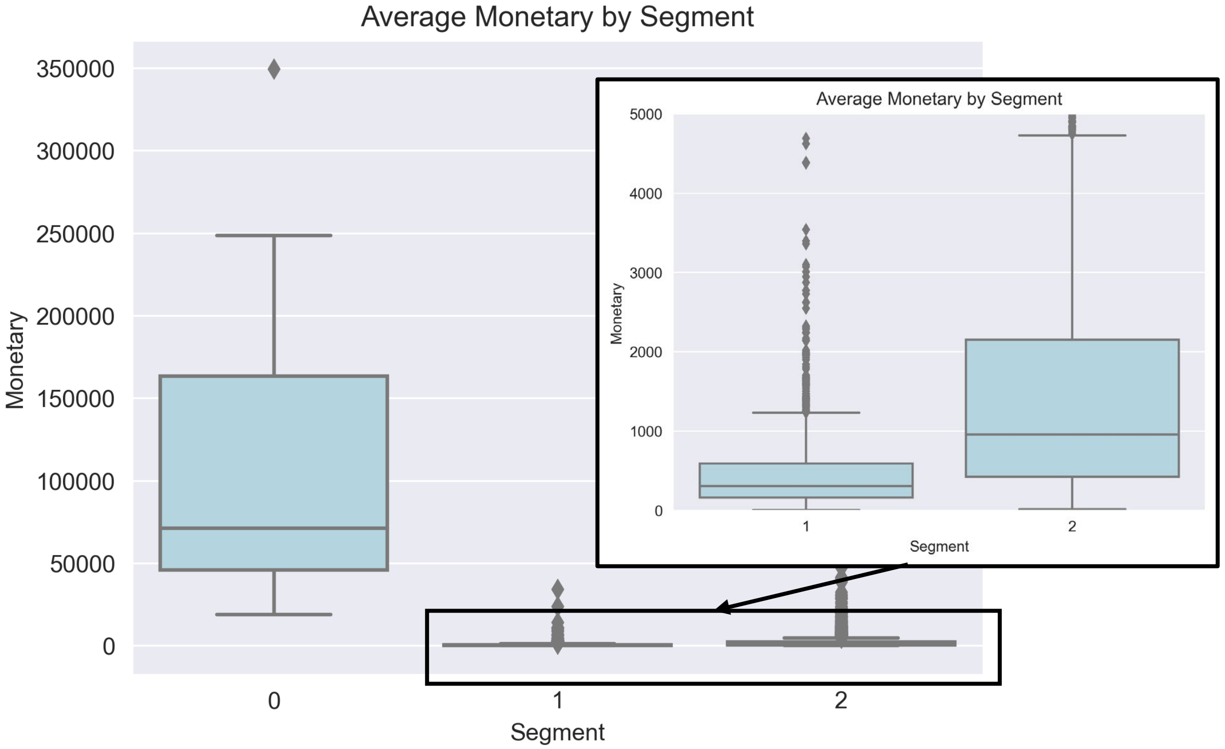




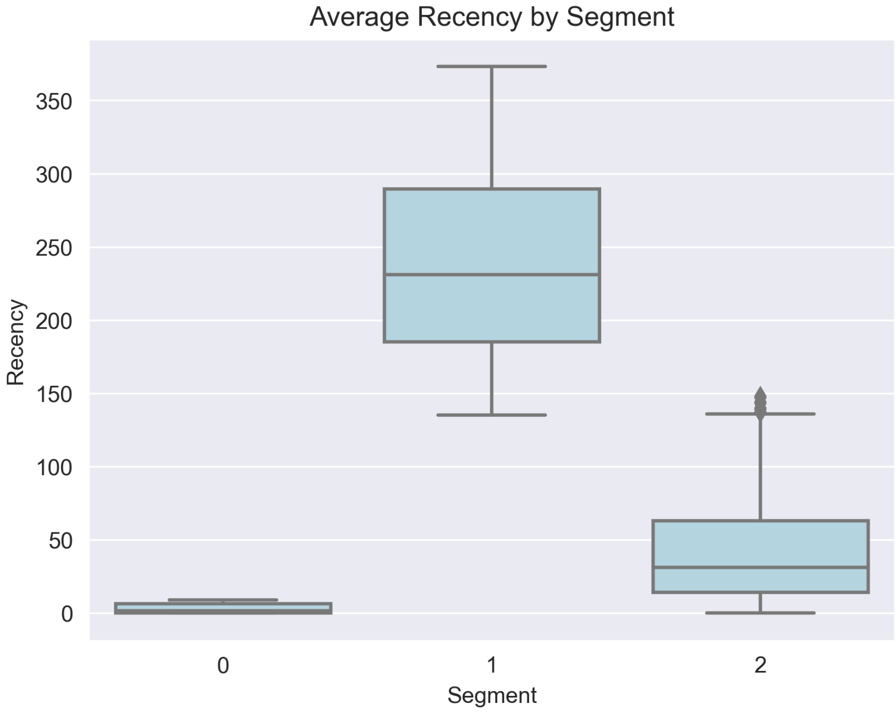




Average Frequency by Segment



Average Monetary by Segment

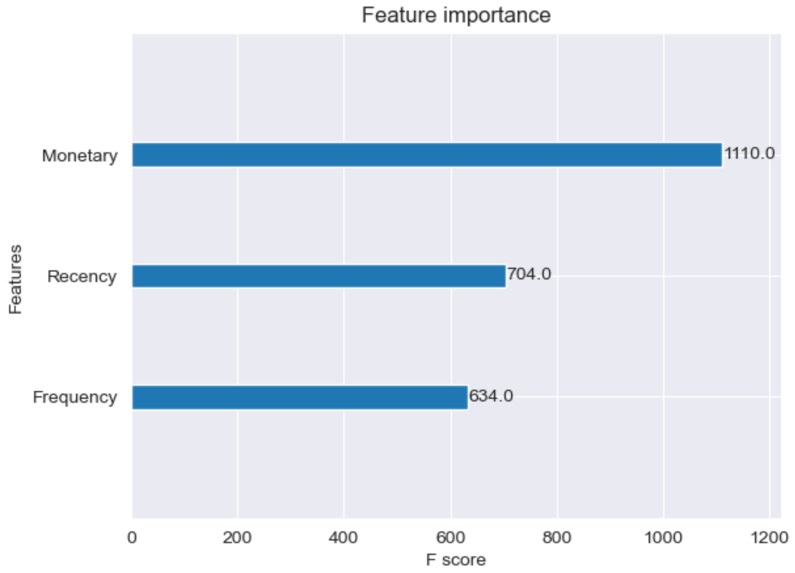


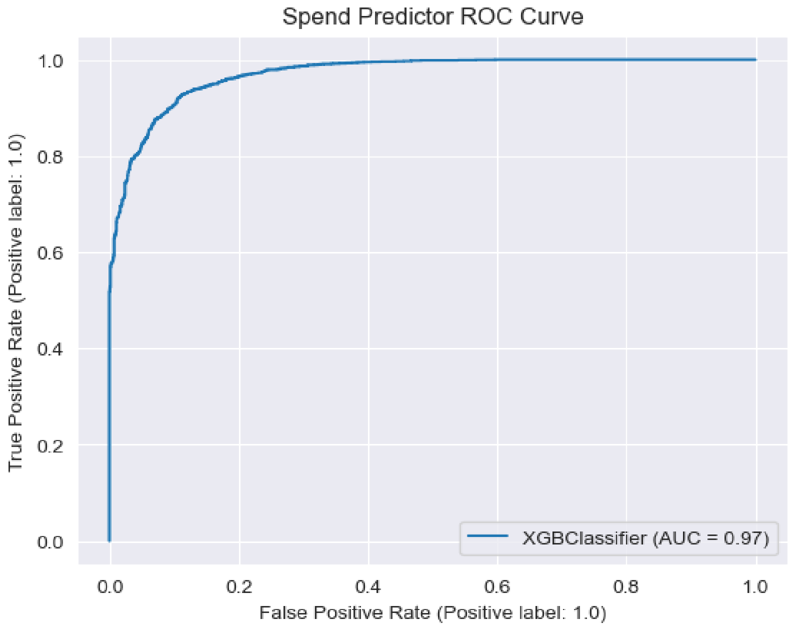
Average Recency by Segment

Segment	Traits	Customer Description
0	High Frequency, Low, High Value	Wild Buyers **(High Spending VIP Customers)**
1	Low Frequency, High Recency, Low Value	Past Customers **(Stopped Spending)**
2	High Frequency, Low Recency, High Value	Active Customers **(Last Purchase within ~2 months)**

Predicting Repeat, High Value Customers

- Built XGBoost Model to determine whether a customer from first 270 days of the dataset would make a purchase in next 90 days.
  - Achieved **90.8% accuracy**, **0.97 AUC** on train data
  - Previous monetary value of customer most predictive of likelihood of another transaction, followed by recency and frequency





- Ran linear regression on recency and frequency of customer purchases on monetary value as target variable. On dataset with outliers removed,
  - Achieved 0.33 R^2 value on dataset 1 day decrease in recency associated with **6.4 GBP** of additional customer monetary value
  - 1 purchase increase in frequency associated with **13.53 GBP** of additional customer monetary value

Next Steps

- Incorporate data from 2010-2011, compare analyses
- Use test set of randomly chosen customers to determine the extent to which XGBoost Classifier will generalize
- If possible, incorporate pre and post 2009-2011 data in order to gain better understanding of entire customer lifetime