CLIENT CLUSTERING & INVENTORY FORECAST

Dimitra Tsagkalidou

PROJECT OVERVIEW

COMPANY

- Online **retail** store
- **UK**-based and 90% of customers in UK
- **Wholesale** and **retail**

DATA

- **Transactional data**
- From 01/12/2010 until 09/12/2011 (**1 year**)
- **541909** instances
- Columns (8): *Invoice No, Stock Code, Description, Quantity, invoice Date, Unit Price, Customer ID, Country*

CAN THE COMPANY OPTIMIZE PRODUCT INVENTORY TO ENSURE TIMELY AVAILABILITY AND MINIMIZE STOCKOUTS?

INVENTORY FORECAST

Implementing **supervised time series machine learning** models to predict inventory for **top-selling** products, considering **trends** and seasonality.

Performance metrics: RMSE, MAE

DATA CLEANING

DUPLICATES

Dropping duplicates

PRODUCT FILTERING

Keeping only instances that refer to products (no shipping, bank fees etc)

RESOLVING MULTIPLE ENTRIES FOR A SINGLE PRODUCT

Addressing cases where a single product has multiple codes, descriptions, or prices. Adjusting StockCode for bundled items and retaining the mode for multiple descriptions, which may have arisen from marketing or typo correction purposes.

CANCELED ORDERS

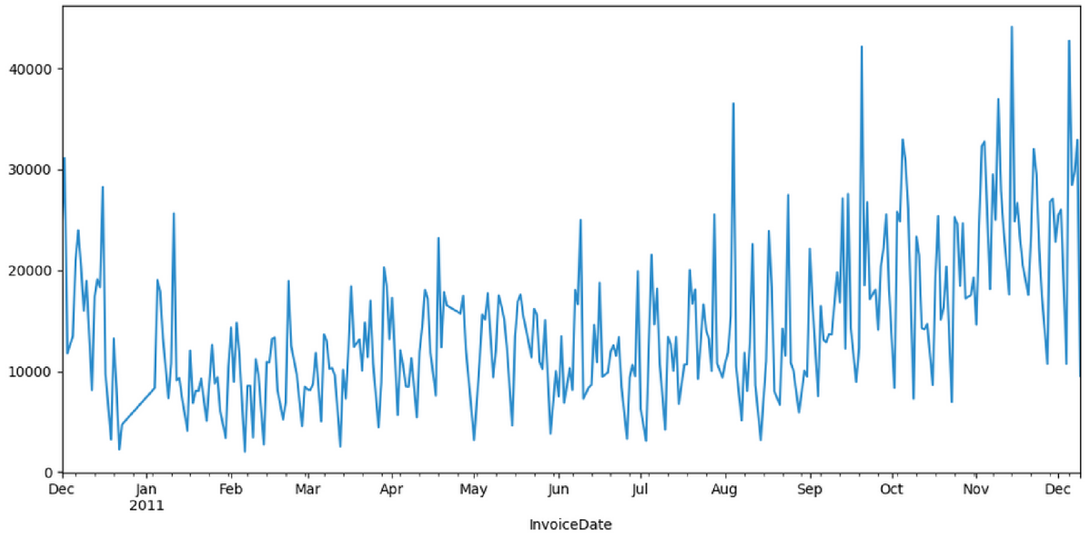
Filtering out canceled orders (not within the scope of the analysis)

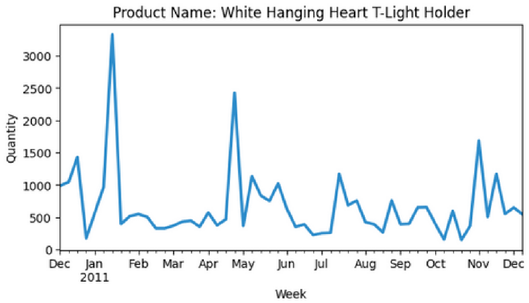
OUTLIERS

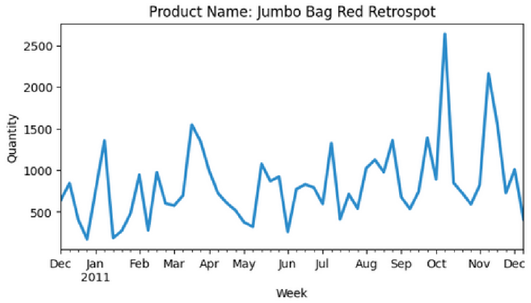
Outlier scaling and clipping when needed

EDA

Main observations and concerns







- Data spans **only 1 year**, hindering clear seasonality
- Presence of numerous **outliers and noise**
- Large **variability in ordering behavior** among clients

CAN THESE CHARACTERISTICS HINDER THE BUILDING OF A RELIABLE FORECASTING MODEL?

~~INVENTORY FORECAST~~

1. CLIENT CLUSTERING

First, let's take a step back and focus on clustering the clients. This approach helps group clients based on **similar ordering behavior, reducing noise** and potentially **improving the accuracy** of the forecasting model. I'll employ an **unsupervised clustering model**, exploring **distance** and **density-based** algorithms.

Performance metrics: Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index, Count of customers per cluster, Visual inspection

CLIENT CLUSTERING
FEATURE ENGINEERING

TOTAL QUANTITY

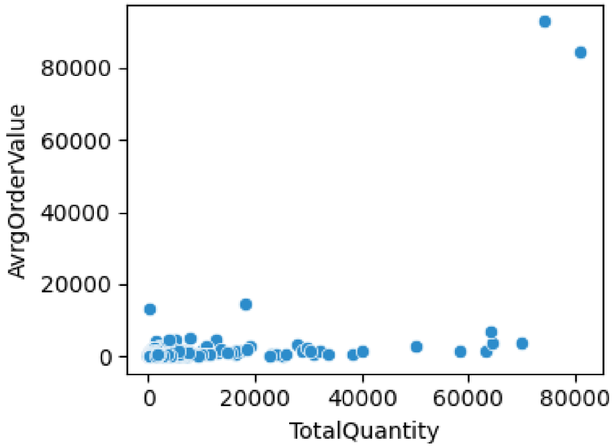
TOTAL VALUE

AVERAGE QUANTITY PER ITEM PER ORDER

AVERAGE ORDER VALUE

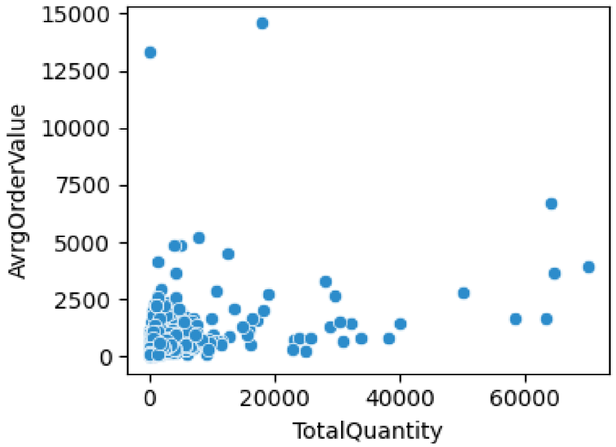
*** ONLY FOR UK**

CLIENT CLUSTERING
OUTLIERS



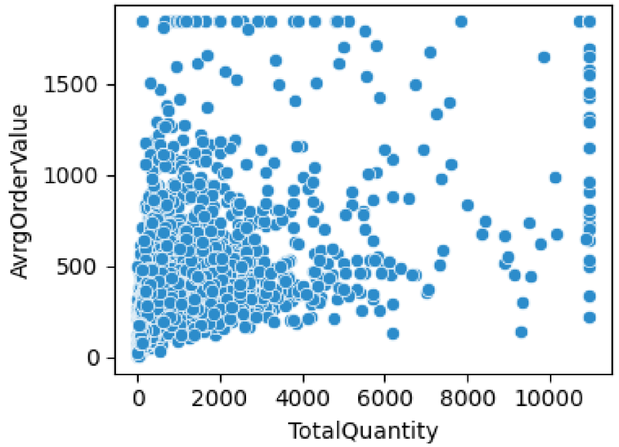
SCALING TO A RANGE

First type of outliers stems from two **unusually large invoices**, from **one-time customers**. To mitigate their impact on the dataset, I'll scale them to the next highest values in the dataset.



CLIPPING

The second type of outliers, though more varied, still contribute significant **noise**. By removing them using the s**tandard deviation method**, keeping data within ±3σ, I aim to achieve a more accurate and representative dataset.



Significant improvement in the overall quality and distribution of the data!

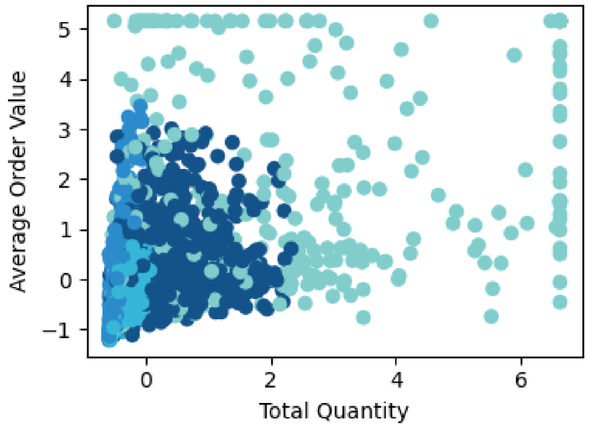
CLIENT CLUSTERING
SCALING & DIMENSIONALITY REDUCTION

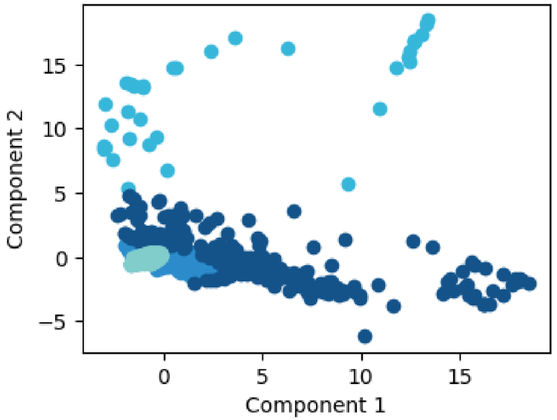
CustomerID	TotalQuantity	TotalValue	AvrgQuantity	AvrgOrderValue
12346	4800	6000.00	471.27	1845.31
12747	1275	4722.51	12.38	429.32

Before scaling

CustomerID				
12346	2.575901	1.312867	10.005464	5.154360
12747	0.256937	0.927293	-0.132466	0.255785

After scaling





DATA SCALING

Data scaled using StandardScaler.
Values are transformed to end up with mean = 0 and std = 1.

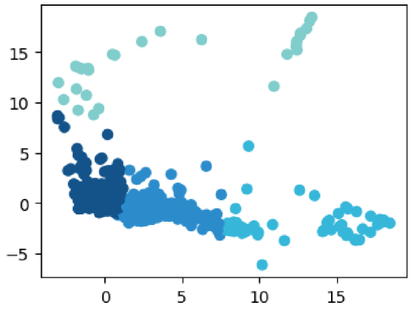
DIMENSIONALITY REDUCTION

Non-linear dimensionality reduction through Isometric Mapping.

1st model:
Silhouette Coefficient: 0.44
Calinski-Harabasz Index: 1497.50
Davies-Bouldin Index: 1.18

2nd model:
Silhouette Coefficient: 0.73
Calinski-Harabasz Index: 2444.32
Davies-Bouldin Index: 0.95

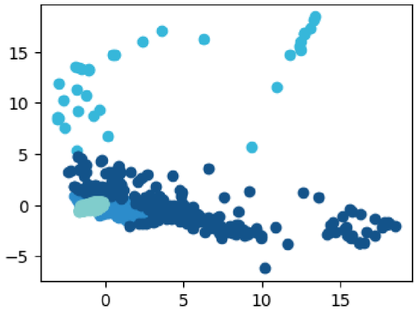
CLIENT CLUSTERING
MODELS



K-Means

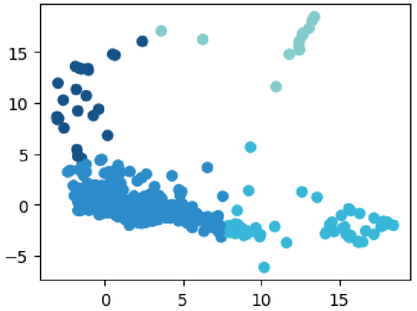
Silhouette Coefficient: 0.87
Calinski-Harabasz Index: 3142.51
Davies-Bouldin Index: 0.44
Percentages for 4 clusters:
96.8%, 2.3%, 0.5%, 0.3%

Optimal cluster number
though Elbow Method: 4



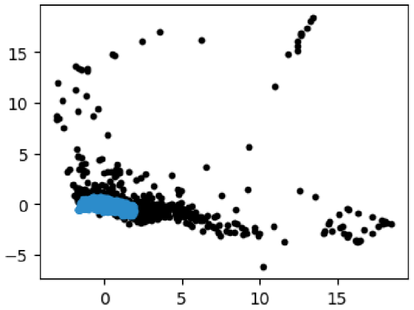
Gaussian Mixture

Silhouette Coefficient: 0.73
Calinski-Harabasz Index: 2444.32
Davies-Bouldin Index: 0.95
Percentages for 4 clusters:
57.2%, 34.9%, 7.0%, 0.9%



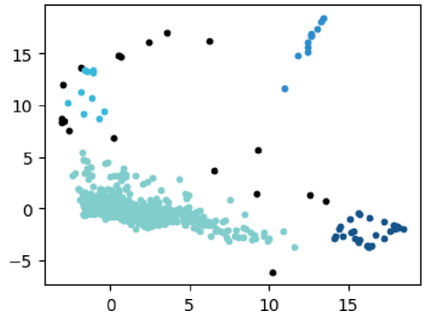
Mean Shft

Silhouette Coefficient: 0.87
Calinski-Harabasz Index: 2866.56
Davies-Bouldin Index: 0.40
Percentages for 4 clusters:
97.7%, 1.3%, 0.6%, 0.3%



DBSCAN

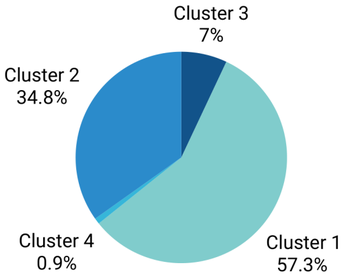
Silhouette Coefficient: 0.71
Calinski-Harabasz Index: 1413.33
Davies-Bouldin Index: 1.25
Percentages for 1 cluster:
88.2%, outliers: 11.8%

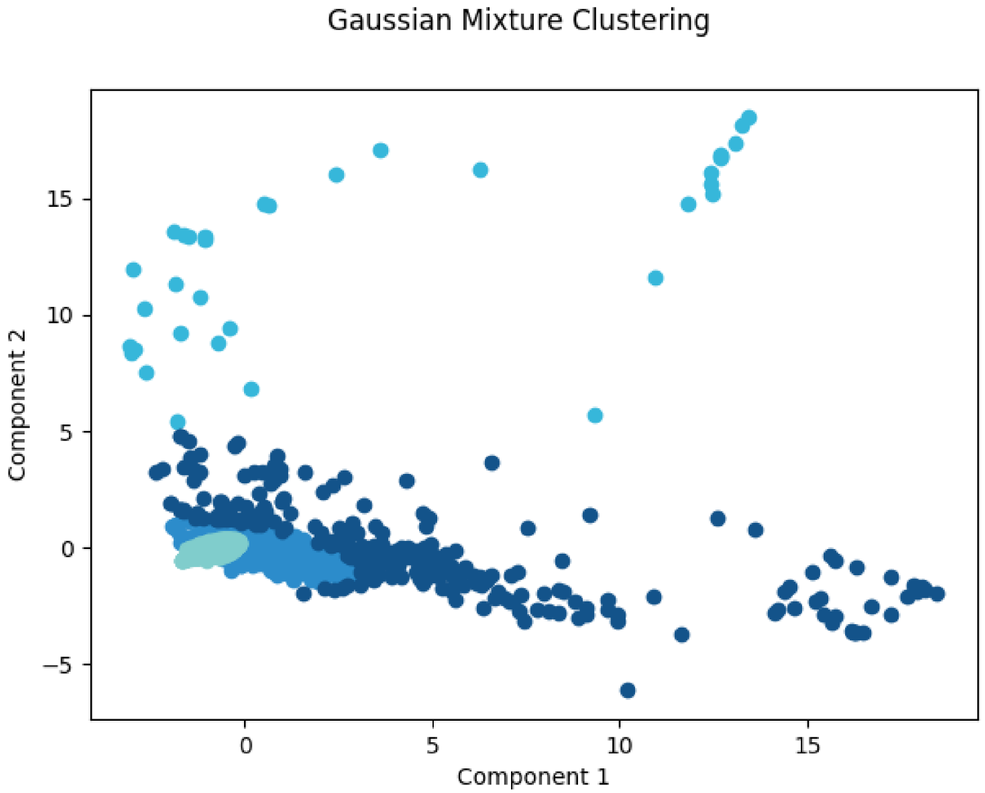


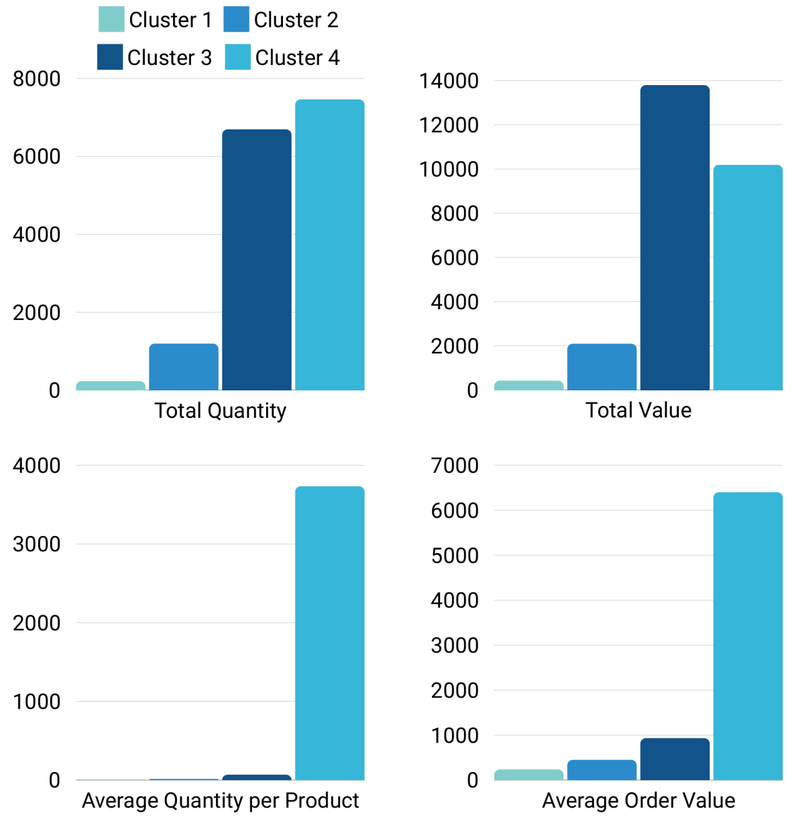
HDBSCAN

Silhouette Coefficient: 0.87
Calinski-Harabasz Index: 1592
Davies-Bouldin Index: 1.15
Percentages for 1 cluster:
98.2%, 0.7%, 0.3%, 0.2% outliers: 0.5%

CLIENT CLUSTERING
CLIENT CLUSTERS







1. CLIENT CLUSTERING

2. INVENTORY FORECAST

Implementing **supervised time series machine learning** model to predict inventory for **top-selling** products, considering **trends** and seasonality.

Performance metrics: RMSE, MAE

INVENTORY FORECAST

FEATURE ENGINEERING & DATA PREP

QUANTITY
OF PRODUCT
SOLD PER DAY

HOLIDAYS

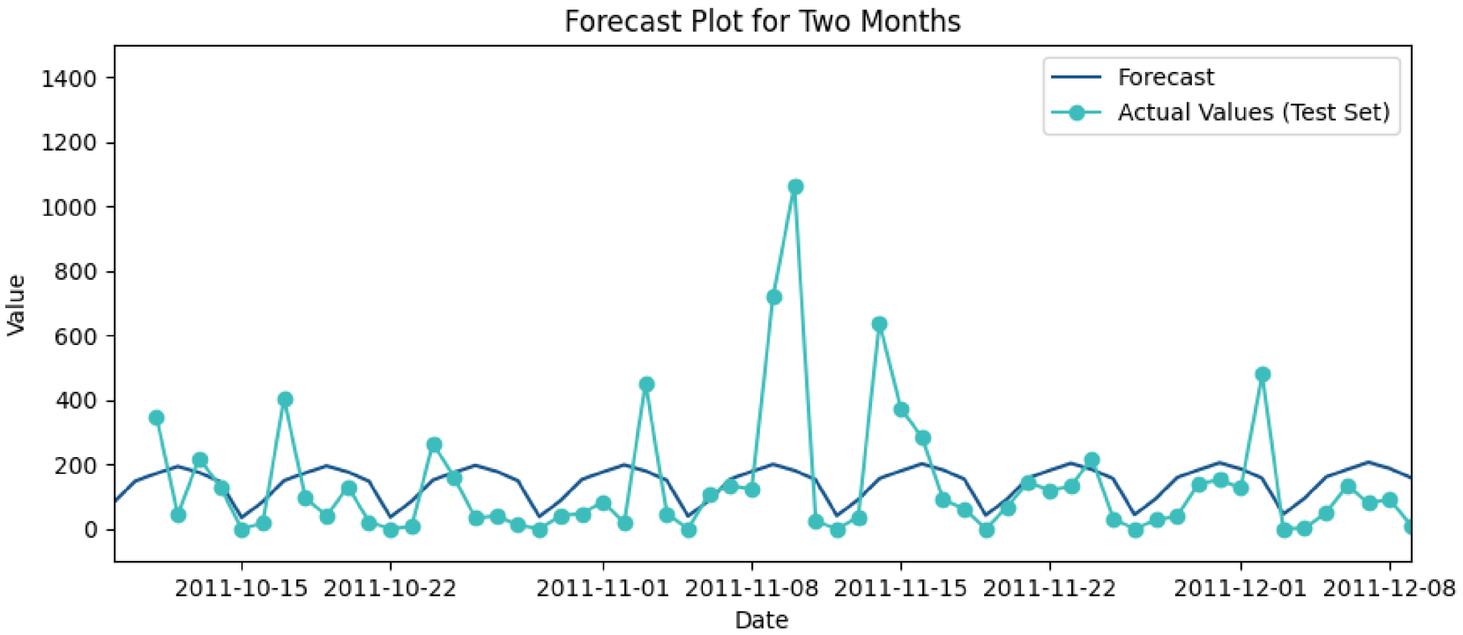
DATE,
EVEN DAYS
WITH 0 ORDERS

TRAIN SET
10 MONTHS

TEST SET
2 MONTHS

*** PREDICTIONS PER PRODUCT**

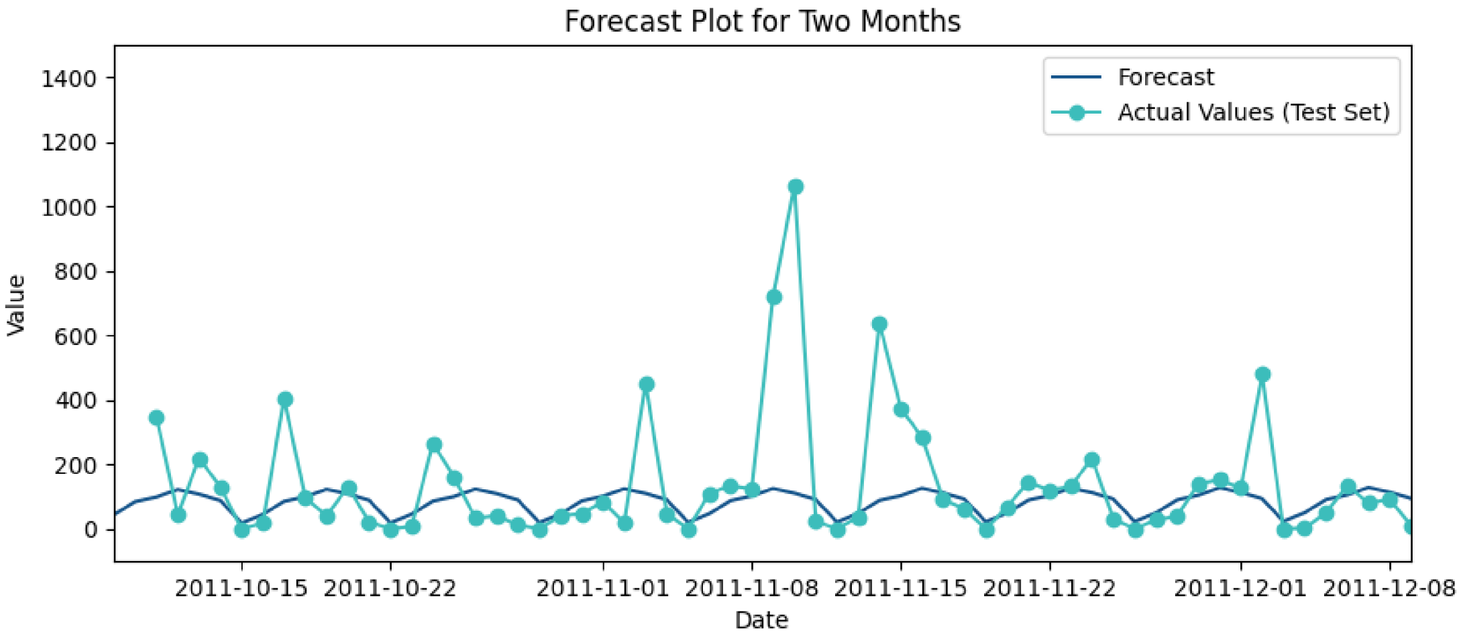
INVENTORY FORECAST
MODEL



Root Mean Squared Error: **182.8239**
Mean Absolute Error: **118.6261**

Prediction for top selling product, Prophet model

INVENTORY FORECAST
MODEL WITH CLIENT CLUSTERING



Root Mean Squared Error: **53.1654**
Mean Absolute Error: **33.5248**

70%

Prediction for top selling product, Prophet model

THANK YOU!