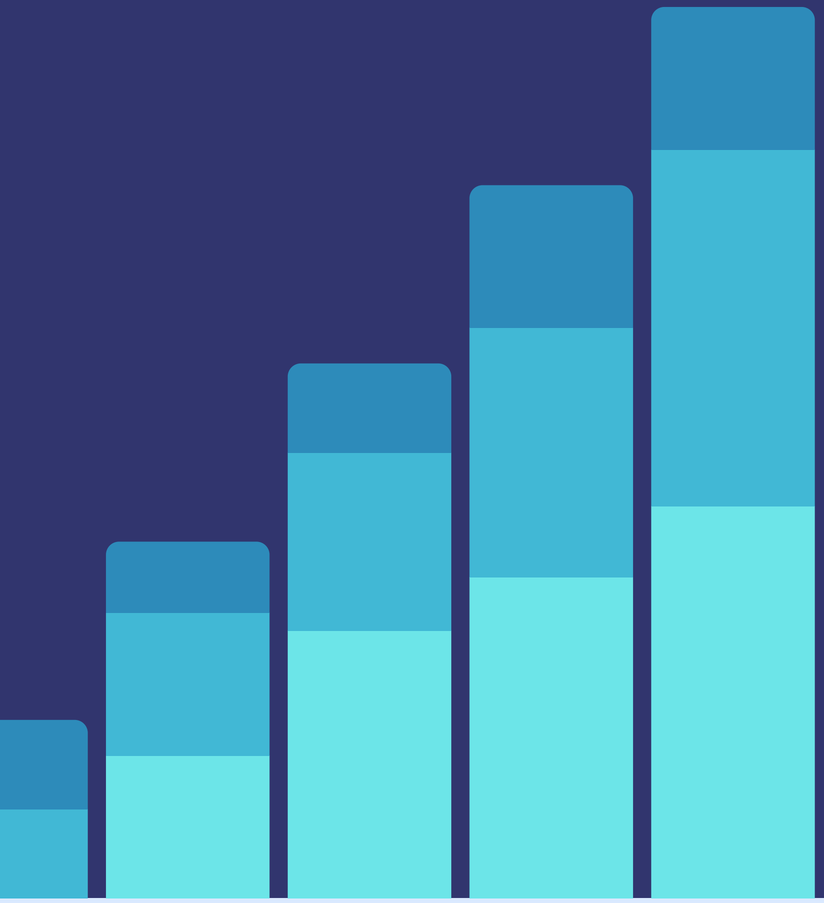
SC1015 mini PROJECT

SERRA ASHAK & DEREN OLGUN

PROBLEM DEFINITION: PREDICT MONTHLY SALES



We used Kaggle dataset about online retail.
https://www.kaggle.com/code/ekrembayar/rfm-analysis-online-retail-ii

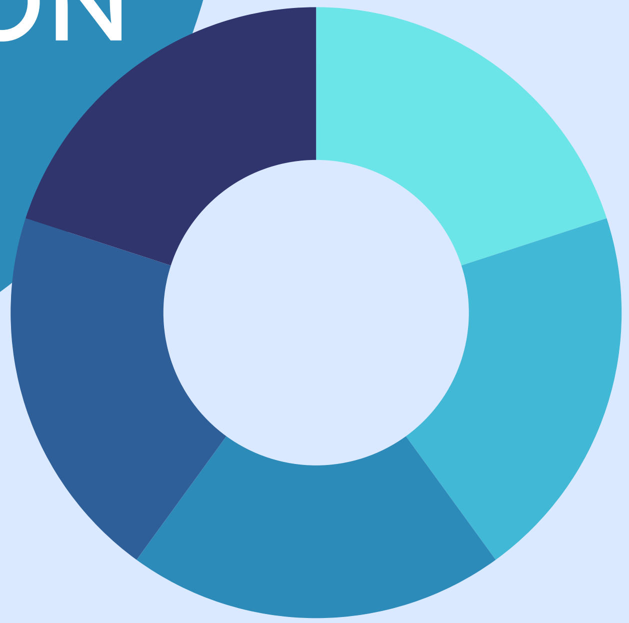
The data includes: Invoice, StockCode, Description, Quantity, InvoiceDate, Price, Customer ID, Country of the sale. And dates between 2009 to 2011.

MOTIVATION

BETTER PLANNING FUTURE

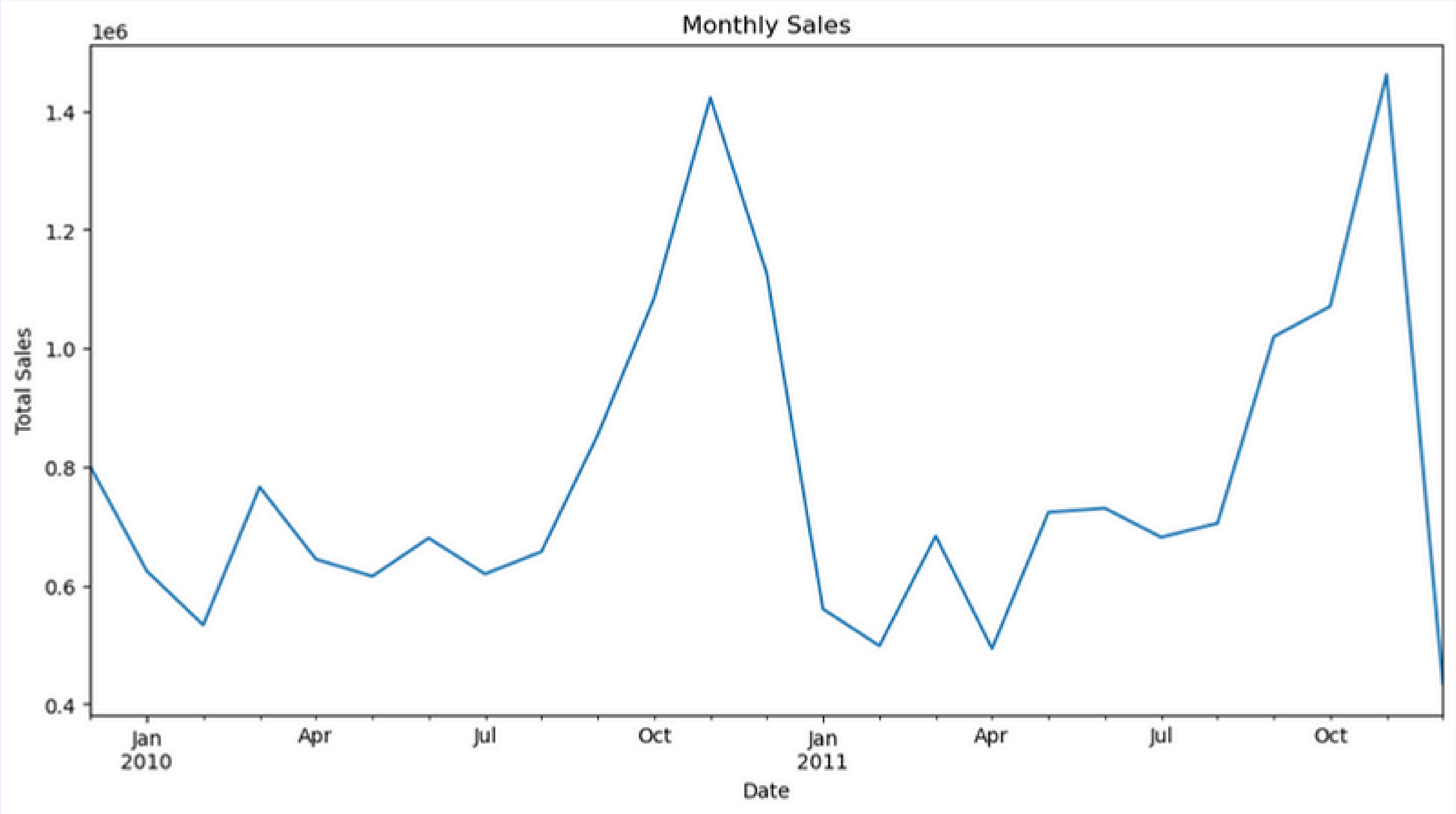
STRATEGIC MARKETING

RESOURCE ALLOCATION & SUPPLY CHAIN OPTIMIZATION



EXPLORE DATA

Total entry in data: 1067365
There are 53622 unique invoice numbers.
There are 5305 unique stock code numbers.



We can see seasonality from monthly sales graph, increased total spend in October and the general pattern look very similar between two years.

PREPARE THE DATA

Handle Missing Data:
- set unkown customer IDs to 0.
- fill missing descriptions from same stock codes.

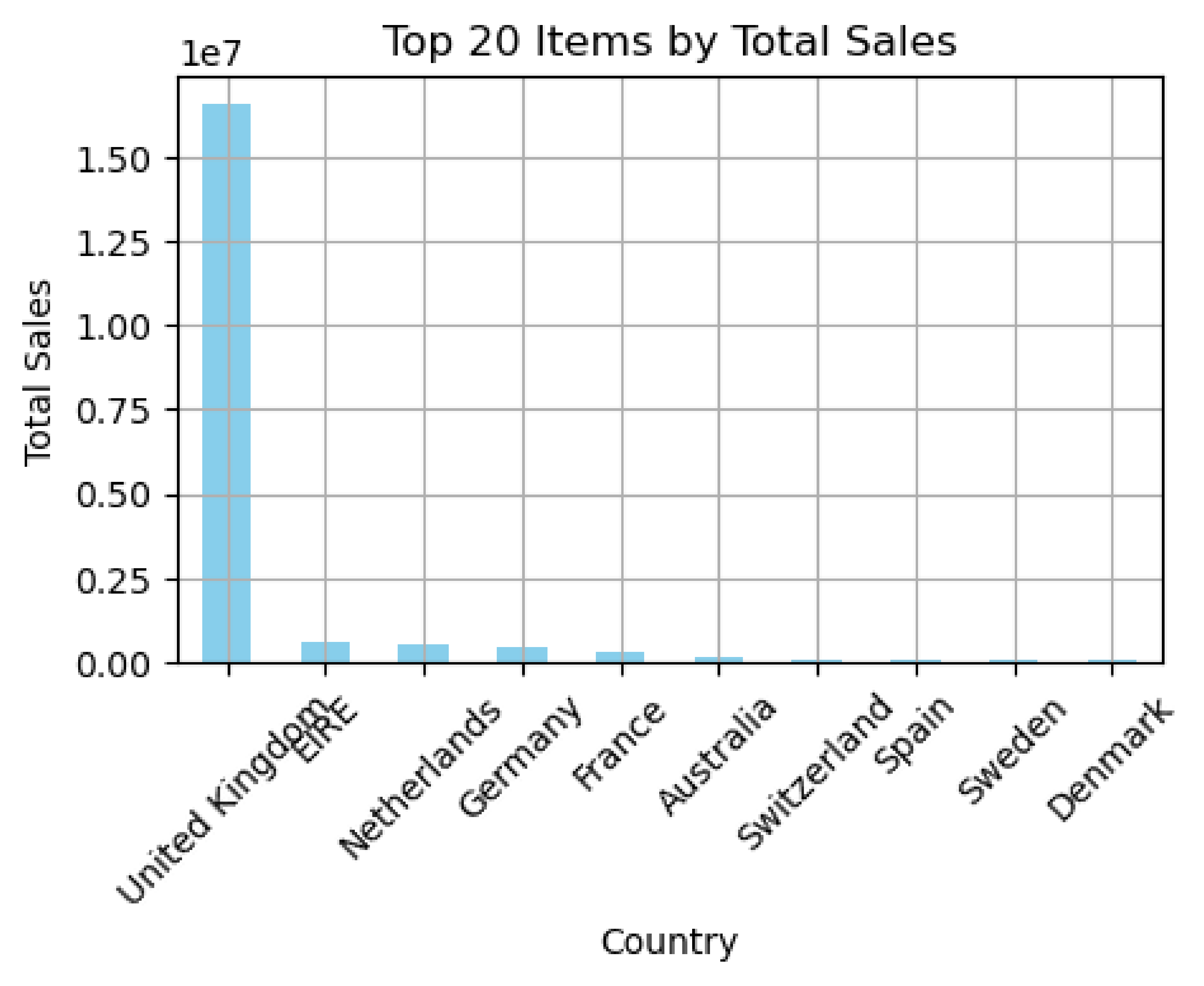
Outlier Handling:
- remove outliers below/above some treshold

Feature Engineering:
- create time features from InvoiceDate.
- like year, month, day of week

COUNTRY TOTAL SPENDS

Most of the sales already from UK, not indicative.

Will not use “Country” in the model.

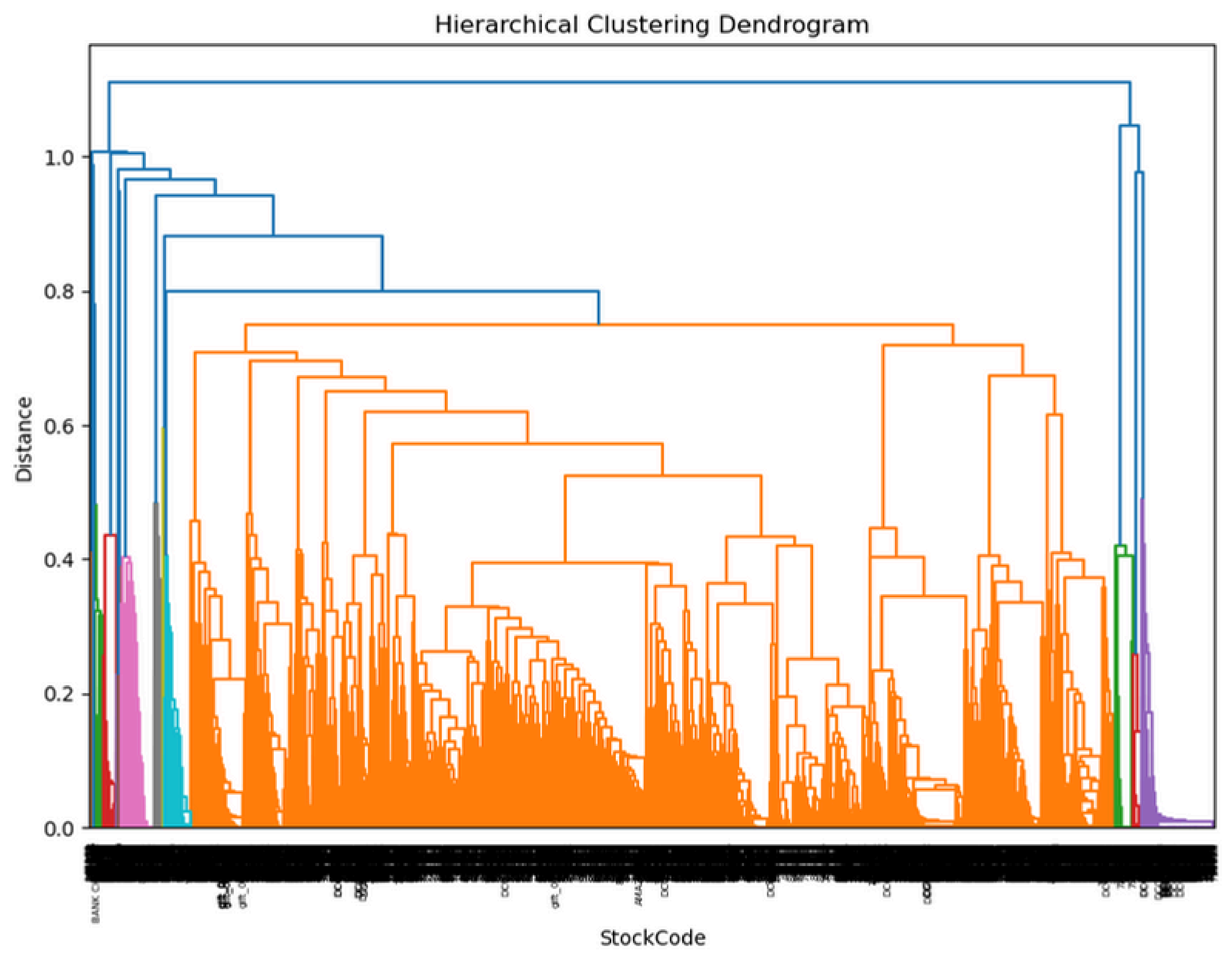


Categorizing Products Based on Trends

- Since each product may follow a distinct sales trend, we analyzed their individual seasonal trends instead of aggregating them cumulatively.

- Given that there are over 5,000 unique product types, products with similar trends were categorized together.

- We wanted to do this clustering because we think looking different products as a whole discard the sales pattern of indiviudal products. One product selling higher in summer cancel out the product selling low in summer for example.

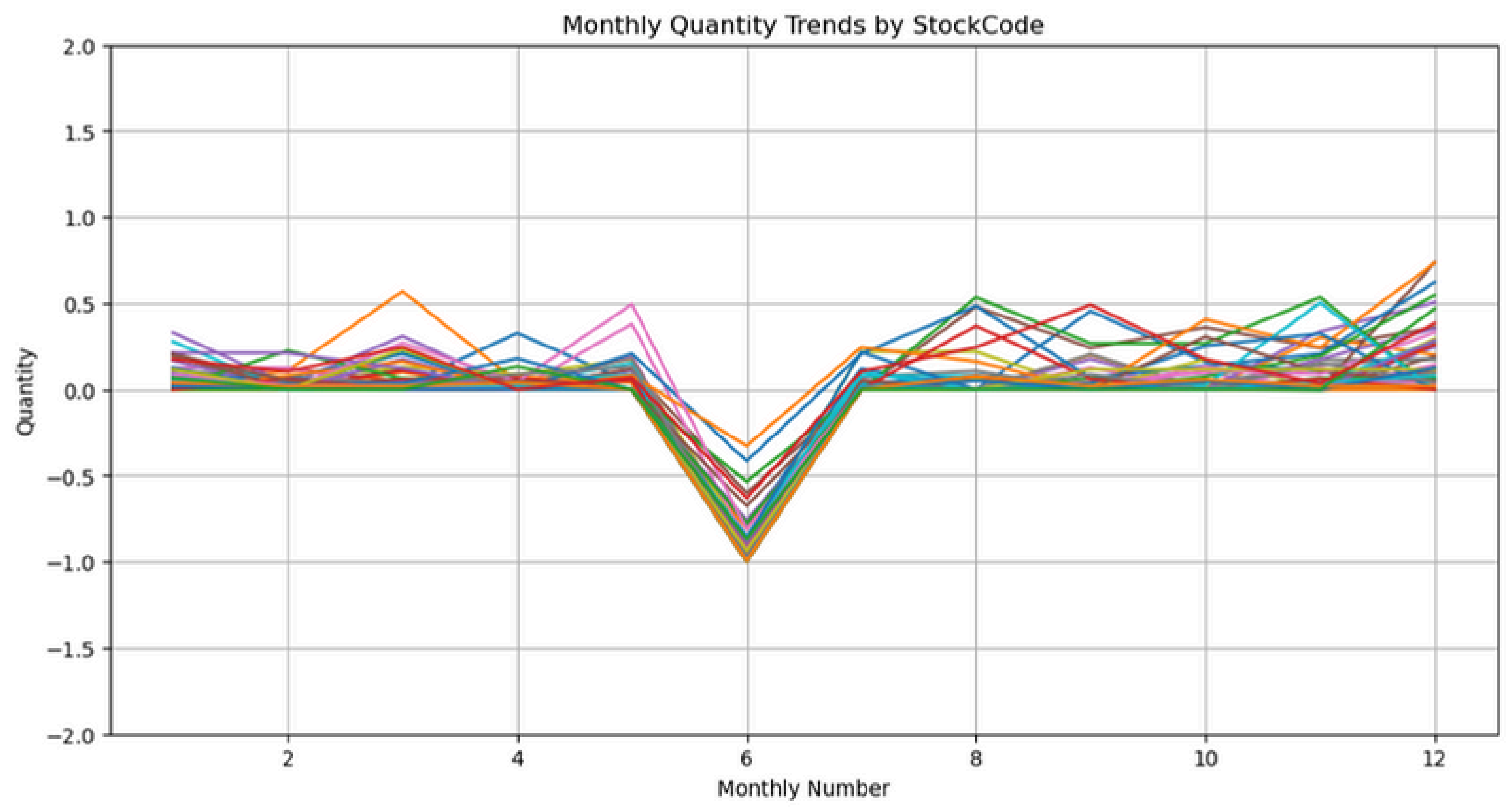


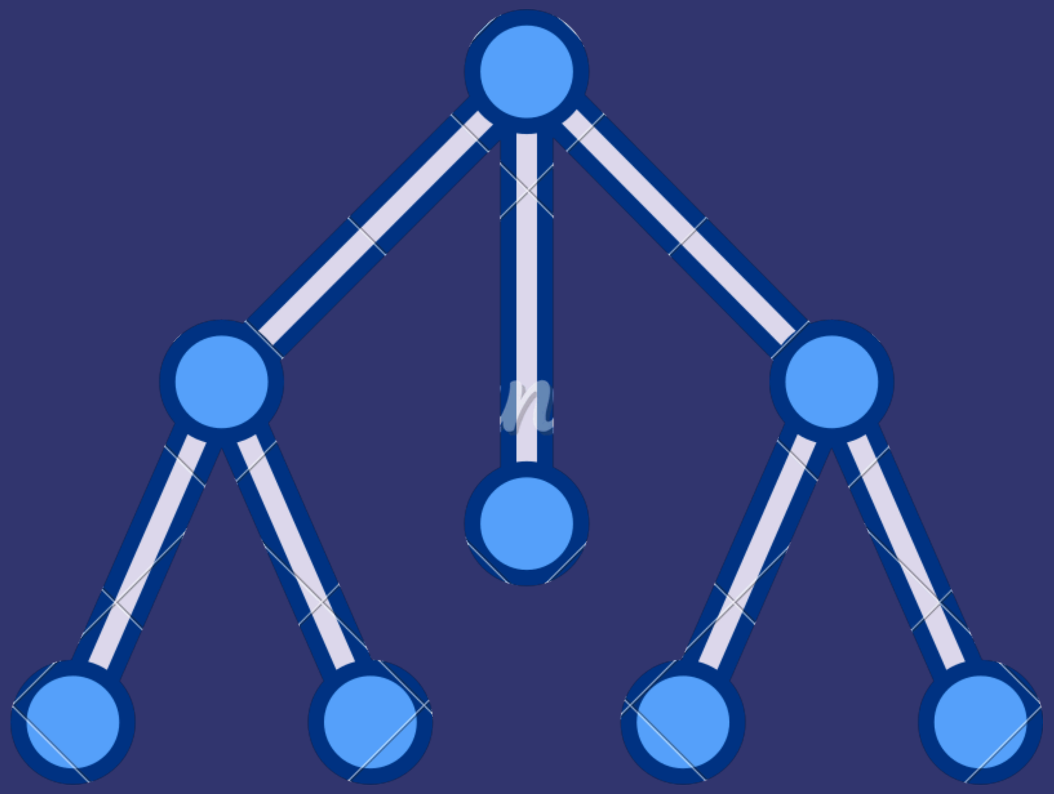
TREND CATEGORIZATION

- Cosine Similarity was utilized to identify similar products.
- For segmentation, a hierarchical clustering method was applied.
- As a result, products were grouped into 5 distinct clusters.

EXAMPLE GROUP

Visual inspection of group number 2.





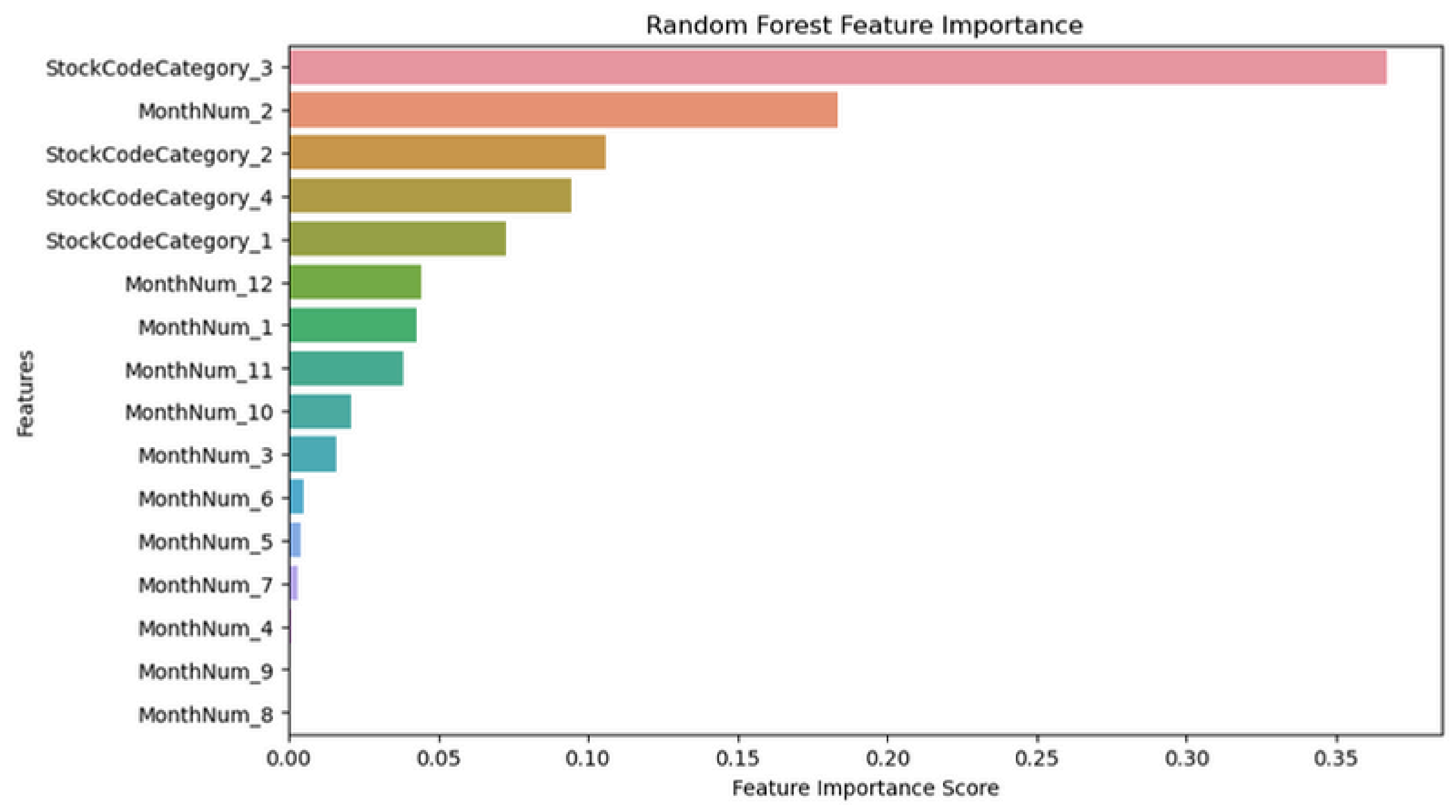
Random Forest Regression

Objective:
To predict the sales quantity for a specific product in a specific month.

Why Random Forest?

- Captures Non-Linear Relationships.
- Handles Categorical Variables.
- Feature Importance Insight.
- Robust to Overfitting by averaging the results of multiple decision trees.

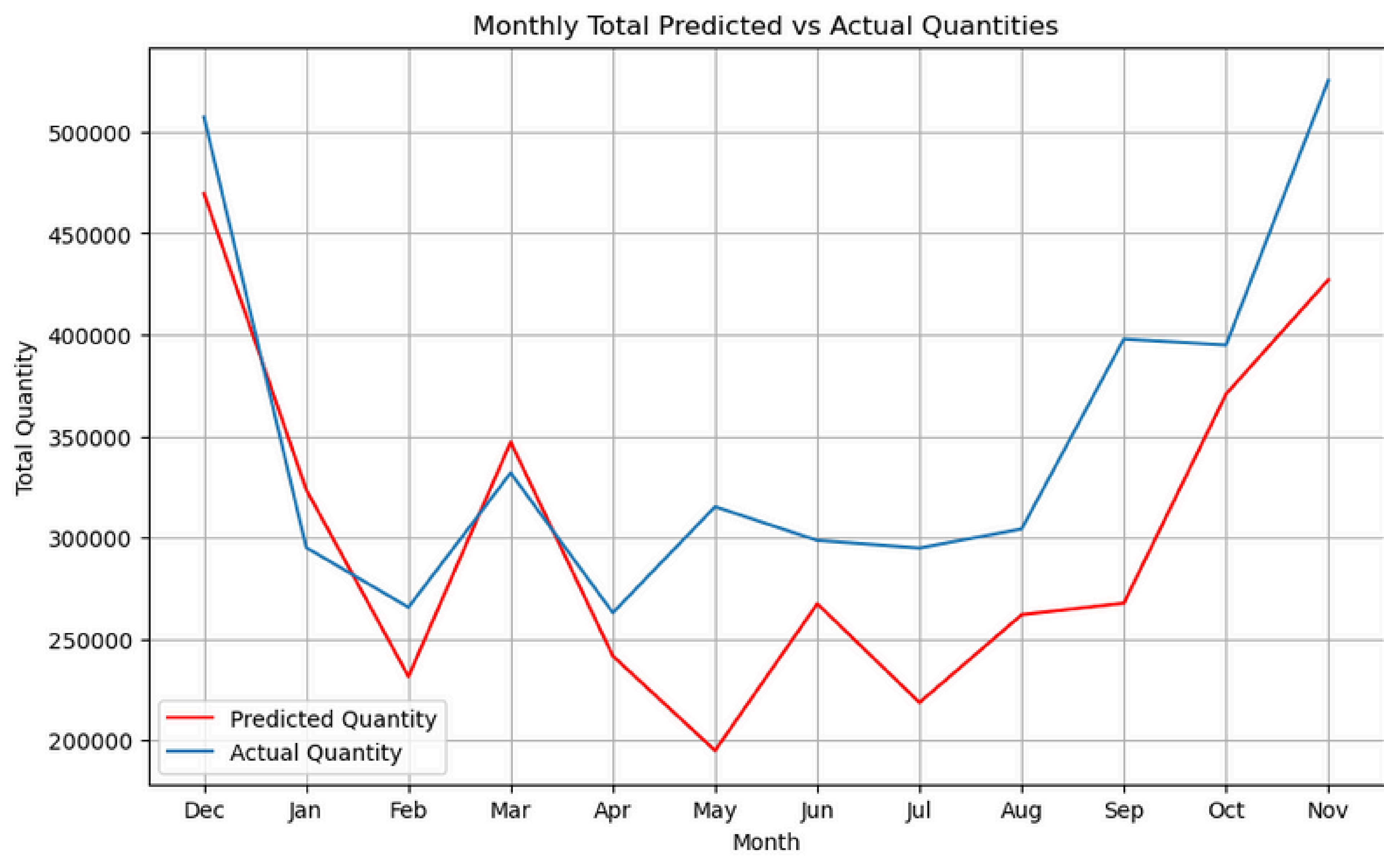
RANDOM FOREST FEATURE IMPORTANCE

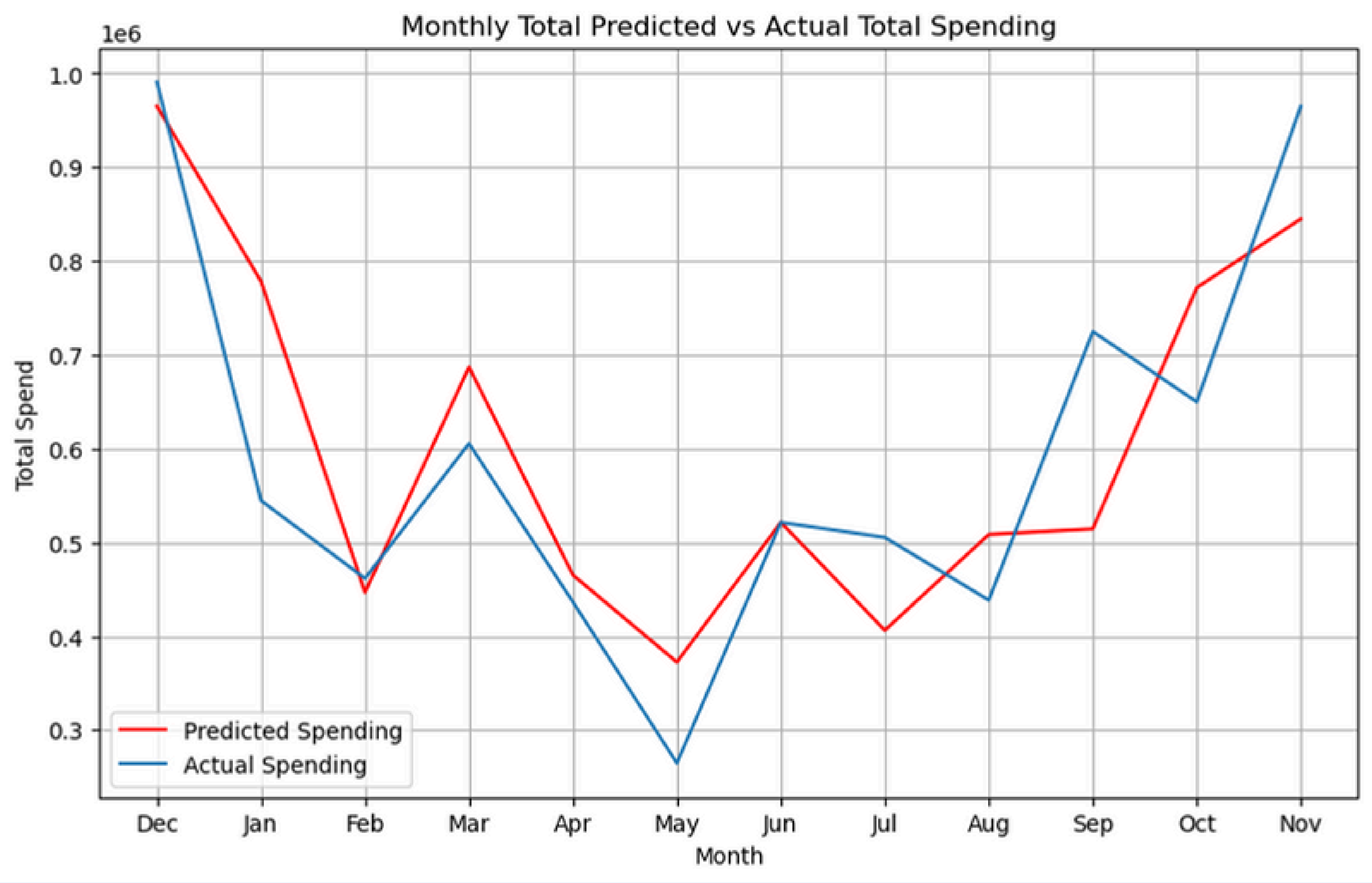


- Products in Category 3 significantly influence sales predictions.
- February stands out as one of the most important features.
- As seen in the earlier monthly sales graph, February have increase in sales, which aligns with its high importance in the model.

COMPARISON OF PREDICTED AND ACTUAL MONTHLY SALES

We observe that the model successfully captures the overall trend in sales quantity.





COMPARISON OF MONETARY SALES VALUE

To estimate the monetary sales value, we can use the predicted sales quantity along with the average monthly product price, allowing us to achieve a close approximation to the actual values.

Outcome

The Random Forest model accurately predicts monthly sales quantities for individual products, capturing the overall trend and seasonality in the data.

Insights

- Some product categories emerged more critical for sales predictions.
- Some specific months' seasonality affect more significant, potential opportunities for targeted campaigns during this period.
- These can be used for tailoring inventory and marketing strategies to match high-demand periods.

RECOMMENDATIONS

Category-Specific Strategy:

- Invest in marketing and supply chain efficiency for Category 3 products, given their significant impact on sales predictions.

Seasonal Planning:

- Focus on February's sales spike with promotions, stock increases, as it is a key driver of sales across multiple categories.

Extend Analysis:

- Analyze external factors (promotions, holidays) that might further explain sales trends.

What We Learned New

- Hierarchical Clustering
- Random forest regressor

THANK YOU!

SERRA ASHAK & DEREN OLGUN