Transaction data insights

FOR DATA FROM 2009 TO 2011

Contents of this Presentation

Understanding the Data Set

What are the Key Dimensions and Facts

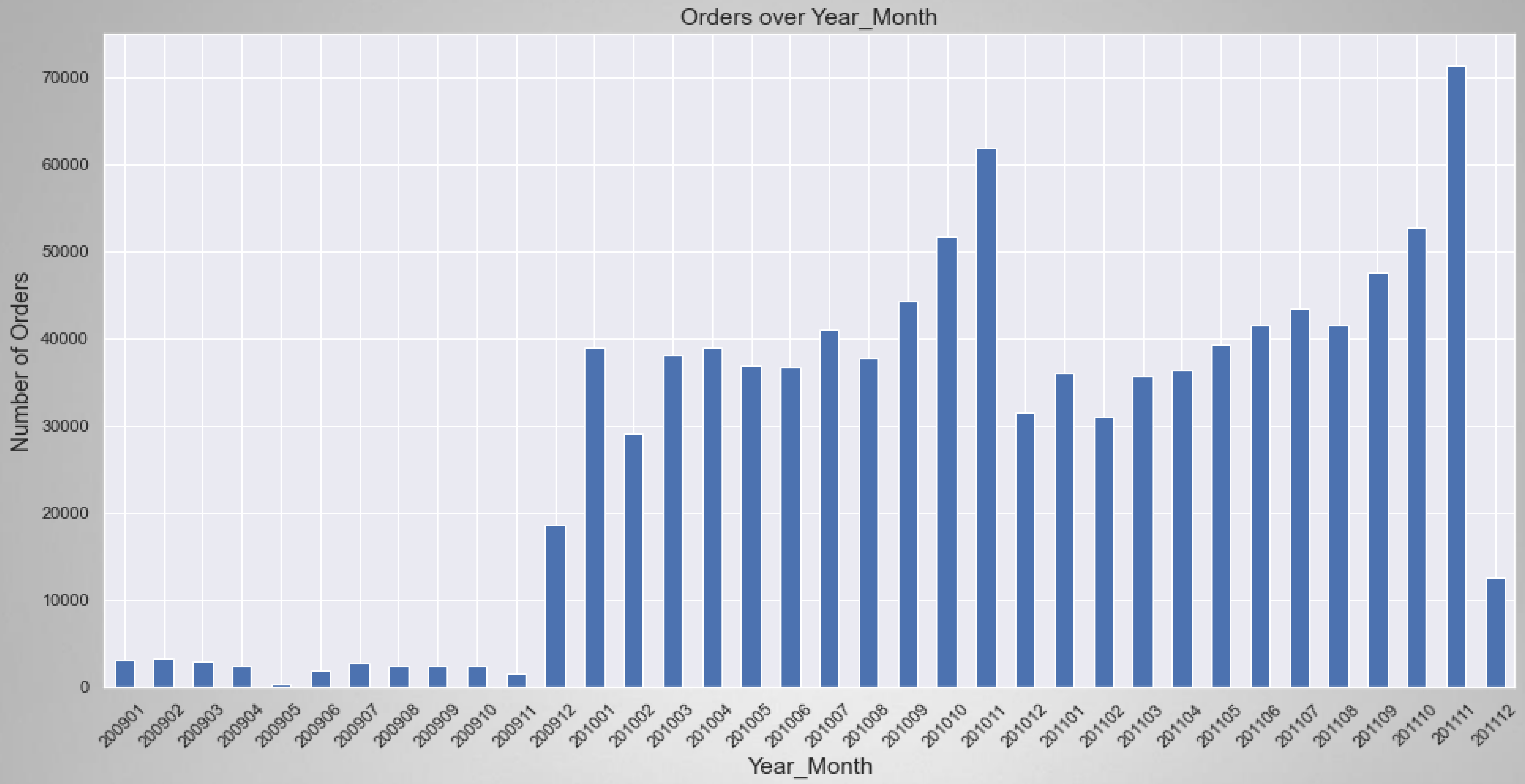
The Data Set contains Transaction / Ledger details for a retail company

- ~1M records from Jan-2009 to Dec-2011
- In essence the Data is at an Invoice-Item Level
- The following Datapoints are available :
    - Invoice : ID , Date , Order Status
    - Product : ID , Quantity , Price , Description
    - Customer : ID , Country

- There are significant Data Integrity issues , which have been detailed in Appendix 1 : Data Integrity

Transactions over Months

From Jan-2009 to Dec-2011



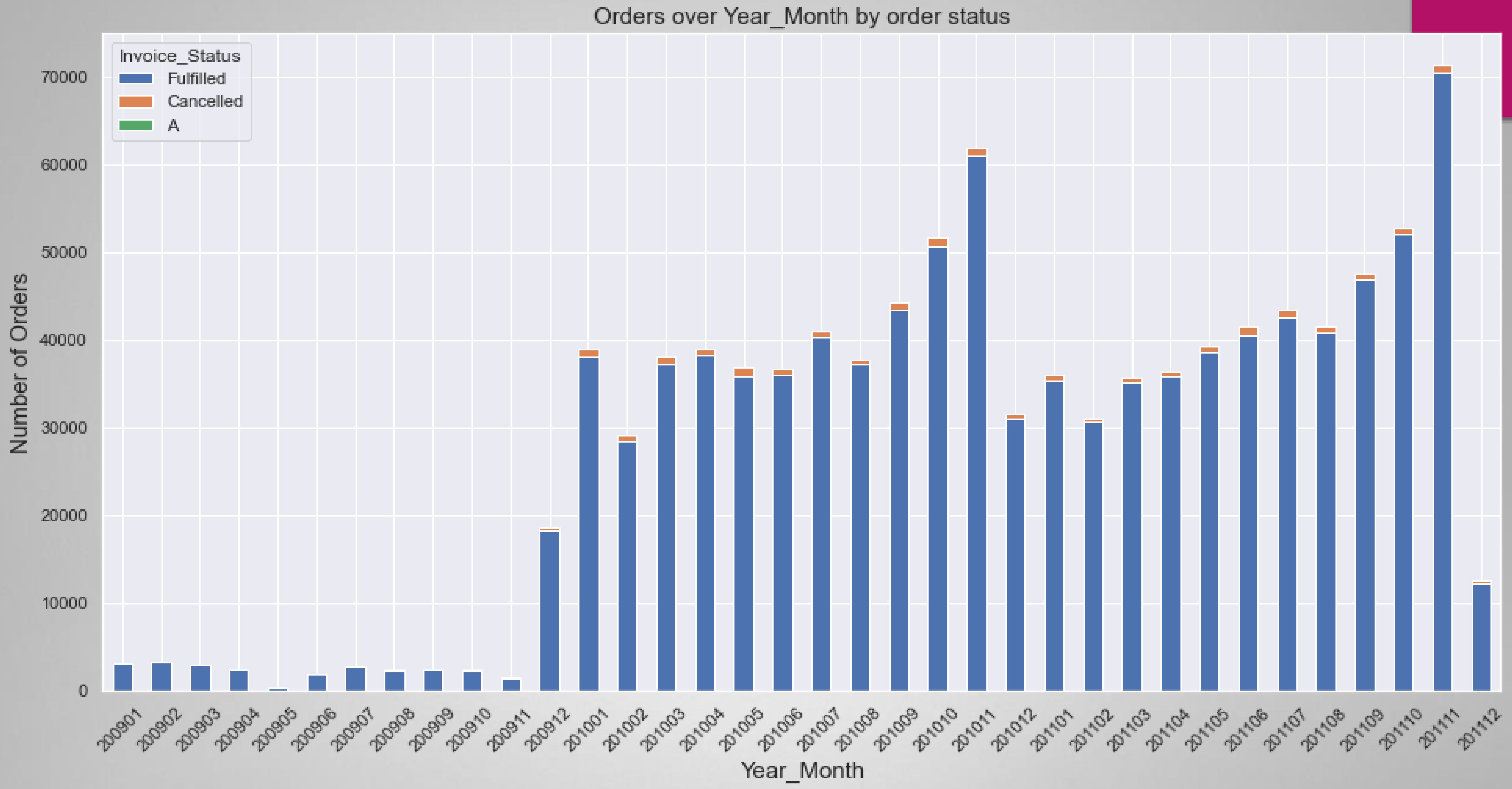
Transactions in 2009 have been very minimal . These Transactions are not uniformly distributed over all weeks

These intermittent Transactions could be because of incomplete data or intermittent business operations

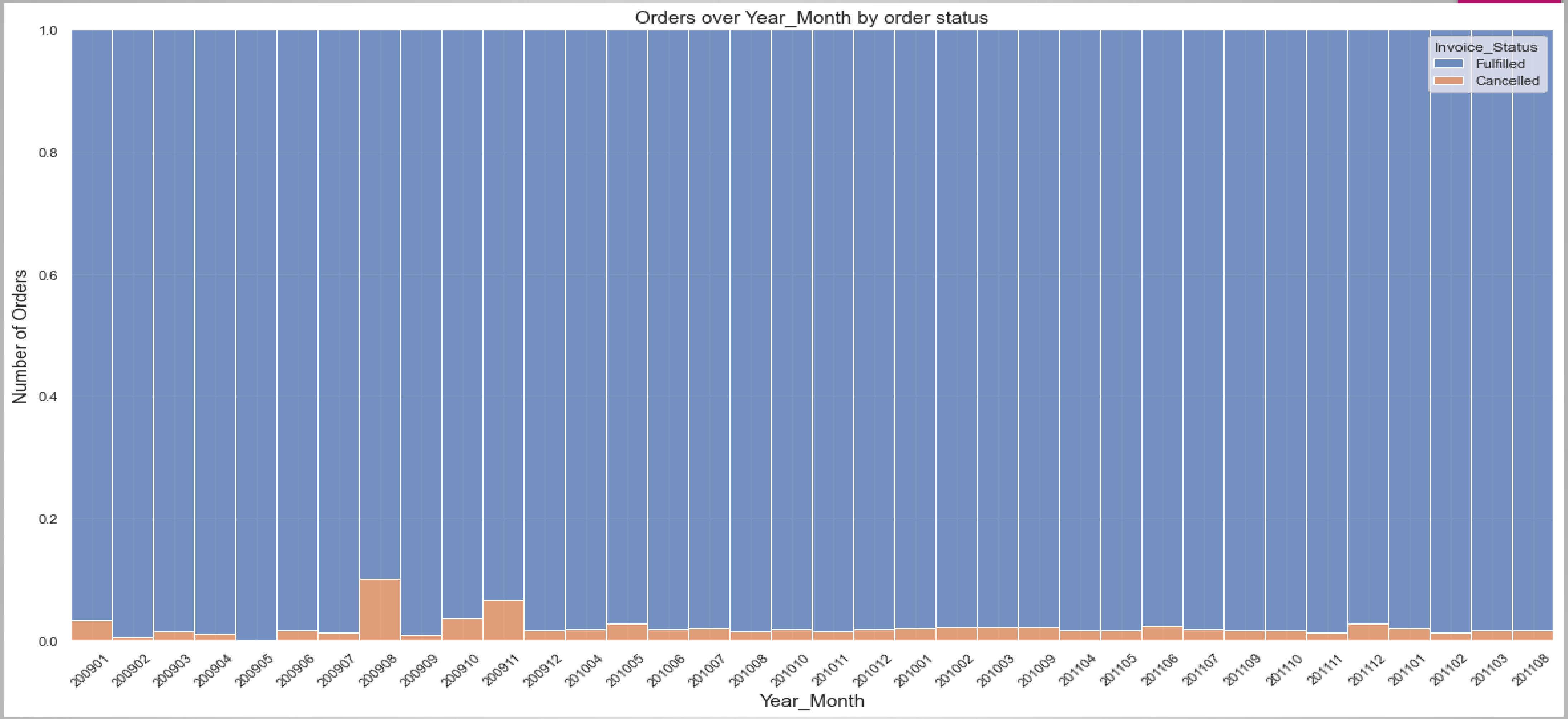
Sales Pickup Near 2009 EOY , before the Christmas/NY period .

A similar seasonal uplift in sales is again observed in 2010 and 2011 in the same Oct/Nov Time period .

A consistent seasonal dip in Sales is observed in Dec after the yearly High , this Dip then recovers over the rest of the Year



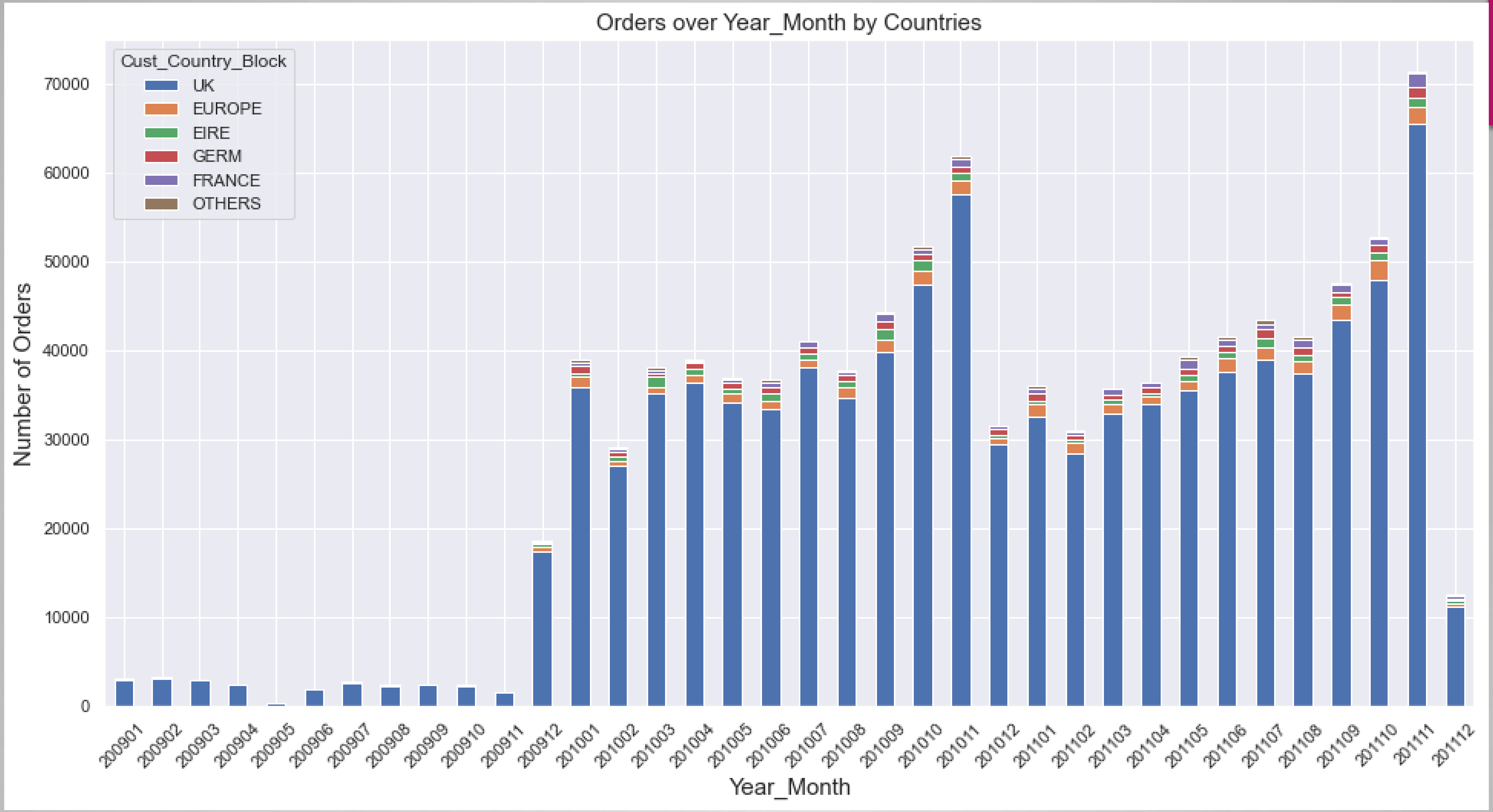
Order Cancellations are usually a minority at < 2% ,
Cancelled Orders also coincide a lot with the datapoints that have poor quality , this is detailed in the Appendix 1



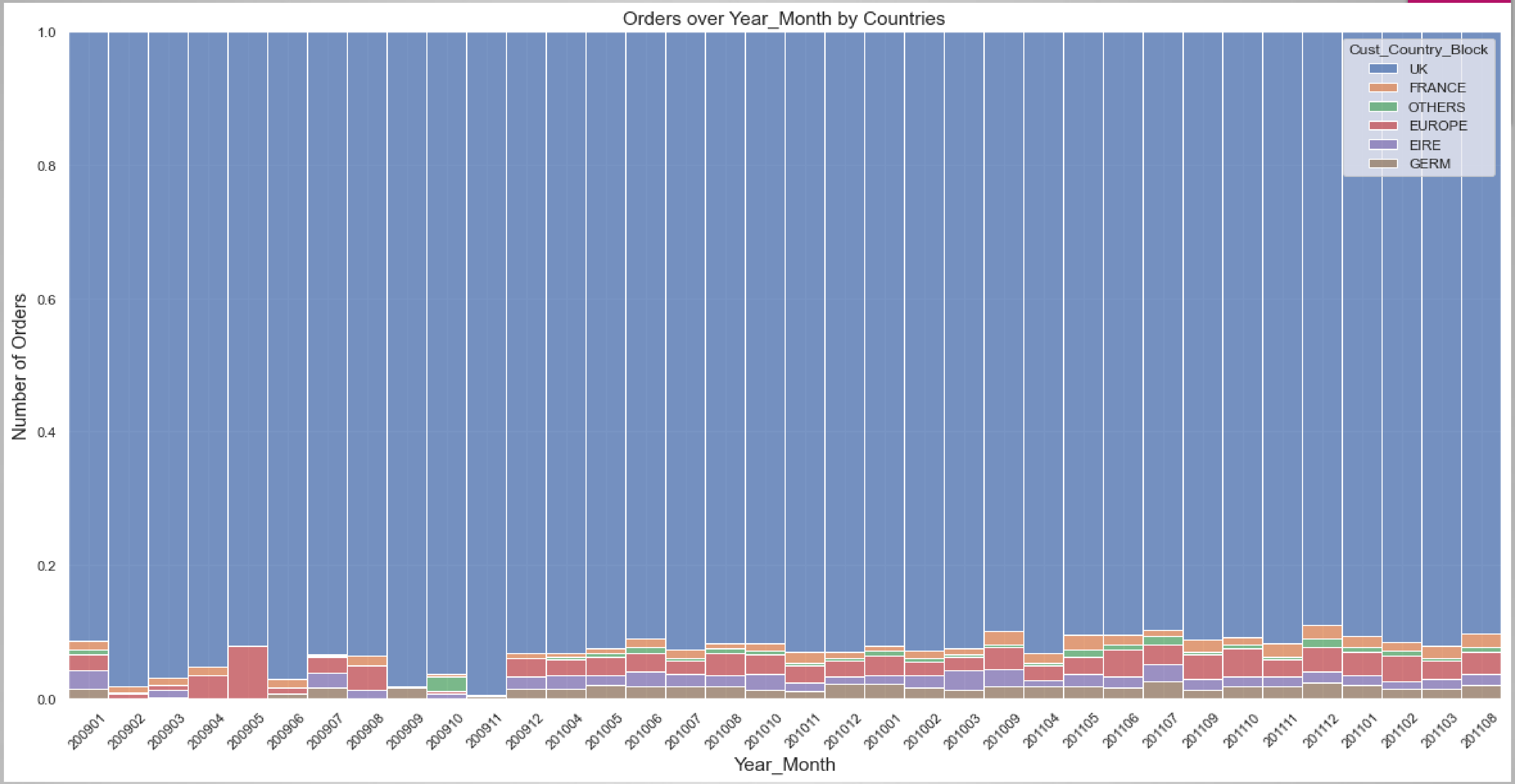
Order Cancellations (Orange) normalized share :

The last time Cancelled Orders peaked was in 2009 Aug and Nov , the data is sparse there

and it is difficult to conclude the reason for this life in Cancellation , without a deeper look and situational information



Majority of the Transactions (92 %) and Sales Volume (85% , $16M) is attributed to UK ,
UK is probably the Domestic Market for this company .

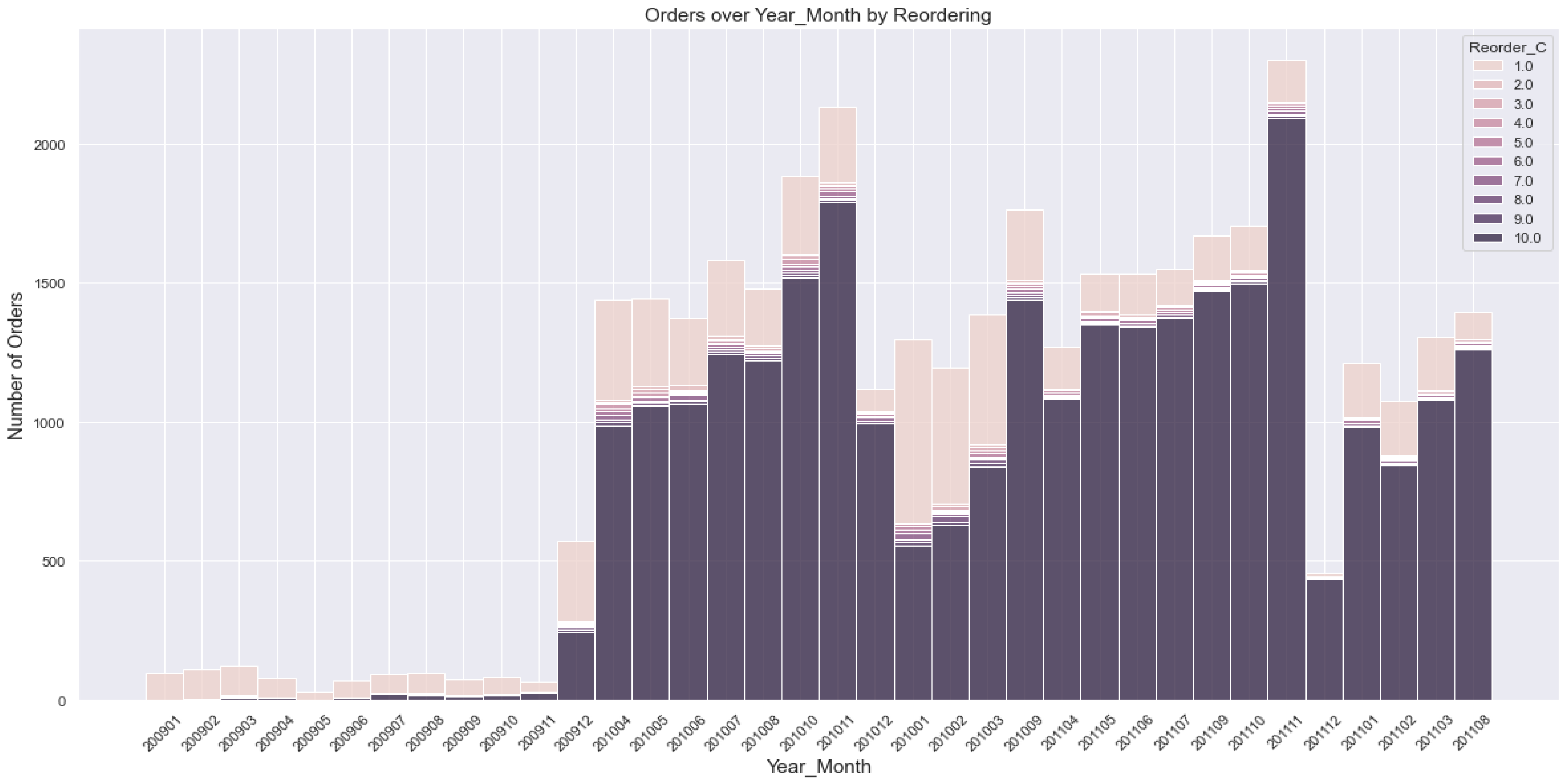


Domestic vs International Share has not changed much since 2009 ,
With France and the Rest of Europe having similar Trn. Volumes

Investigating Customer Loyalty

Order –> ReOrder OR DropOut

- re-ordering can be observed at 2 levels
- one is a Customer Level - where a customer revisits the store ,
  - regardless of whether he buys the same items or not
- the other scenario to study is when customers return to buy the same item multiple times

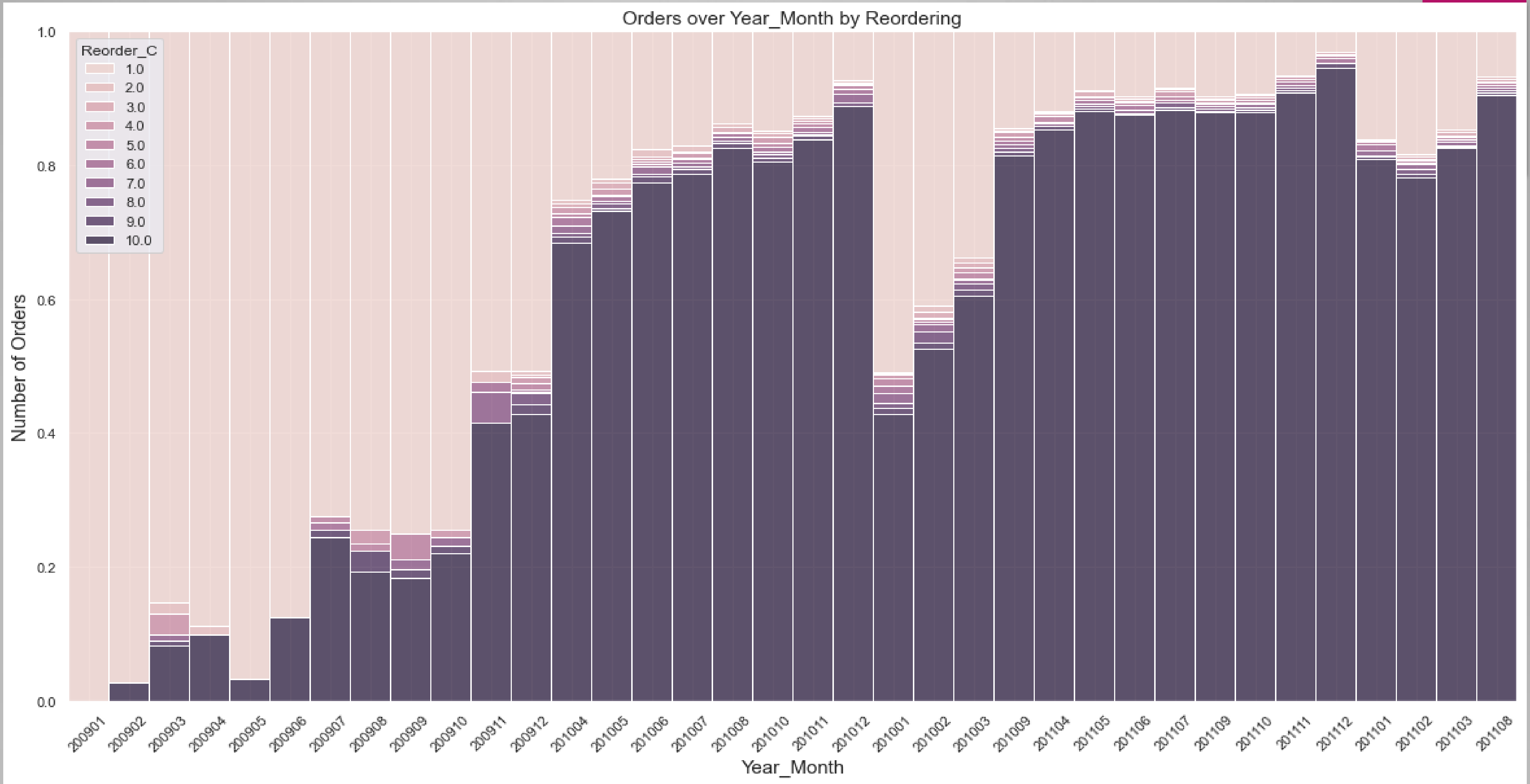


Case 1 : Customer Returns to buy different or even same Items

Peach : New Customers ; Dark Purple : Customers who have placed >= 10 orders with company previously

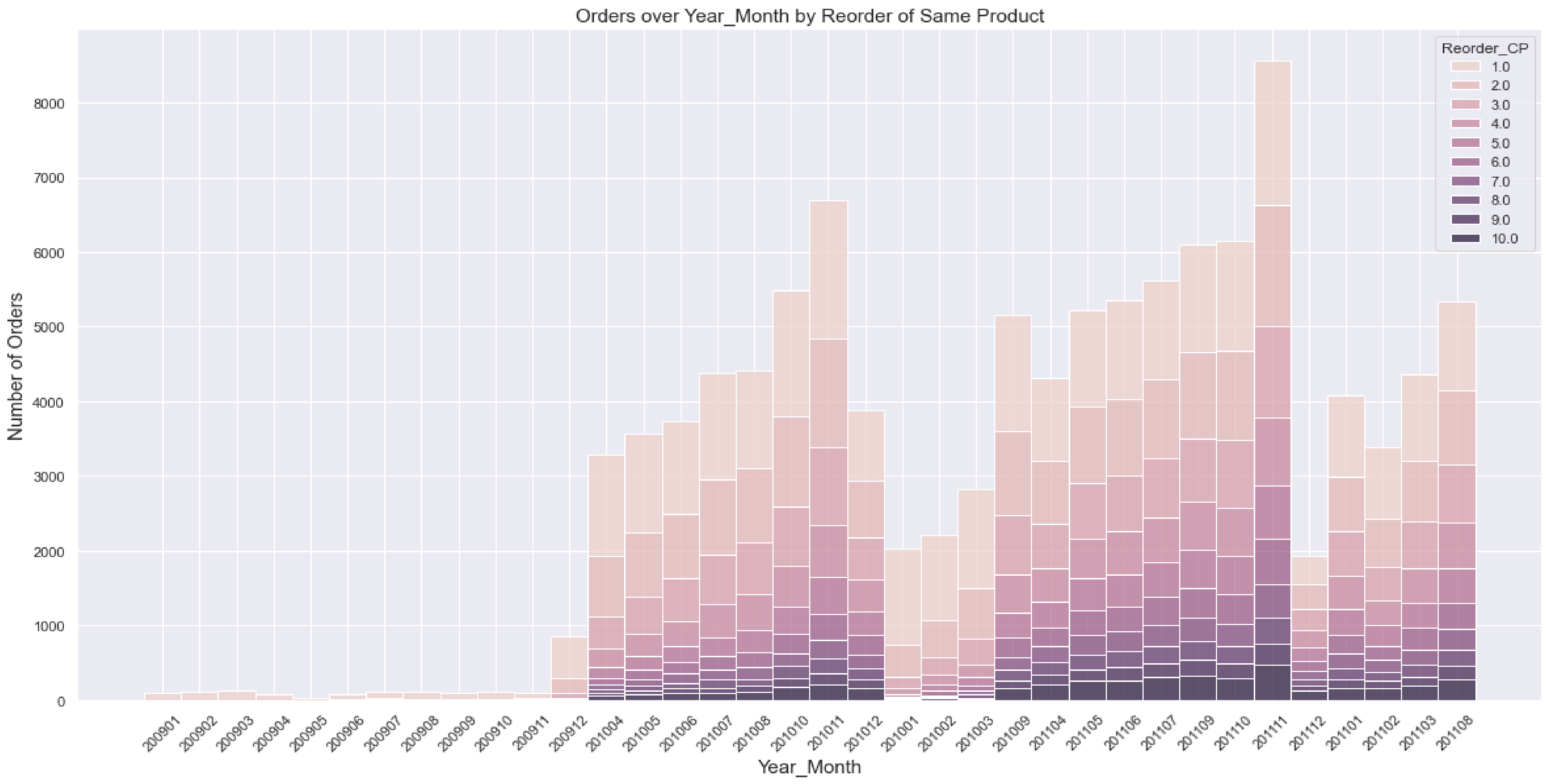
Majority of the Customers are Repeat Buyers ,

Which indicates that the company might be a B2B Supplier who has a small but loyal , recurring customer base



Case 1 : Normalized view : Share of Returning Customers

Except New-Years and Christmas , the Customer Loyalty remains high ,
Dip near NY , might be due to influx of new customers gained from competitors , and some customers lost to competitors

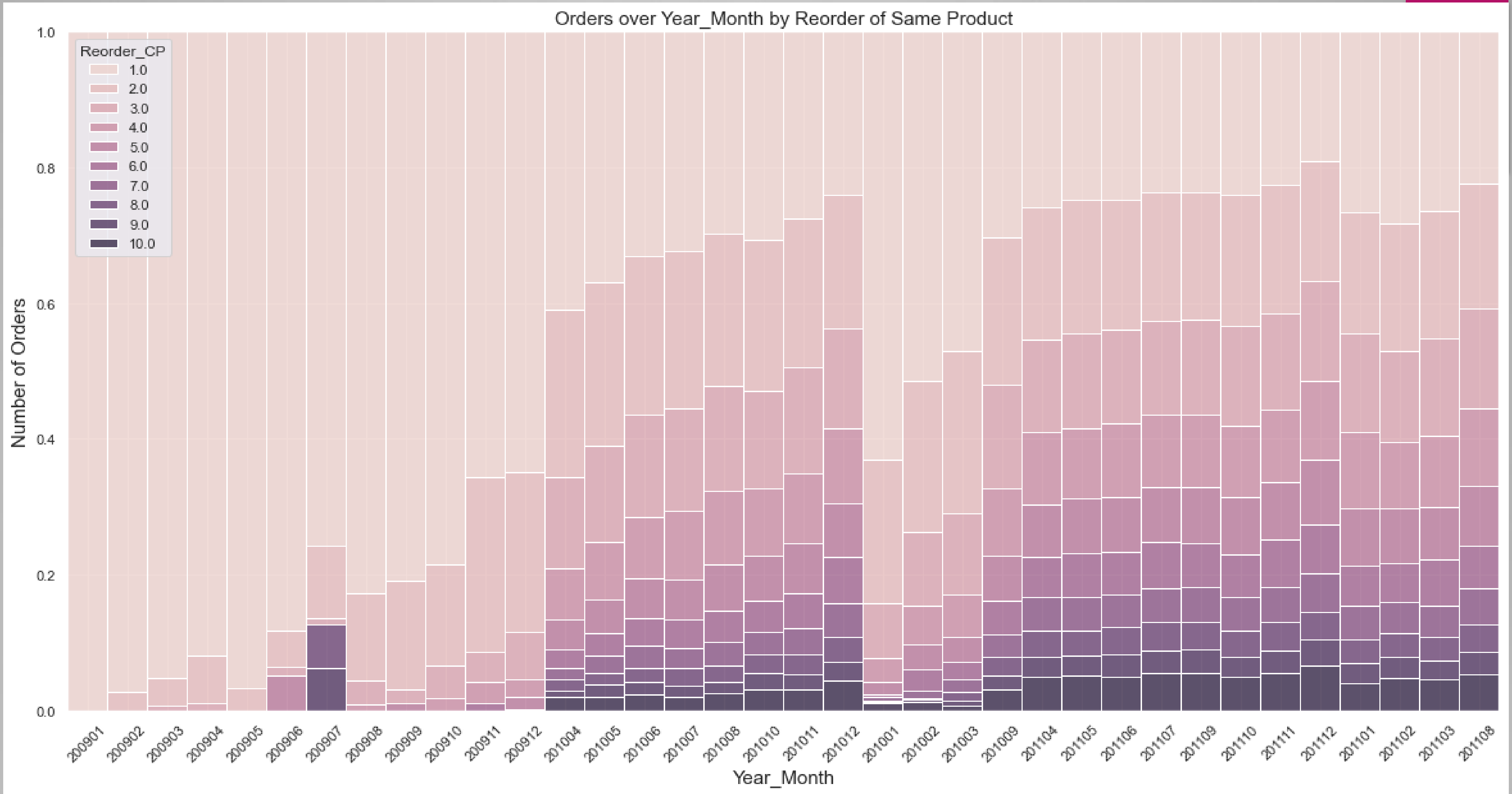


Case 2 : Customer Returns to buy the same Item again

Peach : New Customers ; Dark Purple : Customers who have placed >= 10 orders with company previously

Majority of the Customers are Repeat Buyers for the same Item ,

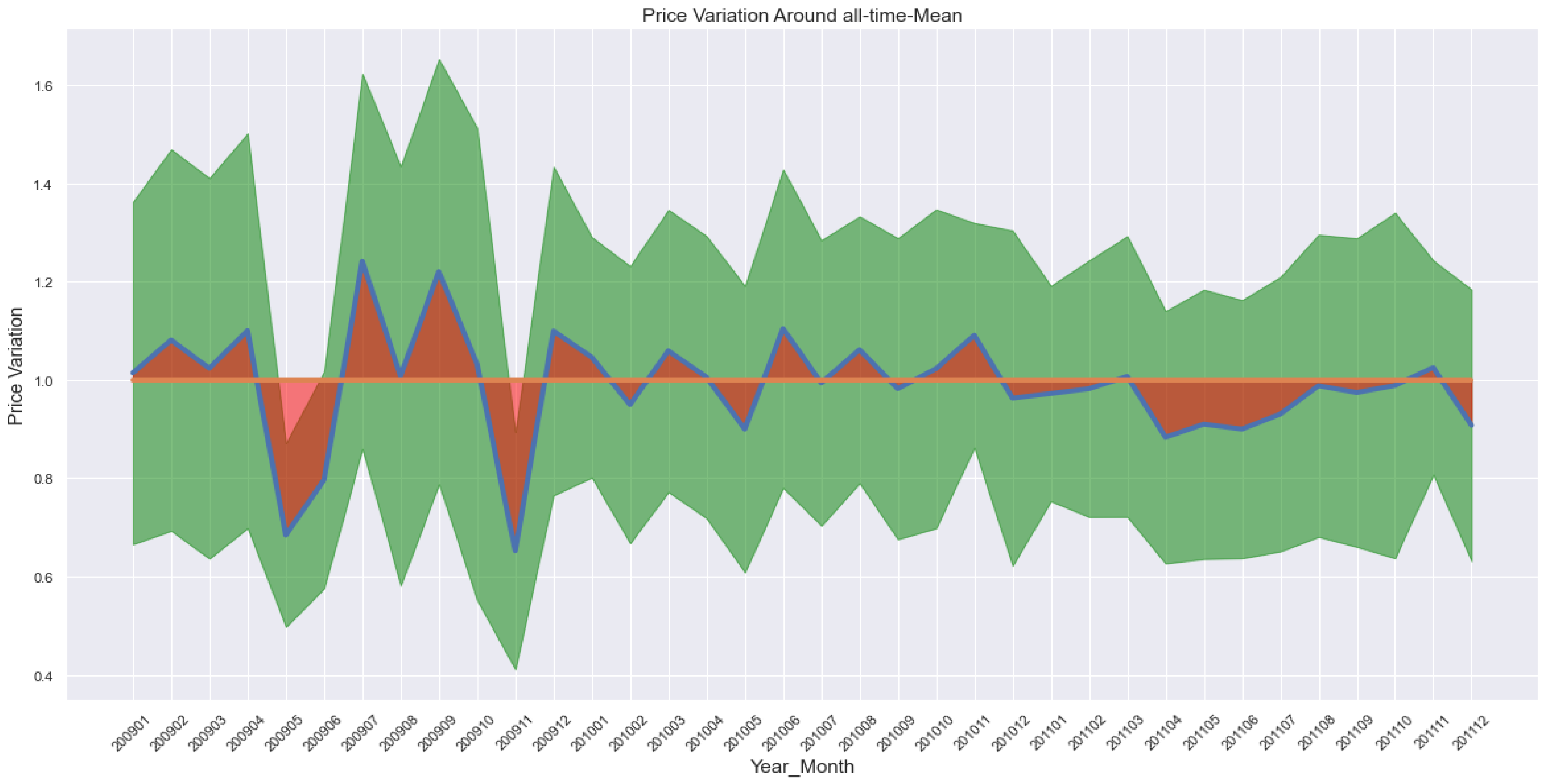
Which indicates that the company might be a selling an AYR/365 Item



Case 2 : Customer Returns to buy the same Item again

Except New-Years and Christmas , the Customer Loyalty remains high ,
Customer share of New vs. repeating Stabilize over time , **indicating lesser customer churn**

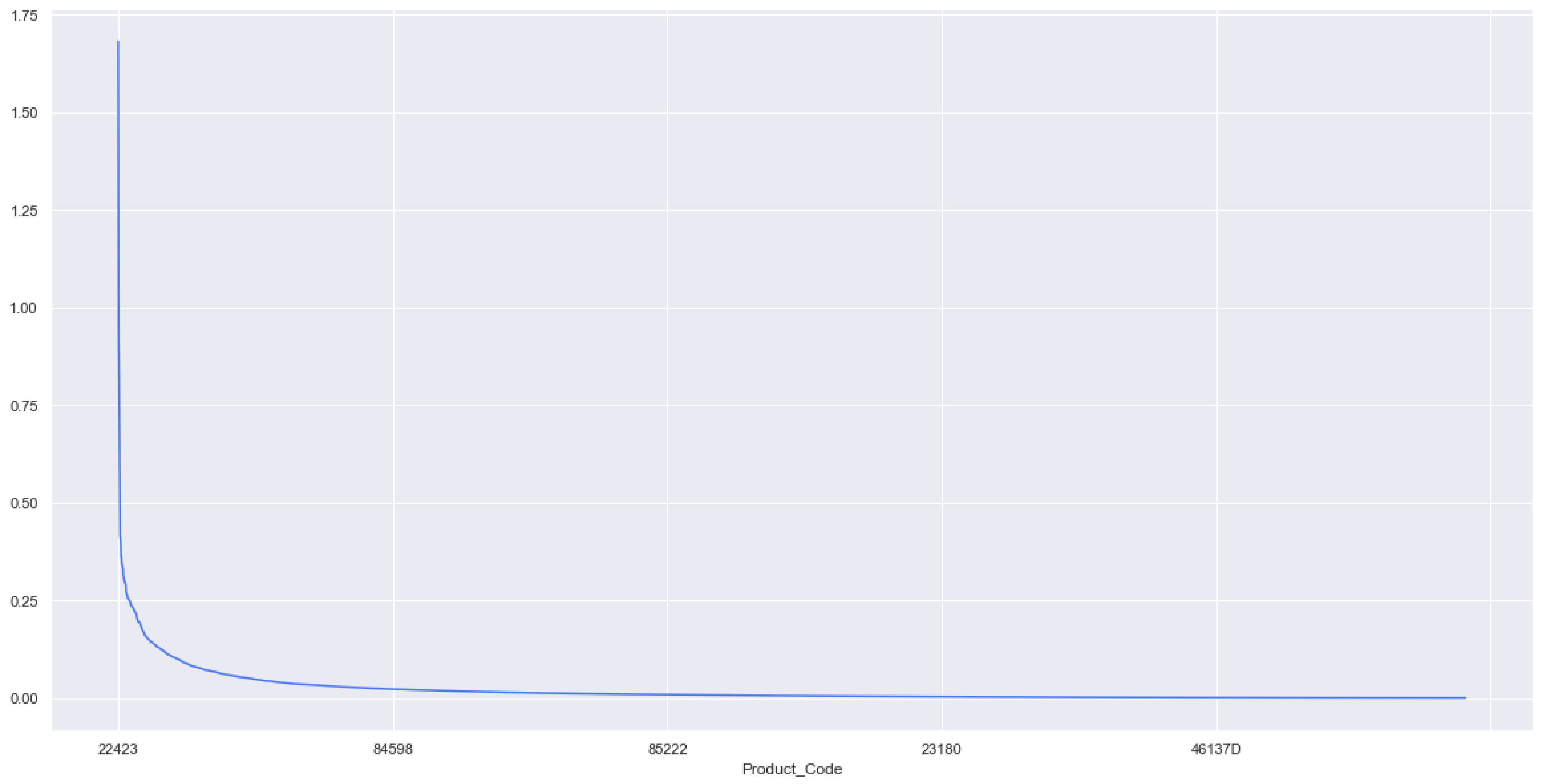
Product and Price Dimension



Mean-Normalized Price variation over Time :
The Product Price has a trend to decrease over time , indicating Discounts . MarkDowns
The Product Price was very Volatile in 2009 , this could be because of Unreliable Data and Intermittent Transactions

Red : Difference from Historical Mean Price at Product level , **Green** : Standard Deviation

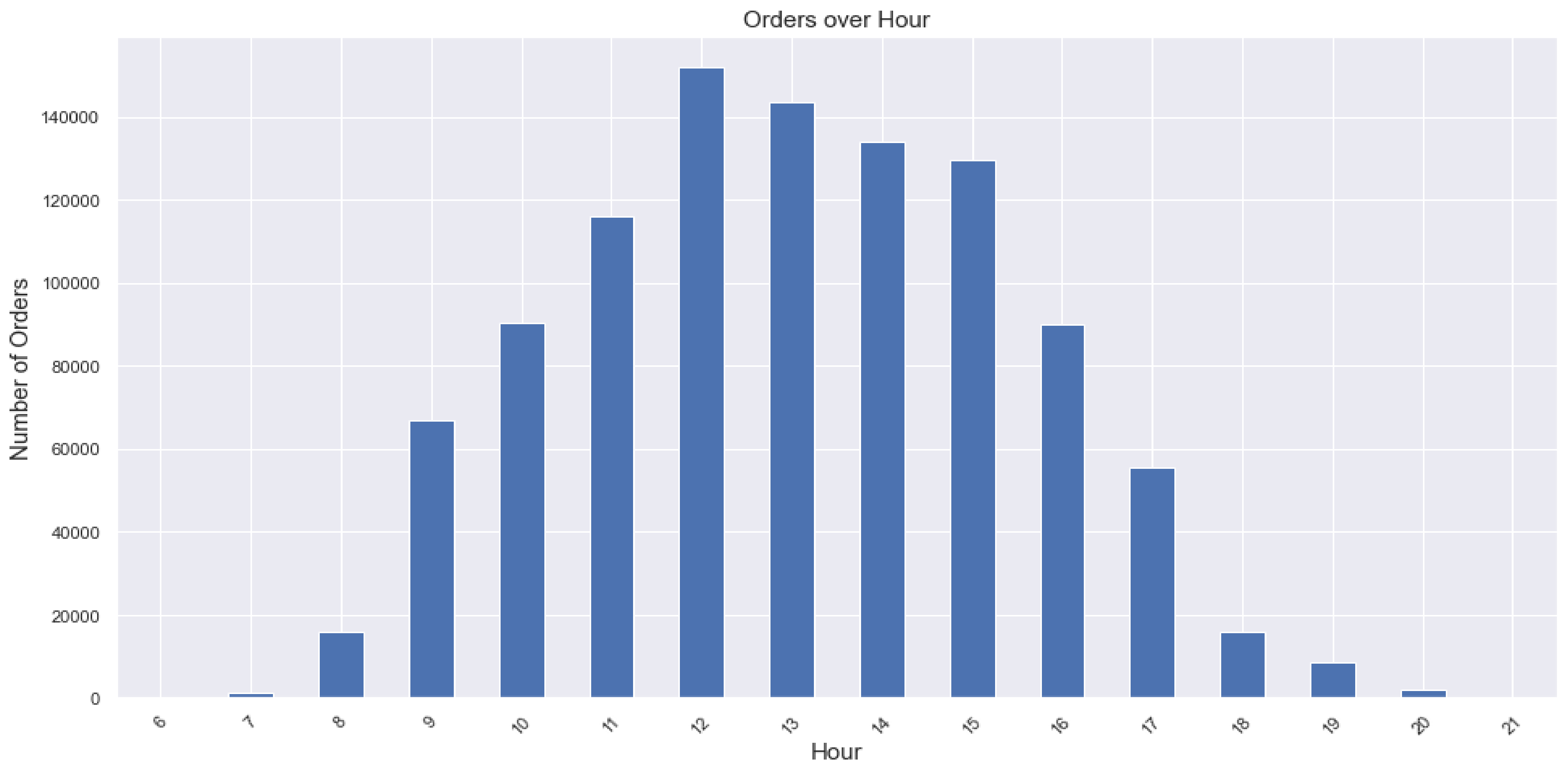
Product Code	Sales Share %
22423	1.68
85123A	1.31
85099B	0.92
23843	0.86
47566	0.75
84879	0.66
22086	0.60
23166	0.42
79321	0.41
22197	0.40



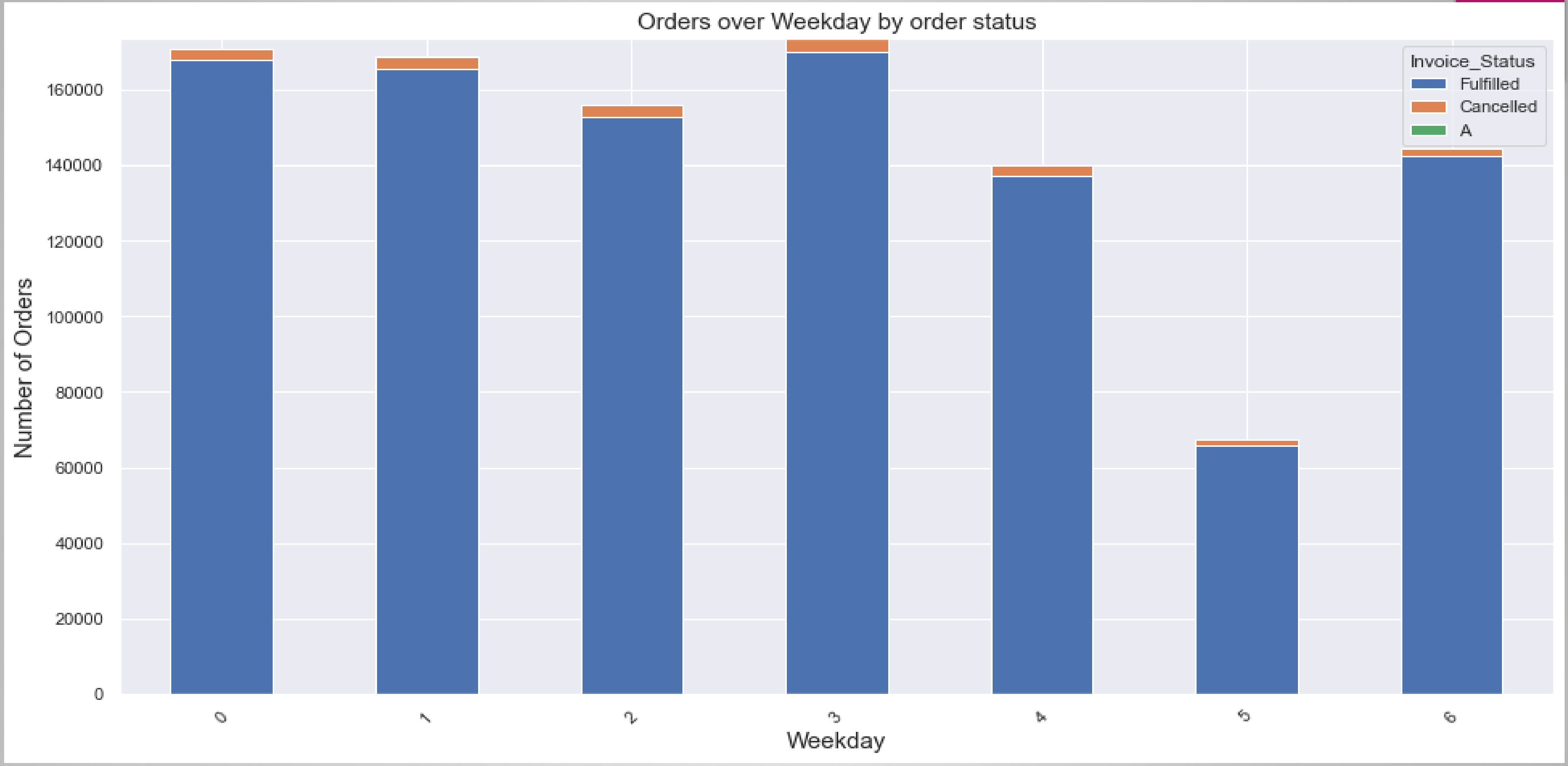
the top selling product only accounts for 1.6 % of the sales ,
which implies that the sales are distributed over a wide product line

but even then , the sales pattern stills follows a sharp **Pareto distribution** ,
with a minority of SKUs responsible for majority of the sales

Transactions Volumes in Day and Week Levels



Orders Peak in the Working Hours ,
which might suggest that the company is an Industrial / B2B supplier , instead of a consumer retailer



There are very Few Orders coming in on Saturdays , while Sunday Seems to be an active Day
It is possible that the company could have an Online Channel too for accepting Orders on Non-Working days

Appendix 1

DATA INTEGRITY DEEP-DIVE

Steps In the Data Pipeline

- Export Data from Incoming Excel to Create CSV dumps
- Append 2 CSVs to create a Pandas DF
- Remove Absolute Duplicates (All rows are same).
- Establish Logical Level of the Data
  - Data is supposed to be at Invoice-Item Level
- Rename Columns for Readability and Ease-of-use
- Remove Cancellation encoding from Invoice_ID as a separate flag column
- Analyse Product Code for Anomalies , and create a Flag to mark these
  - (DeepDive in next slides)
- Type setting of Columns

Anomalous Data Points in Columns

- **Analyse Negative Values in Product_Qty**
  - 86% of Negatives explained by Order Cancellations (maybe to adjust inventory count)
  - Remaining Explained By iNstances where Cust_ID was NULL
  - Took these Product_Desc as added them to the Adjustment Codes List for the Mask Flag
- **Analyse Negative Values in Product_Price**
  - Negative Prices are found only During yearly Stock Adjustments , They Have Product Code ‘B’
  - Special Product_Desc codes explain such discrepancies , added them to the Codes List
- **Handling Null values in Cust_ID :**
  - Since each Invoice_ID is uniquely mapped to only one Cust_ID
  - Used Invoive_ID – Cust_ID map to fill in missing Cust_IDs
  - But this was unsuccessful in filling in , as these missing Cust_IDs were in a block of time

Fixing the Data-Level

- **Analysing Errors in Logical Data Level**
  - Data is supposed to be at Invoice-Item Level
  - But on doing Groupby-Count there are 2.2% Instances which violates this level (i.e. have count > 1)
  - This was NOT because of there being Duplicate entries for Cancelled Orders
    - (i.e. one entry for Original order , then one entry when order was Cancelled , This was NOT the case)
  - There were Multiple Rows with Different Prices and Qtys for the same Invoice-Item Level
    - Possibly to calculate the same Order-Item at different prices for multiple Qtys
  - There were very few instances of Product_Desc column showing level discrepancies,
    - Since that column was not relevant to any analyses , we can ignore those
- **Fixing the Level to Inovice-Item :**
  - All the columns were Aggregated appropriately (Sale : Sum , Qty : Sum , Price : Mean , Rest : Mode)

Creation of Auxiliary Features for Analysis

- Time Dimension Features for Group Bys :
  - created from TimeStamp column : Invoice_Date
  - Levels : Year , Quarter , Month , Week , WeekDay , Day , Hour
- Country column simplification
  - There were 43 distinct countries ,
  - But 'UK' Accounts for 93% of Transactions
  - So grouped Countries into based on Sales and Geography
- Exported Final data for Consumption in Analysis

```
df['Invoice_Date']
df["Year"]
df["Quarter"]
df["Month"]
df["Week"]
df["Weekday"]
df["Day"]
df["Date"]
df["Hour"]
df["Day_of_Year"]
df["Year_Quarter"]
df["Year_Month"]
df["Year_Week"]
df["Year_Day"]
```

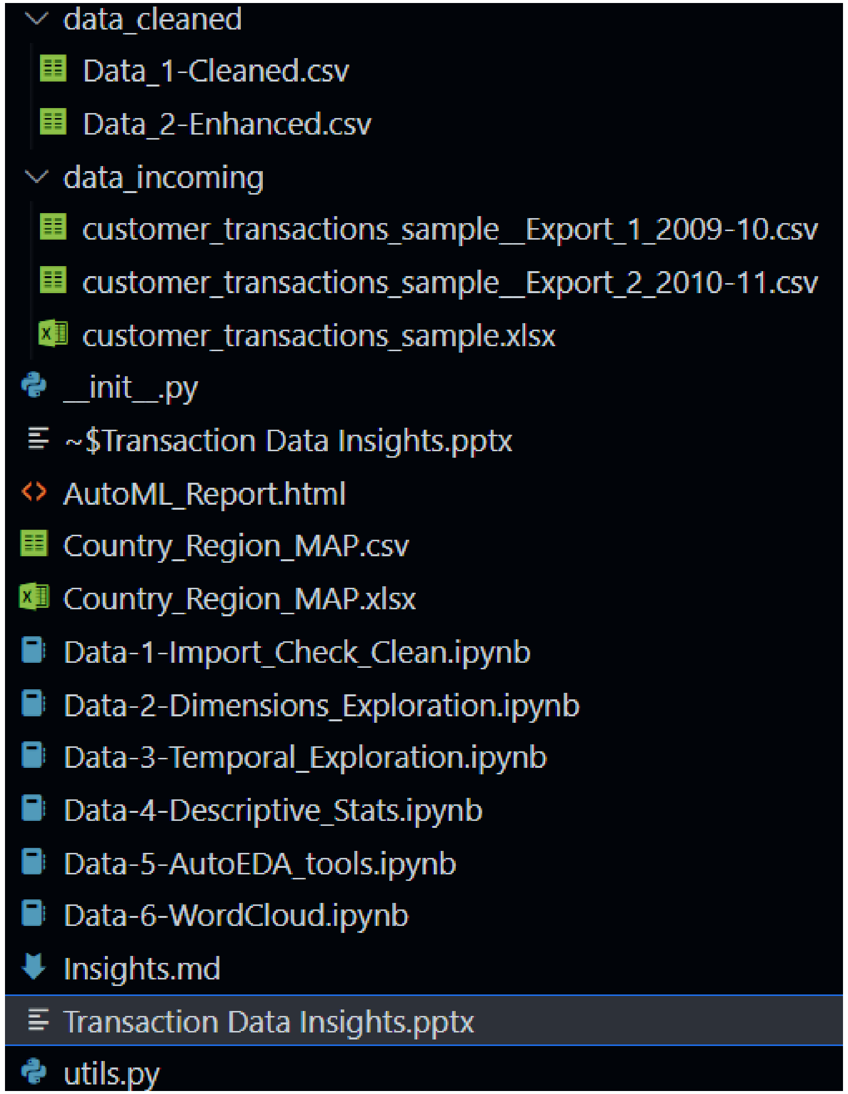
Row Labels	Sum of TRN_Count
UK	936909
EUROPE	29675
EIRE	17653
GERM	17321
FRANCE	13983
OTHERS	5883

Appendix 2

FURTHER IMPROVEMENTS AND NEXT STEPS

Things to try out

- Rate of Sale Analysis :
  - Which item sells fast , which Items sell slow , Their Contribution to Sales
- Order Cancellation RCA
  - What circumstances lead to cancelled orders
- RFM Analysis : Recency , Frequency , Money
- Customer Segmentation (Unsupervised Clustering)
- WordCloud , NLP embeddings



Appendix 3

BRIEF CODE BASE OVER-VIEW

Thank You

Saif Raja
7775029864
saifuddin.raja24@gmail.com