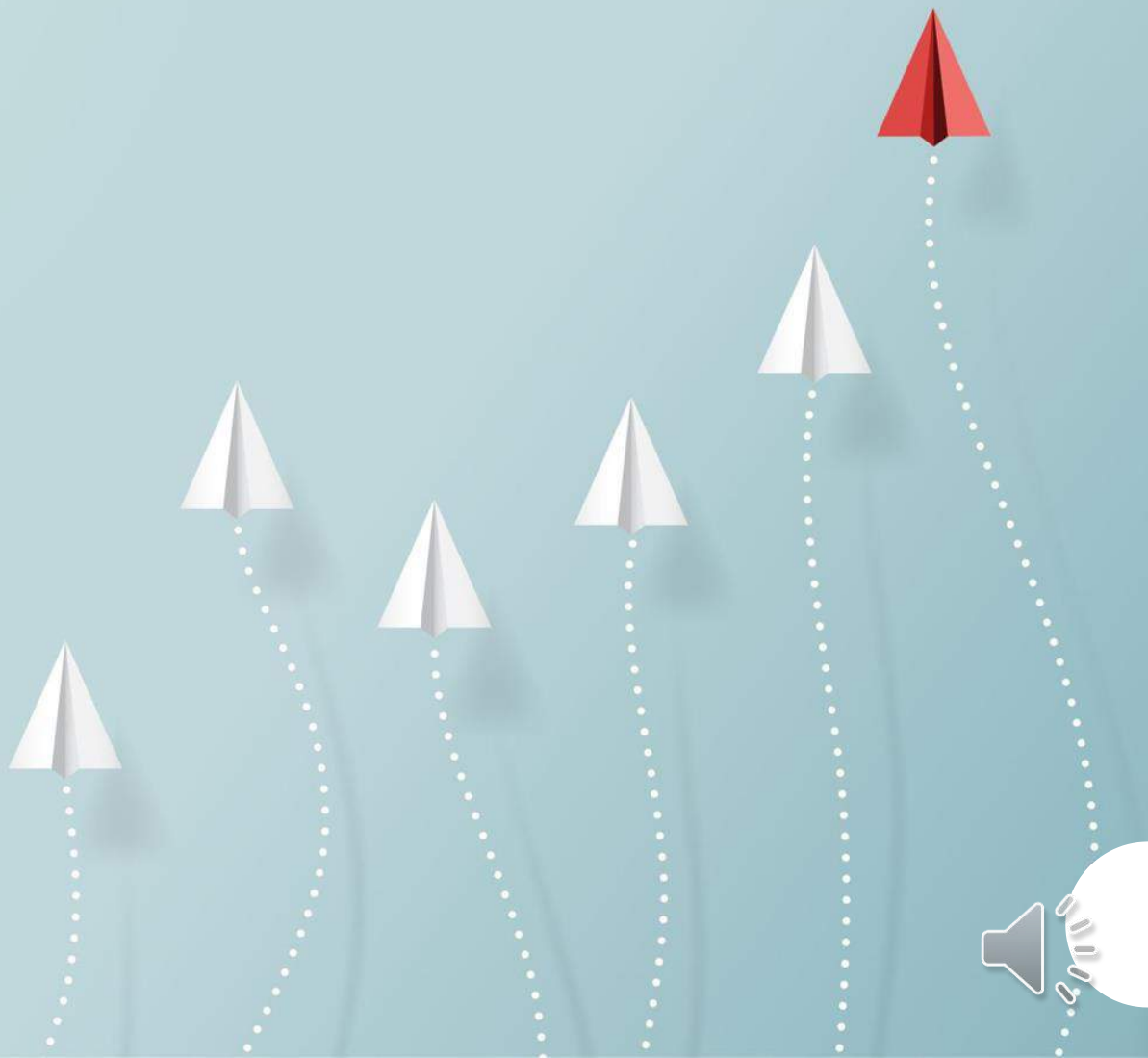
Marketing and Retail Analytics – Capstone Project

Syed Almaas Parvez



Problem Statement

OList is an e-commerce company that is facing some losses recently and they want to manage their inventory very well so as to reduce any unnecessary costs that they might be bearing.

I will be managing the inventory cost of this e-commerce company called OList.

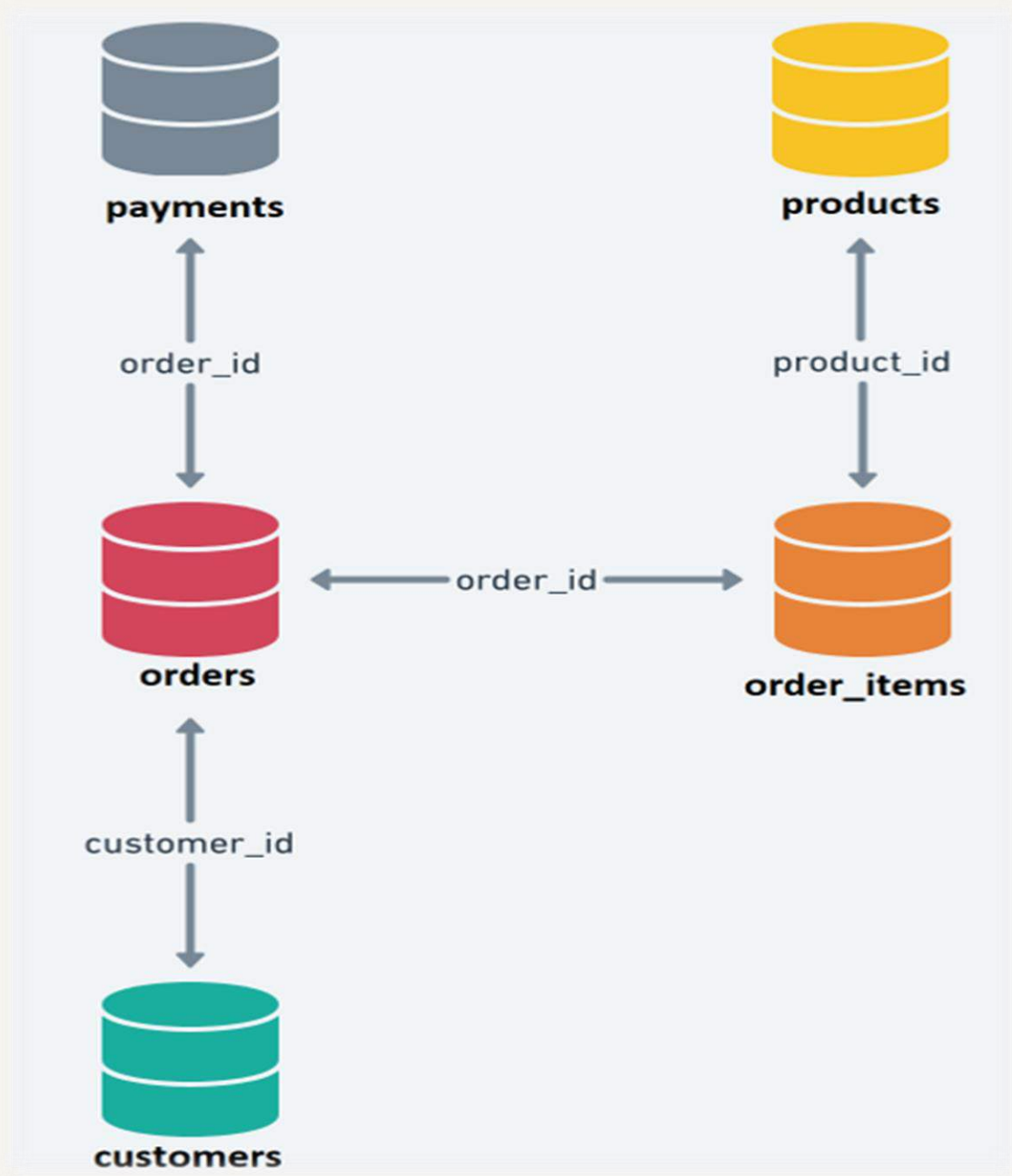
For this purpose I will identify top products that contribute to the revenue and also use market basket analysis to analyse the purchase behavior of individual customers to estimate with relative certainty, what items are more likely to be purchased individually or in combination with some other products.

Olist Dataset- Retail Dataset

The dataset available to us is called "Retail_dataset". It is a xlsx file or simply a spreadsheet.
There are five worksheets in the excel file.
They are namely:
1. Orders
2. Order_Items
3. Products
4. Payments
3. Customers

The total number of Products available at Olist Warehouse is 32950.
There are total 99440 Orders ordered.
Unique customers total at 96096.
There were 3345 repeat customers.
The dataset contains transactions from September 2016 to September 2018.

Entity Relationship Diagram



Step 1: Importing Libraries and Reading Data

I used many python libraries throughout the project. Some of the most important of them are as follows: pandas, numpy, seaborn, Sklearn, matplotlib.pyplot, squarify etc.

I formed a dataframe for each sheet in the excel file using the pandas 'read_excel' function to read all the different sheets in the "Retail_dataset".

Step 2: Data Cleaning and EDA

Since the data available to us contains many duplicates and Null values. We will first make the data suitable for evaluation by performing Exploratory Data Analysis.

As per requirement, we are interested in only orders having statuses as 'delivered'. We will filter down the values in the jupyter notebook and create a separate cleaned file which we can use to create visualizations in Tableau.

If necessary we may create new derived fields which may enhance our understanding of the dataset.

EDA Analysis Observations

We find that 97 % or orders were of 'delivered' status. I dropped all the others orders for the purpose of this project.

There were 160 'order_approved_at' null values and 2965 'order_delivered_timestamp' null values in the order's sheet.

I filled the Nan values of 'order_approved_at' and 'order_delivered_timestamp' with the appropriate values derived from the same dataframe.

Step 3: Creating a Cleaned Excel File

After having gone through the whole dataset and performing necessary cleanup and EDA. I converted the dataframes into a new excel file using the 'ExcelWriter' function of pandas.

I combined all the dataframe into a single sheet to be used further in the Tableau environment.

Step 4: Merging Dataframes

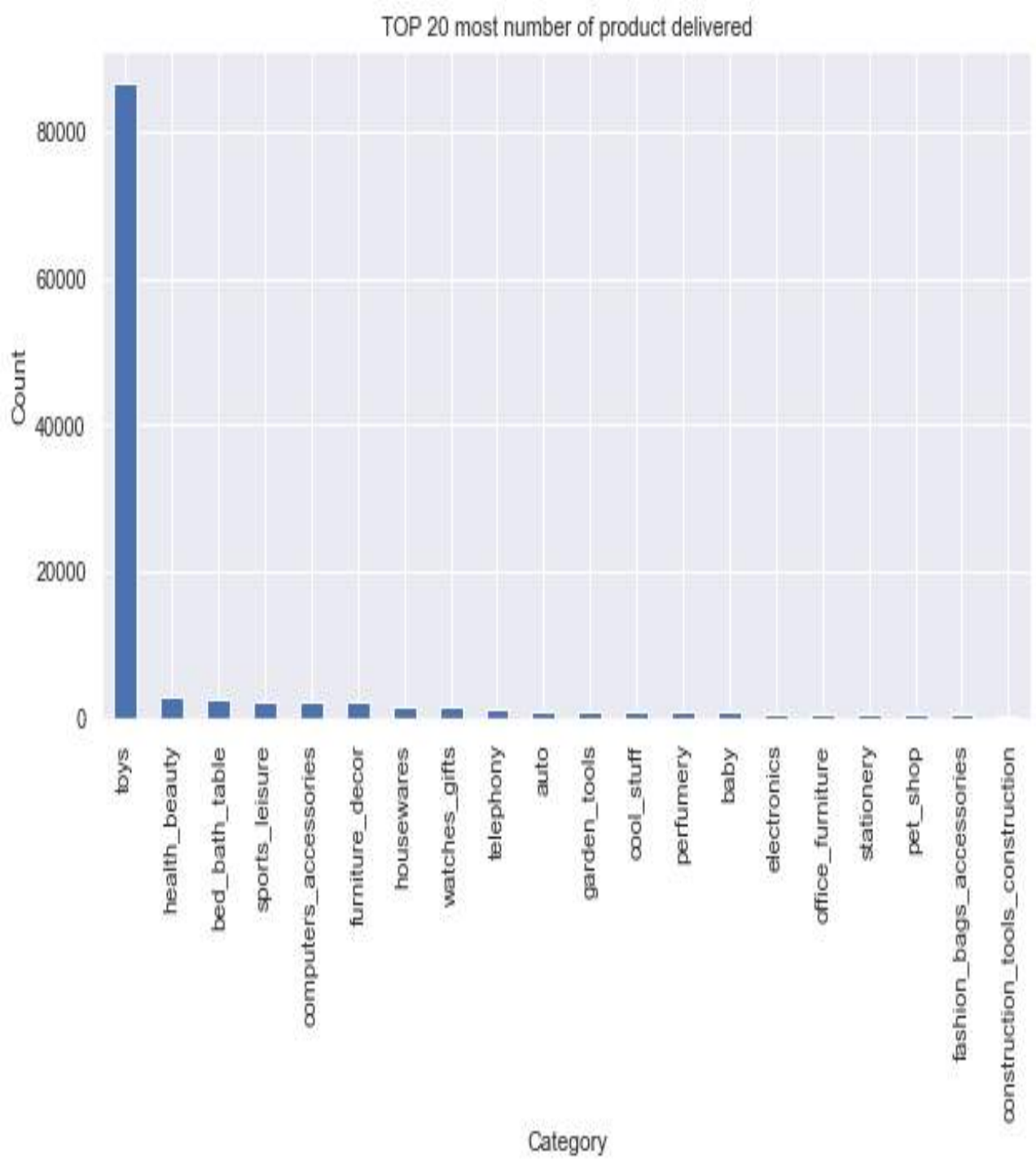
Using pandas merge function to join all the sheets using the common fields as depicted in the Entity Relationship Diagram.

Step 5: Data Visualisation in Python

Using bar plot and pandas value_counts functions, I figured out the:
Top 20 most ordered product categories
Top 20 most revenue genrerating product categories

I saw that the most frequent purchases from a product category are in the following order:
1.Toys 2. Health and Beauty 3. Bed Bath Table 4. Sports and Leisure and 5.Computer Accessories

I observed that Toys health and beauty and watches and gifts make up more than 80% of all revenue.


TOP 20 most number of product delivered
Count
80000
60000
40000
20000
0
toys
health_beauty
bed_bath_table
sports_leisure
computers_accessories
furniture_decor
housewares
watches_gifts
telephony
auto
garden_tools
cool_stuff
perfumery
baby
electronics
office_furniture
stationery
pet_shop
fashion_bags_accessories
construction_tools_construction
Category

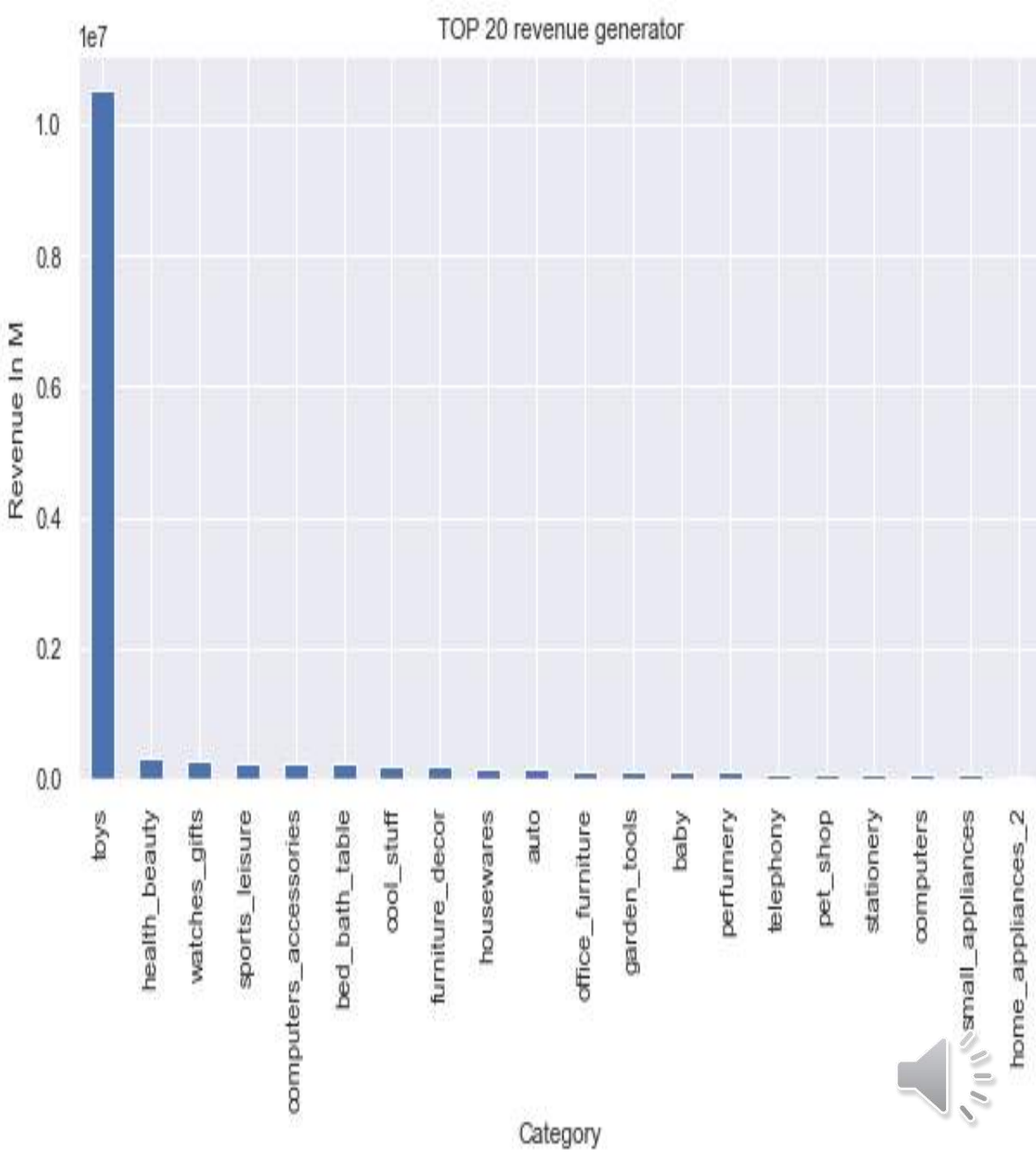
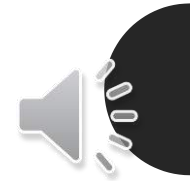
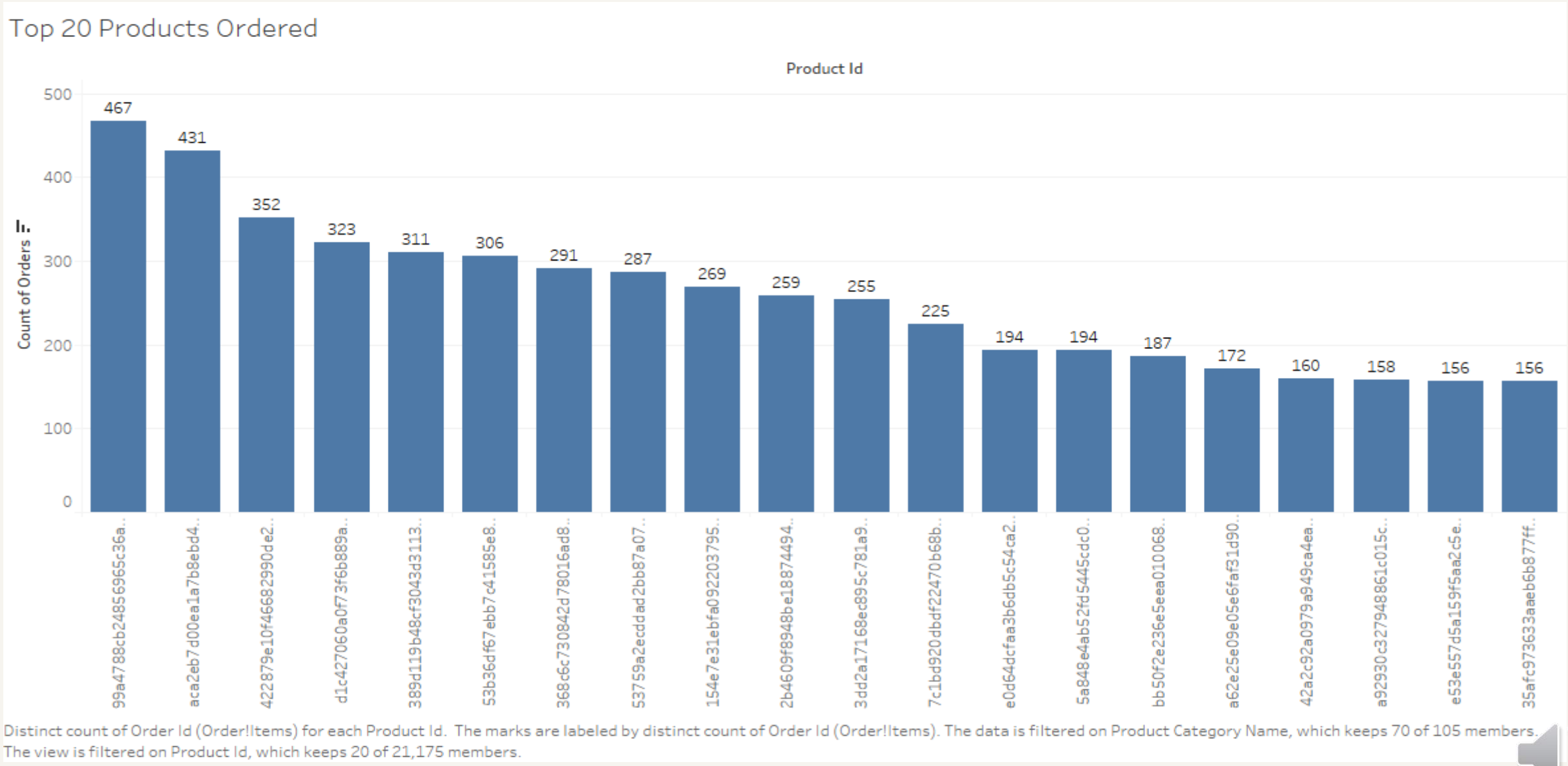

TOP 20 revenue generator
1e7
Revenue In M
1.0
0.8
0.6
0.4
0.2
0.0
toys
health_beauty
watches_gifts
sports_leisure
computers_accessories
bed_bath_table
cool_stuff
furniture_decor
housewares
auto
office_furniture
garden_tools
baby
perfumery
telephony
pet_shop
stationery
computers
small_appliances
home_appliances_2
Category

Tableau Visualizations

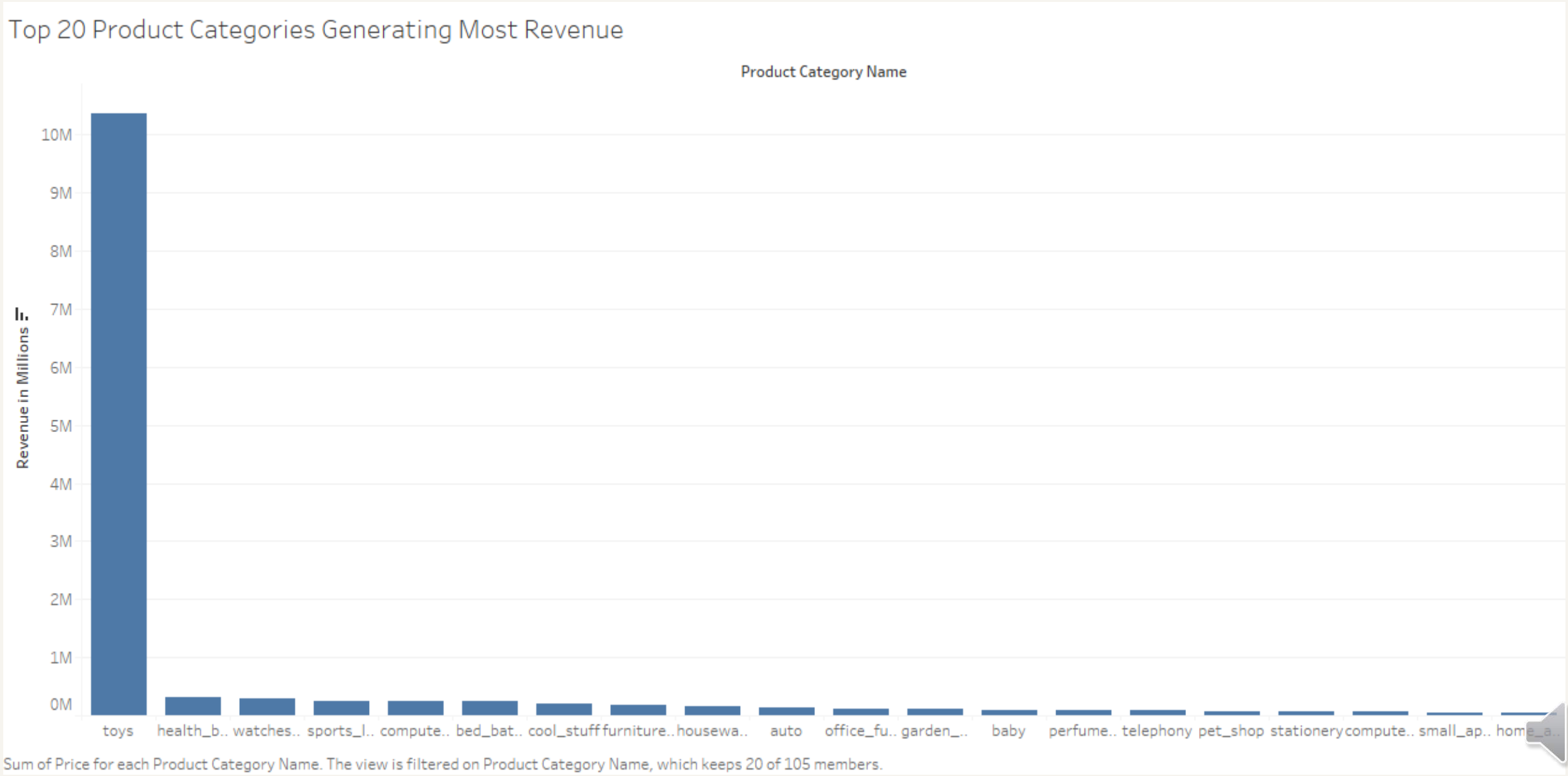


Top 20 Most Ordered Products by Product Id

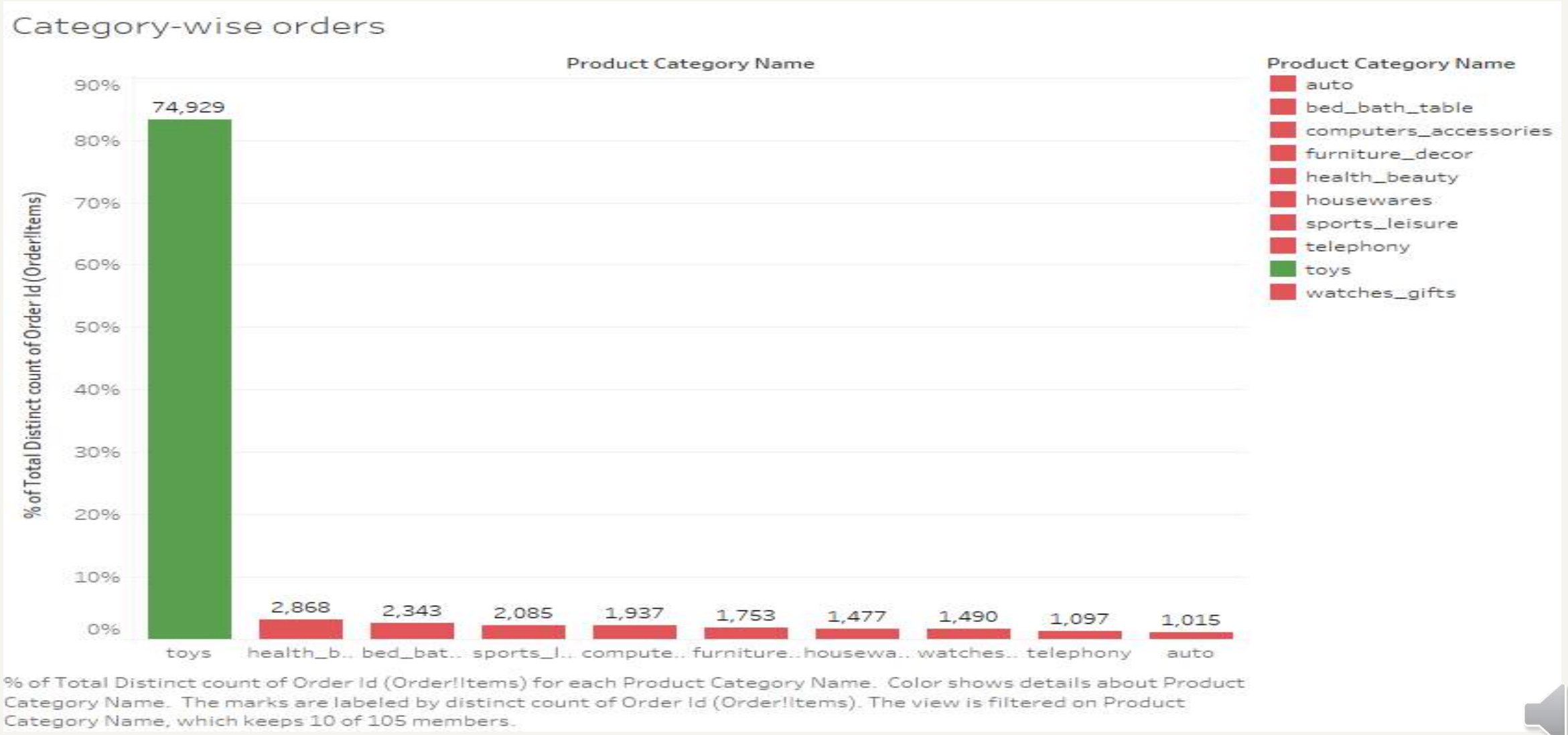


Distinct count of Order Id (Order!Items) for each Product Id. The marks are labeled by distinct count of Order Id (Order!Items). The data is filtered on Product Category Name, which keeps 70 of 105 members. The view is filtered on Product Id, which keeps 20 of 21,175 members.

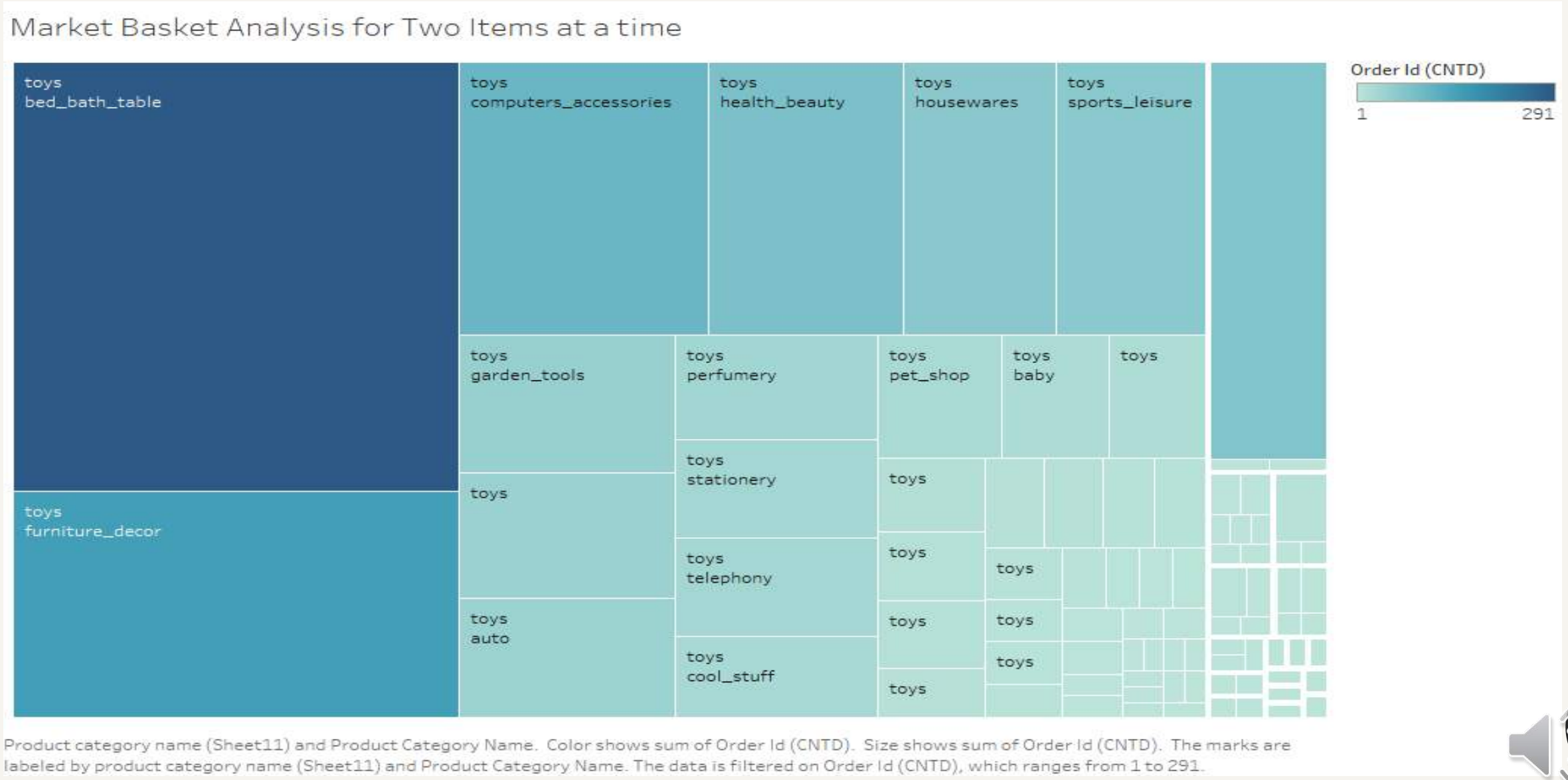
Top 20 Most Revenue Generating Product Categories



Top 10 Most Ordered Product Category



Most Ordered Product Category in Combination of Twos

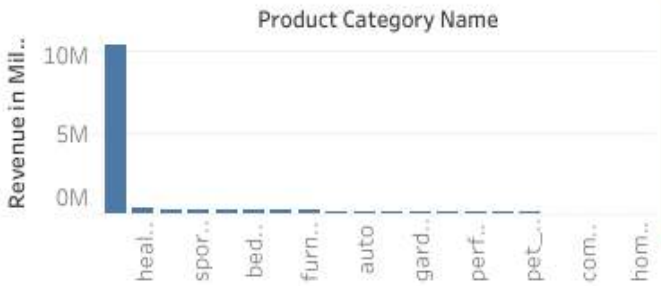


Olist Retail Analytics Dashboard

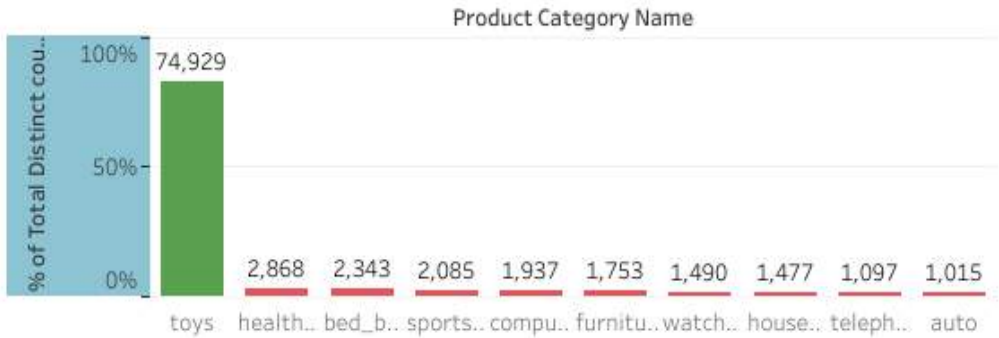
OLIST Retail Analysis

Top 20 Most Ordered Products

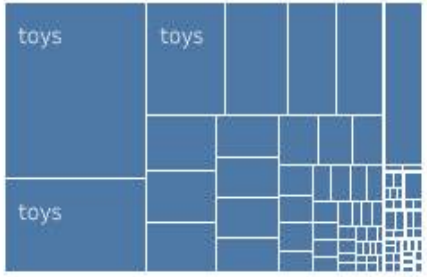
Top 20 Most Revenue Generating Product Categories



Category-wise orders



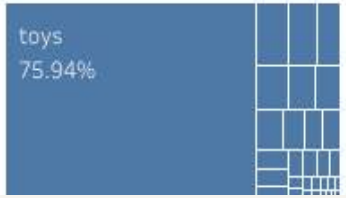
Market Basket Analysis for Two Items at a time



Revenue Pareto

Product Id	Revenue	% of Tota..	% of Tota..
bb50f2e236e5eea010068..	63,885	14.59%	7.14%
d1c427060a0f73f6b889a..	47,215	25.38%	14.29%
99a4788cb24856965c36a..	43,026	35.20%	21.43%
3dd2a17168ec895c781a9..	41,083	44.59%	28.57%
53b36df67ebb7c41585e8..	37,683	61.78%	35.71%
aca2eb7d00ea1a7b8ebd4..	37,609	53.18%	42.86%
e0d64dcfaa3b6db5c54ca..	31,787	69.04%	50.00%
422879e10f46682990de2..	26,577	76.11%	57.14%

Size of Each Product Category by Number of Orders

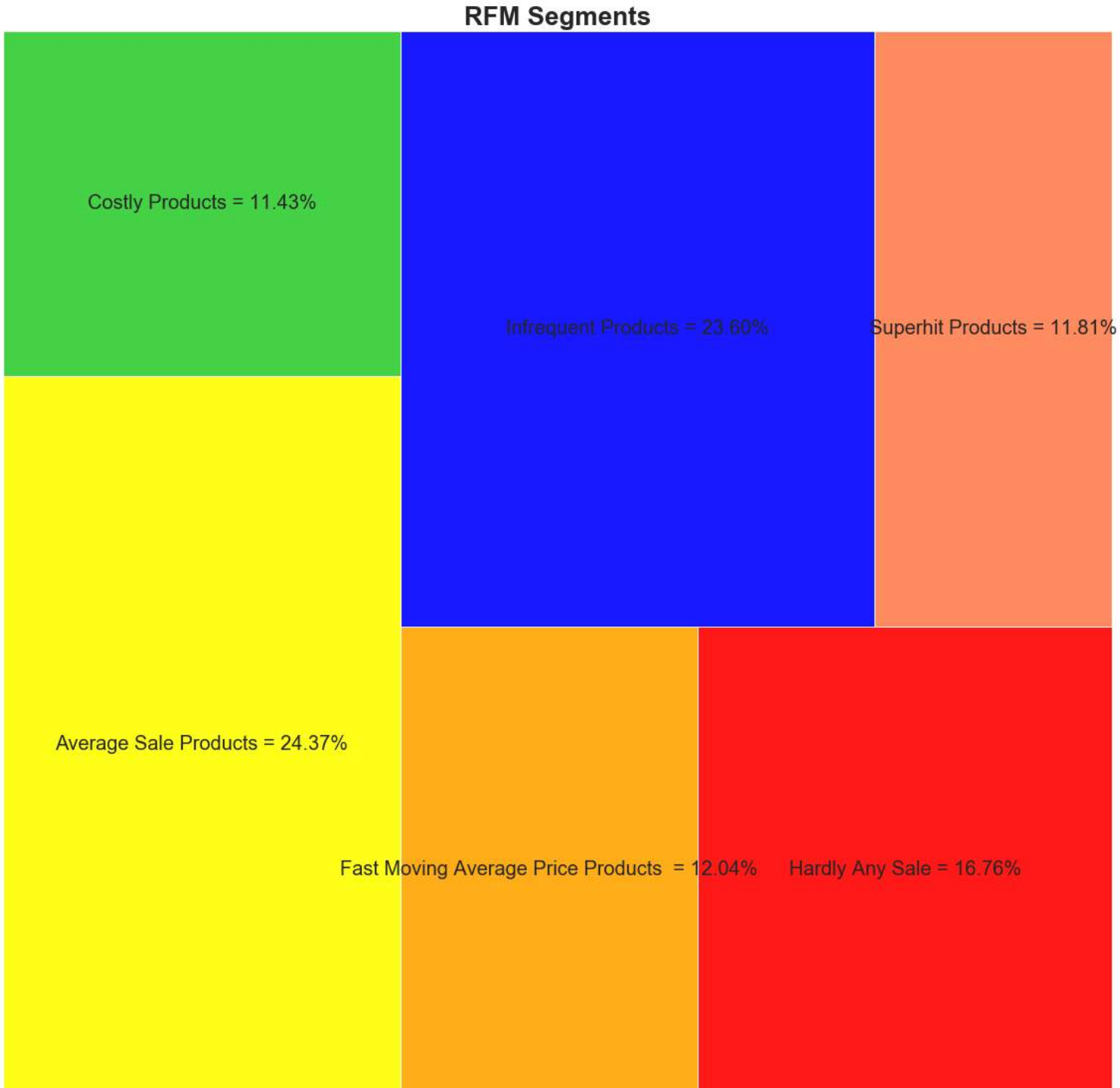


Step 6: RFM Modelling

Using Recency, Frequency and Monetary modelling I divided the products at Olist warehouse into five different categories using a RFM score assignment method. Ranging from 1-4 with the highest number corresponding with the best score on the RFM scale

I got the following product categories:

1. Superhit Products
2 .Average Sale Products
3. Costly Products
4 .Fast Moving Average Price Products
5. Hardly Any Sale Products
6. Infrequent Products



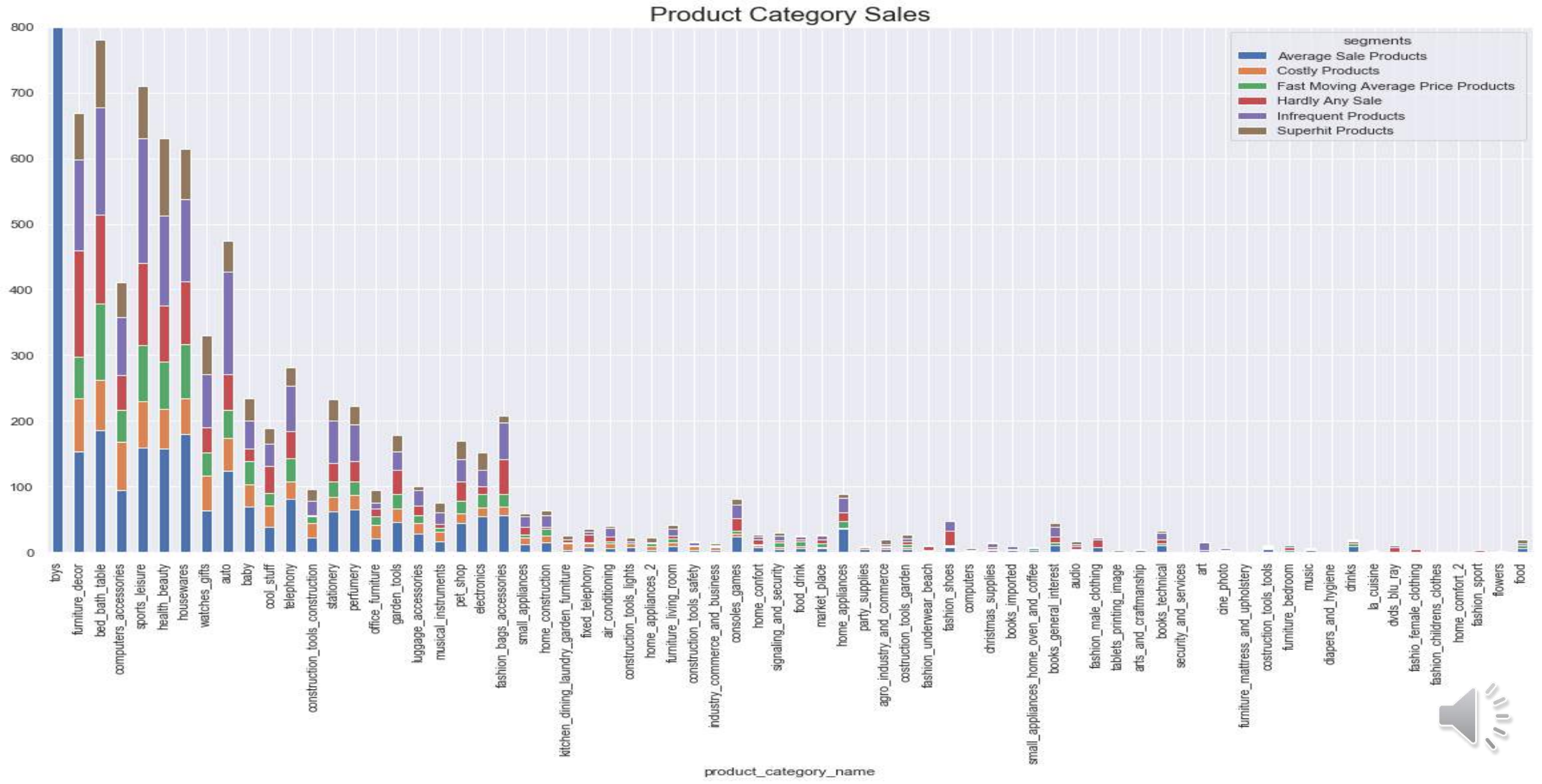
Observation

There is almost equal proportions of average sale and infrequent cheap products. They make up approx. 50% of inventory

Superhit Products and Fast-moving average price products account for 25% of inventory

Hardly any Sale products account for 1 out of every 6 products in the inventory

Stacked Bar Chart of the Product Category Depicting their Product Type



Observation

- It is observed that around 20 product categories account for 80% of Overall Sale count.
- Close to 40-50 % of the sales of furniture decor, bed bath table, sports_leisure, fashion bags and accessories and auto have low frequency or hardly any sale products
- Flowers, home comfort, fashion childrens clothing , furniture mattress and upholstery , security and services and many more have hardly any sale

Step 7: Market Basket Analysis Using Apriori

Since the Apriori function expects data in a one-hot encoded pandas. I created a basket for every order distinguished by the product category using 1-hot encode to assign 0 to all less than 0 and 1 to all positive values.

I created the basket for two products in an order and three products in an order to analyze the most frequent products purchased together in a single order.

Observations In Market Basket Analysis

Top five products categories in groups of twos are:

1. Toys and Bed Bath Table
2. Toys and Fashion Bags Accessories
3. Toys and Auto
4. Toys and Watches Gift
5. Toys and Health & Beauty

Top five products categories in groups of threes are:

1. Toys, Cine photos and Telephony
2. Toys, Home Construction and Computer Accessories
3. Toys, Garden Tools and Computer Accessories
4. Toys Furniture Decor and Electronics
5. Toys, Furniture Decor and Health and Beauty