

Marketing and Retail Analytics – Capstone Project

- Kasyapa Polamraju

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.*
- *For this purpose I will identify top products that contribute to the revenue and also use market basket analysis to analyse the purchase behaviour 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.*
 1. Orders
 2. Order_Items
 3. Products
 4. Payments
 5. Customers
- *The total number of Products available at Olist Warehouse is 32950.*
- *There are total 99440 Orders ordered, unique customer are 96096.*
- *There were 3345 repeated 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.

EDA Analysis Observations

- We find that 97 % of 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: Merging DataFrames

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

Step 4: 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 generating 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 & watches and gifts make up more than 80% of all revenue.

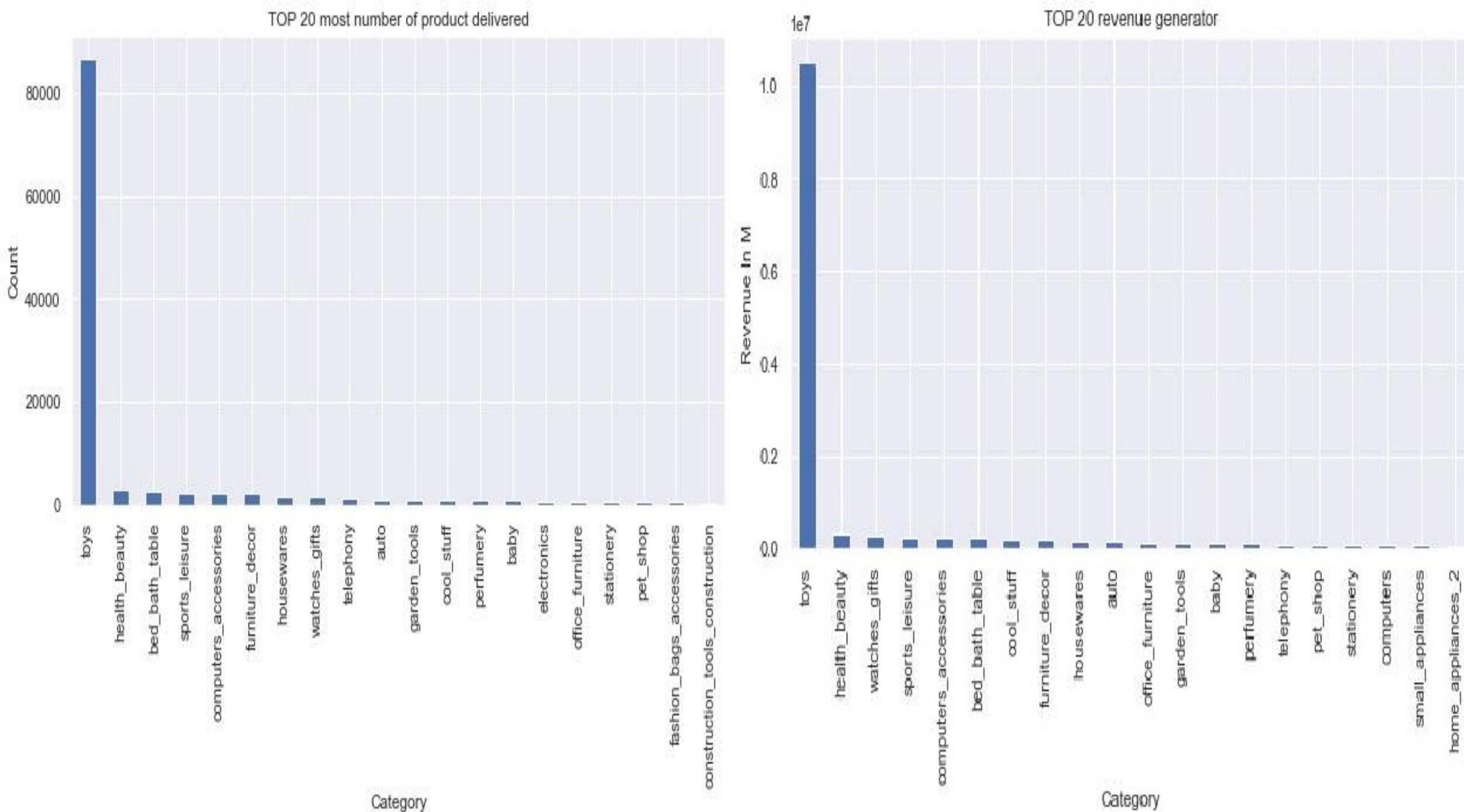
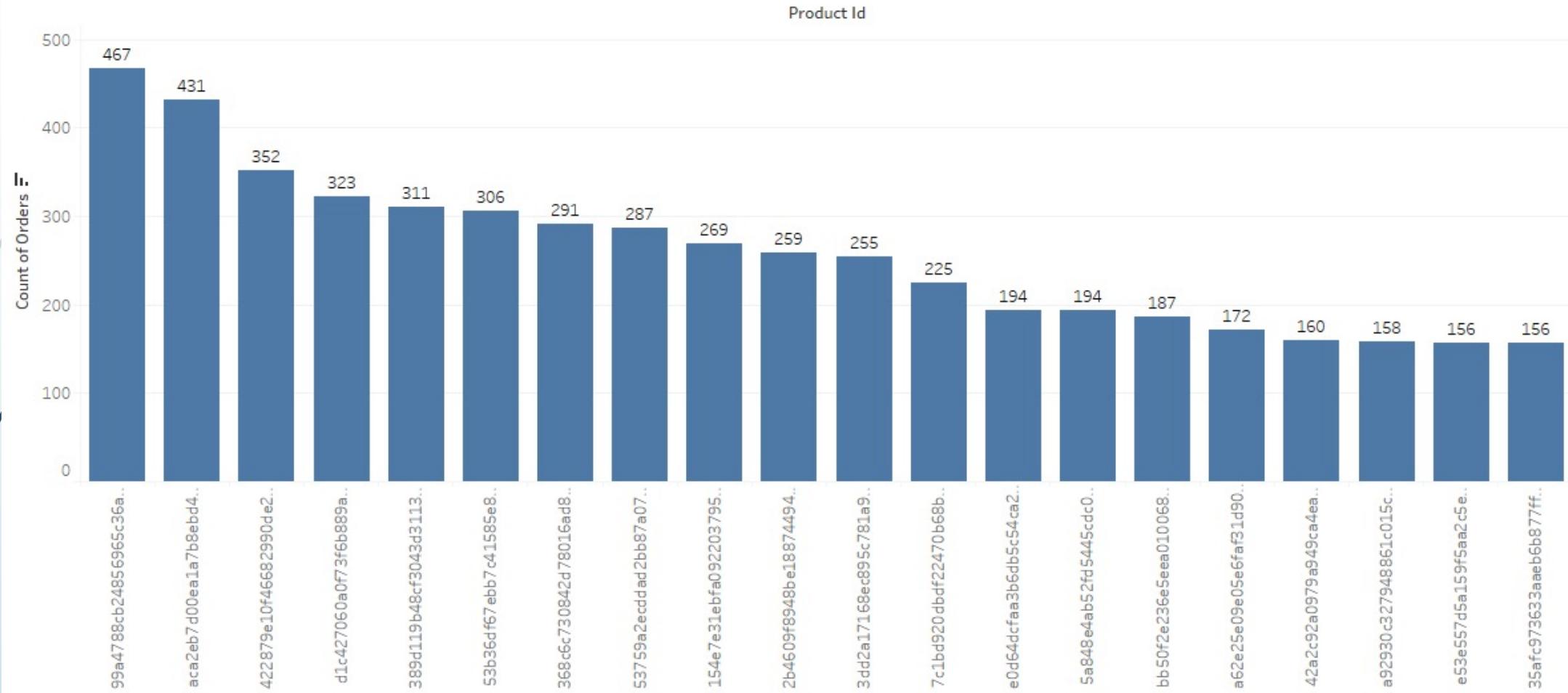




Tableau Visualization

Top 20 Most Ordered Products by Product Id

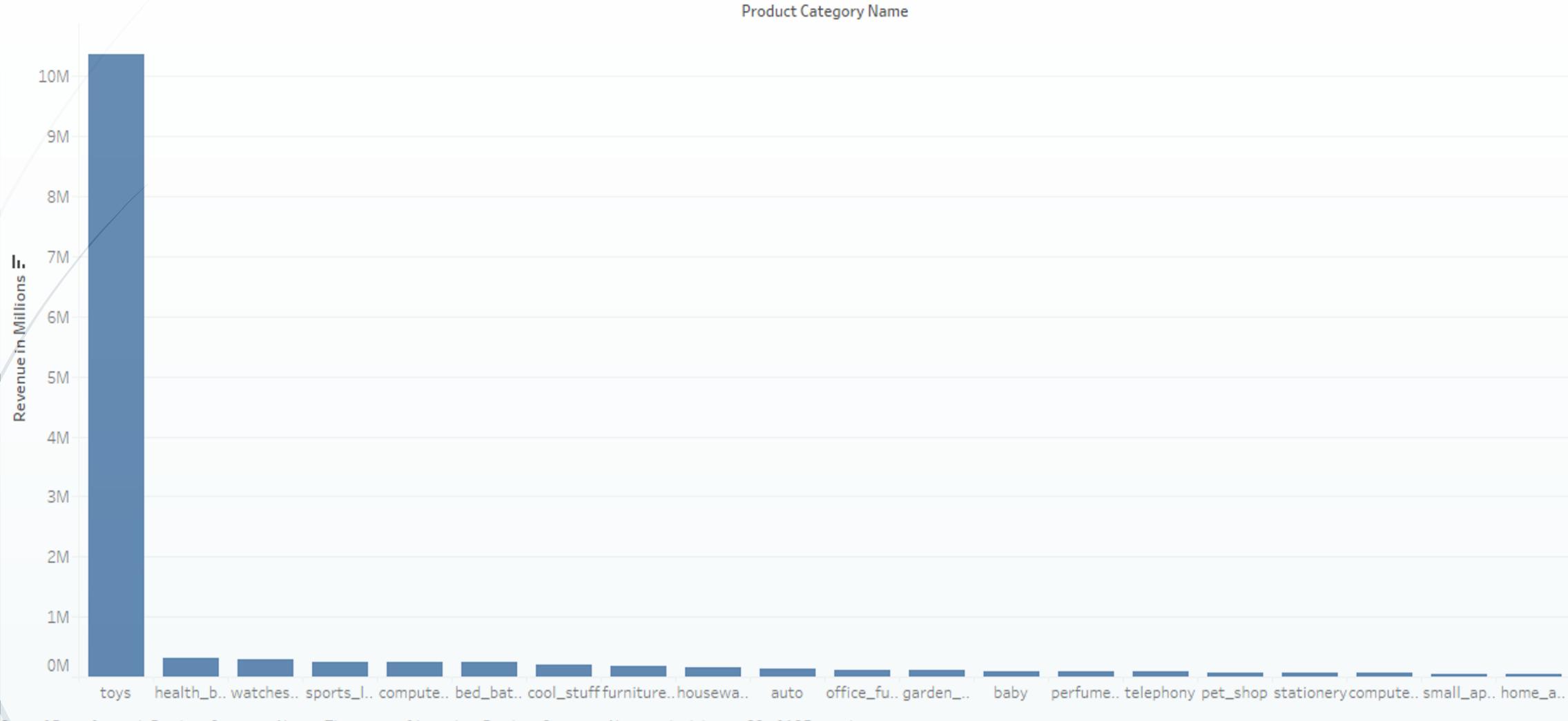
Top 20 Products Ordered



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.

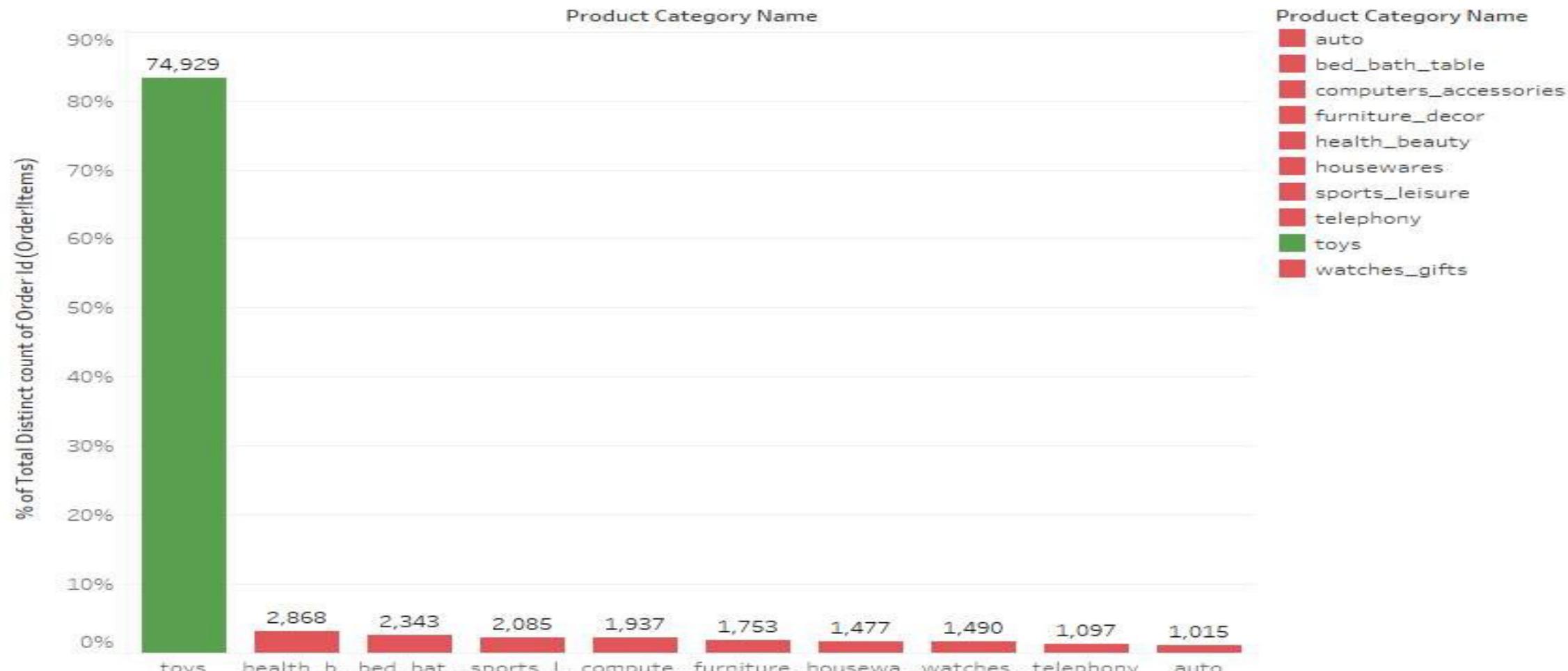
Top 20 Most Revenue Generating Product Categories

Top 20 Product Categories Generating Most Revenue



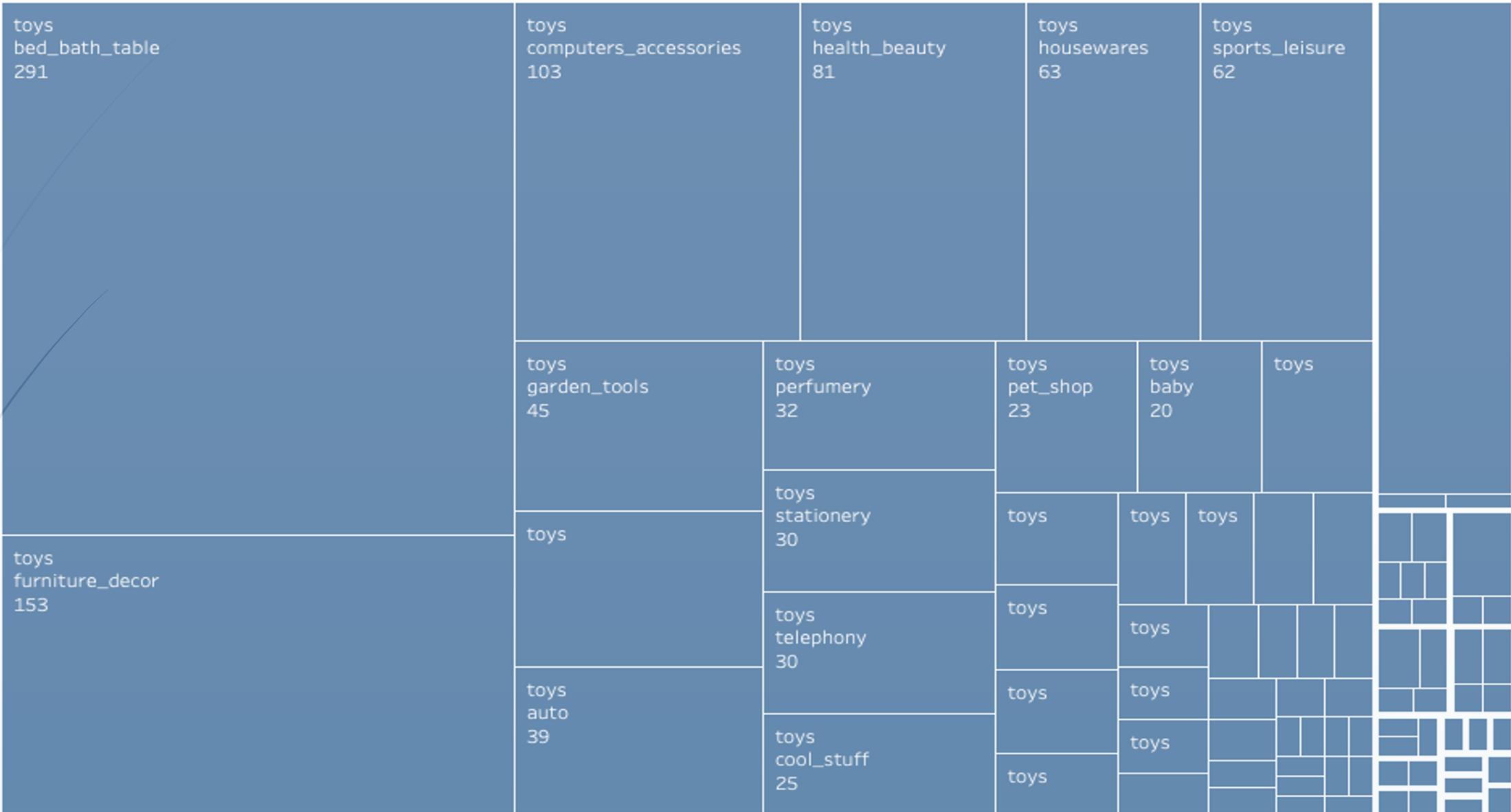
Top 10 Most Ordered Product Category

Category-wise orders



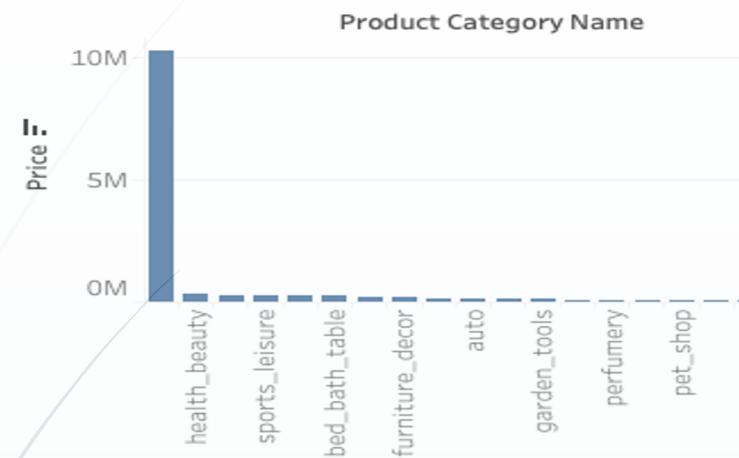
% of Total Distinct count of Order Id (Order!Items) for each Product Category Name. Color shows details about Product Category Name. The marks are labeled by distinct count of Order Id (Order!Items). The view is filtered on Product Category Name, which keeps 10 of 105 members.

Most Ordered Product Category in Combination of Twos



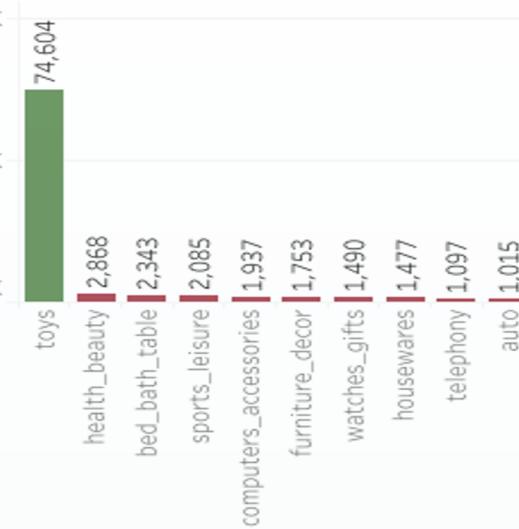
Olist Retail Analytics Dashboard

Top 20 Most Revenue Generating Product Categories



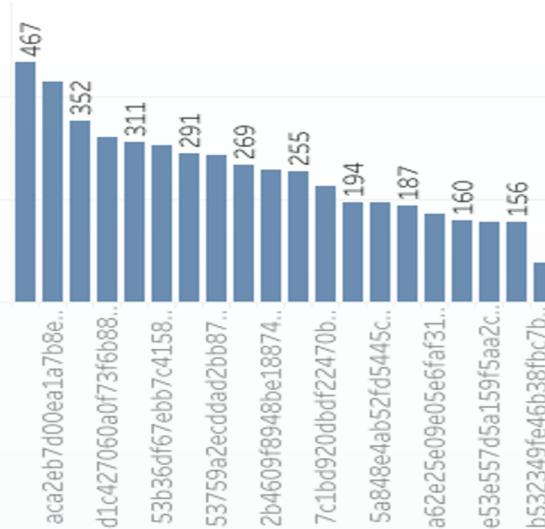
Category-wise orders

Product Category Name

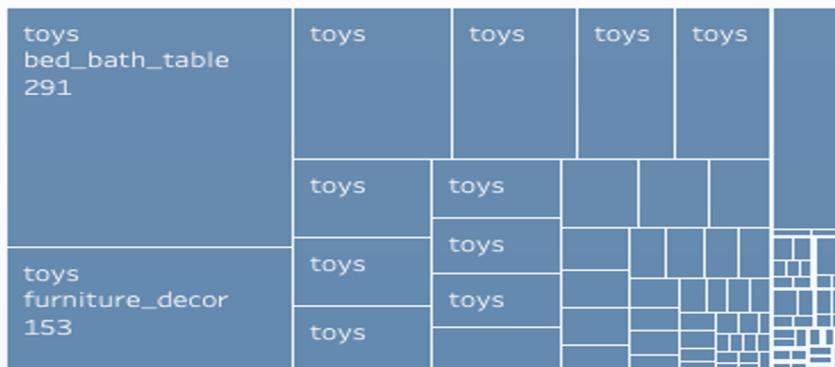


Top 20 Most Ordered Products

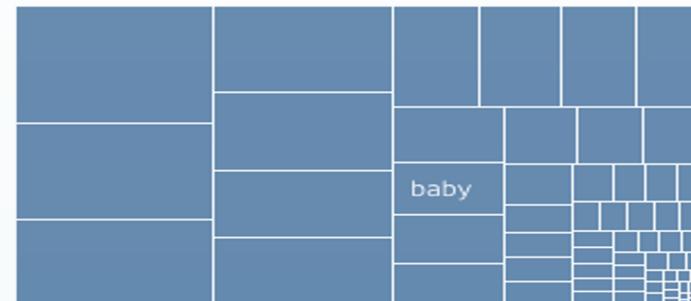
Product Id (Products)



Market Basket Summary



Share of Single Product Category



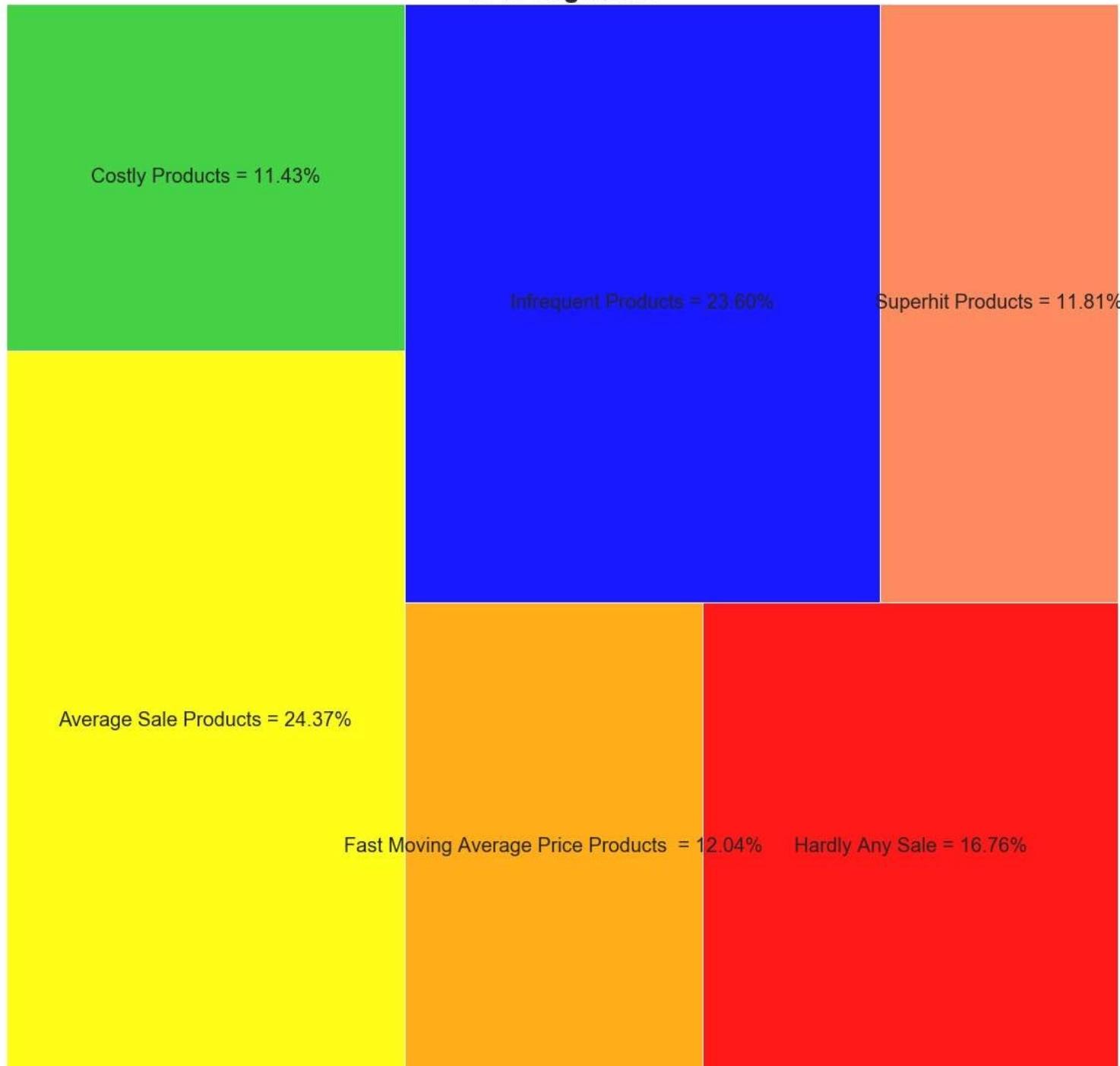
Revenue Pareto

Produ..	% of Total Pr..	Revenue
bb50f2e23..	7.14%	63,885
d1c427060..	14.29%	47,215
99a4788cb..	21.43%	43,026
3dd2a1716..	28.57%	41,083
53b36df67..	35.71%	37,683
aca2eb7d0..	42.86%	37,609
e0d64dcfa..	50.00%	31,787
422879e10..	57.14%	26,577
5a848e4ab..	64.29%	24,229
2b4609f89..	71.43%	22,717
389d119b4..	78.57%	21,441
53759a2ec..	85.71%	20,387
7c1bd920d..	92.86%	13,867

Step 5: 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:
 - Superhit Products
 - Average Sale Products
 - Costly Products
 - Fast Moving Average Price Products
 - Hardly Any Sale Products
 - Infrequent Products

RFM Segments



Observations

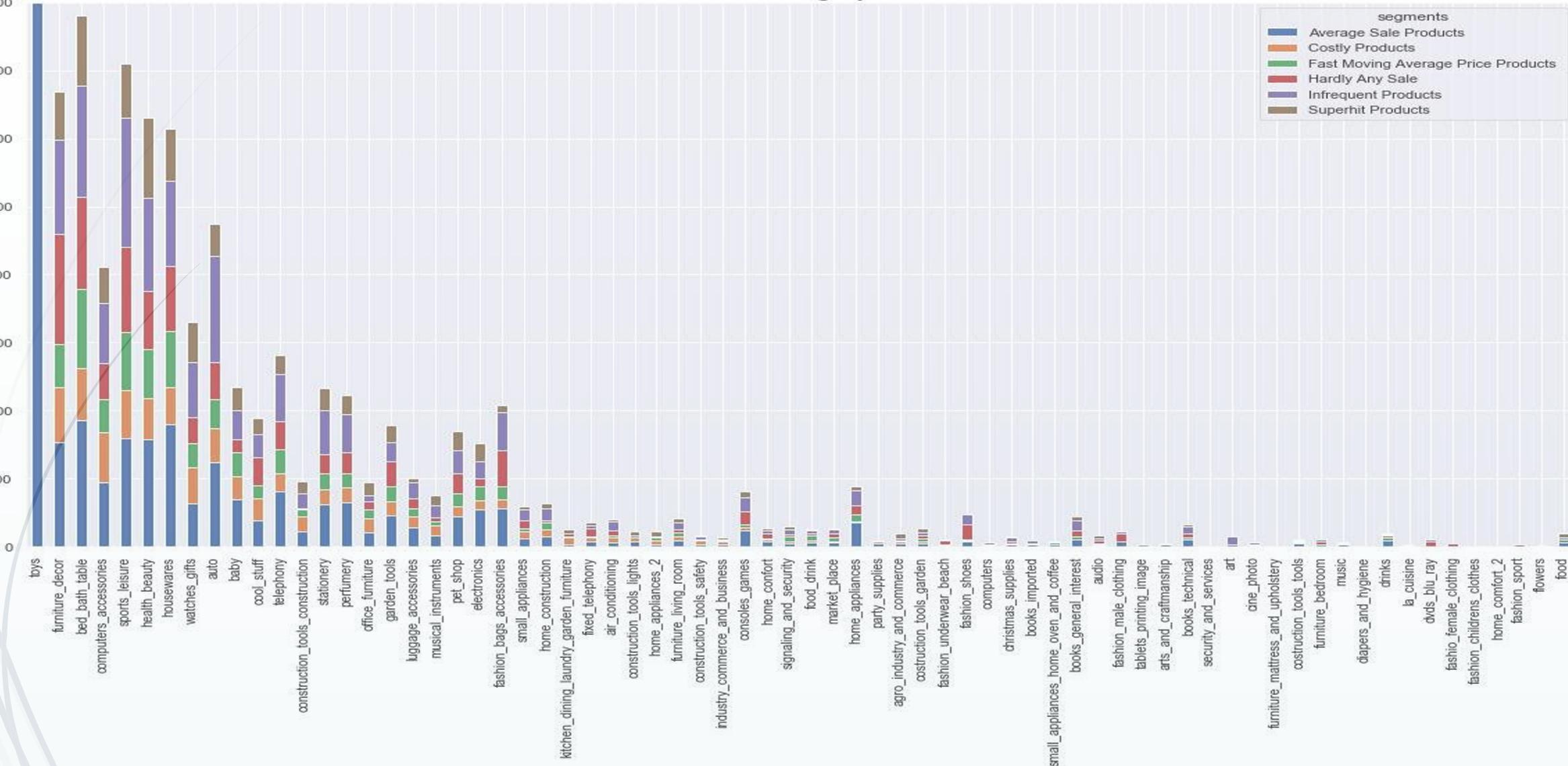
There is almost equal proportions of average sale and infrequent cheap products. They make up approx. 50% of inventory.

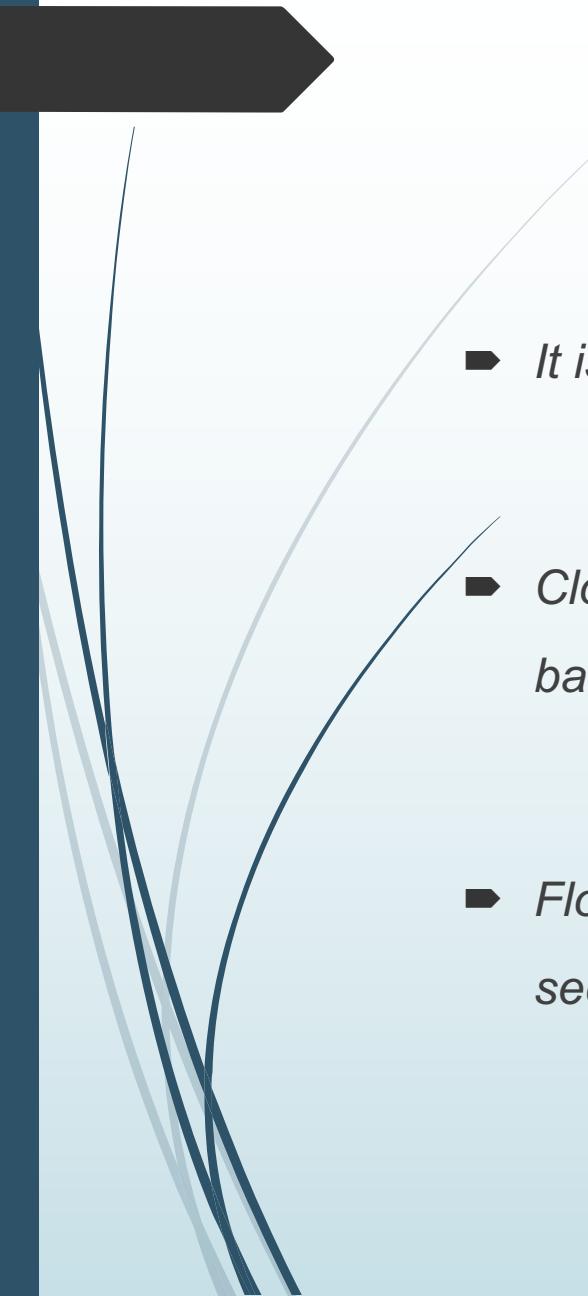
Superhit Products and Fast- moving average price products account for 25% of inventory

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

Product Category Sales





Observations

- ▶ *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_children_clothing, furniture_mattress_and_upholstery, security_and_services and many more have hardly any sale.*

Step 6: 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 analyse the most frequent products purchased together in a single order.

Observations In Marketing 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*