Brazilian e-commerce store: Olist

Mahbubur Rahman

7/17/2021

Introduction

We did the data analysis and data visualization on a large Brazilian e-commerce store named "Olist". We answer the following research questions:

- Which are the explanatory variables or features that influence to calculate the top products of the e-commerce store?
- Calculate and visualize the top products of the e-commerce store using the explanatory variables or features that are selected from the principle component analysis.

Dataset Overview:

There are total 8 tables in the db schema. They are given below:

1. olist_orders_dataset
2. olist_order_items_dataset
3. olist_products_dataset
4. olist_order_customer_dataset
5. olist_order_reviews_dataset
6. olist_order_payments_dataset
7. olist_sellers_dataset
8. olist_geolocation_dataset

For my data analysis, the customer, product, order, order item data set are used. But there are some problems in the db schema. So, as part of data pre processing, data cleaning and at the bottom line to maintain the data quality, We created a refined csv from those data set that is used for later for further data analysis. To make refined csv, We considered only the successful orders which are delivered and shipped.

```
customer_order_product <- read.csv('customer_order_product.csv')
# 2nd row from the refined csv
refined_csv_2nd_row <- customer_order_product[2, ]
row.names(refined_csv_2nd_row) <- "Refined CSV - 2nd Row"
kable(t(refined_csv_2nd_row))
```

	Refined CSV - 2nd Row
order_id	f5eda0ded77c1293b04c953138c8331d
customer_id	68f2b37558e27791155db34bcded5ac0
purchase_year	2017
purchase_month	12
product_id	00088930e925c41fd95ebfe695fd2655
purchase_quantity	1
seller_id	7142540dd4c91e2237acb7e911c4eba2
price	129.9
freight_value	13.93
shipping_year	2017
shipping_month	12
product_weight_g	1225
product_length_cm	55
product_height_cm	10
product_width_cm	26

	Refined CSV - 2nd Row
product_name	p_3491
product_category_name	automotivo
product_photos_qty	4
product_category_name_english	auto
customer_city	franca
customer_state	SP

For example:

The above data row (displayed in vertical order) from the refined csv can be interpreted in textual format as below:

"Using the Olist e-commerce store, a customer Mr. '68f2b37558e27791155db34bcded5ac0' chooses a product '00088930e925c41fd95ebfe695fd2655' from product category 'automotivo' by seeing 4 product photos. Then this customer buys one piece of that product which order number is 'f5eda0ded77c1293b04c953138c8331d' from the seller '7142540dd4c91e2237acb7e911c4eba2' on month '12' and year '2017' at price '129.9'. The order arrives to the customer on month '12' and year '2017'. The shipment price is '13.93' which is billed by the courier service. The shipment price is calculated based on the product's dimensions (width '26 cm', height '10 cm', length '55 cm') and also the product's weight '1225 gm'."

After the refinement, there are total 99,878 observations and each observation contains 21 explanatory variables or features. These explanatory variables are formed into two groups: numeric(12 variables) and factor(9 variables).

External Data Source:

We created another csv that contains product names in english and for this We used external sources because in the existing data set the product name's length and category are present but there is no product name in the product data set. The external source link is given in the reference section. To collect data We used mainly product category and product name's length variables. We did that manually. The few rows from that csv are shown below:

product_category_name	product_name	product_short_name
automotivo	Anti-theft Lock Mcgard Screw Wheels Vitara 2017 TO 2022	Anti-theft Lock
bebes	Kit with 3 Pampers Premium Care P Diapers - 120 Units	Pampers Care
beleza_saude	Hair Stylo CR02 127V Hair Clipper - Mondial	Hair Clipper
beleza_saude	Colgate Slimsoft Black C/4 Toothbrush	Colgate Toothbrush

Interested Persons in Data Visualization:

As this is the data analysis and visualization on an e-commerce store named "Olist". So the intersted persons in the research questions and corresponding data visualization are as follows:

- Olist board of directors: Because, they can easily see from which type of products, the e-commerce store is earning most and which type of products has more sales and revenue in the marketplace. Also, which product category has the maximum sales coverage on overall store.
- Existing and Future Sellers Because, the sellers can learn about the customer's interest on specific products and also on product categories in this marketplace so that they can take better decision on selling products.

Data Visualization for Top Products:

Design Decision:

After the data analysis, We prefer to use waffle chart and spider chart because they are both visually attractive and informative. Also, it is easier to read and detect pattern within the data. The specific advantages are given below:

Waffle chart's advantage over Pie chart:

- The visual representation is more quantitative and countable than pie chart because, in a waffle chart, the counting of cells is easy instead of angle and area in the pie chart. Some cases, the angles in the pie chart are too small that they are very hard to interpret.

- The space utilization is higher because waffle chart uses square to cover the whole area in the chart.

Spider chart's advantage over Bar chart:

- To see the variation and figure out the common characteristics in data, it is better to use the spider chart rather than bar chart.
- In the bar chart, it is very hard to detect key assumptions, causes, effects, or patterns but using the spider chart, it is easy to detect pattern in data.

Principle Component Analysis:

The most important use of PCA is to represent a multivariate data table as smaller set of variables (summary indices) in order to observe trends, jumps, clusters and outliers. This overview may uncover the relationships between observations and variables, and among the variables.

[Source Ref: https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186]

In this research question, principle component analysis is used find out which variables have greater influence on calculating top product and also see the relationships among those variables and observe the data pattern.

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.8666 1.5274 1.0037 0.93212 0.66741 0.63237 0.58461
## Proportion of Variance 0.3871 0.2592 0.1119 0.09654 0.04949 0.04443 0.03797
## Cumulative Proportion  0.3871 0.6463 0.7583 0.85481 0.90430 0.94874 0.98671
##                           PC8     PC9
## Standard deviation     0.33128 0.09925
## Proportion of Variance 0.01219 0.00109
## Cumulative Proportion  0.99891 1.00000
```

From the principle component summary, it is seen that, the principle component 1 has 38.71% variance and the principle component 2 has 25.92% variance.

From the above scree plot, it is seen that there is an elbow formation on PC4 and also in the cumulative proportion variance explained plot, the first principle component that is over 80% is the PC4. So from the above two charts, We decide to take principle component 1 to 4 as the main principle component for further data analysis.

From the above chart, it is seen that there are two clusters. One cluster is top-left corner which includes product width, height, length and weight variables and another cluster includes purchase quantity, price, customer size and freight value. So it can be easily visualize that product dimensions and weights are highly correlated to each others and also same goes for the purchase quantity, price, customer size and freight value. Between this two cluster, the purchase quantity, price, customer size and freight value have much more impact and influence on the main principle components. As a result, We choose the second cluster(the purchase quantity, price, customer size and freight value) for further data analysis.

From the above group(Product Category) wise PCA analysis(PC1 vs PC2), it is seen that most of the product categories are close to each other and has positive dependencies on both PC1 and PC2.

Top Products on Olist:

From the PCA, We choose three explanatory variable or features to visualize the top products. They are:

1. Sales (purchase quantity)
2. Size (customer size)
3. Product Revenue (price)

The following charts show top products on overall store:

In the above spider charts are used to visualize the top products by sales, size and revenues. It is also seen a similar data pattern for the sales, size and revenues which means they are correlated and positively proportional to each others. In this sense, the score plot from PCA is cross validated with the spider plots.

The following two waffle charts explain sales and revenues based on top 6 categories. The top 6 categories are given in the chart legend.

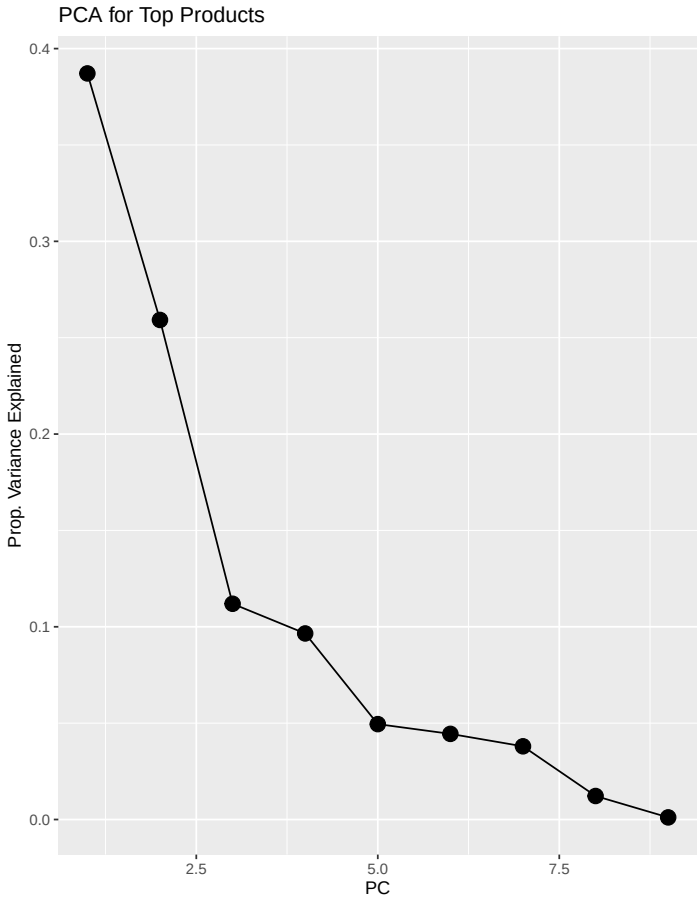

Figure 1.1: Scree Plot

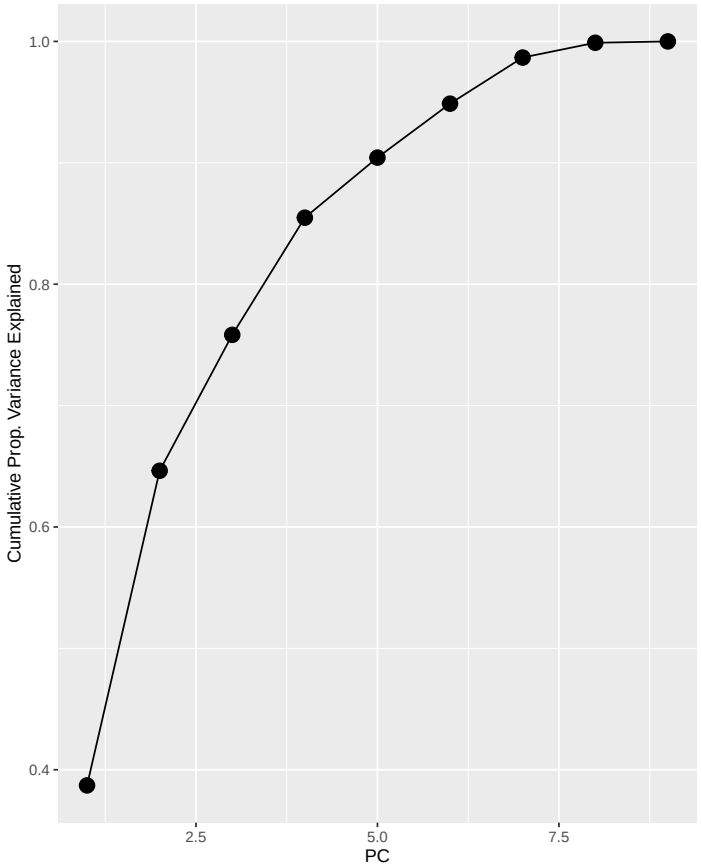

Figure 1.2: Cumulative Sum Plot

Figure 1: Priciple Components Selection

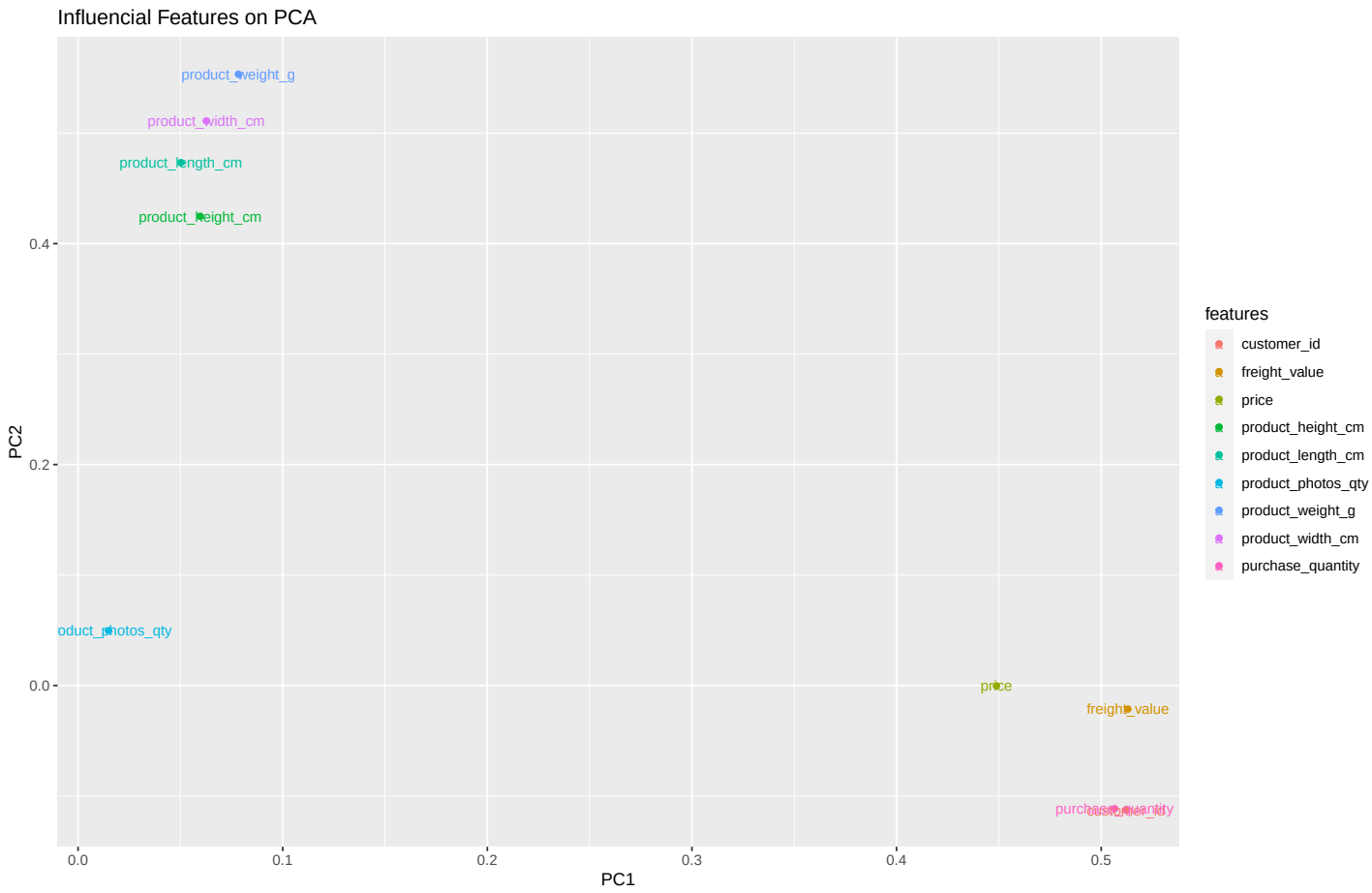

Figure 2.1: Score Plot PC1 vs PC2

Figure 2: Features that influence Principle Components

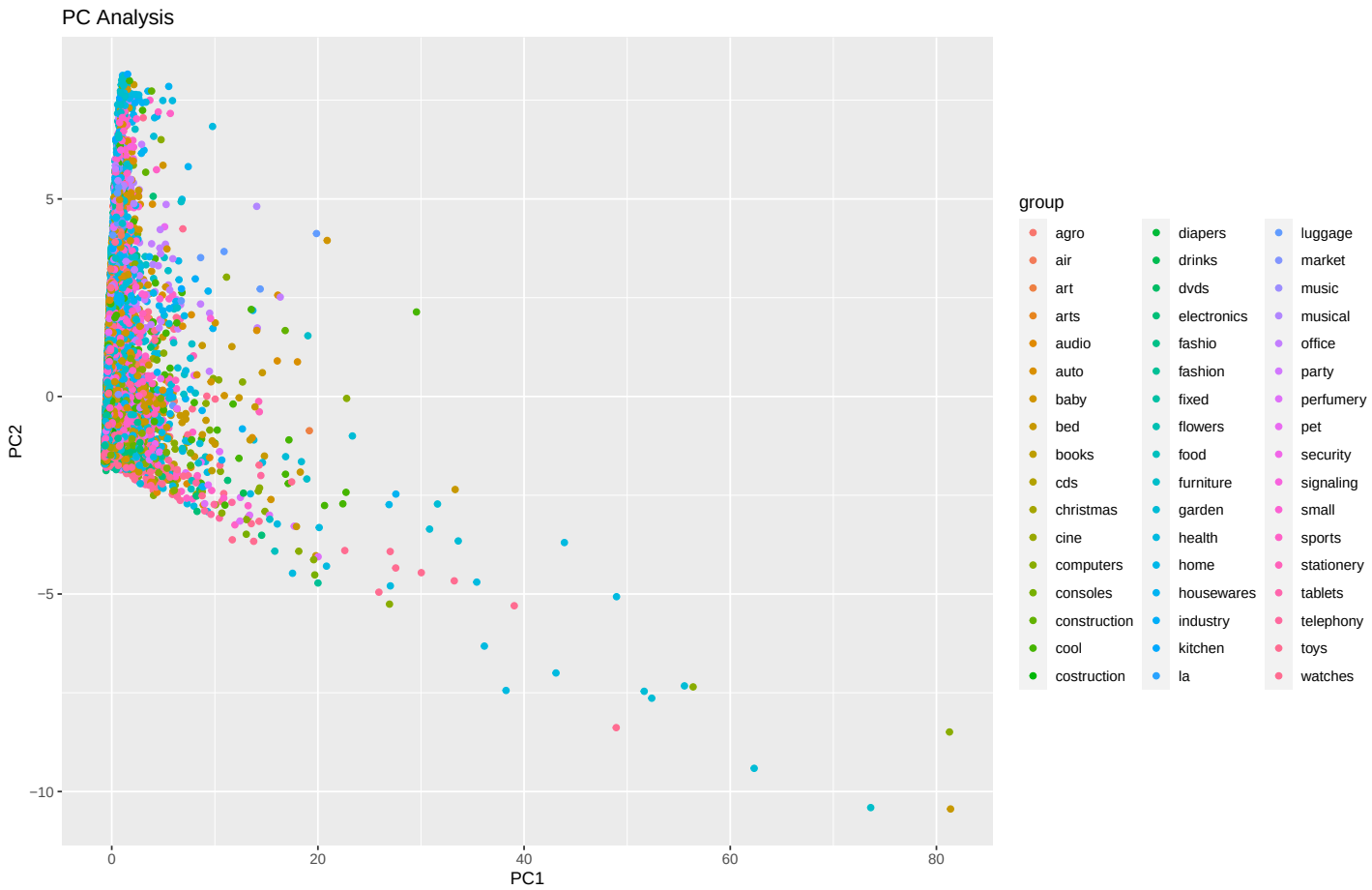

Figure 3.1: Plot PC1 vs PC2 data points by Category

Figure 3: Categories wise Principle Component's values

Figure 4: Top Products on Olist by Sales, Size and Revenue

Figure 5: Waffle charts to visualize Sales and Revenue by Category

In the 1st waffle chart, each square represents 1000 sales and in the 2nd waffle chart, each square represents 100,000 brazillian real.

From the above waffle charts, We select the common category named “Health and Beauty” for the further data visualization because in the “Category wise Sales” waffle chart, the “Health and Beauty” category is the second top most category that have higher number of sales. On the other hand, in the “Category wise Revenues” waffle chart, the “Health and Beauty” category is the first top most category which have the highest number of revenues among other categories.

The following data visualization is for the top products based on the category named “Health and Beauty”:

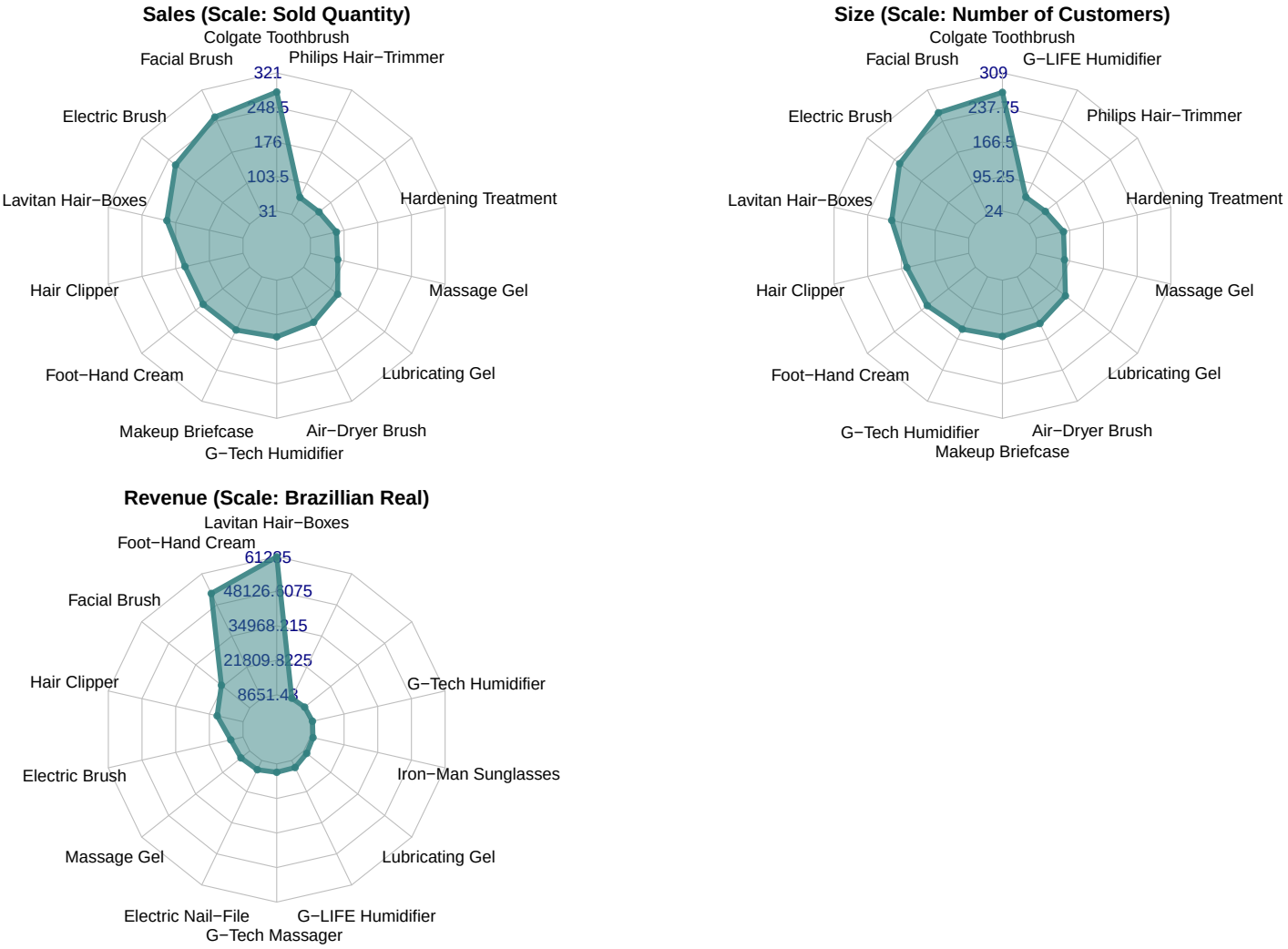

Figure 6: Categorized Top Products on Olist by Sales, Size and Revenue

In the above spider charts are used to visualize the top products by sales, size and revenues. It is also seen a similar data pattern for the sales, size but little bit different data pattern revenues which means they are almost correlated and positively proportional to each others. In this sense, the score plot from PCA is cross validated with the spider plots.

Conclusion:

From the above waffle charts and spider charts, it is concluded that the e-commerce store is earning most from the following product categories:

- bed_bath_table
- health_beauty
- watches_gifts
- sports_leisure
- furniture_decor
- computers_accessories

- housewares

and also the sellers who are selling products from the above categories are earning most. Specially the following products has the maximum coverage on the overall e-commerce store:

- Bunk Bed
- Queen Bed Sheet
- Gasoline Chainsaw
- Electric Trimmer
- Samsung 24 inch LED Monitor
- Pampers Care
- Agricultural Sprayer 2L
- Galvanized Wheelbarrow
- Colgate Toothbrush
- Lavitan Hair Boxes
- Casio Vintage Women's Watch
- Electric Brush

Learnings and Future Work:

After answering the research questions, we experienced about data quality issues on the data set. At the same time, we learned how to overcome these issues through data preparation and pre processing.

The important data quality related issues are given below:

- The order item data set does not have any quantity column and instead of this column, same order item is multiplied according to quantity. For example: a order item has 5 quantity then this order item has 5 copies in the order item set.
- There is no relationship between order item and review data sets. So, the reviews are order specific. In this case, if a order contains two order items which have two different products and a customer scored for one product in the review section, then the review score is applied for both products.

The future work on this data set will be as follows:

- Display which type of payment methods are mostly preferred by the customers. This can be grouped by customer state.
- Visualize the pattern of the product's dimension and weight on shipment cost. This can be grouped by product category, customer state, shipment year and month.
- Visualize how product parameters (for example: category, category, photo quantity, size, width, length, height, weight), customer review scores and geolocations play an important role to make a seller as a top seller among other sellers.
- Display category and geolocation wise customer's buying trend and pattern so that sellers can easily identify about the supply of product in specific category in specific location.