Will Heffernan

INST414

Professor Ai

# Data Exploration Project

## Introduction

The dataset I will be exploring throughout this project is the Brazilian E-Commerce Public Dataset by Olist. Olist Store is an e-commerce platform in Brazil that connects small businesses to consumers by providing them an online marketplace to sell their products. The dataset contains real, though anonymized, commercial data on ~100k orders placed with the Olist Store between 2016 and 2018. Information included in the dataset pertains to customers, payment methods, vendors, products, reviews, and more. The specific attributes of the dataset that I am concerned with are customer locations, types of items ordered, payment values, and order placement times.

To explore relationships between these attributes I will be looking at the files of the dataset regarding customers, orders, and order payments. I would like to investigate the relationships, if any, between customer location and the types of products they buy and how much they typically spend on an order.

Analyzing datasets of a commercial nature have a real world application as it can allow businesses of any size to determine what about their business is working, what customers they are reaching, and where the business should focus their resources. I intend to use the relationships I investigate during this project to provide insight on the most profitable customers and locations for Olist Store and their vendors.

## Data Preparation

In preparing my dataset, I first read in all the files that would be necessary for my analysis later on. The files I read contained information on customers, orders, items in each order, the category of products the items belonged to, and English translations of the product categories. To do some preliminary data cleaning, I viewed each data file and dropped columns that I knew would be unnecessary and would only clutter up my dataset. Such columns were related to things like delivery services and product dimensions.

After dropping unnecessary columns from each file, I merged each file into one dataset, beginning with the orders data file. After merging each file into a singular dataset, I did some more preliminary data cleaning by dropping duplicate columns that were a byproduct of merging files. Additionally, I renamed some columns to allow for easier selection of columns later on.

**Data Cleaning**

Following the merging of my selected data files, I conducted some more thoughtful data cleaning ahead of my exploratory data analysis. I began by dropping duplicate rows so that they would not skew any of my analyses. I also intended to drop any rows that had null values since I considered all of the attributes of my dataset important to my analysis. Finally, considering that I wanted to focus on relationships between cities that customers placed orders from and other attributes of the dataset, I thought I would try to exclude cities that only had a few orders.

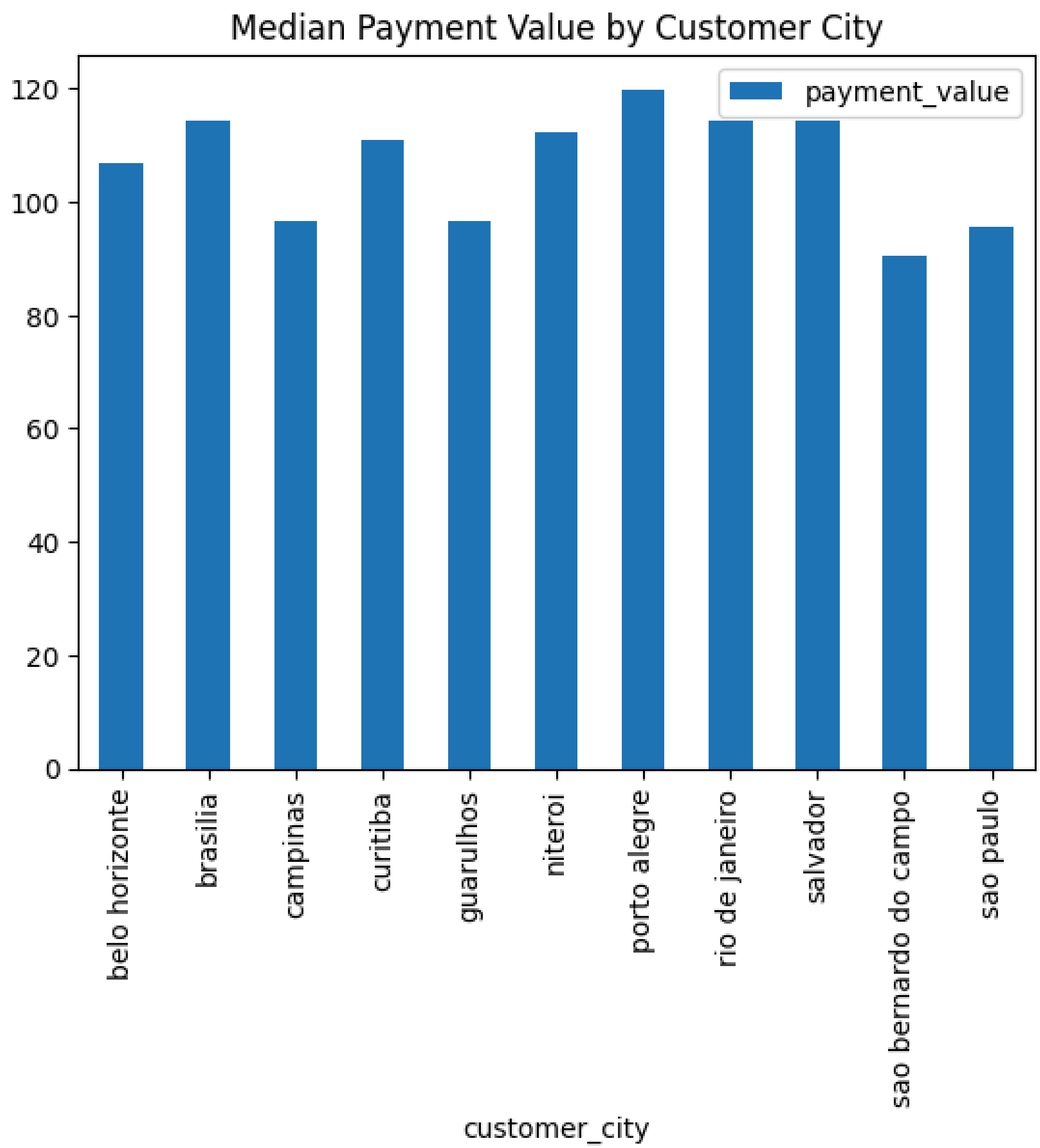
I addressed this by setting a threshold of counts for each city. Many cities had only a few orders. Additionally, including all of the cities in Brazil would not only make the analysis more difficult, but would be an obstacle when trying to draw broader conclusions from this commercial data. I decided on setting a city count threshold of 1000, dropping all transactions from cities that did not have over 1000 orders total.

After dropping transactions from uncommon cities, the data cleaning process was complete and my dataset was left with 41,828 transactions with 12 attributes per transaction.

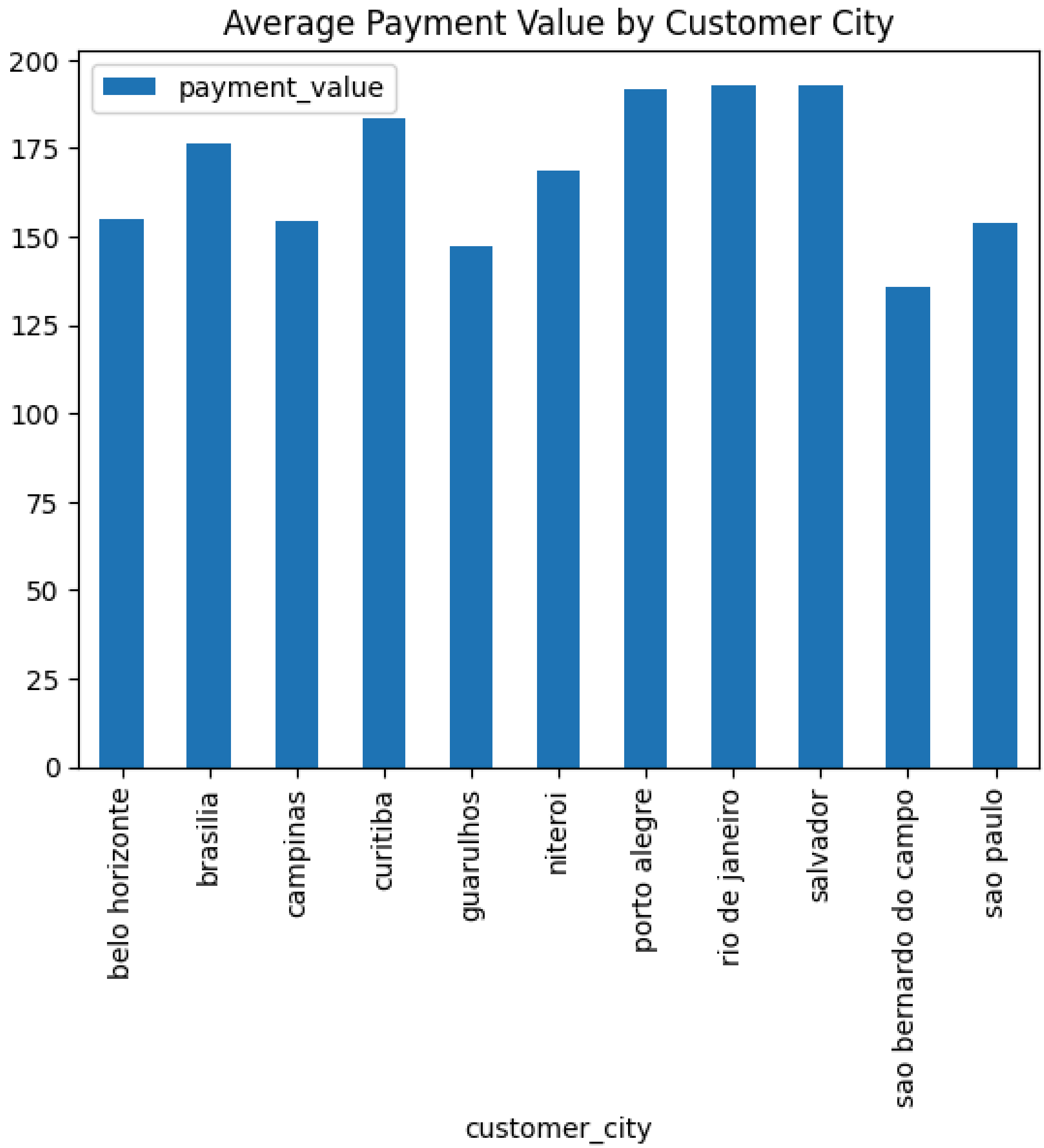
**Exploratory Data Analysis**

I began exploring my dataset through grouping it by customer cities. I then aggregated data and determined the mean and median payment amount in orders from each city, the number of orders placed in each city, the most popular product category to order in each city, and the conditional probabilities of ordering products of a certain category, given that the customer was in Sao Paulo.

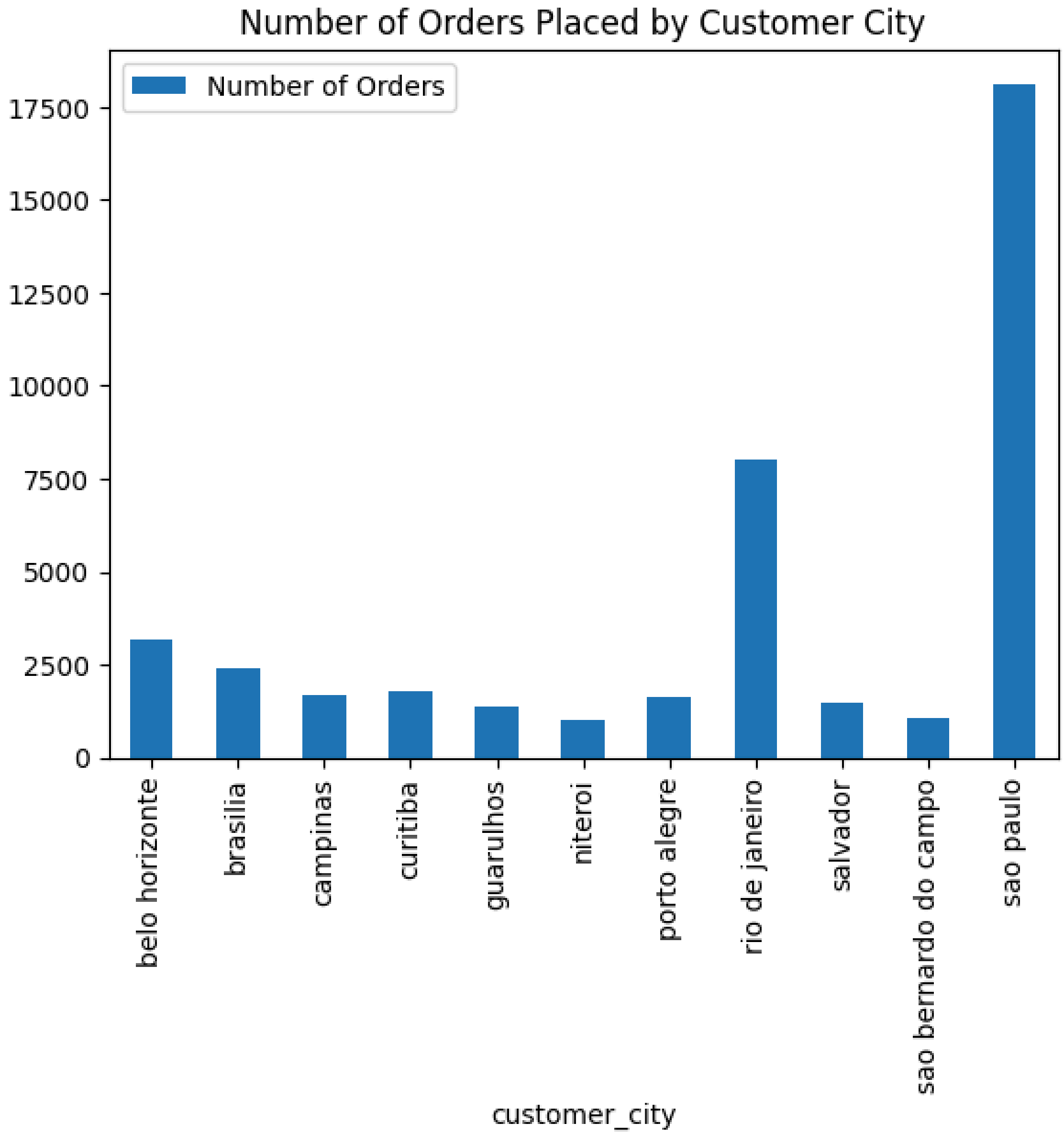
I began with finding the median payment value for each order by city. I accomplished this by grouping the dataset by customer city and applying the median function to the payment value attribute. Based on the results, Porto Alegre had the highest median payment value with 119.94, while Sao Bernardo Do Campo had the lowest median payment value with 90.35. The median payment values of each of the most commonly-ordered from cities are shown below.



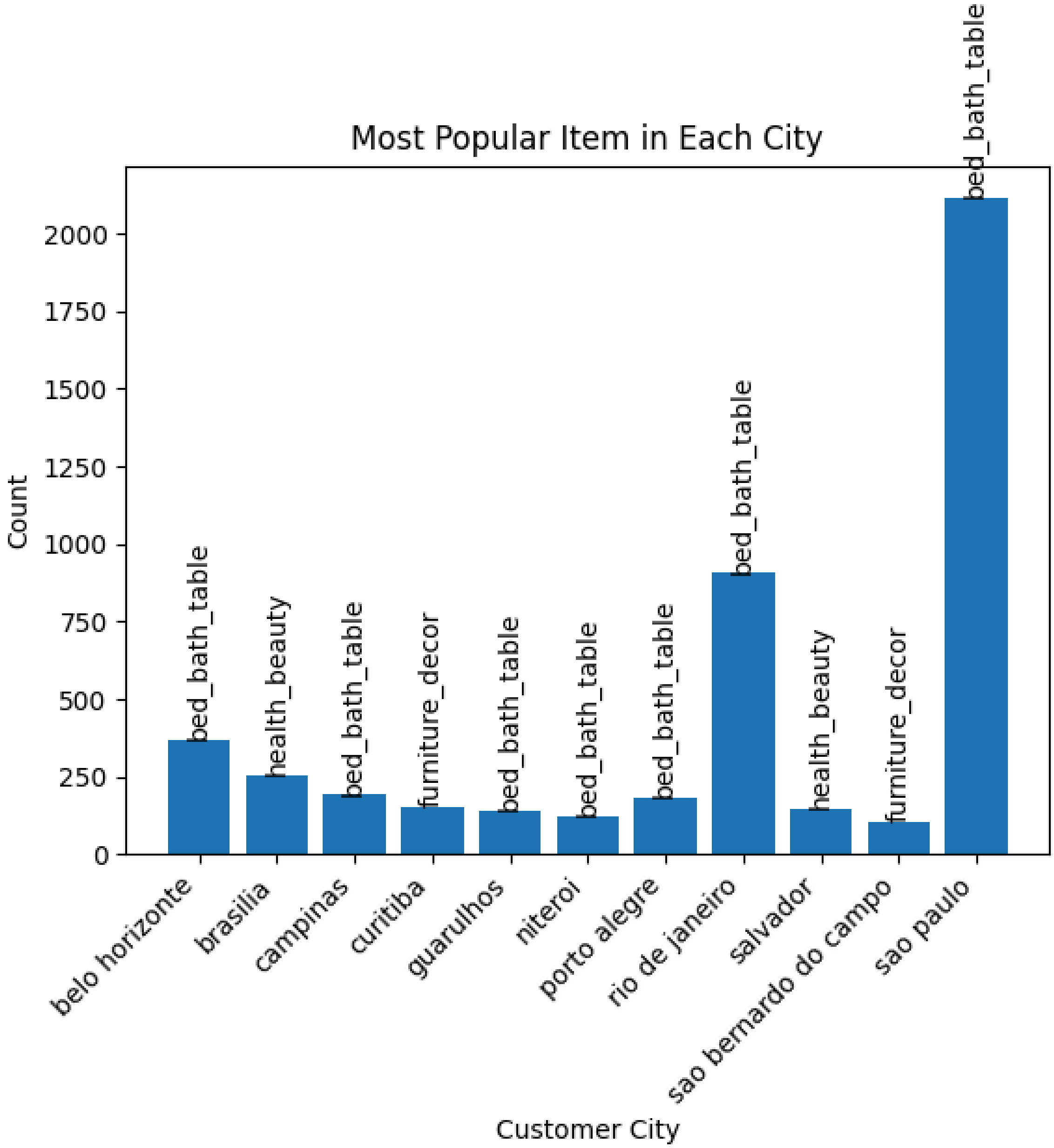
I then used a similar method to find the average payment value by customer city, this time applying the mean function to the payment value attribute after grouping the data by customer city. Based on the results, Rio de Janeiro had the highest average payment value with 193.00, followed closely by Salvador at 192.93 and Porto Alegre with 191.66. The lowest average payment value was from Sao Bernardo do Campo at 135.69, with the next lowest average payment value in Guarulhos at 147.26. A visualization of mean payment values by customer city is shown below.



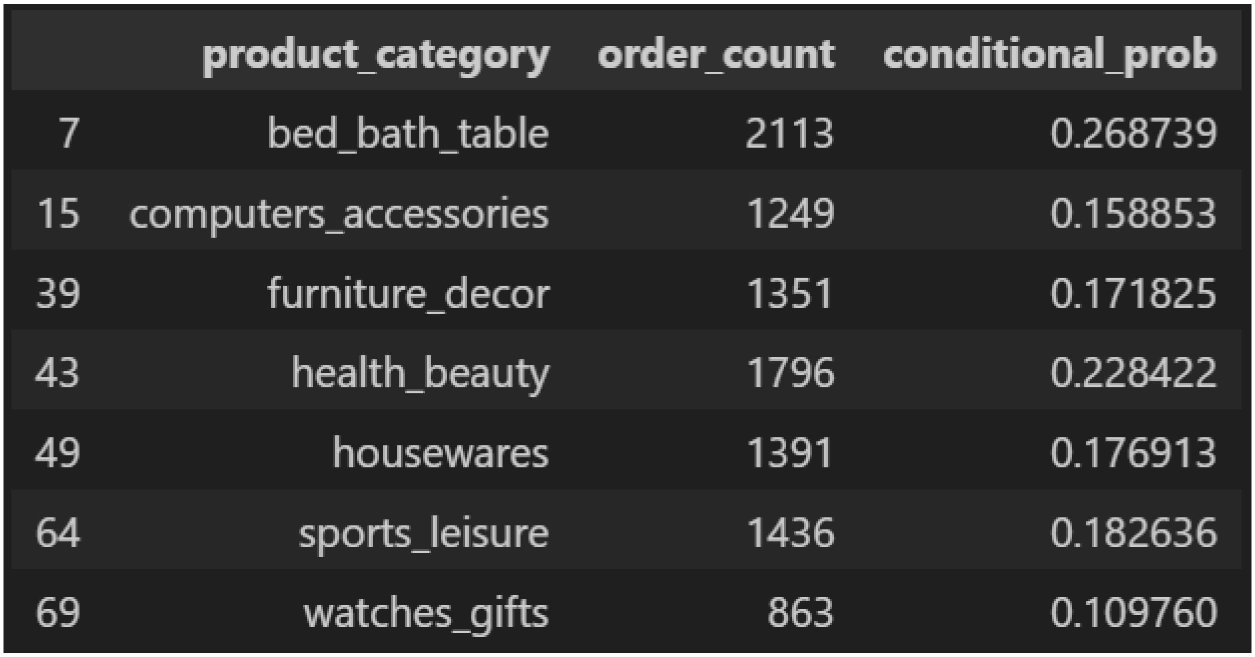
Following the calculation of mean and median payment values by customer city, I continued by analyzing the most popular customer cities. I found the number of orders placed per city by grouping the data by city and applying the count function to the order ID attribute. Sao Paulo and Rio de Janeiro together made up a majority of the orders from the 11 most popular customer cities, with 26,153 orders, or 62.5% of the total. These two cities were followed by Belo Horizonte with 3,185 orders and Brasilia with 2,420 orders, while the rest of the most popular customer cities had less than 2,000 orders recorded. A bar chart of the number of orders placed by customer city is shown below to compare the order counts between the 11 most popular customer cities.



Next, I decided to investigate the most popular categories of items in each customer city. I found the most popular item in each city by first grouping the dataset by city and product category and applying the size function to display the count of each product category for each customer city. I then located the most ordered product category for each city by applying the idmax function to the count of orders per product category per city. Based on the results, a majority of the 11 most popular cities' most ordered product categories were bed, bath, and table products. Two cities had health and beauty as their most popular product category and another two cities had furniture and decor as their most popular product category. A visualization of the most ordered product category in each city and the corresponding count is shown below.



After looking at the most ordered-from customer cities from the dataset, I decided to focus on the city that had the most orders by far, Sao Paulo. I decided to apply conditional probability calculations to each possible product category to gain some more insight on what types of products customers in Sao Paulo are ordering. I accomplished this by first grouping the dataset at large by product category, then grouping the Sao Paulo subset of the data by product category in each order.  I then applied a conditional probability transformation to the Sao Paulo subset using the counts of product categories in orders from Sao Paulo and in the entire dataset. I then filtered the Sao Paulo subset to display product categories that were assigned a conditional probability of greater than 0.1. The dataframe below displays product categories, the total number of that category ordered in Sao Paulo, and the probability that an order contains that product category, given that the customer is from Sao Paulo. The bed bath and table product category had the highest conditional probability at 0.269, followed by the health and beauty product category at 0.228 and the sports and leisure category at 0.183.



| | product_category | order_count | conditional_prob |
|---|---|---|---|
| 7 | bed_bath_table | 2113 | 0.268739 |
| 15 | computers_accessories | 1249 | 0.158853 |
| 39 | furniture_decor | 1351 | 0.171825 |
| 43 | health_beauty | 1796 | 0.228422 |
| 49 | housewares | 1391 | 0.176913 |
| 64 | sports_leisure | 1436 | 0.182636 |
| 69 | watches_gifts | 863 | 0.109760 |

**Conclusion**

Based on the results of my data exploration, I have gained insight into a few trends within the Olist dataset of orders and their customers. First it seems that the most significant base of Olist customers resides in Sao Paulo and Rio de Janeiro. While neither of the cities had the highest mean or median payment values for their orders, Rio and Sao Paulo together made up more than 62% of the total orders placed among the top 11 most popular cities to order from on Olist. Based on the cities' prominence in the number of orders on Olist, I would recommend that advertising resources are distributed accordingly so that the online marketplace can reach their largest bases of customers. Additionally, further analysis may reveal why there are not as many orders being placed in other cities, whether it is due to differences in population, lack of advertising in those areas, or lack of access to the needs or means of ordering products on Olist.

Another insightful relationship between customer cities and database of Olist orders is the difference between the mean and median payment values for orders. Based on the median payment values analysis, I would recommend that Olist advertises larger quantities of products or a larger variety of products to cities such as Porto Alegre. The high median payment value may be indicative of a city's customers tending to place orders of larger quantities when they do buy from Olist. On the other hand, higher average payment values may be indicative of some outlier orders stretching the average, meaning that a few orders in particular cities may be exceptionally expensive. Based on this interpretation, I would recommend that Olist markets more luxury and high-end products to cities with higher average payment values, such as Rio de Janeiro, Salvador, and Porto Alegre.

In a similar vein of interpretation, the analysis of the most popular product categories in each customer city indicates that many of the orders placed on Olist fall under the bed, bath, and table category. Based on these results, it seems that Olist customers already tend to order these types of products over other product categories. Based on this insight, I would recommend that Olist diversifies their selection of products in this category to bring in more sales. Customers who are already interested in this product category may find that new items suit their needs or wants and may choose to purchase the new items as well.

Finally, based on the conditional probabilities of product categories being ordered by Sao Paulo customers, I would recommend that Olist anticipates the demand of particular items in particular cities. For instance, if Olist does or will have a warehouse based in Sao Paulo, it would be beneficial to keep the warehouse stocked with amounts of product categories related to the probability that an order from Sao Paulo will contain that product category. Bed, bath, and table products, for instance, should be the product category with the most stock at a Sao Paulo warehouse, followed by health and beauty products. The practice of keeping the most demanded items in stock closer to where they will be ordered could help cut down on the time it takes to deliver items, enhancing logistics and customer satisfaction, as well as reducing transportation costs and the environmental impacts that comes with it.

Overall, there is still analysis that could be done to gain valuable insight from the data that Olist has on their customer's orders. Anticipating orders of particular products in certain cities, for example, could be applied to all major warehouses that Olist might possess. Other, more in depth analysis could also reveal why exactly the median and mean payment values vary across cities. This might help Olist in determining what markets reside where and advertise accordingly. This project showcased some of the applications of data manipulation and analysis for a commercial impact.