Analysis of Brazilian e-commerce data

Pedro Henrique Figueiredo Magalhaes

Data Analysis capstone project at BitAcademy

Cohort December 2023: Data Analyst Track

Objectives

The project will answer questions about the e-commerce markets in Brazil using a dataset comprised of 100.000 orders made at multiple online marketplaces in the country. The questions I want to answer are:

1. **Which cities and which state buy the most items online?** 
   1. Hypothesis 1a: Brazil is a big country and I expect that people living in bigger cities in the southeast of the country will purchase more items online than in other cities.
      1. Hypothesis 1a is true. Sao Paulo is by far the state that orders the most items online, followed by Minas Gerais and Rio de Janeiro.
      2. Methods:
         1. Merged queries to integrate the customer\_state into the orders dataset.
         2. Created a measure to count number of orders by state (orders\_by\_state).
         3. Imported json file to create a shape map of Brazil that matched the abbreviated state names (map keys).
         4. Created a measure the calculate the amount each state spends in R$ (Brazilian reais).
2. **Which is the most popular form of payment?** 
   1. Hypothesis 4a: Credit cards will account for most of the payments made online. In Brazil we call invoices “boleto”, and I suspect many people will use it because it was a popular payment method many years ago. Nevertheless, credit cards are likely to be the preferred method of payment nowadays because more and more people are getting credit cards.
3. **Who is the top seller in the online marketplace and from where does this seller come from?** 
   1. Hypothesis 2a: Top sellers are in the same cities where people purchase the most items online.
      1. Hypothesis 2a is true. There is a match between the cities top 10 sellers and top 10 customers come from (6 out of 10). Nevertheless, there are discrepancies since smaller cities like Ibitinga and Maringa have a lot of sellers.
   2. Hypothesis 2b: The market is concentrated. There are at most 3 top sellers, and the rest of the market is made up of smaller players with a smaller share of the online marketplace.
      1. Hypothesis 2b can’t be tested in the dataset because there’s unique seller\_id values and they are all distinct, indicating that either all sellers were indeed distinct or that even when orders came from the same seller, the seller\_id value given was distinct each time.
4. **Which categories of products are most sold and the least sold in the online marketplace? [Top 10]** 
   1. Hypothesis 3a: People in Brazil are comfortable purchasing consumer packaged goods online, but for other types of products people prefer to go in store.
5. **When people buy something expensive online, do they use their credit cards, or do they prefer invoices? How much do people pay on average?**
   1. Hypothesis 5a: I suspect that people that purchase something expensive online might prefer to pay via invoice rather than credit card.
      1. Hypothesis 5a is not true. Even for expensive items people still prefer credit card payments when shopping online.
   2. Hypothesis 5b: People spend on average less than 1.000 reais on online purchases. For expensive items, they prefer to go in store.
      1. Hypothesis 5b is not true. There are significant values being transacted online, such as orders valued at R$ 13.700,00. Therefore, people will spend online regardless of the value of the purchase. This could be due to people perceiving online shopping as safer than I assumed they would.

Dataset

The dataset has 8 csv files (1 of them is a translation of product categories from Portuguese to English which I will join in the products table), with features ranging from order status, product prices and categories, information on the location of sellers and buyers and the dates of each transaction.

My plan is to use the STAR model in Power Bi to link the dataset through their common columns, which are ‘order\_id’, ‘seller\_id’, ‘zip\_code\_prefix’ and ‘product\_id’. I expect some files will not be useful apart from a column or two, therefore I expect to perform manipulations and join operations to create new tables containing the data that is most relevant for this specific analysis. I expect to use DAX expressions when calculating average prices for orders (gather all order\_id values that are the same, thus representing a single order, then sum the payment values).

The files attached:

*olist\_customers\_dataset.csv, olist\_geolocation\_dataset.csv, olist\_order\_items\_dataset.csv, olist\_order\_payments\_dataset.csv, olist\_orders\_dataset.csv, olist\_products\_dataset.csv, olist\_sellers\_dataset.csv, product\_category\_name\_translation.csv*