Brazilian e-commerce dataset visualization

Faizuddarain Syam,Junaid Mundichipparakkal

Abstract— E-commerce is becoming one of the most popular business practices in the latest century. In this assignment, we are creating a tool for an e-commerce company to understand its customers and sellers. This tool will help the company to understand the trends in products, sellers, and customers. These trends can be then used by the company to increase its revenue. The Dataset used here is for a Brazillian e-commerce companies year 2018 transaction details and using this tool the company will be able to analyze the trends in states, cities, and each zip code location.

Index Terms—E-commerce,visualization,

1 Introduction

E-commerce is one of the important business practices in this generation. It can be noticed that nowadays most of the people prefer buying almost anything online rather than going to physical shops, as the services are getting better and better. It can also be observed that nowadays E-commerce is one of the most important stocks in this world. Various e-commerce with different backgrounds and niches are emerging in almost every country in the world. This creates a competitive ecosystem which forces online market companies to provide a better user experience on their platform. To achieve such goal, the company must fully understand the background and behavior of their users, as well as recognize the root problem they are facing and act accordingly. By coming up with a simple yet well-thought business decision, it could make a huge difference on why a particular e-commerce platform is preferable than the others.

In this paper, we will demonstrate how an e-commerce business can have deeper insights into their data. This is done by highlighting important information based on their users (customers and sellers) and the orders recorded in the system. The goal of this assignment is to create a tool to help the e-commerce company to understand their customers, sellers, and the product that are sold.

This tool can be used by the e-commerce company to get an insight of their users by featuring the number of customers and sellers in each location and investigate why it is not popular in other locations. By doing this, the company will understand the trend of demands for a certain product category in an particular area. It then can analyze the locations of sellers and analyze how to promote or sustain more supplies in different locations, hence it will open up more potential customers and create a healthier competition amongst the sellers. Furthermore, it can reduce the delivery time spent and charges which will motivate or induce the customers to do more transaction on the platform.

1.1 Problem Description

The e-commerce data contains information mainly about the orders recorded in the system. To support this, it also contains data about the customer, seller, and product that participate in the transaction. It is difficult to make an understanding of how to improve the business for the company. In this data set, the data is from different locations and this can be to analyze the information of each location using a spatial plot to find interesting information about the deliverables. Based on this information and the main idea of our goal mentioned before, the problem is break-downed into sub-problems. Hence, the visualization we will provide must answer the following questions:

- How is the distribution of the customer's and seller's location in the whole country or in a particular state or city?
- Which product categories are the most saleable or marketable in a certain area?
- How does the location of the customers and sellers effect on the delivery time?
- Comparison of turnovers based on location and product categories.

The e-commerce company will be able to answer these questions to improve their services or discover solutions for their existing problems. Hopefully, this will make the platform provide a better user experience, resulting in more engagement and traffic, and also expanding their user-base regions. Consequently, it will increase the company's revenue and foothold in the country to gradually making this one of the most popular e-commerce platform in this region.

2 Data Analysis

2.1 Domain Data Specification

The data sets used for the assignment is a Brazilian e-commerce public dataset of orders made at Olist - an online e-commerce site for sellers, that connects merchants and their products to the main marketplaces of Brazil. The dataset has data about 100k orders ranging from 2016 to 2018. The data sets contains multiple dimensions such as order timestamps, prices, payments, products, reviews, and customer's and seller's location. This is real commercial data that has been anonymized.

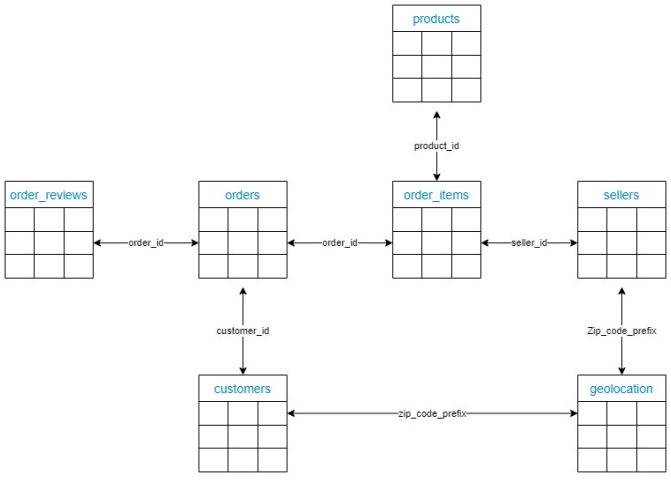

Fig. 1. Data sets network of the given data

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

As shown in Fig. 1, there are 7 data sets that are all connected by variables. This will be very helpful later on by grouping multiple datasets based on these variables. The data sets which are used here are orders, order items, order review, sellers, products, customers. The orders data set contains information about the orders such as customer's ID, order status, and timestamps for the order's purchase, and The order item contains information about the ordered item and seller id and product id. the products data set contains information about the products categories, item code, and etc. sellers data set contains information about seller location and seller products codes. customers dataset contains information about the customer and their location as well the order id. order review dataset contains information about each product's rating and its user's comments..

2.2 Data Abstraction: What

The data set used for this analysis is network type of data which is linking 7 tables which are linked each other with difrrent primary keys. The keys can be observed from Fig. 8.

The dataset consists of 47 columns. Each of these columns contains either spatial, numeric, time, text, or categorical data. 2 columns contains Spatial data, 21 columns contain categorical data, 9 columns contains numeric data, 9 columns contain Date and time data and 2 columns contain ordinal data. We also have a hierarchical connection between states, cities, and zip codes.

Variable	Nunber of Unique values
Product Catgeories	71
States	26
Cities	1412

Table 1. Number of unique values in mainly used categorical values

Table 1 contains the number of unique values for each of the categories which are used in the tool implementation.

3 Task Analysis

To achieve our overall goal for our visualization, we need breakdown our objectives into tasks based one the information provided in the data set.

3.1 Domain Specific Tasks

Based on our explanation before, our main focus in our visualization is on the locations of the customer and seller. From here, we will breakdown how we made our decisions on how the end visualization will look like by defining our tasks.

- Task 1: The distribution of customers and sellers in each location.

  To visualize the distribution of the customer's or seller's location, we will be plotting a geo map of Brazil. In this map, it will show areas where the customer or seller exist. In addition,it will provide more detailed information based on that particular point. There are three key location information in the data set: state, city, and zip code area. To plot this data into a map, a geolocation latitude and longitude is provided in the data set, which is based on the zip code area. Since these location groups are hierarchical, each location group will be a visual point for the location group above it. Which simply means, when selecting a state, the city will be shown as points on the map, and when a city is selected, the zip code areas will be shown.

  The end result should provide a map, starting the view point as the whole country of Brazil and states as the visual points, and an interaction for the user to filter the map based on a state or city. To do this, two drop down list consisting of the list of states and cities is required. Changing the values on this list will result in updating the map based on the selected area. In addition, an 'All' option need to be added in to the drop down list to give an option filter the map based on all of the states or cities.

  To switch between viewing the customers or sellers, since they share the same map, another drop down list needs to be added.

- Task 2: Filtering the viewed data based on a product category.

  Provide a drop down list of product categories. When a product category is selected, the map will only show locations where the product category is ordered or sold

- Task 3: Show popular products based on selected location

  Divided by two types: most ordered, highest rating It will show a bar chart of top 5product categories for each two in addition it shows a mean value for all of the data in the selected location

- Task 4: Show the delivery time and estimated delivery time based on location

  The plot should be able to visualize both the actual and estimated delivery time in a distribution plot. This is to compare the difference between both values. The distribution plot also provides information for the population of orders whether they are mostly delivered shorter or longer than expected. This distribution plot also updates for each update on the location filter, so it will show the population based on the selected location.

- Task 5: Show the amount of orders over time based on location and product Because each order has their order timestamps, we can visualize the history of orders shown for each month. While the timestamps of each month will be the x axis, the amount of orders recorded will be the value for the y axis. In addition, these orders are always updated based one selected location and product category

- Task 6: Show top sellers origin based on location and product The sellers location origin will be shown as the top 5 location, separately based on city and state, that an order came from a customer within the location selected. Not only updated by location, but also the product category should be able to filter these data so the plot will be able to give valuable information of a particular product supplies for a location.

- Task 7: Give hue color to the map the average order price and rating

  To come up with a better decision, the information of the average money spent on orders and the average review score are putted in to consideration. These values should be shown as the color difference on the map, supported by a legend. These values will be next to the customer-seller drop down list to give option whether to view the information of the average price or review s

3.2 Task Abstraction: Why

- Task 1: The number of sellers and customers in each location

  This helps the company to understand the popularity of the e-commerce website in each region. and the company can allocate the recourses based on the number of customers to provide better services.

  As a result, the e-commerce company can understand which areas has potential market and which areas need to be improved.

- Task 2: Filtering the viewed data based on a product category.

  The data set we have is very huge. Filtering the location distribution of the seller and customer based on a product category is an important feature to achieve our goals. It can provide information for the supply and demand of a product type for each level of location selected. This results in understanding the distribution of demands for a product and improve supply near that location or analyze areas with high supplies and market the product around that area.

- Task 3: Show popular products based on selected location

  This data can be used to add more similar products or show recommendation for new products from the company. example like amazon essential.

- Task 4: Show the delivery time and estimated delivery time based on location

  This will help to analyse each location information and help to add more delivery partners or to find why the delivery time is effected.

- Task 5: Show the amount of orders over time based on location and product This will help to understand the total revenue generated and if there is steady drop in revenue in a location. this is serious problem and needed to handled carefully.

- Task 6: Show top sellers origin based on location and product Sellers are the backbone of the e-commerce business hence this analysis will help to add more sellers. Since product based analysis is also available the company can think about adding more lesser for same category or new sellers for new category which are trend in different locations.

- Task 7: Give hue color to the map the average order price and rating The price option is done to analysis the trend of what kind of product is ordered in general for example if the circle is big and the colour is less hue. this shows customers use this to buy common products. while if the color highly hue then it means the customers use this website to purchase expensive items. The review is key feature the company gets from a user which explains the scope of the company's future.

4 Visualization and Interaction Design: How

4.1 Data Prepossessing

example for faiz to refer try to as much as possibel to get page count: we did fro state city

Before looking at the visualization design decisions the dataset needsto be examined. The initial dataset was delivered in a human readablestructure. This made it impossible to make a programming languagerecognize the structure of the dataset. It could not relate which statisticsbelonged to which polling station. This meant preprocessing had totake place on the data before it could be loaded into the visualizationtool. All the empty cells had to be removed, the dataset had to betransposed, the symbols"(","),"."had to be filtered out as d3cannot handle these symbols in the input data. Lastly, the longitudeand latitude had to be added to each postcode to be able to plot thepolling station at the right location on the map. All this preprocessingtook place in python [3] with the help of the library Pandas [?]. Anadditional file with the longitude and latitude per postcode was used aswell. The Python code and the additional file used can be found in theproject folder.

4.2 Encoding and Interactions

This section presents the visualization design of the tasks based on the problem description. For every task abstraction, the choices for visual encoding and interaction design are explained. Furthermore, it illustrates how the design enables users to perform the tasks and answer the question [2]. Each of this plot have house hover information encoding done where the corresponding values will be displayed for user.

- Spatial plot for customer/seller location per location

  Visual Encoding: The map of Brazil is used here since in a spatial encoding and we use a circle of different sizes to denote the number of customers or sellers. The bigger the circle means that the higher the number of customers or sellers in that location.

  Interaction: When the application is started it shows a plot of different circles is of different sizes to denote the number of customers in each state. The user can select a state to get a detailed view of that state where you get an updated plot of the number of customers in each city in that state. If the user selects a city then the update map view will show the detail of each zip code. There is an option for the user to select between the number of customers and sellers. each option will provide the respective counts.

- Spatial plot of review score or price by customers

  Visual Encoding: The same map used in the visualization of a spatial plot of the number of customers or sellers have a color code that ranges from color_ to _ based on price and review score. The color code values are shown as a gradient plot towards the right side of the map.

  Interaction: The user can select the color code of the plot to be based on either price or review score. this also changes its values based on location selections like state and city.

- Most popular product Category based on location

  Visual Encoding: In this task, we are plotting a histogram based on the top 5 categories values. using the count of the price of each category of product.

  Interaction:

  Here the top 5 products will change according to the selected location (city, state).

- Delivery time distribution for each location

  Visual Encoding: This is a histogram and distribution plot where the time of actual delivery time and estimated delivery time are plotted. The color _is used for actual and _ is used for an estimate.

  Interaction: The user can select different locations (state, city), and based on this the given plot will change respectively.

- Count of data based on location and Product Category

  Visual Encoding: This plot shows a bar plot(**not sure with namae, dono what y and x valuea are**) with shows the count of customer, sellers and orders.

  Interaction:

  The user have a complete option to choose based on location(Sate,City) or a specific category.

- Tend of revenue based on location and product category

  Visual Encoding:

  In this task we plot a histogram based on the time and amount of money spend by the customers.

  Interaction: The user have a complete option to choose based on location(Sate,City) or a specific category.

5 Realization

The visualization tool in this assignment is developed using python. The project had two phases. The first phase was Data Preprocessing. for this step, we used a jupyter notebook. and for the second phase of the project, we used the preprocessed dataset to build the visualization tool in python [1].

We are using the pandas library for data manipulation and analysis. pandas is one the most commonly used library in python for this task. if the company is planning to create for a whole level of the bigger dataset it will be advised to use NumPy instead of pandas, this will help to improve the performance of the tool significantly. The Map used is from the Mapbox library. this library provides the maps of any location free for students and another reason for using this library can be because it is a rapidly growing service and well cheaper compared to google maps and other services. This will make the future development of this application cheaper. one of the most important library used in this project is plotly. this library provides graphing, analytics, and statistics tools. We are mainly using this to plot the required encodings.

The reason for using plotly is mainly because it actually using is this library to make it easier to implement visualizations and also in the background this library still uses D3.

In order to make this python code a web project, we are a flask library. This is the framework is used for developing web applications in python. The reason for using flask is only because it is simple compared to Django, Tornado and etc. But one of the difficulties of the flask library is it will lead to developer traps so if the scale of the application is going to have more features it is advised to choose some other web development frameworks.

6 Use Cases

In this section i am using the screen shots from the visualisation and information from these Figures are analysed.

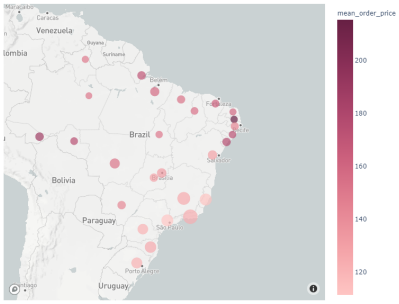

Fig. 2. Customer distribution odf each state

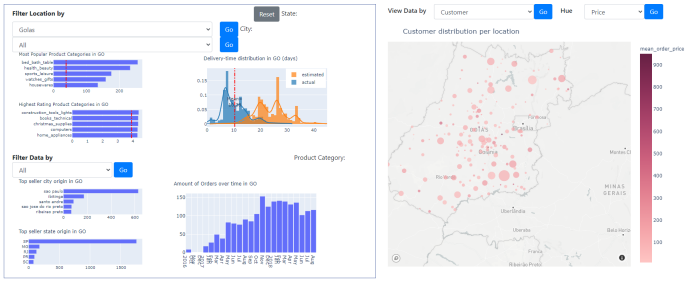

Fig. 3. Customer distribution state golas

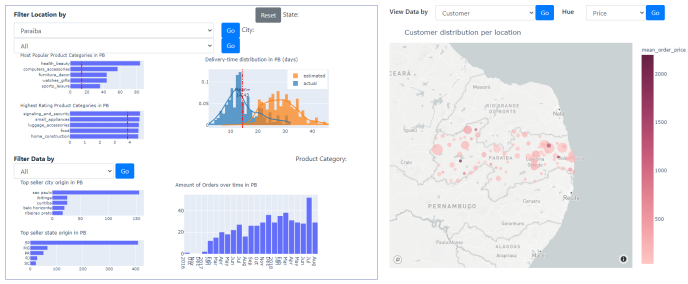

Fig. 4. Customer distribution state parabia

- analysis of several customers and number of customers and sellers. using task 1 we will able to see the number of customers from each location using Figure we will be able to understand the popular cities where this website is used. using this information

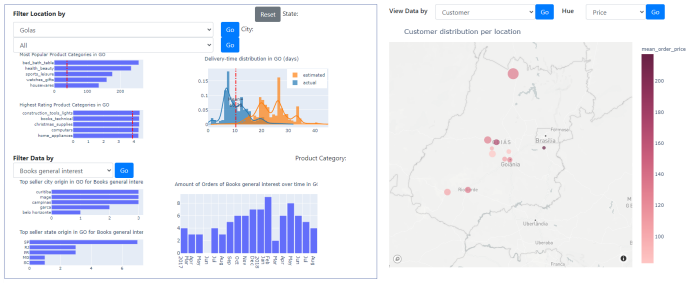

Fig. 5. Customer distribution state golas product category books

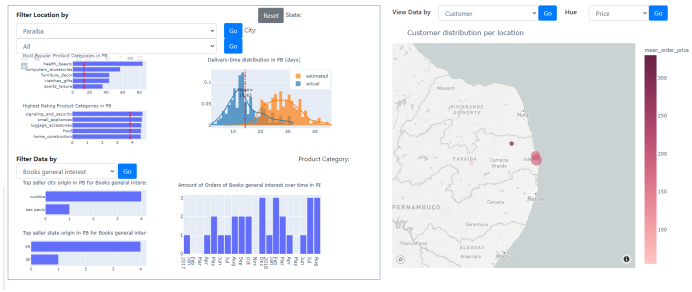

Fig. 6. Customer distribution state parabia product category books

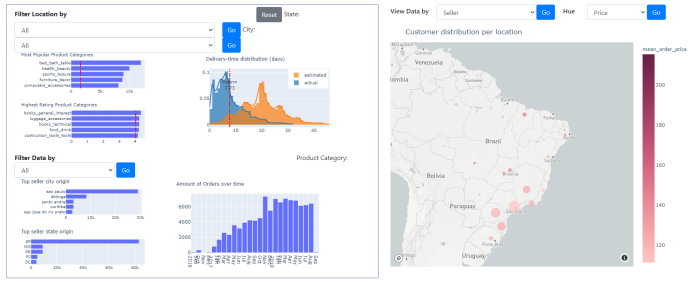

Fig. 7. Sellers distribution all brazil

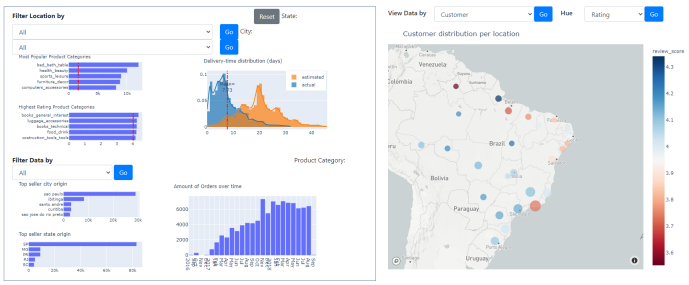

Fig. 8. Customer distribution of Brazil with color code based on review

the e-commerce company can make plans to make plans for advertisements and the number of employees. once this analysis is done it will help to the goal of making this e-commerce website popular in the country of brazil.

- price review will help the user to understand the goals of the user for a region. for example from the figure we know state customer user base is high but the price distribution is less compare to city B where the customer base is less while the price distribution is high. it will help the company to analyze this promote more expensive ports in the region And commonly used products in the region B.

- the delivery time analysis shown here for each city will explain the estimate and actual time. from the figure based on city A and city B, it is always understood the estimate is always higher and all the delivery is actually done fasters than the estimate.

  this analysis comparing cities will help to improve the delivery time in each city by comparing the service the services they are using while it can also be used to improve the individual estimate of each city . so the users who need the product fast will also make order instead of going to a shop.

- The revenue analysis here gives us a graph based on location and each product. from the figures given above it is noted that city a and city b revue total changes and the growth of each revenue while it can also compare the revenues from each product category it will help the company to make new ideas for different product type promotions based on cities for example, bring a similar type of product to the market that is commonly used in that location

- The top 5 product comparison between state and cities will help to make overall recommendations for the company to find in the location base this can also help the company to bring more sellers of those categories to the market to make this e-commerce a solid foundation for those products. this will also help to bring up more new categories which are popular in some to another region to other regions by promotion.

7 Discussion and Conclusion

This assignment will present a tool that can be used to visualize multiple perspectives of e-commerce. The tool visualizes the map of Brazil where the currently used data set scope is. We plotted many plots related to price, review revenue, product category. Using these plots we were able to analyze the problem statement which was chosen by us. this tool will help to summarize the given data and using these visualizations the company will be able to improve their popularity and revue.

One of the limitations of this dataset is that it is only for a single-commerce website so we couldn't compare it with other e-commerce websites to get a better view about how to improve the overall performance of this company comparing other companies. another limitation was the products in orders are mentioned in categories and products. we couldn't find the product name from the product code. it would have been better if we could get those details too it would help us to add new insights into data.

Using this tool it is understood that there are a lot of improvements the company can do to improve the services as described in the use-cases. such as more advertisements in locations with fewer customers. attract more sellers in different locations to get more customers and also this info will help to improve the delivery time. From this assignment, we got an understanding of visualization and how it is important as well as why we use this visualization.

8 Individual Report

8.1 junaid

For this project, I am involved in testing of this project parts like data processing and plotting. The results are then shared with Faiz to implement in the tool. During the beginning of the project, my main focus was on researching how the required functionalities would have to be implemented correctly and what would be the most efficient way to implement it. The stepping stone towards this goal was to choose the dataset to be used for this project. So that we can have a very efficient tool which will provide a great support to the company to improve. For almost all the part of this assignment we were in some way were partially involved. We always discussed each part we working on with each other so that we get different insights. Mainly during the process of transforming theoretical data into actual code. I often would work with Faiz, which often leads to arguments. We all worked on this project equally and the teamwork was amazing. Lastly, I would like to say that I am happy with the result we have produced and therefore agree with the content of our report and other product.

8.2 Faiz

My role in this project was mainly implementing theory and techniques in to practice. So my focus was on the code side while my teammates provide necessary information for me to use. However, my involvement has been balanced equally over the done work. Subsequently this means I have done some different amounts of work on across most of the implemented components and thus share responsibility with my fellow project members. The components for which I consider myself the most responsible for would be implementing the whole tool and better working of this tool. I also worked together with other members on understanding and learning how to implement these theories in to practice. Furthermore, I contributed in writing and explaining the work I have done in the report. Because a lot of work that we have done has been in a group setting I would consider that we each contributed an equal amount of effort into different aspects of the project. As the task of writing this report also has been done in a group setting I agree with the final content of this report and products we developed.

References

[1] A. Mishra. Data visualization with python. *researchgate*, 2019.

[2] T. Munzner. *Visualization analysis and design*. CRC press, 2014.