Analysis of Brazilian E-Commerce Public Data

[COP5725] Database management system

Project Deliverable 1

Kai Jiang:	*6941 - 6949*
Yu Xia:	*5521 - 4691*
Nikhilesh Reddy:	*8350 – 1593*
Jyotik Parikshya:	*0577-1395*

Index

1. Introduction

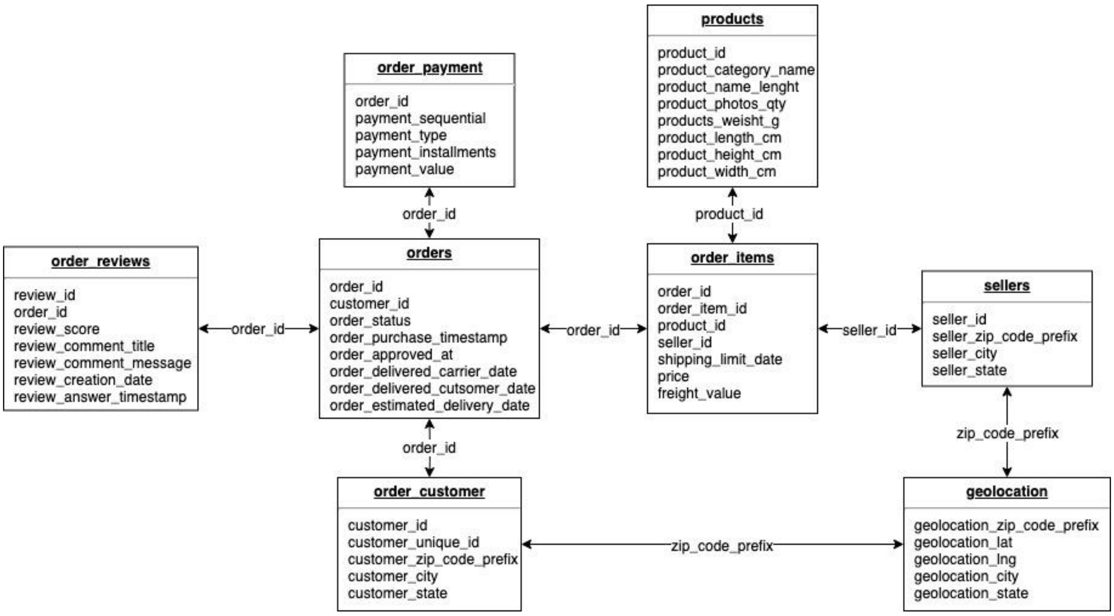
Data Background: The largest department store in Brazilian marketplaces - Olist connects small businesses from all over Brazil to channels without hassle and with a single contract. Those merchants are able to sell their products through the Olist Store and ship them directly to the customers using Olist logistics partners.
After a customer purchases the product from Olist Store a seller gets notified to fulfill that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he/she can give a note for the purchase experience and write down some comments.
Project's data comes from a Brazilian ecommerce public dataset of orders made at Olist Store. The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. There is also a geolocation dataset that relates Brazilian zip codes to latitude/longitude coordinates.
This is all real commercial data, it has been anonymized, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.
Worth Mentioning: 1. An order might have multiple items. 2. Each item might be fulfilled by a distinct seller. 3. All text identifying stores and partners where replaced by the names of Game of Thrones great houses.

System Purpose: There are many kinds of items in the data and purchase records at different times. Therefore, it is difficult to observe the law directly from the data. This system is designed to analyze data and present results in an automated program for users who have different analytical needs for the data.

2. Requirements(software solution)

CISE Oracle(SQL Developer)
Sublime
Language: Java (IDE: Eclipse), HTML5, CSS, JavaScript, SQL
Framework: SpringBoot, BootStrap, Jquery

3. Data Model



customers: customer_id, customer_unique_id, customer_zip_code_prefix, customer_city, customer_state
This dataset contains information about customers and their locations. This is used to identify a unique customer in the order dataset and find the order delivery location.
In the dataset, each order is assigned a unique customer_id. This means that the same customer gets different IDs for different orders. The purpose of customer_unique_id is to identify the customer who repurchase at the store. Otherwise, each order has a different customer associated with it.

geolocation: geolocation_zip_code_prefix, geolocation_lat, geolocation_lng, geolocation_city, geolocation_state
This dataset contains Brazilian postal codes and their latitude/longitude coordinate information. This is used to map and find the distance between the seller and the customer.

order_items: order_id, order_item_id, product_id, seller_id, shipping_limit_date, price, freight_value
This dataset contains data about the items purchased in each order.

order_payments: order_id, payment_sequential, payment_type, payment_installments, payment_value
This dataset includes data about the orders payment options.

order_reviews: review_id, order_id, review_score, review_comment_title, review_comment_message, review_creation_date, review_answer_timestamp
The dataset contains data about reviews made by customers.
When a customer purchases a product from the Olist Store, the seller receives a notification to fulfill the order. When a customer receives a product or the estimated delivery date expires, the customer recieve a satisfaction survey through email where they can write down their purchase experience and comments.

orders: order_id, customer_id, order_status, order_purchase_timestamp, order_approved_at, order_delivered_carrier_date, order_delivered_customer_date, order_estimated_delivery_date
This is the core dataset. All the information can be found for each order.

products: product_id, product_category_name, product_name_lenght, product_description_lenght, product_photos_qty, product_weight_g, product_length_cm, product_height_cm, product_width_cm
This dataset contains data about the products sold by Olist.

sellers: seller_id, seller_zip_code_prefix, seller_city, seller_state
The data set contains data about the sellers who have completed orders in Olist. This is used to find the sellers' location and determine by which seller each product is completed.

category_name_translation (not in structural layout, only for translation): product_category_name, product_category_name_english
This is used to translate product_category_name into English.

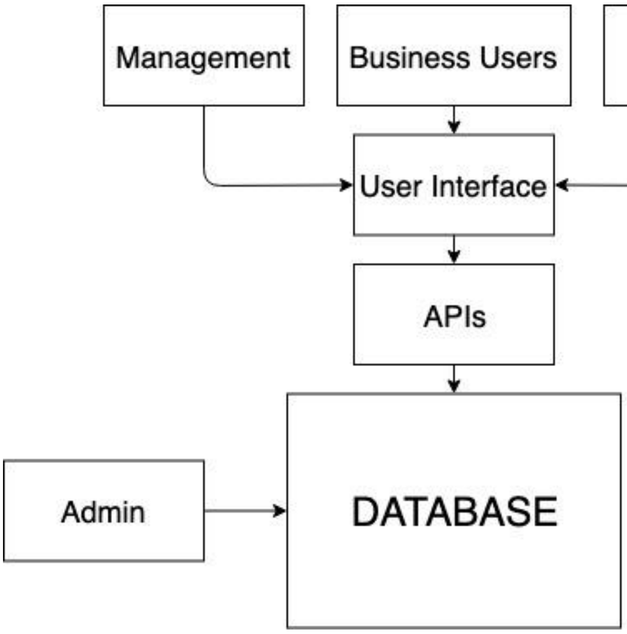
4. Queries

Category	Queries
Trend in Sales	How much amount of sales happened during past few years? What kind of trend in the number of sales can we find across any year or month or week? How much revenue did the company generate during a specified time frame?
Trend of Product and Product Categories	Which categories of product sells best in which time interval? How does the trend in the revenue of a product vary with respect to time?
Season-wise sales Analysis	What is the trend of categories' amount of sales throughout the seasons? Which categories sell best through different seasons?
Geographical Sales Distribution Analysis	What is the shopping order density distribution among different cities in Brazil? What is the total revenue distribution in different location in Brazil?
Delivery Time Analysis	Estimated Delivery time for which region is higher than average? Is there any way to improve it? Does it impact sales and customer satisfaction?
Customer Review Analysis	What can we infer from the customer ratings? What are the factors that are driving the ratings? (For Example: delivery delays, estimated time, product value, freight value etc.)

This is the overview of all the queries.
We will discuss each category in detail in chapter 6.

5. Structural Layout

The above structural layout diagram shows the design of the application we want to develop. For accessing data from Database, different kind of users may have different permissions.



A company will have a single database, using which multiple systems may present their data. In this project we will be showing how a database can be used to develop an application for managers/analyst of the companies. The Database has been used to store data of users, which is also used by their main customer website.

Following commands will be required by the below user groups:
User(Management, Business Users, Analyst): query
Administer: query, create, insert, update, alter

Our basic idea behind building this type of tool is to enable the analyst and managers in a company to make informed decisions/plans for various marketing and operational activities. We will be analyzing trends both graphically and in a tabular manner. There are different sets of data from order details to review scores which is discussed in the next section. Many interactive graphs and maps will be created for these data to analyze trends.

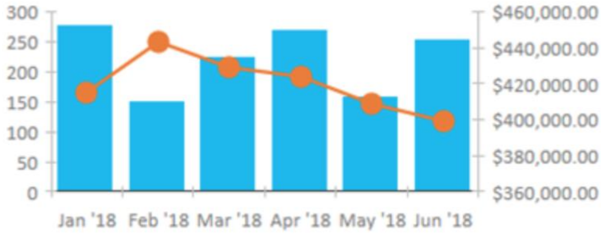
6. Problem Statements Explained

6.1 Section 1 : Trend in Sales

The most basic and major requirement for any E-commerce Organization is to know how many sales are they able to make in a certain period of time so that they can analyze the trend and figure out the strategies for generating more sales or gain edge over competitors.

This information is shown using pictorial representation where we draw graphs by plotting the revenue generated against the time period (either yearly or monthly or daily etc)

Yearly graph shows the number of sales generated over the past few years, monthly graph shows the trend in the number of sales throughout the year or the data from few specified months. The same concept applies for daily and hourly filters. There are some sample graphs here to give you a general idea of the data representation.



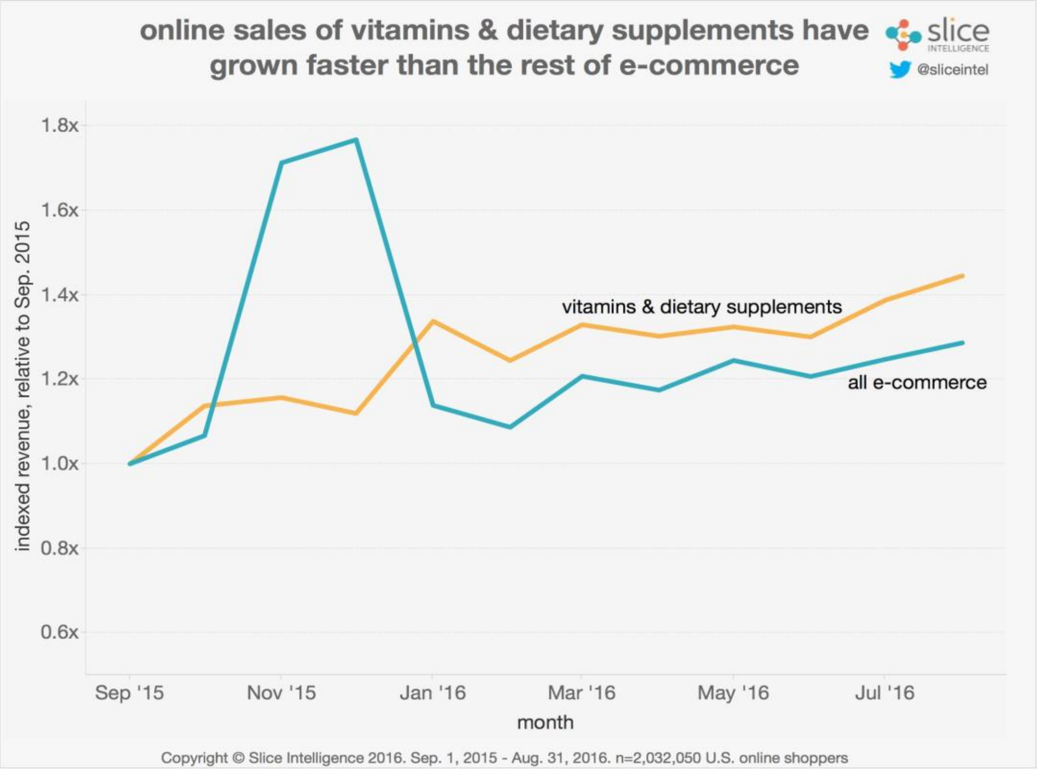
Graphs for Daily sales and hourly sales are presented in the same manner except the values along the X-axis change.

The information shown in these graphs are not bound to any particular department or filtered based on the product categories being sold or mode of payment etc. These graphs only represent the number of sales happened over a period of time and this time variable is adjusted by the organization to view the data using graphs a little more intuitively.

6.2 Section 2: Trend of Product and Product Categories

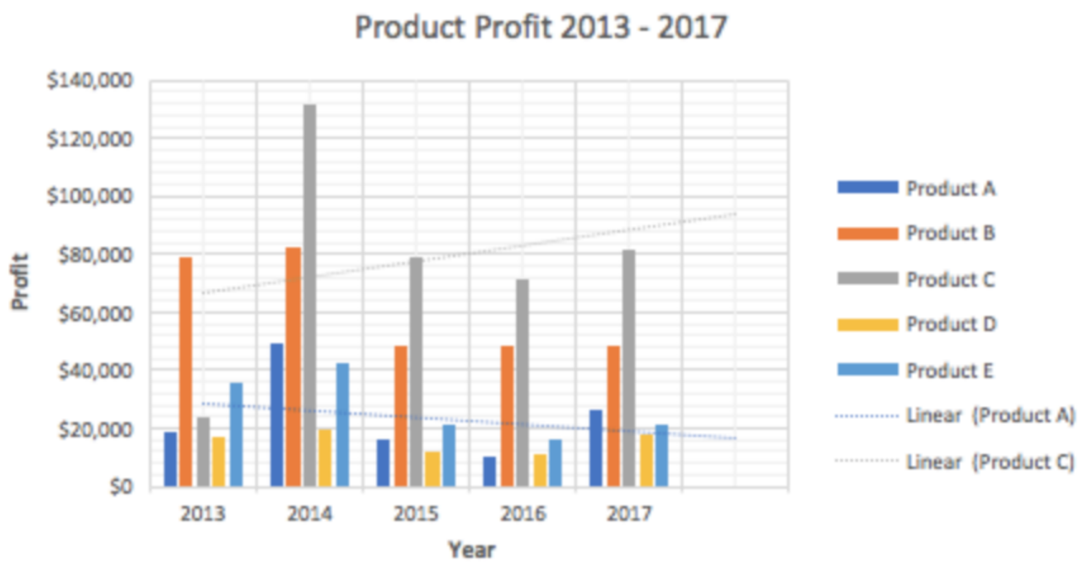
Another step in the same direction is the analysis of number of sales of only certain product categories in a specified amount of time. This gives a lot more understanding of the sales of these products happening across the years and find patterns.

The same concept of representation is used except the data is now restricted to certain product or it's category and not all products included. As an example, consider the e-commerce organization sells vitamins and dietary supplements to its customers and we want to find the trend in the profits of these products over a certain amount of time. The below graph shows the profit margin gained by the organization by selling these products over a specified period of time.



In the earlier case with the total number of sales vs time, we not only show the total number of sales happened over a period, but also the profit or loss incurred by the organization during that time. Applying the same analysis here, we can show the profit or loss trend over a certain products being sold and get a clear picture of the gain and loss time period.
Going forward with this information we can analyze the loss and gain of certain products or product categories over the seasons (summer, winter, rainy etc.) which will be discussed soon.

We also represent the profit or loss margins incurred by the organization. Below are few graphs which help in showing the number of sales and profit or loss, of certain products or product categories varied across a specified amount of time (Yearly, Monthly, Daily, Hourly etc.).



6.3 Section 3: Season-wise sales Analysis

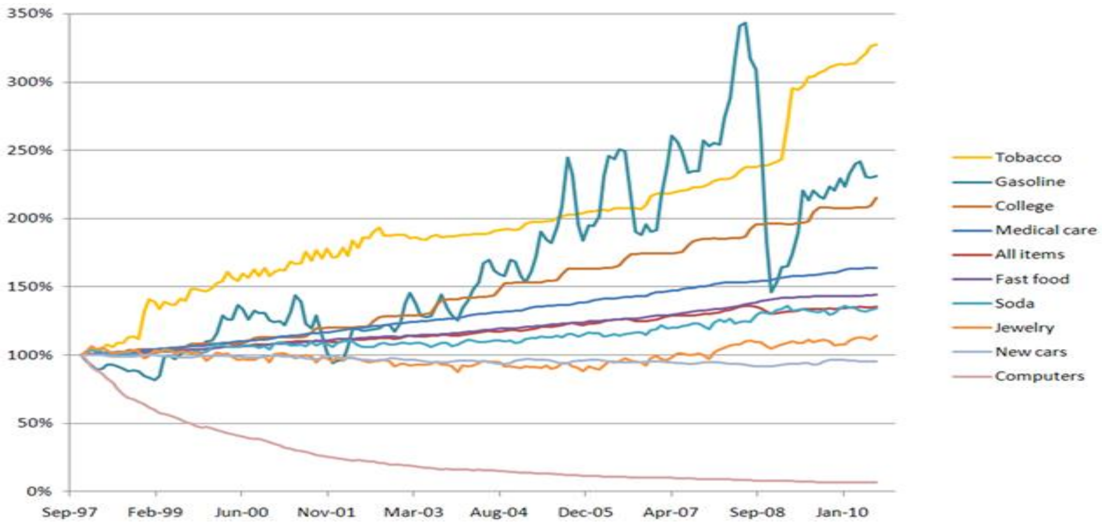
Sellers may want to know in which season (month/week/…) a category sells best and which category sells best in a season (month/week …)

In other words, sellers want to analyze sales, find out seasonal sales peaks and make decisions about shipments and replenishment depending on the analysis.

In conclusion, sellers want to know the law of season goods.

FRUIT	DECEMBER		JANUARY		FEBRUARY		MARCH		APRIL		MAY		JUNE		JULY		AUGUST		SEPTEMBR		OCTOBER		NOVEMBR	
Banana							■	■	■	■	■	■	■	■										
Apricots		■	■																					
Plums		■	■	■	■	■	■																	
Peaches			■	■	■	■	■	■																
Garlic		■	■	■	■	■	■	■	■															
Honey	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Figs					■	■	■	■	■	■														
Apple/Pear					■	■	■	■	■	■	■													
Quince						■	■	■	■															
Walnuts							■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			
Feijoas								■	■	■	■	■												
Persimmons									■	■	■	■	■	■	■									
Casimiroa								■	■	■	■	■	■	■	■	■	■							
Cherimoya																			■	■	■	■	■	
Tamarillo													■	■	■	■	■							
Pomegranate									■	■	■	■												
Grapefruit																■	■	■	■	■	■	■		
Tangelos	■	■	■	■	■	■	■	■	■							■	■	■	■	■	■	■	■	■
Mandarins	■	■	■	■	■	■	■	■					■	■	■	■	■	■	■	■	■	■	■	■
Oranges													■	■	■	■	■	■						
Lemon/Lime	■	■	■									■	■	■	■	■	■	■	■	■	■	■	■	■
Lemonade															■	■	■	■	■	■				
Avocado	■	■	■	■	■	■	■	■								■	■	■	■	■	■	■	■	■
Pumpkins						■	■	■	■	■	■	■	■	■										

Seasonal goods are commodities that have significant seasonal characteristics in production, acquisition, and sales. Such as agricultural and sideline products, summer cool goods, winter goods and so on. Such commodities are seasonally produced, seasonally acquired, and sold year-round; there are perennial production, perennial acquisitions, and seasonal sales; seasonal production, seasonal acquisition, and seasonal sales. In order to ensure the normal supply of seasonal products in the market, enterprises generally purchase in advance according to the characteristics of production and sales, reserve in advance, and prepare for the supply of goods before listing.



There are several types of seasonal goods:
1 single peak type.
This type can be divided into three types. The first is that when the product is at the peak of sales, the price rises, and when the sales are low, the price drops, such as clothing. The second is that when the product is at the peak of sales, the price drops and when the sales are low, the price rises, such as vegetables and fruits. The third is that the price of the product does not change during the peak sales period and the low sales period, such as ice cream and cold drinks.
2 double peak type:
This kind of seasonal goods has reached sales peaks twice in one sales cycle. For example, sales peaks in air conditioners and refrigerators are in winter and summer, and spring and autumn are low in sales.

6.4 Section 4: Geographical Sales Distribution Analysis

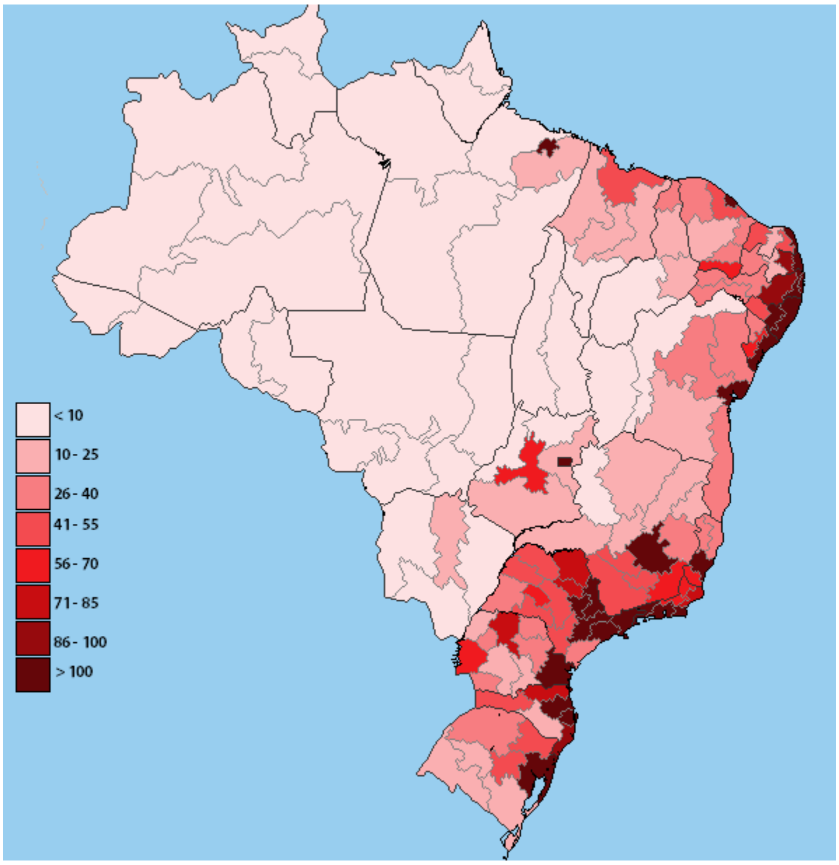
With geolocation information like longitude and latitude and zip code in order information provided in the dataset, we can reflect the quantity information of the corresponding online shopping orders on the map in the form of density. In this way, we can plan logistics distribution stations according to the density information of different regions, thus realizing a more optimized logistics distribution function.

Description:

For online shopping, a significant process is the delivery of goods after shopping. Therefore, it is very important to set up logistics stations in proper places. So how do we plan the distribution of logistics sites?

Using geolocation information such as order zip code, latitude, and longitude in the dataset, we can obtain shopping order density information of each city and each region. This shopping order density information indicates the relative order quantity generated in this area. If the shopping order density is high, it means that there are many people shopping online in this area and there will be relatively more packages to be delivered. In this case, we can choose to set up a logistics delivery station in this area to optimize the delivery of orders and improve the shopping experience of customers.

We can visually display this density information through maps. The way we will show this information is similar to the following figure:



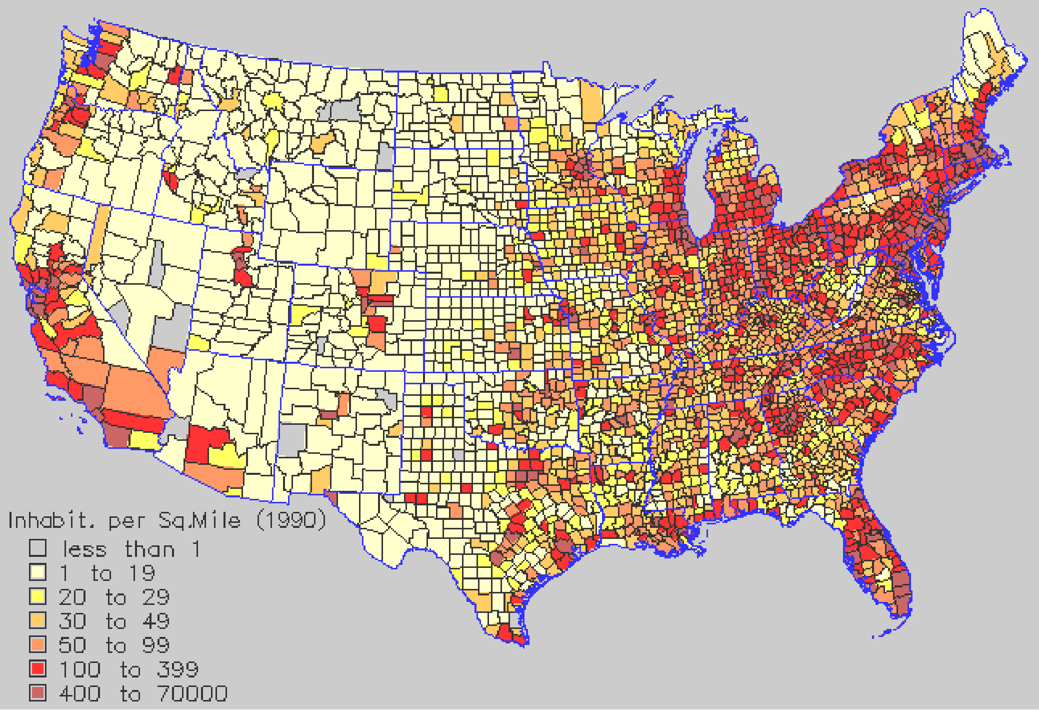
By using the zip code, longitude, and latitude data of orders to determine regions, we count and compare the number of shopping orders in different regions.

If the shopping order density of an area is higher, we can show that with deeper color. In this case, the area that has higher shopping order density can be shown very clearly.

Further, we can do the same thing for a city or even streets. Then according to this shopping order density map, we can set up new logistics station in the areas where have higher quantity of delivery package.

Besides the quantity of shopping orders, the total revenue is also very important for marking analysts. With this information, sellers can find where does the most revenue comes from, in which areas the customers have stronger purchasing power.

Working with zip codes, we can get the sum of products value grouped by zip code. The format we show this information is the same as shopping order density above, using map like this.



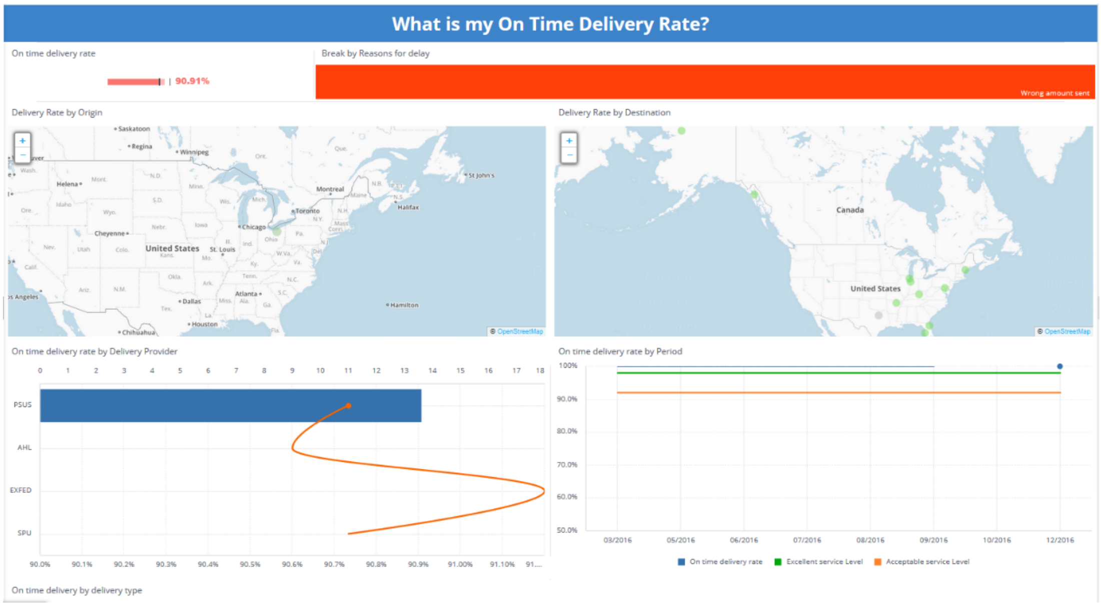
With this information, merchants can choose these areas with stronger purchasing power to carry out promotional shopping activities, thus obtaining higher sales volume.

6.5 Section 5 : Delivery Time Analysis

Analyzing the time taken for processing and delivery based on the seller and location. We will also be able to work through delivery performance and find ways to optimize delivery times. We can use the attributes of data “orders” like delivery date and expected data and use order id to link it with other datasets like order_items, sellers, customer and product.

For every E-Commerce organization it is very important to analyze and rectify all the processing and delivery overheads in terms of both time and cost. Its job of an analyst to give appropriate analysis with help of organized data and graphs, an overview of where things go wrong.
We will create a portal which helps the operations/business team of the company to analyze and pinpoint the exact areas where they can improve, How?

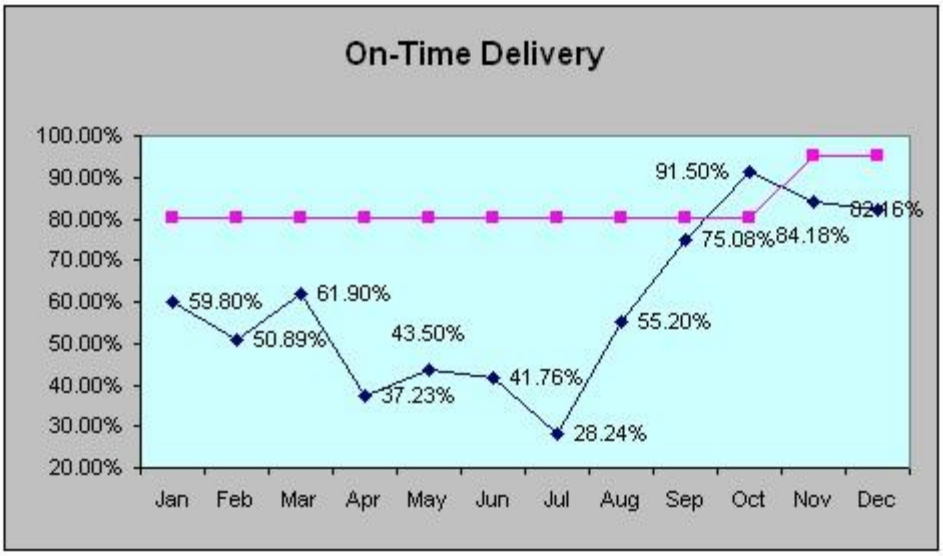
Our Portal will look like the below image with one or two additional graphs.



The data available with the company can be organized to give insights on the above. For Example, the portal will contain a list or a table with processing time mapped with a seller by worst performance (when expected-actual is negative). By this example we can figure out which seller needs to improve its efficiency in terms of processing time. This list can be visualized better on a Map based on seller location. This might give a trend based on geographical conditions.

The Map analysis in the portal will look very similar to the above picture.
Coming to the analysis of the most important part i.e. the delivery time.
On time delivery Is an integral part of customer satisfaction hence a major area of interest. Delivery time analysis not only depends on the seller or location but many other factors like logistics, delivery partner, connectivity from seller to customer etc.
Involving the major factors, we will create a map based distribution of customers and show two things. First is the difference between delivery time and expected delivery time and second is the actual delivery time alone which will be shown by different color shades to make it look for readable.

The Purpose of showing actual delivery time is to enable the companies think tank to plan location of warehouses accordingly. No customer will like to wait for 8 days for delivery of any product, they don't care where the seller is.   In order to reduce the delivery time for higher customer density areas, planning of warehouse can be an interesting field to work on for the company.



We can also have a Month wise analysis on time delivery percentages as well as season wise analysis as shown in the fig above.

6.6 Section 6 : Customer Review Analysis

Analyzing the feedback rating given by the customers:

We can use the data in order reviews by linking it with various other tables using order id, product id and seller id.

According to some marketing experts, below are the major criteria for prediction of review score from customer:
1. Working Days Estimated Delivery Time
Getting the days between order approval and estimated delivery date. A customer might be unsatisfied if he is told that the estimated time is big.
2. Working Days Actual Delivery Time
Gets the days between order approval and delivered customer date. A customer might be more satisfied if he gets the product faster.
3. Working Days Delivery Time Delta
The difference between the actual and estimated date. If the difference is negative, then it means that the delivery was made early and if positive then the delivery was made late. A customer might be more satisfied if the order arrives sooner than expected, or unhappy if he receives the delivery after the deadline
4. Is Late
Binary variable indicates if the order was delivered after the estimated date.
5. Average Product Value
Cheaper products might have lower quality which leaves customers unhappy.
6. Total Order Value
If a customer spends more, he might expect a better order fulfilment.

7. Order Freight Ratio

If a customer pays more for freight, he might expect a better service.

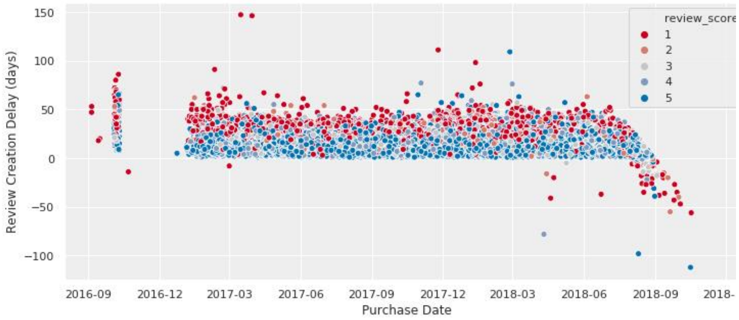
8. Purchase Day of Week

Does it affect how happy the customers are?

We will answer these questions by taking into consideration each factor and do graphical analysis for each of them.

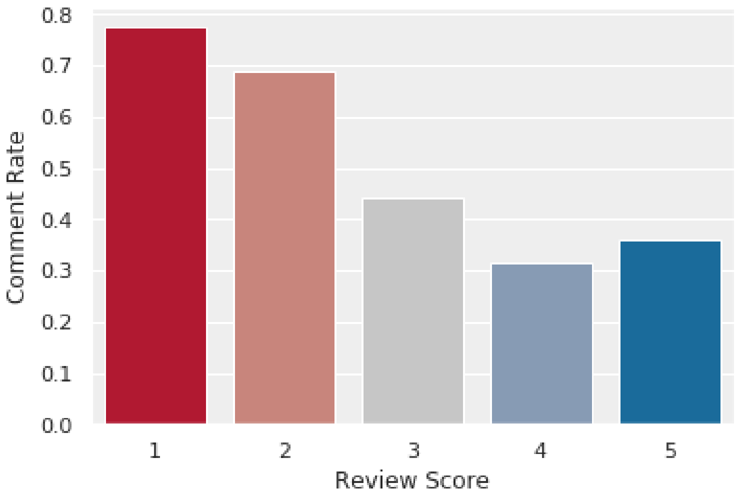
Some more insights into this analysis yields more interesting results:

We can map month wise review scores and analyze by mapping it against the review creation delay. We can get some interesting insights from it.

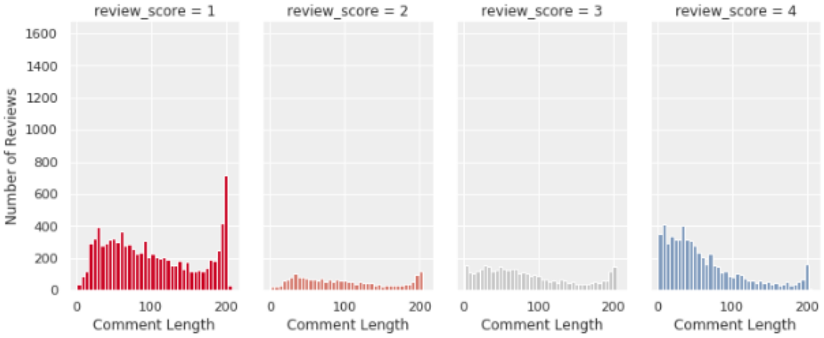


Distribution of ratings based on order status is another way to filter out actual product reviews from the unsatisfied customers due to other reasons.

We can further make some analysis using review comments and see the customer behavior. This particular section will be helpful to the CRM division of the company.



Comment rate as the number of non-NULL comments divided by the number of reviews. Distribution of Comment rate with respect to review score gives an insight on how customers, who are dissatisfied, comment more than the customers who have given good ratings.



We can find out comment length and map it against review score to analyze that angry customer write more. Company can develop categorization based on comment keywords and address the issues in a more organized way.

Further, we can pull out keys words from the comments and categorize them accordingly, this analysis is beyond the scope of this project but is worth mentioning.