The Olist Commercial Dataset -

An analysis of customer satisfaction

by Alin Cristian Preda

Class teacher

Abid Hussain

Copenhagen Business School

Business Data Processing and Business Intelligence

December, 2019

ABSTRACT

WHAT WAS THE TOPIC?

Is prompt delivery of purchased products a staple of customer satisfaction in e-Commerce?

WHAT WAS THE DATASET?

Olist provided a dataset containing an almost two years period of information about its sellers, buyers, orders, products, customer reviews and location data.

WHAT ARE THE MAIN TOOLS AND METHODS?

The main tools are Microsoft SQL and Alteryx, for storing, cleaning and preparing the data, and also Power BI and Tableau, which were used to develop data visualizations and dashboards. The method used for analysis was the observation of the data visualizations.

WHAT WAS THE PROBLEM FORMULATION?

Is Olist doing a good enough job at shipping its seller´s products to the customers?

WHAT WAS THE RESEARCH QUESTION?

Why do customers give low reviews? Are there differences between regions, states, cities?

WHAT WERE THE CONCEPTS?

e-commerce, customer satisfaction, delivery times

WHAT ARE THE MOST IMPORTANT RESULTS?

Customers are affected by long delivery times and delays. More incoming orders correlate with slower delivery times. Statistically speaking, the customers in each region act almost identically.

WHAT ARE THE CONCLUSIONS AND RECOMMENDATIONS?

Olist could make more shipping free of charge; it could contract more couriers to take the load off, especially on busy periods; it could offer more shipping solutions.

KEYWORDS

Brazil; e-Commerce; Transportation; Shipping; Logistics; Customers; Customer satisfaction;

olist
empowering commerce

Table of contents

INTRODUCTION

This research takes a closer look at an online marketplace for local retailers in Brazil. The company Olist has provided the public with a dataset containing commercial information on customers, sellers, orders, products from which we can understand how they interacted with each other. I will attempt to sketch the overall face of Olist and its customers, products and sellers.

This study is most significant to the management of Olist. It mostly suits the interests of the marketing and customer relations departments. The conclusions drawn from this analysis could enable the management to better understand their business. So much so that it can strengthen Olist's bonds with its partners and customers. The study can also be of interest to other companies (possible competitors) in Brazil because it tells an interesting story about the demographical and geographical influences the country can have over a business operating in on-line marketplaces.

I decided to use this dataset because it's from a real company, with a novelty business model, who has been operating for a short period. I was attracted by the complexity of having almost a dozen datasets connected to each other. Finding the right connections between them and then ultimately uniting them in an all-encompassing database proved to be a satisfying learning experience. It also satisfied the data management requirements of the curriculum.

Olist is Brazil's largest department store within the marketplaces. It offers an online platform where small and large businesses from all over the country can sell their products, with a single contract. The sellers can market their products directly through Olist Store and use the company's logistic partners to ship them to the customers. For business owners, looking to sell, Olist offers the opportunity to benefit from great reputation, qualified traffic, great visibility and higher ranking in search engines. Using their services, small or new businesses can profit off of the reputation and visibility of a large community of well-established sellers. Regarding shipping, Olist offers an exclusive partnership with the post offices, boasting more attractive freight values. The company promises shopkeepers a solution to enable them to sell more and better and attract new customers.

There are, as always, unsatisfied customers or areas where the company could do a better job. The main research questions aim to uncover the strengths, weaknesses and opportunities available to Olist. What is behind the bad reviews that some customers gave? And is there any difference between the commercial activity in one region and the others?

METHODOLOGY

DATA COLLECTION: METHODS AND TOOLS

This real commercial data was found on Kaggle, where it was freely publicized by Olist. The datasets are in .csv format and ready for download. The data has been updated or modified multiple times during its history. The version of the dataset is Version 7, updated on the 1st of December 2018. [1]

DATASET DESCRIPTION

For the purpose of this project, I chose to focus on SQL and data visualizations. I simply downloaded the data in delimited flat text type .csv format. JSON format is superior, much more flexible and universally readable. [2] So, a conversion can be made by using, for instance, Python. Multiple methods of achieving this, using modules like csv, json or Pandas, are described on Stackoverflow. [3] Either way, Microsoft's SQL Server Management Studio and Tableau can work with both formats. The data is structured as tabular data, representing information about objects that share common proprieties. My data collection was organized into multiple tables of records and fields. It can also be viewed as an Excel workbook with multiple spreadsheets.

It has actual anonymized commercial data ranging from 12/30/2016 2016 to 9/24/2018. It is organized in multiple specialized datasets that can be merged in different ways, according to the research topic: orders; order items; order payments; order reviews; customers; sellers; products; geolocation and product category translation.

RELATIONAL DATABASE MANAGEMENT SYSTEM

It is fit for relational database management systems, as I am interested in the multitude of entities (customers, sellers) and objects (products, orders, reviews), their special characteristics and the way they interact with each other. Relational operators such as the primary and foreign keys help us manipulate the tables according to our needs and desires. And I thought it best to combine all of these tables into a massive database which gives me access to mine all of the information I need for developing Business Intelligence. I want to identify instances of objects that meet certain criteria. I want to be able to present the dataset in multiple different ways, with the query results being able to be exported and used elsewhere.

SQL QUERIES

A database was created via Microsoft SQL Server Management Studio 18. The files were imported as flat file sources into the tables folder, with the necessary adjustments being made in terms of data type formats and primary/foreign keys.

The Annex 1 – SQL, contains code snippets from the queries that I've used in order to merge the tables, based on their relationships. It also contains queries that I've used to answer specific business questions such as:

- What are the top grossing regions, ordered by state?
- What is the total sum of sales in the South Region?
- What is the total number of customers in the South Region?
- What are the top grossing regions?
- Where do we get the biggest delays (total) in shipment?
- Who are the top 3 sellers?
- And so on…

We can of course also search, sort and filter data. We can perform mathematical operations, for finding statistical indicators such as means, mods, medians, frequency distributions.

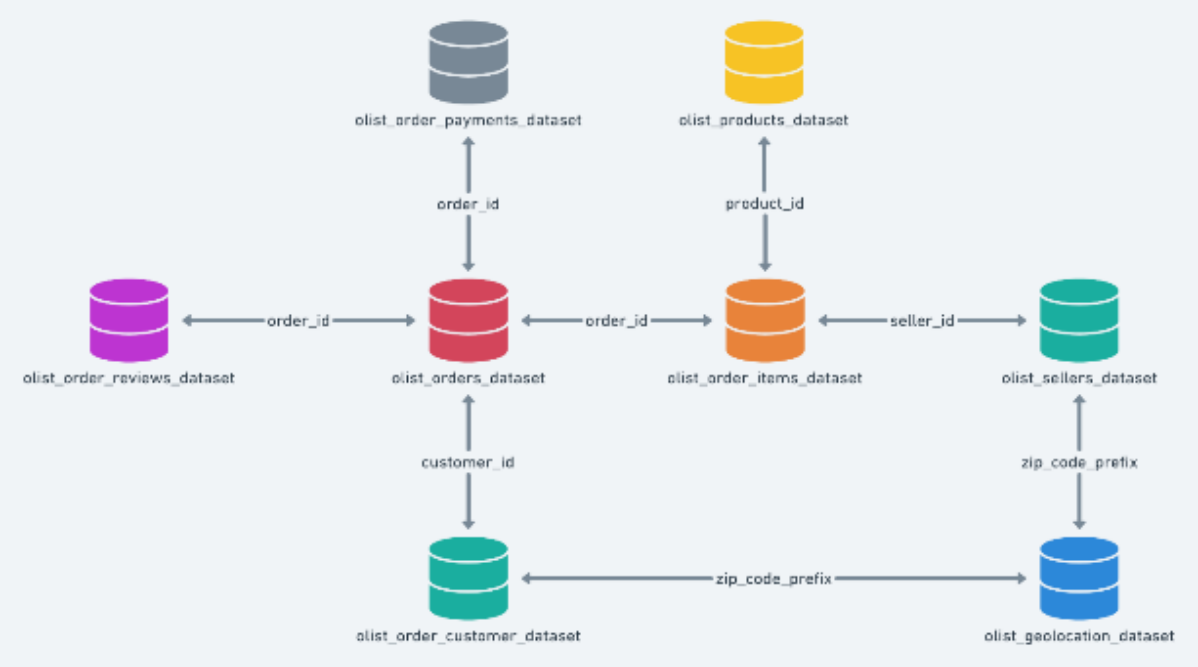
DATA CONTENT AND FORMAT

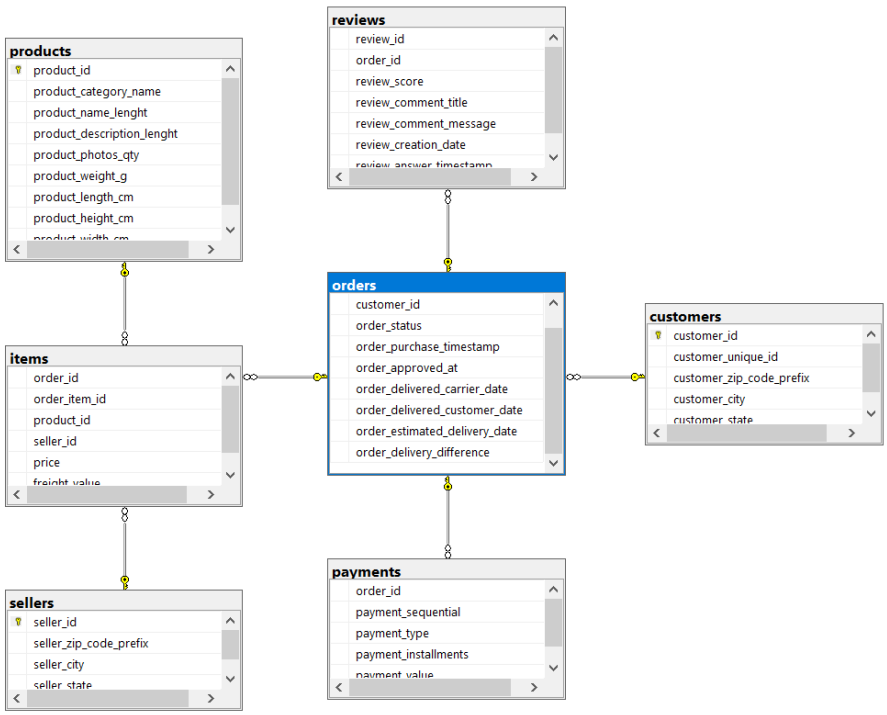
- The customer dataset contains information about the buyers and their location. Each order is assigned to a unique customer_id. This means that the same customer will get different ids for different orders. The purpose of having a customer_unique_id on the dataset is to allow you to identify customers that made repurchases at the store.
- Similarly, the sellers table tells us about their identity and locations.
- The geolocation dataset contains Brazilian zip codes and coordinates
- The orders table stores information about what was purchased.
- The products table contains information about product types and their traits.
- The items dataset contains data about what products were sold in which order.
- The reviews table has information about review score, text, user, order.

The dataset contains many variables of several types: datetime (delivery times, purchase times, order times), string (the IDs, the comments, the product categories), geolocation (latitude, longitude, city, state) and numerical data (price, payment value, freight value).

RELATIONSHIPS

The first schema was provided by Olist. It illustrates all the files available and the relationships are based in common keys or ID´s through which they can be merged. I have opted for creating an ultimate dataset which connects all of these into one pool. Each order might have multiple items. Each item might have a distinct seller. The second schema comes from my database.





DATASET LIMITATIONS

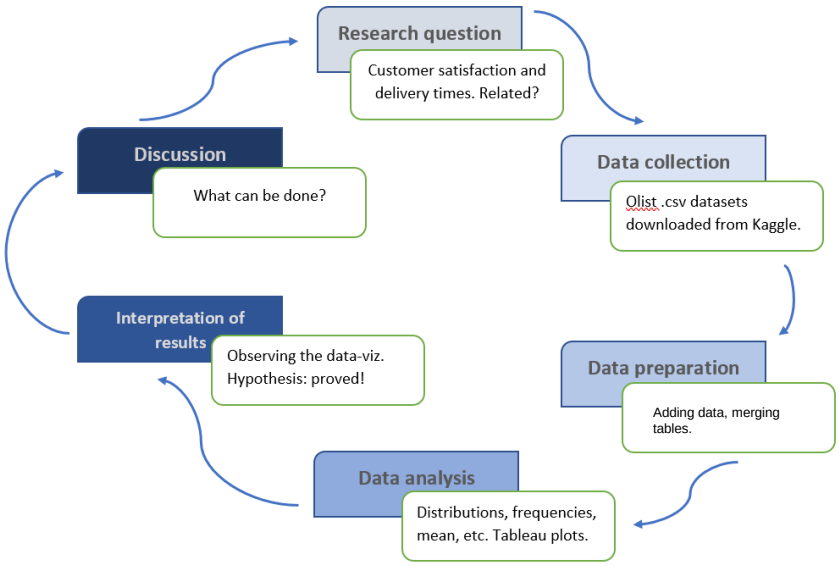
Olist's business is relatively new to Brazil. This is why the dataset is very limited from a temporal point of view. Attempting to identify trends in time series or predicting future KPI is not recommended with having such limited information. But there is still a lot of relevant data to be studied. We can interpret this information such that we can understand the company´s current situation and assess the progress that´s been going on so far. For the purpose of this project, not all features and all data points were necessary so I have limited the scope a bit.

DATA CLEANING AND PREPARATION

The data is already very clean and usable. It just needs a bit of tweaking around. That is why I translated the product categories and added the English names as a new column to the products table. I calculated the difference in days between the estimated delivery time and the actual product delivery time, in order to have a quantifiable indicator of delays. I also added a region column and a product super-category column to the last iteration of the aggregate data table, in order to help in the analysis.

DATA ANALYSIS PROCESS

DATA ANALYSIS PROCESS DIAGRAM

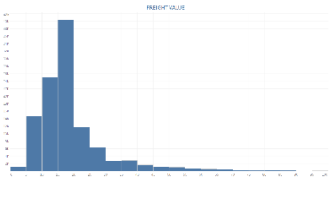


After preparing the data and forming a better understanding of it, visualizations were created using PowerBI and Tableau. Some charts directly attempt to answer the research questions. Other were made to explore which lead to new questions and new answers, some of them uninteresting, some irrelevant, and some that proved surprising.

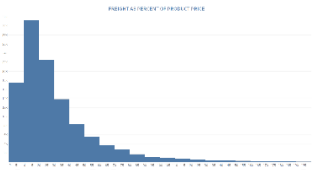
Ultimately, an interactive Tableau dashboard was created which could spark a discussion about the problem and how the company can choose to act based on this research. By making use of filters and highlighters, the dashboard can enable management to pin-point customers, geographical areas or moments in time based on different combinations of variables that describe the 'target'. When analyzing complex datasets of considerable size, it is often recommended to make use of data visualizations in order to better understand the subject matter. Patterns can emerge and subsequent "eureka" moments can give rise to better organizational decision-making.

RESULTS

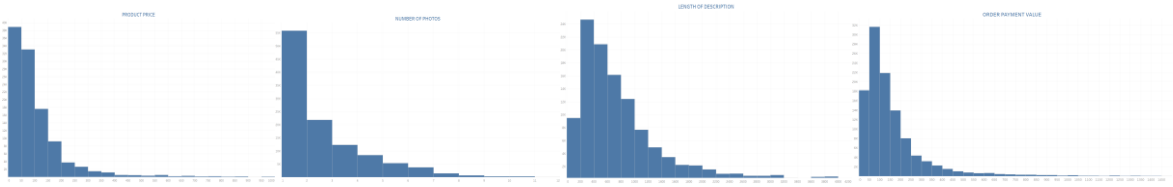
EXPLORATORY ANALYSIS



The freight value follows a right-skewed, positive distribution. That is, most values present themselves towards the lower end. According to the bins, most freights hold values between 10-20 R$. As a percentage of price, most values situate are between 10%-20%.



Price also follows a positive distribution. Most products cost less than 100 R$. Very few of them cost more than 400 R$. Regarding the number of photos in the product's description, we can see a right-skewed histogram, with most of them having just one photo. Most descriptions are short, between 200 and 400 words but some of them can reach over four thousand words.



All the histograms continue with bins to the extreme right, depicting outliers, which I chose to exclude from the visualization and discard as irrelevant. The distributions seem to be normal, considering the nature of the products sold on Olist's platform.

OLIST IN NUMBERS

The final merged dataset contains information about 96k customers from 15k zip codes, 4k cities, 27 states, 5 regions. There are 33k products divided into 73 categories and 11 product super-categories described in over 70k photos. There are records of almost 100k orders with 100k reviews. More exactly, there are 96,096 unique customers. There are 99,441 customer ID´s in the 98,666 orders, which means that 3,345 of the buyers engaged in Olist´s services more than once. That is just over 3%. The customers come from 4,119 different cities. There are 14,994 different ZIP codes. There are 3,095 sellers associated with Olist. They are spread out in 23 of the 27 states and are based in 611 cities, over 2,246 different ZIP codes. Olist´s sellers boast 32,951 different products, presented in 70,794 photographs, spread over 73 product categories. From these 73 product categories, I managed to synthetize a much shorter list of just 11 product super-categories. When people request an order on Olist, they order between 1 and 21 items at a time. The product´s average price is 120.7 Brazilian reals, R$. The median of prices is 74.99 R$. The lowest valued product is just 0.85 while the most expensive one costs 6,735 R$. There are 5,968 different prices. In total, Olist managed to sell products worth a total sum of 13,951,644. In total, the customers paid 2,251,910 in freight value. The average freight is 19.99 R$, with the median being 16.26 R$. There are 6,999 different values for freight. The least a freight can cost is 0 while the most expensive one is 409.7 R$. There are five types of payment available on Olist. A maximum of 24 payment instalments is accepted. Sequential payment is also possible, which means that a customer can choose to pay for a purchase using multiple payment methods. In total, Olist managed to sell products worth a total sum of 16,008,872 R$. The biggest order managed to attract 13,664 R$. The orders averaged 154.1 R$, with a median value of 100.

REGIONAL ANALYSIS – MERGING SOURCES OF INFORMATION

```
UPDATE olistDatabaseFinal
SET    customer_region = 'South'
WHERE customer_state like 'PR'
OR    customer_state like 'SC'
OR    customer_state like 'RS';
```

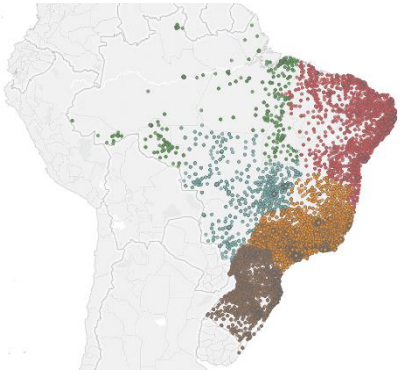
The data can be used to have both a top-down and a bottom-up approach. We can go as broad as regions and states or as narrow as cities and even postcodes, latitudes and longitudes. In order to help me simplify the data analysis, for the purpose of business intelligence, I collected and added, in SQL, information about the Brazilian administrative regions and created super-categories for the products. Below, I've created two visualizations depicting the map of Brazil. Both of them color-code the regions so that they are distinctively represented. The first one is

just a basic administrative map, whereas the second one is plotted using dots. The dots represent customers, so that we can see an actual geographical distribution of Olist's activity.

According to Wikipedia: [4]

NORTH Transportation is mainly done via rivers, of which the region has many. The very few highways are mostly in the east. The cities are spread very far apart. Sometimes airplanes are needed to access remote areas of the region. This is why we see customers are few and the spread is large.

NORTHEAST Transportation is done mostly via the many highways or by sea, along the coast. And indeed, we can see a dense concentration of dots on the cost, fading towards mainland.

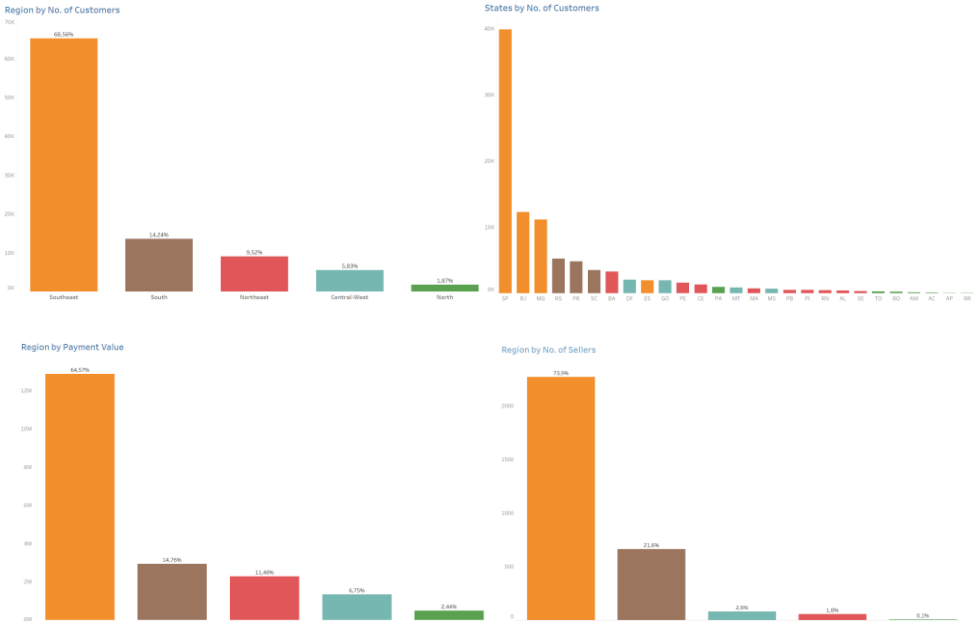


CENTRAL WEST Transport is done by some highways mainly in the east and centre, and the rest of the area is taken care of by nautical transport and sometimes airplanes. Again, one of the areas where the customers are scattered.

SOUTHEAST It is covered by an extensive network of highways and, to some extent, railways which make transportation fairly easy and fast. This region is both the most populous and the richest, being responsible for almost half the country's GDP. It is obviously packed with customers all over its territory.

SOUTH This region also has good highways and railways with the possibility of transportation via rivers, when necessary.

Now if we plot the data more traditionally, we can confirm that the coast is where the money is being made and spent. Almost 7 out of 10 of Olist's customers come from the Southeast. Most of them, from states such as Sao Paolo, Rio de Janeiro and MG. This result was to be expected, given the population of the region and its economic power. The same applies to the rest of the regions: the greater the population and the GDP/capita, the greater their contribution in Olist's customer base.



The color palete of the regions, using both cold and warm hues, helps the viewer easily distinguish between them. Bars help the viewer instantly judge the size of the business.

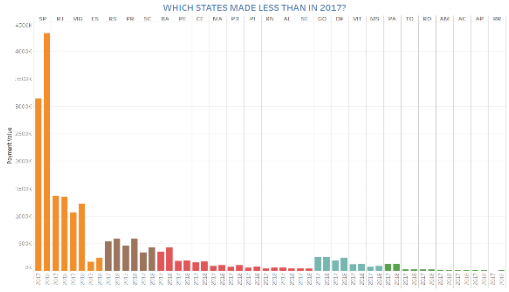
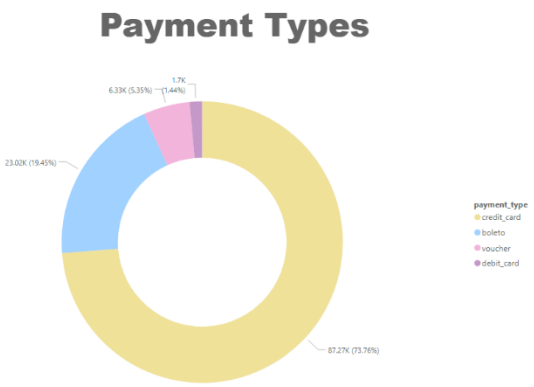


Tableau allows us to also make some useful calculations and integrate them into plots. Most states already performed better in 2018 than in 2017, altough data is still missing from September onwards. Some of them, tough, leave a lot of room for improvement. This chart will become more useful at the end of the year, when all of the accounting is done and we can judge the business' overall performance. That way, management can decide where to focus their attention.

PAYMENT METHODS ANALYSIS

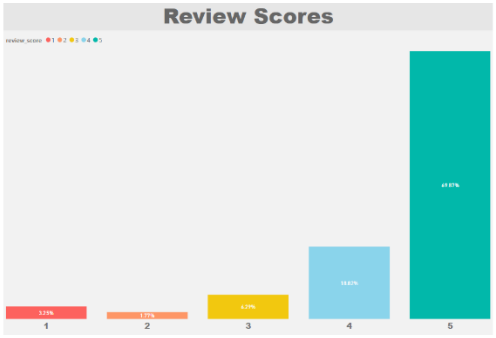
1 Brazilian Real (R$/BRL) = 1.68 Danish Krone (DKK) = 0.22 Euro (EUR/€)



Pie charts are tipically frowned upon among data scientists. Instead, a much more appealing and revealing donut chart was used to study the frequency of payment types (altough a bar chart could have done the job at least as good, if not better). Most choose to pay by credit card, then "boleto", an increasingly popular local payment method. Regardless of region, the proportions stay pretty

much the same. Interestingly, more people make use of vouchers than a debit card. Choosing to supply the customers with the opportunity to pay with boletos was a sound strategy and should be further enforced. Being a Brazilian-only payment method, it is essential on the local markets. Additionally, 81% of Brazilians not having an international credit card, I think Olist could reach more of them if it decided to expand and include international marketplaces to its platform. [4] Boleto Bancario payers are 1 in 5 customers already. I would suggest that every business should integrate the newest, most popular payment methods, for convenience´s sake. There were no significant results between the region groups, all of them recording similar % for each type of payment method.

REVIEW RATINGS ANALYSIS

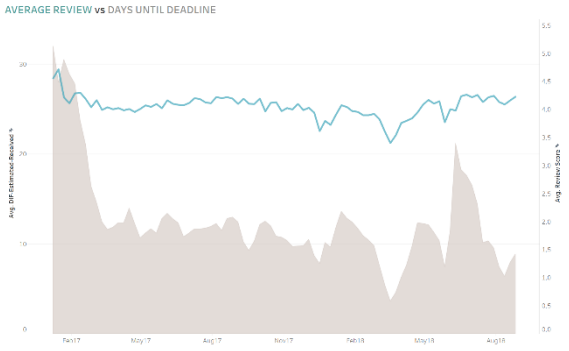


Most customers seem to be satisfied. Indeed, 6 out of 10 orders were rewarded with a review of five stars, 2 out of 10 with four stars. That means 80% of the orders were perceived positively by the buyers. If we take into consideration the fact that the review rating scale has five steps, then a review of three stars can be considered neutral and a two or a one is a negative review, hence the colour scheme choice.

There are not many neutral reviews, with very few of two stars and 12% of orders received the lowest possible score. Overall, 15% of the orders were perceived as a negative experience. What can we learn from these low reviews? Is it something that has to do with the products or the transport? Using BI, we can look at them in the broader sense but also pinpoint those few nasty experiences and learn from them, if we like. There are no significant variations of review preferences between the regions.

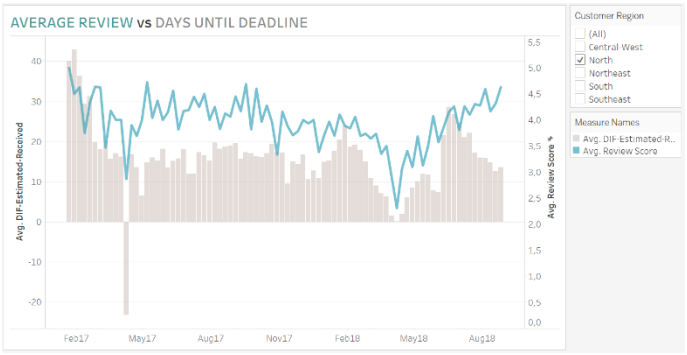
ANALYSIS OF DELIVERY

The dataset made available various timestamps: from when the order was recorded, to when it was approved, when the product was handed in to the courier, to the actual delivery and including an estimate of the maximum date for the arrival time.



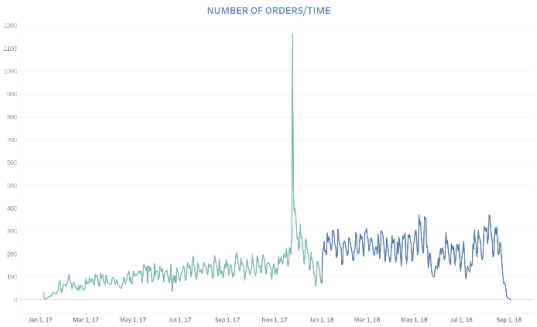
This chart uses an unsynchronized dual axis to analyze the relationship in time between the average review scores and the average time remaining until delivery deadline (Estimated delivery time minus Actual delivery time, measured in days). A larger time difference means that the order got delivered faster than expected. Indeed, it seems that there is a somewhat common trend. Although there is strong correlation, the performance of the couriers cannot fully explain the reasons for all the ratings. A future study using sentiment analysis on the comments written by people who gave low ratings might reveal some patterns and shed light on what might be the main causes of concern.

ANALYSIS OF DELIVERY – GOING REGIONAL

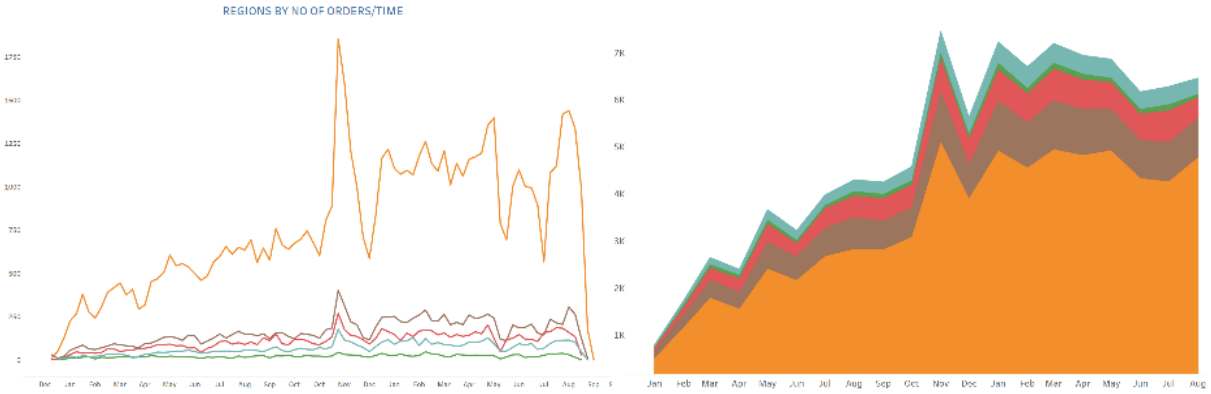


This becomes more evident in certain periods or when we isolate certain areas. For example, here, filtering in the North, we see a pattern emerging. Coincidentally (or not), the review scores suddenly got much lower while at the same time, the delays skyrocketed. The dates are April 17, 2017 and April 18, 2018. Here, the reviews got as low as 2.8 and 2.3 respectively, for average days until deadline of -23 and 0.11 respectively. Although we see an improvement in delivery times, the feedback was worse than last year. Perhaps people remember. Especially since this strange coincidence has a low probability of appearance of 1 in 365 or 0.27%. After all, we are talking about the exact same problem, two years in a row, on basically the same day. It asks for a more thorough analysis of what happens in the spring quarter. Perhaps this unfortunate phenomenon should be expected in 2019. Similar trends appear for all the other regions, reinforcing the fact that it's not just a coincidence or a local nuisance.

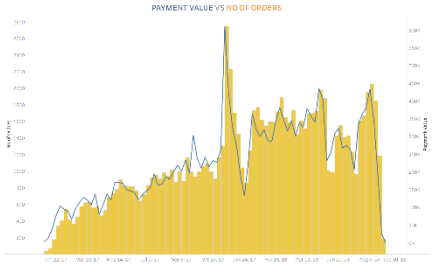
TIME ANALYSIS OF ORDERS



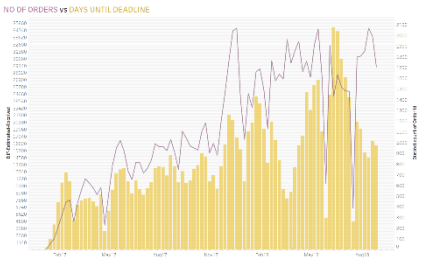
The peak reached by orders/week was reached on Friday, Nov 24, 2017, which was Black Friday in Brazil. 2016 was excluded from the purposes of this study dew to the limited amount of time series data, we can see no seasonality developing but we can expect the same sudden surge in the volume of sales in November 2018. Colors light and dark blue were highlight the passing of the years.



The line chart studies the evolution of orders/week. At a first glance, all regions have the same consumer patterns overall. Lines are best for showing time series. Seasonality would affect each region almost the same, although to very different degrees. The stacked area chart shows the orders per region from month to month. The areas help the viewer understand the volume of orders, putting the relative differences between regions into a better perspective.

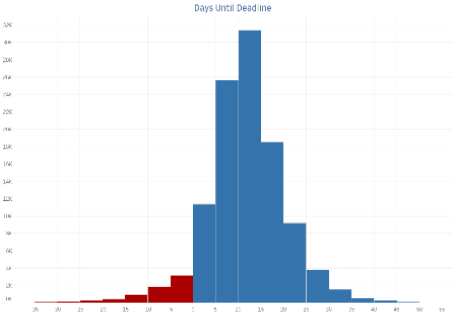


As expected, according to the dual axis plot, the sales amounts are closely tied to the number of orders. But there is also something



contradictory in the data. From this second dual axis chart we can see that, quite expected, when there are many active orders there is also great delay in delivery times. This is generally true except for, most notably, again, spring 2018, when there were both many orders and a period of fast shipment. This means that the couriers sometimes become overwhelmed, for whatever reasons.

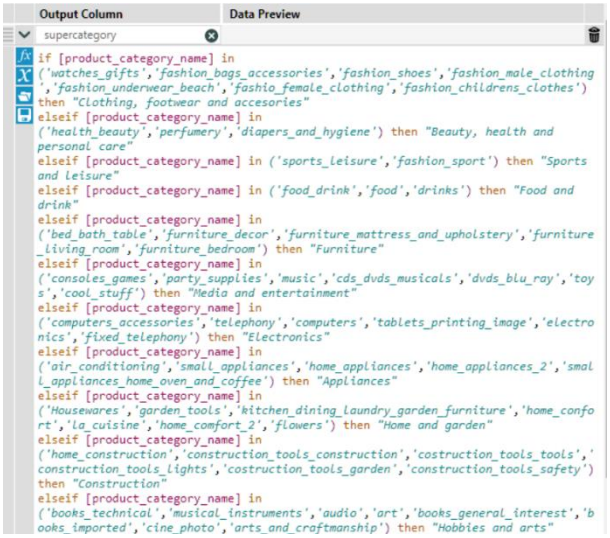
This map explores the cities with the greatest delays. Blue arrive very early, gray is average and the warm colors represent high delay times. We can pick out the problematic cities for further inspection. The method can be further expanded to the level of each individual customer, based on their location. And there are some kernels on Kaggle that actually found reasonable and explainable relationships between postal codes and delivery times.



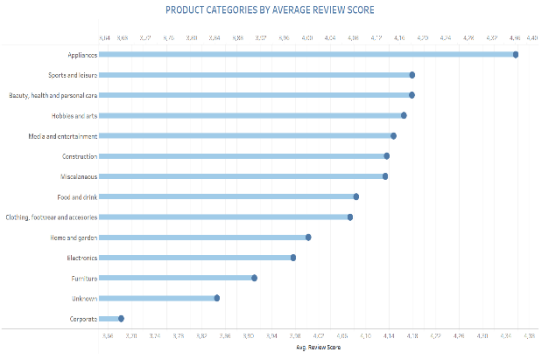
This histogram shows that most orders are actually delivered very early. It would be right to assume that Brazilians, like all humans, value their time dearly and would like to get a hold of their products as fast as possible. In order to appeal to them, the courier services must be improved. Faster deliveries mean faster money-making, more satisfied consumers, a better experience for the sellers as well, and everyone profits.

PRODUCT CATEGORY ANALYSIS

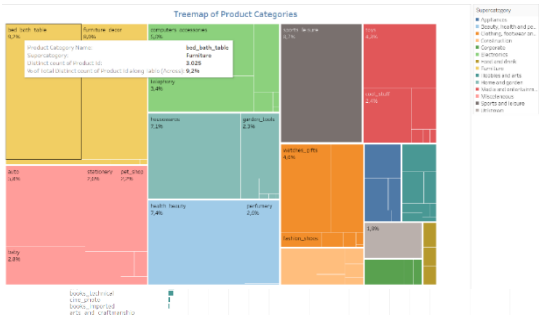
I believed that having over 70 product categories made it very hard to use the data this way and I had to segment it into only 11 lesser "Supercategories" of my own design. In order to do this, I could have used the same method as for the construction of the regions feature, but I chose to instead do it with Alteryx, which has an easier syntax. Alteryx is yet another powerful BI tool which is easier to use by even those who are not so technical or specialized in coding.



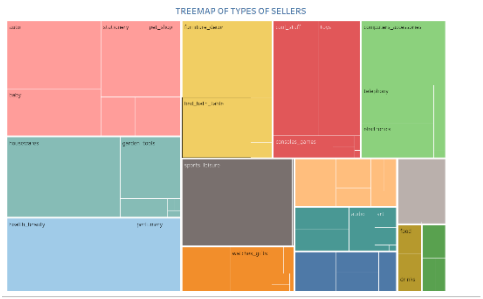
The most unpopular product super-category is Corporate. This is an issue, as it means that our corporate clients are not satisfied with the quality of their products and/or services. Corporate clients are important, as they have greater purchasing power, influence, reputation and can place greater orders. Losing the interest of this client segment could cause mayhem to our overall sales.



I suggest that a separate study should be made. Interestingly, furniture and electronics, some of the most sold products, merely make the bottom of the review score list with average scores of not even 4. The dotted bar chart with visible dual axis excluding zero, was chosen in order to help the viewer find the exact score more quickly.



Both the nested bar chart and the tree-map reveal the structure of the products. The bar chart deals with sales/category and the tree-map showcases the number of purchases from each category (the product popularity). Thus, the company's marketing department can know which products sell the best and which ones need an extra push in order the get those areas selling more. A future analysis of the shopping cart would reveal if certain types of products are often-times bought together. Information such as this could help with targeted ads on the website. This could lead to at least an increased level of customer engagement, or, why not, to an increase in sales of complementary products.

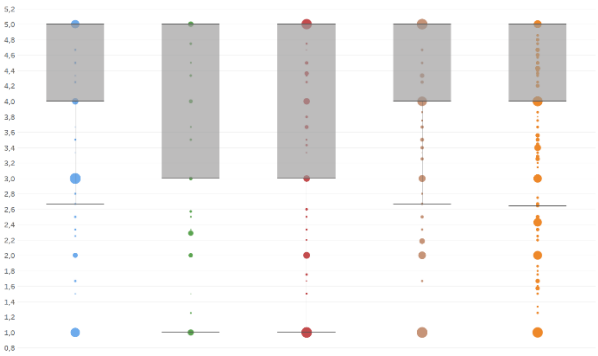


The second tree-map showcases the product super categories and category names according to the number the sellers that have them available. As we can observe, there are visible differences between the strongest and weakest products from each category. We can also study how each product super category and category sold over time (see Appendix).

CUSTOMER ANALYSIS

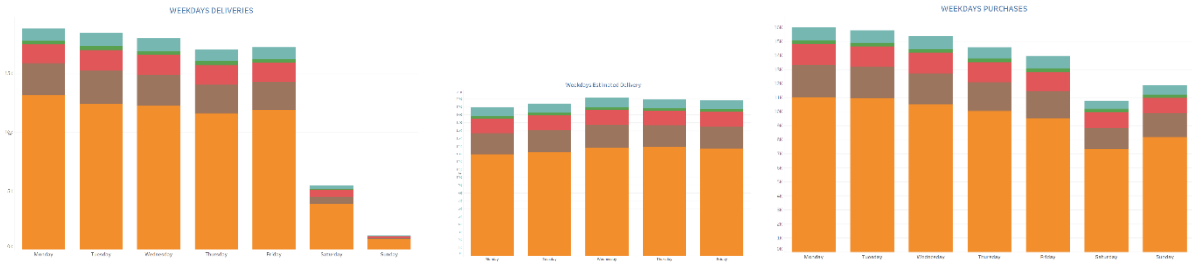
As can be seen from the box plot, some of the biggest clients to date (the size of the circle represents the sum of the payment values) have low average review ratings. These clients with payments in tenths of thousands and extremely negative reviews should be inspected more thoroughly. Our goal should be to attract as many clients as possible and retain those that we already have, with a focus on big spenders. The clients in the North show a wider array of

review scores. That is to say, the prevalence of lower reviews is higher there for some reason. The south has the most outliers, with extremely low scores compared to the majority.



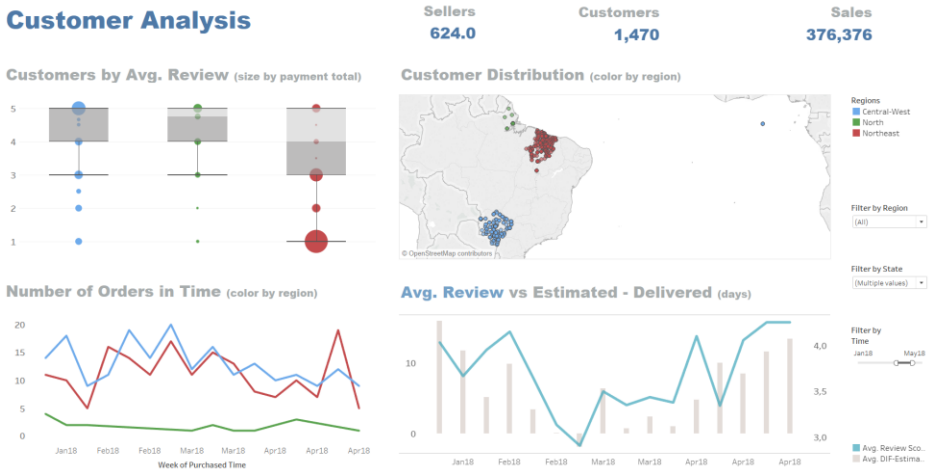
Big clients should be respected as such and I suggest that their satisfaction be prioritized. One way to address this issue is to give them special attention in terms of customer service, targeted marketing and especially delivery services.

WEEKLY TIME ANALYSIS



People tend to issue the most purchases on Mondays and decreasing slowly towards the weekend, with more people issuing orders on Sunday than on Saturday. Regarding the delivery times, again the trend is similar. Most products arrive on Monday and decrease until Thursday, with a slight increase on Friday. As expected, very few products reach their buyers on a Sunday. On the other hand, the estimated delivery times never include weekends, but they do happen.

DASHBOARD



The dashboard is the quintessence of the research work, as it contains those visualizations that can help the viewer decipher the answers to the most important research questions, while at the same time offering sufficient freedom to explore the dataset in new and innovative ways. I thought of the audience as the decision makers involved in Olist´s marketing and customer relationship. What I am trying to achieve with it is transmit the message that Olist could do a better job with its transportation services. It´s not doing horrible, by any means, but it could do better and sometimes it seems to choke under the pressure of its own sales success. Customers with problems should be pinpointed using the dashboard and their case should take further interest.

The color scheme is relatively simple. I used the same palette for the regions, for consistency across the report. Dark blue is used highlight important features. Grey is used for writing secondary information. Font and size are also used in this way. In the bottom-left chart, I also used blue in the title in order to help users identify the line chart more easily. The worksheets and other objects were added in tiled version, vertically or horizontally. Blank spaces were used around charts. To the right of the title, information about the number of sellers, number of customers and the actual sales values is provided for any given filtering or selection.

Four worksheets were selected, not too few, not too many, not to confuse the viewer, who can focus on only a handful of elements at once. I have chosen to introduce the box plot because it enables the viewer to easily identify persons of interest based on their value as a customer and on their overall experience with Olist's system. The chart can be used for highlighting. For example, we can select only those customers with negative reviews. The dual axis line/bar chart was chosen because it showcases how reviews are influenced by long shipment times. The map is relevant because Olist´s management is familiarized with the landscape of their business. It shows us that the Southern customers are all tightly grouped in a high-density area. Northwest shares similar characteristics on its coastline but many customer locations become scattered to the west and center, much like the Nordic regions and the center. This is consistent with the information found on Wikipedia. Selecting customers from the map highlights them. The map is fixed, so that it zooms in detail.

All the charts are dynamic. They can change dramatically through use of filters and highlighters. First, there are the geolocation filters: regions, states, allowing the user to drill down wherever interest rises. The filters allow for multiple choice of selection, which encourages comparisons. The regions legend can be used for highlighting. Secondly, there is a

range filter for time which works on the time series charts. It can be used to study certain moments in more detail.

FACTS, INSIGHTS AND OUTCOMES

MEANINGFUL FACTS

Business is doing good, there is an overall increase in sales and customers. Each region is doing as well as its economies are evolving. The size of the business in each one seems to perfectly relate to the geo-demographical characteristics and the infrastructure of Brazil itself. The average review score is positively tied to the delay time of the couriers. Spring was the weakest quarter for the past two years and the two events seem to be cyclical in nature. Many of Olist's biggest clients to date have had a bad experience with its services for some reason. All in all, there seems to be much predictability in this business.

ACTIONABLE INSIGHTS

The Southeast makes up the bulk of the business while the North pays out the least, as expected. The company should try to watch out for spring 2019 by coordinating more gracefully with the courier services and communicating them the problems that have been experienced in the past.

Olist would be wise to try to connect the markets with the customers more efficiently. It should establish additional contracts with different courier service providers that are specialized locally, in order to deliver its products faster and offer them more shipping options.

Olist should try to stop adding a freight value to each and every product. A cost vs benefit analysis could possibly establish if such a step would be beneficial to the company because it would certainly suit the customers.

According to a paper published by AlixPartners from March 2018, ‘’Consumers are increasingly seeking the convenience of fast, free shipping. Another study by UPS finds similar results. Shoppers are interested in free/cheap fast shipping. It is one of the main reasons why they choose or abandon certain marketplaces.

VALUABLE OUTCOMES

By using the dashboard, connected to a live stream of data, and updated in real-time, the decision-makers could watch out over the business and act with a quick response time.

By constructing the visualizations used in this research, we have found out many things about the progress of the company and established some pattern of expectations from our customers.

CONCLUSIONS AND RECOMMENDATIONS

The dataset already came split into multiple tables, encompassing a wide array of information. The task was to establish relationships between them and manipulate them by making use of keys. Once aggregated, the data presented the advantage of being more suited for BI data visualization tools such as Power BI or Tableau. While there are many things we can learn from separate tables, it actually is the broader picture we are interested in. Regardless, storing the data initially in structured relational systems presents many advantages, especially since there may be several sources of data acting as input.

The remote North has the lowest number of customers and suppliers. It is also the least populated, with a low density, low GDP/capita and is lacking in adequate transportation systems. Unlike the other regions, which benefit from high urbanization levels, with great purchasing power and are traversed by highways, the North is isolated. This is the region where Olist had the greatest delays in delivery times while the average reviews dropped accordingly.

According to the time series studying both the average review rating and the average difference between the estimated delivery time and the actual delivery time, there seems to be a problem in spring. Both in 2017 and in 2018, in the beginning of April. The cause for these delays needs to be further studied and could benefit from additional data. Nevertheless, they seem correlated and special precautions should be taken in this period in the future to avoid further delays and their subsequent complaints from customers. The business has grown considerably since 2016 and sometimes Olist´s courier partners seem to barely reach their deadlines. This is why, along with the increase in sales, we should also consider an increase in courier workers or taking up more options for transporting the goods.

The coastline brings in the most customers and the most revenue and is also the most accessible due to infrastructure. Reaching out to the more remote regions of Brazil does present multiple challenges. On the other hand, these remote regions bear little significance in the overall size

of the business. Thus, before jumping to any actions, a cost / benefit analysis should be done in order to ensure that the effort to connect to the customers does not outweigh the benefits.

Studying the time series, I reached the conclusion that there are some periods during the year which seem to repeat themselves in their effects. Spring brings about a period of decadence, for some reason, while we can also expect during 2018 the same surge in sales as was witnessed during last year's Black Friday. As more time passes and more data is gathered, the picture will become clearer and the company can become equipped

All in all, very few orders didn´t actually meet the deadline. The business is overall growing but there is room for continuous improvement and structured vigilance.

REFERENCES

[1] "Brazilian E-Commerce Public Dataset by Olist," [Online]. Available: https://www.kaggle.com/olistbr/brazilian-ecommerce. [Accessed 13 12 2019].

[2] EDUCBA, "Json vs CSV," [Online]. Available: https://www.educba.com/json-vs-csv/. [Accessed 10 12 2019].

[3] "Stackoverflow - How to Convert CSV into multiline JSON_," [Online]. Available: https://stackoverflow.com/questions/19697846/how-to-convert-csv-file-to-multiline-json. [Accessed 08 12 2019].

[4] Wikipedia, "The regions of Brazil," [Online]. Available: https://en.wikipedia.org/wiki/Regions_of_Brazil . [Accessed 14 12 2019].

[5] Ebanx, "Sell more in Brazil wit hthe Boleto Bancario," [Online]. Available: https://business.ebanx.com/en/brazil/payment-methods/boleto-bancario. [Accessed 13 12 2019].

APPENDIX 1 – SQL CODE SNIPPETS

QUESTIONS

```
-- View database
SELECT * FROM olistDatabaseFinal;

-- What are the top grossing regions, ordered by state?
SELECT customer_region
FROM olistDatabaseFinal
WHERE
ORDER by customer_state
GROUP BY customer_region;

-- What is the total sum of sales in the South Region?
SELECT SUM(price)
FROM olistDatabaseFinal
WHERE customer_region like 'South';
-- A: 2040492.25999973

-- What is the total number of customers in the South Region?
SELECT COUNT(customer_unique_id)
FROM olistDatabaseFinal
WHERE customer_region like 'South';
-- A: 16846

-- What are the top grossing regions?
select top 5 customer_region, sum(price)
from OlistDB
group by customer_region
Order by sum(price) desc;

-- Where do we get the biggest delays (total) in shipment?
select top 5 customer_region, sum(order_delivery_difference_days)
from OlistDB
group by customer_region
Order by sum(order_delivery_difference_days) asc;

-- Where do we get the biggest delays (average) in shipment?
select top 5 customer_region, avg(order_delivery_difference_days)
from OlistDB
group by customer_region
Order by avg(order_delivery_difference_days) asc;

-- Who are the top 3 sellers?
select top 3 seller_id, sum(price)
from OlistDB
group by seller_id
Order by sum(price) desc;

-- Frequency table. The first value is the mode.
SELECT
    payment_value,
    COUNT(*) as count
FROM dbo.payments
GROUP BY payment_value
ORDER BY COUNT(*) DESC;


SELECT SUM SELECT TOP 20 PERCENT payment_value FROM olistDatabaseFinal
WHERE customer_region = 'North'
```

```sql
SELECT TOP 20 PERCENT payment_value FROM olistDatabaseFinal AS top20
WHERE customer_region = 'North'
ORDER BY payment_value DESC
```

ADD COLUMNS

```sql
SELECT *
FROM olistDatabase
ORDER BY pa


SELECT TOP 20 PERCENT
FROM olistDatabase;

SELECT *
FROM olistDatabase
ORDER BY customer_state, customer_city ASC
;

SELECT o_r_p_c.order_id, o_r_p_c.review_id, o_r_p_c.customer_id, o_r_p_c.review_score,
              o_r_p_c.order_status, o_r_p_c.order_delivery_difference_days,
              o_r_p_c.payment_type, o_r_p_c.payment_value,
               o_r_p_c.customer_unique_id, o_r_p_c.customer_city,
o_r_p_c.customer_state,
               o_r_p_c.order_item_id, o_r_p_c.product_id, o_r_p_c.seller_id,
o_r_p_c.price, o_r_p_c.freight_value,
               o_r_p_c.product_category_name, o_r_p_c.seller_city,
o_r_p_c.seller_state,
               o.order_purchase_timestamp

INTO olistDatabaseFinal
FROM olistDatabase AS o_r_p_c
INNER JOIN orders AS o
ON o_r_p_c.order_id = o.order_id;

ALTER TABLE olistDatabaseFinal
ADD customer_region VARCHAR(50), seller_region VARCHAR(50);

UPDATE olistDatabaseFinal
SET   customer_region = 'Northeast'
WHERE customer_state = 'MA'
AND   customer_state = 'BA'
AND   customer_state = 'SE'
AND   customer_state = 'AL'
AND   customer_state = 'PE'
AND   customer_state = 'PB'
AND   customer_state = 'RN'
AND   customer_state = 'CI'
AND   customer_state = 'PI';

UPDATE olistDatabaseFinal
SET   customer_region = 'Southeast'
WHERE customer_state = 'SP'
AND   customer_state = 'RJ'
AND   customer_state = 'ES'
AND   customer_state = 'MG'
AND customer_state IS NULL;

SELECT * from olistDatabaseFinal;

UPDATE olistDatabaseFinal
```

```sql
SET   customer_region = 'Northeast'
WHERE customer_state like 'MA'
OR    customer_state like 'BA'
OR    customer_state like 'SE'
OR    customer_state like 'AL'
OR    customer_state like 'PE'
OR    customer_state like 'PB'
OR    customer_state like 'RN'
OR    customer_state like 'CE'
OR    customer_state like 'PI';

UPDATE olistDatabaseFinal
SET   customer_region = 'Southeast'
WHERE customer_state like 'SP'
OR    customer_state like 'RJ'
OR    customer_state like 'ES'
OR    customer_state like 'MG';

UPDATE olistDatabaseFinal
SET   customer_region = 'Central-West'
WHERE customer_state like 'MT'
OR    customer_state like 'MS'
OR    customer_state like 'GO'
OR    customer_state like 'DF';

UPDATE olistDatabaseFinal
SET   customer_region = 'South'
WHERE customer_state like 'PR'
OR    customer_state like 'SC'
OR    customer_state like 'RS';

UPDATE olistDatabaseFinal
SET   customer_region = 'North'
WHERE customer_state like 'AC'
OR    customer_state like 'AM'
OR    customer_state like 'RR'
OR    customer_state like 'PA'
OR    customer_state like 'AP'
OR    customer_state like 'RO'
OR    customer_state like 'TO';

SELECT * FROM olistDatabaseFinal;

UPDATE olistDatabaseFinal
SET   seller_region = 'Northeast'
WHERE seller_state like 'MA'
OR    seller_state like 'BA'
OR    seller_state like 'SE'
OR    seller_state like 'AL'
OR    seller_state like 'PE'
OR    seller_state like 'PB'
OR    seller_state like 'RN'
OR    seller_state like 'CE'
OR    seller_state like 'PI';

UPDATE olistDatabaseFinal
SET   seller_region = 'Southeast'
WHERE seller_state like 'SP'
OR    seller_state like 'RJ'
OR    seller_state like 'ES'
OR    seller_state like 'MG';
```

```
UPDATE olistDatabaseFinal
SET   seller_region = 'Central-West'
WHERE seller_state like 'MT'
OR   seller_state like 'MS'
OR   seller_state like 'GO'
OR   seller_state like 'DF';

UPDATE olistDatabaseFinal
SET   seller_region = 'South'
WHERE seller_state like 'PR'
OR   seller_state like 'SC'
OR   seller_state like 'RS';

UPDATE olistDatabaseFinal
SET   seller_region = 'North'
WHERE seller_state like 'AC'
OR   seller_state like 'AM'
OR   seller_state like 'RR'
OR   seller_state like 'PA'
OR   seller_state like 'AP'
OR   seller_state like 'RO'
OR   seller_state like 'TO';

SELECT * FROM olistDatabaseFinal;
```

DATEDIFF

```
ALTER TABLE orders
ADD order_delivery_difference_days AS (DATEDIFF (day, order_estimated_delivery_date,
order_delivered_customer_date));

ALTER TABLE orders
DROP COLUMN order_delivery_difference;

SELECT * from orders;

SELECT reviews.order_id, reviews.review_id, reviews.review_score,
              orders.order_status, orders.order_delivery_difference_days,
orders.customer_id
INTO orders_reviews
FROM orders
INNER JOIN reviews
ON orders.order_id = reviews.order_id;

SELECT * FROM orders_reviews;

SELECT COUNT (DISTINCT (order_id)) FROM orders_reviews;

SELECT MIN(order_delivery_difference_days) As SmallestDiff
FROM orders_reviews;
-- -147 (This order came way before the estimated time)

SELECT MAX(order_delivery_difference_days) As LargestDiff
FROM orders_reviews;
-- 188 (This order was very late)

SELECT AVG(CAST (order_delivery_difference_days as float))
FROM orders_reviews;
-- -11.8812427200478 (Overall, the orders were a bit late)

SELECT o_r.order_id, o_r.review_id, o_r.review_score, o_r.customer_id,
              o_r.order_status, o_r.order_delivery_difference_days,
```

```
            p.payment_type, p.payment_value
INTO orders_reviews_payments
FROM orders_reviews AS o_r
INNER JOIN payments AS p
ON o_r.order_id = p.order_id;

SELECT * from orders_reviews_payments;

SELECT o_r_p.order_id, o_r_p.review_id, o_r_p.customer_id, o_r_p.review_score,
            o_r_p.order_status, o_r_p.order_delivery_difference_days,
            o_r_p.payment_type, o_r_p.payment_value,
             c.customer_unique_id, c.customer_city, c.customer_state
INTO orders_reviews_payments_customers
FROM orders_reviews_payments AS o_r_p
INNER JOIN customers AS c
ON o_r_p.customer_id = c.customer_id;

SELECT * FROM orders_reviews_payments_customers;

SELECT o_r_p_c.order_id, o_r_p_c.review_id, o_r_p_c.customer_id, o_r_p_c.review_score,
            o_r_p_c.order_status, o_r_p_c.order_delivery_difference_days,
            o_r_p_c.payment_type, o_r_p_c.payment_value,
             o_r_p_c.customer_unique_id, o_r_p_c.customer_city,
o_r_p_c.customer_state,
             i.order_item_id, i.product_id, i.seller_id, i.price, i.freight_value
INTO orders_reviews_payments_customers_items
FROM orders_reviews_payments_customers AS o_r_p_c
FULL OUTER JOIN items AS i
ON o_r_p_c.order_id = i.order_id;

SELECT * FROM orders_reviews_payments_customers_items;

SELECT COUNT(order_id)
FROM orders_reviews_payments_customers_items;
--57420
SELECT COUNT(DISTINCT order_id)
FROM orders_reviews_payments_customers_items;
--57420
SELECT COUNT(product_id)
FROM orders_reviews_payments_customers_items;
--57420
SELECT COUNT(DISTINCT product_id)
FROM orders_reviews_payments_customers_items;
--57420

SELECT o_r_p_c.order_id, o_r_p_c.review_id, o_r_p_c.customer_id, o_r_p_c.review_score,
            o_r_p_c.order_status, o_r_p_c.order_delivery_difference_days,
            o_r_p_c.payment_type, o_r_p_c.payment_value,
             o_r_p_c.customer_unique_id, o_r_p_c.customer_city,
o_r_p_c.customer_state,
             o_r_p_c.order_item_id, o_r_p_c.product_id, o_r_p_c.seller_id,
o_r_p_c.price, o_r_p_c.freight_value,
             p.product_category_name

INTO orders_reviews_payments_customers_items_products
FROM orders_reviews_payments_customers_items AS o_r_p_c
INNER JOIN products AS p
ON o_r_p_c.product_id = p.product_id;

SELECT * FROM orders_reviews_payments_customers_items_products;

SELECT o_r_p_c.order_id, o_r_p_c.review_id, o_r_p_c.customer_id, o_r_p_c.review_score,
```

```
            o_r_p_c.order_status, o_r_p_c.order_delivery_difference_days,
            o_r_p_c.payment_type, o_r_p_c.payment_value,
             o_r_p_c.customer_unique_id, o_r_p_c.customer_city,
o_r_p_c.customer_state,
             o_r_p_c.order_item_id, o_r_p_c.product_id, o_r_p_c.seller_id,
o_r_p_c.price, o_r_p_c.freight_value,
             o_r_p_c.product_category_name, s.seller_city, s.seller_state

INTO olistDatabase
FROM orders_reviews_payments_customers_items_products AS o_r_p_c
INNER JOIN sellers AS s
ON o_r_p_c.seller_id = s.seller_id;

SELECT * FROM olistDatabase;
```

REVIEWS

```
SELECT * FROM reviews;

SELECT MIN(review_score) As SmallestReview
FROM reviews;

SELECT MAX(review_score) As LargestReview
FROM reviews;

SELECT COUNT(DISTINCT review_id)
FROM reviews;
--99173

SELECT COUNT(review_id)
FROM reviews
WHERE review_comment_message IS NULL;
--58254

SELECT COUNT(review_id)
FROM reviews
WHERE review_comment_title IS NOT NULL;
--11713

SELECT COUNT(review_id)
FROM reviews
WHERE review_score = 5;
--57420

SELECT SUM(CAST (review_score as int))
FROM reviews;
--407089

SELECT AVG(CAST (review_score as float))
FROM reviews;
--4.07089 aprox 4
```

ORDERS

```
ALTER TABLE orders
ADD order_delivery_difference_days AS (DATEDIFF (day, order_estimated_delivery_date,
order_delivered_customer_date));

ALTER TABLE orders
DROP COLUMN order_delivery_difference;

SELECT * from orders;
```

```sql
SELECT reviews.order_id, reviews.review_id, reviews.review_score,
             orders.order_status, orders.order_delivery_difference_days
INTO orders_reviews
FROM orders
INNER JOIN reviews
ON orders.order_id = reviews.order_id;

SELECT * FROM orders_reviews;

SELECT COUNT (DISTINCT (order_id)) FROM orders_reviews;

SELECT MIN(order_delivery_difference_days) As SmallestDiff
FROM orders_reviews;
-- -147 (This order came way before the estimated time)

SELECT MAX(order_delivery_difference_days) As LargestDiff
FROM orders_reviews;
-- 188 (This order was very late)

SELECT AVG(CAST (order_delivery_difference_days as float))
FROM orders_reviews;
-- -11.8812427200478 (Overall, the orders were a bit late)
```

PAYMENTS

```sql
SELECT o_r.order_id, o_r.review_id, o_r.review_score,
             o_r.order_status, o_r.order_delivery_difference_days,
             p.payment_type, p.payment_value
INTO orders_reviews_payments
FROM orders_reviews AS o_r
INNER JOIN payments AS p
ON o_r.order_id = p.order_id;

SELECT * from orders_reviews_payments;



SELECT o_r_p.order_id, o_r_p.review_id, o_r_p.customer_id, o_r_p.review_score,
             o_r_p.order_status, o_r_p.order_delivery_difference_days,
             o_r_p.payment_type, o_r_p.payment_value,
              c.customer_unique_id, c.customer_city, c.customer_state
INTO orders_reviews_payments_customers
FROM orders_reviews_payments AS o_r_p
INNER JOIN customers AS c
ON o_r_p.customer_id = c.customer_id;

SELECT * FROM orders_reviews_payments_customers;

ITEMS

SELECT o_r_p_c.order_id, o_r_p_c.review_id, o_r_p_c.customer_id, o_r_p_c.review_score,
             o_r_p_c.order_status, o_r_p_c.order_delivery_difference_days,
             o_r_p_c.payment_type, o_r_p_c.payment_value,
              o_r_p_c.customer_unique_id, o_r_p_c.customer_city,
o_r_p_c.customer_state,
              i.order_item_id, i.product_id, i.seller_id, i.price, i.freight_value
INTO orders_reviews_payments_customers_items
FROM orders_reviews_payments_customers AS o_r_p_c
FULL OUTER JOIN items AS i
ON o_r_p_c.order_id = i.order_id;

SELECT * FROM orders_reviews_payments_customers_items;
```

```sql
SELECT o_r_p_c.order_id, o_r_p_c.review_id, o_r_p_c.customer_id, o_r_p_c.review_score,
             o_r_p_c.order_status, o_r_p_c.order_delivery_difference_days,
             o_r_p_c.payment_type, o_r_p_c.payment_value,
              o_r_p_c.customer_unique_id, o_r_p_c.customer_city,
o_r_p_c.customer_state,
              i.order_item_id, i.product_id, i.seller_id, i.price, i.freight_value
INTO orders_reviews_payments_customers_items
FROM orders_reviews_payments_customers AS o_r_p_c
FULL OUTER JOIN items AS i
ON o_r_p_c.order_id = i.order_id;

SELECT * FROM orders_reviews_payments_customers_items;

SELECT COUNT(order_id)
FROM orders_reviews_payments_customers_items;
--57420
SELECT COUNT(DISTINCT order_id)
FROM orders_reviews_payments_customers_items;
--57420
SELECT COUNT(product_id)
FROM orders_reviews_payments_customers_items;
--57420
SELECT COUNT(DISTINCT product_id)
FROM orders_reviews_payments_customers_items;
--57420

SELECT o_r_p_c.order_id, o_r_p_c.review_id, o_r_p_c.customer_id, o_r_p_c.review_score,
             o_r_p_c.order_status, o_r_p_c.order_delivery_difference_days,
             o_r_p_c.payment_type, o_r_p_c.payment_value,
              o_r_p_c.customer_unique_id, o_r_p_c.customer_city,
o_r_p_c.customer_state,
              o_r_p_c.order_item_id, o_r_p_c.product_id, o_r_p_c.seller_id,
o_r_p_c.price, o_r_p_c.freight_value,
              p.product_category_name

INTO orders_reviews_payments_customers_items_products
FROM orders_reviews_payments_customers_items AS o_r_p_c
INNER JOIN products AS p
ON o_r_p_c.product_id = p.product_id;

SELECT * FROM orders_reviews_payments_customers_items_products;

SELECT o_r_p_c.order_id, o_r_p_c.review_id, o_r_p_c.customer_id, o_r_p_c.review_score,
             o_r_p_c.order_status, o_r_p_c.order_delivery_difference_days,
             o_r_p_c.payment_type, o_r_p_c.payment_value,
              o_r_p_c.customer_unique_id, o_r_p_c.customer_city,
o_r_p_c.customer_state,
              o_r_p_c.order_item_id, o_r_p_c.product_id, o_r_p_c.seller_id,
o_r_p_c.price, o_r_p_c.freight_value,
              o_r_p_c.product_category_name, s.seller_city, s.seller_state

INTO olistDatabase
FROM orders_reviews_payments_customers_items_products AS o_r_p_c
INNER JOIN sellers AS s
ON o_r_p_c.seller_id = s.seller_id;

SELECT * FROM olistDatabase;
```

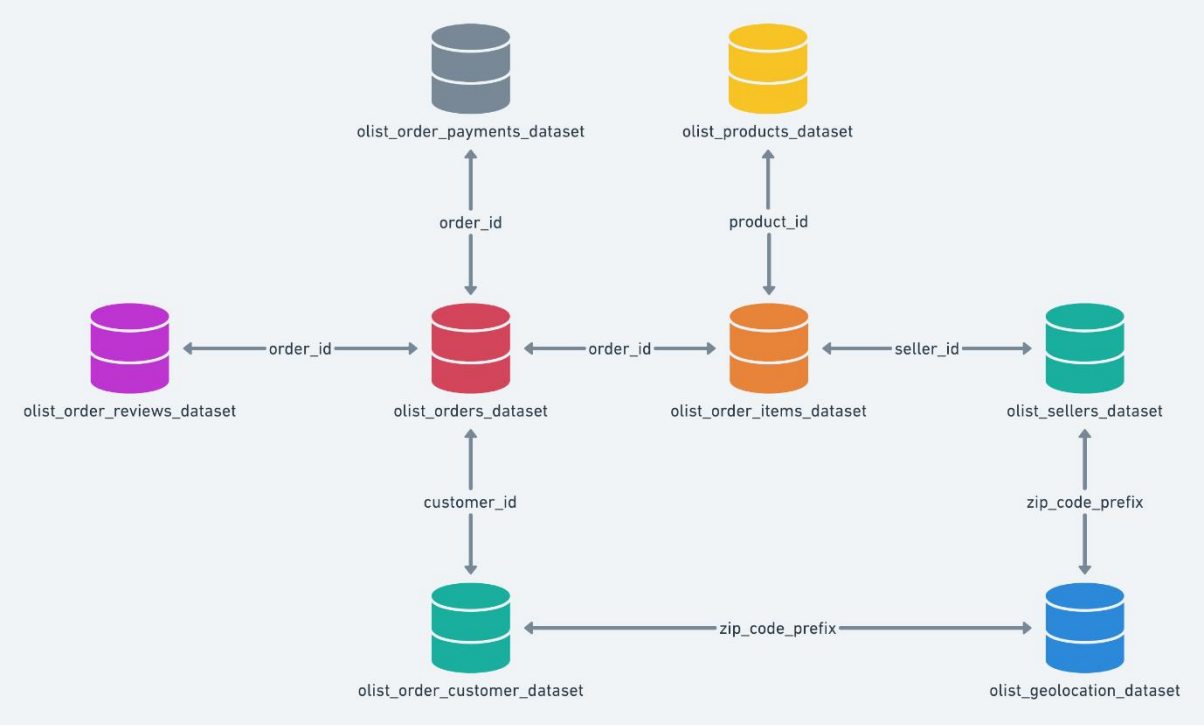
MINOR STATISTICS

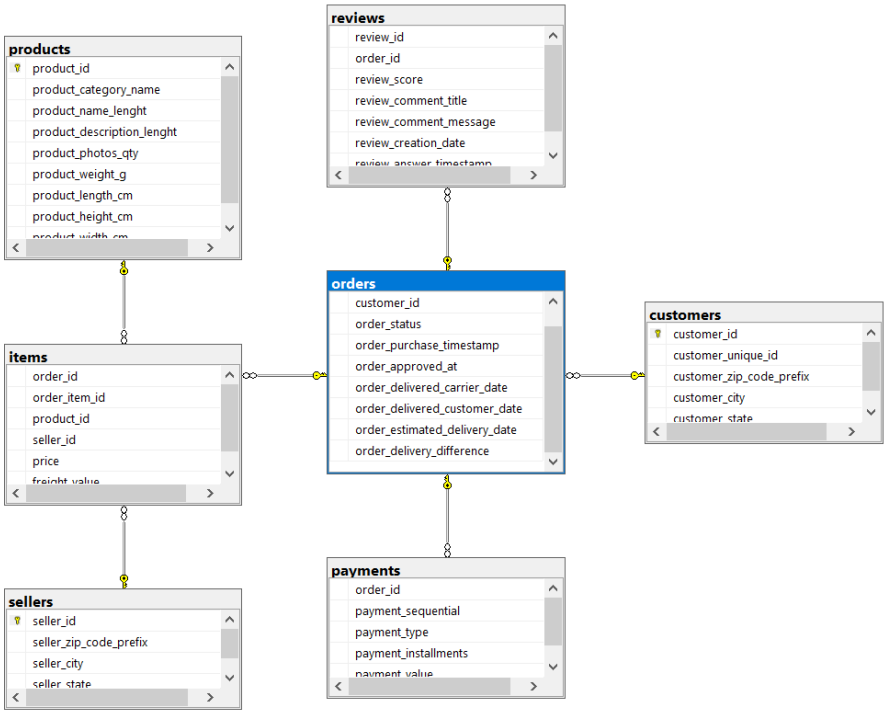
```sql
SELECT * FROM reviews;

SELECT MIN(review_score) As SmallestReview
FROM reviews;

SELECT MAX(review_score) As LargestReview
FROM reviews;

SELECT COUNT(DISTINCT review_id)
FROM reviews;
--99173

SELECT COUNT(review_id)
FROM reviews
WHERE review_comment_message IS NULL;
--58254

SELECT COUNT(review_id)
FROM reviews
WHERE review_comment_title IS NOT NULL;
--11713

SELECT COUNT(review_id)
FROM reviews
WHERE review_score = 5;
--57420

SELECT SUM(CAST (review_score as int))
FROM reviews;
--407089

SELECT AVG(CAST (review_score as float))
FROM reviews;
--4.07089 aprox 4
```

APPENDIX 2 - DATA VISUALIZATIONS

The appendix section covers, in order, the data schema and the data visualizations used across the results chapter.

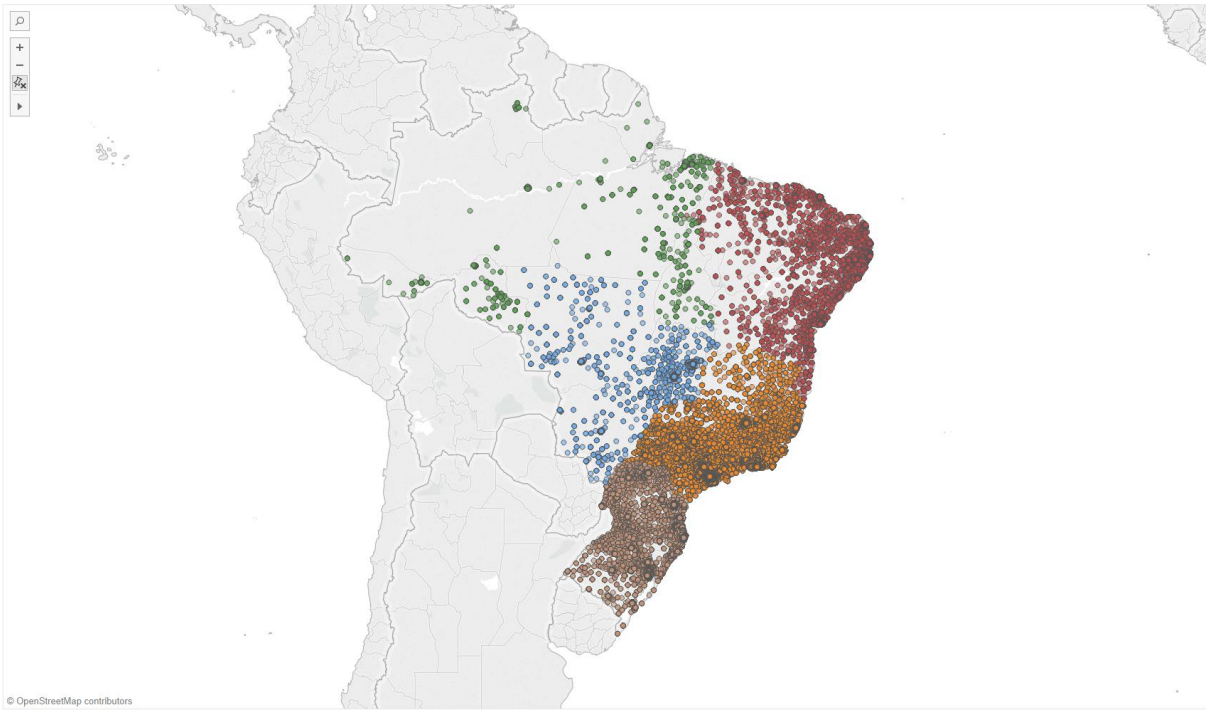
1) Data schemas



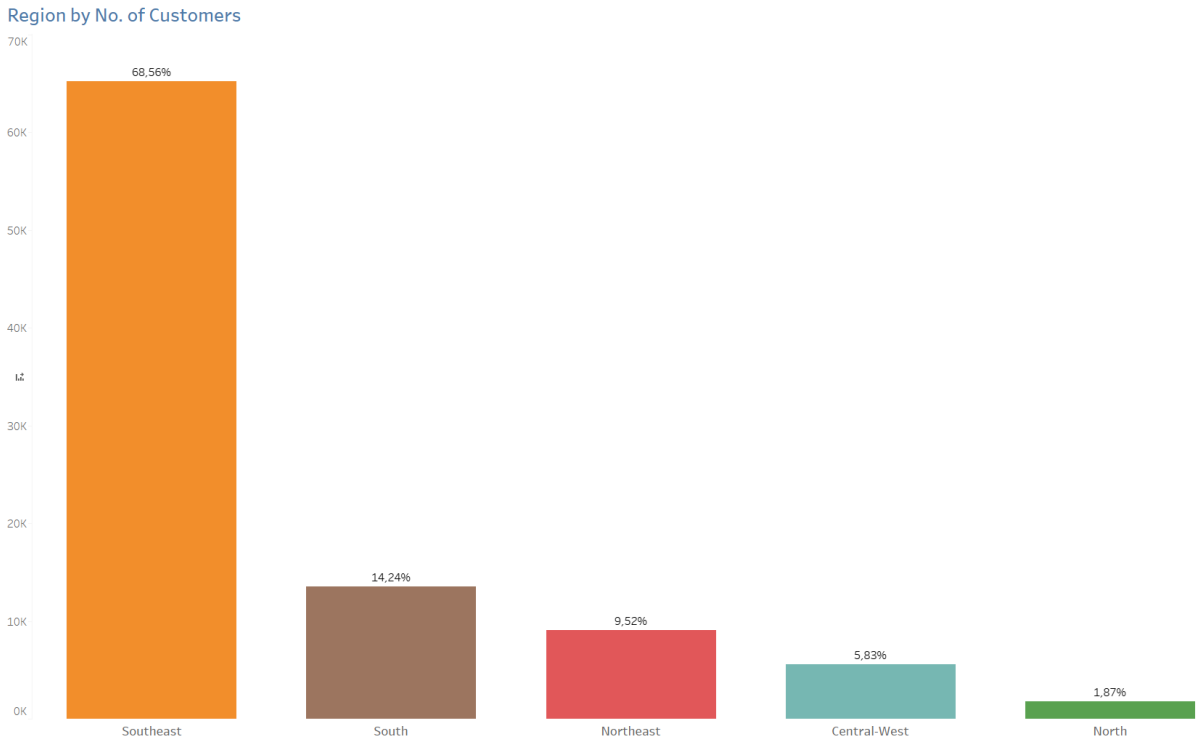


2) Map of Brazil

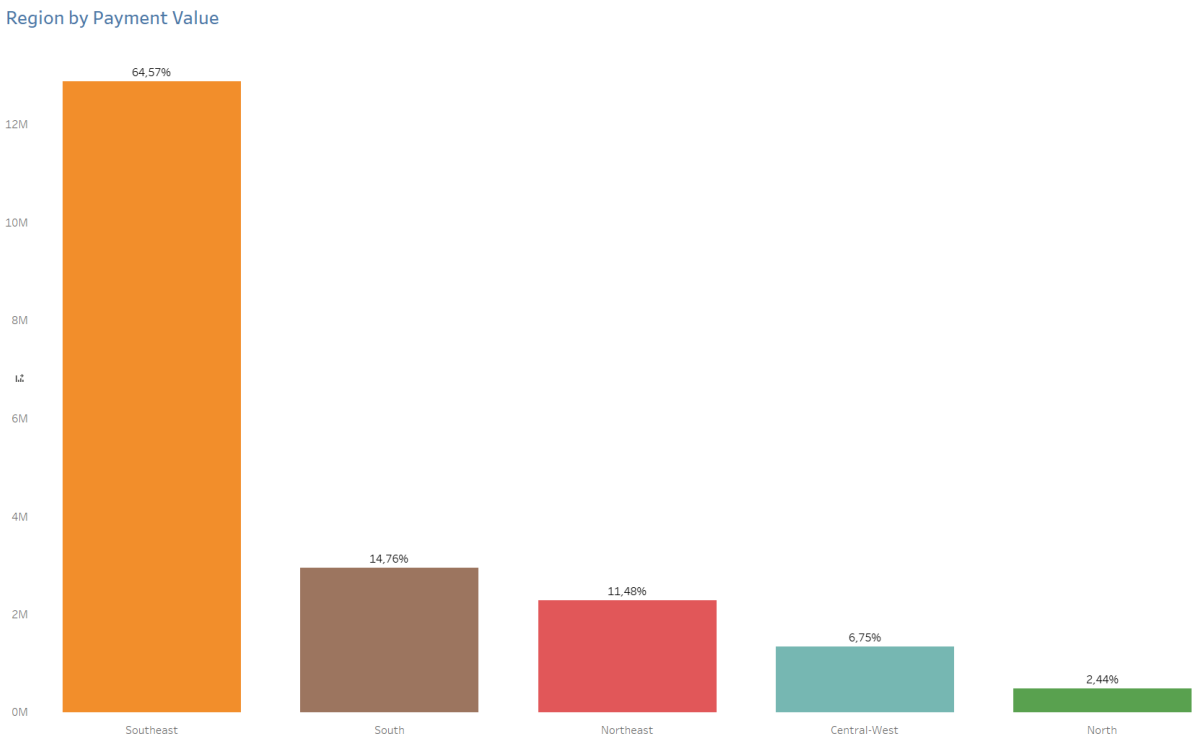
3) Location of customers



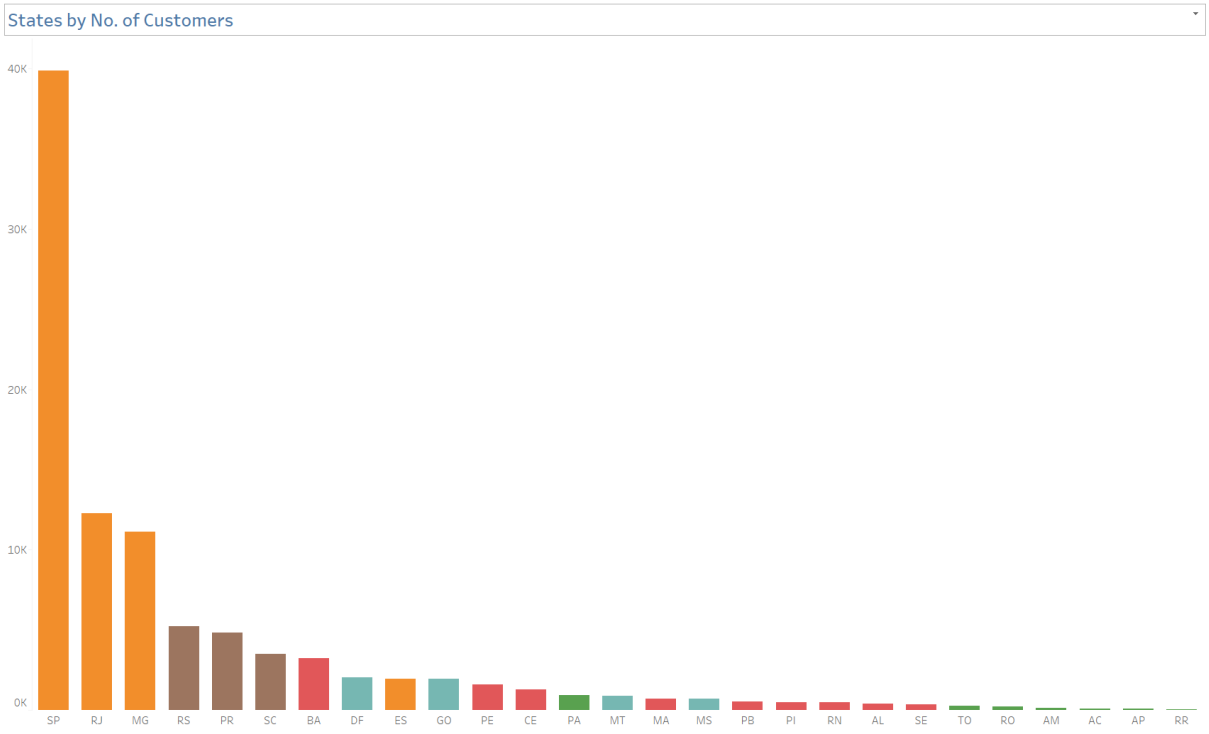
4) Bar chart of region by number of customers



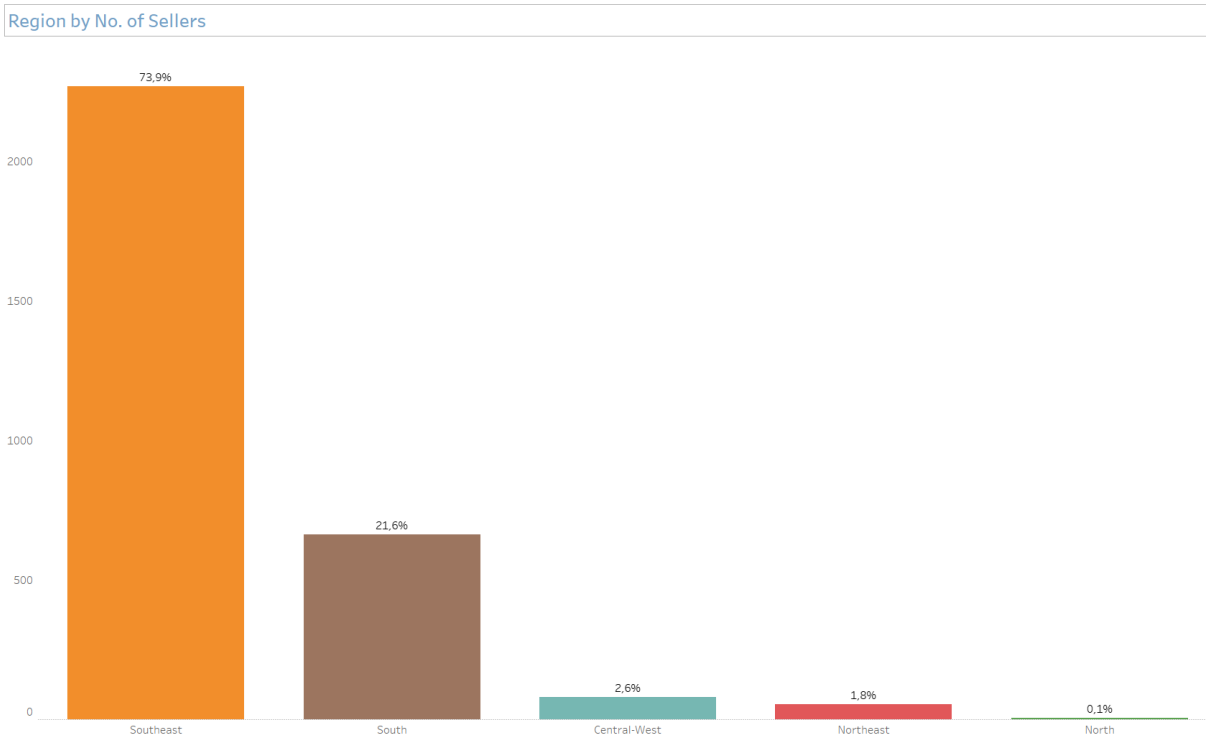
5) Bar chart of region by payment value



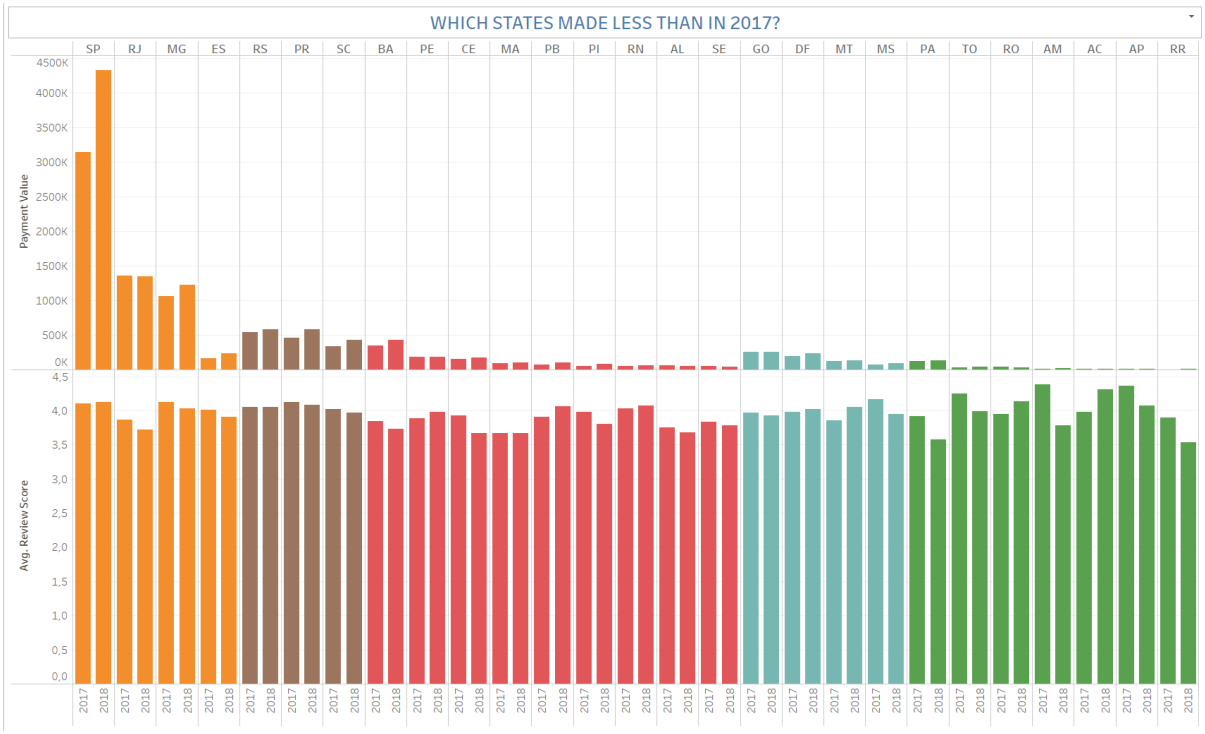
6) Bar chart of states by number of customers



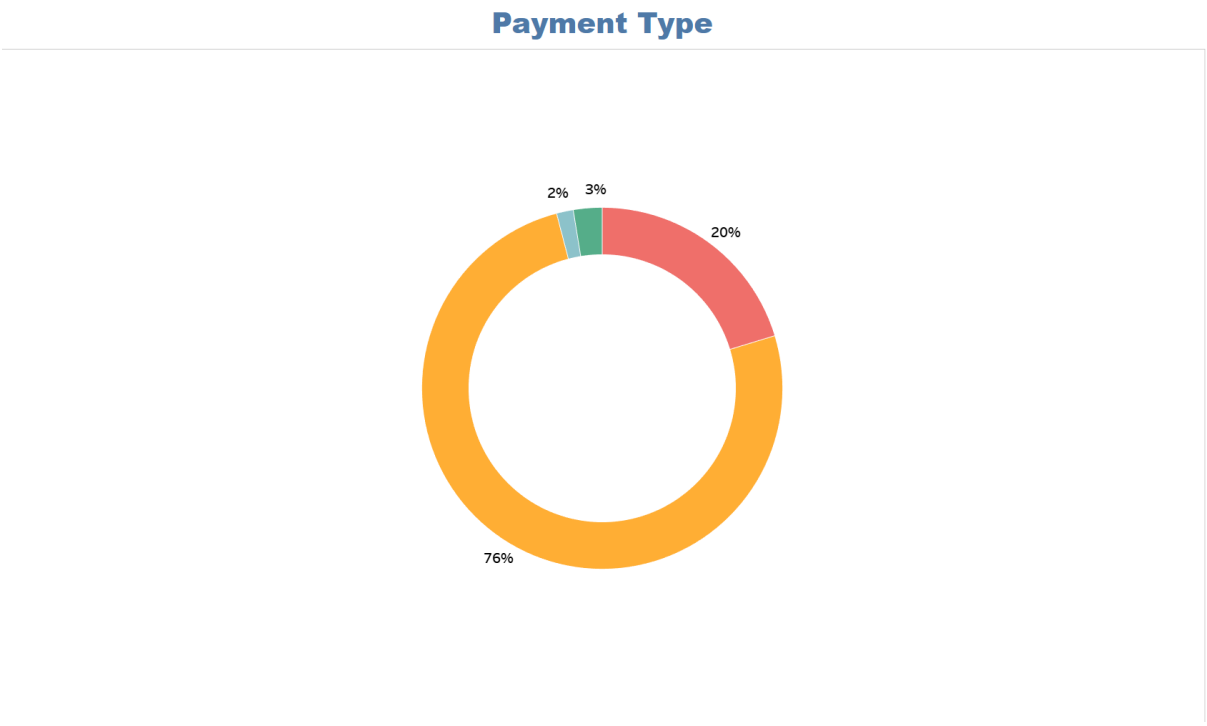
7) Bar chart of regions by number of sellers



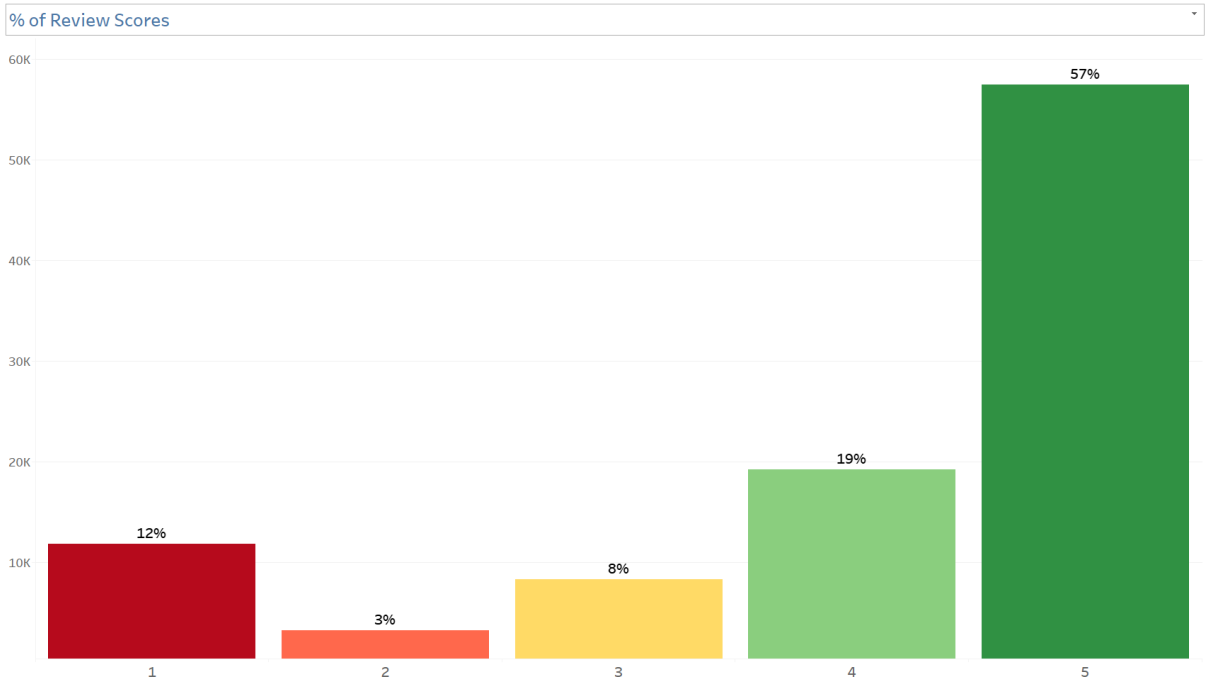
8) Dual Bar chart



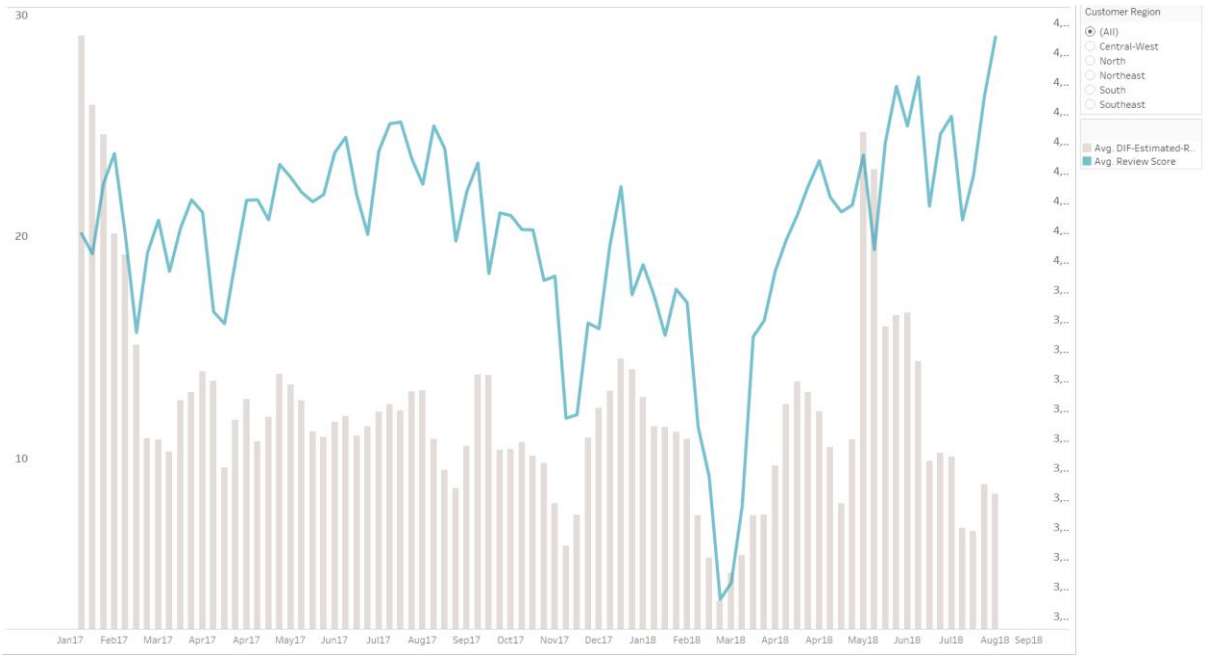
9) Donut chart of payment types



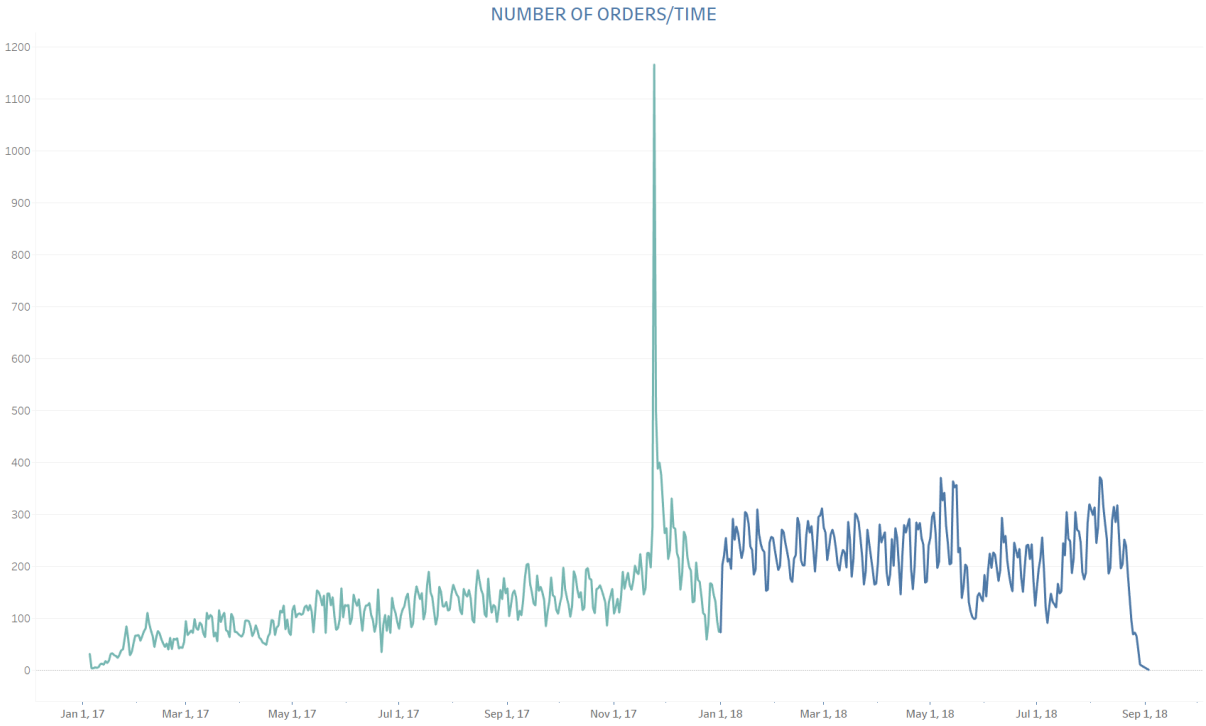
10) bar chart of review scores



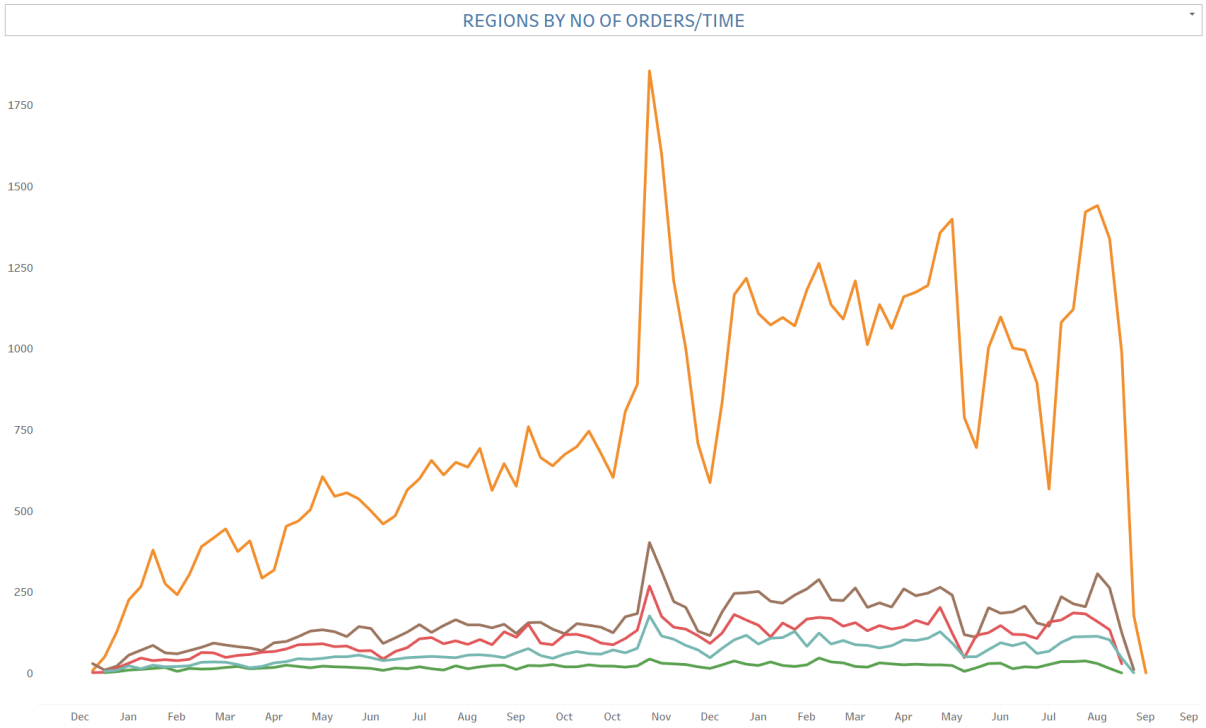
11) Bar/line chart of average delays vs average review scores



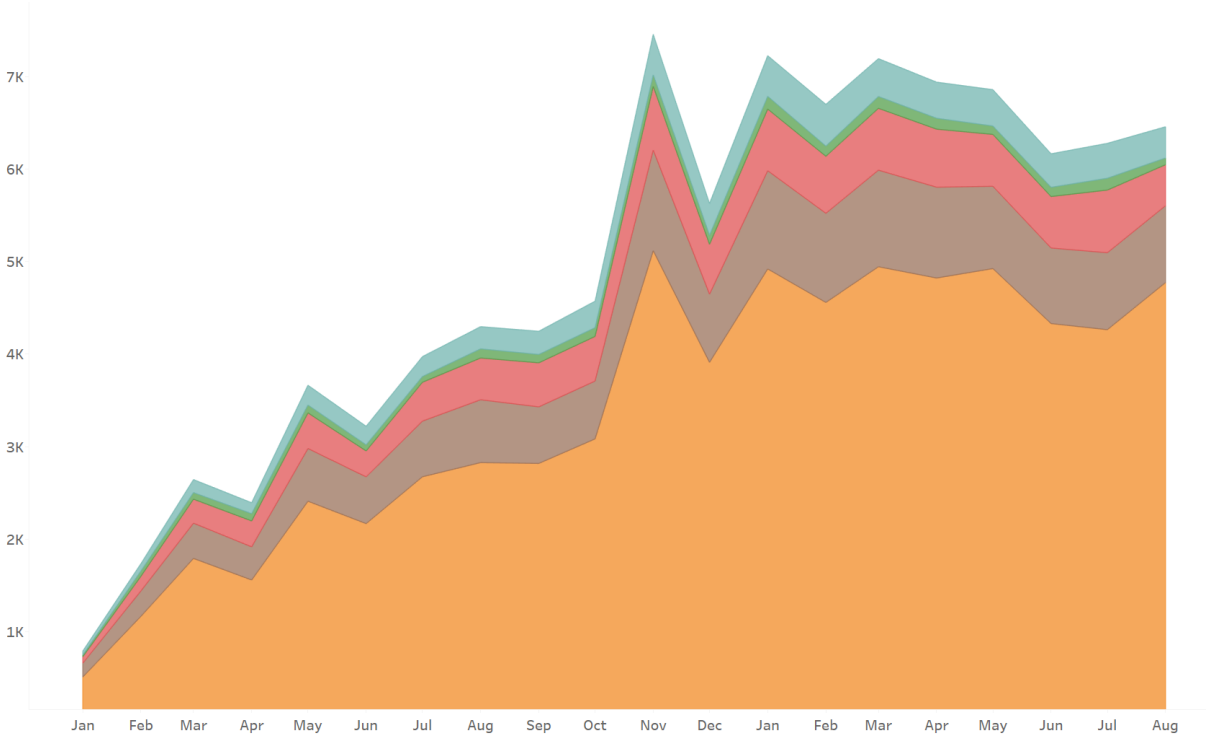
12) Line chart of number of orders



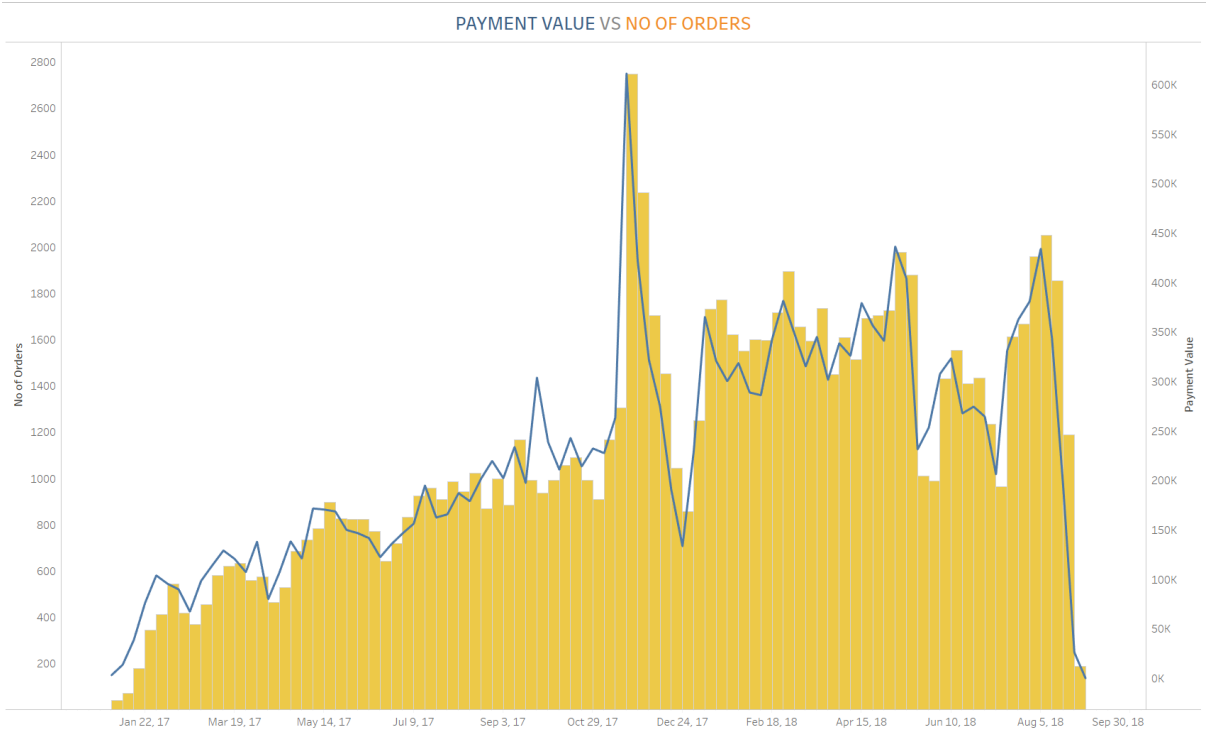
13) Line chart of number of orders per region



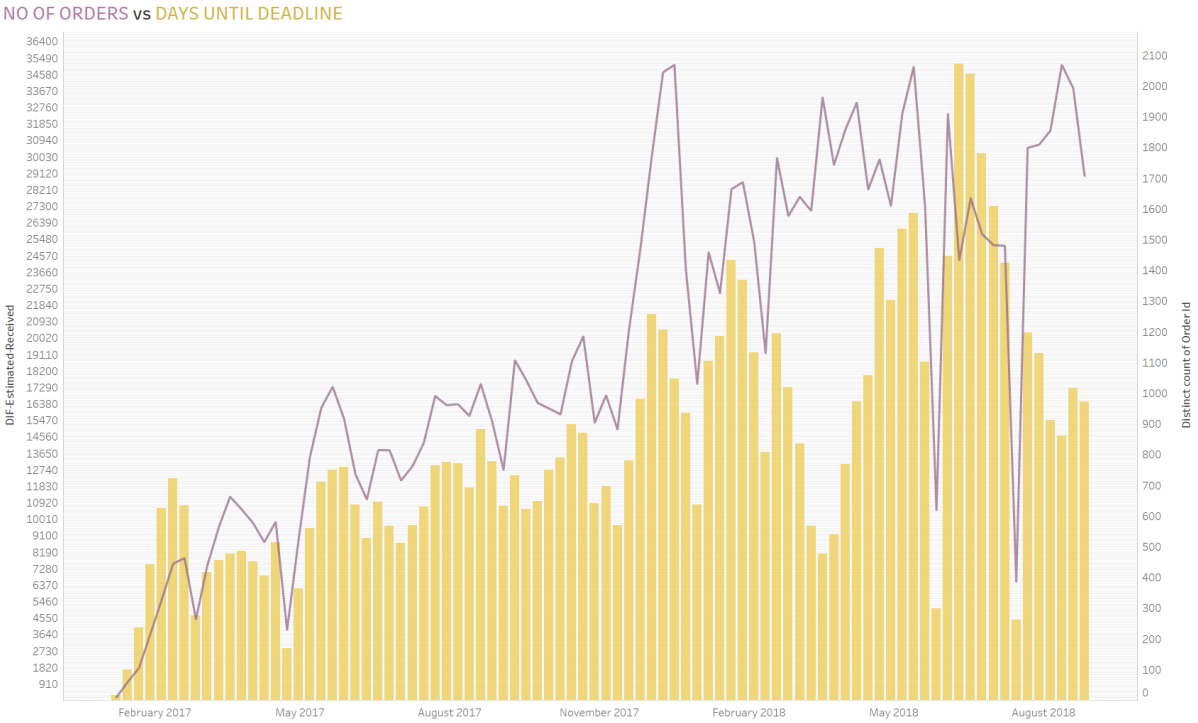
14) Stacked area chart of number of orders per region



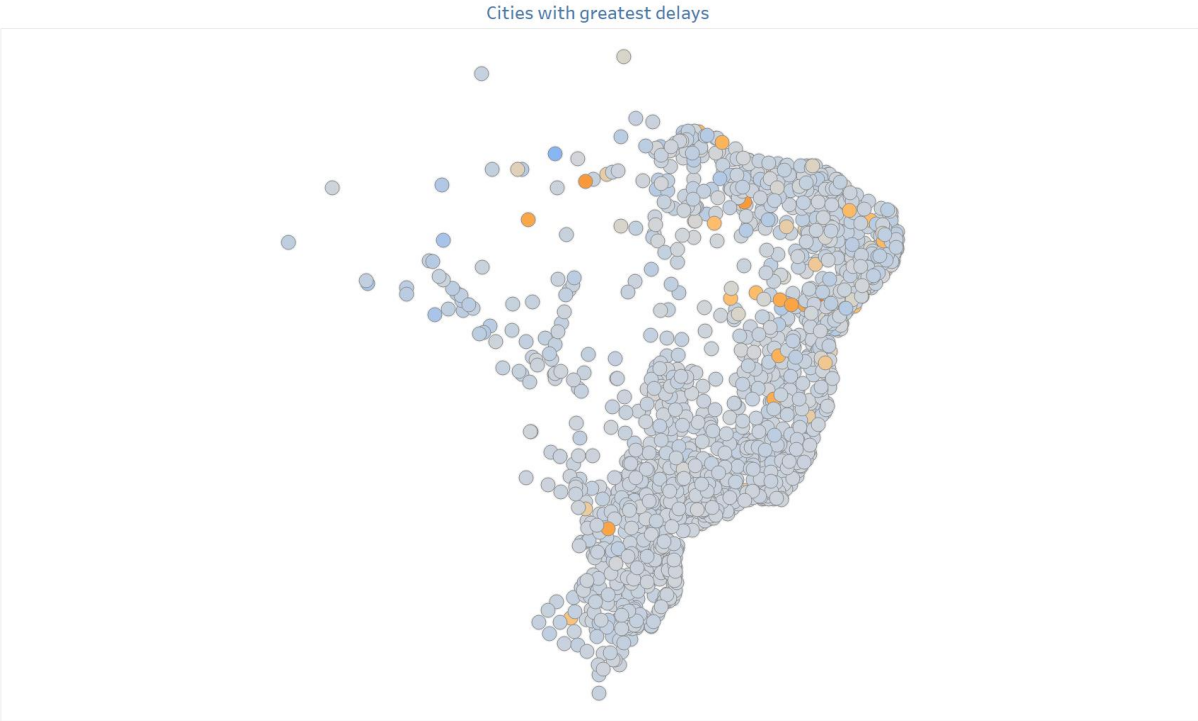
15) Bar/line chart of payment value vs no. of orders



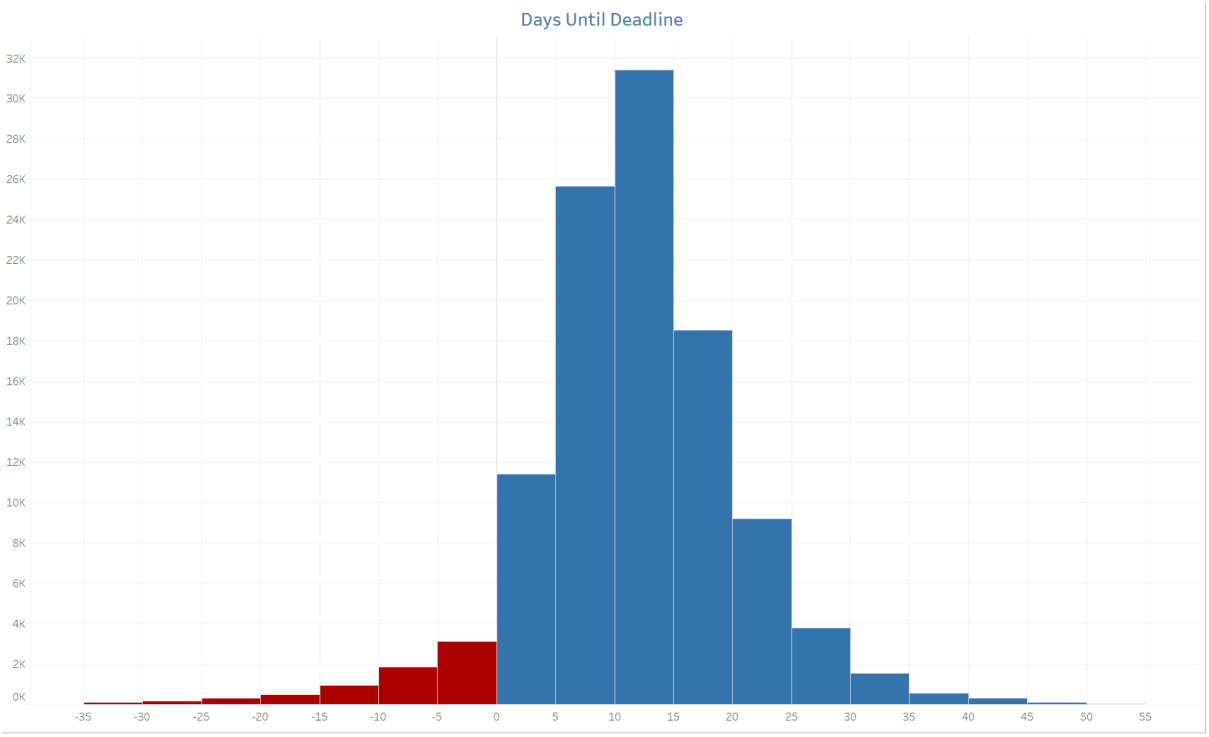
16) Bar/line chart of no. of orders vs days until deadline



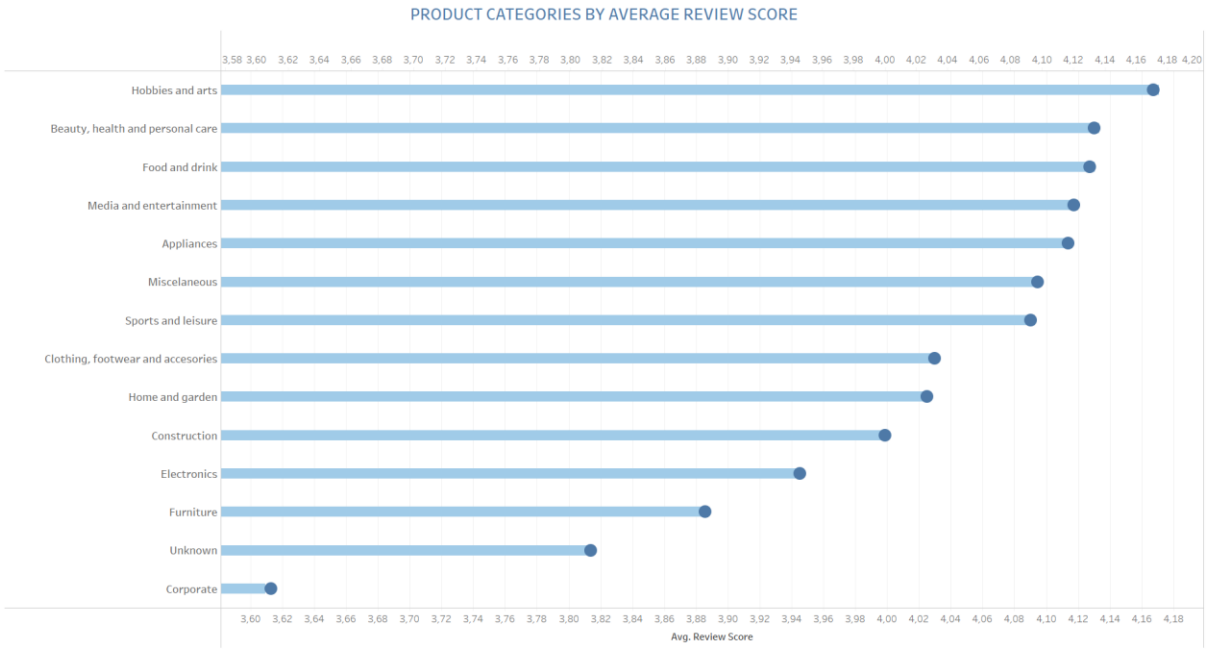
17) Map of delays per city



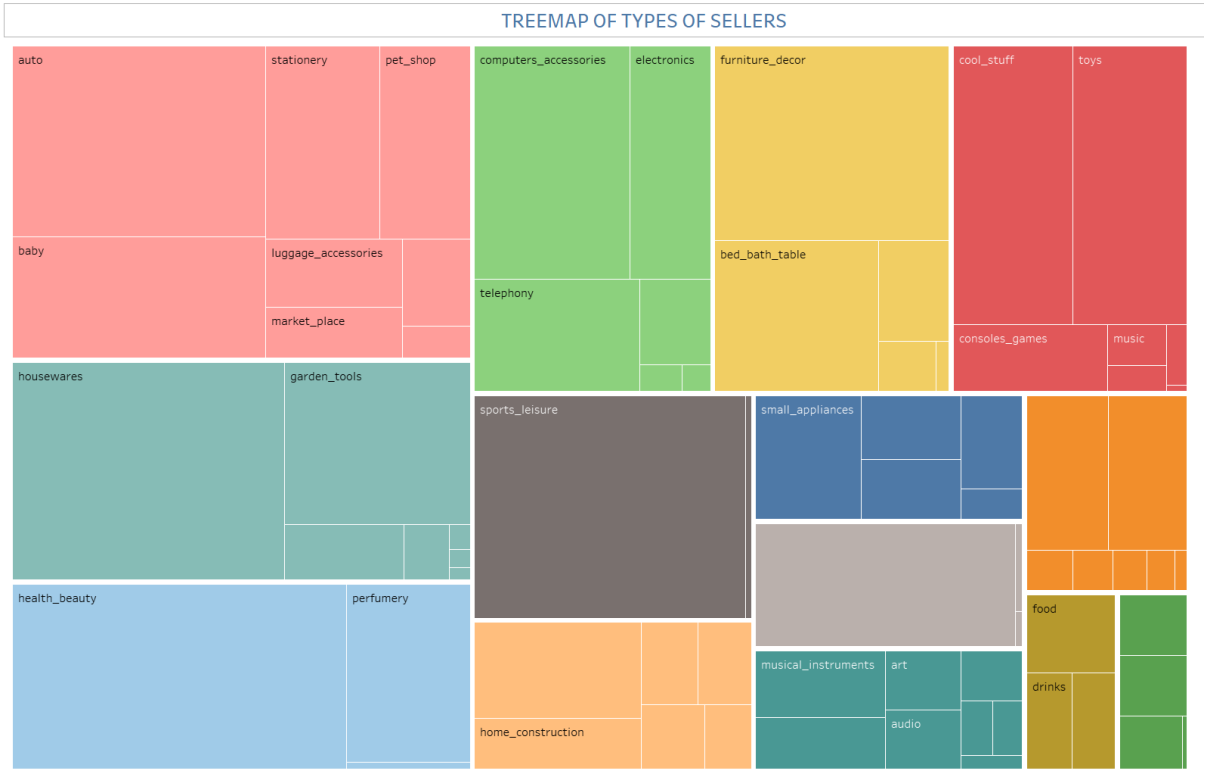
18) Histogram of days until deadline (red is negative)



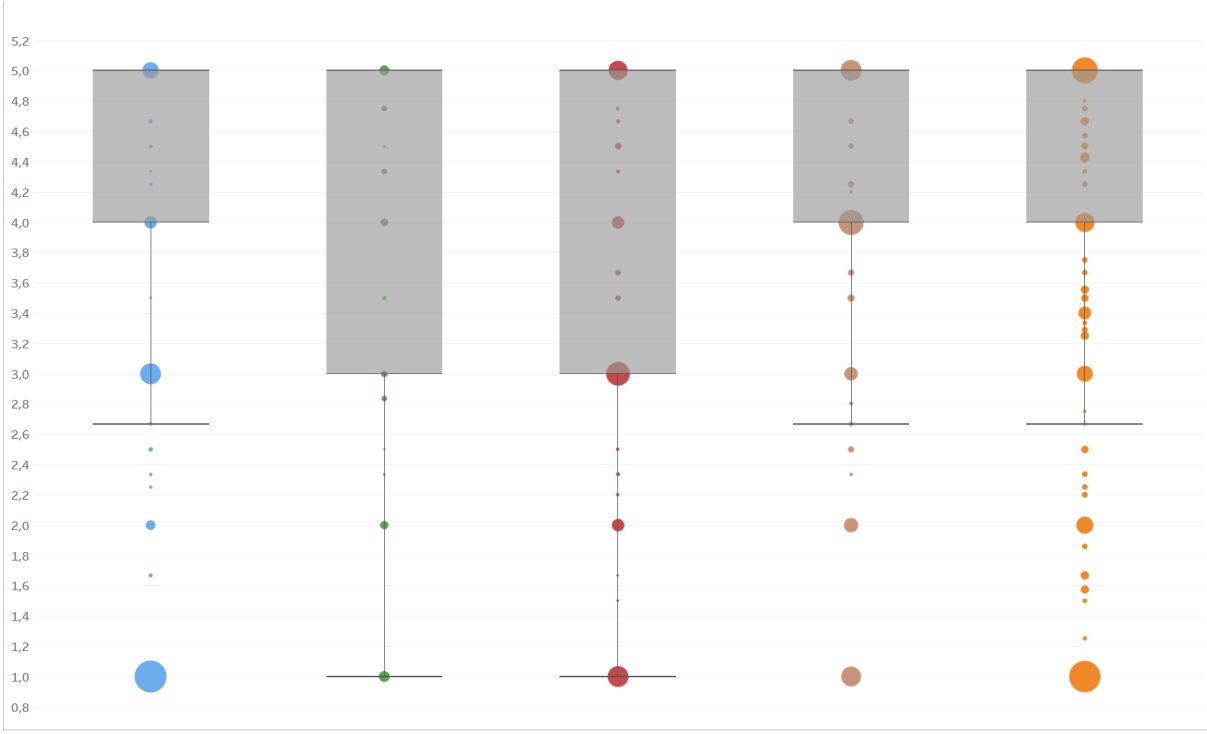
19) Bar chart of product supercategories by average review score



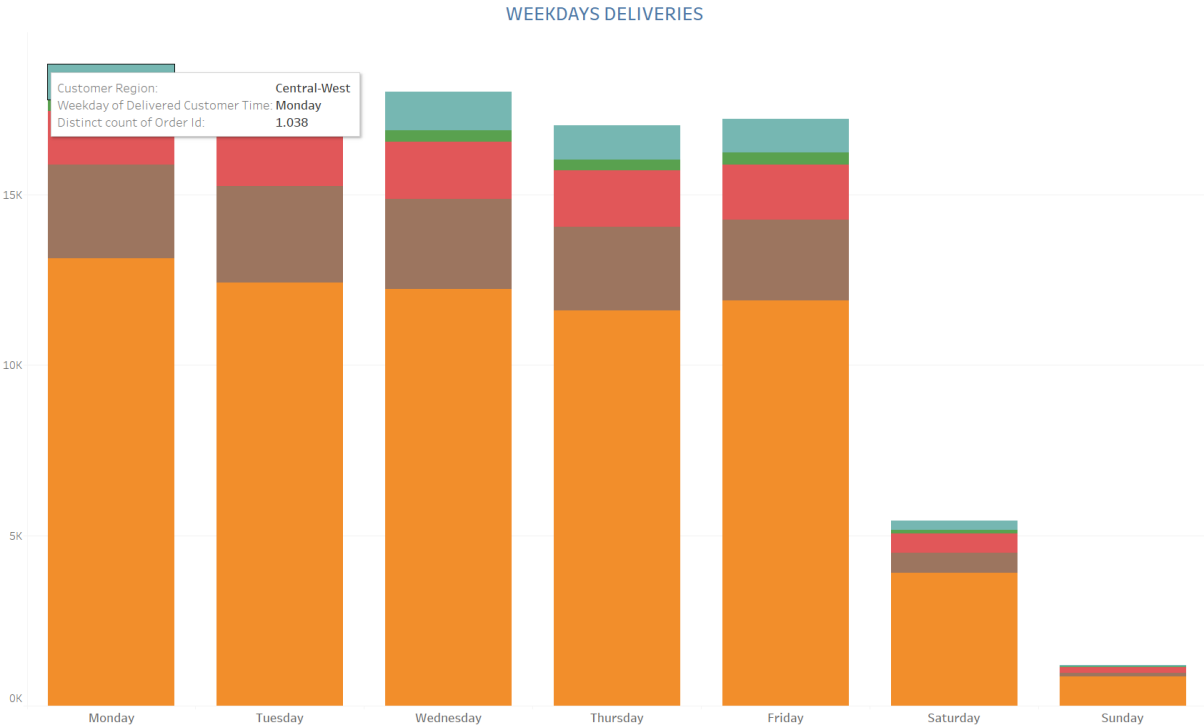
20) Treemap of product categories by sales



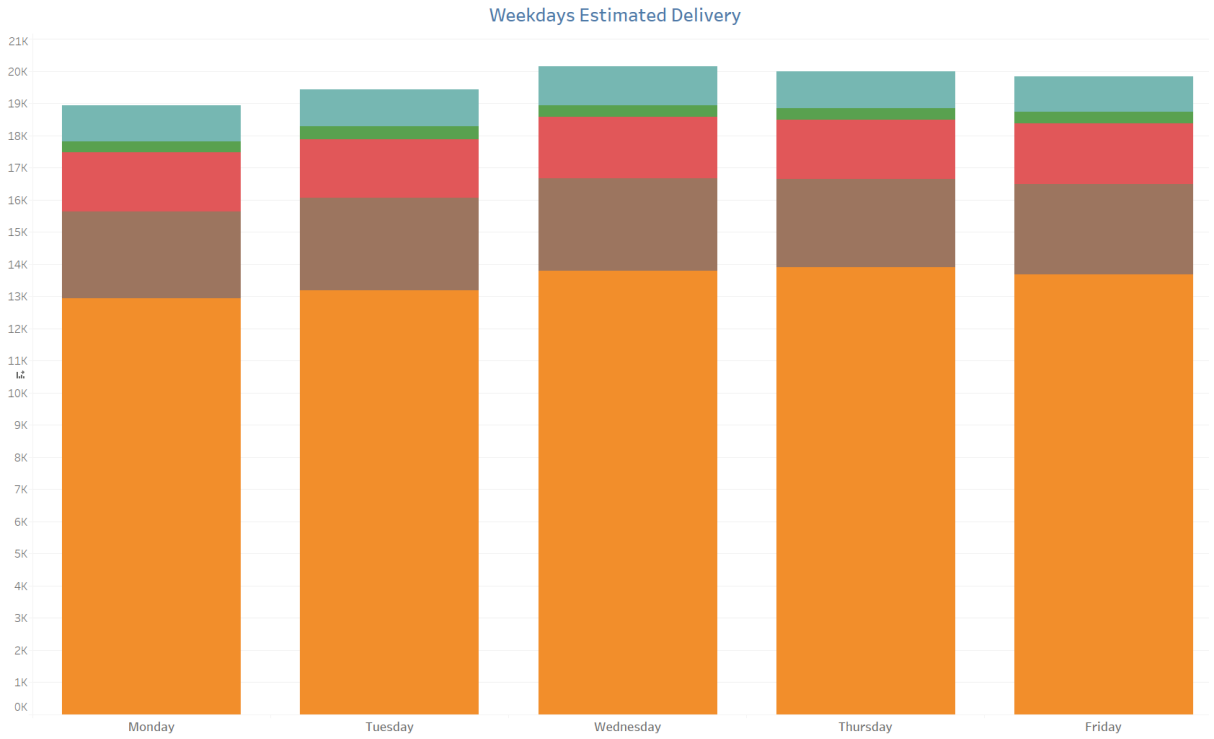
21) Box plot chart of review scores by region



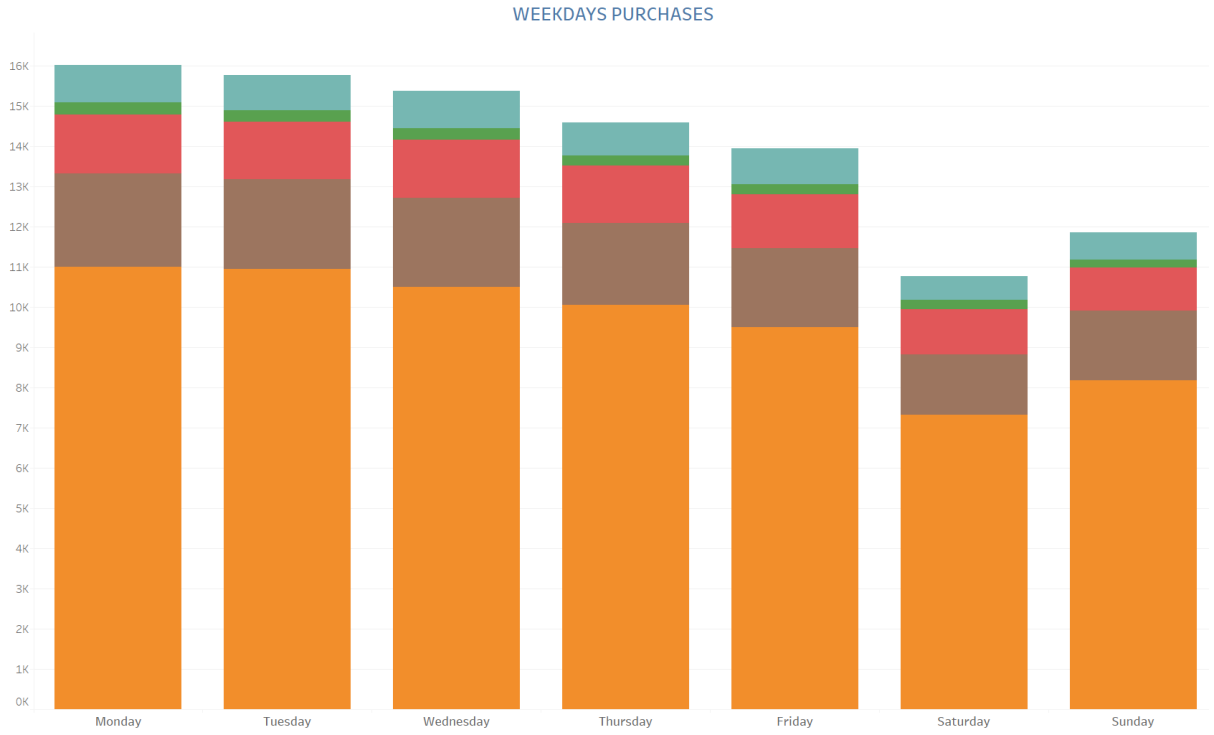
22) Stacked bar chart of deliveries/weekday



23) Stacked bar chart of estimated deliveries/weekday



24) Stacked bar chart of purchases/weekday



25) The dashboard