Case Study 2 notes

Source: [case study example here](https://www.interviewqs.com/blog/case-study-example-1)

The prompt:

You’re a Data Scientist / Business Analyst working for a new eCommerce company called A&B Co. (similar to Amazon) and you’ve been **asked to prepare a presentation** for the Vice President of Sales and the Vice President of Operations **that summarizes sales and operations thus far.** The summary should include (at a minimum) a **summary of current state of the business**, **current customer satisfaction**, and **a proposal of 2-3 areas where the company can improve**. Here are some facts:

Possible questions based on the prompt:

* How many people are using our services?
* How much are we earning from our services?
* Are people enjoying our services?
* Are there any problems that we can identify/resolve?

Possible questions based on the example:

* Does customer satisfaction reflect our current business state? (e.g. if business is down, does customer satisfaction also go down?, if business is boomin’ is customer satisfaction higher?)
* Should I focus on sales area of improvement or customer satisfaction area of improvement?

Given that the datasets were split among 7 separate datasets ([Orders dataset](https://github.com/erood/interviewqs.com_code_snippets/blob/master/Case_1/olist_orders_dataset.csv), [Order items dataset](https://github.com/erood/interviewqs.com_code_snippets/blob/master/Case_1/olist_order_items_dataset.csv), [Order payments dataset](https://github.com/erood/interviewqs.com_code_snippets/blob/master/Case_1/olist_order_payments_dataset.csv), [Product dataset](https://github.com/erood/interviewqs.com_code_snippets/blob/master/Case_1/olist_products_dataset.csv), [Product category name translated dataset](https://github.com/erood/interviewqs.com_code_snippets/blob/master/Case_1/product_category_name_translation.csv), [Order reviews dataset](https://github.com/erood/interviewqs.com_code_snippets/blob/master/Case_1/olist_order_reviews_dataset.csv), [Customers dataset dataset](https://github.com/erood/interviewqs.com_code_snippets/blob/master/Case_1/olist_customers_dataset.csv)),

the first step involved filtering and combining the datasets into a single workable format. This was done by:

1. importing all datasets into PostgreSQL
2. the datasets/tables were then combined together based on primary keys such as **order\_id** and **product\_id**

The following diagram shows how the merge was done in SQL:

Diagram, schematic

Description automatically generated



The final dataset looked something like this:

Graphical user interface, application

Description automatically generated

In short summary, the decisions on the merge were done based on the prompt. Since sales and operations are the primary interests, only the relevant columns such as **payment values** (often equivalent to the **total payment**), **product names**, **reviews**, **customer residence**, **price**, **freight values** were chosen from the batch. **Product ID** and **order ID** were included as well as they are key identifiers of orders and products. The aforementioned columns were retained for these reasons:

* **Order ID/product ID/customer ID**: they are the necessary (unique) identifiers distinguishing each products, orders, and customers
* **Product Category**: this column is identical to product ID, however the actual product name/term is used for **ease of accessibility**
* **Price/freight value/payment type/payment value/total payment**: all these columns are integral to identifying trends of the company’s sales and operations
* **Customer zip code/city/state**: these are important variables that could have in play on how the company’s sales/customer satisfaction may different based on regions
* **Review score**: self-explanatory. Integral for gauging customer satisfaction.

It should be noted that all columns were checked for any missing values. Any duplicates were also left alone since it makes sense in context (duplicate order\_id is possible if customer purchased multiple items under one order)

Based on the EDA, we can observe few exceptional trends immediately

Chart

Description automatically generated

First, we see that the business is generally doing well given the upward trend based on the number of sales/revenues gained per month.

We also see that some states use the services more than others with SP (São Paulo), RJ ([Rio de Janeiro](https://en.wikipedia.org/wiki/Rio_de_Janeiro_(state))) , MG ([Minas Gerais](https://en.wikipedia.org/wiki/Minas_Gerais)), PR ([Paraná](https://en.wikipedia.org/wiki/Paran%C3%A1_(state))), RS ([Rio Grande do Sul](https://en.wikipedia.org/wiki/Rio_Grande_do_Sul)), BA ([Bahia](https://en.wikipedia.org/wiki/Bahia)), SC ([Santa Catarina](https://en.wikipedia.org/wiki/Santa_Catarina_(state))) being the prominent users

* We also see that many of these states record among the highest GDP/density in Brazil. This is especially true for RJ and SP who are 2nd and 3rd highest for GDP (Wikipedia)

Chart

Description automatically generated with medium confidence

Secondly, we see that bed/bath décor, health/beauty items, computer accessories and furniture are among the most popular items in terms of revenue AND general popularity (number of times ordered)

Chart

Description automatically generated

Finally, we can see customer reviews based on states and historic trends (of reviews) which seems to show a generally decent reviews overall apart from October 2017 to March 2018 where we observe a noticeable drop in customer reviews.

* We see that the drops are most likely influenced by the decrease in rating from bigger states such as MG and RJ
* Another observation is that the reviews do not seem to correlate with sales/orders placed
  + Pearson’s correlation between total payment and reviews resulted in little to no correlation (-0.0865)

The Vice president of Sale/Operations wanted to know the **current state of business, and customer satisfaction + areas of improvement.** The first viz answers the first question as it shows the trend of revenue/sales within the past year and shows that the company is generally doing well as seen from the overall positive trend. It does, however, indicate signs of stagnation in recent months as the rate of sale has been decreasing. The second viz also contributes to the gaining further insight on current state of business as it conveys which items/categories are currently popular and garner most revenue for the business. This knowledge can then be used to prioritize sale of these items as they are already popular.

The second question of customer satisfaction can be seen from the third viz which indicates that customers’ reception to the services are positive. It also shows dips in reviews during months from October 2017 to March 2018, which can be further explored to determine the cause of the dip and enforce preventative measures for the future.

Analysis?

<https://medium.com/data-science-on-customer-churn-data/predictive-analysis-using-multiple-linear-regression-b6b3b79b36b6>

<https://github.com/microbhai/CustomerChurnAnalysis/blob/master/MultipleLinearRegression/.ipynb_checkpoints/MultipleRegression-checkpoint.ipynb>