TERM PROJECT 2: Brazilian E-Commerce Public Dataset Analysis

Dataset Description

An ecommerce dataset consisting of 9 tables on purchases on different Brazilian marketplaces in 2016-2018 was chosen for this Term Project 2. More detailed documentation of the project including tables descriptions is provided in the README.md file. The image below represents EER diagram of the RDB:

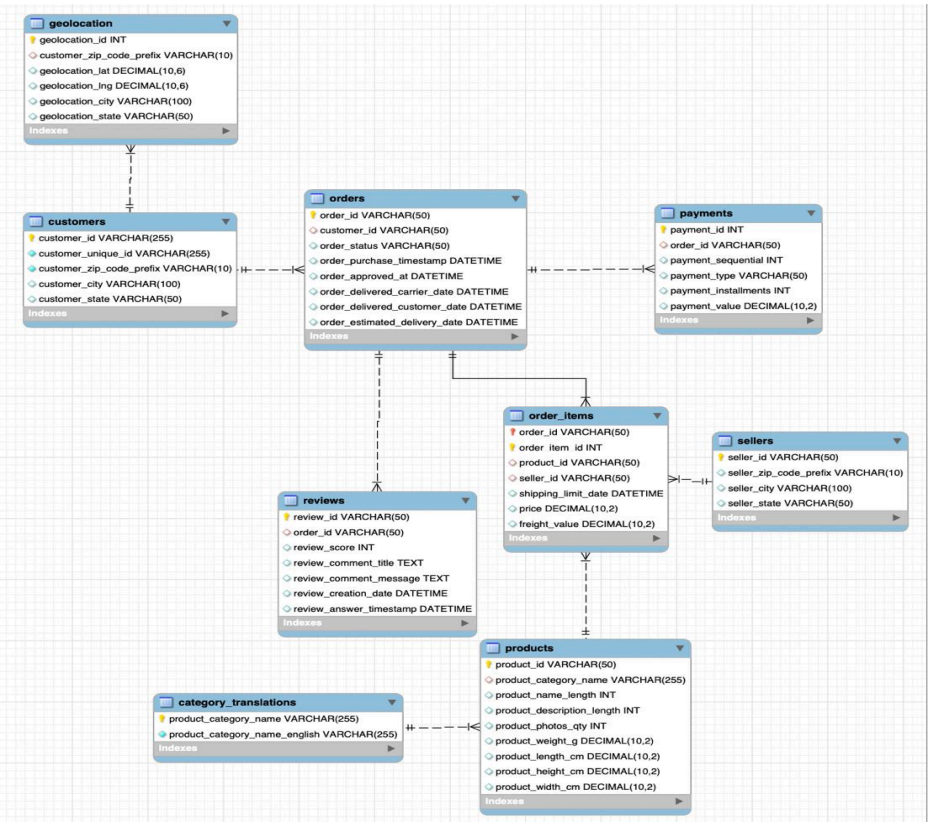

Image 1. EER diagram

Research Questions

For the analytics part we raised 3 research questions as follows:
Q1. How does the delivery performance (on-time vs. delayed) affect customer sentiment in Brazilian e-commerce?
Q2. Is there a clear seasonal trend in order volume? How to promote off-season sales?
Q3. Which categories perform best and worst nationwide?
Q4. Are there noticeable regional differences in the delivery periods?

Solution in Steps & Technical Choices

Firstly, the Python code imports data from CSV files into a MongoDB database, which is connected to Knime. For each research question we have built a separate ETL pipeline (*see Annex, Image A2* or *on Github repo*) in Knime starting with MongoDB Reader Node for joining the data, followed by JSON transformation into tables and data cleaning, filtering – preparation for the further steps of analytics.

For **Q1**, after cleaning, the dataset included approximately 41,000 observations. A representative sample of 3,134 observations was obtained using a 98% confidence level and a 2% margin of error. To assess customer sentiment, we used the Google Cloud NL API, which assigned sentiment scores ranging from -1.0 (negative) to 1.0 (positive) based on

customer reviews. Categorizing the delivery performance as “delayed” if the actual delivery date was after the estimated delivery and as “on-time” otherwise was followed by determining the relationship between delivery performance and sentiment of the comments of the customers.

For **Q2**, duplicates were removed, date columns were formatted, and orders were aggregated by seasons based on months which helped to find seasonal differences.

For **Q3**, after data transformation and cleaning, product categories were analyzed. Through sorting and sampling 10 most and 10 least popular categories were identified.

For **Q4**, delivery times were calculated by comparing order and delivery dates, and joined with geospatial data to examine trends. Geospatial Analytics was used for working with maps and reading geo json files.

Results and Visualizations

Analytics aimed at answering the research questions provided following results:

Q1. As expected, on-time delivered orders tend to have more positive sentiment whereas the negative sentiment in reviews corresponds to delayed orders (*see* *Chart A1*). This shows how important it is for sellers to deliver in a timely manner to satisfy customers.

Q2. As can be seen from *Chart A2*, Winter and Autumn have the highest order volumes, which could be driven by holiday promotions, while Spring has lower sales. Targeted marketing, promotions and different strategies could be helpful to boost sales.

Q3. Products used in households, namely *bed_bath_table* and *health_beauty* categories as well as leisure related categories such as *sports* and *computers* have the largest numbers of orders which can be seen from *Chart A3a*. At the same time, *Chart A3b* shows the least popular categories of products which include *security_and_services*, *fashion_child_clothes* and *CDs_DVDs_musicals*.

Q4. According to the heatmap in *Chart 1*, delivery to the regions further away from the South-East region, namely São Paulo state, is longer, on average. The north region of the country has significantly longer periods of delivery which is coded in red color as a sign of attention.

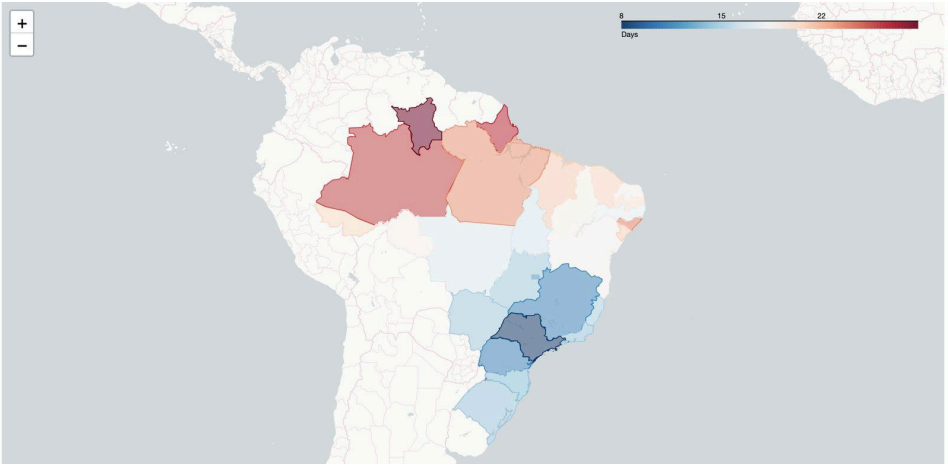

Chart 1. Heatmap of delivery periods

Conclusions

The project results are helpful for understanding key factors influencing customer behavior in the Brazilian e-commerce market. Importance of timely deliveries, understanding seasonal sales, product categories’ performance differences and regional delivery insights can give directions about possible ways of optimizing business strategies of sellers on the e-commerce platforms in Brazil.

Appendix

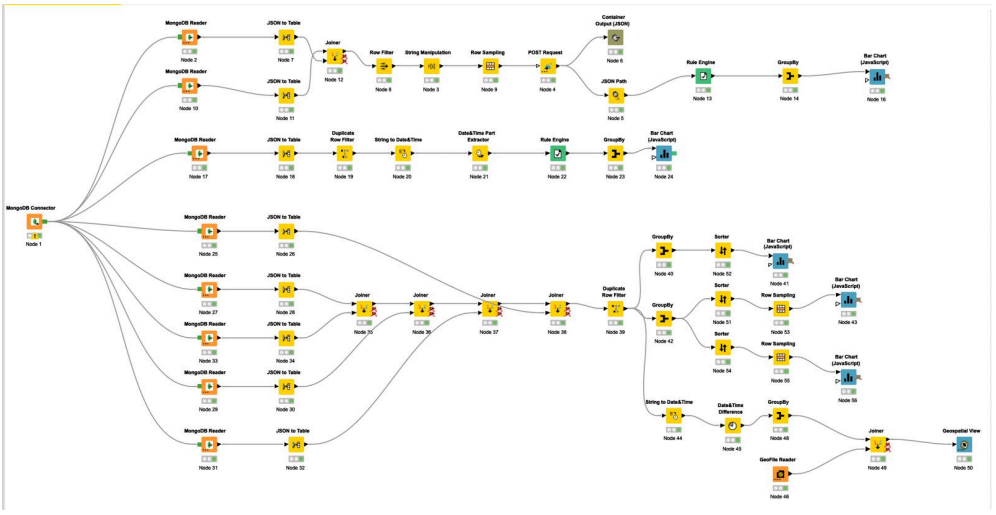

Image A1. ETL pipeline in Knime

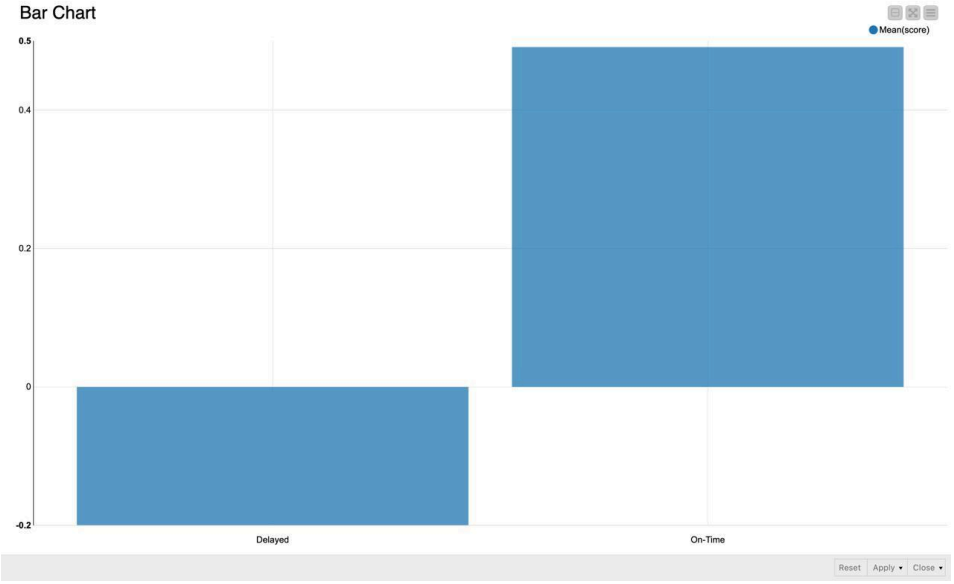

Chart A1. Correlation between sentiment and delivery categorization

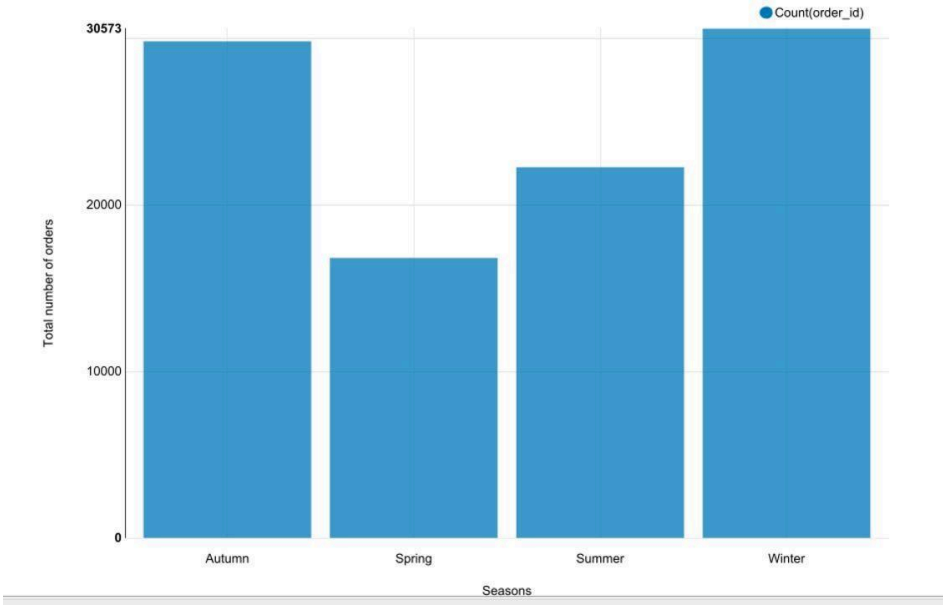

Chart A2. Seasonality character of orders

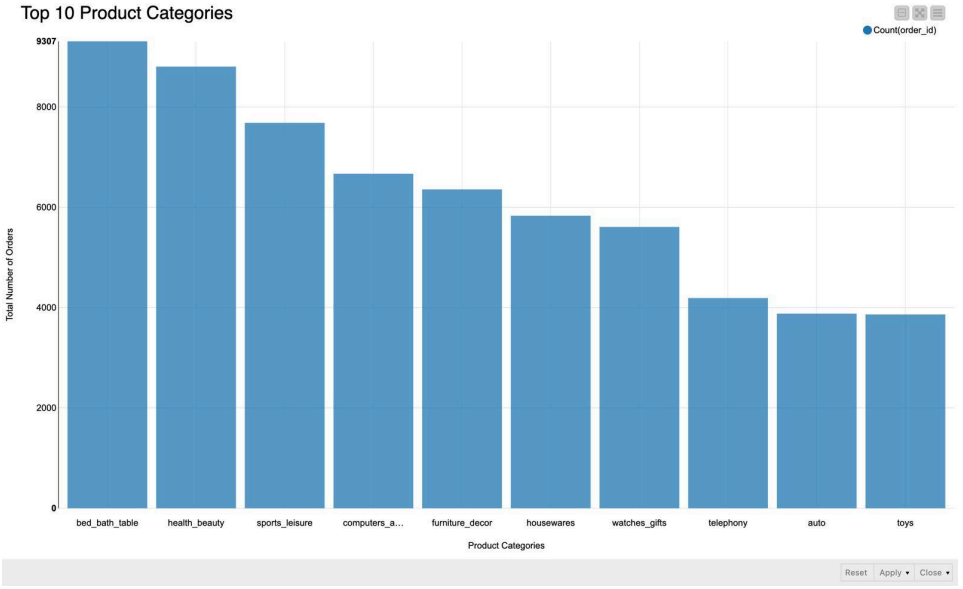

Chart A3a. Top-10 popular product categories

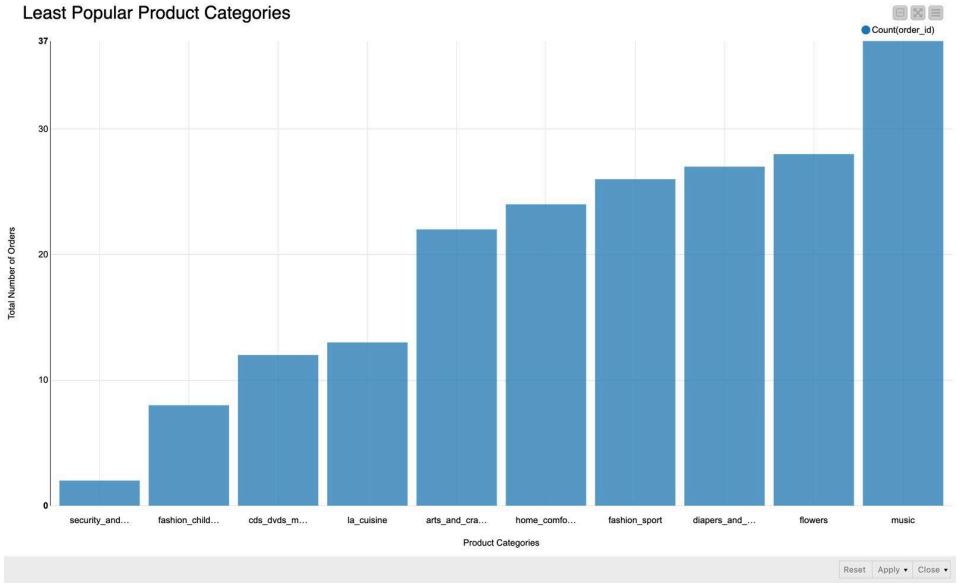

Chart A3b. The least popular 10 product categories