Olist E-Commerce Data Analysis and Estimated Delivery Time Prediction

The **Olist** company, an e-commerce site based in Brazil, shared anonymized data from its actual records between 2016 and 2018 through Kaggle. Let's dive into the details of my project, which uses this data.

Main Technologies Used in the Project:

- SQLite
  - SQL
    - Data Analysis
    - Data Cleaning
    - Feature Engineering
- JupyterLab
  - Python
    - Data Analysis
    - Feature Engineering
    - Data Visualization
    - Machine Learning
- Tableau
  - Data Visualization
  - Data Interpretation

Project Objective:

The dataset used in the project includes data from the **Olist** company, an e-commerce site in Brazil, covering the years 2016 to 2018. The project was approached from 3 perspectives. First, the data was transferred to a database using **SQLite** and analyzed in detail with queries. Errors in the dataset were corrected with **SQL** queries, and new features were added to the data. Secondly, distances between states were calculated using **Python** to show the distance between customers and sellers. Using this prepared data, a final dataset was created for **Machine Learning**. A **Machine Learning** model was then developed to re-predict the estimated delivery time already present in the data.Thirdly, The prediction results, along with other data, were visualized in **Tableau**, creating a dashboard.

Data Details:

The data shared by **Olist** consists of 9 csv files:

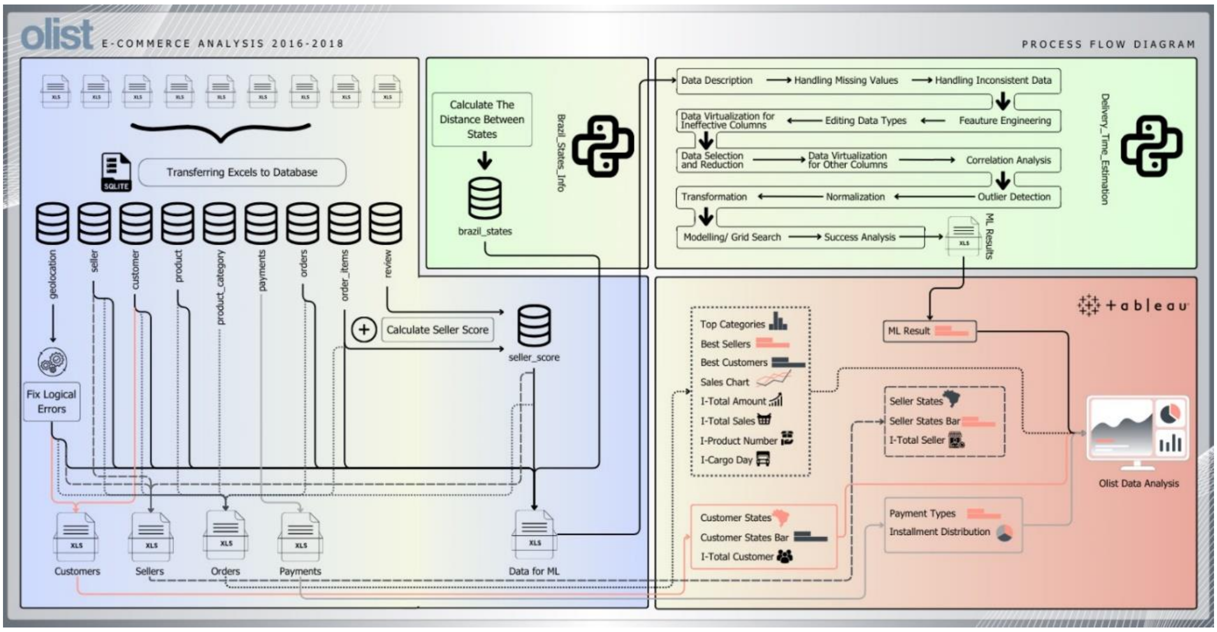
- olist_customer_dataset.csv → Contains information about customers and their addresses.
- olist_sellers_dataset.csv → Contains information about sellers and their addresses.
- olist_geolocation_dataset.csv → Contains geographic locations of addresses.
- olist_order_payments_dataset.csv → Contains payment types, installments, and payment amounts.
- olist_category_name_translation.csv → Contains English translations of product category names.
- olist_products_dataset.csv → Contains information about products and the categories they belong to.
- olist_order_review_dataset.csv → Contains reviews and ratings for orders.

- olist_orders_dataset.csv → Contains general details of orders.
- olist_order_items_dataset.csv → Contains details of items in the orders.

You can find the data, data details and schema at this Kaggle link.

Project Flow Diagram:

Now that we know the data and project objective, let's summarize the project flow in the diagram below. You can access the larger version of the image from the '*Project Process Flow Diagram/Process Flow Diagram.png*' directory.



Functions of the Files in the Project Directory:

To understand the project fully, let's examine the functions of the files in the directory structure:

- SQL:
  - General_Analysis_1: Contains queries for exploring the dataset and creating indexes for faster queries. Not all queries used are included here, but query diversity can be increased according to need.
  - GEO_Update_Wrong_States_2: Logical errors were discovered in the data olist_geolocation_dataset.csv. This includes queries and text normalization that allow for the correction of logical errors.
    - It contains the issue of a single zip code being associated with multiple states. Incorrect states were updated based on the mode of that zip code.
    - It includes adding missing zip codes to the data. Zip codes present in the seller and customer tables but missing from the geolocation table were added to the dataset.
    - It contains the issue of a single zip code being associated with multiple cities. Incorrect cities were updated based on the mode of that zip code. This query is more complex than the state update, as some zip codes were observed to have two mode values.

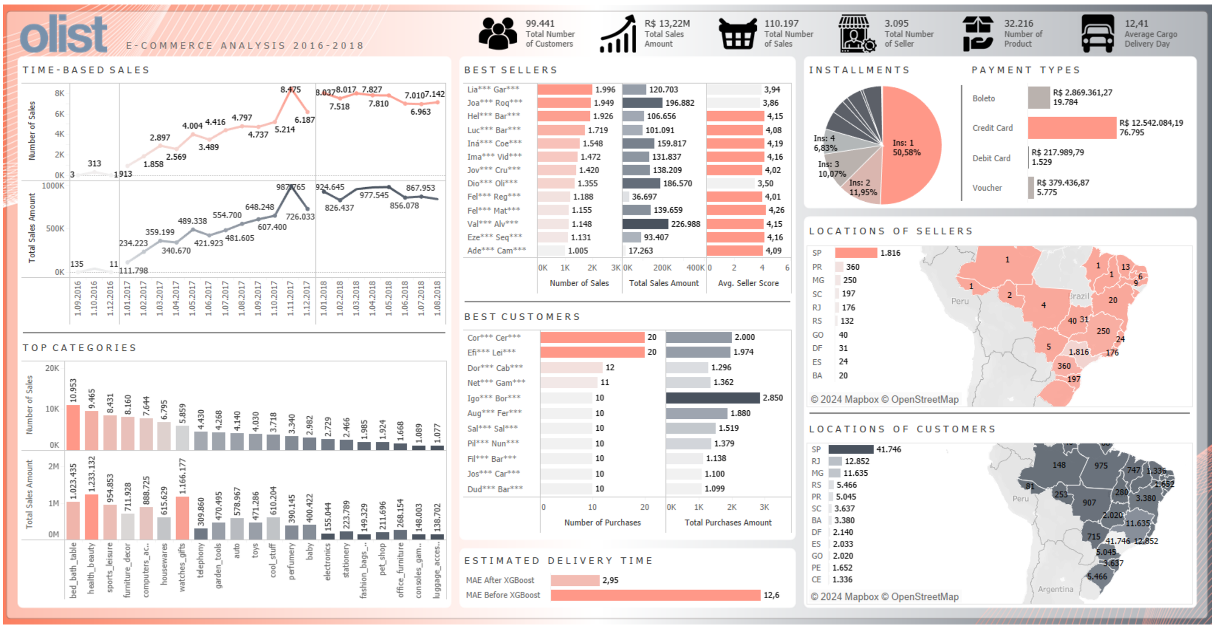
        - Characters outside the Latin alphabet in state names were updated.
        - Punctuation marks were removed from city names containing them.
    - Calculate_Sellers_Score_3: Involves assigning an average score to each seller based on the ratings given to the orders and the sellers of the products in those orders with feature engineering.
    - Create_Main_Data_4: Before, A program was created using the **Python** programming language to calculate the distance between states. Based on this program, the new data is added to the database. After performing join operations with other tables, the main dataset to be used for **Machine Learning** is created. At the same time, indexes are assigned to certain tables to ensure fast query performance.
    - Tableau_Data_5: The data to be used for visualization/analysis in **Tableau** is prepared.
- Python:
    - Brazil_States_Info: It calculates the distance between all states and saves it to an Excel file. This data is used for the main **Machine Learning** dataset prepared with **SQL** queries.
    - Delivery_Time_Estimation: It contains a **Machine Learning** model created to improve the estimated order delivery time already present in the dataset. The following steps were followed during the creation of this model:
        - Data Description: The overall dataset and its statistical values were examined.
        - Handling Missing Values: Missing data in the dataset was handled.
        - Handling Inconsistent Data: Inconsistent data was cleaned and corrected.
        - Feauture Engineering: New features were added to the dataset.
        - Editing Data Types: Data types were adjusted to the appropriate and correct format.
        - Data Virtualization for Ineffective Columns: Data that was decided not to be used in the **Machine Learning** model was visualized before being removed from the dataset.
        - Data Selection and Reduction: Data that would not be used in the **Machine Learning** model was removed from the dataset.
        - Data Virtualization for Other Columns: Data that would be used in training the **Machine Learning** model was visualized.
        - Correlation Analysis for Numerical Values: A correlation graph was created for numerical data.
        - Outlier Detection: Outliers were detected and removed from the dataset.
        - Normalization: Normalization was performed on numerical data.
        - Transformation: Transformation was performed on categorical data.
        - Modelling: **Machine Learning** models were built. Grid search was conducted to obtain the best and most reliable results.
        - Success Analysis: The best model was selected. Success was demonstrated by comparing new predictions with old ones.
- Tableau:
    - Payment Types: Preferred payment types are visualized.
    - Installment Distribution: Preferred installment numbers are visualized.
    - Customer States: The states where customers live are visualized on a map.
    - Customer States Bar: Customer states are visualized with a bar chart.
    - Customer Cities: The cities where customers live are visualized.
    - Seller States: The states where sellers live are visualized on a map.
    - Seller States Bar: Seller states are visualized with a bar chart.
    - Seller Cities: The cities where sellers live are visualized.
    - Seller Score: Seller scores calculated with SQL are visualized.
    - Top Categories: The most purchased categories based on sales are visualized.

- Best Sellers: The success of the best sellers is visualized.
- Best Customers: The orders of the best customers are visualized.
- Sales Chart: Sales over time are visualized.
- I-Total Customer: The total number of customers is displayed.
- I-Total Amount: The total sales amount is displayed.
- I-Total Sales: The total number of sales is displayed.
- I-Total Seller: The total number of sellers is displayed.
- I-Product Number: The total number of unique products is displayed.
- I-Cargo Day: The average number of days it takes for cargo to be delivered to the customer is displayed.
- ML Result: The success of the predicted cargo delivery time after the **Machine Learning** model is visualized.
- Olist Data Analysis: The main dashboard to be presented is created.

Project Results:

In the project, a significant success was achieved in the estimated delivery time parameter as a result of the **Machine Learning** model. The mean absolute error (MAE) of the existing estimated in the dataset is 12.609 days. After the project, the mean absolute error (MAE) of the estimated cargo delivery time is 2.95 days. Therefore, if we consider the implementation of this project in real life, a more accurate delivery time prediction can lead to an increase in customer trust.

Additionally, through the **Tableau** dashboard shown below, the data has been interpreted and visualized. This allows business units to use the dashboard and make various inferences to take the necessary actions. You can access the larger version of the image from the '*Tableau Results*' directory or Tableau Link.



Due to the size of the data, it could not be added. You can access the data from the main Kaggle page. You can generate the other datas by following the process flow diagram.