Olist Analysis

Data source:

This data is open-source data from kaggle.com. It is internal data provided by Olist, a large Brazilian online retailer.

The reason I chose this data was that I originally struggled quite a bit to find something that interested me and met the requirements of the assignment. After about 2 hours of searching I decided to just pick on the suggested datasets from the project brief. I chose this one as I liked the idea of focusing on a geospatial analysis within a country and having many different datasets I had to combine.

Data contents:

The data consists of 9 csv files, varying greatly in content and size. More information in the data profile.

Data profile:

olist_customers_dataset.csv			99441 rows, 5 columns	
Index	**Column**	**Description**	**Time Variance**	**Data Type**
1	customer_id	Key to the orders dataset. Each order has a unique customer_id.	Invariant	Qualitative
2	customer_unique_id	Unique identifier of a customer.	Invariant	Qualitative
3	customer_zip_code_prefix	First five digits of customer zip code	Invariant	Qualitative
4	customer_city	City the customer lives in	Invariant	Qualitative
5	customer_state	State the customer lives in	Invariant	Qualitative

olist_geolocation_dataset.csv			1000163 rows, 5 columns	
Index	**Column**	**Description**	**Time Variance**	**Data Type**
1	geolocation_zip_code_prefix	First five digits of zip code	Invariant	Qualitative
2	geolocation_lat	Latitude	Invariant	Quantitative
3	geolocation_lng	Longitude	Invariant	Quantitative

4	geolocation_city	City	Invariant	Qualitative
5	geolocation_state	State	Invariant	Qualitative

olist_order_items_dataset.csv			112650 rows, 7 columns	
Index	**Column**	**Description**	**Time Variance**	**Data Type**
1	order_id	Order unique identifier	Invariant	Qualitative
2	order_item_id	Sequential number identifying number of items included in the same order	Invariant	Qualitative
3	product_id	Product unique identifier	Invariant	Qualitative
4	seller_id	Seller unique identifier	Invariant	Qualitative
5	shipping_limit_date	Shows the seller shipping limit date for handling the order over to the logistic partner	Invariant	Qualitative
6	price	Item price	Variant	Quantitative
7	freight_value	Item freight value item (if an order has more than one item the freight value is split between items)	Invariant	Quantitative

olist_order_reviews_dataset.csv			99224 rows, 7 columns	
Index	**Column**	**Description**	**Time Variance**	**Data Type**
1	review_id	Unique review identifier	Invariant	Qualitative
2	order_id	Order unique identifier	Invariant	Qualitative
3	review_score	Note ranging from 1 to 5 given by the customer on a satisfaction survey	Invariant	Qualitative
4	review_comment_title	Comment title from the review left by the customer, in Portuguese	Invariant	Qualitative
5	review_comment_message	Comment message from the review left by the customer, in Portuguese	Invariant	Qualitative

6	review_ creation_date	Shows the date in which the satisfaction survey was sent to the customer	Invariant	Quantitative
7	review_answer_timestamp	Shows satisfaction survey answer timestamp	Invariant	Quantitative

olist_orders_dataset.csv			99441 rows, 8 columns	
Index	**Column**	**Description**	**Time Variance**	**Data Type**
1	order_id	Order unique identifier	Invariant	Qualitative
2	customer_id	Key to the customer dataset. Each order has a unique customer_id	Invariant	Qualitative
3	order_status	Reference to the order status (delivered, shipped, etc)	Variant	Qualitative
4	order_purchase_timestamp	Shows the purchase timestamp	Invariant	Quantitative
5	order_approved_at	Shows the payment approval timestamp.	Invariant	Quantitative
6	order_delivered_carrier_date	Shows the order posting timestamp. When it was handed to the logistic partner	Invariant	Quantitative
7	order_delivered_customer_date	Shows the actual order delivery date to the customer	Invariant	Quantitative
8	order_estimated_delivery_date	Shows the estimated delivery date that was informed to the customer at the purchase moment	Invariant	Quantitative

olist_products_dataset.csv			32951 rows, 9 columns	
Index	**Column**	**Description**	**Time Variance**	**Data Type**
1	product_id	Unique product identifier	Invariant	Qualitative

2	product_category_name	Category of product, in Portuguese	Invariant	Qualitative
3	product_name_lenght	Number of characters extracted from the product name	Invariant	Quantitative
4	product_description_lenght	Number of characters extracted from the product description	Invariant	Quantitative
5	product_photos_qty	Number of product published photos	Invariant	Quantitative
6	product_weight_g	Product weight measured in grams	Invariant	Quantitative
7	product_length_cm	Product length measured in centimeters	Invariant	Quantitative
8	product_height_cm	Product height measured in centimeters	Invariant	Quantitative
9	product_width_cm	Product width measured in centimeters	Invariant	Quantitative

olist_sellers_dataset.csv			3095 rows, 4 columns	
Index	**Column**	**Description**	**Time Variance**	**Data Type**
1	seller_id	Seller unique identifier	Invariant	Qualitative
2	seller_zip_code_prefix	First 5 digits of seller zip code	Invariant	Qualitative
3	seller_city	Seller city	Invariant	Qualitative
4	seller_state	Seller state	Invariant	Qualitative

product_category_name_translation.csv			71 rows, 2 columns	
Index	**Column**	**Description**	**Time Variance**	**Data Type**
1	product_category_name	Category name in Portuguese	Invariant	Qualitative
2	Product_category_name_english	Category name in English	Invariant	Qualitative

Data limitations:

The data has very limited information on both customers and sellers. For customers, we only know locations, not age, gender or any socio-economic factors, this will make any kind of targeted demographic analysis essentially impossible. The data also only goes from October 2016 to October 2018.

There should be no ethical considerations in this data; there is no PII data.