A·P·U

ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

GROUP ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

AICT009-4-2-Introduction to Data Analytics

INTRODUCTION TO DATA ANALYTICS

HAND IN DATE: 15 NOVEMBER 2021

NAME: Bryan Hor Jin Hao

TP NUMBER: TP061013

LECTURER: Ms. Hema Latha Krishna Nair

Table of Contents

1. Introduction 1

2. Business Objectives and Scopes 5

Objectives 5

Scopes 5

3. Data Analytics Life Cycle & Methodology 6

Determining Business Objectives 6

Data understanding 6

Data preparation 6

Modelling 7

Evaluation 7

Deployment 7

4. Dataset Understanding 8

olist_orders_dataset 8

olist_customers_dataset 9

olist_order_items_dataset 9

olist_products_dataset 10

product_category_name_translation 10

Dataset schema 11

5. Data Cleaning 12

Loading data into Microsoft Power BI Desktop 12

Data Selection 15

Filtering columns 15

Renaming Columns 17

Merging the datasets 18

Filling in missing values 20

Creating new columns/measures 22

Changing the data format and categorization 22
Grouping data 23
6. Data Visualization and Analysis 24
OLAP 24
Description 24
Drill down and analysis 25
Outlier Analysis 28
7. Data Modelling 30
Regression Analysis 30
Correlation 30
Linear Regression 31
Graphs 35
Product Length 35
Product Weight 35
8. Conclusion & Recommendations 36
Data privacy and security recommendations 36
9. Personal reflection report & Workload matrix 37
Personal Reflection Report 37
Workload matrix 37
10. References 38

1. Introduction

As mentioned on the cover page, the project will be focused on e-commerce activities in Brazil. Before diving into the e-commerce situation in Brazil, we need to understand what e-commerce is first. Although there are many different definitions of e-commerce online, e-commerce is essentially a paperless exchange of business information using electronic data interchange, electronic mail, electronic bulletin boards, electronic funds transfer, the Internet, and other network-based technologies. There are also many different categories of e-commerce. Such categories include business-to-business (B2B), business-to-consumer (B2C), consumer-to-consumer (C2C), consumer-to-business (C2B) and many more (Ingle, Bhalekar, & Pathak, 2014). Each of these categories has different methods of implementation, pricing strategies and requirements. As e-commerce acts as a new medium for exchanging value, there are several benefits over the traditional method. One of them is the convenience e-commerce provides. E-commerce sites can operate 24 hours a day and 7 days a week and this allows customers to browse through the store's catalogue whenever and wherever they are. E-commerce also allows customers to shop without leaving the comfort of their homes. Another benefit of e-commerce is more geared towards the sellers. The sellers now can sell their products to a larger market as e-commerce allows them to reach international customers. This can increase their revenue as there are more customers to serve (Ingle, Bhalekar, & Pathak, 2014). However, e-commerce also has its drawbacks. Customers buying products from e-commerce sites will need to wait for their products to be delivered to them. This can take from a few days to a few weeks. This could decrease the customer's satisfaction as they must wait as compared to when shopping in physical stores where they get their products immediately. Customers may also need to pay the delivery fees when they purchase from e-commerce sites and this could increase the product's price several times, especially so when shopping from other countries (Ingle, Bhalekar, & Pathak, 2014). Although e-commerce has its advantages and disadvantages, when implemented right, it could offset a lot of those negatives.

Brazil has the ninth-largest GDP (Gross Domestic Product) in the world and has a population of 204 million and half of this number is in the middle class of the economy. However, most of them are still underserved in retail sectors and this opens up the window for retailers to jump in and seize this opportunity through the growing e-commerce sector in this country. As the middle-class incomes grow, so does the internet penetration in the country, and this will, in turn, endorse the growth of e-commerce. With growing orders and revenue throughout the years and an expected Compound Annual Growth Rate (CAGR) of 10%, the e-commerce sector in this country cannot be ignored (Lima, 2017).

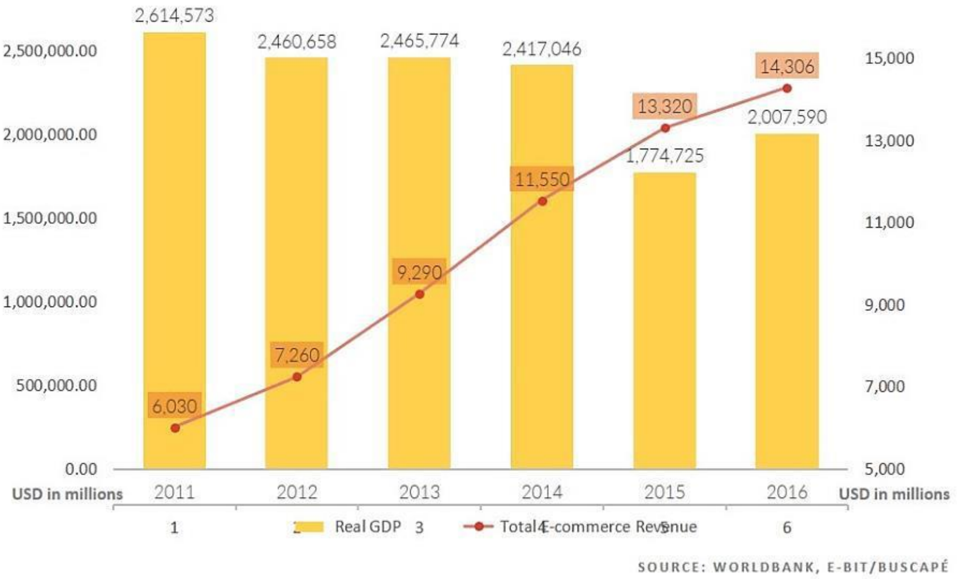
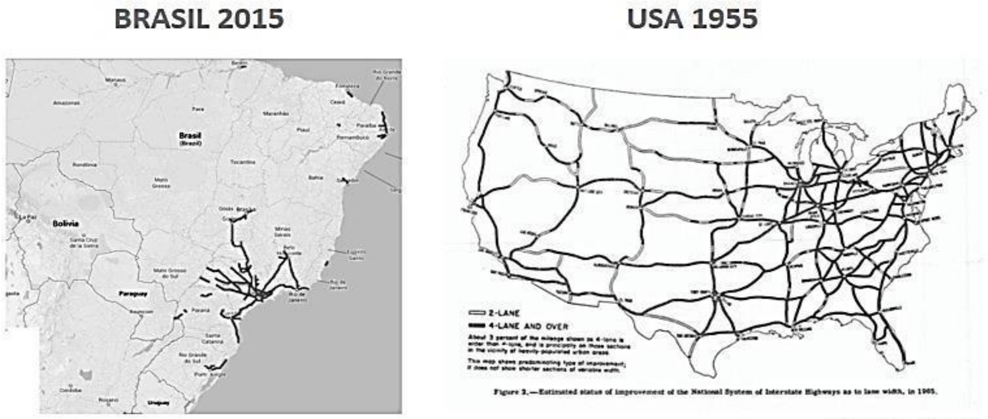


Figure 1 E-commerce growth in Brazil from 2011-2016 (Lima, 2017)

As shown in Figure 1, the e-commerce sector in Brazil has been growing steadily since 2011. However, even though this sector has been growing for the last decade, there are still some difficulties in this sector. One such difficulty is that there are high taxes, and the tax structure is very complicated. Businesses that wish to sell their products will need to understand this structure thoroughly to avoid any legal procedures in the future. Furthermore, the country's underdeveloped distribution infrastructure also proves to be a weak point for the e-commerce sector in the country. The picture below illustrates the comparison between the highway network infrastructure of Brazil in 2015 and the United States of America (Lima, 2017).



BRASIL 2015

USA 1955

SOURCE: DNIT

Figure 2 Highway network of Brazil and the USA

This underdeveloped distribution network will impact the shipping of products by e-commerce as the unpaved and poor conditions of the road in Brazil pose a threat to the safety of the delivery trucks and drivers as well as lengthening the delivery times unnecessarily. Brazil is also infamous for its traffic congestion, with three of its cities in the top ten of most congested cities and Rio de Janeiro coming in fourth in the Traffic Index (Lima, 2017). These factors will affect the delivery schedules immensely. Although the e-commerce sector in Brazil was and is expected to flourish, the problems will still need to be addressed to further boost the growth of e-commerce and attract more foreign investment into this sector.

The analytics that will be done in this project will be used in hopes to solve some if not all of the problems listed in the previous paragraph. The objectives and scopes of the project will be listed in one of the following sections. This project will analyse several aspects of the e-commerce activities in Brazil, such as the geolocations of the customers that bought items from an e-commerce site in Brazil. By analysing this information, courier companies or sellers can plan their routes better to ensure that their customers can receive their items as soon as possible, providing the customers with the best shopping experience and retaining them. Besides that, this project will also focus on the type of products sold on the e-commerce platforms and the trending products at a certain time. With this information, recommendations can be given to the sellers to focus on certain product categories to boost their sales or to governments to provide incentives to certain product categories to grow their sales. This project will also use data analytics to produce a trend of the history of price movements in the e-commerce sector. This can help individuals to plan their expenses, like only purchasing during holiday seasons or when the site is going to provide discounts. This can also help the organizations or sellers to plan their investments into profiting product categories instead of focusing on loss-making products. Furthermore, this project will also try to analyse the suppliers' information. This can help indicate how long it will take for the product to travel from the supplier's warehouse to the seller's warehouse. This can in turn estimate how long it will take for an out-of-stock item to be restocked and solve supply and demand problems.

2. Business Objectives and Scopes

Objectives

To identify the total sales generated by the product categories, and predict the shipping fee of the product based on the product's dimensions

Scopes

- To create an analytical dashboard that contains information about the sales of the products listed on the Olist platform, like the number of sales generated in a certain year, a specific category, and a specific state.

- To create a predictive model to predict the shipping fee based on the product dimensions like the product's height and product's weight. The predictive model that will be used is a Linear Regression model.

3. Data Analytics Life Cycle & Methodology

CRISP-DM stands for cross-industry process for data mining. This methodology includes several phases to analyze data. The steps are as below:

Determining Business Objectives

The first step in this phase is to understand what is needed to accomplish the business perspective. The data analyst would need to decide on the desired outputs of the project by setting up the objectives and then producing the project plan to achieve the data mining and business goals. The analyst should also determine the criteria used to determine the success of the project.

After setting the objectives, the analyst will then need to assess the current situation like the resources available, the requirements, assumptions, constraints, risks, and contingencies of the project, and prepare a project plan accordingly (Smart Vision Europe, 2020).

Data understanding

In this step, the analyst needs to start collecting data that will be used for the data mining project. The analyst will need to prepare a data description report that includes the data's format, quantity, identities of each field, and any other required information on the data that are being collected. Then, the data needs to be explored, identifying the distribution of key attributes, relationships between data, and many more. Aside from the tasks mentioned above, the analyst should also prepare a data quality report to ensure that the data is complete, correct, and suitable for the project (Smart Vision Europe, 2020).

By understanding the datasets that are being used, the duration used to sift through the data, finding useful columns can be greatly reduced.

Data preparation

Here in this step, the analyst will need to select the specific data that is going to be used, clean the data to raise the data quality, construct any required data like derived attributes and generated records, if any, and finally integrate the data. All these steps are used to ensure that the results generated from the data mining project are useful for the organization's goals (Smart Vision Europe, 2020).

Modelling

The first step in this phase is to select the modelling technique. When selecting the modelling technique, the analysts need to make specific assumptions on the data like assumptions on missing values, class attributes, and so on before deciding on the technique. Once the technique is selected, the test design needs to be generated for training, testing and evaluating the models. Then once all has been decided, the model will need to be built by setting the parameters and descriptions. Finally, the model also needs to be assessed (Smart Vision Europe, 2020).

Evaluation

After building the model and running the data mining project using the model, the data mining results need to be assessed and reviewed. Depending on the results of the assessment, the next steps will be decided (Smart Vision Europe, 2020). Once the model is prepared and trained with values from the dataset, the performance of the model will be evaluated through several means to determine its suitability.

Deployment

In this phase, using the evaluation results, the analyst will plan on how to monitor and maintain the project and then produce the final report of the data mining project. The analyst will also need to review the project identifying what went right, what went wrong, and what needs to be improved for the data mining project (Smart Vision Europe, 2020).

4. Dataset Understanding

The datasets that were used to analyze the e-commerce trends in Brazil are all sourced from the Kaggle website. The datasets were prepared by the Olist store located in Brazil that is also an e-commerce website. The datasets that were used are listed as below:

1. olist_orders_dataset
2. olist_customers_dataset
3. olist_order_items_dataset
4. olist_products_dataset
5. product_category_name_translation

Source: Brazilian E-Commerce Public Dataset by Olist | Kaggle

All the above files are stored as CSV files before being imported into the Microsoft Power BI platform. Each of these datasets contains several columns but not all columns will be used in this project. The selection of columns will be done in the next step where data cleaning is carried out. The columns in each dataset are as follows:

olist_orders_dataset

Column name	Details	Values
order_id	The unique identifier of the order	100% valid values
customer_id	The unique identifier of the customer using the platform. A primary key to the customer dataset.	100% valid values
order_status	Reference to the order status (delivered, shipped, etc)	100% valid values
order_purchase_timestamp	Shows the purchase timestamp	100% valid values
order_approved_at	Shows the payment approval timestamp	99.99% valid values <1% missing values
order_delivered _carrier_date	Shows the order posting timestamp when it was handed to the logistic partner`	98% valid values 2% missing values
order_delivered _customer_date	Shows the actual order delivery date to the customer	97% valid values 3% missing values
order_estimated_delivery_date	Shows the estimated delivery date that was informed to the customer at the purchase timestamp	100% valid values

As shown in the above table, the olist_orders_dataset contains several columns but not all columns will be used. Several columns also contain missing values that need to be filled in the Data Cleaning step as well.

olist_customers_dataset

Column name	Details	Values
customer_id	Key to the orders dataset.	100% valid values
customer_unique_id	Uniquely identifies each customer	100% valid values
customer_zip_code_prefix	First 5 digits of customer's zip code	100% valid values
customer_city	Name of the city that the customer is staying in	100% valid values
customer_state	Name of the state that the customer is staying in	100% valid values

The above table shows the details of the olist_customers_dataset. This dataset shows the customers' information, primarily their location in Brazil. With this information, analysis on the duration needed to deliver the items purchased to each customer's location can be shown through map visualizations.

olist_order_items_dataset

Column name	Details	Values
order_id	Uniquely identifies each order	100% valid values
order_item_id	Sequential number identifying the number of items included in the same order	100% valid values
product_id	Uniquely identifies each product	100% valid values
seller_id	Uniquely identifies each seller	100% valid values
shipping_limit_date	Shows the seller shipping limit date for handing the order over to the logistic partner	100% valid values
price	The item price	100% valid values
freight_value	The item freight value, if the order has more than one item with freight values, it is split between the items)	100% valid values

The columns for the above dataset list the number of items that are included in each order. This can be used to aggregate the total amount of items sold based on the product category and the average price of each order. There are no missing values but the price and freight value may need to be rounded to 2 decimal values.

olist_products_dataset

Column name	Details	Values
product_id	Uniquely identifies the product	100% valid values
product_category_name	Root category of the product, in Portuguese	98% valid values 2% missing values
product_name_lenght	Number of characters extracted from the product name	98% valid values 2% missing values
product_description_lenght	Number of characters extracted from the product description	98% valid values 2% missing values
product_photos_qty	The number of product published photos	98% valid values 2% missing values
product_weight_g	Product weight measured in grams	99.99% valid values <1% missing values
product_length_cm	Product length measure in centimeters	99.99% valid values <1% missing values
product_height_cm	Product height measure in centimeters	99.99% valid values <1% missing values
product_width_cm	Product width measure in centimeters	99.99% valid values <1% missing values

As shown in the table above, this dataset contains mostly about information of the product, from the product category to the product's dimensions. However, the column names are spelled incorrectly and there are some missing values albeit small in numbers. Moreover, the units used to measure the products can be normalized to the SI units used internationally for easier comparison.

product_category_name_translation

Column name	Details	Values
product_category_name	Category name in Portuguese	100% valid values
product_category_name_english	Category name in English	100% valid values

This dataset contains the English names for the product categories listed in the products dataset. Therefore, this table can be merged with the products dataset shown previously so that the product categories do need to be translated manually and this will help people who do not understand Portuguese to understand the dataset easier.

Dataset schema

After understanding the columns of each dataset, the star schema for this project can be drawn. The star schema that is shown below contains information about the columns that will be used and after merging them which column will the tables be merged on.

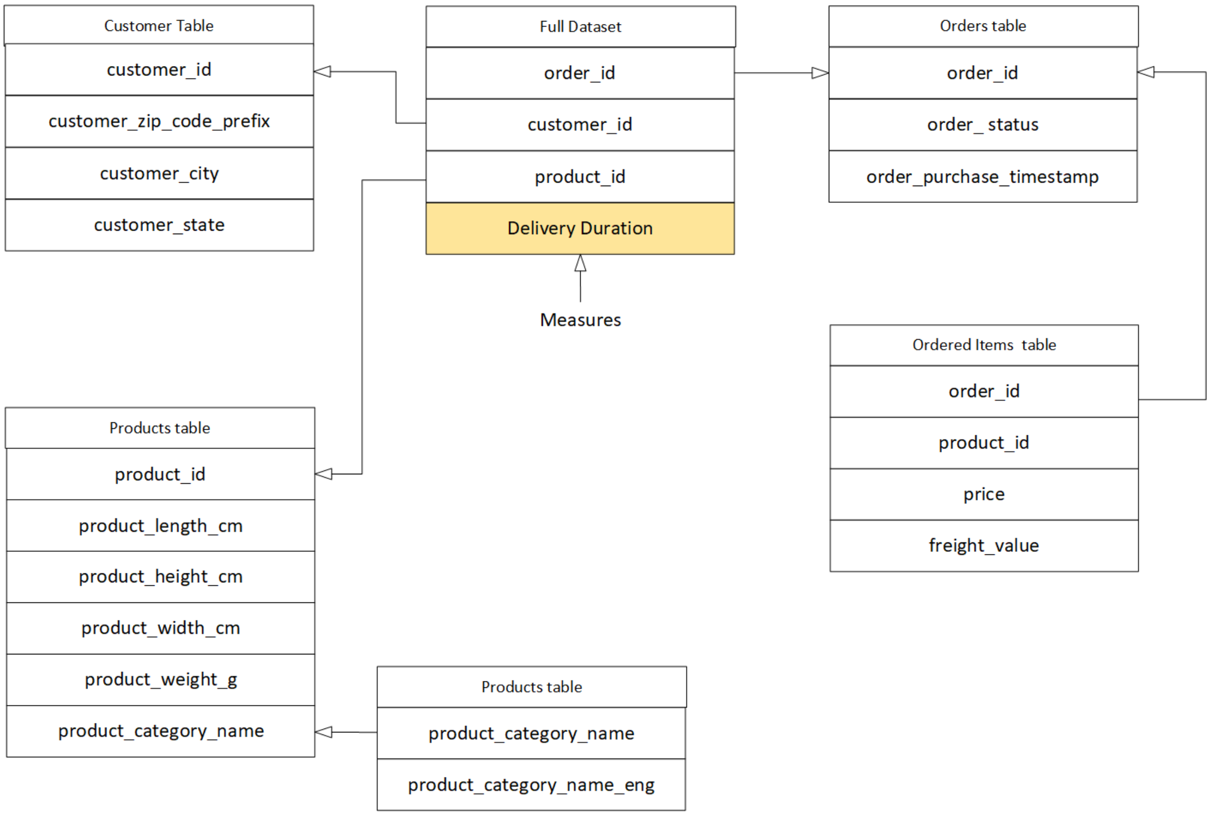
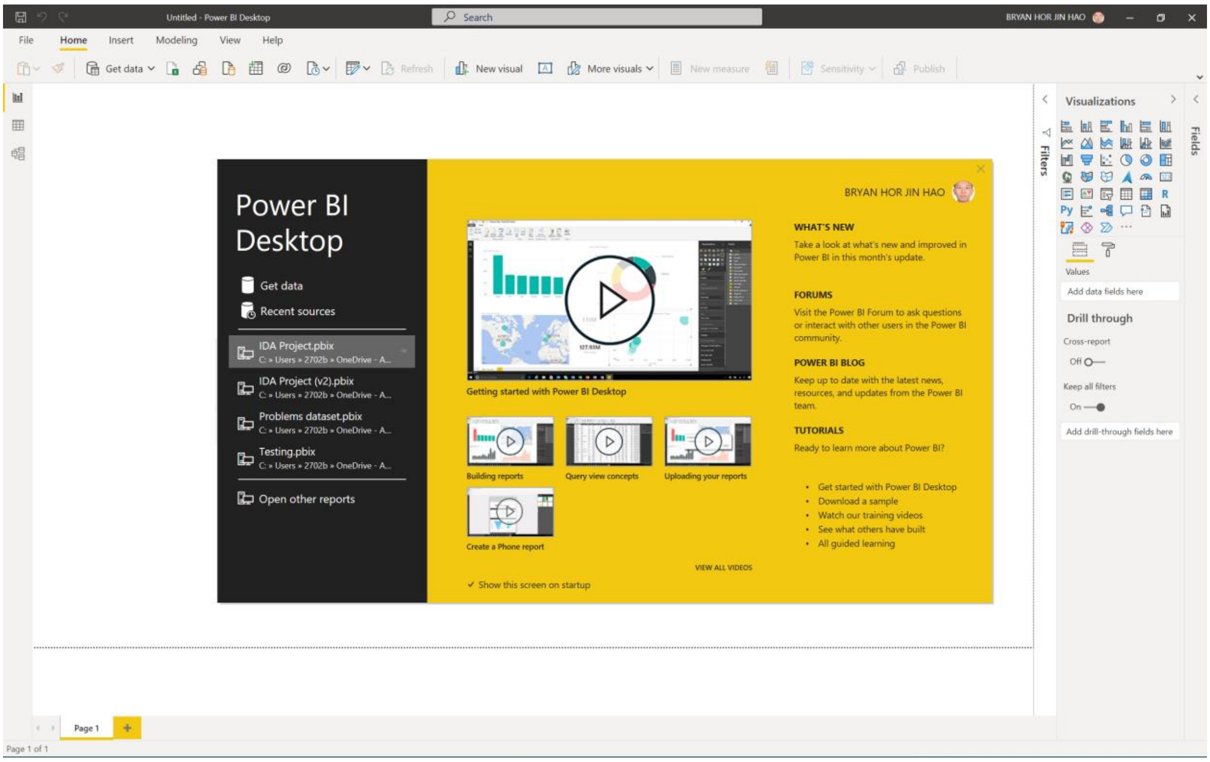


Figure 3 Star Schema

5. Data Cleaning

Loading data into Microsoft Power BI Desktop

To build an analytical dashboard for the datasets that are used, the datasets need to first be loaded into the platform used, which as stated in the title, Microsoft Power BI Desktop. Once Microsoft Power BI Desktop, which will be referred to as Power BI further in the report, there will be a dialogue box shown to the user. In this dialogue box, the user can choose to import new data or use an existing .pbix file that was created previously, as shown in the figure below.



WHAT'S NEW

Take a look at what's new and improved in Power BI in this month's update.

FORUMS

Visit the Power BI Forum to ask questions or interact with other users in the Power BI community.

POWER BI BLOG

Keep up to date with the latest news, resources, and updates from the Power BI team.

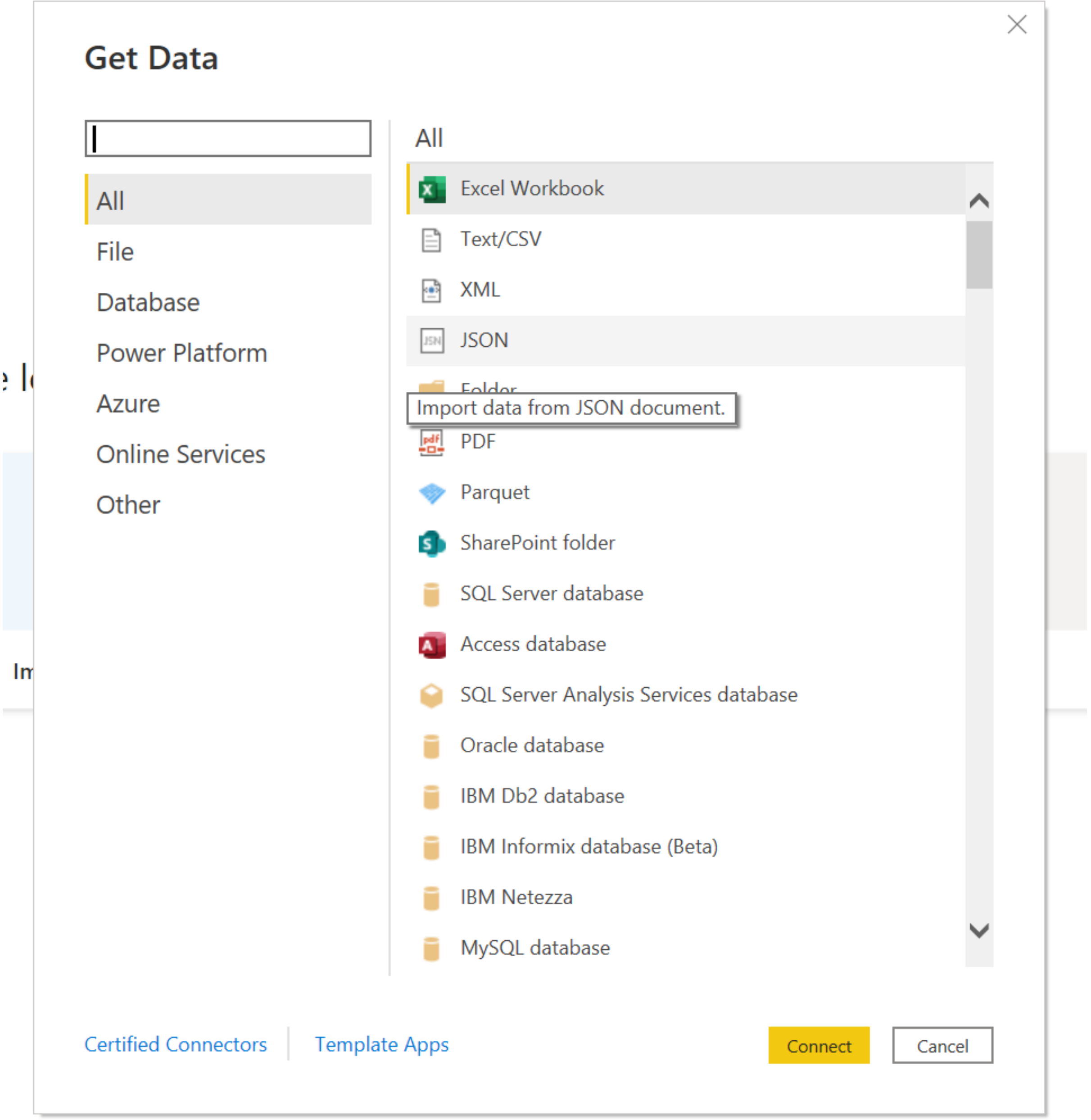
TUTORIALS

Ready to learn more about Power BI?

- Get started with Power BI Desktop
- Download a sample
- Watch our training videos
- See what others have built
- All guided learning

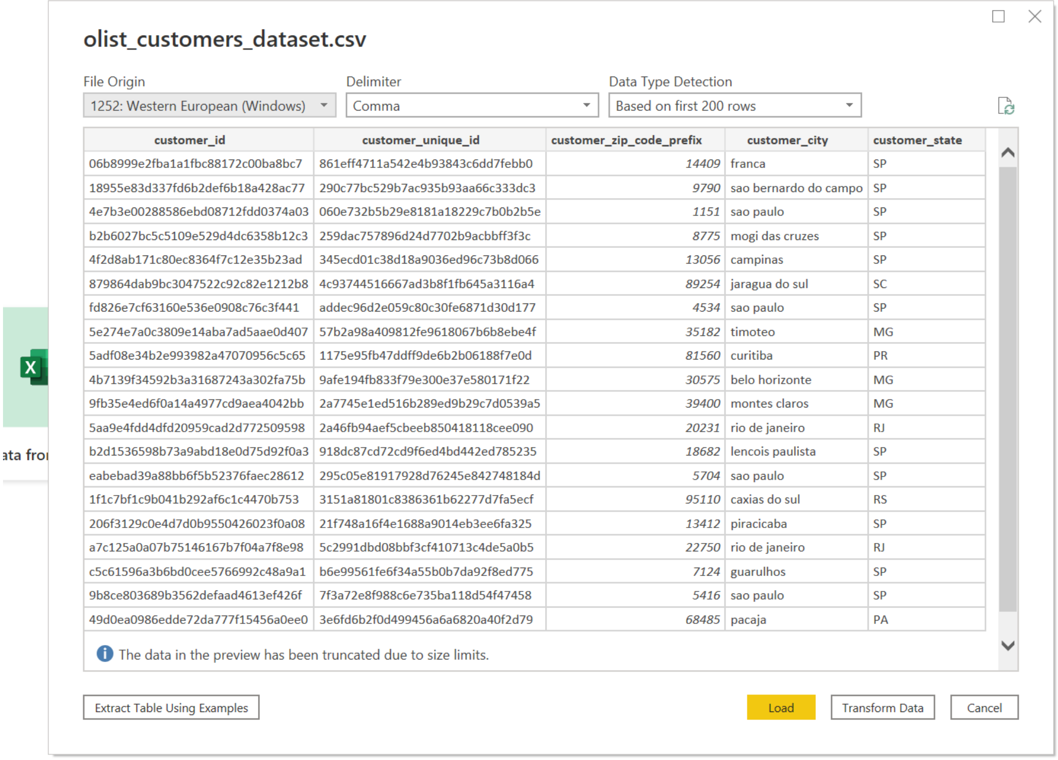
Figure 4 Power BI first dialogue box

Here, the “Get Data” option will be chosen and the datasets that were stored as .csv files will be imported in. However, due to the presence of missing values in the olist_orders_dataset, this specific dataset will be cleaned using python before being loaded for actual use in the platform. When importing the data, the user can choose several different types of files that Power BI can support like Excel files, Access files, CSV, etc. The figures below show the process of importing the data.



Import data from JSON document.

Figure 5 Dialogue box for importing data



olist_customers_dataset.csv

File Origin: 1252: Western European (Windows)

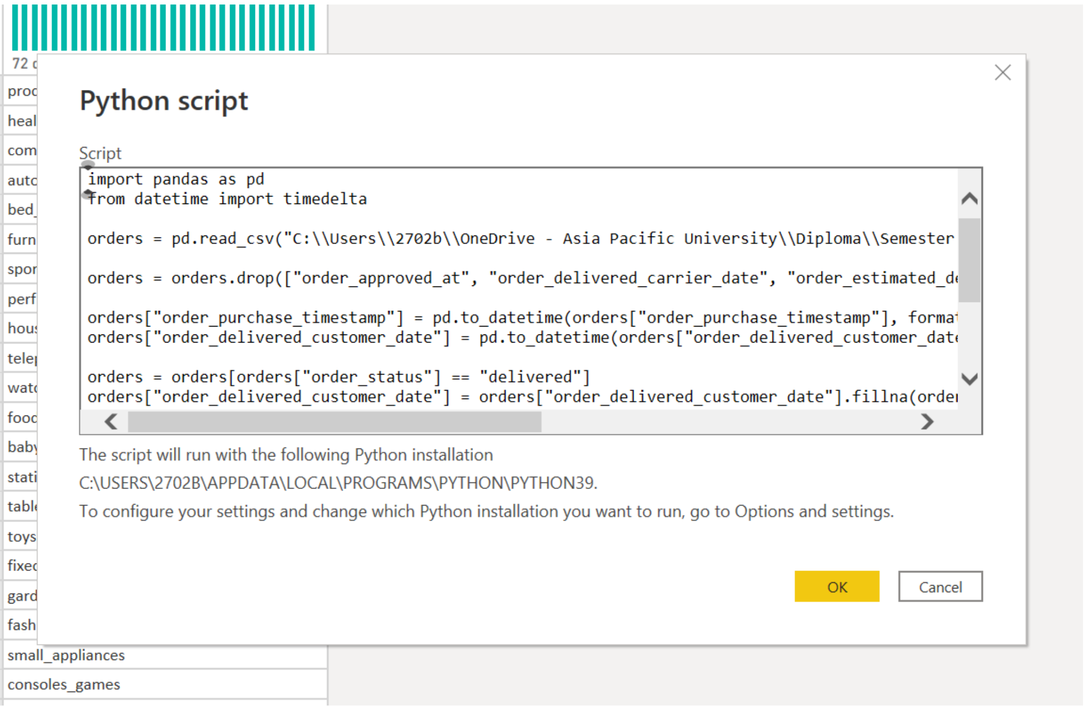
Delimiter: Comma

Data Type Detection: Based on first 200 rows

customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP
18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3	9790	sao bernardo do campo	SP
4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2b5e	1151	sao paulo	SP
b2b6027bc5c5109e529d4dc6358b12c3	259dac757896d24d7702b9acbbff3f3c	8775	mogi das cruzes	SP
4f2d8ab171c80ec8364f7c12e35b23ad	345ecd01c38d18a9036ed96c73b8d066	13056	campinas	SP
879864dab9bc3047522c92c82e1212b8	4c93744516667ad3b8f1fb645a3116a4	89254	jaragua do sul	SC
fd826e7cf63160e536e0908c76c3f441	addec96d2e059c80c30fe6871d30d177	4534	sao paulo	SP
5e274e7a0c3809e14aba7ad5aae0d407	57b2a98a409812fe9618067b6b8ebe4f	35182	timoteo	MG
5adf08e34b2e993982a47070956c5c65	1175e95fb47ddff9de6b2b06188f7e0d	81560	curitiba	PR
4b7139f34592b3a31687243a302fa75b	9afe194fb833f79e300e37e580171f22	30575	belo horizonte	MG
9fb35e4ed6f0a14a4977cd9aea4042bb	2a7745e1ed516b289ed9b29c7d0539a5	39400	montes claros	MG
5aa9e4fdd4dfd20959cad2d772509598	2a46fb94aef5cbeeb850418118cee090	20231	rio de janeiro	RJ
b2d1536598b73a9abd18e0d75d92f0a3	918dc87cd72cd9f6ed4bd442ed785235	18682	lencois paulista	SP
eabebad39a88bb6f5b52376faec28612	295c05e81917928d76245e842748184d	5704	sao paulo	SP
1f1c7bf1c9b041b292af6c1c4470b753	3151a81801c8386361b62277d7fa5ecf	95110	caxias do sul	RS
206f3129c0e4d7d0b9550426023f0a08	21f748a16f4e1688a9014eb3ee6fa325	13412	piracicaba	SP
a7c125a0a07b75146167b7f04a7f8e98	5c2991dbd08bbf3cf410713c4de5a0b5	22750	rio de janeiro	RJ
c5c61596a3b6bd0cee5766992c48a9a1	b6e99561fe6f34a55b0b7da92f8ed775	7124	guarulhos	SP
9b8ce803689b3562defaad4613ef426f	7f3a72e8f988c6e735ba118d54f47458	5416	sao paulo	SP
49d0ea0986edde72da777f15456a0ee0	3e6fd6b2f0d499456a6a6820a40f2d79	68485	pacaja	PA

The data in the preview has been truncated due to size limits.

Figure 6 Preview of the imported dataset



Python script

Script

```
import pandas as pd
from datetime import timedelta

orders = pd.read_csv("C:\\Users\\2702b\\OneDrive - Asia Pacific University\\Diploma\\Semester

orders = orders.drop(["order_approved_at", "order_delivered_carrier_date", "order_estimated_de

orders["order_purchase_timestamp"] = pd.to_datetime(orders["order_purchase_timestamp"], forma
orders["order_delivered_customer_date"] = pd.to_datetime(orders["order_delivered_customer_date

orders = orders[orders["order_status"] == "delivered"]
orders["order_delivered_customer_date"] = orders["order_delivered_customer_date"].fillna(order
```

The script will run with the following Python installation C:\USERS\2702B\APPDATA\LOCAL\PROGRAMS\PYTHON\PYTHON39.

To configure your settings and change which Python installation you want to run, go to Options and settings.

Figure 7 Built-in Python support to import files

Once all the required datasets, which are stated in Chapter 4, are imported, the cleaning process can start.

Data Selection

As described in the previous section, several tables have a few missing values that need to be filled. However, before filling in the values, columns that are not needed will first be dropped. Since analysis will mainly be done on the sales of the product and figuring out the relationship between the product's dimensions with the freight value and delivery duration, columns that are unrelated to these objectives or scopes will be dropped. The columns are:

Columns	Source table
customer_unique_id	olist_customer_dataset
seller_id	olist_order_items_dataset
shipping_limit_date	olist_order_items_dataset
order_approved_at	olist_orders_dataset
order_delivered_carrier_date	olist_orders_dataset
order_estimated_delivery_date	olist_orders_dataset
product_name_lenght	olist_products_dataset
product_description_lenght	olist_products_dataset
product_photos_qty	olist_products_dataset

Filtering columns

As seen in the tables in Chapter 4, several columns are missing data in 2 of the datasets. However, before filling in the missing data with either the mean or median, the relationship between each column within the dataset must first be understood.

In the olist_orders_dataset, there are 3 columns that have missing data, namely columns "order_approved_at", "order_delivered_carrier_date" and "order_delivered_customer_date". However, columns "order_approved_at" and "order_delivered_carrier_date" have been dropped in the previous step and the remaining column needs to be checked with the order status. This is because whether an order has a date in column "order_delivered_customer_date" depends on whether has the order been delivered yet or not, and this information can be found in column "order_status" in the same dataset.

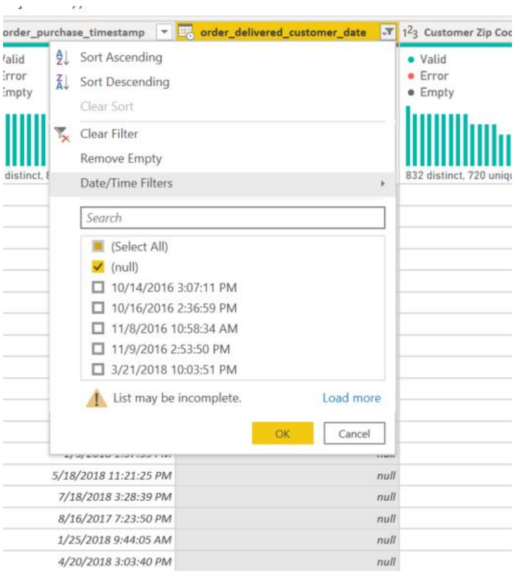
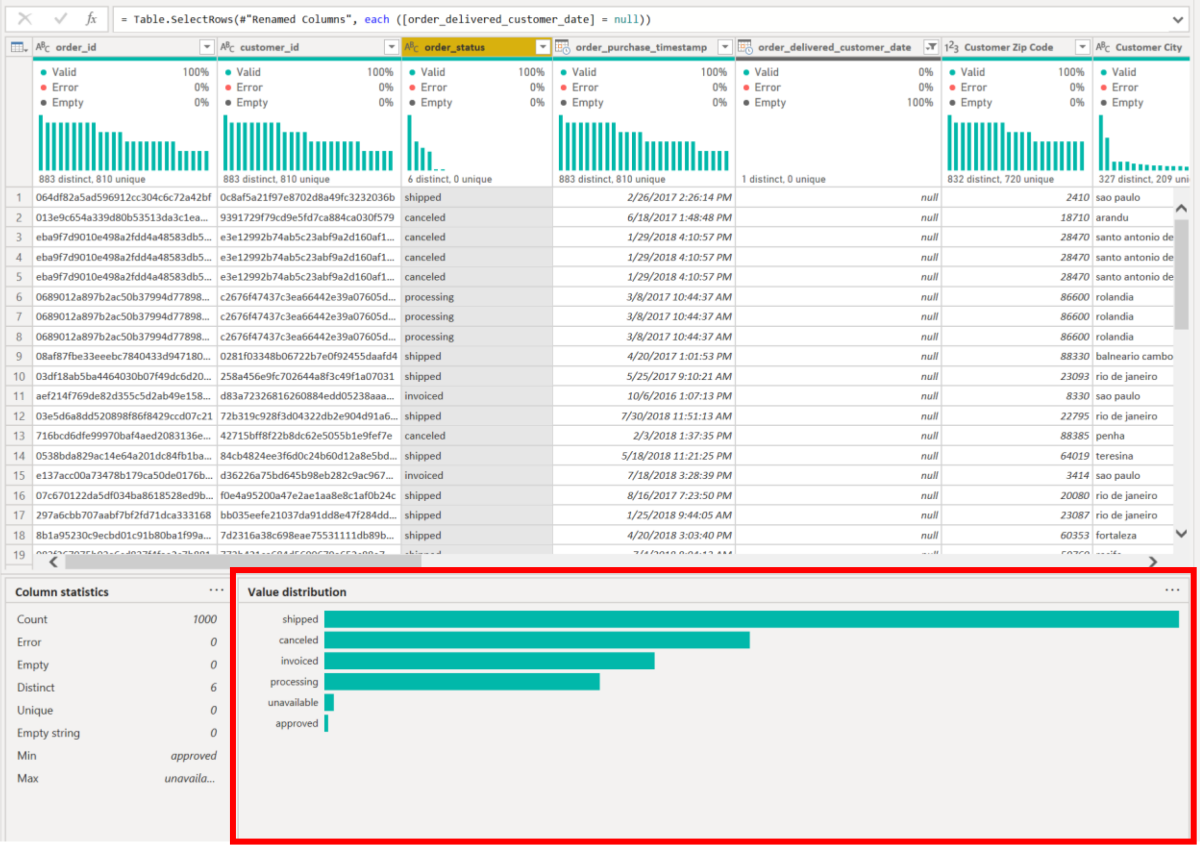


Figure 8 Filtering "order_delivered_customer_date"

By only filtering out the “null” values in the column “order_delivered_customer_date” as shown in the above figure, the data in all other columns will also be filtered. Then, by using the column profiles feature found in Microsoft Power BI, the column “order_status” can be inspected and the value distribution in that column can be seen in Figure 2 below.



```
= Table.SelectRows(#"Renamed Columns", each ([order_delivered_customer_date] = null))
```

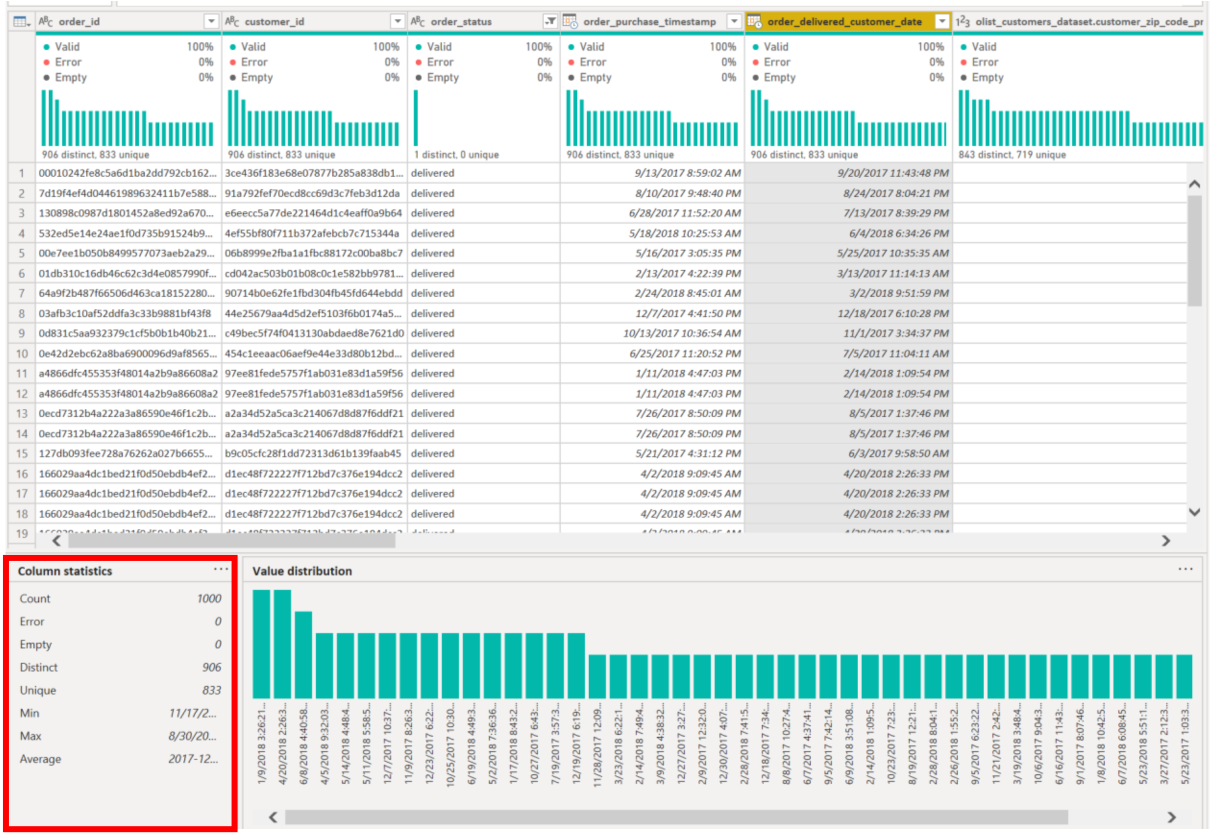
	order_id	customer_id	order_status	order_purchase_timestamp	order_delivered_customer_date	Customer Zip Code	Customer City
	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 0% ● Error 0% ● Empty 100%	● Valid 100% ● Error 0% ● Empty 0%	● Valid ● Error ● Empty
	883 distinct, 810 unique	883 distinct, 810 unique	6 distinct, 0 unique	883 distinct, 810 unique	1 distinct, 0 unique	832 distinct, 720 unique	327 distinct, 209 uni
1	064df82a5ad596912cc304c6c72a42bf	0c8af5a21f97e8702d8a49fc3232036b	shipped	2/26/2017 2:26:14 PM	null	2410	sao paulo
2	013e9c654a339d80b53513da3c1ea...	9391729f79cd9e5fd7ca884ca030f579	canceled	6/18/2017 1:48:48 PM	null	18710	arandu
3	eba9f7d9010e498a2fdd4a48583db5...	e3e12992b74ab5c23abf9a2d160af1...	canceled	1/29/2018 4:10:57 PM	null	28470	santo antonio de
4	eba9f7d9010e498a2fdd4a48583db5...	e3e12992b74ab5c23abf9a2d160af1...	canceled	1/29/2018 4:10:57 PM	null	28470	santo antonio de
5	eba9f7d9010e498a2fdd4a48583db5...	e3e12992b74ab5c23abf9a2d160af1...	canceled	1/29/2018 4:10:57 PM	null	28470	santo antonio de
6	0689012a897b2ac50b37994d77898...	c2676f47437c3ea66442e39a07605d...	processing	3/8/2017 10:44:37 AM	null	86600	rolandia
7	0689012a897b2ac50b37994d77898...	c2676f47437c3ea66442e39a07605d...	processing	3/8/2017 10:44:37 AM	null	86600	rolandia
8	0689012a897b2ac50b37994d77898...	c2676f47437c3ea66442e39a07605d...	processing	3/8/2017 10:44:37 AM	null	86600	rolandia
9	08af87fbe33eeebc7840433d947180...	0281f03348b06722b7e0f92455daafd4	shipped	4/20/2017 1:01:53 PM	null	88330	balneario cambo
10	03df18ab5ba4464030b07f49dc6d20...	258a456e9fc702644a8f3c49f1a07031	shipped	5/25/2017 9:10:21 AM	null	23093	rio de janeiro
11	aef214f769de82d355c5d2ab49e158...	d83a72326816260884edd05238aaa...	invoiced	10/6/2016 1:07:13 PM	null	8330	sao paulo
12	03e5d6a8dd520898f86f8429ccd07c21	72b319c928f3d04322db2e904d91a6...	shipped	7/30/2018 11:51:13 AM	null	22795	rio de janeiro
13	716bcd6dfe99970baf4aed2083136e...	42715bff8f22b8dc62e5055b1e9fef7e	canceled	2/3/2018 1:37:35 PM	null	88385	penha
14	0538bda829ac14e64a201dc84fb1ba...	84cb4824ee3f6d0c24b60d12a8e5bd...	shipped	5/18/2018 11:21:25 PM	null	64019	teresina
15	e137acc00a73478b179ca50de0176b...	d36226a75bd645b98eb282c9ac967...	invoiced	7/18/2018 3:28:39 PM	null	3414	sao paulo
16	07c670122da5df034ba8618528ed9b...	f0e4a95200a47e2ae1aa8e8c1af0b24c	shipped	8/16/2017 7:23:50 PM	null	20080	rio de janeiro
17	297a6cbb707aabf7bf2fd71dca333168	bb035eefe21037da91dd8e47f284dd...	shipped	1/25/2018 9:44:05 AM	null	23087	rio de janeiro
18	8b1a95230c9ecbd01c91b80ba1f99a...	7d2316a38c698eae75531111db89b...	shipped	4/20/2018 3:03:40 PM	null	60353	fortaleza

Column statistics

Count	1000
Error	0
Empty	0
Distinct	6
Unique	0
Empty string	0
Min	approved
Max	unavaila...

Figure 9 Column profile of "order_status" after filtering

Based on the highlighted red box in the figure above, most of the values in "order_status" are "shipped", "cancelled", and "invoiced". None of the values is labelled as "delivered". Therefore, instead of filling in values for the column "order_delivered_customer_date", it is easier to just filter out data that has the value "delivered" in the column "order_status". Figure 3 below displays the column profile of "order_delivered_customer_date" after filtering the column "order_status" to only contain values "delivered". This step is used also because a new column will be created later on which is the Delivery Duration, and only delivered items have a delivery duration. There are no empty values inside the column "order_delivered_customer_date" after placing the filter but this preview only displays the first 1000 rows of the dataset. So there might still be some missing values after loading the entire dataset.



	order_id	customer_id	order_status	order_purchase_timestamp	order_delivered_customer_date	olist_customers_dataset.customer_zip_code_pr
	● Valid 100% ● Error 0% ● Empty 0% 906 distinct, 833 unique	● Valid 100% ● Error 0% ● Empty 0% 906 distinct, 833 unique	● Valid 100% ● Error 0% ● Empty 0% 1 distinct, 0 unique	● Valid 100% ● Error 0% ● Empty 0% 906 distinct, 833 unique	● Valid 100% ● Error 0% ● Empty 0% 906 distinct, 833 unique	● Valid ● Error ● Empty 843 distinct, 719 unique
1	00010242fe8c5a6d1ba2dd792cb162...	3ce436f183e68e07877b285a838db1...	delivered	9/13/2017 8:59:02 AM	9/20/2017 11:43:48 PM	
2	7d19f4ef4d04461989632411b7e588...	91a792fef70ecd8cc69d3c7feb3d12da	delivered	8/10/2017 9:48:40 PM	8/24/2017 8:04:21 PM	
3	130898c0987d1801452a8ed92a670...	e6eecc5a77de221464d1c4eaff0a9b64	delivered	6/28/2017 11:52:20 AM	7/13/2017 8:39:29 PM	
4	532ed5e14e24ae1f0d735b91524b9...	4ef55bf80f711b372afebcb7c715344a	delivered	5/18/2018 10:25:53 AM	6/4/2018 6:34:26 PM	
5	00e7ee1b050b8499577073aeb2a29...	06b8999e2fba1a1fbc88172c00ba8bc7	delivered	5/16/2017 3:05:35 PM	5/25/2017 10:35:35 AM	
6	01db310c16db46c62c3d4e0857990f...	cd042ac503b01b08c0c1e582bb9781...	delivered	2/13/2017 4:22:39 PM	3/13/2017 11:14:13 AM	
7	64a9f2b487f66506d463ca18152280...	90714b0e62fe1fbd304fb45fd644ebdd	delivered	2/24/2018 8:45:01 AM	3/2/2018 9:51:59 PM	
8	03afb3c10af52ddfa3c33b9881bf43f8	44e25679aa4d5d2ef5103f6b0174a5...	delivered	12/7/2017 4:41:50 PM	12/18/2017 6:10:28 PM	
9	0d831c5aa932379c1cf5b0b1b40b21...	c49bec5f74f0413130abdaed8e7621d0	delivered	10/13/2017 10:36:54 AM	11/1/2017 3:34:37 PM	
10	0e42d2ebc62a8ba6900096d9af8565...	454c1eeaac06aef9e44e33d80b12bd...	delivered	6/25/2017 11:20:52 PM	7/5/2017 11:04:11 AM	
11	a4866dfc455353f48014a2b9a86608a2	97ee81fede5757f1ab031e83d1a59f56	delivered	1/11/2018 4:47:03 PM	2/14/2018 1:09:54 PM	
12	a4866dfc455353f48014a2b9a86608a2	97ee81fede5757f1ab031e83d1a59f56	delivered	1/11/2018 4:47:03 PM	2/14/2018 1:09:54 PM	
13	0ecd7312b4a222a3a86590e46f1c2b...	a2a34d52a5ca3c214067d8d87f6ddf21	delivered	7/26/2017 8:50:09 PM	8/5/2017 1:37:46 PM	
14	0ecd7312b4a222a3a86590e46f1c2b...	a2a34d52a5ca3c214067d8d87f6ddf21	delivered	7/26/2017 8:50:09 PM	8/5/2017 1:37:46 PM	
15	127db093fee728a76262a027b6655...	b9c05cfc28f1dd72313d61b139faab45	delivered	5/21/2017 4:31:12 PM	6/3/2017 9:58:50 AM	
16	166029aa4dc1bed21f0d50ebdb4ef2...	d1ec48f722227f712bd7c376e194dcc2	delivered	4/2/2018 9:09:45 AM	4/20/2018 2:26:33 PM	
17	166029aa4dc1bed21f0d50ebdb4ef2...	d1ec48f722227f712bd7c376e194dcc2	delivered	4/2/2018 9:09:45 AM	4/20/2018 2:26:33 PM	
18	166029aa4dc1bed21f0d50ebdb4ef2...	d1ec48f722227f712bd7c376e194dcc2	delivered	4/2/2018 9:09:45 AM	4/20/2018 2:26:33 PM	

Column statistics

Count	1000
Error	0
Empty	0
Distinct	906
Unique	833
Min	11/17/2...
Max	8/30/20...
Average	2017-12...

Value distribution

Figure 10 Column profile of "order_delivered_customer_date" after filtering

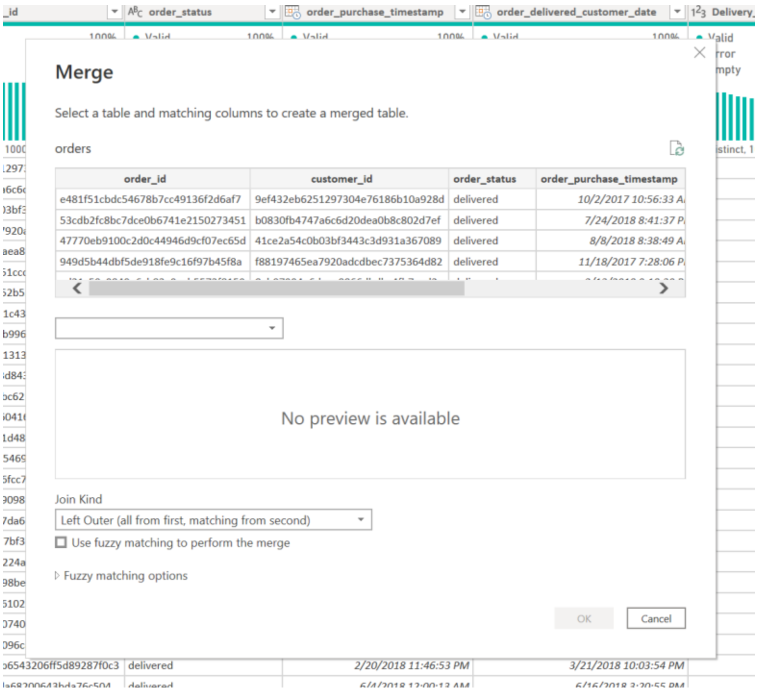
Renaming Columns

Once the datasets have been filtered, all the columns in every imported dataset will be renamed for easier understanding. The table below shows some examples of the name of the new columns:

Original Column Names	Renamed Column Names	Dataset
product_weight_g	Product Weight (g)	olist_products_dataset
customer_zip_code_prefix	Zip Code	olist_customers_dataset
Order_delivered_customer_date	Order receive date by customer	olist_orders_dataset

Merging the datasets

In Power BI, merging the datasets is simple. The user just needs to select the Merge Queries option and then choose the tables and the columns that they are going to merge on. Figure 8 shows the Merge Queries option found in the ribbon of Power BI. Figure 9 shows the dialogue box of the Merge Queries option where the user will choose which tables to merge and which columns to merge the tables on. The highlighted rows shown in Figure 9 are the columns that Power BI will use to merge the 2 different datasets.



Merge

Select a table and matching columns to create a merged table.

orders

order_id	customer_id	order_status	order_purchase_timestamp
e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	10/2/2017 10:56:33 A
53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	7/24/2018 8:41:37 P
47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	8/8/2018 8:38:49 A
949d5b44dbf5de918fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	11/18/2017 7:28:06 P

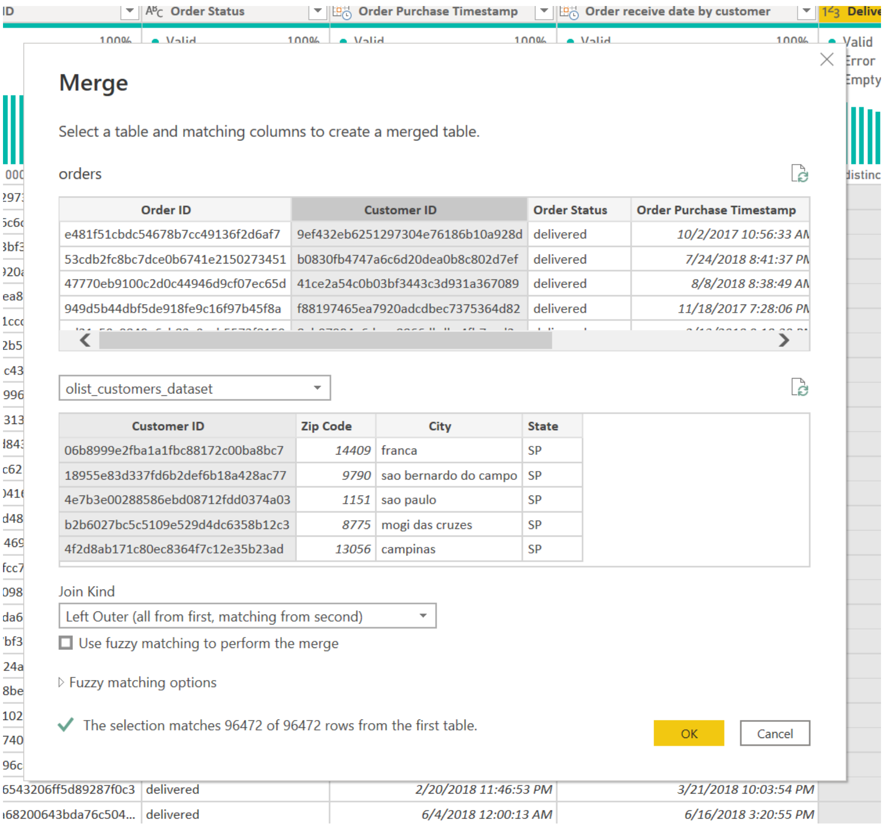
No preview is available

Join Kind

Left Outer (all from first, matching from second)

□ Use fuzzy matching to perform the merge

Figure 11 Merge Queries dialogue box



Merge

Select a table and matching columns to create a merged table.

orders

Order ID	Customer ID	Order Status	Order Purchase Timestamp
e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	10/2/2017 10:56:33 AM
53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	7/24/2018 8:41:37 PM
47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	8/8/2018 8:38:49 AM
949d5b44dbf5de918fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	11/18/2017 7:28:06 PM

olist_customers_dataset

Customer ID	Zip Code	City	State
06b8999e2fba1a1fbc88172c00ba8bc7	14409	franca	SP
18955e83d337fd6b2def6b18a428ac77	9790	sao bernardo do campo	SP
4e7b3e00288586ebd08712fdd0374a03	1151	sao paulo	SP
b2b6027bc5c5109e529d4dc6358b12c3	8775	mogi das cruzes	SP
4f2d8ab171c80ec8364f7c12e35b23ad	13056	campinas	SP

Join Kind

Left Outer (all from first, matching from second)

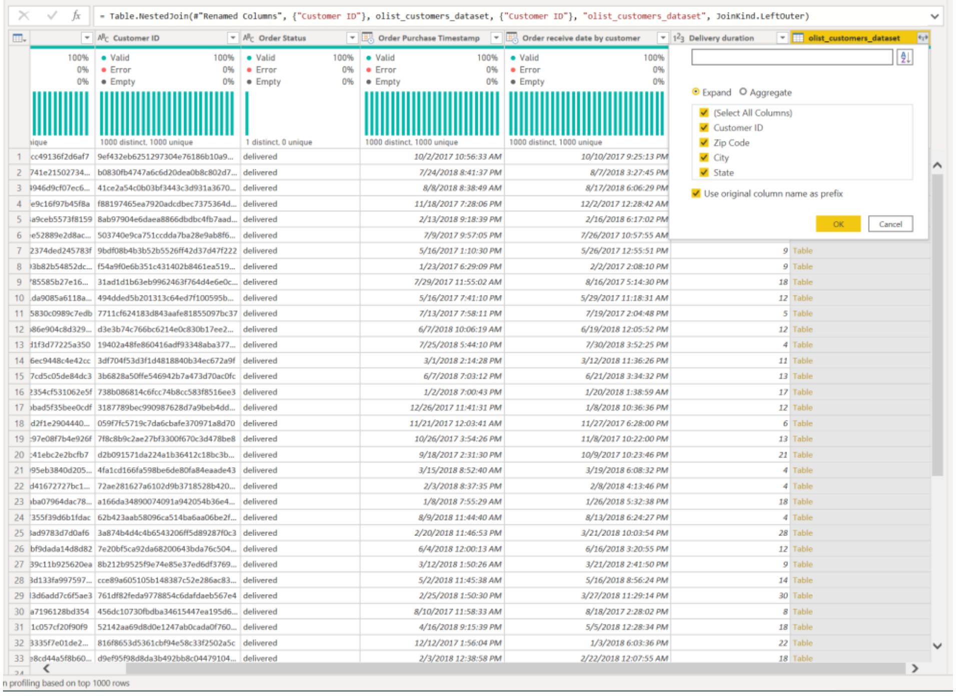
☐ Use fuzzy matching to perform the merge

Fuzzy matching options

✓ The selection matches 96472 of 96472 rows from the first table.

Figure 12 Merging 2 datasets

Once the 2 datasets are merged, the user will need to expand the columns of the merged table. In this step, the user can choose which columns to keep and which to remove as shown in Figure 10.



```
= Table.NestedJoin(#"Renamed Columns", {"Customer ID"}, olist_customers_dataset, {"Customer ID"}, "olist_customers_dataset", JoinKind.LeftOuter)
```

		Customer ID	Order Status	Order Purchase Timestamp	Order receive date by customer	Delivery duration	olist_customers_dataset
	100% 0% 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%		
	iique	1000 distinct, 1000 unique	1 distinct, 0 unique	1000 distinct, 1000 unique	1000 distinct, 1000 unique		
1	cc49136f2d6af7	9ef432eb6251297304e76186b10a9...	delivered	10/2/2017 10:56:33 AM	10/10/2017 9:25:13 PM		
2	741e21502734...	b0830fb4747a6c6d20dea0b8c802d7...	delivered	7/24/2018 8:41:37 PM	8/7/2018 3:27:45 PM		
3	946d9cf07ec6...	41ce2a54c0b03bf3443c3d931a3670...	delivered	8/8/2018 8:38:49 AM	8/17/2018 6:06:29 PM		
4	e9c16f97b45f8a	f88197465ea7920adcdbec7375364d...	delivered	11/18/2017 7:28:06 PM	12/2/2017 12:28:42 AM		
5	a9ceb5573f8159	8ab97904e6daea8866dbdbc4fb7aad...	delivered	2/13/2018 9:18:39 PM	2/16/2018 6:17:02 PM		
6	e52889e2d8ac...	503740e9ca751ccdda7ba28e9ab8f6...	delivered	7/9/2017 9:57:05 PM	7/26/2017 10:57:55 AM		
7	2374ded245783f	9bdf08b4b3b52b5526ff42d37d47f222	delivered	5/16/2017 1:10:30 PM	5/26/2017 12:55:51 PM	9	Table
8	3b82b54852dc...	f54a9f0e6b351c431402b8461ea519...	delivered	1/23/2017 6:29:09 PM	2/2/2017 2:08:10 PM	9	Table
9	85585b27e16...	31ad1d1b63eb9962463f764d4e6e0c...	delivered	7/29/2017 11:55:02 AM	8/16/2017 5:14:30 PM	18	Table
10	da9085a6118a...	494dded5b201313c64ed7f100595b...	delivered	5/16/2017 7:41:10 PM	5/29/2017 11:18:31 AM	12	Table
11	5830c0989c7edb	7711cf624183d843aafe81855097bc37	delivered	7/13/2017 7:58:11 PM	7/19/2017 2:04:48 PM	5	Table
12	86e904c8d329...	d3e3b74c766bc6214e0c830b17ee2...	delivered	6/7/2018 10:06:19 AM	6/19/2018 12:05:52 PM	12	Table
13	d1f3d77225a350	19402a48fe860416adf93348aba377...	delivered	7/25/2018 5:44:10 PM	7/30/2018 3:52:25 PM	4	Table
14	6ec9448c4e42cc	3df704f53d3f1d4818840b34ec672a9f	delivered	3/1/2018 2:14:28 PM	3/12/2018 11:36:26 PM	11	Table
15	7cd5c05de84dc3	3b6828a50ffe546942b7a473d70ac0fc	delivered	6/7/2018 7:03:12 PM	6/21/2018 3:34:32 PM	13	Table
16	2354cf531062e5f	738b086814c6fcc74b8cc583f8516ee3	delivered	1/2/2018 7:00:43 PM	1/20/2018 1:38:59 AM	17	Table
17	bad5f35bee0cdf	3187789bec990987628d7a9beb4dd...	delivered	12/26/2017 11:41:31 PM	1/8/2018 10:36:36 PM	12	Table
18	d2f1e2904440...	059f7fc5719c7da6cbafe370971a8d70	delivered	11/21/2017 12:03:41 AM	11/27/2017 6:28:00 PM	6	Table
19	97e08f7b4e926f	7f8c8b9c2ae27bf3300f670c3d478be8	delivered	10/26/2017 3:54:26 PM	11/8/2017 10:22:00 PM	13	Table
20	41ebc2e2bcfb7	d2b091571da224a1b36412c18bc3b...	delivered	9/18/2017 2:31:30 PM	10/9/2017 10:23:46 PM	21	Table
21	95eb3840d205...	4fa1cd166fa598be6de80fa84eaade43	delivered	3/15/2018 8:52:40 AM	3/19/2018 6:08:32 PM	4	Table
22	d41672727bc1...	72ae281627a6102d9b3718528b420...	delivered	2/3/2018 8:37:35 PM	2/8/2018 4:13:46 PM	4	Table
23	ba07964dac78...	a166da34890074091a942054b36e4...	delivered	1/8/2018 7:55:29 AM	1/26/2018 5:32:38 PM	18	Table
24	355f39d6b1fdac	62b423aab58096ca514ba6aa06be2f...	delivered	8/9/2018 11:44:40 AM	8/13/2018 6:24:27 PM	4	Table
25	ad9783d7d0af6	3a874b4d4c4b6543206ff5d89287f0c3	delivered	2/20/2018 11:46:53 PM	3/21/2018 10:03:54 PM	28	Table
26	bf9dada14d8d82	7e20bf5ca92da68200643bda76c504...	delivered	6/4/2018 12:00:13 AM	6/16/2018 3:20:55 PM	12	Table
27	39c11b925620ea	8b212b9525f9e74e85e37ed6df3769...	delivered	3/12/2018 1:50:26 AM	3/21/2018 2:41:50 PM	9	Table
28	3d133fa997597...	cce89a605105b148387c52e286ac83...	delivered	5/2/2018 11:45:38 AM	5/16/2018 8:56:24 PM	14	Table
29	3d6add7c6f5ae3	761df82feda9778854c6dafdaeb567e4	delivered	2/25/2018 1:50:30 PM	3/27/2018 11:29:14 PM	30	Table
30	a7196128bd354	456dc10730fbdba34615447ea195d6...	delivered	8/10/2017 11:58:33 AM	8/18/2017 2:28:02 PM	8	Table
31	1c057cf20f90f9	52142aa69d8d0e1247ab0cada0f760...	delivered	4/16/2018 9:15:39 PM	5/5/2018 12:28:34 PM	18	Table
32	3335f7e01de2...	816f8653d5361cbf94e58c33f2502a5c	delivered	12/12/2017 1:56:04 PM	1/3/2018 6:03:36 PM	22	Table
33	8cd44a5f8b60...	d9ef95f98d8da3b492bb8c04479104...	delivered	2/3/2018 12:38:58 PM	2/22/2018 12:07:55 AM	18	Table

● Expand ○ Aggregate

☑ (Select All Columns)
☑ Customer ID
☑ Zip Code
☑ City
☑ State

☑ Use original column name as prefix

n profiling based on top 1000 rows

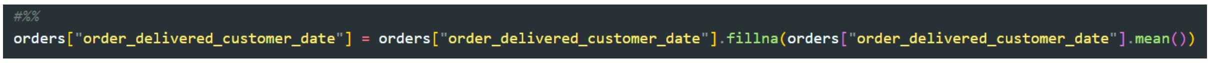
Figure 13 Choosing which columns to expand

Once all the datasets are merged, several columns' data types will be changed to more suitable formats. The table below shows the changes made to the data types of certain columns:

Column name	Original Data Type	Changed Data Type
Order Purchase Timestamp	Date/Time	Date
Order receive date by customer	Date/Time	Date
Price	Decimal Number	Fixed Decimal Number
Freight Value	Decimal Number	Fixed Decimal Number

Filling in missing values

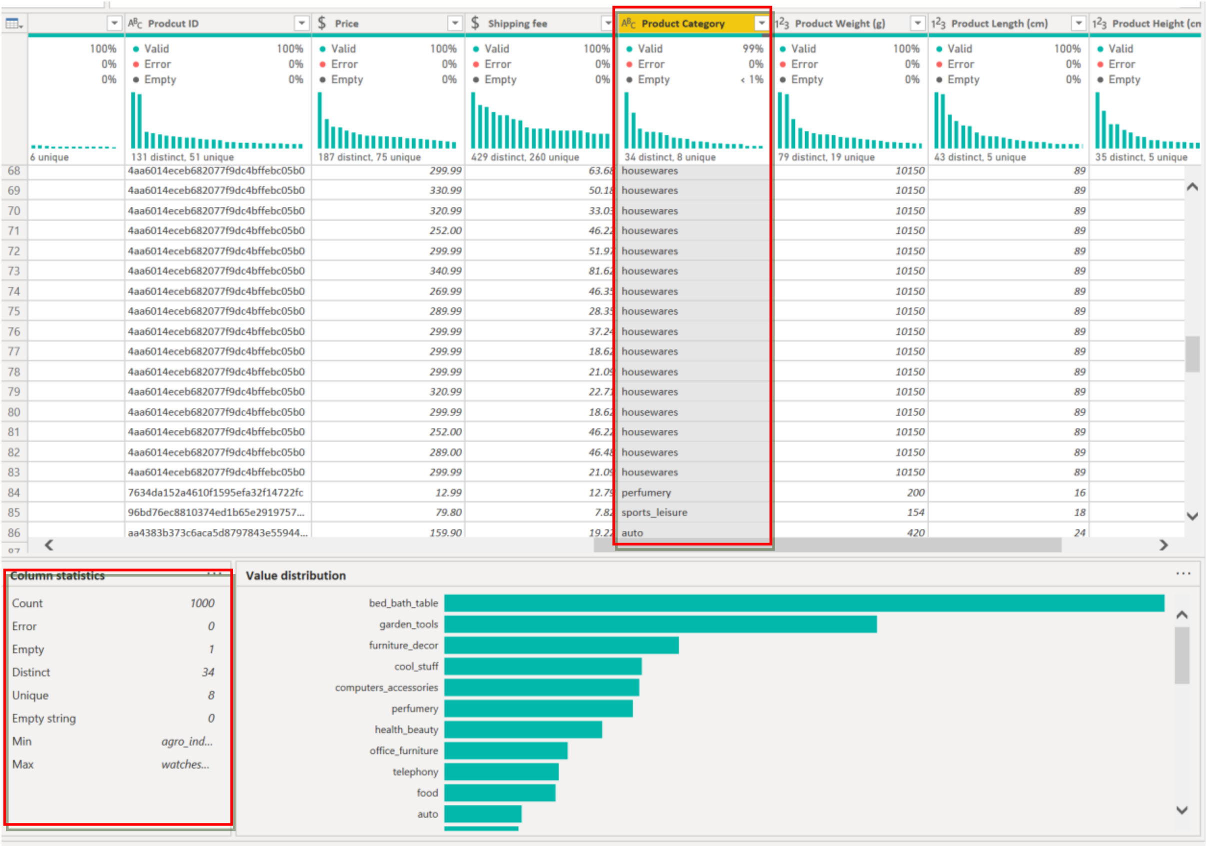
At the start of this Data Cleaning step, the olist_orders_dataset contains some missing values in the "order_delivered_customer_date" column. The missing values in this column need to be filled in because this column will be used to create the calculated column "Delivery Duration" later on. Therefore, the missing values found in this column will be filled in with the mean value of the entire column. This can be done by using Python in Power BI. Power BI allows the user or analyst to run python scripts to achieve functionalities that is not fully supported in Power BI itself. In this case, filling in the missing values with the mean is easier to accomplish with Python rather than using DAX, the programming language in Power BI. The code used to fill in the missing values are saved in the cleaning_ordres.py file and shown below:



```
#%%
orders["order_delivered_customer_date"] = orders["order_delivered_customer_date"].fillna(orders["order_delivered_customer_date"].mean())
```

Figure 14 Filling in missing values for olist_orders_dataset

After merging the datasets, there are still some missing values in the dataset. The column 'Product Category' contains missing values as shown in the figure below.



		Prodcut ID	Price	Shipping fee	Product Category	Product Weight (g)	Product Length (cm)	Product Height (cn
	100% 0% 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 99% ● Error 0% ● Empty < 1%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid ● Error ● Empty
	6 unique	131 distinct, 51 unique	187 distinct, 75 unique	429 distinct, 260 unique	34 distinct, 8 unique	79 distinct, 19 unique	43 distinct, 5 unique	35 distinct, 5 unique
68		4aa6014eceb682077f9dc4bffebc05b0	299.99	63.68	housewares	10150	89	
69		4aa6014eceb682077f9dc4bffebc05b0	330.99	50.18	housewares	10150	89	
70		4aa6014eceb682077f9dc4bffebc05b0	320.99	33.03	housewares	10150	89	
71		4aa6014eceb682077f9dc4bffebc05b0	252.00	46.22	housewares	10150	89	
72		4aa6014eceb682077f9dc4bffebc05b0	299.99	51.97	housewares	10150	89	
73		4aa6014eceb682077f9dc4bffebc05b0	340.99	81.62	housewares	10150	89	
74		4aa6014eceb682077f9dc4bffebc05b0	269.99	46.35	housewares	10150	89	
75		4aa6014eceb682077f9dc4bffebc05b0	289.99	28.35	housewares	10150	89	
76		4aa6014eceb682077f9dc4bffebc05b0	299.99	37.24	housewares	10150	89	
77		4aa6014eceb682077f9dc4bffebc05b0	299.99	18.62	housewares	10150	89	
78		4aa6014eceb682077f9dc4bffebc05b0	299.99	21.09	housewares	10150	89	
79		4aa6014eceb682077f9dc4bffebc05b0	320.99	22.71	housewares	10150	89	
80		4aa6014eceb682077f9dc4bffebc05b0	299.99	18.62	housewares	10150	89	
81		4aa6014eceb682077f9dc4bffebc05b0	252.00	46.22	housewares	10150	89	
82		4aa6014eceb682077f9dc4bffebc05b0	289.00	46.48	housewares	10150	89	
83		4aa6014eceb682077f9dc4bffebc05b0	299.99	21.09	housewares	10150	89	
84		7634da152a4610f1595efa32f14722fc	12.99	12.79	perfumery	200	16	
85		96bd76ec8810374ed1b65e2919757...	79.80	7.82	sports_leisure	154	18	
86		aa4383b373c6aca5d8797843e55944...	159.90	19.22	auto	420	24	

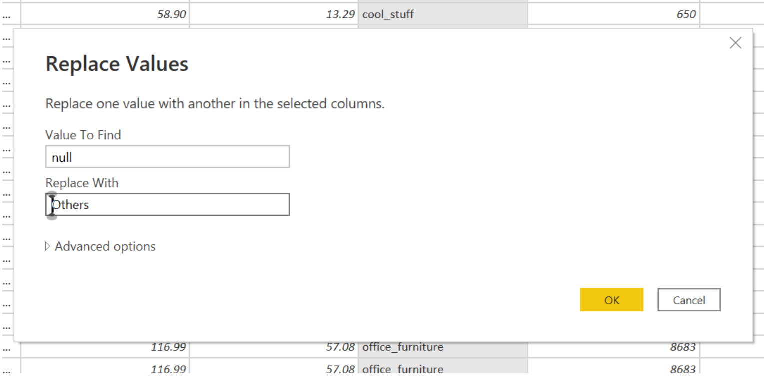
Column statistics

Count	1000
Error	0
Empty	1
Distinct	34
Unique	8
Empty string	0
Min	agro_ind...
Max	watches...

Value distribution

Figure 15 'Product Category' column missing values

Since this column is a categorical value, it does not make sense to fill in the missing values with the mean or median. After reading through the dataset's documentation, these blank values are most likely products that do not fall into any of the existing categories. Therefore, these missing values will be replaced with a new product category named 'Others'. This step can be done through the Replace Values function found in Power BI as shown in the figure below, where the value to find is replaced with the new value.



... 58.90 13.29 cool_stuff 650

Replace Values

Replace one value with another in the selected columns.

Value To Find: null

Replace With: Others

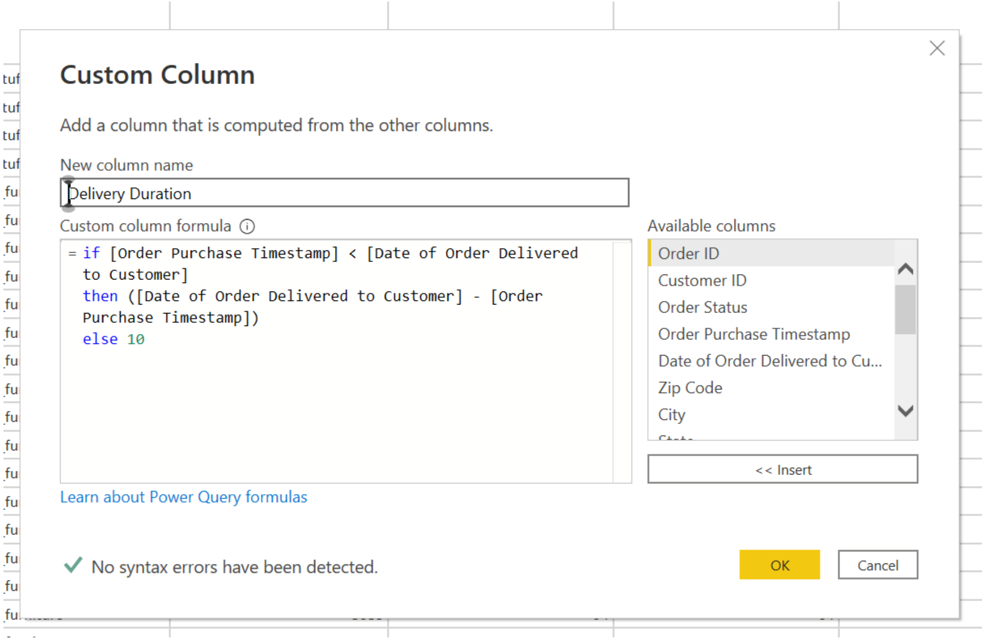
... 116.99 57.08 office_furniture 8683

... 116.99 57.08 office_furniture 8683

Figure 16 Replacing null values

Creating new columns/measures

Since one of the objectives of this analysis is to identify the relationship between the shipping fee and the time needed to deliver the product to the customer, a new column that is related to the time needed to deliver the product is needed. This column can be prepared in Power BI itself. To prepare this new column, the ‘Add Custom Column’ function in Power BI can be used to create the ‘Delivery Duration’ column that tracks the amount of time needed to deliver a product to the customer. The figure below shows the dialogue box that contains the formula to calculate the Delivery Duration.



Custom Column

Add a column that is computed from the other columns.

New column name

Delivery Duration

Custom column formula

```
= if [Order Purchase Timestamp] < [Date of Order Delivered
  to Customer]
  then ([Date of Order Delivered to Customer] - [Order
  Purchase Timestamp])
  else 10
```

Available columns

- Order ID
- Customer ID
- Order Status
- Order Purchase Timestamp
- Date of Order Delivered to Cu...
- Zip Code
- City

Learn about Power Query formulas

✓ No syntax errors have been detected.

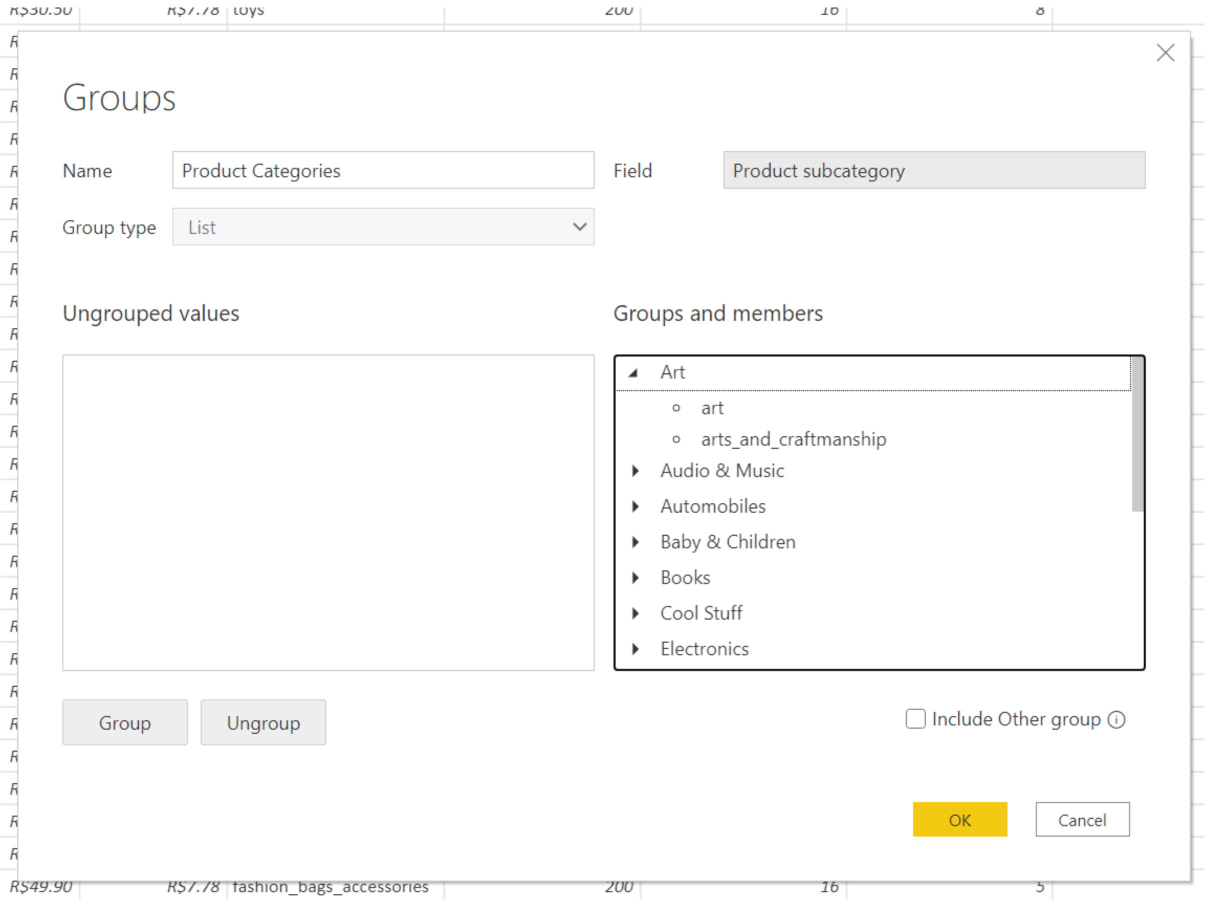
Figure 17 Creating 'Delivery Duration' column

Changing the data format and categorization

Once all the datasets are merged and cleaned, they can then be loaded to Power BI where more tools are available to further transform the data. After loading the data in Power BI, the format of the date columns were changed. Therefore, for easier analysis, the format of the date column will be changed from Long Date to mm/dd/yyyy. Furthermore, columns that contain regional data will also be categorized to their correct category, for example, the Zip Code column will be categorized as Postal Code, and the State column will be categorized as State/Province. Columns that contain price information like the Price and Shipping fee column will also be changed to Currency type to ensure that the calculations involving these columns will be accurate.

Grouping data

After changing the format of the data, there is one column that has data that is very similar and that column is the product categories. Several categories are very similar. For example, the categories arts and arts_and_craftsmanship. These 2 categories sell very similar products, thus they are grouped into a category named 'Art', and they become the subcategories under this larger category. Other product categories are converted into subcategories by grouping them into larger and generalized categories as shown in the figure below:



Groups

Name: Product Categories

Field: Product subcategory

Group type: List

Ungrouped values

Groups and members

- Art
  - art
  - arts_and_craftmanship
- Audio & Music
- Automobiles
- Baby & Children
- Books
- Cool Stuff
- Electronics

☐ Include Other group

Figure 18 Grouping categories

By grouping the categories, it can reduce the number of data points that will be shown in the visuals later on during data visualization. Moreover, if new product categories are introduced, they can also be grouped under these new generalized categories to reduce similarities between the existing product categories.

6. Data Visualization and Analysis

After Data cleaning is done, the next step which is preparing the analytical dashboard with visuals based on the data from the datasets imported can be done. As per the scope detailed in the earlier chapters, the analysis to be done on the imported dataset is to visualize the sales made on the Olist platform from late 2016 to early 2018.

OLAP

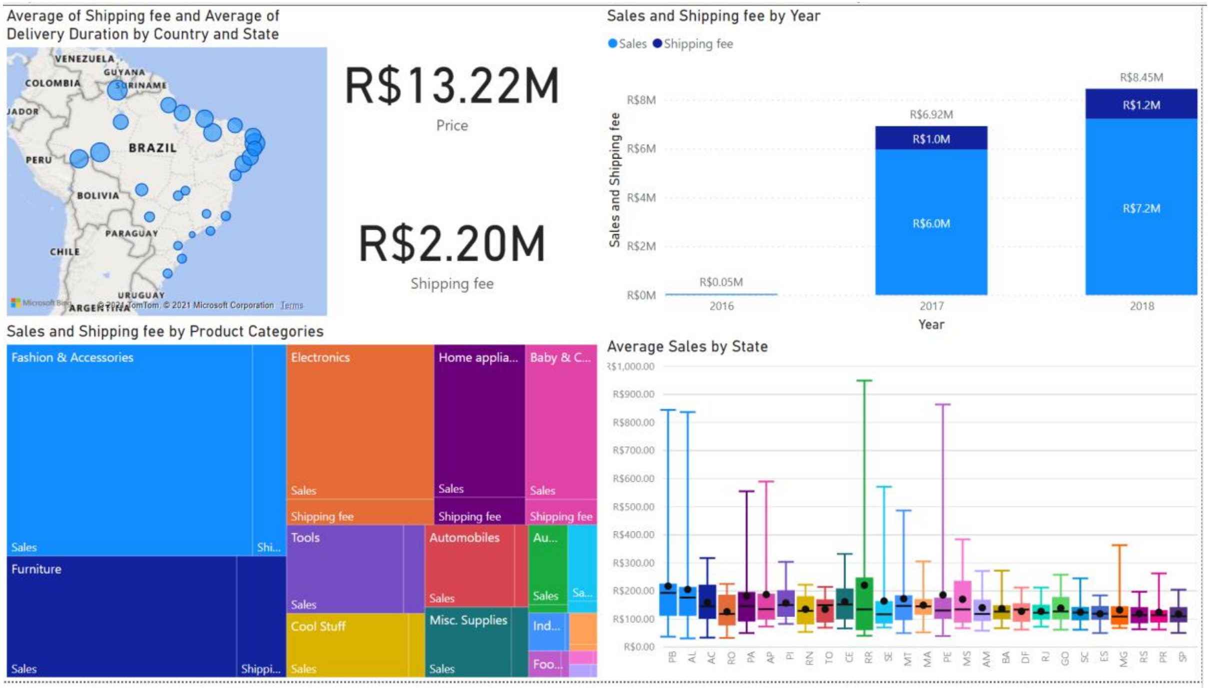


Figure 19 Sales and shipping fee visualization

Description

As seen in the figure above, 4 charts show the Sales and Shipping fee charged on the Olist platform from 2016 to 2018. The stacked column chart shows the total sales and shipping fee for a particular year made by Olist. As observed in the column chart, the sales increase year by year, with a huge growth from 2016 to 2017. This is because the dataset only contains the last 3 months of 2016 rather than the full year. This column chart on the top right corner can be drilled down further to the quarter of the selected year, showing a more specific view of the sales of each quarter and month of that specific year. The drill-down feature will be shown in a later section.

Besides the column chart, the boxplot below the column chart shows the average sales of each state in Brazil. The boxplot shows data about the quantiles, mean, average, maximum and minimum sales generated in each state. From this chart itself, it can be seen that the state

RR generates the highest amount of sales compared to other states, both in terms of the average and the maximum. However, the maximum sales generated does not indicate that the state will constantly generate such high sales. The whiskers at both ends only show the maximum and minimum sales generated by the state. Most of the sales generated will fall under the boxes in the centre of each plot, which shows the interquartile range of the sales generated in that state. For example, the interquartile range of the state RR is between R$62.88 and R$244.44, which means that the bulk of the sales generated in RR is in that range with only occasionally higher or lower sales.

The third chart found in the OLAP dashboard is the treemap which shows the sales and shipping fee of each product category sold in Olist. This treemap shows which product category generates the most sales and charges the highest shipping fee among all product categories. Looking at the treemap, it is evident that Fashion & Accessories is the most profiting product category in Olist out of all other categories, generating over R$ 4 million in sales and R$ 500,000 in shipping fees. This chart can also be drilled down into the subcategories to have a close look.

The fourth chart which is the map visual shows the average shipping fee charged at each state. Since Olist is an e-commerce platform that focuses its operations in Brazil, the locations that they ship their items to are mostly states in Brazil as seen in the map visual. The size of each bubble shows the average shipping fee charged to deliver items to that state. The larger the bubble, the higher the average shipping fee charged. Another 2 cards are also positioned to the right of the map visual to indicate the sales and shipping fee of the selected field. If there are no fields selected, it would show the sum of sales and shipping fees.

Drill down and analysis

As stated earlier, the charts in the OLAP dashboard can be drilled down. Looking at the figure below, the column chart is now drilled down into the 4 quarters of the year 2017. The other charts like the treemap and map visual also changed their data to match the drill down made in the column chart.

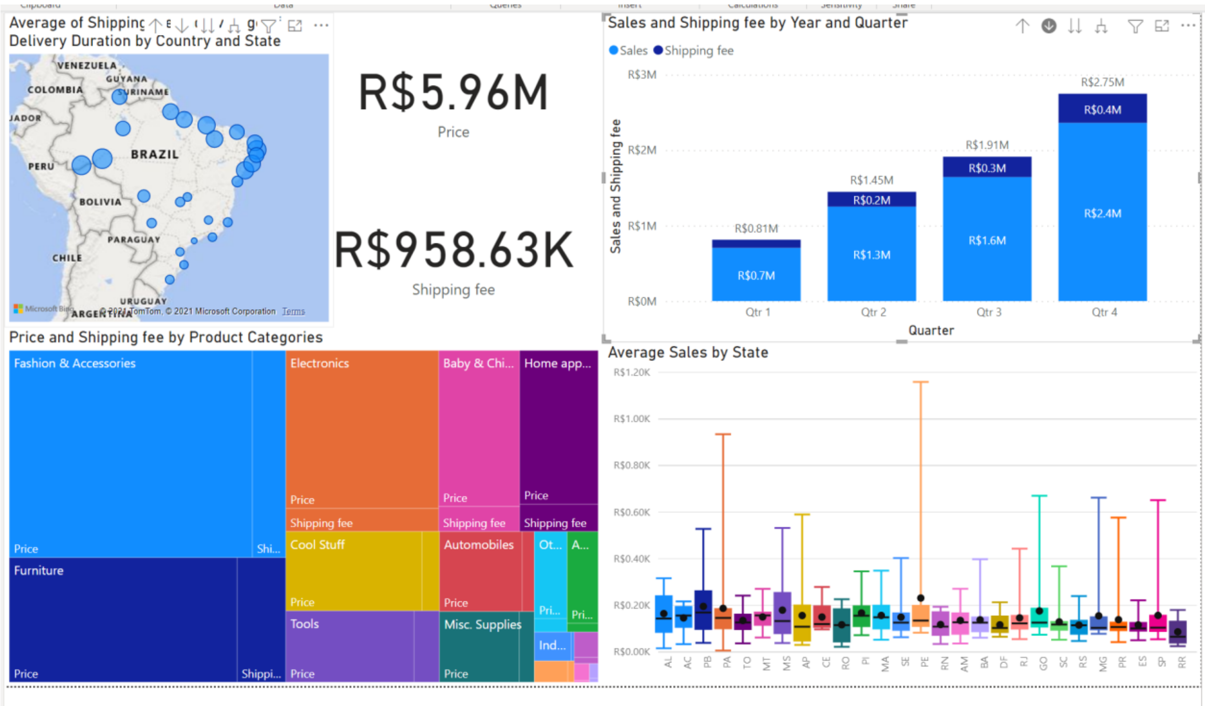


Figure 20 Drilled down OLAP dashboard

As seen in the column chart, the sales generated in each quarter in the year 2017 is also steadily increasing from R$0.7m in the first quarter to R$2.4m in the last quarter. This increase in sales also indicates that there are more and more Brazilians being involved in e-commerce as time goes by. Since more people are purchasing items online, the shipping fees charged will also increase as Olist will need to ship the purchased items to the customers and the shipping fees also increased along with the sales generated in each quarter. This trend clearly shows that Brazil is successful in the adoption of e-commerce technology. Furthermore, looking at the treemap, Fashion & Accessories is still the product category that generated the highest sales followed by Furniture.

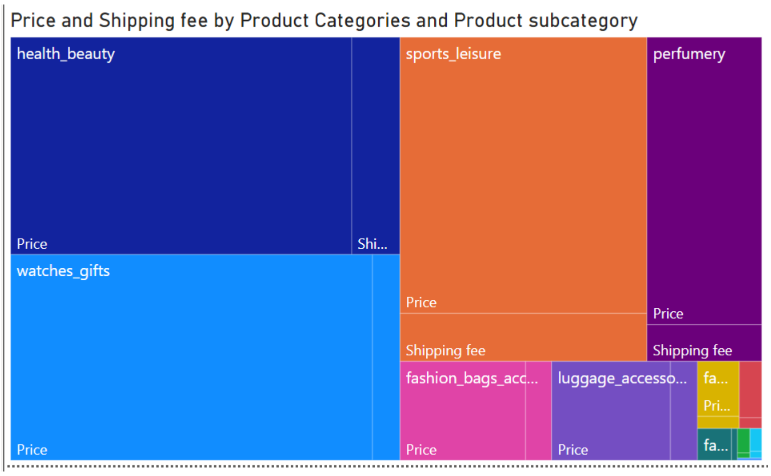


Figure 21 Drilled down treemap

After further drilling down into the product category Fashion & Accessories, it can be seen that the health_beauty subcategory generated the highest sales. This can be interpreted that Brazilians that purchase Fashion & Accessories items from Olist focuses on products that can maintain their health or their beauty, which can explain why most consumers bought health_beuaty products from Olist. Another prominent subcategory is the watches_gifts. This could be because most consumers in Brazil bought these products as a gift to someone during holiday seasons like All Soul's Day. This can be proven by drilling deeper down into the column chart.

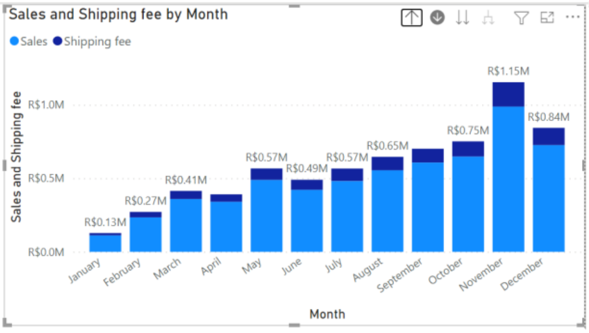


Figure 22 Column chart of months in 2017

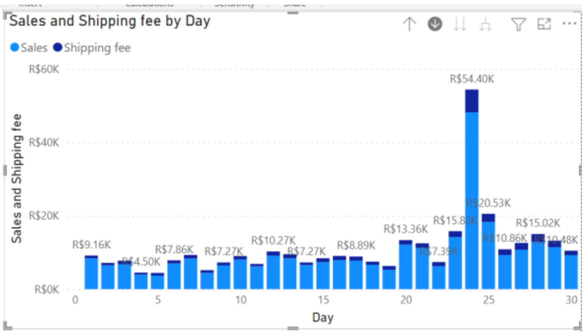


Figure 23 Column chart of November 2017

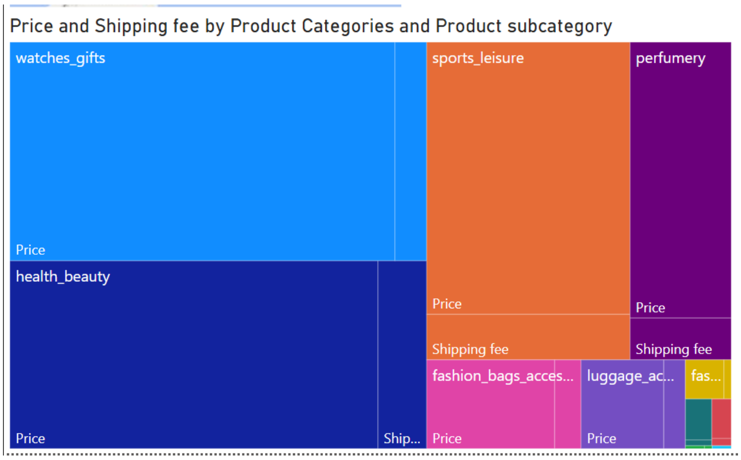


Figure 24 Treemap of November 2017

As shown in the 3 figures above, the sales generated in November 2017 is the highest among all other months in the year 2017. This is because there are numerous national holidays in Brazil in November, like All Soul's Day, Republic Day and so on. Hence, most people would purchase gifts during these holidays to send as gifts to their loved ones.

As for the other charts, in Figure 19, the boxplot has changed to reflect the average sales made by each state in the year 2017. Based on the boxplot it can be seen that the state AL has the highest average sales in 2017 followed closely by state AC. For the map visual on the other hand, most of the states to the north of Brazil has a higher shipping fee compared to the ones toward the south. This might be because Olist warehouses are located more in the south of Brazil, causing the shipping fee charged to deliver products from these warehouses to states in southern Brazil to be much less compared to states in northern Brazil. The higher shipping fee also reflects a higher delivery duration to states located in northern Brazil compared to a lower delivery duration to states located in southern Brazil.

Outlier Analysis

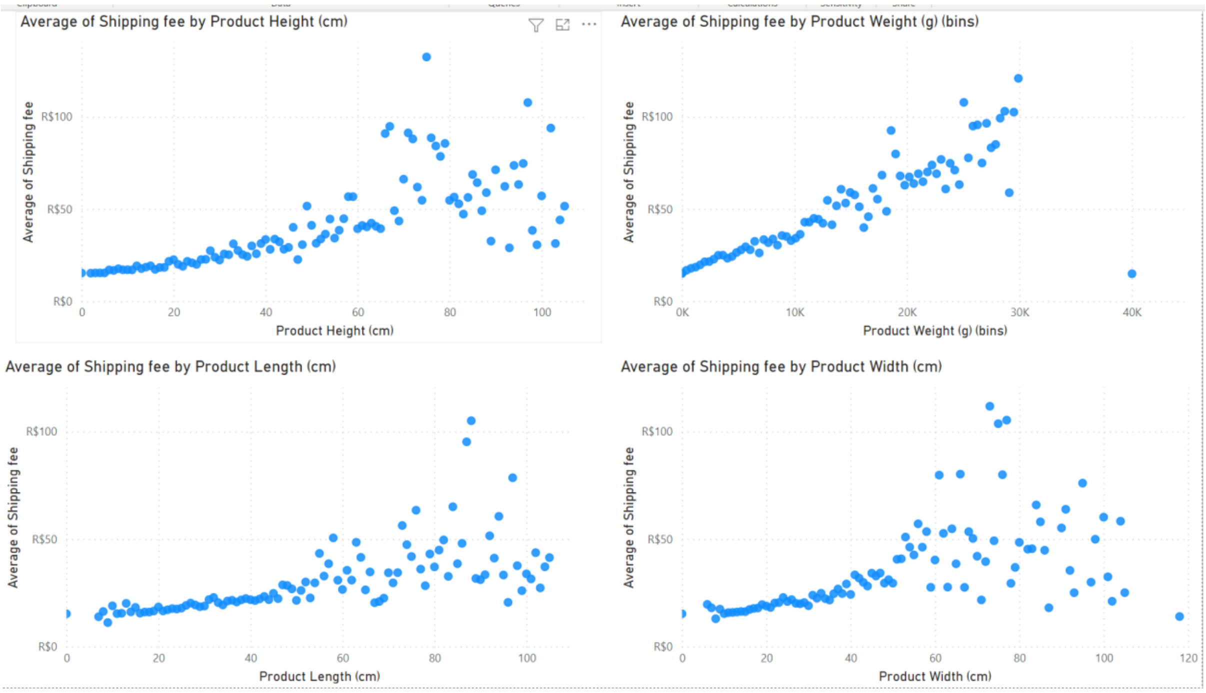


Figure 25 Scatterplots of product dimensions against average shipping fee

The next dashboard shows the 4 different scatterplots of the product's dimensions against the average shipping fee charged on that product. The product dimensions that are used are the product's height, width, length, and weight. The height, length, and width of the product are measured in centimetres (cm) while the weight of the product is measured in grams (g). For the scatterplot of product weight against the average shipping fee, the product weight is grouped into 100 bins. This is because if the product weight is not grouped, there would be too many data points to be plotted out and the graph would be hard to be interpreted by the human eye as shown in Figure 14.

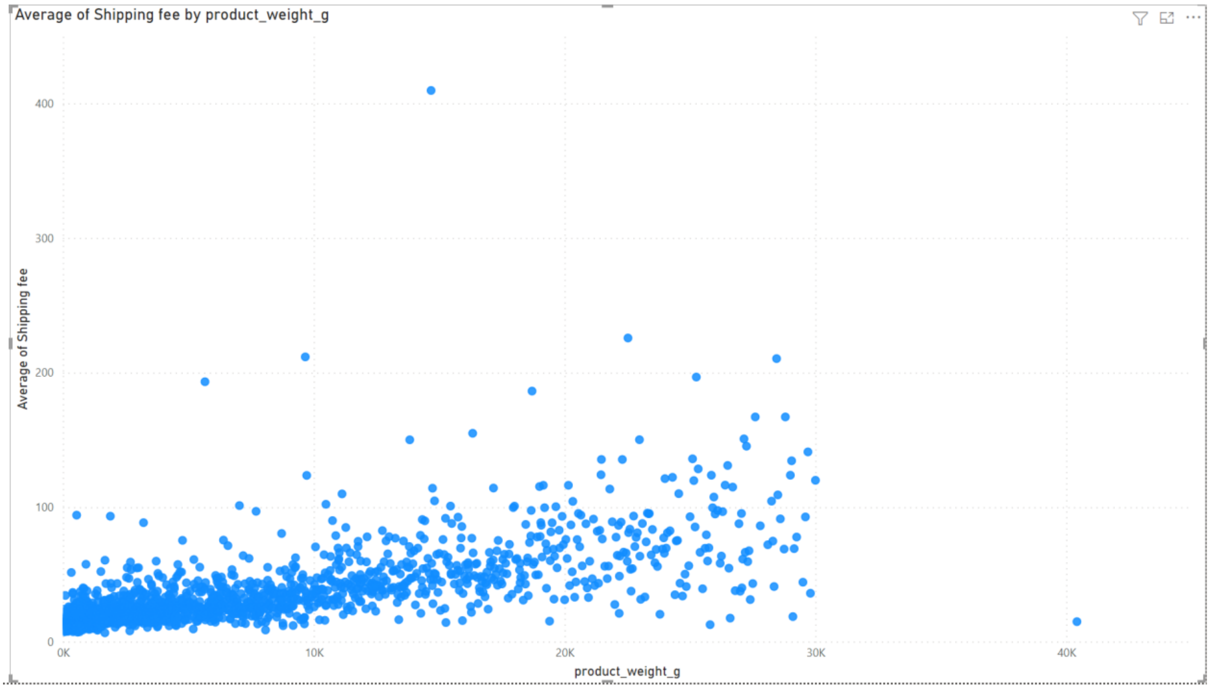


Figure 26 Scatterplot of ungrouped product weight against average shipping fee

As seen in the scatterplots in Figure 13, all the scatterplots show a clear trend of increasing shipping fees as the product dimension increases. The scatterplots about the Product Height (cm), Product Length (cm) and Product Width (cm) shows a slight parabolic curve, which indicates that the relationship between these product dimensions and the shipping fees might not be linear. This will be further investigated in the next section. However, before developing the model, outliers will need to be identified and removed first.

The remaining scatterplot is the Product Weight (g) against the shipping fee. For this scatterplot, the trend is a very clear linear increase but there is one point that does not fit into this linearly increasing trend, which is the data point where Product Weight (g) is equal to 40K. This data point is a clear outlier and must be removed during the development of the linear regression model that will be used to study the effect of the product weight on the shipping fee.

7. Data Modelling

Regression Analysis

Regression analysis is a reliable way of determining variables that have an impact on a topic of interest mathematically. It is a good method to help the user to find out which factor matters the most, which can be ignored and how these factors influence each other (Alchemer, 2021). There are several types of regression analysis, namely Linear Regression, Multiple Linear Regression and Non-linear Regression. For this project, Linear Regression will be used.

Correlation

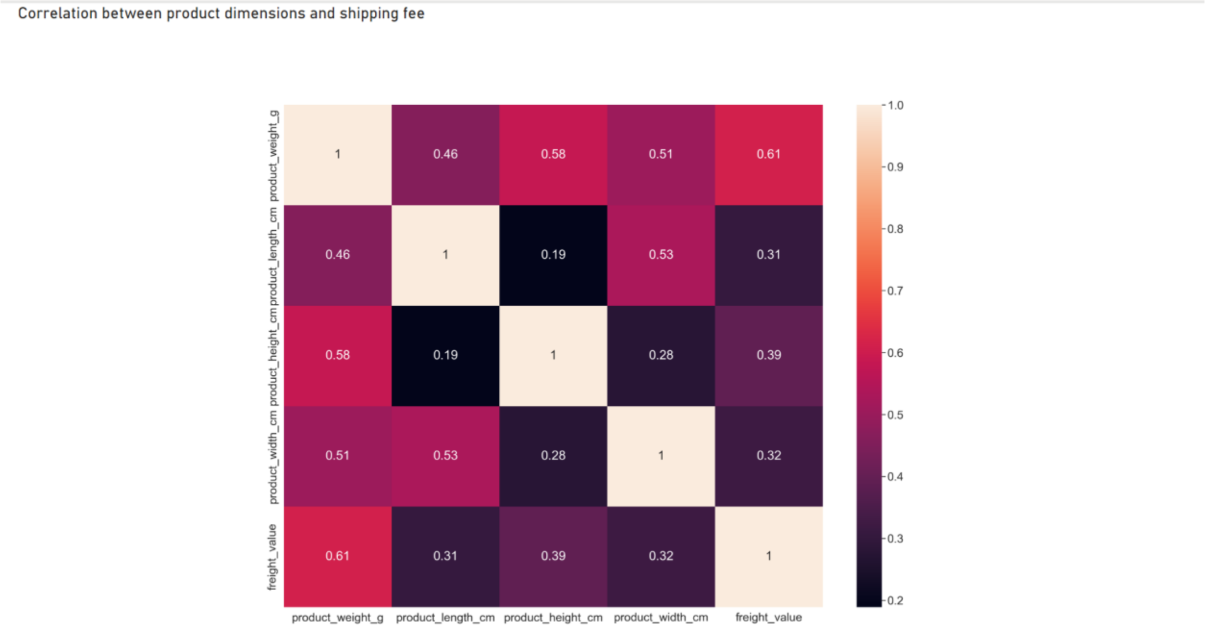

Correlation between product dimensions and shipping fee

Figure 27 Correlation Heatmap

Before starting with the linear regression model, a correlation heatmap is first prepared to analyse which product dimension is the most correlated with the freight value which is also the shipping fee charged for a product. As seen in the heatmap, the freight_value correlates with the product_weight_g the highest, which is 0.61, compared to other product dimensions that correlate lower than 0.35. Based on this observation, it can be said that the product_weight_g affects the freight_value of a product the most and therefore, the linear regression model will focus on this feature the most. However, since the other product dimensions also have a positive correlation with the freight_value, a comparison will be done between the linear regression model of product_weight_g and product_length_cm with the freight_value.

Linear Regression

Linear regression is a very straightforward approach where a quantitative response, Y is predicted based on a predictor X. Mathematically speaking, the formula for a simple linear regression model can be written as (James, Hastie, Witten, & Tibshirani, 2021):

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Y = Dependent variable

β_0 = y-intercept

β_1 = slope

X = Independent variable

ϵ = Random error term

This simple linear regression model is suitable for analyzing data that shows a linear slope, which can be either increasing or decreasing. The figure below shows the scatterplot of the product weight that has been grouped into 100 bins for data smoothing.

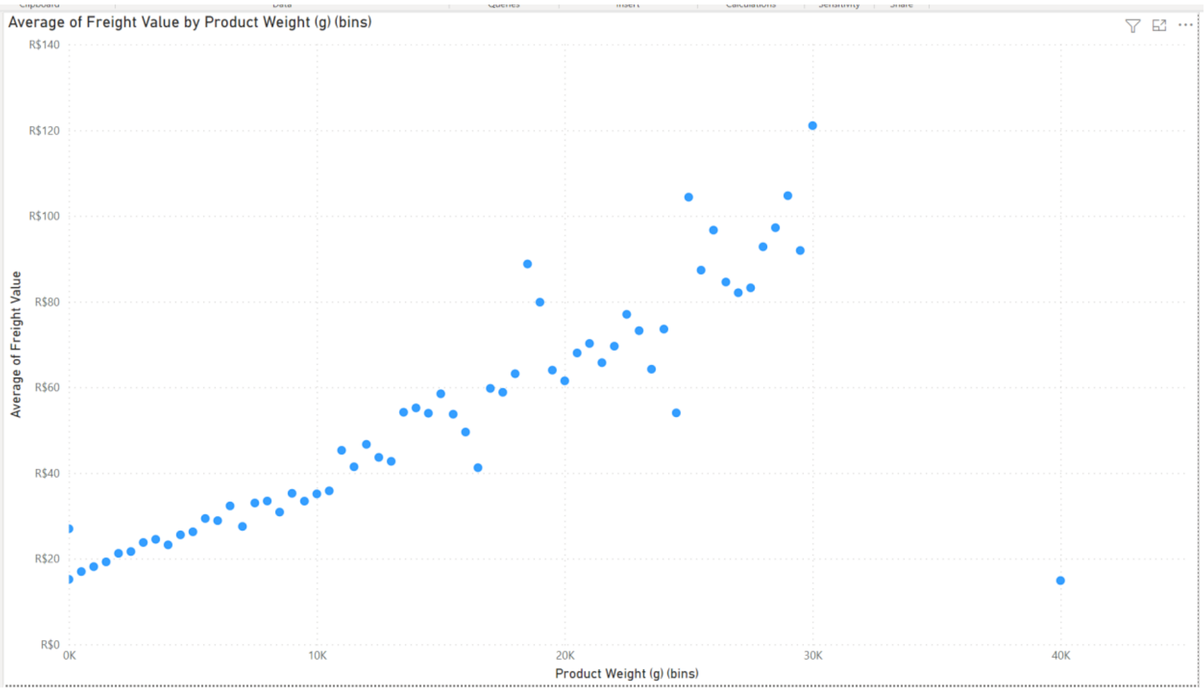
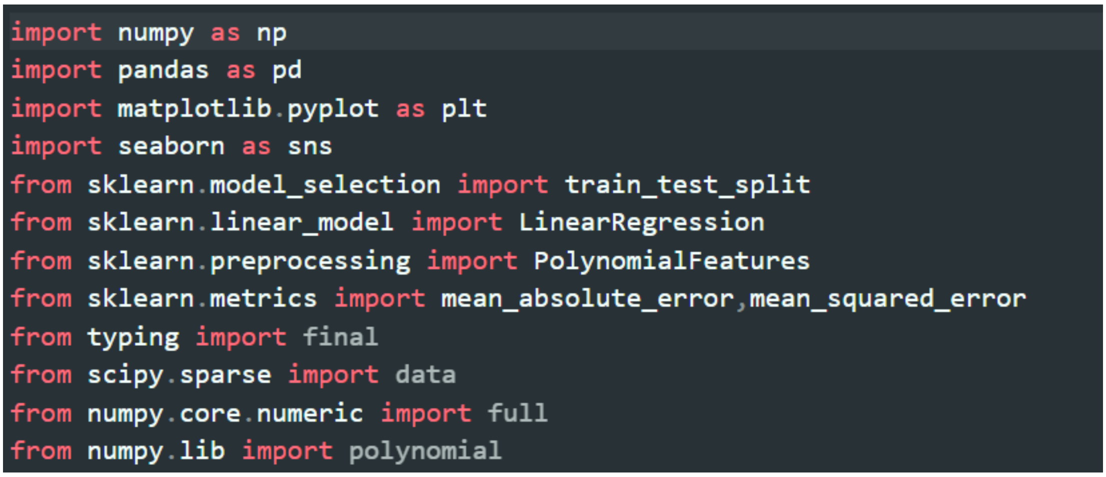


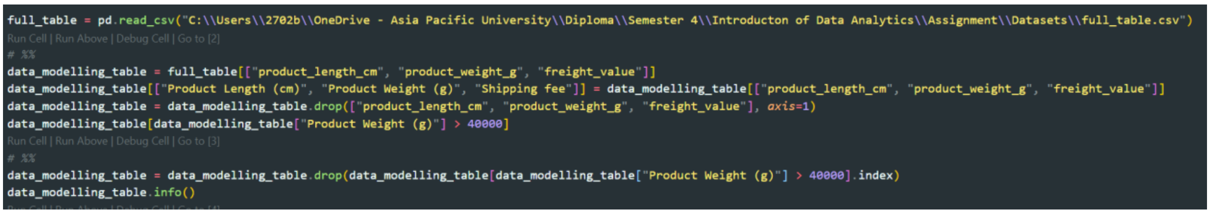
Figure 28 Scatterplot of Freight Value against Product Weight

Before developing the Linear Regression model, all the necessary Python libraries need to be imported first. Libraries like NumPy, pandas and sklearn are imported into the Python environment to prepare the data and carry out linear regression. Once the libraries are imported, the dataset that contains the columns needed for training and testing the linear model will be imported into the environment using pandas.read_csv() method. After importing the dataset, columns that are not needed to build the linear regression model will be dropped using pandas.drop() method. Furthermore, as stated previously, some outlier data needs to be removed. Based on the scatterplots in Figure 24, data that has more than 40000g needs to be removed from the dataset before developing the Linear Regression model because the outliers might affect the accuracy and output of the model. This step can be done through conditional filtering to find the specific columns and then drop them. These steps are shown in Figure 29 below.



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import mean_absolute_error,mean_squared_error
from typing import final
from scipy.sparse import data
from numpy.core.numeric import full
from numpy.lib import polynomial
```

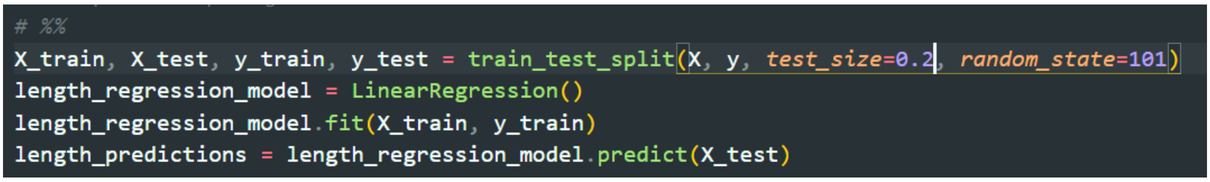
Figure 29 Importing the necessary libraries



```
full_table = pd.read_csv("C:\\Users\\2702b\\OneDrive - Asia Pacific University\\Diploma\\Semester 4\\Introducton of Data Analytics\\Assignment\\Datasets\\full_table.csv")
Run Cell | Run Above | Debug Cell | Go to [2]
# %%
data_modelling_table = full_table[["product_length_cm", "product_weight_g", "freight_value"]]
data_modelling_table[["Product Length (cm)", "Product Weight (g)", "Shipping fee"]] = data_modelling_table[["product_length_cm", "product_weight_g", "freight_value"]]
data_modelling_table = data_modelling_table.drop(["product_length_cm", "product_weight_g", "freight_value"], axis=1)
data_modelling_table[data_modelling_table["Product Weight (g)"] > 40000]
Run Cell | Run Above | Debug Cell | Go to [3]
# %%
data_modelling_table = data_modelling_table.drop(data_modelling_table[data_modelling_table["Product Weight (g)"] > 40000].index)
data_modelling_table.info()
```

Figure 30 Importing, cleaning and dropping unnecessary columns and outliers

Once the libraries and datasets are ready, the training set and testing set can be prepared to train the Linear regression model. Although 2 different Linear regression models with different features (product weight and product length) will be prepared separately, the code used to develop the model is mostly the same, only with some minor changes on the column names used. The first model that will be developed is between the product length and shipping fee. The first step is to only include the features that are needed, which in this case is the product length. Therefore, the product weight will be dropped from the first model. The unnecessary column will be dropped when splitting the dataset into features and labels step, which will be labelled as X and y respectively. Once the features and labels have been assigned, the sklearn.model_selection.train_test_split() function will be used to split the feature and labels into 80% for the training dataset and 20% for the testing dataset. Once this has been done, the Linear Regression model is instantiated and fitted to the training data. After fitting it, the model will be used to predict based on the test labels, which is the X_test data. The code for these steps is shown below.



```
# %%
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=101)
length_regression_model = LinearRegression()
length_regression_model.fit(X_train, y_train)
length_predictions = length_regression_model.predict(X_test)
```

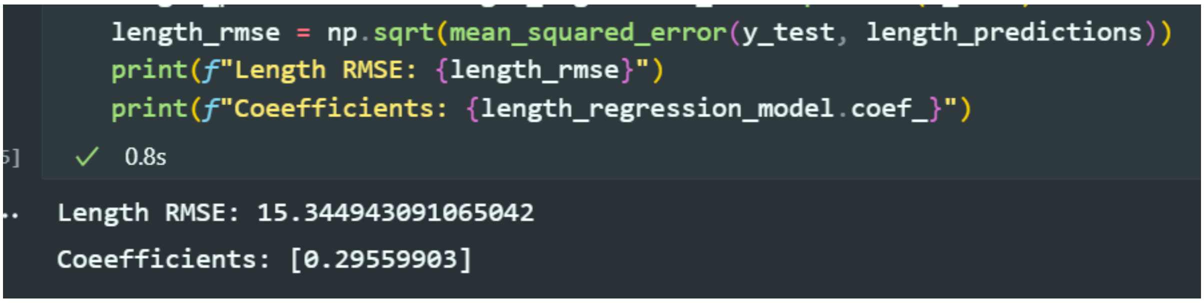
Figure 31 Code used to train the model

Once the model is trained and used to carry out its first prediction, the model is evaluated by calculating the root mean squared error (RMSE). RMSE is one of the most commonly used model evaluation methods to ensure the quality of a prediction (c3.ai, 2021). The formula of RMSE is shown as below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_\imath})^2}$$

where y_i = is the ith measurement, and y^$_i$ is the corresponding prediction.

The RMSE of the Linear Regression model developed using the product length and shipping fee is 15.34. Since the RMSE is lower than the mean value of the shipping fee, which is 19.98, the Linear Regression model developed is acceptable. The RMSE of this model is calculated by utilizing sklearn.metrics.mean_squared_error() function and np.sqrt() function. The coefficient of this model is also calculated by Python itself, which is 0.2956 for the Linear regression between product length and shipping fee. The coefficient indicates that for every 1 cm increase in product length, the shipping fee would increase by R$0.2956. The code and output are shown as below:



```
    length_rmse = np.sqrt(mean_squared_error(y_test, length_predictions))
    print(f"Length RMSE: {length_rmse}")
    print(f"Coeefficients: {length_regression_model.coef_}")
```

```
Length RMSE: 15.344943091065042
Coeefficients: [0.29559903]
```

Figure 32 Code and output for calculation of RMSE and coefficients

Since the model is acceptable, the model will be deployed. For deployment, the finalized model will be saved using the joblib.dump function to save the regression model as a joblib file and future users can use joblib.load function to utilize the linear regression model for their own use cases.

The Linear regression model development between product weight and the shipping fee is the same but the column being dropped in this case will be the product length instead. The other steps are similar to the code used to develop the previous Linear Regression model. As for the RMSE and coefficient for this Linear regression on the product weight and shipping fee, they are as follows:

1. RMSE = 12.55
2. Coefficients = 0.002576

Similarly, the RMSE is still below the mean value of the shipping fee. However, the coefficient of the product weight is significantly lower than the coefficient of product length. However, this coefficient is based on product weight that is measured in grams. If the product weight is converted into kilograms, the coefficient will be increased to 2.576 whereas each 1-kilogram increase in the product weight will cause the shipping fee to increase by R$2.57. The graphs in the next section are plotted using the predicted values from the linear regression models developed in this chapter.

Graphs

Product Length

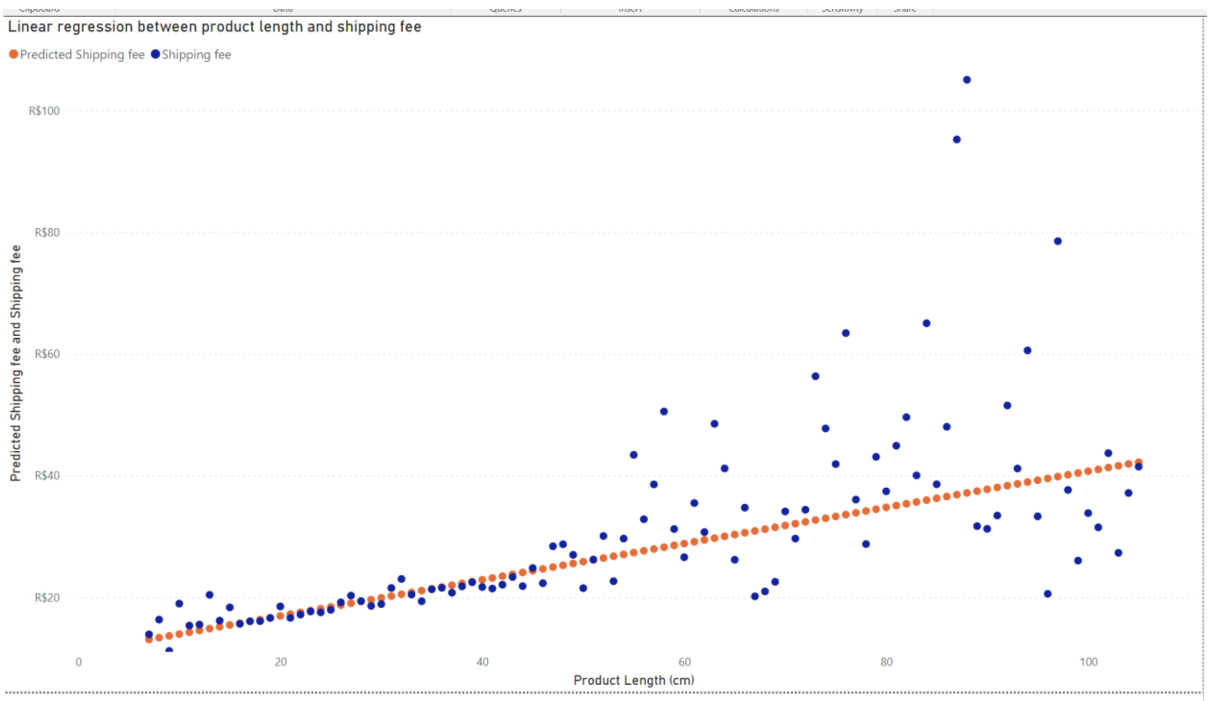


Figure 33 Scatterplot of average shipping fee and average predicted values against product height (cm)

Product Weight

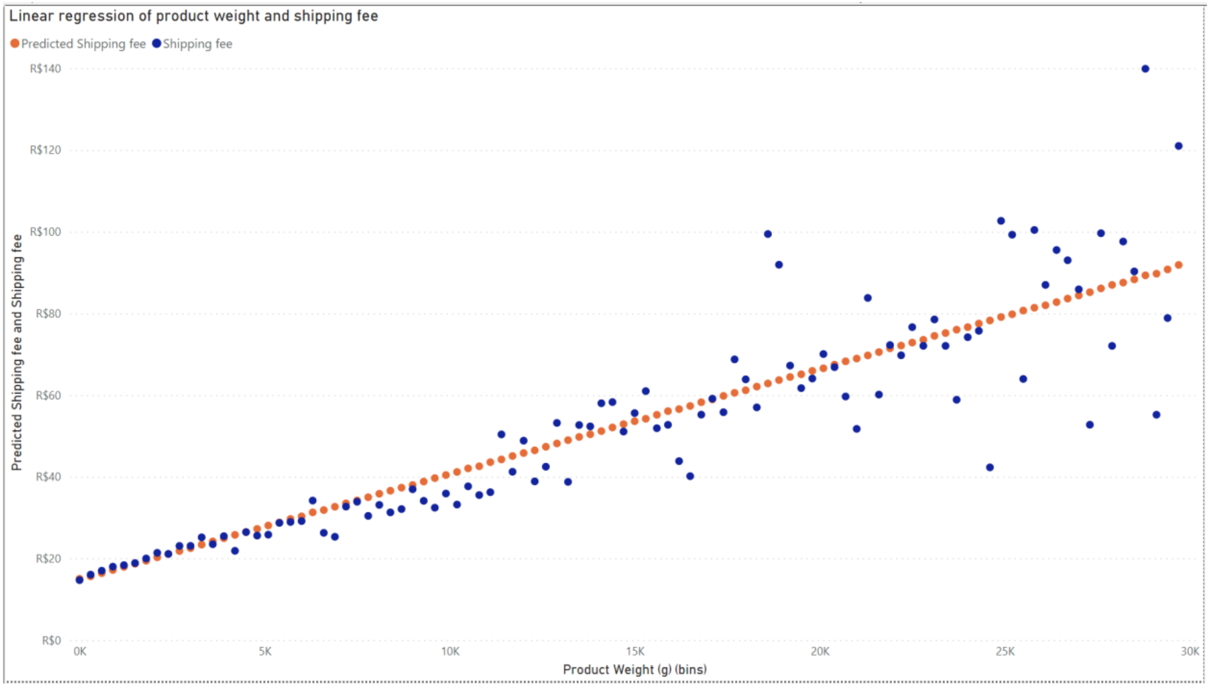


Figure 34 Scatterplot of average shipping fee and average predicted values against product weight (g)

8. Conclusion & Recommendations

After carrying out this project, it can be seen that e-commerce sales in Brazil are truly increasing at a constant rate. Therefore, interested investors should use this opportunity to invest in the e-commerce industry in Brazil before the competition saturates. Aside from that, from the analytical dashboard shown, the average delivery duration to the northern states is comparatively higher than the southern states in Brazil. From this visualization, sellers in Brazil could focus on opening more warehouses in the northern area of Brazil to shorten the delivery duration to increase sales, customer retention and customer satisfaction. Furthermore, it can be concluded that the higher the product dimensions, the higher the freight value, sellers will need to take note of this and plan their delivery expenses carefully to rake in the highest profits.

Data privacy and security recommendations

During the pandemic, e-commerce platforms have a significant siege in user traffic. The E-commerce public dataset in Brazil has been chosen as the main topic to be conducted on analysis. Data security and privacy are the main concerns when it comes to platforms that require a user to fill in personal details. Hence, the article that is used as a reference is titled “Brazilian marketplace integrator Hariexpress exposed 1.75 billion records”. Within the article, it is mentioned that the company responsible for the data leak are refraining from taking steps to secure the data, exposing personal data and available online for the public for more than a month. Since the data was leaked to the public for some time, it is hard for the company to take action to retrieve the data and fix these issues. The consequence faced by the company is the loss of trust of the stakeholders (Waqas, 2021).

Regarding the scope analyzed in the dataset, sensitive data such as customer billing information and address should be disclosed and generalized before being available online to the public. The responsible company should take preventive measurements of users’ personal information by including authentication and encryption service in data protection. Moreover, the company should also notify the customers on what data is gathered from them and what the data is used for. Failure to compromise on respecting users’ information will harm the company reputation, public fear causing distress and facing a lawsuit.

9. Personal reflection report & Workload matrix

Personal Reflection Report

Through this assignment, I have been able to learn new techniques related to data analytics. From the usage of Microsoft Power BI to import datasets to cleaning those datasets for visualization purposes and further analysis by using predictive models like linear regression. From this assignment, I was also able to understand the process of data analytics and the methodologies involved in it like the CRISP-DM used for this assignment. By implementing the methodologies learned, the process of data analytics can be done more systematically and efficiently.

Workload matrix

Name	Task
Bryan Hor Jin Hao	Scope: Sales and Shipping fee Model: Predictive model (Linear Regression)
Chia Wen Xuen	Scope: Popularity of Product Sold Model: Predictive model (Time-Series Forecast)
Ngiam Jie-Hao	Scope: Payment Type Model: Predictive model (Decision Tree)
Tin Eugene	Scope: Bad Reviews Model: Predictive model (Logistic Regression)

10. References

Alchemer. (8 6, 2021). *What is Regression Analysis and Why Should I Use it?* Retrieved from https://www.alchemer.com/resources/blog/regression-analysis/

c3.ai. (12 11, 2021). *Root Mean Square Error (RMSE)*. Retrieved from Data Science: https://c3.ai/glossary/data-science/root-mean-square-error-rmse/

Ingle, S., Bhalekar, P., & Pathak, K. (2014). The Study of E-commerce. *Asian Journal of Computer Science and Information Technology*, 25-27.

International Trade Administration. (22 1, 2021). *eCommerce*. Retrieved from Brazil - Country Commercial Guide: https://www.trade.gov/country-commercial-guides/brazil-ecommerce

James, G., Hastie, T., Witten, D., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R.* Springer Science+Business Media, LLC.

Lima, G. A. (2017). National report on e-commerce development in Brazil. *DEPARTMENT OF POLICY, RESEARCH AND STATISTICS WORKING PAPER 14/2017*, 48.

scikit learn. (4 11, 2021). *sklearn.preprocessing.PolynomialFeatures*. Retrieved from scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

Smart Vision Europe. (17 6, 2020). *CRISP-DM Methodology*. Retrieved from What is the CRISP-DM methodology?: https://www.sv-europe.com/crisp-dm-methodology/

Waqas. (13 10, 2021). *Brazilian Marketplace Integrator Hariexpress exposed 1.75 billion records*. Retrieved from Hackread: https://www.hackread.com/brazilian-marketplace-integrator-hariexpress-records/