

# DSAI Module 2 Final Project Presentation

## Solution Implementation



29 Mar 2025

Group Members  
Austin, Jason, Roy, Tricia,  
Vasanthi

# OVERVIEW OF BRAZILIAN E-COMMERCE

## In 2015...

- Brazil is the world's ninth-largest retail e-commerce market and the only Latin American country in the global top 10, with 80 million digital shoppers in 2015.
- E-commerce retail sales are estimated to hit \$22.5 billion this year and are expected to grow at an 11% CAGR from 2014 to 2019.
- More than half of Brazilians have Internet access and more than 60% of that group connects via smartphone.
- Brazil saw an 87% expansion in median household income from 2003 to 2013, leading to a near doubling of the middle class and spurring regional and global retailers to enter the market.

# Agenda - Build a **Data-Driven** e-Commerce System

---

1. Project Overview
2. Technical and Business Objectives
3. Data Engineering System Design
4. Data Exploration and Understanding
5. Star Schema Design
6. Data Quality Testing Design
7. Pipeline Orchestration
8. Data Visualisation

# 1. Project Overview

Dataset (*source: Kaggle*)

**Brazilian Ecommerce Public Dataset by Olis**

Dataset (*overview*)

100k product orders from **2016 to 2018** marketplaces in Brazil

## 2. Technical and Business Objectives

### Technical Objective

Design a end-to-end **data pipeline** to ingest data from **Kaggle** into **BigQuery**

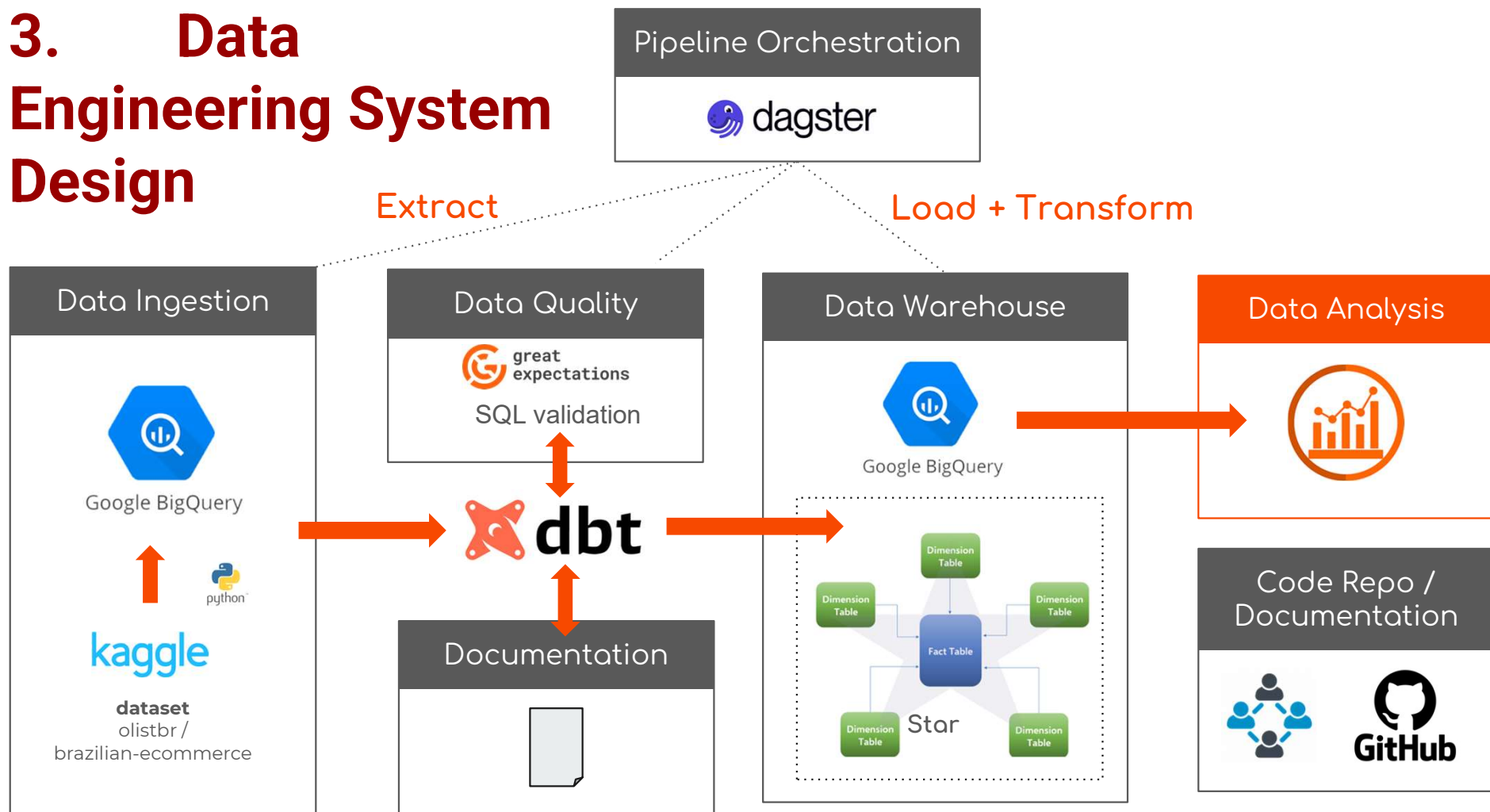
This pipeline will **automate** data cleaning, preprocessing, and quality assurance to ensure accurate and up-to-date data for analysis.

### Business Objectives

**Data-driven** insights into key **business metrics**

Such as: Total **sales** and **product volume** by **Sellers**

### 3. Data Engineering System Design



## 4. Data Exploration and Understanding

---

### Dataset Scope

- 100k orders (2016-2018)
- Real commercial data from multiple marketplaces
- Anonymized Brazilian e-commerce transactions

### Key Components

- 8 Core Tables: Orders, Items, Products, Customers, Sellers, Payments, Reviews, Geolocation
- Nationwide coverage across Brazil
- Complete order journey tracking

### Business Value

- Sales & Customer Behavior Analysis
- Logistics Performance Metrics
- Payment Pattern Insights
- Geographic Distribution Study

### Key Features

- Order status tracking
- Multiple payment methods
- Product categorization
- Delivery performance

## 4. Data Exploration and Understanding

### Data Issue 1

**Different variations** in spelling of seller and customer city

- Massive cleanup, not familiar with Brazilian Cities

### Data Issue 2

**Unknown category name** for over 600 products

- Categorised under "Others"

### Data Issue 3

**Missing values** in various timestamp in orders data

- Due to different stages of the delivery
- *created -> approved -> processing -> invoiced -> shipped-> delivered*

seller_city
sao paulo sp
sao paulo
sao paulo / sao paulo
sao paulop
sao paulo - sp
...
sao paulo
sao paulo
sao paulo
sao paulo
sao paulo

Different variations  
of same city





## 5. Data Quality Testing Design

```
-- brazilcom
|-- dbt_project.yml
|-- tests
|   |-- dbt_test_order_items_price_check.sql
|-- models
|   |-- dimensions
|   |   |-- dim_customer.sql
|   |   |-- dim_geolocation.sql
|   |   |-- dim_order_items.sql
|   |   |-- dim_orders.sql
|   |   |-- dim_products.sql
|   |   |-- dim_sellers.sql
|   |   |-- `-- sources.yml
|   |-- facts
|   |   |-- fact_geolocation_sales.sql
|   |   |-- fact_seller_performance.sql
|   |   |-- fact_top_product_per_seller_geolocation.sql
|   |   |-- fact_top_selling_products.sql
|   |   |-- `-- sources.yml
|   |-- facts.yml
|   |-- facts_orders.sql
|   |-- `-- raw_data
|   |-- `-- sources.yml
|-- myELT.ipynb
|-- profiles.yml
|-- seeds
|-- `-- properties.yml
```

order_items			
Schema			
Field name	Type	Mode	
order_id	STRING	NULLABLE	
order_item_id	INTEGER	NULLABLE	
product_id	STRING	NULLABLE	
seller_id	STRING	NULLABLE	
shipping_limit_date	TIMESTAMP	NULLABLE	
price	FLOAT	NULLABLE	
freight_value	FLOAT	NULLABLE	

```
dbt_test_order_items_price_check.sql
SELECT *
FROM `brazilcom.order_items`
WHERE NOT (price BETWEEN 0.0 AND 10000.0)
```

Run the test under tests folder

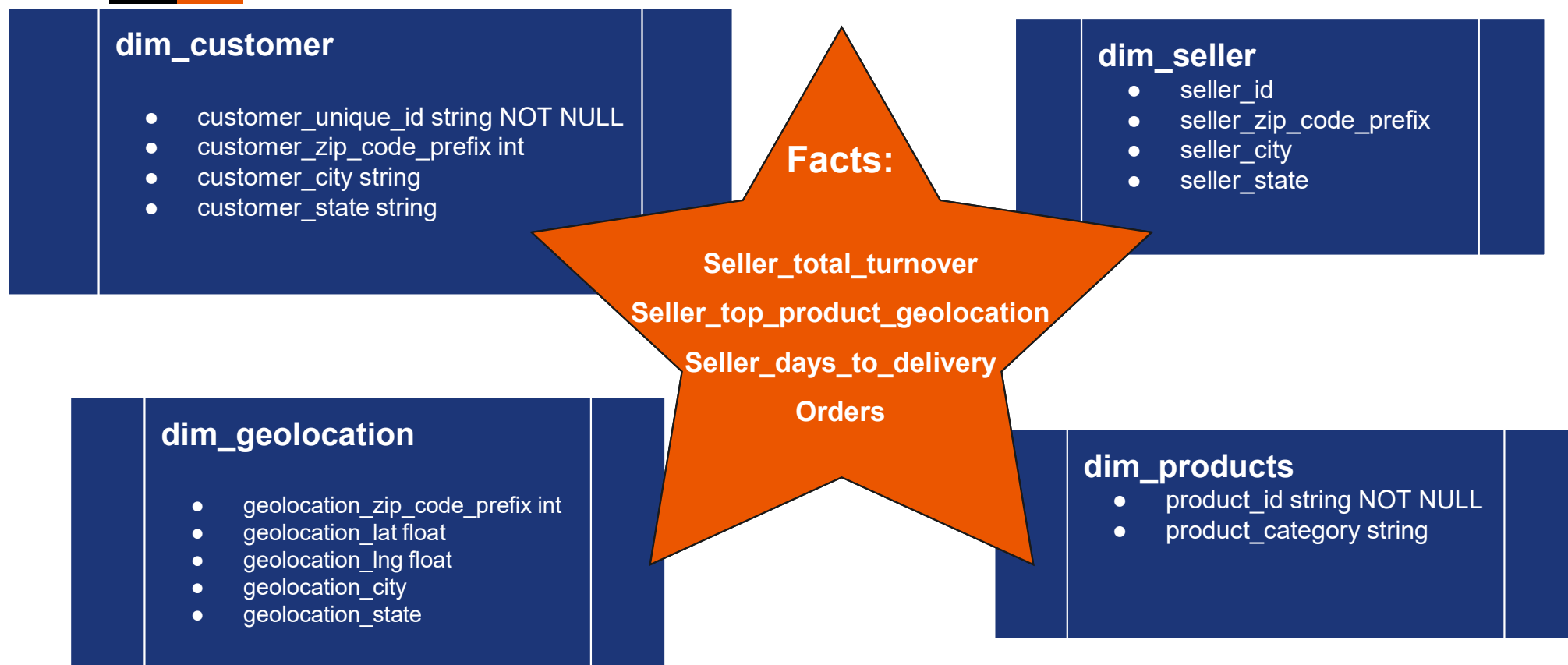
```
% dbt test --select tests/dbt_test_order_items_price_check.sql
```

### Test Results

```
Item price between 0.0 to 10000.0
05:13:27 1 of 21 START test dbt_test_order_items_price_check ..... [RUN]
05:13:28 1 of 21 PASS dbt_test_order_items_price_check ..... [PASS in 0.87s]

Item price between 0.0 to 1000.0 (844 items failed the price check)
05:16:24 1 of 21 START test dbt_test_order_items_price_check ..... [RUN]
05:16:25 1 of 21 FAIL 844 dbt_test_order_items_price_check ..... [FAIL 844 in 0.82s]
```

## 6. Star Schema Design



## 7.1 Dagster-DBT project setup

### Prerequisites

1. Use `dwh` environment: `conda activate dwh`
2. Install required package: `conda install kagglehub`
3. Install dagster-dbt integration: `pip install dagster-dbt`
4. Initiate `dbt` project: `dbt init brazilecom`
5. Set up dagster-dbt project: `dagster-dbt project scaffold --project-name project_dagster --dbt-project-dir path/to/dbt/project`
6. Go to folder `brazilecom` and run the codes below to test `dbt` project setup.
7. For data analysis, need to install the following packages if not done before:
  - 5.1. `google-cloud-bigquery-storage`
  - 5.2. `google-cloud-bigquery`
  - 5.3. `dagster-gcp`
8. Auth with service account. `gcloud auth activate-service-account --key-file D:\Projects\ntu\course\dsai-module-2-group-1\brazilecom\dsai-module-2-pro`

## 7.2 Pipeline Orchestration - Dagster Asset Group

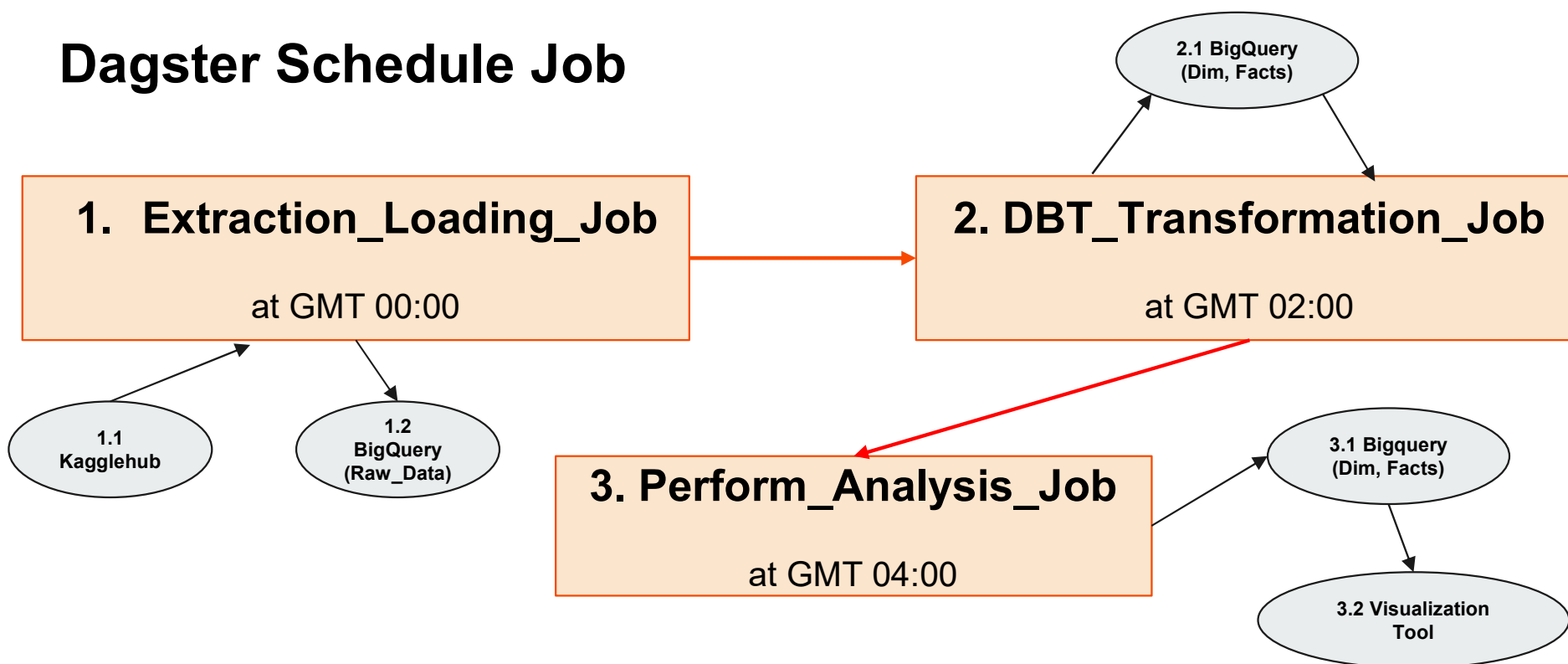


### Asset Groups:

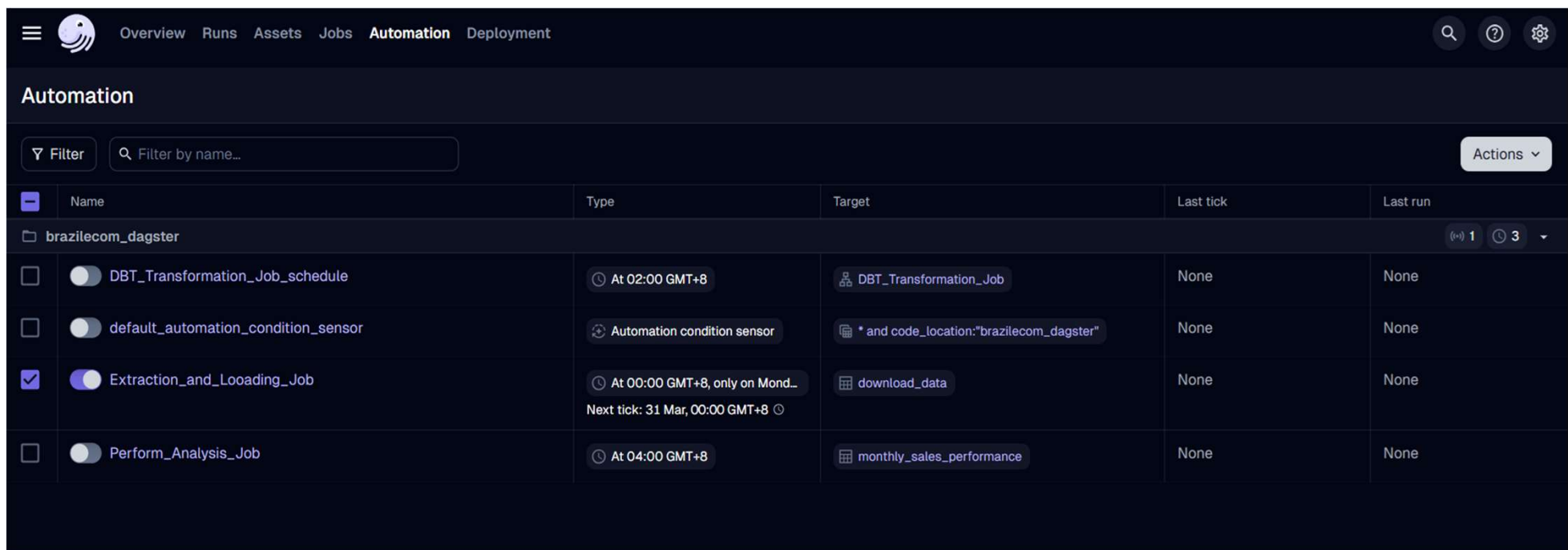
- Raw\_data Group - Assets representing raw data tables (Python Type)
- Upstream Group - Asset to extract data from website and populate into Bigquery under Raw\_data group (Python Asset)
- Star Group - Assets representing STAR dimension tables (DBT Type)
- Facts Group - Assets representing FACT tables (DBT Type)
- Analysis Group - Assets representing data marts for analysis (DBT Type)

## 7.3 Pipeline Orchestration (Pipeline Automation)

### Dagster Schedule Job



## 7. Pipeline Orchestration (Pipeline Automation Schedule Jobs)

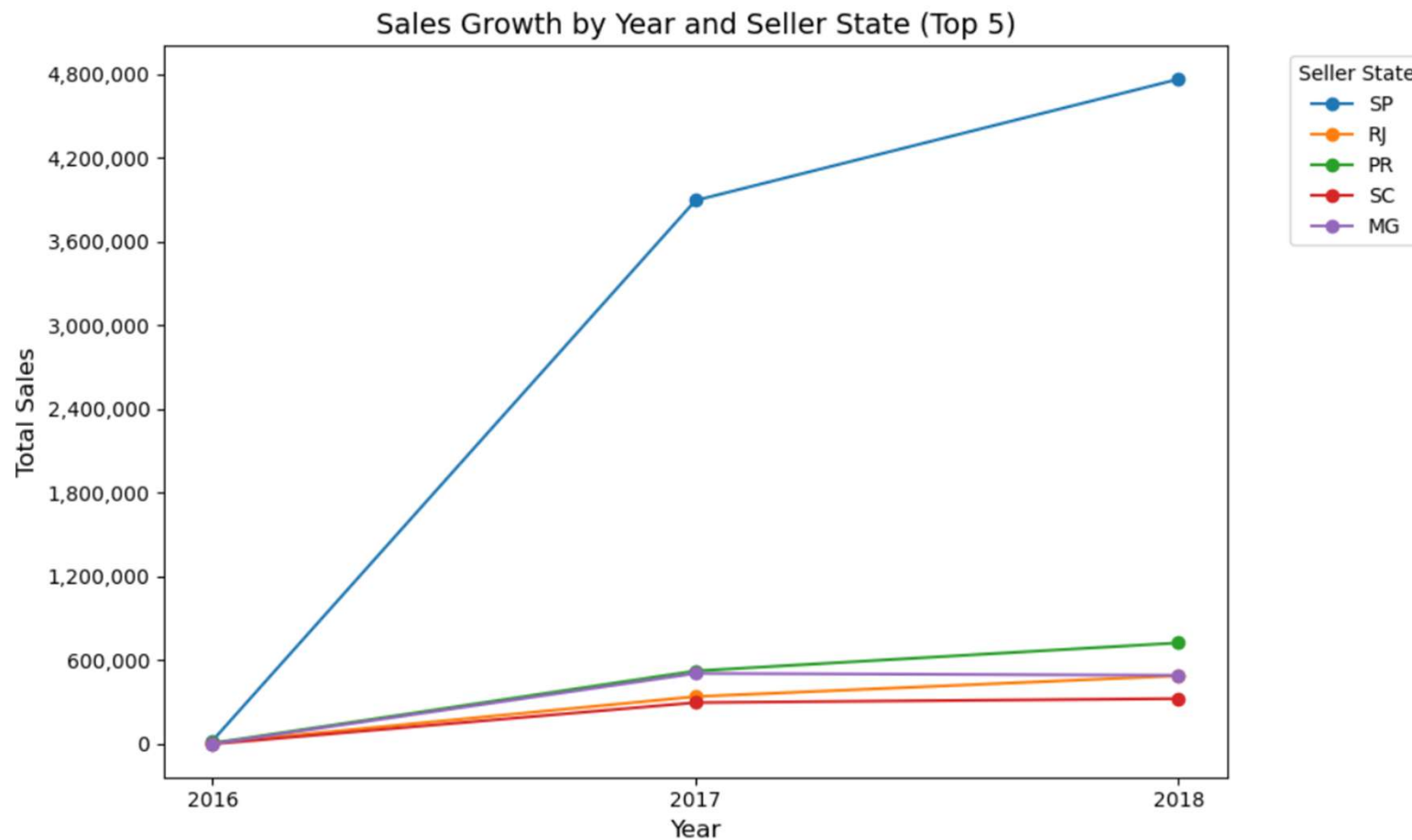


The screenshot displays the Dagster Automation interface. At the top, there's a navigation bar with tabs: Overview, Runs, Assets, Jobs, Automation (selected), and Deployment. Below the navigation bar, the 'Automation' section is active. It features a filter input field with the placeholder 'Filter by name...' and an 'Actions' dropdown menu. The main content is a table listing automation jobs for the 'brazilecom\_dagster' pipeline. The table has columns for Name, Type, Target, Last tick, and Last run. Four jobs are listed: 'DBT\_Transformation\_Job\_schedule', 'default\_automation\_condition\_sensor', 'Extraction\_and\_Loading\_Job' (which is selected with a checkbox), and 'Perform\_Analysis\_Job'. Each job entry includes a toggle switch, a clock icon indicating the schedule, a target icon and name, and 'None' for both 'Last tick' and 'Last run'.

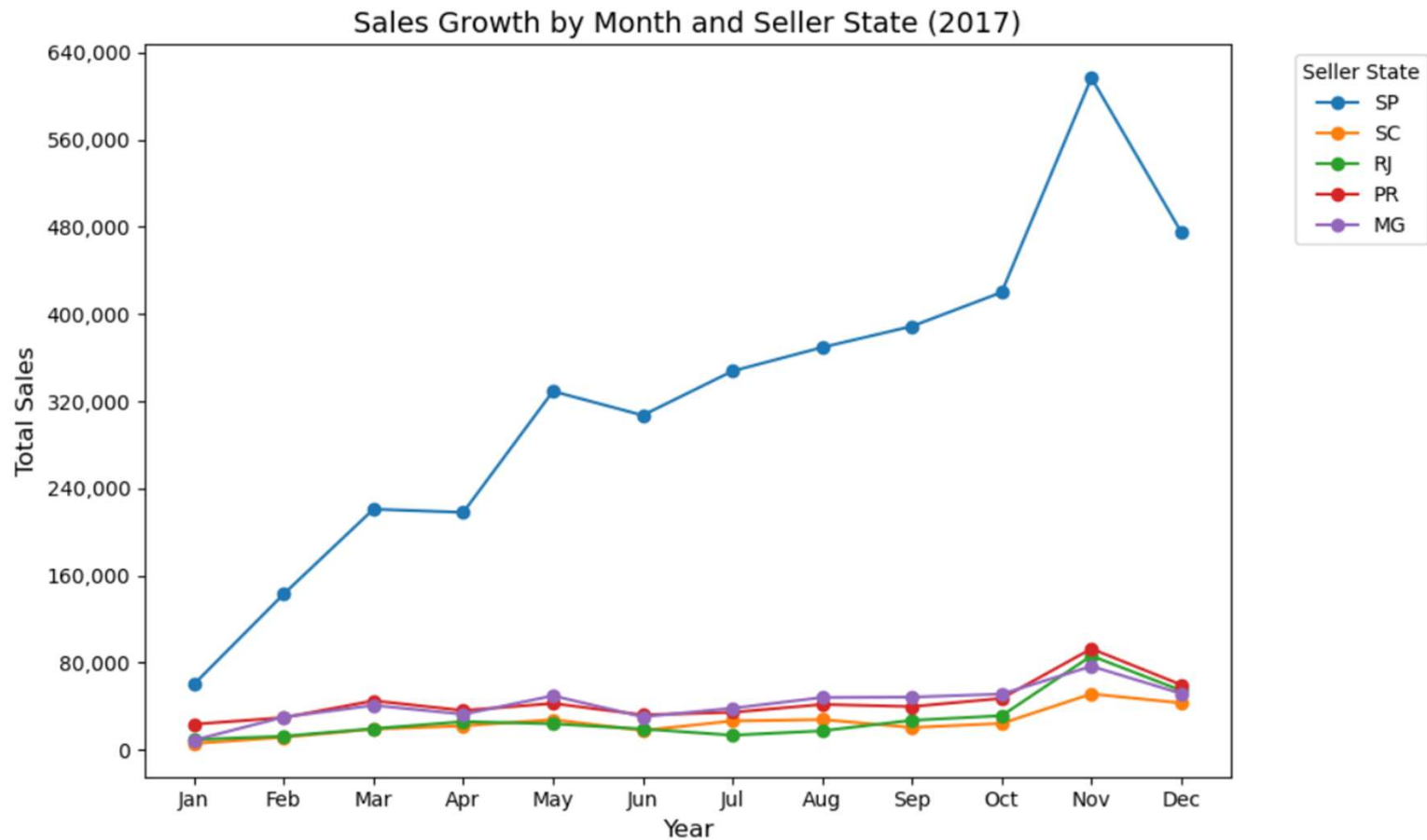
	Name	Type	Target	Last tick	Last run
	brazilecom_dagster				
<input type="checkbox"/>	<input type="checkbox"/> DBT_Transformation_Job_schedule	⌚ At 02:00 GMT+8	📦 DBT_Transformation_Job	None	None
<input type="checkbox"/>	<input type="checkbox"/> default_automation_condition_sensor	⚙️ Automation condition sensor	📄 * and code_location:"brazilecom_dagster"	None	None
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Extraction_and_Loading_Job	⌚ At 00:00 GMT+8, only on Mond... Next tick: 31 Mar, 00:00 GMT+8 ⌚	📄 download_data	None	None
<input type="checkbox"/>	<input type="checkbox"/> Perform_Analysis_Job	⌚ At 04:00 GMT+8	📄 monthly_sales_performance	None	None

Demo at [localhost:3000](http://localhost:3000)

## 8. Data Visualisation

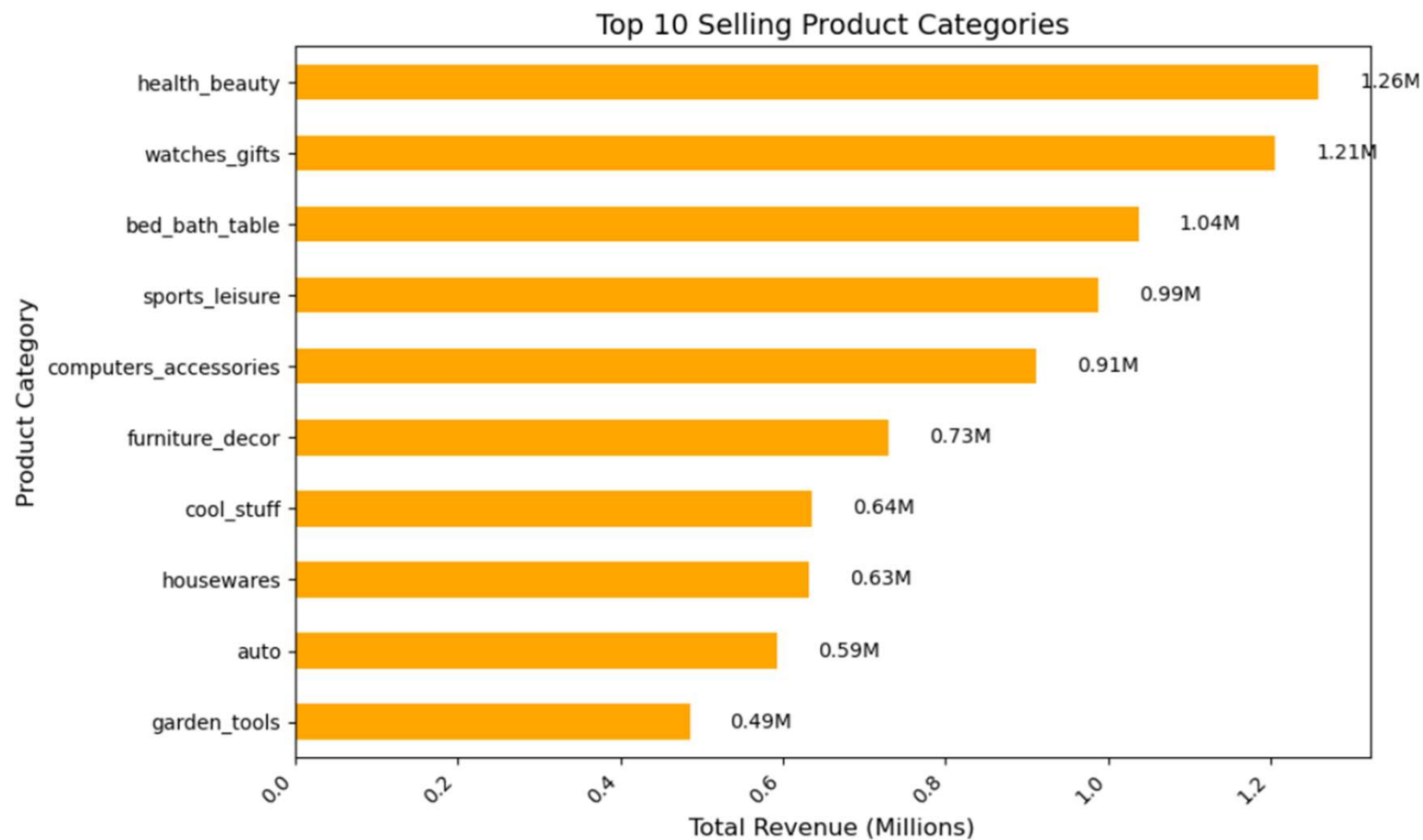


## 8. Data Visualisation





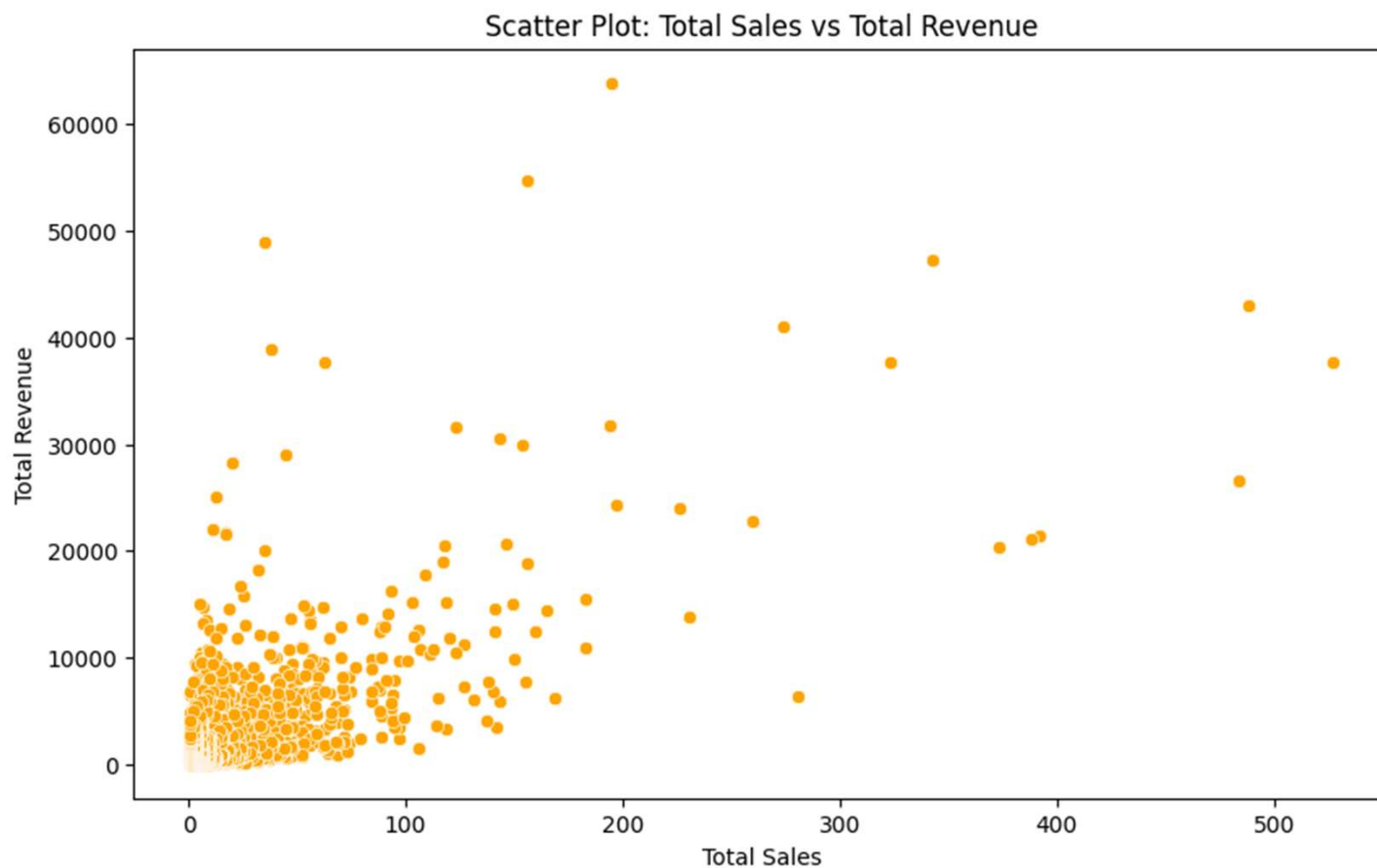
## 8. Data Visualisation



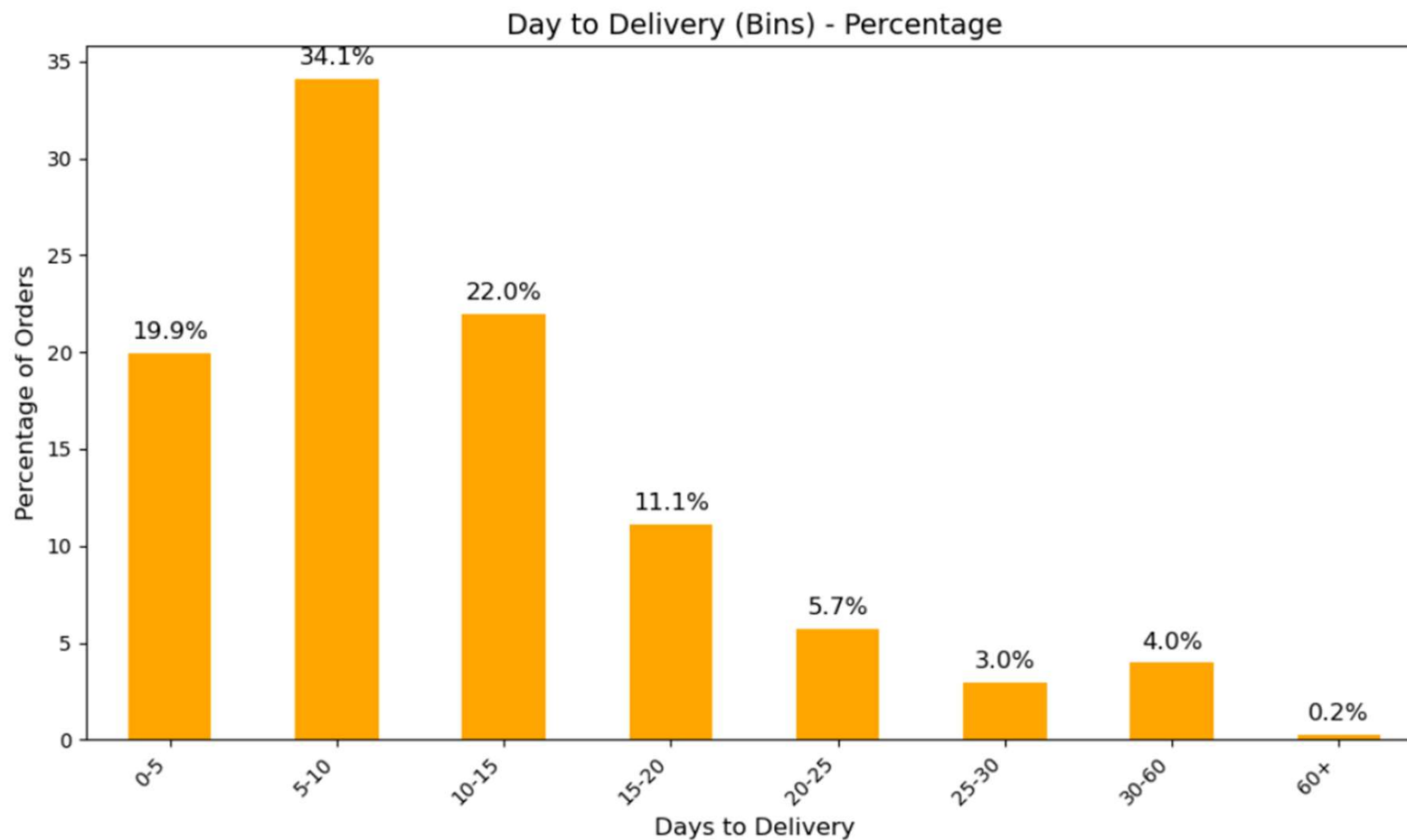
## 8. Data Visualisation



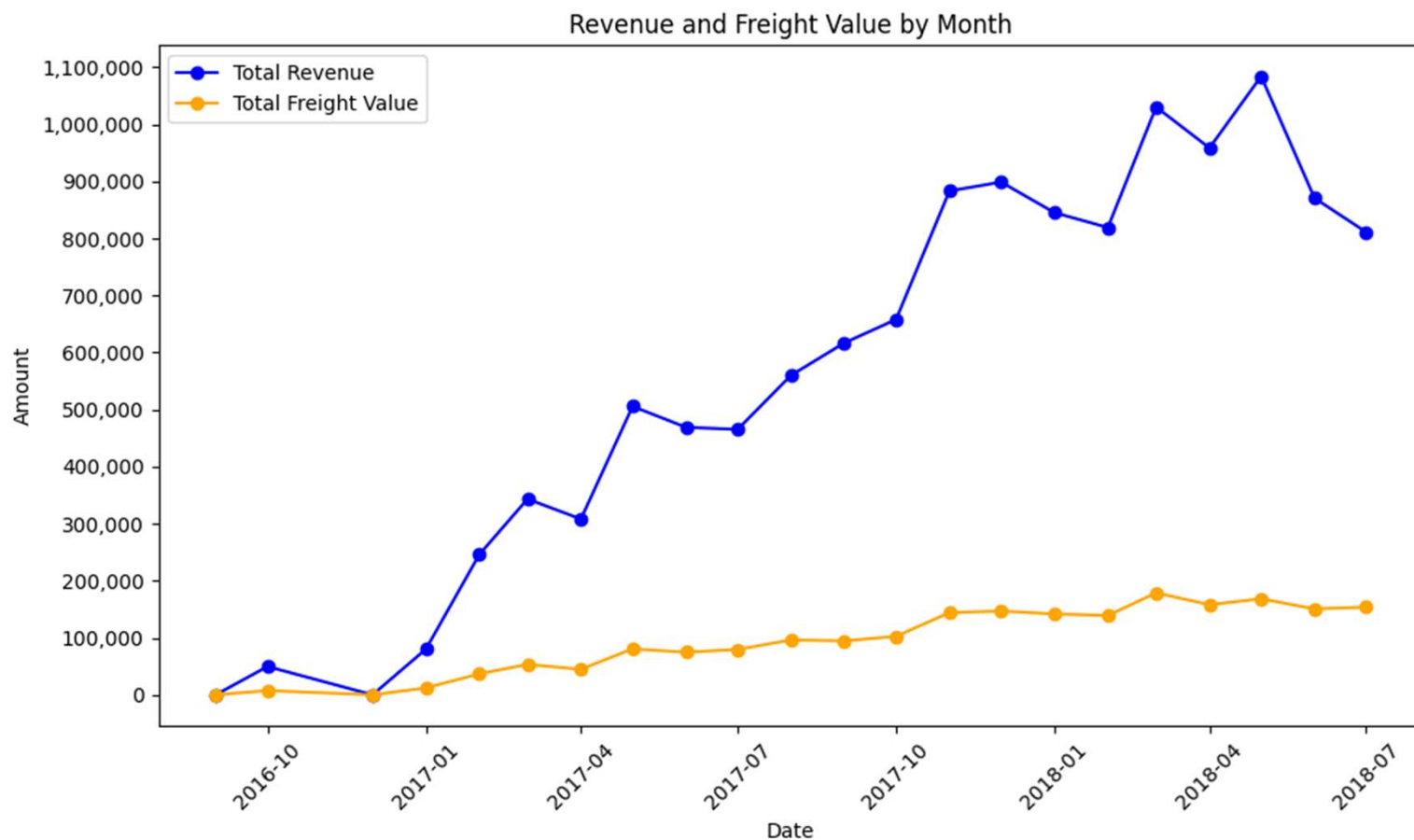
## 8. Data Visualisation



## 8. Data Visualisation



## 8. Data Visualisation

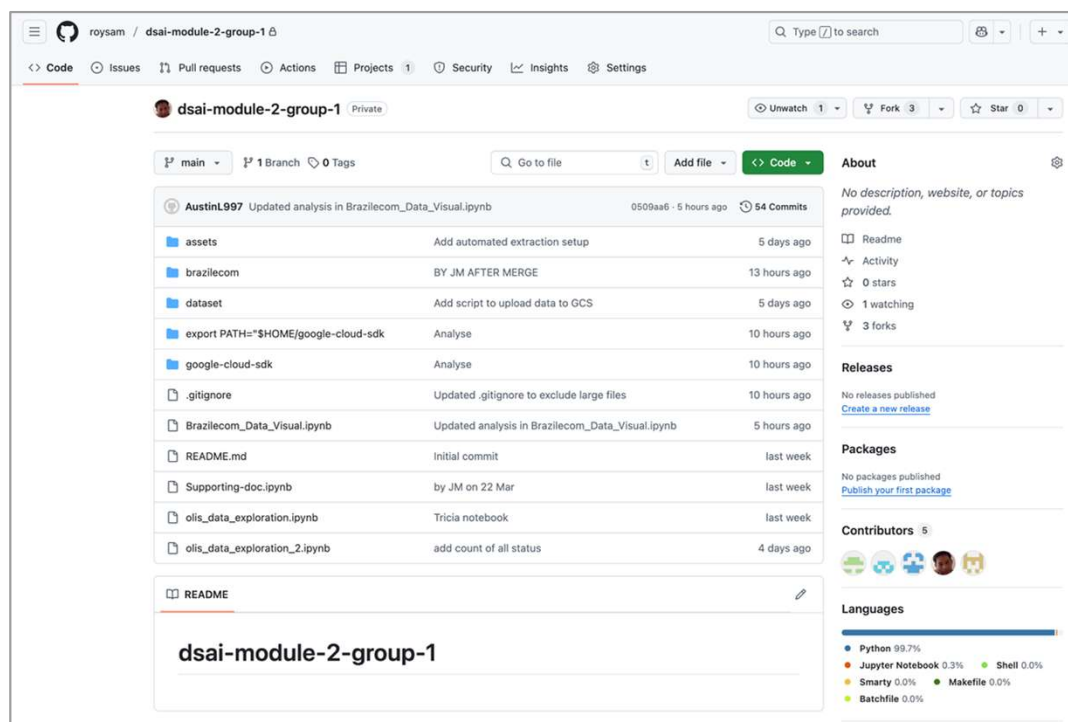




# Documentation

## GitHub Repo

<https://github.com/roysam/dsai-module-2-group-1>



## Data

There are following are the csv files that we will be using:

Description	Filename	No. of columns	No. of records
Customer	olist_customers_dataset	5	99441
Orders	olist_orders_dataset	8	99441
Products	olist_products_dataset	9	32951
Order Items	olist_order_items_dataset	7	112650
Sellers	olist_sellers_dataset	4	3095
Payments	olist_order_payments_dataset	5	103886
Product Category	product_category_name_translation	2	71
Geolocation	olist_geolocation_dataset	5	1000163



## Data Exploration and Understanding

Table Name	Findings	Data Cleanup
Products	<ul style="list-style-type: none"><li>610 records have null values for product_category_name, product_name_lenght, product_description_lenght, product_photos_qty</li></ul>	Categorised under “Others”
Sellers	<ul style="list-style-type: none"><li>Different variations of spelling of the same seller city</li></ul>	No action taken
Order Items	<ul style="list-style-type: none"><li>No quantity column</li><li>Duplicated product_id under the same order_id<ul style="list-style-type: none"><li>customer ordered more than 1 quantity of the same product</li></ul></li><li>freight_value is very much higher than item price<ul style="list-style-type: none"><li>need to make sense of this data</li></ul></li></ul>	No action taken

## Data Exploration and Understanding

Table Name	Findings	Data Cleanup
Orders	<ul style="list-style-type: none"> <li>• Inconsistency in data quality of same status. <ul style="list-style-type: none"> <li>◦ 6 cancelled orders, found with order_delivered_customer_data</li> <li>◦ 23 delivered orders that was found with missing values (Mainly for order_approved_at / order_delivered_carrier_date / order_delivered_customer_date)</li> </ul> </li> <li>• Illogical timestamps for 1395 orders</li> </ul>	<p>Further investigation showed that the missing values were due to the order are still being delivered.</p> <p>Our fact table only take in 4 types of order status: Delivered, Shipped, Invoiced, Processing</p>
order_payment	<ul style="list-style-type: none"> <li>• not_defined payment type are found <ul style="list-style-type: none"> <li>◦ 3 cancelled orders</li> </ul> </li> <li>• Inconsistency in data quality <ul style="list-style-type: none"> <li>◦ 6 additional voucher orders have payment value of \$0</li> </ul> </li> </ul>	No action taken
Customers	<ul style="list-style-type: none"> <li>• Inconsistent customer_zip_code_prefix <ul style="list-style-type: none"> <li>◦ 23995 rows do not have 5 digit prefix, lacking the 0 padding</li> </ul> </li> </ul>	No action taken

## Data Exploration and Understanding

	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value
0	8272b63d03f5f79c56e9e4120aec44ef	2	05b515fdc76e888aada3c6d66c201dff	2709af9587499e95e803a6498a5a56e9	2017-07-21 18:25:23+00:00	1.2	7.89
1	8272b63d03f5f79c56e9e4120aec44ef	3	05b515fdc76e888aada3c6d66c201dff	2709af9587499e95e803a6498a5a56e9	2017-07-21 18:25:23+00:00	1.2	7.89
2	8272b63d03f5f79c56e9e4120aec44ef	4	05b515fdc76e888aada3c6d66c201dff	2709af9587499e95e803a6498a5a56e9	2017-07-21 18:25:23+00:00	1.2	7.89
3	8272b63d03f5f79c56e9e4120aec44ef	5	05b515fdc76e888aada3c6d66c201dff	2709af9587499e95e803a6498a5a56e9	2017-07-21 18:25:23+00:00	1.2	7.89
4	8272b63d03f5f79c56e9e4120aec44ef	6	05b515fdc76e888aada3c6d66c201dff	2709af9587499e95e803a6498a5a56e9	2017-07-21 18:25:23+00:00	1.2	7.89
5	8272b63d03f5f79c56e9e4120aec44ef	7	05b515fdc76e888aada3c6d66c201dff	2709af9587499e95e803a6498a5a56e9	2017-07-21 18:25:23+00:00	1.2	7.89
6	8272b63d03f5f79c56e9e4120aec44ef	8	05b515fdc76e888aada3c6d66c201dff	2709af9587499e95e803a6498a5a56e9	2017-07-21 18:25:23+00:00	1.2	7.89
7	8272b63d03f5f79c56e9e4120aec44ef	9	05b515fdc76e888aada3c6d66c201dff	2709af9587499e95e803a6498a5a56e9	2017-07-21 18:25:23+00:00	1.2	7.89
8	8272b63d03f5f79c56e9e4120aec44ef	10	05b515fdc76e888aada3c6d66c201dff	2709af9587499e95e803a6498a5a56e9	2017-07-21 18:25:23+00:00	1.2	7.89
9	8272b63d03f5f79c56e9e4120aec44ef	11	05b515fdc76e888aada3c6d66c201dff	2709af9587499e95e803a6498a5a56e9	2017-07-21 18:25:23+00:00	1.2	7.89
10	8272b63d03f5f79c56e9e4120aec44ef	1	270516a3f41dc035aa87d220228f844c	2709af9587499e95e803a6498a5a56e9	2017-07-21 18:25:23+00:00	1.2	7.89

Same product id

## Data Exploration and Understanding

Null values in products table:

	product_id	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_lenght
2454	5eb564652db742ff8f28759cd8d2652a	None	<NA>	<NA>	<NA>	<NA>	<
2455	e10758160da97891c2fdbc35f0f031d	None	<NA>	<NA>	<NA>	2200	
2456	39e3b9b12cd0bf8ee681bbc1c130feb5	None	<NA>	<NA>	<NA>	300	
2457	bc815bba008d89458e428078c0b9211	None	<NA>	<NA>	<NA>	150	
2458	212cc0fa7359ab242a697a03a574f719	None	<NA>	<NA>	<NA>	200	
...	...	...	...	...	...	...	
3059	6962734c72522e70e852a2a77d21a730	None	<NA>	<NA>	<NA>	10050	
3060	cee7d5636e59173cc5f484e913db3d	None	<NA>	<NA>	<NA>	30000	
3061	b0a0c5dd78e644373b199380612c350a	None	<NA>	<NA>	<NA>	1800	
3062	7167af17015615b513d5b429758969a2	None	<NA>	<NA>	<NA>	21100	
3063	946cde79b9ebdfc56c52a405cc54dc	None	<NA>	<NA>	<NA>	1600	

## Data Exploration and Understanding

Different variations of the same city

	seller_id	seller_zip_code_prefix	seller_city	seller_state
0	c13ef0cfbe42f190780f621ce81f2234	1207	sao paulo sp	SP
1	5444b12c82f21c923f2639ebc722c1ea	2051	sao pauo	SP
3	71593c7413973a1e160057b80d4958f6	3407	sao paulo / sao paulo	SP
4	6f1a1263039c76e68f40a8e536b1da6a	3581	sao paulop	SP
5	06579cb253ecd5a3a12a9e6eb6bf8f47	4007	sao paulo - sp	SP
...	...	...	...	...
3058	778323240ce2830d68aab11794e00bfb	13600	sao paulo	SP
3059	dace965ca58120f92f8d742a9fa1864b	14015	sao paulo	SP
3060	761681a821d8275bc79f552116d06869	17606	sao paulo	SP
3061	a64e44665225d19dfc0277eeeaaccc57	19400	sao paulo	SP
3062	2a167ca73899c85001a837d8fb4962f6	37540	sao paulo	SP

## Data Exploration and Understanding (Orders)

	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date
0	65d1e226dfaeb8cdc42f665422522d14	70fc57eeae292675927697fe03ad3ff5	canceled	2016-10-03 21:01:41+00:00	2016-10-04 10:18:57+00:00	2016-10-25 12:14:28+00:00	2016-11-08 10:58:34+00:00
1	770d331c84e5b214bd9dc70a10b829d0	6c57e6119369185e575b36712766b0ef	canceled	2016-10-07 14:52:30+00:00	2016-10-07 15:07:10+00:00	2016-10-11 15:07:11+00:00	2016-10-14 15:07:11+00:00
2	dabf2b0e35b423f94618bf965fcb7514	5cdec0bb8cbdf53ffc8fdc212cd247c6	canceled	2016-10-09 00:56:52+00:00	2016-10-09 13:36:58+00:00	2016-10-13 13:36:59+00:00	2016-10-16 14:36:59+00:00
3	8beb59392e21af5eb9547ae1a9938d06	bf609b5741f71697f65ce3852c5d2623	canceled	2016-10-08 20:17:50+00:00	2016-10-09 14:34:30+00:00	2016-10-14 22:45:26+00:00	2016-10-19 18:47:43+00:00
4	2c45c33d2f9cb8ff8b1c86cc28c11c30	de4caa97afa80c8eeac2ff4c8da5b72e	canceled	2016-10-09 15:39:56+00:00	2016-10-10 10:40:49+00:00	2016-10-14 10:40:50+00:00	2016-11-09 14:53:50+00:00
5	1950d777989f6a877539f53795b4c3c3	1bccb206de9f0f25adc6871a1bcf77b2	canceled	2018-02-19 19:48:52+00:00	2018-02-19 20:56:05+00:00	2018-02-20 19:57:13+00:00	2018-03-21 22:03:51+00:00

**Incorrect logic for order\_delivered\_customer\_data**

Cancelled orders should not have the data for order\_delivered\_customer\_data. It could potentially represent a refund upon delivery. Which should be illustrated by creating addition status under 'Refund' instead of clumping into 'canceled' status

## Data Exploration and Understanding (Payment)

	order_id	payment_sequential	payment_type	payment_installments	payment_value	order_status
0	8bcbe01d44d147f901cd3192671144db	4	voucher	1	0.0	delivered
1	fa65dad1b0e818e3ccc5cb0e39231352	14	voucher	1	0.0	shipped
2	6ccb433e00daae1283ccc956189c82ae	4	voucher	1	0.0	delivered
3	4637ca194b6387e2d538dc89b124b0ee	1	not_defined	1	0.0	canceled
4	00b1cb0320190ca0daa2c88b35206009	1	not_defined	1	0.0	canceled
5	45ed6e85398a87c253db47c2d9f48216	3	voucher	1	0.0	delivered
6	fa65dad1b0e818e3ccc5cb0e39231352	13	voucher	1	0.0	shipped
7	c8c528189310eaa44a745b8d9d26908b	1	not_defined	1	0.0	canceled
8	b23878b3e8eb4d25a158f57d96331b18	4	voucher	1	0.0	delivered

**Delivered/Shipped Status**

indicate that the order was successful and the 0 payment\_value represent a data inconsistency

**Canceled Status**

indicate that the order was unsuccessful and could have resulted in 0 payment\_value

Note: Voucher payment should still be represented with how much was paid



## Data Exploration and Understanding (Customers)

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
57415	7ae2a9337aa4bc799723511faa1d6830	0c1a20644f0dc126c3eaff8dbc1bd12c	1003	sao paulo	SP
57416	a09edf8c1e842e94805a206b3d73eed5	968f6d2f674977d88a4b445a5117ccd8	1004	sao paulo	SP
57417	ee9b73e88afb4904ee2322cfc89cf638	095e7c124c5c1ccb1eb9f731152eae6a	1004	sao paulo	SP
57418	5a8b64ee6ccdae09ea823e6aa00e9517	9c84e5193d6ee59b3870e0e4e3a2dad8	1005	sao paulo	SP
57419	6ec2b4682814cfdac8d92bad42b3ddab	57f0ea1c7f6b9ef8615c0a0b8f06fe57	1005	sao paulo	SP
...	...	...	...	...	...
81405	428db965aedef0c8c56f94d005539a9b0	97bc08e526795b17e1d4f642f77ae304	9993	diadema	SP
81406	fd04bf849b36444f719850585a9b0e8a	97bc08e526795b17e1d4f642f77ae304	9993	diadema	SP
81407	9b9024a27b845a8b50ef8d1b7ba89ee8	97bc08e526795b17e1d4f642f77ae304	9993	diadema	SP
81408	10091d0f711745db12815a7935577e26	a8bd559f5b029d6f96e3c9d134288dba	9993	diadema	SP
81409	10567872c1e2e0ba7172faf0a144c21d	461b0e7c11ff521493eaa69ad24e7b3d	9993	diadema	SP

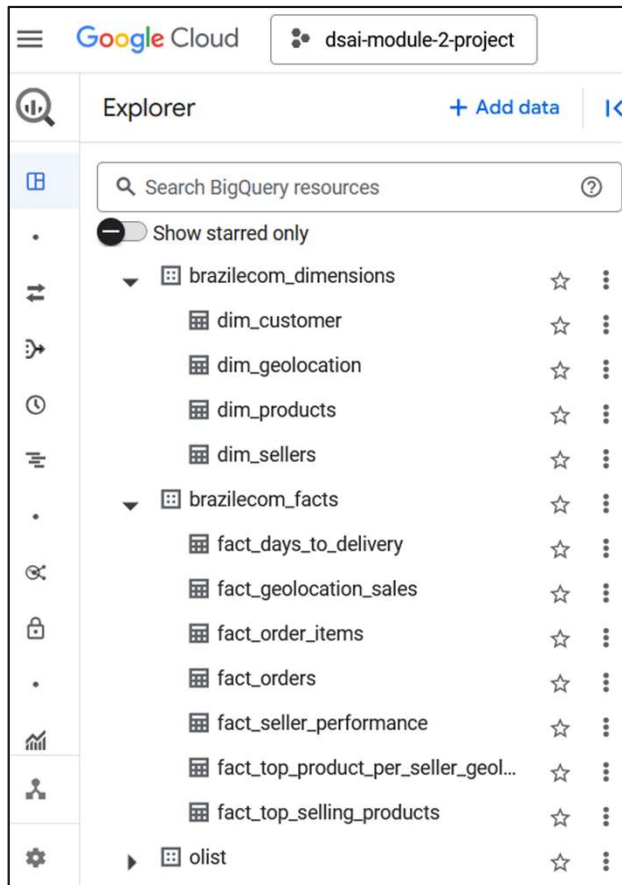
23995 rows x 5 columns

**Represent inconsistency in 5 digit prefix zip code**  
**Data shows a total of 23995 customer\_zip\_code\_prefix that lacks padding to meet the requirement of 5 digit prefix zip code.**





## Final Output in Bigquery Data Warehouse





# Thank You