DSAI Module 2
Final Project Presentation

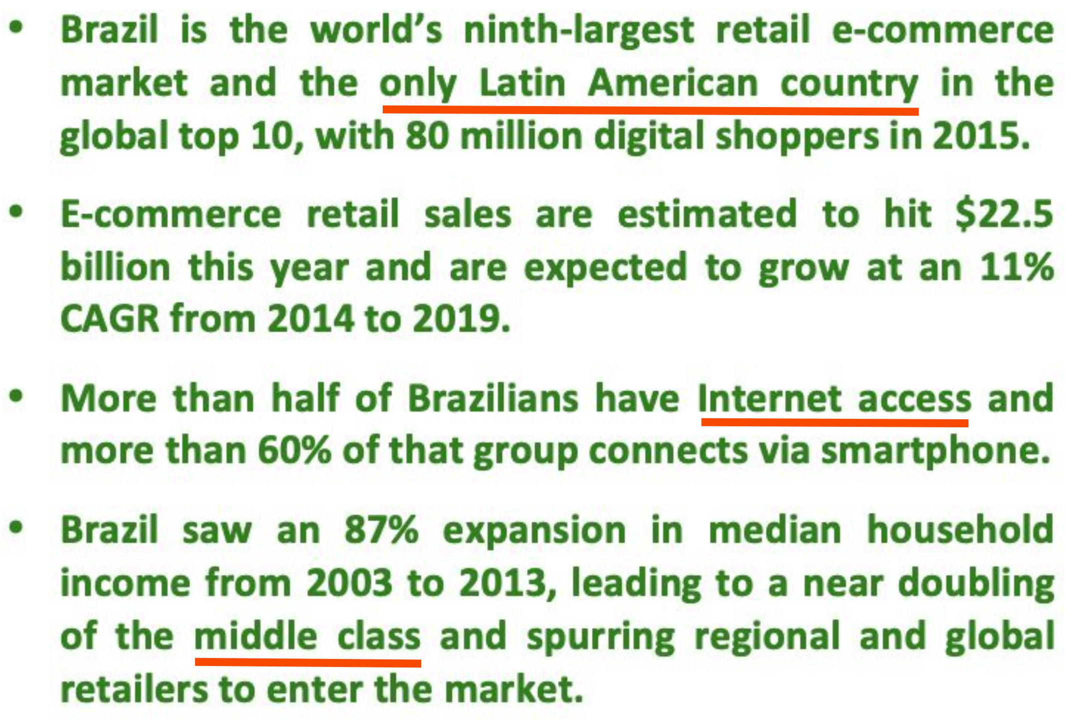
Solution Implementation

Group Members
Austin, Jason, Roy, Tricia, Vasanthi

29 Mar 2025

In 2015...

OVERVIEW OF BRAZILIAN E-COMMERCE



- 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.

source: https://www.lifung.com/wp-content/uploads/2016/03/Overview-of-Brazil-Ecom-by-Fung-Global-Retail-Tech-Mar.-8-2016.pdf

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**

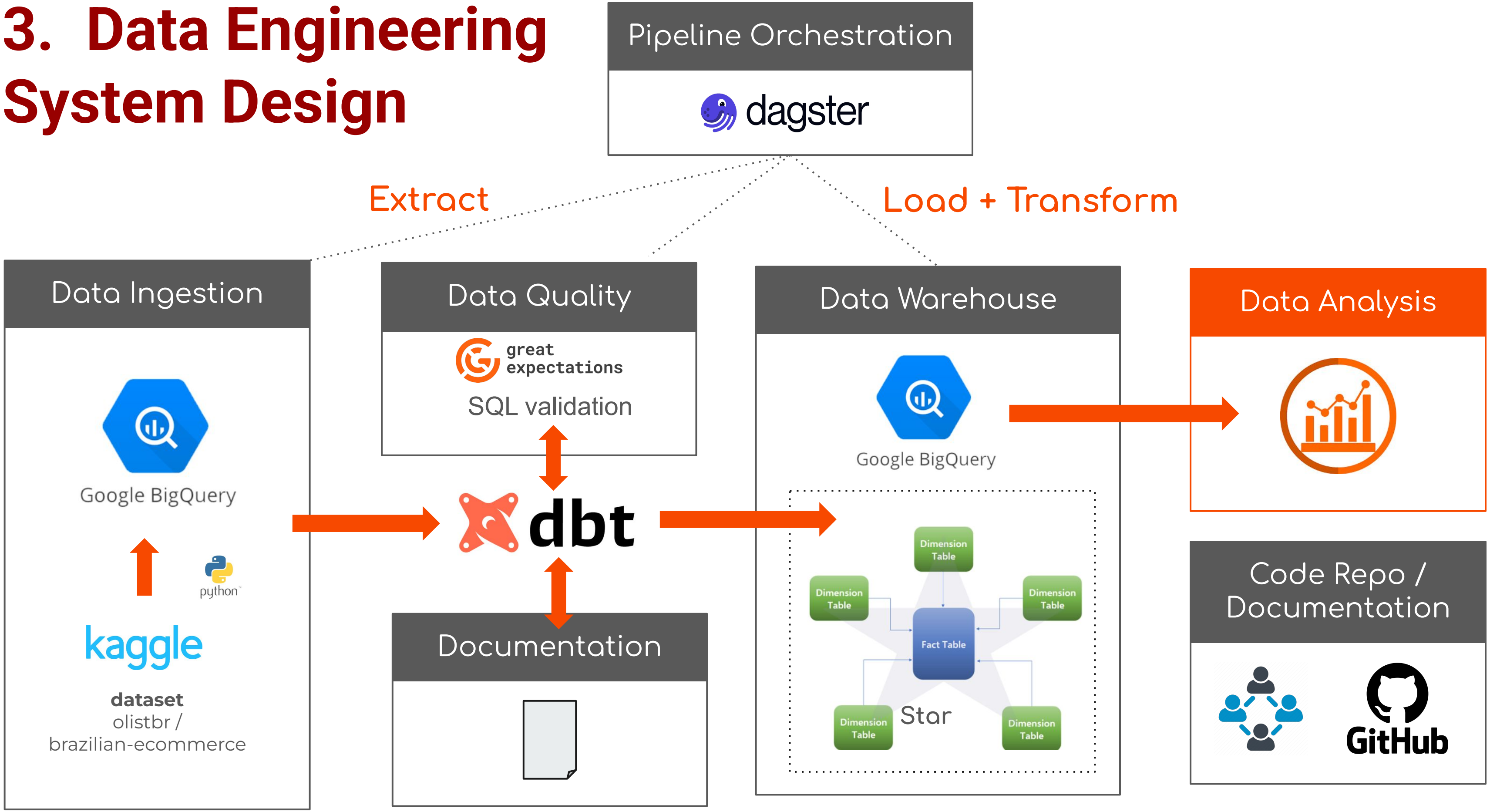
This pipeline will **automate** data cleaning, preprocessing, and quality assurance to ensure accurate and up-to-date data for analysis.

Business Objectives

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

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

3. Data Engineering System Design



Pipeline Orchestration

dagster

Extract

Load + Transform

Data Ingestion

Google BigQuery

python

kaggle

dataset
olistbr /
brazilian-ecommerce

Data Quality

great expectations

SQL validation

dbt

Documentation

Data Warehouse

Google BigQuery

Star

Data Analysis

Code Repo / Documentation

GitHub

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
- Customer satisfaction metrics

4. Data Exploration and Understanding

Data Issue 1

Different variations in spelling of seller and customer city

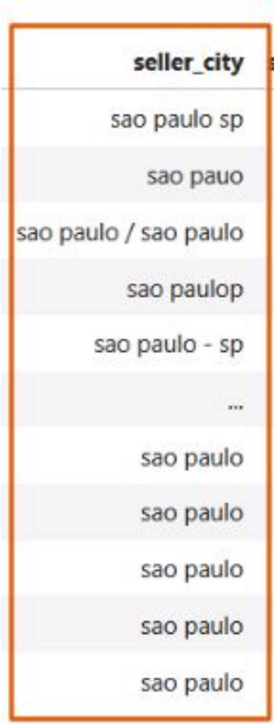
Data Issue 2

Unknown category name for over 600 products

Data Issue 3

Missing values in various timestamp in orders data

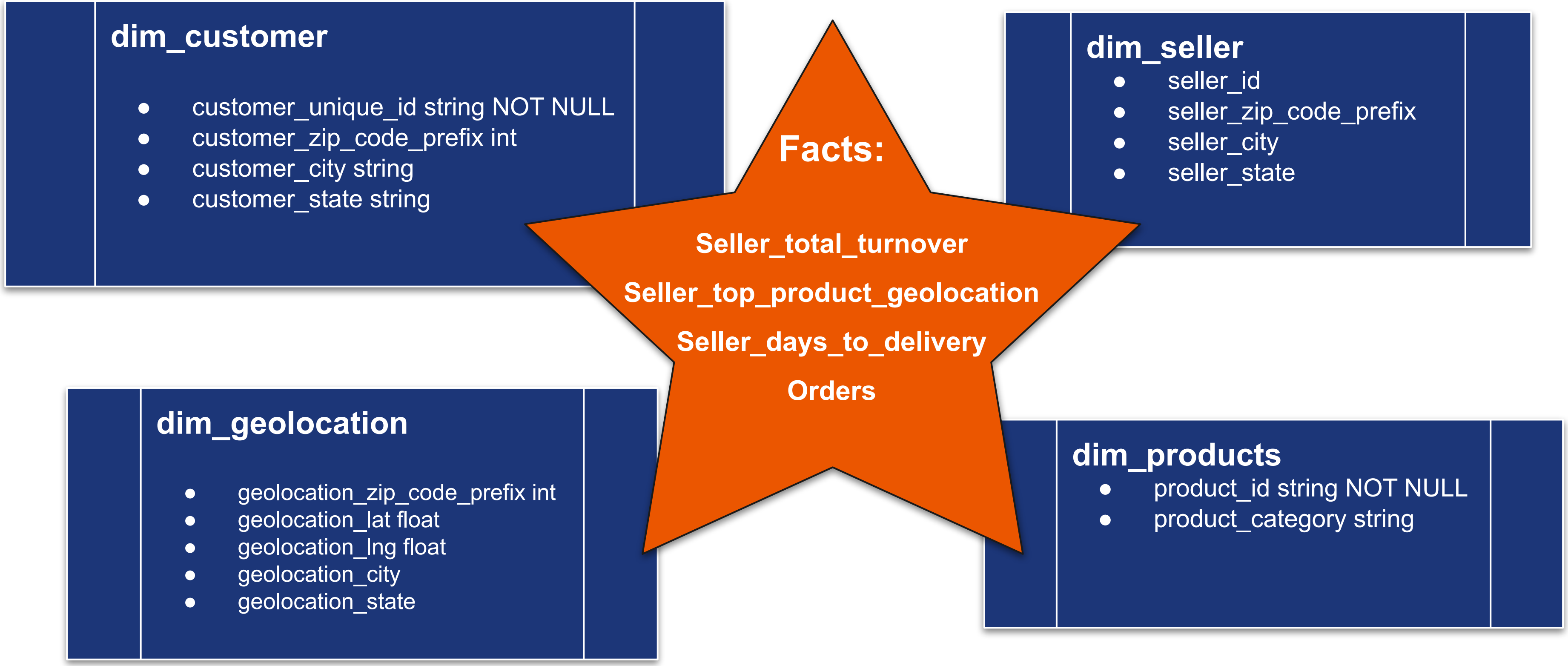
- Due to different stages of the delivery



seller_city
sao paulo sp
sao pauo
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. Star Schema Design



dim_customer

- customer_unique_id string NOT NULL
- customer_zip_code_prefix int
- customer_city string
- customer_state string

Facts:

Seller_total_turnover

Seller_top_product_geolocation

Seller_days_to_delivery

Orders

dim_seller

- seller_id
- seller_zip_code_prefix
- seller_city
- seller_state

dim_geolocation

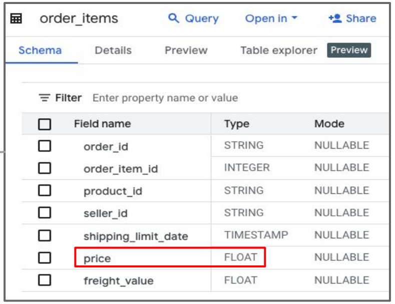
- geolocation_zip_code_prefix int
- geolocation_lat float
- geolocation_lng float
- geolocation_city
- geolocation_state

dim_products

- product_id string NOT NULL
- product_category string

6. Data Quality Testing Design

```
.
|-- brazilecom
|    |-- 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
```



☐	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 `brazilecom.order_items`
WHERE NOT (price BETWEEN 0.0 AND 10000.0)
```

Run the test under tests folder

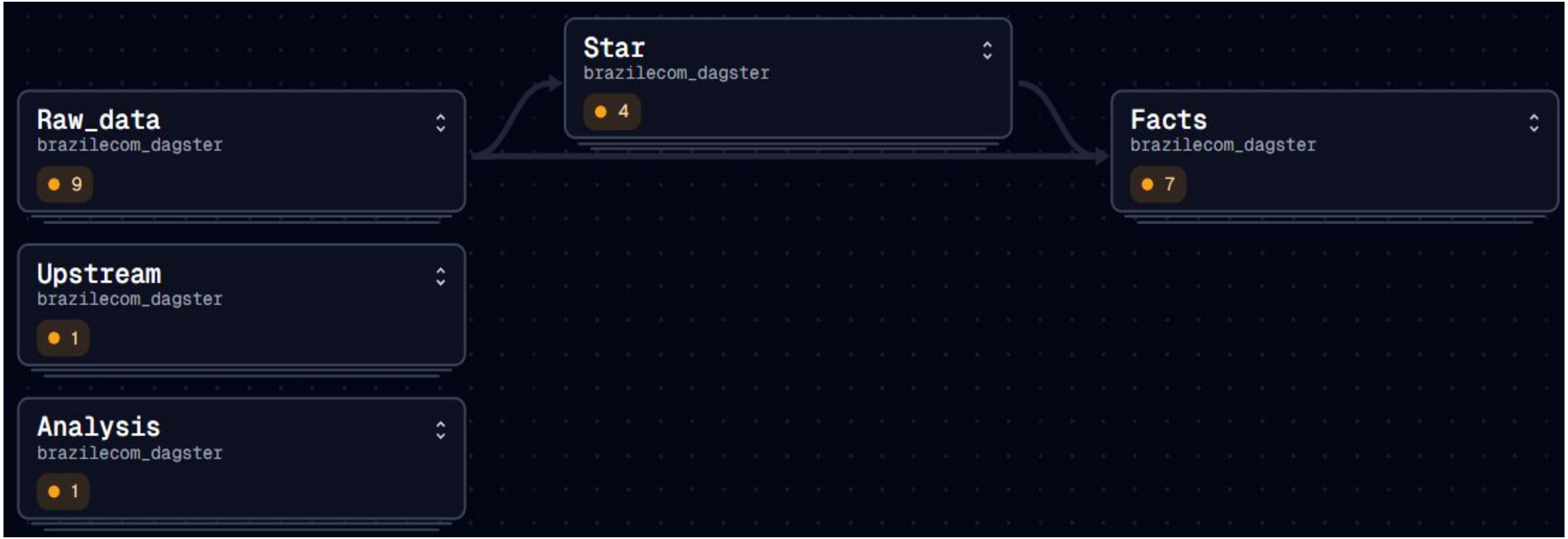
```
% dbt test --select
tests/dbt_test_order_items_price_check.sql
```

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]
```

7. Pipeline Orchestration (Dagster) - Asset

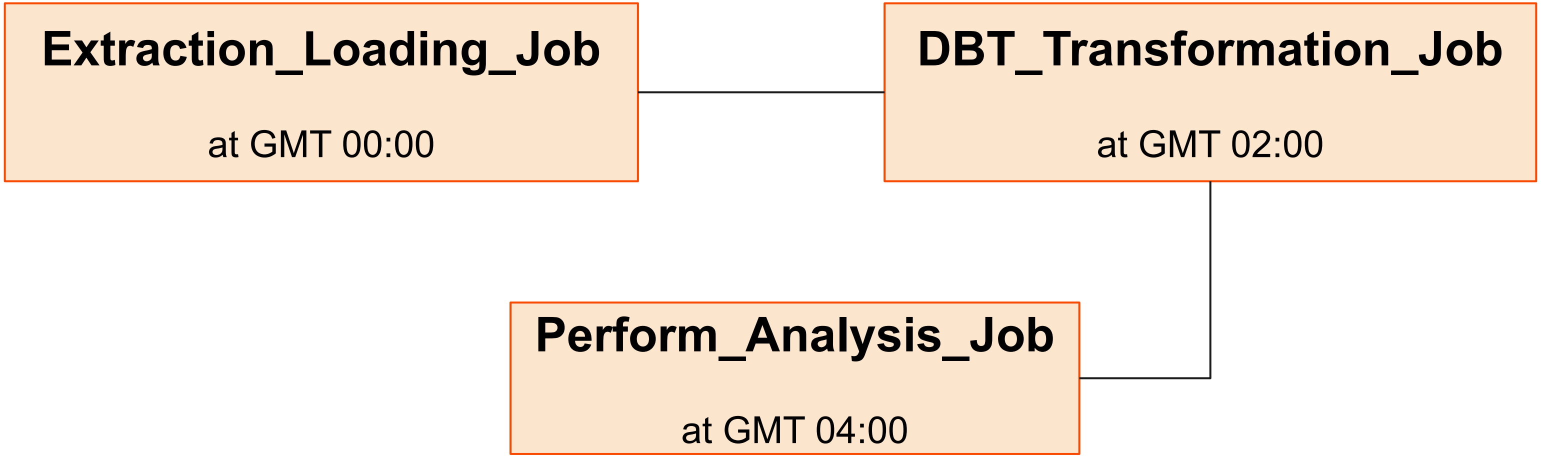


Asset Groups:

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. Pipeline Orchestration (Pipeline Automation)

Dagster Schedule Job



Extraction_Loading_Job

at GMT 00:00

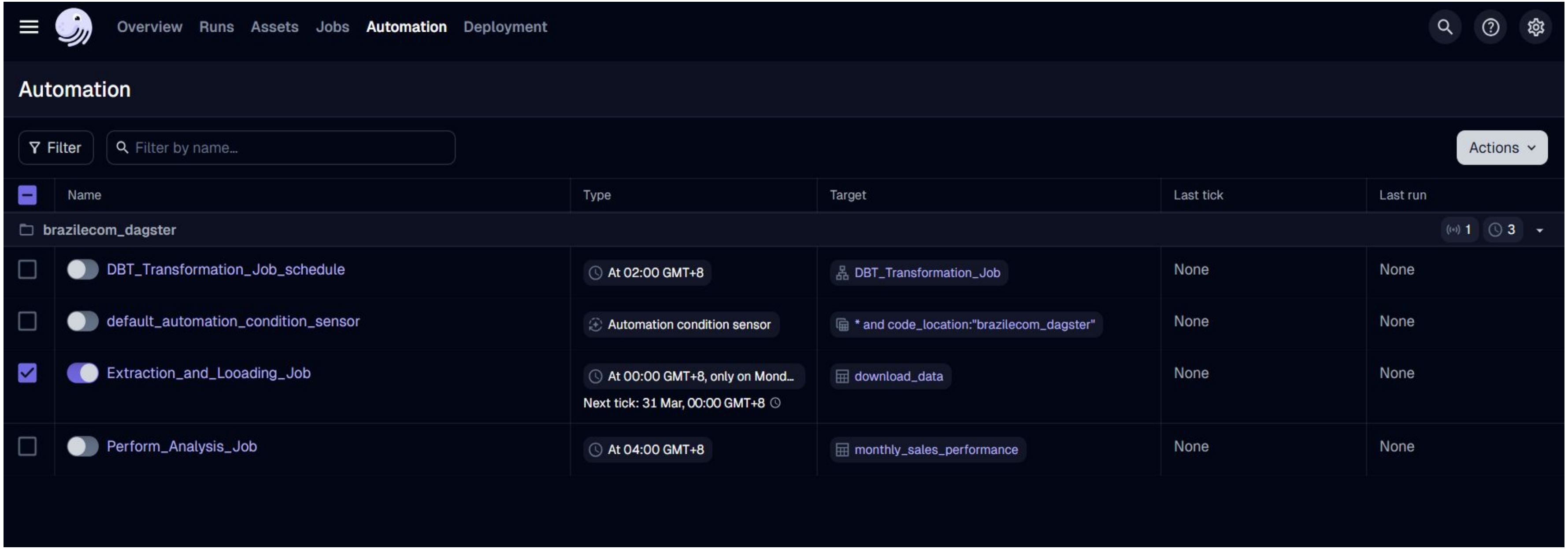
DBT_Transformation_Job

at GMT 02:00

Perform_Analysis_Job

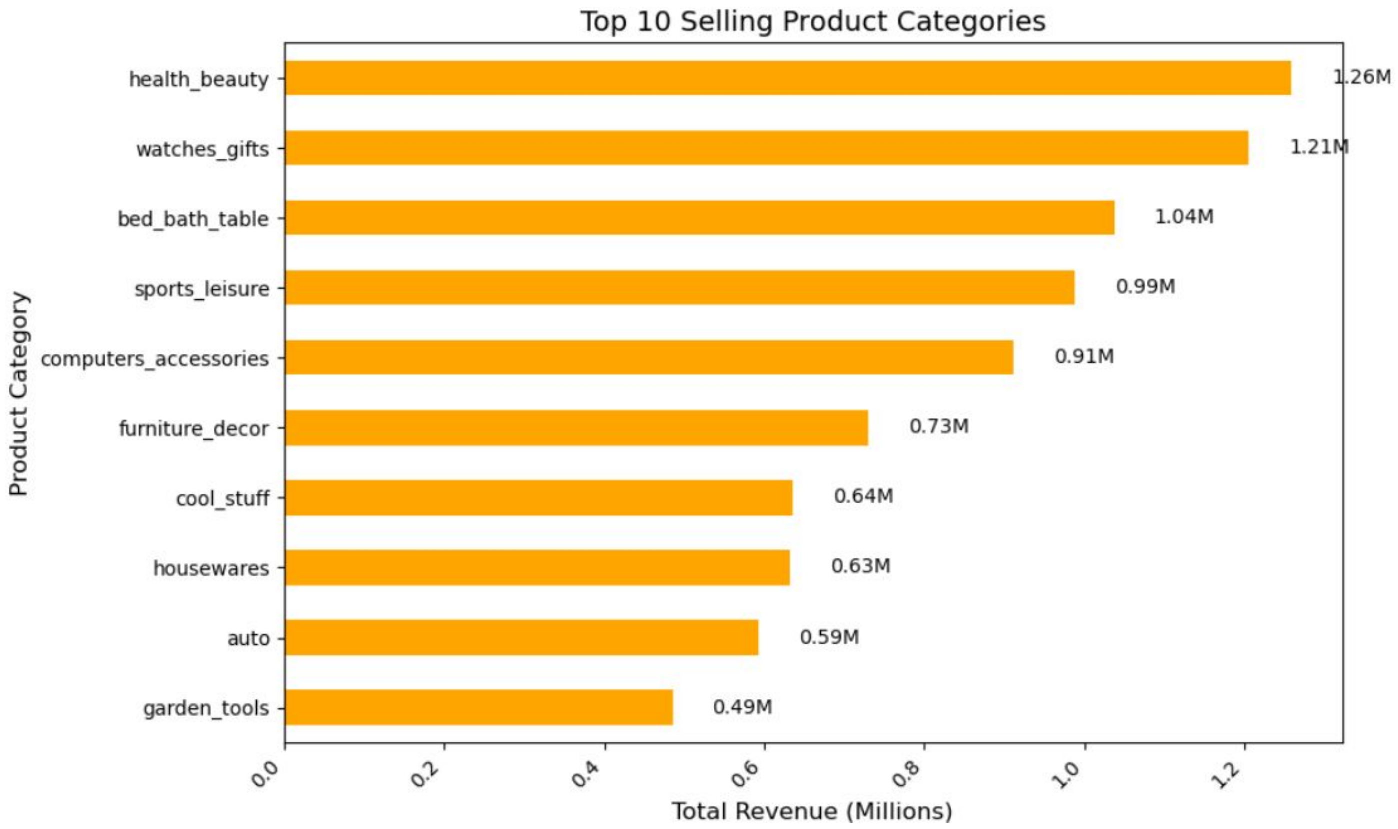
at GMT 04:00

7. Pipeline Orchestration (Pipeline Automation Schedule Jobs)

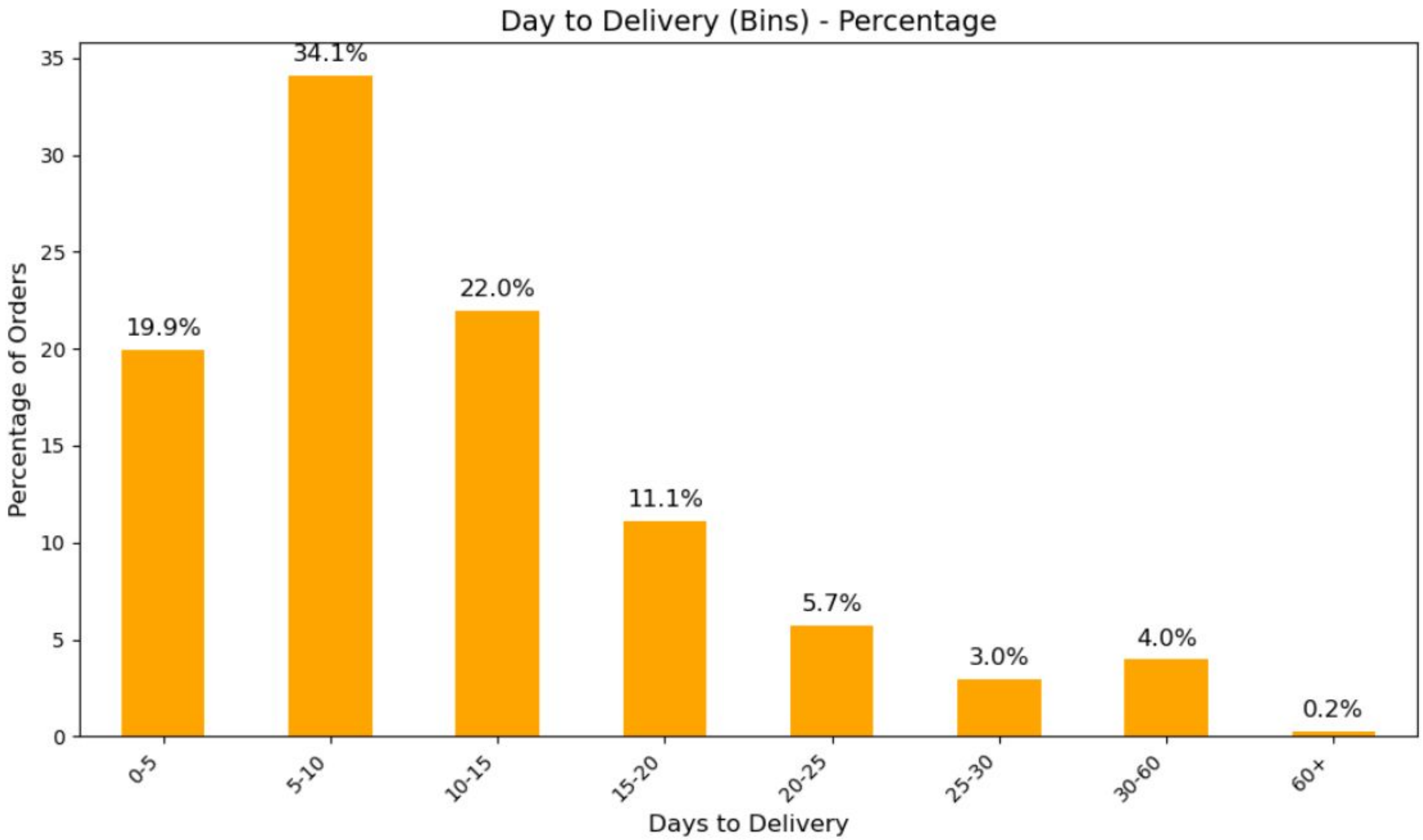


	Name	Type	Target	Last tick	Last run
	brazilecom_dagster				1 3
□	DBT_Transformation_Job_schedule	At 02:00 GMT+8	DBT_Transformation_Job	None	None
□	default_automation_condition_sensor	Automation condition sensor	* and code_location:"brazilecom_dagster"	None	None
☑	Extraction_and_Loooding_Job	At 00:00 GMT+8, only on Mond... Next tick: 31 Mar, 00:00 GMT+8	download_data	None	None
□	Perform_Analysis_Job	At 04:00 GMT+8	monthly_sales_performance	None	None

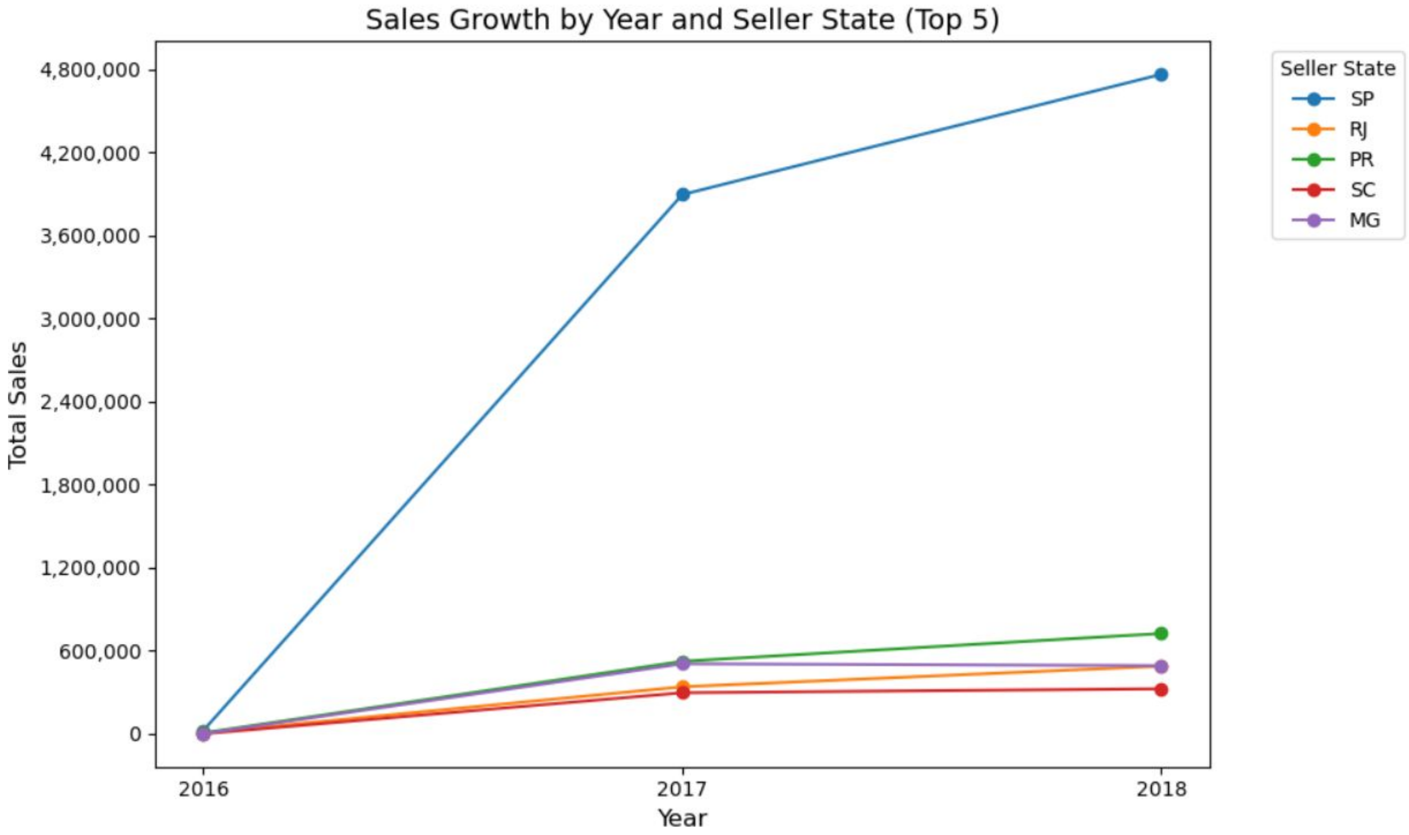
8. Data Visualisation



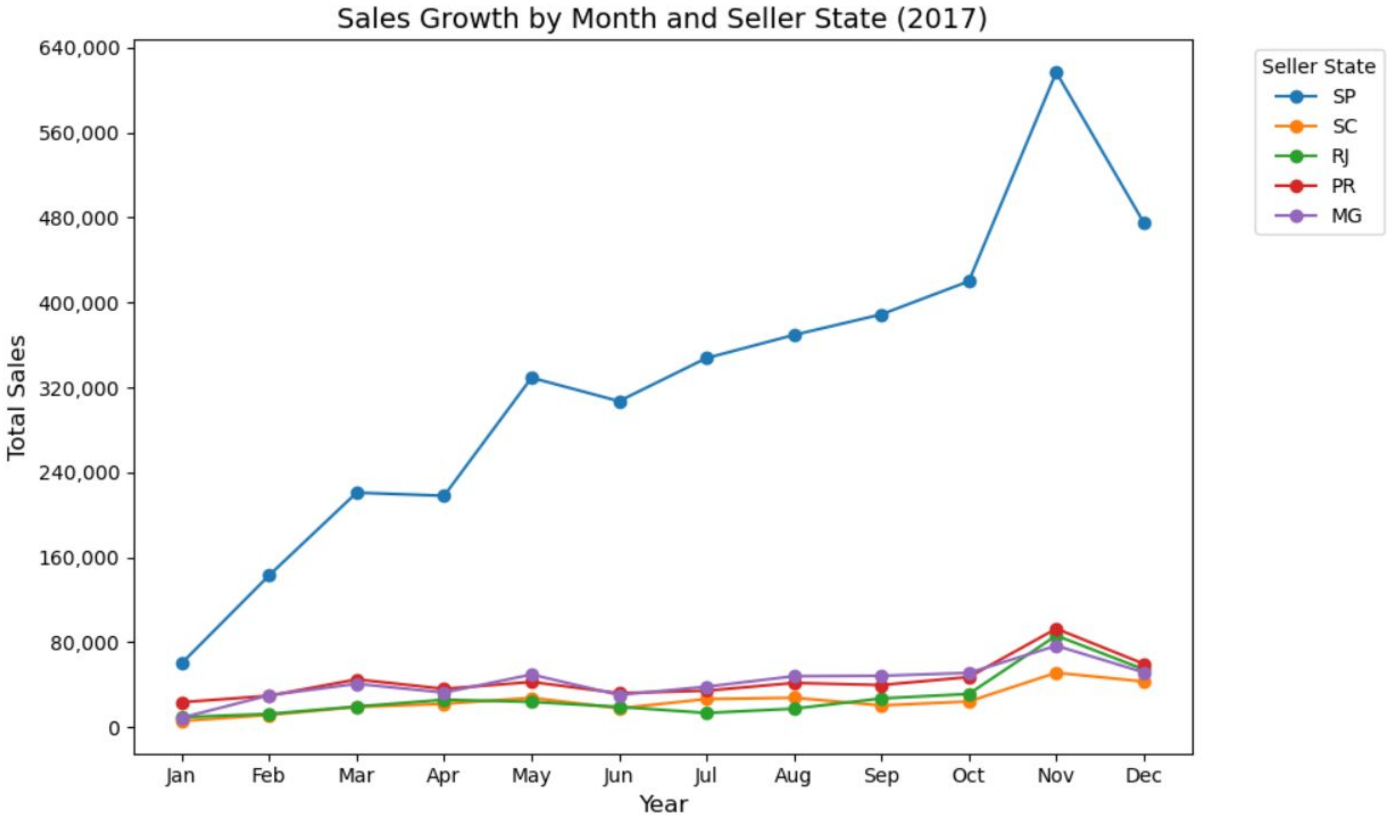
8. Data Visualisation



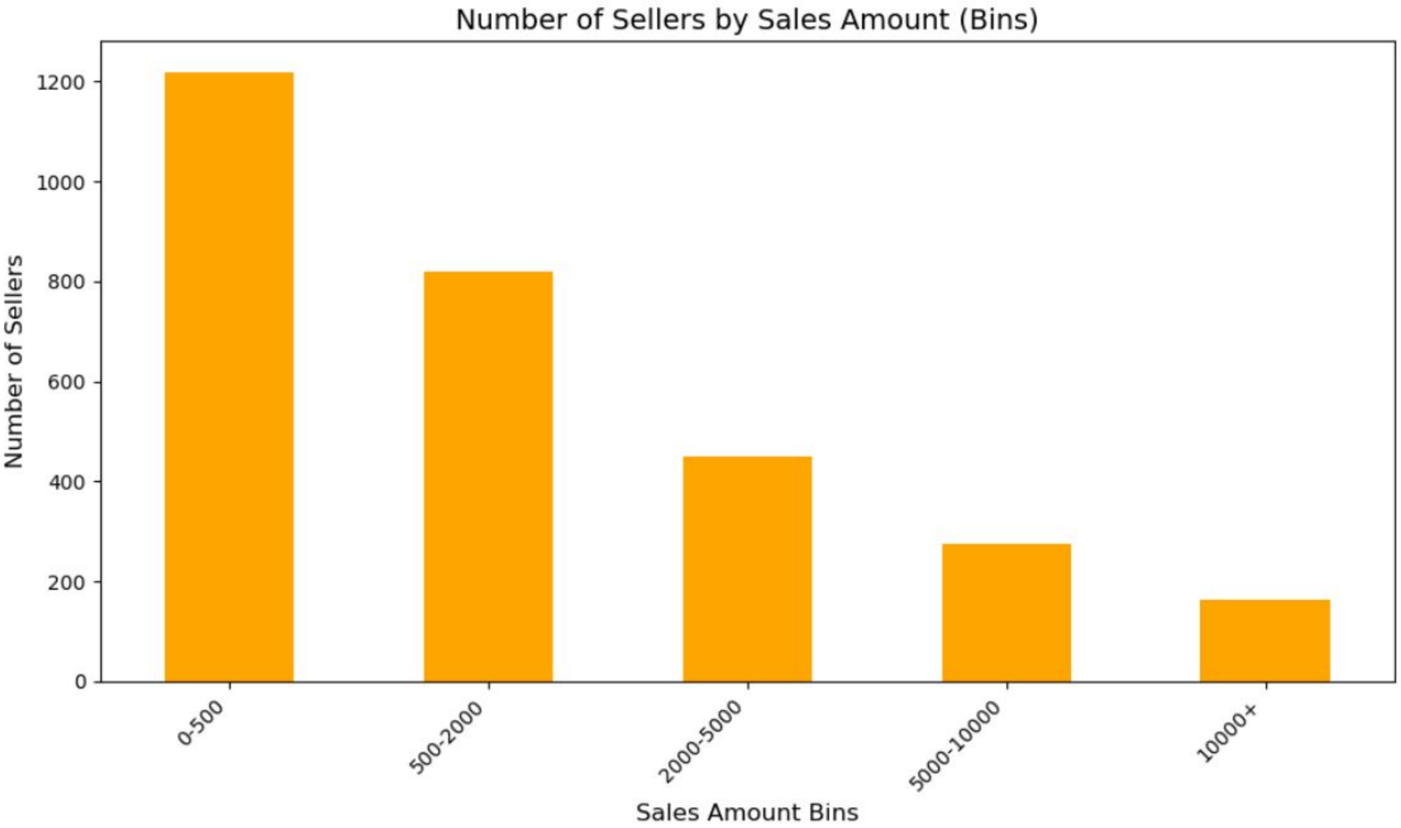
8. Data Visualisation



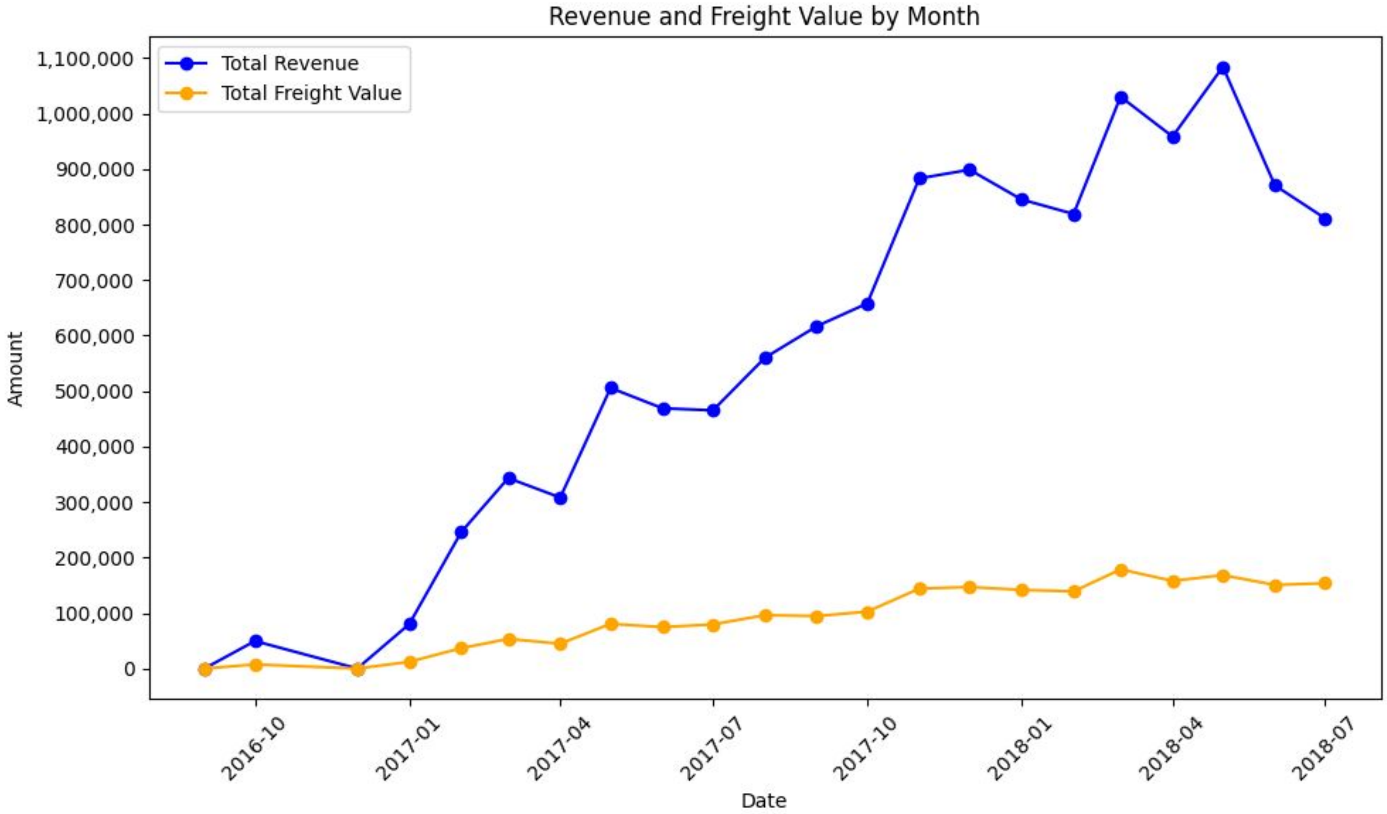
8. Data Visualisation



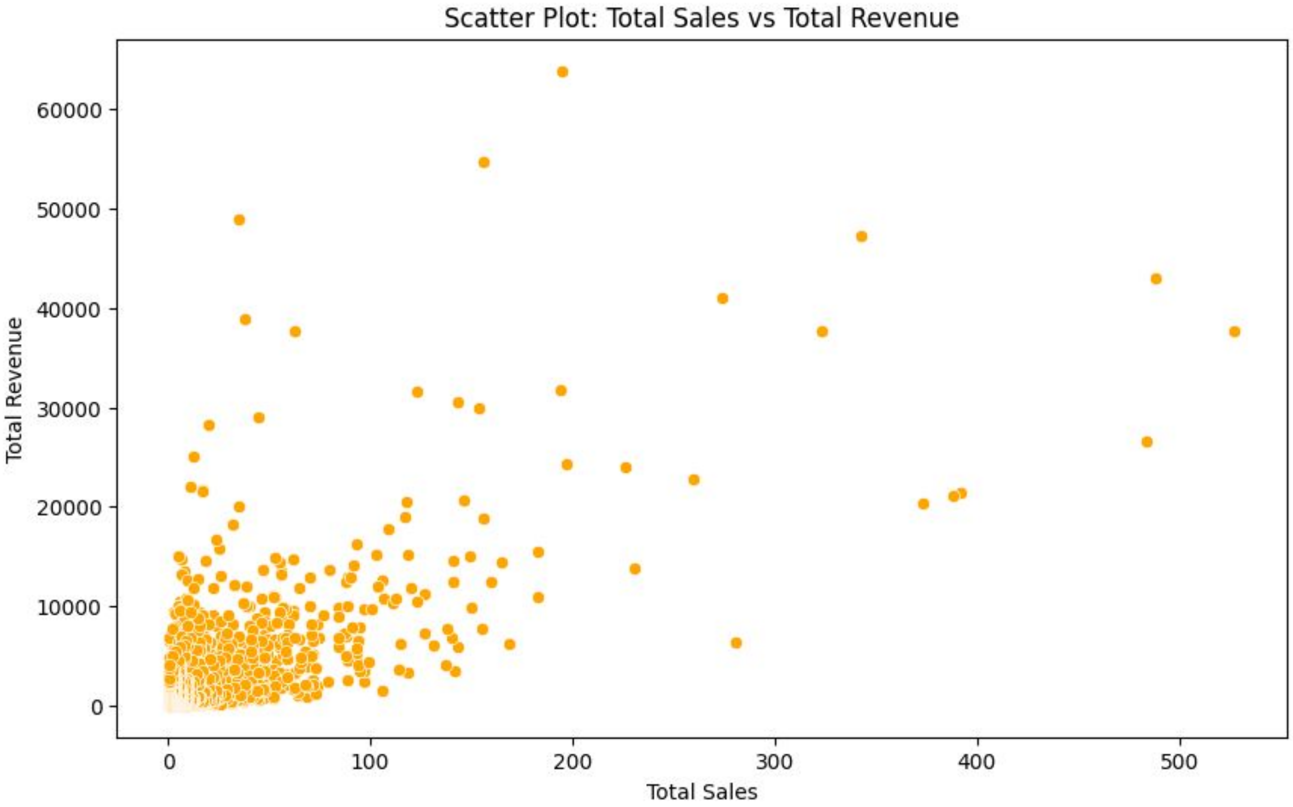
8. Data Visualisation



8. Data Visualisation



8. Data Visualisation



Thank You