ETL Pipeline

for

Grab Delivery

Data Insights

An Interim Project by Byte Buddies

**Saila, Rosni, Lee Hong Eng, Hassan Hosain**

of the

*JDE03 Cohort*

Documentation

**Table of Contents**

[1) Introduction](#_heading=h.fiw3ec9177jc)

[Overview](#_heading=h.se91htthhuzj)

[Problem Statement](#_heading=h.8e1hv8utxyvr)

[2) Data Source](#_heading=h.2ydetfqn2qyz)

[Data Dictionary](#_heading=h.ya936dgkmosu)

[Fields:](#_heading=h.pkir4vtr5621)

[3) Methodology/Approach](#_heading=h.fqwema3646kz)

[Teamwork Approach](#_heading=h.roz4amlks6r6)

[Project Workflow](#_heading=h.3l6lq2xg4l0d)

[Sourcing for datasets](#_heading=h.micwajafhpvc)

[Deliverables Assignment](#_heading=h.sjy58ueq3so4)

[Tools](#_heading=h.em6ei6vqgegj)

[4) ETL Process](#_heading=h.s665333rfuz7)

[Extract](#_heading=h.3vto8ot9argn)

[Transform](#_heading=h.u523w736b19z)

[o establishment\_df: (by Saila)](#_heading=h.6s67qh68eat7)

[o cuisine\_df: (by Hassan)](#_heading=h.bctpjmsmbxv0)

[o delivery\_df: (by Rosni)](#_heading=h.e8resh29isiz)

[o openinghour\_df: (by Hong Eng)](#_heading=h.55zwvhqjiyny)

[Load](#_heading=h.nng6s47uiznb)

[ETL Pipeline Diagram](#_heading=h.xxlz35ldvkzy)

[Database Schema/ERD Diagram](#_heading=h.cht5avt9o2yv)

[5) Challenges (Solutions/Workarounds/Limitations)](#_heading=h.tbmrkivqn74l)

[Challenges](#_heading=h.ucs5e83qi7qh)

[Limitations](#_heading=h.o7r5gedwcl8m)

[6) Findings/Analysis](#_heading=h.2zgfwkoojhkm)

[● Visualization - Top 10 cuisine category](#_heading=h.yvgws6wjnpux)

[● Visualization - All locations of selected cuisine](#_heading=h.wfgh1i9cm4t7)

[7) Next Steps/Conclusion](#_heading=h.lazcm37j1430)

# 1) Introduction

## Overview

This is a Junior Data Engineer Program Interim Project. This document outlines the project documentation for the ETL Pipeline for Grab Delivery Data Insights project, focusing on the ETL process, with the aim to analyze establishment locations and cuisine diversity to identify underserved areas and expansion opportunities for business owners and food delivery services, based on a dataset sourced from Kaggle.com.

## Problem Statement

From a **business perspective**, the analysis aims to identify optimal locations for new restaurant establishments based on the type of cuisine offered. By examining areas with a shortage of certain cuisines, the project will provide valuable insights to business owners on where to establish their restaurants to meet unfulfilled market demands.

From **Grab’s perspective**, this analysis serves as a strategic tool for expansion and competition. It will enable Grab, as well as its competitors, to identify underserved regions and cuisines, thereby informing decisions on where to focus expansion efforts to capitalize on market opportunities.

This dual-faceted approach ensures that the analysis benefits both individual business owners and Grab’s overarching strategic goals.

# 2) Data Source

This dataset is taken from Kaggle: 16000+ Grab Restaurants in Singapore; reference to <https://www.kaggle.com/datasets/polartech/16000-grab-restaurants-in-singapore>. Below is the data dictionary that explains the data source which is a csv format.

## Data Dictionary

The dataset for the ETL Pipeline for Grab Delivery Data Insights project includes detailed attributes of Grab establishments, such as names, locations, cuisines, financials, and operational data. It provides a solid base for market analysis and identifying food delivery industry expansion opportunities.

## Fields:

| **COLUMN NAME** | **TYPE** | **DESCRIPTION** | **Example** |
| --- | --- | --- | --- |
| id\_source | (String) | Unique identifier for the data source. | SGDD01085 |
| name | (String) | Name of the establishment or restaurant. | Texas Chicken |
| address | (String) | General address of the establishment. | Vivocity |
| country | (String) | Country where the establishment is located. | Singapore |
| cuisine | (Array of Strings) | Array of cuisine types offered by the establishment. | ["Western", "Fast Food", "Halal", "Chicken", "...] |
| currency | (String) | Currency in which the prices are listed. | SGD |
| delivery\_cost | (Integer) | Cost of delivery in the establishment’s currency. | 570 |
| lat | (Float) | Latitude coordinate of the establishment’s location. | 1.264792 |
| lon | (Float) | Longitude coordinate of the establishment’s location. | 103.822103 |
| opening\_hours | (JSON Object) | establishment’s opening hours for each day of the week and a boolean indicating if the establishment is open. | {"open": true, "displayedHours": "10:30-21:30"... } |
| image\_url | (String) | URL of an image representing the establishment. | https://d1sag4ddilekf6.cloudfront.net/compress… |
| radius | (Integer) | The delivery radius in meters within which the establishment delivers. | 5000 |
| rating | (Float): | Average customer rating of the establishment. | 3.8 |
| reviews\_nr | (Integer): | Number of reviews the establishment has received. | 1591 |
| delivery\_options | (String): | Delivery options available, such as ‘ONLY\_DELIVERY’. | ONLY\_DELIVERY |
| promo | (String): | Promotional text or codes offered by the establishment. | Use 'TEATIME' for FREE delivery between 3PM to.. |
| loc\_type | (String): | Type of location, such as ‘FOOD’. | FOOD |
| delivery\_by | (String): | Entity responsible for the delivery, such as ‘GRAB’. | GRAB |
| delivery\_time | (Integer): | Estimated delivery time in minutes. | 30 |

# 3) Methodology/Approach

## Teamwork Approach

Our project team decided on an agreed and realistic scope that we can achieve collectively. After researching, it was decided that we will extract a dataset from Kaggle.com. After extraction, the team split up to transform on different parts of the dataset using both Jupyter and Visual Studio Code. Collaborating on Github, eventually, the team combined the individual codes together to load the dataframes on PostgresQL.

## Project Workflow

****

## Sourcing for datasets

Team discussion of datasets suggestions. Each member presented 1 dataset for discussion.

Hassan – Grab dataset

restaurant owner, grab driver, can view our analysis and see where is lacking, and expand in the location.

Saila – Graduates Employment dataset

Employment rate, salary, correlation, ranking, trend

Simpler set, faster

Hong Eng – Microsoft Stock Dataset

Technical analysis on stock trend to for prediction of future stock price

Good for investors

Rosni – Grab dataset

same points as Hassan

After discussion, the team decided on the 16000 Grab Dataset because the team aspired to use the dataset to answer business questions and benefit both individual business owners and Grab’s overarching strategic goals.

## Deliverables Assignment

Documentation > Hassan, Hong Eng

Presentation Deck > Saila, Rosni

## Tools

The team used various tools for coding, collaboration and visualization. Namely,

* Python version 3.11 or higher
* Postgresql
* VS code
* Git
* Github
* Power BI
* Canva

Libraries

* SQLAlchemy: for SQL database interaction and management.
* Nbformat: for working with Jupyter Notebook file formats.
* NumPy: for numerical computing, particularly for handling arrays and matrices.
* Ast: for parsing Python abstract syntax trees (AST).
* Json\_normalize: for normalizing semi-structured JSON data into a flat table.
* Psycopg2: for interacting with PostgreSQL databases from Python.
* Re: for regular expression operations in Python.
* Folium: for creating interactive maps in Python, particularly with Leaflet.js.

# 

# 4) ETL Process

## Extract

The first step involved extracting data by reading API from kaggle dataset on 16000 Grab Restaurants. The dataset was extracted into CSV format.

Extraction was done via a KaggleApi() to download the csv dataset from the 16000-grab-restaurants-in-singapore kaggle dataset.

* prerequisites:
  + Login to your kaggle account>>settings download the API token, this will download the kaggle.json file.
  + Create a directory called .kaggle in C:\Users\<Windows-username>\.kaggle\
  + Save your kaggle.json file that you downloaded into that directory
  + pip install kaggle in the python code
  + Now you can run the code to download the dataset, the dataset will be download to path location C:\Users\<Windows-username>\Documents\grab\_restaurant\_analysis\\data
  + This path location will be where the python code are stored when you git clone the files C:\Users\<Windows-username>\Documents\grab\_restaurant\_analysis\\

(Attachment: api\_kagglecall.ipynb or access on [Github here](https://github.com/HassanHosain/grab_restaurant_analysis/blob/main/api_kagglecall.ipynb).)

## Transform

The transformation phase was critical to ensure the quality and consistency of the data. This involved several sub-steps.

The team split up to transform on each of the 4 different parts of the dataset by Cleaning/Normalizing/Formatting including reconciling different naming conventions in team members’ codes to ensure consistency.

### establishment\_df: (*by Saila*)

* + Data Cleaning on 'address' column.
  + Which involved data inspection, Data correction, Data standardization and Data transformation.
  + Removed duplicates - the address column values were keyed in together with the establishment name separated by "-".
  + Standardized the wordings - values had inconsistent cases. Example : 'Vivocity' , 'VivoCity'.
  + Inconsistent entry - for the same address that was entered inconsistently. Example: 'Vivocity', 'Vivo City' , 'Vivocity Shopping mall'.
  + Filtered and transformed the address entries to avoid unnecessary distinct values.
  + The distinct values were successfully reduced from original 4883 to 3971.
  + Standardization is currently limited to the alphabet 'a'. Moving forward, more comprehensive cleaning efforts are necessary to enhance address standardization, allowing for better categorization and grouping.



### cuisine\_df: (*by Hassan*)

* + Selected columns id\_source, cuisine from the main dataframe to be transformed into a new dataframe cuisine\_df.
  + Data Cleaning of null values in the cuisine category and replaced the values.
  + Cleaned up the cuisine column to remove commas, apostrophes, [] and set it lower case.
  + Sorted the cuisine columns in alphabetical order A-Z.
  + Exploded the cuisine category into multiple columns i.e. cuisine\_1, … , cuisine\_5.
  + Created a new primary key called cuisine\_id as cus\_1, cus\_2 .. etc.
  + Did a check to explore the top 3 popular cuisine category.



### 

### delivery\_df: (*by Rosni*)

* + Selected the following column from the main data frame : id\_source, delivery\_cost,radius, delivery\_options , promo.
  + In the promo columns, all the missing values are replaced by the string “no promotion available”.
  + There were test restaurants with no names, so these rows were dropped..
  + Split the delivery column into 3: delivery, take away and dine in.
  + Added 3 columns with boolean values False.
  + Checked if the delivery column contained the words ’delivery’,’take’, ‘dine’ using .stri.contains( ) function.
  + Using .loc on corresponding column, update to True if the words were present.
  + New index added which will act as primary key.



### 

### openinghour\_df: (*by Hong Eng*)

* + - Subsetting opening hours column.
    - Cleaning opening\_hours which consisted JSON objects/ dictionaries.
    - Expanded the dictionaries into 10 separate columns:

- Using ast module’s ast.literal\_eval function to ensure that the column is in a proper list of dictionaries format.

- Using json\_normalize function to convert the dictionaries into separate columns.

* + - Added index column with prefix.
    - Set new index column to be the first column.
    - Investigated null/NaN values and replaced with relevant values.
    - There were 12849 - 7155 = 5694 establishments with open==NaN and tempClosed==NaN.
    - Based on checking on first 5 establishments at time of interim project, they are currently open, meaning they were not permanently closed at time of dataset. Therefore, for sake of analysis in our project, it is assumed that all entries with open=NaN and tempClosed=NaN are not closed permanently and therefore included in the analysis of the dataset.



## 

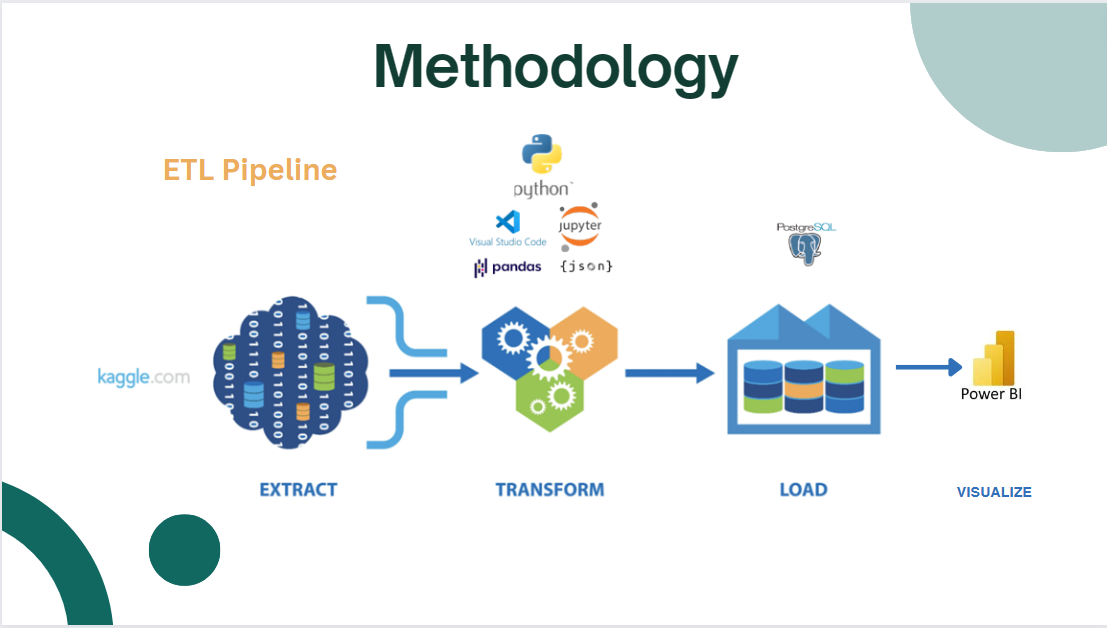
## Load

The final step was to load the transformed data (4 dataframes) into a structured database. We chose PostgreSQL, a relational database management system (RDBMS) for its robustness and ease of querying.

4 tables were created:

1. establishment\_tbl
2. cuisine\_tbl
3. delivery\_tbl
4. openinghour\_tbl

## ETL Pipeline Diagram



## Database Schema/ERD Diagram

## 

**Table Structure**: The final data was structured into a relational database with tables for establishment, cuisine, delivery and opening hours.

**Data Validation**: Validation checks were implemented to ensure data integrity and accuracy throughout the ETL process.

# 5) Challenges (Solutions/Workarounds/Limitations)

## Challenges

* Collaboration: Unfamiliarity and difficulty in using Git Bash  
    
  **Solution:** Used VSCode where commands like committing pulling can be done in a UI method
* Loading to PostgreSQL: Primary Key and Foreign Key were not created on tables.  
    
  **Solution**: Revisited the code on creation of tables to include primary and foreign key references constraint.
* Tech-tools: Unfamiliar with Python libraries. Difficulties combining several notebooks into one unified file.

**Solution**: Leveraged Nbformat to merge disparate files into a cohesive notebook.

* Time: Very limited time to learn new techniques and tools.

**Solution**: : Effective collaboration and teamwork.

## Limitations

Last update of the dataset on kaggle was 2022 which is 2 years before the time of the interim project. Therefore it does not represent the present.

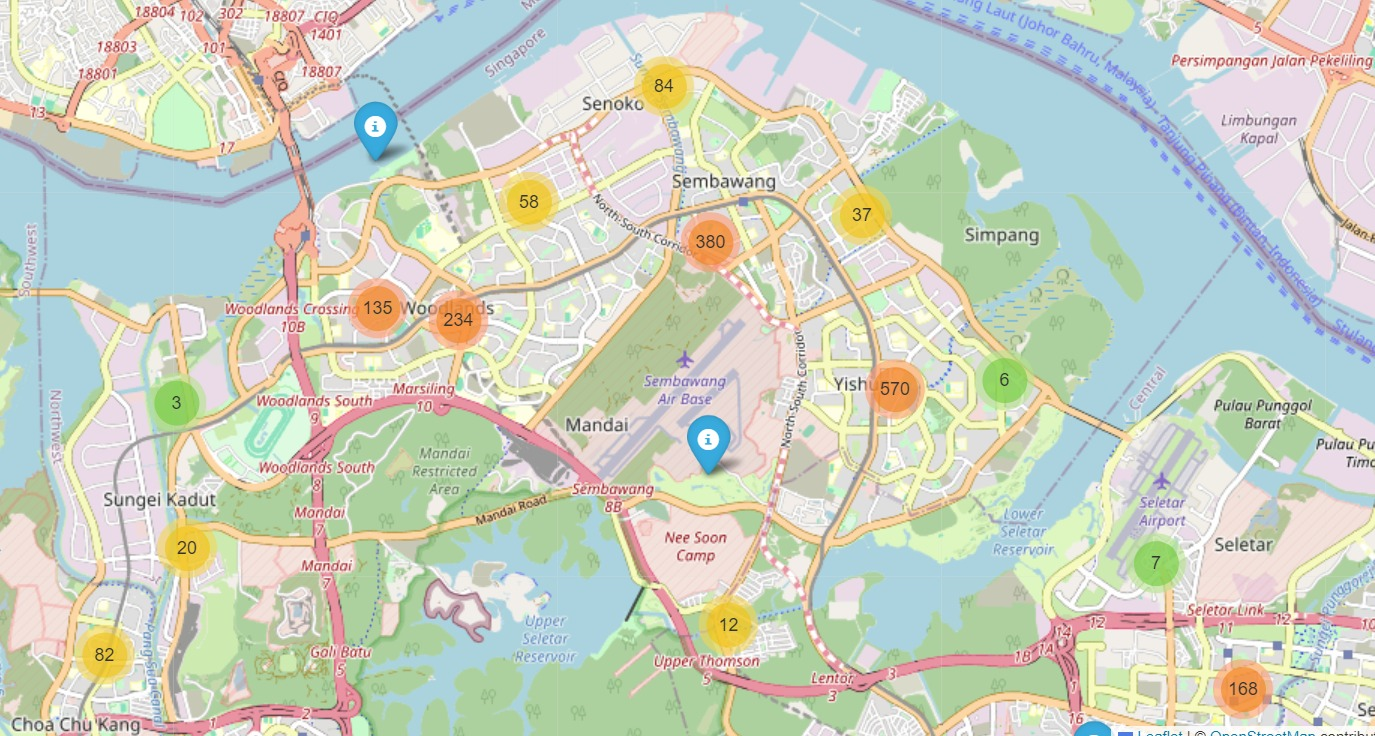
The dataset did not include customer transaction data, limiting the depth of analysis regarding customer behavior and preferences.

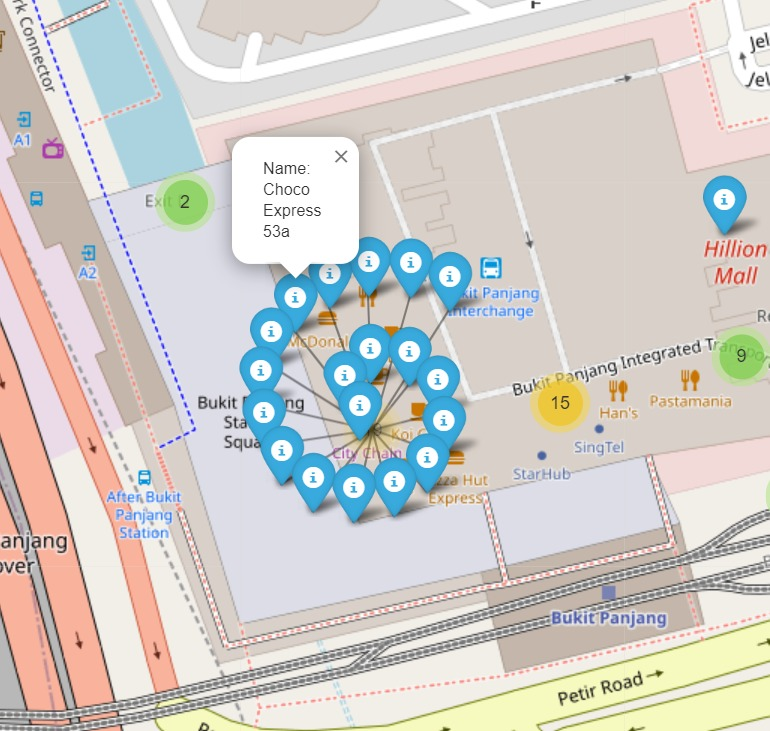
# 

# 6) Findings/Analysis

## Visualization - Interactive map

Distribution of establishments across Singapore .

****

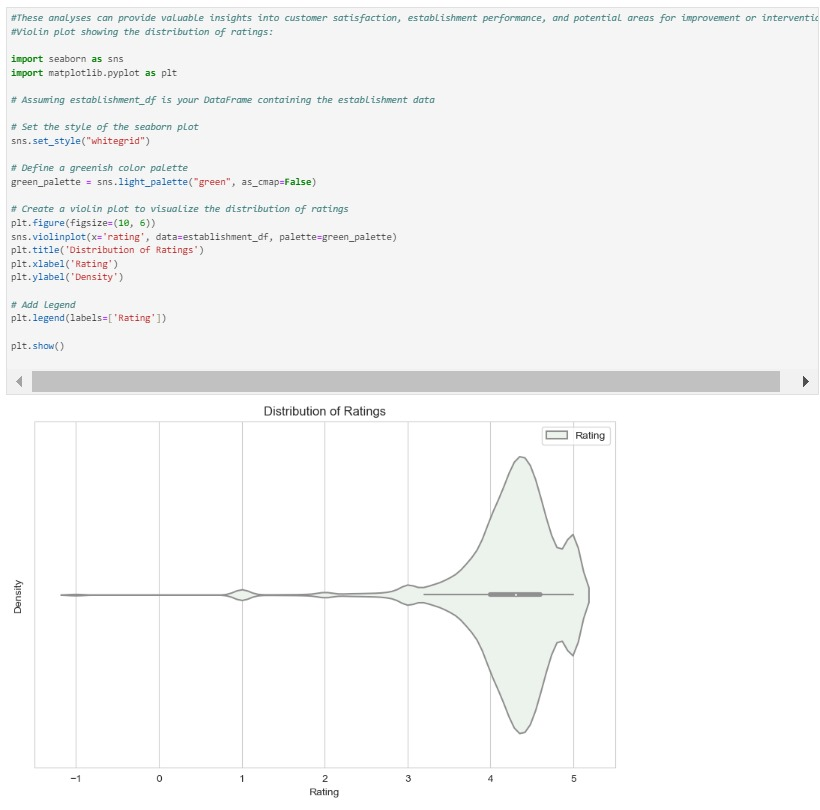


Map with marker cluster layer which prevents overcrowding of the markers while viewing.

## 

## 👉 [Click here for a video of the interactive map](https://drive.google.com/file/d/1sRycdffU456nIePPouAO76qxYHYn86pn/view?usp=sharing)

## Visualization - Violin plot of distribution of ratings

****

These analyses can provide valuable insights into customer satisfaction, establishment performance, and potential areas for improvement or intervention.

## Visualization - Bar plot of Avg Rating by Delivery options

****

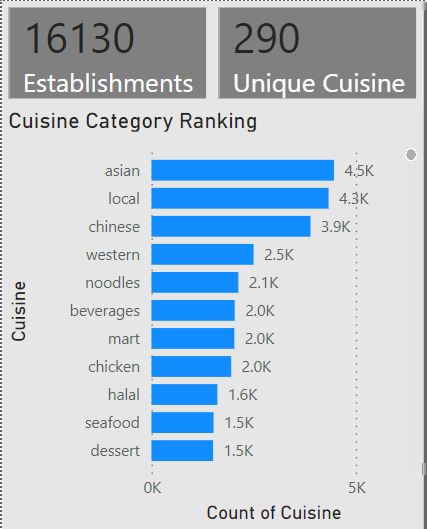
To compare ratings based on delivery type, you can follow a similar approach as before, but this time you'll group by delivery options by grouping the merged data frames.

## 

## Visualization - Cuisine Category Ranking

(total number of establishments in each cuisine category)

(Only the top 10 is shown. Refer to attached Visualisation.pbix for full ranking)

****

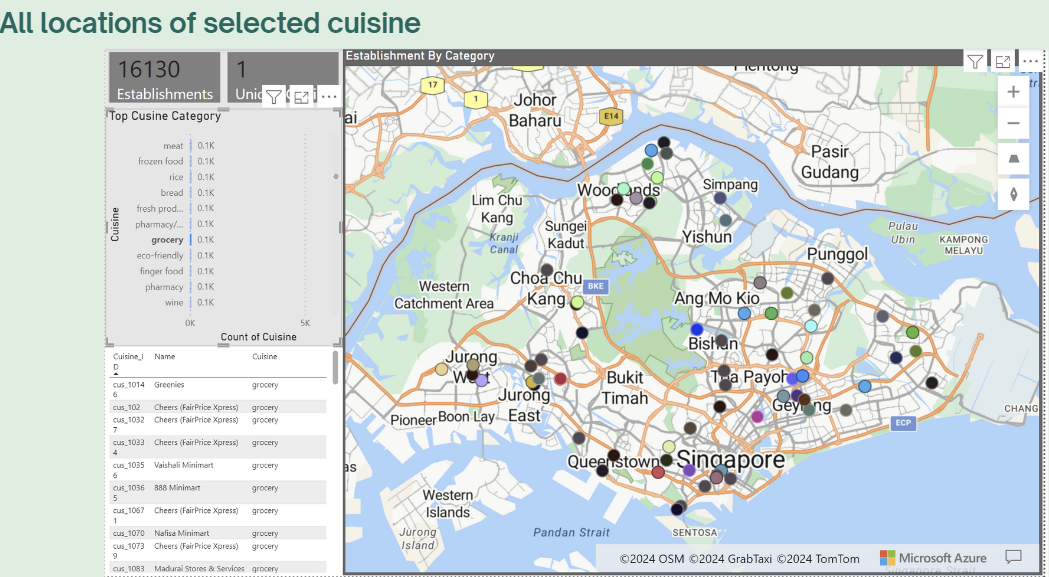
**From prospective business owners’** **perspective**, the higher the ranking, meaning the higher the number of establishments of each category, the more competitive it is. This will offer them the insight to consider which cuisine category is optimal for their business by considering the less competitive category.

**From** **Grab’s perspective**, this analysis serves as a strategic tool for expansion and competition. It will enable Grab, as well as its competitors, to identify underserved regions and cuisines, thereby informing decisions on which cuisine category to focus expansion efforts to capitalize on market opportunities.

## 

## Visualization - All locations of selected cuisine

(example: “grocery” category)



**From prospective business owners’** **perspective**, this will offer them the insight to identify optimal locations for new establishments based on the type of cuisine category selected. By examining areas with a shortage of certain cuisines, business owners can know decide where to start new establishment to meet unfulfilled market demand in that location. For example, looking at the location of Yishun and Simpang, there are only 2 grocery establishments. Starting new grocery establishment at these location will have less competition.

**From** **Grab’s perspective**, this analysis serves as a strategic tool for expansion and competition. It will enable Grab, as well as its competitors, to identify underserved region, thereby informing decisions on which location to focus expansion efforts to capitalize on market opportunities for a particular cuisine.

# 7) Next Steps/Conclusion

**Next Steps**

Further analysis could include customer transaction data (if data is available) to gain deeper insights into purchasing patterns and preferences.

**Conclusion**

The ETL Pipeline for Grab Delivery Data Insights project will provide prospective business owners valuable insights into unfulfilled market demands in terms of optimal location and cuisine category

The ETL Pipeline for Grab Delivery Data Insights project will enable Grab, as well as its competitors, to identify underserved regions and cuisines. It serves as a strategic tool for expansion and competition, thereby informing decisions on where to focus expansion efforts to capitalize on market opportunities.

Overall, this analysis presented in this project will help to answer business questions and benefit both individual business owners and Grab’s overarching strategic goals.

*THE END*