Best State to Open Restaurants

Identification and Preparation of Data for Analytics

By Bootcamp Grp X

**TABLE OF CONTENTS**

1. Introduction
   1. Summary
   2. Scope
   3. Technologies and resource contributions
   4. Definitions, Acronyms and Abbreviations
2. ETL Details
   1. Data Import/Extract Sources and Method
   2. Data Acquisition
   3. Data Transform
   4. Data Integrity
   5. Data Refresh Frequency
   6. Data Security
   7. Data Loading and Availability
3. Data Quality
4. Programs
5. Appendix

|  |  |
| --- | --- |
| **1.** | **INTRODUCTION** |

*The purpose of the Extraction, Transformation, and Load (ETL) Technical Report is to capture details that pertain specifically to ETL portion of the data pipeline that is to be used in a data science project. This however does keep in mind the final target objective while performing the ETL.*

In addition to chosen data sources, available input data formats, and the data storage engine for the ETL, the programming environment, software packages, and developed software are described below.  
Comments on different aspects of data retrieval describe the limitation of usage as well as possible solutions for improvement of usability of retrieved data including data maintenance.

# 1.1 Business Summary

This section summarized the final objective of the project, the business problem definition (problem statement) and the expected outcome of ETL.

The client, a restaurant group offering multiple types of cuisine within its wholly owned holding company is looking to open ten new restaurants within a single state. The client would like to open the locations based on high consumer spending patterns, low market saturation for that type of cuisine, and high ratings for that restaurant category.

The ETL deliverable should allow the restaurant group to view the data to foster a decision in the restaurant openings analysis. Moreover, the client should be able to test at least one hypothesis.

**Hypothesis**

Based on consumer spending, market saturation, and restaurant ratings, one state for each restaurant category is better to open ten restaurants than the remaining forty-nine states.

Null Hypothesis: There is no difference between the fifty states in terms of consumer spending, market saturation, and the restaurant ratings variables.

ANOVA testing could be utilized to test the fifty-state sample for each of the three variables. If the data was cleanly available for each year, a regression analysis could be used to determine the impact of each variable on the other to further predict if a state under consideration might be better suited based on trend (or in the case if a handful of states bubble to the top of the analysis).

# 

# 1.3 Scope

This section explicitly outlines the disparate data sources that are to be integrated, which components of the overall data science project is in the scope for this initiative and lists out the components of the data science project that are not in scope here.

The data sources within scope of this project include Yelp.com, BEA.gov, and other applicable files on the internet containing state related data such as populations and abbreviation conventions. The years under consideration are limited to the available data. For instance, BEA has consumer spending data distributed in the restaurant category only for 2017 with the ability to calculate prior year 2016. Population data and overall consumer spending date back three full years. Conversely, the data from Yelp is current as opposed to historical. For our restaurant group client to decide on the state’s locations, inflation multipliers have to be utilized to transform the historical data to estimated current data as well as retrograde data to coincide with prior years. These inflation multipliers for predictive analysis are out of scope for the project deliverable. Additionally, other components that have been ruled out of scope during the project discussion include the following: (1) employment data by state, and (2) personal income by state as each were deemed loose derivatives of discretionary restaurant spending.

# 1.4 Technologies and resource contributions

This section lists out the team members and their contributions towards the ETL initiative. Use this section to also outline (or list) the tech stack used to obtain the final outcome.

The tech stack includes the following: PostgreSQL, Python, Jupyter Notebook, SQLAlchemy, API calls, CSV files, and various Pandas Libraries to support the development process.

Tessy Azogu (Data Sources, Report Generation)

Martin Hrbac (Database Development & Report Generation)

Sunil Joshi (Statistics Themes, Data Sources, Data Analysis, Report Generation)

Denise McKinley (Data Sources, Report Generation)

Ninglian (Melanie) Wang (Database Development)

# 1.5 Definitions, Acronyms and Abbreviations

List acronyms and terms that need to be defined in this section, such as ETL: Extract, Transform and Load

Market Saturation: Number of restaurant locations in a state by category

PCE: Per Capita Expenditure

FIPS: Federal Information Processing Standard (one- or two-digit code to uniquely identify to each state)

|  |  |
| --- | --- |
| **2.** | **ETL DETAILS** |

*This section outlines a more detailed description of the processes utilized/proposed to achieve the objectives of this initiative.*

# 2.1 Data Import/Extract Sources and Method

section .

This table provides information about the data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data Source | URL | Format | Dimensions | Data Updates | Data Loads | Permissions |
| State | <https://www.nrcs.usda.gov/wps/portal/nrcs/detail/?cid=nrcs143_013696> | MS Excel | Two | Static | Once |  |
| State Population | [https://www.census.gov/newsroom/press-kits/2018/pop-estimates-national-state.html](https://www.census.gov/newsroom/press-kits/2018/pop-estimates-national-state.html" \t "_blank) | MS Excel | Three | Annually | Availability of New Data |  |
| State PCE | <https://www.bea.gov/data/consumer-spending/state> | MS Excel | Five | Annually | Availability of New Data |  |
| Yelp | <https://www.yelp.com/developers> | JSON | Nine | Daily to API Restriction Maximum | Minimum For Each Analysis | 5,000 API calls per 24 hours, resetting every midnight UTC |

# 2.4 Data Acquisition

This section outlines the data needed, such as range and if the data is static or dynamic and needs continuous update. Outline the process to obtain again or update the dataset. The formatting and any special attributes about the data the one should be mindful of while obtaining and processing the raw dataset. How to decide on the selection of data while re-obtaining or updating. Discuss, here the dimension of the obtained dataset and if updated what is the project growth rate of the data. Lastly, address any issues or pre-requisites that needs to be cleared prior to getting the data?

For the state, state population, state personal consumption expenditure sources, the MS Excel files are retrieved from their respective URL’s. The restaurant data from Yelp is retrieved using API calls. Only two tabs from the state personal consumption expenditure MS Excel file were selected to generate the state pce data – table one containing total personal consumption expenditure from 2015 to 2017 and table four containing per capita personal consumption expenditure for the “off-premises food and beverages” category for 2017. There is no immediate concern for growth of any of datasets outside the number of new restaurant openings depicted on Yelp. Each restaurant opening adds a row. Closures are handled as a simple Boolean replacement from “true” if the restaurant is open to “false” if closed so the number of records does not grow.

# 2.3 Data Transform

In this section address any data transformation that needs to be performed to modify, clean, filter or create existing and new parameters. Address any technical analysis performed, include design specification or data models used (example linear interpolation etc.), and any calculations performed for any newly derived fields.

Of the six tables in the relation schema depicted in section 2.7 Data Loading and Availability below, the data in five are mainly imported from data sources. The remaining table restaurant\_category is a relational table between the restaurant table and the category table from Yelp. For all the tables, a new parameter named “id” was created as a surrogate key to uniquely identify each row of data. For every table excluding the state table, a new parameter named “modified\_date” was created to let the user know when that data was loaded. A “source id” column was constructed for all tables excluding the state table to inform the user of the data source – 1 for Yelp, 2 for State PCE, 3 for State Population, and 4 for State. Another interesting twist to data transformation was taking the “price\_range” in the restaurant table and converting strings ranging from “$” to “$$$$” to integers “1” to “4”.

Other notable items include keeping the FIPS number as a unique identifier in the state table and keeping the “alias” as a natural key in the category table.

category Table – keep alias with small letters

restaurant Table – merge address streets into one column, keep business id as unique (surrogate key), Price range can be null, price range is transformed from $-$$$$ to integer, The category has been moved to a separate (dimensional) table (Categtory), a restaurant can have one or more categories.

restaurant\_cabinet Table – connect the a.m. table

state table – as a primary key we use FIPS number

state\_pce – just numbers, we have only one year data since the detail split into category started only in 2015, see their web site

state\_population – just numbers

# 2.4 Data Integrity

In this section discuss the reliability of the extraction source data (e.g., missing data, dates stored as text, invalid code values, text fields with odd characters, etc.). Address the frequency with which the data sources are updated and if it is necessary to update the local data at the same frequency. Lastly, how if any notification can be received when the source data is updated; and what if any notification will be sent to the internal team when the local dataset is updated.

# See the table in 2.1 Data Import/Extract Sources and Method for details.

# 2.5 Data Refresh Frequency

This section explicitly lists the frequency with which this ETL process will refresh the local dataset (Daily, Weekly, Monthly, Quarterly, Semi-Annually, etc.).

# See the table in 2.1 Data Import/Extract Sources and Method for details.

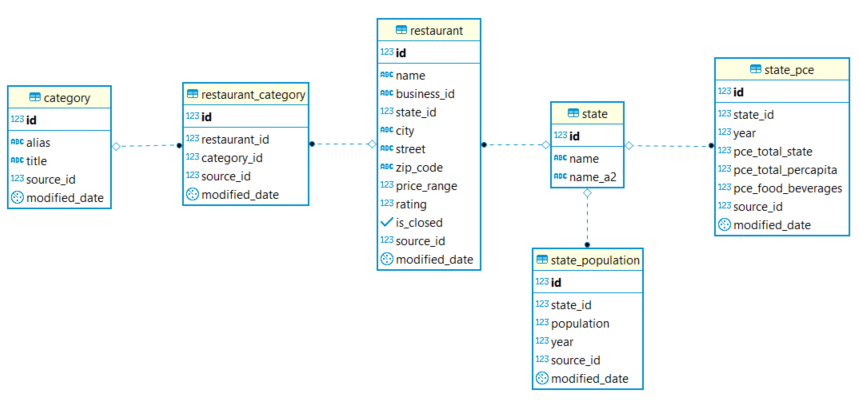
# 2.6 Data Security

There are no issues associated with data security for government and restaurant data for this application.

# 2.7 Data Loading and Availability

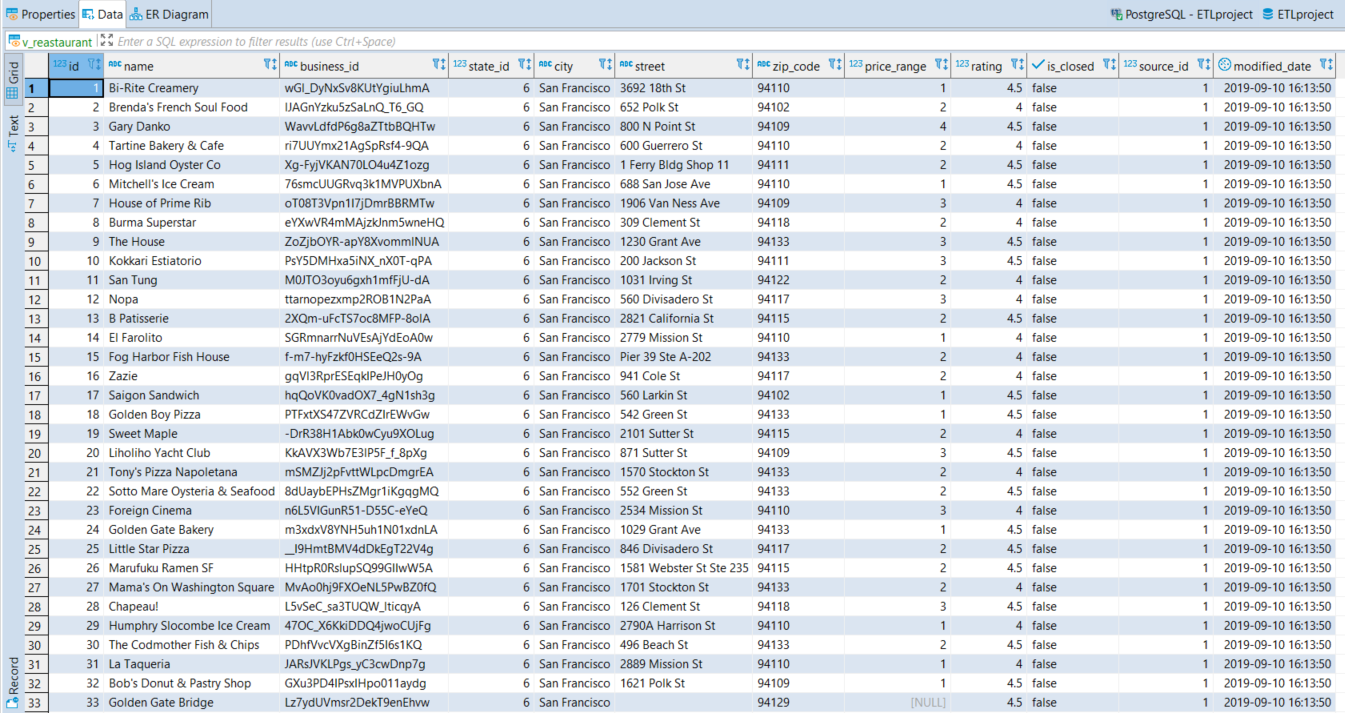
This section addresses the data schema and during of data retention. Discuss the interface that will allow your Client/Users to access the data.

**Database Schema**

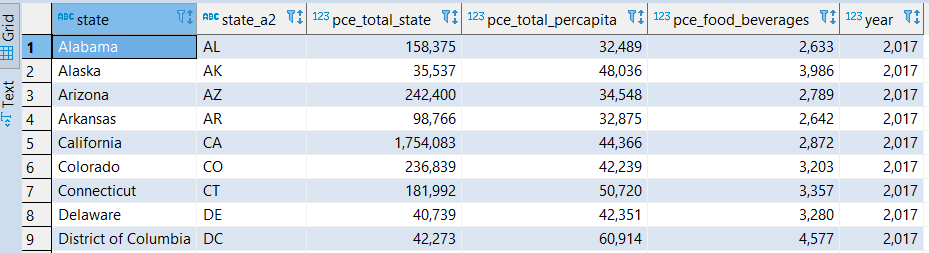


**Database views** have been created in the PostgreSQL database for the client to query data although stored procedures with a secure GUI could serve as a better alternative.

**Restaurant View**



**State PCE View**



|  |  |
| --- | --- |
| **3.** | **DATA QUALITY** |

Address in this section success criteria for this project. Summarize the parameter KPIs such as Totals and expected counts. What user acceptance testing was performed and what were the outcomes. What is the recommended site acceptance testing that your client can perform to ensure the expected outcomes meets their expectations?

Data quality is primarily addressed by input data verification in the loader and by the storage database design, using strict data normalization, and proper data types. The database has a built-in tracking mechanism which holds information about the data source and the date when the values were entered.

|  |  |
| --- | --- |
| **4.** | **PROGRAMS** |

For the original data load and testing 3 Jupyter notebooks were created as follows:  
•    ETL\_StatesData.ipynb – loads state data from Excel files into database,  
•    API\_yelp.ipynb – pulls data from Yelp API, stores data in local files, and  
•    ETL\_RestaurantData.ipynb – loads data from the files, loads it into the database

4.1 State data load

4.1.1 State Basics  
This table is comprised of state name, state id, and two characters state code. This data is considered static and is loaded once only.

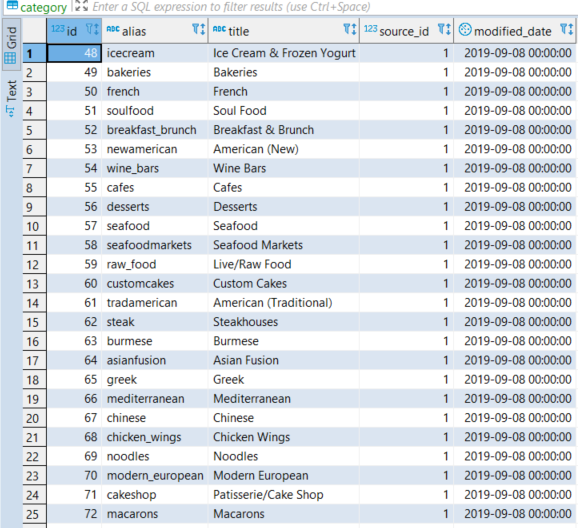
4.1.2 State PCE  
New data is released once a year. The data should be loaded after new data is released.  
Note: For the purpose of the project this data is considered static.

4.2 Restaurant data acquisition – Yelp API call  
This program utilizes and calls Yelp API and stores the raw data into local files. No data manipulation is done by this program. The data is stored in a JSON format.

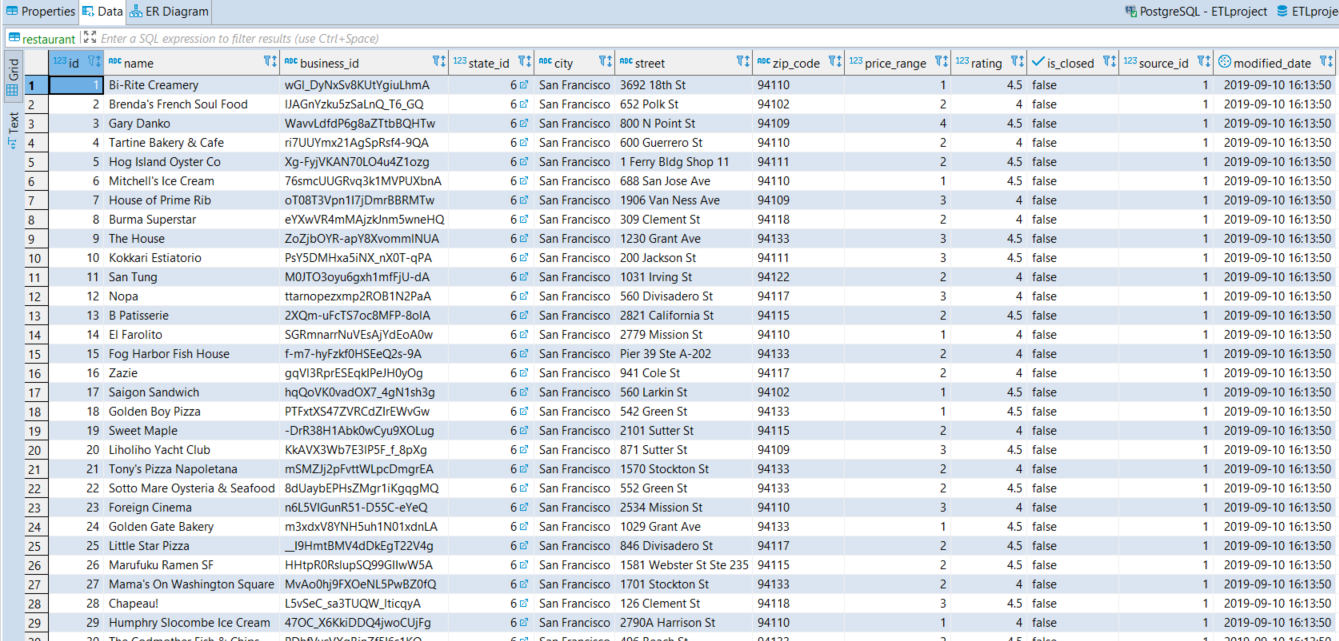
|  |  |
| --- | --- |
| **5.** | **APPENDIX** |

**Tables Snapshots**

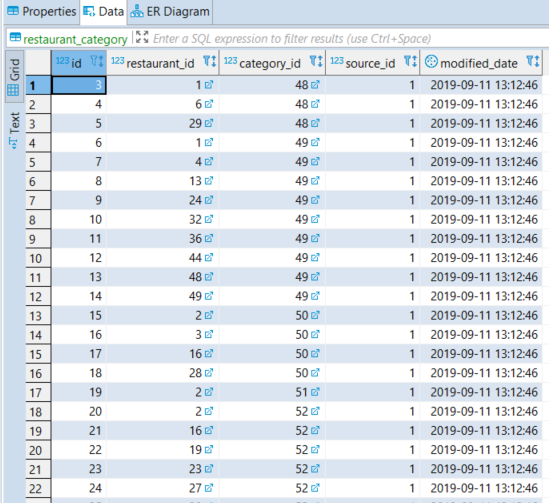
Category Table



Restaurant Table



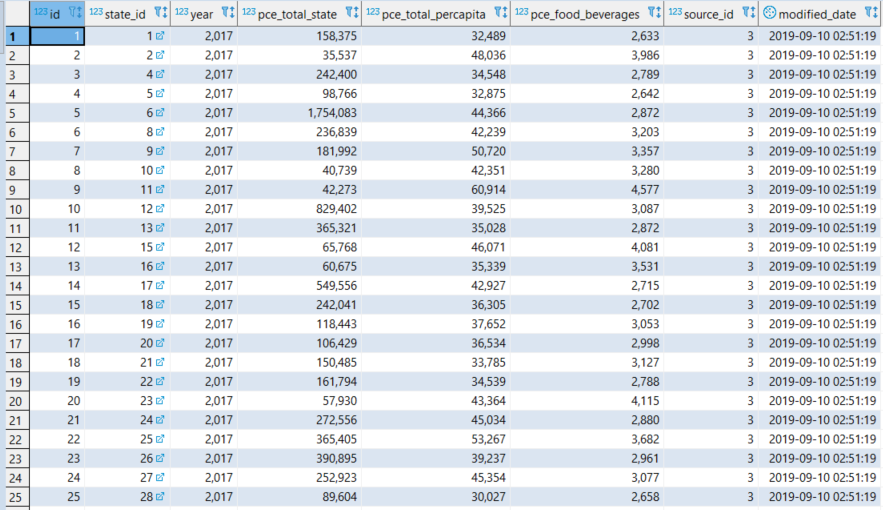
Restaurant\_Category Table



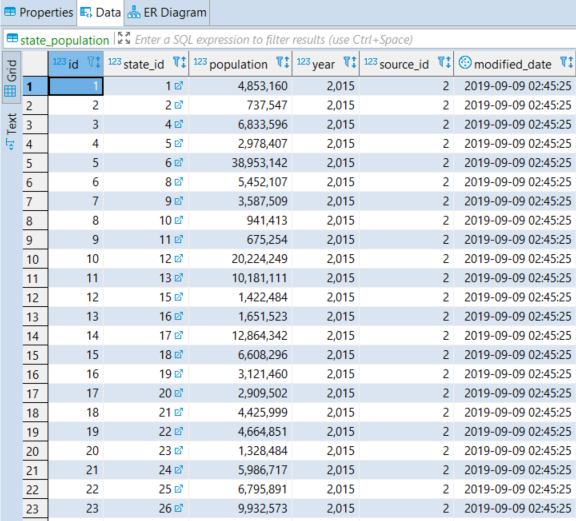
State Table



State\_PCE



State\_Population



**Reference for Create Database SQL:**

[https://github.com/marsion0245/GTATL5Project2/blob/master/Code/CreateDatabase.sql](https://github.com/marsion0245/GTATL5Project2/blob/master/Code/CreateDatabase.sql" \t "_blank)

**Reference for Create View SQL:**

<https://github.com/marsion0245/GTATL5Project2/blob/master/Code/CreateView.sql>