Best States to Open Restaurants

Identification and Preparation of Data for Analytics

By Bootcamp Grp X

**TABLE OF CONTENTS**

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

|  |  |
| --- | --- |
| **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.*

# 1.1 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).

# 1.2 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.

# 1.3 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: 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, Data Analysis, Report Generation)

Denise McKinley (Data Sources, Report Generation)

Ninglian (Melanie) Wang (Database Development)

# 1.4 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

4 data sources

1. State – static data, loaded once only
2. State population – changes one a year, shall be reloaded when new data is available
3. State PCE - changes one a year, shall be reloaded when new data is available
4. Restaurant – changes frequent, data should be reloaded before new analyses is performed

This section provides information about the data and its source. For example, API names and URLs, key parameters available and its subset which will be preserved (loaded). Data extraction protocols (API, FTP, Web scraping etc.), any permissions required to access the said extraction dataset and any restriction placed on the usage and distribution of the acquired dataset.

# 2.2 Data Acquisition

Frequency - see above

Data formats:

1. Excel
2. Excel
3. Excel
4. json

To get new data set

1. Download Excel from …
2. Download Excel from …
3. Download Excel from …
4. Json – call API, …

Be aware of restriction on Yelp – write details

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?

# 2.3 Data Transform

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

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.

# 2.4 Data Integrity

As discussed, 3 type of data as far as update, the most volatile data should be updated before analysis is performed

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.

# 2.5 Data Refresh Frequency

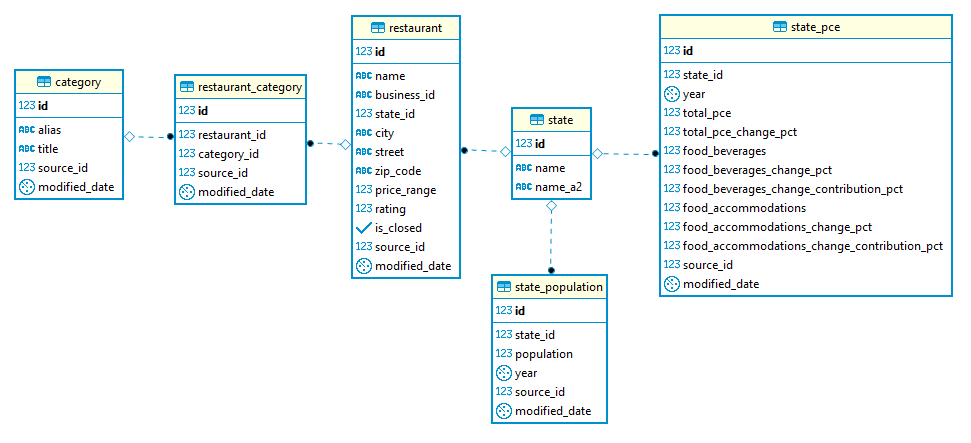
This section explicitly lists the frequency with which this ETL process will refresh the local dataset (Daily, Weekly, Monthly, Quarterly, Semi-Annually, etc.).

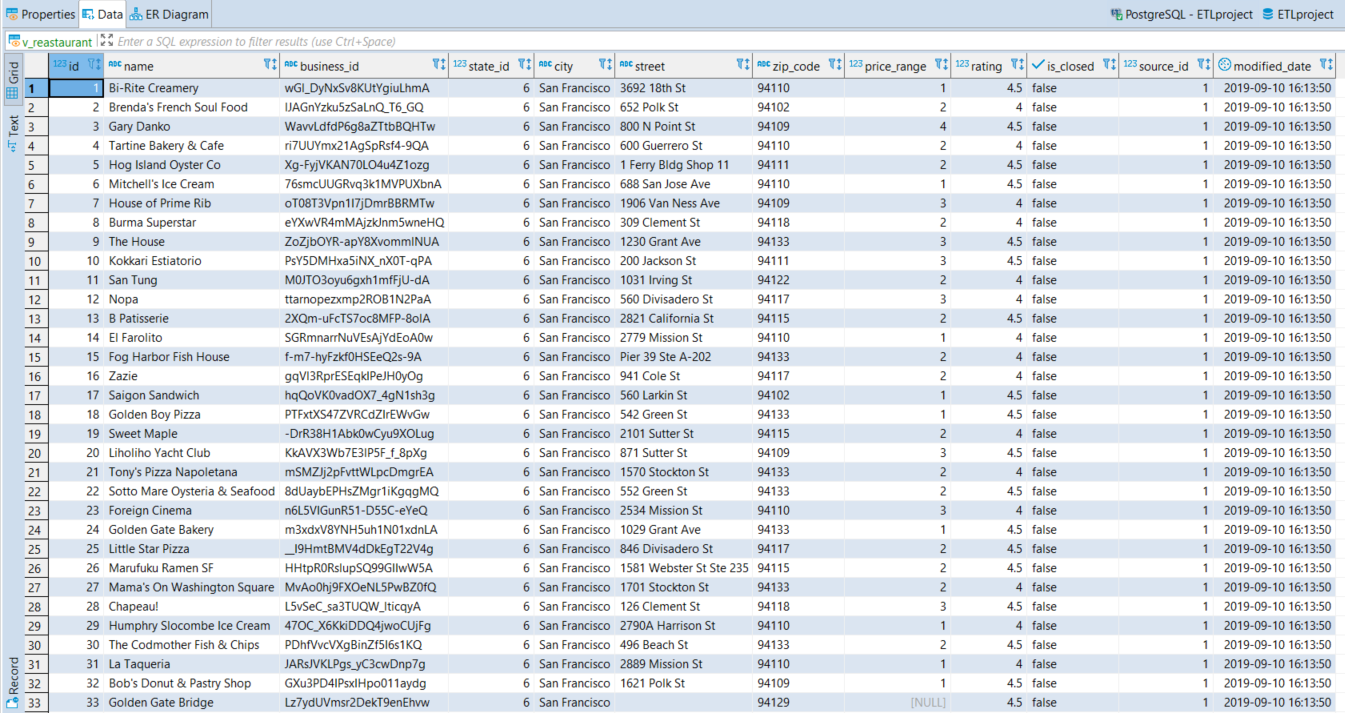
# 2.6 Data Security

This section discusses any data anonymity and security requirements need to be satisfied. Address any federally mandated HIPAA considerations, any need to build in additional privacy, Encryption, Data masking, Auditing, Backups etc.

# 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.





|  |  |
| --- | --- |
| **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?