prn group

*ETL Technical Report*

A technical summary of the ETL process demonstrated by PRN Group to uncover relevant business insights using data from Yelp’s API and the NY Department of Health and Mental Hygiene to understand more about the NYC Restaurant landscape.

Pooja Jaisinghani, Rosarys Esquilin, Nick Bredahl

10/14/2019

prn group

# **INTRODUCTION**

PRN Group was established in October 2019 by co-founders Pooja, Rosie, and Nick. Our business is driven by the mission to help companies thrive with data-driven decision making. We adapt each project to address the unique challenges with a thorough understanding of the data landscape and optimize accordingly. In this report, you will find an example of an ETL Process report that was used to dive deeper into the connections between consumer reviews of restaurants and NYC Inspection ratings with the goal that this will uncover predictive insights on what restaurants will remain open or shut down. The data obtained from the Yelp API and NYDHMH went through a thorough ETL process consisting of cleaning, filtering, joining, transforming and aggregating the datasets and then loading them onto relational databases such as PostGresSQL to create, update, and query connected database tables. A deeper dive into this process is broken down below:

## **STAGE 1: EXTRACTING THE DATA**

The process began with initial research on what business problems we could solve, and whether we could gather credible and relevant data to address the ask. We used Kaggle and found a dataset from the New York Department of Hygiene and Mental Health, which contained violation citations from restaurants and college cafeterias for the past 3 years for up to 24,000 NYC establishments. “Inspectors check for compliance in food handling, food temperature, personal hygiene, and vermin control. Each violation of a regulation gets a certain number of points. At the end of the inspection, the inspector totals the points and this number is the restaurant's inspection score — the lower the score, the better the grade.” (NYDHMH) We believed that we could use this data to connect it to the public perception of restaurants by analyzing restaurant reviews and see if we could find a correlation that would explain factors that affect a business closure. Due to this hypothesis, we then sought to obtain data from Yelp on the same restaurants in NYC that were inspected by using the Yelp API, and implementing parameters to collect the data from restaurants in New York, we believed we could connect the two by using Jupyter notebook. To extract the data onto Jupyter notebook, each dataset had its unique nuances to address. Pulling data from the Yelp API was difficult because we had to create two different codes to extract the data based on the users' computing system. For MAC users, we used the requests library and then had to store outputs into a list, with our specified parameters, and then convert it into a data frame. For Windows users, we sent the API data to a JSON file instead to eliminate object error issues when converting into a data frame. To extract the data from Kaggle/NYDHMH, we used the “pd.read\_csv” command and converted data into a data frame in Jupyter notebook.

## 

## **STAGE 2: TRANSFORMING THE DATA**

Once we extracted all the data into Jupyter/Pandas data frames, we then focused on cleaning and transforming the data as the raw data is not usable in its original form. This step is crucial in the ETL process, as this is how we can add value and change our data to produce relevant insights and seamlessly bring into our relational databases to manipulate as needed. To clean both of our datasets, we used the following codes:

* Filtering the data so we only were using relevant columns. In this step, the Inspection Dataframe went from 26 columns to 6 by creating a new smaller data frame, and the Yelp API data frame was filtered through the parameters created during the JSON requests process and specific API call documentation.
* Each data frame then had to be cleaned to drop any null values and any duplicate values as we would not be able to use SQL to connect the databases if there were multiple rows of the same restaurants since this would be our primary key.
* Since the data was pulled in different formats, we had to match the character formatting by converting all restaurant names in the Yelp data frame to uppercase to match the DBA column formatting in the NYDMH Inspection data.
* We also re-named columns so that corresponding columns were matching so that data could then be merged.
* In this stage, we also mapped out the relationships between the datasets <https://app.quickdatabasediagrams.com/#/> and appointed primary and foreign keys to connect the data into the Postgres database.

## **STAGE 3: LOADING THE DATA**

The third and final stage in the ETL Process is to load our data into a final database which can be used for storage until future analysis for business use. Since we mapped out the relationships between our datasets, we can also extract a PostGres SQL code text from this website to run on the Project Database that we created in the RDBMS. In the database schema, we use the query tool to then create pre-set tables with columns that specify the data types so that we can then bring in our datasets from Jupyter notebook into the database seamlessly. By creating a connection string into Jupyter, we can then load our data using the “.to\_sql” and “pd.read\_sql\_query” commands to populate our pre-set tables as long as our columns, data types and sizes were correctly labeled during the transformation process. We are now able to make queries to carry various SQL commands to dive deeper into the relationship between our NY Inspection Data and Yelp Data.