**ETL PROJECT**

**RESTAURANT INSPECTIONS vs. RAT SIGHTINGS**

**Team**: Alla Zbirun, Hernan Hernandez, Ryan Michel

**EXECUTIVE SUMMARY**:

Rats in NYC persist to be an issue. As residents of NY, we volunteered to investigate this issue and determine if there is a correlation between restaurant inspections by NY Health department and rat sightings in the area. This document summarizes step by set process the team performed from extraction of the data, through data clean up and normalization, resulting in creation of databases finally uploaded into PostgreSQL for convenient SQL quiring.

**EXTRACTION**:

For this project we extracted data from three different sources -

1. NYC Open Data (<https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j/data>) – we were able to download csv files of data containing information on violations determined under inspection conducted during preceding three years for restaurants and college cafeterias
2. Kaggle.com ( <https://www.kaggle.com/new-york-city/nyc-rat-sightings>) – we were able to download csv files on rats sightings in the NYC areas.
3. Yelp.com ( <https://www.yelp.com/dataset>) – we were able to download a JASON file which contained data of restaurant credentials, review ratings and much more info on a global scale.

**TRANSFORMATION**:

The data we received was more than we needed and held number of inconsistencies or missing information. In order to normalize the data we performed following steps in Jupyter notebook to extract only sought-after elements and wrote it into a databases –

1. NYC Open Data - The restaurant dataset consisted of 394,890 rows with 26 columns. In viewing the dataset we noted obvious columns that we removed as the data had nothing to do with the focus of our project. In total we removed 8 columns. We next looked for the sum of nulls in each column. Out of the 18 columns remaining, only 5 didn't have nulls. We decided to drop the rows that had null values. Lastly, the dataset contained 7 rows with duplications and we removed all of them. After the cleaning, the final dataset contains 193,397 rows of the restaurant inspection data.
2. Kaggle data- There were a total of 52 columns found in the original dataset. Since the goal was to see if there was a relationship with health grade ratings and number of rat findings, we only preserved a zip-code column. Further, we used a groupby function to determine number of rat findings by zip-code; which we then saved the total number of observations as a column in the new dataframe. We removed all null values and one entry that listed incomplete zip-code. The final dataset contains 186 rows of rat sightings by zip-code.
3. Yelp data – Dataset file was downloaded in TAR archive format. In order to extract data files we downloaded/utilized 7-Zip, which decompressed the file and extracted contents of the archive into a folder. Out of the data available (see<https://www.yelp.com/dataset/documentation/main> for details) we chose business.json file as it contained most relevant information (i.e. - address, restaurant category, ratings, lat & long, and review counts). After detailed review of the file contents we dropped number of insignificant columns (attributes, hours, is\_open). We re-ordered columns for more convenient read of the data by sending column headers to a list; then changing the order of the appearance and reflecting it in the dataframe. Further, we groupped data by states and noted that the dataset is not exclusive to USA and contains only 22 entries for NYS. We extracted NYS data and were able to determine that this set contained only one entry for NYC which was a beauty salon. At this point we decided to drop this dataset from the analysis.

Please see Data Dictionary table at the bottom or the report for detailed listing and description of the data elements retained for quiring.

**LOAD:**

Based on the retained data we generated following dataframes[[1]](#footnote-1) and uploaded into pgAdmin for convenience of segregation, maintenance and quiring process:

* Based on NYC OpenData set –
  + Violations – holds a list of violations and their descriptions
  + Boro\_Zip – holds a list of zip-codes in the dataset with relative borough name and inspection claim id
  + Restaurant – holds a list of restaurants in the database including name of the business, borough, zip-code, type of cuisine and grade.
  + Inspections – holds a list of inspection claim ids, violation does, zip-codes and inspection date
  + Geo-Location – holds a list of inspection claim ids with relevant latitude and longitude of the restaurant.
* Rats – based on the rat sighting analysis which shows zip-code and associated rat sightings.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **DATA DICTIONARY -** | |  |  |  |
|  |  |  |  |  |
| **FIELD** | **TYPE** | **FIELD SIZE / FORMAT** | **DESCRIPTION** | **EXAMPLE** |
| CAMIS | integer | 8 | This is an unique identifier for the entity (restaurant); 10-digit integer, static per restaurant permit | 50090702 |
| DBA | varchar |  | This field represents the name (doing business as) of the entity (restaurant); Public business name, may change at discretion of restaurant owner | KIND OF SOUL |
| BORO | varchar | 15 | Borough in which the entity (restaurant) is located. NOTE: There may be discrepancies between zip code and listed boro due to differences in an establishment's mailing address and physical location | • 1 = MANHATTAN • 2 = BRONX • 3 = BROOKLYN • 4 = QUEENS • 5 = STATEN ISLAND • Missing |
| STREET | varchar |  | Street name for establishment (restaurant) location | ATLANTIC AVE |
| ZIPCODE | integer | 5 | Zip code of establishment (restaurant) location | 11217 |
| CUISINE DESCRIPTION | varchar |  | This field describes the entity (restaurant) cuisine. ; Optional field provided by provided by restaurant owner/manager | American |
| INSPECTION DATE | date | DD/MM/YYYY | This field represents the date of inspection; NOTE: Inspection dates of 1/1/1900 mean an establishment has not yet had an inspection | 5/1/2019 |
| ACTION | varchar | 225 | This field represents the actions that is associated with each restaurant inspection. | • Violations were cited in the following area(s). • No violations were recorded at the time of this inspection. • Establishment re-opened by DOHMH • Establishment re-closed by DOHMH • Establishment Closed by DOHMH. Violations were cited in the following area(s) and those requiring immediate action were addressed. • "Missing" = not yet inspected; |
| VIOLATION CODE | varchar | 3 | Violation code associated with an establishment (restaurant) inspection | 10F |
| VIOLATION DECRIPTION | varchar | 225 | Violation description associated with an establishment (restaurant) inspection | Food Protection Certificate not held by supervisor of food operations. |
| GRADE | text | 1 | Grade associated with the inspection | • N = Not Yet Graded• A = Grade A• B = Grade B• C = Grade C• Z = Grade Pending• P= Grade Pending issued on re-opening following an initial inspection that resulted in a closure |
| LATITUDE | integer |  | Geographical coordinate | 40.68558614 |
| LONGITUDE | integer |  | Geographical coordinate | -74.00501092 |
| NUMBER OF FINDINGS | integer | 10 | This field represents number of rat sightings per zip-code calculated based on the information pulled from kaggle.com | 1112 |

1. Also see ERD for table details and connections. [↑](#footnote-ref-1)