# RBDA PROJECT REPORT

Team:

| | |
|---|---|
| Rahul Sankar | rrs6684 |
| Rushab Rakesh Shah | rs7477 |
| Varun Ramesh | vr2121 |
| Yashasvi Choukikar | yc4953 |


## Background:

Given how competitive the restaurant business has become in today's world, we understand that identifying correlations among various parameters affecting the decision-making business process would greatly benefit every potential restaurant owner to arrive at better conclusions.

We believe that opening up a new restaurant not only entails cost of raw materials, staff pricing and cuisines , but is also governed by the existing competition as well as real estate property rates and the safety for the customers (crime rate for the locality).

## Problem Statement:

We aim to analyze, considering the competition (data collected from Yelp and NYC health inspection ), real estate property tariff  and crime rates for NYC, given a cuisine and coordinates, will opening a new restaurant business be feasible.

We believe this shall help potential restaurant owners come to a better conclusion given the data-driven analysis.

# Data Collection and Cleaning:

We have identified the following sources of data:

1. Yelp

   Yelp is a perfect aggregation to help analyze not only the competition but also to help gauge the popular cuisines as well as identify how customers are resonating with the same.

   Yelp has API's which produce data in various targeted json and we are targeting NYC data only. Hence the data collection will be limited by adding the filter:'NYC'

   **Business.json**


   As shown above, the data fetched contains various categories like

   Name, **coordinates**, location(entire address), **categories**, phone, **review_count**, etc.

We are concentrating on the highlighted attributes, i.e., coordinates, categories, review_count.

This data collection, profiling & cleaning , translating it to csv shall be done by Rushabh Rakesh Shah.

**Review.json**

Aggregated ratings do not often directly translate to popularity. With this json we aim to consolidate our popularity factor by categorizing the reviews as positive, negative or neutral. This shall be achieved using the openly available sentiment analysis python API which shall output the reviews in three categories described and that shall be appended to the final csv for data analysis. (Note , our aim is to gather the popular consensus for the cuisine and restaurant(locality)  to help measure the correlation between the two, hence it will be included as part of Data Extraction and Combine step only)

This data collection, profiling & cleaning , translating it to csv shall be done by Yashasvi Choukikar.

2. NYC Crime rate:
   https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u

   This is updated daily.
   The file is of 1.3GB
   The dataset has the following attributes, of which the **coordinates** , **latitude and longitude** shall be of main importance.

| Field Name | Description |
| --- | --- |
| ARREST_KEY | Randomly generated persistent ID for each arrest |
| ARREST_DATE | Exact date of arrest for the reported event |
| PD_CD | Three digit internal classification code (more granular than Key Code) |
| PD_DESC | Description of internal classification corresponding with PD code (more granular than Offense Description) |
| KY_CD | Three digit internal classification code (more general category than PD code) |
| OFNS_DESC | Description of internal classification corresponding with KY code (more general category than PD description) |

| | |
|---|---|
| LAW_CODE | Law code charges corresponding to the NYS Penal Law, VTL and other various local laws |
| LAW_CAT_CD | Level of offense: felony, misdemeanor, violation |
| ARREST_BORO | Borough of arrest. B(Bronx), S(Staten Island), K(Brooklyn), M(Manhattan), Q(Queens) |
| ARREST_PRECINCT | Precinct where the arrest occurred |
| JURISDICTION_CODE | Jurisdiction responsible for arrest. Jurisdiction codes 0(Patrol), 1(Transit) and 2(Housing) represent NYPD whilst codes 3 and more represent non NYPD jurisdictions |
| AGE_GROUP | Perpetrator's age within a category |
| PERP_SEX | Perpetrator's sex description |
| PERP_RACE | Perpetrator's race description |
| X_COORD_CD | Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) |
| Y_COORD_CD | Midblock Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) |
| Latitude | Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |
| Longitude | Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |

This data collection, profiling & cleaning , translating it to csv shall be done by Rahul Sankar.

3. NYC Health Inspection
   https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

   This data is updated daily.
   The file is of 139.9 MB

The file has the following attributes:

| |
|---|
| CAMIS |
| DBA |
| BORO |
| **BUILDING** |
| **STREET** |
| **ZIPCODE** |
| PHONE |
| CUISINE DESCRIPTION |
| INSPECTION DATE |
| ACTION |
| VIOLATION CODE |
| VIOLATION DESCRIPTION |
| CRITICAL FLAG |
| SCORE |
| GRADE |
| GRADE DATE |
| RECORD DATE |
| **INSPECTION TYPE** |

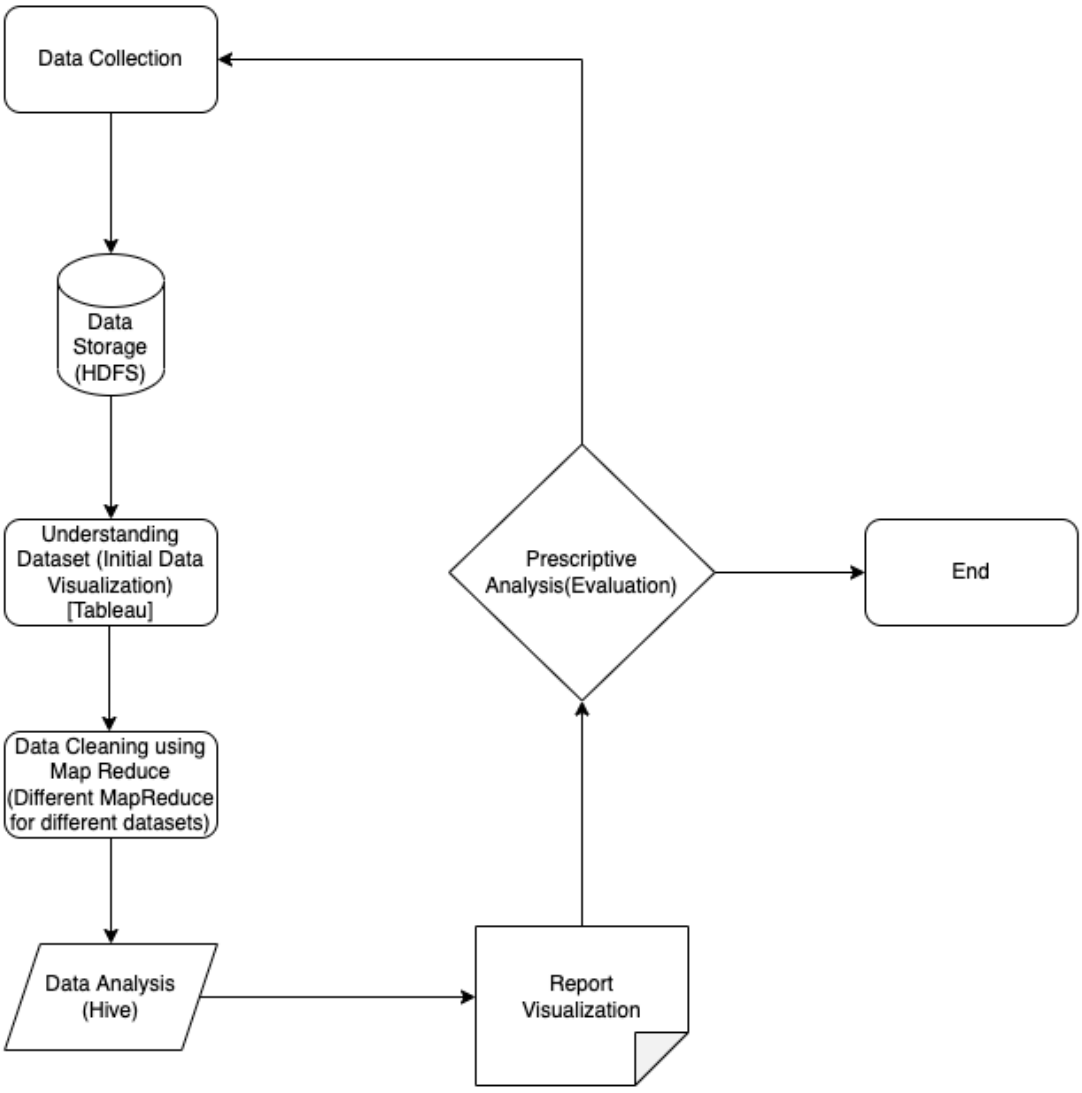
Here, the inspection type is of following values:

- Calorie Posting/ Compliance Inspection
- Calorie Posting/Initial Inspection
- Calorie Posting/ Re-Inspection
- Calorie Posting/ Second Compliance Inspection
- Cycle Inspection/Compliance Inspection
- Cycle Inspection/Initial Inspection
- Cycle Inspection/Re-Inspection
- Cycle Inspection/Reopening Inspection
- Cycle Inspection/Second Compliance Inspection
- Inter-Agency Task Force/Initial Inspection
- Inter-Agency Task Force/Re-Inspection
- Pre-Permit (Non-operational)/ Compliance Inspection
- Pre-Permit (Non-operational)/ Initial Inspection
- Pre-Permit (Non-operational)/ Re-Inspection
- Pre-Permit (Non-operational)/ Second Compliance Inspection
- Pre-Permit(Operational)/Compliance Inspection
- Pre-Permit(Operational)/Initial Inspection
- Pre-Permit(Operational)/Re-Inspection
- Pre-Permit(Operational)/Reopening Inspection
- Pre-Permit(Operational)/Second Compliance Inspection
- Smoke-Free Air Act/Complaint (Initial Inspection)
- Smoke-Free Air Act/Compliance Inspection
- Smoke-Free Air Act/Initial Inspection
- Smoke-Free Air Act/Limited Inspection
- Smoke-Free Air Act/Re-inspection
- Smoke-Free Air Act/Second Compliance Inspection
- Trans Fat/Compliance Inspection
- Trans Fat/Initial Inspection
- Trans Fat/Re-inspection
- Trans Fat/Second Compliance Inspection

Composition of the building, street, zip code will help fetch us the coordinates to understand for the given list of rival restaurants, the health violations prominent in that area(if any) for the potential restaurant owner to be better prepared.

This data collection, profiling & cleaning , translating it to csv shall be done by Varun Ramesh.

# Design Diagram:



# Tools:

Currently, we shall proceed with the following tools:

1. HDFS
2. MapReduce
3. Hive
4. Pandas
5. Tableau (initial visualization)