CAPSTONE PROJECT

BATTLE OF THE NEIGHBORHOODS

NOOH BIN SIKANDER

DATA SET

San-Francisco: Food inspection

Data Description

In this section I will the data that will be used to analyze the problem of food inspection and the source of the data. In order to develop a sufficient prediction system, the data should have the following categories:

- **Weather Data-** In public health, the weather is a key component. Long rains are associated with flooding which predisposes to contamination of food with waterborne microbes.

- **Crime Data-** Higher crime rates have been strongly correlated with poverty due to lack of employment. Poverty has been in turn correlated with low hygiene which tends to predict the occurrence of critical violations of food safety regulations.

- **Places Data-**To help locate food establishments for inspection, there needs to be a way to pinpoint exactly where they are situated and preferably show it on a map. There are different sources of places data each which its set of strength and weakness.

- **Inspection Data-** Inspection data contains information such as previous the history of critical violations, type of facility, whether the establishment has a tobacco license and the length of time the establishment has been operating.

- **Water and Sanitation data-** Garbage and sanitation complaints can be used, together with other data, to try and predict critical violations. A place with frequent sanitation complaints is more likely to have a joint with critical violations as compared to another without any complaint.

- **Demographics data-** Demographics especially health demographics contain data about people living around a place including the age, sex, estimated income, occupation, recent infections all of which can be carefully correlated and used to predict a critical violation.

However, the data I have found is collected from (https://data.sfgov.org/Health-and-Social-Services/Restaurant-Scores-LIVES-Standard/pyih-qa8i). The Health Department has developed an inspection report and scoring system. After conducting an inspection of the facility, the Health Inspector calculates a score based on the violations observed. Violations can fall into:

- **High risk category**: records specific violations that directly relate to the transmission of food borne illnesses, the adulteration of food products and the contamination of food-contact surfaces.

- **Moderate risk category:** records specific violations that are of a moderate risk to the public health and safety.

- **Low risk category:** records violations that are low risk or have no immediate risk to the public health and safety.

The score card that will be issued by the inspector is maintained at the food establishment and is available to the public in this dataset.

The dataset consists of more than 53k rows (inspection cases) and 17 columns (cases features or attributes). The following table give a brief description of each feature:

#	Feature Name	Description
1	business_id	Unique number used for identification of the business
2	business_name	Business Name
3	business_address	The address of the business
4	business_city	The City (here all records have the same city San-Francisco)
5	business_state	The state (here all records have the same state CA)
6	business_postal_code	Zip/postal code of the business
7	business_latitude	The latitude value of the business location
8	business_longitude	The longitude value of the business location
9	business_location	A tuple of the latitude and the longitude values
10	business_phone_no	Business phone number
15	violation_id	Identification of violation
16	violation_description	Short description of the violation if any
17	Risk_category	Classification of the business category, Low, Moderate or High Risk