Food inspection in San Francisco

In this report we will explore and analyze an open dataset collected about San-Francisco businesses inspections. It can be download from: https://data.sfgov.org/Health-and-Social-Services/Restaurant-Scores-LIVES-Standard/pyih-qa8i.

This project will introduce a business inspection predictive analytics report that can help promote business safety and for example food business as part of the many processes put to prevent food-borne illness. Some of these processes include proper handling of food, proper preparation of food and its storage. Food inspection ensures that all these processes are done in such as a manner as to promote and achieve food safety.

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.

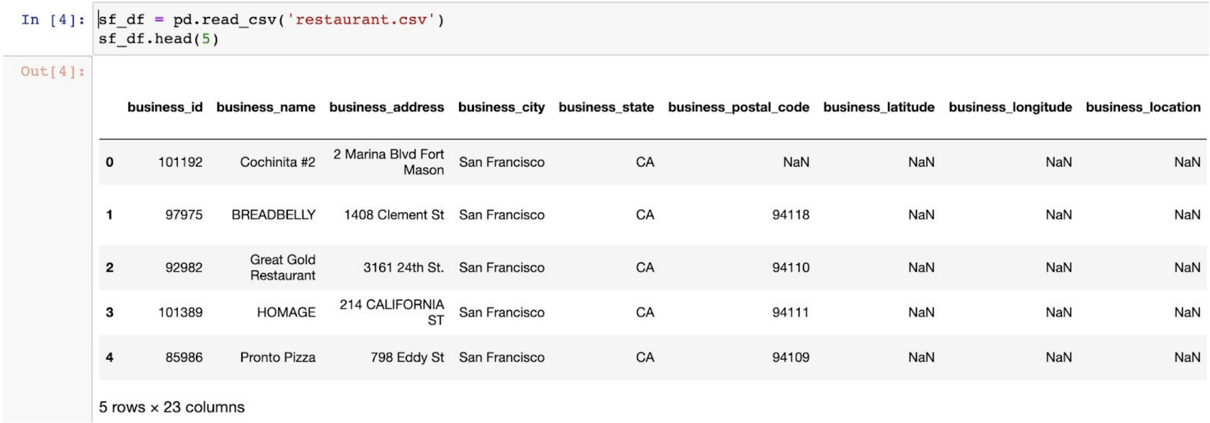
The Health Department has developed an inspection report and scoring system. After conducting an inspection of the facility, the Health Inspector calculates a score based on the violations observed. Violations can fall into:

- 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 foodcontact 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.

First of all we need to download the data from San-Francisco open data website previously given. The collected data are not ready for the analysis approach and need to be explored and organized.

A first view on the date gave us the following information:

- data looks like:



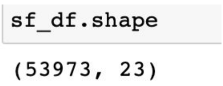
```
In [4]: sf_df = pd.read_csv('restaurant.csv')
        sf_df.head(5)
```

Out[4]:

	business_id	business_name	business_address	business_city	business_state	business_postal_code	business_latitude	business_longitude	business_location
0	101192	Cochinita #2	2 Marina Blvd Fort Mason	San Francisco	CA	NaN	NaN	NaN	NaN
1	97975	BREADBELLY	1408 Clement St	San Francisco	CA	94118	NaN	NaN	NaN
2	92982	Great Gold Restaurant	3161 24th St.	San Francisco	CA	94110	NaN	NaN	NaN
3	101389	HOMAGE	214 CALIFORNIA ST	San Francisco	CA	94111	NaN	NaN	NaN
4	85986	Pronto Pizza	798 Eddy St	San Francisco	CA	94109	NaN	NaN	NaN

5 rows × 23 columns

- we have ~53k rows and 23 features



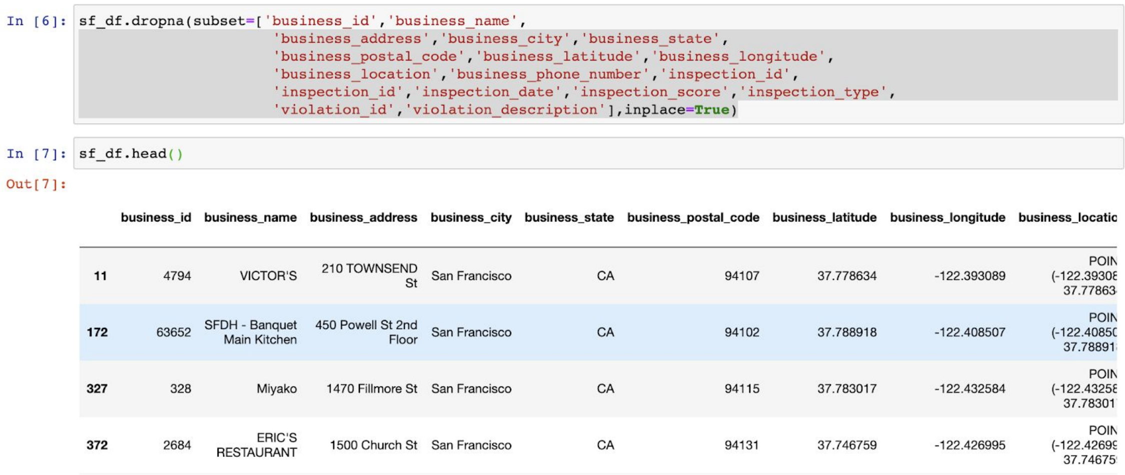
```
sf_df.shape

(53973, 23)
```

The following information represent a brief description of the features:

- business_id - Unique number used for identification of the business
- business_name - Business Name
- business_address - The address of the business
- business_city - The City (here all records have the same city San-Francisco)
- business_state - The state (here all records have the same state CA)
- business_postal_code - Zip/postal code of the business
- business_latitude - The latitude value of the business location
- business_longitude - The longitude value of the business location
- business_location - A tuple of the latitude and the longitude values
- business_phone_no - Business phone number
- inspection_id - Unique number that identifying the inspection case
- inspection_date - The date of the inspection process
- inspection_score - A score out of 100 that the business got after the inspection
- inspection_type - Routine-Unscheduled, complaint, New ownership, new construction or Non-inspection site visit. In our dataset this feature has only one value "Routine-Unscheduled"
- violation_id - Identification of violation
- violation_description - Short description of the violation if any
- risk_category - Classification of the business category, Low, Moderate or High Risk

The next step includes the preprocessing and the preparation of the data. In order to give the data to a model, we first need to have it in a proper format:

- delete the NaN values:



```
In [6]: sf_df.dropna(subset=['business_id','business_name',
                             'business_address','business_city','business_state',
                             'business_postal_code','business_latitude','business_longitude',
                             'business_location','business_phone_number','inspection_id',
                             'inspection_id','inspection_date','inspection_score','inspection_type',
                             'violation_id','violation_description'],inplace=True)
```

```
In [7]: sf_df.head()
```

Out[7]:

	business_id	business_name	business_address	business_city	business_state	business_postal_code	business_latitude	business_longitude	business_locatio
11	4794	VICTOR'S	210 TOWNSEND St	San Francisco	CA	94107	37.778634	-122.393089	POIN (-122.39308 37.77863
172	63652	SFDH - Banquet Main Kitchen	450 Powell St 2nd Floor	San Francisco	CA	94102	37.788918	-122.408507	POIN (-122.40850 37.78891
327	328	Miyako	1470 Fillmore St	San Francisco	CA	94115	37.783017	-122.432584	POIN (-122.43258 37.78301
372	2684	ERIC'S RESTAURANT	1500 Church St	San Francisco	CA	94131	37.746759	-122.426995	POIN (-122.42699 37.74675

We will use summarize the inspection data by risk_category. The general process involves the following steps:

1. **Split:** Splitting the data into groups based on the risk_category.
2. **Apply:** Applying the count and mean function to each group independently:
3. **Combine:** Combining the results into a data structure.

```
In [8]: df_risk = sf_df.groupby('risk_category', axis=0).count()
        df_risk.head(10)
```

Out[8]:

	business_id	business_name	business_address	business_city	business_state	business_postal_code	business_latitude	business_longitude	busine
risk_category									
High Risk	735	735	735	735	735	735	735	735	
Low Risk	2538	2538	2538	2538	2538	2538	2538	2538	
Moderate Risk	1942	1942	1942	1942	1942	1942	1942	1942	

3 rows × 22 columns