**The final report**

Determining a feature vector for classification the food serving venue according to risk for a public health

## Introduction/Business problem

Sanitary inspection of various cities in United States publish the results of checking the sanitary and other relevant conditions of venues that serve a food in the form of dataset with list of venues’ metadata and the result of the inspection (ranking by the risk) on the healthdata.gov web site (<http://www.healthdata.gov>). On the other side, the Foursquare dataset contains an attribute “likes” that contains the number of times the users gave a like (prefer, vote up) that restaurant among others. It is interesting to see how objective the users in their likes are/could be from the perspective of sanitary conditions, or how much are sanitary conditions relevant to users (guests) in their decision to like or not to like a restaurant.

Idea of the project is to correlate this dataset with Foursquare database for one city (Chicago), and check if attribute “likes” can be a good predictor for a risk category of a restaurant according to the inspection results . Based on data for the reference city (Chicago) determine a set of features relevant to classify the restaurant in another city as risky or another class, according to available list of risk classes in the results of inspections (multivalue classification). As evaluation dataset will be used the exact results of inspection control for the second city (San Francisco), acquired from healthdata.gov.

The audience for the problem could be optimisation of the resources in inspections by targeted approach to inspection sample or to rise the frequency of inspection on highly risky samples.

## Data

Foursquare dataset contains, among the others, these data that are relevant for the insight in data and solving a problem:

|  |  |
| --- | --- |
| Foursquare.com | |
| Column name | Description |
| Venue name | Name of the venue (restaurant in this case) |
| Latitude | Latitude part of geolocation |
| Longitude | Longitude part of geolocation |
| Address | Postal address (geocoded) |
| City | City |
| State | State |
| Zip | Postal code |
| Likes | Number of users’ likes for the restaurant |

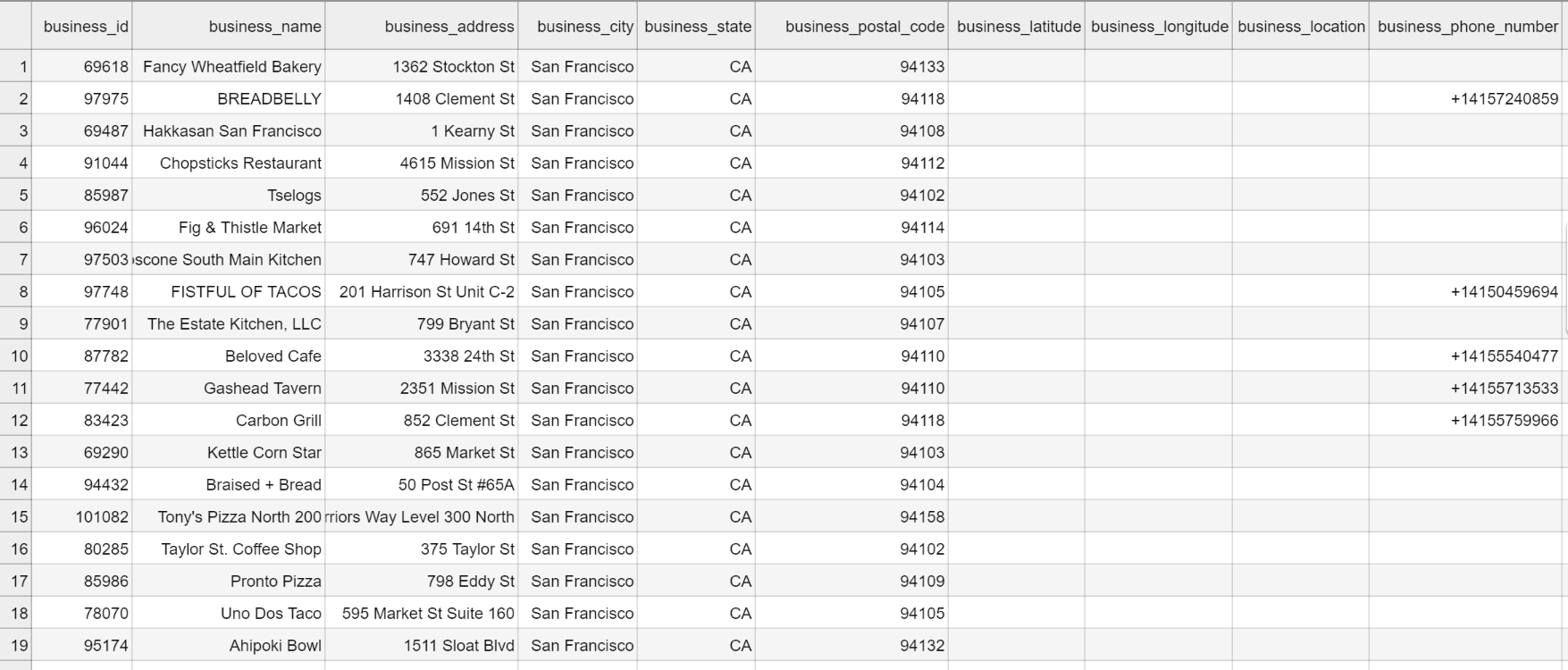
Datasets that arise as results of inspection control contain, among others, these data that are relevant for the solution of the problem:

|  |  |
| --- | --- |
| Food control results | |
| Column name | Description |
| Restaurant name | Name of the restaurant |
| Address | Postal address (geocoded) |
| City | Name of the city |
| State | Name of the state |
| Zip | Postal code |
| Risk | Risk level value |
| Latitude | Latitude part of geolocation |
| Longitude | Longitude part of geolocation |

Example of the cells from the reference (training) dataset:



Example of the cells from the evaluation (test) dataset:



Data are obviously in the need for some data wrangling: wrongly parsed csv data need to be properly aligned, missing location data need to be determined out of the geocoded address (or cells deleted).

Pairing the datasets will be done according to the tuples (Name, Address, City, State, Zip, Latitude, Longitude), in order to mitigate a possible ambiguity in data.

Merged dataset with clearly marked source of data is given below:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Merged dataset | | | | | | | | | |
| Source: | Foursquare.com | | | | | | | | Food control |
| Feature name | Name | Likes | Latitude | Longitude | Address | City | State | Zip | Risk |
| 1 |  |  |  |  |  |  |  |  |  |