**San Francisco restaurants dataset analysis**

1.Summary

My task was to analyze POI data from San Francisco restaurant dataset for our [www.navigator.ba](http://www.navigator.ba) project. At Navigator, we are currently expanding our location coverage and are in search of quality datasets which would enable us to put local San Francisco restaurants on the map along with some basic info.

The supplier provided us with San Francisco restaurant inspections dataset which has all the information we need to set up the coverage expansion. Since we at Navigator look only for highest quality data, I used basic data analysis principles and tools to separate that kind of data from low quality one.

Following metrics were used to categorize data from the dataset:

* High quality records – Full row, validated and verified data.
* Medium quality records – Full row, lacks proper validation or verification.
* Low quality records – Rows with missing values.

After performing analysis, presented in Jupyter Notebook: Restaurant-analysis.ipynb, the following conclusions were drawn:

* Number of high quality records: 804
* Number of medium quality records: 55
* Number of low quality records: 4563

2. Analysis

Tools used to perform the analysis were Jupyter Notebook and pandas Python 3.9 library.

Upon loading data into pandas dataframe first observation was that 52315 records were present. Since we, at Navigator, concern ourselves only with the basic description of restaurants all the columns regarding health inspections were immediately dropped.

Many restaurants appeared multiple times through the dataset because each was subject to varying number of inspections. After removing duplicates 5422 records of either low, medium or high quality remained.

To separate medium and high from low quality records we filtered records with missing values and ended up with 859 records. It was then proceeded with verification and validation of the remaining records to filter high quality from medium quality ones. To achieve that every column of the dataset apart from restaurant name, address and phone number was checked. Verification of these fields required a process too great and complex for our task.

Upon validating the ‘**business city**’ column it was confirmed that every record was labeled ‘San Francisco’ which is verified since our dataset is about restaurants located in San Francisco, CA, USA.

The ‘**business state**’ column showed three different labels: ‘CA’, ‘California’ and ‘IL’. Two of those: ‘CA’ and ‘IL’ were valid since ‘California’ is not a state code in USA. The external verification showed the ‘IL’ state code stands for the state of Illinois so these records were filtered out.

Next was the ‘**business postal code**’ column. External verification showed that all the ZIP codes of San Francisco area have 5 digits and start with ‘94’. Listing out the unique values in our dataset showed us 4 anomalous codes: ‘941102019’, ‘941’, ‘CA’, ‘92672’ with the last one being a valid code but verified not to be one for the SF area.

Lastly, it was proceeded with the verification of geographical coordinates. A decision was made to map the records and visually confirm if all the restaurants were in San Francisco area. For this, ‘folium’ Python library was imported. Upon mapping the records it was visually clear that many coordinates do not respond to places in San Francisco. Google maps were used to draw a rectangle around the city of San Francisco. Coordinates of rectangle were used to further filter out the records whose ‘**business** **latitude**’ and ‘**business** **longitude**’ did fall within rectangle. It is important, however, to note that random sample of 10 restaurants with false coordinates showed they do, in fact, exist in San Francisco but, because of this fallacy, couldn’t meet our criteria for high quality records.

In the end 804 records or 14.82% of unique restaurant records were validated and verified to fall into the high quality category.

Medium quality records, 55 of them, which can be rectified into high quality with minimum effort – mostly basic googling, made up 1.02% of unique restaurant records.

The rest were low quality records – 4563 of them. They made 84.15% of unique records.

Attached to this document are Jupyter Notebook with the detailed analysis and a CSV file which contains the dataset used.

Radovan Simikićstaurants were in San Francisco area.rdinates. A decision was made to map the records and visually