The final report

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

Autor:

Nihad Hasic

December 2019.

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 and attribute "rating" which represents the average of all ratings given by users. It is interesting to see how objective the users in their likes and ratings 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 or how to rate a restaurant.

Idea of the project is to correlate this dataset with Foursquare database for one city (Chicago), and check if attributes "likes" and "rating" 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
Rating	The rating of the venue according to users

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:

	DBA Name	AKA Name	License #	Facility Type	Risk	Address	City	ate	Zip	Inspection Date	Inspection Type	Results	
1	SALSA'S GRILL	SALSA'S GRILL	2590110	Restaurant	Risk 1 (High)	3808 W LAWRENCE AVE	CHICAGO	IL	60625	10/29/2019	Canvass	Pass w/ Conditions	3. MAN
2	A BEAUTIFUL RIND	A BEAUTIFUL RIND	2670348		Risk 3 (Low)	2211 N MILWAUKEE AVE	CHICAGO	IL	60647	10/28/2019	License	Not Ready	
3	A BEAUTIFUL RIND	A BEAUTIFUL RIND	2670349		Risk 3 (Low)	2211 N MILWAUKEE AVE	CHICAGO	IL	60647	10/28/2019	License	Not Ready	
4	A BEAUTIFUL RIND	A BEAUTIFUL RIND	2670347		Risk 1 (High)	2211 N MILWAUKEE AVE	CHICAGO	IL	60647	10/28/2019	License	Not Ready	
5	MARISCOS ALMADA		2698332		Risk 1 (High)	9485 S EWING AVE	CHICAGO	IL	60617	10/25/2019	License	Not Ready	
6	ALAM RESTAURANT	SALAM RESTAURANT	2002822	Restaurant	Risk 1 (High)	4634-4636 N KEDZIE AVE	CHICAGO	IL	60625	10/25/2019	int Re-Inspection	Pass	
7	TRINO'S PIZZERIA	TRINO'S PIZZERIA	2142757	Restaurant	Risk 2 (Medium)	1013 W 18TH ST	CHICAGO	IL	60608	10/25/2019	Canvass	No Entry	
8	BEST BBQ	BEST BBQ	1575975	Restaurant	Risk 1 (High)	1648 W 115TH ST	CHICAGO	IL	60643	10/25/2019	t Form Complaint	Pass	
9	LA CATRINA CAFE	LA CATRINA CAFE	2185072	Restaurant	Risk 2 (Medium)	1011 W 18TH ST	CHICAGO	IL	60608	10/25/2019	Canvass	Out of Business	
10	OVER RICE'N BREA	CORP	N BREAD	2451495	Restaurant	Risk 1 (High)	IELD AVE	GO	IL	60657	10/25/2019	nvass Re-Inspection	
11	EXTRA VALUE CO	EXTRA VALUE CO	29630		Risk 3 (Low)	7300 N WESTERN AVE	CHICAGO	IL	60645	10/25/2019	Canvass	Out of Business	
12	OLLOP COFFEE CO	DOLLOP COFFEE CO	2698381	Restaurant	Risk 1 (High)	1636 W MONTROSE AVE	CHICAGO	IL	60613	10/24/2019	License	No Entry	
13	JORDAN DISCOUNT	JORDAN DISCOUNT	2694525	Grocery Store	Risk 3 (Low)	5254 W MADISON ST	CHICAGO	IL	60644	10/24/2019	se Re-Inspection	Pass	
14	FIE'S CANTEEN INC	EN (TAXI/LIMO AREA)	29570	Restaurant	Risk 1 (High)	11601 W TOUHY AVE	CHICAGO	IL	60666	10/24/2019	Canvass	Pass	AL FACILITIES I
15	ELOPMENT CENTER	VELOPMENT CENTER	2215757	(2 - 6 Years)	Risk 1 (High)	5900 W IOWA ST	CHICAGO	IL	60651	10/24/2019	License	Pass	
16	R MEAT & GROCERY	OR MEAT &GROCERY	2341419	Grocery Store	Risk 2 (Medium)	2507 W DEVON AVE	CHICAGO	IL	60659	10/24/2019	Canvass	Out of Business	
17	CERMAK PRODUCE	CERMAK PRODUCE	2693885	Grocery Store	Risk 1 (High)	4810 W DIVERSEY AVE	CHICAGO	IL	60639	10/24/2019	se Re-Inspection	Pass w/ Conditions	AL FACILITIES I
18	MEXI-TACOS	MEXI-TACOS	2694744	ood Preparer	Risk 2 (Medium)	2300 S THROOP ST	CHICAGO	IL	60608	10/23/2019	se Re-Inspection	Pass	
19	GIANT	GIANT	2442931	Restaurant	Risk 1 (High)	3209 W ARMITAGE AVE	CHICAGO	IL	60647	10/23/2019	Canvass	No Entry	

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

	business_id	business_name	business_address	business_city	business_state	business_postal_code	business_latitude	business_longitude	business_location	business_phone_number
1	69618	Fancy Wheatfield Bakery	1362 Stockton St	San Francisco	CA	94133				
2	97975	BREADBELLY	1408 Clement St	San Francisco	CA	94118				+14157240859
3	69487	Hakkasan San Francisco	1 Kearny St	San Francisco	CA	94108				
4	91044	Chopsticks Restaurant	4615 Mission St	San Francisco	CA	94112				
5	85987	Tselogs	552 Jones St	San Francisco	CA	94102				
6	96024	Fig & Thistle Market	691 14th St	San Francisco	CA	94114				
7	97503	iscone South Main Kitchen	747 Howard St	San Francisco	CA	94103				
8	97748	FISTFUL OF TACOS	201 Harrison St Unit C-2	San Francisco	CA	94105				+14150459694
9	77901	The Estate Kitchen, LLC	799 Bryant St	San Francisco	CA	94107				
10	87782	Beloved Cafe	3338 24th St	San Francisco	CA	94110				+14155540477
11	77442	Gashead Tavern	2351 Mission St	San Francisco	CA	94110				+14155713533
12	83423	Carbon Grill	852 Clement St	San Francisco	CA	94118				+14155759966
13	69290	Kettle Corn Star	865 Market St	San Francisco	CA	94103				
14	94432	Braised + Bread	50 Post St #65A	San Francisco	CA	94104				
15	101082	Tony's Pizza North 200	rriors Way Level 300 North	San Francisco	CA	94158				
16	80285	Taylor St. Coffee Shop	375 Taylor St	San Francisco	CA	94102				
17	85986	Pronto Pizza	798 Eddy St	San Francisco	CA	94109				
18	78070	Uno Dos Taco	595 Market St Suite 160	San Francisco	CA	94105				
19	95174	Ahipoki Bowl	1511 Sloat Blvd	San Francisco	CA	94132				

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	Rating	Latitude	Longitude	Address	City	State	Zip	Risk
1										

Training dataset

Inspection control dataset for Chicago contains about 195000 historical records of the inspection controls of various facility types in the 12-years period from 2003 – 2015.

In order to optimize processing of data, only the records of the last inspection control were considered and other records needed to be deleted, because of the limited processing capabilities of the platform used.

Foursquare dataset returned a values for 100 venues that were used as metadata and needed to be merged with the rest data from inspection dataset. Unfortunately, these two datasets don't contain a common unique identifier that could be used as a key for merging (joining) the datasets. The only

common data in these two datasets were name of the venue and location data: address and geolocation data. Moreover, name and address were not standardised and written in the same manner in both datasets, and geolocation data for the same object were not equal. Short form of the parts of names or address were used, as well as the long names with addition of the type of the facility, for example:

1. Column City contained a lot of different values although the results of inspection controls were explicitly declared for Chicago:

```
CHICAGO         194243
Chicago            320
NaN                138
chicago             97
CCHICAGO            46
                   ...
LANSING              1
BURNHAM              1
alsip                1
NEW HOLSTEIN         1
BROADVIEW            1
Name: City, Length: 72, dtype: int64
```

   These data were resolved by setting the one single value of “CHICAGO” for the whole column.

2. Column State contained missing and dirty values:

```
IL      195036
NaN         42
IN           1
WI           1
NY           1
Name: State, dtype: int64
```

   This example was resolved by setting the one single value of “IL” for the whole column.

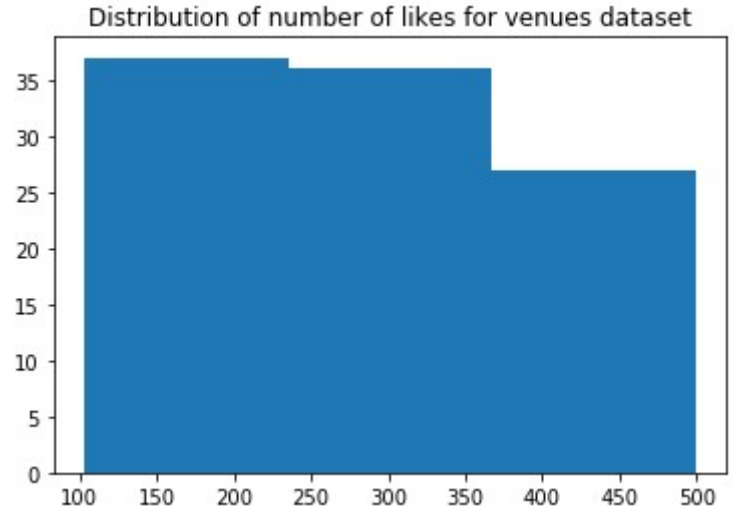
3. Columns Latitude, Longitude and Address had missing values for 683 rows. Because Address was one of the fields for pairing the datasets, these rows were deleted.

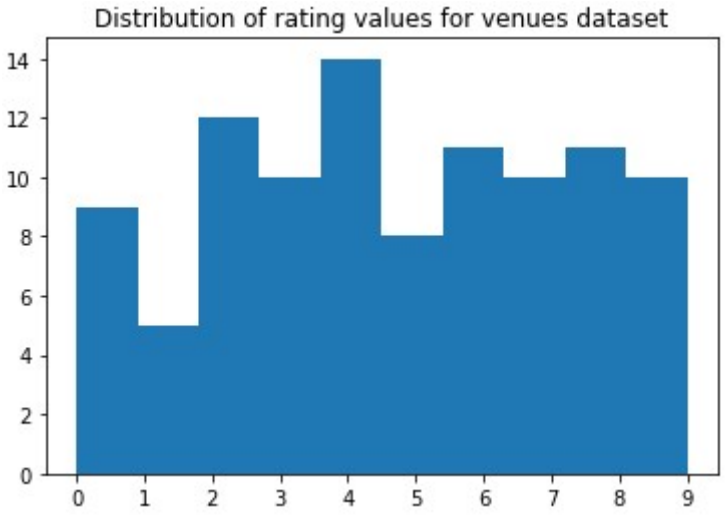
4. Column Risk had missing values in 71 rows. These values were filled with the most frequent value for column Risk – Risk 1 (High)

```
Risk 1 (High)      139442
Risk 2 (Medium)     37926
Risk 3 (Low)        16928
NaN                    71
All                    31
Name: Risk, dtype: int64
```

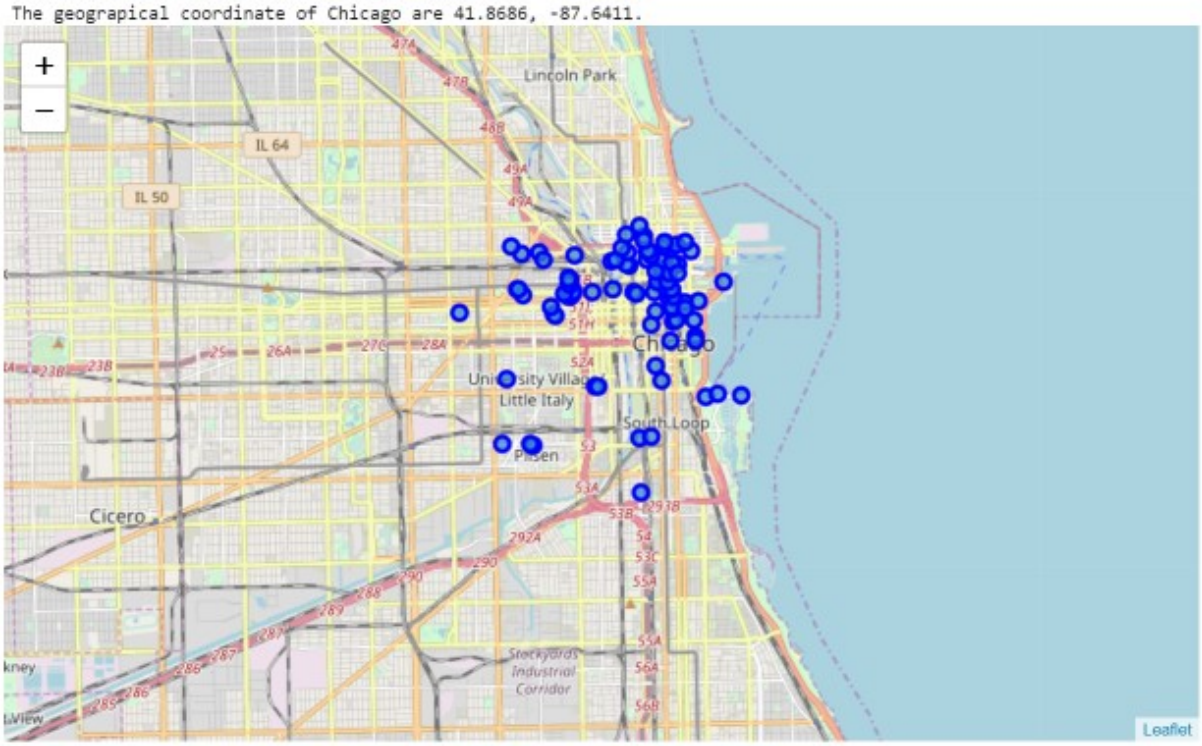
5. Datatype of column Inspection Date needed to be converted to type datatime in order to be able to sort the data by this column.

Distribution of data in training dataset is given below:



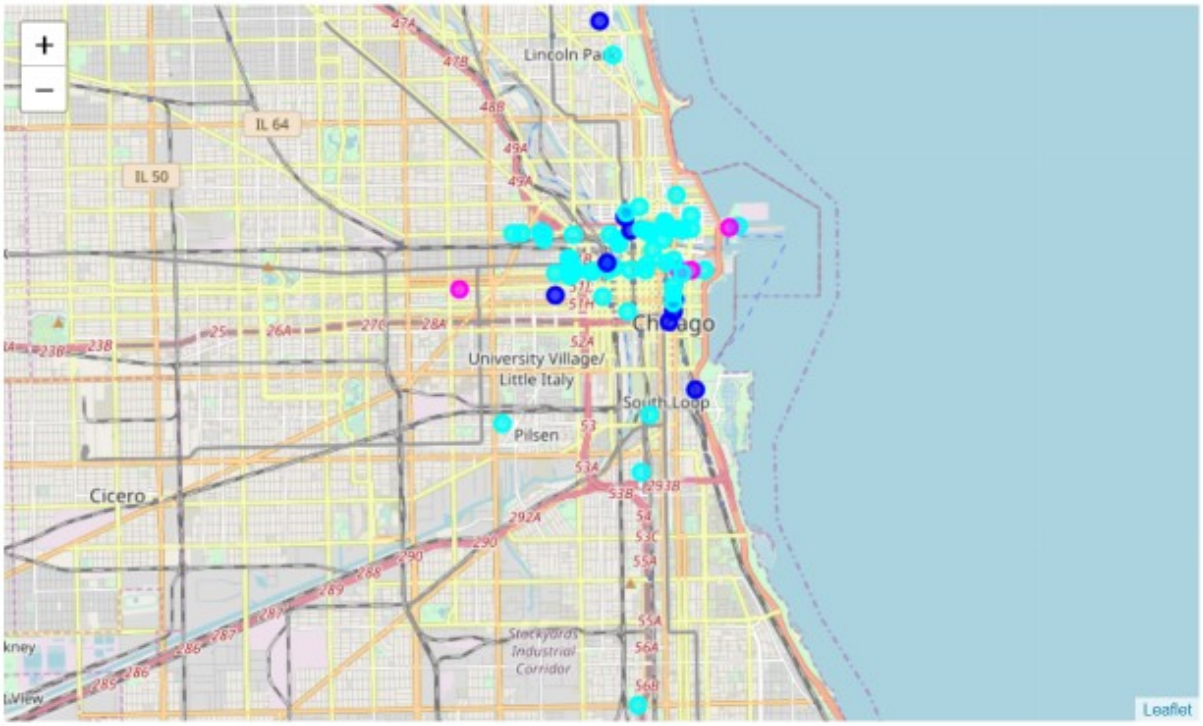


Spatial distribution of venues from Foursquare is depicted on the map below:



After joining the Foursquare and inspection control dataset venues are represented on the map below with coloring according to the risk level by following schema:

Color	Coded risk value	Risk
cyan	1	High
magenta	2	Moderate
blue	3	Low



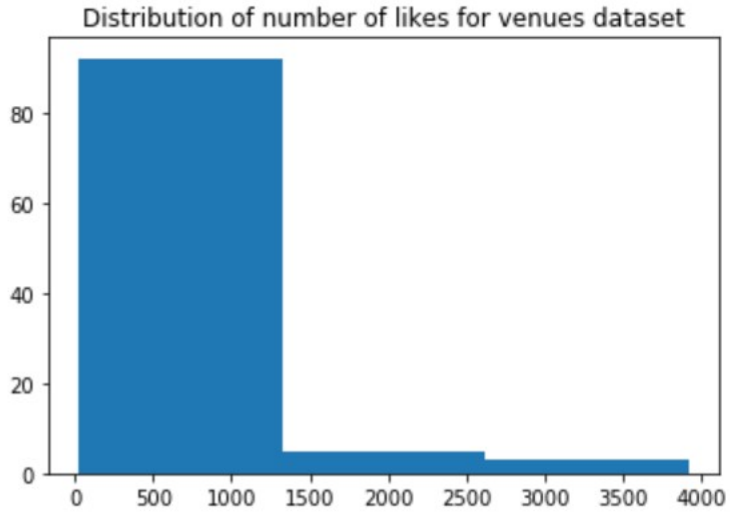
Test dataset

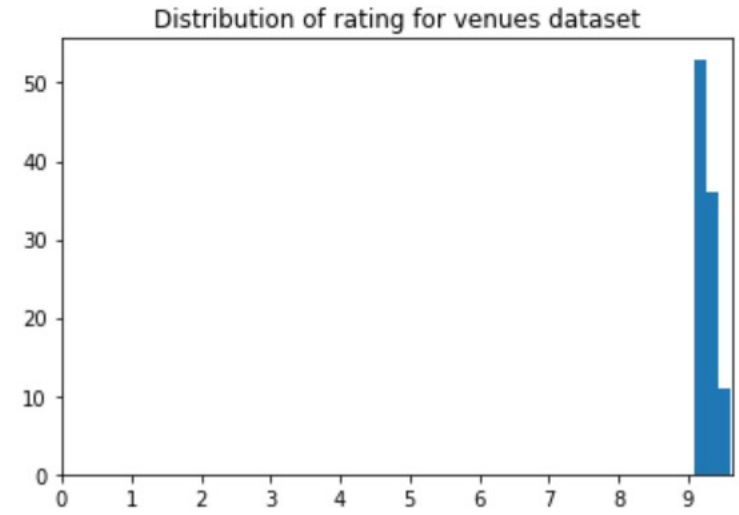
Inspection control dataset for San Francisco contains about 54000 historical records of the inspection controls of various facility types in the 4-years period from 2016 – 2019.

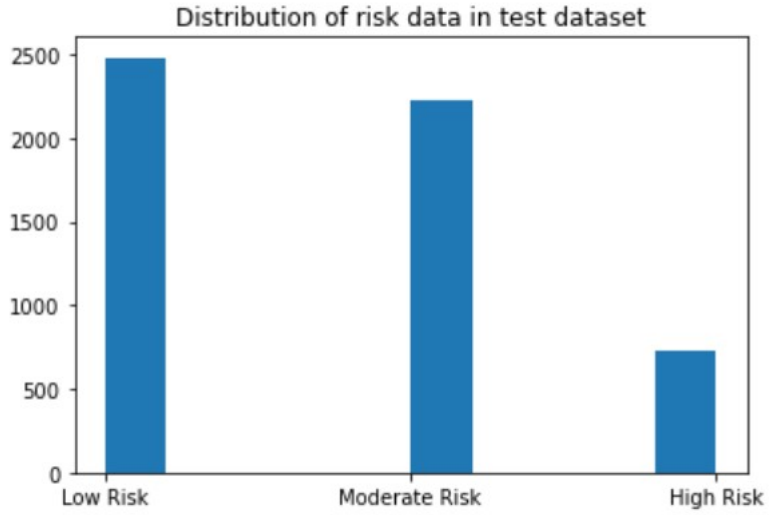
As in the case of training dataset, only the records of the last inspection control were considered and other records needed to be deleted, because of the limited processing capabilities of the platform used.

The problems of dirty and missing data and missing common unique identifier were also present here and were overcome as described in the case of training data.

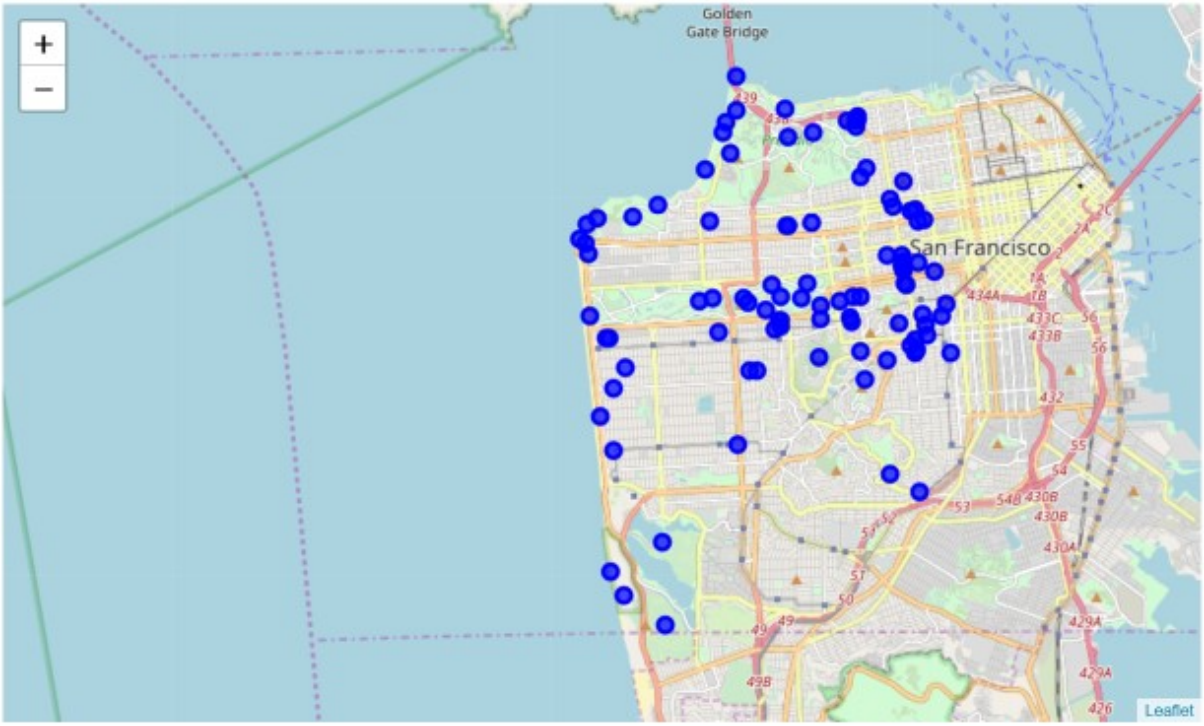
Data in test dataset were distributed as shown below:





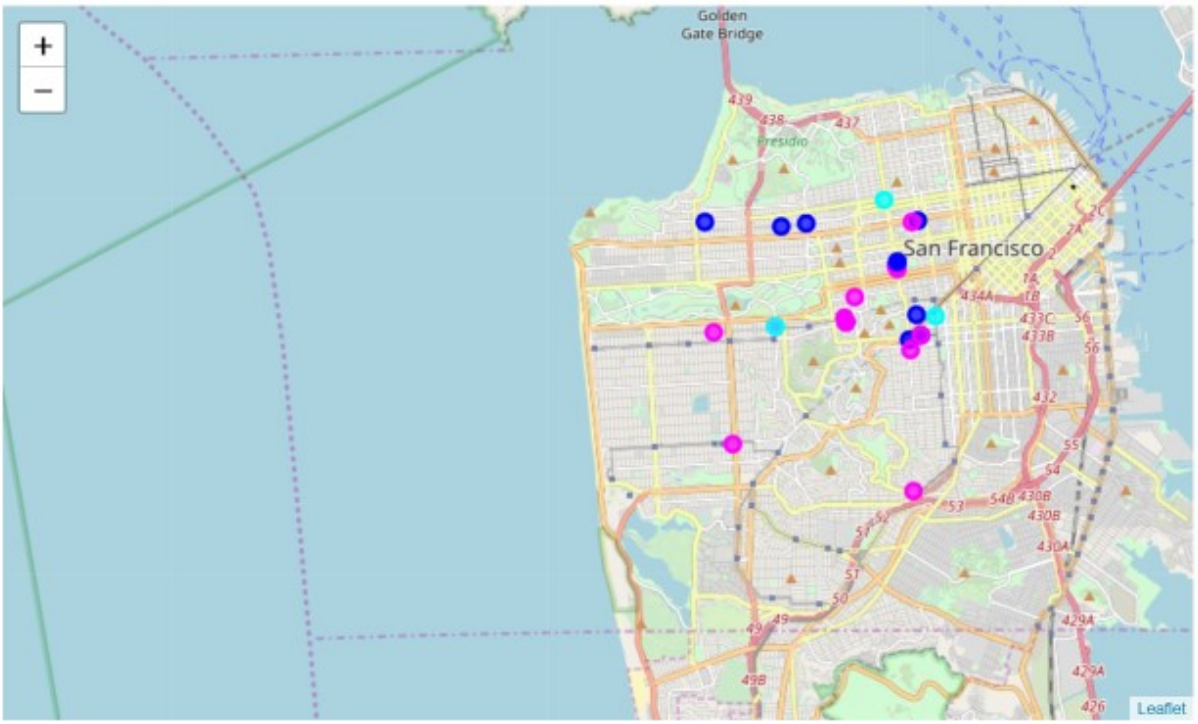


Venues from the Foursquare for a test dataset are given on the map below:



Data from merged dataset consisting of Foursquare dataset and inspection control dataset are depicted on the map below with coloring according to following schema:

Color	Coded risk value	Risk
cyan	1	High
magenta	2	Moderate
blue	3	Low



Methodology

All results of the data insight given above mean that it was dealt with dirty data, and the task of merging datasets was the problem of matching the strings to the acceptable level of the difference.

The definition of “acceptable level of difference” needed to be discovered empirically from the datasets, but for the measuring of difference was needed to choose a known standard algorithm for the problems of matching strings. Two most known measures in matching the strings are Levenshtein and Hamming distance. Formal definitions will be omitted here, but what is important is that Hamming distance has a disadvantage in this case that strings need to be of the same length. Because of that the Levenshtein distance was chosen as a measure of similarity of the strings for the complex key for merging two datasets, consisting of name of the object and the given postal address.

By analysing the test merges of data was founded that the optimal values of Levenshtein distance between the name and address in Foursquare dataset and the name and address in the inspection control dataset were: 5 for name and 8 for address.

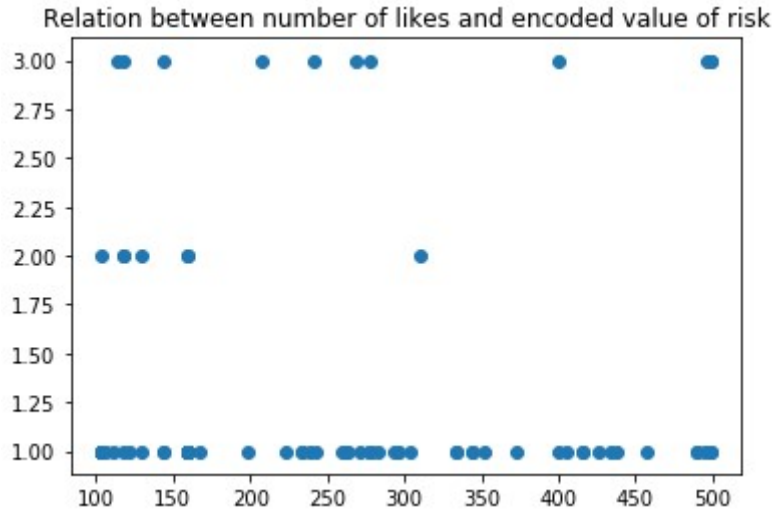
Considering the geolocation distance computed from the values of latitude and longitude from the Foursquare and the inspection control dataset brought only unnecessary computing load and not the better results of matching. Therefore, was not used as criteria for matching.

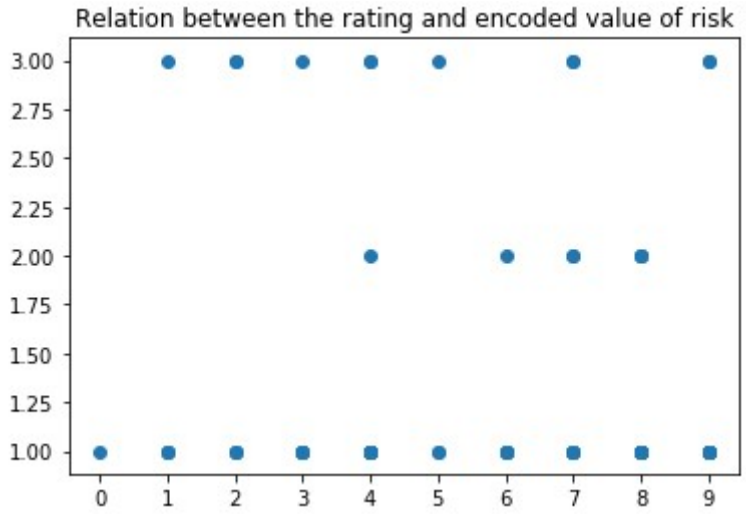
After the matching, 82 matches between two datasets were found. This seems to be quite a small dataset for training any serious classifier, but it is the result of the limitation of Foursquare API that returns maximally 100 records. This doesn’t necessarily reduce the generality of the solution. With the more generous API and enough computation power, solution is easily scalable on those sources of data.

In the case of test dataset, the parameters for Levenshtein distance for name and address were the same: 5 and 8 respectively. After the matching, 56 matches between two datasets with test data were found.

For the purpose of classification text values in column Risk were encoded using the LabelEncoder.

The scatter diagrams which represent relationship between risk and number of likes and risk and rating are given below:





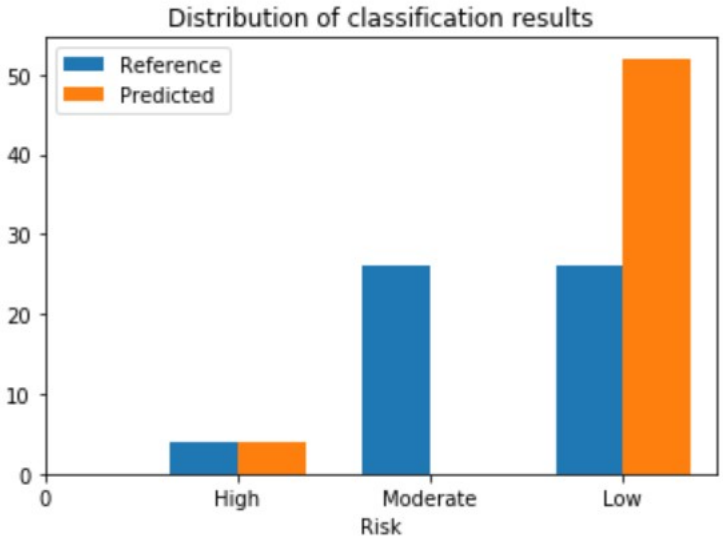
From the scatter diagrams is easy to conclude that within the available data, correlation between those variables is weak. That means that generalisation potential of the resulting classifier will be weak and that it is only possible to examine what is the best what can be achieved under these conditions, at least to check what the overfitting brings in this case.
The problem of limitations of Foursquare API that returns only 100 records also contributes to the small number of matches.

The machine learning algorithm that will be used is the multi class logistic regression since the problem represented here is a multiclass classification problem and the simplicity of solution needed to be preserved.

Additionally two other classification algorithms: K-nearest neighbours and decision trees will be evaluated, and results compared.

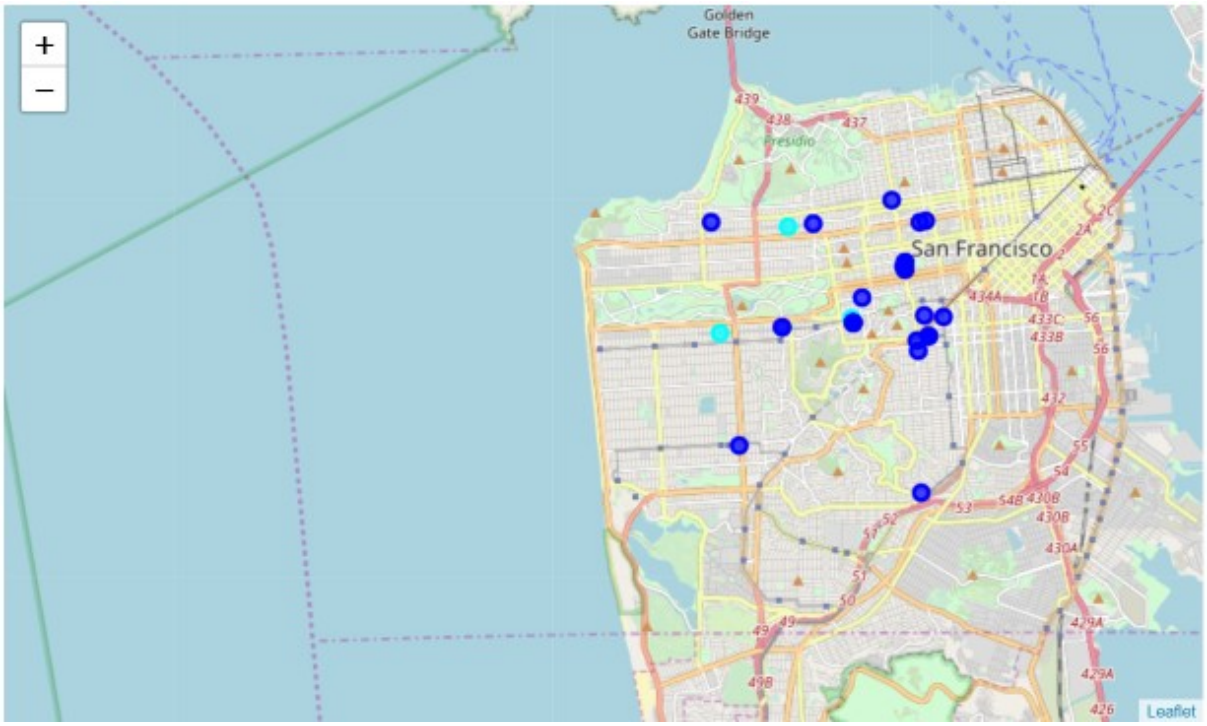
Results

On the bar chart below is presented the distribution of classification results with logistic regression classifier in terms of correct value (reference - blue) and predicted value for that venue (predicted - orange):

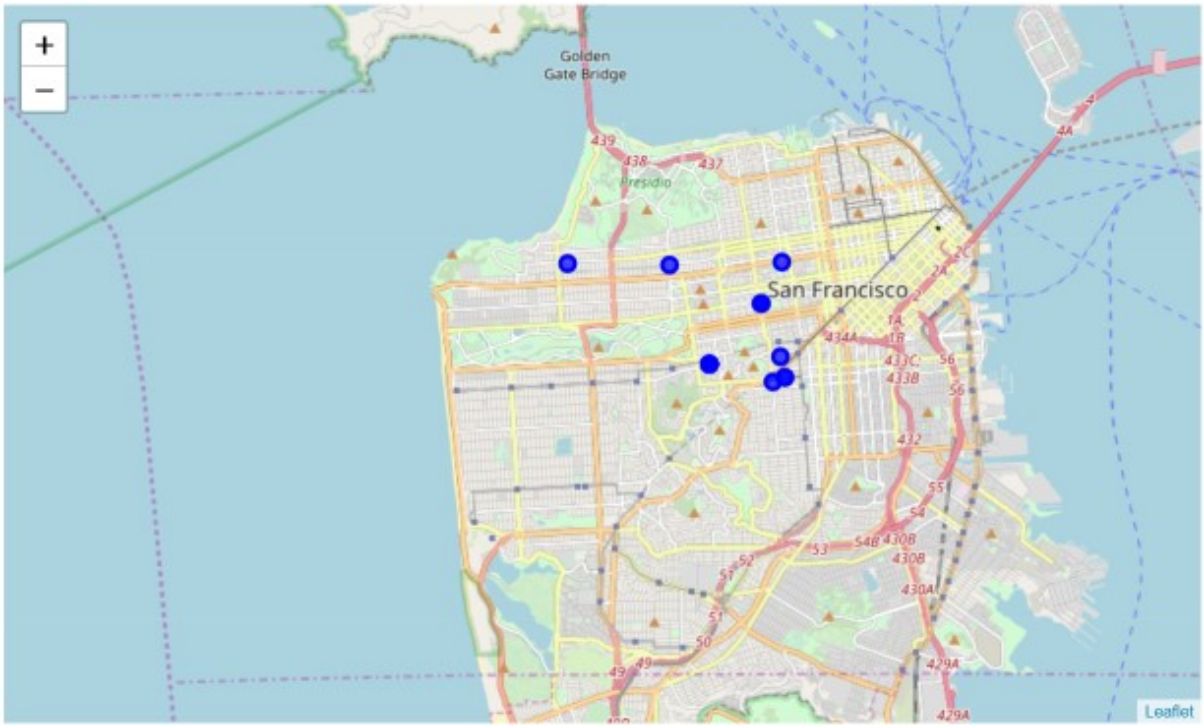


Results of classification – spatial distribution with coloring according to the legend:

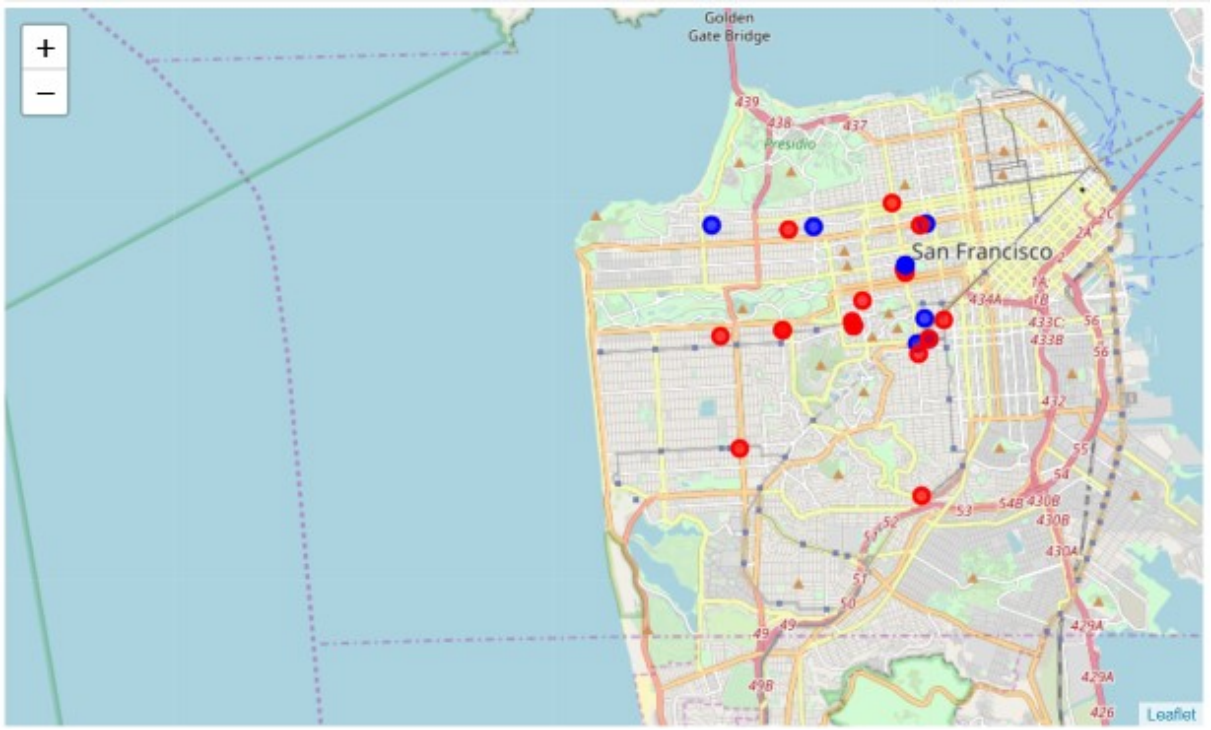
Color	Coded risk value	Risk
cyan	1	High
magenta	2	Moderate
blue	3	Low



On the map below are given the hits i.e. venues with successfully predicted risk values:



Map below presents the comparison of number of hits (blue circles) and fails (red circles), I.e. visually represents success rate of classification:



Calculated values of common classification evaluation metrics are given in the table below for chosen logistic regression classifier, as well as for two other classification algorithms that were evaluated:

Algorithm	Jaccard similarity index	F1 score
Logistic regression	0.44643	0.29762
KNN (K=4)	0.08929	0.04424
Decision trees (max depth = 6)	0.07143	0.00969

Discussion

As already discussed above, the problem of dirty and missing data in inspection control datasets, limitation of number of results on Foursquare API, and missing common unique identification keys for two datasets lead to the training set that didn't contain enough data for successful training the classifier. Although three classifiers were used in comparison, none of them gave satisfying results on the test dataset. Even variation of parameters didn't bring a better precision.

Taking into consideration that chosen features number of likes and rating can also be very subjective, inconsistent and error-prone, finding the relationship or generalisation of the rules can be very difficult even with the more training samples.

Conclusion

In this analysis it couldn't be proven that number of likes and rating on the Foursquare could be good predictors of sanitary conditions or results of inspection controls. Value of Jaccard's similarity index of 0.44643 in the case of logistic regression tells that there is a possibility with the bigger training sample to prove a relationship between variables and significance of number of likes and rating as the features in the classification of venues according to sanitary risk.