**San-Francisco Business Inspection - Food Inspections**

We will explore and analyze a dataset collected about San-Francisco businesses inspections. I have applied most of the stages of the data science methodology that I have studied in this specialization. This project will introduce a business inspection-predictive analytics report that can help promote business safety. For example, food businesses use many processes to prevent food-borne illnesses. 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 a manner to promote and achieve food safety.

Food inspection involves not only sampling and testing of products, but also assessing food centers to ensure compliance with food safety management systems and laws. This minimizes the occurrence of public health food safety problems. The Food and Drug Administration (FDA) publishes the Food Code that sets guidelines and procedures to assist in food control jurisdictions. The Food Code provides a scientifically and legally backed basis for regulating the retail and food service industries. These include restaurants, grocery stores, and institutional food service providers. In the past, food inspection was done in a reactive manner whereby officers waited for reports of possible non-compliance. However, it has been shown through research that food inspection should be done in a more proactive manner. Currently, some cities in the US, such as San Francisco, are implementing a technologically-driven approach to food inspection in attempt to predict food establishments that are more likely to be non-compliant to food safety regulation. This is driven in part by the low Inspector to Food place ratio making it difficult to efficiently inspect all the food places.

We will use Foursquare, which comes with venue data that contains key descriptors of different venues including the category and popularity. This will show categories like food establishments along with attributes such as name, address, ratings, and reviews from millions of points of interest. This report would be beneficial to public health specialists and every stakeholder working to alleviate public health concerns through preventive measures. The solution is not to introduce food inspection since these professionals are already carrying out food inspections in the relevant jurisdictions, but to make the process more efficient.

In San-Francisco, it is estimated that one business inspector needs to efficiently inspect more than 500 business establishments given that there are only about 4 dozen inspectors to cover all business establishments. It is in this statistic that the city saw an opportunity to make the process of food inspection more efficient by utilizing data analytics. In San-Francisco, through the Department of Public Health, they collected food inspection data from close to 100,000 sanitation inspections. Using this data, along with data on weather, related complaints, and business characteristics, the city’s analytics team helped predict the food establishments that are more likely to violate food safety regulations. The food inspectors can then have a “Critical first” inspection approach where the places that have been predicted to have critical violations are inspected first. Some of the factors that tend to predict critical violations include previous violations, high temperatures, nearby sanitation complains, nearby burglaries, etc.

# Data Description

This section data will be used to analyze the problem of food inspection and the source of the data. To develop a proper prediction system, the data should have the following categories:

* **Weather Data-** In public health, the weather is a key component. Rains is 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 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, income, occupation, and recent infections, all of which can be carefully correlated and used to predict a critical violation.

However, the data collected is from ([https://data.sfgov.org/Health-andSocialServices/Restaurant-Scores-LIVES-Standard/pyih-qa8i).](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.

# Methodology

In this part of the report we are going to describe the main components of our analysis and predication system. Our methodology consists of 5 components as shown below.



Collect Inspection Data



Explore and Understand



the Data



Data preparation and



preprocessing



Modeling



Evaluation and Testing

# Collect Inspection Data

We downloaded the data from San-Francisco open data website as follows



The collected data is not ready for the analysis approach and needs to be explored and organized.

# Explore and Understand the Data

We read the dataset that we collect about San-Francisco business inspection into a pandas’ data frame and display the first 5 rows of it as follows:





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 |
| 11 | inspection\_id | Unique number that identifying the inspection case |
| 12 | inspection\_date | The date of the inspection process |
| 13 | inspection\_score | A score out of 100 that the business got after the inspection |
| 14 | 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*” |
| 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 |

We visualize the dataset to get more insight about it and discovering some pattern that might help in the modeling section.

# Data Preparation and Preprocessing

In this component, we prepare the dataset for the modeling process where we choose the machine learning algorithms. To do that, we have cleaned the data from NaN values as follows:



We have extracted some new features from some fields. For example, from inspection\_date we got the year, month and day and added them into the data frame as follows:



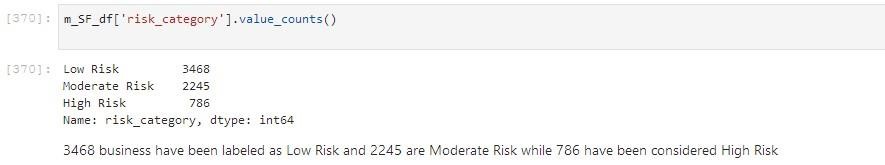
We clean up the address to have the neighborhood and so grouping businesses with respect to the neighborhoods.

We also have dropped some features and selected some features to be in the modeling process. So, before starting the modeling stage, we need to:

1. Convert the inspection\_date field to date time object
2. Convert categorical features to numerical values
3. Drop unnecessary fields from the data set
4. Find the number of each class is in our data set

# Modeling

After exploring the dataset and have a deep insight, we will apply a machine learning technique to classifying the inspection to have a better understanding of inspection process. We have three classes in this dataset as follows:



To perform a machine-learning technique on this dataset, we have selected two main algorithms to do the classification:

* K Nearest Neighbor (KNN)
* Logistic Regression

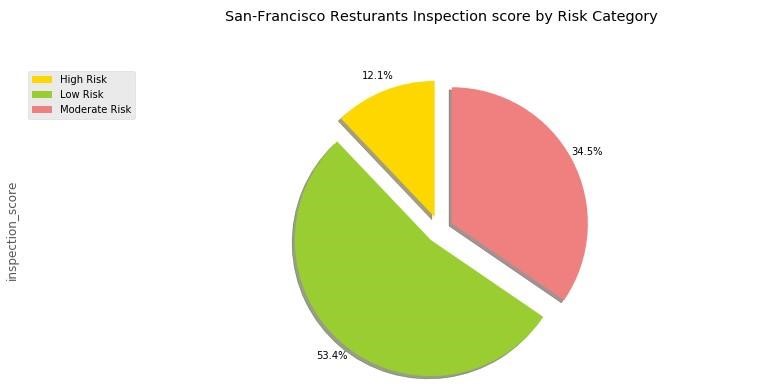
We have trained these two algorithms and tested them regarding our dataset.

# Evaluation and Testing

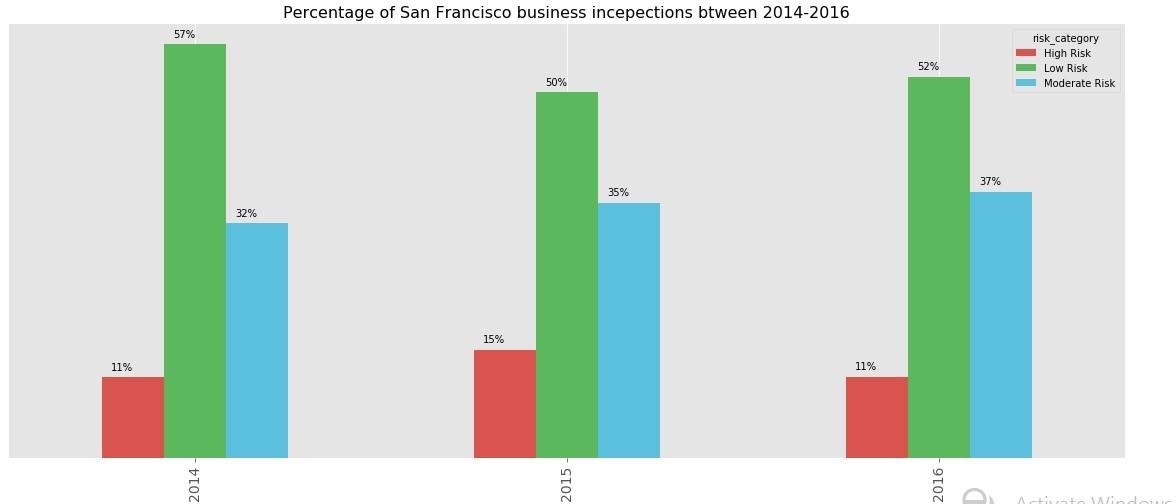
In this part we test the modeling algorithms by calculating the accuracy and f1-measure. We have also search for the best k that can give us the best classification model.

# Results

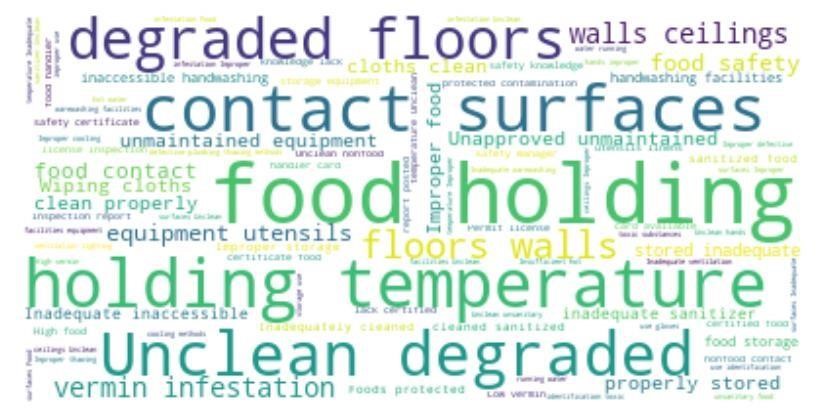
In this section, we can discuss some results that we have got from the analysis and modeling sections. We have started by examining the categories of the inspections that we have in the dataset. We found that, in general, 53.4% of the businesses are considered in low risk, 34.5% are in moderate risk, while the high-risk businesses are 12.1% as depicted below.



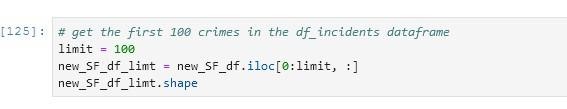
We grouped the inspections by year for each category low, moderate, and high risk. We have found that the High-Risk category increase by 4% from 11% in 2014, to 15% in 2015, and that is very interesting since it should be decreased not increased. Then, it deceased into 11% in 2016. This might lead to a conclusion that there was a deficiency of controlling the violation from 2014 to 2015, despite the lessening in 2016 because this percentage is not significant. Another observation that prove this conclusion is that the moderate category is always increasing from 32% in 2014, to 35% in 2015, to 37% in 2016 as illustrated below.



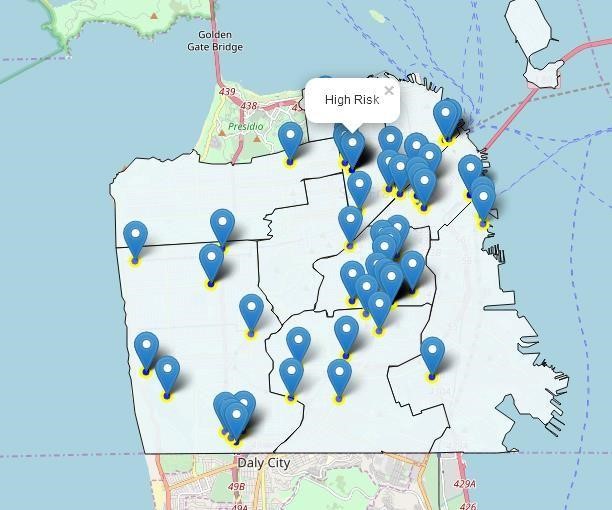
Using violation description, we count each violation's description words based on how much they contribute to the total inspections. We removed all stop-words here and created the word cloud below.



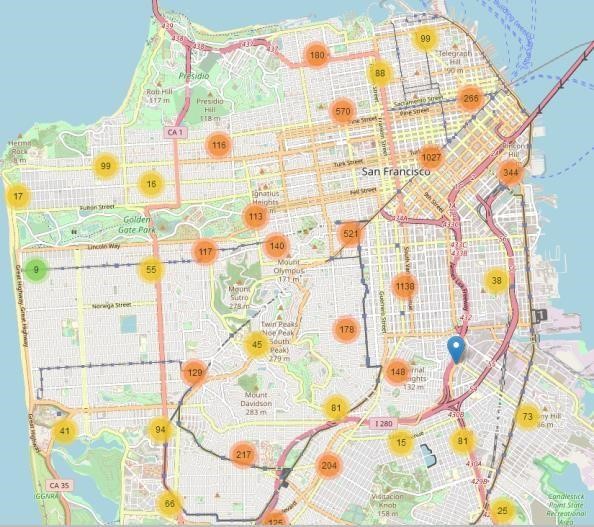
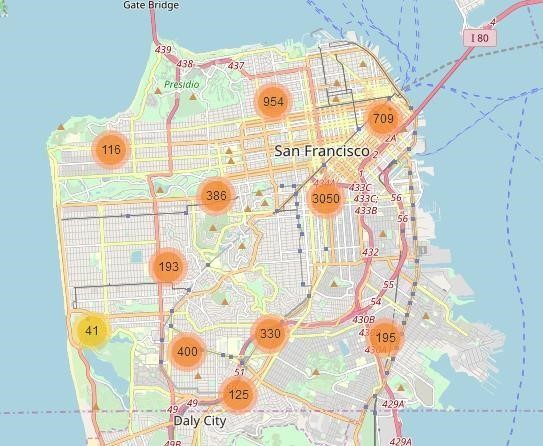
This cloud will give an indication of the most used terms that describe the violation problem. As we can see food, unclean, temperature, degraded, surfaces, contact, and floors are the main descriptions. We have used folium to visualize the locations of the inspections. To reduce computational cost, let's just work with the first 100 inspections in this dataset.



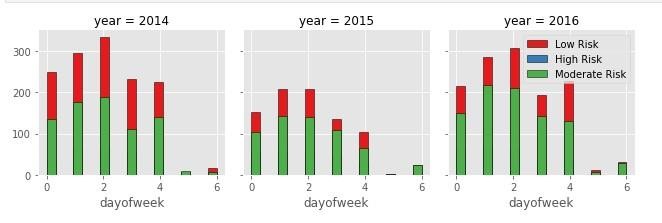
Now that we reduced the data, let's visualize where these inspections took place in San Francisco. We will use the default style and initialize the zoom level to 12. We superimpose the locations of the crimes onto the map. To do that in **Folium**, create a *feature group* with its own features and style and then add it to the sanfran\_map. We can also add some pop-up text that would get displayed when hovering over a marker. Let's make each marker display the category of the inspection when hovered over. The results were as depicted below.



We have grouped all business according to their categories. A clean and categorized copy of the map of San Francisco is shown below.



When we looked at the day of the week businesses were getting inspected, we have found that the inspection is very active in the beginning of the week and sharply decreased on Friday then increases little bit on Saturday as shown below.



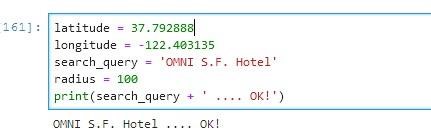
After modeling, we can see that in each year the low risk businesses are more than 50%. On the other hand, the High-Risk businesses are almost the same in both 2014 and 2016 while increasing in 2015 from 11% to almost 15%. Moreover, the Moderate Risk is increasing every year from 32% in 2014, to 35% in 2015, and to 36.5% in 2016. With kNN, the best accuracy was with 0.52. The following table show the results accuracy of our classification model.

|  |  |  |
| --- | --- | --- |
|  | kNN | LR |
| Training Set Accuracy | 0.6332499518953242 | 0.5358860881277661 |
| Test Set Accuracy | 0.5215384615384615 | 0.5246153846153846 |
| F1 Accuracy | 0.47777033142713926 | 0.36103702553753003 |

From the results in the table above, we can see that the accuracy is not that good and needs more features to get better. However, kNN perform better than LR in the training set and in accuracy of the F1 score as well.

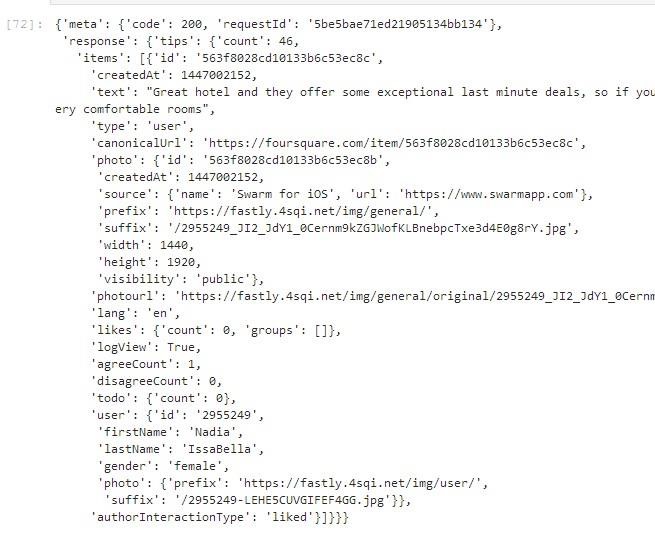
We also used Foursquare to analyze the neighborhoods of the inspected businesses. The Foursquare dataset comes with venue data which contains key descriptors of different venues including the category and popularity. This will show categories such as food establishments along with attributes like name, address, ratings, and reviews from millions of points of interest.

With Foursquare, you can also explore new destinations through learning likely interest of the user. In our case, our inspectors would be interested in food establishments and can get suggestions for new places where they can use to learn and even recognize food establishments around them that they never knew existed. To reduce computational cost, let's just work with the first 100 inspections in this dataset. Also, we investigate one of the businesses to check the venue and how people rate this business. We choose “OMNI S.F. Hotel”, it is Low Risk business and it got a 96 score rating from our collected dataset as follows:





Afterwards, we check the business rating. It falls in the Low Risk category, with an 8.5/10 rating. We can see it has a better rating, let's explore it further. This hotel has also 46 tips, however, because of the limits that Foursquare API gives us we are able to only show one tip.



We also examine the user who made that tip and we found that:

1. She is female
2. Here first name is Nadia and Last Name: IssaBella, Home City: Vaughan, Canada
3. Nadia is very active in Foursquare as we can see she has 598 tips. Let us explore them.

# Discussion and Importance of Food inspection

Food inspection helps promote food safety 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.

The large amount of diseases varies, most of which are infections acquired via contaminated food. The World Health Organization have scientifically proved over the years that preventive medicine is better than curative. Like many health matters, food safety is important to everyone involved. Here are a few people who would benefit from better food inspection:

* States and governments need better food inspection, hence food safety, to reduce financial burdens in the long run. Furthermore, food safety leads to a healthier population and a better workforce for the government.
* Citizens also directly benefit from food inspection because they can be protected from unnecessary life-threatening infections. They use their health for the betterment of themselves and other.
* Hospitals and medical practitioners are also happy when infections are prevented. Their workload reduces, and their patients get better. They can, in turn, dedicate their time and energy to other issues such as cancer research and technological innovations.

# Conclusion

To promote health, stakeholders in the healthcare industry need to continuously innovate to make the process more efficient. In food inspection, technology can be used to predict a likely critical violation using data analytics, instead of inspecting everything blindly given the lack of enough manpower for this. The data used to predict critical violation include weather, crime, and inspection data. Afterwards, Foursquare is used to locate the food establishment for physical inspection.