San-Francisco Food Inspection Report

In this report we will explore and analyze a dataset collected about San-Francisco food-businesses inspections. We analyze the collected given results of inspections in the last 3 years, from 2016 till 2018.

This project will introduce a business inspection predictive analytics report that can help promote business safety. It's also for the 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 food storage. Food inspection ensures that all these processes are done in such as a manner as to promote and achieve food safety.

Food inspection involves not only sampling and testing of end products but also assessing food centers to ensure compliance with food safety management systems. This minimizes the occurrence of public health food safety problems.
Food inspection dates back to ancient times as part of the history of public health. 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 e.g. nursing homes.

In the past, food inspection was done in a reactive manner whereby officers waited for reports of joints with 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 united states e.g. San Francisco are implementing a technologically driven approach to food inspection to try and 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.

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 waking of 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, systematically collected food inspection data from close to 100,000 sanitation inspections. Using this data, together with metadata on weather, related complaints e.g. sanitation, business characteristics, the city's advanced 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 violation include previous critical violations, high temperatures, nearby sanitation complains, nearby burglaries, etc

This report would be beneficial to public health specialists and every stakeholder working to alleviate public health concerns through preventive measures. It's 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.

Data

In this section I will explain the data that will be used to analyze the problem of food inspection and the source of the data. In order to develop a sufficient prediction system, the data should have the following categories:

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

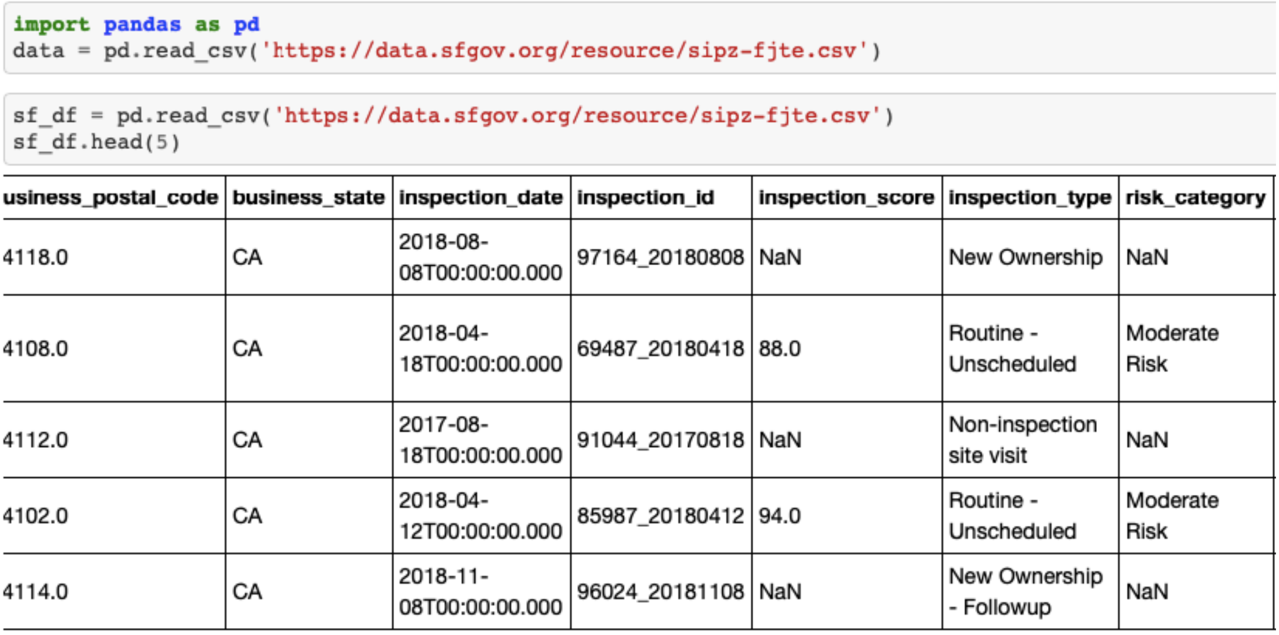
The data is collected from (https://data.sfgov.org/Health-and-SocialServices/Restaurant-Scores-LIVES-Standard/pyih-qa8i). The Health Department has developed an inspection report and scoring system. After conducting an inspection of the facility, the Health Inspector calculates a score based on the violations 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.

We downloaded the data from San-Francisco open data website. We explore and understand the data and 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:



```
import pandas as pd
data = pd.read_csv('https://data.sfgov.org/resource/sipz-fjte.csv')
```

```
sf_df = pd.read_csv('https://data.sfgov.org/resource/sipz-fjte.csv')
sf_df.head(5)
```

usiness_postal_code	business_state	inspection_date	inspection_id	inspection_score	inspection_type	risk_category
4118.0	CA	2018-08-08T00:00:00.000	97164_20180808	NaN	New Ownership	NaN
4108.0	CA	2018-04-18T00:00:00.000	69487_20180418	88.0	Routine - Unscheduled	Moderate Risk
4112.0	CA	2017-08-18T00:00:00.000	91044_20170818	NaN	Non-inspection site visit	NaN
4102.0	CA	2018-04-12T00:00:00.000	85987_20180412	94.0	Routine - Unscheduled	Moderate Risk
4114.0	CA	2018-11-08T00:00:00.000	96024_20181108	NaN	New Ownership - Followup	NaN

The following table give a brief description of the most important features in the dataframe:

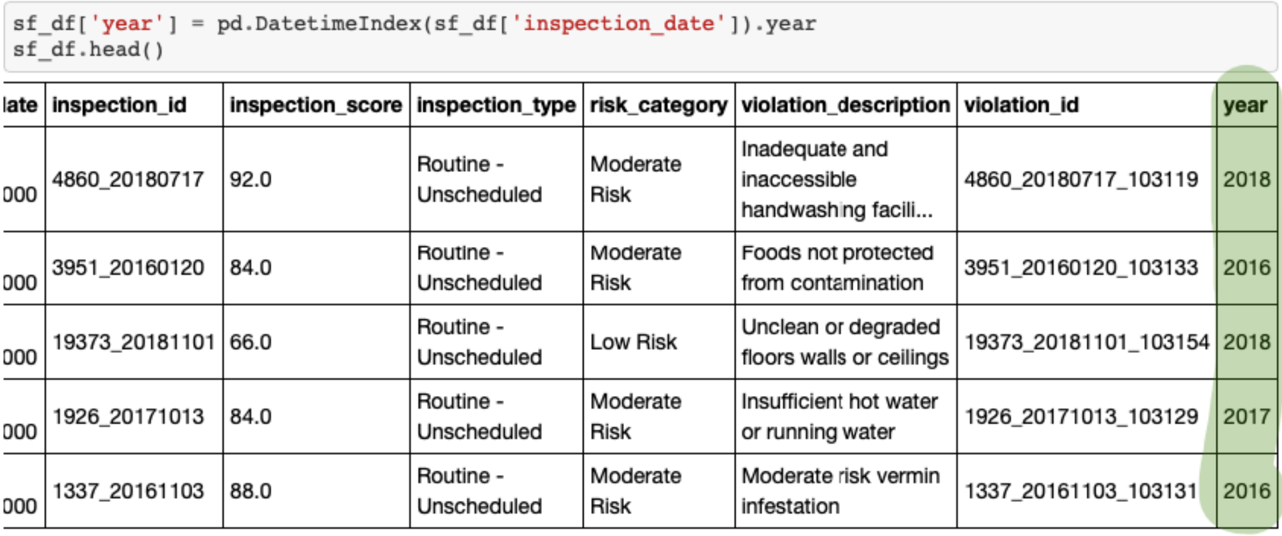
#	Feature Name	Description
1	business_id	Unique number used for identification of the business
2	business_name	Business Name
3	business_address	The address of the business
4	business_city	The City (here all records have the same city San-Francisco)
5	business_state	The state (here all records have the 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

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:

For example we have extracted some new features from some fields. For example, from inspection_date we got the year filtered out and added them as an extra column into the dataframe as follows:



```
sf_df['year'] = pd.DatetimeIndex(sf_df['inspection_date']).year
sf_df.head()
```

late	inspection_id	inspection_score	inspection_type	risk_category	violation_description	violation_id	year
000	4860_20180717	92.0	Routine - Unscheduled	Moderate Risk	Inadequate and inaccessible handwashing facili...	4860_20180717_103119	2018
000	3951_20160120	84.0	Routine - Unscheduled	Moderate Risk	Foods not protected from contamination	3951_20160120_103133	2016
000	19373_20181101	66.0	Routine - Unscheduled	Low Risk	Unclean or degraded floors walls or ceilings	19373_20181101_103154	2018
000	1926_20171013	84.0	Routine - Unscheduled	Moderate Risk	Insufficient hot water or running water	1926_20171013_103129	2017
000	1337_20161103	88.0	Routine - Unscheduled	Moderate Risk	Moderate risk vermin infestation	1337_20161103_103131	2016

Results

In this section we can discus 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, 44,6% of the businesses are considered in low risk, 47,7% are in moderate risk, while the high risk businesses are 7.7% as depicts in figure 2.

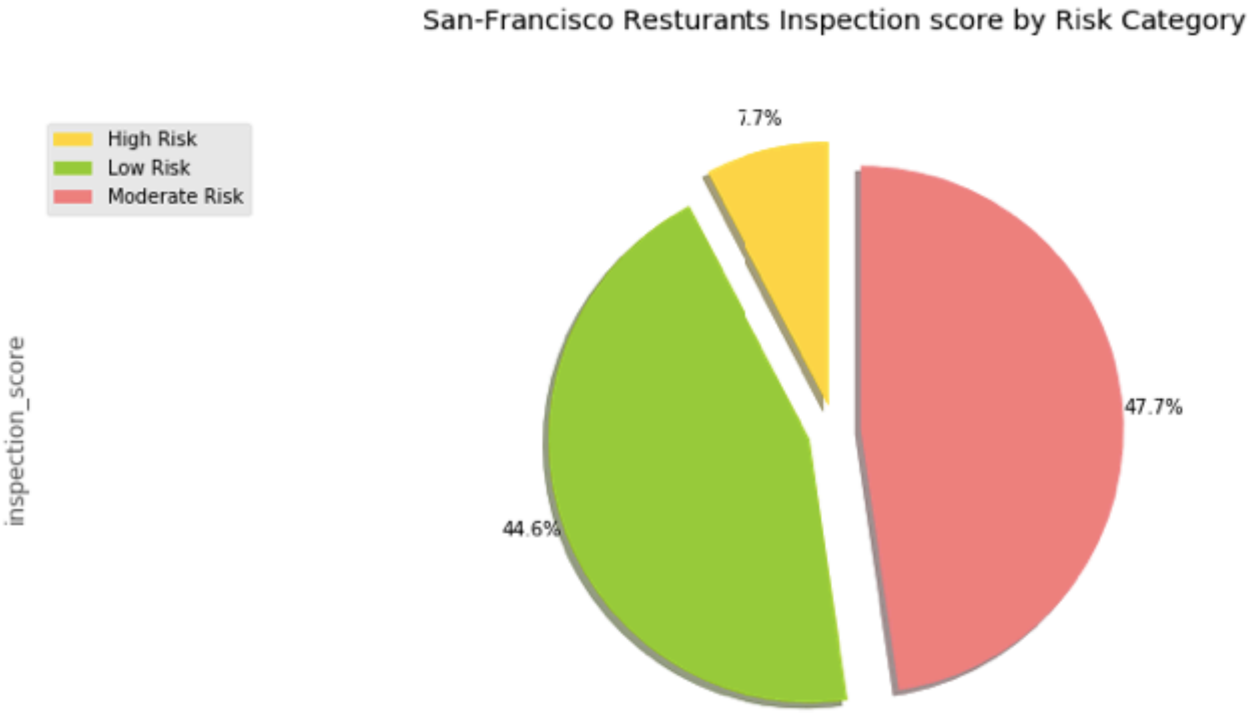
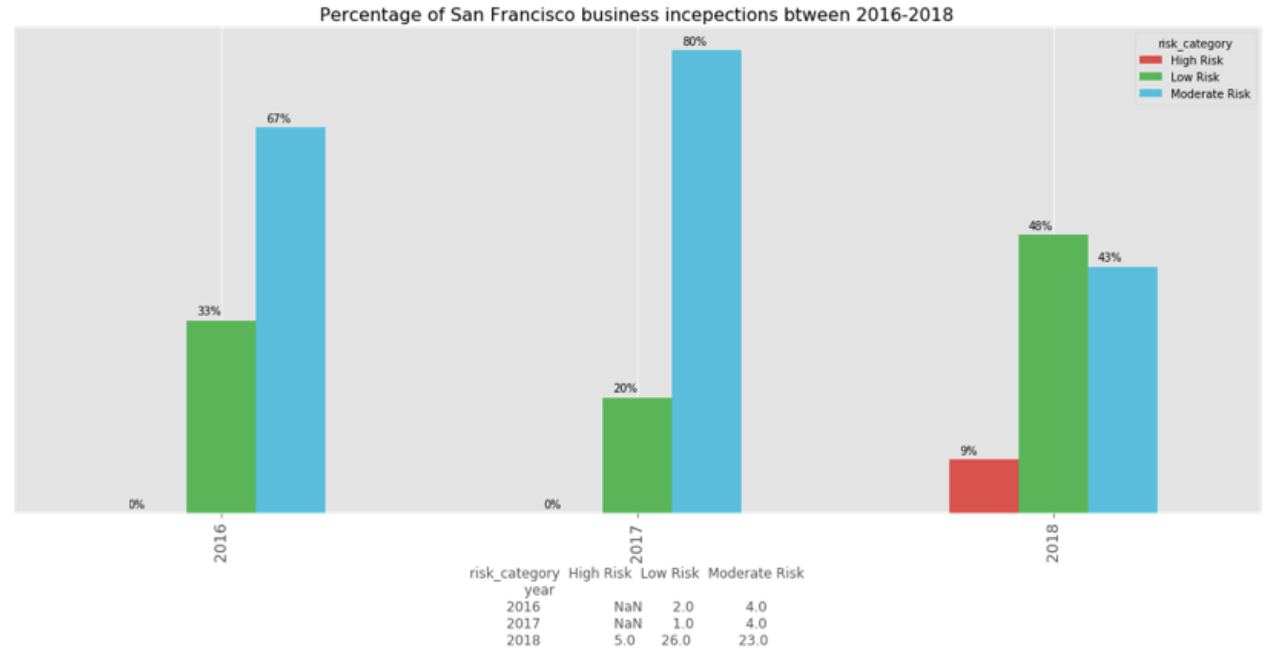


Figure 2 Risk category for SF businesses

We grouped the inspections by year for each category low, moderate and high risk. We have found that the Moderate category increase by 13% from 67% in 2016 to 80% in 2018. Then, it deceased into 43% in 2018, but the High Risk category increased by 9%.
Another observation is the moderate category is increasing from 20% in 2016 to 48% in 2017 and decreased to 37% in 2016 as illustrated in figure 3.



```
risk_category  High Risk  Low Risk  Moderate Risk
year
2016                 NaN       2.0            4.0
2017                 NaN       1.0            4.0
2018                 5.0      26.0           23.0
```

Figure 3: Inspections by year

Now let's visualize where these inspections took place in the city of San Francisco. We will use the default style and we will initialize the zoom level to 12. We superimpose the locations of the crimes onto the map. The way to do that in **Folium** is to 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 we hover over a marker. Let's make each marker display the category of the inspection when hovered over. The results were as depicted in figures 5 and 6.

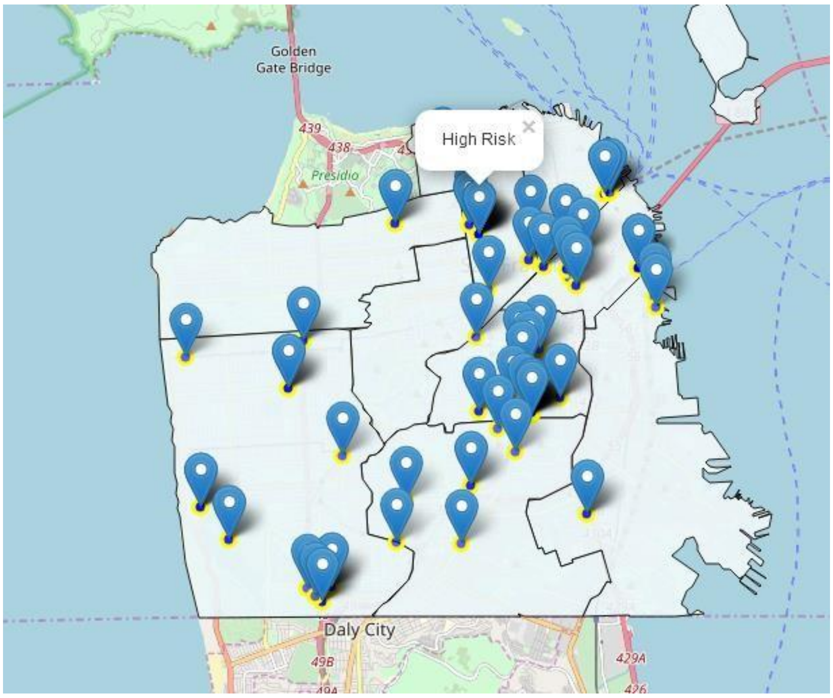


Figure 5 San-Francisco Inspection Map

We have grouped all business according to their categories. A clean and categorized copy of the map of San Francisco is shown in Figure 6.

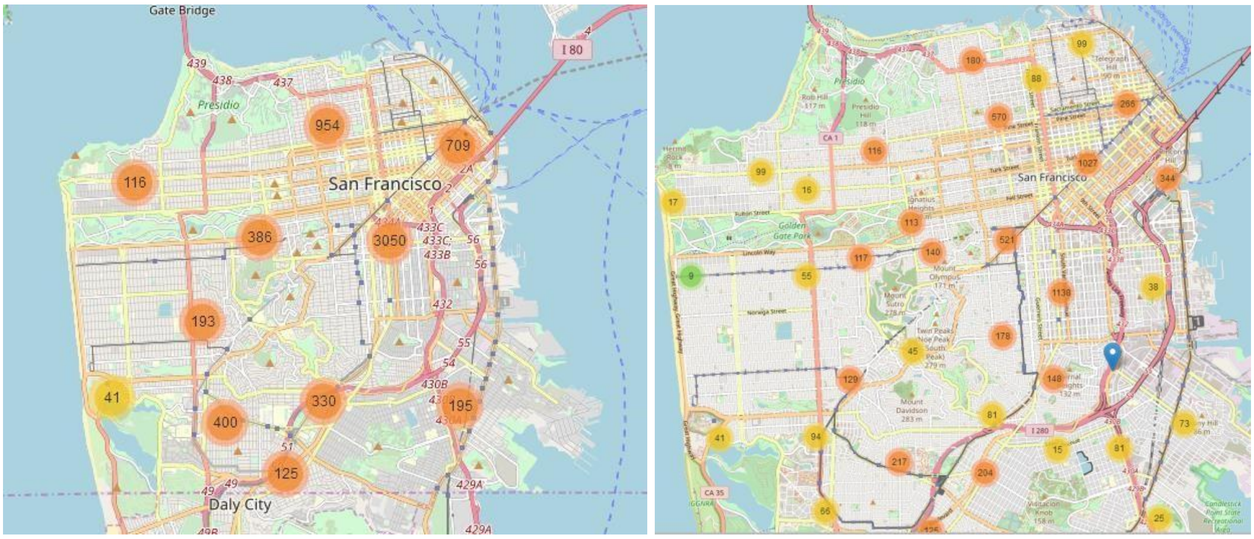


Figure 6:A clean and categorized copy of the map of San Francisco

When we looked at the day of the week businesses we're getting inspected, we have found that in the inspection is only active in weekdays and and last year sharply decreased on Thursday and Friday. But on these days there are more High Risk's.

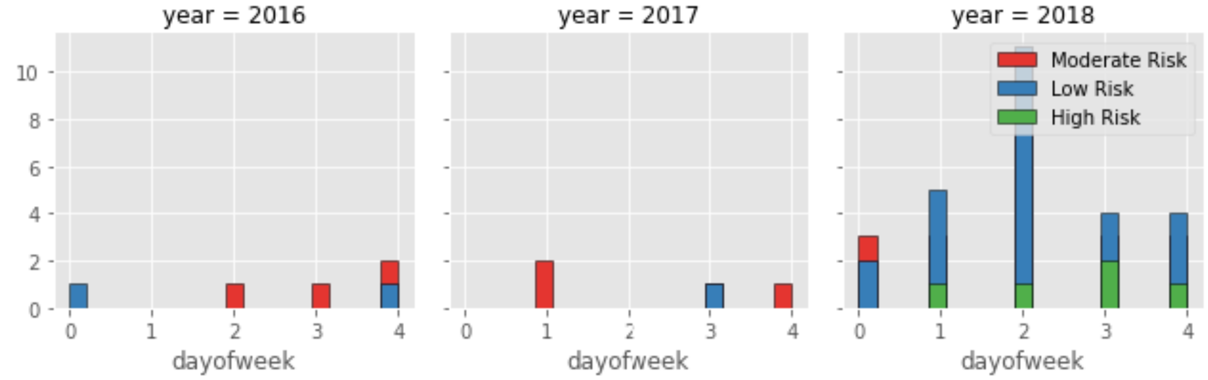


Figure 7: Inspection activities days of the week

The following table show the results accuracy of our classification model.

	kNN	LR
Train set Accuracy	0.6332499518953242	0.5358860881277661
Test set Accuracy	0.5215384615384615	0.5246153846153846
F1 Accuracy	0.47777033142713926	0.36103702553753003

From the result 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.

Discussion and Importance of Food inspection

Food inspection helps to promote food safety as part of the many processes 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.

Quite a big chunk of diseases repertoire is infection many of which are acquired via contaminated food. The World Health Organization has 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 and 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 are able to 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 minds and energy to other more pertinent issues like cancer research and technological innovations.

Conclusion

In the past, food inspection was done in a reactive manner whereby officers waited for reports of joints with possible non-compliance. Currently, some cities in the united states e.g. San Francisco, are implementing a technologically driven approach to food inspection to try and predict food establishments that are more likely to be non-compliant to food safety regulation.

To promote health, stakeholders in the healthcare industry need to continuously innovate to make this process more efficient. In food inspection, it's proven that technology can be used to predict a likely critical violation through the use of data analytics. This instead of inspecting every joint blindly given the lack of enough manpower for this.

Through the data in 2018 the Moderate Risk business decreased by 37% and the Low Risk Business increased by 20%. Because the inspectors knew where to locate the High Risk food establishment for physical inspection, the results of this category increased firmly last year.

That proves the available data that is used to predict critical violation helped in San Francisco to increase health.