Battle of Neighborhoods: Case Study of San Francisco Businesses by Anuj Nadig

In this report 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 and for example 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 its storage. Food inspection ensures that all these processes are done in such as a manner as to promote and achieve food safety.

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 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.

Data Description

In this section I will 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.

However, the data I have found 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 foodcontact 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.

Model Selection

There were several models that could be used to get good accuracy using this dataset. After performing Exploratory Data Analysis and testing different models on the dataset, I concluded that KNN and Logistic Regression were the best models. These two models work

the best because of the simplistic nature of the dataset wherein the features have strong correlation with the target variable.

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 in figure 1.

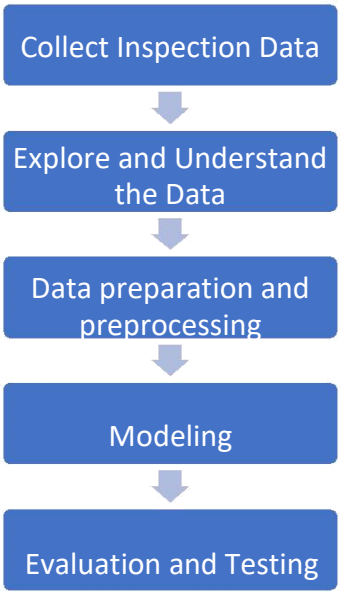


Figure 1 : Main Components of the Methodology

1. **Collect Inspection Data**

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

```
In [1]: !wget -q -O 'SF_Inspection.csv' https://data.sfgov.org/resource/sipz-fjte.csv
        print('Data downloaded!')

        Data downloaded!
```

The collected data are not ready for the analysis approach and need to be explored and organized.

2. **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:

```
[5]: sf_df = pd.read_csv('SF_Inspections.csv')
     sf_df.head(5)
```

[5]:

	business_id	business_name	business_address	business_city	business_state	business_postal_code	business_latitude	business_longitude	business_location	business_phone_number	inspection_id	inspection_date
0	10	Tiramisu Kitchen	033 Belden Pl	San Francisco	CA	94104	37.791116	-122.403816	(37.791116, -122.403816)	NaN	10_20140114	01/14/2014 12:00:00 AM
1	10	Tiramisu Kitchen	033 Belden Pl	San Francisco	CA	94104	37.791116	-122.403816	(37.791116, -122.403816)	NaN	10_20140114	01/14/2014 12:00:00 AM
2	10	Tiramisu Kitchen	033 Belden Pl	San Francisco	CA	94104	37.791116	-122.403816	(37.791116, -122.403816)	NaN	10_20140114	01/14/2014 12:00:00 AM
3	10	Tiramisu Kitchen	033 Belden Pl	San Francisco	CA	94104	37.791116	-122.403816	(37.791116, -122.403816)	NaN	10_20140729	07/29/2014 12:00:00 AM
4	10	Tiramisu Kitchen	033 Belden Pl	San Francisco	CA	94104	37.791116	-122.403816	(37.791116, -122.403816)	NaN	10_20140729	07/29/2014 12:00:00 AM

```
[88]: sf_df.shape

[88]: (53555, 17)
```

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. For more detail, we explain this section in details in ipython notebook, please check the file "*Week2_My_Capstone_Project_V10.ipynb*"

3. 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:

```
[178]: copy_sf_df.dropna(subset=['business_id','business_name',
                                 'business_address','business_city','business_state',
                                 'business_postal_code','business_latitude','business_longitude',
                                 'business_location','business_phone_number','inspection_id',
                                 'inspection_id','inspection_date','inspection_score','inspection_type',
                                 'violation_id','violation_description'],inplace=True)
```

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 dataframe as follows:

```
[185]: copy_sf_df['year']= copy_sf_df['inspection_date'].apply(lambda x: getYear(str(x)))
       copy_sf_df['Month']= copy_sf_df['inspection_date'].apply(lambda x: getMonth(str(x)))
       copy_sf_df['day']= copy_sf_df['inspection_date'].apply(lambda x: getDay(str(x)))
       copy_sf_df.head(10)
```

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 start the modeling stage We need to:

- convert inspection_date field to date time object
- Convert Categorical features to numerical values
- drop unnecessary fields from the dataset
- how many of each class is in our dataset

4. Modeling

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

```
[370]: m_SF_df['risk_category'].value_counts()

[370]: Low Risk         3468
       Moderate Risk    2245
       High Risk         786
       Name: risk_category, dtype: int64
```

3468 business have been labeled as Low Risk and 2245 are Moderate Risk while 786 have been considered High Risk

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

5. 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 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, 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 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 High Risk category increase by 4% from 11% in 2014 to 15% in 2015 and that is very interesting where it should be decreased not increase. 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 proof this conclusion is the moderate category is always increasing from 32% in 2014 to 35% in 2015 to 37% in 2016 as illustrated in figure 3.

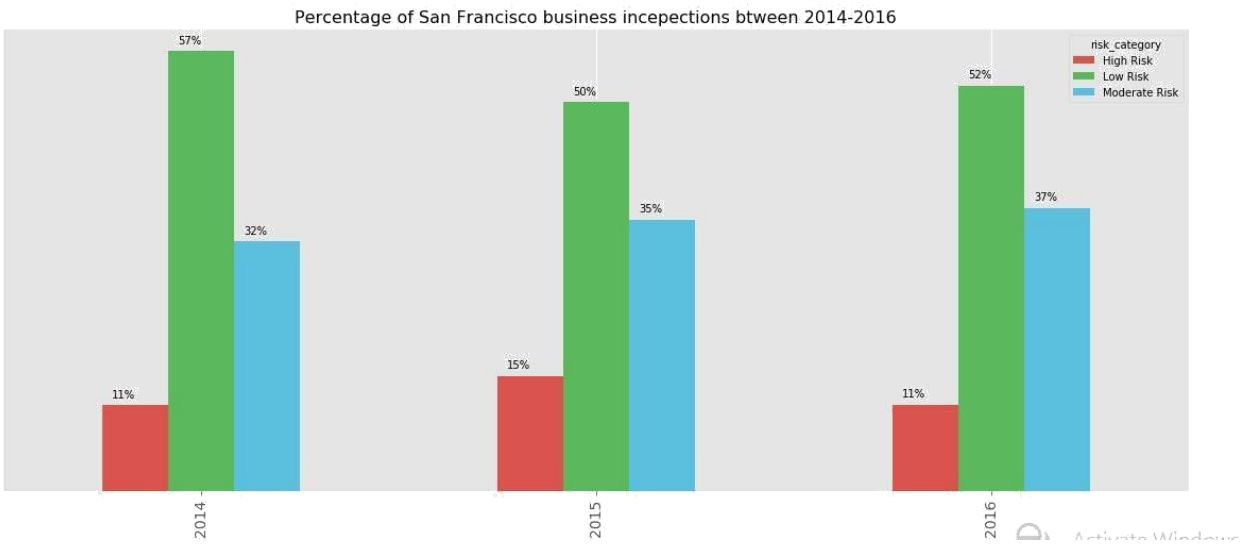


Figure 3: Inspections by year

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 as shown in figure 4.

Figure 4: Violation description cloud words

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. **In order to reduce computational cost, let's just work with the first 100 inspections in this dataset**.

```
[125]: # get the first 100 crimes in the df_incidents dataframe
limit = 100
new_SF_df_limt = new_SF_df.iloc[0:limit, :]
new_SF_df_limt.shape
```

Now that we reduced the data a little bit, 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.

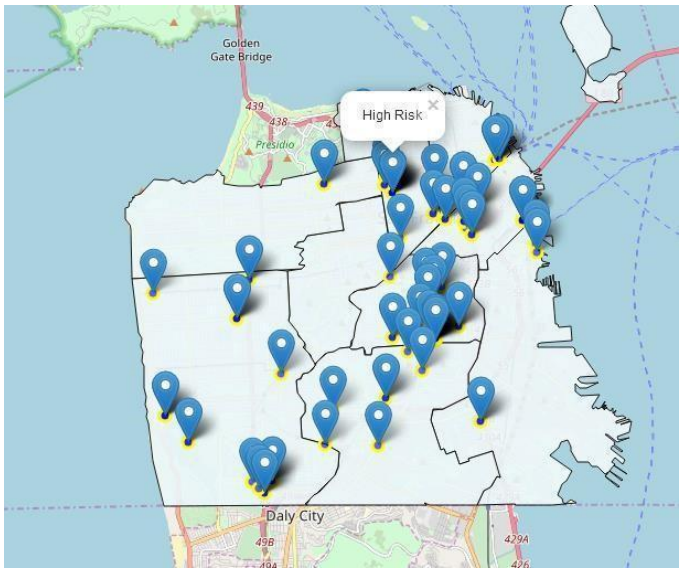


Figure 5 San-Francisco Inspection Map

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.

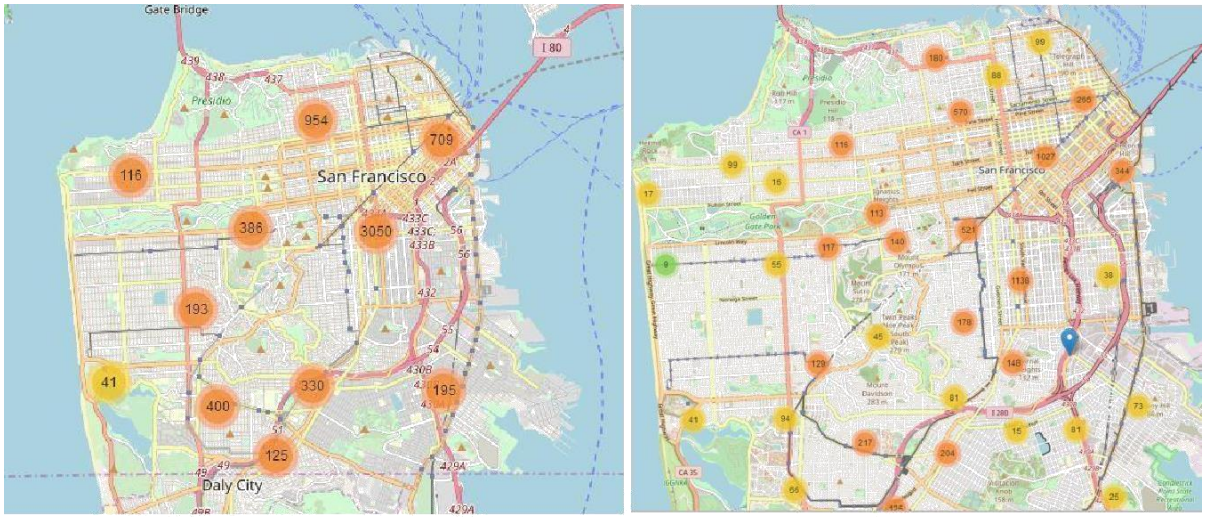


Figure 6:A clean and categorized copy of the map of San Francisco

When we looked at the day of the week businesses 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 in figure 7.

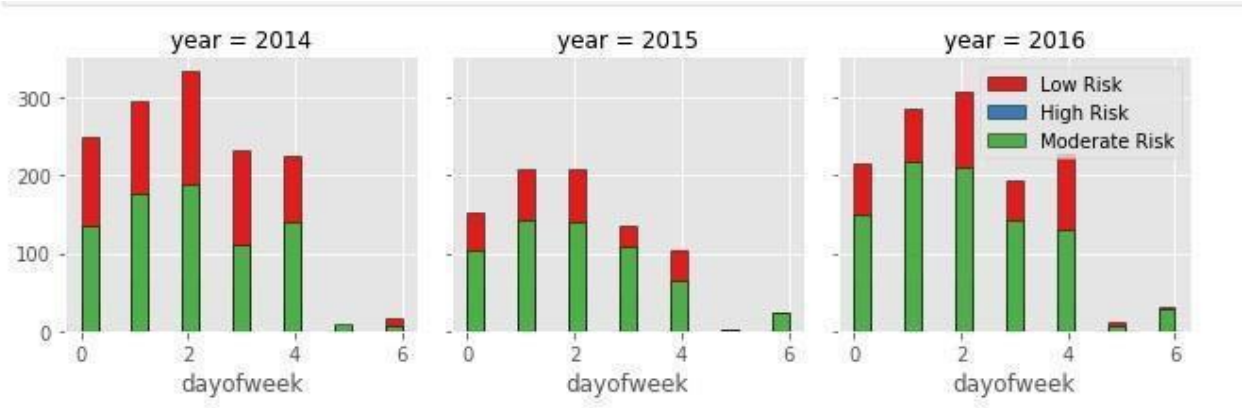


Figure 7: Inspection activities days of the week

After modeling we can see that in all years 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 36.5% in 2016. With kNN the best accuracy was with 0.52 when k= 6. 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.

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

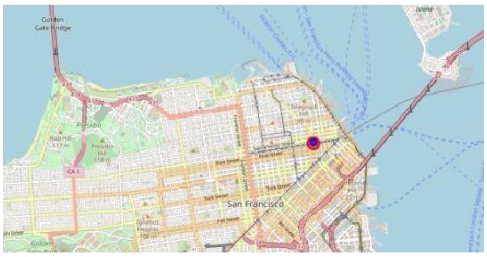
With Foursquare, you can also explore new destinations through learning likely interest of the user. In our case, our inspectors would be interested in food establishments and can get suggestions for new places where they can use to learn and even recognize food establishments around them that they never knew existed. **In order 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

inspection and it got 96 score from our collected dataset as follows:

```
161]: latitude = 37.792888
      longitude = -122.403135
      search_query = 'OMNI S.F. Hotel'
      radius = 100
      print(search_query + ' .... OK!')
```

```
OMNI S.F. Hotel .... OK!
```

```
[164]: {'meta': {'code': 200, 'requestId': '5be6fc239fb6b71ed1998912'},
        'response': {'venues': [{'id': '4a5ae9a1f964a520e0ba1fe3',
           'name': 'Omni San Francisco Hotel',
           'location': {'address': '500 California St',
            'crossStreet': 'at Montgomery St',
            'lat': 37.793119745957455,
            'lng': -122.4031025916338,
            'labeledLatLngs': [{'label': 'display',
              'lat': 37.793119745957455,
              'lng': -122.4031025916338}],
            'distance': 25,
            'postalCode': '94104',
            'cc': 'US',
            'city': 'San Francisco',
            'state': 'CA',
            'country': 'United States',
            'formattedAddress': ['500 California St (at Montgomery St)',
             'San Francisco, CA 94104',
             'United States']},
           'categories': [{'id': '4bf58dd8d48988d1fa931735',
             'name': 'Hotel',
             'pluralName': 'Hotels',
             'shortName': 'Hotel',
             'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/travel/hotel_',
              'suffix': '.png'},
             'primary': True}],
           'venuePage': {'id': '497545373'},
           'referralId': 'v-1541864483',
           'hasPerk': False}]}}
```



After we check this business rating since it has a Low Risk category, it has 8.5/10 We can see it has a better rating, let's explore it further. This hotel has also 46 tips and because of the limitation that Foursquare API gives us we are able to only show one tip.

```
[72]: {'meta': {'code': 200, 'requestId': '5be5bae71ed21905134bb134'},
       'response': {'tips': {'count': 46,
         'items': [{'id': '563f8028cd10133b6c53ec8c',
           'createdAt': 1447002152,
           'text': "Great hotel and they offer some exceptional last minute deals, so if you
      ery comfortable rooms",
           'type': 'user',
           'canonicalUrl': 'https://foursquare.com/item/563f8028cd10133b6c53ec8c',
           'photo': {'id': '563f8028cd10133b6c53ec8b',
            'createdAt': 1447002152,
            'source': {'name': 'Swarm for iOS', 'url': 'https://www.swarmapp.com'},
            'prefix': 'https://fastly.4sqi.net/img/general/',
            'suffix': '/2955249_JI2_JdY1_0Cernm9kZGJWofKLBnebpcTxe3d4E0g8rY.jpg',
            'width': 1440,
            'height': 1920,
            'visibility': 'public'},
           'photourl': 'https://fastly.4sqi.net/img/general/original/2955249_JI2_JdY1_0Cernm
           'lang': 'en',
           'likes': {'count': 0, 'groups': []},
           'logView': True,
           'agreeCount': 1,
           'disagreeCount': 0,
           'todo': {'count': 0},
           'user': {'id': '2955249',
            'firstName': 'Nadia',
            'lastName': 'IssaBella',
            'gender': 'female',
            'photo': {'prefix': 'https://fastly.4sqi.net/img/user/',
             'suffix': '/2955249-LEHE5CUVGIFEF4GG.jpg'}},
           'authorInteractionType': 'liked'}]}}}
```

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.

Please go to the python notebook for more details.

Discussion and Importance of Food inspection

Food inspection help 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.

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

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 through the use of data analytics instead of inspecting every joint blindly given the lack of enough manpower for this. The data used to predict critical violation include

weather, crime and inspection data. Afterward, places data e.g. Foursquare is used to locate the food establishment for physical inspection.