DINESAFE

GROUP 6

Hwarang Kim

Raaid Batarfi

Jiarun He

Mehdi Ostadhashem

David Scott

3250-Term Project Presented to

UNIVERSITY of TORONTO

in partial fulfillment of the requirements for

SCS 3250-021 Foundation of Data Science

Toronto, Ontario, Canada, 2018

DINESAFE

EXECUTIVE SUMMARY

DineSafe is Toronto Public Health's food safety program that inspects all establishments serving and preparing food. Each inspection results in a pass, a conditional pass or a closed notice. Food premises are inspected at a frequency established by the Ministry of Health according to identified risk levels. The DineSafe program aims to reduce the incidence of foodborne illness among consumers in Toronto. The purpose of this analysis is to shed light on the state of Toronto's Food Preparation industry and the effectiveness of the DineSafe program.

The DineSafe program has not been effective at increasing the passing rate of establishments in Toronto. The pass rate of establishments has been decreasing over the 25 months analyzed. It is clear that the high risk establishments are the category that is driving the decrease in pass rate. It has also been shown that the number of inspections has an impact on the Establishment Safety Score. The more frequently an establishment is inspected, the more likely that establishment is to receive a pass on any given inspection.

We recommend that Toronto Public Health increase the inspection frequency of high risk establishments in order to achieve their desired outcome of an increased inspection pass rate.

CONTENTS

1. Project Overview

The objective of this analysis is to shed light on the effectiveness of the DineSafe inspection system. The system is prepared by the Toronto Public Health's food safety program and is based on the Provincial and Municipal regulations, which is released by the city of Toronto as public information. The program focuses on the inspection of all establishments serving and preparing food in the city of Toronto. Each inspection of an establishment results in a pass, a conditional pass or a closed notice. Food premises are inspected at a frequency established by the Ministry of Health according to identified risk levels. Every establishment in the City of Toronto receives a minimum of 1, 2, or 3 inspections each year depending on the specific type of establishment, the processes and preparation of the food, the type of food served, the volume of food served, and other related criteria. There are three different risk levels that are used to categorize each establishment.These risk levels are as follows:

A. *High Risk Level*: which includes any eating or drinking establishments that prepare hazardous food and meet at least one of the following criteria:
   - Serve a high-risk population.
   - Use processes involving many preparation steps and foods frequently implicated as the cause of foodborne illness.
   - Implicated or confirmed as a source of foodborne illness/outbreak.
   - Minimum inspections: three times per year.

B. *Moderate Risk Leve*l: which includes any eating or drinking establishments that meet one or more of the following criteria:
   - Prepare hazardous food without meeting the criteria for high risk.
   - Prepare non-hazardous foods with extensive handling or high volume.
   - Minimum inspections: two times per year.

C. *Low Risk Level*: This risk level includes any eating or drinking establishments that do not prepare hazardous food and meet one or more of the following criteria:
   - Serve pre-packaged hazardous foods
   - Prepare and/or serve non-hazardous foods without meeting the criteria for moderate risk
   - Establishments that are used as a food storage facility for non-hazardous foods only
   - Public health concerns related primarily to sanitation and maintenance
   - Minimum inspections: one time per year

2. Data Preparation

What was the data source?

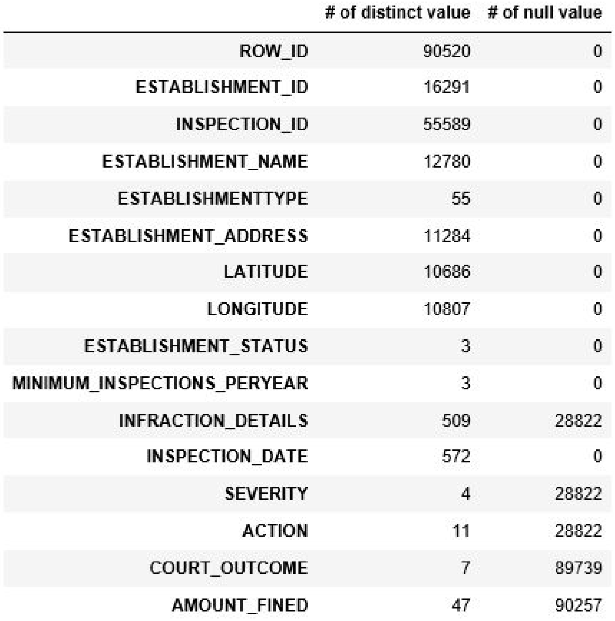
The dataset in this report is owned by the Toronto Public Health's food safety program and is available to download from the City of Toronto open data catalogue: website. The dataset had around 90,520 records, and covered over 16,291 establishments which are all located in Toronto, Ontario. The dataset had records for a period of 25 months (from September 2016 to September 2018). Each row of the dataset represented registered records of one establishment and one inspection result (pass, conditional pass, or closed) and one infraction detail (if inspection finds no infraction, infraction detail has null value) by an inspection date. Note that one establishment could have multiple infraction details in a particular date. For example, if an establishment had an inspection on an inspection date and the inspection led to three infractions, there would be three records for this instance.

How good was the data quality?

As we had worked on the dataset, we concluded that the quality of the dataset was decent. Most columns in the dataset contained non-numeric values. We treated most of these non-numeric values as strings. ESTABLISHMENT_ID and INSPECTION_ID are numeric values but we treated them as strings since no calculation was applied to these two columns.

Moreover, there were several missing values or details in the dataset (see Table 1). Out of 16 columns in the dataset, the following columns have null values: INFRACTION_DETAILS, SEVERITY, ACTION, COURT_OUTCOME, and AMOUNT_FINED (see data dictionary in Appendix **[A-2.2]**). These columns allowed missing values intentionally when there was no infraction, no court order or no fine.

Table 1 - DineSafe dataset overview (count of distinct value and null value)



	# of distinct value	# of null value
ROW_ID	90520	0
ESTABLISHMENT_ID	16291	0
INSPECTION_ID	55589	0
ESTABLISHMENT_NAME	12780	0
ESTABLISHMENTTYPE	55	0
ESTABLISHMENT_ADDRESS	11284	0
LATITUDE	10686	0
LONGITUDE	10807	0
ESTABLISHMENT_STATUS	3	0
MINIMUM_INSPECTIONS_PERYEAR	3	0
INFRACTION_DETAILS	509	28822
INSPECTION_DATE	572	0
SEVERITY	4	28822
ACTION	11	28822
COURT_OUTCOME	7	89739
AMOUNT_FINED	47	90257

How did you procure the data?

The dataset was downloaded in **xml** format from the website mentioned above. For the purpose of our analysis the dataset was converted to **csv** format. The command used to read the dataset was *pandas.read_csv* method which allows us to load the full dataset into a *DataFrame*. See Appendix **[A-2.1]**.

What tools/code did you use to prepare for the analysis?

In this project we used Python (pandas, numpy, datetime, and matplotlib.pyplot). For example some data preparation used in this project include the following but are not limited to:

- Created eight extra columns (all extracted from the **INSPECTION_DATE** column) which was used to do further analysis to the dataset. Four of these extra columns separate the **INSPECTION_DATE** column into individual columns (e.g. ***Year, Quarter, Month*** *and* ***Week***). The other four columns included were "***Year and Quarter", "Year and Month"*** *and* ***"Year and Week"*** and ***"Week day"***. See **[A-2.3]** for the sample code.
- In order to prepare data for the analysis, we created the functions to generate the aggregated datasets and add the calculated columns. See **[A-2.4]** for the code:
  - **count_inspection()**
  - **count_infraction()**
  - **ration_infr_insp()**
  - **pass_rate_cal()**
- The aggregated datasets were ranked by number of infraction to identify top 15 establishment types and names by using "value_counts()" and ".head(15)()". See **[A-2.5]** for the sample code

What challenges did you face?

While preparing the data for analysis we faced some challenges. These challenges include, but are not limited to, the following:

- When we downloaded the dataset, INSPECTION_DATE was in text format. In order to prepare the dataset for time series analysis, this date columns needed to be converted into date-time format.

- The data did not have the city or postal codes, while instead it had the address, latitude and longitude. To fetch the city and postal codes, a Google API was used. However, approximately 100,000 records needed to be processed, with the average call rate on a PC being 6000 calls per hour. For enhancing the process, another dataframe has been created with unique latitude and longitude. Postal codes and areas (cities) have been added to this dataframe and saved as a separate csv file for any further usage.

- Any interruption or error in calling the Google API could break "apply()" function, and it rolls back all the records and does not keep the fetched data. For resolving the issue, the process was broken into 50-record segments so that any break would result in the loss of <50 records. Finally, we found there was an error in the longitude of the data which has been fixed.

- We also noticed that the dataset has multiple establishment names for one establishment ID. For example, Starbucks Coffee has two different names "Starbucks Coffee" and "Starbucks". Some other establishments has double space in the establishment name. These have made it a very challenging to perform an analysis on the establishments name.

3. Analysis: Trends, Correlations, & Patterns

3.0 Overview

This section contains all details of the analysis performed on the DineSafe dataset. The DineSafe program aims to reduce the incidence of foodborne illness among consumers in Toronto. The purpose of this analysis is to shed light on the state of Toronto's Food Preparation industry and the effectiveness of the DineSafe program.

The following analysis sections are centered around three key metrics:

1. Number of safety infractions per inspection (also referred to as infraction ratio),
2. Inspection pass rate, and
3. Establishment safety score

A low infraction per inspection ratio and high pass rate correspond to increased food safety. The establishment safety score is an engineered attribute that quantifies the relative safety of each establishment.

3.1 Infraction Ratio

The infraction ratio refers to the number of safety infractions found per inspection. The following 5 attributes (all available in the DineSafe dataset) were used to create and analyze the infraction ratio:

1. Establishment Name
2. Establishment Type (ie Food Court, Bakery, Restaurant etc)
3. Minimum Required Number of Inspections per Year (Based on the establishment risk level)
4. Infraction Details
5. Inspection Date

The infraction ratio was analyzed to get an idea of how many safety violations are typically found for different types of segments. By plotting the infraction ratio for the entire dataset we were able to visualize the general month-over-month trend (Figure 3.1.1). The infraction ratio reached a maximum in April, 2018 at 1.4 infractions per inspection and has been decreasing since. Another interesting observation is the low number of inspections that occured in December, 2016. When further segmenting the infraction ratio into high, medium, and low risk establishments, it is clear that the high risk establishment segment is the driver of a high infraction ratio (Figure 3.1.2). Medium risk establishments have an infraction ratio of ~1 while low risk establishments have a ratio of ~0.5. This indicates that low risk establishments are just as likely to have no infractions as they are to have 1 or more infractions. In contrast, the high risks establishments are more likely to have multiple infractions per inspection.

Figure 3.1.1 - Infraction Ratio Trend

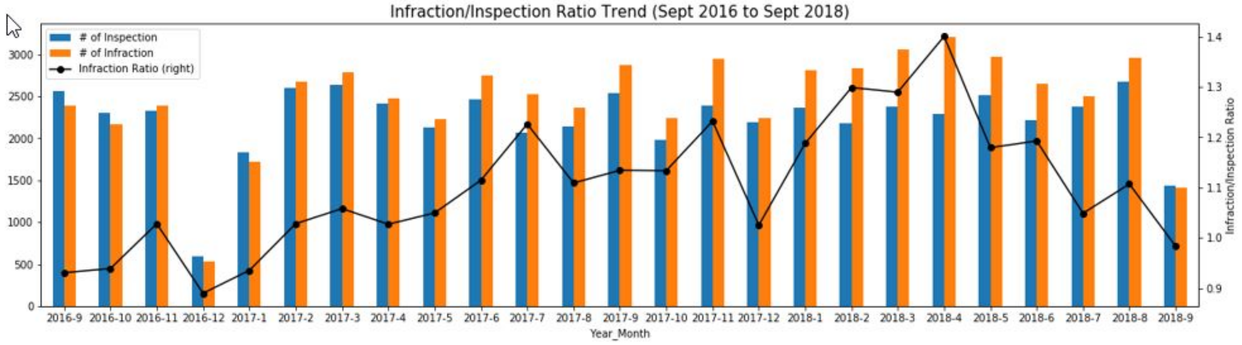
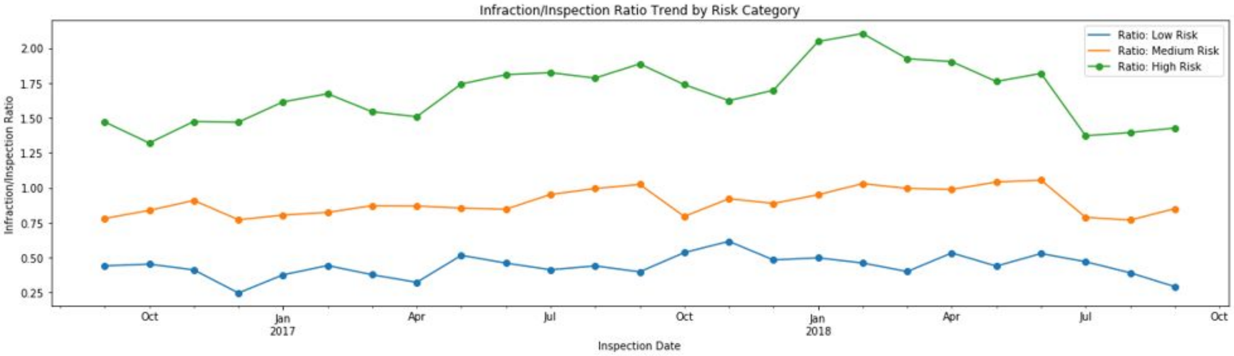


Figure 3.1.2 - Infraction Ratio by Establishment Risk Level



3.2 Pass Rate

The pass rate refers to the number of inspections that received a "pass" compared to the total number of inspections. This sections answers some important questions that arose during our analysis.

Is the DineSafe program working?

The pass rate has been decreasing over the period from September 2016 to September 2018 (Figure 3.2.1). In an attempt to uncover a trend in the pass rate, our team plotted the autocorrelation (Figure 3.2.2). As shown, there is no significant self correlation of the data after a few months of lag (plot remains within the two dashed lines). Similar to the analysis of the infraction ratio, it is useful to segment the pass rate by establishment risk level. Figure 3.2.3 shows that the decreasing trend in pass rate can be attributed to a drop in the passing rate of high risk establishments over time. Overall, the DineSafe program has not been effective at increasing the passing rate of establishments in Toronto.

Figure 3.2.1 - Pass Rate

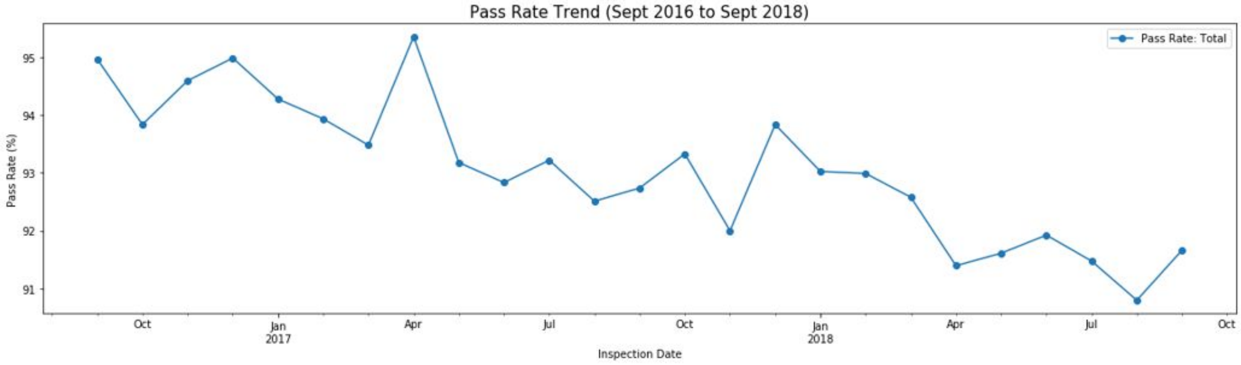
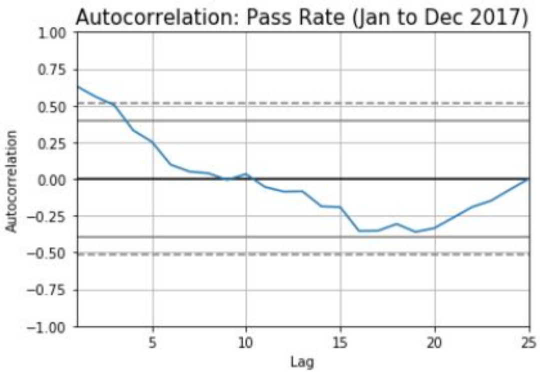
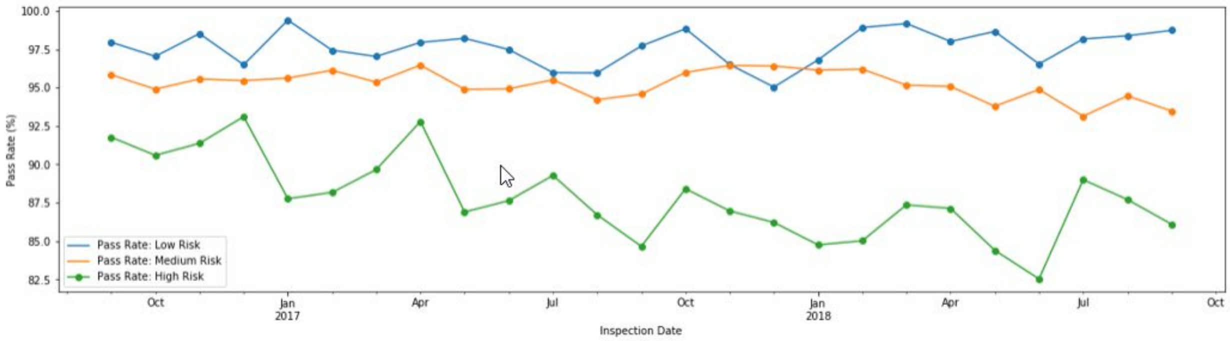


Figure 3.2.2 - Autocorrelation of Pass Rate



3.2.3 - Pass Rate by Establishment Risk Level



What are the safest food preparation establishments? Are some areas safer than others?

Out of the most commonly inspected establishments, Second Cup and Thai Express have surprisingly low pass rates. Other notable franchises such as Tim Hortons and Subway have exceptionally high pass rates of 98% and 99% respectively (Figure 3.2.4). By dividing the GTA into 6 sublocations, Etobicoke is revealed as the safest place to eat while North York has the lowest passing rate of 90% (Figure 3.2.5).

Figure 3.2.4 - Pass Rates of Most Common Establishment Names

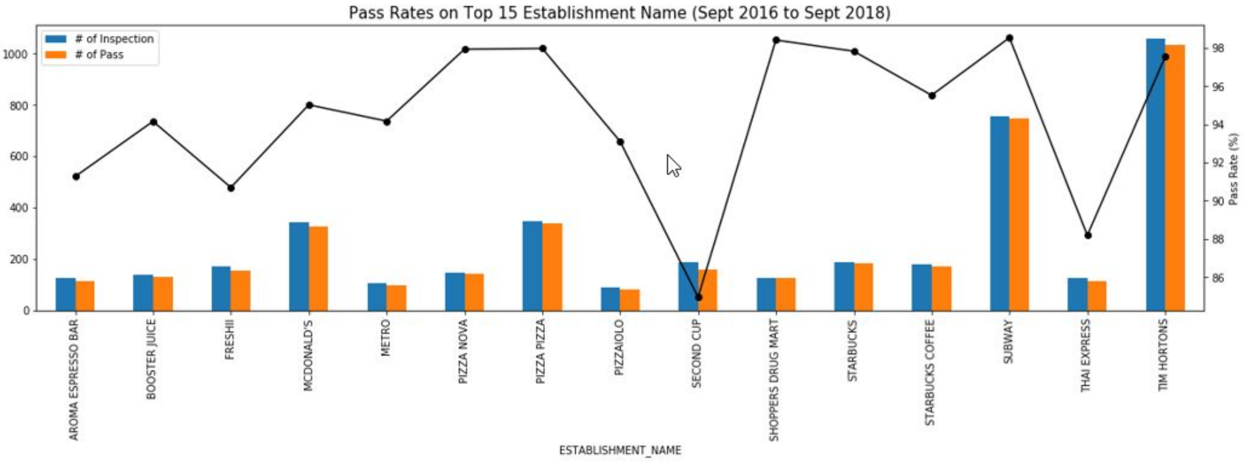
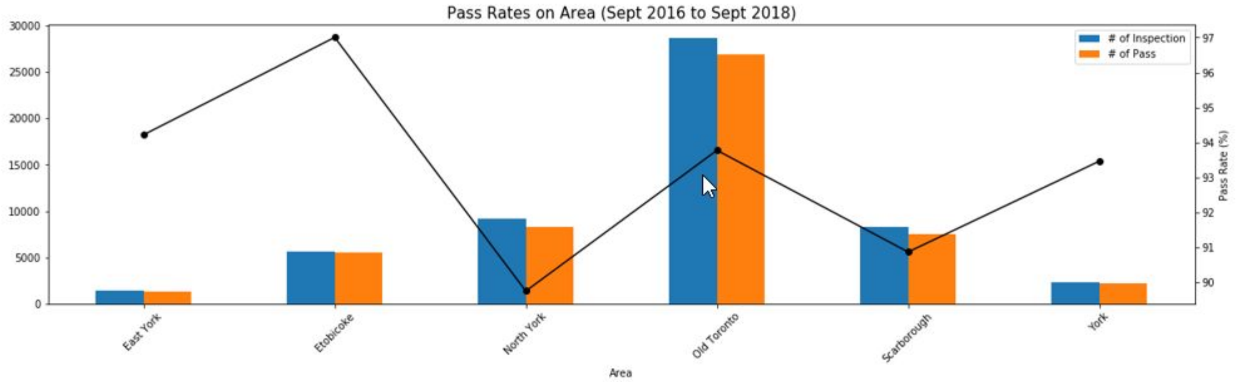


Figure 3.2.5 - Pass Rate by Location



Does the day of the inspection affect the result?

Most of the inspections take place on Monday through Friday. Friday is shown to have a minor (though significant) increase in passing rate over the other 4 days of the week (Figure 3.3.6). Interestingly, the pass rate also appears to be impacted by seasonality. The number of inspections each month remains fairly constant, but the passing rate decreases into the Summer and increases dramatically in September (Figure 3.2.7).

Figure 3.3.6 - Pass Rate by Day of the Week

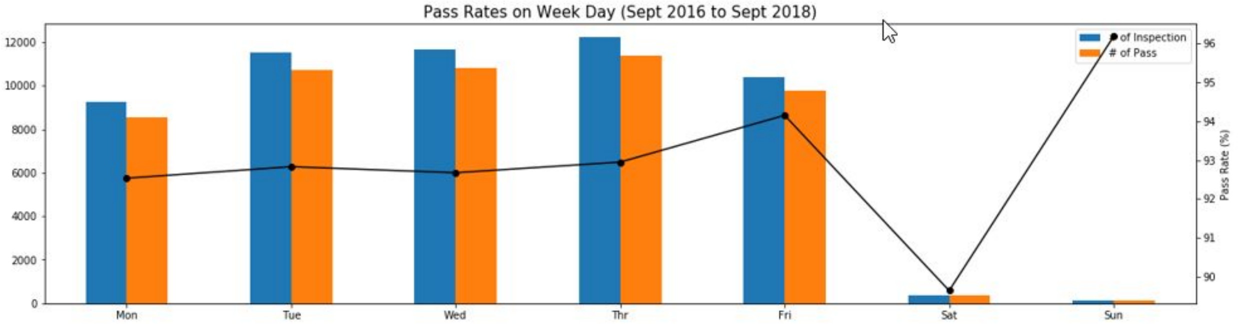
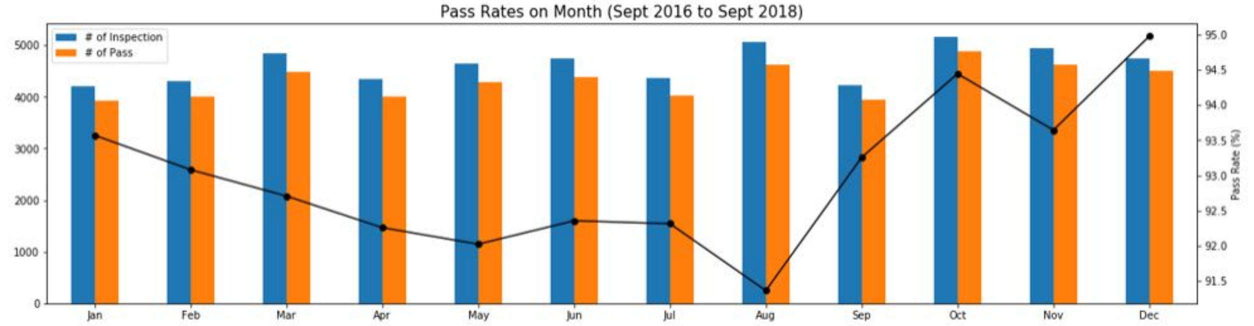


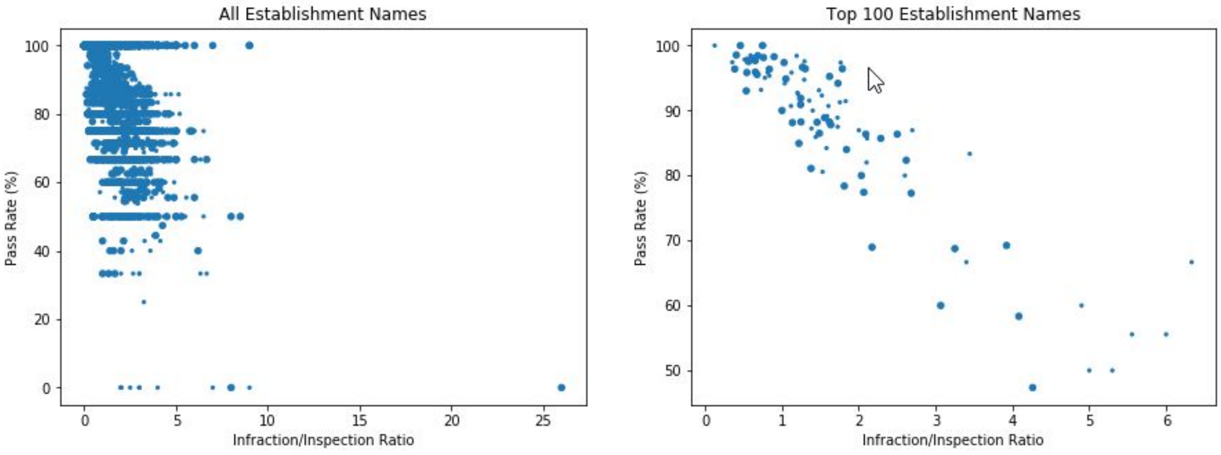
Figure 3.2.7 - Pass Rate by Month

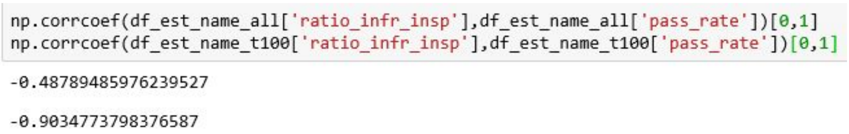


3.3 Infraction Ratio Compared to Pass Rate

In order to understand the correlation of infraction ratio and pass rate, data was aggregated on Establishment Name. Figure 3.3.1 shows that there is weak correlation (-0.4878) when analyzing the entire data set. However, when analyzing the top 100 establishments based on number of inspections, there is a strong negative correlation (-0.9034) between infraction ratio and pass rate.

Figure 3.3.1 - Correlation of Pass Rate and Inspection Ratio





```
np.corrcoef(df_est_name_all['ratio_infr_insp'],df_est_name_all['pass_rate'])[0,1]
np.corrcoef(df_est_name_t100['ratio_infr_insp'],df_est_name_t100['pass_rate'])[0,1]
```

```
-0.48789485976239527

-0.9034773798376587
```

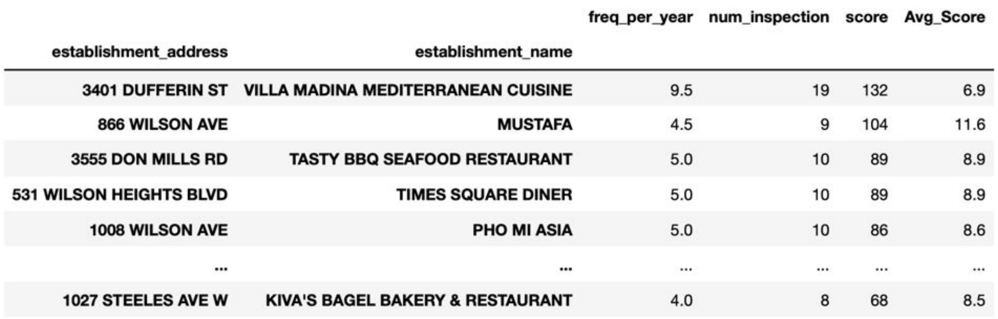
3.4 Establishment Safety Scoring

According to the Ministry of Health, three types of infractions can be ticketed to establishments. They are minor, significant and crucial. Minor infractions represent a minimal risk to the public, and they must be corrected in the next inspection. Significant infractions can lead to safety hazards and have to be corrected within 24-48 hours after the initial inspection. Crucial infractions are dangerous as they present immediate health hazards, and have to be corrected on the spot or an “order to close” of the premises will be issued.

Creating the safety score attribute

In order to promote a better food safety, assess the performance of establishments and analyze the consistency of safe food preparation, the team decided to invent a food safety scoring system. The first thing we did is to assign a numerical score to each infraction based on the category of severity. A custom lambda function was implemented to perform this operation. For example, the function will check the string in the ‘severity’ column of each row, and if the entry was ‘M-Minor’, it will assign a score of ‘1’ the infraction record (row) ; see Appendix **[A-3.5.1]**. For other infraction categories, we assigned scores of ‘2’ to ‘S – Significant’ and ‘3’ to ‘C – Crucial’. Therefore, the new scoring attributes can help the Ministry of Health analyze the performance of each establishment in the DineSafe system.

Table 2 - Establishment Safety Score Result

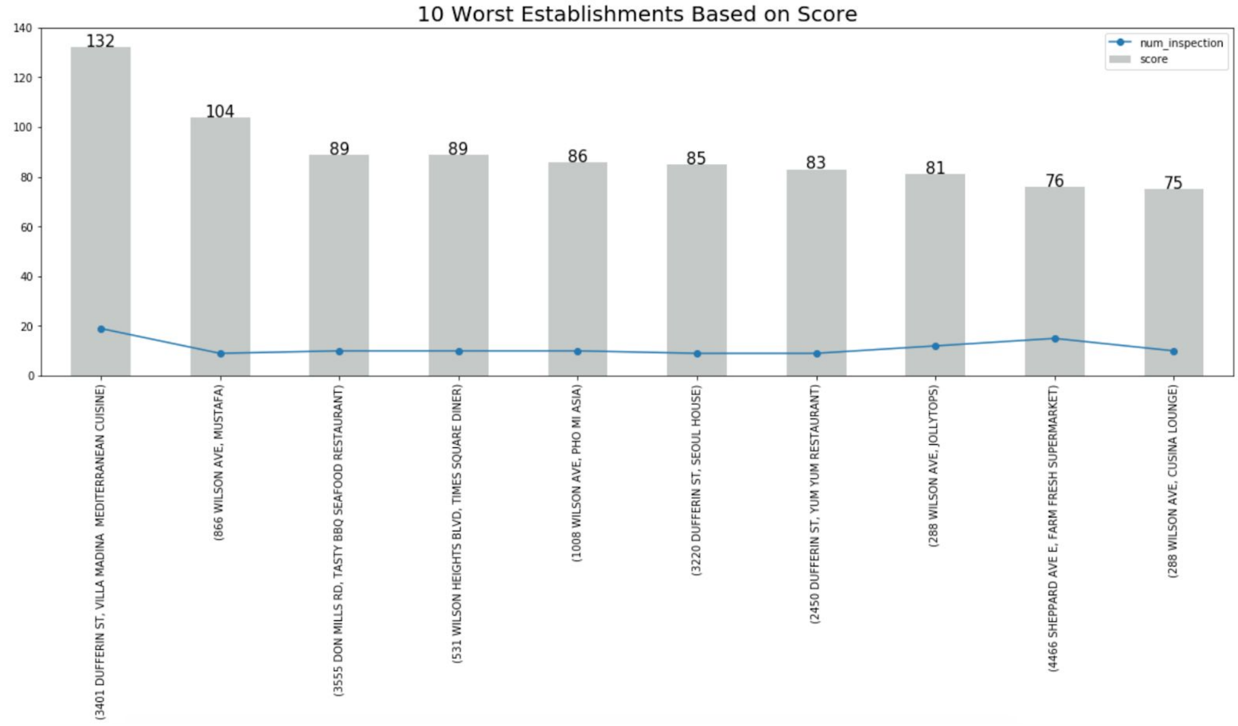


		freq_per_year	num_inspection	score	Avg_Score
establishment_address	**establishment_name**				
3401 DUFFERIN ST	**VILLA MADINA MEDITERRANEAN CUISINE**	9.5	19	132	6.9
866 WILSON AVE	**MUSTAFA**	4.5	9	104	11.6
3555 DON MILLS RD	**TASTY BBQ SEAFOOD RESTAURANT**	5.0	10	89	8.9
531 WILSON HEIGHTS BLVD	**TIMES SQUARE DINER**	5.0	10	89	8.9
1008 WILSON AVE	**PHO MI ASIA**	5.0	10	86	8.6
...	**...**	...	...	...	...
1027 STEELES AVE W	**KIVA'S BAGEL BAKERY & RESTAURANT**	4.0	8	68	8.5

What are the least safe food preparation establishments?

In order to retrieve the top 10 establishments with highest infraction scores (lower is better), we needed to first create a new score data frame. The team used Groupby function (see Appendix **[A-3.5.2]**) to derive two sub-data frames: the first one is the Total Score per Establishment, and the second is the Total Number of Inspection per Establishment. After that, we used inner join to merge the two data frames into the infraction score data frame. Finally, the data frame was sorted by the total number of infractions score.

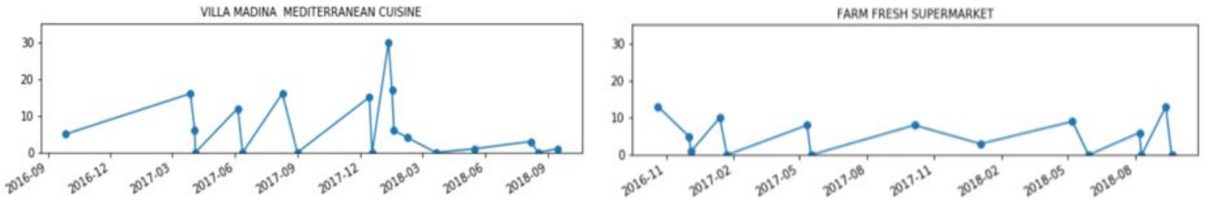
Figure 3.4.1 - 10 Worst Establishments Based on Score



According to the previous bar plot, we can see that the top 10 establishments with the highest infraction scores can range from 75 to 132 points in the 2-year inspection window. In general, a high infraction score is caused by either high frequency of inspection or a severe inspection result. It will be interesting to investigate what is the cause of high infraction scores for the listed establishment. Therefore, we created individual subplots to analyze the time series of infraction scores for those top 10 establishments, as shown in Appendix **[A-3.5.3]**.

Here, we chose two of the ten subplots to do a visual comparison of the time series plots. The first one is Villa Madina, the establishment with the highest infraction score. The second is Farm Fresh Supermarket.

Figure 3.4.2 - Villa Madina & Farm Fresh Time Series of Safety Score

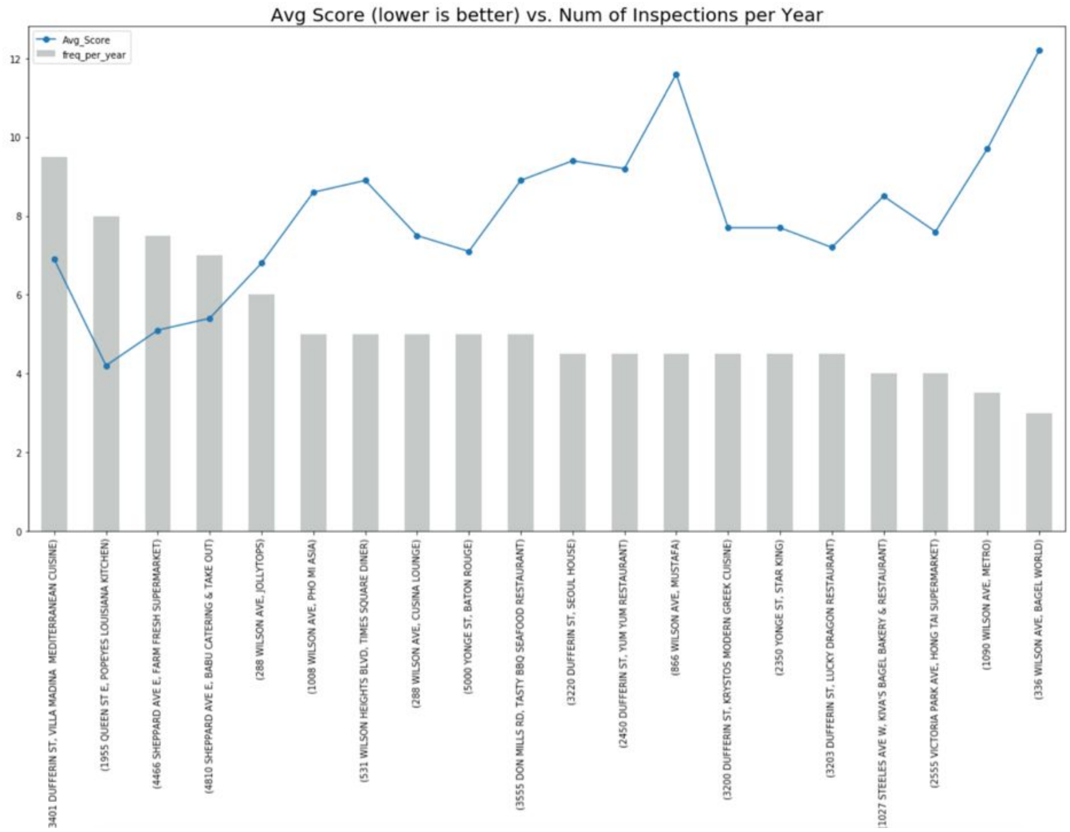


As we can tell by the establishment name, Villa Madina is a food court vendor and Farm Fresh is a supermarket, but they both belong to the high-risk type of establishment which requires at least 3 inspections per year. In addition, they both have been inspected around 20 times in the past two years. As we can see from the time series plots, Villa Madina has a much higher volatility than Farm Fresh, and because of the higher infraction score per inspection, Villa Madina has invited a more frequent second inspection following the initial assessment. However, its inspection frequency and score have significantly reduced after March 2018 due to the better inspection result (lower infraction score). On the other hand, Farm Fresh's infraction score is not perfect, but the volatility is low. The relatively lower and more consistent infraction score can help the establishment to avoid second inspections within 48 hours, and result in lower overall infraction scores.

Does the number of inspections impact the safety score of the establishment?

The improvement of inspection results for Villa Madina motivated the team to create a new KPI, which is the Average Score. Average Score can alleviate the inflation of infraction scores due to the high number of follow-up inspections, and it is calculated by dividing the total score by total number of inspections (see Appendix **[A-3.5.4]**). In addition, we were also interested in analyzing the correlation between Avg. Score and Number of Inspections per Year, as our intuition tells us that Villa Madina's good food safety practice in the recent 6 months might be motivated by the high frequency of inspections it has experienced. Here, we plot the Avg. Score vs. Number of Inspections per Year graph, which includes all avg. scores for top 20 most inspected establishments in the past two years (see Appendix **[A-3.5.5]).**

Figure 3.4.3 - Establishment Safety Score vs. Number of Inspections



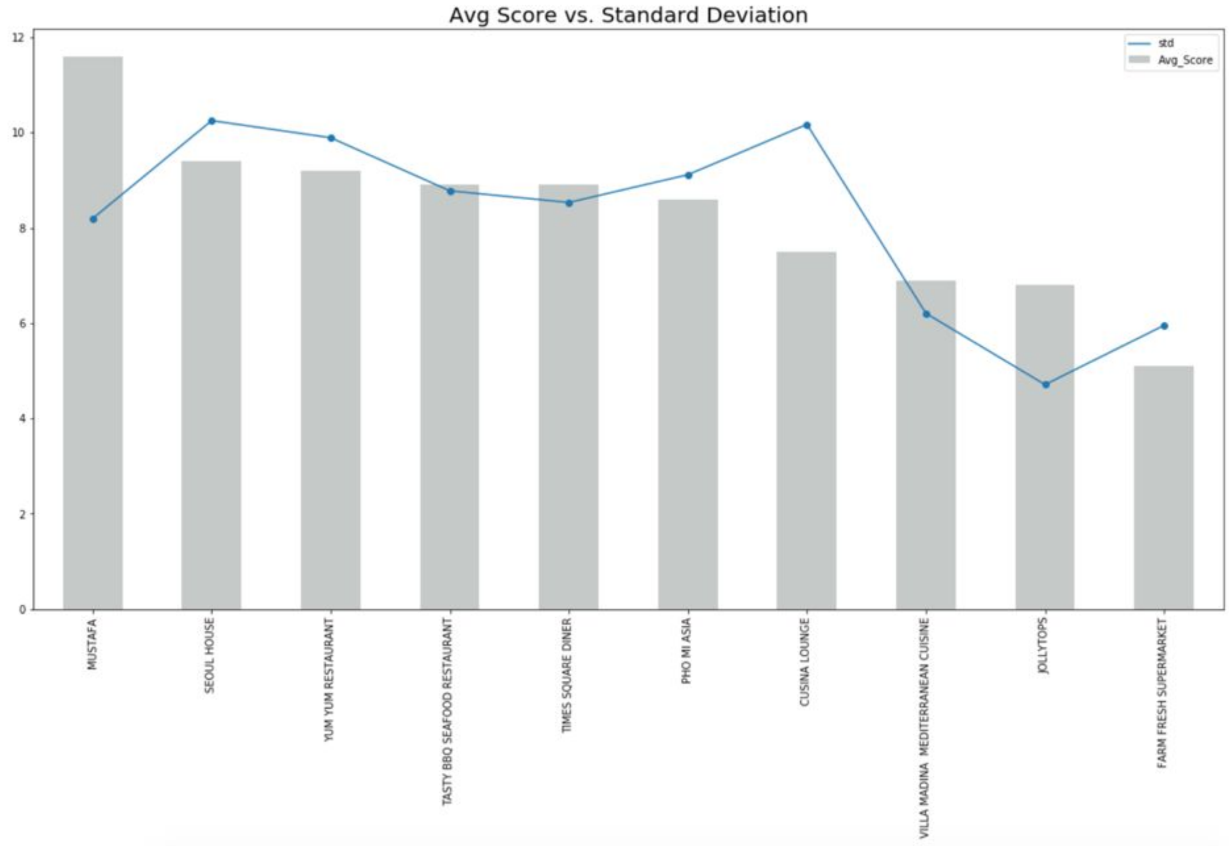
As we can see, Villa Madina has ranked No.1 in term of number of inspections per year, but its average infraction score is the 3rd lowest among top 20 most inspected places. Therefore, we believe the Avg. Score and Num of Inspections have a strong negative correlation, as we can see the lower the frequency of inspection, the higher the infraction score for establishments. In other words, the food safety inspections are making positive impacts to improve the food safety, and more frequent inspections are likely to generate better food safety results.

Statistical Analysis of the Establishment Safety Score

As previously discussed in Section 5.2, we have compared the time series of scores between Villa Madina and Farm Fresh. Farm Fresh has an overall infraction score due to the fact that it has better food safety control (less volatility) than Villa Madina in the months before Mar. 2018. This finding has inspired the team to perform a comparison between Avg. Score and Standard Deviation of Inspection Score.

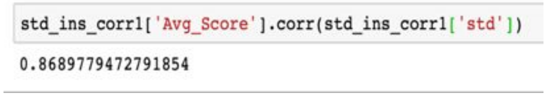
In order to calculate the Standard Deviation (std) for the top most frequently inspected establishments, we embedded two list objects 'std_list=[ ]' and 'str_list=[ ]' in the for loop that was used to create the infraction score time series subplots. Later on, we constructed a new 'std' dataframe that consists of the name of the establishment and the 'std' of each establishment (see Appendix **[A-3.5.6]**).

Figure 3.4.4 - Establishment Safety Score Standard Deviation



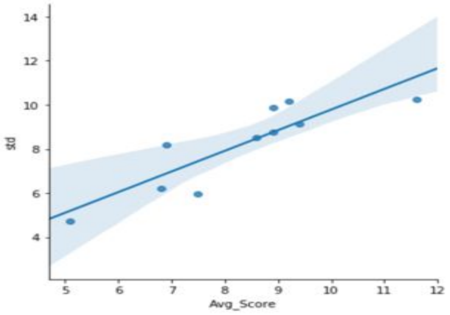
As we can observe from the Avg. Score vs. Standard Deviation graph, the Avg. Infraction Score is likely to decrease as the Standard Deviation (or Volatility) decreases. To help us to confirm this finding, we calculated the correlation between ‘std’ and ‘Avg_Score’ and made a regression line to fit the scatter plot.

Figure 3.4.5 - Correlation of Standard Deviation and Average Score



```
std_ins_corr1['Avg_Score'].corr(std_ins_corr1['std'])
```

```
0.8689779472791854
```



As we can see, the correlation score is 0.87, which indicates a strong positive correlation between ‘std’ and ‘Avg. Score’. In addition, the regression line shows a good fit for the points in the scatter plot. As a result, the more consistent the food safety infraction score (lower std), the lower avg. infraction score for establishments.

4. Conclusion & Recommendation

For this term project, our group analyzed the dataset containing records of the DineSafe inspection program. The program focuses on the inspection of all establishments serving and preparing food in the city of Toronto. There are three results that may occur after each inspection of an establishment (a pass, a conditional pass or a closed notice). Each establishment is inspected at a frequency established by the Ministry of Health and according to three identified risk levels (High risk, Moderate risk and Low risk). Every establishment receives a minimum of 1 to 3 inspections each year depending on the specific type of establishment.

The dataset was downloaded from the City of Toronto open data catalogue and had around 90,520 records, and over 16,291 establishments which are all located in the city of Toronto. The dataset had records of 25 months (September 2016 to September 2018). In the dataset, each row represented a registered record of one establishment, one inspection result and one infraction detail. In other words, each establishment could have multiple infraction details for a particular date, and each infraction would be listed as a separate record in the data set.

Overall, the DineSafe program has not been effective at increasing the passing rate of establishments in Toronto. The pass rate of establishments has been decreasing over the 25 months analyzed. It is clear that the high risk establishments are the category that is driving the decrease in pass rate. It has also been shown that the number of inspections has an impact on the Establishment Safety Score. The more frequently an establishment is inspected, the more likely that establishment is to receive a pass on any given inspection.

We recommend that Toronto Public Health increase the inspection frequency of high risk establishments in order to achieve their desired outcome of an increased inspection pass rate.

5. APPENDICES

A-2 Sample Codes for Section 2

[A-2.1] load the data into a pandas DataFrame .

```
#read dataset into python
df = pd.read_csv("dinesafe.csv”)
```

[A-2.2] Data dictionary.

- **ROW_ID** - Represents the Row Number
- **ESTABLISHMENT_ID** – Unique identifier for an establishment
- **INSPECTION_ID** - Unique identifier for each Inspection
- **ESTABLISHMENT_NAME** – Business name of the establishment
- **ESTABLISHMENTTYPE** – Establishment type ie restaurant, mobile cart
- **ESTABLISHMENT_ADDRESS** – Municipal address of the establishment
- **LONG/LAT**– Longitude & Latitude coordinates of an establishment
- **ESTABLISHMENT_STATUS** – Pass, Conditional Pass, Closed
- **MINIMUM_INSPECTIONS_PERYEAR** – Every eating and drinking establishment in the City of Toronto receives a minimum of 1, 2, or 3 inspections each year depending on the specific type of establishment, the food preparation processes, volume and type of food served and other related criteria
- **INFRACTION_DETAILS** – Description of the Infraction
- **INSPECTION_DATE** – Calendar date the inspection was conducted
- **SEVERITY – Level** of the infraction, i.e. S – Significant, M – Minor, C – Crucial
- **ACTION** – Enforcement activity based on the infractions noted during a food safety inspection
- **COURT_OUTCOME** – The registered court decision resulting from the issuance of a ticket or summons for outstanding infractions to the Health Protection and Promotion Act
- **AMOUNT_FINED** – Fine determined in a court outcome

[A-2.3] Adding data dimension to the dataset.

```
#adding date dimensions
df['INSPECTION_DATE'] = pd.to_datetime(df['INSPECTION_DATE'])
df['Year'] = pd.DatetimeIndex(df['INSPECTION_DATE']).year
df['Quarter'] = pd.DatetimeIndex(df['INSPECTION_DATE']).quarter
df['Month'] = pd.DatetimeIndex(df['INSPECTION_DATE']).month
df['Week'] = pd.DatetimeIndex(df['INSPECTION_DATE']).week
df['Year_Quarter'] = df['Year'].astype(str) +'-'+ df['Quarter'].astype(str)
df['Year_Month'] = df['Year'].astype(str) +'-'+ df['Month'].astype(str)
df['Year_Week'] = df['Year'].astype(str) +'-'+ df['Week'].astype(str)
df['Week_Day'] = pd.DatetimeIndex(df['INSPECTION_DATE']).weekday #The day of the week with
Monday=0, Sunday=6
```

[A-2.4] Functions.

```
def count_inspection(dim, val, interval):
   cnt_insp = df[df[dim]==val].groupby([interval]).INSPECTION_ID.nunique()
```

```
    return cnt_insp

def count_infraction(dim,val,interval):
    cnt_infr = df[df[dim]==val].groupby([interval]).INSPECTION_ID.size()
    return cnt_infr
def ratio_infr_insp(dim,val,interval):
    cnt_infr = df[df[dim]==val].groupby([interval]).INSPECTION_ID.size()
    cnt_insp = df[df[dim]==val].groupby([interval]).INSPECTION_ID.nunique()
    ratio_infr_insp = cnt_infr/cnt_insp
    return ratio_infr_insp
def pass_rate_cal(dim, val, interval):
    df_pass = df[df['ESTABLISHMENT_STATUS']=='Pass']
    cnt_pass = df_pass[df_pass[dim]==val].groupby([interval]).INSPECTION_ID.nunique()
    cnt_insp = df[df[dim]==val].groupby([interval]).INSPECTION_ID.nunique()
    pass_rate = pd.Series(cnt_pass/cnt_insp*100)
    return pass_rate
```

[A-2.5] Getting top 15 Establishment Type

```
df_est_type = df.ESTABLISHMENTTYPE.value_counts()
t15 = df_est_type.head(15)
```

A-3 Sample Codes for Section 2

[A-3.5.1] Converting severity level to integer values.

```
def label_severity(row):
 if row['severity'] == 'M - Minor':
   return 1
 if row['severity']== 'S - Significant':
   return 2
 if row['severity']== 'C - Crucial':
   return 3
 return 0
 data['score']=data.apply(lambda row: label_severity(row),axis=1)
```

[A-3.5.2] Getting the total score per establishment.

```
sum_score = data.groupby(['establishment_address', 'establishment_name'])['score'].sum()
total_score = sum_score.to_frame(name = 'score')
total_score=total_score.sort_values(['score'],ascending = False)
# get the total number of inspection per establishment
total_inspection = data.groupby(['establishment_address',
'establishment_name'])['inspection_id'].nunique()
total_ins = total_inspection.to_frame(name = 'num_inspection')
total_ins=total_ins.sort_values(['num_inspection'],ascending = False)
 # join total score table with the total inspection table
temp = pd.merge(total_ins, total_score, on=('establishment_address','establishment_name'),
how='inner')
# create  a avg_score column
temp['Avg_Score'] = temp['score']/temp['num_inspection']
avgScore = temp.sort_values(['Avg_Score'],ascending = False)
```

```
# round avg score to 1 decimal
avgScore=avgScore.round(1)
# reset index for both freq_order and avgScore table
freq_order.reset_index()
avgScore.reset_index()
 # Join freq_order and avgScore table to get the scoring system table
ScoreData = pd.merge(freq_order, avgScore, on=('establishment_address','establishment_name'),
how='inner')
ScoreData=ScoreData.sort_values(['score'],ascending = False)
ScoreData
```

[A-3.5.3] Getting the total score per establishment.

```
loc_index = ScoreData.index.values[:10]
len(loc_index)
num=1
#var_data=pd.DataFrame(columns=['establishment_name','variance'])
std_list=[]
str_list=[]
import seaborn as sns
for i in range(len(loc_index)):
    plt.subplot(5,2, num)
    num+=1
    #if num in range(14) :
    #   plt.tick_params(labelbottom=False)
    #if num not in [1,4,7] :
     #  plt.tick_params(labelleft=True)
    plt.subplots_adjust(hspace=1)
    #plt.tick_params(labelbottom=True)
    Villa_Trend = data[(data['establishment_address']==loc_index[i][0]) &
(data['establishment_name']==loc_index[i][1])].sort_values(['inspection_date'],ascending = True)
   Villa_Trend=Villa_Trend.groupby('inspection_date').sum()
    x=np.std(Villa_Trend['score'])
    #list1=[loc_index[i][1],x]
    std_list.append(x)
    str_list.append(loc_index[i][1])
   ax_v=Villa_Trend['score'].plot(linestyle='-', marker='o',figsize=(20,20))
    ax_v.set_ylim(0,35)
    plt.title(loc_index[i][1], fontsize = 10)
ax_v.xaxis.label.set_visible(False)
```

[A-3.5.4] Getting the total score per establishment.

```
# create  a avg_score column
temp['Avg_Score'] = temp['score']/temp['num_inspection']
```

[A-3.5.5] Getting the total score per establishment.

```
sort_score_data = top20_score_data.sort_values(['num_inspection'],ascending = False)
ax = sort_score_data[['Avg_Score']].plot(linestyle='-', marker='o',figsize=(20,10))
sort_score_data[['freq_per_year']].plot(kind='bar', ax=ax,color='xkcd:silver')
ax.set_title("Avg Score (lower is better) vs. Num of Inspections per Year", fontsize=20)
plt.legend()
```

[A-3.5.6] Getting the total score per establishment.

```
std_df = pd.DataFrame({'establishment_name':str_list, 'std':std_list})
std_ins_corr = pd.merge(ScoreData, std_df, on=('establishment_name'), how='inner')
std_ins_corr1=std_ins_corr.sort_values(['Avg_Score'],ascending = False)
std_ins_corr1
ax_std1 = std_ins_corr[['std']].plot(linestyle='-', marker='o',figsize=(20,10))
std_ins_corr1[['Avg_Score']].plot(kind='bar',ax = ax_std1, color='xkcd:silver')
ax_std1.set_title("Avg Score vs. Standard Deviation", fontsize=20)
ax_std1.set_xticklabels(std_ins_corr1.establishment_name)
```