Data Analytics Project Report

Time Series Analysis on Restaurant Inspection Data for forecasting Violation Counts

Shivani Vogiral

Abstract
Restaurants and markets in Los Angeles are regularly inspected for health code violations. The data collected by the county is publicly available and accessible. The goal of this project is to perform Time Series analysis on the data and forecast violation counts.

Introduction
Food safety and hygiene is a global health concern. Restaurant inspections are often conducted to make sure food products are handled and according to the state and local regulations to protect the public. The *Environment Health Agency* operating as a part of *LA County Public Health Department* regularly inspects restaurants for health code violations. By making the data collected publicly available and accessible, the agency enables a transparency.

Dataset
The dataset "*LA County Restaurant Inspections and Violations"* has been downloaded from Kaggle. It covers information in two .csv files :

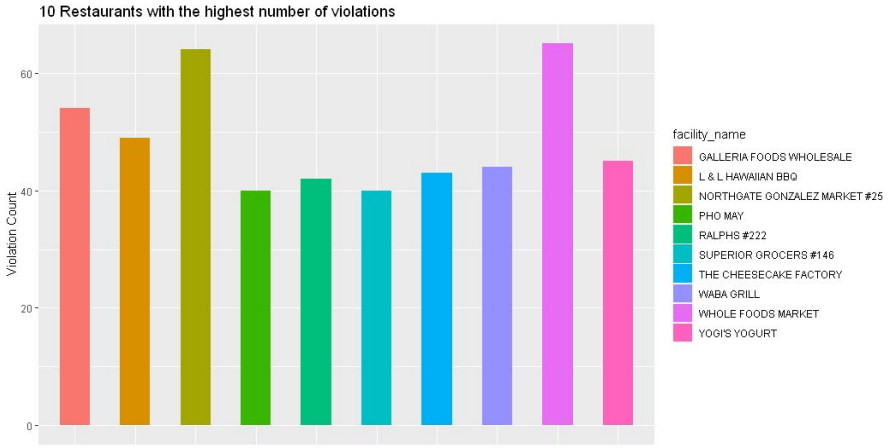
1. violations.csv: This dataset contains Environmental Health Violations for Restaurants and Markets in Los Angeles County. Attributes in the dataset: *points, serial_number, violations_code, violation_status*
2. Inspections.csv: This dataset contains results of the Restaurant Inspections by the Los Angeles County Environmental Health Agency from the year 2015-2017. Each row represents an inspection. Attributes in the dataset:*activity_date, employee_id, facility_address, facility_city, facility_id, facility_name, facility_state, facility_zip, grade, owner_id, owner_name, pe_description, program_element_pe, program_name, program_status, record_id, score, serial_number, service_code, service_description*

For the project two datasets have been merged by the attribute "serial_number" and only inspection data of 15 cities in Los Angeles from the year 2016-2017 has been considered for analysis and the cities included are: Hawaiian Gardens, Avalon, Wilmington, Marina Del Rey, Hermosa Beach, Sherman Oaks, Hawthorne, Duarte, Industry, La Verne, Maywood, Vernon, Eagle Rock, Lake Balboa, Arcadia.

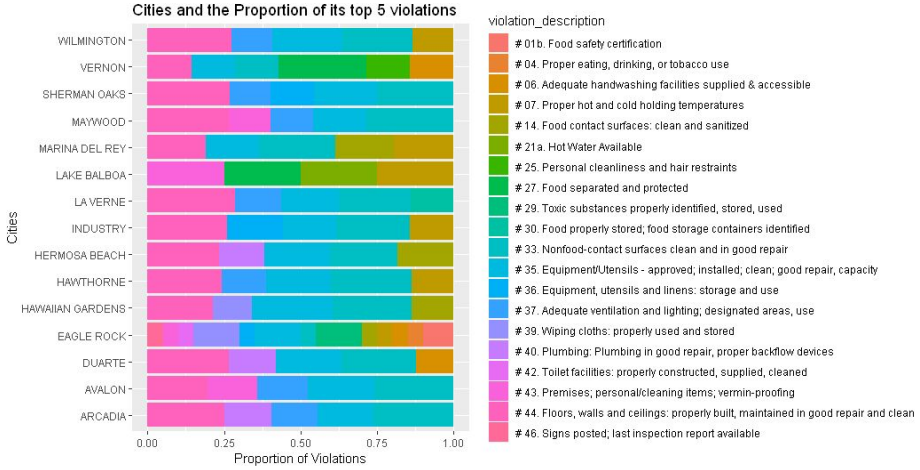
Exploratory Analysis

Top 10 Restaurant/Market Health Code Violations

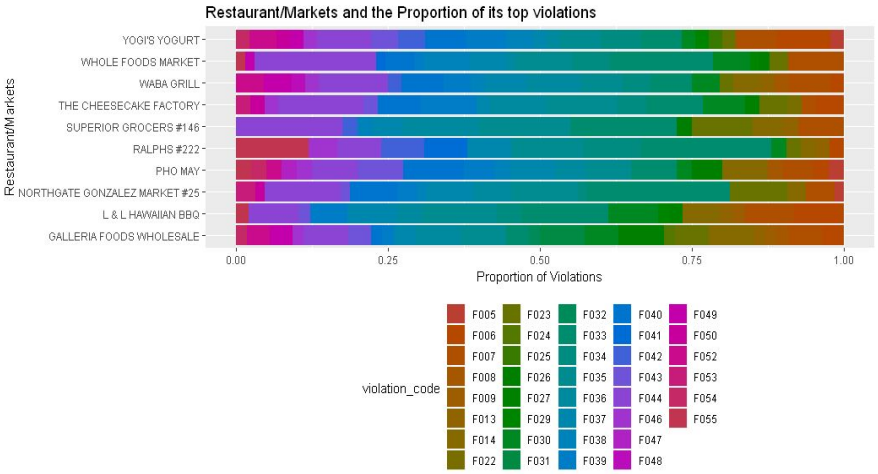
Restaurants with the highest number of violations



Cities in LA and the proportion of the commonly observed violations in Restaurants and Markets



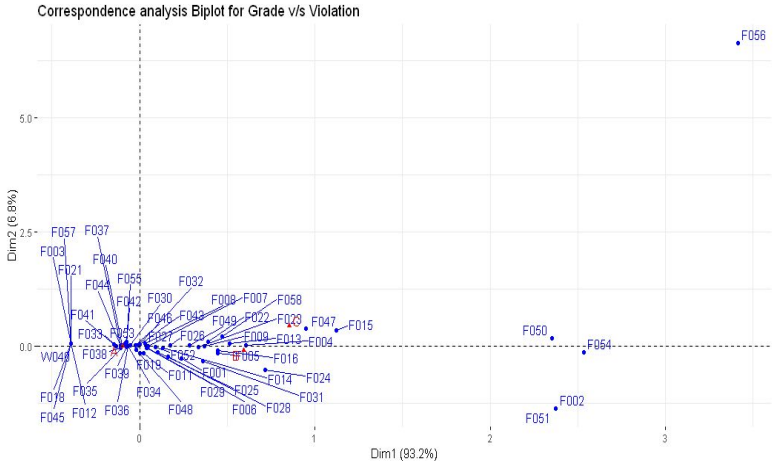
Restaurant/Market wise violation analysis



Correspondence Analysis

A technique to analyse the measure of correspondence between the rows and columns of a contingency table. The following analysis is used to understand the existence of a relationship between restaurants with a certain grade and the violations.

	A	B	C
F001	44	13	0
F002	0	1	0
F003	2	0	0
F004	35	15	3
F005	80	30	3
F006	553	88	11
F007	843	130	20
F008	96	13	3
F009	128	38	7
F011	17	3	0
F012	1	0	0



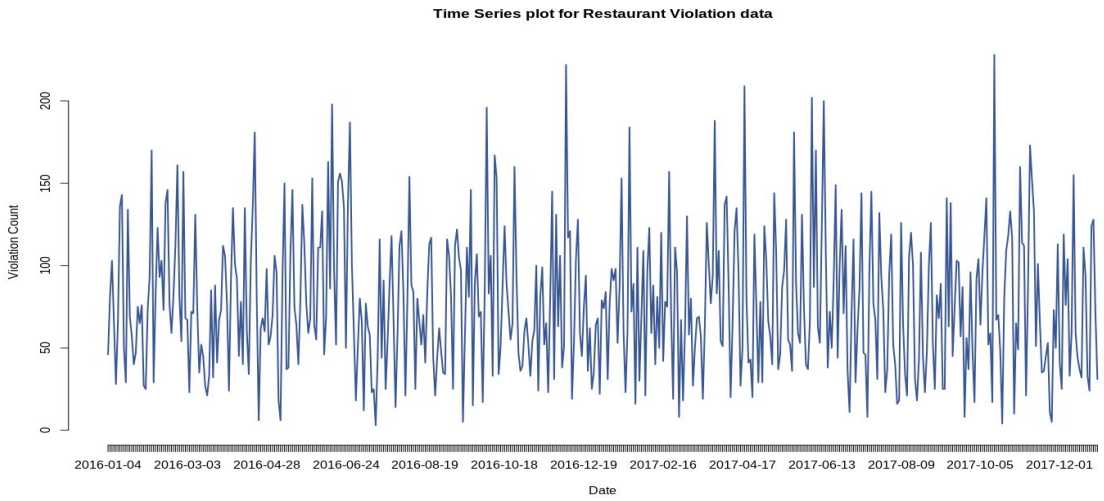
Time Series Analysis and Modelling

The following is the time series data obtained from the dataset. It is a frequency table which has the date and the number of violations found on that day.

2017-06-01	2017-06-02	2017-06-05	2017-06-06	2017-06-07	2017-06-08	2017-06-09	2017-06-12
40	37	67	202	87	170	63	53
2017-06-13	2017-06-14	2017-06-15	2017-06-16	2017-06-19	2017-06-20	2017-06-21	2017-06-22
123	200	109	38	72	50	91	149
2017-06-23	2017-06-26	2017-06-27	2017-06-28	2017-06-29	2017-06-30	2017-07-03	2017-07-05
44	100	134	71	112	36	11	71
2017-07-06	2017-07-07	2017-07-10	2017-07-11	2017-07-12	2017-07-13	2017-07-14	2017-07-17
116	29	61	86	144	47	46	8
2017-07-18	2017-07-19	2017-07-20	2017-07-21	2017-07-24	2017-07-25	2017-07-26	2017-07-27
97	145	77	67	31	132	93	73
2017-07-28	2017-07-31	2017-08-01	2017-08-02	2017-08-03	2017-08-04	2017-08-07	2017-08-08
23	37	95	119	52	40	16	18
2017-08-09	2017-08-10	2017-08-11	2017-08-14	2017-08-15	2017-08-16	2017-08-17	2017-08-18
126	64	34	21	105	120	100	31
2017-08-21	2017-08-22	2017-08-23	2017-08-24	2017-08-25	2017-08-28	2017-08-29	2017-08-30
18	43	108	45	23	48	100	126
2017-08-31	2017-09-01	2017-09-05	2017-09-06	2017-09-07	2017-09-08	2017-09-11	2017-09-12
53	25	82	68	89	25	25	141

Plot of the time series data

From the plot it is clearly evident that the time series data is random. The peaks and troughs which occur at irregular intervals. There is no trend, seasonality nor cyclicality in the data.



Test for stationarity

In order to determine whether the time series data is stationary or non-stationary, the Augmented Dickey Fuller Test is performed. It is a t-statistic based test. Smaller p-values suggest that the series is stationary.

Hypothesis:

Null Hypothesis(H0): The time series is not stationary

Alternate Hypothesis(H1): The time series is stationary

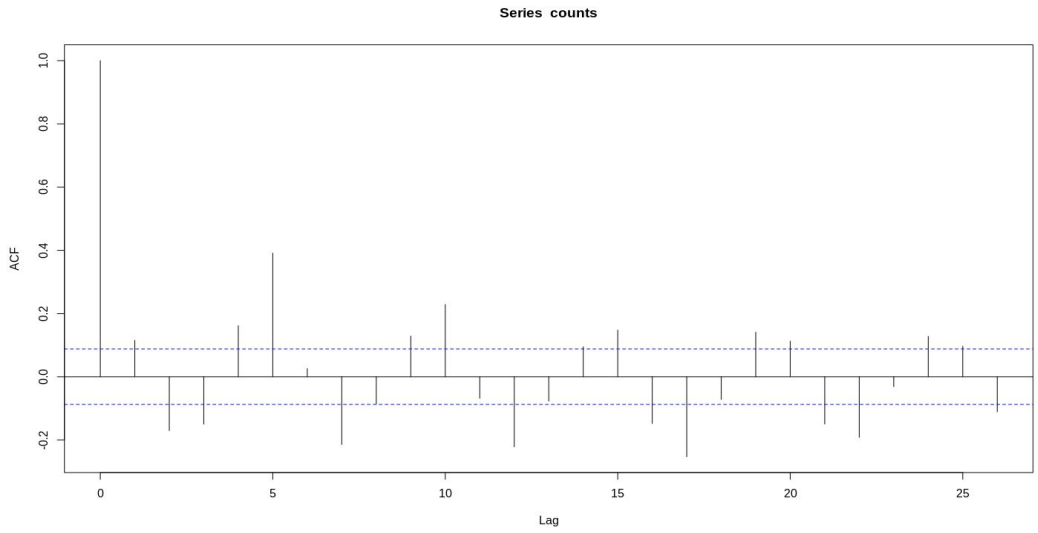
```
        Augmented Dickey-Fuller Test

data:  counts
Dickey-Fuller = -6.7174, Lag
order = 7, p-value = 0.01
alternative hypothesis: stationary
```

The p-value is very small and the null hypothesis is rejected. This proves our alternate hypothesis that the series is stationary and hence we reject the null hypothesis

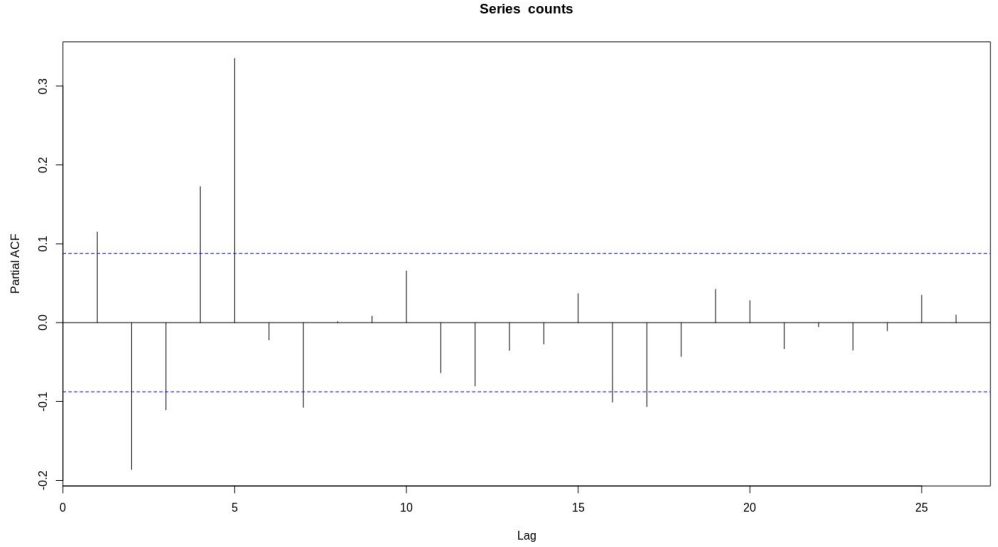
Auto-correlation Function and Partial Auto-Correlation Function Plots

From the plots, both plots display significant values. So an ARMA time series model will serve the needs. The ACF can be used to estimate MA-part, and the PACF can be used to estimate the AR-part

```
Autocorrelations of series ‘counts’, by lag

     0      1      2      3      4      5      6      7      8      9     10     11     12
 1.000  0.115 -0.171 -0.150  0.161  0.391  0.026 -0.214 -0.086  0.129  0.229 -0.068 -0.222
    13     14     15     16     17     18     19     20     21     22     23     24     25
-0.077  0.095  0.148 -0.148 -0.253 -0.072  0.141  0.113 -0.150 -0.192 -0.031  0.128  0.096
    26
-0.111
```




```
Partial autocorrelations of series ‘counts’, by lag

     1      2      3      4      5      6      7      8      9     10     11     12     13
 0.115 -0.186 -0.111  0.173  0.335 -0.022 -0.107  0.001  0.008  0.066 -0.064 -0.080 -0.035
    14     15     16     17     18     19     20     21     22     23     24     25     26
-0.027  0.037 -0.101 -0.107 -0.043  0.042  0.028 -0.033 -0.005 -0.035 -0.010  0.035  0.010
```



Building the ARIMA model

In order to obtain the best ARIMA model, the auto.arima() function in R has been used. This function conducts a search over possible model within the order constraints provided. The model which is suggested by the auto.arima() function is ARMA(4,3). These models will help us attempt to capture more of the serial correlation present within an instrument. Ultimately they will provide a means of forecasting the future violation counts.

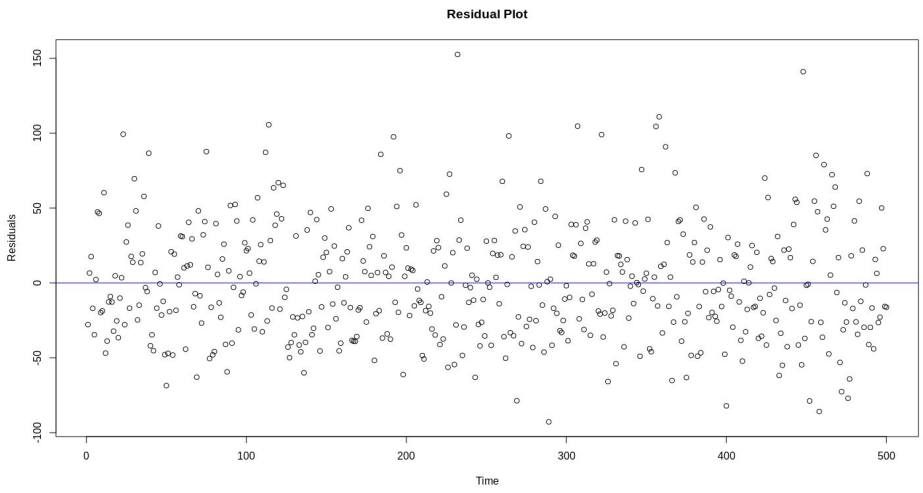
```
#forecasting violations for 20 days
arma_forecast <-forecast(arma,h=20)
plot(arma_forecast)
plot.ts(arma_forecast$residuals,type = "p")
```

The number of periods being forecasted is 20 days. Below is a plot of the forecasted violation counts which has a shaded region specifying the 80% and 95% confidence intervals.

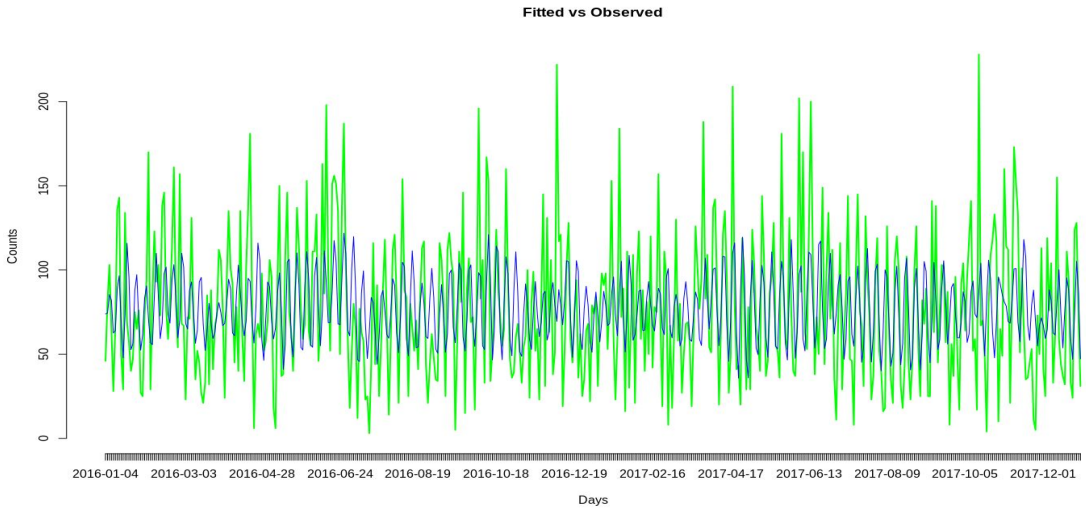
Summary of the chosen ARMA model

```
> summary(arma)
Series: counts
ARIMA(4,0,3) with non-zero mean

Coefficients:
         ar1      ar2     ar3     ar4      ma1     ma2      ma3     mean
      1.1368  -1.1224  0.5058  0.1312  -1.0682  1.0187  -0.4534  77.2925
s.e.  0.1600   0.1713  0.1808  0.0726   0.1578  0.1410   0.1217   2.4230

sigma^2 estimated as 1478:  log likelihood=-2530.56
AIC=5079.11   AICc=5079.48   BIC=5117.04

Training set error measures:
                     ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 0.04856021 38.13027 30.49089 -52.34147 76.84429 0.6630476 -0.001566367
```

Residual Plot

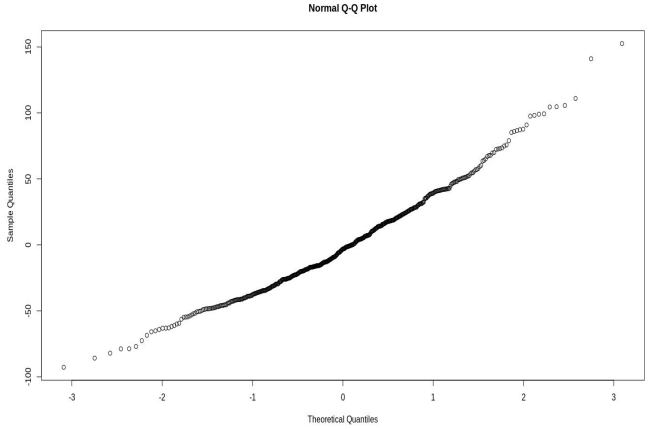


Plot of Observed Values vs Fitted Values



Normal Q-Q plot of the residuals

From the plot we can observe that the residuals are normally distributed

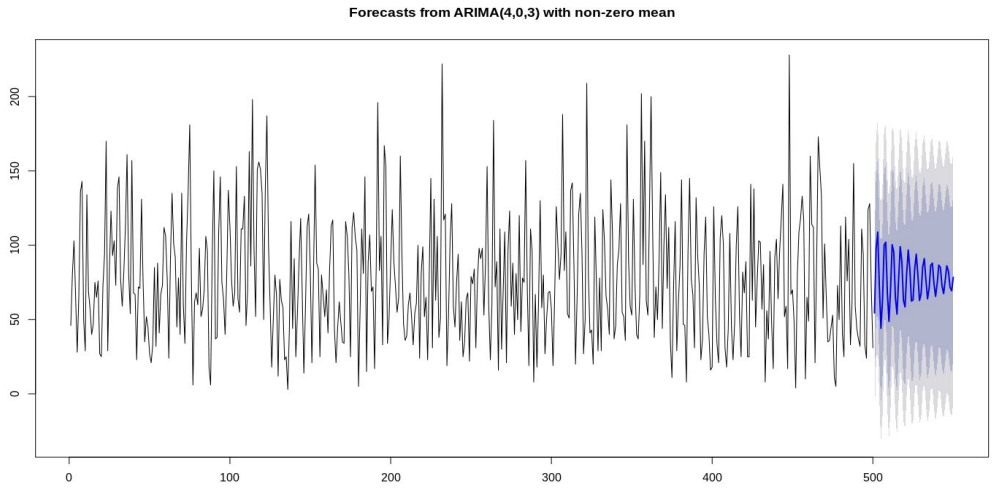


Forecasting Violation Counts

The series has violation counts of 500 days. With the help of forecast() function in R package, forecast, the violation counts for 50 days starting from the 501st day to the 550th day has been forecasted.

Forecasted Violation Counts

```
Time Series:
Start = 501
End = 550
Frequency = 1
 [1]  54  97 108  73  44  63 100 102  68  48  70 100  95  64  53  76  99  89  62  58  80  96  84  62  63  83  94  79  62  67  85  90  76  63  71  86  87  74  65  74  86  85  72  67  76  86  82
[48]  71  69  78
```



Conclusion

The forecasted violation counts may not represent the actual situation. But, it may be helpful for the Health Inspectors to evaluate restaurant code violations efficiently. It is because the time series data is a random series, there is no seasonality, no trend, or cyclicality in the data. The data which has been analyzed includes the violations found in Food Markets and Restaurants. Many factors affect the number of violations found on a particular day, such as, the areas where the facility is located, the people employed to work at the restaurant, the day the inspection took place and how strict the Health Inspector is in evaluating the health code violations.