San Francisco Restuarants' Inspection Data Analysis

Resham Vyas, Sindhuja Ambadasugari & Shruthi Bhat

24 August 2017

Introduction

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: 1.high risk category 2.moderate risk category 3.low risk category

Violations that are directly related to the transmission of food borne illnesses, the adulteration of food products and the contamination of food-contact surfaces fall into high risk category

Records specific violations that are of a moderate risk to the public health and safety fall into moderate risk category.

Record violations that are low risk or have no immediate risk to the public health and safety fall into low risk category.

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.

Field names and it's description:

Field Name	Data Type	Description
business_id	string	Unique identifier for the business. For many cities, this may be the license number.
business_name	string	Common name of the business.
business_address	string	Street address of the business. Example: 706 Mission St.
business_city	string	City of the business. This field must be included if the file contains businesses from multiple cities.
business_state	string	State or province for the business. In the U.S. this should be the two-letter code for the state.
business_postal_code	string	Zip code or other postal code.
business_latitude	number	Latitude of the business. This field must be a valid WGS 84 latitude. Example: 37.7859547
business_longitude	number	Longitude of the business. This field must be a valid WGS 84 longitude. Example: -122.4024658
business_location	Location	For geospatial API capabilities or for geocode addresses this Location datatype column is needed. Examples: • (37.7859547, -122.4024658) • 600 Fourth Ave, Seattle, WA 98104
business_phone_number	string	Phone number for a business including country specific dialing information. Example: +14159083801
inspection_id	string	A unique identifier for a given inspection
inspection_date	date	Date of the inspection in YYYY-MM-DD format. Example: 2015-08-22
inspection_score	number	Calculated inspection score, may be either graded (0-5, 0-100), or cumulative, and this should be defined in your feed metadata.
inspection_result	string	For jurisdictions that do not capture a score, this string represents the non-numeric result of the inspection, for example "Pass" or "Fail". The original LIVES standard requires this field to contain 4 characters or fewer. For broader use of LIVES data, we suggest shortened terms for this field. * If inspections are unscored, this value must be provided.
inspection_description	string	Single line description containing details on the outcome of an inspection.
inspection_type	string	String representing the type of inspection. Must be one of: initial, routine, follow-up, complaint
violation_id	string	A unique identifier for a given violation

violation_description	string	One line description of the violation. * If violation data is provided then this field is required
violation_code	string	Code for the violation. It is recommended that this be based on the FDA Food Code. However, municipalities can decide to use pre-existing codes for this field.
violation_critical	boolean	Describes whether the violation is critical (i.e., if it would cause the restaurant to fail their inspection) Must be one of: true, false

Motivation:

Eating out is a way of life in the busy city of San Fransisco. Be it hanging out for fun or saving time from cooking at home , most people do eat at restaurants either once in a while or more frequently. We will analyze the available restaurants data to make certain data based findings

Data Source

URL: https://data.sfgov.org/Health-and-Social-Services/Restaurant-Scores-LIVES-Standard/pyih-qa8i/data

Individual restaurant performance over years

1.Find The restuarants that have been performing consistently well from 2014 to 2016 2.Find the restuarants that have been inconsistent but improving from 2014 to 2016 3.Find the restuarants that been been worsened from 2014 to 2016

```r
setwd("/Users/shruthi/RWorkspace/")

# Read the Data
raw_data = read.csv("Restaurant_Scores_-_LIVES_Standard.csv")

# Choose the columns needed for the analysis
YoY_data = raw_data[c(2,12,13)]

# Choose the dates column and chnage format
dates <- YoY_data$inspection_date
tmp <-  as.Date(dates,'%m/%d/%Y')
year = format(tmp,'%Y')

# Add the year column to the data frame
YoY_data$inspection_year = year

# Reorder the data columns and remove previous date column
YoY_data = YoY_data[c(1,4,3)]

# Remove the NAs from inspection scores column
YoY_data = YoY_data[complete.cases(YoY_data[,3]), ]

# Group the data by business and year
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

YoY_summary = group_by(YoY_data, business_name, inspection_year)
YoY_summary = summarize(YoY_summary, avg_score =
mean(inspection_score))

# Represent the data in the form of a table
library(reshape)

##
## Attaching package: 'reshape'

## The following object is masked from 'package:dplyr':
##
##     rename

YoY_table = cast(YoY_summary, business_name ~ inspection_year, value =
"avg_score")

# Remove the NAs from 2014-2016. 2017 has too many NAs. Removing them
would mean excluding lot of important information.
YoY_table = YoY_table[complete.cases(YoY_table[,2:4]), ]

# Considering only data 2014-2016
YoY_table = YoY_table[c(1:4)]

# Finding variance of the restaurant scores over time
library(matrixStats)

##
## Attaching package: 'matrixStats'

## The following object is masked from 'package:dplyr':
##
##     count

YoY_table$variance = rowVars(data.matrix(YoY_table[,2:4]))

# Choosing the rows/restaurants with 0 variance
YoY_consistent = YoY_table[YoY_table$variance == 0, ]
YoY_consistent
```

```
##                                           business_name 2014 2015
2016
## 85                                   Acme Bread Company  100  100
100
## 134                                 Alex Gourmet Burrito  100  100
100
## 228                 ARGONNE ALTERNATIVE ELEMENTARY SCHO  100  100
100
## 262              AT&T - (CART 14) BAYSIDE BREWS [145102]  100  100
100
## 279               AT&T - (CART 35) DOGGIE DINER [145122]  100  100
100
## 367                         AT&T PARK - ANCHOR PLAZA  100  100
100
## 397                 AT&T PARK - MARGARITA CART (P 10)  100  100
100
## 416                            Auntie Anne's Pretzels   97   97
97
## 479                                       Bar Tartine  100  100
100
## 493                                              BASK  100  100
100
## 546                                 Beluga Restaurant  100  100
100
## 572                                               BHL  100  100
100
## 573                                       BHUK Burger  100  100
100
## 658                                  Borderlands Cafe  100  100
100
## 805                                CAFE INTERNATIONAL   96   96
96
## 898                                        Cane Rosso  100  100
100
## 966                            Cerveceria De Mateveza  100  100
100
## 1025                                      Chile Lindo   96   96
96
## 1039          Chinese Education Center Elementary School  100  100
100
## 1040                                 Chinese Hospital  100  100
100
## 1118                        CLEVELAND ELEMENTARY SCHOOL  100  100
100
## 1129                    Coastside Farms and Specialties  100  100
100
## 1179                                 COSTCO WHOLESALE   96   96
96
## 1194                    Cowgirl Creamery/Artisan Cheese  100  100
100
```

```
## 1234                               Crystal Jade Jiang Nan  100  100
100
## 1400                             East & West Gourmet Food  100  100
100
## 1416                                         Eclipse Cafe  100  100
100
## 1504                           Escape from New York Pizza  100  100
100
## 1512                               Esther's German Bakery  100  100
100
## 1609                                   flourChylde Bakery  100  100
100
## 1638                        Four Star Theatre - Snack Bar   98   98
98
## 1651                      FRANK MCCOPPIN ELEMENTARY SCHOOL  100  100
100
## 1689                                   G. L. Alfieri, LLC  100  100
100
## 1798                                 Golden Gate Meat Co  100  100
100
## 1944                                         Happy Lounge  100  100
100
## 2003                           Hilton Financial District  100  100
100
## 2004       Hilton Financial District- Banquet & Catering  100  100
100
## 2005   Hilton Financial District- Restaurant Seven Fifty  100  100
100
## 2006 Hilton Financial District-Flute Coffee and Wine Bar  100  100
100
## 2032                             Home Maid Ravioli Co Inc  100  100
100
## 2039                                  Hon's Wun Tun House   98   98
98
## 2080                                  Hotwok Express, Inc   84   84
84
## 2105                                     Humphry Slocombe  100  100
100
## 2118   Hyatt Regency - Main Kitchen, Employees Cafeteria  100  100
100
## 2173              International Hotel Senior Housing, Inc  100  100
100
## 2283                   Joey & Pat's Bakery & Coffee Shop  100  100
100
## 2315                      JUNIPERO SERRA ELEMENTARY SCHOOL  100  100
100
## 2343                                 Kara's Cupcakes Inc.  100  100
100
## 2367                                KEY ELEMENTARY SCHOOL  100  100
100
```

```
## 2454                                    La Corneta   94   94
94
## 2507                                 Ladle and Leaf  100  100
100
## 2630                                 LITTLE HENRY'S   98   98
98
## 2786                     Mariposa Baking Co., Inc.  100  100
100
## 2808                  MARTIN L. KING MIDDLE SCHOOL  100  100
100
## 2884                                 Mi Lindo Peru'   90   90
90
## 2885                            Mi Pueblito Market   88   88
88
## 2896                                        Miette  100  100
100
## 2903                       Mikes Grocery & Liquor  100  100
100
## 2965                             Mojo Bicycle Cafe   94   94
94
## 3011                                Moyo's Yogurt  100  100
100
## 3041                                    MV Taurus  100  100
100
## 3228                     Oceanview Market and Deli  100  100
100
## 3242                            OMI Senior Center  100  100
100
## 3260                                 Onigilly LLC  100  100
100
## 3429                               Pepples Donuts  100  100
100
## 3443                              PETE'S UNOCAL 76  100  100
100
## 3543                       POPEYES-GENEVA & MISSION   92   92
92
## 3560                       Prather Ranch Meat Co.  100  100
100
## 3594                          PUERTO ALEGRE NO. 2   98   98
98
## 3608                     QUICK N EASY INDIAN FOODS  100  100
100
## 3610                     Quick-N-Ezee Indian Foods  100  100
100
## 3666                        Regency Club 18th Floor  100  100
100
## 3715                       Ritz-Carlton SF - Bakery   98   98
98
## 3718            Ritz-Carlton SF - Employee Cafeteria   96   96
96
```

```
## 3929                                      Schilling & Co.  100  100
100
## 3953                            SELF-HELP FOR THE ELDERLY  100  100
100
## 4043                                             Sidekick  100  100
100
## 4146                                              SPRESSA  100  100
100
## 4177                                     Stanley Steamers  100  100
100
## 4255                                       Subway #61240  100  100
100
## 4333                                       Sushi In, LLC   93   93
93
## 4380                                             T-WE TEA  100  100
100
## 4494                                         Ten Ren Tea   96   96
96
## 4706                                      The Sweet House  100  100
100
## 4719                         The Village Market & Pizza   96   96
96
## 4796                                          Toyose INC  100  100
100
## 4879                                             Upcider   96   96
96
## 4996                               Western Sunset Market  100  100
100
## 5132   Zaida T. Rodriguez (ZTR) Child Development Center  100  100
100
##      variance
## 85          0
## 134         0
## 228         0
## 262         0
## 279         0
## 367         0
## 397         0
## 416         0
## 479         0
## 493         0
## 546         0
## 572         0
## 573         0
## 658         0
## 805         0
## 898         0
## 966         0
## 1025        0
## 1039        0
```

```
## 1040         0
## 1118         0
## 1129         0
## 1179         0
## 1194         0
## 1234         0
## 1400         0
## 1416         0
## 1504         0
## 1512         0
## 1609         0
## 1638         0
## 1651         0
## 1689         0
## 1798         0
## 1944         0
## 2003         0
## 2004         0
## 2005         0
## 2006         0
## 2032         0
## 2039         0
## 2080         0
## 2105         0
## 2118         0
## 2173         0
## 2283         0
## 2315         0
## 2343         0
## 2367         0
## 2454         0
## 2507         0
## 2630         0
## 2786         0
## 2808         0
## 2884         0
## 2885         0
## 2896         0
## 2903         0
## 2965         0
## 3011         0
## 3041         0
## 3228         0
## 3242         0
## 3260         0
## 3429         0
## 3443         0
## 3543         0
## 3560         0
## 3594         0
```

```
## 3608        0
## 3610        0
## 3666        0
## 3715        0
## 3718        0
## 3929        0
## 3953        0
## 4043        0
## 4146        0
## 4177        0
## 4255        0
## 4333        0
## 4380        0
## 4494        0
## 4706        0
## 4719        0
## 4796        0
## 4879        0
## 4996        0
## 5132        0

# All the restaurants that have 0 variance are the ones which are doing
good. There are no restaurants that are consistently bad.

#Choosing the top 50 rows/restaurants with highest variance
YoY_inconsistent = YoY_table[order(-YoY_table$variance), ][1:50,]
YoY_inconsistent

##                               business_name      2014      2015
2016
## 74                          ABC Bakery Cafe  46.00000  94.00000
72.00000
## 4583                               The Crew  96.00000  61.00000
100.00000
## 1840                        Gourmet Kitchen  55.00000  92.00000
92.00000
## 4900                        Velo Rouge Cafe  98.00000  66.00000
98.00000
## 5118       Yummy Dim Sum & Fast Food, LLC  50.00000  82.00000
81.00000
## 2732                             M.Y. China  90.00000 100.00000
66.00000
## 4980                            WE BE SUSHI  98.00000  82.00000
65.00000
## 492           BASIL THAI RESTAURANT & BAR  85.00000  54.00000
77.00000
## 4542                               The Bell  92.00000  62.00000
85.28571
## 4941                        Volcano Kitchen  92.80000  62.00000
83.00000
```

```
## 659                     Borodudur Restaurant  90.00000  96.00000
67.00000
## 1809                 Golden River Restaurant  73.60000  81.00000
52.00000
## 1590                              First Cake  58.91304  79.00000
88.00000
## 63                               A La Turca  90.00000  86.00000
63.00000
## 3425              Penang Garden  Restaurant  69.00000  94.00000
94.00000
## 4862                             Uncle Cafe  82.00000  76.00000
55.00000
## 206                            Another Cafe  98.00000  70.00000
87.85714
## 2045                       Hong Kong Lounge  90.00000  79.38462
61.94444
## 1835                   Gourment Noodle House  93.66667  98.00000
71.61538
## 1102                    City View Restaurant  96.00000  84.00000
68.00000
## 1762                   Glide Memorial Church  87.00000  72.00000
100.00000
## 5010               Whitcomb Hotel Bar & Grill  72.00000 100.00000
85.00000
## 534                           Begoni Bistro  96.00000  74.00000
70.00000
## 4406                     Tai Chi Restaurant  72.00000  58.00000
86.00000
## 798                             Cafe Europa  96.00000  94.00000
71.00000
## 3846                       SAJJ Street Eats  71.00000  93.00000
96.00000
## 1861                Great Eastern Restaurant  70.00000  76.00000
96.00000
## 2373                    Khan Toke Thai House  65.37500  90.00000
68.00000
## 4091                    Sodini's Restaurant  72.00000  65.00000
91.00000
## 1794               Golden Gate Dim Sum Seafood  92.00000  90.00000
68.00000
## 1937                              Happy Cafe  90.00000  92.00000
68.00000
## 3497                            Piraat Pizza  70.00000  94.00000
92.00000
## 5116               Yummy Bakery & Restaurant  86.00000  68.00000
94.00000
## 1021                           Chico's Grill  85.00000  65.00000
90.00000
## 3006                            MOULIN ROUGE  94.00000  69.00000
89.00000
```

```
## 831                               CAFE PICARO  93.00000  96.00000
72.00000
## 5067                         Wow Naan-N-Curry  75.00000  64.00000
90.00000
## 2058                       Horizon Restaurant  86.00000  63.00000
85.00000
## 2696                               Love Berry  96.00000  96.00000
73.60000
## 3237                         Olea Restaurant  86.00000  93.00000
68.00000
## 4089    Sodexo at Academy of Art University  76.00000 100.00000
96.00000
## 3460                              Pho Express  96.00000  88.00000
71.00000
## 4438                Taqueria El Buen Sabor #2 100.00000  75.00000
92.00000
## 464                               Bamboo Asia  79.00000 100.00000
77.00000
## 2582                              Les Joulins  53.00000  77.00000
72.00000
## 4321                            Supreme Pizza 100.00000  96.00000
76.61538
## 1081                        Chutney USA, Inc.  69.00000  82.00000
94.00000
## 2224                              Jade Garden  76.00000  83.00000
59.00000
## 1501 Equinox Fitness SC San Francisco, Inc.  76.00000  92.00000
100.00000
## 3775                                Ruby Skye  78.00000  94.00000
70.00000
##      variance
## 74   577.3333
## 4583 460.3333
## 1840 456.3333
## 4900 341.3333
## 5118 331.0000
## 2732 305.3333
## 4980 272.3333
## 492  259.0000
## 4542 247.8844
## 4941 247.6133
## 659  234.3333
## 1809 227.0533
## 1590 221.7561
## 63   212.3333
## 3425 208.3333
## 4862 201.0000
## 206  200.9592
## 2045 200.6600
## 1835 200.1975
```

```
## 1102 197.3333
## 1762 196.3333
## 5010 196.3333
## 534  196.0000
## 4406 196.0000
## 798  193.0000
## 3846 186.3333
## 1861 185.3333
## 2373 182.8802
## 4091 181.0000
## 1794 177.3333
## 1937 177.3333
## 3497 177.3333
## 5116 177.3333
## 1021 175.0000
## 3006 175.0000
## 831  171.0000
## 5067 170.3333
## 2058 169.0000
## 2696 167.2533
## 3237 166.3333
## 4089 165.3333
## 3460 163.0000
## 4438 163.0000
## 464  162.3333
## 2582 160.3333
## 4321 156.4339
## 1081 156.3333
## 2224 152.3333
## 1501 149.3333
## 3775 149.3333
```

```
# Choosing restaurants that have improved YoY
YoY_inconsistent_good = subset(YoY_inconsistent,
YoY_inconsistent$`2015` > YoY_inconsistent$`2014` &
YoY_inconsistent$`2016` > YoY_inconsistent$`2015`)
YoY_inconsistent_good
```

```
##                                business_name     2014 2015 2016
variance
## 1590                              First Cake 58.91304   79   88
221.7561
## 3846                        SAJJ Street Eats 71.00000   93   96
186.3333
## 1861                Great Eastern Restaurant 70.00000   76   96
185.3333
## 1081                      Chutney USA, Inc. 69.00000   82   94
156.3333
## 1501 Equinox Fitness SC San Francisco, Inc. 76.00000   92  100
149.3333
```

```
# There are 6 out of 50 which have improved considerably YoY

#Choosing restaurants that worsened YoY
YoY_inconsistent_bad = subset(YoY_inconsistent, YoY_inconsistent$`2015`
< YoY_inconsistent$`2014` & YoY_inconsistent$`2016` <
YoY_inconsistent$`2015`)
YoY_inconsistent_bad

##                        business_name 2014     2015     2016 variance
## 4980                     WE BE SUSHI   98 82.00000 65.00000 272.3333
## 63                        A La Turca   90 86.00000 63.00000 212.3333
## 4862                      Uncle Cafe   82 76.00000 55.00000 201.0000
## 2045                Hong Kong Lounge   90 79.38462 61.94444 200.6600
## 1102            City View Restaurant   96 84.00000 68.00000 197.3333
## 534                    Begoni Bistro   96 74.00000 70.00000 196.0000
## 798                      Cafe Europa   96 94.00000 71.00000 193.0000
## 1794 Golden Gate Dim Sum Seafood       92 90.00000 68.00000 177.3333
## 3460                     Pho Express   96 88.00000 71.00000 163.0000
## 4321                   Supreme Pizza  100 96.00000 76.61538 156.4339

# There are 7 out of 50 which have dropped considerably YoY
```

Finding:

1.Restuarants that have been performing consistently well from 2014 to 2016 are - Totally 50 restuarants have been doing well over the years.Top three are Acme Bread Company, Alex Gourmet Burrito and ARGONNE ALTERNATIVE ELEMENTARY SCHO.

2.Restuarants that have been inconsistent but improving from 2014 to 2016 are - Totally 6 restuarants have improved over the years and top three are First Cake,SAJJ Street Eats and Great Eastern Restaurant

3.Restuarants that been been worsened from 2014 to 2016 -Totally 7 restuarants have worsened over the years and top three are WE BE SUSHI, A La Turca and Uncle Cafe.

Consistency based on location

Find the areas in the SF city where the restuarants have been doing consistently well and worsened over years?

```
setwd("/Users/shruthi/RWorkspace/")
# Read the Data
pin_data = read.csv("/Users/shruthi/RWorkspace/Restaurant_Scores_-
_LIVES_Standard.csv")

# Choose the columns needed for the analysis
pin_data =
pin_data[c("business_postal_code","inspection_date","inspection_score")
```

```
]

# Choose the dates column and chnage format
dates <- pin_data$inspection_date
tmp <-  as.Date(dates,'%m/%d/%Y')
year = format(tmp,'%Y')

# Add the year column to the data frame
pin_data$inspection_year = year

# Reorder the data columns and remove previous date column
pin_data = pin_data[c(1,4,3)]

# Remove the NAs from inspection scores column
pin_data = pin_data[complete.cases(pin_data[,3]), ]

# Group the data by pin and year
library(dplyr)
pin_summary = group_by(pin_data, business_postal_code, inspection_year)
pin_summary = summarize(pin_summary, avg_score =
round(mean(inspection_score)))

# Represent the data in the form of a table
library(reshape)
pin_table = cast(pin_summary, business_postal_code ~ inspection_year,
value = "avg_score")

# Remove NAs once again (Since it is the form of table, we remove pins
with no score during any year)
pin_table = pin_table[complete.cases(pin_table[,2:5]), ]
pin_table

##    business_postal_code 2014 2015 2016 2017
## 1                         88   90   88   94
## 9                 94102   85   86   85   88
## 10                94103   87   85   85   87
## 12                94104   89   93   85   89
## 14                94107   91   90   92   88
## 15                94108   83   86   83   82
## 16                94109   86   82   82   87
## 17                94110   89   88   88   90
## 19                94111   89   91   86   90
## 20                94112   89   88   85   89
## 21                94114   87   86   92   90
## 22                94115   87   84   87   81
## 23                94116   86   88   86   90
## 24                94117   89   84   89   92
## 25                94118   89   87   83   82
## 27                94121   86   84   81   92
```

```
## 28                  94122   83   82   85   89
## 30                  94124   90   89   92   96
## 34                  94131   91   93   87   86
## 35                  94132   92   93   88   88
## 36                  94133   82   83   84   83
## 37                  94134   88   88   82   82

# Finding variance of the scores over time
pin_table$variance = rowVars(data.matrix(pin_table[,2:5]))

# Choosing the rows/pins with 0 variance
pin_consistent = pin_table[pin_table$variance == 0, ]
pin_consistent

## [1] business_postal_code 2014                 2015
## [4] 2016                 2017                 variance
## <0 rows> (or 0-length row.names)

# No pin codes with 0 variance

# Choosing rows/pins that have improved YoY
pin_inconsistent_good = subset(pin_table, pin_table$`2015` >
pin_table$`2014` & pin_table$`2016` > pin_table$`2015` &
pin_table$`2017`>pin_table$`2016`)
pin_inconsistent_good

## [1] business_postal_code 2014                 2015
## [4] 2016                 2017                 variance
## <0 rows> (or 0-length row.names)

# There are no pin codes which have improved YoY

# Choosing rows/pins that worsened YoY
pin_inconsistent_bad = subset(pin_table, pin_table$`2015` <
pin_table$`2014` & pin_table$`2016` < pin_table$`2015` &
pin_table$`2017`< pin_table$`2016`)
pin_inconsistent_bad

##    business_postal_code 2014 2015 2016 2017 variance
## 25                  94118   89   87   83   82 10.91667

# There are 1 pin code which has dropped considerably YoY

# Aggregating the scores of pins over all years
aggregate_pin_summary = pin_data[c(1,3)]
aggregate_pin_summary$len <-
nchar(as.character(aggregate_pin_summary$business_postal_code))
aggregate_pin_summary =
aggregate_pin_summary[(aggregate_pin_summary$len==5),]
aggregate_pin_summary = group_by(aggregate_pin_summary,
business_postal_code)
```

```
aggregate_pin_summary = summarize(aggregate_pin_summary, avg_score =
mean(inspection_score))
library (ggplot2)
ggplot(aggregate_pin_summary, aes(x=business_postal_code, y=avg_score))
+
  geom_bar(stat='identity', width=0.5, fill = "#FF6666") + labs(y =
"Inspection Score", x = "Postal Codes", title = "Inspection Scores for
different Postal Codes")+
  coord_flip()+theme(text = element_text(size=7), axis.text.x =
element_text(angle=0, hjust=1))
```

Finding:

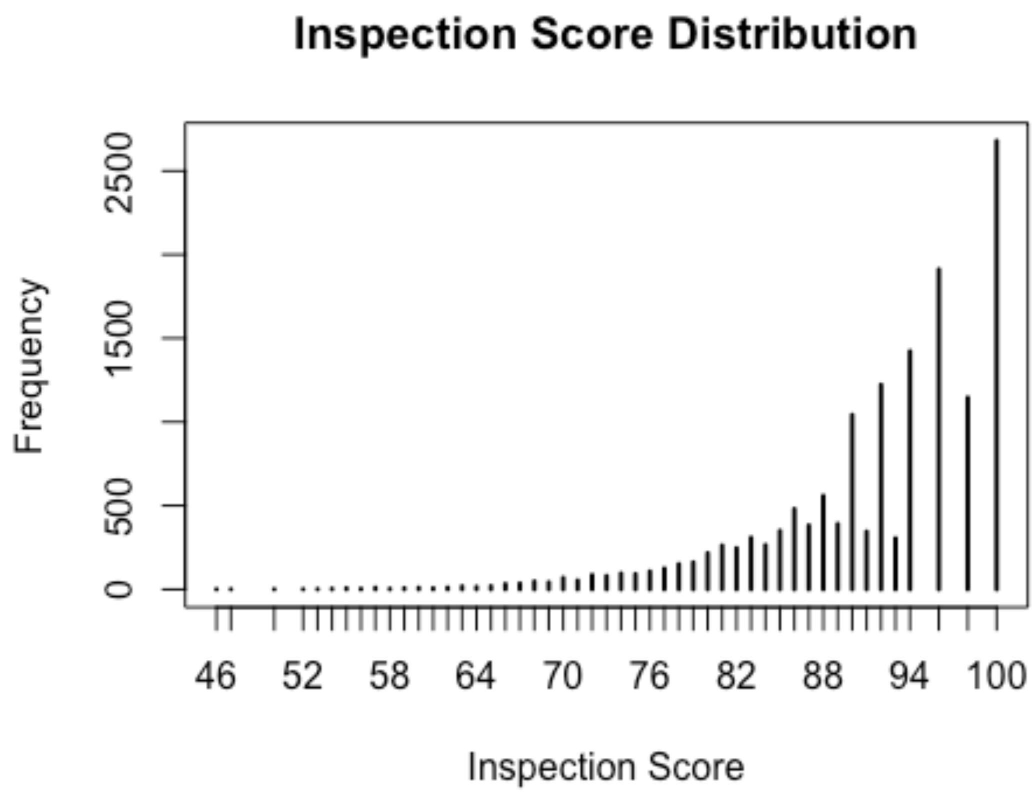
The best area in the SF bay area to eat out based on three year data is Hayward(94545) and San Bruno(94066)

The worst area to eat out is Daly City(94014)

Inspection score distribution

For the past three years what is the inspection score distribution like for SF Restuarants?

```r
setwd("/Users/shruthi/RWorkspace/")
# Read the Data
raw_data = read.csv("/Users/shruthi/RWorkspace/Restaurant_Scores_-
_LIVES_Standard.csv")

# Choose the columns needed for the analysis
inspec_data =
raw_data[c("business_id","inspection_score","inspection_date")]

# Choose the dates column and change format
dates <- inspec_data$inspection_date
dates <-  as.Date(dates,'%m/%d/%Y')
year = format(dates,'%Y')

# Add the date, year column to the data frame
inspec_data$inspection_date = dates
inspec_data$inspection_year = year

# Remove the NAs/blanks from inspection scores column
inspec_data = inspec_data[complete.cases(inspec_data[,2]), ]

# Group the data by business and date
library(dplyr)
inspec_summary = group_by(inspec_data, business_id, inspection_date)
inspec_summary = summarize(inspec_summary, avg_score =
mean(inspection_score))

# Create a frequency table for inspection scores
inspec_table = table(inspec_summary$avg_score)

# Plot the inspection scores
plot(inspec_table, xlab = "Inspection Score", ylab = "Frequency", main
= "Inspection Score Distribution")
```




```
#inspec_summary =
inspec_summary[!duplicated(inspec_summary$inspection_score),]
```

Finding:

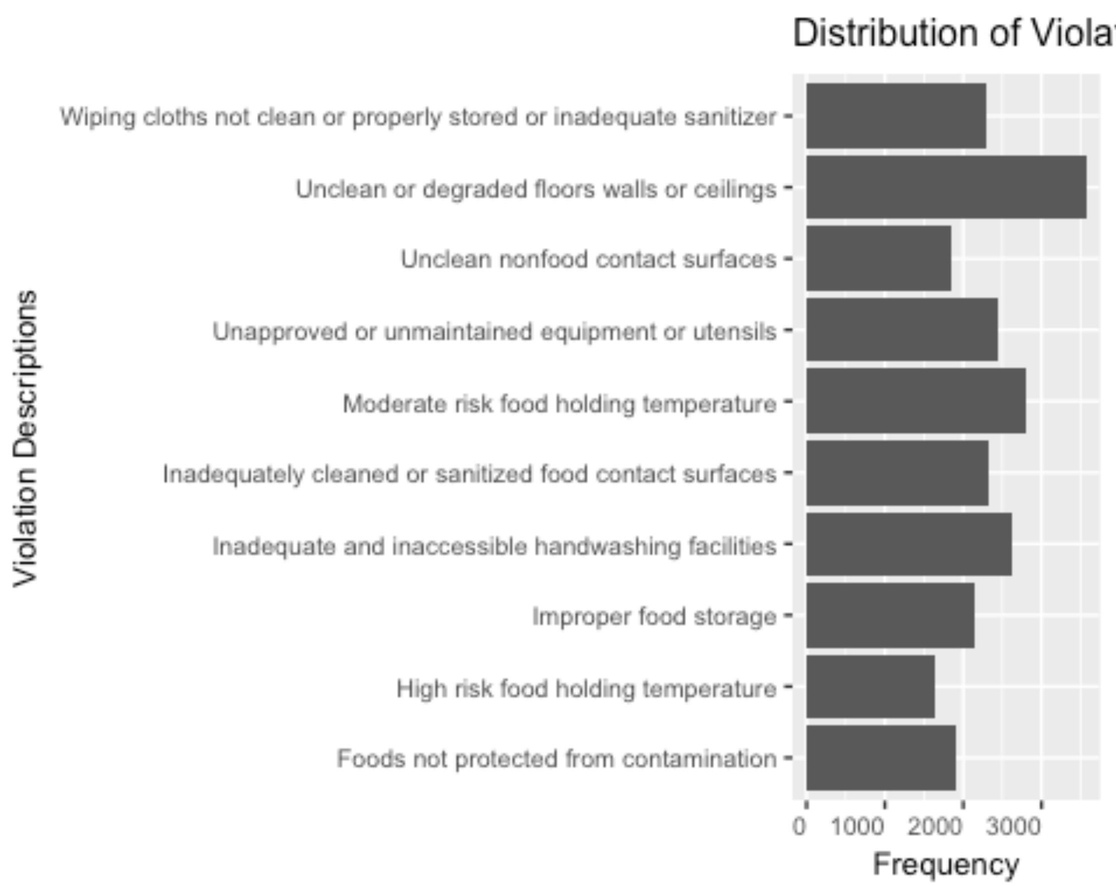
Over the three years,more number of restuarants have scored 96 and 100.

Violation description distribution

For the past three years what is the inspection score distribution like for SF restuarants?

```
setwd("/Users/shruthi/RWorkspace/")
# Read the Data
raw_data = read.csv("/Users/shruthi/RWorkspace/Restaurant_Scores_-
_LIVES_Standard.csv")
# Choose the columns needed for the analysis
violation_data = raw_data["violation_description"]

# Remove the NAs/blanks from inspection scores column
violation_data = violation_data[complete.cases(violation_data[,1]), ]

# Group the data by business and date
```

```
# library(dplyr)
# inspec_summary = group_by(inspec_data, business_id, inspection_date)
# inspec_summary = summarize(inspec_summary, avg_score =
mean(inspection_score))

# Create a frequency table for inspection scores
violation_table = table(violation_data)

# Sort the table based on frequencies and plot the frequencies for top
10 violations
violation_df = as.data.frame(violation_table)
violation_df = violation_df[(violation_df$violation_data != ""),]
violation_sorted = violation_df[order(-violation_df$Freq),]
library (ggplot2)
ggplot(violation_sorted[1:10,], aes(x=violation_data, y=Freq)) +
  geom_bar(stat='identity') + labs(y = "Frequency", x = "Violation
Descriptions", title = "Distribution of Violation Descriptions")+
  coord_flip()+theme(text = element_text(size=10), axis.text.x =
element_text(angle=0, hjust=1))
```



Finding: Over the past three years, the violation that was committed most by the SF restuarants is 'Unclean or degraded floors, walls or ceilings'

For a score of 100, what are the violations?

Even though a restaurant scores 100 there could be different types of violations.Here we are trying to find the various violations committed by the restuarants in the past three years

```
setwd("/Users/shruthi/RWorkspace/")

# Read the Data
raw_data = read.csv("/Users/shruthi/RWorkspace/Restaurant_Scores_-
_LIVES_Standard.csv")

# Choose the columns needed for the analysis
topscore_data = raw_data[c("inspection_score",
"violation_description")]

# Choose the data where inspection score is 100 and no blanks
topscore_data = topscore_data[complete.cases(topscore_data[,1]), ]
topscore_data = topscore_data[topscore_data$inspection_score == 100,]
topscore_data = topscore_data[topscore_data$violation_description !=
"",]

# Create a frequency table for violation descriptions
topscore_table = table(topscore_data$violation_description)

# Sort the table based on frequencies
topscore_df = as.data.frame(topscore_table)
colnames(topscore_df) = c("Violation Description", "Frequency")
topscore_sorted = topscore_df[order(-topscore_df$Frequency),]
head(topscore_sorted)

##                                                  Violation
Description
## 22                     Inadequate and inaccessible handwashing
facilities
## 37                                 Moderate risk food holding
temperature
## 68 Wiping cloths not clean or properly stored or inadequate
sanitizer
## 31                Inadequately cleaned or sanitized food contact
surfaces
## 38                                       Moderate risk vermin
infestation
## 14                                                  Improper food
storage
##    Frequency
## 22        11
## 37        11
## 68         6
```

```
## 31          5
## 38          5
## 14          4
```

Finding:

The top 2 violations committed by restuarants who have scored 100 are 'Inadequate and inaccessible handwashing facilities' and 'Moderate risk food holding temperature'.

Conclusion:

Based on our exploratory analysis we found out that even though all the restuarants have made the violation of not having convenient hand washing facility,other high risk violations have been found less.Although it is safe to eat in most areas in SF, the best restuarants to dine out are in the areas of Hayward and San Bruno and the worst area to eat would be Daly City.