INFORMAC: Analytics Data Science Challenge

...

- Placed Runner Up in University Challenge

Overview

The city of Los Angeles publishes data on Environment Health inspection and enforcement results from restaurants in the Los Angeles county. These data cover 85 of 88 cities and all unincorporated areas in the LA county.

There are two datasets available: (i) market inspection dataset: contains results of inspection; and (ii) market violations dataset: contains information on health code violations in restaurants. These data were last updated on January 16, 2019. Data dictionaries for the above two data sets are included below. Feel free to supplement the above information with other publicly-available information.

Understanding the problem: Problem Statement

Q 1	Q 2	Q 3	Q 4
What are the key factors in predicting health “scores” of the restaurants in Los Angeles county?	What are the most important factors in classifying restaurants into different “grades”?	Are there any relationships between various types of health code violations and scores/grades of a restaurant?	Are there any patterns in terms of how health scores of restaurants change over time?

Approach:

Timeline

Data Preprocessing

- Removing NaN, duplicates
- Transforming categorical into one hot encoding

Visualize data for Anomalies

Heatmap for correlation | box plots for outliers

Ensemble Learning

- Random Forest Feature Importance
- Chi Square for categorical variable comparison

Understanding important aspects of Dataset

Activity Date	• Date of health inspection • First inspection date taken as 1, and every month followed with increment of 1, including the year
Serial Number	• Foreign key - used to join inspection and violation datasets
Zip Code /Location	• Calculate distance (in miles) from one reference lat long
Score - Grade	• Correlated variables • Score used for regression; Grade for classification

Pre-Processing

Calculated relative distance (in miles) of the restaurants and used them as interval data for modeling.

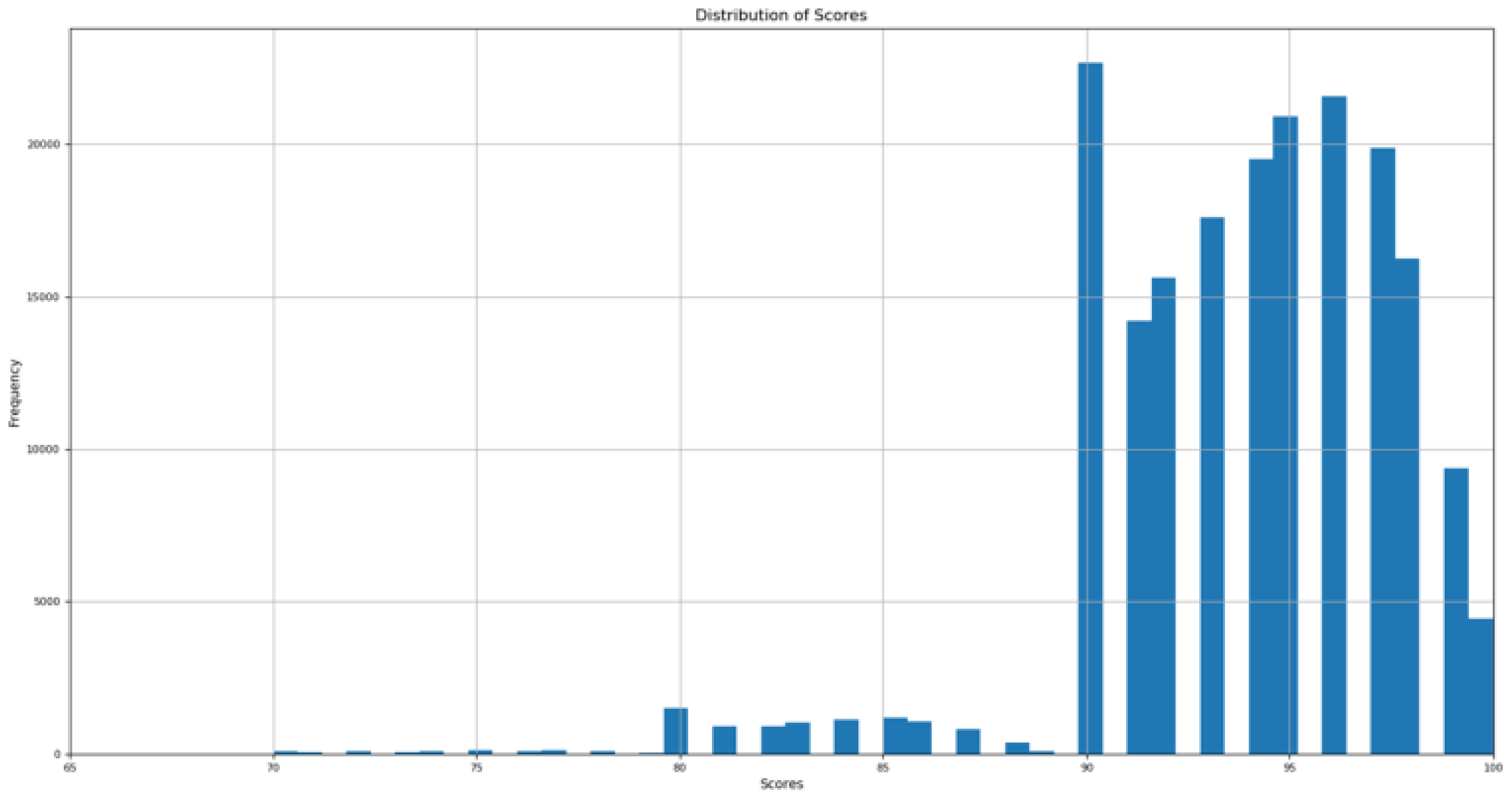
Program description is parsed to get establishment type, seating capacity and the risk factor, used regex in python

Longitude	Latitude	Distance in miles
-118.295	34.04771	0
-118.326	34.098	3.906000639
-118.233	33.94311	8.05014473
-118.536	34.1935	17.0901967
-118.3	34.17459	8.773421187
-118.311	34.0617	1.351031373
-118.078	33.90229	15.98451445
-118.102	34.07957	11.24438867
-118.369	34.06378	4.374914302
-118.255	34.04535	2.28303536
-118.236	34.22171	12.48491766
-118.352	33.89278	11.20538452
-118.091	33.91717	14.75363495
-118.48	33.99446	11.24842351
-117.89	34.07505	23.23861435
-118.26	34.07317	2.673523817
-118.015	34.09101	16.27768292
-118.288	34.09816	3.505837149

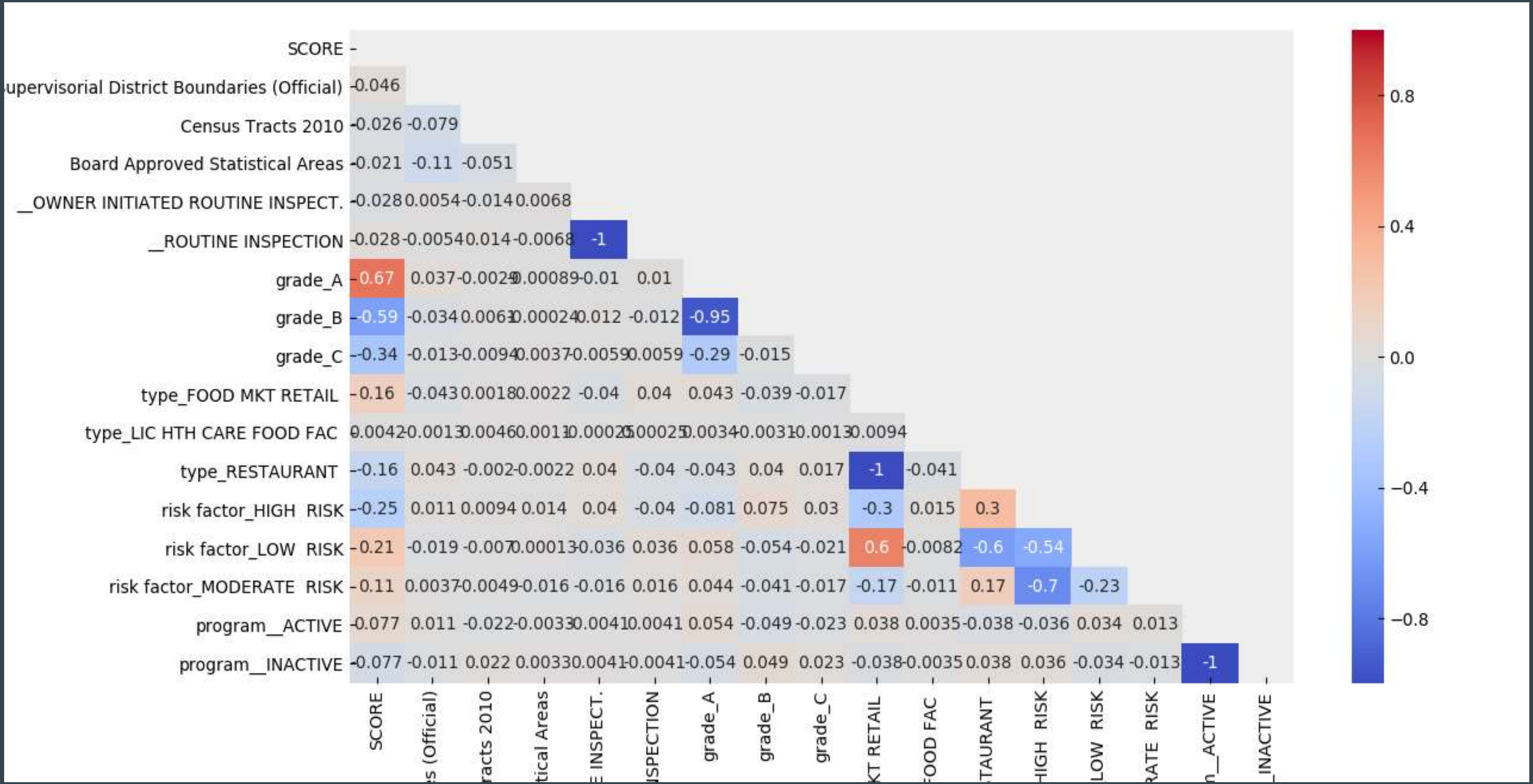
PE DESCRIPTION
RESTAURANT (0-30) SEATS HIGH RISK
RESTAURANT (0-30) SEATS HIGH RISK
FOOD MKT RETAIL (1-1,999 SF) LOW RISK
RESTAURANT (61-150) SEATS MODERATE RISK
RESTAURANT (0-30) SEATS HIGH RISK
RESTAURANT (31-60) SEATS MODERATE RISK
RESTAURANT (61-150) SEATS HIGH RISK
FOOD MKT RETAIL (1-1,999 SF) MODERATE RISK
RESTAURANT (0-30) SEATS HIGH RISK
RESTAURANT (0-30) SEATS MODERATE RISK
RESTAURANT (31-60) SEATS MODERATE RISK
RESTAURANT (0-30) SEATS HIGH RISK
RESTAURANT (0-30) SEATS MODERATE RISK
FOOD MKT RETAIL (2,000+ SF) LOW RISK
FOOD MKT RETAIL (2,000+ SF) MODERATE RISK
FOOD MKT RETAIL (1-1,999 SF) LOW RISK
RESTAURANT (151 +) SEATS HIGH RISK
RESTAURANT (61-150) SEATS MODERATE RISK
RESTAURANT (0-30) SEATS HIGH RISK
RESTAURANT (31-60) SEATS HIGH RISK
RESTAURANT (0-30) SEATS HIGH RISK

type	seating size	risk factor
1	0-30	1
1	0-30	1
2	1-1,999 SF	3
1	0-30	1
1	31-60	2
1	61-150	1
1	0-30	1
1	0-30	2
1	31-60	2
1	0-30	1
1	0-30	2
2	2,000+ SF	2
2	1-1,999 SF	3
1	151 +	1
1	61-150	2
1	0-30	1
1	31-60	1

EXPLANATORY DATA ANALYSIS


Distribution of Scores
Frequency
Scores
0
5000
10000
15000
20000
65
70
75
80
85
90
95
100

CORRELATION CHART



RESTAURANT COUNT ACCORDING TO RISK FACTOR

```
risk factor
HIGH  RISK        119771
LOW  RISK          28414
MODERATE  RISK     43703
```

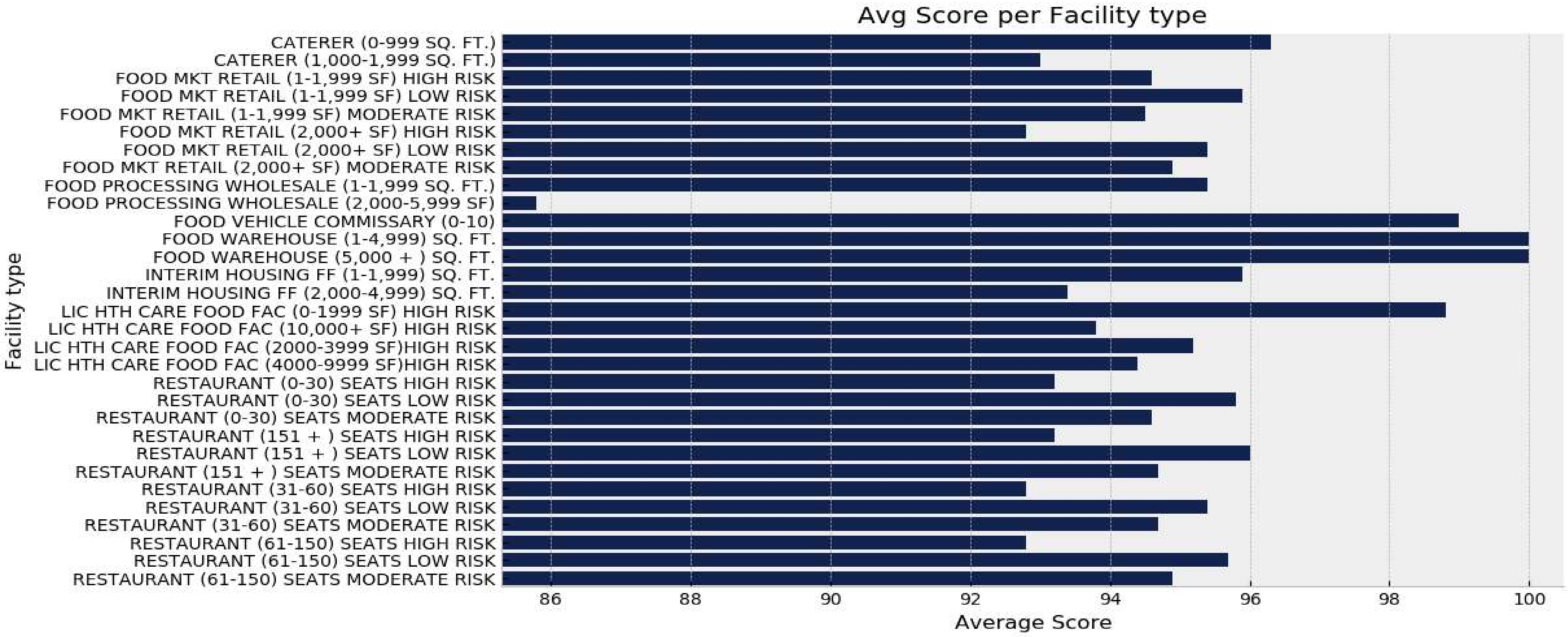
TOP 10 RESTAURANTS WITH HIGH SCORE

	FACILITY NAME	SCORE
101592	IHOP	100
147153	COSTCO WHOLESALE #117	100
125832	LOS ANGELES CITY COLLEGE BO	100
90194	PRIME LIQUOR MARKET	100
111703	NORTHGATE MARKET #4	100
24335	STARBUCKS COFFEE #5708	100
66397	GRIFFIN CLUB LOS ANGELES	100
184881	SUPER VALUE + EXPRESS	100
111656	BAJA FRESH #130	100
66423	POMONA FISH MARKET	100

BOTTOM 10 RESTAURANTS

	FACILITY NAME	SCORE
183329	CAFE CON LECHE BAKERY AND CAFE	70
99056	MARISCOS CHENTE	70
41351	MANDARIN DISH	70
55449	GOLD HIBACHI BUFFET	70
153624	EL OAXAQUENO BAKERY	70
33689	PONCE'S BAKERY	70
124683	LIVE BASIL PIZZA-SMASH BURGER-TOM'S URBAN	70
20633	SPARE TIRE	70
28960	PALETERIA Y NEVERIA FIESTA MICHOACAN	70
135318	GARAGE KITCHEN	70

AVERAGE SCORE PER FACILITY TYPE



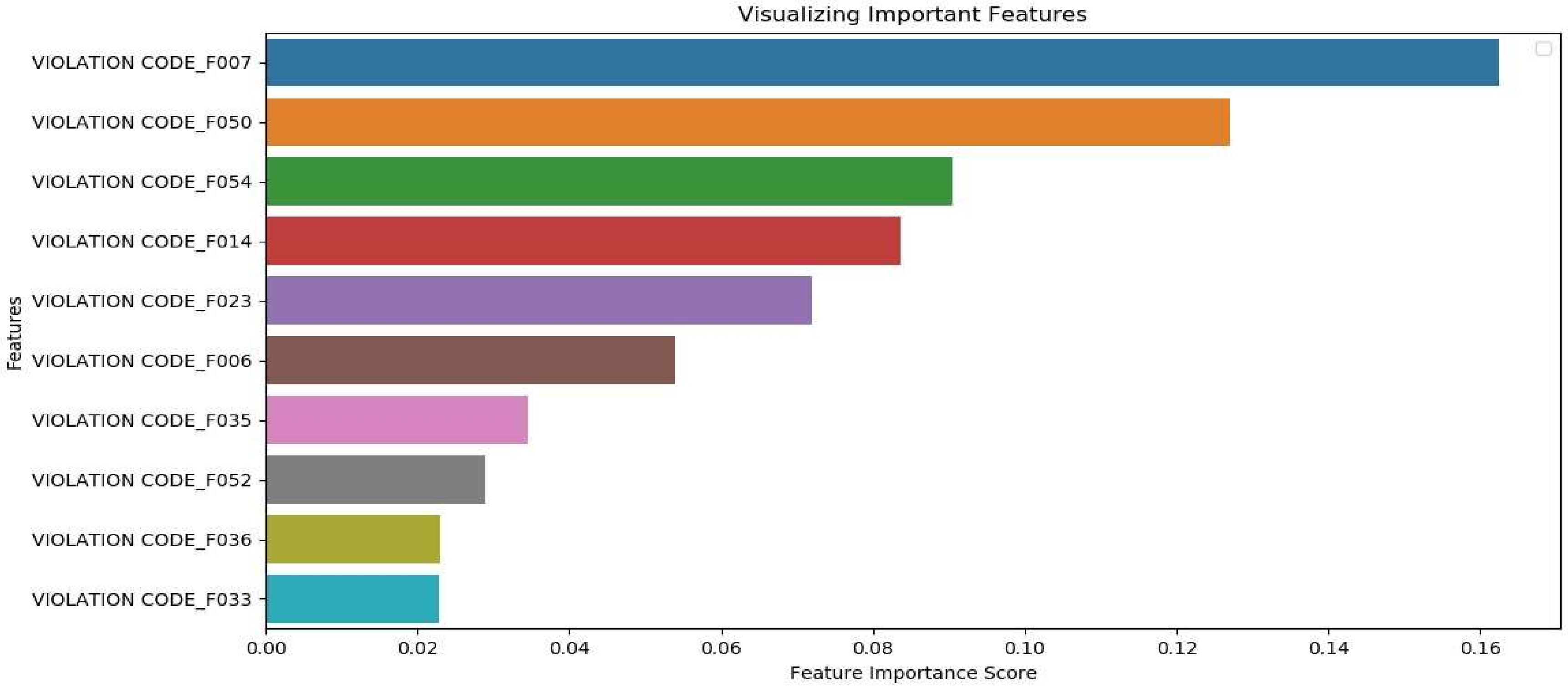
Top 15 violation codes

VIOLATION DESCRIPTION	VIOLATION CODE	Count
# 44. Floors, walls and ceilings: properly buil...	F044	106608
# 33. Nonfood-contact surfaces clean and in goo...	F033	103926
# 35. Equipment/Utensils - approved; installed;...	F035	83417
# 40. Plumbing: Plumbing in good repair, proper...	F040	52977
# 36. Equipment, utensils and linens: storage a...	F036	52451
# 37. Adequate ventilation and lighting; design...	F037	50165
# 43. Premises; personal/cleaning items; vermin...	F043	45496
# 07. Proper hot and cold holding temperatures	F007	42511
# 30. Food properly stored; food storage contai...	F030	40213
# 14. Food contact surfaces: clean and sanitized	F014	38802
# 39. Wiping cloths: properly used and stored	F039	36336
# 06. Adequate handwashing facilities supplied ...	F006	35572
# 23. No rodents, insects, birds, or animals	F023	32734
# 34. Warewashing facilities: Adequate, maintai...	F034	20760
# 29. Toxic substances properly identified, sto...	F029	19902

Q1:

- Problem Statement : to predict the scores of the restaurants using random forest regression.
- Target Variable : Score
- Predictors : ‘PROGRAM STATUS’, ‘SERVICE CODE’, ‘Distance in miles’, ‘2011 Supervisorial District Boundaries Official ‘, ‘Census Tracts 2010’, ‘Board Approved Statistical Areas’, ‘type’, ‘risk factor’, ‘month’, and all one-hot encoded violation codes
- Accuracy : 94.32%
- Used bagging regressor and classifier random forest feature selection to display top 10 features

Key Factors In predicting health scores.

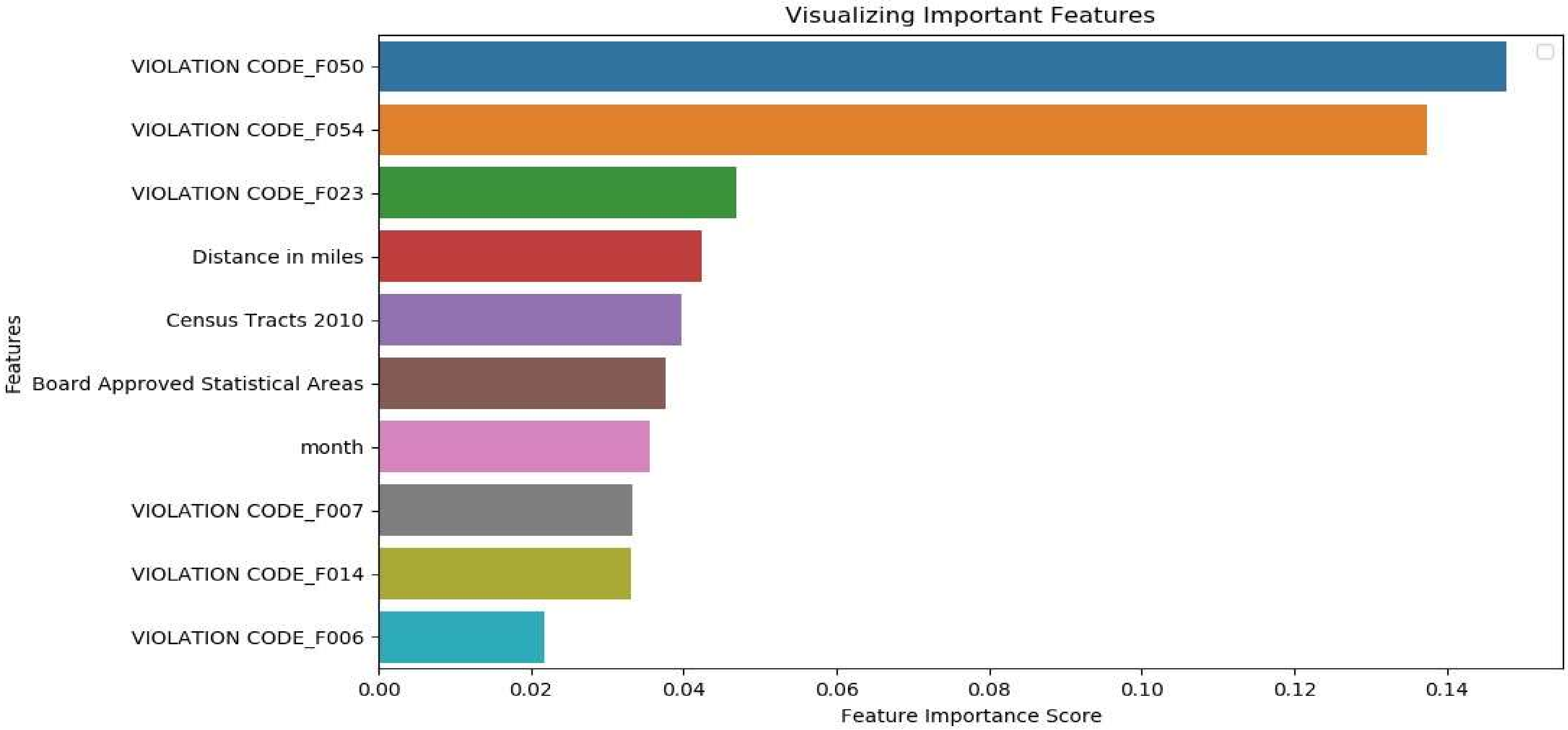


Q2 :

- Problem statement : to classify the restaurants based on grade
- Model : Building a classifier using Random Forest
- Target Variable : Grade
- Predictors : ‘PROGRAM STATUS’, ‘SERVICE CODE’, ‘Distance in miles’, ‘2011 Supervisorial District Boundaries Official ‘, ‘Census Tracts 2010’, ‘Board Approved Statistical Areas’, ‘type’, ‘risk factor’, ‘month’, and all one-hot encoded violation codes
- Accuracy : 96.70 %

```
inf
"C:\Users\Amit Mutgi\AppData\Local\Programs\Python\Python37\python.exe" "C:/Users/Amit Mutgi/Documents/informac/inf.py"
Accuracy: 0.9670277989174584

Process finished with exit code 0
```

Key Factors In classifying restaurants into grades.



- Selecting the top 10 features based on importance score
- Re-generating the model on selected features :
- Accuracy : 97.56 %

```
"C:\Users\Amit Mutgi\AppData\Local\Programs\Python\Python37\python.exe" "C:/Users/Amit Mutgi/Documents/informac/inf.py"
Accuracy: 0.9756983240223464

Process finished with exit code 0
```

- Removing the least important features results increased the accuracy. This is because of removal of misleading data and noise. Also, reduced training time increases the efficiency.

Q3.

Count of restaurant grade totals	
GRADE	
A	182171
B	8986
C	880

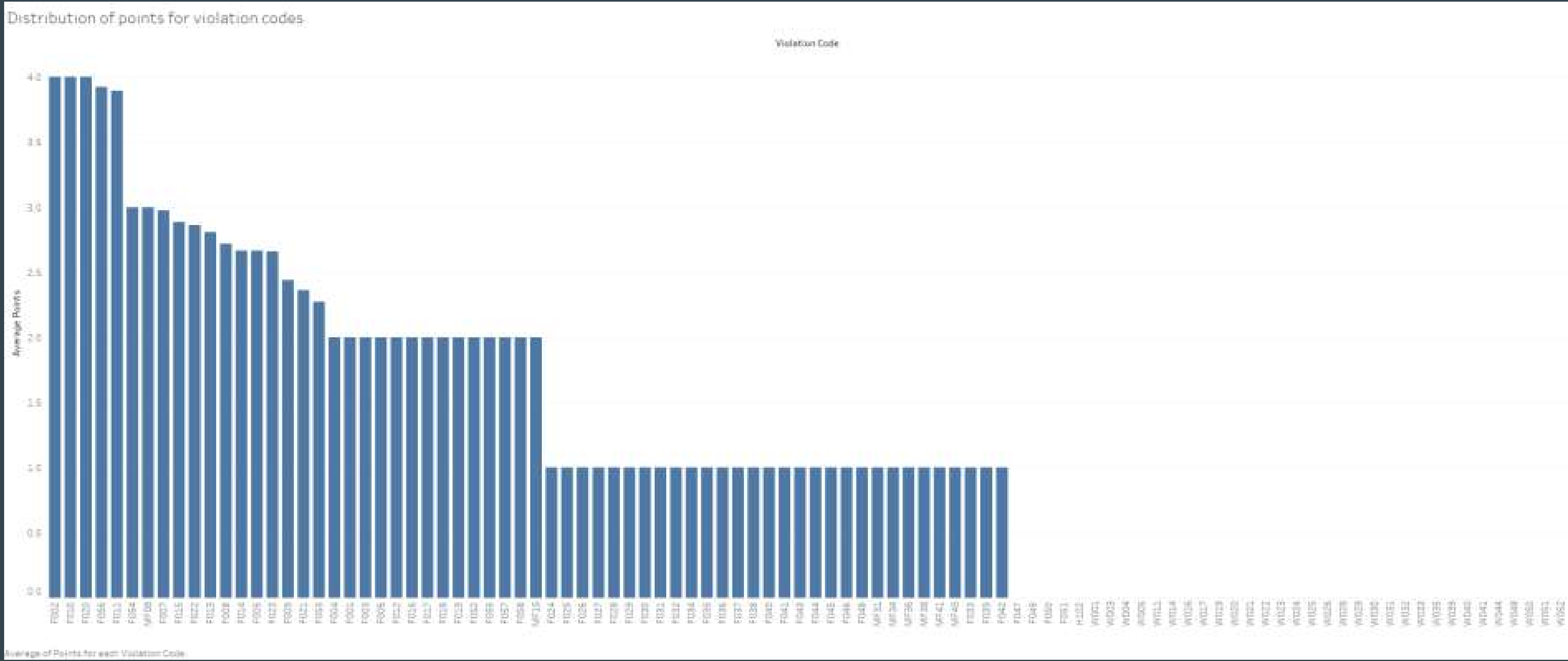
```
Chi Square Stats for Grade of restaurant:
Power_divergenceResult(statistic=1341701.079781985, pvalue=0.0)


Chi Square Stats for Violation Codes :
Power_divergenceResult(statistic=4295120.612272515, pvalue=0.0)


Chi Square Stats
54918.0092155784


Degrees of Freedom
194


P-Value
0.0


Contingency Table
[[2.46873205e+03 3.11035356e+01 3.11035356e+01 3.17522665e+03
  4.44780559e+03 3.19184482e+04 3.70691937e+04 4.81216129e+03
  6.91831496e+03 2.13281387e+01 6.59394954e+02 2.13281387e+01
  4.35052030e+03 3.50732354e+04 6.62060971e+02 1.10373115e+03
  3.91015876e+01 7.10937956e+01 1.06640693e+02 8.88672445e-01
  1.29746177e+03 1.58094828e+03 2.72742460e+04 7.39375474e+02
  1.10488645e+04 7.36620590e+03 1.62644831e+04 1.70625109e+02
  1.79556267e+04 3.58660155e+04 5.74002399e+02 5.87501353e+03
  9.25816953e+04 1.87207737e+04 7.41028405e+04 4.71476279e+04
  4.50148140e+04 1.00348892e+04 3.16669052e+04 4.74871008e+04
  4.17587182e+03 1.55642092e+04 4.07634050e+04 9.51803735e+04
  4.35449498e+02 1.56735159e+04 6.62949644e+02 5.45111678e+03
  8.53036680e+03 2.57004071e+03 4.35449498e+01 1.51198730e+04
  8.62367741e+03 2.39497224e+03 4.59887990e+03 4.26562774e+01
  1.89287231e+02 7.37558129e+02 8.88672445e-01 1.77734489e+00
  1.77734489e+00 8.88672445e-01 8.88672445e-01 4.44336222e+00
  8.88672445e-01 8.88672445e-01 4.44336222e+00 8.88672445e-01
  1.77734489e+00 1.77734489e+00 5.33203467e+00 1.77734489e+00
  1.77734489e+00 8.88672445e-01 2.66601733e+00 4.44336222e+00
  2.66601733e+00 2.66601733e+00 8.88672445e-01 1.06640693e+01
  8.88672445e-01 8.88672445e-01 2.66601733e+00 8.88672445e-01
  8.88672445e-01 8.88672445e-01 2.66601733e+00 8.88672445e-01
  8.88672445e-01 8.88672445e-01 8.88672445e-01 8.88672445e-01
  8.88672445e-01 2.66601733e+00 8.88672445e-01 3.55468978e+00
  7.99805200e+00 8.88672445e+00]
 [2.72031586e+02 3.42732379e+00 3.42732379e+00 3.49880798e+02
  4.90107302e+02 3.51711968e+03 4.08468450e+03 5.30255953e+02
  7.62334735e+02 2.35016489e+00 7.26592644e+01 2.35016489e+00
  4.68368277e+02 3.86474823e+03 7.29530350e+01 1.21621033e+02
```

- Chi – square statistical test used since we are dealing with categorical variables
- The results of the test will help us to understand whether a relation exists between the grades of the restaurant and violation codes.
- Hypothesis :

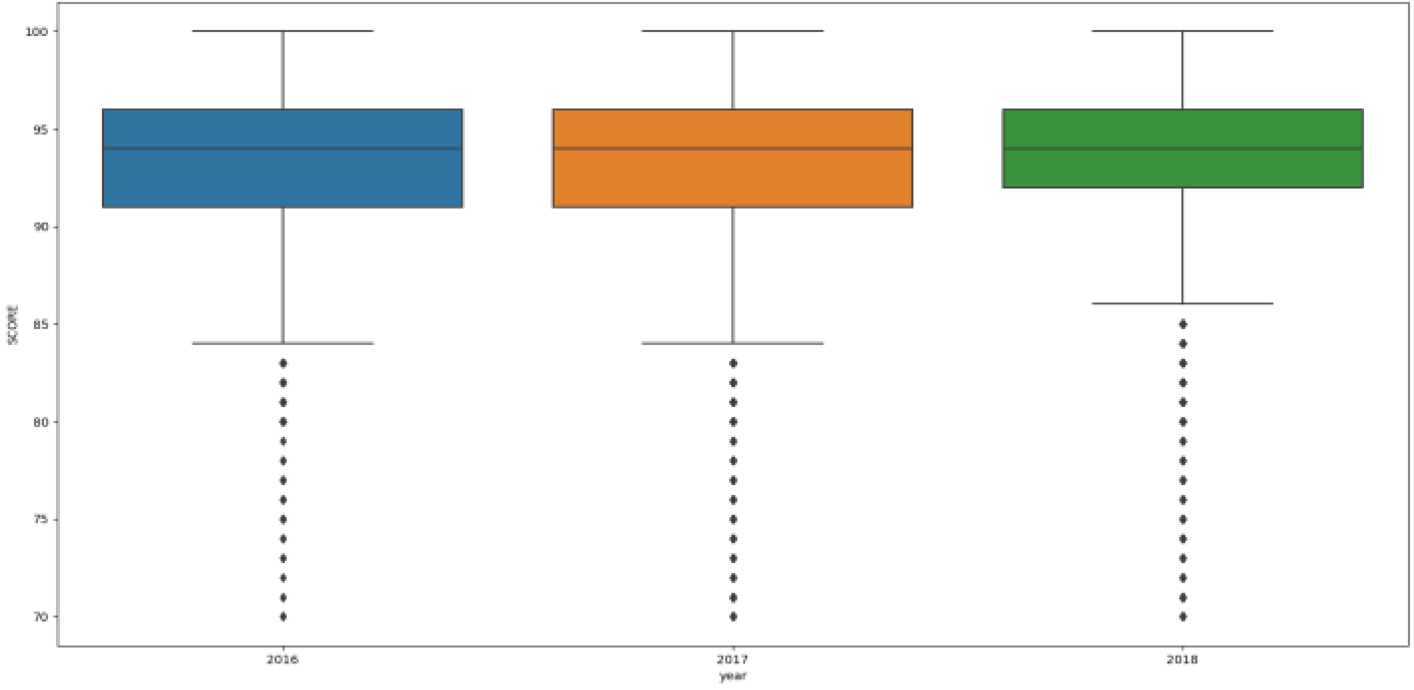
    Null hypothesis : Grades of the restaurant and violation codes are not related

    Alternate hypothesis : There exists a relation between the grades & violation codes
- The chi square statistic value compares the counts of categorical responses between the two independent group.
- Considering 5% significance level, p value has a value less than 0.05, showing that it is significant and null hypothesis can be rejected.
- We can conclude that there is a relationship between grades of the restaurant and violation codes.

Distribution of points for violation codes



Q4. Variation of Score across Years

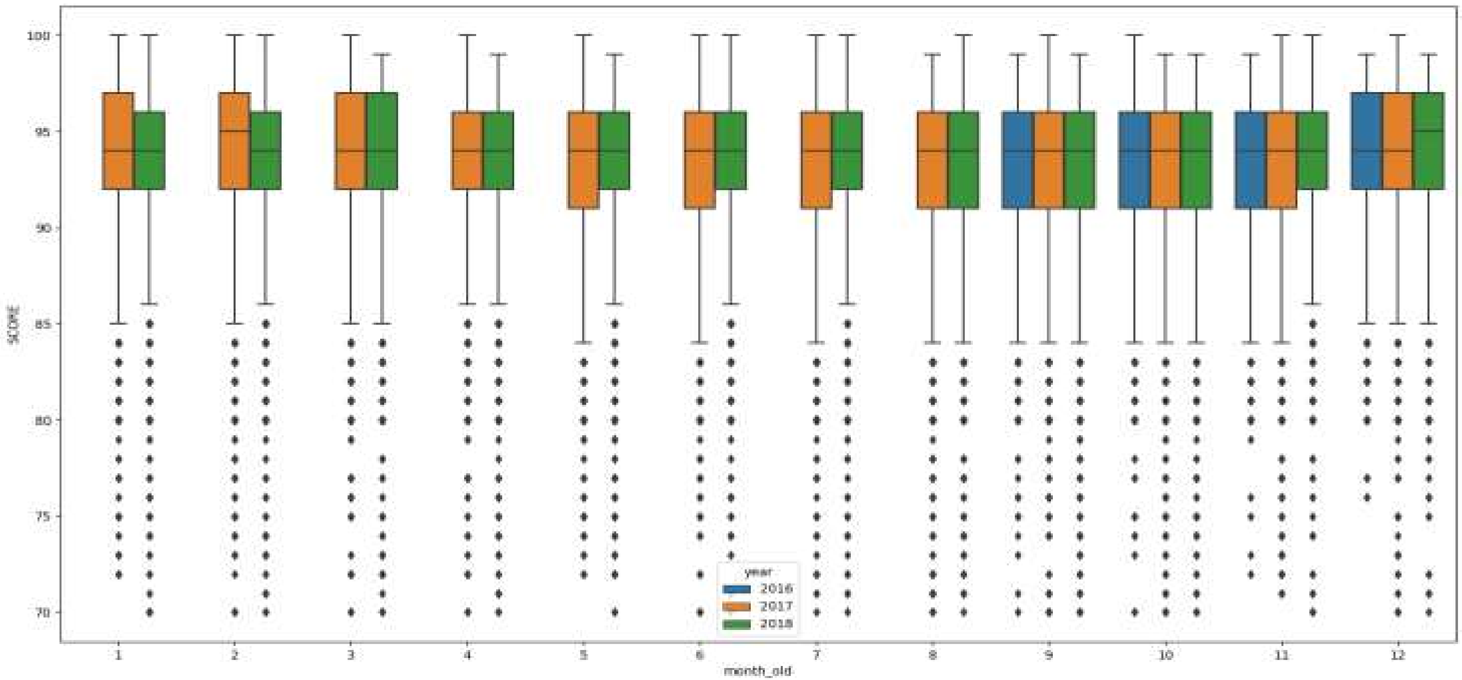


1. The scores have remained constant over the three years.

2. However in the year 2018 ,the restaurants having lower scored have improved .

3. This can be due to the fact that people prefer more hygienic restaurants.

Variation of Scores Across Months and Year



1. January and February has seen a decline in score as compared to the previous year.

2. This can be due to weather changes in Los Angeles during January and February in the year 2018.

3. The median of score in December for 2018 has seen an increase.

Variation of score in Top 5 Facilities over the years.

FACILITY ID	FACILITY NAME
FA0006427	LEVY PREMIUM FOODSERVICE LIMITED PARTNERSHIP
FA0019271	LEVY Premium foodservice ,LP
FA0065100	Legends Hospitality ,LLC
FA0170678	Magic Mountain ,LLC
FA0170909	Universal City Studios ,LLC


2016
100
95
90
Avg. Score
September
October
November
December
Month of Activity Date [2016]
Facility Id
FA0006427
FA0019271
FA0065100
FA0170678
FA0170909
2017
100
95
90
Avg. Score
January
February
March
April
May
June
July
August
September
October
November
December
Month of Activity Date [2017]
2018
100
95
90
85
80
Avg. Score
January
February
March
April
May
June
July
August
September
October
November
December
Month of Activity Date [2018]

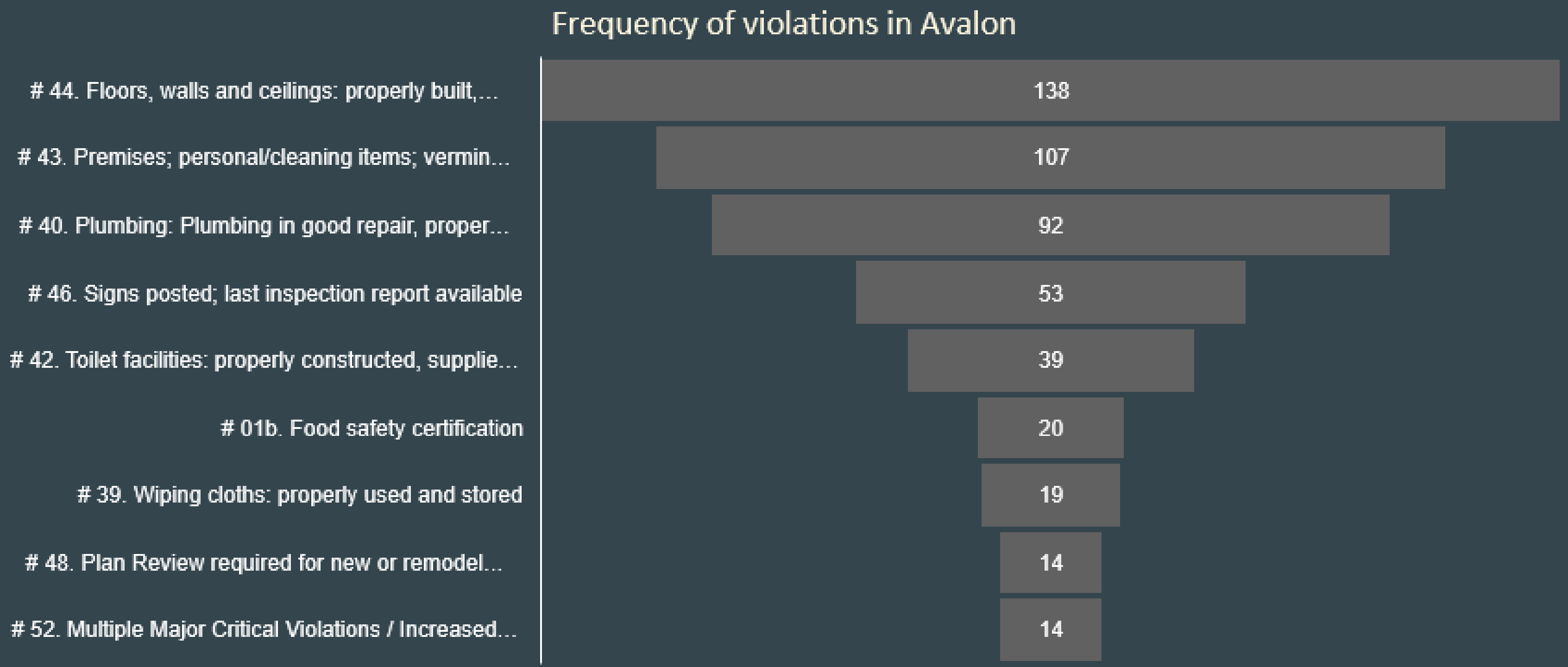
1.In the year 2016 Legends Hospitality ,LLC has seen a continuous increase in Score over the year and Levy Premium foodservice Limited Partnership has seen a decrease in the score.

2.Most of the Facilities see a dip of score in the month of June and July.

3.The scores have remained constant across the different facilities in the year 2018.

Avg Score across Different areas

Business Recommendations for Avalon



General Recommendations

- If your establishment receives a poor health inspection score, you can schedule a re-inspection in 5 - 45 days. This will give you time to correct the violations.
- Figure out how each violation occurred and how you can prevent it from happening again.
- Just like with your own self-inspection, review any violations and their proper corrective action with your staff.

Thank You