Final Presentation - 4381 - Restaurant Health Violations (SFO)

Thao Wells

	business_id	business_name	business_address	business_city	business_state	business_postal_code	business_latitude	business_longitude	business_location
0	85936	Laurel Court	950 Mason St	San Francisco	CA	94108	NaN	NaN	NaN
1	5827	HILLCREST ELEMENTARY SCHOOL	810 SILVER Ave	San Francisco	CA	94134	37.729016	-122.419253	POINT (-122.419253 37.729016)
2	94910	Ike's Kitchen	800 Van Ness Ave	San Francisco	CA	94109	NaN	NaN	NaN
3	64667	Jasmine Rae Bakery	1890 Bryant St #309	San Francisco	CA	94110	37.763156	-122.410351	POINT (-122.410351 37.763156)
4	97722	THE CHURRO FACTORY	PIER 39 K-01	San Francisco	CA	94133	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...
53968	70220	Trader Joe's #200	1095 Hyde St	San Francisco	CA	94109	NaN	NaN	NaN
53969	95021	Wing Wings	422 Haight St	San Francisco	CA	94117	NaN	NaN	NaN
53970	78289	Sam Jordans Bar	4004 03rd St	San Francisco	CA	94124	NaN	NaN	NaN
53971	100887	ASIA CHINESE FOOD	350 BAY ST.	San Francisco	CA	94133	NaN	NaN	NaN
53972	15120	Nordstrom Espresso Bar	865 Market Street	San Francisco	CA	94103	37.784317	-122.407563	POINT (-122.407563 37.784317)

business_phone_number	...	inspection_score	inspection_type	violation_id	violation_description	risk_category	Neighborhoods	SF Find Neighborhoods	Current Police Districts	Current Supervisor Districts	Analysis Neighborhoods
1.415578e+10	...	100.0	Routine - Unscheduled	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1.415546e+10	...	NaN	Reinspection/Followup	NaN	NaN	NaN	92.0	92.0	2.0	2.0	7.0
NaN	...	NaN	New Ownership - Followup	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	...	NaN	Reinspection/Followup	NaN	NaN	NaN	53.0	53.0	3.0	2.0	20.0
NaN	...	96.0	Routine - Unscheduled	97722_20181217_103154	Unclean or degraded floors walls or ceilings	Low Risk	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...
1.415530e+10	...	NaN	Complaint	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1.415584e+10	...	92.0	Routine - Unscheduled	95021_20190228_103119	Inadequate and inaccessible handwashing facili...	Moderate Risk	NaN	NaN	NaN	NaN	NaN
NaN	...	NaN	Reinspection/Followup	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1.415582e+10	...	NaN	New Ownership - Followup	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	...	NaN	Non-inspection site visit	NaN	NaN	NaN	32.0	32.0	5.0	10.0	34.0

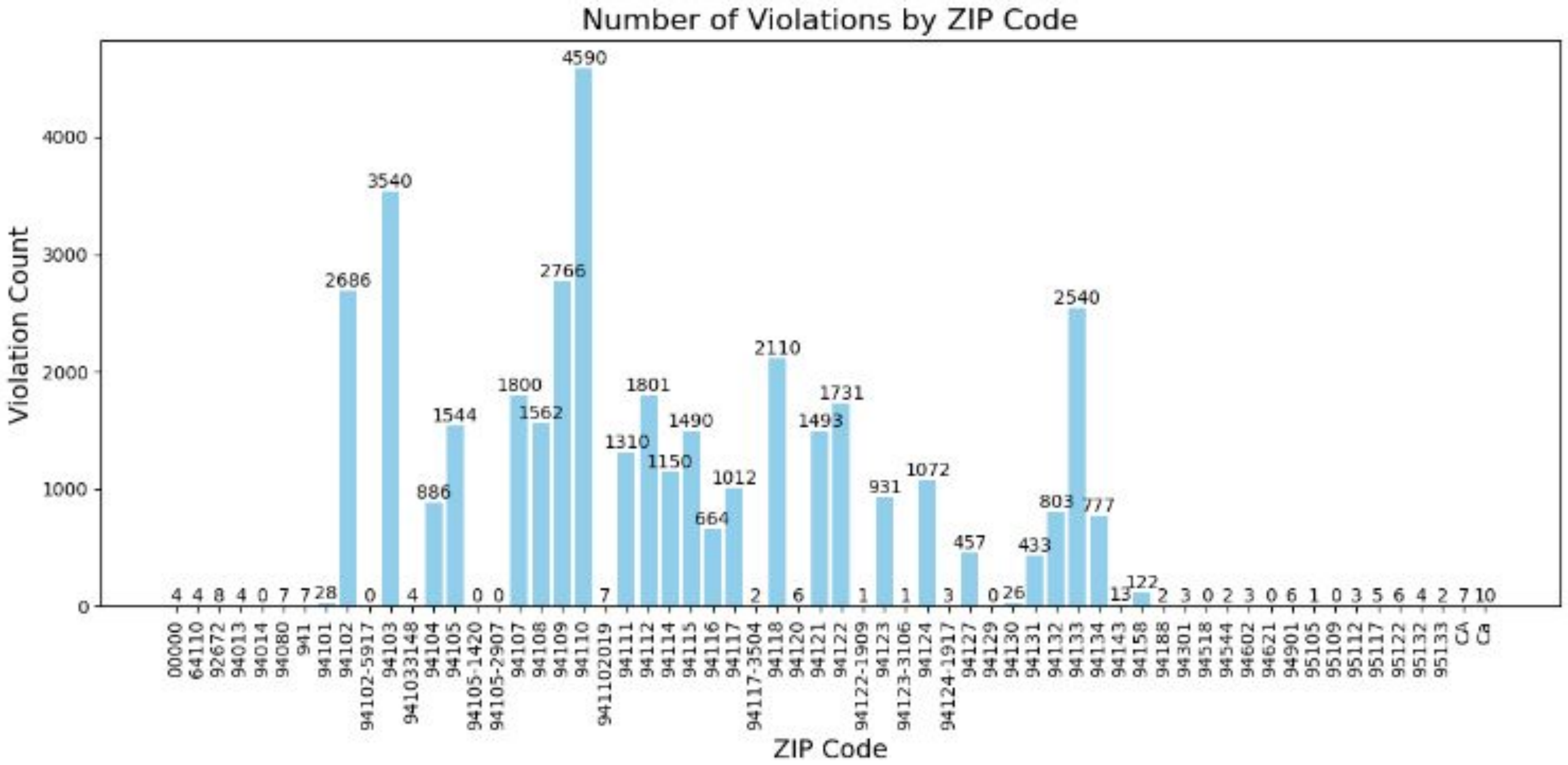
Models:

- Random Forest (was working, needs to be redone)
- Gradient Boosting Machines (was working, needs to be redone)
- Support Vector Machines (SVM) (unfinished)
- Baseline: Logistic Regression (done, will be redone when secondary datasets are added)

Evaluation Metrics:

- Accuracy, Precision, Recall, F1-Score, ROC-AUC.
- Special focus on precision/recall for high-risk violations.

Analysis of Violations and Highway Proximity

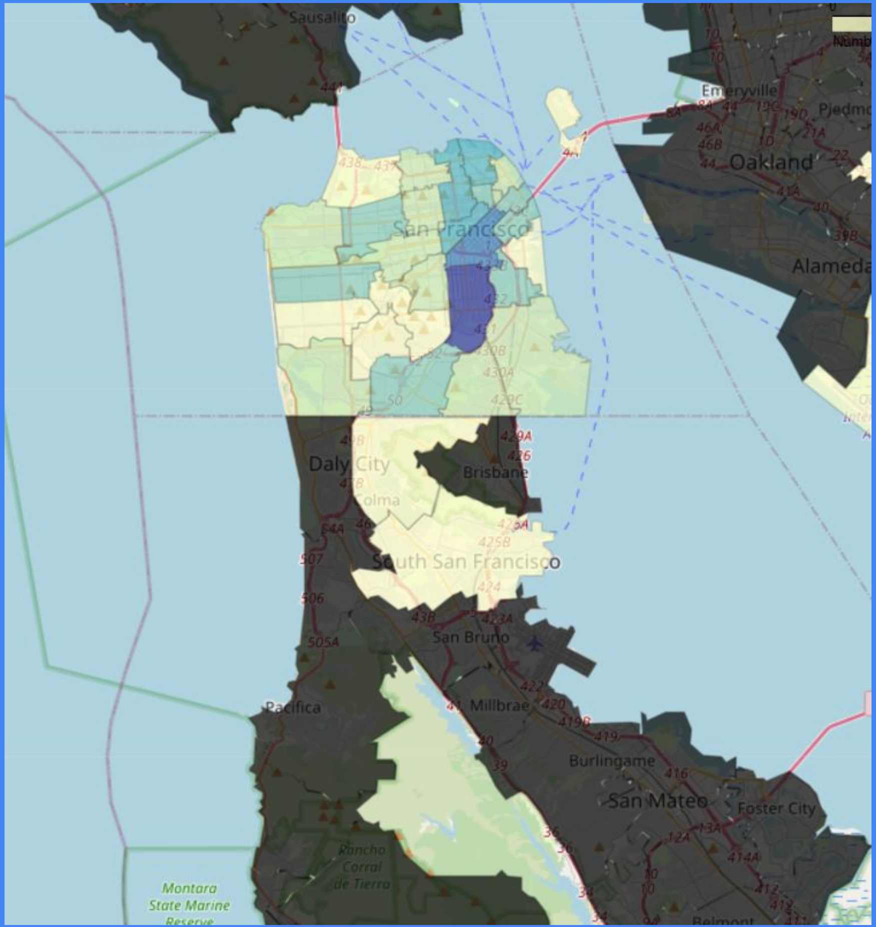


Number of Violations by ZIP Code
Violation Count
ZIP Code
0
1000
2000
3000
4000
00000 64110 92672 94013 94014 94080 941 94101 94102 94102-5917 94103 941033148 94104 94105 94105-1420 94105-2907 94107 94108 94109 94110 941102019 94111 94112 94114 94115 94116 94117 94117-3504 94118 94120 94121 94122 94122-1909 94123 94123-3106 94124 94124-1917 94127 94129 94130 94131 94132 94133 94134 94143 94158 94188 94301 94518 94544 94602 94621 94901 95105 95109 95112 95117 95122 95132 95133 CA Ca
4 4 8 4 0 7 7 28 2686 0 3540 4 886 1544 0 0 1800 1562 2766 4590 7 1310 1801 1150 1490 664 1012 2 2110 6 1493 1731 1 931 1 1072 3 457 0 26 433 803 2540 777 13 122 2 3 0 2 3 0 6 1 0 3 5 6 4 2 7 10

Number of Violations by Type and ZIP Code:

violation_description	00000	64110	92672	94013	94080	941	94101	94102	94103	941033148	94104	94105	94107	94108	94109	9411
Consumer advisory not provided for raw or undercooked foods	0	0	0	0	0	0	0	4	1	0	0	0	1	1	1	0
Contaminated or adulterated food	0	0	0	0	0	0	0	11	11	0	0	3	6	8	14	20
Discharge from employee nose mouth or eye	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1
Employee eating or smoking	0	0	0	0	0	0	0	10	20	0	4	2	7	4	7	53
Food in poor condition	0	0	0	0	0	0	0	8	9	0	1	2	6	1	3	7
Food safety certificate or food handler card not available	0	1	0	0	0	1	2	80	100	0	21	20	51	47	97	123
Foods not protected from contamination	0	0	1	0	0	0	0	129	184	0	40	87	103	85	160	157
High risk food holding temperature	0	1	0	0	0	0	2	116	171	0	52	89	61	72	119	182
High risk vermin infestation	0	0	0	0	0	0	0	50	88	0	9	14	31	49	88	107
Improper cooking time or temperatures	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0
Improper cooling methods	0	0	0	0	0	0	0	67	104	0	10	98	43	36	43	148
Improper food labeling or menu misrepresentation	0	0	0	0	0	0	0	12	14	0	7	6	6	5	20	24
Improper food storage	0	0	1	0	0	1	1	129	124	0	29	35	54	88	115	170
Improper or defective plumbing	0	0	0	0	0	0	1	54	64	0	35	43	23	44	57	69
Improper reheating of food	0	0	0	0	0	0	0	11	15	0	7	9	7	3	6	13
Improper storage of equipment utensils or linens	0	0	0	1	0	1	1	66	66	0	28	22	30	44	64	108
Improper storage use or identification of toxic substances	0	0	0	0	0	1	1	40	26	0	15	10	17	14	33	26
Improper thawing methods	0	0	0	0	1	0	1	44	48	0	9	9	30	36	45	57
Improperly displayed mobile food permit or signage	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
Improperly washed fruits and vegetables	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	2
Inadequate HACCP plan record keeping	0	0	0	0	0	0	0	4	3	0	1	2	3	1	0	3
Inadequate and inaccessible handwashing facilities	2	0	1	0	0	0	1	194	242	1	38	128	171	102	174	254
Inadequate dressing rooms or improper storage of personal items	0	0	0	0	0	0	0	13	43	0	6	31	22	6	13	44
Inadequate food safety knowledge or lack of certified food safety manager	0	0	0	0	0	0	2	54	86	1	42	54	49	29	49	162
Inadequate or unsanitary refuse containers or area or no garbage service	0	0	0	0	0	0	0	10	3	0	0	2	2	4	8	20
Inadequate procedures or records for time as a public health control	0	0	0	0	0	0	0	7	31	0	3	10	11	3	7	15
Inadequate sewage or wastewater disposal	0	0	0	0	0	0	0	5	13	0	1	8	1	2	8	4
Inadequate ventilation or lighting	0	0	0	0	0	0	0	27	15	0	6	11	16	18	33	32
Inadequate warewashing facilities or equipment	0	0	0	0	0	0	2	34	45	0	33	11	9	17	33	16
Inadequately cleaned or sanitized food contact surfaces	1	1	0	1	1	0	1	181	310	1	81	137	142	123	154	303
Insufficient hot water or running water	0	0	0	0	0	0	2	59	48	0	49	37	35	43	78	51
Low risk vermin infestation	0	0	0	0	0	0	0	82	82	0	15	34	40	44	102	175
Mobile food facility not operating with an approved commissary	0	0	0	0	0	0	0	0	10	0	0	0	0	0	2	0
Mobile food facility stored in unapproved location	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mobile food facility with unapproved operating conditions	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0
Moderate risk food holding temperature	0	0	1	1	2	0	1	163	231	0	59	109	173	96	151	292


Violation Types Distribution per ZIP Code
0: Inadequate food safety knowledge or lack of certified food safety manager
0: Improper food storage
0: Foods not protected from contamination
0: Improper cooking time or temperatures
0: Improper food labeling or menu misrepresentation
0: Food safety certificate or food handler card not available
0: Improper cooling methods
0: Improper thawing methods
0: Unclean or unsanitary food contact surfaces
0: Improper reheating of food
0: Contaminated or adulterated food
1: Unclean hands or improper use of gloves
1: Unclean nonfood contact surfaces
1: Wiping cloths not clean or properly stored or inadequate sanitizer
1: Inadequately cleaned or sanitized food contact surfaces
1: Unclean or degraded floors walls or ceilings
1: Unclean or unsanitary refuse containers or area or no garbage service
1: Inadequate warewashing facilities or equipment
2: High risk vermin infestation
2: Moderate risk vermin infestation
2: Low risk vermin infestation
3: Improper or defective plumbing
3: Unapproved or unmaintained equipment or utensils
3: No thermometers or uncalibrated thermometers
3: Mobile food facility with unapproved operating conditions
3: Improper storage of equipment utensils or linens
4: Inadequate and inaccessible handwashing facilities
4: Employee eating or smoking
4: Discharge from employee nose mouth or eye
4: Sewage or wastewater contamination
4: Non service animal
4: Worker safety hazards
5: Permit license or inspection report not posted
5: Unpermitted food facility
5: No plan review or Building Permit
5: Noncompliance with shell fish tags or display
5: Noncompliance with HAACP plan or variance
00000
1.0
50.0%
50.0%
4.0
64110
0.0
33.3%
66.7%
1.0
92672
4.0
14.3%
0.0
57.1%
28.6%
1.0
94013
94080
941


Sausalito
Emeryville
Oakland
Alameda
San Francisco
Daly City
Brisbane
Colma
South San Francisco
San Bruno
Pacifica
Millbrae
Burlingame
San Mateo
Foster City
Belmont
Rancho Corral de Tierra
Montara State Marine Reserve

Exploring Correlations Between Business Locations, Violations, and Highway Distance

Investigate the relationship between business location, violation codes, and proximity to highways.

Examine whether being closer to highways affects the frequency and type of violations.

*Initially, I was working on models that were just based off of a geographical data set and the restaurant dataset, but I wanted to analyse it compared against proximity to a highway (implying traffic)

Overview

Data Sources:

- Violation data (Business postal codes, violation codes, and counts of violations).
- Geographic data (Highway shapefiles for the San Francisco area).

Key Variables:

- `business_postal_code`: Location of businesses.
- `violation_code`: Type of violation (e.g., parking violations).
- `Count`: Number of violations.
- `distance_to_highway`: Proximity of each business to the nearest highway.

Issues

Problem: Unable to parse the HTML file (violations_by_zip.html) containing the relevant data.

Attempted Solutions:

Tried reading the file using pd.read_html(), but no tables were found in the HTML.

Tried extracting JSON data from the HTML using BeautifulSoup, but the data format was not as expected.

Solution: Reached out to various methods, including examining the raw HTML content and adjusting the extraction code, but failed to get usable data directly from the HTML file.

Issues

GeoJSON File for Map Visualization:

Problem: GeoJSON data used for Choropleth mapping needed to match ZIP codes in violation data.

Issue with handling mismatched formats.

Solution: Successfully created Choropleth maps after resolving format issues, and a separate readable file as the html file did not have tables

Issues

Issues with Correlation Matrix:

Data Preprocessing Challenges:

Issue: Some columns contained non-numeric data, making it impossible to compute a correlation matrix directly.

Solution: Dropped non-numeric columns and encoded categorical features using One-Hot Encoding.

Key Takeaways

Importance of data format consistency (JSON, CSV, GeoJSON).

Careful preprocessing is essential before performing statistical analyses like PCA and correlation matrices.

Handling missing values and non-numeric data is crucial for generating meaningful insights.

preprocessing

- Cleaned and merged violation data with geographic coordinates.
- Used geographic shapefiles for highway data.
- Calculated the distance of each business from the nearest highway using GeoPandas.
- Re-projected spatial data to ensure matching CRS (Coordinate Reference System).

	business_postal_code	violation_code	Count	distance_to_highway
business_postal_code	1.000000	-0.024614	0.737425	-0.976476
violation_code	-0.024614	1.000000	-0.693375	0.239595
Count	0.737425	-0.693375	1.000000	-0.865718
distance_to_highway	-0.976476	0.239595	-0.865718	1.000000

Correlation Analysis

Key Findings from Correlation Matrix:

- **Business postal code vs. distance to highway**: Strong negative correlation (-0.98), indicating that businesses further from city centers tend to be closer to highways.
- **Violation code vs. distance to highway**: Weak positive correlation (0.24), suggesting some types of violations are slightly more frequent near highways.
- **Count vs. distance to highway**: Strong negative correlation (-0.87), suggesting that higher numbers of violations tend to occur near highways.

Proximity to highways plays a significant role in the occurrence of violations, with more violations occurring closer to highways.
Certain violation types may be more associated with locations near highways.

Further Analysis

Further Analysis:

- Explore more advanced modeling techniques (e.g., regression analysis or spatial clustering) to predict violations based on proximity to highways.
- Investigate the spatial distribution of violations to identify specific patterns and trends.

Potential Applications:

- Urban planning and traffic management.
- Policy development for improved enforcement and safety measures.