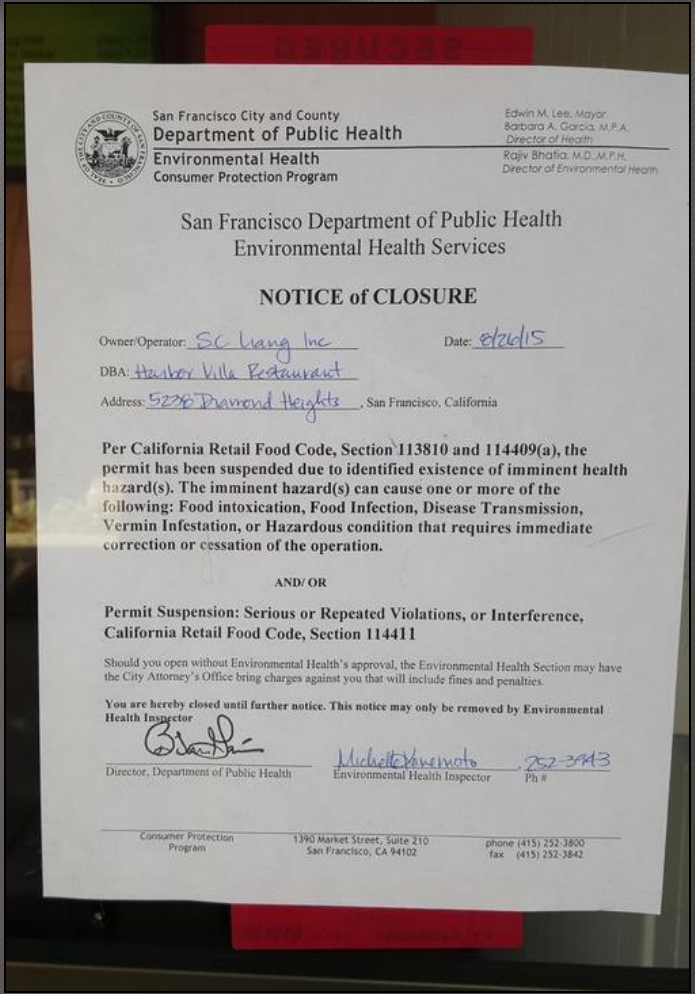
SFDPH + yelp

Predicting San Francisco restaurant safety scores from Yelp data

Background

- The Food Safety Program enforces health code regulations, which may result in administrative actions and suspension or revocation of the Permit to Operate when violations are identified.
- To enforce the health code regulations, Environmental Health Inspectors inspect over 7,000 locations in San Francisco that serves food to the public including restaurants, bars, markets, bakeries, pushcarts, and stadium food.
  - Inspection score > 80: requires 2 routine inspections per year
  - Inspection score < 80: requires 3 routine inspections per year

- Recent articles have pointed out that San Francisco, as well as other cities, have shortages of health (food/restaurant) inspectors.
- Many establishments have not received a single follow-up inspection in over 2 years.



San Francisco City and County
Department of Public Health
Environmental Health
Consumer Protection Program

Edwin M. Lee, Mayor
Barbara A. Garcia, M.P.A.
Director of Health
Rajiv Bhatia, M.D., M.P.H.
Director of Environmental Health

San Francisco Department of Public Health
Environmental Health Services

NOTICE of CLOSURE

Owner/Operator: SC Liang Inc Date: 8/26/15

DBA: Harbor Villa Restaurant

Address: 5238 Diamond Heights, San Francisco, California

Per California Retail Food Code, Section 113810 and 114409(a), the permit has been suspended due to identified existence of imminent health hazard(s). The imminent hazard(s) can cause one or more of the following: Food intoxication, Food Infection, Disease Transmission, Vermin Infestation, or Hazardous condition that requires immediate correction or cessation of the operation.

AND/ OR

Permit Suspension: Serious or Repeated Violations, or Interference, California Retail Food Code, Section 114411

Should you open without Environmental Health's approval, the Environmental Health Section may have the City Attorney's Office bring charges against you that will include fines and penalties.

You are hereby closed until further notice. This notice may only be removed by Environmental Health Inspector

Director, Department of Public Health

Michelle Kanemoto
Environmental Health Inspector

252-3943
Ph #

Consumer Protection
Program

1390 Market Street, Suite 210
San Francisco, CA 94102

phone (415) 252-3800
fax (415) 252-3842

Objective

- Determine if SF health department food inspection scores can be predicted based on Yelp data.

- This could allow better resource allocation for health inspectors.

Data Sources

1. SF Public Health Department (SFPHD): Inspection data
   - https://extxfer.sfdph.org/food/SFBusinesses.zip [last accessed 8/23/2016]
   - Businesses.csv: list of business ids (primary key), name, location, phone
   - Inspections.csv: score, date

SF HEALTH NETWORK
SAN FRANCISCO DEPARTMENT OF PUBLIC HEALTH

2. Yelp
   - Used Yelp API's to 'manually' assemble the Yelp dataset
   - Two different versions of the API had to be run
     - Version 2 required to return the 'neighborhood' feature
     - Version 3 (in beta) required to return the 'price' feature
   - At least 2 different implementations of each had to be used to collect all the data (phone & location via latitude/longitude) making for 2x2 = 4 calls over all the businesses

yelp

Process

- Data collection
  - Obtained SFDPH inspection data set - straightforward
  - Created a database linking SFDPH data to generate list of query criteria for Yelp APIs
  - Programmed Yelp API queries
  - Two different versions of the API had to be run to obtain desired features
    - V2 required to return the ‘neighborhood’ feature
    - V3 (in beta) required to return the ‘price’ feature
  - At least 2 different implementations of each version had to be used to collect all the data (phone & location) making for 2x2 = 4x calls per business (~4k+) = 16k+ calls
  - Yelp JSON data parsed to Pandas DataFrames
  - Yelp data added to database

Process

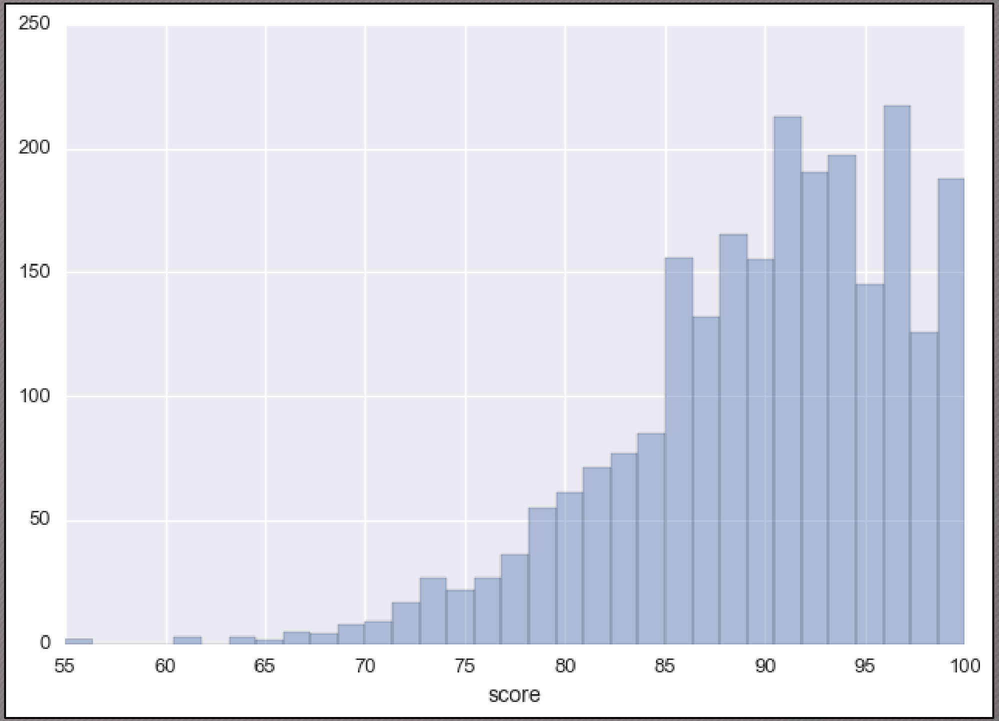
- Cleaning
  - Businesses with less than 10 Yelp reviews were dropped – Yelp data likely not meaningful
  - Businesses operating at AT&T Park (SF Giants) were dropped – too difficult to find via Yelp APIs
  - Businesses with SFDPH name and Yelp API return name that do not match were dropped
  - Initial # of rows 3418 -> cleaning -> 2997 (88% remained)
- Data split into training and testing sets
- Exploratory data analysis
- Model development

Data

- Inspection score
  - 0-100, integer
- Rating (Yelp)
  - Categorical, 1-5 in 0.5 steps
- Counts
  - # Ratings, integer # of Yelp ratings
  - # Inspections, integer # of health inspections
- Chain
  - Binary, 0 not a chain, 1 chain
  - Calculated - defined as having more than 1 duplicate business name
- Price
  - Categorical, 1 to 4 ($ to $$$$)
- Category, type of establishment
  - Categorical, condensed from a list of 152 down to 25
- Neighborhood, location of establishment
  - Categorical, condensed from a list of 63 down to 21

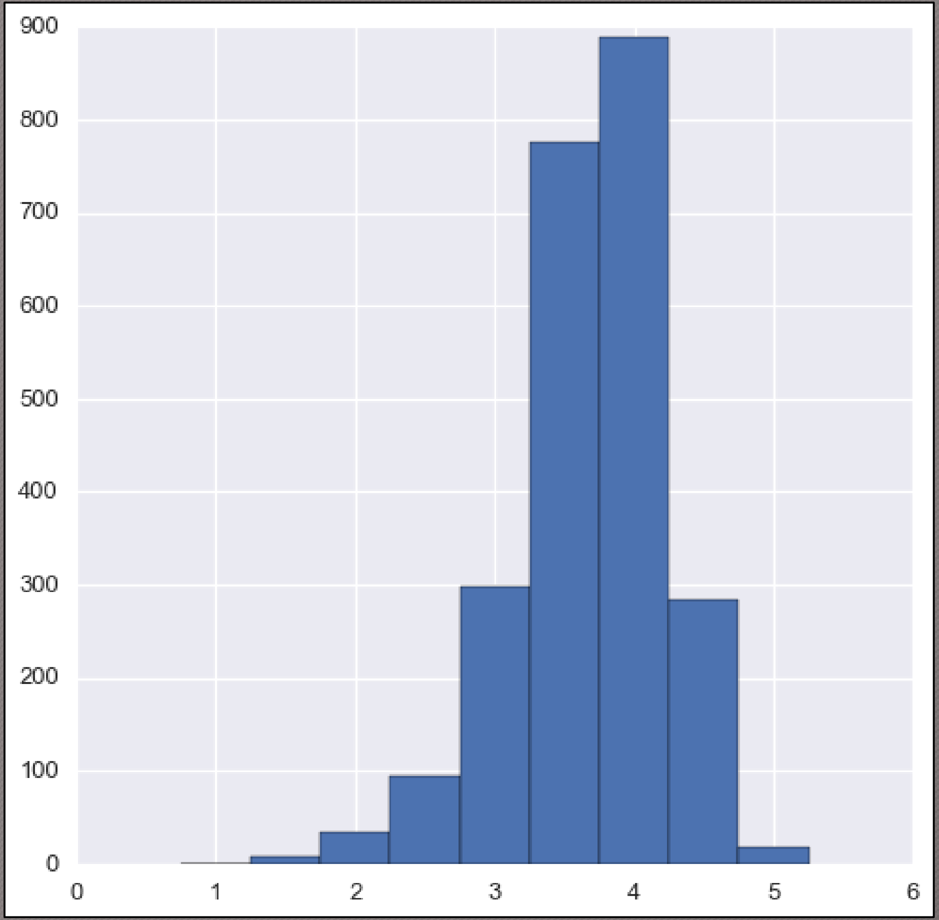
- Totals
  - # of features: 60
  - # of samples (after cleaning): 2998

Inspection Score

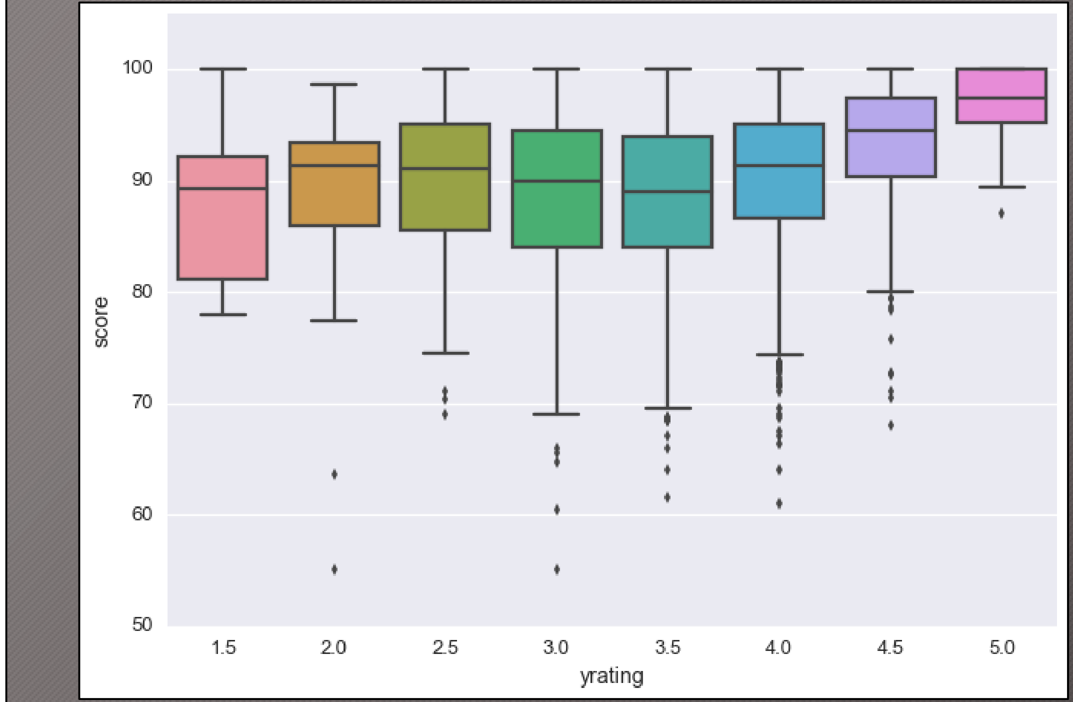


Count	2997
Mean	90
Std	7.1
Min	55
25%	86
50%	91
75%	95
Max	100

Yelp Rating

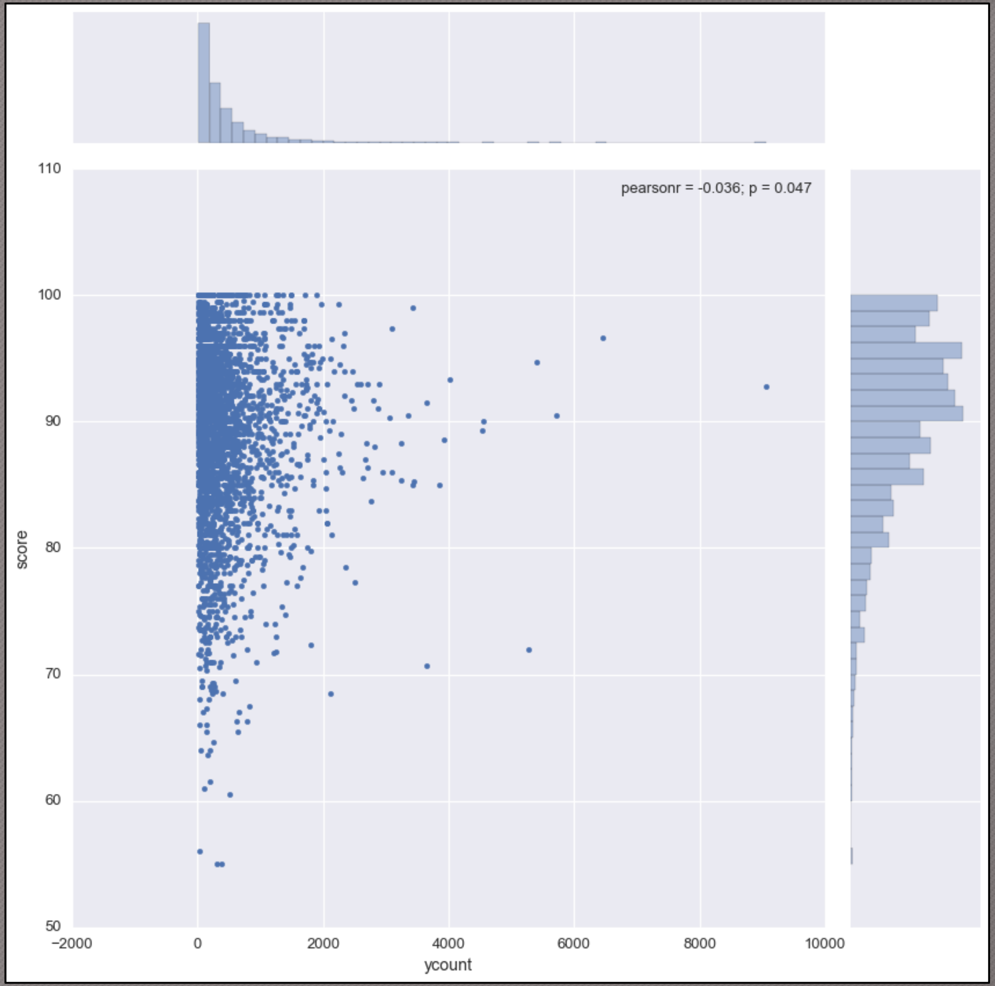


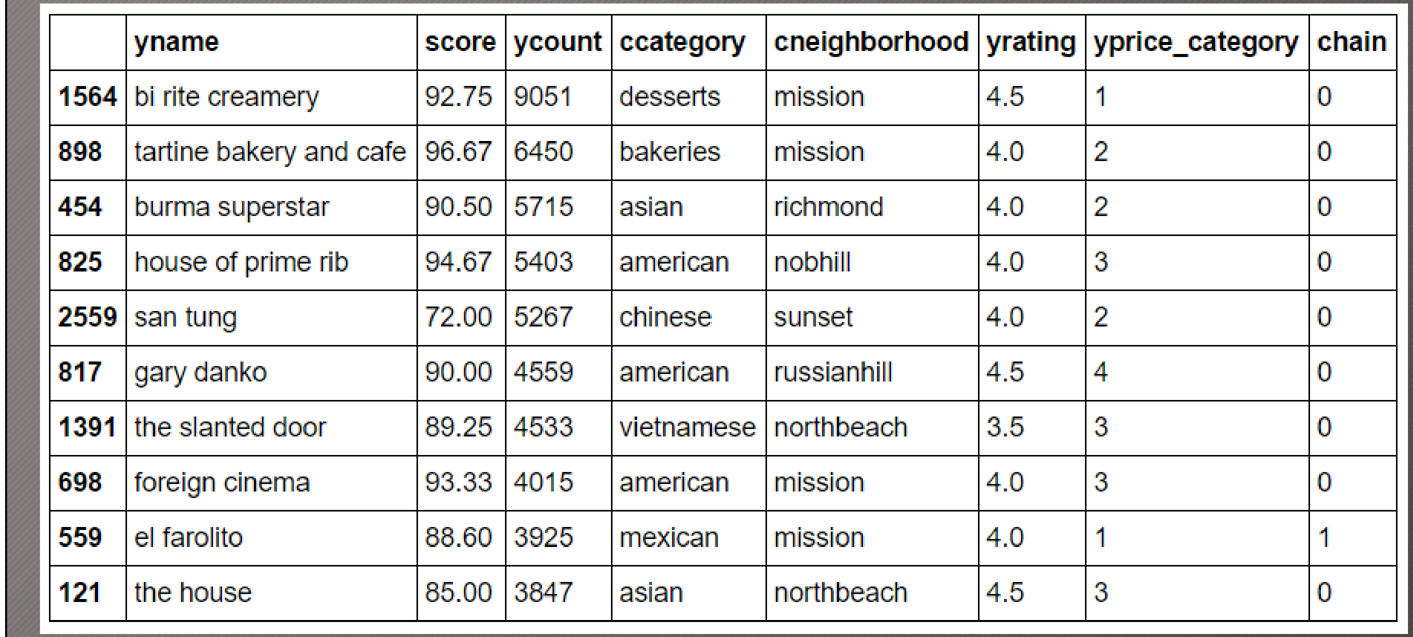
900
800
700
600
500
400
300
200
100
0
0
1
2
3
4
5
6



score
100
90
80
70
60
50
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
yrating

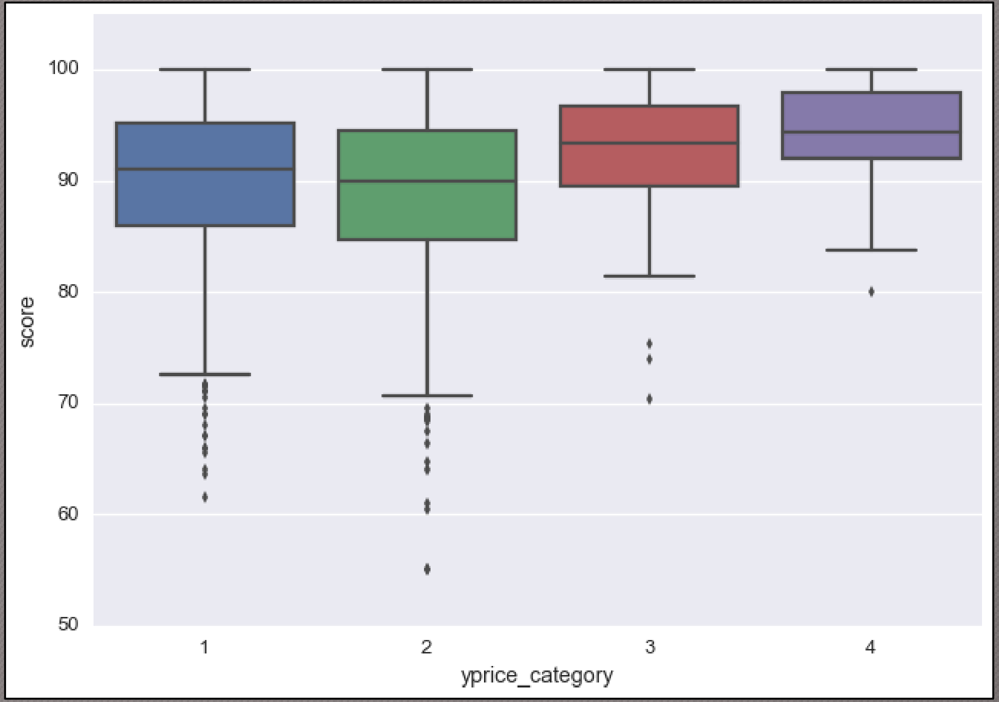
# of Yelp Reviews

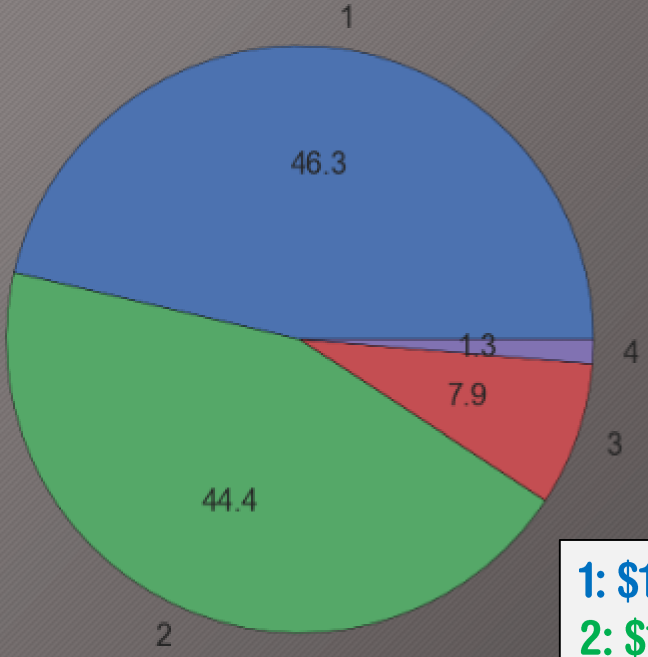




	yname	score	ycount	ccategory	cneighborhood	yrating	yprice_category	chain
1564	bi rite creamery	92.75	9051	desserts	mission	4.5	1	0
898	tartine bakery and cafe	96.67	6450	bakeries	mission	4.0	2	0
454	burma superstar	90.50	5715	asian	richmond	4.0	2	0
825	house of prime rib	94.67	5403	american	nobhill	4.0	3	0
2559	san tung	72.00	5267	chinese	sunset	4.0	2	0
817	gary danko	90.00	4559	american	russianhill	4.5	4	0
1391	the slanted door	89.25	4533	vietnamese	northbeach	3.5	3	0
698	foreign cinema	93.33	4015	american	mission	4.0	3	0
559	el farolito	88.60	3925	mexican	mission	4.0	1	1
121	the house	85.00	3847	asian	northbeach	4.5	3	0

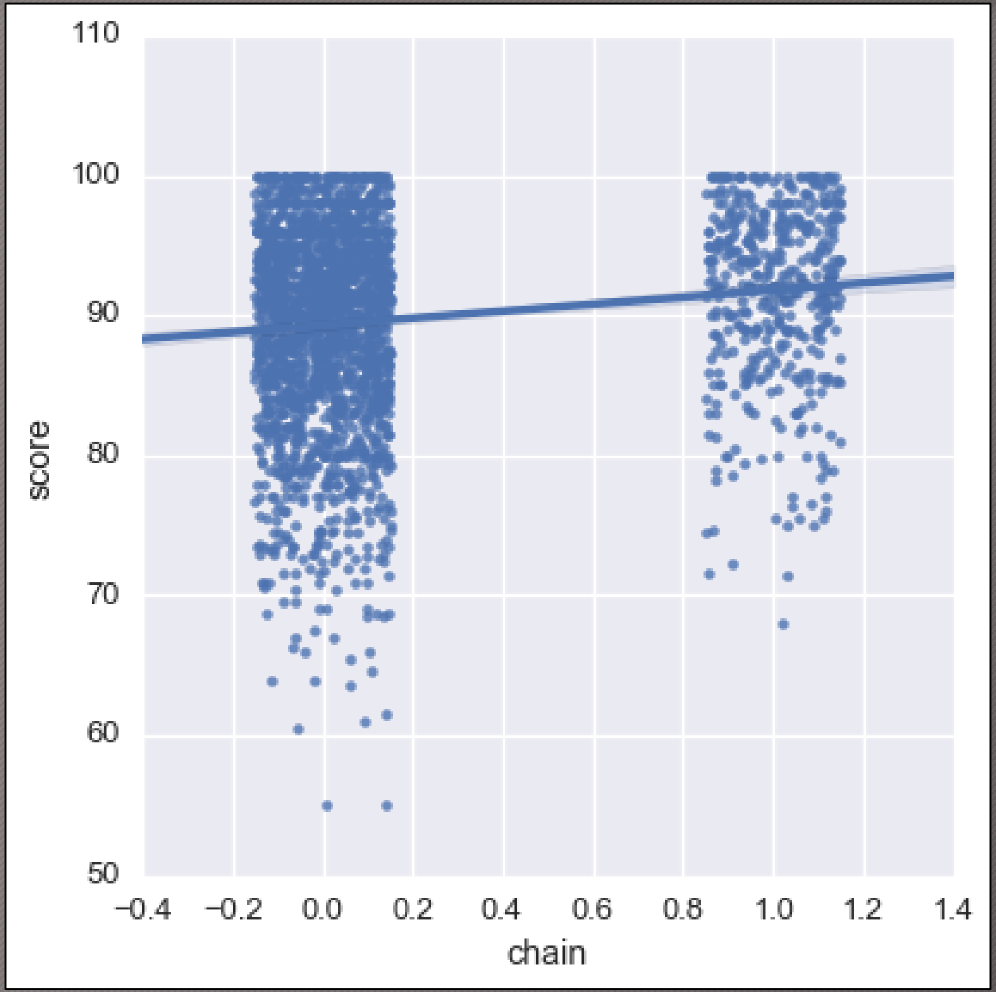
Price





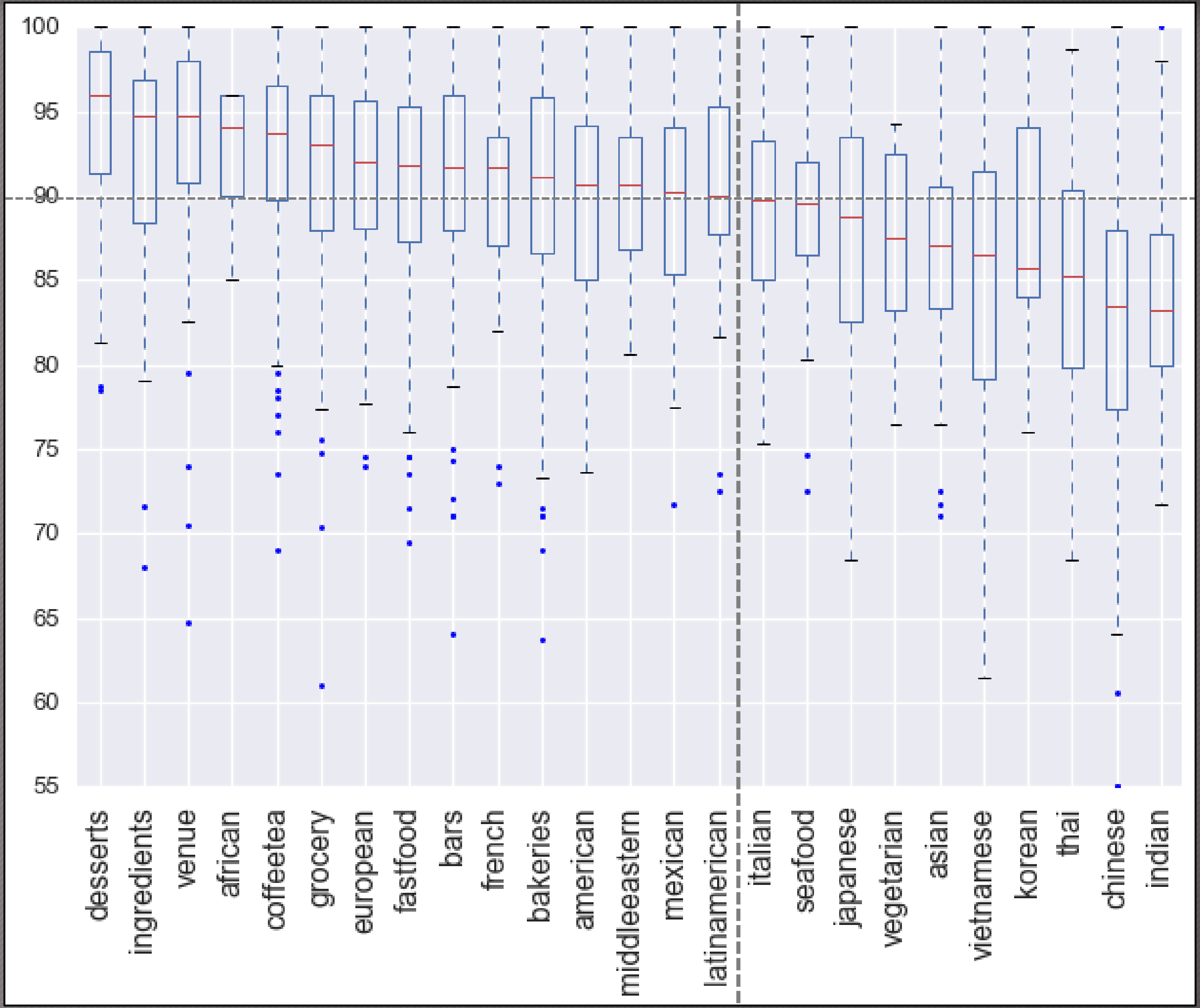
1: $10 or less
2: $11-$30
3: $31-$60
4: $61+

Chain



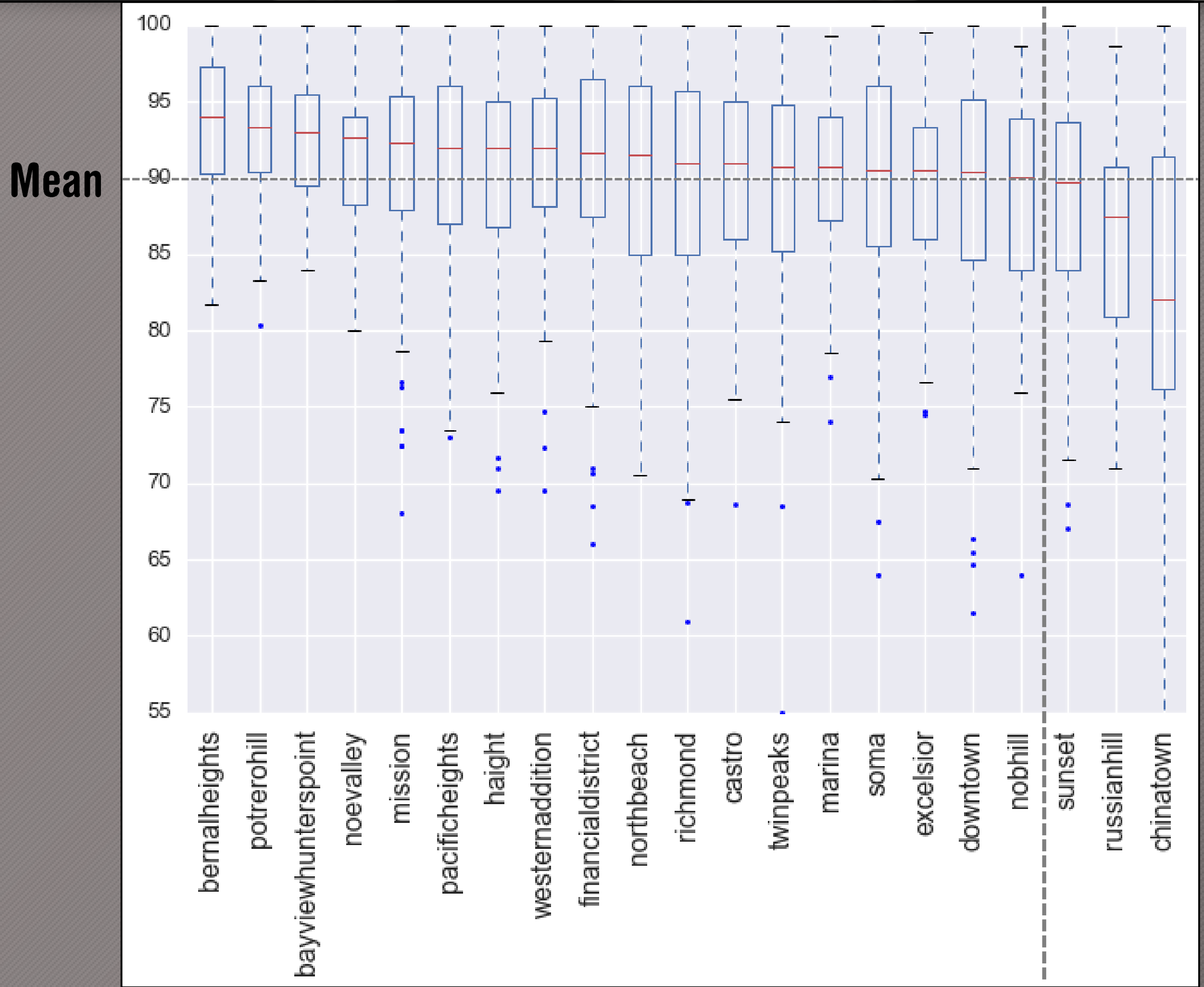
Category

Mean

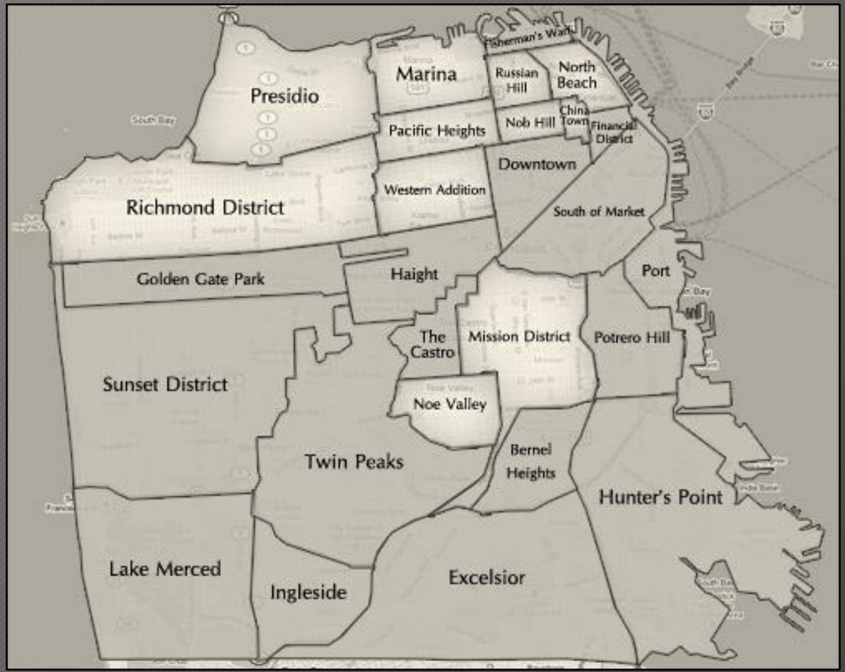


fastfood	450
coffeetea	369
chinese	221
american	203
bars	192
grocery	171
bakeries	168
venue	162
mexican	149
japanese	148
italian	111
thai	84
desserts	77
vietnamese	76
european	71
seafood	52
ingredients	45
latinamerican	43
indian	40
asian	40
french	38
middleeastern	34
korean	29
vegetarian	14
african	10

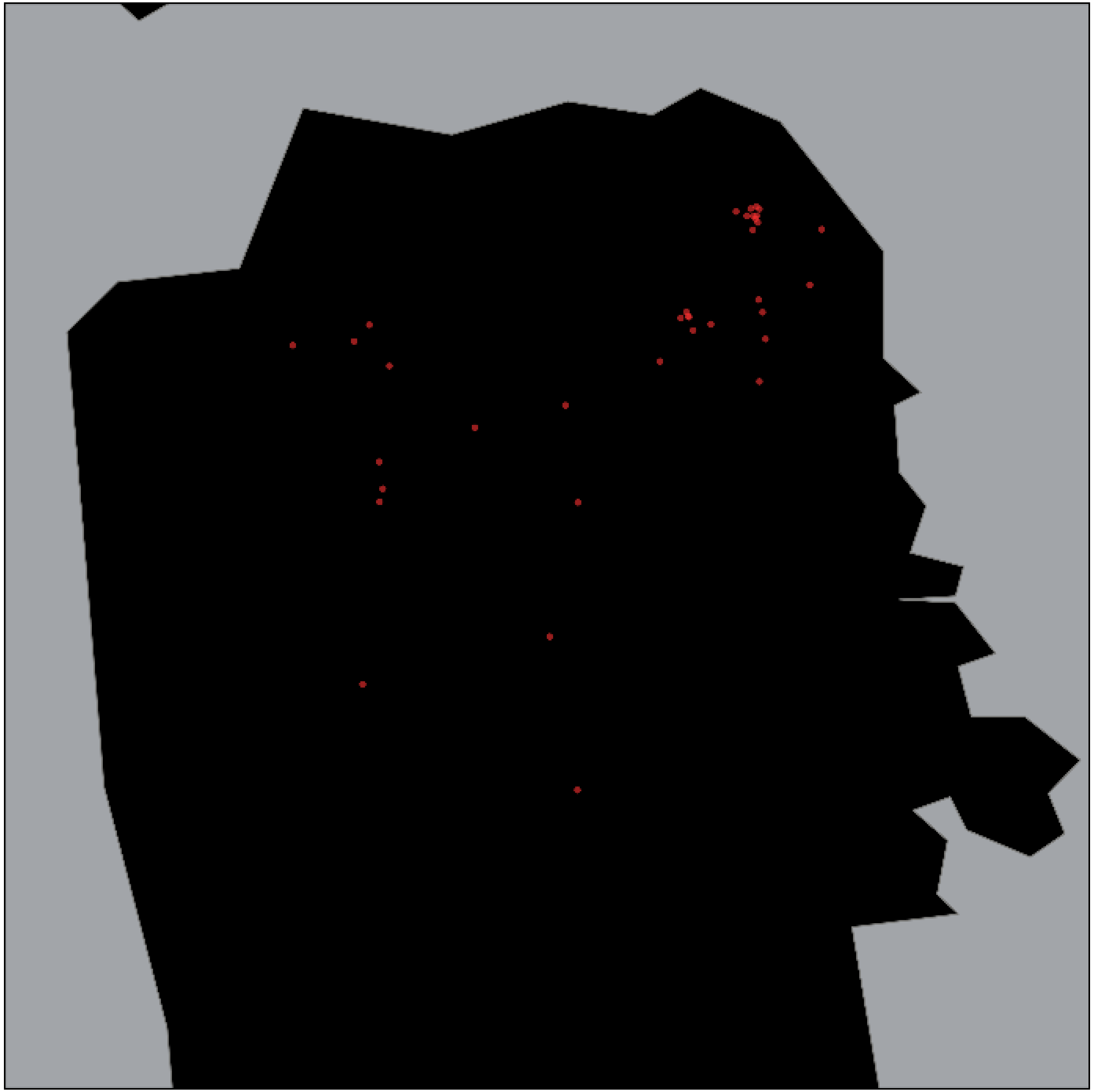
Neighborhood



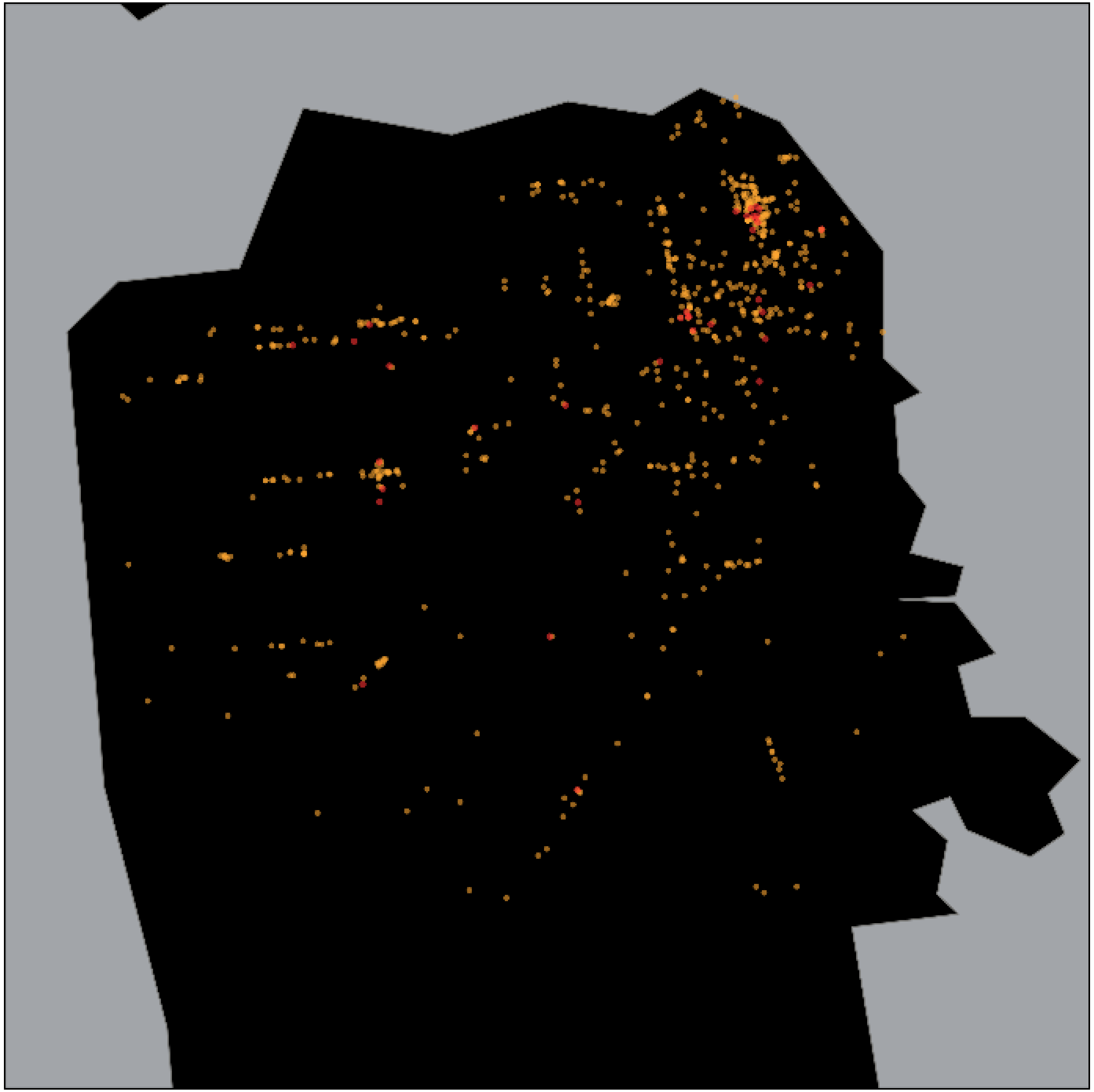
mission	356
financialdistrict	305
downtown	257
sunset	253
richmond	229
soma	192
northbeach	172
pacificheights	152
nobhill	142
marina	111
haight	108
russianhill	105
chinatown	98
excelsior	94
castro	88
twinpeaks	76
westernaddition	63
bernalheights	63
noevalley	51
bayviewhunterspoint	43
potrerohill	39



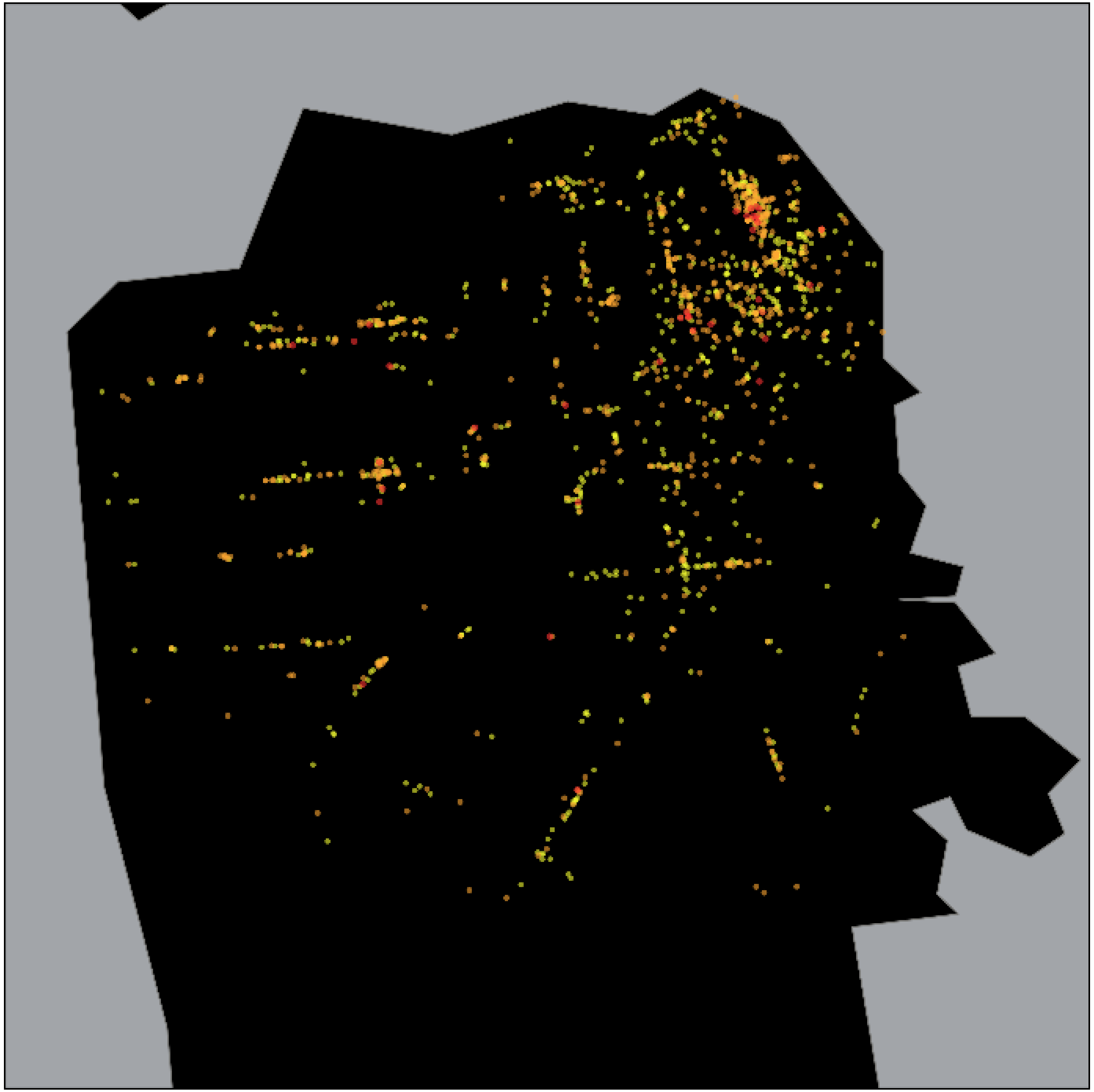
Score	Operating Condition Category	Inspection Findings
>90	Good	• Typically, only lower-risk health and safety violations observed • May have high-risk violations
86-90	Adequate	• Several violations observed • May have high-risk violations
71-85	Needs Improvement	• Multiple violations observed • Typically, several high-risk violations
Less than or equal to 70	Poor	• Multiple violations observed • Typically, several high-risk violations



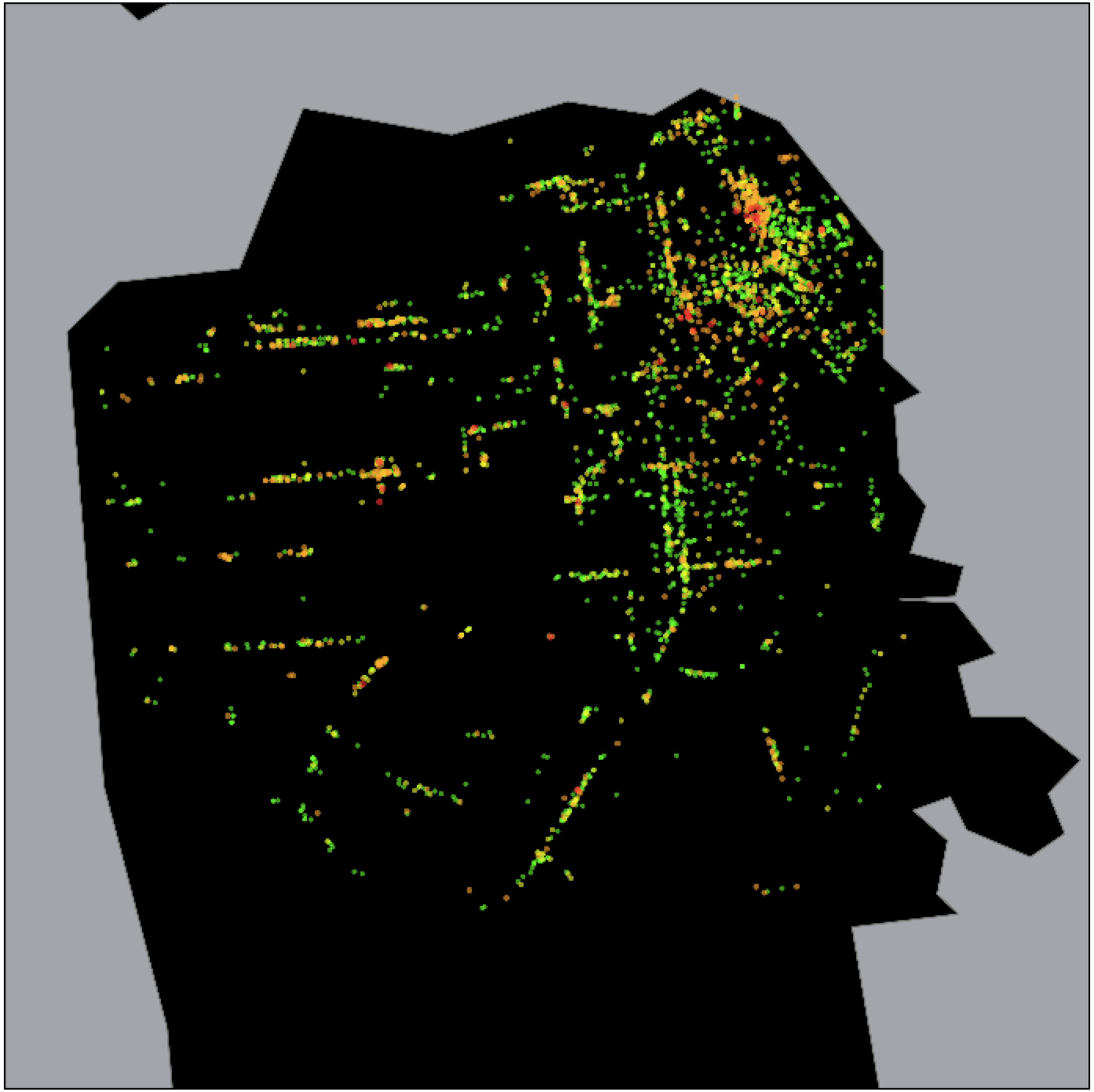
Score	Operating Condition Category	Inspection Findings
>90	Good	- Typically, only lower-risk health and safety violations observed - May have high-risk violations
86-90	Adequate	- Several violations observed - May have high-risk violations
71-85	Needs Improvement	- Multiple violations observed - Typically, several high-risk violations
Less than or equal to 70	Poor	- Multiple violations observed - Typically, several high-risk violations



Score	Operating Condition Category	Inspection Findings
>90	Good	- Typically, only lower-risk health and safety violations observed - May have high-risk violations
86-90	Adequate	- Several violations observed - May have high-risk violations
71-85	Needs Improvement	- Multiple violations observed - Typically, several high-risk violations
Less than or equal to 70	Poor	- Multiple violations observed - Typically, several high-risk violations

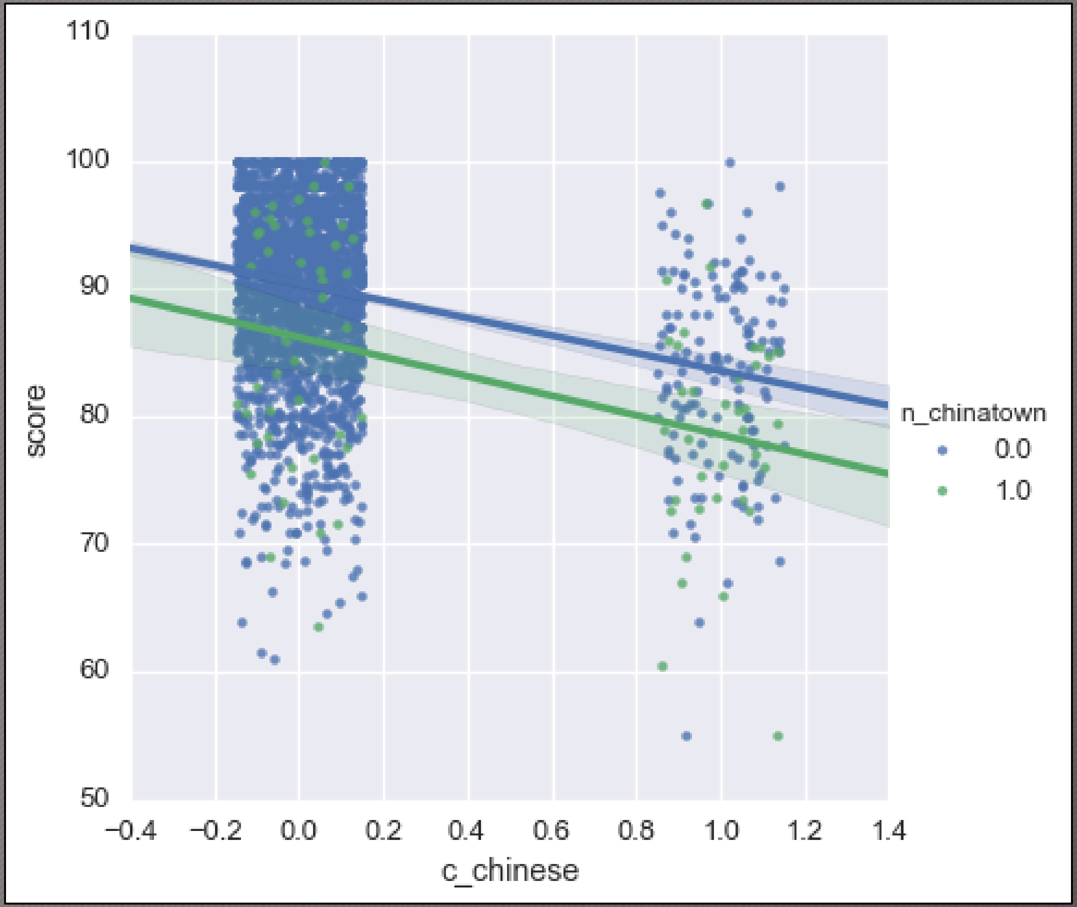


Score	Operating Condition Category	Inspection Findings
>90	Good	• Typically, only lower-risk health and safety violations observed • May have high-risk violations
86-90	Adequate	• Several violations observed • May have high-risk violations
71-85	Needs Improvement	• Multiple violations observed • Typically, several high-risk violations
Less than or equal to 70	Poor	• Multiple violations observed • Typically, several high-risk violations



Score	Operating Condition Category	Inspection Findings
>90	Good	• Typically, only lower-risk health and safety violations observed • May have high-risk violations
86-90	Adequate	• Several violations observed • May have high-risk violations
71-85	Needs Improvement	• Multiple violations observed • Typically, several high-risk violations
Less than or equal to 70	Poor	• Multiple violations observed • Typically, several high-risk violations

Combined Effects

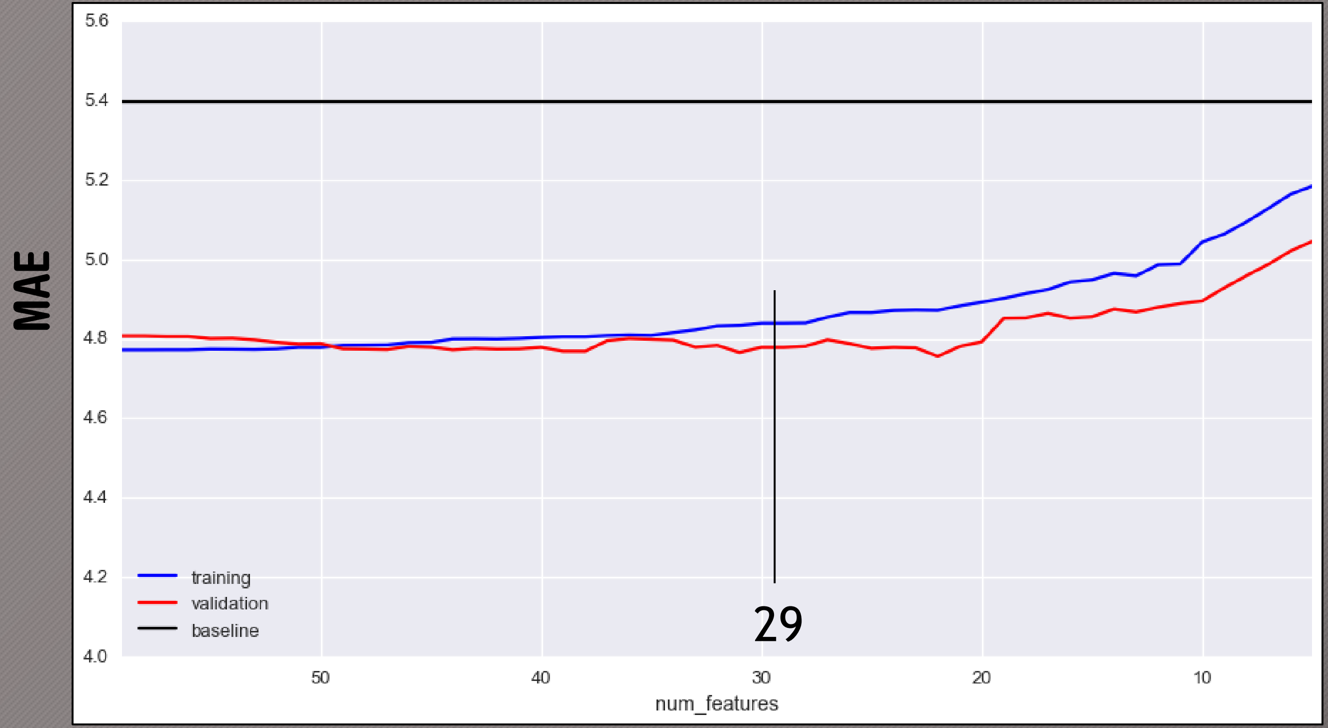


Modeling Approach

- Optimizing for Mean Absolute Error
  - The average of the absolute errors, where f_i is the predicted value (yhat) and y_i is the actual value.
- Data split
  - 80% training/validation, 20% test
  - Of the training set: 70% model training, 30% model validation
- 60 features -> All-in approach
  - All features fit
  - Ranked
  - Lowest contributor tossed out
  - Re-fit
  - Repeat until all features are exhausted or the training MAE is > the baseline MAE
- The feature set to use is where the test/training MAE starts to significantly deviate

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|f_i - y_i\right| = \frac{1}{n}\sum_{i=1}^{n}\left|e_i\right|.$$

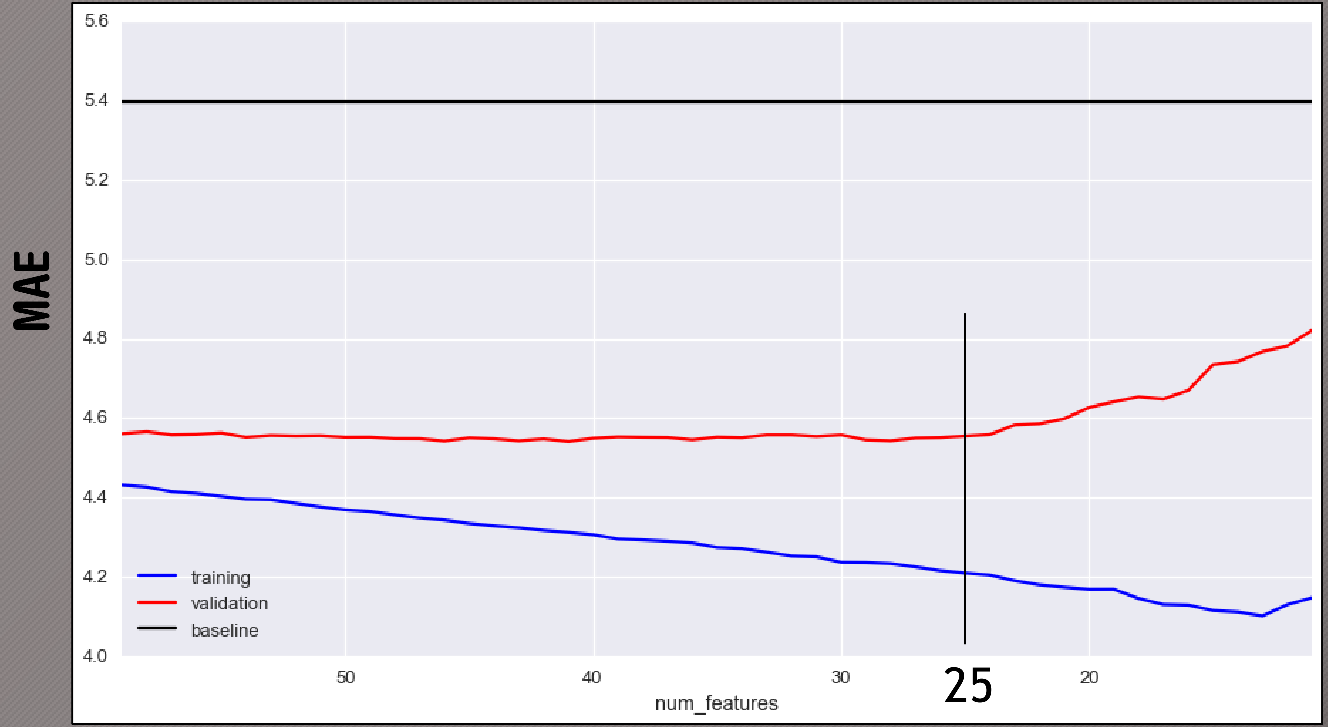
Linear Regression



```
1.          1.810894717139152e-46, 'c_chinese'
2.          7.2285477679285451e-20, 'n_chinatown'
3.          3.0320814941594692e-18, 'rating_4.5'
4.          3.9937047952153105e-14, 'c_coffeetea'
5.          1.0833529253459801e-11, 'chain'
6.          3.7961835364454637e-11, 'rating_3.5'
7.          2.8215529367003095e-10, 'c_venue'
8.          1.005588355444725e-09, 'c_thai'
9.          1.0717760243754801e-09, 'price_3'
10.         1.5998268265255936e-08, 'c_vietnamese'
11.         9.006843133751674e-08, 'price_2'
12.         1.3686555097935e-06, 'c_indian'
13.         2.0700553858507486e-06, 'c_desserts'
14.         4.0915394735824555e-06, 'n_russianhill'
15.         7.5880180054662327e-05, 'rating_5.0'
16.         0.00023091715880025794, 'n_mission'
17.         0.00040909974942875712, 'n_bernalheights'
18.         0.00064214295416450951, 'n_financialdistrict'
19.         0.00076069604792101497, 'price_4'
20.         0.0011117470575030132, 'c_japanese'
21.         0.0013196357937397 38, 'c_fastfood'
22.         0.0016738527093267197, 'insp_count'
23.         0.0038915225910515867, 'rating_3.0'
24.         0.0059144242908699 65, 'c_asian'
25.         0.014441365848194495, 'c_grocery'
26.         0.015171515966960245, 'c_bars'
27.         0.017443323923044839, 'rating_4.0'
28.         0.023421443595428235, 'n_potrerohill'
29.         0.27815701064558435, 'n_haight'


count   2398
mean     4.795664
std      3.799989
min      0.002107
25%      1.815817
50%      4.087439
75%      6.792585
max     27.895557
```

Random Forest Regression



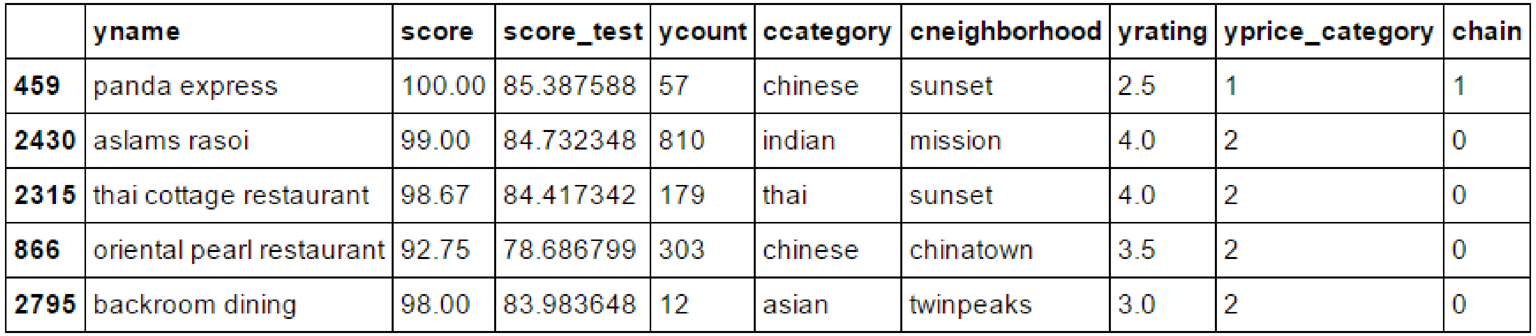
n_features = 10, min_samples_leaf = 5

```
1.    0.24727621331786329, 'ycount'
2.    0.1536387867101322, 'c_chinese'
3.    0.080218216633282227, 'insp_count'
4.    0.049499623659676265, 'n_chinatown'
5.    0.048760944031403734, 'rating_4.5'
6.    0.039798033848302279, 'chain'
7.    0.031893551363132648, 'rating_3.5'
8.    0.030674919560013794, 'c_thai'
9.    0.030343857314512959, 'c_coffeetea'
10.   0.029003944810917746, 'rating_4.0'
11.   0.028910887586159912, 'c_vietnamese'
12.   0.024836747237302611, 'price_2'
13.   0.0239275656249869, 'c_venue'
14.   0.019494771910807007, 'price_1'
15.   0.018774190011046427, 'n_sunset'
16.   0.018043212903155997, 'n_russianhill'
17.   0.016353538501653223, 'n_downtown'
18.   0.01601880840380349, 'price_3'
19.   0.015660240587556988, 'n_mission'
20.   0.015472136326288662, 'c_japanese'
21.   0.014890384430931297, 'rating_3.0'
22.   0.013798707580006769, 'c_fastfood'
23.   0.01291440092037606, 'n_financialdistrict'
24.   0.011315910083317615, 'c_bakeries'
25.   0.0084804066433696538, 'n_richmond'
```

```
count    2398
mean        4.114636
std         3.300645
min         0.001760
25%         1.609043
50%         3.507397
75%         5.758480
max        27.902103
```

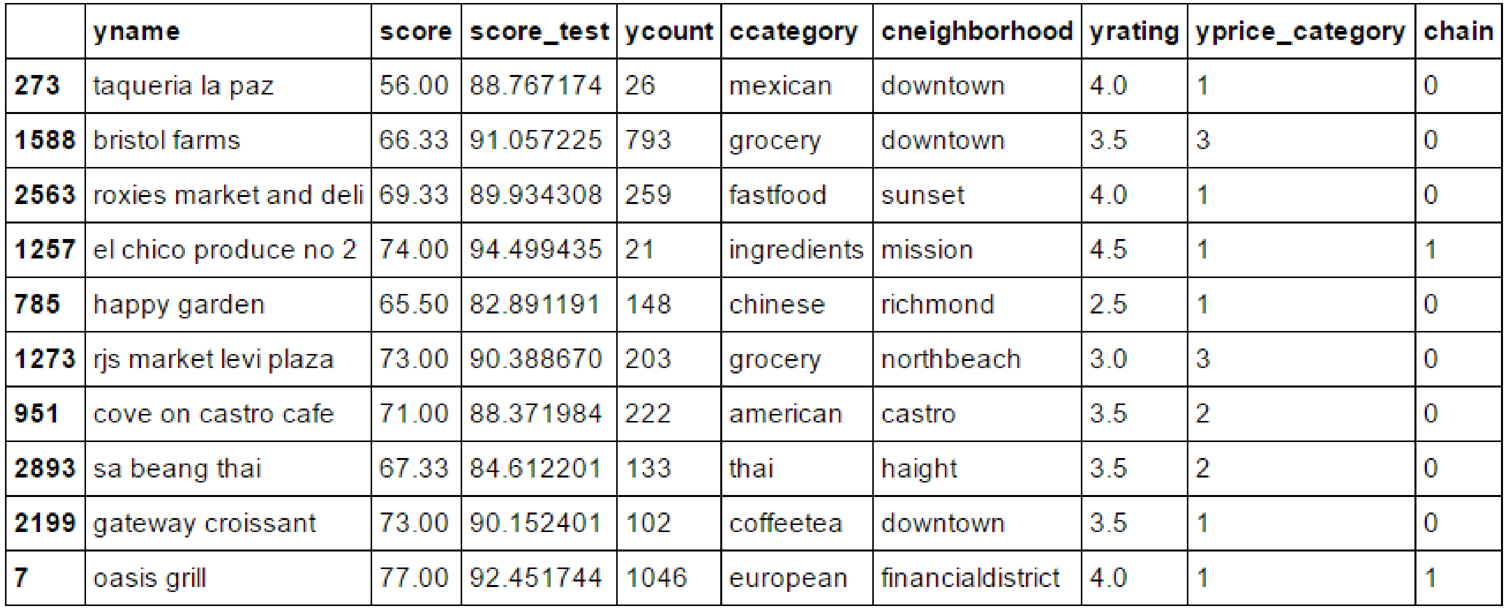
Linear Regression: Outliers

Under-predicted



	yname	score	score_test	ycount	ccategory	cneighborhood	yrating	yprice_category	chain
459	panda express	100.00	85.387588	57	chinese	sunset	2.5	1	1
2430	aslams rasoi	99.00	84.732348	810	indian	mission	4.0	2	0
2315	thai cottage restaurant	98.67	84.417342	179	thai	sunset	4.0	2	0
866	oriental pearl restaurant	92.75	78.686799	303	chinese	chinatown	3.5	2	0
2795	backroom dining	98.00	83.983648	12	asian	twinpeaks	3.0	2	0

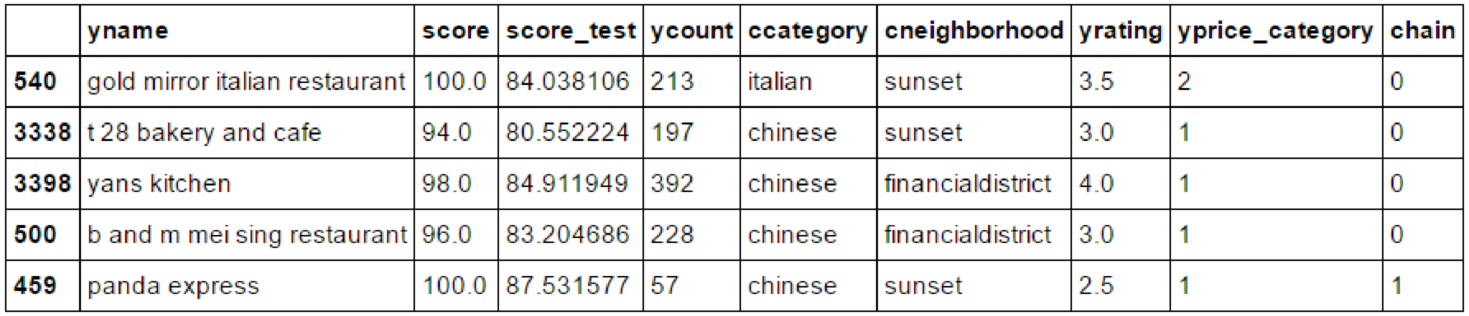
Over-predicted



	yname	score	score_test	ycount	ccategory	cneighborhood	yrating	yprice_category	chain
273	taqueria la paz	56.00	88.767174	26	mexican	downtown	4.0	1	0
1588	bristol farms	66.33	91.057225	793	grocery	downtown	3.5	3	0
2563	roxies market and deli	69.33	89.934308	259	fastfood	sunset	4.0	1	0
1257	el chico produce no 2	74.00	94.499435	21	ingredients	mission	4.5	1	1
785	happy garden	65.50	82.891191	148	chinese	richmond	2.5	1	0
1273	rjs market levi plaza	73.00	90.388670	203	grocery	northbeach	3.0	3	0
951	cove on castro cafe	71.00	88.371984	222	american	castro	3.5	2	0
2893	sa beang thai	67.33	84.612201	133	thai	haight	3.5	2	0
2199	gateway croissant	73.00	90.152401	102	coffeetea	downtown	3.5	1	0
7	oasis grill	77.00	92.451744	1046	european	financialdistrict	4.0	1	1

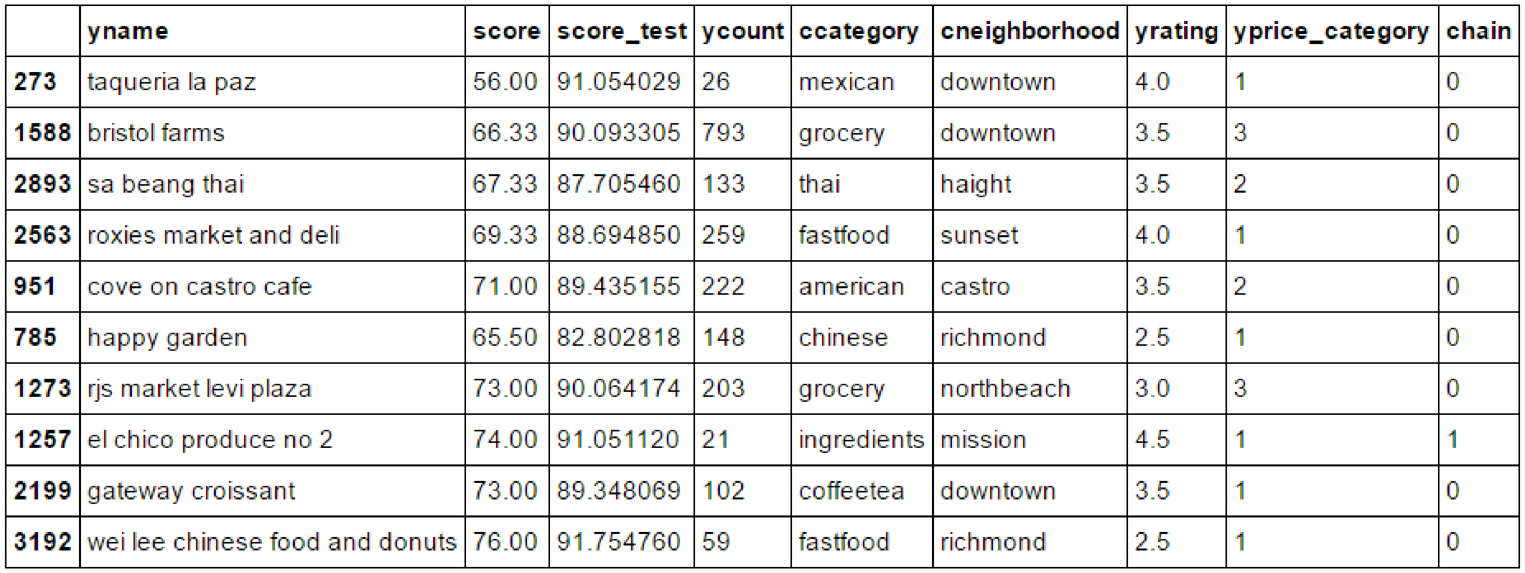
Random Forest: Outliers

Under-predicted



	yname	score	score_test	ycount	ccategory	cneighborhood	yrating	yprice_category	chain
540	gold mirror italian restaurant	100.0	84.038106	213	italian	sunset	3.5	2	0
3338	t 28 bakery and cafe	94.0	80.552224	197	chinese	sunset	3.0	1	0
3398	yans kitchen	98.0	84.911949	392	chinese	financialdistrict	4.0	1	0
500	b and m mei sing restaurant	96.0	83.204686	228	chinese	financialdistrict	3.0	1	0
459	panda express	100.0	87.531577	57	chinese	sunset	2.5	1	1

Over-predicted



	yname	score	score_test	ycount	ccategory	cneighborhood	yrating	yprice_category	chain
273	taqueria la paz	56.00	91.054029	26	mexican	downtown	4.0	1	0
1588	bristol farms	66.33	90.093305	793	grocery	downtown	3.5	3	0
2893	sa beang thai	67.33	87.705460	133	thai	haight	3.5	2	0
2563	roxies market and deli	69.33	88.694850	259	fastfood	sunset	4.0	1	0
951	cove on castro cafe	71.00	89.435155	222	american	castro	3.5	2	0
785	happy garden	65.50	82.802818	148	chinese	richmond	2.5	1	0
1273	rjs market levi plaza	73.00	90.064174	203	grocery	northbeach	3.0	3	0
1257	el chico produce no 2	74.00	91.051120	21	ingredients	mission	4.5	1	1
2199	gateway croissant	73.00	89.348069	102	coffeetea	downtown	3.5	1	0
3192	wei lee chinese food and donuts	76.00	91.754760	59	fastfood	richmond	2.5	1	0

Model Comparison

N = 599	Baseline	Linear (OLS)	Random Forest
Improvement (mean/mean)	-	8.3%	11.5%
mean	5.40	4.95	4.78
std	4.43	3.98	3.94
25%	2.06	2.01	1.84
50%	4.39	4.05	4.00
75%	7.94	6.87	6.47

Model Comparison

N = 599	Baseline	Linear (OLS)	Random Forest
Improvement (mean/mean)	-	8.3%	11.5%
mean	5.40	4.95	4.78
std	4.43	3.98	3.94
25%	2.06	2.01	1.84
50%	4.39	4.05	4.00
75%	7.94	6.87	6.47

- High Risk
  - -7 pts
  - Violations that directly relate to the transmission of food borne illnesses, the adulteration of food products and the contamination of food-contact surfaces.
- Moderate Risk
  - - 4 pts
  - Violations that are of a moderate risk to the public health and safety.
- Low Risk
  - -2 pts
  - Violations that are low risk or have no immediate risk to the public health and safety.

Conclusions

- Be careful eating Chinese in Chinatown

- Exploratory data analysis was really interesting
  - What is the highest rated & most reviewed business with the lowest inspection score?
    - Arizmendi Bakery, 4.5 Yelp rating and 1682 ratings, w/ mean score of 78.5 over 2 inspections
  - What is the most expensive business with the lowest inspection score?
    - Campton Place (2 Michelin Stars!!) w/ mean score of 80 over 4 inspections

- Model could be useful for resource allocation but it is not good at capturing extremes

Conclusions

- High degree of variability in scores due to the random nature of violations
  - Off/busy day
  - Bad employees
  - Lack of training
  - Age/condition of kitchen

- All of the factors above would likely not be apparent to Yelp reviewers
  - It would be interesting to add a feature if the kitchen is visible to patrons (an open kitchen); currently doesn't exist in Yelp data but could possibly be text mined from reviews

Next Steps

- Cleaner data = More data
  - Need a better link between SFDPH and Yelp data
  - It would be best if the Yelp ID was imbedded in the SFDPH dataset
  - The Yelp ID (their primary key) is the only real unique way to identify a business in Yelp

- NLP / latent variable
  - Analysis of inspection violation types
  - Sentiment analysis in Yelp reviews

- Collect new data over fixed time intervals
  - New Yelp data + historical inspection and Yelp data -> predict new inspection score
  - This type of model would likely be more useful for better resource allocation of inspectors

End

Sample Data: SF Health Department

business_id	name	address	city	state	postal_code	latitude	longitude	phone_number
10	TIRAMISU KITCHEN	033 BELDEN PL	San Francisco	CA	94104	37.791116	-122.403816	14154217044
19	NRGIZE LIFESTYLE CAFE	1200 VAN NESS AVE, 3RD FLOOR	San Francisco	CA	94109	37.786848	-122.421547	14157763262
24	OMNI S.F. HOTEL - 2ND FLOOR PANTRY	500 CALIFORNIA ST, 2ND FLOOR	San Francisco	CA	94104	37.792888	-122.403135	14156779494
31	NORMAN'S ICE CREAM AND FREEZES	2801 LEAVENWORTH ST	San Francisco	CA	94133	37.807155	-122.419004	
45	CHARLIE'S DELI CAFE	3202 FOLSOM ST	San Francisco	CA	94110	37.747114	-122.413641	14156415051
48	ART'S CAFE	747 IRVING ST	San Francisco	CA	94122	37.764013	-122.465749	14156657440
50	SUSHI ZONE	1815 MARKET ST.	San Francisco	CA	94103	37.771437	-122.423892	14156211114
54	RHODA GOLDMAN PLAZA	2180 POST ST	San Francisco	CA	94115	37.784626	-122.437734	14153455060
56	CAFE X + O	1799 CHURCH ST	San Francisco	CA	94131	37.742325	-122.426476	14158263535

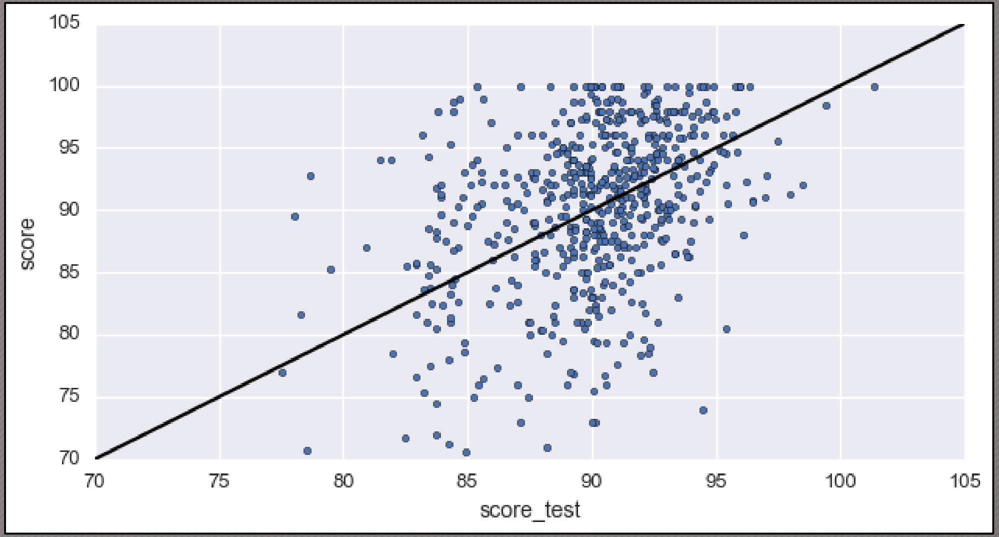
↕

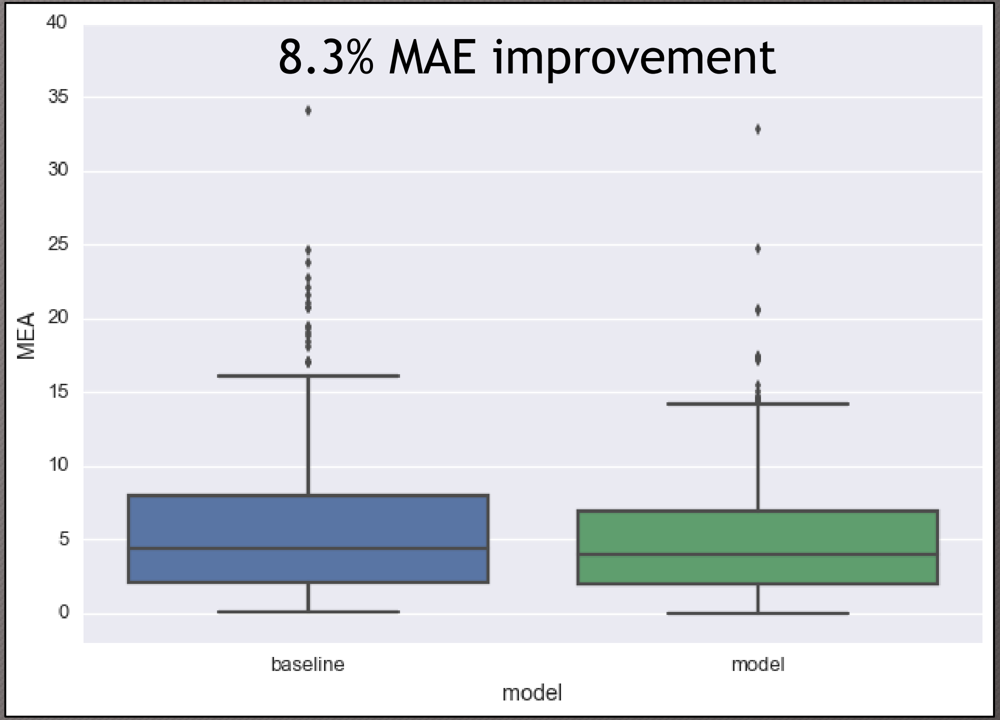
business_id	score	date	type
10	82	20160503	routine
10	94	20140729	routine
10	92	20140114	routine
19	94	20160513	routine
19	94	20141110	routine
19	94	20140214	routine
19	96	20130904	routine
24	96	20160311	routine
24	96	20141124	routine
24	96	20140612	routine
24	100	20131118	routine

Sample Data: Yelp API’s

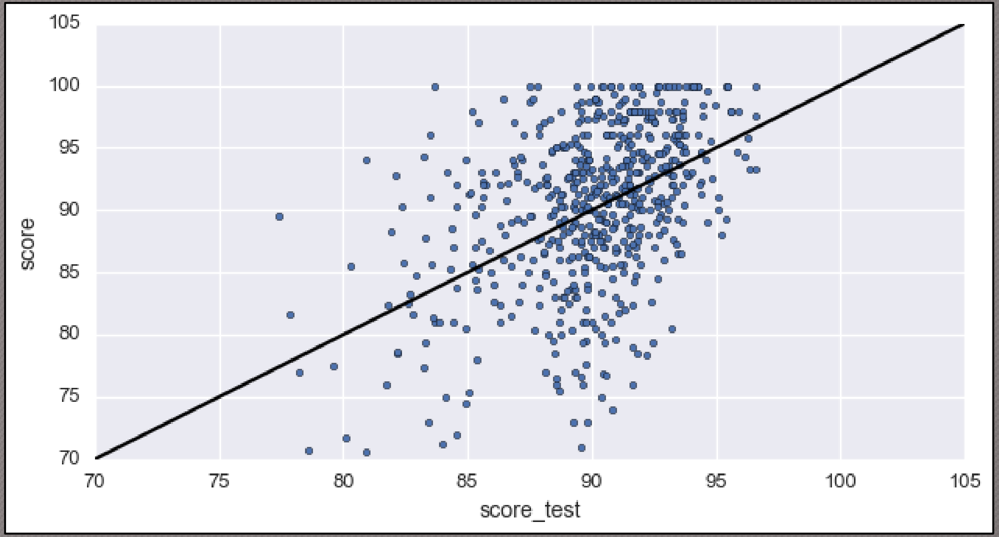
- JSON output
- [{"rating": 4.0, "review_count": 1046, "name": "Oasis Grill", "photos": ["https://s3-media3.fl.yelpcdn.com/bphoto/ct6p78kOUf6u6JVES6dTNQ/o.jpg", "https://s3-media4.fl.yelpcdn.com/bphoto/Sg5cz9hi_p6G22TFja2xRA/o.jpg", "https://s3-media2.fl.yelpcdn.com/bphoto/FlbKcFrbuwyqSgsOA2_4PA/o.jpg"], "url": "https://www.yelp.com/biz/oasis-grill-san-francisco?adjust_creative=wXkLRioDMWxUkNynLkngJg&utm_campaign=yelp_api_v3&utm_medium=api_v3_business_lookup&utm_source=wXkLRioDMWxUkNynLkngJg", "price": "$", "coordinates": {"latitude": 37.7944464239893, "longitude": -122.396736520868}, "hours": [{"hours_type": "REGULAR", "open": [{"is_overnight": false, "end": "2200", "day": 0, "start": "0900"}, {"is_overnight": false, "end": "2200", "day": 1, "start": "0900"}, {"is_overnight": false, "end": "2200", "day": 2, "start": "0900"}, {"is_overnight": false, "end": "2200", "day": 3, "start": "0900"}, {"is_overnight": false, "end": "2200", "day": 4, "start": "0900"}, {"is_overnight": false, "end": "2200", "day": 5, "start": "1000"}, {"is_overnight": false, "end": "2200", "day": 6, "start": "1000"}], "is_open_now": true}], "phone": "+14157810313", "image_url": "https://s3-media4.fl.yelpcdn.com/bphoto/ct6p78kOUf6u6JVES6dTNQ/o.jpg", "location": {"city": "San Francisco", "address1": "91 Drumm St", "address2": "", "address3": "", "state": "CA", "country": "US", "zip_code": "94111"}, "id": "oasis-grill-san-francisco", "categories": [{"alias": "greek", "title": "Greek"}, {"alias": "mediterranean", "title": "Mediterranean"}, {"alias": "tradamerican", "title": "American (Traditional)"}]

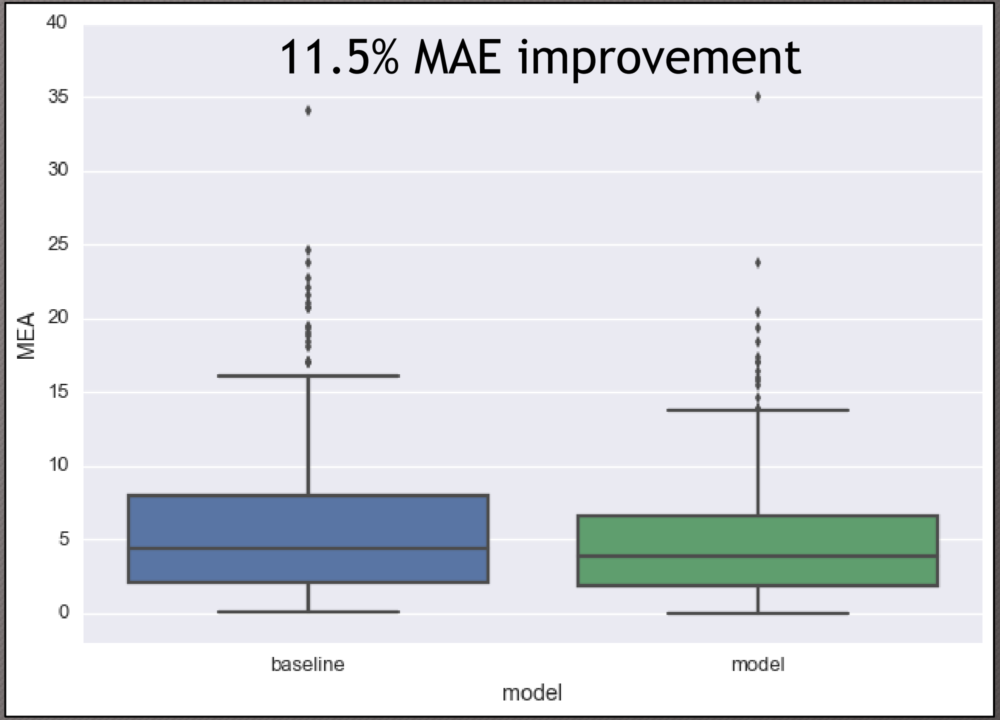
Results: Linear Regression





Results: Random Forest Regression





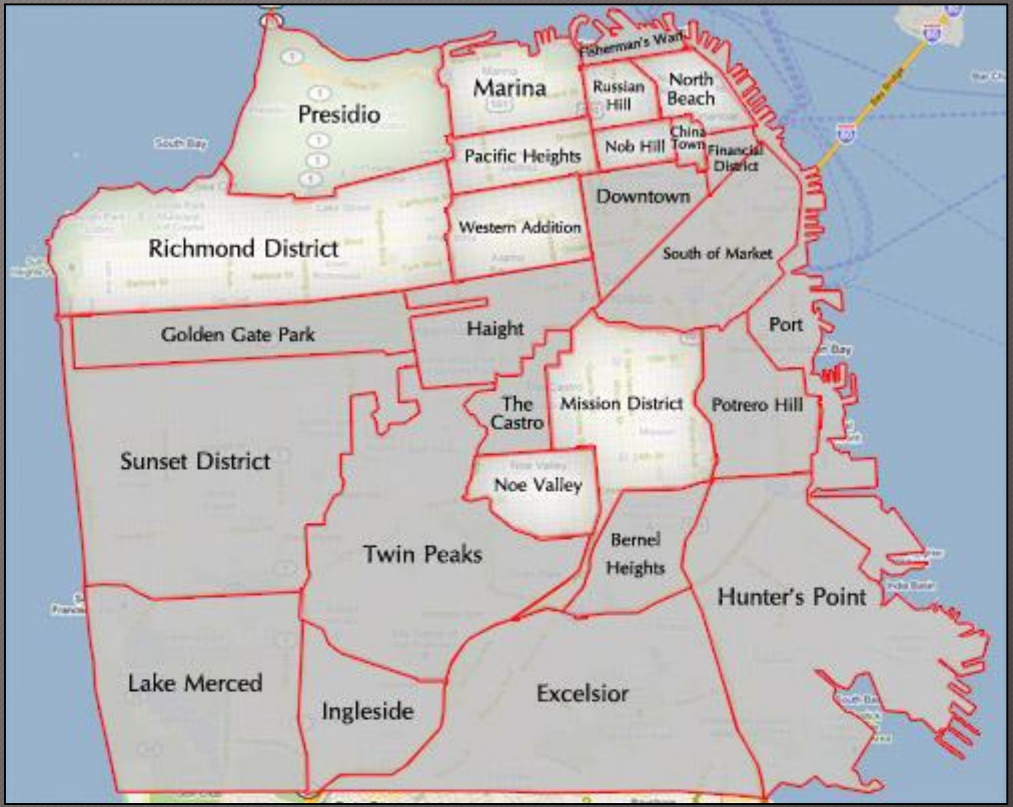
Interesting Data

Total Yelp reviews by neighborhood

mission	168495
financialdistrict	116963
downtown	108509
northbeach	108137
richmond	99474
sunset	97861
soma	91942
pacificheights	66837
nobhill	66135
russianhill	57943
haight	52343
marina	49518
westernaddition	42838
castro	36007
chinatown	31163
twinpeaks	21677
noevalley	18706
potrerohill	18513
bernalheights	18287
excelsior	12163
bayviewhunterspoint	3781

Mean Yelp reviews by neighborhood

westernaddition	680
northbeach	629
russianhill	552
haight	485
soma	479
potrerohill	475
mission	473
nobhill	466
marina	446
pacificheights	440
richmond	434
downtown	422
castro	409
sunset	387
financialdistrict	383
noevalley	367
chinatown	318
bernalheights	290
twinpeaks	285
excelsior	129
bayviewhunterspoint	88


Fisherman's Wharf
Presidio
Marina
Russian Hill
North Beach
China Town
Pacific Heights
Nob Hill
Financial District
Downtown
Richmond District
Western Addition
South of Market
Golden Gate Park
Haight
Port
The Castro
Mission District
Potrero Hill
Sunset District
Noe Valley
Twin Peaks
Bernel Heights
Hunter's Point
Lake Merced
Ingleside
Excelsior