Chicago Restaurant Risk Analysis

Yelp & City of Chicago Restaurant data

Chunyang Zhang

Pipeline

background introduction

Topic, Research Questions

Data

Data Source, Data Collection, Data Cleaning & Wrangling, EDA

Results

Data visualization, Analysis

Conclusion

Take Home points

Background Introduction

-----Restaurant sanitary level

- Consumer health and safety
- Public health policy
- industry standards
- Improve business practices

"CDC estimates 48 million people get sick, 128,000 are hospitalized, and 3,000 die from foodborne diseases each year in the United States"

"Over half of all foodborne disease outbreaks reported to the Centers for Disease Control and Prevention are associated with eating in restaurants or delicatessens"

Background Introduction

Important Definitions

- risk level
- features of restaurants
- customer segmentation

Research Questions

- How is customer segmentation(zip codes, income level) and a restaurant's risk level(sanitary level) correlated ?
- How features of restaurants correlated with its risk level?

Data Sources and Collection

yelp fusion

- **Dataset size**: 39,158 raw data, **13,159** data after deduplication (**16 features**)

- **Advantages：** huge amount of active users, a good representation of how most consumers perceive a restaurant
- Not a sensitive API, returns clear structured data

- **Challenges**: Limit of 1000 businesses per request; solved with changing search criteria from 'city' to 'zipcodes.'

CHICAGO DATA PORTAL

Food Inspections

- **Dataset size**: **156,379 data, 16 features**

- **Advantages：** an **official government website**; a reliable source of information that has been **verified and validated** by the city government
- Open and easy to download file

- **Challenges**: Some features don't have clear definitions; Some records are not up-to-date...
-

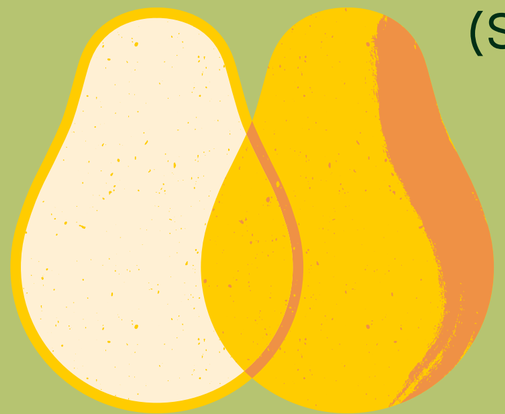
simplemaps
Interactive Maps & Data

- **Dataset size: 60*2**
- **Advantages：simple**
- **Challenges**:

Other Data Sources https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-ZIP-Codes/gdcf-axmw

Data Cleaning

- drop rows with **missing** values
- drop **duplicate** rows
- Convert the address of the restaurant to **lowercase**
- **Extract** Address, City, State and zip code from location in dedup_yelp
- use the **clean function** to clean address, city and state
- Only keep **up-to-date** records
  (Standard: Results: 'Out of business', 'Business not located'.)



Record Linkage

CHICAGO DATA PORTAL

yelp

Income Household Median

city	Chicago, IL
field	Income Household Median
fieldname	`income_household_median`

Linking Method

1. use address, city, and state to link Yelp data and Chicago government data
2. use zip code to link income household median with the above data frame

Exploratory Data Analysis

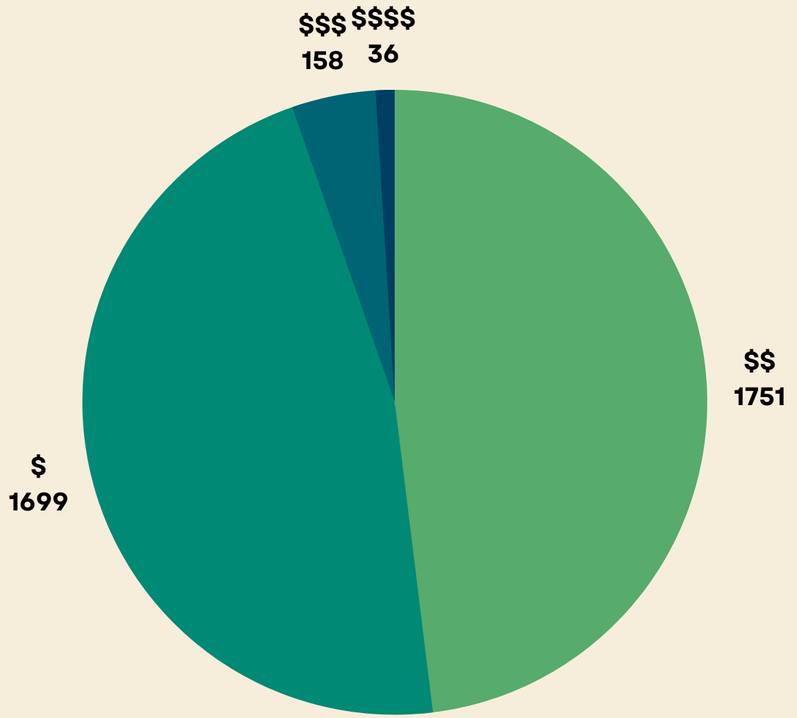
1. Select caring features and do EDA
2. Inspect each feature's data type
3. Do feature engineering if it's necessary
4. Select features that are important but hard to deal with

Data after linking and basic preprocessing: 3644 rows × 36 columns

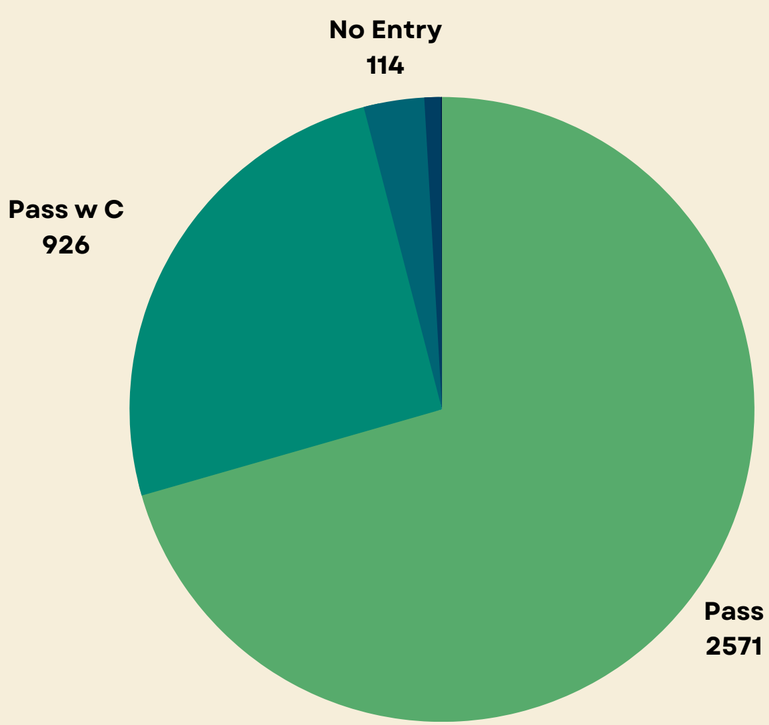
	id	alias	name	image_url	is_closed	url	review_count	categories	rating	transactions	...	City	State	Zip	Inspection Date	Inspection Type	Results	Violations	Latitude	Longitude	income_household_median
0	LPwAwxEjetjdNh7Uadro3g	smoque-bbq-chicago	Smoque BBQ	https://s3-media3.fl.yelpcdn.com/bphoto/wBfhSd...	False	https://www.yelp.com/biz/smoque-bbq-chicago?ad...	4622	[{'alias': 'bbq', 'title': 'Barbeque'}]	4.5	[]	...	chicago	il	60641	7/27/22	Canvass	Pass	53. TOILET FACILITIES: PROPERLY CONSTRUCTED, S...	41.950073	-87.727657	63545

Exploratory Data Analysis

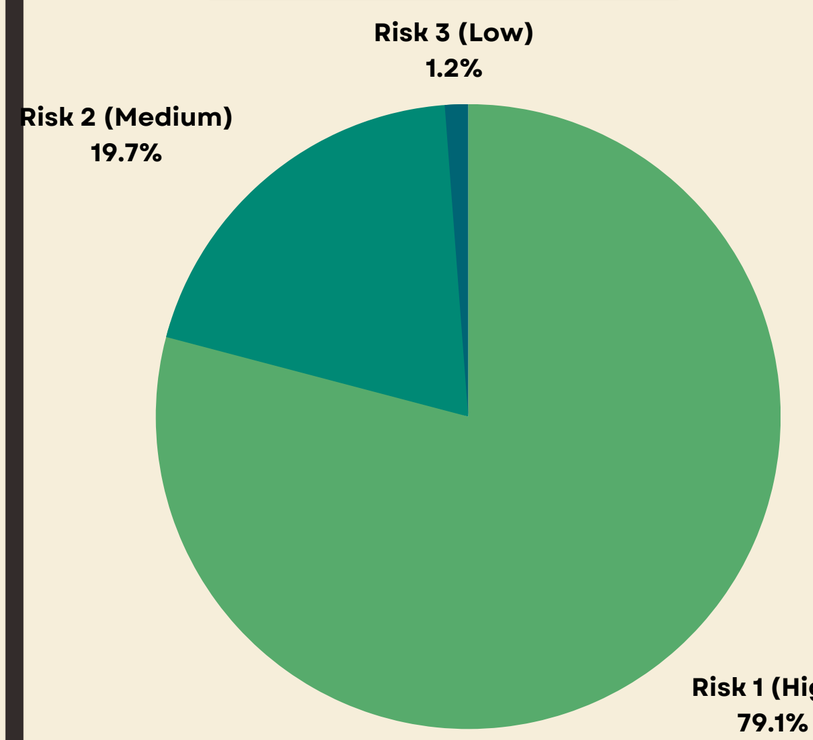
Price level



Inspection Results



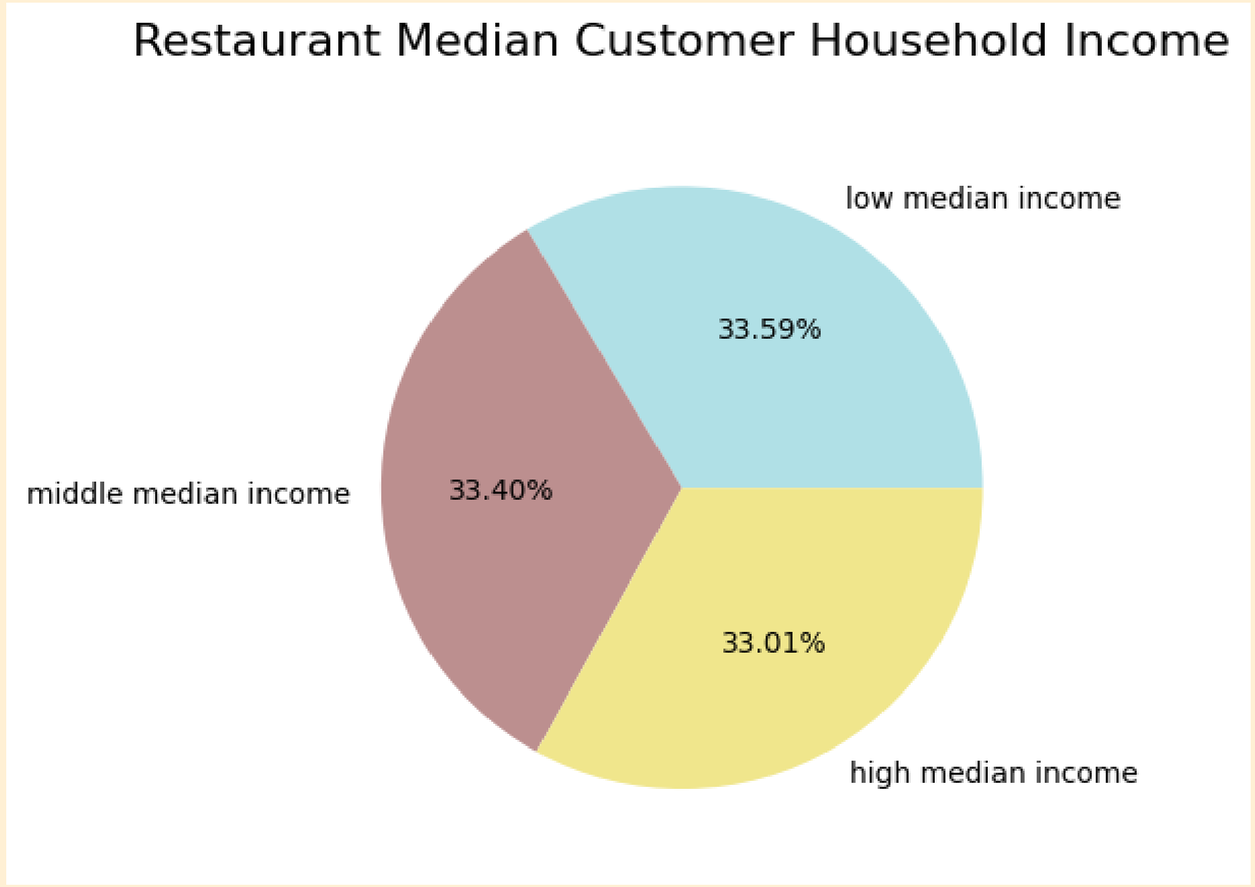
Risk



Exploratory Data Analysis

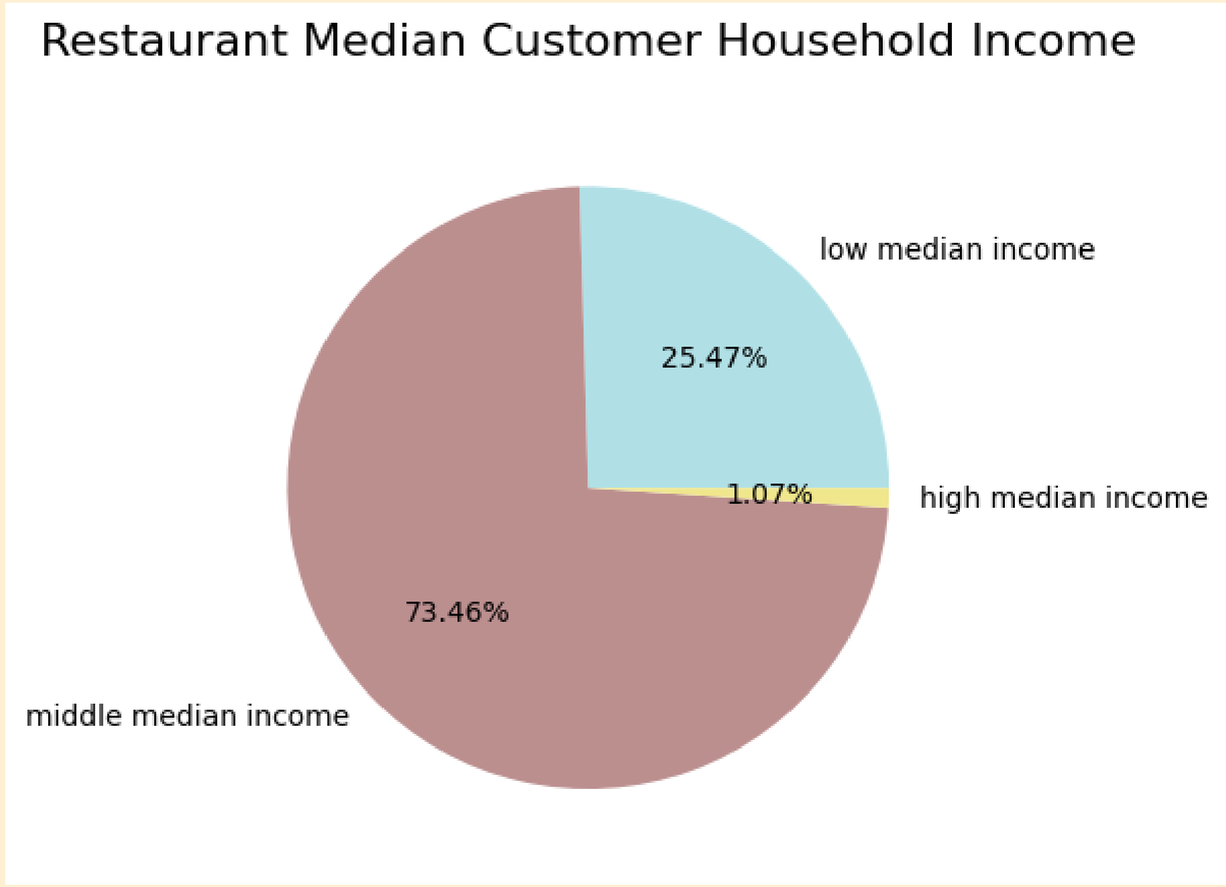
Income Level

classification-1



i > 61625 and i < 101091

classification-2

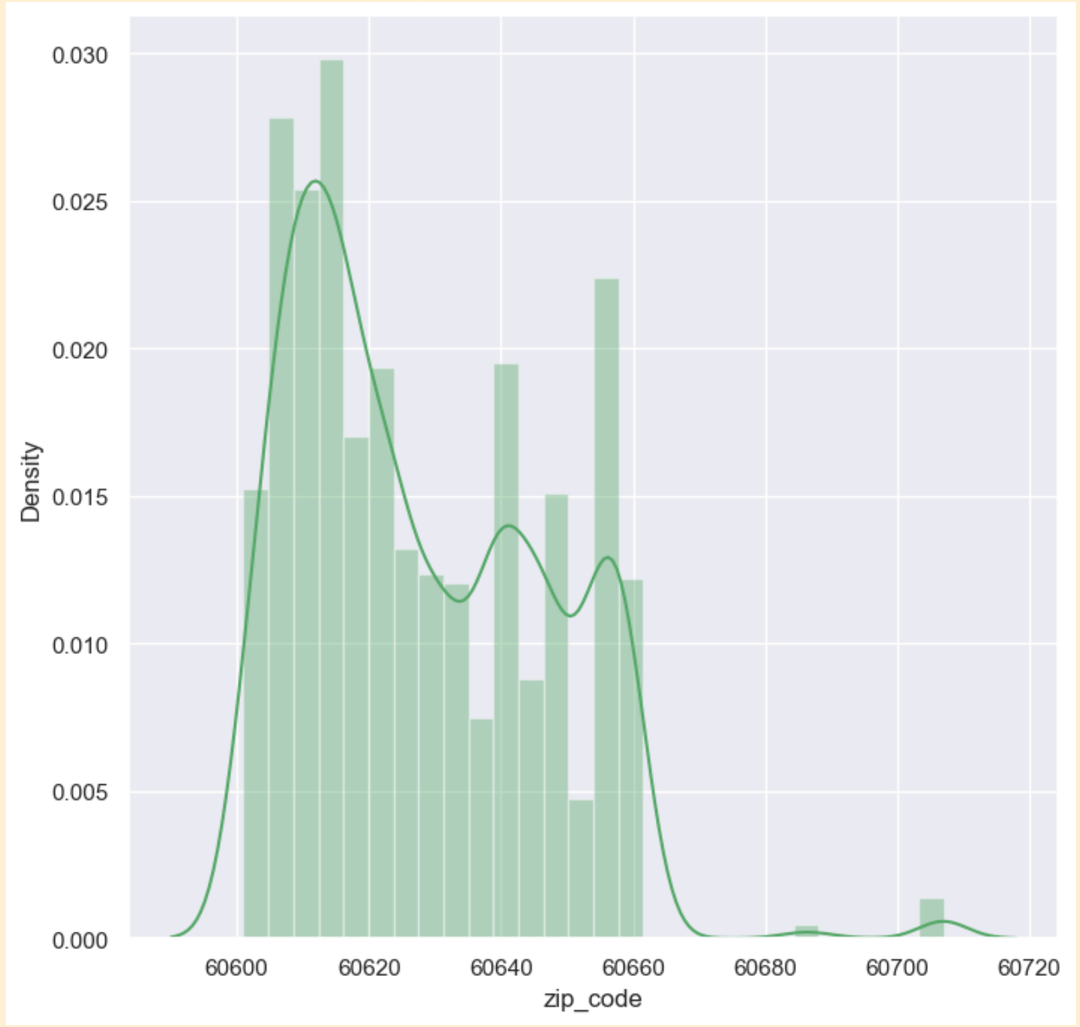


i > 52000 and i <= 156000

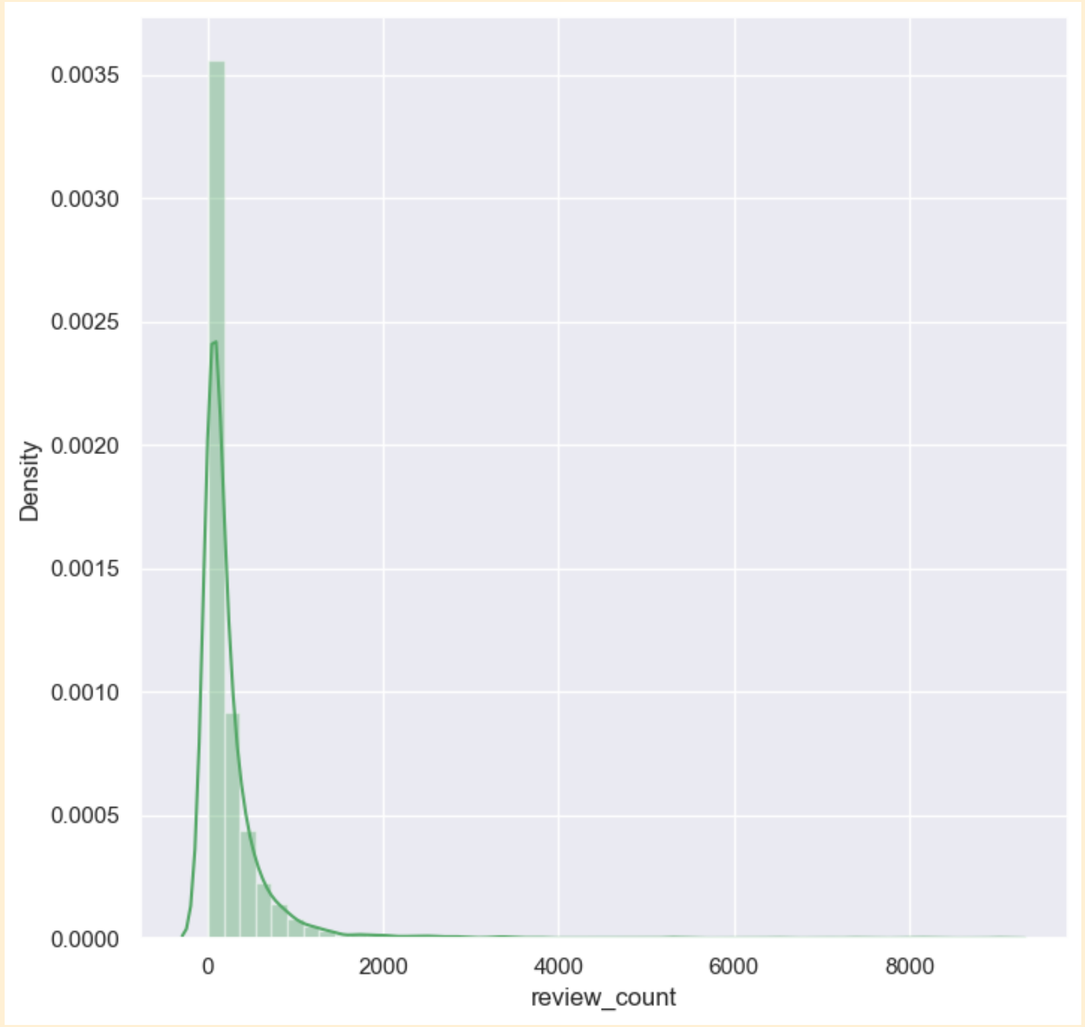
Exploratory Data Analysis

Continuous

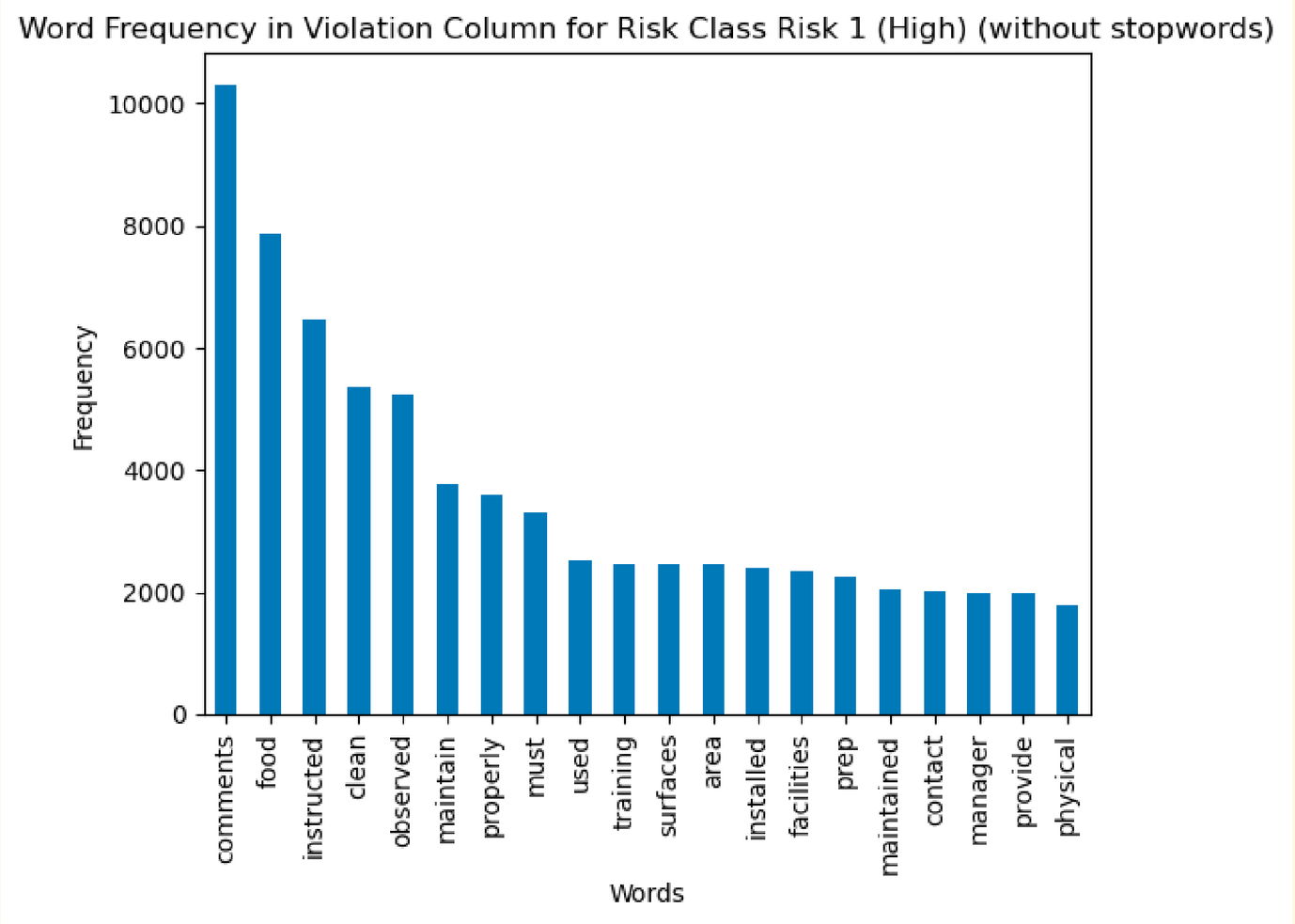
zip_code



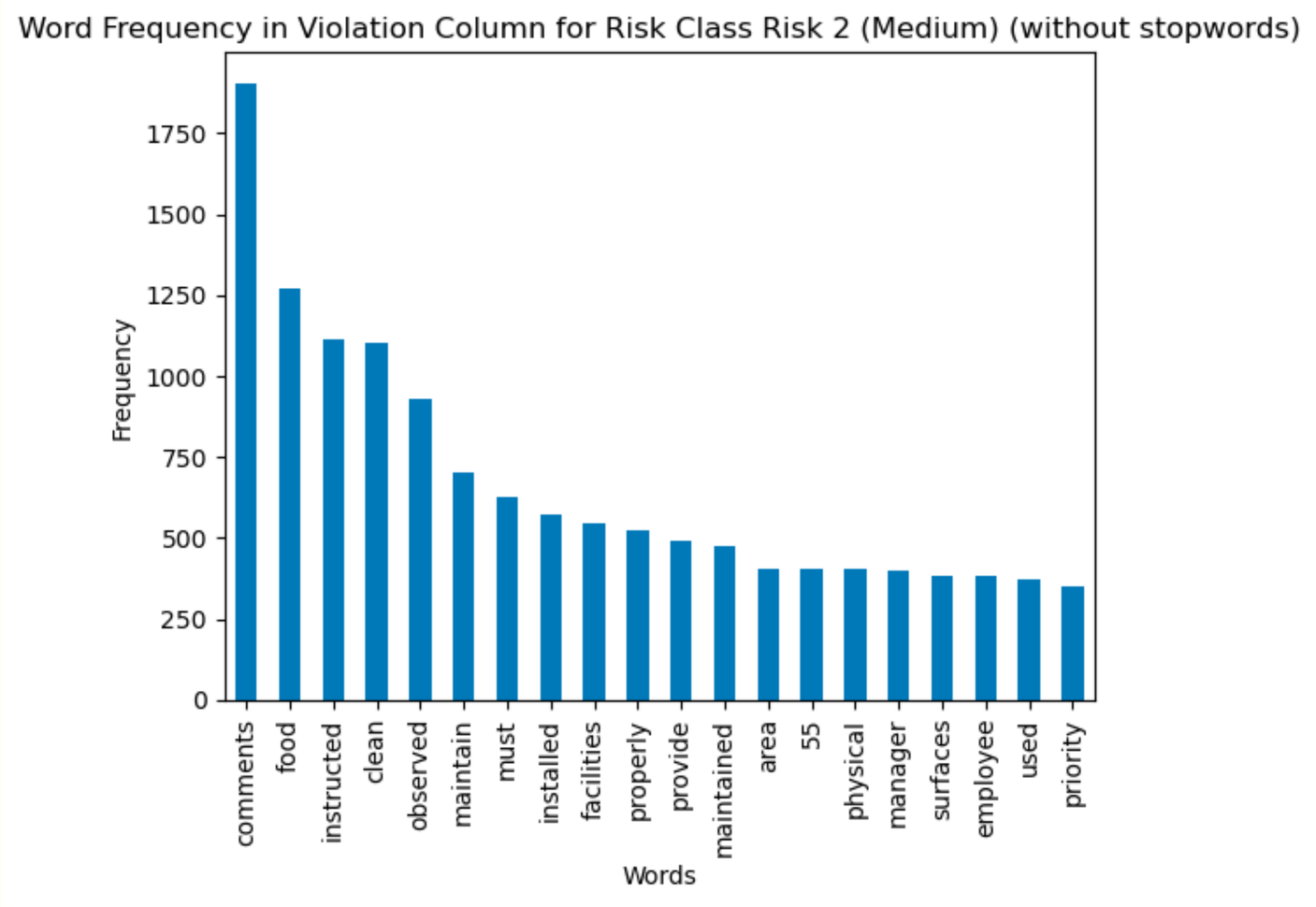
review_count



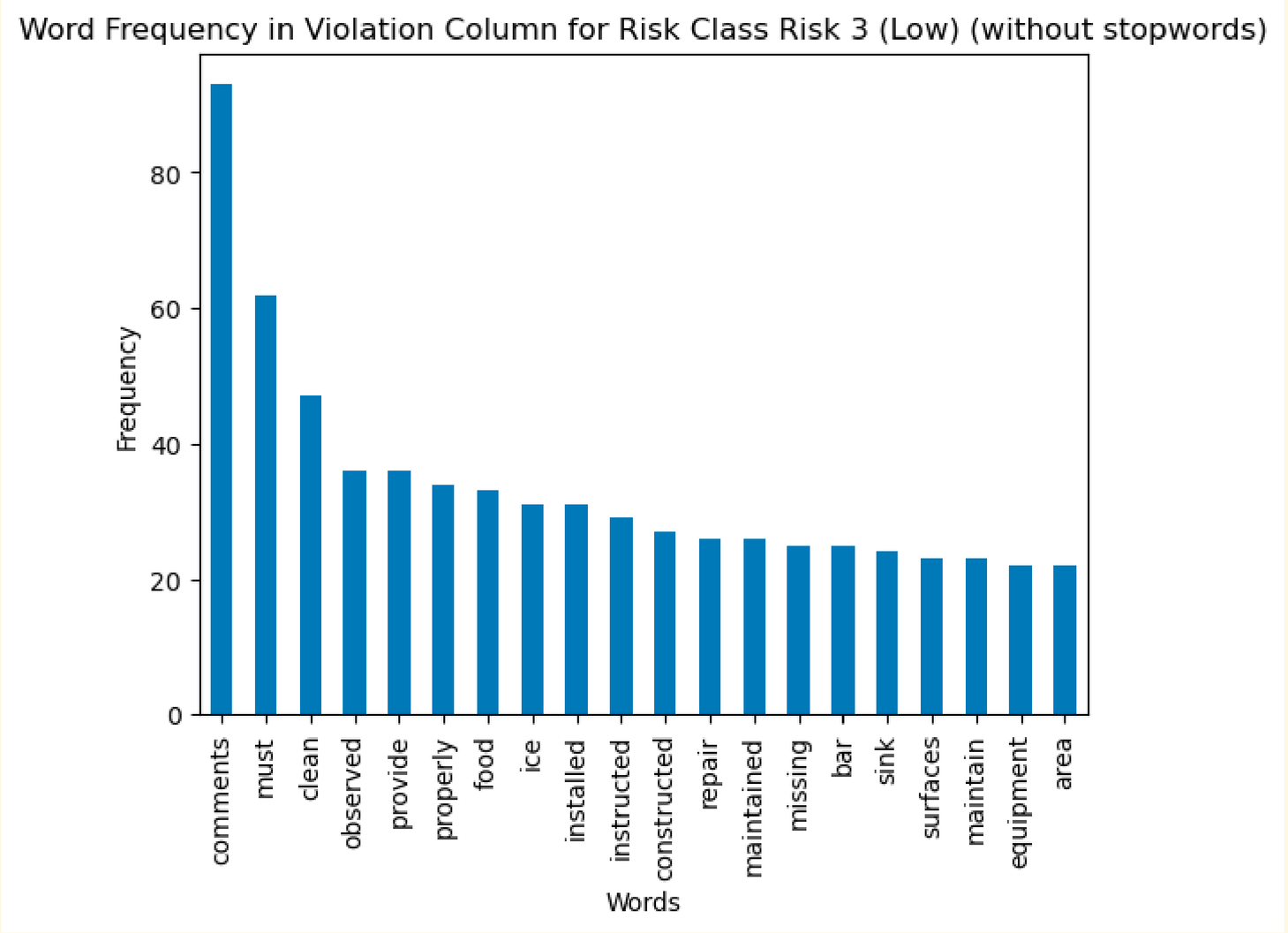
Worldcloud
frequency distribution of Risk1



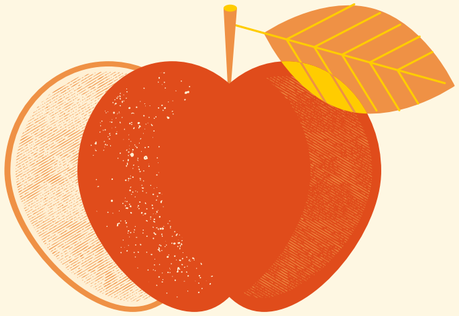
Worldcloud
frequency distribution of Risk2



Worldcloud
frequency distribution of Risk3



Feature Selection



1. Target Variable: 'Risk'
2. Variables important but hard to deal with : 'categories', 'violations'
3. Selected independent Variables:

Risk	price	rating	transactions	income_household_median	review_count	delivery	pickup	restaurant_reservation
Risk 1 (High)	$$	4.5	['pickup', 'restaurant_reservation', 'delivery']	63545	595	1	1	1

Feature Engineering

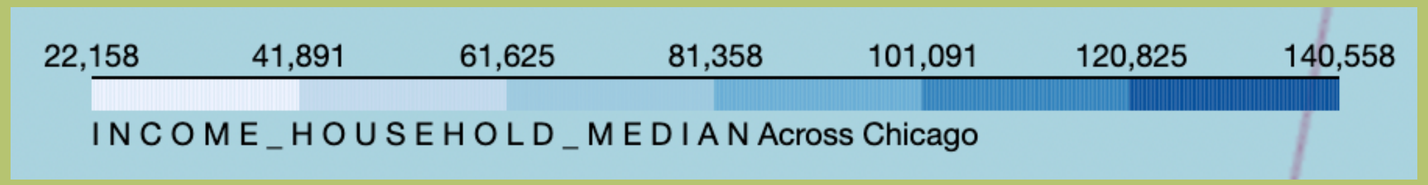
1. List, Dictionary-like str variables: use regex extract useful information

[{'alias': 'burgers',

'title': 'Burgers'}, {'alias': 'hotdogs', 'title': 'Fast Food'}, {'alias':

'hotdog', 'title': 'Hot Dogs'}]

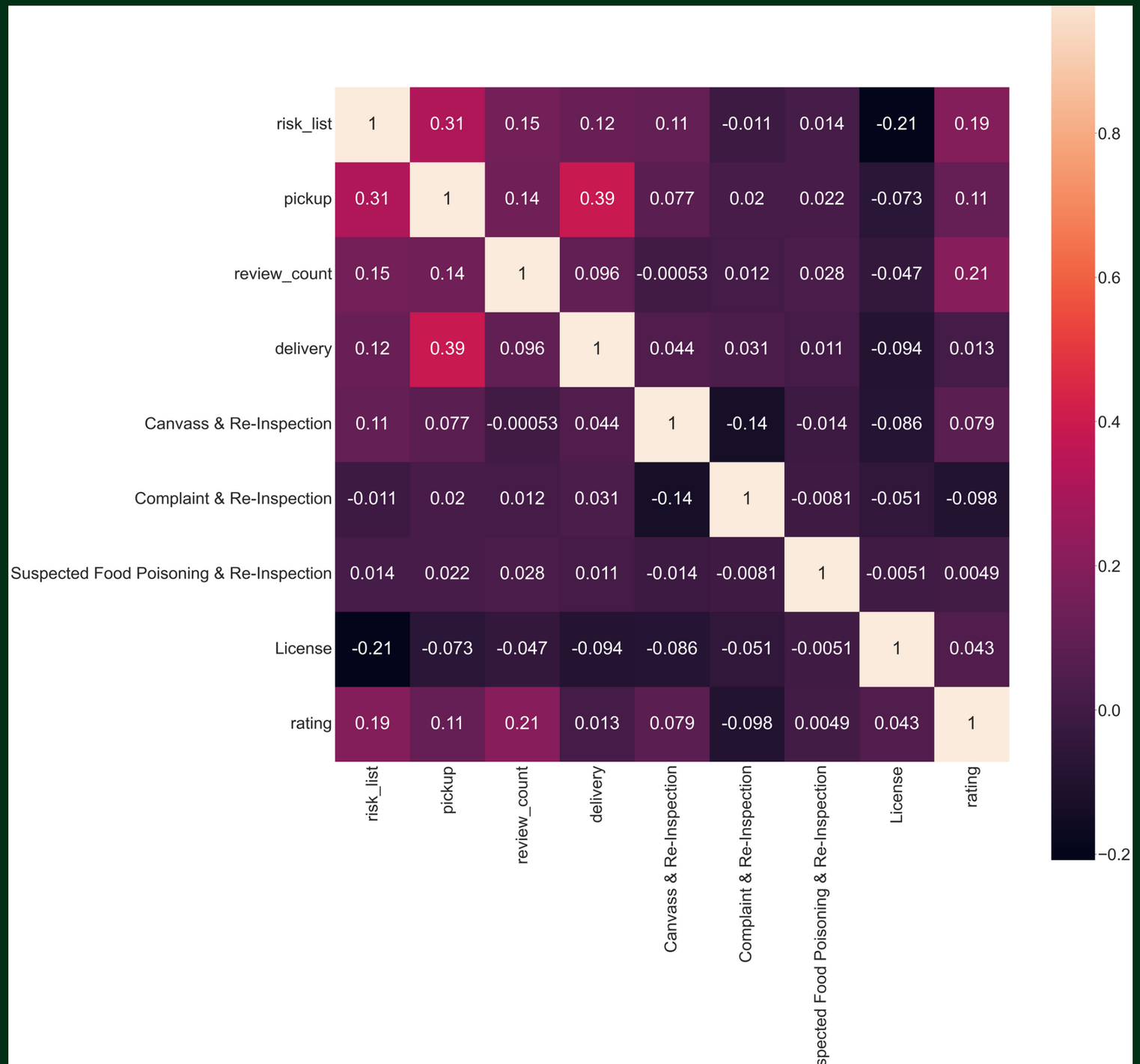
2. According to data distribution, categorize 'income_household_median' into 3 levels



3. Categorical Variable: assign numerical values (direct assign or get dummies)

	rating	review_count	delivery	pickup	restaurant_reservation	income_level	risk_num	Results_Fail	Results_No Entry	Results_Not Ready	...	Complaint	Complaint Re-Inspection	License	License Re-Inspection	Non-Inspection	Recent Inspection	Short Form Complaint	Suspected Food Poisoning	Suspected Food Poisoning Re-inspection	Price_encoded
0	4.5	4622	0	0	0	2	3	0	0	0	...	0	0	0	0	0	0	0	0	0	2

Results



According to this correlation heat map, we find:

1/ pickup & delivery

2/pickup & risk

3/license & risk

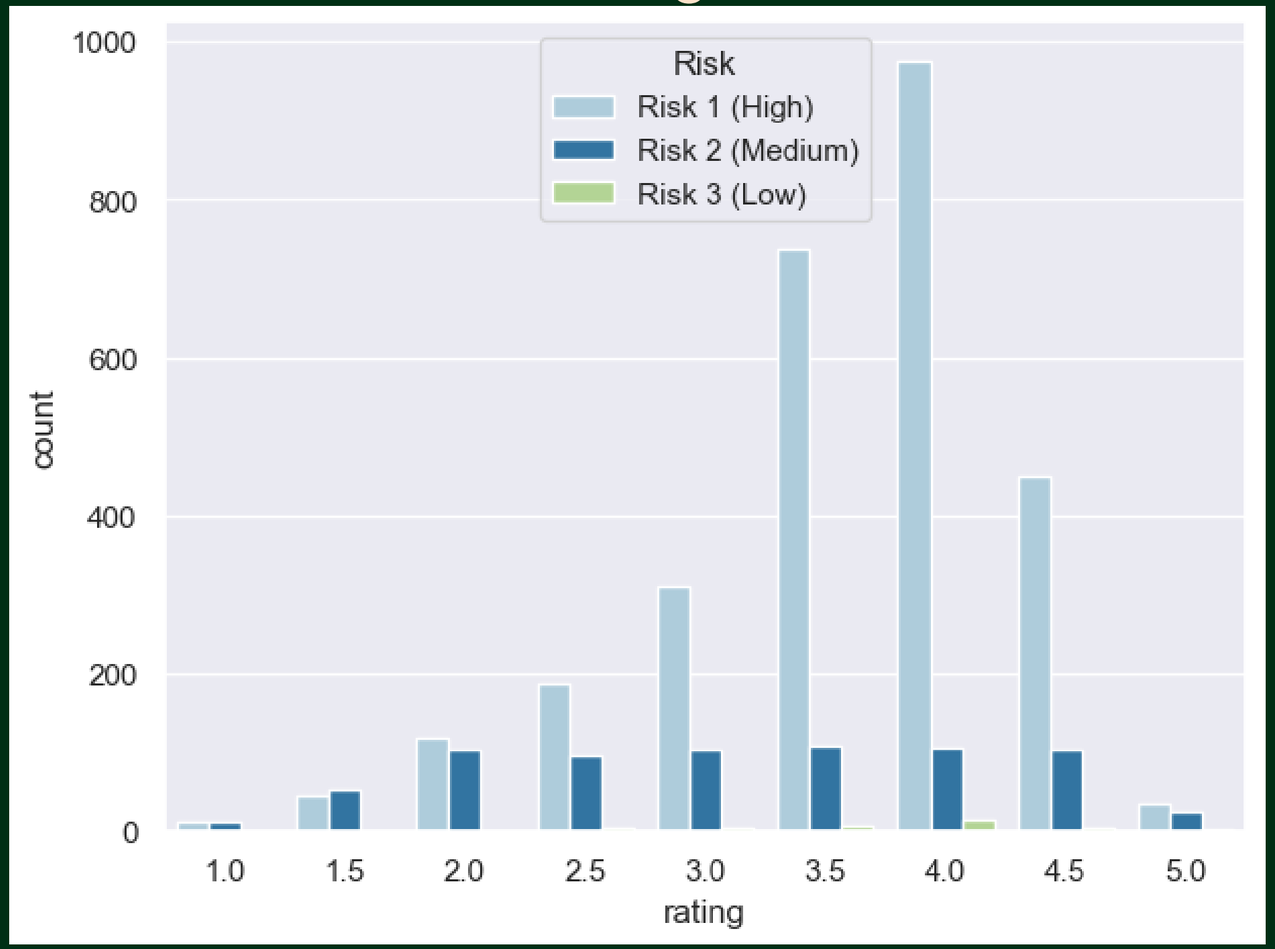
4/rating & risk

5/review counts & risk

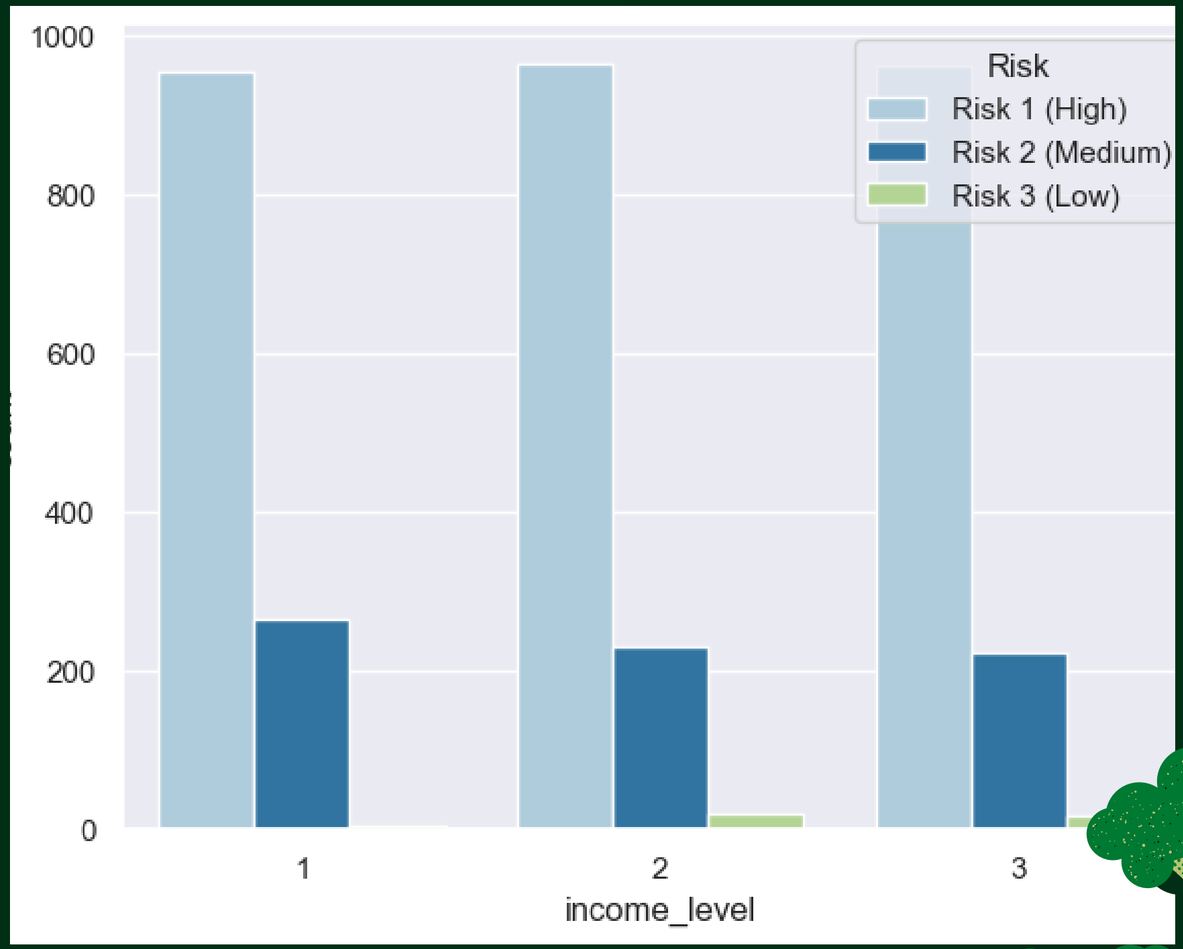
6/delivery & risk

Results

Rating

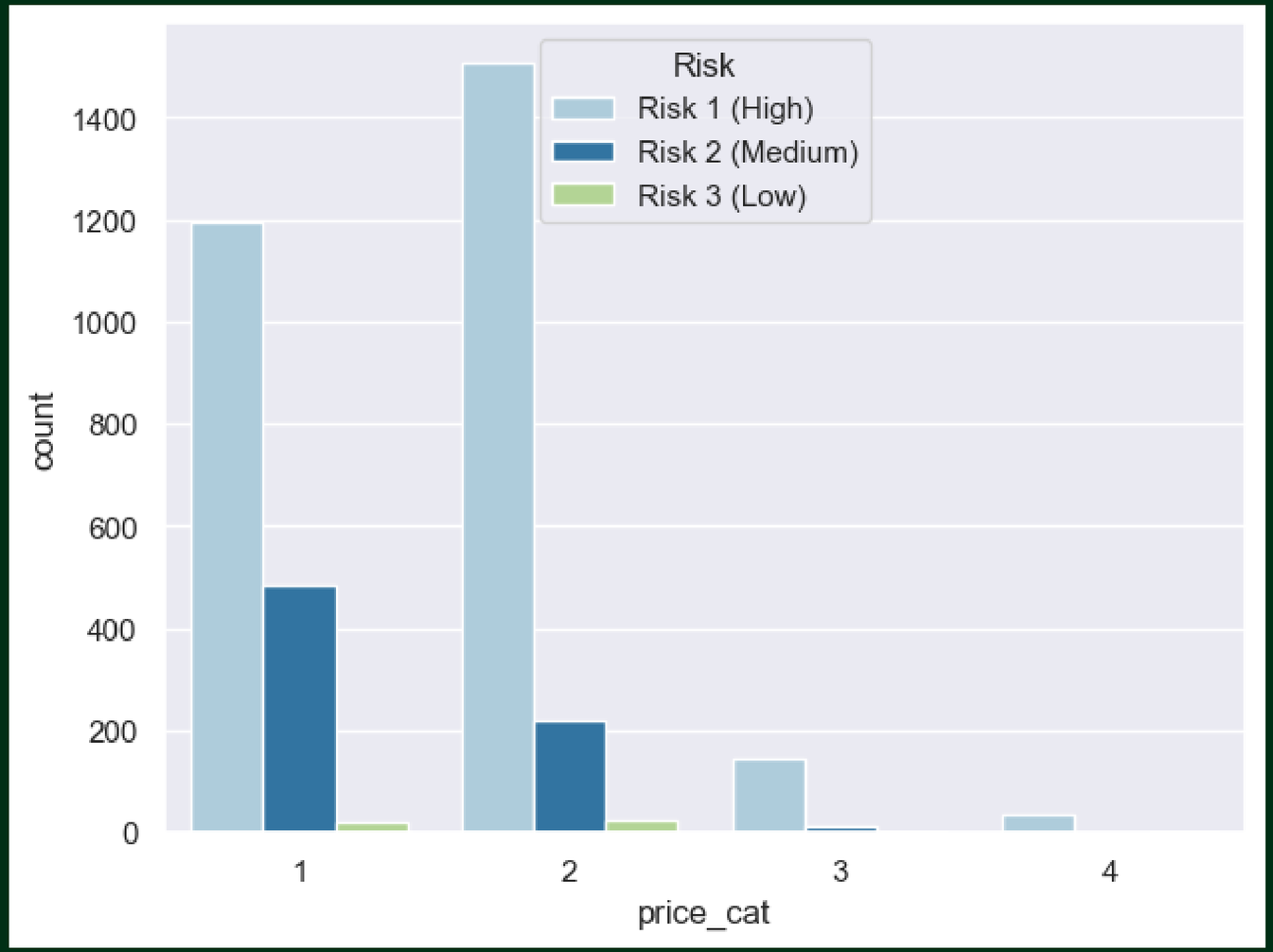


Income_level



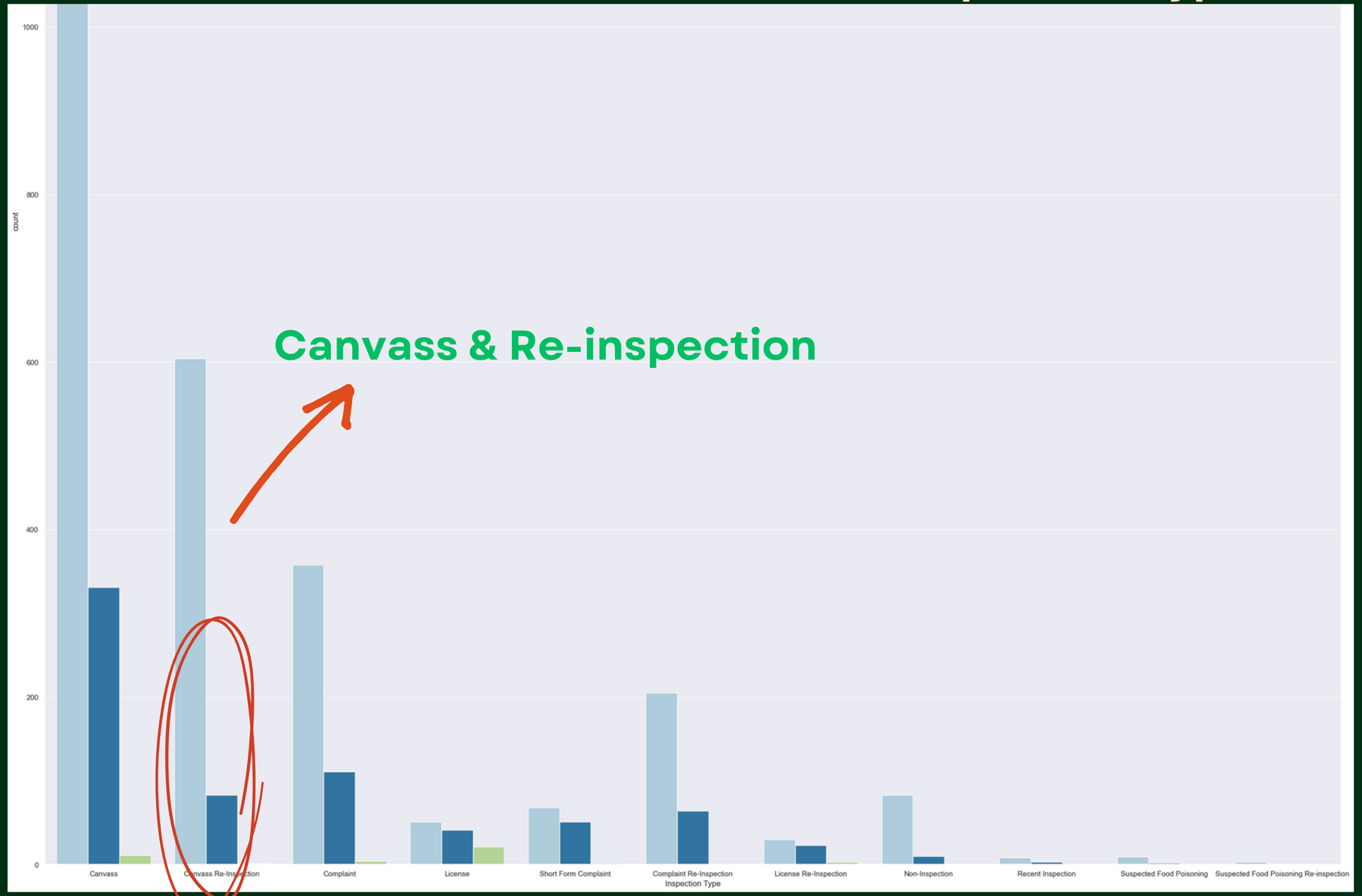
Results

Price Level



Results

Inspection Type

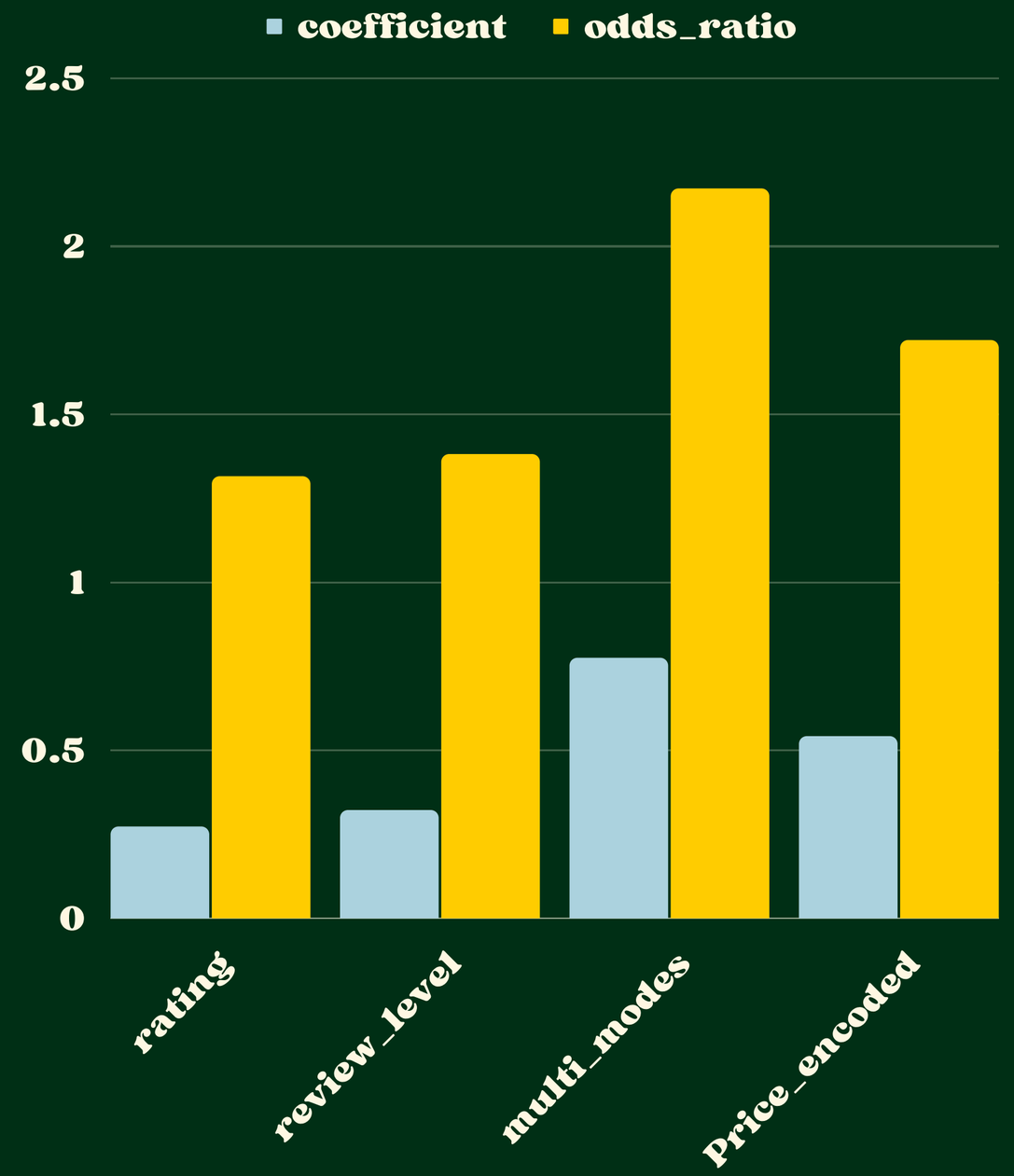


Yelp Ordered Logistic Regressin

Target Variable: **Risk**

	coefficient	odds_ratio
rating	0.273096	1.314027
review_counts	0.322022	1.379916
multi_modes	0.774915	2.170408
Price_encoded	0.541875	1.719228

Odds Ratio:
1 unit increase in X will result in
(exp(coefficient) - 1)*100 percent increase in
the odds of Y outcome



Income Ordered Logistic Regressin

Target Variable: **Risk**

income_household_median

OrderedModel Results

Dep. Variable:	risk_num	Log-Likelihood:	-2028.8
Model:	OrderedModel	AIC:	4064.
Method:	Maximum Likelihood	BIC:	4082.
Date:	Tue, 28 Feb 2023		
Time:	22:02:16		
No. Observations:	3644		
Df Residuals:	3641		
Df Model:	3		

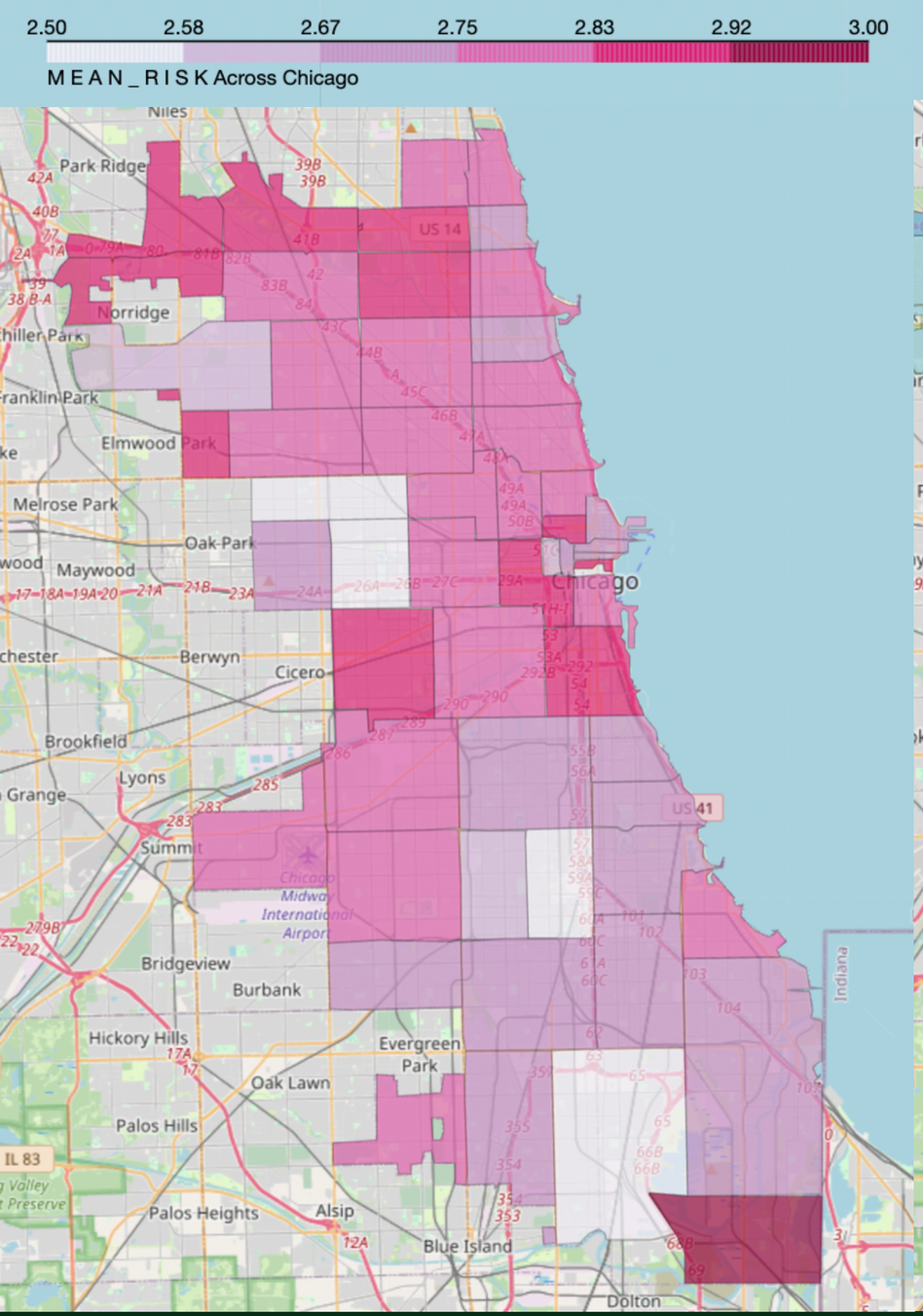
	coef	std err	z	P>|z|	[0.025	0.975]
income_household_median	1.761e-06	1.42e-06	1.239	0.215	-1.02e-06	4.55e-06
0.0/0.5	-4.3133	0.194	-22.287	0.000	-4.693	-3.934
0.5/1.0	1.1379	0.049	23.356	0.000	1.042	1.233

income_level

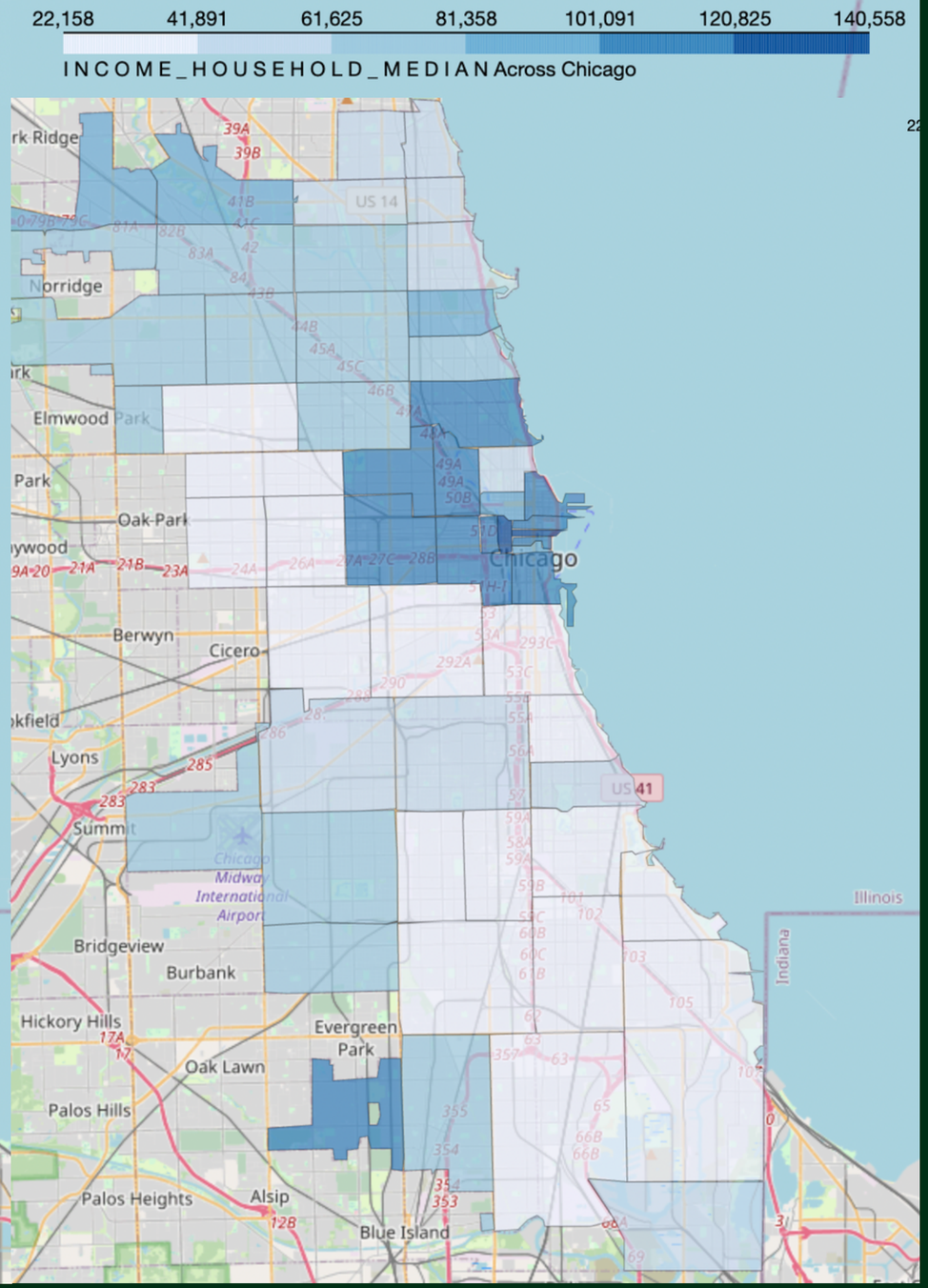
OrderedModel Results

Dep. Variable:	risk_num	Log-Likelihood:	-2029.1
Model:	OrderedModel	AIC:	4064.
Method:	Maximum Likelihood	BIC:	4083.
Date:	Tue, 28 Feb 2023		
Time:	22:03:16		
No. Observations:	3644		
Df Residuals:	3641		
Df Model:	3		

	coef	std err	z	P>|z|	[0.025	0.975]
income_level	0.0555	0.050	1.113	0.266	-0.042	0.153
0.0/0.5	-4.3422	0.183	-23.679	0.000	-4.702	-3.983
0.5/1.0	1.1378	0.049	23.362	0.000	1.042	1.233



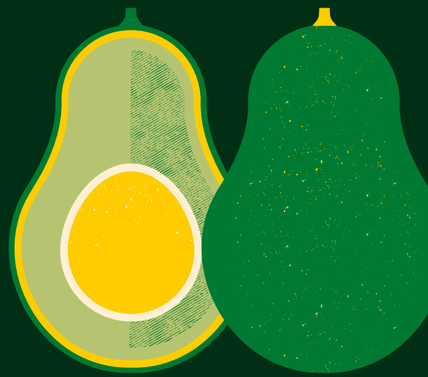
Mean Risk



Income_Household_Median

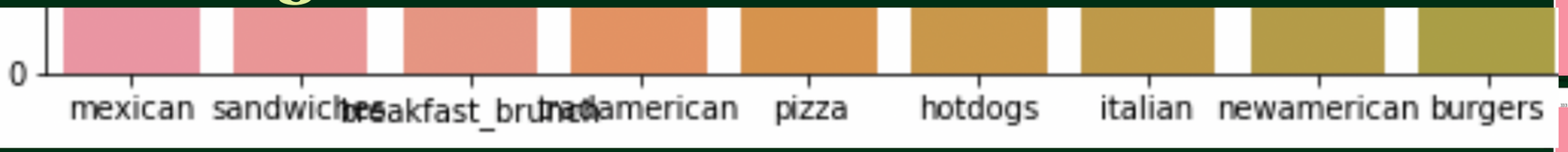
Results

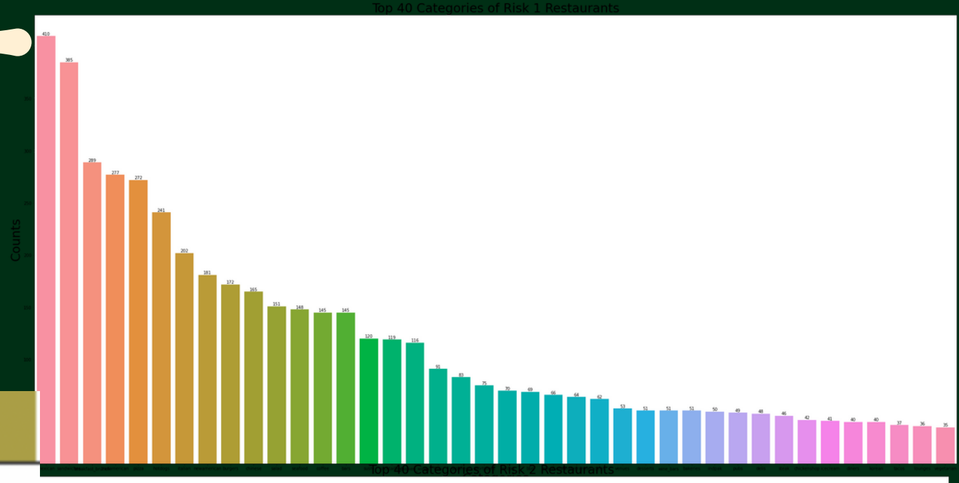
- **Risk and income_household_median**



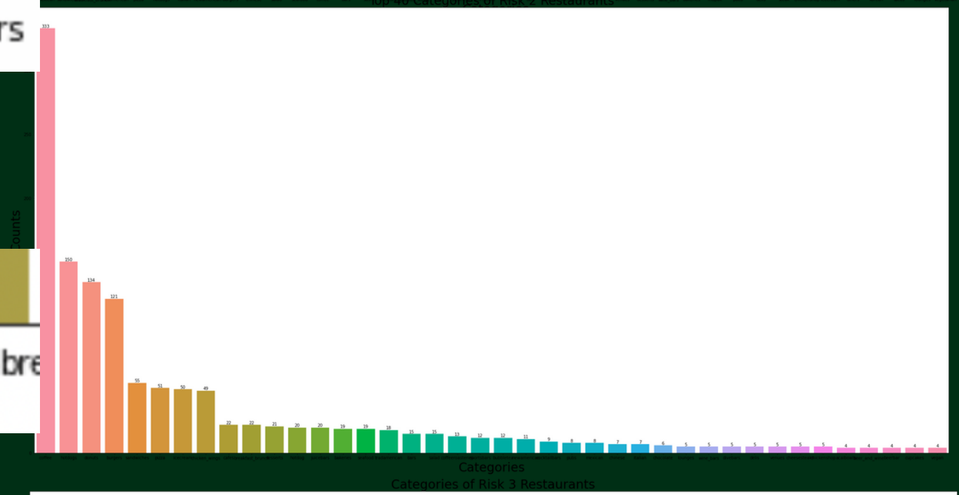
Categories 211 categories

Risk 1 (High):

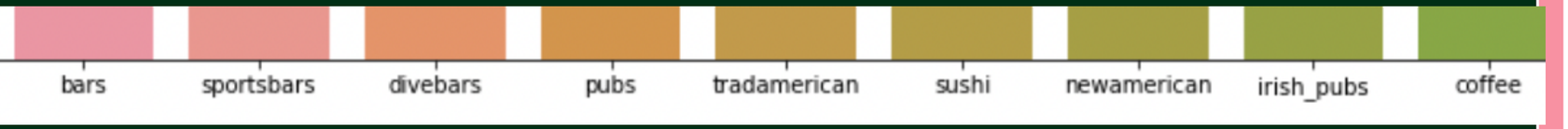


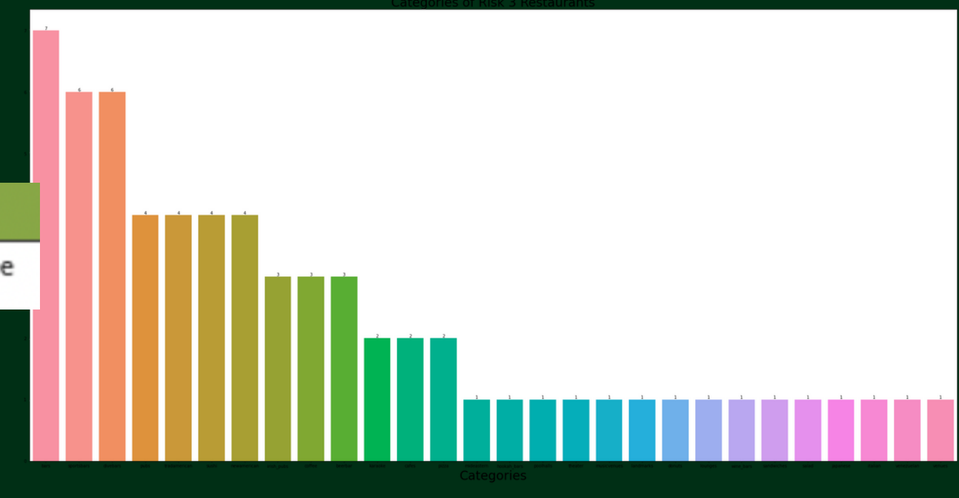


Risk 2(Medium):



Risk 3 (Low):





Risk is also related to categories of the business

Take-Home Points

Implications

1/It may be because ratings means more popularity of the restaurant, thus more difficult to keep sanitary.

2/Multi-mode operation is more difficult to keep sanitary.
（Many restaurants offer pickup options together with the delivery service.

3/More review count may also mean more popularity of the restaurant, so more difficult to keep sanitary.

Take-Home Points

Implications

4/ The re-inspection usually means a re-assessment of a restaurant's risk level.
It could be more rigid in the procedure, which may correlate with the result of a high-risk level.

Take-Home Points

For future improvements

- The number of restaurants in the data is **not evenly distributed** across zip codes, so the results may be more meaningful for economically active regions.
- In this project, we only analyzed the risk level but not the **different types of violations**, which may give us more insights into food safety.

References

- Gould, L Hannah et al. “Contributing factors in restaurant-associated foodborne disease outbreaks, FoodNet sites, 2006 and 2007.” Journal of food protection vol. 76,11 (2013): 1824-8. doi:10.4315/0362-028X.JFP-13-037
- Qiao, F. (2018, August 27). Visualizing data at the ZIP code level with folium. Medium. Retrieved February 26, 2023, from https://towardsdatascience.com/visualizing-data-at-the-zip-code-level-with-folium-d07ac983db20
- Russell Falcon, Nexstar Media Wire. “You Need to Make This Much to Be Considered 'Middle Class' in These Cities.” WGN Radio 720 - Chicago's Very Own, WGN Radio 720 - Chicago's Very Own, 7 Jan. 2023, wgnradio.com/news/you-need-to-make-this-much-to-be-considered-middle-class-in-these-cities/.
- Team, The Healthline Editorial. “17 Of the Worst Foodborne Illness Outbreaks in U.S. History.” Healthline, Healthline Media, 5 Oct. 2018, www.healthline.com/health/worst-foodborne-illness-outbreaks.

THANK YOU