Final Project Report

Project Goal:

Our project goal is to discover what attributes are associated with a restaurant's risk level(sanitary level). We consider restaurant's sanitary level is worthy for researchers because: CDC estimates "48 million people get sick, 128,000 are hospitalized, and 3,000 die from foodborne diseases each year in the United States" and "Over half of all foodborne disease outbreaks reported to the Centers for Disease Control and Prevention are associated with eating in restaurants or delicatessens"(Gould et al.).

We believe this goal can profit from various social science perspectives:

1. Consumer health and safety: The research findings can help determine if consumers can trust the information provided on the Yelp platform regarding the sanitary conditions of restaurants. This information is crucial for consumers to make informed decisions about where they dine and can help improve overall food safety.

2. Industry standards: The results of this research can also help set industry standards for restaurants in terms of maintaining proper sanitary conditions and accurately representing them on online platforms.

3. Improved business practices: The research can also help restaurants improve their business practices by highlighting the importance of maintaining good sanitary conditions and accurately representing them to customers.

4. Public health policy: The findings of the research can inform public health policies and regulations, particularly in regard to food safety and the accuracy of information provided to consumers.

5. Knowledge advancement: This research can also contribute to the body of knowledge in the fields of public health, food safety, and consumer behavior by providing insights into the relationship between restaurant sanitary conditions and the information provided on Yelp.

Overall, researching the relationship between restaurant sanitary conditions and their information on the Yelp platform can have a positive impact on consumer health and safety, industry standards, business practices, public health policy, and the advancement of knowledge in relevant fields.

Research Questions:

1. How are income level segmentation(zip codes) and a restaurant's risk level(sanitary level) correlated？
2. How are the features of restaurants correlated with their risk level?

(We select these features of restaurants according to the result of our Exploratory Data Analysis. The hypothesis section and Exploratory Data Analysis sections will provide details of "features of restaurants")

Hypothesis:

1. There is a positive correlation between risk level and median household income level.
2. There is a positive correlation between risk level and price, rating, transactions, and review counts.

- Dependent variable: (restaurant risk level)
- Independent variables: restaurant features: 'price', 'rating', 'transactions'(for 'pickup' and 'delivery' particularly), 'income_household_median'

Important Definitions:

1. risk level(high, medium, low): it is the restaurant sanitary level that is regularly assessed by the Chicago Department of Public Health. The criteria include good hygienic practices, protection from contamination, employee health, etc.(This info is provided by: https://www.chicago.gov/city/en/depts/cdph/provdrs/food_safety/svcs/understand_healthcoderequirementsforfoodestablishments.html)

2. Income level segmentation: We group restaurant locations according to their income level.(Income level will be explained more during the exploratory data analysis section)
3. Differences between high risk and low risk(for getting a sense of risk level differences): high risks mean main concerns are about general sanitation(e.g.: "Baking of non-potentially hazardous food "), comparing low risks with main concerns for room temperature(less affecting sanitary).
4. Re-inspection: Re-inspection occurs when there are specific areas of the inspection criteria that need to be checked(decided case-by-case). It happens if staff need to be ensured that some violations are fixed by the restaurant.

Data Sources:

Data Sources 1(downloading)**:**

https://data.cityofchicago.org/Health-Human-Services/Restaurant/5udb-dr6f

- time frame: Year of 2023
- Dataset size: 156379*16
- We collected business name, food risk level, inspection date, Inspection Type, Results, Violations, Latitude, Longitude, Location. The food risk level is the main target variable here because we will see how other features correlate with the risk level. This data is from cityofchicago.org. As an official government website, cityofchicago.org is a primary source of information about the city and its various departments and initiatives. This makes it a reliable source of information that has been verified and validated by the city government. Therefore, we believe it has little reliability/validity issue.

Data Sample:

DBA Name	AKA Name	License #	Facility Type	Risk	Address	City	State	Zip	Inspection Date	Inspection Type	Results	Violations	Latitude	Longitude	Location
LA MICHOACANA ICE CREAM SHOP	LA MICHOACANA ICE CREAM SHOP	2698396	Restaurant	Risk 1 (High)	3591-3597 N MILWAUKEE AVE	CHICAGO	IL	60641	02/14/2023	Canvass	Fail	38. INSECTS	41.946140053442800	-87.735183019952700	(41.946140053442802, -87.735183019952704)

Data Sources 2(downloading)**:**

https://simplemaps.com/city/chicago/zips/income-household-median

- time frame: Not Sure
- Dataset size: 60*2

- This dataset contains two features, zip code and median household income in USD.We use this dataset as a reference to divide zip codes into different income levels so that we can use zip codes as a measurement of income to see if this is a feature that affects the restaurant food safety level. Even though the data source does not provide the exact time frame, based on Chicago's economic development these years, we think it is reliable to use it to do income stratification.

Data Sample:

zip	income_household_median
60018	64429

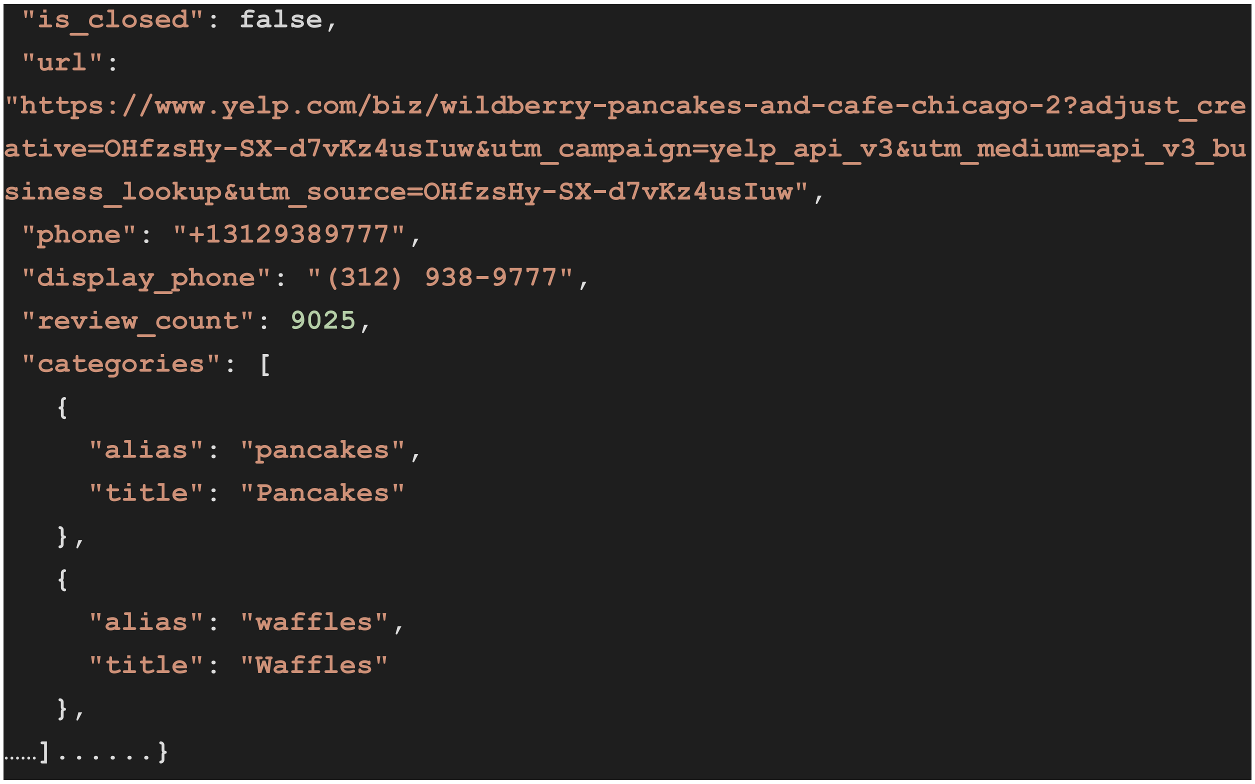
Data Sources 3(Scraping**):**
https://docs.developer.yelp.com/docs/fusion-intro

- time frame: Year of 2023 (Data scraping date: Feb 10, 2023)
- Dataset size: 39,158 raw data, 13,159 data after deduplication (16 features)
- This dataset contains information about restaurants in Chicago; the features we will use our name, reviews, is_closed, review_count, categories, rating, location, zip_code, price, hours, and transactions. Yelp.com has a large and diverse user base that provides a wealth of information about customer behavior and opinions. This data can be analyzed to uncover patterns and trends in customer behavior that can be used to inform social science research. Yelp.com has measures in place to verify the authenticity of reviews, such as using algorithms to detect fake or biased reviews. This helps to ensure that the data is reliable and trustworthy. We will use that information to reflect Chicago's restaurant information and presumably link this with the food safety level.
- Due to limitations of the Yelp API, we scraped data based on zip codes corresponding to data in Data Sources 2.

Data Sample:

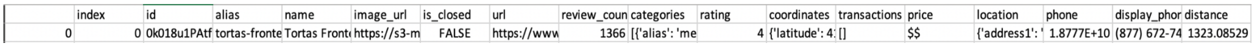
Scraped data structure:

```
{
 "id": "cKZNbMvoqJaUe7n6lf6i7w",
 "alias": "wildberry-pancakes-and-cafe-chicago-2",
 "name": "Wildberry Pancakes and Cafe",
 "image_url":
"https://s3-media2.fl.yelpcdn.com/bphoto/43XNyVUbPJNtC6IFobGRMw/o.jpg",
 "is_claimed": true,
```



```
 "is_closed": false,
 "url":
"https://www.yelp.com/biz/wildberry-pancakes-and-cafe-chicago-2?adjust_cre
ative=OHfzsHy-SX-d7vKz4usIuw&utm_campaign=yelp_api_v3&utm_medium=api_v3_bu
siness_lookup&utm_source=OHfzsHy-SX-d7vKz4usIuw",
 "phone": "+13129389777",
 "display_phone": "(312) 938-9777",
 "review_count": 9025,
 "categories": [
   {
     "alias": "pancakes",
     "title": "Pancakes"
   },
   {
     "alias": "waffles",
     "title": "Waffles"
   },
……] . . . . . . }
```

CSV file data sample:



	index	id	alias	name	image_url	is_closed	url	review_coun	categories	rating	coordinates	transactions	price	location	phone	display_phor	distance
0	0	0k018u1PAtf	tortas-fronte	Tortas Front	https://s3-m	FALSE	https://www	1366	[{'alias': 'me	4	{'latitude': 4	[]	$$	{'address1': '	1.8777E+10	(877) 672-74	1323.08529

Data preprocessing and linking

After we scraped Yelp data via its official API Yelp Fusion, and download other necessary data from sources like Chicago Data Portal (as shown in the "data sources" section), we did some fundamental data preprocessing:

1. Dealing with missing values:
   After we checked missing values for datasets from the three sources if the missing variables are crucial for linking (e.g., restaurant address), we chose to drop them. If the missing variables are not that important (e.g., 'AKA names,' 'violations'), we simply fill the missing values with the string "missing."

2. Drop duplicate rows:

   For each dataset, we drop duplicate rows based on their unique id, which is provided by the original data.

3. We standardize the address of each restaurant; for example, we convert the address of the restaurant to lowercase. We use the clean function to clean addresses, cities, and states.

4. We extract the address, city, state, and zip code from ‘location’ in the yelp data file ‘dedup_yelp.’
5. We also only keep up-to-date records and drop others. In the government health inspection dataset, some restaurants’ results show 'Out of business' or ‘'Business not located. Correspondingly, their most recent inspection dates are also shown as years ago. To make the two datasets (Yelp and government) have the same timing, we decide to drop those out-of-dates.
6. **Linking Data:**

   For our project, since we got enough data when linking the two main datasets, we prioritize accuracy, which means we only want an exact match to keep a high matching quality.

   For our linking method, we use address, city, and state to link Yelp and Chicago government data to form our main data frame. And then, we use zip code to link income household median with the data frame.

7. Feature Engineering:

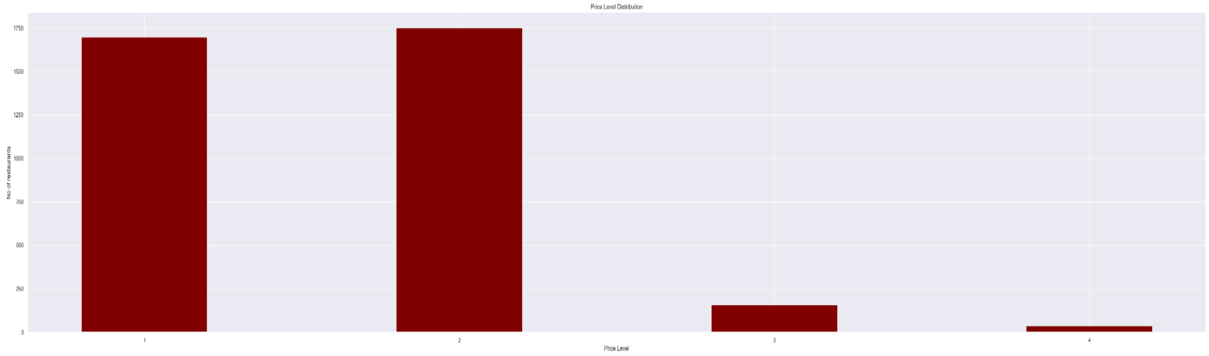
   To facilitate the data analysis process, we further deal with our variables. For our categorical variables, we focus on doing one-hot encoding or original encoding (depending on whether the categories are in a meaningful order). For hard-to-deal-with variables, such as ‘categories’ (with a complex structure and too many types), and ‘violations’(contains long and complex texts), we separate them and do analysis separately.

Exploratory Data Analysis(EDA) & Visualization:

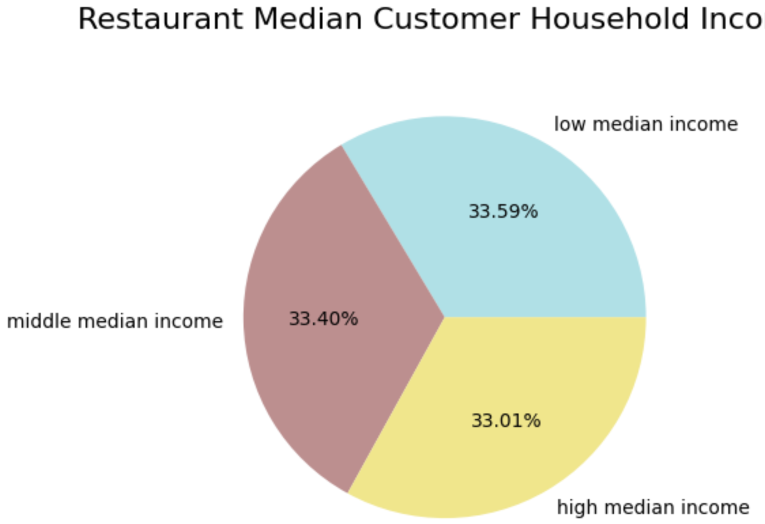
1. Dataset description:
   After the fundamental data preprocessing, the size of our final data is 3644 rows × 36 columns, which we regard as an appropriate size for our further exploration and analysis.

2. Variables Exploration
   In this part, due to space limitation, we will display a part of our visualization.
   1/ 'Price': This is a categorical variable that contains four categories: ‘$’, ‘$$’, ‘$$$’, and ‘$$$$’.According to Yelp, ‘$’ is for average price per person under $10, and ‘$$$$’ is for average price per person above $61.(For the bar graph below, we change dollar signs to 1, 2, 3, and 4, so that 1 represents ‘$’ and 4 represents ‘$$$$’. It might be too small to see,

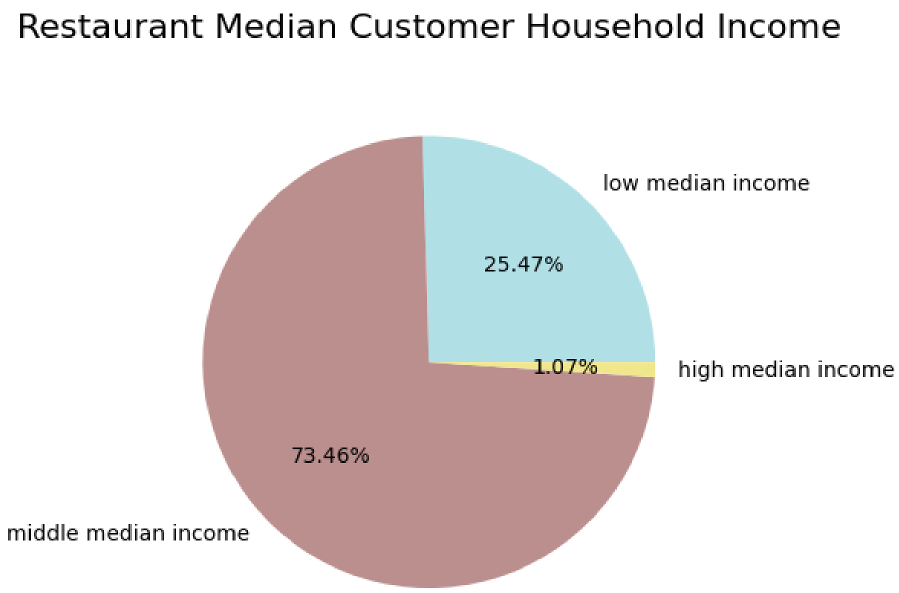
but the first two columns represent 1 and 2. Thus, most restaurants are in low price level for average per person.



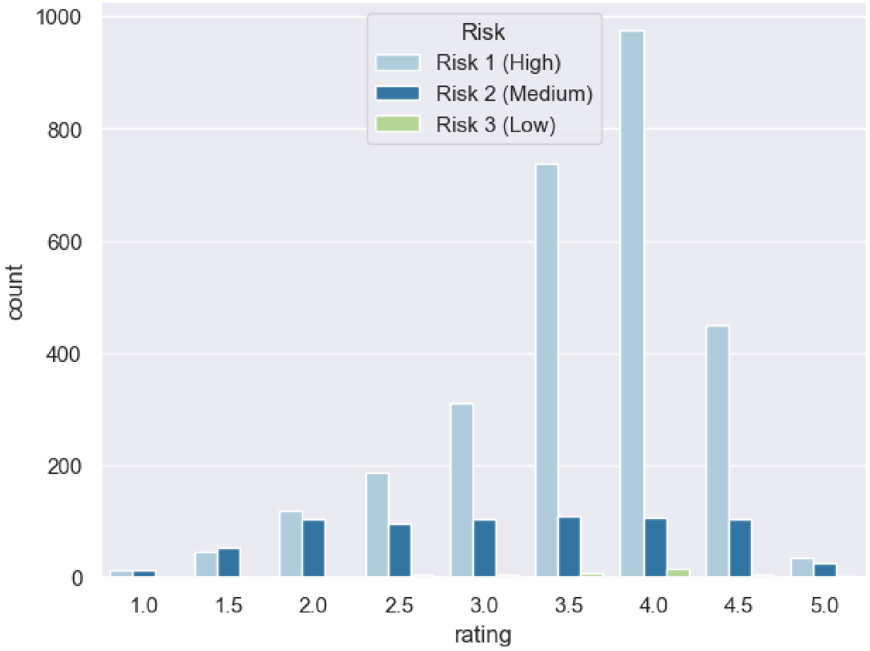
2/ 'income_household_median': This is a numerical variable that is for the median of the restaurant's location's household income. For segmenting restaurant locations by income level, we decided to use three categories, low, middle, and high. There are two ways to set the middle-income level: Classification one is based on our data, so we did data describing this variable and evenly divided it into three income levels. The middle level is for any value larger than 61625 and smaller than 101091.



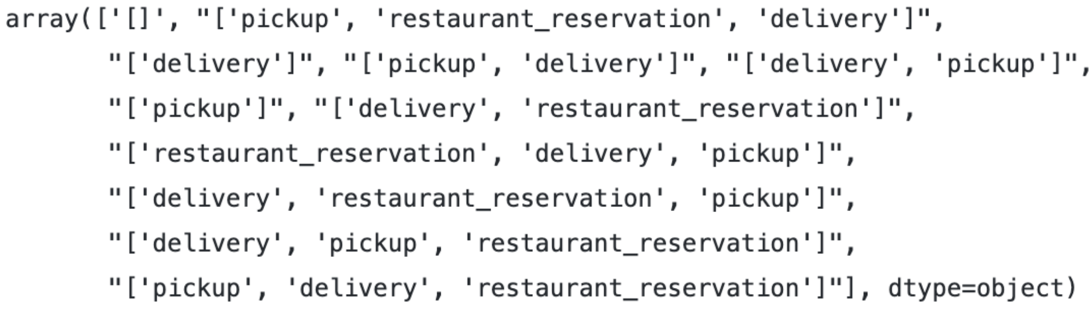
Second way is to use the US. Census Bureau's criteria in 2021. The middle income is for any value larger than 52000 and smaller than or equal to 156000. Then the low income level is below this range, high income is above this range, we can see that more than 70% of our data belongs to the middle income group.



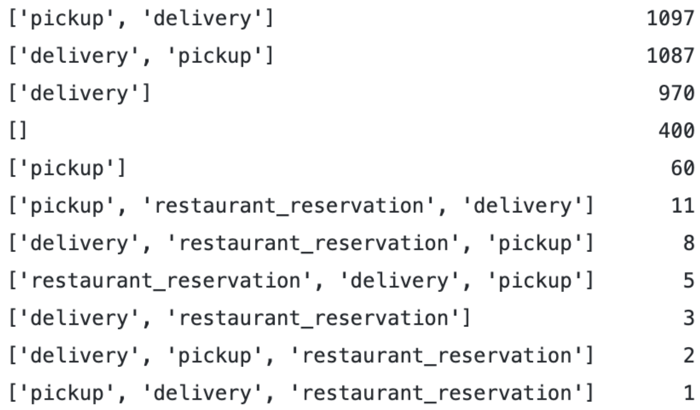
3/'rating': It is a categorical variable that has values: 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5. The rating implies the lowest satisfaction will be 1, and the rating that implies the highest satisfaction will be 5. We can see the interaction between the dependent variable of risk level and the independent variable of rating below: It seems like the number of restaurants in the high-risk level increases gradually as the rating increases. However, we can see that the number of restaurants in high-risk levels decreases as the rating approaches 4.0. It needs to be stressed that such a graph does not explain the correlation, we still need to use regression models for correlation conclusions.



4/ 'transaction': Restaurants may provide various types of services. It is a categorical variable that contains the combination set of either of the three: 'pickup ', 'delivery ', and 'restaurant reservation '. We can see that restaurants are in one of the categories of transaction-type combination set by:



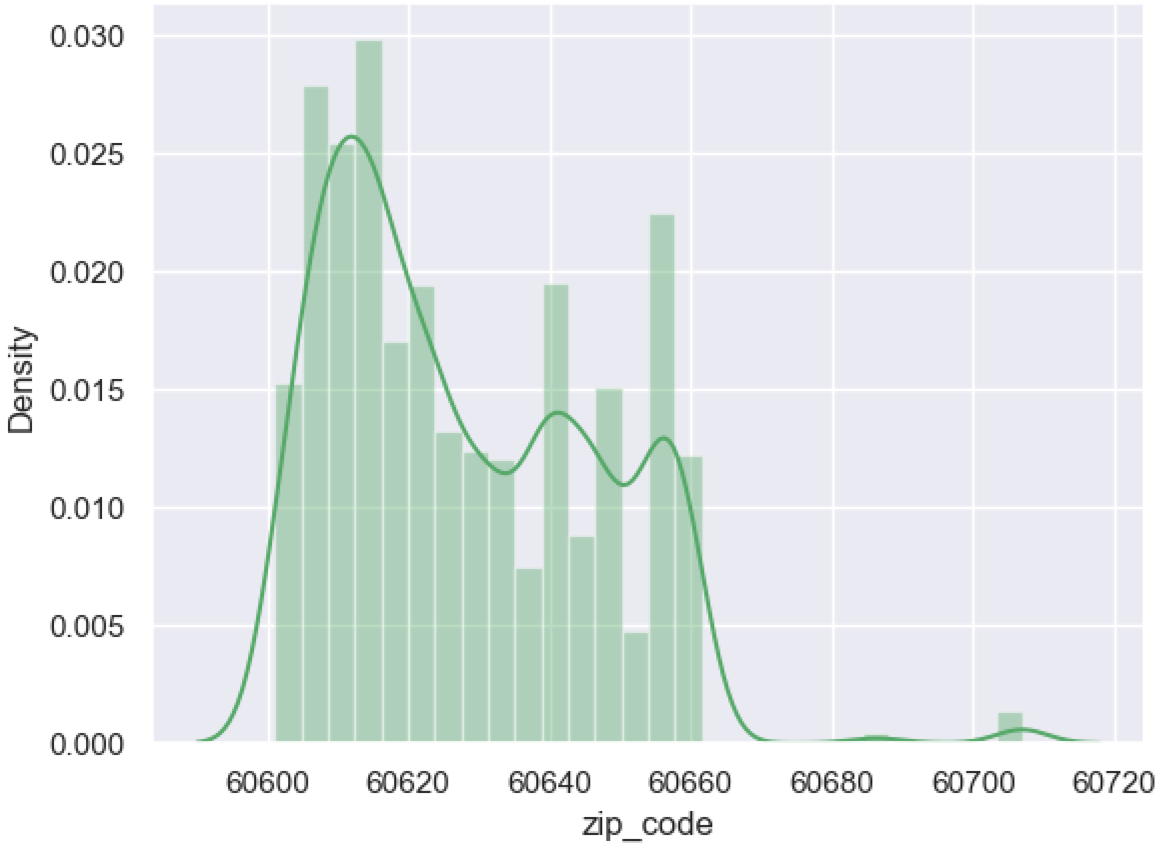
```
array(['[]', "['pickup', 'restaurant_reservation', 'delivery']",
       "['delivery']", "['pickup', 'delivery']", "['delivery', 'pickup']",
       "['pickup']", "['delivery', 'restaurant_reservation']",
       "['restaurant_reservation', 'delivery', 'pickup']",
       "['delivery', 'restaurant_reservation', 'pickup']",
       "['delivery', 'pickup', 'restaurant_reservation']",
       "['pickup', 'delivery', 'restaurant_reservation']"], dtype=object)
```



```
['pickup', 'delivery']                              1097
['delivery', 'pickup']                              1087
['delivery']                                         970
[]                                                   400
['pickup']                                            60
['pickup', 'restaurant_reservation', 'delivery']      11
['delivery', 'restaurant_reservation', 'pickup']       8
['restaurant_reservation', 'delivery', 'pickup']       5
['delivery', 'restaurant_reservation']                 3
['delivery', 'pickup', 'restaurant_reservation']       2
['pickup', 'delivery', 'restaurant_reservation']       1
```

We can see that many restaurants offer delivery and pickup services.

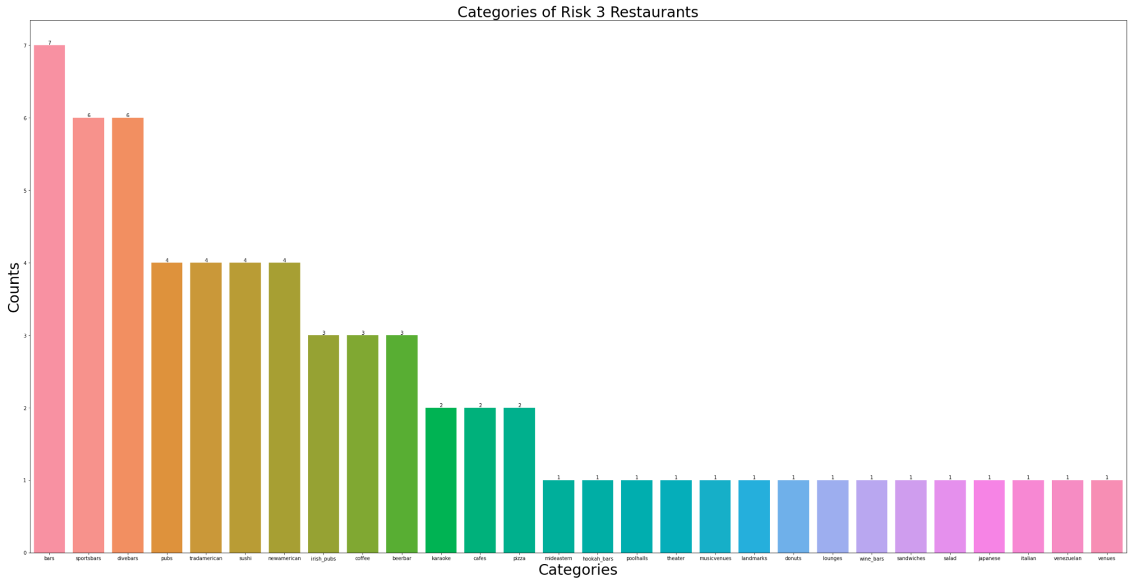
5/ 'zip_code ': it is also a continuous variable representing the zip code location of restaurants. Below is the distribution of zip_code:



We can see that Chicago restaurants are mostly clustered in zip_code[60600, 60660].
Density in distribution plot meaning:

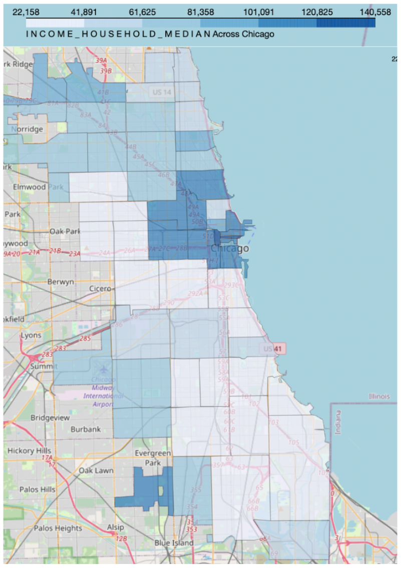
“A density plot can be seen as an extension of the histogram. As opposed to the histogram, the density plot can smooth out the distribution of values and reduce the noise. It visualizes the distribution of data over a given period, and the peaks show where values are concentrated.” (Synergy Codes, 2022)

6/ 'categories ':
Since there are about 200 types of category labels in our dataset, we decide to analyze this variable separately. We use histograms to visualize this variable. For example, for risk 3 (low risk) business:



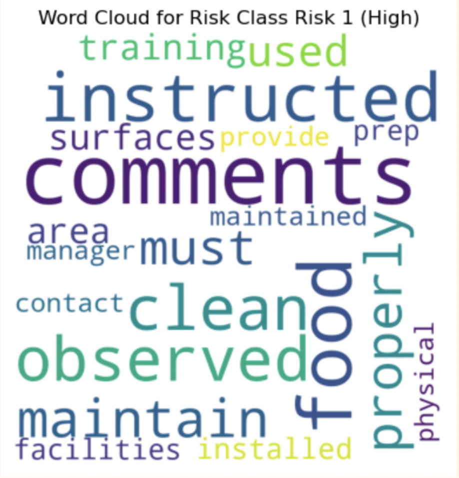
From these top labels for risk 3 business, we can conclude that since bars and pubs barely serve hot foods, it is reasonable that they are easier to keep sanitary conditions at a good level.

We also tried other visualization libraries and methods, for example:
7/ 'income_household_median ':



We use the folium library to plot maps to more directly visualize many of our variables. The example above is the map of the distribution of the Chicago income household median. We can clearly observe that the loop and north area show a higher economic level.

8/ 'violations':

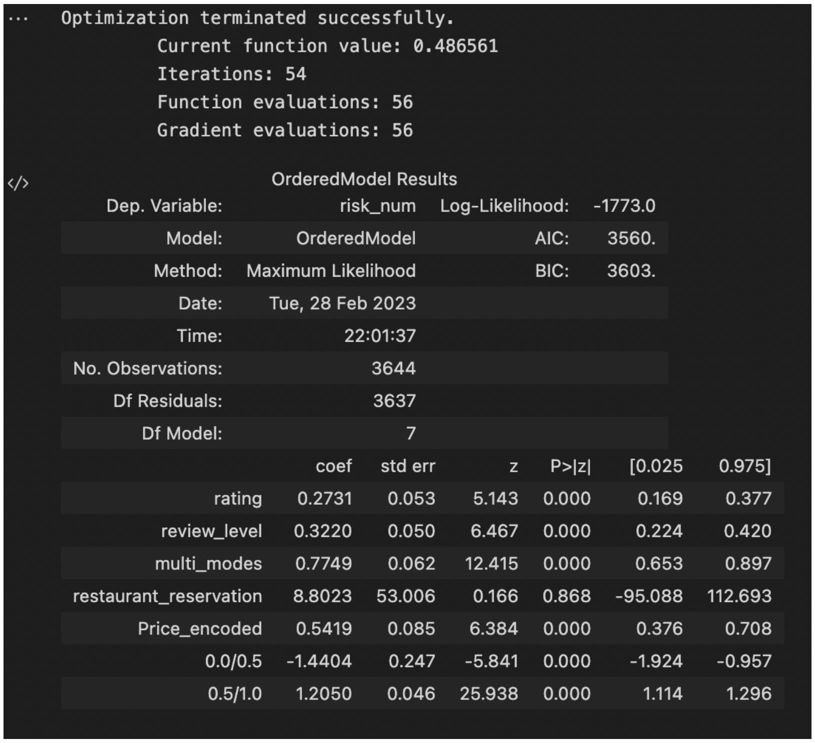


We also visualize the ‘violations’ variable using the library word cloud based on different risk levels. For example, from the above word cloud for risk 1 violation, we can see that risk 1 sanitary issues are usually related to physical contact and surface cleanliness maintenance. PS: We provided the above examples of EDA and related visualizations; more are present in the Jupyter Notebook.

Models:

To further answer our research questions and test our hypothesis, we run a regression model. Since our dependent or target variable is the risk level of each business, which is categorical and in a meaningful order, we choose to do the ordered logistic regression.

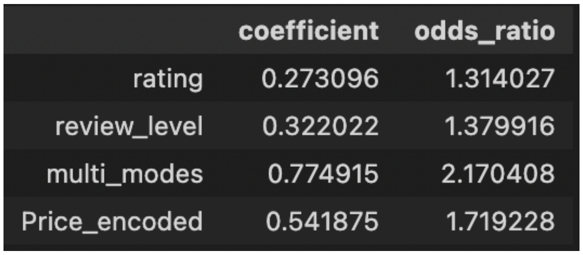
Result:



```
...   Optimization terminated successfully.
               Current function value: 0.486561
               Iterations: 54
               Function evaluations: 56
               Gradient evaluations: 56

                              OrderedModel Results
       Dep. Variable:             risk_num   Log-Likelihood:  -1773.0
              Model:          OrderedModel              AIC:    3560.
             Method:    Maximum Likelihood              BIC:    3603.
               Date:      Tue, 28 Feb 2023
               Time:              22:01:37
   No. Observations:                  3644
       Df Residuals:                  3637
           Df Model:                     7
                             coef   std err        z   P>|z|    [0.025    0.975]
                  rating   0.2731     0.053    5.143   0.000     0.169     0.377
            review_level   0.3220     0.050    6.467   0.000     0.224     0.420
             multi_modes   0.7749     0.062   12.415   0.000     0.653     0.897
 restaurant_reservation    8.8023    53.006    0.166   0.868   -95.088   112.693
           Price_encoded   0.5419     0.085    6.384   0.000     0.376     0.708
                 0.0/0.5  -1.4404     0.247   -5.841   0.000    -1.924    -0.957
                 0.5/1.0   1.2050     0.046   25.938   0.000     1.114     1.296
```

First, we do a regression with risk as the dependent variable and multiple yelp data as independent variables (rating, review_level, multi_modes, restaurant_reservation and price_encoded). The regression result is shown above. We can observe that, rating, review_level, multi_modes, and price_encoded are statistically significant with a p value equals to zero. To further interpret their coefficients, we transform them into odds ratios:

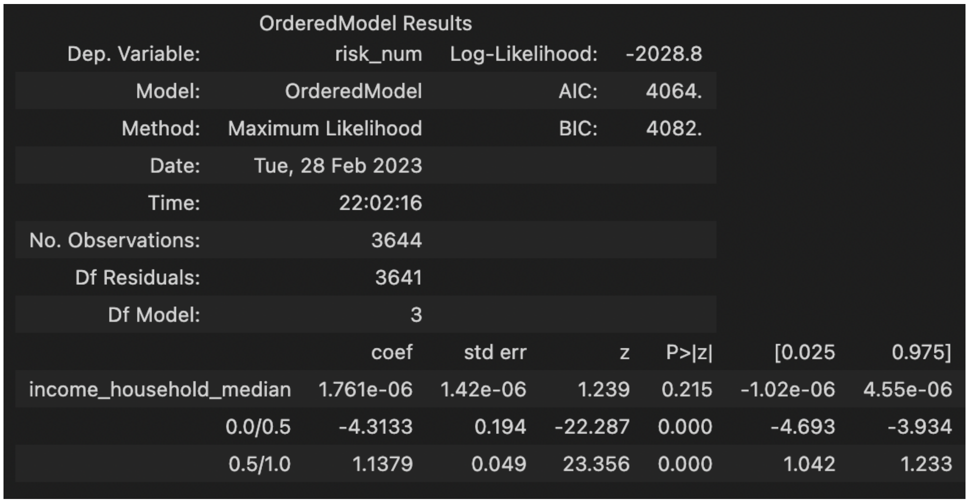


	coefficient	odds_ratio
rating	0.273096	1.314027
review_level	0.322022	1.379916
multi_modes	0.774915	2.170408
Price_encoded	0.541875	1.719228

Odds ratio are more direct and could be interpreted using the formula: 1 unit increase in X will result in (exp(coefficient) - 1)*100 percent increase in the odds of Y outcome. For example, from our result, 1 unit increase in rating will result in about 31.4% increase in the odds of risk outcome.

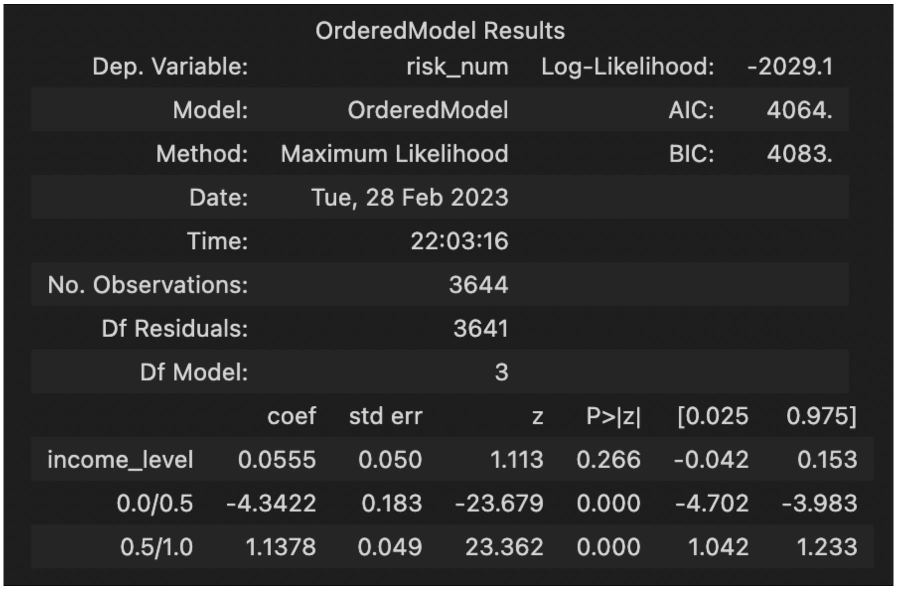
Then to test our first hypothesis, we also conduct a ordered logistic regression with risk as the dependent variable and income_household_median as the independent variable. Since the income data comes from a completely different souce from Yelp, we decide not to put them in the same regression, but to examine the correlation separately.

Result:
We represent the income_household_median in two ways. The first way is to use its original value as the independent variable:



```
                          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
```

The second way is to use the three levels that we categorized, as mentioned previously in the Exploratory Data Analysis part:



```
                        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
```

From both regression results shown above we can see that neither income household median nor the income level we categorized are statistically significant. Hence we can conclude that the variable or feature of income household median does not show a correlation with restaurants' risk level.

Conclusions:

1. Restaurants' rating, review, operation modes and price level all have statistically significant positive correlation with our target variable: restaurants' risk levels. Among these features, whether the restaurant is operated in multi modes shows the highest correlation, with a coefficient of 0.775 and odds ratio of 2.170. We can conclude that whether a restaurant is operated in multi modes (offering pickup and delivery services) is the most important feature to impact its risk level.
   The result supports our second hypothesis.
2. Income household median does not have a correlation with restaurants' risk level. Our first hypothesis is not supported.

Take-home points:

- Implications from our conclusions:

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

- Possible future improvements:

1. 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.
2. 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:

- "Burden of Foodborne Illness: Findings." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 5 Nov. 2018, www.cdc.gov/foodborneburden/2011-foodborne-estimates.html.
- 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.
- Synergy Codes. (2022, May 24). *What is a density plot? definition, importance, and examples – glossary*. Synergy Codes. Retrieved March 5, 2023, from https://synergycodes.com/glossary/what-is-density-plot/