COMPLETE DATA SPEC

In our project, we are analyzing the restaurant scene in New York City. We are using three main sources of data:

* Median Incomes of different neighborhoods in NYC from the Citizens’ Committee for Children.
* A Kaggle dataset that has been made by scraping Google to get information about Prices, Reviews, and Ratings of restaurants in NYC.
* A dataset from DOHMH that contains health inspection reports of various restaurants in NYC

The above datasets are all CSV files. We extract them into panda data frames using the pandas’ inbuilt read\_csv() function and then perform our cleaning and combining. We also did a bit of cleaning on Excel to get from the raw to the clean versions of each CSV.

At first, we tried combining all three datasets, but very quickly realized that combining the last two datasets would be very tedious as they are not indexed by any similar keys. We tried to index them by their name, but there were different spellings across the two or they had additional words like bar and restaurant, etc. We tried to index them by address, but they had different addresses. For example, one restaurant had 95 W Broadway in the Kaggle dataset and 113 Reade St in the DOHMH one – both these addresses refer to the same lot. We even tried Latitude and Longitude and that did not work. We finally tried to manually match the two on Excel but punching in the values took way too long. It took about an hour just for 15 entries as we had to aggregate all the violations per restaurant and match it with the columns from the other dataset.

We finally made the decision not to combine the two datasets as the properties we are testing are self-contained in each dataset. We want to test the correlations between price and ratings which is in the Kaggle dataset. We also want to test the relationship between neighborhood income and ethnicity of cuisine with health violations which is contained in the combination of the first and third datasets.

Since we are using two main datasets I will explain the important attributes from both:

KAGGLE DATASET:

* **NAME:** It is the name of the restaurant, and it is represented by a string. There is no default value here. These values are by and large unique. We might have chain restaurants with the same name, but the addresses are different. This is a required value. It is an important value but will not directly feature in our analysis and it is not a sensitive value
* **RATING:**  It is the rating of this restaurant on Google, and it is represented by float. These is no default value here as well. We drop rows without this value as it is an important one. The range is from 0-5. It is a required value and is used in our analysis. We are going to see how it correlates with another field namely Price Category. It doesn’t contain sensitive information as it is public.
* **PRICE CATEGORY:** Once again this is the price of going to the restaurant and it is represented by an int. Its range is from 1-4 and if the value is not present it is populated by NaN and we drop the row from our analysis. It is a required value and it is used in our price-review analysis
* **RATING COUNT:**  This is an int that captures how many reviews have been posted. It is not an essential field as we are not currently planning on using it for analysis, but nevertheless, we feel it is an important one as it gives validity to the **RATING** field which we do indeed use. The values are not unique and are usually in the range of several hundred. Some popular restaurants have 1000-2000 reviews. Not unique or sensitive. There is another field called detailed ratings that also elaborates on the ratings

SAMPLE OF THE DATAA screenshot of a computer

Description automatically generated

DOHMH Dataset:

* **DBA:** It is a string and it is the name of the Dataset. It is very much akin to the name column in the Kaggle dataset. It is required and usually unique barring chain restaurants. It is not used in analysis directly.
* **BORO:** It is a string and it takes values form a fixed list namely the boroughs in new York city. It is required and there is no default value. It is used to determine the median income of the neighborhood/borough which is used in the analysis of violations and ethnicity of restaurants.
* **CUISINE DESCRIPTION:** It is a string yet again. It has no default value. It is used to determine the ethnicity of the food. Some sample values are – Chinese, Pizza, Mexican, etc. We then will use this to determine if it is an established cuisine or an up-and-coming one or not. Values are not unique but are required and will be used to conduct analysis.
* **CRITICAL FLAG:** Binary value that takes one of two strings – critical or non-critical. It helps us understand whether the restaurant is sanitary or not depending on the kind of violation. We want to understand if there is a relationship between the affluence of the surrounding neighborhood or the ethnicity of cuisine on cleanliness standards and this is the variable that will help us do that. This is required and not unique. The information is not sensitive as it is published by the city of new York. There are many other columns that better help us understand the nature of this violation
* **2021 Median Income:**  it is a float and it represents the median income of the neighborhood. It is not unique and it is used in analysis. It is required and there is no default value.

Median Income sample

A screenshot of a computer

Description automatically generated

DOHMH sampleA screenshot of a table

Description automatically generated