CS 638 REPORT

Stage II

Team Group

Shixuan Fan: sfan33@wisc.edu

Yuting Liu: liu487@wisc.edu

Zhenyu Zhang: zzhang546@wisc.edu

PART I. List the schema of the two tables

Restaurant (Name, Phone, Rate, Price, Zip-code, State, City, Street address, Has Delivery, Has take-out)

PART II. List the attributes of the two tables

- Restaurant Name
- Phone
- Rate
- Price
- Zip code
- State
- City
- Street address
- Has delivery
- Has take-out

PART III. List the attributes of the two tables

(1) Restaurant Name

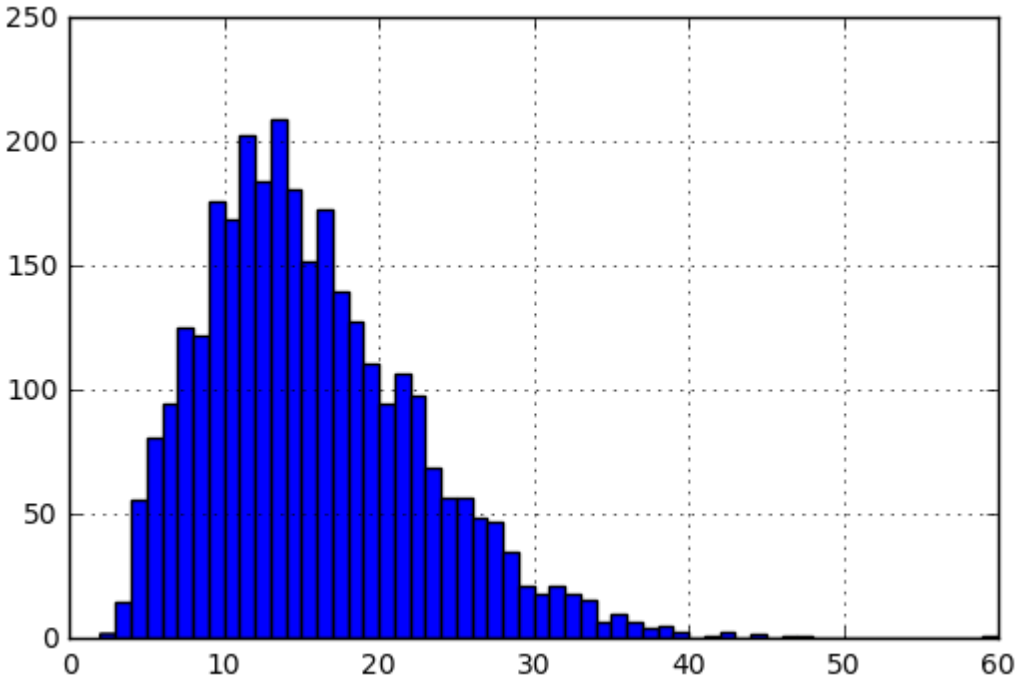
I. Missing values

- # tuples with missing value: 0

II. Classification: textual

- Average length = 15.19
- Minimum length = 60
- Maximum length = 2

III. Outliers and anomalies



- Most name has the length between 2 ~ 40. There are several restaurants with the name length between 40 ~ 50. We would still consider them as acceptable even though their lengths are longer than most of other restaurants.
- There is a restaurant with name "Yelp Elite Battle Of The Band T's @ RAISED Urban Rooftop Bar", that has the length of 60. This one would be considered at the outlier.

IV. Synonyms: NO

V. Sprinkled data: NO

VI. Other data quality problems

- There are a lot problems relevant to UTF-8 encoding when crawling data from websites. There are many special symbols in the restaurant name, sometimes there will be some French, Spanish letters which would have a UTF-8 encoding with comma, which will lead to many problems in store this information in CSV file, because it would separate different columns by ‘,’ firstly. It caused us a lot of time to reorganize the table schema.

(2) Phone

I. Missing values

- # tuples with missing value: 131/3076 (4.2588%)
- Solutions to fill in missing values
    - It’s really hard to refill the exact phone number to corresponding restaurant if we could not crawl this information from the website. Instead, we would use the area code based on its address and a default zero number to refill in this information, such as (ABC) 000-0000, where ABC is its area code.
    - It’s really hard to refill the exact phone number to corresponding restaurant if we could not crawl this information from the website. Instead, we would use the area code based on its address and a default zero number to refill in this information, such as (ABC)-000-0000, where ABC is its area code.

II. Classification: textual

- Average length = Minimum length = Maximum length = 14
- This attribute follows the certain format: (XXX) XXXX-XXXX
    - All tuples except those with missing values follow this format

III. Outliers and anomalies: NO

IV. Synonyms: NO

V. Sprinkled data: NO

VI. Other data quality problems: NO

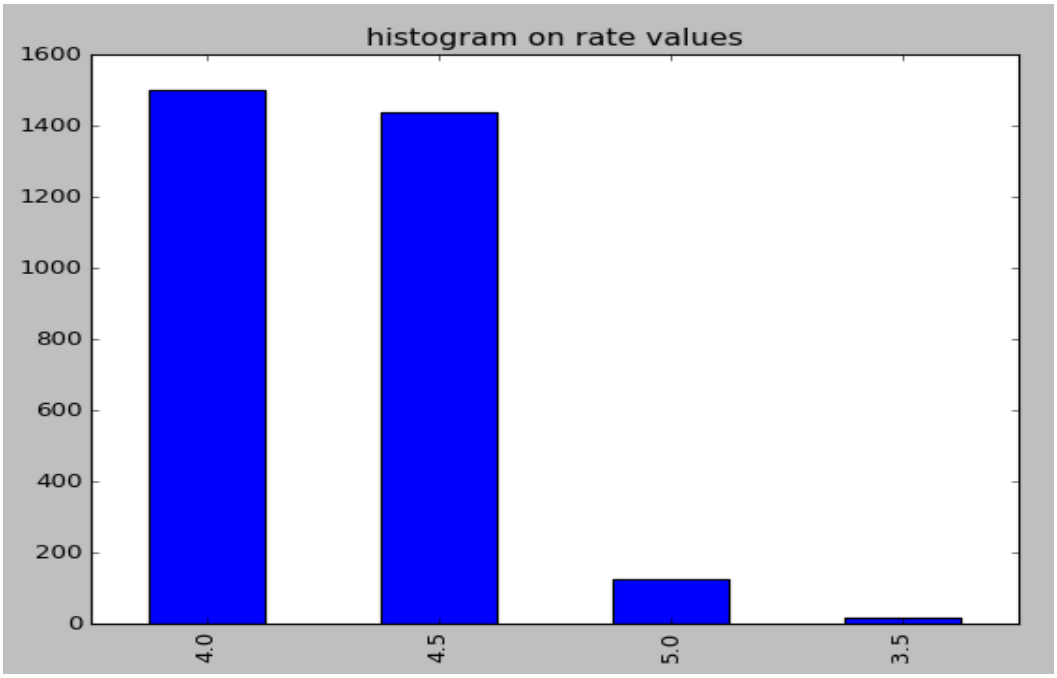
(3) Rate

I. Missing values

- # tuples with missing value: 0 → 0%

II. Classification: numerical

III. Outliers and anomalies: NO



- There will be no restaurants considered as outliers no matter what rate they are.

IV. Synonyms: NO

V. Sprinkled data: NO

VI. Other data quality problems: NO

(4) Price

I. Missing values

- # tuples with missing value: 154/3076 (5.0065%)
- Solutions to fill in missing values
    - It's possible that we re-search the page of corresponding restaurants to see whether there is some other information on the page that we could deduct the price of this restaurant.

- Sometimes, there will be an electronical menu on the page that we could check the average price on its appetizer, main course, and desserts, then we could calculate the average price.
- We could also go through reviews by customers left on YELP to see those comments on whether they think this restaurant is cheap or the exact price they spent on having food in this restaurant to get the price of the restaurant

II. Classification: categorical

III. Outliers and anomalies: NO

IV. Synonyms: NO

IV. Sprinkled data: NO

VI. Other data quality problems: NO

(5) Zip-code

I. Missing values

- # tuples with missing value: 4/3076 (0.1300%)
- Solutions to fill in missing values
  - If the street address is valid, we could find its zip-code by searching its street address directly with some search engine, like Google map.
  - If the street address is not valid, we could use the default value like “00000” to imply that we could know the address directly

II. Classification: textual

- Average length = Maximal length = Minimal length = 5
- This attribute follows the same format: XXXXX
  - All tuples except those with missing values follow this format

III. Outliers and anomalies: NO

IV. Synonyms: NO

IV. Sprinkled data: NO

VI. Other data quality problems: NO

(6) State

I. Missing values

- # tuples with missing value: 0

II. Classification: categorical

III. Outliers and anomalies: NO

IV. Synonyms

- Sometime restaurant will use State code. However, some others will use the full name of the State or its abbreviation. For example, we found that some of our restaurants with State field filled with “California”, “Cal”, and “CA”. All these names corresponds to the same one → “California”

IV. Sprinkled data: NO

VI. Other data quality problems: NO

(7) City

I. Missing values

- # tuples with missing value: 0

II. Classification: categorical

III. Outliers and anomalies: NO

IV. Synonyms

- It has synonyms similar to State attribute. Cities like “Los Angeles” and “San Francisco” would have commonly used abbreviation “LA” and “SF”.

IV. Sprinkled data

- Sometimes the city names could also be found in the address field.

VI. Other data quality problems

- For New York, there are many districts. So it would get “Brooklyn”, “Long Island City”, which are not the city name, but the area.

(8) Street Address

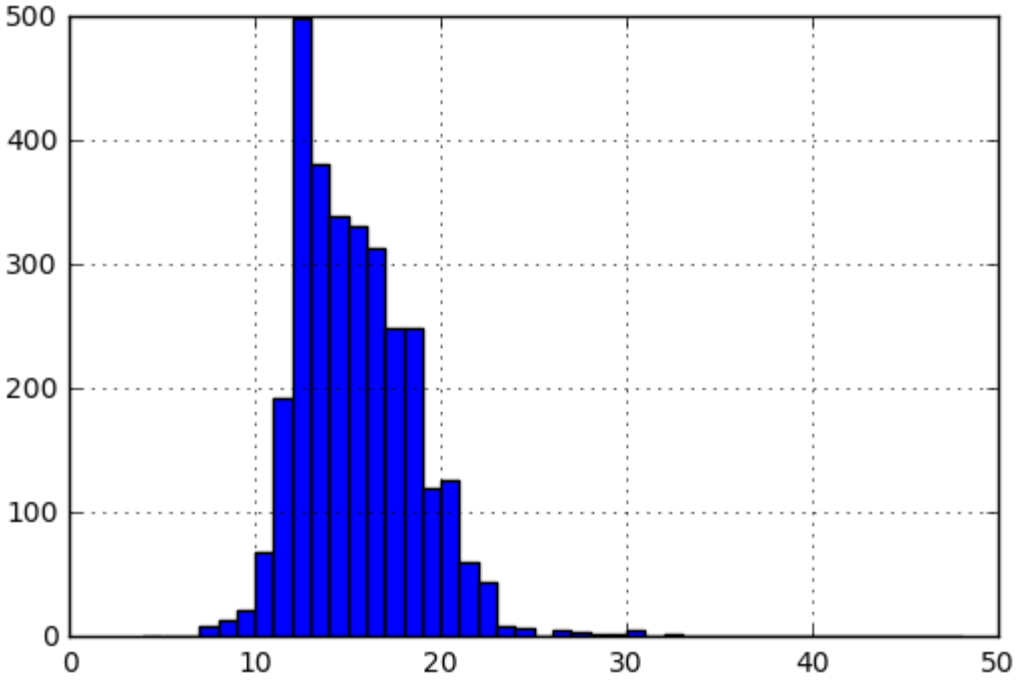
I. Missing values

- # tuples with missing value: 17/3076 (0.5527%)
- Solution to fil in the missing values
    - If the zip code is valid, we could search this zip code directly to get its street address.
    - If the zip code is also invalid, we would use “NA” as default to imply that we could not get the street address from the website.

II. Classification: textual

- Average length
- Maximal length
- Minimal length

III. Outliers and anomalies



- Most name has the length between 7 ~ 30. There are several restaurants with the name length a little bit greater than 30, which are still considered to be acceptable.
- There are three restaurants that we consider to be outliers:
    - “47-38 Vernon Blvd, Long Island City”
    - “University of Washington, King Lane &; Pierce Ln”
    - “Near Park Entrance At Haight and Stanyan St”.

IV. Synonyms: NO

IV. Sprinkled data: NO

VI. Other data quality problems

- There will be some addresses that are incomplete and you could not get the information of the addresses from its

(9) Has delivery

I. Missing values

- # tuples with missing value: 0 → 0%
- Solutions to fill in missing values
    - Actually, we assume that if the restaurant has the service of delivery, it would be a great highlight. However, if this is left blank, we use “NO” by default to assume that it has no such service.

II. Classification: Boolean

III. Outliers and anomalies: NO

IV. Synonyms: NO

IV. Sprinkled data: NO

VI. Other data quality problems: NO

(10) Has take-out

I. Missing values

- # tuples with missing value: 0 → 0%
- Solutions to fill in missing values
    - Actually, we assume that if the restaurant has the service of delivery, it would be a great highlight. However, if this is left blank, we use “NO” by default to assume that it has no such service.

II. Classification: Boolean

III. Outliers and anomalies: NO

IV. Synonyms: NO

IV. Sprinkled data: NO

VI. Other data quality problems: NO

PART IV. Software tools

- Jupyter
- Microsoft Excel
- Python
- Panda