Introduction

The dataset represents approximately 1 million businesses. To analyze the characteristics of these businesses, the data must be cleaned by removing duplicate entries as well as entries with insufficient information. In order to decide what constitutes insufficient information, it is first necessary to compute the properties of the fields.

Dataset Properties

The dataset consists of a million records for businesses with the following fields. Below is what the Github Readme file defines the fields as:

- **name**: The name of the business
- **address**: The street address of the business
- **city**: The city the business is in
- **zip**: The businesses zip code
- **state**: The state which the business is in.
- **time_in_business**: The years the company has been in business
- **phone**: The businesses phone number
- **category_code**: The NAICS code for the business
- **headcount**: The number of people employed by the business
- **revenue**: The revenue (in thousands) of the business

Fill Rate

To effectively clean the data, we first need to find the fill rate for each field. The fill rate is defined as the percent of records which have a value. The fill rate for each field is given below:

Table 1: Fill rate for each field

Field	Fill Rate
City	999950
Name	999964
Zip	999955
category_code	999962
Revenue	943066
Phone	590859
State	999961
Address	999957
Headcount	962334
time_in_business	916107

This dataset seems very Google Maps friendly. The business name, address (with the city, state, and zip code), and the NAICS code are filled in such that none of these fields are missing more than fifty entries. In contrast, the "phone" field is relatively empty. This is likely a protection against spammers and other bots which might be willing to web crawl Google's data. The "revenue", "headcount", and "time_in_business" fields have a greater than 90 percent fill rate. Even here, the usefulness of the information to scammers likely correlates with the fill rate. Measuring the correlation is outside the scope of this report.

True-Value Fill Rate

Even though many fields may have data, there's no guarantee that the data itself is relevant. In other words, leaving the field blank is not the only way to avoid inserting data. Analyzing the dataset, it is apparent that certain elements have "none", "null", an empty string, an empty space, and a simple "0" instead of inputted data. The zero is easy to spot in the following fields: "name", "address", "city", "zip", "category_code", "phone", and "headcount". None of those fields should realistically have simply a zero value; "headcount" should have a minimum of 1 (the business owner).

It is much trickier in the fields of "revenue", and "time_in_business". In those fields, a simple zero may indicate revenue of $0, or indicate that the business has just opened. In such a large dataset, a business bringing in $0 dollars in revenue would be an unusual outlier. In addition, businesses that are open for less than a year have indicated so in their submission to the field. More information is required to guarantee that all "0" values are irrelevant, but this report will act on the side of caution and assume that all "0" values absent any other symbols or text represent irrelevant data.

Table 2: True-value fill rate for each field

Field	True-value Fill Rate
City	999895
Name	999910
Zip	999890
category_code	999910
Revenue	943001
Phone	590798
State	999896
Address	999898
Headcount	962273
time_in_business	916048

The true-value fill rates for the fields reinforce the observations teased by the fill rate. The Google Maps friendly fields have fewer than a hundred elements with irrelevant or empty data, or a true-value fill rate of greater than 99%. The "phone" field remains the only field to have a true-value fill rate less than 90% percent.

Cardinality

Even if the elements are filled with relevant data, there's no guarantee that the data is entirely useful. Errors such as including a business twice are a fact of life when dealing with large datasets. Finding the cardinality of each field is a first-pass attempt at identifying these duplications. According to Technopedia, cardinality is defined as the percentage of unique values. "High cardinality means that the column contains a large percentage of totally unique values. Low cardinality means that the column contains a lot of "repeats" in its data range." The cardinality of each field is given in the table below:

Table 3: Cardinality for each field

Field	Number of Unique Elements	Cardinality
City	13714	0.013715440121212728
Name	890717	0.8907971717454571
Zip	26391	0.026393903329366232
category_code	1178	0.0011781060295426588
Revenue	11	1.1664886887712739e-05
Phone	575148	0.973510404571444
State	53	5.300551257330762e-05
Address	892114	0.8922050049105009
Headcount	9	9.352855166880917e-06
time_in_business	5	5.458229263095384e-06

The fields "name", "phone", "address" have a higher cardinality, as expected. The phone number is unique for each business, making it the best field to use to remove duplicates. The field "name" has a lower cardinality than the previous fields, also as expected. Popular names are routinely reused with no malicious intent. The repeats there would have to be investigated individually to confirm that they are indeed the same business. The fact that the phone field has a high cardinality despite having a low number of unique elements is due to the fact that the phone field contains a lot of empty or irrelevant data. However, "address" doesn't have a cardinality of 1. Since the address field indicates both the number and the street, this should be an individual value.

As expected, the "state", "zip", and "city" fields have low cardinality. There are only 50 states in the US. The data has captured businesses in all fifty states plus three territories. Even though the US has far more zip codes, the limited geographic footprint of the dataset ensures a certain repetition of elements in the "zip" field. The same holds true for the "city" field. The difference in cardinality between the "city" and "zip" field could be due to US metropolitan areas' natural political fragmentation. More data will be needed to analyze this.

The low cardinality of the "time_in_business" field is a surprise at first. Randomly investigating the data shows that this field contains categorical data. The same holds true for the fields "revenue" and "headcount". As an aside, the categorical nature of the elements in these fields further supports the assertion that an element with a value of "0" should be treated as empty.

Using the cardinality of each field is not a useful measure to determine data duplication. Most fields have low cardinality due to the fact that they are categories or due to the limited possible inputs. Measuring the cardinality of fields such as “state” is practically useless due to the limited number of existing states, and fields which contain categorical data have low cardinality. Fields with a high cardinality may also be misleading, since two businesses with the same name may not necessarily be duplicates.

Miscellaneous Observations

Table 4 shows that most of the elements which are filled in correctly. None of the fields have more than 65 irrelevant non-empty elements.

Table 4: Non-empty irrelevant data for each field

Field	Difference between Fill Rate and True-value Fill Rate
City	55
Name	54
Zip	65
category_code	65
Revenue	59
Phone	61
State	65
Address	59
Headcount	61
time_in_business	52

The data exclusively covers the fifty states of the US, DC, Puerto Rico (PR), and the US Virgin Islands. American Samoa (‘AS’), Guam (‘GU’), and the Marianna Islands (‘MP’) are excluded. The US Minor Outlying Islands are a nature reserve, so it would make sense for them to be excluded. Below is a dictionary which shows how many times each state appears. It would be interesting in the future to scale this by the state’s population, but that’s outside of the scope of the project.

For the fields which contain categorical elements, we want to see what the business distribution is among the elements. From looking at a simple count of each category, it is shown that a few categories predominate. The log of the count of each category was taken in order to better show the relationship between categories. Figure 1 below shows those relationships for the

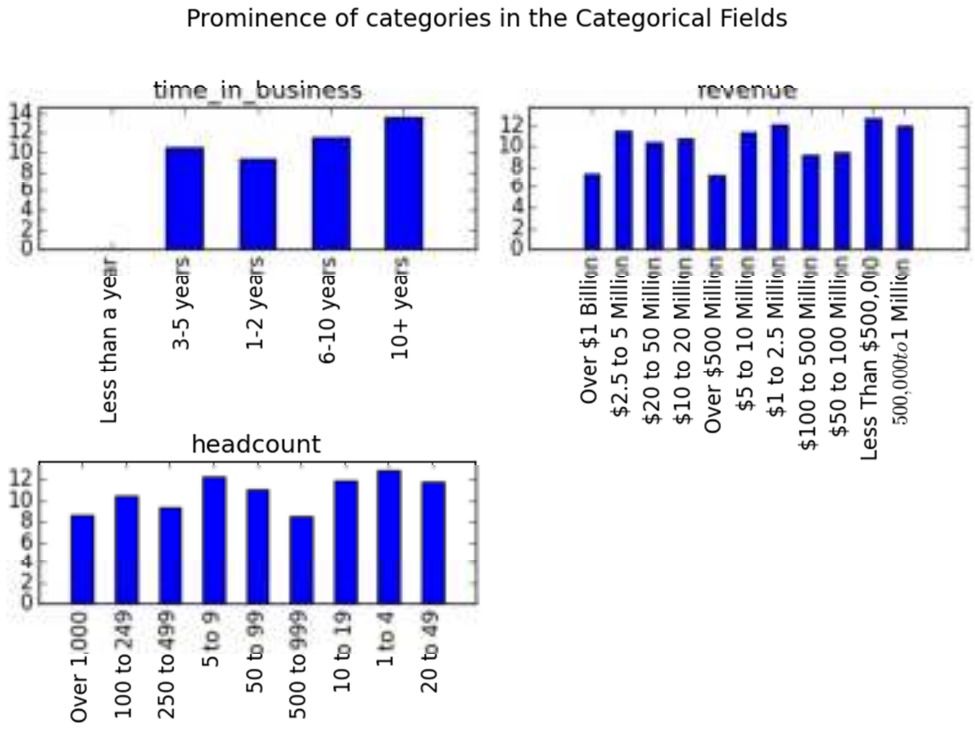

Figure 1: Prominence of categories in the Categorical Fields

Figure 1 above shows the distribution for the fields “revenue”, “headcount”, and “time_in_business”. From Figure 1, we see that the businesses in this dataset trend towards older businesses. The businesses in this dataset also have a range of revenue and headcount. However, the dataset does contain a bias against businesses with revenue over $1 billion and a headcount of over 250. Since most businesses in the US are classified as small businesses, these biases are to be expected. The dataset contains only one business which has existed for less than a year, creating an outlier in that particular field.

Since the ‘phone’ field is an outlier, it would be best to exclude it from analysis when searching for duplicates. There are 830820 entries which have relevant data except in the phone field. Looking through that subset of entries, there are 830820 unique entries. In other words, the data contains no duplicates.

Conclusion

The Google Maps friendly dataset is relatively clean, as seen by the high fill rate and true-value fill rate. Most of the data is entered correctly. All but two fields have low cardinality, but that is to be expected due to the nature of the fields themselves. That is true either because they contain category values or because the range of possible values is limited. To narrow down the search for duplicates, entries with empty fields other than ‘phone’ were ignored. With the data cleaned up, the categorical fields were investigated to show the characteristics of the usable businesses. Now the data is open for further analysis, perhaps by machine learning algorithms?

Reference

https://github.com/RadiusIntelligence/datascience-cc-1

https://www.techopedia.com/definition/18/cardinality-databases

https://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations