Radius Intelligence Exercise

Hanqiu Xia

May 2017

1 Data Overview

The dataset contains 1,000,000 observations, each observation represents a business. Each business has the following 10 fields:

- **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 (5 digits)
- **time_in_business**: The years the company has been in business
- **phone**: The businesses phone number
- **category_code**: The NAICS code for the business (8 digits)
- **headcount**: The number of people employed by the business
- **revenue**: The revenue (in thousands) of the business

The data type is string in each field. However, there are some missing data and uninformative data entries in each field, which are the focus in the following sections.

2 Fill Rate

In this section, I try to figure out that, for each field, how many records have a value. In other words, I need to filter out the records that with null values. For each field, number of records with values is the amount of observation minus the amount of null values. The fill rate is defined as the ratio of number of not null records to amount of observation. Figure 1 depicts the details.

	Fill rate	Number of not null vaules	Number of null values
address	0.999986	999986	14
category_code	0.999986	999986	14
city	0.999986	999986	14
headcount	0.962352	962352	37648
name	0.999986	999986	14
phone	0.590889	590889	409111
revenue	0.943092	943092	56908
state	0.999986	999986	14
time_in_business	0.916125	916125	83875
zip	0.999988	999988	12

Figure 1: Fill rate for each field

Most features have more than 95% fill rate.“zip” has the highest fill rate among all fields, while “phone” has the lowers fill rate, which is only 60%.

3 True-valued Fill Rate

Although most of records have values filled, some of them are filled with irrelevant values. For example, a field which has string valued entries may contain empty string like ‘ ’. This is a string but may not be ’good’ data depending on the field. Therefore, I need to investigate the all possible bad entries for each field carefully.

I start with “address”:

- First, I check if there exists the more than one data type, then I found besides null data, there exits both string and integer records
- Second, I check if there exists ‘empty’ string like ‘’, ‘ ’, etc. These string should have 0 length after splitting by space.
- Third, since address is less likely to have record with only one character, I filter out the entries with length as one after splitting
- In summary, besides null values, ‘address’ contains bad values includes: “”, “ ”, “0”, “none”, “null”, 0. (“Address” also contains values as “2ND” which is not informative.)

I use the similar logic to check other fields and found other fields have the same problem with “address”. Figure2 illustrates the true-values fill rate.

	Number of irrelevant vaules	Number of null values	Number of true values	True-Valued Fill Rate
address	89	14	999897	0.999897
category_code	76	14	999910	0.999910
city	91	14	999895	0.999895
headcount	79	37648	962273	0.962273
name	76	14	999910	0.999910
phone	91	409111	590798	0.590798
revenue	91	56908	943001	0.943001
state	90	14	999896	0.999896
time_in_business	77	83875	916048	0.916048
zip	98	12	999890	0.999890

Figure 2: True-values fill rate for each field

The variance of number of bad entries is small, the range is from 75 to 100. Comparing to the fill rate I computed from last section, the true-valued fill rate does not change a lot. Most features have more than 95% true-valued fill rate."Address" has the highest true-valued fill rate among all fields, while "phone" has the lowers true-valued fill rate, which is only 60%.

4 Cardinality

In data analysis, cardinality means the uniqueness of data values contained in a particular column of a database table.

I use data structure "set" to get the unique entries of each field, since "set" cannot take multiple occurrences of the same element. I compute the cardinality for all records first, then ignore the bad entries mentioned in section 3 and compute the cardinality for good records only. Figure 3 shows the details.

	cardinality for all records	cardinality for good records
address	892120	892114
category_code	1184	1178
city	13720	13714
headcount	15	9
name	890723	890717
phone	575154	575148
revenue	17	11
state	59	53
time_in_business	11	5
zip	26397	26391

Figure 3: Cardinality for each field

The variance for cardinality is large for all fields. "time in business" has the lowest cardinality

while "address" has the highest. The lower the cardinality, the more duplicated elements in a column. Thus, a column with the lower cardinality would have the same value for many rows. In SQL database, we could apply cardinality to help determine the optimal query plan for a given query.

5 Something Interesting

5.1 Inconsistent Format in Zip

The zip code in U.S. should have 5 digits, but there are some entries have 4 digits. I check them online and figure out that is because these zips all start with '0', they chop off the leading zeros. We need to be careful when we deal with these data in later analysis. Some software is friendly enough to accept just the last 4 digits, others may assume it as error so we have to add the leading zeros back.

5.2 Relationship Among Headcount, Revenue and Time in Business

I try to investigate relationship among headcount, revenue and time in business in quantitative way. Since the original data entries are all string, I replace them with the middle point of the range. For example, if the "headcount" is "1 to 4", then I use 2.5 instead. Based on this simple substitution, I draw the scatter plot of headcount and revenue with best fit line in Figure 4.

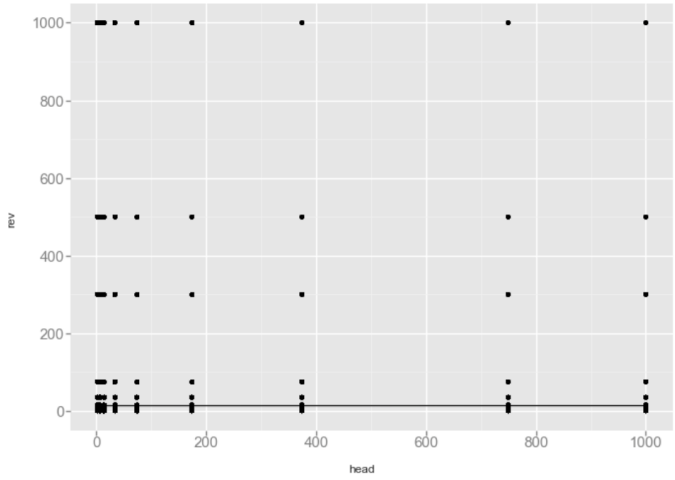

Figure 4: Scatter point of headcount and revenue with best fitted line

Unfortunately, there does not exist an obviously increasing or decreasing relationship between headcount and revenue.

In addition, I also explore the distribution of headcount and revenue in different years in business.

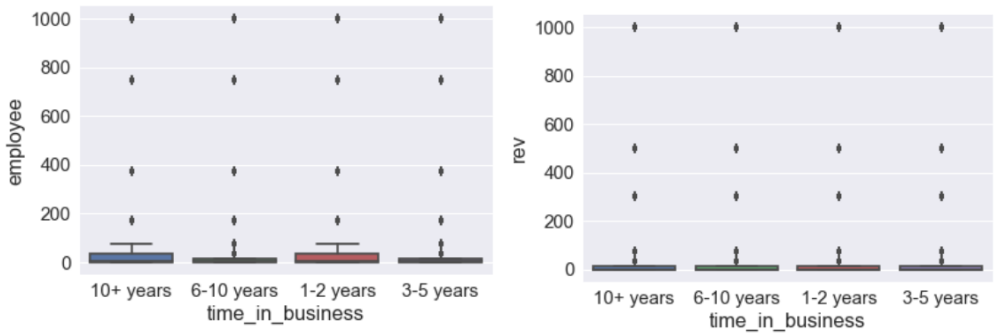

Figure 5: Left: Boxplot of headcount for different years in business; Right: Boxplot of revenue for different years in business

Neither headcount or revenue have dramatic difference for different time in business. Headcount has slightly larger interquartile range for businesses running 1 - 2 years and more than 10 years, revenue performs similarly in different company running time.

5.3 Top Popular States and Cities for Business

By counting the frequency of business occurred in each state, I plot the picture below to illustrate the top 10 popular states:

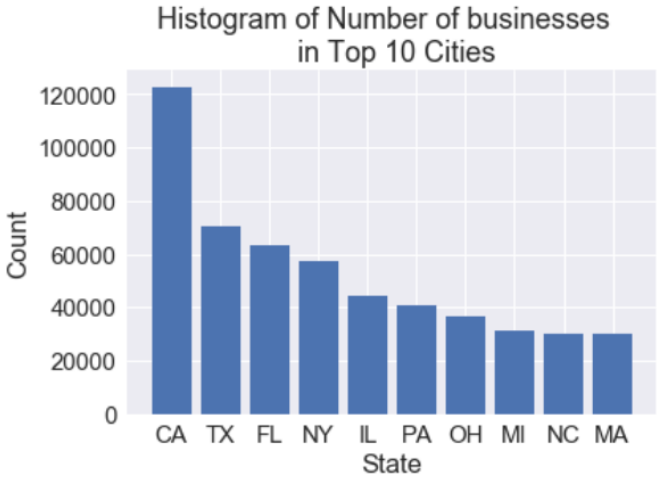

Figure 6: Scatter point of headcount and revenue with best fitted line

California, as the most populous state in the United States and the third most extensive by area, has the most companies in our dataset, which overwhelms the succeeding states, Texas and Florida. Within California, most businesses are located in the following cities:

	Number of Businesses
SAN DIEGO	5987
LOS ANGELES	5518
SAN FRANCISCO	4867
SAN JOSE	3215
SACRAMENTO	2466
IRVINE	2077
OAKLAND	1637
FRESNO	1488
LONG BEACH	1185
ANAHEIM	1182

Figure 7: Scatter point of headcount and revenue with best fitted line

Most of cities are located in the Greater Los Angeles Area and the San Francisco Bay Area, which are the nation's second- and fifth-most populous urban regions, respectively.