THE PROBLEM

In this challenge, I was given some business data that is representative of the type of data received by Radius Intelligence from its external data providers. My objective was to investigate this data, compute some basic properties (fill rate, true fill rate, and cardinality), and discover interesting findings.

In this report, I will explain my process of data cleaning and data analysis, and I will present my results.

EXPLORATORY DATA ANALYSIS

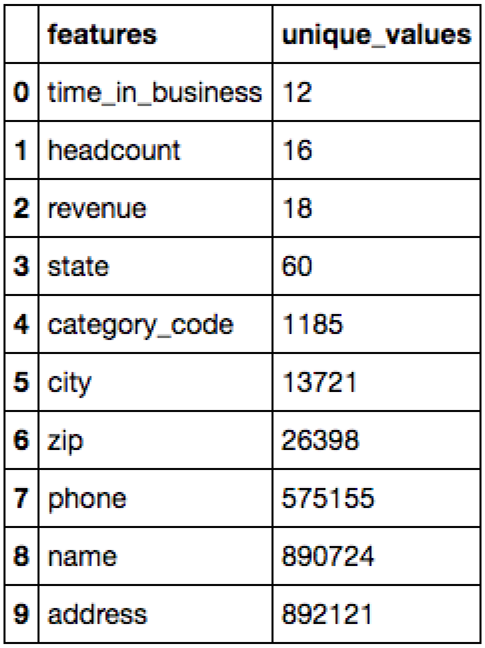
- **Data Overview**

Once the data was loaded, I started exploring its basic characteristics. Regarding size, the data set has 1,000,000 rows (each row representing a business) and 10 columns (each column representing a business feature).

Going more in depth on the features present on the data, the column names are as follows: ‘address’, ‘category_code’, ‘city, headcount, name’, ‘phone, revenue’, ‘state’, ‘time_in_business’, and ‘zip’. Most of these features refer to the business location or contact information (address, city, state, zip, phone, name), and at first glance all of the data seems to be string type and categorical.

- **Data Cleaning**

Exploring the data more, I computed the quantity of unique values of each field to learn more about each of them (see Table 1). The field ‘time in business’ presents the lowest amount of unique values, at the total number of 12, whereas ‘address’ holds the position of highest amount of unique values with a total of 892,121. For now, these results seem reasonable for it is expected that our data have duplicate information in each field, addresses being more unique than other fields.



	features	unique_values
0	time_in_business	12
1	headcount	16
2	revenue	18
3	state	60
4	category_code	1185
5	city	13721
6	zip	26398
7	phone	575155
8	name	890724
9	address	892121

Table 1: Features and Unique Values

Because ‘time in business’ is the field with the lowest unique values, I decided to analyse this field first. Looking at the values within this field, these are the results:

u'10+ years', u'6-10 years', None, u'1-2 years', u'3-5 years', u'null', u'0', u'Less than a year', u' ', u'', u'none', 0.

As expected, most of the data are string type representing categories. However, it is possible to detected some troublesome values that do not seem to belong in the data set or that do not give us useful information (None, ‘none’, ‘null’, ‘0’, 0, ‘’, ‘ ‘).

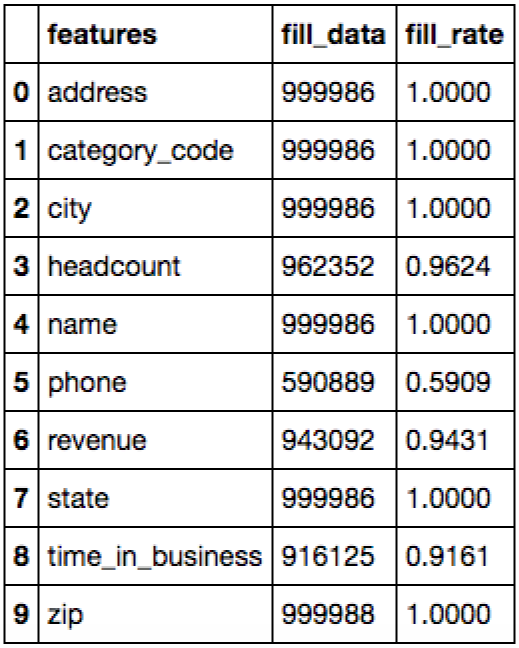
The most intriguing peculiar value is 0, for it could be miscategorization for the specific case of ‘time in business’, in which it could have been used in lieu of ‘Less than a year’. However, after further inspection of all other fields, zero is present in all of them. Including in fields where it is clearly wrong, for example ‘name’.

After cleaning the data set, I could move forward on my analysis and compute some basic properties. I will talk about each property separately in the following sections.

- **Fill Rate**

Fill Rate refers to the number of records that have a value for each field. More precisely, the total number of records from each field discounting the missing data (or null values).

Because this property it is not interested in the correctness or validity of the data, it should be calculated before cleaning the data.



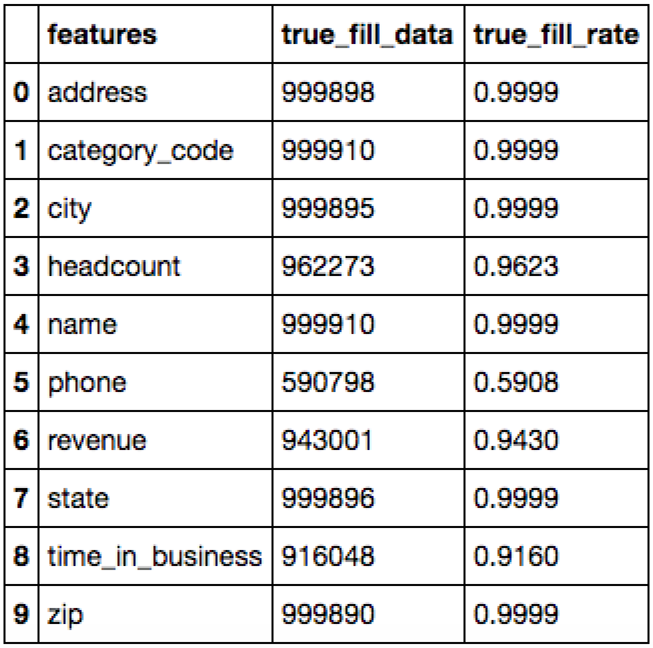
	features	fill_data	fill_rate
0	address	999986	1.0000
1	category_code	999986	1.0000
2	city	999986	1.0000
3	headcount	962352	0.9624
4	name	999986	1.0000
5	phone	590889	0.5909
6	revenue	943092	0.9431
7	state	999986	1.0000
8	time_in_business	916125	0.9161
9	zip	999988	1.0000

Table 2: Fill Rate

- **True-Valued Fill Rate**

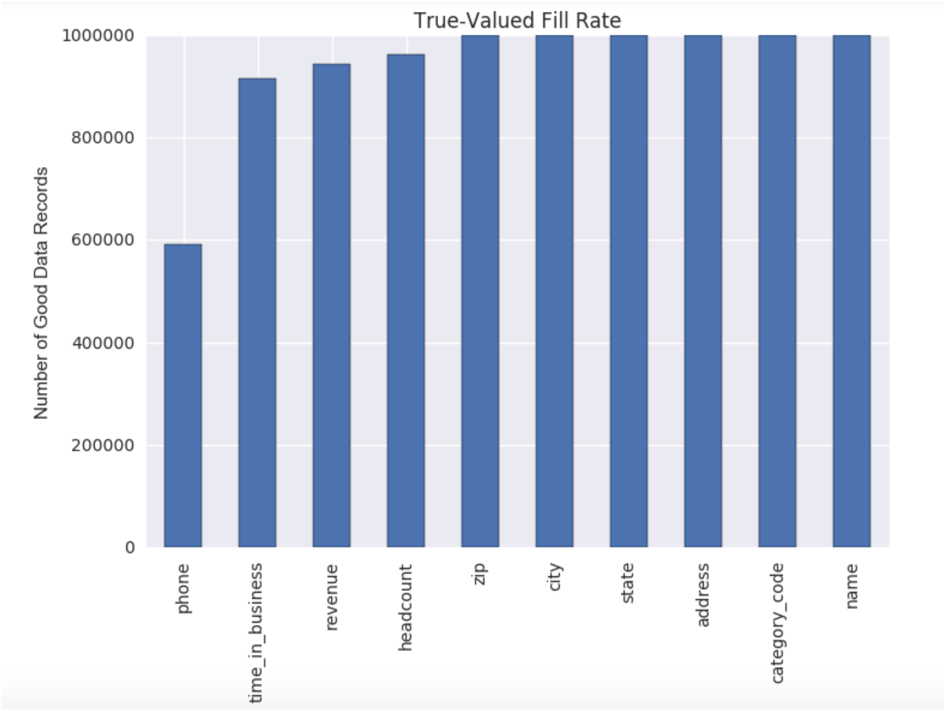
The True-Valued Fill Rate refers to number of records with relevant or good data. Not only is it necessary to exclude the missing values, but it is also needed to exclude data that was badly filled in, such as the examples mentioned during the data cleaning process (None, ‘none’, ‘null’, ‘0’, 0, ‘’, ‘ ‘).

On Table 3 it shows the values of true-valued fill rate for each field, and it is possible to see the decrease in the numbers. We are also able to analyse the true wellness of the data now.



	features	true_fill_data	true_fill_rate
0	address	999898	0.9999
1	category_code	999910	0.9999
2	city	999895	0.9999
3	headcount	962273	0.9623
4	name	999910	0.9999
5	phone	590798	0.5908
6	revenue	943001	0.9430
7	state	999896	0.9999
8	time_in_business	916048	0.9160
9	zip	999890	0.9999

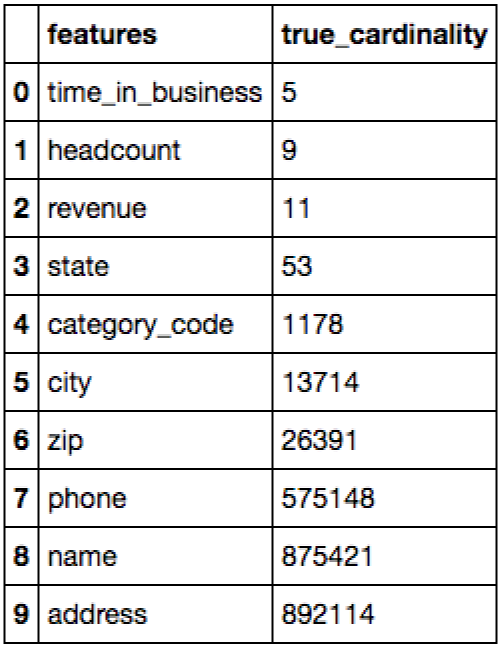
Table 3: True Fill Rate



By the True-Valued Fill Rate graph above, it is noticeable that the phone numbers is the most incomplete field present in the data by a big margin. More than 40% of its data is missing. Otherwise, all the data has a very high true-valued fill rate (> 91%).

- **Cardinality**

Cardinality refers to the number of unique elements in each field. I have calculated it in the beginning of my explorations, however, these numbers are expected to decrease after the data cleaning process.



	features	true_cardinality
0	time_in_business	5
1	headcount	9
2	revenue	11
3	state	53
4	category_code	1178
5	city	13714
6	zip	26391
7	phone	575148
8	name	875421
9	address	892114

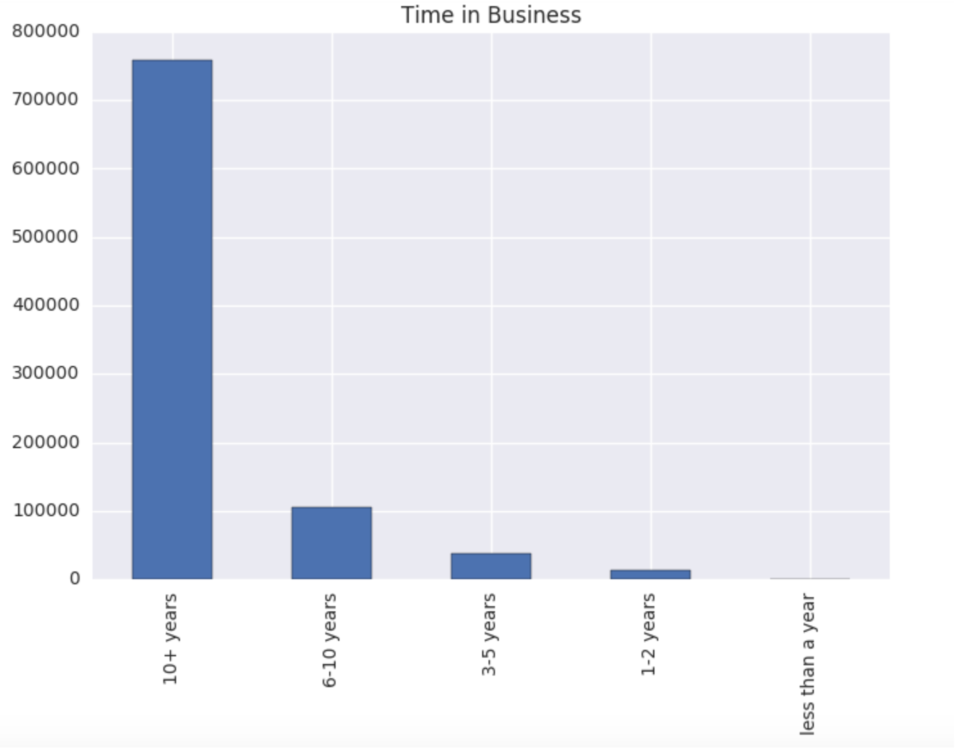
Table 4: Cardinality

- **Interesting Facts**

In this section of the report I am going to talk about some interesting or odd facts discovered in the data.

  - Time in Business

As mentioned before, ‘time in business’ is the field with the smallest cardinality, therefore, I decided to focus my attention in it.

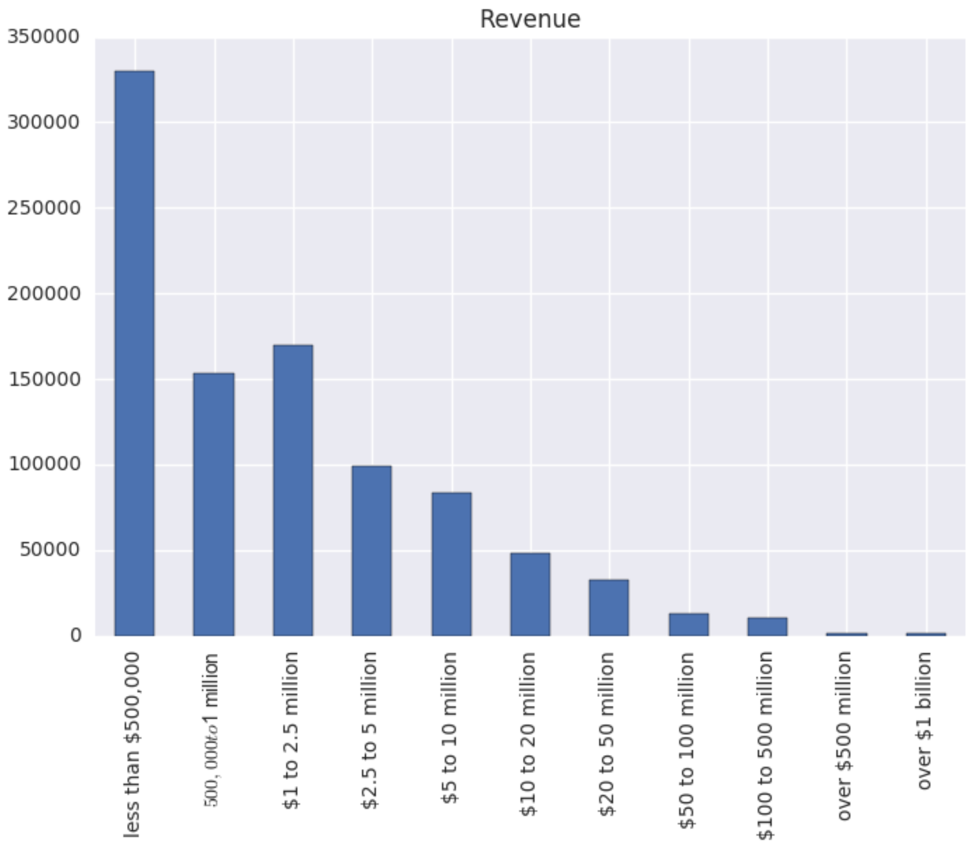


The graph above shows the distribution of the businesses in the data set in the five categories of ‘time in business’. The majority of companies (75%) in the data set has been operational for longer than 10 years. This is interesting, however it seems very reasonable that the senior companies have their data more available for the external data providers.

The most peculiar part about this graph is in the ‘less than a year’ category. There is only a single company. It is the “Mind Body Wellness Center”, located in Hurst, Texas and they are already in $2.5 to 5 million revenue bracket.

- Revenue

Revenue is a very important metric for any business. The graph below shows the revenue distribution among the business in the data set.



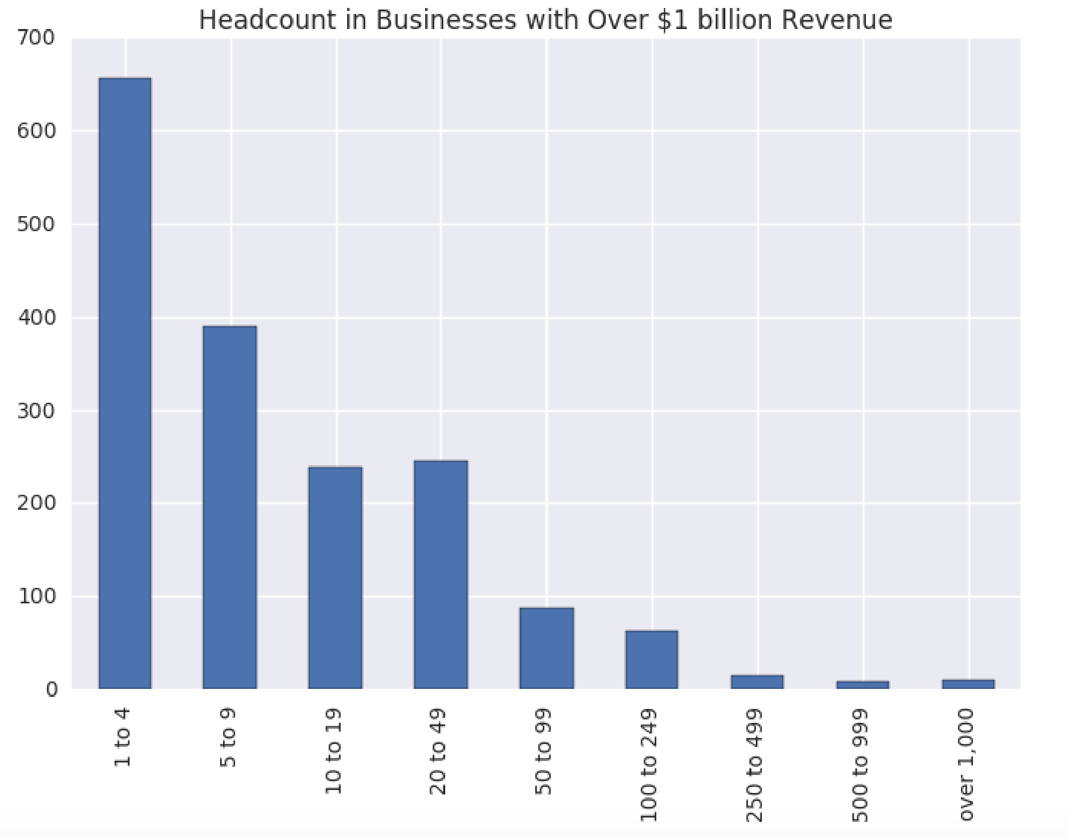
It is interesting to see that there is an almost linear pattern to the data. The majority of the companies belong to the first category ('less than $500,000'), and there is a trend of the number of companies decreasing the higher the revenue category.

This trend is broken in the beginning with the third category ('$2.5 to 5 million') having a larger number of businesses than the former ($500,000 to 1 million).

- Billion Dollar Companies

From the previous graph, we can verify that companies making over $1 billion are a minority, so I wanted to study their data and find any trends.

While analysing the distribution of headcount within the businesses making over $1 billion, I was expecting the majority of them falling under the last category ('over 1,000') because usually bigger companies require more workforce. However, the highest number of occurrences was on '1 to 4' category, the lowest headcount available.



It seemed really odd for a company with a over $1 billion revenue to have only 1 to 4 employees. Investigating this data, I encountered some odd entries. I list a few examples below.

- San Francisco Hypnotherapy Center (Lancaster, VA)
- Volunteers for Outdoor Colorado (North Oxford, MA)
- Sugar Creek Elementary School (Norcross, GA)
- Albion Elementary School (Phoenix, AZ)
- Sayen Elementary School (New Orleans, LA)

From these businesses names, it seems unlikely they actually have a revenue of over $1 billion. Moreover, it seems odd that San Francisco Hypnotherapy Center is not located in San Francisco, CA. The same strangeness regarding location is present on Volunteers for Outdoor Colorado, not located in Colorado.

We can assume that these entries probably have wrong data. Therefore, they should be removed from our data set.

FUTURE WORK

For future work, I would like to focus on a few points.

- Investigate more the companies with over $1 billion dollar revenue and find more patterns.
- Do more cleaning of ‘bad’ data, for instance, identifying data that was poorly entered or that has wrong information.
- Address has some duplicate data, so I would like to check if every data with the same address are good data points.
- I would like to have more precise time information about the company. For instance, date of creation.
- Explore more the NAICS category code and see if there is any relationship between it and other data (revenue, headcount, state).