Radius Data Analysis
By Jarrod Valentine

The data for this analysis is a collection of 1 million rows of business data representative of what Radius would receive from an external data provider. Each business has 10 fields associated with it, including: Address; NAICS Category Code; City; Headcount (the number of employees); Name; Phone; Revenue; State; Time In Business; and Zip Code. In my analysis, I sought to answer questions about the quality of the data, the amount of missing data for each field, and the degree to which the data in each field is correct. The following table gives a summary of each field. The Fill Rate is the percentage of records in each field that have a value out of the entire population for that field, while the True-Valued Fill Rate gives the percentage of records in each that have a meaningful and seemingly correct value.

	Fill Rate	True-Valued Fill Rate	% Of Unique TVF Values	Cardinality
Address	99.999%	99.63%	89.19%	High
Category Code	99.999%	91.37%	0.07%	Low
City	99.999%	99.99%	1.37%	Low
Headcount	96.235%	96.23%	0.0009%	Low
Name	99.999%	99.99%	89.08%	High
Phone	59.089%	59.08%	97.31%	High
Revenue	94.309%	94.30%	0.0012%	Low
State	99.999%	99.54%	0.01%	Low
Time In Business	91.613%	91.60%	0.0005%	Low
Zip	99.999%	95.34%	2.56%	Low
Total Entries	85.567%	84.28%	25.7406%	N/A

To find the **Fill Rate** for each field, I found how many records are missing an entry as indicated in a Pandas DataFrame by the **NaN** (Not a Number) or **None** data type. This rate does not include the strings **"null"** and **"None"** which are likely used to indicate missing data, but overlooked as they do not count as missing data.

Finding the **True-Valued Fill Rate** for each field required a slightly different approach depending on how I expected the data for that field to behave.

Zip

To find the True-Valued Fill Rate for the Zip field, I ensured that each entry was a string of length 5, as is standard for U.S. postal zip codes. Since zip codes are not static and I was unsure of the time period of the dataset, I did not check each entry against a master list of zip codes. Given a year, however, such verification would be possible.

Address

To check that each entry in the Address field was of the format [Street Address] + [Street Name], I made each entry a string, split it into a list, and checked whether the first component of the list was a digit and that the list had 2 or more components. This is a field that seems likely to need more human evaluation since addresses may be of the right form, but could still be incorrect.

Category Code

The NAICS Category Code is a 2-6 digit number designating the type of industry to which each business belongs. The codes range in specificity, with 2 digit codes being the most general and 6 digit codes the most specific. The data in this field, however, is 8 digits, and is likely a reflection of the use of codes from the NAICS-based codes used in the Economic Census and are 7- and 8-digit "kind of business codes." To check for validity, I obtained a copy of the NAICS codes and checked each entry against this list. I first truncated the 8 digit codes in the dataset to 6 digits and then padded all the codes in the NAICS list with trailing zeroes to make 6 digit codes.

City

I did some data exploration to find the minimum length of a City name as a string and no valid cities had fewer than 3 characters. There is a very real possibility that various City names are misspelled or non-existent, so further human evaluation or a check against a larger database of city names might be required to ensure the accuracy of the data in this field.

Headcount, Revenue, Time In Business

Since these fields have so few unique possible values, I made a list of the desired values and checked each entry against it. These fields were very simple to verify.

Phone

A standard U.S. phone number is a 10-digit number. To ensure that numbers were of this format, I converted the data type to string and removed all extraneous punctuation and spaces. This field has the highest amount of missing data, but of the data that is present, only 91 phone numbers are of an incorrect format.

State

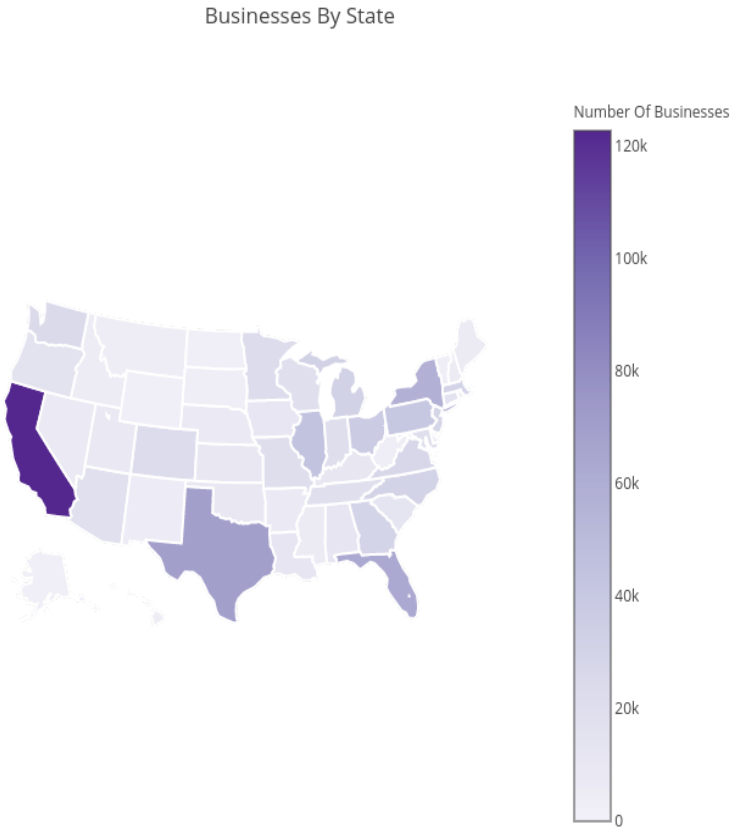
To find the number of records with relevant State data, I capitalized each state record and compared it against a list of capitalized state abbreviations. I included ‘PR’ for Puerto Rico and ‘VI’ for the U.S. Virgin Islands since they were present in the data.

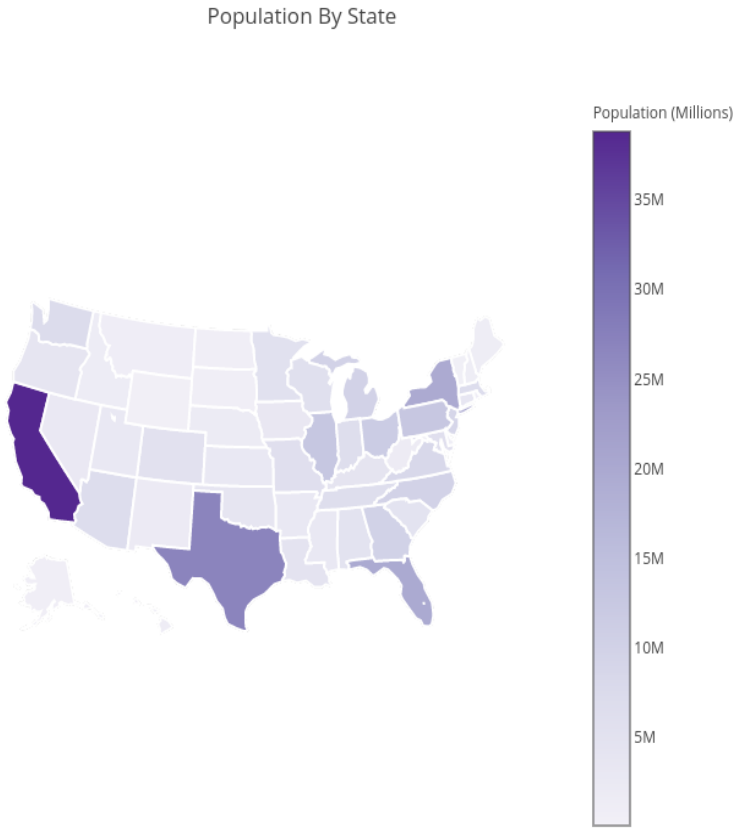
Name

The Name field is the most difficult to verify. There are many numerically named businesses (76, 2307), business names with 2, 3, and 4 letters that don’t parse as words but are nonetheless business names (BP, AHS, CBSI), and businesses with punctuation or special characters (AT&T). As such, further human evaluation is likely needed to further verify the accuracy of these business names.

To determine the **Cardinality** of each field, I found the number of unique true-filled values for each field as a percentage of the total number of true field values for that field. A low percentage of unique TFV entries indicates that the field has a low cardinality while a high percentage indicates it has a high cardinality.

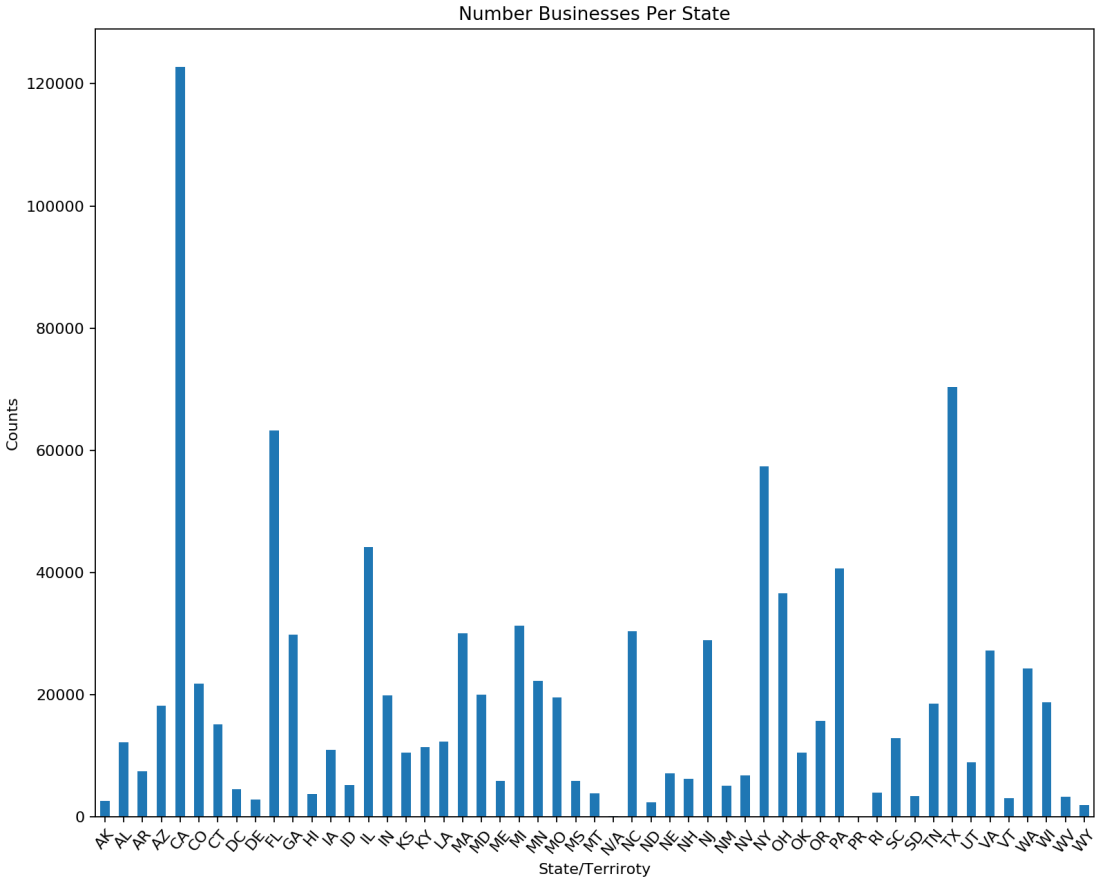
I checked the feature independence of each city, headcount, revenue, state, time in business, and zip code fields and found no significant relationship among them aside from the relationships between location based data.





After this initial data analysis, I decided to look at the breakdown of businesses according to state. I made a chloropleth map of the U.S. (excluding the very small number of businesses in both Puerto Rico and the Virgin Islands) and compared it to a map of U.S. population by state. As one would expect, the number of businesses in a state pretty accurately reflects the relative population of that state.

I then created a histogram to see the numeric breakdown of businesses by state and California, the most populous state in the U.S., has the highest number of businesses by a large margin. When we look at the number of businesses in each state as a percentage of the total number of businesses in the dataset, it is interesting that it pretty closely reflects the actual breakdown of the size of of each state's economy as a percentage of the U.S. economy [see link at end for reference].



Looking specifically into the California data, I wanted to see the breakdown of these businesses according to their revenue. The number of small businesses in California (those with revenue <$500,000) is far greater than those in other revenue categories. If we look further into this subset of business we data, we see that most of these small businesses cluster around major metropolitan areas like Los Angeles, The Bay Area, and San Diego. These small businesses operate primarily in their local economies and are more likely to do business with other small, local businesses in their metropolitan areas. Of the data we have concerning these businesses, we are missing about 4% of their entries, lower than the overall ~14% for the entire dataset.

This California small business data is highly valuable. Because Radius' product and business model benefits from the connectivity of businesses and providing them with the ability to segment and target others, both the company and the businesses it serves will benefit from minimizing the amount of missing data from this subset of businesses. A rate of 4% missing data seems small, but considering how large of a market share these small businesses represent, Radius should put a concerted effort into minimizing this rate after this data has been incorporated into The Network of Record. Some human evaluation of the data may be required since fields like addresses and names are more difficult to verify algorithmically, but the high percentage of missing telephone data that is missing can easily be found using web scraping techniques.

Link to State Economies as Percentage of U.S. Economy:
https://blogs.voanews.com/all-about-america/files/2015/09/US-Economy-by-State-Flag-be95-2-974x974.png