Insights into the Business data provided by Radius Intelligence

Hemalatha Vakade(Author)

Abstract – The goal of this report is to detail the information learnt from the data provided by Radius. This report contains a thorough exploratory analysis on the dataset. Exploratory analysis reveals the concentration of the business by some of the features in the dataset such as category, headcount, revenue, time in business etc. Analysis indicates that for the dataset provided California state has the maximum share of companies – 1.3%, with a headcount of "'1 to 4", Time in business of "10+ years" and a revenue of "Less Than $500,000"

Introduction – Some of the most important questions that businesses [1] ask are – "What do I need to succeed in my business?", "What kind of profits can I expect?", "Are some locations better than the others?". While there are many more, some of them are very specific to the type of the industry that the business works in, the above questions are few of the general questions that a business owner(s) would ask.

"I consider each business investment based on concept and revenue."

-Daymond John

To elaborate a little more on the features a business use to succeed or measure success, one of the essential areas is the revenue. [2] gives some very good pointers on why businesses should record their margins, which can be calculated using revenue.

While finding ways to increase revenue there may be other factors which influence it and in-turn the success of the business. Names of some of the businesses in this dataset includes "The UPS store", "Discount Drug Mart Inc.". It is intuitive that if these stores are close to residential areas they will draw more customers compared to similar stores farther away. [3] explains how location can be crucial for retail businesses.

Moving on to the longevity of the company and its influence on businesses according to a research [3] "a new firm is added and an existing one removed roughly every two weeks!". This study [4] suggests that "61-year tenure for average firm in 1958 narrowed to 25 years in 1980—to 18 years now". The same article mentions for successfully continuing the business the following are the key requirements 1) running operations effectively, 2) creating new businesses which meet customer needs, and 3) shedding business that once might have been core but now no longer meet company standards for growth and return. One of the companies that can be cited as an example here is Kodak [5] which was popular because it introduced digital photography but could not keep up with the trend of new emerging technologies.

Revenue per employee is one of the key performance indicators for a company because it gives the idea about how well a company is using its resources. According to an article [6], “smart headcount growth that optimizes technology is essential. Throwing headcount at a problem is not.” Helping workers operate productively is like asset utilization, an accounting term measuring how well a company uses assets to grow revenue. [7]

Sometimes changes in a country’s economy can directly impact certain industries and sectors and this has a direct bearing on the revenue. A company should always be aware of what is the status of the industry or the sector it is working in. An example would be the demonetization that recently affected the Indian economy. Certain sectors were impacted more than the others. According to this article [8] some companies such as smartphone based, or automobile companies were not impacted as much. Banks gained a lot while there seemed to have been negative impact on Power and Coal.

We can see that location, longevity of company, headcount, industry/sector of the company are having some direct or indirect impact on the revenue of the company. I further explore the dataset, courtesy Radius, to analyze these features.

Data

“In God, we trust; all others must bring data.”

– William Edward Deming

I absolutely agree with Dr. Deming. Data should act as the foundation for your decision-making process, not as a substitute for your own judgement [9]. The hard part after acquiring the data is to see if we can understand it, assimilate it and draw some insight out of it.

The dataset given to me consists of 1,000,000 data points. The data points here, refer to the rows in the excel provided. The rows correspond to the various businesses within USA. The following details, henceforth referred to as features or columns, of each business is made available.

- **Name** – Name of the company
- **Address** – The street address of the company
- **City** – City corresponding to the address of each company’s location
- **State** – The state in which the company is located.
- **Zip code** – Postal code corresponding to the location of the company
- **Phone** – The phone number of the business
- **Category Code** – The NACIS code for business.
- **Headcount**: The number of people employed by the business
- **Revenue**: The revenue (in thousands) of the business

Some of the statistics of the data is presented in the below table. This helps us understand the missing information in our dataset.

	address	category_code	city	headcount	name	phone	revenue	state	time_in_business	zip
count	999986	999986	999986	962352	999986	590889	943092	999986	916125	999988
unique	892120	1184	13720	15	890723	575154	17	59	11	26397
top	1 S DEARBORN ST	61111000	NEW YORK	1 to 4	Farmers Insurance	3037705531	Less Than $500,000	CA	10+ years	10001
freq	76	39461	14264	358207	821	88	329635	122812	758867	1151

Table-1: Shows the various statistics about the dataset.

Table-1 shows the various statistics of all the columns in the dataset. It can be useful to note that the label “count” indicates the total number of rows for which the data in a column is *not empty.* This can include meaningless values such as just a whitespace or the value ‘none’ etc. The label “unique” indicates the count of unique value within the column. Even this count includes values which are meaningless. The label “top” includes the most common value encountered in a column. The label “freq” corresponds to number of times the “top” value occurs in the dataset.

Its intuitive that the columns “address”, “name”, “phone”, “zip” and “city” have more number of unique values. The number of unique values in a dataset is also called Cardinality. The above-mentioned columns are therefore *high-cardinality* columns which means they have a lot of unique values. While columns such as “category_code” are *normal cardinality* and , “headcount”, “revenue”, “state” and “time_in_business” are *low-cardinality* columns. This is probably because these columns have limited values and each of the companies can be categorized using the values. The number of states is of course fixed. Although it is important to note that USA has 50 states only but some websites and articles include District of Columbia, and Puerto Rico as states. Although incorrect this information has been retained in my analysis as is [10].

The category code column gives the classification of the sector/industry that the business belongs to. Although the description that came with the data mentions it to be NACIS code, further research led to the conclusion the codes were not as per standard NACIS [11] codes (the length of the codes differed). On further investigation, I found that NACIS codes were adopted from SIC codes, but the a legitimate SIC code was 4-digit. The 8-digit code was a created unofficially by private companies. Even with the unoffical list I found that for example this code “31490000” corresponded to “FOOTWEAR, EXCEPT RUBBER, NEC” but in the dataset the name of the business was listed against the name “Real Hope Real Estate Inc.” which clearly was a mismatch

or incorrect . This means either the name is listed incorrectly or the category code. The category codes did not seem to match for other data points to considering the names which was intuitive to figure out the sector/industry. Therefore, I conclude that this column would need more research to figure out the veracity of the values. I do not consider the SIC code in further analysis.

Analysis –

Fill Rate

Fill Rate is described as “how many records have a value.” This is to be calculated for each column or feature of the Dataset. The Fill Rate for each column is nothing but the “count” label in Table-1. Fig-1. gives the visual representation of these counts for each column. As we can see the “phone” has some missing data, followed by time_in_business, revenue etc. Missing data can be problematic while developing insights because the data can be skewed and this affects the results or conclusions drawn from the data.

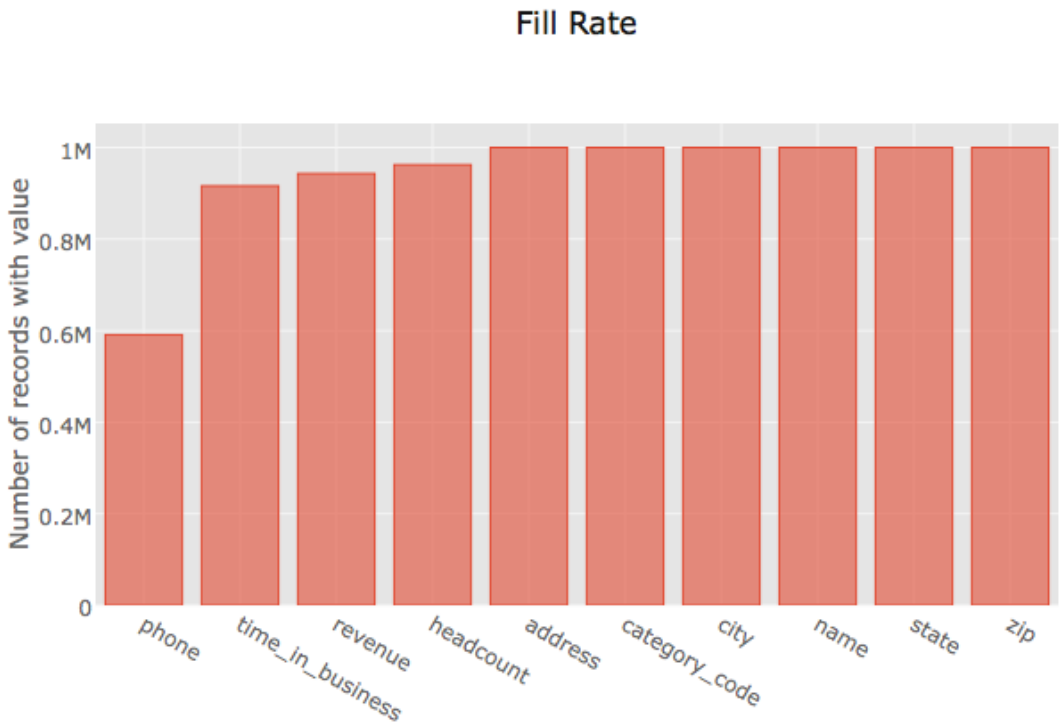


Fig- 1 Fill Rate for each column.

True-Valued Fill Rate

As mentioned earlier insights ca be drawn from relevant data. Data must go through a process of Extraction and Transformation so that it can be fed into models to get results that can answer the questions asked. If the data fed to the model is questionable then ultimately the predictions can be misleading and wrong. Therefore, process of data cleaning, extraction and transformation become the most crucial part of the analysis.

For this dataset the columns headcount, revenue and time_in_business has fewer unique values and therefore I chose them to find about the irrelevant columns. This revealed that the following values were not “good” values. – None, ‘0’ (number zero in quotes), ‘‘(whitespace), ‘’(empty string), 0 (number zero), ‘none’ (value ‘none’ in quotes). This was consistent with the headcount, revenue and time_in_business columns. When extended to the other columns in the dataset it showed the presence same set of values. These were filtered out the all the columns in the dataset and new statistics were found.

	address	category_code	city	headcount	name	phone	revenue	state	time_in_business	zip
count	999898	999910	999895	962273	999910	590798	943001	999896	916048	999890
unique	892114	1178	13714	9	890717	575148	11	53	5	26391
top	1 S DEARBORN ST	61111000	NEW YORK	1 to 4	Farmers Insurance	3037705531	Less Than $500,000	CA	10+ years	10001
freq	76	39461	14264	358207	821	88	329635	122812	758867	1151

Table-2: Shows the various statistics for the filtered dataset.

Table-2 shows the statistics after removing the irrelevant values. Fig-2 shows the True Valued Fill Rate for each column. Although the most common values or each of the columns haven’t changed there is small difference we can see against the count for each column. When plotted the variation is so small that it is not very conspicuous. (Fig-3) But the data set is free of irrelevant values.

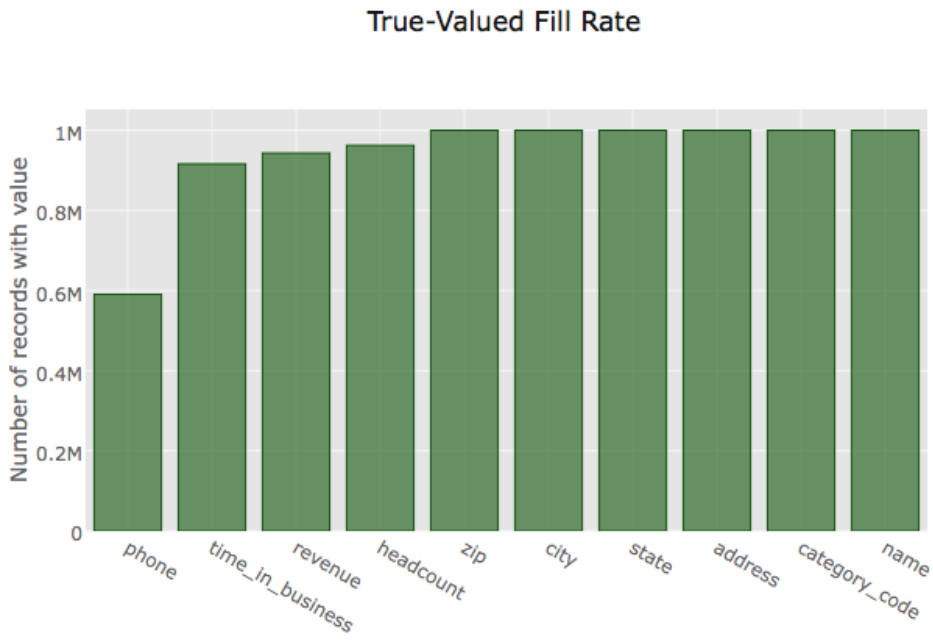


Fig- 2 True Valued Fill Rate for each column.

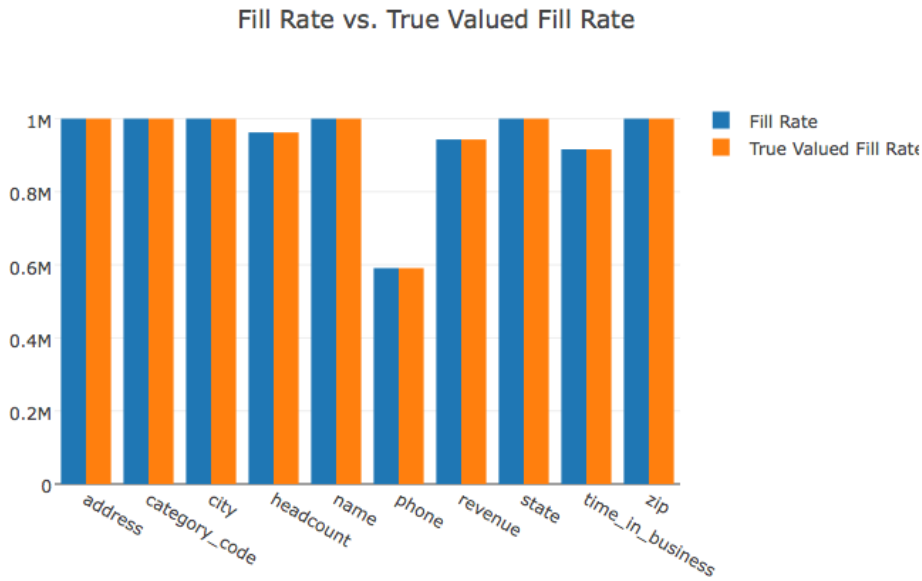


Fig- 3 True-Valued Fill Rate vs Fill Rate for each column.

Cardinality:

Cardinality as described earlier, is the count of unique values in each column.

It makes sense for some columns in the dataset to have high cardinality, such as name and address. These columns do not contribute too much to aggregations or various statistics to be derived. Although a normal cardinality column category code puts businesses into groups which helps in analysis. The cardinality refers to the “unique” labels under the Table-2. The visual representation can be seen in Fig-4.

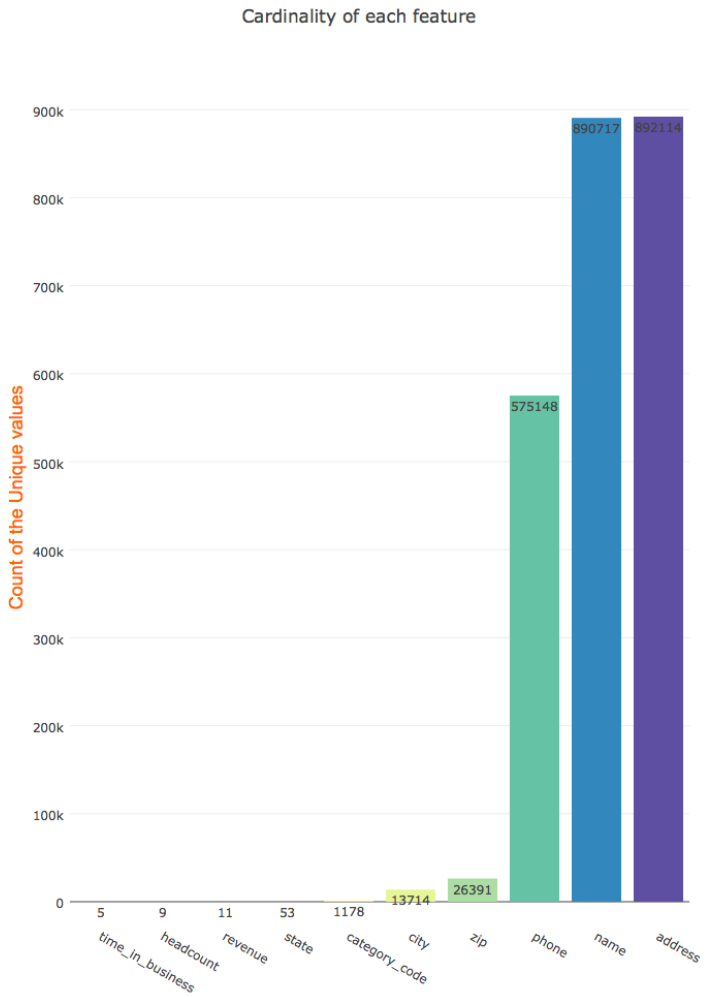


Fig- 4 True-Valued Fill Rate vs Fill Rate for each column.

Further Analysis –

<u>Most common value:</u>

From Table -2 it is also interesting to note that the most common value of state is “CA” while the city is “New York”. This while the count of the values in their respective columns is the same. The reason why this might occur because businesses in California state constitute 12.2 % while New York is low at 1.5% But “New York” is clear winner when it comes to the favorite city for business. It would be interesting to understand the

category of the businesses because as I mentioned in the introduction, industry/sector is important for a business. Among the businesses “Farmers Insurance” takes the top spot.

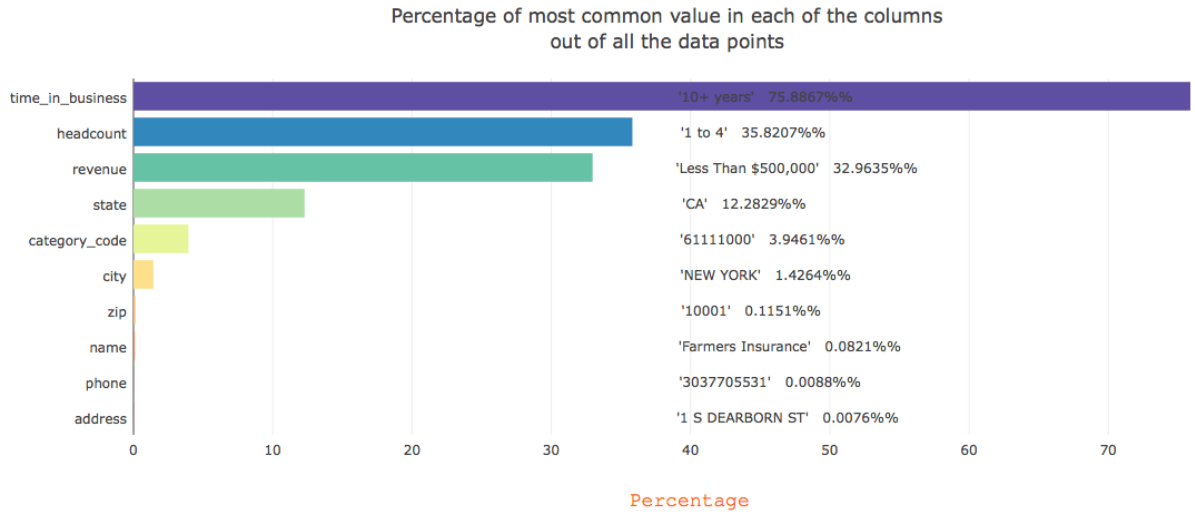


Fig-5 Most common value in each column and their percentages taken for the whole dataset.

Fig-5 gives an idea about the percentages of the most common value for each column in the dataset. The percentage is calculated by considering all the data points (1,000,000). We can see that most of the business, a whopping 76%, have a longevity of 10+ years. This does not necessarily mean that the study mentioned in [4] in incorrect but just means that for the dataset here the companies with 10+ years are more in number.

It is also interesting to note that the headcount is only “1-4”, which seems to be in line with what is mentioned in [6]. The revenue per employee in these companies considering may be therefore higher.

Even though I don’t have enough information about the category code I can say that it constitutes about 4% companies in that category.

“New York” city has about 1.4% of the total businesses and “CA” about 12.3% of the businesses.

The fig(4) and fig(5) makes sense together each validating the other, because the high cardinality columns have lower percentage of the most common value in the data.

<u>Distribution of values in fields:</u>

To see how the data is distributed between the values in various columns I plotted the Bar plots for headcount, revenue, Time in business columns.

From Fig – 6 we can see the distribution of data among various categories of Revenue. We can 35% of the revenue in less than $500,000. It is interesting to note that the data is skewed in a way that is does not have many companies listed which earn Over 1 billion or 500 million.

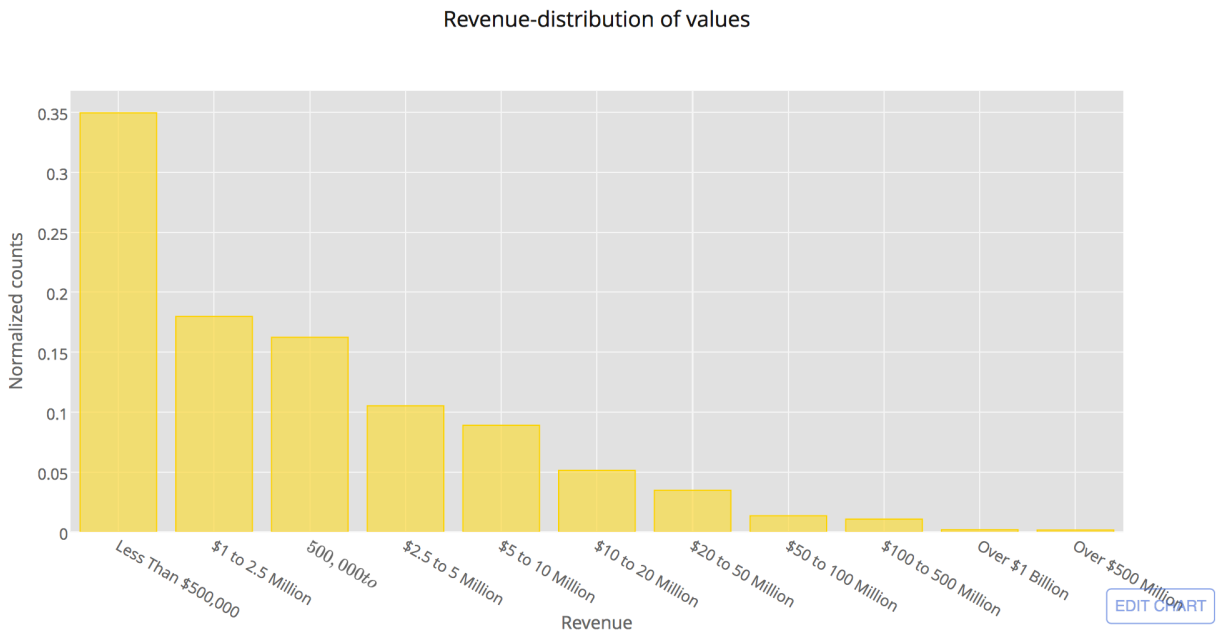


Fig-6 Distribution of data among various categories of Revenue.

Fig 7. Suggests the domination of companies having a small headcount. A little less than 40% of the companies have a small headcount. The dataset does not give the definite value for headcount and revenue but provides a range. It will be interesting to note the distribution of companies between these two features, revenue and headcount. As mentioned earlier a lower headcount indicates a better usage of the assets of the company.

Fig 8 shows a heatmap which summarizes the distribution of companies between Headcount and Revenue. The way to read a heatmap is to relate the colors on the graph to the colorbar which map the color to a count/score. For example, we can see for the headcount of “1 to 4” and revenue of “Less than $500,000” the color is bright yellow which maps to 12% of the companies . Therefore, according to our dataset most of the companies in this slot may be small businesses with a small headcount. It is also interesting to note that headcount of “1 to 4” also maps to 8%, have a revenue of $1 to 2.5 million, which leads to higher revenue per employee.

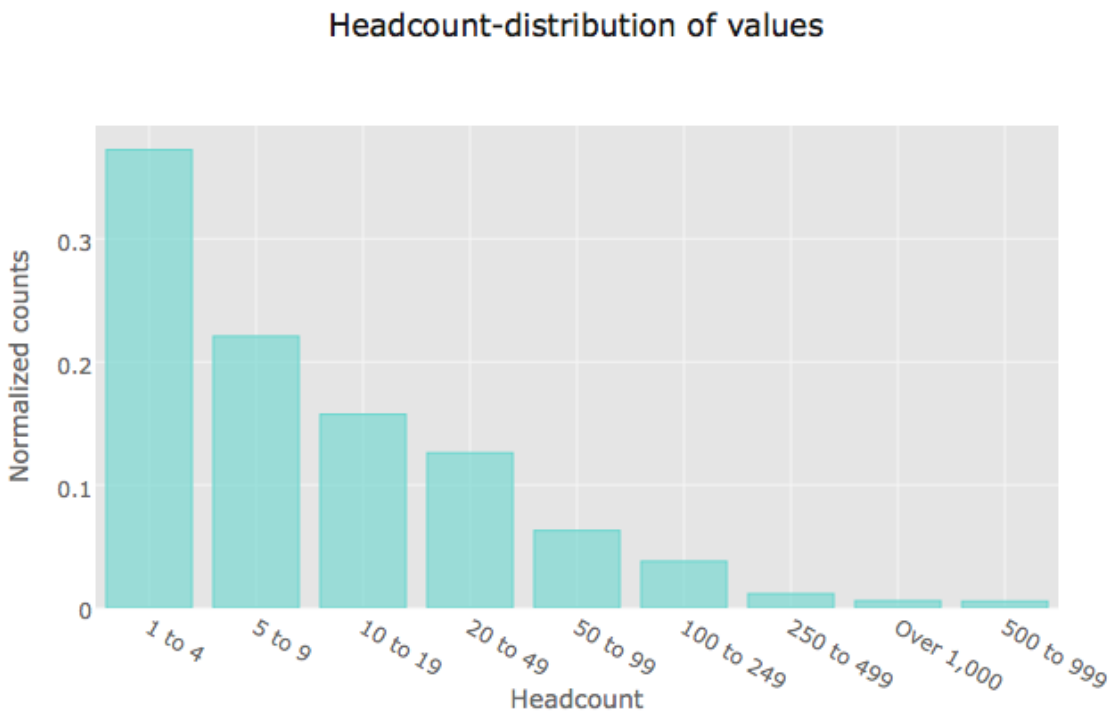


Fig-7 Distribution of data among various categories of Headcount

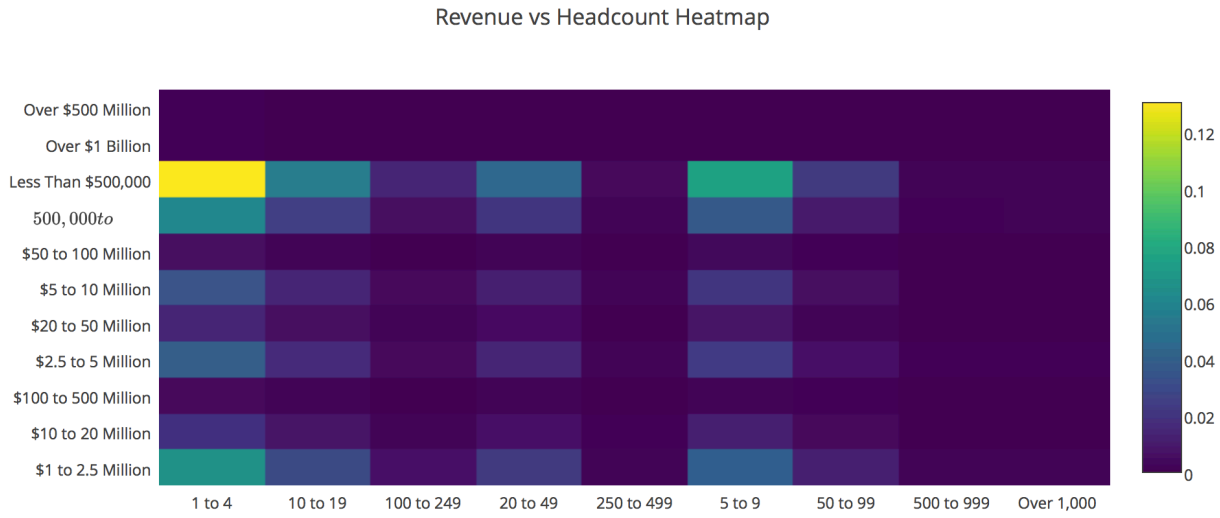


Fig-8 Distribution of data among various categories of Headcount and revenue

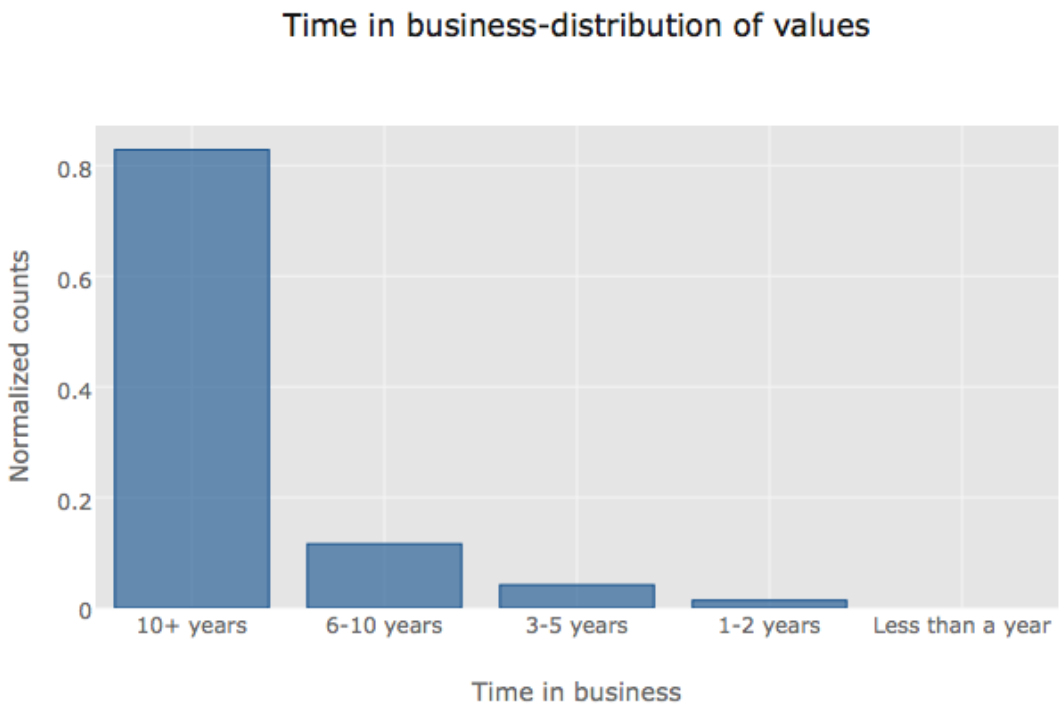


Fig-9 Distribution of data among various categories of years in business

The above Fig-9 shows the distribution of companies between various categories of Time in Business. It is interesting to note that there are barely any companies in the dataset with a longevity of less than a year. That would mean less number of “start-up” companies. 80% of the companies have longevity of more than 10 years.

On exploring the longevity of companies against Headcount the businesses with smaller headcount have been in business the longest.(Fig 10). Looking at the heatmap in Fig10 and Fig 8 we can say these business with the longevity of 10+ years are small businesses which have a revenue of less than $500,000

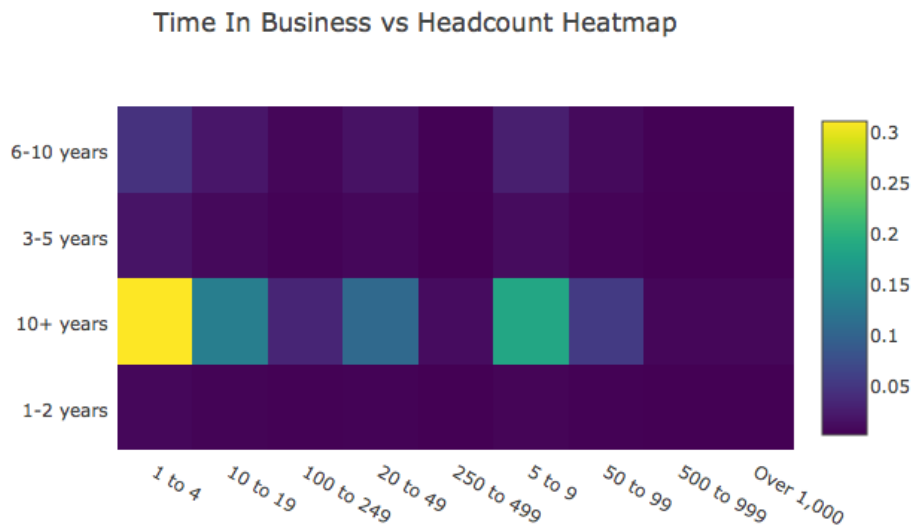


Fig-10 Distribution of data among various categories of years in business and headcount

I plotted similar heatmap for “state” vs “time_in_business” and “state” vs “headcount”.

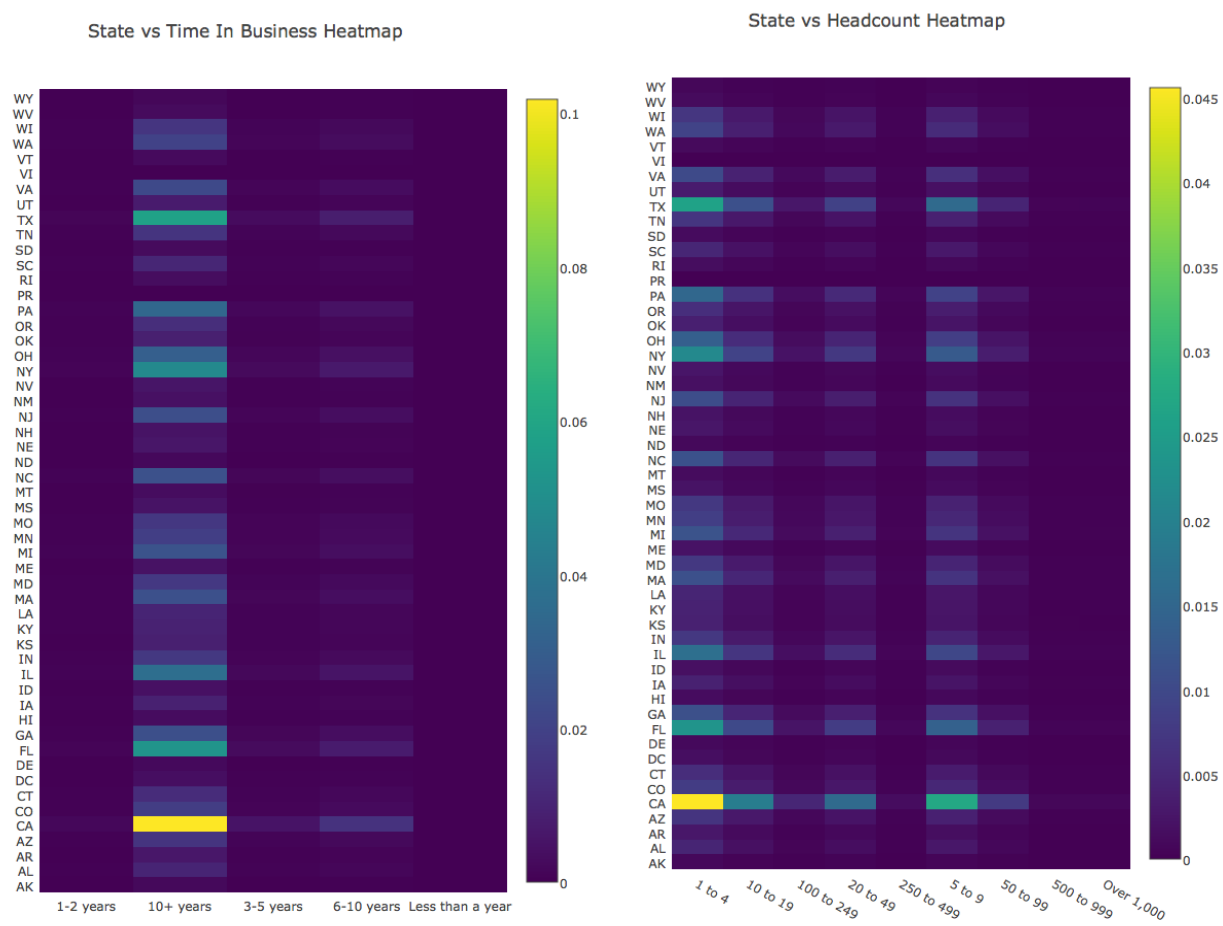
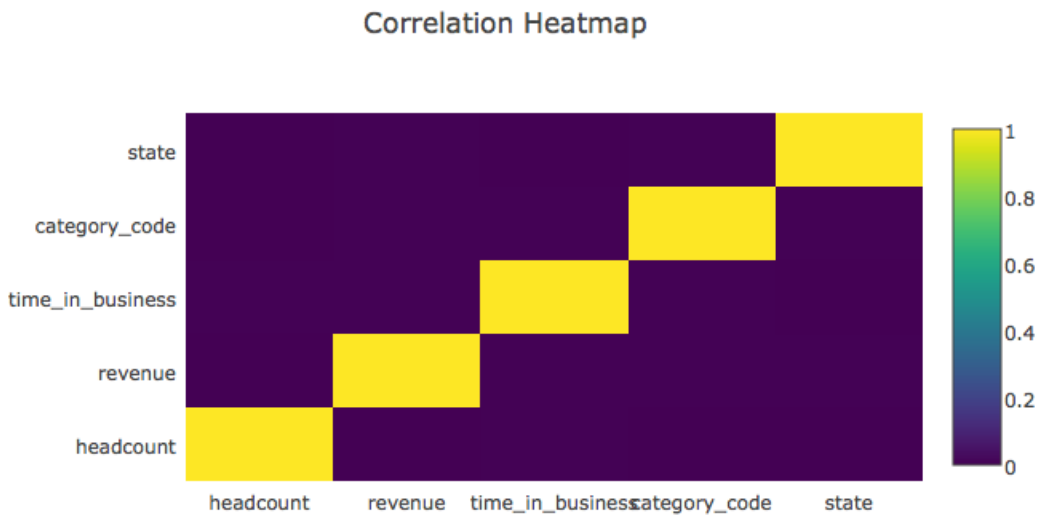


Fig-11(a) Left: Distribution of data among various categories of years in business and state
11(b) Right: Distribution of data among various categories of headcount and state

The figures 11(a) and 11(b) show that California has a major share of the companies with headcount “1 to 4” (4.5%) and time in business for “10+ years” (10%).

Correlation

To get a summary of how all the field correlate with each other I plot a heatmap of the correlation. I am only interested in the correlation between “state”, “time_in_business”, “revenue”, “headcount” and “category_code”.

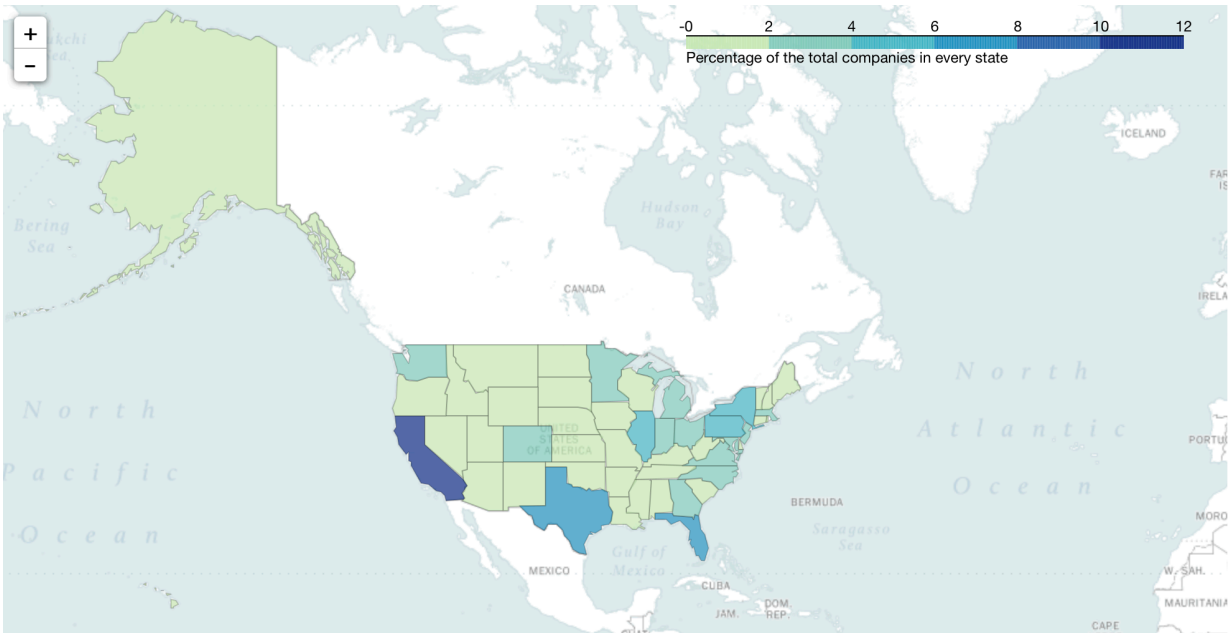


The correlation heatmap suggests that for this data there seems to be no correlation between two different columns. Correlation is the measure of the strength of a relationship between two different variables. This value can be positive or negative. Two variables are positively correlated if an increase of one causes an increase of another and they are negatively correlated if the increase in one decreases the other. It is clear from the map that no such relationship exists between any two different variables.

Location based aggregations

To see how the location influences the business I did analyses state wise grouping them at various levels. But before that there were few missing values in the state column. This could be found out programmatically using a zip code lookup through Google Map APIs. For the missing values in the zip code column I did a lookup using I used SmartyStreets API with the help of the street address combined with the city and state. At the end, I had a dataset with no missing data in “state” column.

To see the distribution of companies only by “state” I plotted a choropleth.



Evidently, California leads with 12% of the total companies, followed by Texas, Florida, New York etc.

I also explored heatmaps for “state”, “revenue”, “headcount” and “time_in_business”. (Please refer the attached Jupyter notebook which consists of the interactive maps)

I found that California with a revenue of “Less than $500,000”, headcount of “1 to 4”, a time_in_business of “10+ years”, had a maximum share of 1.3% of the companies.

I tried to figure out the industry for the companies for the above values, since “category_code” was not clear. Some of the company names were “Salem Family Medicine”, “Cannons Methodist Church” , “California University Of Pennsylvania”, “US Mattress”, “Talley Insurance Services”. It makes sense that these companies have a headcount of “1 to 4”. The other values of the data make sense too. Business in sectors such as education, ,healthcare, real estate services don’t make much in terms of revenue and usually have a small headcount in offices while their time in business is also long.

Conclusion – After the analysis on dataset, I can conclude that the findings show the maximum share of companies for the combination of “CA” state , in business for 10+ years with a headcount of 1 to 4 and a low revenue of “Less Than $500,000”.

The data is not balanced among the various classes of a column. There is a huge skew in the dataset. For example, more than 75% are running businesses for 10+ years. If I had more time, I would like to collect more data to explore algorithms to predict the revenue, given state, headcount, time in business and category code. I would also like to gather data about the

sectors (category code) in which these, companies are working, because strategy to run businesses differs according to industries.

References -

[1] http://www.ala.org/ascla/asclapubs/surviving/thirtymostasked/thirtymostasked
[2] https://www.jeffalytics.com/understanding-profit-margins/
[3] https://hbr.org/2013/12/research-most-large-companies-cant-maintain-their-revenue-streams
[4] https://www.innosight.com/insight/creative-destruction-whips-through-corporate-america-an-innosight-executive-briefing-on-corporate-strategy/
[5] http://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/reflections-on-corporate-longevity
[6] https://hbr.org/2016/02/why-digital-companies-grow-without-adding-headcount
[7] http://www.investopedia.com/terms/r/revenueperemployee.asp
[8] http://economictimes.indiatimes.com/slideshows/investments-markets/how-demonetisation-will-impact-top-11-sectors-of-economy/capital-goods/slideshow/55432930.cms
[9] https://medium.com/@adambreckler/in-god-we-trust-all-others-bring-data-96784d01e9be
[10] https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States
[11] https://classcodes.com/naics-code-list/
[12] http://www.stssamples.com/sic-code.asp