RADIUS INTELLIGENCE
DATA CHALLENGE

Zainab Danish

Table of Contents

Executive Summary

This article uses an external dataset on 1 million businesses from various industries in the US and is divided into two sections: Section 1 talks about issues pertaining to unclean data and measures taken to overcome such issues, Section 2 focuses on the healthcare industry and presents an in-depth analysis with graphics. The investigation concludes that California is a flourishing state in terms of healthcare, with many new businesses as well as private practices. California also ranks at the top in terms of revenue generated from healthcare.

Section 1: Data Quality

1.1 Fill Rate, True Valued Fill Rate and Cardinality

The difference between fill rate and true-valued fill rate presents a big challenge since the latter goes further than just the removal of empty values; it adds additional filters to ensure the data contained is valid and represents a value that actually belongs to that particular field. One of the most fundamental problems at the beginning of every data science exercise is the quality of data. Data cleaning really begins after the obvious missing values have been removed.

The dataset at hand contains various fields defining key facts about business belonging to various different industries as defined by their respective NAICS codes. The fields and possible checks for data validity have been listed below for posterity:

Field	Type Check	Length Check
name	string	-
address	string	-
city	string	-
zip	string	5 (US Postal Code length)
time_in_business	string	-
phone	string	10 (country code excluded)
category_code	string	8 (NAICS standard)
headcount	string	-
revenue	string	-
state	string	2 (state abbreviations)

Table 1.1

Other than the stipulations above, it is important that none of the values should contain empty strings, the words ‘none’ or ‘null’ in string form, empty values or zero in either string or integer form. Pertinent to note here is that a headcount, time_in_business or revenue of 0 does not make sense because the lowest of ranges (e.g. ‘< $500,000) should incorporate 0 if such a business exists. Hence, the ‘0’ values are also excluded when calculating the true-valued fill rate.

Lastly, cardinality of a field refers to the number of unique values that field contains. The fill-rate, true-valued fill rate and cardinality for each field was computed and the results are presented in the dataframe that follows:

	Columns	Fill Rate	True Value Fill Rate	Cardinality
0	address	999986	999898	892114
1	category_code	999986	999910	1178
2	city	999986	999895	13714
3	headcount	962352	962273	9
4	name	999986	999910	890717
5	phone	590889	581380	565731
6	revenue	943092	943001	11
7	state	999986	999896	53
8	time_in_business	916125	916048	5
9	zip	999988	953374	24410

Fig 1.1

It is interesting to note that out of the 581,380 phone numbers there are only 565,731 unique values, entailing that 15,649 businesses share phone numbers. However, over 107,000 businesses share their address with another business. The disagreement between number of shared addresses and shared phone numbers can be attributed to the missing values in the phone number field amongst other things.

Overall, the true-valued fill rate presents a promising picture being in the high 900,000's for all fields except phone number.

1.2 Inconsistencies

The NAICS provided for each business includes various levels of industries encoded in a single string of digits. A quick run of value_counts for this field indicated that the highest number of business in the same unique category came from code '61111000' which, as the NAICS dictionary indicates, represents high schools and elementary schools. The following figure shows the head of the filtered dataframe which includes data on high schools and elementary schools. This dataframe contains over 35,000 rows.

address	category_code	city	headcount	name	phone	revenue	state	time_in_business	zip
1005 W EHRINGHAUS ST	61111000	ELIZABETH CITY	10 to 19	ACUSTREAM	None	$2.5 to 5 Million	NC	None	27909
220 INFO HWY	61111000	SLINGER	50 to 99	SHAW-LUNDQUIST ASSOCIATES INC	None	$20 to 50 Million	WI	10+ years	53086
4201 BASELINE RD	61111000	LITTLE ROCK	1 to 4	City of Watauga	2126555411	$1 to 2.5 Million	AR	10+ years	72209
105 BRIGHTON AVE	61111000	ALLSTON	1 to 4	The Inn At Dupont Circle North	9547646099	$1 to 2.5 Million	MA	10+ years	02134

Fig 1.2.1

A quick look at the name field indicates a serious inconsistency. Coincidentally, none of the entries in the head of the dataframe appear to be schools, e.g. Shaw Lundquist Associates Inc., Acustream, etc. A quick google search further confirms this. A very crude way of filtering out actual schools would be to select

entries wherein the name contained the words ‘school’ or ‘elementary’ or ‘high’ or ‘academy’ or ‘charter’. When these words were used for filtering, the number of entries dwindled to a mere 1556. The head of the filtered dataframe only containing schools is presented here:

address	category_code	city	headcount	name	phone	revenue	state	time_in_business	zip
100 N BOKELMAN ST APT 422	61111000	ROSELLE	10 to 19	Swinburn Elementary School	5032358655	$1 to 2.5 Million	IL	10+ years	60172
667 WHITEHALL RD	61111000	LONG ISLAND	1 to 4	Ivy League Academy	None	Less Than $500,000	VA	10+ years	24569
3890 HIGHWAY 81	61111000	LOGANVILLE	20 to 49	Pontiac Middle School	None	Less Than $500,000	GA	10+ years	30052
3498 MAHAN DR	61111000	TALLAHASSEE	1 to 4	Murtaugh Grade School	None	$1 to 2.5 Million	FL	None	32308

Fig 1.2.2

The case presented above represents only one of the many inconsistent labels present in the dataset. While it is possible in some scenarios, such as the one presented above, to explicitly filter out relevant entries using indicator words, it is much harder to do so in others. Take the clothing retail industry, for instance. If the business ‘Williams Sonoma’ is included under clothing retail, there is no indicator pointing to the fact that this is not a clothing store, other than common knowledge perhaps. Hence, for the next part of this analysis, the NAICS was not used as the primary field for filtering.

Section 2: Data Analysis

2.1 Introduction

This section analyzes the healthcare sector which can also be filtered out using explicit keywords. Three different types of healthcare services are explored - hospitals, clinics and private practices. Each healthcare service is unique in terms of size and functionality making for an interesting analysis. The keywords used for filtering each are listed below.

1. **Hospitals:** hospital
2. **Clinics:** clinic
3. **Private Practice:** MBBS, MD, DDS, BDS, OD, DPM, DMD, therapist

The list of words for private practices was taken from a wikipedia page defining medical credentials. It is important to note here that the approach taken can be improved on immensely by accounting for various other keywords that might define the profession as well as different variations of each keyword provided. For instance, MD can be written as M D or M.D..

2.2 NAICS Discrepancies

For this analysis, the NAICS code was not taken into account because there were a lot more entries which were mislabeled when they should have been under the healthcare category. There were close to 20,000 entries containing the keywords mentioned above after filtering out entries with missing values in key variables - headcount, state, time_in_business.

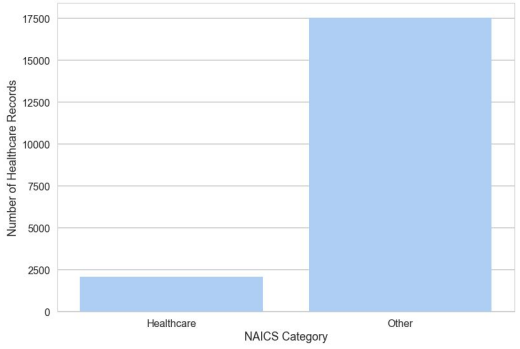

Fig 2.2.1

Fig 2.2.1 above shows the healthcare operations (filtered using keywords) broken down by their NAICS label used. The bar labelled 'Other' represents all the healthcare operations we would have missed had we only been using the NAICS code for filtering out healthcare services. Out of 19,592 filtered healthcare operations, only 2051 were actually under the 2-digit NAICS code representing the healthcare industry.

The example above represents the perils of relying on a single variable to provide key information about entries. 90% of the data about healthcare services could have been left unused.

2.3 Distribution Analysis

This section investigates the distribution of data in terms of years in service, revenue and states. The first two graphs take a quick look at the revenue distribution and the distribution by years of service for the healthcare industry.

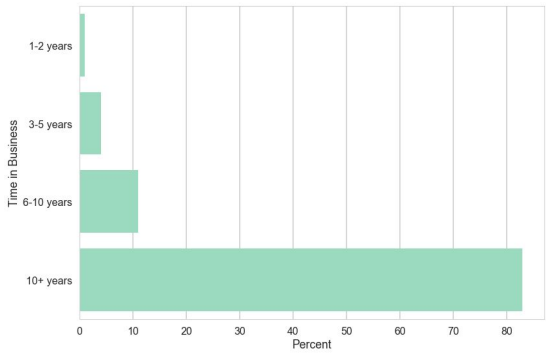

Fig 2.3.1

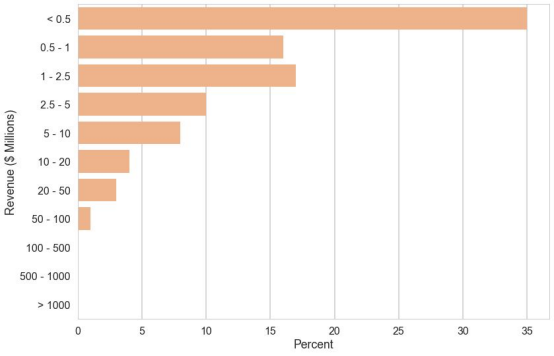

Fig 2.3.2

Fig 2.3.1 on the left shows that the vast majority of healthcare operations in this dataset are older than ten years. Interestingly, the revenue graph on the right, Fig 2.3.2, shows that about 35% of the businesses make less than $500,000 a year. This was rather curious, so I decided to see what the statewide distribution looked like in terms of revenue and where that revenue was really coming from.

For the next bit of investigation, I created two new variables. The first one was average revenue. This was created by taking the midpoint of the upper and lower limit of every range. For the upper extreme range, namely 'Greater than 1 billion', $ 1 billion was used. The use of average values might not offer a good approximation of the actual data especially since the sizes of the ranges provided vary considerably.

The second variable I created was type of healthcare service. As stated earlier, these operations came from three primary fields: hospitals, clinics and private practices. It would be interesting to see which states had the highest revenue and where that revenue was really coming from (click on link for the interactive version):

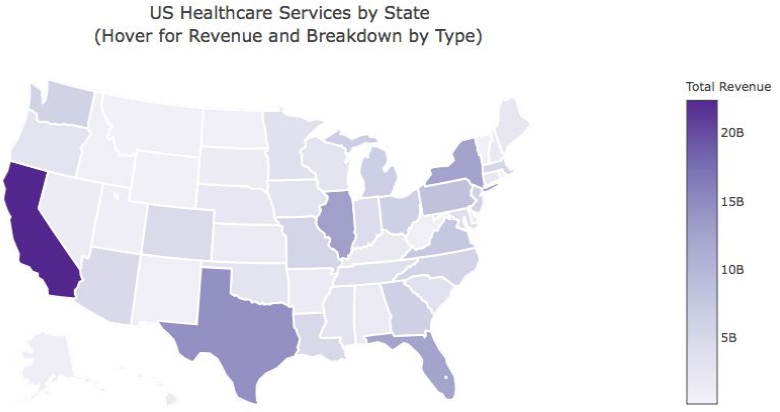

Fig 2.3.3

The highest revenue from healthcare comes from California over $20 Billion as shown in Fig 2.3.3. The next graph (Fig 2.3.4) shows the hospital to private practice ratio. It was originally hypothesized that a higher ratio would imply higher revenues, since hospitals are the biggest service providing body.

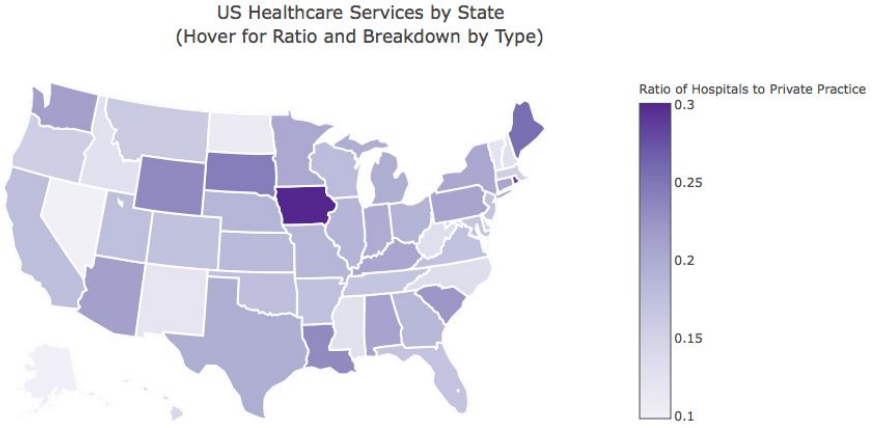

Fig 2.3.4

Interestingly, California has a lower hospital to private practice ratio, implying that private practices form the bulk of this revenue. In fact, a comparison of these figures shows that the highest revenue comes from places with lower hospital to private practice ratios. Stated differently, private practices appear to be cash cows and hospitals lag behind in terms of revenue. It is pertinent to note here that the private practices outnumber hospitals in every state. Therefore, a higher ratio does not imply a greater number of hospitals than private practices.

2.4 Comparing Hospitals, Clinics and Private Practices

It would also be interesting to see how years in business impacted the revenue per year. To get a proper understanding of this, I used the average revenue variable. However, due to a high degree of variability, the values were all over the place and the relationship was not very clear. I then proceeded to take the log of revenue, essentially creating a new variable that defined revenue in terms of powers of 10. So if a hospital make 10 million in a year, the power would be 7.

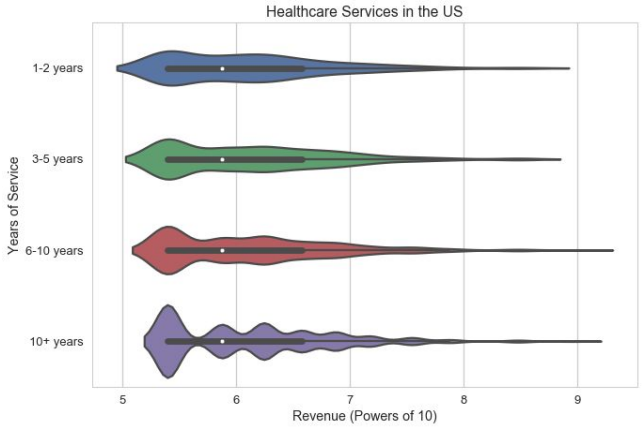

Fig 2.4.1

Fig 2.4.1 is an all-encompassing version of revenue versus years of service and captures all types of healthcare services. One can see that as the number of years in business increases, the mass of the distribution becomes more and more focused around one point. At the bottom of the graph, the purple violin representing ten years balloons at the beginning implying that a large chunk of the businesses older than 10 years earn around $10^(5.5). This revenue seems rather low for a healthcare business and this might point to inaccuracies in recording.

To further examine any differences that may exist between the three distinct types of healthcare services, the next figure (Fig 2.4.2) presents a small multiples plot showing the relationship between years in service and log(revenue) for each type of business. The length of the vertical line at each point represents the range of values and the dot represents a point estimate (mean).

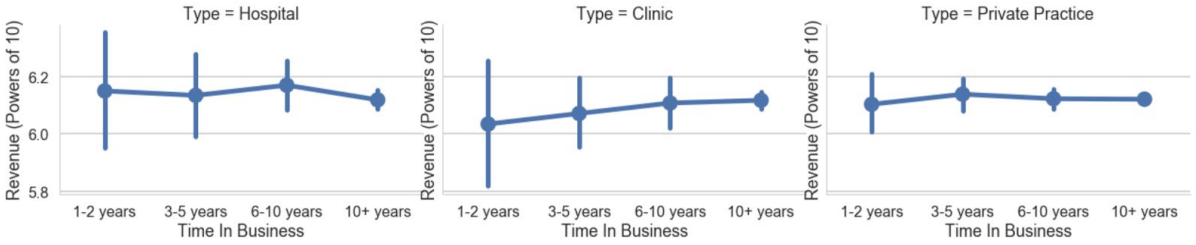

Fig 2.4.2

One thing that is common between each type of operation is the variability. The revenue seems to stabilize after 10 years. However, this might be attributed to the fact that we have much more data for 10+ years than we do for 1-2 years. It is also interesting to check if the top three states in terms of revenue follow the same pattern as exhibited above.

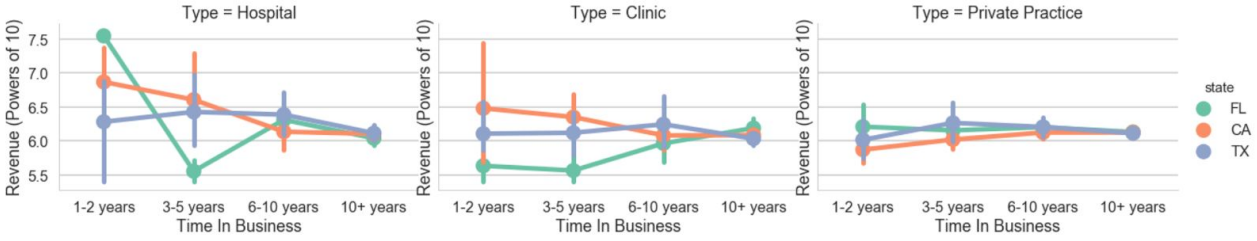

Fig 2.4.3

While the trend for private practices is very similar for each of Florida, California and Texas, the trend for clinics and hospitals is very different. In fact, the graph for Hospitals shows Florida experiencing a sharp decline in revenue for hospitals as the years in business increase from 1-2 to 3-5 years. Similarly, a decline in revenue is also seen for California as it moves towards the 3-5 years range. All types of healthcare services seem to converge to the same amount of revenue after 10+ years.

Another relationship of interest would be between the number of employees and average revenue, as presented in Fig 2.4.4.

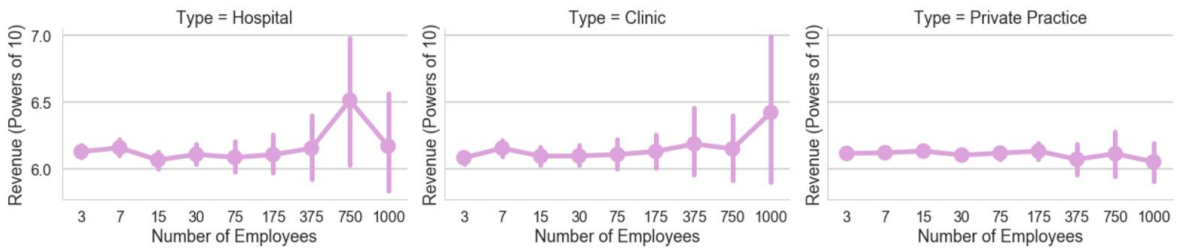

Fig 2.4.4

There isn't a very obvious trend in revenue with an increase in the number of employees. On the other hand, an increase in variability in revenues is more evident for hospitals and clinics as compared to private practices which manage to stay stable throughout. The values towards the left extreme for hospitals and towards the right extreme for private practices cast doubt on the integrity of this dataset. The number of employees for a private practice should not be anywhere close to 1000 and the number of employees for hospitals should be far above a mere 3.

The chloropleth-map in this link as well as those presented at the start indicate higher revenues in healthcare in states closer to the coast as compared to those inland. California appears to be a booming market for the healthcare industry. It has the highest number of healthcare service businesses overall as well as the highest number of recently established healthcare services. Private practice appears to be very high in demand considering the high ratio of private practices to hospitals. The next table shows various rankings for the top 3 cities in California in terms of number of healthcare services.

City	Rank by No. of Services	Rank by Revenue	Rank by Average Revenue
San Diego	1	1	29
Los Angeles	2	2	51
San Francisco	3	3	48

Table 2.4.1

San Diego has the highest number of healthcare services, followed by Los Angeles, and San Francisco. These cities also happen to have the highest numbers of private practices which account for the majority of revenue. Since private practices account for a large part of the services, the average revenue here is lower than places which have fewer services but larger bodies.

Conclusion

While the dataset used in the analysis contained 1 million rows and few missing values in comparison, the integrity of a large portion seemed doubtful. A lot of the data was mislabeled, additionally the headcount, revenue and even the addresses listed seemed outdated. A very important lesson learnt from this exercise was to never rely on one indicator variable to provide all the information. The true-valued fill rate is a very important consideration that is often disregarded. It goes further than just filtering very obviously wrong values. It is essential to validate data further by ensuring that each entry contains what it is supposed to contain, 5 numeric digits in the case of zip code for example.

The methods of filtering out data used in this exercise are very crude and resulted in a loss of information. There might be a lot of healthcare services whose names do not explicitly indicate that they are in fact such, e.g. Kaiser Permanente. Having said that, the use of string manipulation will always be key in data science exercises.

Lastly, a lot of this analysis has relied on the use of average values for the revenue ranges provided. Using these averages was necessary to create some of the visualizations presented. The analysis might change a great deal had actual values been used.