OCTOBER 4, 2019

ANALYSIS OF THE NATIONAL PLAN AND PROVIDER ENUMERATION SYSTEM

HEALTHCARE DATA MINING AND DATA ANALYSIS

QINGYUE SU
MSBA, BRANDEIS UNIVERSITY
qingyuesu@brandeis.edu

Table of Contents

Analysis of the National Plan and Provider Enumeration System

Introduction

This paper provides an overview of the US healthcare providers based on the states. Due to the large size of the raw dataset, I firstly filtered the data and narrowed it to a workable size. I will need a total of eight columns and more than six million records after omitting the NA. values, and the dataset is ready for further manipulation towards to each question. First, the information of the healthcare providers that related to our group members are collected, and the state in which they are first licensed were reported. The second part of the essay examines if the gender difference of building up a sole proprietorship exists in the healthcare industry. Only Female and male, and sole proprietorship on the field was tested during the data mining process. The gender analysis of choosing the higher risk position is being tested in part three, and I am going to find out from the filtered database (specific states that are assigned to our group) that if male doctors are more likely to choose the practice that is associated with a higher risk for a higher reward. The states that I manipulated are New York, New Jersey, Massachusetts, Alabama, Oklahoman, Kansas, Idaho, Delaware, and the District of Columbia. The last part of the report visualizes the MRI clients' density per one million population, per state on healthcare facilities bases all over the United States.

Find our own doctor

The table below shows the healthcare provider of six classmates, most of them are licensed in Massachusetts while one is licensed in Kansas. The name and other personal information are not recorded in order to protect our and doctors' privacy.

Group member (ordered LN)	Providers license state
Steve Hsu	MA
Louise Liu	KS
Sue Su	MA
Yuan Tian	MA
Sharon Tsai	MA
Chuyue Wu	MA

Table 1

Exploring gender difference in practicing as a "Sole Proprietor"

To test if there is a difference, I first need to manipulate the dataset and retrieve the

related data. By using Python, I filter the database to select the columns 'Provider Gender Code' and 'Is Sole Proprietor'. Then clean the data and remove empty rows and nasty data such as nan or “X”. Group the compiled data into four separate groups by both 'Provider Gender Code' and 'Is Sole Proprietor', and the number of variables is counted separately in each group as it shows in table 1. The result is falls into four categories: Female Sole Proprietor, Male Sole Proprietor, Female Not Sole Proprietor, and Male Not Sole Proprietor.

Provider Gender Code	Is Sole Proprietor	
F	N	269759
	Y	174937
M	N	155012
	Y	94744

Table 2

Table 2 is formatted into a 2*2 table to present the result clearer. After applying the Fisher’s Exact Test function (stats.fisher_exact()) to the table, the gender preferences of the observations are presented by the statistical result (table 3).

	N	Y
F	269759	174937
M	155012	94744

(0.9424984943590091, 9.729532887167638e-31)

Table 3

As I can see from the results above, the odds ratio is 0.94 which is below 1. P-value is statistically significant at the 1% level, in which the null hypothesis can be rejected. As a result, I can conclude that sole proprietor is different between genders, and females are being less likely to establish a solo practice office.

Note:

- Odd Ratio >1 indicates that females are more likely to establish a solo practice office
- Odd Ratio <1 indicates that females are less likely to establish a solo practice office

Gender differences in choosing practices

Now I am going to test is a doctor’s gender an influencing factor of his / her choices between practices with low risk and reward levels and those with high risk and reward levels.
The answer to the question above is yes, which means that gender really affects one’s decision to choose practices with different risk levels. Besides, our group supposes that male doctors are more

likely than their female peers to choose the practices that are associated with a higher risk for a higher reward.

Since it is a test about whether one factor has an influence on another factor, our group chooses to use Python to realize the Fishers' Exact Test on these two variables to testify (Database from NPPES website). Fisher's exact test is a statistical test used to determine if there are nonrandom associations between two categorical variables, which is suitable for this situation. Here exists two such variables gender and category of practice with the different risk level. Speaking of the category of gender, there are 68,556 females and 64,226 males in the margin in our group's sample, and in another category of risk level, there are 109,074 doctors with low risk and reward level and 23,708 doctors with high risk and reward level in the margin in these states assigned to our group. (Statistics are from calculations of code as below)

```
npiallrisk=npirisk.loc[npirisk['Healthcare Provider Taxonomy Code_1'].isin (['207V00000X', '208000000X', '208600000X ', '207X00000X'])]
npiallrisk = npiallrisk.dropna(axis=0,how='any')

npif = npiallrisk[npiallrisk['Provider Gender Code'] == 'F']
npif = npif.dropna(axis=0,how='any')

npim = npiallrisk[npiallrisk['Provider Gender Code'] == 'M']
npim = npim.dropna(axis=0,how='any')



npiallgender=npirisk.loc[npirisk['Provider Gender Code'].isin (['F', 'M'])]
npiallgender = npiallgender.dropna(axis=0,how='any')

npigenderlow = npiallgender[npiallgender['Healthcare Provider Taxonomy Code_1'].isin(['207V00000X','208000000X'])]
npigenderlow = npigenderlow.dropna(axis=0,how='any')

npigenderhigh = npiallgender[npiallgender['Healthcare Provider Taxonomy Code_1'].isin(['208600000X ', '207X00000X'])]
npigenderhigh = npigenderhigh.dropna(axis=0,how='any')
```

Since the marginal value of these two categorical variables is fixed, our group begins to test the observational value of the specific distribution of this relationship sheet, which means that I have to calculate the numbers in four situations of the combination of these two categories, which is like the table 4 below:

	High risk and high reward level	Low risk and low reward level	Margin
Female	a	b	68,556
Male	23,708 – a	109,074 – b	64,226
Margin	23,708	109,074	132,782

Table 4

Based on our hypothesis that gender really affects the choice to practices with different risk and reward level, which is the theory I want to testify, our group sets our null hypothesis as the effect is zero, which means that no matter what gender these subjects are, the proportion of high risk and reward level is the same. Then, our group uses the original dataset to verify whether this null hypothesis is acceptable or not. (Below are some of our group's code)

```
##question 3
npirisk = npi_data[['Provider Gender Code','Healthcare Provider Taxonomy Code_1']]
print(npirisk)
npirisk = npirisk.dropna(axis=0,how='any')
#low:Obstetrics & Gynecology is 207V00000X,"Pediatrics"-208000000X
#high risk: Surgery - 208600000X   Orthopaedic Surgery - 207X00000X
npilowrisk=npirisk.loc[npirisk['Healthcare Provider Taxonomy Code_1'].isin (['207V00000X', '208000000X'])]
print(npilowrisk)
b = npilowrisk.groupby(by=['Provider Gender Code'])['Healthcare Provider Taxonomy Code_1'].size()
print(b)

npihighrisk=npirisk.loc[npirisk['Healthcare Provider Taxonomy Code_1'].isin (['208600000X ', '207X00000X'])]
print(npihighrisk)
c = npihighrisk.groupby(by=['Provider Gender Code'])['Healthcare Provider Taxonomy Code_1'].size()
print(c)

obs2 = pd.DataFrame([[66868,1688], [42206,22020]])
new_col = ['low','high']
obs2.columns=new_col
obs2.index=['F','M']
print(obs2)

fisher_result = stats.fisher_exact(obs2)
print(fisher_result)
```

From the code above, I can tell that our group proceeds our testing in three steps. First, our group searches four specific disease names which are the domains of doctors on the website to find the Taxonomy code of these domains. Here, I use these domains to represent the level of different risks and rewards. The sum of Obstetrics & Gynecology and Pediatrics represents the practice with low risk and low reward. On the opposite, the sum of Surgery and Orthopaedic Surgery represents the practice with high risk and high reward.

Then, I use Python to filter the specific value of those four different situations as mentioned above, which are 66,868, 1,688, 42,206 and 22,020 respectively. (Shown as table 5 below, which is gained from the console of Python)

	low	high
F	66868	1688
M	42206	22020

Table 5

Last, I put these values into Fishers' Exact Test to check the degree I can reject or accept the null hypothesis that these two factors have no effects on each other. The results are shown below.

```
In [45]: fisher_result = stats.fisher_exact(obs2)
    ...: print(fisher_result)
(20.667550693023813, 0.0)
```

The number on the left is the odds ratio, which quantifies the strength of the association between these two variables, and the number on the right is the p-value of our sample, which means that when all these four margin values are fixed as before, the probability of showing the distribution of our sample or of much extremer situation is around 0.0, which is a very small value, especially compared with 0.5, which means that I can reject the null hypothesis under 95% confidence level, no matter

using two-sided test or one- sided test. So, I can tell that gender is really a factor to influence one's decision about choosing practices with a higher or lower level of risk and reward.

The p-value of this sample's Fishers' Exact Test is 0.0, which means that a very small probability event happens. Therefore, I have a very sound reason to reject the null hypothesis and verify our assumption, which means that male doctors are more likely than their female peers to choose the practices that are associated with a higher risk for a higher reward.

The density of national MRI centers by state population

I use python to extract data and visualize the result by heat maps that the performance of each state is clearly shown in both statistical and geographical aspects.

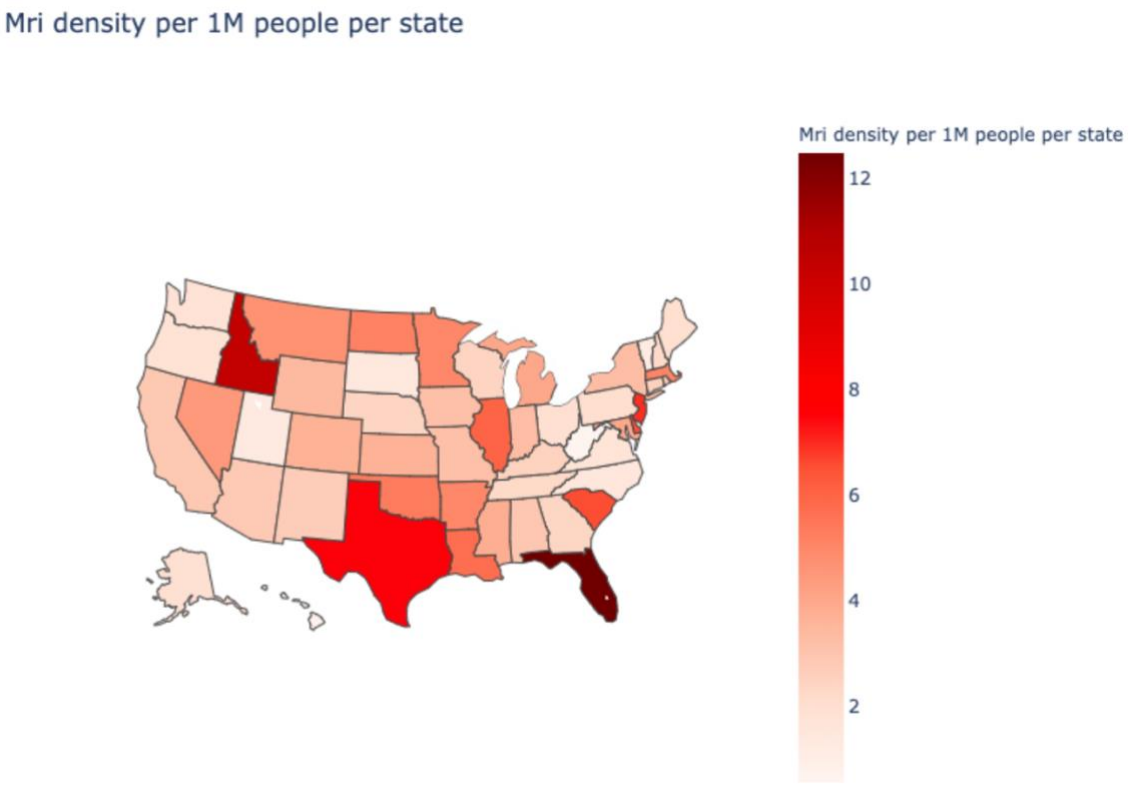
Firstly, I create a data frame in python where "Entity Type Code" is limited to 2 which represents that the data is limited on healthcare facilities, and "Healthcare Provider Taxonomy Code_2" is limited to "261QM1200X" that filtered only Magnetic Resonance Imaging (MRI) center. In this way, I extract all MRI centers in the US. To match with the map I will use later, the data of Puerto Rico and Guam have been excluded.

Second, I compute MRI density in python. I count the MRI centers in each state and match these data to the total population in each state (Wikipedia). By doing division, each state's MRI density has been figured out and shown below.

	state	count	population	mridensity
9	FL	266	21299325	12.49
43	TX	220	28701845	7.67
4	CA	113	39557045	2.86
14	IL	77	12741080	6.04
31	NJ	64	8908520	7.18
22	MI	40	9995915	4.00
35	OH	38	19542209	1.94
34	NY	38	11689442	3.25
19	MA	35	6902149	5.07

Table 6

In the last step, I use the package called *'plotly'* in python to draw the heat maps. The figure is shown below.



From the heat map, I find that most states in the US have less than 6 MRI centers per 1,000,000 people. But in some states, such as Florida, Idaho, Texas and South Carolina, there are more MRI centers than other states. Especially in Florida, the MRI density reaches to 12.49 per 1,000,000 people.

It makes sense because, on the one hand, Florida is a retirement destination for many Americans. Building more MRI centers could provide those people with better healthcare services, which could bring more and more retired people there to boost economic development in Florida. On the other hand, Florida's medical device manufacturing ranks second in the United States and it has developed medical systems, which could also explain why the density of Florida is so high. In addition, I found that the MRI in Florida is less expensive than the other states in the U.S. Hence, the high MRI density in Florida may because the facilities, taxes are inexpensive there so that the suppliers could build up more MRI centers with a less expensive service.

Conclusion

The dataset provides us with a good example of the supply side of the US healthcare model. According to the tests and visualization I did in the report, I have an overall understanding of today's healthcare providers. By retrieving our own doctor from the dataset, a real-world connection between us and the healthcare industry is built up.

Then I explored the practical issue about the relationship between gender and sole proprietorship that I am able to dig into the dataset and apply our data analysis skills to real-world problems. I find out that females are less likely to establish their own office, and I suppose that the difference not only shows in the healthcare industry. More analysis of this social condition could be done in the future. Meanwhile, for the chosen states that are included in part three, I conclude that more male health providers are willing to choose the practices of higher risks but with higher rewards. This is another examination based on the gender differences in the healthcare industry that this result may be influenced by many aspects. One possible reason may come from culturally instilled. For instance, men are viewed to be more masculine is they take more risk, and the social pressure is pushing people to obey gender norms.

The analysis of the density of MRI centers in the United States is more intuitive because I use the visualization of heat maps to present our results. As I concluded in the last section, I am aware of the distribution of MRI centers. Florida is the place that has the highest density of MRI per 1000,000 people, followed by Idaho, Texas, and New Jersy. However, the test is based on the facility bases instead of the MRI machine bases, the real situation may still differ from the result.

Work Cited

"List of U.S. states_by_population." 3rd. November 2019. Web. 4th. November 2019. https://simple.wikipedia.org/wiki/List_of_U.S._states_by_population

"The perfect climate for business in Florida" https://www.enterpriseflorida.com/wp-content/uploads/Floridas-Business-Advantages- Chinese.pdf