Analysis Report of National Plan and Provider Enumeration System (NPPES)

Team 2

Declaration

The data of states Florida, Georgia, Arizona, Colorado, Connecticut, Mississippi, Hawaii, Rhode Island was explored for Q2-Q3.

Q1- Find your own doctor!

Last Name of Team Member	PCP's First License State
Bai	MA
Cao	MA
Chen	MA
Dai	MA
Han	MA
LI	MA

Question 2 - Gender difference in practicing as a “Sole Proprietor”

Data Processing Steps: To test the hypothesis, we first extracted the columns of gender and sole proprietor from the giant dataset through python. Then we counted the number of people in each of the four groups: (Male, Yes), (Female, Yes), (Male, No), (Female, No), which was conducted by SQL. Finally, we constructed the 2 * 2 table below, and conduct the Fisher Exact Test in Python, using “fisher_exact” function in “scipy.stats” package.

Results:

Item	Male	Female
Sole Proprietor YES	87,461	154,862
Sole Proprietor NO	171,578	272,773

P-value: 1.89e-94

Conclusion: From the results of Fisher Exact Test, whose p-value is less than 5% that we can reject the null hypothesis and conclude that Gender has impacts on practicing as a “Sole Proprietor”. Generally, in healthcare, male and tends not to be Sole Proprietor.

Question 3 - Hypothesis test

Data Processing Steps: In order to test the hypothesis, firstly we are committed to clean and filter the data. We extract the data that indicates individuals, gender and assigned states into a new data set. Then we count the number of people group by gender who are conducting low or high risk-award practices by filtering low risk-award practices including “Obstetrics & Gynecology” and “Pediatrics” with Taxonomy code '207V00000X','208000000X' and high risk-award practices including “Surgery” and “Orthopaedic Surgery” with Taxonomy code '208600000X','207X00000X' .

As a result, we get the following table from the data set:

	Female	Male
Low risk-award	3341	2930
High risk-award	371	2898

Fisher's Exact Test for Count Data

data: Test

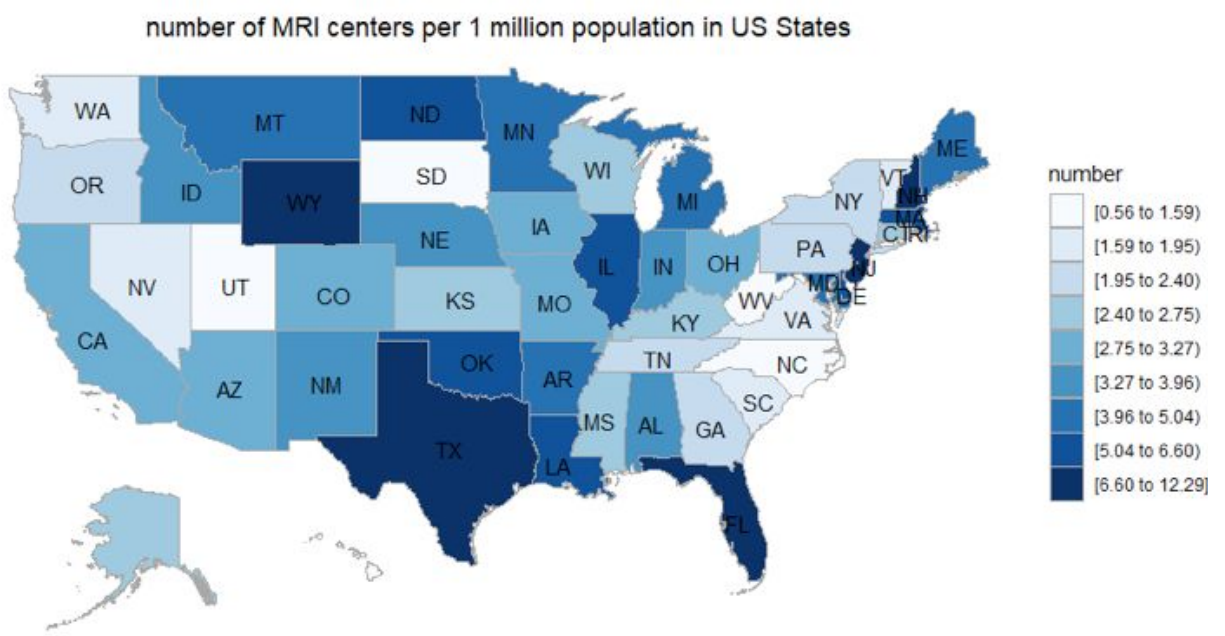
p-value < 2.2e-16

Conclusion: By conducting Fisher Exact Test, we are happy to see that the p-value is so small that we can reject the null hypothesis and conclude 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.

Question 4 - National heat map of MRI centers per 1 million population

Data Processing Steps: We firstly split whole dataset file into 8 csv files using shell and then uploaded to sqlite database for data exploration and filtering the data we needed. After that we downloaded the data and used R for data visualization. For the US population data, we

downloaded the data from world population review and employed the 2019 US population data.



Interpretation: From the above heatmap, we can know that states Florida, Wyoming, Texas and New Jersey have outstanding higher MRI centers per 1 million population while states Utah and South Dakota have obvious low MRI centers per 1 million population. From the data, we can also know Florida has the highest number of MRI centers per 1 million population with most MRI numbers (266), which is 12.29.

Possible Reasons: One reason for the high number might be that Florida has one of America's first outpatient MRI centers and the founder Robert Kagan has long been committed to evangelize and promote MRI scanning in Diasonics and Technicare. Besides, Florida has a vibrant health care sector that encompasses one of the largest and most sophisticated health care systems in the country. According to the FDA, Florida ranks 2nd in the U.S. for the number of FDA-registered medical device establishments. In addition, Florida's two leading research universities, Florida State University and the University of Florida provide advanced healthcare services and promote health and technology research.

Appendix – Code

Question2

Python

```
import pymysql
```

```
import pandas as pd
from sqlalchemy import create_engine
df1 =
pd.read_csv("/Users/baixuhui/Desktop/Healthcare/NPPES_Data_Dissemination_October_2019/
healthcare_full.csv", sep=',', usecols=[41,307])

db = pymysql.connect(host='localhost',
user='root',
password='******',
db='db_healthcare',
charset='utf8mb4',
cursorclass=pymysql.cursors.DictCursor)
engine = create_engine('mysql+pymysql://root:******@localhost/db_healthcare')

pd.io.sql.to_sql(df1, 'problem2', con=engine, index=False, if_exists='replace')
```

SQL:

```
SELECT count(*) FROM problem WHERE gender = 'M' and sole = 'Y';
SELECT count(*) FROM problem WHERE gender = 'M' and sole = 'N';
SELECT count(*) FROM problem WHERE gender = 'F' and sole = 'Y';
SELECT count(*) FROM problem WHERE gender =  'F' and sole = 'N';
```

Python (For Fisher's Exact Test):

```
import scipy.stats as stats
oddsratio, pvalue = stats.fisher_exact([[87461,154862],[171578,272773]])
pvalue
```

Question 3

R

```
library(tidyverse)

npidata_pfile_20050523.20191013<-read.csv('npidata_pfile_20050523-20191013.csv')

npi=npidata_pfile_20050523.20191013

npi_filter<-npi%>%filter(npi$Entity.Type.Code==1 &
npi$Provider.Business.Practice.Location.Address.State.Name %in%
c("FL","GA","AZ","CO","CT","MS","HL","RI")& npi$Provider.Gender.Code %in% c("M","F"))
```

```
low_M<-npi_filter%>%filter((npi_filter$Healthcare.Provider.Taxonomy.Code_1 %in%
c('207V00000X','208000000X')) & (npi_filter$Provider.Gender.Code== "M") )%>% tally()

low_F<-npi_filter%>%filter((npi_filter$Healthcare.Provider.Taxonomy.Code_1 %in%
c('207V00000X','208000000X')) & (npi_filter$Provider.Gender.Code== "F") )%>% tally()

high_M<-npi_filter%>%filter((npi_filter$Healthcare.Provider.Taxonomy.Code_1 %in%
c('208600000X','207X00000X')) & (npi_filter$Provider.Gender.Code== "M") )%>% tally()

high_F<-npi_filter%>%filter((npi_filter$Healthcare.Provider.Taxonomy.Code_1 %in%
c('208600000X','207X00000X')) & (npi_filter$Provider.Gender.Code== "F") )%>% tally()

Test <-
  matrix(c(low_F$n,high_F$n,low_M$n, high_M$n),
         nrow = 2,
         dimnames = list(c("low risk-award", "high risk-reward"),
                         c("Female", "Male")))
fisher.test(Test)
```

Question 4

PowerShell:

Split the csv file into 8 files

```
$i=0; Get-Content E:\MBA@Brandeis\csvsplit\npidata_pfile_20050523-20191013.csv
-ReadCount 800000 | %{$i++; $_ | Out-File E:\MBA@Brandeis\csvsplit\npidata_sp_$i.csv}
```

SQLite

create MRI list:

```
create table npi_MRI_list as
SELECT NPI,
t.[ProviderOrganizationName(LegalBusinessName)] org_name,
t.ProviderBusinessPracticeLocationAddressStateName practice_state
FROM npidata_sp_1_new t
WHERE entitytypecode = '2' AND
HealthcareProviderTaxonomyCode_1 = '261QM1200X'
UNION
SELECT NPI,
t.[ProviderOrganizationName(LegalBusinessName)] org_name,
```

```
t.ProviderBusinessPracticeLocationAddressStateName practice_state
FROM npidata_sp_2_new t
WHERE entitytypecode = '2' AND
HealthcareProviderTaxonomyCode_1 = '261QM1200X'
UNION
SELECT NPI,
t.[ProviderOrganizationName(LegalBusinessName)] org_name,
t.ProviderBusinessPracticeLocationAddressStateName practice_state
FROM npidata_sp_3_new t
WHERE entitytypecode = '2' AND
HealthcareProviderTaxonomyCode_1 = '261QM1200X'
UNION
SELECT NPI,
t.[ProviderOrganizationName(LegalBusinessName)] org_name,
t.ProviderBusinessPracticeLocationAddressStateName practice_state
FROM npidata_sp_4_new t
WHERE entitytypecode = '2' AND
HealthcareProviderTaxonomyCode_1 = '261QM1200X'
UNION
SELECT NPI,
t.[ProviderOrganizationName(LegalBusinessName)] org_name,
t.ProviderBusinessPracticeLocationAddressStateName practice_state
FROM npidata_sp_5_new t
WHERE entitytypecode = '2' AND
HealthcareProviderTaxonomyCode_1 = '261QM1200X'
UNION
SELECT NPI,
t.[ProviderOrganizationName(LegalBusinessName)] org_name,
t.ProviderBusinessPracticeLocationAddressStateName practice_state
FROM npidata_sp_6_new t
WHERE entitytypecode = '2' AND
HealthcareProviderTaxonomyCode_1 = '261QM1200X'
UNION
SELECT NPI,
t.[ProviderOrganizationName(LegalBusinessName)] org_name,
t.ProviderBusinessPracticeLocationAddressStateName practice_state
FROM npidata_sp_7_new t
WHERE entitytypecode = '2' AND
HealthcareProviderTaxonomyCode_1 = '261QM1200X'
UNION
SELECT NPI,
t.[ProviderOrganizationName(LegalBusinessName)] org_name,
t.ProviderBusinessPracticeLocationAddressStateName practice_state
```

```
FROM npidata_sp_8_new t
WHERE entitytypecode = '2' AND
HealthcareProviderTaxonomyCode_1 = '261QM1200X';
```

Population Table upload

Population_data

R

```
setwd("E:/MBA@Brandeis/Syllabus/193HS-256F Healthcare Data Analytics and Data
Mining/Hw1")
getwd()
library(ggplot2)
library(dplyr)
library(openintro)
library(choroplethr)

mri <- read.csv("MRI_List.csv", header = T)
pop <- read.csv("population_data.csv", header = T)
str(mri)
mri_num <- mri %>% group_by(practice_state) %>% summarise(n = n())
mri_num$practice_state <- abbr2state(mri_num$practice_state)
mri_num$practice_state <- tolower(mri_num$practice_state)
pop$State <- tolower(pop$State)
mri_pop <- merge(mri_num,pop,by.x = "practice_state",by.y = "State")
map_mri <- mri_pop %>%
            select(practice_state,n,Pop) %>%
            mutate(mri_perm =  round(n/(Pop/1000000),2))
colnames(map_mri) <- c('region','n','pop','value')
state_choropleth(map_mri,
                 title = "Number of MRI centers per 1 million population in US States",
                 legend = 'number',
                 num_colors = 9)+
  theme(plot.title = element_text(hjust = 0.5))
```