Report on Provider Data Analysis

Introduction

In this project, our team used Python, SQL and Tableau for analyzing and visualizing the data. The dataset consists of healthcare provider information, which was extracted from the National Plan and Provider Enumeration System (NPPES) and we used the updated Full Replacement Monthly NPI File from January 2019 for the United States. As a team, we were assigned to study nine states - California, Ohio, Washington, Maryland, Louisiana, Iowa, New Mexico, Maine, Alaska, Texas, and Georgia.

Question 1

Student Name	**Provider Name**	**NPI**	**State**
Malavika Andavilli	Tracy Haradon	1023011228	MA
Zhiqi Chen	Martha Brady	1336147966	MA
Debarati Mazumdar	Linda Chung	1053314211	CA
Salil Redkar	Debra Reid	1104813856	MA
Nkosingiphile Shongwe	Hyejon Ko	1639211154	TX

Question 2

For this question, we used python to perform data analysis using cross tabulation & the chi-square test.

Is Sole Proprietor	**N**	**X**	**Y**	**All**
Provider Gender Code				
F	326444	9089	187486	523019
M	206026	12480	120554	339060

All	532470	21569	308040	862079

Null hypothesis: There is no difference in type of practice based on gender.

Alternative hypothesis: There is a difference in type of practice based on gender.

The chi-square statistic is 3199.5370053443967 and the p-value is 0.0,
with degrees of freedom as 2 (refer to appendix). The null hypothesis for this question was that there is no difference in types of practice based on gender, which was rejected with a significantly low p-value obtained after performing the chi-square test. In other words, women tend to choose to be a sole proprietor than men. Further investigation may be needed to determine the causes that dictate these preferences.

Question 3

Healthcare Provider Taxonomy Code_1	**hrhr**	**lrlr**	**All**
Provider Gender Code			
F	**1874**	**15091**	**16965**
M	**9832**	**9368**	**19200**
All	**11706**	**24459**	**36165**

Key:

hrhr = High risk, high reward

lrlr = Low risk, low reward

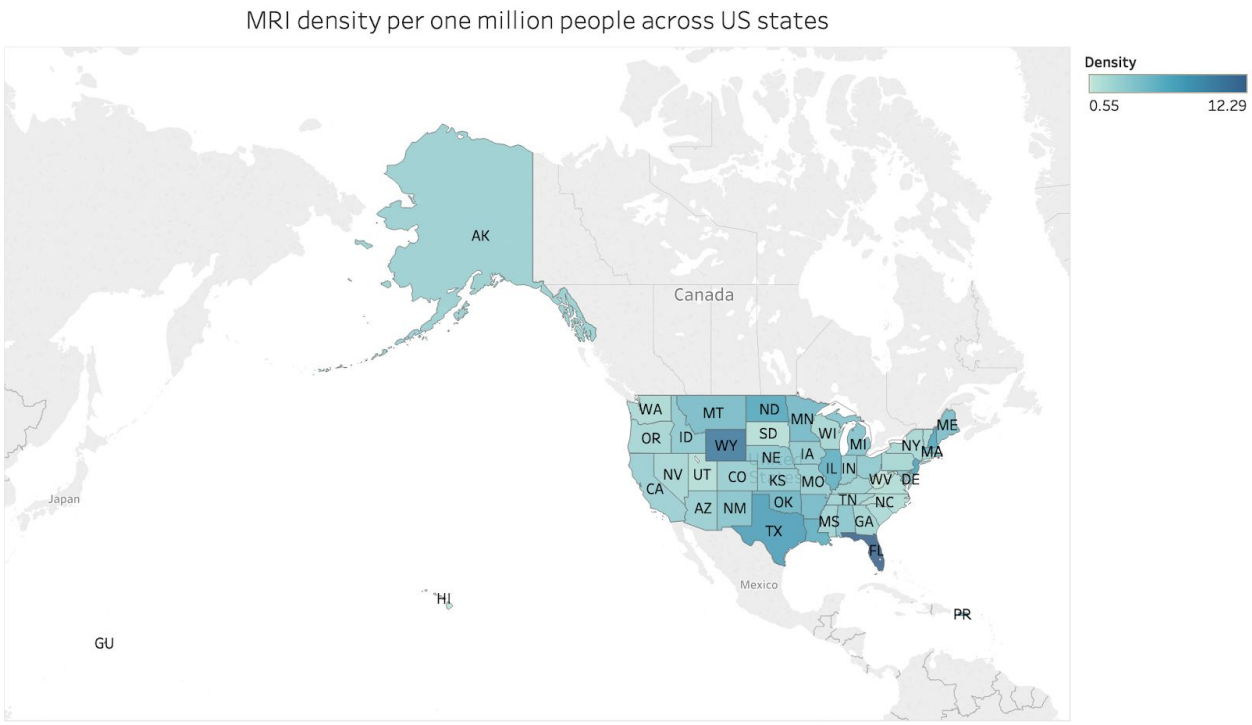
Null hypothesis: There is no difference between the genders in choosing between high risk, high reward and low risk, low reward medical specializations.

Alternative hypothesis: There is a statistically significant difference in the genders when choosing between high risk, high reward and low risk, low reward medical specializations.

Men prefer high risk high reward practices than women based on the states assigned (refer to appendix). In other words, on average, men choose medical specializations that have a higher risk and also offer a substantially larger compensation compared to women. This was confirmed using the chi-square statistic, which is 6634.50 with 1 degree of freedom (refer to appendix). The p-value is approximately 0.00, which suggests that there is a difference between the genders when choosing specialization. It will be worthwhile to determine what factors deter women from choosing the high risk, high reward specializations. This might be evidence of an unfortunate systemic pressure on women to choose certain specializations.

Question 4

For this question, we first used Tableau to filter "Entity Type Code = 2" and "Health Care Provider Taxonomy Code 1 = 261QM1200X" for MRI. We deleted all other columns except for one column called "Provider Business Practice Location Address State Name" and exported the dataset to a csv file. Using this new dataset, we used R to group and summarize the MRI centers per state. We then used Excel to calculate the number of MRI clinics per 1,000,000 people using state population figures from world population review (http://worldpopulationreview.com/states/). And the population of Puerto Rico and Guam were retrieved from worldometers (http://www.worldometers.info/world-population/puerto-rico-population/). We then used this dataset in Tableau to construct a heatmap. We represented MRI density intensity by color in each state.



Summary Statistics

Sum:	193.64
Average:	3.72
Minimum:	0.55
Maximum:	12.29
Median:	3.11
First quartile:	2.05
Third quartile:	4.97

The MRI density has a range from 0.55 to 12.29 in the United States. The average MRI density is 3.72, which means that most states have a density of above four MRI's per one million population. Even among the states with higher populations, density values vary considerably from state to state.

If we compare the MRI density with the population density in each state, we can say that California has the highest population density but is low on MRI density (2.79); this may be caused by younger people living there since it has more companies hence more working population. Another state is Texas which has high population density and high MRI density (7.28); this can be because of many senior people living in Texas since age is a factor of many illnesses (where MRI centers are used).

The MRI density is a statistic that shows the number of magnetic resonance imaging (MRI) units in selected the 52 states including Guam and Puerto Rico. As of 2019, there were nearly 3.72 on average such devices per one million inhabitants in the United States, where twenty-five percent of the states have about 2.05, and seventy-five percent around 4.97. Looking at the summary statistics, Florida has one of the highest densities of such devices at 12.29, and Alaska the lowest densities at 0.55. The intensity of each density is shown by the different colors, and from the map, we can see that only two states are above ten, and four states between five and nine. All other states are below five.

Florida has a largely older population, mostly retired and pensioners (http://worldpopulationreview.com/states/florida-population/). There may be more MRI's because falls and fractures are more common in a geriatric population so there's more demand for them.

Conclusion

In this project, we used The National Plan and Provider Enumeration System (NPPES) database that reports the status of the US healthcare providers. This database is updated frequently to reflect the most recent status of each and every healthcare provider in terms of their demographics such as address, license number, and nature of practice. In this project, we downloaded the Full Replacement Monthly NPI File as of January 2019. We used this data to investigate questions regarding gender, reward, and density of MRIs. In some questions, specifically two and three, we used nine distinct states for the analysis. These states are

California, Ohio, Washington, Maryland, Louisiana, Iowa, New Mexico, Maine, Alaska, Texas, and Georgia. We filtered the data to investigate individual providers or healthcare facility depending on what question we were answering. In order to learn more about the data, we first looked for our doctors using SQL, and the license number. This was useful in introducing us to the data, and being familiar with the codes.

For question two, we explored whether there is a difference in gender if the individual providers are operating as sole proprietors using the Fisher's Exact statistical test. We discovered that based on the p-value, there is a gender difference if a provider is considered a sole proprietor, whereby about 60.8% of females operate as sole proprietors. For the third question, we expanded the question to investigate whether the type of practice in terms of risk and reward has a gender difference. The results show that there is a statistical difference in genders when choosing a practice. Men tend to choose the high risk high reward practice while women prefer the low risk low reward practice.

Research shows that the use of MRIs in the United States is rapidly increasing when compared to other developed countries (https://www.healthsystemtracker.org/chart/density-mri-units-increased-rapidly-u-s-comparable-countries/#item-start). As of 2017, there were nearly 38 such devices per one million inhabitants in the United States, making it one of the highest densities of such devices (https://www.statista.com/statistics/282401/density-of-magnetic-resonance-imaging-units-by-country/). As a result, our team further invested the density of MRIs per one million people in healthcare facilities. In healthcare facilities, the average MRI density per one million people in approximately 3.72, and the number varies for each state from 0.55 to 12.29. Most interestingly, we found that Florida has the highest density because of the population demographics, whereby Florida has a high proportion of the elderly compared to other states.

Appendix

Question 2

In [107]:
```
import numpy as np
import pandas as pd
import scipy.stats as stats

# Importing the data
df = pd.read_csv('npidata_pfile_20050523-20190113.csv',usecols=['Provide
r Gender Code','Is Sole Proprietor','Entity Type Code','Provider License
 Number State Code_1'])
```

In [108]:
```
# Filtering based on entity code:
df=df.loc[df['Entity Type Code']== 1.0]

#Filtering the Data based on states:

state_list= ['CA','OH','WA','MD','LA','IA','NM','ME','AK']
df=df.loc[df['Provider License Number State Code_1'].isin(state_list)]
```

In [109]:
```
df.head(n=10)
```

Out[109]:

	Entity Type Code	Provider Gender Code	Provider License Number State Code_1	Is Sole Proprietor
16	1.0	M	OH	N
28	1.0	F	LA	Y
53	1.0	M	OH	Y
56	1.0	F	OH	N
59	1.0	F	OH	N
87	1.0	M	ME	N
94	1.0	M	CA	X
98	1.0	F	CA	N
120	1.0	M	IA	N
123	1.0	M	NM	Y

In [110]:
```
# Making the Cross- Tabulation
contingency_table = pd.crosstab(
    df['Provider Gender Code'],
    df['Is Sole Proprietor'],
    margins = True
)

contingency_table
```

Out[110]:

Is Sole Proprietor	N	X	Y	All
Provider Gender Code				
F	326444	9089	187486	523019
M	206026	12480	120554	339060
All	532470	21569	308040	862079

In [111]:
```
f_obs = np.array([contingency_table.iloc[0][0:3].values,
                  contingency_table.iloc[1][0:3].values])
f_obs[0:2]
```

Out[111]:
```
array([[326444,   9089, 187486],
       [206026,  12480, 120554]])
```

In [112]:
```
ans=stats.chi2_contingency(f_obs)[0:3]
"The chi-square statistic is {} and the p-value is {}, with degrees of f
reedom as {}".format(ans[0],ans[1],ans[2])
```

Out[112]:
```
'The chi-square statistic is 3199.5370053443967 and the p-value is 0.0,
with degrees of freedom as 2'
```

Question 3

In [113]:
```
df2 = pd.read_csv('npidata_pfile_20050523-20190113.csv',usecols=['Provid
er Gender Code','Healthcare Provider Taxonomy Code_1','Entity Type Code'
,'Provider License Number State Code_1'])
df2.head()
```

Out[113]:

	Entity Type Code	Provider Gender Code	Healthcare Provider Taxonomy Code_1	Provider License Number State Code_1
0	1.0	M	207X00000X	NE
1	1.0	M	207RC0000X	FL
2	2.0	NaN	251G00000X	NC
3	1.0	M	2085R0202X	TX
4	1.0	M	174400000X	TX

In [114]:
```
# Filtering based on entity code:
df2=df2.loc[df2['Entity Type Code']== 1.0]

#Filtering the Data based on states:

state_list= ['CA','OH','WA','MD','LA','IA','NM','ME','AK']
df2=df2.loc[df2['Provider License Number State Code_1'].isin(state_list
)]

df2.head(n=10)
```

Out[114]:

	Entity Type Code	Provider Gender Code	Healthcare Provider Taxonomy Code_1	Provider License Number State Code_1
16	1.0	M	207Q00000X	OH
28	1.0	F	207Q00000X	LA
53	1.0	M	208100000X	OH
56	1.0	F	207R00000X	OH
59	1.0	F	363L00000X	OH
87	1.0	M	208600000X	ME
94	1.0	M	213ES0103X	CA
98	1.0	F	363AM0700X	CA
120	1.0	M	207X00000X	IA
123	1.0	M	174400000X	NM

Conversion of codes

In [115]:
```
code_list= ['207V00000X','208000000X','208600000X','207X00000X']

df2=df2.loc[df2['Healthcare Provider Taxonomy Code_1'].isin(code_list)]

high_rr=['207X00000X','208600000X' ]
low_rr=['207V00000X', '208000000X']

df2.head()
```

Out[115]:

	Entity Type Code	Provider Gender Code	Healthcare Provider Taxonomy Code_1	Provider License Number State Code_1
87	1.0	M	208600000X	ME
120	1.0	M	207X00000X	IA
223	1.0	F	208000000X	OH
279	1.0	M	207X00000X	CA
330	1.0	M	208000000X	CA

In [116]:
```
df2['Healthcare Provider Taxonomy Code_1']= df2['Healthcare Provider Tax
onomy Code_1'].replace(high_rr,'hrhr')
df2['Healthcare Provider Taxonomy Code_1']= df2['Healthcare Provider Tax
onomy Code_1'].replace(low_rr,'lrlr')
df2.head()
```

Out[116]:

	Entity Type Code	Provider Gender Code	Healthcare Provider Taxonomy Code_1	Provider License Number State Code_1
87	1.0	M	hrhr	ME
120	1.0	M	hrhr	IA
223	1.0	F	lrlr	OH
279	1.0	M	hrhr	CA
330	1.0	M	lrlr	CA

In [117]:
```
# Making the Cross- Tabulation
contingency_table = pd.crosstab(
    df2['Provider Gender Code'],
    df2['Healthcare Provider Taxonomy Code_1'],
    margins = True
)

contingency_table
```

Out[117]:

Healthcare Provider Taxonomy Code_1	hrhr	lrlr	All
Provider Gender Code			
F	1874	15091	16965
M	9832	9368	19200
All	11706	24459	36165

In [118]:
```
f_obs = np.array([contingency_table.iloc[0][0:2].values,
                  contingency_table.iloc[1][0:2].values])
f_obs[0:2]
```

Out[118]:
```
array([[ 1874, 15091],
       [ 9832,  9368]])
```

In [119]:
```
ans=stats.chi2_contingency(f_obs)[0:3]

"The chi-square statistic is {} and the p-value is {}, with degrees of f
reedom as {}".format(ans[0],ans[1],ans[2])
```

Out[119]:
```
'The chi-square statistic is 6634.5011218984555 and the p-value is 0.0,
with degrees of freedom as 1'
```

Question 4 Code in R

```
library(readxl)
Q4_DATA <- read_excel("Desktop/Q4_DATA.xlsx")
MRI = Q4_DATA %>%
  group_by(`Provider Business Practice Location Address State Name`) %>%
  summarise(n = n())
write.csv(MRI, "Desktop/Q4_Tableau.csv")
```