Econometrics 322 Lab #3

Descriptive Statistics in Econometrics

Prof. Paczkowski

Enter your Name in the Next Cell

Maanya Tandon

Grading Rubric

Score:___: Max(0, 20 - Total Deductions)

Content Area	Deduction	Times Deducted	Check	Comments
Abstract				
Missing	5		[ ]	
Insufficient/Wrong Focus	1		[ ]	
Data Dictionary (Metadata)				
Missing	5		[ ]	
Insufficient/Wrong Form or Wording	1		[ ]	
Graphs				
Missing	5		[ ]	
Missing Title	1		[ ]	
Missing/Wrong Labels	1		[ ]	
Pre-Lab				
Missing	5		[ ]	
Insufficient/Wrong Answer	2 Each		[ ]	
No/Incorrect/Insufficient Model Specification	2		[ ]	
No/Incorrect Statistical Hypothesis Statement	2 Each		[ ]	
Post-Lab				
Missing	5		[ ]	
Insufficient/Wrong Answer	2 Each		[ ]	
Correlations				
Missing	5		[ ]	
Insufficient/Wrong Analysis	2		[ ]	
Missing Graph	2		[ ]	
Estimations				
Missing	5		[ ]	
No or incorrect discussion/interpretation of...				
Hypothesis tests and p-values	2 Each		[ ]	
R^2	2		[ ]	
F-Statistic	2		[ ]	
Multicollinearity/VIF	2		[ ]	
Heteroskedasticity/Test	2		[ ]	
Autocorrelation/Test	2		[ ]	
No/insufficient model selection	2		[ ]	
Elasticities				
Missing	5		[ ]	
Incorrect Interpretation	2		[ ]	
Missing Summary Table	2		[ ]	
Model Portfolio			[ ]	
Missing	5		[ ]	
General Comments:				

Contents

Collaboration Policy

Back to Contents

```
1. Study groups are allowed but I expect students to understand and complete their own assignmen
ts and to hand in one
assignment per student.
2. If you worked in a group, please put the names of your study group in the following table.
3. Just like all other classes at Rutgers, the student Honor Code is taken seriously.
```

Collaborator(s) Name(s)
name(s) here

Introduction

Back to Contents

Purpose

Back to Contents

This lab will introduce you to descriptive statistics in Pandas.

At the end of this lab, you will be able to:

- merge two DataFrame;
- calculate summary statistics;
- graph data; and
- interpret the results.

Problem

Back to Contents

Health-care is an important topic in our society and a major area of study and analysis in economics. One issue is the distribution of physicians and nurses, the *health-care providers*, around the country. Do some areas, e.g., states snd regions, have more health-care providers than others? What is the geographic distribution of health-care providers?

Assignment

Back to Contents

Use the following data:

- Physicians and nurses by state; and
- State-Region mapping.

The data files are available in the *Resouces* tab for this lab.

Use the references in Lesson 3.

Documentation

Back to Contents

Abstract

Back to Contents

In doing this lab I was able to learn a lot about how to analyze data when comparing it to a population. I had to adjust the number of nurses and physicians in comparison to the population so that I could compare the number of nurses and physicians in a state or region accurately. I also learned about testing a set of data for normality and interpreting the results of this. I did this by interpreting the skewness and kurtosis of a set of data points.

Data Dictionary

Back to Contents

Variable	Values	Source	Mnemonic
Physicians	Number	Resources on Sakai	Physicians
Nurses	Number	Resources on Sakai	Nurses
Physicians (per million people)	Number	Calculated by Physicians / (10^6)*Population	Physicians (per mil)
Nurses (per million people)	Number	Calculated by Nurses / (10^6)*Population	Nurses (per mil)
State	Each State is a separate word	Resources on Sakai	State
Region	Each region is a separate word	Resources on Sakai	Region
Population	Number	Resources on Sakai	Population
Population (per million people)	Number	Calculated by Population / (10^6)	Population (per mil)

Pre-lab Questions

Back to Contents

Before you do any work, please think about the relationship among these variables. In particular, think how you would answer the following if called on in class.

What type of data is this and why (i.e., source and domain)?

This is a secondary data source because it is data that was collected by the US government. I sourced the data from Sakai. Someone else collected this data for their own purposes and understanding, and I am using the data for my own purposes.

What pattern do you expect to see for physicians by state? Explain your answer.

In states with a higher population, I expect to see a higher number of physicians and nurses. I would also expect to see more physicians and nurses in the northeast because the northeast is more urban and has more industrial production than the other regions. Because the northeast has more urban areas and industrial production, there is a higher amount of wealh in flow, so there is more money to pay for doctors.

What should you do to the physician and nurse data before you do any analytical work? Think carefully about this.

The physician and nurse data should be adjusted to account for the differences in population in each state or region. This can be done by adding the physicians to make it an amount per capita. Dividing the number of physicians by the total population of a state will give us a decimal. Instead, I will use the number of physicians per a million people for a larger number that is easier to analyze. I will do the same for the nurse data.

Tasks and Questions

Back to Contents

Load the Pandas and Seaborn packages and give them aliases.

In [47]:

```
##
import pandas as pd
import numpy as np
##
##  load graphic packages
##
import seaborn as sns
sns.set(rc={'figure.figsize':(11.7,8.27)})
import matplotlib.pyplot as plt
%matplotlib inline
##
##  modeling packages
##
import statsmodels as sm
##
from statsmodels.graphics.gofplots import qqplot
```

Import the physician and nurse data. Set the row index to the state names.

In [48]:

```
path = r'/Users/maanyatandon/Documents/fall2020/econometrics/Lab3/'
df_phy = pd.read_excel(path+'physicians.xlsx', sheet_name = '2009' )
df_phy.set_index('State', inplace = True);
```

Print the first five (5) records.

In [49]:

```
df_phy.head()
```

Out[49]:

State	Physicians	Nurses
Alabama	10265	42880
Alaska	1574	5010
Arizona	14051	38570
Arkansas	5902	23050
California	100131	233030

Import the state information (see Lesson #3). Set the row index to the state names.

In [50]:

```
path = r'/Users/maanyatandon/Documents/fall2020/econometrics/Lab3/'
df_srm = pd.read_excel(path+'statesRegionsMapping.xlsx', sheet_name = 'data' )
df_srm.set_index('State', inplace = True);
```

Print the first five (5) records.

In [51]:

```
df_srm.head()
```

Out[51]:

State	Region	Population
Alabama	South	4779736
Alaska	West	710231
Arizona	West	6392017
Arkansas	South	2915918
California	West	37253956

Merge the physician/nurse data and the state data as described in Lesson #3.

In [52]:

```
df_all = pd.merge(df_srm, df_phy, how = 'inner', left_index = True, right_index=True)
df_all.head()
#df_all.dtypes
```

Out[52]:

State	Region	Population	Physicians	Nurses
Alabama	South	4779736	10265	42880
Alaska	West	710231	1574	5010
Arizona	West	6392017	14051	38570
Arkansas	South	2915918	5902	23050
California	West	37253956	100131	233030

Recall your answer to the question above regarding what you should do to the physician and nurse data before you do any analytical work. Make the correction here. Be sure to use the corrected data for the following tasks.

In [53]:

```
df_all['Population (mil)'] = df_all['Population'] / 1000000.0
df_all['Physicians (per mil)'] = df_all['Physicians'] / df_all['Population (mil)']
df_all['Nurses (per mil)'] = df_all['Nurses'] / df_all['Population (mil)']
pd.set_option('display.float_format', lambda x: '%.2f' % x)
df_all.head()
```

Out[53]:

State	Region	Population	Physicians	Nurses	Population (mil)	Physicians (per mil)	Nurses (per mil)
Alabama	South	4779736	10265	42880	4.78	2147.61	8971.21
Alaska	West	710231	1574	5010	0.71	2216.18	7054.04
Arizona	West	6392017	14051	38570	6.39	2198.21	6034.09
Arkansas	South	2915918	5902	23050	2.92	2024.06	7904.89
California	West	37253956	100131	233030	37.25	2687.80	6255.17

Create summary statistics for the physician and nurse data.

In [54]:

```
## Create summary statistics for the physician and nurse data.
#df_all.style.format("{:,.0f}")
df_all.describe()
```

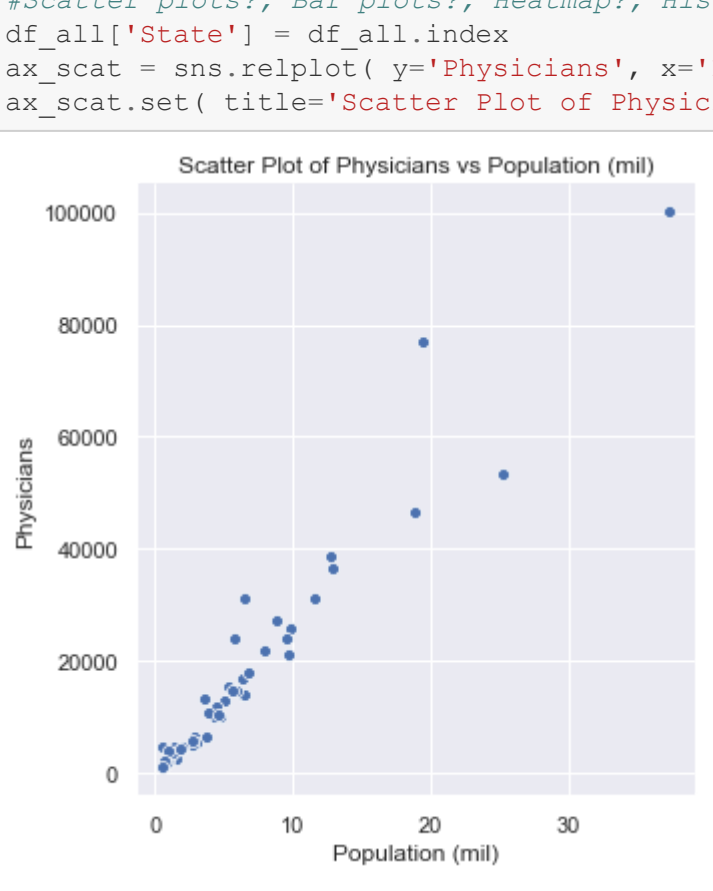
Out[54]:

	Population	Physicians	Nurses	Population (mil)	Physicians (per mil)	Nurses (per mil)
count	51.00	51.00	51.00	51.00	51.00	51.00
mean	6053834.08	16440.25	50662.16	6.05	2721.31	8931.81
std	6823984.27	19257.69	49553.46	6.82	1013.43	1842.39
min	563626.00	1020.00	4700.00	0.56	1689.86	5961.75
25%	1696961.50	4655.50	13870.00	1.70	2196.84	7963.07
50%	4339367.00	10403.00	38020.00	4.34	2480.94	8940.77
75%	6636084.50	21600.00	62055.00	6.64	2826.86	10053.87
max	37253956.00	100131.00	233030.00	37.25	8143.28	14774.24

Plot the physician and nurse data using graphs you learned in Stat 101.

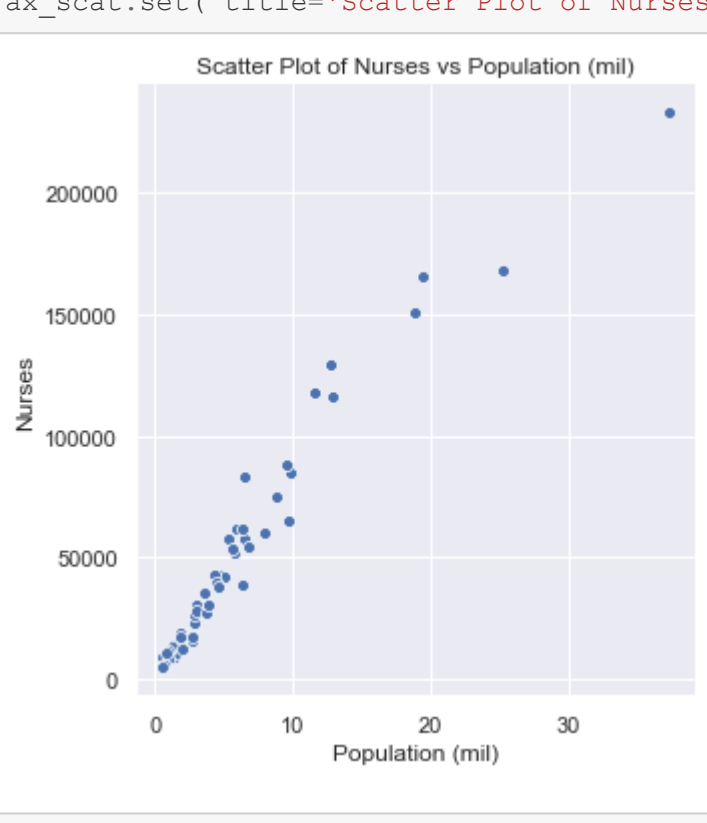
In [55]:

```
##
## Enter code here.  You can insert more code cells for more graphs below
## this one by clicking on the thick edge to the far left and hitting the b
## key or by just clicking on the + sign on the menu bar.  Create as many graphs as you need.
##
#Scatter plots?, Bar plots?, Heatmap?, Histogram?
df_all['State'] = df_all.index
ax_scat = sns.relplot( y='Physicians', x='Population (mil)',data=df_all)
ax_scat.set( title='Scatter Plot of Physicians vs Population (mil)');
```



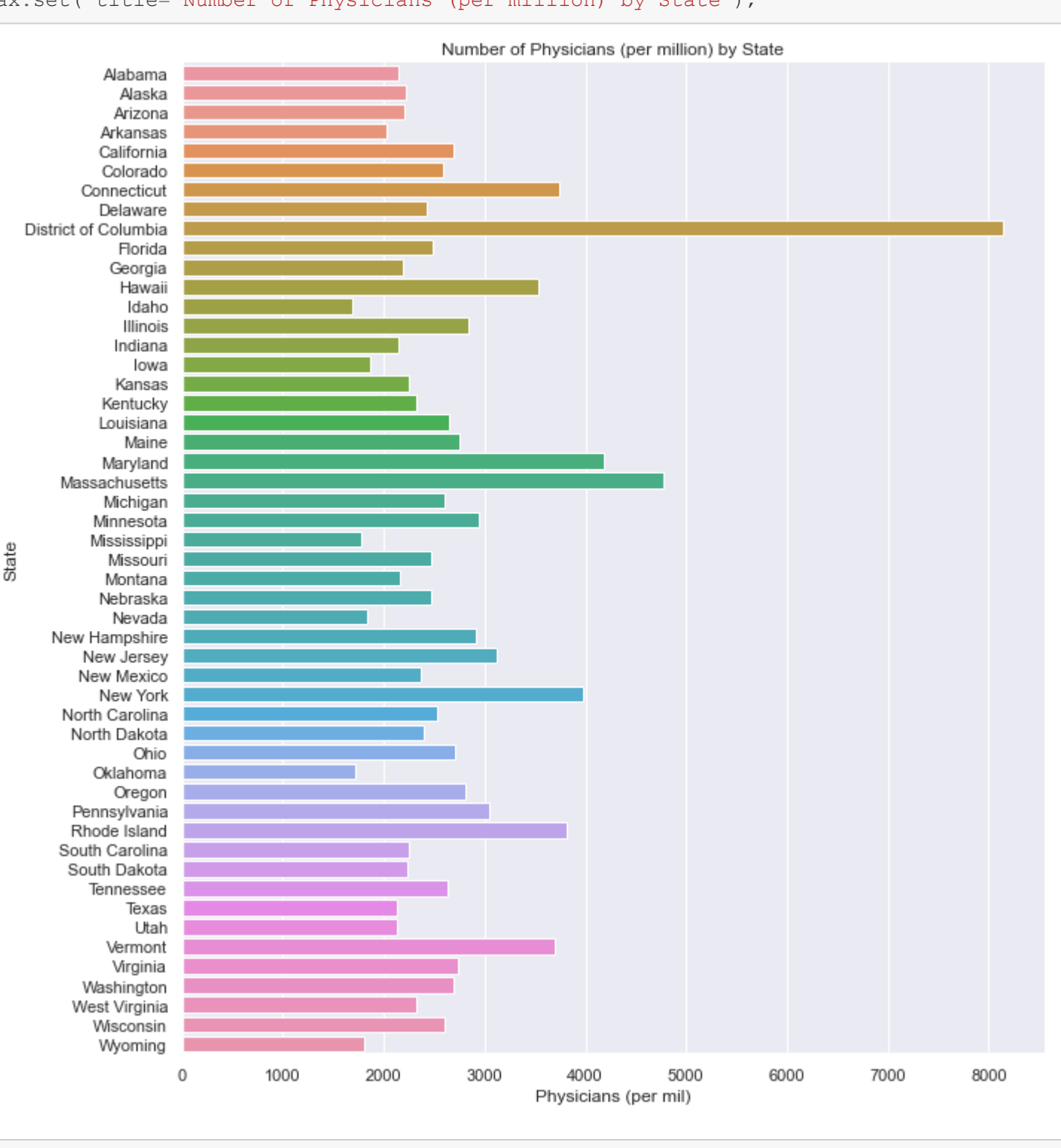
In [56]:

```
#Population vs nurses scatter plot
ax_scat = sns.relplot( y='Nurses', x='Population (mil)',data=df_all)
ax_scat.set( title='Scatter Plot of Nurses vs Population (mil)');
```



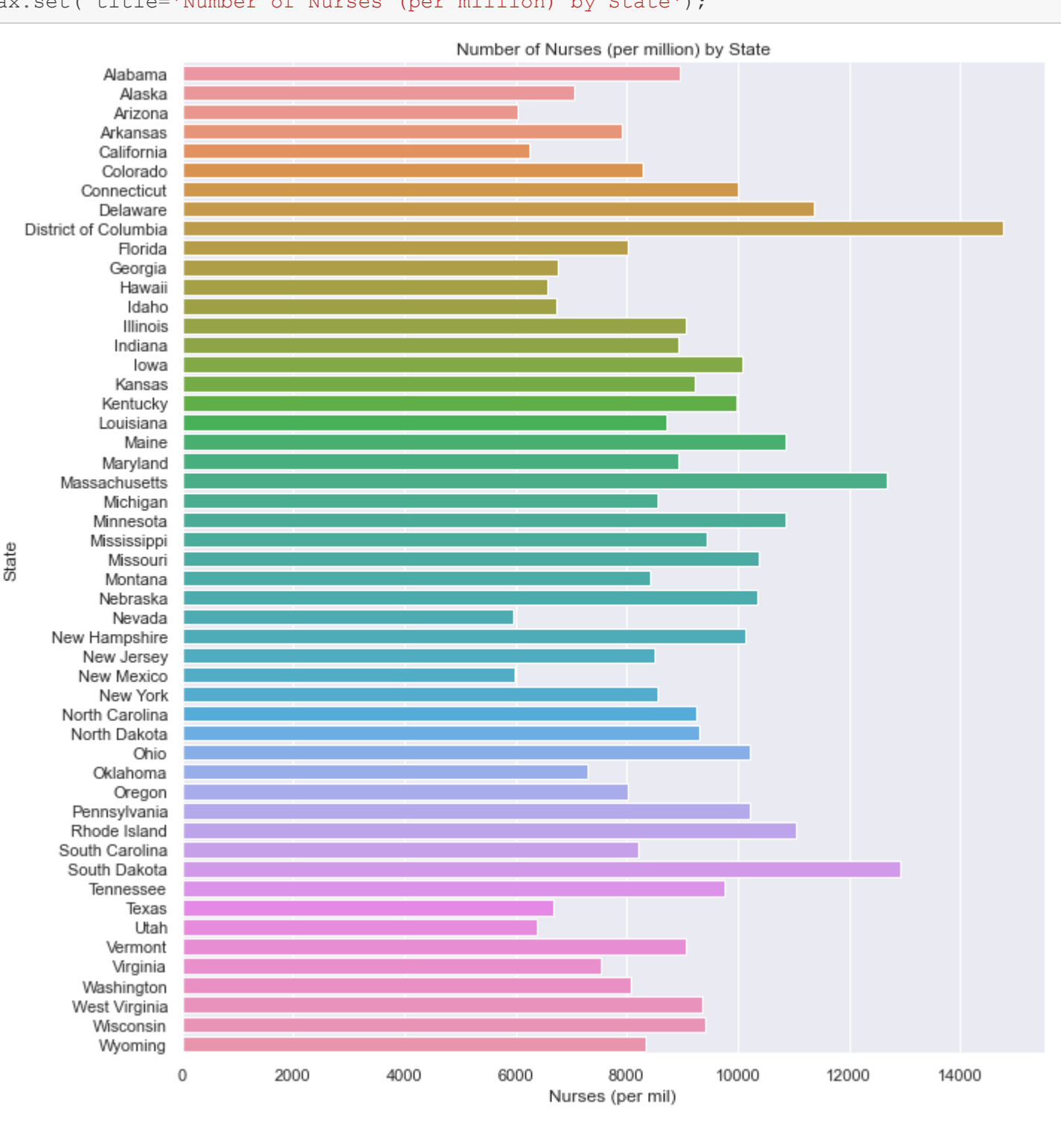
In [57]:

```
ax = sns.catplot(y='State', x='Physicians (per mil)', kind='bar', data=df_all, height=10, aspect=1);
ax.set( title='Number of Physicians (per million) by State');
```



In [58]:

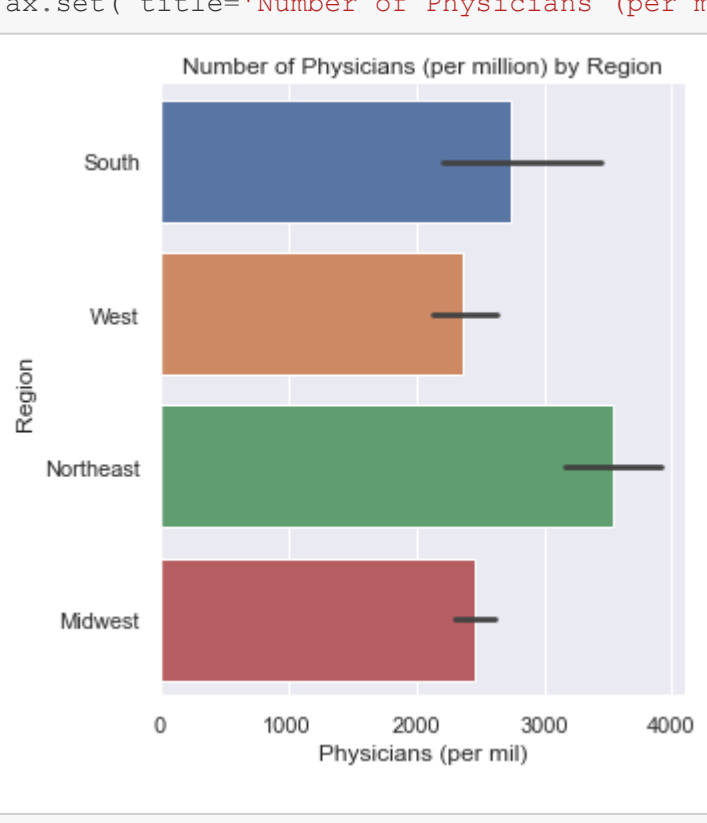
```
ax = sns.catplot(y='State', x='Nurses (per mil)', kind='bar', data=df_all, height=10, aspect=1);
ax.set( title='Number of Nurses (per million) by State');
```



Plot the physicians by region. What graph type would you use?

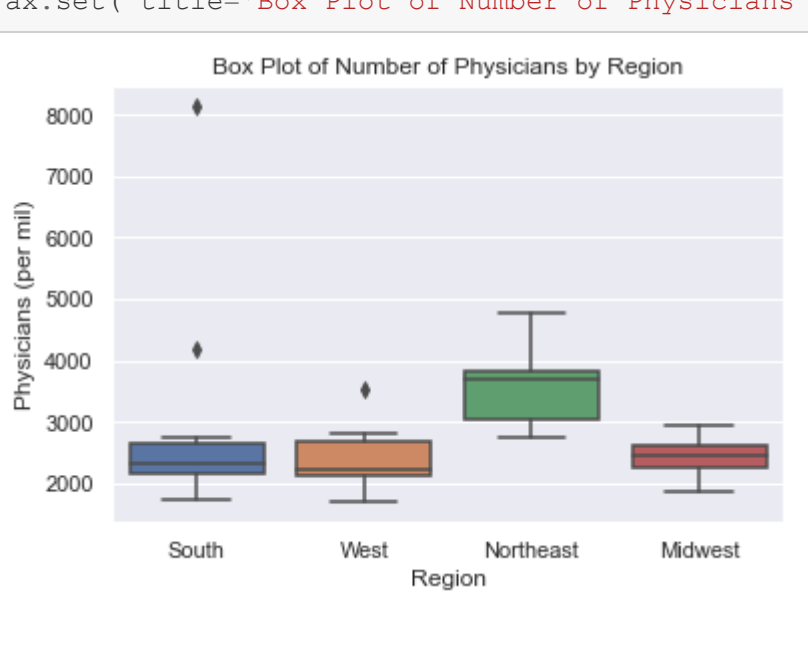
In [59]:

```
## Group By Region
df_region = df_all[['Region', 'Physicians', 'Nurses', 'Population (mil)']].groupby(['Region']).sum()
## df_region.columns
df_region['Physicians (per mil)'] = df_region['Physicians'] / df_region['Population (mil)']
df_region['Nurses (per mil)']= df_region['Nurses']/ df_region['Population (mil)']
#df_region.head()
ax = sns.catplot(y='Region', x='Physicians (per mil)', kind='bar', data=df_all, height=5, aspect=1);
ax.set( title='Number of Physicians (per million) by Region');
```



In [60]:

```
ax = sns.boxplot( x='Region', y='Physicians (per mil)', data=df_all)
ax.set( title='Box Plot of Number of Physicians by Region');
```



Test the physicians by region. What do you conclude?

In [61]:

```
## Group By Region
#df_region = df_all[['Region', 'Physicians', 'Nurses', 'Population (mil)']].groupby(['Region']).agg(['s
um'])
df_region = df_all[['Region', 'Physicians', 'Nurses', 'Population (mil)']].groupby(['Region']).sum()
## df_region.columns
df_region['Physicians (per mil)'] = df_region['Physicians'] / df_region['Population (mil)']
df_region['Nurses (per mil)']= df_region['Nurses']/ df_region['Population (mil)']
df_region
```

Out[61]:

Region	Physicians	Nurses	Population (mil)	Physicians (per mil)	Nurses (per mil)
Midwest	172781	642700	66.93	2581.63	9603.00
Northeast	201597	534170	55.32	3644.38	9656.48
South	280075	924930	114.56	2444.88	8074.06
West	184000	481970	71.95	2557.49	6699.09

In [62]:

```
#test physicians by region
```

Print and graph a correlation matrix.

In [63]:

```
correl = df_all.corr()
correl
```

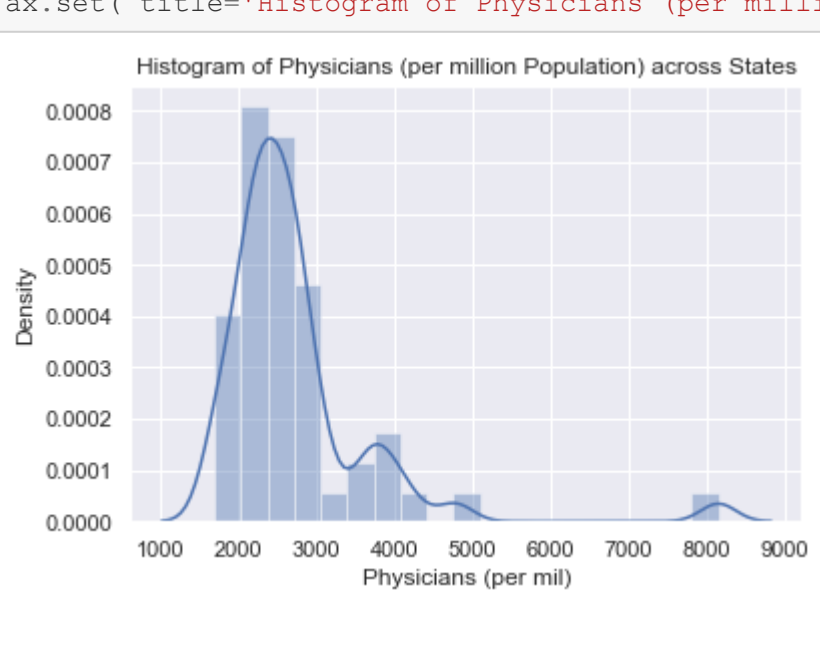
Out[63]:

	Population	Physicians	Nurses	Population (mil)	Physicians (per mil)	Nurses (per mil)
Population	1.00	0.97	0.97	1.00	-0.01	-0.28
Physicians	0.97	1.00	0.97	0.97	0.13	-0.18
Nurses	0.97	0.97	1.00	0.97	0.05	-0.15
Population (mil)	1.00	0.97	0.97	1.00	-0.01	-0.28
Physicians (per mil)	-0.01	0.13	0.05	-0.01	1.00	0.56
Nurses (per mil)	-0.28	-0.18	-0.15	-0.28	0.56	1.00

Plot a histogram of physicians.

In [64]:

```
ax = sns.distplot( df_all['Physicians (per mil)'], hist=True, kde=True )
# ax.set_title('Histogram of Physicians (per million Population) across States');
ax.set( title='Histogram of Physicians (per million Population) across States', ylabel= 'Density');
```



Perform a normality test of the physicians distribution.

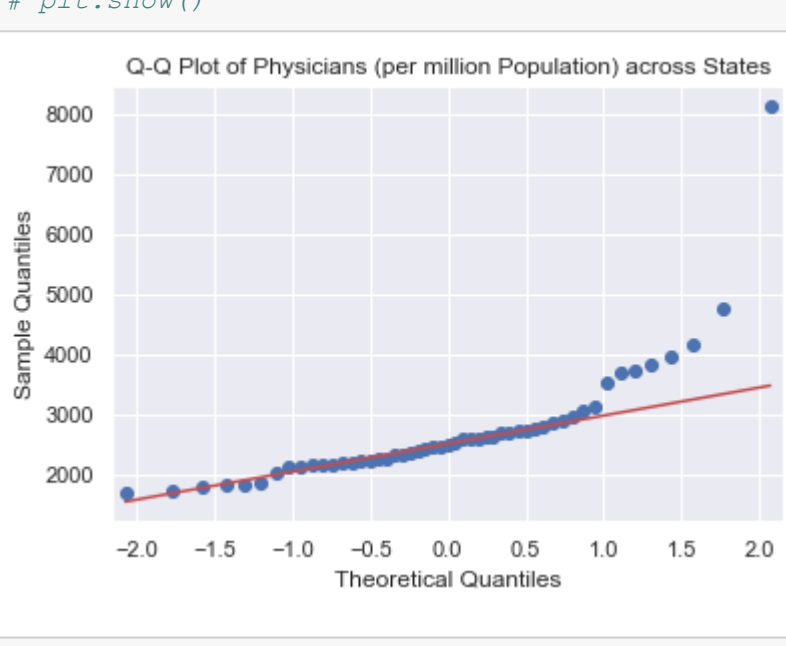
In [65]:

```
phy_mean = df_all['Physicians (per mil)'].mean()
phy_stdev=df_all['Physicians (per mil)'].std()
phy_skew=df_all['Physicians (per mil)'].skew()
phy_kurt=df_all['Physicians (per mil)'].kurtosis()
phy_count=df_all['Physicians (per mil)'].count()
print('phy_mean ' + repr(phy_mean) + ', phy_stdev', repr(phy_stdev), ', phy_count', repr(phy_count))
## A truly symmetrical data set has a skewness equal to 0.
## The rule of thumb is:
##    If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
##    If the skewness is between -1 and - 0.5 or between 0.5 and 1, the data are moderately skewed.
##    If the skewness is less than -1 or greater than 1, the data are highly skewed.
## The kurtosis parameter is a measure of the combined weight of the tails relative to the rest of the
 distribution.
##    If the kurtosis is close to 0, then a normal distribution is often assumed.
##    If the kurtosis < 0, then the distribution has light tails.
##    If the kurtosis > 0, then the distribution has heavier tails.
print('phy_skewness'+ repr(phy_skew) + ', phy_kurtosis', repr(phy_kurt))
```

```
phy_mean 2721.3076567881612, phy_stdev 1013.4297190820889 , phy_count 51
phy_skewness3.3945503075632852, phy_kurtosis 16.027127996805078
```

In [66]:

```
# q-q plot: A test of normal distribution
# Unknown library: from seaborn_qqplot import qqplot
from statsmodels.graphics.gofplots import qqplot
ax = qqplot( df_all['Physicians (per mil)'], line='q')
#ax.set( title='Q-Q Plot of Physicians (per million Population) across States');
h = plt.title('Q-Q Plot of Physicians (per million Population) across States')
# plt.show()
```



In [67]:

```
# normality test
# The D'Agostino's K^2 test is available via the normaltest() SciPy function and returns the test stati
stic and the p-value.
# https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/
from scipy.stats import normaltest
stat, p = normaltest( df_all['Physicians (per mil)'] )
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
    print('Sample looks Gaussian (fail to reject H0)')
else:
    print('Sample does not look Gaussian (reject H0)')
```

```
Statistics=64.396, p=0.000
Sample does not look Gaussian (reject H0)
```

Post-lab Questions

Back to Contents

Interpret the summary statistics.

The Northeast region has the highest number of physicians per capita with 3,644 physicians per a million people. The South region has the lowest number of physicians per capita with 2,449 physicians per million population. The Northeast has almost 50% more physicians per capita compared to South. From greatest to least number of physicians per capita, the Notheast is followed by Midwest, West, and finally South. Also, we see that the number of physicians (per mil) is highly variable in South with a standard deviation of 1491, as compared to other regions which have standard deviation in the range 306-641. This suggests that the reach of healthcare is not only low, but also higly uneven in the South. In South, Oklahoma has lowest with 1724 while DC has the highest (8143) number of physicians per million of population.

Interpret the graphs. What do they tell you about the distribution of physicians around the country?

From the Scatter Plots of Physicians and Nurses versus Population for States, we see that they are almost linearly related. Hence we can use Physicians and Nurses per million of population for further analysis.

From Box Plot of Number of Physicians (per mil) by Region - we see that the Northeast region has highest number of Physicians per Capita and highest variability, while South has the lowest Physicians per Capita and low variability. South has two outlier states with exceptionally high number of Physicians (per mil) compared to the rest of South - these being District of Columbia and Maryland.

Also, Northeast region has highest Nurses per Capita, while West region has the lowest number of Nurses per Capita.

From the Bar Chart for the Number of Physicians (per mil), District of Columbia has highest number of Physicians followed by Massachusetts.

From the histogram of Physicians (per million Population) above, the distribution does not look normally distributed - since it is not symmetric, and has a longer and thicker right tail, compared to the left tail.

What are the Null and Alternative Hypotheses for the normality test of physicians? Explain your answer.

The distribution of Physicians (per million Population) has a skewness of 3.3945, while symmetric distributions like Normal distribution have skewness of zero. This indicates a positively skewed tail (non-symmetric), as visually observed - instead of Normal distribution
The distibution of Physicians (per million Population) has a kurtosis of 16.0271. This indicates a heavier tails compared to the peak, compared to the Normal distribution which has zero kurtosis.
However, the sample size of 51 may be too small to conclusively establish non-normal disrtibution based on skewness and kurtosis alone. Hence to do a statistical test of normality via D\'Agostino\'s K^2 test which is available via the normaltest() SciPy function. We test the following hypothesis with $\alpha = 0.05$: Null Hypothesis H_0: Population (per mil) is normally distributed
Alternative Hypothesis H_a: Population (per mil) is not normally distributed

Are physicians normally distributed? That is, do you reject or fail to reject your Null Hypothesis? Explain you answer.

The q-q plot above plots the actual qunatiles in which the observations for 'Physicians (per mil)' fall versus the theoretical quantiles for Normal distribution. We see that the the sample quantiles do not fall on the theoretical line (red line) for it to have normal distribution. Also from the normality test of the null hypothesis H_0 that Physicians (per million Population) is normally distributed with $\alpha = 0.05$, we get a p value of 0.000. Hence we reject the hypothesis that Physicians per capita is Normally distributed.

What can you observe about the correlation matrix? Explain.

Both Physicians and Nurses have a correlation of 0.97 with the Population - as expected - indicating that states with larger population have more physicians and nurses compared to states with smaller population.
The correlation between the population of a state and the Physicians (per mil) is -0.01 (close to zero). This indicates that the number of physicians per capita are not related to population of the state.
However, the correlation between the population of a state and the Nurses (per mil) is -0.28 (reasonably negative). This indicates that in general there are less nurses per capita for states with a larger population.
Also, the correlation between Nurses (per mil) and Physicians (per mil) is 0.56, which indicates that this relationship is not as strong as the relationship between the absolute number of Physicians and Nurses which has a correlation of 0.97. So as the number of Physicians increase with the size of population of a state, the number of nurses do not increase as much proportionately.

Well done!

Make sure your name is on this notebook at the top and on the file.
Please submit this notebook as a PDF file. Nothing else will be accepted.

[ ]: