In [79]:
```
#GOALS
#1. Load the python libraries
#2. Create dummy variables and account for missing data
#3. Describe your data
#4. Which of our variables are potentially collinear?
#5. Create a exploratory analysis plan of your data
#6. What is your hypothesis?
#7. Bonus: Test your hypothesis with Logistic Regression
```

In [80]:
```
import pandas as pd
import numpy as np
```

In [81]:
```
SHD = pd.read_csv('Sexual_Health_Discussions 7_22_18.csv')
```

In [82]:
```
SHD.head()
```

Out[82]:

	User Id	%PatsDiscSexHlth	%PatsDiscSTDsSTIs	ComfDiscSxHlthPatsAdol	ComfDiscSxHlthPatsAdlts
0	1	50	50	7.0	7
1	2	100	60	7.0	7
2	3	60	50	1.0	7
3	4	60	60	7.0	7
4	5	20	20	6.0	6

In [105]:
```
#peeking at null values, future analysis may not work if there are nulls
SHD.isna().any()
```

Out[105]:
```
User Id                    False
%PatsDiscSexHlth           False
%PatsDiscSTDsSTIs          False
ComfDiscSxHlthPatsAdol      True
ComfDiscSxHlthPatsAdlts    False
ComfDiscSxHlthPatsSnrs      True
Years in Practice          False
Age                        False
Practice Setting           False
Patient Volume             False
Gender                     False
Female                     False
Male                       False
Prefer not to answer       False
dtype: bool
```

In [84]:
```
#Gender will be the dummy variable
dummy = pd.get_dummies(SHD['Gender'])
```

In [85]:
```
dummy.head()
```

Out[85]:

	Female	Male	Prefer not to answer
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

In [86]:
```
#adding into larger dataset
SHD = pd.concat([SHD, dummy], axis=1)
```

In [87]:
```
#woo! it worked
SHD.head(10)
```

Out[87]:

	User Id	%PatsDiscSexHlth	%PatsDiscSTDsSTIs	ComfDiscSxHlthPatsAdol	ComfDiscSxHlthPatsAdlts
0	1	50	50	7.0	7
1	2	100	60	7.0	7
2	3	60	50	1.0	7
3	4	60	60	7.0	7
4	5	20	20	6.0	6
5	6	15	1	7.0	7
6	7	10	30	7.0	7
7	8	0	0	4.0	6
8	9	100	100	6.0	6
9	10	5	50	NaN	7

In [88]:
```
#For my exploratory analysis, I will get to know the data by describing it,
#looking for null values, and determining any key correlations. Once the co
#try to find an interesting or relevant aspect for hypothesis and visualizat
SHD.describe()
```

Out[88]:

	User Id	%PatsDiscSexHlth	%PatsDiscSTDsSTIs	ComfDiscSxHlthPatsAdol	ComfDiscSxHlt
count	129.000000	129.000000	129.000000	110.000000	
mean	65.000000	44.310078	39.519380	5.290909	
std	37.383151	35.752141	32.325276	2.019835	
min	1.000000	0.000000	0.000000	1.000000	
25%	33.000000	10.000000	10.000000	4.000000	
50%	65.000000	50.000000	30.000000	6.000000	
75%	97.000000	80.000000	60.000000	7.000000	
max	129.000000	100.000000	100.000000	7.000000	

In [89]:
```
SHD.dtypes
```

Out[89]:
```
User Id                       int64
%PatsDiscSexHlth              int64
%PatsDiscSTDsSTIs             int64
ComfDiscSxHlthPatsAdol      float64
ComfDiscSxHlthPatsAdlts       int64
ComfDiscSxHlthPatsSnrs      float64
Years in Practice             int64
Age                           int64
Practice Setting             object
Patient Volume                int64
Gender                       object
Female                        uint8
Male                          uint8
Prefer not to answer          uint8
dtype: object
```

In [90]:
```
SHD.isna().any()
```

Out[90]:
```
User Id                     False
%PatsDiscSexHlth            False
%PatsDiscSTDsSTIs           False
ComfDiscSxHlthPatsAdol       True
ComfDiscSxHlthPatsAdlts     False
ComfDiscSxHlthPatsSnrs       True
Years in Practice           False
Age                         False
Practice Setting            False
Patient Volume              False
Gender                      False
Female                      False
Male                        False
Prefer not to answer        False
dtype: bool
```

In [91]:
```
SHD.ComfDiscSxHlthPatsAdol.mean()
```

Out[91]:
```
5.290909090909091
```

In [92]:
```
SHD.ComfDiscSxHlthPatsAdlts.mean()
```

Out[92]:
```
6.232558139534884
```

In [93]:
```
SHD.ComfDiscSxHlthPatsSnrs.mean()
```

Out[93]:
```
5.621848739495798
```

In [94]:
```
#based on the above data, on average physicians are most comfortable speakin
#about general sexual health when compared to their comfort level speaking t
import matplotlib.pyplot as plt
```

In [95]:
```
#Looking at this data, I would anticipate that colinear variables would be
#the comfort level discussing sexual health and the % of patients which whon
#physicians discuss sexual health at annual appointments (across all ages),
#age of physician and comfort discussing sexual health with seniors,
#and an inverse relationship between patient volume and % of patients
#with which physician discuss both sexual health and STDs/STIs. I also might
#expect that some of the overlapping variables will correlate, like the comf
#physicians in discussing sexual health with adults, and the comfort level c
#in discussing sexual health with seniors.
```

In [96]:
```
SHD.corr()
```

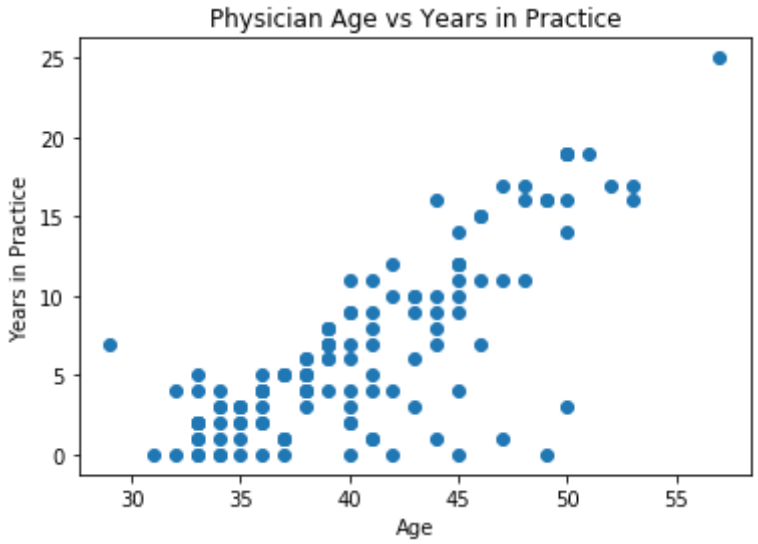
Out[96]:

	User Id	%PatsDiscSexHlth	%PatsDiscSTDsSTIs	ComfDiscSxHlthPatsAd
User Id	1.000000	-0.049124	-0.006898	-0.00340
%PatsDiscSexHlth	-0.049124	1.000000	0.696368	0.46598
%PatsDiscSTDsSTIs	-0.006898	0.696368	1.000000	0.33816
ComfDiscSxHlthPatsAdol	-0.003408	0.465985	0.338160	1.00000
ComfDiscSxHlthPatsAdlts	-0.001060	0.346209	0.419026	0.38343
ComfDiscSxHlthPatsSnrs	-0.120030	0.303725	0.369548	0.26653
Years in Practice	-0.950208	0.059436	0.015474	0.06756
Age	-0.687598	0.135773	0.113364	0.19188
Patient Volume	0.098479	0.135757	0.092043	0.08031

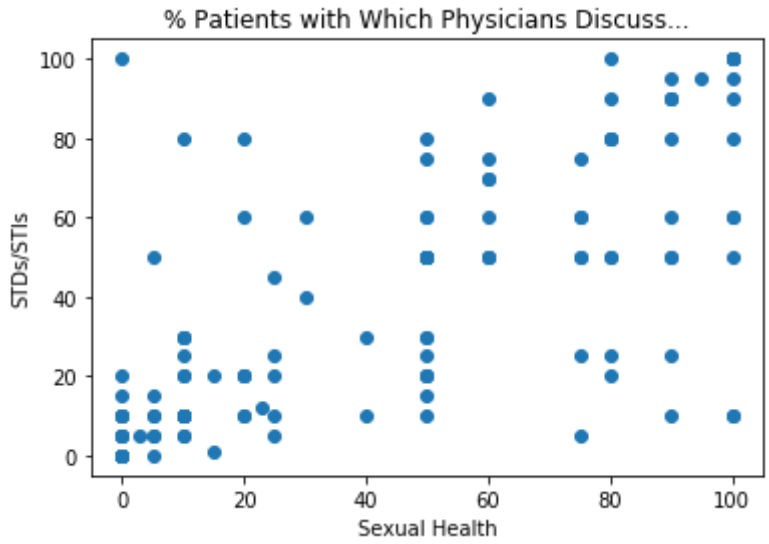
In [97]:
```
#I was very surprised that patient volume and physician age hardly
#correlated with any of the sexual health variables!
#I was also surprised at the lack of consistency in correlation between
#adolescent, adult, and senior variables. If one of those age groups
#correlates strongly, it is not particularly likely that the others will
#as well.

#Happily, *per my hypothesis*, there is a strong correlation between physici
#are comfortable discussing sexual health with adults and those
#who are comfortable discussing those topics with seniors (.66).
#There are several other strong correlating variables of interest.
#The % of patients with which physicians discuss sexual health at their
#annual appointments is correlated with the % of patients with which
#physicians discuss STDs/STIs (.70) at those appointments.
#The % of patients with which physicians
#discuss sexual health at their annual appointments is also correlated
#with physician comfort levels in discussing sexual health with
#adolescent patients (.47). We do not see equally strong correlations with
#physician comfort levels in speaking to adults (.35) or seniors (.30).
#Lastly, physicians who are more comfortable discussing sexual health
#with adults, also seem somewhat more likely to discuss STDs/STIs
#with a higher percentage of patients at their annual appointments (.42).
#Physician age and years in practice were colinear as well, but that's borir
```

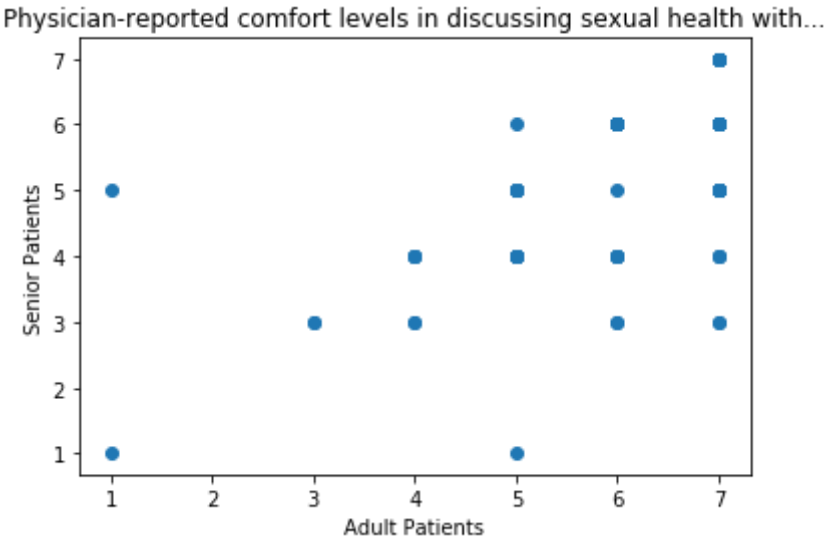
In [98]:
```
#plotting practice
plt.scatter(SHD[['Age']], SHD[['Years in Practice']])
plt.xlabel('Age')
plt.ylabel('Years in Practice')
plt.title('Physician Age vs Years in Practice')
plt.show()
```



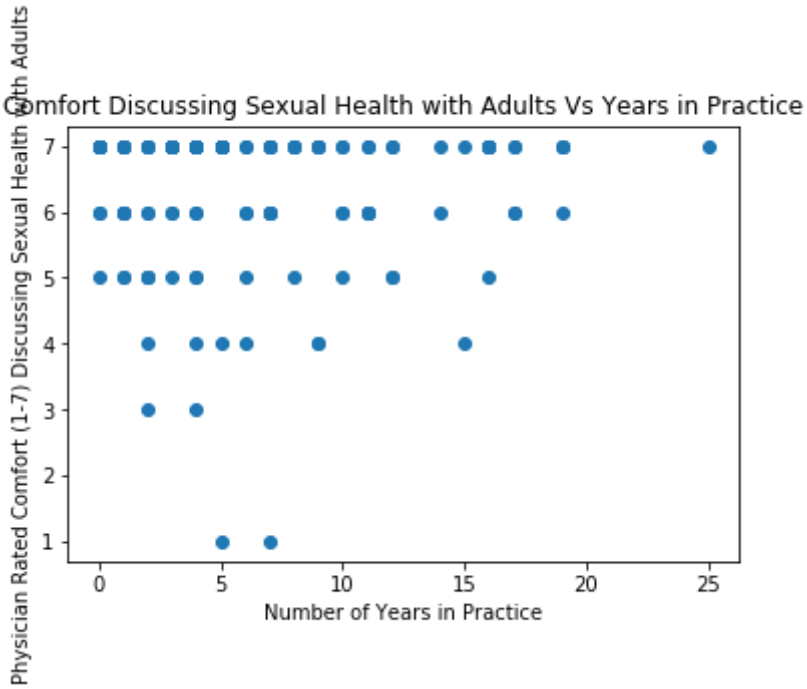
In [99]:
```
#taking a look at another correlation
plt.scatter(SHD[['%PatsDiscSexHlth']], SHD[['%PatsDiscSTDsSTIs']])
plt.xlabel('Sexual Health')
plt.ylabel('STDs/STIs')
plt.title('% Patients with Which Physicians Discuss...')
plt.show()
```



In [100]:
```
#visualizing hypothesis data
plt.scatter(SHD[['ComfDiscSxHlthPatsAdlts']], SHD[['ComfDiscSxHlthPatsSnrs']
plt.xlabel('Adult Patients')
plt.ylabel('Senior Patients')
plt.title('Physician-reported comfort levels in discussing sexual health wit
plt.show()
```



In [101]:
```
#I tried to run this quite a few ways, but did not have success. I also look
#time to investigate two variables with low correlation for best fit
plt.scatter(SHD[['Years in Practice']], SHD[['ComfDiscSxHlthPatsAdlts']])
plt.xlabel('Number of Years in Practice')
plt.ylabel('Physician Rated Comfort (1-7) Discussing Sexual Health with Adul
plt.title('Comfort Discussing Sexual Health with Adults Vs Years in Practice
plt.show()
```



In [106]:
```
#through scatterplot and histogram, data appears limited, but I think there
#overlap as this question only allowed for a 7 point rating response. Note,
#without null values (pretty sure it wont work if there are nulls)
```

In [103]:
```
#to perform linear regression and test the model
#sklearn import linear model
from sklearn import linear_model
```

In [104]:
```
lm = linear_model.LinearRegression()
lm.fit(SHD[['Years in Practice']], SHD[['ComfDiscSxHlthPatsAdlts']])
```

Out[104]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=Fal
se)
```

In [107]:
```
lm.intercept_
```

Out[107]:
```
array([6.15544585])
```

In [108]:
```
lm.coef_
```

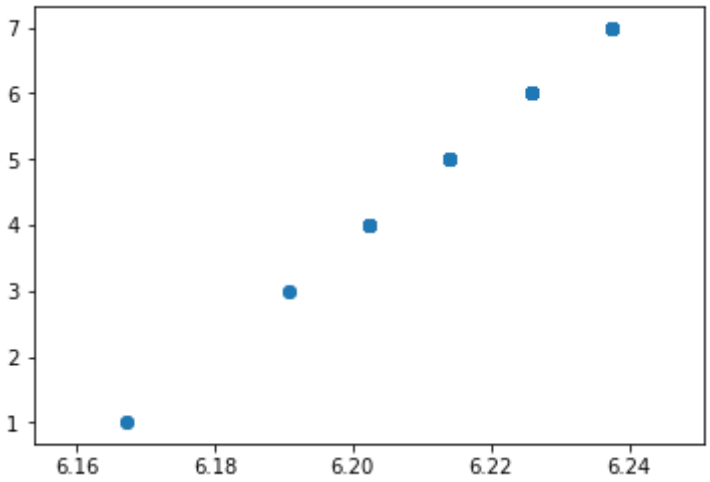
Out[108]:
```
array([[0.01173053]])
```

In [109]:
```
y_pred = lm.predict(SHD[['ComfDiscSxHlthPatsAdlts']])
```

In [110]:
```
y_pred
```

Out[110]:
```
array([[6.23755953],
       [6.23755953],
       [6.23755953],
       [6.23755953],
       [6.225829  ],
       [6.23755953],
       [6.23755953],
       [6.225829  ],
       [6.225829  ],
       [6.23755953],
       [6.23755953],
       [6.23755953],
       [6.23755953],
       [6.23755953],
       [6.21409848],
       [6.23755953],
       [6.20236795],
       [6.23755953],
       [6.225829  ],
```

In [111]:
```
# truth vs predicted
# scatterplot of Physician-reported comfort levels in discussing sexual heal
plt.scatter(y_pred, SHD['ComfDiscSxHlthPatsAdlts'])
plt.show()
```

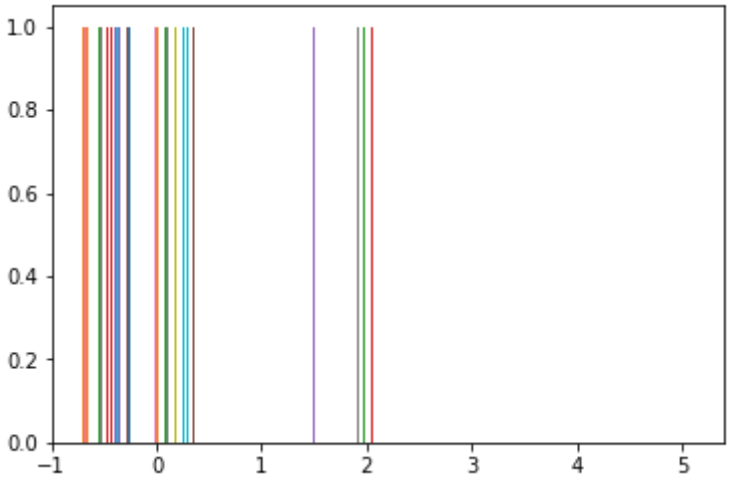


In [113]:
```
#gorgeous!
#if your Years in Practice are 10, then your predicted value for
#Comfort discussing sexual health with adult patients (on the 1-7 scale) =
lm.predict(10)
```

Out[113]:
```
array([[6.2727511]])
```

In [114]:
```
resid = y_pred - SHD[['ComfDiscSxHlthPatsAdlts']]
```

In [119]:
```
#checking for errors
plt.hist(resid)
plt.show()
```



In [ ]:
```
#this looks V weird..I cannot get the histograms to come out correctly!
```

In [121]:
```
lm.score(SHD[['Years in Practice']], SHD['ComfDiscSxHlthPatsAdlts'])
```

Out[121]:
```
0.003000442305399642
```

In [122]:
```
#only 3% of that comfort rating is explained by physician experience (aka Ye
```

In [124]:
```
from sklearn import metrics
metrics.mean_squared_error(SHD['ComfDiscSxHlthPatsAdlts'], y_pred)
```

Out[124]:
```
1.3554254397829866
```

In [ ]:
```
#this is pretty low..at least compared to our class data. so that is pretty
```