First Name:

Last Name:

In [1]:

```
#import pandas & numpy
import pandas as pd
import numpy as np
import scipy.stats #I usually keep scipy as scipy because you will need to access it libra
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:

```
#read in csv file into
nesarc = pd.read_csv('nesarc.csv', low_memory=False) #increase efficiency
pd.set_option('display.float_format', lambda x:'%f'%x)
```

In [3]:

```
#setting variables you will be working with to numeric
nesarc['S2AQ5B'] = pd.to_numeric(nesarc['S2AQ5B'], errors='coerce') #convert variable to nu
nesarc['S2AQ5D'] = pd.to_numeric(nesarc['S2AQ5D'], errors='coerce') #convert variable to nu
nesarc['S2AQ5A'] = pd.to_numeric(nesarc['S2AQ5A'], errors='coerce') #convert variable to nu
nesarc['S2BQ1B1'] = pd.to_numeric(nesarc['S2BQ1B1'], errors='coerce') #convert variable to
nesarc['AGE'] = pd.to_numeric(nesarc['AGE'], errors='coerce') #convert variable to numeric
```

In [4]:

```
#subset data to adults age 26 to 50 who have consumed beer in the past 12 months
sub1=nesarc[(nesarc['AGE']>=26) & (nesarc['AGE']<=50) & (nesarc['S2AQ5A']==1)]
```

In [5]:

```
sub2=sub1.copy()
```

In [6]:

```
#SETTING MISSING DATA
sub2['S2AQ5D']=sub2['S2AQ5D'].replace(99, np.nan)

sub2['S2AQ5B']=sub2['S2AQ5B'].replace(8, np.nan)
sub2['S2AQ5B']=sub2['S2AQ5B'].replace(9, np.nan)
sub2['S2AQ5B']=sub2['S2AQ5B'].replace(10, np.nan)
sub2['S2AQ5B']=sub2['S2AQ5B'].replace(99, np.nan)

sub2['S2BQ1B1']=sub2['S2BQ1B1'].replace(9, np.nan)
```

In [7]:

```
#recoding number of days consumed beer in the past month
recode2 = {1:30, 2:26, 3:14, 4:8, 5:4, 6:2.5, 7:1}
sub2['BEER_FEQMO']= sub2['S2AQ5B'].map(recode2)

recode3 = {2:0, 1:1}
sub2['S2BQ1B1']= sub2['S2BQ1B1'].map(recode3)
```

contingency table of observed counts - between beer dependence (S2BQ1B1) and beer drinking frequency (BEER_FEQMO)

Use sub2

In [8]:

```
ct1=pd.crosstab(sub2['S2BQ1B1'], sub2['BEER_FEQMO'])
print (ct1)
```

```
BEER_FEQMO  1.000000   2.500000   4.000000   8.000000   14.000000  26.000000
  \
S2BQ1B1

0.000000         1172       1477       1390       1189        842        313

1.000000           40         80         82        114         78         51


BEER_FEQMO  30.000000
S2BQ1B1
0.000000          343
1.000000           65
```

Contingency table that shows relation between beer dependency and beer drinking frequency. It is noticeable that as interviewee drinks more beer, the higher percentage of them grow beer dependency.

contingency table of observed percentages - between beer dependence (S2BQ1B1) and beer drinking frequency (BEER_FEQMO)

Use ct1 calculated in the above cell

In [9]:

```
colsum=ct1.sum(axis=0)
colpct=ct1/colsum
print(colpct)
```

```
BEER_FEQMO  1.000000   2.500000   4.000000   8.000000   14.000000  26.000000
  \
S2BQ1B1

0.000000     0.966997   0.948619   0.944293   0.912510   0.915217   0.859890

1.000000     0.033003   0.051381   0.055707   0.087490   0.084783   0.140110


BEER_FEQMO  30.000000
S2BQ1B1
0.000000     0.840686
1.000000     0.159314
```

Table that uses percentage to describe relation between beer drinking frequency and beer dependency.

chi-square analysis between beer dependence (S2BQ1B1) and beer drinking frequency (BEER_FEQMO)

Use ct1

Apply chi-square analysis to test hypothesis between beer dependency and beer drinking frequency. The p value is 2.06 e-24, and expected counts is 6

In [10]:

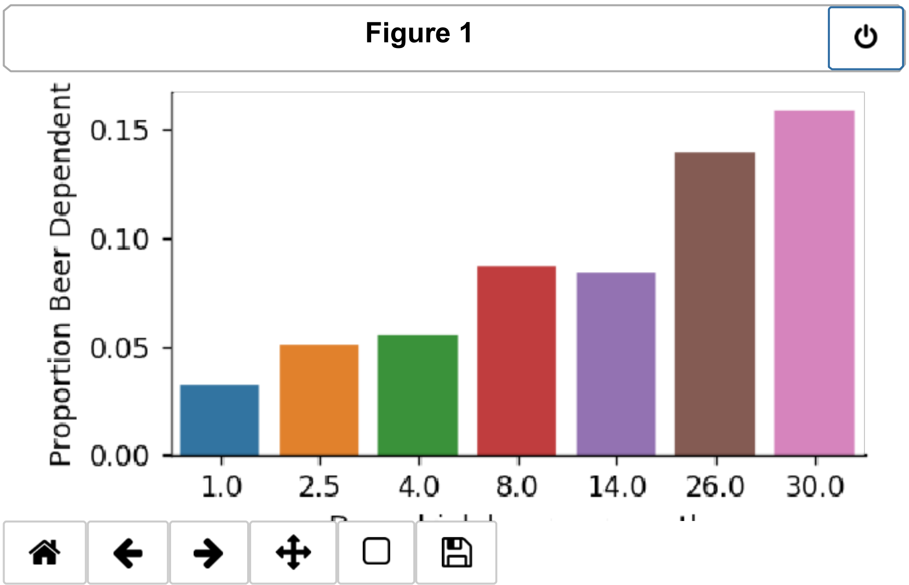
```
print ('chi-square value, p value, expected counts')
cs1= scipy.stats.chi2_contingency(ct1)
print (cs1)
```

```
chi-square value, p value, expected counts
(124.26789738394885, 2.0662068579068001e-24, 6, array([[ 1126.57711443,  144
7.26119403,  1368.25207297,  1211.16334992,
          855.15754561,   338.34494196,   379.24378109],
       [   85.42288557,   109.73880597,   103.74792703,    91.83665008,
           64.84245439,    25.65505804,    28.75621891]]))
```

Bar plot to show relationship between beer dependence (S2BQ1B1) and beer drinking frequency (BEER_FEQMO)

In [11]:

```
%matplotlib notebook
sns.factorplot(x="BEER_FEQMO", y="S2BQ1B1", data=sub2, kind="bar", ci=None)
plt.xlabel('Days drink beer per month')
plt.ylabel('Proportion Beer Dependent')
```



A bar chart that shows as interviewee's beer drinking frequency increase, there is a higher chance they will become beer dependent.

Out[11]:

In [12]:

```
recode2 = {1: 1, 2.5: 2.5}
sub2['COMP1v2']= sub2['BEER_FEQMO'].map(recode2)
```

In [13]:

```
# contingency table of observed counts
ct2=pd.crosstab(sub2['S2BQ1B1'], sub2['COMP1v2'])
print (ct2)
```

```
COMP1v2   1.000000  2.500000
S2BQ1B1
0.000000      1172      1477
1.000000        40        80
```

Comparing beer dependency between groups with beer drinking frequency of 1 and 2.5.

In [14]:

```
# column percentages
colsum=ct2.sum(axis=0)
colpct=ct2/colsum
print(colpct)
```

```
COMP1v2   1.000000  2.500000
S2BQ1B1
0.000000  0.966997  0.948619
1.000000  0.033003  0.051381
```

Compare beer dependency percentage between two groups with beer drinking frequency of 1 and 2.5.

In [5]:

```
print ('chi-square value, p value, expected counts')
cs2= scipy.stats.chi2_contingency(ct2)
print (cs2)
```

```
chi-square value, p value, expected counts
(5.117284954394778, 0.023688651519463009, 1, array([[ 1159.47562297,  1489.5
2437703],
       [   52.52437703,    67.47562297]]))
```

Chi-square analysis result shows chi-square value is 5.11, p value is 0.0237, and expected counts is 1.

Post-hoc Analysis - Concise Code

In [16]:

```
sub3=sub2.copy()
cat = [1,2.5,4,8,14,26,30]

for x in range(0,len(cat)-1):
    for y in range(x+1,len(cat)):
        recode = { cat[x]:cat[x], cat[y]:cat[y]}
        sub3['temp'] = sub3['BEER_FEQMO'].map(recode)
        cont=pd.crosstab(sub3['S2BQ1B1'], sub3['temp'])
        cs= scipy.stats.chi2_contingency(cont)
        print("\n", cat[x], " versus ", cat[y],
              "Chi value: ", cs[0], "\tp value: ", cs[1])
```

```
 1  versus  2.5 Chi value:  5.11728495439       p value:  0.0236886515195

 1  versus  4 Chi value:  7.38180981336         p value:  0.00658868347191

 1  versus  8 Chi value:  31.489708359  p value:  2.00500133257e-08

 1  versus  14 Chi value:  25.8381672411        p value:  3.71273750161e-0
7

 1  versus  26 Chi value:  57.0712721169        p value:  4.20300560446e-1
4

 1  versus  30 Chi value:  78.2738078076        p value:  8.97034162448e-1
9

 2.5  versus  4 Chi value:  0.200755296546      p value:  0.654111881913

 2.5  versus  8 Chi value:  14.0623750892       p value:  0.0001768463156

 2.5  versus  14 Chi value:  10.2518760701      p value:  0.00136545664799

 2.5  versus  26 Chi value:  35.1701032845      p value:  3.02126345992e-0
9

 2.5  versus  30 Chi value:  53.5356271939      p value:  2.53945509183e-1
3

 4  versus  8 Chi value:  10.158373116  p value:  0.00143647327389

 4  versus  14 Chi value:  7.2097999272         p value:  0.00725065764704

 4  versus  26 Chi value:  29.6976634042        p value:  5.04956489946e-0
8

 4  versus  30 Chi value:  46.1496260739        p value:  1.09557923269e-1
1

 8  versus  14 Chi value:  0.0216664812731      p value:  0.882977803406

 8  versus  26 Chi value:  8.25308960637        p value:  0.00406826989653

 8  versus  30 Chi value:  16.3527743019        p value:  5.25791380981e-0
5

 14  versus  26 Chi value:  8.23247785293       p value:  0.00411473134878
```

```
  14  versus  30 Chi value:  15.5740908695       p value:  7.93342824608e-0
 5

  26  versus  30 Chi value:  0.415412986821      p value:  0.51923479448
```

Post hoc analysis that conduct chi-square analysis with different combinations of two groups with different beer drinking frequency

Result indicates as two groups’ beer drinking frequency difference enlarges, the chi-square value and p value increase as well.