First Name:

Last Name:

In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import statsmodels.formula.api as smf
import statsmodels.stats.multicomp as multi
import matplotlib.pyplot as plt
```

From Prac 1 to 3

In [2]:

```
nesarc = pd.read_csv('nesarc.csv', low_memory=False)
pd.set_option('display.float_format', lambda x:'%f'%x)
```

In [3]:

```
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
```

In [4]:

```
sub1=nesarc[(nesarc['AGE']>=26) & (nesarc['AGE']<=50) & (nesarc['S2AQ5A']==1)]
sub2=sub1.copy()
```

In [5]:

```
#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)
```

In [6]:

```
recode2 = {1:30, 2:26, 3:14, 4:8, 5:4, 6:2.5, 7:1}
sub2['BEER_FEQMO']= sub2['S2AQ5B'].map(recode2)
sub2['BEER_FEQMO']= pd.to_numeric(sub2['BEER_FEQMO'])
```

In [2]:

```
# Creating a secondary variable multiplying the days consumed beer/month and the number of
sub2['NUMBEERMO_EST']=sub2['BEER_FEQMO'] * sub2['S2AQ5D']
sub2['NUMBEERMO_EST']= pd.to_numeric(sub2['NUMBEERMO_EST'])
```

In [8]:

```
ct1 = sub2.groupby('NUMBEERMO_EST').size()
print (ct1)
```

```
NUMBEERMO_EST
1.000000      477
2.000000      407
2.500000      414
3.000000      172
4.000000      429
5.000000      623
6.000000       36
7.000000        5
7.500000      267
8.000000      635
10.000000     119
12.000000     296
12.500000      48
14.000000     160
15.000000      87
16.000000     561
17.500000       5
18.000000       1
20.000000      81
22.500000       3
24.000000     410
25.000000       6
26.000000      51
27.500000       1
28.000000     242
30.000000      62
32.000000     168
35.000000       1
36.000000       3
37.500000       2
             ...
98.000000       9
104.000000     37
112.000000     21
120.000000     39
130.000000     13
140.000000      5
144.000000      2
150.000000     18
156.000000     54
168.000000     27
180.000000     77
182.000000      6
192.000000      3
208.000000     10
210.000000      5
234.000000      2
240.000000     13
252.000000      5
260.000000      3
270.000000      4
300.000000      6
312.000000     14
360.000000     25
468.000000      1
```

Result shows how many interviewees drank from 1 to 900 bottles of beer per month

```
510.000000      1
520.000000      1
540.000000      2
624.000000      1
720.000000      2
900.000000      1
Length: 75, dtype: int64
```

Categorical -> Quantitative - ANOVA

In [9]:

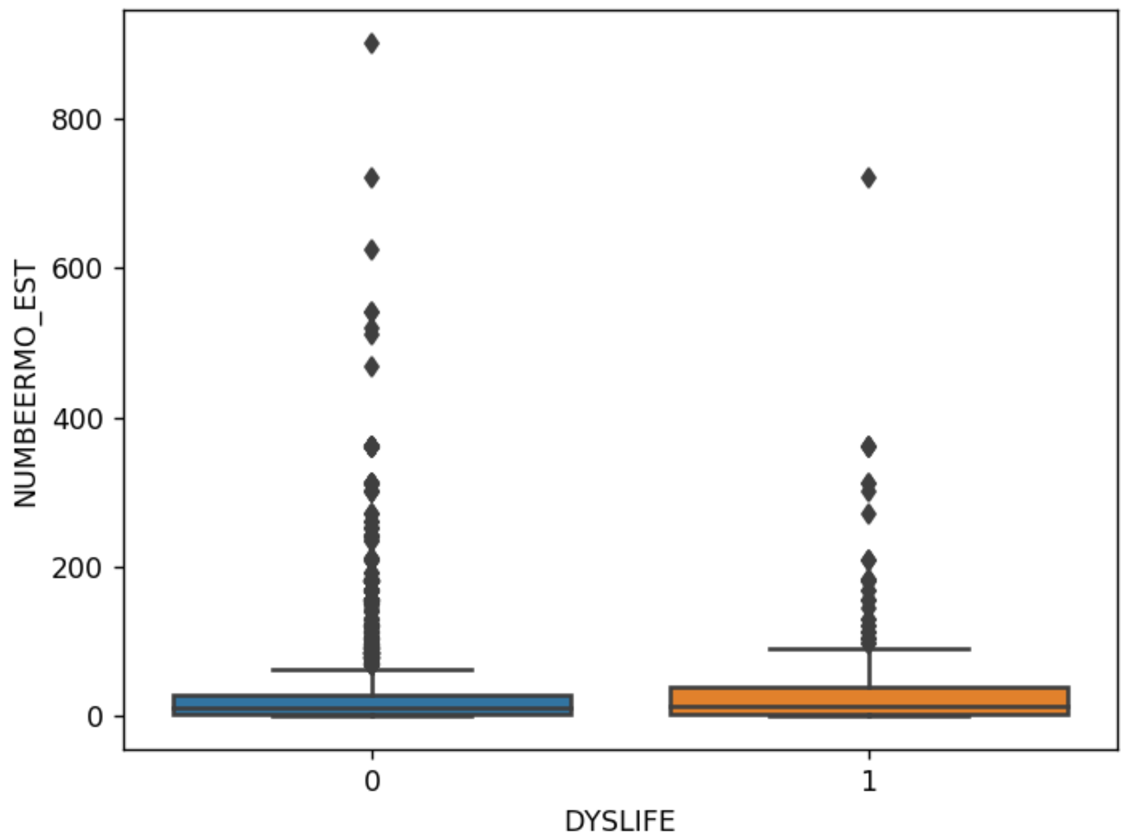
```
sub2['DYSLIFE'] = sub2['DYSLIFE'].astype('category')
```

Draw boxplot to show relationship between minor depression status (DYSLIFE (categorical)) and estimated number of beer consumed (NUMBEERMO_EST (quantitative))

In [22]:

```
%matplotlib notebook
sns.boxplot(x='DYSLIFE', y='NUMBEERMO_EST', data=sub2)
plt.xlabel('DYSLIFE')
plt.ylabel('NUMBEERMO_EST')
```

Figure 1



The box plot indicates people who are slightly depressed drank higher volume of beer per month.

Out[22]:

```
Text(0,0.5,'NUMBEERMO_EST')
```

Perform ANOVA analysis between minor depression status (DYSLIFE (categorical)) and estimated number of beer consumed (NUMBEERMO_EST (quantitative))

In [11]:

```
model1 = smf.ols(formula='NUMBEERMO_EST ~ C(DYSLIFE)', data=sub2).fit()
print (model1.summary())
```

```
                            OLS Regression Results

==============================================================================
==
Dep. Variable:          NUMBEERMO_EST   R-squared:                       0.0
03
Model:                            OLS   Adj. R-squared:                  0.0
03
Method:                 Least Squares   F-statistic:                     20.
23
Date:                Fri, 27 Apr 2018   Prob (F-statistic):           6.99e-
06
Time:                        14:58:59   Log-Likelihood:                -3880
4.
No. Observations:                7303   AIC:                       7.761e+
04
Df Residuals:                    7301   BIC:                       7.763e+
04
Df Model:                           1

Covariance Type:            nonrobust

==============================================================================
=======
                      coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------------------------
Intercept          27.2277      0.587     46.361      0.000      26.076
28.379
C(DYSLIFE)[T.1]    12.9670      2.883      4.497      0.000       7.315
18.619
==============================================================================
==
Omnibus:                     7622.371   Durbin-Watson:                   2.0
26
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           640965.7
69
Skew:                           5.150   Prob(JB):                         0.
00
Kurtosis:                      47.725   Cond. No.                         5.
02
==============================================================================
==

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
```

A summary for ANOVA analysis of relation between minor depression and estimated bottle of beer drank per month.

In [12]:

```
sub3 = sub2[['NUMBEERMO_EST', 'DYSLIFE']].dropna()
```

print the mean of number of beer consumed grouped

by minor depression status

In [13]:

```
print ('means for NUMBEERMO_EST by minor depression status')
m1= sub3.groupby('DYSLIFE').mean()
print (m1)
```

```
means for NUMBEERMO_EST by minor depression status
         NUMBEERMO_EST
DYSLIFE
0            27.227714
1            40.194719
```

People with minor depression on average drank 40 bottles of beer per month, while people with no depression drank 27 on average.

print the standard deviation (std) of number beer consumed grouped by minor depression status

In [14]:

```
print ('standard deviations for NUMBEERMO_EST by minor depression status')
sd1 = sub3.groupby('DYSLIFE').std()
print (sd1)
```

```
standard deviations for NUMBEERMO_EST by minor depression status
         NUMBEERMO_EST
DYSLIFE
0            47.678467
1            75.407118
```

Group with depression has higher standard deviation, which indicates the values in that group are more volatile.

Categorical (>2) -> Quantitative - ANOVA

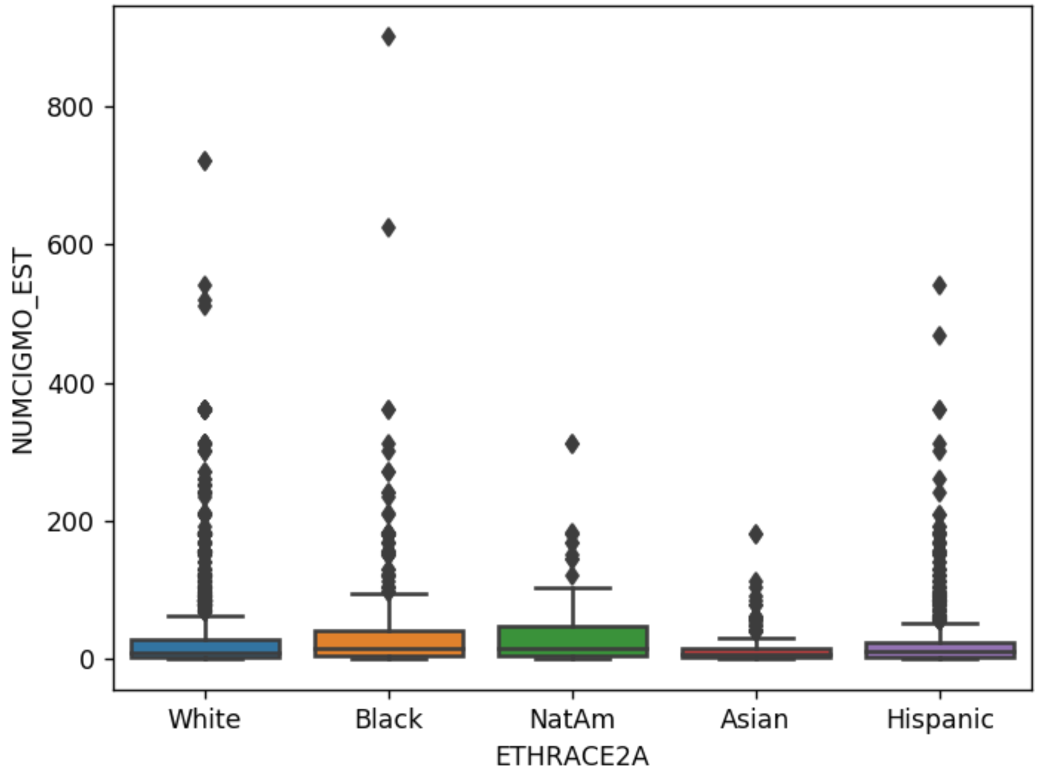
In [15]:

```
sub2['ETHRACE2A'] = sub2['ETHRACE2A'].astype('category')
sub2['ETHRACE2A']=sub2['ETHRACE2A'].cat.rename_categories(["White", "Black", "NatAm", "Asia
```

Draw boxplot to show relationship between ethinicity (ETHRACE2A (categorical)) and estimated number of beer consumed (NUMBEERMO_EST (quantitative))

In [16]:

```
%matplotlib notebook
sns.boxplot(x='ETHRACE2A', y='NUMBEERMO_EST', data=sub2)
plt.xlabel('ETHRACE2A')
plt.ylabel('NUMCIGMO_EST')
```



A box plot that shows number of beer drank per month among 5 ethnic groups.

Black and Native Americans are two groups drank most beer on average.

Asians and Hispanic are two groups with lowest amount of beer consumption.

Out[16]:

```
Text(0,0.5,'NUMCIGMO_EST')
```

In [17]:

```
sub4 = sub2[['NUMBEERMO_EST', 'ETHRACE2A']].dropna()
```

Perform ANOVA analysis between ethinicity (ETHRACE2A (categorical)) and estimated number of beer consumed (NUMBEERMO_EST (quantitative))

In [18]:

```
model2 = smf.ols(formula='NUMBEERMO_EST ~ C(ETHRACE2A)', data=sub4).fit()
print (model2.summary())
```

```
                            OLS Regression Results                            
==============================================================================
==
Dep. Variable:          NUMBEERMO_EST   R-squared:                       0.0
05
Model:                            OLS   Adj. R-squared:                  0.0
04
Method:                 Least Squares   F-statistic:                     8.2
61
Date:                Fri, 27 Apr 2018   Prob (F-statistic):           1.21e-
06
Time:                        14:59:00   Log-Likelihood:                -3879
7.
No. Observations:                7303   AIC:                         7.760e+
04
Df Residuals:                    7298   BIC:                         7.764e+
04
Df Model:                           4                                         

Covariance Type:            nonrobust                                         
==============================================================================
================
                               coef    std err          t      P>|t|      
[0.025      0.975]
------------------------------------------------------------------------------
------------------------------------
Intercept                   27.8589      0.742     37.535      0.000      2
6.404      29.314
C(ETHRACE2A)[T.Black]        4.5843      1.656      2.768      0.006      
1.338       7.831
C(ETHRACE2A)[T.NatAm]       11.6496      4.581      2.543      0.011      
2.670      20.629
C(ETHRACE2A)[T.Asian]      -11.2589      3.594     -3.133      0.002     -1
8.304      -4.214
C(ETHRACE2A)[T.Hispanic]    -3.2403      1.464     -2.213      0.027      -
6.111      -0.370
==============================================================================
==
Omnibus:                     7639.304   Durbin-Watson:                   2.0
27
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             646522.8
33
Skew:                           5.167   Prob(JB):                         0.
00
Kurtosis:                      47.921   Cond. No.                         8.
28
==============================================================================
==

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
```

A summary for ANOVA analysis of relation between ethnic group and monthly beer consumption

print the mean of number of beer consumed grouped by ethinicity

In [19]:

```
print ('means for NUMBEERMO_EST by ethinicity')
m2= sub4.groupby('ETHRACE2A').mean()
print (m2)
```

```
means for NUMBEERMO_EST by ethinicity
           NUMBEERMO_EST
ETHRACE2A
White          27.858922
Black          32.443182
NatAm          39.508475
Asian          16.600000
Hispanic       24.618638
```

Native Americans has the highest average of monthly beer consumption - 39.5 bottles.

And Asians has the lowest number - 16.6 bottles.

print the standard deviation (std) of number of beer consumed grouped by ethinicity

In [20]:

```
print ('standard deviations for NUMBEERMO_EST by ethnicity')
sd2 = sub4.groupby('ETHRACE2A').std()
print (sd2)
```

```
standard deviations for NUMBEERMO_EST by ethnicity
           NUMBEERMO_EST
ETHRACE2A
White          50.537013
Black          55.289755
NatAm          57.231386
Asian          25.572698
Hispanic       41.073842
```

White, Black, and Native American groups have a std value that is over 50.

Hispanic group is slightly better with a standard deviation of 41.

Asians are the most consistent dataset, with standard deviation of 25.6.

Perform Tukey's Honestly Significant Difference (Post hoc) test

In [21]:

```
mc1 = multi.MultiComparison(sub4['NUMBEERMO_EST'], sub4['ETHRACE2A'])
res1 = mc1.tukeyhsd()
print(res1.summary())
```

```
Multiple Comparison of Means - Tukey HSD,FWER=0.05
==================================================
 group1   group2  meandiff  lower    upper  reject
--------------------------------------------------
 Asian    Black   15.8432   5.4332  26.2532  True
 Asian   Hispanic  8.0186  -2.1752  18.2124 False
 Asian    NatAm   22.9085   7.2827  38.5343  True
 Asian    White   11.2589   1.4533  21.0646  True
 Black   Hispanic -7.8245  -13.1332 -2.5159  True
 Black    NatAm    7.0653  -5.9129  20.0435 False
 Black    White   -4.5843   -9.103  -0.0655  True
Hispanic  NatAm   14.8898   2.0844  27.6953  True
Hispanic  White    3.2403  -0.7553   7.2359 False
 NatAm    White   -11.6496 -24.1482  0.8491 False
--------------------------------------------------
```

A summary for Turkey's Honestly Significant Difference (Post hoc) Test. The summary shows comparison and describes differences between each ethnic group.