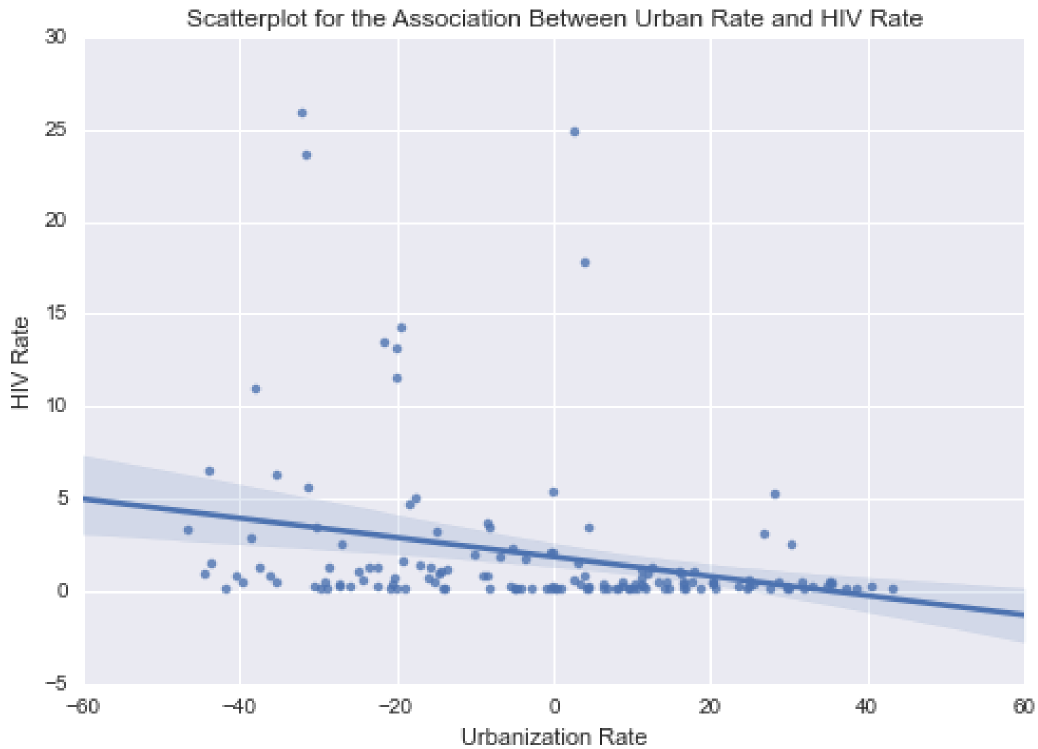
Assignment 2: HIV Rate vs. Urbanization Rate

I looked at association between HIV rate (quantitative response variable) and urban rate (quantitative explanatory variable). Since my explanatory variable is quantitative, I first centered it and then calculate the mean to check my centering.

Mean of urban rate: 56.76935960591131
Mean of centered urban rate: 1.8446109445594718e-14 (very close to zero)

Next, I rested a linear regression model to understand the relationship between the HIV rate and urban rate. Results are shown in the table and in the graph (for the centered explanatory variable). p-value = 0.001 < 0.05, then we can reject the null hypothesis and get that HIV rate is associated with the urban rate. Regression coefficients are the following (see in the table): hivrate= 1.86 – 0.05*urbanrate. The countries with high urban rate have a lower rate of HIV infection (hivrate=1.86-0.05*40= - 0.14). The countries with low urban rate have a higher rate of HIV infection (hivrate=1.86-0.05*(-40) = 3.86). In this analysis, the outliers were not removed as they represented the low urban countries with the highest HIV rates, such as Swaziland, Namibia (parts of Africa), South Africa and others.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                hivrate   R-squared:                       0.072
Model:                            OLS   Adj. R-squared:                  0.066
Method:                 Least Squares   F-statistic:                     11.28
Date:                Mon, 30 Nov 2015   Prob (F-statistic):            0.00100
Time:                        20:11:13   Log-Likelihood:                -419.59
No. Observations:                 147   AIC:                             843.2
Df Residuals:                     145   BIC:                             849.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      1.8586      0.350      5.315      0.000         1.168     2.550
urbanrate     -0.0524      0.016     -3.358      0.001        -0.083    -0.022
==============================================================================
Omnibus:                      141.274   Durbin-Watson:                   1.984
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1678.634
Skew:                           3.663   Prob(JB):                         0.00
Kurtosis:                      17.845   Cond. No.                         22.4
==============================================================================
```



Python Code:

```
"""
Created on Mon Nov 30 2015

@author: violetgirl
"""
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf
import statsmodels.stats.multicomp as multi
import scipy
import scipy.stats
import statsmodels.api
import seaborn as sb
import matplotlib.pyplot as plt

# load gapminder dataset
data = pd.read_csv('gapminder.csv',low_memory=False)
# lower-case all DataFrame column names
data.columns = map(str.lower, data.columns)
```

```
# bug fix for display formats to avoid run time errors
pd.set_option('display.float_format', lambda x:'%.2f'%x)

# remove missing values
data_copy=data.copy()
data_copy=data_copy.dropna()

# convert variables to numeric format using convert_objects function
# quantitative reponse variable - HIV rate; quantitative explanatory variable - urbanrate
data_copy['hivrate'] = data_copy['hivrate'].convert_objects(convert_numeric=True)
data_copy['urbanrate'] = data_copy['urbanrate'].convert_objects(convert_numeric=True)

####################################################################
BASIC LINEAR REGRESSION
###################################################################
# listwise deletion for calculating means for regression model observations
sub1 = data_copy[['hivrate', 'urbanrate']]

# calculate mean for quantitative explanatory variable
print ("Mean of Urban Rate")
print(sub1['urbanrate'].mean())
# center a quantitative explanatory variable
sub1['urbanrate']=sub1['urbanrate']-sub1['urbanrate'].mean()
# check the mean of centered variable
print ("Mean of Centered Urban Rate")
print(sub1['urbanrate'].mean())

sub1=sub1.dropna()

scat2 = sb.regplot(x="urbanrate", y="hivrate", scatter=True, data=sub1)
plt.xlabel('Urbanization Rate')
plt.ylabel('HIV Rate')
plt.title ('Scatterplot for the Association Between Urban Rate and HIV Rate')
print(scat2)

print ("OLS regression model for the association between urban rate and HIV
rate")
# resonse variable (y) ~ explanatory variable (x)
reg2 = smf.ols('hivrate ~ urbanrate', data=sub1).fit()
print (reg2.summary())
```