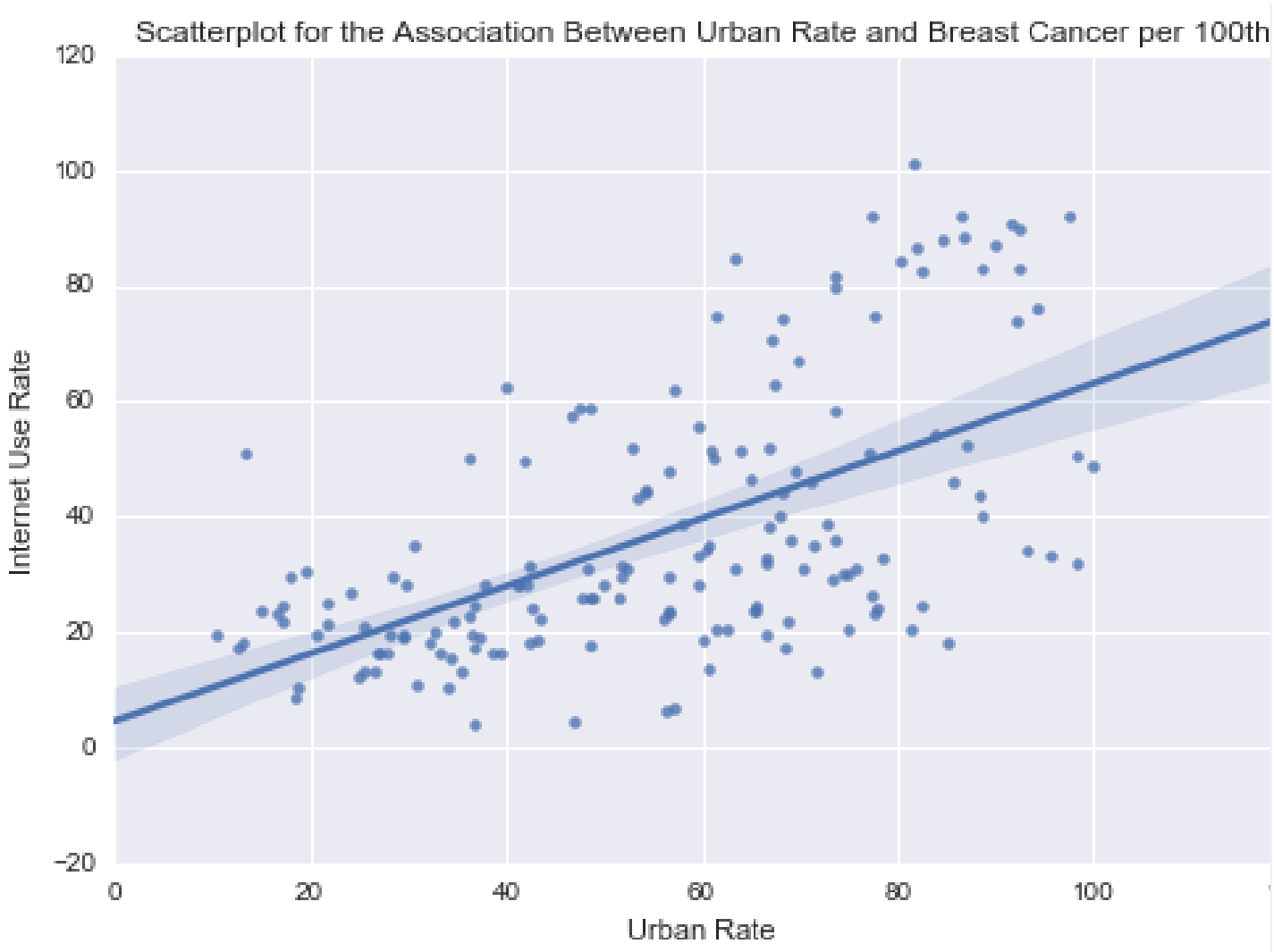
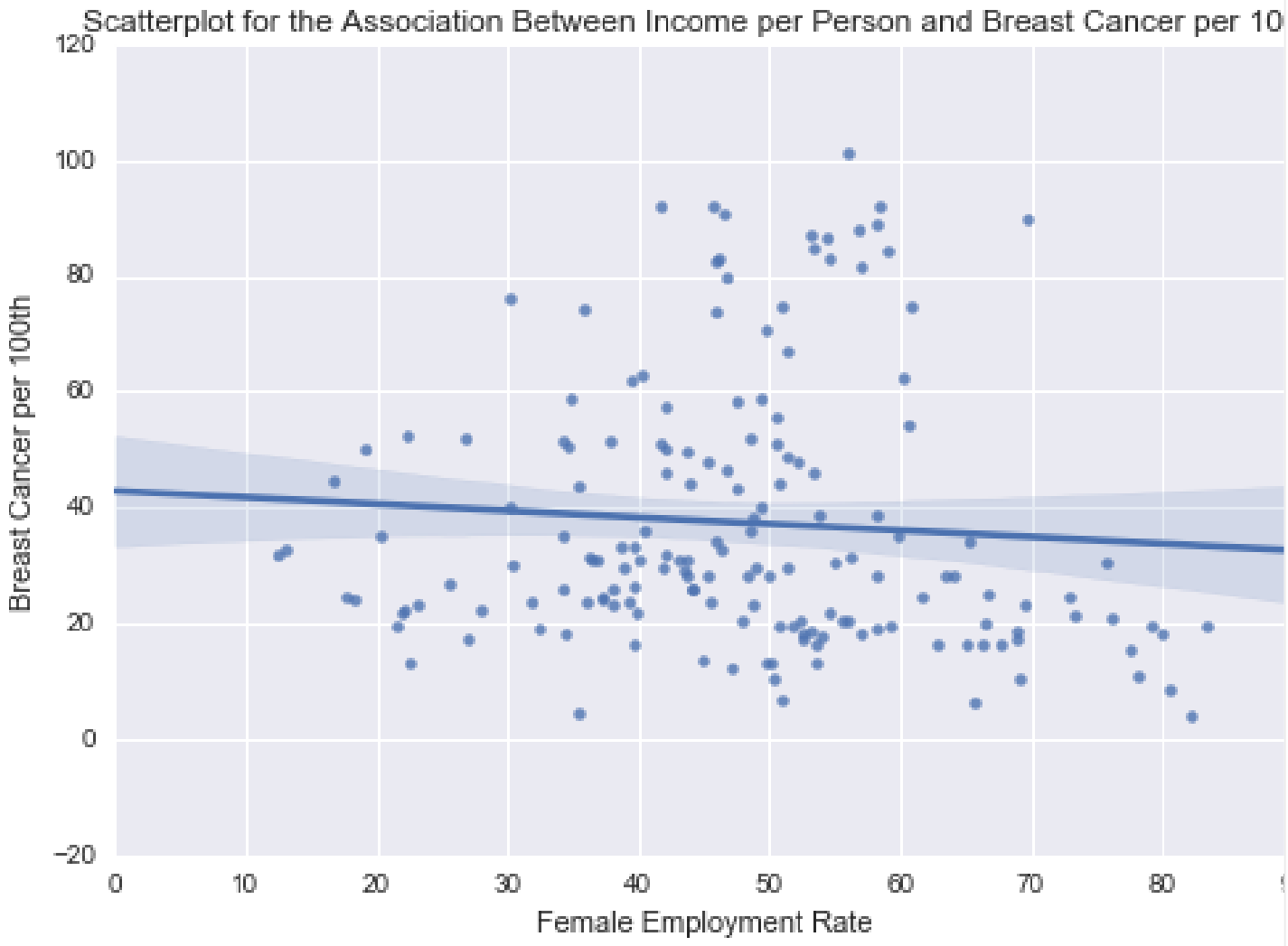
Untitled

ARCHIVE


Scatterplot for the Association Between Urban Rate and Breast Cancer per 100th
Internet Use Rate
Urban Rate
120
100
80
60
40
20
0
−20
0
20
40
60
80
100


Scatterplot for the Association Between Income per Person and Breast Cancer per 10
Breast Cancer per 100th
Female Employment Rate
120
100
80
60
40
20
0
−20
0
10
20
30
40
50
60
70
80

Code: See Below

Dataset: GapMinder

Alternate Hypothesis: Urban Rates affect breast cancer rates

Output results:

> **association between urbanrate and breastcancerper100th**
>
> (0.58186863212416406, 1.6316781907187033e-16)
>
> r2 = 0.3364

> **association between incomeperperson and breastcancerper100th**
>
> (-0.072974651960486414, 0.34865183373948994)
>
> r2 = 0.0049

Analysis: After reviewing the scatter plot, correlation coefficient and p-value for the relationship between income per person and breast cancer, I have found that there is a very weak, negative linear relationship between these two variables.

However, after reviewing the scatter plot, correlation coefficient for the relationship between urban rate and breast cancer, I have found that there is a pretty strong , positive linear relationship between these two variables. The correlation coefficient for these two variables is 0.58 which isn't very strong but it is pretty strong. The p-value for this relationship is significantly below 0.05. Additionally, by looking at the scatter plot itself, I can visually see this assumption being confirmed. After squaring the coefficient, I can predict 33.64% of the variability we will see in the rate of internet use, by knowing the urban rate.

—— Code ——

```
#authored by tumblr blog mestupmxpxfan10

# library import
import pandas
import numpy
import seaborn
import scipy
import matplotlib.pyplot as plt

# dataset import
data = pandas.read_csv('gapminder.csv', low_memory=False)
```

```
# convert variables to numbers
data['breastcancerper100th’] =
data['breastcancerper100th’].convert_objects(convert_numeric=True)
data['femaleemployrate’] = data['femaleemployrate’].convert_objects(convert_numeric=True)
data['urbanrate’] = data['urbanrate’].convert_objects(convert_numeric=True)

# creating of subsets of data that only includes breast cancer data with values
datausing = data[['breastcancerper100th’,'femaleemployrate’, 'urbanrate’]]
data_clean = datausing.dropna()
data_clean2 = data_clean.copy()

scat1 = seaborn.regplot(x=“urbanrate”, y=“breastcancerper100th”, fit_reg=True,
data=data_clean2)
plt.xlabel('Urban Rate’)
plt.ylabel('Internet Use Rate’)
plt.title('Scatterplot for the Association Between Urban Rate and Breast Cancer per 100th’)

scat2 = seaborn.regplot(x=“femaleemployrate”, y=“breastcancerper100th”, fit_reg=True,
data=data_clean2)
plt.xlabel('Female Employment Rate’)
plt.ylabel('Breast Cancer per 100th’)
plt.title('Scatterplot for the Association Between Income per Person and Breast Cancer per
100th’)

print ('association between urbanrate and breastcancerper100th’)
print (scipy.stats.pearsonr(data_clean2['urbanrate’], data_clean2['breastcancerper100th’]))

print ('association between incomeperperson and breastcancerper100th’)
print (scipy.stats.pearsonr(data_clean2['femaleemployrate’],
data_clean2['breastcancerper100th’]))
```

Feb 28th, 2016