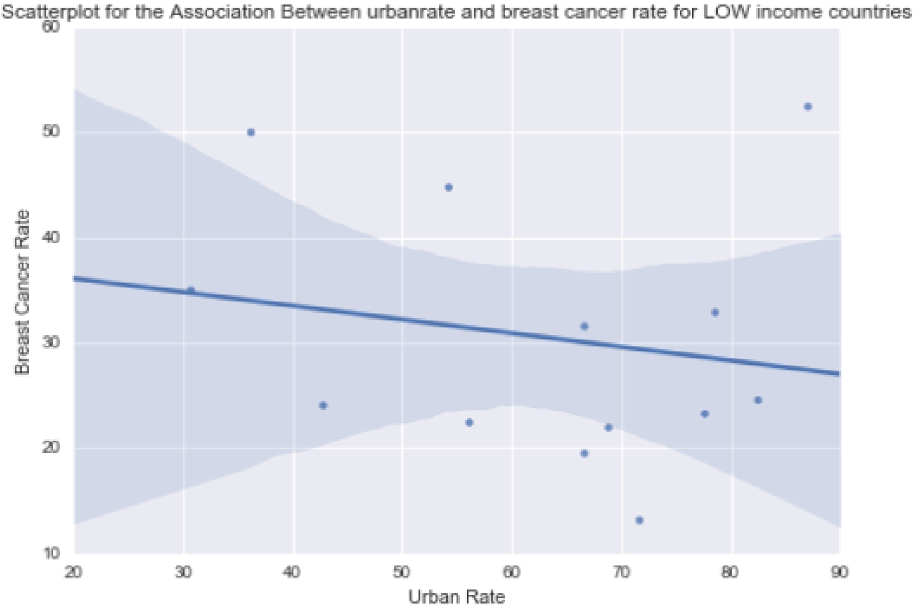
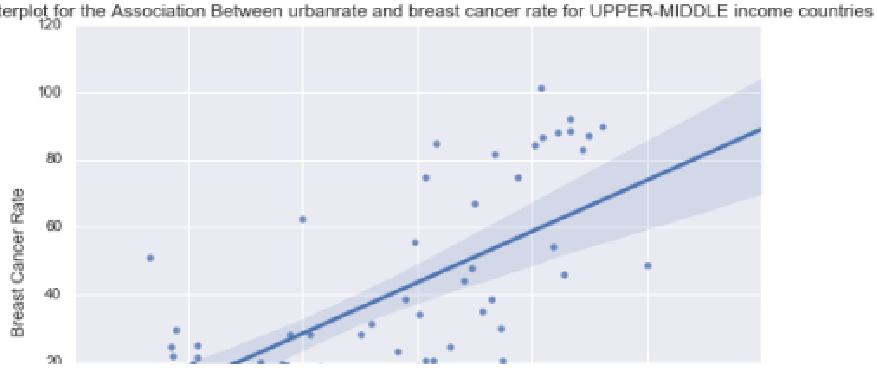
Data Analysis and Interpretation Specialization

Scatterplot for the Association Between urbanrate and breast cancer rate for LOW income countries

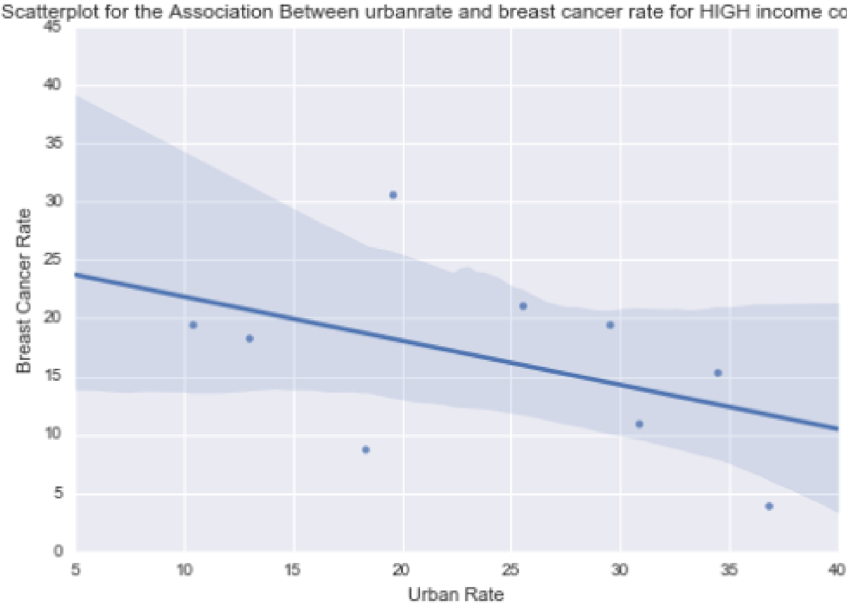


Scatterplot for the Association Between urbanrate and breast cancer rate for LOW-MIDDLE income countries

Scatterplot for the Association Between urbanrate and breast cancer rate for UPPER-MIDDLE income countries



Scatterplot for the Association Between urbanrate and breast cancer rate for HIGH income countries



Dataset: GapMinder

Python Code: See Below

Output (images above and text below):

> association between urban rate and breast cancer rate for LOW employment rate
> (-0.1906961863964324, 0.53259163859845038)
>
> association between urban rate and breast cancer rate for LOW-MIDDLE employment rate
> (**0.52672937480451154**, **4.3912034773430283e-07)**
>
> association between urban rate and breast cancer rate for UPPER-MIDDLE employment rate
> (**0.66909103154213945**, **1.5077849400082526e-09**)
>
> association between urban rate and breast cancer rate for HIGH employment rate
> (-0.4542943325503318, 0.21927741877360313)

Alternate Hypothesis: Urban Rates affect breast cancer rates

Summary: Initially in my earlier analysis between urban rates and breast cancer rates, I found that I had a p-value significantly below 0.05, and a significant chi-square value. which led me to believe that I could reject the null hypothesis of no correlation between urban rate and breast cancer rate. To help me understand this data better and to see if there was a potential moderator, I ran a correlation coefficient test, using the variable "female employment rate". I chose this variable, because it was initially a part of my original hypothesis, of which I removed due to lack of support for that variable.

In my correlation coefficient test I found that there was no significance between my main 3 variables, in regards to countries with a **low employment rate**, and those countries with a **high employment rate**. However, I was able to determine that with **countries that have a low-middle and upper-middle employment rate, that there is a significant correlation rate and p-value and that I can reject the null hypothesis**. Additionally, after reviewing the graphs themselves (graph 2 and 3), I can visually confirm this significance as well.

Post-hoc Tests: Not needed due to use of pearson correlation

—– CODE —–

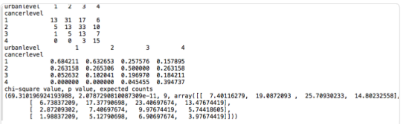
```
# -*- coding: utf-8 -*-
“”“
Created on Tue Mar  1 17:11:20 2016
@author: tumblr blog mestupmxpxfan10
”“”

# library import
import pandas
```

```
import numpy
import scipy.stats
import seaborn
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 contain values
datausing = data[['breastcancerper100th', 'urbanrate', 'femaleemployrate']]
data_clean = datausing.dropna()
data_clean2 = data_clean.copy()

#creating categorical variable out of female employment rate
def employgrp (row):
  if (row['femaleemployrate'] <= 25):
    return 1
  elif (row['femaleemployrate'] <= 50) & (row['femaleemployrate'] > 25):
    return 2
  elif (row['femaleemployrate'] <= 75) & (row['femaleemployrate'] > 50):
    return 3
  elif (row['femaleemployrate'] > 75):
    return 4

data_clean2['employgrp'] = data_clean2.apply (lambda row: employgrp (row),axis=1)
chk1 = data_clean2['employgrp'].value_counts(sort=False, dropna=False)
print(chk1)

#data frames that include only 1 employment group each
sub1=data_clean[(data_clean2['employgrp']== 1)]
sub2=data_clean[(data_clean2['employgrp']== 2)]
sub3=data_clean[(data_clean2['employgrp']== 3)]
sub4=data_clean[(data_clean2['employgrp']== 4)]

#pearson correlation measuring association between urban rate and cancer rate, as well as p-value
print ('association between urbanrate and breast cancer rate for LOW employment rate')
print (scipy.stats.pearsonr(sub1['urbanrate'], sub1['breastcancerper100th']))
print ('      ')
print ('association between urbanrate and breast cancer rate for LOW-MIDDLE employment rate')
print (scipy.stats.pearsonr(sub2['urbanrate'], sub2['breastcancerper100th']))
print ('      ')
print ('association between urbanrate and breast cancer rate for UPPER-MIDDLE employment rate')
print (scipy.stats.pearsonr(sub3['urbanrate'], sub3['breastcancerper100th']))
print ('      ')
print ('association between urbanrate and breast cancer rate for HIGH employment rate')
print (scipy.stats.pearsonr(sub4['urbanrate'], sub4['breastcancerper100th']))
#%%
scat1 = seaborn.regplot(x="urbanrate", y="breastcancerper100th", data=sub1)
plt.xlabel('Urban Rate')
plt.ylabel('Breast Cancer Rate')
plt.title('Scatterplot for the Association Between urbanrate and breast cancer rate for LOW income
countries')
print (scat1)
#%%
scat2 = seaborn.regplot(x="urbanrate", y="breastcancerper100th", data=sub2)
plt.xlabel('Urban Rate')
plt.ylabel('Breast Cancer Rate')
plt.title('Scatterplot for the Association Between urbanrate and breast cancer rate for LOW-MIDDLE
income countries')
print (scat2)
#%%
scat3 = seaborn.regplot(x="urbanrate", y="breastcancerper100th", data=sub3)
plt.xlabel('Urban Rate')
plt.ylabel('Breast Cancer Rate')
plt.title('Scatterplot for the Association Between urbanrate and breast cancer rate for UPPER-MIDDLE
```

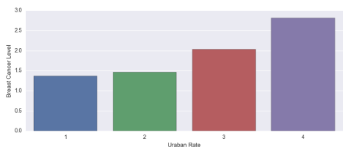
income countries')
print (scat3)
#%%
scat4 = seaborn.regplot(x="urbanrate", y="breastcancerper100th", data=sub4)
plt.xlabel('Urban Rate')
plt.ylabel('Breast Cancer Rate')
plt.title('Scatterplot for the Association Between urbanrate and breast cancer rate for HIGH income countries')
print (scat4)

Mar 3rd, 2016

MORE YOU MIGHT LIKE



```
urbanlevel    1   2   3   4
cancerlevel
1            13  31  17   6
2             5  13  33  10
3             1   5  13   7
4             0   0   3  15
urbanlevel          1         2         3         4
cancerlevel
1            0.684211  0.632653  0.257576  0.157895
2            0.263158  0.265306  0.500000  0.263158
3            0.052632  0.102041  0.196970  0.184211
4            0.000000  0.000000  0.045455  0.394737
chi-square value, p value, expected counts
(69.310196924193988, 2.0787290810087309e-11, 9, array([[  7.40116279,  19.0872093 ,  25.70930233,  14.80232558],
       [  6.73837209,  17.37790698,  23.40697674,  13.47674419],
       [  2.87209302,   7.40697674,   9.97674419,   5.74418605],
       [  1.98837209,   5.12790698,   6.90697674,   3.97674419]]))
```



Code: See Below

Dataset: GapMinder

Alternate Hypothesis: Urban Rates affect breast cancer rates

Chi-Square Analysis 1: Chi-Square Value = 69.310196924193988, p-value =2.0787290810087309e-11, df = 4-1 = 3

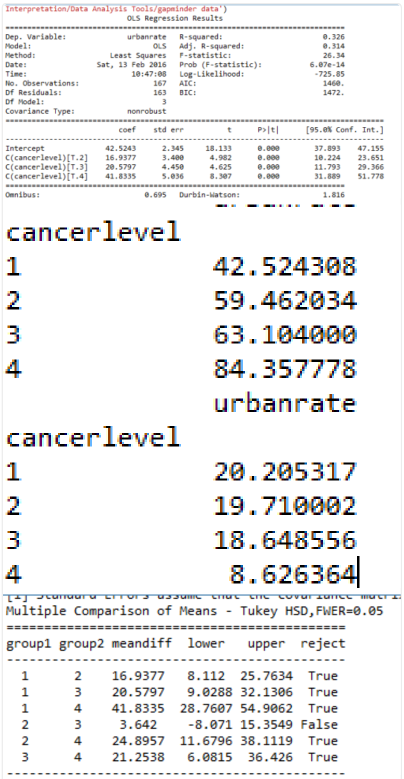
There was a significant chi-square value, which suggests that there is a high probability of independence between my variables and i should reject the null hypothesis. My p-value is extremely low, as well. So for this reason, I have also done a post-hoc Chi-Square analysis test

Post-Hoc Chi-Square analysis: bonferroni adjustment = 0.003125

Post hoc comparisons of cancer rates by urban rate revealed that the lowest cancer rates were seen among those with the lowest urban rates. We see major differences in cancer rates between group 1 and all other groups. Differences between group 2 and 3 were found, but differences between group 2 and 4 and groups 3 and 4 were not found to be significant enough to reject the null hypothesis.

Output results:

COMP1v2 1 2
cancerlevel
1 13 31
2 5 13
3 1 5
COMP1v2 1 2
cancerlevel
1 0.684211 0.632653



```
Interpretation/Data Analysis Tools/gapminder data')
                            OLS Regression Results
==============================================================================
Dep. Variable:              urbanrate   R-squared:                       0.326
Model:                            OLS   Adj. R-squared:                  0.314
Method:                 Least Squares   F-statistic:                     26.34
Date:                Sat, 13 Feb 2016   Prob (F-statistic):           6.07e-14
Time:                        10:47:08   Log-Likelihood:                -725.85
No. Observations:                 167   AIC:                             1460.
Df Residuals:                     163   BIC:                             1472.
Df Model:                           3
Covariance Type:            nonrobust
=================================================================================
                        coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------
Intercept            42.5243      2.345     18.133      0.000        37.893    47.155
C(cancerlevel)[T.2]  16.9377      3.400      4.982      0.000        10.224    23.651
C(cancerlevel)[T.3]  20.5797      4.450      4.625      0.000        11.793    29.366
C(cancerlevel)[T.4]  41.8335      5.036      8.307      0.000        31.889    51.778
==============================================================================
Omnibus:                        0.695   Durbin-Watson:                   1.816

cancerlevel
1              42.524308
2              59.462034
3              63.104000
4              84.357778
               urbanrate
cancerlevel
1              20.205317
2              19.710002
3              18.648556
4               8.626364

Multiple Comparison of Means - Tukey HSD,FWER=0.05
=============================================
group1 group2 meandiff  lower   upper  reject
---------------------------------------------
  1      2    16.9377   8.112  25.7634  True
  1      3    20.5797   9.0288 32.1306  True
  1      4    41.8335  28.7607 54.9062  True
  2      3     3.642   -8.071  15.3549 False
  2      4    24.8957  11.6796 38.1119  True
  3      4    21.2538   6.0815  36.426  True
---------------------------------------------
```

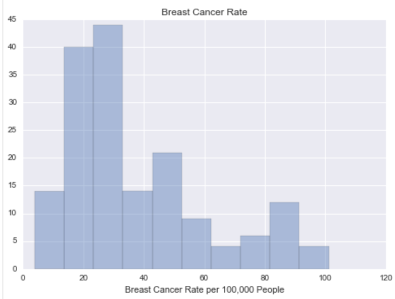
Code: See Below

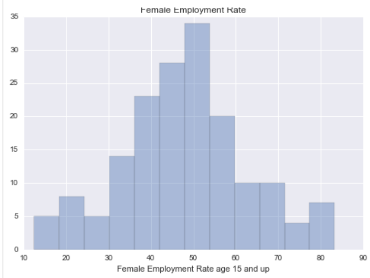
Dataset: GapMinder

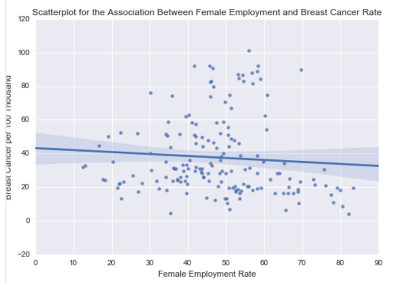
Alternate Hypothesis: Urban Rates affect breast cancer rates

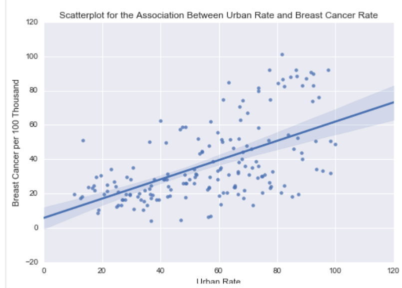
Summary: I set out to see if there was a correlation between breast cancer rates, and urban rates. After running the OLS program, i found that I had a p-value significantly below 0.05, which led me to believe that I could reject the null hypothesis of no correlation between urban rate and breast cancer rate. After running the Tukey Honestly Significant Difference test, I later confirmed my data to confirm my alternate hypothesis as plausible, especially among tier 1 and tier 4 of the breast cancer rates. While, tiers 2 and 3 are not that significantly different to accept my alternate hypothesis

— CODE —



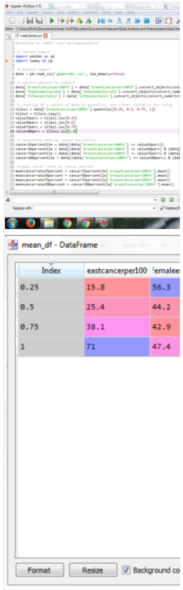






Research Question: Is there a correlation between breast cancer per 100,000 values and female employee rates?

Data management techniques: Create a subset of data that removes all nan values. Create univariate graphs for all 3 variables in this new data set (breast cancer per 100th, female employment rate, urban rate) and create 2 bivariate graphs. The first graph had female



*The attached im
from my program

Summary:

Due to the nature
question and the
working, I had to
approach to revie

Since my questio
breast cancer rat
(breastcancerper
possible correlati

my data more rea
try to break the d
which are based
(25%, 50%, 75%,
based off of the g
Having then brok
4 manageable ro
a count would no
counts would be
what i ended up v
which was based
the data given to

2 0.263158 0.265306
3 0.052632 0.102041
chi-square value, p value, expected counts
(0.43528735258058576, **0.80441202168159909**, 2, array([[12.29411765, 31.70588235],
[5.02941176, 12.97058824],
[1.67647059, 4.32352941]]))
COMP1v3 1 3
cancerlevel
1 13 17
2 5 33
3 1 13
4 0 3
COMP1v3 1 3
cancerlevel
1 0.684211 0.257576
2 0.263158 0.500000
3 0.052632 0.196970
4 0.000000 0.045455
chi-square value, p value, expected counts
(12.189150108125176, **0.0067625168252241829**, 3, array([[6.70588235, 23.29411765],
[8.49411765, 29.50588235],
[3.12941176, 10.87058824],
[0.67058824, 2.32941176]]))
COMP1v4 1 4
cancerlevel
1 13 6
2 5 10
3 1 7
4 0 15
COMP1v4 1 4
cancerlevel
1 0.684211 0.157895
2 0.263158 0.263158
3 0.052632 0.184211
4 0.000000 0.394737
chi-square value, p value, expected counts
(19.588815789473685, **0.00020652162048918167**, 3, array([[6.33333333, 12.66666667],
[5. , 10.],
[2.66666667, 5.33333333],
[5. , 10.]]))
COMP1v5 2 3
cancerlevel
1 31 17
2 13 33
3 5 13
4 0 3
COMP1v5 2 3
cancerlevel
1 0.632653 0.257576
2 0.265306 0.500000
3 0.102041 0.196970
4 0.000000 0.045455
chi-square value, p value, expected counts
(17.197302532123963, **0.00064368244549237359**, 3, array([[20.45217391, 27.54782609],
[19.6 , 26.4],

```
#authored by tumblr blog mestupmxpxfan10
# library import
import pandas
import numpy
import statsmodels.formula.api as smf
import statsmodels.stats.multicomp as multi

# 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()

#new variables, based off rate or per 100,000 value. This is categorical 1-4 Value
#Value Index: 1=0-24.99, 2=25-49.99, 3=50-74.99 4=75+
def cancerlevel (column):
  if (column['breastcancerper100th'] < 25):
    return 1
  if (column['breastcancerper100th'] > 25) & (column['breastcancerper100th'] < 50):
    return 2
  if (column['breastcancerper100th'] > 50) & (column['breastcancerper100th'] < 75):
    return 3
  if (column['breastcancerper100th'] > 75) :
    return 4
data_clean2['cancerlevel'] = data_clean2.apply (lambda row: cancerlevel (row),axis=1)

# data frame that includes only variables that I am using
dataset1 = data_clean2[['urbanrate','cancerlevel']]

# using ols function for calculating the F-statistic and associated p-value
model1 = smf.ols(formula='urbanrate ~ C(cancerlevel)', data=dataset1).fit()
print(model1.summary())
```

employment rate as the x-axis, and breast cancer per 100th as the y-axis. The second graph had urban rate as the x-axis, and breast cancer per 100th as the y-axis.

Initial Findings: There is no correlation between female employment and breast cancer rates, in fact there is a weak negative correlation between the two. There is, however, a weak positive correlation between urban rates and breast cancer.

——-PYTHON CODE ————————————————————

```
#authored by tumblr blog mestupmxpxfan10
# library import
import pandas
import numpy
import seaborn
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
data_clean = data[(data['breastcancerper100th']>=0.01) & (data['femaleemployrate']>=0.01)]
data_clean2 = data_clean.copy()

#univariate bar graph for cancer level variable
seaborn.distplot(data_clean2["breastcancerper100th"].dropna(), kde=False);
plt.xlabel('Breast Cancer Rate per 100,000 People')
plt.title('Breast Cancer Rate')

#univariate bar graph for employment rate variables
seaborn.distplot(data_clean2["femaleemployrate"].dropna(), kde=False);
plt.xlabel('Female Employment Rate age 15 and up')
plt.title('Female Employment Rate')

#univariate bar graph for employment rate variables
seaborn.distplot(data_clean2["urbanrate"].dropna(), kde=False);
plt.xlabel('Urban Rate')
plt.title('Urban Rate')
```

Some of the first
that as cancer ra
employment rate
with it. However,
expectancy (lifee
up. How odd is th
that as cancer ris
would decrease.
hypothesis is see

Below is my
which I crea

———————
———————

```
#authored by tum
mestupmxpxfan1

# library import
import pandas as
import numpy as

# dataset import
data = pd.read_c
low_memory=Fal

# convert objects
data['breastcance
data['breastcance
bjects(convert_nu
data['femaleemplo
data['femaleemplo
s(convert_numer
data['lifeexpectan
data['lifeexpectan
onvert_numeric=

# creating of a su
quantiles, and cre
tiles1 =
data['breastcance
.25, 0.5, 0.75, 1])
tiles2 = tiles1.cop
value25perc = tile
value50perc = tile
value75perc = tile
value100perc = ti

# subsetting data
cancer25percent
data[(data['breast
value25perc)]
cancer50percent
data[(data['breast
value50perc) &
(data['breastcanc
value25perc)]
cancer75percent
data[(data['breast
value75perc) &
(data['breastcanc
value50perc)]
cancer100percen
data[(data['breast
value100perc) &
(data['breastcanc
value75perc)]
```

```
 [ 7.66956522,  10.33043478],
 [ 1.27826087,   1.72173913]]))
COMP1v6       2   4
cancerlevel
1            31   6
2            13  10
3             5   7
4             0  15
COMP1v6             2         4
cancerlevel
1            0.632653  0.157895
2            0.265306  0.263158
3            0.102041  0.184211
4            0.000000  0.394737
chi-square value, p value, expected
counts
(31.733017231260114,
5.957298782478261e-07, 3, array([[
20.83908046,  16.16091954],
 [ 12.95402299,  10.04597701],
 [  6.75862069,   5.24137931],
 [  8.44827586,   6.55172414]]))
COMP1v7       3   4
cancerlevel
1            17   6
2            33  10
3            13   7
4             3  15
COMP1v7             3         4
cancerlevel
1            0.257576  0.157895
2            0.500000  0.263158
3            0.196970  0.184211
4            0.045455  0.394737
chi-square value, p value, expected
counts
(21.374034958708471,
8.8028672912121767e-05, 3, array([[
14.59615385,   8.40384615],
 [ 27.28846154,  15.71153846],
 [ 12.69230769,   7.30769231],
 [ 11.42307692,   6.57692308]]))
>>>
```

—— CODE ——

```
#authored by tumblr blog
mestupmxpxfan10
# library import
import pandas
import numpy
import scipy.stats
import seaborn
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_o
bjects(convert_numeric=True)
data['femaleemployrate'] =
data['femaleemployrate'].convert_object
s(convert_numeric=True)
```

```
#means and standard deviation,
compared
m1 =
dataset1.groupby('cancerlevel').mean()
sd1 =
dataset1.groupby('cancerlevel').std()
print(m1)
print (sd1)

#Multi-Comparison using tukey's
honestly significant difference
mc1 =
multi.MultiComparison(dataset1['urbanr
ate'], dataset1['cancerlevel'])
res1 = mc1.tukeyhsd()
print(res1.summary())
```

My Research Project

I have chosen to use the GapMinder Dataset. After reviewing the code book, I was immediately drawn to learn more about breast cancer rates. As, I looked further into the data, I wondered if female employment rate would be associated with these rates. I wondered this because I have a suspicion that western culture has a higher propensity at receiving breast cancer, when compared to the rest of the world, and that the working conditions in industrialized societies may produce a greater likelihood towards breast cancer in women.

After doing research on this topic (querying google scholar on breast cancer and employment rates), I believe that employment rate will reflect a person's socio-economic condition, and that a person's socio-economic condition will be likely tied into their breast cancer rate.

Sources for hypothesis:

"American Journal of Epidemiology." *SOCIAL CLASS AND THE BLACK/WHITE CROSSOVER IN THE AGE-SPECIFIC INCIDENCE OF BREAST CANCER: A STUDY LINKING CENSUS-DERIVED DATA TO POPULATION-BASED REGISTRY RECORDS*. Web. 11 Jan. 2016.

"International Agency for Research on Cancer" Social Inequalities of Cancer Web. 11 Jan. 2016.


```
#bivariate measuring if female
employment affects breast cancer
scat1 =
seaborn.regplot(x="femaleemployrate",
y="breastcancerper100th",data=data)
plt.xlabel('Female Employment Rate')
plt.ylabel('Breast Cancer per 100
Thousand')
plt.title('Scatterplot for the Association
Between Female Employment and
Breast Cancer Rate')

#bivariate measuring if urban rate
affects breast cancer
scat2 = seaborn.regplot(x="urbanrate",
y="breastcancerper100th", data=data)
plt.xlabel('Urban Rate')
plt.ylabel('Breast Cancer per 100
Thousand')
plt.title('Scatterplot for the Association
Between Urban Rate and Breast
Cancer Rate')
```

```
# mean cancer ra
meancancerrate2
cancer25percent
0th'].mean()
meancancerrate5
cancer50percent
0th'].mean()
meancancerrate7
cancer75percent
0th'].mean()
meancancerrate1
cancer100percer
00th'].mean()

# mean empmloy
percent
meanemployrate2
cancer25percent
mean()
meanemployrate5
cancer50percent
mean()
meanemployrate7
cancer75percent
mean()
meanemployrate1
cancer100percer
].mean()

# mean life expec
cancer percent
meanlifeexpec25
cancer25percent
an()
meanlifeexpec50
cancer50percent
an()
meanlifeexpec75
cancer75percent
an()
meanlifeexpec10
cancer100percer
ean()

# creating a datas
that i took
mean_d = {'breas
pd.Series([meanc
meancancerrate5
meancancerrate7
meancancerrate1
['0.25', '0.5', '0.75
          'femalee
pd.Series([meane
meanemployrate5
meanemployrate7
meanemployrate1
['0.25', '0.5', '0.75
          'lifeexpe
pd.Series([meanli
meanlifeexpec50
meanlifeexpec75
meanlifeexpec10
['0.25', '0.5', '0.75
mean_df = pd.Da

print(mean_df)
```

```
data['urbanrate'] =
data['urbanrate'].convert_objects(conve
rt_numeric=True)

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

#new variables, based off rate or per
100,000 value. This is categorical 1-4
Value
#Value Index:  1=0-24.99, 2=25-49.99,
3=50-74.99 4=75+
def cancerlevel (column):
 if (column['breastcancerper100th'] <
25):
    return 1
 if (column['breastcancerper100th'] >=
25) & (column['breastcancerper100th']
< 50):
    return 2
 if (column['breastcancerper100th'] >=
50) & (column['breastcancerper100th']
< 75):
    return 3
 if (column['breastcancerper100th'] >=
75) :
    return 4
data_clean2['cancerlevel'] =
data_clean2.apply (lambda row:
cancerlevel (row),axis=1)

def urbanlevel (column):
 if (column['urbanrate'] < 25):
    return 1
 if (column['urbanrate'] >= 25) &
(column['urbanrate'] < 50):
    return 2
 if (column['urbanrate'] >= 50) &
(column['urbanrate'] < 75):
    return 3
 if (column['urbanrate'] >= 75) :
    return 4
data_clean2['urbanlevel'] =
data_clean2.apply (lambda row:
urbanlevel (row),axis=1)

# contingency table of observed counts
ct1=pandas.crosstab(data_clean2['can
cerlevel'], data_clean2['urbanlevel'])
print (ct1)

# column percentages
colsum=ct1.sum(axis=0)
colpct=ct1/colsum
print(colpct)

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

# graph percent with nicotine
dependence within each smoking
frequency group
```

```
seaborn.factorplot(x='urbanlevel',
y='cancerlevel', data=data_clean2,
kind="bar", ci=None)
plt.xlabel('Urban Rate')
plt.ylabel('Breast Cancer Level')

#post-hoc
recode2 = {1: 1, 2: 2}
data_clean2['COMP1v2']=
data_clean2['urbanlevel'].map(recode2)

# contingency table of observed counts
ct2=pandas.crosstab(data_clean2['can
cerlevel'], data_clean2['COMP1v2'])
print (ct2)

# column percentages
colsum2=ct2.sum(axis=0)
colpct2=ct2/colsum2
print(colpct2)

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

recode3 = {1: 1, 3: 3}
data_clean2['COMP1v3']=
data_clean2['urbanlevel'].map(recode3)

# contingency table of observed counts
ct3=pandas.crosstab(data_clean2['can
cerlevel'], data_clean2['COMP1v3'])
print (ct3)
# column percentages
colsum3=ct3.sum(axis=0)
colpct3=ct3/colsum3
print(colpct3)

print ('chi-square value, p value,
expected counts')
cs3= scipy.stats.chi2_contingency(ct3)
print (cs3)

recode4 = {1: 1, 4: 4}
data_clean2['COMP1v4']=
data_clean2['urbanlevel'].map(recode4)

# contingency table of observed counts
ct4=pandas.crosstab(data_clean2['can
cerlevel'], data_clean2['COMP1v4'])
print (ct4)

# column percentages
colsum4=ct4.sum(axis=0)
colpct4=ct4/colsum4
print(colpct4)

print ('chi-square value, p value,
expected counts')
cs4= scipy.stats.chi2_contingency(ct4)
print (cs4)

recode5 = {2: 2, 3: 3}
data_clean2['COMP1v5']=
data_clean2['urbanlevel'].map(recode5)

# contingency table of observed counts
ct5=pandas.crosstab(data_clean2['can
cerlevel'], data_clean2['COMP1v5'])
print (ct5)
# column percentages
```

```
colsum5=ct5.sum(axis=0)
colpct5=ct5/colsum5
print(colpct5)

print ('chi-square value, p value,
expected counts')
cs5= scipy.stats.chi2_contingency(ct5)
print (cs5)

recode6 = {2: 2, 4: 4}
data_clean2['COMP1v6']=
data_clean2['urbanlevel'].map(recode6)

# contingency table of observed counts
ct6=pandas.crosstab(data_clean2['can
cerlevel'], data_clean2['COMP1v6'])
print (ct6)

# column percentages
colsum6=ct6.sum(axis=0)
colpct6=ct6/colsum6
print(colpct6)

print ('chi-square value, p value,
expected counts')
cs6= scipy.stats.chi2_contingency(ct6)
print (cs6)

recode7 = {3: 3, 4: 4}
data_clean2['COMP1v7']=
data_clean2['urbanlevel'].map(recode7)

# contingency table of observed counts
ct7=pandas.crosstab(data_clean2['can
cerlevel'], data_clean2['COMP1v7'])
print (ct7)

# column percentages
colsum7=ct7.sum(axis=0)
colpct7=ct7/colsum7
print(colpct7)

print ('chi-square value, p value,
expected counts')
cs7= scipy.stats.chi2_contingency(ct7)
print (cs7)
```

Show more