Testing a Potential Moderator (Confounder) in between the response variable life expectancy and the explanatory variable internet user rate for different countries from the GapMinder dataset

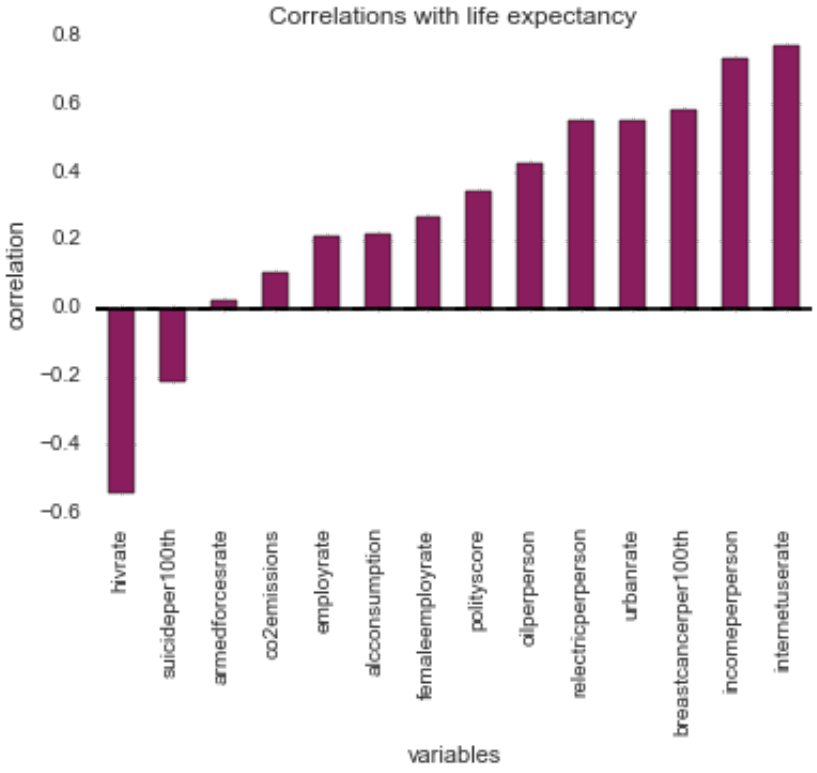
First, a *correlation analysis* was conducted on the *GapMinder* dataset to understand the association of 14 explanatory variables (including income per person, alcohol consumption, armed forces rate, breast cancer per 100th, co2 emissions, female employment rate, hiv rate, internet use rate, oil per person, polity score, relectric per person, suicide per 100th, employment rate, urbanization rate) with the variable *life expectancy*.

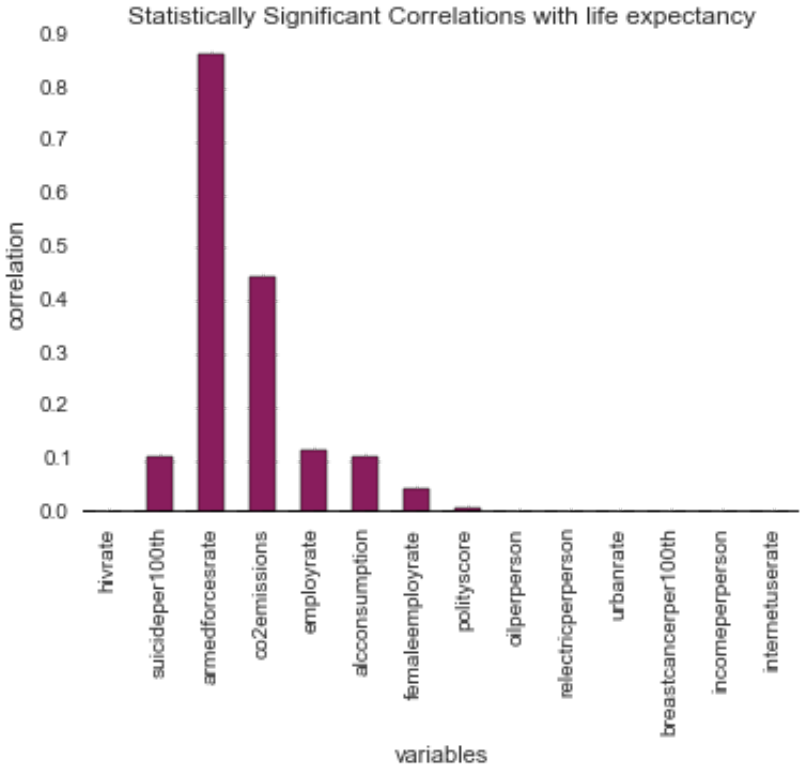
After removing the obeservations with missing values the pearson correlation coefficient is computed. As can be seen from the below results, the variable *internetuserate* has a strong positive correlation (with correlation coefficient ~ 0.77 and significant at 1% level, with a very low p-value) with the variable *life expectancy*.

Out[17]:

	variables	pearson-r	p-value
0	hivrate	-0.542506	1.566318e-05
0	suicideper100th	-0.218335	1.059663e-01
0	armedforcesrate	0.023648	8.626540e-01
0	co2emissions	0.103990	4.456349e-01
0	employrate	0.210334	1.197189e-01
0	alcconsumption	0.218541	1.056298e-01
0	femaleemployrate	0.268129	4.571763e-02
0	polityscore	0.344843	9.248381e-03
0	oilperperson	0.422911	1.165352e-03
0	relectricperperson	0.551581	1.052532e-05
0	urbanrate	0.552084	1.029253e-05
0	breastcancerper100th	0.580247	2.769328e-06
0	incomeperperson	0.732452	1.400123e-10
0	internetuserate	0.769160	4.381504e-12

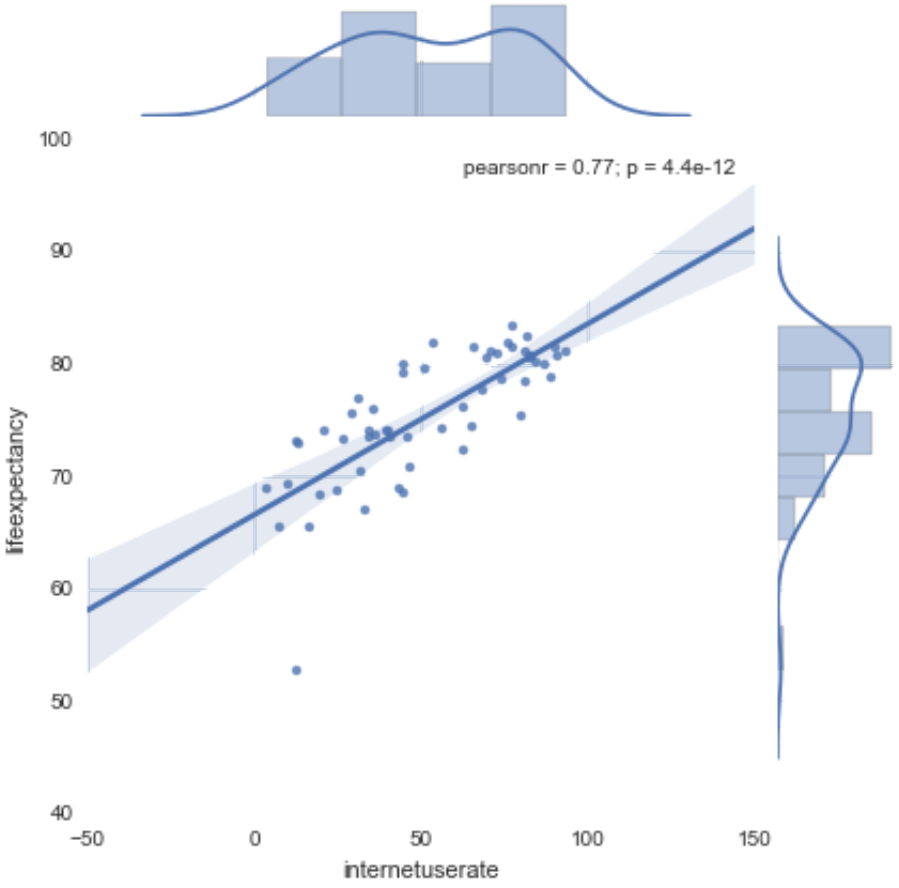
<matplotlib.figure.Figure at 0x61e85d0>


Correlations with life expectancy
correlation
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
hivrate
suicideper100th
armedforcesrate
co2emissions
employrate
alcconsumption
femaleemployrate
polityscore
oilperperson
relectricperperson
urbanrate
breastcancerper100th
incomeperperson
internetuserate
variables


Statistically Significant Correlations with life expectancy
correlation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
hivrate
suicideper100th
armedforcesrate
co2emissions
employrate
alcconsumption
femaleemployrate
polityscore
oilperperson
relectricperperson
urbanrate
breastcancerper100th
incomeperperson
internetuserate
variables

```
Out[5]:
```

```
<seaborn.axisgrid.JointGrid at 0x301ee70>
```



Analysis 1 (with correlation coefficient)

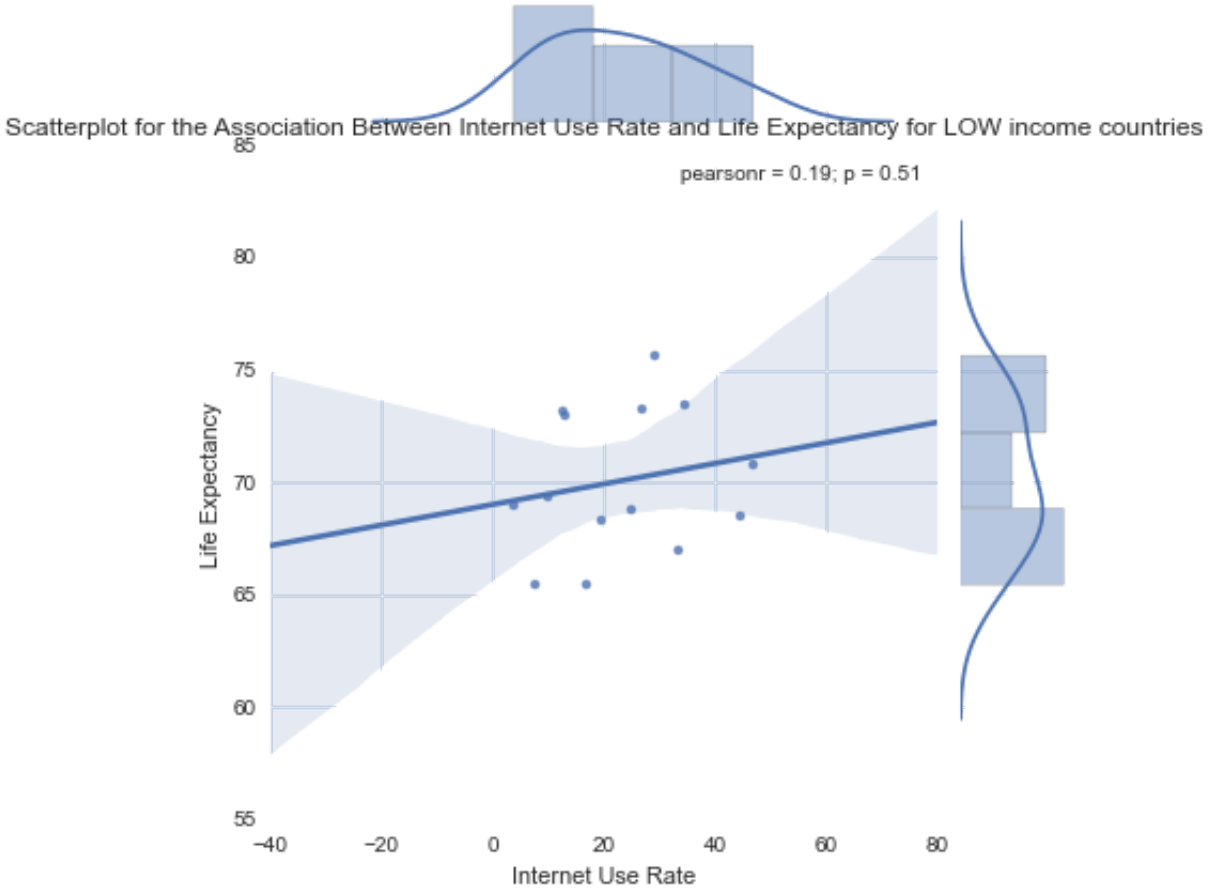
We want to answer the following question: does the variable *income per person* **moderate** the relationship between the qunatitative explanatory variable *internet use rate* and the quantitative response variable *life expectancy*? To answer this, the income per person variable was partitioned into **4 groups** (LOW, MID, HIGH, VERY HIGH) using the quartiles as shown below, after removing the NA values.

```
count       56.000000
mean     12982.654643
std      12712.681024
min        558.062877
25%       2532.598585
50%       6219.692968
75%      25373.478550
max      39972.352768
Name: incomeperperson, dtype: float64

association between internetuserate and lifeexpectancy for LOW income countries
(0.19318865715608682, 0.50814385277502172)

association between internetuserate and lifeexpectancy for MIDDLE income countr
ies
(0.49179543859770558, 0.07407099884980306)

association between internetuserate and lifeexpectancy for HIGH income countrie
s
(0.26833836982846088, 0.35361855231884515)

association between internetuserate and lifeexpectancy for VERY HIGH income cou
ntries
```

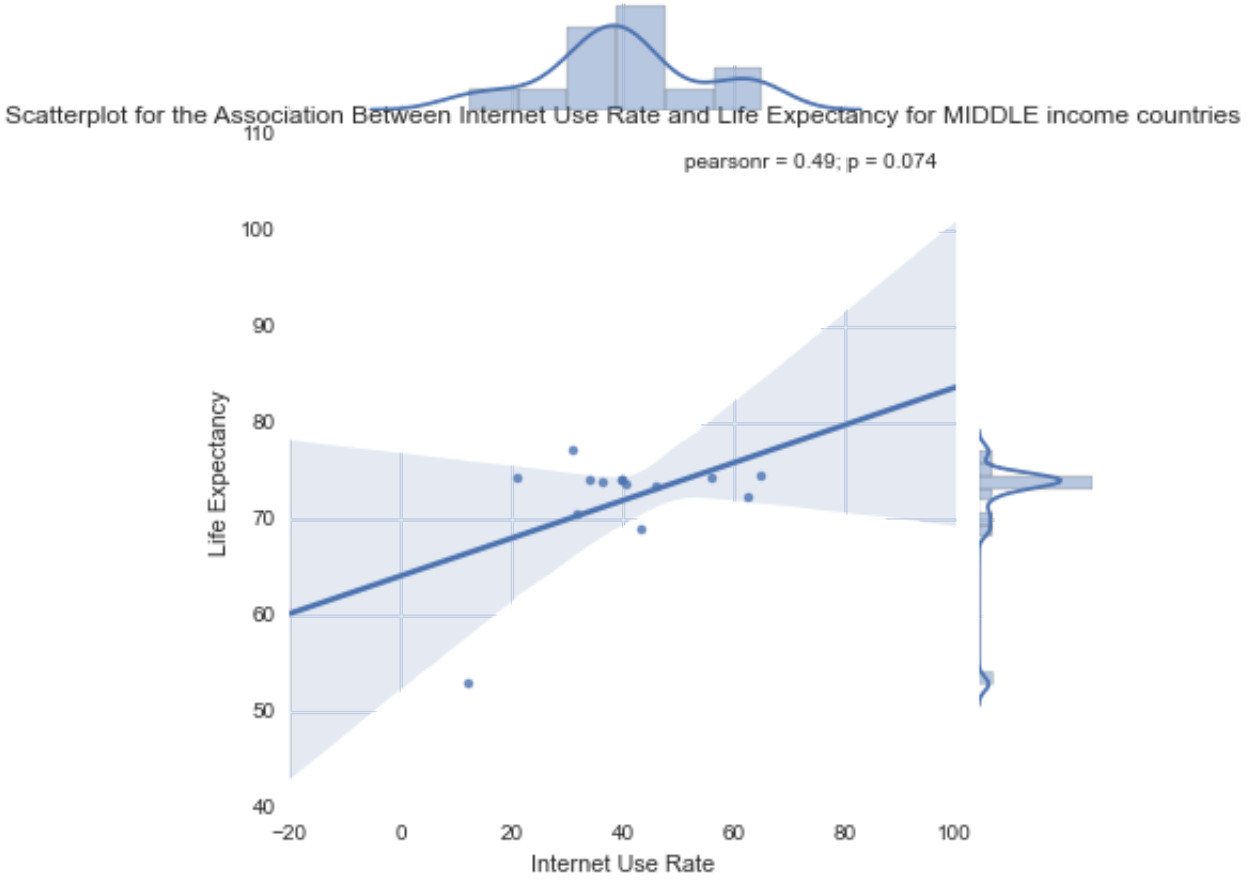
```
C:\Python27\lib\site-packages\IPython\kernel\__main__.py:13: SettingWithCopyWarnin
g:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the the caveats in the documentation: http://pandas.pydata.org/pandas-docs/sta
ble/indexing.html#indexing-view-versus-copy
```

As can be seen from the below scatterplots for different income groups, the income groups LOW, MID and HIGH appears to have somewhat positive corrletaion in between the variables *internet user rate* and *life expectancy* (altough none of them statisticallly significant at 5% level), but the income group VERY HIGH almost have no correlation (very weak negative correlation). It implies that the variable *income per person* moderates the relation in between *internet user rate* and *life expectancy* significantly.

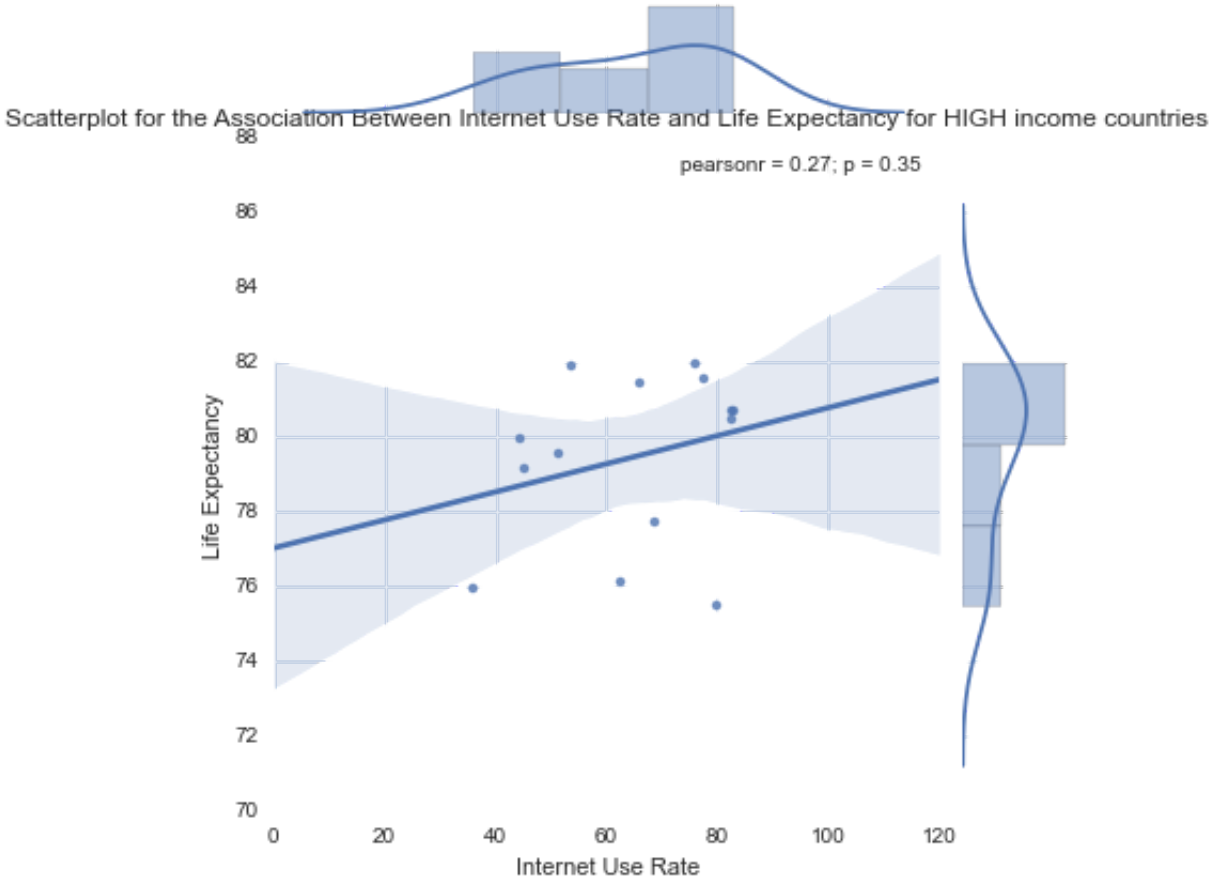
```
<seaborn.axisgrid.JointGrid object at 0x068F5E30>
```



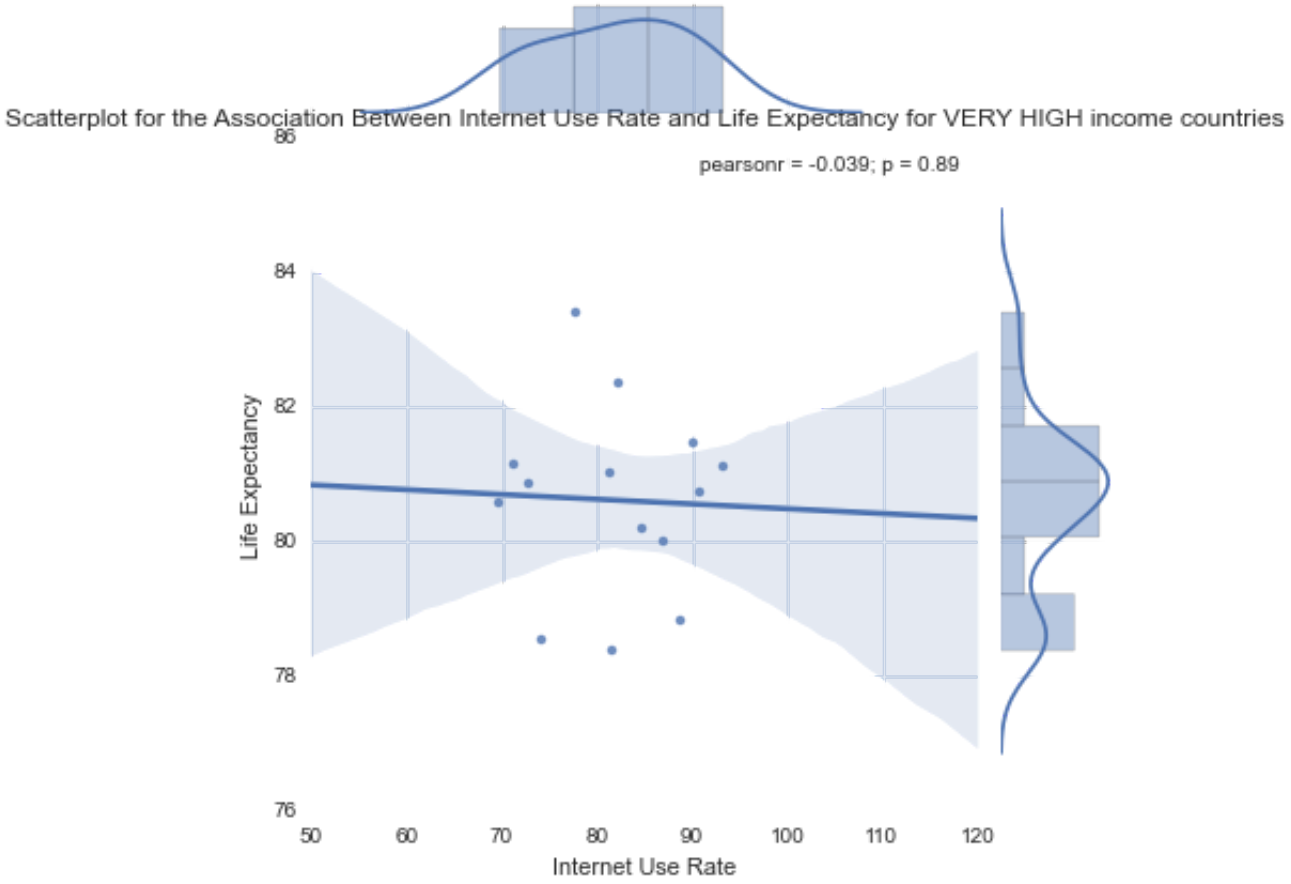
```
<seaborn.axisgrid.JointGrid object at 0x065FA530>
```



```
<seaborn.axisgrid.JointGrid object at 0x07F38C90>
```



```
<seaborn.axisgrid.JointGrid object at 0x0817A7F0>
```



Analysis 2 (with one-way ANOVA)

First, let's convert the quantitative variable *internet use rate* to a categoriecal variable by dividing it into **2 groups**: 1=LOW (below 50%) and 2=HIGH (otherwise). First we shall use one-way ANOVA to find whether the association between the categorical explanatory variable *internet use rate* and the quantitative response variable *life expectancy* is significant and also compute mean *life expectancy* for the *internet use rate* groups. As can be seen, the relationship in between these two variables was found to be statistically significant at 5% level (with low *p-value* and *F-statistic* 52.11).

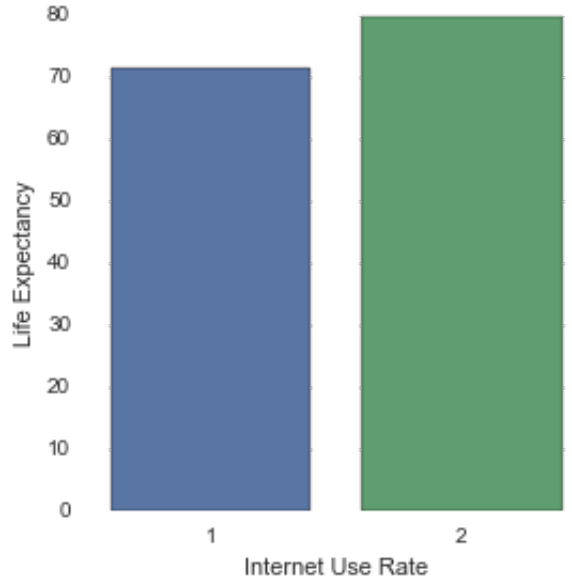
```
count    56.000000
mean     52.464245
std      26.218205
min       3.700003
25%      33.049632
50%      48.980090
75%      77.533598
max      93.277508
Name: internetuserate, dtype: float64
```

```
C:\Python27\lib\site-packages\IPython\kernel\__main__.py:6: SettingWithCopyWarnin
g:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the the caveats in the documentation: http://pandas.pydata.org/pandas-docs/sta
ble/indexing.html#indexing-view-versus-copy
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         lifeexpectancy   R-squared:                       0.491
Model:                            OLS   Adj. R-squared:                  0.482
Method:                 Least Squares   F-statistic:                     52.11
Date:                Mon, 29 Feb 2016   Prob (F-statistic):           1.82e-09
Time:                        17:41:03   Log-Likelihood:                -158.36
No. Observations:                  56   AIC:                             320.7
Df Residuals:                      54   BIC:                             324.8
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
=========
                               coef    std err          t      P>|t|      [95.0% Co
nf. Int.]
--------------------------------------------------------------------------------
---------
Intercept                   71.4784      0.787     90.777      0.000        69.900
73.057
C(internetuserate)[T.2]     8.0382      1.114      7.218      0.000         5.806
10.271
==============================================================================
Omnibus:                       34.166   Durbin-Watson:                   1.801
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              108.341
Skew:                          -1.649   Prob(JB):                     2.98e-24
Kurtosis:                       8.962   Cond. No.                         2.62
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
pecified.
means for lifeexpectancy by internetuserate LOW vs. HIGH
                 lifeexpectancy
internetuserate
1                     71.478357
2                     79.516536
standard deviation for mean lifeexpectancy by internetuserate LOW vs. HIGH
                 lifeexpectancy
internetuserate
1                      5.217157
2                      2.738937
```

Out[12]:

<matplotlib.text.Text at 0x871c790>



Now, we want to answer the following question: does the variable *income per person* **moderate** the relationship between the (converted) categorical explanatory variable *internet use rate* and the quantitative response variable *life expectancy*? To answer this, the income per person variable was partitioned into **2 groups** (LOW=1, HIGH=2) using the median, after removing the NA values. Then one-way ANOVA test (and Tukey post-hoc test) was conducted on each of the groups. As can be seen, the association between the categorical explanatory variable *internet use rate* and the quantitative response variable *life expectancy* is not significant inside both the income groups, which implies that the variable *income per person* **moderates** the relationship.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         lifeexpectancy   R-squared:                       0.040
Model:                            OLS   Adj. R-squared:                  0.003
Method:                 Least Squares   F-statistic:                     1.070
Date:                Mon, 29 Feb 2016   Prob (F-statistic):              0.311
Time:                        17:41:04   Log-Likelihood:                -81.965
No. Observations:                  28   AIC:                             167.9
Df Residuals:                      26   BIC:                             170.6
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
=========
                              coef    std err          t      P>|t|      [95.0% Co
nf. Int.]
-------------------------------------------------------------------------------
---------
Intercept                   70.6583      0.938     75.330      0.000        68.730
72.586
C(internetuserate)[T.2]     2.9637      2.866      1.034      0.311        -2.927
8.854
==============================================================================
Omnibus:                       28.841   Durbin-Watson:                   2.106
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               64.278
Skew:                          -2.095   Prob(JB):                     1.10e-14
Kurtosis:                       9.127   Cond. No.                         3.27
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
pecified.
means for lifeexpectancy by incomegrp
                lifeexpectancy
internetuserate
1                    70.65832
2                    73.62200
standard deviations for lifeexpectancy by incomegrp
                lifeexpectancy
internetuserate
1                    4.868961
2                    1.208500

C:\Python27\lib\site-packages\IPython\kernel\__main__.py:7: SettingWithCopyWarnin
g:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the the caveats in the documentation: http://pandas.pydata.org/pandas-docs/sta
ble/indexing.html#indexing-view-versus-copy
```

```
Multiple Comparison of Means - Tukey HSD,FWER=0.05
============================================
group1 group2 meandiff  lower  upper reject
--------------------------------------------
  1      2     2.9637  -2.9266 8.854 False
--------------------------------------------


                            OLS Regression Results
==============================================================================
Dep. Variable:         lifeexpectancy   R-squared:                       0.096
Model:                            OLS   Adj. R-squared:                  0.062
Method:                 Least Squares   F-statistic:                     2.769
Date:                Mon, 29 Feb 2016   Prob (F-statistic):              0.108
Time:                        17:41:38   Log-Likelihood:                -56.372
No. Observations:                  28   AIC:                             116.7
Df Residuals:                      26   BIC:                             119.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=========
                               coef    std err          t      P>|t|      [95.0% Co
nf. Int.]
------------------------------------------------------------------------------
---------
Intercept                   78.3120      1.086     72.140      0.000        76.081
80.543
C(internetuserate)[T.2]     1.9119      1.149      1.664      0.108        -0.450
4.273
==============================================================================
Omnibus:                        6.132   Durbin-Watson:                   1.869
Prob(Omnibus):                  0.047   Jarque-Bera (JB):                4.559
Skew:                          -0.957   Prob(JB):                        0.102
Kurtosis:                       3.495   Cond. No.                         5.95
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
pecified.
means for lifeexpectancy by incomegrp
                 lifeexpectancy
internetuserate
1                     78.31200
2                     80.22388
standard deviations for lifeexpectancy by incomegrp
                 lifeexpectancy
internetuserate
1                     2.125487
2                     1.858336


Multiple Comparison of Means - Tukey HSD,FWER=0.05
============================================
group1 group2 meandiff  lower  upper  reject
--------------------------------------------
  1      2     1.9119  -0.4496 4.2734 False
--------------------------------------------
```