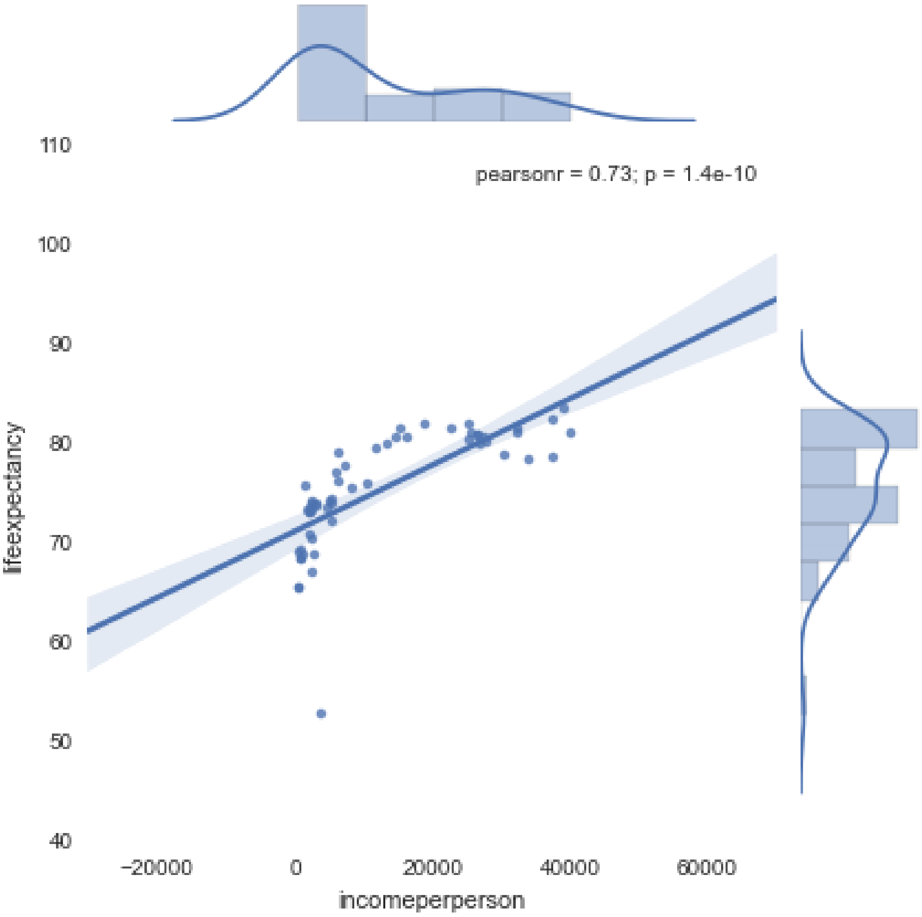
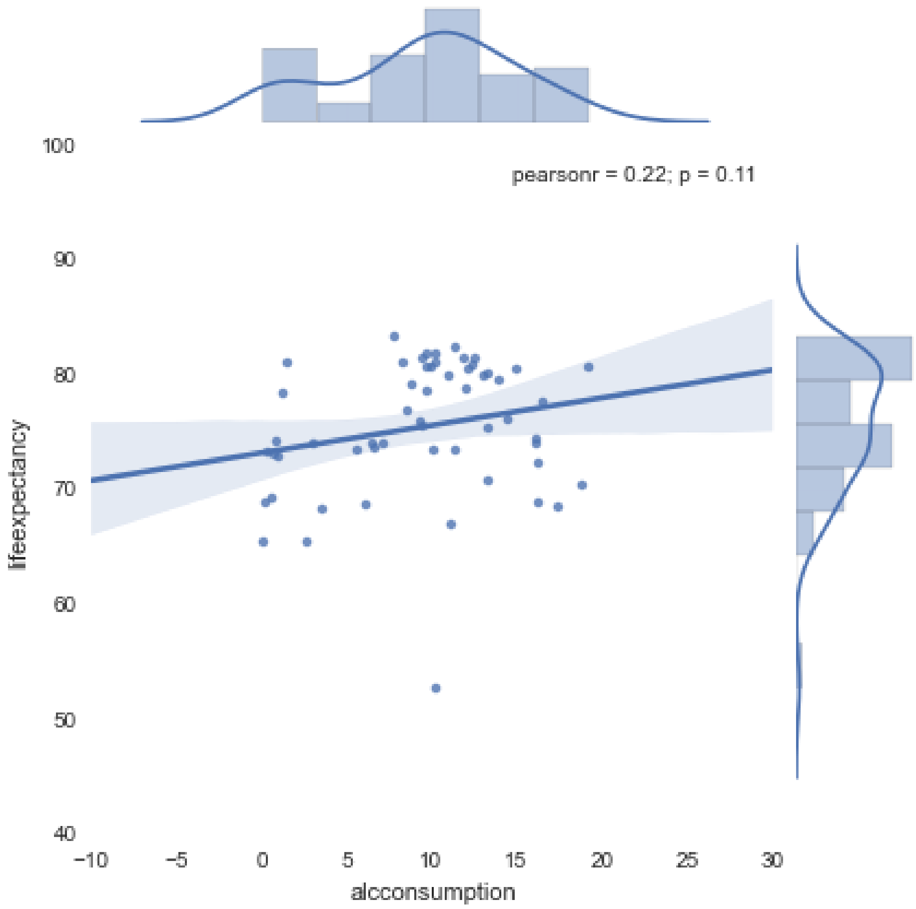
Association of life expectancy with other explanatory variables for different countries from the GapMinder dataset

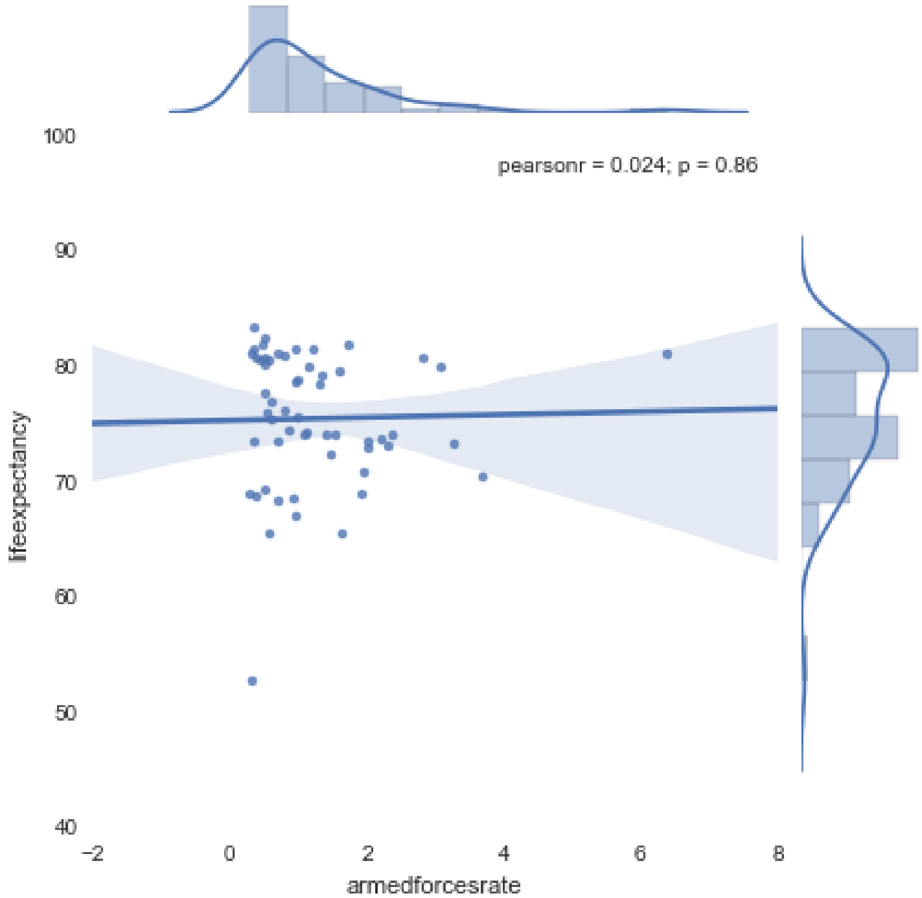
A **correlation analysis** was conducted on the GapMinder dataset to understand the association of 14 explanatory variables (including income per person, alcohol consumption, armed forces rate, breast cancer per 100th, co2 emissions, female employment rate, hiv rate, internet use rate, oil per person,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

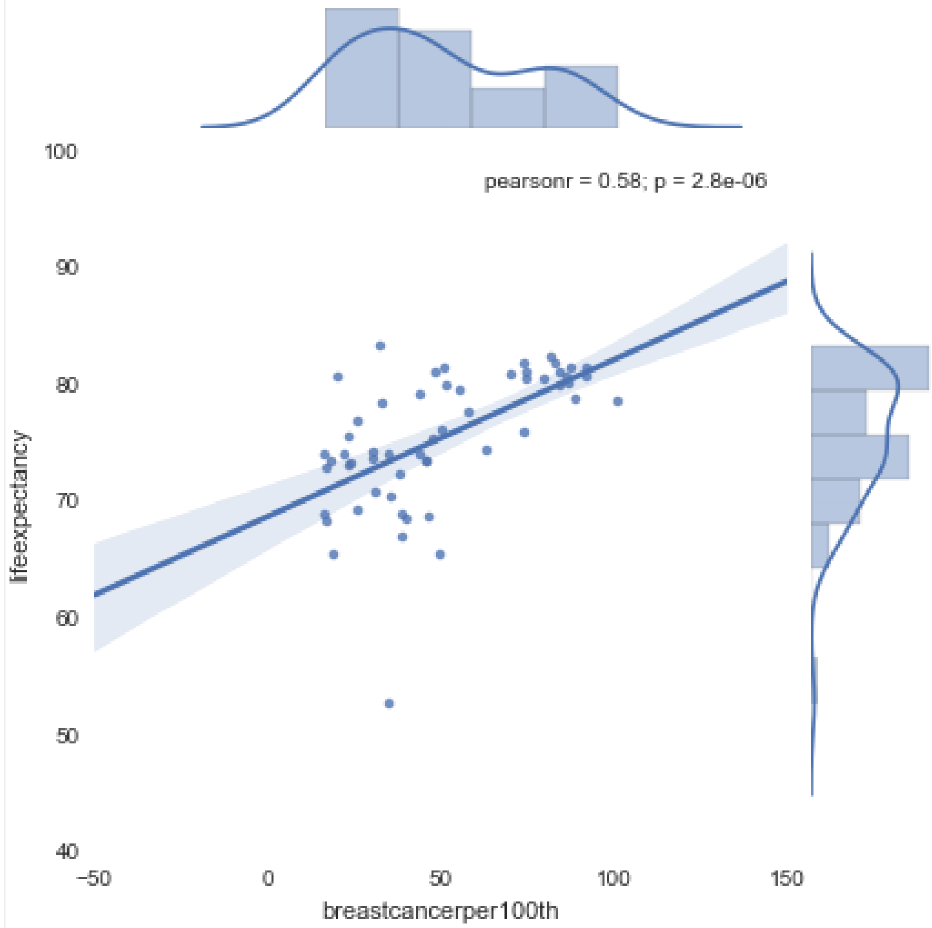
After removing the observations with missing values the **pearson correlation coefficient** is computed. As can be seen from the below results, the variable *internetuserate* has a very strong positive correlation with the variable *life expectancy*. The variable *incomeperperson* also has a strong positive correlation with *life expectancy*. The variable *hivrate* is the variable most negatively associated with the variable life expectancy. The variable *armedforcesrate* has the least correlation with life expectancy. Also, the corresponding p-values (with the null hypothesis that the variables are not correlated) are reported. All the variables except *armedforcesrate* and *co2emissions* have **statistically significant correlations** at **5% level of significance**.

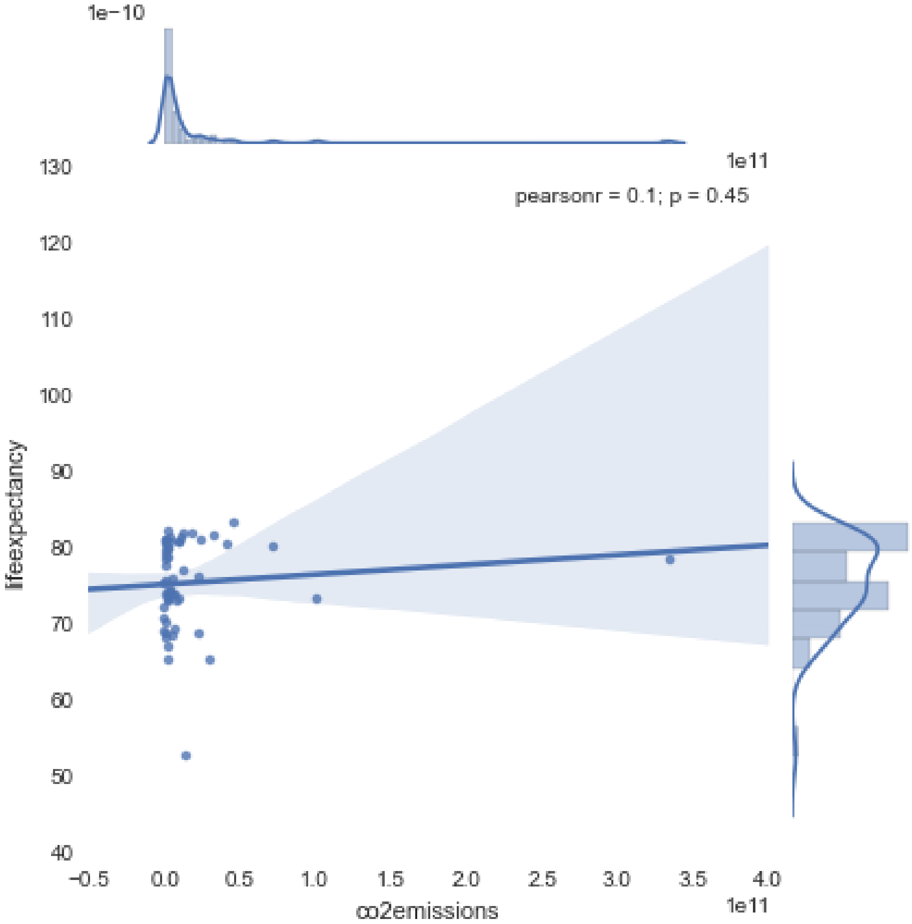
variables	pearson-r	p-value
hivrate	-0.542506	1.566318e-05
suicideper100th	-0.218335	1.059663e-01
armedforcesrate	0.023648	8.626540e-01
co2emissions	0.103990	4.456349e-01
employrate	0.210334	1.197189e-01
alcconsumption	0.218541	1.056298e-01
femaleemployrate	0.268129	4.571763e-02
polityscore	0.344843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

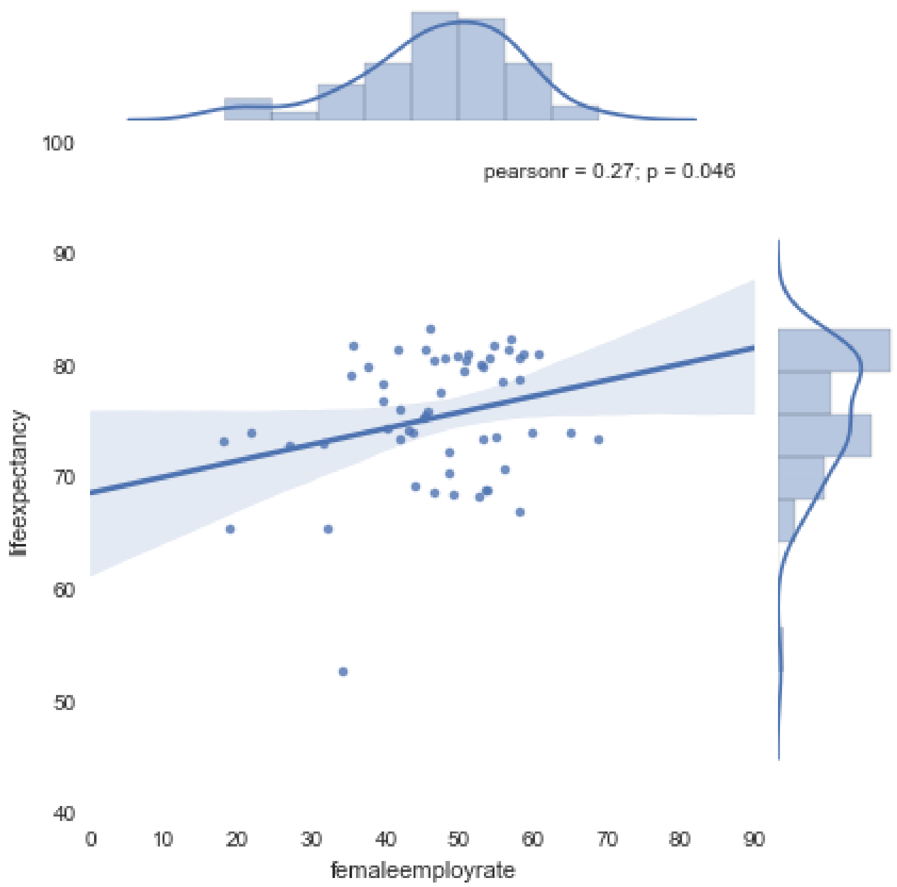

pearsonr = 0.73; p = 1.4e-10
lifeexpectancy
incomeperperson

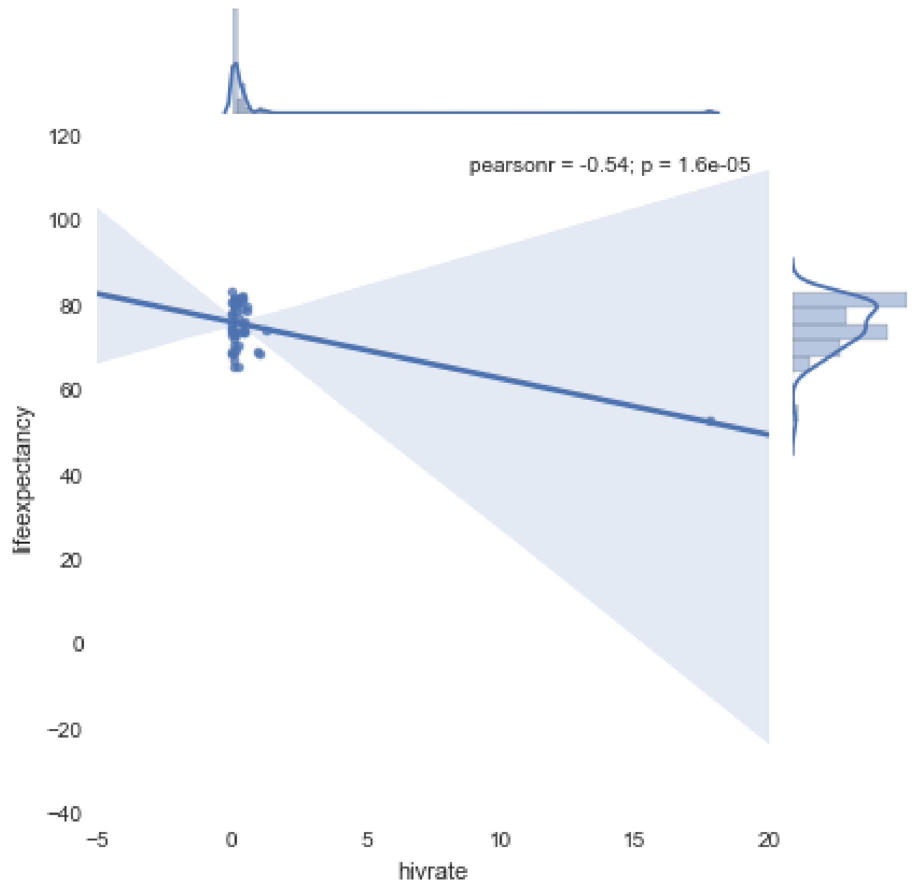

pearsonr = 0.22; p = 0.11
lifeexpectancy
alcconsumption


pearsonr = 0.024; p = 0.86
lifeexpectancy
armedforcesrate
100
90
80
70
60
50
40
−2
0
2
4
6
8

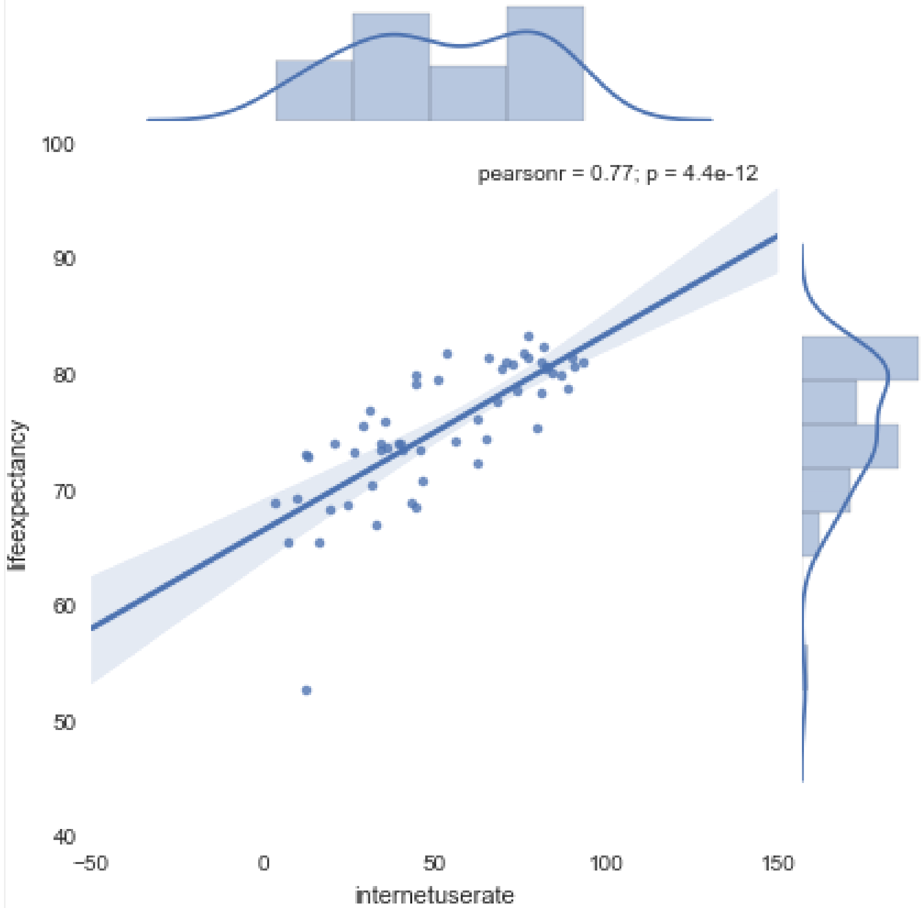

pearsonr = 0.58; p = 2.8e-06
lifeexpectancy
breastcancerper100th
100
90
80
70
60
50
40
−50
0
50
100
150


1e−10
1e11
pearsonr = 0.1; p = 0.45
130
120
110
100
90
80
70
60
50
40
lifeexpectancy
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
co2emissions
1e11

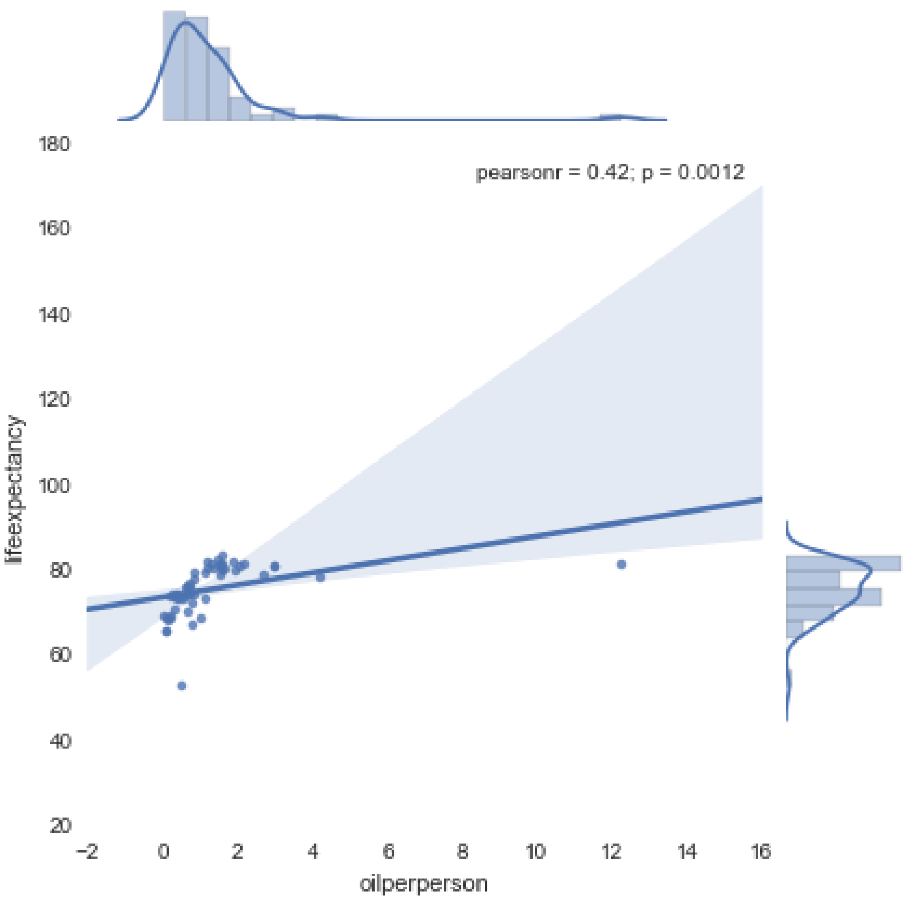

pearsonr = 0.27; p = 0.046
100
90
80
70
60
50
40
lifeexpectancy
0
10
20
30
40
50
60
70
80
90
femaleemployrate

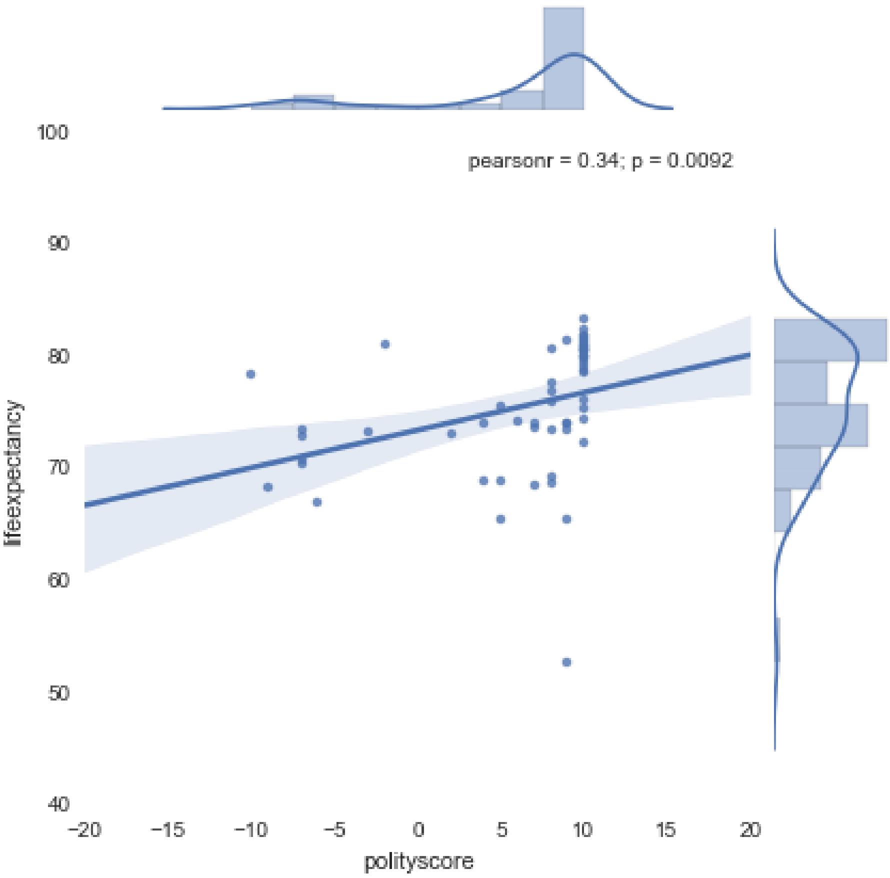


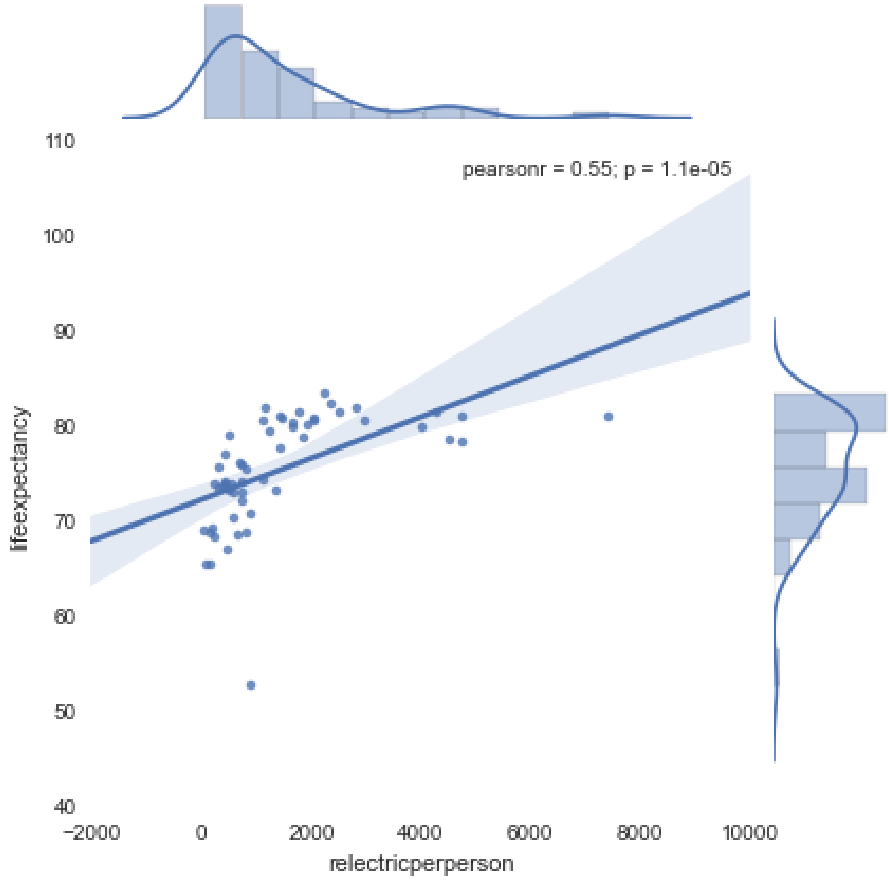
pearsonr = -0.54; p = 1.6e-05
lifeexpectancy
hivrate
120
100
80
60
40
20
0
−20
−40
−5
0
5
10
15
20

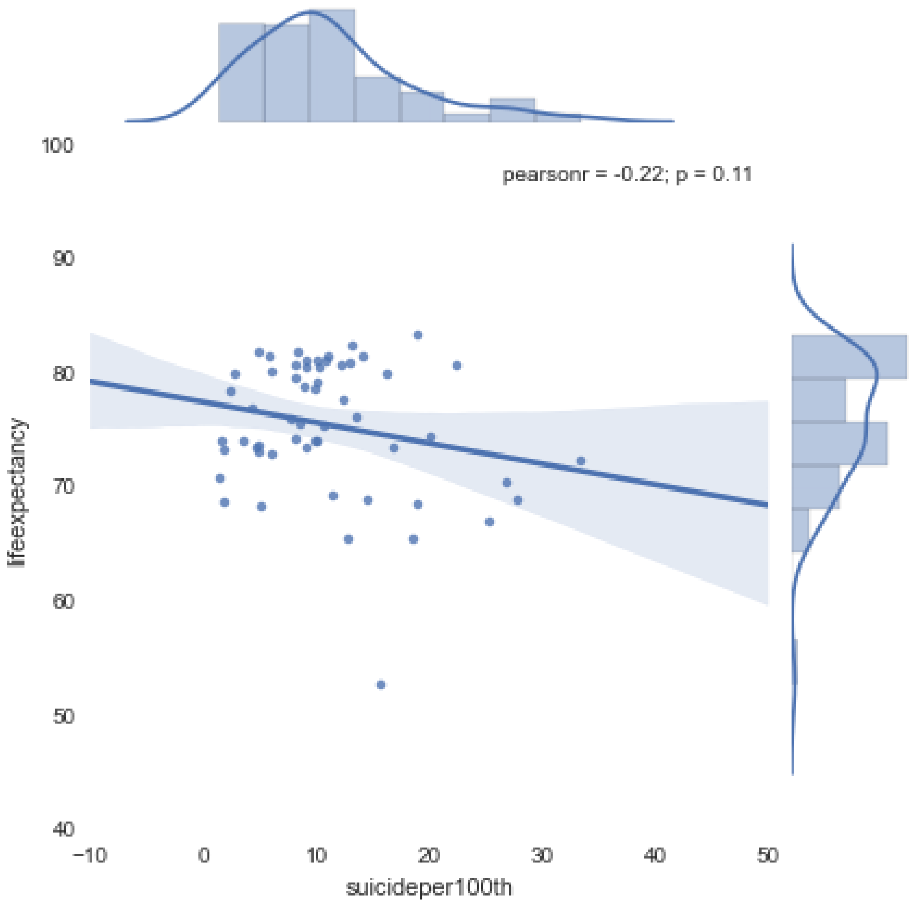


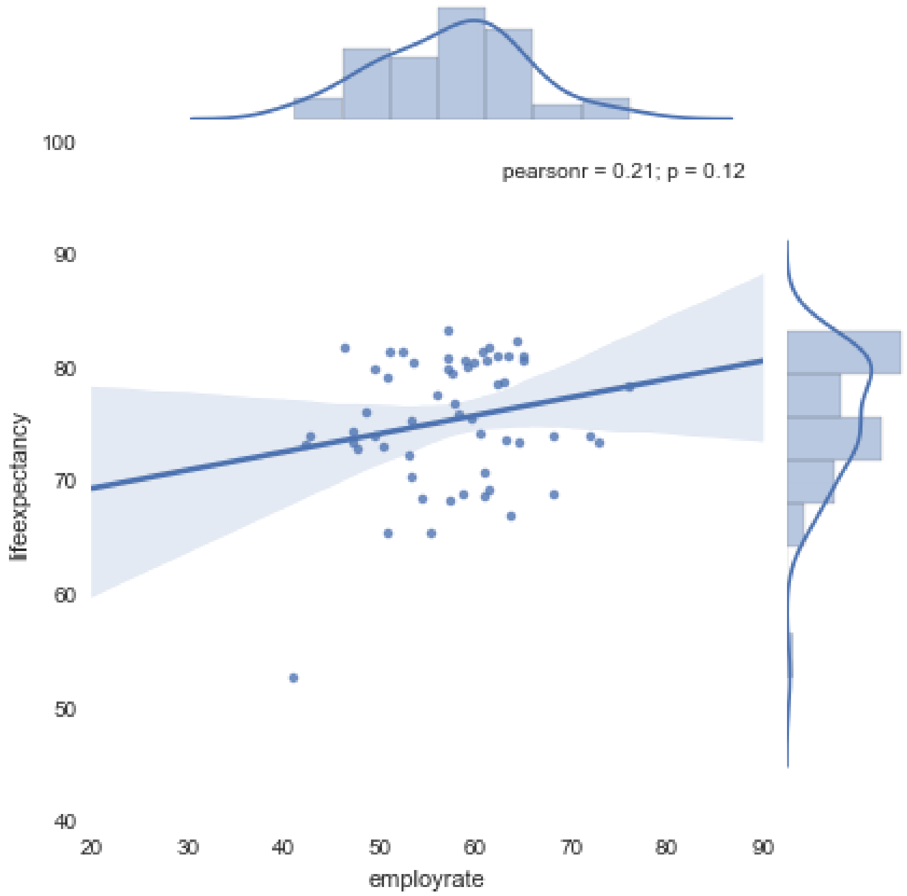
pearsonr = 0.77; p = 4.4e-12
lifeexpectancy
internetuserate
100
90
80
70
60
50
40
−50
0
50
100
150

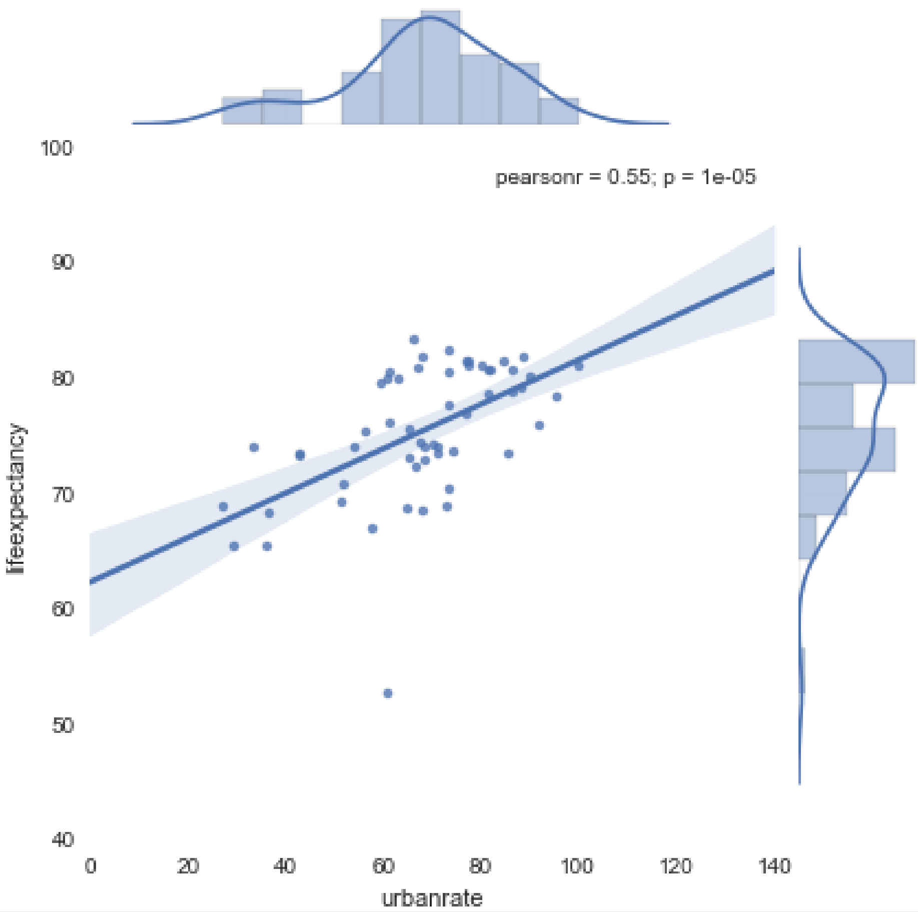

pearsonr = 0.42; p = 0.0012
lifeexpectancy
oilperperson

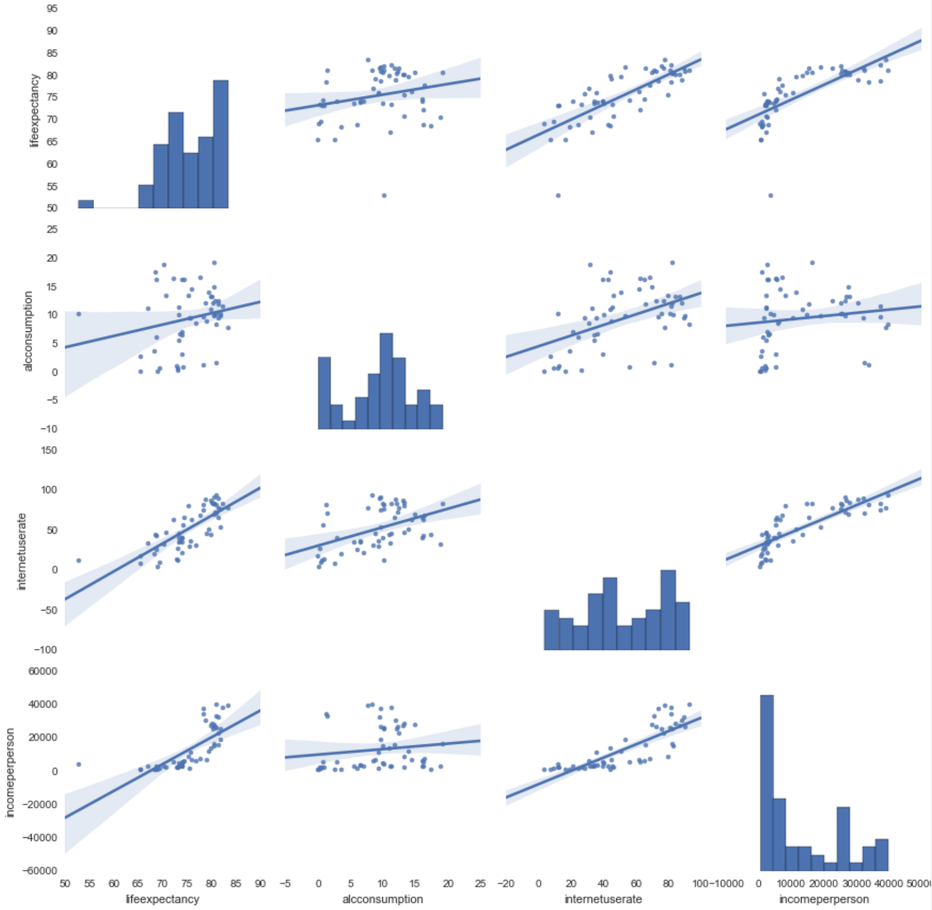

pearsonr = 0.34; p = 0.0092
lifeexpectancy
polityscore

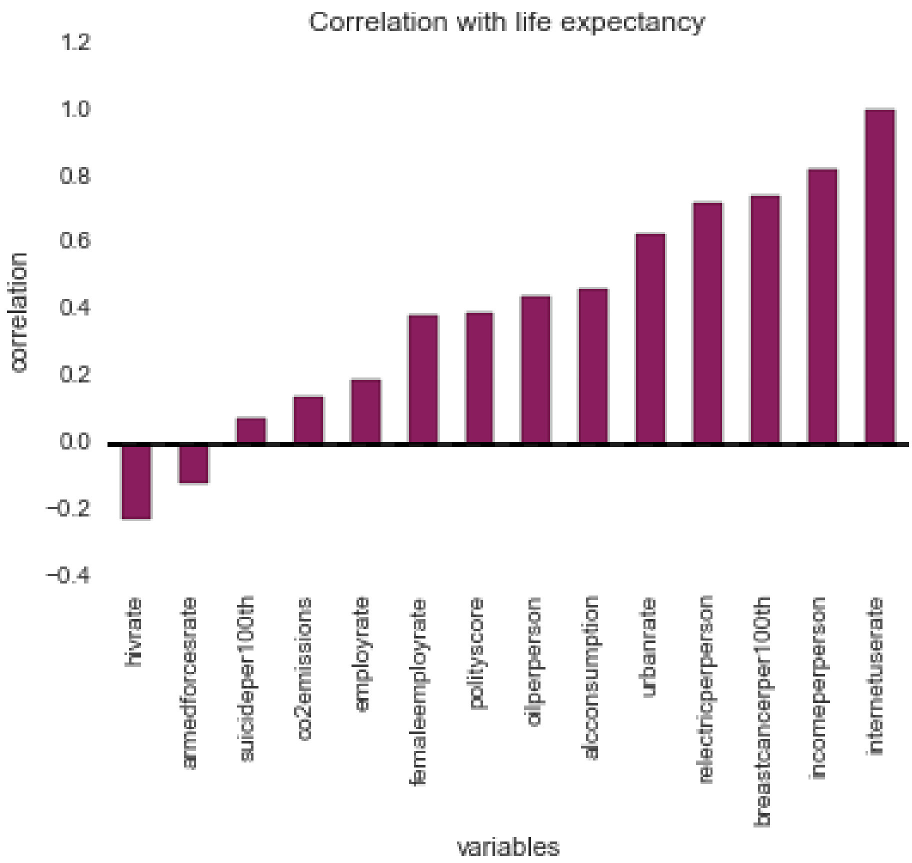

pearsonr = 0.55; p = 1.1e-05
lifeexpectancy
relectricperperson
110
100
90
80
70
60
50
40
−2000
0
2000
4000
6000
8000
10000

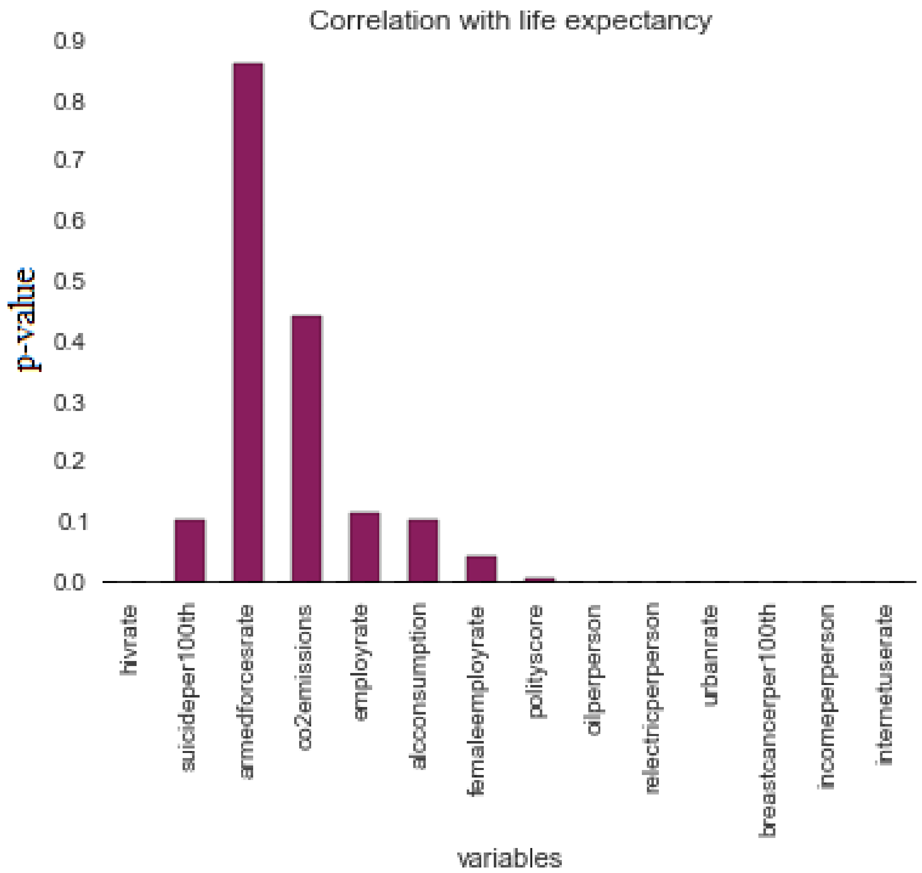

pearsonr = -0.22; p = 0.11
lifeexpectancy
suicideper100th
100
90
80
70
60
50
40
−10
0
10
20
30
40
50


pearsonr = 0.21; p = 0.12
lifeexpectancy
employrate
100
90
80
70
60
50
40
20
30
40
50
60
70
80
90


pearsonr = 0.55; p = 1e-05
lifeexpectancy
urbanrate
100
90
80
70
60
50
40
0
20
40
60
80
100
120
140


lifeexpectancy
alcconsumption
internetuserate
incomeperperson


Correlation with life expectancy
correlation
variables
hivrate
armedforcesrate
suicideper100th
co2emissions
employrate
femaleemployrate
polityscore
oilperperson
alcconsumption
urbanrate
relectricperperson
breastcancerper100th
incomeperperson
internetuserate



Python code fragment shown below:

```
import pandas
import numpy
import seaborn
import scipy
import matplotlib.pyplot as plt

data = pandas.read_csv('C:\\courses\\Coursera\\Current\\Data Analysis Tools\\Week3\\gapminder.csv', low_memory=False)

data_clean = data.drop('country', 1)
data_clean = data_clean.convert_objects(convert_numeric=True) #.dtypes
data_clean = data_clean.replace(' ', numpy.nan)
data_clean = data_clean.dropna()


import seaborn as sns
g = sns.pairplot(data=data_clean,
                 x_vars=['incomeperperson','alcconsumption','armedforcesrate',
                    'breastcancerper100th','co2emissions','femaleemployrate','hivrate',
                    'internetuserate','oilperperson','polityscore','relectricperperson',
                    'suicideper100th','employrate','urbanrate', 'lifeexpectancy'],
                 y_vars=['lifeexpectancy'])

fig, axes = plt.subplots(ncols=14)
for i, xvar in enumerate(['incomeperperson','alcconsumption','armedforcesrate',
                    'breastcancerper100th','co2emissions','femaleemployrate','hivrate',
                    'internetuserate','oilperperson','polityscore','relectricperperson',
                    'suicideper100th','employrate','urbanrate', 'lifeexpectancy']):
    axes[i].scatter(data[xvar],data['lifeexpectancy'])

for x in ['incomeperperson','alcconsumption','armedforcesrate', \
                    'breastcancerper100th','co2emissions','femaleemployrate','hivrate',  \
                    'internetuserate','oilperperson','polityscore','relectricperperson', \
                    'suicideper100th','employrate','urbanrate', 'lifeexpectancy']:
   seaborn.regplot(x=x, y="lifeexpectancy", fit_reg=True, data=data_clean)


scat1 = seaborn.regplot(x="urbanrate", y="internetuserate", fit_reg=True, data=data)
plt.xlabel('Urban Rate')
plt.ylabel('Internet Use Rate')
plt.title('Scatterplot for the Association Between Urban Rate and Internet Use Rate')

scat2 = seaborn.regplot(x="incomeperperson", y="internetuserate", fit_reg=True, data=data)
plt.xlabel('Income per Person')
plt.ylabel('Internet Use Rate')
plt.title('Scatterplot for the Association Between Income per Person and Internet Use Rate')
```

Similar analysis in **R** can be found here: http://rpubs.com/sandipan/155374

Feb 24th, 2016

MORE YOU MIGHT LIKE

Segmenting different countries in the GapMinder dataset with KMeans Clustering using Python Scikit Learn and Pandas

A **k-means cluster** analysis was conducted on the GapMinder dataset to identify underlying subgroups of countries based on their similarity of responses on 14 variables that represent characteristics that could have an impact on **life expectancy**. Clustering variables included the following variables 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. The variable **life expectancy** was *not used* in clustering, it was used later as *ground truth*, to verify whether the clusters obtained were significantly different by comparing *mean life expectancy* across the clusters (with **ANOVA** and **Tukey HSD** tests). All clustering variables were standardized (z-score normalized) to have a mean of 0 and a standard deviation of 1.

After removing the obeservations with missing values in the variable life expectancy, the data were first imputed (all the missing values in other variables were replaced by the corresponding median values) and then randomly split into a training set that included 70% of the observations (N=133) and a test set that included 30% of the observations (N=58).

A series of **k-means cluster** analyses were conducted on the training data specifying k=1-9 clusters, using Euclidean distance. The variance in the clustering variables that was accounted for by the clusters (r-square) was plotted for each of the nine cluster solutions in an elbow curve to provide guidance for choosing the number of clusters to interpret.The elbow curve was inconclusive, suggesting that the

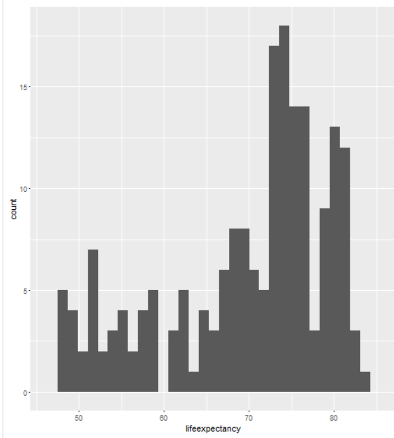
Finding the most important predictors for life expectancy and making predictions with the GapMinder dataset with Random Forest and ExtraTree Forest Ensemble Classifiers using Python Scikit Learn and Pandas

sandipanumbc:

Finding patterns (decision rules) and predicting the life expectancy from the GapMinder dataset with Decision Tree (CART) with R

Decision tree analysis was performed to test nonlinear relationships among a set of explanatory variables and a binary, categorical response variable. All possible separations (categorical) or cut points (quantitative) are tested. For the present analyses, the entropy “goodness of split” criterion was used to grow the tree and a cost complexity algorithm was used for pruning the full tree into a final subtree.

The following explanatory variables were included as possible contributors to a classification tree model evaluating life expectancy (my response variable, which is a continuous numeric variable (the histogram below shows the distribution of life expectancy, which I used to decide the cut point) but was binned into 2 categories: if life expectancy > 70, then life expectancy = High otherwise Low), 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.



The following shows the decision tree model learnt using CART algorithm in R:

df$lifeexpectancy.factor <-

Testing associa betwee expecta the alco consum differen countri the Gap dataset Chi-squ Post-hc with Py Scikit Le

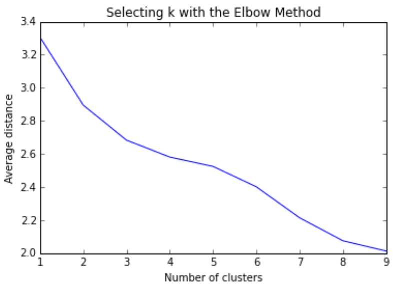
We are intereste assoication betw expectancy

and alcohol cons Gapminder datas variables are nun converted both o categorical varia around the media examining the as expectancy (cate alcohol consump explanatory), a c independence re countries with lo rates ((0-6] litres have higher life e them with (0-70] those with high a rates (63% of the years), χ2=18.61 p-value=1.60227

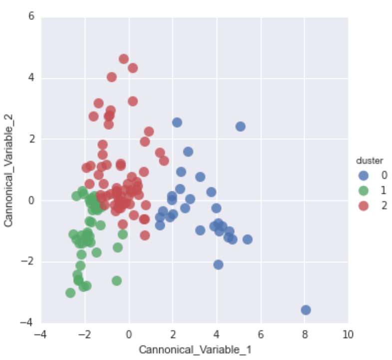
The df or degree is the number of l explanatory varia since the alcohol levels (df 2-1=1).

	lifeexpectan
count	176.000000
mean	69.143682
std	9.828267
min	47.794000
25%	62.646000
50%	72.558500
75%	75.985000
max	83.394000

2, 3, 4 and 8-cluster solutions might be interpreted. The results below are for an interpretation of the 3-cluster solution.



Canonical discriminant analyses was used to reduce the 14 clustering variable down a few variables that accounted for most of the variance in the clustering variables. A scatterplot of the first two canonical variables by cluster (as shown below in the next figure) indicated that the observations in clusters were packed with relatively low within-cluster variance, and did not overlap much with the other clusters. Cluster 2 was generally distinct and densely packed and the observations had low spread suggesting low within-cluster variance. Observations in cluster 0 were spread out more than the other clusters, showing high within-cluster variance. The results of this plot suggest that the best cluster solution may have 3 or more than 3 clusters, so it will be especially important to also evaluate the cluster solutions with more than 3 clusters.



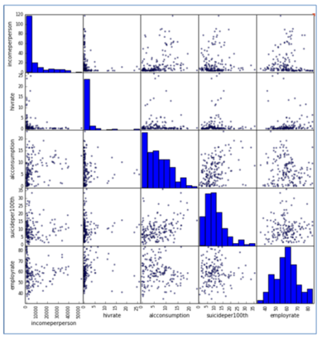
Clusters 0, 1 and 2 contained 28, 42, 63 observations respectively. The means on the clustering variables showed that, compared to the other clusters, countries in cluster 0 had high levels on most of the clustering variables. They had a relatively high income per person, alcohol consumption, breastcancer percentage, co2 emissions, internet use rate, oil per person, urban rate, suicide percentage, but moderate levels of armed forces rate, employment rate and female employment rate. They also appeared to have the lowest levels of hiv rate. Similarly, we can describe the other 2 clusters by the means of the clustering variables as shown below.

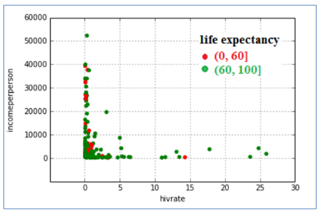
variables

Ensemble learning using **Random Forest** and **ExtraTree Forest** were performed to evaluate the importance of a series of explanatory variables in predicting a binary, categorical response variable with the GamMinder dataset.

The following explanatory variables were included as possible contributors to a classification tree model evaluating **life expectancy** (my response variable, which is a continuous numeric variable but was binned into 2 categories: (0-60] and (60-100]), 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.

The following figure shows relations in between some of the predictors used and some exploratory visualizations:

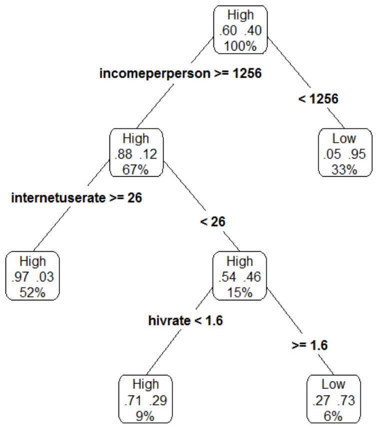




After removal of the NA values in the life expectancy variable, the predictor variables in the original dataset were imputed, the missing values in the numeric columns were replaced with median values. Then the dataset was divided (by taking a random sample of size 60% of the entire dataset) into training dataset with 114 data tuples and test dataset with 77 data tuples. Then the Random Forest and Extra Tree classifiers were trained on the trainign dataset and the models were used to predict on the test dataset. Ans ensemble of 25 decision trees were used to build the random forest predictor and gini index measure was used for the best feature selection at each round

as.factor(ifelse(df$lifeexpectancy > 70, 'High', 'Low'))

tr <- rpart(lifeexpectancy.factor~.-lifeexpectancy, df)



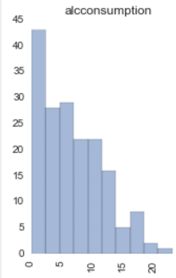
The original dataset contained 191 data tuples (each row representing a country) after removal of the NA values, which was then divided into training (by taking a random sample of size 60% of the entire dataset) and test dataset (the rest part).

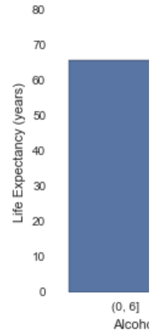
As can be seen from above 3 of the predictors were used by the decision tree for classification of the life expectancy binary class: income per person, internet user rate and hiv rate.

The income per person score was the first variable to separate the training sample into two subgroups. If a country has income per person more than 1256 (per capita in constant 2000 US $) and the internet user rate is more than 26%, the country is more likely (97% of the time) to have High life expectancy. On the other hand, if a country has income per person less than 1256, it is likely to have Low life expectancy (in 95% of the cases present in the leaf node, where that rightmost leaf node itself contained 33% of the data tuples).

Another rule (pattern) found was: if the income per person for a country is higher than 1256 and the internet user rate is below 26% and the hiv rate is below 1.6% then also the country is likely to have High life expectancy.

The model learnt from the training dataset was used to predict the life expectancy for the countries in the test dataset. The confusion matrix (contingency table) on the test dataset is shown below, which shows that we obtained ~88.3% accuracy on the held-out unseen dataset.





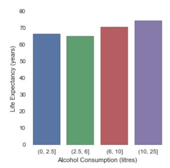

```
lifeexpectancy  (0, 70]  (70, 100]
alcconsumption
(0, 6]               53         37
(6, 25]              22         64

lifeexpectancy   (0, 70]  (70, 100]
alcconsumption
(0, 6]          0.706667   0.366337
(6, 25]         0.293333   0.633663

chi-square value, p value, expected cour
(18.61177373551309, 1.6022793011516588e-
       [ 36.64772727,  49.35227273]]))
```

```
# contingency table of observec
ct1=pandas.crosstab(df2['alccor
print (ct1)
```

```
# chi-square
print ('chi-square valu
cs1= scipy.stats.chi2_c
print (cs1)
```

Model Interpreta
Chi-Square Test

Now in order to u association bette numeric explanat consumption into the quartiles) and test. The Chi Squ independence ag alcohol consump ordered categorie expectancy (bina variable) were sig χ−square=21.534 p-value=8.15187




```
lifeexpectancy  (0, 70]  (70, 100]
alcconsumption
(0, 2.5]             25         19
(10, 25]              8         34
(2.5, 6]             28         18
(6, 10]              14         30

lifeexpectancy   (0, 70]  (70, 100]
alcconsumption
(0, 2.5]        0.333333   0.188119
(10, 25]        0.106667   0.336634
(2.5, 6]        0.373333   0.178218
(6, 10]         0.186667   0.297030

chi-square value, p value, expected c
(21.534561903395307, 8.15187041667602
       [ 17.89772727,  24.10227273],
       [ 19.60227273,  26.39772727],
       [ 18.75      ,  25.25      ]])
```

As we can see fr the p-value < 0.0 the null hypothesi and) conclude tha variableslife expe consumption are

0. incomeperperson
1. alcconsumption
2. armedforcesrate
3. breastcancerper100th
4. co2emissions
5. femaleemployrate
6. hivrate
7. internetuserate
8. oilperperson
9. polityscore
10. relectricperperson
11. suicideper100th
12. employrate
13. urbanrate

Clustering variable means by cluster

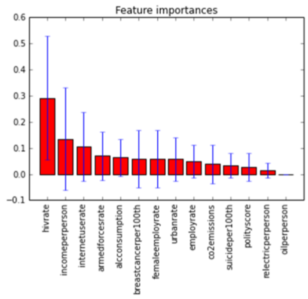
```
variables           0
   1              2
   3               4
cluster


0          1.538475  0.760138
 0.011147  1.306950  0.569876
1         -0.573720 -0.577666
-0.418215 -0.676802 -0.092707
2         -0.363449 -0.029717
 0.193216 -0.152813 -0.146593

                        5
      6                 7
      8                 9
     10           11
cluster


0        0.172466 -0.340029
 1.486122  1.002050  0.491119
 1.248254  0.1896
1        0.698932  0.261283
-0.919388 -0.225554 -0.479426
-0.435333  0.0319
2       -0.581180 -0.025354
-0.130583 -0.247051  0.002781
-0.311070 -0.3972

                      12
 13
cluster
0        0.062857  1.101800
1        0.795325 -0.924312
2       -0.585185  0.165104
```

In order to externally validate the clusters, an Analysis of Variance (**ANOVA**) was conducted to test for significant differences between the clusters on life expectancy. A tukey test was used for post hoc comparisons between the clusters. Results indicated significant differences between the clusters on life expectancy (F(2, 130)=58.08, p<.00000001). The **Tukey post hoc** comparisons showed **significant differences** (rejecting the null hypothesis of no

for the decision trees.

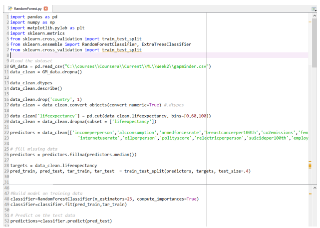
As can be seen from the 3 most important predictors selected by the ExtraTree Forest model were: hiv rate, income per person and internet user rate.

The model learnt from the training dataset was used to predict the life expectancy for the countries in the test dataset. The confusion matrix (contingency table) on the test dataset is shown below, which shows that we obtained ~**90.9% accuracy** on the held-out unseen dataset.



Predicted	(0, 60]	(60, 100]
Actual		
(0, 60]	13	7
(60, 100]	0	57

Part of the python code attached:



sandipanumbc

1 note

	Predicted	
Actual	High	Low
High	43	5
Low	4	25

1 note

Using One-way Analysis of Variance with R and Python to find the Association between quantitative response variable Life expectancy and the converted categorical explanatory variable Income per person / Alcohol consumption in the GapMinder Dataset

Model Interpretation for ANOVA:

When examining the association between the **life expectancy** in number of years (*quantitative response*) and the variable **income per person** (which is the GDP per capita in constant 2000 US$) categorized into *2 ordered categories* (if income per person is in between (0, 2385], it's *low*, otherwise it's *high*, where 2385 is approximately the median value of the variable, splitting around which we got *categorical explanatory variable*) for different countries from the Gapminder dataset, a (one-way) Analysis of Variance (ANOVA) revealed that among the countries with high (2385-52302] income per person, reported to have significantly more life expectancy (Mean=75.74 s.d. ±6.08) compared to the countries with low (0-2385] income per person (Mean=63.57, s.d. ±8.86), F(1, 174)=113.0, p = 1.8 x 10^(-20) .

Note that the degrees of freedom that I report in parentheses) following 'F' can be found in the OLS table as the DF model and DF residuals. In this example 113.0 is the actual F value from the OLS table and we commonly report a very very small p value as simply = 1.8 x 10^(-20).

Now, we need to
comparisons to te
between different
consumption. The
levels, so we nee
square tests for e
alcohol consump
expectancy, with
correction on p-v
0.05/6=0.08333 a
significance).

Post hoc compar
expectancy by pa
consumption cate
higher life expect
the countries with
alcohol consump
with (high) alcoho
between (10,25] l
(statistically) sign
expectancy (with
value≈0.00073<0
countries with (lo
consumption rate

```
lifeexpectancy        (0, 70]  (70, 100
alcconsumption2.5v25
(0, 2.5]                   25         1
(10, 25]                    8         3

lifeexpectancy         (0, 70]  (70, 10
alcconsumption2.5v25
(0, 2.5]              0.757576   0.3584
(10, 25]              0.242424   0.6415

chi-square value, p value, expected cou
(11.415351134476344, 0.000728397289296
       [ 16.11627907,  25.88372093]]))
```

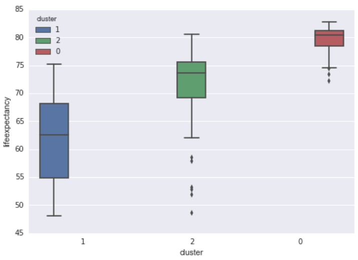
In comparison, po
of life expectancy
expectancy is sta
among those cou
consumptions (0,
p-value greater th
in the following re

```
lifeexpectancy        (0, 70]  (70, 10
alcconsumption2.5v6
(0, 2.5]                   25
(2.5, 6]                   28

lifeexpectancy         (0, 70]  (70, 1
alcconsumption2.5v6
(0, 2.5]              0.471698   0.513
(2.5, 6]              0.528302   0.486

chi-square value, p value, expected c
(0.031042583339548117, 0.860145418596
       [ 27.08888889,  18.91111111]])
```

difference) between all of the 3 clusters on *life expectancy*. Countries in cluster 0 had the highest life expectancy (mean=79.32, sd=2.85), and cluster 1 had the lowest life expectancy (mean=61.6, sd=8.38).

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          lifeexpectancy   R-squared:                       0.472
Model:                             OLS   Adj. R-squared:                  0.464
Method:                  Least Squares   F-statistic:                     58.08
Date:                 Tue, 23 Feb 2016   Prob (F-statistic):           9.48e-19
Time:                         16:44:08   Log-Likelihood:                -444.11
No. Observations:                  133   AIC:                             894.2
Df Residuals:                      130   BIC:                             902.9
Df Model:                            2
Covariance Type:             nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------
Intercept          79.3245      1.304     60.823      0.000        76.744    81.905
C(cluster)[T.1]   -17.7251      1.684    -10.528      0.000       -21.056   -14.394
C(cluster)[T.2]    -7.8677      1.567     -5.019      0.000       -10.969    -4.767
==============================================================================
Omnibus:                        18.309   Durbin-Watson:                   2.036
Prob(Omnibus):                   0.000   Jarque-Bera (JB):               21.294
Skew:                           -0.885   Prob(JB):                     2.38e-05
Kurtosis:                        3.844   Cond. No.                         4.67
==============================================================================
```

```
means for lifeexpectancy by cluster
         lifeexpectancy
cluster
0             79.324464
1             61.599333
2             71.456794

standard deviations for lifeexpectancy by cluster
         lifeexpectancy
cluster
0              2.847906
1              8.382371
2              7.061208
```




```
Multiple Comparison of Means - Tukey HSD,FWER=0.05
============================================
group1 group2 meandiff  lower    upper  reject
--------------------------------------------
  0      1    -17.7251 -21.7172 -13.733  True
  0      2    -7.8677  -11.5841 -4.1512  True
  1      2     9.8575   6.5979  13.117   True
--------------------------------------------
```

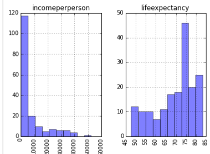
The following table shows the **mean** values for clustering variables in the **test** dataset:

```
Test data clustering variable means by cluster
             index         0         1         2         3         4
cluster
0        30.083333  1.665091  0.706143 -0.057190  1.387202  0.120254
1        30.200000 -0.562181 -0.256216 -0.042824 -0.725118 -0.188946
2        26.461538 -0.185431  0.057726  0.254732 -0.026375 -0.018908

                5         6         7         8         9        10        11
cluster
0        0.145309 -0.263278  1.470878  0.323087  0.638233  1.092147  0.315470
1        0.975337  0.556141 -0.708892 -0.167463 -0.288587 -0.378082  0.623598
2       -0.723859 -0.300740  0.067574 -0.136450  0.166242 -0.100531  0.081538

               12        13
cluster
0        0.074531  0.874936
1        0.976838 -1.131510
2       -0.720312  0.373080
```

Python code fragment for the analysis

```python

# Data Management
data_clean = data.drop('country', 1)
data_clean = data_clean.convert_objects(convert_numeric=True) #.dtypes
data_clean = data_clean.dropna(subset = ['lifeexpectancy'])

# subset clustering variables
cluster=data_clean[['incomeperperson','alcconsumption','armedforcesrate',
                    'breastcancerper100th','co2emissions','femaleemployrate','hivrate',
                    'internetuserate','oilperperson','polityscore','relectricperperson',
                    'suicideper100th','employrate','urbanrate', 'lifeexpectancy']]
cluster.describe()

# standardize clustering variables to have mean=0 and sd=1
clustervar=cluster.copy()
clustervar = clustervar.fillna(clustervar.median())
clustervar['incomeperperson']=preprocessing.scale(clustervar['incomeperperson'].astype('float64'))
clustervar['alcconsumption']=preprocessing.scale(clustervar['alcconsumption'].astype('float64'))
clustervar['armedforcesrate']=preprocessing.scale(clustervar['armedforcesrate'].astype('float64'))
clustervar['breastcancerper100th']=preprocessing.scale(clustervar['breastcancerper100th'].astype('float64'))
clustervar['co2emissions']=preprocessing.scale(clustervar['co2emissions'].astype('float64'))
clustervar['femaleemployrate']=preprocessing.scale(clustervar['femaleemployrate'].astype('float64'))
clustervar['hivrate']=preprocessing.scale(clustervar['hivrate'].astype('float64'))
clustervar['internetuserate']=preprocessing.scale(clustervar['internetuserate'].astype('float64'))
clustervar['oilperperson']=preprocessing.scale(clustervar['oilperperson'].astype('float64'))
clustervar['polityscore']=preprocessing.scale(clustervar['polityscore'].astype('float64'))
clustervar['relectricperperson']=preprocessing.scale(clustervar['relectricperperson'].astype('float64'))
clustervar['suicideper100th']=preprocessing.scale(clustervar['suicideper100th'].astype('float64'))
clustervar['employrate']=preprocessing.scale(clustervar['employrate'].astype('float64'))
clustervar['urbanrate']=preprocessing.scale(clustervar['urbanrate'].astype('float64'))
clustervar['lifeexpectancy']=preprocessing.scale(clustervar['lifeexpectancy'].astype('float64'))

# split data into train and test sets
clus_train, clus_test = train_test_split(clustervar, test_size=.3, random_state=123)

# k-means cluster analysis for 1-9 clusters
from scipy.spatial.distance import cdist
clusters=range(1,10)
meandist=[]

for k in clusters:
    model=KMeans(n_clusters=k)
    model.fit(clus_train)
    clusassign=model.predict(clus_train)
    meandist.append(sum(np.min(cdist(clus_train, model.cluster_centers_, 'euclidean'), axis=1))
    / clus_train.shape[0])

```

The results from python are shown below.



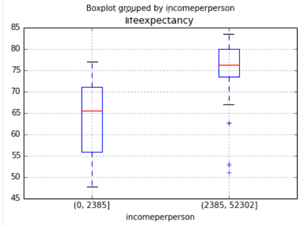
Summary

	lifeexpectancy	incomeperperson
count	176.000000	176.000000
mean	69.654733	7327.444414
std	9.729521	10567.304022
min	47.794000	103.775857
25%	63.041500	702.366463
50%	73.126500	2385.184105
75%	76.569500	8497.779228
max	83.394000	52301.587179

```python
# using ols function for calculating the F-statistic and associated p value
model1 = smf.ols(formula='lifeexpectancy ~ C(incomeperperson)', data=df1)
results1 = model1.fit()
print (results1.summary())
```

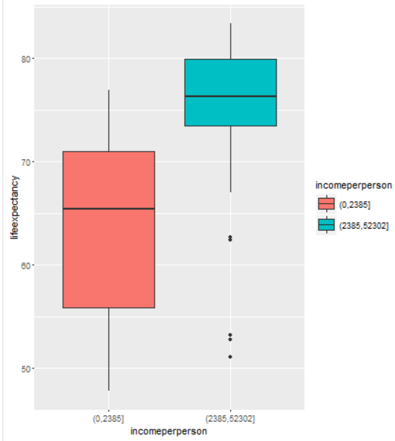
```
                            OLS Regression Results
==============================================================================
Dep. Variable:          lifeexpectancy   R-squared:                       0.394
Model:                             OLS   Adj. R-squared:                  0.390
Method:                  Least Squares   F-statistic:                     113.0
Date:                 Sat, 13 Feb 2016   Prob (F-statistic):           1.18e-20
Time:                         23:34:38   Log-Likelihood:                -605.62
No. Observations:                  176   AIC:                             1215.
Df Residuals:                      174   BIC:                             1222.
Df Model:                            1
==============================================================================================
                                  coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------------------
Intercept                      63.5669      0.810     78.489      0.000        61.968    65.165
C(incomeperperson)[T.(2385, 52302]]  12.1757      1.145     10.631      0.000         9.915    14.436
==============================================================================
Omnibus:                        16.918   Durbin-Watson:                   1.812
Prob(Omnibus):                   0.000   Jarque-Bera (JB):               18.825
Skew:                           -0.776   Prob(JB):                     8.17e-05
Kurtosis:                        3.395   Cond. No.                         2.62
==============================================================================
```

```
means for lifeexpectancy by incomeperperson categories
                lifeexpectancy
incomeperperson
(0, 2385]            63.566886
(2385, 52302]        75.742580
```



The following are the same results with R

```
                 Df Sum Sq Mean Sq F value                Pr(>F)
incomeperperson   1   6523    6523     113 <0.0000000000000002 ***
Residuals       174  10043      58
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

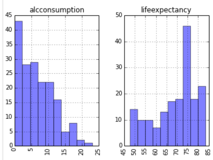


Model Interpretation for post hoc ANOVA results:

When examining the association between the **life expectancy** in number of years (*quantitative response*) and another explanatory variable **alcohol consumption** (avg in litres) categorized into *4 ordered categories* (splitting around the quartiles we got *categorical explanatory variable* with 4 levels (0,3], (3-6], (6-10], (10-25]) for different countries from the same dataset, (one-way) ANOVA revealed that among daily, the life expectancy (quantitative response variable) and alcohol consumption were significantly associated, F (3, 172)=8.927, p=1.57x10^-5.

Post hoc comparisons of the alcohol consumption by pairs of categories revealed that the countries with alcohol consumption level (10,25] (group 1) reported significantly more life expectancy compared to those with level (0,3] (group 0). Similarly, the countries with alcohol consumption level (10,25] reported significantly more life expectancy compared to those with level (3,6]. And the countries with alcohol consumption level (6,10] reported significantly more life expectancy compared to those with level (3,6]. All other comparisons were statistically similar.

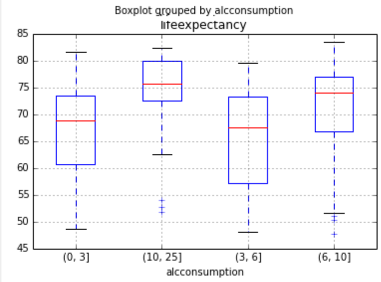
The results from python are shown below.



Summary

	lifeexpectancy	alcconsumption
count	176.000000	176.000000
mean	69.143682	6.646250
std	9.828267	4.977415
min	47.794000	0.030000
25%	62.646000	2.495000
50%	72.558500	5.865000
75%	75.985000	9.810000
max	83.394000	23.010000

```
means for lifeexpectancy by alcconsumption
                     lifeexpectancy
alcconsumption
(0, 3]                    66.850458
(10, 25]                  74.411119
(3, 6]                    64.897810
(6, 10]                   70.670250
```

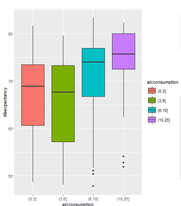



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         lifeexpectancy   R-squared:                       0.135
Model:                            OLS   Adj. R-squared:                  0.120
Method:                 Least Squares   F-statistic:                     8.927
Date:                Sat, 13 Feb 2016   Prob (F-statistic):           1.57e-05
Time:                        23:57:07   Log-Likelihood:                -638.70
No. Observations:                 176   AIC:                             1285.
Df Residuals:                     172   BIC:                             1298.
Df Model:                           3
===============================================================================================
                                coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------------------
Intercept                    66.8505      1.331     50.225      0.000        64.223    69.478
C(alcconsumption)[T.(10, 25]]  7.5607      1.948      3.880      0.000         3.715    11.407
C(alcconsumption)[T.(3, 6]]  -1.9526      1.948     -1.002      0.318        -5.799     1.893
C(alcconsumption)[T.(6, 10]]  3.8198      1.925      1.985      0.049         0.021     7.619
==============================================================================
Omnibus:                       15.572   Durbin-Watson:                   1.853
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               17.547
Skew:                          -0.753   Prob(JB):                     0.000155
Kurtosis:                       2.646   Cond. No.                         4.62
==============================================================================
```

```
mc1 = multi.MultiComparison(df2['lifeexpectancy'], df2['alcconsumption'])
res1 = mc1.tukeyhsd()
print(res1.summary())
```

```
Multiple Comparison of Means - Tukey HSD,FWER=0.05
==============================================
group1 group2 meandiff  lower    upper  reject
----------------------------------------------
  0      1     7.5607   2.5055  12.6158  True
  0      2    -1.9526  -7.0078   3.1025 False
  0      3     3.8198  -1.1737   8.8133 False
  1      2    -9.5133 -14.7342  -4.2924  True
  1      3    -3.7409  -8.9021   1.4204 False
  2      3     5.7724   0.6112  10.9337  True
```

The following are the same results with R




```
> summary(aov.out)
                Df Sum Sq Mean Sq F value    Pr(>F)
alcconsumption   3   2277   759.1   8.927 0.0000157 ***
Residuals      172  14627    85.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lifeexpectancy ~ alcconsumption, data = df2)

$alcconsumption
                      diff        lwr       upr     p adj
(3,6]-(0,3]      -1.952649 -7.0075007  3.102203 0.7483361
(6,10]-(0,3]      3.819792 -1.1734134  8.812997 0.1977879
(10,25]-(0,3]     7.560661  2.5058088 12.615513 0.0008481
(6,10]-(3,6]      5.772440  0.6114781 10.933403 0.0216368
(10,25]-(3,6]     9.513310  4.2926809 14.733938 0.0000278
(10,25]-(6,10]    3.740869 -1.4200933  8.901831 0.2402665
```

Show more