Introduction to Programming

Data Science and Advanced Analytics

NOVA IMS

2018/2019

VITOR MANITA
M20180054

RODRIGO UMBELINO
M20180060

RITA FRANCO
M20180081

ADOLESCENTS ALCOHOL CONSUMPTION

Adolescents Alcohol Consumption

Vitor Manita, Rita Franco, Rodrigo Umbelino
[1] M20180054; m20180054@novaims.unl.pt
[2] M20180081; m20180081@novaims.unl.pt
[3] M20180060; m20180060@novaims.unl.pt

Abstract:

The purpose of this research is to study and identify behavior patterns regarding adolescents' alcohol consumption, as well as other risk substances. The presented data was collected from the World Health Organization dataset entitled "Health Behavior in School Aged Children". We used plenty of different data analysis methods and models using Python, a high-level programming language.

All scripts are openly available at: *https://github.com/m20180054/IP-Project*

Keywords:

Alcohol; Python; Gender

Statement of Contribution:

The group worked together and evenly for the accomplishment of this project. Due to a medical condition, element m20180060 could not be present for some part of the work but followed the process, so we believe there shouldn't be any penalty. Thus, elements m20180054 and m20180081 were responsible by coding the analysis made. Regarding the report and overall justifications and conclusions, all three elements were fundamental.

Index

I. Introduction 3
II. Data and Pre-processing 4
III. Process and Results 5
a. Correlations 5
b. Gender Gap..7
c. Extra Variables 9
d. Regressions..10
e. PCA 14
IV. Conclusions and Future objectives 18
V. References 18
VI. Attachments..19
f. Importing tables..19
g. Aggregating columns into one (example)..20
h. Creating a DataFrame with required variables and separating by gender..20
i. Explained variance of each variable in each Principal Component 21

Figure 1. Correlation Matrix for Girls 5
Figure 2. Correlation Matrix for Boys 6
Figure 3. Percentage of adolescents that drink daily or weekly by gender 7
Figure 4. Percentage of adolescents that drink 3 or more drinks by gender. 8
Figure 5. Correlation Matrix for all variables 9
Figure 6. Code for Regressions 10
Figure 7. Variable importance based on R-squared for been_drunk..11
Figure 8. OLS Regression Results 11
Figure 9. Variable importance based on R-squared..12
Figure 10. OLS Regressions Outputs 13
Figure 11. PCA elbow graph 14
Figure 12.Hierarchical Clustering Dendrogram 14
Figure 13. k-means elbow graph 14
Figure 14. Cluster centroids..16
Figure 15. Geographic clusters representation..17

Table 1. Description of the variables 4
Table 2. Description of extra variables 9
Table 3. Clusters composition 15

I. Introduction

The World Health Organization is a specific department within the United Nations that focuses on international public health. In this case, data comes from a collaborative cross-national study, which aggregates information from participant countries and regions across Europe and North America, in efforts to understand young people’s health in their social context - living conditions, school, family and friends. In addition to this, behaviors that are adopted during the adolescence have a direct influence in issues like alcohol and tobacco use, drug abuse, physical activity levels, mental health, health complaints, along with many others.

With the aim of withdrawing new and even more relevant conclusions from this vast repository of information, it was decided that it would be interesting to collect this data and process and analyze the correlations between the different variables. In the making of this project, the focus was producing user-friendly and appellative visualizations, which aim to reduce the gap between the content itself and a reader with not a lot of experience in analytical or statistical fields.

Finally, and always taking into consideration the alcohol consumption behavior, it was deemed important to know how the gender gap varies in different countries. Furthermore, having found that there were indeed several variables related with adolescents’ behavior in different countries, it was only logical to make regression analysis and build several data models to confirm if there are variables that affect each other. Last, but not least, it is provided an exhaustive cluster analysis and visualization, for effective reader comprehension.

II. Data and Pre-processing

As it was stated before, the used data belongs to a W.H.O. dataset, built in the year of 2013/2014. The sampling procedure for this dataset consisted on a selection of students through random selection of classes within targeted school years/grades, and the mode of data collection was a paper/pencil questionnaire. Within this dataset it is recognized that there are indeed some topics which are highly sensitive in some countries, such as illicit drug use or sexual health. For these reasons, it is reported that the inclusion of such topics might not have been possible in some cases.

In a dataset as large as this one, it is advisable to reduce the amount of information, therefore a choice was made of filtering the initial dataset to 12 final tables. This was a choice based on our common sense and social understanding of the problem, as well as researching about the topic in additional readings [1].

table 1 After importing the data, we had the long task of pre-processing it. Many of the tables had from 5 to 10 different categories, and, for efficiency sake, it would be more suitable to reduce the entire table to a single variable. Having said that, the table **first_drunk** (which contains information about the age from 11 to 16 when the respondent first got drunk) and its categories 11, 12 and 13 were aggregated by summing the percentages. This allowed the transformation of the table into a variable that represents the percentage of children with ages below 13 (in each country) that had already been drunk. Similar methods were applied individually for each table and the result was the following:

Variable	Description	Type
alcopops	Adolescents that consume alcohol at least weekly	float, %
been_drunk	Adolescents that have been drunk twice or more	float, %
first_drunk	Adolescents that were 13 years old or under when they first got drunk	float, %
drink_day	Adolescents that drink 3 or more drinks when they go out	float, %
cannabis	Adolescents that have smoked cannabis	float, %
smoking	Adolescents who smoke at least weekly	float, %
sex	Adolescents that have had sexual intercourse	float, %
exercise	Adolescents that do exercise at least 3 hours per week	float, %
evening_friends	Adolescents that meet friends outside school time after 8PM at least weekly	float, %
school_achiev	Adolescents' perception of their teachers' opinion about their school performance, compared to other classmates	float, %
friends_help	Adolescents' opinion on the help that they receive from their friends	float, average (1-7)
family_help	Adolescents' opinion on the help that they receive from their family	float, average (1-7)
family_well	Adolescents' opinion on the state of their family	float, average (1-5)
life_sas	Adolescents' opinion on the state of their own lives (life satisfaction)	float, average (0-10)

Table 1. Description of the variables

Example of the imports and pre-process of the data can be viewed in the source code or in the attachments.

[1] Currie C et al., eds. Social determinants of health and well-being among young people. Health Behavior in School-aged Children (HBSC) study: international report from the 2009/2010 survey. Copenhagen, WHO Regional Office for Europe, 2012 (Health Policy for Children and Adolescents, No. 6)

III. Process and Results

In this chapter, the processing of the data and the subsequence results will be explained and presented. There are five main different types of analysis on the data that reveal different perspectives and knowledge about the way adolescents consume alcohol in these countries.

a. Correlations

Initially a simple correlation matrix was made between all variables for each gender. Although the main purpose of this step was to create important information for further analysis as, for example, which variables should not be included in regressions due to multicollinearity, the results found were rather interesting. Two correlation matrixes were made to find out how the behaviors and relation between various aspects would differ with gender.

(Figure 1, Figure 2) The data shows that adolescents with high family support tend to have high friends' support and vice-versa. The correlation between **friends_help** and **family_help** is 0.78 for girls and even higher for male adolescents, 0.85. An expected result is the high positive correlation between the percentage of kids that have already been drunk when answering the survey and the percentage of kids whose first time getting drunk was before or at the age of 13 (0.84 for girls and 0.81 for boys).

(Figure 1, Figure 2) The most interesting results were found in the correlations that differ with gender, although these correlations are not as strong as the previous ones. Female adolescents that have had sexual relations before were very likely to have been drunk twice or more (correlation between **been_drunk** and **sex** is 0.63). In addition to this, there is a small negative relation between **sex** and **life_sas** (-0.53) which indicates (statistically) that having sexual relations at these young ages leads girls to see their lives worse in average. Regarding boys, the phenomenon is different, instead of being correlated with being drunk, having sexual relations is correlated with going out with friends at least weekly (correlation between **evening_friends** and **sex** is 0.64).

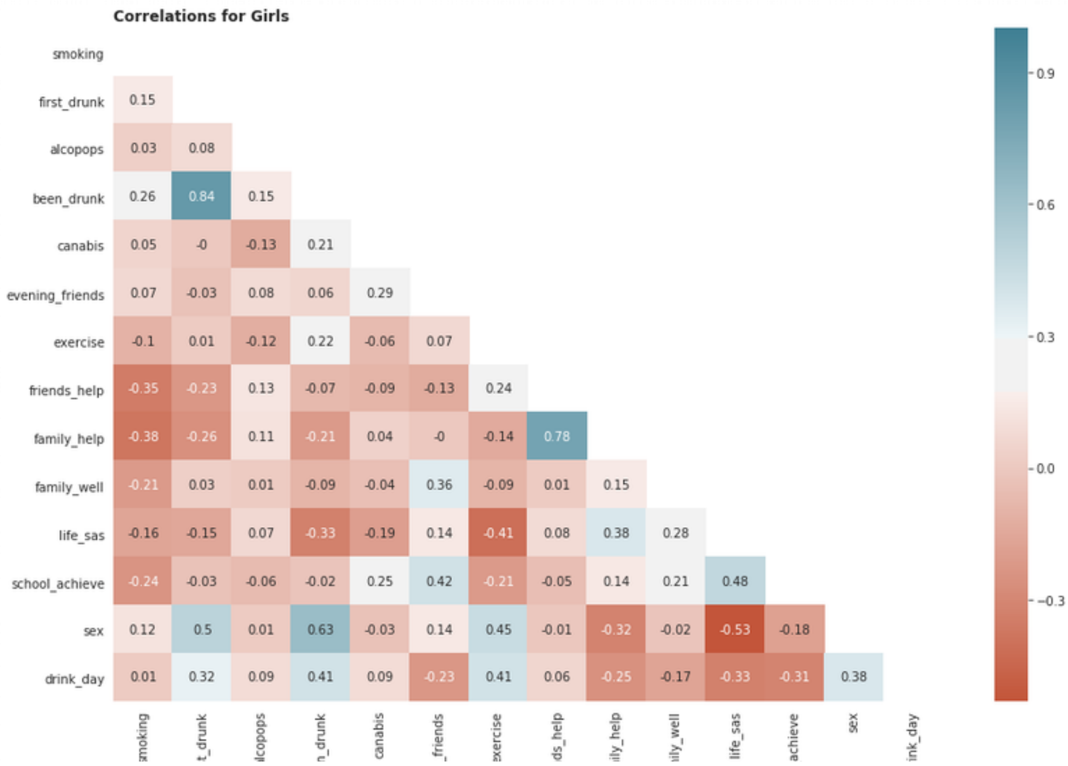


Figure 1. Correlation Matrix for Girls

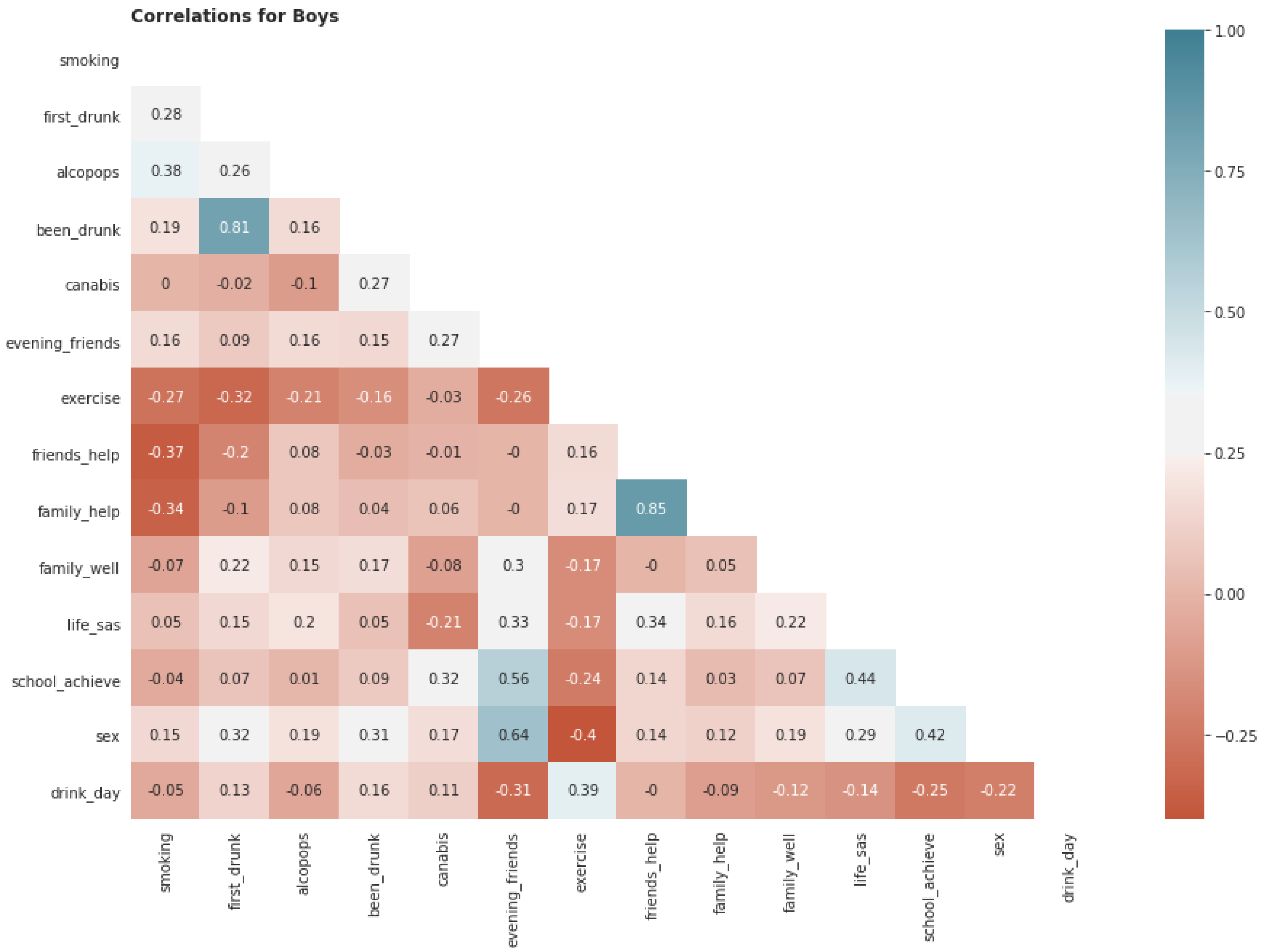


Figure 2. Correlation Matrix for Boys

b. Gender Gap

In this chapter, we have one of the most visually challenging and best-looking visualizations we were able to build using *Python*. There was this question in mind of how the gender gap behaved in different countries, so, the focus was on two variables: **Alcopops** and **Drink_Day,** which regard the percentage of adolescents that drink alcohol daily or weekly and the percentage of adolescents that drink 3 or more drinks when they drink. Highlighted in red, we have Portugal.

Drinking Frequency

(Figure 3) In the following figure, the gap between boys and girls of 11 to 15 years old in terms of alcohol frequency is clear. The biggest difference is in Ukraine and Israel, where the gap in each one is around four percentual points, from 9.6% to 5.7% in Ukraine and 7.9% to 3.4% on Israel. On the opposite side, in countries like Norway, Slovakia and England, the gap is smaller and presents the lowest values for boys and girls, for example, in Norway, the percentage of adolescents that drink daily or weekly is around 1.3% for boys and 0.6% for girls.

It is important to notice that male adolescents show themselves as being more frequent drinkers, except in Canada, Scotland and Netherlands, where they are surpassed by female adolescents. Respectively, in these regions, the percentage of frequent male drinkers is 3%, 2.2% and 2.2% and female drinkers is 3.2%, 2.6% and 2.8%. In this category, Portugal reveals to have 3.1% of its male adolescents drinking very frequently and 2.1% of female adolescents doing so.

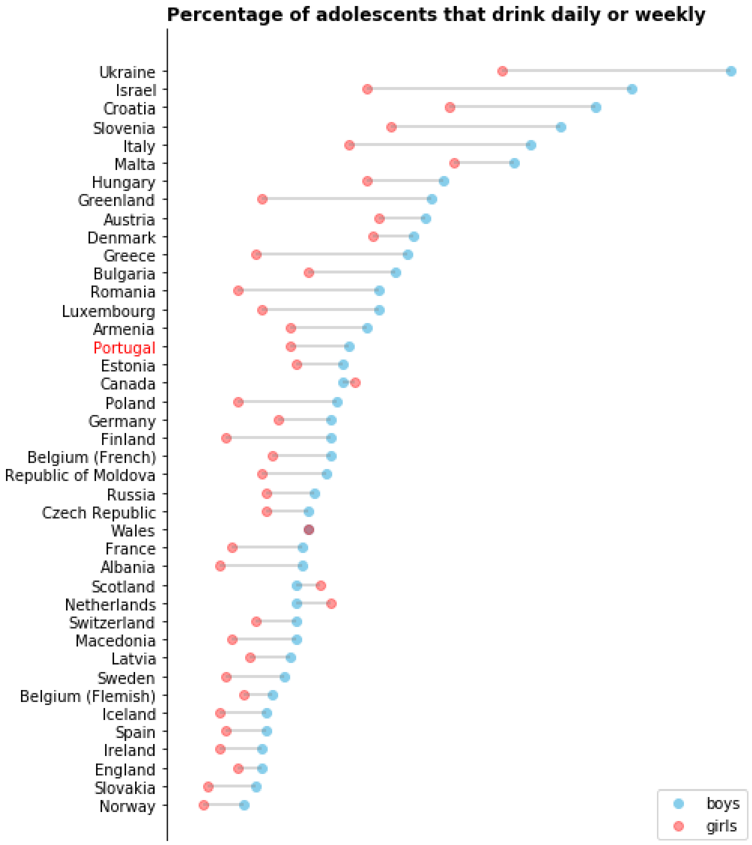


Figure 3. Percentage of adolescents that drink daily or weekly by gender

Top 5 Lower		
Country	Boys	Girls
Norway	1.3%	0.6%
Slovakia	1.5%	0.7%
England	1.6%	1.2%
Ireland	1.6%	0.9%
Spain	1.7%	1%

Top 5 Highest		
Country	Boys	Girls
Ukraine	9.6%	5.7%
Israel	7.9%	3.4%
Croatia	7.3%	4.8%
Slovenia	6.7%	3.8%
Italy	6.2%	3.1%

Drinking Quantities

(Figure 4) When speaking about drinking quantities, in number of drinks, there is an unbalanced view in terms of gender, where boys apparently tend to drink more quantities of alcohol as compared to girls.

Before analyzing, it is important to notice that the scale of the figure, even though it is in the same units as the previous one, the values are more widely spread in this one. We see the highest values in Belgium (French side), Austria and Canada, even though the gap is small. In Belgium (FR), 48% of adolescent boys admit to drink 3 or more drinks when they drink whereas this value increases to 48.6% in the case of adolescent girls. For Austria and Canada, the values are 31.5% and 29.4% for boys and 31.8% and 28.9% for girls.

On the opposite side, Iceland, Italy and Moldova have the lowest values with 3.4%, 4.6% and 5.8% for boys and 3.6%, 3.8% and 2.4% for girls. Iceland follows the example of the top 3 countries where the gap is not so significant. In this component, it is important to notice how there are much more frequent values of female adolescents surpassing boys. Notice that the regions where girls drink with more frequency than boys are as well regions where they drink more when compared to boys.

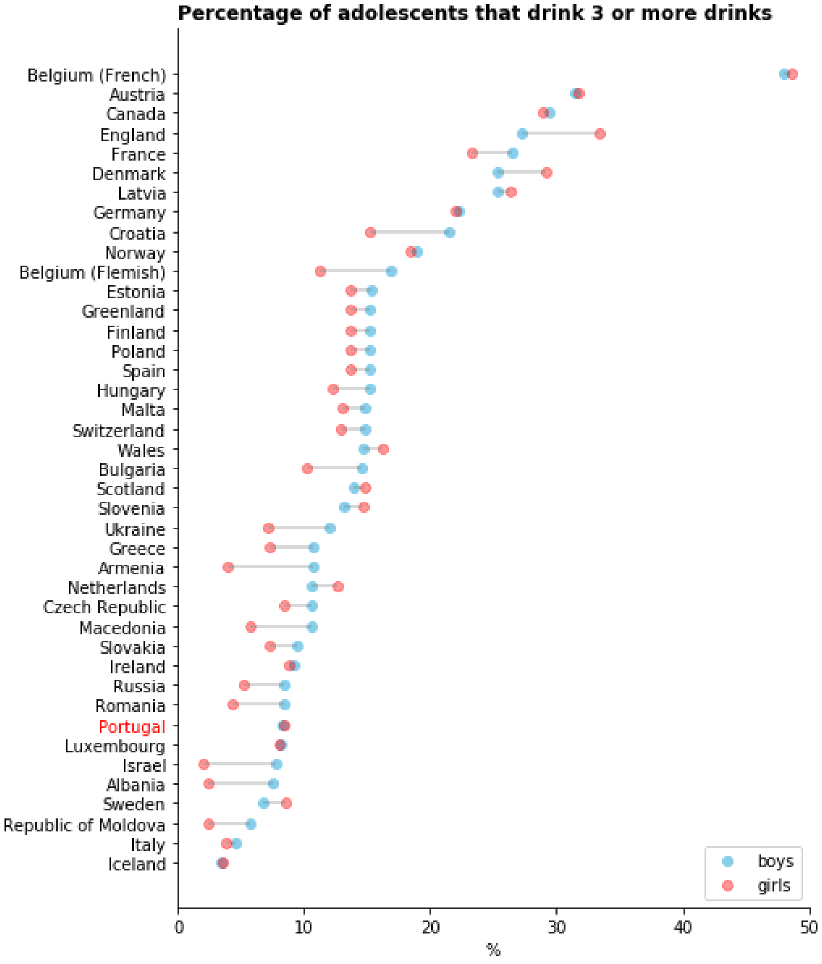


Figure 4. Percentage of adolescents that drink 3 or more drinks by gender.

In Portugal, Female adolescents are the ones that drink more quantities when they do, with values of 8.5% when compared to 8.3% of boys.

Top 5 Lower		
Country	Boys	Girls
Iceland	3.4%	3.6%
Italy	4.6%	3.8%
Moldova	5.8%	2.4
Sweden	6.8%	8.6%
Albania	7.6%	2.4%

Top 5 Highest		
Country	Boys	Girls
Belgium (FR)	48%	48.6%
Austria	31.5%	31.8%
Canada	29.4%	28.9%
England	27.3%	33.4%
France	26.5%	23.3%

c. Extra Variables

Having a set of behaviors of demographic characteristics of our target population, adolescents from 11 to 15 years old, there was a need to explore if some country characteristics had any impact or relationship with what was observed in the previous chapters. All extra variables, as we called them, were collected from *The World Bank.org*.

Variable	Description	Type
GDPpc	Gross Domestic Product per Capita	US $
Rural_Pop	Percentage of Population living in Rural areas	float, %
Urban_Pop	Percentage of Population living in Urban areas	float, %
Teenmom	Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant)	float, %
OutschoolT	Adolescents out of school (% of lower secondary school age)	float, %
Expeduc	Expenses on Education (% of GDP)	float, %

Table 2. Description of extra variables

As seen in the initial data, a pre-process of the data was necessary for its use in the analysis. First, the data was converted into a *DataFrame* format, since its initial format was of a row per indicator, per country, with only one column containing the value, thus repeating the country names for each indicator. Having done this, two indicators were dropped, the percentage of male and female adolescents out of school, that had too many *null* values for viable use, so the result was just one variable for the out of school percentage, that had the total values. Then, the *Null* values in the **Teenmom, OutschoolT** and **Expeduc** indicators were cleaned, since there were not that many, filling the blank spaces with the indicator mean. A conversion to numeric values was also necessary for the use of the data. (Figure 5) In this phase, it is shown how the extra collected variables relate with the initial ones. It is curious to notice that **GDPpc** is positively correlated (0.78) with **Exercise** but negatively (-0.65) with **Teenmom**, which may be explained by bigger expenses on sexual education and free contraceptives. **GDPpc** is also somehow positively correlated (0.57) with the percentage of population living in urban areas, where the exact opposite effect occurs with the rural areas, since **urban_pop** and **rural_pop** have a perfect negative correlation (1-the other).

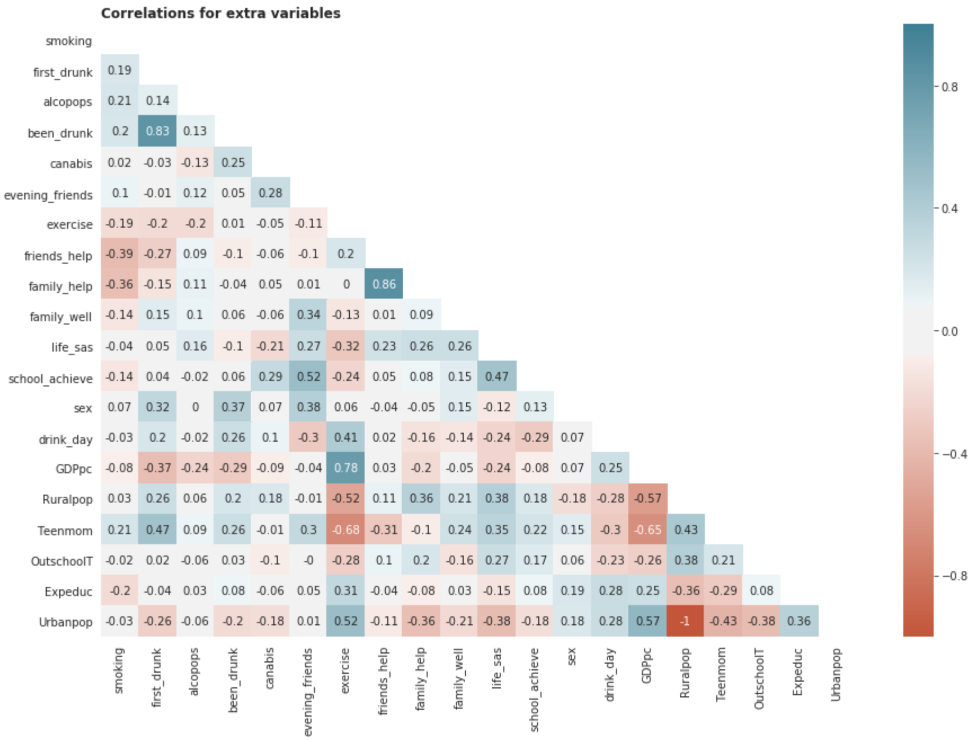


Figure 5. Correlation Matrix for all variables

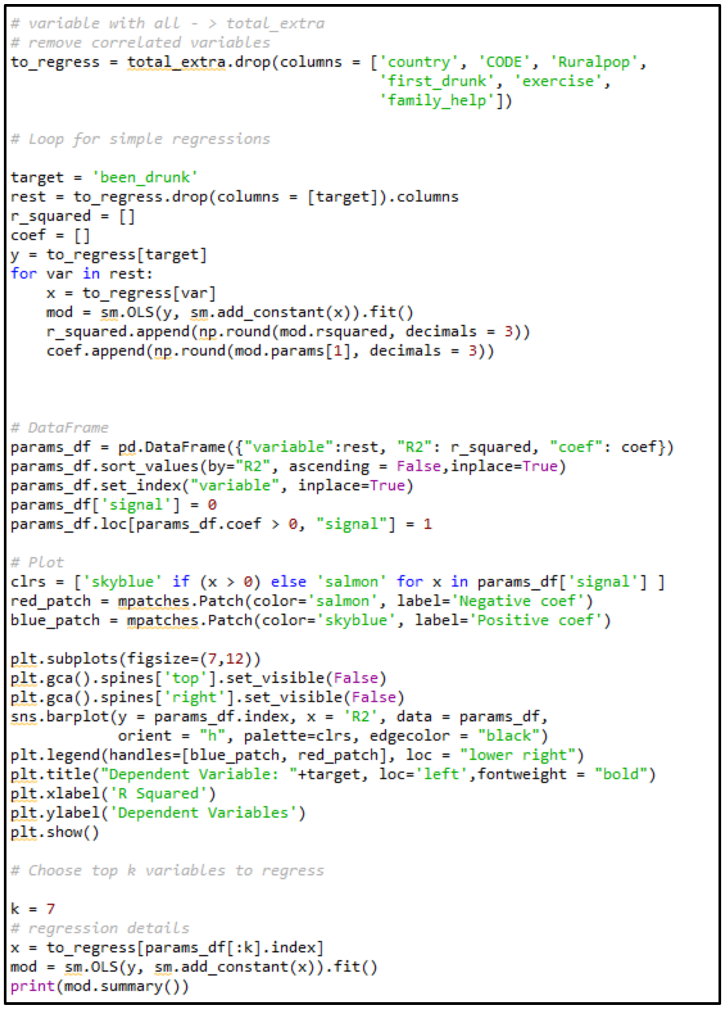
d. Regressions

Having demographic and behavior variables related with the behavior of adolescents in each country, it would be interesting to observe how some variables affect other, therefore, regressions were built. The standard questions emerged: Which is the variable we want to explain and which variables to use with that objective in mind?

For the dependent variable, it was chosen the most interesting ones from our point of view, then, the information obtained from the correlations discovered in the past chapters was applied to remove strongly correlated variables that would affect the analysis by means of multicollinearity.

(Figure 6) To answer the questions of which variables to use, iterative lines of code were made, using *Loops*, that, by changing just one word – the regression target variable – a simple linear regression with just one independent variable is made for each one of all other variables, resulting in *K* regressions where *K* is the number of independent variables. This way, there is an R^2 for each one of the regressions and, by means of a simple bar plot, it is possible to order the variables which, statistically speaking, are the most relevant to explain the target. Therefore, it is now possible, in a simple and visual way, choose the ones with more explaining power.

Then, it was established another *K* representing the maximum number of variables included, by decreasing order, according to their relative importance to the target variable, whose calculations are described in the last paragraph. So, the regressed variables were the most important ones, like: **Been_drunk**, **Drink_day**, **Alcopops**, **Cannabis** and **Sex.** For the in-depth regression statistics, it is only described for one variable, in this report, due to the limit of words.



```
# variable with all - > total_extra
# remove correlated variables
to_regress = total_extra.drop(columns = ['country', 'CODE', 'Ruralpop',
                                         'first_drunk', 'exercise',
                                         'family_help'])

# Loop for simple regressions

target = 'been_drunk'
rest = to_regress.drop(columns = [target]).columns
r_squared = []
coef = []
y = to_regress[target]
for var in rest:
    x = to_regress[var]
    mod = sm.OLS(y, sm.add_constant(x)).fit()
    r_squared.append(np.round(mod.rsquared, decimals = 3))
    coef.append(np.round(mod.params[1], decimals = 3))



# DataFrame
params_df = pd.DataFrame({"variable":rest, "R2": r_squared, "coef": coef})
params_df.sort_values(by="R2", ascending = False,inplace=True)
params_df.set_index("variable", inplace=True)
params_df['signal'] = 0
params_df.loc[params_df.coef > 0, "signal"] = 1

# Plot
clrs = ['skyblue' if (x > 0) else 'salmon' for x in params_df['signal'] ]
red_patch = mpatches.Patch(color='salmon', label='Negative coef')
blue_patch = mpatches.Patch(color='skyblue', label='Positive coef')

plt.subplots(figsize=(7,12))
plt.gca().spines['top'].set_visible(False)
plt.gca().spines['right'].set_visible(False)
sns.barplot(y = params_df.index, x = 'R2', data = params_df,
            orient = "h", palette=clrs, edgecolor = "black")
plt.legend(handles=[blue_patch, red_patch], loc = "lower right")
plt.title("Dependent Variable: "+target, loc='left',fontweight = "bold")
plt.xlabel('R Squared')
plt.ylabel('Dependent Variables')
plt.show()

# Choose top k variables to regress

k = 7
# regression details
x = to_regress[params_df[:k].index]
mod = sm.OLS(y, sm.add_constant(x)).fit()
print(mod.summary())
```

Figure 6. Code for Regressions

(Figure 7) Starting with the percentage of 13 year old adolescents or under that have already been drunk, it is observed that the most impactful characteristics to explain this phenomena are if they already had sexual intercourse, if they drink more than 3 drinks when they do, the GDP per capita of the region they live in, the number of teen moms in that region, the percentage of adolescents that has already smoked cannabis, the percentage of population living in urban areas and the percentage of adolescents that have smoked tobacco at least weekly.

From these variables, all have a positive coefficient/impact on the target variable, except for the **GDPpc** and Urban pop. Notice how the other risk behaviors variables like **drink_day, cannabis** and **smoking** are present in the top 7 most important variables.

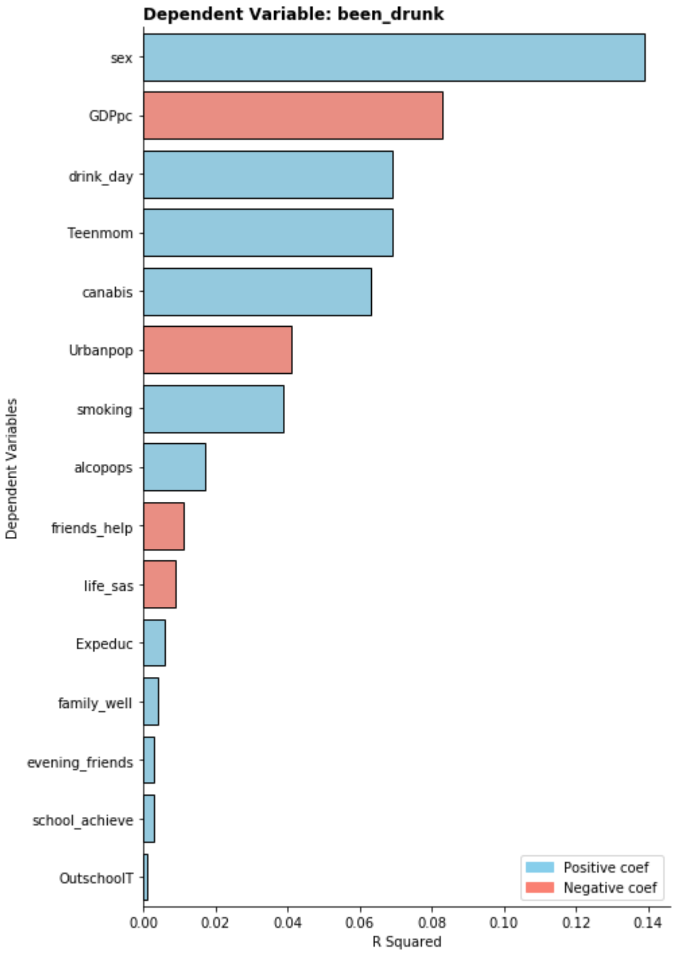
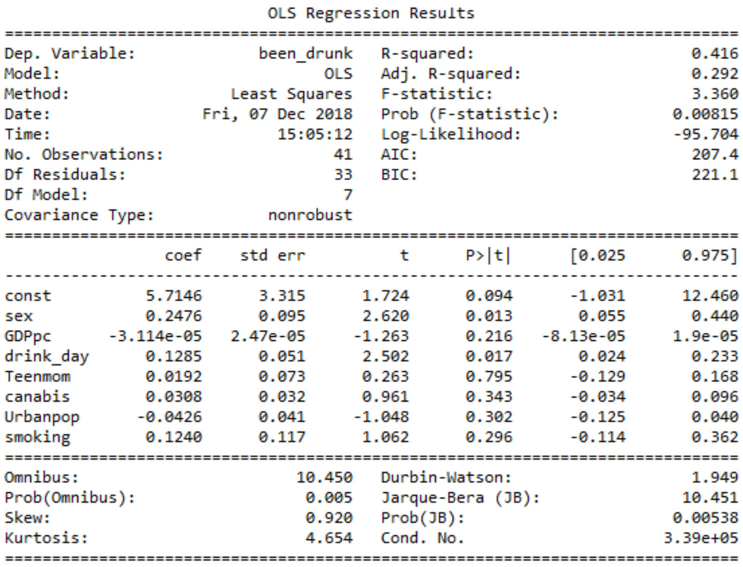


Figure 7. Variable importance based on R-squared for been_drunk

Analysing the model

(Figure 8) Proceeding to an in-depth analysis of the model, it is observed that the target variable, **been_drunk**, described with the 7 top variables, has 41.4% of its variations explained.

Furthermore, by analyzing the *P-value*, only **sex** and **drink_day** are statistically significant at 2% significance level. All others are not significant at the usual levels of 1%, 5% and 10%. Thus, it is only impactful to analyze these two variables. For **sex**, for each percentual point higher in the percentage of adolescents who already had sex, the percentage of adolescents that have already been drunk twice or more increases by 0.13 percentual points. Having as well a positive relationship, for each percentual point higher in the percentage of adolescents that drink 3 drinks or more when they do, the percentage of adolescents that have already been drunk at least twice increases by 0.09 percentual points. On the next page, there are other interesting regressions to make, each one ordered by the most important variables. The statistic descriptions of them are like the one made in the previous paragraph.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             been_drunk   R-squared:                       0.416
Model:                            OLS   Adj. R-squared:                  0.292
Method:                 Least Squares   F-statistic:                     3.360
Date:                Fri, 07 Dec 2018   Prob (F-statistic):            0.00815
Time:                        15:05:12   Log-Likelihood:                -95.704
No. Observations:                  41   AIC:                             207.4
Df Residuals:                      33   BIC:                             221.1
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          5.7146      3.315      1.724      0.094      -1.031      12.460
sex            0.2476      0.095      2.620      0.013       0.055       0.440
GDPpc      -3.114e-05   2.47e-05     -1.263      0.216   -8.13e-05     1.9e-05
drink_day      0.1285      0.051      2.502      0.017       0.024       0.233
Teenmom        0.0192      0.073      0.263      0.795      -0.129       0.168
canabis        0.0308      0.032      0.961      0.343      -0.034       0.096
Urbanpop      -0.0426      0.041     -1.048      0.302      -0.125       0.040
smoking        0.1240      0.117      1.062      0.296      -0.114       0.362
==============================================================================
Omnibus:                       10.450   Durbin-Watson:                   1.949
Prob(Omnibus):                  0.005   Jarque-Bera (JB):               10.451
Skew:                           0.920   Prob(JB):                      0.00538
Kurtosis:                       4.654   Cond. No.                     3.39e+05
==============================================================================
```

Figure 8. OLS Regression Results

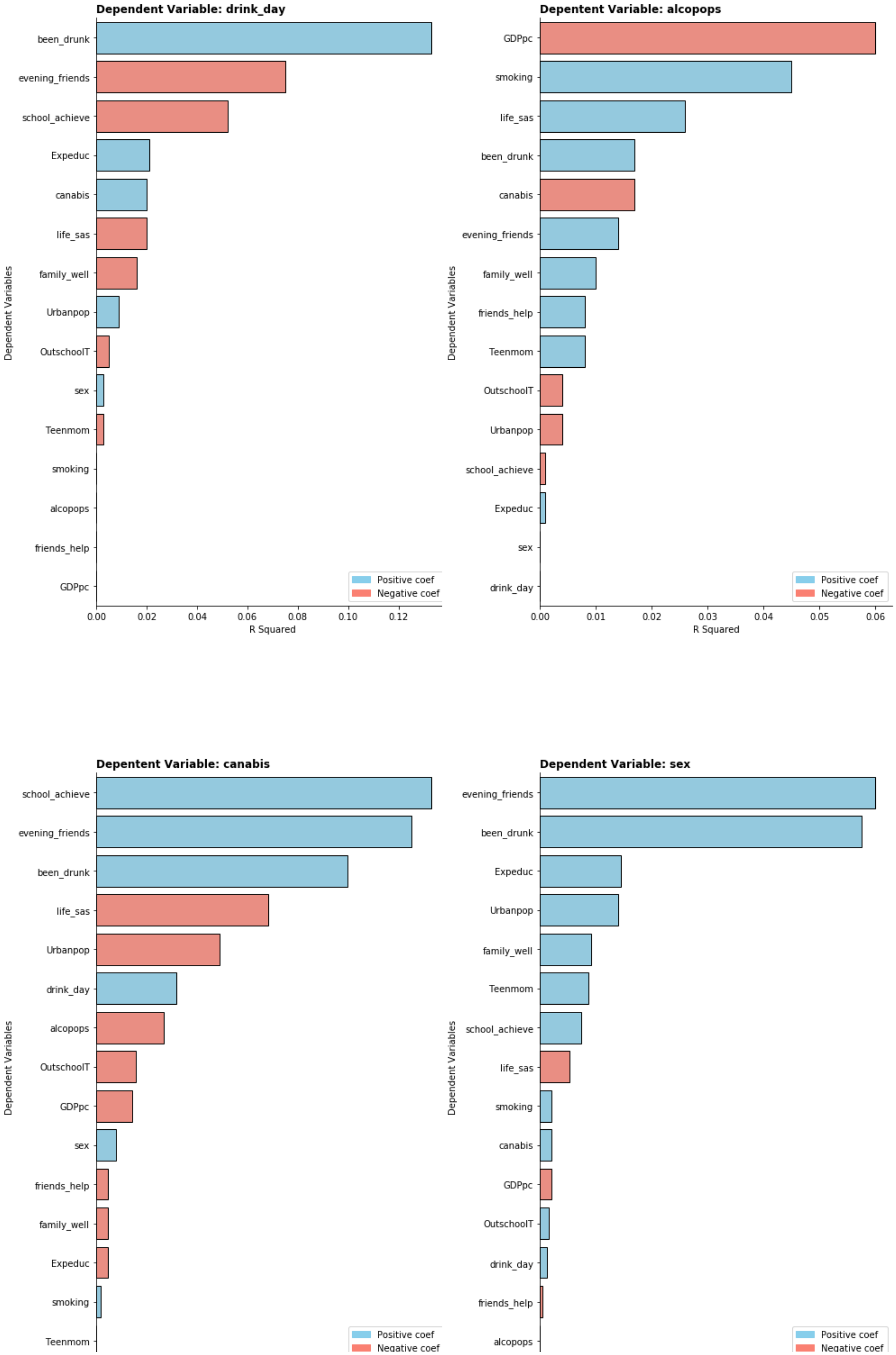
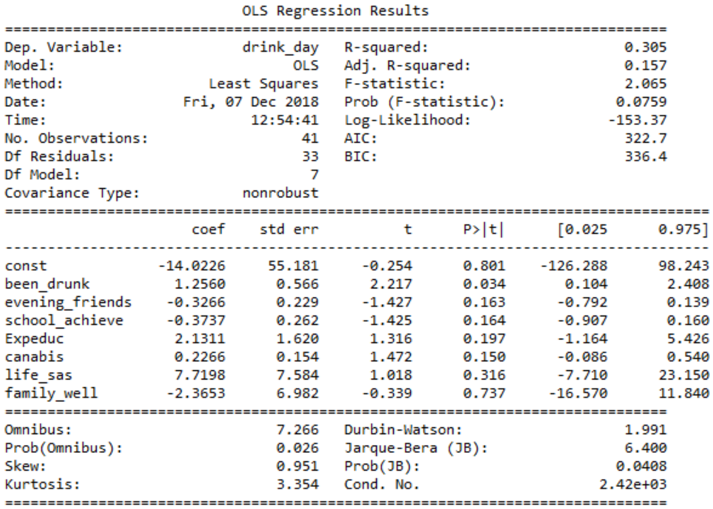
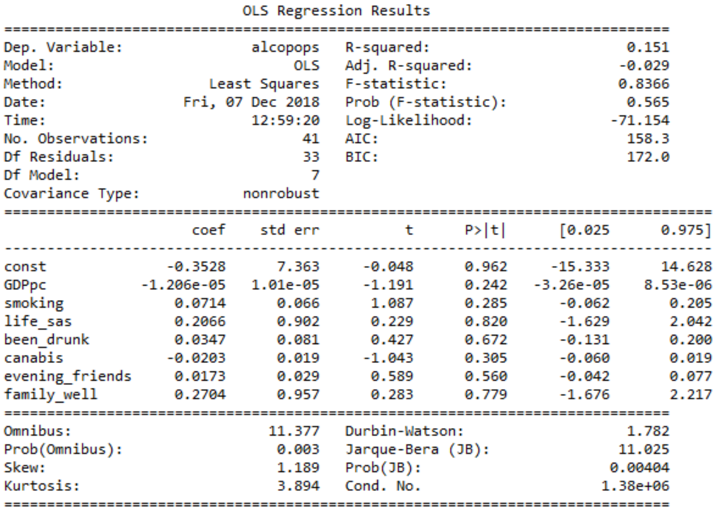


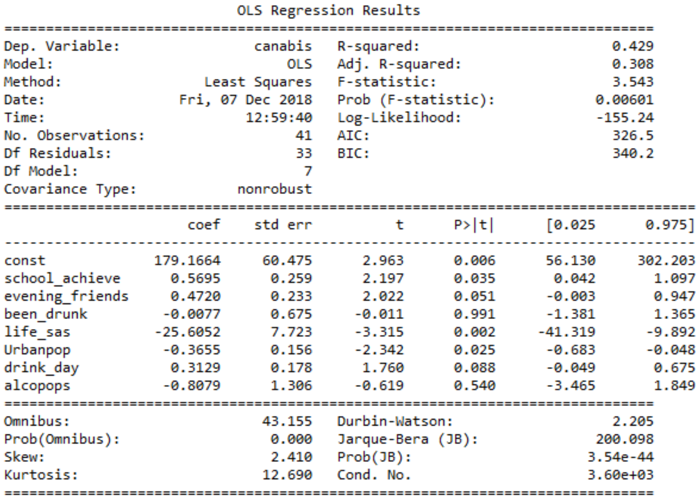
Figure 9. Variable importance based on R-squared



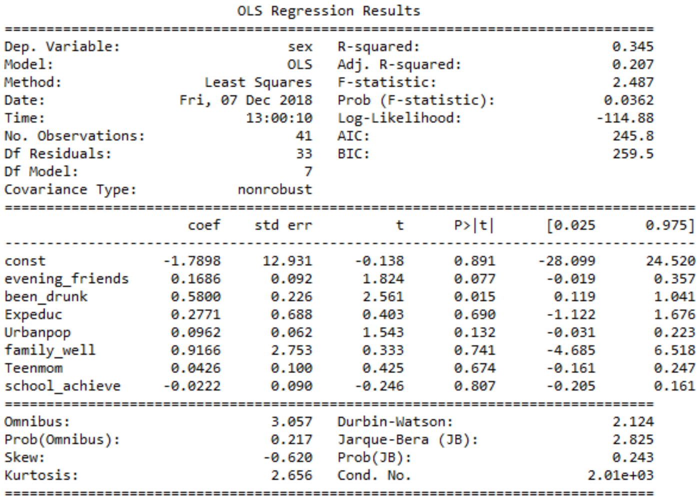
```
                            OLS Regression Results
==============================================================================
Dep. Variable:          drink_day   R-squared:                           0.305
Model:                        OLS   Adj. R-squared:                      0.157
Method:             Least Squares   F-statistic:                         2.065
Date:            Fri, 07 Dec 2018   Prob (F-statistic):                 0.0759
Time:                    12:54:41   Log-Likelihood:                    -153.37
No. Observations:              41   AIC:                                 322.7
Df Residuals:                  33   BIC:                                 336.4
Df Model:                       7
Covariance Type:        nonrobust
=================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            -14.0226     55.181     -0.254      0.801    -126.288     98.243
been_drunk         1.2560      0.566      2.217      0.034       0.104      2.408
evening_friends   -0.3266      0.229     -1.427      0.163      -0.792      0.139
school_achieve    -0.3737      0.262     -1.425      0.164      -0.907      0.160
Expeduc            2.1311      1.620      1.316      0.197      -1.164      5.426
canabis            0.2266      0.154      1.472      0.150      -0.086      0.540
life_sas           7.7198      7.584      1.018      0.316      -7.710     23.150
family_well       -2.3653      6.982     -0.339      0.737     -16.570     11.840
==============================================================================
Omnibus:                        7.266   Durbin-Watson:                       1.991
Prob(Omnibus):                  0.026   Jarque-Bera (JB):                    6.400
Skew:                           0.951   Prob(JB):                           0.0408
Kurtosis:                       3.354   Cond. No.                         2.42e+03
==============================================================================
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           alcopops   R-squared:                           0.151
Model:                        OLS   Adj. R-squared:                     -0.029
Method:             Least Squares   F-statistic:                        0.8366
Date:            Fri, 07 Dec 2018   Prob (F-statistic):                  0.565
Time:                    12:59:20   Log-Likelihood:                    -71.154
No. Observations:              41   AIC:                                 158.3
Df Residuals:                  33   BIC:                                 172.0
Df Model:                       7
Covariance Type:        nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const              -0.3528      7.363     -0.048      0.962     -15.333     14.628
GDPpc           -1.206e-05   1.01e-05     -1.191      0.242   -3.26e-05   8.53e-06
smoking             0.0714      0.066      1.087      0.285      -0.062      0.205
life_sas            0.2066      0.902      0.229      0.820      -1.629      2.042
been_drunk          0.0347      0.081      0.427      0.672      -0.131      0.200
canabis            -0.0203      0.019     -1.043      0.305      -0.060      0.019
evening_friends     0.0173      0.029      0.589      0.560      -0.042      0.077
family_well         0.2704      0.957      0.283      0.779      -1.676      2.217
==============================================================================
Omnibus:                       11.377   Durbin-Watson:                       1.782
Prob(Omnibus):                  0.003   Jarque-Bera (JB):                   11.025
Skew:                           1.189   Prob(JB):                          0.00404
Kurtosis:                       3.894   Cond. No.                         1.38e+06
==============================================================================
```



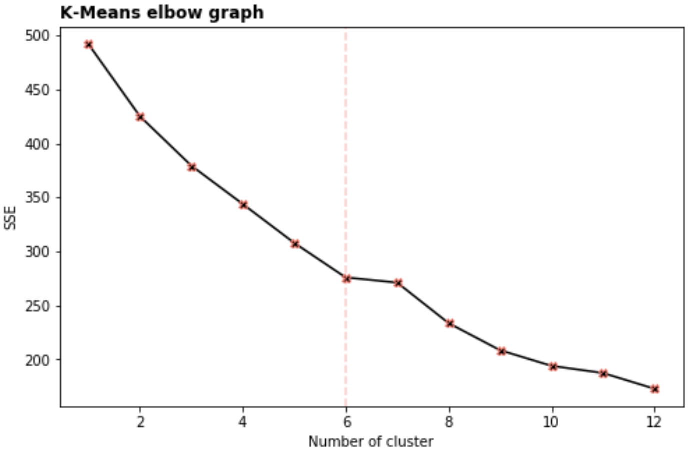
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            canabis   R-squared:                           0.429
Model:                        OLS   Adj. R-squared:                      0.308
Method:             Least Squares   F-statistic:                         3.543
Date:            Fri, 07 Dec 2018   Prob (F-statistic):                0.00601
Time:                    12:59:40   Log-Likelihood:                    -155.24
No. Observations:              41   AIC:                                 326.5
Df Residuals:                  33   BIC:                                 340.2
Df Model:                       7
Covariance Type:        nonrobust
=================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            179.1664     60.475      2.963      0.006      56.130    302.203
school_achieve     0.5695      0.259      2.197      0.035       0.042      1.097
evening_friends    0.4720      0.233      2.022      0.051      -0.003      0.947
been_drunk        -0.0077      0.675     -0.011      0.991      -1.381      1.365
life_sas         -25.6052      7.723     -3.315      0.002     -41.319     -9.892
Urbanpop          -0.3655      0.156     -2.342      0.025      -0.683     -0.048
drink_day          0.3129      0.178      1.760      0.088      -0.049      0.675
alcopops          -0.8079      1.306     -0.619      0.540      -3.465      1.849
==============================================================================
Omnibus:                       43.155   Durbin-Watson:                       2.205
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                  200.098
Skew:                           2.410   Prob(JB):                         3.54e-44
Kurtosis:                      12.690   Cond. No.                         3.60e+03
==============================================================================
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                sex   R-squared:                           0.345
Model:                        OLS   Adj. R-squared:                      0.207
Method:             Least Squares   F-statistic:                         2.487
Date:            Fri, 07 Dec 2018   Prob (F-statistic):                 0.0362
Time:                    13:00:10   Log-Likelihood:                    -114.88
No. Observations:              41   AIC:                                 245.8
Df Residuals:                  33   BIC:                                 259.5
Df Model:                       7
Covariance Type:        nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const              -1.7898     12.931     -0.138      0.891     -28.099     24.520
evening_friends     0.1686      0.092      1.824      0.077      -0.019      0.357
been_drunk          0.5800      0.226      2.561      0.015       0.119      1.041
Expeduc             0.2771      0.688      0.403      0.690      -1.122      1.676
Urbanpop            0.0962      0.062      1.543      0.132      -0.031      0.223
family_well         0.9166      2.753      0.333      0.741      -4.685      6.518
Teenmom             0.0426      0.100      0.425      0.674      -0.161      0.247
school_achieve     -0.0222      0.090     -0.246      0.807      -0.205      0.161
==============================================================================
Omnibus:                        3.057   Durbin-Watson:                       2.124
Prob(Omnibus):                  0.217   Jarque-Bera (JB):                    2.825
Skew:                          -0.620   Prob(JB):                            0.243
Kurtosis:                       2.656   Cond. No.                         2.01e+03
==============================================================================
```

Figure 10. OLS Regressions Outputs

e. PCA



Lastly, Principal Components and Cluster analysis were made in order to find out which countries share the same type of behavior and which do not. To perform this analysis, only the initial variables were considered excluding **friends_help** and **first_drunk** given their high correlation with **family_help** and **been_drunk**, respectively.

(Figure 11) Initially, a Principal Components Analysis (PCA) was made to reduce the number of variables used for clustering and to turn the interpretation easier. Analyzing the elbow graph of explained variable and taking in consideration a minimum cumulative explained variance of 80%, a total of seven principal components were consider. Using the 7 principle components, all variables were analyzed to see how much of their variations were explained by each component. The conclusion was that most of them are not explained by these components, resulting in unexplained components and requiring a higher number of them. In addition to this, assigning a meaning to these components would not be possible given that they did not explain the majority of one or two variables but a small percentage of a big group of variables. Thus, the PCA was not worth to use and the original variables were used for clustering. (see attachment 06.01)

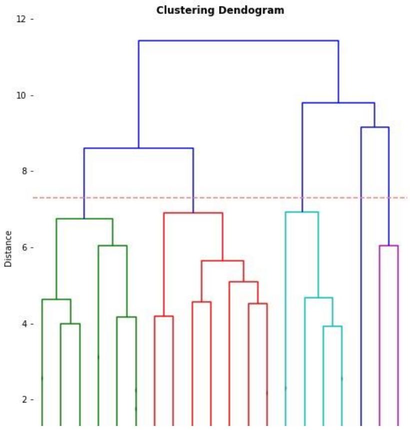


Figure 12.Hierarchical Clustering Dendrogram

(Figure 12, figure 13) Regarding the clustering analysis, four methods were made: k-means, hierarchical, expectation-maximization and k-means with the centroids of hierarchical clustering. After running the four algorithms, the k-means was the chosen one to perform the analysis. Thus, only a small introduction will be made for hierarchical clustering and then the analysis of the k-means clusters will be presented. Hierarchical clustering algorithm was run using a ward linkage and given its dendrogram, five clusters would be used to make the analysis. For the k-means method one additional cluster would be needed. This was considered important given that the two major outliers in the population would be almost alone and did not distort the centroids for other clusters. These outliers are: Macedonia, with high percentage of kids that have smoked cannabis (95.4%) and Greenland, with a high percentage of adolescents that smoke weekly or daily (25.4%), and as shown in *table 3*, cluster 4 and 5 are composed by mainly these two countries.

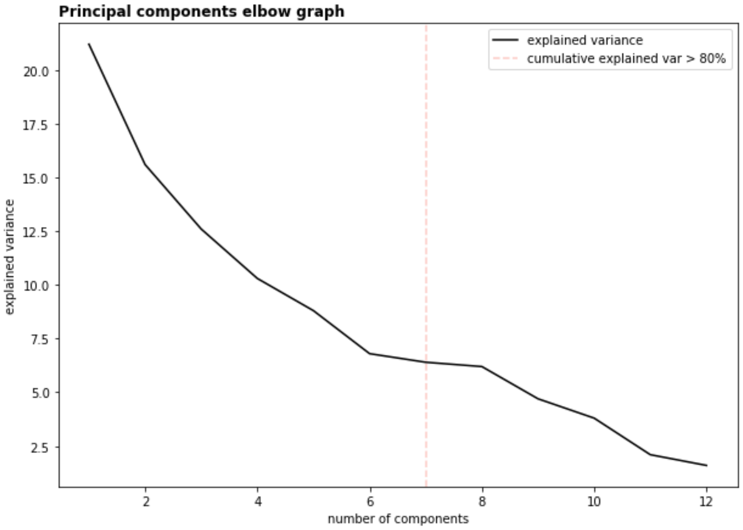


Figure 13. k-means elbow graph

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Albania	Belgium (French)	Belgium (Flemish)	Austria	Macedonia	Bulgaria
Armenia	Canada	Czech Republic	Croatia		Greenland
Estonia	Denmark	Ireland	Finland		
Iceland	England	Luxembourg	Greece		
Norway	France	Netherlands	Hungary		
Republic of Moldova	Germany	Poland	Israel		
Romania	Latvia	Portugal	Italy		
Slovakia	Wales	Russia	Malta		
Sweden		Spain	Scotland		
		Switzerland	Slovenia		
			Ukraine		

Table 3. Clusters composition

(Figure 14) The major differences between clusters are seen in **smoking**, **cannabis, sex, drink_day, alcopops** and **been_drunk.** Next, a characterization of the clusters will be made based on their centroids.

Cluster 0

The cluster with the highest average of life satisfaction, although very close to the other clusters. Also, the countries where less adolescents have tried cannabis, one of the lowest percentage of adolescents that have been drunk twice or more, that drink weekly and that drink more than 3 drinks. On average, it is the healthiest cluster.

Cluster 1

This cluster is composed mostly by countries from Northern Europe that have higher GDP and are wealthier. In this group of countries, more adolescents exercise frequently (around 56%) but they are also the ones that usually drink more when going out, almost 30% of the individuals said they would drink 3 or more drinks and have a high percentage of kids affirming that they have smoked cannabis. It is important to mention that this cluster has the lowest amount of kids going out with friends at least weekly (**evening_friends**).

Cluster 3

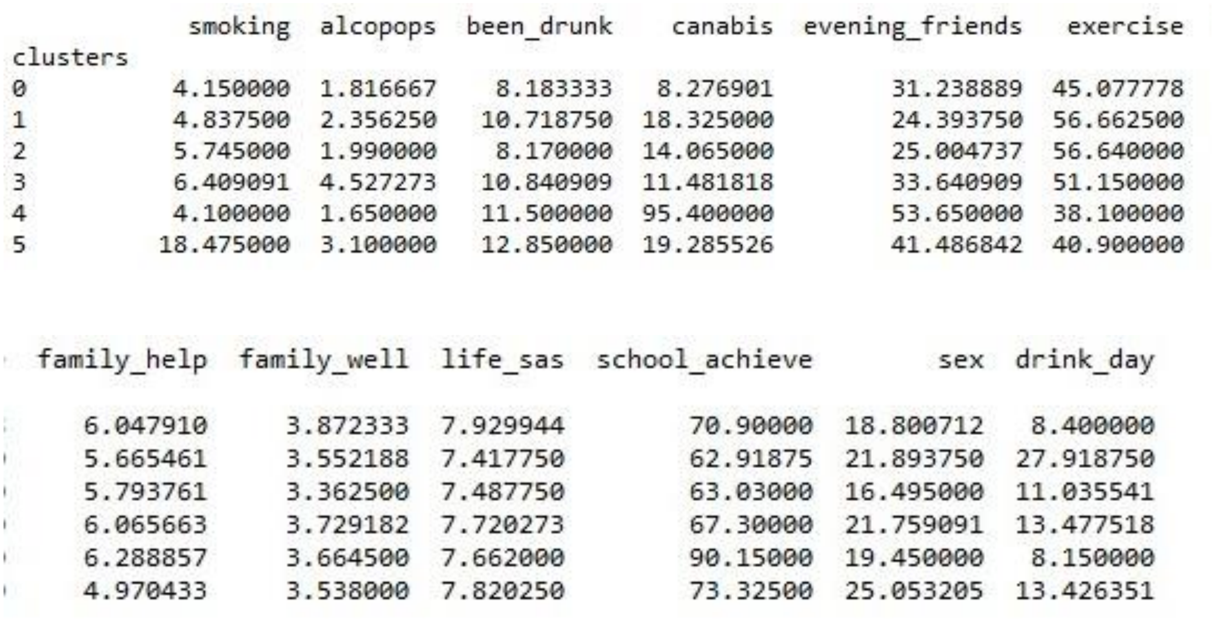
Composed by some of South Europe and countries from the Balkans, this cluster is characterized by having the highest percentage of kids that drink at least weekly, followed by having 6% of adolescents smoking. In these countries, 11% of kids have smoked cannabis and have been drunk more than twice. On the bright side, these are the countries with lower percentage of kids affirming that they had sexual relations, on average.

Cluster 4

As mentioned before, Macedonia is an outlier when it comes to the amount of kids that have smoked cannabis before and it is because of this that is a cluster by itself. Furthermore, this country also shows the highest percentage of adolescents that go out at least weekly, more than half of the individuals, and the highest percentage of students that believe they are "good" or "very good" (90%). Macedonians show the lowest average of adolescents that exercise and also one of the lowest percentage of teens that smoke weekly or daily.

Cluster 5

This cluster has only two countries and this happens due to their high percentage of adolescents that smoke at least weekly (almost 20%). Cluster 5 is also characterized by having the highest percentage of adolescent that have been drunk twice or more and 3% of the individuals drink weekly. Of these children, 25% affirm that they have had sexual relations before which is the highest cluster average. Furthermore, in these countries, kids that go out weekly represent 41%, the second largest amount. Another aspect of these countries is that they have the lowest average for **family_help** and **family_well.**



```
          smoking  alcopops  been_drunk    canabis  evening_friends   exercise
clusters
0        4.150000  1.816667    8.183333   8.276901        31.238889  45.077778
1        4.837500  2.356250   10.718750  18.325000        24.393750  56.662500
2        5.745000  1.990000    8.170000  14.065000        25.004737  56.640000
3        6.409091  4.527273   10.840909  11.481818        33.640909  51.150000
4        4.100000  1.650000   11.500000  95.400000        53.650000  38.100000
5       18.475000  3.100000   12.850000  19.285526        41.486842  40.900000
```

```
 family_help  family_well  life_sas  school_achieve        sex  drink_day

    6.047910     3.872333  7.929944        70.90000  18.800712   8.400000
    5.665461     3.552188  7.417750        62.91875  21.893750  27.918750
    5.793761     3.362500  7.487750        63.03000  16.495000  11.035541
    6.065663     3.729182  7.720273        67.30000  21.759091  13.477518
    6.288857     3.664500  7.662000        90.15000  19.450000   8.150000
    4.970433     3.538000  7.820250        73.32500  25.053205  13.426351
```

Figure 14. Cluster centroids

After analyzing the clusters, a map was created to see if there was any geographic pattern. It is shown that some clusters are geographically close but not most of the countries.

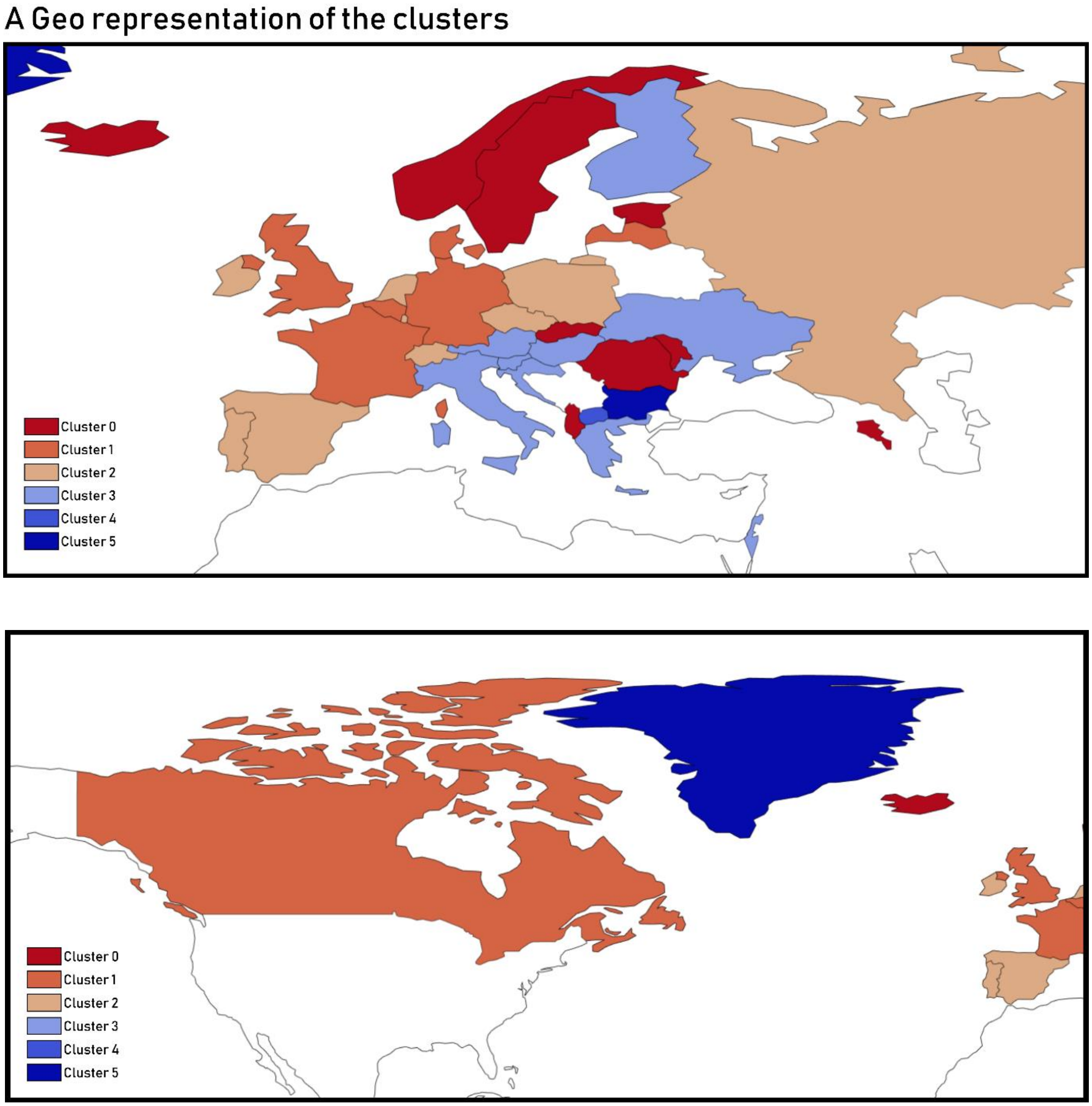


Figure 15. Geographic clusters representation

IV. Conclusions and Future objectives

An initial question was made *apriori* to all processes described in this report, which was how big a risky behavior (like alcohol drinking) present in adolescents in several world regions was and how was it affected by other behaviors or characteristics. The results discovered were beyond our predictions.

Overall, we are very satisfied with the results obtained in this project. After a rough start in finding the necessary data, a long and important process of collecting, pre-processing and merging data was made in order to fit our needs.

It was interesting to see how male adolescents' behavior of spending evening with friends was positively correlated with sex, while in female adolescents' sex was positively correlated with having been drunk at least twice. On the other hand, having been drunk at least twice is also positively correlated with adolescents that first got drunk at 13 years old or under. We discovered in which regions the gender gap was bigger in alcohol risky behaviors and which demographic characteristics were associated with each behavior. By means of regressions, we were also able to quantify the impact of some variables in others.

If we had more time, it would have been interesting to explore this subject even deeper. For example, we would like to make a similar analysis but, instead of focusing on the gender gap, to observe how risky behaviors change in ages 11, 13 and 15-years old intervals. Also, even though we could not find it, more data regarding different ages would provide a different perspective on how, in each different region, does each behavior or demographic characteristic affect a respective risky behavior.

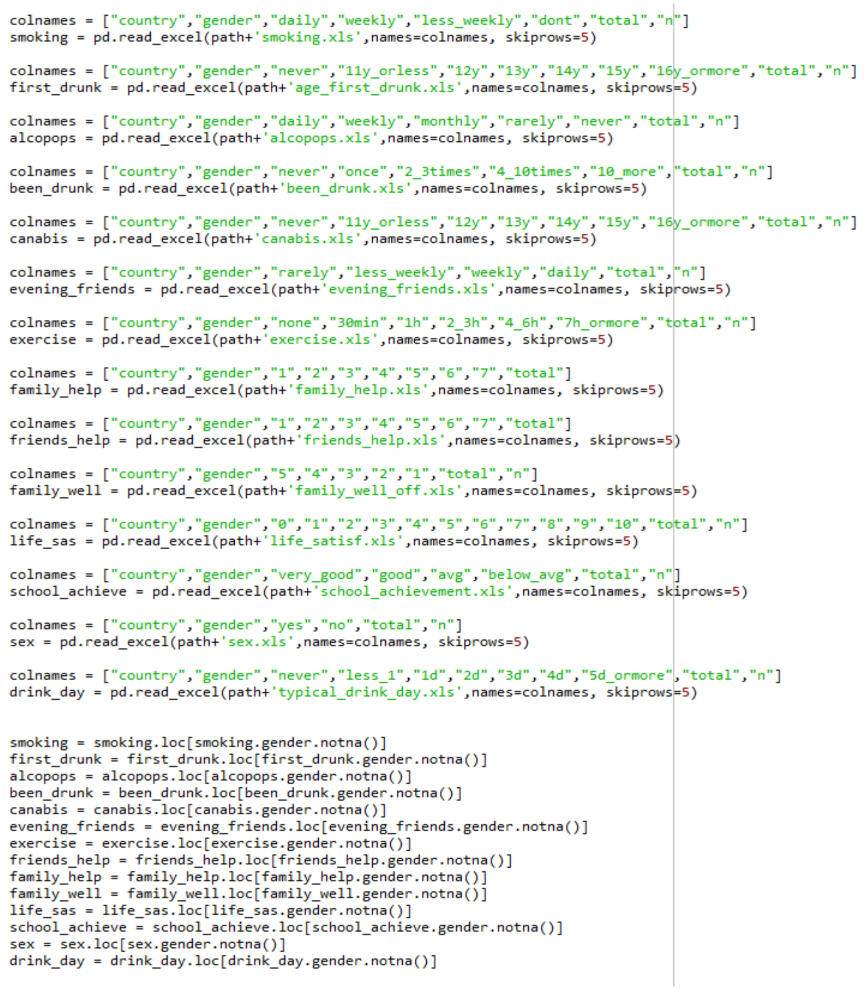
On another note, the research made on this project and the process of getting from an abstract idea to collect, transform and analyze data provided us with the necessary tools to do something similar with whatever problem we come across, which we believe to be truly the essence of Data Science.

V. References

1. Hbsc. 2013/2014. Social determinants of health and well-being among young people. Available at: https://bit.ly/2BJrDaC
2. The American Physical Society. 2011. Physical Review Style and Notation Guide. Available at: https://cdn.journals.aps.org/files/styleguide-pr.pdf

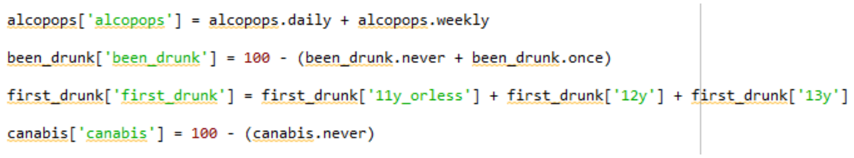
VI. Attachments

f. Importing tables



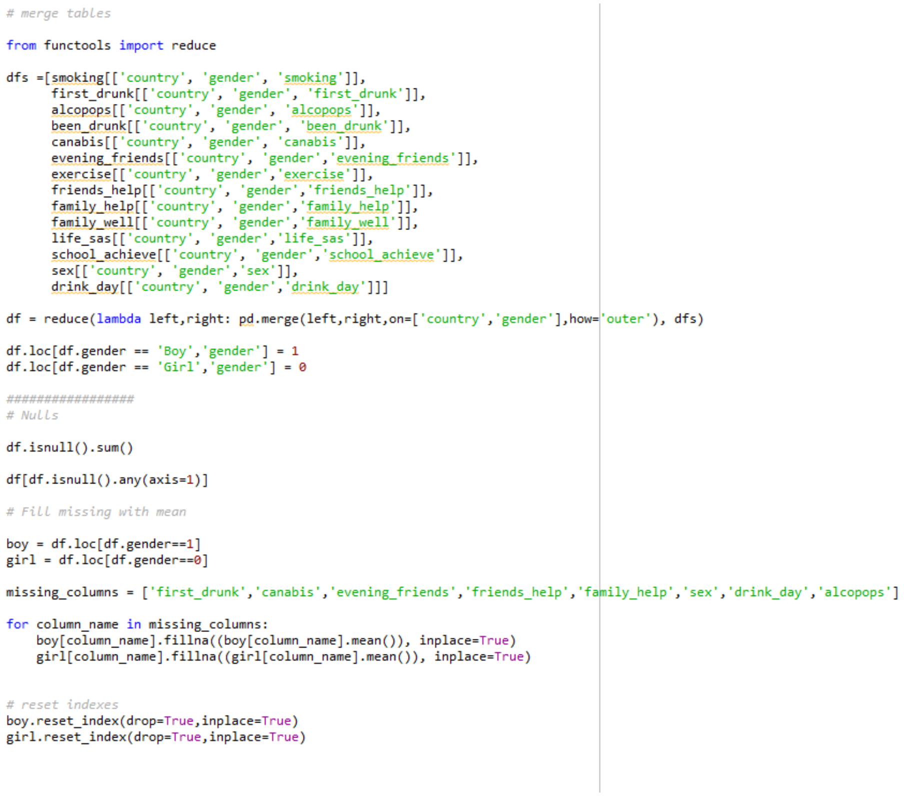
```
colnames = ["country","gender","daily","weekly","less_weekly","dont","total","n"]
smoking = pd.read_excel(path+'smoking.xls',names=colnames, skiprows=5)

colnames = ["country","gender","never","11y_orless","12y","13y","14y","15y","16y_ormore","total","n"]
first_drunk = pd.read_excel(path+'age_first_drunk.xls',names=colnames, skiprows=5)

colnames = ["country","gender","daily","weekly","monthly","rarely","never","total","n"]
alcopops = pd.read_excel(path+'alcopops.xls',names=colnames, skiprows=5)

colnames = ["country","gender","never","once","2_3times","4_10times","10_more","total","n"]
been_drunk = pd.read_excel(path+'been_drunk.xls',names=colnames, skiprows=5)

colnames = ["country","gender","never","11y_orless","12y","13y","14y","15y","16y_ormore","total","n"]
canabis = pd.read_excel(path+'canabis.xls',names=colnames, skiprows=5)

colnames = ["country","gender","rarely","less_weekly","weekly","daily","total","n"]
evening_friends = pd.read_excel(path+'evening_friends.xls',names=colnames, skiprows=5)

colnames = ["country","gender","none","30min","1h","2_3h","4_6h","7h_ormore","total","n"]
exercise = pd.read_excel(path+'exercise.xls',names=colnames, skiprows=5)

colnames = ["country","gender","1","2","3","4","5","6","7","total"]
family_help = pd.read_excel(path+'family_help.xls',names=colnames, skiprows=5)

colnames = ["country","gender","1","2","3","4","5","6","7","total"]
friends_help = pd.read_excel(path+'friends_help.xls',names=colnames, skiprows=5)

colnames = ["country","gender","5","4","3","2","1","total","n"]
family_well = pd.read_excel(path+'family_well_off.xls',names=colnames, skiprows=5)

colnames = ["country","gender","0","1","2","3","4","5","6","7","8","9","10","total","n"]
life_sas = pd.read_excel(path+'life_satisf.xls',names=colnames, skiprows=5)

colnames = ["country","gender","very_good","good","avg","below_avg","total","n"]
school_achieve = pd.read_excel(path+'school_achievement.xls',names=colnames, skiprows=5)

colnames = ["country","gender","yes","no","total","n"]
sex = pd.read_excel(path+'sex.xls',names=colnames, skiprows=5)

colnames = ["country","gender","never","less_1","1d","2d","3d","4d","5d_ormore","total","n"]
drink_day = pd.read_excel(path+'typical_drink_day.xls',names=colnames, skiprows=5)


smoking = smoking.loc[smoking.gender.notna()]
first_drunk = first_drunk.loc[first_drunk.gender.notna()]
alcopops = alcopops.loc[alcopops.gender.notna()]
been_drunk = been_drunk.loc[been_drunk.gender.notna()]
canabis = canabis.loc[canabis.gender.notna()]
evening_friends = evening_friends.loc[evening_friends.gender.notna()]
exercise = exercise.loc[exercise.gender.notna()]
friends_help = friends_help.loc[friends_help.gender.notna()]
family_help = family_help.loc[family_help.gender.notna()]
family_well = family_well.loc[family_well.gender.notna()]
life_sas = life_sas.loc[life_sas.gender.notna()]
school_achieve = school_achieve.loc[school_achieve.gender.notna()]
sex = sex.loc[sex.gender.notna()]
drink_day = drink_day.loc[drink_day.gender.notna()]
```

g. Aggregating columns into one (example)



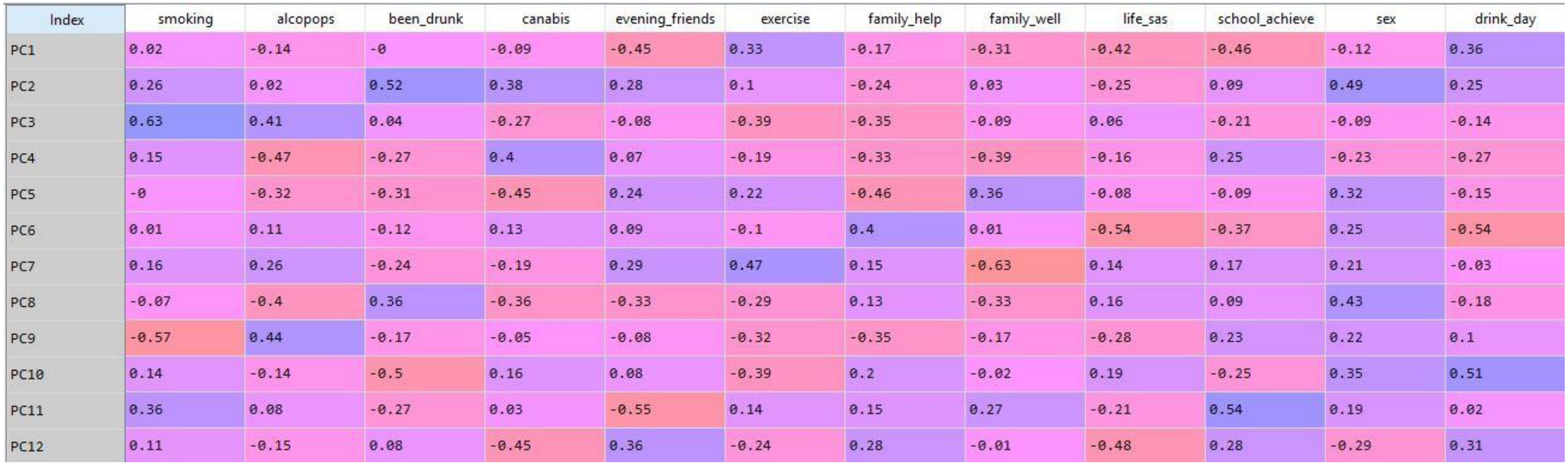
```
alcopops['alcopops'] = alcopops.daily + alcopops.weekly

been_drunk['been_drunk'] = 100 - (been_drunk.never + been_drunk.once)

first_drunk['first_drunk'] = first_drunk['11y_orless'] + first_drunk['12y'] + first_drunk['13y']

canabis['canabis'] = 100 - (canabis.never)
```

h. Creating a DataFrame with required variables and separating by gender



```
# merge tables

from functools import reduce

dfs =[smoking[['country', 'gender', 'smoking']],
      first_drunk[['country', 'gender', 'first_drunk']],
      alcopops[['country', 'gender', 'alcopops']],
      been_drunk[['country', 'gender', 'been_drunk']],
      canabis[['country', 'gender', 'canabis']],
      evening_friends[['country', 'gender','evening_friends']],
      exercise[['country', 'gender','exercise']],
      friends_help[['country', 'gender','friends_help']],
      family_help[['country', 'gender','family_help']],
      family_well[['country', 'gender','family_well']],
      life_sas[['country', 'gender','life_sas']],
      school_achieve[['country', 'gender','school_achieve']],
      sex[['country', 'gender','sex']],
      drink_day[['country', 'gender','drink_day']]]

df = reduce(lambda left,right: pd.merge(left,right,on=['country','gender'],how='outer'), dfs)

df.loc[df.gender == 'Boy','gender'] = 1
df.loc[df.gender == 'Girl','gender'] = 0

#################
# Nulls

df.isnull().sum()

df[df.isnull().any(axis=1)]

# Fill missing with mean

boy = df.loc[df.gender==1]
girl = df.loc[df.gender==0]

missing_columns = ['first_drunk','canabis','evening_friends','friends_help','family_help','sex','drink_day','alcopops']

for column_name in missing_columns:
    boy[column_name].fillna((boy[column_name].mean()), inplace=True)
    girl[column_name].fillna((girl[column_name].mean()), inplace=True)


# reset indexes
boy.reset_index(drop=True,inplace=True)
girl.reset_index(drop=True,inplace=True)
```

i. Explained variance of each variable in each Principal Component



Index	smoking	alcopops	been_drunk	canabis	evening_friends	exercise	family_help	family_well	life_sas	school_achieve	sex	drink_day
PC1	0.02	-0.14	-0	-0.09	-0.45	0.33	-0.17	-0.31	-0.42	-0.46	-0.12	0.36
PC2	0.26	0.02	0.52	0.38	0.28	0.1	-0.24	0.03	-0.25	0.09	0.49	0.25
PC3	0.63	0.41	0.04	-0.27	-0.08	-0.39	-0.35	-0.09	0.06	-0.21	-0.09	-0.14
PC4	0.15	-0.47	-0.27	0.4	0.07	-0.19	-0.33	-0.39	-0.16	0.25	-0.23	-0.27
PC5	-0	-0.32	-0.31	-0.45	0.24	0.22	-0.46	0.36	-0.08	-0.09	0.32	-0.15
PC6	0.01	0.11	-0.12	0.13	0.09	-0.1	0.4	0.01	-0.54	-0.37	0.25	-0.54
PC7	0.16	0.26	-0.24	-0.19	0.29	0.47	0.15	-0.63	0.14	0.17	0.21	-0.03
PC8	-0.07	-0.4	0.36	-0.36	-0.33	-0.29	0.13	-0.33	0.16	0.09	0.43	-0.18
PC9	-0.57	0.44	-0.17	-0.05	-0.08	-0.32	-0.35	-0.17	-0.28	0.23	0.22	0.1
PC10	0.14	-0.14	-0.5	0.16	0.08	-0.39	0.2	-0.02	0.19	-0.25	0.35	0.51
PC11	0.36	0.08	-0.27	0.03	-0.55	0.14	0.15	0.27	-0.21	0.54	0.19	0.02
PC12	0.11	-0.15	0.08	-0.45	0.36	-0.24	0.28	-0.01	-0.48	0.28	-0.29	0.31