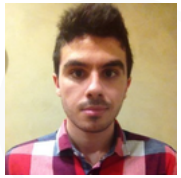


Jose Ignacio Hervás

I'm a data analytics fanatic. I belive that we could see every aspect of our daily lifes represented thought graphs, and that this could help us to understand a bit better our world.

ARCHIVE

(Data management and Visualization) 4th assignment: Creating graphs for my data

This is my fourth and last assignment for the Wesleyan University's program on "Data management and Visualization":

In this assignment, my classmates and me were asked to examinate the distributions of our variables by univariate graphs.

In my project, I'm testing if there are any differences between the alcohol consumption ratio of the tennagers whose parents let them make their own decisions about the people they hang around with and whose parent's don't do it.

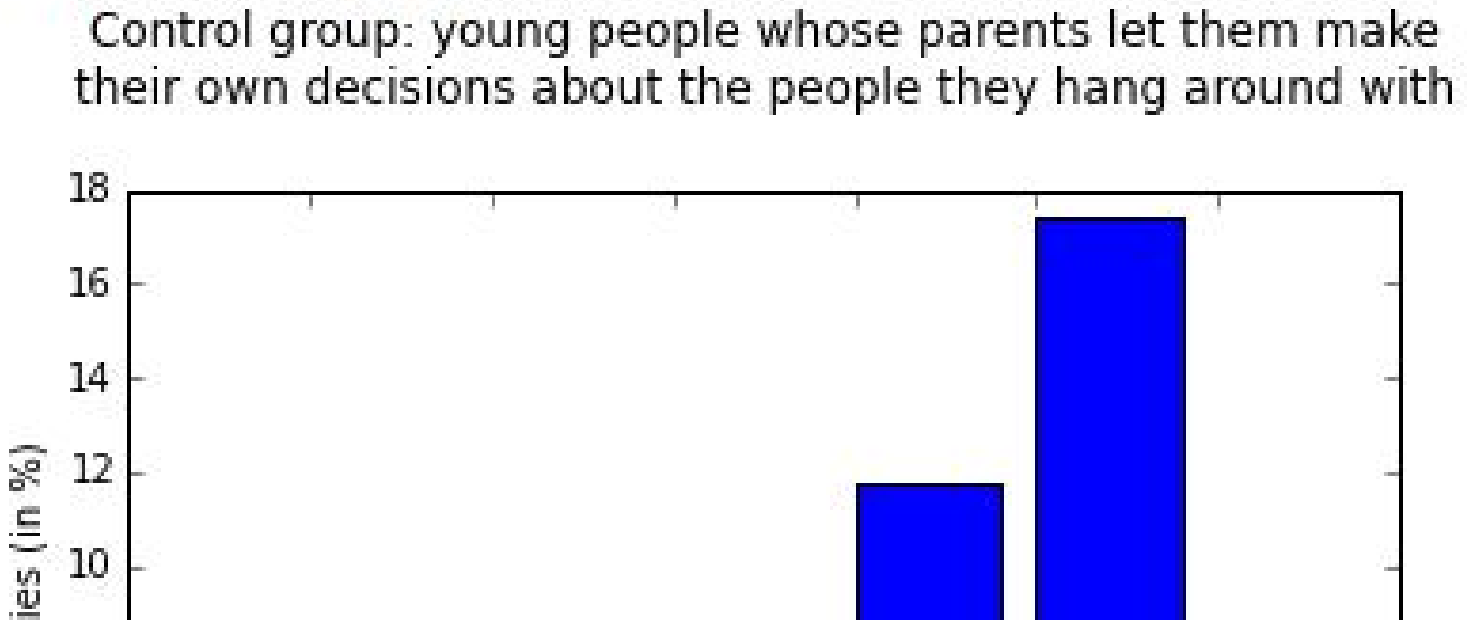
My program looks like this:

```
subgroup2[“H1TO15”]=subgroup2[“H1TO15”].replace(96,np.nan)

subgroup2[“H1TO15”]=subgroup2[“H1TO15”].replace(97,np.nan)

subgroup2[“H1TO15”]=subgroup2[“H1TO15”].replace(98,np.nan)

p8=subgroup2[“H1TO15”].value_counts(sort=False,normalize=True,dropna=False)

p8=p8*100

plt.bar(p8.index,p8)

plt.title(‘Control group: young people whose parents let them make’ + ’\n’ + 'their own decisions about the people they hang around with’,y=1.08)

plt.xlabel('Number of days of alcohol consumption’)

plt.ylabel('Frequencies (in %)’)

p8.describe()

subgroup1[“H1TO15”]=subgroup1[“H1TO15”].replace(96,np.nan)

subgroup1[“H1TO15”]=subgroup1[“H1TO15”].replace(97,np.nan)

subgroup1[“H1TO15”]=subgroup1[“H1TO15”].replace(98,np.nan)

p3 = subgroup1[“H1TO15”].value_counts(sort=False, normalize=True, dropna=False)

p3=p3*100

plt.bar(p3.index,p3)

plt.title('Study group: young people whose parents do not let them make’ + ’\n’ + 'their own decisions about the people they hang around with’,y=1.08)
```

```
plt.xlabel('Number of days of alcohol consumption')

plt.ylabel('Frequencies (in %)')

p3.describe()
```

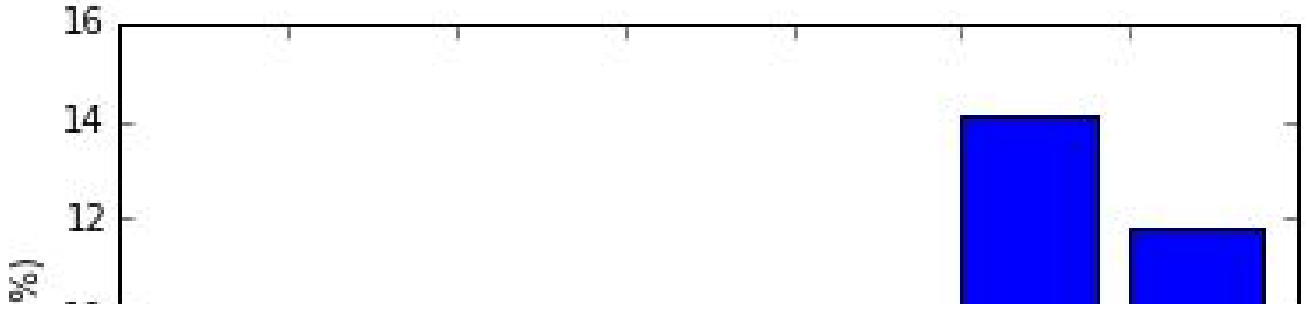
The output of my program are two graphs and two description tables:

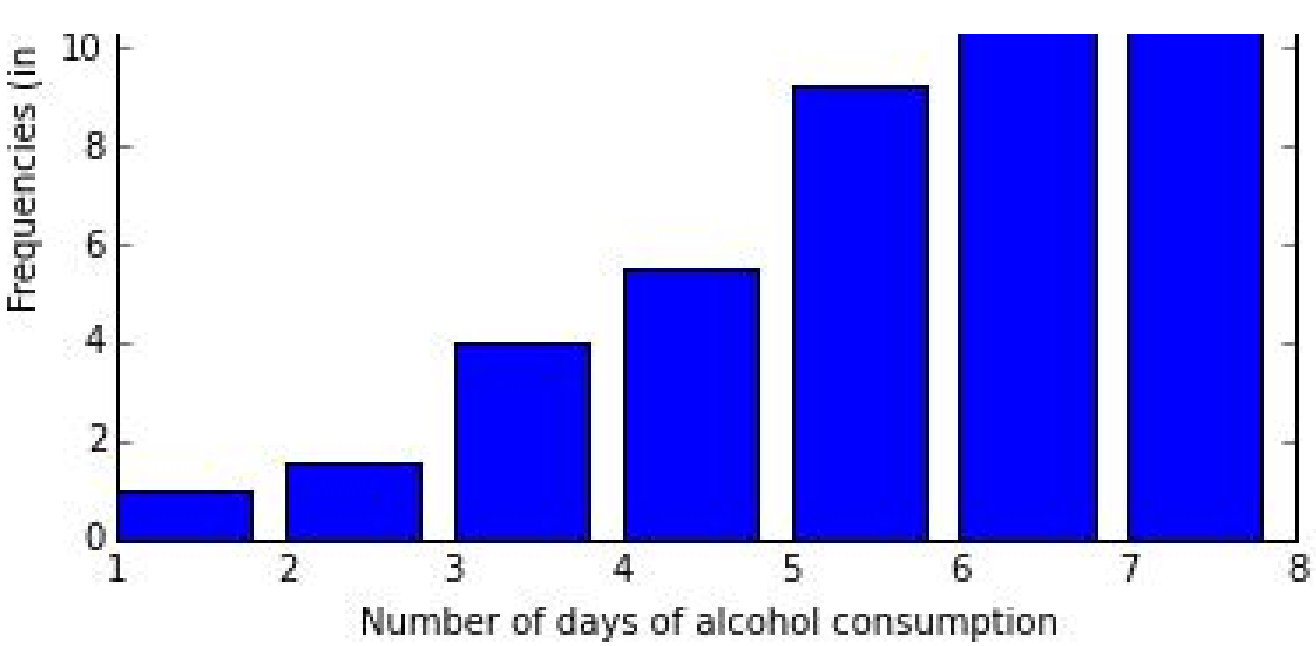
```
count      8.000000
mean      12.500000
std       13.911324
min        0.848708
25%        5.327491
50%        8.311808
75%       13.178044
max       44.483395
dtype: float64
```




```
count      8.000000
mean      12.500000
std       16.933949
min        0.955414
25%        3.423567
50%        7.377919
75%       12.367304
max       52.760085
dtype: float64
```

Study group: young people whose parents do not let them make their own decisions about the people they hang around with





COMMENTS: As we can see on the bar charts, the alcohol consumption ratio of the young people in the study and the control group are not very much different from each other. The mean is the same for the both distributions, althought the standard deviation is slightly bigger on the study group, owing to the fact that in the control group there are significantly more people on the 6th value (almost a 4% more of people that in the study group).

May 15th, 2016

My 3th assignment for the Data Management and Visualization course.

Hi again! In the third week of this course, my coursemates and me were asked to use Python for encoding and processing data. This week I'm going to continue using the Addhealth dataset with the variables H1WP2 and H1WP2. Here's my code:

Getting my research project started

Hello! My name is Jose, I'm 21 years old, and I'm from Spain. I work as data analyst for a Market Research company in Barcelona, so my daily life is closely related to the analysis and visualization of data. So I decided to start the "Data Analysis and Interpretation" specialization on the Wesleyan University and learn everything about this field. This is my first assignment for the Data Management and Visualization program.

In this assignment, me and my classmates have been asked to find our own research question. After taking a look at a large list of datasets, I have decided that I would like to study the math achievements of the kids around the world, and, particularly, I am

My first analytics program with Python

Today I'm really excited. I'm going to share my first script for statistic's analysis of data, as part of my Data Management and Visualization course.

In this second assignment, me and my coursemates have been asked to create an analytic script with Python and use it over some frequency distrbutions.

Unfortunately, today I'm not able to access the dataset that I wanted to use on my research. As I explained in my first post, I wanted to use the PISA dataset with the students' math results, but when I woke up this morning, the website in which I had planned to download the dataset wasn't available (the website is still down at this moment), so I'm forced to look for another alternative dataset to do this assignment.

```
import pandas as pd

data =
pd.read_csv('addhealth_pds.csv',lo
w_memory=False)

data["H1WP2"]=data["H1WP2"].con
vert_objects(convert_numeric=True
)
data["H1TO1"]=data["H1TO1"].conv
ert_objects(convert_numeric=True)

print 'Global data'
case = len(data)
print 'Total nº cases: ' + str(case)

print
'_____
_____'
print
'_____
_____'

print 'Do your parents let you make
your own decisions about the
people you hang around with?
(values in %)'
p1 = data["H1WP2"]
codesp1=
{1:"Yes",0:"No",6:"refused",7:"legitim
ate skip",8:"don't know",9:"not
```

interested in the association between the maths results of the kids in school and the self-concept that each kid has.

Some important psychologists, like Robert W. Lent, Steven D. Brown and Kevin C. Larkin (1986) have studied the relations between the perceived self-efficacy of the kids in school and their perfomance in their proffesional career in the future. Other psychologists, like Frank Pajares and Laura Graham (1999) have also determinated that the level of self-efficacy of the kids in school may affect their perfomance in maths at some level. I would like to continue with this investigations, and for that purpose, I will use the free dataset of the PISA tests.

This dataset is a recopilation of a large sum of variables related to the results of the kids in many different countries on the PISA test. For my investigation, I have decided to use the variables that in some way form part of the self-concepts of the kids in the scholar enviorement (like the self-efficacy in maths and their self-confidence with them). I have taken a look at the

Anyway, this is what my script looks like:

```
applicable"}
print
'_____
_____'
p1=p1.map(codesp1)
p1 = p1.value_counts(sort=False,
normalize=True, dropna=False)
p1=p1*100
print p1
print
'_____
_____'
print
'_____
_____'

print 'Control group: young people
whose parents let them make their
own decisions about the people
they hang around with'
subgroup3 =
data[data["H1WP2"]==1]
print '————————————o
————————————-'

print 'Have you ever tried cigarette
smoking? (values in %)'
print
'_____
_____'
```

codebook for the PISA dataset, and I have taken the variables in which I'm interested. This is going to be my personal codebook for this research. This personal codebook will be uploaded attached with this post.

According to the ideas of Pajares and Graham (1999), my initial hypotesis will be that the self-concept of the school kids is directly related to their results in maths. On the next weeks I will try to test if this relation is real, and which forms it could take. The results of this research could show us the importance that the self-steem factor plays in the scholar enviorement, and the importance of the teachers to maximize this psychological factor in order to help our childs to develop all their academical and personal potential.

Thanks for joining me on this exciting travel,

See you soon!

References:

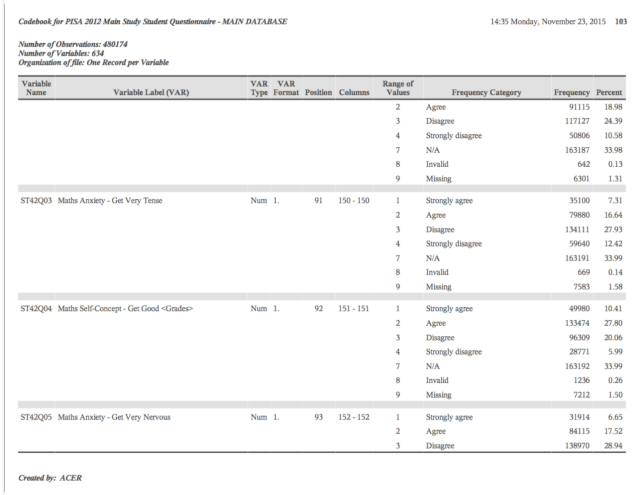
```
import pandas as pd

data =
pd.read_csv('addhealth_pds.csv',lo
w_memory=False)

data =
pd.read_csv('addhealth_pds.csv',lo
w_memory=False)data["H1WP2"]=
data["H1WP2"].convert_objects(con
vert_numeric=True)

data["H1TO1"]=data["H1TO1"].conv
ert_objects(convert_numeric=True)
data["H1TO15"]=data["H1TO15"].co
nvert_objects(convert_numeric=Tru
e)
data["H1TO41"]=data["H1TO41"].co
nvert_objects(convert_numeric=Tru
e)

print 'Number os cases: ' +
str(len(data))

print 'Do your parents let you make
your own decisions about the
people you hang around with? 1 =
yes'
```

```
p3 = subgroup3[“H1TO1”]
codesp3=
{1:“Yes”,0:“No”,6:“refused”,8:“don’t
know”,9:“not applicable”}
p3=p3.map(codesp3)
p3 = p3.value_counts(sort=False,
normalize=True, dropna=False)
p3=p3*100
print p3
print
’_____
_____’
print
’_____
_____’

print 'Study group: young people
whose parents do not let them make
their own decisions about the
people they hang around with’
subgroup1 =
data[data[“H1WP2”]==0]
print ’————————o
———————_’

print 'Have you ever tried cigarette
smoking? (values in %)’
print
’_____
_____’
```

“Self-efficacy in the prediction of academic performance and perceived career options”. Lent, Robert W.; Brown, Steven D.; Larkin, Kevin C. Journal of Counseling Psychology, Vol 33(3), Jul 1986, 265-269.

“Self-Efficacy, Motivation Constructs, and Mathematics Performance of Entering Middle School Students”. Pajares, Fank & Graham, Laura. Contemporary Educational Psychology. Volume 24, Issue 2, April 1999, Pages 124–139.

Codebook:



Codebook for PISA 2012 Main Study Student Questionnaire - MAIN DATABASE 14:35 Monday, November 23, 2015 103

Number of Observations: 480174
Number of Variables: 634
Organization of file: One Record per Variable

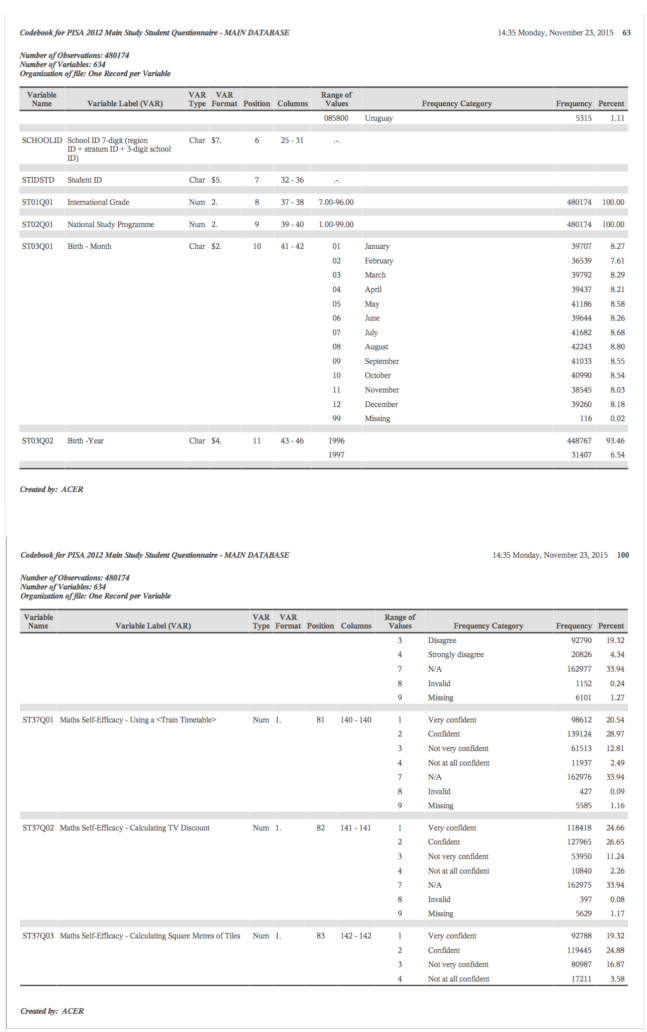
Variable Name	Variable Label (VAR)	VAR Type	VAR Format	Position	Columns	Range of Values	Frequency Category	Frequency	Percent
						2	Agree	91115	18.98
						3	Disagree	117127	24.39
						4	Strongly disagree	50806	10.58
						7	N/A	163187	33.98
						8	Invalid	642	0.13
						9	Missing	6301	1.31
ST42Q03	Maths Anxiety - Get Very Tense	Num	1.	91	150 - 150	1	Strongly agree	35100	7.31
						2	Agree	79880	16.64
						3	Disagree	134111	27.93
						4	Strongly disagree	59640	12.42
						7	N/A	163191	33.99
						8	Invalid	669	0.14
						9	Missing	7583	1.58
ST42Q04	Maths Self-Concept - Get Good <Grades>	Num	1.	92	151 - 151	1	Strongly agree	49980	10.41
						2	Agree	133474	27.80
						3	Disagree	96309	20.06
						4	Strongly disagree	28771	5.99
						7	N/A	163192	33.99
						8	Invalid	1236	0.26
						9	Missing	7212	1.50
ST42Q05	Maths Anxiety - Get Very Nervous	Num	1.	93	152 - 152	1	Strongly agree	31914	6.65
						2	Agree	84115	17.52
						3	Disagree	138970	28.94

Created by: ACER

```
p1 =
data[“H1WP2”].value_counts(sort=
False, normalize=True,
dropna=False)
print p1

print
’_____
_____’

print 'Study group: young people
whose parents do not let them make
their own decisions about the
people they hang around with’

subgroup1 =
data[data[“H1WP2”]==0]

print
’_____
_____’

print 'Have you ever tried cigarette
smoking? 1 = yes’
p2 =
subgroup1[“H1TO1”].value_counts(
sort=False, normalize=True,
dropna=False)
print p2
```

```
p2 = subgroup1[“H1TO1”]
codesp2=
{1:“Yes”,0:“No”,6:“refused”,8:“don’t
know”,9:“not applicable”}
p2=p2.map(codesp2)
p2 = p2.value_counts(sort=False,
normalize=True, dropna=False)
p2=p2*100
print p2
```

And this is the resulting data:



Codebook for PISA 2012 Main Study Student Questionnaire - MAIN DATABASE 14:35 Monday, November 23, 2015 63

Number of Observations: 480174
Number of Variables: 634
Organization of file: One Record per Variable

Variable Name	Variable Label (VAR)	VAR Type	VAR Format	Position	Columns	Range of Values	Frequency Category	Frequency	Percent
						085800	Uruguay	5315	1.11
SCHOOLID	School ID 7-digit (region ID + stratum ID + 3-digit school ID)	Char	$7.	6	25 - 31	.-			
STIDSTD	Student ID	Char	$5.	7	32 - 36	.-			
ST01Q01	International Grade	Num	2.	8	37 - 38	7.00-96.00		480174	100.00
ST02Q01	National Study Programme	Num	2.	9	39 - 40	1.00-99.00		480174	100.00
ST03Q01	Birth - Month	Char	$2.	10	41 - 42	01	January	39707	8.27
						02	February	36539	7.61
						03	March	39792	8.29
						04	April	39437	8.21
						05	May	41186	8.58
						06	June	39644	8.26
						07	July	41682	8.68
						08	August	42243	8.80
						09	September	41033	8.55
						10	October	40990	8.54
						11	November	38545	8.03
						12	December	39260	8.18
						99	Missing	116	0.02
ST03Q02	Birth -Year	Char	$4.	11	43 - 46	1996		448767	93.46
						1997		31407	6.54

Created by: ACER

Codebook for PISA 2012 Main Study Student Questionnaire - MAIN DATABASE 14:35 Monday, November 23, 2015 100

Number of Observations: 480174
Number of Variables: 634
Organization of file: One Record per Variable

Variable Name	Variable Label (VAR)	VAR Type	VAR Format	Position	Columns	Range of Values	Frequency Category	Frequency	Percent
						3	Disagree	92790	19.32
						4	Strongly disagree	20826	4.34
						7	N/A	162977	33.94
						8	Invalid	1152	0.24
						9	Missing	6101	1.27
ST37Q01	Maths Self-Efficacy - Using a <Train Timetable>	Num	1.	81	140 - 140	1	Very confident	98612	20.54
						2	Confident	139124	28.97
						3	Not very confident	61513	12.81
						4	Not at all confident	11937	2.49
						7	N/A	162976	33.94
						8	Invalid	427	0.09
						9	Missing	5585	1.16
ST37Q02	Maths Self-Efficacy - Calculating TV Discount	Num	1.	82	141 - 141	1	Very confident	118418	24.66
						2	Confident	127965	26.65
						3	Not very confident	53950	11.24
						4	Not at all confident	10840	2.26
						7	N/A	162975	33.94
						8	Invalid	397	0.08
						9	Missing	5629	1.17
ST37Q03	Maths Self-Efficacy - Calculating Square Metres of Tiles	Num	1.	83	142 - 142	1	Very confident	92788	19.32
						2	Confident	119445	24.88
						3	Not very confident	80987	16.87
						4	Not at all confident	17211	3.58

Created by: ACER

```
print 'During the past 12 months, on
how many days did you drink
alcohol?’
p3 =
subgroup1[“H1TO15”].value_counts
(sort=False, normalize=True,
dropna=False)
print p3

print
'Duringyourlife,howmanytimeshavey
ouusedanyofthesetypes of illegal
drugs?’
p4 =
subgroup1[“H1TO41”].value_counts
(sort=False, normalize=True,
dropna=False)
print p4
```

The output resulting from this script over my AddHealth database is the next:

Global data
Total nº cases: 6504

———————————————————

———

———————————————————

———

Do your parents let you make your own decisions about the people you hang around with? (values in %)

———————————————————

———

not applicable 0.015375
Yes 83.333333
refused 0.046125
No 14.483395
don’t know 0.107626
legitimate skip 2.014145
Name: H1WP2, dtype: float64

———————————————————

———

———————————————————

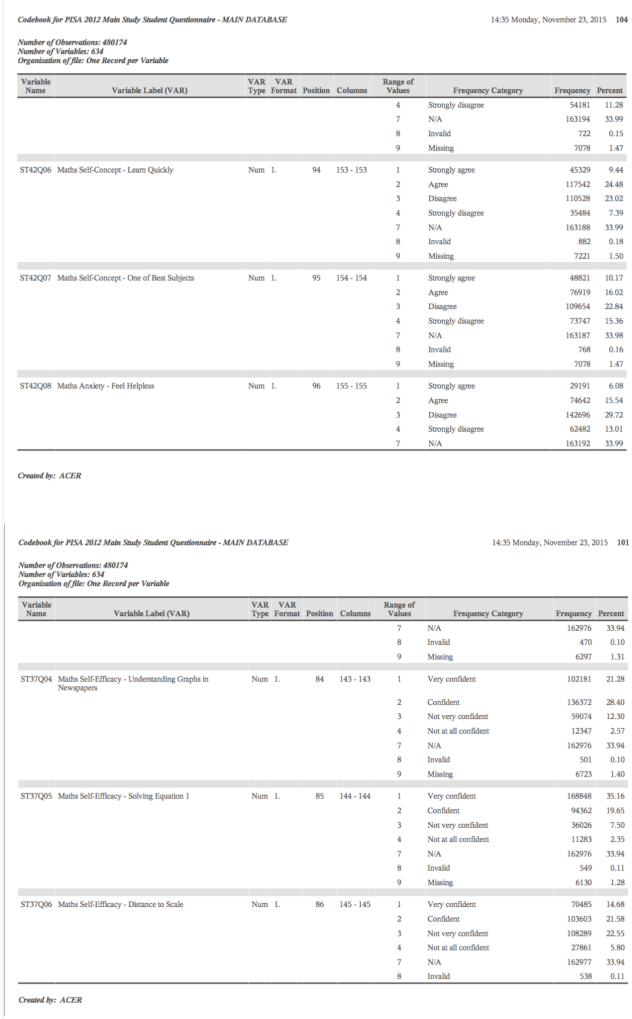
———

Control group: young people whose parents let them make their own decisions about the people they hang around with

————————o

———————————-

Have you ever tried cigarette



Codebook for PISA 2012 Main Study Student Questionnaire - MAIN DATABASE 14:35 Monday, November 23, 2015 **104**

Number of Observations: 480174
Number of Variables: 634
Organization of file: One Record per Variable

Variable Name	Variable Label (VAR)	VAR Type	VAR Format	Position	Columns	Range of Values	Frequency Category	Frequency	Percent
						4	Strongly disagree	54181	11.28
						7	N/A	163194	33.99
						8	Invalid	722	0.15
						9	Missing	7078	1.47
ST42Q06	Maths Self-Concept - Learn Quickly	Num	1.	94	153 - 153	1	Strongly agree	45329	9.44
						2	Agree	117542	24.48
						3	Disagree	110528	23.02
						4	Strongly disagree	35484	7.39
						7	N/A	163188	33.99
						8	Invalid	882	0.18
						9	Missing	7221	1.50
ST42Q07	Maths Self-Concept - One of Best Subjects	Num	1.	95	154 - 154	1	Strongly agree	48821	10.17
						2	Agree	76919	16.02
						3	Disagree	109654	22.84
						4	Strongly disagree	73747	15.36
						7	N/A	163187	33.98
						8	Invalid	768	0.16
						9	Missing	7078	1.47
ST42Q08	Maths Anxiety - Feel Helpless	Num	1.	96	155 - 155	1	Strongly agree	29191	6.08
						2	Agree	74642	15.54
						3	Disagree	142696	29.72
						4	Strongly disagree	62482	13.01
						7	N/A	163192	33.99

Created by: ACER

Codebook for PISA 2012 Main Study Student Questionnaire - MAIN DATABASE 14:35 Monday, November 23, 2015 **101**

Number of Observations: 480174
Number of Variables: 634
Organization of file: One Record per Variable

Variable Name	Variable Label (VAR)	VAR Type	VAR Format	Position	Columns	Range of Values	Frequency Category	Frequency	Percent
						7	N/A	162976	33.94
						8	Invalid	470	0.10
						9	Missing	6297	1.31
ST37Q04	Maths Self-Efficacy - Understanding Graphs in Newspapers	Num	1.	84	143 - 143	1	Very confident	102181	21.28
						2	Confident	136372	28.40
						3	Not very confident	59074	12.30
						4	Not at all confident	12347	2.57
						7	N/A	162976	33.94
						8	Invalid	501	0.10
						9	Missing	6723	1.40
ST37Q05	Maths Self-Efficacy - Solving Equation 1	Num	1.	85	144 - 144	1	Very confident	168848	35.16
						2	Confident	94362	19.65
						3	Not very confident	36026	7.50
						4	Not at all confident	11283	2.35
						7	N/A	162976	33.94
						8	Invalid	549	0.11
						9	Missing	6130	1.28
ST37Q06	Maths Self-Efficacy - Distance to Scale	Num	1.	86	145 - 145	1	Very confident	70485	14.68
						2	Confident	103603	21.58
						3	Not very confident	108289	22.55
						4	Not at all confident	27861	5.80
						7	N/A	162977	33.94
						8	Invalid	538	0.11

Created by: ACER

Number of cases: 6504
Do your parents let you make your own decisions about the people you hang around with?
1 = yes
0 0.144834
8 0.001076
9 0.000154
6 0.000461
7 0.020141

———————————————————

———

Control group: young people whose parents let them make their own decisions about the people they hang around with

———————————————————

———

Have you ever tried cigarette smoking?
1 = yes
0 0.434871
8 0.001476
1 0.559594
6 0.004059

During the past 12 months, on how many days did you drink alcohol?

smoking? (values in %)

———————————————————————

———

Yes 55.959410
refused 0.405904
No 43.487085
don’t know 0.147601
Name: H1TO1, dtype: float64

———————————————————————

———

———————————————————————

———

Study group: young people whose parents do not let them make their own decisions about the people they hang around with

————————————o

————————————-

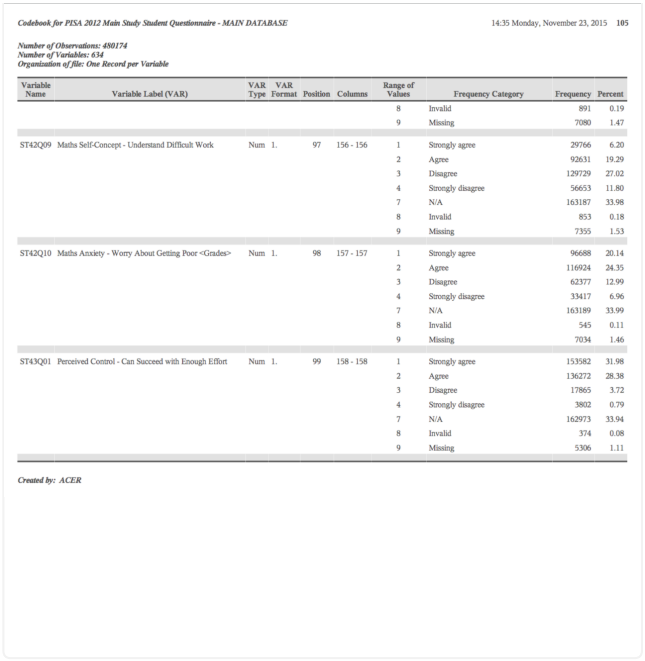
Have you ever tried cigarette smoking? (values in %)

———————————————————————

———

No 49.787686
refused 0.636943
don’t know 0.424628
Yes 49.150743
Name: H1TO1, dtype: float64

Comments:



Codebook for PISA 2012 Main Study Student Questionnaire - MAIN DATABASE 14:35 Monday, November 23, 2015 **105**

Number of Observations: 480174
Number of Variables: 634
Organization of file: One Record per Variable

Variable Name	Variable Label (VAR)	VAR Type	VAR Format	Position	Columns	Range of Values	Frequency Category	Frequency	Percent
						8	Invalid	891	0.19
						9	Missing	7080	1.47
ST42Q09	Maths Self-Concept - Understand Difficult Work	Num	1.	97	156 - 156	1	Strongly agree	29766	6.20
						2	Agree	92631	19.29
						3	Disagree	129729	27.02
						4	Strongly disagree	56653	11.80
						7	N/A	163187	33.98
						8	Invalid	853	0.18
						9	Missing	7355	1.53
ST42Q10	Maths Anxiety - Worry About Getting Poor <Grades>	Num	1.	98	157 - 157	1	Strongly agree	96688	20.14
						2	Agree	116924	24.35
						3	Disagree	62377	12.99
						4	Strongly disagree	33417	6.96
						7	N/A	163189	33.99
						8	Invalid	545	0.11
						9	Missing	7034	1.46
ST43Q01	Perceived Control - Can Succeed with Enough Effort	Num	1.	99	158 - 158	1	Strongly agree	153582	31.98
						2	Agree	136272	28.38
						3	Disagree	17865	3.72
						4	Strongly disagree	3802	0.79
						7	N/A	162973	33.94
						8	Invalid	374	0.08
						9	Missing	5306	1.11

Created by: ACER

4 0.082103
96 0.000923
1 0.008487
5 0.117712
97 0.442620
2 0.026568
6 0.173985
98 0.001292
3 0.062177
7 0.084133

During your life, how many times have you used any of these types of illegal drugs?

NaN 0.000185
1.0 0.015498
2.0 0.011808
3.0 0.006273
4.0 0.004244
5.0 0.005351
6.0 0.002030
7.0 0.001661
8.0 0.000923
9.0 0.000185
10.0 0.004428
12.0 0.001476
13.0 0.000185
14.0 0.000369
15.0 0.001661

As you can see, I'm looking at the relationship between the variables "level of freedom that the parents give to their children about the people with they hang around" and the "use of cigarette smoking".

The 83% of the parents let their children make their own decisions about the people with they hang around. These children have a lower percentage of cigarette smoking use (44%) that those children whose parent's try to control the decisions about the people with they hang around (49%). There's no missing data on anyone of the studied variables.

See you on the next week!

16.0	0.000369
17.0	0.000369
20.0	0.002768
22.0	0.000369
23.0	0.000369
25.0	0.001292
26.0	0.000185
29.0	0.000185
30.0	0.002214
31.0	0.000185
32.0	0.000185
40.0	0.000738
45.0	0.000185
50.0	0.002583
55.0	0.000185
59.0	0.000185
60.0	0.000369
65.0	0.000185
67.0	0.000369
75.0	0.000369
78.0	0.000185
80.0	0.000185
99.0	0.000185
100.0	0.002952
101.0	0.000185
110.0	0.000185
150.0	0.000369
200.0	0.000738
250.0	0.000185

333.0	0.000185
450.0	0.000185
500.0	0.000554
600.0	0.000185
700.0	0.000185
900.0	0.000185
996.0	0.007196
997.0	0.907749
998.0	0.007011
999.0	0.001476

Show more

As you can see, I have taken four variables on my script: the level of freedom and control that the parents perform over their children (1st variable), and the level of relationship that their children have with tabaco (2nd variable), alcohol drinks (3rd variable) and drugs (4th variable). On my research, I'm looking at the frequencies distribution of the last three variables from children whose parents try to perform a big control over their lives (variable 1 == 0 on my script).

I have not uploaded the full code and the results of my research to this post in order to make it shorter, but if somebody would like to have access to the full script and see the full data, you

can send me a message to my Tumblr profile and I gladly will share them with you.

Taking a quick review of my results, we can see that the values for the three variables (relationship with tabaco, alcohol and drugs) look pretty bigger on my study group rather than in the control group. The distribution of the results shows that the children whose parents try to control the people they hang around with say that they have had less contact with tabaco, alcohol and drugs.

In the next assignments, I will have to use some significance test like the chi square or t student test to see if these differences are statistically significants and if we can extract real conclusions from my research.

See you soon!