Data Analysis and Statistical Inference. Data Analysis Project Part I - Proposal

Eduardo Díez Báez

March, 2014

ABSTRACT

This study is dedicated to examine if there exist association between the sexual orientation of respondent and attitude facing that gay or lesbian couples should be legally permitted to adopt children, according to the data from the 2012 ANES survey.

This report is the proposed project, to complain the Computational Track for the COURSERA course "Data Analysis and Statistical Inference" and it aims to outline the scope of the project according to the submission guidelines, therefore will be followed in the order and content.

Additionally, in preparation for the Part II, this proposal will be published in markdown format, and also translate into a PDF file directly from RStudio in order to be attached as it is state for that last submission.

1. RESEARCH QUESTION. In one sentence, what is your research question?

(Evaluation/feedback note: Do you understand the research question? How can the question be made clearer? Is it clear how data can be used to answer this research question as its phrased?)

With the information extracted from the NAES 2012 survey, Is there an association between the ***sexual orientation*** and the attitude of the respondents facing the ***legal adoption of children by homosexual couples***?

2. DATA - Citation. Include a citation for your data, and if your data set is online, provide a link to the source.

(Evaluation/feedback note: Is there a citation for the data? If there is a link provided, does it bring you to the data?)

Data Citation. The American National Election Studies (ANES; www.electionstudies.org). The ANES 2012 Time Series Study [dataset]. Stanford University and the University of Michigan [producers].

The dataset has been modified slightly to make them easier to use as part of this course and it is the one to use in the present report. The dataset can be obtained from:

- Use the following link to download, directly to you PC, the statistics anes modified dataset file.
- (As indicated in this COURSERA course) Use the following code to load the COURSERA modified ANES data set into R:

```
load(url("http://bit.ly/dasi_anes_data"))
```

```
## Error: all connections are in use
```

3. DATA - Collection. Describe how the data were collected.

(Evaluation/feedback note: Is the data collection explained clearly?)

The ANES 2012 Time Series is a dual-mode survey (face-to-face and Internet) with two independent samples. Cases selected for the face-to -face sample could not be interviewed on the Internet, and cases selected for the Internet survey could not be interviewed in person. Design criteria also included having sufficient numbers of black and Hispanic respondents to enable analysis of those subgroups.

The variables of interest for this study correspond to respondents' answers to the following two questions:

1. Do you consider yourself to be heterosexual or straight, homosexual or gay, or bisexual? (variable: ORIENTN_RGAY)
2. Do you think gay or lesbian couples should be legally permitted to adopt children? (variable: GAYRT_ADOPT)

4. DATA - Cases (observational/experimental units). What are the cases? (Remember: case = units of observation or units of experiment)

(Evaluation/feedback note: Are the cases (the units of observation or experimental units) explained clearly?)

The cases are individuals, each RESPONDENT of the surveys is a unit of observation.

5. DATA - Variables. What are the two variables you will be studying? State the type of each variable.

(Evaluation/feedback note: Are the variable types identified accurately?)

The two variables we are interested are:

GAYRT_ADOPT From the original ANES data set.

Data type: numeric Minimum code defined as valid: 0 Record/columns: 1/213-214

VALUE	LABEL
NA	Refused
NA	Don't know
NA	Not asked, unit nonresponse
NA	Error
NA	Restricted access
NA	Inapplicable
1	Yes
2	No

From the course recommended modified data set we have:

Data type: categorical (Factor) with 2 levels

Categories:
{Yes}
{No}

ORIENTN_RGAY From the original ANES data set.

Data type: numeric Minimum code defined as valid: 0 Record/columns: 1/243-244

VALUE	LABEL
NA	Refused
NA	Don't know
NA	Not asked, unit nonresponse
NA	Error

VALUE	LABEL
NA	Restricted access
NA	Inapplicable
1	Heterosexual or straight
2	Homosexual or gay (or lesbian)
3	Bisexual

From the course recommended modified data set we have:

Data type: categorical (Factor) with 3 levels

Categories:
{Heterosexual Or Straight
{Homosexual Or Gay (Or Lesbian)}
{Bisexual}

6. DATA - Type of study. What is the type of study? Is it an observational study or an experiment? Explain how you've arrived at your conclusion using information on the sampling and/or experimental design.

> *(Evaluation/feedback note: Is the type of study identified correctly? Is the supporting information on the sampling and/or experimental design of the study satisfactory for making the decision on the type of study?)*

The proposal is a **retrospective observational study** based on data already collected in ANES 2012

The study is observational because there is no control intervention on our part and is limited to make use of the measurements of the two variables of interest from ANES 2012.

7. DATA - Scope of inference - generalizability. Identify the population of interest, and whether the findings from this analysis can be generalized to that population, or, if not, a subsection of that population. Explain why or why not. Also discuss any potential sources of bias that might prevent generalizability.

> *(Evaluation/feedback note: Did the writer correctly identify the population of interest? Did the writer correctly decide whether the findings from this analysis can be generalized to that population, or, if not, a subsection of that population? Is their explanation satisfactory to make this decision? Are potential sources of bias discussed, and if so, is the discussion satisfactory?)*

The population is the contemporary United States society in general, U.S. citizens age 18 or older.

There could exist several possible reasons of bias from the respondents of the survey (young, poor, residentially mobile, and many more) and also we must take into account that we have no control of any possible confounding factors.

Despite the foregoing, the findings could be generalize to explain the trend of this population because it is an observational study based on a well know dataset as the ANES.

8. DATA - Scope of inference - causality. Can these data be used to establish causal links between the variables of interest? Explain why or why not.

> *(Evaluation/feedback note: Did the writer identify correctly whether these data be used to establish causal links between the variables of interest. Is the explanation satisfactory?)*

We cannot establish causal links between the variables because the study is observational and not experimental. That's the principal reason our findings could be generalized, but that does not mean we could establish a causality relations between the variables of interest.

9. EXPLORATORY DATA ANALYSIS. Perform a brief exploratory data analysis - just one or two relevant descriptive statistics and visualizations of the data. Also address what the exploratory data analysis suggests about your research question.

> You can include summary statistics in your write up in the space provided below by copying and pasting the R output, however you must upload your plot(s) as a separate PDF or image file. Make sure to provide a discussion/interpretation of any summary statistic or plot you include.
>
> Note: This is not required to be an exhaustive exploratory data analysis (you can save that for the second phase), instead it's just intended to ensure that you have successfully imported your data set into R such that you can produce plots and numerical summaries.
>
> *(Evaluation/feedback note: Have relevant summary statistics been produced and discussed?)*

We load the ANES data set in order to work with some of the variables and look at the dim command.

```
load(url("http://bit.ly/dasi_anes_data"))
```

```
## Error: all connections are in use
```

```
dim(anes)
```

```
## [1] 5914  205
```

The ANES data set has 5914 cases and 205 variables.

But we are interested in the variables (as category A) "orientn_rgay" and (as category B) "gayrt_adopt" and one extra variable to control the resulting dataset as is the original "caseid".

We follow with the creation of a subset called "myanes" containing those variables of interest, and then we take out all cases with NAs values.

```
myanestmp <- subset(anes, select = c(caseid, orientn_rgay, gayrt_adopt))
myanes <- na.omit(myanestmp)
```

We are ready to give a look at the subset that we obtained and start to work with it.

```
dim(myanes)
```

```
## [1] 5579    3
```

```
names(myanes)
```

```
## [1] "caseid"       "orientn_rgay" "gayrt_adopt"
```

```
str(myanes)
```

```
## 'data.frame':    5579 obs. of  3 variables:
##  $ caseid      : int  1 2 3 4 5 6 9 10 11 12 ...
##  $ orientn_rgay: Factor w/ 3 levels "Heterosexual Or Straight",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ gayrt_adopt : Factor w/ 2 levels "Yes","No": 2 1 2 2 1 1 2 1 1 1 ...
##  - attr(*, "na.action")=Class 'omit'  Named int [1:335] 7 8 22 35 48 68 70 74 76 80 ...
##   .. ..- attr(*, "names")= chr [1:335] "7" "8" "22" "35" ...
```

```
summary(myanes$orientn_rgay)
```

```
##       Heterosexual Or Straight Homosexual Or Gay (Or Lesbian)
##                           5341                             124
##                       Bisexual
##                            114
```

```
summary(myanes$gayrt_adopt)
```

```
##  Yes   No
## 3451 2128
```

```
head(myanes)
```

```
##   caseid             orientn_rgay gayrt_adopt
## 1      1 Heterosexual Or Straight          No
## 2      2 Heterosexual Or Straight         Yes
## 3      3 Heterosexual Or Straight          No
## 4      4 Heterosexual Or Straight          No
## 5      5 Heterosexual Or Straight         Yes
## 6      6 Heterosexual Or Straight         Yes
```

```
tail(myanes)
```

```
##      caseid             orientn_rgay gayrt_adopt
## 5909   6859 Heterosexual Or Straight         Yes
## 5910   6860 Heterosexual Or Straight         Yes
## 5911   6861 Heterosexual Or Straight         Yes
## 5912   6862 Heterosexual Or Straight          No
## 5913   6863 Heterosexual Or Straight          No
## 5914   6864 Heterosexual Or Straight         Yes
```

The resulting data set has 5579 cases and the 3 variables we already mentioned them.

With the command str() we get the type of variable and the summary command help us to identify the levels of each variable as we mentioned before and the number of cases in the subset for each level.

We continue finding the relationship of the two variables of interest in order to get some statistics. We start creating a contingency table and showing the results

```
Sex_orientation <- myanes$orientn_rgay
Attitude_pro_adoption <- myanes$gayrt_adopt
mytable <- table(Sex_orientation, Attitude_pro_adoption)
mytable
```

```
##                                 Attitude_pro_adoption
## Sex_orientation                   Yes   No
##   Heterosexual Or Straight       3230 2111
##   Homosexual Or Gay (Or Lesbian)  120    4
##   Bisexual                        101   13
```

It is time to show those value in a graphical manner.

```
# see Figure 1
png(filename = "figure/mosaic01.png", width = 6, height = 6, units = "in", res = 266,
    bg = "transparent")
mosaicplot(mytable, main = "Pro Adoption - Sex Orientation")
dev.off()
```

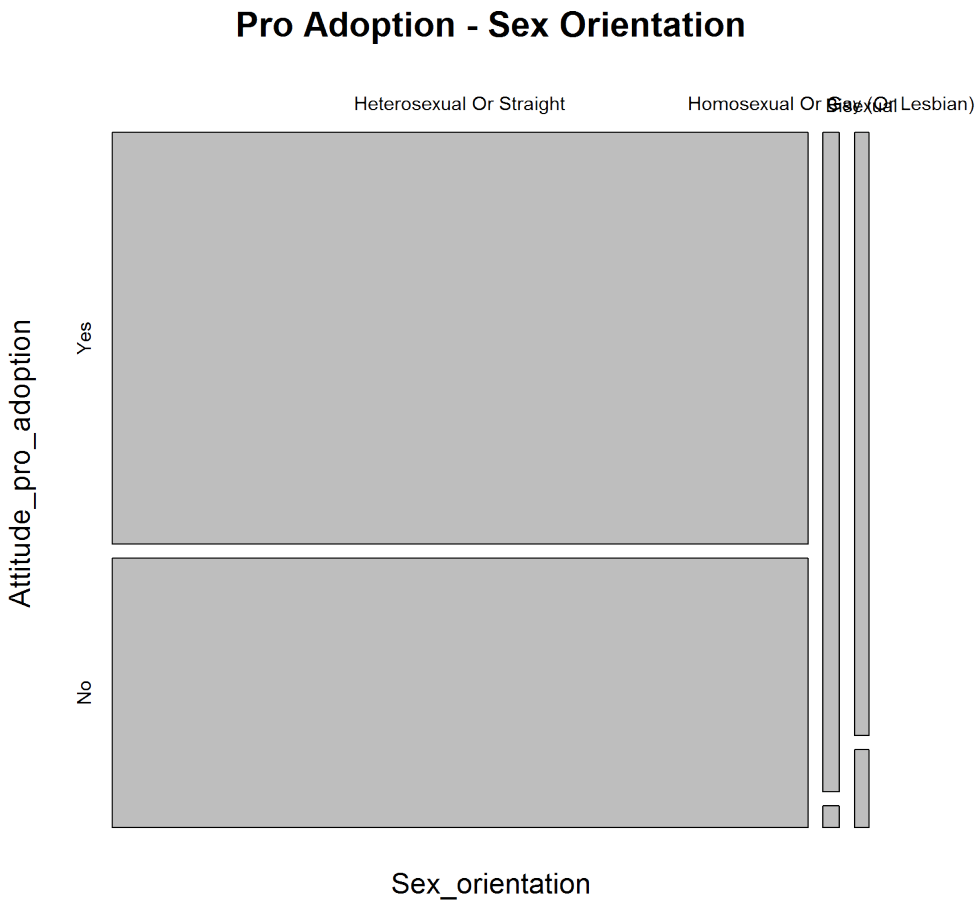

Figure 1: Standard mosaic plot

```
## pdf
##   2
```

In figure 1, the variables seem to be strong correlated. We observe a tend to increse the frecuency of "Yes" answer as we go from heterosexuels to homosexuals through bisexuals.

From the last two, the contingency table and the mosaic plot, seems that there exist a relationship between the two variables. We can give a deeper look and calculating the chi-sq from the contingency table.

```
chisq.test(mytable, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  mytable
## X-squared = 102.9, df = 2, p-value < 2.2e-16
```

We observe a X-sq value over 100 and a p-value very low, almost zero.

Before to conclude, we make another test. The Kendall Tau-B and to do so, we need to prepare our data (the two variables)

```
Orientation <- as.numeric(factor(myanes$orientn_rgay, levels = c("Homosexual Or Gay (Or Lesbian)"
    "Bisexual", "Heterosexual Or Straight")))
Adoption <- as.numeric(factor(myanes$gayrt_adopt, levels = c("Yes", "No")))
```

Now, we combine the two variables in a new contingency table and show the result.

```
mytabla <- table(Orientation, Adoption)
mytabla
```

```
##             Adoption
## Orientation    1    2
##           1  120    4
##           2  101   13
##           3 3230 2111
```

We check the Kendall correlation

```
mymm <- cbind(Orientation, Adoption)
cor(mymm, method = "kendall", use = "pairwise")
```

```
##              Orientation Adoption
## Orientation       1.0000   0.1344
## Adoption          0.1344   1.0000
```

We observe a Kendall correlation value equal to 0.1344 between the variables, once they have been valuated with levels equally spaced.

Let's check the alternative hypothesis

```
cor.test(Orientation, Adoption, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  Orientation and Adoption
## z = 10.09, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##    tau
## 0.1344
```

Last, we make and extra mosaicplot, but this time in a different way

```
# see Figure 2
require(openintro)
require(vcd)
require(vcdExtra)
png(filename = "figure/mosaic02.png", width = 6, height = 6, units = "in", res = 266,
    bg = "transparent")
mosaic(mytabla)
dev.off()
```

```
## pdf
##   2
```

Figure 2 was obtained using the "vcd" library. Basically it is the same as mosaicplot but this time we used the already prepare data where changes the factors levels into equally space values and also changes the order the are show from the data set, and so we can use other technics as the ones we can use with numerics variables. In the figure we observe the same tendency.

We must have into account our objetive; we are interesting in the association of the variables, so let's draw an association plot.

The association plot (Cohen, 1980[1] and Friendly, 1991[2]) puts deviations from independence in the foreground, and so, the area of each box is made proportional to observed - expected frequency.

- In the association plot, each cell is shown by a rectangle.

[1] Cohen, A. (1980). On the graphical display of the significant components in a two-way contingency table. Communications in Statistics - Theory and Methods, A9, 1025-1041.

[2] Friendly, M. (1991). The SAS System for Statistical Graphics. Cary, NC: SAS Institute Inc.

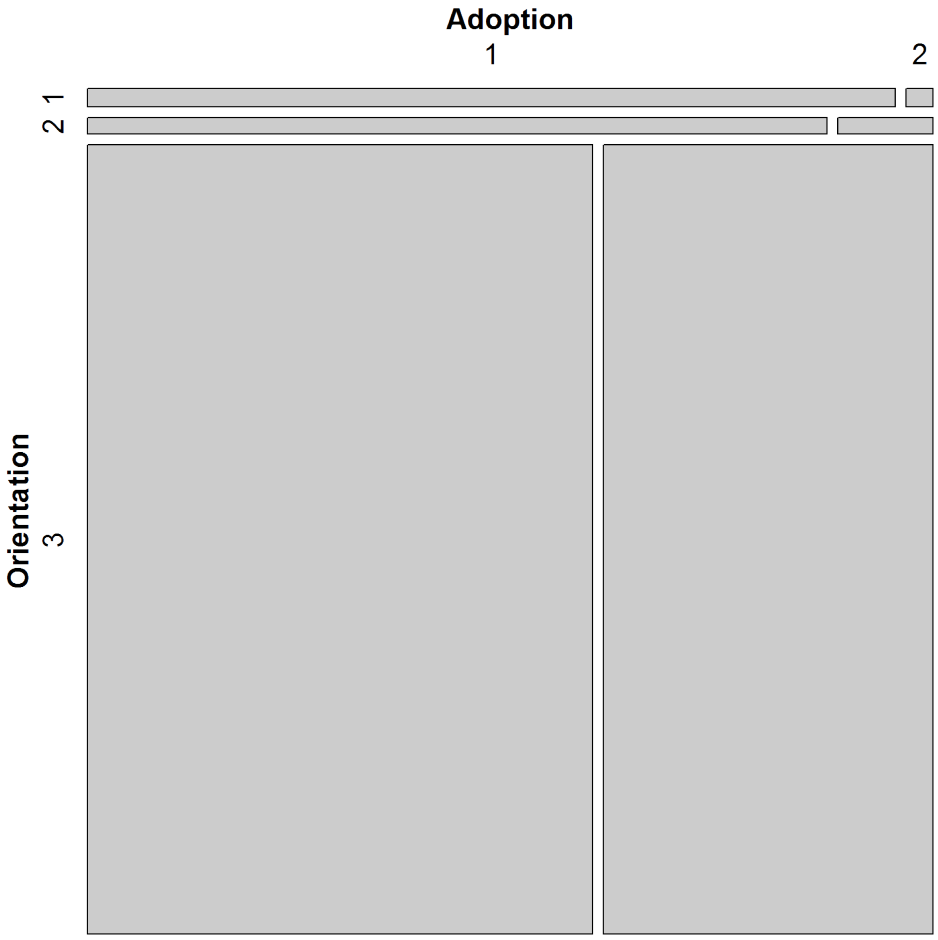

Figure 2: Mosaic plot using vcd with numerics variables

- The rectangles for each row in the table are positioned **relative to a baseline representing independence** shown by a dotted line.
- Cells with **observed > expected frequency** rise above the line (colored black);
- Cells that contain **less than the expected frequency** fall below it (shaded red).

In R we can plot this kind of informatition with the following code:

```
assocplot(the_contingency_table)
```

Let's keep save our plot to show it later inside the documment.

```
# see Figure 3
png(filename = "figure/associ01.png", width = 6, height = 6, units = "in", res = 266,
    bg = "transparent")
assocplot(mytabla, main = "Relation: Sx Orientation and Adoption Attitude")
dev.off()
```

```
## pdf
##   2
```

Figure 3 shows; Adoption-1=Yes (Pro Adoption) and Adoption-2=No (Anti Adoption), while the levels of Sexual Orientation are related as follow; 1=Homosexual, 2=Bisexual and 3=Heterosexual. The attitude Pro Adoption overcome the expectations from the colective of respondents that considerer themself as Homosexuals and Bisexual. At the same time the attitude Anti Adoption from this colective is under the expectations. The Heterosexual colective tend in opposite way; The attitude Pro Adoption is under the expentancies and the Anti Adoption attitude of this colective is over the expentancies.

After we observed the chi-sq, p-value, Kendall correlation, mosaic and association plots we arrive to the next find:

> Having H_0 -the null hypothesis- that the variables are independents, and H_a -the alternative hypothesis- that the variables are not independents; we have found evidences that make us reject H_o in favor of H_a, because the Chi-sq is high, and the p-value almost cero, so compare with $\alpha = 0.01$ ($\Rightarrow confident\ level\ = 99\%$), there is a very low probability that the variables are independents
>
> > Note:We are going to learn about this issue in future weeks in this course and so, one could make the phase II of the computational track project.

10. Attach a page of your data set that includes all of the relevant columns. You do not need to include all of the data; one page is sufficient

> *(Evaluation/feedback note: Is there a data set present? Can you see all of the relevant columns?)*

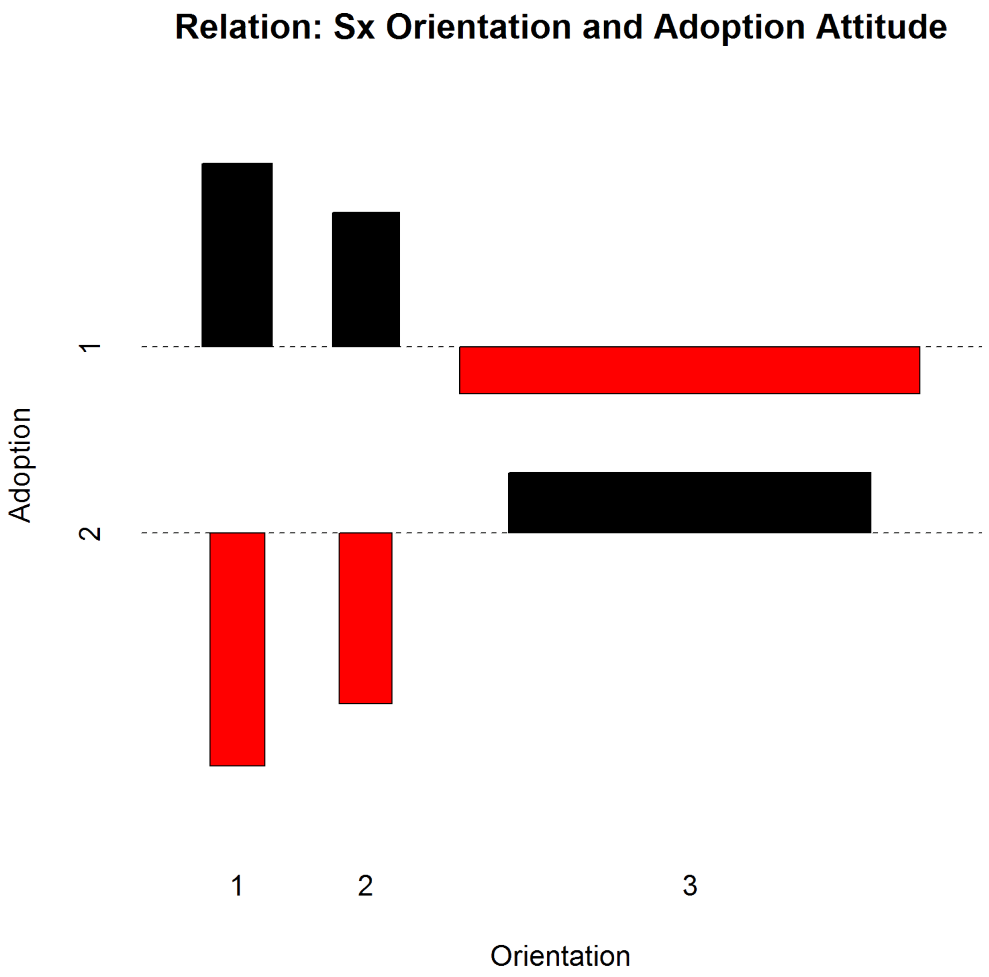

Figure 3: Association plot "Cohen-Friendly".

Let's print one page long of the dataset in an old fashion style with 43 lines.

```
print(myanes[1:43, ])
```

```
##    caseid             orientn_rgay gayrt_adopt
## 1       1 Heterosexual Or Straight          No
## 2       2 Heterosexual Or Straight         Yes
## 3       3 Heterosexual Or Straight          No
## 4       4 Heterosexual Or Straight          No
## 5       5 Heterosexual Or Straight         Yes
## 6       6 Heterosexual Or Straight         Yes
## 9       9 Heterosexual Or Straight          No
## 10     10 Heterosexual Or Straight         Yes
## 11     11 Heterosexual Or Straight         Yes
## 12     12 Heterosexual Or Straight         Yes
## 13     13 Heterosexual Or Straight         Yes
## 14     14 Heterosexual Or Straight         Yes
## 15     15 Heterosexual Or Straight         Yes
## 16     16 Heterosexual Or Straight          No
## 17     17 Heterosexual Or Straight          No
## 18     18 Heterosexual Or Straight         Yes
## 19     19 Heterosexual Or Straight         Yes
## 20     20 Heterosexual Or Straight          No
## 21     21 Heterosexual Or Straight         Yes
## 23     23 Heterosexual Or Straight         Yes
## 24     24 Heterosexual Or Straight          No
## 25     25 Heterosexual Or Straight          No
## 26     26 Heterosexual Or Straight          No
## 27     27 Heterosexual Or Straight          No
## 28     28 Heterosexual Or Straight         Yes
## 29     29 Heterosexual Or Straight          No
## 30     30                 Bisexual         Yes
## 31     31 Heterosexual Or Straight         Yes
## 32     32 Heterosexual Or Straight         Yes
## 33     33 Heterosexual Or Straight         Yes
## 34     34 Heterosexual Or Straight          No
## 36     36 Heterosexual Or Straight         Yes
## 37     37 Heterosexual Or Straight         Yes
## 38     38                 Bisexual         Yes
## 39     39 Heterosexual Or Straight         Yes
## 40     40 Heterosexual Or Straight         Yes
## 41     41 Heterosexual Or Straight         Yes
## 42     42 Heterosexual Or Straight         Yes
## 43     43 Heterosexual Or Straight         Yes
## 44     44 Heterosexual Or Straight         Yes
## 45     45                 Bisexual         Yes
## 46     46 Heterosexual Or Straight          No
## 47     47 Heterosexual Or Straight          No
```

Overall evaluation/feedback

> *Please make a general statement about the appropriateness of the dataset to answer the question. In addition, summarize any issues with this submission.*

Apendix A

> This apendix is intended to guide fellow students in generate one file with all the information in order to submit just one file as is state in the submission page for the part II of this project, where is requested to submit all the content in one file edited in R-markdown.

In order to avoid the use of the command or system console to run pandoc, the R-code that follow will help you to do the job.

All this documents was created as a R markdown file in RStudio. below is the listing of an extra R-code file that do all necesaries calls to make the translation form R-markdonw (using knit to trnaslate from *.Rmd to* .md) and then calls Pandoc to translate from *.md to the format you deside, in this case to pdf throuht the use of Latex, in version MiKTex for Windows.

Once one desire to make the translation the only to do is run this code. You can make changes in it to custom your needs. Very important to have the correct path to work without problem:

In Windows 7, Pandoc is usually installed at,

```
"C:\Users\Eduardo\AppData\Local\Pandoc"
```

But Pandoc, by default, thinks as user directory:

```
"C:\Users\YOUR_USER_NAME\AppData\Roaming\pandoc\"
```

If you use this subdirectory as your working one, everything will work fine. You can use the next R code to see wich directory is actually your working one:

```
getwd()
```

And you can use the next R code to change it, for example;

```
setwd("C:/Users/YOUR_USER_NAME/AppData/Roaming/pandoc")
```

Here is the extra R-script:

```
# Define your report
system("RMDFILE=myreport")

# Knit the Rmd to an Md file
# Convert the MD file to Html
system("Rscript -e 'require(knitr);
        require(markdown);
        knit('$RMDFILE.rmd', '$RMDFILE.md');
        markdownToHTML('$RMDFILE.md', '$RMDFILE.html',
        options=c(\"use_xhml\"))'")

require(knitr)
require(markdown)

knit("myreport.Rmd")
markdownToHTML('myreport.md','myreport.html', options=c("use_xhml"))

# convert the generated md to pdf
system("pandoc -s myreport.md -o myreport_md.pdf")
# this is the one I used
# but you have many many many diferent format with pandoc.
```

In order to do the job, R or RStudio require "knitr" and "markdown". The former should be installed as base. Anyway you must check.

Aditionally, your installation machine (PC) need Pandoc and MiKTeX. The later is a fantastic big stuff but sometimes a difficult one. Hopefully, it is supose that you install MiKTex and Pandoc and everything should go easily. In MiKTex (I try to remember) should be an option when you are installing that indicate that any needed library can be downloaded on the fly, as it is needed in compilation time; say "yes" to this option and so, you don't need to care about and you will produce fantastic good looking documents with LaTeX.

References: