Nationality and Close Relationships

Anshu Chen

Abstract: Does a person's culture and national origin influence the way they behave in close relationships? To answer this question, I examine anonymous results from the Experiences in Close Relationships Scale psychological survey (Brennan et al). My analysis reveals a somewhat significant association between national origin and anxiety/avoidance scores. Based on this association, I ranked the countries in terms of their aggregated anxiety/avoidance scores. I believe my results provide some insight into possible cultural differences that cause these score differences.

1. Exploration and cleaning

This was an immense data set that began with 17387 observations. The raw data can be found here. http://personality-testing.info/_rawdata/ I relegate most of my data cleaning to the appendix. But I include treating missing values, dropping bad values, and compiling the avoidance and anxiety scores, because I made these decisions directly based on my final analysis approach.

1.1. Missing values

If a respondent skipped a question, the survey recorded 0. This would distort the anxiety and avoidance total, so I dropped all observations containing missing values.

```
df1[, 1:36] <- apply(df1[, 1:36], 2, function(df1) gsub("0", NA, df1))
df1 <- df1[!rowSums(is.na(df1[, 1:36])), ]
df1[, 1:37] <- apply(df1[, 1:37], 2, as.numeric)
```

Some observations were missing gender values as well. The data's documentation states that 1 and 2 represent male and female respectively, but the gender values contained 0s and 3s.

```
table(df1$gender)
```

```
## 
##      0      1      2      3 gender 
##    112   5037  10403    196      0 
```

I initially wanted to remove such unorthodox points. In my later analysis, however, I average each country's observations into a single representative observation for that country. It would be impossible to fairly assign a gender to each country, so I had to drop gender as a variable. Thus I can keep these unorthodox genders and include people who may not identify as either gender-or, alternatively, forgot to answer the question.

1.2 Dropping bad values

Many observations contain implausible ages.

```
summary(df1$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   -90.0    18.0    22.0    25.9    30.0  2150.0 
```

I discard ages greater than 100.

```
df1 <- df1[-c(which(df1$age > 100)),]
```

As for a lower bound, I assume that nobody under 10 would take a survey about close relationships.

```
df1 <- df1[-c(which(df1$age < 10)),]
```

These restrictions remove PA (Panama) and VG (Virgin Islands) from the data. I'll remove their levels.

```
levels(df1$country)[levels(df1$country) == "PA"] <- NA
levels(df1$country)[levels(df1$country) == "VG"] <- NA
```

While I'm sure some of these remaining people are lying about their age, there's really no way to further filter the data. I suppose it's conceivable that a 99-year-old grandma or grandpa somewhere has taken this quiz. I'd hate to boot them out of the data set because of a few bad apples.

1.3 Summing avoidance and anxiety scores

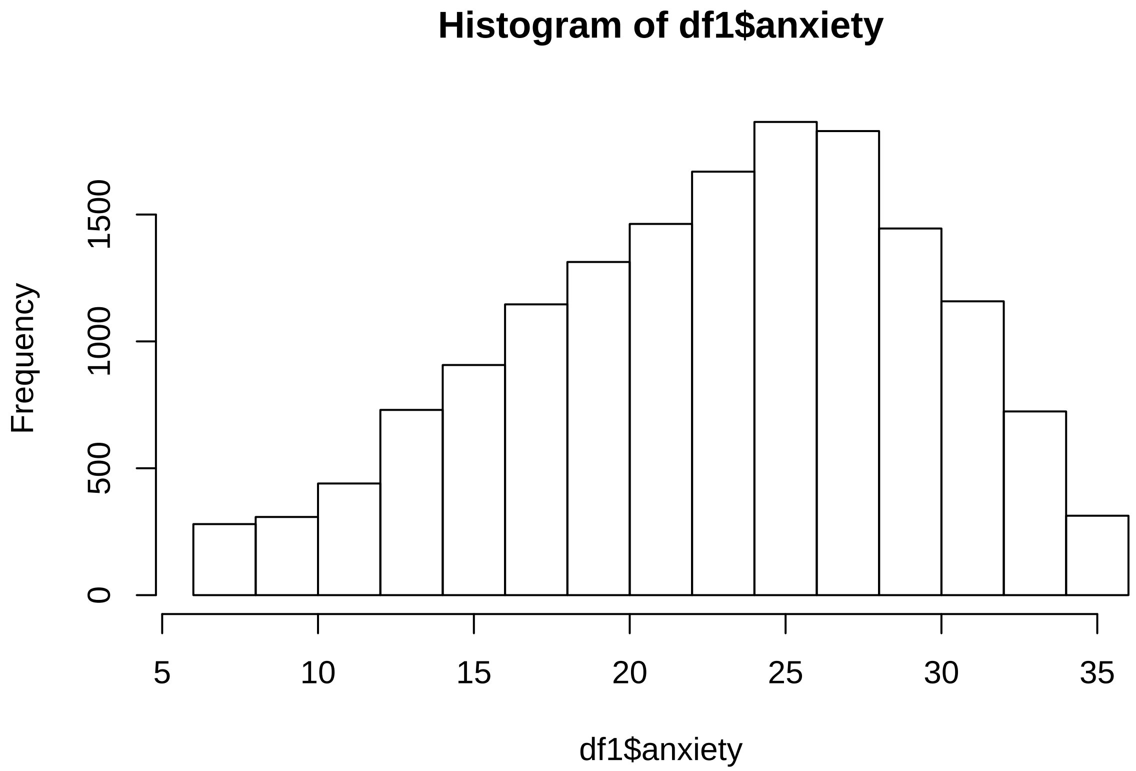
Based on the documentation, I categorized the questions into "anxiety" and "avoidance". Then I assigned them as positive or negative, based on whether a positive response indicated higher or lower anxiety/avoidance.

```
df1$avoidance <- df1$Q.1 - df1$Q.3 + df1$Q.5 + df1$Q.7 + df1$Q.9 + df1$Q.11 + df1$Q.13
- df1$Q.15 + df1$Q.17 - df1$Q.19 + df1$Q.21 + df1$Q.23 - df1$Q.25
-    df1$Q.27 - df1$Q.29 - df1$Q.31 - df1$Q.33 - df1$Q.35
df1$anxiety <- df1$Q.2 + df1$Q.4 + df1$Q.6 + df1$Q.8 + df1$Q.10 + df1$Q.12 +    df1$Q.14
+ df1$Q.16 + df1$Q.18 + df1$Q.20 - df1$Q.22 + df1$Q.24 + df1$Q.26
+   df1$Q.28 + df1$Q.30 + df1$Q.32 + df1$Q.34 + df1$Q.36
```

2. Regression

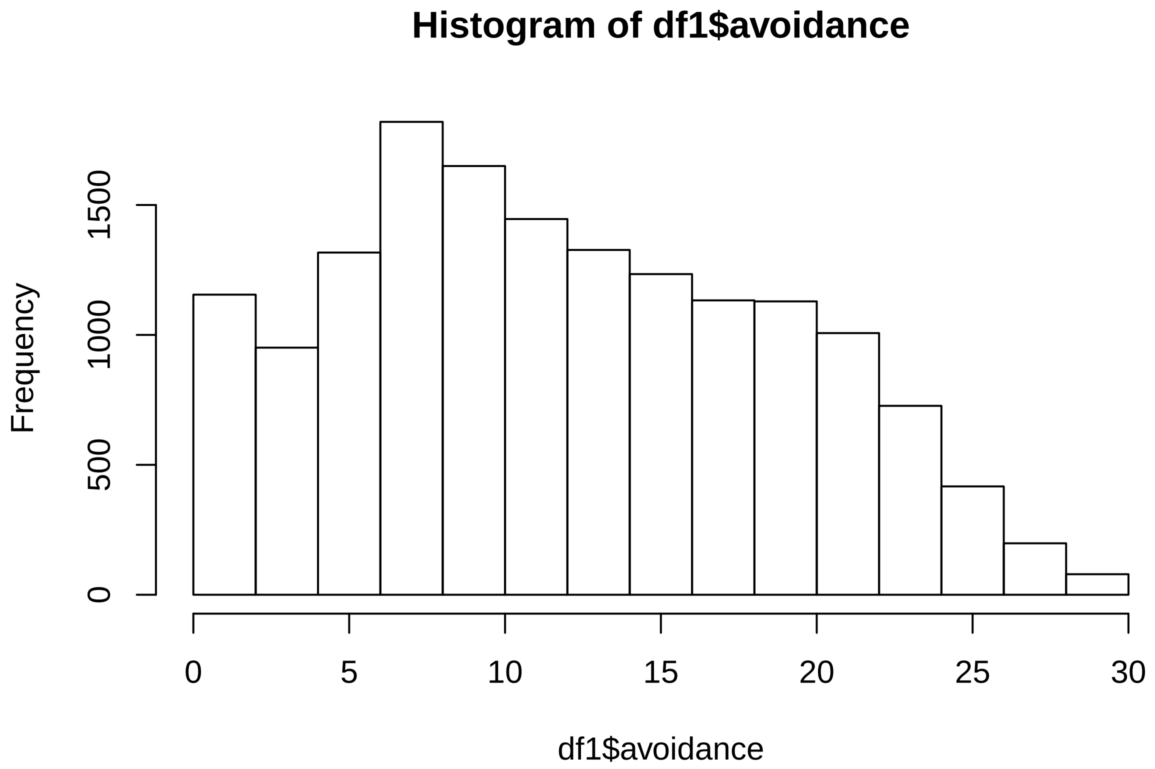
Now that I have created the necessary variables, I will explore the anxiety and avoidance variables to see if they are normal.

```
hist(df1$anxiety)
```



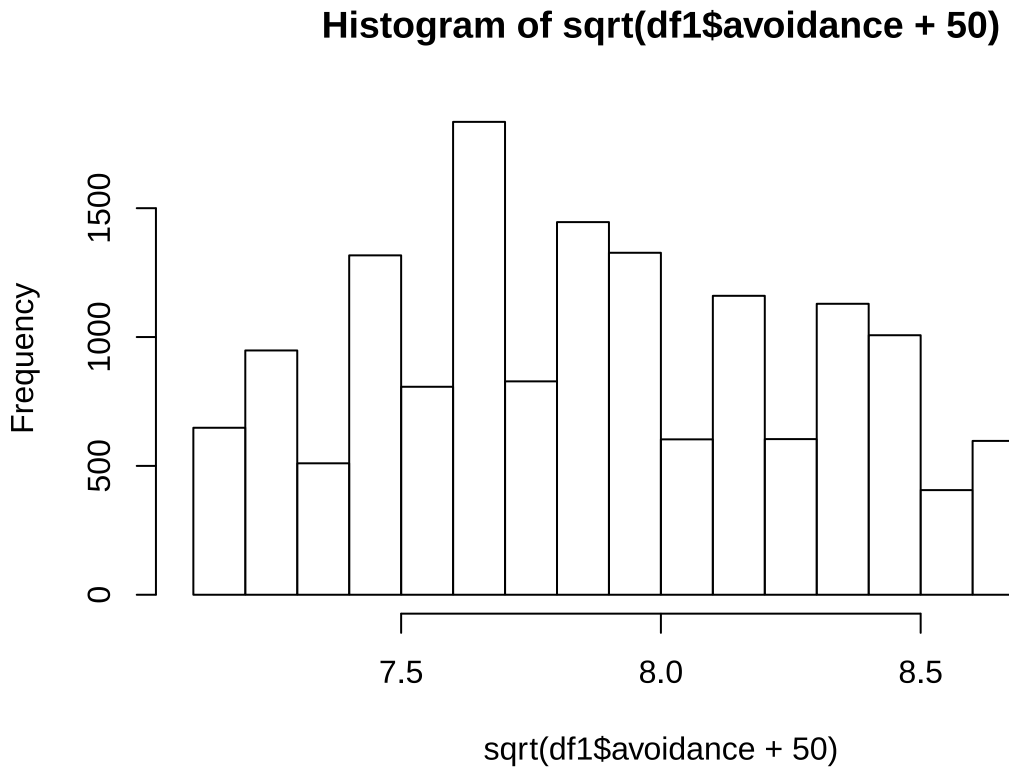
A little skewed, but altogether not bad.

```
hist(df1$avoidance)
```



This is quite skewed. I will try adding 50 (enough to make all measurements positive) and taking the square root.

```r
hist(sqrt(df1$avoidance + 50))
```

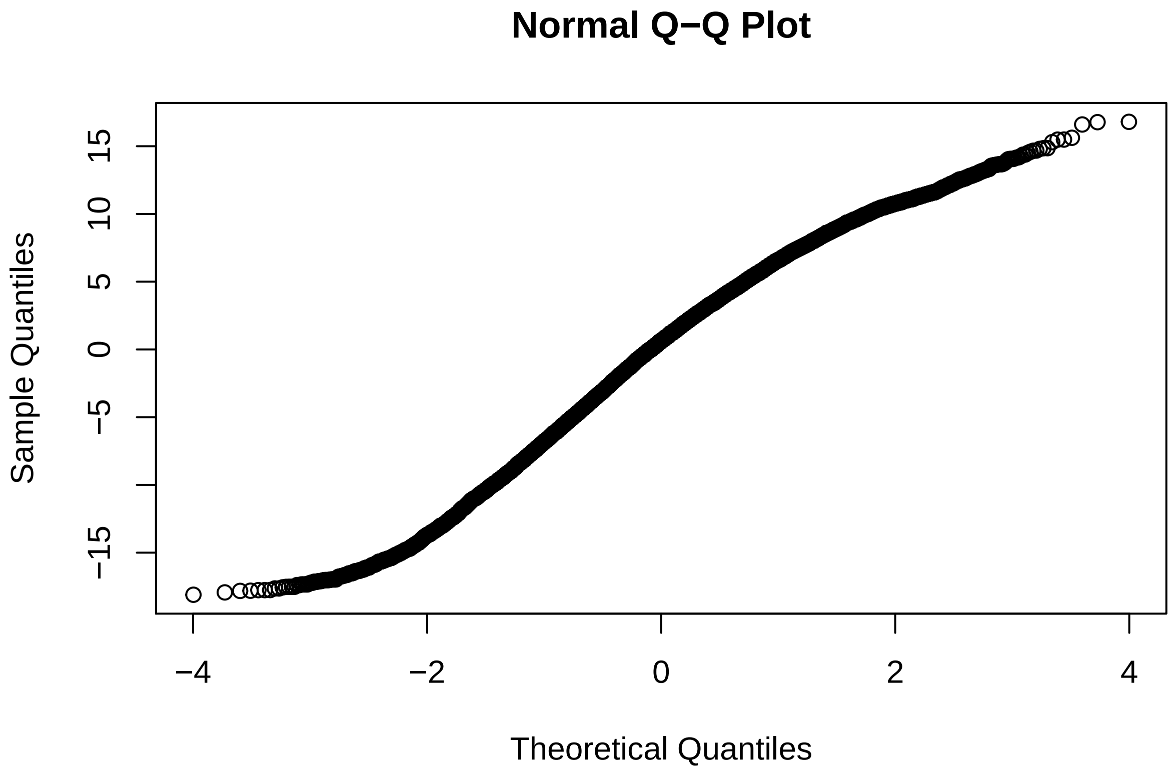


This is better.

```
df1$avoidance.1 <- sqrt(df1$avoidance + 50)
```

2.1. Anxiety

```
lm.0 <- lm(anxiety ~ age + gender + country, data=df1)
qqnorm(lm.0$resid)
```



A fairly straight qqnorm plot, all things considered.

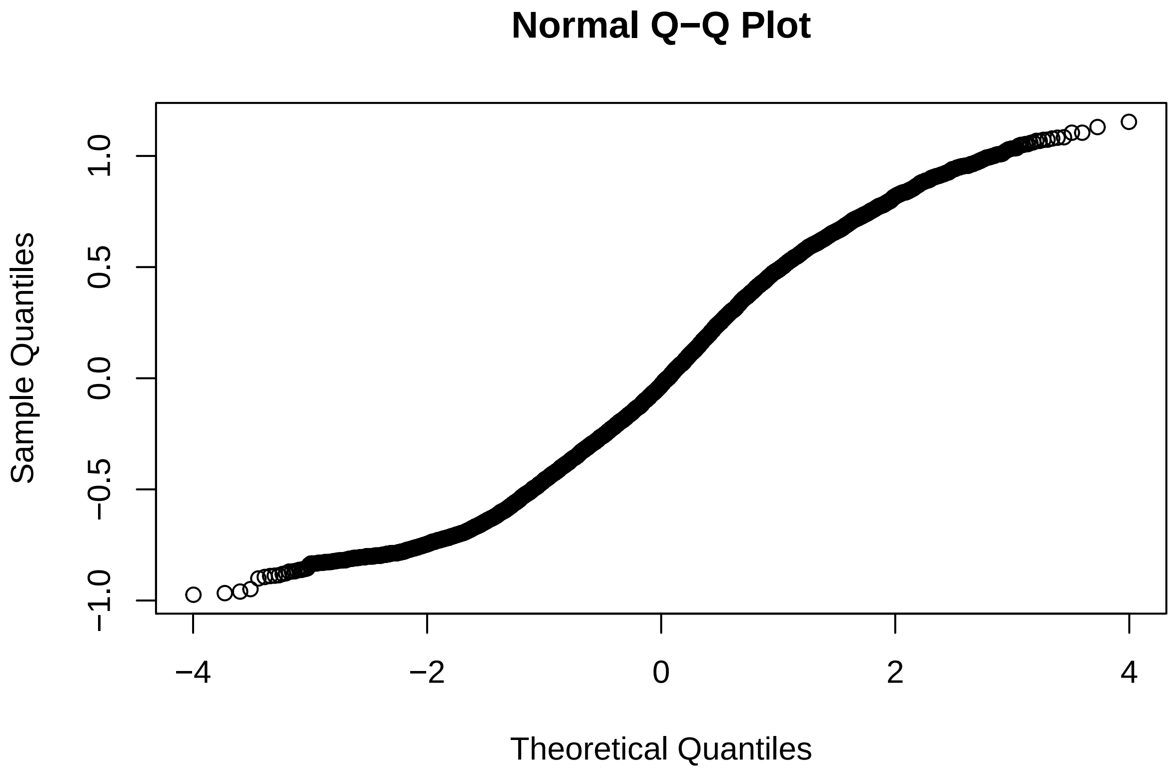
```
anova(lm.0)
```

```
## Analysis of Variance Table
##
## Response: anxiety
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## age           1  29939 29939.2 732.4303 < 2.2e-16 ***
## gender        3    381   127.0   3.1068  0.025350 *
## country     138   7982    57.8   1.4150  0.001029 **
## Residuals 15447 631421    40.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Country has a p-value of .002306. Respectably significant, I'd say, even if age's p-value is lower.

2.2. Avoidance

```
lm.1 <- lm(avoidance.1 ~ age + gender + country, data=df1)
qqnorm(lm.1$resid)
```



Worse-looking than the avoidance plot, but it could still be worse.

```
anova(lm.1)
```

```
## Analysis of Variance Table
##
## Response: avoidance.1
##              Df  Sum Sq Mean Sq  F value  Pr(>F)
## age           1   28.00 27.9993 150.5636 < 2e-16 ***
## gender        3   35.86 11.9520  64.2707 < 2e-16 ***
## country     138   31.61  0.2291   1.2318 0.03411 *
## Residuals 15447 2872.58  0.1860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ouch. Country's p-value takes a definite hit down to 0.0211. But I'll keep going.

3. Aggregation

I have established that country is associated with anxiety and avoidance. The observations are unequally distributed between countries; some have thousands while others have one.

```
table(df1$country)
```

```
## 
##   AE   AF   AG   AL   AM   AP   AR   AT   AU   BA   BB   BD   BE   BF   BG 
##   27    1    3    2    1    7   29   22  739    8    1    4   40    1   19 
##   BH   BN   BO   BR   BS   BW   BZ   CA   CH   CI   CL   CM   CN   CO   CR 
##    1    3    2  117    0    1    1  931   29    1   10    1   21   15    3 
##   CW   CY   CZ   DE   DK   DO   DZ   EC   EE   EG   ES   EU   FI   FR   GB 
##    1    3   21  158   47    4    1    5   18   15   52   14   94   82 1400 
##   GD   GE   GF   GH   GI   GR   GU   HK   HN   HR   HT   HU   ID   IE   IL 
##    1    5    1   10    1   41    4   19    1   42    1   26   62   96   29 
##   IM   IN   IQ   IR   IS   IT   JE   JM   JO   JP   KE   KG   KN   KR   KW 
##    1  379    1   10    8   71    2   11    2   31   16    1    1   23    2 
##   KY   LB   LK   LT   LU   LV   LY   MA   MC   MD   ME   MG   MK   MM   MN 
##    1   11    7    9    4   10    1    3    1    5    6    1    6    1    2 
##   MO   MR   MT   MU   MV   MW   MX   MY   NG   NL   NO   NP   NZ   OM   PE 
##    1    1    8    1    1    3   73   90   15   86   48    3  134    2    8 
##   PH   PK   PL   PR   PS   PT   QA   RO   RS   RU   SA   SB   SD   SE   SG 
##  219   68   62    9    1   42    2   50   38   18   15    1    2  111   96 
##   SI   SK   SV   SY   TH   TN   TR   TT   TW   TZ   UA   UG   US   UY   UZ 
##   21    8    3    3   11    2   31   18   12    3   11    3 9214    4    2 
##   VE   VN   WS   ZA   ZW 
##    8   10    1   93    3
```

To compare the countries, I average each country's measurements to create a representative observation. In other words, this creates a "typical" American, Chinese, German, and Afghan respondent.

```
library(data.table)
agg <- setDT(df1)[, lapply(.SD, mean), by = country, .SDcols = -("gender")]
dagg <- as.data.frame(agg)
```

A possible problem: this takes age out of the equation entirely, since the "average" person also has an "average" age-and countries with only one observation are stuck with their original age. Since age is so effective at predicting avoidance/anxiety, a country represented by one old person will distort the model. Then I should only aggregate if I can prove that the world's population is approximately equally aged. In other worse, that one cannot predict age using country. I will use multiple regression to model age with the other variables.

```
lm.2 <- lm(age ~ gender + avoidance.1 + anxiety + country, data=df1)
anova(lm.2)
```

```
## Analysis of Variance Table
## 
## Response: age
##                Df  Sum Sq Mean Sq  F value    Pr(>F)    
## gender          3    6266    2089  18.4563 6.009e-12 ***
## avoidance.1     1   16000   16000 141.3842 < 2.2e-16 ***
## anxiety         1   77316   77316 683.2145 < 2.2e-16 ***
## country       138   26795     194   1.7158 4.054e-07 ***
## Residuals   15446 1747959     113                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It turns out that country is a pretty good predictor of age. But on the bright side, it's the weakest predictor out of all the ones in the model. Also, the multiple R-squared is only about 0.05. So, I don't feel terrible about mixing the ages together. Gender may also present a similar problem: I suspect that the observations may cluster based on gender. But since many countries only have one observation, and therefore one gender, it's difficult to separate groups based on gender. I eliminate gender as a variable. I realize that this is an extremely sloppy move, but since I'm creating a representative person for each country, that person can't encompass all possible gender groups. Plus, this allows me to include non-binary people who identify as neither or both!

4. Results

4.1. Anxiety (see fig.1)

Most anxious countries: Columbia, Brazil, Venezuela, Egypt, Russia, India, Pakistan, Iraq, Iran.

Least anxious countries: Peru, the Czech Republic, Libya, Sudan, Mauritania, Morocco, Myanmar, Laos, Cambodia, Vietnam.

Interesting. The Western countries are in the middle of the spectrum, but that's likely because they had the most responders and thus are average by default. But I am intrigued by a few things:

1. The most anxious countries are loosely concentrated in the Middle East and parts of Asia. These countries also have strong religious traditions–Catholic for Columbia, Brazil, and Venezuela, and Islamic for Egypt, Pakistan, Iraq, and Iran. India's culture is highly structured around marriage and status, which may impact this anxiety.

2. The Czech Republic, on the other hand, has always viewed relationships liberally and without great anxiety–hence the term "bohemian" used to describe carefree romances. The Southeast Asian countries, stereotypically, also tend to be laid back about relationships: a 2014 study found that Southeast Asian youths were far more likely to express and engage in romantic love, as opposed to the arranged marriages of South Asia.

4.2. Avoidance (see fig.2)

Most avoidant countries: Russia, China, France Libya, Algeria, Mexico, Peru. It's interesting that many countries who were extremly not anxious are on the avoidant list… I suppose it makes sense that someone who is not clingy at all would be on the opposite extreme. France is interesting. I would want to investigate it further–it seems that France has not shed its formal hierarchisms after all. Britain is also fairly reticent compared to Spain and India. Least avoidant countries: India, Romania, Sudan, Bolivia.

Appendix

1. Graphics

2. Preliminary data cleaning

```r
x <- scan("raw data.csv", what="", sep="\n")
y <- gsub("\t", ",", x) # replaces separating \t's with commas
as.vector(y, mode="any") # turns each observation into a vector
f <- strsplit(y, ",") # splits the vectors along the columns
df <- data.frame(matrix(unlist(f), nrow=17387, byrow=T))
# transforms the list into a data frame
names(df) = c(paste("Q", 1:36, sep = "."), "age", "gender", "country")
# creates proper variable names
df <- df[-1,]
# eliminates first, extraneous variable (number ranking)
df1 <- df
df1$Q.1 <- as.numeric(df$Q.1) - 1
```

```
# For some reason, my cleaning added 1 to the 1st question responses... so I'm fixing that.
df1$country <- gsub("\"", "", df1$country)
# Wiped out leading forward slashes
df1$country <- as.factor(df1$country)
# Since country categorizes, it is a factor
df1 <- subset(df1, !(df1$country %in% c("A1","A2","O1")))
# A1, A2, and O1 are unknown countries and satellite providers. So I remove
# them.
levels(df1$country)[levels(df1$country) == "A1"] <- NA
levels(df1$country)[levels(df1$country) == "A2"] <- NA
levels(df1$country)[levels(df1$country) == "O1"] <- NA
```

3. Flaws

I acknowledge that my study had many flaws. If I had more time and could do this study myself, I would design it differently.

2.1. The data contained unreliable parameters, such as age and gender, that the participants could easily ignore or abuse. Many participants wrote inappropriate negative or extremely high ages, rendering their observations unusable. Still others did not indicate any gender, making it impossible to cluster observations based on possible gender differences.

2.2. Unequal data: some countries had thousands of observations, wheras some had only one. The sparse quantity, again, makes it difficult to control for other variables such as age or gender.

2.3. It is uncertain whether we can associate country with anxiety and avoidance. The anova test gave us significant values, for a threshold of 0.05, but 0.05 is arbitrary…and the resulting p-values were certainly not zero. So, I was not too comfortable with putting all my trust in 'country' and aggregating everyone into one representative.

All in all, however, I learned a lot from this project. I am happy that I attempted it.

4. Acknowledgements

Babalievsky, Fil; for helping me think of a title.

Brenner et al for their survey.

http://personality-testing.info/_rawdata/, for their data.

StackOverflow, for introducing me to data.table, echo=FALSE, and dev.off()'s basics.

https://www.worldpulse.com/fr/node/34544, for the basics of romantic relationships contrasted between Southeastern Asia and South Asia.