Homework 2

Simon G. Brauer

Sunday, November 08, 2015

Instructions

1. Take a data set that you are currently working on, and select a small number of variables of interest from it-fewer than ten. You can do this manually if you like-e.g. in Stata, or Excel-or you can use the subset() function to select data from the full dataset. Either way is fine.
2. Using RStudio, create either an RMarkdown file (preferred) or an .R file (also OK) that does the following:

- Get the data into R so you can work with it
- Produce numerical summaries of the variables with summary() or similar
- Use ggplot to produce histograms or some other appropriate 1-D distributional summary of each variable separately
- Use ggplot and GGally packages together to produce a pairs or generalized pairs plot of all the variables together
- Select two or three variables and look at their bivariate relationships more closely. E.g. you might produce a scatterplot & smoother of two continuous variables, or see how the distribution of some continuous variable (e.g. income) varies by a categorical variable (e.g. level of education)

3. Produce a PDF of the results and send it to me, OR use RStudio's publication feature to make a publicly accessible HTML version on its RPubs service. If you use RMarkdown this should be straightforward. If you use an .R script, save the graphical output into individual files and include them in a single PDF document. Your document should contain brief linking information (i.e. literally one or two sentences) explaining what each variable and plot is. You can do as many plots as you like, but at a minimum please show me summaries of each variable individually, a pairs plot, and at least two bivariate or multivariate plots.

```
setwd("C:/Users/Simon/OneDrive/Documents/Github/Data-Visualization/")
library(foreign)
library(ggplot2)
library(GGally)
NCS <- read.dta("NCS.dta")
attach(NCS)
```

I am using data from the National Congregations Study (NCS). I've included variables measuring religious tradition, size of the congregation, clergy's race, congregation's income, the percentage of homosexual attendees, restrictions on homosexual membership and leadership, and how liberal or conservative the congregation is.

```
head(NCS)
```

```
##         DENCODE3 numadlts                 clerrace  income gaypct mbrgay
## 1       Lutheran      350                    White  291000     NA   <NA>
## 2        Baptist       50 Black or African American   65000     NA   <NA>
## 3      Methodist      155                    White  233558     NA   <NA>
## 4 Roman Catholic      275                    White  159924     NA   <NA>
## 5 Roman Catholic     2200                    White 1136253     NA   <NA>
## 6       Lutheran      600                    White 1000000     NA   <NA>
##   ldrgay                        libcon
## 1   <NA> More on the conservative side
## 2   <NA>           Right in the middle
## 3   <NA> More on the conservative side
## 4   <NA> More on the conservative side
## 5   <NA> More on the conservative side
## 6   <NA> More on the conservative side
```

Univariate statistics

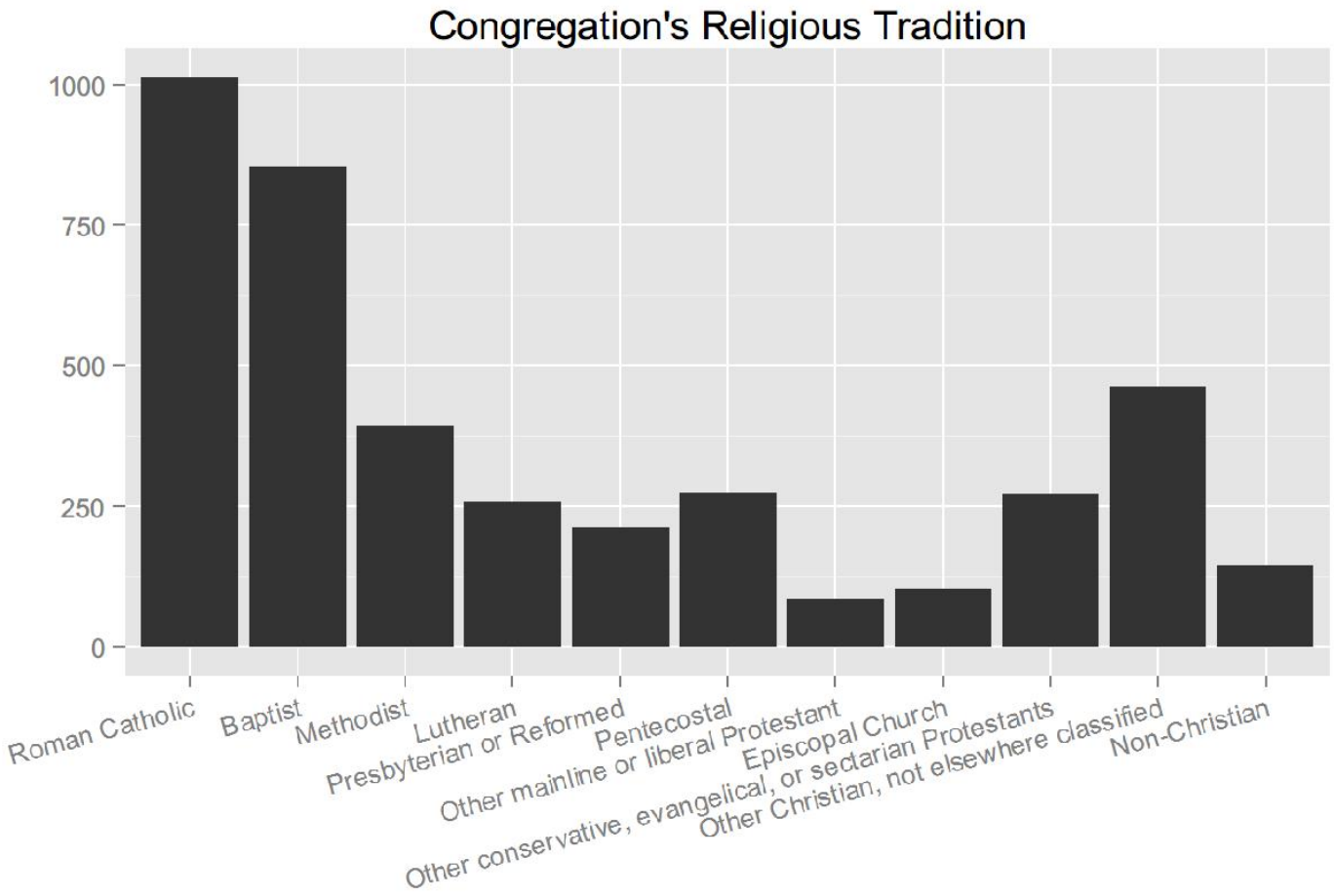
Note that all results are unweighted. Because the NCS is a hypernetwork sample, the distribution is representative of *attendees* and not congregations.

For example, the average attendee has 776 adults in their congregation (`numadlts`). Likewise, roughly half (1,388 of 2816) of attendees in the US are in congregations in which homosexuals are allowed to become members (`mbrgay`)

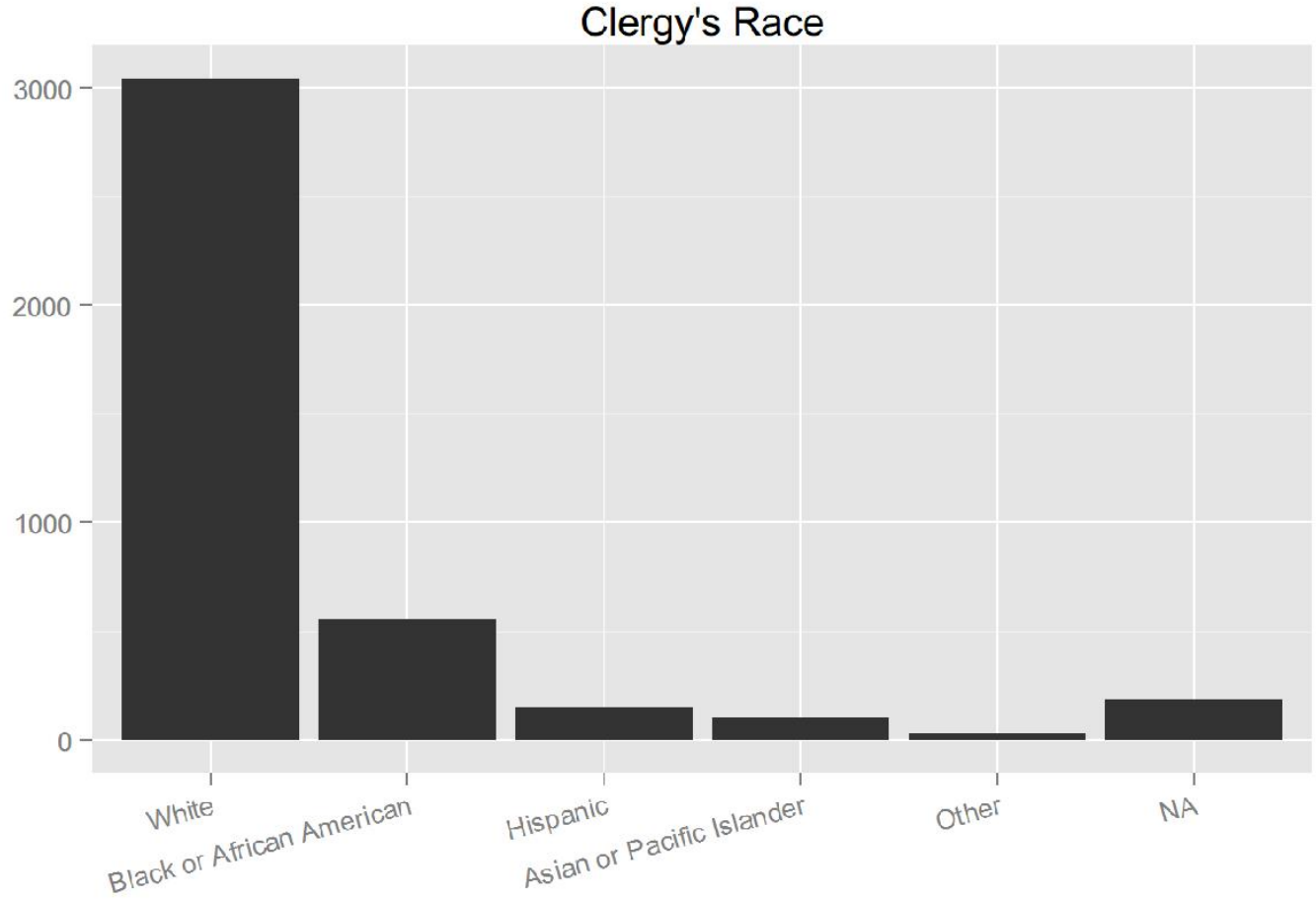
```
counter <- 1
angle1 <- c(15, 15, 0, 0, 15)
horz <- c(1, 1, 0.5, 0.5, 1)
for(i in list(DENCODE3, clerrace, mbrgay, ldrgay, libcon)){
  print(discrete.title[counter])
  print(discrete.description[counter])
  print(summary(i))

  temp.fig <- ggplot(data = NCS) +
    geom_histogram(aes(x = i)) +
    ggtitle(discrete.title[counter]) +
    ylab("") +
    xlab("") +
    theme(axis.text.x = element_text(angle = angle1[counter], hjust = horz[counter]))
  print(temp.fig)
  counter <- counter + 1
}
```

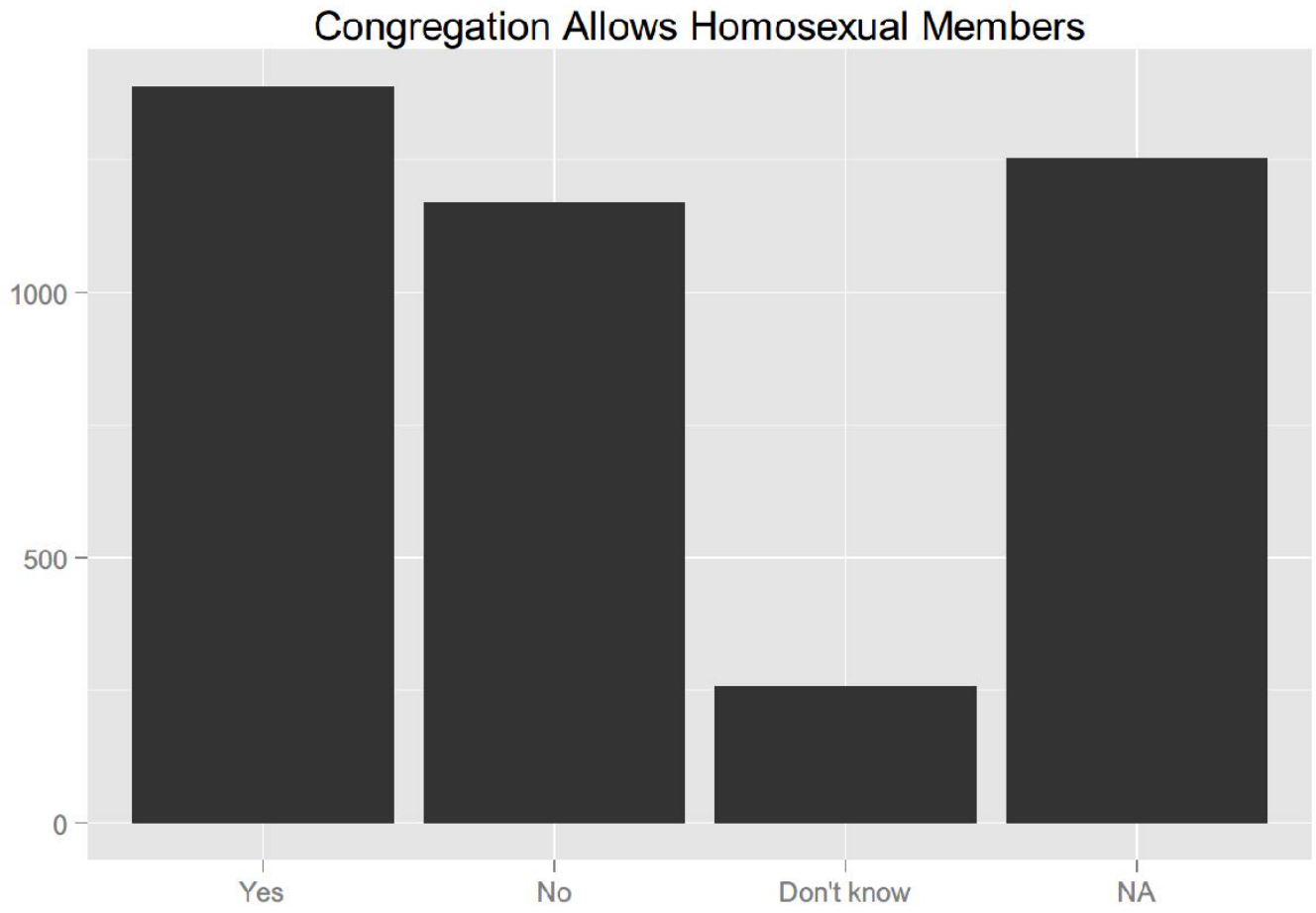
```
## [1] "Congregation's Religious Tradition"
## [1] "Catholic parishes make up the majority of the NCS data, followed by Baptists. No
other group has more than 500 congregations represented."
##                                                 Roman Catholic
##                                                           1013
##                                                       Baptist
##                                                            854
##                                                      Methodist
##                                                            392
##                                                     Lutheran
##                                                            258
##                                      Presbyterian or Reformed
##                                                            213
##                                                    Pentecostal
##                                                            274
##                         Other mainline or liberal Protestant
##                                                             85
##                                               Episcopal Church
##                                                            103
## Other conservative, evangelical, or sectarian Protestants
##                                                            272
##                     Other Christian, not elsewhere classified
##                                                            462
##                                                  Non-Christian
##                                                            145
```



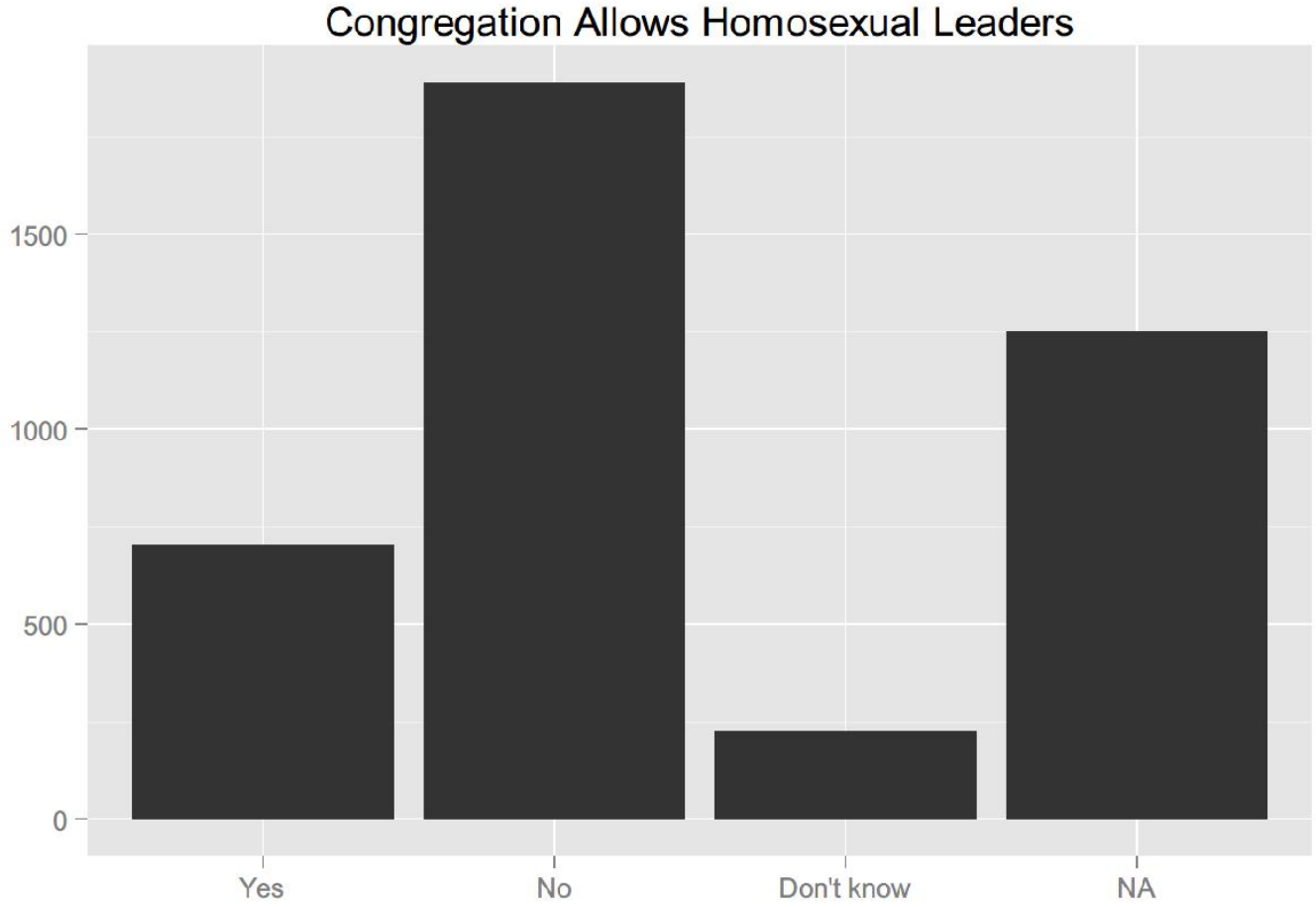

```
## [1] "Clergy's Race"
## [1] "Clergy are overwhelmingly white in the NCS."
##                    White Black or African American
##                     3045                      556
##                 Hispanic Asian or Pacific Islander
##                      151                      102
##                    Other                     NA's
##                       34                      183
```



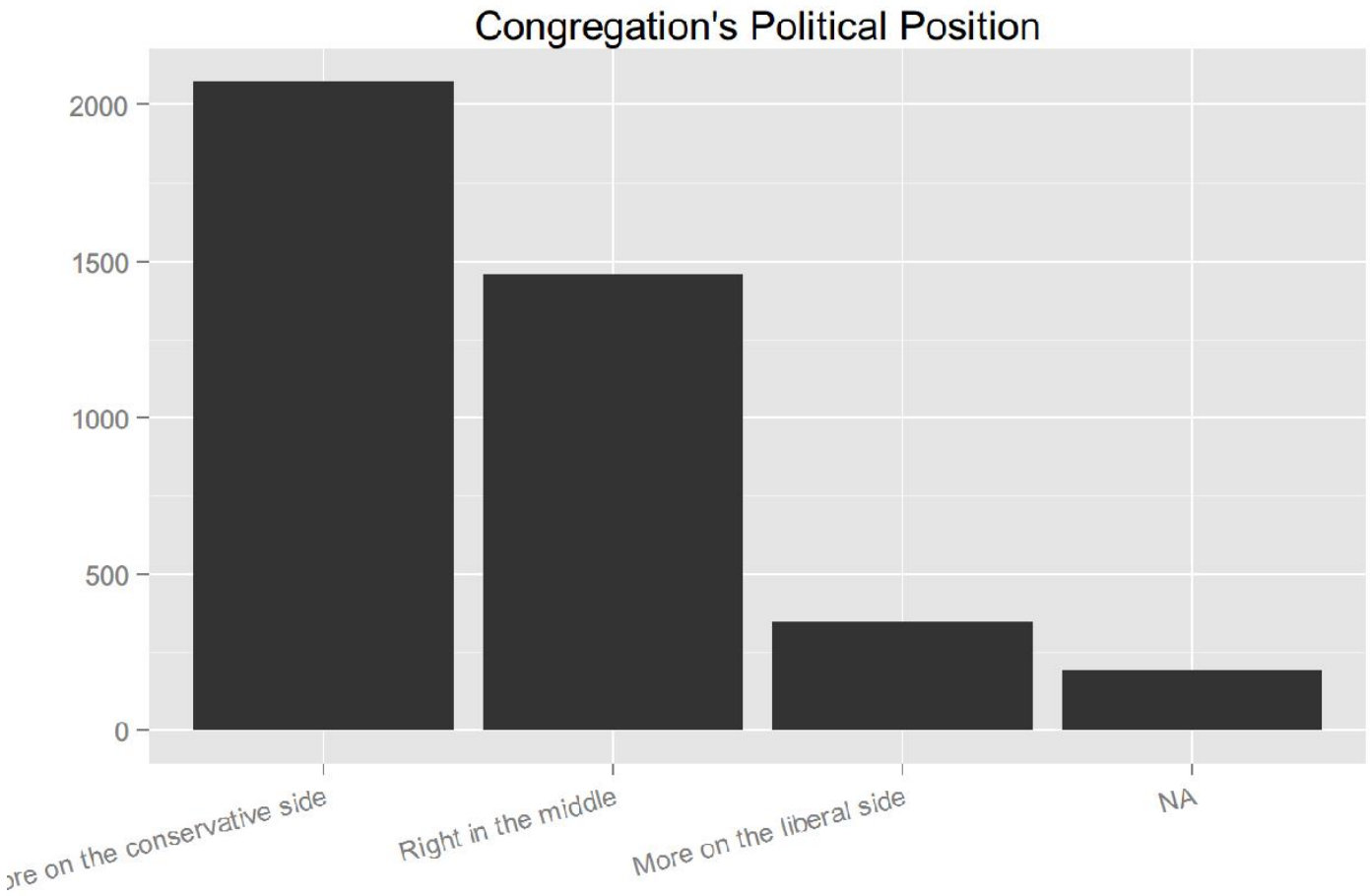

```
## [1] "Congregation Allows Homosexual Members"
## [1] "Nearly half of individuals are in a congregation that allows homosexual members."
##        Yes         No Don't know       NA's
##       1388       1170        258       1255
```




```
## [1] "Congregation Allows Homosexual Leaders"
## [1] "Most individuals are in a congregation that restricts homosexuals from becoming l
eaders."
##        Yes         No Don't know       NA's
##        705       1889        226       1251
```



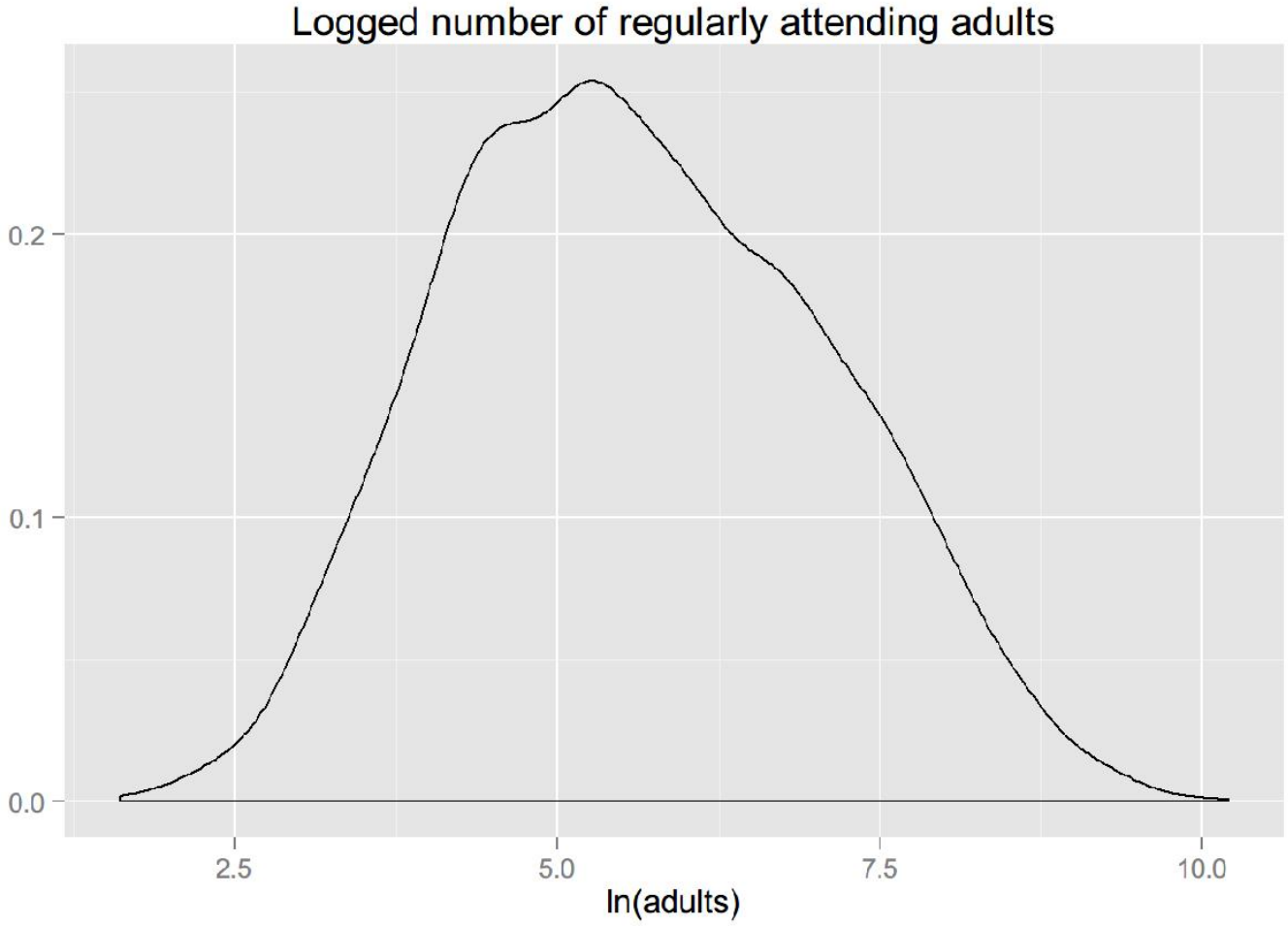

```
## [1] "Congregation's Political Position"
## [1] "Most individuals are in congregations that are conservative or moderate. Relative
ly few are in liberal congregations."
## More on the conservative side          Right in the middle
##                          2073                         1459
##      More on the liberal side                         NA's
##                           347                          192
```



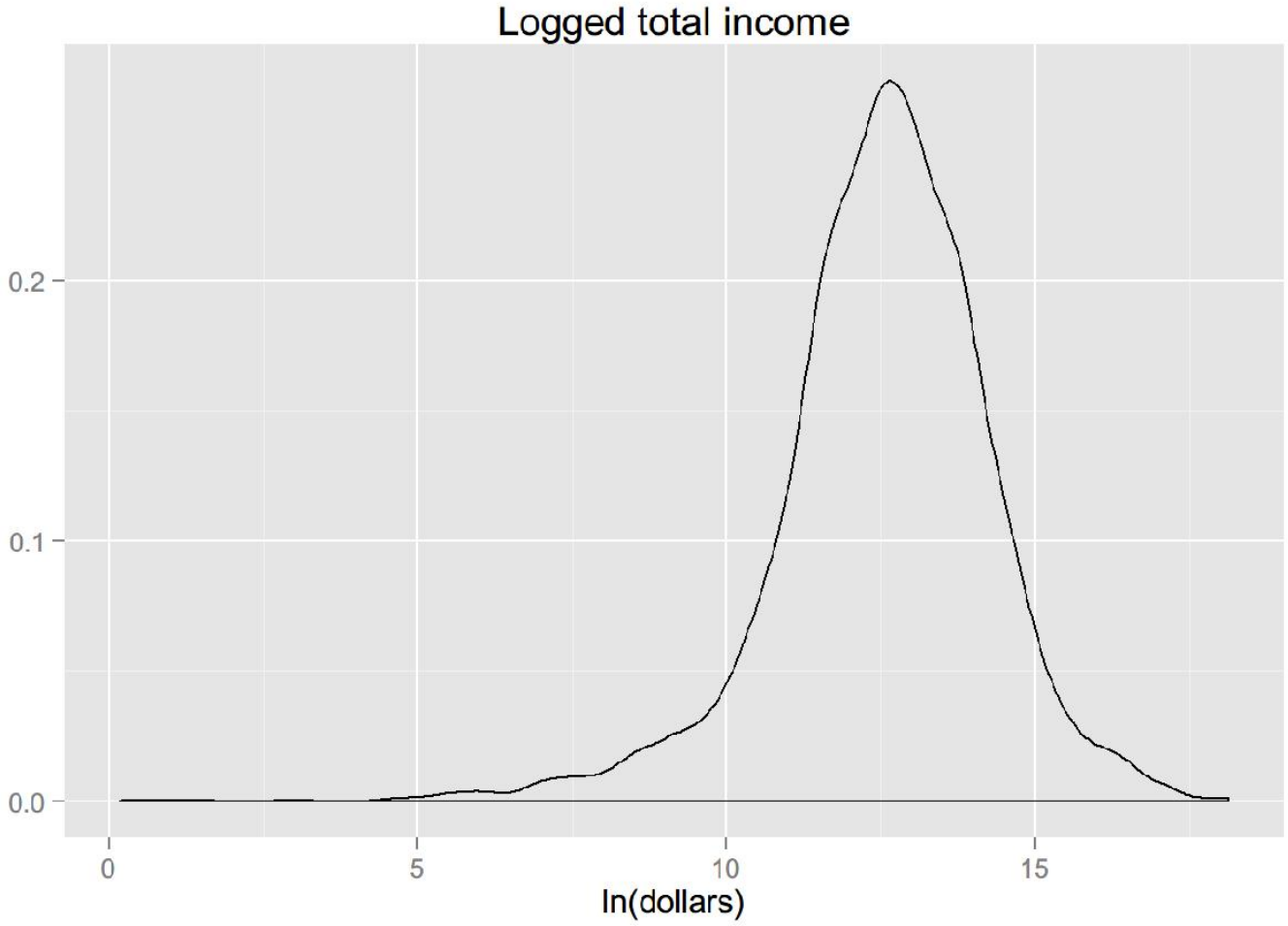

```
counter <- 1
for(i in list(numadlts, income, gaypct)){
  print(continuous.title[counter])
  print(continuous.description[counter])
  print(summary(i))

  temp.fig <- ggplot(data = NCS) +
    geom_density(aes(x = log(i))) +
    ggtitle(continuous.title[counter]) +
    ylab("") +
    xlab(continuous.label[counter])
  print(temp.fig)
  counter <- counter + 1
}
```

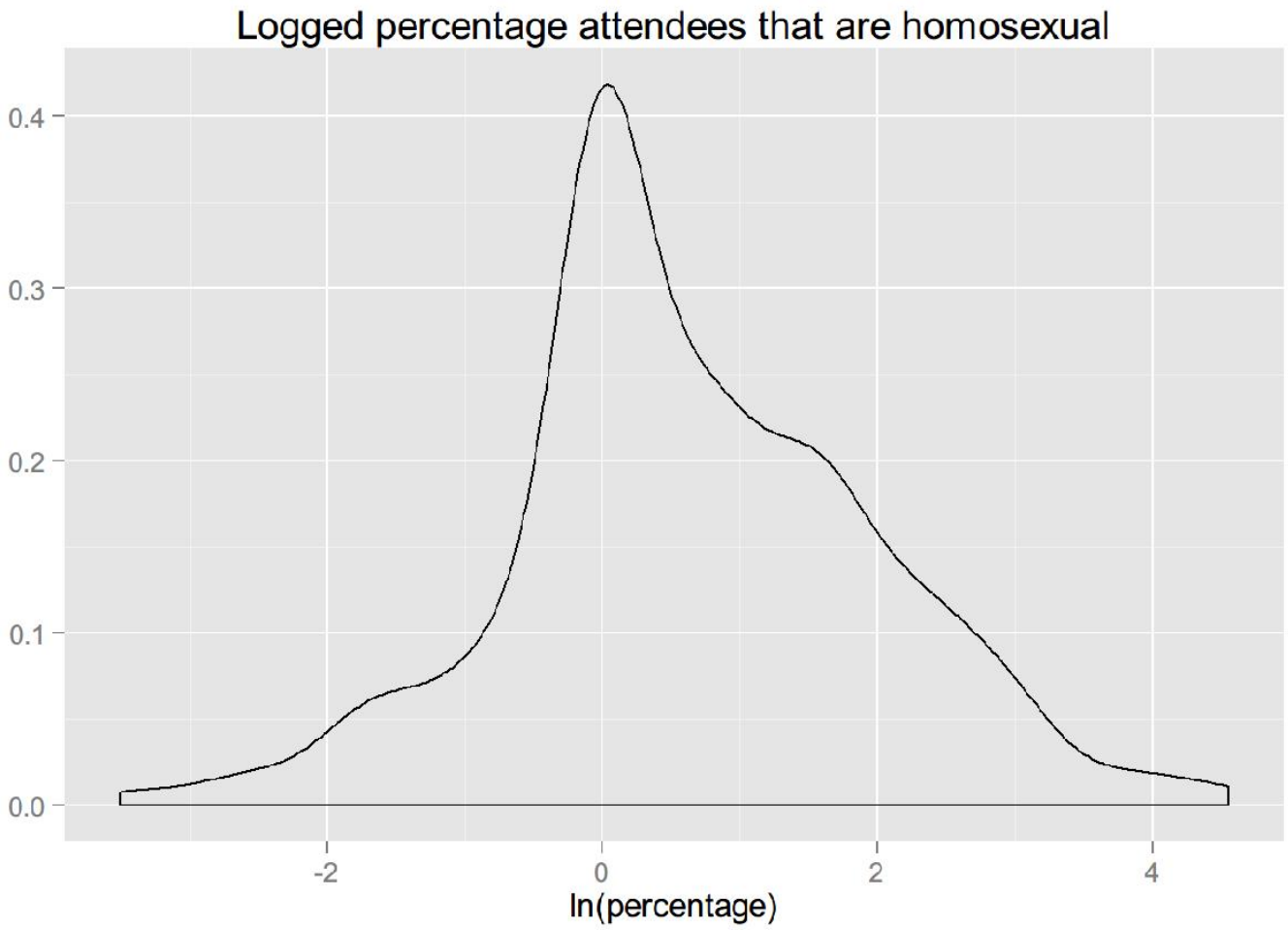
```
## [1] "Logged number of regularly attending adults"
## [1] "The average atendee has 776 adults that regularly attend their congregation. The
median attendee has 250, indicating that the distribution is highly right-skewed."
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     5.0    90.0   250.0   775.9   800.0 27000.0
```




```
## [1] "Logged total income"
## [1] "The average attendee is in a congregation that brings in just under $1,000,000 a
year, though the median attendee is in a congregation that brings in $300,000"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##        0   109400   300000   991300   800000 75000000      886
```




```
## [1] "Logged percentage attendees that are homosexual"
## [1] "While the average attendee's congregation is 1.4% made up of homosexual members,
the median is 0, indicating that most attendees have no homosexual members in their congr
egation, or at least that the respondent did not know of or chose not to report any."
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  1.4220  0.6667 95.0000    2824
```

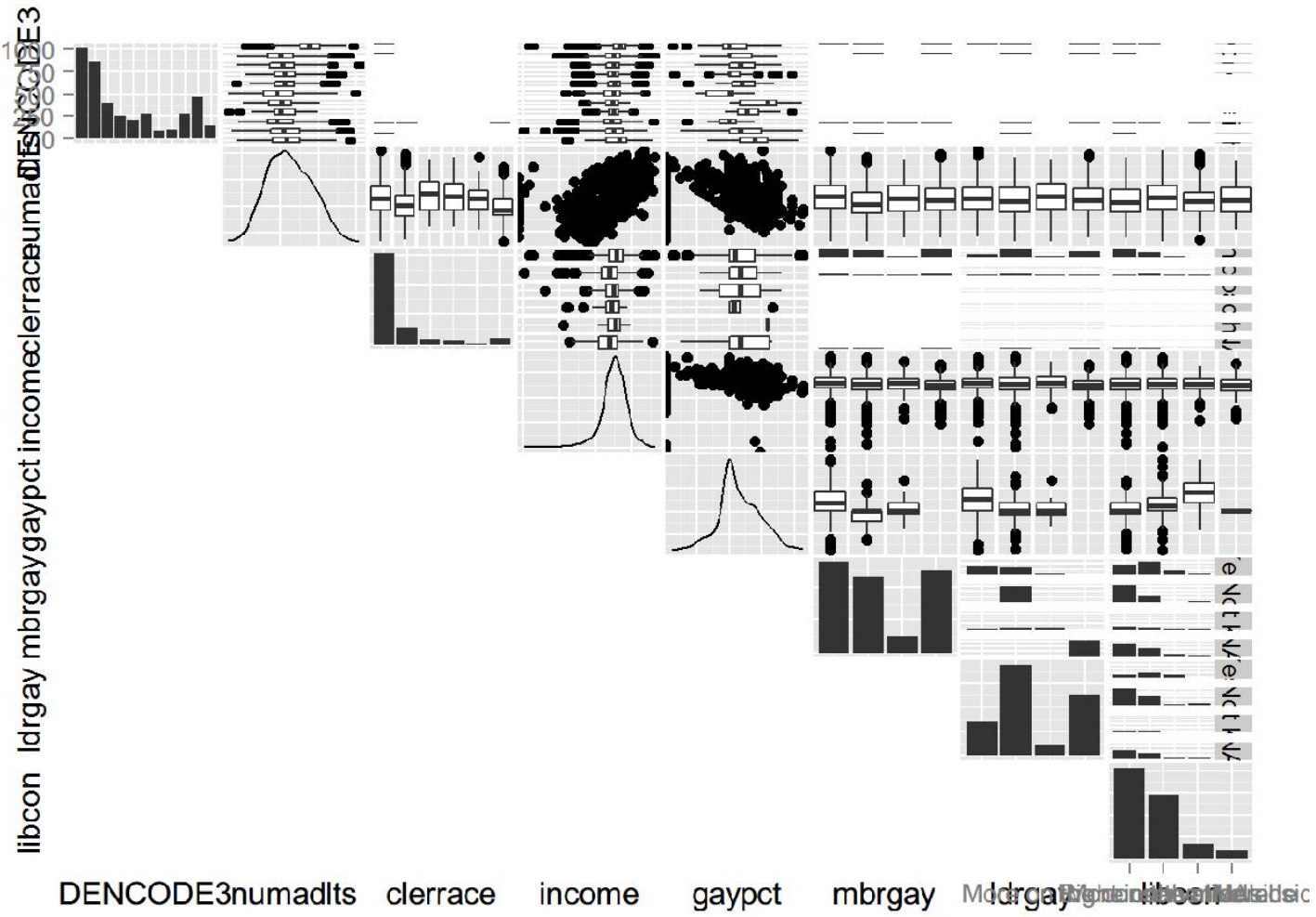


Bivariate relationships

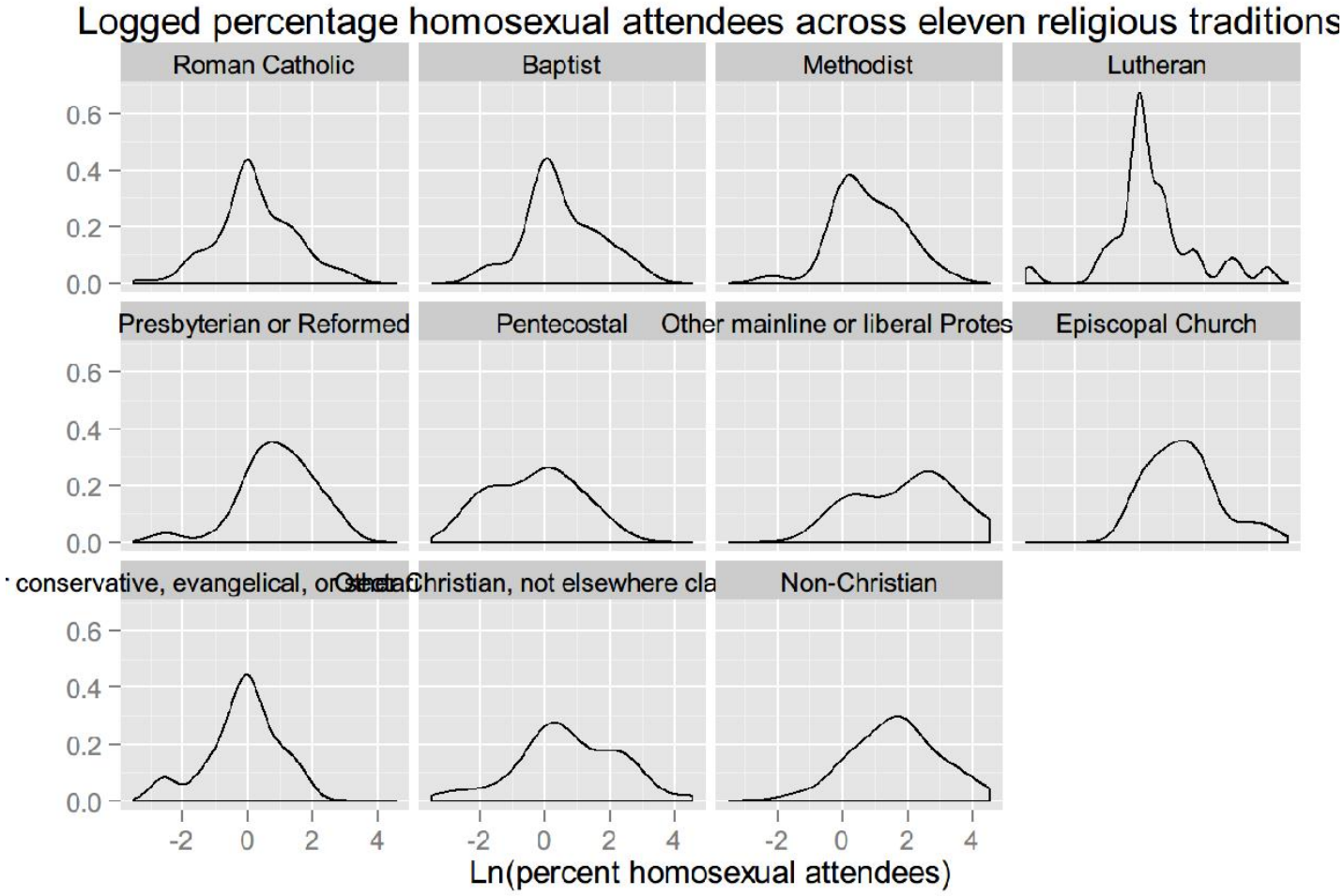
First, I produce a pairs plot of all of my variables to look for concerning patterns.

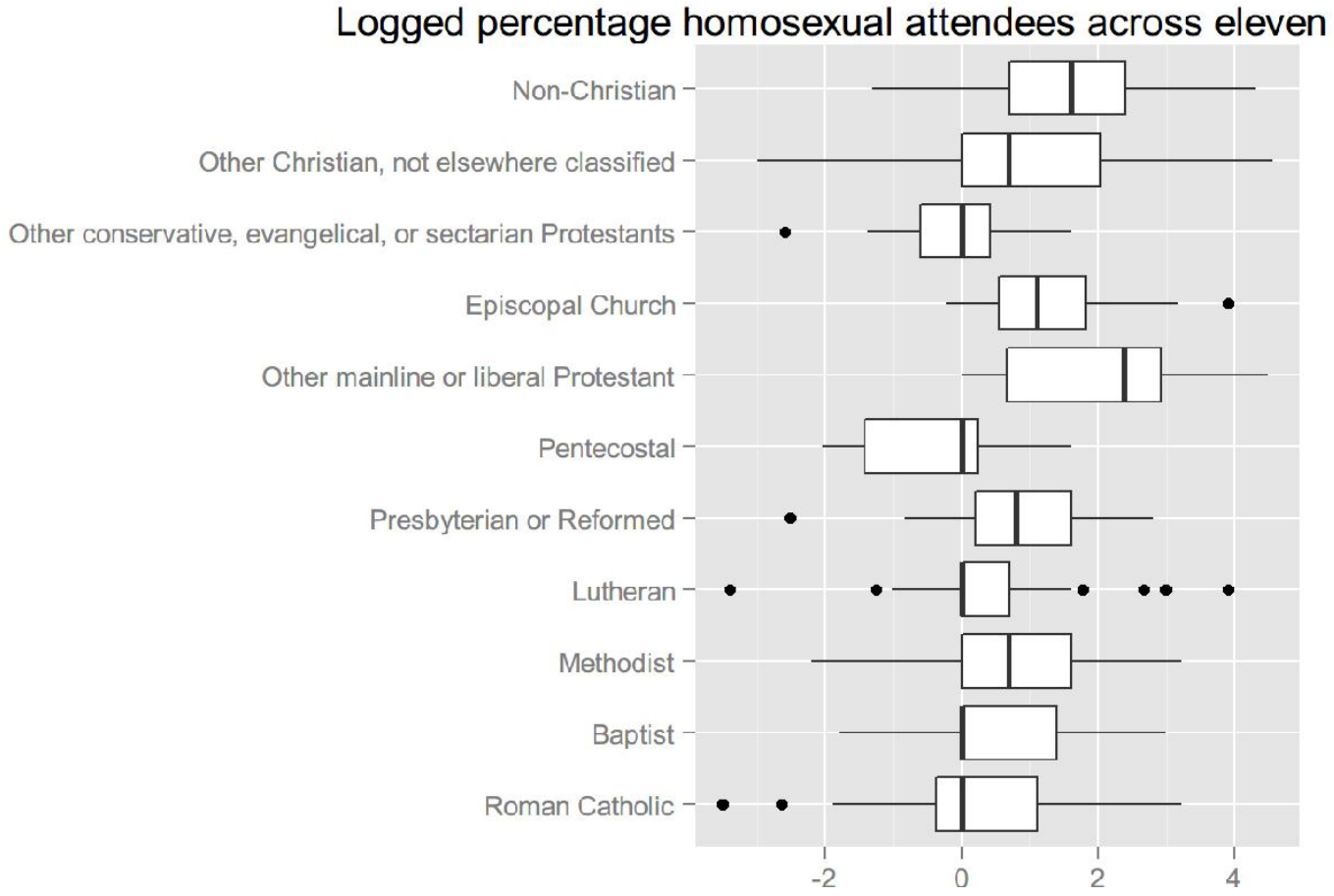
```
NCS.log <- NCS
NCS.log$numadlts <- log(NCS$numadlts)
NCS.log$income <- log(NCS$income)
NCS.log$gaypct <- log(NCS$gaypct)

ggpairs(NCS.log,
        upper = list(continuous = "smooth", combo = "box", discrete = "facetbar"),
        lower = list(continuous = "blank", combo = "blank", discrete = "blank"))
```



While not seemingly significantly different, it looks like there's some variation in the percentage of homosexuals in congregations by religious tradition. I look at this further, first, with a density plot to show the distributions, and second, with box plots.


Logged percentage homosexual attendees across eleven religious traditions
Roman Catholic
Baptist
Methodist
Lutheran
Presbyterian or Reformed
Pentecostal
Other mainline or liberal Protes
Episcopal Church
conservative, evangelical, or sectarian
Other Christian, not elsewhere cla
Non-Christian
0.0
0.2
0.4
0.6
-2
0
2
4
Ln(percent homosexual attendees)


Logged percentage homosexual attendees across eleven
Non-Christian
Other Christian, not elsewhere classified
Other conservative, evangelical, or sectarian Protestants
Episcopal Church
Other mainline or liberal Protestant
Pentecostal
Presbyterian or Reformed
Lutheran
Methodist
Baptist
Roman Catholic
-2
0
2
4