project proposal-DAaSI

A. Student

sep-2014

Contents

Research Question (in one sentence)

Is education level (highest degree attained) a significant factor in determining income earning ability?

Citation

The analysis will utilize an extract of the General Social Survey data set obtained from the National Opinion Research Center at the University of Chicago `http://www.norc.org/Research/Projects/Pages/general-social-survey.asp`
The data can be downloaded from: `http://bit.ly/dasi_gss_data`.

Data Collection

The responses from this survey are collected face-to-face with an in-person interview by the National Opinion Research Center at the University of Chicago, of adults (18+) in randomly selected households. The survey was conducted every year from 1972 to 1994 (except in 1979, 1981, and 1992). Since 1994, it has been conducted every other year. As of 2010 28 national samples with 55,087 respondents and 5,417 variables had been collected. The data collected about this survey includes both demographic information and respondents' opinions on matters ranging from government spending to the state of race relations to the existence and nature of God.

Data Cases

There are a total of 57,061 cases and 114 variables in this modified dataset. This is a cumulative data file for surveys conducted between 1972 and 2012 and not all respondents answered all questions in all years.

The analysis of this data set will consist of a subset representing complete cases from 2008, 2010, and 2012. There are 5309 cases where both degree and coninc (constant income in year 2000 dollars) exist in this subset.

Any statement of sampling error assumes that the bias in quota sampling due to the lack of control over respondent availability is slight for the study under consideration. For details see: `http://publicdata.norc.org:41000/gss/documents//BOOK/GSS_Codebook_AppendixA.pdf`

Data Variables

The variables used in this analysis are:

- coninc: [(ORDINAL CATEGORICAL) FAMILY INCOME IN CONSTANT DOLLARS] Literal Question: Inflation-adjusted family income.

Income variables (INCOME72, INCOME, INCOME77, INCOME82, INCOME86, INCOME91) are recoded in six-digit numbers and converted to year 2000 dollars. The collapsed numbers above are for convenience of display only. Since this variable is based on categorical data, income is not continuous, but based on categorical mid-points and imputations.

- degree: [(ORDINAL CATEGORICAL) RESPONDANT'S HIGHEST DEGREE] Literal Question: Do you have any college degrees? (IF YES: What degree or degrees?) CODE HIGHEST DEGREE EARNED.

Categories are: Less than high school, High school, Associate/Junior College, Bachelor's, Graduate

Type of Study

This is an analysis of an observational study. As mentioned in the Data Collection section, the responses from this survey were collected face-to-face with an in-person interview. An observational study finds the relationship between changes that already exist whereas an experiment requires one to observe what happens when they make some sort of change. The subjects are selected randomly, and are sufficient in size but not overly large to question the independence assumption.

Scope of Inference - Generalizability

Because the survey respondents are randomly selected from the population of the U.S.A, is sufficiently large without being too large compared to the population, it can be said to be generalizable to all persons in the United States.

Scope of Inference - Causality

The results of this analysis will determine whether there is an association between the variables that cannot be explained by chance, given the available data. It will also determine the strength of any association. This analysis will not determine causation though. In order to demonstrate causation, an experiment would be necessary. A researcher would need to randomly assign subjects to pursue different educational goals, and then compare income with educational level. Additionally, some method such as blocking may be necessary to adjust for factors such as socio-economic status, parental education, or other well-known predictors of innate cognitive ability.

Exploratory Data Analysis

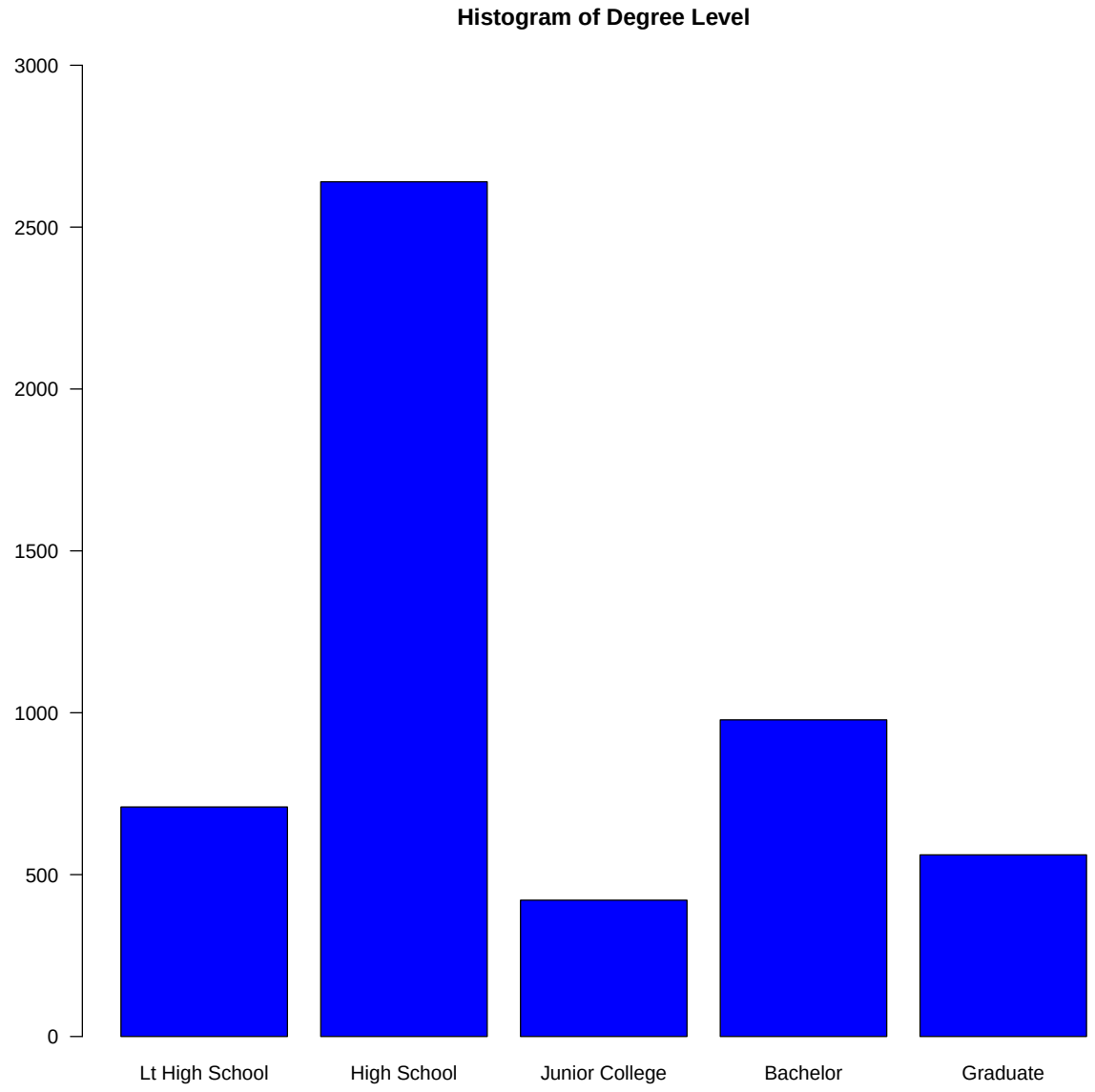
```
##             degree         coninc
##  Lt High School: 709   56059  : 201
##  High School   :2640   54203  : 173
##  Junior College: 421   51705  : 153
##  Bachelor      : 978   34470  : 151
##  Graduate      : 561   63195  : 151
##                        36135  : 147
##                        (Other):4333
```

```
##
## Lt High School    High School Junior College       Bachelor       Graduate
##            709           2640            421            978            561
```

```
##
## Lt High School    High School Junior College       Bachelor       Graduate
##      0.1335468      0.4972688      0.0792993      0.1842155      0.1056696
```

```
## [1] 5309
```

All 5309 records are accounted for and there are no zero values. The proportions are reasonable, although the proportion of High School graduates comprises 49.7 percent of the observations. This should not be a hinderance to the analysis.




```
## png
##   3
```

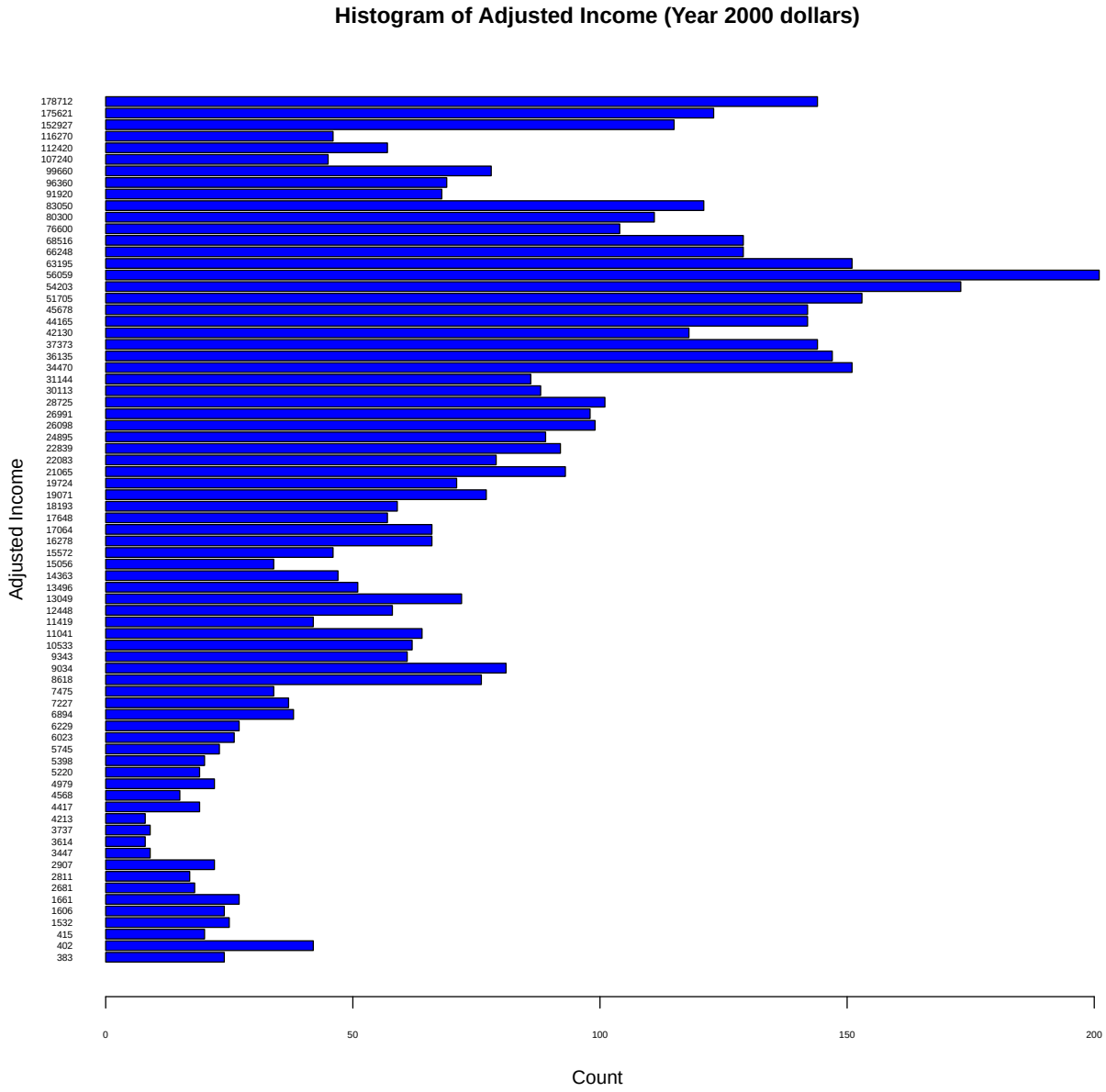
```
## pdf
##   2
```

This is a visual representation of the same data. It is included only as a convenience to reviewers.

```
##    383    402    415   1532   1606   1661   2681   2811   2907   3447   3614   3737   4213   4417
```

```
##    24    42    20    25    24    27    18    17    22     9     8     9     8    19
##  5398  5745  6023  6229  6894  7227  7475  8618  9034  9343 10533 11041 11419 12448 13
##    20    23    26    27    38    37    34    76    81    61    62    64    42    58
## 15056 15572 16278 17064 17648 18193 19071 19724 21065 22083 22839 24895 26098 26991 28
##    34    46    66    66    57    59    77    71    93    79    92    89    99    98
## 34470 36135 37373 42130 44165 45678 51705 54203 56059 63195 66248 68516 76600 80300 83
##   151   147   144   118   142   142   153   173   201   151   129   129   104   111
## 99660 107240 112420 116270 152927 175621 178712
##    78    45    57    46   115   123   144
```

The minimum value in any one coninc bucket is 8, which exceeds the (arbitrary) minimum of 5 necessary for inclusion. I may increase the subset of data to include 2006 and compare the results for differences.

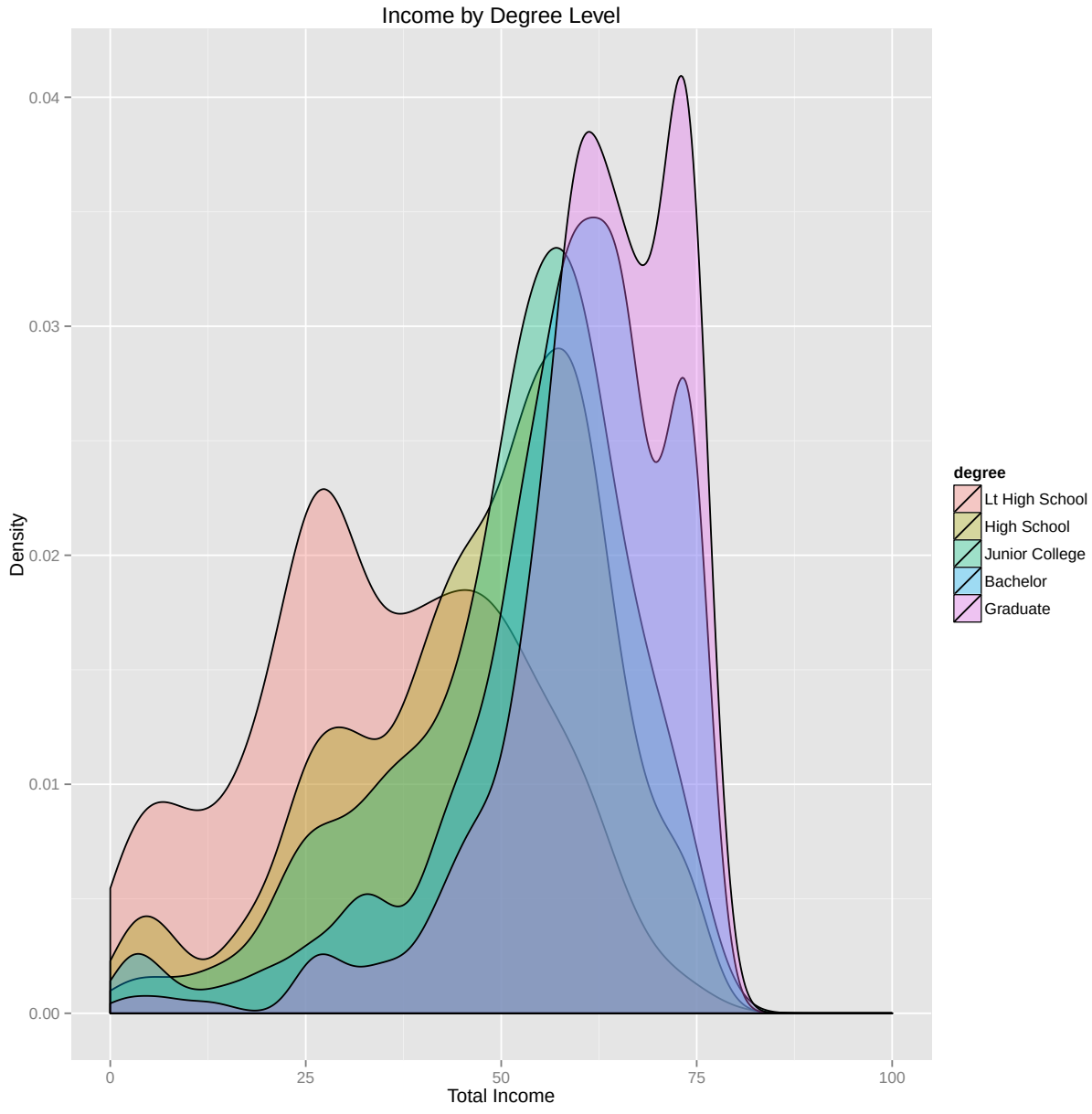



```
## png
```

```
##   3
```

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

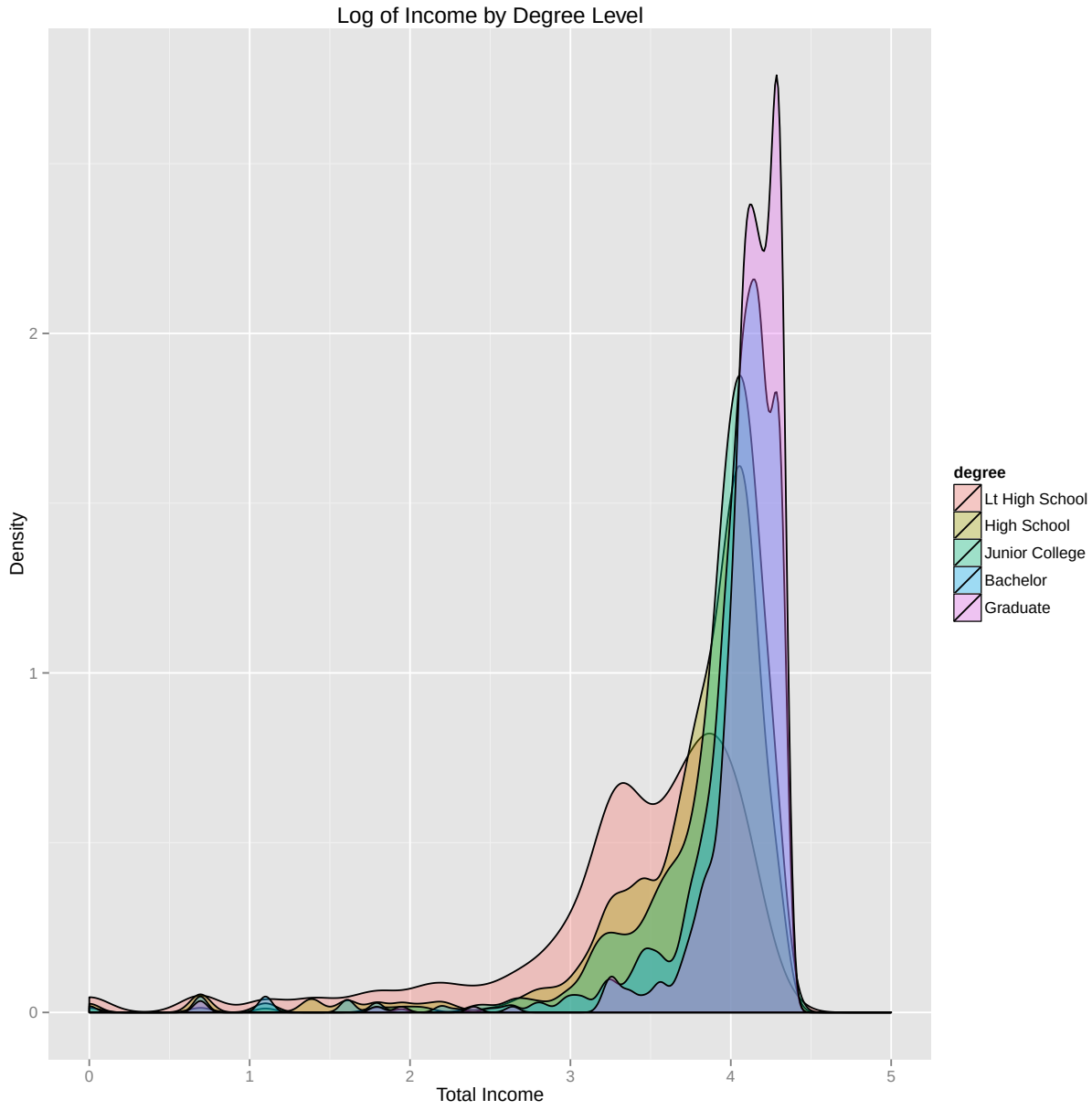
The barplot for coninc indicates a left-skewed distribution indicating a transformation might be necessary or desirable.




```
## png
##   3
```

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

This plot again, reveals the left-skewed distribution, but indicates that it exists for only those with only a High School education.




```
## png
##   3

## pdf
##   2
```

This plot is a log transformation of the coninc data by degree level. The transformation exacerbates the skewness, indicating that it is an improper transformation. The exact nature of data modification will be addressed in the detailed analysis.

```
##
```

```
## Call:
## lm(formula = log(as.numeric(coninc)) ~ degree, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.971 -0.084  0.143  0.320  0.985
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.3325     0.0226   147.5   <2e-16 ***
## degreeHigh School       0.3887     0.0254    15.3   <2e-16 ***
## degreeJunior College    0.5214     0.0370    14.1   <2e-16 ***
## degreeBachelor          0.6383     0.0297    21.5   <2e-16 ***
## degreeGraduate          0.7591     0.0340    22.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.602 on 5304 degrees of freedom
## Multiple R-squared:  0.112,  Adjusted R-squared:  0.111
## F-statistic:  167 on 4 and 5304 DF,  p-value: <2e-16
```

This is the least-squares regression analysis. As indicated by the plots, the relationship between the chosen variables in not linear. The results, therefor, are pretty much useless until a proper transformation is discovered that linearizes the relationship between them. This will be addressed in the detailed analysis.

Data Page

These are the first 30 observations used in this proposal.

```
##                degree coninc
## 51021     High School   6229
## 51022        Graduate  99660
## 51023     High School  37373
## 51024     High School  31144
## 51025     High School  26991
## 51031        Graduate 175621
## 51032        Graduate  99660
## 51033        Bachelor  17648
## 51034     High School  31144
## 51037     High School  68516
## 51038     High School  56059
## 51039        Bachelor 175621
## 51040 Junior College  45678
## 51041     High School 175621
## 51042 Lt High School  37373
## 51043     High School  56059
## 51047     High School  31144
## 51048        Graduate 175621
## 51049 Junior College  45678
## 51050        Graduate  68516
## 51051        Graduate  68516
## 51054     High School  17648
## 51056        Bachelor  68516
```

```
## 51060       Bachelor  68516
## 51066       Graduate  99660
## 51067       Graduate  83050
## 51068       Bachelor 175621
## 51069    High School  19724
## 51070       Graduate  68516
## 51071       Bachelor  56059
```

References

Wikipedia: http://en.wikipedia.org/wiki/General_Social_Survey