What is the relationship between education level and income?

Introduction:

“An investment in education pays the best interest” - Benjamin Franklin.

It is often said that education is the key to success. Education broadens one’s mind, builds confidence to make decisions, face challenges and accept failures, and opens the door to new and better job opportunities. (Notwithstanding, some of the world’s richest and most successful people were school dropouts. Steve Jobs never graduated from college while Bill Gates dropped out of university; they are the outliers who prove that success is not completely dependent on education.)

On the other hand, consider the cost of education. College costs have surged 500% in the US since 1985 (Jamrisko & Kolet, 2013). Average tuition at private schools was $30,094 in 2013 - 2014, up from $18,060 in 2002 - 2003 (Gage & Lorin, 2014). Education debt exceeded $1 trillion in the third quarter of 2013 (Gage & Lorin, 2014) and the average debt load for the class of 2012 was $29,400 (Ellis, 2013). Given the state of the economy today, a college education is by no means a guarantee for a stable and decent paying job.

In light of the above, this paper will examine the following question: *“What is the relationship between one’s highest education level attained and current income?”*. Do all levels of education lead to higher income? Or do certain education qualifications lead to greater increases in income?

Data:

To examine the research question, data from the General Social Survey (“GSS”) was used. The GSS is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. While the GSS provides data from 1972 - 2012, this paper will examine only data from 2012 to control for possible confounding variables including time, changes in the education system, and rising levels of income.

Data collection for the GSS was conducted through (i) computer-assisted personal interviews, (ii) face-to-face interviews, (iii) and telephone interviews. For the 2012 GSS data, the cases were a sample of all English and Spanish speaking people age 18 and over who were living in households at the time of the survey (or non-institutionalised) in the US.

For this paper, the two variables studied are the highest level of education attained (“education”) and total family income in constant dollars (“income”). Given that there is no data collected on *personal* income, *total family* income will be examined as a proxy. In addition, while a measure of income in *current dollars* is available, this paper will examine income in *constant dollars* (i.e., inflation-adjusted income) to allow for comparison across time with other studies. Education is a categorical variable with 5 levels (i.e., “Less than High School”, “High School”, “Junior College”, “Bachelor”, “Graduate” (i.e., Masters and above)) and is labeled “degree” in the dataset. Income is a continuous variable ranging from $383 - $178,712, with a median of $34,470, and is labeled “coninc” in the data set.

The study is an observational study given that there was no random assignment of individuals to different conditions/treatments. Full probability sampling, where every individual had a chance of being selected, was conducted. Notwithstanding, there were exceptions that will be discussed below. The sampling method was stratified sampling; the population was stratified first by region followed by country. With regard to experimental design, there was no random assignment of individuals to different conditions or treatments.

The population of interest is the working US population. As full probability sampling was conducted, the findings can be generalised to the entire working US population. Potential sources of bias may arise given that the GSS 2012 did not sample from (i) minors and (ii) people who do not speak either English and Spanish. For (i), the bias is likely to be minor (pun intended) given that our interest is examining the working population's income, assuming that minors are still pursuing an education and do not have an income. With regard to (ii), the 2011 census on language use suggests that only 0.294% of the US population do not speak English and/or Spanish (Ryan, 2013). Thus, the biases in the 2012 GSS will have a negligible impact on the generalizability of this study.

The data cannot be used to establish causal links between the variables of interest as there was no random assignment to the explanatory/independent variable (i.e., education).

Exploratory data analysis:

Education by Count and Percentage

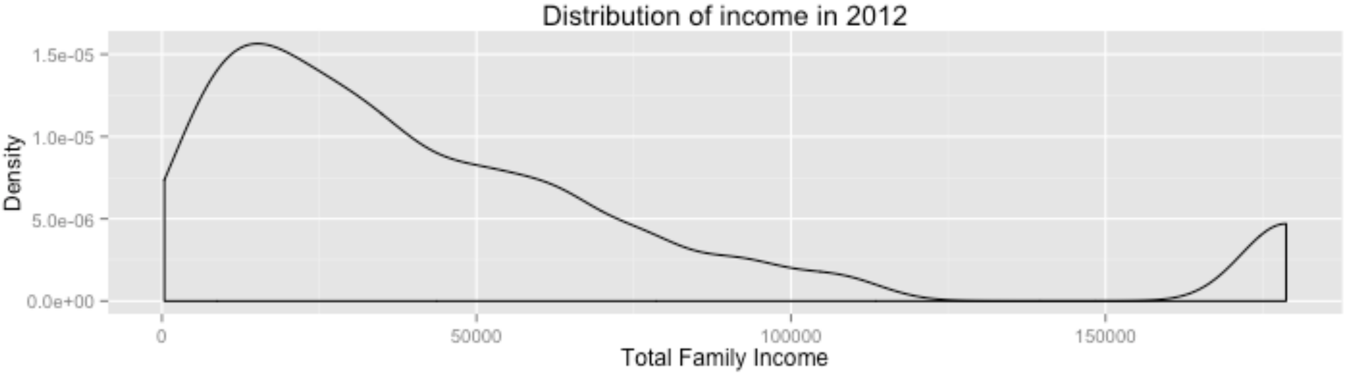
```
## Lt High School    High School Junior College       Bachelor       Graduate 
##            222            869            130            319            180 
```

```
## Lt High School    High School Junior College       Bachelor       Graduate 
##        0.12907        0.50523        0.07558        0.18547        0.10465 
```

A majority of the US population has an education level of high school level and below, with approximately 29% having a bachelor degree and above.

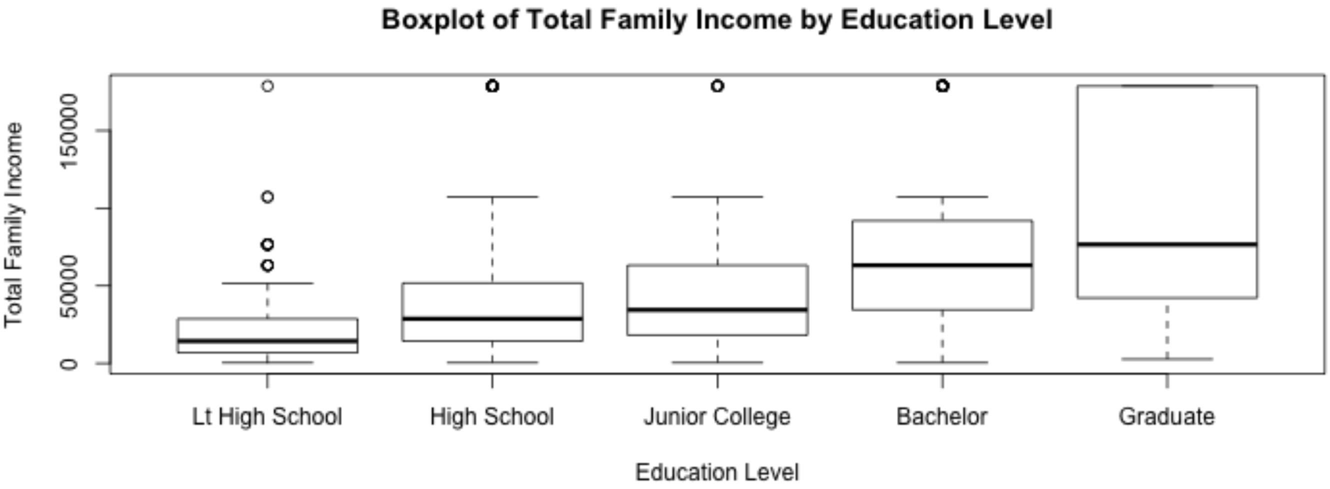
Summary and Density Distribution of 2012 GSS Current Income

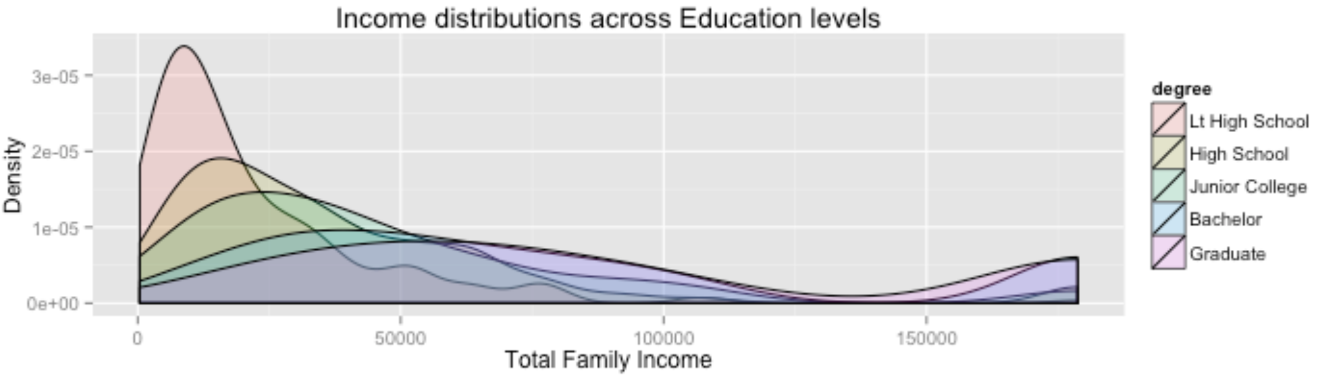
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     383   16300   34500   48800   63200  179000 
```



The median income in 2012 is $34,450, with a mean of $48,800, and range of $383 - $179,000. Income distribution is bimodal and right skewed, with a peak at approximately $15,000 and another at the extreme right tail, with a gap between $125,000 and $160,000.

Boxplot and Overlapping Density Distribution of Current Income across Education





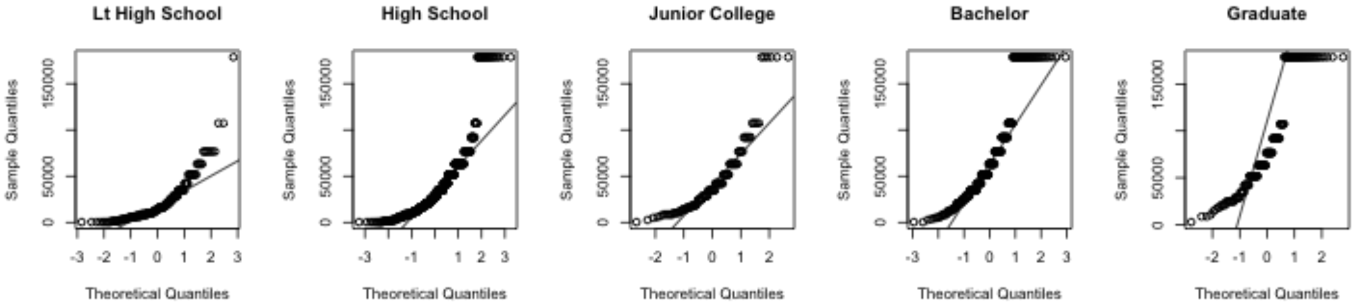
The box plots suggest a significant and positive relationship between higher education and income. The overlapping distribution plots further hint at the strong relationship between education and income, warranting a deeper investigation of the research question.

Inference:

The hypotheses for this study are as stated below:
Null Hypothesis: The mean income is the same across all levels of education.
Alternative Hypothesis: At least one pair of mean incomes are different from each other.

There are three conditions for analysis of variance ("ANOVA"), namely (i) independence, (ii) approximate normality, and (iii) equal variance. For (i), the data was randomly sampled with full probability sampling, and the sample size of each education group is less than 10% of the population and independent of each other. For (ii), while the normal probability plots (below) for each education group show that the data is right skewed and deviates from normality, this is mitigated by the large sample sizes for each education group. For (iii), the previous box plots of income across education levels show roughly equal variance for the High School, Junior College, and Bachelor groups, while the Less than High School group has lower variance and the Graduate group has higher variance. To address this, a non-parametric test such as the Kruskal-Wallis test can be used; however, this is not covered under the class syllabus. Thus, this study will proceed with the ANOVA analysis.

Normal Probability Plots of Current Income at each Education level



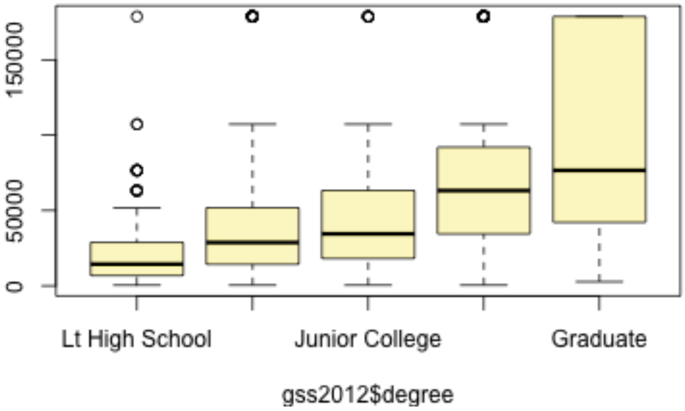
As the means between more than two groups (i.e., five) will be compared, the study will use the ANOVA. The ANOVA analysis will compare the means across the five groups and determine if the observed differences are due to between-group variability (i.e., education) or within-group variability (other factors).

Anova of Current Income and Education

```
# anova of gss2012$coninc ~ gss2012$degree
inference(y = gss2012$coninc, x = gss2012$degree, est = "mean", type = "ht",
    null = 0, alternative = "greater", method = "theoretical")
```

```
## Response variable: numerical, Explanatory variable: categorical
## ANOVA
## Summary statistics:
## n_Lt High School = 222, mean_Lt High School = 21657, sd_Lt High School =
22693
## n_High School = 869, mean_High School = 37665, sd_High School = 35146
## n_Junior College = 130, mean_Junior College = 46221, sd_Junior College =
39487
## n_Bachelor = 319, mean_Bachelor = 75871, sd_Bachelor = 55549
## n_Graduate = 180, mean_Graduate = 90371, sd_Graduate = 58309
```

```
## H_0: All means are equal.
## H_A: At least one mean is different.
## Analysis of Variance Table
##
## Response: y
##             Df   Sum Sq  Mean Sq F value Pr(>F)
## x            4 8.17e+11 2.04e+11     118 <2e-16
## Residuals 1715 2.98e+12 1.74e+09
##
## Pairwise tests: t tests with pooled SD
##                 Lt High School High School Junior College Bachelor
## High School                  0          NA             NA       NA
## Junior College               0      0.0291             NA       NA
## Bachelor                     0      0.0000              0       NA
## Graduate                     0      0.0000              0    2e-04
```



```
## [1] "Bonferroni Correction: Modified alpha level = 0.5/((5*4)/2) = 0.005"
```

Income Quantiles at each Education

```
## gss2012$degree: Lt High School
##     0%    25%    50%    75%   100%
##    383   6894  14363  28725 178712
## ------------------------------------------------------------
## gss2012$degree: High School
##     0%    25%    50%    75%   100%
##    383  14363  28725  51705 178712
## ------------------------------------------------------------
## gss2012$degree: Junior College
##     0%    25%    50%    75%   100%
##    383  18193  34470  63195 178712
## ------------------------------------------------------------
## gss2012$degree: Bachelor
##     0%    25%    50%    75%   100%
##    383  34470  63195  91920 178712
## ------------------------------------------------------------
## gss2012$degree: Graduate
##     0%    25%    50%    75%   100%
##   2681  42130  76600 178712 178712
```

The p-value from the ANOVA is almost 0 (i.e., less than 2.2e-16). Thus, we reject the null hypothesis, at the 5% significance level, and conclude that the data provides convincing evidence that at least one pair of income means are different from each other.

To determine which education levels differ in mean incomes, we examine the pairwise tests with a modified significance level of 0.5% (based on the Bonferroni correction). At the 0.5% significance level, p-values from all the pairwise tests are significant, except for the high school-junior college pair. Thus, we conclude the data provides convincing evidence that mean income is different across all education pairs except for the high school-junior college pair. The box plots of income for high school and junior college education, with the medians close to each other, alluded to this. There is no associated confidence interval for the ANOVA technique and thus there is nothing to compare the ANOVA results with.

Conclusion:

To summarise the findings, in 2012, there is a significant and positive relationship between higher education level and income (i.e., higher education qualifications lead to higher income). Notwithstanding, it should be noted that there is no significant difference in income between the high school and junior college education levels.

Is getting a bachelor's degree worth the cost? For this, we examine income quantiles across education levels. Median income for bachelor's degree holders is nearly twice that of junior college graduates, with a difference of $28,735. In the introduction, it was shared that the average education debt load for 2012 was $29,400. Assuming a buoyant economy and decent job, the increase in median income from a bachelor's degree should pay off the education debt incurred within a year.

Next, we examine the incomes of between bachelor's degree and graduate degree holders. Based on median income, graduate degree holders earn $13,405 more than bachelor's degree holders; this may not seem like much relative to the cost of a graduate education. However, examining income at the 75th percentile, graduate degree holders earn nearly twice that of bachelor's degree holders, with a difference of $86,762. It seems that for the top 25%, a graduate degree pays better interest than a bachelor’s degree.

However, this analysis does not imply that income is dependent *solely* on education level. Referring to the box plots, there are outliers at every education level that have extremely high income. This is also seen in the overlapping distribution plots, where high income earners at the right tail of the distribution consists of all education levels (though predominantly bachelor's degree and graduate degree holders).

One shortcoming of the study is the current data not including people who do not speak either English or Spanish. While this is only 0.294% of the population, future research could try to include this segment of the population. Another limitation is that *total family income*, instead of *personal income*, was used in the study as the measure for income; perhaps data on personal income could be collected and analysed in future studies. Another shortcoming is the lack of equal variance in income across education levels; to address this issue, the Kruskal-Wallis test can be used in further research and analysis.

The current analysis does not take into account possible extraneous variables such as age, gender, and family background (i.e., family income at the age of 16). Future research could examine the relationship between these variables and current income in a multiple regression model (see below).

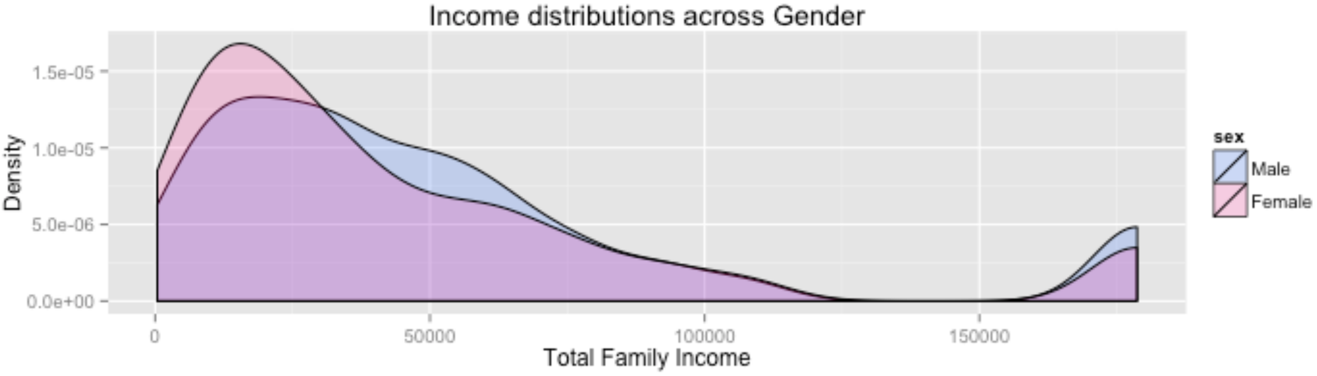
Further Analysis

“Education, beyond all other devices of human origin, is the great equaliser of the conditions of men, the balance-wheel of the social machinery” - Horace Mann

A multiple regression analysis has been done with (i) education, (ii) age, (iii) gender, and (iv) family income at the age of 16 as explanatory variables and current income as the response variable. Based on the analysis, (i) higher education continues to be strongly and positively related to current income, (ii) age does not have a significant relationship with income, (iii) gender is significantly related to income, with females earning less, and (iv) family income is significantly related to current income *only* if family income was "above average”, but not “far above average”.

It seems that while education is able to lift millions out of poverty, it has not been able to completely level the playing field for those coming from a poor family background and women. It will be interesting to observe how massive online open courses ("MOOCs”) will have an impact on this. An overlapping plot of income across gender and a summary of the regression analysis is appended below.

Overlapping Density Distribution of Current Income across Gender, and multiple regression model of Education, Family Income, and Gender on Current Income



```
## 
## Call:
## lm(formula = gss2012$coninc ~ gss2012$degree + gss2012$incom16 + 
##     gss2012$sex)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -92835 -24853  -9533  14056 154903 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                         22427       4106    5.46  5.4e-08 ***
## gss2012$degreeHigh School           15057       3153    4.77  2.0e-06 ***
## gss2012$degreeJunior College        23795       4593    5.18  2.5e-07 ***
## gss2012$degreeBachelor              51527       3704   13.91  < 2e-16 ***
## gss2012$degreeGraduate              66001       4218   15.65  < 2e-16 ***
## gss2012$incom16Below Average         1152       3811    0.30   0.7624    
## gss2012$incom16Average               2916       3662    0.80   0.4259    
## gss2012$incom16Above Average        13024       4228    3.08   0.0021 ** 
## gss2012$incom16Far Above Average     7504       7162    1.05   0.2949    
## gss2012$sexFemale                   -6123       2007   -3.05   0.0023 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 41400 on 1710 degrees of freedom
## Multiple R-squared:  0.227,  Adjusted R-squared:  0.223 
## F-statistic: 55.8 on 9 and 1710 DF,  p-value: <2e-16
```

```
## [1] "Note: Age was not included in the regression model as a prior t-test
had showed no significant relationship between age and current income"
```

References:

Smith, Tom W., Michael Hout, and Peter V. Marsden. General Social Survey, 1972-2012 [Cumulative File]. ICPSR34802-v1. Storrs, CT: Roper Center for Public Opinion Research, University of Connecticut /Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributors], 2013-09-11. doi:10.3886/ICPSR34802.v1. URL: http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/34802/version/1 Dataset URL: http://bit.ly/dasi_gss_data

Jamrisko, M., and Kolet, I. (2013, Aug 2013). College Costs Surge 500% in U.S. Since 1985: Chart of the Day. Retrieved from http://www.bloomberg.com/news/2013-08-26/college-costs-surge-500-in-u-s-since-1985-chart-of-the-day.html

Gage, C. S., and Lorin, J. (2014, Jan 15). Fed Student-Loan Focus Shows Recognition of Growth Risk. Retrieved from http://www.bloomberg.com/news/2014-01-15/fed-student-loan-focus-recognizes-threat-to-u-s-economy.html

Ellis, B. (2013, Dec 5). Average student loan debt: $29,400. Retrieved from http://money.cnn.com/2013/12/04/pf/college/student-loan-debt/

Appendix

```
##       caseid year age    sex educ         degree coninc           incom16
## 55088  55088 2012  22   Male   16       Bachelor 178712     Above Average
## 55089  55089 2012  21   Male   12    High School 178712     Above Average
## 55090  55090 2012  42   Male   12    High School  91920     Below Average
## 55091  55091 2012  49 Female   13    High School 107240 Far Above Average
## 55092  55092 2012  70 Female   16       Bachelor  42130     Below Average
## 55094  55094 2012  35 Female   15 Junior College  24895           Average
## 55095  55095 2012  24 Female   11 Lt High School   4213           Average
## 55096  55096 2012  28 Female    9 Lt High School    383           Average
## 55097  55097 2012  28 Female   17       Bachelor  24895 Far Below Average
## 55098  55098 2012  55   Male   10 Lt High School    383     Above Average
## 55099  55099 2012  36 Female   16       Bachelor  42130     Below Average
## 55100  55100 2012  28 Female   12    High School   6894           Average
## 55101  55101 2012  59 Female   12    High School  18193 Far Below Average
## 55103  55103 2012  35 Female   13    High School  42130           Average
## 55104  55104 2012  36   Male   12    High School  42130           Average
## 55105  55105 2012  47 Female   13    High School  34470     Above Average
## 55106  55106 2012  55   Male   12    High School  51705     Below Average
## 55107  55107 2012  18 Female   12    High School  18193     Below Average
## 55109  55109 2012  39   Male   10 Lt High School  34470     Below Average
## 55110  55110 2012  54   Male   14 Junior College  76600           Average
## 55111  55111 2012  45 Female   16 Junior College 107240     Below Average
## 55112  55112 2012  71   Male   12    High School  91920     Below Average
## 55114  55114 2012  22   Male   15    High School 178712     Above Average
## 55116  55116 2012  81 Female   16       Bachelor  34470     Below Average
## 55117  55117 2012  44 Female   13    High School   6894           Average
## 55118  55118 2012  78   Male   16       Bachelor  63195     Above Average
## 55119  55119 2012  63 Female   14 Junior College  42130           Average
## 55120  55120 2012  73   Male   19       Graduate 178712 Far Below Average
## 55121  55121 2012  40   Male   16       Bachelor  51705     Below Average
## 55122  55122 2012  42 Female   14    High School  51705     Below Average
## 55123  55123 2012  62   Male   18       Graduate  51705           Average
## 55124  55124 2012  52   Male   11    High School  76600           Average
## 55125  55125 2012  49 Female   12    High School  34470           Average
## 55126  55126 2012  27 Female   17       Bachelor  24895           Average
## 55127  55127 2012  30 Female   14    High School  34470           Average
## 55128  55128 2012  29 Female   18       Graduate  76600           Average
## 55129  55129 2012  69 Female   14    High School  91920     Below Average
## 55130  55130 2012  51 Female   18       Graduate 178712           Average
## 55131  55131 2012  57 Female   16       Bachelor 178712     Below Average
## 55132  55132 2012  44   Male   16       Bachelor 178712     Above Average
## 55133  55133 2012  73 Female   16       Bachelor 178712     Above Average
## 55134  55134 2012  73 Female   16       Bachelor 178712           Average
## 55135  55135 2012  68   Male   16       Graduate 178712 Far Above Average
## 55136  55136 2012  84 Female   16       Bachelor  51705           Average
## 55137  55137 2012  63   Male   19       Graduate 178712     Below Average
## 55138  55138 2012  57   Male   16       Bachelor  51705           Average
## 55139  55139 2012  42 Female   16       Bachelor 178712     Below Average
## 55140  55140 2012  45 Female   20       Graduate 178712     Below Average
## 55141  55141 2012  38   Male   14 Junior College 107240     Below Average
## 55142  55142 2012  46   Male   20       Graduate 178712     Above Average
## 55144  55144 2012  41 Female   14    High School  28725 Far Below Average
## 55145  55145 2012  75 Female   12    High School  21065     Below Average
## 55146  55146 2012  81 Female    8 Lt High School   6894     Below Average
## 55148  55148 2012  67 Female   19       Graduate  51705     Below Average
## 55149  55149 2012  71   Male   19       Graduate  21065 Far Below Average
## 55151  55151 2012  45 Female   17       Bachelor  34470           Average
## 55154  55154 2012  52 Female   11 Lt High School  24895           Average
## 55155  55155 2012  49 Female   16       Bachelor  51705     Below Average
## 55156  55156 2012  46   Male   16       Bachelor  63195           Average
## 55157  55157 2012  51 Female   14    High School  24895 Far Below Average
```