Predicting whether an individual can earn more than $100,000 annually with age and education level

Yuyu Fei, Biqi Jiang, Jiayi Yang, Yuwen Wu

October 19, 2020

Contents

Relevant files are stored in the following GitHub repository: https://github.com/felixfei1998/STA304-PS3

0.1 Abstract

We investigated if an individual's age and educational level significantly attribute to his/her income, particularly if he/she makes over $100,000 Canadian dollars in Canada. We analyzed survey data from the 2017 General Social Survey (GSS) on the Family and built a formal statistical model to answer this research question and our results were not surprising. We discovered that age and educational level are significant in predicting whether an individual makes more than $100,000 annually. We found out that the older the individual, the higher the chance of making over $100,000 and that those with a certificate from either an educational institution or a professional body have a higher chance of making over $100,000 per year than those only with a high school diploma or less. Our results are informative of the current socio-economic trends in Canada in most provinces. The next steps are to investigate whether there are other factors that affect whether an individual can make over $100,000 per year and account for them to generalize our results.

0.2 Introduction

The Sunshine list is a public sector disclosure in Ontario that reports individuals in the public sector that earns more than $100,000 a year. These individuals come from all kinds of sectors, fields and profession, and have different kinds of ethnic and socio-economic backgrounds. Many deem being able to make over $100,000 a year as a recognized success and many children and adolescents today aim to achieve this very goal. Inspired by this, we were interested in predicting whether an individual can make over $100,000 a year or not. Based on literature, age (Hedström and Ringen 1987) and education level (Houthakker 1959) are the most important factors that affect an individual's income.

We aimed to build a formal statistical model suitable for modeling such binary data using variables age and education level from the 2017 General Social Survey (GSS) on the Family, a nationwide survey that was designed to record socio-demographic variables of Canadians and capture socio-demographic and socio-economic trends.

The factors that contribute to whether an individual is able to earn more than $100,000 annually, which is considered high income and more than twice the national personal income average of Canada, are of importance to policy makers in the government, researchers in academia and the wider population and media.

0.3 Data

The dataset comes from the 2017 General Social Survey (GSS) on the Family. The 2017 GSS, coordinated from February 2, 2017 to November 30, 2017, is a sample survey of cross-sectional survey design. The target population includes all non-institutionalized persons over 15 years old, living in the ten chosen major provinces of Canada. The survey uses an updated frame, designed back in 2013, that encompasses telephone numbers with Statistics Canada's Address Register, and conducts data collection over telephone. Data are susceptible to both sampling and non-sampling errors.

The GSS program was established in 1985. It conducts telephone surveys across the ten provinces across Canada. The GSS is known for its ongoing collection of cross-sectional data that can be used for trend analysis, and its capability to validate and develop novel concepts that deal with current or surfacing issues. The two primary objectives of the General Social Survey are: 1) to collect data on societal trends in order to oversee nuances and changes in the living conditions and well-being of residents of Canada across time; and 2) to give information on particular societal policy problems of current or surfacing interest.

The crucial role family plays in people's lives cannot be disputed. Today's family, however, must steer through changing marital, family, and professional trajectories. While our understanding of families in Canada has improved considerably over the past few years, the future of families remains a topic of great interest. As we see that families are getting more diverse. The GSS on families will inform researchers on the types and characteristics of families in Canada to deepen our understanding of families.

The survey collected a tremendous amount of data for each sampled respondent and also relevant information about each family member of the respondent's household.

For sampling purposes, each of the ten provinces is divided into different strata (ie geographic regions). Many Census Metropolitan Areas 1 (CMA) are considered separate strata. This is the case for Saint John, Halifax, Saint John, Montreal, Quebec City, Toronto, Ottawa, Hamilton, Winnipeg, Regina, Saskatoon, Calgary, Edmonton and Vancouver. Though all CMAs in this list are located in Quebec, Ontario and British Columbia. Moncton. By grouping the remaining CMA (except Moncton), three other levels are formed Quebec, Ontario and British Columbia. Finally, the non-CMA areas in each of the ten provinces are also grouped to form another ten levels, for a total of 27 levels. Moncton has been added to the non-CMA layer New Brunswick. Stratified random sampling was then used, specifically selecting a simple random sample from each of the constructed stratum.

The survey frame was created using two different components: 1) various lists of telephone numbers in use (both mobile and landline) available to Statistics Canada from various sources (telephone companies, Census of population, etc.) and 2) the Address Register (AR) which is a list of all dwellings within the ten provinces.

The response rate of the 2017 GSS was 52.4%, which is sufficient for achieving sample representativeness.

0.4 Model

The response variable is the main outcome of interest denoted as Y_i: whether an individual makes over \$100,000 per year. The two main predictors of interest are age and educaiton level. If an individual makes over \$100,000 per year, then $Y_i = 1$, otherwise $Y_i = 0$. We model Y_i with a Bernoulli distribution such that $Y_i \sim Bernoulli(p_i)$, where $p_i = Pr(Y_i = 1)$.

The statistical model we used is the logistic regression model. It is a class of generalized linear models in which a function of the mean response is modeled as a function of the predictors, where the coefficients are linear. Logistic regression model in particular models the log odds of the mean response. The odds of the mean response is the $\frac{p_i}{1-p_i}$.

A one unit increase in the value of predictor variable X_i leads to a β_i increase in $\log \frac{p}{1-p}$, keeping all other predictor variables fixed at a specific value.

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 X_{age,i} + \beta_2 X_{education-certificate,i} + \beta_3 X_{education-university,i}$$

0.5 Results

0.6 Exploratory data analysis

In this section we visually assess the main outcome of interest as well as the predictor variables of interest and their relationships.

We see that in Figure 1 our sample approximately 8% made more than $100,000 a year.

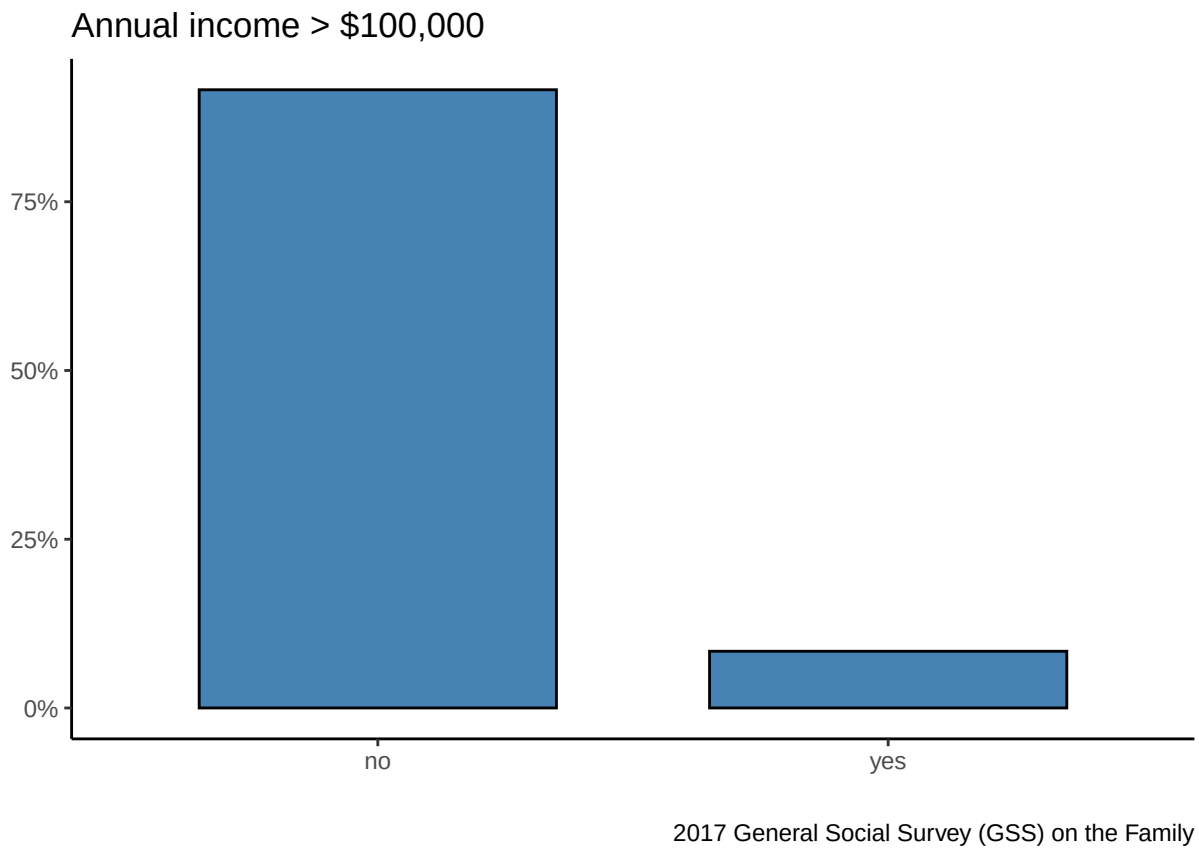

Figure 1: Proportion of sampled individuals who make more than $100,000 a year in the 2017 CSS survey

We see that in Figure 2 our sample close to 40% only had a high school diploma or without one. A little over 40% had some sort of certificate. And only approximately 19% had a university degree.

The distribution of age is slightly left skewed and the median is about 55 (Figure 3).

We see that in Figure 4 education does not really have an association with whether an individual can make over $100,000 a year or not.

We see that in Figure 5 the age distribution for those who make more than $100,000 a year is less spread out.

0.7 Model results

Table 1 contains the logistic regression model output. We see that the P-values for age and education level are < 0.05 = level of significant. Age and education level are statistically significant in predicting whether an individual makes over $100,000 per year.

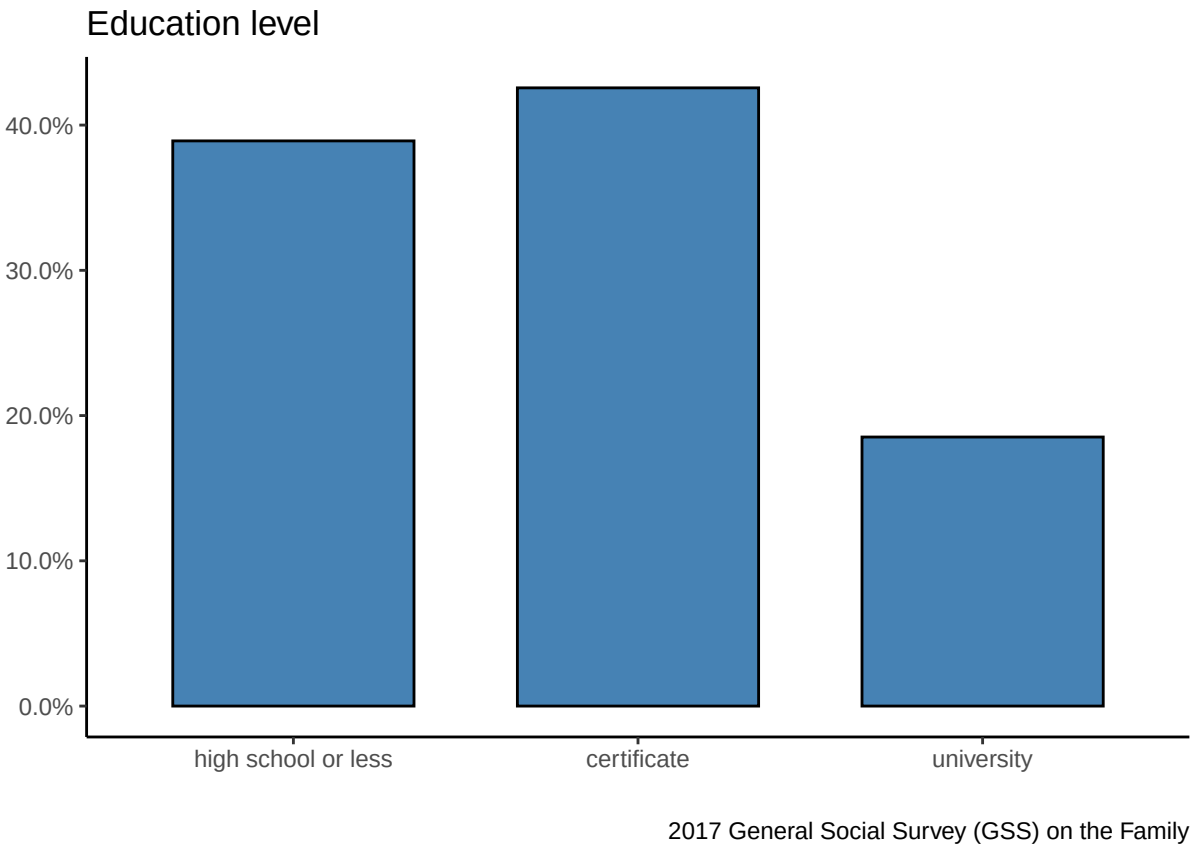

Figure 2: Breakdown of education level of sampled individuals in the 2017 CSS survey

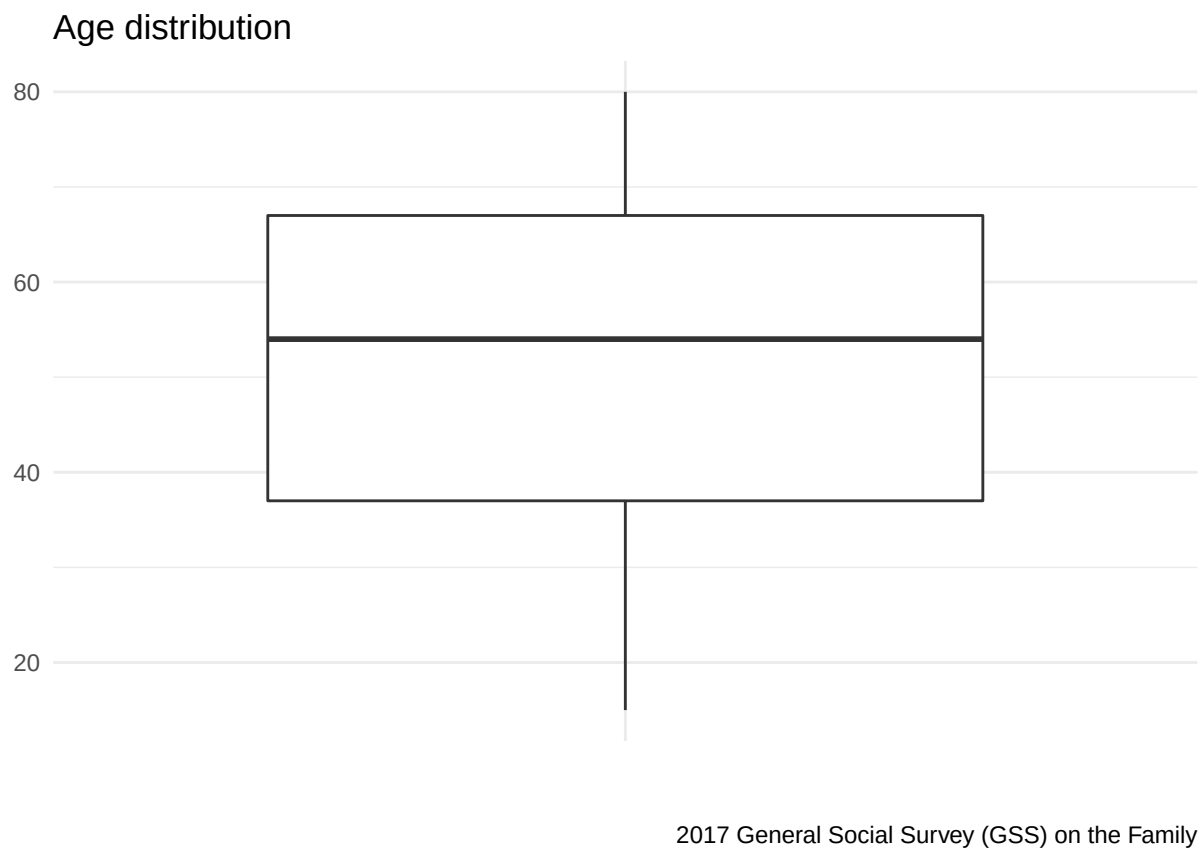

Figure 3: Distribution of age of sampled individuals in the 2017 CSS survey

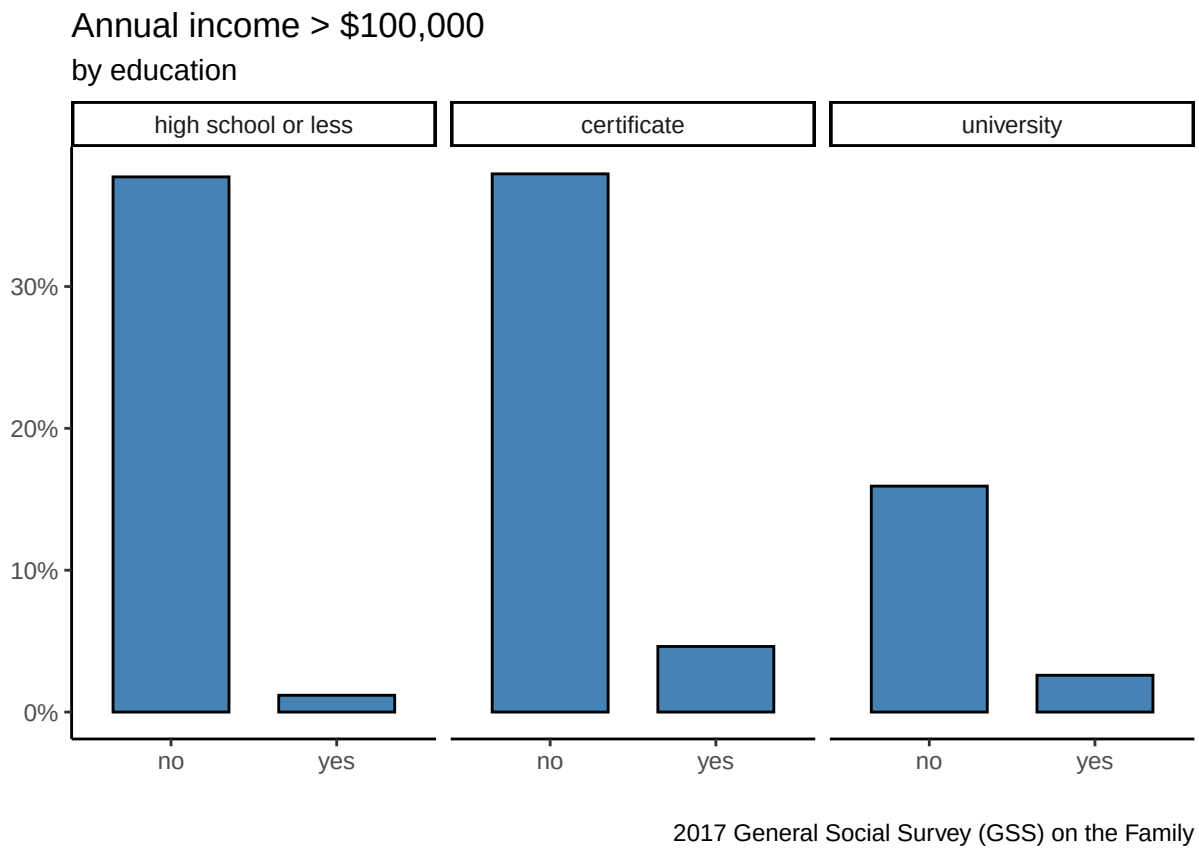

Figure 4: Proportion of sampled individuals who make more than $100,000 a year by in the 2017 CSS survey

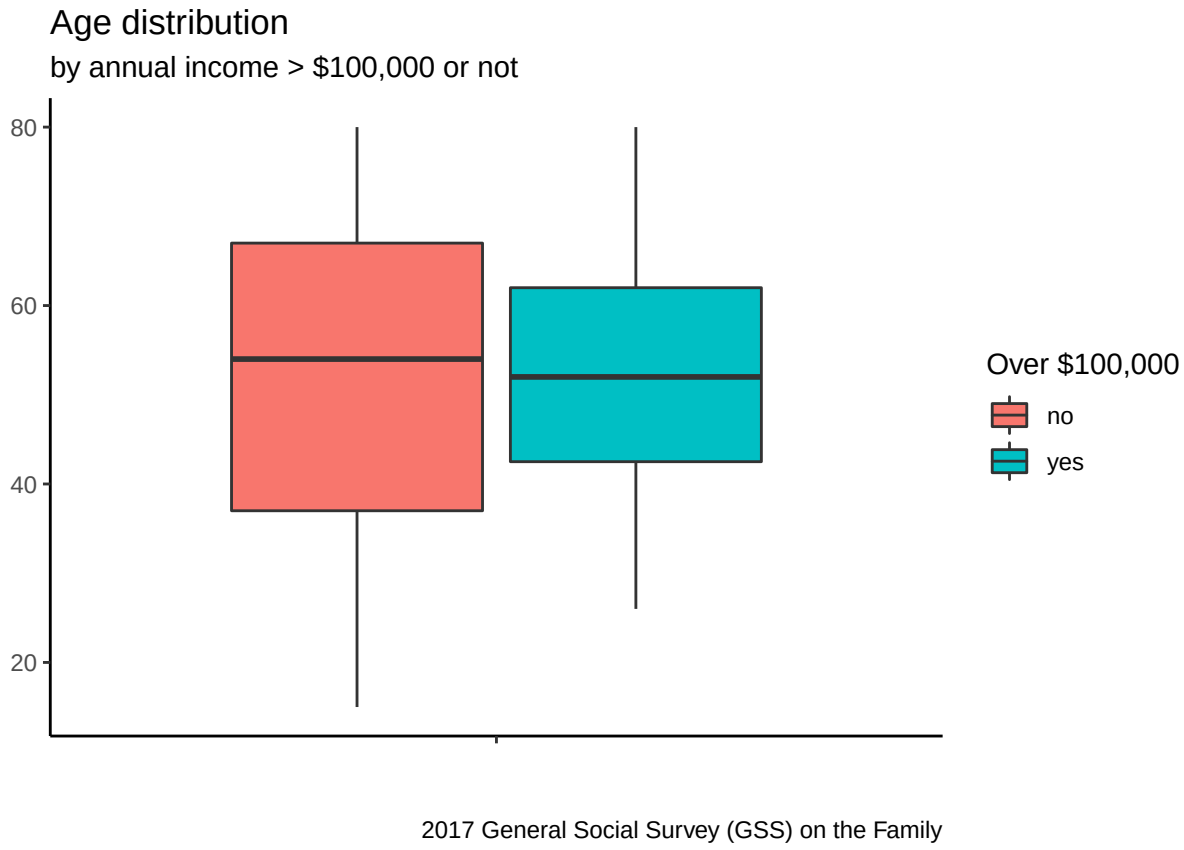

Figure 5: Distribution of age for sampled individuals who make more than vs. less than $100,000 a year in the 2017 CSS survey

Table 1: Logistic Regression Model Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.778	0.109	-34.527	0
educertificate	1.374	0.074	18.480	0
eduuniversity	1.682	0.081	20.677	0
age	0.006	0.002	3.685	0

Table 2 contains the estimated odds ratio with the 95% confidence interval from the logistic regression model. We see that individuals with either some sort of certificate or a university degree are more likely to make over $100,000 per year compared to individuals whose highest education degree obtained is a high school diploma.

Table 2: Estimated Odds Ratio [95% CI]

	OR	2.5 %	97.5 %
(Intercept)	0.023	0.018	0.028
educertificate	3.951	3.421	4.579
eduuniversity	5.379	4.592	6.318
age	1.006	1.003	1.009

0.8 Discussion

We analyzed data from the 2017 General Social Survey (GSS) on the Family. We were interested in finding out whether an individual's age and education level would affect his/her ability to earn over $100,000 per year. Our results were not surprising but could inform future related studies. The odds of earning over $100,000 increases by 1.006 times when you get one year older while keeping education level fixed. The odds of earning over $100,000 for those with a certificate is almost 4 times higher compared to those who only have a high school diploma or less than that. The odds of earning over $100,000 is over 5 times higher for those with a university degree compared to those who only have a high school diploma or less than that.

Weaknesses

One major weakness in our study is the lack of ability to generalize the results. We did not consider a lot of factors that could affect whether an individual can make over $100,000 per year. In addition, the sample design only included 10 major southern provinces of Canada and excluded Yukon, Northwest Territories and Nunavut. Residents in these regions are inherently different than the 10 included provinces in regards to age distribution, education level, profession available and culture. These could all affect our outcome of interest and personal income overall. Lastly, few observations with missing values were removed from our analysis sample. These could have caused a very slight bias in our analysis results but were unlikely to change the analysis results or conclusion.

Next Steps

Although age and education level have been proven in numerous literature to be predictive of an individual's income level and that our analysis results have shown that age and education level are predictive of whether an individual is able to earn more than $100,000 per year, there are other factors that could affect this outcome of interest. Factors like gender, ethnicity, language ability, citizenship status and many more could all be predictive of personal income. The next steps include building a predictive model that accounts for these factors which could potentially be confounding. In addition, we should explore methods of missing value imputation to impute missing values that may have arisen from non-response or erroneous data entry. Deleting records because of missing values is not cost-effective and can lead to bias results. Should data not be available from the 2017 CSS survey, we could try to find other national surveys that capture the information we would require to conduct further analysis.

References

Hedström, Peter, and Stein Ringen. 1987. "Age and Income in Contemporary Society: A Research Note." *Journal of Social Policy* 16 (2): 227–39.

Houthakker, Hendrik S. 1959. "Education and Income." *The Review of Economics and Statistics* 41 (1): 24–28.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Muller. 2020. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Dana Seidel. 2020. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.