Does education level, gender and being a minority affect an individual's income level?

Smriti Kannan

October 19th 2020

Does education level, gender and being a minority affect an individual's income level?

Smriti Kannan

October 19h 2020

Abstract

There are numerous factors that affect an individual's income. There are factors that can be controlled by the individual such as education, years of experience and the type of job. There are also other factors that cannot be controlled such as the current state of the economy. In this report, we will specifically look at how and if education level, gender and being a minority affect an individual's income. The data used for this analysis is from the most recent General Social Survey on Family (2017) that took place in Canada. A logistic regression model is used to determine if education levels, gender and being a minority affect a person's income being $75,000 or greater. From the model and some calculations from the model, we find out the levels of education seem to have the largest effect in whether or not a person will have an income greater than $75,000. Gender also seems to affect the probability of having an income greater than $75,000, but for minority we cannot make a claim as confidently.

Introduction

Various factors influence an individual's income level. Some factors can be controlled to some extent by individuals - we can choose the type of job we apply for, choose how much education we receive. However, there are factors that an individual cannot really control such as their sex and whether or not they are a minority. In this report, we are going to investigate if these factors, ones we can control and ones we cannot affect an individuals income. The median income of individuals was $61,400 in 2018, virtually unchanged from 2017 (Government of Canada, Statistics Canada). The data we are using is from 2017 and we are going to investigate how certain factors affect an individual's chance of having an income of $75,000 or greater which is a bit higher than the median income in Canada in 2018. The reason for looking at income greater than $75,000 is mainly because in the data set, we are not given the actual income but are given which income group the individual falls under.

One of the factors we are going to look at is sex and this is quite interesting because there has always been a gender pay gap. The gender pay gap has narrowed since the 1980s but has been relatively stable in the past 15 years or so (Graf et al.). One of the goals of this paper is to determine how much more likely it is for men, with the same education level as women to receive an income greater than $75,000. Another goal is to analyze whether being a visible minority affects the probability of an individual to have a higher income.

The final factor which might be more obvious to having an effect to an individual's income is education level. As an undergraduate student myself, I find it quite interesting to look at how much more likely it is to get a higher paying job with a post-graduate degree than an undergraduate degree. Does it really make a significant difference by having a post-graduate degree than just an undergraduate degree to have a higher income? What is an individual doesn't go to college or university? These are some things we are going to try and investigate in this report.

Data

The data used in the analysis is from the CHASS website and in particular it is the general social survey on family, 2017. The target population in our study is all Canadians aged 15 and over in the ten provinces of Canada (Government of Canada, Statistics Canada). The sampling frame for the survey is the Statistics Canada's common telephone frame, which combines land line and cellular telephone numbers from the Address Register (Government of Canada, Statistics Canada).

An advantage of this data is that it is the most recent general social survey (GSS) so the data we are using is relatively new. Another advantage of this data set is that it has numerous variables (81) and is also a large data set containing 20,602 data points. The data set contains various categorical variables such as, education-level, marital status, self rated mental health etc. This allows us to perform various types of analysis and gives us a lot of options for building a model.

A potential drawback to the data is that is has fewer quantitative variables. For example, with each individual's income, we are only given which income bracket they fall under but not the actual income value itself. If we had data such as actual income values, we might may be able to perform linear regression analysis. This is because in linear regression, the response variable must be numerical. Another drawback of the data is that there are quite a lot of missing values for some of the variables. Luckily the variables being focused on for this report do not have a lot of missing values. But if we wanted to further investigate something with different variables, it could cause some accuracy in results to be lost.

After obtaining the data set from the CHASS website, the data set was cleaned and variable names were changed to make it easier to understand. Some variables were also changed to make the data more meaningful. For example, the original data set for gender had values of 1 or 2 indicating male or female. After modifying the data set, now gender takes the value of either male or female. After cleaning the data, we cleaned it to only include specific columns that were of interest to the study. The variables chosen are vis_minority (indicating if a person is a visible minority or not), education (indicating the furthest education a person has received), income_respondent (which tells us the person's income group) and gender. After this, any data point with a missing value was also removed. Data points with 'Don't know' for the vis_minority variable were also removed. A new column was created to indicate whether or not the individual's income was greater than or equal to $75,000. This will be used later when we are building the model.

Model

The model being used to analyze the data is a logistic regression model. Logistic regression is a type of statistical model that uses a logistic function to model a binary response variable. The main thing that needs to be checked before using this model is that the response variable is binary. The response variable that is going to be used in this model is whether or not an individual's income is above $75,000. This can only have two values, yes or no and therefore it is appropriate to use a logistic regression model. The equation for our logistic model is:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0+\beta_1 X_{vis\_minority}+\beta_2 X_{college}+\beta_3 X_{highschool}+\beta_4 X_{>highschool}+\beta_5 X_{trade}+\beta_6 X_{uni}+\beta_7 X_{unimaster}+\beta_8 X_{male}$$

where:

- vis_minority = Visible minority

- college = College, CEGEP or other non-university certificate or diploma
- highschool = High school diploma or high school equivalency certificate
- >highschool = Less than high school diploma or its equivalent
- trade = Trade certificate or diploma
- uni = University certificate or diploma below the bachelor's level
- unimaster = University certificate, diploma or degree above the bachelor's level
- male = Male

and

$$\hat{p}$$

is the probability of having an income of 75,000 or greater.

Using the R function glm, a simple logistic regression model is created. The values from the model can be found in figure 5 in the results section of the report.

Results

Before we look at the model results, let's take a closer look at the data and plot some bar graphs with our variables of interest. First looking at if the gender of a person affects their income.

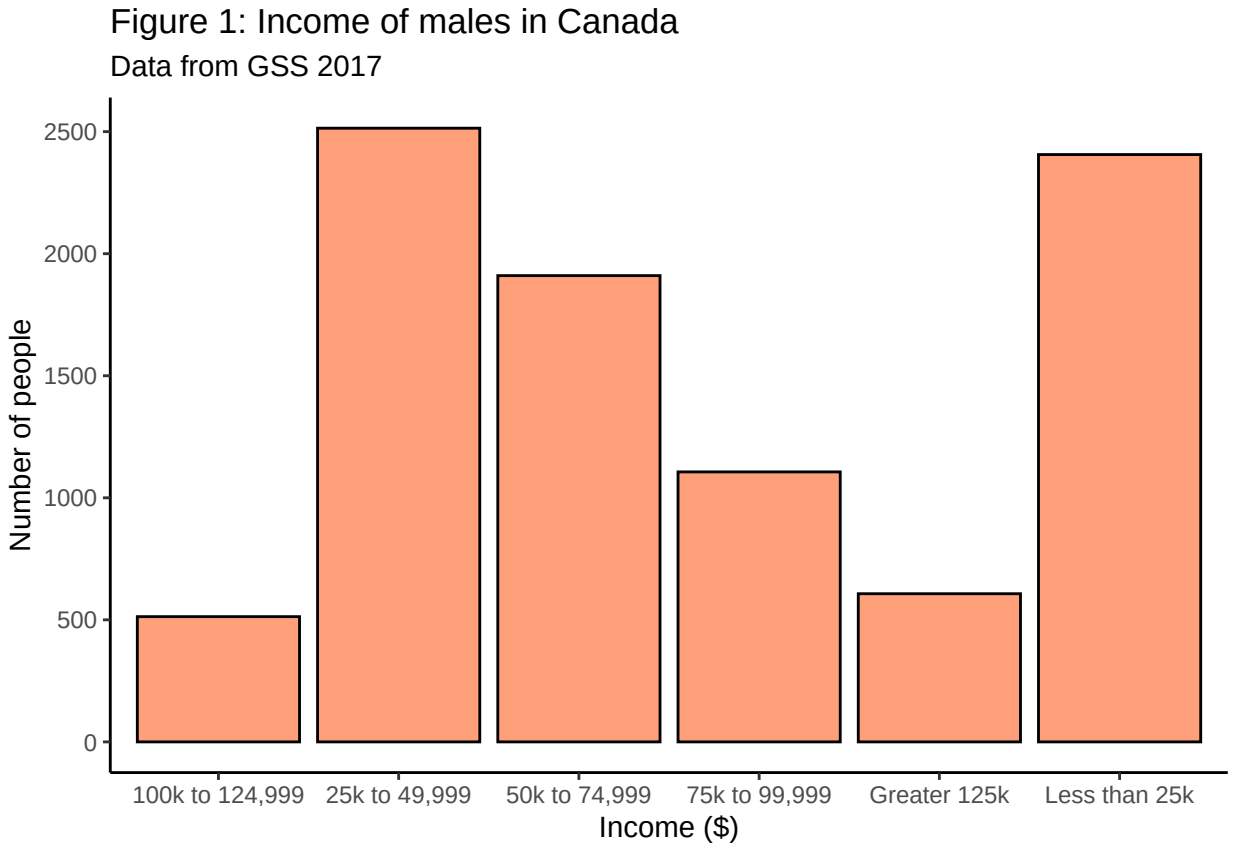
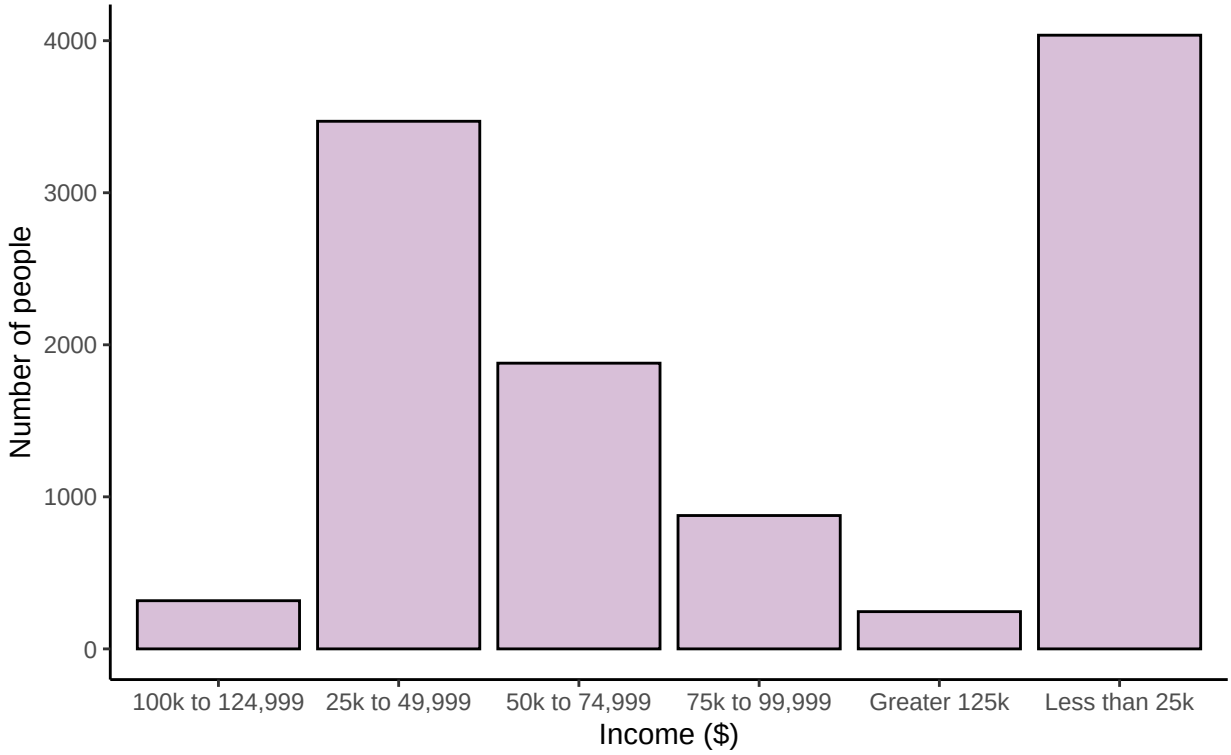


Figure 2: Income of females in Canada

Data from GSS 2017



Next looking at another variable of interest, whether or not being a visible minority affects income.

Figure 3: Income of non−minorities in Canada

Data from GSS 2017

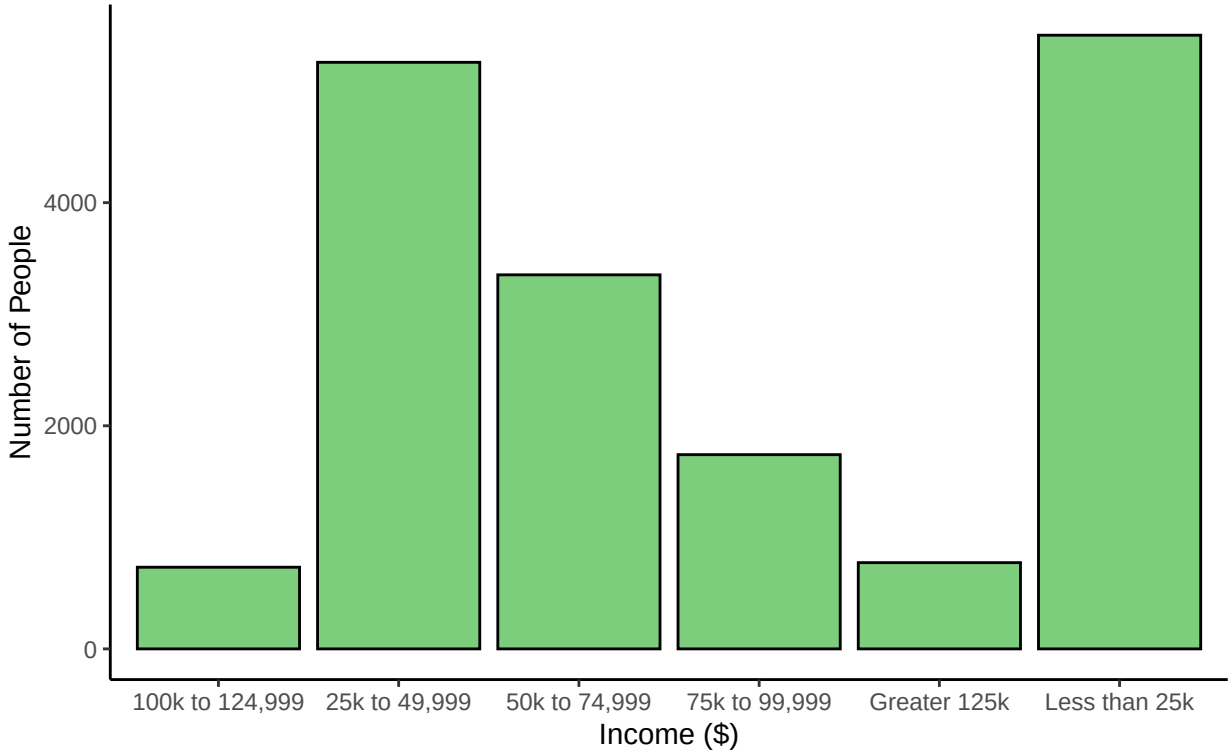


Figure 4: Income of visible minorities in Canada

Data from GSS 2017

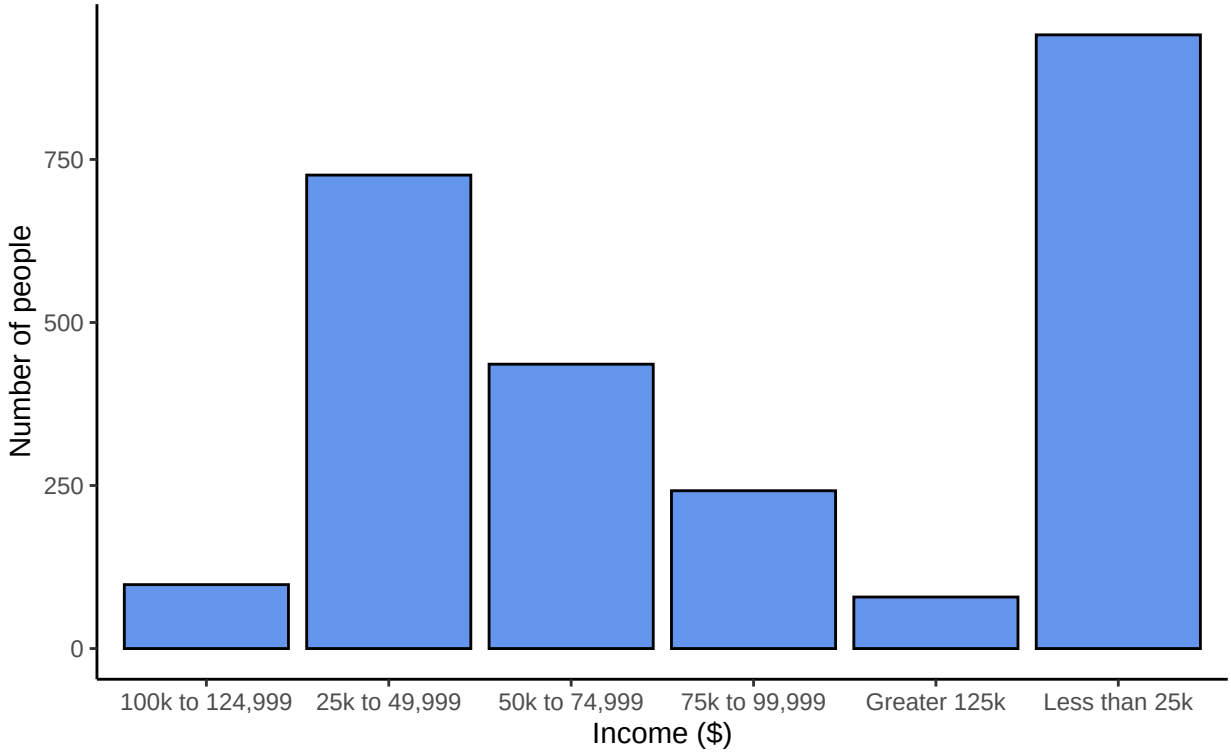


Figure 5: Coefficients from the logistic regression model

Variable names	Estimate	Std. Error.	z value	Pr>|z|
(Intercept)	-1.08268	0.04281	-25.292	< 2e-16
Visible minority	-0.58026	0.06140	-9.450	< 2e-16
College,CEGEP or other non-university certificate or diploma	-0.96866	0.05587	-17.339	< 2e-16
High school diploma or high school equivalency certificate	-1.56127	0.06130	-25.469	< 2e-16
Less than high school diploma or its equivalent	-2.72570	0.10709	-25.453	< 2e-16
Trade certificate or diploma	-1.05621	0.07894	-13.380	< 2e-16
University certificate or diploma below the bachelor's level	-0.39772	0.09628	-4.131	3.62e-05
University certificate, diploma or degree above the bachelor's level	0.56944	0.06131	9.288	< 2e-16
Male	0.89146	0.04029	22.128	< 2e-16

Figure 6: Probability calculations using the model and model equation

Value of variables selected	Probability of income being over $75k
Male, University certificate or diploma below the bachelor's level, not a minority	0.3569
Male, University certificate or diploma below the bachelor's level, visible minority	0.2370
Female, University certificate or diploma below the bachelor's level, not a minority	0.1854
Male, University certificate or diploma above the bachelor's level, not a minority	0.5934
Male, High school diploma or high school equivalency certificate, not a minority	0.1477
Male, Less than high school diploma or its equivalent, not a minority	0.0513

Discussion

The first factor we are going to discuss is whether the sex of an individual affects their chances of having an income greater than or equal to $75,000. Figures 1 and 2 are bar graphs that show us the number of people within each income group for each sex. First looking at the lowest income brackets, we can see that in females, more individual's have incomes less than $25,000 than individuals who have an income between $25,000 and $49,000. The opposite is true for the males. More males have an income between $25,000 and $49,000 than incomes less than $25,000, when we are looking the two lowest brackets of income. While looking at the upper two income brackets ($100,000 to $124,999 and $125,000 or more), it is very clear that the proportion of males in these two income levels is much larger than the females. We can see this by just looking the Figure 1 and Figure 2 and noticing the heights of the bars for these income levels. This suggests that as we make calculations using our model, we should find some significant differences. This is true as we can see from Figure 5 that the estimate of the slope for males is 0.89146 when compared to females. This means that if we keep everything else constant or the same, we would expect the log odds of having an income greater than $75,000 for males to increase by 0.89146 when compared to females. To put this in probabilities we understand a bit better, from Figure 6 we can see that the probability of a male having an income over

$75,000 is 0.3569 and the probability for females is only 0.1854 (when we take education for both to be a university certificate or diploma below the bachelor's level and both are also not minorities). The probability of men having an income over $75,000 is 92.5% greater than the probability of a women having an income over $75,000. This is very high and it just reconfirms our thinking in the beginning that there still is a gender gap in incomes.

The second factor we looked at was whether being a visible minority or not affected the probability of that individual to have an income greater than $75,000. Figures 3 and 4 are bar graphs that show us the number of people withing each income group for non-minorities and visible minorities. At a firt glance, the bar graphs look pretty similar and we cannot see an obvious difference. If we look a bit closer, it appears that the proportion of people in the top two income brackets ($100,000 to $124,999 and $125,000 or more) is a bit higher in non-minorities than in visible minorities. However, this difference is much smaller than it was for males and females and therfore we need to look at our model and calculations before we can make any conclusions. From Figure 5, we can see that the slope estimate for visible minorities is -0.58026. Therefore, we can predict that the log odds of a visible minority person having an income greater than $75,000 decrease by 0.58026 when compared to a non-visible minority individual. In Figure 6, we can see if there is a significant difference in the probabilities. The probability of having an income over $75,000 for a non-visible minority individual is 0.3569. The same probability for an individual who is a visible minority is 0.2370 (in these calculations the other variables are kept the same - male and university certificate or diploma below the bachelor's level). The difference in probability is 0.1199 which is not too large but is still not small enough or large enough for us to make a definite conclusion.

The last and final factor we looked at was education level and its impact on income levels. The data set contained a lot of different values for the education variable so I decided to focus only on a few of them. The values focused on are university certificate or diploma above and below the bachelor level, high school diploma or high school equivalence certificate and less than high school diploma or its equivalent. From figure 6 we can see that the probabilities of having an income over $75,000 vary quite a lot based on different levels of education. The probability of having an income over $75,000 is 0.3569 and 0.5934 for university certificate or diploma below and above the bachelor level respectively. There is quite a large difference between these probabilities and difference in probabilities is 0.2365. If we look at the two extremes, the difference in probability for an education level of less than a high school diploma and education level of university certificate above the bachelor's level is 0.5421. This is the largest difference in probability we have come across while looking at our factors. Therefore we can be quite positive that education level makes and impact on how likely an individual will have an income greater than $75,000.

Weaknesses

One of the weaknesses of the study is that the data is a survey and therefore there might be a bias as our sample is a non-probability sampling technique is used. Each person in the target population (which is all Canadian aged 15 or over) does not have an equal probability of being selected. This could lead to certain people being under-represented or not represented at all. The general social survey data is collected using a combination of self-completed online questionnaires and telephone interviews since 2013 (Government of Canada, Statistics Canada). This might lead to some inaccuracy in the data collected as people fill these out themselves and they might not be completely honest. However, Statistics Canada began asking respondents for permission to link their survey information to additional data sources such as personal tax records (Government of Canada, Statistics Canada) which allows for data such as income to be much more accurate.

Another weakness of our study is that we are not looking at the type of job and just assuming that the types of jobs are approximately equally split between genders. However there in our dataset, there could be a higher proportion of men with higher paying type jobs than women which would cause some bias in our results. To improve our analysis, we could break down the data further and sort it into job categories and then analyze the difference in income between men and women in each category. This could also lead to some further findings such as do certain types of jobs have a larger difference in income between men and women.

Next Steps

Some helpful next steps would be to in data collection receive numerical data for incomes of each individual instead of just the income range. This will allow for more quantitative analysis such as constructing a liner regression model. Another step that can increase the accuracy of the finding is to split up the data based on job types. By analyzing the data in each job group, we can look for similar trends within groups and also look at which group has the largest differences. To further understand if these differences in incomes have changed, we should look at data from previous General Social Surveys and see if there are any trends or if the findings are similar to what we found with the 2017 data set.

References

Alexander, Rohan and Caetano, Sam. "gss_cleaning.R". Accessed 15 Oct. 2020

"CHASS Data Centre." Utoronto.Ca, 2015, datacentre.chass.utoronto.ca/. Accessed 15 Oct. 2020.

Rstudio. Graphical Primitives Data Visualization with Ggplot2 Geoms -Use a Geom to Represent Data Points, Use the Geom's Aesthetic Properties to Represent Variables. Each Function Returns a Layer. One Variable Geoms -Use a Geom to Represent Data Points, Use the Geom's Aesthetic Properties to Represent Variables. Accessed 17 Oct. 2020.

"Ggplot2 Quick Reference: Colour (and Fill) | Software and Programmer Efficiency Research Group." Inf.Usi.Ch, 2020, sape.inf.usi.ch/quick-reference/ggplot2/colour. Accessed 17 Oct. 2020.

Graf, Nikki, et al. "The Narrowing, but Persistent, Gender Gap in Pay." Pew Research Center, Pew Research Center, 22 Mar. 2019, www.pewresearch.org/fact-tank/2019/03/22/gender-pay-gap-facts/. Accessed 18 Oct. 2020.

Government of Canada, Statistics Canada. "The Daily — Canadian Income Survey, 2018." Statcan.Gc.Ca, 2018, www150.statcan.gc.ca/n1/daily-quotidien/200224/dq200224a-eng.htm. Accessed 18 Oct. 2020.

"General Social Survey (GSS): An Overview." Statcan.Gc.Ca, Government of Canada, Statistics Canada, 20 Feb. 2019, www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm. Accessed 18 Oct. 2020.