Relationship between Income, and Gender and Level of Education in Canada

Aditi Khetwal, Ananya Jha, Sukhmani Khaira, Medha Srivastava

19th October 2020

Abstract

Financial differences between communities have long been the centre of research and studies, and the dependence of income levels on social factors has led to further disparities between people. This paper tries to examine the likelihood of having high income (CAD 100K or more) on grounds of a person's gender and level of education. We find that men are significantly more likely to fall in the high income category as compared to women, and that there is a positive correlation between having a higher level of education and earning more. As people go from less than a high school diploma, to some university diploma, to degree above bachelors, we see that their prospects of earning CAD 100K or more also go up.

Introduction

Social and demographic information about families in Canada is largely varied due to the increasing diversity in the country, and can be indicative of significant social trends present. Income is a social factor that varies substantially within populations [1], and is known to be influenced by a multitude of factors such as sex or education level [2,3]. As a factor shown to greatly contribute to quality of life[4], it is a common variable of interest in studying patterns within populations and relationships between social factors. Previous research conducted on the relationship between gender and income has shown notably lower average and median incomes for employed women than employed men in Canada, suggesting gender could be a crucial indicator for individual income levels [2]. Additionally, education levels have also been shown to positively correlate with income levels, leading to the prediction that higher education leads to higher incomes [3]. This report investigates the relationship between sex, education, and income of Canadians. Specifically, both sex and education are explored as predictors for high or low income levels, to predict the likelihood of individuals having high incomes (> CAD 100,000) given their sex and education level. These relationships have crucial applications for national policies and addressing prevalent social issues. Disparities between sexes and/or education levels with regards to income could instigate policies that decrease gender inequality in wages or that increase education accessibility. The dataset analyzed was the Canadian 2017 General Social Survey (GSS), which studies social trends within Canada annually and focused on families.

To explore the goal of predicting income level with sex and education variables, a logistic regression is performed predicting income as a binary response variable (less than CAD 100,00 being the lower income class and greater than $100,000 being the higher income class) based on sex and education level as predictor variables. The income responses in the GSS dataset were provided as categorical values of ranges of income, such as "CAD 50,000 to CAD 74,999", instead of continuous numerical data. The lack of continuous data prevented a well-fitting linear model, and supported the selection of a logistic regression model to analyze the relationship between the data. Additionally, the aim of the report was to investigate the likelihood of an individual having a high income or not given base characteristics of sex and education, so a binary response variable was fitting as specific incomes were not as necessary as the classification of high income or not. In addition to the regression model, the report discusses and visualizes relationships between each of the predictor variables and an ANOVA table is made.

Data

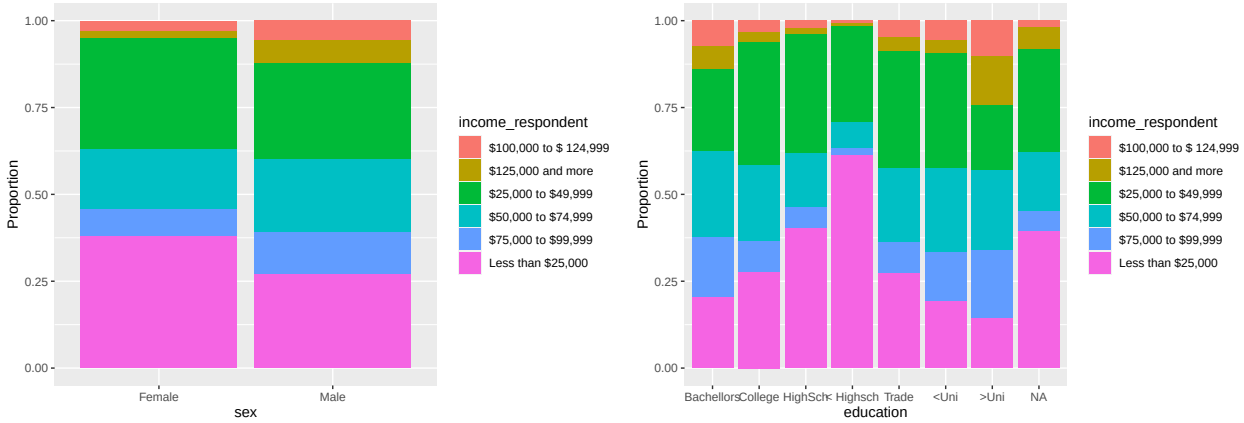
The report aimed to analyze recent sex, education levels, and income for all individuals Canada, so 2017 GSS was chosen as the most recent and representative dataset of the entire population of Canada, obtained through the CHASS data centre by the University of Toronto.The GSS collects data on various social factors within Canada annually and the GSS on Family is repeated approximately every 5 years. The survey's target population was Canadians in all 10 provinces that are 15 years or older, using a survey frame of telephone numbers provided by Statistics Canada's Address Register [5]. The study sample was then all the telephone numbers attached to households sorted to have one number per household, reaching a sample size of 20,602 participants. The data primarily contains categorical variables and some continuous numerical data on social factors in Canadian families. The categorical nature of most of the variables proved to be a weakness of the dataset, as visualization and analysis of the data was less accurate without numerical data. The variables focused on were the sex of the respondent and the education level of the respondent, as well as the income level. There were 2 variables regarding income, family income and individual income of the respondent, but the individual income was chosen to be the focus since the goal of this report is to determine how likely an individual is to earn a high income given sex and education levels.

The presence of sampling and non-sampling errors in the survey impact the strength of the dataset. A major non-sampling error present in the survey data was non-responses in the data. To minimize the effect of the non-responses to the survey, the weight of those responses was redistributed within other households in the strata that did respond. However, partial non-responses, where only certain questions were not answered, were not redistributed and were counted as valid responses with some missing values [5]. This is a major weakness of the data, as the analyses were performed by removing missing values for the variables of interest (sex and education level) and therefore decreasing the sample size by 341 observations. A potential minor error is that some telephone numbers belonged to the same household, and although these were attempted to be grouped together and duplications were prevented through only contacting one number in the group, it is possible some numbers were incorrectly grouped together and actually contacted different households, or numbers considered different did contact the same household. Additionally, households that did not have any telephone number associated with them were not considered part of the sample or the population, therefore limiting a part of the target population of all 15 years or older individuals in Canada. Another weakness of the data set is the fact that income information was not obtained through questions in the interview as the rest of the data was, and was instead taken from both fiscal files and the GSS [5]. This allows secondary errors such as delayed tax filing or unreported incomes to impact the data, which is highly relevant to this report as the income variable was the basis of the conclusions [6]. Since the survey data was obtained through a telephone questionnaire, the quality of the questions also impacts the results. The majority of the questions only allowed for limited answers that fit certain categories, and no free-written answers. This is beneficial for the data analysis requirements, but limits the answers obtained to fit in certain categories.

The raw data of proportion of males and females per income group, and of the proportion of respondents in different education levels per income group, shows some relationships between the predictive factors and income. However, the presence of missing values made the raw data plots not very informative, so the data was cleaned for a proper analysis and visualization of the data.

Plotting raw data

Proportion of sex against the income of respondent and *Proprtion of type of education and the income of respondent*



Model

```
##
## Call:
## glm(formula = as.factor(income_logistic) ~ as.factor(sex) + as.factor(education),
##     family = "binomial", data = data_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9096  -0.4664  -0.2988  -0.2129   3.0962
##
## Coefficients:
##                                                                          Estimate
## (Intercept)                                                             -2.35169
## as.factor(sex)Male                                                       1.01460
## as.factor(education)College, CEGEP or other non-university certificate or di...  -0.82655
## as.factor(education)High school diploma or a high school equivalency certificate -1.42365
## as.factor(education)Less than high school diploma or its equivalent     -2.43329
## as.factor(education)Trade certificate or diploma                        -0.73477
## as.factor(education)University certificate or diploma below the bachelor's level -0.43150
## as.factor(education)University certificate, diploma or degree above the bach...  0.66840
##                                                                         Std. Error
## (Intercept)                                                                0.05882
## as.factor(sex)Male                                                         0.05546
## as.factor(education)College, CEGEP or other non-university certificate or di...    0.07777
## as.factor(education)High school diploma or a high school equivalency certificate   0.08784
## as.factor(education)Less than high school diploma or its equivalent        0.15631
## as.factor(education)Trade certificate or diploma                           0.10408
## as.factor(education)University certificate or diploma below the bachelor's level   0.13730
## as.factor(education)University certificate, diploma or degree above the bach...    0.07323
##                                                                         z value
## (Intercept)                                                             -39.982
## as.factor(sex)Male                                                       18.296
## as.factor(education)College, CEGEP or other non-university certificate or di...  -10.628
## as.factor(education)High school diploma or a high school equivalency certificate -16.207
## as.factor(education)Less than high school diploma or its equivalent     -15.567
## as.factor(education)Trade certificate or diploma                         -7.060
## as.factor(education)University certificate or diploma below the bachelor's level  -3.143
## as.factor(education)University certificate, diploma or degree above the bach...   9.127
##                                                                         Pr(>|z|)
## (Intercept)                                                              < 2e-16
## as.factor(sex)Male                                                       < 2e-16
## as.factor(education)College, CEGEP or other non-university certificate or di...   < 2e-16
## as.factor(education)High school diploma or a high school equivalency certificate  < 2e-16
## as.factor(education)Less than high school diploma or its equivalent      < 2e-16
## as.factor(education)Trade certificate or diploma                        1.67e-12
## as.factor(education)University certificate or diploma below the bachelor's level  0.00167
## as.factor(education)University certificate, diploma or degree above the bach...   < 2e-16
##                                                                            ***
## as.factor(sex)Male                                                         ***
## as.factor(education)College, CEGEP or other non-university certificate or di...  ***
## as.factor(education)High school diploma or a high school equivalency certificate ***
## as.factor(education)Less than high school diploma or its equivalent        ***
## as.factor(education)Trade certificate or diploma                           ***
## as.factor(education)University certificate or diploma below the bachelor's level **
## as.factor(education)University certificate, diploma or degree above the bach...  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11693  on 20260  degrees of freedom
## Residual deviance: 10331  on 20253  degrees of freedom
## AIC: 10347
##
## Number of Fisher Scoring iterations: 6
```

Regression Equation:

$$\widehat{Inc_{>100k}} = -2.3516 + 1.0146 Sex_M - 0.8266 Ed_{Col} - 1.4236 Ed_{HS} - 2.4333 Ed_{<HS} - 0.7348 Ed_{Tra} - 0.4315 Ed_{<Uni} + 0.6684 Ed_{>Uni}$$

$Inc_{>100k}$: Income logistic i.e income of the respondent being more than or equal to CAD 100,000.

Sex_M : Identifies the sex of the respondent. 'M' stands for male and the alternative is female.

Ed_{Col} : Level of education being "College, CEGEP or other non-university certificate"

Ed_{HS} : Level of education being "High school diploma or a high school equivalency certificate"

$Ed_{<HS}$: Level of education being "Less than high school diploma or its equivalent"

Ed_{Tra} : Level of education being "Trade certificate or diploma"

$Ed_{<Uni}$: Level of education being "University certificate or diploma below the bachelor's level"

$Ed_{>Uni}$: Level of education being "University certificate, diploma or degree above the bachelors"

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: as.factor(income_logistic)
##
## Terms added sequentially (first to last)
##
##
##                      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                 20260      11693
## as.factor(sex)        1   334.97     20259      11358 < 2.2e-16 ***
## as.factor(education)  6  1027.11     20253      10331 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model

A binomial logistic regression model was used to explain the relationship between sex, education, and income. Logistic regression is the primary choice of model when response variables are categorical, and since the income responses were noted as categorical ranges in the survey the model was most appropriate. The goal was to predict high or low incomes generally rather than specific numerical incomes, so a binomial classification of upper-level income categories are "high" and lower-level income categories as "low" allowed the binomial logistic regression model to significantly fit the data of interest.

The coefficients of variables in logistic regression equations indicate the change in log odds of the response variable occurring given a unit increase in the variable. The logistic value, $Inc_{>100k}$, is mathematically denoted as $log(p/1-p)$, where p is the probability of the event occurring (i.e. the probability of the income being high income, > CAD 100,000). In this context, the coefficient of 1.0146 for the sex variable being male suggests that males have relatively the higher probability of being in the high-income (> CAD 100,000) group. The -0.8266 coefficient for the level of education being "College, CEGEP or other non-university certificate" indicates that people with College or other non-university certificate education have a slightly lower probability of earning high-income than do other education levels. The coefficient of -2.4333 for the less than high school education level variable suggests someone who has less than high school education is significantly less likely than other education levels to be in the high-income group. People who have a University certificate or diploma below the bachelor's level of education would be very slightly less likely to earn high incomes, since the coefficient is -0.4315, a smaller negative coefficient than the other coefficients. Additionally, the variables male and education level of University certificate, diploma or degree above the bachelors are the only 2 variables with a positive coefficient, suggesting they are the only 2 factors that result in a greater probability of earning high income than the base level. The base level, or intercept when y = 0, in this analysis is a female with a Bachelor's degree. The intercept of -2.3516 implies that the probability of earning in the high income group is approximately 9.1%, for a female with a bachelor's degree.

ANOVA ANALYSIS

Since we are working with categorical data, we run a logistic regression in R. To run a binomial ANOVA for our response variable, we use a chi-squared test to calculate p-values. As the response variable can take either 1 or 0 as value, 1 indicates presence in high income group ($100K or more) while a 0 means not being in a high income group. Both our predictor variables are highly significant at predicting success (i.e value of 1). We looked at other variables that were significant in the model, but not all of them worked together, so we decided to drop them. (for e.g number of children does not correlate with income the way education does, but that is something worth looking into in our next steps) Adding other variables did not increase our

model significance, so we choose the two most relevant and significant predictor variables. (both of those were supported by existing literature)

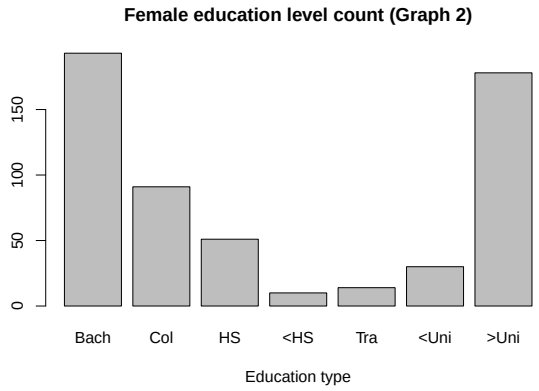
Results

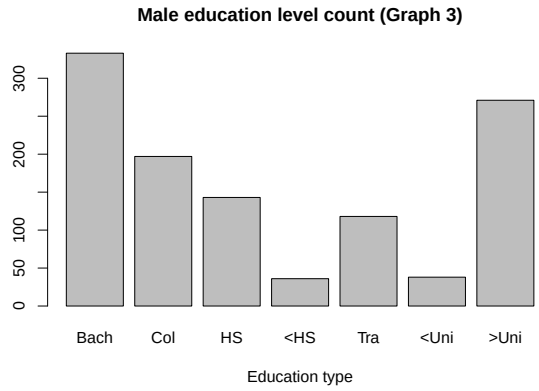
Our model tries to express the likelihood of earning CAD 100K or more as depending on characteristics like sex and level of education. Table ___ shows the Analysis of Deviance Table for both factors, and supports the significance of our model. When we compare the proportion of men and women who earn more than CAD 100K in income, we see huge disparities. Men constitute 2/3rd of the people who earn more than CAD 100K, while women only make up 1/3rd of that size. The distribution of income by level of education varies considerably. We notice a sharp drop in the proportion of people who earn CAD 100K or more, when we reduce the level of education. People with trade certificates or diplomas constitute a very small fraction of those people, while those with a university diploma or some form of bachelor's degree make up the most part. An interesting insight is that between the categories of trade certificate, high school education or less than high school education, we see significant changes, but when we move from university diploma to some bachelor's degree, the numbers are much more comparable (bachelor's degree has a slight lead).

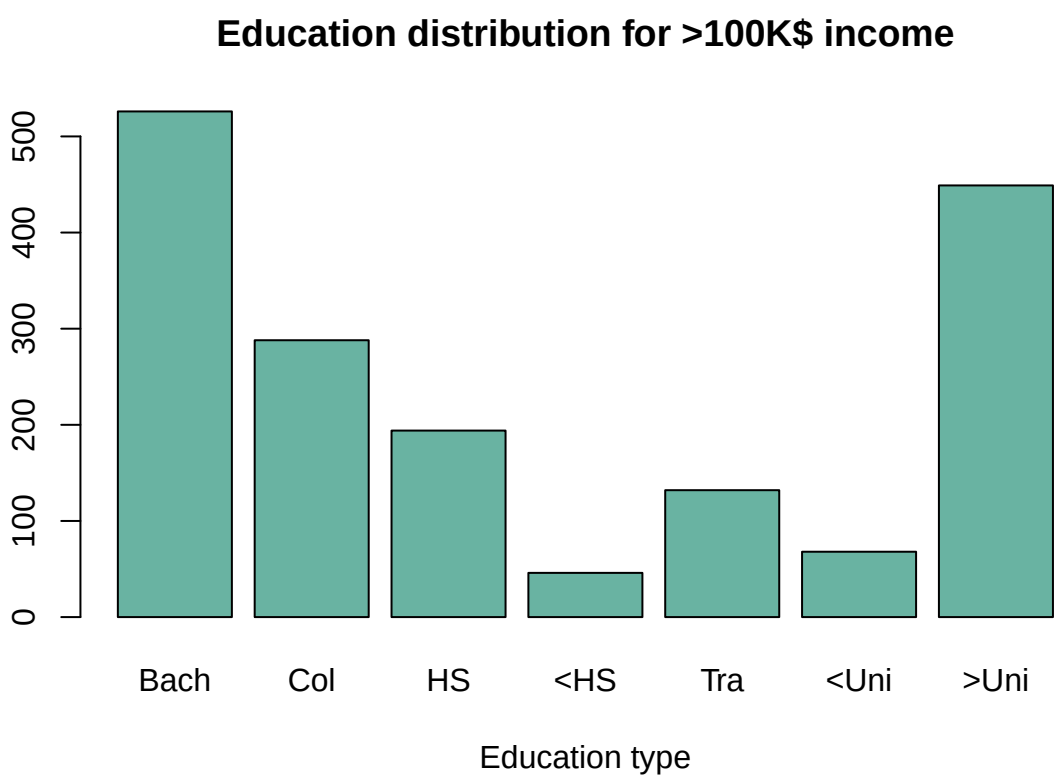
Prportion of male and female with incomes higher than CAD 100,000 respectively:

```
##           n
## 1 0.6670581
```

```
##           n
## 1 0.3329419
```







Where :

Col : Level of education being "College, CEGEP or other non-university certificate"

HS : Level of education being "High school diploma or a high school equivalency certificate"

<HS: Level of education being "Less than high school diploma or its equivalent"

Tra : Level of education being "Trade certificate or diploma"

(<Uni) : Level of education being "University certificate or diploma below the bachelor's level"

(>Uni) : Level of education being "University certificate, diploma or degree above the bachelors"

Discussion and Weaknesses

Our results highlight the disparities in income between people, depending on various social factors. Our model is able to predict the likelihood of a person having high income (CAD 100K or more), based on their sex and education level, and both factors seem to significantly affect our response variable.

Much has been written about the Gender Pay Gap between both men and women- in 2020, the purchasing power parity of the world average of a man's income is CAD 21,000, while that of a woman is significantly lower at CAD 11,000 000 (Global Gender Gap Report, 2020).[7] From our data we can see that men are considerably more likely to earn a higher income than women are. A 2:1 ratio of men to women, is indicative of a wage gap in our sample. Since we are also looking at education levels as well, it is important to note that women MAY be further disadvantaged in access to education, depending on their cultural and social backgrounds. Quite a bit of our data includes migrants and people of diverse cultures, so some women may be earning less because of both, their sex and lower education levels.

When we analyse income depending on levels of education, we see positive correlations between higher income

and higher levels of education. This is consistent with our hypothesis and is backed up by existing research studies. Education is the highest form of investment in human capital, and having a higher level of education could translate to having more skills, knowledge and technical competencies that support a person in having a higher paying job.

The data only looks at male and female genders, and is not inclusive as it should be. For the sample to represent the actual population, sex should not have been a dummy variable, but a choice for people to choose outside of male and female. This could mean that our numbers may not actually represent gender the way they should. Secondly, our data includes a lot of immigrants. For a lot of people who immigrate to Canada, they are unable to transfer their educational qualifications from their home country over to the Canadian system, and it could mean that level of education isn't indicative of a person's skills and may relate to their income differently. Additionally as our predictor variables were categorical, we were limited in what we could do in our statistical analysis.

Next Steps

Since education has been shown to significantly impact income, a possible next step would be to evaluate what factors lead to an individual being more likely to obtain a bachelor's degree and higher or not. Another logistic regression could be performed predicting likelihood of obtaining higher education with predictors such as family income, sex, immigrant status, household type, etc. This would potentially be able to determine how to increase education levels and accessibility. Also, since these relationships/trends have implications for policies and national improvement in disparities or social issues, another potential next step would be an intergenerational analysis over the years comparing income disparities in men and women throughout the years etc. This would help determine the efficacy of certain policies and to determine if the situations are being improved or not over time. A strong predictive model helps identify gaps in policies, and the ability to identify a person's financial situation could support creation of government policies that better target the predictor variables.

Further analysis could also be done with the total children per household as a factor in income levels, as it was found to be significant in the model and may provide further insights on the factors contributing to high income in households. Finally, for a more informative look at the impact of predictive factors on individual income, performing a multinomial logistic regression may be a more efficient algorithm that allows the response variable of income to be non-binary, therefore allowing an anlalysis of the relationship between other factors and specific income categories rather than just high or low income-leading to more accurate conclusions since binary analysis of high income versus low income can be limiting in its conclusions.

Appendix:

Evaluation of sampling methods used in GSS 2017:

The sampling technique used was stratified random sampling, where the total population is divided into sub-groups (strata) of a shared characteristic to allow better and more convenient sampling, and random samples are then taken within the strata. In the 2017 GSS, the 10 provinces were divided into 27 total strata by geographic location[5]. A simple random sample was taken from each of the strata, and each household with eligible respondents (at least one person 15 years or older) was given a telephone interview. The main benefit of the use of stratified sampling in this context is that it allows a representative sample to be taken for the whole target population of all Canadians 15 years or older, which would not be possible through simple random sampling or other methods. The response rate for the survey overall was 52.4%, so the use of strata to represent the entire country using households that did respond was useful considering the voluntary response aspect of surveys[5]. A major trade-off with the use of stratified sampling in this context is that the strata were based on geographic location, whereas the variables measured regarding family are primarily unrelated to geographic relation and therefore this may have decreased homogeneity within the strata, making them not fully representative of the population. The benefit of stratifying samples usually comes from ensuring each subgroup is accurately represented, but in this case geographic location does not necessarily form a homogeneous subgroup for income levels or for sex.

References

1. Government of Canada, S. (2020, February 24). Canadian Income Survey, 2018. Retrieved October 19, 2020, from https://www150.statcan.gc.ca/n1/daily-quotidien/200224/dq200224a-eng.htm

2. Moyser, M. (2019, August 30). Measuring and Analyzing the Gender Pay Gap: A Conceptual and Methodological Overview. Retrieved October 19, 2020, from https://www150.statcan.gc.ca/n1/pub/45-20-0002/452000022019001-eng.htm

3. Porter, E. (2014, September 10). A Simple Equation: More Education = More Income. Retrieved October 19, 2020, from https://www.nytimes.com/2014/09/11/business/economy/a-simple-equation-more-education-more-income.html

4. Eurostat. (n.d.). Quality of life indicators - measuring quality of life. Retrieved October 19, 2020, from https://ec.europa.eu/eurostat/statistics-explained/index.php/Quality_of_life_indicators_-_measuring_quality_of_life

5. Statistics Canada, & Beaupré, P. (2020). Public Use Microdata File Documentation and User's Guide. General Social Survey Cycle 31: Families, 45250001(2019001).

6. Statistics Canada, & Messacar, D. (2018, January 11). Big Tax Data and Economic Analysis: Effects of Personal Income Tax Reassessments and Delayed Tax Filing. Retrieved October 19, 2020, from https://www150.statcan.gc.ca/n1/pub/11-633-x/11-633-x2018012-eng.htm

7. Payscale. (2020). Gender Pay Gap Statistics for 2020. Retrieved October 19, 2020, from https://www.payscale.com/data/gender-pay-gap