An examination of educational qualification on income among male and female graduates

Roy McPherson

2022

Research Question

This report seeks to answer the following research question: "What influence does one's educational qualification have on income among male and female graduates?"

Summary

In summary, the following can be said:

1. Males reported a higher mean median income when compared to females at all levels of education.
2. The differences in mean median income earned was most pronounced among those who had a professional degree.
3. All the differences between mean median income for each level of education was statistically significant, which means they did not occur by chance.
4. The strongest association of median income between sexes was seen among those with professional degrees, in that, as median income for males increased, the median income for females also increased.
5. There was progress towards a decrease in the mean median income gap for males and females at all levels of education as all differences for the period 2010-2015 were trending towards 0.

Methodology

Variables of Interest

The following variables were utilized to answer the research question:

- Educational qualification
- Sex
- Year
- Median income

Data Cleaning

The steps below outlined the data cleaning procedures:

- All missing values were replaced with NA for ease of calculation and preparation of charts.

- The data was transformed from wide to long format, thus abiding by the principles of 'tidy data' where each row represents an observation, and each column represents a variable. This was necessary for the data set to function effectively as a data frame in R. The new column headings were: "ID" (Identification number), "Education qualification", "Area of study", "Sex", "Year", "sum of graduates", and "sum of median income".
- "Sex", "Education qualification" and "Area of study" were changed into factors.

Analysis of findings

Demographics

The largest number of graduates reported that they had an undergraduate degree (853,010). On the other hand, the lowest number of graduates had a doctoral degree (30150). The mean median salary was 62509 annually with those from the doctoral degree deviating the most from the mean when compared to other qualifications (See Figure 1).

Educational qualification	No. of graduates	SD of median income
Undergraduate degree	853010	10410
Professional degree	42790	11834
Master's degree	210700	17114
Doctoral degree	30150	19913
Diploma	323760	8756
Certificate	193200	8912
	N = 1653610	**Mean** = 62509

Figure 1: Number of graduates and standard deviation of median income by educational qualification

Bivariate and inferential analysis

For all levels of qualification, males reported a higher mean median income when compared to females (See Figure 2).

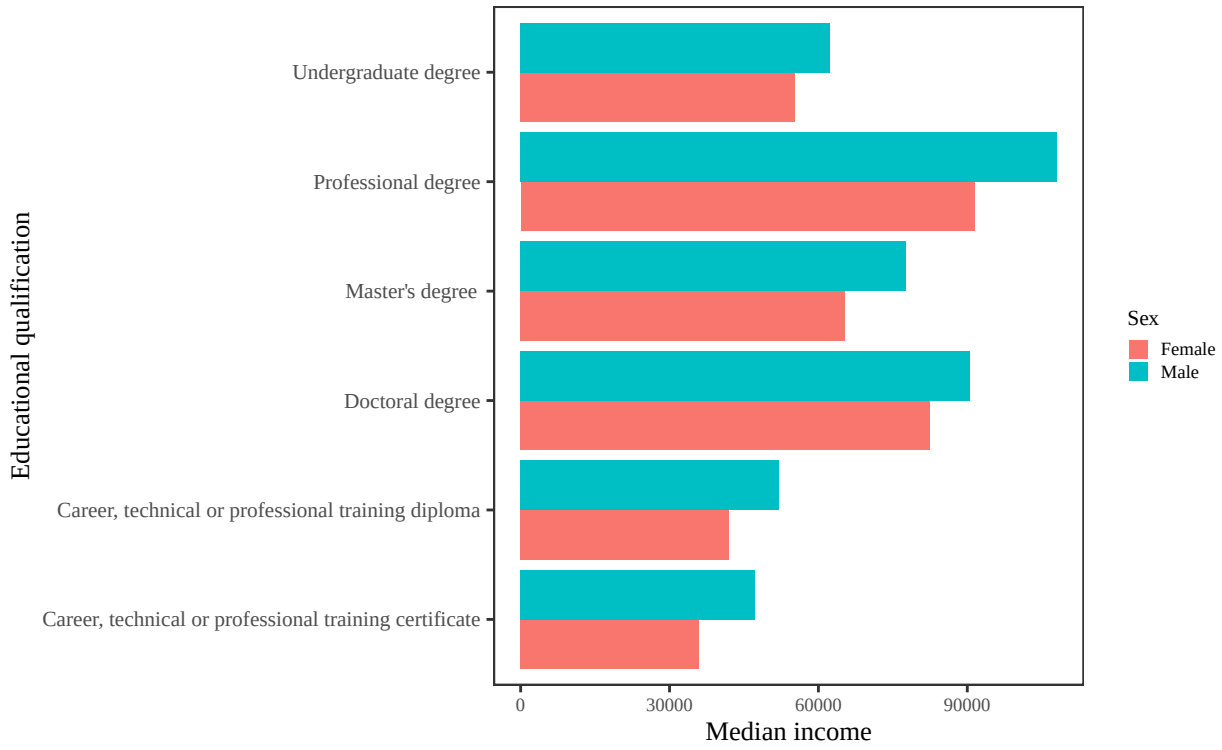

Figure 2: Mean median income by sex and educational qualification

The differences in mean median income between male and female by educational qualification was largest among those who held professional degrees, where females earned 16600 less than males. The smallest difference was between those who held undergraduate degrees, where females earned 7029 less than males (See Figure 3).

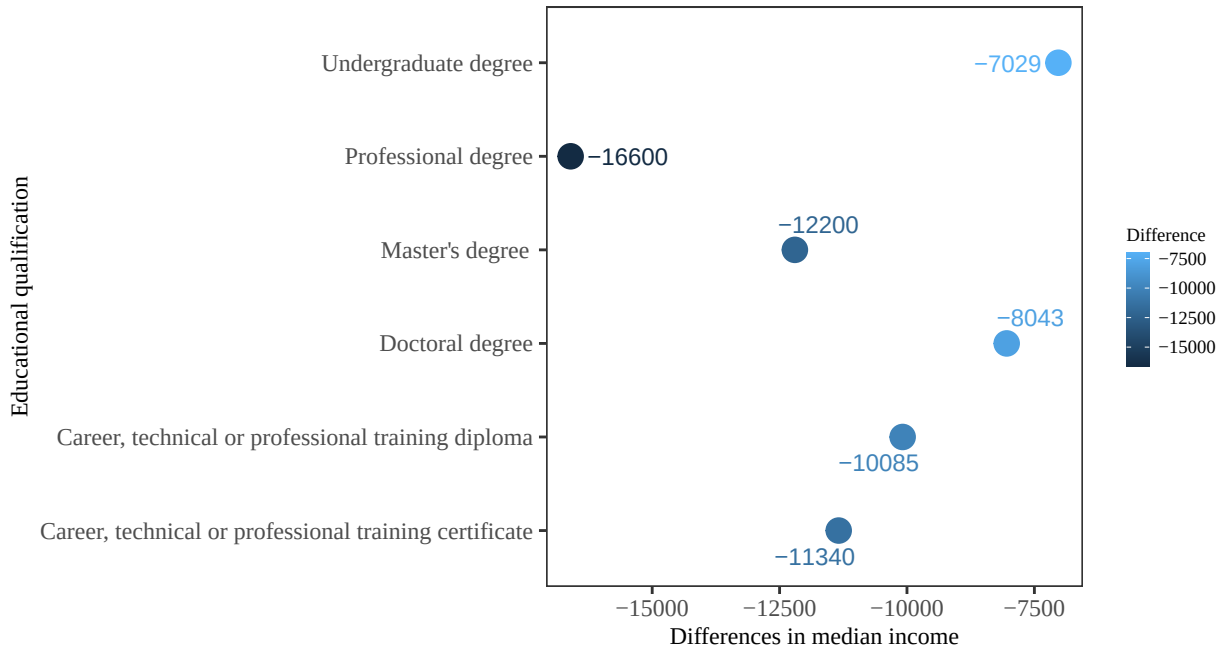

Figure 3: Mean median differences in income by educational qualification

A regression analysis using dummy variables was done to determine if the differences recorded between male and female mean median income were statistically different for each education level. The results provide statistical evidence that there was a significant difference for all levels of education with the doctoral degree coming close to not being statistically significant (See Figure 4).

Educational qualification	**Model**	**Significant**
Undergraduate degree	y = 55258 + 7029x	0.000
Professional degree	y = 91450 +16600x	0.000
Master's degree	y = 65421 + 12200x	0.000
Doctoral degree	y = 82480 + 8043x	0.022
Diploma	y = 42037 + 10085x	0.000
Certificate	y = 35882 + 11340x	0.000
		B_1=Male

Figure 4: Regression models on the mean median income differences by sex and educational qualification

The strongest association of median income between male and female graduates was seen among those with professional degrees, in that, as median income for males increased, the median income for females also increased (r = 0.95). The weakest association was seen between graduates with a certificate (r = 0.44). In addition, all six models recorded a significant p value (See Figure 5).

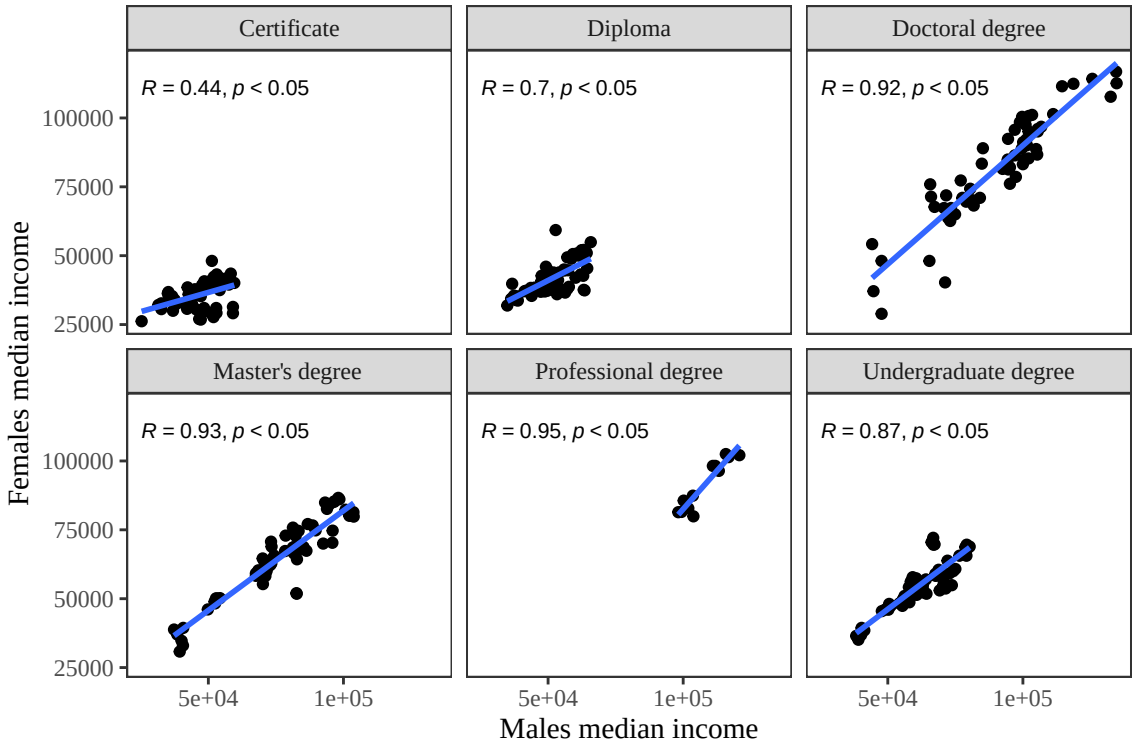

Figure 5: Linear regression of male and female median income by educational qualification

There is a general trend towards a decrease in mean median income differences over the period 2010-2015 among all educational qualification. The differences between those with a doctorate experienced the sharpest decrease in income differences (-54 percent) for the period (See Figure 6) (See percentage change table in Appendix that corroborates the observed trend).

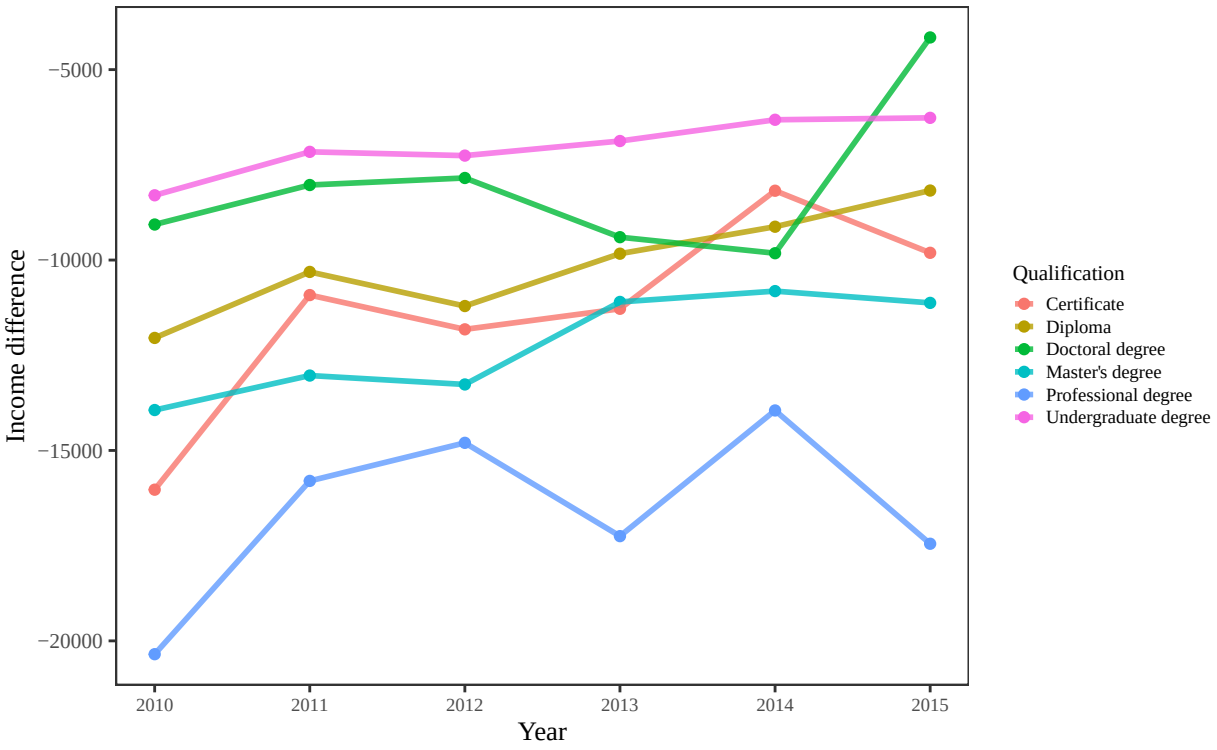

Figure 6: Mean median income differences by education qualification and year

Conclusion

In conclusion, the following can be said:

1. Males reported a higher mean median income when compared to females at all levels of education.
2. The differences in mean median income earned was most pronounced among those who had a professional degree with females earning 16600 less than their male counterpart. The difference was least pronounced among those who had an undergraduate degree with females earning 7029 less than males.
3. All the differences between mean median income for each level of education was statistically significant, which means they did not occur by chance.
4. The strongest association of median income between sexes was seen among those with professional degrees, in that, as median income for males increased, the median income for females also increased.
5. There was progress towards a decrease in the mean median income gap for males and females at all levels of education as all differences were trending towards 0. Of note, the differences between those with a doctorate experienced the sharpest decrease in income differences (-54 percent) for the period.

One of the main limitations of the study is the period of interest was limited to five years, 2010-2015. The study would have been more insightful and relevant if data from more

recent years were available for analysis. Future studies could also focus on variables such as immigration status (e.g. Are you or either parents an immigrant?) and race (e.g. Select the option that best describes your race - Caucasian, African-American, etc.). With the discovery and subsequent apology of systematic racism towards the black community by Toronto's police chief (Loriggio, 2022) and the reality that 23 % of Canada's population are immigrants (Statistics Canada, 2022), such a study is vital in the current Canadian climate.

Appendix

Codebook

Variable name	Type	Short description	Values	Value labels
ID	double	Unique identifier for each observation.	1:864	NA
Education qualification	factor	Measures level of education	NA	• Career, technical or professional training certificate • Career, technical or professional training diploma • Doctoral degree • Master's degree • Professional degree • Undergraduate degree
Area of study	factor	Captures the area of study	NA	• Agriculture, natural resources and conservation • Architecture, engineering, and related technologies • Business, management and public administration • Education • Health and related fields • Humanities • Mathematics, computer and information sciences • Other instructional programs • Personal, protective and transportation services • Physical and life sciences and technologies • Social and behavioural sciences and law • Visual and performing arts, and communications technologies

Variable name	Type	Short description	Values	Value labels
Sex	factor	Captures participants sex	NA	• Male • Female
Year	double	Captures the year of each observation	2010:2015	NA
sum_of_graduates	double	Captures the total number of graduates	0:22430	NA
sum_of_median_income	double	Captures the median income of participants	25300:134700	NA

*Format adopted from the National Victimisation Survey, Bureau of Justice 2022.

R code

Demographics

```
final_dataset %>%
  group_by(`Education qualification`) %>%
  summarize(Sum = sum(sum_of_graduates, na.rm = T))
```

```
final_dataset %>%
  group_by(`Education qualification`) %>%
  summarize(Sum = round(sd(sum_of_median_income, na.rm = T),0))
```

Bivariate and inferential analysis

```
# Median income by sex and educational qualification (Figure 2)
final_dataset %>%
  group_by(`Education qualification`, Sex) %>%
  summarise(`Median income` =
  round(mean(sum_of_median_income, na.rm=TRUE),0)) %>%
  ggplot(aes(x = `Education qualification`,
  y = `Median income`, fill = Sex)) +
  geom_col(position = "dodge") + coord_flip() +
  theme_bw() +
  theme(legend.key.height= unit(3, 'mm'),
        legend.key.width= unit(3, 'mm'),
        legend.title = element_text(size=8),
        legend.text = element_text(size = 7),
        axis.text.y = element_text(size = 8),
        axis.text.x = element_text(size = 7),
        text = element_text(family = "serif", color = "black"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title = element_text(size=10))+
  labs(y = "Median income", x = "Educational qualification")
```

```r
# The median differences in income by educational
#qualification were calculated and saved in
#excel and plotted in R (Figure 3)

library(ggrepel)

Mean_income_differences_gender <-
  read_csv("Mean_income_differences_gender.csv")

ggplot(data=Mean_income_differences_gender,
       aes(x= `Educational qualification`, y= Difference,
           color = Difference)) + geom_point(size = 4) +
  coord_flip() + theme_bw() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        text = element_text(family = "serif"),
        legend.key.height= unit(3, 'mm'),
        legend.key.width= unit(3, 'mm'),
        legend.title = element_text(size=7),
        legend.text = element_text(size = 7),
        axis.title = element_text(size = 8.5)) +
  labs(x = "Educational qualification",
       y = "Differences in median income") +
  geom_text_repel(aes(label = Difference),
                  box.padding   = 0.35,
                  point.padding = 0.5,
                  segment.color = 'grey50', size = 3)
```

```r
#Regression with dummy variables for certificate programs

certificate_only <- final_dataset %>%
  filter(`Education qualification` ==
           "Career, technical or professional training certificate") %>%
  lm(sum_of_median_income ~ Sex, data =.) %>%
  summary()

certificate_only$coefficients[,c(1,4)]
```

```r
#Regression with dummy variables for diploma programs

diploma_only <- final_dataset %>%
  filter(`Education qualification` ==
           "Career, technical or professional training diploma") %>%
  lm(sum_of_median_income ~ Sex, data =.) %>%
```

```
  summary()

diploma_only$coefficients[,c(1,4)]
```

```
#Regression with dummy variables for dotorate programs

doctorate_only <- final_dataset %>%
  filter(`Education qualification` == "Doctoral degree") %>%
  lm(sum_of_median_income ~ Sex, data =.) %>%
  summary()

doctorate_only$coefficients[,c(1,4)]
```

```
#Regression with dummy variables for masters programs

masters_only <- final_dataset %>%
  filter(`Education qualification` == "Master's degree") %>%
  lm(sum_of_median_income ~ Sex, data =.) %>%
  summary()

masters_only$coefficients[,c(1,4)]
```

```
#Regression with dummy variables for professional programs

professional_only <- final_dataset %>%
  filter(`Education qualification` == "Professional degree") %>%
  lm(sum_of_median_income ~ Sex, data =.) %>%
  summary()

professional_only$coefficients[,c(1,4)]
```

```
#Regression with dummy variables for undergraduate programs

undergraduate_only <- final_dataset %>%
  filter(`Education qualification` == "Undergraduate degree") %>%
  lm(sum_of_median_income ~ Sex, data =.) %>%
  summary()

undergraduate_only$coefficients[,c(1,4)]
```

```
#Correlation between male and female median income by
#educational qualification (Figure 5)
```

```
library(ggpmisc)
library(ggpubr)


Male_median_income <- final_dataset%>%
  dplyr::filter(Sex == "Male") %>%
  dplyr::mutate(`Education qualification` =
                  forcats::fct_recode(`Education qualification`,
'Certificate' = "Career, technical or professional training certificate",
'Diploma' = "Career, technical or professional training diploma",
'Undergraduate degree' = "Undergraduate degree",
'Professional degree' = "Professional degree",
'Doctoral degree' = "Doctoral degree",
"Master's degree" = "Master's degree"))


Female_median_income <- final_dataset%>%
  dplyr::filter(Sex == "Female")

ggplot(Male_median_income, aes(x = sum_of_median_income,
    y = Female_median_income$sum_of_median_income)) +
    geom_point()  +  stat_cor(method = "pearson",
    p.accuracy = 0.05, size = 2.9, position = "jitter") +
  facet_wrap(~Male_median_income$`Education qualification`) +
  stat_poly_line(se =FALSE) +
  labs( y = "Females median income", x = "Males median income") +
  theme_bw() + theme(panel.grid.major = element_blank(),
                     panel.grid.minor = element_blank(),
                     text = element_text(family = "serif"))
```

```
#Income difference by educational qualification and year (Figure 6)

final_dataset %>%
  pivot_wider(names_from = "Sex",
              values_from = sum_of_median_income) %>%
  group_by(Year, `Education qualification`) %>%
  summarize(Male = round(mean(Male, na.rm =T),0),
            Female = round(mean(Female, na.rm=T),0)) %>%
  mutate(Difference = Female - Male, Qualification =
           forcats::fct_recode(`Education qualification`,
'Certificate' = "Career, technical or professional training certificate",
'Diploma' = "Career, technical or professional training diploma",
'Undergraduate degree' = "Undergraduate degree",
'Professional degree' = "Professional degree",
```

```
'Doctoral degree' = "Doctoral degree",
"Master's degree" = "Master's degree")) %>%
  ggplot(aes(x = Year, y = Difference,
             group= Qualification, color = Qualification)) +
  geom_point() + geom_line(size=1, alpha=.8) + theme_bw() +
  theme(legend.key.height= unit(3, 'mm'),
        legend.key.width= unit(3, 'mm'),
        legend.title = element_text(size=8),
        legend.text = element_text(size = 7),
        axis.text.y = element_text(size = 8),
        axis.text.x = element_text(size = 7),
        text = element_text(family = "serif", color = "black"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title = element_text(size=10)) +
  labs(y = "Income difference")
```

```
#Percentage change in differences in median income by qualification
#for the period 2010-2015

x <- final_dataset %>%
  pivot_wider(names_from = "Sex",
              values_from = sum_of_median_income) %>%
  group_by(Year, `Education qualification`) %>%
  summarize(Male = round(mean(Male, na.rm =T),0),
            Female = round(mean(Female, na.rm=T),0)) %>%
  mutate(Difference = Female - Male, Qualification =
           forcats::fct_recode(`Education qualification`,
'Certificate' = "Career, technical or professional training certificate",
'Diploma' = "Career, technical or professional training diploma",
'Undergraduate degree' = "Undergraduate degree",
'Professional degree' = "Professional degree",
'Doctoral degree' = "Doctoral degree",
"Master's degree" = "Master's degree")) %>%
  group_by(Year, Qualification) %>%
  summarize(Difference = sum(Difference)) %>%
  pivot_wider(names_from = "Year",
              values_from = Difference) %>%
  mutate(Percentage_change = round((`2015`-`2010`)/`2010`*100, 0)) %>%
  select(Qualification, Percentage_change) %>%
  arrange(Percentage_change)

data.frame(x)
```

```
         Qualification Percentage_change
1      Doctoral degree               -54
2          Certificate               -39
3              Diploma               -32
4 Undergraduate degree               -25
5      Master's degree               -20
6  Professional degree               -14
```

References

1. Statistics Canada (2022). *Immigrants make up the largest share of the population in over 150 years and continue to shape who we are as Canadians.* Retrieved November 22, 2022 from https://www150.statcan.gc.ca/n1/daily-quotidien/221026/dq221026a-eng.htm

2. Loriggio, P., (2022). *Toronto police chief apologizes to Black community as race-based data released.* Retrieved November 22, 2022 from https://globalnews.ca/news/8922183/toronto-police-chief-apologizes-black-community-race-based-data/

3. Kassambara, A. (2018). *Machine Learning Essentials: Practical Guide in R.* CreateSpace Independent Publishing Platform.