Visible Minority, Visible Gaps in Education and Income? An Analysis on Income and Education Through the Lens of Visible Minority in Toronto

Faria Khandaker, Asel Kushkeyeva, and Shabrina Mardevi

02/24/2020

Abstract

In this report we analyzed education and income through the lens of diversity using two National Household Survey Datasets from Open Data Toronto. We found a slightly positive correlation between visible minorities and diplomas held from colleges and trade schools. We also found a slightly negative correlation between an increase in population diversity, income and post-secondary education. This may be indicative of systematic and financial barriers faced by people of diverse backgrounds or it may be indicative of people seizing the opportunity to work in high demand, low competition positions.

Introduction

There is often conflicting discourse in the media regarding minority populations. Some claim that minorities are free loaders who only collect welfare checks contribute to national debt. Others claim that minorities "are stealing our jobs". We were interested to see what the data says. We conducted two different analysis: one comparing income with population diversity and one comparing education with population diversity. We used The Demographics NHS dataset and The Education NHS dataset from the Toronto's Open data portal. We compared 4 different education levels: apprenticeship and trades, Non-University Certificates, Bachelor's degree and Graduate degree. There is no clear comparison between minority group and education level. While the Apprenticeship and Non-University categories show a weak positive correlation between percentage of visible minority in the population and the completion of the program, the bachelor's and graduate degrees has a slightly negative slope, but the data points are evenly dispersed in both graphs. The factors that contribute to these negative slopes of university-level academic qualifications can be because of financial and systematic barriers. Our study concludes that minority populations are overall educated but perhaps underpaid/underemployed; they are diverse not only in their culture but also in the level of skills that is present in the communities to fit the employment market.

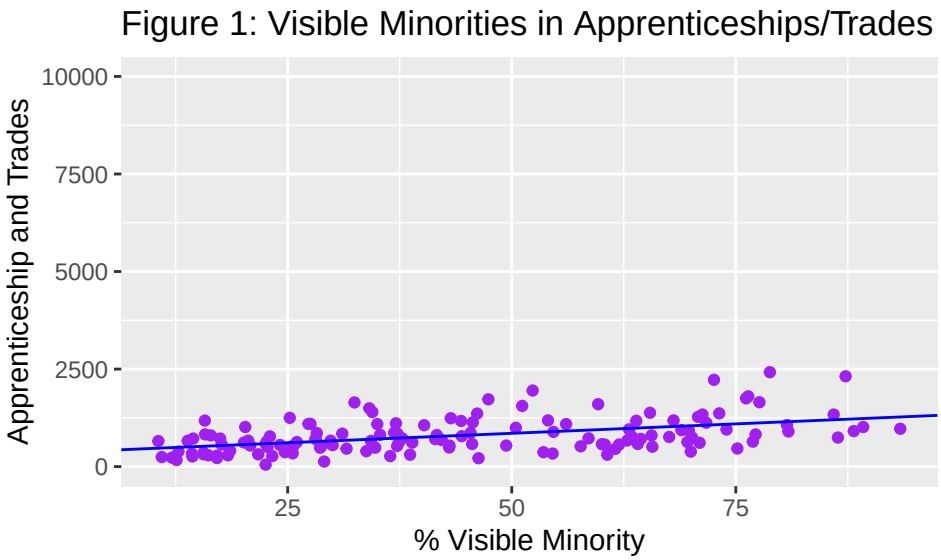
Dataset

The datasets used both belong to Wellbeing Toronto - Demographics: NHS Indicators. One dataset has demographic information such as breakdown of population based on culture, income, education, employment, marital status, etc. The education dataset has a more detailed focus on education and a larger breakdown of educational categories. Both contain information on neighbourhood (Hood) number in Toronto and 140 rows of data. The dataset was last updated on April 30th, 2016.

Analysis and Discussions

We compared the percentage of minorities in a neighbourhood to the four different levels of education. Later, we analysed neighbourhood median net income by percent visible minority population. Taking the percentage minimizes the bias and skew of data as the numbers for each educational category was for the total population each hood.

Figure 1 shows a slight positive correlation, which means that with the increase of visible minority, the higher likeliness of apprenticeship and trades being the highest level of education attainment in that neighbourhood. Characteristically, it has no outliers and all of its points are clustered along the regression line, showing that observed values are close to predicted ones.



Similar to Figure 1, Figure 2 shows a slight positive correlation. With the increase of visible minority, the higher it is the likeliness of non-university certificates being held in that neighbourhood. It has some outliers and the points are more loosely dispersed surrounding the regression line.

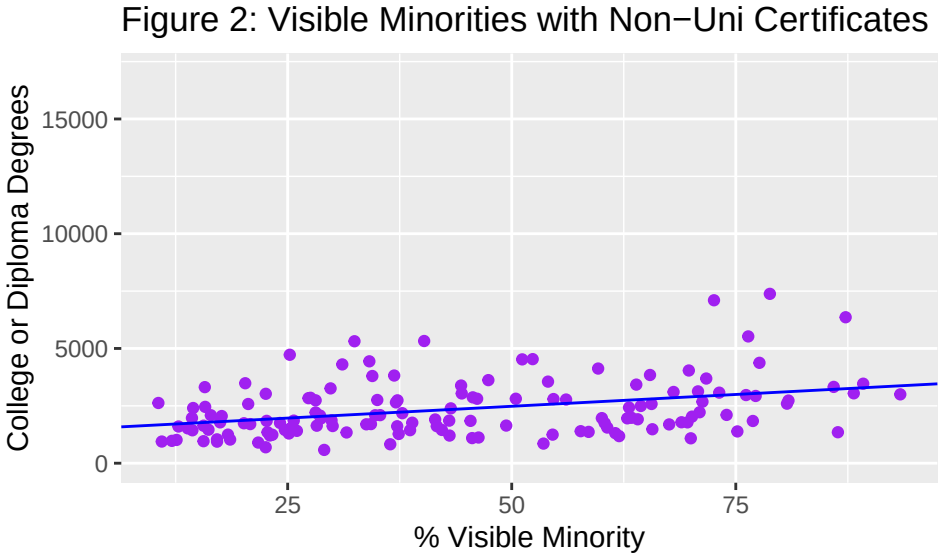


Figure 3 and 4 show slight negative relationships, with Figure 3 having most outliers. Figure 6 (Appendix 2) shows a version of Figure 3 that is cleaned off of outliers, yet it does not show significant difference. Based

on these graphs, highest attainment of education for both university levels decrease as the population's percentage of visible minorities increases in a neighbourhood. It is likely that there are less people from diverse backgrounds that are able to or decide to pursue higher education, since they more likely to belong in lower income brackets (Huffman, 2017). As costs for post-secondary education continues to rise, many end up looking to other quicker and cheaper options to avoid a life of debt (Huffman, 2017).

Figure 3: Visible Minorities with Bachelor's Degree

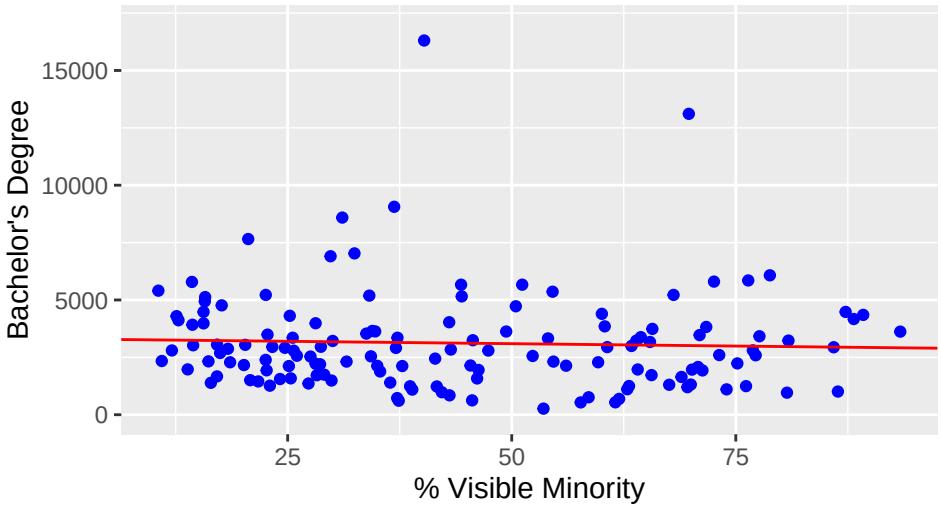
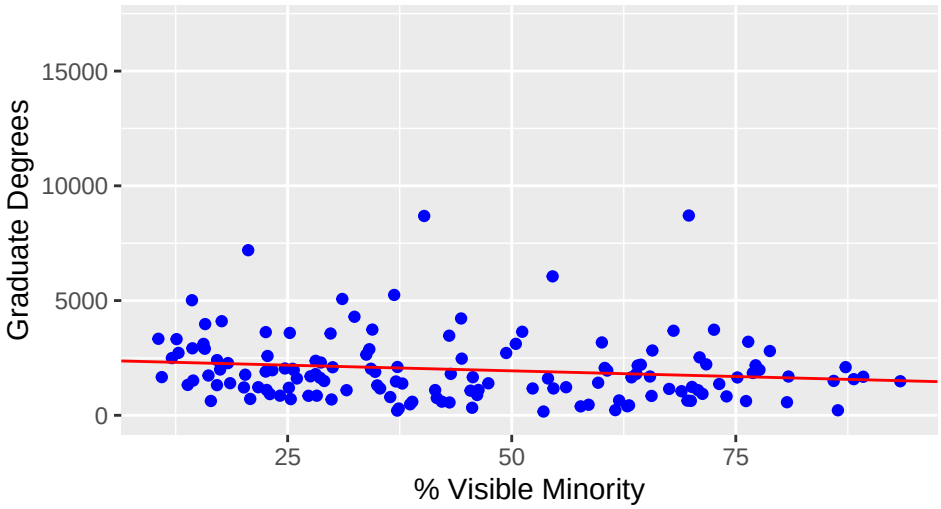
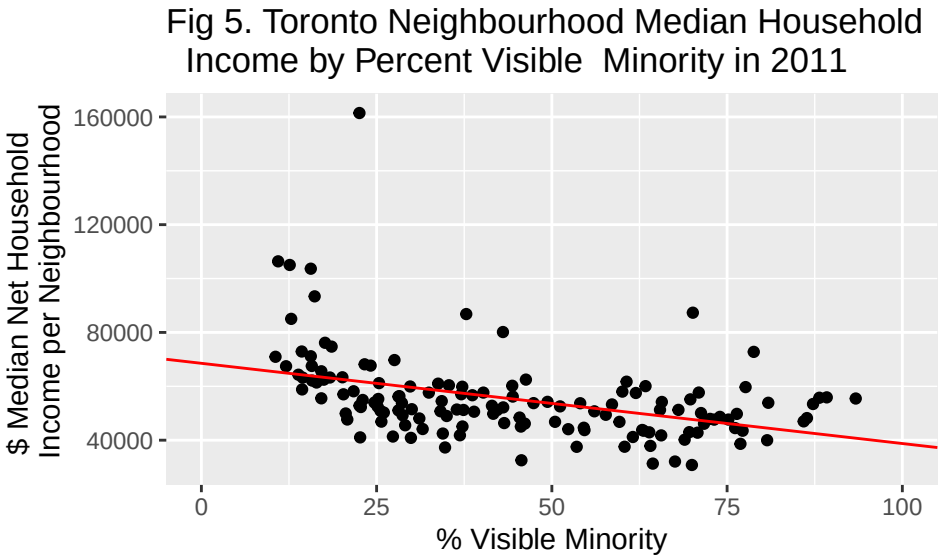


Figure 4: Visible Minorities with Graduate Degrees



High diversity and low-income assumption in this context is first proven by Figure 5. In Summary 1, the lowest neighbourhood median income is 30,794 CAD. Figure 5 illustrates that such median income is situated in a neighbourhood with relatively high (at approximately 70%) population of visible minority. To provide a context against an extreme, highest median net income quartile starts from 59,963 CAD. In Figure 5, it shows that neighbourhoods with ~20% or below population of visible minority are exclusively in this income quartile. Figure 7 (Appendix 2) was produced to check for any significant changes in the model if the outlier of $162,000 was removed. However, no noticeable change was revealed.



Summary 1. Summary of Neighbourhood After-Tax Median Income

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30794   46690   52660   55426   59963  161448
```

Further, in Summary 2, a median household net income of 68,538.80 CAD is expected for neighbourhoods with hypothetical 0.00% of visible minority, which is a number above highest median income quartile. With one percent increase of visible minority population in a neighbourhood, median household net income decreases by 297.83 CAD. This correlation is shown as significant at 99.99% confidence interval. Rsquared, however, is low at 0.168; this means that roughly only 16.85% of the variance found in the response variable of Median net Income can be explained by the predictor variable of Percent Visible Minority. Despite low Rsquared, it should not be disregarded that changes in the population of visible minority are still negatively associated with changes in median net income.

Summary 2. Residuals and Regression Coefficients of Neighbourhood Median After-Tax Income by Percent Visible Minority

```
##
## Call:
## lm(formula = medi_net_income ~ percent_minority)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22401  -7953  -1818   3996  99625
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      68538.80    2774.96  24.699  < 2e-16 ***
## percent_minority  -297.83      56.31  -5.289 4.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14750 on 138 degrees of freedom
```

```
## Multiple R-squared:  0.1685, Adjusted R-squared:  0.1625
## F-statistic: 27.97 on 1 and 138 DF,  p-value: 4.707e-07
```

One explanation to address this income inequality is that immigrants to Canada face systematic barriers to high-level jobs. Employers often look for "Canadian Experience" on resumes or as in the case of health professions, look for Canadian credentials from medical or pharmacy schools (CBC News, 2016). Networking is another important aspect to the job hunt, an aspect which immigrants and other visible minority groups face whether because of language, discrimination, or not knowing how to go about networking (CBC News, 2016).

Consistent with the relationship shown in Figure 1, ethnic/diversity barriers in income and employment often lead highly educated immigrants to seek out other career paths through trade school or college. It is possible for people belong to more than one educational category, but our dataset is unable to differentiate that. Many times, people finish their bachelor's degree but cannot find jobs in that field (Matti, 2019). Therefore, according to a 2019 news report by CBC, Canadian employers have a hard time filling in Tradesmen roles like electricians or welders. Instead of returning to 4 more years of school, some decide to get a diploma or a certificate in fields that are less competitive or are facing a labour shortage (Matti, 2019).

Limitations

A limitation of this dataset is the age of the data. This dataset is from 2011 and since then, many policies have changed, job trends have changed, many waves of skilled immigrants have migrated to Canada, etc. The results we have may not provide an accurate depiction of the experiences of today's visible minorities. Other major limitations lie upon the choices and forms of the data/variables. Filtering out populations in non-active years (outside 24-64 years old) are essential practice in measuring highest education attainments and income to reduce age bias. For the bachelor and graduate graphs, the pattern of dispersion of the data and number of outliers can also indicate a push for post-secondary education in a certain minority community versus others. Moreover, the grouping all visible minorities together is an overall weakness of this dataset and our results.

Assumptions of Linear Regression

According to Bansal (n.d), the four assumptions of linear regression are: 1 - Linearity of residuals 2 - Independence of residuals 3 - Normal distribution of residuals 4 - Equal variance of residuals

We drew scatterplots with our dependent variables and residuals to check for linearity and equal variance assumptions (Figures 8 to 12 in Appendix 3). Both are met in our models.

Cross-sectional and longitudinal datasets are the two common types of datasets. Cross-sectionals are the ones where the entity data is collected once. Whereas, longitudinal datasets contain information about the same entity over time. Since we worked with cross-sectional datasets, the independence of the variables is assumed to be met.

Normal distribution of residuals is checked by drawing a histogram of the residuals. Our residuals are distributed normally, that means the assumption is met (Figures 13 to 17 in Appendix 3). Although, in some cases the histograms are slightly skewed, they are not far deviated from being a normal distribution.

We can conclude, that our models conform to the assumptions of our chosen analysis method.

Ethics and the Importance of Open Datasets

Concerns around privacy and security are raised when the topic of having datasets open for public access are discussed. "Data is the new oil" is a common phrase used to describe the craze behind the storage, retrieval

and analysis of data that has risen within the past decade and is shaping the future of the workforce (The Economist, 2017).

There are many ethical considerations that should come into play when working with data concerning living beings. Firstly, while it is easy to make assumptions on about groups of people based on numbers and graphs, we have to remember that these are real people who have to make real choices based on a number of factors. It is important to look at the numbers in a broader context to look for the reasons behind the results. Looking at our graphs, policy makers can assume that children from minority backgrounds are less likely to pursue higher education and therefore more funding should be allocated to after school programs. In reality, the problem maybe more related to rising cost of both post-secondary education and living costs within the city. If data is a force behind policy changes, data analysts have to be careful about shaping results to fit certain narratives.

Open datasets allow for transparency and accountability of officials, innovation and enhancement of public policy and improve efficiency in services (European Data Portal, n.d). Transparency and accountability of government offices are important for gaining and retaining public trust. Within a society, it encourages research and innovation; as students like us are able to take real life raw data, clean it, analyze it, see the problems within our own environments and try to come up with a solution.

Our datasets are under the consultations of Open Government License. This license states that the Information Provider grants royalty-free and non-exclusive permission to use, copy, modify, and disseminate the data. Several exemptions of the type of data not for use, in this project and general, are personal information, health information, and other items tied onto intellectual property rights. Source of information should be credited whenever possible and not be used as a form of endorsement.

References

- City of Toronto. (2013, August 19). Open Government License - Toronto. Web. Retrieved from https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-licence/
- Bansal, G. (n.d). What are the four assumptions of linear regression?. University of Winsconsin Green Bay Blog. Web. Retrieved from https://blog.uwgb.edu/bansalg/statistics-data-analytics/linear-regression/what-are-the-four-assumptions-of-linear-regr ession/
- CBC News. (2016, April 14). Lack of 'Canadian experience' an obstacle to immigrant employment. Web. Retrieved from https://www.cbc.ca/news/canada/kitchener-waterloo/immigrants-jobs-skills-conference-multi-cultural-1.3534121
- The Economist. (2017, May 6). The world's most valuable resource is no longer oil, but data. Web. Retrieved from https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data)
- Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2019). skimr: Compact and Flexible Summaries of Data. R package version 2.0.2. https://CRAN.R-project.org/package=skimr
- European Data Portal. (n.d). Benefits of Open Data. Web. Retrieved from https://www.europeandataportal.eu/en/using-data/benefits-of-open-data
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr
- Hadley Wickham and Edgar Ruiz (2019). dbplyr: A 'dplyr' Back End for Databases. R package version 1.4.2. https://CRAN.R-project.org/package=dbplyr
- Hadley Wickham and Lionel Henry (2019). tidyr: Tidy Messy Data. R package version 1.0.0. https://CRAN.R-project.org/package=tidyr
- Huffman, M. (2017, May 10). Survey finds debt may be discouraging students from college careers. Consumer Affairs. Web. Retrieved from https://www.consumeraffairs.com/news/survey-finds-debt-may-be-discouraging-students-from-college-careers-100517.html
- Matti, M. (2019, October 4). Canada's skilled labour shortage: What does it mean for workers and employers. CTV News. Web Retrieved from https://www.ctvnews.ca/canada/canada-s-skilled-labour-shortage-what-does-it-mean-for-workers-and-employers-1.4623996
- Sharla Gelfand (2019). opendatatoronto: Access the City of Toronto Open Data Portal. R package version 0.1.1. https://CRAN.R-project.org/package=opendatatoronto
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Appendix 1: Data Preparation Codes

Setup

```
knitr::opts_chunk$set(echo = TRUE)
```

```
library(opendatatoronto)
library(dbplyr)
library(tidyr)
library(ggplot2)
library(tidyverse)
library(stringr)
library(skimr)
```

Dataset Download

```
demo_nhs_package<-search_packages("Wellbeing Toronto - Demographics: NHS Indicators")
demo_nhs_resources<-demo_nhs_package %>% list_package_resources()
demo_nhs_resources
```

```
demo_edu_dataset<-get_resource("c5bffd35-da12-48d9-9fb7-3ec61f0edddb")
demo_edu_dataset
head(demo_edu_dataset)
```

```
skim(demo_edu_dataset)
```

Cleaning

```
demo_nhs_dataset<-get_resource("8d838d5c-20da-48bd-b4a7-3e66b1c25b55")
demo_nhs_dataset = demo_nhs_dataset[-1,]
```

```
demo_edu<-demo_edu_dataset

# Clean dataset to include neighborhood names and hood numbers
demo_edu<-demo_edu%>%
  separate(Geography, into= c("name", "geonum"), sep = "\\(")%>% #separating numbers
  separate(name, into=c("rest", "Neighborhood"), sep = "D - ")%>% #separating names
  #getting rid of extra parenthesis
  separate(geonum, into= c("geonum", "paren"), sep = "\\)")%>%
  #getting rid of extra columns
  select(-paren)%>%
  select(-rest)

#convert geonum to numeric
demo_edu$geonum<- as.numeric(as.character(demo_edu$geonum))

#create tibble w/specifc values for analysis
```

```r
#cleaning practice with other dataset
hoodnum<-hood_data %>%
  separate(Geography, into= c("name", "number"), sep = "\\(")
  stringr::str_extract_all(hood_data, "\\d")

hoodnum<-hoodnum%>%
  separate(name, into=c("rest", "neighborhood"), sep = "D - ")
hoodnum<-hoodnum %>% select(-name)

hoodnumc<-hoodnum%>%
  separate(number, into= c("num", "paren"), sep = "\\)")

hoodnum<-hoodnumc%>%
  select(-paren)
hoodnum<- as.numeric(as.character(hoodnum$num))  # Convert to numeric
hoodnum<-as.tibble(hoodnum)

new_hood_data<-merge(demo_edu_dataset,hoodnum)
```

Data Creation

```r
#education by hood
edu_tab<-
  tibble(Hood=demo_edu$geonum, Neighborhood=demo_edu$Neighborhood,
         Appren_or_Trade=demo_edu$`Apprenticeship or trades certificate
         or diploma`,
         Non_Uni_Cert=demo_edu$`College, CEGEP or other non-university
         certificate or diploma`,
         Bachelors=demo_edu$`Bachelor's degree`,
         Above_Bachelors=demo_edu$`University certificate,
         diploma or degree above bachelor level`)
#total visible minorities by hood
min_tab <-tibble(Hood=demo_nhs_dataset$`Hood#`,
                 Vis_Min=demo_nhs_dataset$`Total visible minority population`)
min_tab$Hood<- as.numeric(as.character(min_tab$Hood))
min_tab$Vis_Min<- as.numeric(as.character(min_tab$Vis_Min))
#merge edu and min
edu_by_min<-merge(edu_tab, min_tab, by="Hood")
#income and population proportion
income_prop<-tibble(Hood=demo_nhs_dataset$`Hood#`,
                    Tot_Pop=demo_nhs_dataset$`Total Population`,
                    Med_Inc=demo_nhs_dataset$`Median after-tax household income $`)

income_prop$Hood<-as.numeric(as.character(income_prop$Hood))
income_prop$Tot_Pop<-as.numeric(as.character(income_prop$Tot_Pop))
income_prop$Med_Inc<-as.numeric(as.character(income_prop$Med_Inc))

edu_inc_by_min<-merge(edu_by_min, income_prop, by="Hood")
edu_inc_by_min<- edu_inc_by_min %>%
  mutate(Vis_Min_Per= (Vis_Min/Tot_Pop)*100)
summary(edu_inc_by_min)
```

Appendix 2: Summaries and Figures

Summary 1

```
summary(medi_net_income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30794   46690   52660   55426   59963  161448
```

Summary 2

```
reg_incdiv <- lm(medi_net_income~percent_minority)
summary(reg_incdiv)
```

```
##
## Call:
## lm(formula = medi_net_income ~ percent_minority)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22401  -7953  -1818   3996  99625
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      68538.80    2774.96  24.699  < 2e-16 ***
## percent_minority  -297.83      56.31  -5.289 4.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14750 on 138 degrees of freedom
## Multiple R-squared:  0.1685, Adjusted R-squared:  0.1625
## F-statistic: 27.97 on 1 and 138 DF,  p-value: 4.707e-07
```

Figure 1

```
lmapp<-lm(Appren_or_Trade~Vis_Min_Per, data=edu_inc_by_min)
summary(lmapp)
```

```
##
## Call:
## lm(formula = Appren_or_Trade ~ Vis_Min_Per, data = edu_inc_by_min)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -661.79 -263.90  -87.86  183.66 1287.89
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  370.201     74.772   4.951 2.12e-06 ***
## Vis_Min_Per    9.668      1.517   6.372 2.60e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 397.5 on 138 degrees of freedom
## Multiple R-squared:  0.2273, Adjusted R-squared:  0.2217
## F-statistic:  40.6 on 1 and 138 DF,  p-value: 2.598e-09
```

```
ggplot(edu_inc_by_min, aes(x=Vis_Min_Per, y=Appren_or_Trade)) +
  geom_point(color = 'purple') +
  ylim(0,10000) +
  geom_abline(intercept= 370.201, slope=9.668, colour="blue") +
  labs(title = "Figure 1. Visible Minorities in Apprenticeships/Trades",
       x = "% Visible Minority",
       y = "Apprenticeship and Trades")
```

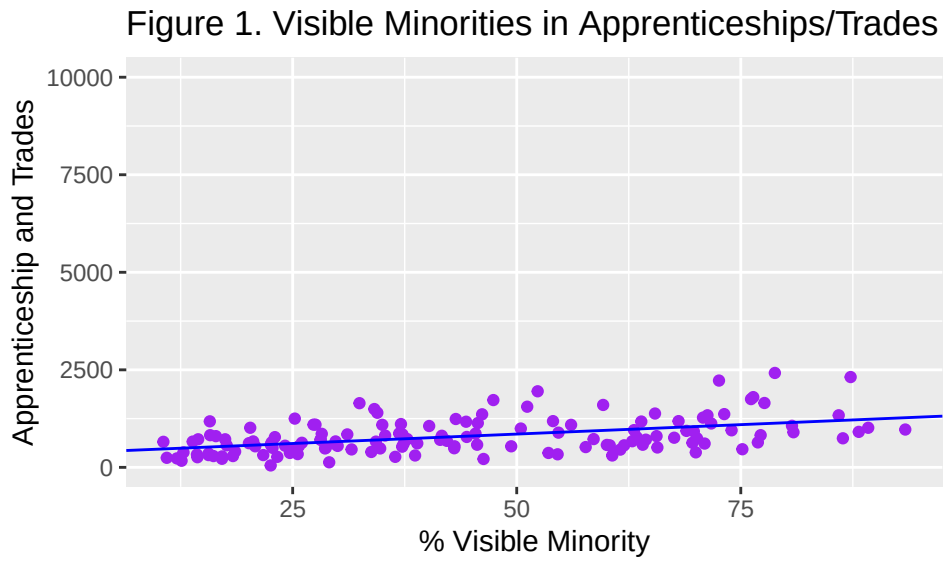


Figure 2

```
nonuni<-lm(Non_Uni_Cert~Vis_Min_Per, data=edu_inc_by_min)
summary(nonuni)
```

```
##
## Call:
## lm(formula = Non_Uni_Cert ~ Vis_Min_Per, data = edu_inc_by_min)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1880.0  -766.9  -221.7   536.9  4306.2
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1451.570    219.716   6.607 7.88e-10 ***
## Vis_Min_Per   20.585      4.459   4.617 8.82e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1168 on 138 degrees of freedom
## Multiple R-squared:  0.1338, Adjusted R-squared:  0.1275
## F-statistic: 21.32 on 1 and 138 DF,  p-value: 8.822e-06
```

```
ggplot(edu_inc_by_min, aes(x=Vis_Min_Per, y=Non_Uni_Cert)) +
  geom_point(color = 'purple') +
  geom_abline(intercept= 1451.570, slope=20.585, colour="blue") +
  ylim(0,17000) +
  labs(title = "Figure 2. Visible Minorities with Non-Uni Certificates",
       x = "% Visible Minority",
       y = "College or Diploma Degrees")
```

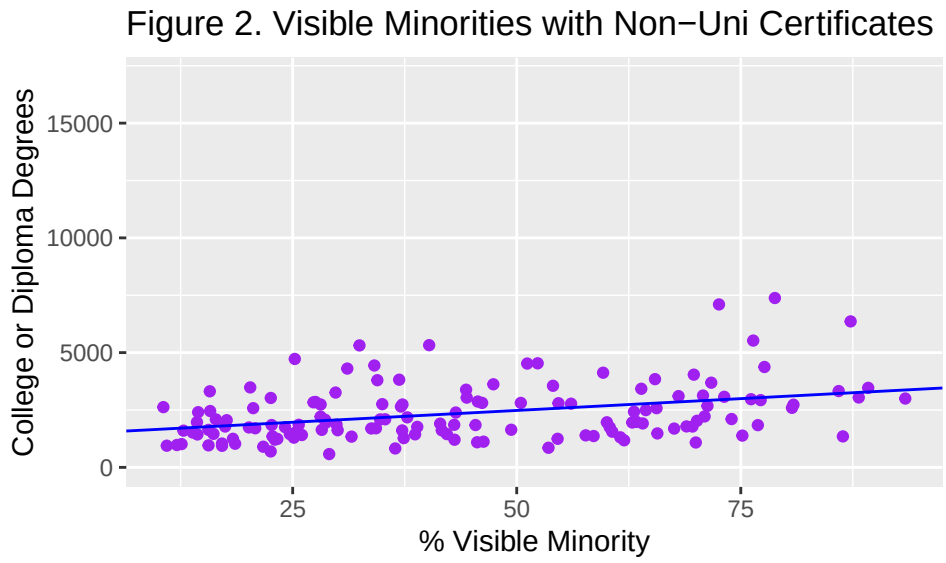


Figure 3

```
bach<-lm(Bachelors~Vis_Min_Per, data=edu_inc_by_min)
summary(bach)
```

```
##
## Call:
## lm(formula = Bachelors ~ Vis_Min_Per, data = edu_inc_by_min)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2815.4 -1387.8  -395.5   756.0 13169.3
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3302.765    411.579   8.025 3.96e-13 ***
## Vis_Min_Per   -4.153      8.352  -0.497     0.62
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2188 on 138 degrees of freedom
## Multiple R-squared:  0.001788,   Adjusted R-squared:  -0.005445
## F-statistic: 0.2472 on 1 and 138 DF,  p-value: 0.6198
```

```
ggplot(edu_inc_by_min, aes(x=Vis_Min_Per, y=Bachelors)) +
  geom_point(color = 'blue') +
  geom_abline(intercept= 3302.765, slope=-4.153, colour="red") +
  ylim(0,17000) +
  labs(title = "Figure 3. Visible Minorities with Bachelor's Degree",
       x = "% Visible Minority",
       y = "Bachelor's Degree")
```

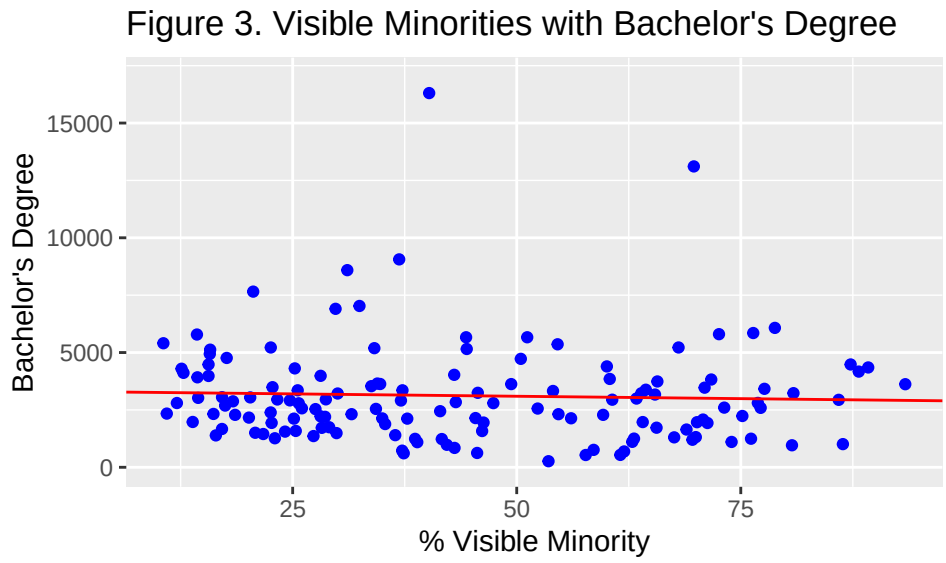


Figure 4

```
grad<-lm(Above_Bachelors~Vis_Min_Per, data=edu_inc_by_min)
summary(grad)
```

```
##
## Call:
## lm(formula = Above_Bachelors ~ Vis_Min_Per, data = edu_inc_by_min)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1859.4  -973.8  -278.9   515.9  6962.4
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2432.406    275.726   8.822 4.43e-15 ***
## Vis_Min_Per   -9.887      5.595  -1.767   0.0794 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1466 on 138 degrees of freedom
## Multiple R-squared:  0.02213,    Adjusted R-squared:  0.01504
## F-statistic: 3.123 on 1 and 138 DF,  p-value: 0.07942
```

```
ggplot(edu_inc_by_min, aes(x=Vis_Min_Per, y=Above_Bachelors)) +
  geom_point(color = 'blue') +
  geom_abline(intercept=2432.406, slope=-9.887, colour="red") +
  ylim(0,17000) +
  labs(title = "Figure 4. Visible Minorities with Graduate Degrees",
       x = "% Visible Minority",
       y = "Graduate Degrees")
```

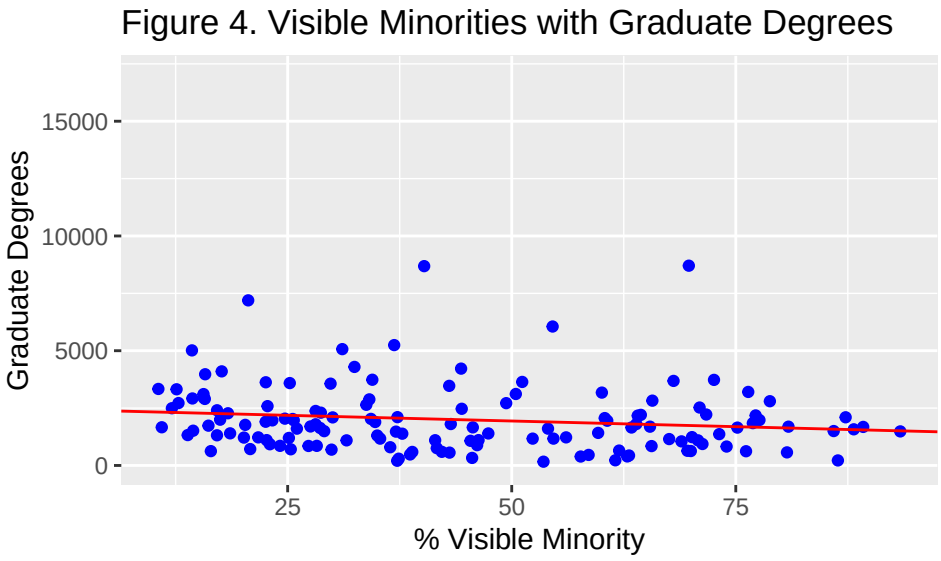
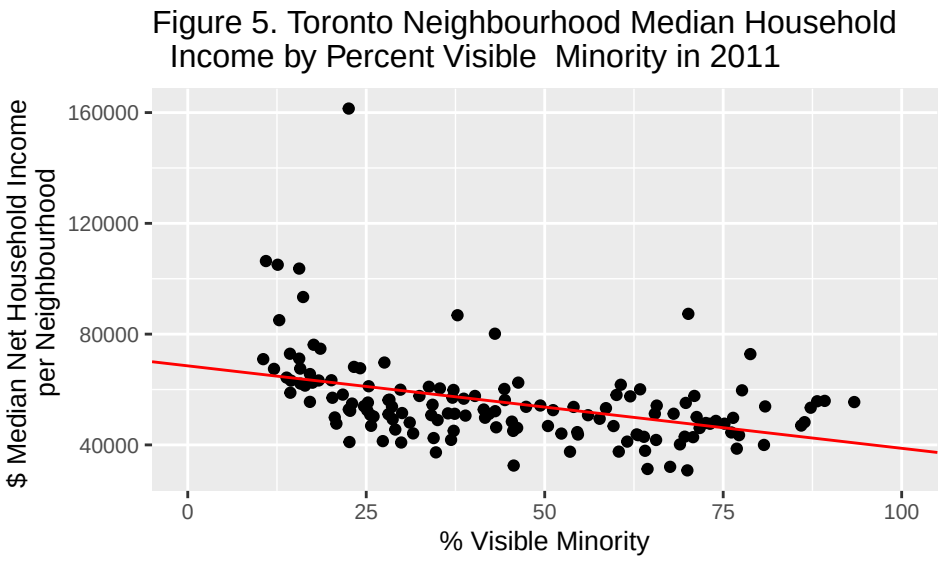


Figure 5

```
medi_net_income<-as.numeric(demo_nhs_dataset$`Median after-tax household income $`)
percent_minority<-((as.numeric(demo_nhs_dataset$`Total visible minority population`))/
                     (as.numeric(demo_nhs_dataset$`Total Population`)))*100
```

```
ggplot(demo_nhs_dataset, aes(x=percent_minority, y=medi_net_income)) +
  geom_point(color = 'black') +
  geom_abline(intercept=68538.80 , slope=-297.83, colour="red") +
  xlim(0,100) +
  ylim(30000,162000) +
  labs(title = "Figure 5. Toronto Neighbourhood Median Household
  Income by Percent Visible  Minority in 2011", x = "% Visible Minority",
  y = "$ Median Net Household Income
```

```
  per Neighbourhood") +
  theme(text=element_text(size=10))
```

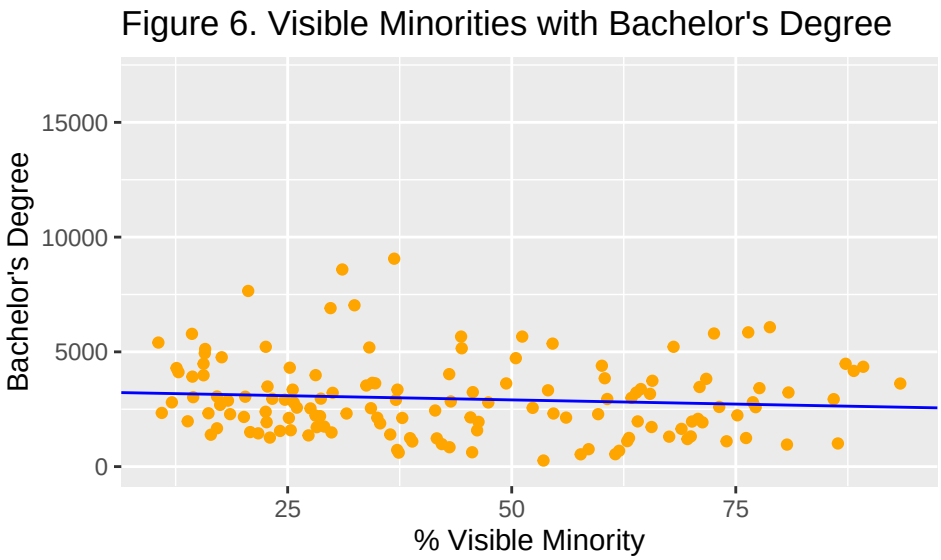


Remove outliers and plot Visible Minorities with Bachelor's Degree

```
clean_edu <- subset(edu_inc_by_min, `Bachelors` < 13000)

clean_vis_min <- clean_edu$Vis_Min_Per
clean_bachelors <- clean_edu$Bachelors

clean_lmbach<-lm(clean_bachelors~clean_vis_min, data=clean_edu)
summary(clean_lmbach)
```

```
##
## Call:
## lm(formula = clean_bachelors ~ clean_vis_min, data = clean_edu)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2616  -1285   -279    894   6057
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3272.062    315.491  10.371   <2e-16 ***
## clean_vis_min   -7.296      6.418  -1.137    0.258
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1673 on 136 degrees of freedom
## Multiple R-squared:  0.009414,   Adjusted R-squared:  0.002131
## F-statistic: 1.293 on 1 and 136 DF,  p-value: 0.2576
```

```r
ggplot(clean_edu, aes(x=clean_vis_min, y=clean_bachelors)) +
  geom_point(color = "orange") +
  geom_abline(intercept= 3272.062, slope=-7.296, colour="blue") +
  ylim(0,17000) +
  labs(title = "Figure 6. Visible Minorities with Bachelor's Degree",
       x = "% Visible Minority",
       y = "Bachelor's Degree")
```



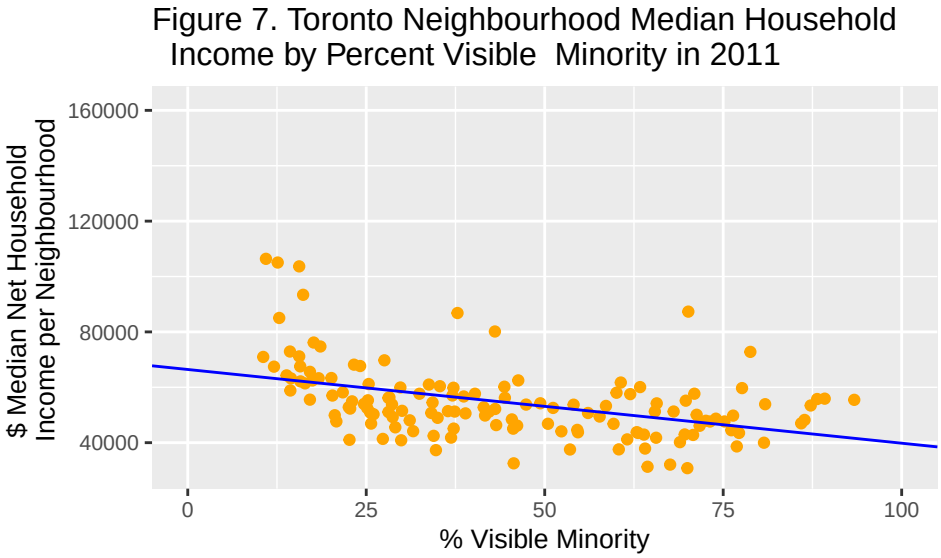
Remove outliers and plot Toronto Neighbourhood Median Household Income by Percent Visible Minority in 2011

```r
test2 <- subset(demo_nhs_dataset, medi_net_income < 160000)

test_medi_net_income<-as.numeric(test2$`Median after-tax household income $`)
test_percent_minority<-((as.numeric(test2$`Total visible minority population`))/
                          (as.numeric(test2$`Total Population`)))*100

test_reg_incdiv <- lm(test_medi_net_income~test_percent_minority)
summary(test_reg_incdiv)
```

```
##
## Call:
## lm(formula = test_medi_net_income ~ test_percent_minority)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21731  -7297  -1641   3972  42877
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           66424.90    2285.00  29.070  < 2e-16 ***
```

```
## test_percent_minority  -266.21      46.24  -5.757 5.37e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12070 on 137 degrees of freedom
## Multiple R-squared:  0.1948, Adjusted R-squared:  0.1889
## F-statistic: 33.15 on 1 and 137 DF,  p-value: 5.368e-08
```

```
ggplot(test2, aes(x=test_percent_minority, y=test_medi_net_income)) +
  geom_point(color = 'orange') +
  geom_abline(intercept=66424.90, slope=-266.21, colour="blue") +
  xlim(0,100) +
  ylim(30000,162000) +
  labs(title = "Figure 7. Toronto Neighbourhood Median Household
  Income by Percent Visible  Minority in 2011",
  x = "% Visible Minority", y = "$ Median Net Household
  Income per Neighbourhood") +
  theme(text=element_text(size=10))
```

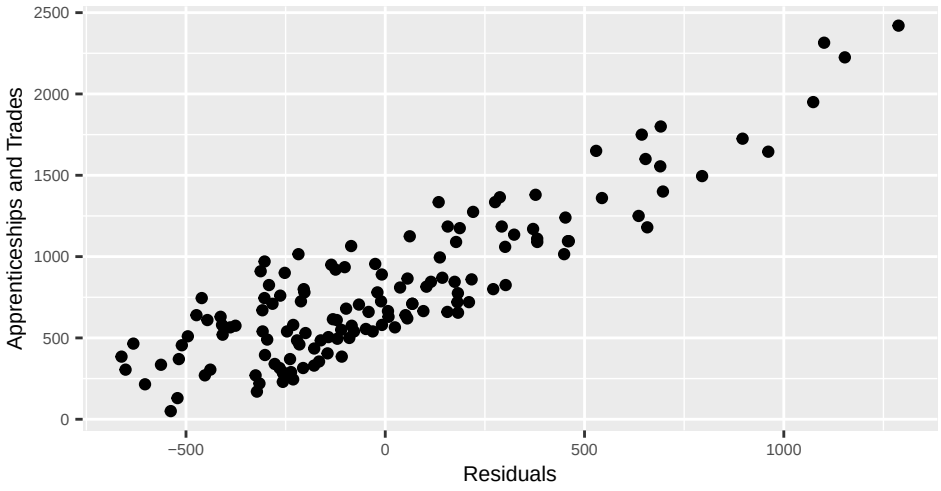


Appendix 3. Assumptions of Linear Regression

Checking for linearity and equal variance assumptions in education levels and median income

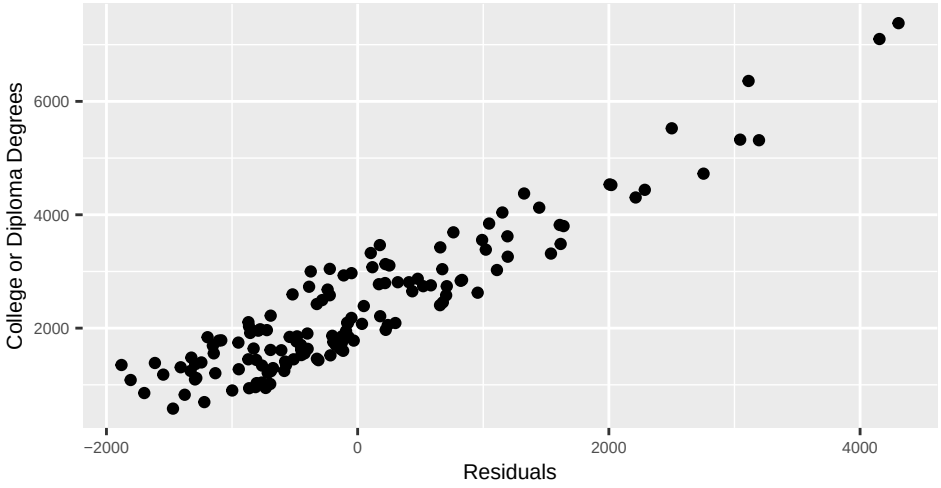
```
ggplot(data = edu_inc_by_min) +
  geom_point(mapping = aes(x = residuals.lm(lmapp), y = edu_inc_by_min$Appren_or_Trade)) +
  labs(title = "Figure 8. Checking for linearity and equal variance assumptions
       for apprenticeships and trades",
       x = "Residuals",
       y = "Apprenticeships and Trades") +
  theme(text=element_text(size=8))
```

Figure 8. Checking for linearity and equal variance assumptions for apprenticeships and trades



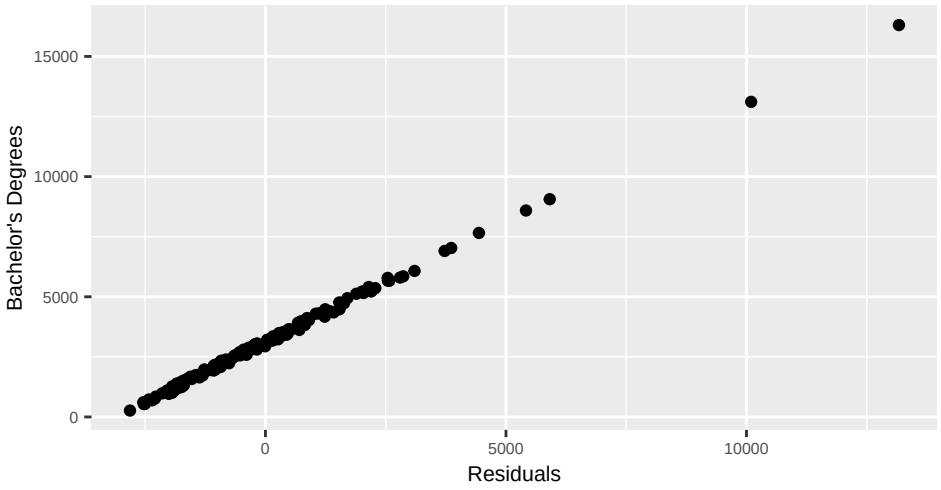

```
ggplot(data = edu_inc_by_min) +
  geom_point(mapping = aes(x = residuals.lm(nonuni), y = edu_inc_by_min$Non_Uni_Cert)) +
  labs(title = "Figure 9. Checking for linearity and equal variance assumptions
       for college or diploma degrees",
       x = "Residuals",
       y = "College or Diploma Degrees") +
  theme(text=element_text(size=8))
```

Figure 9. Checking for linearity and equal variance assumptions
for college or diploma degrees



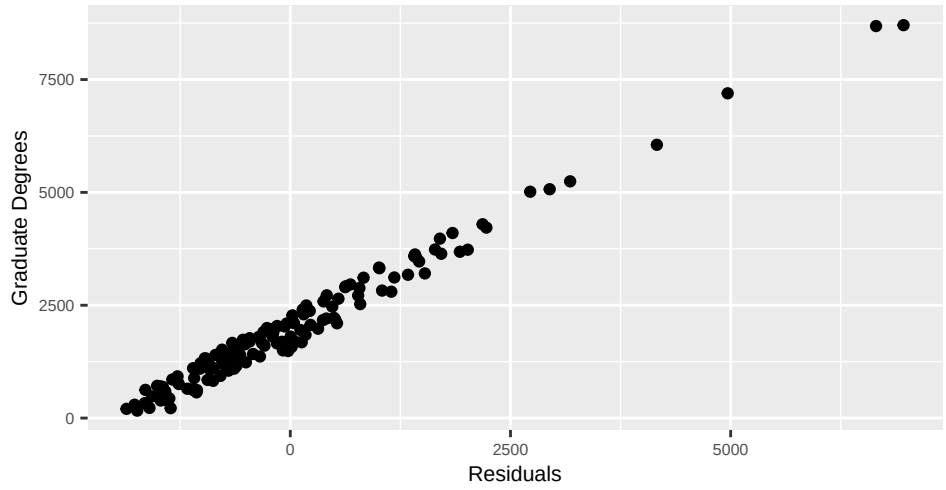

```
ggplot(data = edu_inc_by_min) +
  geom_point(mapping = aes(x = residuals.lm(bach), y = edu_inc_by_min$Bachelors)) +
  labs(title = "Figure 10. Checking for linearity and equal variance assumptions
       for bachelor's degrees",
       x = "Residuals",
       y = "Bachelor's Degrees") +
  theme(text=element_text(size=8))
```

Figure 10. Checking for linearity and equal variance assumptions
for bachelor's degrees



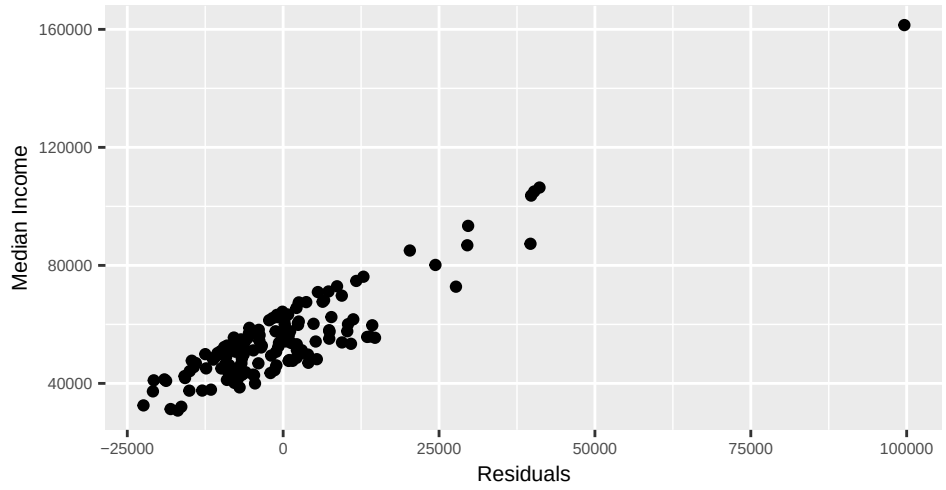

```
ggplot(data = edu_inc_by_min) +
  geom_point(mapping = aes(x = residuals.lm(grad), y = edu_inc_by_min$Above_Bachelors)) +
  labs(title = "Figure 11. Checking for linearity and equal variance assumptions
       for graduate degrees",
       x = "Residuals",
       y = "Graduate Degrees") +
  theme(text=element_text(size=8))
```

Figure 11. Checking for linearity and equal variance assumptions for graduate degrees




```
ggplot(data = demo_nhs_dataset) +
  geom_point(mapping = aes(x = residuals.lm(reg_incdiv), y = medi_net_income)) +
  labs(title = "Figure 12. Checking for linearity and equal variance assumptions
       for median income",
       x = "Residuals",
       y = "Median Income") +
  theme(text=element_text(size=8))
```

Figure 12. Checking for linearity and equal variance assumptions for median income

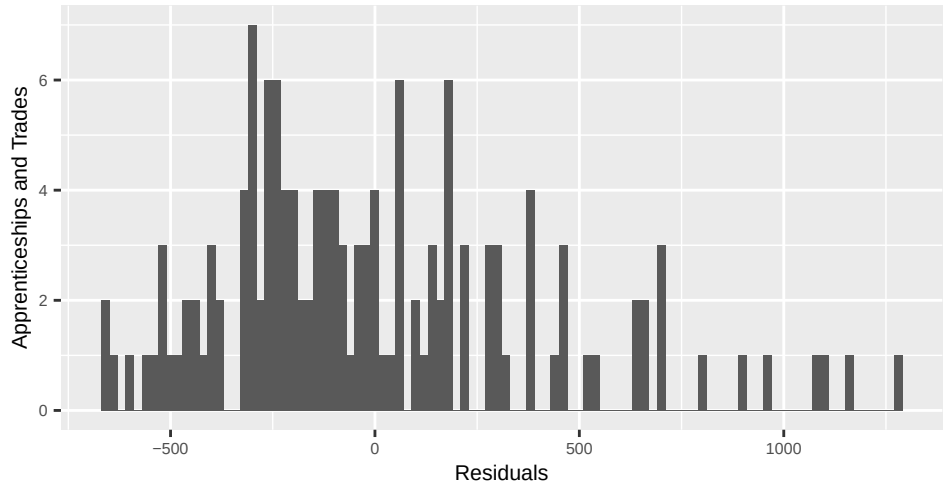


Checking for normality for education levels and median income

```
ggplot(data = edu_inc_by_min) +
 geom_histogram(binwidth = 20, mapping = aes(x = residuals(lmapp))) +
  labs(title = "Figure 13. Checking for normality assumptions
```

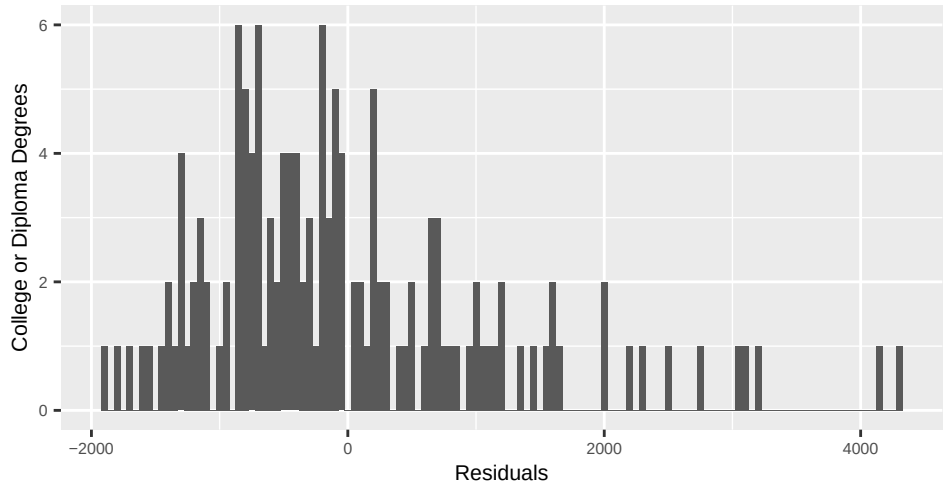
```
       for apprenticeships and trades",
       x = "Residuals",
       y = "Apprenticeships and Trades") +
  theme(text=element_text(size=8))
```

Figure 13. Checking for normality assumptions for apprenticeships and trades



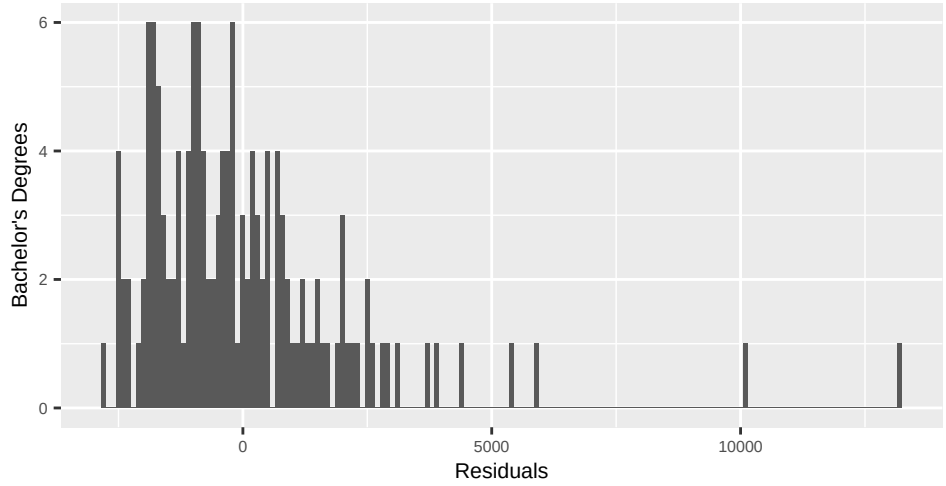

```
ggplot(data = edu_inc_by_min) +
 geom_histogram(binwidth = 50, mapping = aes(x = residuals(nonuni))) +
  labs(title = "Figure 14. Checking for normality assumptions
       for college or diploma degrees",
       x = "Residuals",
       y = "College or Diploma Degrees") +
  theme(text=element_text(size=8))
```

Figure 14. Checking for normality assumptions for college or diploma degrees



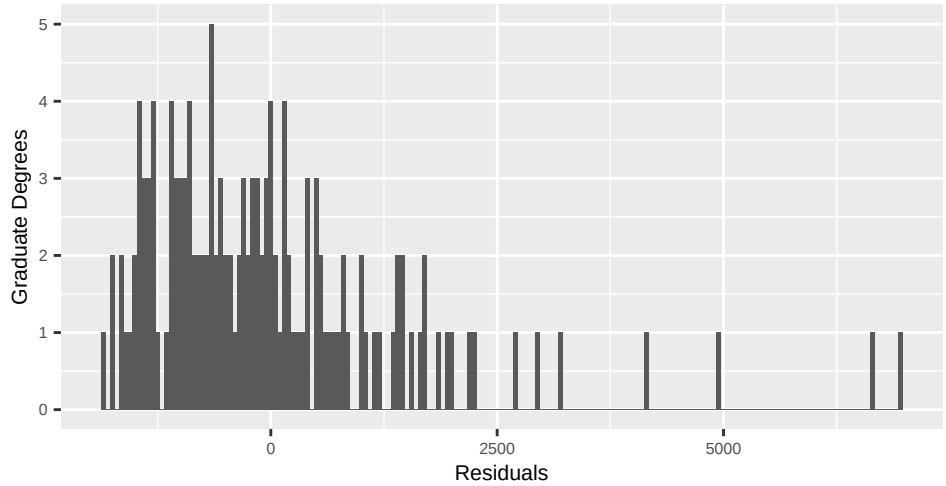
```
ggplot(data = edu_inc_by_min) +
 geom_histogram(binwidth = 100, mapping = aes(x = residuals(bach))) +
  labs(title = "Figure 15. Checking for normality assumptions
       for bachelor's degrees",
       x = "Residuals",
       y = "Bachelor's Degrees") +
  theme(text=element_text(size=8))
```

Figure 15. Checking for normality assumptions
for bachelor's degrees




```
ggplot(data = edu_inc_by_min) +
 geom_histogram(binwidth = 50, mapping = aes(x = residuals(grad))) +
  labs(title = "Figure 16. Checking for normality assumptions
       for graduate degrees",
       x = "Residuals",
       y = "Graduate Degrees") +
  theme(text=element_text(size=8))
```

Figure 16. Checking for normality assumptions
for graduate degrees



```
ggplot(data = edu_inc_by_min) +
 geom_histogram(binwidth = 500, mapping = aes(x = residuals(reg_incdiv))) +
  labs(title = "Figure 17. Checking for normality assumptions
       for median income",
       x = "Residuals",
       y = "Median Income") +
  theme(text=element_text(size=8))
```

Figure 17. Checking for normality assumptions
for median income