ProblemSet2_new

Ke-li Chiu & Diego Mamanche Castellanos

06/02/2020

Abstract

Education is a contributor to many beneficial socio-economic outcomes. We intend to exam the the relationship between academic majors and household income in Toronto's neighbourhoods—can we find higher or lower percentage population in certain academic majors in higher or lower income neighbourhoods? The result shows that higher income neighbourhoods also have higher percentage of residents with majors in business, The data set is retrieved from the package *City of Toronto Neighbourhood Profiles* from the year 2011. ###### We intend to use the comparison to bring awareness of inequality in income and education existing in the City of Toronto.

Introduction

Education is a contributor to many beneficial socio-economic outcomes. In this study, we intend to examine the relationship between household income and educational profile in terms of residents' academic majors in Toronto's 140 neighbourhoods. The majors are categorized as follows— "education", "visual and performing arts and communications technologies," "humanities," "social and behavioural sciences and law," "business_management_and_public_administration," "physical and life sciences and technologies," "mathematics computer and information sciences," "architecture engineering and related technologies," "agriculture natural resources and conservation," "personal protective and transportation services" and "no postsecondary certificate diploma or degree." The question we want to answer is if we can find a higher population percentage in specific academic majors in higher-income neighbourhoods or vice versa? The question is relevant because the academic major is a significant factor for people's occupations and their associated earning; thus, it is deemed as an essential tool for social mobility.

Dataset and Method Description

City of Toronto Neighbourhood Profiles data set is sourced from several Census tables released by Statistics Canada every five years. The dataset uses this Census data to provide a portrait of the demographic, social and economic characteristics of the people and households in each City of Toronto neighbourhood. Each data point in this file is presented for the City's 140 neighbourhoods, as well as for the City of Toronto as a whole. The data set consists of 144 columns for neighbourhoods and 1537 rows for characteristics of the neighbourhoods. Data cleaning and reshaping are performed to have a new data frame that has 14 characteristics regarding topics of income as columns and education and the 140 neighbourhoods as rows.

Exploratory data analysis is conducted to obtain insight from the dataset. To get a picture of a neighbourhood's economic status, we look at the median total household income which divides the income distribution into two equal groups, half having income above that amount, and half having income below that amount. As for the education profile, the academic majors of the residents are the variables in which we are interested. Each field for the major is the number of people who took this major in the neighbourhood. To be able to compare the neighbourhood regardless of the difference in population, we turn the value to percentages by dividing the total population between 25 to 64 years old by the population in each major. All codes are available in Appendix A.

Limitation of the Dataset and Approach

Because we chose to look at the median instead of the mean of household income, we are excluding the outliers in the income distribution. The median value would give us a more accurate picture of the neighbourhoods' economic status but also prohibit us from investigating further the outliers that are skewing the distributions. Moreover, although there are datasets for 2001,2006, 2011 and 2016, the structure and available information of the datasets are inconsistent. For example, median total household income for all neighbourhoods is only available in the dataset from 2011; therefore, this study is limited to the year 2011 and is unable to provide a comparison between Census years regarding the changes in income and education profile of the neighbourhoods. Finally, we are also not able to find out the income distribution by academic fields. Therefore, earnings differences linking to differences in residents' fields of study are not included in our analysis. This prevents from knowing if specific academic majors are more significant contributors to the neighbourhoods' total household income.

Ethical Considerations

Population percentage in specific majors have correlations with median household income in the 140 neighbourhoods. For example, higher income neighbourhoods have more percentage of business majors, and lower income neighbourhoods have more percentage of transportation majors. This observation, without further anaylysis, potentially projects social status and ranking among academic fields.

Income Distribution Map

Darker areas in the above map indicate high levels of unemployment, while lighter areas indicate low levels of unemployment. As Figure 1 shows, we observe that higher-income neighbourhoods tend to be in the central part of Toronto, neighbouring each other.

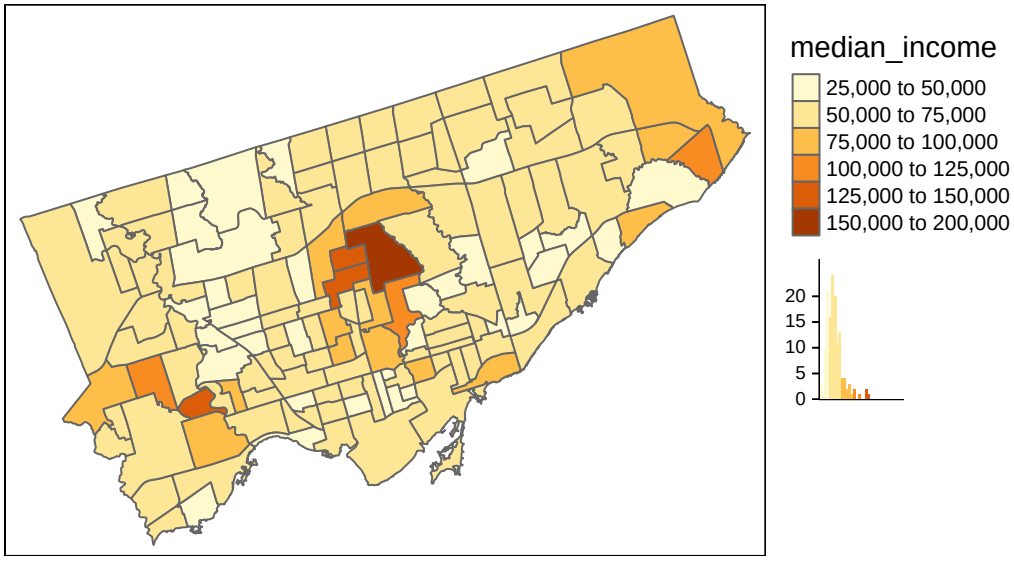

Figure 1: Median household income map

Correlation Between Household Income and Academic Majors

We conducted a correlation matrix table to see if certain majors correlate with household income. As a result, shown in Table 1, "business management and public administration", "education", "social and behavioural sciences and law", have stronger positive correlations with household income. In contrast, "no certificate" has a strong negative correlation.

major	correlation
no_certificate	-0.49
transportation	-0.46
architecture_engin	-0.09
agriculture	-0.05
mathematics_&_it	-0.01
arts_&_com	0.06
physical_&_life	0.26
health	0.36
humanities	0.44
social_&_law	0.46
education	0.54
business	0.55

Table 1: Correlation score of income and academic majors

Majors Distribution Maps

In Figure 2, the four categories that have the most significant correlation scores are plotted into choropleth maps to be compared with Figure 1. The distribution of "business management and public administration," "education," "social and behavioural sciences and law," have a similar pattern with Figure 1, where darker colours tend to be around the central part of Toronto. The distribution of "no certificate" is contrasting with Figure 1, where the lighter area locates in the central part of Toronto.

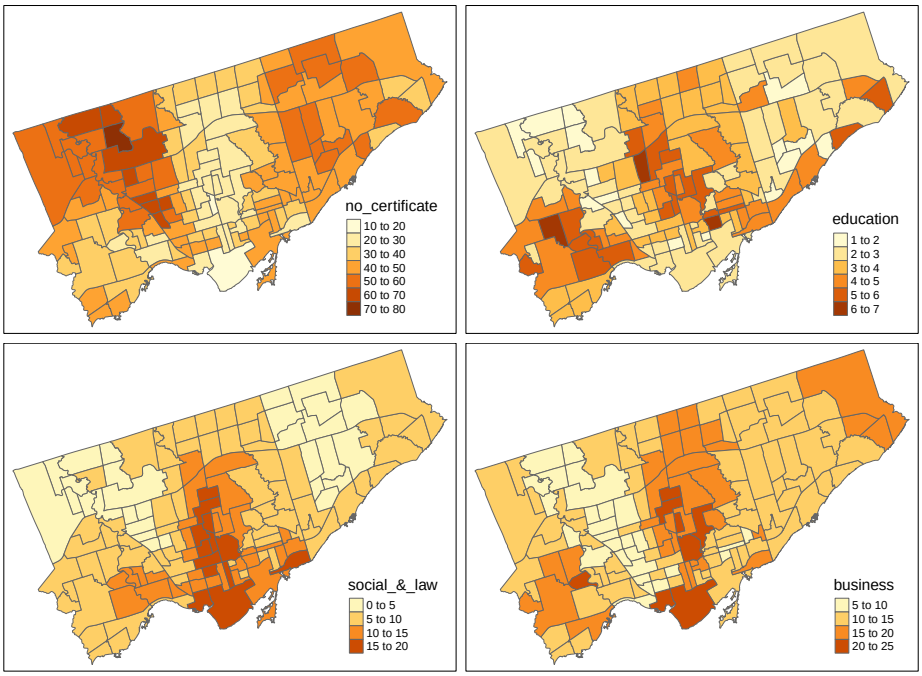

Figure 2: Distribution maps for selected majors

Comparison Between Polarized Neighbourhoods and the City of Toronto

Finally, we gathered the population percentage in different majors for the neighbourhoods that have the highest income (Bridle Path Sunnybrook York) and the lowest income (Regent Park) and compared them with the City of Toronto. As Figure 3 shows, the shapes of the academic major distributions in the three groups are similar to each other; most population percentage lies in "no certification," and the second-largest portion lies in "business management and public administration." However, the scales of population percentages in different majors have apparent differences in the three groups. Bridle Path Sunnybrook York, the percentage of the population who have no certificate is 5.47 % higher than than the percentage of the population who major in "business management and public administration". In Regent Park, the percentage of the population who have no certificate is drastically higher than the percentage of the population who major in "business management and public administration" by 37.31%.

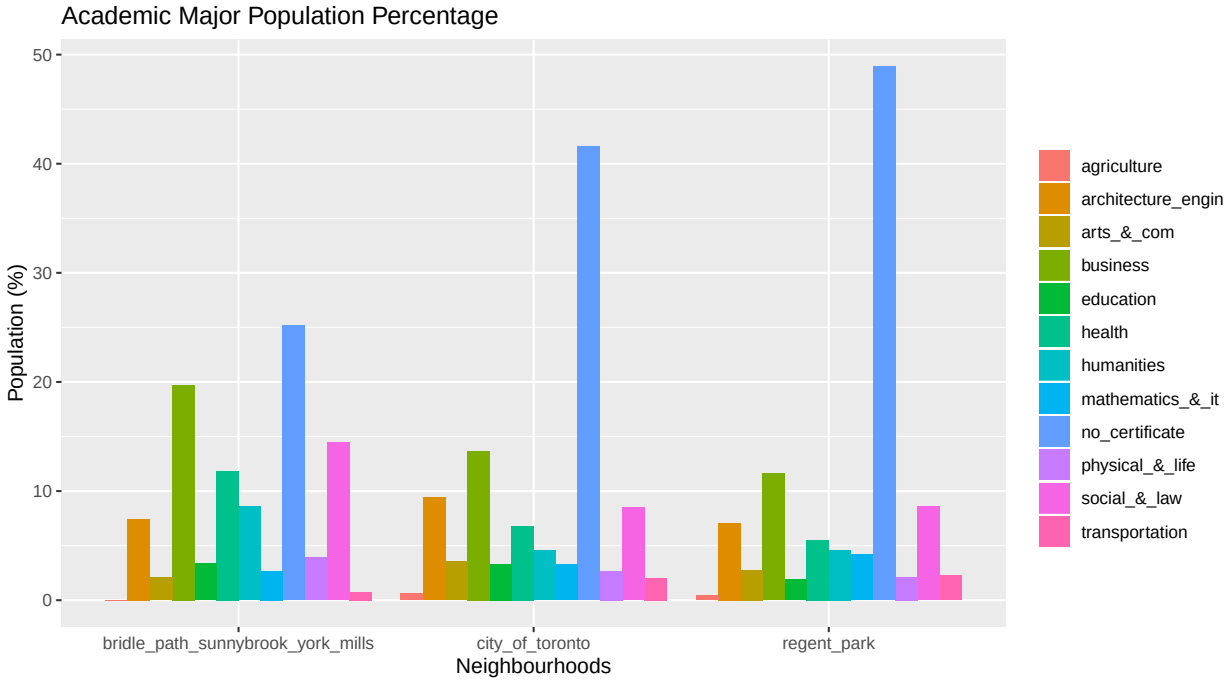

Figure 3: Population percentage in each academic major in different neighbourhoods

Appendix A

```r
knitr::opts_chunk$set(echo = TRUE)
#Set up the environment
library(opendatatoronto)
library(dplyr)
library(tidyr)
library(tidyverse)
library(ggplot2)
library(knitr)
library(data.table) #Useful to transponse a dataframe
library(sf)
library(tmap)
library(tmaptools)
library(leaflet)
#Set working Directory
#setwd("~/Experimental Design for Data Science/ProblemSet2")
# Search packages (this returns a table)
neighbourhood_packages <- search_packages("Neighbourhood")
# Filter the neighbourhood demographic dataset
neighbourhood_demographics_package<- neighbourhood_packages %>%
  filter(title == "Neighbourhood Profiles")
#Create main dataset for year 2011
main_df_2011 <-  neighbourhood_demographics_package %>% # Start with the package
  list_package_resources() %>% # List the resources in the package
  filter(name == "Neighbourhood Data 2001, 2006, 2011.xlsx") %>% # Only keep the resource we want
  get_resource()
### GEO-LOCATION DATASET TORONTO BY NEIGHBOURHOODS ###
# get package
package <- show_package("4def3f65-2a65-4a4f-83c4-b2a4aed72d46")
package
# get all resources for this package
resources <- list_package_resources("4def3f65-2a65-4a4f-83c4-b2a4aed72d46")
# identify datastore resources; by default, Toronto Open Data sets datastore resource format to CSV for
datastore_resources <- filter(resources, tolower(format) %in% c('csv', 'geojson'))
# load the first datastore resource as a sample
geo_data <- filter(datastore_resources, row_number()==1) %>% get_resource()
##### Distribution by education 2011####
#Select sheet 2011
clean_header_df_2011 <- main_df_2011$`2011`
#Clean header using janitor
clean_header_df_2011 <- janitor::clean_names(clean_header_df_2011)
#Select only Education by major from the main dataset
education_df_2011 <- filter(clean_header_df_2011, category == "Education" & topic == "Major field of stu
# Remove row with the totals
education_df_2011 <-
  education_df_2011 %>%
    filter(!str_detect(attribute, "Total"))
#Remove first two columns
education_df_2011 <- education_df_2011[,3:length(education_df_2011)]
#Change class of columns containing numbers as.numeric
education_df_2011[,2:length(education_df_2011)] <- sapply(education_df_2011[,2:length(education_df_2011
#Create new dataframe
```

```r
scaled_education_df_2011 <- education_df_2011
#Scale dataframe by column
scaled_education_df_2011 <-
  scaled_education_df_2011 %>%
    janitor::adorn_percentages(denominator = "col")
#Multiply by 100
scaled_education_df_2011[,2:length(education_df_2011)] <- sapply(scaled_education_df_2011[,2:length(scal
#Round new scales to 2 digits
scaled_education_df_2011 <-
  scaled_education_df_2011 %>%
    janitor::adorn_rounding(digits = 2)
#Transform the result to dataframe
scaled_education_df_2011 <- as.data.frame(scaled_education_df_2011)
#Save neighbourhoods into a dataframe
col_names_2011 <- as.data.frame(colnames(scaled_education_df_2011))
col_names_2011 <- col_names_2011[2:nrow(col_names_2011),]
col_names_2011 <- as.data.frame(col_names_2011)
colnames(col_names_2011) <- "neighbourhoods"
col_names_2011
#Reshape dataframe
edu_df_2011_reshaped <- transpose(scaled_education_df_2011, make.names=1)
#Include neighbourhoods column
edu_df_2011_reshaped <- mutate(edu_df_2011_reshaped, neighbourhoods = col_names_2011$neighbourhoods)
#Clean header names
edu_df_2011_reshaped <- janitor::clean_names(edu_df_2011_reshaped)
#Reorder columns
edu_df_2011_reshaped <- select(edu_df_2011_reshaped, neighbourhoods, no_postsecondary_certificate_diplo
##### Distribution by income 2011####
#Select sheet 2011
clean_income_df_2011 <- main_df_2011$`2011`
#Clean header using janitor
clean_income_df_2011 <- janitor::clean_names(clean_income_df_2011)
#Select only median income from the main dataset
income_df_2011 <- clean_income_df_2011[756,]
#Remove first two columns
income_df_2011 <- income_df_2011[,3:length(income_df_2011)]
#Change class of columns containing numbers as.numeric
income_df_2011[,2:length(income_df_2011)] <- sapply(income_df_2011[,2:length(income_df_2011)], function
#Create new dataframe
col_names_2011_income  <- as.data.frame(colnames(income_df_2011))
#Save neighbourhoods into a dataframe
col_names_2011_income <- col_names_2011_income[2:nrow(col_names_2011_income),]
col_names_2011_income <- as.data.frame(col_names_2011_income)
colnames(col_names_2011_income) <- "neighbourhoods"
#Reshape dataframe
inco_df_2011_reshaped <- transpose(income_df_2011, make.names=1)
#Include neighbourhoods column
inco_df_2011_reshaped <- mutate(inco_df_2011_reshaped, neighbourhoods = col_names_2011_income$neighbourh
#Clean header names
inco_df_2011_reshaped <- janitor::clean_names(inco_df_2011_reshaped)
#Reorder columns
inco_df_2011_reshaped <- select(inco_df_2011_reshaped, neighbourhoods, median_household_total_income)
### Cleaning geo-location dataset
```

```r
clean_geo_data <- janitor::clean_names(geo_data)
clean_geo_data <- extract(clean_geo_data, area_name, into = "neighbourhoods" , regex = "([^(0-9)]+)")
clean_geo_data["neighbourhoods"] <-
  janitor::make_clean_names(as.matrix(clean_geo_data["neighbourhoods"]))
clean_geo_data <- janitor::clean_names(clean_geo_data)
clean_geo_data <- select(clean_geo_data, neighbourhoods, longitude, latitude, geometry)
filter(clean_geo_data, neighbourhoods %in% c("mimico","weston_pellam_park"))
clean_geo_data["neighbourhoods"][c(17,67),] <- c("mimico_includes_humber_bay_shores","weston_pelham_park")
clean_geo_data
#Merge income and education 2011
merged_df <- merge(edu_df_2011_reshaped, inco_df_2011_reshaped, by = 'neighbourhoods')
merged_df <- merge(merged_df, clean_geo_data, by = 'neighbourhoods', all.x = TRUE)
#Create a sf version
sf_merged_df <- st_sf(merged_df, sf_column_name = "geometry")
colnames(sf_merged_df) <- c("neighbourhoods" ,
"no_certificate" ,
"education" ,
"arts_&_com",
"humanities" ,
"social_&_law" ,
"business" ,
"physical_&_life" ,
"mathematics_&_it" ,
"architecture_engin" ,
"agriculture" ,
"health" ,
"transportation" ,
"other" ,
"median_income",
"longitude",
"latitude",                                                                    "geometry")
colnames(merged_df)
### Explore relationship between income and every major percentage ###
#Create a new dataframe and change column names. Remove last three columns
education_percentage_only <- merged_df[,1:(length(merged_df)-3)]
colnames(education_percentage_only) <- c("neighbourhoods" ,
"no_certificate" ,
"education" ,
"arts_&_com",
"humanities" ,
"social_&_law" ,
"business" ,
"physical_&_life" ,
"mathematics_&_it" ,
"architecture_engin" ,
"agriculture" ,
"health" ,
"transportation" ,
"other" ,
"median_income")
#Plot Median Income
tmap_mode("plot")
tm_shape(sf_merged_df) +
```

```
  tm_polygons(col = "median_income",
              breaks = c(25000,50000,75000,100000,125000,150000,200000),
              #style = "quantile",
              legend.hist = TRUE,
              palette = "seq") +
  tm_layout(legend.outside = TRUE)
### Correlation table
correlation_df <- as.data.frame(cor(education_percentage_only[,2:(length(education_percentage_only)-2)]
correlation_df <- round(correlation_df, digits = 2)
colnames(correlation_df) <- "correlation"
correlation_df <- mutate(correlation_df, major = rownames(correlation_df))
#Reorder columns
correlation_df <- select(correlation_df, major, correlation)
correlation_df <- correlation_df[order(correlation_df$correlation),]
row.names(correlation_df)<-1:nrow(correlation_df)
kable(correlation_df)
#Plot all majors without palette
tmap_mode("plot")
tm_shape(sf_merged_df) +
tm_layout(legend.show = TRUE, legend.position = c("right", "bottom"), title.size = 2, title.position = c
  tm_polygons(c("no_certificate" ,
"education" ,
"social_&_law" ,
"business")) +
  tm_facets(sync = TRUE, ncol = 2)
data_plot_group <-
  education_percentage_only %>%
  filter(neighbourhoods == "bridle_path_sunnybrook_york_mills" | neighbourhoods == "regent_park" |neighb
  pivot_longer(cols = "no_certificate":"transportation", names_to = "major")
ggplot(data_plot_group, aes(x = neighbourhoods, y = value, fill = major)) +
  # Specify that we want a bar graph:
  geom_bar(stat="identity", position=position_dodge()) +
  # Add graph title and axis labels:
  labs(title = "Academic Major Population Percentage", x = "Neighbourhoods",
       y = "Population (%)") +
  # Rename the legend labels to be more readable
  scale_fill_discrete(name = "")
```

References

Open Data Dataset. (n.d.). Retrieved January 30, 2020, from https://open.toronto.ca/dataset/ wellbeing-toronto-demographics-nhs-indicators/

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

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

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr

Hadley Wickham and Lionel Henry (2019). tidyr: Tidy Messy Data. R package version 1.0.0. https://CRAN.R-project.org/package=tidyr

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Yihui Xie (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.25.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible

Matt Dowle and Arun Srinivasan (2019). data.table: Extension of `data.frame`. R package version 1.12.6. https://CRAN.R-project.org/package=data.table Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009

Tennekes M (2018). "tmap: Thematic Maps in R." *Journal of Statistical Software*, *84*(6), 1-39. doi: 10.18637/jss.v084.i06 (URL: https://doi.org/10.18637/jss.v084.i06).

Martijn Tennekes (2019). tmaptools: Thematic Map Tools. R package version 2.0-2. https://CRAN.R-project.org/package=tmaptools

Joe Cheng, Bhaskar Karambelkar and Yihui Xie (2019). leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.3. https://CRAN.R-project.org/package=leaflet