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— "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 descriptiion

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.

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.

Correlation between income and academic majors

We conducted a correlation matrix table to see if certain majors have correlation with household income. As a result, we find that business, education, and [sss] have stronger positive correlations, which no certificate has strong negative correlation.

Majors distribution map

The four categories that have the largest correlation scores are plotted into choropleth maps to be compared with Figure 1. The distribution of business, education and asfd have the similar pattern with

Comparison between polarized neighbourhoods and the City of Toronto

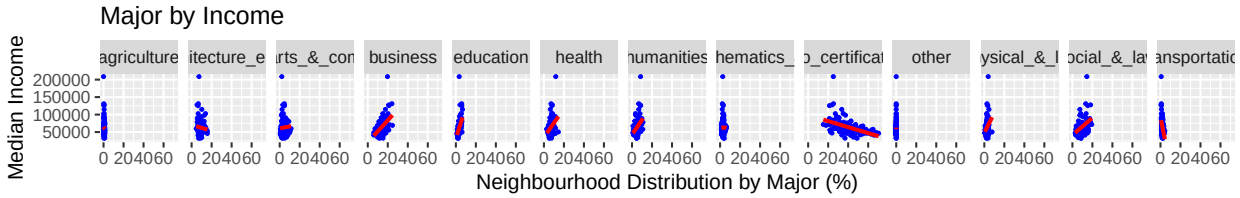
```
### Cleaning geo-location dataset
clean_geo_data <- janitor::clean_names(geo_data)
clean_geo_data <- extract(clean_geo_data, area_name, into = "neighbourhoods" , regex = "([^(0-9)]+)")
clean_geo_data["neighbourhoods"] <-
  janitor::make_clean_names(as.matrix(clean_geo_data["neighbourhoods"]))
clean_geo_data <- 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"))
```

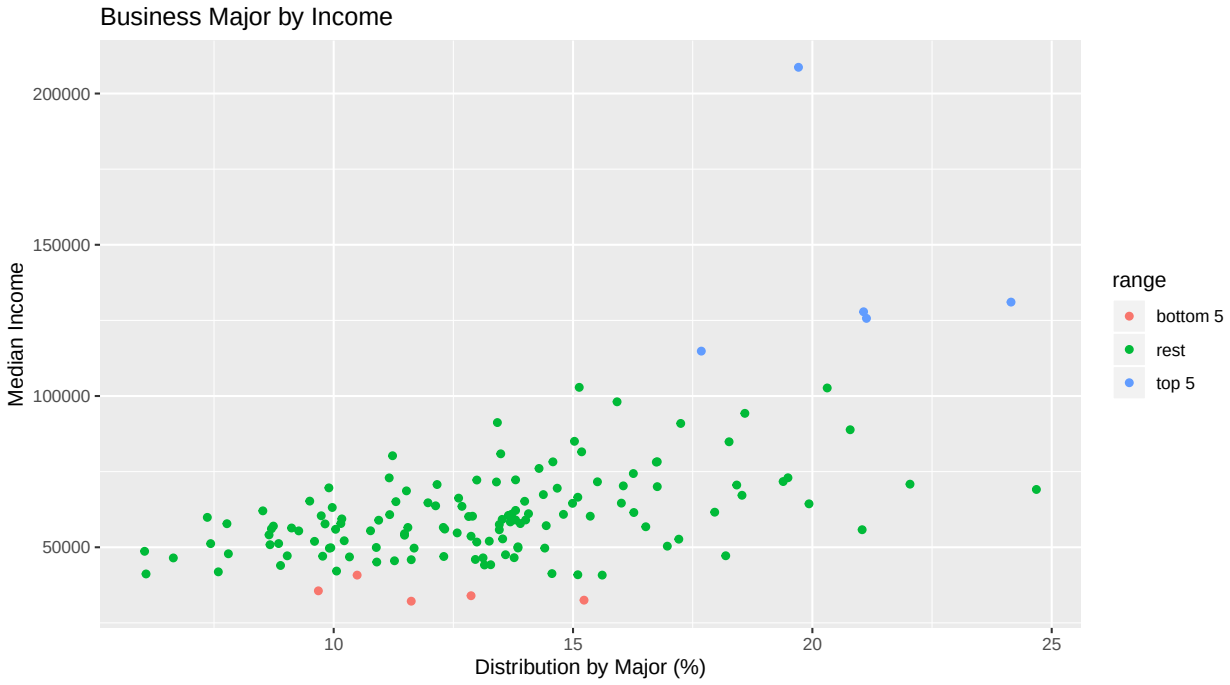
```
## # A tibble: 2 x 4
##   neighbourhoods   longitude latitude                                  geometry
##   <chr>                <dbl>    <dbl>                               <POLYGON [°]>
## 1 mimico               -79.5     43.6 ((-79.4804 43.62107, -79.48033 43.62107, ~
## 2 weston_pellam_p~     -79.5     43.7 ((-79.46005 43.66723, -79.46092 43.66811,~
```

```
clean_geo_data["neighbourhoods"][c(17,67),] <- c("mimico_includes_humber_bay_shores","weston_pelham_park
clean_geo_data
```

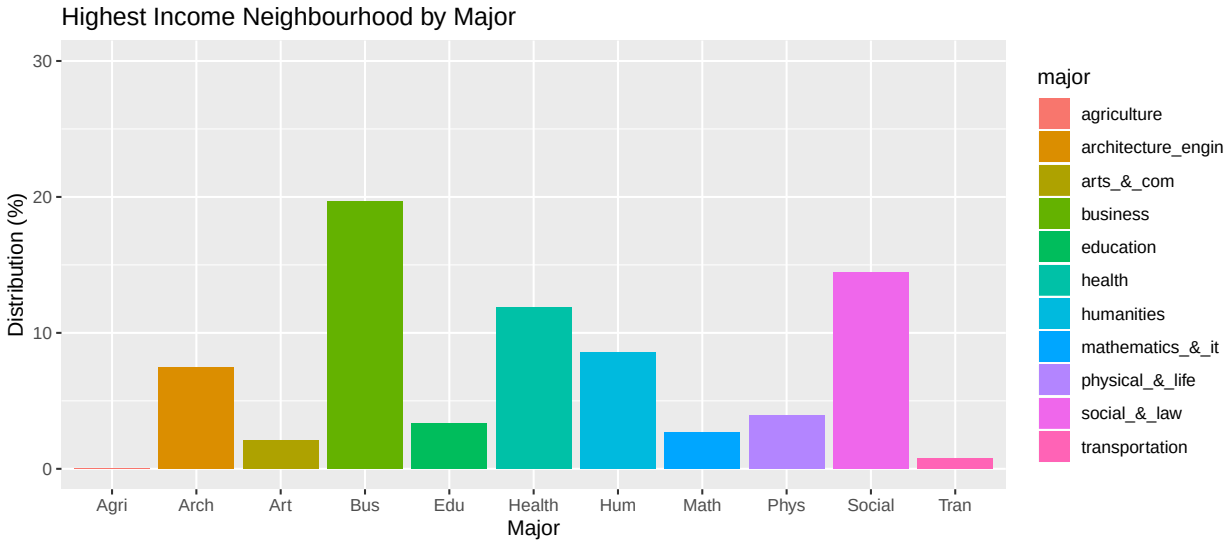
```
## # A tibble: 140 x 4
##    neighbourhoods    longitude latitude                                  geometry
##    <chr>                 <dbl>    <dbl>                               <POLYGON [°]>
##  1 wychwood              -79.4     43.7 ((-79.43592 43.68015, -79.43492 43.6803~
##  2 yonge_eglinton        -79.4     43.7 ((-79.41096 43.70408, -79.40962 43.7043~
##  3 yonge_st_clair        -79.4     43.7 ((-79.39119 43.68108, -79.39141 43.6809~
##  4 york_university_~     -79.5     43.8 ((-79.50529 43.75987, -79.50488 43.7599~
##  5 yorkdale_glen_pa~     -79.5     43.7 ((-79.43969 43.70561, -79.44011 43.7055~
##  6 lambton_baby_poi~     -79.5     43.7 ((-79.50552 43.66281, -79.50577 43.6629~
##  7 lansing_westgate      -79.4     43.8 ((-79.43998 43.76156, -79.44004 43.7617~
##  8 lawrence_park_no~     -79.4     43.7 ((-79.39008 43.72768, -79.39199 43.7272~
##  9 lawrence_park_so~     -79.4     43.7 ((-79.41096 43.70408, -79.41165 43.7039~
## 10 leaside_benningt~     -79.4     43.7 ((-79.37749 43.71309, -79.37762 43.7138~
## # ... with 130 more rows
```

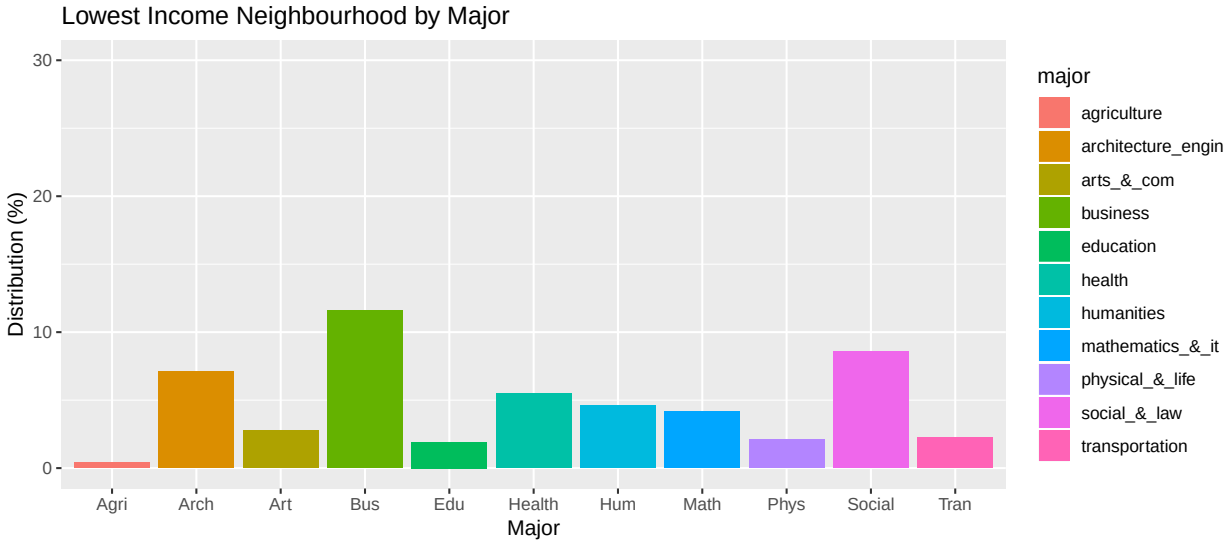
Income and Education by Major 2011

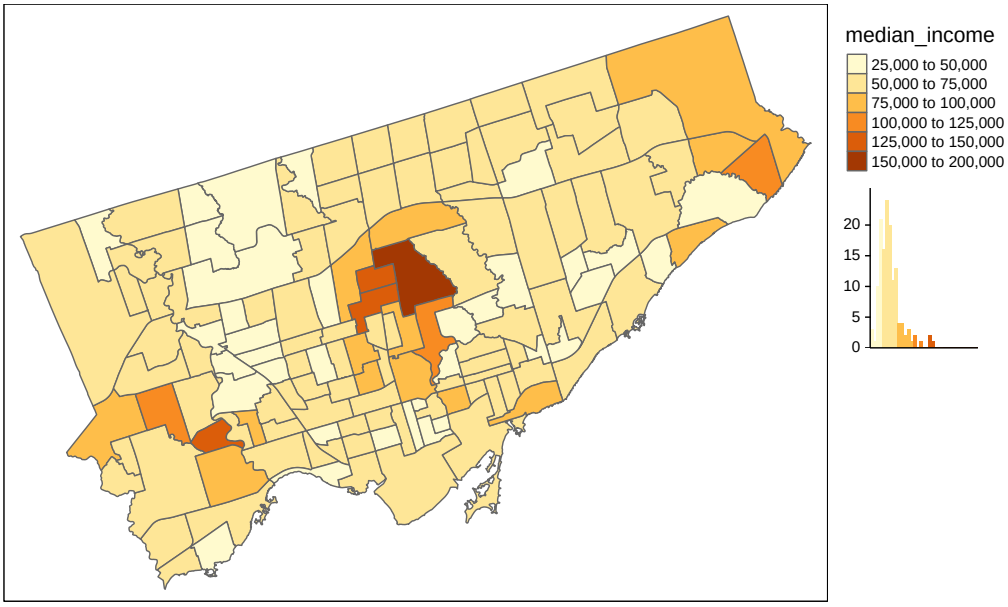


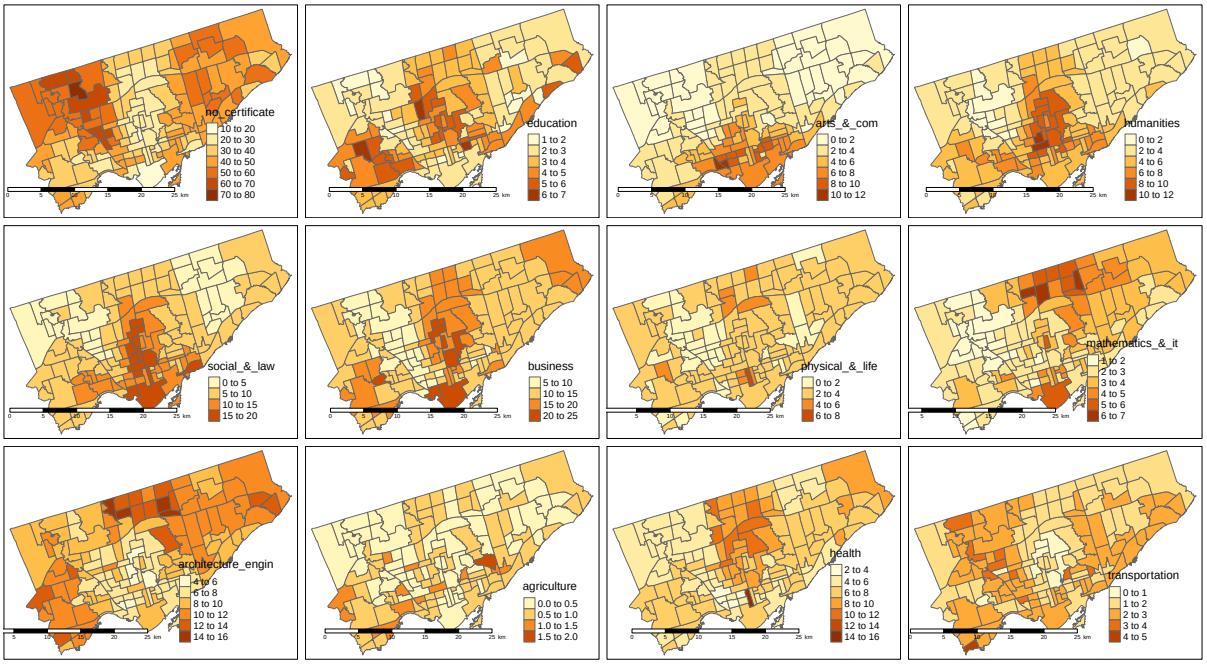



```
##
## Call:
## lm(formula = business ~ median_income, data = education_percentage_only)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3977 -2.1601 -0.1705  1.6585 10.7048
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.473e+00  8.122e-01   9.200  4.8e-16 ***
## median_income 9.409e-05  1.227e-05   7.671  2.7e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.101 on 139 degrees of freedom
## Multiple R-squared:  0.2974, Adjusted R-squared:  0.2924
## F-statistic: 58.84 on 1 and 139 DF,  p-value: 2.699e-12
```


Highest Income Neighbourhood by Major
Distribution (%)
30
20
10
0
Agri
Arch
Art
Bus
Edu
Health
Hum
Math
Phys
Social
Tran
Major
major
agriculture
architecture_engin
arts_&_com
business
education
health
humanities
mathematics_&_it
physical_&_life
social_&_law
transportation


Lowest Income Neighbourhood by Major
Distribution (%)
30
20
10
0
Agri
Arch
Art
Bus
Edu
Health
Hum
Math
Phys
Social
Tran
Major
major
agriculture
architecture_engin
arts_&_com
business
education
health
humanities
mathematics_&_it
physical_&_life
social_&_law
transportation





major	correlation
no_certificate	-0.49
transportation	-0.46
architecture_engin	-0.09
agriculture	-0.05
other	-0.02

major	correlation
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