Problem Set 1

Ilke Sun

02/10/2020

Question 1

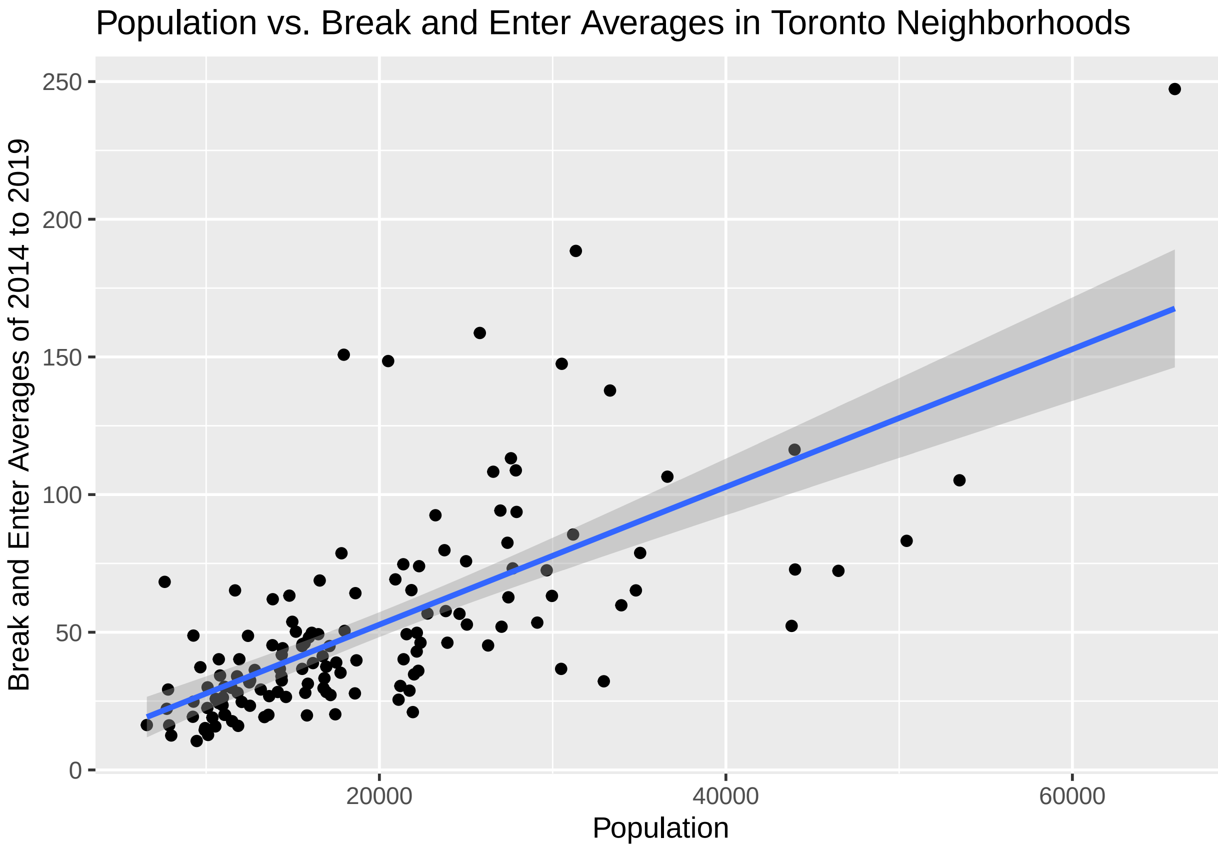
Part a

The data set includes the Crime Data by neighborhood in Toronto starting from 2014 up to, and including, 2019. There are 62 variables and 140 observations in the data set. The observations are 140 non-overlapping areas, which are the different neighborhoods. The variables include name of the neighborhood, population, assaults counts, robbery counts and other various crime rates, counts and averages, some of these variables are represented below. The averages and the populations of the neighborhoods are based on 2016 Census Population, hence, the population changes between 2014 and 2019 are neglected.

This data is interesting because it reports different crime counts in different neighborhoods in Toronto. I am interested in analyzing the relationship between population and average crime rates. There are various social crimes reported in this data set which are robbery, break and enter, auto theft, theft over. This data is good for analyzing because it reports counts, rates ,and averages, hence, it does not require much mutation. In addition, the data is numeric which is also beneficial because it is easier to visualize the data. The data set can also be improved in some ways. For example, the population could have been reported for each year so that we can a more in-depth analysis to see how population growth is associated with social crime counts in those neighborhoods but it would have been more costly for the city to obtain population for all neighborhoods in each year.

```
##  Neighbourhood        Population    Assault_Rate_2019 AutoTheft_2019
##  Length:140         Min.   : 6577   Min.   : 161.1    Min.   :  3.00
##  Class :character   1st Qu.:12020   1st Qu.: 392.3    1st Qu.: 15.75
##  Mode  :character   Median :16750   Median : 592.6    Median : 27.00
##                     Mean   :19511   Mean   : 714.3    Mean   : 37.04
##                     3rd Qu.:23854   3rd Qu.: 875.7    3rd Qu.: 42.00
##                     Max.   :65913   Max.   :3550.8    Max.   :482.00
##  BreakandEnter_2019 Homicide_2019     Robbery_2019              geometry
##  Min.   :  8.00     Min.   :0.0000   Min.   :  1.00   POLYGON      :140
##  1st Qu.: 26.00     1st Qu.:0.0000   1st Qu.: 11.00   epsg:4326    :  0
##  Median : 45.00     Median :0.0000   Median : 16.00   +proj=long...:  0
##  Mean   : 59.90     Mean   :0.5571   Mean   : 24.41
##  3rd Qu.: 67.25     3rd Qu.:1.0000   3rd Qu.: 30.25
##  Max.   :336.00     Max.   :3.0000   Max.   :143.00
```

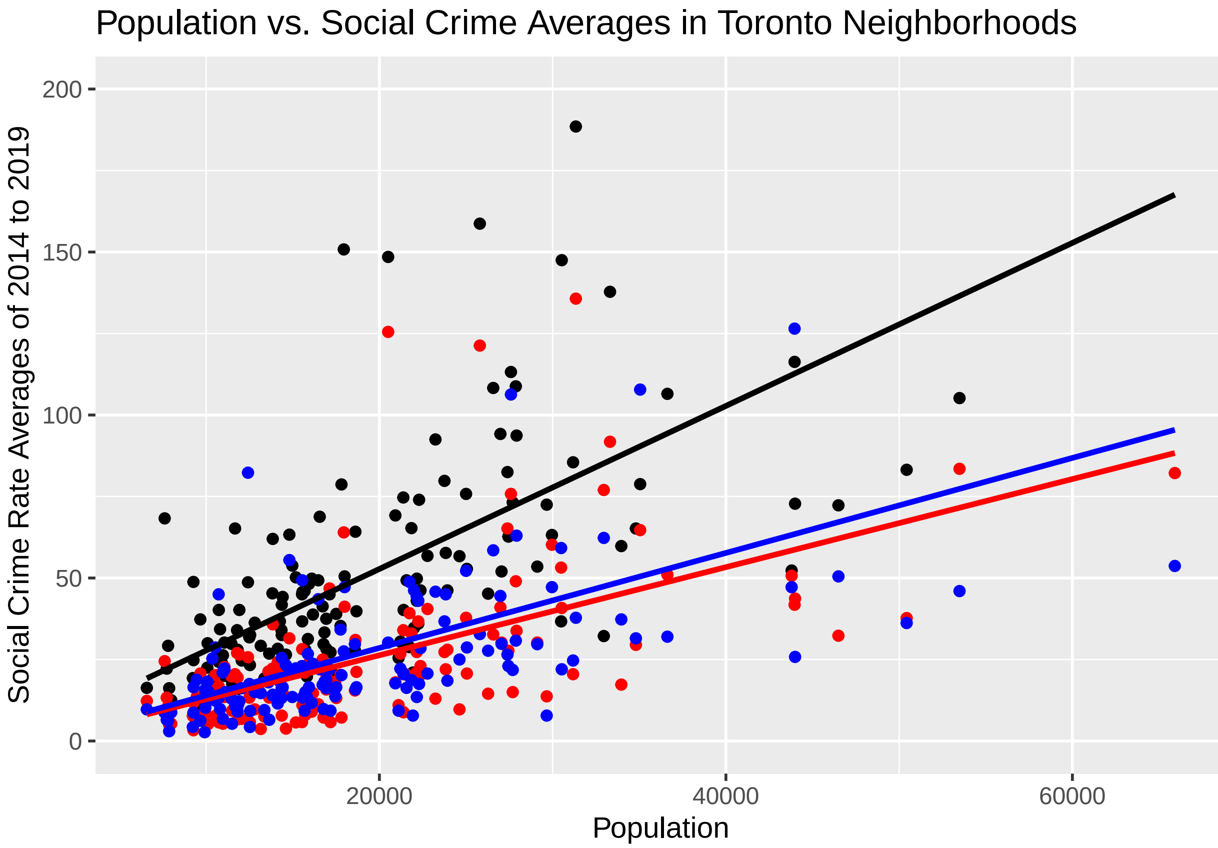
Part b



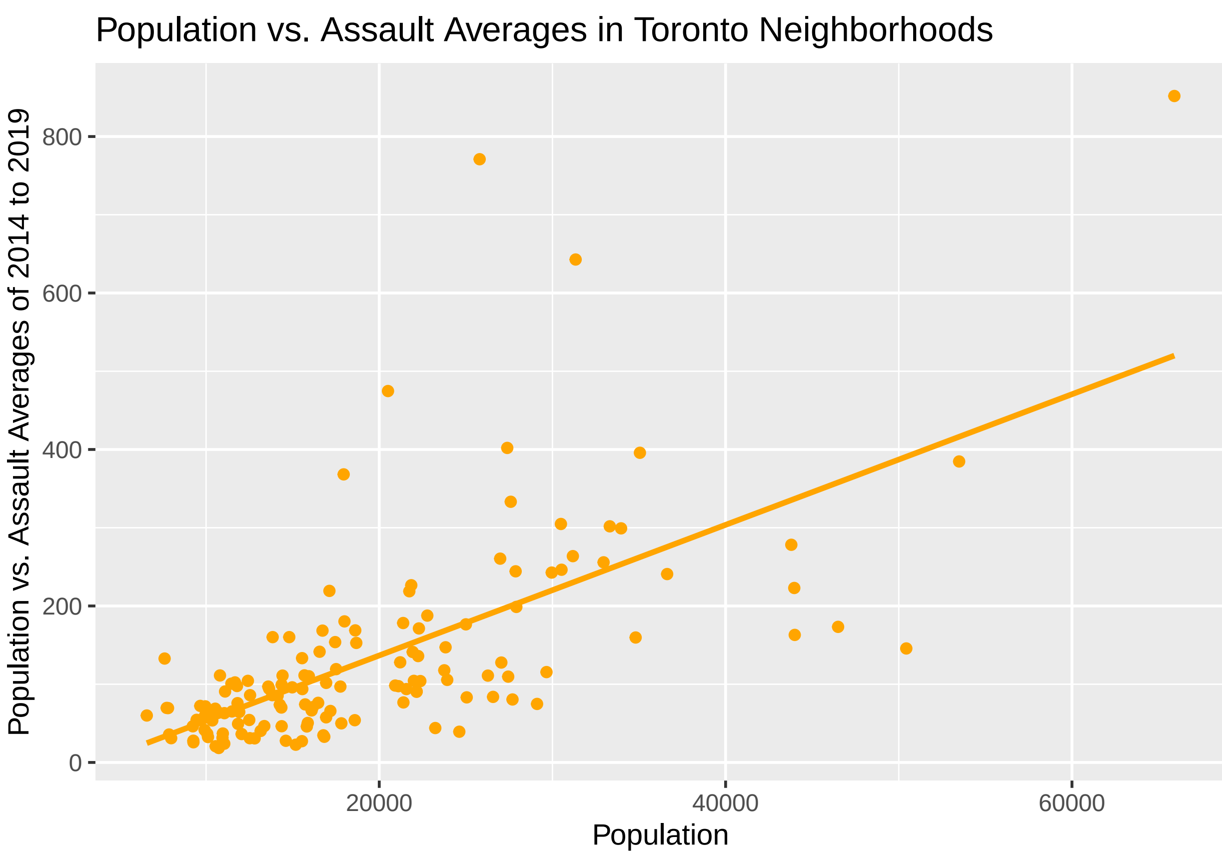
There are 140 data points in the scatter plot above. These data points are the different neighborhoods in Toronto. The scatter plot above has population of the neighborhoods as the explanatory variable and break and enter average between 2014 and 2019 as the outcome variable. This bivariate analysis is interesting because higher population is associated with higher crime rates, especially for some type of crimes which are usually the social crimes such as robbery, car theft, and break and enter. Vicky Chuqiao Yang, author of Bigger cities boost 'social crimes', says that "bigger cities are a double-edged sword, what leads to more innovation and wealth in bigger cities also makes them more dangerous at the same time." Hence, I have graphed population of the neighborhoods, and break and enter crime averages over six years of these neighborhoods to observe if this trend also holds in Toronto neighbors.

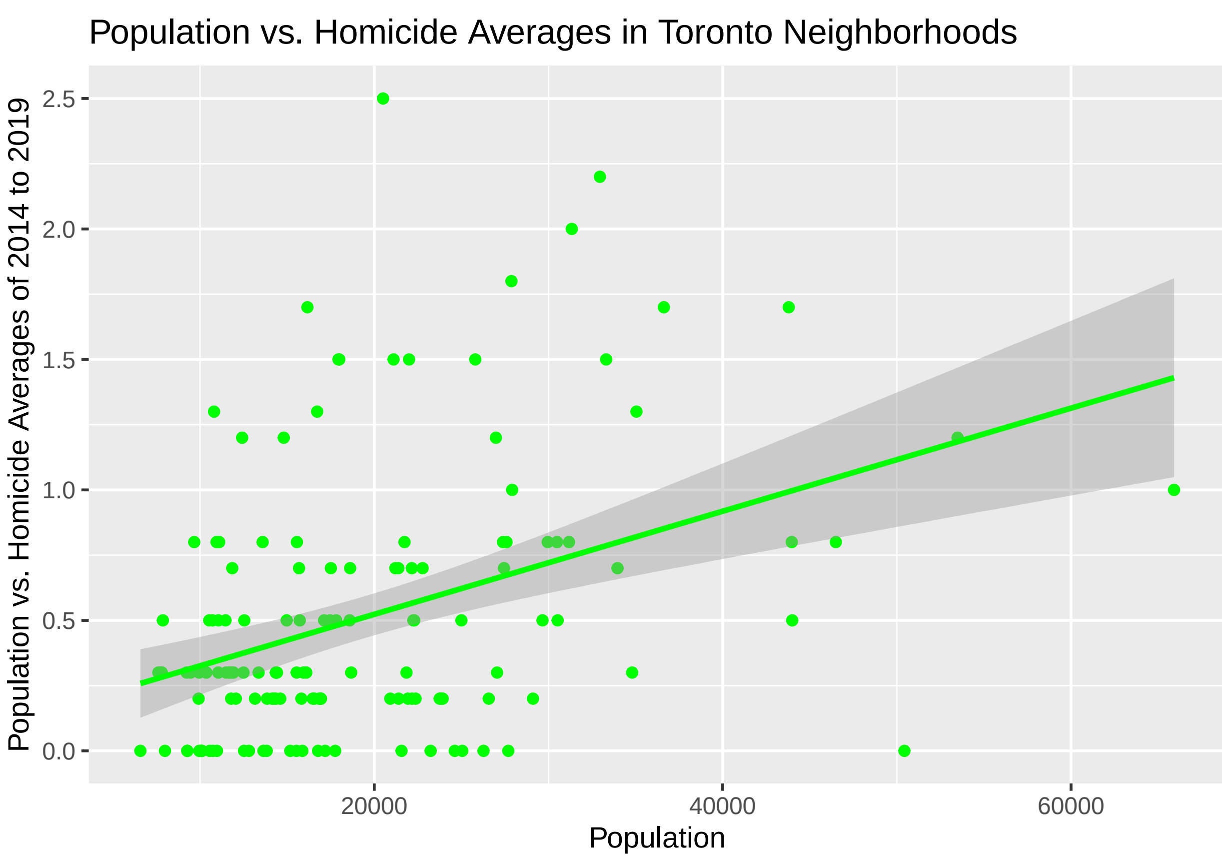
Initial observation of the linear model suggests that there is a positive correlation between the population and break and enter averages. However, most of the neighborhoods have less than 30,000 people living in it, hence, most of the points are concentrated on the bottom left side of the graph which makes it hard to derive an outcome visually. In order to better understand and visualize this relationship between population and social crimes, we may need to filter our data. In addition, we must analyze other types of crimes both social and other (e.g. homicide, assault) that were reported in the data set.

Part c



The graph above visualizes population and social crime averages of neighborhoods between 2014 and 2019. Black line and points represent break and enter averages, red line and points represent robbery averages, and blue line and points are represent auto theft averages. We see that there is a positive correlation for all types of crimes under social crimes. Break and enter averages increase the most as the population gets larger. Slopes for both robbery and auto theft are pretty similar, almost parallel to one another. In addition, y-axis is limited to 200 so that the data is better visualized and easier to observe.


Population vs. Assault Averages in Toronto Neighborhoods
Population vs. Assault Averages of 2014 to 2019
0
200
400
600
800
20000
40000
60000
Population


Population vs. Homicide Averages in Toronto Neighborhoods
Population vs. Homicide Averages of 2014 to 2019
0.0
0.5
1.0
1.5
2.0
2.5
20000
40000
60000
Population

Last two graphs visualize other crimes averages on population. Crimes categorized as other are assault and homicide. For homicide we observe similar slope to auto theft and robbery, however, linear model captures almost no data. It seems like homicide is not correlated with population. We observe that highest homicide rates are for mid-size neighborhoods. For assault, averages we observe the highest slope, including social crimes. If assault averages over those six years, have higher positive correlation with population, this would cause us to fail in rejecting the null hypothesis. We consider null hypothesis for this study to be there are no differences between the correlation between population and social crime averages and between population and other crime averages. Whereas the alternative hypothesis that we are testing for is that social crimes have higher positive correlation with population compared to other crimes.

Part d

In this study, we aimed to observe the correlation between population and crime rates in neighborhoods in Toronto. However, the study would be more in-depth and would able to provide more information if instead of population we used population density. I have provided prior information regarding how population and population growth may be associated with higher crime rates in the city. We have seen in the analysis that higher population is associated with higher crime in Toronto neighborhoods but initially I suggested that this increase would be more apparent, would have a higher positive correlation, for social crimes such as robbery and auto theft compared to other reported crimes which are assault and homicide. However, data visualizations have showed that this hypothesis is likely to be false. Assault crime averages between 2014 and 2019 was associated with larger increase as the population gets larger. With this background research we can conclude that that initial hypothesizes are likely to be untrue.

Better hypothesizes may be concern about population density rather than population. Thus, the alternative hypothesis would be that the higher population density is associated with higher social crime rates in Toronto neighborhoods. In contrast, the null hypothesis would be higher population density is not associated with higher crime rates in Toronto neighborhoods. In order to analyze, these hypothesis we must obtain data regarding the surface area of those non-overlapping neighborhoods so that we can obtain population densities for each of the observations. Then we can visualize data and use regression to provide evidence for the alternative hypothesis.

Part e

Bibliography:

1. Wu, Changbao, and Mary E. Thompson. "Basic Concepts in Survey Sampling." Sampling Theory and Practice. Springer, Cham, 2020. 3-15.
2. Waclawski, Eugene. "How I use it: Survey monkey." Occupational Medicine 62.6 (2012): 477-477.
3. "Open Data Dataset." City of Toronto Open Data Portal, open.toronto.ca/dataset/neighbourhood-crime-rates/.
4. Santa Fe Institute. (2019, September 18). Bigger cities boost 'social crimes'. ScienceDaily. Retrieved October 2, 2020 from www.sciencedaily.com/releases/2019/09/190918093101.htm