# Final Project - Auto Theft

## Analysis of Auto Thefts Occurred Focusing on Representative Neighbourhoods

Yuchen Zeng (Rachel), Leyi Wang (Amanda),Chen Yang (Karen),Jessie Lin, TUT0210, Group 3

# Introduction

- According to the Toronto Police Service, the occurrence of auto theft increased in Toronto over the past few years.
- We look at patterns in the distribution of auto thefts to determine how to better distribute the police force to prevent these crimes.

# Objectives

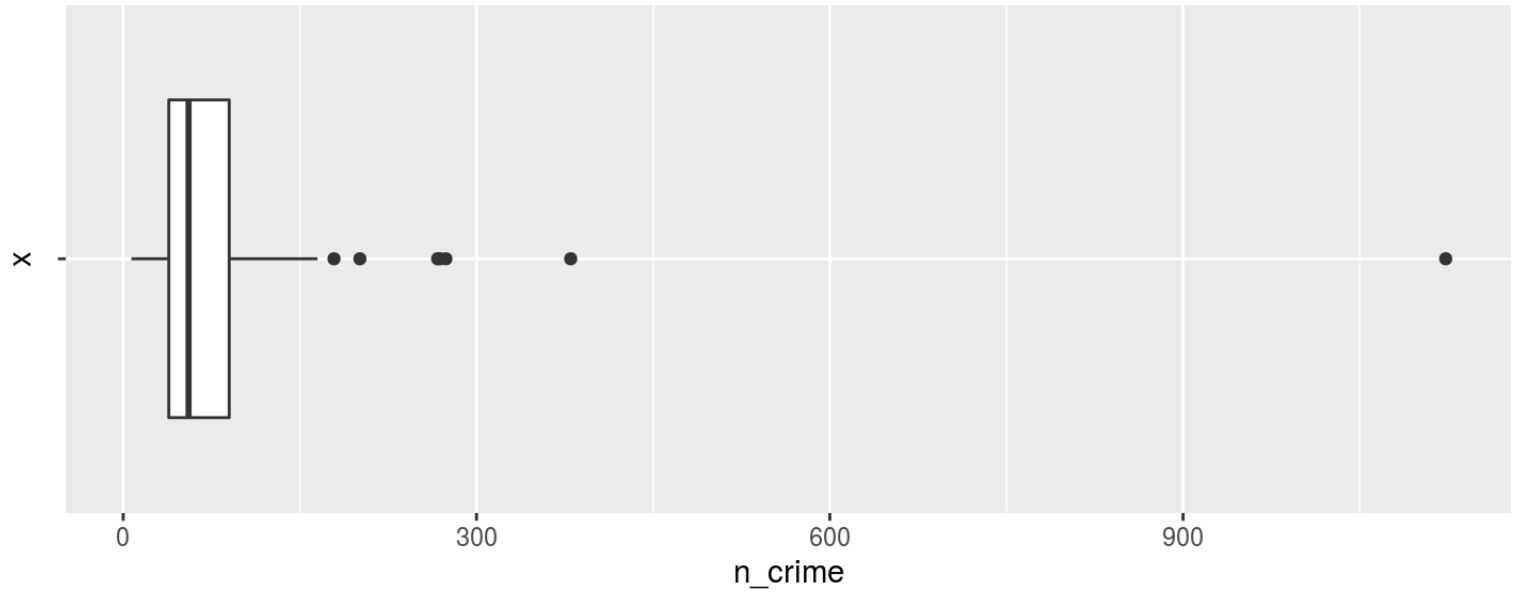
- What is the distribution of auto thefts rate looks like among for major neighbourhoods in Toronto?
- Are there any outliers?
- What factors can be used to predict the auto theft occurrence among neighbourhoods in Toronto?

# Data Summary - auto_thefts

**More Representative:**

- Only look at data from the past 3 years
- Filtered out all observations that occurred before 2016 in auto_thefts
- Phonomenon in some neighbourhoods are abnormal - May be caused by other factors that we cannot find out yet
- We focus on neighbourhoods that have commonalities & filtered out outliers

# Data Summary - Outliers




- The majority: in the range of 60 to 130
- Observed some outliers that have extremely larger crime occurrrence(7)

# Data Summary - Elders

**neighbourhood_profiles_2016**

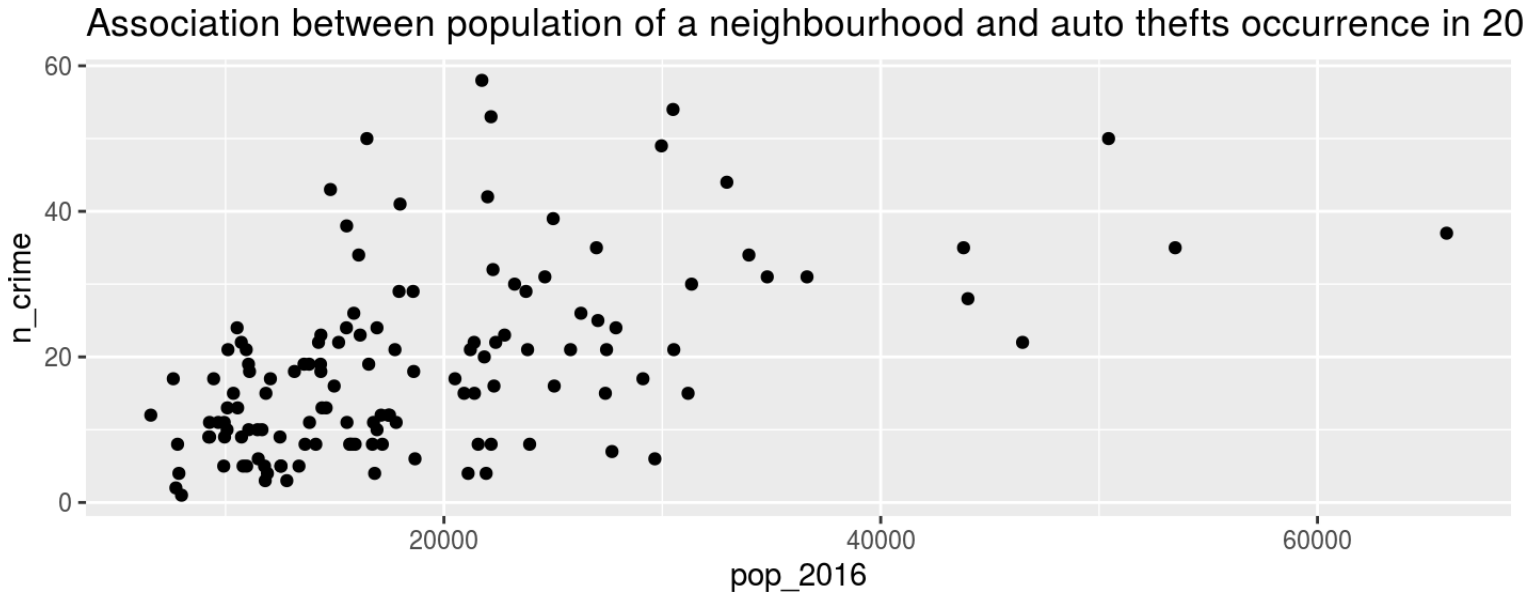
- We want to see if there is a relation between the number of auto thefts and the number of elders in the neighbourhood
- We create a new variable called **n_elders** contains the number of residences older than 64

# Data Summary - n_crime, auto_theft_rate

According to Toronto Police Service, a crime rate is calculated by dividing the number of reported crimes by the total population, and the result is multiplied by 100,000. We also created a variable called auto_theft_rate.

- Created 3 summaries of auto theft occurrence among neighbourhood for each year
- New variable: n_crime(auto thefts occurrence)
- New variable: auto_theft_rate(crime rate)

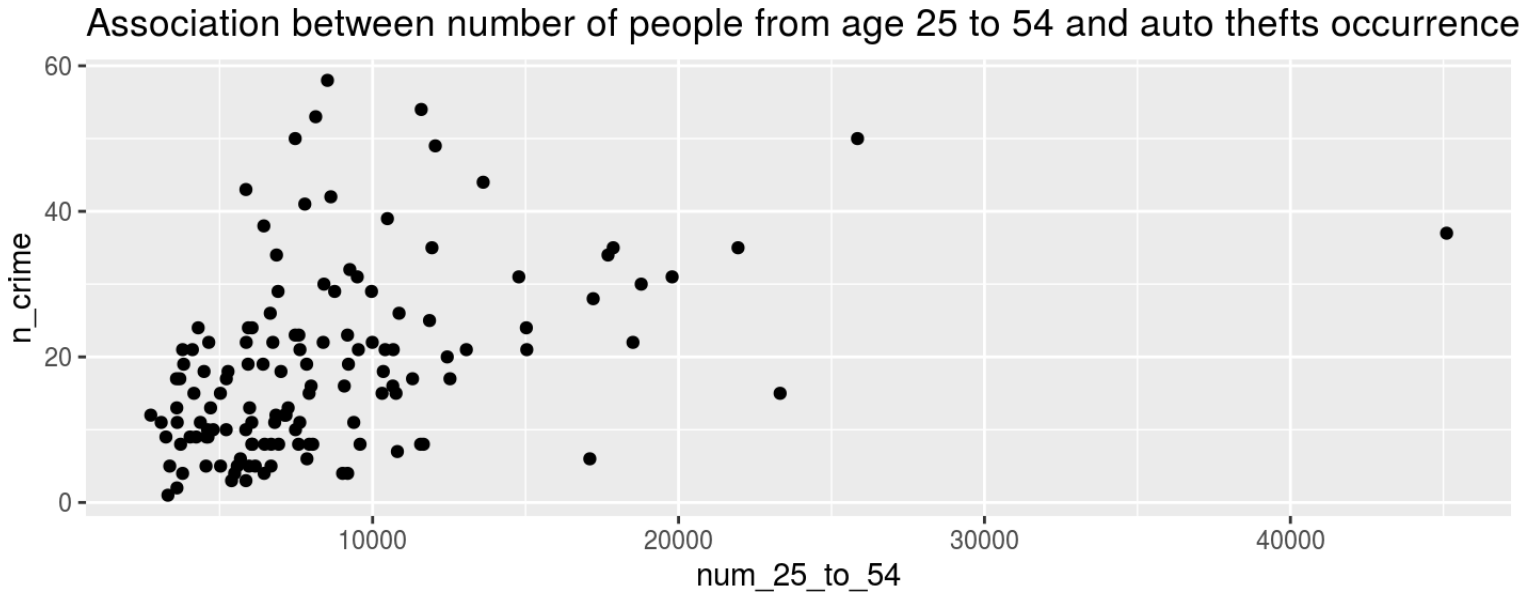
# Data Summary




## pop_2016 and n_crime

- The scatterplot of population of a neighbourhood and auto thefts occurrence in 2016
- strong positive linear association.

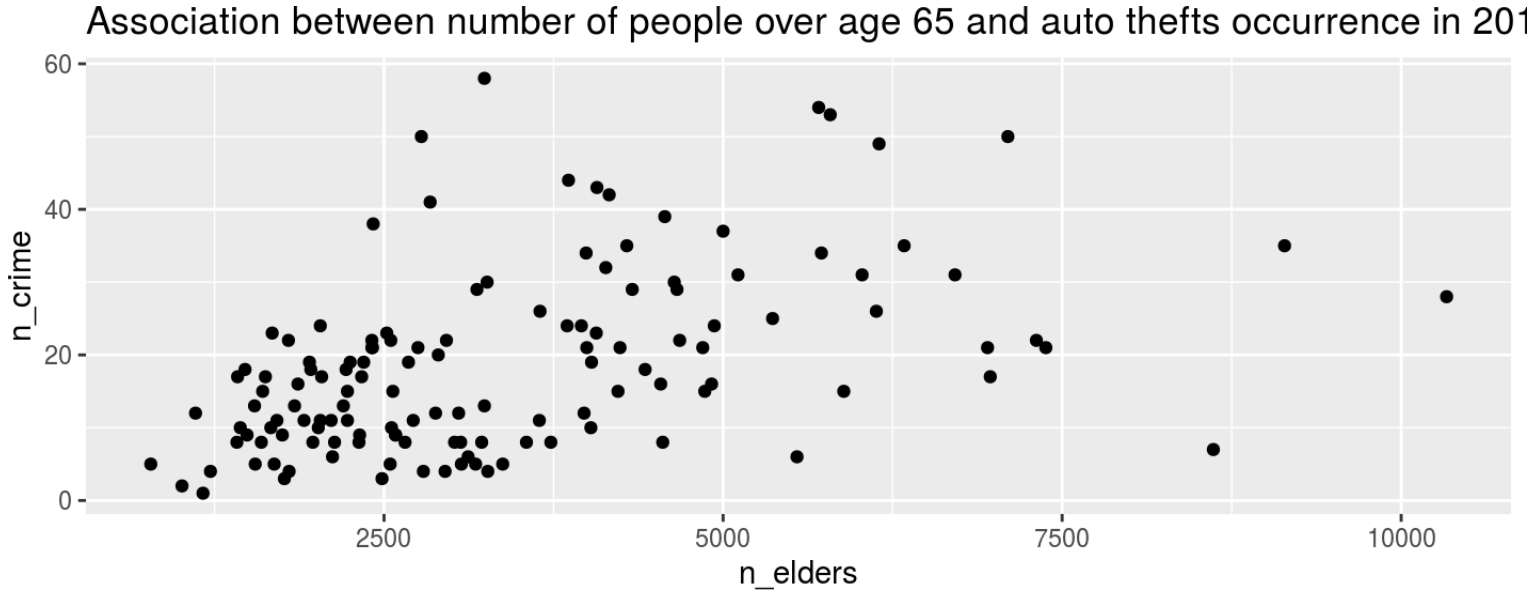
# Data Summary




**num_25_to_54 and n_crime**

- The scatterplot of number of people from age 25 to 54 in a neighbourhood and auto thefts occurrence in 2016
- strong positive linear correlation.

# Data Summary




## n_elders and n_crime

- The scatterplot of number of people over age 65 in a neighbourhood and auto thefts occurrence in 2016
- moderately strong positive linear correlation.

# Statistical Methods

## Confidence Interval, Bootstrap Sampling

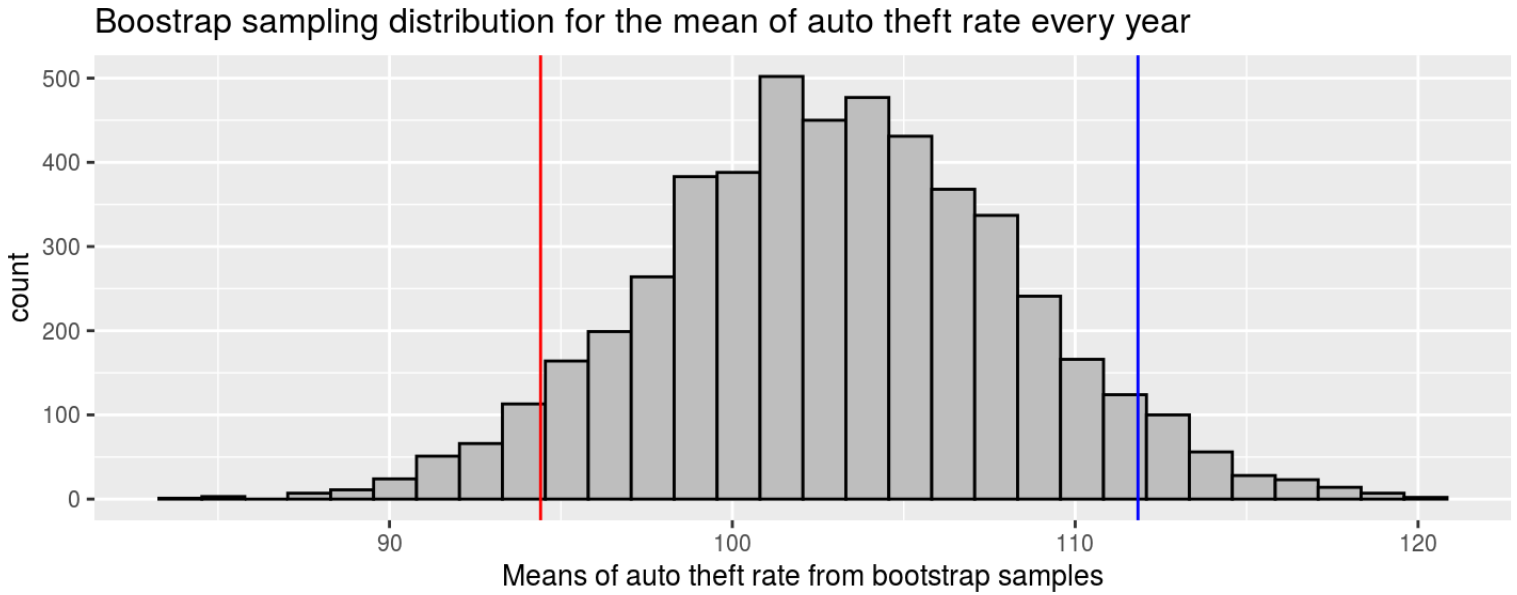
- Estimate the true mean of auto thefts rate among representative neighbourhoods every year
- Simulated 5000 bootstrap samples for auto theft rate of 2016 among neighbourhoods
- 90% confidence level

# Statistical Methods

**Linear Regression**
Fitted linear regression models using data from 2016 to predict auto theft rate among neighbourhoods.

- Training data: 80% of data from 3 summaries combined(2016-2018)
- Testing data: 20% of data from 3 summaries combined(2016-2018)
- Model A: num_25_to_54 as predictor
- Model B: n_elders as predictor
- Model C: pop_2016 as predictor
- Model D: pop_2016, n_elders, num_25_to_54

# Results




```
##        5%       95%
##  94.40746 111.83130
```

# Results - Models

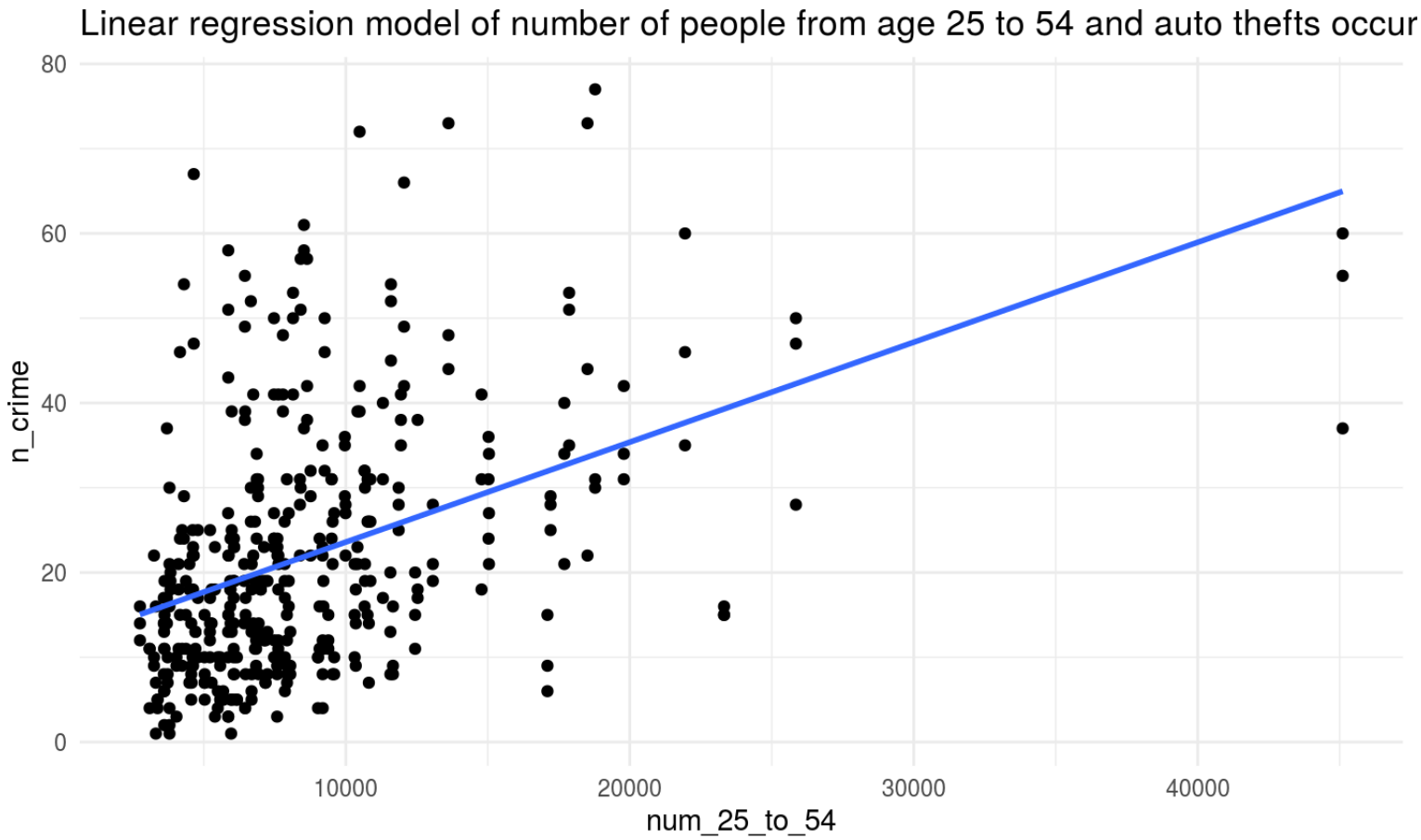
```
##                  Estimate     Pr(>|t|)
## (Intercept)  11.726510440 9.746478e-17
## num_25_to_54  0.001161843 2.382425e-17
```

```
##                Estimate     Pr(>|t|)
## (Intercept) 10.159630867 5.563762e-10
## n_elders     0.003342758 3.546517e-15
```

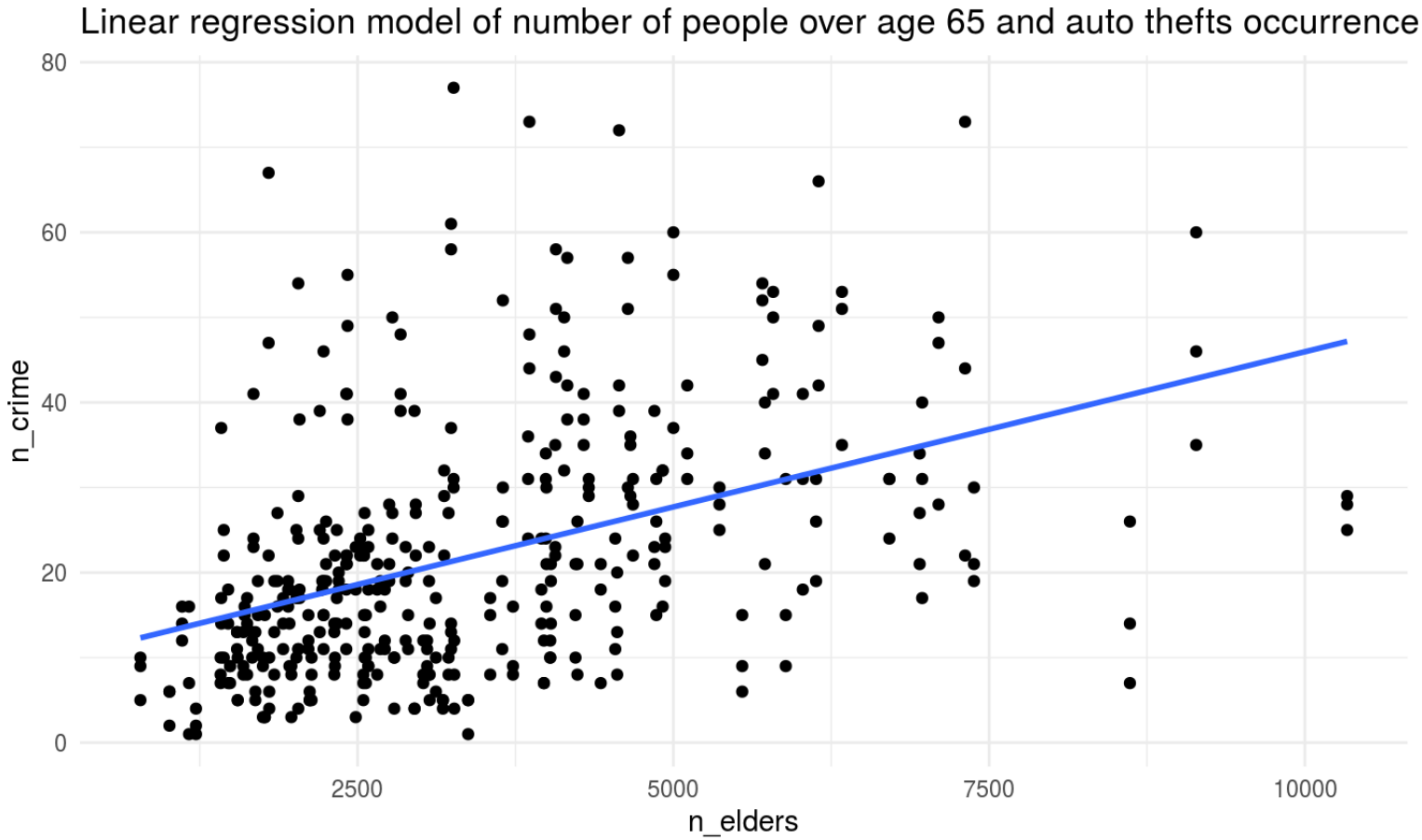
```
##                Estimate     Pr(>|t|)
## (Intercept) 6.9195719494 4.002670e-06
## pop_2016    0.0007736125 2.045868e-25
```

```
##                  Estimate     Pr(>|t|)
## (Intercept)   5.836513198 1.427045e-04
## pop_2016      0.002313230 1.251133e-10
## n_elders     -0.002539183 1.767208e-03
## num_25_to_54 -0.002268608 4.029742e-06
```

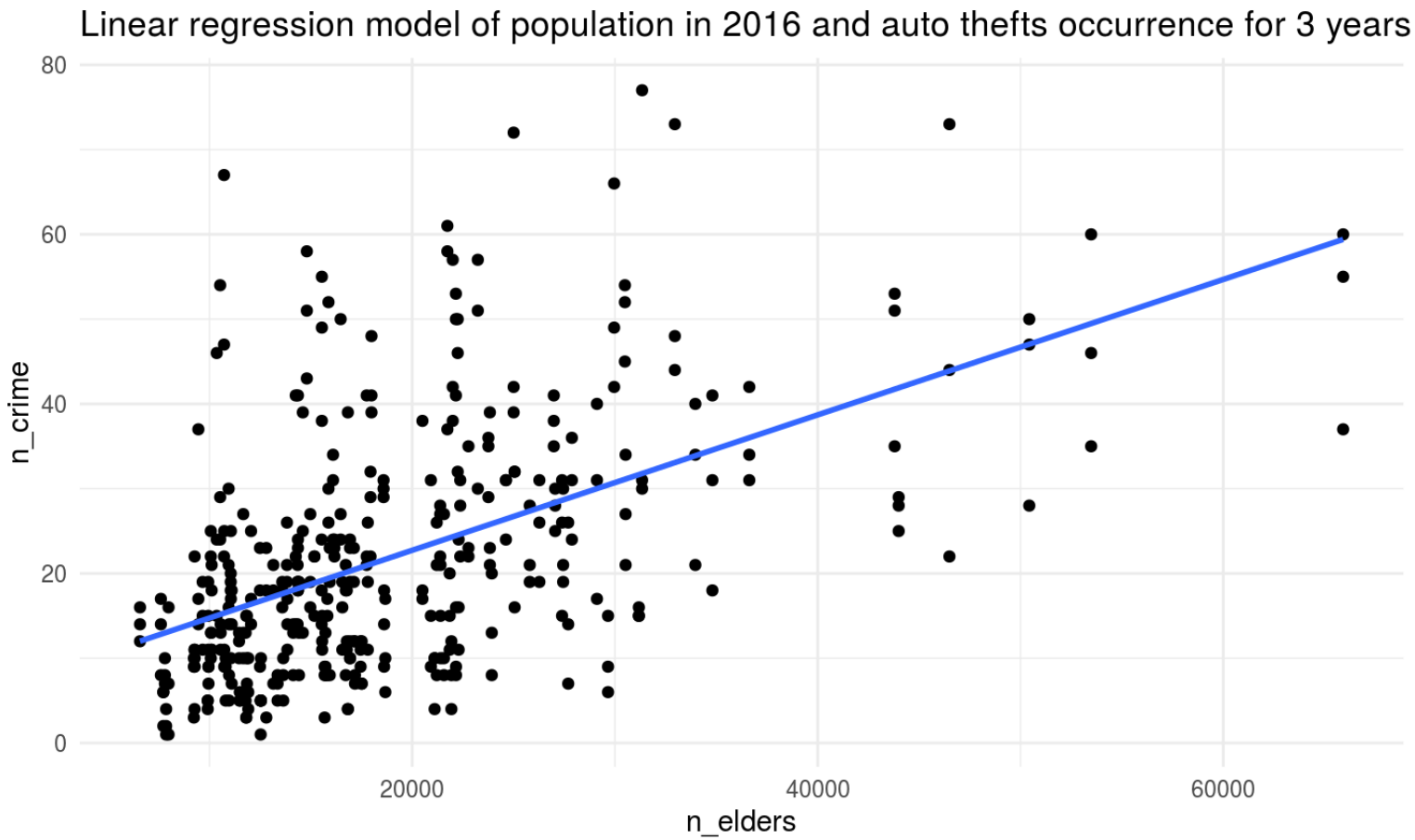
# Results - Models A



# Results - Models B



# Results - Models C



# Discussion - Models

- n_elders vs n_crime has the greatest slope(affects the crime occurrence more than the other factors for each unit increase in population.
- The more elderly, the more the crime occurrence
- Populations with a high elderly count most at risk for crime(require more security personnel)

# Results - RMSE

```
## # A tibble: 4 x 4
##   Model RMSE_train RMSE_test ratio_of_RMSEs
##   <chr>      <dbl>     <dbl>          <dbl>
## 1 A           15.6      12.8           1.22
## 2 B           14.6      13.0           1.13
## 3 C           14.5      12.1           1.20
## 4 D           14.1      11.7           1.21
```

# Discussion - Best Model

**Model D**

- With n_elders, pop_2016, num_25_to_54 as predictors
- Combines the effects of all three predictors
- Takes the effects of all the different scenarios into account
- Lowest RMSE
- Most useful in predicting future scenarios

Equation:

$$n\_crime = \beta_0 + \beta_1 pop\_2016 + \beta_2 n\_elders + \beta_3 num\_25\_to\_54$$

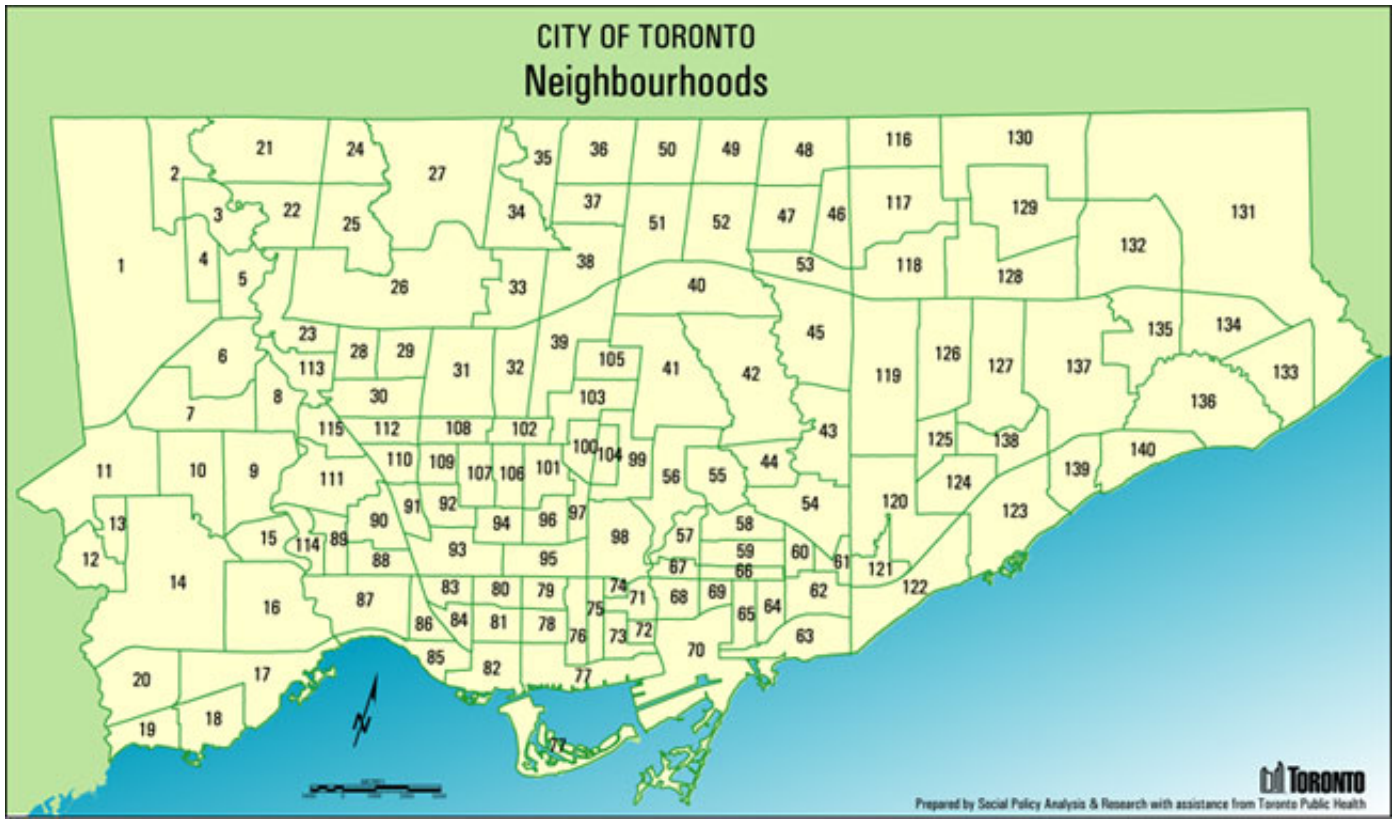
$$n\_crime = 5.836513198 + 0.002313230 pop\_2016 + 0.002539183 n\_elders + 0.002268608 num\_25\_to\_54$$

# Conclusion

## Outliers

- Hood ID: 1, 14, 21, 26, 27, 119, 130
- Contains similar values for other variables, but considerably higher auto thefts occurrence
- There should always be police stationed in these locations
- The possible factors that may be responsible for the abnormal outliers are the income, insurance, and proximity to crime inducing businesses (clubs, bars, etc).

# Outliers - Overview of Neighbourhoods




Credit: City of Toronto

# Conclusion

- The range for mean crime occurrence can help the police make budgets and recruitment accordingly
- E.g. Tackle 80 crimes each year - find which neighbourhood need more protection using estimated mean
- More police should be stationed in populated communities especially in those with a higher proportion of the elderly