Steps and Timeline

1. Data Preprocessing
    a. Data Cleaning
        i. Assign 1 or 2 people per dataset
    b. Data Integration
        i. For the Major Crime Dataset and Shelter Dataset ensure that the neighbourhoods and dates have the same format so that it's easier to integrate datasets. Dates should be in the yyyy-mm-dd format. For neighbourhoods, it should be neighbourhood name in one column and neighbourhood ID in another column
2. Azure Analysis
3. Presentation

Data Preprocessing

Datasets:

1. **MCI Dataset (Major Crimes)**
    a. *occurrence date, offence type, and neighbourhood*
2. **Shelter Dataset**
    a. *Date, neighbourhood, capacity, occupancy*
3. **Neighbourhood Profiles Dataset**
    a. *"Population density of each neighbourhood", "population by age", "average total income of individuals by neighbourhood", "employment rate", "rate of core housing need", "rate of unaffordable housing", "ratio of households Spending 30% or more/less of income on shelter costs", "average household size", and "highest level of education count" by neighborhood.*

Sunday, Nov 17th (Meeting Notes)

- **Shelter + overlap**
- **Overlap**

Cleaning Shelter Dataset:

6 shelters without postal code

1. We have added neighborhoods(& its IDs) from the addresses.
2. To find the neighbourhood and ID from the addresses, we used this website: https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/

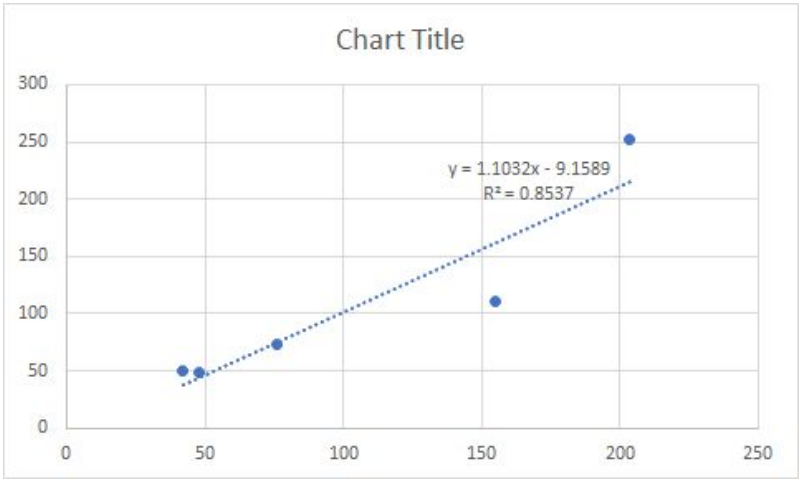
3. Used text to Columns to format the date to yyyy-mm-dd

4. There were 7 shelters/programs with zero/blank capacity, however, it does not make sense to have zero/blank capacity, thus we came up with a linear regression model using the average capacities and occupancies of the shelters in the same neighbourhood and sector (e.g. families, co-ed, men) with known/non zeros occupancy and capacity values. We then used the linear equation to predict the unknown capacity of the shelter (x=average of occupancy).

   First, we deleted the samples with zero/blank capacity and occupancy (2234 samples). The shelters affected:

   - ❖ Family Residence - AL Site (Some: 241)
   - ❖ Family Residence - GA Site (Some: 267)
   - ❖ Family Residence - CoEd Building (Some: 89). No zero/blank after deleting the zero/blank occupancy and capacity. No further work needed
   - ❖ Family Residence - ID Site (All: 278)
   - ❖ Second Base Youth Shelter (All: 410)
   - ❖ S.A. Hope - 167 College St (All: 730)
   - ❖ YWCA - Beatrice House (All: 219)

   - 8.99 occup & unknown cap for Family Residence AL Site ->0.76->1 (103 samples)
   - 3.91 occup & unknown cap for Family Residence GA Site ->-4.86 ->0 (11 samples)
   - 0 for Family Residence ID Site (all capacity & occupancy for this is 0 -> deleted)
   - 203.48 occup & 252 cap for Family Residence LI Site
   - 47.69 occup & 48 cap for Family Residence MA Site
   - 155.14 occup & 110 cap for Family Residence Main Building
   - 42.03 occup & 50 cap for Family Residence Roycroft Site
   - 75.88 occup & 72.5 cap for Family Residence TO Site
   - Linear Graph with the equation & R2:



The predicted capacity for AL Site is 1 (rounding 0.75). For GA Site, the predicted capacity is negative (-4.86). Since it does not make sense to have a negative capacity, we will enter 0 as capacity. As a result, we predicted 114 capacity values (104 for AL Site and 11 for GA Site).

5. There were 13 shelters/programs with some/all zero/blank occupancy.
    a. Fort York Extreme Weather Program
    b. Red Door - Family Shelter
    c. University Settlement - Out of the Cold
    d. University Settlement: Extreme Weather Program
    e. Covenant House - Transitional Safe Beds for women: the capacity is 2. Hence zero occupancy is normal
    f. Family Residence - Co-Ed
    g. Seaton House - Hostels Extreme Weather Program
    h. Eva's Satellite Extreme Weather Program: the capacity is 1. Hence zero occupancy is normal
    i. COSTI - Quality Suites Singles Refugee Program
    j. COSTI Radisson Hotel Family Program.
    k. Sojourn House - Queens Drive: The zeros and other low occupancy are associated with low capacity, no change
    l. Fife-Sherbourne Transitional Program
    m. Christie Refugee Welcome Centre - Singles

We smoothed outliers, filled in the blanks and replaced the zero occupancies with the monthly average of the non-zero occupancy numbers (rounded to whole number) of the given shelter.

For instance, the average number of the non-zero occupancy for the Fort York Extreme Weather Program in January is 10.8 (round to 11), Therefore, we filled in the blanks, replaced the zeros.

We filled in/ replaced or smoothed the occupancy values of 322 samples.
We also searched for duplicated entries such as same date, shelter name and program name, however, there was none of them existing.

R Code for Aggregation: Average and Total Occupancy and Capacity

```
#Aggregating and Merging by Week and Neighbourhood
library(readxl)
library(writexl)
daily_shelter<-read_excel("Shelter_Final.xlsx")
Table1<-aggregate(daily_shelter$OCCUPANCY,
by=list(daily_shelter$WEEK,daily_shelter$NEIGHBOURHOOD,daily_shelter$NEIGHBOURHOOD_ID),FUN = sum)
Table2<-aggregate(daily_shelter$CAPACITY,
by=list(daily_shelter$WEEK,daily_shelter$NEIGHBOURHOOD,daily_shelter$NEIGHBOURHOOD_ID),FUN = sum)
Table3<-merge(Table1,Table2,by=c("Group.1","Group.2","Group.3","Group.3"),all=TRUE)
write_xlsx(Table3,"Aggregate_Shelter.xlsx")
```

```
Table4<-aggregate(daily_shelter,
by=list(daily_shelter$WEEK,daily_shelter$NEIGHBOURHOOD,daily_shelter$SHELTER_NAME),FUN =length)
write_xlsx(Table4,"Shelter_Count.xlsx")

library(readxl)
library(writexl)
MCI<-read_excel("MCI_Assault.xlsx")
Table1<-aggregate(MCI$ASSAULT, by=list(MCI$WEEK,MCI$NEIGHBOURHOOD_ID,MCI$NEIGHBOURHOOD),FUN = sum)
write_xlsx(Table1,"Aggreagate_MCI_Assault.xlsx")

MCI<-read_excel("MCI_breakenter.xlsx")
Table2<-aggregate(MCI$BREAKANDENTER, by=list(MCI$WEEK,MCI$NEIGHBOURHOOD_ID,MCI$NEIGHBOURHOOD),FUN =
sum)
write_xlsx(Table2,"Aggreagate_MCI_breakenter.xlsx")

MCI<-read_excel("MCI_autotheft.xlsx")
Table3<-aggregate(MCI$AUTOTHEFT, by=list(MCI$WEEK,MCI$NEIGHBOURHOOD_ID,MCI$NEIGHBOURHOOD),FUN = sum)
write_xlsx(Table3,"Aggreagate_MCI_autotheft.xlsx")

MCI<-read_excel("MCI_robbery.xlsx")
Table4<-aggregate(MCI$ROBBERY, by=list(MCI$WEEK,MCI$NEIGHBOURHOOD_ID,MCI$NEIGHBOURHOOD),FUN = sum)
write_xlsx(Table4,"Aggreagate_MCI_robbery.xlsx")

MCI<-read_excel("MCI_theftover.xlsx")
Table5<-aggregate(MCI$THEFTOVER, by=list(MCI$WEEK,MCI$NEIGHBOURHOOD_ID,MCI$NEIGHBOURHOOD),FUN = sum)
write_xlsx(Table5,"Aggreagate_MCI_theftover.xlsx")

MCI<-read_excel("MCI_total.xlsx")
Table6<-aggregate(MCI$TOTAL, by=list(MCI$WEEK,MCI$NEIGHBOURHOOD_ID,MCI$NEIGHBOURHOOD),FUN = sum)
write_xlsx(Table6,"Aggreagate_MCI_total.xlsx")

#Merging the sub_datasets
table1<-read_excel("Aggreagate_MCI_Assault.xlsx")
table2<-read_excel("Aggreagate_MCI_breakenter.xlsx")
Table<-merge(table1,table2,by=c("WEEK","NEIGHBOURHOOD_ID","NEIGHBOURHOOD"), all=TRUE)

table3<-read_excel("Aggreagate_MCI_autotheft.xlsx")
Tablebis<-merge(Table,table3,by=c("WEEK","NEIGHBOURHOOD_ID","NEIGHBOURHOOD"), all=TRUE)

table4<-read_excel("Aggreagate_MCI_robbery.xlsx")
Tableter<-merge(Tablebis,table4,by=c("WEEK","NEIGHBOURHOOD_ID","NEIGHBOURHOOD"), all=TRUE)

table5<-read_excel("Aggreagate_MCI_theftover.xlsx")
Tablequart<-merge(Tableter,table5,by=c("WEEK","NEIGHBOURHOOD_ID","NEIGHBOURHOOD"), all=TRUE)

table6<-read_excel("Aggreagate_MCI_total.xlsx")
TableFinal<-merge(Tablequart,table6,by=c("WEEK","NEIGHBOURHOOD_ID","NEIGHBOURHOOD"), all=TRUE)
write_xlsx(TableFinal,"Aggregate_MCI.xlsx")

#Merging MCI and Shelter by Week and Neighbourhood
library(readxl)
library(writexl)

table1<-read_excel("Aggregate_Shelter.xlsx")
table2<-read_excel("Aggregate_MCI.xlsx")
Table<-merge(table1,table2,by=c("WEEK","NEIGHBOURHOOD","NEIGHBOURHOOD_ID"),all.x=TRUE)
```

```
write_xlsx(Table,"Aggregate_Dataset_Mid_BIS.xlsx")

table3<-read_excel("Neighborhood_Cleaned_Final.xlsx")
Table1<-merge(Table,table3,by=c("NEIGHBOURHOOD","NEIGHBOURHOOD_ID"),all.x=TRUE)
write_xlsx(Table1,"Aggregate_Dataset_Final.xlsx")
```

In between aggregating and merging datasets in R, In Excel, we also:

- deleted duplicated columns
- edited column names
- matched shelter count with neighbourhood and week
- calculated the over_under_capacity and the shelter_demand

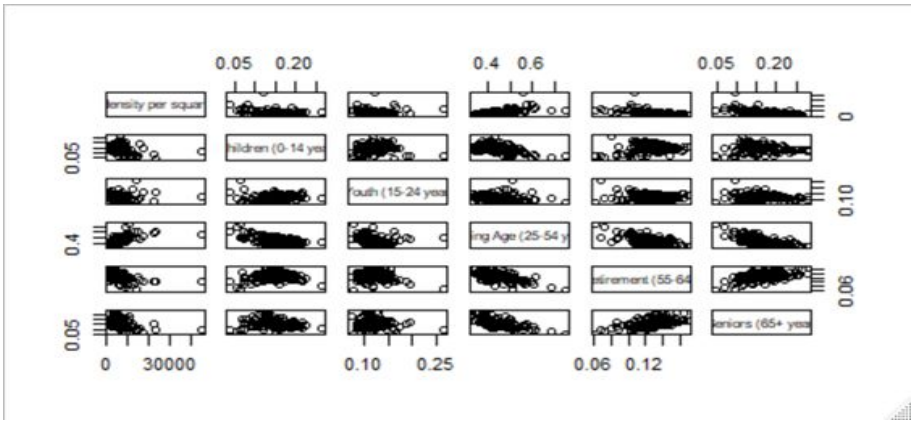
AVERAGE_UNDER_OVER_CAPACITY = AVERAGE_CAPACITY - AVERAGE_OCCUPANCY
TOTAL_UNDER_OVER_CAPACITY = TOTAL_CAPACITY - TOTAL_OCCUPANCY
SHELTER_DEMAND= (TOTAL_OCCUPANCY*100)/TOTAL_CAPACITY. We always get the same result, no matter if we use the total or the average.

Cleaning the Neighbourhood Dataset: Extracting the Variables and Data reduction: dimensionality reduction by high correlation (linear regression) and low variance filters

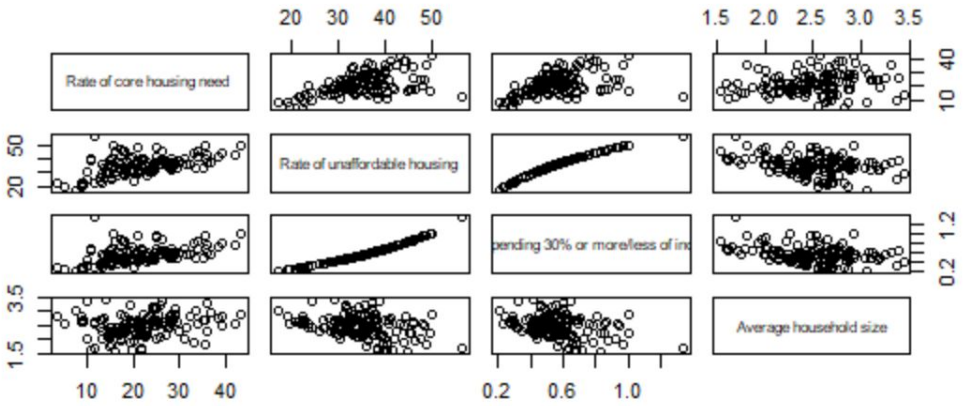
Example R code for graphs:

```
Table1<- read_excel("Aggregate_Neighbourhood_Population_Variable_Reduction.xlsx")
attach(Table1)
plot(Table1)
detach(Table1)
```

1. **Population**
    - 1.1. Population, 2016
    - 1.2. Population density per square kilometre
    - 1.3. %Children (0-14 years)
    - 1.4. %Youth (15-24 years)
    - 1.5. %Working Age (25-54 years)
    - 1.6. %Pre-retirement (55-64 years)
    - 1.7. %Seniors (65+ years)



2. **Household**
    2.1. Rate of core housing need
    2.2. Rate of unaffordable housing
    2.3. Ratio of households Spending 30% or more/less of income on shelter costs (re)
    2.4. Average household size

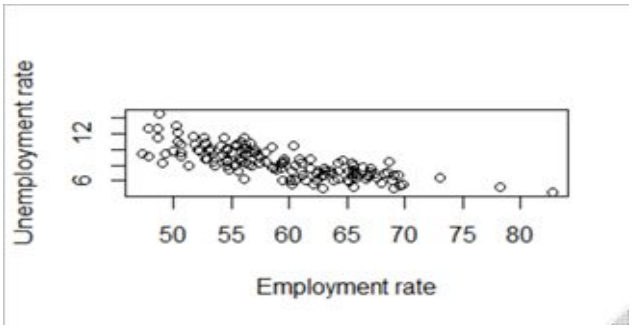


VAR(%Private dwellings occupied by usual residents) = 0.00065. Hence, we can drop it.

There is a **high correlation between the Rate of unaffordable housing** and the **Ratio of households Spending 30% or more/less of income on shelter costs**.
VAR(**Rate of unaffordable housing**) = 47.49
VAR(**Ratio of households Spending 30% or more/less of income on shelter costs**) = 0.031

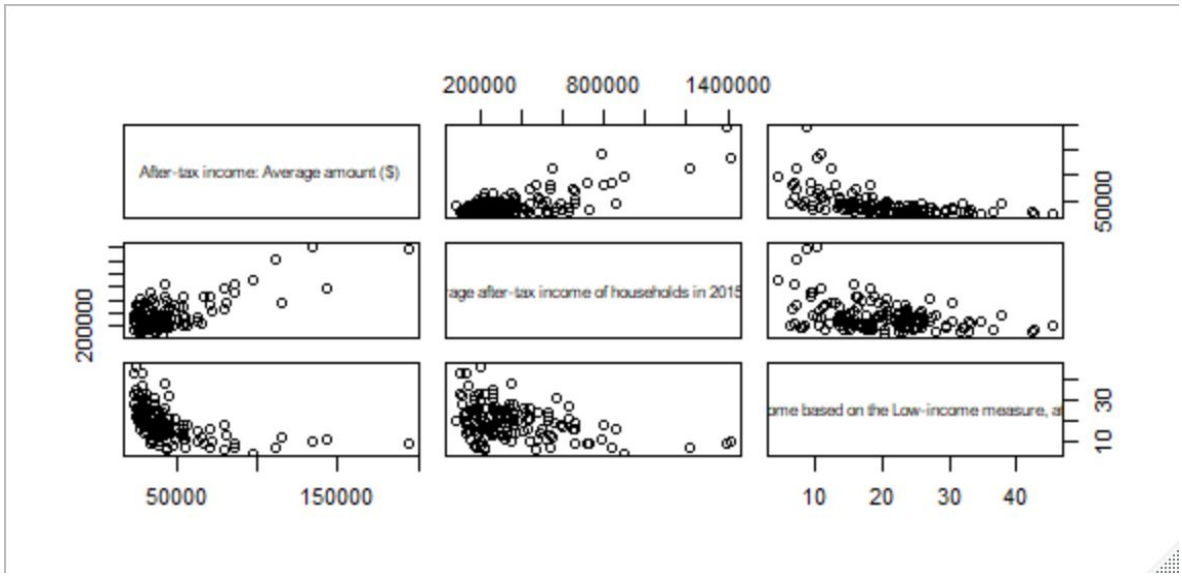
There is a **correlation between the Rate of unaffordable housing** and the **Rate of core housing needs.**
VAR(**Rate of unaffordable housing**) = 47.49
VAR(**Rate of core housing need**) = 60.60

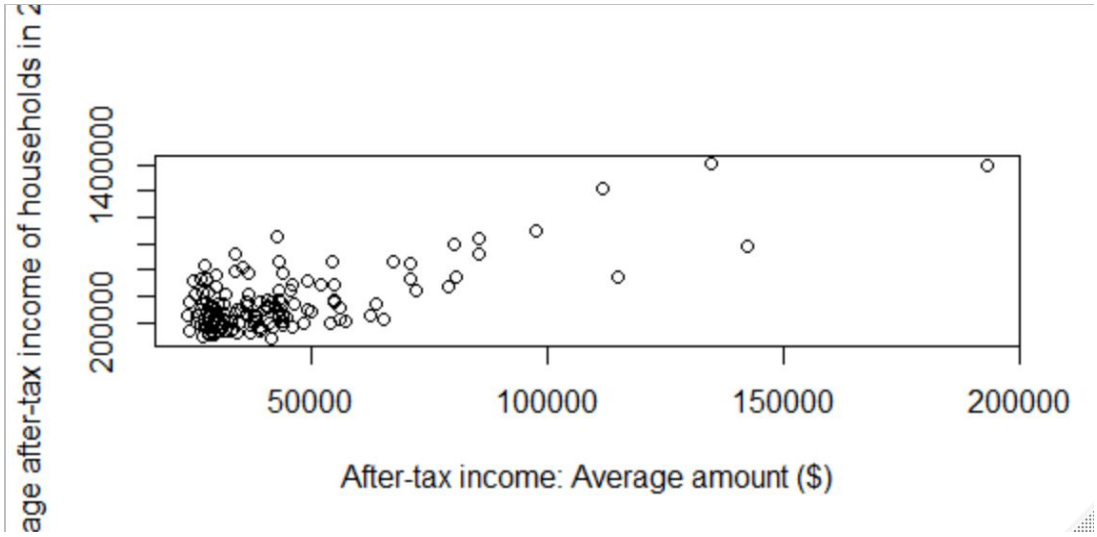
3. **Employment**
    3.1. Employment rate
    3.2. Unemployment rate



The graph shows a **correlation between Employment rate** and **Unemployment rate**
VAR(**Employment rate**)=41.85
VAR(**Unemployment rate**)=3.55

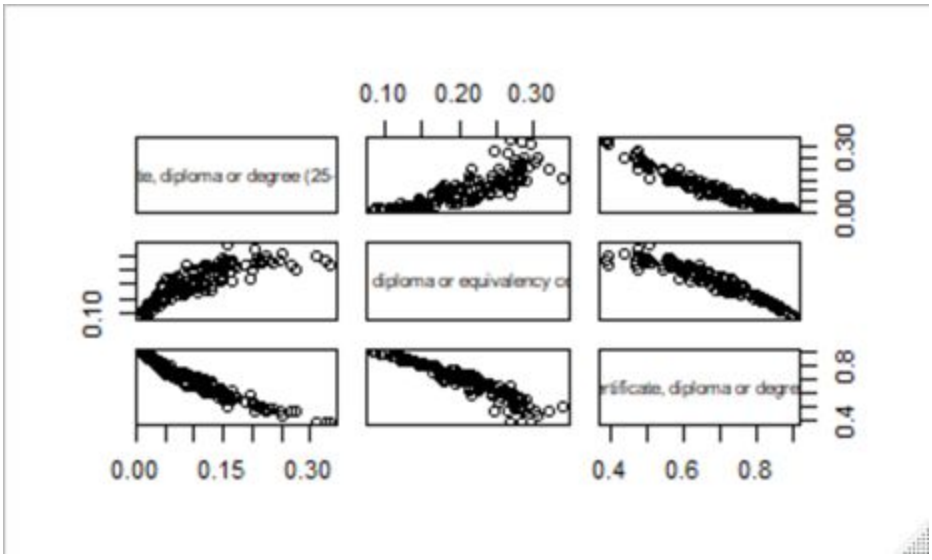
4. **Income**
   - 4.1. After-tax income: Average amount ($)
   - 4.2. Average after-tax income of households in 2015 ($)
   - 4.3. Prevalence of low income based on the Low-income measure, after tax (LIM-AT) (%)





The graph above shows a **correlation for After-tax income: Average amount ($)** and **Average after-tax income of households in 2015 ($).**
VAR(**After-tax income: Average amount**) = 5.72E + 08
VAR(**Average after-tax income of households in 2015**) = 5.31E + 10

5. **Education**
   - 5.1. % of No certificate, diploma or degree (25-64 years old)
   - 5.2. % of Secondary (high) school diploma or equivalency certificate (25-64 years old)
   - 5.3. % of Postsecondary certificate, diploma or degree (25-64 years old)



The graphs above appears to show a **strong correlation between all three variables.**
VAR(**%No certificate, diploma or degree**) = 0.0051
VAR(**%Secondary school diploma or equivalency certificate**) = 0.0037
VAR(**Postsecondary certificate, diploma or degree**) = 0.016