# Investigation of $NO_x$ (in ppb), $O_3$ (in ppb), and $PM_{2.5}$ (in µg/m3) trends in Ontario from 2003 to 2022

Mohamad Damaj - 10109, Moham An, Jo Zhu

2025-03-18

## 1. Description of Datasets

The datasets analyzed in this report are three, one for each of $NO_x$ (in ppb), $PM_{2.5}$ (in µg/m3), and $O_3$ (in ppb), all containing the same variables. These datasets contain records of the pollutant levels from 2003 (considered the earliest of all three) to 2022. Each dataset contains detailed information of pollutant levels in the form of the following variables:

```
##  [1] "Year"            "Station Number"  "City"            "Location"
##  [5] "Type"            "Valid Hour"      "10th Percentile" "30th Percentile"
##  [9] "50th Percentile" "70th Percentile" "90th Percentile" "99 Percentile"
## [13] "Mean"            "1-Hour Maximum"  "24-Hour  Maximum"
```

1. Year: The calendar year when the pollutant data was recorded.
2. Station Number: A unique numerical identifier for the monitoring station.
3. City: The name of the city where the pollutant was measured.
4. Location: A more specific description of the monitoring location (street, area).
5. Type: The type of monitoring station or measurement protocol used.
6. Valid Hour: The number of valid hourly measurements recorded for that pollutant in the given year.
7. Percentiles: Percentile statistics showing the distribution of pollutant levels (in ppb, µg/m3, ppb) throughout the year.
8. Mean: The annual average concentration of the pollutant (in ppb, µg/m3, ppb).
9. 1-Hour Maximum: The highest recorded concentration within a one-hour period during the year.
10. 24-Hour Maximum: The highest average concentration recorded over a 24-hour period in that year.

## 2. Background of the Data

The data used in this analysis was collected across various cities in Ontario, throughout the period from 2003 to 2022. These measurements were compiled and made available by the Ontario Ministry of the Environment, Conservation and Parks.

The information is publically avaliable and is helpful in assiting researchers, policymakers, and public health officials to investigate air quality trends, evaluate the impact of environmental regulations, and inform policy decisions. It provides a long-term view of pollutant levels in various cities in Ontario, allowing for improving environmental and public health outcomes in Ontario.

# 3. Overall Research Question

The aim of this paper is to investigate the spatial and temporal trends of air pollutants ($NO_x$, $O_3$, and $PM_{2.5}$) across various Ontario cities from 2003 to 2022.

1. How have annual mean concentrations (in ppb, µg/m3, ppb) for each pollutant changed from 2003 to 2022, and which years show the most significant shifts?

2. Which cities consistently rank among the highest (or lowest) in terms of average pollutant levels, and do these rankings shift over time?

3. How do the four pollutants correlate with each other across different cities and years, and what might this indicate about broader air quality patterns?

4. How do the four pollutants project into future years based on current trends?

# 4. Tables

## 4.1 The Top 10 Cities With the Highest Pollutant Concentration

| $NO_x$ Table (in ppb) | | $O_3$ Table (in ppb) | | $PM_{2.5}$ Table (in µg/m3) | |
|---|---|---|---|---|---|
| City | Conc. | City | Conc. | City | Conc. |
| Toronto West | 31.32 | Port Stanley | 32.84 | Sarnia | 9.57 |
| Toronto East | 21.82 | Tiverton | 31.93 | Windsor West | 9.03 |
| Toronto Downtown | 20.41 | Grand Bend | 31.28 | Hamilton Downtown | 8.91 |
| Toronto North | 19.67 | Parry Sound | 30.56 | Etobicoke West | 8.44 |
| Hamilton Downtown | 19.36 | Chatham | 29.88 | Windsor Downtown | 8.36 |
| Windsor Downtown | 18.00 | Belleville | 29.63 | Hamilton West | 8.06 |
| Burlington | 17.74 | Kingston | 29.61 | Hamilton Mountain | 7.91 |
| Hamilton West | 17.63 | Newmarket | 29.13 | Toronto West | 7.86 |
| Windsor West | 16.89 | Hamilton Mountain | 28.61 | Toronto North | 7.59 |
| Brampton | 15.95 | Peterborough | 28.55 | Kitchener | 7.52 |

Table 1: Top 10 Cities with Highest Concentration of Each Pollutant

The table above is three separate tables each one for a respective pollutant, they are sorted in descending Mean and only the top 10 are being shown based on the overall Mean of the City, over all years from 2003 to 2022; as such, the cities with the high concentration will be listed first. Therefore, we can draw a few key observations about pollutant concentrations across these Ontario cities:

- The highest mean $NO_x$ readings (31.32 ppb) appear at Toronto West, followed closely by other Toronto stations (Toronto East, Toronto Downtown) and industrial/urban areas like Hamilton and Windsor.

- Port Stanley shows the highest $O_3$ levels (32.84 ppb), with other high concentrations at Tiverton, Grand Bend, and Parry Sound—generally smaller or semi-rural communities.

- Sarnia tops the $PM_{2.5}$ list (9.57 µg/m3), followed by Windsor (West and Downtown) and Hamilton stations, reflecting the influence of industrial facilities and cross-border pollution.

## 4.2 The Top 10 Regions and Years with Highest Concentration of Each Pollutant

In order to view the pollutant concentration changes by region, we grouped the cities by region based on the map of Ontario from the Ministry of Natural Resources and Forestry, and calculated the mean of the cities in each region grouped by year.

$NO_x$ Table (in ppb)

| Year | Region | Conc. |
|------|--------|-------|
| 2003 | Central Ontario | 31.45 |
| 2005 | Central Ontario | 28.17 |
| 2003 | Western Ontario | 26.98 |
| 2004 | Central Ontario | 26.94 |
| 2006 | Central Ontario | 23.80 |
| 2007 | Central Ontario | 21.68 |
| 2004 | Western Ontario | 21.55 |
| 2008 | Central Ontario | 20.07 |
| 2005 | Western Ontario | 20.00 |
| 2004 | Northern Ontario | 19.04 |

$O_3$ Table (in ppb)

| Year | Region | Conc. |
|------|--------|-------|
| 2010 | Western Ontario | 29.54 |
| 2022 | Western Ontario | 29.49 |
| 2007 | Eastern Ontario | 29.46 |
| 2010 | Eastern Ontario | 29.20 |
| 2021 | Western Ontario | 29.10 |
| 2018 | Eastern Ontario | 28.97 |
| 2012 | Western Ontario | 28.82 |
| 2016 | Western Ontario | 28.79 |
| 2013 | Western Ontario | 28.57 |
| 2008 | Eastern Ontario | 28.54 |

$PM_{2.5}$ Table (in µg/m3)

| Year | Region | Conc. |
|------|--------|-------|
| 2005 | Western Ontario | 9.48 |
| 2014 | Western Ontario | 9.27 |
| 2003 | Western Ontario | 8.93 |
| 2015 | Western Ontario | 8.79 |
| 2013 | Western Ontario | 8.72 |
| 2005 | Central Ontario | 8.56 |
| 2004 | Western Ontario | 8.44 |
| 2014 | Central Ontario | 8.33 |
| 2007 | Western Ontario | 8.30 |
| 2003 | Central Ontario | 8.25 |

Table 2: Top 10 Regions and Years with Highest Concentration of Each Pollutant

This table follows a similar format to the Table 1, listing the highest concentrations first. The concentrations of $NO_x$ (in ppb) and $PM_{2.5}$ (in µg/m3) were highest in the early to mid-2000s, particularly in Western and Central Ontario, and have since declined. Meanwhile, $O_3$ (in ppb) levels are highest in Western Ontario dominating the top 10 with Northern Ontario not occupying a single spot. The $O_3$ concentrations show minor decrease, likely reflecting $O_3$ long lifespan in the atmosphere.

## 4.3 The Top 10 Years with Highest Concentration of Each Pollutant

$NO_x$ Table (in ppb)

| Year | Conc. |
|------|-------|
| 2003 | 25.08 |
| 2004 | 23.42 |
| 2005 | 22.43 |
| 2006 | 17.87 |
| 2007 | 15.80 |
| 2008 | 15.04 |
| 2009 | 13.91 |
| 2011 | 12.36 |
| 2010 | 12.35 |
| 2012 | 11.23 |

$O_3$ Table (in ppb)

| Year | Conc. |
|------|-------|
| 2010 | 28.41 |
| 2021 | 27.97 |
| 2018 | 27.88 |
| 2012 | 27.85 |
| 2007 | 27.83 |
| 2022 | 27.83 |
| 2013 | 27.73 |
| 2016 | 27.67 |
| 2017 | 27.43 |
| 2008 | 27.41 |

$PM_{2.5}$ Table (in µg/m3)

| Year | Conc. |
|------|-------|
| 2005 | 8.36 |
| 2003 | 8.04 |
| 2014 | 7.96 |
| 2015 | 7.73 |
| 2013 | 7.61 |
| 2004 | 7.30 |
| 2006 | 7.04 |
| 2007 | 7.03 |
| 2021 | 6.92 |
| 2018 | 6.78 |

Table 3: Top 10 Years with the Highest Concentration

The table reveals that $NO_x$ levels peaked in 2003 – the first year of the dataset – and has steadily decreased since. Additionally, $PM_{2.5}$ concentrations were highest in the mid-2000s before decreasing, and $O_3$ levels, which peaked around 2010–2012, have remained relatively high before a minor decrease in later years.

# 5. Graphs

## 5.1 Line Charts of Yearly Concentration by Region
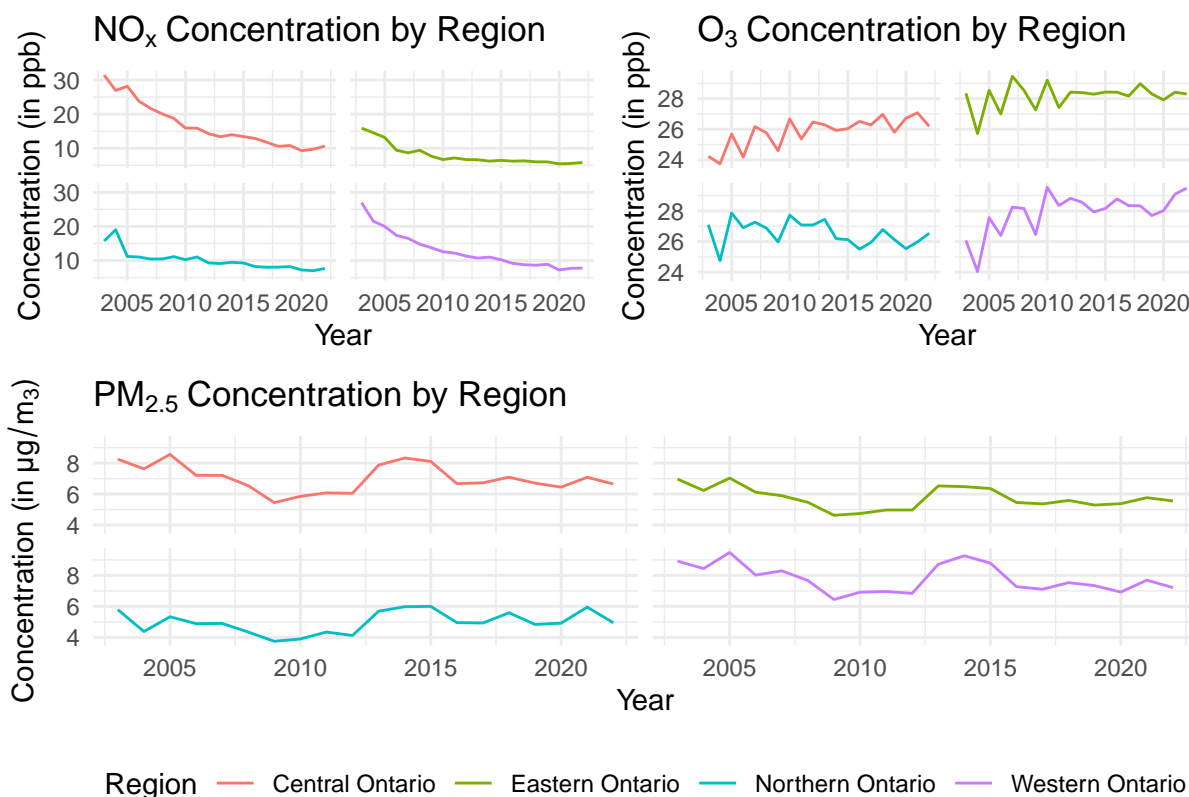


The Line Charts of Yearly Concentration by Region display the concentration of each pollutant faceted by region over 2003 till 2022. The x-axes represent the year and the y-axes represent the concentration of the pollutant at that year in its respective unit.

The graphs reveal a notable improvement in air quality over the years, with clear declines in both $NO_x$ and $PM_{2.5}$ levels across all regions, suggesting that emission controls and cleaner technologies have had a significant impact. $NO_x$ concentrations show a consistent downward trend from higher levels in the early years, converging towards lower values by 2020, while $PM_{2.5}$ levels exhibit a marked decrease, with the initial difference between regions decreasing over time. In contrast, $O_3$ concentrations appear relatively stable – with some fluctuations – which reflects the complex nature of ozone formation that depends on multiple factors such as sunlight, temperature, and precursor emissions.

## 5.2 Scatter Plot of Overall Pollutant Concentration by Year

Total Concentration by Year
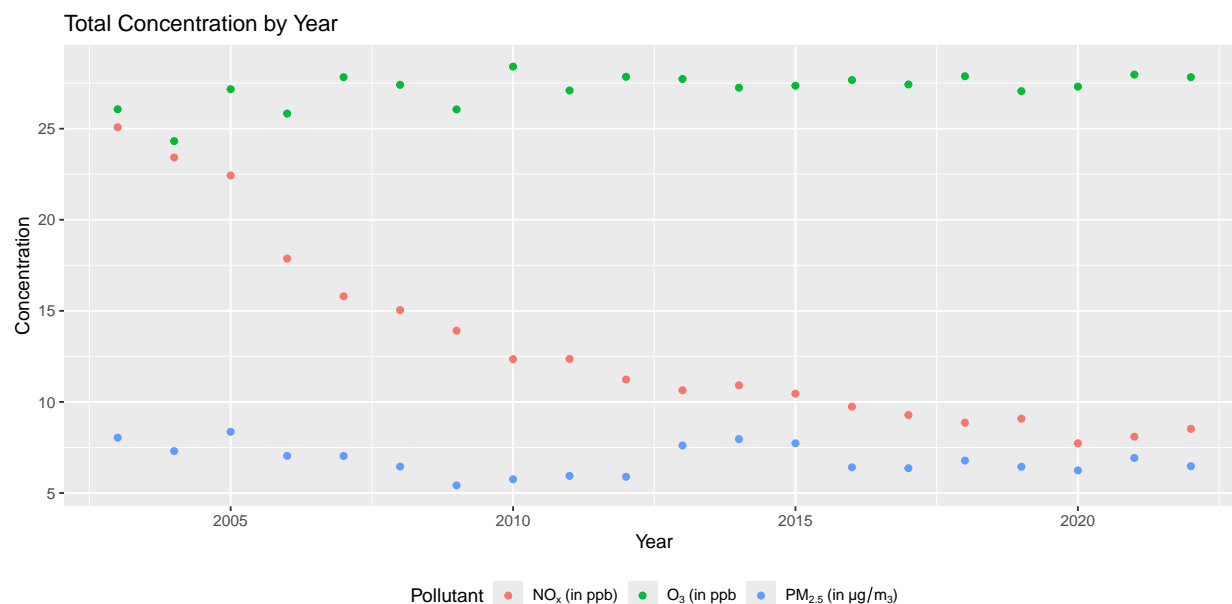


This Scatterplot of Overall Pollutant Concentration by Region displays the overall concentration, calculated by taking the mean of each year for all cities in the dataset, of each pollutant over 2003 till 2022. The x-axes represent the year and the y-axes represent the concentration of the pollutant at that year in its respective unit.

The plot reveals a downward trends of $NO_x$ until 2020, with concentrations appearing to drop from around 25 ppb to near 8 ppb in 2020. This substantial decrease suggests that measures aimed at reducing emissions through regulations and increased standards have been effectively implemented over the years. On the other hand, $PM_{2.5}$ and $O_3$ do not show a steady downward trend; instead, their concentrations fluctuate over the period analyzed. These fluctuations show the challenges in the efforts of controlling pollutants that are not directly emitted but are produced by chemical interactions in the atmosphere.

## 5.3 Heatmap

# 6. Hypothesis Testing

## Estimating the Average Concentration of pollutants in 2022

To estimate the average concentration of pollutants in 2022, we can use bootstrapping to find a 95% confidence interval of the pollutant concentration of the cities.

In order to use bootstrapping to find a 95% confidence interval, we will repeated take samples of the pollutants filtered for only the year 2022 and calculate the mean of each sample. This process of re-sampling will be repeated 1000 times to produce a sampling distribution for each pollutant. From the sample distribution, we can find 2.5% and 97.5% quantile.

```
## 95% confidence interval for NOX is [ 7.219514 , 10.0246 ]
## 95% confidence interval for O3 is [ 27.10182 , 28.6084 ]
## 95% confidence interval for PM25 is [ 6.140158 , 6.821632 ]
```

# 7. Bootstrapping

**Estimating the Average Concentration of Pollutants in 2022**

To estimate the average concentration of pollutants in 2022, we can use bootstrapping to find a 95% confidence interval of the mean pollutants of the cities.

In order to use bootstrapping to find a 95% confidence interval, we will repeated take samples of the pollutants filtered for only the year 2022 and calculate the mean of each sample. This process of re-sampling will be repeated 1000 times to produce a sampling distribution for each pollutant. From the sample distribution, we can find 2.5% and 97.5% quantile.

```
## 1 ) Mean NOX concentration of: 8.523611 , 95% confidence interval for NOX is [ 7.219514 , 10.0246 ]
## 2 ) Mean O3 concentration of: 27.83395 , 95% confidence interval for O3 is [ 27.10182 , 28.6084 ]
## 3 ) Mean PM25 concentration of: 6.474474 , 95% confidence interval for PM25 is [ 6.140158 , 6.821632
```

Based on the result, we have:

1) A mean $NO_x$ concentration of 8.523611. A 95% confidence interval of [7.219514 , 10.0246] for the concentration of $NO_x$. This means that out of 100 samples, 95 samples will have a mean $NO_x$ concentration between [7.219514 , 10.0246].

2) A mean $O_3$ concentration of 27.83395. A 95% confidence interval of [27.10182 , 28.6084] for the concentration of $O_3$. This means that out of 100 samples, 95 samples will have a mean $O_3$ concentration between [27.10182 , 28.6084].

3) A mean $PM_{2.5}$ concentration of 6.474474. A 95% confidence interval of [6.140158 , 6.821632] for the concentration of $PM_{2.5}$. This means that out of 100 samples, 95 samples will have a mean $PM_{2.5}$ concentration between [6.140158 , 6.821632].

# 8. Regression

## 8.1 Analyzing the Relationship Between Pollutants and Year

In order to analyze the relationship of a pollutant with other pollutants and the year, we would treat the year and two pollutants as continuous variables to fit a regression. For example, if we were examining the relationship between the overall yearly $NO_x$ concentration and the year, the $O_3$ concentration and the overall yearly $PM_{2.5}$ concentration, we will take the logarithm of the $NO_x$ concentration as the dependent variable and the year, the overall yearly $O_3$ concentration and the overall yearly $PM_{2.5}$ concentration all as the independent variable.

```
##
## Call:
## lm(formula = formula(log(Conc.NOX) ~ Year + Conc.PM25 + Conc.O3),
##     data = Combined_Yearly_Mean)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.137814 -0.034320 -0.005249  0.022683  0.173198
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 105.538569    7.485517  14.099 1.93e-10 ***
## Year          -0.050618    0.003861 -13.110 5.65e-10 ***
## Conc.PM25      0.062989    0.023487   2.682   0.0164 *
## Conc.O3       -0.058474    0.023494  -2.489   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08013 on 16 degrees of freedom
## Multiple R-squared:  0.9584, Adjusted R-squared:  0.9506
## F-statistic: 122.9 on 3 and 16 DF,  p-value: 2.925e-11
```

From the linear model, it suggests that their is a strong relationship between the logarithm of the $NO_x$ concentration and the year, the $PM_{2.5}$ concentration and the $O_3$ concentration. In particular, by analyzing the p-values of each parameter, it can be seen that the year parameter is especially significant and that the mean $PM_{2.5}$ and $O_3$ levels have moderate significance on the logarithm of the mean $NO_x$ concentration in comparison. The coefficient of determination ($R^2$) of 0.9584 suggests that 95.84% of the variables in the logarithm of the $NO_x$ concentration can be explained by the model.

## 8.2 Interpreting regression parameters

Here, the regression parameters refer to the coefficients of the independent variables of the model.

1) Intercept:
   - The intercept coefficient expresses the value of the dependent variable (logarithm of the $NO_x$ concentration) when all other features are equal to 0. In this case, an intercept of ~105.54 suggests that the logarithm of the $NO_x$ concentration is equal to ~105.54 when all independent variables are equal to 0.

2) Year, Mean $PM_{2.5}$ Concentration, and Mean $O_3$ Concentration:
   - The coefficient for Year expresses the linear effect of the Year parameter; specifically, a coefficient of -0.050621 suggests that as the Year increases by 1, the logarithm of the $NO_x$ concentration decreases by 0.050621. Similarly, a coefficient for the $PM_{2.5}$ level of 0.062955 and a coefficient for the $O_3$ level of -0.058456 suggests that as the $PM_{2.5}$ concentration increases by 1 and the $O_3$ level increases by 1 there will be an increase in the logarithm of the $NO_x$ level by 0.062955 and a decrease in the logarithm of the $NO_x$ level by 0.058456, respectively.

Overall, this model suggests a strong linear relationship between the logarithm of the mean $NO_x$ level and the year, the mean $PM_{2.5}$ level and the mean $O_3$ level. In other words, this model also explains that the $NO_x$ concentration has an exponential relationship with the features used, in particular an exponential decay relationship with the year.

# 9. Cross Validation

Previously, we analyzed the regression parameters to see how well our model performed on our data set. In this case, we will further analyze the relationships of pollutants with each other and the year by performing cross validation.

Specifically, we will analyze the relationship between $PM_{2.5}$ and the other pollutants alongside the year by using k-fold cross validation. To perform k-fold cross validation, we will split the data set into k even pieces and use each $i^{th}$ fold to check the validity of our model that is trained on the remaining $k-1$ folds by taking predictions using the data in the $i^{th}$ fold. This process will be repeated for each fold of the data set, a mean squared error (MSE) will be calculated for each fold, and an average of the MSE's will be taken at the end.

```
## The average MSE is: 0.5452378
```

**Conclusion on Results**

The average MSE from the k-fold cross validation was 0.5452296 which is relatively low and maybe indicate the moderate predicting power of the linear model where the $PM_{2.5}$ level was the dependent variable with year, $NO_x$ concentration and the $O_3$ concentration were the independent variables. This suggests that there is likely a linear relationship between the $PM_{2.5}$ concentration, the year, $NO_x$ concentration and $O_3$ concentration.

# 10. Summary of Research

# 11. Appendix

```r
# Loading libraries
library(tidyverse)
library(readxl)
library(patchwork)
library(lubridate)
library(kableExtra)
library(ggplot2)
library(lubridate)
library(knitr)
library(stringr)
library(tidygeocoder)
library(gridExtra)
# Reading in all data excel files
NOX <- read_excel("./data/NOX.xlsx", skip = 3)
O3 <- read_excel("./data/O3.xlsx", skip = 3)
PM25 <- read_excel("./data/PM25.xlsx", skip = 5)
# 1. Description of the Dataset Columns
names(NOX)
# 2. Background of the Data: No code for this section
# 3. Overall Research Question
# Create a dataframe that contains the mapping of each city to it's region
city_to_region <- data.frame(City = c(
    # Southwest
    "Windsor West","Windsor Downtown","Windsor North","Grand Bend","London",
    "Chatham","Sarnia","Brantford","Kitchener",
    "St. Catharines","St. Catherines","Port Stanley","Essex","Hamilton Mountain",
    "Hamilton Downtown","Hamilton West","Guelph",
    # GTA (Greater Toronto Area) - including Halton/Peel/York/Durham
    "Toronto Downtown","Toronto East","Toronto North","Toronto West","Burlington",
    "Oakville","Oshawa","Brampton","Mississauga",
    "Newmarket","Milton","Etobicoke West","Stouffville","Barrie","Parry Sound",
    "Peterborough",
    # Eastern Ontario
    "Ottawa Downtown","Ottawa Central","Kingston","Tiverton","Belleville",
    "Cornwall","Dorset","Petawawa","Morrisburg",
    # Northern Ontario
```

```r
    "Sudbury","Sault Ste. Marie","North Bay","Thunder Bay"
  ),Region = c(
    # Southwest
    "Western Ontario","Western Ontario","Western Ontario","Western Ontario",
    "Western Ontario","Western Ontario","Western Ontario","Western Ontario",
    "Western Ontario",
    "Western Ontario","Western Ontario","Western Ontario","Western Ontario",
    "Western Ontario","Western Ontario","Western Ontario","Western Ontario",
    # GTA
    "Central Ontario","Central Ontario","Central Ontario","Central Ontario",
    "Central Ontario","Central Ontario","Central Ontario","Central Ontario",
    "Central Ontario","Central Ontario","Central Ontario","Central Ontario",
    "Central Ontario",
    # Central Ontario
    "Central Ontario",
    "Central Ontario",
    # Eastern Ontario
    "Eastern Ontario","Eastern Ontario","Eastern Ontario","Eastern Ontario",
    "Eastern Ontario","Eastern Ontario","Eastern Ontario","Eastern Ontario",
    "Eastern Ontario","Eastern Ontario",
    # Northern Ontario
    "Northern Ontario","Northern Ontario","Northern Ontario","Northern Ontario"
  ),stringsAsFactors = FALSE)
# Mutate datasets to have numeric means, years, and only contain data from 2003 onwards
NOX = NOX %>% mutate(Mean = as.numeric(Mean)) %>%
  filter(Year >= 2003)
O3 = O3 %>% mutate(Mean = as.numeric(Mean)) %>%
  filter(Year >= 2003)
PM25 = PM25 %>% mutate(Mean = as.numeric(Mean)) %>%
  mutate(Year = as.integer(Year)) %>%  filter(Year >= 2003)
# Remove all rows with a non-numeric mean
O3 = O3 %>% drop_na(Mean)
NOX = NOX %>% drop_na(Mean)
PM25 = PM25 %>% drop_na(Mean)
# Join all data to city-region-mapping to convert cities into regions
NOX = NOX %>% left_join(city_to_region, by = "City")
O3 = O3 %>% left_join(city_to_region, by = "City")
PM25 = PM25 %>% left_join(city_to_region, by = "City")

# Make tables
# Mean per years for that particulate matter
NOX_City_mean <- NOX %>% group_by(City) %>% summarise(Conc. = mean(Mean))
O3_City_mean <- O3 %>% group_by(City) %>% summarise(Conc. = mean(Mean))
PM25_City_mean <- PM25 %>% group_by(City) %>% summarise(Conc. = mean(Mean))

# 4. Tables
# 4.1 The Top 10 Cities With the Highest Pollutant Concentration
# Display tables of most relevant cities from datasets arranged in descending order
NOX_table <- kable(NOX_City_mean %>% mutate(Conc. = round(Conc., 2)) %>%
                   arrange(desc(Conc.)) %>% head(10), "latex")
O3_table <- kable(O3_City_mean %>% mutate(Conc. = round(Conc., 2)) %>%
                   arrange(desc(Conc.)) %>% head(10), "latex")
PM25_table <- kable(PM25_City_mean %>% mutate(Conc. = round(Conc.,2)) %>%
```

```r
                        arrange(desc(Conc.)) %>% head(10), "latex")
# Summarise data by determining the mean of each region over all years
NOX_Region = NOX %>% group_by(Year, Region) %>% summarise(Conc. = mean(Mean))
O3_Region = O3 %>% group_by(Year, Region) %>% summarise(Conc. = mean(Mean))
PM25_Region = PM25 %>% group_by(Year, Region) %>% summarise(Conc. = mean(Mean))

# 4.2 The Top 10 Regions and Years with Highest Concentration of Each Pollutant
# Display tables from datasets of the most relevant regions arranged in descending order
NOX_tab  <- kable(NOX_Region  %>% mutate(Conc. = round(Conc., 2)) %>%
                    arrange(desc(Conc.)) %>% head(10), "latex")
O3_tab   <- kable(O3_Region   %>% mutate(Conc. = round(Conc., 2)) %>%
                    arrange(desc(Conc.)) %>% head(10), "latex")
PM25_tab <- kable(PM25_Region %>% mutate(Conc. = round(Conc., 2)) %>%
                    arrange(desc(Conc.)) %>% head(10), "latex")

# 4.3 The Top 10 Years with Highest Concentration of Each Pollutant
# Summarise data by determining mean of all regions/ctiies over every year
NOX_Yearly_Mean = NOX %>% group_by(Year) %>% summarise(Conc. = round(mean(Mean), 2))
O3_Yearly_Mean = O3 %>% group_by(Year) %>% summarise(Conc. = round(mean(Mean),2))
PM25_Yearly_Mean = PM25 %>% group_by(Year) %>% summarise(Conc. = round(mean(Mean) ,2))
# Display tables from datasets of the yearly mean of each respective particle
NOX_tab  <- kable(NOX_Yearly_Mean  %>% mutate(Conc. = round(Conc., 2)) %>%
                    arrange(desc(Conc.)) %>% head(10), "latex")
O3_tab   <- kable(O3_Yearly_Mean   %>% mutate(Conc. = round(Conc., 2)) %>%
                    arrange(desc(Conc.)) %>% head(10), "latex")
PM25_tab <- kable(PM25_Yearly_Mean %>% mutate(Conc. = round(Conc., 2)) %>%
                    arrange(desc(Conc.)) %>% head(10), "latex")

# 5. Graphs
# 5.1 Line Charts of Yearly Concentration by Region
# Display graphs of concentration of each respective particle over the years per region
plot_NOX  <- ggplot(NOX_Region, aes(x = Year, y = Conc., col = Region)) +
  geom_line() +
  facet_wrap(~ Region) +
  labs(title = expression(NO[x] ~ "Concentration by Region"), x = "Year",
       y = "Concentration (in ppb)") +
  theme_minimal() +
  theme(legend.position = "none", strip.text = element_blank())

plot_O3 <- ggplot(O3_Region, aes(x = Year, y = Conc., col = Region)) +
  geom_line() +
  facet_wrap(~ Region) +
  labs(title = expression(O[3] ~ "Concentration by Region"), x = "Year",
       y = "Concentration (in ppb)") +
  theme_minimal() +
  theme(legend.position = "none", strip.text = element_blank())

plot_PM25 <- ggplot(PM25_Region, aes(x = Year, y = Conc., col = Region)) +
  geom_line() +
  facet_wrap(~ Region) +
  labs(title = expression(PM[2.5] ~ "Concentration by Region"), x = "Year",
       y = expression("Concentration (in" ~ µg/m[3] *")")) +
  theme_minimal() +
```

```r
    theme(legend.position = "none", strip.text = element_blank()))


(plot_NOX | plot_O3) / plot_PM25 +
  plot_layout(guides = "collect") & theme(legend.position = "bottom")


# 5.2 Scatter Plot of Overall Pollutant Concentration by Year
# Store all yearly means for each respective particle
Yearly_Means = list(NOX_Yearly_Mean, PM25_Yearly_Mean, O3_Yearly_Mean)
# Merge all data by year
Combined_Yearly_Mean = reduce(Yearly_Means, full_join, by = "Year")
# Rename columns for ease of use
Combined_Yearly_Mean = Combined_Yearly_Mean %>%
  rename(Conc.NOX = Conc..x, Conc.PM25 = Conc..y, Conc.O3 = Conc.)
# Reshape data into longer format
Graphing_Yearly_Means <- Combined_Yearly_Mean %>%
              pivot_longer(cols = starts_with("Conc."),
              names_to = "Pollutant",
              values_to = "Value")
# Graph yearly mean concentration of each particle over time
ggplot(Graphing_Yearly_Means,
      aes(x = Year, y = Value,
          col = factor(Pollutant,
                      levels = c("Conc.NOX", "Conc.O3", "Conc.PM25")))) +
      geom_point() +
      labs(col = "Pollutant", y = "Concentration", title = "Total Concentration by Year ") +
  scale_color_discrete(labels = c(expression(NO[x] ~ "(in ppb)"),
      expression(O[3] ~ "(in ppb"), expression(PM[2.5] ~ "(in" ~ µg/m[3] *
  ")"))) + theme(legend.position = "bottom")


# 5.3 Heatmap of Mean Particle Concentration over all Time Accross Ontario

# 6. Hypothesis Testing
# Estimating the average of the Mean NOX over year X
# Set random seed so that results are reproduceable
set.seed(123)

# Filter datasets for only the year 2022
NOX_2022 <- NOX %>% filter(Year == 2022) %>% select(Mean)
O3_2022 <- O3 %>% filter(Year == 2022) %>% select(Mean)
PM25_2022 <- PM25 %>% filter(Year == 2022) %>% select(Mean)

# Bootstrap function that will repeated take samples
boot_function <- function() {
  boot_sample_NOX = sample_n(NOX_2022, nrow(NOX_2022), replace=T)
  boot_sample_O3 = sample_n(O3_2022, nrow(O3_2022), replace=T)
  boot_sample_PM25 = sample_n(PM25_2022, nrow(PM25_2022), replace=T)
  return (c(mean(boot_sample_NOX$Mean), mean(boot_sample_O3$Mean),
          mean(boot_sample_PM25$Mean)))
}

# Combine all bootstrap means into a dataframe
bootstrap_bar = as.data.frame(replicate(1000, boot_function()))
columns = c("NOX", "O3", "PM25")
```

```r
for (i in 1:3) {
  # Flatten rows of dataframe into a column vectors
  vec = as.vector(as.matrix(bootstrap_bar[i,]))
  # Find the 2.5% and 97.5% quantile
  conf = quantile(vec, c(0.025, 0.975))
  cat("95% confidence interval for", columns[i], "is", "[", conf[1], ",",
      conf[2], "]", "\n")
}


# 7. Bootstrapping
# Estimating the Average Concentration of Pollutants in 2022
# Estimating the average of the Mean NOX over year X
# Set random seed so that results are reproduceable
set.seed(123)

# Filter datasets for only the year 2022
NOX_2022 <- NOX %>% filter(Year == 2022) %>% select(Mean)
O3_2022 <- O3 %>% filter(Year == 2022) %>% select(Mean)
PM25_2022 <- PM25 %>% filter(Year == 2022) %>% select(Mean)

# Bootstrap function that will repeated take samples
boot_function <- function() {
  boot_sample_NOX = sample_n(NOX_2022, nrow(NOX_2022), replace=T)
  boot_sample_O3 = sample_n(O3_2022, nrow(O3_2022), replace=T)
  boot_sample_PM25 = sample_n(PM25_2022, nrow(PM25_2022), replace=T)
  return (c(mean(boot_sample_NOX$Mean), mean(boot_sample_O3$Mean), mean(boot_sample_PM25$Mean)))
}

# Combine all bootstrap means into a dataframe
bootstrap_bar = as.data.frame(replicate(1000, boot_function()))
columns = c("NOX", "O3", "PM25")
datasets = c(NOX_2022, O3_2022, PM25_2022)
for (i in 1:3) {
  # Flatten rows of dataframe into a column vectors
  vec = as.vector(as.matrix(bootstrap_bar[i,]))
  # Find the 2.5% and 97.5% quantile
  conf = quantile(vec, c(0.025, 0.975))
  cat(i, ")", "Mean", columns[i], "concentration of:", mean(datasets[i]$Mean), ",",
  "95% confidence interval for", columns[i], "is", "[", conf[1], ",", conf[2], "]", "\n")
}


# 8. Regression
# 8.1 Analyzing the Relationship Between Pollutants and Year
NOX_other_pollutant_model = lm(formula(log(Conc.NOX)~Year + Conc.PM25 + Conc.O3),
                               data = Combined_Yearly_Mean)
summary(NOX_other_pollutant_model)
# 8.2 Interpreting regression parameters: No code for this section
# 9. Cross Validation
set.seed(123)
k_fold_mse = rep(0, 5)
# Split data into 5 folds
Combined_Yearly_Mean <- Combined_Yearly_Mean %>%
  mutate(group_ind = sample(c(1:5),size=nrow(Combined_Yearly_Mean), replace=TRUE))
```

```r
for (i in 1:5) {
  # train on fold all fold not i then test on fold i
  train_set <- Combined_Yearly_Mean %>% filter(group_ind != i)
  test_set <- Combined_Yearly_Mean %>% filter(group_ind == i)
  NOX_other_pollutant_model = lm(formula(Conc.PM25~Year + Conc.NOX + Conc.O3), data = train_set)
  pred = predict(NOX_other_pollutant_model, newdata=test_set)
  mse = mean((test_set$Conc.PM25 - pred)^2)
  k_fold_mse[i] = mse
}
cat("The average MSE is:", mean(k_fold_mse), "\n")
# 10. Summary of Research: No code for this section
```