AIR POLLUTION IN SEOUL

Sophie Youk (sy5qm)

STAT 6130

4 May 2020

BACKGROUND & DATA

Background

- Data obtained on the website Kaggle
  - *Air Pollution in Seoul.* https://www.kaggle.com/bappekim/air-pollution-in-seoul
- Seoul Metropolitan Government (SMG) has collected and provided many public data including air pollution information. There are several stations measuring air pollution in South Korea including Seoul

Data

- The 25 mearing stations in Seoul has measured air pollutants (SO_2, NO_2, O_3, CO, PM_{10}, $PM_{2.5}$) hourly
  - Used data by 8 hours since I can't run the full dataset in my computer
- Datasets from 12 AM on 1 January 2017 to 11 PM on 31 December 2019
- Every station has its own code (101 to 125)
- Address, latitude, and longitude indicate where the stations are located

Variables:

- Measurement date: Measurement date and time
- Station code: Measuring station code
- Address: Address of measuring station
- Latitude: Latitude of address
- Longitude: Longitude of address
- SO2: Sulfur dioxide
- NO2: Nitrogen dioxide
- O3: Ozone
- CO: Carbon monoxide
- PM10: Particulate matter
- PM2.5: Particulate matter

PROBLEM, MODEL, METHOD & ANALYSIS

Problem

- Relationship between 4 air pollutants and PMs
- How the 4 air pollutants are related to the location (latitude, longitude, address, or stations code) and time (measurement date)

Model

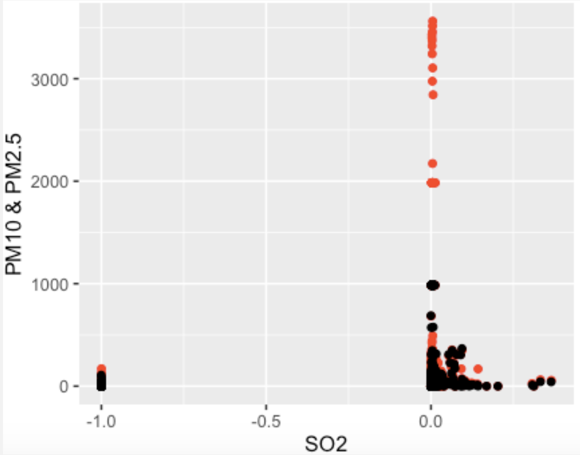
- Multivariate Regression Model 1
  - Responses: Measurement date, Station code, Latitude, Longitude
  - Predictors: SO_2, NO_2, O_3, CO, PM_{10}, $PM_{2.5}$
- Multivariate Regression Model 2
  - Responses: PM_{10}, $PM_{2.5}$
  - Predictors: SO_2, NO_2, O_3, CO

Method

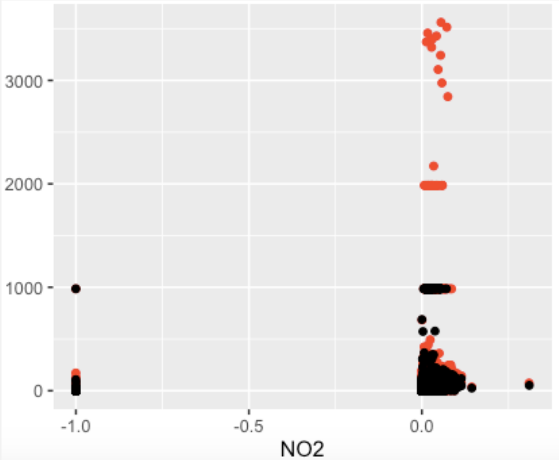
- Exploratory Data Analysis
  - Plot each of 4 air pollutants vs. PM_{10} and $PM_{2.5}$
- Clustering
  - Used data by one week since I can't run the full dataset in my computer
  - Focus on SO_2, NO_2, O_3, CO pollutants

REMARKABLE PLOTS

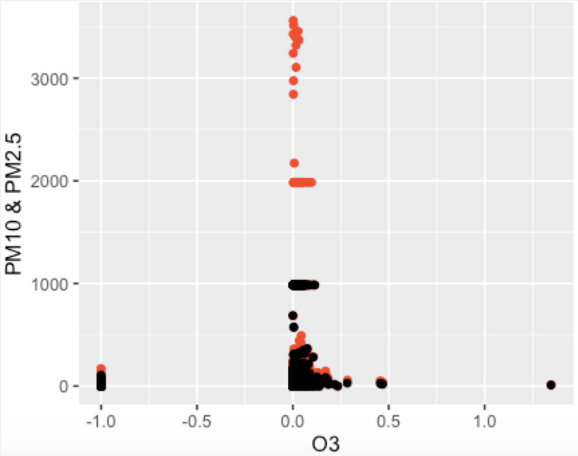
- Figure 3: SO_2 vs. PM_{10} & $PM_{2.5}$



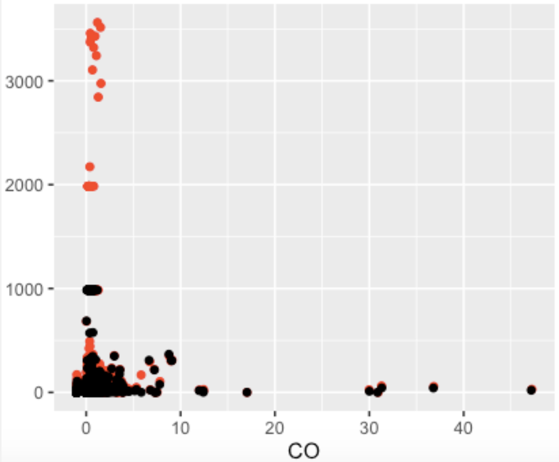
- Figure 4: NO_2 vs. PM_{10} & $PM_{2.5}$

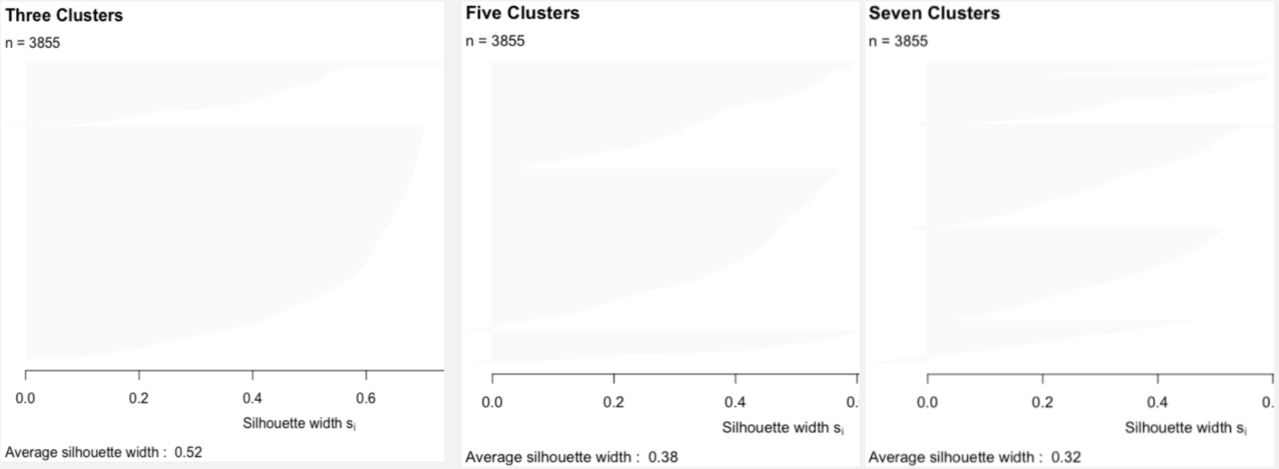


- Figure 4: O_3 vs. PM_{10} & $PM_{2.5}$



- Figure 4: CO vs. PM_{10} & $PM_{2.5}$





- Figure 8: Average Silhouettes under 3, 5, and 7 Clusters

CONCLUSION & DISCUSSION

Conclusion

- Only NO_2 out of other pollutants and PMs was positively related to latitude
  - NO_2 values were recorded higher in northern counties.
- SO_2 was negatively related to PM_{10} and $PM_{2.5}$ values
  - When SO_2 values were high, values of PM_{10} and $PM_{2.5}$ were recorded low
- 51 datasets have very high PM_{10} values higher than approximately 2,000
  - According to the code information, four stations (116, 117, 121, and 122) are located southern or southwestern part of Seoul

Discussion

- One county has only one measuring station
  - but the size, population, and number of factories are very various and random
- Counties with high density of population and factories may have higher air pollutants values than other counties even though they have fewer population and factories in reality.
- More information of predictors/factors can improve analyzing the air pollution data.
- Rain sometimes decreases the air pollution measurements
  - Better to analyze data collected in the same condition of weather

Sophie Youk (sy5qm)