Investigation of Air Pollution in NYC

Yutong Wu

12/8/2019

1. Introduction

1.1 Aim of the Project

NYC Health states that air pollution is one of the most significant environmental problem in New York and causes more than 3000 death every year. This is partially accredited to the prosperous tourism, relatively high population density, and heavy traffic in the city. Besides, as NYC Health demonstrates, traffic contributes to 17% of all emissions.

Intrigued from this fact, the study aims to help the government better predict and monitor the air quality as well as control the key factors leading to air pollution, by evaluating the relationship between air pollution and factors, including traffic, air pollutant, and weather, and investigating how those factors contribute to the overall level air pollution in New York.

1.2 Project Roadmap

First, the air quality within different counties of New York State in 2016 is summarized and compared. Then, based on the result that New York (NY) contributes most to the pollution level, the focus is narrowed down towards NY City. After that, a Simple Linear Regression is ran to examine the overall relationship between traffic and pm 2.5. Based on this, air pollutant and weather factors are included to build 3 regression models using machine learning. Finally, the optimal prediction model will be selected, providing a simplified regression model to estimate the air pollution; and the relationship between different

predictors will be analyzed, to better understand the direct and indirect patterns of those factors in contributing to air pollution.

2. Problem Statement and Background

2.1 Aim of Analysis

In this study, I analyze the traffic, air pollutant, and meteoroManhattan factors within New York City in 2016. Using these data, the analysis aims to investigate how traffic, air pollutant, and weather condition affect air quality by adopting a combination of statistic learning models to (1) evaluate how different features synergistically affect the pm2.5 concentration and interact with each other; (2) predict future air pollution level; (3) identify predictors that are most important. The analysis results are expected to guide policy makers better understand the significance in controlling traffic to improve air quality and provide a simplified prediction model to monitor and predict the air pollution level.

2.2 Related Work Review

Rybarczyk and Zalakeviciute (2017) conducted a similar project aiming to provide an affordable air pollution predicted model based on machine learning method for developing countries. The authors also used transportation, weather, and air pollutant variable as predictors and PM2.5 as outcome. However, their study mainly focus on discussing how to accurately predict Air Pollution using affordable sources and devices for developing counties, rather than evaluating the relationship between features and outcome in an air pollution model for a specific city in developed country.

3. Data

3.1 Overview and Data Sources

The unit of observation is New York City (NYC). Daily average concentration of PM2.5 in 2016 is selected as the outcomes representing the level of air pollution. Besides, there are three types of predictors, traffic amount, air pollutant density, and meteorological factors. The restructured traffic data represents the daily amount of vehicles (traffic amount) traveled in NY city. Air pollutants include the daily concentration of SO2, CO, NO2, and Ozone and is collected from website of U.S. Environmental Protection Agency. Meteorological factors involve daily Temperature in Celsius (tempm), Dew point in Celsius (dewptm), Humidity percentage (hum), Wind speed in kph (wspdm), and Pressure in mBar (pressurem).

Regarding the data source, the air pollutant data is collected from the website of U.S. Environmental Protection Agency (EPA). Weather data of NYC in 2016 is downloaded from Kaggle (kaggle). Traffic data is gathered from New York Government official website. (NY Gov)

3.2 Data Wrangling

The data wrangling process for this study can be summarized as five parts: a) restructure data sets such as selecting related variables; b) transform the data type such as from character to date type. c) uniform the unit of observation; d) merge different data sets of predictors and outcome; e) clean missing value;

Specifically, regarding the traffic data, the raw data describes the hourly amount of vehicles passing through the tunnels and bridges operated by the MTA (Metropolitan Transportation Authority) in the NY city from 2012 to 2019. The first step is to transform the Date variable from character to date type using *mutate* and *as.Date* functions and then select the data from 2016/1/1 to 2016/12/31. Then, a new variable total_vehicles is created by adding up

the ETC and cash column using *mutate*; besides, out of the 2 direction, only direction "in" is filtered to avoid double count and only plaza ID from 1 to 11 are filtered to represent NYC. Finally, the unit of analysis is transformed to NYC through adding up all hourly vehicle volume passing different tunnels by date using *group_by* and *sum* function.

For the air pollutant data, data sets of 5 type air pollutants are separately acquired from EPA website. The first step is to restructure each data set using *select* function in *dplyr* package, to keep the variables that will be imputed in analytic models in order to reduce potential issues in joining the different datasets later. The second step is to join the restructured data sets together using *full_join* by Date, County, and Site variables. I use *full_join* here since want to keep all the observation. After that, missing value was dropped.

For the meteorological data, row data of hourly weather information in New York city are used. Similarly, first step is to select the relevant variables and transformed the date variable into date class. Different from daily-based pollutant data, weather data is hourly base. Thus the next step is to transform the weather data into daily base by summarizing the mean of the hourly data for each day. I average the hourly data instead of adding them up as did for traffic data because the traffic data is measured by volumn. After that, the unit of analysis is transformed to NYC by filtering New York, Bronx, Kings, and Queens from the previous emission data set and running the air pollutant average value of the 4 counties for each day in 2016.

Finally, the 3 type of variable was merged by *full_join* and missing value was dropped using *na.omit* function.

4. Analysis

4.1 Method and Tools

Primary analytic methods include correlation analysis, single linear regression, and machine learning for multivariable regression. Three models are generated for machine learning regression, including linear regression, K-nearest neighbor(knn), and random forest(rf). Statistic metrics are adopted to assess the accuracy of these models. Specifically, Rsquared is used to examine the predictability. To evaluate the goodness of fit, since RMSE tends to be larger than MAE as the sample size getting larger and is more sensitive to outliers (Chai&Draxler, 2014), I use RMSE instead of Mean Absolute Error (MAE) for conservative purpose and expect a good RMSE as lower than 0.2. Data visualization is also adopted to demonstrate analysis and results, with specific tools such as table (*kable*), scatter plots(*ggscatter*), and dotplot.

4.2 Analysis Process

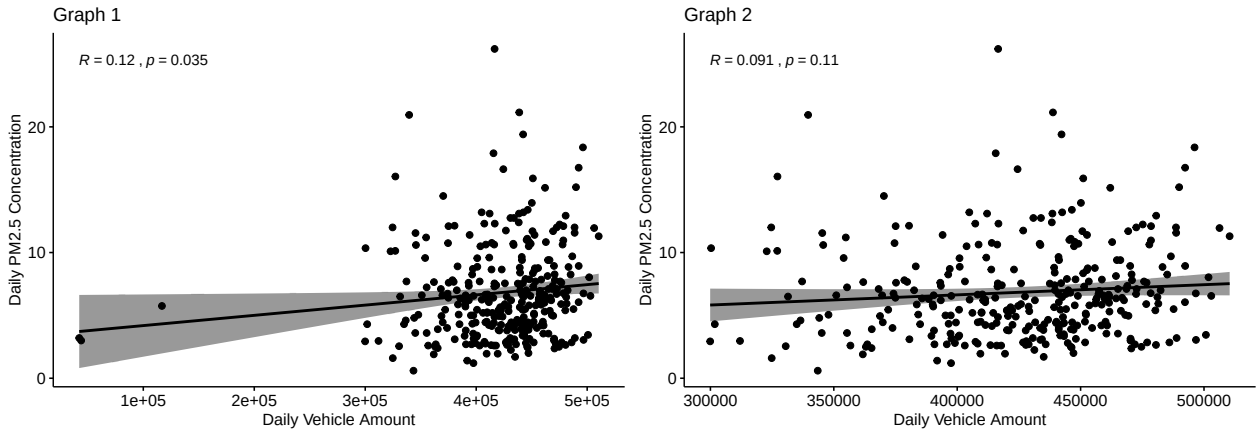
4.2.1 Air Pollution Performance Evaluation by Counties in NY State

I start by taking an overview of the PM2.5 level among 10 counties of NY State in 2016. This is achieved by summarizing the mean of daily PM2.5 for each county through the year using *summarize* and *mean* and ranking the value by *arrange*. The result is then demonstrated as table using `_kable_`. The result reveals that 4 counties in New York City is most polluted. Among those counties, Manhattan has the highest average daily PM2.5 value. Based on this result, the following analysis will focus on NY City.

4.2.2 Simple Linear Regression for Air Pollution and Traffic

In the next stage, the relationship between PM2.5 and Traffic is analyzed by adopting simple linear regression method. Firstly, I use *cor()* function to compute the Pearson's correlation

coefficient of estimated daily vehicles volume and the PM2.5 concentration. The result is 0.1185, indicating that there is a small positive correlation between traffic and pm2.5. To examine this result, I used *ggscatter* in *inggpubr* package to visualize the relationship by scatter plots (Graph 1), which shows that the outliers affect the accuracy of the correlation. To deal this the outliers effect, I use *filter* to exclude the outliers and re-run the correlation result by the same procedure (Graph 2). The new result suggests a slightly lower coefficient of approximately 0.1. One of the possible explanations why the statistic result is weaker than expected is that other relationships such as from hidden variables drive different forces toward the outcome. Thus, more variables are included in the following analysis.

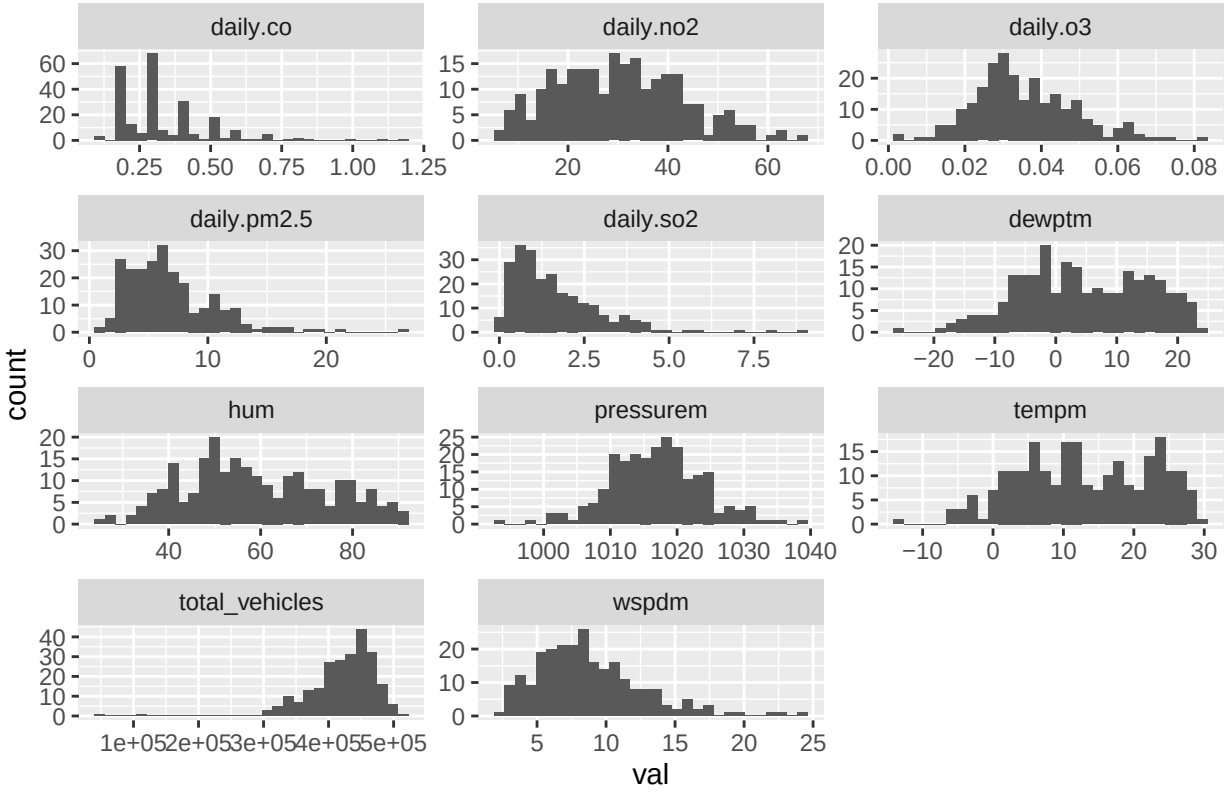


4.2.3 Multivariables Machine Learning Model

Air pollutant and meteorological factors are now introduced as new independent variables in addition to the traffic predictor. To predict the air pollution and to examine the relationship between features, machine learning technique is adopted. To achieve this, the first step is to split 75% of the data as training data and 25% as testing data, applying *createDataPartition* in *caret* package. After that, a thorough examination on training data is conducted through observing the distribution of each variable by applying *ggplot* (Shown Below). *facet_wrap* in *ggplot* is used to create subplots for each variable. Through examination, three issues are detected. First, the distribution of the outcome variable pm2.5 and predictor SO2 is skewed, violating the normal distribution as-

sumption. Second, the scale range of variables considerably varies. Variables with larger range will outweigh those with lower scale, causing the value not comparable and disturbing the accuracy of the regression model(Lakshmanan, 2019). Third, there are outliers exist in SO2 and vehicle variable.

Graph 3. Data Examination

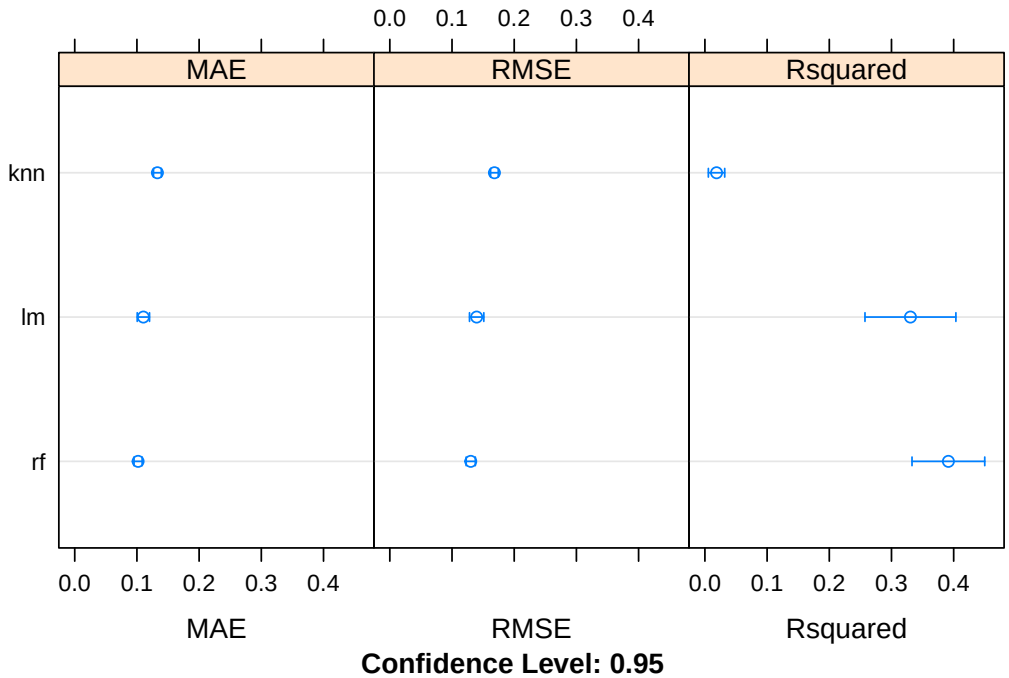


Targeting on those issues, a data pre-processing process is conducted. To standardize the distribution, I manipulate the skewed variables into log value using *mutate* and *log* and store in new train and test objective. After that, a recipe is built as a container that holds data pre-processing steps to feed into model (Thoen, 2018), including step to normalizing the range of data here using *step_range*. Then, I apply the recipe to the train and test data set respectively to obtain the final training and testing data set.

Next, before building model, the first thing is to randomly split the training data into 5 groups by *createFolds* in *caret* package for cross validation purpose in machine learning. The 5-group list are stored as an index and then applied into *trainControl* function in

order to set up the control condition that can be applied across different model. With data and environment settled, 3 different models are ran by applying *train* function, linear model, k-nearest neighbors, and random forest model. To select the optimal one, a list is created by *list* function and is applied to *resamples* and *dotplot* function to summarize performance by different metrics (MAE, RMSE, Rsquared) and to visualize the comparison result. The result (figure 4) shows that Ramdon Forest (rf) model performs best with the lowest RMSE of 0.13 and with the higest Rsquared of 0.39. 0.13 means that square root of the residuals variance under rf model is only 0.13 and the fitness is better than other models. 0.39 indicates that the model explains 39% of the variability of the response data around its mean thus the ability to predict is relatively higher (Rieuf, 2017). To test the accuracy of the optimal model, I predict the outcome using the test data under rf model by *predict* function and calculate the corresponding RMSE. The RMSE based on testing data is 0.11, which is lower than 0.2 and indicates a good fitness.

Graph 4. Comparison Between Models



Finally, based on the optimal model, I compare the importance of predictors by *varImp* function which returns the relative importance scores of each features. Besides, a correlation

coefficients metric between each variables are generated using *cor* function and visualized by *ggcorrplot* function. The result will be intepreted in details in the following section.

5. Summary of Results

First, the overview of air Pollution Performance within different counties tells us that the top 4 polluted counties are from NY City, including New York(Manhattan), Bronx, Kings, and Queens (As shown in the table below). Among those counties, Manhattan has the highest average daily PM2.5 concentration, which is 8.09 ug/mˆ3. Based on this result, I suggest state government set goals and re-allocate resources targeting air pollution with a prioritized focus on the most polluted counties.

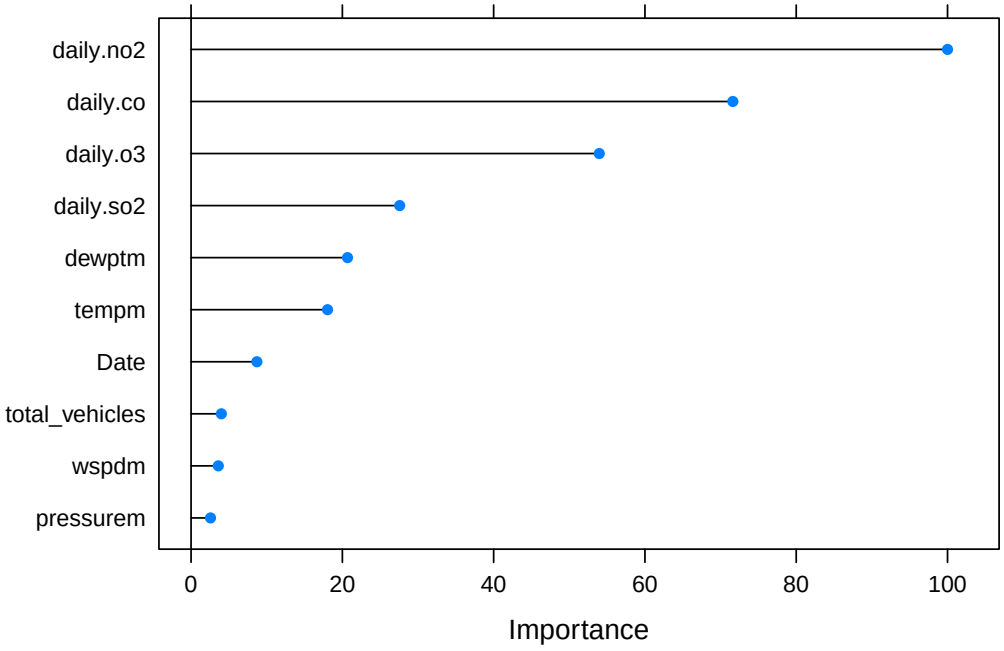
COUNTY	mean_daily_pm2.5	var_daily_pm2.5
New York	8.093111	16.32449
Bronx	6.725038	16.70366
Kings	6.625294	13.65937
Queens	6.555060	14.94248
Nassau	6.239437	13.97347
Albany	6.024579	13.63432
Westchester	6.006407	13.67965
Oneida	5.562707	13.62384

Second, the simple linear regression for air pollution and traffic shows that there is a positive correlation between vehicle amount and PM2.5 concentration. The more vehicles travel in the city, the higher the PM2.5 concentration might be. Although the correlation is not as high as expected, we cannot say the relationship between traffic and air pollution is weak. The result is probably due to the limitation of the data set and the hidden variable effect.

Third, concerning the multivariables Machine Learning Model, the Random Forest model is selected as the best prediction model due to a relatively better goodness of fit (lower RMSE) and stronger explanatory power (higher Rsquared). Based on rf model, an analysis of the variable importance reveals that air pollutant factors are most important type of variable and

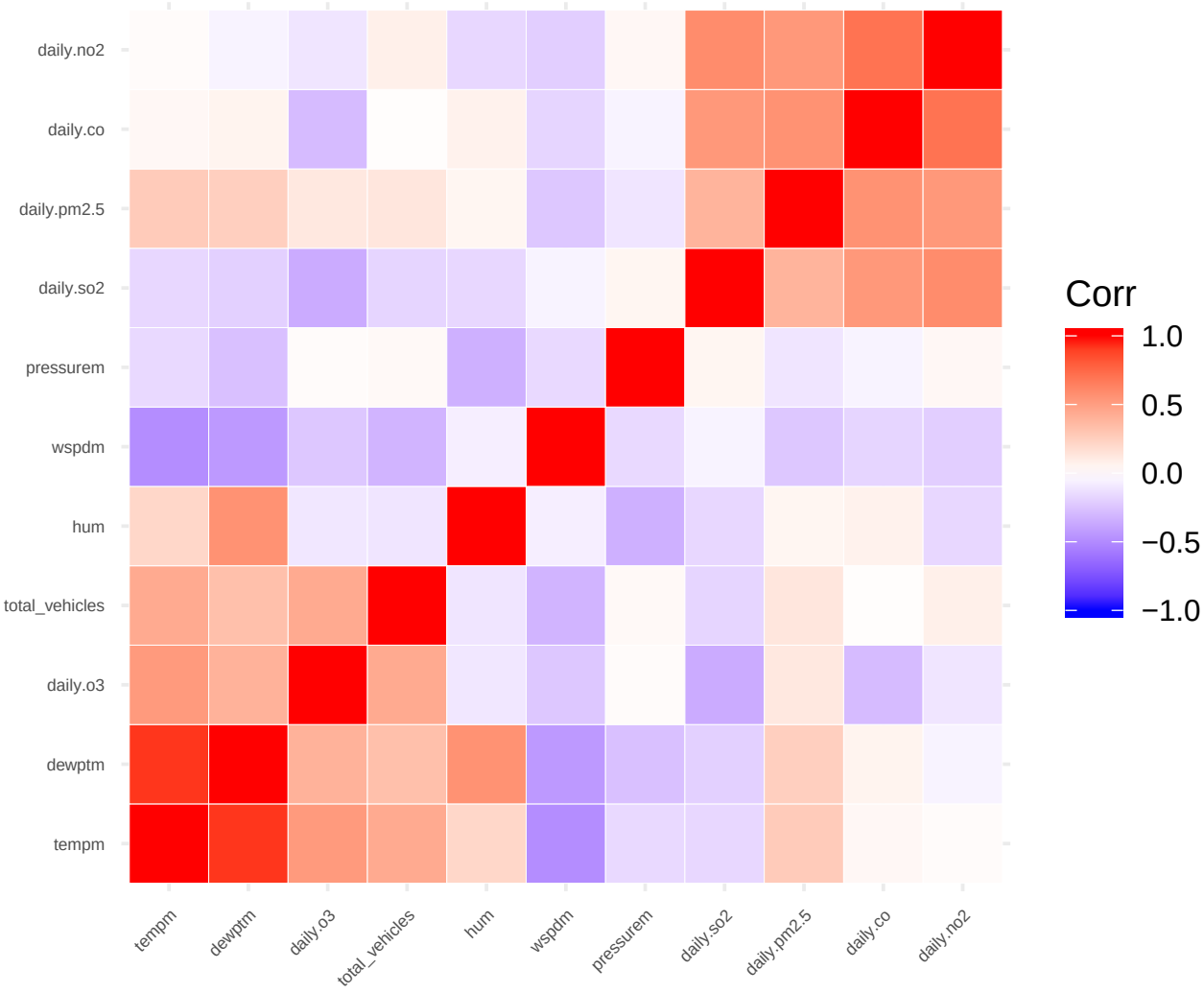
most related with PM2.5. Using this result, government could target on resident behavior or industry emission that contribute most to, such as, NO2 and Ozone.

Graph 5. Importance Score of Variables



Regarding the relationship between different features (Graph 6), I find that air pollutants have the strongest positive correlation with each other compared and matter most in this PM2.5 prediction model. Weather data such as temperature and dew point also show moderate level positive correlation with PM2.5 concentration. Although we didn't see a strong direct relationship between traffic and pm2.5 as expected, the result of correlation between features suggest that the correlation between traffic and Ozone is around 0.5, which is relatively strong, and O3 is an important feature according to the above importance analysis. This indicates that indirect effect of traffic on PM2.5 is worthwhile to be further examined in the future research.

Graph 6. Correlation between Features



6. Discussion

This project is conducted in a logical manner and successfully finds some patterns between PM2.5 and different type of factors and builds a simplified prediction model for better air pollution Monitoring.

However, several limitation should not be overlooked. First, concerning the limitation of data source, the prediction model could be more reliable if the sample size is larger. Besides, the way I calculate daily traffic amount (add up the hourly amount of vehicle passing through the tunnels and bridges operated by the MTA by one direction) is discussable and might

partially result in the weak direct correlation between traffic amount and PM2.5.

Second, regarding examination, I will dig more into the outliers and investigate how those outliers affect the accuracy of the model. Besides, I will evaluate the different ways in dealing with and compare the effect of imputing and omitting missing value on the prediction model.

Third, in terms of the feature limitation, the air pollutant data is the aggregate outcome resulted from not only traffic, but also resident and business behavior. Given more time, I would investigate deeper into those compositions by involving more features, such as heating oil consumption and industrial emission.

Reference

E.Thoen, 2018. *A recipe for recipes.* [Available via: https://edwinth.github.io/blog/recipes_blog/]

E.Rieuf, 2017. *How To Interpret R-squared and Goodness-of-Fit in Regression Analysis.* [Available via:https://www.datasciencecentral.com/profiles/blogs/regression-analysis-how-do-i-interpret-r-squared-and-assess-the]

NYC Health. *Air Pollution and the Health of New Yorkers: The Impact of Fine Particles and Ozone.* [Avaiable via: https://www1.nyc.gov/assets/doh/downloads/pdf/eode/eode-air-quality-impact.pdf]

NYC Health. *The Public Health Impacts of PM2.5 from Traffic Air Pollution* [Available via: http://a816-dohbesp.nyc.gov/IndicatorPublic/traffic/index.html]

S.Lakshmanan, 2019. *How, When and Why Should You Normalize / Standardize / Rescale Your Data?* [Available via: https://medium.com/@swethalakshmanan14/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff]

T. Chai & R. R. Draxler, 2014. *Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature.* [Available via: https://www.geosci-model-dev.net/7/1247/2014/gmd-7-1247-2014.pdf]

Y.Rybarczyk & R.Zalakeviciute, 2017. *Regression Models to Predict Air Pollution from Affordable Data Collections.* [Available via: https://www.intechopen.com/books/machine-learning-advanced-techniques-and-emerging-applications/regression-models-to-predict-air-pollution-from-affordable-data-collections]

- `Data Source`

EPA, 2018. *Outdoor Air Quality Data.* [Available via: https://www.epa.gov/outdoor-air-quality-data/download-daily-data]

kaggle, *NYC Hourly Weather Data.* [Available via: https://www.kaggle.com/meinertsen/nyc-hourly-weather-data/data].

NY Gov, [Available via: https://data.ny.gov/widgets/qzve-kjga]

- Citation for R Package

citation (“tidyverse”); citation(“tidyr”); citation (“caret”); citation (“recipes”); citation(“inggpubr”); citation(“ggplot”); citation (“knitr”); citation(“kableExtra”);