

# **Air Quality Predictions in Beijing, China**

By: Alex Calametti, Brian Guenther, Gabriela Delgado, and Naseema Omer

# Overview

**Background**

1

**Data  
Preparation**

2

**Exploration**

3



4

**Modeling**

5

**Conclusions**

6

**Next Steps**

# Background & Scope

- This project examines how air quality in Beijing, China is affected by
  - Time
  - Pollutants
  - Weather conditions
- Objectives:
  - Build models to predict the  $O_3$  values

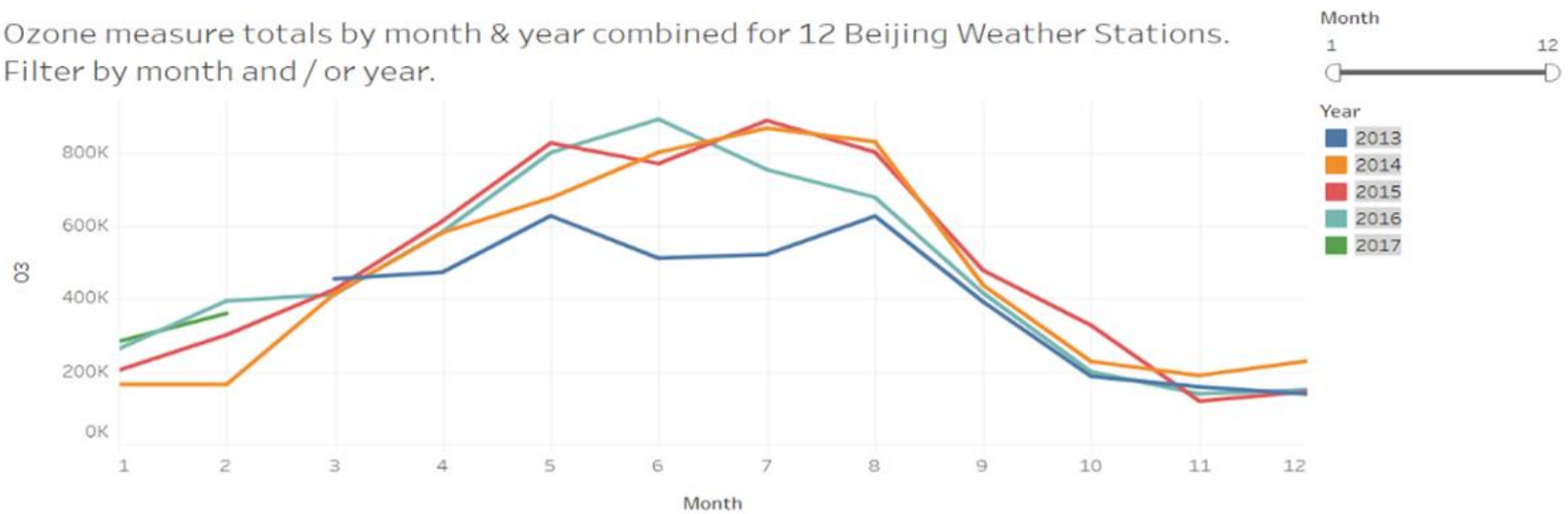


# Data Preparation

- Data gathered from UCI Machine Learning Repository
  - Includes data collected from 12 different sites in Beijing, China from 2013-2017
  - Includes different chemicals and weather conditions that affect the air quality
- Clean data is concatenated and put into S3 buckets
  - Data with NaN values dropped
  - Data with NaN values replaced with the median value for each station
  - Data with NaN values and station names dropped

# Data Exploration (O<sub>3</sub>)

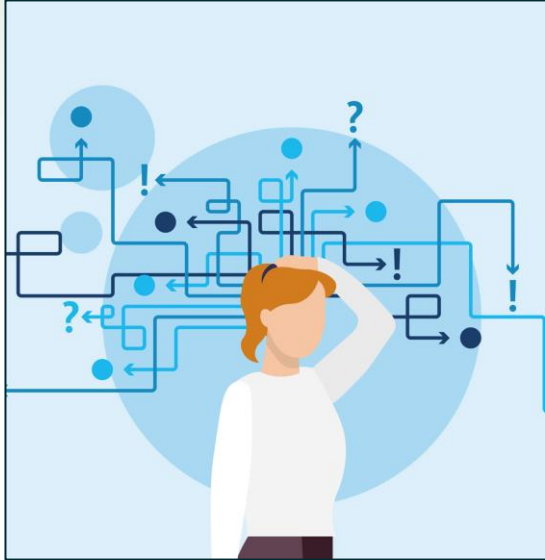
Ozone measure totals by month & year combined for 12 Beijing Weather Stations.  
Filter by month and / or year.



# Data Exploration (CO)

Carbon Monoxide measure totals by month & year combined for 12 Beijing Stations.  
Filter by month and / or year.





# **Modeling Implementation and Optimization**

# Models Used/Attempted

- Neural Network
- Simple Linear Regression
- Multivariable Linear Regression
- Decision Tree Regressor



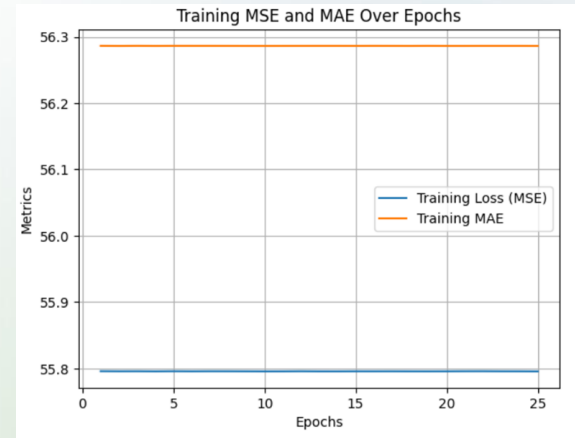
# Neural Network

This model uses

- O3 as the target variable (y)
- The rest of the columns of interest in the dataset as the predictor (X)
- Layers and nodes with different values to train the model

The results of this model are not statistically significant:

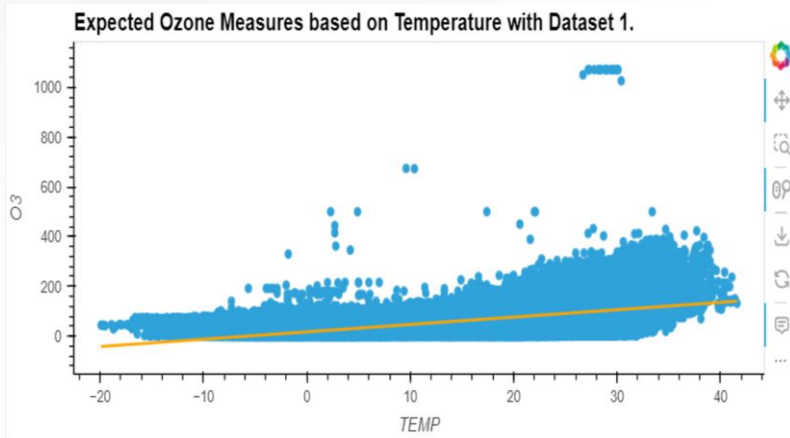
- Loss = ~55
- MSE = ~6500
- MAE = ~56



# Linear Regression

## Simple

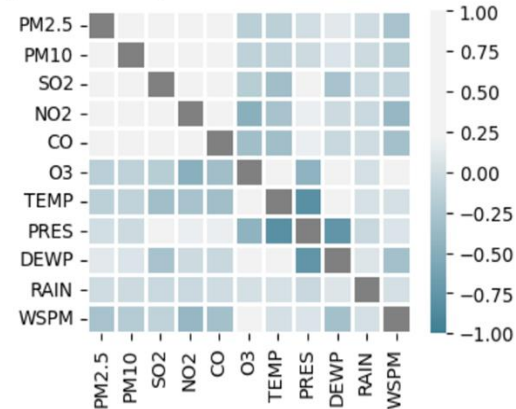
- Target  $y = O_3$ ,  $x = Temp$
- 1: NAN dropped  
 $y = 17.31191562019925 + 2.963656055355665x$   
 $r^2$  is 0.3565257008834236
- 2: NAN median  
 $y = 18.118847608324636 + 2.866120914930229x$   
 $r^2$  is 0.34434693063035227



## Multivariable

- Model 1:  $y = O_3$   
 $r^2$  is 0.5724457157382516
- Model 2:  $y = CO$   
 $r^2$  is 0.723158568713435
- Correlation coefficient of the variables

Heatmap visualizing Pearson Correlation Coefficient Matrix



# Decision Tree Regressor

- Variables:
  - Date info (year, month, day, hour)
  - Other pollutants(NO2, SO2, etc.)
  - Weather conditions (temp, rain, wind speed, etc.)
- Ozone was best predicted by date
- R-Squared value of 0.87

```
#Previously attempted data inputs and their associated R-squared (R2) values.

#air_data_df.drop(["wd"],axis=1,inplace=True)
#(R2 = 0.83) --> ran model with all columns containing numerical values

#air_data_df.drop(["wd","year","month","day","hour",,axis=1,inplace=True)
#(R2 = 0.67) --> evaluated chemical compounds and weather variables as pred

#air_data_df.drop(["wd","year","month","day","hour","PM2.5","PM10","SO2","NO2",
#(R2 =0.63) --> evaluated weather variables as predictor of O3

#air_data_df.drop(["wd","TEMP","PRES","DEWP","RAIN","WSPM","year","month","day",
#(R2 = 0.03) --> evaluated other chemical compounds as predictor of O3

#air_data_df.drop(["wd","PM2.5","PM10","SO2","NO2","CO","PRES","DEWP","RAIN","WSPM",
#(R2 = 0.84) --> evaluated date and temperature as predictor of ozone

##Attempt to see if calculating relative humidity optimizes model --> R2 = 0.87
#formula obtained from https://bmcnoldy.earth.miami.edu/Humidity.html retrieved
```

# Results / Conclusions

## **O<sub>3</sub> varies across time**

O<sub>3</sub> values are affected by the months and show a similar trend across years

## **Weather Impacts**

Temperature has a weak correlation with O<sub>3</sub> & CO



## **Chemicals are correlated**

O<sub>3</sub> & CO presented a weak correlation with pollutants present in the air

## **Decision Tree: Best model**

Decision tree was the best predictor of O<sub>3</sub>

# Next Steps

- Explore how well models can predict values for other compounds
- Modify parameters
- Seasonal time series analysis



# THANK YOU!



Thank you to Hunter, Sam, Randy, tutors, and our amazing classmates for all of their help on this project and over the course of the semester!