Data Appendix to /The Effects of Bus Rapid Transit Systems on Reducing Ozone Pollution in Sao Paulo, Brazil/

Abby Hanna

5/7/2020

Sao Paulo Air Pollution description

Citation: Companhia Ambiental do Estado de São Paulo and Danilo Lessa Bernardineli, "Air Pollution at São Paulo, Brazil, since 2013." Kaggle, https://www.kaggle.com/danlessa/air-pollution-at-so-paulo-brazil-since-2013/metadata
DOI: 10.34740/KAGGLE/DSV/627405.**
Date Downloaded:** April 20, 2020
Filename(s): raw_data/cetesb.csv **Unit of observation:** pollutants recorded at stations in region of Sao Paulo, Brazil
Dates covered: 2013-2019

To obtain a copy

This data has been webscraped from the Companhia Ambiental do Estado de São Paulo website for easy access on Kaggle.com by Danilo Lessa Bernardelli. To obtain the data, navigate to the link provided above.

Importable version (if necessary)

This data needed to be processed in order for me to work with it on my computer. The file originally contains air pollution data for different pollutants, but many of the data are missing. Given this, I chose to focus on ozone, as there were the most observations. Due to time constraints and computer malfunctions I was not able to mutate the data on my own, so professor Sayre performed the data extraction. The raw data file was mutated into a summary table of daily ozone levels:

Filename(s): importable_data/ozone_daily.csv

Variable descriptions

- **date** Date ozone was recorded.
- **municipality:** City ozone recorded in.
- **mean_ozone:** Average daily level of ozone recorded.
- **max_ozone:** Maxoimum daily level of ozone recorded.
- **min_ozone:** Minimum daily level of ozone recorded.

Data import code and summary

```
ozone_daily <- read_csv("ozone_daily.csv")
```

```
## Parsed with column specification:
## cols(
##   date = col_date(format = ""),
##   id = col_double(),
##   municipality = col_character(),
##   mean_ozone = col_double(),
##   max_ozone = col_double(),
##   min_ozone = col_double()
## )
```

Sao Paulo historical climate data

Citation: Doutorado and Mestrado. "Hourly Weather Surface - Brazil (Southeast Region)." Kaggle, April 6, 2018. https://www.kaggle.com/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region/metadata. Original data source: Dataset Source: INMET (National Meteorological Institute - Brazil). **Date Downloaded:** April 30, 2020 **Filename(s):** raw_data/sudeste.csv **Unit of observation:** Hourly climate data for southeast region of Brazil
Dates covered: 2000-2016

To obtain a copy

The easiest way to obtain the data is to follow the above link to Kaggle.com.

Importable version (if necessary)

This data also needed to be modified in order for my computer to work with it properly. Professor Sayre extracted the data from the Sao Paulo weather stations and averages for maximum temperature, minumum temperature, average temperature, average humidity, and average windspeed as control variables for the regression. The data was combined into one dataset by geographical id.

Filename(s): importable_data/recent_weather_daily.csv, importable_data/nearest_weather_Station.csv

Variable descriptions

- **max_temp:** Maximum daily temperature recorded.
- **min_temp:** Minimum daily temperature recorded.
- **avg_temp:** Average daily temperature recorded.
- **avg_humidity:** Average daily humidity recorded.
- **mean_windspeed:** Average daily windspeed recorded.

Data import code and summary

```
nearest_weather_Station <- read_csv("nearest_weather_Station.csv")
```

```
## Parsed with column specification:
## cols(
##   wsid = col_double(),
##   id = col_double()
## )
```

```
recent_weather_daily <- read_csv("recent_weather_daily.csv")
```

```
## Parsed with column specification:
## cols(
##   wsid = col_double(),
##   wsnm = col_character(),
##   city = col_character(),
##   date = col_date(format = ""),
##   max_temp = col_double(),
##   min_temp = col_double(),
##   avg_temp = col_double(),
##   mean_humidity = col_double(),
##   mean_windspeed = col_double()
## )
```

Data processing and combination

```
## Air Data
#Ozone readings from stations in Sao Paulo since 2013
table_one <- ozone_daily %>%
  filter(municipality == "SAO PAULO") %>%
  select(date, municipality, mean_ozone, max_ozone, min_ozone)

## Weather Data
#Daily weather averages in Sao Paulo since 2013
table_two <- recent_weather_daily %>%
  filter(city == "São Paulo") %>%
  select(date, city, max_temp, min_temp, avg_temp, mean_humidity, mean_windspeed)

 #Combine the data into one frame
combined_data_sp <- merge(table_one, table_two)
  #Table 2: Summary statistics for ozone and weather variables in state of Sao Paulo
stargazer(combined_data_sp)
```

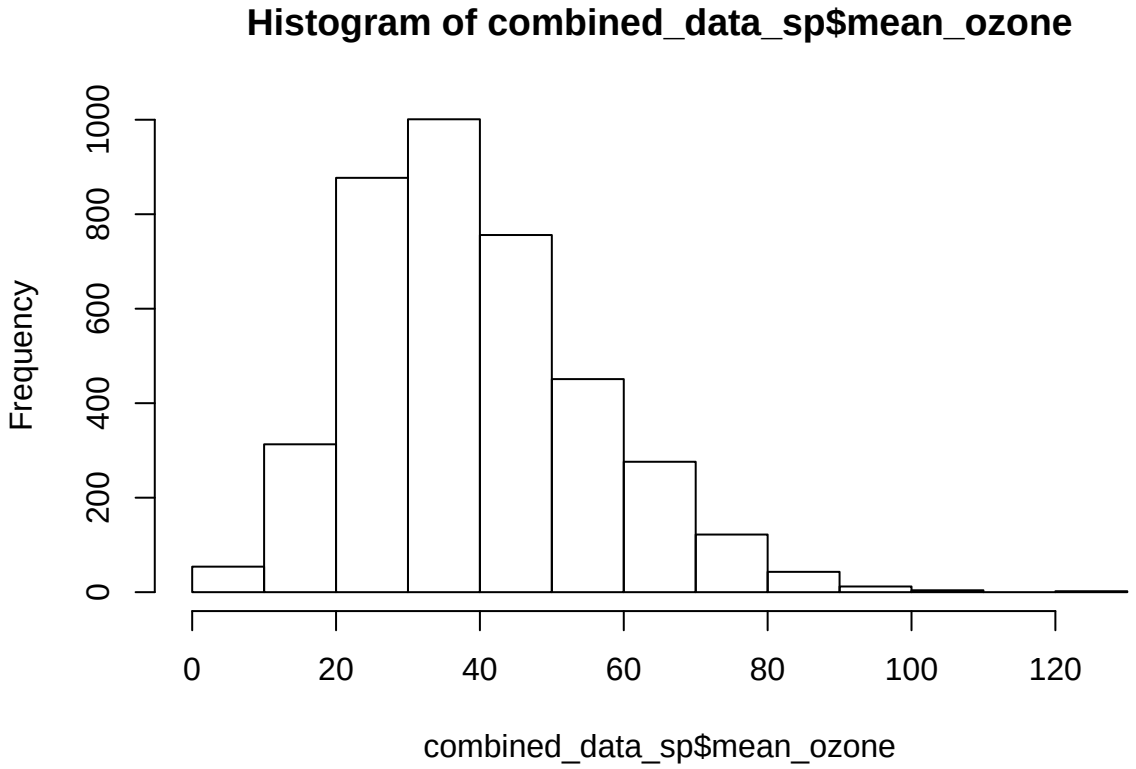
% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, May 07, 2020 - 02:04:27 PM

```
sp_data <- as_tibble(combined_data_sp)
weather_data <- as_tibble(nearest_weather_Station)
ozone_data <- as_tibble(ozone_daily)
```

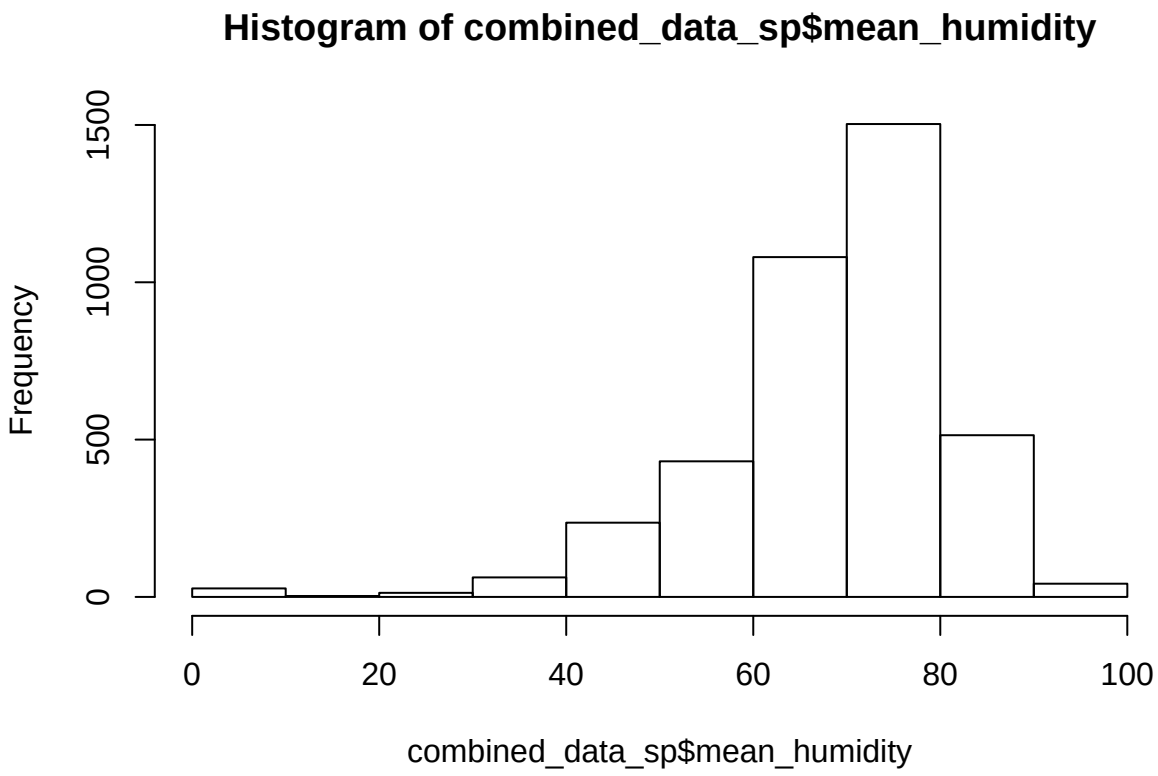
```
hist(combined_data_sp$mean_ozone)
```

Table 1:

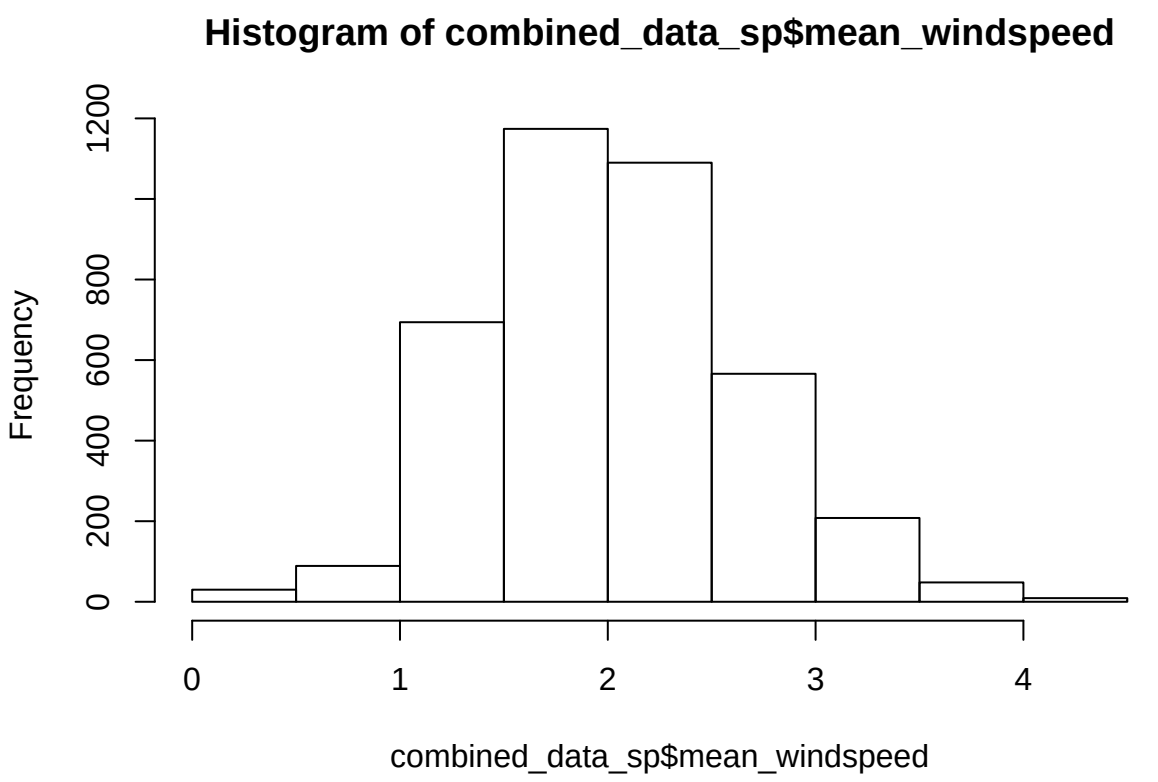
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
mean_ozone	3,911	39.132	16.440	0.000	27.356	48.913	125.478
max_ozone	3,911	86.773	41.315	0	55	111	347
min_ozone	3,911	7.711	9.940	0	0	11	76
max_temp	3,911	25.309	5.050	0.000	22.150	29.000	37.000
min_temp	3,911	16.406	4.199	0.000	14.300	19.400	25.400
avg_temp	3,911	20.092	4.032	0.000	17.550	22.812	29.117
mean_humidity	3,911	68.387	12.835	0.000	62.958	76.375	96.292
mean_windspeed	3,908	2.022	0.625	0.000	1.587	2.432	4.358



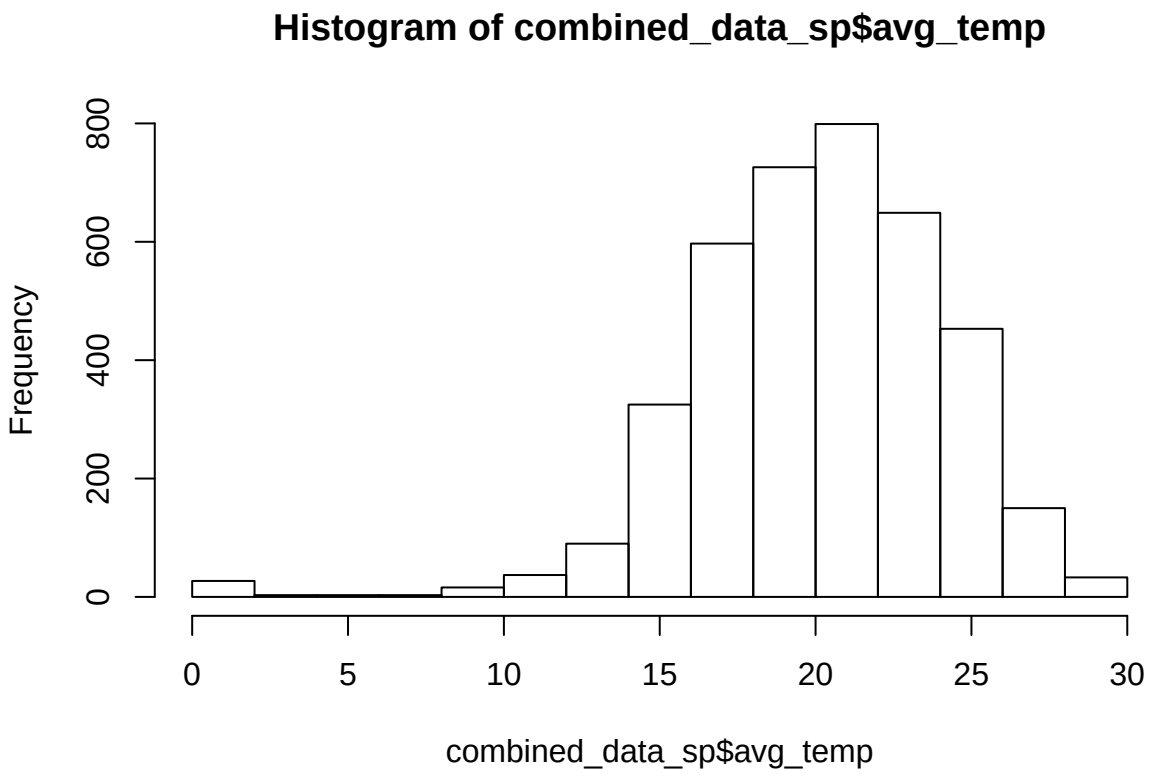

```
hist(combined_data_sp$mean_humidity)
```




```
hist(combined_data_sp$mean_windspeed)
```

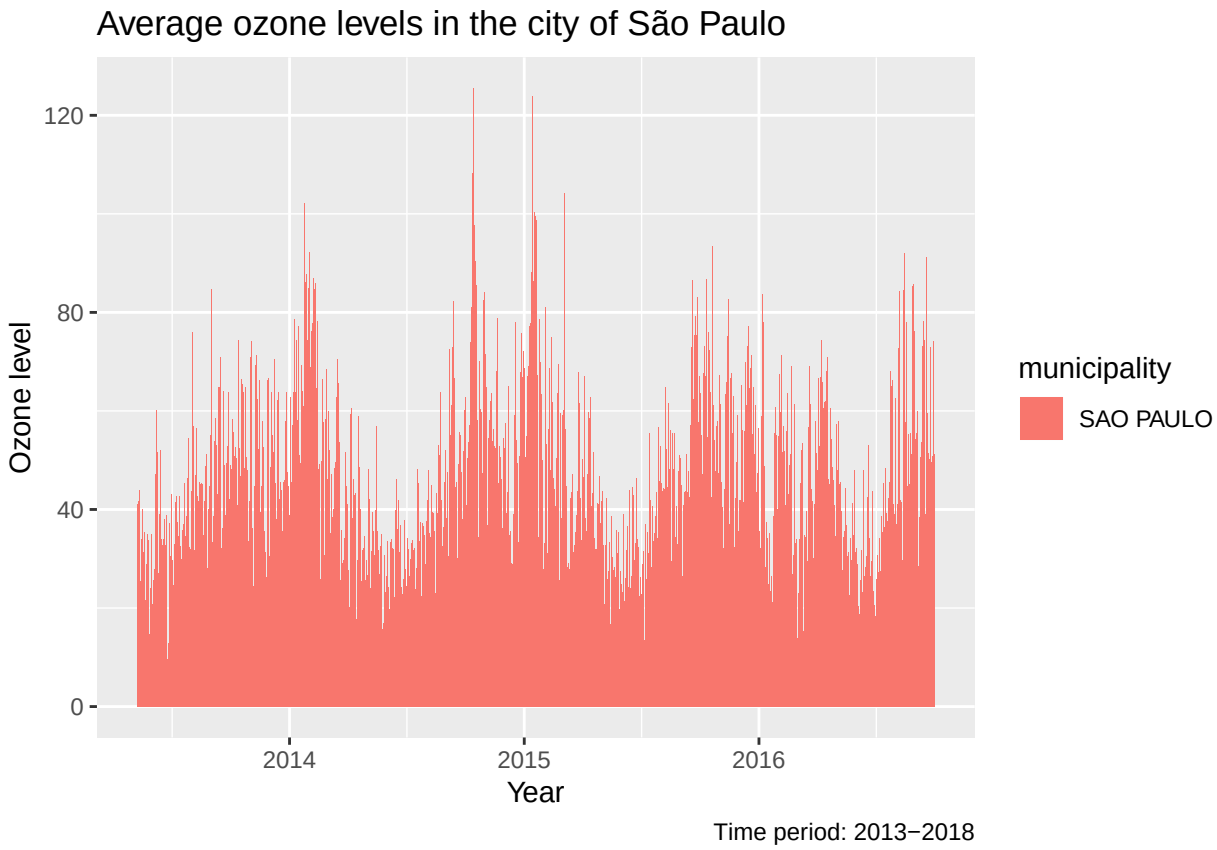



```
hist(combined_data_sp$avg_temp)
```

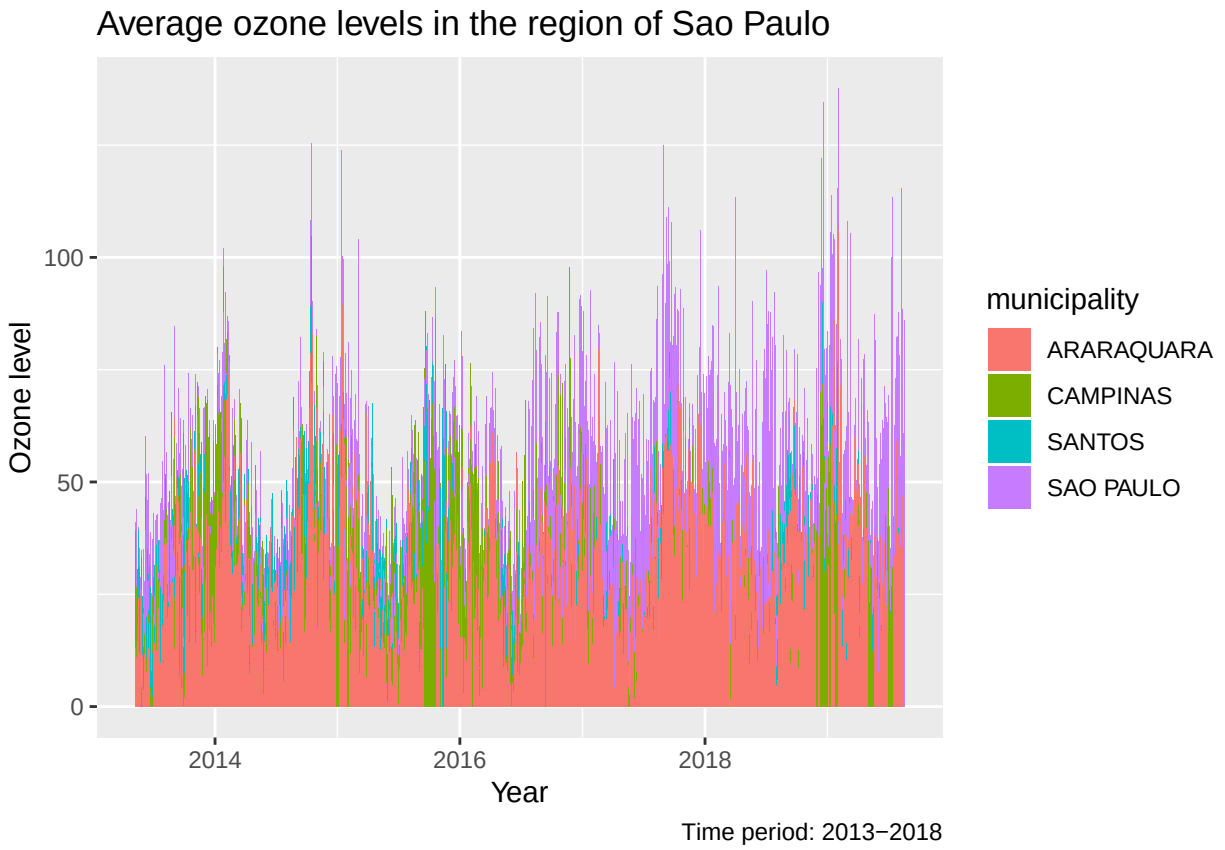


Create plots for city and state of Sao Paulo, Figure 1 and Figure 2

```
# Figure 1
ggplot(combined_data_sp, aes(x=date, y=mean_ozone, fill = municipality)) + geom_col(position = "dodge")
```




```
# Figure 2
ggplot(ozone_daily, aes(x=date, y=mean_ozone, fill = municipality)) + geom_col(position = "dodge") + lab
```




```
#create variables for after 2015 and in Sao Paulo (city)
sp_data$after_wc <- ifelse(sp_data$date >= "2015-01-01",1,0)
new_ozone_data <- ozone_data %>%
  mutate(log_ozone = log(mean_ozone+1),
         after_wc =ifelse(date >= "2015-01-01",1,0),
         in_sp =ifelse(municipality == "SAO PAULO",1,0), after_wc_did = after_wc*in_sp)
new_weather_data <- merge(recent_weather_daily, weather_data)
#Combine the data into one frame
combined_data <- new_ozone_data %>%
  left_join(new_weather_data)
```

```
## Joining, by = c("date", "id")
```

```
# Table 2
model_sp <- lm(log_ozone ~ after_wc + in_sp + after_wc_did + max_temp + min_temp + avg_temp + mean_humi
stargazer(model_sp, omit = "as.factor")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, May 07, 2020 - 02:04:32 PM

notes = Monitor fixed effects included for id (weather station), but not reported here.

Analysis Variables

Note: this code comes out nicely in R, but it does not knit well into the PDF document. The code chunk is included in the rmd of the appendix and the variable description table is in the pdf version of the final paper.

Table 2:

	Dependent variable:
	log_ozone
after_wc	0.053^{***}
	(0.015)
in_sp	0.696^{***}
	(0.031)
after_wc_did	-0.095^{***}
	(0.021)
max_temp	0.041^{***}
	(0.003)
min_temp	0.017^{***}
	(0.003)
avg_temp	-0.006
	(0.005)
mean_humidity	-0.012^{***}
	(0.0004)
mean_windspeed	0.093^{***}
	(0.009)
Constant	2.552^{***}
	(0.049)
Observations	7,259
R^2	0.384
Adjusted R^2	0.383
Residual Std. Error	0.437 (df = 7244)
F Statistic	322.880^{***} (df = 14; 7244)
Note:	*p<0.1; **p<0.05; ***p<0.01

Discussion of Data

Both climate and air pollution datasets had to be modified by professor Sayre, as the original raw data files did not read properly into my computer. The datasets were downloaded from the website above and extracted using the methods and code described.