COVID-19 and Urban Air Pollution

Big Data Final Project - Group 4

Tianrun Wang, Raymond Dee, Jonathan Pun

CHINA
Wuhan
Chongqing
Hong Kong
CHINA
Lanzhou
Beijing
Shanghai
500 km

Source: NASA Earth Observatory

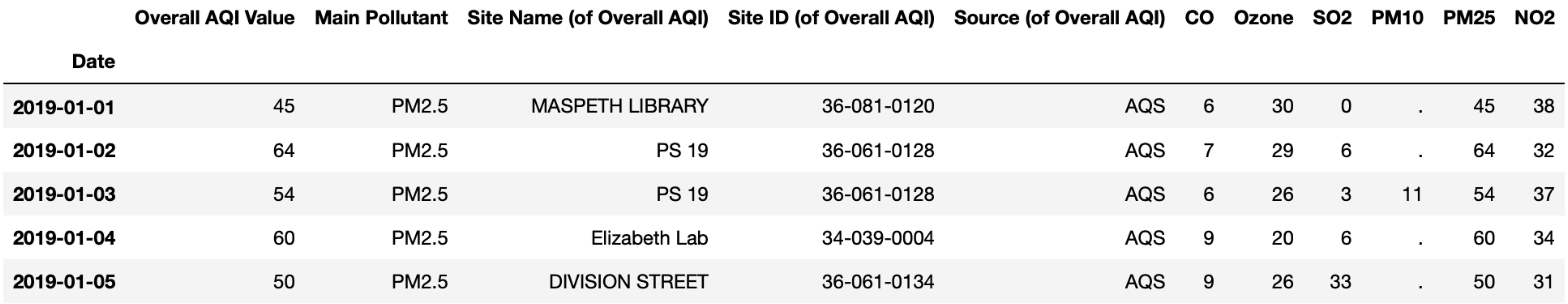
Outline

1. Datasets Introduction
2. Time Series Correlation
3. Regression Analysis
4. Near Term Air Quality Projection
5. Challenges Faced

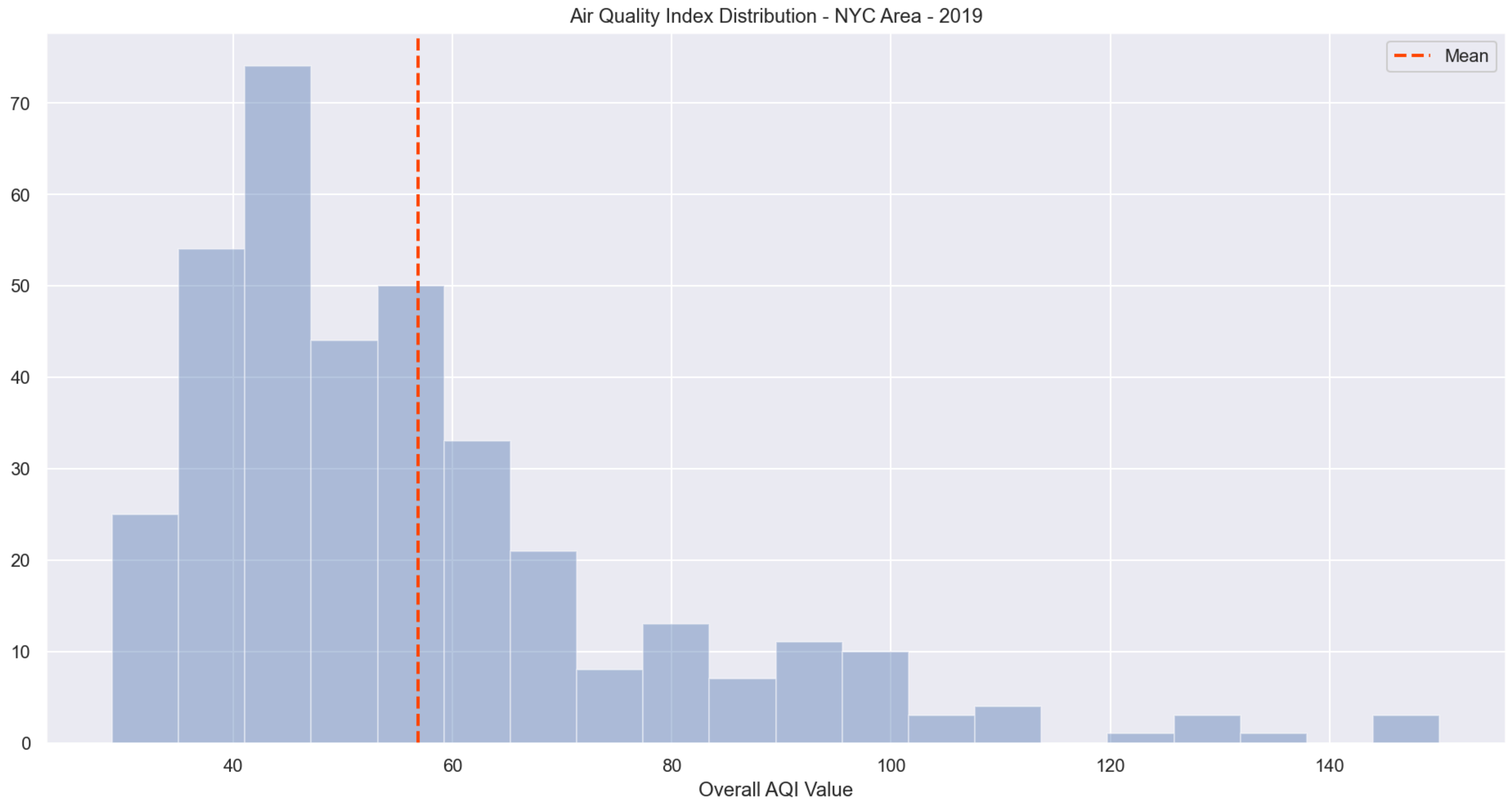
Datasets Introduction

EPA Daily Air Quality Report

- The EPA dataset provides daily readings of Air Quality Index and major pollutants such as PM2.5, NO2, and Ozone
- Temporal Resolution: Daily
- Spatial Resolution: New York Metropolitan Area
- Temporal Availability: At least 10 years to date



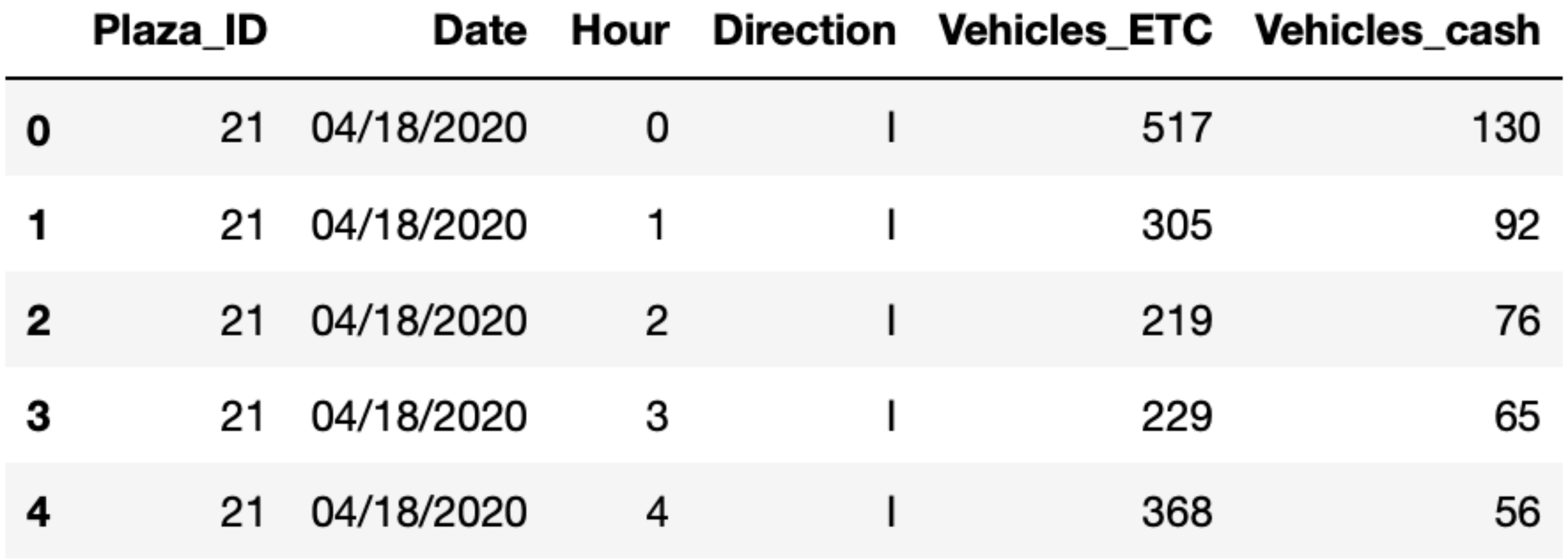
	Overall AQI Value	Main Pollutant	Site Name (of Overall AQI)	Site ID (of Overall AQI)	Source (of Overall AQI)	CO	Ozone	SO2	PM10	PM25	NO2
Date											
2019-01-01	45	PM2.5	MASPETH LIBRARY	36-081-0120	AQS	6	30	0	.	45	38
2019-01-02	64	PM2.5	PS 19	36-061-0128	AQS	7	29	6	.	64	32
2019-01-03	54	PM2.5	PS 19	36-061-0128	AQS	6	26	3	11	54	37
2019-01-04	60	PM2.5	Elizabeth Lab	34-039-0004	AQS	9	20	6	.	60	34
2019-01-05	50	PM2.5	DIVISION STREET	36-061-0134	AQS	9	26	33	.	50	31


Air Quality Index Distribution - NYC Area - 2019
Mean
0
10
20
30
40
50
60
70
40
60
80
100
120
140
Overall AQI Value

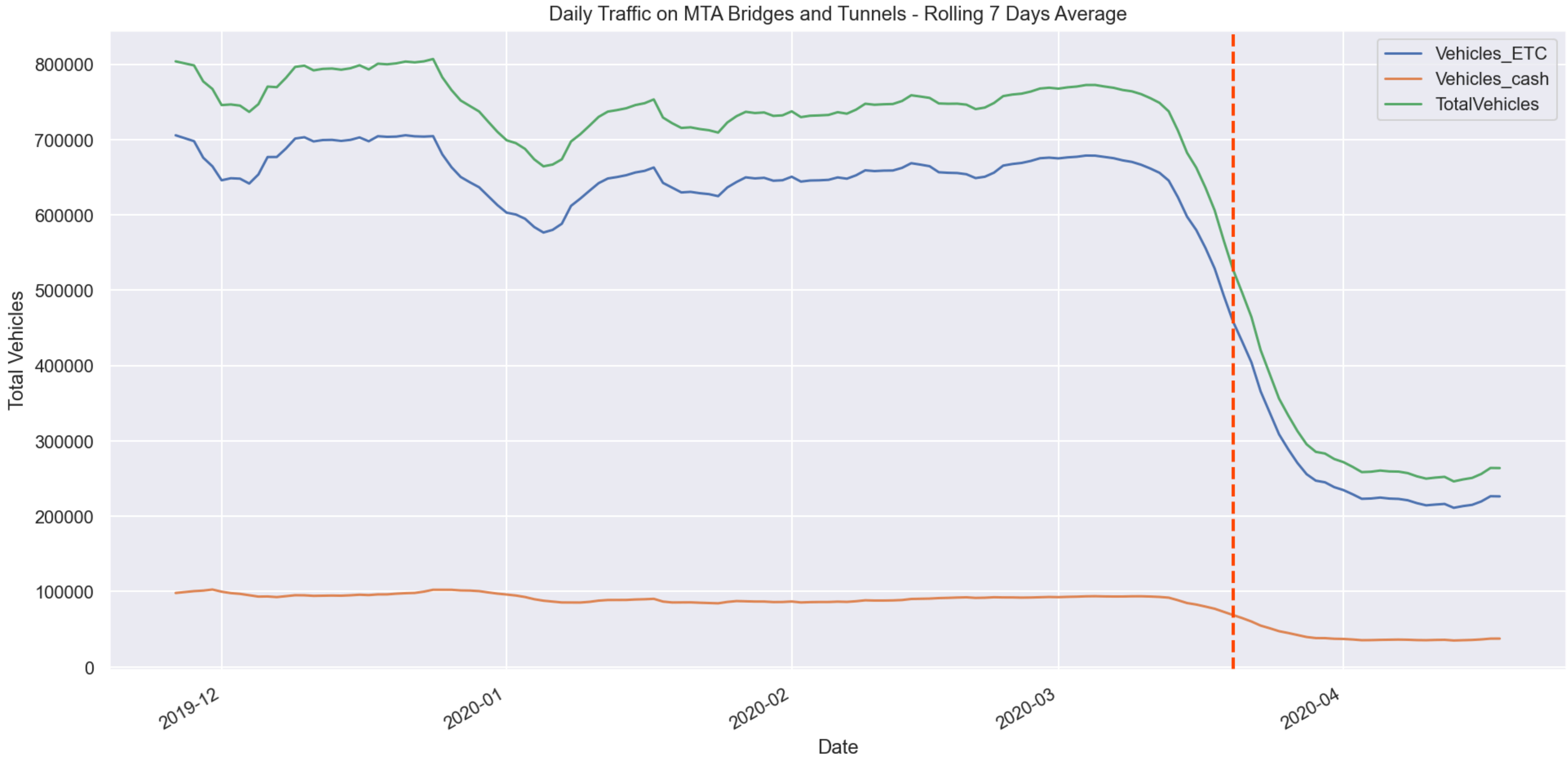
Air Quality Index and Major Pollutants - NYC Area - Rolling 30 Days Average
Overall AQI Value
Ozone
PM25
Index Value
55
50
45
40
35
30
Jan
2020
Feb
Mar
Apr
Date

Hourly Traffic on MTA Bridges and Tunnels

- Total traffic throughput on MTA bridges and Tunnels each hour
- Aggregated to total daily throughput (Spark SQL)
- Temporal Resolution: hourly
- Spatial Resolution: each MTA bridge or tunnel
- Temporal Availability: 2010 to date

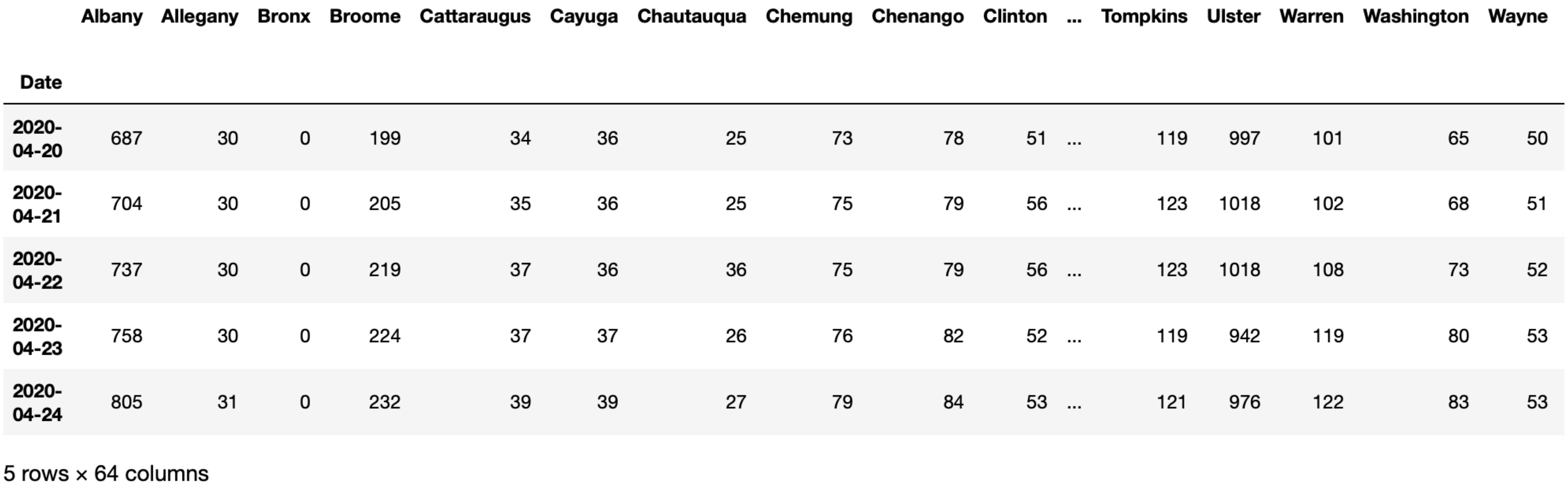


	Plaza_ID	Date	Hour	Direction	Vehicles_ETC	Vehicles_cash
0	21	04/18/2020	0	I	517	130
1	21	04/18/2020	1	I	305	92
2	21	04/18/2020	2	I	219	76
3	21	04/18/2020	3	I	229	65
4	21	04/18/2020	4	I	368	56


Daily Traffic on MTA Bridges and Tunnels - Rolling 7 Days Average
Vehicles_ETC
Vehicles_cash
TotalVehicles
Total Vehicles
800000
700000
600000
500000
400000
300000
200000
100000
0
2019-12
2020-01
2020-02
2020-03
2020-04
Date

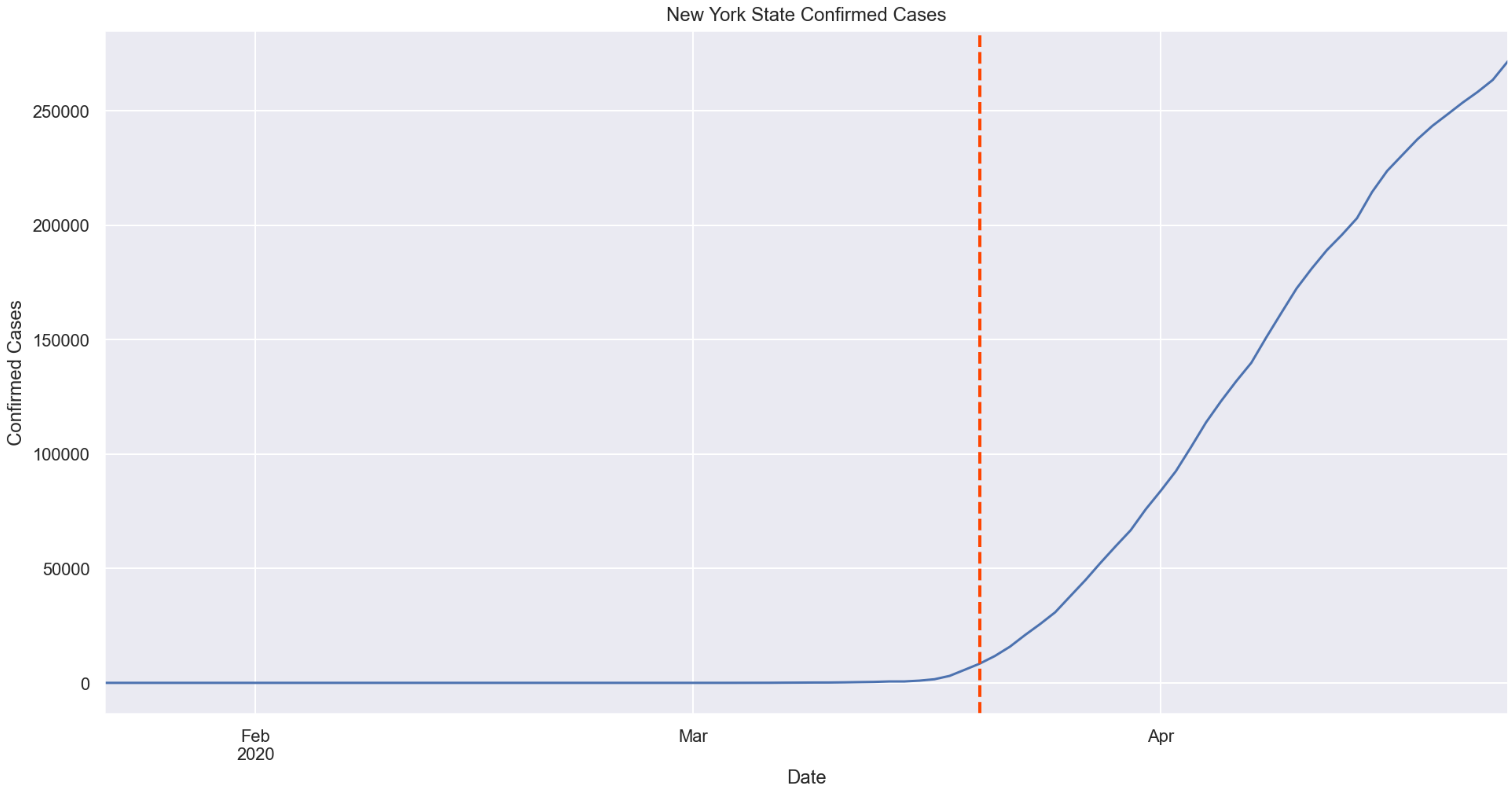
Johns Hopkins COVID-19

- Daily time series of confirmed cases
- Temporal Resolution: Daily
- Spatial Resolution: County
- Temporal Availability: 1/22/2020 to date



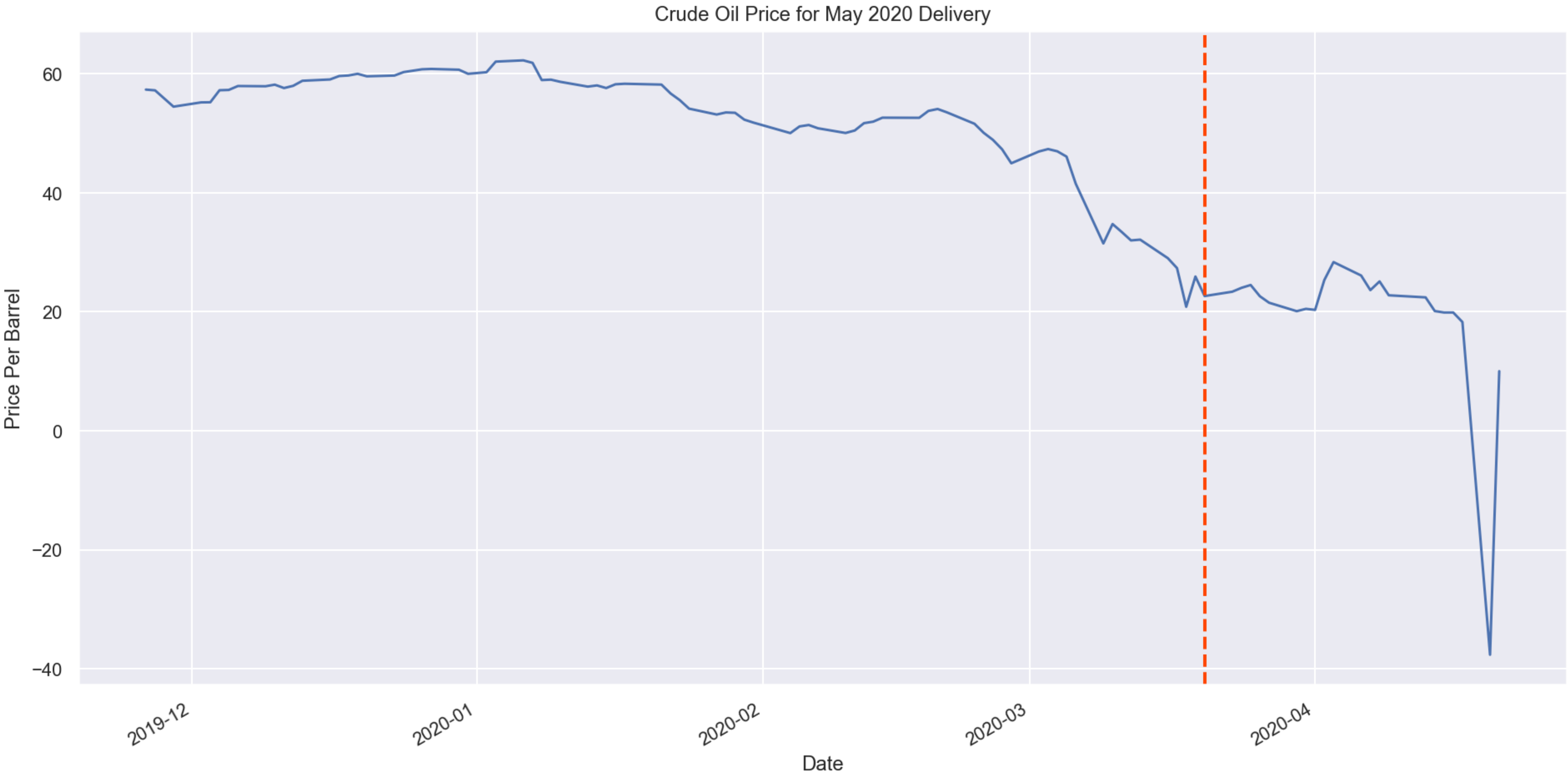
	Albany	Allegany	Bronx	Broome	Cattaraugus	Cayuga	Chautauqua	Chemung	Chenango	Clinton	...	Tompkins	Ulster	Warren	Washington	Wayne
Date																
2020-04-20	687	30	0	199	34	36	25	73	78	51	...	119	997	101	65	50
2020-04-21	704	30	0	205	35	36	25	75	79	56	...	123	1018	102	68	51
2020-04-22	737	30	0	219	37	36	36	75	79	56	...	123	1018	108	73	52
2020-04-23	758	30	0	224	37	37	26	76	82	52	...	119	942	119	80	53
2020-04-24	805	31	0	232	39	39	27	79	84	53	...	121	976	122	83	53

5 rows × 64 columns


New York State Confirmed Cases
Confirmed Cases
250000
200000
150000
100000
50000
0
Feb
2020
Mar
Apr
Date

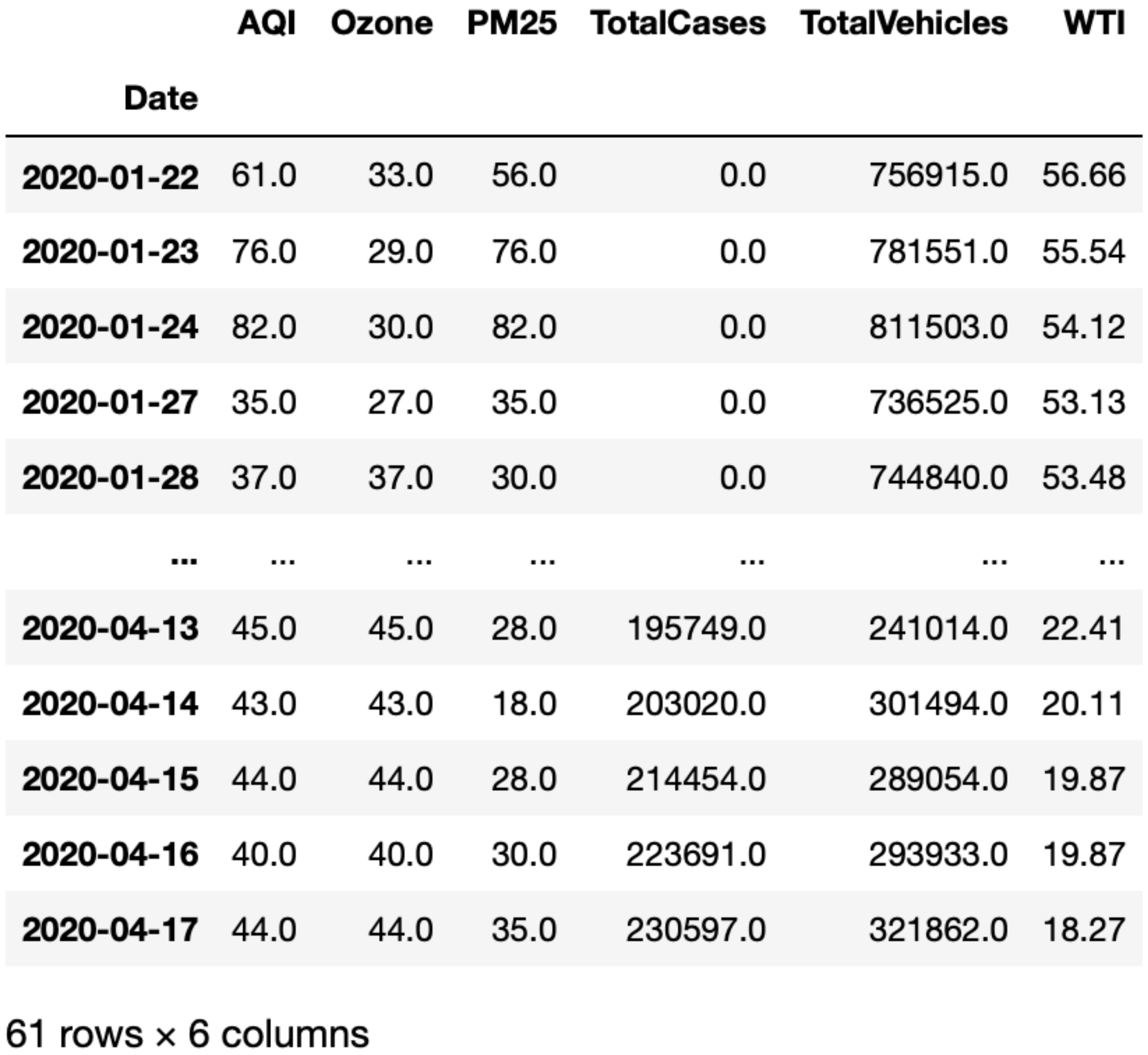
CME Crude Oil Futures

- Reflects crude oil demand/supply dynamic, used as proxy for energy consumption and industrial activity
- May 2020 delivery, trading terminated at end of April
- Temporal Resolution: Daily
- Temporal Availability: 11/20/2014 to 4/24/2020


Crude Oil Price for May 2020 Delivery
Price Per Barrel
60
40
20
0
−20
−40
2019-12
2020-01
2020-02
2020-03
2020-04
Date

Integrated Dataframe

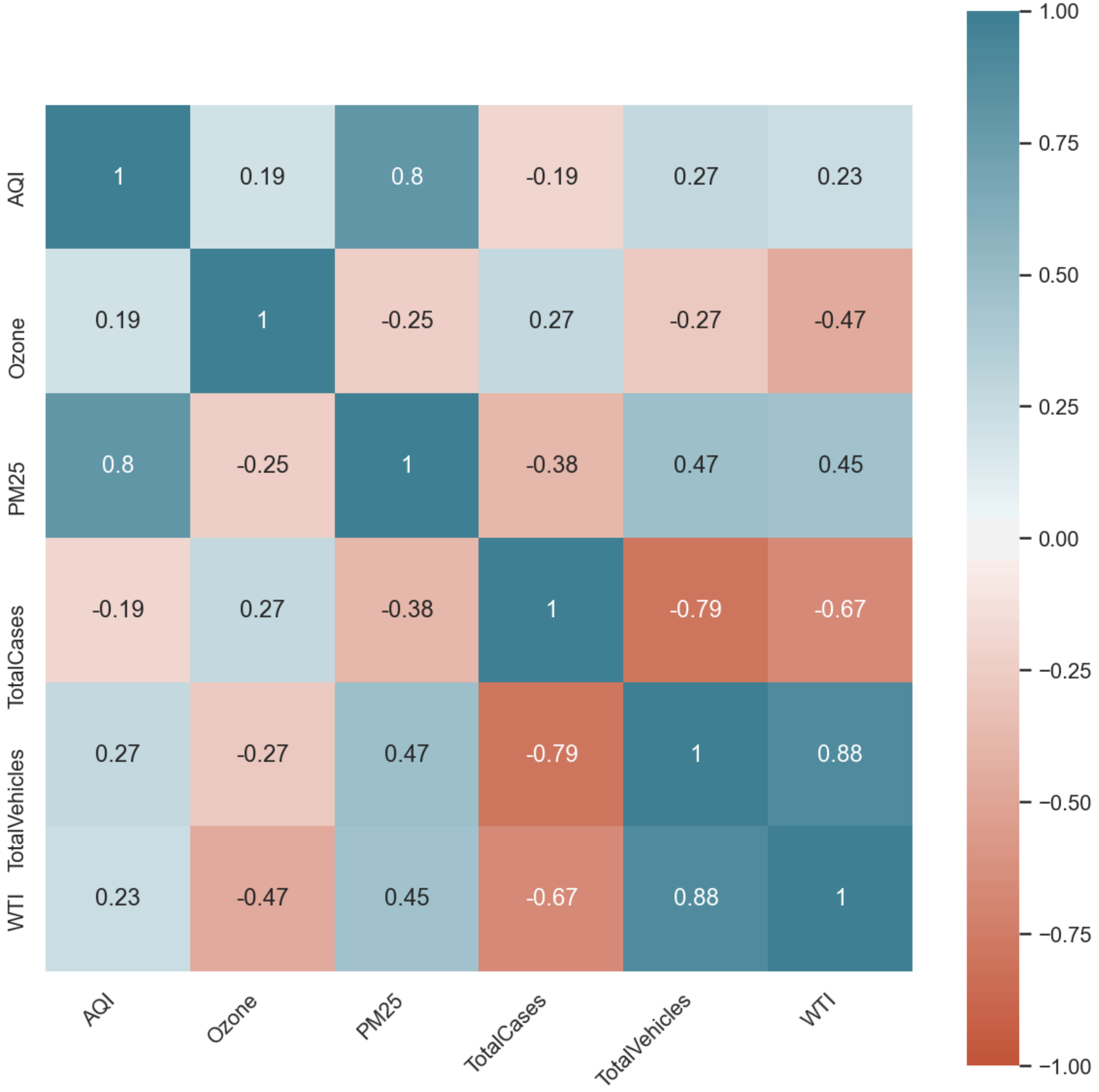
- Inner-joined on date time index
- Time series from 1/22 to 4/17



	AQI	Ozone	PM25	TotalCases	TotalVehicles	WTI
Date						
2020-01-22	61.0	33.0	56.0	0.0	756915.0	56.66
2020-01-23	76.0	29.0	76.0	0.0	781551.0	55.54
2020-01-24	82.0	30.0	82.0	0.0	811503.0	54.12
2020-01-27	35.0	27.0	35.0	0.0	736525.0	53.13
2020-01-28	37.0	37.0	30.0	0.0	744840.0	53.48
...	...	...	...	...	...	...
2020-04-13	45.0	45.0	28.0	195749.0	241014.0	22.41
2020-04-14	43.0	43.0	18.0	203020.0	301494.0	20.11
2020-04-15	44.0	44.0	28.0	214454.0	289054.0	19.87
2020-04-16	40.0	40.0	30.0	223691.0	293933.0	19.87
2020-04-17	44.0	44.0	35.0	230597.0	321862.0	18.27

61 rows × 6 columns

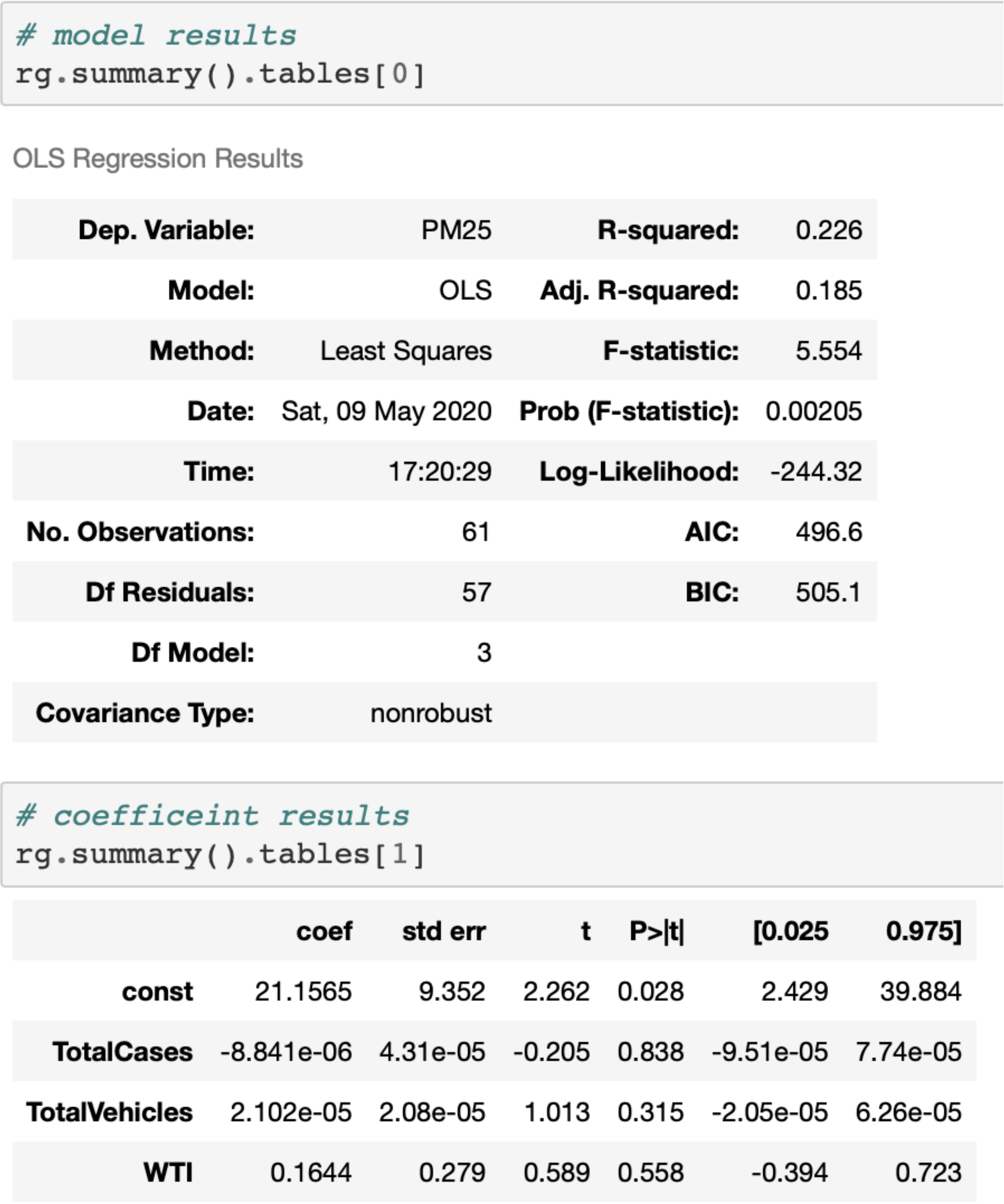
Time Series Correlation


AQI
Ozone
PM25
TotalCases
TotalVehicles
WTI
1
0.19
0.8
-0.19
0.27
0.23
0.19
1
-0.25
0.27
-0.27
-0.47
0.8
-0.25
1
-0.38
0.47
0.45
-0.19
0.27
-0.38
1
-0.79
-0.67
0.27
-0.27
0.47
-0.79
1
0.88
0.23
-0.47
0.45
-0.67
0.88
1
AQI
Ozone
PM25
TotalCases
TotalVehicles
WTI
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Regression Analysis

Effects on PM2.5

- PM2.5 on TotalCases, TotalVehicles and WTI
- TotalVehicles and WTI load positively, TotalCases loads negatively



```
# model results
rg.summary().tables[0]
```

OLS Regression Results

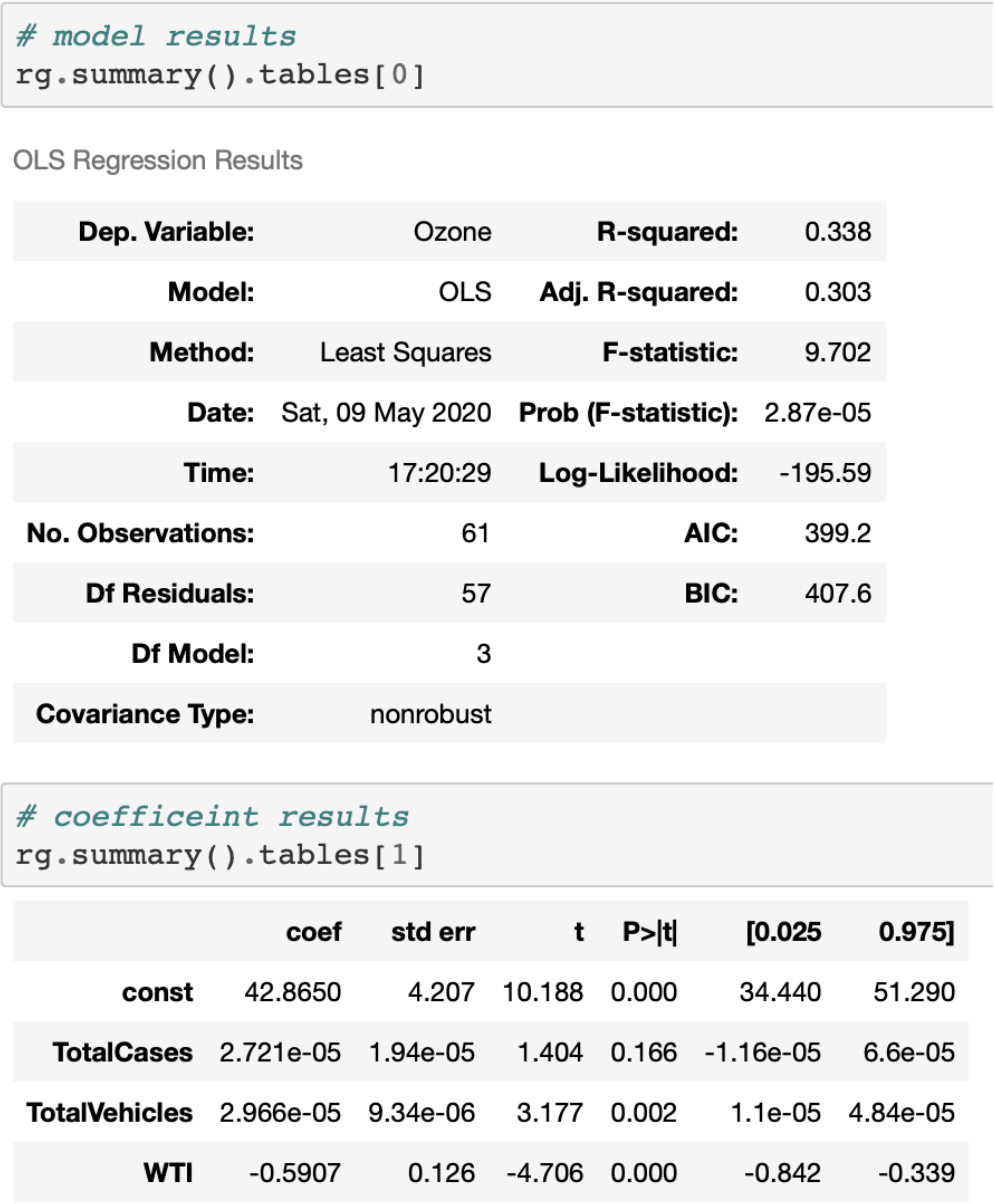
Dep. Variable:	PM25	**R-squared:**	0.226
Model:	OLS	**Adj. R-squared:**	0.185
Method:	Least Squares	**F-statistic:**	5.554
Date:	Sat, 09 May 2020	**Prob (F-statistic):**	0.00205
Time:	17:20:29	**Log-Likelihood:**	-244.32
No. Observations:	61	**AIC:**	496.6
Df Residuals:	57	**BIC:**	505.1
Df Model:	3		
Covariance Type:	nonrobust		

```
# coefficeint results
rg.summary().tables[1]
```

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	21.1565	9.352	2.262	0.028	2.429	39.884
TotalCases	-8.841e-06	4.31e-05	-0.205	0.838	-9.51e-05	7.74e-05
TotalVehicles	2.102e-05	2.08e-05	1.013	0.315	-2.05e-05	6.26e-05
WTI	0.1644	0.279	0.589	0.558	-0.394	0.723

Effects on Ozone

- Ozone on TotalCases, TotalVehicles and WTI
- TotalCases and TotalVehicles load positively, WTI loads negatively



```
# model results
rg.summary().tables[0]
```

OLS Regression Results

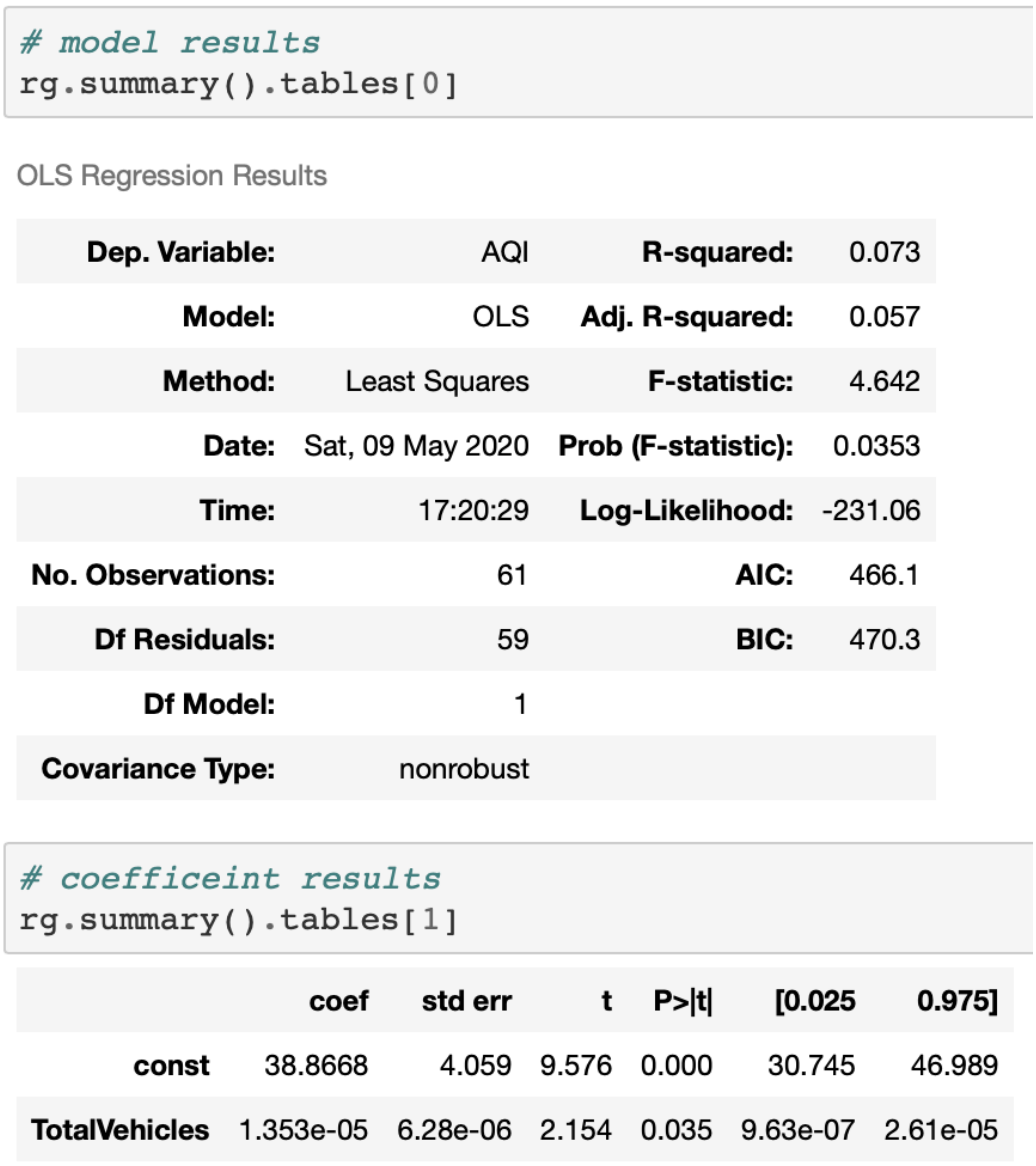
Dep. Variable:	Ozone	**R-squared:**	0.338
Model:	OLS	**Adj. R-squared:**	0.303
Method:	Least Squares	**F-statistic:**	9.702
Date:	Sat, 09 May 2020	**Prob (F-statistic):**	2.87e-05
Time:	17:20:29	**Log-Likelihood:**	-195.59
No. Observations:	61	**AIC:**	399.2
Df Residuals:	57	**BIC:**	407.6
Df Model:	3		
Covariance Type:	nonrobust		

```
# coefficeint results
rg.summary().tables[1]
```

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	42.8650	4.207	10.188	0.000	34.440	51.290
TotalCases	2.721e-05	1.94e-05	1.404	0.166	-1.16e-05	6.6e-05
TotalVehicles	2.966e-05	9.34e-06	3.177	0.002	1.1e-05	4.84e-05
WTI	-0.5907	0.126	-4.706	0.000	-0.842	-0.339

Effects on AQI

- AQI on TotalCases, TotalVehicles, and WTI
- TotalCases and WTI insignificant, dropped from model
- TotalVehicles loads positively



```
# model results
rg.summary().tables[0]
```

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.073
Model:	OLS	**Adj. R-squared:**	0.057
Method:	Least Squares	**F-statistic:**	4.642
Date:	Sat, 09 May 2020	**Prob (F-statistic):**	0.0353
Time:	17:20:29	**Log-Likelihood:**	-231.06
No. Observations:	61	**AIC:**	466.1
Df Residuals:	59	**BIC:**	470.3
Df Model:	1		
Covariance Type:	nonrobust		

```
# coefficeint results
rg.summary().tables[1]
```

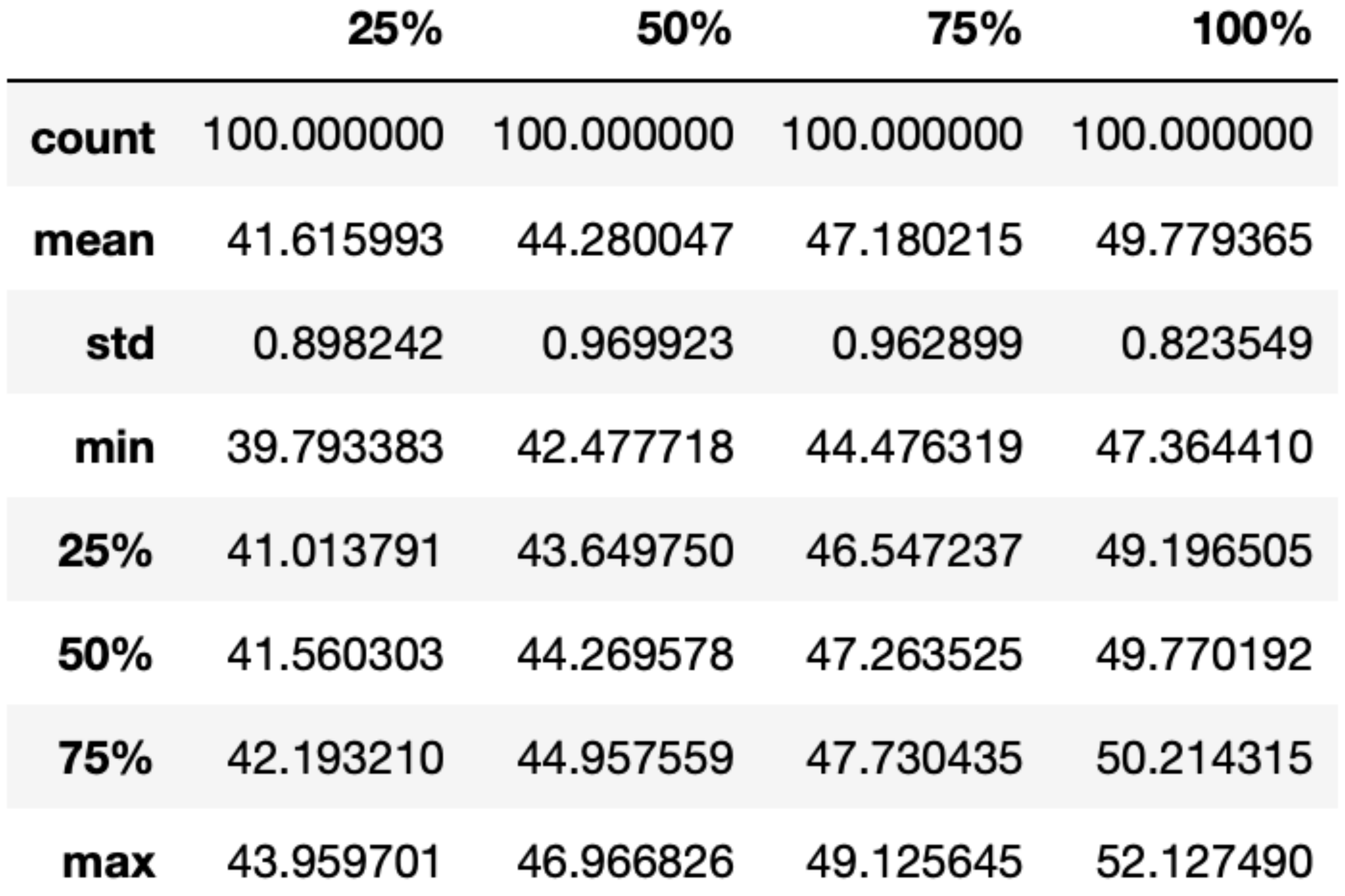
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	38.8668	4.059	9.576	0.000	30.745	46.989
TotalVehicles	1.353e-05	6.28e-06	2.154	0.035	9.63e-07	2.61e-05

Near Term Air Quality Index Projection

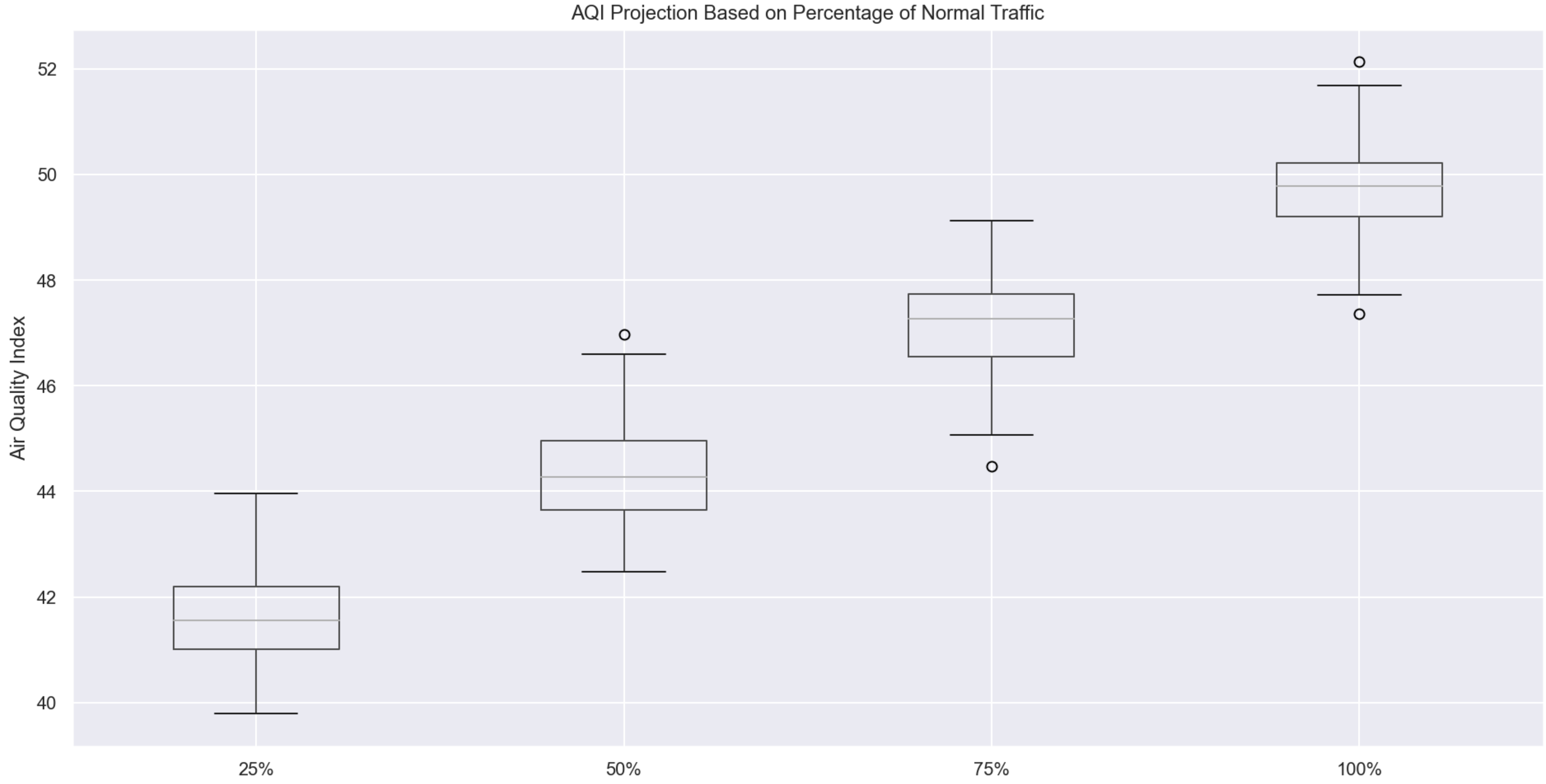
AQI Projection

Based on Percentage of Normal Traffic

- Use one-variable (TotalVehicles) model
- Projections on 25%, 50%, 75% and 100% of normal traffic (2019 mean)
- For each percentage level, run 100 projections
- Vehicles sampled from normal distribution with 2019 mean and std



	25%	50%	75%	100%
count	100.000000	100.000000	100.000000	100.000000
mean	41.615993	44.280047	47.180215	49.779365
std	0.898242	0.969923	0.962899	0.823549
min	39.793383	42.477718	44.476319	47.364410
25%	41.013791	43.649750	46.547237	49.196505
50%	41.560303	44.269578	47.263525	49.770192
75%	42.193210	44.957559	47.730435	50.214315
max	43.959701	46.966826	49.125645	52.127490


AQI Projection Based on Percentage of Normal Traffic
Air Quality Index
52
50
48
46
44
42
40
25%
50%
75%
100%

Challenges in the Project

- Finding the right datasets: appropriate frequency and publishing schedule
- Cleaning the data: datetime parsing is a pain, standardized “%Y-%m-%d” easiest to work with, e.g. 2020-05-11
- COVID-19 relatively novel phenomenon, time series analysis prone to influence of outliers due to scarce data

Questions and Comments