1. Business Understanding

a. Define business requirements and objectives

The business goal is defined from the perspective of the U.S. Environmental Protection Agency (EPA), which is a public organization responsible for the protection of human health and the environment. The end goal of EPA is to decide whether to move further with the issue regarding the current air pollution in the U.S and plan the team cycle for developing any regulations needed. Then, as they have been doing for the past, they will enforce new regulations for national standards and help companies follow the requirements. With the goal of developing the most effective and impactful regulations to the public, the project will mainly focus on forecast and analysis of US air pollution data, however, the project assumes that EPA is currently investigating how impactful the eclectic vehicles are to air pollution as well, therefore will include the analysis portion of EVs if and only if the rate of air pollution decline as time passses. Therefore, it will be decided once the modeling of the air pollution data is done. Theoretically, owning the EVs cars will reduce air pollution since EVs reduce the emission of greenhouse gases and various air pollutants, and EPA wants to see if there are any correlations.

For initial strategy before trying different model algorithms, one of the possible models could be derived from tensorflow's machine learning, organizing the input and output, implementing hidden layers, and training the model.

b. **Problem statement**

With rapid global warming, regulating emission of air pollutants, one of the top 3 causes of global warming is more imperative than ever. Build a machine learning model to forecast the US air pollution, analyze the trend, then leverage the insight from the forecast to decide whether to proceed with publication of the data and regulatory enforcements to the public.

2. Data Understanding

a. Collect data

After evaluating the data sets based on how valuable their attributes are to the business goals of this data analysis, U.S. Air Pollution Dataset from Kaggle.com is selected.

	Unnamed: 0	State Code	County Code	Site Num	Address	State	County	City	Date Local	NO2 Units	NO2 Mean	NO2 1st Max Value	NO2 1st Max Hour	NO2 AQI	O3 Units	O3 Mean	O3 1st Max Value	O3 1st Max Hour	O3 AQI	SO2 Units	SO2 Mean	SO2 1st Max Value	SO2 1st Max Hour	SO2 AQI	CO Units	CO Mean	CO 1st Max Value	CO 1st Max Hour	CO AQI
0	0	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	2000-01-01	Parts per billion	19.041667	49.0	19	46	Parts per million	0.022500	0.040	10	34	Parts per billion	3.000000	9.0	21	13.0	Parts per million	1.145833	4.2	21	NaN
1	1	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	2000-01-01	Parts per billion	19.041667	49.0	19	46	Parts per million	0.022500	0.040	10	34	Parts per billion	3.000000	9.0	21	13.0	Parts per million	0.878947	2.2	23	25.0
2	2	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	2000-01-01	Parts per billion	19.041667	49.0	19	46	Parts per million	0.022500	0.040	10	34	Parts per billion	2.975000	6.6	23	NaN	Parts per million	1.145833	4.2	21	NaN
3	3	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	2000-01-01	Parts per billion	19.041667	49.0	19	46	Parts per million	0.022500	0.040	10	34	Parts per billion	2.975000	6.6	23	NaN	Parts per million	0.878947	2.2	23	25.0
4	4	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	2000-01-02	Parts per billion	22.958333	36.0	19	34	Parts per million	0.013375	0.032	10	27	Parts per billion	1.958333	3.0	22	4.0	Parts per million	0.850000	1.6	23	NaN

Table 1-1. U.S. Pollution Data from Kaggle.com

The data set above is composed of 1.7 million rows containing the content for 28 columns. 28 atomic attributes represent the location, date, and measured 4 major pollutants: NO2, SO2, O3, and CO. 19 Columns are in numerical types, and 9 are in categorical values.

```
 #   Column             Dtype
---  ------             -----
 0   Unnamed: 0         int64
 1   State Code         int64
 2   County Code        int64
 3   Site Num           int64
 4   Address            object
 5   State              object
 6   County             object
 7   City               object
 8   Date Local         object
 9   NO2 Units          object
 10  NO2 Mean           float64
 11  NO2 1st Max Value  float64
 12  NO2 1st Max Hour   int64
 13  NO2 AQI            int64
 14  O3 Units           object
 15  O3 Mean            float64
 16  O3 1st Max Value   float64
 17  O3 1st Max Hour    int64
 18  O3 AQI             int64
 19  SO2 Units          object
 20  SO2 Mean           float64
 21  SO2 1st Max Value  float64
 22  SO2 1st Max Hour   int64
 23  SO2 AQI            float64
 24  CO Units           object
 25  CO Mean            float64
 26  CO 1st Max Value   float64
 27  CO 1st Max Hour    int64
 28  CO AQI             float64
dtypes: float64(10), int64(10), object(9)
memory usage: 386.5+ MB
```

Table 1-2. Data Types of U.S. Air Pollution Data

Categorical values can be converted into numerical values, but unnecessary columns such as “address”, “country” and “city” and “unit” of pollutants in an object type will be dropped in the cleaning process. Number of pollutants measured differs by state, with California being the most measured for 576,142 records and Washingtonn being the least measured state for 962 records which is 0.1% of records measured in California. Also, according to the unique values of the “State” column, the data only covers 47 states. AQI for pollutants are also numerical values, and they report the daily air quality, for example, the day with a lower AQI value has a better air quality than days with a higher AQI value.

b. descriptive statistic

	Unnamed: 0	State Code	County Code	Site Num	NO2 Mean	NO2 1st Max Value	NO2 1st Max Hour	NO2 AQI	O3 Mean	O3 1st Max Value	O3 1st Max Hour	O3 AQI	SO2 Mean	SO2 1st Max Value	SO2 1st Max Hour	SO2 AQI	CO Mean	CO 1st Max Value	CO 1st Max Hour	CO AQI
count	1746661.00	1746661.00	1746661.00	1746661.00	1746661.00	1746661.00	1746661.00	1746661.00	1746661.00	1746661.00	1746661.00	1746661.00	1746661.00	1746661.00	1746661.00	873754.00	1746661.00	1746661.00	1746661.00	873338.00
mean	54714.14	22.31	71.69	1118.21	12.82	25.41	11.73	23.90	0.03	0.04	10.17	36.05	1.87	4.49	9.66	7.12	0.37	0.62	7.88	6.00
std	33729.08	17.26	79.48	2003.10	9.50	16.00	7.88	15.16	0.01	0.02	4.00	19.78	2.76	7.68	6.73	11.94	0.31	0.64	7.98	5.85
min	0.00	1.00	1.00	1.00	-2.00	-2.00	0.00	0.00	0.00	0.00	0.00	0.00	-2.00	-2.00	0.00	0.00	-0.44	-0.40	0.00	0.00
25%	25753.00	6.00	17.00	9.00	5.75	13.00	5.00	12.00	0.02	0.03	9.00	25.00	0.26	0.80	5.00	1.00	0.18	0.29	0.00	2.00
50%	53045.00	17.00	59.00	60.00	10.74	24.00	9.00	23.00	0.03	0.04	10.00	33.00	0.99	2.00	8.00	3.00	0.29	0.40	6.00	5.00
75%	80336.00	40.00	97.00	1039.00	17.71	35.70	20.00	33.00	0.03	0.05	11.00	42.00	2.33	5.00	14.00	9.00	0.47	0.80	13.00	8.00
max	134575.00	80.00	650.00	9997.00	139.54	267.00	23.00	132.00	0.10	0.14	23.00	218.00	321.62	351.00	23.00	200.00	7.51	19.90	23.00	201.00

Table 1-3 Descriptive Statistics of Numerical Data Values of U.S Air Pollution Date

According to the standard deviation value from descriptive statistics from Table 1-4, NO2 mean values out of all 4 pollutants are the most dispersed values. Therefore, more descriptive and dramatic changes for NO2 over the years are assumed.

The four major pollutants, NO2 (Nitrogen Dioxide), SO2 (Sulphur Dioxide), CO (Carbon Monoxide), and OZ (Ozone), are measured, and the cause of NO2 emission is burning of fossil fuels. Poor data quality might negatively impact the data analysis, hence the assessment of data quality is crucial.

c. Assess Data Quality

Though human intervention in data generation is expected, the data has been recorded by the government organization, EPA, therefore the data is considered credible. No modification of data is detected throughout the rows, and the data meets an expected consistency, except a slightly uneven number of measurements for each states, and interpretability of the data as well.

```
Unnamed: 0                0
State Code                0
County Code               0
Site Num                  0
Address                   0
State                     0
County                    0
City                      0
Date Local                0
NO2 Units                 0
NO2 Mean                  0
NO2 1st Max Value         0
NO2 1st Max Hour          0
NO2 AQI                   0
O3 Units                  0
O3 Mean                   0
O3 1st Max Value          0
O3 1st Max Hour           0
O3 AQI                    0
SO2 Units                 0
SO2 Mean                  0
SO2 1st Max Value         0
SO2 1st Max Hour          0
SO2 AQI              872907
CO Units                  0
CO Mean                   0
CO 1st Max Value          0
CO 1st Max Hour           0
CO AQI               873323
dtype: int64
```

Table 1-5 Number of null values of U.S Air Pollution Data

Null values are observed in two different columns, SO2 AQI and CO AQI. They are manageable during the cleaning phase of the data, and the data is not necessarily damaged by these missing values. The only concern is that the data covers the 2000 until 2016 which is not the most recent, however, this was the most optimal data set available for the public and covers over tens years of time span, so the data will be processed further.

d. Find Insights

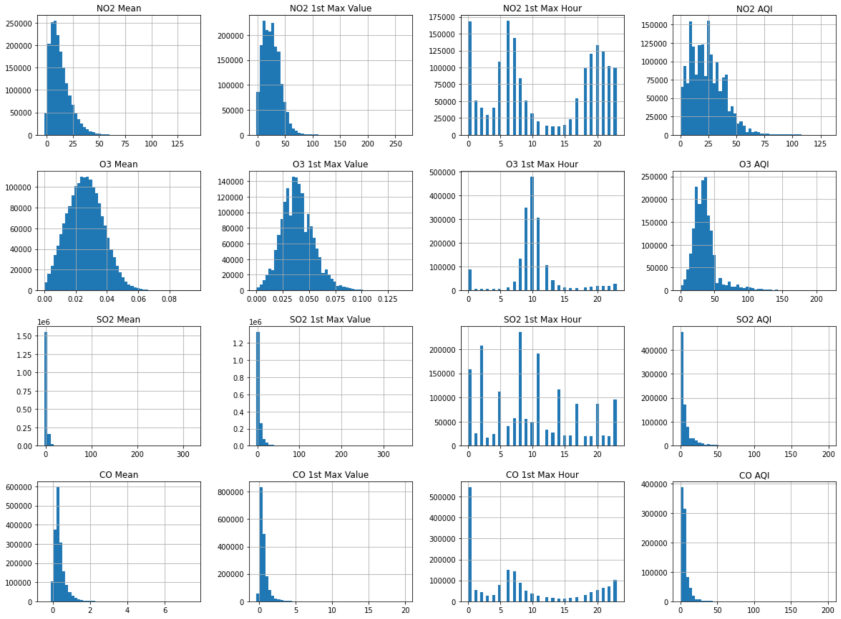

Figure 1-6 Histograms for Numerical value columns in U.S Air Pollution Data

NO2 Mean	13.510848
NO2 1st Max Value	26.345511
NO2 1st Max Hour	11.858057
NO2 AQI	24.791960
O3 Mean	0.025380
O3 1st Max Value	0.038284
O3 1st Max Hour	10.163525
O3 AQI	35.006299
SO2 Mean	1.996077
SO2 1st Max Value	4.739460
SO2 1st Max Hour	9.668549
SO2 AQI	7.530025
CO Mean	0.395403
CO 1st Max Value	0.677170
CO 1st Max Hour	8.045104
CO AQI	6.534192

Figure 1-7 Mean Values Grouped by Each Pollutant Columns

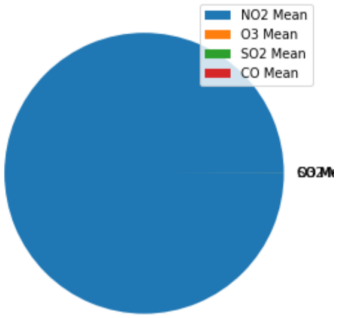

Figure 1-8 A Pie Plot of Composition of 4 Pollutants In Air Pollution

Few things to note here is that the unit of NO2 is in billion while others are in million. Therefore, the NO2 mean value is plotted after being multiplied by a factor of 1,000. Considering the composition weight of NO2 amongst three values, it is reasonable to dig into the values of NO2.

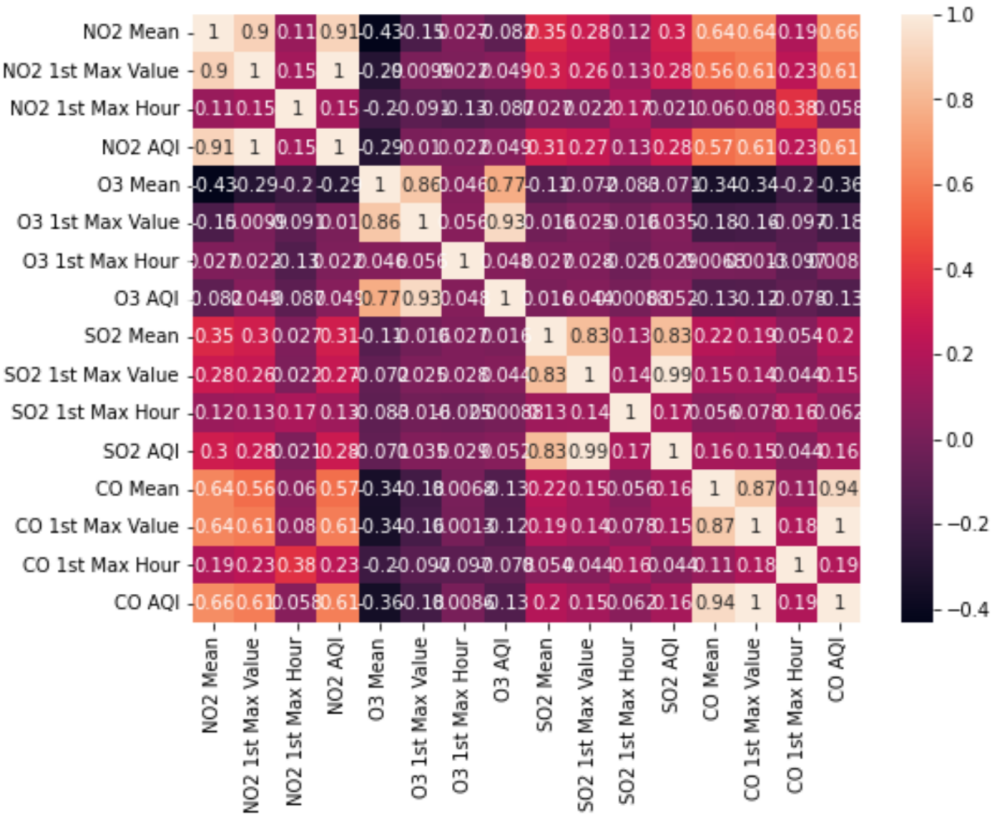

Figure 1-9 Correlation between 4 pollutants from U.S Air Pollution Data

Overall, a high correlation between 4 pollutants are observed, and the optimal correlations are observed in 4x4 blocks for each pollutant. However, this correlation itself doesn't give a meaningful insight. Also, null values in SO2 AQI and CO AQI might have affected the correlation.

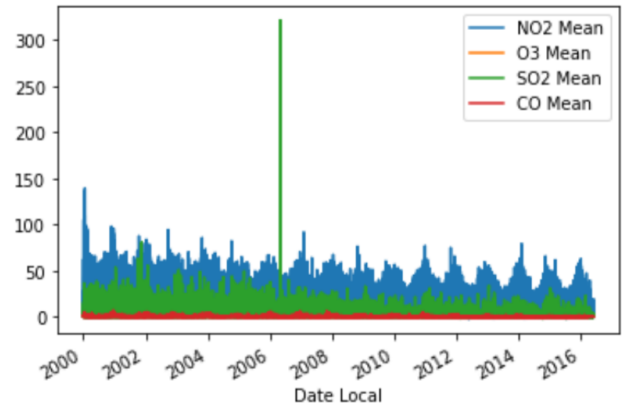

Figure 1-10 Measure of Each Pollutants from 2000 to 2016

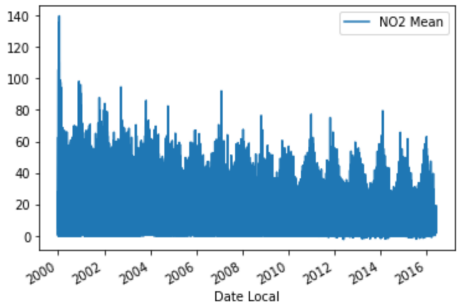

Figure 1-11 Measure of NO2 Mean from 2000 to 2016

The sudden spike in the values of SO2 Mean in Figure 1-10 tells us that there was an interruption in data recording that was missed in the earlier process, hence it revealed a meaningful information about the data that needed to be taken care of in the preparation process of the data.

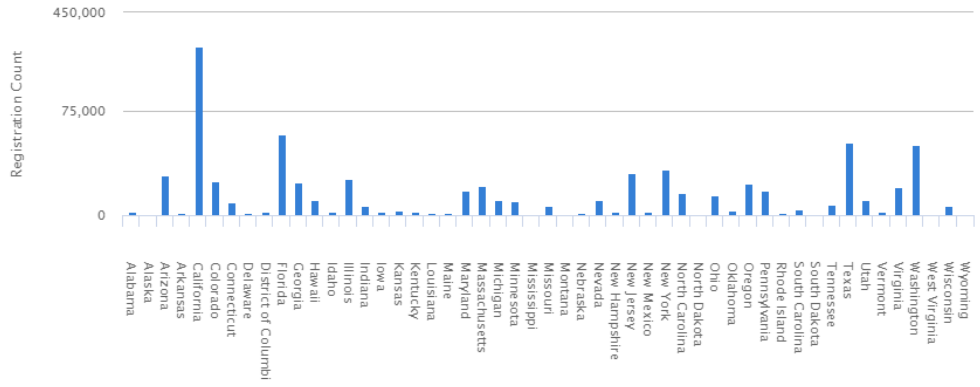

Figure 1-12 Electric Vehicle Registrations by State from the National Renewable Energy Laboratory

California	576142
Pennsylvania	188892
Texas	123208
New York	70487
Arizona	69840
Illinois	50116
North Carolina	37126
Virginia	36422
Arkansas	35332
Colorado	35188
Oklahoma	34420
Kansas	31480
Connecticut	29933
New Jersey	26732
Florida	25918
Iowa	25850
District Of Columbia	25696
Louisiana	23874
Maine	23623
Maryland	23538
Ohio	22934
Massachusetts	21572
Hawaii	20276
Missouri	19778
Kentucky	14686
Indiana	13926
Wyoming	13048
Oregon	11794
North Dakota	11018
Nevada	9698
Country Of Mexico	9506
New Hampshire	9294
Utah	8668
South Dakota	8316
Michigan	8182
Georgia	7722
New Mexico	7130
South Carolina	6536
Rhode Island	6324
Tennessee	5842
Delaware	3630
Minnesota	3558
Alabama	3126
Alaska	1974
Idaho	1828
Wisconsin	1516
Washington	962

Name: State, dtype: int64

Table 1-13. Value Count for “State” Columns of U.S Air pollution Data

Considering that California is a state with the most records from U.S Air Pollution Data and with the highest number of electric vehicles registered for states based on Figure 1-12, the California Electric Vehicles Registration Data sourced from California Energy Commission, which has been recently updated and credible has been selected for evaluating further correlation of EVs and air pollution.

	Vehicle ID	County GEOID	Registration Valid Date	DMV ID	DMV Snapshot	Registration Expiration Date	State Abbreviation	Geography	Vehicle Name
0	CA-002-03597\r	06099	2011-01-01	2	CA Registration Data from CA (12/31/2011)	NaN	CA	County	Chevrolet Volt
1	CA-002-03598\r	06105	2011-01-01	2	CA Registration Data from CA (12/31/2011)	NaN	CA	County	Nissan Leaf
2	CA-002-03599\r	06103	2011-01-01	2	CA Registration Data from CA (12/31/2011)	NaN	CA	County	Chevrolet Volt
3	CA-002-03600\r	06099	2011-01-01	2	CA Registration Data from CA (12/31/2011)	NaN	CA	County	Tesla Roadster
4	CA-002-03601\r	06099	2011-01-01	2	CA Registration Data from CA (12/31/2011)	NaN	CA	County	Tesla Roadster

Table 1-14 California Electric Vehicles Registration Data

```
 #   Column                       Dtype
---  ------                       -----
 0   Vehicle ID                   object
 1   County GEOID                 object
 2   Registration Valid Date      object
 3   DMV ID                       int64
 4   DMV Snapshot                 object
 5   Registration Expiration Date  float64
 6   State Abbreviation           object
 7   Geography                    object
 8   Vehicle Name                 object
dtypes: float64(1), int64(1), object(7)
```

Table 1-15 Data Type of California Electric Vehicles Registration Data

However, the more sufficient understanding and cleaning process will be on hold until the models for air pollution are assessed.

3. Data Preparation

a. Cleaning Null Values Categorical Values and Filtering

```
0   2000-01-01
1   2000-01-01
2   2000-01-01
3   2000-01-01
4   2000-01-02
Name: Date Local, dtype: datetime64[ns]
```

Table 2-1 First 5 rows of "Date Local" column after converting data type to datetime64

By utilizing pandas's .to_datetime()the object type of "Date Local", a crucial value that will enable analyzing the level of pollutant over the year is now converted to the data type of datetime64.

To clean up the data, unnecessary columns for modelling, "State Code", "County Code", "Site Num", "Address", "County", and "City" are dropped. The units measured in for each pollutants are also dropped, but it's noted that NO2 is measured in "parts per billion", while all other 3 pollutants are measured in "parts per million"

	State	Date Local	NO2 Mean	NO2 1st Max Value	NO2 1st Max Hour	NO2 AQI	O3 Mean	O3 1st Max Value	O3 1st Max Hour	O3 AQI	SO2 Mean	SO2 1st Max Value	SO2 1st Max Hour	SO2 AQI	CO Mean	CO 1st Max Value	CO 1st Max Hour	CO AQI
0	Arizona	2000-01-01	19.041667	49.0	19	46	0.022500	0.040	10	34	3.000000	9.0	21	13.0	1.145833	4.2	21	NaN
1	Arizona	2000-01-01	19.041667	49.0	19	46	0.022500	0.040	10	34	3.000000	9.0	21	13.0	0.878947	2.2	23	25.0
2	Arizona	2000-01-01	19.041667	49.0	19	46	0.022500	0.040	10	34	2.975000	6.6	23	NaN	1.145833	4.2	21	NaN
3	Arizona	2000-01-01	19.041667	49.0	19	46	0.022500	0.040	10	34	2.975000	6.6	23	NaN	0.878947	2.2	23	25.0
4	Arizona	2000-01-02	22.958333	36.0	19	34	0.013375	0.032	10	27	1.958333	3.0	22	4.0	0.850000	1.6	23	NaN

Table 2-2 U.S Air Pollution Data After Initial Cleaning of Numerical Values

To accesses the null values in AQI columns, inconsistency in SO2 mean and to focus on the pollutant that its emission is directly related to burning fossil fuels for cars, the data has been condensed into columns of "State, Date"

	State	Date Local	NO2 Mean	NO2 1st Max Value	NO2 1st Max Hour	NO2 AQI
0	Arizona	2000-01-01	19.041667	49.0	19	46
1	Arizona	2000-01-01	19.041667	49.0	19	46
2	Arizona	2000-01-01	19.041667	49.0	19	46
3	Arizona	2000-01-01	19.041667	49.0	19	46
4	Arizona	2000-01-02	22.958333	36.0	19	34

Table 2-3 U.S Air Pollution Data After Second Cleaning of Numerical Values

The EVs dataset is available only for California, so to have a high quality analysis and to observe more accurate correlation between the selected datasets the data set will be focused on the measurements taken in California.

b. Final Data Set

	State	Date Local	NO2 Mean	NO2 1st Max Value	NO2 1st Max Hour	NO2 AQI
3516	California	2000-01-01	14.782609	26.0	4	25
3517	California	2000-01-01	14.782609	26.0	4	25
3518	California	2000-01-01	14.782609	26.0	4	25
3519	California	2000-01-01	14.782609	26.0	4	25
3520	California	2000-01-02	16.043478	30.0	21	28

Table 2-3 Final Data Set After Two Cleaning of Numerical Values and Filtering measurements taken in from California

After selecting the rows the dataset with the value of “State” being California. The final dataset includes datetime type for “Local Date” rather than an object type.

```
Int64Index: 576142 entries, 3516 to 1729196
Data columns (total 6 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   State              576142 non-null  object
 1   Date Local         576142 non-null  datetime64[ns]
 2   NO2 Mean           576142 non-null  float64
 3   NO2 1st Max Value  576142 non-null  float64
 4   NO2 1st Max Hour   576142 non-null  int64
 5   NO2 AQI            576142 non-null  int64
dtypes: datetime64[ns](1), float64(2), int64(2), object(1)
```

Table 2-4. Data Types of Final Data Set From U.S. Air Pollution Data

c. Test Sets

i. stratified sampling

Test sets are selected by stratified sampling using the StratifiedShuffleSplit class from Scikit-Learn's package in order to They are randomly selected but cover all range of because they are splitted based on the "date_category" in which the final data is cut into 5 parts, for example the date measured in between year 2000 to 2004, 2004 to 2008, 2008 to 2012, and so on. This is to minimize any testing bias due to skewed representation of the data.

Sources

1. U.S Air Pollution Data

   https://www.kaggle.com/sogun3/uspollution

2. California Electric Vehicle Registration Data:

   https://www.atlasevhub.com/materials/state-ev-registration-data/#data

3. Jupyter WorkSpace Document for Tables and Figures

   https://colab.research.google.com/drive/1GkI6hc9F5VtUFvG1fSpr1UrY14N2QpkB#scrollTo=Dab8uvNWeyhG