EM 623 - Data Science and Knowledge Discovery

Final Project Report

Understanding U.S Air Pollution & Electric Vehicles in U.S

Kristin Kim

"I pledge in my honor that I have not given or received any unauthorized assistance on this assignment/examination."

Table of Contents

Abstract

Over the last few years, the government and organizations have been spending their time and money to advocate environmental awareness. With the goal of protecting the earth, the initiative to ban single-use plastic products has been taken in several states of the U.S, and people have been trying to be more cautious on their purchases of the products. One of the ways to help reducing air pollutant emission is to own electric cars rather than typical petrol cars that are known to emit 4.6 tons of carbon dioxide per year

The study is aimed to provide a direction to the U.S. Environmental Protection Agency (EPA) whether to proceed with publication of the data and regulatory enforcement to the public or not. Also, the overarching call was to not only study the relationship between two different factors but also to uncover hidden patterns across the sets of data by analyzing them with data science approaches learned in classes. The study is conducted with the date sets obtained from Kaggle and the state website, and both are recent and reliable.

Data science approaches such as predictive approaches and descriptive approaches will be used to explore the potential patterns in characteristics of the electric vehicle population which can be marketable business insight for companies that sell electric vehicles. Exploration of the unseen patterns will be done without any hypothesis, and the data will be analyzed through the data mining approaches to figure out the valuable patterns amongst attributes of the data.

The expected result of the project will be the inverse relationship between the number of electric vehicles and the severity of air pollution in the U.S, and the decrease in pollutant levels as the time passes. Linear regression modeling, decision tree modeling with one hot encoder, and decision tree modeling with label encoder were utilized to validate the insight from data, and

various python libraries and methods support the procedures of them. The major concern and the threats to validity from the expected result will be that the data might have been affected in a greater time scope due the ongoing pandemic that has altered many systems such as production in factories.

CRISP-DM Analysis

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 EVs cars will reduce air pollution since EVs reduce the emission of greenhouse gasses 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.

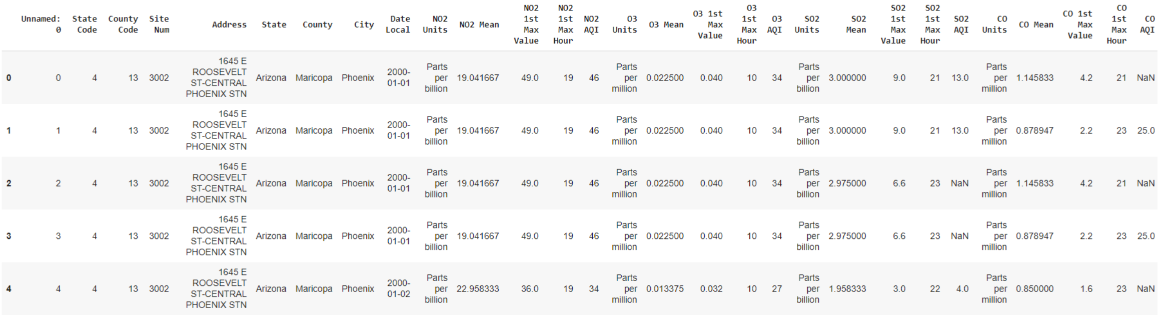
b. Problem statement

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. Data Collection

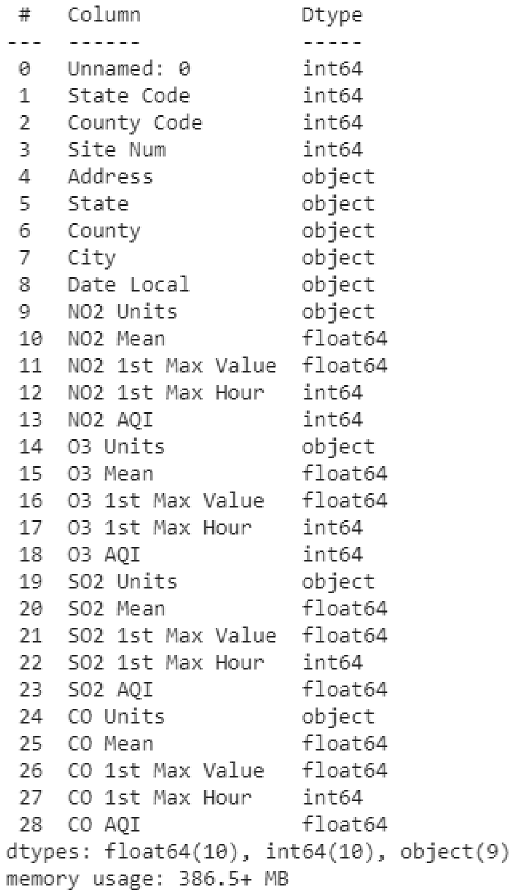
After evaluating the data sets based on how valuable their attributes are to the business goals of this data analysis, U.S. Air Pollution Dataset from Kaggle.com is selected.



	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.

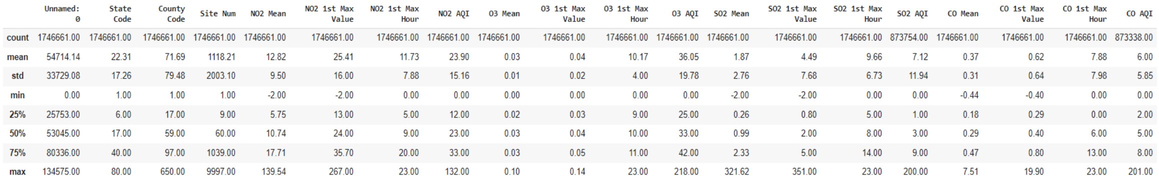


```
 #   Column            Dtype  
---  ------            -----  
 0   Unnamed: 0        int64  
 1   State Code        int64  
 2   County Code       int64  
 3   Site Num          int64  
 4   Address           object 
 5   State             object 
 6   County            object 
 7   City              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 Statistics



	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

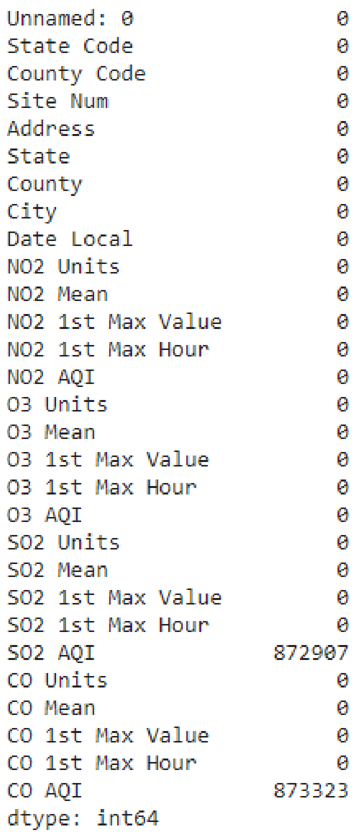
Table 1-3 Descriptive Statistics of Numerical Data Values of U.S Air Pollution Date

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 state, 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

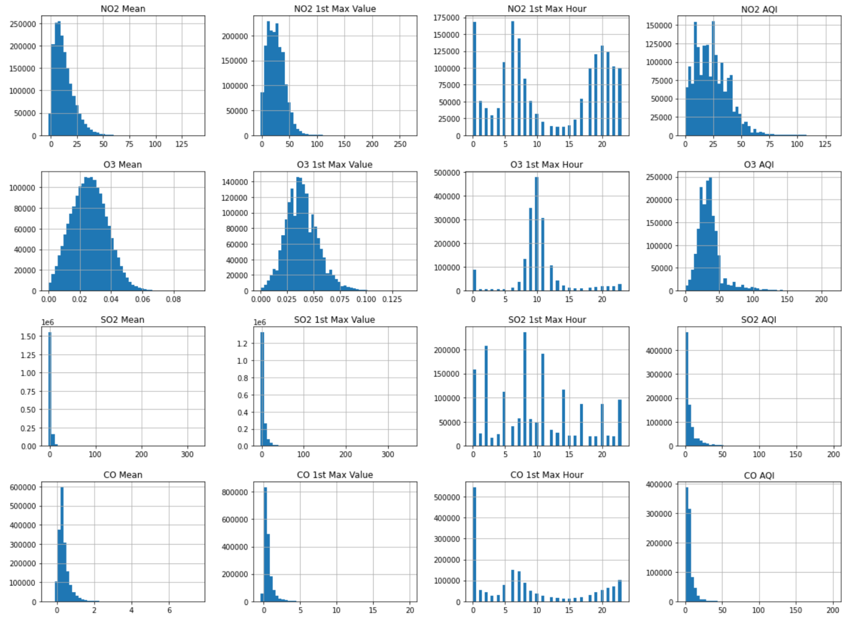
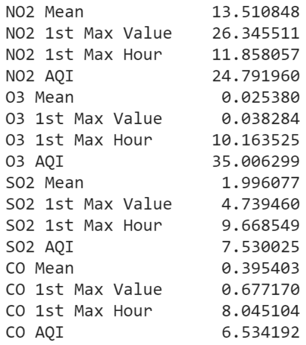


Figure 1-6 Histograms for Numerical value columns in U.S Air Pollution Data



```
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
```

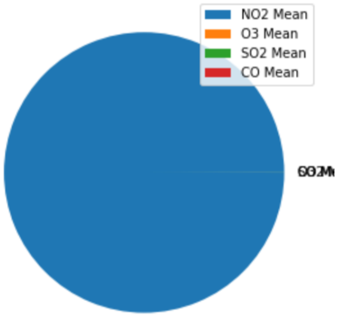


Figure 1-7 Mean Values Grouped by Each Pollutant Columns

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.

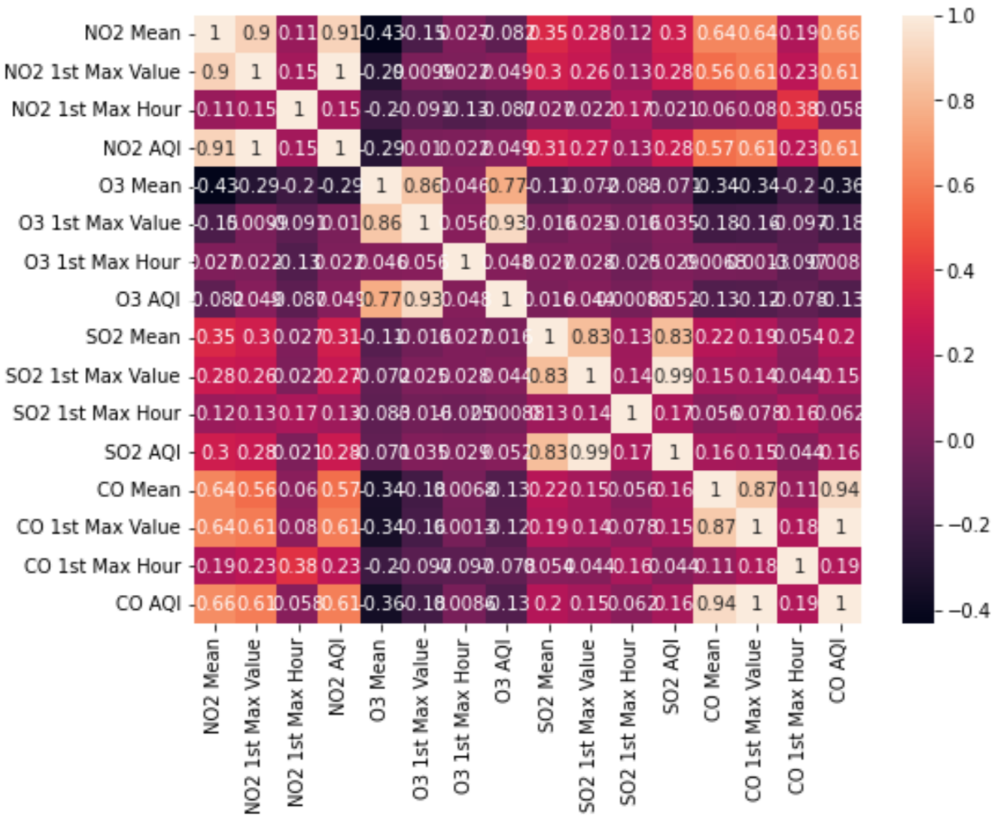


Figure 1-9 Correlation between 4 pollutants from U.S Air Pollution Data

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.

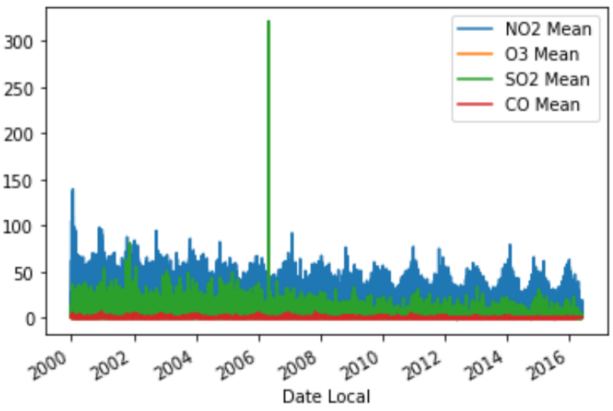


Figure 1-10 Measure of Each Pollutants from 2000 to 2016

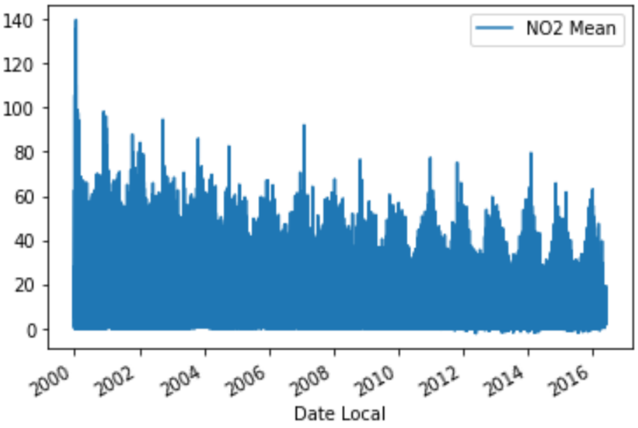


Figure 1-11 Measure of NO2 Mean from 2000 to 2016

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 meaningful information about the data that needed to be taken care of in the preparation process of the data.

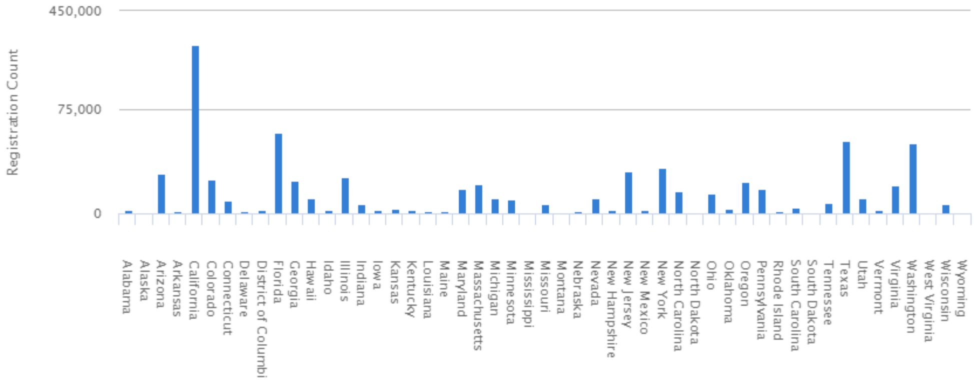
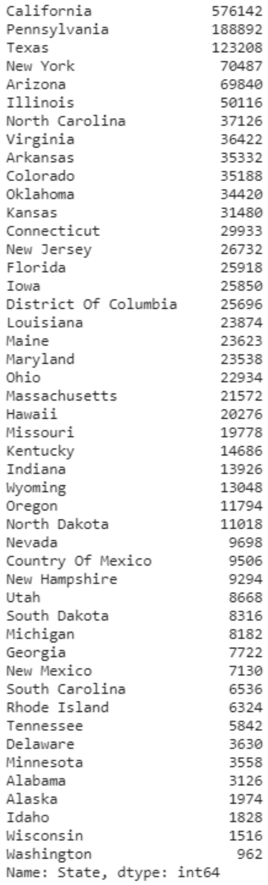


Figure 1-12 Electric Vehicle Registrations by State from the National Renewable Energy Laboratory

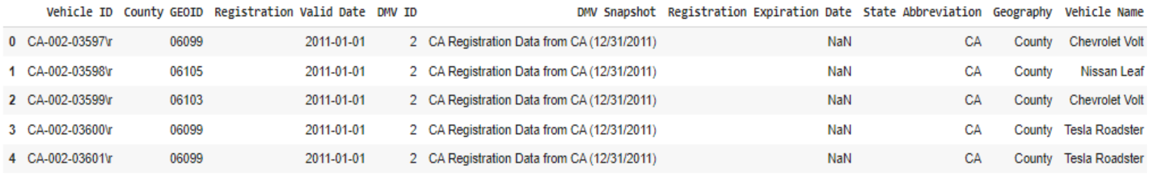


```
California              576142
Pennsylvania            188892
Texas                   123208
New York                 70487
Arizona                  69840
Illinois                 50116
North Carolina           37126
Virginia                 36422
Arkansas                 35332
Colorado                 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

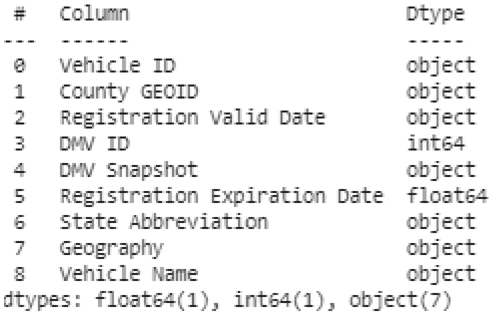
Considering that California is a state with the most records from U.S Air Pollution Data and with the highest number of electric vehicles registered for states based on Figure 1-12, the California Electric

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



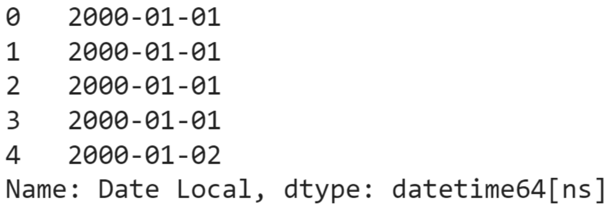
```
 #   Column                        Dtype  
---  ------                        -----  
 0   Vehicle ID                    object 
 1   County GEOID                  object 
 2   Registration Valid Date       object 
 3   DMV ID                        int64  
 4   DMV Snapshot                  object 
 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

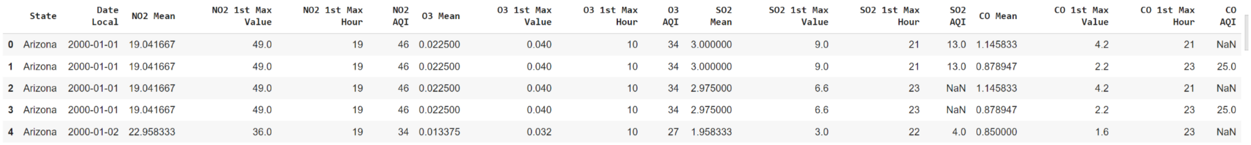


```
0   2000-01-01
1   2000-01-01
2   2000-01-01
3   2000-01-01
4   2000-01-02
Name: Date Local, dtype: datetime64[ns]
```

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.

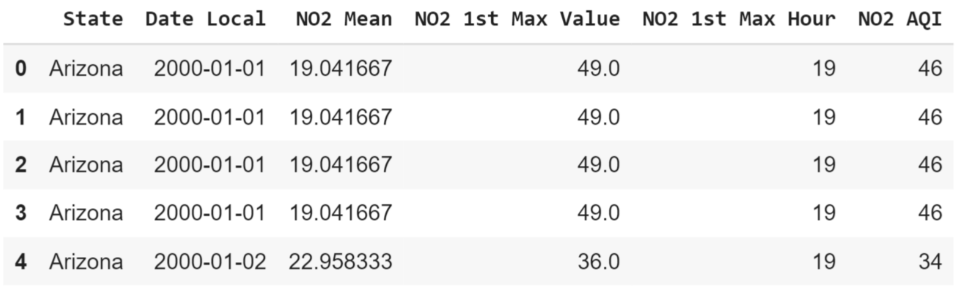
To clean up the data, unnecessary columns for modeling, "State Code", "County Code", "Site Num", "Address", "County", and "City" are dropped. The units measured in for each pollutants are also dropped, but it's noted that NO2 is measured in "parts per billion", while all other 3 pollutants are measured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

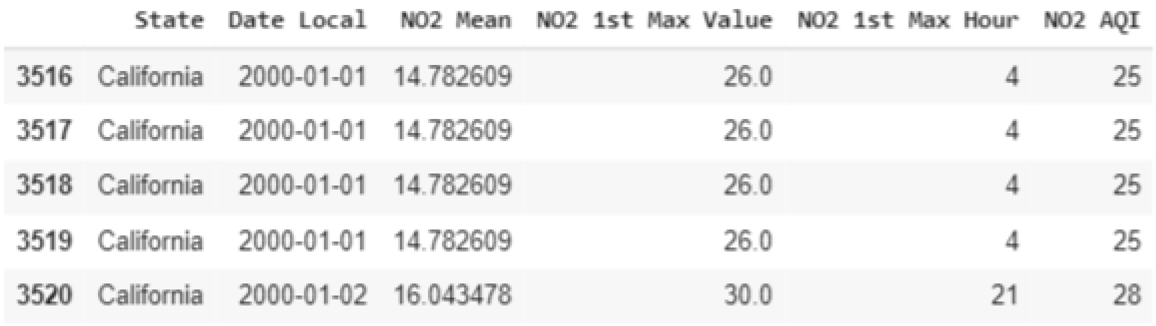


	State	Date Local	NO2 Mean	NO2 1st Max Value	NO2 1st Max Hour	NO2 AQI
0	Arizona	2000-01-01	19.041667	49.0	19	46
1	Arizona	2000-01-01	19.041667	49.0	19	46
2	Arizona	2000-01-01	19.041667	49.0	19	46
3	Arizona	2000-01-01	19.041667	49.0	19	46
4	Arizona	2000-01-02	22.958333	36.0	19	34

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.

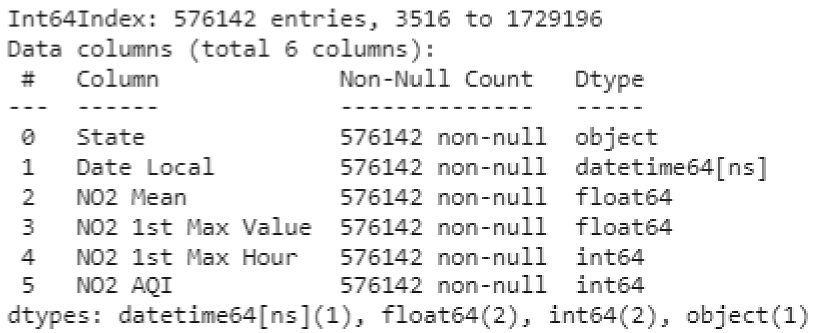
b. Final Data Set



	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

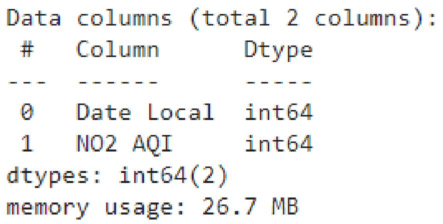
Test sets are selected by stratified sampling using the StratifiedShuffleSplit class from Scikit-Learn's package in order to They are randomly selected but cover all range of because they are splitted based on the "date_category" in which the final data is cut into 5 parts, for example the date measured 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.

CRISP-DM modeling

In order for the U.S. The Environmental Protection Agency (EPA) to make a better decision on whether to enforce law or policies on air pollution, there must be an accurate prediction on future trends of air pollution in the U.S. Even though the best, the most fundamental algorithm to predict is to use Linear Regression. Having NO2 data points for over 15 years spanned several states in the U.S, California with the most number of data sets.

1. Linear Regression

After splitting the data by using the train_test_split class from Scikit-Learn's model selection, the LinearRegression class from Scikit-Learn's linear_model has been imported. By utilizing fit(), the model uses a machine learning library to train on air pollution data. Various trials and errors have been made in modeling this linear regression model, especially the aspects of fitting values and predicting with a test set to result in a correct data frame format. The "Date Local" column is ensured to be converted into the integer values.

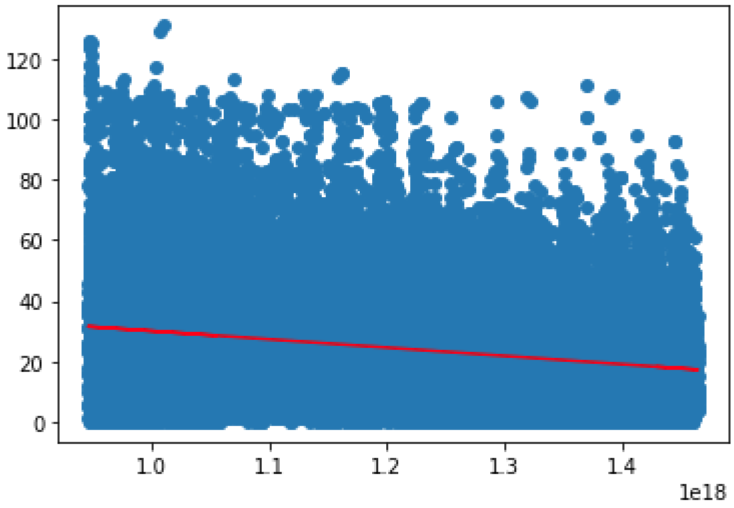


```
Data columns (total 2 columns):
 #   Column      Dtype
---  ------      -----
 0   Date Local  int64
 1   NO2 AQI     int64
dtypes: int64(2)
memory usage: 26.7 MB
```

Figure 4-1 Type Information of Finalized Data Set

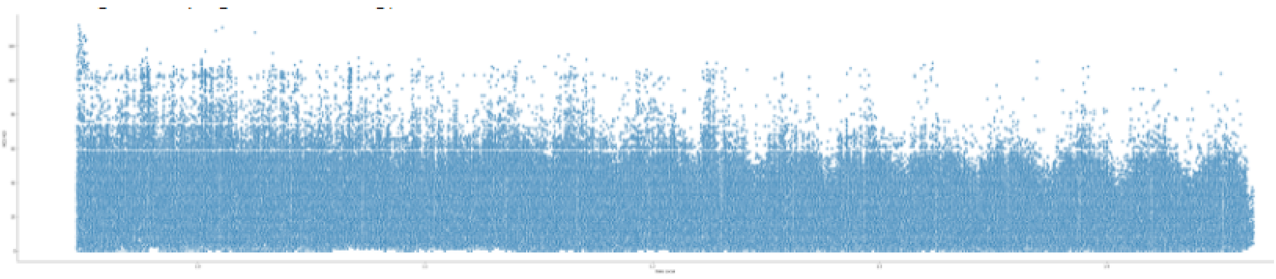
To provide more details on the model, the Intercept of the model, 57.628, and the Coefficient of the model, -2.76*10^-7 have been obtained by utilizing .intercept_() and.coef_() methods of the

LinearRegression class. These supplementary values will be discussed deeply in the Evaluation section.



Figures 4-2 Best Fit Line of the U.S Air Pollution Data Set

As shown in the **Figures 4-2** above, the red trendline of the data successfully indicates the decrease in NO2 AQI scores, and this is expected from the plot below which has been generated before modeling to obtain a better understanding. However, there still exists a lack of accuracy in predicting the values which can be observed in aforementioned outcomes of this model, and this will be mentioned again in the Evaluation section.



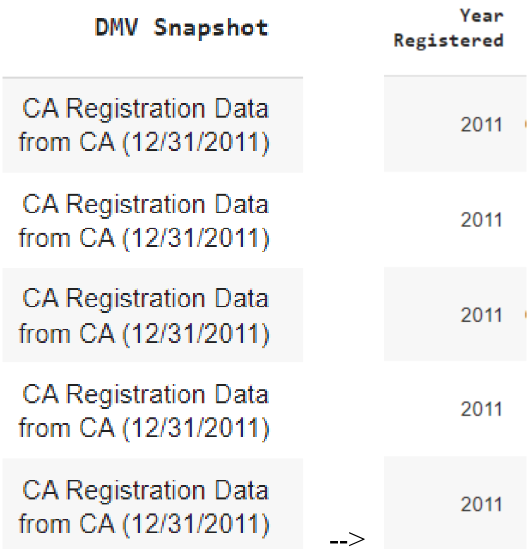
Figures 4-3 Plot drawn by seaborn's .pairplot() methods with x variables of "Date Local" attributes, and y variables of "NO2 AQI" score attribute

2. Decision Tree Classification

Regardless of the accuracy in predicting the outcomes in a relationship of two attributes, “Date Local” and “NO2 AQI”, there exists a decreasing trend clearly in relation, so will move forward to analyze and build a classification model on the Electric Vehicle in California data set as mentioned.

i. Data Pre-Prosessing

Sufficient amount of data pre-prosessing is done to move forward to the modeling process. Null values are cleaned and meaningless columns are dropped while a new column with more insightful values is added. For example, new column to indicate a specific year registered for vehicle is generated from the existing column of “DMV Snapshot"



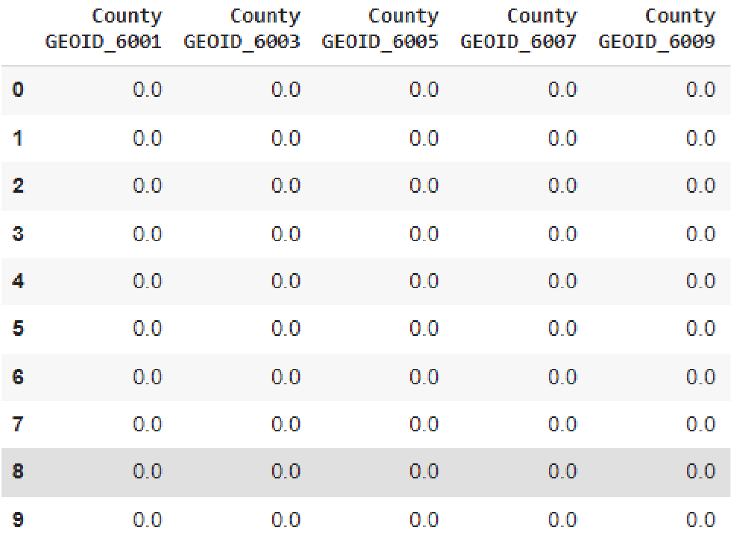
DMV Snapshot	Year Registered
CA Registration Data from CA (12/31/2011)	2011
CA Registration Data from CA (12/31/2011)	2011
CA Registration Data from CA (12/31/2011)	2011
CA Registration Data from CA (12/31/2011)	2011
CA Registration Data from CA (12/31/2011)	2011

Figure 4-4 Year of Registration Extracted from “DMV Snapshot” Column

Also, from a long and numerous values from the “Vehicle Name” column, only the brand has been extracted to efficiently classify instances by the brand of car, year registered, and the county geographical ID.

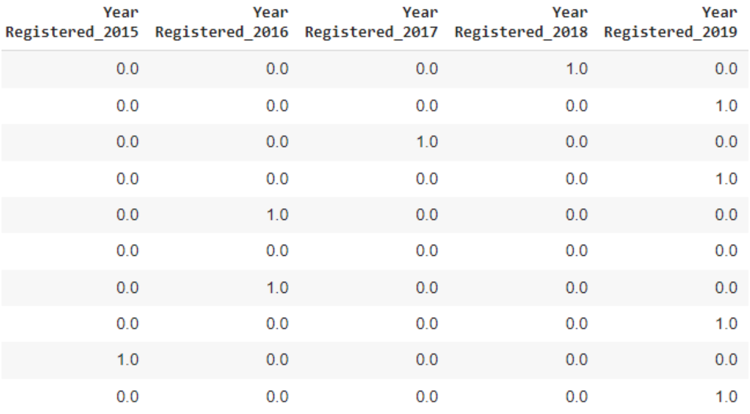
ii. One Hot Coder

In order to handle categorical values, one hot coder class was applied to cover 58 different values of “County GEOID” of [6099, 6105, 6103, 6097, 6067, …. 6011, 6027, 6091]). One hot Encoder was chosen in specific to capture the categorical meaning behind these values of column, since GEOID of 6055 would mean any higher ranking than GEOID of 6005. Also, the difference between two numbers does not mean anything beside their uniqueness. Also, One Hot Encoder is similarly applied to the column of “Registered Year”r as well.



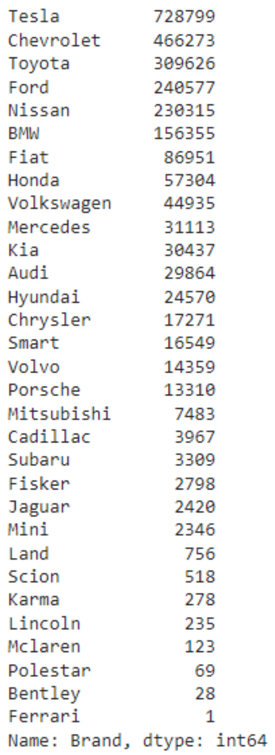
	County GEOID_6001	County GEOID_6003	County GEOID_6005	County GEOID_6007	County GEOID_6009
0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0

Figure 4-5 One Hot Coder Applied To “County GEOID” Column



Year Registered_2015	Year Registered_2016	Year Registered_2017	Year Registered_2018	Year Registered_2019
0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	1.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	0.0	0.0	1.0
0.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	1.0
1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	1.0

Figure 4-6 One Hot Coder Applied To “Year Registered” Column

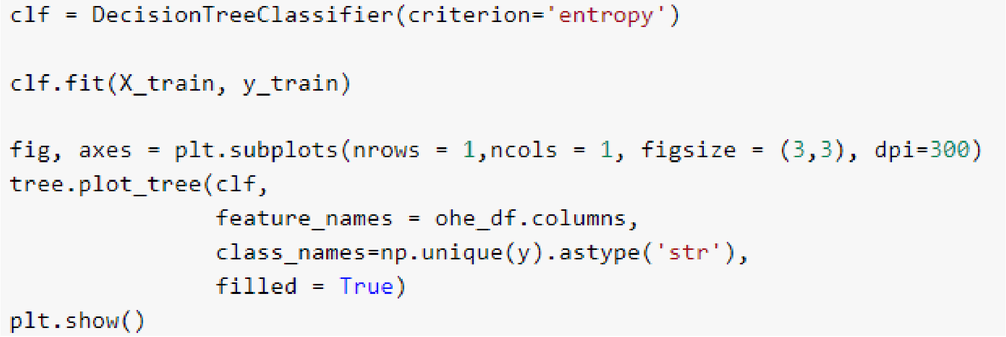


```
Tesla         728799
Chevrolet     466273
Toyota        309626
Ford          240577
Nissan        230315
BMW           156355
Fiat           86951
Honda          57304
Volkswagen     44935
Mercedes       31113
Kia            30437
Audi           29864
Hyundai        24570
Chrysler       17271
Smart          16549
Volvo          14359
Porsche        13310
Mitsubishi      7483
Cadillac        3967
Subaru          3309
Fisker          2798
Jaguar          2420
Mini            2346
Land             756
Scion            518
Karma            278
Lincoln          235
Mclaren          123
Polestar          69
Bentley           28
Ferrari            1
Name: Brand, dtype: int64
```

Figure 4-7 Value counts of "Brand" column of California Electric Vehicle Registration Data

iii. Modeling (One Hot Encoder & Label Hot Encoder)

The DecisionTreeClassifier class from Scikit-Learn's tree is imported after successfully splitting the data by using the train_test_split class from Scikit-Learn's model selection. By utilizing fit() class, the pre-defined classifier itself performs an accurate classification on the data set. Also, 'entropy' is chosen for the criterion attribute of the DecisionTreeClassifier. After specifying the attribute of the plot, the plot is displayed to represent the modeled classification.



```
clf = DecisionTreeClassifier(criterion='entropy')

clf.fit(X_train, y_train)

fig, axes = plt.subplots(nrows = 1,ncols = 1, figsize = (3,3), dpi=300)
tree.plot_tree(clf,
               feature_names = ohe_df.columns,
               class_names=np.unique(y).astype('str'),
               filled = True)
plt.show()
```

Figure 4-8 Decision Tree Classification In Python Using Sklearn's Class

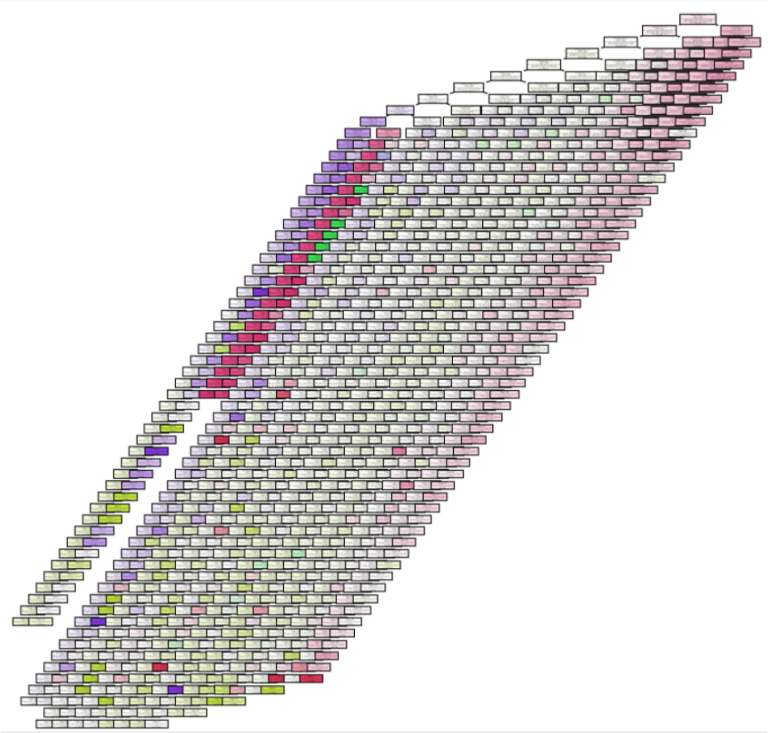


Figure 4-1 Result of Decision Tree Classification (One Hot Encoder on X)

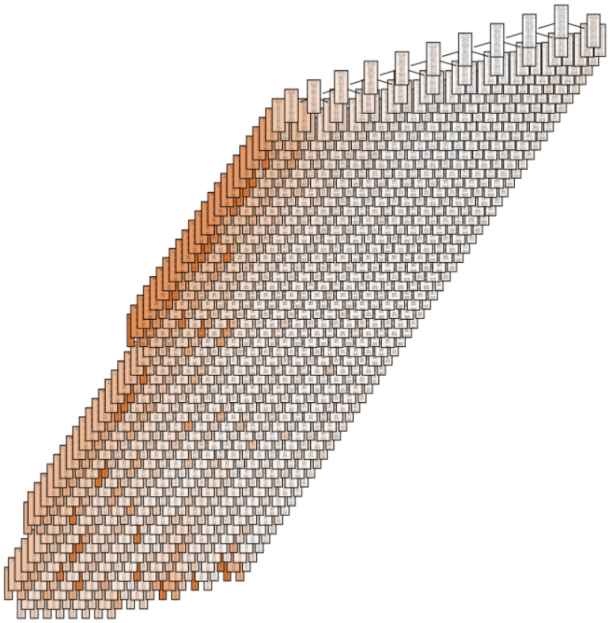


Table 4-1 Result of Decision Tree Classification (One Hot Encoder on both X and y)

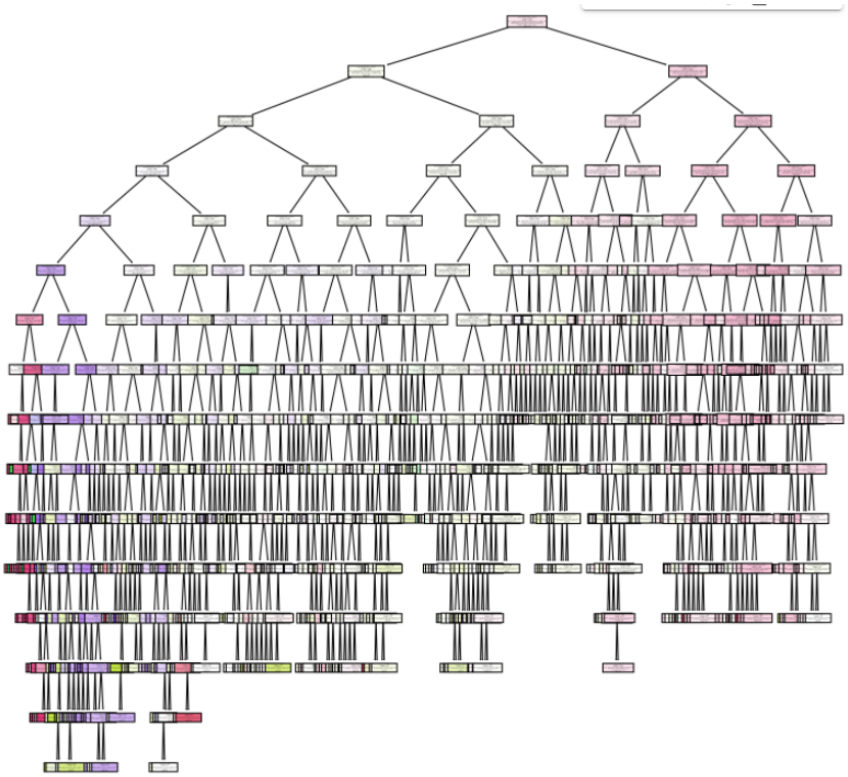
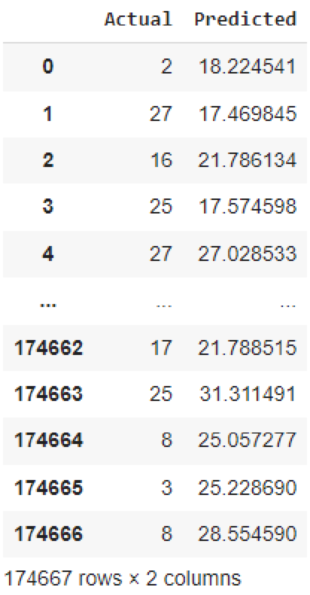


Table 4-1 Result of Decision Tree Classification (Label Encoder on y)

CRISP-DM Evaluation

1. Evaluation of Linear Regression

Predicted values for the test data are generated by .predict() method of Linear Regressor, then .mean_absolute_error(), metrics.mean_squared_error(), and .sqrt() methods from Scikit Learn’s metrics class evaluated the model’s accuracy in depth.



	Actual	Predicted
0	2	18.224541
1	27	17.469845
2	16	21.786134
3	25	17.574598
4	27	27.028533
...	...	...
174662	17	21.788515
174663	25	31.311491
174664	8	25.057277
174665	3	25.228690
174666	8	28.554590

174667 rows × 2 columns

Table 5-1 Actual and Predicted NO2 AQI Values

a. **Mean Absolute Error**: 11.64

b. **Mean Squared Error:** 213.43

c. **Root Mean Squared Error:** 14.60

d. **R square**: 0.07

e. **Score:** 0.07135808351314676

Based on 5 different reliable evaluation values, this model is analyzed to be not accurate, hence will not help predicting the future scores with the. To elaborate more on those values, Mean Absolute Error indicates the mean of the absolute value of the errors, Mean Squared Error explains the mean of the squared errors, and Root Mean Squared Error is the square root of the mean of the squared errors and. Also, the 5th evaluation values, score is obtained from the .score() method of Linear Regression classand and it represents coefficient of determination of the prediction.

2. Evaluation of Decision Tree Classification

a. **Model generated with One Hot Encoder**

   - Accuracy: 0.32

   After the First round of evaluation on the model, to enhance the accuracy, another encoding method was applied when building the model. However, this model with Label Encoder gives the same accuracy as the one utilized byOne Hot Encoder.

To restate the accuracy score,

f. Model generated with Label Encoder

   - Accuracy: 0.32

Though two previously shown Tree Classification generated with encoding methods seem to be not as effective as wanted to be, there are still other encoding methods that are worth giving a shot. For example, Binary Encoding could be used. According to the **Figure 5-2** below which showcases the accuracy of decision trees modeled by different encoding methods, Binary Encoding can be tried. Also, the major trend emphasized by Soukhavong along with his **Figure 5-2**, is that the more unbalanced dataset will result in a higher accuracy of encoding, while a more balanced dataset will result in a lower accuracy of encoding.

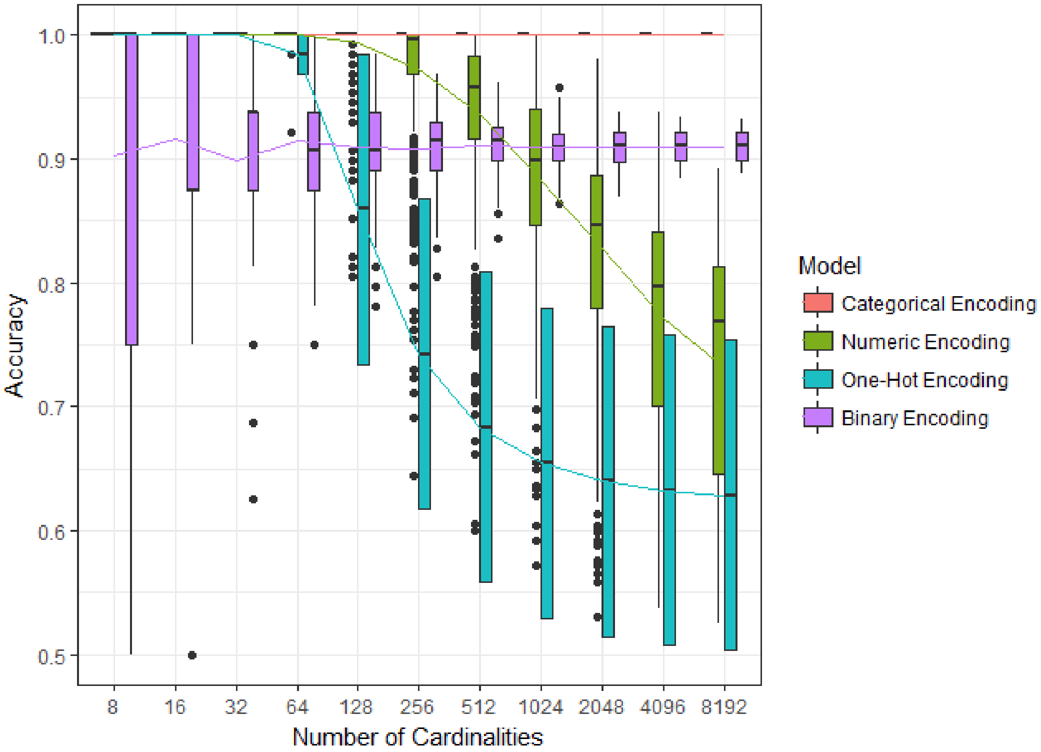


Figure 5-2 Accuracy of Decision Tree model by Encoding by Damien Soukhavong, Data Scientist/ Architect

Though the Linear Regression model and three different Decision Tree Classification models are generated to be less accurate than desired to be, overall, evaluation of the models was successfully done since it evaluates the models in an objective manner with reliable values. More precise modeling, perhaps, more advanced data analysis in pre-processing models is encouraged in the future to enable these models to predict even highly accurate values.

Deployment

After successful modeling and evaluations, the current models can be used for accessing the similar data, in other words, the model can return the expected air pollutant in future years of 2022 or 2023 based on the training data set. The linear regression model will be very helpful to estimate the air pollution level caused by NO2 in an upcoming year and even after. If it is possible to develop a model with a greater range of data in terms of time and state location, the execution of the enhanced model will be highly in demand by the U.S. Environmental Protection Agency in predicting the future pollutant level in the U.S. Also, the development of future models can be done by utilizing the techniques used in this report for the current model. The decision tree models with both one hot encoder and label encoder will not be as successful as the linear regression model, but the sampling process and the procedure of applying the testing data set to the trained model will be useful for setting up the future procedures.

Conclusion

Future research with more sophisticated modeling and evaluation is much needed to access this environmental issue and hidden trend better. Though the current model successfully shows a distinct decrease in the data points for the first data set, the U.S Air Pollution data, from the researcher's point of view, further research could be done by modeling the data with different machine learning algorithm, such as Naive Bayesian, also by hyper cleaning the first data set to enhance the accuracy in linear regression model. Overall, the major goal of understanding the U.S Air Pollution from the U.S. Environmental Protection Agency's perspective was successfully met, also more importantly, hands on experience of utilizing various python libraries and methods, Scatter_matrix, Pandas, Numpy, Sklearn, Train_test_split, DecisionTreeClassifier, LinearRegression metric, s r2_score, accuracy_score, OneHotEncoder, LabelEncoder, Seaborn, Matplotlib and StratifiedShuffleSplit, have gained.

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