EV Registration Analysis: Assignment

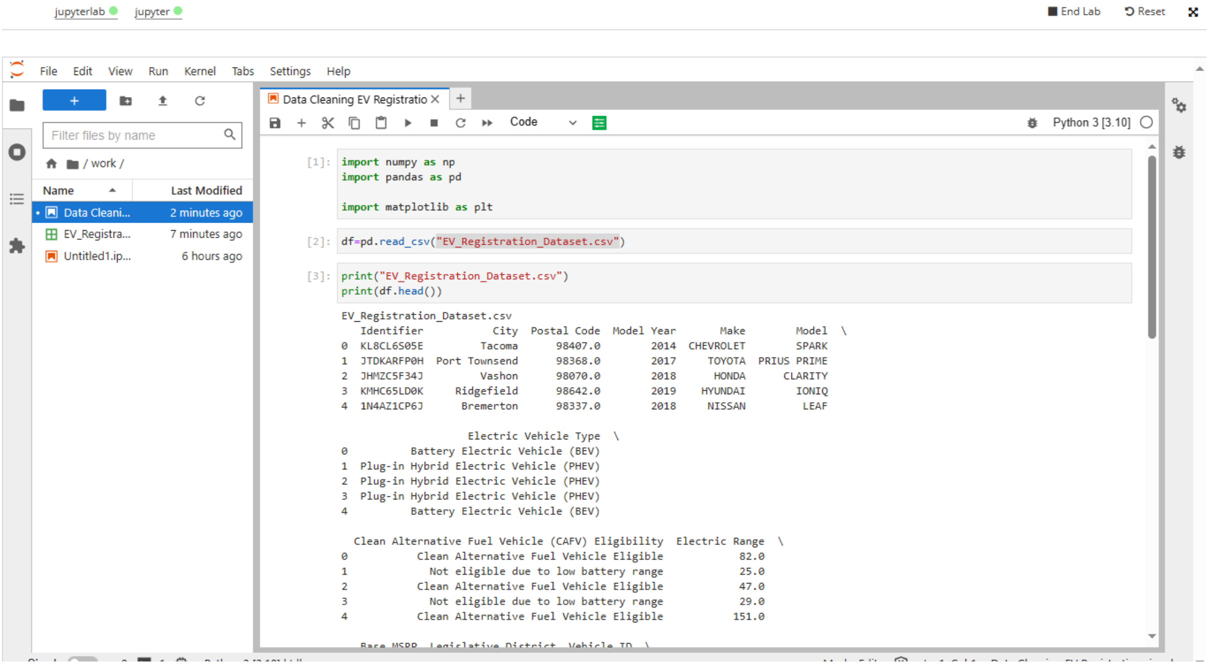
Data Science: Python Project: DATA Cleaning

Name: Chumani Mazwi

Email: cmazwi5@gmail.com

Institution: Deviare| MICT SETA| Belong|

Clean the data by removing incorrect, duplicate or missing data.

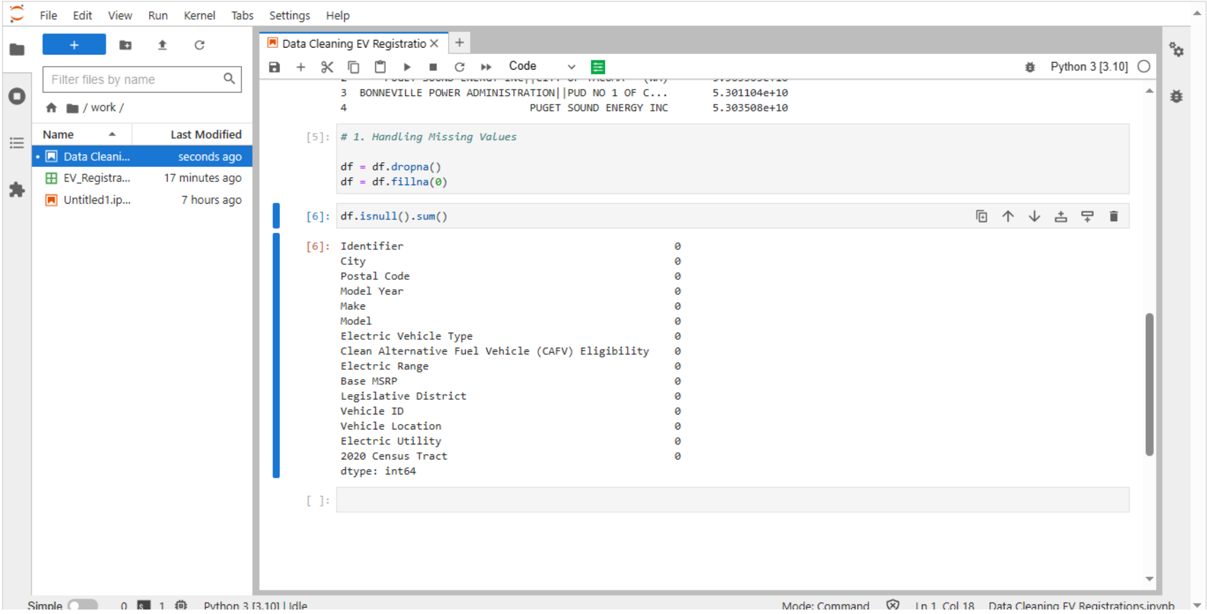


```
[1]: import numpy as np
     import pandas as pd

     import matplotlib as plt
```

```
[2]: df=pd.read_csv("EV_Registration_Dataset.csv")
```

```
[3]: print("EV_Registration_Dataset.csv")
     print(df.head())
```

```
EV_Registration_Dataset.csv
   Identifier           City  Postal Code  Model Year       Make        Model  \
0  KL8CL6S05E         Tacoma      98407.0        2014  CHEVROLET        SPARK   
1  JTDKARFP0H  Port Townsend      98368.0        2017     TOYOTA  PRIUS PRIME   
2  JHMZC5F34J         Vashon      98070.0        2018      HONDA     CLARITY   
3  KMHC65LD0K     Ridgefield      98642.0        2019    HYUNDAI        IONIQ   
4  1N4AZ1CP6J      Bremerton      98337.0        2018     NISSAN         LEAF   

                    Electric Vehicle Type  \
0          Battery Electric Vehicle (BEV)   
1  Plug-in Hybrid Electric Vehicle (PHEV)   
2  Plug-in Hybrid Electric Vehicle (PHEV)   
3  Plug-in Hybrid Electric Vehicle (PHEV)   
4          Battery Electric Vehicle (BEV)   

  Clean Alternative Fuel Vehicle (CAFV) Eligibility  Electric Range  \
0           Clean Alternative Fuel Vehicle Eligible            82.0   
1             Not eligible due to low battery range            25.0   
2           Clean Alternative Fuel Vehicle Eligible            47.0   
3             Not eligible due to low battery range            29.0   
4           Clean Alternative Fuel Vehicle Eligible           151.0   

   Base MSRP  Legislative District  Vehicle ID  \
```

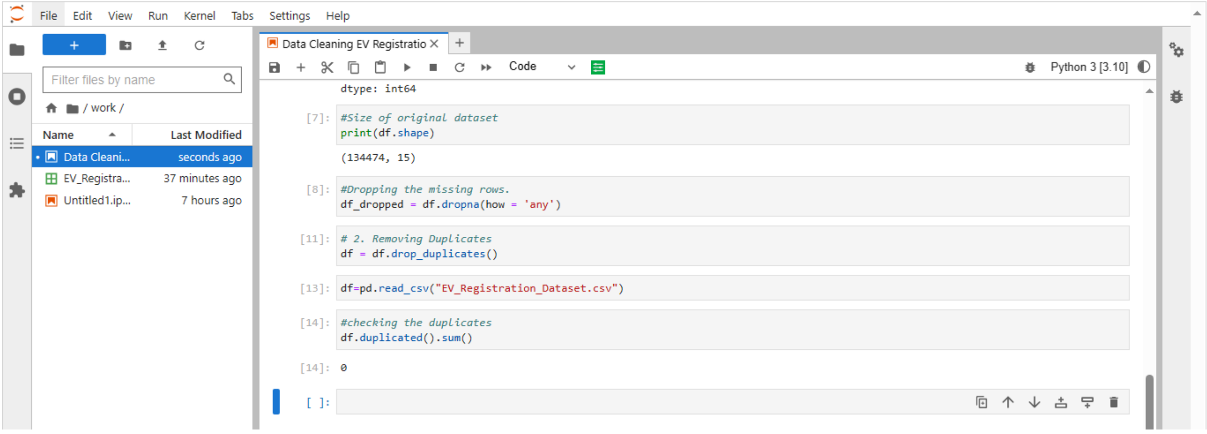


```
3  BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF C...       5.301104e+10
4                               PUGET SOUND ENERGY INC       5.303508e+10
```

```
[5]: # 1. Handling Missing Values

     df = df.dropna()
     df = df.fillna(0)
```

```
[6]: df.isnull().sum()
```

```
[6]: Identifier                                           0
     City                                                 0
     Postal Code                                          0
     Model Year                                           0
     Make                                                 0
     Model                                                0
     Electric Vehicle Type                                0
     Clean Alternative Fuel Vehicle (CAFV) Eligibility    0
     Electric Range                                       0
     Base MSRP                                            0
     Legislative District                                 0
     Vehicle ID                                           0
     Vehicle Location                                     0
     Electric Utility                                     0
     2020 Census Tract                                    0
     dtype: int64
```



```
dtype: int64
```

```
[7]: #Size of original dataset
     print(df.shape)

     (134474, 15)
```

```
[8]: #Dropping the missing rows.
     df_dropped = df.dropna(how = 'any')
```

```
[11]: # 2. Removing Duplicates
      df = df.drop_duplicates()
```

```
[13]: df=pd.read_csv("EV_Registration_Dataset.csv")
```

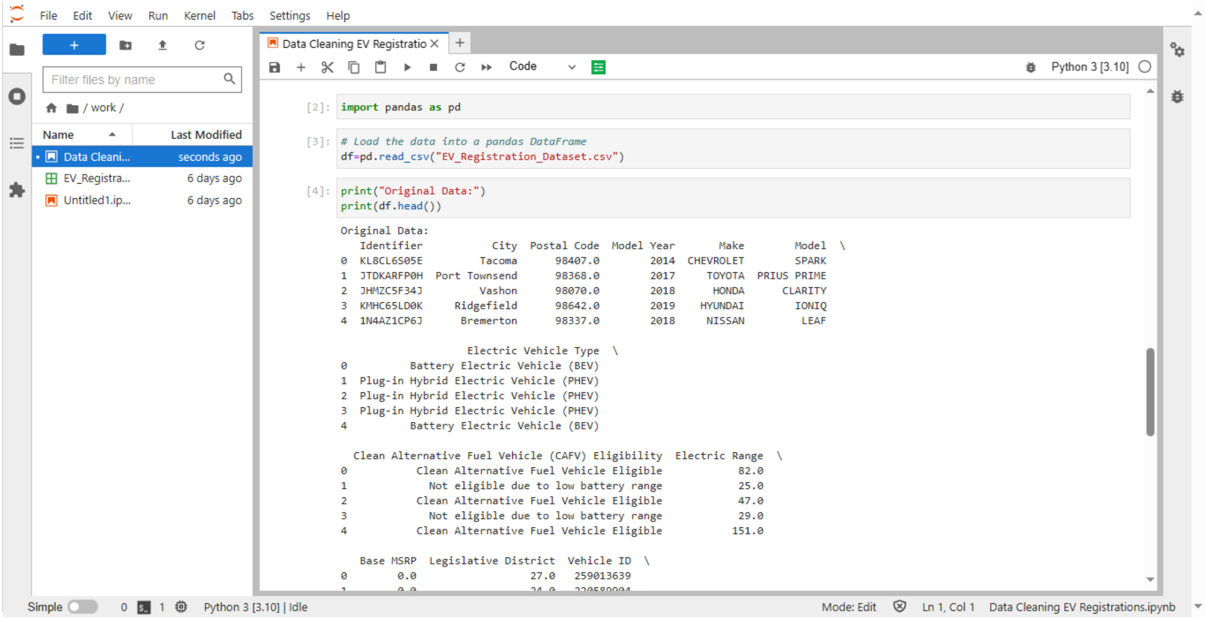
```
[14]: #checking the duplicates
      df.duplicated().sum()
```

```
[14]: 0
```

Clean the data by removing incorrect, duplicate or missing data:

Step 1:

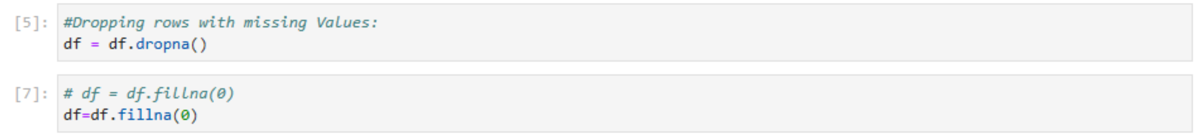
Importing the necessary libraries and loading the data set:



```
[2]: import pandas as pd

[3]: # Load the data into a pandas DataFrame
     df=pd.read_csv("EV_Registration_Dataset.csv")

[4]: print("Original Data:")
     print(df.head())

     Original Data:
        Identifier          City  Postal Code  Model Year       Make         Model  \
     0  KL8CL6S05E        Tacoma      98407.0        2014  CHEVROLET         SPARK
     1  JTDKARFP0H  Port Townsend      98368.0        2017     TOYOTA  PRIUS PRIME
     2  JHMZC5F34J        Vashon      98070.0        2018      HONDA      CLARITY
     3  KMHC65LD0K    Ridgefield      98642.0        2019    HYUNDAI        IONIQ
     4  1N4AZ1CP6J     Bremerton      98337.0        2018     NISSAN          LEAF

                   Electric Vehicle Type  \
     0          Battery Electric Vehicle (BEV)
     1  Plug-in Hybrid Electric Vehicle (PHEV)
     2  Plug-in Hybrid Electric Vehicle (PHEV)
     3  Plug-in Hybrid Electric Vehicle (PHEV)
     4          Battery Electric Vehicle (BEV)

       Clean Alternative Fuel Vehicle (CAFV) Eligibility  Electric Range  \
     0           Clean Alternative Fuel Vehicle Eligible            82.0
     1             Not eligible due to low battery range            25.0
     2           Clean Alternative Fuel Vehicle Eligible            47.0
     3             Not eligible due to low battery range            29.0
     4           Clean Alternative Fuel Vehicle Eligible           151.0

        Base MSRP  Legislative District  Vehicle ID  \
     0        0.0                  27.0   259013639
```

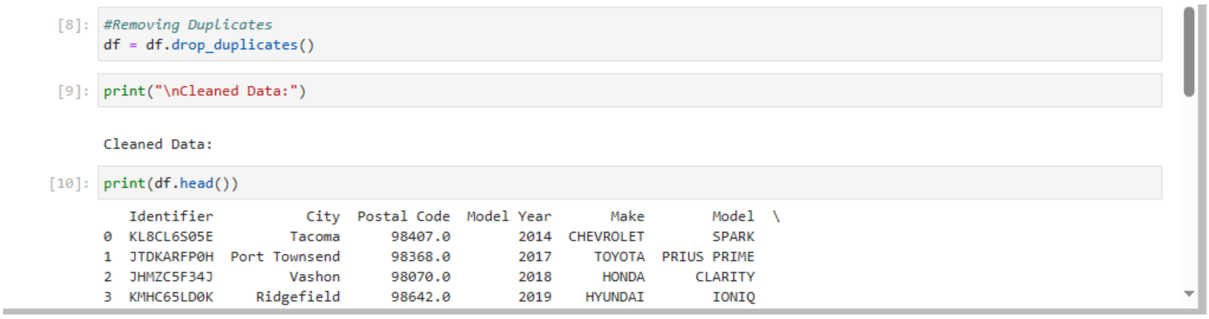
Step 2: Data Cleaning Steps: 1 { # Handling Missing Values}

Dropping rows with missing values| Alternatively, fill missing values with a specific value e.g (0):



```
[5]: #Dropping rows with missing Values:
     df = df.dropna()

[7]: # df = df.fillna(0)
     df=df.fillna(0)
```

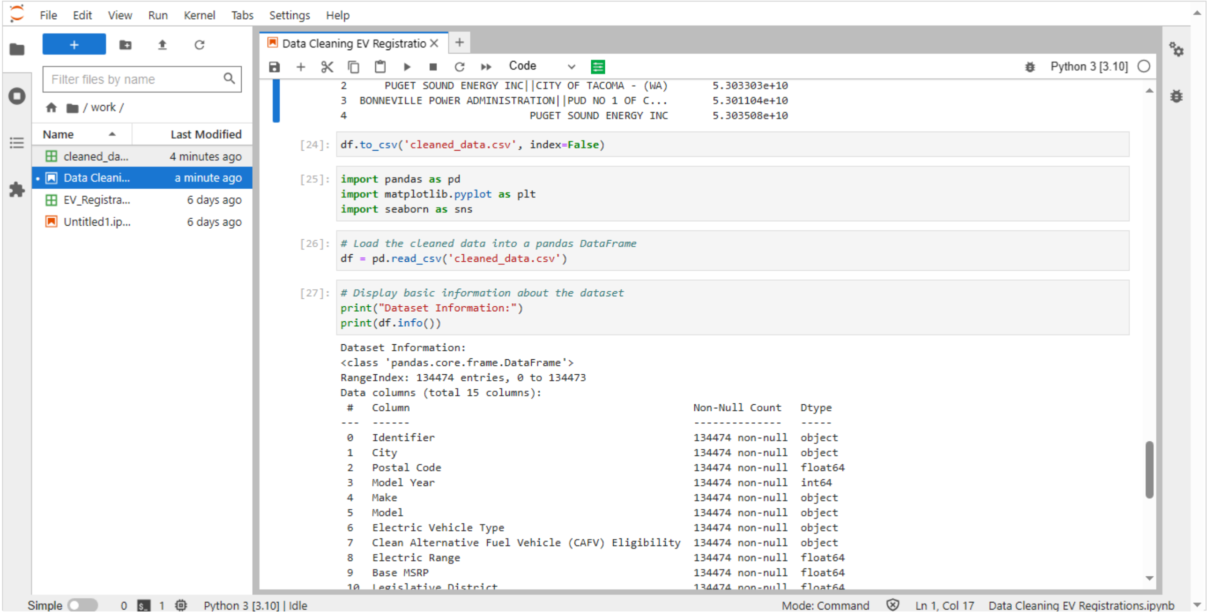
Step 3: Removing Duplicates and Printing the Cleaned Data



```
[8]: #Removing Duplicates
     df = df.drop_duplicates()

[9]: print("\nCleaned Data:")

     Cleaned Data:

[10]: print(df.head())

        Identifier          City  Postal Code  Model Year       Make         Model  \
     0  KL8CL6S05E        Tacoma      98407.0        2014  CHEVROLET         SPARK
     1  JTDKARFP0H  Port Townsend      98368.0        2017     TOYOTA  PRIUS PRIME
     2  JHMZC5F34J        Vashon      98070.0        2018      HONDA      CLARITY
     3  KMHC65LD0K    Ridgefield      98642.0        2019    HYUNDAI        IONIQ
```

Analyse the data and understand the nature of the trends present in the data:

The first step is to import the necessary libraries, load the data into the pandas Data frame and display the basic information about the dataset:



```
2      PUGET SOUND ENERGY INC||CITY OF TACOMA - (WA)       5.303303e+10
3  BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF C...       5.301104e+10
4                             PUGET SOUND ENERGY INC       5.303508e+10
```

```
[24]: df.to_csv('cleaned_data.csv', index=False)
```

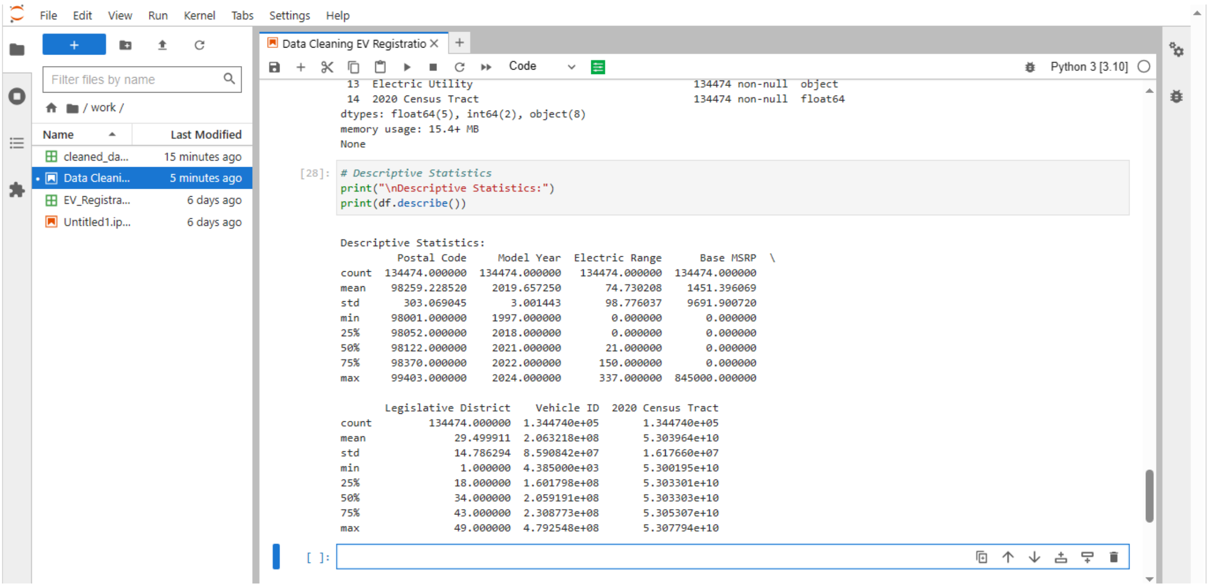
```
[25]: import pandas as pd
      import matplotlib.pyplot as plt
      import seaborn as sns
```

```
[26]: # Load the cleaned data into a pandas DataFrame
      df = pd.read_csv('cleaned_data.csv')
```

```
[27]: # Display basic information about the dataset
      print("Dataset Information:")
      print(df.info())
```

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 134474 entries, 0 to 134473
Data columns (total 15 columns):
 #   Column                                             Non-Null Count   Dtype
---  ------                                             --------------   -----
 0   Identifier                                         134474 non-null  object
 1   City                                               134474 non-null  object
 2   Postal Code                                        134474 non-null  float64
 3   Model Year                                         134474 non-null  int64
 4   Make                                               134474 non-null  object
 5   Model                                              134474 non-null  object
 6   Electric Vehicle Type                              134474 non-null  object
 7   Clean Alternative Fuel Vehicle (CAFV) Eligibility  134474 non-null  object
 8   Electric Range                                     134474 non-null  float64
 9   Base MSRP                                          134474 non-null  float64
 10  Legislative District                               134474 non-null  float64
```

The next step involves applying descriptive statistics about the dataset:


File Edit View Run Kernel Tabs Settings Help
Filter files by name
/ work /
Name
Last Modified
cleaned_da... 15 minutes ago
Data Cleani... 5 minutes ago
EV_Registra... 6 days ago
Untitled1.ip... 6 days ago
Data Cleaning EV Registratio
Code
Python 3 [3.10]

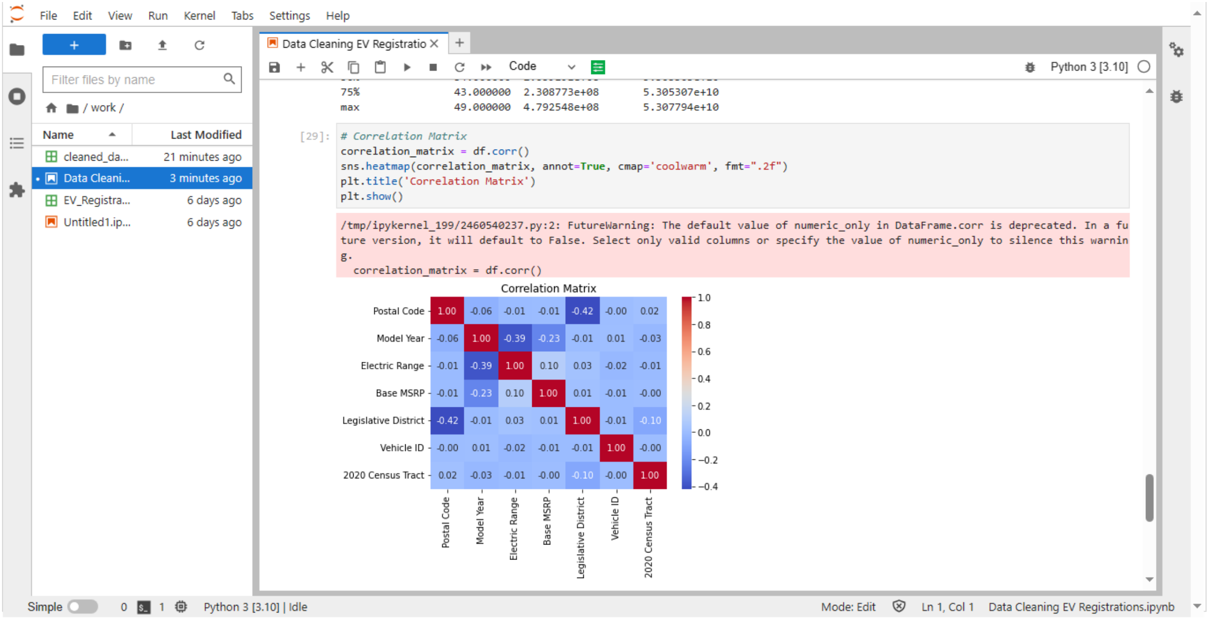
```
 13  Electric Utility                                   134474 non-null  object
 14  2020 Census Tract                                  134474 non-null  float64
dtypes: float64(5), int64(2), object(8)
memory usage: 15.4+ MB
None
```

```
[28]: # Descriptive Statistics
      print("\nDescriptive Statistics:")
      print(df.describe())
```

```
Descriptive Statistics:
         Postal Code     Model Year  Electric Range      Base MSRP  \
count  134474.000000  134474.000000   134474.000000  134474.000000
mean    98259.228520    2019.657250       74.730208    1451.396069
std       303.069045       3.001443       98.776037    9691.900720
min     98001.000000    1997.000000        0.000000       0.000000
25%     98052.000000    2018.000000        0.000000       0.000000
50%     98122.000000    2021.000000       21.000000       0.000000
75%     98370.000000    2022.000000      150.000000       0.000000
max     99403.000000    2024.000000      337.000000  845000.000000

       Legislative District    Vehicle ID  2020 Census Tract
count         134474.000000  1.344740e+05       1.344740e+05
mean              29.499911  2.063218e+08       5.303964e+10
std               14.786294  8.590842e+07       1.617660e+07
min                1.000000  4.385000e+03       5.300195e+10
25%               18.000000  1.601798e+08       5.303301e+10
50%               34.000000  2.059191e+08       5.303303e+10
75%               43.000000  2.308773e+08       5.305307e+10
max               49.000000  4.792548e+08       5.307794e+10
```

```
[ ]:
```

The next step involves using the correlation matrix to interpret the dataset:


File Edit View Run Kernel Tabs Settings Help
Filter files by name
/ work /
Name
Last Modified
cleaned_da... 21 minutes ago
Data Cleani... 3 minutes ago
EV_Registra... 6 days ago
Untitled1.ip... 6 days ago
Data Cleaning EV Registratio
Code
Python 3 [3.10]
Correlation Matrix
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Simple
Python 3 [3.10] | Idle
Mode: Edit
Ln 1, Col 1
Data Cleaning EV Registrations.ipynb

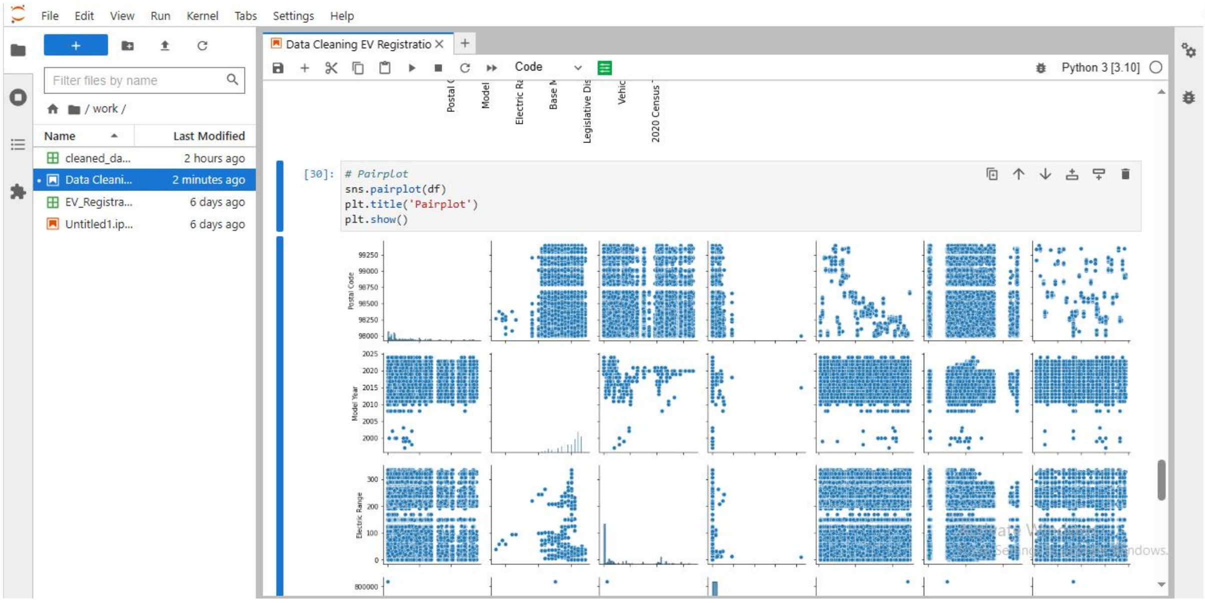
```
75%               43.000000  2.308773e+08       5.305307e+10
max               49.000000  4.792548e+08       5.307794e+10
```

```
[29]: # Correlation Matrix
      correlation_matrix = df.corr()
      sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
      plt.title('Correlation Matrix')
      plt.show()
```

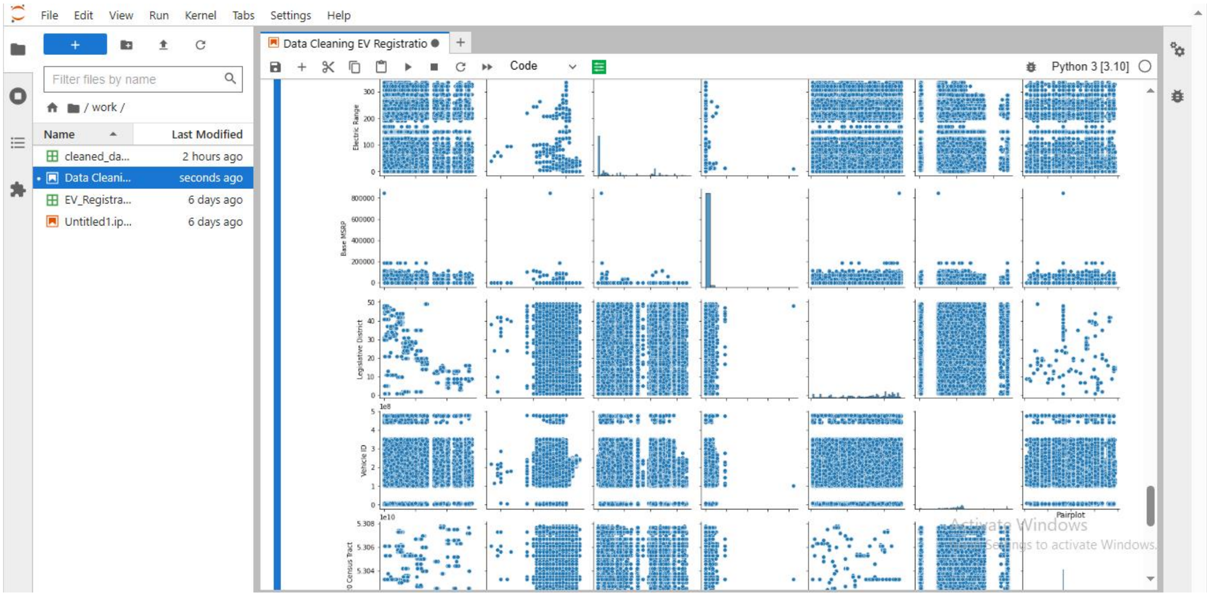
```
/tmp/ipykernel_199/2460540237.py:2: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a fu
ture version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warnin
g.
  correlation_matrix = df.corr()
```

	Postal Code	Model Year	Electric Range	Base MSRP	Legislative District	Vehicle ID	2020 Census Tract
Postal Code	1.00	-0.06	-0.01	-0.01	-0.42	-0.00	0.02
Model Year	-0.06	1.00	-0.39	-0.23	-0.01	0.01	-0.03
Electric Range	-0.01	-0.39	1.00	0.10	0.03	-0.02	-0.01
Base MSRP	-0.01	-0.23	0.10	1.00	0.01	-0.01	-0.00
Legislative District	-0.42	-0.01	0.03	0.01	1.00	-0.01	-0.10
Vehicle ID	-0.00	0.01	-0.02	-0.01	-0.01	1.00	-0.00
2020 Census Tract	0.02	-0.03	-0.01	-0.00	-0.10	-0.00	1.00

The next step involves using the pair plot to interpret the dataset:



```
[30]: # Pairplot
      sns.pairplot(df)
      plt.title('Pairplot')
      plt.show()
```


File
Edit
View
Run
Kernel
Tabs
Settings
Help
Filter files by name
/ work /
Name
Last Modified
cleaned_da... 2 hours ago
Data Cleani... seconds ago
EV_Registra... 6 days ago
Untitled1.ip... 6 days ago
Data Cleaning EV Registratio
Code
Python 3 [3.10]
Electric Range
Base MSRP
Legislative District
Vehicle ID
1e8
0 Census Tract
1e10
Pairplot
Activate Windows
Go to Settings to activate Windows.