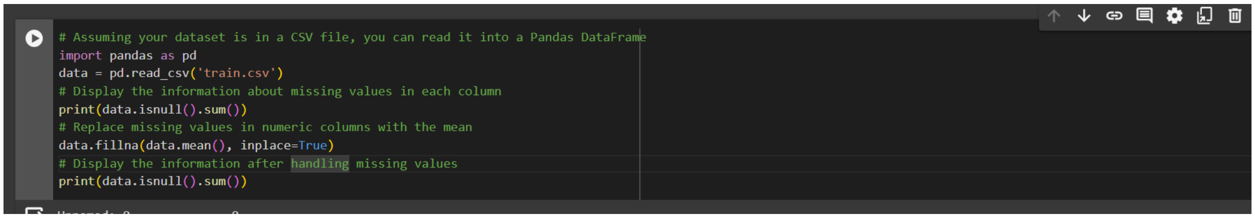
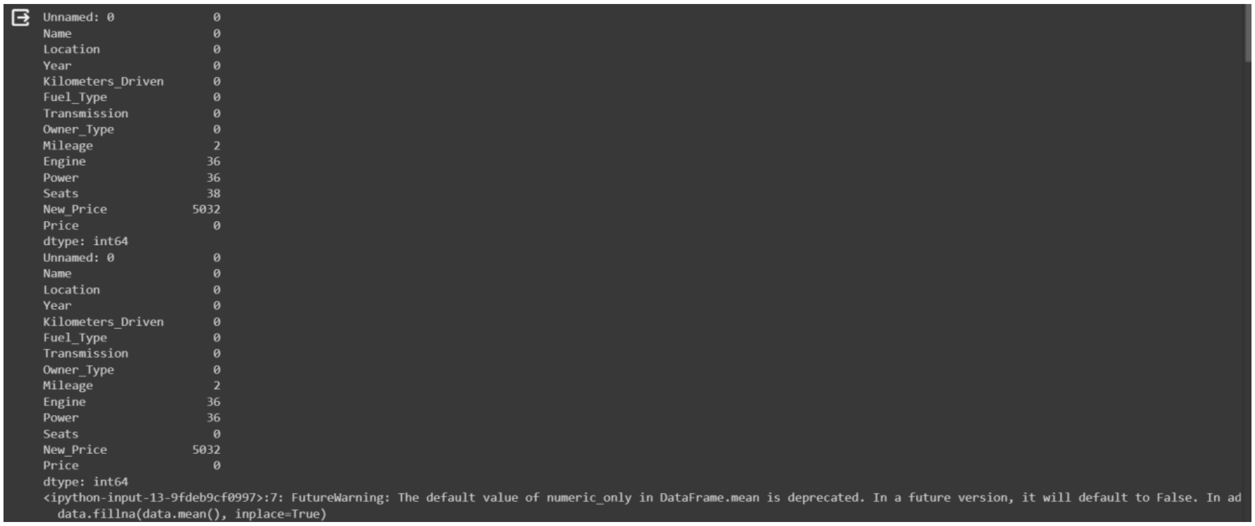
PRINCIPLES OF DATA SCIENCE (5530)-ASSIGNMENT 2

Name: Satya Sai Pramod Burgula
ID : 26342013

(a) This Python code reads a dataset from a CSV file into a Pandas DataFrame named 'data'. It then identifies missing values in each column, replaces missing values in numeric columns with the mean, and finally, displays the updated information. This imputation strategy using the mean ensures a balanced treatment of missing data, enhancing the dataset's completeness for subsequent analysis while maintaining the statistical integrity of numeric features.

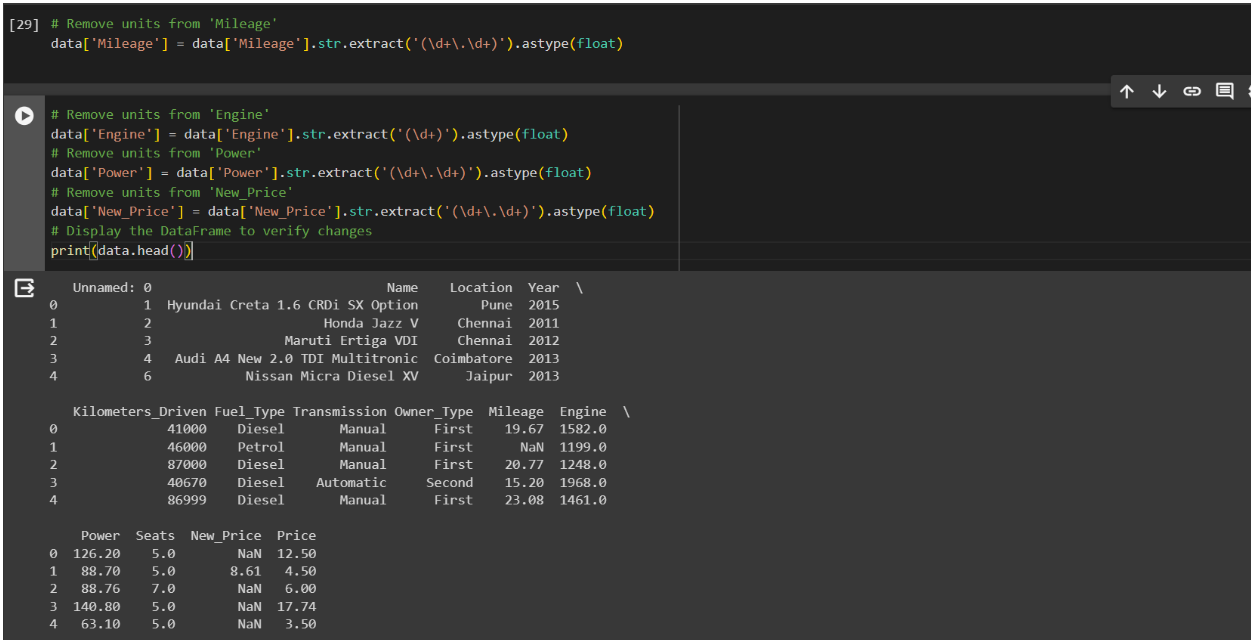


```
# Assuming your dataset is in a CSV file, you can read it into a Pandas DataFrame
import pandas as pd
data = pd.read_csv('train.csv')
# Display the information about missing values in each column
print(data.isnull().sum())
# Replace missing values in numeric columns with the mean
data.fillna(data.mean(), inplace=True)
# Display the information after handling missing values
print(data.isnull().sum())
```



```
Unnamed: 0              0
Name                    0
Location                0
Year                    0
Kilometers_Driven       0
Fuel_Type               0
Transmission            0
Owner_Type              0
Mileage                 2
Engine                 36
Power                  36
Seats                  38
New_Price            5032
Price                   0
dtype: int64
Unnamed: 0              0
Name                    0
Location                0
Year                    0
Kilometers_Driven       0
Fuel_Type               0
Transmission            0
Owner_Type              0
Mileage                 2
Engine                 36
Power                  36
Seats                   0
New_Price            5032
Price                   0
dtype: int64
<ipython-input-13-9fdeb9cf0997>:7: FutureWarning: The default value of numeric_only in DataFrame.mean is deprecated. In a future version, it will default to False. In ad
  data.fillna(data.mean(), inplace=True)
```

(b)

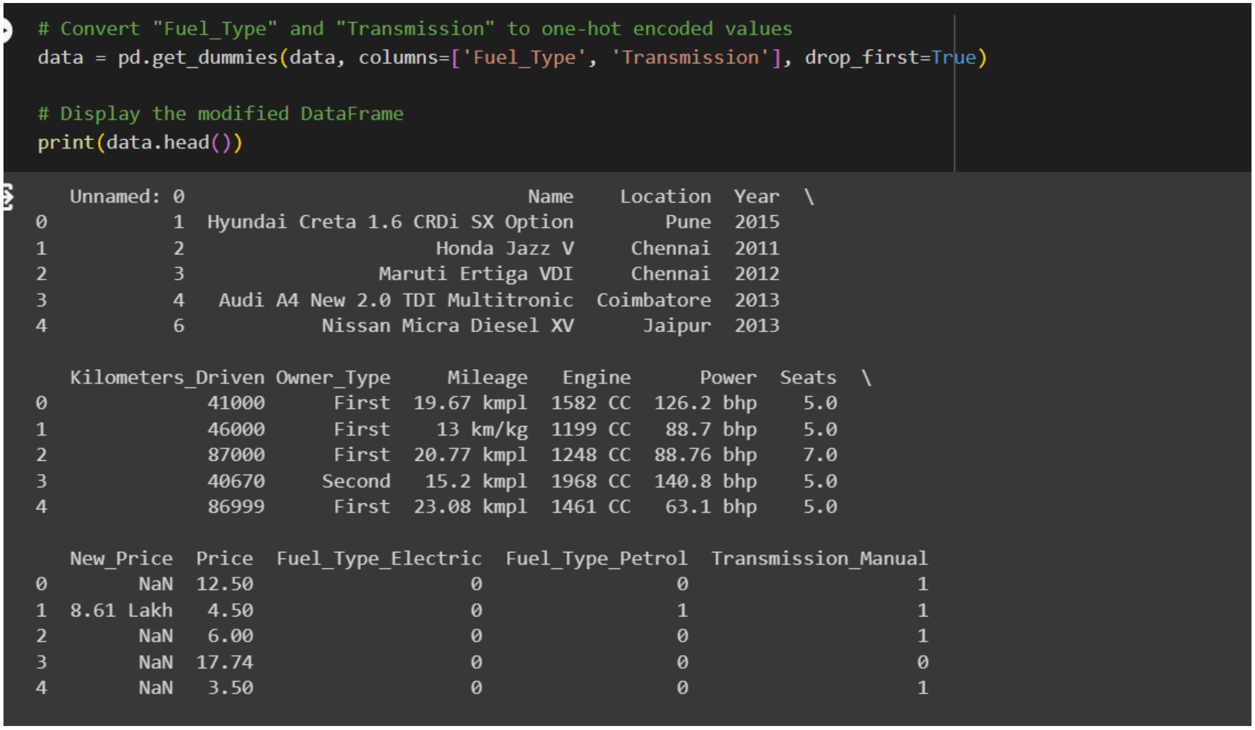


```
# Remove units from 'Mileage'
data['Mileage'] = data['Mileage'].str.extract('(\d+\.\d+)').astype(float)
```

```
# Remove units from 'Engine'
data['Engine'] = data['Engine'].str.extract('(\d+)').astype(float)
# Remove units from 'Power'
data['Power'] = data['Power'].str.extract('(\d+\.\d+)').astype(float)
# Remove units from 'New_Price'
data['New_Price'] = data['New_Price'].str.extract('(\d+\.\d+)').astype(float)
# Display the DataFrame to verify changes
print(data.head())
```

```
   Unnamed: 0                              Name    Location  Year  \
0           1  Hyundai Creta 1.6 CRDi SX Option        Pune  2015   
1           2                      Honda Jazz V     Chennai  2011   
2           3                 Maruti Ertiga VDI     Chennai  2012   
3           4   Audi A4 New 2.0 TDI Multitronic  Coimbatore  2013   
4           6            Nissan Micra Diesel XV      Jaipur  2013   

   Kilometers_Driven Fuel_Type Transmission Owner_Type  Mileage  Engine  \
0              41000    Diesel       Manual      First    19.67  1582.0   
1              46000    Petrol       Manual      First      NaN  1199.0   
2              87000    Diesel       Manual      First    20.77  1248.0   
3              40670    Diesel    Automatic     Second    15.20  1968.0   
4              86999    Diesel       Manual      First    23.08  1461.0   

    Power  Seats  New_Price  Price  
0  126.20    5.0        NaN  12.50  
1   88.70    5.0       8.61   4.50  
2   88.76    7.0        NaN   6.00  
3  140.80    5.0        NaN  17.74  
4   63.10    5.0        NaN   3.50  
```

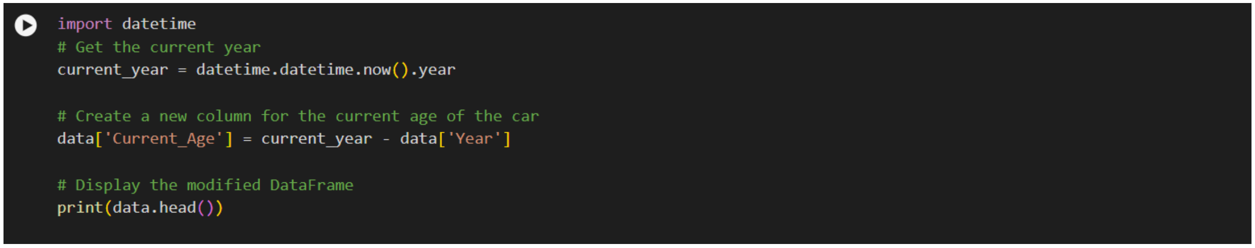
(c)



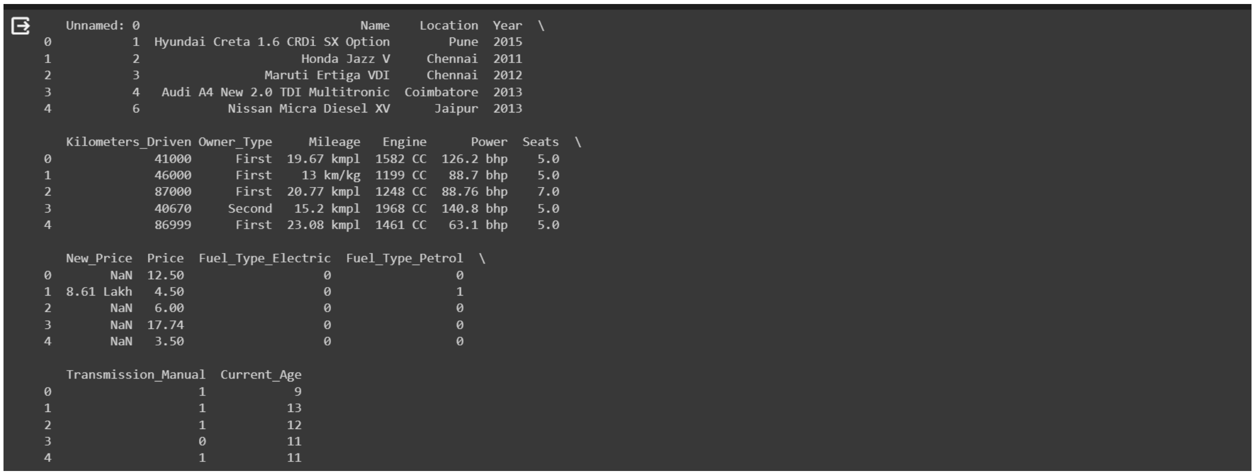
```
# Convert "Fuel_Type" and "Transmission" to one-hot encoded values
data = pd.get_dummies(data, columns=['Fuel_Type', 'Transmission'], drop_first=True)

# Display the modified DataFrame
print(data.head())
```

```
   Unnamed: 0                              Name    Location  Year  \
0           1  Hyundai Creta 1.6 CRDi SX Option        Pune  2015   
1           2                      Honda Jazz V     Chennai  2011   
2           3                 Maruti Ertiga VDI     Chennai  2012   
3           4   Audi A4 New 2.0 TDI Multitronic  Coimbatore  2013   
4           6            Nissan Micra Diesel XV      Jaipur  2013   

   Kilometers_Driven Owner_Type     Mileage   Engine      Power  Seats  \
0              41000      First  19.67 kmpl  1582 CC  126.2 bhp    5.0   
1              46000      First    13 km/kg  1199 CC   88.7 bhp    5.0   
2              87000      First  20.77 kmpl  1248 CC  88.76 bhp    7.0   
3              40670     Second   15.2 kmpl  1968 CC  140.8 bhp    5.0   
4              86999      First  23.08 kmpl  1461 CC   63.1 bhp    5.0   

   New_Price  Price  Fuel_Type_Electric  Fuel_Type_Petrol  Transmission_Manual  
0        NaN  12.50                   0                 0                    1  
1  8.61 Lakh   4.50                   0                 1                    1  
2        NaN   6.00                   0                 0                    1  
3        NaN  17.74                   0                 0                    0  
4        NaN   3.50                   0                 0                    1  
```

(d)

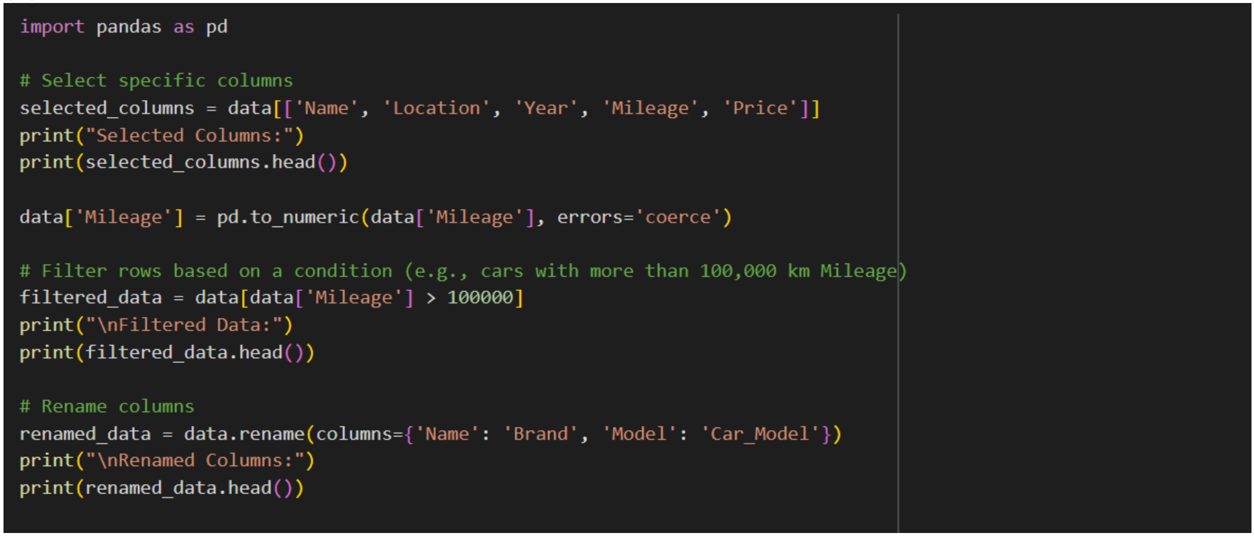


```
import datetime
# Get the current year
current_year = datetime.datetime.now().year

# Create a new column for the current age of the car
data['Current_Age'] = current_year - data['Year']

# Display the modified DataFrame
print(data.head())
```

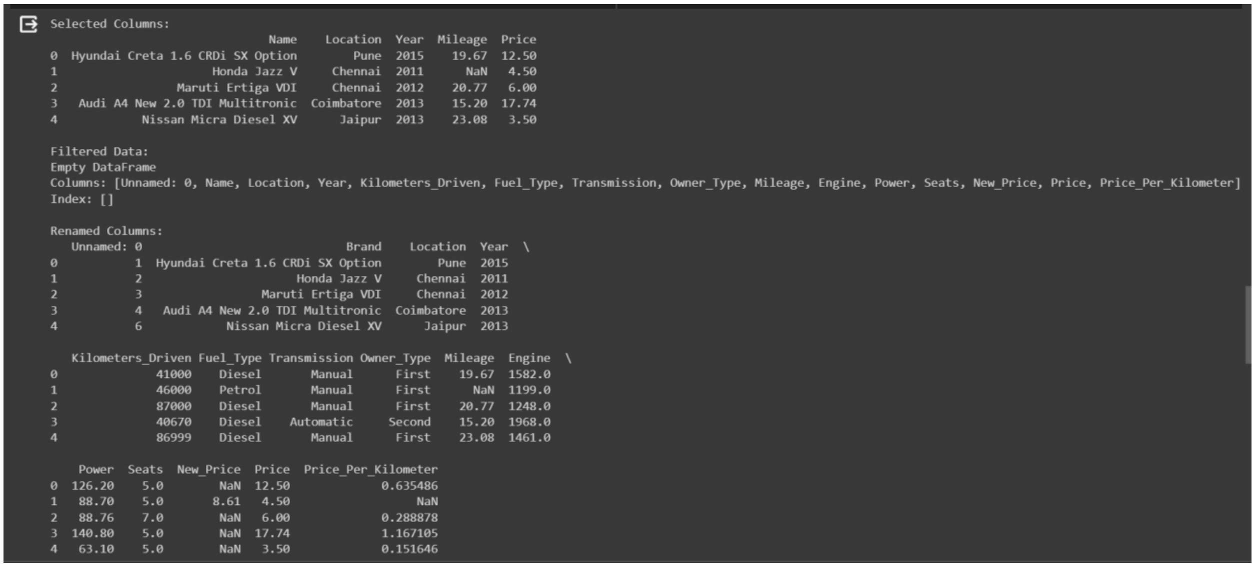


```
   Unnamed: 0                          Name    Location  Year  \
0           1  Hyundai Creta 1.6 CRDi SX Option        Pune  2015
1           2                      Honda Jazz V     Chennai  2011
2           3                 Maruti Ertiga VDI     Chennai  2012
3           4   Audi A4 New 2.0 TDI Multitronic  Coimbatore  2013
4           6            Nissan Micra Diesel XV      Jaipur  2013

   Kilometers_Driven Owner_Type     Mileage   Engine      Power  Seats  \
0              41000      First  19.67 kmpl  1582 CC  126.2 bhp    5.0
1              46000      First    13 km/kg  1199 CC   88.7 bhp    5.0
2              87000      First  20.77 kmpl  1248 CC  88.76 bhp    7.0
3              40670     Second   15.2 kmpl  1968 CC  140.8 bhp    5.0
4              86999      First  23.08 kmpl  1461 CC   63.1 bhp    5.0

   New_Price  Price  Fuel_Type_Electric  Fuel_Type_Petrol  \
0        NaN  12.50                   0                 0
1  8.61 Lakh   4.50                   0                 1
2        NaN   6.00                   0                 0
3        NaN  17.74                   0                 0
4        NaN   3.50                   0                 0

   Transmission_Manual  Current_Age
0                    1            9
1                    1           13
2                    1           12
3                    0           11
4                    1           11
```

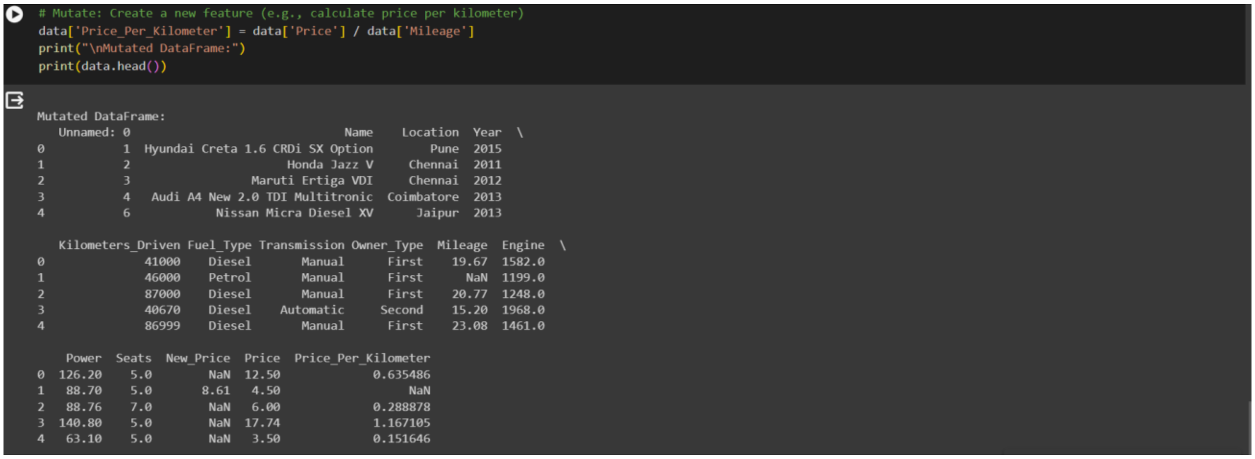
(e)



```
import pandas as pd

# Select specific columns
selected_columns = data[['Name', 'Location', 'Year', 'Mileage', 'Price']]
print("Selected Columns:")
print(selected_columns.head())

data['Mileage'] = pd.to_numeric(data['Mileage'], errors='coerce')

# Filter rows based on a condition (e.g., cars with more than 100,000 km Mileage)
filtered_data = data[data['Mileage'] > 100000]
print("\nFiltered Data:")
print(filtered_data.head())

# Rename columns
renamed_data = data.rename(columns={'Name': 'Brand', 'Model': 'Car_Model'})
print("\nRenamed Columns:")
print(renamed_data.head())
```

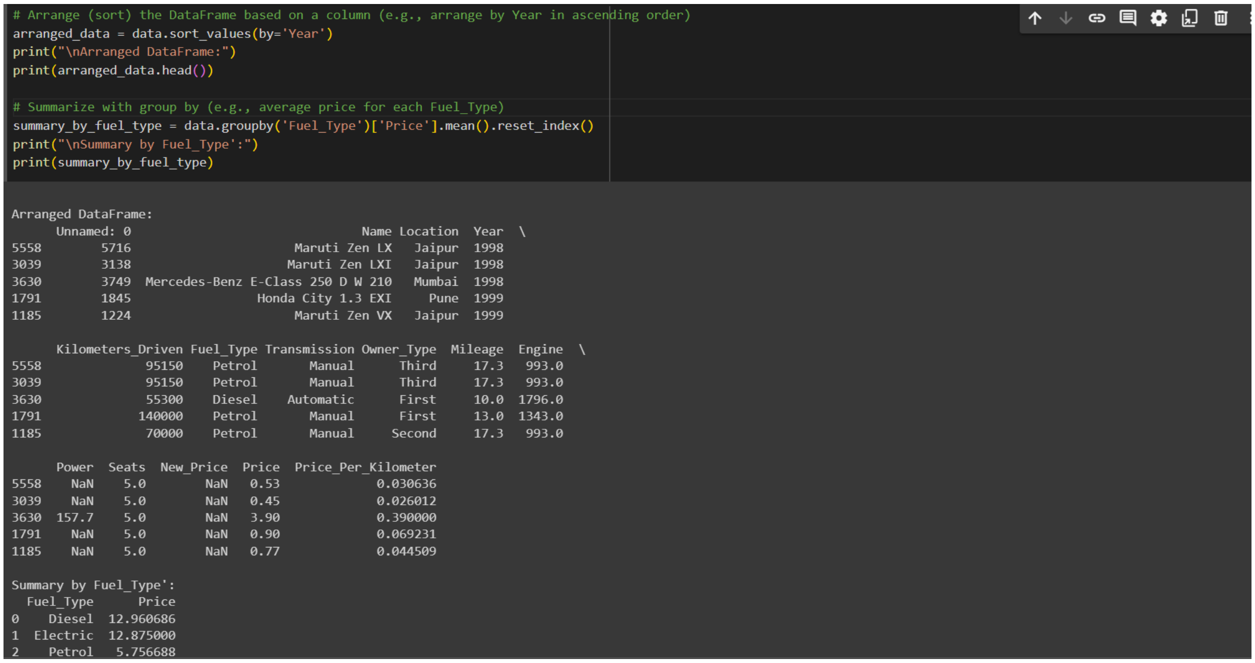


```
Selected Columns:
                              Name    Location  Year  Mileage  Price
0  Hyundai Creta 1.6 CRDi SX Option        Pune  2015    19.67  12.50
1                     Honda Jazz V     Chennai  2011      NaN   4.50
2                Maruti Ertiga VDI     Chennai  2012    20.77   6.00
3   Audi A4 New 2.0 TDI Multitronic  Coimbatore  2013    15.20  17.74
4           Nissan Micra Diesel XV      Jaipur  2013    23.08   3.50

Filtered Data:
Empty DataFrame
Columns: [Unnamed: 0, Name, Location, Year, Kilometers_Driven, Fuel_Type, Transmission, Owner_Type, Mileage, Engine, Power, Seats, New_Price, Price, Price_Per_Kilometer]
Index: []

Renamed Columns:
   Unnamed: 0                             Brand    Location  Year  \
0           1  Hyundai Creta 1.6 CRDi SX Option        Pune  2015
1           2                      Honda Jazz V     Chennai  2011
2           3                 Maruti Ertiga VDI     Chennai  2012
3           4   Audi A4 New 2.0 TDI Multitronic  Coimbatore  2013
4           6            Nissan Micra Diesel XV      Jaipur  2013

   Kilometers_Driven Fuel_Type Transmission Owner_Type  Mileage  Engine  \
0              41000    Diesel       Manual      First    19.67  1582.0
1              46000    Petrol       Manual      First      NaN  1199.0
2              87000    Diesel       Manual      First    20.77  1248.0
3              40670    Diesel    Automatic     Second    15.20  1968.0
4              86999    Diesel       Manual      First    23.08  1461.0

    Power  Seats  New_Price  Price  Price_Per_Kilometer
0  126.20    5.0        NaN  12.50             0.635486
1   88.70    5.0       8.61   4.50                  NaN
2   88.76    7.0        NaN   6.00             0.288878
3  140.80    5.0        NaN  17.74             1.167105
4   63.10    5.0        NaN   3.50             0.151646
```



```
# Mutate: Create a new feature (e.g., calculate price per kilometer)
data['Price_Per_Kilometer'] = data['Price'] / data['Mileage']
print("\nMutated DataFrame:")
print(data.head())
```

```
Mutated DataFrame:
   Unnamed: 0                             Name    Location  Year  \
0           1  Hyundai Creta 1.6 CRDi SX Option        Pune  2015
1           2                      Honda Jazz V     Chennai  2011
2           3                 Maruti Ertiga VDI     Chennai  2012
3           4   Audi A4 New 2.0 TDI Multitronic  Coimbatore  2013
4           6            Nissan Micra Diesel XV      Jaipur  2013

   Kilometers_Driven Fuel_Type Transmission Owner_Type  Mileage  Engine  \
0              41000    Diesel       Manual      First    19.67  1582.0
1              46000    Petrol       Manual      First      NaN  1199.0
2              87000    Diesel       Manual      First    20.77  1248.0
3              40670    Diesel    Automatic     Second    15.20  1968.0
4              86999    Diesel       Manual      First    23.08  1461.0

    Power  Seats  New_Price  Price  Price_Per_Kilometer
0  126.20    5.0        NaN  12.50             0.635486
1   88.70    5.0       8.61   4.50                  NaN
2   88.76    7.0        NaN   6.00             0.288878
3  140.80    5.0        NaN  17.74             1.167105
4   63.10    5.0        NaN   3.50             0.151646
```



```python
# Arrange (sort) the DataFrame based on a column (e.g., arrange by Year in ascending order)
arranged_data = data.sort_values(by='Year')
print("\nArranged DataFrame:")
print(arranged_data.head())

# Summarize with group by (e.g., average price for each Fuel_Type)
summary_by_fuel_type = data.groupby('Fuel_Type')['Price'].mean().reset_index()
print("\nSummary by Fuel_Type':")
print(summary_by_fuel_type)
```

```
Arranged DataFrame:
      Unnamed: 0                               Name Location  Year  \
5558        5716                      Maruti Zen LX   Jaipur  1998   
3039        3138                     Maruti Zen LXI   Jaipur  1998   
3630        3749  Mercedes-Benz E-Class 250 D W 210   Mumbai  1998   
1791        1845                 Honda City 1.3 EXI     Pune  1999   
1185        1224                      Maruti Zen VX   Jaipur  1999   

      Kilometers_Driven Fuel_Type Transmission Owner_Type  Mileage  Engine  \
5558              95150    Petrol       Manual      Third     17.3   993.0   
3039              95150    Petrol       Manual      Third     17.3   993.0   
3630              55300    Diesel    Automatic      First     10.0  1796.0   
1791             140000    Petrol       Manual      First     13.0  1343.0   
1185              70000    Petrol       Manual     Second     17.3   993.0   

      Power  Seats  New_Price  Price  Price_Per_Kilometer  
5558    NaN    5.0        NaN   0.53             0.030636  
3039    NaN    5.0        NaN   0.45             0.026012  
3630  157.7    5.0        NaN   3.90             0.390000  
1791    NaN    5.0        NaN   0.90             0.069231  
1185    NaN    5.0        NaN   0.77             0.044509  

Summary by Fuel_Type':
  Fuel_Type      Price
0    Diesel  12.960686
1  Electric  12.875000
2    Petrol   5.756688
```