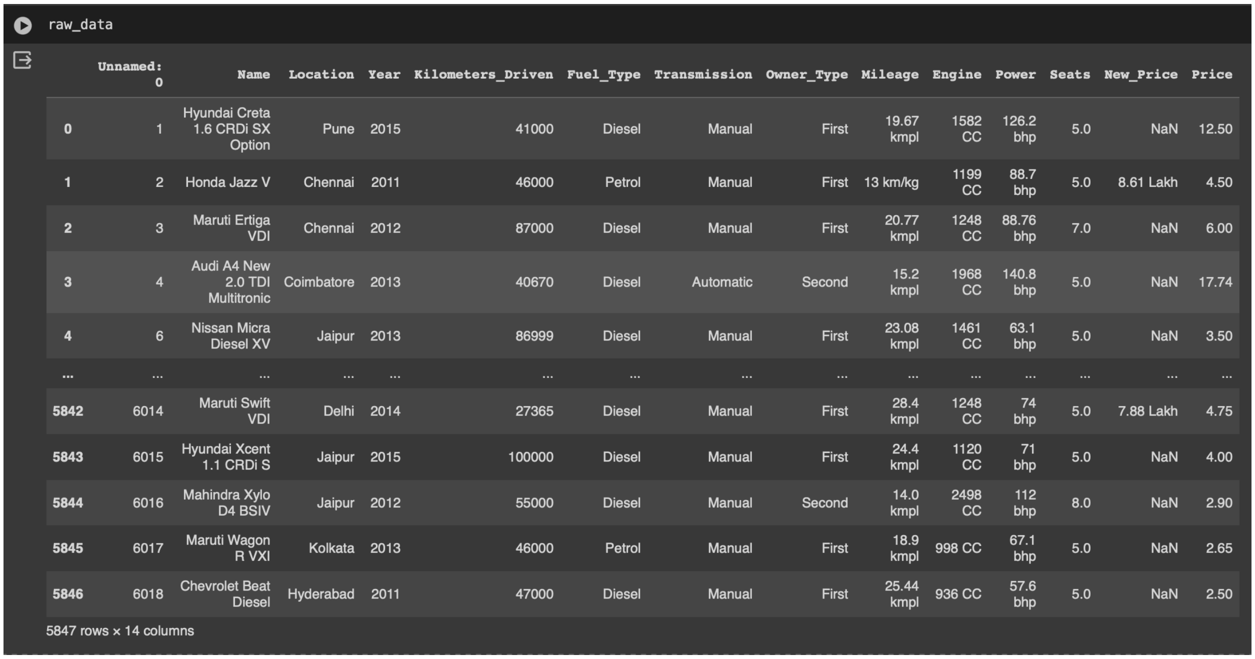
Assignment – 2

Name: Sai Pavan Pratapagiri
ID: 16343743

A. Look for the missing values in all the columns and either impute them (replace with mean, median, or mode) or drop them. Justify your action for this task.

- Load the dataset using pandas. The cars dataset contains 5847 observations and 14 variables.

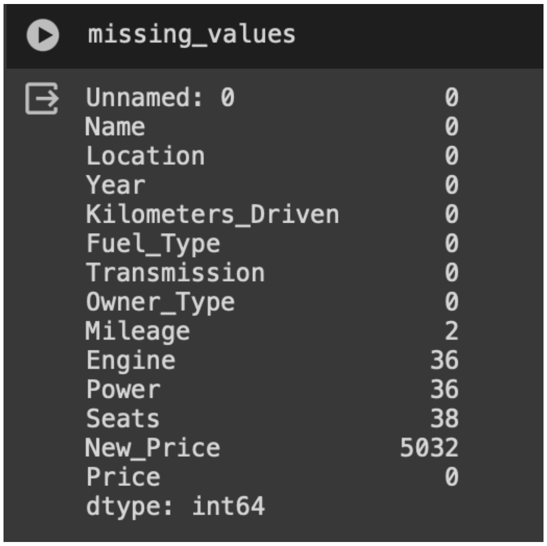


```
raw_data
```

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	NaN	12.50
1	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	13 km/kg	1199 CC	88.7 bhp	5.0	8.61 Lakh	4.50
2	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN	6.00
3	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	NaN	17.74
4	6	Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5.0	NaN	3.50
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5842	6014	Maruti Swift VDI	Delhi	2014	27365	Diesel	Manual	First	28.4 kmpl	1248 CC	74 bhp	5.0	7.88 Lakh	4.75
5843	6015	Hyundai Xcent 1.1 CRDi S	Jaipur	2015	100000	Diesel	Manual	First	24.4 kmpl	1120 CC	71 bhp	5.0	NaN	4.00
5844	6016	Mahindra Xylo D4 BSIV	Jaipur	2012	55000	Diesel	Manual	Second	14.0 kmpl	2498 CC	112 bhp	8.0	NaN	2.90
5845	6017	Maruti Wagon R VXI	Kolkata	2013	46000	Petrol	Manual	First	18.9 kmpl	998 CC	67.1 bhp	5.0	NaN	2.65
5846	6018	Chevrolet Beat Diesel	Hyderabad	2011	47000	Diesel	Manual	First	25.44 kmpl	936 CC	57.6 bhp	5.0	NaN	2.50

5847 rows × 14 columns

- Now we need to find the missing values in the dataset.

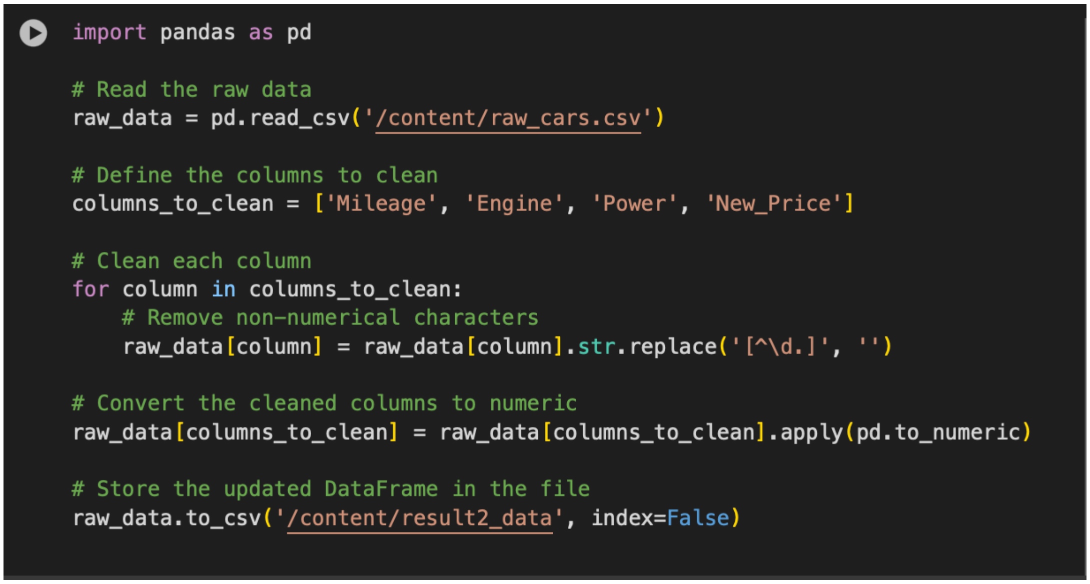


```
missing_values
```

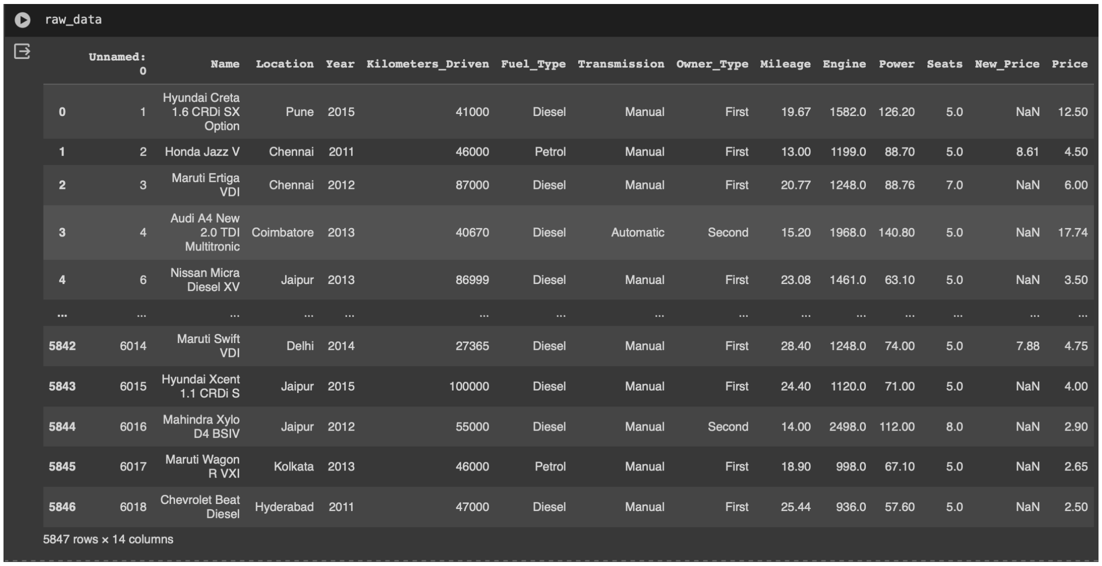
```
Unnamed: 0              0
Name                    0
Location                0
Year                    0
Kilometers_Driven       0
Fuel_Type               0
Transmission            0
Owner_Type              0
Mileage                 2
Engine                 36
Power                  36
Seats                  38
New_Price            5032
Price                   0
dtype: int64
```

- Before imputing the missing values, we need to perform a proper cleaning process. Mean operations can be performed only on numeric datatype.
- Dropping the first column and the New_Price column as they won't add any weightage to our analysis.
- As all the missing values are of numeric type, replacing with Mean values.
- The cleaned dataset is stored in a file in results folder a s "clean_data".

B. Remove the units from some of the attributes and only keep the numerical values (for example remove kmpl from "Mileage", CC from "Engine", bhp from "Power", and lakh from "New_price").



```
import pandas as pd

# Read the raw data
raw_data = pd.read_csv('/content/raw_cars.csv')

# Define the columns to clean
columns_to_clean = ['Mileage', 'Engine', 'Power', 'New_Price']

# Clean each column
for column in columns_to_clean:
    # Remove non-numerical characters
    raw_data[column] = raw_data[column].str.replace('[^\d.]', '')

# Convert the cleaned columns to numeric
raw_data[columns_to_clean] = raw_data[columns_to_clean].apply(pd.to_numeric)

# Store the updated DataFrame in the file
raw_data.to_csv('/content/result2_data', index=False)
```

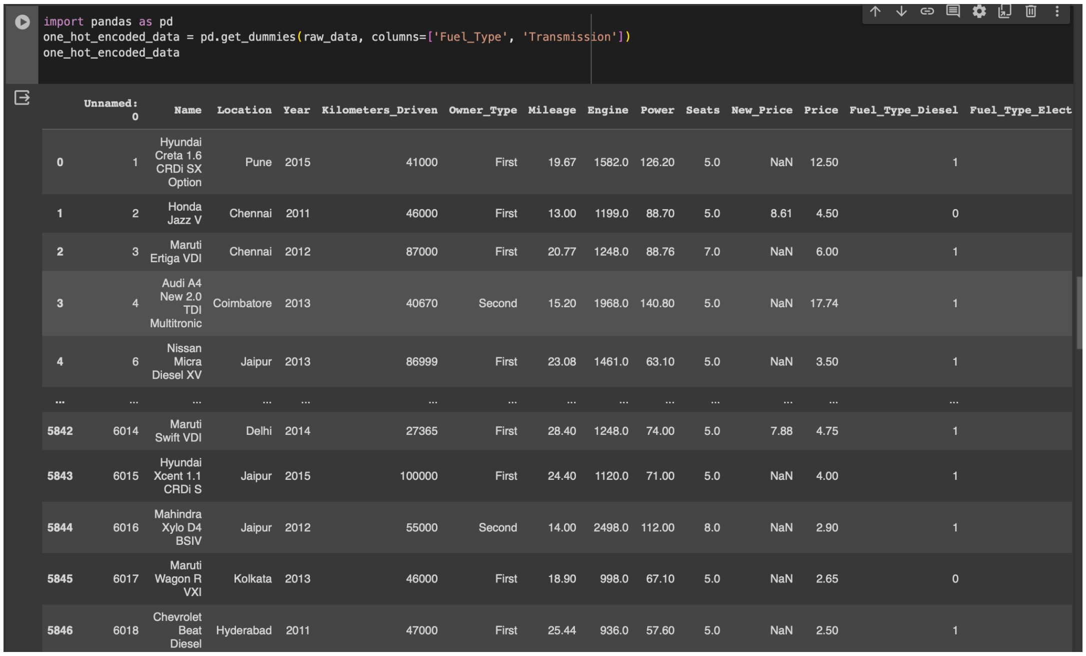


```
raw_data
```

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67	1582.0	126.20	5.0	NaN	12.50
1	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	13.00	1199.0	88.70	5.0	8.61	4.50
2	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77	1248.0	88.76	7.0	NaN	6.00
3	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.20	1968.0	140.80	5.0	NaN	17.74
4	6	Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.08	1461.0	63.10	5.0	NaN	3.50
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5842	6014	Maruti Swift VDI	Delhi	2014	27365	Diesel	Manual	First	28.40	1248.0	74.00	5.0	7.88	4.75
5843	6015	Hyundai Xcent 1.1 CRDi S	Jaipur	2015	100000	Diesel	Manual	First	24.40	1120.0	71.00	5.0	NaN	4.00
5844	6016	Mahindra Xylo D4 BSIV	Jaipur	2012	55000	Diesel	Manual	Second	14.00	2498.0	112.00	8.0	NaN	2.90
5845	6017	Maruti Wagon R VXI	Kolkata	2013	46000	Petrol	Manual	First	18.90	998.0	67.10	5.0	NaN	2.65
5846	6018	Chevrolet Beat Diesel	Hyderabad	2011	47000	Diesel	Manual	First	25.44	936.0	57.60	5.0	NaN	2.50

5847 rows × 14 columns

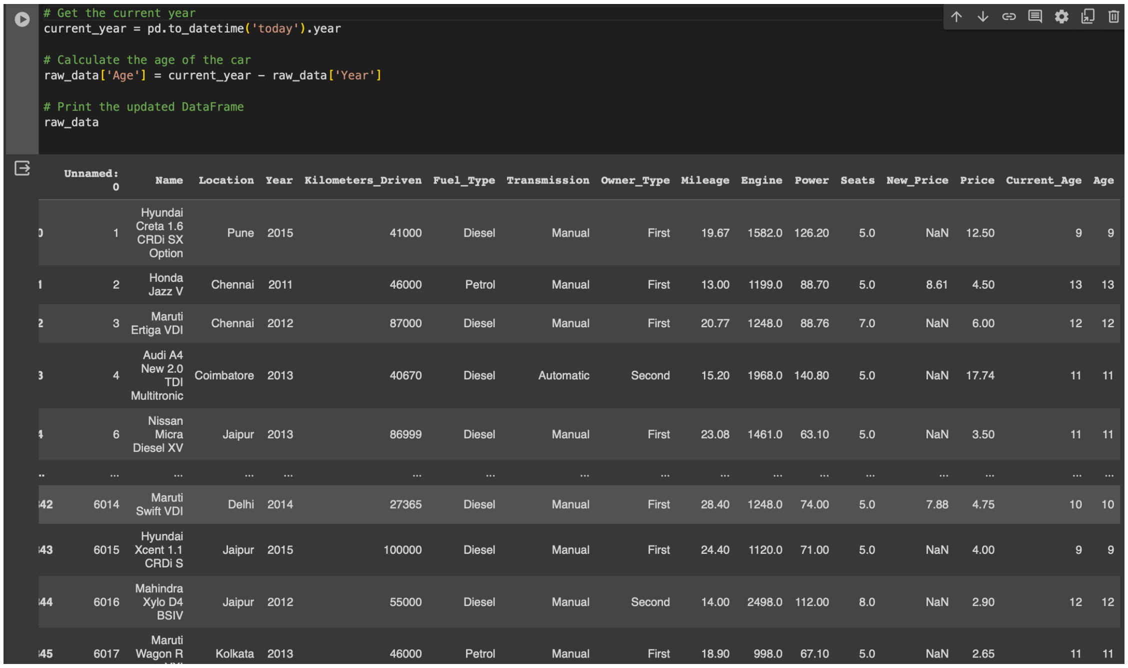
C. Change the categorical variables ("Fuel_Type" and "Transmission") into numerical one-hot encoded value.



```python
import pandas as pd
one_hot_encoded_data = pd.get_dummies(raw_data, columns=['Fuel_Type', 'Transmission'])
one_hot_encoded_data
```

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price	Fuel_Type_Diesel	Fuel_Type_Elect
0	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	First	19.67	1582.0	126.20	5.0	NaN	12.50	1	
1	2	Honda Jazz V	Chennai	2011	46000	First	13.00	1199.0	88.70	5.0	8.61	4.50	0	
2	3	Maruti Ertiga VDI	Chennai	2012	87000	First	20.77	1248.0	88.76	7.0	NaN	6.00	1	
3	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Second	15.20	1968.0	140.80	5.0	NaN	17.74	1	
4	6	Nissan Micra Diesel XV	Jaipur	2013	86999	First	23.08	1461.0	63.10	5.0	NaN	3.50	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5842	6014	Maruti Swift VDI	Delhi	2014	27365	First	28.40	1248.0	74.00	5.0	7.88	4.75	1	
5843	6015	Hyundai Xcent 1.1 CRDi S	Jaipur	2015	100000	First	24.40	1120.0	71.00	5.0	NaN	4.00	1	
5844	6016	Mahindra Xylo D4 BSIV	Jaipur	2012	55000	Second	14.00	2498.0	112.00	8.0	NaN	2.90	1	
5845	6017	Maruti Wagon R VXI	Kolkata	2013	46000	First	18.90	998.0	67.10	5.0	NaN	2.65	0	
5846	6018	Chevrolet Beat Diesel	Hyderabad	2011	47000	First	25.44	936.0	57.60	5.0	NaN	2.50	1	

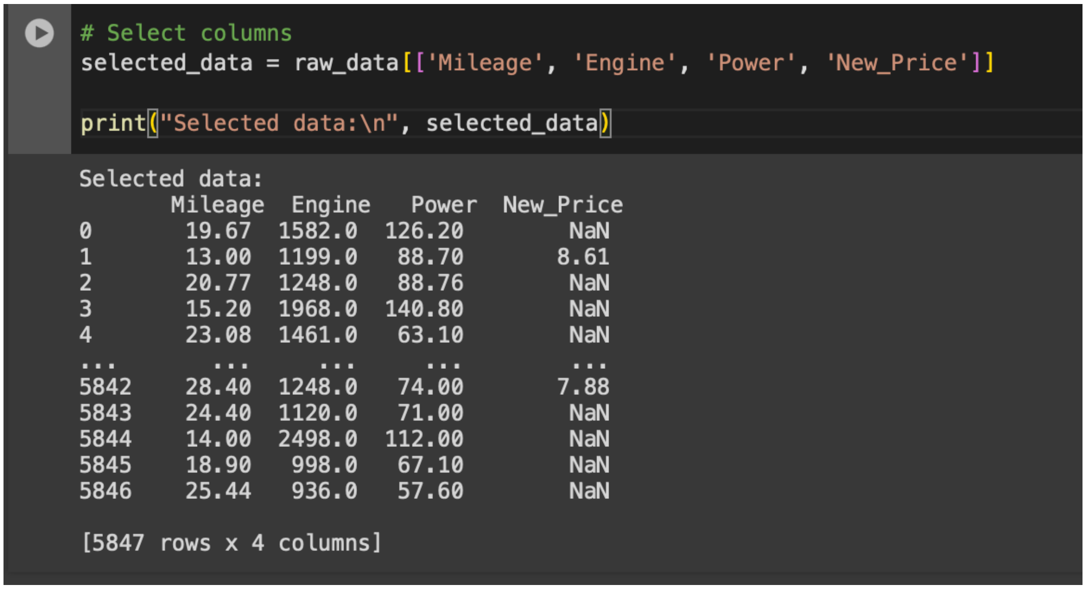
D. **Create one more feature and add this column to the dataset (you can use mutate function in R for this). For example, you can calculate the current age of the car by subtracting "Year" value from the current year.**



```python
# Get the current year
current_year = pd.to_datetime('today').year

# Calculate the age of the car
raw_data['Age'] = current_year - raw_data['Year']

# Print the updated DataFrame
raw_data
```

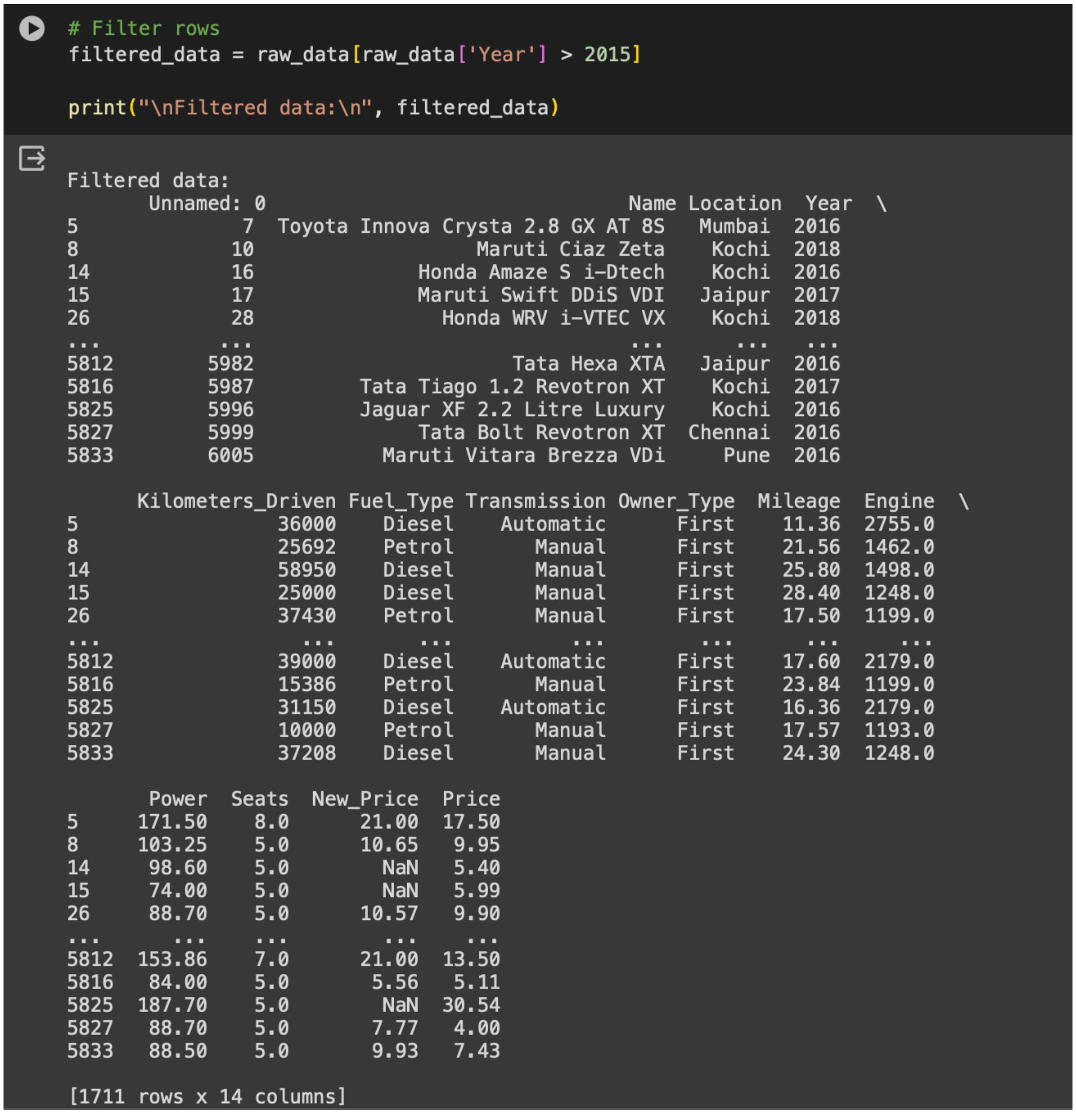
	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price	Current_Age	Age
0	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67	1582.0	126.20	5.0	NaN	12.50	9	9
1	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	13.00	1199.0	88.70	5.0	8.61	4.50	13	13
2	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77	1248.0	88.76	7.0	NaN	6.00	12	12
3	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.20	1968.0	140.80	5.0	NaN	17.74	11	11
4	6	Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.08	1461.0	63.10	5.0	NaN	3.50	11	11
..	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
42	6014	Maruti Swift VDI	Delhi	2014	27365	Diesel	Manual	First	28.40	1248.0	74.00	5.0	7.88	4.75	10	10
43	6015	Hyundai Xcent 1.1 CRDi S	Jaipur	2015	100000	Diesel	Manual	First	24.40	1120.0	71.00	5.0	NaN	4.00	9	9
44	6016	Mahindra Xylo D4 BSIV	Jaipur	2012	55000	Diesel	Manual	Second	14.00	2498.0	112.00	8.0	NaN	2.90	12	12
45	6017	Maruti Wagon R	Kolkata	2013	46000	Petrol	Manual	First	18.90	998.0	67.10	5.0	NaN	2.65	11	11

E. **Perform select, filter, rename, mutate, arrange, and summarize with group by operations (or their equivalent operations in python) on this dataset.**



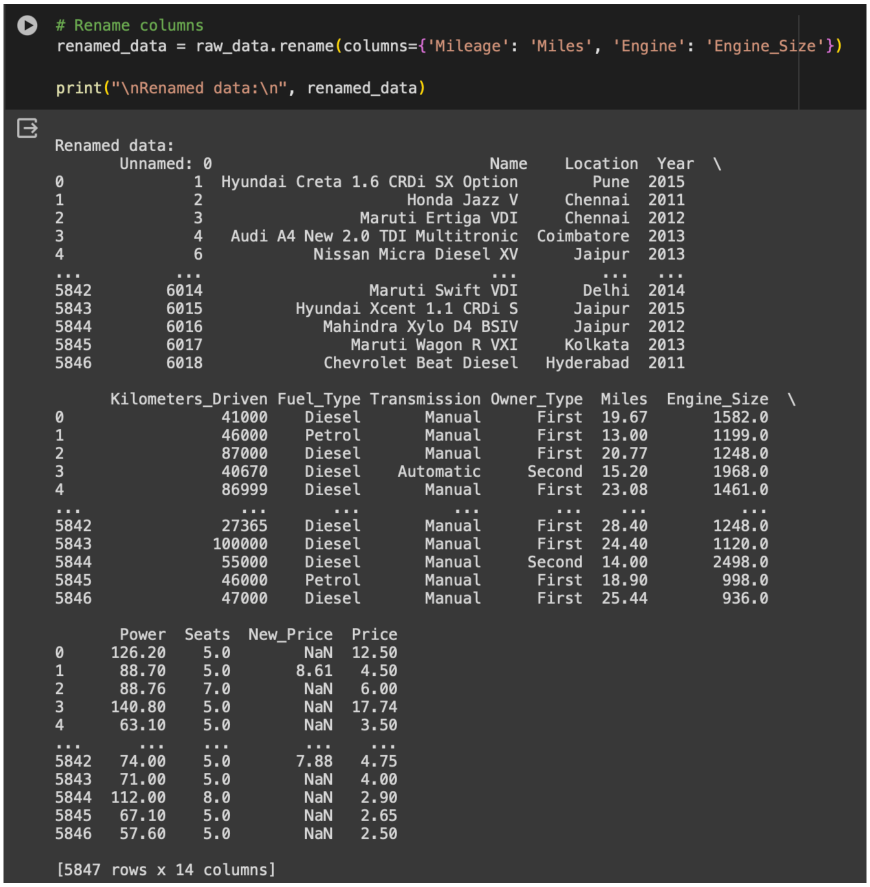
```python
# Select columns
selected_data = raw_data[['Mileage', 'Engine', 'Power', 'New_Price']]

print("Selected data:\n", selected_data)
```

```
Selected data:
       Mileage  Engine   Power  New_Price
0       19.67  1582.0  126.20        NaN
1       13.00  1199.0   88.70       8.61
2       20.77  1248.0   88.76        NaN
3       15.20  1968.0  140.80        NaN
4       23.08  1461.0   63.10        NaN
...       ...     ...     ...        ...
5842    28.40  1248.0   74.00       7.88
5843    24.40  1120.0   71.00        NaN
5844    14.00  2498.0  112.00        NaN
5845    18.90   998.0   67.10        NaN
5846    25.44   936.0   57.60        NaN

[5847 rows x 4 columns]
```



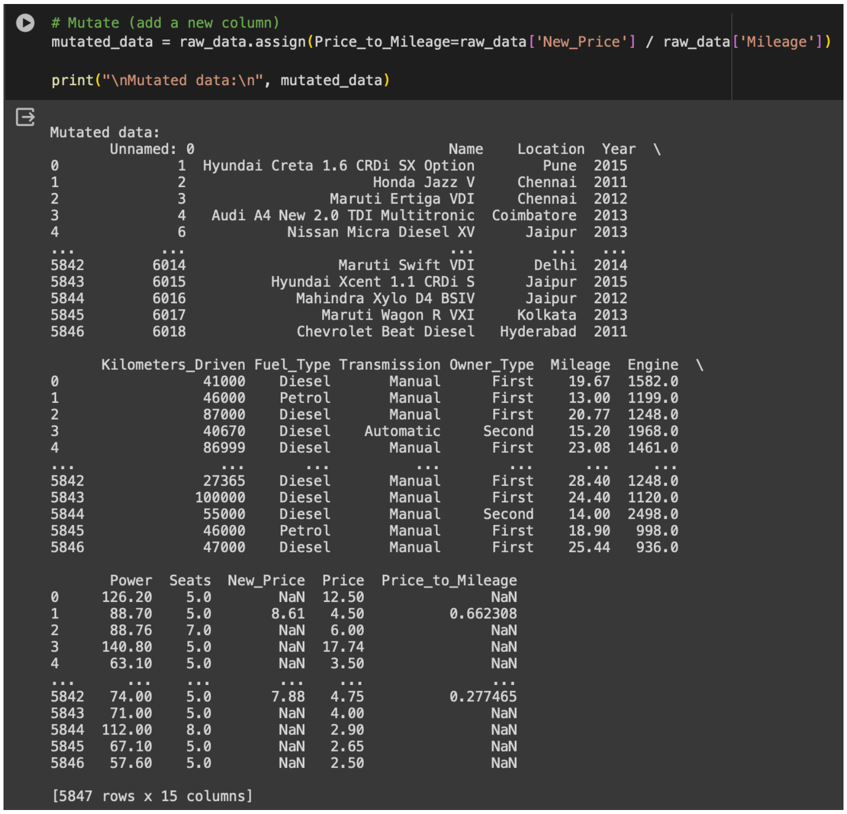
```python
# Filter rows
filtered_data = raw_data[raw_data['Year'] > 2015]

print("\nFiltered data:\n", filtered_data)
```

```
Filtered data:
       Unnamed: 0                               Name Location  Year  \
5              7  Toyota Innova Crysta 2.8 GX AT 8S   Mumbai  2016   
8             10                   Maruti Ciaz Zeta    Kochi  2018   
14            16              Honda Amaze S i-Dtech    Kochi  2016   
15            17             Maruti Swift DDiS VDI   Jaipur  2017   
26            28                Honda WRV i-VTEC VX    Kochi  2018   
...          ...                                ...      ...   ...   
5812        5982                     Tata Hexa XTA   Jaipur  2016   
5816        5987         Tata Tiago 1.2 Revotron XT    Kochi  2017   
5825        5996        Jaguar XF 2.2 Litre Luxury    Kochi  2016   
5827        5999           Tata Bolt Revotron XT  Chennai  2016   
5833        6005          Maruti Vitara Brezza VDi     Pune  2016   

      Kilometers_Driven Fuel_Type Transmission Owner_Type  Mileage  Engine  \
5                 36000    Diesel    Automatic      First    11.36  2755.0   
8                 25692    Petrol       Manual      First    21.56  1462.0   
14                58950    Diesel       Manual      First    25.80  1498.0   
15                25000    Diesel       Manual      First    28.40  1248.0   
26                37430    Petrol       Manual      First    17.50  1199.0   
...                 ...       ...          ...        ...      ...     ...   
5812              39000    Diesel    Automatic      First    17.60  2179.0   
5816              15386    Petrol       Manual      First    23.84  1199.0   
5825              31150    Diesel    Automatic      First    16.36  2179.0   
5827              10000    Petrol       Manual      First    17.57  1193.0   
5833              37208    Diesel       Manual      First    24.30  1248.0   

       Power  Seats  New_Price  Price  
5     171.50    8.0      21.00  17.50  
8     103.25    5.0      10.65   9.95  
14     98.60    5.0        NaN   5.40  
15     74.00    5.0        NaN   5.99  
26     88.70    5.0      10.57   9.90  
...      ...    ...        ...    ...  
5812  153.86    7.0      21.00  13.50  
5816   84.00    5.0       5.56   5.11  
5825  187.70    5.0        NaN  30.54  
5827   88.70    5.0       7.77   4.00  
5833   88.50    5.0       9.93   7.43  

[1711 rows x 14 columns]
```



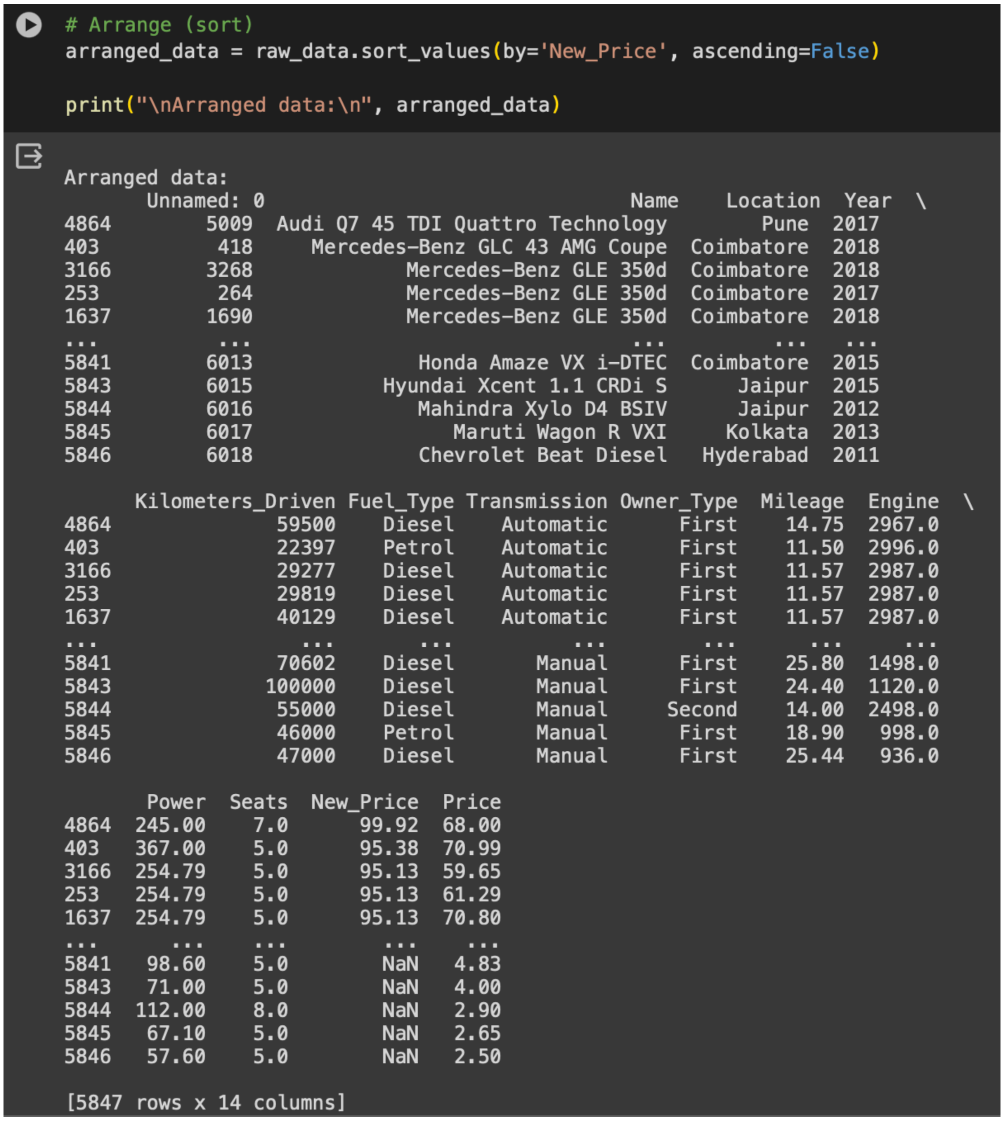
```
# Rename columns
renamed_data = raw_data.rename(columns={'Mileage': 'Miles', 'Engine': 'Engine_Size'})

print("\nRenamed data:\n", renamed_data)
```

```
Renamed data:
        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   
...          ...                               ...         ...   ...   
5842        6014                  Maruti Swift VDI       Delhi  2014   
5843        6015          Hyundai Xcent 1.1 CRDi S      Jaipur  2015   
5844        6016             Mahindra Xylo D4 BSIV      Jaipur  2012   
5845        6017                Maruti Wagon R VXI     Kolkata  2013   
5846        6018             Chevrolet Beat Diesel   Hyderabad  2011   

      Kilometers_Driven Fuel_Type Transmission Owner_Type  Miles  Engine_Size  \
0                 41000    Diesel       Manual      First  19.67       1582.0   
1                 46000    Petrol       Manual      First  13.00       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   
...                 ...       ...          ...        ...    ...          ...   
5842              27365    Diesel       Manual      First  28.40       1248.0   
5843             100000    Diesel       Manual      First  24.40       1120.0   
5844              55000    Diesel       Manual     Second  14.00       2498.0   
5845              46000    Petrol       Manual      First  18.90        998.0   
5846              47000    Diesel       Manual      First  25.44        936.0   

       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  
...      ...    ...        ...    ...  
5842   74.00    5.0       7.88   4.75  
5843   71.00    5.0        NaN   4.00  
5844  112.00    8.0        NaN   2.90  
5845   67.10    5.0        NaN   2.65  
5846   57.60    5.0        NaN   2.50  

[5847 rows x 14 columns]
```



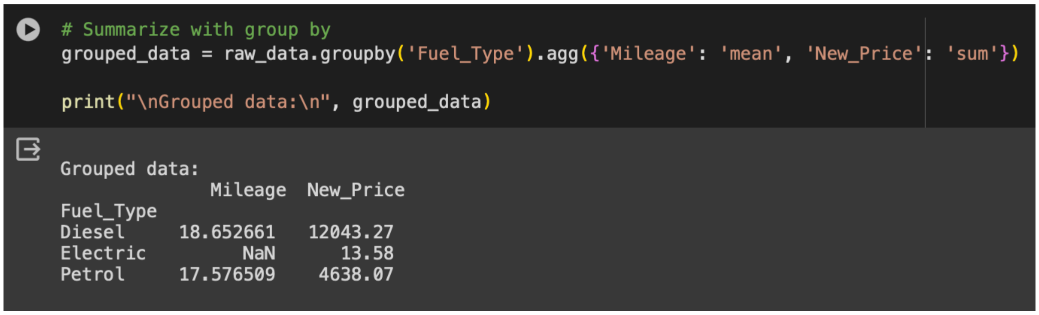
```
# Mutate (add a new column)
mutated_data = raw_data.assign(Price_to_Mileage=raw_data['New_Price'] / raw_data['Mileage'])

print("\nMutated data:\n", mutated_data)
```

```
Mutated data:
        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   
...          ...                               ...         ...   ...   
5842        6014                  Maruti Swift VDI       Delhi  2014   
5843        6015          Hyundai Xcent 1.1 CRDi S      Jaipur  2015   
5844        6016             Mahindra Xylo D4 BSIV      Jaipur  2012   
5845        6017                Maruti Wagon R VXI     Kolkata  2013   
5846        6018             Chevrolet Beat Diesel   Hyderabad  2011   

      Kilometers_Driven Fuel_Type Transmission Owner_Type  Mileage  Engine  \
0                 41000    Diesel       Manual      First    19.67  1582.0   
1                 46000    Petrol       Manual      First    13.00  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   
...                 ...       ...          ...        ...      ...     ...   
5842              27365    Diesel       Manual      First    28.40  1248.0   
5843             100000    Diesel       Manual      First    24.40  1120.0   
5844              55000    Diesel       Manual     Second    14.00  2498.0   
5845              46000    Petrol       Manual      First    18.90   998.0   
5846              47000    Diesel       Manual      First    25.44   936.0   

       Power  Seats  New_Price  Price  Price_to_Mileage  
0     126.20    5.0        NaN  12.50               NaN  
1      88.70    5.0       8.61   4.50          0.662308  
2      88.76    7.0        NaN   6.00               NaN  
3     140.80    5.0        NaN  17.74               NaN  
4      63.10    5.0        NaN   3.50               NaN  
...      ...    ...        ...    ...               ...  
5842   74.00    5.0       7.88   4.75          0.277465  
5843   71.00    5.0        NaN   4.00               NaN  
5844  112.00    8.0        NaN   2.90               NaN  
5845   67.10    5.0        NaN   2.65               NaN  
5846   57.60    5.0        NaN   2.50               NaN  

[5847 rows x 15 columns]
```



```python
# Arrange (sort)
arranged_data = raw_data.sort_values(by='New_Price', ascending=False)

print("\nArranged data:\n", arranged_data)
```

```
Arranged data:
        Unnamed: 0                               Name    Location  Year  \
4864        5009  Audi Q7 45 TDI Quattro Technology        Pune  2017   
403          418     Mercedes-Benz GLC 43 AMG Coupe  Coimbatore  2018   
3166        3268             Mercedes-Benz GLE 350d  Coimbatore  2018   
253          264             Mercedes-Benz GLE 350d  Coimbatore  2017   
1637        1690             Mercedes-Benz GLE 350d  Coimbatore  2018   
...          ...                                ...         ...   ...   
5841        6013              Honda Amaze VX i-DTEC  Coimbatore  2015   
5843        6015           Hyundai Xcent 1.1 CRDi S      Jaipur  2015   
5844        6016              Mahindra Xylo D4 BSIV      Jaipur  2012   
5845        6017                 Maruti Wagon R VXI     Kolkata  2013   
5846        6018              Chevrolet Beat Diesel   Hyderabad  2011   

      Kilometers_Driven Fuel_Type Transmission Owner_Type  Mileage  Engine  \
4864              59500    Diesel    Automatic      First    14.75  2967.0   
403               22397    Petrol    Automatic      First    11.50  2996.0   
3166              29277    Diesel    Automatic      First    11.57  2987.0   
253               29819    Diesel    Automatic      First    11.57  2987.0   
1637              40129    Diesel    Automatic      First    11.57  2987.0   
...                 ...       ...          ...        ...      ...     ...   
5841              70602    Diesel       Manual      First    25.80  1498.0   
5843             100000    Diesel       Manual      First    24.40  1120.0   
5844              55000    Diesel       Manual     Second    14.00  2498.0   
5845              46000    Petrol       Manual      First    18.90   998.0   
5846              47000    Diesel       Manual      First    25.44   936.0   

       Power  Seats  New_Price  Price  
4864  245.00    7.0      99.92  68.00  
403   367.00    5.0      95.38  70.99  
3166  254.79    5.0      95.13  59.65  
253   254.79    5.0      95.13  61.29  
1637  254.79    5.0      95.13  70.80  
...      ...    ...        ...    ...  
5841   98.60    5.0        NaN   4.83  
5843   71.00    5.0        NaN   4.00  
5844  112.00    8.0        NaN   2.90  
5845   67.10    5.0        NaN   2.65  
5846   57.60    5.0        NaN   2.50  

[5847 rows x 14 columns]
```



```python
# Summarize with group by
grouped_data = raw_data.groupby('Fuel_Type').agg({'Mileage': 'mean', 'New_Price': 'sum'})

print("\nGrouped data:\n", grouped_data)
```

```
Grouped data:
              Mileage  New_Price
Fuel_Type                    
Diesel     18.652661   12043.27
Electric         NaN      13.58
Petrol     17.576509    4638.07
```