

PDS- Assignment 2

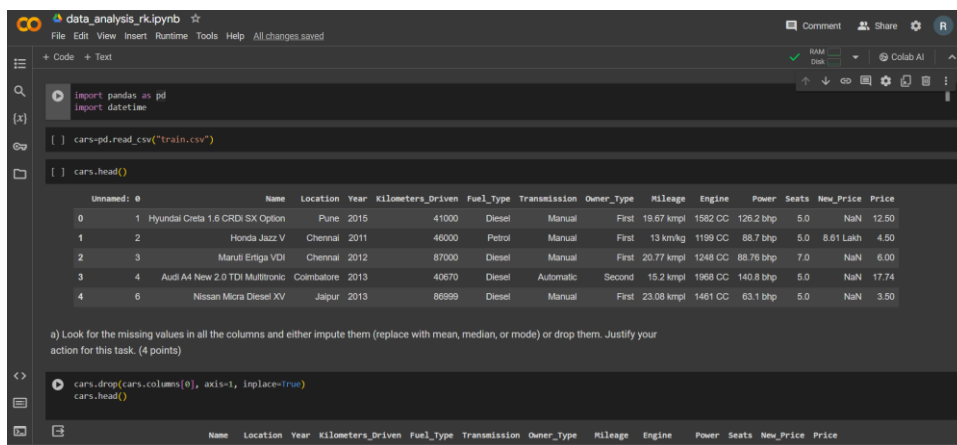
Name: SiramNagaSaiRadhakrishna
Student id: 16356525

- Handling Missing Values:**

Identified and addressed missing values in all columns.

Imputed missing values for categorical columns with the mode to preserve data integrity.

Imputed missing values for numerical columns with the mean to maintain overall data distribution.



```
import pandas as pd
import datetime

cars=pd.read_csv("train.csv")

cars.head()
```

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

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. (4 points)

```
cars.drop(cars.columns[0], axis=1, inplace=True)
cars.head()
```

Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
------	----------	------	-------------------	-----------	--------------	------------	---------	--------	-------	-------	-----------	-------

```
cars.drop(cars.columns[0], axis=1, inplace=True)
cars.head()
```

Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price	
0	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	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	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN	6.00
3	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	Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5.0	NaN	3.50

```
[ ] missing = cars.isna().sum()
for column in missing.index:
    if missing[column] > 0:
        if cars[column].dtype == 'object':
            # Replace missing values with the mode (most common value), for categorical columns
            mode_value = cars[column].mode()[0]
            cars[column].fillna(mode_value, inplace=True)
        else:
            # Replace missing values with mean, for numerical columns
            mean_value = cars[column].mean()
            cars[column].fillna(mean_value, inplace=True)
cars.head()
```

- **Removing Units:**

Stripped units from specified attributes, such as removing "kmpl" from "Mileage," "CC" from "Engine," "bhp" from "Power," and "lakh" from "New_price." Retained only the numerical values for a cleaner dataset.

```
[ ] # Remove units from attributes
cars['Mileage'] = cars['Mileage'].str.replace('kmpl', '').str.replace('km/kg', '').astype(float)
cars['Engine'] = cars['Engine'].str.replace('CC', '').astype(float)
cars['Power'] = cars['Power'].str.replace('bhp', '').astype(float)
cars['New_Price'] = cars['New_Price'].str.replace('lakh', '').str.replace('cr', '').astype(float)
```

cars.head()

	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	Hyundai Creta 1.6 CRDI SX Option	Pune	2015	41000	Diesel	Manual	First	19.67	1582.0	126.20	5.0	4.78	12.50
1	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	13.00	1199.0	88.70	5.0	8.61	4.50
2	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77	1248.0	88.76	7.0	4.78	6.00

- **Categorical to Numerical Transformation:**

Converted categorical variables, specifically "Fuel_Type" and "Transmission," into numerical one-hot encoded values.

Facilitated further analysis by representing categorical information in a numerical format.

```
[ ] cars= pd.get_dummies(cars, columns=['fuel_type', 'transmission'])
```

cars.head()

	Name	Location	Year	Kilometers_Driven	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price	Fuel_Type_Diesel	Fuel_Type_Electric	Fuel_Type_Petrol	Transmission_Autom
0	Hyundai Creta 1.6 CRDI SX Option	Pune	2015	41000	First	19.67	1582.0	126.20	5.0	4.78	12.50	1	0	0	
1	Honda Jazz V	Chennai	2011	46000	First	13.00	1199.0	88.70	5.0	8.61	4.50	0	0	1	
2	Maruti Ertiga VDI	Chennai	2012	87000	First	20.77	1248.0	88.76	7.0	4.78	6.00	1	0	0	
3	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Second	15.20	1968.0	140.80	5.0	4.78	17.74	1	0	0	

- **Creating Additional Feature:**

Created a new feature using the mutate function in R, calculating the current age of each car by subtracting the "Year" value from the current year.

Added the newly calculated current age column to the dataset.

```
[ ] current_year = datetime.datetime.now().year
cars['current_Age'] = current_year - cars['Year']
```

```
cars.head()
```

	Name	Location	Year	Kilometers_Driven	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price	Fuel_Type_Diesel	Fuel_Type_Electric	Fuel_Type_Petrol	Trans
0	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	First	19.67	1582.0	126.20	5.0	4.78	12.50	1	0	0	
1	Honda Jazz V	Chennai	2011	46000	First	13.00	1199.0	88.70	5.0	8.61	4.50	0	0	1	
2	Maruti Ertiga VDI	Chennai	2012	87000	First	20.77	1248.0	88.76	7.0	4.78	6.00	1	0	0	
3	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Second	15.20	1968.0	140.80	5.0	4.78	17.74	1	0	0	
4	Nissan Micra														

• Data Manipulation Operations:

Utilized various data manipulation operations like select, filter, rename, mutate, arrange, and summarize with group by.

Selected specific columns for analysis, filtered data based on specific conditions, renamed columns for clarity, created new features through mutation, arranged rows based on certain criteria, and performed summarization with group by to extract meaningful insights.

```
selection = cars[['Name', 'Year', 'Price']]#select
print(selection)
```

	Name	Year	Price
0	Hyundai Creta 1.6 CRDi SX Option	2015	12.50
1	Honda Jazz V	2011	4.50
2	Maruti Ertiga VDI	2012	6.00
3	Audi A4 New 2.0 TDI Multitronic	2013	17.74
4	Nissan Micra Diesel XV	2013	3.50
...
5842	Maruti Swift VDI	2014	4.75
5843	Hyundai Xcent 1.1 CRDi S	2015	4.00
5844	Mahindra Xylo D4 BSIV	2012	2.90
5845	Maruti Wagon R VXI	2013	2.65
5846	Chevrolet Beat Diesel	2011	2.50

[5847 rows x 3 columns]

```
[ ] filters = cars[cars['Location'] == 'Chennai']#filtered Data
print(filters)
```

	Name	Location	Year	\
1	Honda Jazz V	Chennai	2011	
2	Maruti Ertiga VDI	Chennai	2012	
7	Tata Indica Vista Quadrajet LS	Chennai	2012	
37	Volkswagen Polo Diesel Trendline 1.2L	Chennai	2013	
52	Hyundai Grand i10 Sportz	Chennai	2015	
...	
5762	Maruti Swift Dzire VXI Optional	Chennai	2015	
5764	Maruti Alto 800 2016-2019 LXI	Chennai	2016	
5781	Toyota Fortuner 4x4 AT	Chennai	2015	

```
cars_rename = cars.rename(columns={'New_Price': 'NewPriceofCAR'})#rename
cars_rename
```

	Name	Location	Year	Kilometers_Driven	Owner_Type	Mileage	Engine	Power	Seats	NewPriceofCAR	Price	Fuel_Type_Diesel	Fuel_Type_Electric	Fuel_Type_Petrol	Trans
0	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	First	19.67	1582.0	126.20	5.0	4.78	12.50	1	0	0	
1	Honda Jazz V	Chennai	2011	46000	First	13.00	1199.0	88.70	5.0	8.61	4.50	0	0	1	
2	Maruti Ertiga VDI	Chennai	2012	87000	First	20.77	1248.0	88.76	7.0	4.78	6.00	1	0	0	
3	Audi A4 New 2.0 TDI	Coimbatore	2013	40670	Second	15.20	1968.0	140.80	5.0	4.78	17.74	1	0	0	