DATA 240-21, Fall 2024

Assignment #1

Release on Sept 11th, 2024
Due 11:59pm on Sept 24th, 2024

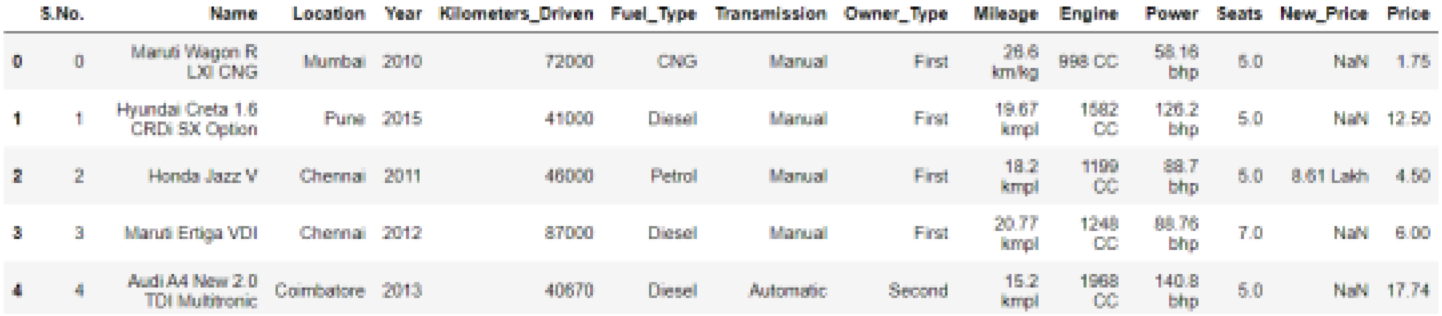
Notes

This assignment should be submitted in Canvas as a format of ipython notebook (assignment1.ipynb).

No late assignments will be accepted. Do not accept any other format. Minimum penalty is 2pts with acceptable excuse.

You may collaborate on homework but must **independently** write code/solutions. Copying and other forms of cheating will not be tolerated and will result in a **zero score** for the homework (minimal penalty) or a failing grade for the course. Your work will be graded in terms of correctness, completeness, and clarity, not just the answer. Thus, correct answers with no or poorly written supporting steps may receive very little credit.

Please download used_cars_data.csv. This is a dataset consisting of used car sales prices.



	S.No.	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0	NaN	1.75
1	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
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.61 Lakh	4.50
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN	6.00
4	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

1. (1 pt) Data cleaning

Please remove the following columns: ‘S.No’ and ‘New_Price’.

Please create ‘Car_Age’ feature which is defined as the difference between the current year and the year the car was built. E.g. “Car_Age” for the first record is 14.

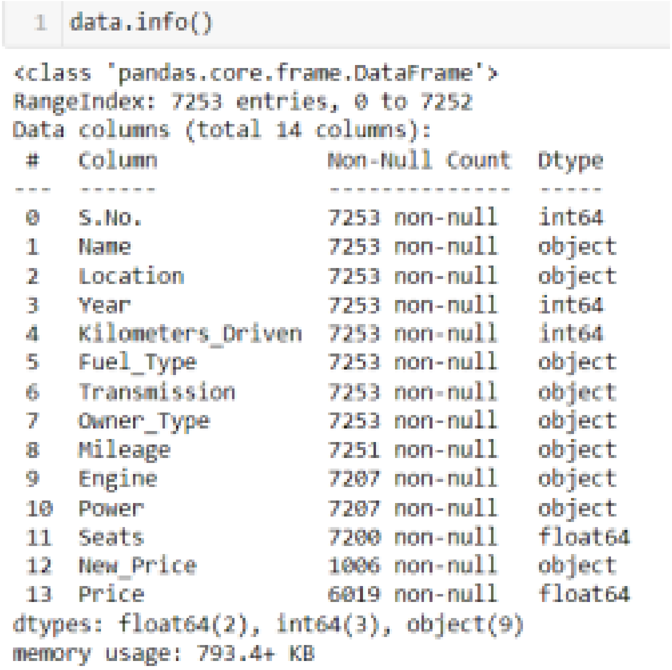
There are two records for Electric vehicles. The corresponding mileage columns are empty. Please search internet what will be the reasonable value and fill out it. Please provide the logics with the reference.

2. (2 pts) Transformation

Among the columns in the dataset, python datatype of the ‘Mileage’, ‘Engine’, ‘Power’ columns are ‘object’.

Please convert them to numerical datatype. Remove unit and convert string to numerical value (floating point or integer)

NOTE:
You should check the unit of the three columns. If there is more than two units in a column, you should find dominant unit and perform unit conversion to achieve consistency within the column.
Please describe step by step for the unit conversion. Please include the reference.



```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7253 entries, 0 to 7252
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   S.No.              7253 non-null   int64
 1   Name               7253 non-null   object
 2   Location           7253 non-null   object
 3   Year               7253 non-null   int64
 4   Kilometers_Driven  7253 non-null   int64
 5   Fuel_Type          7253 non-null   object
 6   Transmission       7253 non-null   object
 7   Owner_Type         7253 non-null   object
 8   Mileage            7251 non-null   object
 9   Engine             7207 non-null   object
 10  Power              7207 non-null   object
 11  Seats              7200 non-null   float64
 12  New_Price          1006 non-null   object
 13  Price              6019 non-null   float64
dtypes: float64(2), int64(3), object(9)
memory usage: 793.4+ KB
```

3. (2.5 pts) Outlier detection and box-plot

Please check whether the data is in normal distribution or non-normal distribution for the following numerical columns: ‘Car_Age’, ‘Kilometers_Driven’, ‘Mileage’, ‘Engine’, ‘Power’, ‘Seats’, ‘Price’

Please detect outliers based on the data distribution type. For outlier detection, please calculate step by step.
Please count(print) how many outliers for each column.

Please draw box-plot for the columns. Please draw box-plot together if the scales of the columns are in similar range. Otherwise, please draw box-plot separately.

Please draw box-plot for ‘Mileage’ with ‘Fuel_Type’.

4. (1.5 pts) Pearson correlation coefficient and scatter plot

Please calculate Pearson correlation coefficient between two columns for the following columns: ‘Car_Age’, ‘Kilometers_Driven’, ‘Mileage’, ‘Engine’, ‘Power’, ‘Seats’, ‘Price’.

NOTE:
Before calculating the coefficient, you need to exclude the outliers.
You should calculate the coefficient from scratch.

Please draw scatterplots between two columns for the following columns: ‘Car_Age’, ‘Kilometers_Driven’, ‘Mileage’, ‘Engine’, ‘Power’, ‘Seats’, ‘Price’. Please include the coefficient value inside of the figures.

5. (3 pts) Handling missing values

There are missing values in the following columns: ‘Mileage’, ‘Engine’, ‘Power’, ‘Seats’, ‘Price’.
Please treat the outliers also as missing values.

(1.5 pt) Please count(print) missing values for each column in the columns of ‘Engine’, ‘Power’, and ‘Seats’.
Please impute the missing values based on subclass (subgroups).
Please draw histogram(distribution) for each column and use different color for the imputed missing values.

NOTE:
For categorical or discrete features, use mode. For continuous features, use mean for all samples belonging to the same subclass.
If imputing using a subclass or multiple subclasses does not impute all the missing values, please impute using the subclass as much as possible. Then, impute using the global constant for the remaining rows.

(1.5 pt) Please count(print) missing values for each column in the columns of ‘Mileage’ and ‘Price’.
Please impute the missing values using linear regression.
Please draw histogram(distribution) for each column and use different color for the imputed missing values

NOTE:
You need to find which columns have strong correlations with ‘Mileage’ or ‘Price’. Then, build the linear regression model using scikit-learn library and apply the model to impute the missing values.