CMPE 255-03, Spring 2024

Assignment #1

Release on Feb 15th, 2023
Due 11:59pm on Sunday, Feb 25th, 2023

Notes

This assignment should be submitted in Canvas as a format of ipython notebook (assignment_1_yourFirstName_LastnName.ipynb).

No late assignments will be accepted.

You may collaborate on homework but must write independent 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. You may remove the following columns: 'S.No' and 'New_Price'.

	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. (3 pts) Transformation

Among the columns in the dataset, the datatype of the 'Mileage', 'Engine', 'Power' columns is 'object'. Please convert them to numerical columns (float or integer). This may involve performing unit conversions to achieve consistency within each column.

2. (2 pts) Outliers and Correlation

NOTE:
'Car_Age' feature needs to be created and 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.
Please calculate Lower Limit and Upper Limit based on IQR from scratch.
You can use any package for calculating the correlation coefficient.

Please check and print outliers based on IQR and draw box-plots accordingly for the following columns: ‘Car_Age’, ‘Kilometers_Driven’, ‘Mileage’, ‘Engine’, ‘Power’, ‘Seats’, ‘Price’

If there are data point(s) that are clearly an error (for example they do not make sense in the context of the column. Additionally, being an outlier based on IQR does not make a data point a clear error necessarily), then the data point(s) should be dealt with in an appropriate fashion. If needed, outliers should be checked and printed again, and the boxplots should be redrawn.

Please calculate correlation coefficient and create scatterplot against ‘Price’ for the following columns: ‘Car_Age’, ‘Kilometers_Driven’, ‘Mileage’, ‘Engine’, ‘Power’, ‘Seats’.

Please describe/summarize your observations with respect to data distribution, outliers, and correlation.

```
1 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
```

```
1 data.isnull().sum()
```

```
S.No.                 0
Name                  0
Location              0
Year                  0
Kilometers_Driven     0
Fuel_Type             0
Transmission          0
Owner_Type            0
Mileage               2
Engine               46
Power                46
Seats                53
New_Price          6247
Price              1234
```

3. (3 pts) Handling missing values

There are missing values in the following columns: ‘Mileage’, ‘Engine’, ‘Power’, ‘Seats’, ‘Price’. Please impute the missing values using subclass (subgroups). As discussed during lecture, categorical or discrete features use mode and continuous features use mean or median for all samples belonging to the same subclass. Please justify your choice of mode, mean, or median in each case.

NOTE:
If imputing using a subclass or multiple subclasses does not get rid of all the missing values, please impute using the subclass as much as you can. Then, the remaining rows can be used without using subclass.

4. (2 pts) Outliers and Correlation with the imputed data points

Please repeat #2 with the imputed data points. The resulting plots should contain both original and imputed data points in different colors to distinguish one from the other; please use a clear legend in each plot.

Please describe/summarize your observations with respect to data distribution, outliers, and correlation after imputation.