Team 4
Assignment-1

Questions 1:

A. Summarize the ToyotaCorolla Dataset:

Number of columns in the dataset - 39 and Number of rows in the dataset - 1436.
The Dataset has 28 Categorical variables and 10 continuous variables.

The Dataset has no null values and has data types which includes integers and objects.

```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Id                 1436 non-null   int64
 1   Model              1436 non-null   object
 2   Price              1436 non-null   int64
 3   Age_08_04          1436 non-null   int64
 4   Mfg_Month          1436 non-null   int64
 5   Mfg_Year           1436 non-null   int64
 6   KM                 1436 non-null   int64
 7   Fuel_Type          1436 non-null   object
 8   HP                 1436 non-null   int64
 9   Met_Color          1436 non-null   int64
 10  Color              1436 non-null   object
 11  Automatic          1436 non-null   int64
 12  CC                 1436 non-null   int64
 13  Doors              1436 non-null   int64
 14  Cylinders          1436 non-null   int64
 15  Gears              1436 non-null   int64
 16  Quarterly_Tax      1436 non-null   int64
 17  Weight             1436 non-null   int64
 18  Mfr_Guarantee      1436 non-null   int64
 19  BOVAG_Guarantee    1436 non-null   int64
 20  Guarantee_Period   1436 non-null   int64
 21  ABS                1436 non-null   int64
 22  Airbag_1           1436 non-null   int64
 23  Airbag_2           1436 non-null   int64
 24  Airco              1436 non-null   int64
 25  Automatic_airco    1436 non-null   int64
 26  Boardcomputer      1436 non-null   int64
 27  CD_Player          1436 non-null   int64
 28  Central_Lock       1436 non-null   int64
 29  Powered_Windows    1436 non-null   int64
 30  Power_Steering     1436 non-null   int64
 31  Radio              1436 non-null   int64
 32  Mistlamps          1436 non-null   int64
 33  Sport_Model        1436 non-null   int64
 34  Backseat_Divider   1436 non-null   int64
 35  Metallic_Rim       1436 non-null   int64
 36  Radio_cassette     1436 non-null   int64
 37  Parking_Assistant  1436 non-null   int64
 38  Tow_Bar            1436 non-null   int64
dtypes: int64(36), object(3)
memory usage: 437.7+ KB
```

```
ToyotaCorolla_df.isna().sum()
```

```
Id                  0
Model               0
Price               0
Age_08_04           0
Mfg_Month           0
Mfg_Year            0
KM                  0
Fuel_Type           0
HP                  0
Met_Color           0
Color               0
Automatic           0
CC                  0
Doors               0
Cylinders           0
Gears               0
Quarterly_Tax       0
Weight              0
Mfr_Guarantee       0
BOVAG_Guarantee     0
Guarantee_Period    0
ABS                 0
Airbag_1            0
Airbag_2            0
Airco               0
Automatic_airco     0
Boardcomputer       0
CD_Player           0
Central_Lock        0
Powered_Windows     0
Power_Steering      0
Radio               0
Mistlamps           0
Sport_Model         0
Backseat_Divider    0
Metallic_Rim        0
Radio_cassette      0
Parking_Assistant   0
Tow_Bar             0
dtype: int64
```

Summarizing the dataset:

	count	mean	std	min	25%	50%	75%	max
Price	1436.0	10730.824513	3626.964585	4350.0	8450.0	9900.0	11950.00	32500.0
Age_08_04	1436.0	55.947075	18.599988	1.0	44.0	61.0	70.00	80.0
KM	1436.0	68533.259749	37506.448872	1.0	43000.0	63389.5	87020.75	243000.0
CC	1436.0	1576.855850	424.386770	1300.0	1400.0	1600.0	1600.00	16000.0
Quarterly_Tax	1436.0	87.122563	41.128611	19.0	69.0	85.0	85.00	283.0
Guarantee_Period	1436.0	3.815460	3.011025	3.0	3.0	3.0	3.00	36.0
Mfg_Year	1436.0	1999.625348	1.540722	1998.0	1998.0	1999.0	2001.00	2004.0
Mfg_Month	1436.0	5.548747	3.354085	1.0	3.0	5.0	8.00	12.0
HP	1436.0	101.502089	14.981080	69.0	90.0	110.0	110.00	192.0
Weight	1436.0	1072.459610	52.641120	1000.0	1040.0	1070.0	1085.00	1615.0

1. The dataset Toyota Corolla has 1436 entries based on the various features which impacts the price and performance in the market.
2. Price ranges from 4350 to 32500 with an average price of 10730. Median indicates that half of the vehicles are below 9900.
3. The variation in age from 1 month to 80 months old signifies that there are latest vehicles as well as old vehicles are included which is affecting the price of the vehicle. Similarly, the minimum age value of 1.0 means that these models are the latest ones.
4. The wide range(1- 243000) in Kms indicates the usage of the vehicles from the time they were manufactured and used. Similarly, Cars with higher kilometers suggest that they've been driven for a long time, which means that they might have been manufactured and sold early on.
   Mean (68533) and median (63389) indicate that the variable KM is skewed to the left.
5. For CC the mean is around 1500 but the maximum value is 16000 suggesting that it's an outlier, or it can be a data error or sampling error. This also indicates that the car might be a newer model because of the high CC.
   Cars can also be categorized based on their HP and Weight. For example, Cars with higher HP might be put into the sports car category, and more heavy Cars might be SUVs.
6. For Mfg- year, 75% of the cars were manufactured in 2001 or before.
7. The horsepower of cars has a mean of 101. 5 and std dev of 14, which indicates most of the vehicles have horsepower nearly around mean value, 86.5 and 116.5. Also, (min HP - 69.5 has price of 4350 and max HP 192 has price of 32500), which indicates that with the increase in horsepower ranges the price of the vehicle increases.
   Median and 3rd quartile have same hp values, which might indicate that the majority of the vehicles are close to or equal to 110 hp.
8. The mean price is higher than the median price, which suggests that there are some outliers (e.g. premium models).

B. Normalizing the Variable KM:

```
from sklearn.preprocessing import MinMaxScaler, StandardScaler


# using sklearn:
scaler = StandardScaler()
ToyotaCorolla_df[['KM']] = pd.DataFrame(scaler.fit_transform(ToyotaCorolla_df[['KM']]))
```

Used standard scaler from sklearn to normalize the KM Variable.

Output:

```
ToyotaCorolla_df['KM'].head()
```

```
0   -0.574695
1    0.117454
2   -0.715386
3   -0.547650
4   -0.801028
Name: KM, dtype: float64
```

C. Create dummies for the variable Fuel Type:

Fuel_Type_CNG	False	False	False	False	False
Fuel_Type_Diesel	True	True	True	True	True
Fuel_Type_Petrol	False	False	False	False	False

D. Partition the data into three sets.

70% is training data and 30% is divided between validation and testing data with random state of 1. Out of the 30% of remaining data 10% of it is validation data and 20% is test.

```
# using sklearn
trainData, temp = train_test_split(ToyotaCorolla_df, test_size=0.3, random_state=1)
validData, testData = train_test_split(temp, test_size=0.2, random_state=1)
print('Training   : ', trainData.shape)
print('Validation : ', validData.shape)
print('Test       : ', testData.shape)
```

```
Training   :  (1005, 39)
Validation :  (344, 39)
Test       :  (87, 39)
```