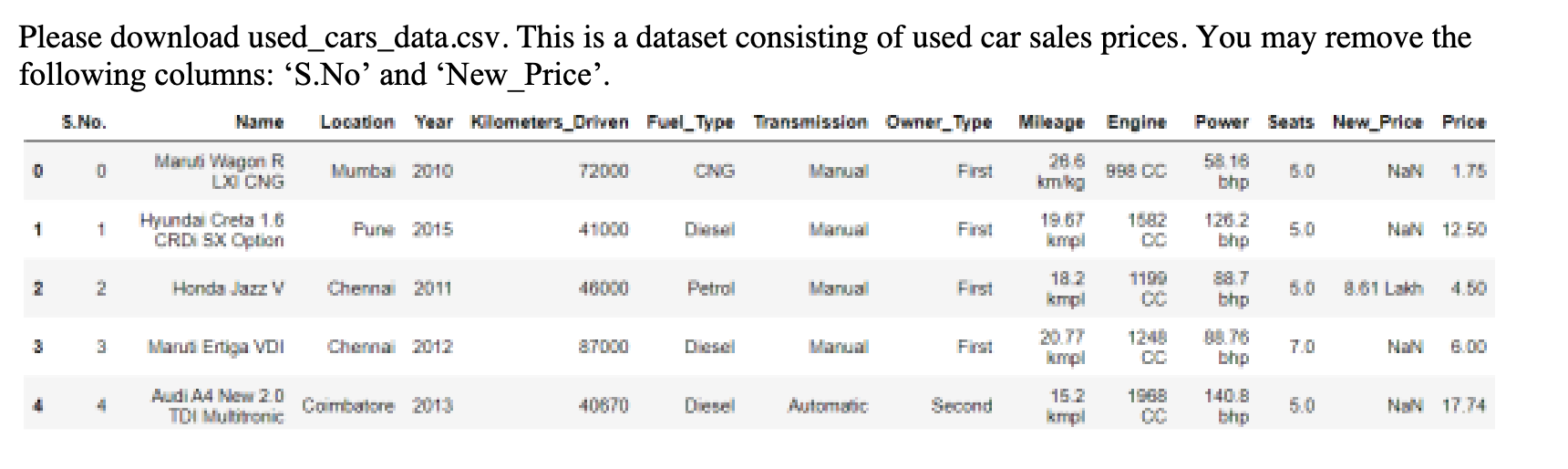
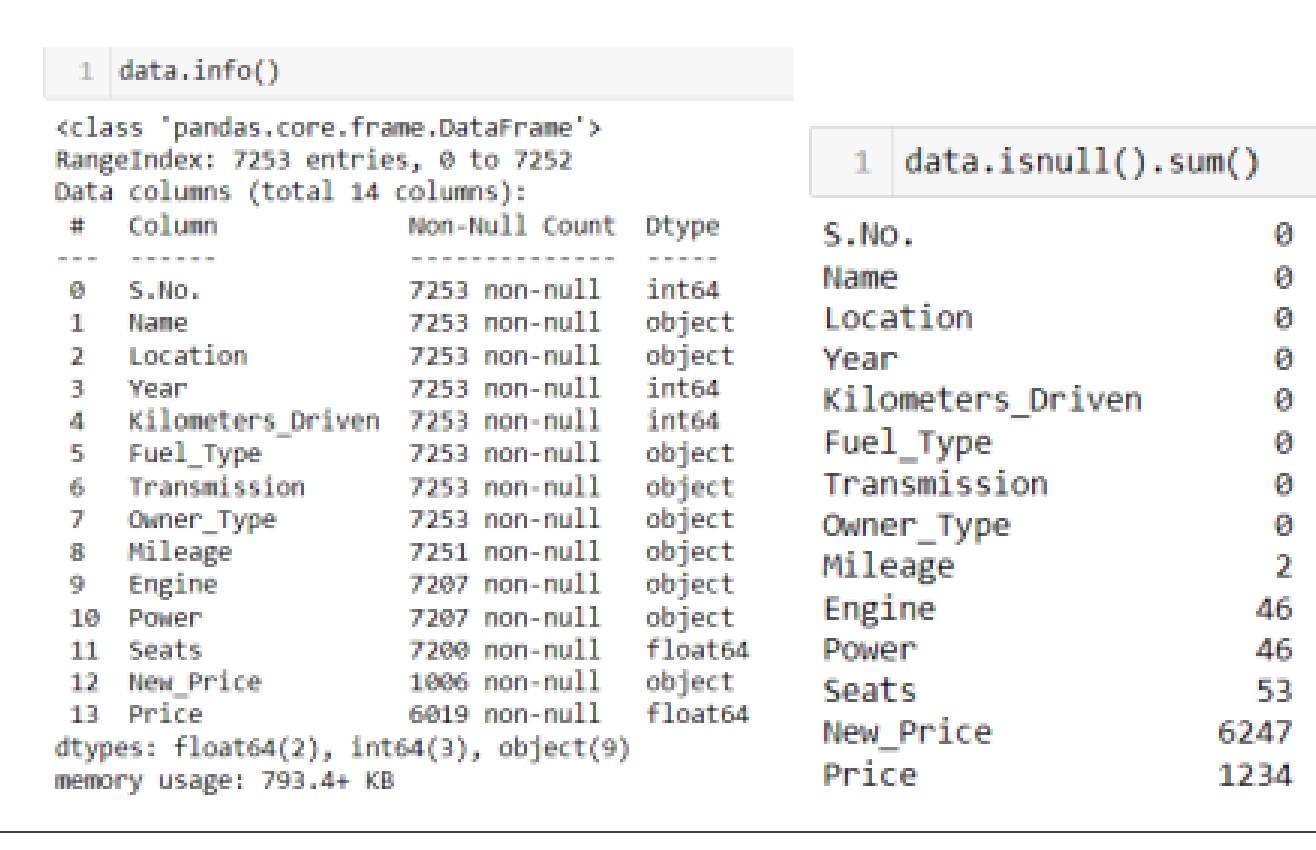
**DATA 240 ASSIGNMENT**

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’.  
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.

DATA 240, Spring 2024 Page 2 of 3  
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’. When printing the outliers,  
please print the number of outliers in each column, and the number of records with at least one outlier.  
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.

  
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.