**Advance Data Analysis Project-1**

**Data Cleaning**

· After loading the dataset. It includes some features which do not provide any useful information, or have very less amount of data with a lot of missing values.

*Dropping unwanted features :*

· ‘notRepairedDamage’ have many NaN values more than 20% and ‘nrOfPictures’ column have zero variance and all of the values in this column were zeroes. Hence both these columns were dropped.

*Dropping null values:*

· After dropping the columns, all the records that have null values in any of the columns were dropped.

*Structural Errors:*

· Features with datetime data type where the time and seconds are consistently set to "00:00:00" have been removed.

· There is some inaccurate data in the 'yearOfRegistration' and 'powerPS' features, where values that fall below the required minimum or exceed the expected maximum have been removed.

*Outliers:*

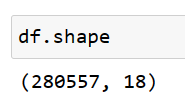
Price was zero, below 500 dollars and above 4,00,000 dollars, those rows were identified as outliers and were dropped.

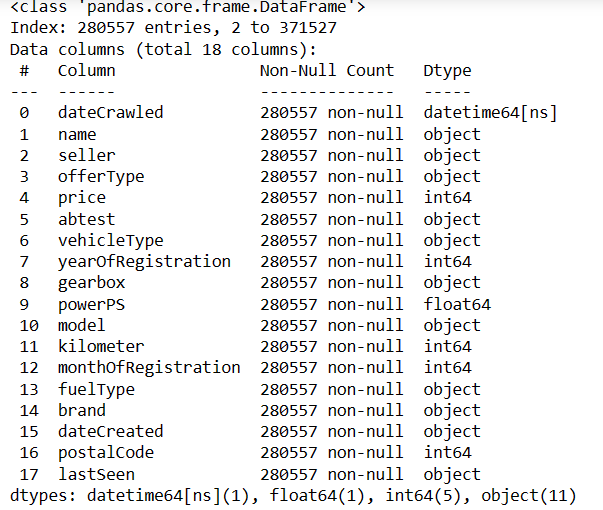
*Duplicates:*

· Duplicated rows were identified and were deleted and there were 280557 rows and 18 columns that are useful and accurate for the analysis.

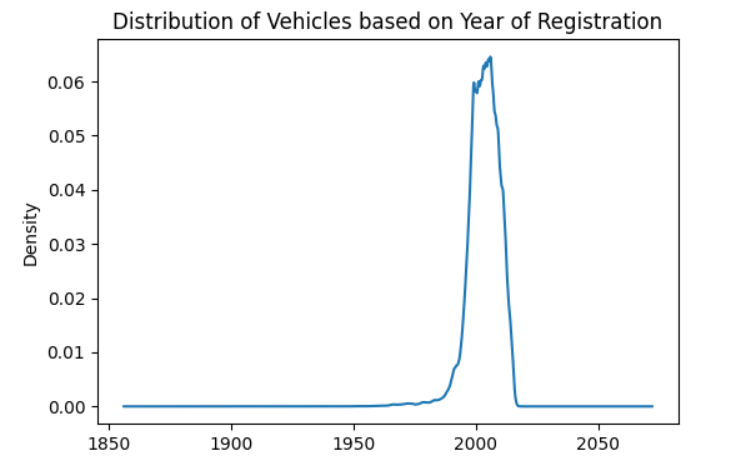
**Analysis 1**

1) Perform general Data analysis



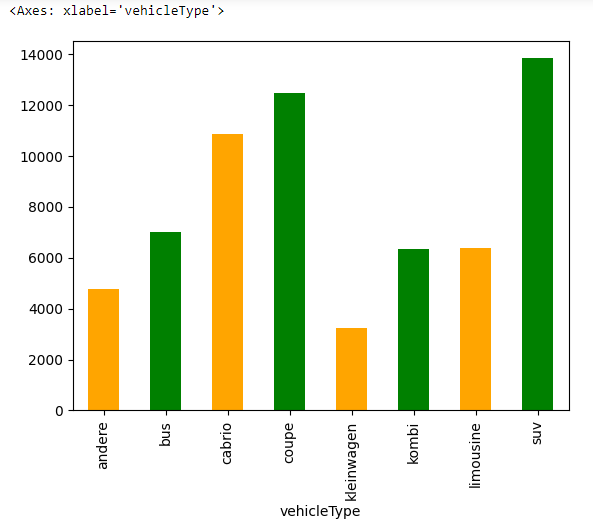


2)Can you tell me the Distribution of Vehicles based on Year of Registration with the help of a plot.

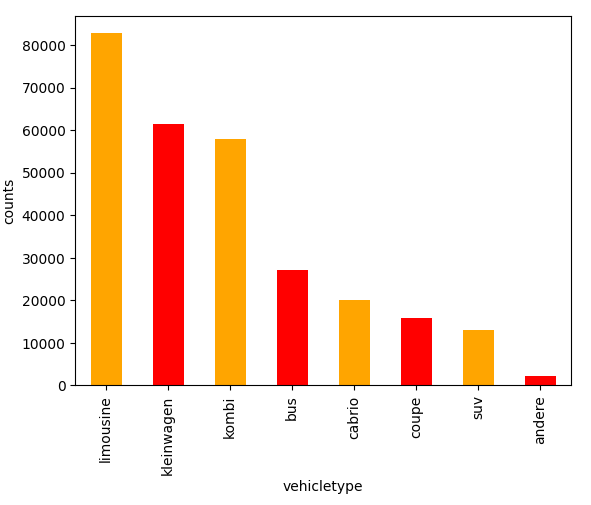


* The 'yearOfRegistration' variable exhibits a normal distribution, as evidenced by its bell-shaped curve with a single peak.

3) Create a plot based on the Variation of the price range by the vehicle type

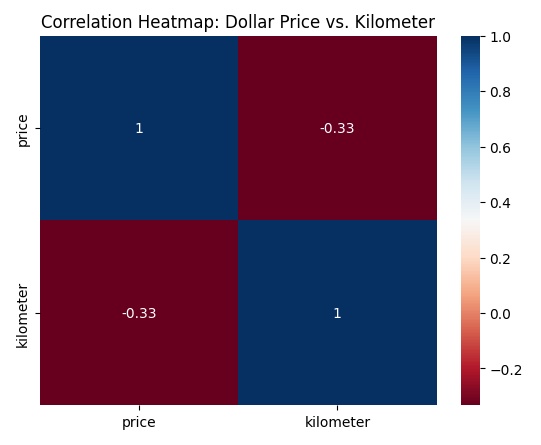
* 
* When we group the rows based on the 'vehicleType' column and look at the average prices, SUVs and coupes turn out to be more expensive, while kleinwagens are the most budget-friendly compared to other types of vehicles.

4) Find out Total count of vehicles by type available on ebay for sale.As well as create a visualization for the client.



* limousine brand offers a wider variety of vehicle types, while the andere brand has a more limited range of vehicle types.

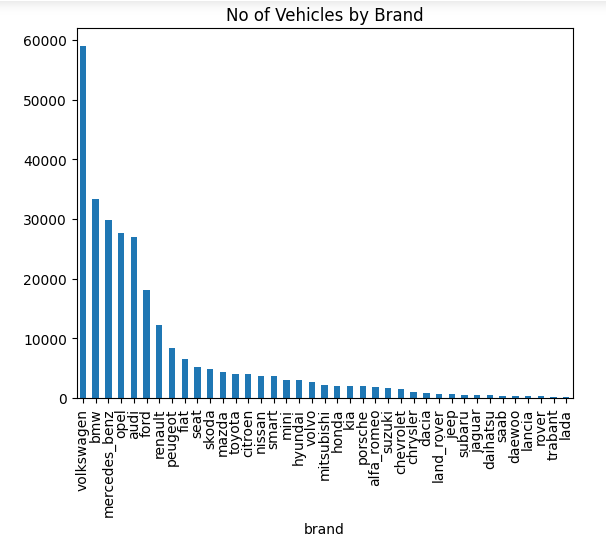
5) Is there any relationship between dollar\_price and kilometer? (Explain with appropriate analysis)



* There is a negative correlation of -0.332 between the price and the kilometer.

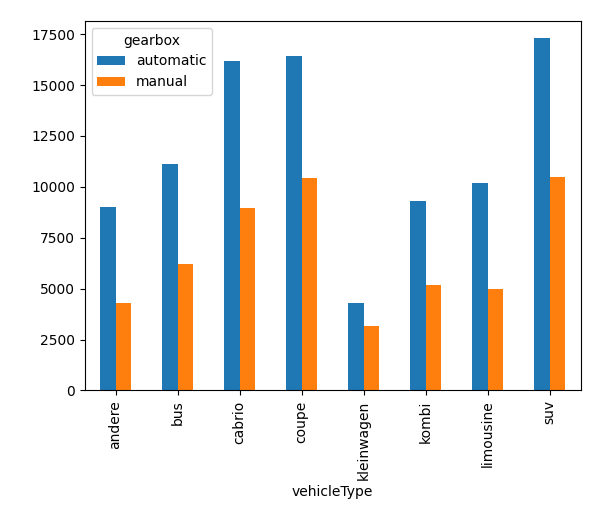
**Analysis - 2**

1. Can you tell me No of Vehicles by Brand Available on ebay for sale with the help of visualization



* volkswagen brand has highest number of sales on ebay

2. What is the Average price for vehicles based on the type of vehicle as well as on the type of gearbox.Explain me with both numerical and visualization analysis



* Above graph shows Average price of vehicles based on the geartype and SUV has highest price in automatic gearbox, price of cabrio and coupe has almost equal price on automatic geartype.

3. What is the marginal probability of private seller.

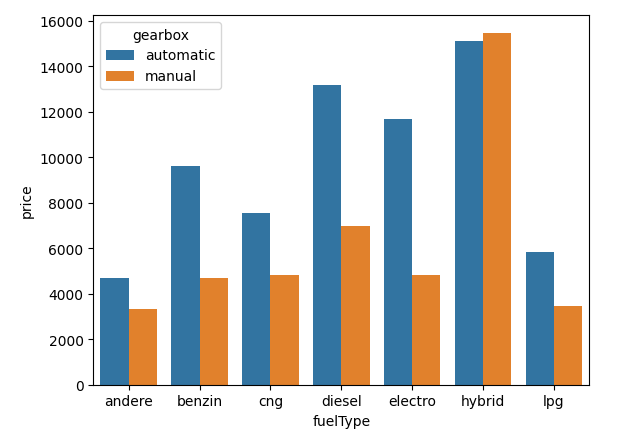
Ans) The Marginal probability of Privat seller is 0.999993.

**Analysis 3**

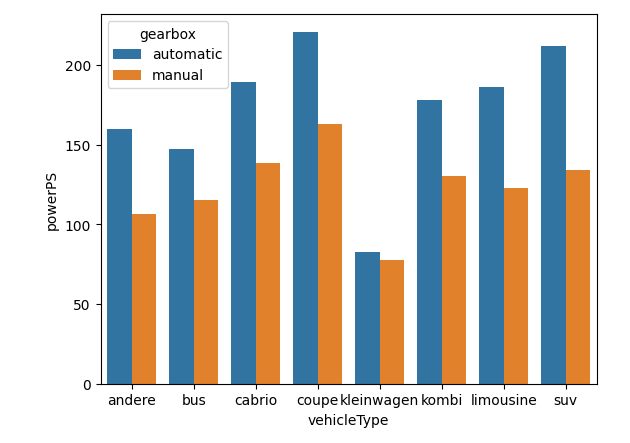
1. The memory usage of the data is around 6.1 mb.How can we reduce the memory usage of the data set?

Reducing memory usage can be done through by converting data types from integer to float and vice versa, optimizing text data, and eliminating not reqired columns to streamline data size for analysis.

1. What is the Average price of vehicle by fuel type and gearbox type.Give a plot

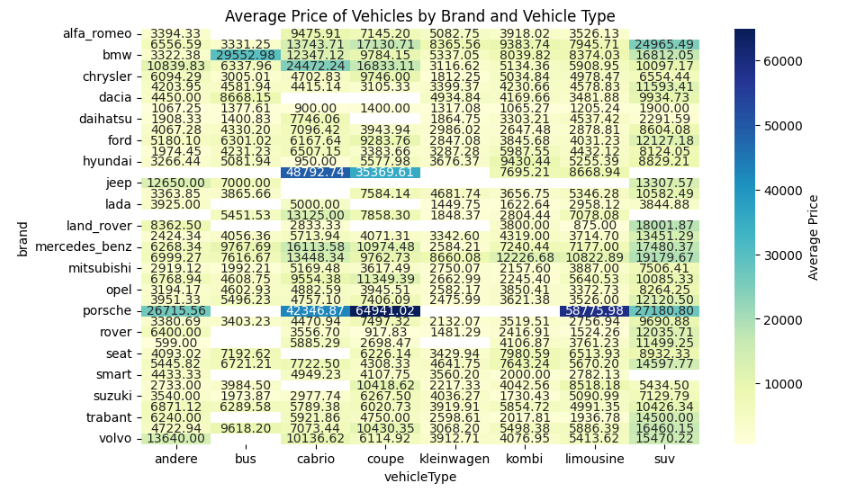
'Hybrid' vehicles with both automatic and manual gearboxes have the highest prices compared to other fuel types.

1. What is the Average power of a vehicle by vehicle type and gearbox type.Give a plot



* The average powerPS of vehicles based on the vehicle type and gearbox, where coupes exhibit the highest average power, followed by SUVs, while kleinwagens has the lowest average power.

4. What is the Average price of a vehicle by brand as well as vehicle type.Use heatmap to explain this

· Most of the cars in all brands lie below 10000 dollars and Porshe’s coupe is the most expensive in this dataset. Jaguar and Porshe have some of the most expensive cars in the dataset.