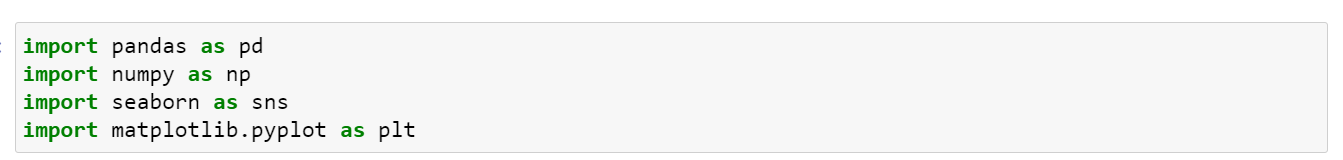
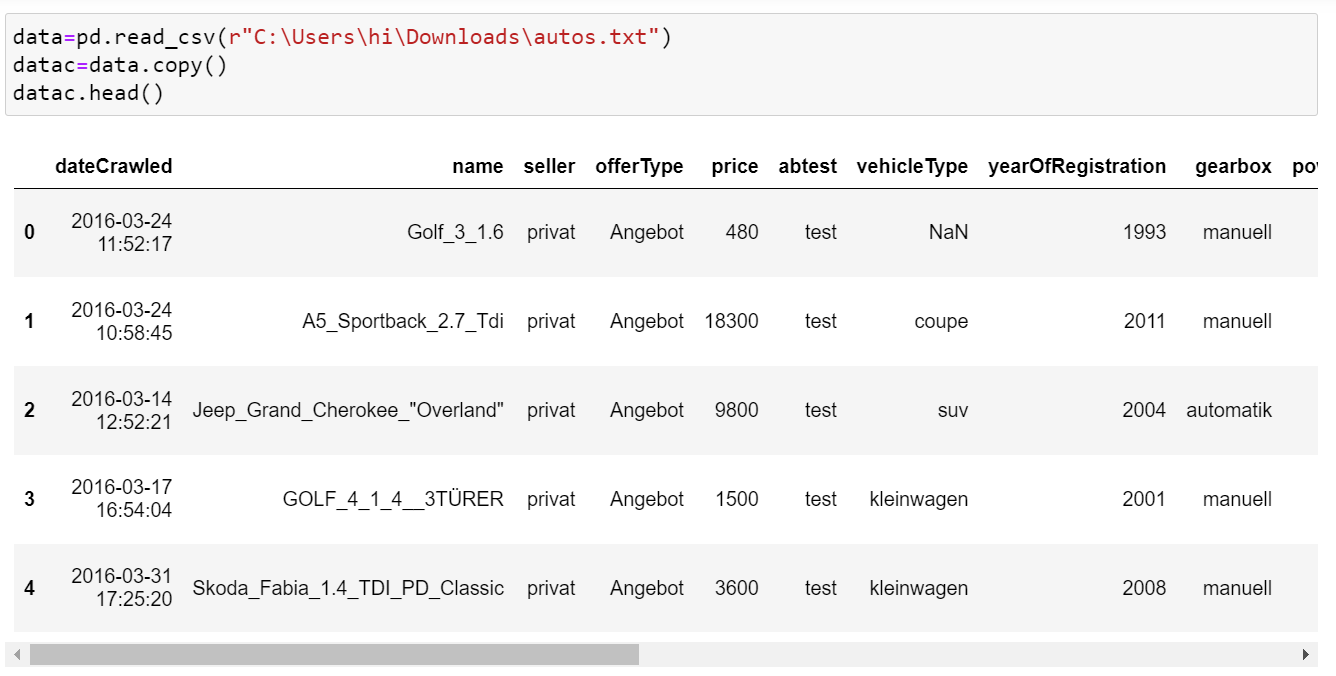
**Report**

Firstly, we have import python libraries for analyze the data and visualize the data. The python libraries as follows:



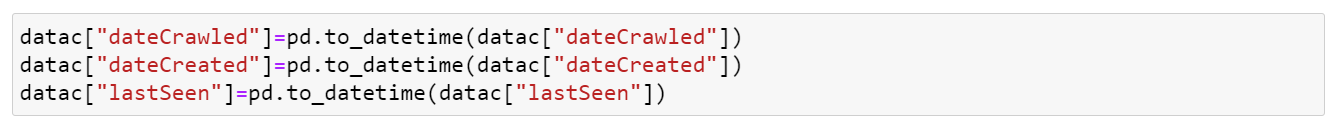
Next step is loading the dataset. After loading the data save the copy of the data as datac. So, the original won’t get effected. To see preview of the dataset use datac.head (). We can see the code below:



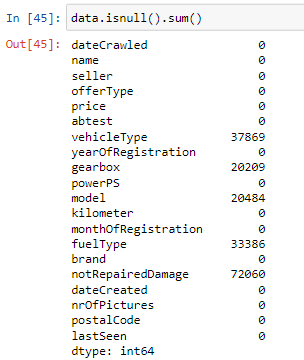
**Analysis-1:**

1. **Perform General Data Analysis:**

* First check the dataset by datac.info ()
* we can see few columns data types have to be change according to the requirement. So, we changed the data types columns by below code:

****

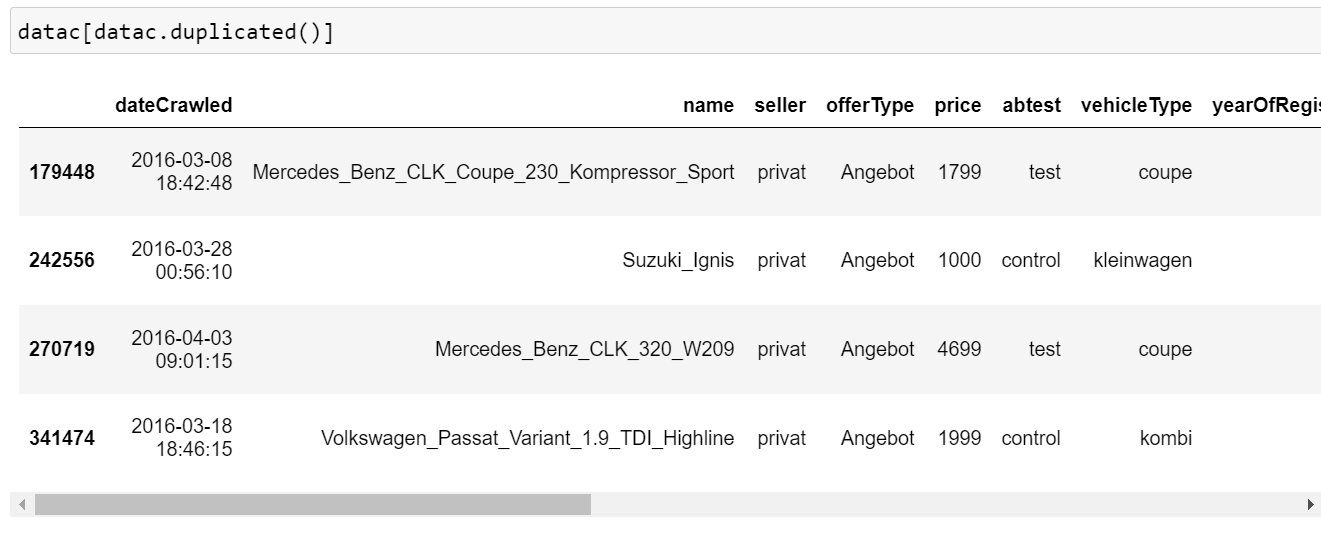
* Check if we have any null values. To find the null values the Code is:



* So, we have 37869 null values in vehicle type column, 20209 null values in gearbox column ,20484 null values in model column, 33386 null values in fuel type column and 72060 null values in not repaired damage column.
* Replaced the missing values with appropriate mode value according to that particular column. The code is:

**datac.fillna ({"notRepairedDamage":data["notRepairedDamage"].mode()[0],"fuelType":data["fuelType"].mode()[0],"vehicleType":data["vehicleType"].mode()[0],"gearbox":data["gearbox"].mode()[0],"model":data["model"].mode()[0]},inplace=True)**

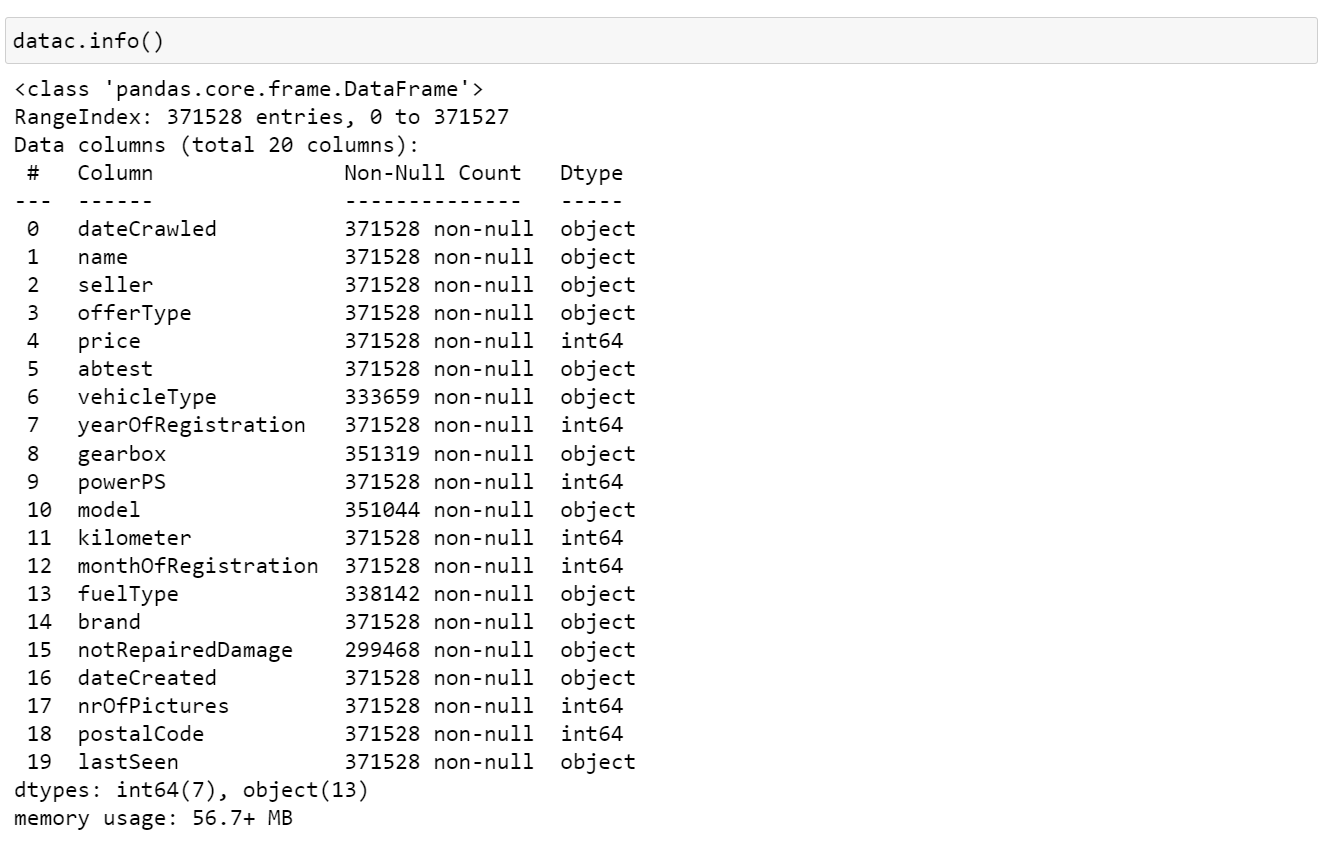
* There are 4 duplicates rows in given data set. To find duplicates rows use below code:



* We dropped all 4 duplicate rows and dropped the unwanted column by below code:

****

* Once again use data.info () to check whether we missed anything. After checking we can say given data is clear.



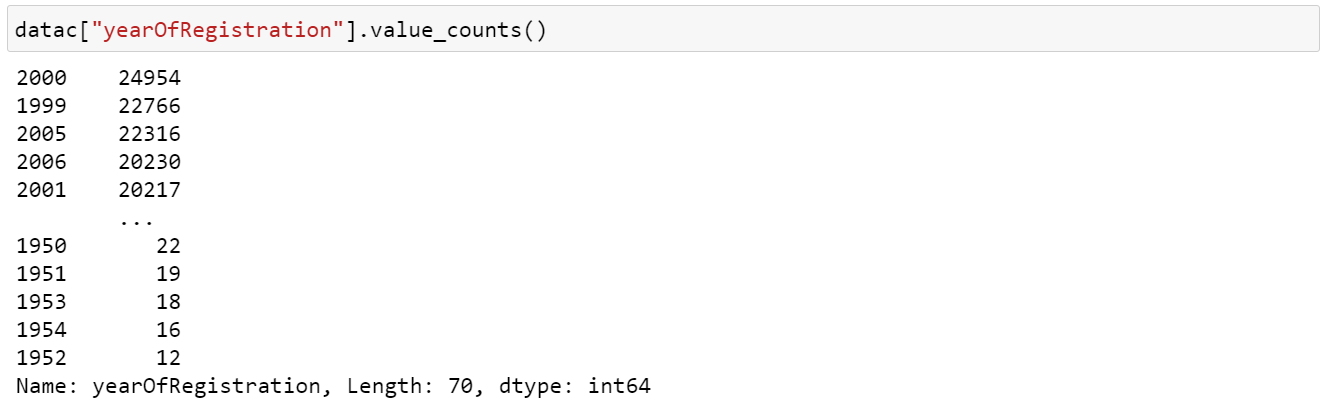
* So, the general data analysis is done!

1. **Can you tell me the Distribution of Vehicles based on Year of Registration with the help of a plot:**

* The distribution of vehicles based on year of registration provides the insights of no of vehicles that are registered in particular year.
* As we have lot of years, we pick a period of time tell distribution of vehicles based on year of registration. For that the code is:

**datac["yearOfRegistration"]=datac["yearOfRegistration"].where((datac["yearOfRegistration"]>=1950) & (datac["yearOfRegistration"]<=2023),datac["yearOfRegistration"].mode()[0])**

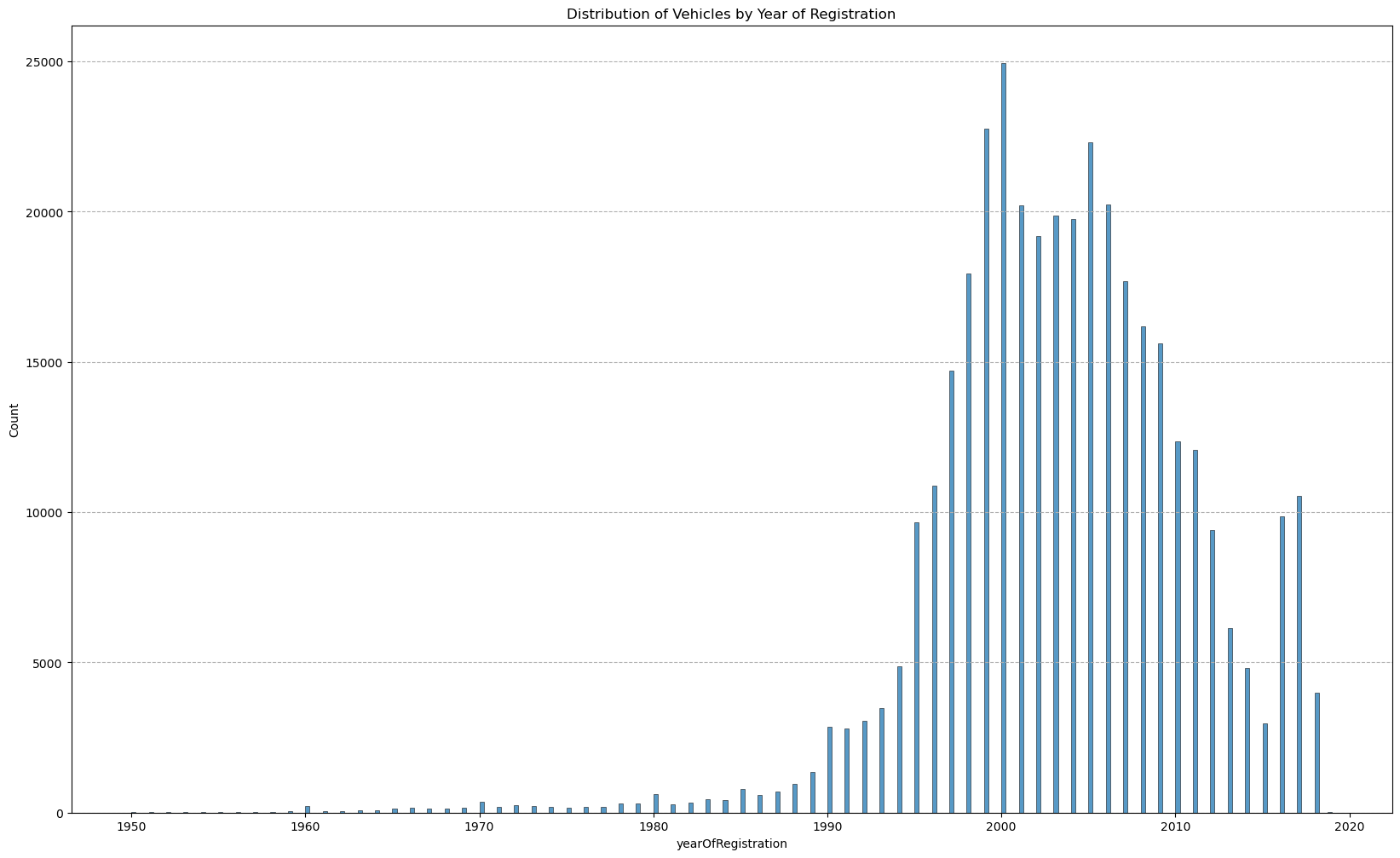
* Now we can see no of vehicles that are registered in each year, below is code and output:



* In the year 2000, 24954 no of vehicles are registered which is highest among the all years in the data set. 1999 year have the second highest registrations with 22766 no of vehicles. 1952 year has lowest registrations with 12 no of vehicles.
* For visualization, we used histogram plot to visualize the distribution of vehicles based on the year of registration. The code is:

****

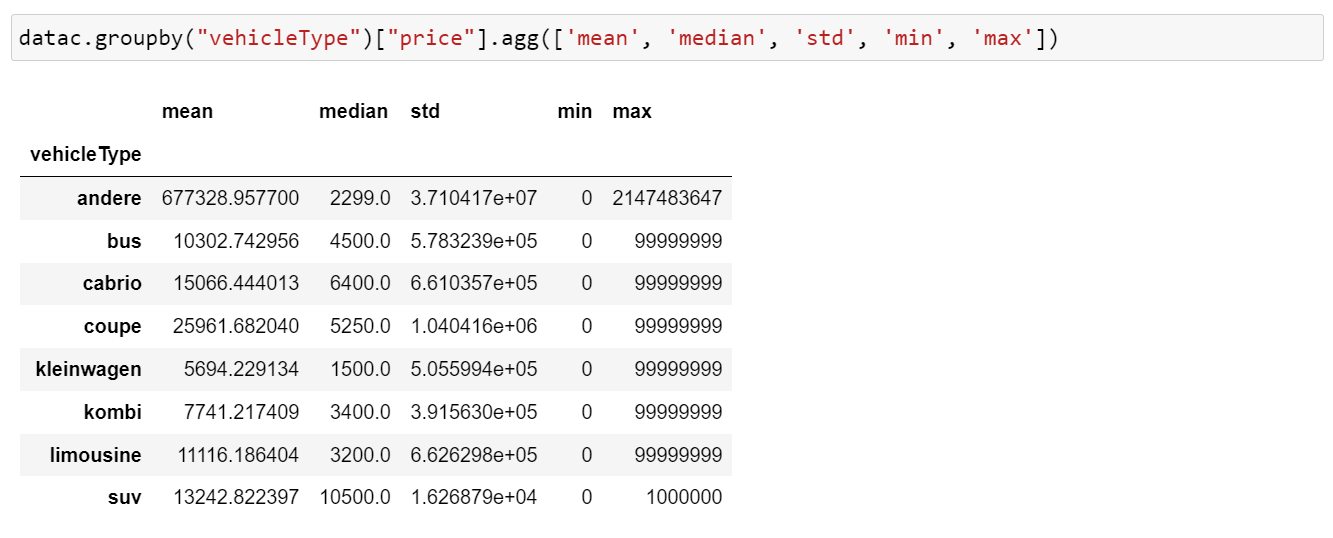
* The visualization for distribution of vehicles by year of registration is



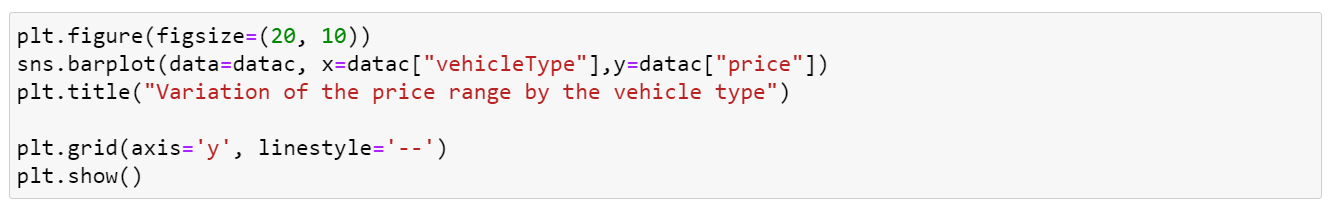
* The x-axis represents the years, and y-axis represents the count of vehicles registered in each year.

1. **Create a plot based on the Variation of the price range by the vehicle type:**

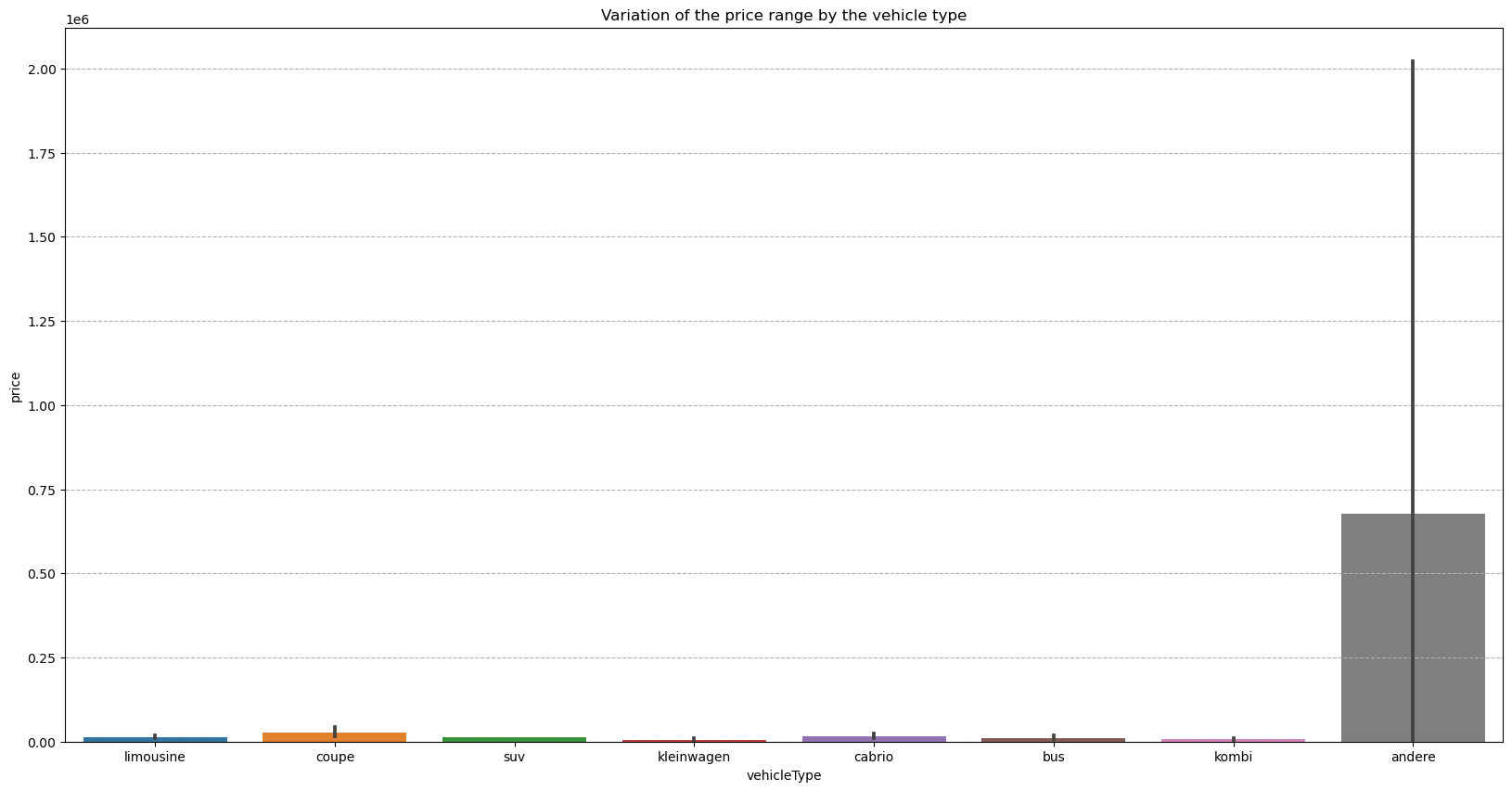
* Variation of price range by vehicle type provides the information about differences in prices of different vehicle types.
* Now we divide the all-vehicle types by groupby and access price column and apply aggregate function types like mean, median, max, min to see the variation of the price range by the vehicle type. The code and output are:

****

* Out of all vehicle types, “andere” vehicle type has high difference in price range. This vehicle type vehicles are high in price.
* Remaining vehicle type price ranges have moderate variation when compared with “andere” vehicle type.
* This helps us to see if some vehicle type is more expensive than other vehicle type by comparing them.
* We used bar plot to visualize the Variation of price range by vehicle type. The code is:

****

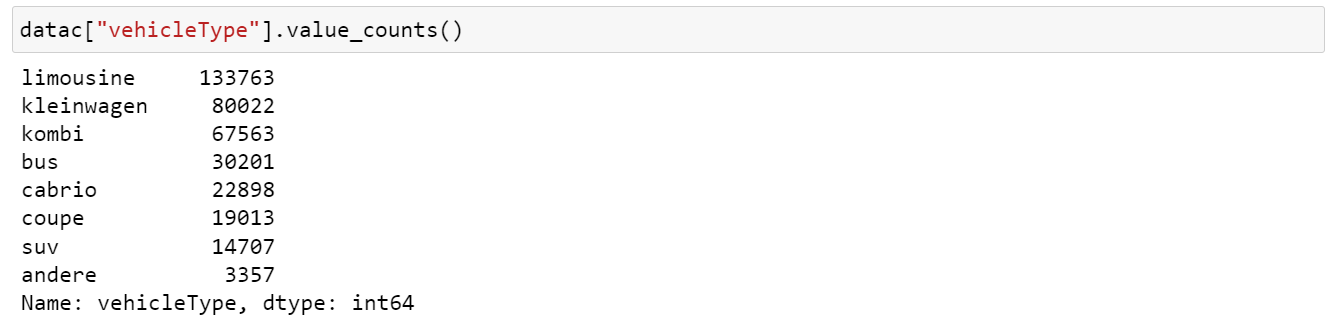
* The visualization for the Variation of price range by vehicle type is



* The y-axis represents the price range, and the x-axis will represent the vehicle type.
* We can see outliers by the visualization in the given data.
* Because of outliers there’s significant difference in price variation of vehicle types.

1. **Find out Total count of vehicles by type available on eBay for sale. As well as create a visualization for the client:**

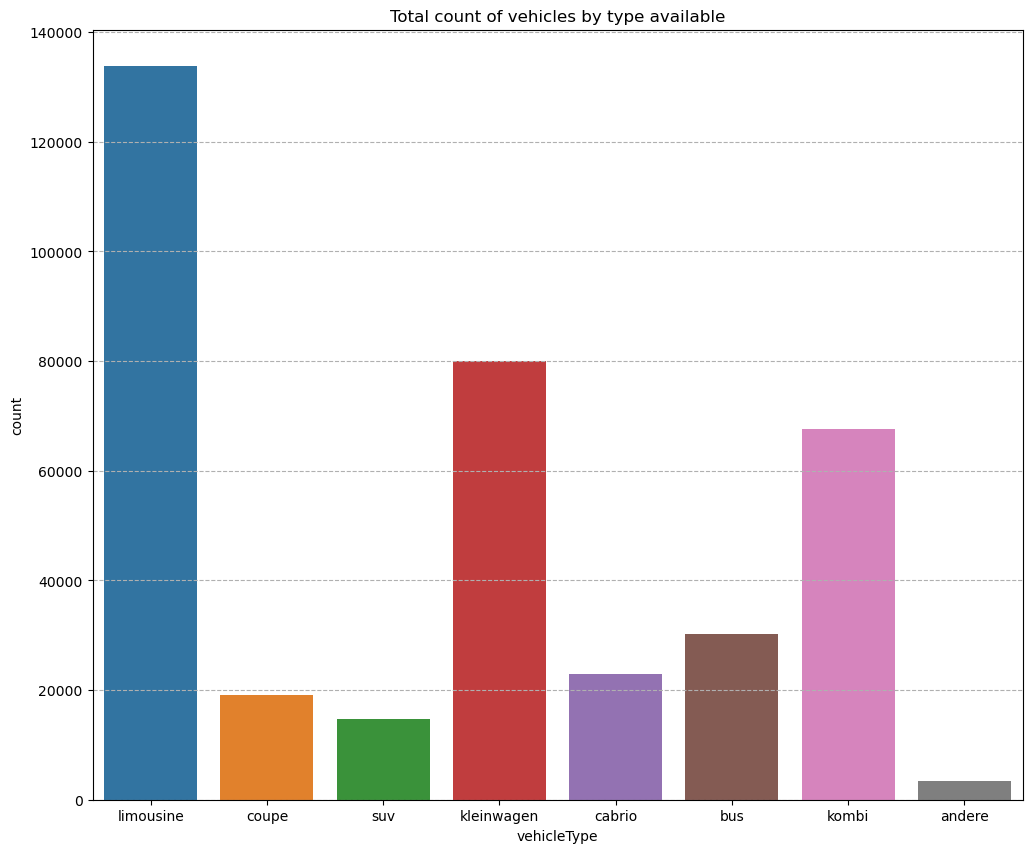
* In total count of vehicles by type available, you can find the total count of vehicles for each vehicle type.
* To know the total count of vehicles we will access vehicle type column and apply value\_counts ().
* Now we can see no of vehicles that are registered in each year, below is code and output:

****

* Vehicle type “limousine” has highest count of 133763 for sale.
* Vehicle type “andere” has lowest count of 3357 for sale.
* This will help us to see how many kinds of vehicles are available on eBay.
* For visualization, we used bar plot to visualize the total count of vehicles by vehicle type. The code is:

****

* To get to clear visualization, we can increase the figure size.
* The visualization for the total count of vehicles by vehicle type is



* The x-axis represents the vehicle type, and y-axis represents the count of vehicles.

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

* To analyse the relationship between dollar price and kilometer, you can use one of the statistical method “Correlation coefficient”.
* By Correlation coefficient, we can check if there’s a relationship between how much a car costs and how many kilometers vehicle have been driven.

### Firstly, we have to access both price and kilometer column and apply correlation function to know the relation. The code is:

### 

* If the correlation lies between 0 to 0.5, we can say that moderately positively correlated.
* If the correlation lies between 0 to -0.5, we can say that moderately negatively correlated.
* If the correlation lies between 0.5-1, we can say that highly positively correlated.
* If the correlation lies between -0.5 to -1, we can say that highly negatively correlated.
* The correlation between “price” and “kilometer” is -0.37305. It indicates that the relationship between two variables is moderately negatively correlated.
* In other words, if one variable(price) increases, another variable (kilometer) decreases and vice versa.
* Correlation coefficient (-0.37305) is not very close to 1, which means correlation between two variables is little weak. Correlation coefficient (-0.37305) is not very far from 1 either, so we can say there is a moderate relationship between two variables, but not

strong one.

### For visualization, we used heatmap to visualize the relation between price and kilometer. The code is:

### 

### The visualization for relationship between price and kilometer is

### 