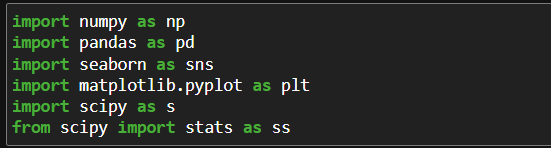
REPORT ON ANALYSIS - 1

Importing the data:-

* To do the analysis of data we have to import the data in our IDE.
* We are using Python IDE to do the analysis, In the IDE we have to import different libraries to perform analysis like Pandas, Numpy, Stats, Matplotlib and Seaborn for visualization.
* The syntax to import libraries is given in the image shown below.



* After importing the libraries we have to load the dataset in our IDE like shown below.



* We are naming our data set as data1 and copying a duplicate of dataset named as data.
* Now we have our dataset loaded in our IDE. So, let’s start our analysis.

GENERAL ANALYSIS ON THE DATA

* To perform general analysis we have to know the general details about the dataset like,

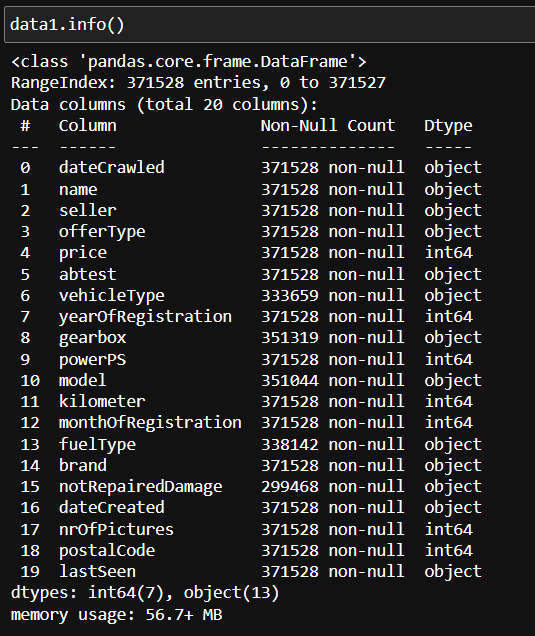
Number of rows in the data.

Number of columns in the data.

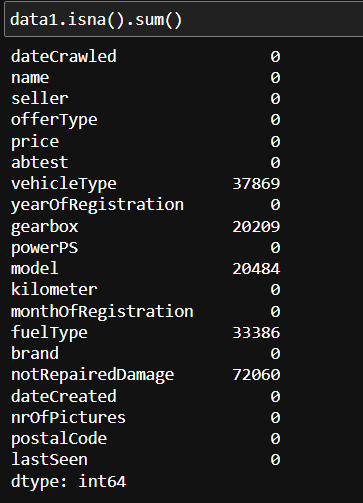
Number of null values in the data.

Type of the data given.

* We can know the basic details by using a simple keyword “info”.



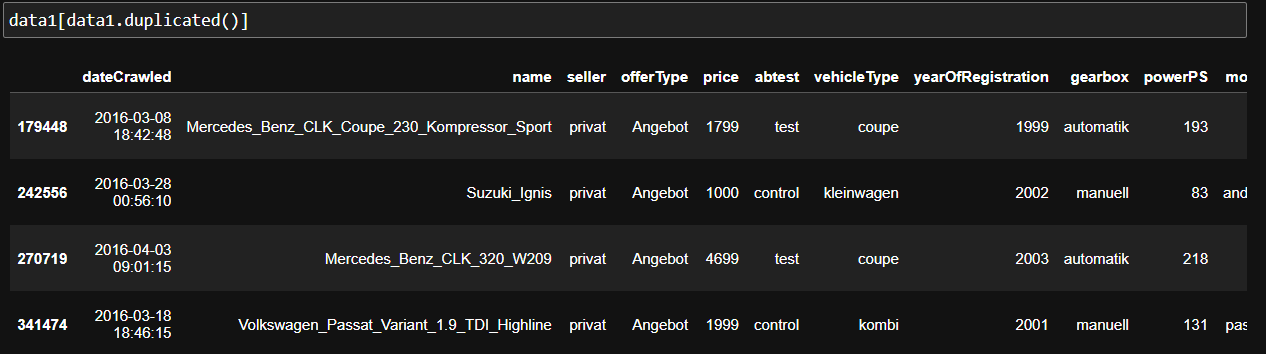
* There are total 20 columns in the data and 371528 rows.
* There are some null values in the data which we can identify by seeing the columns and non-null values less than 371528.
* Let’s find the null values.



* The number of null values in each and every column are shown above by using a keyword “isna()” and “sum()”
* There are 37869 null values in the vehicletype column, 20209 null values in the gearbox column, 20484 null values in the model column, 33386 null values in the fueltype column, 72060 null values in notrepairedamage column.
* We can fill the null values in the dataset with our own values according to the type of data in the column with a keyword called “fillna()”.

REMOVING DUPLICATES :-

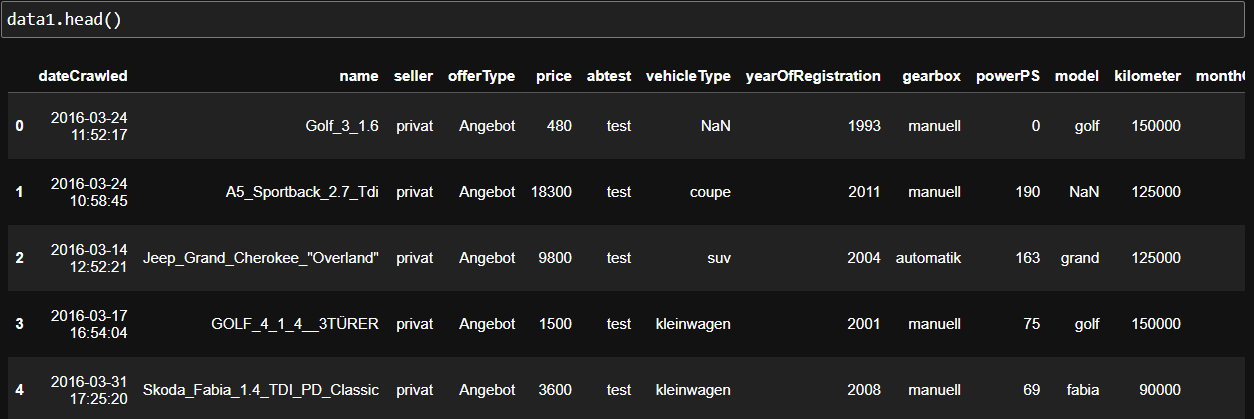
* Duplicates are the rows which are same rows which are repeated and unnecessary in the dataset.
* We can see the duplicates by using the keyword called “data.duplicated()”.



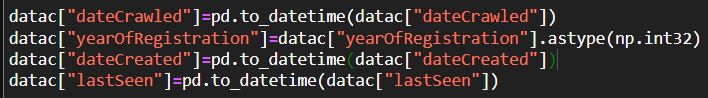
* The above shown rows are duplicate rows which we can remove if we want by using keyword called “drop\_duplicates()”.
* After using the above syntax the duplicated values will be eliminated in the dataset.

Datatype conversion of the columns:

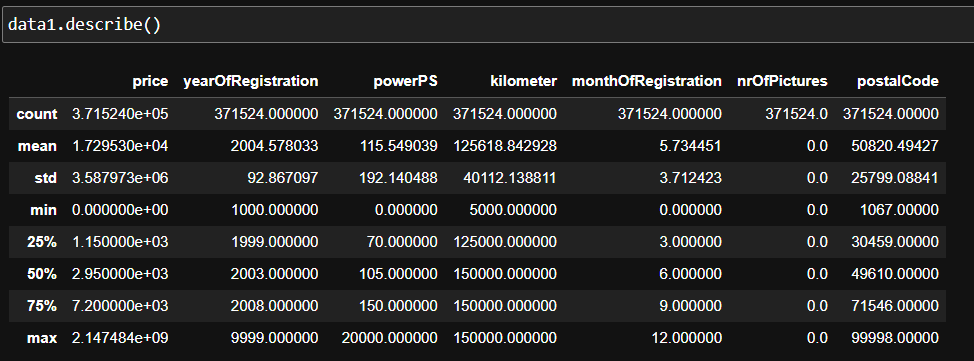
* There will be some different data types in the columns instead of original, for example, the data will be numeric and the type of that data will be in object(string) format. In that case the analysis will be very much affected.
* So, to get rid of this case we can convert the data type into original as shown below



* In the info we see that some columns that contains numeric values, but the data type is object.



* The data type of the columns have been changed.
* We can know some mathematical values like minimum, maximum, mean, etc., by using keyword called “describe()”.
* But, It works only on numeric datatype columns.

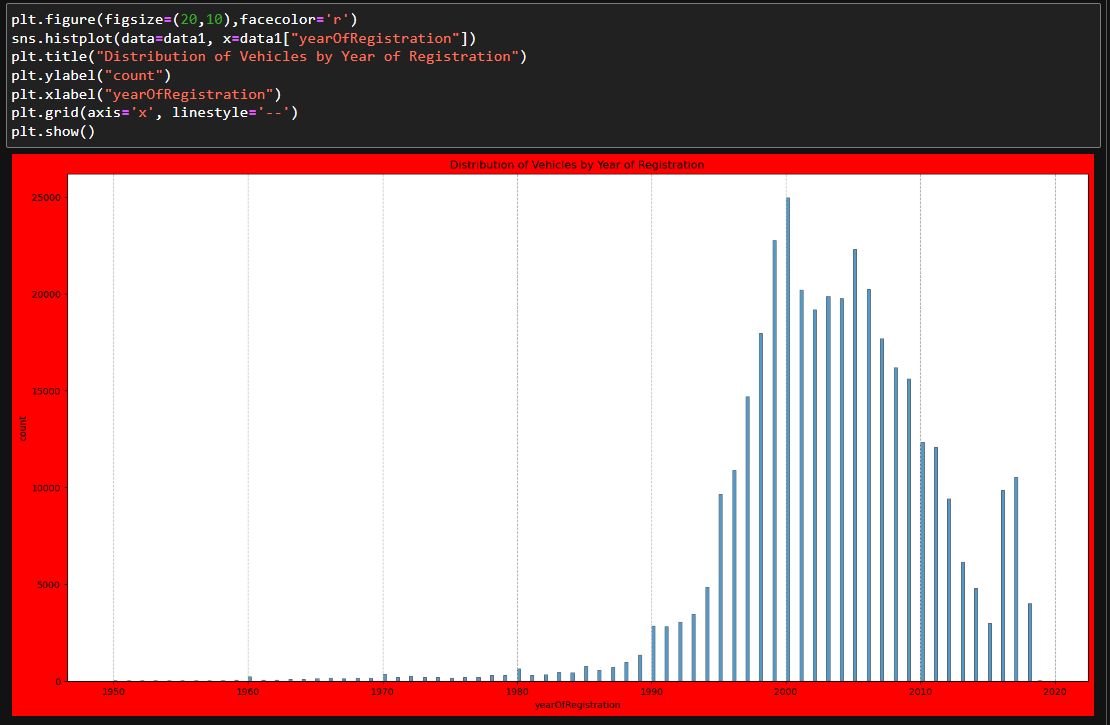


* The above values are the count, mean, standard deviation, min, max, quartiles(25,50,75).
* We have completed knowing the general details about the data now, let us know about the data in deep according to the given problem statements.

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

* To represent any distribution of any plot we can use histogram or bar plot.
* We have to show how many no of vehicles have registered at that particular year so first we have to extract the data of registered year column and fit in a plot.
* But the problem here is there are many years out of bound. So, we have to handle this by ranging the data to a certain value.
* And replacing the other bounded data to to the mode of its data values in the data by using the code.

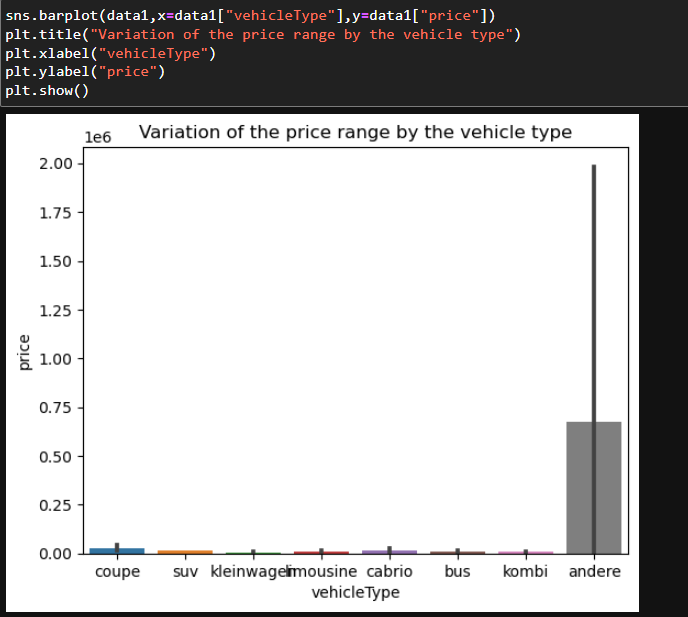
data1["yearOfRegistration"] = data1["yearOfRegistration"].where((data1["yearOfRegistration"] >= 1950) & (data1["yearOfRegistration"] <= 2023),data1["yearOfRegistration"].mode()[0])



* here we replaced the data of the column name year of registration which is not lying between 1950 and 2023 will be replaced with the mode of the data.
* By using the matplotlib we visualized the above code.
* The above image represents that the distribution of vehicles by the year of registration.
* At x-axis we consider the year of registration and on y-axis we consider the count of vehicles at that particular year on bar graph.
* From this distribution we can conclude that a max no of vehicles are registered between the years 1980-2018.
* And there are very less count of vehicles which are registered on that particular years.

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

* To create a plot variation of price differ from one vehicle type to another vehicle we can select the bar plot
* Because in bar plot we can differentiate the count difference from one vehicle to another vehicle type
* Before that we should take the consideration based on vehicle type and price by using the code

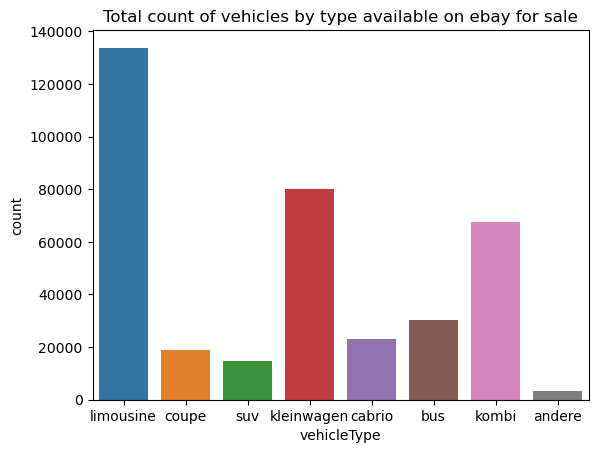


* From the figure we can say that the “andere” vehicle have the highest price of $0.75 \*10^6 and all other vehicles are very least price in the range of $0.12\*10^6-$0.0001\*10^6.
* Kleinwagen vehicle have the least cost price among all the vehicle types

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

To create a plot on the count of vehicles by its type we can use the count plot for the data

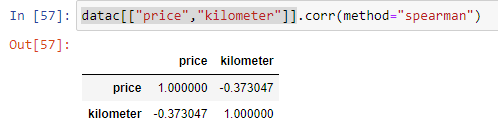
* Because in count plot we can get the count frequency on y-axis and vehicle type on x-axis
* Before that we should take the consideration based on vehicle type by using the code
* Datac[“vehicletype”]
* To plot the data of the count of vehicle type we should use the code as below.



* From the figure we can define the total count of vehicles by its type which are available on ebay.
* From the figure we can say that limousine have the highest count of 138900 cars sold.
* Also, Andere car has least count sold which is around 5000 car.

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

* Yes, there a relation between the column price and kilometer which can be explained as below. To know the relationship between the columns.
* In the statistics there is a concept called correlation, here we can find the relation between two or more columns.



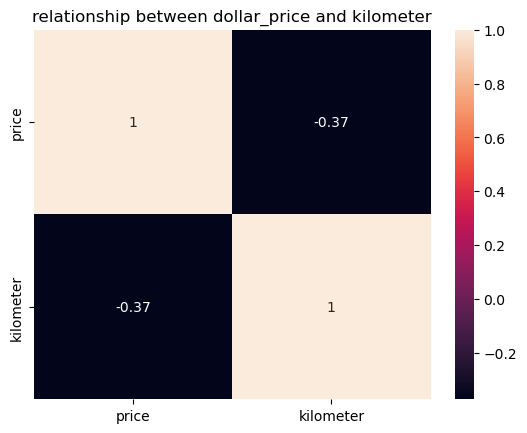
* To represent this information in a plot we can use the heat map which can shows your result in a better effective and understanding
* For that an analysis we have to write the code as,

sns.heatmap(datac[["price","kilometer"]].corr(method="spearman"),annot=True)

plt.title("relationship between dollar\_price and kilometer")

plt.show()

* After running this code this will gives you the plot as,



* From the figure we can define that price and kilometer are negatively moderately correlated
* And the correlation between the price column and kilometer is -0.37
* Why because the correlation always between -1&1
* 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
* From our figure we can say there is an inversely proportional relation between price and kilometer due to its negative correlation.