**Summary**

**Analysis-1:**

1. **General Data Analysis:**

* First, we have to load the data. Copy the data so that original won’t get effected.
* For basic data analysis, mainly we have to understand about our data.
* There are 4 duplicates rows in given data set. We dropped all 4 duplicate rows.
* In the given data we have lots of missing (or) NaN values. Replaced the missing values with appropriate mode value according to that particular column.
* Next, we have to understand the data type of each column. By using data.info () we can see few columns have to be change according to the requirement.
* We have one unwanted column in the data set. So, we have to drop the “nrOfPictures” column which is not useful for analysis.
* 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. **Distribution of Vehicles Based on Year of Registration:**

* The distribution of vehicles based on year of registration provides the insights of no of vehicles that are registered in particular year.
* For visualization, you can create histogram plot (or) count plot to visualize the distribution of vehicles based on the year of registration.
* The x-axis represents the years, and y-axis represents the count of vehicles registered in each year.
* By seeing the visualization, we can say we’re trying to find out how many vehicles are registered from different years.
* 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.
* By the visualization, we can say if there are more old cars or new vehicles are there for the sale.

1. **Variation of Price Range by Vehicle Type:**

* Variation of price range by vehicle type provides the information about differences in prices of different vehicle types.
* You can create histogram plot (or) bar plot to visualize the Variation of price range by vehicle type.
* 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.
* This helps us to see if some vehicle type is more expensive than other vehicle type by comparing them.
* 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 prices have moderate variation when compared with “andere” vehicle type.

1. **Total Count of Vehicles by Type Available:**

* In total count of vehicles by type available, you can find the total count of vehicles for each vehicle type.
* You can create bar plot (or) pie plot to visualize the total count of vehicles by vehicle type.
* The x-axis represents the vehicle type, and y-axis represents the count of vehicles.
* This will help us to see how many kinds of vehicles are available on eBay.
* Vehicle type “limousine” has highest count of 133763 for sale.
* Vehicle type “andere” has lowest count of 3357 for sale.

1. **Relationship Between Dollar Price and Kilometer:**

* 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.
* In correlation, Spearman rank correlation is preferred when the two variables are non-linear. Pearson correlation is suitable when the two variables are linear. It is biased towards linearity.
* So, Spearman rank correlation is better measure than Pearson correlation.
* 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.

* We can use heat map to visualize the relationship between dollar price and

kilometer