Analysis-1 Report

- This dataset contains details of the used cars in Germany which are on sale on ebay.
- This dataset is not clean and hence a lot of data cleaning should be carried out. Then we will follow several steps for cleaning the dataset.

  Those are:

  1. Check the missing values,if any missing values are occure then replace those missing values with appropriate values.
  2. Check the datatype of columns,if datatype of columns are wrong then modify it.
  3. Check any wrong values are placed with in column,then remove those values from Column or replace with suitable value.
  4. Finally check any duplicate values occure in dataset.then remove those records from Dataset if it exists.
  5. Convert cleaned dataset into new csv file and save it in my folder.

- Levels of Measurements for each column.

1. Date trawled column belongs to ordinal data.
2. Name column belongs to nominal data.
3. Seller column belongs to nominal data.
4. Offer Type column belongs to nominal data.
5. Price column belongs to ratio data.
6. Abtest column belongs to nominal data.
7. Vehicle Type column belongs to nominal data.
8. Year of Registration column belongs to ordinal data.
9. Gearbox column belongs to nominal data.
10. PowerPs column belongs to ratio data.
11. Model column belongs to nominal data.
12. Kilometer column belongs to ratio data.
13. Month of Registration column belongs to ordinal data.
14. Fuel type column belongs to nominal data.
15. Brand column belongs to nominal data.
16. Not repaired damage column belongs to nominal data.
17. Date created column belongs to ordinal data.
18. Nrof pictures column belongs to ratio data.
19. Postel code column belongs to nominal data.

20. Last seen column belongs to ordinal data.

Q1) Perform general Data analysis

Performing general data analysis involves several steps.

Those are:

1.Data Collection:

- Obtain the dataset from a reliable source. This could be in the form of a CSV file.

2.Data Cleaning:

- Check for missing values,inconsistencies in the data.
- Handle missing data through imputation or removal.
- Standardize data formats.

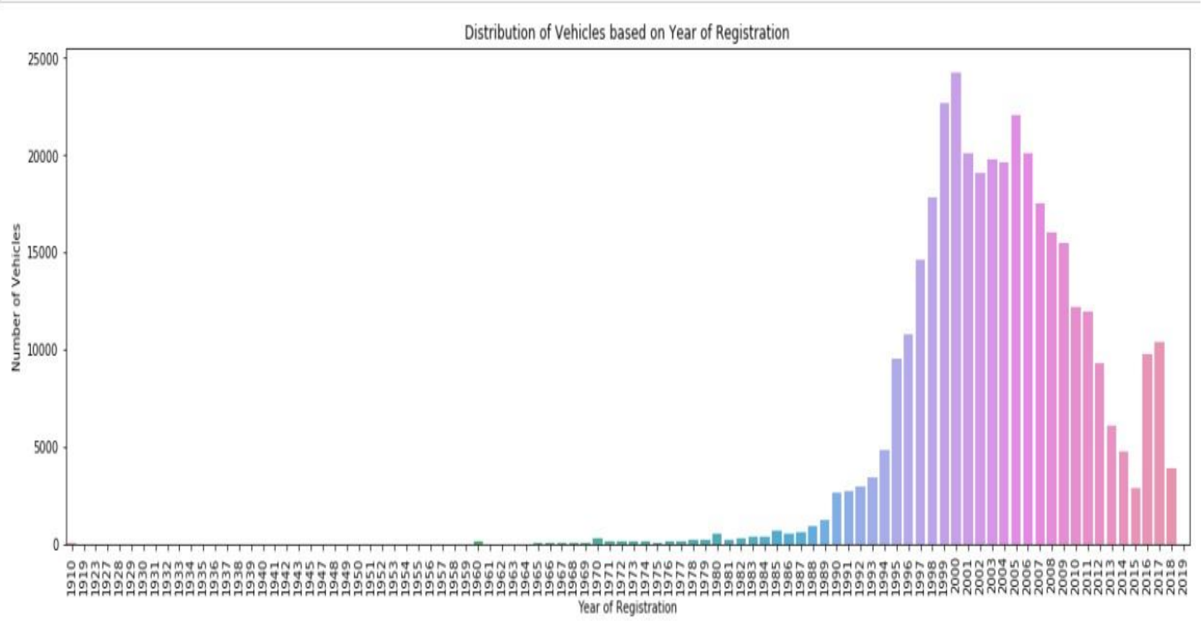
3.Exploratory Data Analysis (EDA):

- Visualize the data using graphs, histograms, box plots, scatter plots etc.,
- to understand the distribution,relationships and patterns.
- Calculate summary statistics (mean, median, standard deviation, etc.) to describe the data.

4.Reporting:

- Communicate the findings effectively, using visualizations, reports, or presentations.

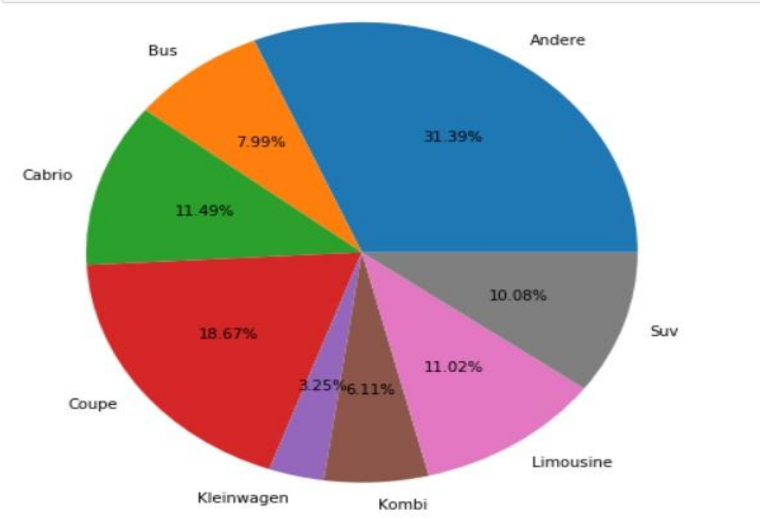
Q2) Can you tell me the Distribution of Vehicles based on Year of Registration with the help of a plot



Summary:

- Based on the plot, the highest car sales were held in the year 2000 and second highest year is 1999.
- The lowest car sales were held in 1910 to 1959 compare to remaining.

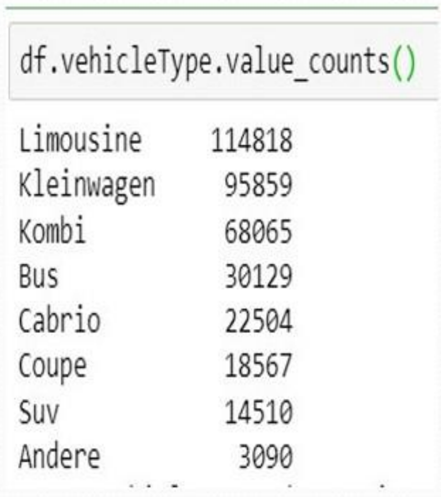
Q3) Create a plot based on the Variation of the price range by the vehicle type



Summary:

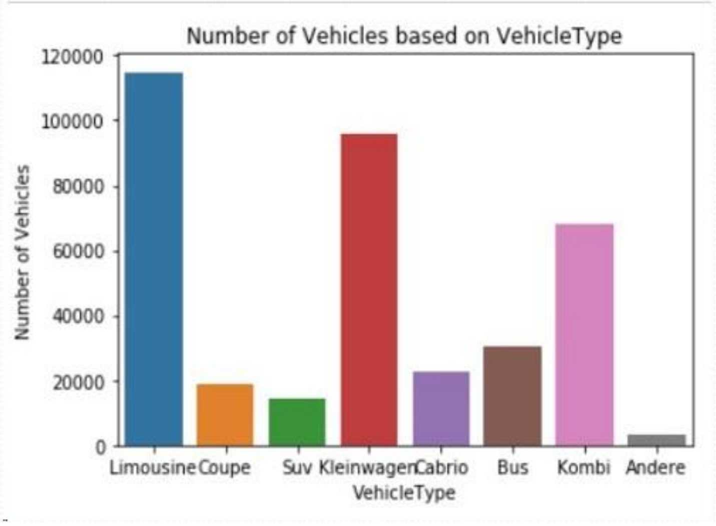
- Based on this plot, Andere vehicle type has the highest price (31.39%) compared to other vehicle types.
- In my opinion Andere vehicle type is more costliest than other types.so many are not willing to buy this type of vehicles because it's price is very high.
- Kleinwagen vehicle type has the lowest price range (3.25%). Many peoples are willing to buy this type of vehicles.

Q4) Find out Total count of vehicles by type available on ebay for sale.As well as create a visualization for the client.



```
df.vehicleType.value_counts()
```

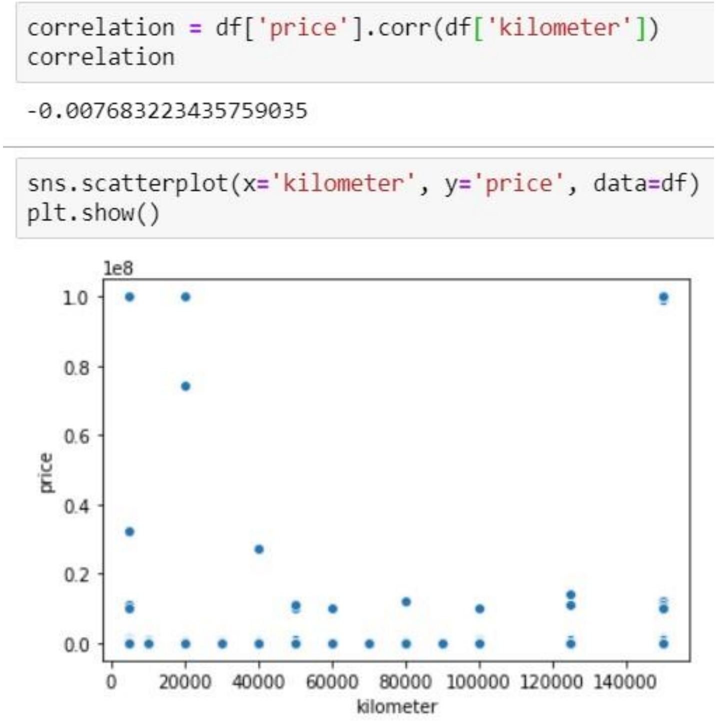
```
Limousine     114818
Kleinwagen     95859
Kombi          68065
Bus            30129
Cabrio         22504
Coupe          18567
Suv            14510
Andere          3090
```



Summary:

- Based on this plot, Limousine vehicle type has the highest available vehicle type (114818) compared to other vehicle types.
- Because these type of vehicle's cost is less compare to andere type and suv type.so many people are willing to buy this type of vehicles.
- So the company mostly manufactures Limousine type of vehicles.
- The company less manufactures andere type of vehicles and suv type of vehicles.

5) Is there any relationship between dollar_price and kilometer? (Explain with appropriate analysis)



```
correlation = df['price'].corr(df['kilometer'])
correlation
```

```
-0.007683223435759035
```

```
sns.scatterplot(x='kilometer', y='price', data=df)
plt.show()
```

Summary:

- I can find the correlation coefficient between 'price' and 'kilometer'. The correlation value ranges from -1 to 1.

- - If the value is close to 1, it indicates a strong positive correlation.

  - If the value is close to -1, it indicates a strong negative correlation.

  - If the value is close to 0, it indicates no correlation.

- The scatter plot visualizes the relationship between the two variables. If the points on the plot show a clear pattern, it suggests a relationship between the 'price' and 'kilometer'.
- finally caluculated correlation coefficient is -0.0076 (approximately),then we will clarify that it indicates no correlation between "price" and "kilometer",because correlation coefficient is close to the 0.