**REPORT**

**prepared by: varshini**

**importing libraries:**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

**Reading csv file:**

🡪data= pd.read\_csv(r"C:\Users\seeth\Downloads\autos.csv",encoding="latin1")

🡪df=data.copy()

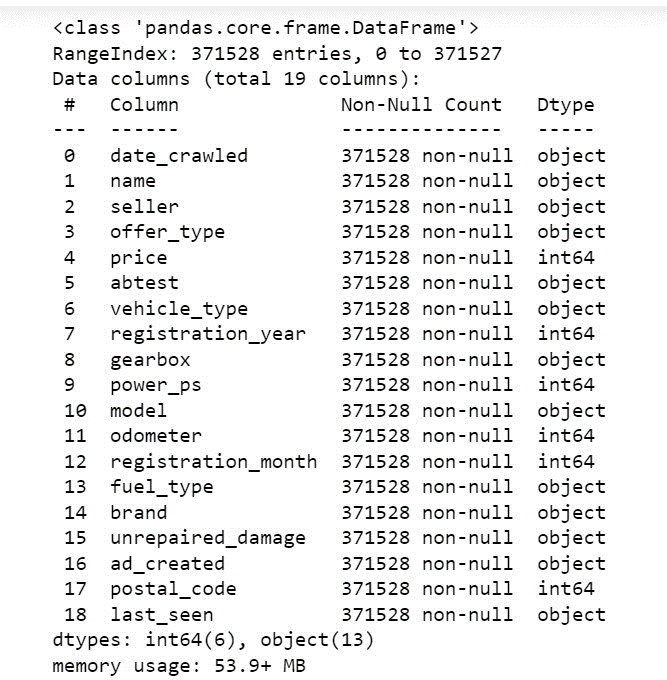
🡪df

**1.Perform general Data analysis.**

**Code:**

df.info()

**Output:**



**Explanation:** In the autos dataset there are 371528 rows and 20 columns and in the above

output there are 20 columns names which are present in the dataset. There is non-null count and the data type of each column. There are 7-int64, 13-object in dtype. The memory usage of autos dataset is 56.7+MB.

**Data Cleaning:** Cleaned the data by changing column names, filling null values with some statistical measures (mean, median, mode) or some null values with “not mentioned” and dropping the unwanted columns in which the original dataset doesn’t affect and dropping

duplicates. Changing the datatype of the columns.

🡪After cleaning saving the dataset into another csv file and performing the analysis on

cleaned dataset.

🡪data=pd.read\_csv(r"C:\Users\seeth\python 232\Statistics\cleaned data of autos.csv")

🡪df=data.copy()

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

**Code:**

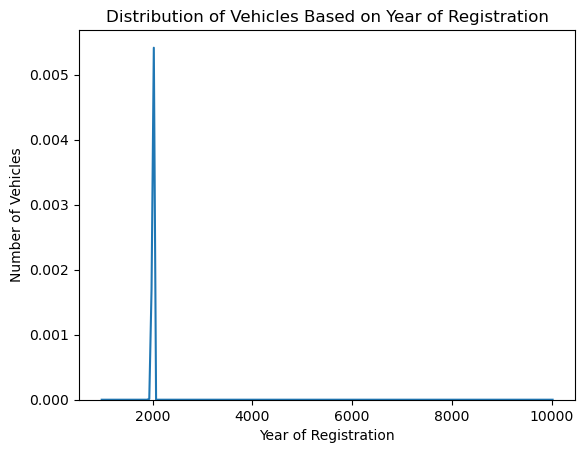
sns.kdeplot(data=df,x=df["registration\_year"])

plt.xlabel("Year of Registration")

plt.ylabel("Number of Vehicles")

plt.title("Distribution of Vehicles Based on Year of Registration")

plt.show()

**Output:** 

**Explanation:** In the above graph, from 1900 to 2020 most of the vehicles got registered.

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

**Code:**

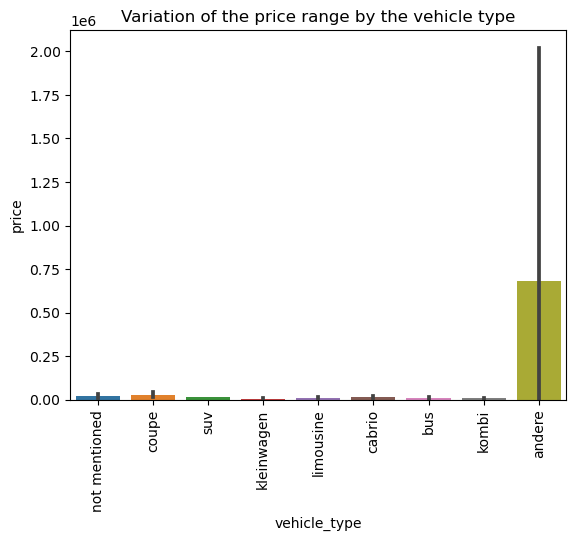
sns.barplot(data=df,x=df["vehicle\_type"],y=df["price"])

plt.xticks(rotation=90)

plt.title("Variation of the price range by the vehicle type")

plt.show()

**Output:**



**Explanation:** The above output shows that there is a high variation of price for andere compared to another vehicles.

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

**Code:**

sns.countplot(data=df,x="vehicle\_type")

plt.xticks(rotation=90)

plt.title("Total count of vehicles by type available on ebay for sale ")

plt.show()

**Output:**

limousine 95508

kleinwagen 79708

kombi 67264

not mentioned 37773

bus 30090

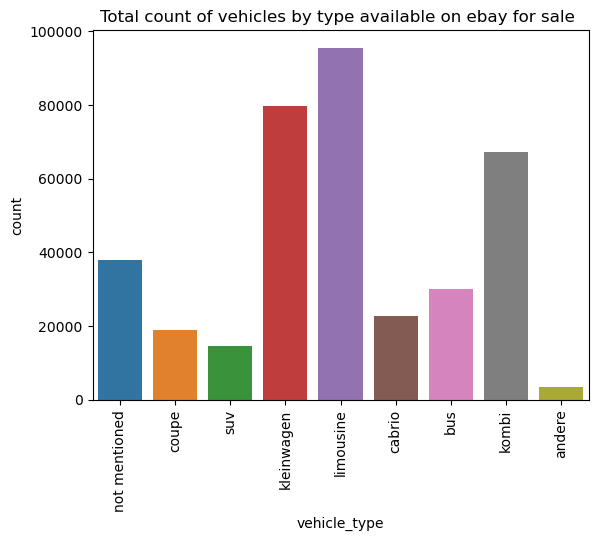
cabrio 22829

coupe 18954

suv 14604

andere 3350

Name: vehicle\_type, dtype: int64



**Explanation:** From the above graph 95508 is the highest count of limousine vehicle when

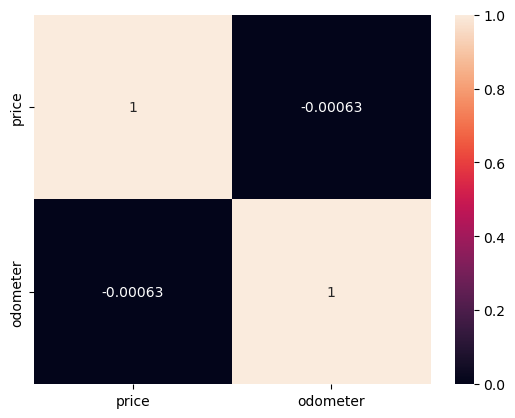
compared with all other vehicles.

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

**Code:**

sns.heatmap(df[["price","odometer"]].corr(method="pearson"),annot=True)

**Output:**



**Explanation:** There is a less relation between price and odometer.