ADVANCED DATA ANALYSIS

ON

GERMAN USED CARS

Name: KADIYALA UDAY

Date: 23-12-2023

Analysis -1

1. General Data Analysis:

- In this analysis, I have imported all the modules that are required for the analysis of given data set. The modules are NumPy, pandas, matplotlib, seaborn.
- Then imported the data set using pandas for the cleaning and analysis process.

1.1 Data Cleaning Steps:

The Data Cleaning steps that are involved in the analysis are:

Missing Values: Filling the missing values lead to the incorrect analysis and interpretations, so dropping all the missing values.
Duplicate Values: There are 4 duplicate records, dropping the duplicate rows for the accurate analysis.
Outlier Detection: To detect the outliers in the data set boxplot is used.
Outlier Treatment: The outliers in the dataset are replaced and outliers in the price column is removed because we can not replace the outliers it may give biased result.
Structural Errors: The structural errors should be removed or replaced. In monthofRegistration column there is 0 which can be treated as structural error because it is considered as object data type. So, 0 is replaced with mode.
Type Casting: Using the appropriate data types for the columns can reduce the memory usage. We have to Choose the smallest possible data type that can represent the data accurately.
Delete Unnecessary columns: Deleting the unnamed columns and nrOfPictures column that are not useful for the analysis.

2. The distribution of vehicles based on year of registration:

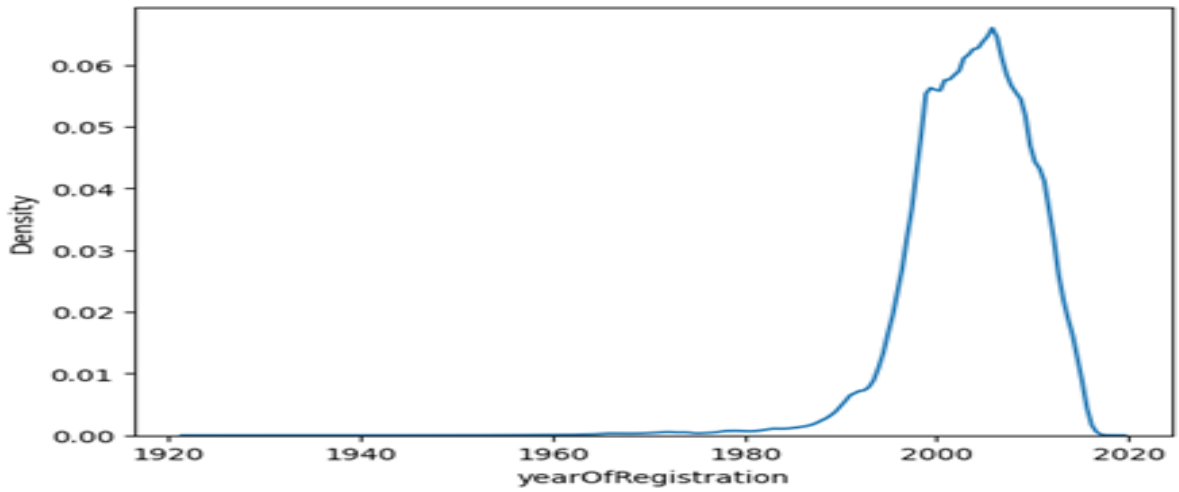


Fig: Left Skewed Distribution

- Density plot is useful for finding the distribution of the vehicles based on the year of registration.
- Year of registration is Left Skewed Distribution because more data points on the right side and the long tail extending to the left side.
- Less vehicles are registered in between the years 1920 and 1980 in Germany.
- Highest number of vehicles are registered in between the years 1980 to 2020.

3. Variation of the price range by the vehicle type:

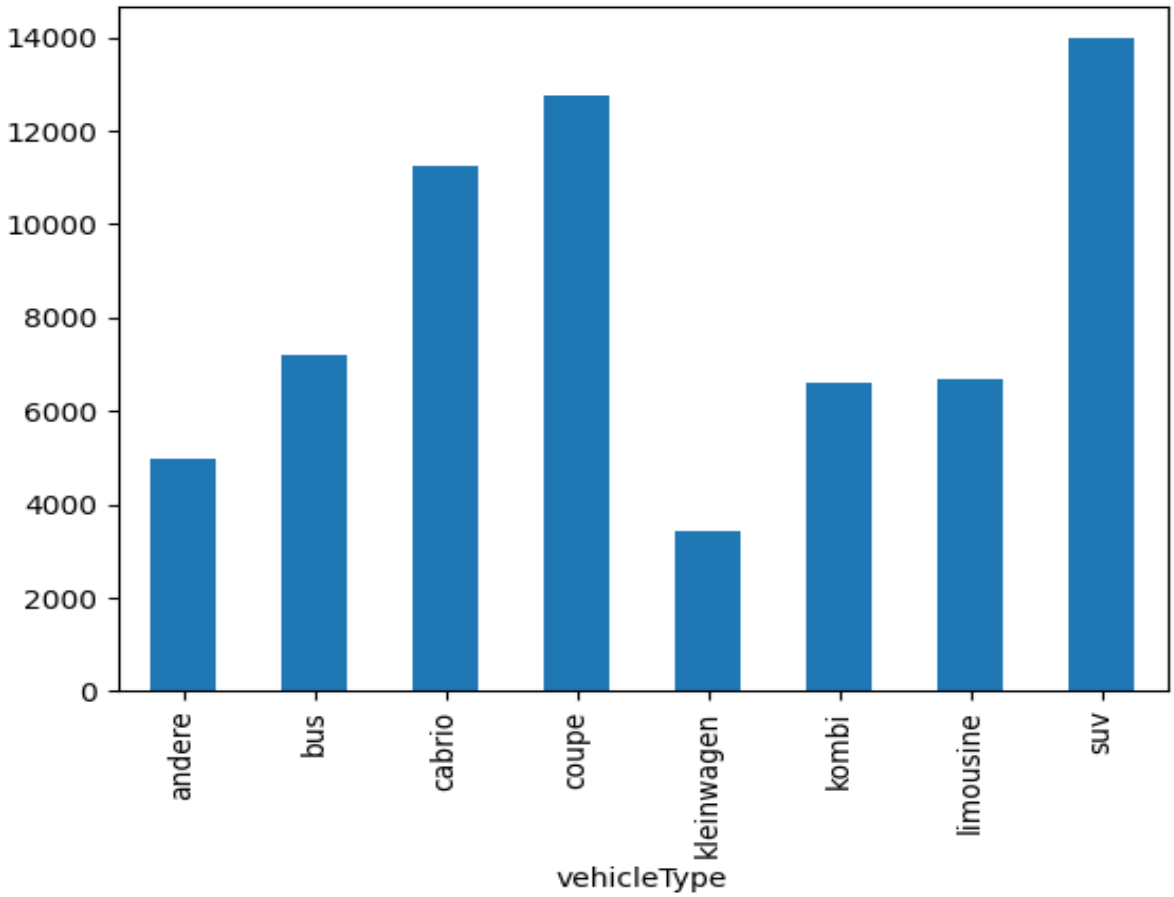


Fig: Variation of the price range by vehicle type

- On an average, the highest price for a vehicle is 14000 for SUV and lowest price for a vehicle is around 3000 for Klein Wagen.
- Coupe is slightly lesser than SUV with an average price of around 12000.

4. Total count of vehicles by type available on eBay for sale:

- As per the data, limousine is the most repeated vehicle type and its count is around 75000.
- Andere(others) are the least repeated vehicle type and its count is around 1600.

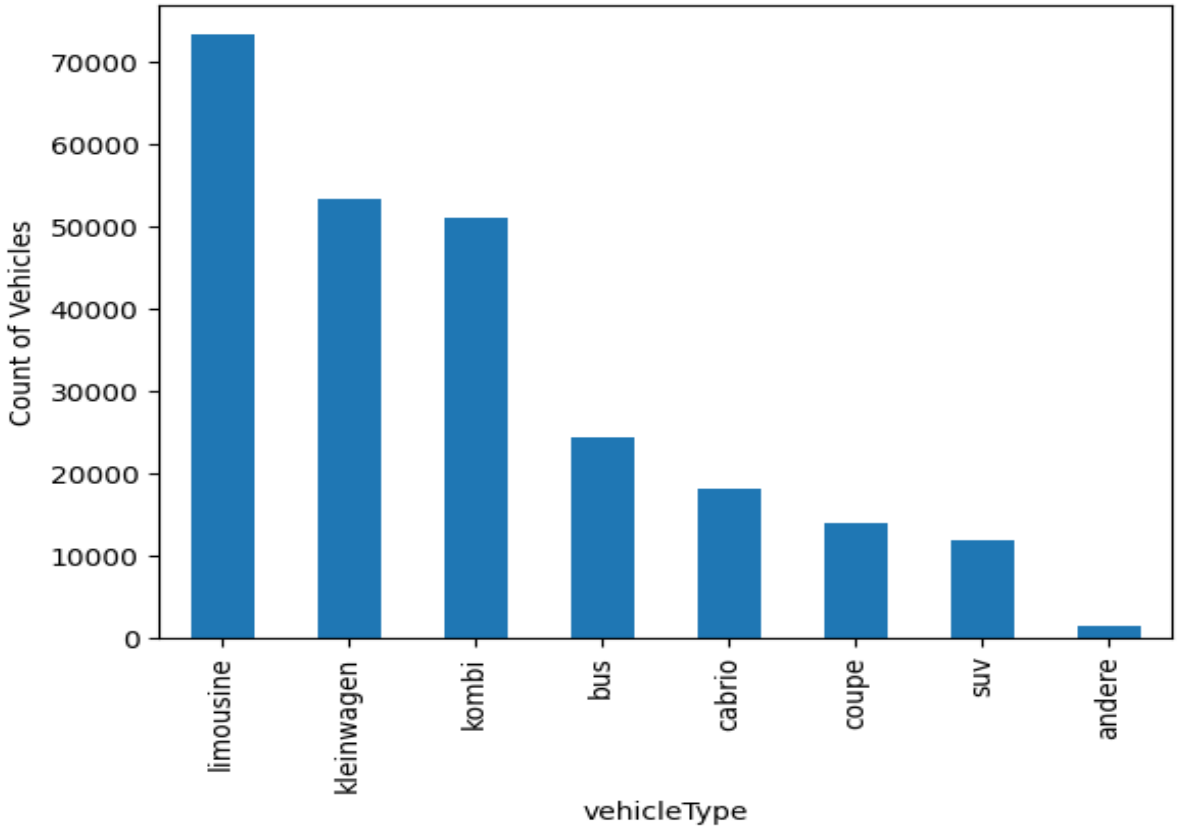


Fig: Total count of vehicles by type available on eBay

5. Relationship between price and kilometre:

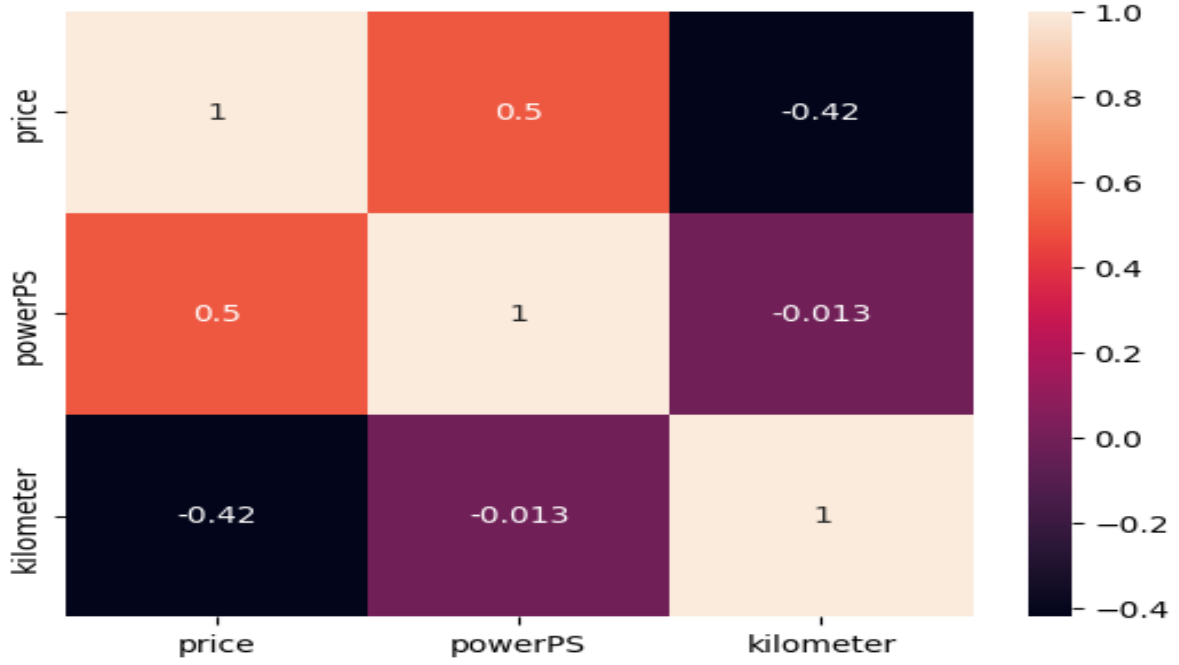


Fig: Heatmap

- According to the given data, the relationship is negative weak relationship.
- Which means if the kilometres that car travelled is increased the price decreases and if the kilometres travelled by the car is less than the price of the car is high.
-

Analysis -2

1. Number of vehicles by brand available on eBay for sale:

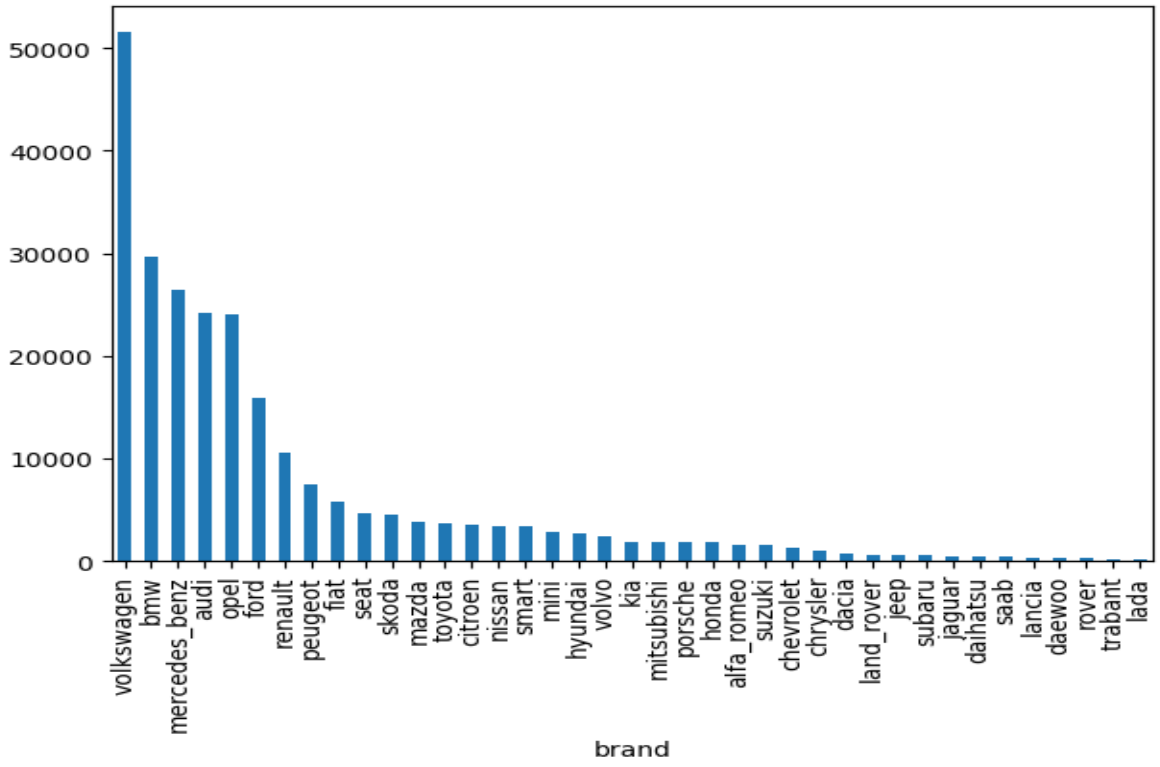


Fig: Bar plot for number of vehicles by brand available

- As per the data, Volkswagen is the most repeated brand and its count is around 55000.
- Volkswagen, BMW, Mercedes-Benzes, Audi, Opel are the top 5 brands in the list which are most repeated.
- Rover, Trabant, Lada are the least repeated vehicle type and its count is less than 300.

2. The Average price for vehicles based on the type of vehicle as well as on the type of gearbox:

- The average price of a vehicle type SUV with gearbox type automatik and manuell is high among all the vehicle types.

- Coupe, cabrio, bus, limousine, kombi, andere and kleinwagen are in the next places respectively.
- The average price of a vehicle type kleinwagen with gearbox type automatik and manuell is low among all the vehicle types.

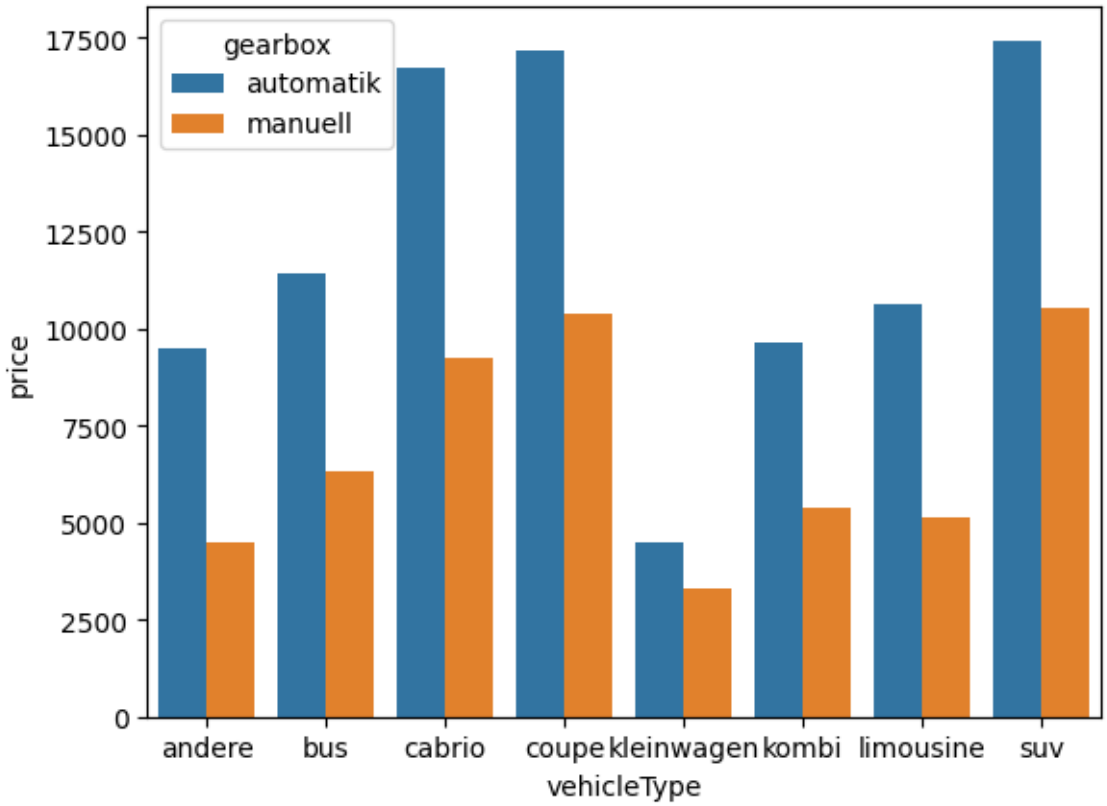


Fig: The Average price for vehicles based on the vehicle type and gearbox

3. The marginal probability of private seller:

- Marginal Probability P(A) = $\frac{\textbf{Number of times private seller occurs}}{\textbf{Total number of outcomes}}$

- P (private seller) = 0.9999919383444584

Analysis -3

1. How can we reduce the memory usage of the data set:

To reduce the memory usage:

1.1 Dropping the Missing Values:

- Dropping the missing values can reduce the memory usage, if we try to fill the missing values that will lead to the incorrect analysis and interpretations.

1.2 Dropping the Duplicate Values:

- Dropping the duplicate values can be useful for the accurate analysis and to remove the redundancy.

1.3 Type Casting:

- Using the appropriate data types for the columns can reduce the memory usage. We have to Choose the smallest possible data type that can represent the data accurately.

1.4 Dropping the Unwanted Columns:

- Deleting the unnamed columns and the columns that are not useful for the analysis can reduce memory size.

1.5 Dealing with the Outliers:

- Replacing the outliers or removing the outliers can reduce the memory usage.

2. The Average price of vehicle by fuel type and gearbox type:

- The fuel type of the vehicle, which has an electro engine and an automatik gearbox, is very expensive compared to other vehicles.
- The fuel type of the vehicle, which has hybrid engine and an manuell gearbox, is very expensive compared to other vehicles.
- The fuel type lpg with gearbox type automatik or manuell vehicle price is less compared to others.

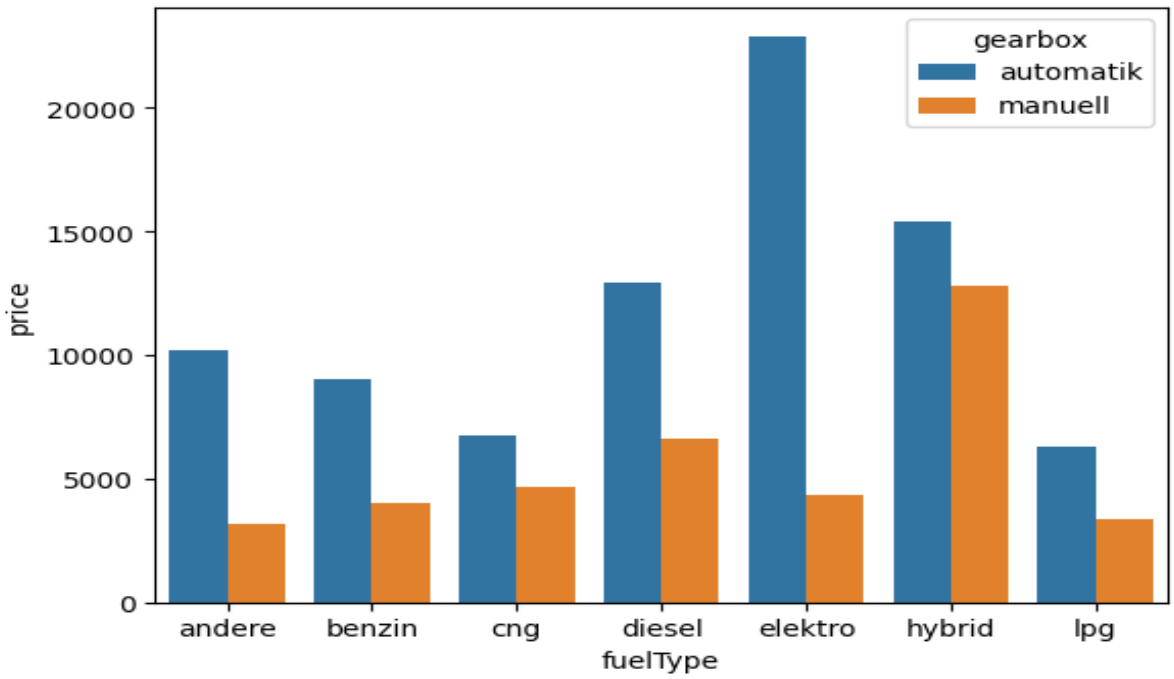
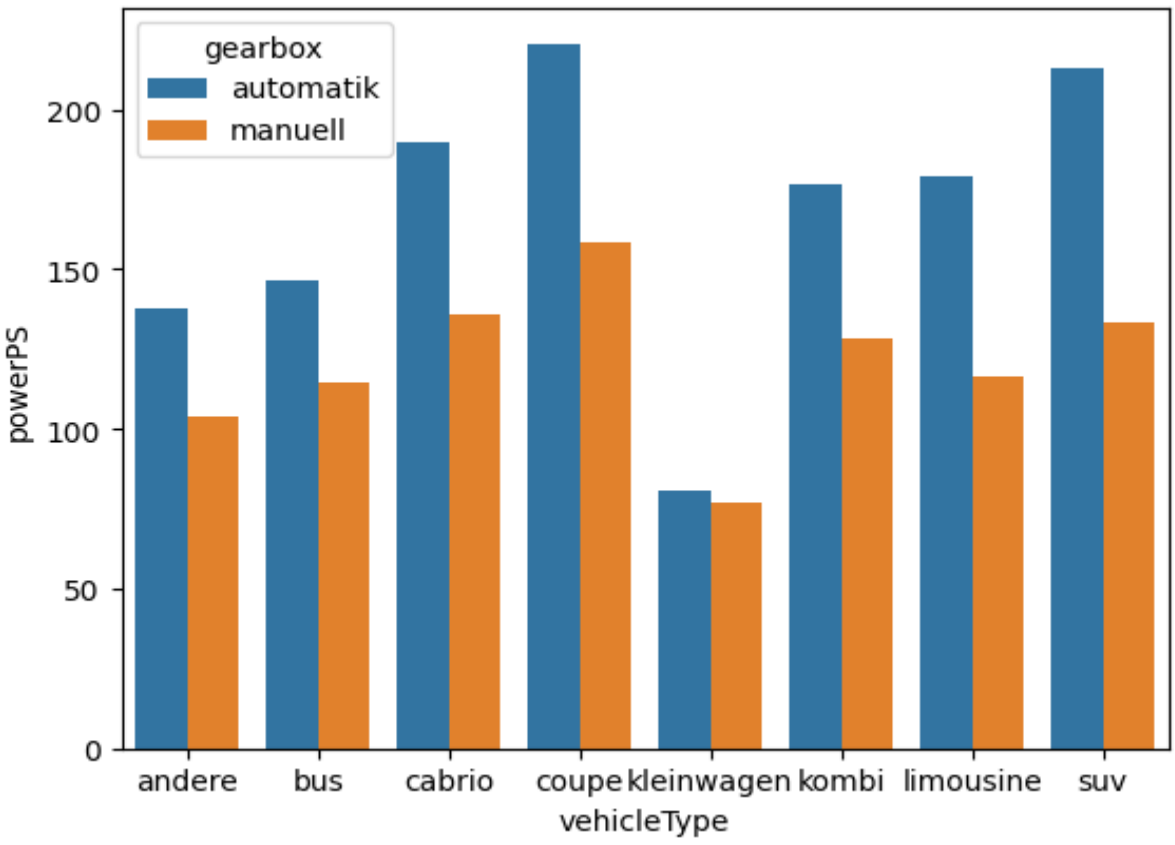


Fig: Average price of vehicle by fuel type and gearbox type

3. The average power of a vehicle by gear box type and vehicle type:

- The vehicle coupe with gearbox type automatik and manuell have an average power greater than 200 and around 150 which is greater than other vehicles.



4. The average price of a vehicle by brand as well as vehicle type. Use heatmap to explain this:

- All the blanks are referred to null values because some brands do not have some types of vehicles like land rover does not have vehicle type, bus.

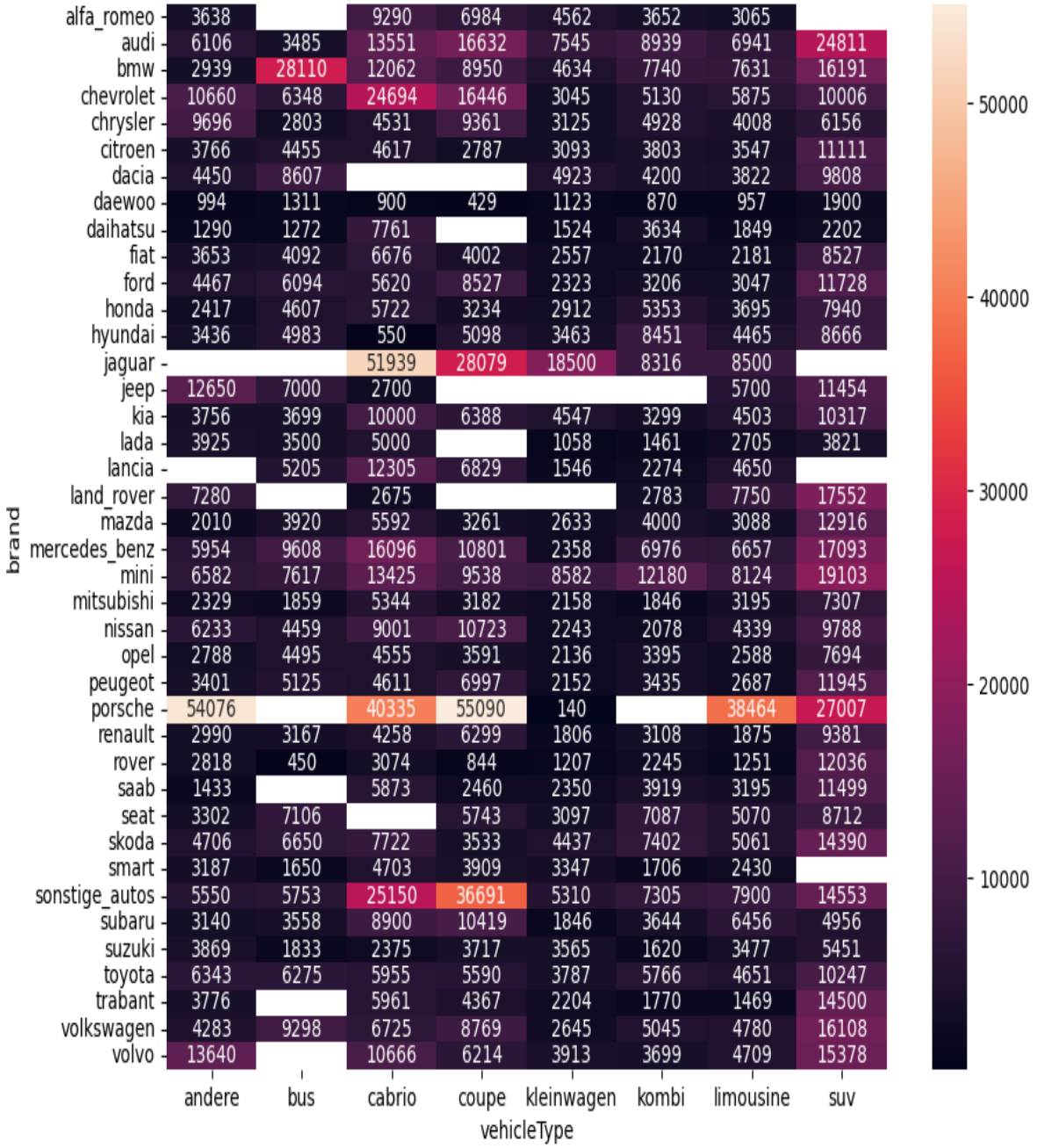


Fig: Heatmap