Assignment Overview

Visulaizzare statistiche basate su un dataset di auto usate.

Assignment Deliverable

Assignment Background

The goal of our project is to analyse the availability of used cars in the market.

We want to help the user to understand the

* Analysing whether the most convenient are diesel or benzine-used cars
* Compare the range of prices for different brand
* Identify the most frequent models for each brand
* Analysing the distribution of used cars based on the registration year and calculating the numbers of vehicles available for each range.

Steps to Do Before Running the Code

Install libraries

DATA SET DESCRIPTION:

This dataset is scraped from Ebay. The content of the dataset is in German, but it should not impose critical issues in understanding the data. The fields included in the dataset are as following:

\* dateCrawled: when this ad was first crawled, all field-values are taken from this date

\* name: ”name” of the car

\* seller: private or dealer

\* offerType

\* price: the price in euro on the ad to sell the car

\* abtest

\* vehicleType

\* yearOfRegistration : at which year the car was first registered

\* gearbox

\* powerPS: power of the car in PS

\* model

\* kilometer: how many kilometers the car has driven

\* monthOfRegistration: at which month the car was first registered

\* fuelType: vehicle fuel type

\* brand: vehicle brand

\* notRepairedDamage: if the car has a damage which is not repaired yet

\* dateCreated: the date for which the ad at ebay was created

\* nrOfPictures: number of pictures in the ad

\* postalCode

\* lastSeenOnline: when the crawler saw this ad last online

The goal of this project is to use Python and Jupyter notebook to explore, analyze and visualize the *Used Cars* dataset. To solve the assignment we applied the knowledge gained from the theoretical and practical exercises. We used Seaborn, Matplotlib, and Pandas visualization libraries.

Step 1:

In the Used Cars dataset, we identified the missing and invalid values for the columns: `vehicle type`, `price`, `brand`, and `month of registration`. We standardized the information and converted it to a unique value. We identified for each column the number of missing or invalid instances. The prices are in euros and the range of accepted prices is between €1000 and €100000.

Once we identified missing/invalid values for the given columns, we removed all rows where one or more columns had invalid/missing data.

Step 2:

We wanted to analyze for a given type of vehicle whether the price of diesel is greater than the one of benzine. We created a histogram that provides a representation of the average price of the used cars based on fuel types.

We can see on the plot that the price of diesel cars is higher than that of benzine.

Step 3:

We created a box plot to represent the distribution of prices for the following brands: *mercedes\_benz*, *fiat*, *volvo*, *alfa\_romeo* and *lancia*. The box plot showed that the price of the Mercedes is the highest followed by Volvo, Alfa Romeo, Lancia and Fiat.

Step 4:

We created a tabular representation of price ranges, per brand, year of registration (only consider the interval 1960-2020 and create bins), and powerful/not powerful (above/below median powerPS of the entire dataset). The table shows the different brands as rows, and powerful/not powerful, years of registration as hierarchical columns. The value of a cell represents the range of prices for the cars with the brand identified by the row, interval of registration year, and powerful/not powerful identified by the column. Then we created a dataframe with the 3 most frequent models for each brand.

Step 5:

For the price range 0 – 100’000 and the year of registration 1960 – 2020, we splitted the price range into 10 bins and for each bin we showed the distribution of the year of registration with a box plot.

Step 6:

We splitted the kilometer in 6 bins and created a dataframe where the columns are such bins, and the rows are the different vehicle types. A cell represents the number of cars with the kilometer in the range identified by the column, and the type of vehicle identified by the row. We create the data frame using groupby.