Saudi Used Cars Dataset Analysis

Khalid Alqahtani

Abstract

The goal of this project was to EDA (Exploration Data Analysis) approach to figure this data out and discover valuable information. The used car dataset is downloaded from the popular website (Kaggle) and I used various charts to get many paramount insights. In addition, I handled missing data by using four different techniques.

Design

As mentioned previously the data was collected from Kaggle under name "Saudi Arabia Used Cars Dataset", which is about used car for sale in Saudi Arabia. I applied EDA approaches trying to discover this kind of data by showing a variety of charts like bar charts and pie chart. Moreover, I applied "mean", "median", "mode", and KNN (K-Nearest-Neighbor) to handle missing data. I follow these step for missing data imputation:

1) Removing all rows in the column price which equal zeros
2) Choosing a car type with highest number of rows.
3) Making about 25% of the price column as zero value.
4) Imputing these missing data with the four techniques.
5) Evaluating the techniques

Data

The dataset contains 8035 records with 13 features, 4 of which are numerical. The features are: car brand called in the dataset 'Make', car's name as 'Type', Year, country origin of the car as 'Origin', Color, car option such as standard option, semi-full option, or full option as 'Options', Engine_Size, fuel type gas or diesel as 'Fuel_Type', gear type manual or automatic as 'Gear_Type', Mileage, the city or region where the car is currently there as 'Region', 'Price', and 'Negotiable' is True when the owner want to negotiate the price with the buyer, and False if the owner do not.

Algorithms

Feature Engineering

I appended a new column to the dataset called “Mileage_Scale”, to scale a mileage number like the following:

- (0-50,000) represented as ‘0’
- (50,000 – 100,000) represented as ‘1’
- And so on, until if the mileage number greater than 500,000 Km will be represented as 11. The reason for adding this column to consider it as a feature for missing data methods.

EDA Approach

I used different charts which are bar chart, pie chart, and tree map chart for visualizing and better understanding of the data.

Missing Data Imputation Techniques

I applied four methods: mean, median, mode, and KNN.

Tools

- Numpy and Pandas for data manipulation
- Matplotlib and Squarify for plotting
- Sklearn for imputers and matrices

Communication

All the results and insights will be embedded on the slides.