Project description

You're an analyst at Crankshaft List. Hundreds of free advertisements for vehicles are published on your site every day. You need to study data collected over the last few years and determine which factors influence the price of a vehicle.

Instructions for completing the project

Step 1. Open the data file and study the general information

File path: /datasets/vehicles_us.csv. Download dataset

Step 2. Data preprocessing

— Identify and study missing values:

- In some cases there's an obvious way to replace missing values. For instance, if a Boolean field contains only `True` values, it's reasonable to assume that the missing values are `False`. There aren't such obvious fixes for other data types, and there are cases when the fact that a value is missing is significant. In such instances, don't fill in the values.
- When appropriate, do fill in the values. Explain why you chose to do so and how you selected the replacement values.
- Describe the factors that may have resulted in missing values.

— Convert the data to the required types:

- Indicate the columns where the data types need to be changed and explain why.

Step 3. Calculate and add to the table the following:

- Day of the week, month, and year the ad was placed
- The vehicle's age (in years) when the ad was placed
- The vehicle's average mileage per year

In the `condition` column, replace string values with a numeric scale:

- new = 5
- like new = 4
- excellent = 3
- good = 2
- fair = 1
- salvage = 0

Step 4. Carry out exploratory data analysis, following the instructions below:

- Study the following parameters: price, vehicle's age when the ad was placed, mileage, number of cylinders, and condition. Plot histograms for each of these parameters. Study how outliers affect the form and readability of the histograms.
- Determine the upper limits of outliers, remove the outliers and store them in a separate DataFrame, and continue your work with the filtered data.
- Use the filtered data to plot new histograms. Compare them with the earlier histograms (the ones that included outliers). Draw conclusions for each histogram.
- Study how many days advertisements were displayed (`days_listed`). Plot a histogram. Calculate the mean and median. Describe the typical lifetime of an ad. Determine when ads were removed quickly, and when they were listed for an abnormally long time.
- Analyze the number of ads and the average price for each type of vehicle. Plot a graph showing the dependence of the number of ads on the vehicle type. Select the two types with the greatest number of ads.
- What factors impact the price most? Take each of the popular types you detected at the previous stage and study whether the price depends on age, mileage, condition, transmission type, and color. For categorical variables (transmission type and color), plot box-and-whisker charts, and create scatterplots for the rest. When analyzing categorical variables, note that the categories must have at least 50 ads; otherwise, their parameters won't be valid for analysis.

Step 5. Write an overall conclusion

Format: Complete the task in a Jupyter notebook. Put your code in the code cells and your text explanations in markdown cells, then apply formatting and headings.

Description of the data

The dataset contains the following fields:

- `price`
- `model_year`
- `model`
- `condition`
- `cylinders`
- `fuel` — gas, diesel, etc.
- `odometer` — the vehicle's mileage when the ad was published
- `transmission`
- `paint_color`
- `is_4wd` — whether the vehicle has 4-wheel drive (Boolean type)
- `date_posted` — the date the ad was published
- `days_listed` — from publication to removal