Course › Exploring Data › Lab › Visualizing Data for Regression

Visualizing Data for Regression

Visualizing Data for Regression

There are two goals for data exploration and visualization. First to understand the relationships between the data columns. Second to identify features that may be useful for predicting labels in machine learning projects. Additionally, redundant, colinear features can be identified. Thus, visualization for data exploration is an essential data science skill. This process is also known as **exploratory data analysis**.

In this lab, your first goal is to explore a dataset that includes information about automobile pricing. In other labs you will use what you learn through visualization to create a solution that predicts the price of an automobile based on its characteristics. This type of predictive modeling, in which you attempt to predict a real numeric value, is known as **regression**; and it will be discussed in more detail later in the course. For now, the focus of this lab is on visually exploring the data to determine which features may be useful in predicting automobile prices.

By the completion of this lab, you will:

1. Use summary statistics to understand the basics of a data set.
2. Use several types of plots to display distributions.
3. Create scatter plots with different transparency.
4. Use density plots and hex bin plots to overcome overplotting.
5. Apply aesthetics to project additional dimensions of categorical and numeric variables onto a 2d plot surface.
6. Create pair-wise scatter plots and conditioned plots to create displays with multiple axes.

Lab Steps

1. Make sure that you have completed the setup requirements as described in the Lab Overview section.
2. Now, run jupyter notebook and open the “VisualizingDataForRegression.ipynb” notebook under Module 2 folder.
3. Examine the notebook and answer the questions along the way.

Question 1

1.0/1.0 point (graded)

What can you conclude from the plotted histograms?

○ There are more cars with high miles per gallon than low miles per gallon

○ Most of the cars have larger than 3000 curb weight

○ There are only a few cars with engine size lower than 150

● Most of the cars cost less than 30,000 ✓

Submit   You have used 2 of 2 attempts

Question 2

1.0/1.0 point (graded)

From the kde plots, which feature shows the closest resemblance, in terms of distribution, to price?

○ curb_weight

● engine_size ✓

○ city_mpg

Submit You have used 2 of 2 attempts

Question 3

1.0/1.0 point (graded)

Select three relationships that are now apparent in the scatter plots:

☑ Both gas and diesel turbo cars are generally more expensive than standard cars.

☑ Turbo cars appear to have worse city_mpg at a given price point than standard cars.

□ Standard cars generally has diesel engine.

☑ Turbo cars have greater horsepower at a given price point.

✓

Submit You have used 2 of 2 attempts

Question 4

1.0/1.0 point (graded)

Select three relationships that are now apparent in the conditioned plots:

☑ The distribution of the values generally increases for length and curb_weight for real wheel drive (rwd) cars, with the values for 4 wheel drive (4wd) and real wheel drive (rwd) overlapping.

□ Generally, 4wd cars have the highest engine size.

☑ Cars with fwd have the highest city_mpg, whereas, 4wd and rwd in a similar range.

☑ Generally, 4wd cars have the lowest price, with rwd cars having the widest range.

✓

Submit You have used 2 of 2 attempts

Question 5

1.0/1.0 point (graded)

Which combination produces blank plots?

○ fwd and convertible

○ 4wd and hatchback

○ fwd and hardtop

● 4wd and convertible ✓

Submit You have used 2 of 2 attempts

Learn About Verified Certificates