Bucketing

Let's start with a quick review of a key idea from Machine Learning Crash Course (/machine-learning/crash-course/representation/cleaning-data). Look at the distribution in the chart below.

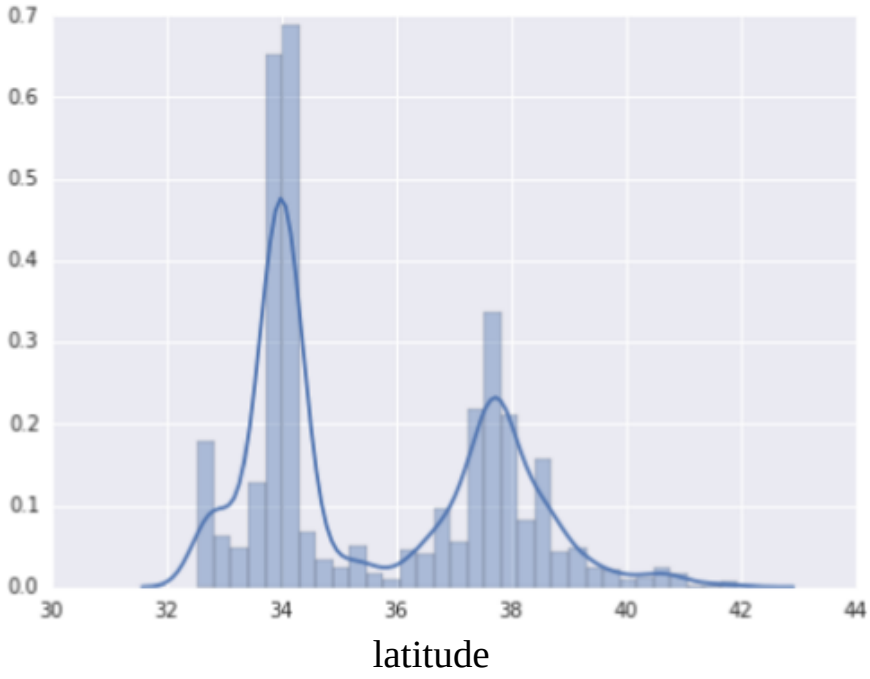

Figure 1: House prices versus latitude.

For the following question, click the desired arrow to check your answer:

Consider Figure 1. If you think latitude might be a good predictor of housing values, should you leave latitude as a floating-point value? Why or why not? (Assume this is a linear model.)

No — there's no linear relationship between latitude and the housing values.

Yes — if latitude is a floating-point value in the dataset, you shouldn't change it.

In cases like the latitude example, you need to divide the latitudes into buckets to learn something different about housing values for each bucket. This transformation of numeric features into categorical features, using a set of thresholds, is called **bucketing**

(/machine-learning/glossary#bucketing) (or binning). In this bucketing example, the boundaries are equally spaced.

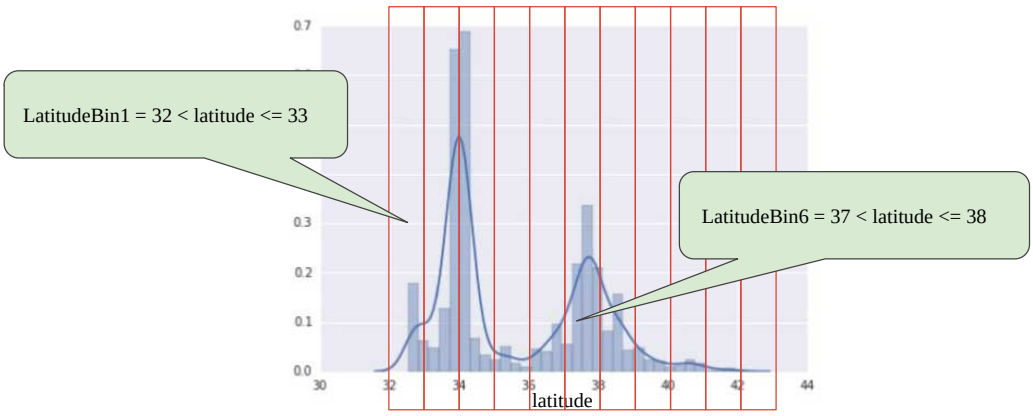

Figure 2: House prices versus latitude, now divided into buckets.

Quantile Bucketing

Let's revisit our car price dataset with buckets added. With one feature per bucket, the model uses as much capacity for a single example in the >45000 range as for all the examples in the 5000-10000 range. This seems wasteful. How might we improve this situation?

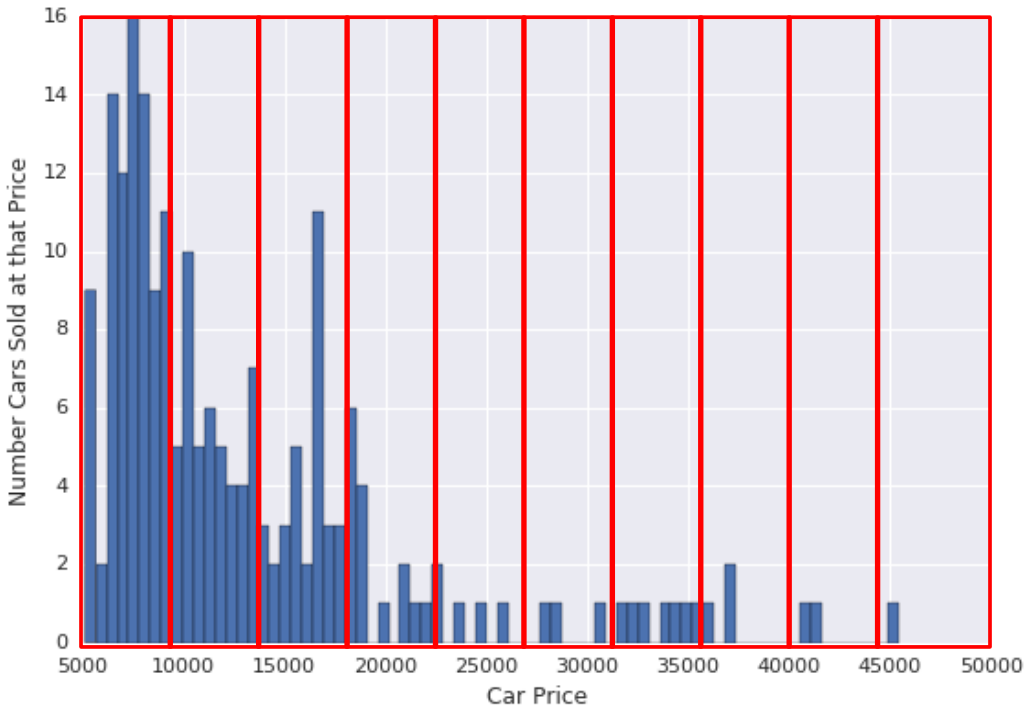

Figure 3: Number of cars sold at different prices.

The problem is that equally spaced buckets don't capture this distribution well. The solution lies in creating buckets that each have the same number of points. This technique is called **quantile bucketing** (/machine-learning/glossary#quantile_bucketing). For example, the following figure divides car prices into quantile buckets. In order to get the same number of examples in each bucket, some of the buckets encompass a narrow price span while others encompass a very wide price span.

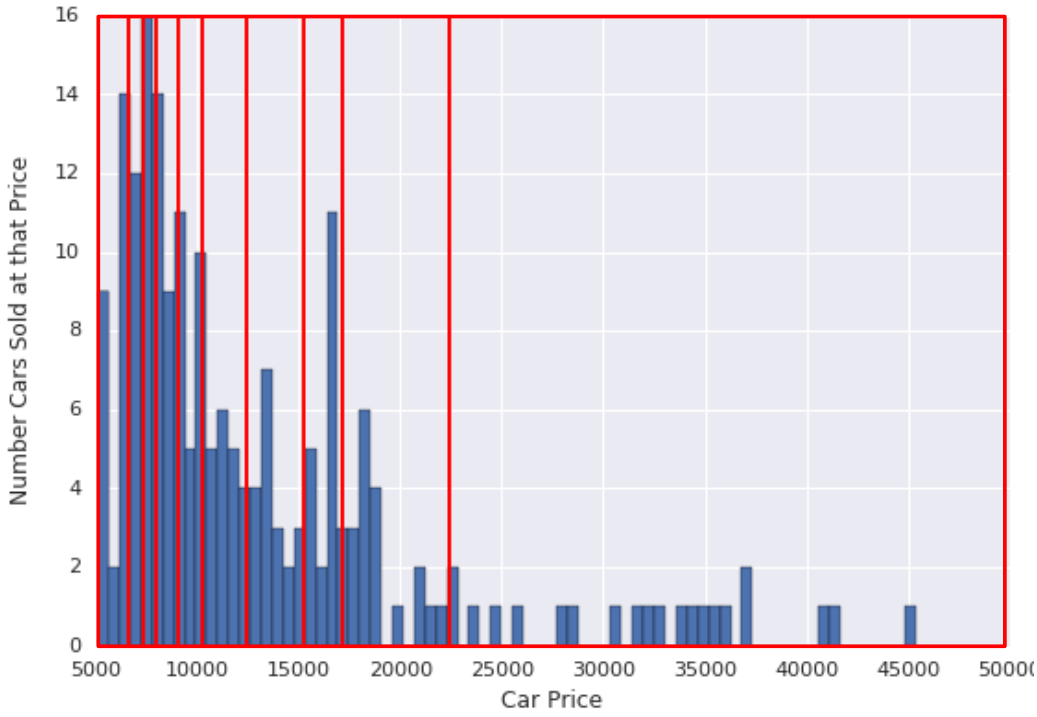

Figure 4: Quantile bucketing gives each bucket about the same number of cars.

Bucketing Summary

If you choose to bucketize your numerical features, be clear about how you are setting the boundaries and which type of bucketing you're applying:

- **Buckets with equally spaced boundaries**: the boundaries are fixed and encompass the same range (for example, 0-4 degrees, 5-9 degrees, and 10-14 degrees, or $5,000-$9,999, $10,000-$14,999, and $15,000-$19,999). Some buckets could contain many points, while others could have few or none.
- **Buckets with quantile boundaries**: each bucket has the same number of points. The boundaries are not fixed and could encompass a narrow or wide span of values.

ting with equally spaced boundaries is an easy method that works for a lot of data distributions. For skew however, try bucketing with quantile bucketing.

rms:

cketing (/machine-learning/glossary#bucketing)

Previous
← Normalization (/machine-learning/data-prep/transform/normalization)

Next
Transforming Categorical Data →
(/machine-learning/data-prep/transform/transform-categorical)