Python Assignment: Understand Binning

Peng Wang

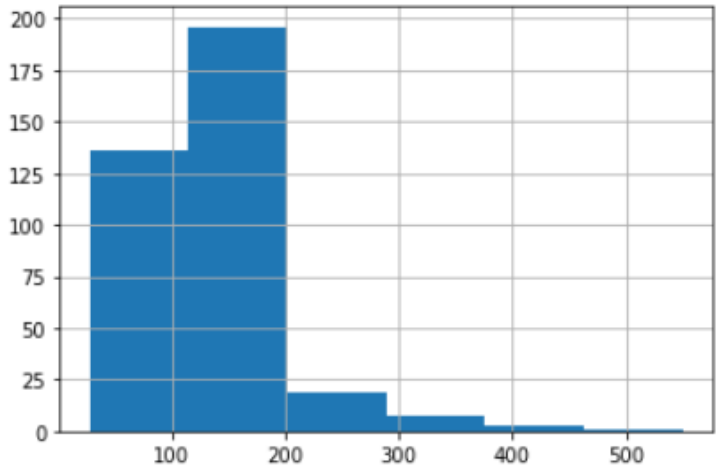
June 19, 2021

I have a dataset that contains a feature names' LoanAmount'. The values are across from 28 to 550, inclusive.

I visualized this feature with the histogram, having parameter bins=6:

```
data['LoanAmount'].hist(bins=6)
```

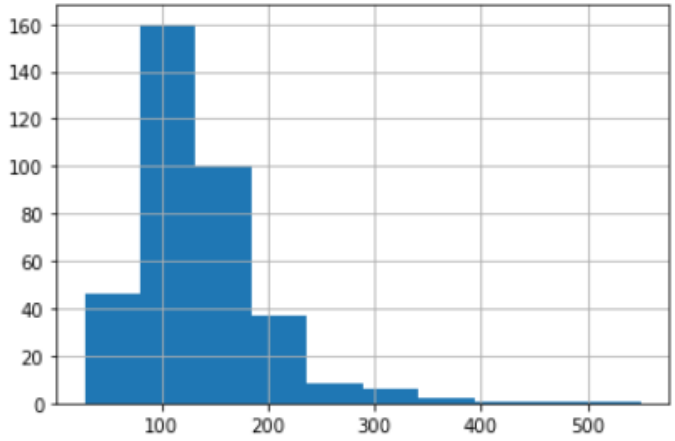
```
<AxesSubplot:>
```



Then with bins=10:

```
data['LoanAmount'].hist(bins=10)
```

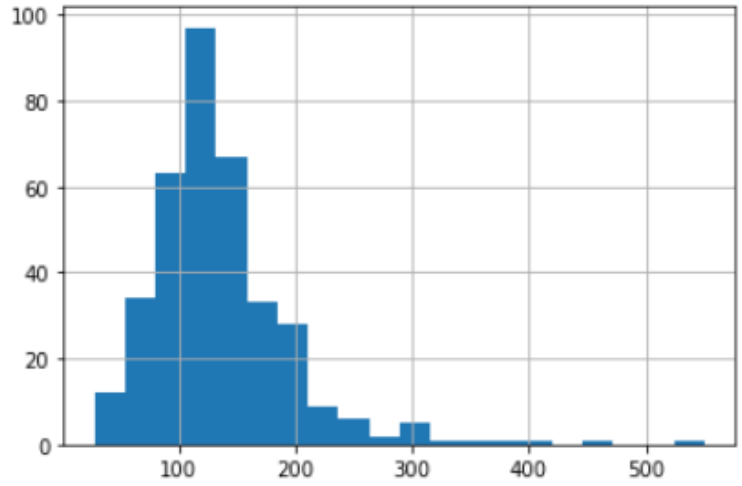
```
<AxesSubplot:>
```



Or, bins = 20:

```
data['LoanAmount'].hist(bins=20)
```

```
<AxesSubplot:>
```



So, what is the Bin?

The definition of Bin, or Binning

Data Binning is the process that segments continuous values by dividing them into several segments, and each plays the same role as a category. The method of converting continuous values into discrete values is called Binning.

For example, the scores of the class:

- below 60 points segment as failing,
- those between 60 and 70 segments as good,
- those between 70 and 85 segments as better, and
- those between 85 and 100 segments as excellent.

Why Binning Important

It is common to use linear regression and logistic regression to segment the continuous features. When building a classification model, discretizing the continuous variables will make the model more stable and avoid overfitting.

For example, a dataset has the age of 20-30 years old. Assume that 5 bins between 20 and 30 and the next time If there are features that are over 30 years old during training, directly add a column of 0-1 to indicate whether the user is over 30 years old or not. If age>30, value is 1, otherwise 0. Without this segmentation, if there's an abnormal "age 300 years" will cause great impact on the model.

How to Process Bin

Supervised Binning

Chi-square bin

The low chi-square value indicates that two adjacent intervals have similar class distributions.

Adjacent intervals with the smallest chi-square value are merged until meeting a particular stopping criterion. The relative class frequencies should be utterly consistent and if two adjacent intervals have very similar class distributions, these two intervals can be merged; otherwise, they should be kept separate.

Unsupervised binning

Isometric Bin

From the minimum to the maximum, it is divided into N equal parts, so if A and B are the minimum and maximum, then the length of each interval is W=(B−A)/N, and the interval boundary value is A +W,A+2W,….A+(N−1)W.

Equal frequency Bin

The boundary-value of the interval must be selected so that each interval contains approximately the same number of instances. For example, N=10, each interval should have about 10% of instances.

The drawbacks of the above two algorithms: for example,

- equal-width interval division, divided into five intervals, and the maximum salary is 50,000, then all people with wages less than 10,000 are divided into the same interval.
- The equal frequency interval may be just the opposite. All people with a salary higher than 50,000 will be divided into 50,000 intervals. Both of these two algorithms ignore the type of instance, and the chance of falling in the correct interval is excellent.

Parameters Reference in Matplotlib.pyplot.hist

binsint or sequence or str, default: `rcParams["hist.bins"]` (default: `10`)

If *bins* is an integer, it defines the number of equal-width bins in the range.

If *bins* is a sequence, it defines the bin edges, including the left edge of the first bin and the right edge of the last bin; in this case, bins may be unequally spaced. All but the last (righthand-most) bin is half-open. In other words, if *bins* is:

```
[1, 2, 3, 4]
```

then the first bin is `[1, 2)` (including 1, but excluding 2) and the second `[2, 3)`. The last bin, however, is `[3, 4]`, which *includes* 4.

If *bins* is a string, it is one of the binning strategies supported by `numpy.histogram_bin_edges`: 'auto', 'fd', 'doane', 'scott', 'stone', 'rice', 'sturges', or 'sqrt'.

References:

https://www.huaweicloud.com/articles/c5eb0bb01a4bc79e02a2fe53003bf03f.html

https://zhuanlan.zhihu.com/p/57509701

https://zhuanlan.zhihu.com/p/68865422

https://www.zhihu.com/question/31989952/answer/54184582

https://www.jianshu.com/p/0805f185ecdf

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html

https://stackoverflow.com/questions/43005462/pandas-bar-plot-with-binned-range