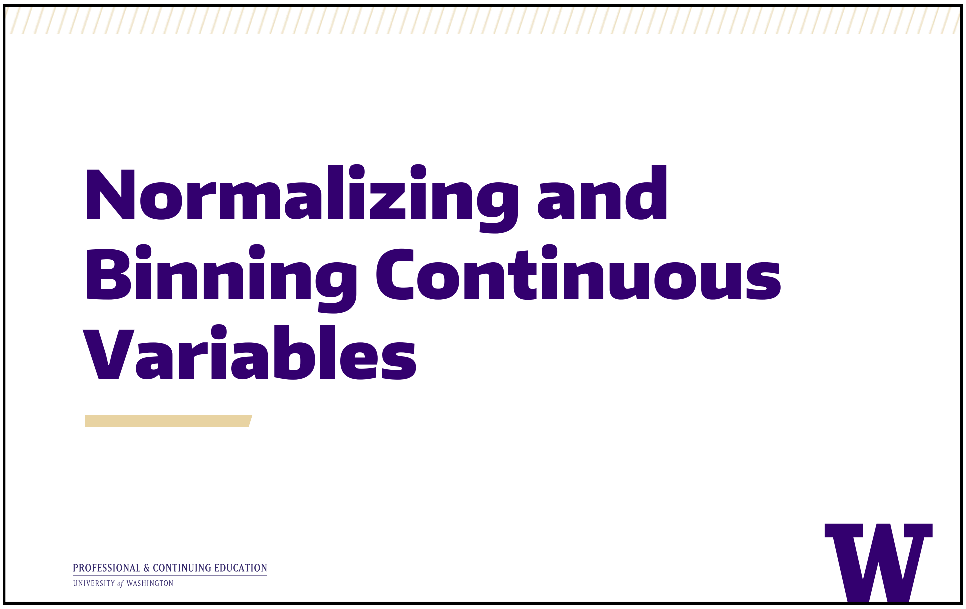
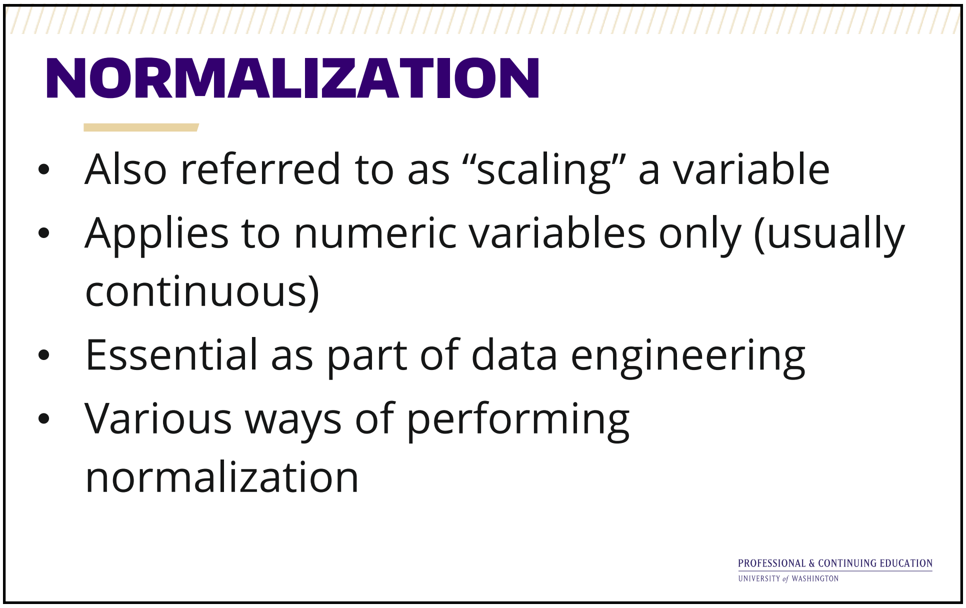


Normalizing and Binning Continuous Variables

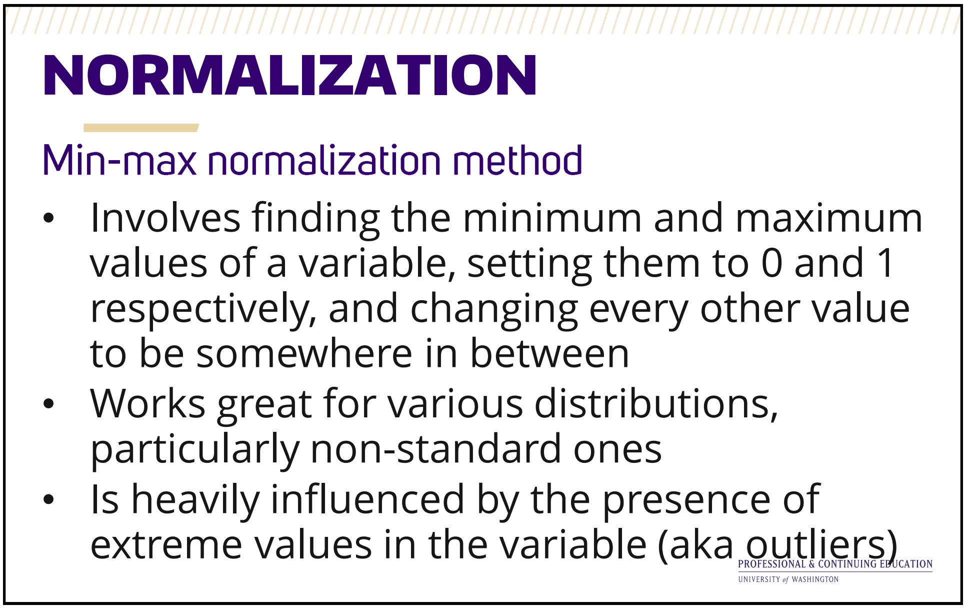
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



NORMALIZATION

- Also referred to as “scaling” a variable
- Applies to numeric variables only (usually continuous)
- Essential as part of data engineering
- Various ways of performing normalization

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

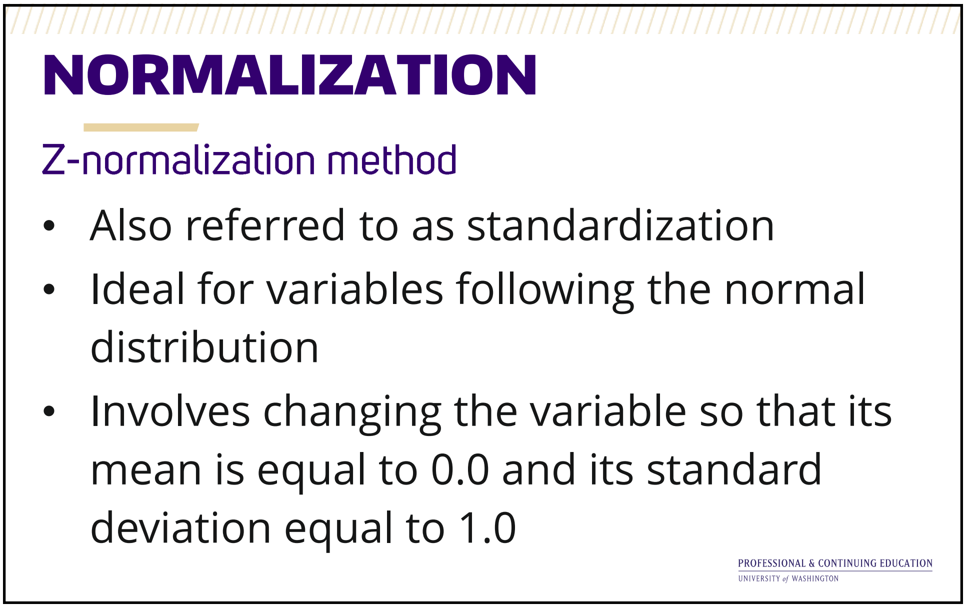


NORMALIZATION

Min-max normalization method

- Involves finding the minimum and maximum values of a variable, setting them to 0 and 1 respectively, and changing every other value to be somewhere in between
- Works great for various distributions, particularly non-standard ones
- Is heavily influenced by the presence of extreme values in the variable (aka outliers)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

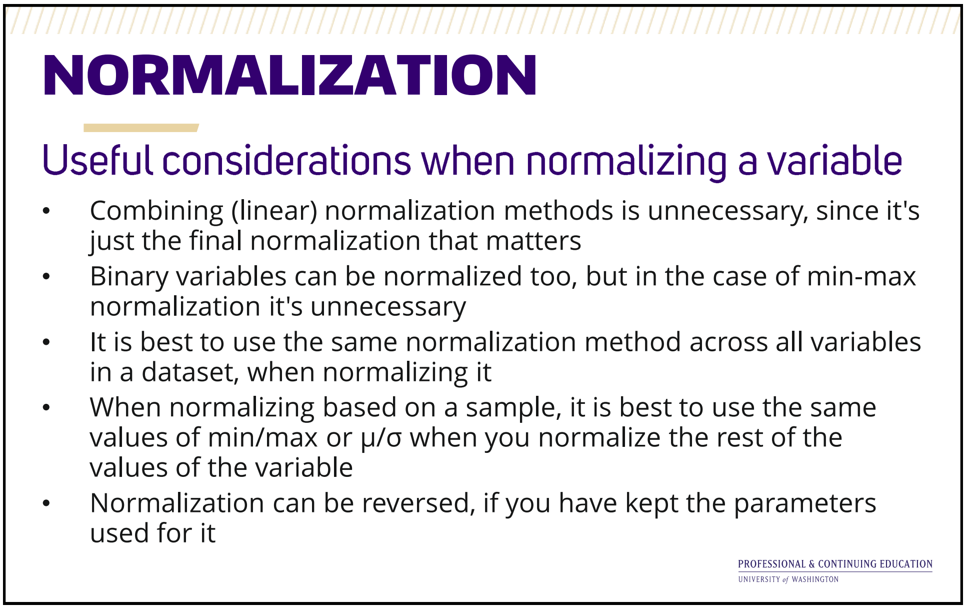


NORMALIZATION

Z-normalization method

- Also referred to as standardization
- Ideal for variables following the normal distribution
- Involves changing the variable so that its mean is equal to 0.0 and its standard deviation equal to 1.0

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

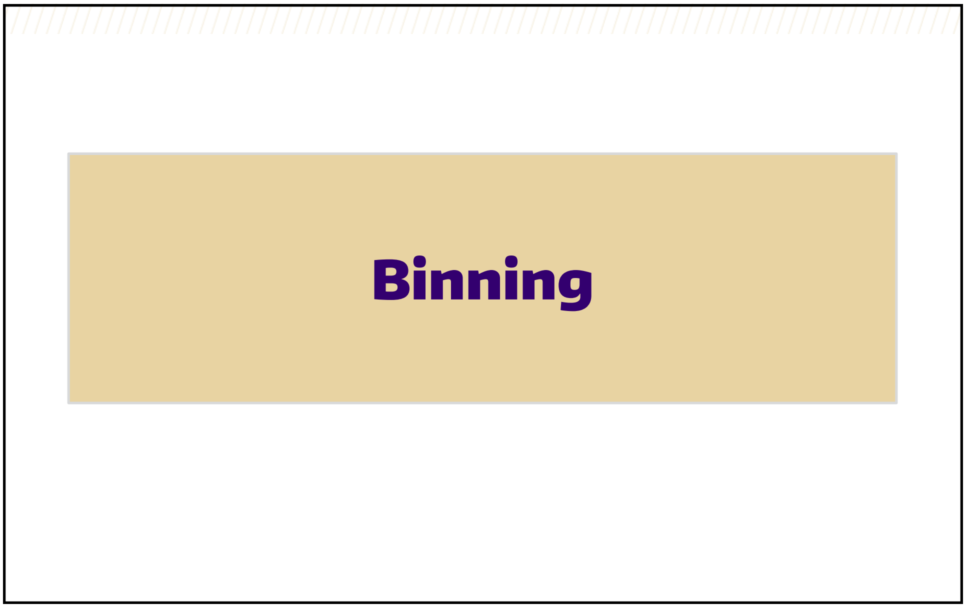


NORMALIZATION

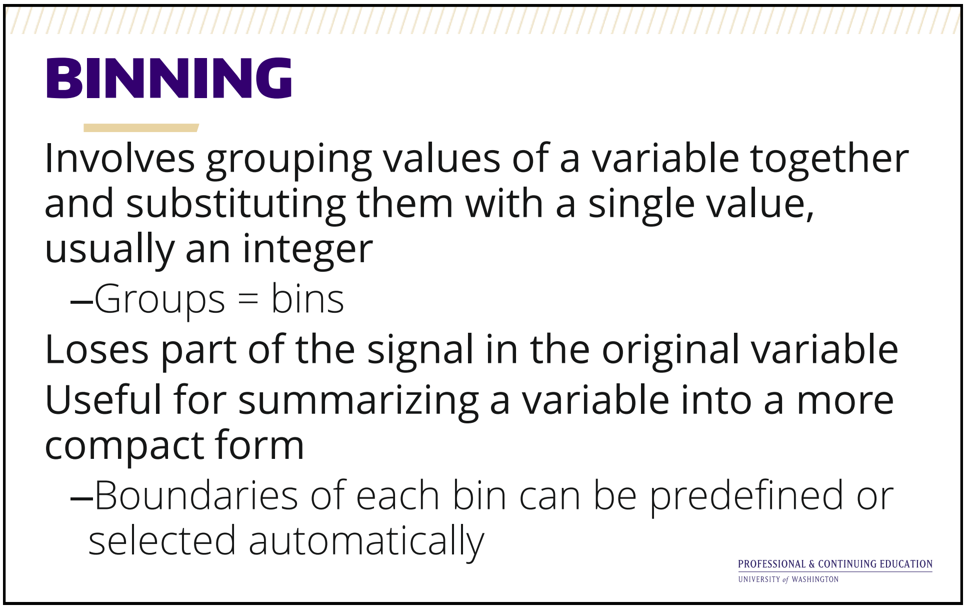
Useful considerations when normalizing a variable

- Combining (linear) normalization methods is unnecessary, since it's just the final normalization that matters
- Binary variables can be normalized too, but in the case of min-max normalization it's unnecessary
- It is best to use the same normalization method across all variables in a dataset, when normalizing it
- When normalizing based on a sample, it is best to use the same values of min/max or μ/σ when you normalize the rest of the values of the variable
- Normalization can be reversed, if you have kept the parameters used for it

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



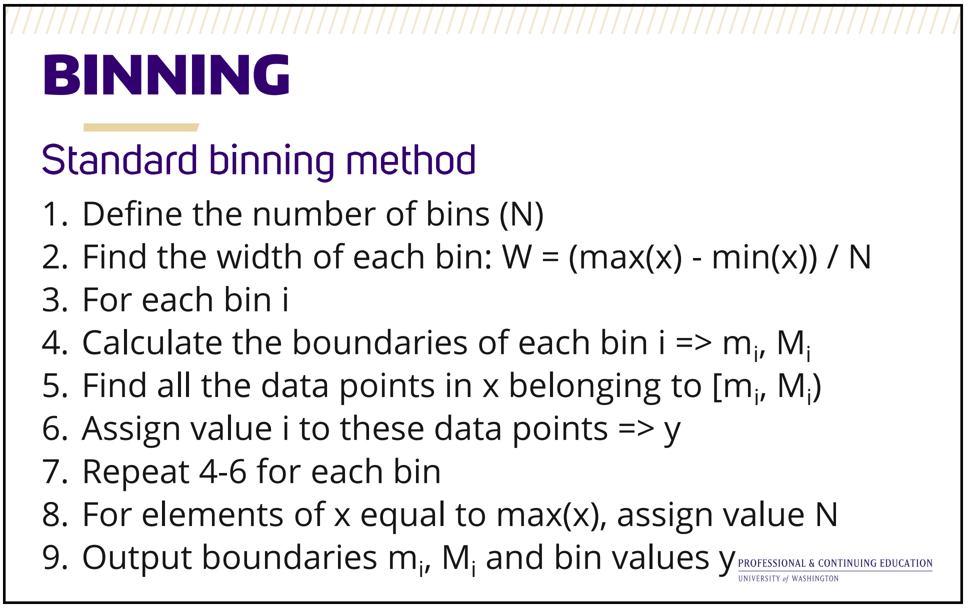
Binning



BINNING

- Involves grouping values of a variable together and substituting them with a single value, usually an integer
  - Groups = bins
- Loses part of the signal in the original variable
- Useful for summarizing a variable into a more compact form
  - Boundaries of each bin can be predefined or selected automatically

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

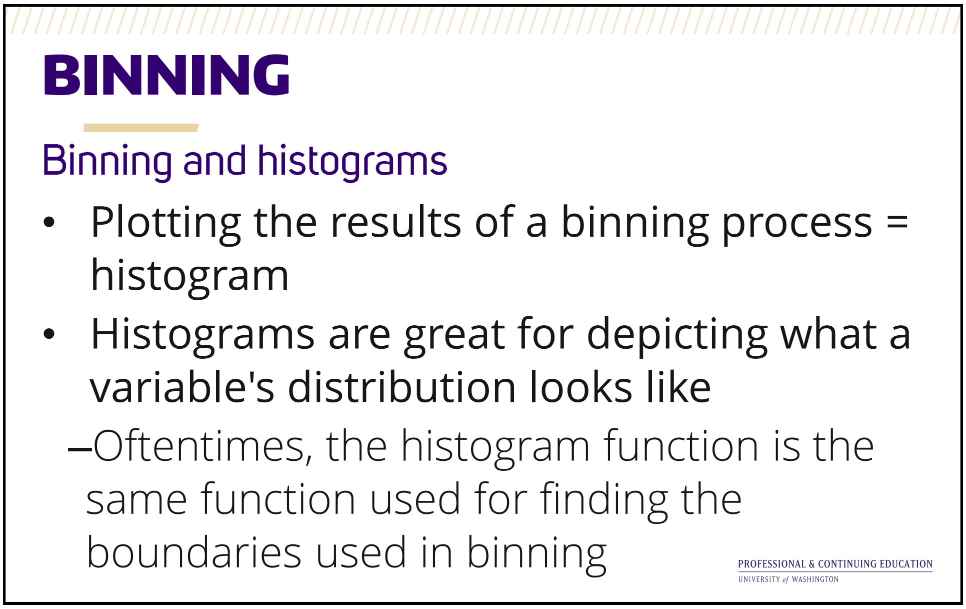


BINNING

Standard binning method

1. Define the number of bins (N)
2. Find the width of each bin: W = (max(x) - min(x)) / N
3. For each bin i
4. Calculate the boundaries of each bin i => m_i, M_i
5. Find all the data points in x belonging to [m_i, M_i)
6. Assign value i to these data points => y
7. Repeat 4-6 for each bin
8. For elements of x equal to max(x), assign value N
9. Output boundaries m_i, M_i and bin values y

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

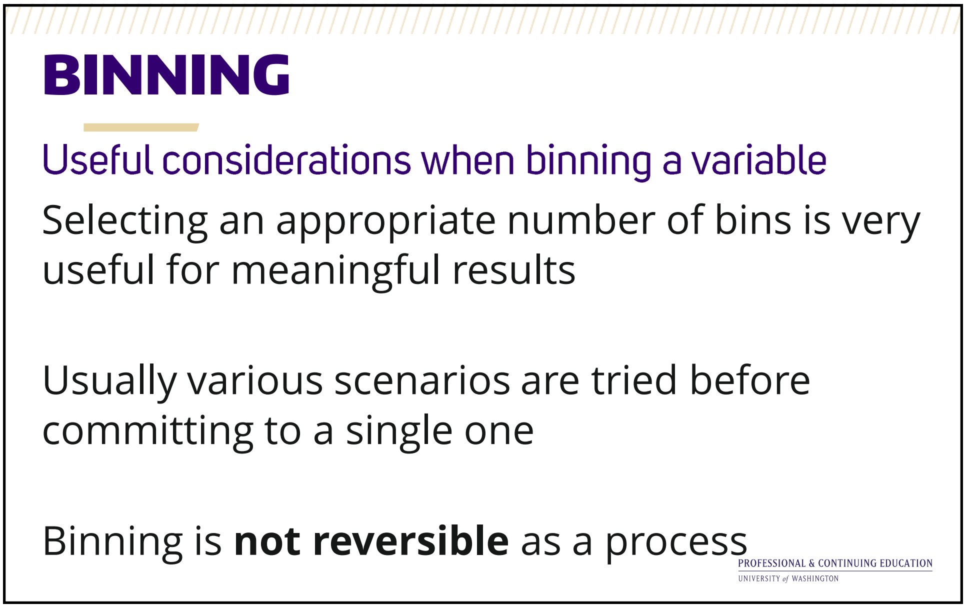


BINNING

Binning and histograms

- Plotting the results of a binning process = histogram
- Histograms are great for depicting what a variable's distribution looks like
  - Oftentimes, the histogram function is the same function used for finding the boundaries used in binning

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



BINNING

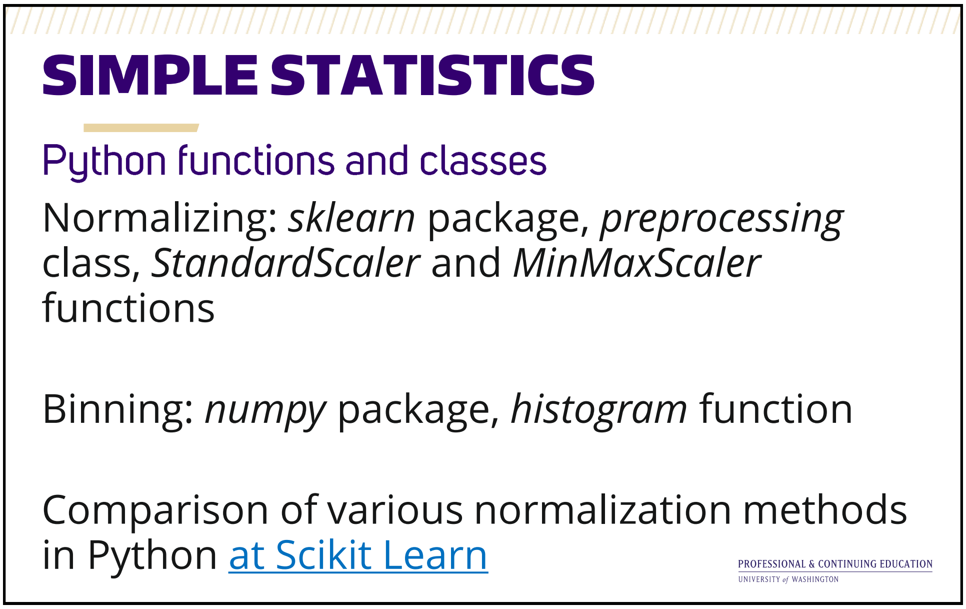
Useful considerations when binning a variable

Selecting an appropriate number of bins is very useful for meaningful results

Usually various scenarios are tried before committing to a single one

Binning is **not reversible** as a process

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



SIMPLE STATISTICS

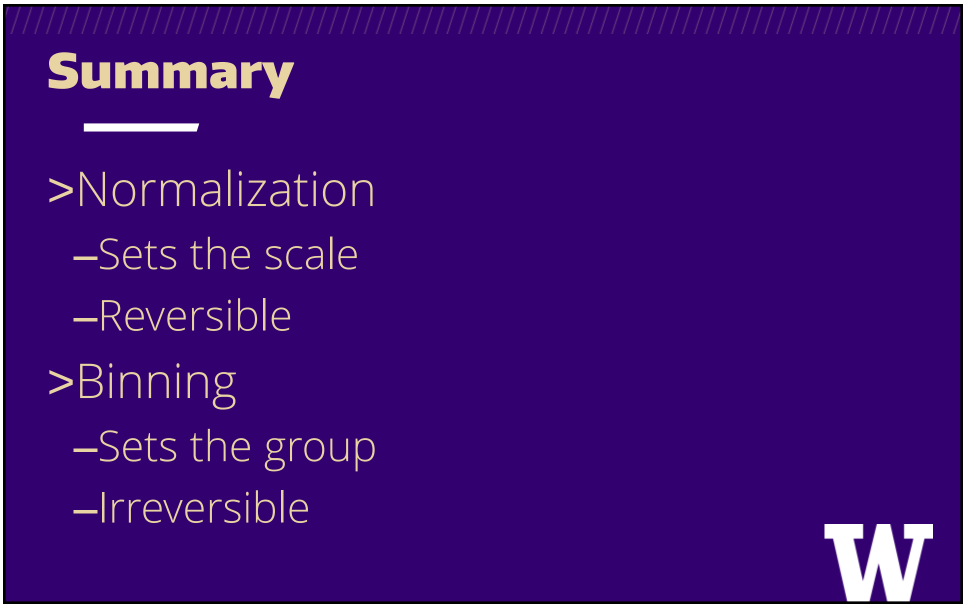
Python functions and classes

Normalizing: *sklearn* package, *preprocessing* class, *StandardScaler* and *MinMaxScaler* functions

Binning: *numpy* package, *histogram* function

Comparison of various normalization methods in Python at Scikit Learn

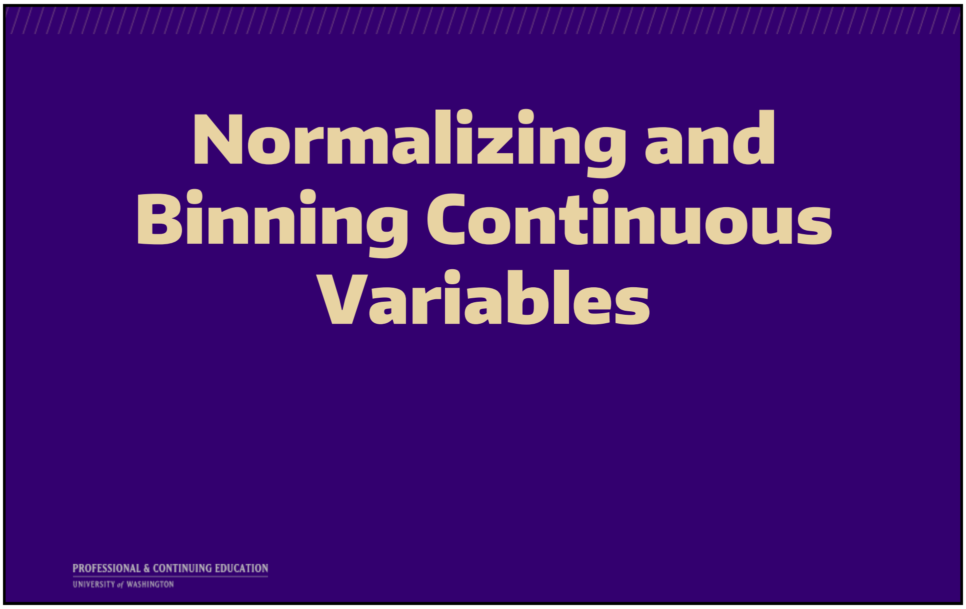
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



Summary

>Normalization
- Sets the scale
- Reversible

>Binning
- Sets the group
- Irreversible



Normalizing and Binning Continuous Variables

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON