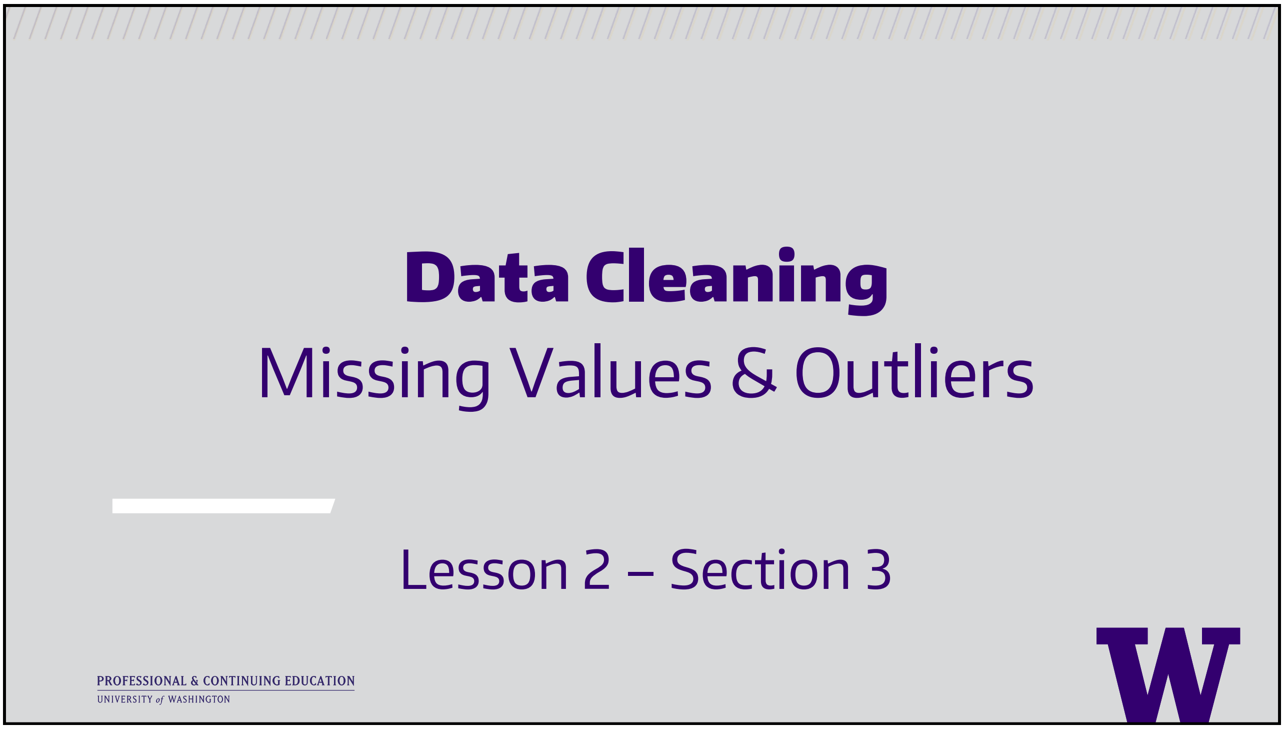
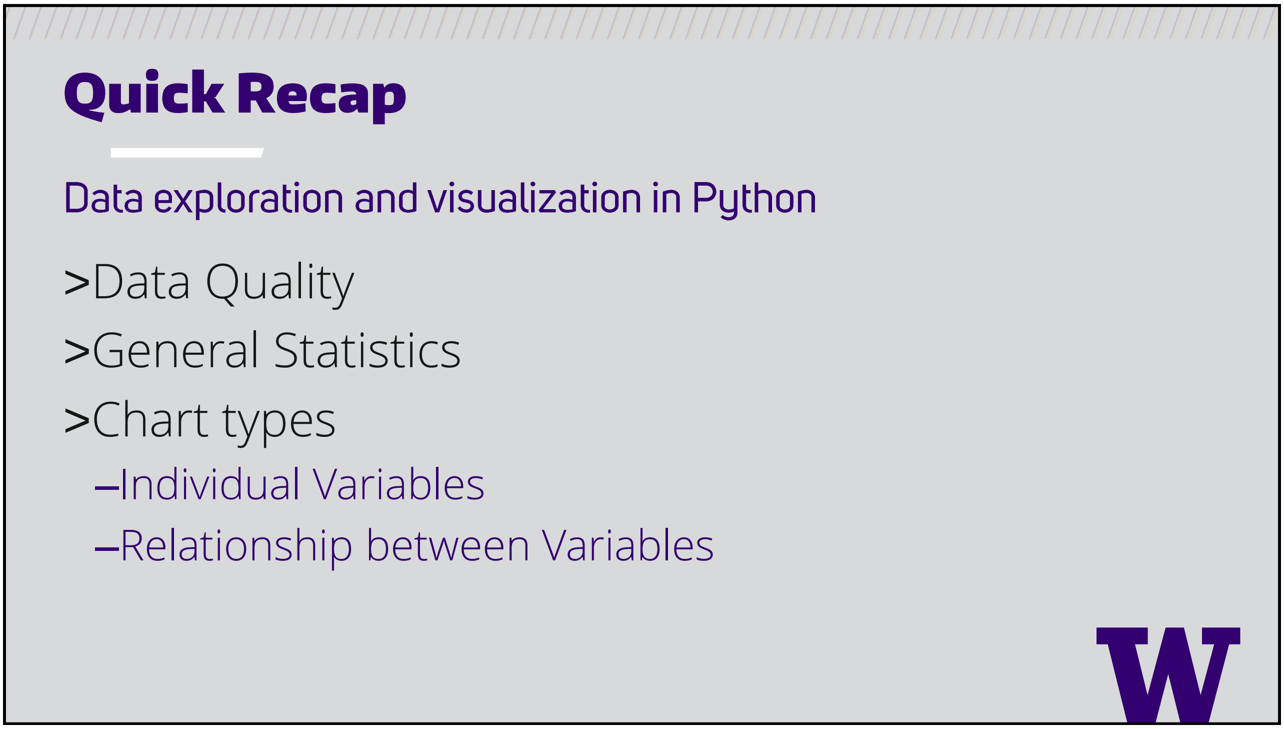


Data Cleaning

Missing Values & Outliers

Lesson 2 – Section 3

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



Quick Recap

Data exploration and visualization in Python

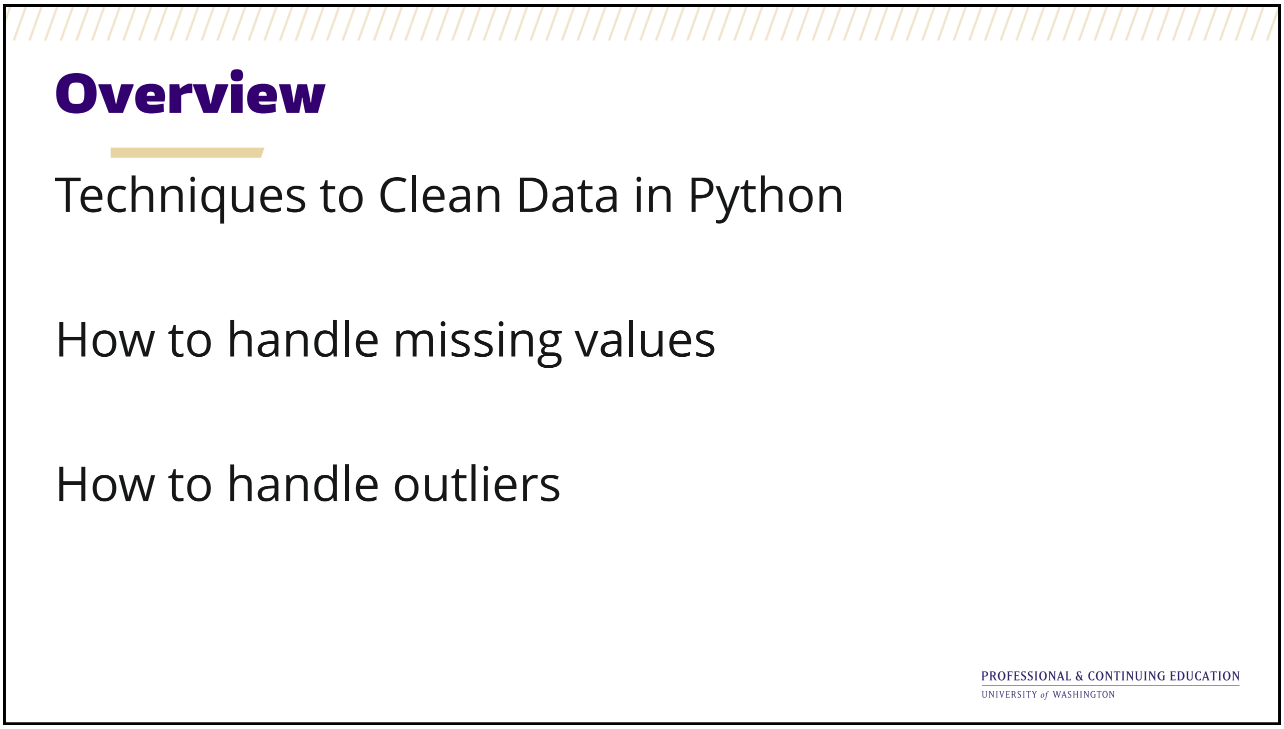
>Data Quality

>General Statistics

>Chart types

–Individual Variables

–Relationship between Variables

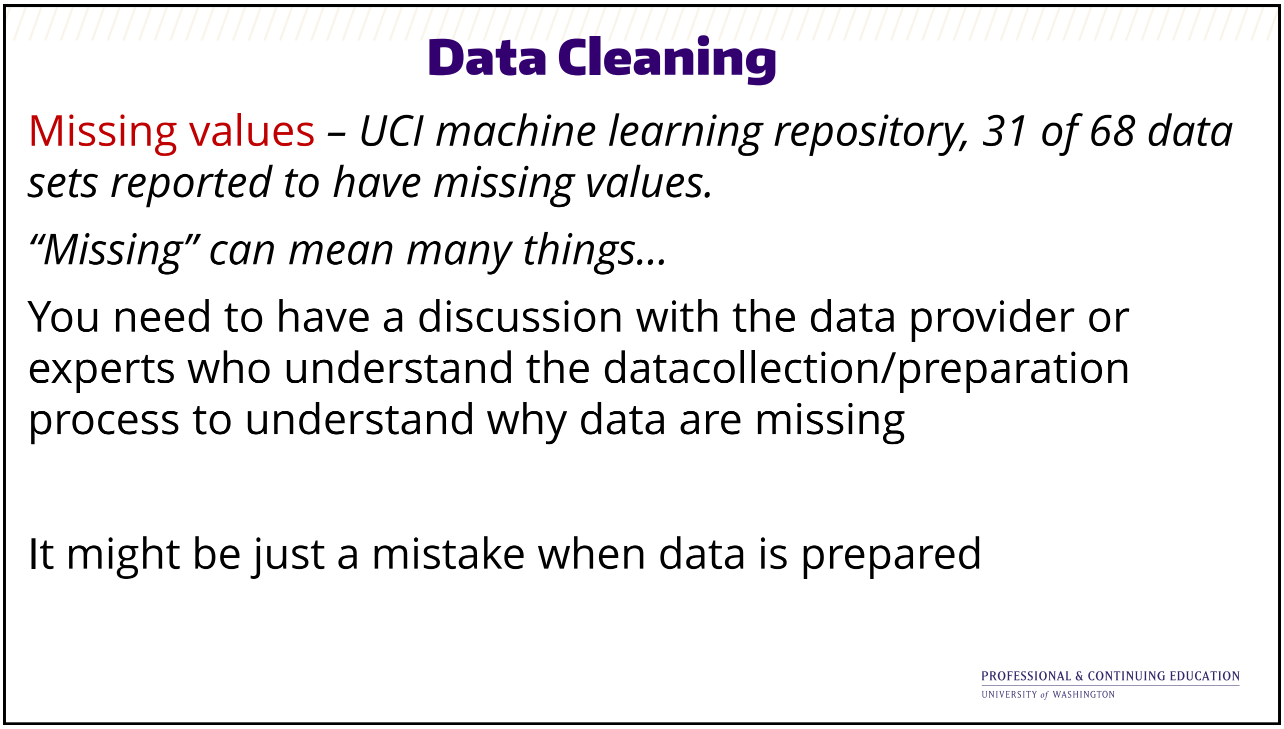


Overview

Techniques to Clean Data in Python

How to handle missing values

How to handle outliers



Data Cleaning

Missing values – *UCI machine learning repository, 31 of 68 data sets reported to have missing values.*

"Missing" can mean many things…

You need to have a discussion with the data provider or experts who understand the datacollection/preparation process to understand why data are missing

It might be just a mistake when data is prepared

Dealing With Missing Data – 1

Throw away cases with missing values

- in some data sets, most cases get thrown away
- if not missing at random, throwing away cases can bias sample towards certain kinds of cases

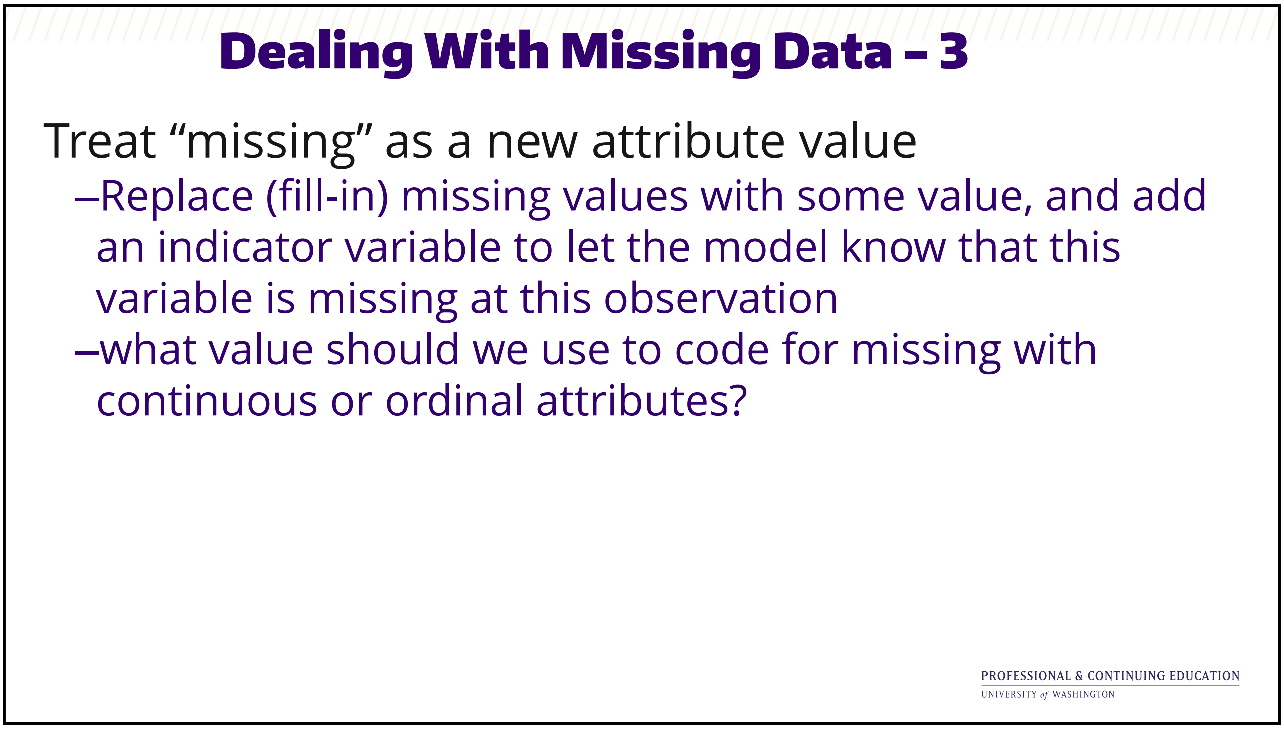
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Dealing With Missing Data – 2

Impute (fill-in) missing values

- –Once filled in, data set is easy to use
- –However, if missing values poorly predicted, may hurt performance of subsequent uses of data set

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

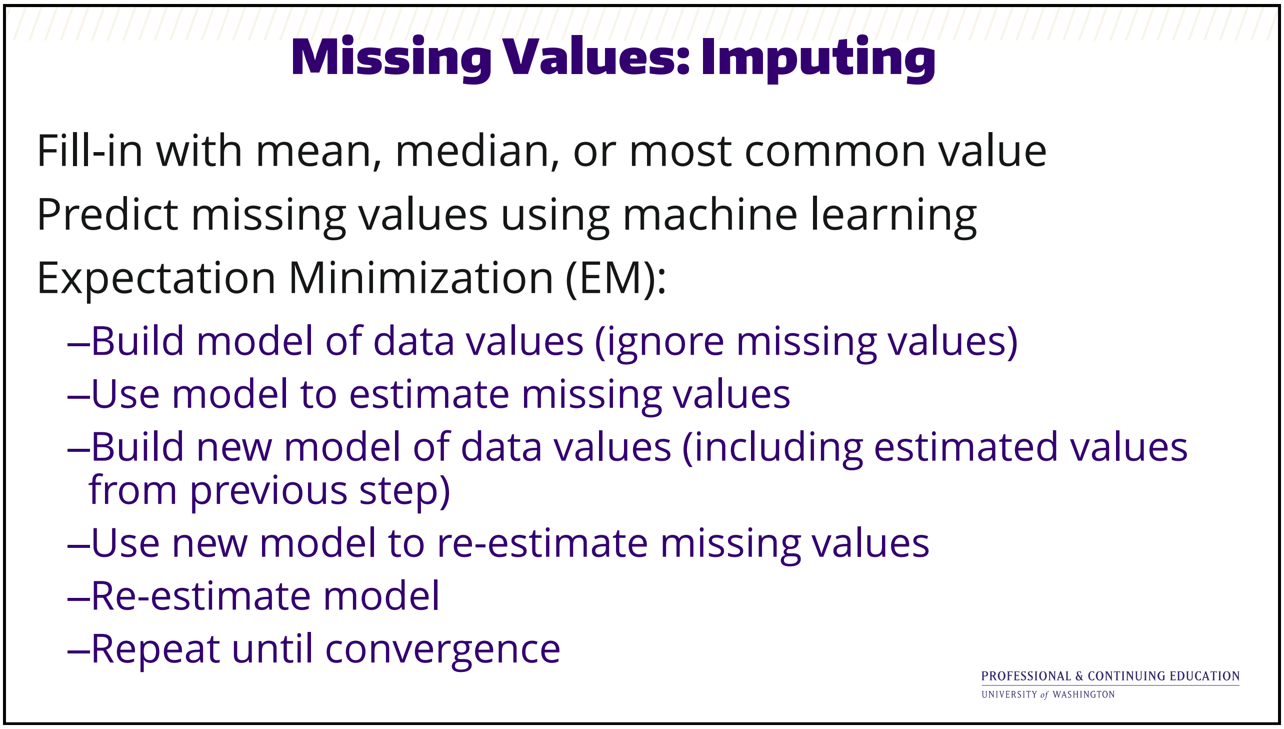


Dealing With Missing Data – 3

Treat “missing” as a new attribute value

- –Replace (fill-in) missing values with some value, and add an indicator variable to let the model know that this variable is missing at this observation
- –what value should we use to code for missing with continuous or ordinal attributes?

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



Missing Values: Imputing

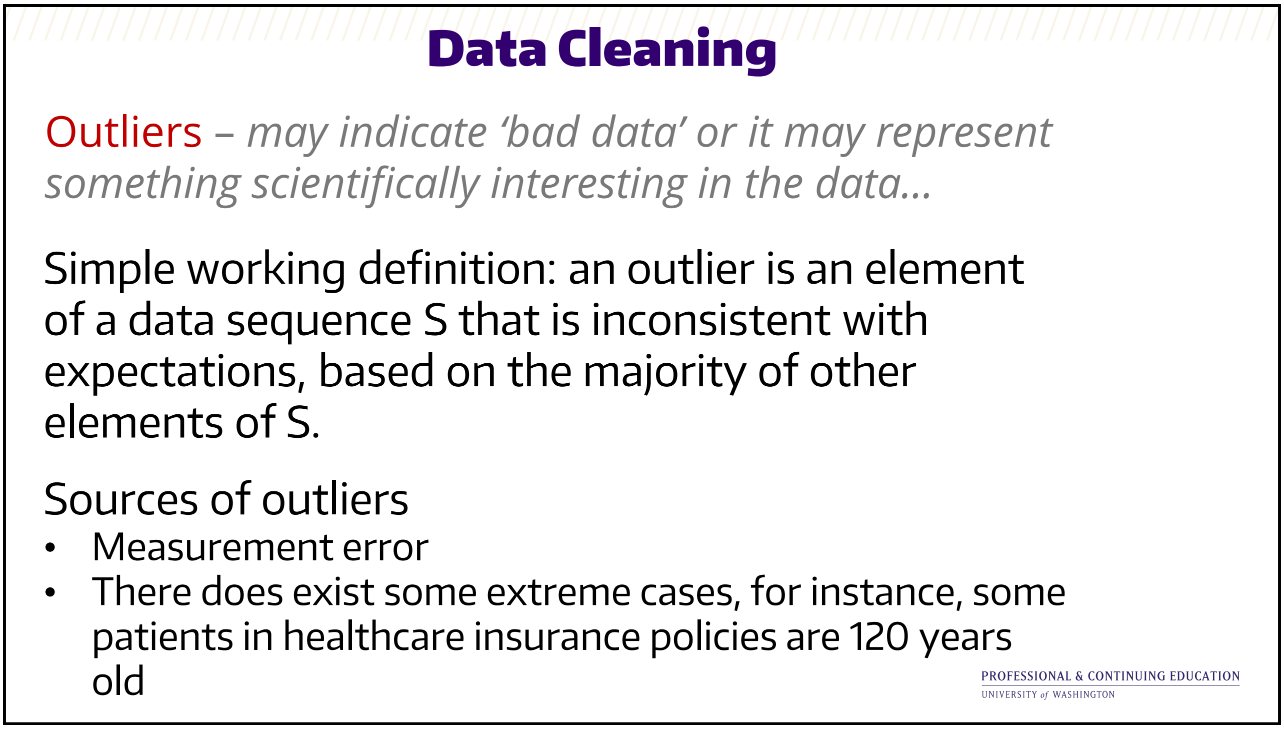
Fill-in with mean, median, or most common value

Predict missing values using machine learning

Expectation Minimization (EM):

- –Build model of data values (ignore missing values)
- –Use model to estimate missing values
- –Build new model of data values (including estimated values from previous step)
- –Use new model to re-estimate missing values
- –Re-estimate model
- –Repeat until convergence

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



Data Cleaning

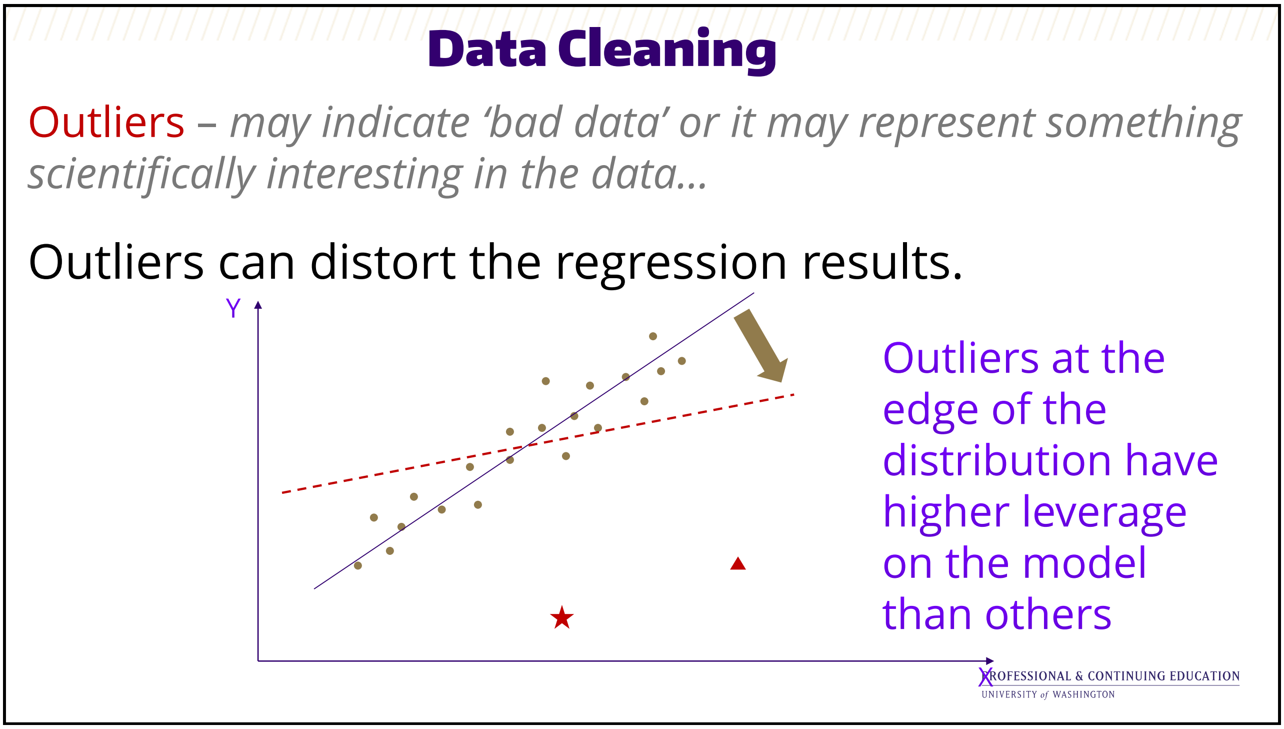
Outliers – *may indicate 'bad data' or it may represent something scientifically interesting in the data…*

Simple working definition: an outlier is an element of a data sequence S that is inconsistent with expectations, based on the majority of other elements of S.

Sources of outliers

- Measurement error
- There does exist some extreme cases, for instance, some patients in healthcare insurance policies are 120 years old

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



Data Cleaning

Outliers – *may indicate 'bad data' or it may represent something scientifically interesting in the data…*

Outliers can distort the regression results.

Outliers at the edge of the distribution have higher leverage on the model than others

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Data Cleaning

Outliers – *may indicate 'bad data' or it may represent something scientifically interesting in the data…*

Identify outliers

- Question origin, domain knowledge invaluable
- Dispersion – *"spread" of a data set, departure from central tendency, use a box plot…*

Deal with outliers

- Winsorize – Set all outliers to a specified percentile of the data. Not equivalent to trimming, which simply excludes data. In a Winsorized estimator, extreme values are instead replaced by certain percentiles (the trimmed minimum and maximum). Same as clipping in signal processing.

{92,19,**101**,58,**1053**,91,26,78,10,13,**-40**,**101**,86,85,15,89,89,28,-5,41}

Winsorize

-40
-5 ← 10th percentile
10
13
15
19
26
28
41
58
78
85
86
89
89
91
92
101 ← 90th percentile
101
1053

{92,19,**101**,58,**1053**,91,26,78,10,13,**-40**,**101**,86,85,15,89,89,28,**-5**,41}

Mean =101.5

{92,19,**101**,58,**101**,91,26,78,10,13,**-5**,**101**,86,85,15,89,89,28,**-5**,41}

Mean = 55.65

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Deal with outliers: Robust statistics, and Transformation

- If you are only going to model with some statistics of a sequence of data, where outliers might exist
  - Median is more robust than mean
  - Median Absolute Deviation (MAD) is more robust than standard deviation

$$MAD = median(|\, x_i - median(X)\,|)$$

$$\text{where } X = [x_1, x_2, \Lambda\ , x_n]$$

  - Relationship between MAD and Standard Deviation? For normal distribution, SD = 1.4826*MAD
- Data transformation can eliminate the extreme tendency of the outlier e.g. transforming to Log scale converts extreme values to acceptable range

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Summary

Two data cleaning techniques:

>How to handle missing data?

>How to handle outliers?

Practiced in Python