Feature Engineering

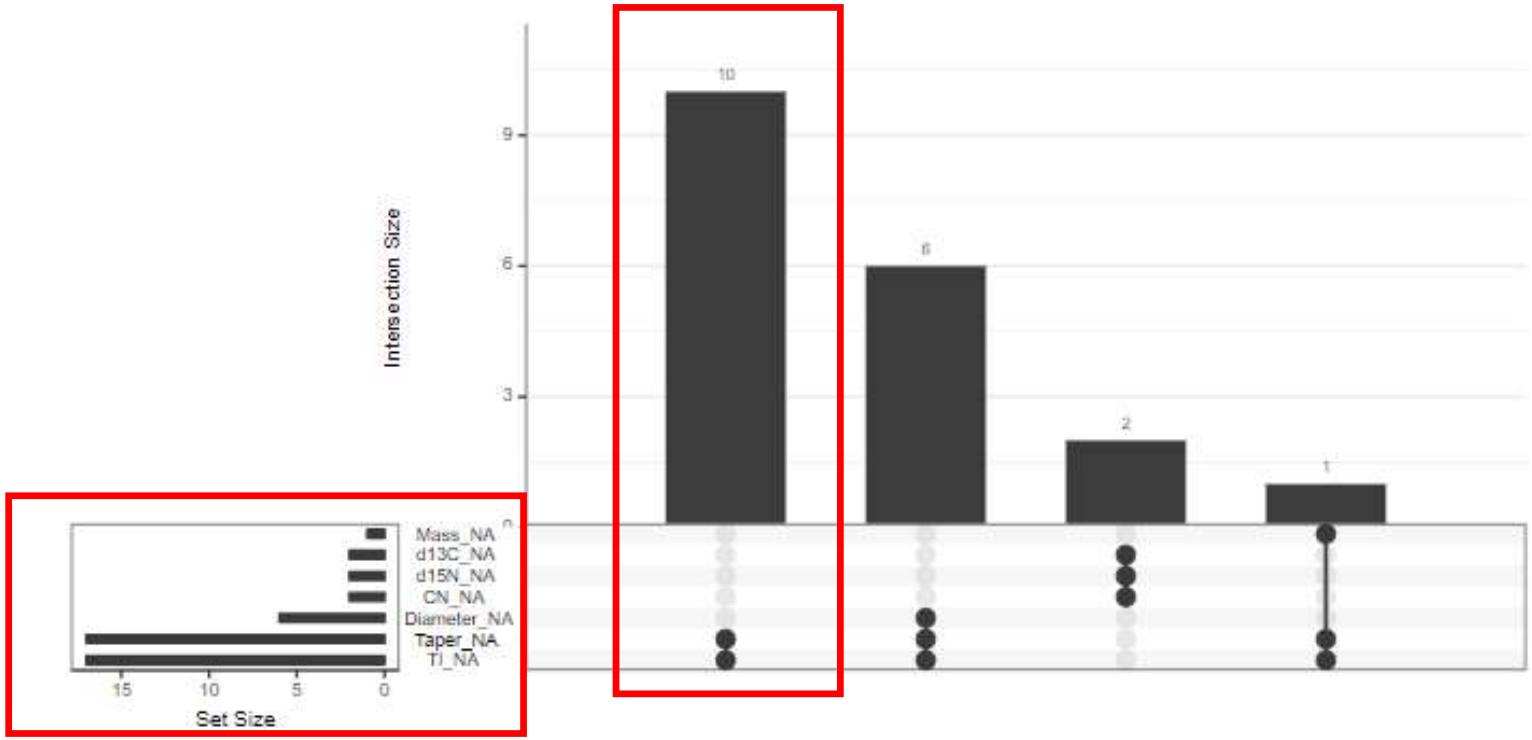
Chapter 8

Stephen Kimel

Reasons for Missing Data

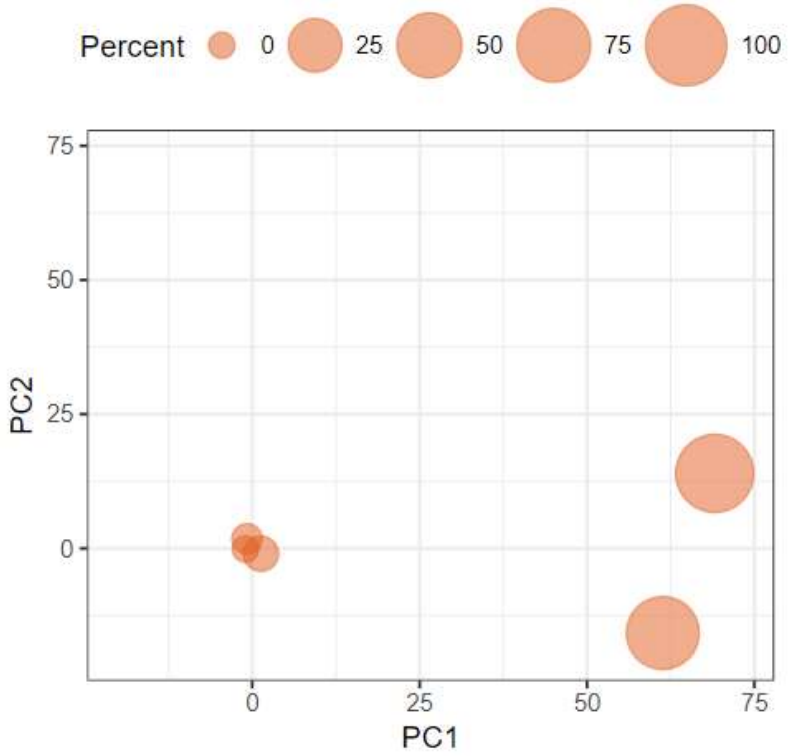
- Structural issues
- Random occurrences, or
- Specific causes

Visualizing Missing Data

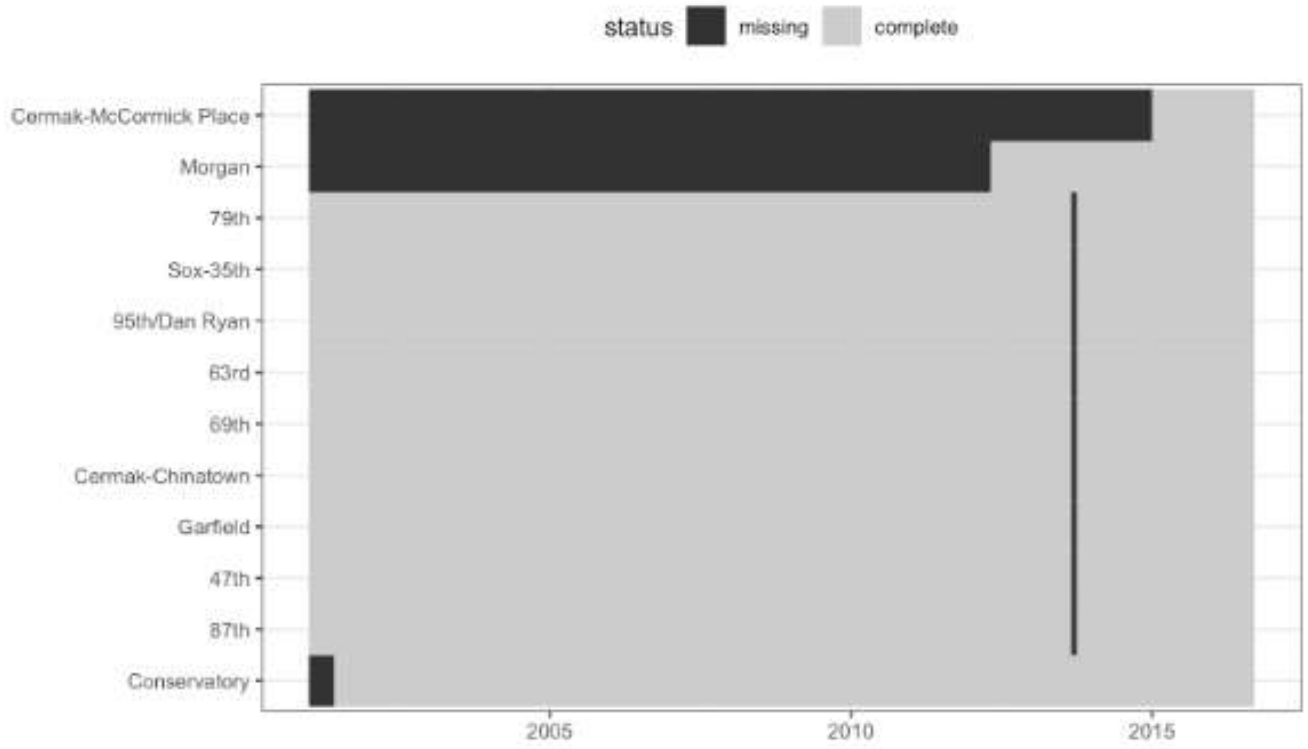


Using PCA to Identify Missing Data

0 is non-missing and 1 is missing



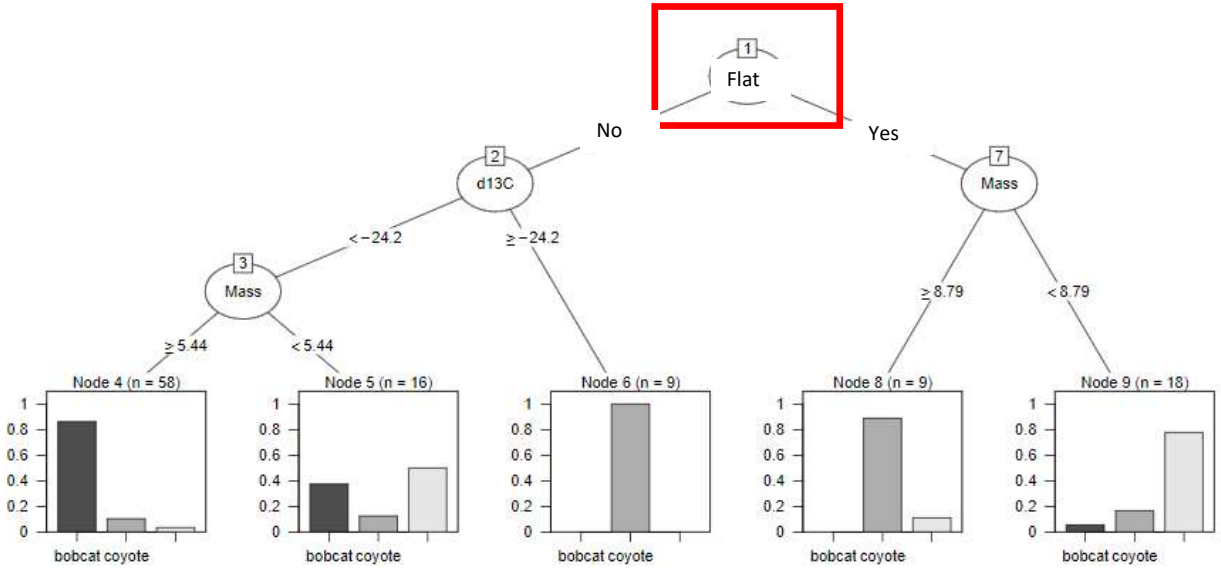
Visualizing Missing Data



Models Resistant to Missing Data

Bolds are Variables

- CART (Tree-based) surrogate split
  - Find variable highly correlated (**Flat**) with the variable you want to split on (**CN**)
  - If data point is missing (**CN**) for a certain observation, split on the highly correlated variable (



Can We Just Observations/Columns with Delete Missing Data?

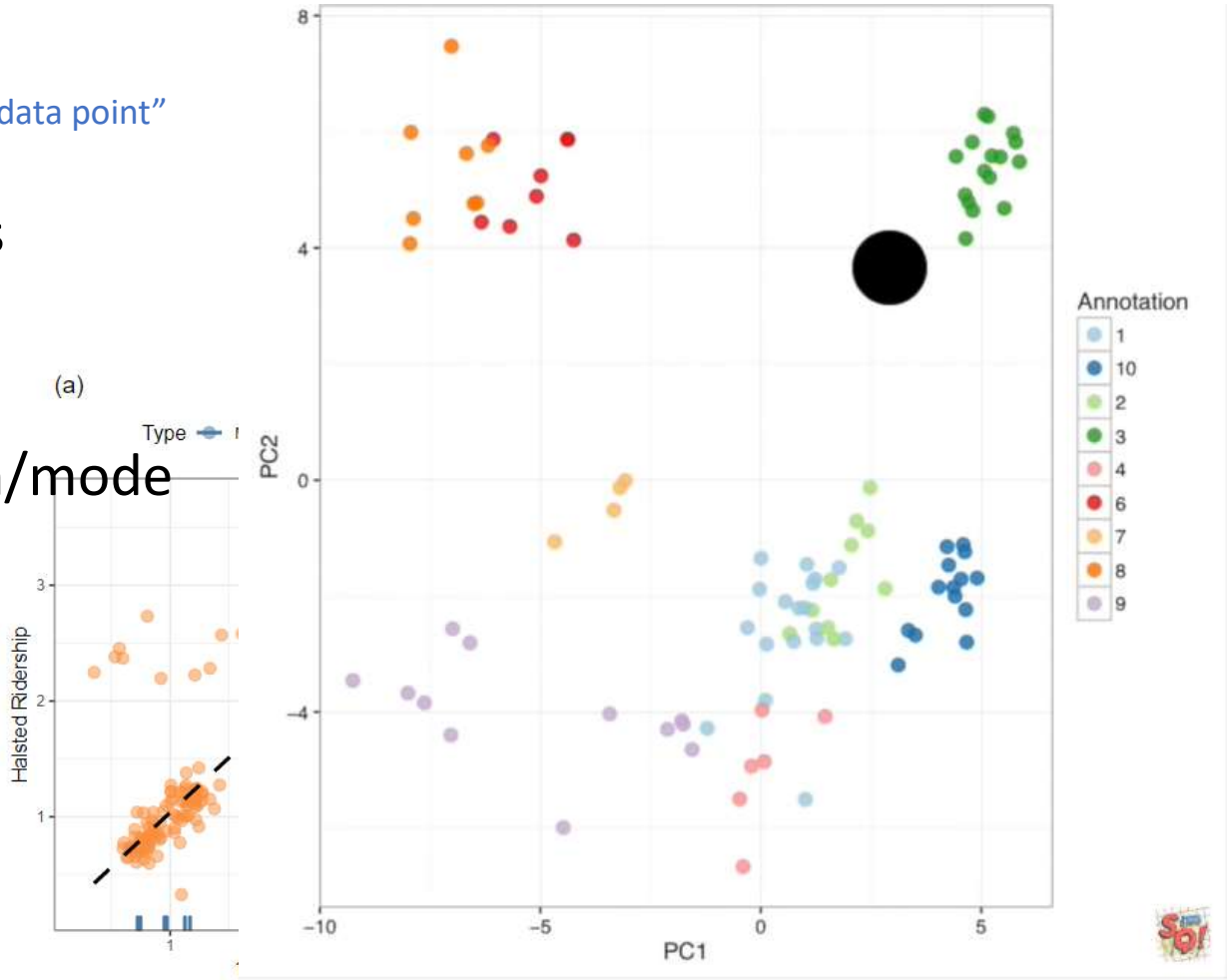
Encode Missing Data for Categorical Variable

- Create a category of “Missing”

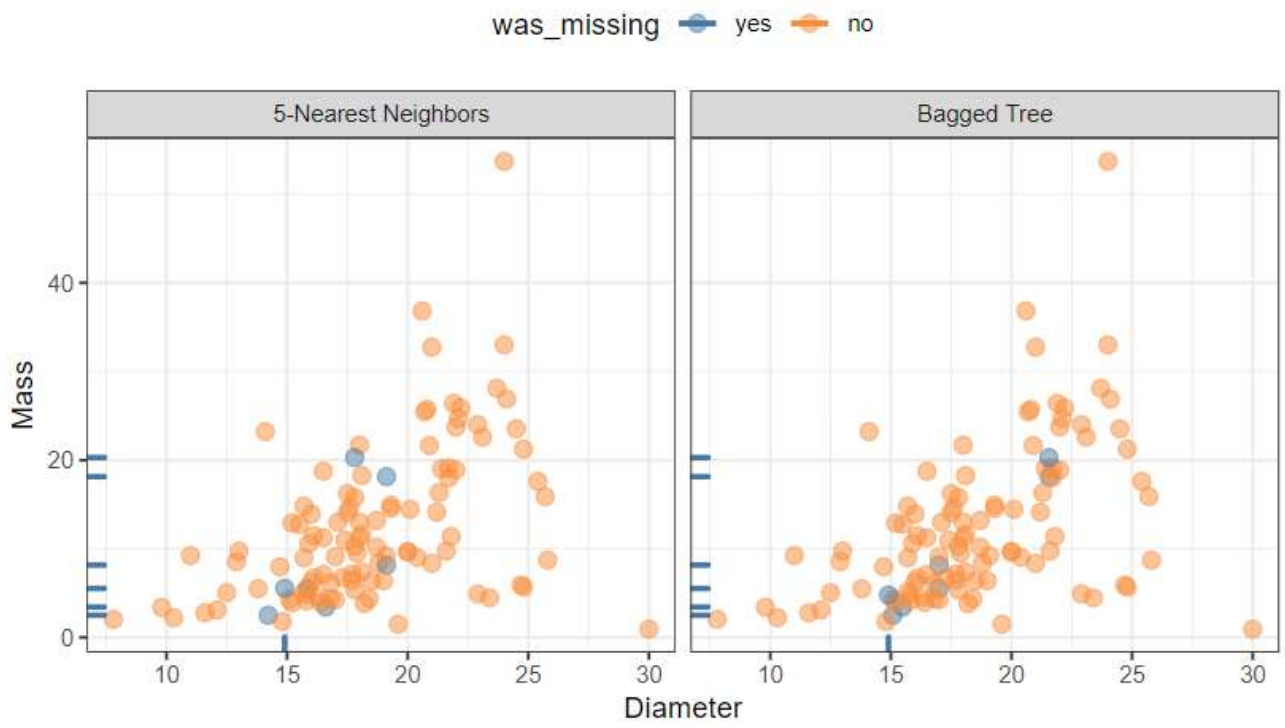
Imputation

"produce the most accurate prediction of the missing data point"

- At what stage in the preprocess should you impute?
- 20% missing rule-of-thumb
- Replace with the mean/median/mode
- K-Nearest Neighbors
- Tree-based methods
- Linear Method



Imputation Method Comarisons



Adding a Missing Column for Continuous Variables

- If imputing any values in a column, create a new column that indicates whether or not the value was imputed.

Special Cases

Time-based Variables

- Data censoring
  - Durations are often *right* censored since the terminating value is not known.
  - *Left* censoring can occur. For example, laboratory measurements may have a *lower limit of detection*, which means that the measuring instrument cannot reliably quantify values below a threshold X.