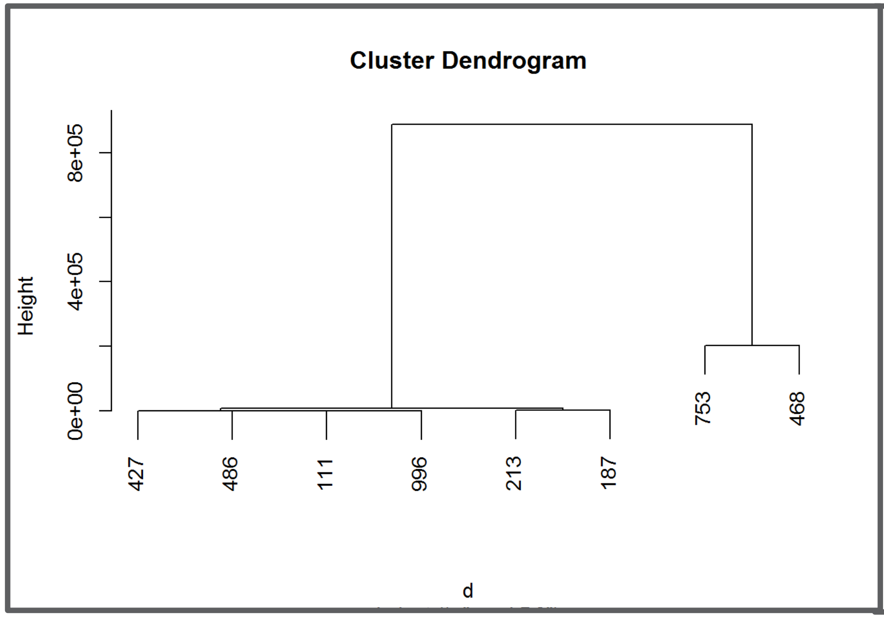
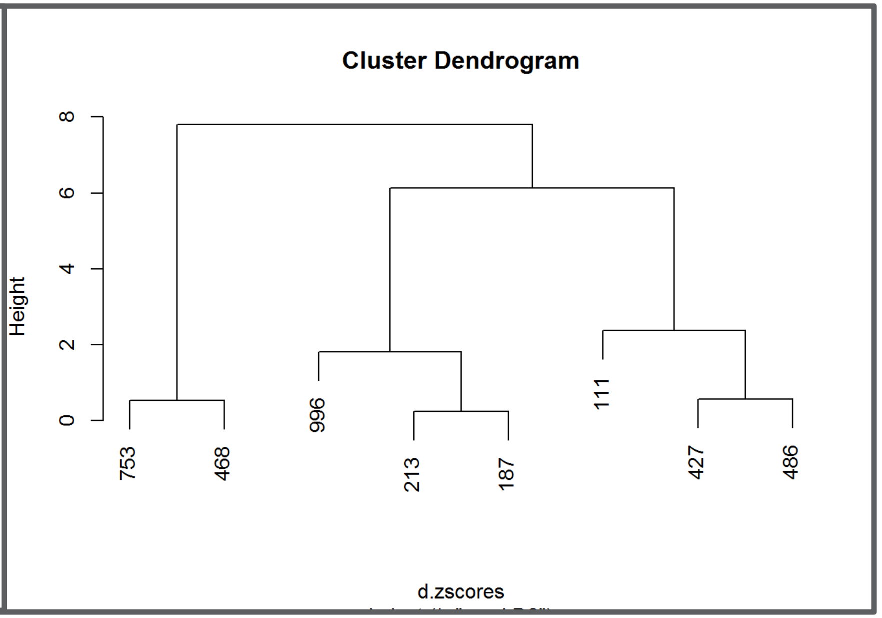
Data Preparation: Linear Transformations
Normalizing, Standardizing, and Rescaling



Before Z-score Standardization



After Z-score Standardization

- **Centering**
- **Standardizing**
- **Z-scores**
- **Normalizing**
- **Rescaling**

Prof. Nagiza F. Samatova
samatova@csc.ncsu.edu
Department of Computer Science
North Carolina State University

NC STATE Executive Education

February 2019

Linear Transformations

- Linear transformations of variables often do not affect the accuracy of predictive models such as linear regression
  - E.g.: Linear regression: any change in the x or y variables will be compensated for in corresponding changes in the β values
- However, linear transforms can still be important for at least 3 reasons:
  - Avoiding nonsensical values by centering
    - centering: subtracting the mean
    - the mean of centered data is always 0
  - Increasing comparability by Z-Score Standardization
    - e.g., distance calculations for clustering
    - Z-score standardization: dividing the centered variable by its standard deviation
    - the means of Z-scores are always 0 and their standard deviations are always 1,
    - so differences are always on the same scale
  - Reducing collinearity of predictors

Mean

Let $x = (x_1, x_2, \dots, x_n)$ be the quantitative variable over n observations

The mean of the variable:

$$\boxed{\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

```python
import numpy as np
```

```python
x = (2.1, 2.5, 4.0, 3.6)
x_bar = np.mean(x)
x_bar
```

```
3.05
```

$$\bar{x} = \frac{2.1 + 2.5 + 4.0 + 3.6}{4} = 3.05$$

Economic Growth % (x_i)	S & P 500 Returns % (y_i)
2.1	8
2.5	12
4.0	14
3.6	10

Centering

Let $x = (x_1, x_2, \ldots, x_n)$ be the column: quantitative variable over n observations

The mean of the vector: $$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{1}\sum_{i=1}^{n} x_i$$ ← **scalar, number**

Centering the variable: center x at its mean

$$\boxed{x_c = x - \bar{x} = (x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x})}$$ ⟵ **centered variable**

```
import numpy as np
```

```
x = (2.1, 2.5, 4.0, 3.6)
x_bar = np.mean(x)
x_bar
```

```
3.05
```

```
x_c = x - x_bar
x_c
```

```
array([-0.95, -0.55,  0.95,  0.55])
```

```
print("{:.2f} : mean of x_c".format(np.mean(x_c)))
```

```
0.00 : mean of x_c
```

Note: The mean of the centered vector is zero: $\overline{x_c} = 0$

$$\bar{x} = \frac{2.1 + 2.5 + 4.0 + 3.6}{4} = 3.05$$

$$x_c = (2.1 - 3.05, 2.5 - 3.05, 4.0 - 3.05, 3.6 - 3.05)$$
$$= (-.95, -0.55, 0.95, 0.55)$$

Economic Growth % (x_i)	S & P 500 Returns % (y_i)
2.1	8
2.5	12
4.0	14
3.6	10

Standardizing and Z-scores

Economic Growth % (x_i)	S & P 500 Returns % (y_i)
2.1	8
2.5	12
4.0	14
3.6	10

Let $x = (x_1, x_2, \ldots, x_n)$ be the column: variable over n observations

Centered variable: $\boxed{\boldsymbol{x_c = x - \bar{x}} = (x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x})}$

Variance: $\boldsymbol{var\,(x)} = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{\boldsymbol{n-1}}$

Standard Deviation: $\boldsymbol{sd(x)} = \sqrt{\boldsymbol{var(x)}}$

Standardizing using standard deviation: $\boxed{\boldsymbol{x_s = \frac{x}{sd(x)}}}$

Standardizing using mean & standard deviation (Z-score):

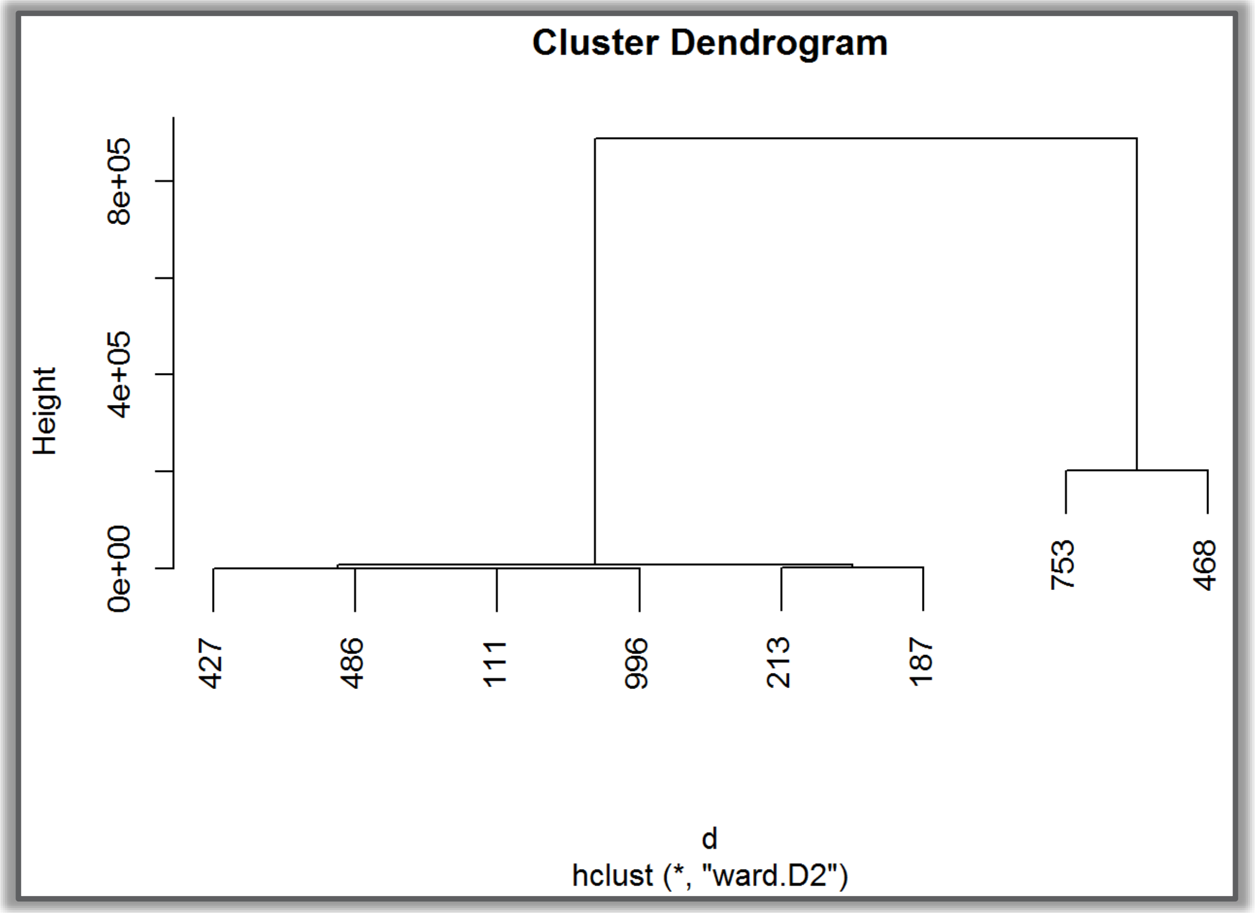
$$\boxed{Z\text{-score} = \frac{\boldsymbol{x-\bar{x}}}{\boldsymbol{sd(x)}} = \frac{\boldsymbol{x_c}}{\boldsymbol{sd(x)}}}$$

Z-score Standardization

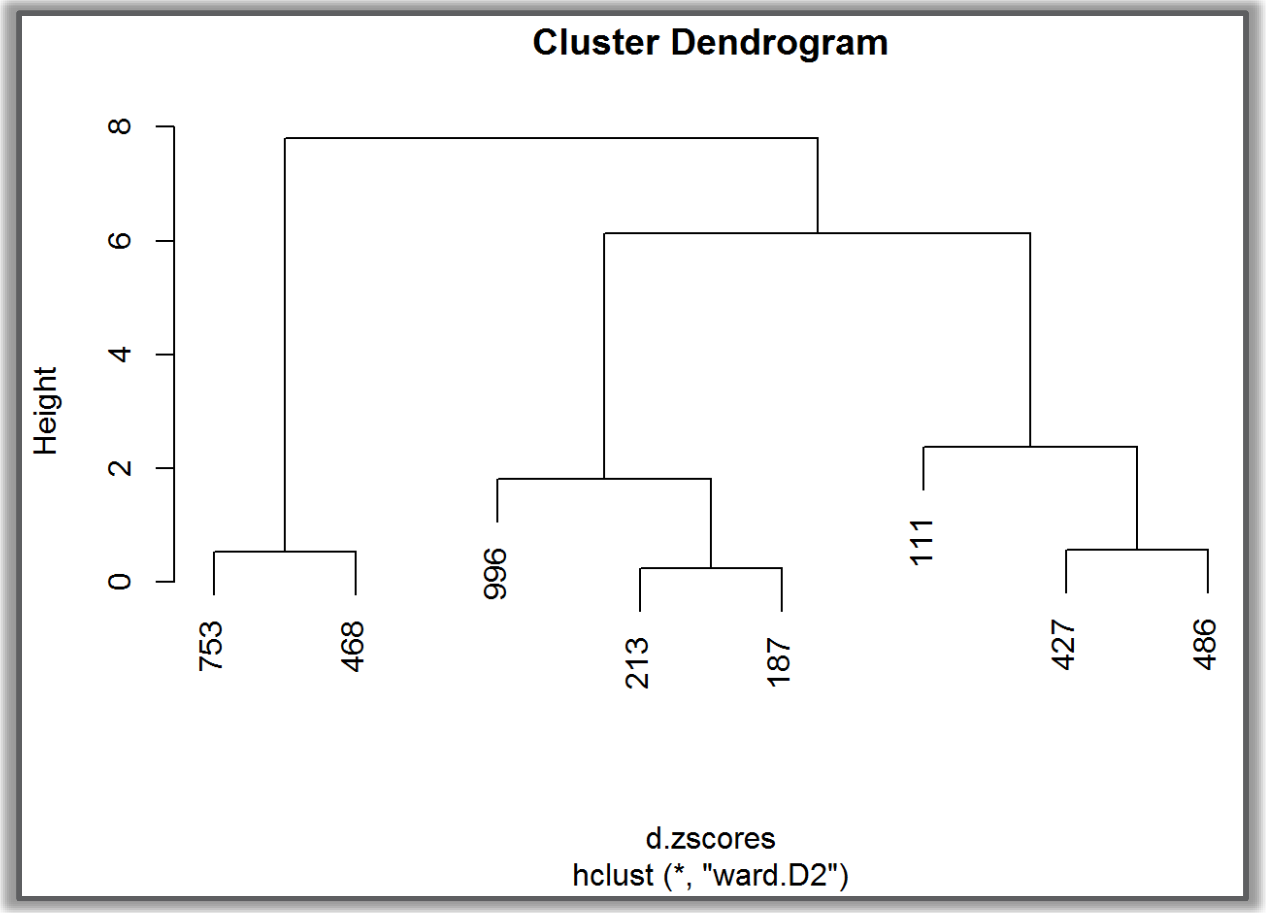
- Applied to symmetrically distributed data, such as normal distribution data
- If the distribution is skewed or wide:
  - then transformations for skewed and wide distributions should be applied first
  - before z-scores are computed
- Z-score values less than -1 signify values smaller than typical
- Z-score values greater than 1 signify values greater than typical

Z-scores Standardization for Increased Comparability

Before Z-score Standardization



After Z-score Standardization



Normalization

- **Useful when absolute quantities are less meaningful than relative ones:**
  - **the meaningful quantity can be external (came from analyst's domain knowledge) or internal (derived from the data)**
- **Examples:**
  - **normalizing income relative to another meaningful quantity: median income**
  - **rather than considering customer's absolute age, consider how old or young they are relative to a "typical" customer**
    - the mean age of customers can be treated as the typical age