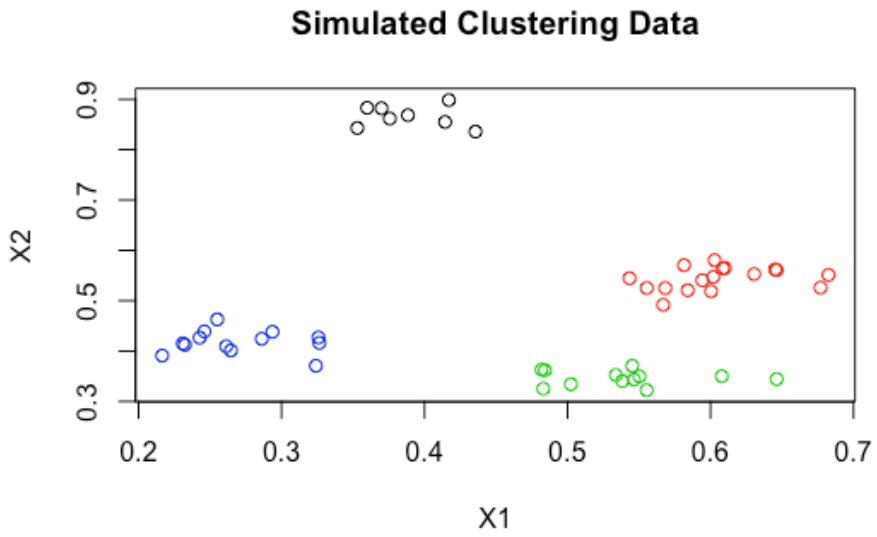
# K-means

- **K-means Clustering**
  1. Specify the number of clusters (K).
  2. Define what is **similar** or **dissimilar**
     → **Euclidean Distance!**

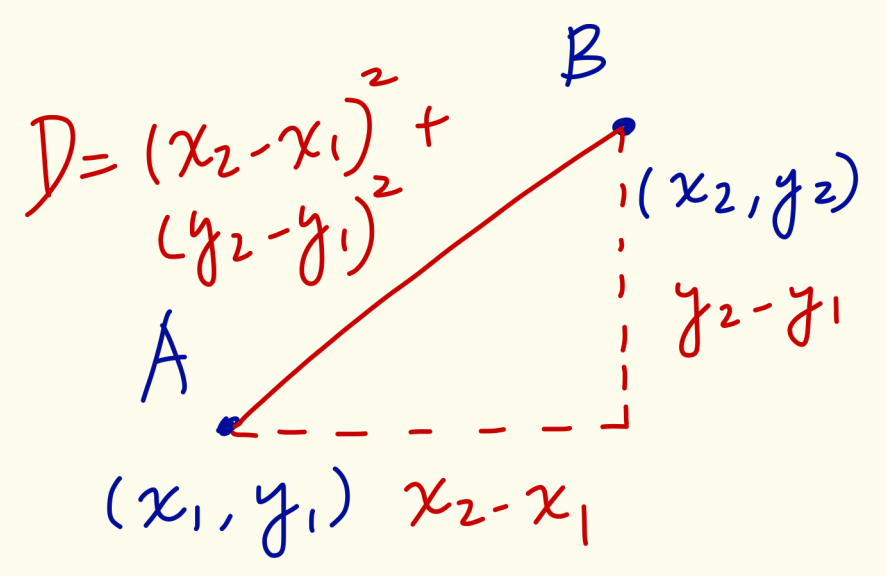


# Definition for (Dis)similarity

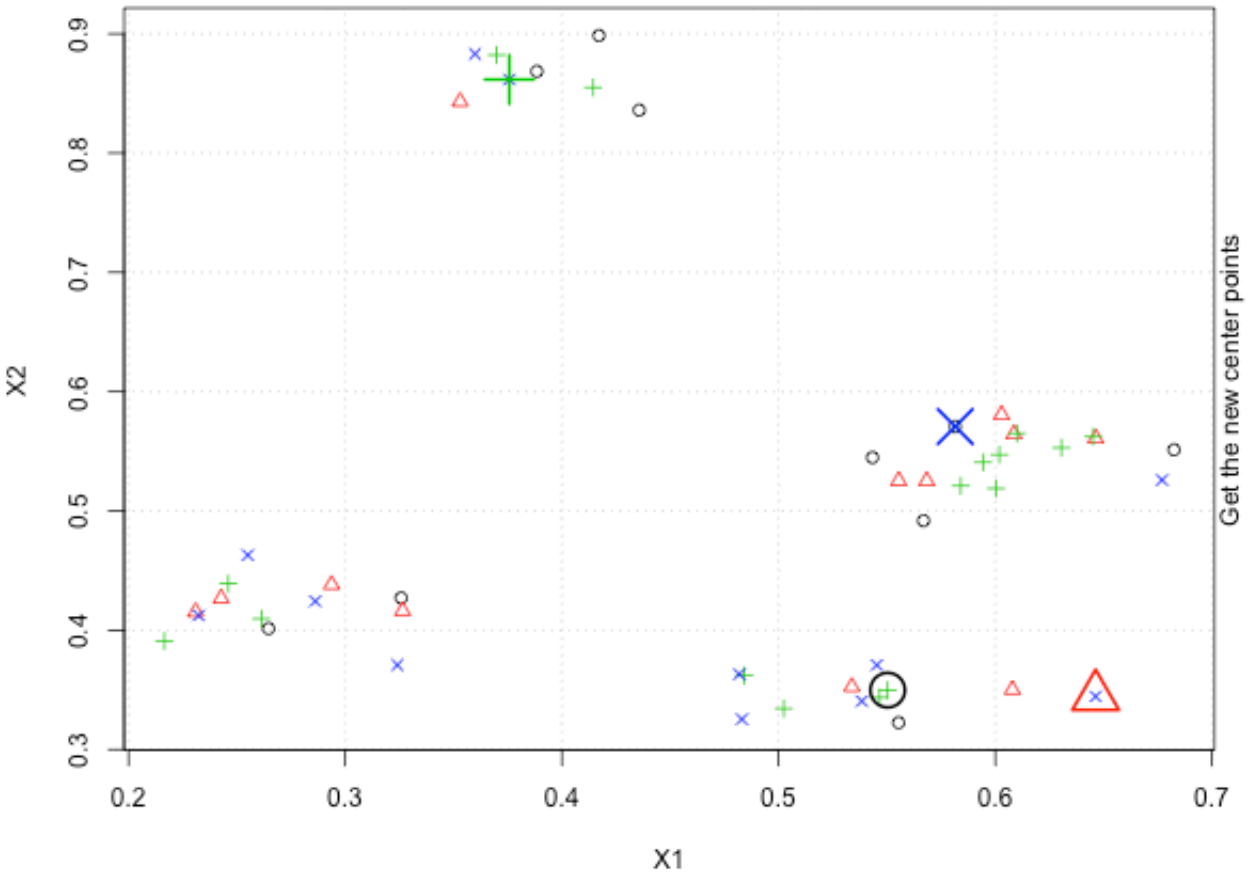
- Euclidean Distance

$$d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2.$$

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2$$



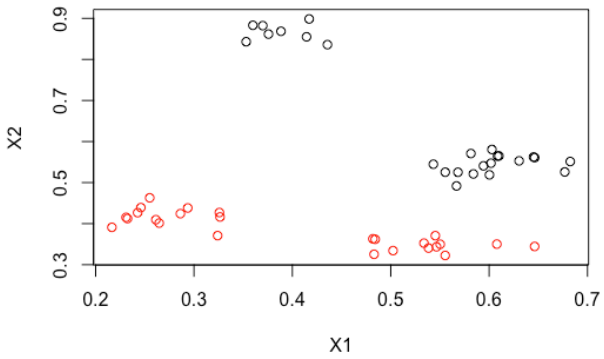
# K-means Algorithm

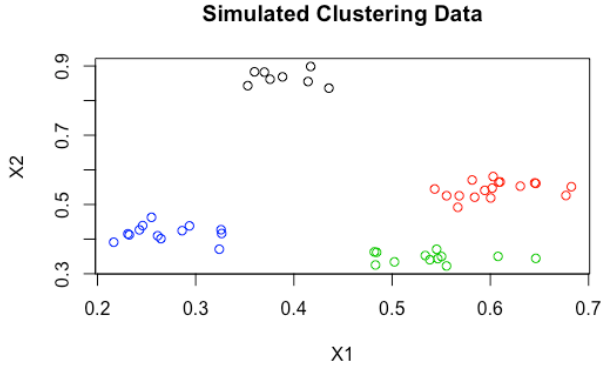


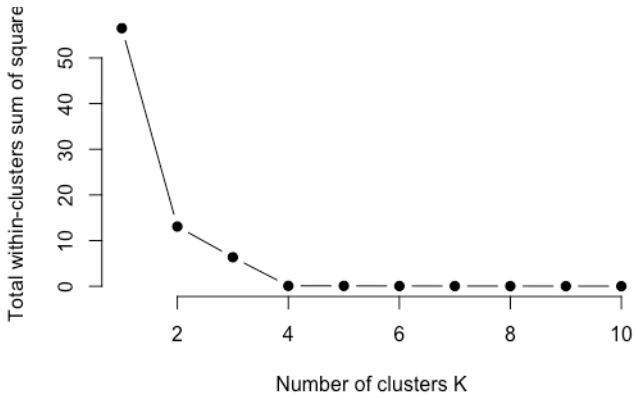
# Finding the best k

**TOTAL** Sum of Squares = **WITHIN** Groups SS + **BETWEEN** Groups SS

(**TOTAL** variance = **WITHIN** variance + **BETWEEN** variance)

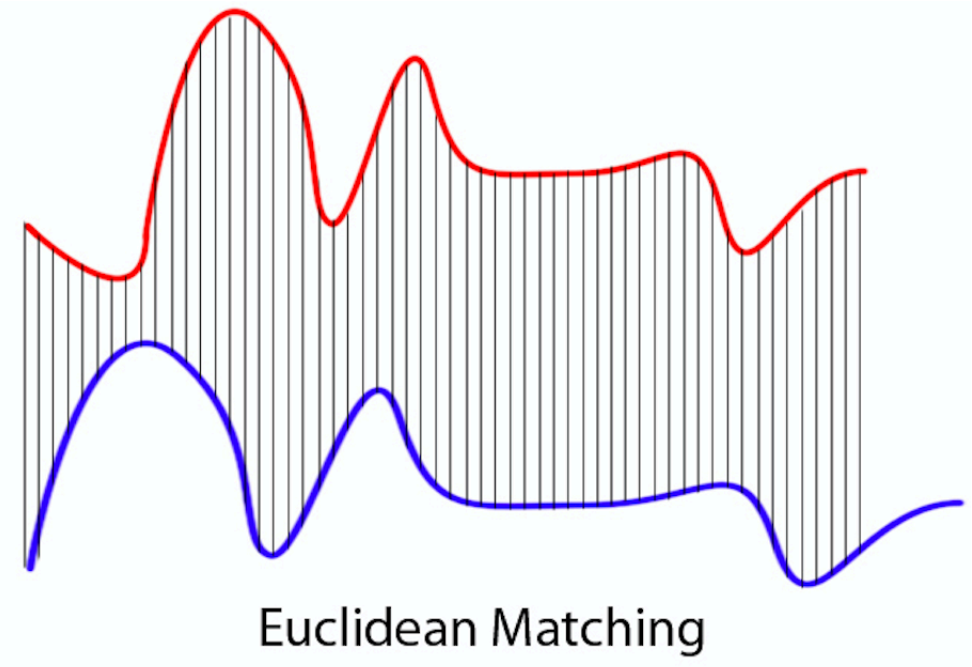


# Distance Between Two Time Series

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2$$

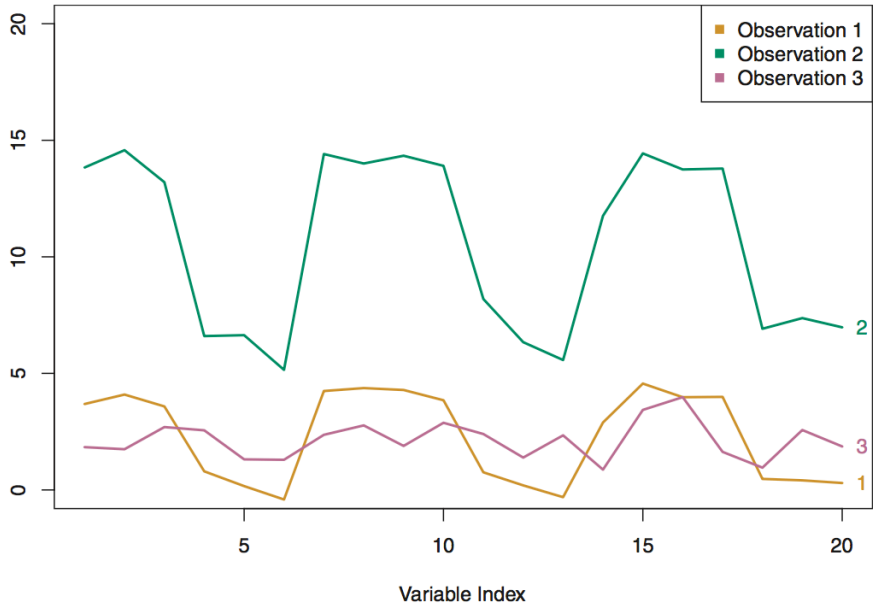


Euclidean Matching

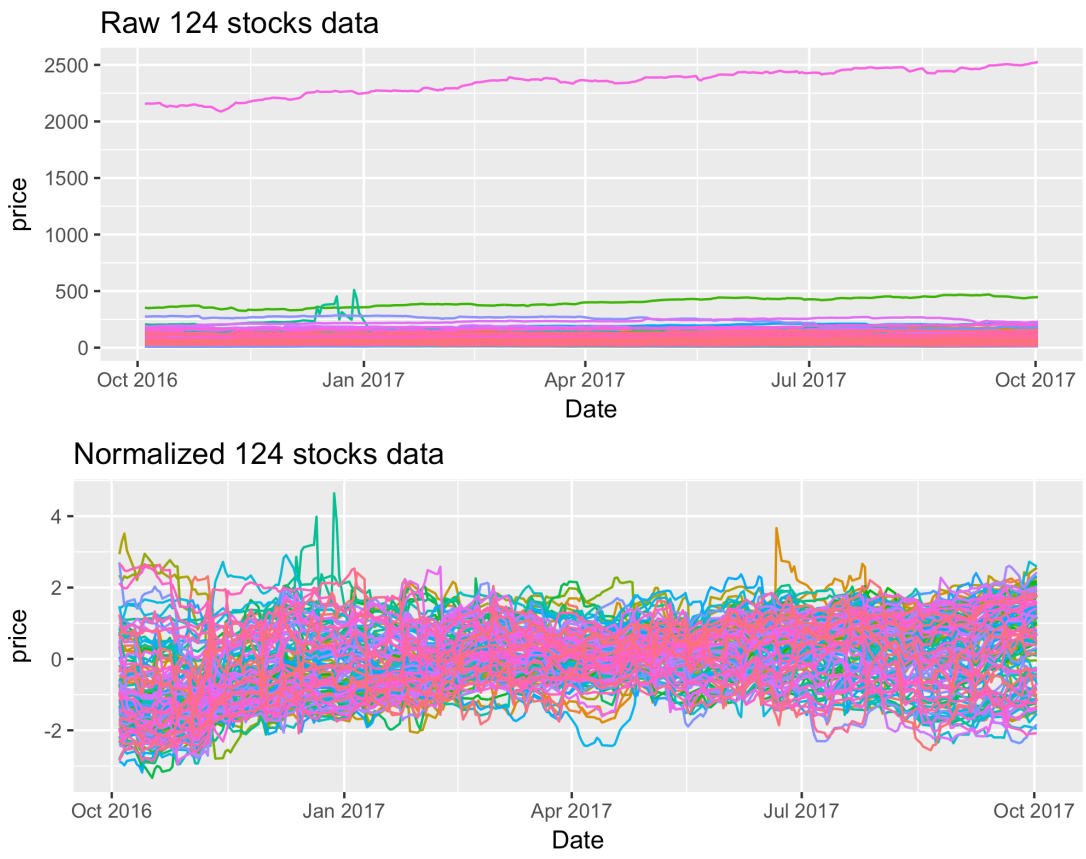
# Distance between Time Series

- **Correlation Distance may be a better idea!**

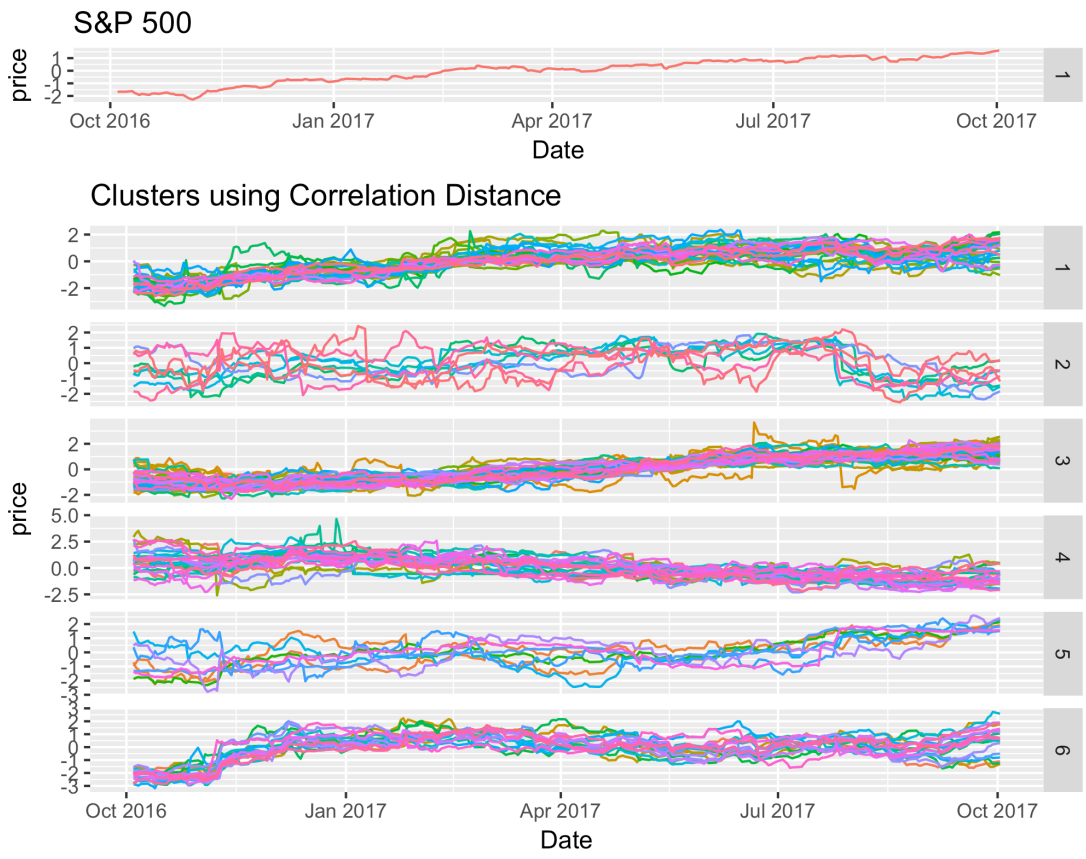
$$\rho(x_i, x_{i'}) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{i'j} - \bar{x}_{i'})^2}}$$



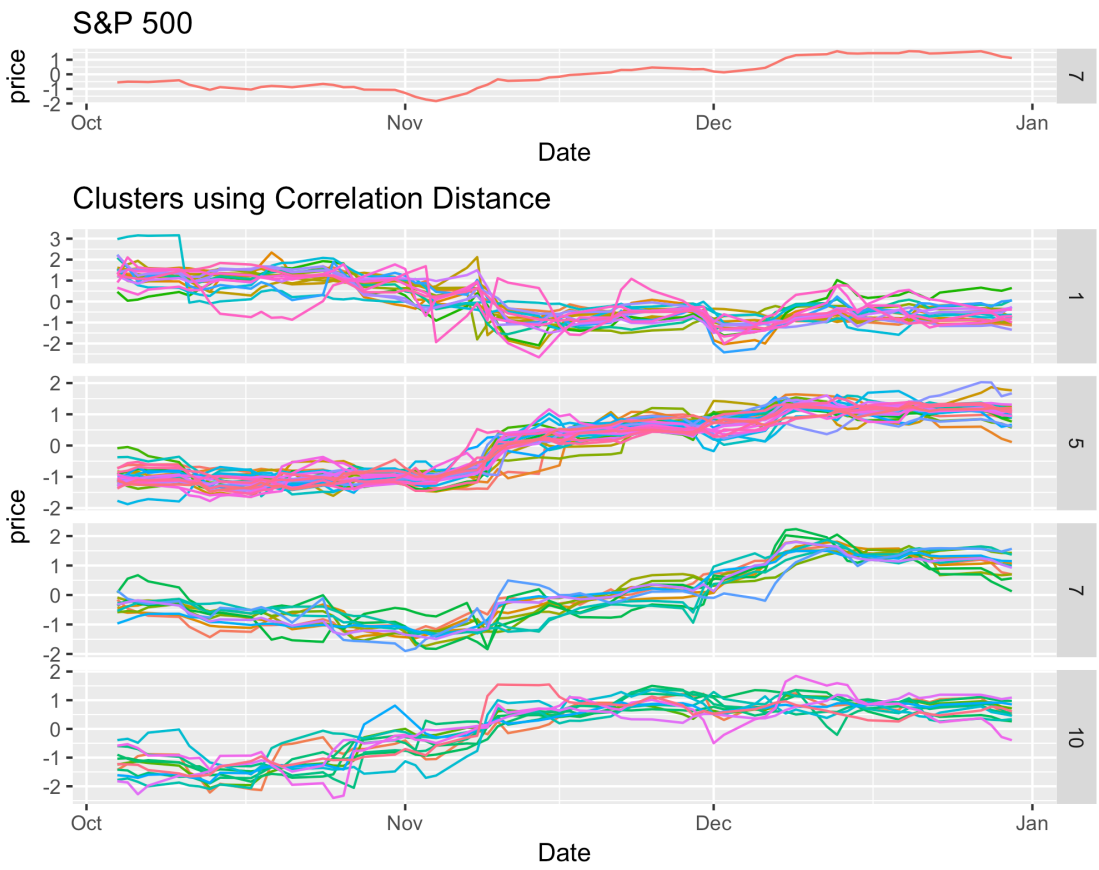
# All 124 stocks Data



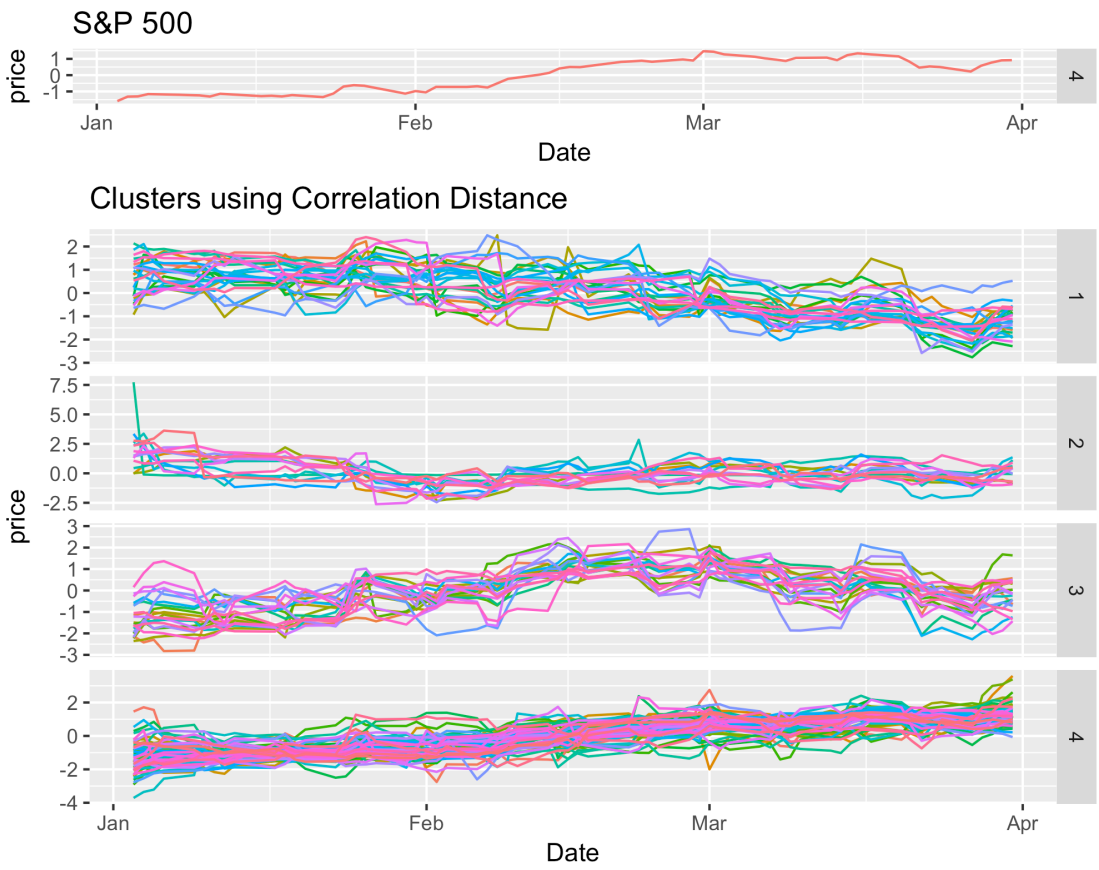
# One Year Data



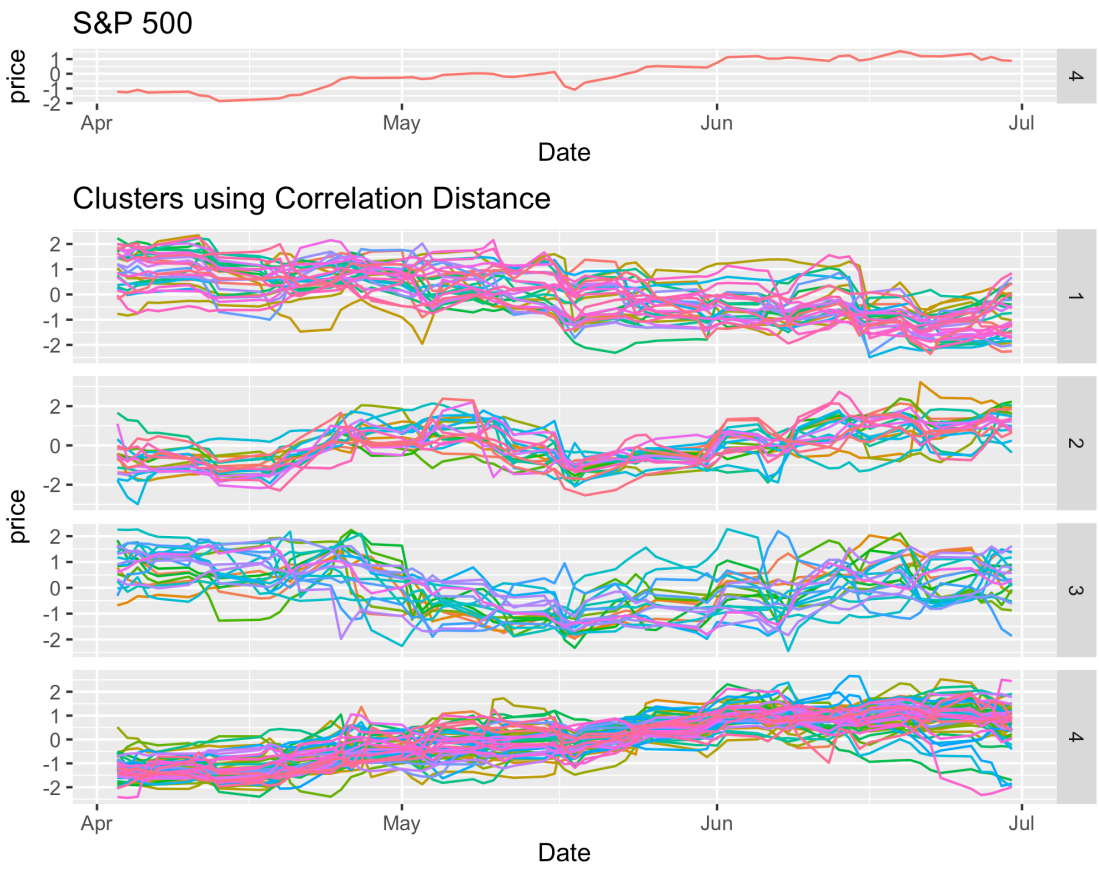
Oct. 2016 ~
Dec. 2016



Jan. 2017 ~
Mar. 2017



Apr. 2017 ~
Jun. 2017



Jul. 2017 ~
Sep. 2017