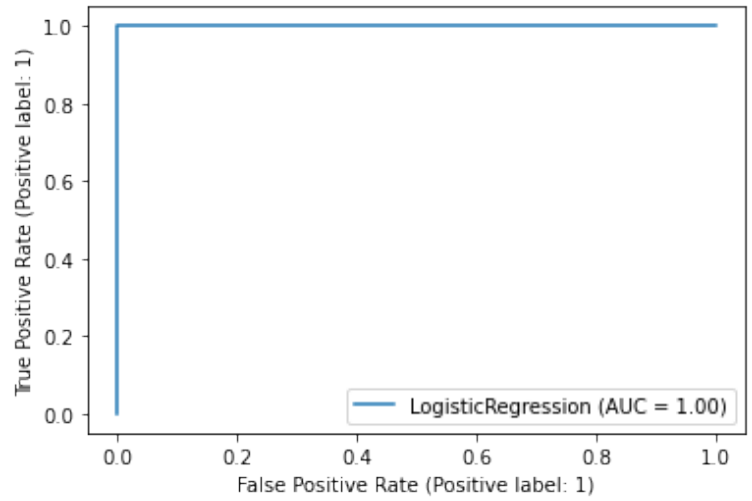
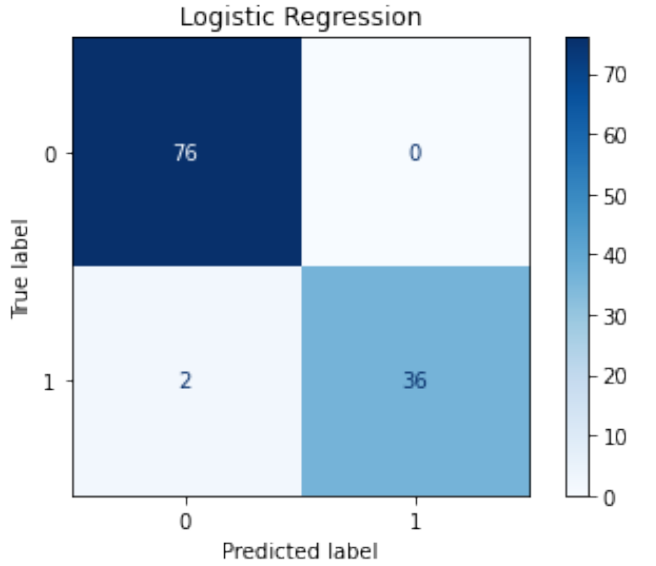
Week 06 Report

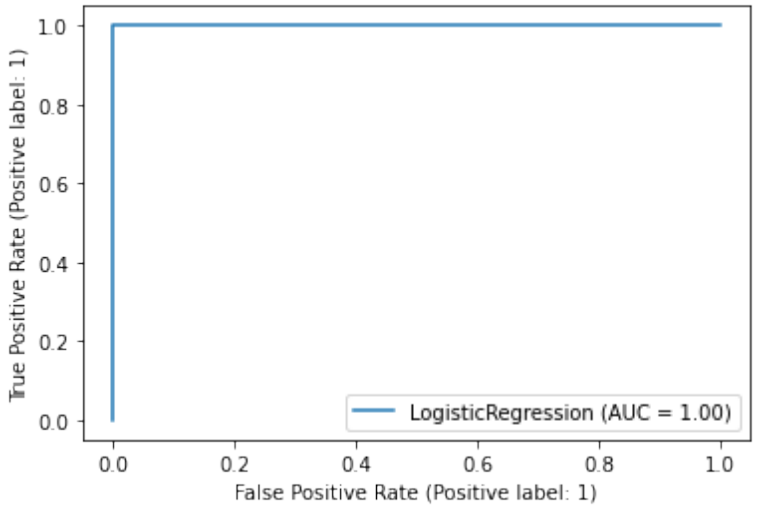
Jeffrey Li

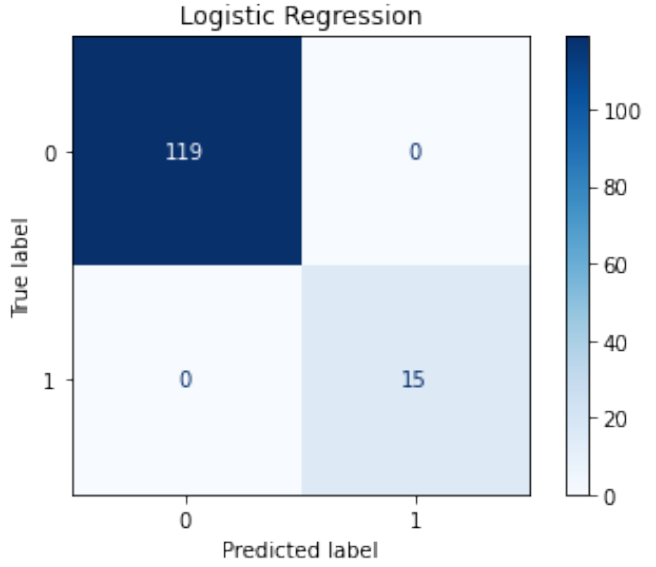
Previous Week

- Using Artificial Data
  - Artificial data was generated by taking the original data and adding Gaussian noise.
- Idea 1: Use artificial data in the testing process to see how model performs on more unseen data (above).
- Idea 2: Use artificial data in the training process to regularize the model (below).
- Results: Model achieved perfect Area Under the Curve (AUC) score.
  - Labels may need to be reconfigured.







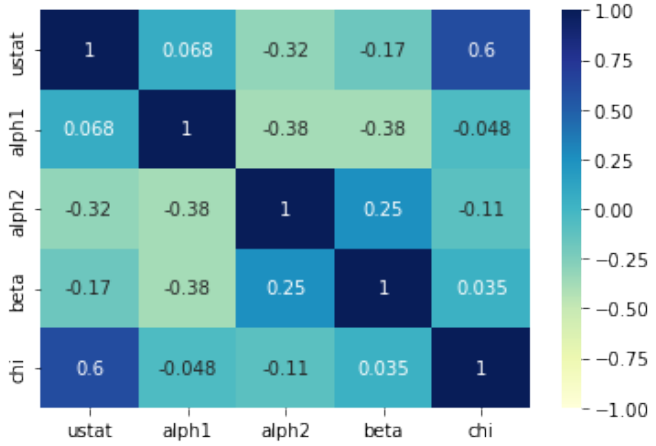


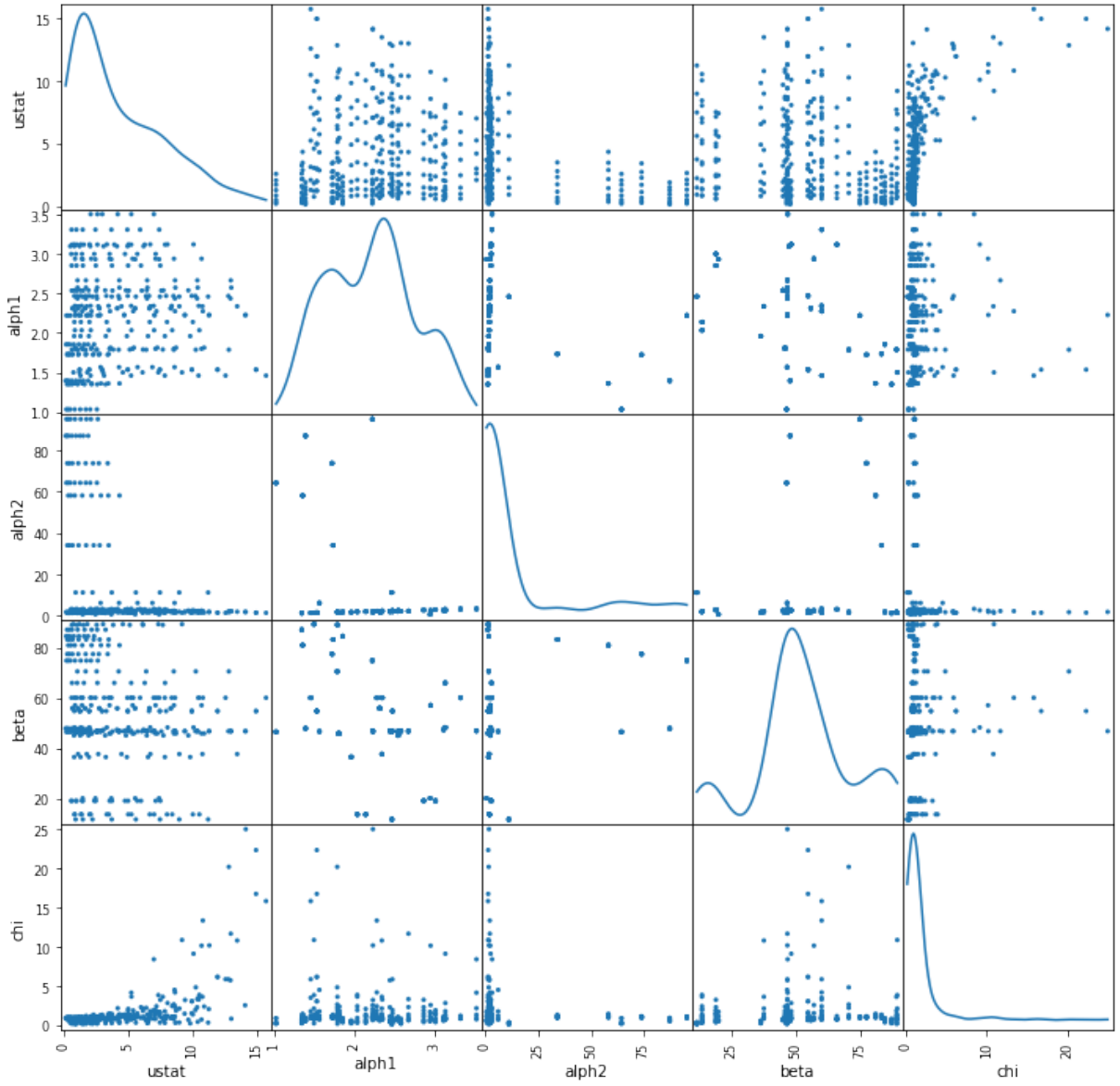
Updates

- Examined updated purged dataset.
- Preprocessed data using Power Transform function.
- Performed k-Means clustering to find new labels.

Exploration of Purged Dataset

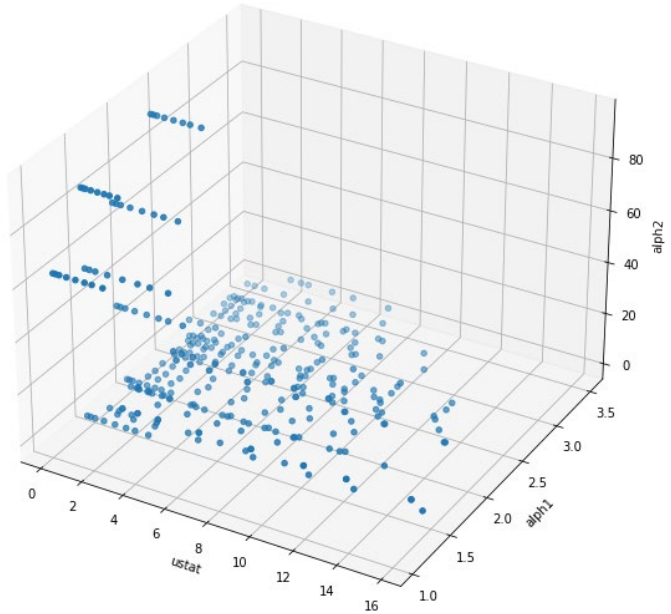
- 344 data instances.
- Distributions are mostly right skewed.
- Correlation and distributions resemble those from previous dataset.



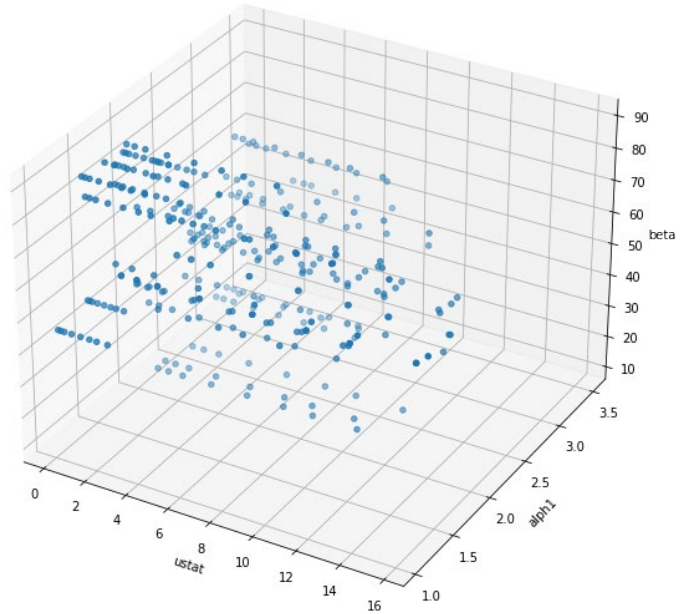


3D Features Scatter

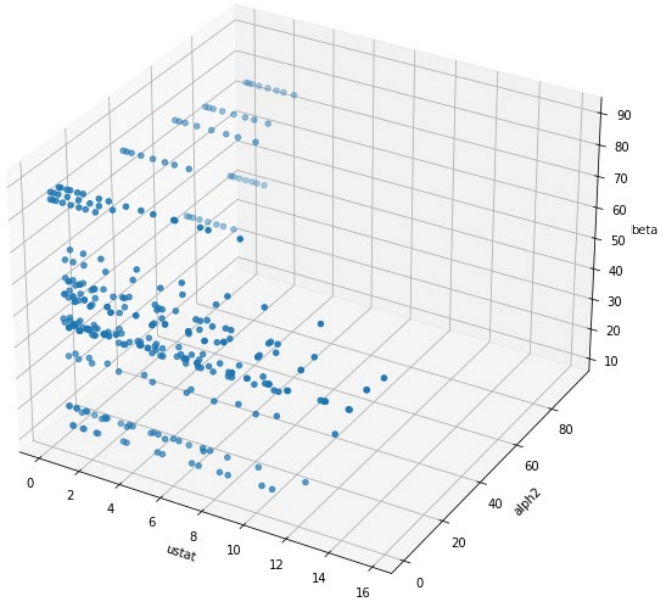
Feature Scatter: ustat, alph1, alph2



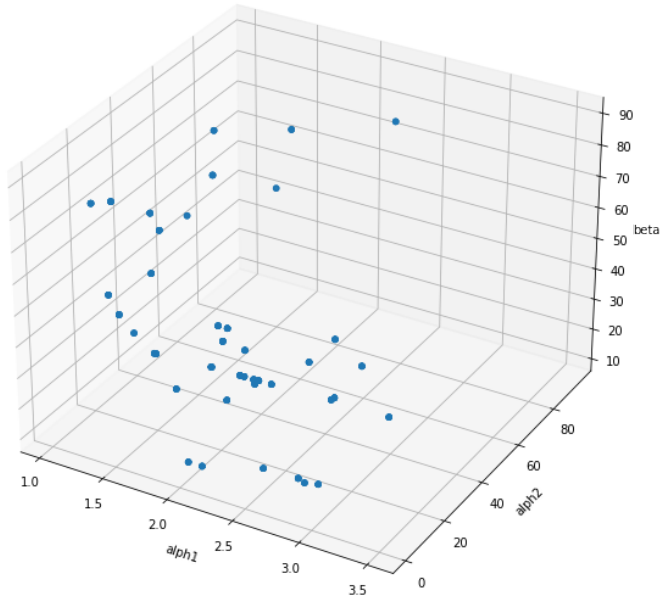
Feature Scatter: ustat, alph1, beta



Feature Scatter: ustat, alph2, beta



Feature Scatter: alph1, alph2, beta

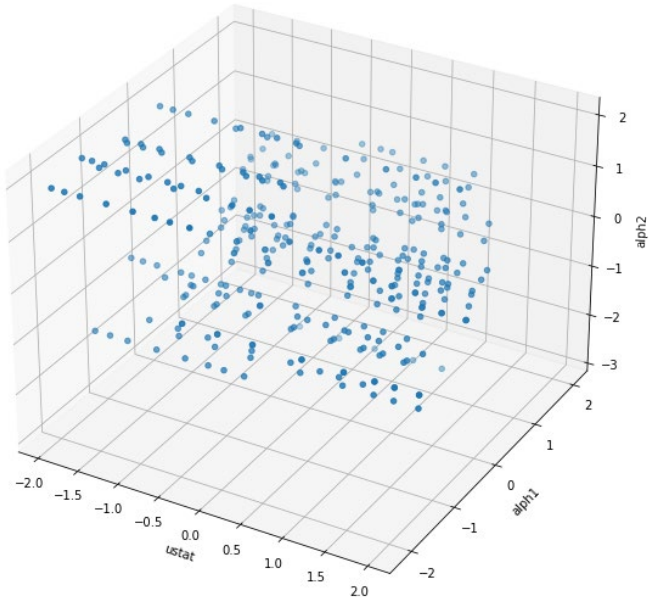


Power Transform

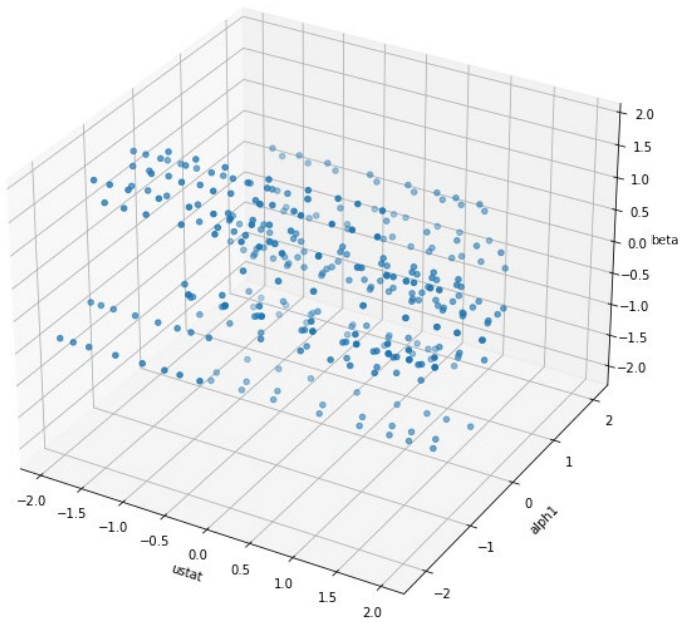
- Class of techniques that use a power function to make the probability distribution of a variable more-Gaussian like.
- Purpose: Increase the symmetry of the distribution of the features.
- Distanced-based models (k-NN, DBSCAN, k-Means) may not work properly if the features are skewed. Power transformations will symmetrize the features without affecting their predictive power too much.
- **Box-Cox Transform:** assumes the values are strictly positive
- **Yeo-Johnson Transform:** does not require the values to be strictly positive

3D Features Scatter (Transformed)

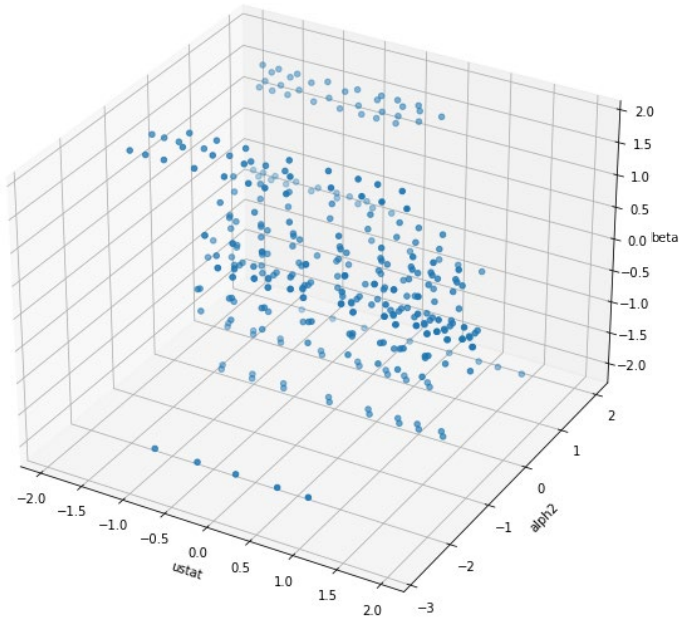
Feature Scatter: ustat, alph1, alph2



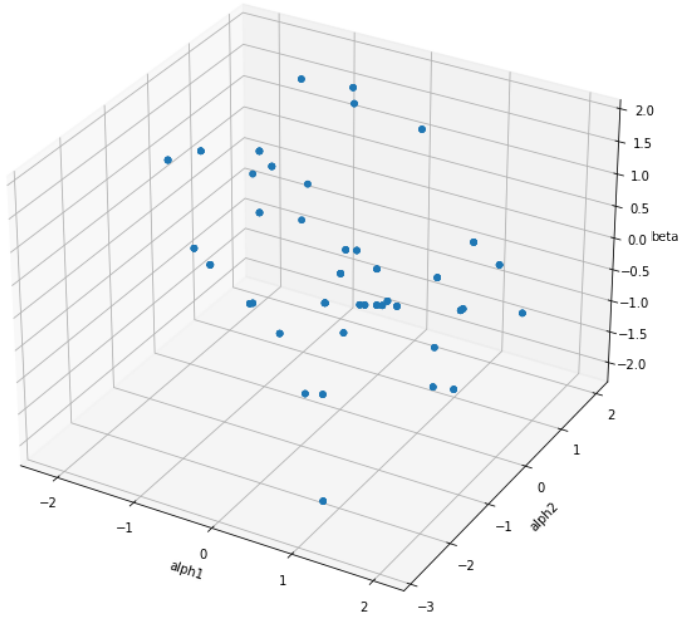
Feature Scatter: ustat, alph1, beta



Feature Scatter: ustat, alph2, beta



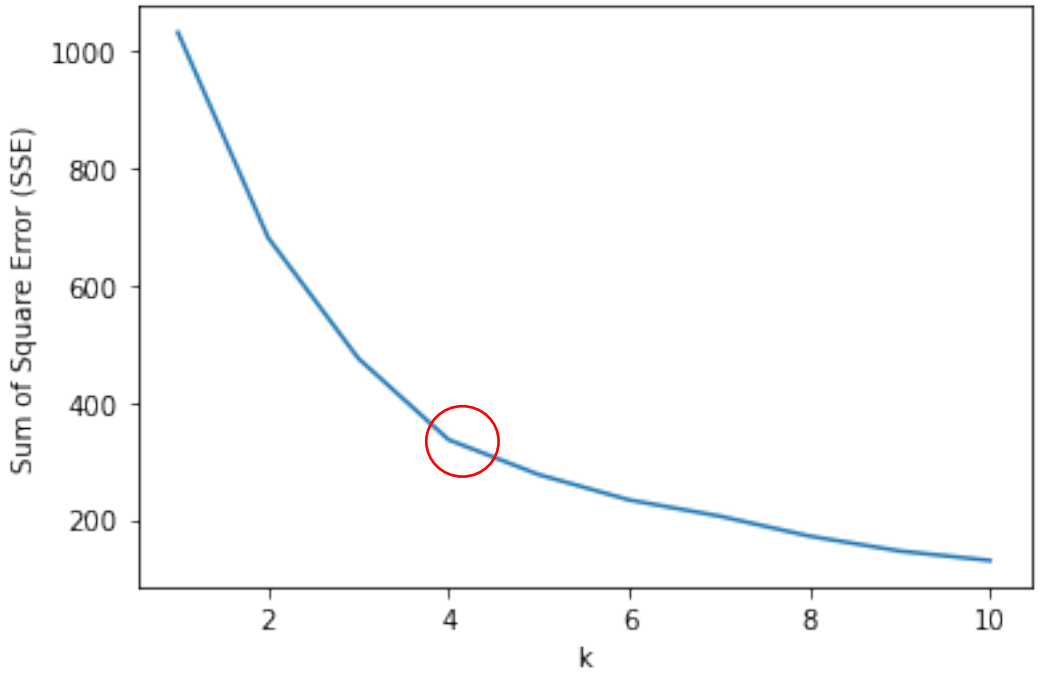
Feature Scatter: alph1, alph2, beta

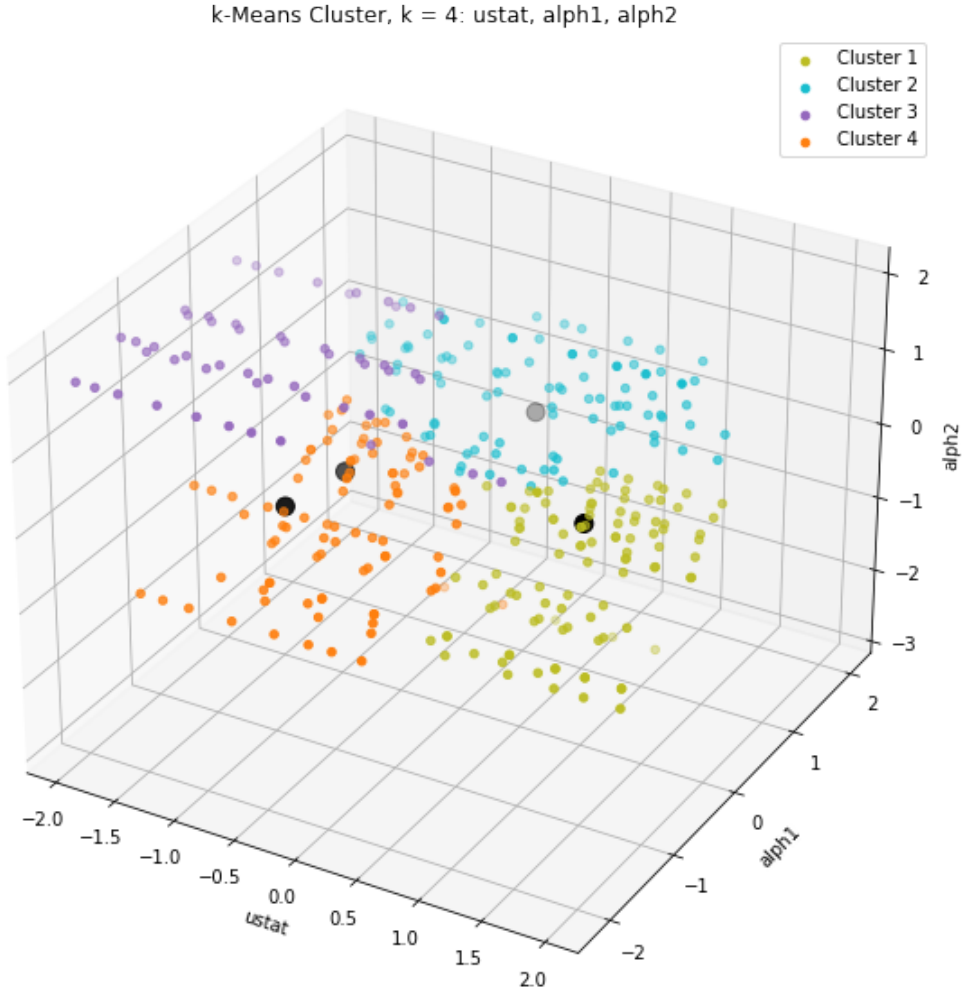


K-Means Clustering

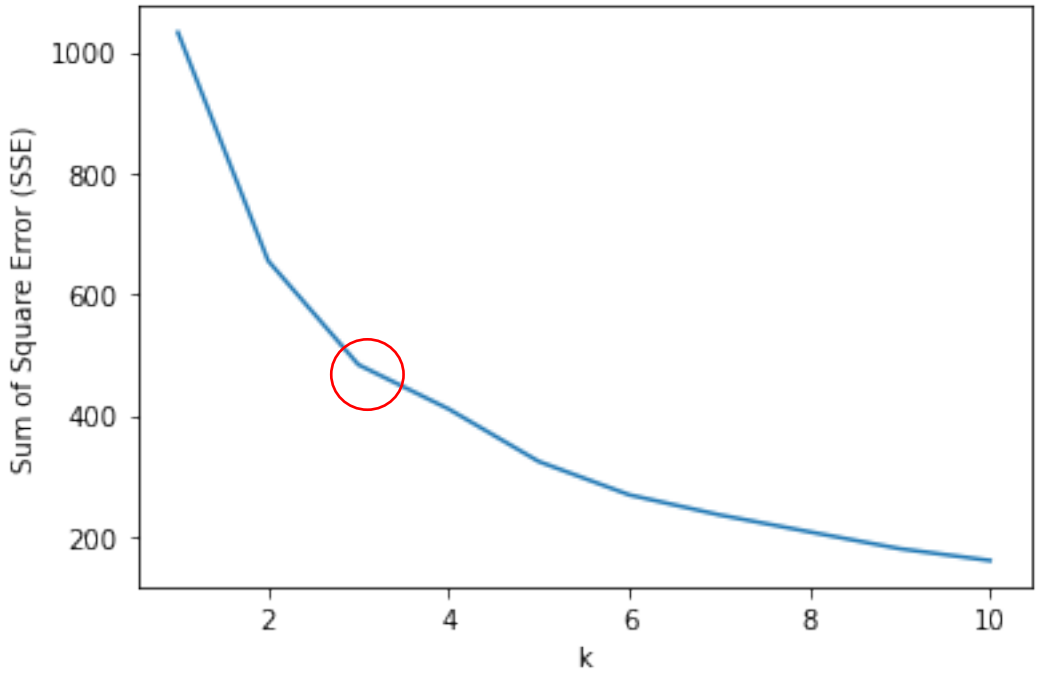
- Groups similar items in the form of clusters. The number of groups is represented by the k value.
  - Iterative Approach: Initialize centroids and assign data instances to centroids to form clusters.
  - Goal: Minimize the variance within each cluster.
- **Elbow Method:** Select the optimal k value.
  - Sum of Squared Error (SSE): squaring each points' distance to its respective clusters' centroid and then summing everything up.
  - When the value of k is 1, the within-cluster sum of the square will be high. As the value of k increases, the within-cluster sum of square value will decrease.
  - Find point (k) where graph rapidly changes and moves parallel to X-axis.

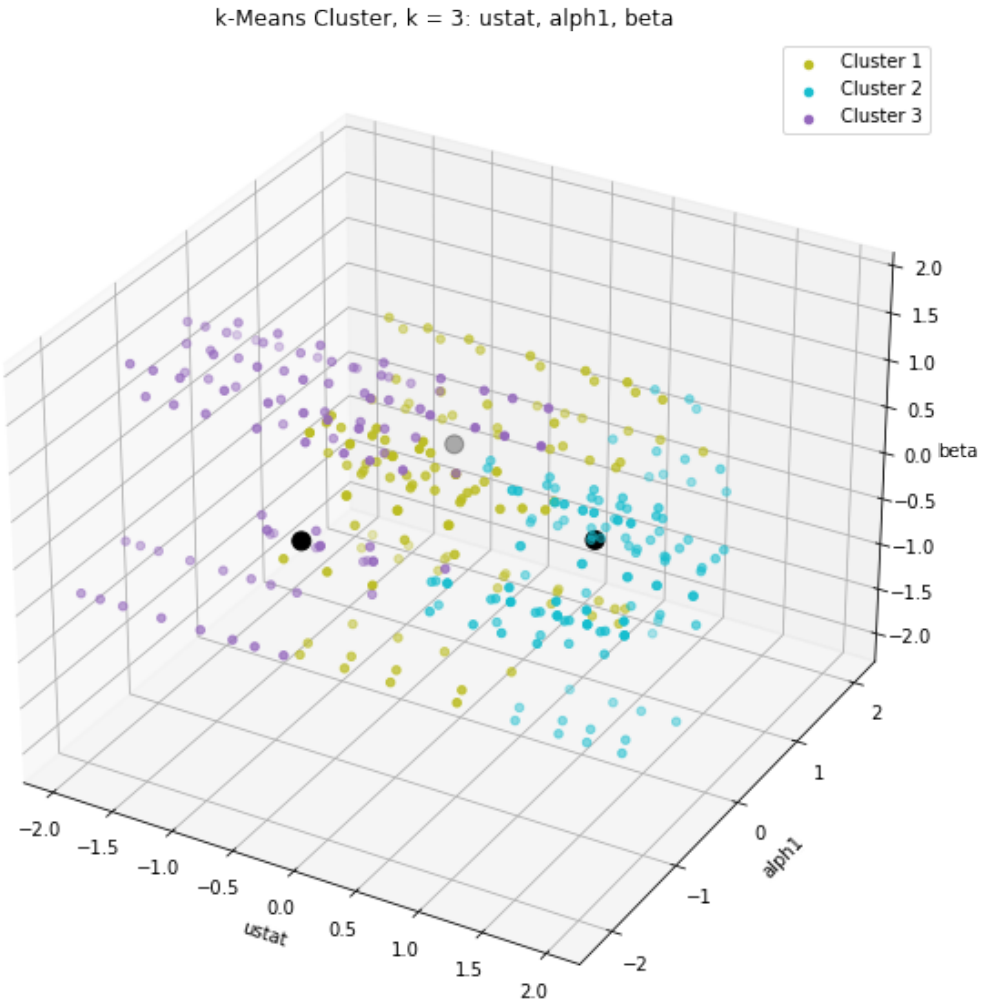
k = 4: ustat, alph1, alph2



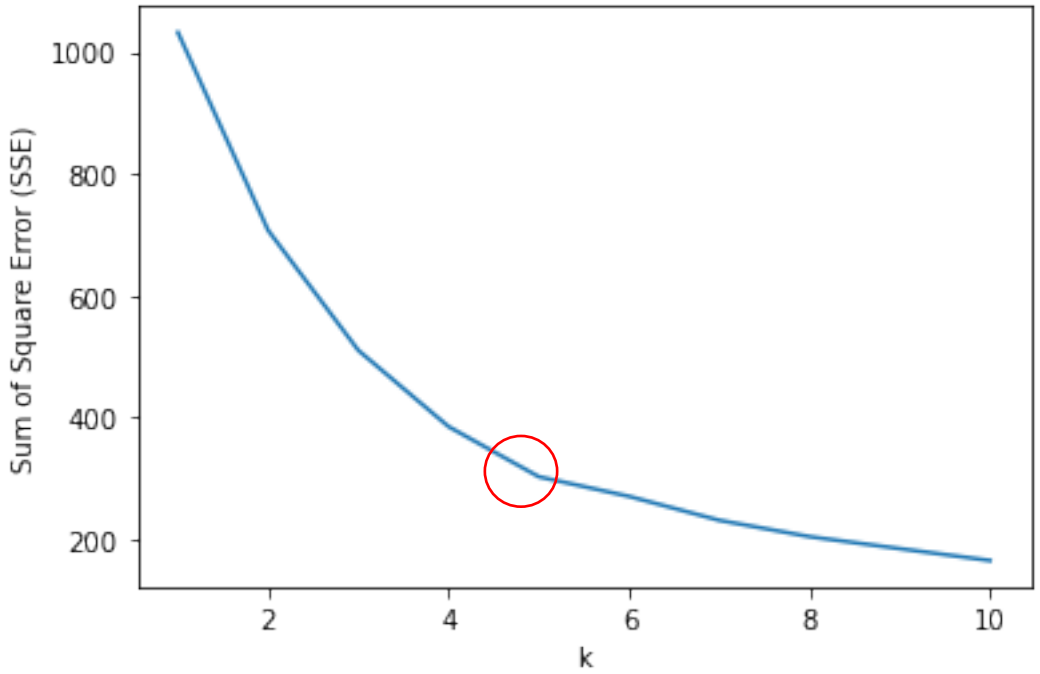


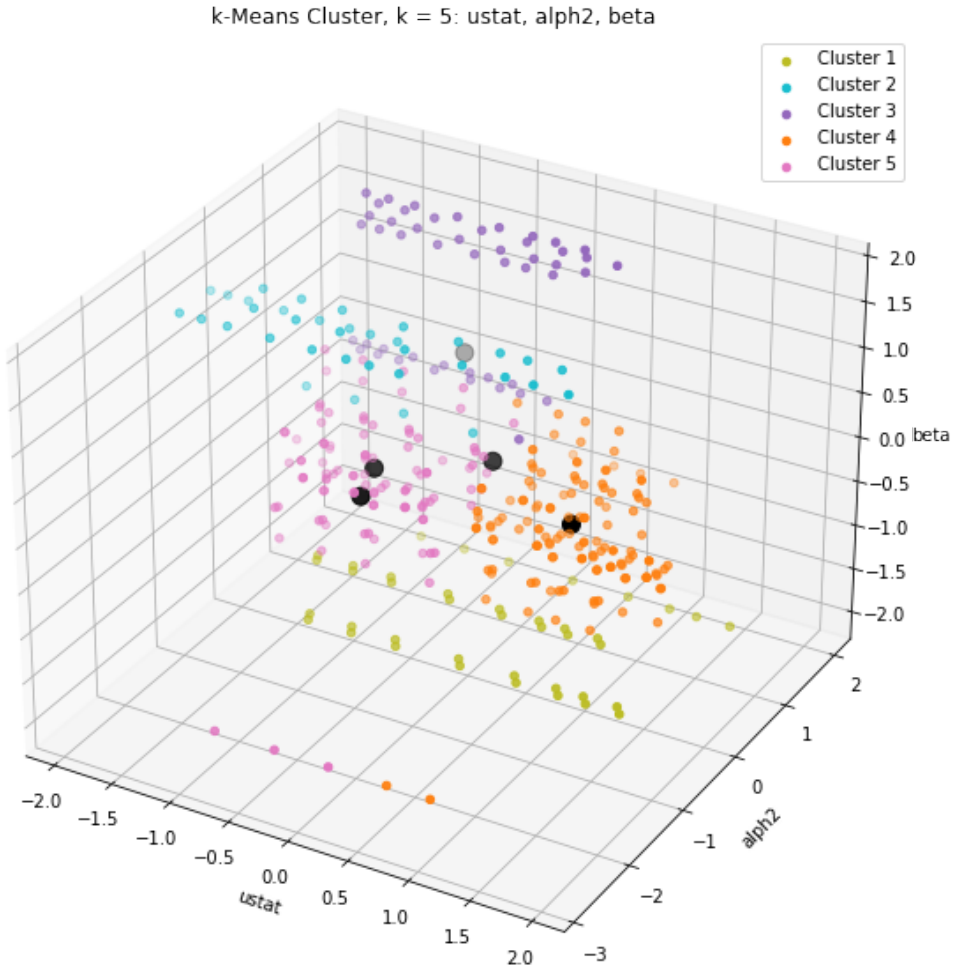
k = 3: ustat, alph1, beta





k = 5: ustat, alph2, beta





k = 3: alph1, alph2, beta