**D212 Performance Assessment Task 1**

Renzo Espinoza

College of Information Technology, Western Governors University

D212: Data Mining II

Dr. Keiona Middleton

January 27, 2023

**A1:Proposal of Question**

The research question that I will answer using k-means clustering is: is the length of a customer’s tenure related to their survey answers where they rated 8 factors based on their personal importance.

**A2:Defined Goal**

In this scenario, I am performing analysis for a telecommunications company that is interested in retaining their customers because it costs 10 times more to obtain customers than keep existing ones. For this reason, the goal of my analysis is to identify the potential features that influence the length of tenure of a customer. The dataset contains each customers responses to a survey where they were asked to rate importance of eight aspects of the company’s services. With this opinion data, I can look for a patterns in their answers that could influence how long they stay with the company.

**B1:Explanation of Clustering Technique**

My chosen clustering technique is k-means clustering. According to Urdan (2022), k-means clustering analyzes datasets by dividing a sample based on the similarity of different clustering variables. The number of groups, or clusters, is k.

This technique will cluster the customers with similar responses to the eight survey variables together. My expected outcome is that a group of customers who highly value certain factors of the company’s services will be identified, and they will have a statistically significant different average length of tenure than the others.

**B2:Summary of Technique Assumption**

According to Thomas Delatte (2020), an assumption of k-means clustering is that each instance of a cluster is more close to its own centroid than the centroids of nearby clusters. This is one of the foundations of the algorithm.

**B3:Packages or Libraries List**

|  |  |
| --- | --- |
| **Library** | **Usage** |
| Numpy and Pandas | Data storage and manipulation |
| Matplotlib and Seaborn | Data visualization |
| Scikit-learn | Implement k-means clustering and principal component analysis |
| Pingouin | Perform ANOVA |

**C1:Data Preprocessing**

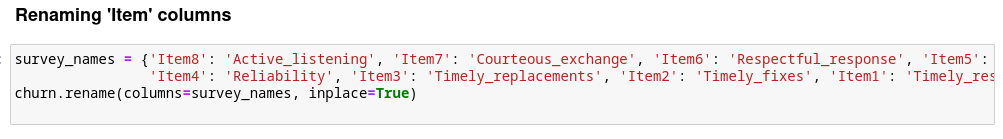
My main data preprocessing goal for this analysis is dimension reduction. Visualizing the resulting clusters of my analysis will be difficult because of the high dimensionality of the independent variables. To achieve dimension reduction I will use principal component analysis.

**C2:Dataset Variables**

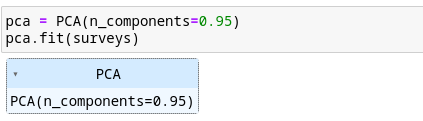
|  |  |
| --- | --- |
| Tenure | Continuous |
| Item1 | Continuous |
| Item2 | Continuous |
| Item3 | Continuous |
| Item4 | Continuous |
| Item5 | Continuous |
| Item6 | Continuous |
| Item7 | Continuous |
| Item8 | Continuous |

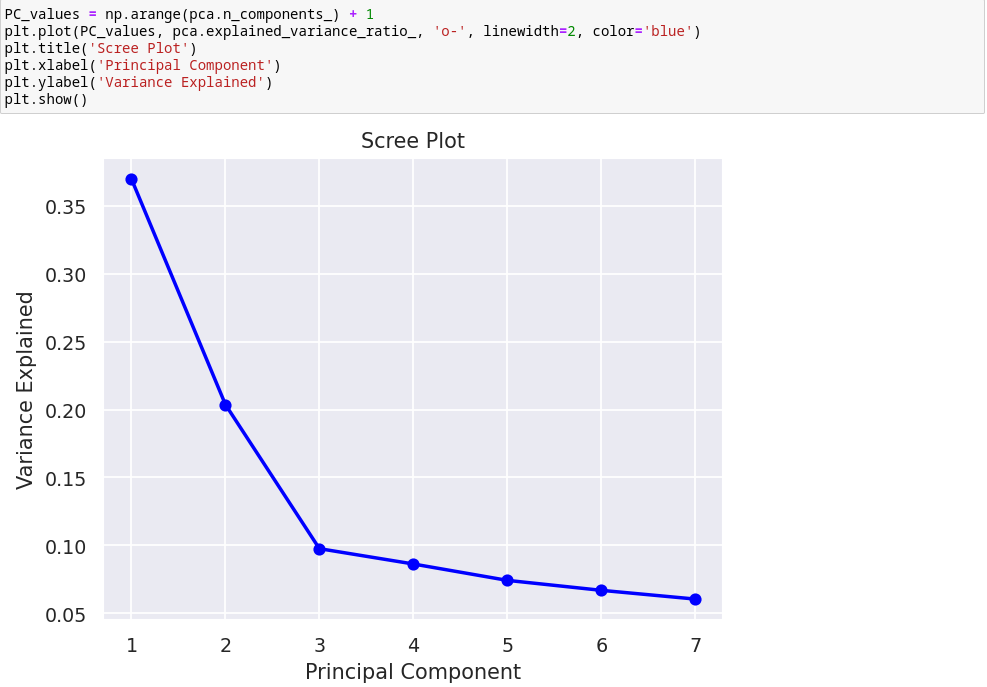
**C3:Steps for Analysis**

My first step to prepare the data for analysis was to rename the independent variables so they are easier to interpret.

 For the next step, I checked for missing values and values outside of the range of 1-8.

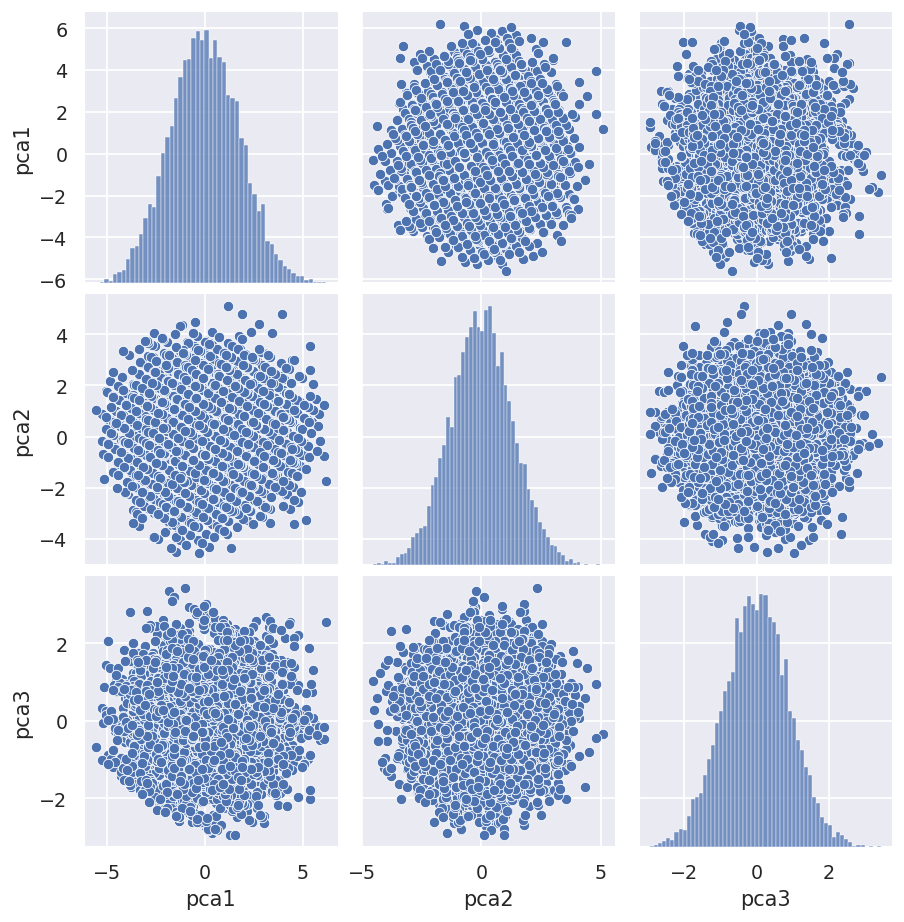
 The variables are all on the same scale of 1-8, so feature scaling was not necessary. My final step for preparing the data for clustering was to perform dimension reduction. For this I used the PCA module from the Scikit-learn library.

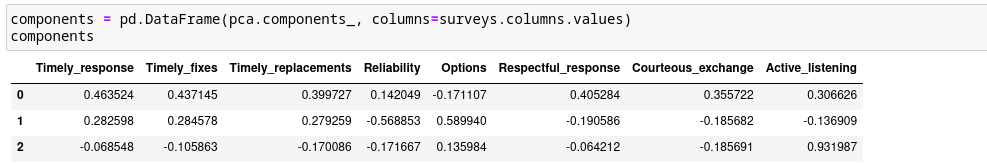


****

After looking at the scree plot, I decided that the number of principal components would be 3.

****

****

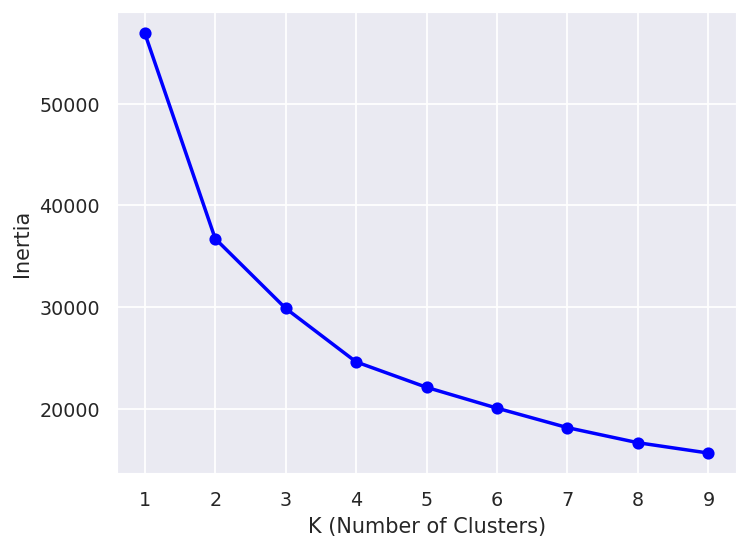
****

**C4:Cleaned Dataset**

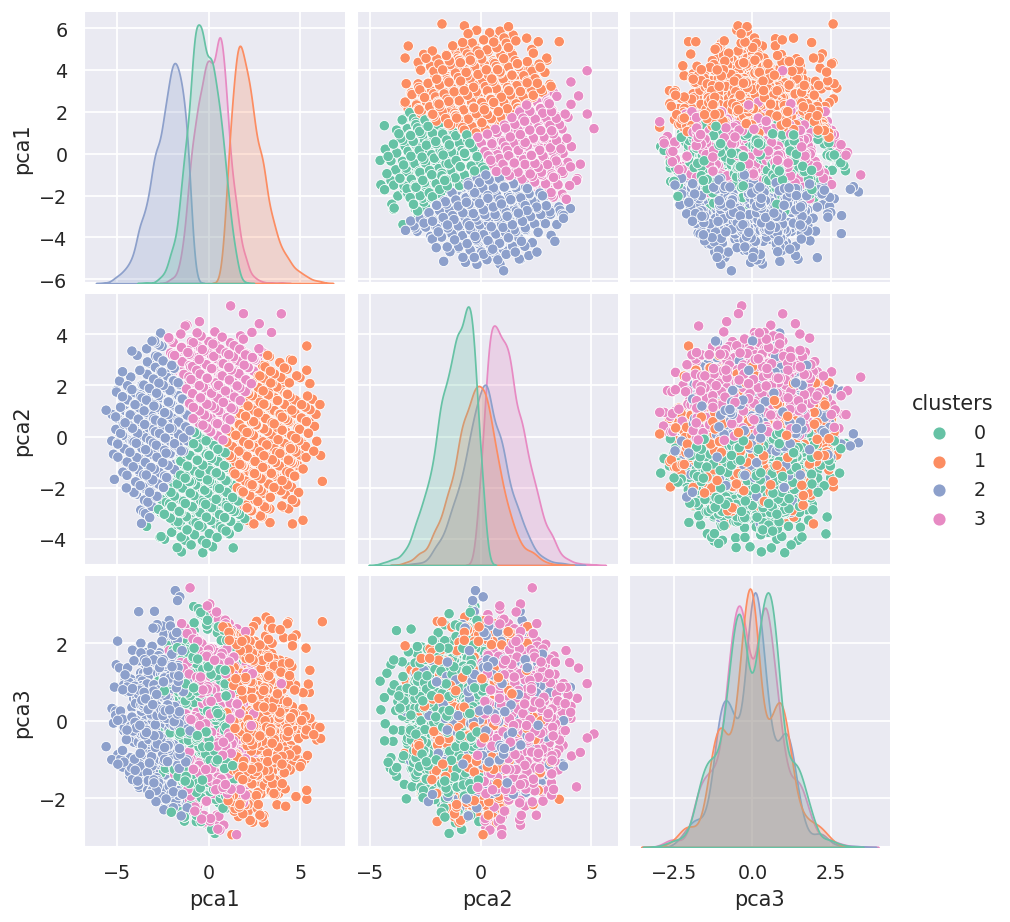
The cleaned dataset will be attached to my submission as ‘PA1\_cleaned\_dataset.csv’.

**D1:Output and Intermediate Calculations**

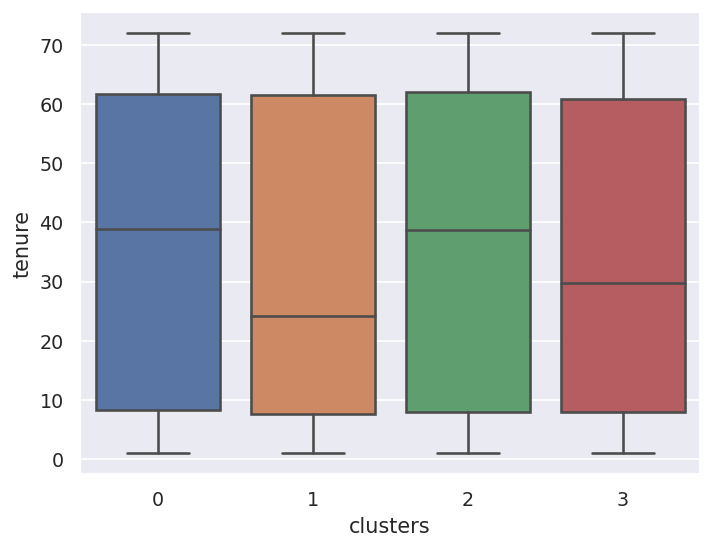
To determine how many clusters to use in my analysis, I plotted the inertia of the k-means model at different amounts of k and tried to find an elbow in the plot.

****

I chose 4 as the amount of clusters for the model. Then I plotted the clustering results and printed the inertia.

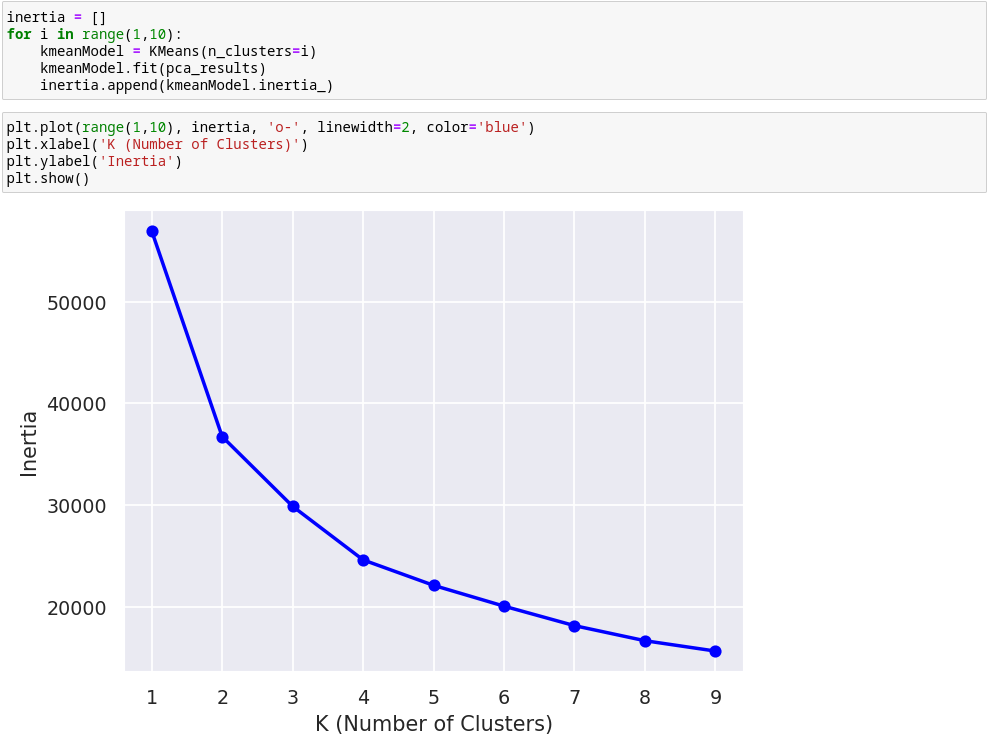
****

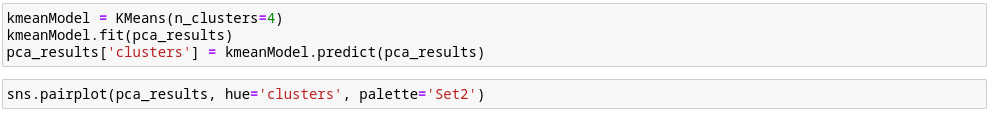
To see if there is a difference between the clusters’ length of tenure, I constructed a boxplot.

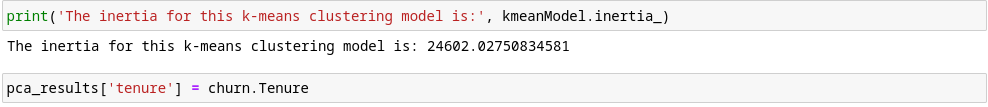
 From this you can see that cluster 1 has the lowest median tenure length, and clusters 0 and 2 have very median values. To check if these clusters have a statistically significant difference from one another, I performed an ANOVA test.

**D2:Code Execution**

For the scree plot:

To create the final model, add the results to a dataframe, and then visualize the results:

To print the inertia of the k-means clustering model and add the ‘Tenure’ column to the dataframe holding the results:

To display the boxplot of each cluster’s tenure distribution:

To perform the ANOVA test:

**E1:Accuracy of Clustering Technique**

For this analysis, I am using inertia as my accuracy metric. Inertia is the sum of the squared distances between each data point and its assigned centroid. An inertia value of 24,602 is very high and suggests the clusters are not very tight.

**E2:Results and Implications**

The ANOVA test results show a p-value of 0.64. This means we cannot reject the null hypothesis using the common significance level of 0.05. Therefore, there is not a statistically significant difference between the clusters’ mean value for ‘Tenure’. This implies that the eight survey answers are not solely related to a customer’s length of tenure.

**E3:Limitation**

One limitation of my analysis is that I performed k-means clustering on the principal components of the eight survey variables instead of the variables themselves. The clustering results on the new variables generated by PCA aren’t directly comparable to the original features and makes interpreting the results difficult.

**E4:Course of Action**

While my analysis unfortunately did not result in uncovering a group features that influence a customer’s tenure, there are still plenty of possible variables that I did not test. My recommended course of action would be to perform further cluster analysis using a different combination of features than the one I used for this experiment.

**F:Panopto Recording**

**G & H: Sources**

Delatte, T. (2020, April 15). *Understanding K-means clustering*. Thomas Delatte. Retrieved January 25, 2023, from https://thomasdelatte.com/2020/04/kmeans/

Urdan, T. C. (2022). *Statistics in plain english* (5th ed.). Routledge.