Assignment

Static Data Analysis

Feature Extraction

We Normalize Numerical Columns to scale the data into a common range. This ensures that no single numerical feature disproportionately influences the model due to its scale. For example, a feature with a larger range could dominate others, skewing the model's learning process. Normalization makes all numerical features comparable and improves the performance and convergence speed of many machine learning algorithms.

Categorical variables represent distinct categories, and assigning arbitrary numerical values to them (like 1, 2, 3) could imply a non-existent ordinal relationship between them. One-hot encoding avoids this by creating binary vectors where each category is represented by a unique bit position. This ensures that the model interprets the categories correctly without introducing any unintended ordinal relationships.

Visual Analysis

In the scatter plot we can see the 2 first columns of U, which corresponds to the principal component with the largest singular value. It captures the direction of maximum variance in the original data.

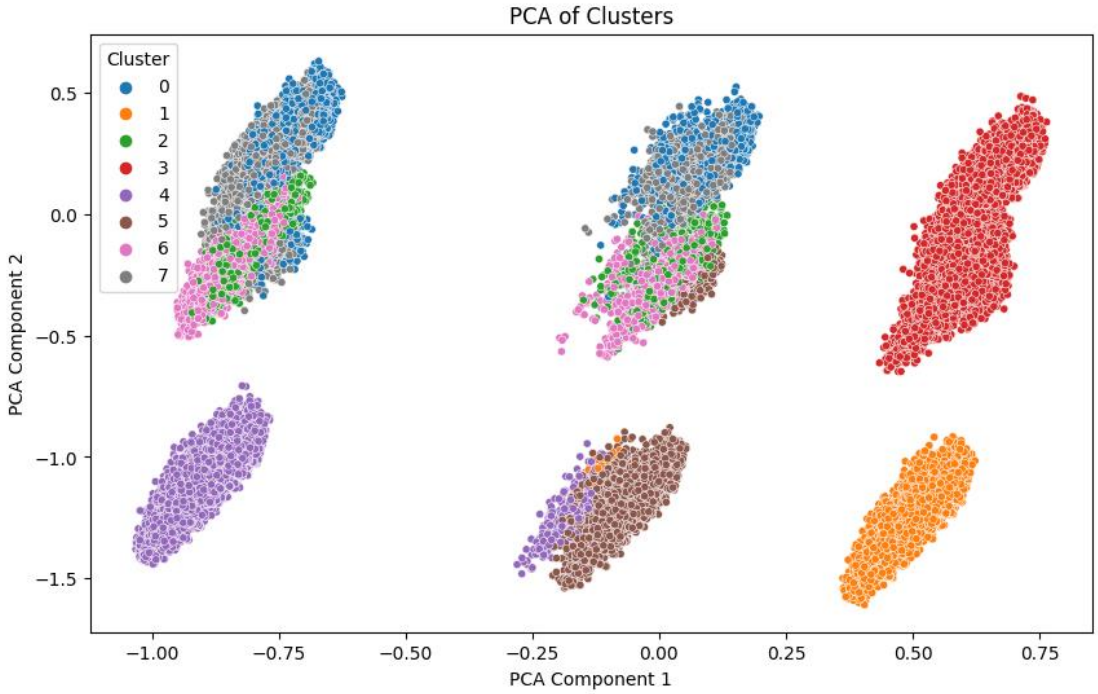
In a dataset where you've applied SVD to reduce the dimensionality to 2 (k=2), U1 and U2 represent the two new dimensions (principal components) in the reduced space. These dimensions are linear combinations of the original features, capturing the most significant patterns in the data.

In the plot we can see 3 clusters.

Visual Clustering

After using PCA for reducing the data points' dimensions, the scatter plot shows varying levels of overlap. Some clusters are tightly packed, while others are more spread out. In the original, higher-dimensional space, the data might have more nuanced distinctions between clusters. When reducing dimensions (like from many features down to just 2 with PCA), you're trying to retain as much of the variance (important information) as possible. However, this often means that some details get compressed, and clusters that were distinct in high dimensions may appear closer or overlapping in the reduced space.

With fewer clusters, each cluster is larger and likely encompasses more variation within it. This can help better capture the "big picture" structure of the data, making it easier to distinguish between clusters, even after dimensionality reduction. Therefore, I would suggest to decrease the number of clusters and more specifically decrease it to k = 6 (we recognize 6 well separated groups at 2D). So that, we might achieve a more meaningful grouping where each cluster is more distinct in the 2D space.



Cluster's Viewing Analysis

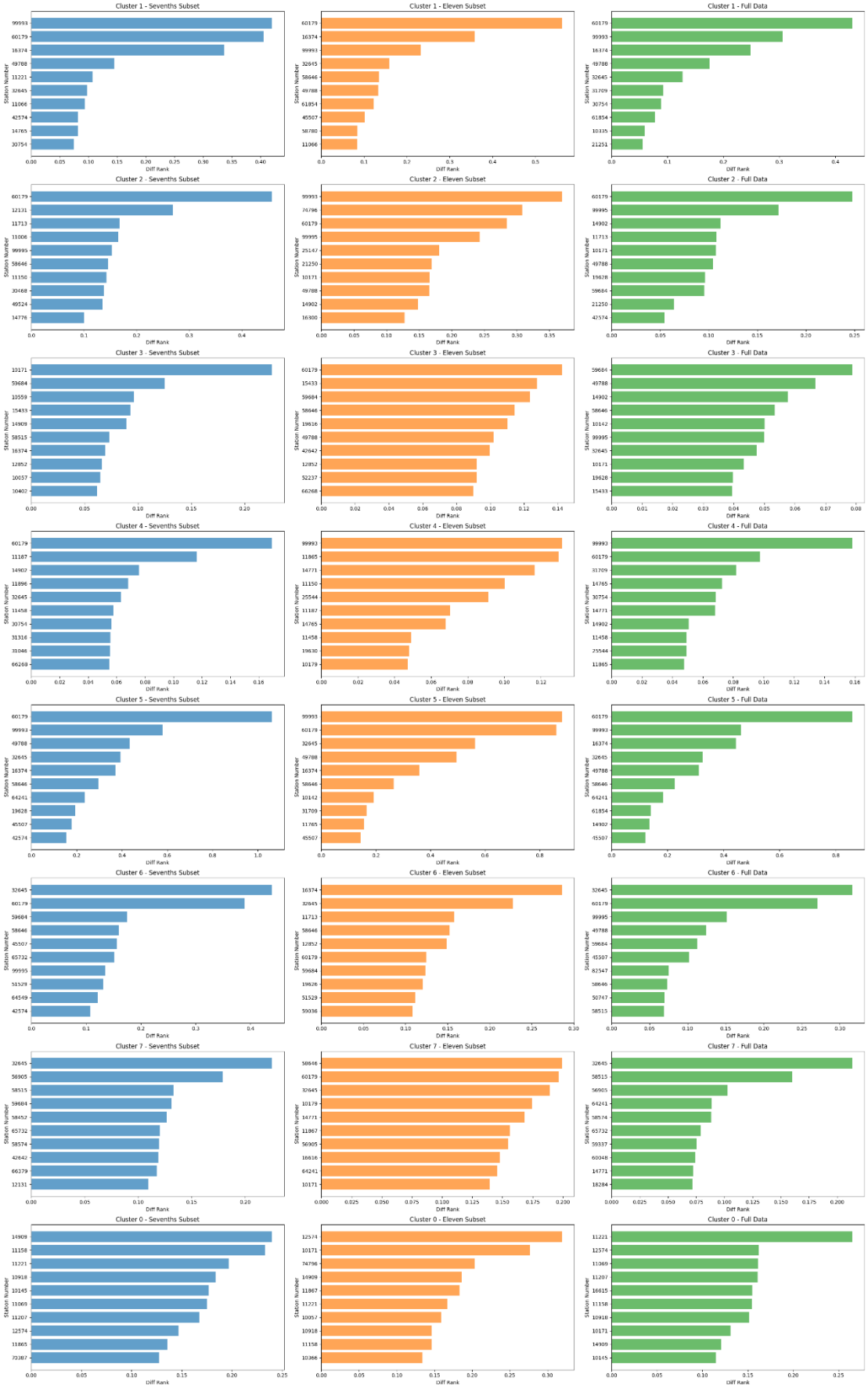
A positive 'diff rank' indicates that the station is more popular within the specific cluster (or subset of the cluster) compared to the general population. This suggests that the viewing habits of this cluster are different from the average household, showing a preference for this station. The opposite is true for a negative diff rank.

The results across different subset types within the same clusters show consistency. The stations with the highest 'diff rank' are prominently featured in all subsets, indicating that even when selecting 1/7 or 1/11 of the data, the proportion of preference remains similar. This suggests that as the data volume increases, the proportion of station preference remains stable.

The results across different clusters are less consistent. Generally, the 'diff rank' values are low, frequently falling below 0.8, which suggests there is no strong preference for any particular station across the clusters. For instance, cluster 1, which was considered favorable in the previous section, has a lower 'diff rank' compared to more dispersed clusters.

Overall, the dimensionality reduction using PCA and the chosen number of clusters do not provide clear insights into specific viewing preferences within the clusters. The data does not reveal distinct viewing preferences for specific stations among the different clusters.

Top 10 Stations by 'Diff Rank' for Each Cluster and Subset



Dynamic Data Analysis – Streaming

In the results, we can see various levels of similarity between the triggers for each cluster but the general similarity level is quite high. At most clusters, the different triggers shared the most of top-10 stations with order's change. The similarity observed suggests that the viewing patterns within a cluster's subsets remain relatively consistent over time. As the data accumulates across triggers, the viewing preferences of households in each cluster become more pronounced.

A addition significant observation is the decrease in 'diff rank' as we progress through the triggers. It suggests that the relative connection between each cluster and its preferred stations is weakening, i.e., the stations preferences in each cluster are gradually aligning with the general population's preferences. As more data accumulates across triggers, the initial distinctions between cluster-specific viewing habits and general trends might blur, causing the 'diff rank'—which measures how different a cluster's station preference is from the general population's—to decrease. Essentially, the unique viewing patterns within clusters become less distinct as the dataset grows, reflecting a convergence towards broader viewing trends and lower relative connection between clusters and its preferred stations.

**The multi-panel visualization is designed for convenience, allowing for optimal data analysis. The code for generating this panel is included as a comment in the IPython Notebook file we submitted.*