STATIC DATA ANALYSIS:

Feature Extraction:

Why do we normalize numerical columns?

1. We normalize numerical columns because later we'll use PCA (in visual analysis) and clustering algorithms that rely on Euclidean distance. Without normalization, features with larger scales may dominate the distance calculation, leading to biased results. Normalizing ensures that all features contribute equally.

2. After normalization, each value is scaled relative to the minimum and maximum values of its respective feature, placing it on a comparable scale with other features.

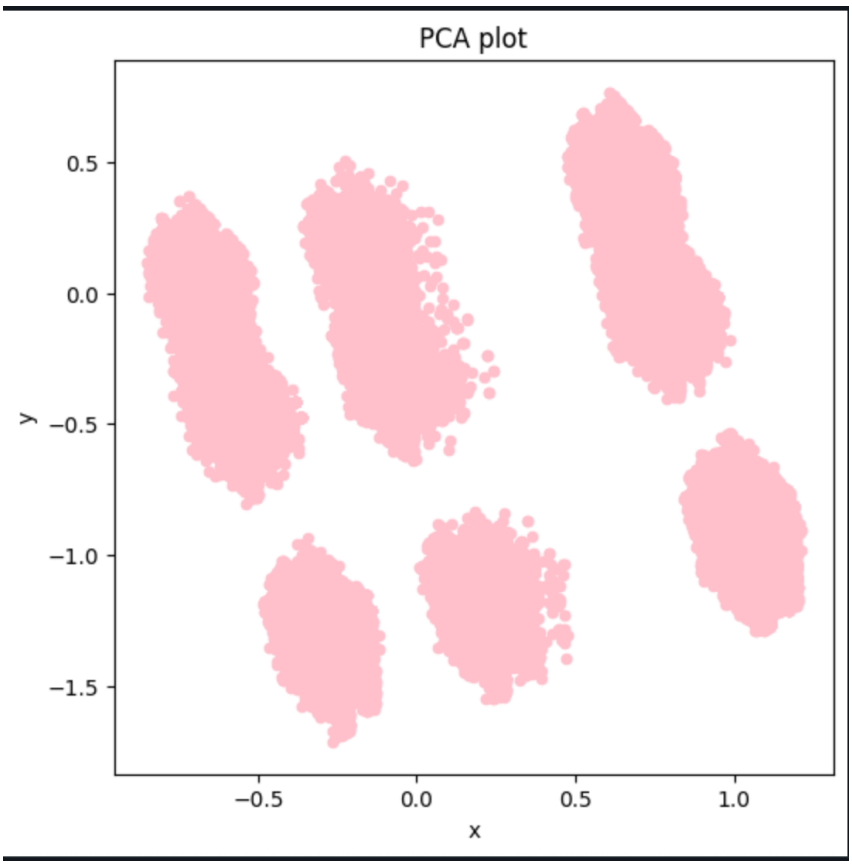
Why do we use one-hot encoding instead of assigning numerical indices to the categories?

One-hot encoding creates a binary vector where the number of dimensions equals the number of possible categories. For each instance, the corresponding category is represented by a 1 in its position and 0s in all others. This approach is useful when using algorithms that rely on Euclidean distance, as it ensures that all categories are equally distant from each other. In contrast, assigning numerical indices can introduce unintended ordinal relationships between categories, which may mislead the model.

Visual Analysis:

What do you observe in the scatter plot? How many clusters do you think are visible? We will denote the number of clusters you observed with c

As clearly seen in the scatter plot, we can see that there are 6 dividing clusters .



Cluster's Viewing Analysis

What does a positive or negative diff_rank mean for a station?

What insights can it give us about the viewing prefrences of a specific cluster compared to the general population?

diff_rank is defined as:
diff_rank = subset_rating - general_rating
Where:

- subset_rating is the percentage of viewers within a specific cluster (or subset) who watch a given station.
- general_rating is the overall percentage of viewers who watch that station, regardless of cluster.

A **positive diff_rank** (i.e., diff_rank > 0) indicates that a station is more popular within the specific subset than in the general population. This suggests that the station aligns well with the preferences or interests of that particular cluster.

Conversely, a **negative diff_rank** (i.e., diff_rank < 0) suggests that the station is less relevant or appealing to the subset compared to the general public. This can indicate a mismatch between the station's content and the preferences of that specific population.

Compare the same subset types across clusters, and also compare different subset types within the same cluster

Do we observe meaningful differences?

In general, we want to explore whether households that belong to the same cluster share similar preferences in TV stations.

For example, station **16374** appears in all three subsets with only slight differences in diff_rank, suggesting consistent viewing behavior across these subsets.

On the other hand, when we look at station **12131** in **Cluster 2**, we notice a significant difference in diff_rank between the **Full** and **17th** subsets (1.08 and 0.98, respectively) compared to the **3rd** subset (1.6). This indicates that, even within the same cluster, the **3rd** subset had a noticeably stronger preference for that station compared to the other subsets.

Does the clustering appear to reflect actual differences in content preference ?

Yes, let's look again at station **12131**. For the **17th** subset in **Cluster 0**, we observe a diff_rank of **0.19**, whereas in **Cluster 2**, the same subset has a diff_rank of **0.98**. This shows that the same station, even for the same subset type, can have significantly different values across clusters—indicating varying preferences within similar household types depending on the cluster.

Project Part B - Streaming

- Compare the results across clusters for each batch
  We can notice that most of the time , the clusters drawn to similar stations, and also the rank stays the same (not always but a lot).
- We can also notice that some clusters has a larger "diff_rank", that means the household in this cluster are preferred this station.