Static Data Analysis

Feature Extraction:

Answer:

The reason that we normalize the numerical columns is so that our values in the data is bounded and to represent it in a graph, since if we don't normalize theoretically the data is not bounded and if try to show it in a graph then the representation could get stretched and it would be hard to infer from the graph. Furthermore, we might want each feature to have equal impact on the representation and therefore each feature needs to be bound in the same range.

Similarly, for categorical columns, firstly, it's not enough to convert to numerical values since we'll have the issue that the range is unbound.

Furthermore, if we convert the categorical columns directly to numerical values, we can get misleading relationships since we assign a different value to each category which we might say one category is better/lesser than another category which is not necessarily true. Thus, converting to vectors would be better since each unique category will get its own binary feature such that each category is independent and equidistant and therefore keeping a proper preservation of the original data.

Visual Analysis:

We can see in the scatter plot a distribution of points where each point represents a household. We can clearly see that the graph has two distinct clusters. Thus, we can see in the scatterplot that we have two different groups/patterns of households given the clusters.

Visual Clustering:

We can see 6 distinct clusters in the scatter plot where each color is a different cluster that a point belongs to. We can get a better color plot division if we lower the number of clusters from 8 to 6 since we can see 6 distinct clusters on the graph and therefore it makes sense for this specific graph result that there are 6 clusters rather than 8.

Cluster Viewing Analysis:

Positive value in 'diff_rank' of a station for a cluster would tell us that the specific subset that we're looking at has a higher popularity rating relative to the popularity rating of all the data combined.

Negative value in 'diff_rank' is the exact vice versa. Meaning that if the values are close to zero then we can infer that the viewing habits of the ppl in that cluster are like the average station. If a lot further away from 0, for example, close to 100 (-100) then that station is a lot higher (lower) rating/popularity (which could mean watching more shows or popular shows for that station) compared to the average station viewing habits.

Average View Percentage:

For the metric showing the average view percentage for each cluster, the cluster with the highest value identifies the cluster with the most engagement, that reveals to us which cluster has the highest interest in specific stations. The middle value represents an average level of engagement, providing a benchmark for average cluster behaviour. The lowest value highlights the cluster with the least engagement, indicating lower interest in those stations.

Average Rank Difference:

The metric showing the average difference between a cluster's view percentage and the general benchmark helps assess how unique a cluster is. The highest value indicates the cluster with the highest positive std from the general benchmark, suggesting that it has a notably distinct preference in relation to the average population's viewing habits. The middle value represents an average std, reflecting a typical difference from the benchmark. The lowest value shows the cluster with the least positive or most negative std, indicating a cluster with viewing habits most like or less favourable than the general population.

Standard Deviation of View Percentage:

For the metric capturing variability in view percentages within each cluster, the highest value denotes the cluster with the most fluctuation in engagement, indicating high diversity in viewing habits within that cluster. The middle value shows average variability, representing a typical level of consistency in viewing habits. The lowest value identifies the cluster with the most consistent engagement, suggesting less variation in viewing behaviour across different stations within that cluster.

These metrics help determine if the clusters we identified have distinct viewing habits, validating the effectiveness of our clustering approach. By comparing the 'diff rank' measure, we can assess how each cluster's viewing preferences differ from the general population and evaluate the meaningfulness of our clusters in capturing unique viewing behaviours.

Each subgroup represents its own cluster. The ratio between the rate of views between all 7 subgroups (between the different clusters) is similar to the ratio between the rate of views between all 11 subgroups between the different clusters, and is similar to the ratio between the rate of views between the different clusters themselves.

In addition, the viewing rates between the different subgroups within the same cluster are similar.

From this it can be concluded that the division into clusters is indeed representative, and also that the sampling of the subgroups is good enough. Thus, each subgroup really represents the whole cluster, from which it was taken.

Using code here are the statistics calculated for each subset group:

<u>Sevenths Results</u>

- **Average View Percentage**
  - **Highest: Cluster 0 with a value of 0.07457121551081283**
  - **Middle: Cluster 3 with a value of 0.05219206680584552**
  - **Lowest: Cluster 2 with a value of 0.04474272930648769**
- **Average Rank Difference**
  - **Highest: Cluster 0 with a value of 0.003730304883270004**
  - **Middle: Cluster 1 with a value of 0.00032378425847299613**
  - **Lowest: Cluster 2 with a value of 8.022534457184039e-05**
- **Standard Deviation of View Percentage**
  - **Highest: Cluster 0 with a value of 0.2790316701994497**
  - **Middle: Cluster 4 with a value of 0.21148151216782468**
  - **Lowest: Cluster 2 with a value of 0.20662652581793584**

Elevenths Results

- **Average View Percentage**
  - **Highest: Cluster 0 with a value of 0.08718395815170005**
  - **Middle: Cluster 3 with a value of 0.05803830528148577**
  - **Lowest: Cluster 2 with a value of 0.047687172150691466**
- **Average Rank Difference**
  - **Highest: Cluster 0 with a value of 0.005805506731320231**
  - **Middle: Cluster 3 with a value of 0.0005921444137890031**
  - **Lowest: Cluster 2 with a value of 0.00016773857665559476**
- **Standard Deviation of View Percentage**
  - **Highest: Cluster 0 with a value of 0.2846552020454802**
  - **Middle: Cluster 4 with a value of 0.22173434329984332**
  - **Lowest: Cluster 3 with a value of 0.21397350755876457**

Full Results

- **Average View Percentage**
    - **Highest: Cluster 0 with a value of 0.04800768122899664**
    - **Middle: Cluster 1 with a value of 0.04001600640256105**
    - **Lowest: Cluster 2 with a value of 0.03762227238525205**
- **Average Rank Difference**
    - **Highest: Cluster 0 with a value of 0.0002922443893280914**
    - **Middle: Cluster 1 with a value of 1.751327751211325e-05**
    - **Lowest: Cluster 2 with a value of 4.275576822975792e-06**
- **Standard Deviation of View Percentage**
    - **Highest: Cluster 0 with a value of 0.2236639645428829**
    - **Middle: Cluster 6 with a value of 0.1896351898283524**
    - **Lowest: Cluster 1 with a value of 0.184020177071707**

Using statistical calculations on the df's that we found throughout the project here are the results regarding the clusters given different subsets.

The numbers printed for the highest, middle, and lowest clusters represent the **average rank_diff** within each subset (full, sevenths, or elevenths) for a particular cluster (centroid).

- **Interpretation**: The value of avg_rank_diff indicates the cluster (within a particular subset) that has the most pronounced negative/positive/average difference in viewing preferences compared to the general population.
- **What It Represents**: This number shows how much less/more/around (on average) the cluster favours certain stations compared to the general viewing habits. A lower/higher/same value suggests that the cluster is less/more/is interested in stations that are generally popular, indicating a divergence from mainstream preferences.

Cluster 0

- **Highest:** Subset *elevenths* with a value of **0.006116228711717073**
- **Middle:** Subset *sevenths* with a value of **0.00335894301450459**
- **Lowest:** Subset *full* with a value of **0.0002922443893280915**

Cluster 1

- **Highest:** Subset *elevenths* with a value of **0.0006195528319193443**
- **Middle:** Subset *sevenths* with a value of **0.00036779899041705313**
- **Lowest:** Subset *full* with a value of **1.7513277512113305e-05**

Cluster 2

- **Highest:** Subset *elevenths* with a value of **0.00017490784361082502**
- **Middle:** Subset *sevenths* with a value of **8.61818327995327e-05**
- **Lowest:** Subset *full* with a value of **4.2755768229758135e-06**

Cluster 3

- **Highest:** Subset *elevenths* with a value of **0.0007396059289912993**
- **Middle:** Subset *sevenths* with a value of **0.00033831289031265277**
- **Lowest:** Subset *full* with a value of **2.1590783266098677e-05**

Cluster 4

- **Highest:** Subset *elevenths* with a value of **0.00034526105308819755**
- **Middle:** Subset *sevenths* with a value of **0.00016328565491403147**
- **Lowest:** Subset *full* with a value of **9.964509047706438e-06**

Cluster 5

- **Highest:** Subset *elevenths* with a value of **0.000293305463378187**
- **Middle:** Subset *sevenths* with a value of **0.00020256245249885313**
- **Lowest:** Subset *full* with a value of **1.2042837741329807e-05**

Cluster 6

- **Highest:** Subset *elevenths* with a value of **0.0003444971322411404**
- **Middle:** Subset *sevenths* with a value of **0.00021769167750667316**
- **Lowest:** Subset *full* with a value of **1.2420200213627749e-05**

Cluster 7

- **Highest:** Subset *elevenths* with a value of **0.0018159906871739926**
- **Middle:** Subset *sevenths* with a value of **0.000963656988696112**

- **Lowest:** Subset *full* with a value of **7.329639475749983e-05**

In regard to whether we see any values in the clustering done earlier, in a way we do because the statistical values that we found are based on the clustering thus being the values that we are checking. We're essentially finding different ways to represent the data that was clustered and capture the patterns in the data whether in all the data or the subset groups that we built.

Dynamic Data Analysis - Streaming:

Looking throughout clusters as we move through processing more batches, we can see that the rank_diff remains more or less consistent throughout the same cluster as we process more batches. For example, if we look at the rank_diff of centroid 0 for the first and third batch there are multiple values around 0.35, and even after adding more data dynamically we can still see similar behaviour given the demographic between the same clusters in different batches.

Furthermore, each subgroup represents its own cluster. The ratio between the rate of views between all 7 subgroups (between the different clusters) is similar to the ratio between the rate of views between all 11 subgroups between the different clusters and is similar to the ratio between the rate of views between the different clusters themselves.

In addition, the viewing rates between the different subgroups within the same cluster are similar.

From this it can be concluded that the division into clusters is indeed representative, and that the sampling of the subgroups is good enough. Thus, each subgroup really represents the whole cluster, from which it was taken.