**1 cluster on Driver Demographics & Risk Profiles**

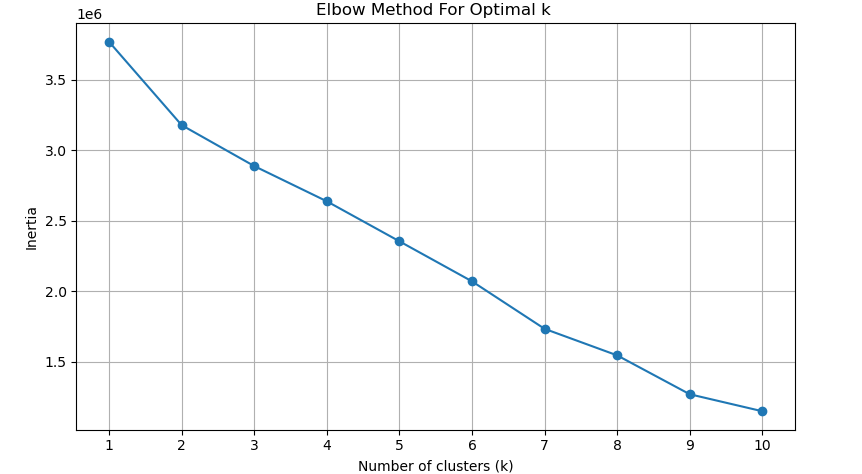
Based on the statistics obtained from the previous steps, we can directly perform cluster analysis on Driver Demographics & Risk Profiles. The required data attributes include: 'AGE\_GROUP', 'SEX', 'ROAD\_USER\_TYPE\_DESC', and 'INJ\_LEVEL', which can be directly extracted from the df\_person dataset.

**1.1 Clustering Step-by-Step Procedure**

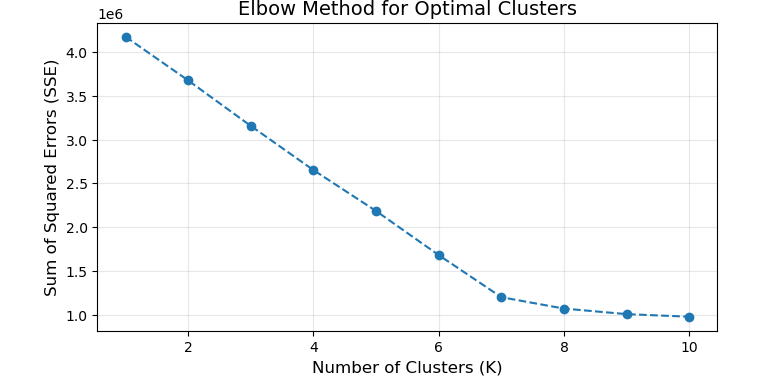
1. Extract data from the dataset.
2. Encode each feature.
3. Normalize the encoded data
4. Plot an elbow curve using the normalized data to select the cluster count
5. Conduct K-means clustering
6. Reduce data dimensionality
7. Plot a scatter plot of the clusters (after dimensionality reduction)
8. Generate a correlation heatmap

**1.2 Improvements in the code implementation process**

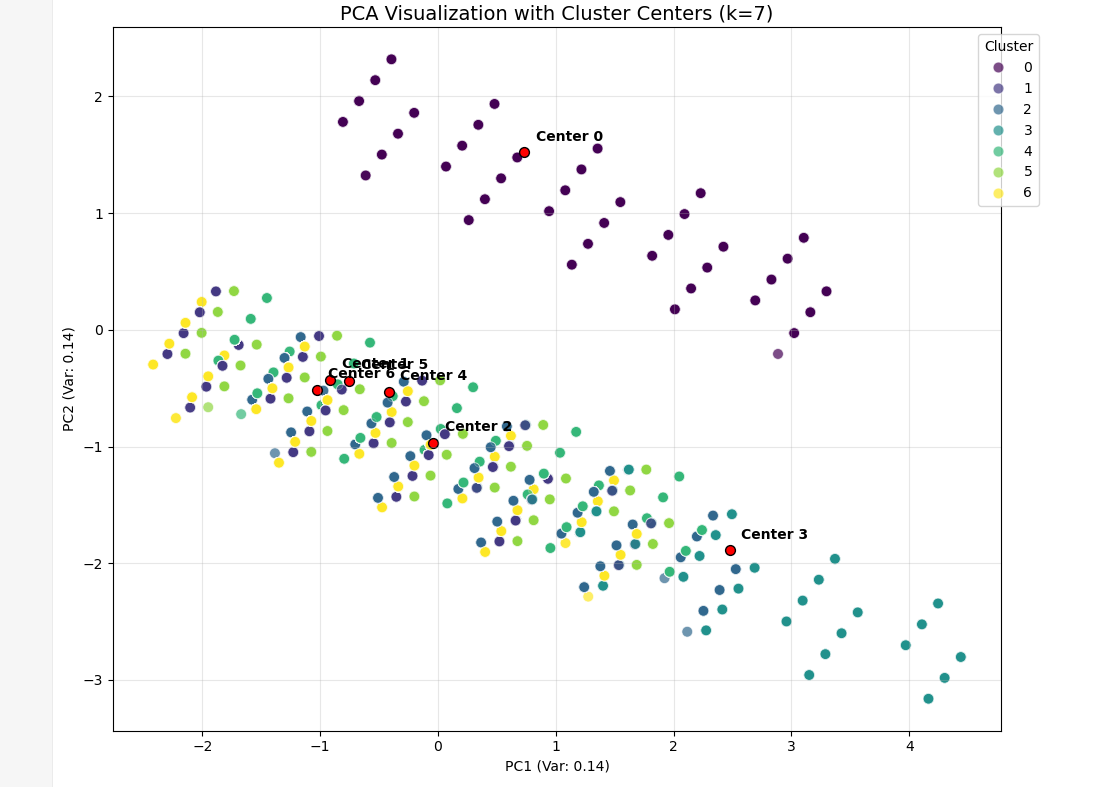
1. Adjust Encoding Methods
   1. Initially, one-hot encoding was applied to all features, but the resulting elbow plot showed an unclear inflection point (as illustrated below).



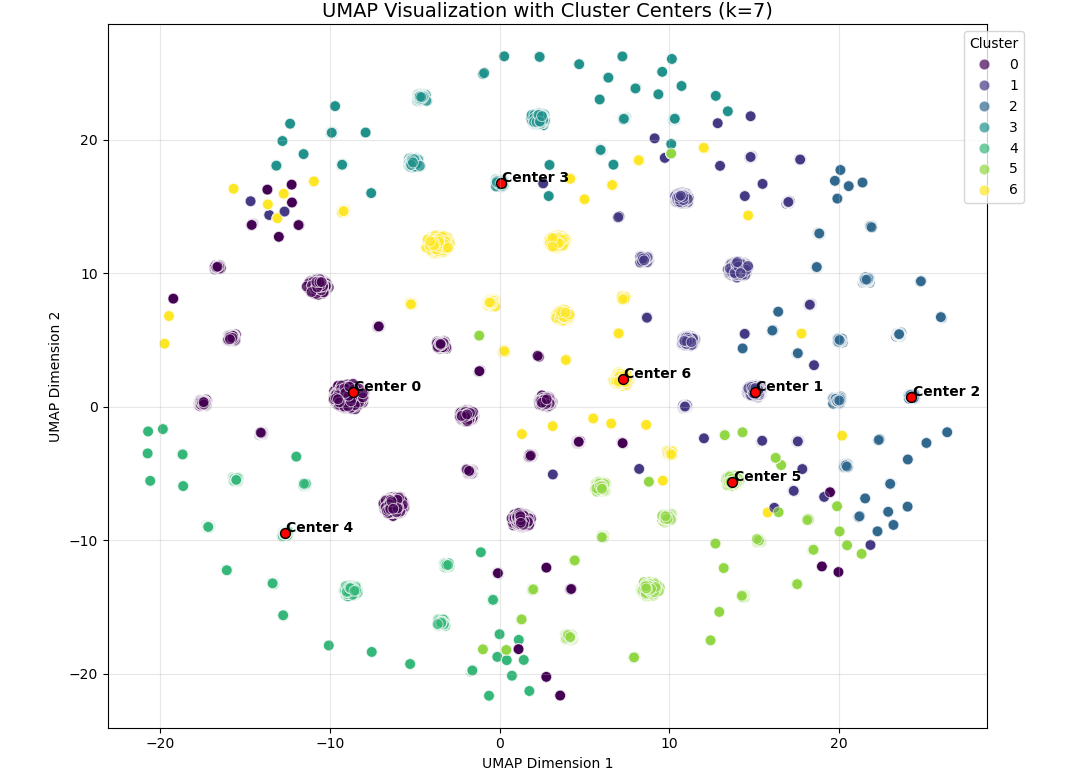
* 1. After researching other visualization cases, different encoding methods were applied to different attributes. Specifically, one-hot encoding was used for AGE\_GROUP because it is numerical; 'AGE\_GROUP', 'SEX', 'ROAD\_USER\_TYPE\_DESC', and 'INJ\_LEVEL' were processed using the LabelEncoder() tool, as they are all ordered categorical variables. After these adjustments, the resulting elbow plot showed significant improvement.



1. The dimensionality reduction method was changed from PCA to UMAP.
   1. After reducing the dimensionality to two dimensions using PCA, it was found that both PC1 and PC2 had cumulative variance contributions below 0.15. This indicated low correlation and an inability to accurately represent the characteristics of the multidimensional data, as shown in the figure:



* 1. Subsequently, dimensionality reduction and visualization were performed using the UMAP method, which showed a significant reduction in overlap between clusters.



**1.3** **Final result analysis**

1. Elbow plot: The elbow point is clearly identified at 7, indicating that using 7 different clusters is most appropriate.
2. Cluster scatter plot: The scatter plot after UMAP dimensionality reduction shows much less overlap between clusters, indicating better visualization performance compared to PCA-reduced plots.
3. From the correlation heatmap between features and UMAP components (UMAP1, UMAP2), the age variable has the greatest impact: ages 18-29 show a strong negative correlation with UMAP1 (-0.55), ages 40-49 (0.44) and over 70 (0.50) show positive correlation; ages 0-17 show a strong positive correlation with UMAP2 (0.53); ages 50-59 and 60-69 are negatively correlated on both dimensions. Gender (SEX\_encoded) and injury level (INJ\_LEVEL\_encoded) have very weak correlations (close to 0), while road user type (ROAD\_USER\_encoded) has a slight positive correlation with UMAP2 (0.18). This indicates that for this clustering, Driver Demographics & Risk Profiles age is the primary influencing factor, while other variables have lesser impact.

