Homework 19

1. What's the best value for k? Based on the elbow plot, silhouette scores, and VRC scores, k=4 is the best choice. While k=2 has a higher silhouette score, k=4 shows a significant elbow in the plot and a jump in VRC score.
2. What's the total explained variance of the three principal components? As you calculated, the three principal components explain approximately 89.5% of the total variance in the data.
3. What's the best value for k when using the scaled PCA DataFrame? Does it differ from the best value for k that you found by using the original scaled DataFrame? Similar to the original data, k=4 appears optimal for the PCA data, showing a clear elbow and good silhouette and VRC scores. The best k value (k=4) is the same for both the original and PCA-reduced data.
4. Based on visually analyzing the cluster analysis results, what's the impact of using fewer features to cluster the data by using K-means? When using PCA to reduce features, the clusters appear more distinct in your t-SNE visualization. The evaluation metrics (silhouette score, VRC) are generally higher. And, the elbow in the k vs. inertia plot is more pronounced. This suggests that dimensionality reduction helped remove noise and made the cluster structure clearer.