CPS 803 – Assignment 4 Report

Background

For this project, I worked with a video game sales dataset that includes the top 100 best-selling games worldwide. Each game has sales numbers from different regions, including North America, Europe, Japan, and other international markets. I chose this dataset because video games are a global industry, and sales often vary a lot depending on the region. This makes it interesting to see whether games naturally fall into groups based on their sales patterns.

The goal of the assignment is to apply clustering techniques to discover these hidden patterns. Before clustering, the data needed some basic preprocessing, such as selecting numeric sales features, handling missing values, and scaling the numbers so the algorithms can compare them properly. After preparing the data, I used two clustering methods - K-Means and Hierarchical Clustering, to see how the games group together and what insights these groups might reveal.

Methods

To explore patterns in the video game dataset, I followed a simple and structured approach. I started by selecting the numeric sales columns - North America, Europe, Japan, Other regions, and Global sales because clustering algorithms rely on numerical values to measure similarities between data points. Before applying any clustering, I checked for missing values and filled them using each column's meaning, which ensures the algorithms won't break due to gaps in the data.

Next, I standardized all numeric features using **StandardScaler**, which transforms the data, so each column has a consistent scale. This step is important because the raw sales numbers vary a lot between regions and could bias the clustering algorithm if not normalized. After the data was cleaned and scaled, I applied two clustering methods: **K-Means** and **Hierarchical Clustering**.

For **K-Means**, I used the elbow method to decide how many clusters to use. This involved running K-Means with different values of *k* and plotting the inertia (or within-cluster variance). The "elbow" in the graph appeared at **k = 3**, meaning that three clusters provided a good balance between simplicity and accuracy. I then ran K-Means again using *k = 3* and saved each game's cluster label.

The second method I used was **Agglomerative Hierarchical Clustering** with Ward linkage. This method builds clusters from the bottom up and groups the most similar items first. I generated a dendrogram to visualize how clusters merge at different distances and, for consistency with the K-Means results, I also chose to divide the data into three clusters.

Finally, to help visualize the results, I used **Principal Component Analysis (PCA)** to reduce the dataset from five features down to two main components. This allowed me to create scatter plots that clearly show how the different clusters separate in two-dimensional space. I also calculated the silhouette score for both clustering methods to evaluate how well the data points fit within their assigned clusters.

Results

After applying both clustering methods to the dataset, several patterns became clear. Using **K-Means with k = 3**, the algorithm separated the games into three distinct groups based on their sales performance. One cluster mainly contained globally successful games with very high sales across multiple regions. The second cluster represented mid-level sellers with moderate performance, and the third cluster included games with relatively low sales worldwide. This separation was also visible in the PCA plot, where the high-selling games stood out clearly as their own group.

The **Hierarchical Clustering** model, also set to three clusters, produced a similar structure. The dendrogram showed that many low-selling games were grouped together early, while the highest-selling games formed a separate branch. Although the overall grouping resembled the K-Means results, the hierarchical clusters were slightly less clean in the PCA visualization, with some overlap between the mid-range and lower-selling games.

To compare the two methods more formally, I calculated the **silhouette scores**. K-Means achieved a higher score, indicating that its clusters were more compact and well-separated. Hierarchical clustering had a slightly lower score, meaning its clusters were less distinct. In both cases, however, the results still reflected meaningful differences in game sales patterns.

Overall, the results showed that both clustering methods could identify similar trends in the data: a clear separation between top sellers and the rest, and additional subgroups within mid- and lower-performing games. These findings suggest that sales behavior naturally forms clusters, and that K-Means provided the sharper separation in this dataset.

Conclusion

In this project, clustering helped reveal meaningful patterns in video game sales across different regions. Both K-Means and Hierarchical Clustering showed that the dataset naturally separates into three main groups: top-selling games, mid-level performers, and lower-selling titles. While both methods produced similar patterns, K-Means gave slightly cleaner and more separated clusters, supported by a higher silhouette score and clearer PCA visuals. Overall, the analysis showed that clustering is a useful way to understand sales trends and compare game performance on a global scale.

References

Video Game Sales Dataset. Retrieved from Kaggle/UCI [https://www.kaggle.com/datasets/volodymyrpivoshenko/video-game-sales-dataset]

Pedregosa, F., et al. “Scikit-learn: Machine Learning in Python.” Journal of Machine Learning Research, 2011

Ward, J. H. “Hierarchical Grouping to Optimize an Objective Function.” Journal of the American Statistical Association, 1963