Dachuan He

10x10 Correlation Matrix heatmap
- Pros: directly shows correlation between pairs of attributes, positive/negative correlation, the strength.
- Cons: cannot tell us which attributes have the highest correlation sum. Also, cannot show details of a pair of attributes.

5 x 5 Scatter plot matrix
- Pros: directly shows the relationship between pairs of attributes. For instance, we can easily find there is a linear relationship between Global_Sales and Other_Sales.
- Cons: we cannot see a continuous influence by attributes.

Parallel coordinates display
- Pros: we can see some patterns beyond the positive/negative correlation. For instance, we can see two patterns between User_Score and Critic_Score. First, increases with the increasing Critic_Score. Second, many video games' Critic_Score are higher than User_Score.
- Cons: when the observations are too much, the diagram could lose usage.

PCA with scree plot
-Pros:
1. Removes Correlated Features.
2. Improves Visualization: PCA transforms 10-dimensional data into 2 dimensions so that the Video Games dataset can be visualized easily.
2D Screen Plot tells us which Principal Components result in high variance and have more impact as compared to other Principal Components.

-Cons:
1. Principal Components are not as readable and interpretable as original features.
2. Data standardization must be performed before PCA, otherwise, PCA will not be able to find the optimal Principal Components.

Also, for standardization, all the categorical features are required to be converted into numerical features before PCA can be applied.
(categorical features are not used in the project implementation)

PCA is affected by scale, so we need to scale the features in data before applying PCA. (I've used StandardScaler from Scikit Learn to standardize the dataset features)

3. Information Loss: Although Principal Components try to cover maximum variance among the features in a dataset, if we don’t select the number of Principal Components with care, it may miss some information as compared to the original list of features.

Biplot with 10 projected axes
- Pros: Showing how the original features are combined with Principal components, like NA_Sales, Global_Sales, Other_Sales, PAL Sales are grouped together.
- Cons: None.

MDS of data with euclidean distance
- Pros: Multidimensional scaling (MDS) allows us to visualize how near points are to each other for many kinds of distance, such as euclidean distance in this project.
It produces a representation of Video Games data (10 dimensions) in 2 dimensions.
MDS does not require raw data, but only a matrix of pairwise distances or dissimilarities.

- Cons: I tried to get some clustering information based on the diagram by the ESRB_Rating attribute. It shows some M (Mature 17+) games are grouped on the right-bottom corner of the plot. But the other ESRB_Rating such as E, E10, and T are not clustered very well. It may be caused by the original dataset's attribute lack of clustering relationship.

MDS of attributes with 1-|correlation| distance
- Pros: Higher correlation leads to a smaller distance. The diagram shows how attributes are close to each other. It tells NA_Sales, PAL_Sales, Global_Sales, Rank, Other_Sales are neighbors. The rest of the attributes lay alone.
- Cons: None.

From the MDS of attributes, I find Global_Sales, NA_Sales, PAL_Sales, Other_Sales, Rank attributes are really close to each other. However, JP_Sales are not. It means Japan market is very small compared to the game platforms included in the dataset. Meanwhile, Critic_Score and User_Score are not important to Global_Sales. As I thought, customers buy games that they're interested in, although the games' scores are low. The other interesting thing is.From PCA biplot, I find JP_Sales, Critic_Score and User_Score are really close. It means Japanese customers really care about video games' scores. I guess they'll read every game purchase comment carefully before buying it. It's Japan's market target customers' behavior, which game selling platforms may take into consideration.Also, from the parallel coordinates diagram, I find many games' Critic_Score are higher than User_Score. It means sometimes customers don't like a fancy game that gets high scores from critics. These kinds of games may cost a huge investment but get low returns. This case is common.

These displays show attributes relationships well, like how the other attributes can influence a video game's Global_Sales.

The original dataset has some categorical columns such as Genre and ESRB_Rating. The genre has 14 unique values. ESRB_Rating has 5 unique values.
I didn't use them in this project. But if I'll do it again, I would try to convert them to a numeric type, seeing how they influence Global_Sales. These categorical attributes may provide some other interesting findings.