**Lab 2 – Feature Selection**

**Revisiting Data Preprocessing:**

After conducting an analysis and brainstorming session, I decided to revisit my previous data preprocessing steps. I came to realize that replacing missing values in certain columns would impact my model's future results. Additionally, dropping these missing values and data points didn't significantly reduce the dataset size, indicating that I wasn't losing crucial information. Moreover, I noticed that the columns with missing values were consistently missing in the same samples, suggesting that these samples didn't contribute much information. Consequently, I made the decision to drop the missing values. Furthermore, there was a specific column called "Loaned From" that had missing values in all data points, so I opted to remove that column as well.

**Converting Data Types:**

After addressing the missing values, I encountered several columns that were supposed to be in float or int format but were stored as objects due to the presence of units alongside the numbers, such as "$123K." To rectify this, I replaced "K" with 1000, "M" with 1000000, and removed the '$' sign. Subsequently, I converted these values into either int or float data types. This process was applied to approximately seven columns with similar data formatting issues. Additionally, there was one column called "Joined" that needed to be in datetime format, so I adjusted it accordingly.

**Visualizing:**

After handling missing values and addressing data types, I proceeded to create visualizations for each column in the dataset to examine their distribution, basic curves, and identify any outliers. These visualizations can be found in the Figure folder. Most columns displayed a normal or skewed normal curve, with a few outliers present in the data. Notably, some columns, such as the "Value" column, had a substantial number of outliers. However, I chose to retain these outliers as the column exhibited a bimodal distribution. These visualizations significantly improved my understanding of the dataset.

**Label Encoding:**

To handle the categorical data in the dataset, I performed label encoding, transforming them into numerical values. This step was necessary because machine learning models exclusively accept numerical inputs.

**Feature Selection:**

Once I had transformed the data into a format suitable for modeling, I initiated the feature selection process. My first step was to calculate the correlation between each column and the target variable, which in this case was "Release Clause." The visualizations for this can be found in the figures folder. These graphs revealed that eight columns exhibited a strong correlation with the target variable, including "Value," "Wage," "Potential," "Overall," and others.

Following the correlation analysis, I employed various feature selection algorithms from the Sklearn library, including "Variance Threshold," "SelectKBest," "Mutual Info," "Percentile," "RandomForest," "Chi-Squared," "PCA," and "Genetic Algorithm." I also visualized some of the results from these algorithms. After applying these techniques, I determined that "Overall," "Potential," "Value," "Wage," "Reactions," and "Composure" were the most significant features for this dataset.