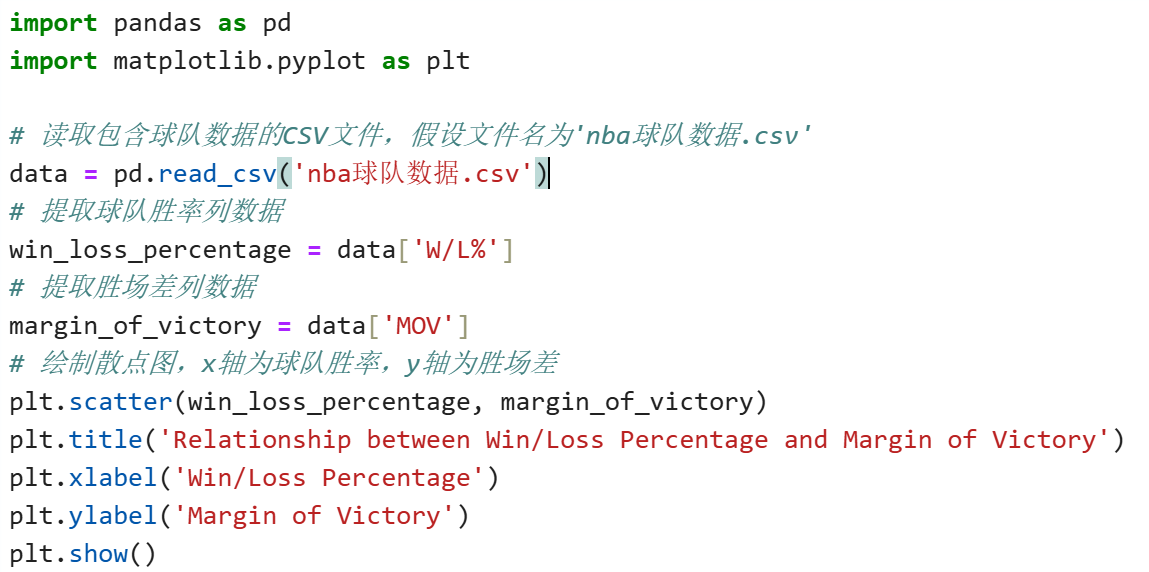
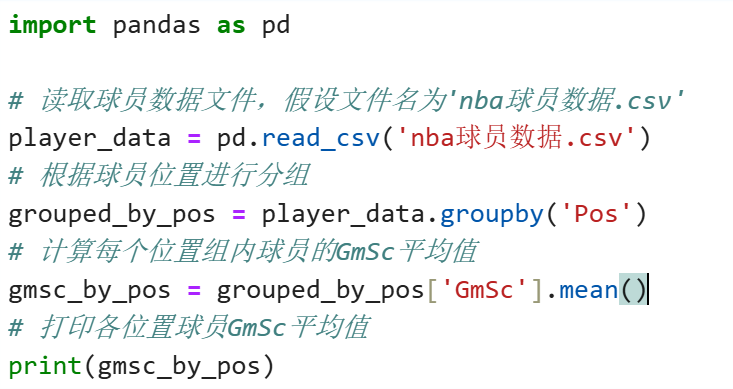
NBA Player Performance and Trend Analysis Report

1. Data Acquisition and Preprocessing
2. Data Sources:
3. Acquisition of average data for the 2022 - 2023 season: The average data for this season was obtained from Basketball Reference (https://www.basketball-reference.com/leagues/NBA\_2023\_per\_game.html). During the data acquisition process, since the data on the Kaggle platform can be directly downloaded as a CSV file, no specific code was used for web data scraping operations.
4. Acquisition of historical data: Historical player statistical data was obtained through Kaggle (https://www.kaggle.com/datasets/drgilermo/nba-players-stats), and the CSV file was also directly downloaded without involving complex data acquisition code.
5. Data Preprocessing:
6. Missing value handling: For the blank columns in the historical player data, they were uniformly filled with "Unknown" to clarify the data missing situation and facilitate subsequent analysis and processing. In the 2022 - 2023 season data, some blank columns were filled with spaces. In terms of data inspection, it was ensured that the letter columns only contained letters and the number columns only contained numbers, thus ensuring the consistency and accuracy of the data format and laying a foundation for subsequent analysis.
7. Data consistency correction: Some possible inconsistent expressions in the data were preliminarily processed. Although no complex data field conversions were involved, through the above standardization of the data format, the overall consistency of the data was improved to a certain extent.
8. Exploratory Data Analysis
9. Relationship between Player Data and Team Success
10. Analysis method: A correlation analysis method was adopted, and scatter plots (implemented with the help of the `matplotlib.pyplot` library) were drawn to visually present the relationships between different variables, thereby exploring the connections between player data and variables related to team success. This visual means is helpful for preliminarily judging the correlation trends between variables.
11. Example code and explanation 

Through this scatter plot, the distribution relationship between the team's winning percentage and the winning margin can be visually observed, and it can be preliminarily judged whether there is a certain linear or nonlinear correlation trend between them, providing a visual basis for further in-depth analysis.

1. Identifying Progressive Players
2. Definition of "significant progress": By measuring the average GmSc values of the first 20 games and the last 20 games in an 82-game season (this index comprehensively considers the player's overall performance), the difference in the player's state in the front and back sections of a season is evaluated, thereby determining whether the player has progressed or regressed during the season.
3. Implementation logic: 

In the code, first, the data needs to be filtered according to the player ID and the number of games. The average GmSc values of the first 20 games and the last 20 games of each player are calculated respectively, and then the difference between the two is calculated. If the difference is positive and exceeds a certain threshold (determined according to the actual data distribution and analysis requirements), it is considered that the player has made significant progress during the season; if the difference is negative and is lower than a certain threshold, it is considered that the player has a regression trend. In this way, the development trend of the player within the season can be more comprehensively evaluated from the perspective of overall performance.

1. Impact of Player Position on Performance Indicators
2. Grouping and analysis method: The `groupby` function in the `pandas` library was used to group the player data according to the player position (Pos column). For each position group, the GmThe GmSc value of the players in the group was calculated, so as to obtain the differences in player performance among different positions. Further, by combining the team winning percentage data, the distribution of players in the top 3 GmSc values for each team was analyzed, thereby exploring the approximate relationship between the winning percentage and the player position.
3. Code example and explanation (simplified schematic) The above code first reads the player data and then uses the `groupby` function to group according to the player position. Then, the average GmSc value of the players in each group is calculated, so that the differences in the comprehensive performance indicator (GmSc) among players in different positions can be visually compared. Through further correlation analysis with the team winning percentage data, the potential impact of the player position on the team's performance can be further explored.
4. Modeling
5. Predicting Player Future Performance
6. Algorithm selection: A random forest algorithm was adopted to construct a model for predicting player future performance. The random forest is an ensemble learning algorithm based on decision trees, which can handle complex nonlinear relationships and has good robustness and generalization ability when dealing with a large number of features and data.
7. Model training and evaluation implementation Training process: In the code, first, the data set was divided into a training set and a test set (for example, in a ratio of 70:30 or 80:20, specifically determined according to the data characteristics and analysis requirements). Then, the random forest model was trained using the training set. During the training process, the model would learn the patterns and relationships in the data according to the input features (such as the player's historical average data per game, the number of games, age, etc. related factors), and construct multiple decision tree models.

Evaluation process: After the model was trained, the test set was used to evaluate the performance of the model. In this analysis, variance was used as an evaluation indicator. Variance can reflect the dispersion degree of the model's prediction results. The smaller the variance, the higher the stability of the model and the more reliable the prediction results. By calculating the variance between the prediction results and the true results on the test set, the performance of the model on new data was quantified and evaluated to determine whether the model could effectively predict the player's future performance.

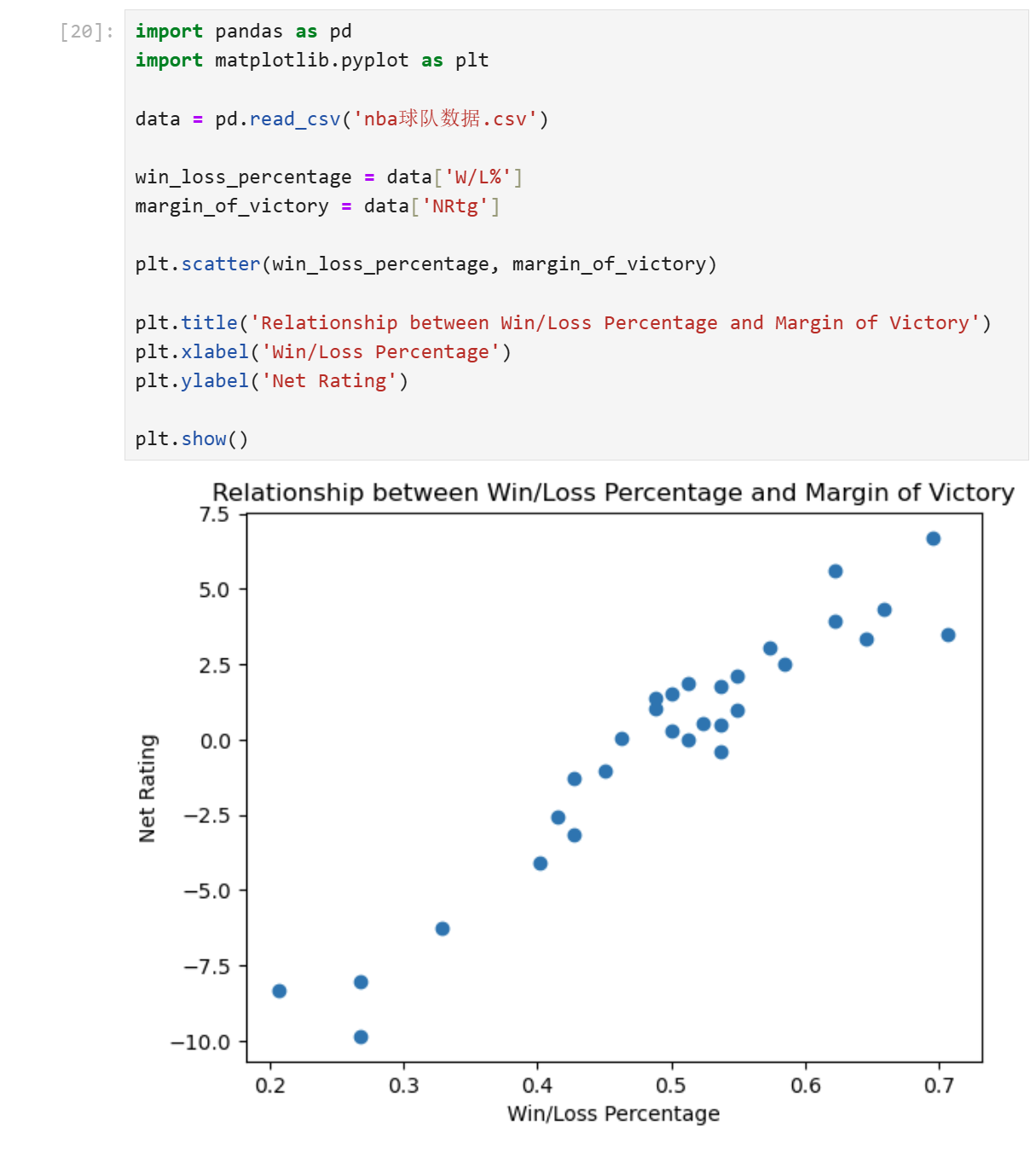
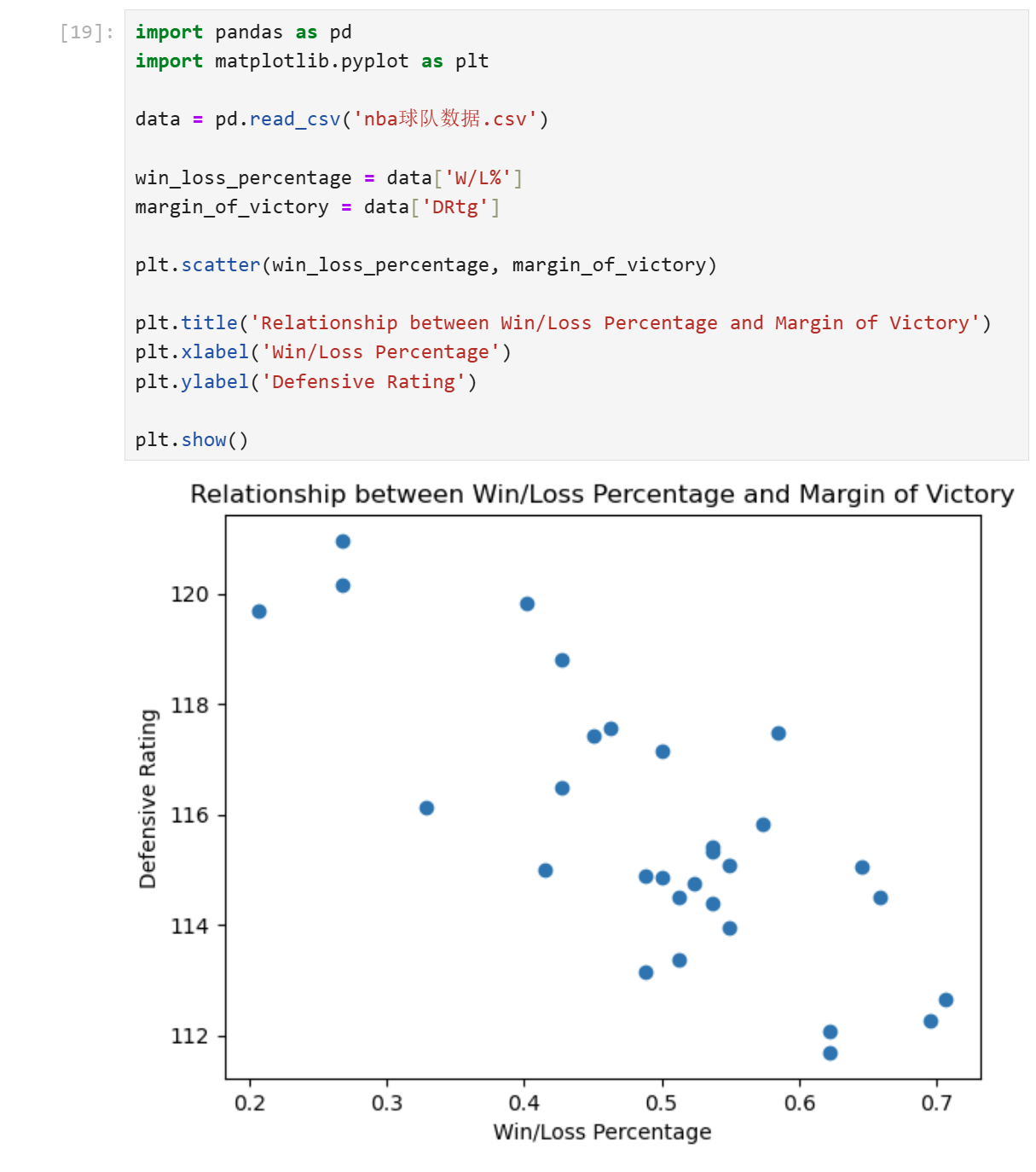
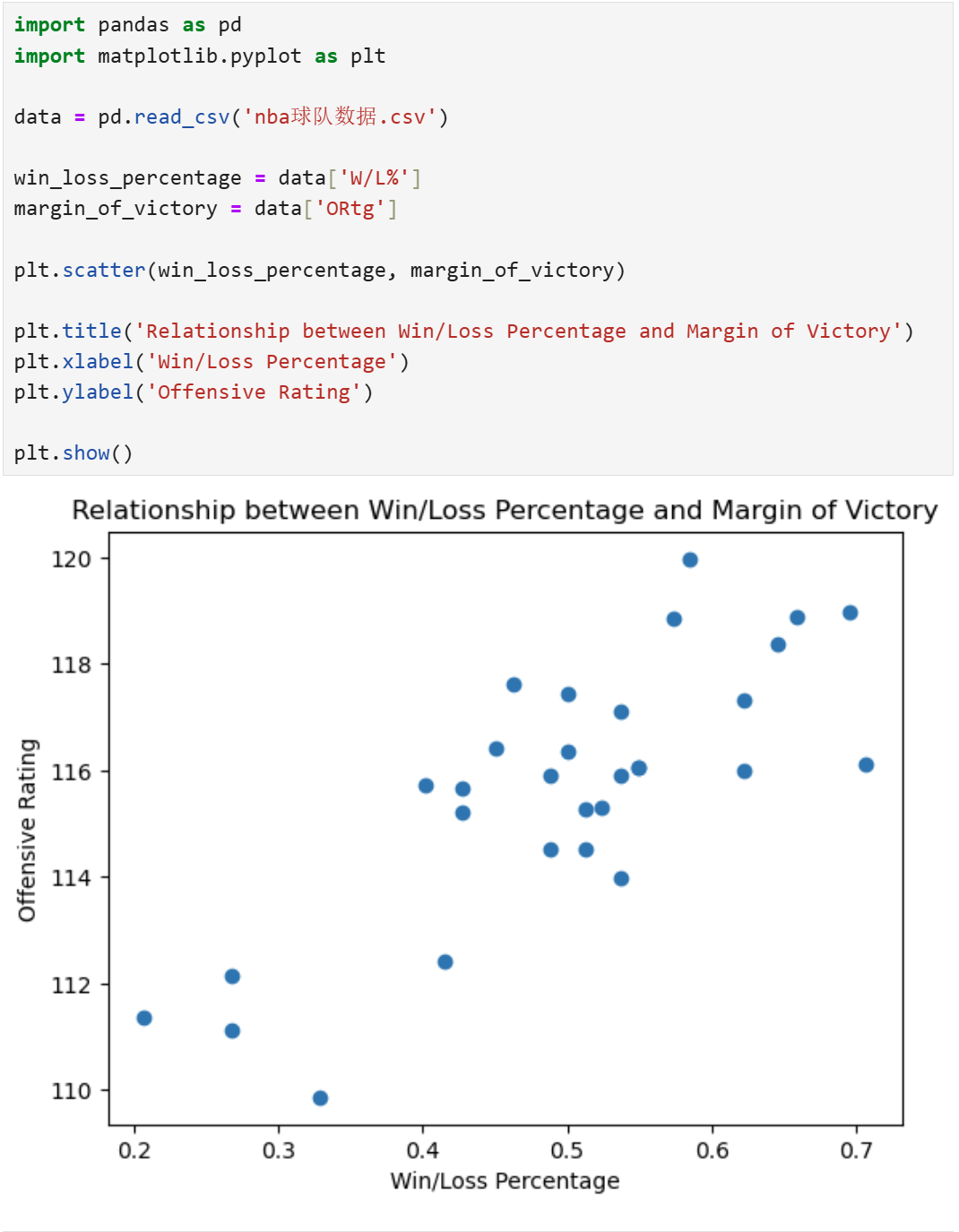
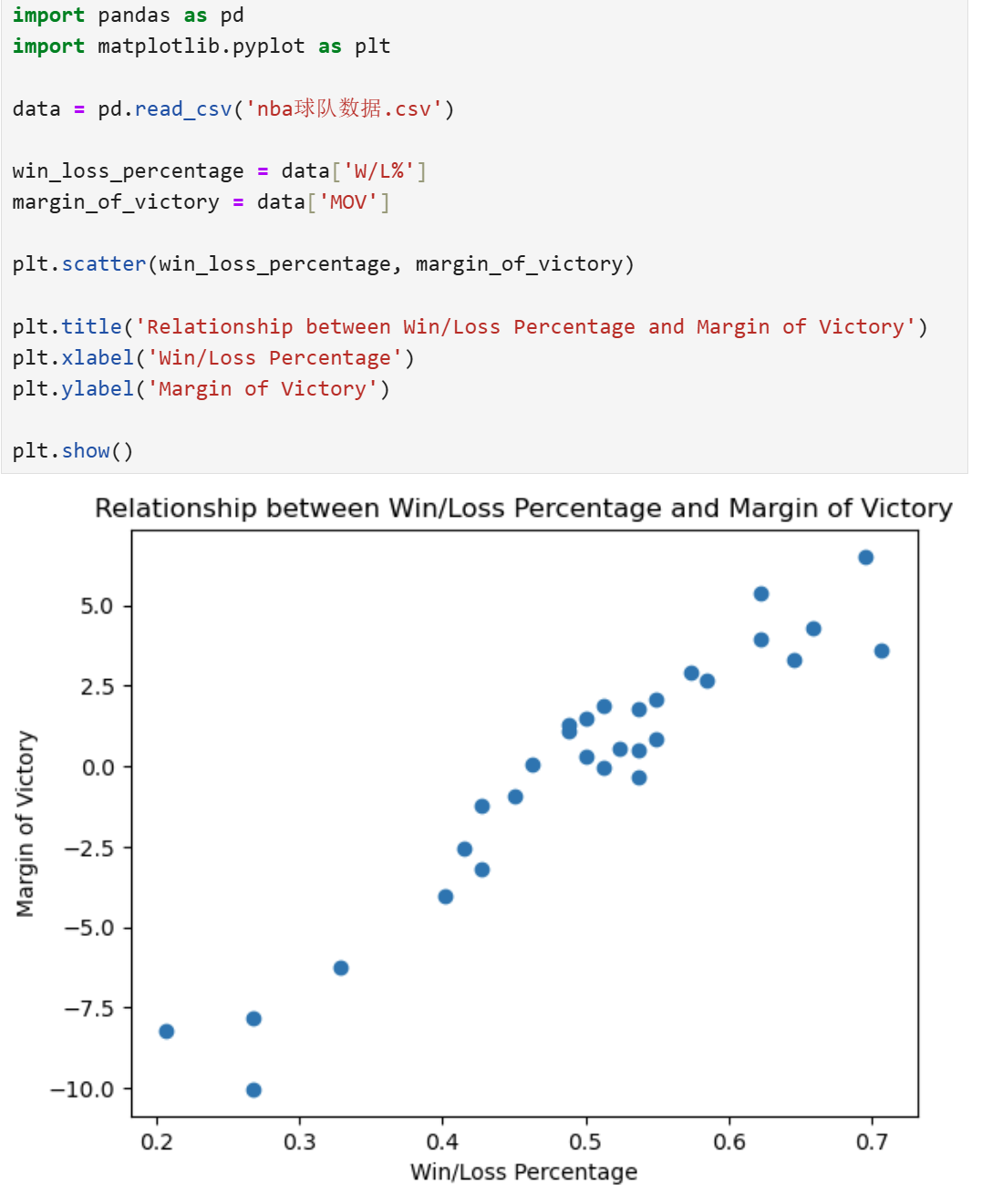
1. Cluster Analysis
2. Algorithm selection: A hierarchical clustering algorithm was selected to group similar players. The hierarchical clustering algorithm does not need to pre-determine the number of clusters. It constructs a clustering hierarchy to group the data and can intuitively show the hierarchical relationships between different categories, which is suitable for exploratory analysis scenarios where the data distribution situation is not very clear.
3. Clustering process and result evaluation

Clustering process: In the code, the relevant data of the players (such as average score per game, rebounds, assists, etc. feature data) was input into the hierarchical clustering algorithm. The algorithm would gradually merge or split the data points according to the similarity between the data points (usually using a distance metric, such as Euclidean distance), forming a clustering hierarchy. During this process, the data points would gradually form different levels of clustering clusters according to the similarity. Data points with higher similarity would first form small clusters, and then the small clusters would gradually merge into larger clusters until all data points were merged into one cluster or a preset stopping condition (such as a distance threshold) was reached.

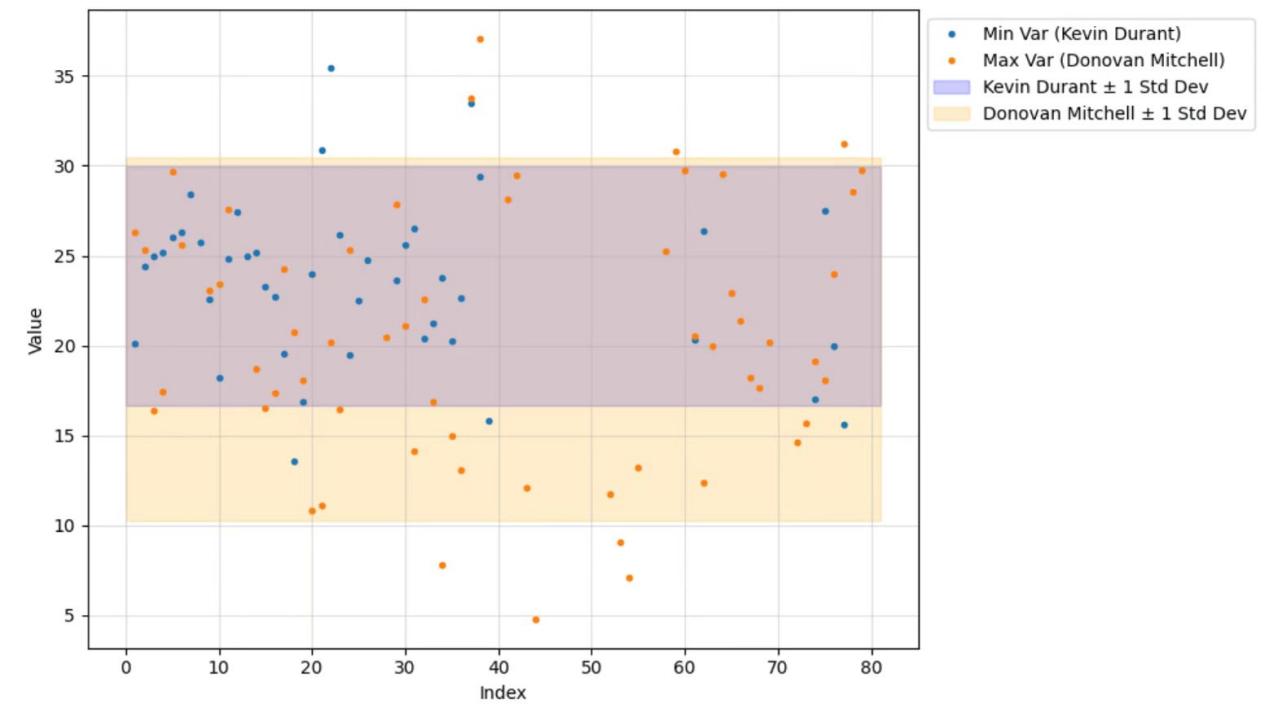
Result evaluation: Although the hierarchical clustering algorithm does not need to pre-determine the number of clusters, it is still necessary to evaluate the rationality of the clustering results. Common evaluation methods include observing the clustering dendrogram (Dendrogram), and judging the clustering effect under different numbers of clusters according to the length and structure of the branches in the dendrogram. The parts with shorter branch lengths and relatively stable clustering structures usually indicate more reasonable clustering divisions. In addition, it is possible to combine professional domain knowledge to qualitatively analyze the clustering results, and judge whether the clustering results conform to the understanding and expectation of the player type. For example, whether the players in different clustering clusters have obvious similarities and differences in their performance characteristics in the game. If the clustering results do not conform to the expectation or need further optimization, it is possible to try to adjust the distance metric method or other parameter settings, and re-perform the clustering analysis to obtain more reasonable clustering results.

1. Summary and Outlook
2. Summary
3. This project has conducted in-depth research on the performance and trends of NBA players through systematic data acquisition, preprocessing, exploratory data analysis, and modeling operations. During the data acquisition stage, appropriate acquisition methods were selected according to the characteristics of different data sources to ensure the integrity and usability of the data.
4. In the data preprocessing stage, effective treatment was carried out for missing values and data consistency problems, improving the data quality and providing a solid foundation for subsequent analysis. Exploratory data analysis has revealed the relationships between player data and team success, player progress, and player position from multiple angles. Through visual and statistical analysis methods, data support and new insights have been provided for understanding the complex phenomena in NBA games.
5. In the modeling part, the application of the random forest prediction model and the hierarchical clustering model has provided new perspectives and methods for player performance analysis. Although the models have certain limitations, they still have certain reference values in practical applications and can provide meaningful decision-making bases and development direction guidance for team management, coaching teams, and players themselves.
6. Outlook
7. Data cleaning optimization: With the continuous increase in the scale of data sets, traditional data processing methods may face efficiency challenges. In the future, more advanced distributed data processing frameworks (such as Apache Spark) can be considered to introduce to improve the efficiency of data cleaning and preprocessing, ensuring that large-scale data can be processed quickly. For data imbalance problems, in addition to the existing simple treatment methods, more effective over-sampling or under-sampling techniques (such as SMOTE and its variant algorithms) can be further studied and applied to balance the distribution of different classes of data and improve the accuracy and generalization ability of the model.
8. Visualization enhancement: In data visualization, although some basic plotting tools (such as `matplotlib.pyplot`) have been used to achieve the visual display of some data relationships, in order to provide a more rich and interactive data exploration experience, more advanced interactive visualization libraries (such as Plotly or Bokeh) can be further tried. These tools can create dynamic, interactive charts that allow users to explore data details through operations such as zooming, hovering, and filtering, and more intuitively discover the hidden patterns and trends in the data.
9. Analysis technology expansion: In order to more comprehensively and deeply mine the information in the data, more advanced analysis techniques can be explored in the future. For example, time series analysis can be used to study the dynamic change trends of player performance over time, capturing the key turning points and cycle characteristics in the growth trajectory of players. Principal component analysis (PCA) and other dimension reduction techniques can retain the key information of the data while reducing the data dimension, simplifying the model construction process and improving the model's interpretability, helping to extract the main influencing factors from the complex data relationships. In addition, deep learning algorithms (such as neural networks) can be combined to perform more complex pattern recognition and prediction tasks, such as predicting the performance of players in specific game scenarios or injury risks.
10. Indicator system improvement: When evaluating the impact of player position on the game result, currently mainly relying on traditional average data per game indicators and simple comprehensive indicators (such as GmSc). In the future, the indicator system can be further expanded, introducing more advanced statistical indicators (such as efficiency value PER, true plus-minus value RPM, victory contribution value WS, etc.), measuring the contribution of players in the game from multiple dimensions more comprehensively and accurately. At the same time, combining game video analysis and player tracking data (such as motion trajectories, speed, acceleration, etc.), a more in-depth understanding of the actual performance of players on the field and their impact on the game can be achieved, providing more scientific bases for team tactical arrangements and player cultivation. Through this project practice, not only has the performance and trends of NBA players been deeply analyzed, but also rich experience has been accumulated in data processing, analysis, and modeling, providing an important reference for future similar data analysis projects. In the continuously developing field of sports data analysis, continuous exploration and innovation will help to further mine the potential value of the data and provide more powerful support for sports decisions.

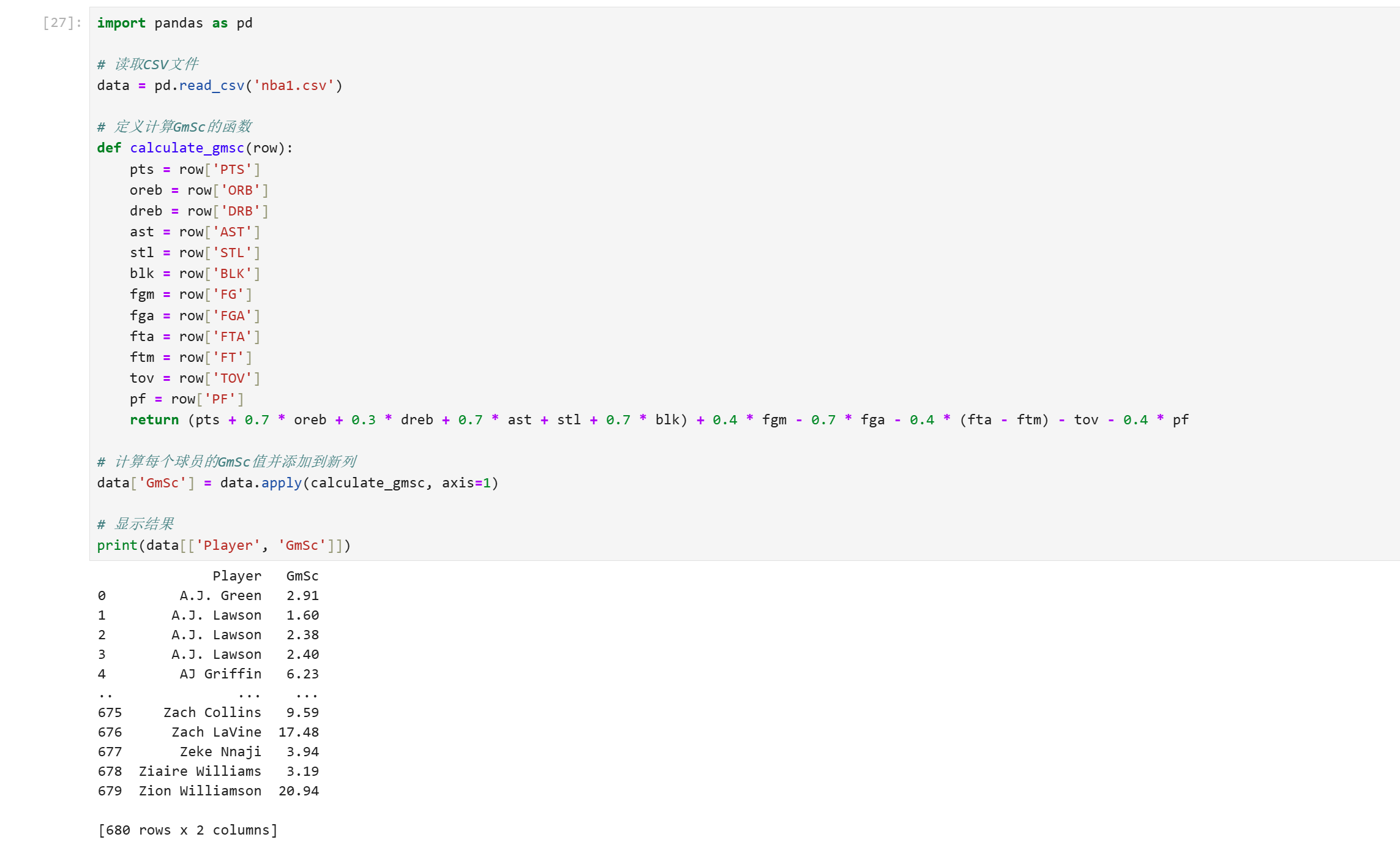
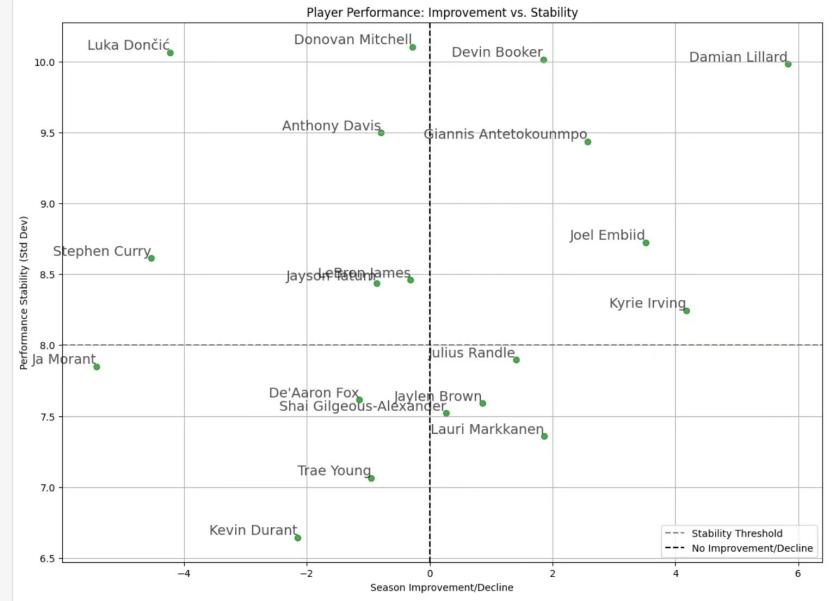
And The following will show some code and processed images.



The following image can be drawn through this table.



The following image can be drawn through this table.



There is too much code here, so I will only show the results below.

