Clemson Athletics Technical Project Report

Pandu Ranga Avinash Srikhakollu
Clemson University
psrikha@clemson.edu

ABSTRACT

This project aims to analyze Clemson Athletics data, with a specific focus on the 2023-24 inaugural season of Clemson Gymnastics. It consists of two main parts: Part 1 focuses on examining customer attendance and the factors influencing it, while Part 2 focuses on analyzing Clemson Gymnastics data, encompassing both individual and team performance over the course of the season.

1. DATASETS

Multiple datasets were utilized for each part of the project. In Part 1, a gymnastics dataset was used, consisting of records of 280,995 customers. This dataset included fields such as user ID, account ID, athletics event ID, athletics event name, athletics event date, scan minutes before start, scan time, scan method, sport name, season, row name, section name, seat number, ticket acquisition type, price, and distance traveled in miles.

For Part 2, 11 datasets representing different gymnastics meets were utilized. These datasets contained fields such as round, order, name, team, event, score, difficulty, execution, neutral deduction, stick bonus, inquiry, edited, date, event type, sv, judge 1, judge 2, source, sv1, sv2, sv3, sv4, judge 3, judge 4, place, and player.

2. TOOLS AND LIBRARIES USED

The project is implemented using the Jyputer Notebook software for both data analysis and visualization. Additionally, a few visualizations are previsualized using Tableau. The libraries used for the project are - pandas, numpy, matplotlib and seaborn.

3. PART -1 ATTENDANCE

1. The first analysis was to identify the number of customers scanned into a meet during the inaugural season. To achieve this, I first verified the total number of customers attending the event since each account ID had multiple UIDs.

   The gymnastics dataset consisted of all the sports (Baseball, Basketball, etc) as well. Given our focus on Clemson Gymnastics, the data was filtered to include only records related to gymnastics. Then I filtered the records with no scan data and displayed the count of total customers scanned and the number of customers scanned into each meet.

   Result:

```
In [10]: df_gymnastics = df[df['SPORT_NAME'] == 'Women\'s Gymnastics']
         df_gymnastics_scanned = df_gymnastics[df_gymnastics['SCAN_METHOD'].notna()]
         customers_scanned = len(df_gymnastics_scanned)
         customers_scanned

Out[10]: 27606
```

```
Out[50]: Clemson v Air Force         5986
         Clemson v North Carolina    5906
         Clemson v NC State          5673
         Clemson v Pittsburgh        5529
         Clemson v William & Mary    4512
         Name: ATHLETICS_EVENT_NAME, dtype: int64
```

2. The second analysis was to determine the percentage of seats scanned and the percentage of seats scanned in after the start of each event for every meet. To achieve this, I have first calculated the percentage of seats scanned for each meet out of the total number of seats scanned. Following this, a data frame was created by filtering the seats that had been scanned after the start of the event, and the percentage was calculated accordingly.

   Result:

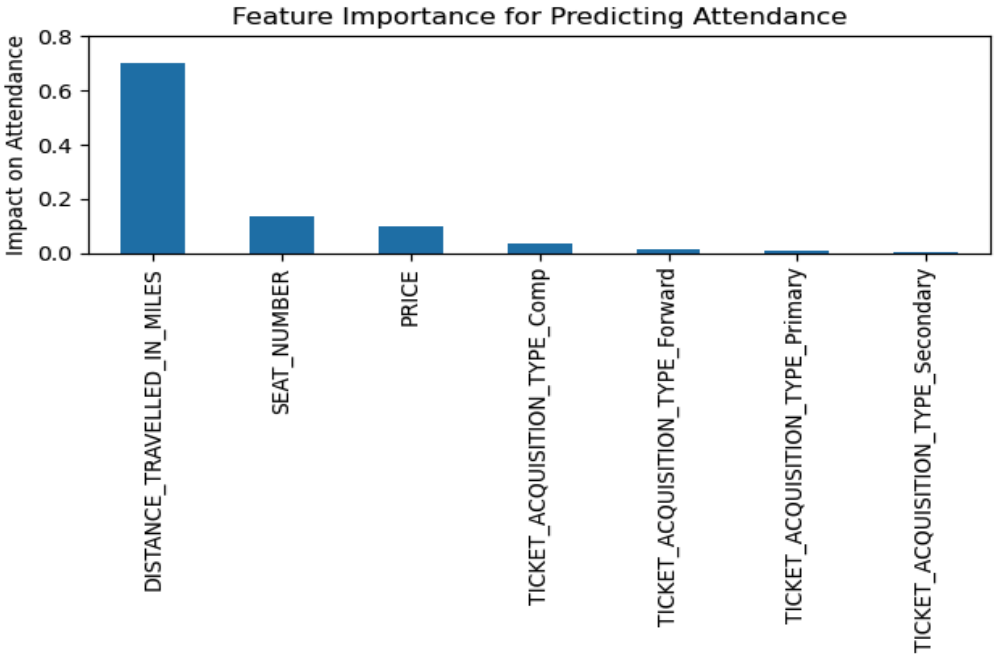
```
Percentage of seats scanned for each meet:

Out[16]: Clemson v Air Force         21.683692
         Clemson v North Carolina    21.393900
         Clemson v NC State          20.549880
         Clemson v Pittsburgh        20.028255
         Clemson v William & Mary    16.344273
         Name: ATHLETICS_EVENT_NAME, dtype: float64
```

```
Percentage of seats scanned after the start of event for each meet:

Out[21]: Clemson v Air Force         12.763114
         Clemson v NC State          21.928433
         Clemson v North Carolina    12.834406
         Clemson v Pittsburgh         8.808103
         Clemson v William & Mary    22.805851
         Name: ATHLETICS_EVENT_NAME, dtype: float64
```

3. The third analysis was to identify the factors which had the biggest impact on customer attendance. To accomplish this, I have implemented the feature importance of a random forest classifier which provides insights into the importance of each variable on the target variable. However, for this calculation, all the values are needed to be in a numerical datatype.

   In the dataset, the categories that could be analyzed for attendance impact are distance_travelled_in_miles, price, seat_number and Ticket_acquisition_type. Out of these, Ticket_Aquisition_Type was a categorical variable which need to be converted into numerical value. This is done using the “One Hot Encoding” process that converts the categorical variables to numerical values. Further, the data is split into train and test data and feature importance is calculated with the following result:

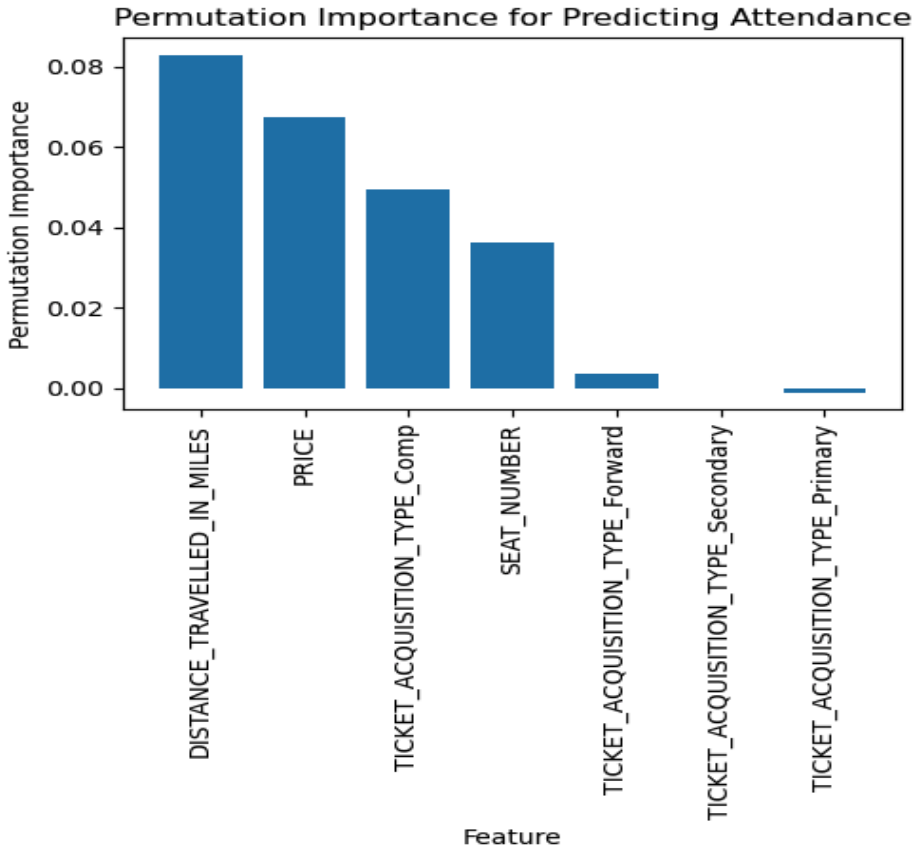
```
Feature Importance of the factors affecting customer attendance:

DISTANCE_TRAVELLED_IN_MILES          0.700385
SEAT_NUMBER                          0.136072
PRICE                                0.101781
TICKET_ACQUISITION_TYPE_Comp         0.034149
TICKET_ACQUISITION_TYPE_Forward      0.015419
TICKET_ACQUISITION_TYPE_Primary      0.010531
TICKET_ACQUISITION_TYPE_Secondary    0.001663
dtype: float64
```



From the above output, we can understand that the metric "**Distance_travelled_in_miles**" had the biggest impact on customer attendance followed by "**Seat_Number**" and "**Price**"

To verify the result, I have calculated permutation importance which involves randomly shuffling values of a single feature and then measuring the reduction in model performance criterion. The greater the reduction is, the more important the feature is. The difference between feature importance and permutation importance is that permutation importance is performed on test data while feature importance is performed on train data. The result of permutation importance is presented below:

```
Permutation Importance of the factors affecting customer attendance:

DISTANCE_TRAVELLED_IN_MILES: 0.08298084291187736
PRICE: 0.06755938697318002
TICKET_ACQUISITION_TYPE_Comp: 0.04930651340996165
SEAT_NUMBER: 0.03624904214559382
TICKET_ACQUISITION_TYPE_Forward: 0.003333333333333277
TICKET_ACQUISITION_TYPE_Secondary: -0.00011494252873567682
TICKET_ACQUISITION_TYPE_Primary: -0.0011839080459770594
```



Final Result: In both calculations, it is evident that "**Distance_Travelled_In_Miles**" had the biggest impact on customer attendance along with "**Price**" and "**Seat_Number**".

4. PART -2 SPORTS COMMUNICATIONS

1. For the sports communications, there were a total of 11 data files representing 11 different meets of the Clemson gymnastics team over the season. The first task was to combine the data from the meets into a single dataset using programmatic methods. The following steps are undertaken to implement the task:

    a. **Combining**: Initially, all the data files were merged into a single file by gathering the data files in a directory and iterating through the directory.
    b. **Data Cleaning**: To optimize the organization of the data, I performed data preprocessing on specific portions of the data files. This involved addressing numerous empty values in critical fields and resolving instances where multiple column names were assigned to the same columns. Additionally, certain data files lacked event dates, which I retrieved from the official Clemson gymnastics website and incorporated into the respective data files.

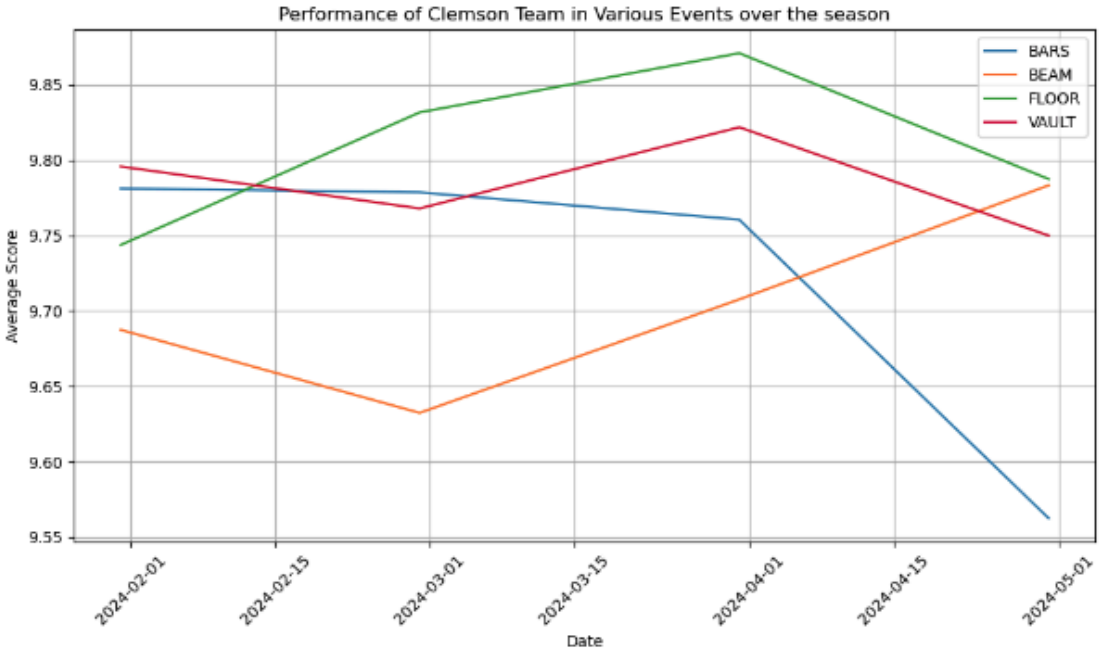
c. **Merging**: Following the completion of cleaning and preprocessing steps, I have merged the data into a singular file, and a CSV was generated for this consolidated dataset.

2. The second task was to find out the specific event in which UCLA scored the highest and their score. For this, I created a data frame by filtering the data with the UCLA team and returned their highest score using the inbuilt Python function.

   Result:

```
UCLA scored the highest on the following events:

     event  score
394  FLOOR   10.0
418   BARS   10.0
430  VAULT   10.0
```

3. The third task was to identify 3 unique insights from the combined data about the Clemson team's or individual gymnasts' performance. Following are a few insights I have identified -
   a. **Performance of the Clemson team in various events over the season:** The gymnastics season lasted for around 4 months and I have analyzed Clemson's team performance in various events (BARS, BEAM, FLOOR, VAULT) over the time period using a line chart as shown below -



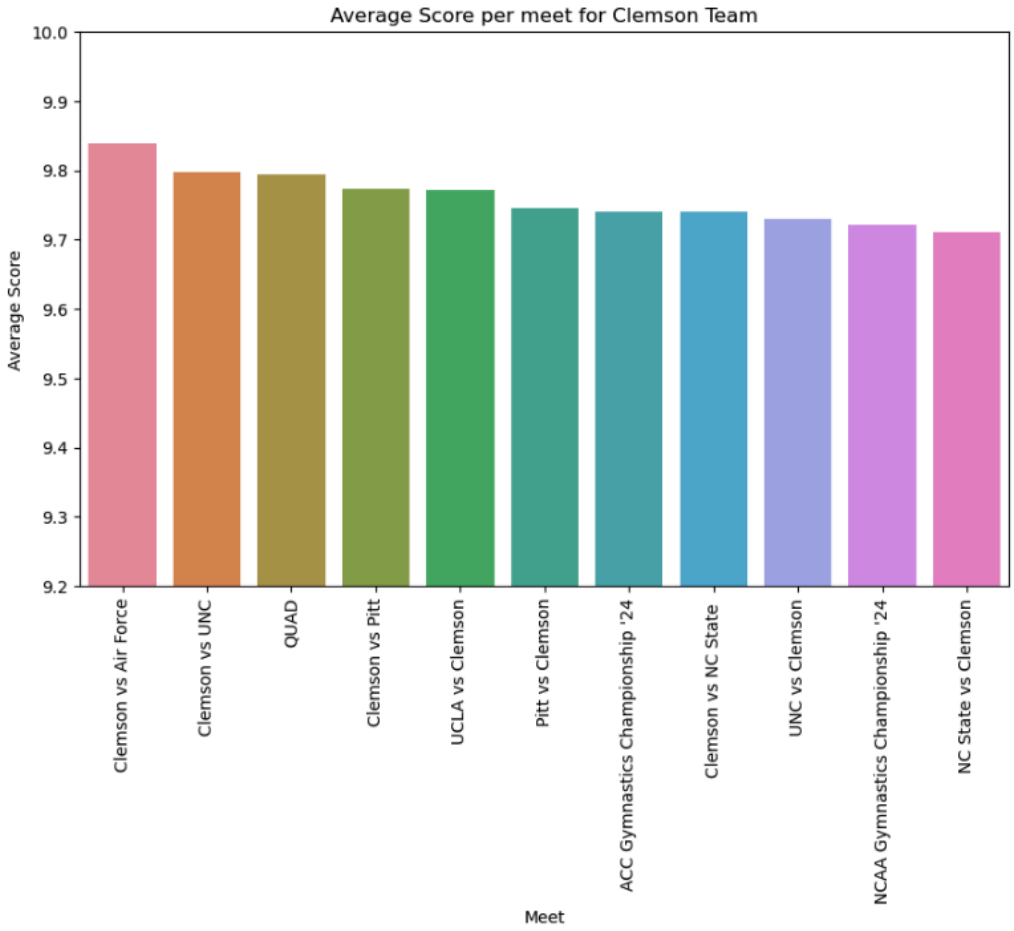
From the above graph, we can understand that the gymnastics team started the season well in all the events over the first two months except for the "BEAM" event. Over a period of time performance of the team in "BARS", "FLOOR" and "VAULT" remained good until the beginning of April from which the performance in these three events gradually decreased whereas the performance in "BEAM" significantly increased from the beginning of March.

b. **Event with the top score in each fixture for the Clemson team:** The gymnastics team participated in 11 meets over the season and I have tried to analyze the best gymnast against each opposition with the specific event and score. The result is as follows -

fixture	player/gymnast	event	score
ACC Gymnastics Championship '24	Molly Arnold	FLOOR	9.950
Clemson vs Air Force	Rebecca Wells	BEAM	9.950
Clemson vs NC State	Molly Arnold	FLOOR	9.950
Clemson vs Pitt	Brie Clark	FLOOR	9.925
Clemson vs UNC	Rebecca Wells	FLOOR	9.950
NC State vs Clemson	Rebecca Wells	BEAM	9.950
NCAA Gymnastics Championship '24	Kaitlin DeGuzman	BARS	9.900
Pitt vs Clemson	Rebecca Wells	BEAM	9.925
QUAD	Madison Minner	VAULT	9.900
UCLA vs Clemson	Molly Arnold	VAULT	9.975
UNC vs Clemson	Lauren Rutherford	VAULT	9.925

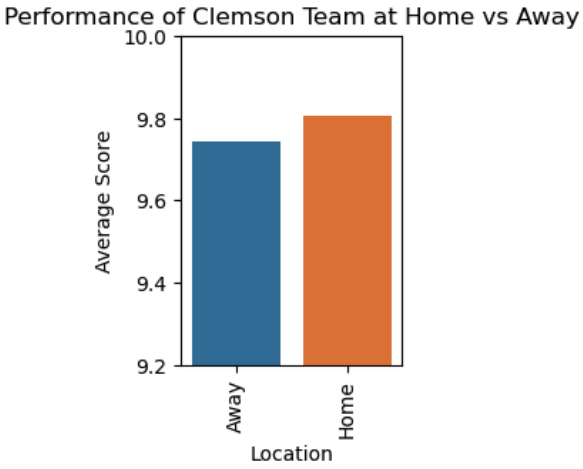
From the above output, we can identify the best gymnast in each meet with the specific event that demonstrates Clemson's strength against each opposition.

c. **Average Score per Event for Clemson Team:** The below bar chart describes the average score secured by the Clemson team in each meet.

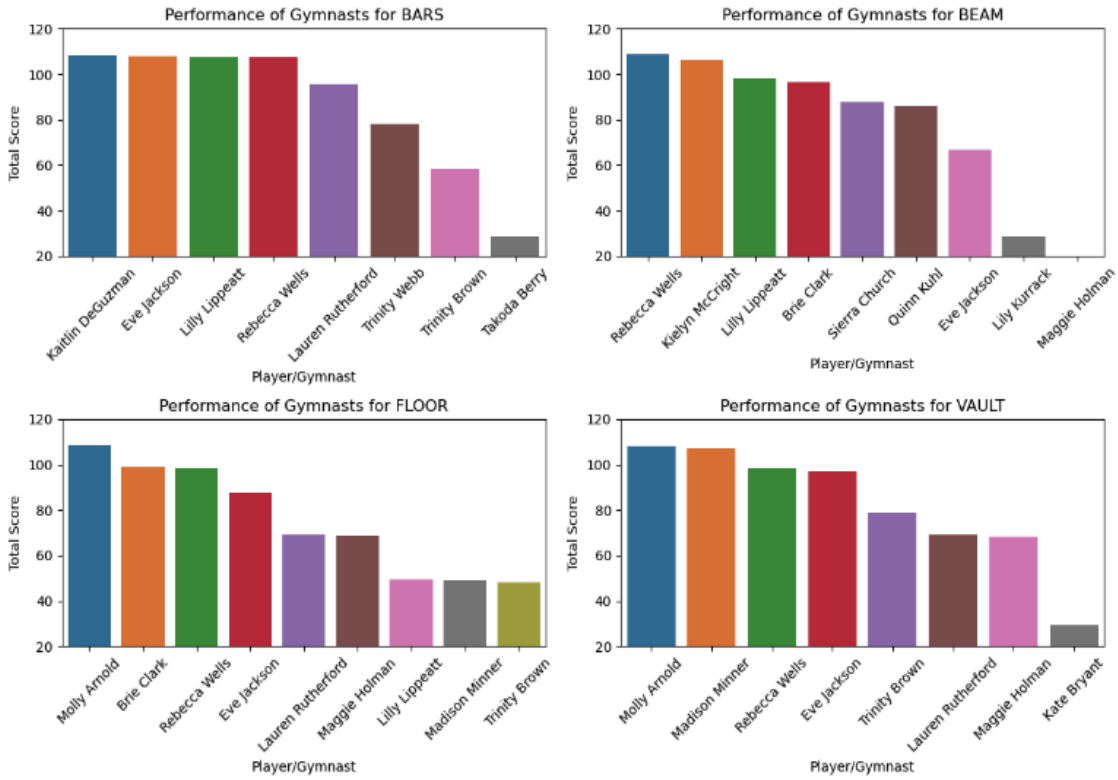


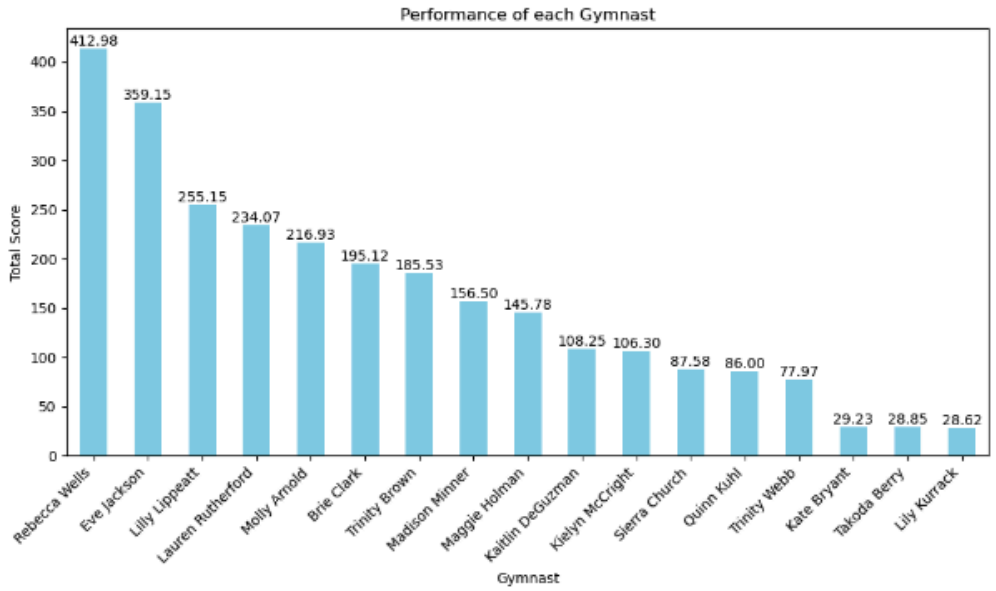
From the above chart, we can say that Clemson has the best average score against the **"Air Force"** and the lowest average score against "**NC State"**.

d. **Performance of Clemson Team at Home vs Away**: Out of the 11 available meets, Clemson participated in 4 meets at home and 7 meets away from home. Analyzing the average team score at these meets shows that Clemson had a better performance at home meets compared to the away meets, as illustrated below.



e. **Additional Insights:** I have also worked on in-depth analysis of individual gymnasts' performances across different events (BARS, BEAM, FLOOR, VAULT) and throughout the entire season for the Clemson team. Here are the key statistics gathered from the analysis**.**





From the above graphs, we can conclude that "**Rebecca Wells**" was the highest scorer for the Clemson Gymnastics team followed by "**Eve Jackson**" and "**Lilly Lippeatt**"

4. CONCLUSION

Through this project, I have worked on the Data Analysis of the Clemson Athletics with a specific focus on the **Clemson Gymnastics** team. In part 1, I worked on analyzing customer attendance for a particular meet, identifying key factors influencing attendance. For part 2, I merged the data from all the meets into a single dataset by following proper data pre-processing and data cleaning techniques. Also, I have generated visualizations that give valuable insights regarding the Clemson team/individual gymnasts' performance.