Team Agenda for Midterm Preparation

Meeting Minutes

Datasheet/Metadata N/A

Motivation to Select Data

We chose this dataset because we are all moderately interested in baseball. Sports lends itself nicely to data science because of how much data can be produced. In baseball, each pitch produces an abundance of data: pitch speed, swing speed, ball spin rate, and many others. Not to mention that each of these pitches happens alongside many other unique variables, such as the batter's batting average. Our favorite aspect of baseball is the massive hits, like home runs. Therefore, we chose to analyze the batting statistics of all batters in the league.

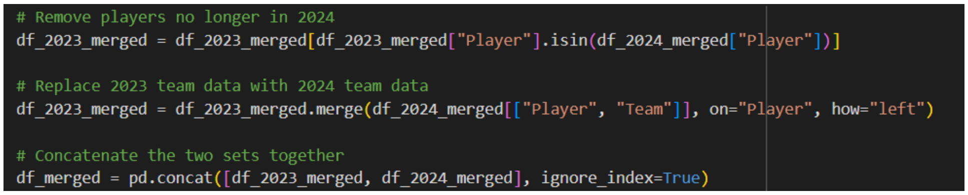
Description of Cleaning, Merging, and Wrangling

Our first step was to add a column for the year of the dataset. Since the plan is to merge both datasets and update all "Team" entries to reflect a player's most recent team, we need some way to differentiate two entries of the same player. Additionally, we can use these two years to see if a player has improved or not between years and analyze small trends that could help us predict how well a team will perform in 2025.

Once we verified the column names and dataframe shapes were correct, we dealt with missing values. The three columns that contained missing values were sacrifice bunts (SH), sacrifice flies (SF), and games played (G). SH and SF are not extremely important values for measuring a player's offensive prowess and they tend to be relatively slow, so we just replaced all missing values with 0. G, however, is important. If a player has a low number of games played (less than 20 or so), their batting average, among other stats, will not be representative since they are based on a low sample size. To fill in missing values, we computed another statistic, plate appearances (PA), and divided that by 4 to get games played.

data_cleaning_merging_wrangling_11.ipynb

For duplicate players within a dataset, we merged them into one player and aggregated their stats to represent their entire season. After this, we standardized column formats and then merged the two datasets. This required a few steps. First, we dropped the team column from the 2023 dataset and all players that did not exist in the 2024 dataset. Then, we performed a left merge on "Player" to add the 2024 teams column to the 2023 dataset. We lastly did a long merge to concatenate the two datasets.

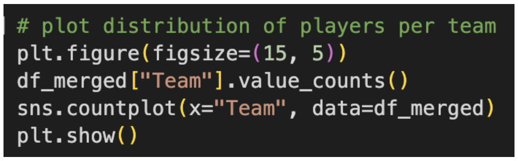


```
# Remove players no longer in 2024
df_2023_merged = df_2023_merged[df_2023_merged["Player"].isin(df_2024_merged["Player"])]

# Replace 2023 team data with 2024 team data
df_2023_merged = df_2023_merged.merge(df_2024_merged[["Player", "Team"]], on="Player", how="left")

# Concatenate the two sets together
df_merged = pd.concat([df_2023_merged, df_2024_merged], ignore_index=True)
```

Finally, we did a little bit of filtering. We checked for games played (G) outliers and batting average (AVG) outliers. Even the best players rarely have above 0.400 average, so we removed all entries above 0.400 and at 0.000. For games, as we previously mentioned, players with 20 or less games played do not provide a representative sample of their skill, so they should be removed.

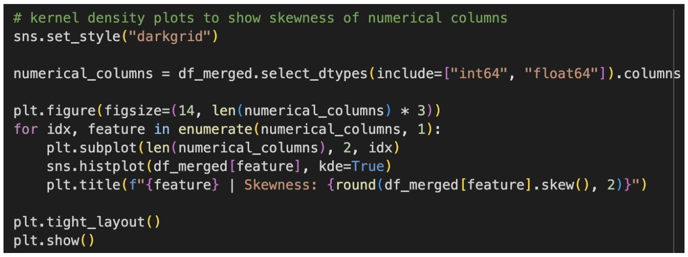
Descriptive Exploratory Analysis

We started with a univariate analysis to examine individual variables and their distributions. Checking the number of players assigned to each team to understand team composition and highlight any imbalances, as well as kernel density plots to evaluate the skewness and distribution shape of numerical columns.



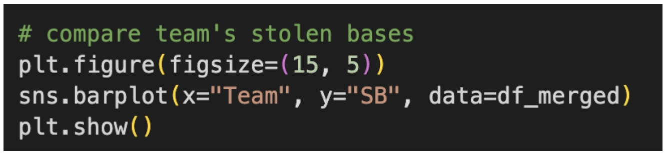
```
# plot distribution of players per team
plt.figure(figsize=(15, 5))
df_merged["Team"].value_counts()
sns.countplot(x="Team", data=df_merged)
plt.show()
```

exploratory_data_analysis_11.ipynb

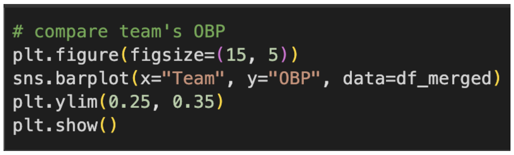


```
# kernel density plots to show skewness of numerical columns
sns.set_style("darkgrid")

numerical_columns = df_merged.select_dtypes(include=["int64", "float64"]).columns

plt.figure(figsize=(14, len(numerical_columns) * 3))
for idx, feature in enumerate(numerical_columns, 1):
    plt.subplot(len(numerical_columns), 2, idx)
    sns.histplot(df_merged[feature], kde=True)
    plt.title(f"{feature} | Skewness: {round(df_merged[feature].skew(), 2)}")

plt.tight_layout()
plt.show()
```

Then, we conducted a bivariate analysis to look for patterns in batting statistics based on teams. We looked at stolen bases (SB), on-base percentage (OBS), batting average (AVG), and home runs (HR) to help determine whether certain teams were consistently better in specific areas. exploratory_data_analysis_11.ipynb

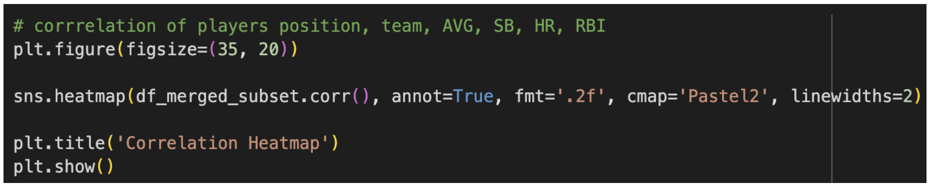


```
# compare team's stolen bases
plt.figure(figsize=(15, 5))
sns.barplot(x="Team", y="SB", data=df_merged)
plt.show()
```



```
# compare team's OBP
plt.figure(figsize=(15, 5))
sns.barplot(x="Team", y="OBP", data=df_merged)
plt.ylim(0.25, 0.35)
plt.show()
```

That was followed by a multivariate analysis to see interactions and correlation between several variables simultaneously. We looked at a player's position (Pos), team, batting average (AVG), stolen bases (SB), home runs (HR), and runs batted in (RBI) to see the relationship of their performance. exploratory_data_analysis_11.ipynb



```
# corrrelation of players position, team, AVG, SB, HR, RBI
plt.figure(figsize=(35, 20))

sns.heatmap(df_merged_subset.corr(), annot=True, fmt='.2f', cmap='Pastel2', linewidths=2)

plt.title('Correlation Heatmap')
plt.show()
```

Exploratory Analysis Findings, Challenges, and Next Steps

During the univariate analysis, the distribution of the data appeared relatively normal, with no significant outliers or unusual patterns. However, the bivariate analysis revealed some trends. Stolen Bases (SB) and Batting Average (AVG) showed a large degree of variability across teams, suggesting differences in performance between teams in these areas. Home Runs (HR) also showed considerable variance, indicating that some teams have many strong hitters. On the other hand, on-base percentage (OBP) offered low variability, meaning that teams were relatively consistent at getting on base. The multivariate analysis and correlation plot highlighted relationships within the data. A moderate correlation of 0.49 was found between Batting Average (AVG) and Runs Batted In (RBI), indicating that players with higher batting averages also tend to accumulate more RBIs. There was another moderate correlation of 0.39 between Batting Average (AVG) and Home Runs (HR), showing a relationship between hitting average and power. An interesting correlation of 0.23 was found between a player's Position (Pos) and Stolen Bases (SB), suggesting that players in certain positions might be more likely to steal bases than others.

Some challenges are missing players from the 2024 season who did not play in the 2023 season. Some players have changed teams for the 2025, and their statistics are still attributed to their old teams. We do not have data on player injuries or other status updates, which could have an impact on the performance of certain statistics.

The next steps are to break down this data more so that each team can compare the team's performance and identify trends. As well as look deeper into correlations since some statistics have a high variance. We will also look at a broader range of time to understand how teams evolved and forecast the future team performance better.

GitHub Link
https://github.com/csc442-team11/MLB-Data-Analysis.git
Cleaning, Merging, and Wrangling Google Colab Notebook
https://colab.research.google.com/drive/19WI1HbwaBem45U1j-MAMEhIr5qI1qyMa?usp=sharing
Exploratory Data Analysis Google Colab Notebook
https://colab.research.google.com/drive/1GY8oFHB6yD_9rchl7iA8kN6CLqczOZGu?usp=sharing