MDA 720 Capstone Project: Fantasy Baseball Analytics Platform

By: Eoin Gawronski

- ➔ Background
- ➔ Objectives/Goal of the Project
- ➔ Data Extraction
- ➔ Data Exploration
- ➔ Data Visualizations
- ➔ Data Analysis/Data-Text Mining
- ➔ Conclusions/Recommendations
- ➔ Works Cited

Background

“Fantasy Baseball is a game in which the participants serve as owners and general managers of virtual baseball teams. The competitors select their rosters by participating in a draft in which all relevant Major League Baseball (MLB) players are available. Fantasy points are awarded in weekly matchups based on the actual performance of baseball players in real-world competition”. This definition provides a quick understanding of fantasy baseball and is ultimately important in understanding the business idea I will recommend. As a baseball fan and fantasy player, I thought of an idea to create a platform that allows fantasy baseball users to enhance their teams and performances through an analytical approach. The approach was to look at Google Trends and data from Baseball Savant to create a platform.

Objective/Goals

Ultimately, the business objective is to develop a fantasy baseball analytics platform that allows users to see optimal lineups, receive trade recommendations, and see player trends based on batting statistics over five years. The player stats over the five year period will provide detail of

how the player has performed over the years, with the most recent data in 2024 being the most important.

Data Extraction

Statcast Custom Leaderboards | baseballsavant.com (mlb.com)

Above is the link to the dataset I used.

The data extraction process began with an attempt to web scrape from the baseball savant data. I attempted to web scrape the data from the 2020-2024 season in attempts to use the statistics provided by the website. I was able to access the data and use the dataset which included 723 rows and 27 columns. The rows include MLB player names ranging from 2020-2024, and the columns include different batting statistics. I chose to use this data because it is relatable data to fantasy baseball and would help with player selection. I also extracted data using Google trends.This search allowed me to hone in on data regarding fantasy baseball, more than player data. I looked specifically at search interest for the term ‘fantasy baseball’ and different fantasy baseball applications.

Some of these included ESPN, Yahoo, CBS, and Fantrax.

last_name, first_name	player_id	year	player_age	ab	pa	hit	single	double	triple	...	on_base_plus_slg	woba	xwoba	sweet_spot_percent	barrel_batted_rate	h
Cabrera, Miguel	408234	2020	37	204	231	51	37	4	0	...	0.746	0.323	0.379	36.8	9.7	
Cruz Jr., Nelson	443558	2020	39	185	214	56	34	6	0	...	0.992	0.411	0.383	39.4	15.0	
Peralta, David	444482	2020	32	203	218	61	45	10	1	...	0.772	0.333	0.299	29.4	5.0	
Longoria, Evan	446334	2020	34	193	209	49	31	10	1	...	0.722	0.308	0.364	29.9	11.5	
Cabrera, Asdrúbal	452678	2020	34	190	213	46	26	9	3	...	0.752	0.319	0.317	30.5	6.5	

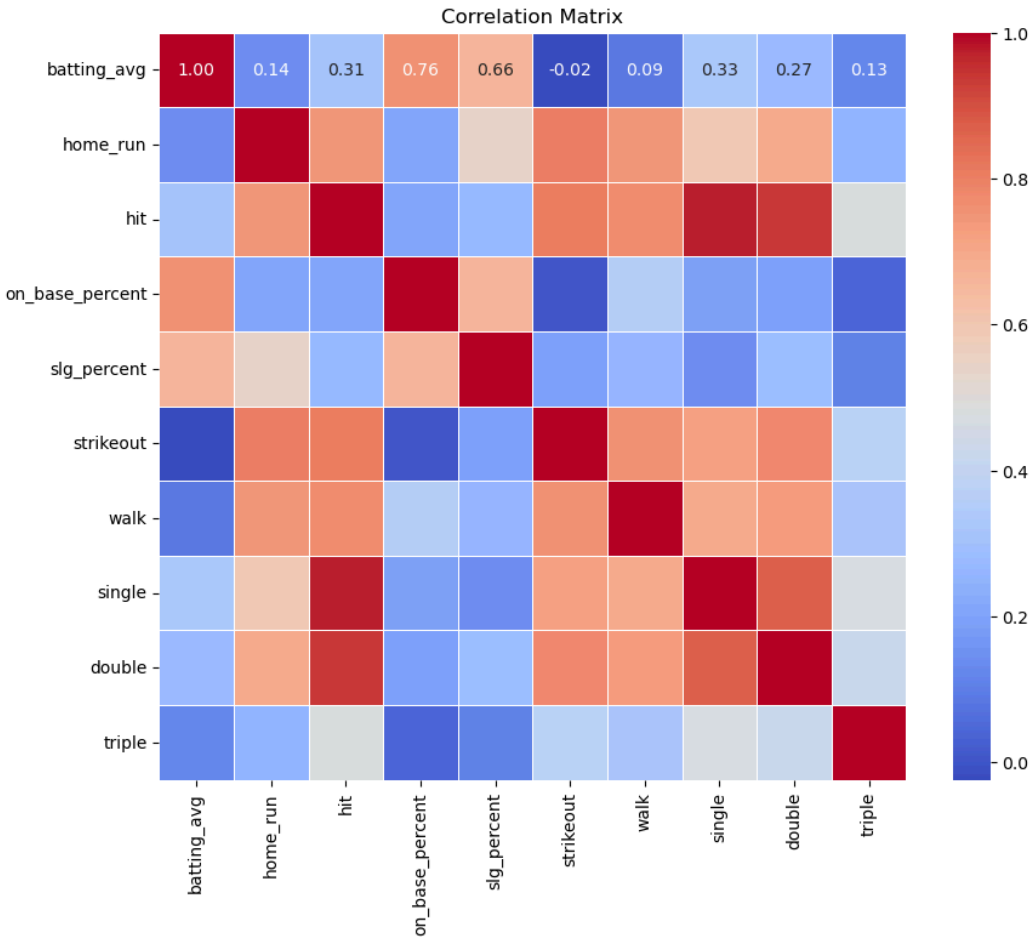
Data Exploration

Data exploration started by checking the data types, exploring missing values, and analyzing the summary statistics. I also used feature engineering to aggregate selected season stats that seemed important to a player's overall statistics. The aggregation included player name and year along with the sum and averages of hit, home_run, batting_avg, and on_base_percent. I have attached the result of this process below.

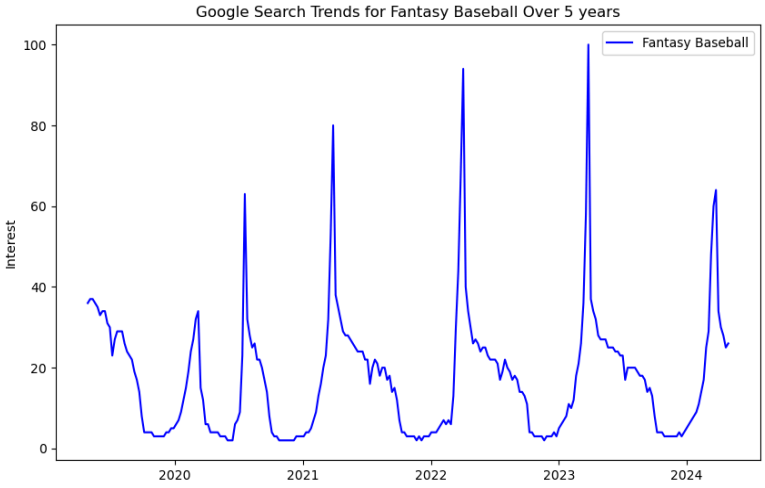
		hit	home_run	batting_avg	on_base_percent
last_name, first_name	year				
Abrams, CJ	2023	138	18	0.245	0.300
	2024	28	6	0.301	0.369
Abreu, José	2020	76	19	0.317	0.370
	2021	148	30	0.261	0.351
	2022	183	15	0.304	0.378
...	...	...	...	...	...
Yelich, Christian	2020	41	12	0.205	0.356
	2022	145	14	0.252	0.355
	2023	153	19	0.278	0.370
Yoshida, Masataka	2023	155	15	0.289	0.338
	2024	21	2	0.269	0.345

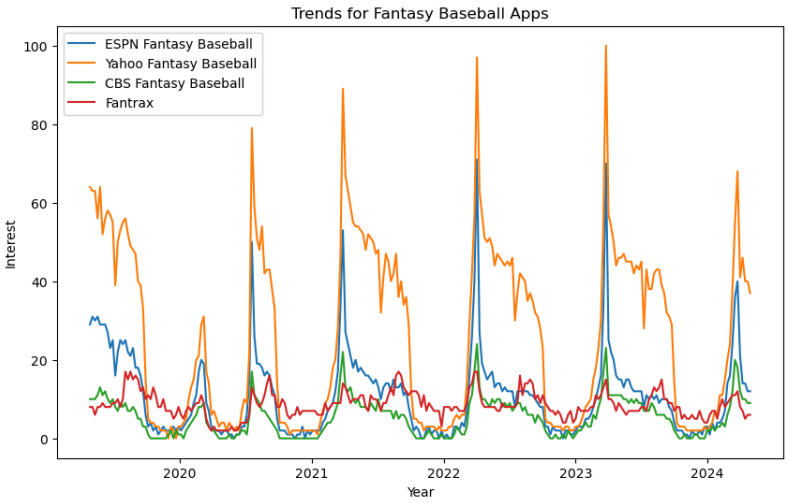
Data Visualization

The correlation matrix shows relationships between different features from the dataset. The darker red shows a high correlation between variables, but it's important to take note of multicollinearity. This matrix helps narrow down features when selecting for models.

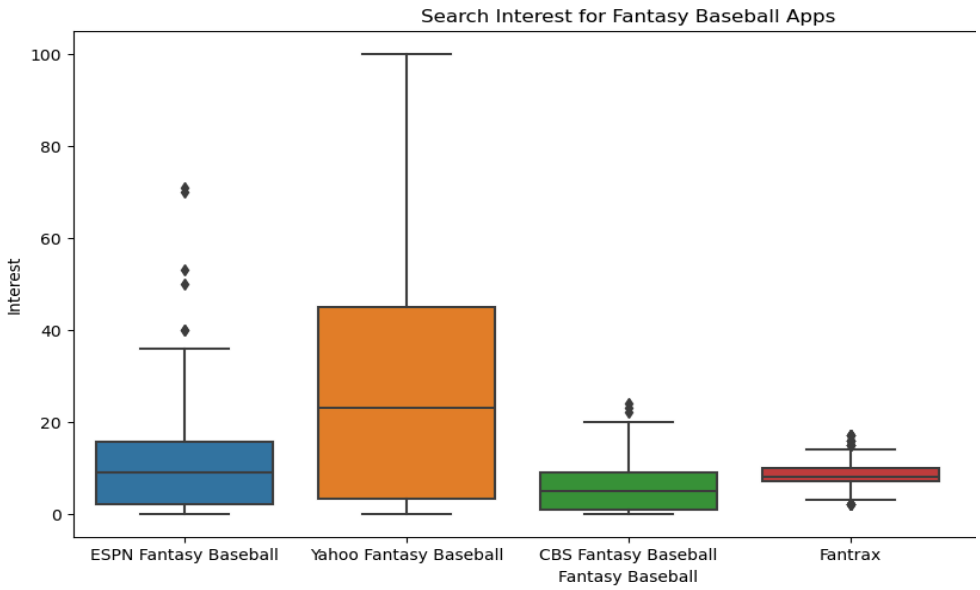


These two line graphs show different Google trend topics that I selected to relate to my project. The left graph shows the term “Fantasy baseball”, which shows the varying up and down interest of the term. It was interesting to see how the graph changes based on baseball season. The right chart shows four different fantasy baseball apps, which is also important data for my business idea. It’s important to see competitors when trying to create a platform.





The boxplots of the different apps are below and help show the mean interest for each app, but also show the outliers for each app. The outliers are always important when analyzing the data and seeing if it's skewed.



Data Analysis/Data Mining/Text Mining

Some of the data analysis that I looked at was running a linear regression model with training and testing data. I narrowed down features such as the x train, and I assigned a y variable to test the model. The features used as x predictors were “hits”, “home runs”, “at bats”, “walks”, “strikeouts”, and “singles”. The y variable chosen was “batting average”. This test was run to see how the x variables impact the batting average percentage. After training the data, I ran some tests that included R-squared, RMSE, MAE, and MSE. These metrics are important to determine the accuracy of the selected linear model.

```
Mean Squared Error: 0.0004977768812271636
R-squared: 0.5945255061263812
Mean Absolute Error: 0.015060944734466822
Mean Squared Error: 0.0004977768812271636
```

Conclusions/Recommendations

Based on my research and analysis of my data, the creation of a platform for fantasy baseball statistics is complex, but it was very interesting to try and apply google trends and the data to create models and graphs. I learned that to create a successful platform it's important to understand and analyze player statistics to see optimal lineups and other key fantasy features. I believe that there was deeper analysis to be done if there was a way to access more data related to trades, player health, and other factors that may impact lineup decisions and fantasy decisions. I would recommend looking into more data scraping to get this information on the web.