Darren Parry
Dr. Ning Wang
DSCI 549
25 April 2025

Homework 7

Project Description

My project that I proposed in the first homework of this course was something baseball related because of my keen interest in sports and more specifically baseball. As all of my data is from the Lahman Baseball Database, there are tons of information and statistics that need to be sorted and cleaned throughout. The Lahman Baseball Database is a comprehensive tabular form of a dataset that analyzes relationships between individual statistics and team statistics. To sort through the extensive data, I first need to find out the relative statistics needed for my project. I want to figure out if there is a relationship between individual statistics and team statistics, and in particular, which ones are statistically significant, whether it be all-star appearances, team win totals, or championships. This project will utilize quantitative analysis techniques and machine-readable data, with a focus on predictive modeling and quantitative analysis, using correlation or regression analysis, given that the independent and dependent variables are more likely to be continuous variables. My plan for this project is to determine whether team statistics including wins or championships affect player performance, like batting average, homeruns, runs batted in, on-base percentage, etc. We can then use these statistics to build other predictive models including projecting future performance based on historical averages or predicting binary outcomes, like all-star appearances. Here is the Dropbox link to the .csv files:
https://www.dropbox.com/scl/fi/hy0sxw6gaai7ghemrshi8/lahman_1871-2023_csv.7z?rlkey=edw1u63zzxg48gvpcmr3qpnhz&dl=0

Elevator Pitch

This project explores how individual baseball player statistics relate to team success by analyzing data from the Lahman Baseball Database. Using statistical modeling and predictive analytics, it aims to identify which individual performance metrics, like home runs or OPS, most significantly predict team success, helping to forecast outcomes such as wins, championships, and All-Star selections.

Project Objectives

Regarding the input datasets, there are multiple CSV files that make up the Lahman Baseball Datasets. These datasets that I plan on using are Batting.csv, People.csv, and Teams.csv, Managers.csv, and AwardsPlayers.csv. These are the main files that I will be using in the project because of the relevant information of statistics needed in all the files. As seen in *Figure 1,* the batting data represents each player in MLB history and their seasonal hitting stats. Each player is represented by a unique player, where it can be traced to their full name with personal details in the People data set, as seen in *Figure 2*. In Batting.csv, each row represents a unique player-season with numerical attributes linking each player to the season. This table has a playerID, which is a primary key to link this

player to other tables, a yearID, teamID, games played, and relevant batting statistical information, like at bats, runs, hits, doubles, triples, homeruns, runs batted in, stolen bases, walks, strikeouts, etc. Since it is comprised of hundreds of thousands of rows, it considered big data due to the sheer volume of the table. People.csv has relevant information about the players name, birthdate, and demographics. This table is only relevant because it links the players statistics to their full name. I will not be testing categorical data like demographic factors to player success, although that could be an interesting project for the future. Teams.csv, Managers.csv, and AwardsPlayers.csv include the names of teams and managers throughout, as well as the awards the players have won. They all link through primary keys, like playerID, teamID, and yearID. These datasets are good to use due to their interoperability because their tables are standardized with easy to understand primary and foreign key relationships. This database is licensed under a Creative Commons Attribution ShareAlike 3.0, which means when distributing this data, you must share it with the same license that the original data is from. This allows users to use, adapt, and share the data, as long as they credit the original creator as needed.

To prepare the data for analysis, I plan to implement certain data pre-processing steps to ensure a smoother analysis workflow. There are numerous ways to pre-processing data and there is not one correct approach, like you can use tokenization, stop word removal, removing punctuation and special characters, normalization, stemming, lemmatization, etc. For my project, I first plan on data cleaning to ensure that duplicate records are removed, figure out which columns or rows have missing values, and standardizing column names. I plan on then using data integration to merge multiple datasets using relational joins on keys like playerID and teamID. If I had the Batting.csv and I wanted to know how many wins their team had or the first name of the player, I could easily have that all in one table if I were to merge the certain tables to one big table. The most exhaustive part of pre-processing this data would be filtering players with insufficient data out of the table. This would mean that I would eliminate players on the batting table if they didn't have a certain number of at-bats because it means that they didn't play that much, meaning they didn't contribute a lot to their team. Another thing I could do is add relevant stats to their table based on the current stats in the table. For example, a predictive stat used a lot in baseball lingo is on-base plus slugging (OPS), so I would add a mathematical equation to get OPS, which is on-base percentage plus slugging percentage. On-base percentage is walks divided by plate appearances where slugging percentage is a bit more complicated, (1*singles + 2*doubles + 3*triples + 4*homeruns) / at-bats. I will also make sure the data is normalized and make sure the numeric features are scaled to standardize the data for regression analysis. I can also use data transformation to make the data easier to read, by converting categorical variables, like team names, player positions, etc. into dummy variables. The primary goal of this analysis is to discover and try and quantify patterns in individual player performance and team outcomes by applying a combination of descriptive statistics, inferential statistical modeling, and predictive analytics techniques.

playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	IBB	HBP	SH	SF	GIDP	G_old
aardsda01	2004	1	SFN	NL	11		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
aardsda01	2006	1	CHN	NL	45		2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
aardsda01	2007	1	CHA	AL	25		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
aardsda01	2008	1	BOS	AL	47		1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
aardsda01	2009	1	SEA	AL	73		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
aardsda01	2010	1	SEA	AL	53		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
aardsda01	2012	1	NYA	AL	1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
aardsda01	2013	1	NYN	NL	43		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
aardsda01	2015	1	ATL	NL	33		1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
aaronha01	1954	1	ML1	NL	122		468	58	131	27	6	13	69	2	2	28	39		3	6	4	13	
aaronha01	1955	1	ML1	NL	153		602	105	189	37	9	27	106	3	1	49	61	5	3	7	4	20	
aaronha01	1956	1	ML1	NL	153		609	106	200	34	14	26	92	2	4	37	54	6	2	5	7	21	
aaronha01	1957	1	ML1	NL	151		615	118	198	27	6	44	132	1	1	57	58	15	0	0	3	13	

Figure 1. Batting.csv

ID	playerID	birthYear	birthMonth	birthDay	birthCity	birthCountry	birthState	deathYear
1	aardsda01	1981	12	27	Denver	USA	CO	
2	aaronha01	1934	2	5	Mobile	USA	AL	2021
3	aaronto01	1939	8	5	Mobile	USA	AL	1984
4	aasedo01	1954	9	8	Orange	USA	CA	
5	abadan01	1972	8	25	Palm Beach	USA	FL	
6	abadfe01	1985	12	17	La Romana	D.R.	La Romana	
7	abadijo01	1850	11	4	Philadelphia	USA	PA	1905
8	abbated01	1877	4	15	Latrobe	USA	PA	1957
9	abbeybe01	1869	11	11	Essex	USA	VT	1962
10	abbeych01	1866	10	14	Falls City	USA	NE	1926
11	abbotco01	1995	9	20	San Diego	USA	CA	
12	abbotda01	1862	3	16	Portage	USA	OH	1930

Figure 2. People.csv

Project Scope

This project is focused on analyzing the relationship between offensive batting statistics and success metrics at both the individual and team levels using a polished form of the Lahman Baseball Database. While the study will explore certain offensive statistics like batting averages, home runs, runs batted in, OPS, and other offensive metrics and how they correlate with team wins, championships and other team centered achievements, it will exclude several aspects of the team game. Defensive and pitching metrics will be intentionally left out because I want to solely focus on one aspect of the game, even though pitching and defensive are equally as important as the offensive side of the ball. Demographic or categorical variables will not be in the scope of this project because I only want to focus on numerical data, so I can perform the necessary analysis within. If I wanted to focus on these categorical variables, that would be for another project. Nationality and which way a batter hits, right or left, is also an equally metric in determining

success of an individual player or team. There are other certain temporal factors that will not be included in this project like rule changes or variation in season length that can affect player or team performance. For example, batting stats during the steroid era of baseball, from the late 1980s to the late 2000s, will be skewed positively towards the hitters because of the immense number of players using steroids in that era. Of course, the external factors in the data will not be included, but it will be addressed in this section, like the managerial strategies of teams and certain weather conditions where each team plays. The focus on this project remains strictly on offensive performance and its quantifiable connection to individual and team success in Major League Baseball.

Data Analysis and Solution

In this project the dependent or target variables will be team-level and player-level success metrics, like total wins in a season, winning percentage, number of postseason appearances, and championships for team-level and number of All-Star selections and accumulated awards for player-level metrics. These are the metrics I intend to predict or explain as causes. The independent variables or predictor variables in this project will be in the individual statistics from the batting dataset. These continuous statistics include at bats, runs, hits, doubles, triples, homeruns, runs batted in, stolen bases, walks, strikeouts, and my derived metrics like OPS. My first analysis I will perform is exploratory data analysis by using this data to create visualizations, like scatterplots and heatmaps to visually identify outliers or other variables distributions. For the independent variables, I will also compute summary statistics, like them mean, median, standard deviation, so I can set a baseline for the metrics. I will then perform linear regression where team wins or championships are predicted using batting statistics and I can evaluate the model using confidence intervals, p-values, etc. to determine statistical significance. For players with long enough careers, I plan on using expected value estimation for the predictive analysis part of this project by using the historical data to compute weighted averages for players to estimate future seasons performances. As seen in *Figure 3*, I sketched a workflow of the entire data preprocessing and data analysis workflow through the entire project. The inputs and outputs of the dataset throughout the project are represented with the blue rectangles, while the orange ovals represent the functions that will be used throughout. As you can see, there is a clean workflow that goes from inputs to functions to outputs, while the outputs are the inputs for the next function.

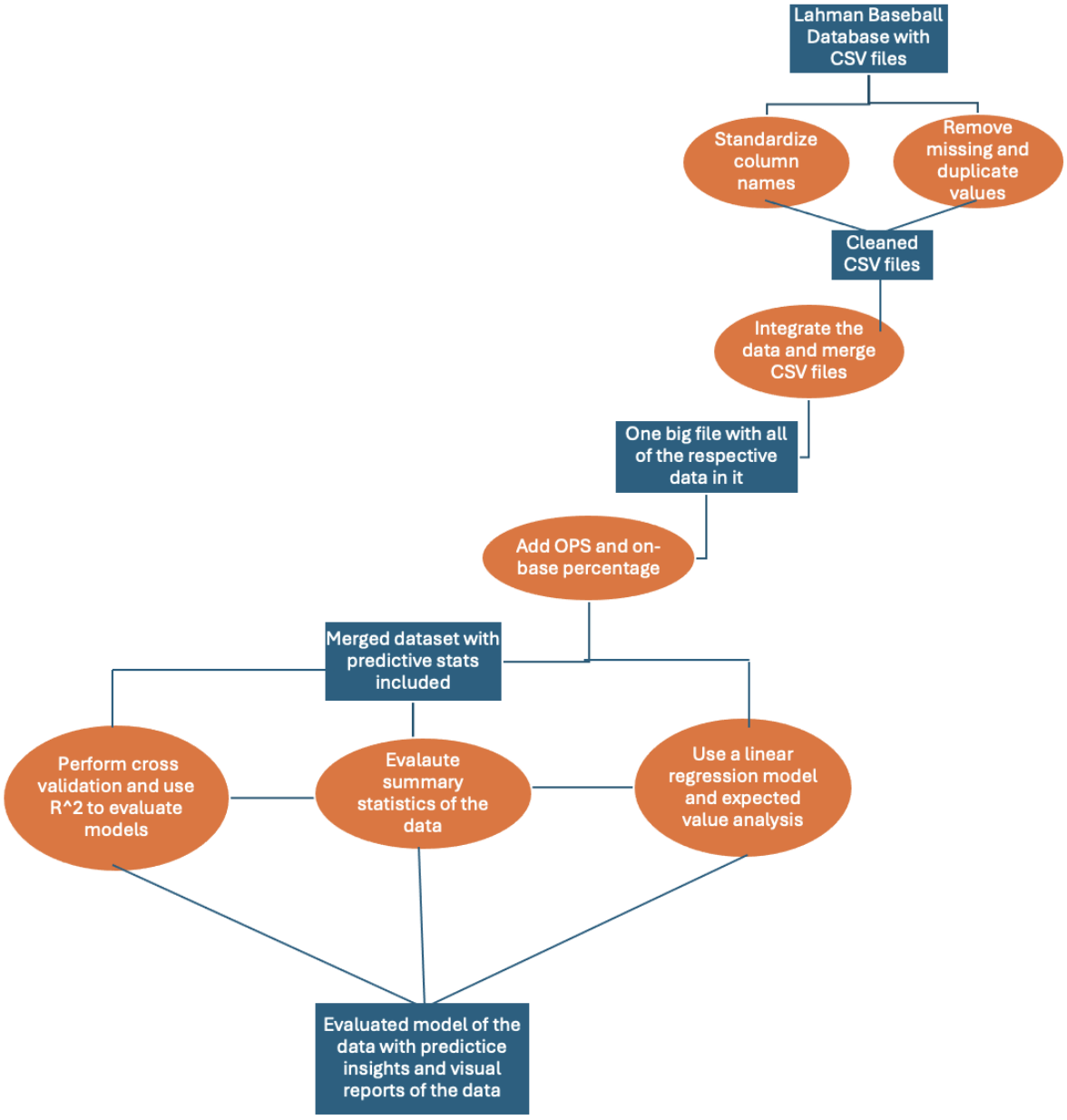

Figure 3. Workflow

Project Outcome

The anticipated outcome of this project includes the development of a well-structured, cleaned, and merged dataset derived from multiple CSV files in the Lahman Baseball Database. This dataset will serve as the foundation for exploratory visualizations that highlight patterns and correlations in batting performance with team and individual performance. Using regression analysis, statistical modeling techniques, and predictive analysis, this project aims to identify which offensive statistics have the strongest predictive power regarding team success metrics like wins and championships and individual success factors as well with All-Star selections and awards. This project will also

illustrate feature importance, indicating which batting metrics most reliably affect future performance and whether players will have a long, successful, consistent tenure in the Major Leagues. In the end, the project is meant to be used in a set of predictive insights and models that could inform scouting, team building and player evaluation efforts in the scope of baseball analytics as a whole.

Appendix I Homework 1

I want to center my project on something baseball related, so in this paragraph I will explain a couple of baseball data sets. All of my data sets are from the Lahman Baseball Database. This data represents every single statistic and player in Major League Baseball (MLB) history. It includes hitting, pitching, fielding, win percentages and many more statistical baseball data from 18721 to 2023. This database includes the following data sets: People, Teams, Teams Franchises, Parks, Batting, Pitching, Fielding OF, Fielding OF split, Appearances, Managers, All Star Full, Batting Postseason, Pitching Postseason, Fielding Postseason, Series Postseason, Home Games, Managers Half, Teams Half, Awards Managers, Awards Players, Awards Share Managers, Awards Share Players, Hall of Fame, College Playing, Salaries, and Schools. The half stats represent how each preforms in half of a season. As seen in *Figure 1,* the batting data represents each player in MLB history and their seasonal hitting stats. Each player is represented by a unique player, where it can be traced to their full name with personal details in the People data set, as seen in *Figure 2*. This data is publicly accessible as I accessed it through the Sean Lahman's personal website, http://seanlahman.com. The .csv files can be downloaded through a Dropbox link, where it has the necessary legal information. This database is licensed under a Creative Commons Attribution ShareAlike 3.0, which means when distributing this data, you must share it with the same license that the original data is from. This allows users to use, adapt, and share the data, as long as they credit the original creator as needed. This data is also machine processable because it is structured in rows and columns and can be automatically read and processed in certain applications within the computer. On my Mac computer, I opened the .csv file in Numbers, however, when we are working on the tabular data and cleaning it, I probably need to use a different interface because the Batting file is considered "Big Data", as all of the cells cannot be loaded into the application. This dataset is considered big data because it has volume, variety, and velocity. The Numbers application can only process 65,535, where my data has 113,805 rows. This data is also simple to use because it is structured in rows and columns and there are primary and candidate keys within each dataset to compare them, which proves that this data is interoperable as well. The data quality seems to match up to the best because of well-organized the format is throughout the information. The only thing I am worried about is filtering out the players who only played one season or who did not have a lot of game experience. Here is the Dropbox link to the .csv files: https://www.dropbox.com/scl/fi/hy0sxw6gaai7ghemrshi8/lahman_1871-2023_csv.7z?rlkey=edw1u63zzxg48gvpcmr3qpnhz&dl=0

playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	IBB	HBP	SH	SF	GIDP	G_old
aardsda01	2004	1	SFN	NL	11		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
aardsda01	2006	1	CHN	NL	45		2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
aardsda01	2007	1	CHA	AL	25		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
aardsda01	2008	1	BOS	AL	47		1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
aardsda01	2009	1	SEA	AL	73		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
aardsda01	2010	1	SEA	AL	53		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
aardsda01	2012	1	NYA	AL	1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
aardsda01	2013	1	NYN	NL	43		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
aardsda01	2015	1	ATL	NL	33		1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
aaronha01	1954	1	ML1	NL	122		468	58	131	27	6	13	69	2	2	28	39		3	6	4	13	
aaronha01	1955	1	ML1	NL	153		602	105	189	37	9	27	106	3	1	49	61	5	3	7	4	20	
aaronha01	1956	1	ML1	NL	153		609	106	200	34	14	26	92	2	4	37	54	6	2	5	7	21	
aaronha01	1957	1	ML1	NL	151		615	118	198	27	6	44	132	1	1	57	58	15	0	0	3	13	

Figure 1. Batting.csv

ID	playerID	birthYear	birthMonth	birthDay	birthCity	birthCountry	birthState	deathYear
1	aardsda01	1981	12	27	Denver	USA	CO	
2	aaronha01	1934	2	5	Mobile	USA	AL	2021
3	aaronto01	1939	8	5	Mobile	USA	AL	1984
4	aasedo01	1954	9	8	Orange	USA	CA	
5	abadan01	1972	8	25	Palm Beach	USA	FL	
6	abadfe01	1985	12	17	La Romana	D.R.	La Romana	
7	abadijo01	1850	11	4	Philadelphia	USA	PA	1905
8	abbated01	1877	4	15	Latrobe	USA	PA	1957
9	abbeybe01	1869	11	11	Essex	USA	VT	1962
10	abbeych01	1866	10	14	Falls City	USA	NE	1926
11	abbotco01	1995	9	20	San Diego	USA	CA	
12	abbotda01	1862	3	16	Portage	USA	OH	1930

Figure 2. People.csv

This data can be analyzed in many different ways because of the vast nature of this nature and how almost every single statistic can be found here in the game of baseball. One can either go the individual player route, where you can analyze how each player performs season after season, either offensively, defensively, or pitching. Or one can analyze a team's performance, where you can predict how often a team is projected win and if there are any common factors associated with certain statistics. Here are some of the individual centered questions I have come up with: Which statistic accurately predicts more success within a batter? Which statistic accurately predicts more success within a pitcher? What factors affect batting performance within a batter? What factors affect pitching performance within a pitcher? What factors go into all-star appearances? Here are some of the team centered questions I have come up with: Do ballparks have any indication to how successful a team is? Do ballparks correlate with championships or awards? If a team is successful in the first half of the season, will they be successful in the second half? What factors going into a successful franchise? I also may try and combine team performance with individual performance, keeping the following questions in mind: What batting performances indicate more championships for a team? Do more all-star appearances indicate more successful seasons for a franchise? These questions can be analyzed from the various .csv files while looking at specific players' batting stats. Their

primary key is their player ID, and you can compare which teams they are on, evaluating each stat and comparing it number of championships won, games won, all-star appearances, etc. Regarding the number of championships won and wins, you could use an OLS regression analysis to see if some batting factors are statistically significant in the sense of each team's performance. For example, we could have a regression analysis on championships won and have the batting variables as batting average, homeruns, singles, doubles, on-base percentage, etc. The outcome for this statistical test would have values that would tell if each predictor variable was statistically significant or not. In the end, my analysis will aim to analyze and predict both team level success metrics and player level success metrics using batting statistics as the primary predictor variables. For team level analysis, this project will explore the question of which team batting statistics best predict a higher number of wins or championships. While the player level analysis will be which offensive statistics most accurately predict a player's All-Star selection or award likelihood and are specific batting metrics consent indicators of future performance across seasons. Ultimately, the goal is to identify statistically significant patterns liking offensive performance to both individual accolades and team achievements through the seasons and try to quantify these relationships using data science.

Appendix II Homework 5

My project that I proposed in the first homework of this course was something baseball related because of my keen interest in sports and more specifically baseball. I still want to expand on that topic for my project, so I will be explaining more to my project in the rest of this assignment. As all of my data is from the Lahman Baseball Database, there are tons of information and statistics that need to be sorted and cleaned throughout. The Lahman Baseball Database is a comprehensive tabular form of a dataset that analyzes relationships between individual statistics and team statistics. To sort through the extensive data, I first need to find out the relative statistics needed for my project. I want to figure out if there is a relationship between individual statistics and team statistics, and in particular, which ones are statistically significant, whether it be all-star appearances, team win totals, or championships. This project will utilize quantitative analysis techniques and machine-readable data, with a focus on predictive modeling and quantitative analysis, using correlation or regression analysis, given that the independent and dependent variables are more likely to be continuous variables. My plan for this project is to determine whether team statistics including wins or championships affect player performance, like batting average, homeruns, runs batted in, on-base percentage, etc. We can then use these statistics to build other predictive models including projecting future performance based on historical averages or predicting binary outcomes, like all-star appearances.

Regarding the input datasets, there are multiple CSV files that make up the Lahman Baseball Datasets. These datasets that I plan on using are Batting.csv, People.csv, and Teams.csv, Managers.csv, and AwardsPlayers.csv. These are the main files that I will be using in the project because of the relevant information of statistics needed in all the files. In Batting.csv, each row represents a unique player-season with numerical attributes linking each player to the season. This table has a playerID, which is a primary key to link this player to other tables, a yearID, teamID, games played, and relevant batting statistical information, like at bats, runs, hits, doubles, triples, homeruns, runs batted in, stolen bases, walks, strikeouts, etc. Since it is comprised of hundreds of thousands of rows, it considered big data due to the sheer volume of the table. People.csv has relevant information about the players name, birthdate, and demographics. This table is only relevant because it links the players statistics to their full name. I will not be testing categorical data like demographic factors to player success, although that could be an interesting project for the future. Teams.csv, Managers.csv, and AwardsPlayers.csv include the names of teams and managers throughout, as well as the awards the players have won. They all link through primary keys, like playerID, teamID, and yearID. These datasets are good to use due to their interoperability because their tables are standardized with easy to understand primary and foreign key relationships. This database is licensed under a Creative Commons Attribution ShareAlike 3.0, which means when distributing this data, you must share it with the same license that the original data is from. This allows users to use, adapt, and share the data, as long as they credit the original creator as needed.

To prepare the data for analysis, I plan to implement certain data pre-processing steps to ensure a smoother analysis workflow. For my project, I first plan on data cleaning to ensure that duplicate records are removed, figure out which columns or rows have

missing values, and standardizing column names. I plan on then using data integration to merge multiple datasets using relational joins on keys like playerID and teamID. If I had the Batting.csv and I wanted to know how many wins their team had or the first name of the player, I could easily have that all in one table if I were to merge the certain tables to one big table. The most exhaustive part of pre-processing this data would be filtering players with insufficient data out of the table. This would mean that I would eliminate players on the batting table if they didn't have a certain number of at-bats because it means that they didn't play that much, meaning they didn't contribute a lot to their team. Another thing I could do is add relevant stats to their table based on the current stats in the table. For example, a predictive stat used a lot in baseball lingo is on-base plus slugging (OPS), so I would add a mathematical equation to get OPS, which is on-base percentage plus slugging percentage. On-base percentage is walks divided by plate appearances where slugging percentage is a bit more complicated, (1*singles + 2*doubles + 3*triples + 4*homeruns) / at-bats. I will also make sure the data is normalized and make sure the numeric features are scaled to standardize the data for regression analysis. I can also use data transformation to make the data easier to read, by converting categorical variables, like team names, player positions, etc. into dummy variables.

The primary goal of this analysis is to discover and try and quantify patterns in individual player performance and team outcomes by applying a combination of descriptive statistics, inferential statistical modeling, and predictive analytics techniques. In this project the dependent or target variables will be team-level and player-level success metrics, like total wins in a season, winning percentage, number of postseason appearances, and championships for team-level and number of All-Star selections and accumulated awards for player-level metrics. These are the metrics I intend to predict or explain as causes. The independent variables or predictor variables in this project will be in the individual statistics from the batting dataset. These continuous statistics include at bats, runs, hits, doubles, triples, homeruns, runs batted in, stolen bases, walks, strikeouts, and my derived metrics like OPS. My first analysis I will perform is exploratory data analysis by using this data to create visualizations, like scatterplots and heatmaps to visually identify outliers or other variables distributions. For the independent variables, I will also compute summary statistics, like them mean, median, standard deviation, so I can set a baseline for the metrics. I will then perform linear regression where team wins or championships are predicted using batting statistics and I can evaluate the model using confidence intervals, p-values, etc. to determine statistical significance. For players with long enough careers, I plan on using expected value estimation for the predictive analysis part of this project by using the historical data to compute weighted averages for players to estimate future seasons performances.

As seen in *Figure 1*, I sketched a workflow of the entire data preprocessing and data analysis workflow through the entire project. The inputs and outputs of the dataset throughout the project are represented with the blue rectangles, while the orange ovals represent the functions that will be used throughout. As you can see, there is a clean workflow that goes from inputs to functions to outputs, while the outputs are the inputs for the next function.

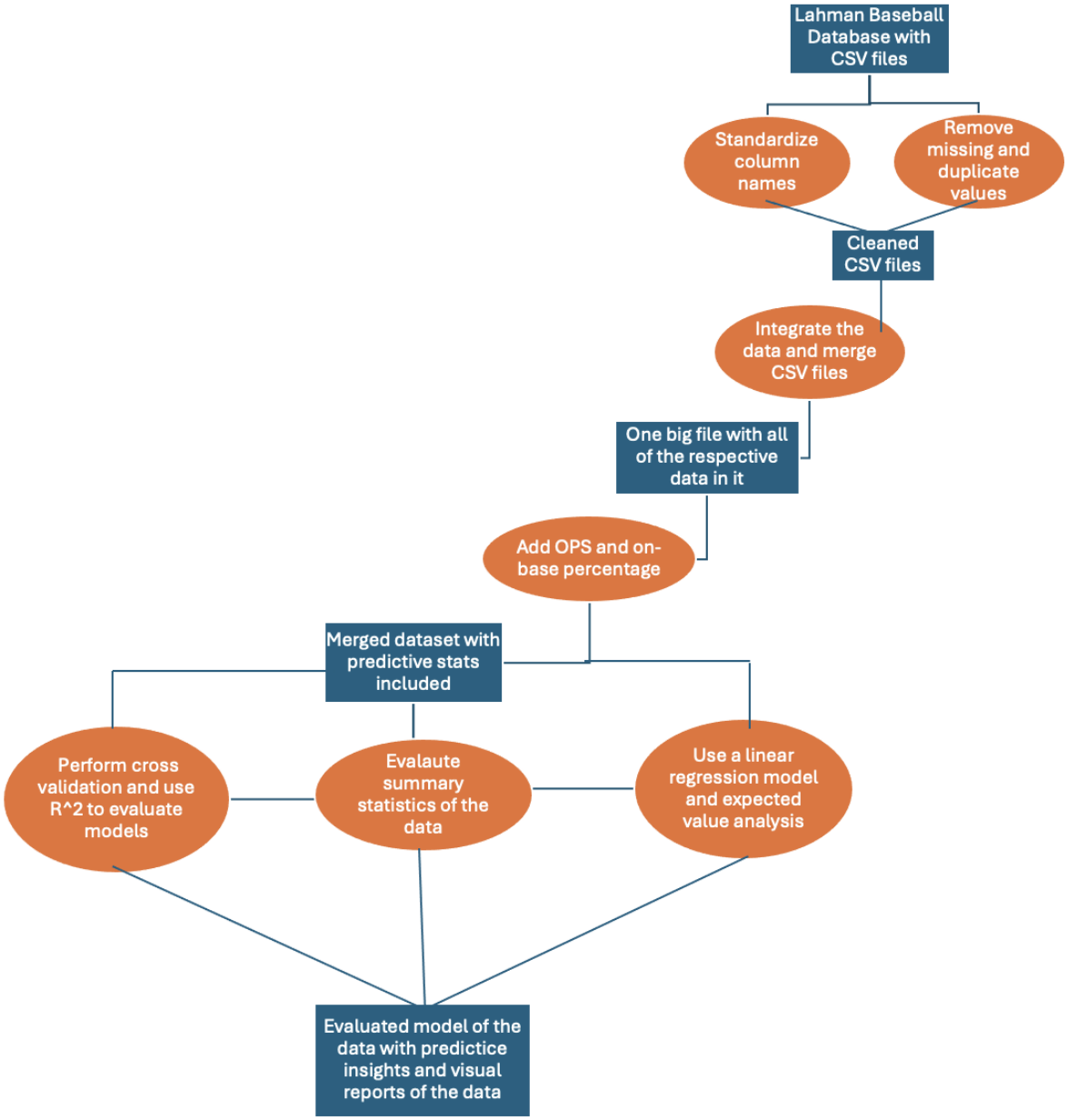

Figure 1. Workflow