Donald McDowell
AIT 580
M7 Project Deliverable 2

This Begins Project Deliverable 1

For this assignment I plan to use a dataset that was listed as one of the possible datasets to use for the M5 A1B assignment, or exercise 6-2 from the textbook. Specifically, I plan to use The Sean Lahman's Baseball Database. I think this data is sufficiently large to get good results while also being small enough to manage the analysis on my local computer. I know some of the files are quite large, but I think as long as I either don't do incredibly complex commands or filter the data beforehand, it will be quite manageable. Additionally, the data is comprised of several csv files which will give me a great opportunity to use SQL within this assignment. The total combined size of all files together is 32.1 MB, but I don't necessarily intend on using all of that. Also, some of the data is duplicated across files so the actual size should be a little smaller. As far as the data itself, it includes anything, and everything related to Major League Baseball. This includes player stats, team stats, descriptive player, managerial, and team data, and individual player seasons. All this data types are included and was collected from 1871-2014.

1. **Who**

This data was collected by a plethora of individuals and corporations over the years. It was compiled in this version by Sean Lahman. According to him, he wanted to make baseball statistics freely available to the general public. He distributes this content in a variety of formats including the csv version I'm going to go with.

2. **Need**

They contributors to this database I'm sure collected this data in part to take on projects themselves and perhaps solve big data problems. There is not only one big data problem, but many embedded in this data set. This dataset has both large volume and a large variety of data. Advanced sports statistics have taken off with advances in modern technology, especially in baseball. Theoretically with data of this type there could be potential ethical or copyright issues. However, Mr. Lahman has licensed this data set to specifically be used by individuals for free. If I were a company rather than an individual, I could have some issues I would need to work out first. As far as any privacy concerns for the people included in the data set, there really aren't any as all of this data is publicly available.

3. What potential questions could be answered by studying this data?

This is the interesting part of this data set. Because it is so expansive, the options are limitless. I do intend to go a little off the beaten path and look at where all the players were born and see if there are areas that produce more MLB players per capita than others. I think examining this question also opens the door for some very interesting visualizations including choropleth maps. I also intend to examine if there are certain stadiums that have a impactful "home field advantage". As this data includes all records from 1871-2014, I should be able to analyze whether or not certain ballparks help the home team. I know solving this question will require a very technical approach and in-depth statistical analysis, but I feel that it is manageable.

4. HW/SW Resources

For this project I will definitely use R and SQL in a large part. I intend to work Python in as well. With multiple data files SQL will be helpful in storing data as well as binding tables together. R is my preferred tool for statistical analysis and high-level visualizations. Python will also come in handy with Pandas and helping gain insights from the data.

This Begins Project Deliverable 2

The dataset I chose to analyze for this project is provided by The Lahman Baseball Database which freely distributes statistics on Major League Baseball (MLB) to the public. This dataset includes: unique player id (nominal character), birth and death date (numeric interval), birth and death location (nominal character), name (nominal character), height and weight (numeric ratio), batting and throwing hand (nominal character), and debut and final game date (numeric interval). I also want to mention that one of the questions I posed in Project Deliverable 1 that I intended to answer here I was unable to. One question I intended to answer was whether there were any ballparks that provided a distinct home field advantage. However, I misunderstood a data type and as such, that question is unanswerable by this data. Nonetheless, I was still able to produce good results on the question of if there were particular geographic locations that produced more players than others. I also found some other interesting trends and was able to utilize some statistical tests to provide concrete evidence for some findings.

SQL Schema

In accordance with the prompt, the SQL schema for this dataset is as follows:

```
CREATE TABLE BASEBALL  (PLAYERID VARCHAR(100),
                        BIRTHYEAR INTEGER,
                        BIRTHMONTH INTEGER,
                        BIRTHDAY INTEGER,
                        BIRTHCOUNTRY VARCHAR(100),
                        BIRTHSTATE VARCHAR(100),
                        BIRTHCITY VARCHAR(100),
                        DEATHYEAR INTEGER,
                        DEATHMONTH INTEGER,
                        DEATHDAY INTEGER,
                        DEATH COUNTRY VARCHAR(100),
                        DEATH STATE VARCHAR(100),
                        DEATHCITY VARCHAR(100),
                        FIRSTNAME VARCHAR(100),
                        LASTNAME VARCHAR(100),
                        GIVENNAME VARCHAR(100),
                        WEIGHT INTEGER,
                        HEIGHT INTEGER,
                        BATS CHARACTER(1),
                        THROWS CHARACTER(1),
                        DEBUT VARCHAR(100),
                        FINALGAME VARCHAR(100),
                        RETROID VARCHAR(100),
                        BBREFID VARCHAR(100);
```

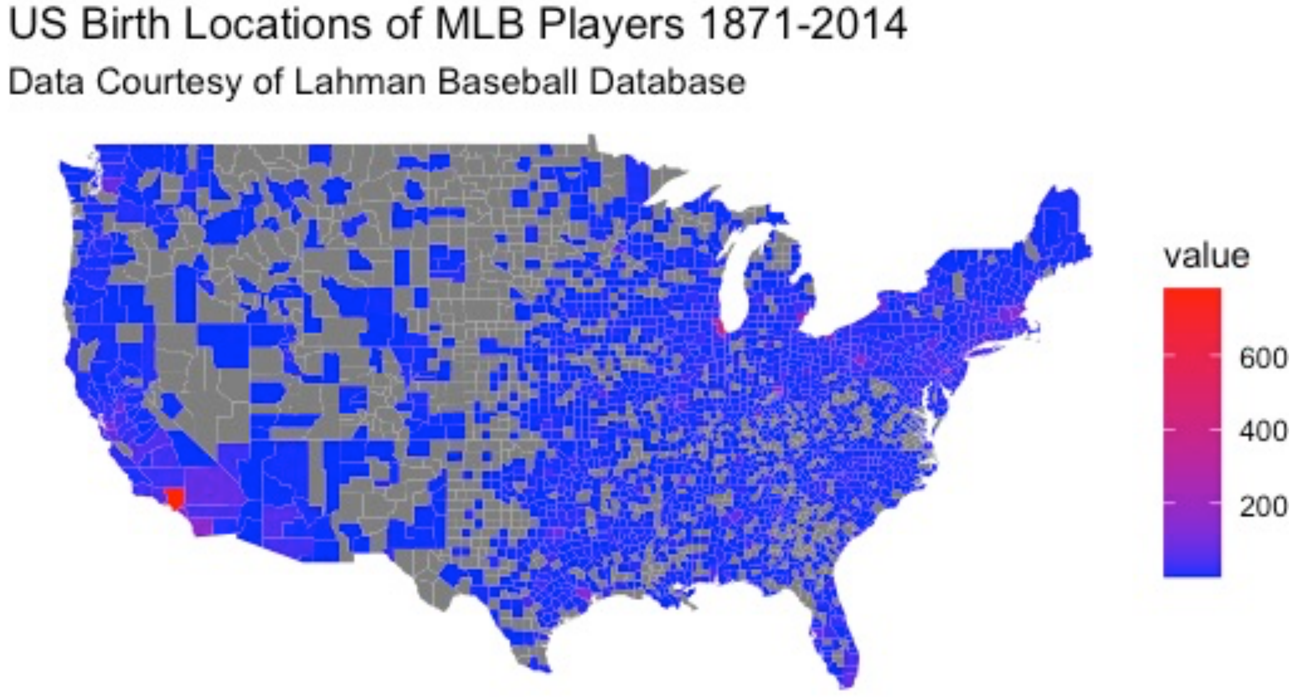
One thing I learned during this project that was key was the “VARCHAR” designation for a indeterminant length. I also did some basic queries just to make sure I had inputted the data correctly.

```
SELECT COUNT(*) FROM BASEBALL

SELECT * FROM BASEBALL WHERE BIRTHSTATE = “OK”

SELECT AVG(WEIGHT) FROM BASEBALL

SELECT AVG(HEIGHT) FROM BASEBALL

SELECT DISTINCT BIRTHCOUNTRY FROM BASEBALL

SELECT DISTINCT BIRTHSTATE FROM BASEBALL
```

Most of these queries were just exploring the data. I was born in, and currently live in Oklahoma so I wanted to see if there were any players from my hometown (there were not). I also had a thought that height and weight probably vary over time but I wasn’t sure how. I figured players would get taller over time and because height and weight are naturally correlated, weight should increase also. I also wanted to get an idea of the birth locations since I really wanted to dig deeper into that variable. Most of the in-depth analysis was conducted in R as I am more comfortable with it for visualization and analysis purposes.

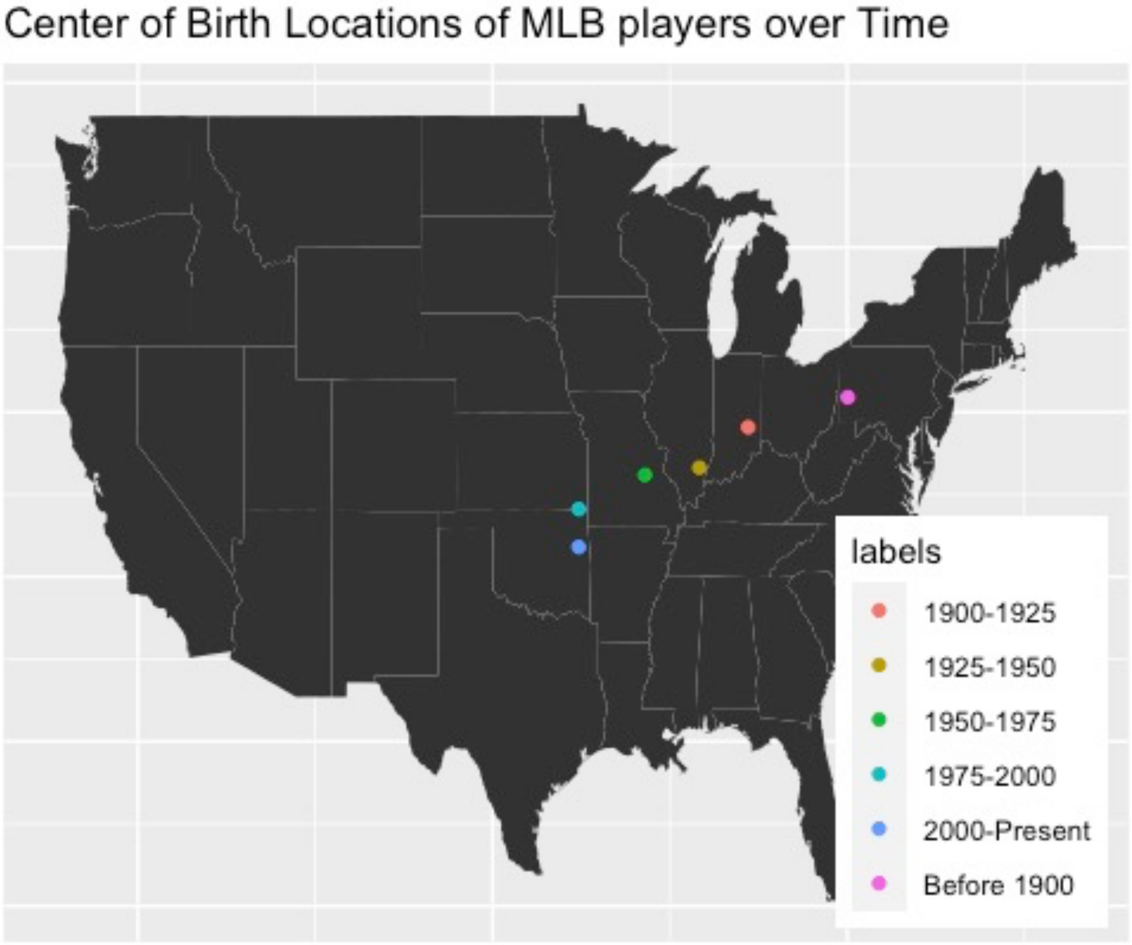
Geographic Analysis

One aspect of this dataset I really wanted to explore was birth location. Because this dataset contains all MLB players who played between 1871 and 2014, any findings can be pretty definitive. I started by focusing only on players born in the United States. This only reduced the number of players from 18,040 to 16,504, giving me plenty of data to work with. I wanted to build a choropleth map of birth locations among US born players. In order to do this, the birth locations had to be converted from city-state format to FIPS format. I accomplished by utilizing an intermediate step of converting the city-state format to GPS coordinates using Google's Geocoding API. From there I was able to take the GPS coordinates and convert them to FIPS codes by using an API for census block information courtesy of the Federal Communications


US Birth Locations of MLB Players 1871-2014
Data Courtesy of Lahman Baseball Database
value
600
400
200

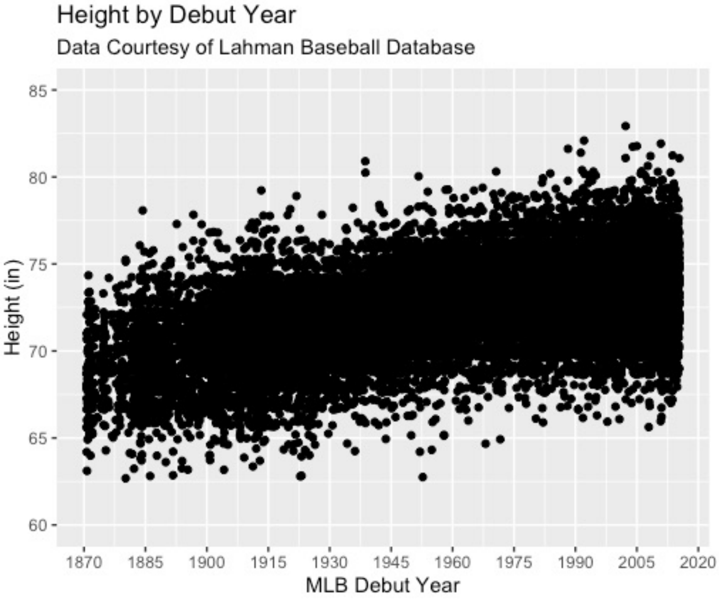
Commission. These two steps took a substantial amount of time to research and implement but also to run on the data, even with removing duplicates to improve efficiency. Once I had the locations as FIPS codes, I was able to build a choropleth map of the birth locations.

One thing that is a little surprising about this map to me is just the dominance of two counties. These would be Los Angeles County and Cook County (home to Chicago, IL). I'm not too surprised by LA being home to the most players, but I also thought some other southern areas would have a better showing such as Miami, Atlanta, Dallas, and Houston. While Dallas and Houston at least show some representation here, they're still behind cities like Boston, Pittsburgh, Detroit, and Philadelphia. The flip side of this conversation is how sparse the representation is from western states. I understand these counties are less populous but there is a lot of grey on the map indicating no players from that location. Looking at the data a little more closely, I had a feeling that some concentration of Northeastern US birth locations could be due the early days of baseball. The data collection in this dataset began with the formation of the National Association of Professional Baseball Players in 1871, only 6 years after the Civil War. I wanted to see just how much this data was skewing the results. I decided to graph the "center of population" for players' birth locations binned by year. The way I did this was by averaging the GPS coordinates. Looking at the map on page 5, we can see that over time the center of population has shifted from Pittsburgh, PA to just south of Tulsa, OK. In the nearly 150 years since the data collection began. I decided that a 25-year block was sufficient to see the westward movement of the center of population. It is also worth noting that these dates were taken from when a player made his MLB debut, not his birth date. For example, the first year-range contains the birth locations of all players who made their debut prior to 1900. Another aspect of note is that in the last bin of years (2000-Present), the westward movement stopped. The southward movement continued, however. I would imagine this is due to more involvement from the South and Southeastern US.



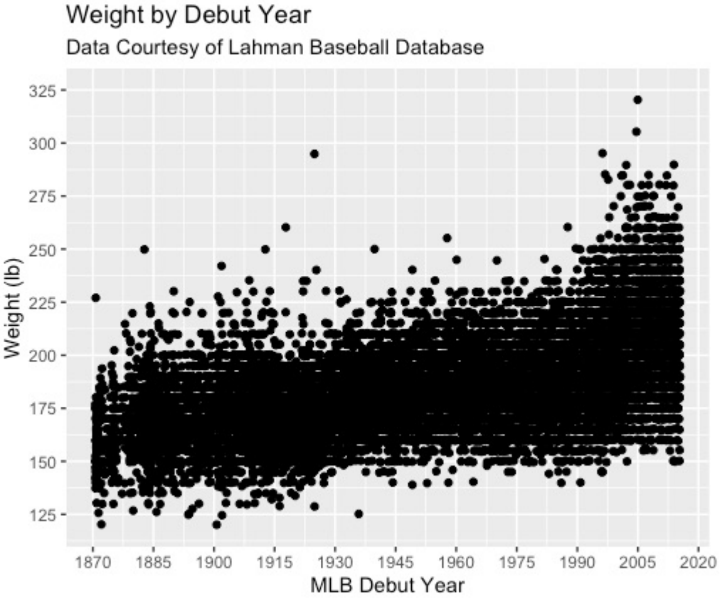
Analysis of Height and Weight

Another aspect of this dataset worth analysis is how certain variables have changed over time. In this project, I’m going to dig more into how height and weight vary over time. The first step for me was to build scatterplots of both these variables against the debut year of a given player. Looking at these plots, there definitely a trend. Both height and weight seem to increase as time goes on. Before I started made these plots, I thought height might trend upwards but as



Height by Debut Year

Data Courtesy of Lahman Baseball Database



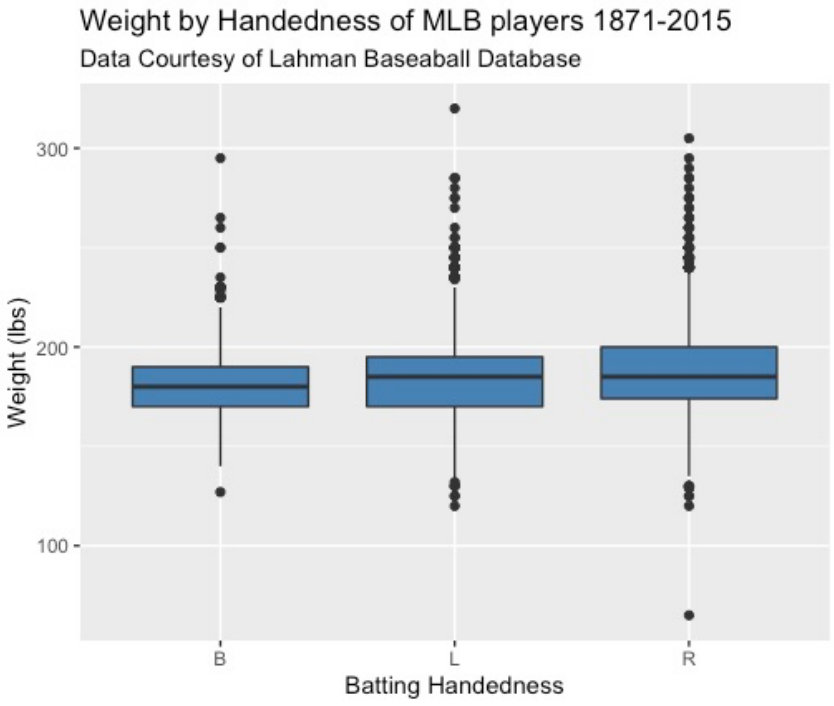
Weight by Debut Year

Data Courtesy of Lahman Baseball Database

players would be getting in better shape, I wasn't sure if that would translate to higher weights or not. One interesting aspect about baseball is that it is one of the few sports where you don't have to have the physical intangibles to be good. In fact, the average baseball player included in this data set is 6'0'' tall and weighs 186 lbs. These intangibles do appear to become more important in the modern game of baseball though. In order to determine if this correlation is significant, I computed a Pearson's correlation coefficient for height and weight vs. debut year. I discovered that there is significant correlation for height and debut year, r = .518 (p<.0001) and also significant correlation for weight and debut year, r = .566 (p<.0001). While we can say that there is a significant correlation between year and both height and weight, if we split the data roughly in two, is there a significant difference in the mean height and weight for the two groups? For this I chose a debut year of 1950 as the cut date because it very nearly perfectly bisected the data in terms of total players and time. On these two groups, I conducted a Student's T-test for means to determine if there was a significant difference in the means of the two groups. As it turns out, the t-test showed there is a significant difference in the means for height (p<.0001) and weight (p<.001) between players debuting before and after 1950.

Analysis of Handedness

If you follow baseball, you're probably familiar with the idea of wanting a left-handed power hitter in your lineup. While this dataset doesn't contain batting statistics to determine if left-handers have more success hitting homeruns, I think there is still some analysis that can be done here. It may not be a perfect stand in, but I think there can be a case made for using weight to account for one aspect of power hitting. There is definitely more to having success hitting homeruns than weight (for example left-handers having the advantage of hitting off more right-handed pitchers), but as we've seen, weighing more generally means being taller and perhaps stronger as well. With the groundwork established, I wanted to see if more, heavier left-handed batters made it into MLB by being able to hit for power. As you can see in the boxplot below, there actually doesn't appear to be any trend here. Even if you look at the outliers for the groups, there appears to be more outliers for right-handed batters than for left-handed batters or players who bat both left and right-handed. I must say I was a little surprised by this. In addition to what was mentioned above, players who play second base, third base, and shortstop are



usually smaller, and almost always right-handed. I wanted to make sure I was thorough in my analysis though and I decided to run a Pearson's Correlation Test. To do this I had to create a numeric binary variable. I did this by considering batting right-handed to be a '1' and batting left-handed or both to be a '0'. The results of the correlation test actually produce a significant, albeit small, correlation r = .048 (p<.0001). Knowing this, I then wanted to see if it was possible to model weight by handedness. To do this, I also incorporated debut year into the model since we know that weight varies by debut year. The resulting linear regression analysis yielded an equation of:

Estimated Weight = .309*DebutYear + 2.832*BatHand – 420.587

Maybe the most interesting thing here is that the model shows if you bat right-handed that contributes to a higher weight, since the coefficient is positive. As for the diagnostics of the model, they're not great. Both variables are significant predictors, but our model has an Adjusted R^2 of just .3414. This means that 34.14% of the variability in weight is explained by debut year and batting handedness.

Value of Study

While there were some aspects of this study that didn't yield the results I was hoping for, there is still a great deal of positive here. This study showed the movement of MLB player birth locations over time and that there are a few concentrated areas that churn out players at a high rate. While the past is important, I think this study has shown that perhaps further analysis can be limited to the "recent past" rather than incorporating data from as far back as 1871 when baseball and the United States as a whole was a very different place.

Data Source:

http://www.data-manual.com/data

additionally can be found at

http://www.seanlahman.com/baseball-archive/statistics