Assignment #1 – Information Seeking:

1. Lahman Baseball Statistics Database 2016

This dataset is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License, and more information regarding that type of license terms of use can be found <http://creativecommons.org/licenses/by-sa/3.0/>.  
 This dataset is interesting in that it contains American year and team level baseball statistics for players, teams, and team managers from the years 1871 through 2016, for what is now commonly referred to as Major League Baseball (MLB). These data contain pitching, batting, and fielding statistics for historic American baseball players such as Babe Ruth, Hank Aaron, Pete Rose, Jackie Robinson, Ty Cobb, and many more. This data set also contains information regarding every baseball team franchise, including the various team name iterations and the cities in which the franchises have been located. The Baseball Hall of Fame yearly voting records and inductees are also kept in this collection. The dataset is available in three formats: Microsoft Access, comma-delimited file, or SQL file.   
 Journalists, baseball archivists or historians, and general baseball hobbyists or enthusiasts would be the most likely users or decision-makers for this data collection. Journalists, hobbyists, or general enthusiasts may take advantage of this data in order to compare specific statistics of current players to past historical players. Those looking to create some type of predictive analytics system may also use this historical data to help create their models. Historians or archivists may also use this data in order to record and preserve records of what is generally considered America’s past-time (baseball).   
 Three questions that this data might help answer could be: How does a current player compare to specific historically well-known players with respect to several batting statistical categories at the same point in time of their careers?; How many years did it take a particular player, team manager, or general baseball figure to be inducted into the Baseball Hall of Fame?; Which are the ten oldest baseball franchises in American baseball history, and in what order from oldest to youngest?

Citation:

Lahman, Sean. (2017). Lahman Baseball Database 2016. Retrieved September 11, 2017, from

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

1. 1999 to 2015 NFL (National Football League) Combine Data

This dataset is “compiled from publicly-available NFL play-by-play data on the internet” by Daren Willman and is free to download for website users (<http://www.nflsavant.com/about.php>). Darren states on his website a commonly asked question from website users, “Can I have you data?”, and the responding answer in “Yes!”. The exact extent of the terms of use are not elsewise explicitly defined, but Darren provides his Twitter handle and email on his website in order to answer any “questions, concerns, or issues”. He has been sent an email in order to better understand his dataset’s re-usability terms.   
 This dataset is interesting because it contains information regarding individual football players that were invited and participated in the annual NFL combine from the years 1999 to 2015. The exact individual information regarding these players contains details regarding combine year, simple biographic information (name, position, college), body measurements (height, weight, arm length, hand size), performances in the standard combine drills (forty yard dash, vertical leap, broad jump), as well as what round and overall pick they were chosen in the NFL Draft. This data would be interesting in order to draw correlations between specific data elements mentioned above and what overall draft pick the player is selected in (if at all). This could potentially help predict when current NFL Combine players could expect to be drafted based on draft results of players with similar previous combine results.  
 Sports journalists, predictive statistical data analysts, and general NFL football hobbyists or enthusiasts would be the most likely users or decision-makers for this data collection. Journalists, hobbyists, or general enthusiasts may take advantage of this data in order to compare specific statistics of current players to past historical players. Those looking to create some type of predictive draft analytics system may also use this historical data to help create their models. Personnel employed by NFL teams trying to determine which players that they should draft may use this dataset in order to compare their different available choices.   
 Three questions that this data might help answer could be: How does a current player compare to specific historically well-known players with respect to several combine performance tests, such as the forty yard dash or vertical leap?; Which player has ran the fastest forty yard dash time at the NFL Draft Combine between the years of 1999 and 2015?; Which college had the most players drafted in the first three rounds of the NFL Draft in 2014 and 2015?

Citation:

Willman, Darren. (2017). 1999 to 2015 Combine Data. Retrieved September 11, 2017, from <http://www.nflsavant.com/dump/combine.csv?year=2015>

1. US Department of Education College Scorecard Data

According to a notice (<https://www2.ed.gov/notices/copyright/index.html>) by the U.S. Department of Education, “Unless specifically stated otherwise, all information on the U.S. Department of Education's (ED's) website at www.ed.gov is in the public domain, and may be reproduced, published or otherwise used without ED's permission”. This notice details that the data downloaded from the College Scorecard is publically available and is free to use or reuse.  
 This dataset is interesting because it contains a variety of variables that describe higher education institutions in the United States, which could be used to help prospective students decide which school that they may want to attend. This data may include data items such as the average net price of attendance for students whose families earned between $30,001 - $75,000, in-state tuition and fees, average faculty salary, proportion of faculty that are full-time, completion rate for first-time full-time students, and many more. The list of available variables can be viewed at <https://collegescorecard.ed.gov/assets/CollegeScorecardDataDictionary.xlsx>. All of this data is available about Title IV aid receiving higher education institutions, so tax payers should be satisfied that there is some level of transparency involved with the performance of the institutions that receive their tax money.  
 Prospective higher education students and parents, higher education administrators, as well as public policy makers would be the most likely users or decision-makers for this data collection. Prospective students and parents would benefit from using this data in order to compare a group of schools that the prospective student is considering matriculating to. Administrators at higher education institutions would be likely to use this data in order to see how their schools compares to their peer institutions and also to shape their education strategies in order to improve specific institutional outcomes. Public policy makers would also most likely use this data in order to help shape local, state, or federal level policy that affects the citizens in their respective jurisdiction and their respective budgets.  
 Three questions that this data might help answer could be: Which institution should a certain prospective student enroll at, if certain data variables such as faculty-student ratio, are to be highly considered?; Which areas are a specific higher education institution performing poorly in, when compared to peer institutions?; Which Maryland public institutions are most-deserving of a state-wide academic grant, based on specific degree-attaining outcomes?

Citation:

US Department of Education. (2017). College Scorecard Data. Retrieved September 11, 2017, from

<https://ed-public-download.app.cloud.gov/downloads/CollegeScorecard_Raw_Data.zip>

*Word Count: 1181*