Unit 02 Relational Databases: Project

The goal of this project is to use a real-world relational database with a comprehensive dataset comprised of a significant number of tables and to combine our understanding of how to acquire data through SQL with our knowledge of Pandas and Tableau to investigate and explore the data to answer some interesting questions.

Each team of two students can focus their efforts on one of two provided relational databases:

1. `imdb`: Movie/film database
2. `lahman2016`: Baseball statistics database

Movie Database

Data Tables Reference

As many of you know, IMDb is an online resource that maintains information on movies, tv, and celebrities for cinema produced from pretty much the beginning of movies up to the present (`Black Panther`, for instance, is present in the database).

The link above gives the so-called **code dictionary** that documents the tables and meaning of the columns within the tables.

Baseball Database

Data Tables Reference

Sean Lahman, an award winning database journalist and author, has compiled the most comprehensive baseball database to date. The baseball archive has been active since 1995, making it the oldest running baseball website on the web. Every year Lahman makes updates to the Lahman Database, a free relational database version of the archive that covers the game of baseball back to 1871. Giving free access to detailed statistics starting just two years after the first professional baseball team was founded in Cincinnati in 1869, 30 years after the reported invention of the game. This database has over 25 relational tables which give novice and professional analysts a rich dataset.

Project Information

As in the first project, I want you to synthesize the elements of what we have learned in this unit to ask interesting questions and use the data to create visualizations that help answer those questions. In this case, we want to use the techniques of asking for data from multiple tables in a relational database. We want to do so programmatically, and to define and use functions that provide good abstractions of the query operations we wish to perform.

I am not prescriptive: you get to decide which database you are interested in using, and you get to be creative and postulate relationships between dependent and independent variables provided in the data and ask interesting questions that explore those relationships. I am not asking for machine learning or advanced statistical methods to be employed … there are plenty of interesting relationships that can be found directly by combining tables, grouping data in various ways, and building graphical visualizations of the results.

Required elements

- Need at least **four** SQL queries of moderate to high complexity that are issued to the database to get the data to answer your questions
- The "heavy lifting" must be done by the SQL, not by `pandas` subsequent manipulation, nor by work (joins, grouping) in `Tableau`
- We **must** use good functional abstraction, which is best if we can include parameters that allow us to use the same function in more general ways through manipulation of the function arguments.
- The end product is an essay, targeted at a non-Python-SQL expert, describing the questions, the queries and strategies employed in solving the query problem with your SQL, and the visualizations and their interpretation.

Visualization

Some basic requirements about your visualizations:
1. They should clearly answer questions about the data
2. They **must** have clear axis, labels, titles, and units, so that someone who did not generate the visualization can interpret it.
- Some may need values within the plotting area itself.

Evaluation and Grading

Your task is to write the Python and SQL code to acquire the data for each of your questions and analysis, writing (good, clean, coherent) functions to make your queries and bring the data into `Pandas` dataframes. You should demonstrate the usability of the data by showing some of the data and exploration in pandas. The data should then be written to a csv for import into Tableau and the construction of your visualization.

For this synthesis assignment, you may not, in fact, need tons of Python code. So your Project assignments are not just about “writing the program”. I want you, through the Markdown and Notebook, as well as the visualization, to **clearly communicate** all of your steps in the start-to-finish progression and what you **learned about the datasets**. As outlined in the rubric below, your grade is about all of these parts.

I want to see you put together a coherent essay in an iPython Notebook that describes the steps taken, presents the code and rationale for why your functions coherently decompose and solve the overall problem, and uses graphs and English description to describe and **interpret** the results. So the end of the notebook will be including the graphs/visualizations developed in Tableau and leading the reader through the interpretation.

A good notebook should allow an audience/reader who is *not associated with this course* to understand the development and your conclusions.

Your grade will be determined as follows:

- 25 points for a well written and coherent essay and visualizations that answer good questions about the data
- 20 points for well constructed, documented, explained SQL of moderate to high complexity for querying the database and doing the real work in getting the data to answer your questions
- 10 points for the Python code, functional decomposition, code documentation
- 5 points for self assessment, partner assessment, and instructor assessment of an equitable parnership

As in all assignments, documentation for the **code** requires
- Complete docstrings for each function that include description, parameters, and return value
- Comments in the code that describe (for a reader that might *only* see your code) *how* the code is working.

Submission

Your submission will consist of a single iPython Notebook containing the code and documentation as described above. This will by a `.ipynb` file. You will also have multiple picture files (`.jpg`, `.png` etc.) showing the results of your visualization in Tableau. Note that the picture files should be incorporated *in your notebook* as part of your documentation of the questions you asked about the dataset and how the visualizations help you in interpreting the data and answering those questions.

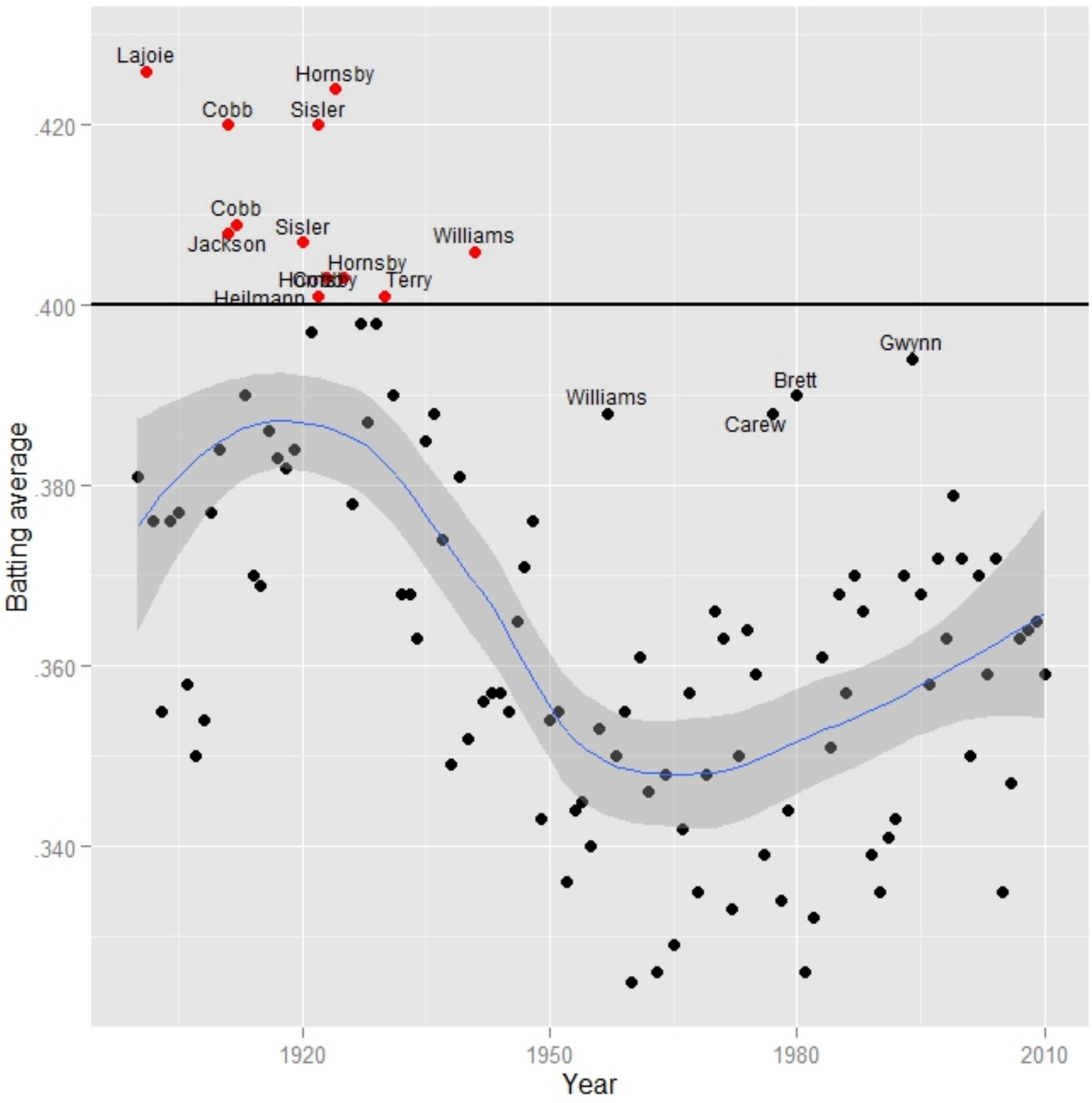
Thus, you will upload into the course `Notebowl` the files:
1. A `.ipynb` Notebook,
2. The set of `.jpg` or `.png` files for the figures generated in Tableau and included in your notebook.

Example Baseball Analysis

Since Michael Lewis's book, **Moneyball**, published in 2003, that looked at the use of Data Analysis by certain teams to maximize their benefit relative to salary expense to assemble a better team for the money, interest in using similar techniques has flourished. Through the Lahman database we have access to much of that same data. In the following section, we will consider some of the following example graphs exploring baseball data that I found by searching the web.

Batting Average by Year

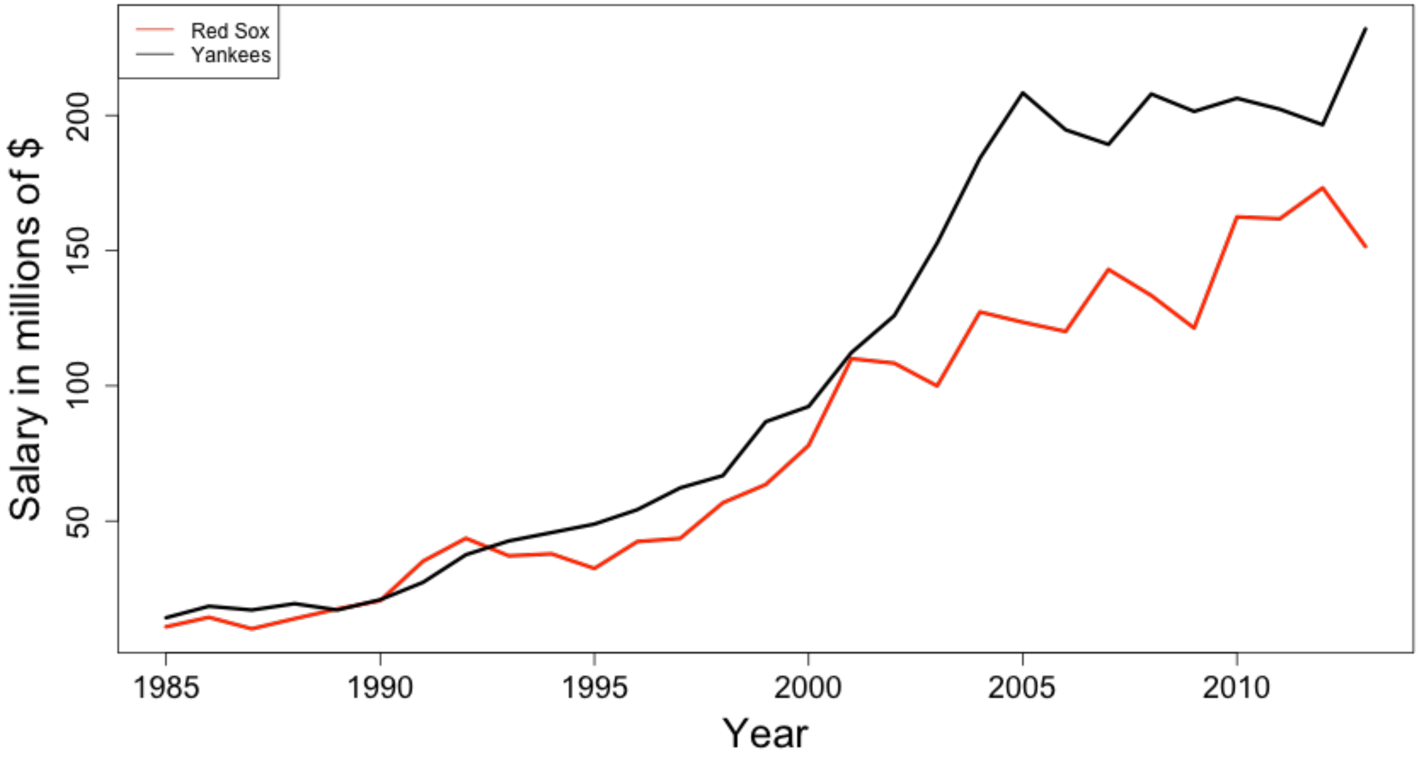
The first example shows a scatter plot of Batting Average against year, to see the change over time, and to then plot some well known exceptional hitters.



The grouping of individual batting averages for a given year might be filtered by players with at least a minimum number of times at bat.

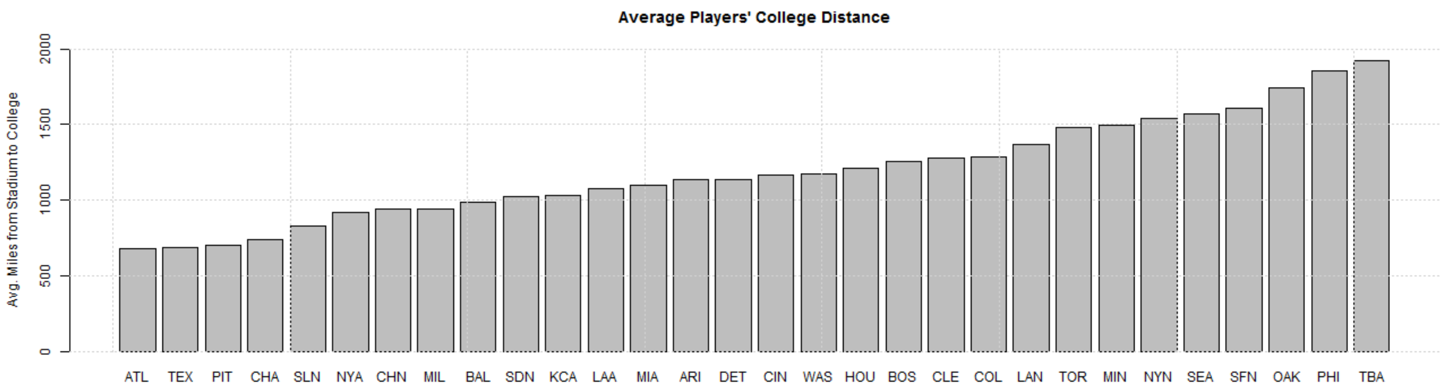
The Red Sox / Yankees Salary Story

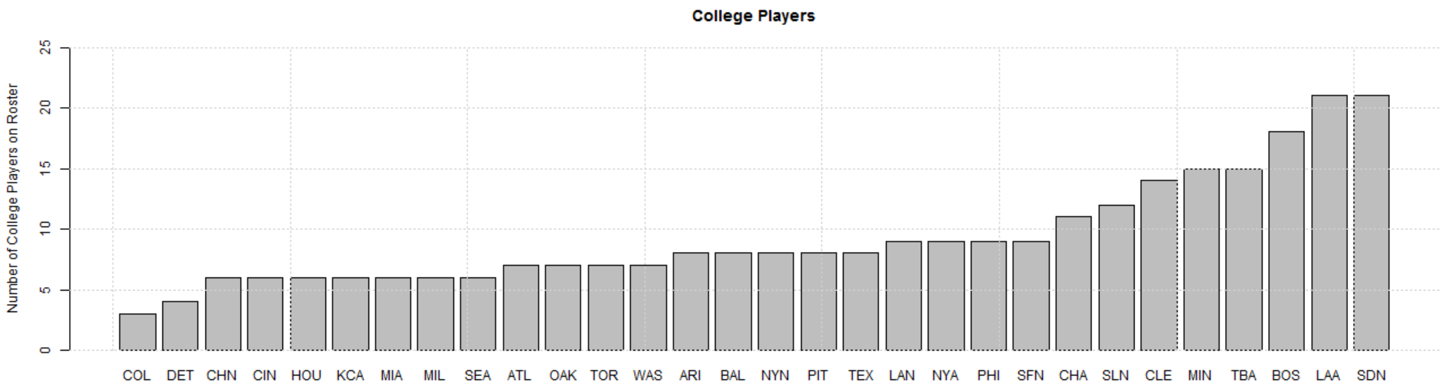
The next example was a simple visualization of cumulative salary for two teams over the years, focusing on the Boston Red Sox and the New York Yankees.



Players Proximity

The next analysis was exploring how “regional” different teams are, as measured by the distance between player’s college locale and the team’s stadium. Here the lower graph tells us how many college players are on the various teams' roster, and the upper graph calculates an average distance (relative to those players that have a college listed).

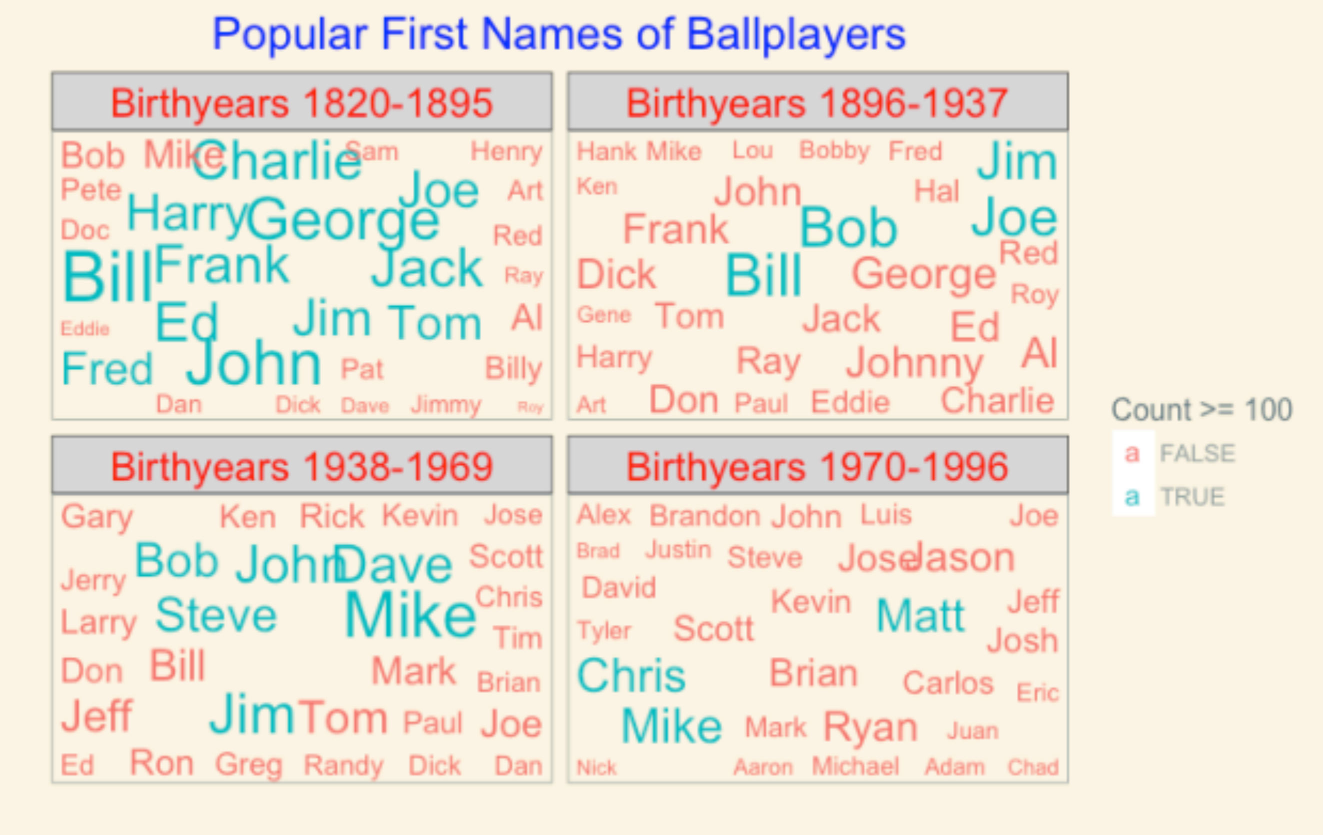




Note that this analysis needs a way to ask for a distance between two points on a map, and requires an additional API to accomplish this.

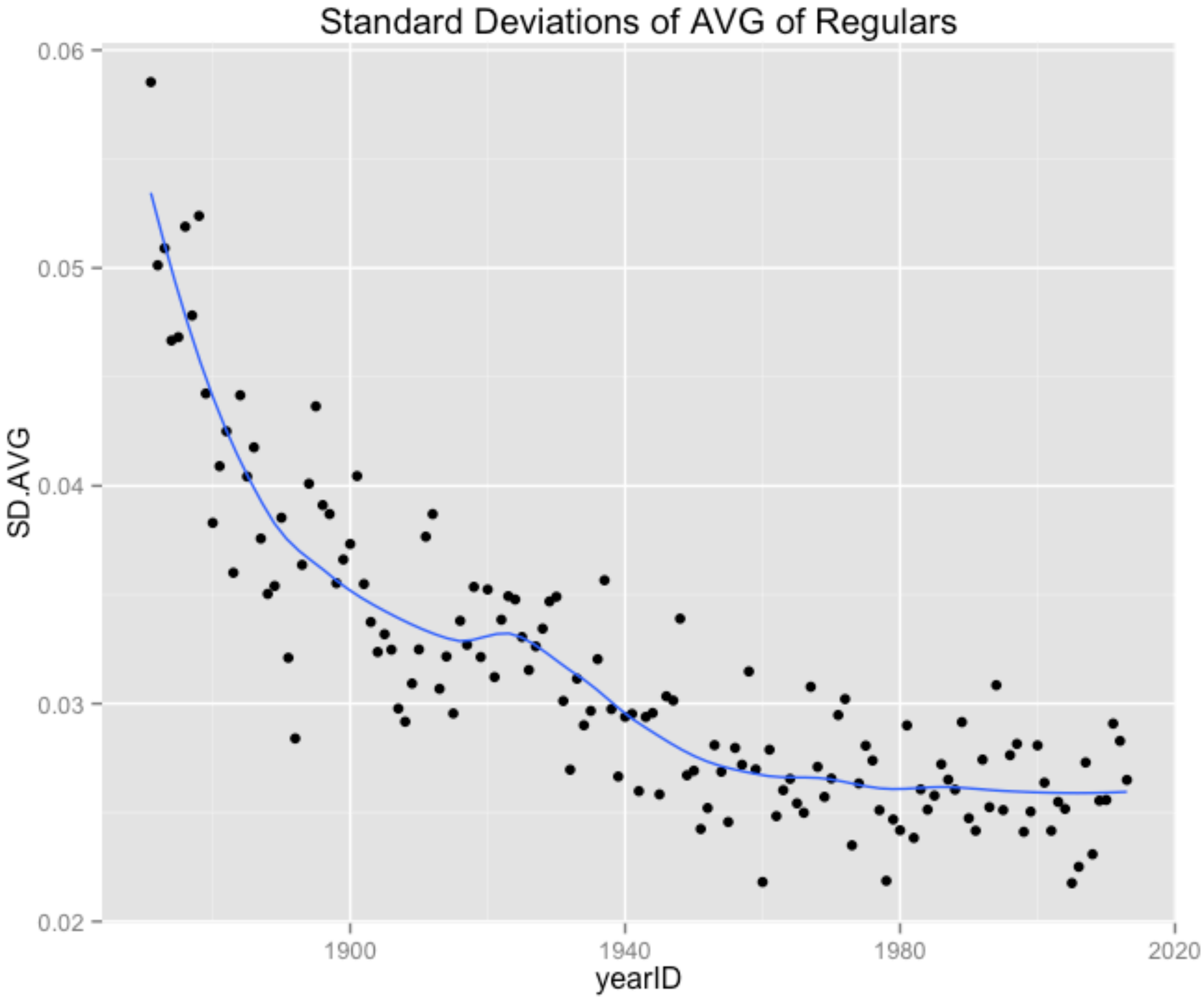
Popular First Names

Given the popularity of word clouds (which, btw, Tableau is capable of), this next example looked at popular first names of ball players, but segmented time into different eras and presented the word cloud for each of four different time spans.



Decreasing Variation in Players' Batting Averages

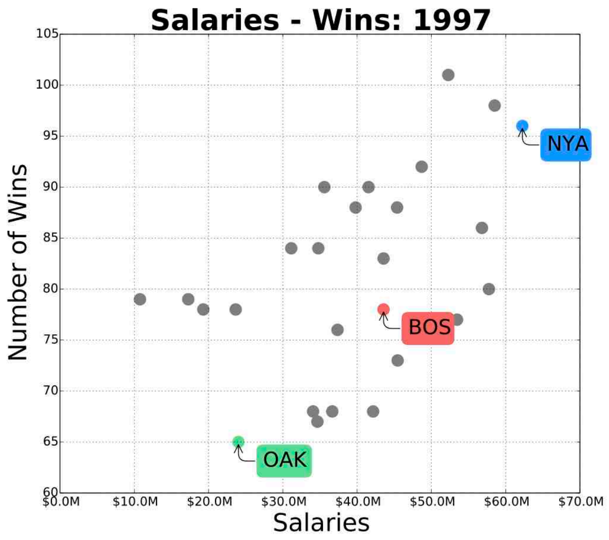
The next analysis looks at the statistic of standard deviation in players' batting averages and plots this measure of variation as a function of time. This decreasing trend is then used to argue the reason for the rarity of players who hit “over 400”, (i.e greater than 40%) in modern times.

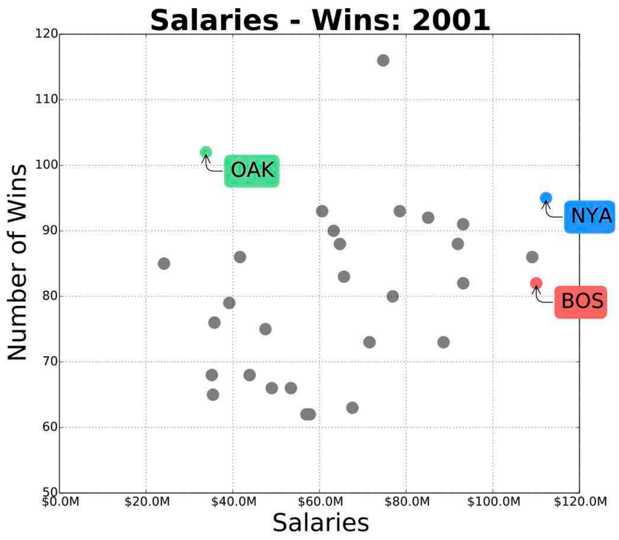


For inclusion in this analysis, there was again filtering against a certain number of opportunities at bat.

Salaries vs Wins

The last example shows the effect of OAK in their moneyball efforts between 1997 and 2001, by building a scatter plot of wins against salary for the two years, and highlighting a couple of the teams.





Baseball Analysis Tutorials/Examples

The following itemized list provides references for some of the above examples:

- Introduction to Using R Packages for Basebass Research
- GitHub CS-109 Course Slides
- Intro to Sabermetrics with Baseball Intro