How-To and Source Advice

1 How To Sign Up For Slack:

We'll be using Slack as our primary communication tool. There is one for each baseball team and the channels will be further subdivided into individual projects. [REDACTED] will be a part of their channel; there's no confirmation yet that [REDACTED] will be in theirs.

Click on one of the following links for whichever team you're on:

- REDACTED
- REDACTED

If you decide to switch teams or projects, please let me know so that I can deactivate your account and/or be added to a different project channel.

2 How to Access Data:

For [REDACTED], it is very likely that we will not get any data for our projects from them. This will make us rely on publicly available data. This isn't a problem however! There is an incredible amount of publicly available baseball data, and what is public is the primary information sources the teams use anyways (though of course cleaned up and modified many different ways). Here's how to access them:

- **Recommendation #1 (Hardest)**: Find a copy of the book *Analyzing Baseball Data with R*. Chapter 1 of this book as well as the Appendix will instruct you on how to access data from RETROSHEET, Lahman, and MLB Advanced Media. This will allow you to become familiar with the basics of R and will let you access whatever data you want at any time.
- **Recommendation #2 (Simpler)**: Follow the guidelines in section 3 to connect to the *Saberbase* MySQL Server. You won't have to figure out the various mechanisms from the Baseball textbook to gather data, but you will have to know some basic MySQL to interact with it.

- **Not Recommended, but available if you're stuck**: If you don't feel you have enough time to go through *Analyzing Baseball Data with R* or to learn some MySQL, I have a few sample files of data available at my Big Problems Github page. Feel free to download them and request more data from me if you have an idea you'd like to pursue.

3 How to Access the MySQL Server:

There is approximately 15GB worth of information currently sitting in the MySQL server from RETROSHEET, Lahman, and the Gameday sources. There are also two empty databases that are team-specific. To access it, go through the following steps:

1. Let me know you intend to use the MySQL server so that I can create a username and password for you.
2. Download and install both MySQL and MySQL Workbench. If you're a Windows user, MySQL or MySQL Workbench may prompt you to download Visual Studio for it's C++ packages. Download the community version of Visual Studio here.
3. Boot up MySQL Workbench.
4. Hit the plus button where it says "MySQL Connections" to create a new connection.
5. Create a "Connection Name" for yourself as your own name.
6. For "Hostname," copy and paste the following:
saberbase.cn2snhhvsjfa.us-east-2.rds.amazonaws.com
7. For username, type out your given username. For password, press "Store in Keychain" to type out your password. Press OK afterwards.
8. Click "Test Connection" to see if we have success. If yes, hit "OK" to open up the SQL editor.

The first problem we'll face actually isn't modeling or what algorithms to use, but cleaning up the database. There is way more data than we need, not all of it is useful, and not all of what we'll need conveniently sits in the same table. There is also a 20GB limit on the amount of data we can have sit in the database in any given point in time. So, we will need to make choices on what can be removed and what can be added in.

I don't have a good book to recommend to you for MySQL, but I don't think it's necessary to. Much like LaTeX, you can learn by googling things as you encounter new problems.

4 Major-Specific Useful Sources

4.1 For Everyone:

So that everyone has a "shared language" so to say of data science, I'd like everyone to start by finding a copy of the following two books:

- An Introduction to Statistical Learning, by Gareth James + others. This book also has a series of lecture videos that are organized here, and an online course going over the material exists here (free if you choose to audit) if you'd like a little more structure to your learning.
- Analyzing Baseball Data with R, 2nd Edition, by Max Marchi + others.

An Introduction to Statistical Learning will both provide a great introduction to the field of Data Science and a good introduction to R programming. It is a streamlined version of the graduate-level text *The Elements of Statistical Learning* and is meant to be accessible to a much wider variety of backgrounds.

Analyzing Baseball Data with R is a direct application book; it takes knowledge/techniques from both Data Science and Mathematical Modeling and provides an introduction to how they are used in Sabermetrics. This will provide a great introduction to data visualization, data mining, etc. in the *tidyverse* package of R and how Sabermetricians analyze MLB data. I'm pointing to the 2nd edition specifically because the newer edition uses the *tidyverse* package of R, while the 1st edition does not.

A good personal target to shoot for by the end of this semester is skim-reading the entirety of *Analyzing Baseball Data with R*, even if not fully understanding the material, and up to Chapter 7 of ISLR.

The following recommended sources after these two are additional options for people wanting extra material to go through, want a different but still useful skillset to provide to the project, etc.

4.1.1 Grad Students

Do whatever you want.

4.1.2 Undergrad Stats Majors

If you have experience working with R, you're pretty much all set with respect to technical skills. I'd recommend going further into some Sabermetrics knowledge by reading one of the books listed off this page, with "The Book" being highly recommended.

4.1.3 Undergrad Math Majors

Unfortunately you likely have no prior experience with stats or with computer programming. On the bright side though, your background as a math major means that you can teach yourself enough material by the end of the semester to make valuable contributions. Here are some alternative sources that you might find more helpful:

- CMU 10-701 Source Material - This is a first semester graduate course in Machine Learning at CMU, complete with lecture videos. It quickly goes through a long series of introductions through different developing areas of the field. For Math Majors I think it would be worthwhile to go through all material up until Semi-Supervised Learning.
- Machine Learning with Andrew Ng - This provides an introduction to the field of machine learning (sans any deep dives) and an introduction to the MATLAB/Octave programming language. Since it doesn't have much probability theory, I don't recommend it nearly as much as the CMU material. However, this would be a much better fit for someone with limited time, and/or has little prior programming experience.
- Stanford's CS229, a more in-depth version of Andrew Ng's course.
- Machine Learning: An Applied Mathematics Introduction, by Paul Wilmott. A great introduction to Machine Learning purely through it's mathematics. If your time is limited and don't think you can learn a programming language during the semester, this can still provide you with enough context to contribute in a group conversation or in a presentation.

4.1.4 Undergrad CS Majors or Computer Engineers

Take a shot at going through CS1675 material as well as the prerequisites listed on the page. If you suspect you won't have time to go through the material throughly, some quick and dirty versions of the same material can be found in:

- Hands-on Machine Learning with Scikit-Learn and Tensorflow, by Aurelion Geron. Gives a nice overview of the models in machine learning and the tools available without going too in-depth.
- Deep Learning with Python, by Francois Chollet, if you're already familiar with most material and decide Neural Nets is how you'd like to contribute.

4.1.5 Other Majors

Please provide me with some feedback at the end of this semester! Aside from a handful of Business majors, I don't have much experience in working with majors outside of STEM. That of creates a communication gap not unlike what is found in many modern businesses.

I have a few books recommendations that I think are still very valuable for learning Data Science / Sabermetrics without getting into nitty-gritty details. Here are a few of them:

- Mathletics, by Wayne Winston - A great book of ab introduction of Sports Statistics, with applications to Excel. Does a very good job of showing the "modeling" mindset that analysts use, and answers some very important questions in relating runs to wins to finances to each other through it's modeling showcases. It doesn't come with it's own exercises though, so make sure to play with a few excel spreadsheets or SQL commands to make sure you're following along.
- Machine Learning For Absolute Beginners, by Oliver Theobald - A great short read which gives a broad overlook of the field of machine learning, and finishes off with an exercise in the Python programming language.

4.2 Other Useful Books

This is a catch-all suggestion space for special cases.

- Advanced R, by Hadley Wickham - despite the name, its level of difficulty is somewhere between CS401 and CS445. It's a great introduction to the data structures of R and to the Functional Programming paradigm.
- R For Data Science, by Garrett Grolemund and again Hadley Wickham. A good reference book for the tools at one's disposal in the *tidyverse* package of R, and some introduction to how a data scientist thinks about problem solving.
- Web Scraping with Python, by Ryan Mitchell. Incredible book on tools available for Web Scraping in Python.
- Pattern Recognition and Machine Learning by Christopher Bishop and The Elements of Statistical Learning by Rob Tibshirani and others - these are graduate textbooks on the subject of Data Science. Although they're great textbooks and are usually the first ones people point to, they're down here in "other" because they can be incomprehensible to the uninitiated. You really need to know upper-level Stats, Calc 3, Linear Algebra, and some CS to make sense of the books. If you have all 4 of those things or lots of free time to work on catching up on 1 one those things, knock yourself out.
- Machine Learning, by Tom Mitchell - this book is very similar to the ISLR book in terms of valuable content and intentions of being an accessible version of what's usually a graduate-level subject. I chose ISLR though because even though it's aim is to be accessible, it is still mostly reserved for advanced Math and CS undergrads. Those with that kind of background could substitute ISLR with this book if they want to get a little more in-depth.

5 Online Sources for Baseball News and Data

1. The Athletic (Subscriber Only)
2. FanGraphs
3. The Hardball Times (Subsection of Fangraphs)
4. Baseball Savant
5. MLB
6. Baseball Prospectus
7. Baseball America (Subscriber Only)
8. Tango Tiger
9. A Syllabus of A Sabermetrics Class, at Williams College.