Playing the Game:

**QUICK VIEW**

|  |
| --- |
| Company: |
| Role Played: Analyst and Content Manager |
| Learning Objectives: |
| Technical Skills   * + Probability and Statistics   + R   + Data visualization |
| Soft Skills   * Research skills * Presentation skills * Data contextualization skills   Brief Case Summary:The following xCase will require you to analyze and present sports data in a way that will attract sports leaders to your website. The long term goal is to establish analytics and consulting partnerships between these sports leaders and Stattleship.  Deliverables:   * Commented .R files with code for retrieving, manipulating, analyzing and visualizing data * Strategy presentation that demonstrates the following:   + Methodology (metrics development, etc.)   + Story of your results/process   + Lessons learned and Recommendations |
|  |
|  |

Become a Trusted Voice in Sports Using Novel Data Analysis

A LevelEdu xCase

Developed in conjunction with Stattleship

1. Background

1.1 What is an xCase?These engaging lab exercises will require you to apply the skills you are gaining as a part of your Level coursework to a research or business problem that is rooted in a real organization. After exploring and analyzing datasets provided by these companies, you will derive insights from the data and create a work product that will be presented to the rest of the class.

1.2 Your Company: [Stattleship](https://www.stattleship.com/) helps connect brands with sports fans through social media. The company helps brands engage in social media conversations at the speed of the sports. This means providing company social media teams with compelling sports content they can use before, during and after games to engage consumers. The company, comprised of data scientists and software engineers, is based in Cambridge, MA.

1.3 Your Problem (Generally): Stattleship is hoping to establish itself as a destination for quality sports data and analysis. They are able to reach and interact with a diverse group of fans and followers through their API, which allows users to access their rich database of sports data. The organization would like to extend their reach and relevancy in the field, with the ultimate goal of working directly with teams and owners. To do this, they must establish themselves as reliable thought-leaders who offer commentary and insight on topics that are meaningful to players, coaches, managers and owners. Data analysis expertise is of paramount importance for these stakeholders. Thus, Stattleship has an opportunity to appeal to potential clients if they can prove that they do not merely know how to provide data, but can also make sense of it and translate it into strategic decisions for team members and owners.

You are a member of the data analytics team with a specific focus on baseball analytics. As summer approaches and teams begin to make their waiver decisions1, you would like to explores and compares the stats of the best teams and individual players in the league. Rather than using standard metrics to compare teams and players, such as RBI and batting average, you will develop novel sabermetrics2. Examples of other sabermetrics include WAR3 and the techniques used in *Moneyball*4*.* Your team sees an opportunity to differentiate themselves by developing novel metrics to compare teams using data gathered with the R Wrapper for the Stattleship API.

***You believe that each player contributes equally to a baseball team, so you have decided to develop normalized metrics for all players in the league, specific to their position.***

At the very least, your boss at Stattleship determines that your final overview should include how your favorite baseball team compares to other teams in your league.

2. Approach

First, you will gather, clean and analyze the data. You will develop metrics for each position in the league, and then calculate these metrics for each player.

Remember to document all code in one or more R scripts so that it can be revisited throughout the exercise, and reviewed at the end. **Assume for this article that we are interested in all of the data from the 2016 season**.

Suggested breakdown of time:

2.1 Week 1: Research and gathering data (continued on page 4)

* Read the case from beginning to end to understand the data that may be of interest.
* Refer to the next page and choose a path for gathering the Stattleship data:

**Data retrieval options**

If you find yourself spending more than 30-60 minutes on a more challenging approach without much progress, try the next step down.

**Most challenging – I want to figure this out on my own**

* Install the Stattleship package in R. (Request one month access to the API from Teri (@teri or teri@leveledu.com), then go to their website and read the instructions)
* Explore the documentation to see what data is available and to figure out how to gather the appropriate data by testing different inputs to the functions in the package.
* Save the raw Stattleship data so you don’t have to download it again.
* Choose a baseball team upon which to focus your data gathering and analyses.
* Begin gathering the data that you need in one or more data frames. Make sure your data frames are well labeled and clear. Adjust the data types in each column as needed.

**Challenging – I want to figure this out with some hints**

* Install the Stattleship package in R. (Request one month access to the API from Teri (@teri or teri@leveledu.com))
  + HINT: [Install Instructions](https://www.dropbox.com/s/iphp72qdebtk662/Installing%20the%20R%20Wrapper%20for%20the%20Stattleship%20API.pdf?dl=1)
* Explore the documentation to see what data is available and to figure out how to gather the appropriate data by testing different inputs to the functions in the package.
  + HINT: [ss\_get\_result](https://www.dropbox.com/s/flpuv26osf8rjty/Using%20the%20R%20Wrapper%20for%20the%20Stattleship%20API.pdf?dl=1)
* Save the raw Stattleship data so you don’t have to download it again.
* Choose a baseball team upon which to focus your data gathering and analyses.
* Begin gathering the data that you need in one or more data frames. Make sure your data frames are well labeled and clear. Adjust the data types in each column as needed.

**Moderate to challenging – Just give me the data, I’ll figure out what to do with it**

* Choose a baseball team upon which to focus your data gathering and analyses.
* Download the 2016 game logs:
  + gls2016: [gls2016](https://www.dropbox.com/s/8xqovro9p5qz2g3/gls2016.RData?dl=1)
* Download the 2016 player logs:
  + pls2016: [pls2016](https://www.dropbox.com/s/1bl97khzr8ys0p7/pls2016.RData?dl=1)
* Save them in your working directory
* Load the two datasets into R, e.g. **load("gls2016.RData")**
* Explore the data, what type of data object is it? Can you convert it to a data frame? Make sure your data frames are well labeled and clear. Adjust the data types in each column as needed.

**Moderate – Give me the data, and help me understand its structure**

* Choose a baseball team upon which to focus your data gathering and analyses.
* Download the 2016 game logs:
  + gls2016: [gls2016](https://www.dropbox.com/s/8xqovro9p5qz2g3/gls2016.RData?dl=1)
* Download the 2016 player logs:
  + pls2016: [pls2016](https://www.dropbox.com/s/1bl97khzr8ys0p7/pls2016.RData?dl=1)
* Save them in your working directory
* Load the ***two*** datasets into R, e.g. **load("gls2016.RData")**
* Let us explore the data together using the following script and starting with pls2016:
  + R Script to understand player data: [understand stattleship data](https://www.dropbox.com/s/4d5kg02t5xgd4y7/understand_stattleship_data.R?dl=1)
* Do a similar process to gls2016
* Make sure your data frames are well labeled and clear. Adjust the data types in each column as needed.
* Hint: merge your player and game log data frame by player id (check the column names and change as necessary)

**Before moving on: data retrieval options**

After you’ve completed one of the processes above, please review the R Script from the “Moderate” data retrieval option to ensure you understand the data structure:

[understand stattleship data](https://www.dropbox.com/s/4d5kg02t5xgd4y7/understand_stattleship_data.R?dl=1)

Suggested breakdown of time:

2.1 Before Week 1 ends: Research and gathering data (continued)

* Research the different positions in baseball to determine what skills are important for each type of player or position and which standard statistics are used to assess these skills. Don’t know anything about baseball?
  + HINT document: [Baseball Overview](https://www.dropbox.com/s/0fulenpxtduz34e/Baseball%20Overview.pdf?dl=1)
  + HINT 30 min video: [Level Student Explains Baseball](https://www.dropbox.com/s/i18k92c223z31z4/Baseball%2BStats%2BOverview.mp4?dl=1)

2.2 Before Weeks 2-3 ends: Metric development, calculation and basic plotting

* Develop a metric for ***each baseball position*** based on your research. Each metric should contain at least two statistics, combined in some way. You can stick with the popular. ***Before deciding on a metric, think about the data:***
  + Gather any additional data needed to calculate statistics for every player in the league. Calculate the statistics you are interested in for each position:
    - Think about the following:
      * Not every player played in every game, is there data for every player in every game? How can we eliminate data for days when a player was benched or injured?
      * Should a given statistic be averaged or summed? What about a new statistic that looks at how a player performed over time?
      * How does using cumulative statistics vs. average statistics affect the meaning of the metric?
      * Should the players that play multiple positions, but fewer games (i.e. not starters) be treated differently? Make a decision here and document your reasoning.
      * What about pitchers? Should relief pitchers be treated as a different position than starters?
  + Are these statistics random variables? Are they relatively normally distributed across the league? Confirm this. If the data is not normal (look at the histogram, skewness and kurtosis, look at the normal probability plot), see if simple transformations can improve normality. Settle on a final version of the statistic, and calculate descriptive statistics for all variables you plan to use in your metrics.
  + How are the statistics from the data best combined into position metrics? E.g.:
    - If you choose salary in dollars and RBI as your statistics, and you add them, what influence is RBI having on the total metric?
    - What if you add a low variance statistic and a high variance statistic?
  + If you think it is necessary (if the statistics have fairly different expected values or variances), normalize your variables. One approach to this is:
    - Subtract the mean from all data for that statistic (i.e. for each player in the league), such that the mean is equal to zero for that statistic across the league. Divide this difference by the standard deviation associated with the distribution for each statistic. Check that the mean and standard deviation of the resulting data are equal to zero and one, respectively, for each statistic. (This is creating *Z*-scores!) (N(0,1))
  + Combine your statistics into a metric for each position
    - Weight your variables intentionally to ensure that your metric is measuring what you want it to measure
* Document these metrics and your thought process in a clear and concise way. Make sure you are able to explain your reasoning beyond each metric.
* Calculate the relevant metrics for each player in each position across the entire league. Look at the distributions of each metric for each position across the league. Calculate a few descriptive statistics that are interesting for each distribution (at least: mean and standard deviation, at most: skewness and kurtosis).
* For each position metric, normalize as:
  + For a given position metric, subtract the mean for that metric such that the new mean for that metric is zero. Divide this difference by the standard deviation associated with the position metric distribution. Check that the mean and standard deviation of the resulting data are equal to zero and one, respectively, for each position metric. (N(0,1))
* Now that each player has an associated metric, rearrange the dataset by teams. Plot the distributions of the standardized metrics for your team, as well as the other teams in your division.

2.3 Before Week 4 ends: Statistics, advanced plotting and begin to create your presentation

* Determine the “best team” in your **division** using these distributions. What makes it the best team, define how you measure “best” using these distributions. Is the difference between that team and your team statistically significant? What about that team and the worst team in the league?
* Rank all of the teams using your metric, compare your team rankings to the current MLB rankings, how are they similar or different? Why do you think that is?
* If you have time, see Section 3.0 below for more analysis ideas
* Create informative visualizations based on your data, as well as tables and text when needed, to tell a story with your data. This story should include your methodology, findings, and a pitch for or against your metrics/approach. Include lessons learned, if any. Put this information into a PowerPoint deck for a 10-minute presentation that pitches your approach to your coworkers at Stattleship. This presentation could translate into a presentation for players, coaches, managers and owners.
* You will present your PowerPoint deck to the class. Be prepared to ask and answer questions about your process and results.

2.4 Before the presentation: Finalize your presentation

3. Optional questions to answer

3.0 Additional challenge questions

**If you have extra time in lab, or future work, please include any of the following outcomes in your presentation. You may pick and choose from this list, you do not need to address these problems in order:**

* **Challenge**:What happens if you do the same analysis for 2015 data? Can you predict the outcomes from the 2015 season?
* **Challenge:** Assume all players are equally important to your team. What is the relationship between their normalized metric and their salaries? Is this similar to the rest of the league? Propose a way to use this information to inform trades and improve your team. Is salary normally distributed across the league? What about for each position? Does this analysis change if you normalize the salary data by position to have a mean of zero and standard deviation of one?
* **Challenge:** Pick your favorite position and look at the relationship between salaries and the metric that you’ve developed for that position, which player do you think is the “best deal?” Which player do you think is the “worst deal?” Why?
* **Challenge:** Look at a few players on your team over time, either using your metric for their position or another set of statistics. Do you see any players that are trending up?
* **Big Challenge:** Build your dream team on a budget—who can create the “best team?” In this case, assume the best team has the following properties: high mean of metric, low variance of metric, and low total salaries. Normalize the salary data by position to have a mean of zero and a standard deviation of one. Your team should include the following 22 players (based on ESPN fantasy baseball):
  + C, 1B, 2B, 3B, SS, OF, OF, OF, OF, OF, 2B/SS, 1B/3B, UTIL (OF can be LF, RF, CF or OF)
  + P, P, P, P, P, P, P, P, P (at least 4 SP, the rest can be RP)
  + Note: UTIL can be filled by DH or anything other than P
* Your dream team will be evaluated by the following metric:

(metric mean) – (metric standard deviation) – (average cost)

* **Challenge:** Pose your own question and answer it with the data.

4. Your Toolkit

4.1 The dataset:

* Your dataset will come from the Stattleship R Wrapper.

4.2 References:

1 <http://www.baseball-reference.com/bullpen/Waivers>

2 <https://en.wikipedia.org/wiki/Sabermetrics>

3 <https://en.wikipedia.org/wiki/Wins_Above_Replacement>

4 <https://en.wikipedia.org/wiki/Moneyball>

5 <http://games.espn.go.com/flb/resources/help/content?name=roster-settings-standard>