NBA Statistics and Winning
Fall 2019
CGT 270 – Final Project Notebook

Final Poster

Basketball Bros
By: Josh Knull, Jitesh Motati, and Richard Wu
CGT 270 Fall 2019

Introduction

The topic we decide to explore relates to how various statistics used to measure performance in the NBA correlates with team success. The intended audience includes basketball enthusiasts and sports analysts as they can utilize the findings to draw conclusions in a more accurate manner. Prior work has been done on a topic similar to ours by Yuanhao Yang who studied the correlation between individual player's performance and their team's success level. The variables used in Yang's study revolves around normal statistics such as points, rebounds, and assists. Together as a group, we believe that these statistics do not accurately represent the value of individual players and their impact on the game. Therefore, we think our work is necessary as it provides more insight on team success by analyzing in-depth performance-related variables such as efficiency ratings and field goal percentages. The result of this project is significant to its intended audience because it provides insight on effective team compositions, assists analysts to provide more accurate analysis of teams, and helps organizations to trade players and formulate their teams in an efficient way.

Background

The variables that are used in this study includes efficiency ratings, shooting percentages, and r-squared values. The efficiency ratings are divided into three categories: offensive, defensive, and net efficiency. The formula for each is listed below.

Offensive Efficiency Rating = 100*Points Scored/Possessions(self)
Defensive Efficiency Rating = 100*Points Allowed/Possessions(opponent)
Net Efficiency Rating =
Offensive Efficiency Rating - Defensive Efficiency Rating

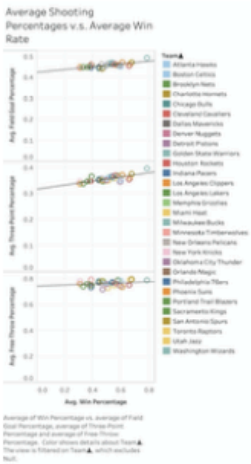
*Possession is defined as Field goal attempts - Offensive rebounds + Turnovers + .475*Free throw attempts

Shooting percentages are divided into three-point percentages, field goal percentages, and free-throw percentages. These variables measure the accuracy of each type of shot which can be calculated by
Number of shots made/Number of shots attempted.

R-squared values are statistical measures that describe the strength of a correlation between two variables. R-squared values ranges from 0 to 1 with 0 indicating no correlation and 1 indicating very strong correlation.

Figure 1

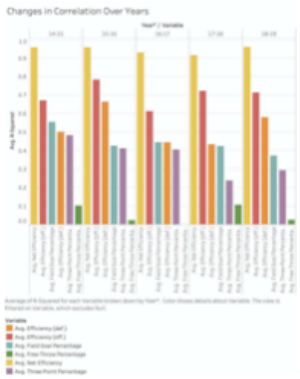
Figure 2



1. How do different stats correlate with team success? Which variable has the strongest correlation? Which has the weakest?

Figures 1 and 2 collectively addresses the question by analyzing two different sets of variables: efficient ratings and shooting percentages. Both visualizations are created using the average of all the values for each variable across five consecutive NBA regular seasons between the years of 2014 and 2019. The x-axis of each graph measures the average win ratio while the y-axis displays the average values for a specific variable. The data for each team is plotted in a different color and a line of best fit is created for each scatter plot using the functionalities provided by Tableau. These lines can be used to analyze the strength and direction of the correlation for a specific variable. A negative slope indicates that the variable negatively correlates with a team's success while a positive slope indicates the opposite. The steepness of the slope represents the strength of the correlation. A slope close to 1 indicates a strong correlation and a slope close to 0 indicates a weak correlation. As the two figures suggests, average net efficiency has the strongest correlation with team success due to its nearly-perfect diagonal slope. In contrast, the slope for average free-throw percentage reveals that it correlates the weakest with the successfulness of a team.

Figure 3



2. How do the correlations change over time?

Figure 3 addresses the question by comparing the r-squared values for each variable during different regular seasons. These r-squared values are calculated by dividing each scatter plot in Figures 1 and 2 into five separate ones with each representing a particular NBA season and analyzing the line of best fit descriptions generated by Tableau. The x-axis of Figure 3 displays the variables for each season while the y-axis provides a scale for the average r-squared values. The bars are color coded based on the variable they represent. The height of the bar suggests the strength of the correlation between a specific variable and the average win percentage. As the visualization suggests, average net efficiency has the highest r-squared value across all five season and average free-throw percentage has the lowest. In other words, average net efficiency has a consistent strong correlation with a team's success and average free-throw percentage has a consistent weak correlation. It is also noticed that the r-squared values for average defensive efficiency ratings fluctuates the most out of the six variables investigated in the study.

Conclusion

In our study, we explored the correlations between NBA teams' average win ratio over the five most recent regular seasons and six different variables that measure performance level. Based on our findings from the visualizations we created, we concluded that net efficient ratings correlates the strongest with a team's success in the league while free-throw percentage correlates the weakest. This finding is not surprising to us as we expected free-throw percentages to have the least impact on a team chances to win a game. We hope that the results of our study will highlight the significance of advanced statistical measures used in sports analysis such as efficiency ratings in comparison to regular stats (points per game, rebounds, etc.), which often fails to accurately evaluate a team's performance level. For future extensions to this study, we hope to incorporate more variables and increase the number of regular season used to generate a comprehensive version of the analysis we did in the current project. Furthermore, this will allow us to analyze any existing trends regarding the changes in the correlation strengths of each variable and apply it to the overall changes occuring in the league.

References

Data source: NBA & ABA League Index. (n.d.). Retrieved November 8, 2019, from https://www.basketball-reference.com/leagues/.

Related work: Yang, Y. S. (2015). *Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics* (Doctoral dissertation, PhD thesis, UC Berkeley).

Additional research:
Hayes, A. (2019, May 8). R-Squared. Retrieved November 28, 2019, from https://www.investopedia.com/terms/r/r-squared.asp.

Jones, K. (2018, November 4). The five factors behind the NCAA's NET ranking system. Retrieved November 8, 2019, from https://www.si.com/college/2018/11/04/college-basketball-rankings-net-system-explain.

Final Paper

NBA Team Statistics and Their Effect on Team Success

CGT 270—Final Project Paper

By: Basketball Bros (Josh Knull, Richard Wu, Jitesh Motati)

Abstract

Our project looks at various statistics in the NBA and how strongly (or weakly) they correlate with team success. The three of us are big fans of the National Basketball Association (NBA), and wanted to do something with the many statistics associated with the game. But rather than looking into the individual player stats, we decided to try something, that we thought, was more unique: team statistics.

Besides from the obvious measure of a team's performance—points scored, we wanted to consider the large variety of factors that affect how a team performs: rebounds, assists, efficiency, shot percentage, etc.

Related Work

Upon initial research into the topic, we found a study done by Yuanhao Yang, from UC Berkeley. In his study, he looked at how various individual statistics and performances affected a team's success level. He focused mainly on "basic" basketball stats, including points, rebounds, and assists. He used this information to make models to predict how many wins a team would get in a particular season, given the individual performances of its players.

But for our project, we wanted to answer the following questions:

(1) How strong do various team stats correlate with team success?

(2) How consistent have these correlations been over the last 5 years?

Introduction

There are many age-old sayings and facts when it comes to basketball: "Defense wins championships", "make sure to grab your rebounds", "always make your free throws", "find the open guy for the better shot" (just to name a few). As a team, wanted to investigate these to see how true they are, based on the numbers, in the highest level of basketball. Everyone talks about how good the *players* are, how many points (or blocks, or rebounds) a *player* had one game or one season, but the media does not seem to cover how a *team* does. The only thing people seem to care about when it comes to the team? Who won, and who lost.

The goal of our project was to get our audience thinking about the bigger picture—the *team*. Given that we were looking at simple and advanced stats, we expected our audience to be, at least, a common basketball fan. Of course, more diehard fans will probably be more interested in our work, but we did not consider much the interests of a non-sports fan. Nevertheless, we have provided a description of most of the statistics we used in our research and analysis.

We expect to have, after having done our analysis, a clearer view as to what makes an NBA team successful in the current state of the game, and an idea as to how true the "sayings" are mentioned above.

Methods

When starting the project, there were four things we needed to figure out: (1) How are we going to define "team success"? (2) What years of the NBA are we going to look at? (3) What specific team statistics are we going to analyze? (4) Where are we going to find this data?

Firstly, we decided to decide team success by one thing: win percentage during the regular season. We figured that since the name of the game is winning/losing, that should be our metric for measuring "success". We are aware that winning in the regular season is not the only goal in the NBA (winning the championship is the ultimate goal), but since the vast majority of NBA games are regular season games, we figured this would give us the best picture. Also, we did not consider other factors that can be associated with "team success" in basketball, such as attendance, financial growth, TV ratings, or merchandise sales. Although these measures are important for an NBA franchise, they are not important for our "on-the-court" view of a team's performance.

Next, we had to narrow down the scope of the NBA to a specific range of years. The NBA started back in 1949, but the game was completely different than it is now. In fact, we would argue that in the last five years (starting with the dominance of the Golden State Warriors) the league has evolved significantly, in terms of team strategy, focus on three-point shooting, high ball-screens, etc. For this very reason, we decided to look at the *last five years*, starting with the 2014-15 season, for our analysis.

We then had to decide which statistics we would look at. Since we were aiming to look at correlations between various statistics and winning percentage, we found it most logical to look at statistics that had percentages involved. This includes field goal percentage, as one can see in our visuals below. We also looked at efficiency, since that statistic has been known to be very "important" to teams. Below, we define the statistics we have in our visualizations:

Field Goal: An attempt to shoot the ball into the basket. When behind the 3-point arc, worth 3 points if successful. Otherwise, worth 2.

Free Throw: After a foul, a player may be given the chance to shoot an uncontested shot from 15 feet away from the basket. Worth 1 point

Possession: (field goal attempts) – (offensive rebounds) + (turnovers) + .475 * (free throw attempts)

Offensive Efficiency: 100 * (points scored) / (offensive possession)

Defensive Efficiency: 100 * (points allowed) / (defensive possession)

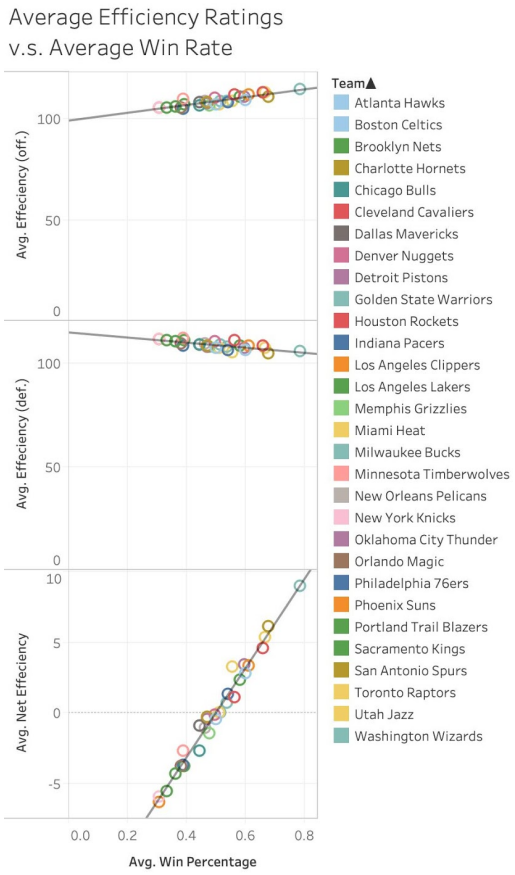
Net Efficiency: (offensive efficiency) – (defensive efficiency)

Shooting Percentage: (shots made) / (shot attempts). Typically divided into three categories: free throws, 2-point field goals, and 3-point field goals

Lastly, we had to figure out where we would get data. We considered both the NBA official website and basketball-reference, a sport-statistic site. We went with the latter, as the data was already downloadable and was more detailed than that of the NBA site.

Results

After analyzing efficiency ratings over the five-year period, we made this visualization:



Average of Win Percentage vs. average of Effeciency (off.), average of Effeciency (def.) and average of Net Effeciency. Color shows details about Team▲. The view is filtered on Team▲, which excludes Null.

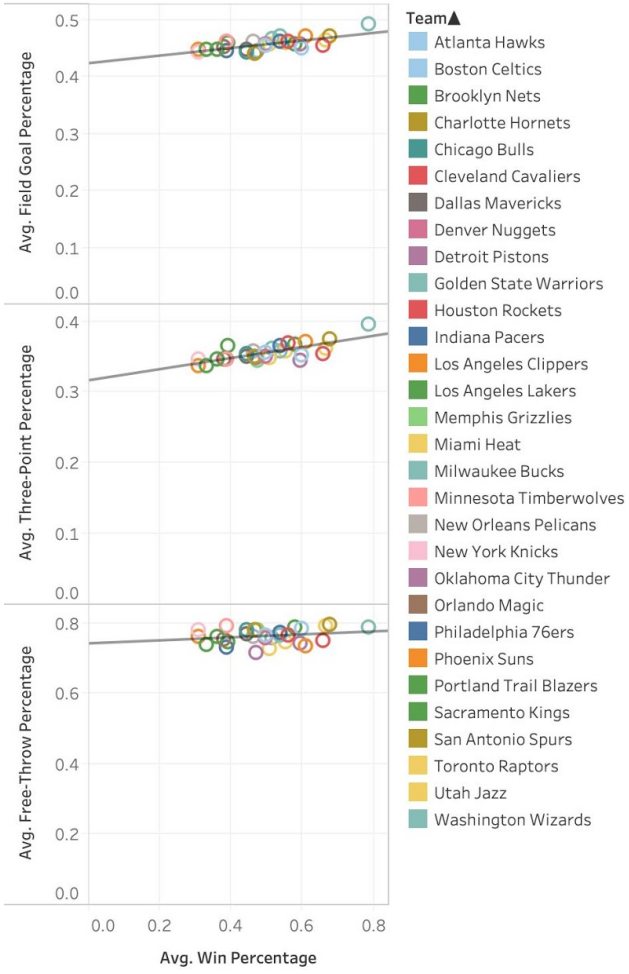
The x-axis is each team's average win percentage (our marker for "success") over those 5 years. The y-axes represent the three types of efficiencies: offensive, defensive, and net, as explained earlier. The lines represent the correlations between these efficiencies and win percentage, and the individual marks (circles) show each team's averages. The higher the slope of the regression line, the stronger the correlation.

The first thing to notice here is how strong the correlation is between net efficiency and win percentage. In fact, this is the strongest correlation among all the statistics we analyzed, as we will show later on. But this does make at least some sense: if you score more points than you give up per 100 possessions, then you should win, right? Asking this out-loud makes one feel that this should be true. And the numbers agree.

The second takeaway from this visual is the middle line: there is a *negative* correlation between defensive efficiency and win percentage. In short, this suggests that if a team is *worse* defensively, they do *better* in terms of winning. This may seem counter-intuitive, but the NBA has been an offense-based league, especially recently. Teams that do better on offense do not need to do as well on defense—as long as they are scoring more than they give up. Thus, the age-old saying that "defense wins championships" may not be so true.

Next, we looked at the various shooting-percentages and how those correlated with win percentages.

Average Shooting Percentages v.s. Average Win Rate



Average of Win Percentage vs. average of Field Goal Percentage, average of Three-Point Percentage and average of Free-Throw Percentage. Color shows details about Team▲. The view is filtered on Team▲, which excludes Null.

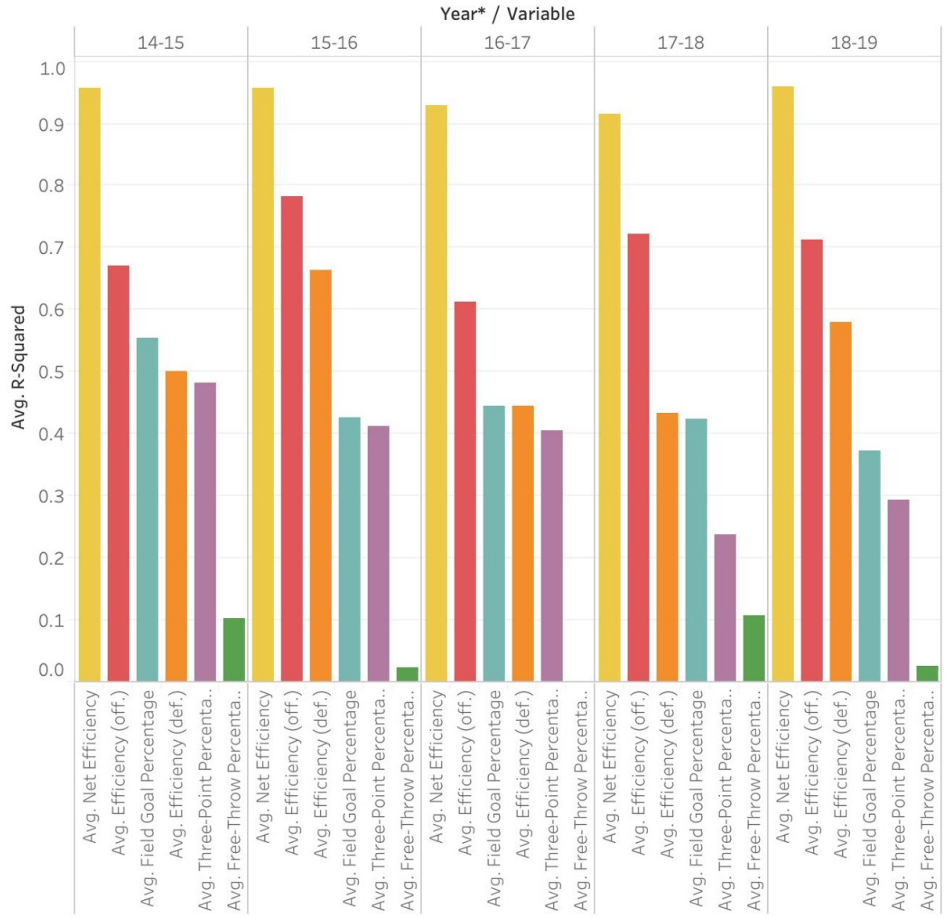
Similar to the first visual, this graph shows the correlations of free-throw, two-point, and three-point percentages and win percentage. All three are positive (as seen by the positive slope), which makes sense, since if you make more of your shots, you'll score more points, which should help you win games.

The interesting insights from these analyses come from the differences in the correlations. 3-point percentage is the strongest, while free-throws is the weakest. We believe that 3's are the strongest because (1) they are worth more points and (2) they carry more momentum than a two-pointer. Although every team is worse at shooting 3's than 2's. the better you are at 3's, the better-off you will be. This is especially true in today's game, in which teams tend to "space-out" to get shooters open behind the arc.

While the 3-point correlation is the strongest, that of the free throw is the weakest. In fact, there is almost no correlation at all. Every team shoots around 75%, give or take a few percent. They always say, "make your free throws", but in reality, this is not that important. As long as you shoot around what everyone else does, you'll be just fine. Free throws, as a whole, have very little to winning games, although making a big free throw late in the game can still "decide" the game at that point. But in reality, that last-second point is worth the same as every other point throughout the game.

Lastly, we looked at how these correlations changed over the 5-year period.

Changes in Correlation Over Years



Average of R-Squared for each Variable broken down by Year*. Color shows details about Variable. The view is filtered on Variable, which excludes Null.

Variable
- Avg. Efficiency (def.)
- Avg. Efficiency (off.)
- Avg. Field Goal Percentage
- Avg. Free-Throw Percentage
- Avg. Net Efficiency
- Avg. Three-Point Percentage

The bars represent the R-squared values for the various stats. R-squared is a measure of the strength of a correlation, with 1 being the strongest and 0 being the weakest.

As can be seen, net efficiency has by far the strongest correlation each year over the past five years, and free-throws have by far the weakest. Offensive efficiency is consistently the second strongest, and the rest have some variation.

Discussion

We believe that these results all highlight the changes in today's NBA vs. that of the past: it's all about offense. Even if a team is giving up a lot of points, as long as they score more, then they will win games. And that makes sense!

In today's game, the 3-pointer is becoming more and more valuable. Our second visual supports this, as it has the highest correlation with winning of the three shot types.

We expect that this trend will grow stronger and stronger in the near future.

Net efficiency is consistently strong, and we do not expect that to (ever) change. The name of the game is scoring more points than your opponent. Unless the NBA becomes like golf (lowest score wins), then this correlation will always be strong.

Future Work

There are two aspects which we think can be expanded on to gain further insight:

The first: take playoffs into consideration. How do stats in the playoffs differ from those from the regular season? Since teams are known to play "harder" defense, we would be interested in how shooting percentages and defensive efficiencies are affected.

The second: consider the different schedules of different teams. Each division (5 teams) plays a different schedule. Does what it take to win in the western conference differ from what it takes in the east? What about throughout the 6 divisions? It would be interesting to analyze this, to see if there are different trends throughout different regions in the league.

References:

-Data source: NBA & ABA League Index. (n.d.). Retrieved November 8, 2019, from https://www.basketball-reference.com/leagues/

-Related work: Yang, Y. S. (2015). *Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics* (Doctoral dissertation, PhD thesis, UC Berkeley).

Additional research:
-Hayes, A. (2019, May 8). R-Squared. Retrieved November 28, 2019, from https://www.investopedia.com/terms/r/r-squared.asp.

-Jones, K. (2018, November 4). The five factors behind the NCAA's NET ranking system. Retrieved November 8, 2019, from https://www.si.com/college/2018/11/04/college-basketball-rankings-net-system-explain.

Visualization Process

Name: Josh Knull Course: CGT 27000-LC Term: Fall 2019 Worksheet ID: WS005

Semester Topic Proposal

Goal: Develop semester project proposal.
Objectives: Students will define a semester topic and clearly define the problem to be addressed, provide a brief description of what's been done before, the significance of the problem and their solution (what they plan to do). **Outcomes:** Students will submit a project proposal for approval describing their approach for addressing the defined problem.

Your proposal will be reviewed, and your semester topic evaluated along with your previously submitted Topics to Data Sources worksheets. If your proposal and worksheets support a *Good Research Topic/Question* with access to adequate data to work with, your semester topic will be approved, and you will be encouraged to proceed with collection, processing and visualization of data. If your proposal and worksheets *do not* support a Good Research Topic and/or if there is insufficient data available, your proposal will be denied, and you will be encouraged to refine your research question. You will have one week to refine your research question, update your proposal to reflect the revised topic and submit the revised proposal for approval. If your proposal is denied a second time, you will be given a topic by Dr. Byrd for your semester project.

Answer each question below in 250 words or less and complete sentences. Your answer must be single space, Times New Roman, 11-point font and, there should be a blank line between each question/answer block. DO NOT change the margins, DO NOT re-order the questions. Failure to follow instructions will result in loss of points.

1. 1) **Describe your research problem/question.**

   I plan to research how Major League Baseball statistics have changed over time. I know that batting average (the typical stat associated with offensive performance) has been going down, but total runs scored have not. I want to show that both of those statements are true, but mainly, I want to evaluate where the runs have been coming from. What statistics have improved to compensate for a decrease in batting average?

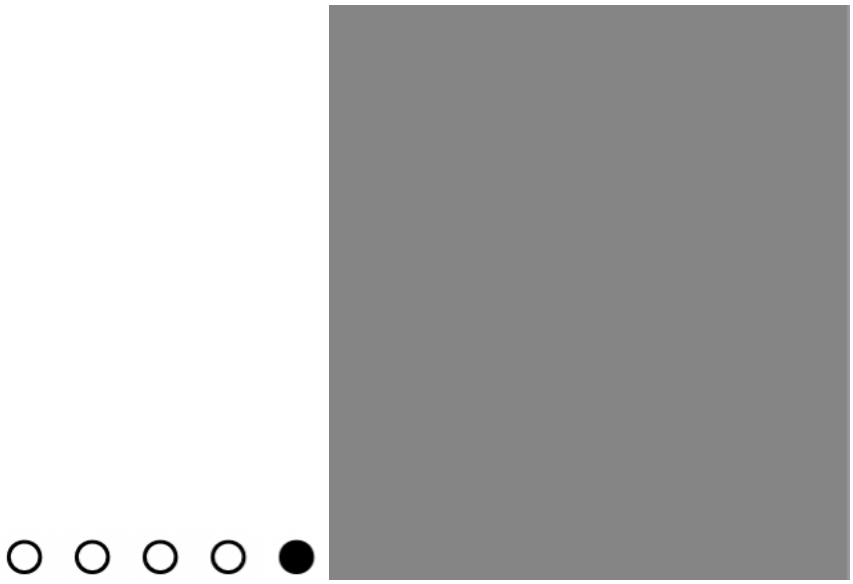
2. **2) Who is your audience?**

   Most baseball experts already know what has caused the runs scored to remain steady, or even increase. But I would guess that the average baseball fan (casual, hardcore, or in between) does not. So ideally, I would like my audience to be any fan of baseball. For someone unfamiliar with the game, the statistics I plan to evaluate will not make sense to them, making them a less "relevant" audience member for my purposes.

3. **3) Describe why the work is needed and/or is important?**

   Historically, the first thing people ask when they want to know how good a hitter is, is "What's his batting average?" That has been going down recently. The league-average batting average has been decreasing, which naturally leads one to think that offense has been going down. But this is not the case. I want to investigate where the "production" has been coming from. The obvious

answer is just runs scored, but which statistics (common or uncommon) have players been improving at?



4) Explain what remains to be done and what you plan to do differently.

I need to look at the statistics and how they have changed over the years, and I need to put these *changes* side by side. I want to see which stats have increased/decreased the most. This should give me a good idea concerning the trends of today's game vs in the past.

5. **5) What does the data look like? Have you downloaded the data? Is it publicly available?**

   I have found a "batting encyclopedia" on baseball reference that has all sorts of raw baseball statistics. I should be able to look up any player's stats during any year, league leaders any year, and hopefully league averages. I should not have to download it. It is public and free!

6. **6) What visualization tool(s) do you plan to use?**

   If I can figure out Tableau, I will try to implement graphs on there. If I have to, I can probably get by just using some Excel line graphs, but I hope to implement something a bit more complex.

7. **7) Explain the significance/importance of your work (why the problem/question is important).**

   Relative to the "real" issues of the world, a topic concerning baseball is not too important. But within the world of baseball, I would argue that where the run production comes from is quite important! Upon quick glance, one would see players with their batting averages and notice that they are considerably "lower" than they used to be. But the numbers on the scoreboard are increasing. How is this happening?

8. **8) Explain what the expected outcomes of your work will be.**

   I expect to have comparisons in the change over time of certain statistics. I plan to indicate answers as to what has made up for the lack in batting average. I hope to find a good way to visualize these changes. I hope my audience gains an appreciation for the game, or at the very least, learns something about the sport!

Name: Josh Knull Course: CGT 270 Term: Fall 2019 Worksheet ID: WS001

Topics

Goal: To generate topics for semester project.
Objectives: Students will identify a list of topics of interest.
Outcomes: Students will generate, compare and rank topics of interest and identify who the audience is.

Generate

1. **1) List three topics that interests you**
   1. *How statistics in Major League Baseball have changed over time.*
   2. *How Procter & Gamble, and other manufacturing companies, use and collect shelf data to optimize shelf*

      layout

2. **2) Which topic interests you the most?**

a. How statistics in Major League Baseball have changed over time.

3. **3) What do you know about each topic?**
   1. *Topic 1: Basic stats, insights, how teams look at stats. Basic reasons for why certain stats have changed*
   2. *Topic 2: Basic organization format, what planograms look like and their purpose*
4. **4) What do you not know about each topic?**
   1. *Topic 1: Any higher-level sabermetrics, why basic stats have changed and when the changes started*
   2. *Topic 2: How shelf data is collected and how planograms are made*

Evaluate

5) Compare and rank topics

a. In this space evaluate your topics, what you know what you do not know about them. Rank the topics according to your preference then answer the following questions about each: Is this the right challenge to tackle? Is there enough detail or is there too much?

1. There is certainly “unlimited” detail. If I could figure out which stats to look at, I could get a lot out of this.
2. There are a lot of benefits to looking at shelf data, which I know hardly anything about. I can definitely look into how data is collected / what it is used for / how much of a difference it makes.

Byrd Data Visualization Lab © 2019

Name: Josh Knull Course: CGT 270 Term: Fall 2019 Worksheet ID: WS002

From Topics to Questions

Goal: To articulate what the student hopes to find out about the Topic of interest (Box 2) in the Topics Worksheet. **Objectives:** Students will transition from a broad topic to a specific question.
Outcomes: Students will answer who/what/when/where/whether/why/how about their topic

You are no longer a mere data collector, but a researcher interested in understanding something better. List your topic of interest from the Topics Worksheet) below in a complete sentence.

Topic:

1) The topic I am studying is: Major League Baseball’s “Statcast”, along with other higher-level baseball sabermetrics.

Because:

2. **2) I want to find out what/why/how:**

   I want to find out **how** Statcast collects its data. Exit velocity? **How** does it measure something that happens so fast? How does it store the data of a player’s hit tendencies? **Who** is keeping track of it? **How** do MLB teams use this information to come up with defensive shifts? **How** effective have shifts actually been?

   I want to know to **what** degree Statcast data actually affects a team’s tactics/decisions. Does the ‘route efficiency’ stat help coaches/players, or is it just something for fans to gaze at?
   What types of sabermetrics are actually relevant? **What** does Major League Baseball plan to do with Statcast in the future? **What** kinds of stats does MLB record for its umpires, if any?

   Why are certain stats measured (ex: WAR)? **Why** are there two (or more?) different versions of WAR? **Why** aren’t these stats common knowledge to the average baseball fan? **What** would it have to take?

Why is batting average becoming less important? **Why** is OPS the "better" measurer of a hitter's success these days? And **why** wasn't it 20 years ago?

3. **3) In order to help my reader understand:**

   Although baseball might seem like a mundane, repetitive, simple game, there is so much more that goes into it. In addition to the hours of training by athletes, the miles traveled by scouts to find the next generation of talent, and the negotiations made by owners/general managers to improve their team, this is the roll of the statistician.

   There are tons of numbers associated with baseball, much more than in any other sport: Batting average, RBI, OPS, ERA, WHIP, FIP, BABIP, BARISP, WAR, VORP, just to name a few. These statistics are meticulously recorded and calculated, and when analyzed correctly, can help increase the probability of a team's chance of success.

I want my reader to come to a new appreciation as to what there is behind the game played with just a white ball and a stick. The next time they see a defensive shift, I want them to come up with an educated guess as to why the shortstop is playing on the other side of the base (no more "What? Did he forget where the shortstop plays?"). I want to introduce my reader to a stat they probably have never heard of, as well as clear up what a certain stat means.

All in all, I want my reader to understand the value of numbers in baseball. And if they're anything like me, these new insights will make them even more fascinated with a game they've always claimed to love.

Resource:

Booth, W. C., Colomb, G. G., Colomb, G. G., Williams, J. M., & Williams, J. M. (2003). *The craft of research*.

University of Chicago press. (Chapter 3)

Byrd Data Visualization Lab © 2019

Name: Josh Knull Course: CGT 290 Term: Fall 2019 Worksheet ID: WS003

From Questions to a Problem

Goal: To generate questions based on the Topic of interest (Box 2) in the Topics Worksheet. **Objectives:** Students will transition from a broad topic to a specific question.
Outcomes: Students will answer who/what/when/where/whether/why/how about their topic.

You are no longer a mere data collector, but a researcher interested in understanding something better. List your topic of interest (from box 2) on the Topics Worksheet) below (in a complete sentence):

1. **1) Identify the significance of your topic by completing the following sentences**
    1. *Topic: The topic I am studying is baseball statistics: how they have evolved and changed over time and how players have done over time.*
    2. *Question: because I want to find out what/why/how players have been evaluated over time, which statistics have lost/gained relevance, and how players have improved/gotten worse over the years.*
    3. *Significance: in order to help my reader understand that baseball is not the same as it used to be and is always changing. There is no simple way to define how "good" a player is.*
2. **2) Use the space below to determine if your problem is practical or a conceptual problem by answering the following questions about your topic.**
    1. *What's the situation or condition you desire to address, and why is it important?*

       I want to address why batting average (and other statistics) are not as important in today's game as newer stats, such as OPS. On a game level, it is important to know why these new stats lead more directly to the most important stat in baseball: winning.

    2. *What's the undesirable consequences caused by that condition including costs that you (or your readers) don't want to pay? Why is it important?*

       I don't want people to think that the only way to evaluate players in by the numbers. While they are important, they don't measure effort level, team chemistry, or other important factors to a player and team's success.

    3. *What do we do about it?*

       I'll make sure to reinforce that fact, that the numbers don't tell everything!

3. **3) Practical vs Conceptual Problems: complete one of the two questions about your problem**
    1. *Practical Problem: What do you want your reader to do?*
    2. *Conceptual Problem: What should the reader think? I want them to understand how/why these stats are*

       important and why we use them to measure a player's success.

4. **4) Significance**

a. Explain why you want your reader to know and care about the topic. There is a lot that goes into baseball! Although I am sure much of my audience finds the game boring and slow, there are tons of opportunities to gather information and analyze the players and teams. The game is not so simple after all

Resource:

Booth, W. C., Colomb, G. G., Colomb, G. G., Williams, J. M., & Williams, J. M. (2003). *The craft of research*.

University of Chicago press. (Chapter 4)

Byrd Data Visualization Lab © 2019

Name: Josh Knull

Course: CGT 270 Term: Fall 2019 Worksheet ID: WS004

From a Problem to Data Sources

Goal: Acquire data for semester topic.
Objectives: Students will identify appropriate data sources transition from a broad topic to a specific question. **Outcomes:** Students will plan, generate, and determine the availability of data for the semester project.

1. **1) Describe your semester topic below**
    1. I am working on the topic of: *MLB baseball statistics and how they have changed over time*
    2. Because I want to find out: *how certain stats have improved, while others have gone down. How do these*

       relate to the change in runs scored/runs allowed each year?

    3. In order to help my reader, understand better: *how the game has evolved, how more runs have been*

       scored despite less hits, etc. And for the non-baseball fans: basic knowledge of basic stats!

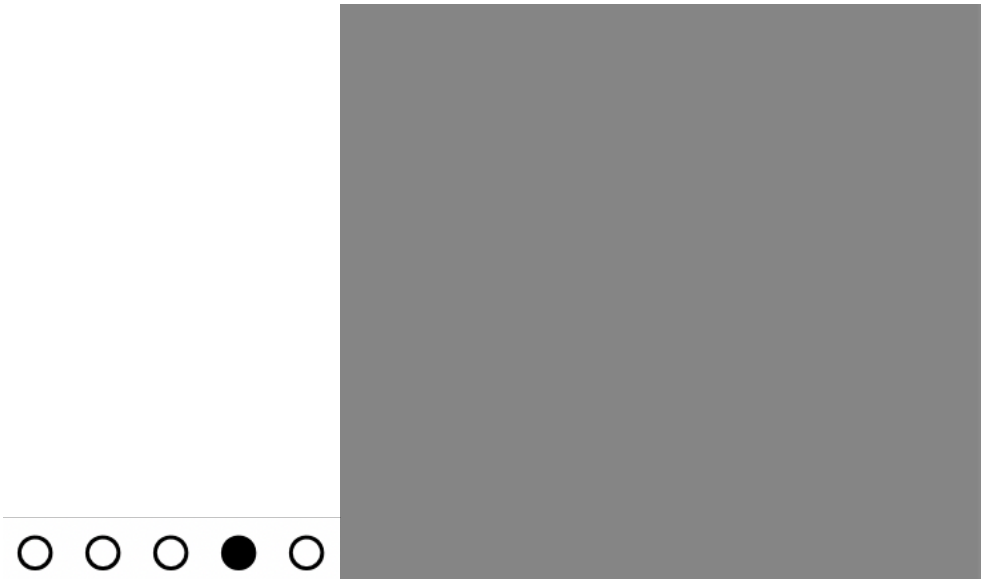
2. I need data to visualize to support my topic.
3. **2) Talk to a Librarian to help identify appropriate data sources for semester topic.**

   You have a minimum of 3 data sources. If all of your data sources are internet web sites, this page will be returned to you to find additional data sources.

Primary Sources: Baseball-reference Hitting Encyclopedia
Secondary Sources: The Major League Baseball Ultimate Book of Records, by Major League Baseball *Tertiary Sources: Books and Articles that synthesize secondary data for general readers*

4. **3) How did you evaluate the data sources provided above for relevance and reliability?**

   The encyclopedia on baseball-reference is very reliable, since the mission of the site is just to provide info. Thus, it can be trusted. It is also quite relevant since it has data from over 100 years and is updated daily.



Name: ____Josh Knull_____ CGT 270 Section (001 or LC): LC_____
[List three sites relevant to your topic (**that you chose and listed in the Identifying Topics Worksheet**) that offer

different perspectives on the same issue.] Site 1: baseball-reference.com
Site 2: MLB.com
Site 3: ESPN.com

1) Analyze each web page by asking yourself the following questions.
Site One Site Two Site Three

Who is responsible for creating this website?	Sean Forman and his team	Major League Baseball Advanced Media	ESPN
When was this website created?	April 2000	1995	1995
What are the strengths of this website?	Gives up-to-date and historical statistics from almost every baseball game in MLB history	Gives live updates of games, players, the current season. Has some history stats as well	Gives live updates of games, players, the current season. Has some history stats as well. Allows fans to interact and play games
How even-handed is this website? Does it have any potential for bias?	No bias. Nothing would be gained from lying about stats	The journalists of specific teams are certainly biased, but the site as a whole is not	The journalists of specific teams are certainly biased, but the site as a whole is not

2) Compare the three sites; how are they different?

Site One Site Two Site Three

Who is the intended audience?	Any baseball fan. More specifically, hard-core baseball fans looking for very specific statistical insights	Any baseball fan looking for live news	Any sports fan looking for live news
What is its primary focus?	Listing baseball statistics over time	Giving live updates about the league	Giving live updates about sports
What evidence is provided for claims?	It is implicit that the stats are correct	It is implicit that the stats and news are correct	It is implicit that the stats and news are correct

3) Using the database of your choice, find a recent scholarly article on the same topic and note the author(s), title, journal, volume, date, and page numbers. How does the article compare to the information you found on the three websites?

I found an article looking at how baseball players perform at different ages. This is a very specific insight, something that my listed websites do not usually look into. They just have the hard/raw data, whereas this list has filtered that for its study.

4). If you were going to cite one of these web pages as a source, you would need to seek out the following information. Choose one of the sites and see if you can identify these elements.

► Author(s); check bottom of page for a personal link; look for “about” links (Note: For some websites, like those of government agencies and non-profit organizations, the author may be the office or organization, rather than a specific person.)

Sean Forman
► Title; if there is no obvious title, use: Home page

Sports Refrence
► Title of the entire website, if the page is part of a larger site Data the page was created or updated; check the bottom of the age

Updated Daily ► Web address

Write a citation for one of the pages in APA format. If you do not know APA format: visit Purdue Writing Lab.

Forman, S. (2000, April). MLB Stats, Scores, History, & Records. Retrieved from https://www.baseball-reference.com/

Name: Richard Wu Course: CGT 27000-LC3 Term: Fall 2019 Worksheet ID: WS005

Semester Topic Proposal **Goal:** Develop semester project proposal.

Objectives: Students will define a semester topic and clearly define the problem to be addressed, provide a brief description of what's been done before, the significance of the problem and their solution (what they plan to do).

Outcomes: Students will submit a project proposal for approval describing their approach for addressing the defined problem.

Your proposal will be reviewed, and your semester topic evaluated along with your previously submitted Topics to Data Sources worksheets. If your proposal and worksheets support a *Good Research Topic/Question* with access to adequate data to work with, your semester topic will be approved, and you will be encouraged to proceed with collection, processing and visualization of data. If your proposal and worksheets *do not* support a Good Research Topic and/or if there is insufficient data available, your proposal will be denied, and you will be encouraged to refine your research question. You will have one week to refine your research question, update your proposal to reflect the revised topic and submit the revised proposal for approval. If your proposal is denied a second time, you will be given a topic by Dr. Byrd for your semester project.

Answer each question below in 250 words or less and complete sentences. Your answer must be single space, Times New Roman, 11-point font and, there should be a blank line between each question/answer block. DO NOT change the margins, DO NOT re-order the questions. Failure to follow instructions will result in loss of points.

1) **Describe your research problem/question.**

The proposed research problem revolves around analyzing NBA player tracking data across different regular seasons and creating different performance categories (ex. all-star level, outstanding, mediocre, and poor) based on the average performance level of the league during each specific season. The average standards of these different categories will then be compared across several seasons to conclude trends regarding the overall competitiveness of the league.

2) Who is your audience?

The audience for this research problem would include NBA players, professional trainers, and sports analysts.

3) Describe why the work is needed and/or is important?

The work is important because the players and trainers will need this information in order to effectively plan and prepare themselves for the competitiveness of the league in the future. Sports analysts also need this information to provide accurate analysis of specific player's performance level and make predictions on seasons ahead.

4) Explain what remains to be done and what you plan to do differently.

Analysis of all-star and non-all-star players' statistics have been done previously. Further categorization of non-all-star players into "outstanding", "mediocre", and "poor" performance levels are needed. In addition, the standards of these categorizations will also be compared across different regular seasons.

5) What does the data look like? Have you downloaded the data? Is it publicly available?

The data is publicly available on the NBA website. It is in the format of a table. Parts of the data have already been downloaded.

6) What visualization tool(s) do you plan to use?

Tableau will be the main visualization tool used for this project.

7) Explain the significance/importance of your work (why the problem/question is important).

The result of the project will provide a better insight on the performance of different players and the league in general to not only professional analysts but also anyone who is interested in basketball. Furthermore, it will also benefit the players, coaches, and trainers, as they have a sense of other players' level of performance.

8) Explain what the expected outcomes of your work will be.

The expected outcome of the project is an interactive visualization from which the audience can easily understand and manipulate the variables to obtain the information they need. It is also expected that the overall appearance is simple and straightforward.

Name: Richard Wu Course: CGT270-001 Term: F2019 Worksheet ID: WS001

Topics

Goal: To generate topics for semester project.
Objectives: Students will identify a list of topics of interest.
Outcomes: Students will generate, compare and rank topics of interest and identify who the audience is.

Generate

1. **1) List three topics that interests you**
    1. *Analyzing NBA player tracking data to conclude trends regarding players' performances across different seasons.*
    2. *The influence that genres have on each movie's rating on IMDb.*

3. *The relationship between the length of a song and its rating on Billboard Top 100.*

2. **2) Which topic interests you the most?**

a. Analyzing NBA player tracking data to conclude trends regarding players' performances across different seasons.

3. **3) What do you know about each topic?**
    1. *Topic 1: Players who are in the all-star team are players with outstanding performance levels.*
    2. *Topic 2: A movie is considered to be poor when the rating is below 5.*
    3. *Topic 3: There is a correlation between the popularity of the song and the amount of time it stays on the*

       top 100 list.

4. **4) What do you not know about each topic?**
    1. *Topic 1: The factors that separates a good player from a bad/underperforming player.*
    2. *Topic 2: The criteria for the evaluation of each movie.*
    3. *Topic 3: The process for determining the popularity of each song.*

Evaluate

5) Compare and rank topics

a. In this space evaluate your topics, what you know what you do not know about them. Rank the topics according to your preference then answer the following questions about each: Is this the right challenge to tackle? Is there enough detail or is there too much?

Analyzing NBA player tracking data to conclude trends regarding players' performances across different seasons.

Known: Players who are in the all-star team are players with outstanding performance levels.
Unknown: The factors that separates a good player from a bad/underperforming player.
I believe this is a right challenge to tackle because findings to this challenge can potentially help increase players' future performance.

There appears to be enough details for this topic.

The influence that genres have on each movie's rating on IMDb.
Known: A movie is considered to be poor when the rating is below 5.

Unknown: The criteria for the evaluation of each movie.

I believe this is a right challenge to tackle because findings to this challenge can help improve the quality of upcoming movies.

There appears to be enough details for this topic.

The relationship between the length of a song and its rating on Billboard Top 100.
Known: There is a correlation between the popularity of the song and the amount of time it stays on the top 100 list.
Unknown: The process for determining the popularity of each song.

I believe this is a right challenge to tackle because findings to this challenge allow artists to create more songs that will receive a lot of popularity.

There appears to be enough details for this topic.

Byrd Data Visualization Lab © Worksheet Due: Week 2

Name: Richard Wu Course: CGT 270 Term: Fall 2019 Worksheet ID: WS002

From Topics to Questions

Goal: To articulate what the student hopes to find out about the Topic of interest (Box 2) in the Topics Worksheet. **Objectives:** Students will transition from a broad topic to a specific question.
Outcomes: Students will answer who/what/when/where/whether/why/how about their topic

You are no longer a mere data collector, but a researcher interested in understanding something better. List your topic of interest from the Topics Worksheet) below in a complete sentence.

Topic:

1) The topic I am studying is: Analyzing NBA player tracking data to conclude trends regarding players' performances across different seasons

Because:

2. **2) I want to find out what/why/how: What separates a good player from a bad/underperforming player?**
3. **3) In order to help my reader understand: How to properly categorize players' performance levels based on**

   their statistics.

Resource:

Booth, W. C., Colomb, G. G., Colomb, G. G., Williams, J. M., & Williams, J. M. (2003). *The craft of research*.

University of Chicago press. (Chapter 3)

Byrd Data Visualization Lab © Worksheet Due: Week 2

Name: Richard Wu Course: CGT270 Term: Fall 2019 Worksheet ID: WS003

From Questions to a Problem

Goal: To generate questions based on the Topic of interest (Box 2) in the Topics Worksheet.
Objectives: Students will transition from a broad topic to a specific question.
Outcomes: Students will answer who/what/when/where/whether/why/how about their topic.

You are no longer a mere data collector, but a researcher interested in understanding something better. List your topic of interest (from box 2) on the Topics Worksheet) below (in a complete sentence):

1. **1) Identify the significance of your topic by completing the following sentences**
    1. Topic: Analyzing NBA player tracking data to conclude trends regarding players' performances across different seasons.
    2. Question: How can we reliably categorize players based on their performances?
    3. Significance: in order to help my reader understand the factors that contribute toward a

       professional player's performance level.

2. **2) Use the space below to determine if your problem is practical or a conceptual problem by answering the following questions about your topic.**

a. What's the situation or condition you desire to address, and why is it important?

NBA players' performance levels are changing over time as a result of general problems that they face throughout their career. This is important because the impact done by some of these problems can be reduced or even eliminated through creative solutions.
b. What's the undesirable consequences caused by that condition including costs that you (or your

readers) don't want to pay? Why is it important?

The undesirable consequences might include shorter careers for average players and lower level of competitiveness in the league. This is important because it hinders the players and the professional organizations.
c. What do we do about it?

We can analyze the data and create potential solutions that maintain or improve the performance of

the athletes in the long run.

3) Practical vs Conceptual Problems: complete on of the two questions about your problem

a. Practical Problem: What do you want your reader to do?

The question intends to make the reader acknowledge the general problems that professional athletes face and generate possible solutions to these problems to improve players' level of performance.

b. Conceptual Problem: What should the reader think?

The reader should think about the impact of the problems NBA players face and construct a solution

that responds to these problems.

4) Significance

a. Explain why you want your reader to know and care about the topic.

The topic is important to look into because it can potentially help improve professional athletes' performances by avoiding or neutralizing the factors that hinder their capacity to play.

Resource:

Booth, W. C., Colomb, G. G., Colomb, G. G., Williams, J. M., & Williams, J. M. (2003). *The craft of research*.

University of Chicago press. (Chapter 4)

Byrd Data Visualization Lab © Worksheet Due: Week 3

Name: Richard Wu Course: CGT 270 Term: Fall 2019 Worksheet ID: WS004

From a Problem to Data Sources

Goal: Acquire data for semester topic.
Objectives: Students will identify appropriate data sources transition from a broad topic to a specific question. **Outcomes:** Students will plan, generate, and determine the availability of data for the semester project.

1. **1) Describe your semester topic below**
    1. I am working on the topic of: *describe your topic*
       Analyzing NBA player tracking data to conclude trends regarding players' performances across different

       seasons.

    2. Because I want to find out: *describe research goal*

       To distinguish the differences between each player's performance level and group them into poor, mediocre, outstanding, and all-star level based on the overall data of a particular season.

    3. In order to help my reader understand better: *describe intention for reader*
       The league's performance level for different seasons and how each player is categorized based on their

       performance with respect to other players' performances during a particular season.

2. I need data to visualize to support my topic.
3. **2) Talk to a Librarian to help identify appropriate data sources for semester topic.**

   You have a minimum of 3 data sources. If all of your data sources are internet web sites, this page will be returned to you to find additional data sources.

   Primary Sources: *Raw Data; where application*
   Dataset provided by NBA website:
   https://stats.nba.com/players/traditional/?sort=PTS&dir=-1
   Secondary Sources: *Literature; must have at least 1 secondary source, best secondary sources are books from university press and/or articles that have been peer reviewed.*

   Article from Purdue Online Database:

   PLoS ONE, 10
   Tertiary Sources: *Books and Articles that synthesize secondary data for general readers*

Article from Purdue Online Database:

PLoS ONE, 10

4. **3) How did you evaluate the data sources provided above for relevance and reliability?**

To evaluate the relevance of the data sources above, the publishing date of each data source can be analyzed and compared to other sources of the same topic to determine its state of relevancy. As for the reliability of the data

Jaime Sampaio, Tim Mcgarry, Julio Calleja-González, Sergio Jiménez Sáiz, Xavi Schelling I Del Alcázar,

& Mindaugas Balciunas. (2015). Exploring Game Performance in the National Basketball Association

Using Player Tracking Data.

Byrd Data Visualization Lab © Worksheet Due: Week 4

(7), E0132894.

Brian Skinner, & Stephen J Guy. (2015). A Method for Using Player Tracking Data in Basketball to

Learn Player Skills and Predict Team Performance.

(9), E0136393.

sources, each author can be researched regarding their background information and the citations used can also be checked for their credibility.

Byrd Data Visualization Lab © Worksheet Due: Week 4

Web Page Evaluation

Name: ___Richard Wu_____ CGT 270 Section (001 or LC): _____LC___ [List three sites relevant to your topic (**that you chose and listed in the Identifying Topics**

Worksheet) that offer different perspectives on the same issue.]
Site 1: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4501835/
Site 2: https://towardsdatascience.com/which-nba-players-are-most-similar-machine-learning-provides-the-answers-r-project-b903f9b2fe1f
Site 3: https://bleacherreport.com/articles/1813902-advanced-nba-stats-for-dummies-how-to-understand-the-new-hoops-math#slide2
1) Analyze each web page by asking yourself the following questions.

Site One Site Two

Site Three

Who is responsible for creating this website?	The National Center for Biotechnology Information is responsible for creating this website.	The website is made by Toward Data Science Journal.	Bleacher Report is responsible for creating this website.
When was this website created?	The web page is published on July 14th, 2015.	The web page is published on November 10th, 2018.	This article is published on October 18th, 2013.
What are the strengths of this website?	The website has credibility since it's a government site. The content of the article is also quite relevant.	The article is part of a credible journal. In addition, the publish date reveals that the article is also recent.	The website is a popular sports website that is trusted by the general public.
How even-handed is this website? Does it have any potential for bias?	The website is even-handed because it uses a scientific approach backed with evidences.	Potential bias does not exist since the article revolves around a statistical analysis of the topic.	The site does not contain any potential bias because it

2) Compare the three sites; how are they different?
Site One Site Two

Site Three

Who is the intended audience?	The intended audience include people who are interested in analyzing the performance of different NBA players.	The intended audience include people who are interested in comparing different players in the NBA.	The intended audience include people who are unfamiliar with some of the advanced concepts used to analyze player performance in basketball.
What is its primary focus?	The primary focus is to introduce a method of statistical analysis	The primary focus is to educate the reader about several	The primary focus is to introduce some of the terminologies and

Adapted from Folke Bernadette Memorial Library, Gustavas Adolphus College -

http://gustavus.edu/academics/library/exercies/webevaluation.doc

Adapted from Folke Bernadette Memorial Library, Gustavas Adolphus College -

http://gustavus.edu/academics/library/exercies/webevaluation.doc

Web Page Evaluation

	for the performance of different athletes.	methods of analysis and the variables used for each method.	variables that can be used for statistical analysis.

What evidence is provided for claims?	Data from the 2013- 2014 regular season is used for evidence.	Data from the 2017-2018 regular season is used for evidence.	Data from the 2012-2013 regular season is used for evidence.

Web Page Evaluation

3) Using the database of your choice, find a recent scholarly article on the same topic and note the author(s), title, journal, volume, date, and page numbers. How does the article compare to the information you found on the three websites?
Author: Scott Bruce
Title: A scalable Framework for NBA Player and Team Comparison Using Player Tracking Data
Journal: Journal of Sports Analytics
Volume: 2
Date: 2016
Page Numbers: 107-119
The article provides an in-depth analysis using the Principle Component Analysis (PC) method to distinguish several components that determine the variation in the performance of NBA players. The article is similar to the information found on the three websites in the way that its purpose is to introduce a method of statistical analysis to the reader. However, the scholarly article is more detailed and provides a full-scale example of one specific type of analysis.

4). If you were going to cite one of these web pages as a source, you would need to seek out the following information. Choose one of the sites and see if you can identify these elements.

► Author(s); check bottom of page for a personal link; look for "about" links (Note: For some websites, like those of government agencies and non-profit organizations, the author may be the office or organization, rather than a specific person.)
Jaime Sampaio, Tim McGarry, Julio Calleja-González, Sergio Jiménez Sái, Xavi Schelling i del Alcázar, and Mindaugas Balciunas.

► Title; if there is no obvious title, use: Home page
Exploring Game Performance in the National Basketball Association Using Player Tracking Data
► Title of the entire website, if the page is part of a larger site
National Center for Biotechnology Information
Data the page was created or updated; check the bottom of the age
July 14, 2015
► Web address
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4501835/
Write a citation for one of the pages in APA format. If you do not know APA format: visit Purdue Writing Lab.

PloS one, *10*
Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4501835/

[redacted]

Sampaio, J., McGarry, T., Calleja-González, J., Jiménez Sáiz, S., Schelling I Del Alcázar, X., &

[redacted]

Balciunas, M. (2015). Exploring Game Performance in the National Basketball Association

[redacted]

Using Player Tracking Data.

Adapted from Folke Bernadette Memorial Library, Gustavas Adolphus College -

http://gustavus.edu/academics/library/exercies/webevaluation.doc

(7), e0132894. doi:10.1371/journal.pone.0132894

Name: Jitesh Motati Course: CGT 27000-LC3 Term: Fall 2019 Worksheet ID: WS005

Semester Topic Proposal **Goal:** Develop semester project proposal.

Objectives: Students will define a semester topic and clearly define the problem to be addressed, provide a brief description of what's been done before, the significance of the problem and their solution (what they plan to do).

Outcomes: Students will submit a project proposal for approval describing their approach for addressing the defined problem.

Your proposal will be reviewed, and your semester topic evaluated along with your previously submitted Topics to Data Sources worksheets. If your proposal and worksheets support a *Good Research Topic/Question* with access to adequate data to work with, your semester topic will be approved, and you will be encouraged to proceed with collection, processing and visualization of data. If your proposal and worksheets *do not* support a Good Research Topic and/or if there is insufficient data available, your proposal will be denied, and you will be encouraged to refine your research question. You will have one week to refine your research question, update your proposal to reflect the revised topic and submit the revised proposal for approval. If your proposal is denied a second time, you will be given a topic by Dr. Byrd for your semester project.

Answer each question below in 250 words or less and complete sentences. Your answer must be single space, Times New Roman, 11-point font and, there should be a blank line between each question/answer block. DO NOT change the margins, DO NOT re-order the questions. Failure to follow instructions will result in loss of points.

1. 1) **Describe your research problem/question.**
   My research question for the semester project is: How does player efficiency rating impact win percentage in the

   NBA?

2. **2) Who is your audience?**

   My audience for this research question is any person interested in sports analytics or looking to research the

   impact of an individual player on a sports team.

3. **3) Describe why the work is needed and/or is important?**

   This work is important because it can help analyze the effect of a single player on a team's success. For the average fan, they usually see the player that scores the most points as most impactful. This project will help show that basketball is a team game and that certain players being more efficient can make a greater impact without scoring the most.

4. **4) Explain what remains to be done and what you plan to do differently.**

   I need to figure out a way to show if there is a correlation between a range of PER and win percentage. I have to provide visualizations that show concretely how much impact each player has on games over the season. I plan to look at more data and see if there is an easy way to put PER and win percentage together and make a common factor/number.

5. **5) What does the data look like? Have you downloaded the data? Is it publicly available?**

   The data is in .csv files or Excel file format. I have downloaded the data and it is publicly available to download.

   The data is available at no cost.

6. **6) What visualization tool(s) do you plan to use?**

   I plan to use Excel to make some basic visualizations and then use Tableau to refine those visualizations. I would

   use Tableau to make the visualizations more interactive.

7. **7) Explain the significance/importance of your work (why the problem/question is important).**

My work is significant because it will help prove to the average fan that players that score less points can have a greater impact on the game. It will allow people to understand how there is more than one aspect to the game and that there are other ways to make an impact without just scoring.

8) Explain what the expected outcomes of your work will be.

The expected outcomes of the work will be to have a visualization that proves whether higher PER leads to a better win percentage or not.

Byrd Data Visualization Lab © 2019

Name: Jitesh Motati Course: CGT270 Term: Fall 2019 Worksheet ID: WS001

Topics

Goal: To generate topics for semester project.
Objectives: Students will identify a list of topics of interest.
Outcomes: Students will generate, compare and rank topics of interest and identify who the audience is.

Generate

1. **1) List three topics that interests you**
   1. *I want to visualize major statistics in the sport of baskeball.*
   2. *Visualize popularity of media platforms such as YouTube and Twitch and look at average revenue change*

of content creators over time.

3. *The visualization of popularity of programming languages over time.*

2. **2) Which topic interests you the most?**

a. The topic that interests me the most is visualizing certain statistics in basketball,.

3. **3) What do you know about each topic?**
    1. *Topic 1: There are huge datasets and historical data that can easily be accessible.*
    2. *Topic 2: The popularity of certain platforms often depend on the type of content creators using that*

       platform.

    3. *Topic 3: The popularity usually usually depends on which language is more powerful or ease of use.*
4. **4) What do you not know about each topic?**
    1. *Topic 1: I do not know what specific dataset want to look at because there are multiple different statistics for this topic.*
    2. *Topic 2: I do not know if there is enough historical data to make an accurate conclusion.*
    3. *Topic 3: I do not know where I would get the data from and if the data would be similar enough to*

       compare.

Evaluate

5) Compare and rank topics

1. *I think this topic might be the right challenge because it has the largest dataset and we can visualize this in many different ways. There are definitely enough data on this topic.*
2. *The data can be easily accessible and is public. Visualizing this might be an interesting challenge and there is enough detail.*
3. *This challenge may not be the best because there isn't necessarily an easy way to access data and may not have enough data.*

Byrd Data Visualization Lab © Worksheet Due: Week 2

Name: Jitesh Motati Course: CGT270-LC3 Term: Fall
Worksheet ID: WS002

From Topics to Questions

Goal: To articulate what the student hopes to find out about the Topic of interest (Box 2) in the Topics Worksheet. **Objectives:** Students will transition from a broad topic to a specific question.
Outcomes: Students will answer who/what/when/where/whether/why/how about their topic

You are no longer a mere data collector, but a researcher interested in understanding something better. List your topic of interest from the Topics Worksheet) below in a complete sentence.

Topic:

1) The topic I am studying is per game statistics over the years in the NBA.

Because:

2. 2) **I want to find out what/why/how:** why there is an increase in one of the statistics (Pace) and if the other statistics are the reason that there is an increase in pace each year.
3. **3) In order to help my reader understand:** that there is a major difference in how basketball is played compared to 10 years ago.

Resource:

Booth, W. C., Colomb, G. G., Colomb, G. G., Williams, J. M., & Williams, J. M. (2003). *The craft of research*.

University of Chicago press. (Chapter 3)

Byrd Data Visualization Lab © Worksheet Due: Week 2

Name: Jitesh Motati Course: CGT270-LC3 Term: Fall 2019 Worksheet ID: WS003

From Questions to a Problem

Goal: To generate questions based on the Topic of interest (Box 2) in the Topics Worksheet.
Objectives: Students will transition from a broad topic to a specific question.
Outcomes: Students will answer who/what/when/where/whether/why/how about their topic.

You are no longer a mere data collector, but a researcher interested in understanding something better. List your topic of interest (from box 2) on the Topics Worksheet) below (in a complete sentence):

1. **1) Identify the significance of your topic by completing the following sentences**

1. Topic: The topic I am studying is per game statistics over the years in the NBA.
2. Question: because I want to find out what/why/how

   What have players done to increase the amount of scoring and pace each game?

3. Significance: in order to help my reader understand

   How the game of basketball has evolved, and which aspects of the game are focused on more

   today?

2. **2) Use the space below to determine if your problem is practical or a conceptual problem by answering the following questions about your topic.**
   1. What's the situation or condition you desire to address, and why is it important?

      The situation I want to address is the increase of scoring and how it is occurring. It is important because it allows fans to have a better understanding of how the game has changed over time.

   2. What's the undesirable consequences caused by that condition including costs that you (or your readers) don't want to pay? Why is it important?

      Undesirable consequences caused by the condition include shorter game time and less viewing time for fans.

   3. What do we do about it?
      We cannot change this unless the style of play changes again.
3. **3) Practical vs Conceptual Problems: complete on of the two questions about your problem**
   1. Practical Problem: What do you want your reader to do?
   2. Conceptual Problem: What should the reader think?
4. **4) Significance**

a. Explain why you want your reader to know and care about the topic.
I want my reader to know and care about the topic because it helps them understand why games

today feel a lot shorter than in the past. This is important because fans pay a lot of money to watch games and if average game time is shorter, they don't get the most viewing time for their money.

Resource:

Booth, W. C., Colomb, G. G., Colomb, G. G., Williams, J. M., & Williams, J. M. (2003). *The craft of research*.

University of Chicago press. (Chapter 4)

Byrd Data Visualization Lab © Worksheet Due: Week 3

Name: Jitesh Motati Course: CGT270-LC3 Term: Fall
Worksheet ID: WS004

From a Problem to Data Sources

Goal: Acquire data for semester topic.
Objectives: Students will identify appropriate data sources transition from a broad topic to a specific question. **Outcomes:** Students will plan, generate, and determine the availability of data for the semester project.

1. **1) Describe your semester topic below**
    1. I am working on the topic of: *player efficiency in the NBA.*
    2. Because I want to find out: *how player efficiency has an impact on win percentage over a season of NBA*

       basketball.

    3. In order to help my reader understand better: *why players with high player efficiency are usually best in*

       the game and make the most impact on the team.

2. I need data to visualize to support my topic.
3. **2) Talk to a Librarian to help identify appropriate data sources for semester topic.**

   You have a minimum of 3 data sources. If all of your data sources are internet web sites, this page will be returned to you to find additional data sources.

   Primary Sources: *https://www.nbastuffer.com/2018-2019-nba-player-stats/*
   Secondary Sources: *Literature; must have at least 1 secondary source, best secondary sources are books from university press and/or articles that have been peer reviewed.*
   Basketball Analytics: Objective and Efficient Strategies for Understanding How Teams Win
   Tertiary Sources: *Books and Articles that synthesize secondary data for general readers*

   Article: Visualizing One Million NCAA Basketball Shots

https://minimaxir.com/2018/03/basketball-shots/

3) How did you evaluate the data sources provided above for relevance and reliability?

To evaluate the data, I looked through each statistical column and figured out which fields may be most useful to make a visualization that supports my question. For the secondary source, I read through a couple chapters of the book to better understand how the analytics are tracked and looked at some of the datasets. The datasets gave me a better idea of what I needed to do to get reliable and useful statistics. The article *Visualizing One Million NCAA Basketball Shots* gave me some inspiration on what kind of visualization I could possibly make for my semester project.

Byrd Data Visualization Lab © Worksheet Due: Week 4

Web Page Evaluation
Name: ____Jitesh Motati_____ CGT 270 Section (001 or LC): _____LC_____

[List three sites relevant to your topic (**that you chose and listed in the Identifying Topics Worksheet**) that offer different perspectives on the same issue.]

Site 1: https://www.kaggle.com/drgilermo/nba-players-stats
Site 2: http://www.nbaminer.com/margins-comebacks/
Site 3: https://www.nbastuffer.com/2018-2019-nba-team-stats/
1) Analyze each web page by asking yourself the following questions.

Site One Site Two

Site Three

Who is responsible for creating this website?	Omri Goldstein on Kaggle.com. was responsible for creating this website.	NBA Miner is responsible for creating this website, it is not credited to a single author.	NBAstuffer is the creator of this website and is not affiliated with the NBA.
When was this website created?	This website was last updated a year ago.	This website was created in 2014, which was the last copyright date. However, it seems to be updated every season/year.	This website was created in 2007 and updated daily during the NBA season.
What are the strengths of this website?	This website gives data that is scraped from Basketball-reference in an easy to read csv format.	The strengths of this website are that it shows data on a deeper level. It has stats such as margins won/lost by and comebacks. It also has an easy to use export function.	The strengths of this website include explanations of different analytics that are shown in its data sets and an easy to export option.
How even-handed is this website? Does it have any potential for bias?	It does not have any potential bias because it shows the raw statistics.	The website seems to not show any bias as it just uses raw statistics as a basis for its analytics.	This website is not biased because it uses statistics from the NBA as a basis for any analysis.

2) Compare the three sites; how are they different?
Site One Site Two

Adapted from Folke Bernadette Memorial Library, Gustavas Adolphus College -

http://gustavus.edu/academics/library/exercies/webevaluation.doc

Site Three

Who is the intended audience?	The intended audience is anyone looking to visualize or manipulate data.	This website is intended more for people looking to	This site is more for the casual sport statistics enthusiast

Adapted from Folke Bernadette Memorial Library, Gustavas Adolphus College -

http://gustavus.edu/academics/library/exercies/webevaluation.doc

Web Page Evaluation

		research analytics on a deeper level.	and does provide useful datasets.
What is its primary focus?	Primary focus is to give raw data to those that would like to manipulate it.	The primary focus is to use the datasets to support certain claims/analysis.	The primary focus is to have a centralized place to view all NBA statistics over a certain time period.
What evidence is provided for claims?	The evidence for the claims is in the data, and this website does not make many claims.	The website uses their data to prove some of their claims regarding NBA teams/players performance.	The website does not really make any claims and it just presents itself as a centralized data storage website.

Web Page Evaluation

3) Using the database of your choice, find a recent scholarly article on the same topic and note the author(s), title, journal, volume, date, and page numbers. How does the article compare to the information you found on the three websites?

https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Stanley_Yang%20_Thesis.pdf

Author: Yuanhao (Stanley) Yang
Title: Predicting Regular Season Results of NBA Teams Based on Regression Analysis of

Common Basketball Statistics

Journal: Berkeley Statistics Page Numbers: 11-28

This article has a lot more analysis on the raw data. Most of the 3 websites just provide data and some basic analysis. This article goes in depth on how certain statistics effect win percentage and has a discussion of results.

4). If you were going to cite one of these web pages as a source, you would need to seek out the following information. Choose one of the sites and see if you can identify these elements.

► Author(s); check bottom of page for a personal link; look for "about" links (Note: For some websites, like those of government agencies and non-profit organizations, the author may be the office or organization, rather than a specific person.)

Author: Omri Goldstein

► Title; if there is no obvious title, use: Home page

Title: NBA Players stats since 1950

► Title of the entire website, if the page is part of a larger site

Kaggle.com

Data the page was created or updated; check the bottom of the age

Last updated approximately a year ago.

► Web address

Write a citation for one of the pages in APA format. If you do not know APA format: visit Purdue Writing Lab.

Adapted from Folke Bernadette Memorial Library, Gustavas Adolphus College -

http://gustavus.edu/academics/library/exercies/webevaluation.doc

Goldstein, O. (2018, April 27). NBA Players stats since 1950. Retrieved from

https://www.kaggle.com/drgilermo/nba-players-stats