

DS 3000 – Topic Proposals

The first deliverable of the project is a topic proposal for your DS project. You should start by thinking about a real-world problem that interests you and also has clear connections to an area to which data science can contribute. **Each group** will submit 2-3 topic proposals (depending on whether your group consists of two or three members). For instance, if your group has two members, you should submit two different ideas **as a team** (not individually). Then your professor and/or TAs will help evaluate which topic seems most promising for this project.

For each of the proposed ideas, the deliverable should include a description of the problem you would like to tackle, why you think it matters, and whether you have identified any potential datasets that might be useful for the project.

Please refer to the sample project papers posted on Canvas (Final Project > Sample Projects).

What to submit

- ✓ Complete the following sections for each topic idea.
- ✓ Before submitting your proposals, **delete this first page** containing the instructions.
- ✓ Submit your completed document (as a PDF) on Canvas.
- ✓ One submission per team (no individual submissions necessary)

Topic Idea 1: <Small Ball Lineups>

1. Problem Statement

- ✓ Describe the problem you would like to tackle.
The problem we want to tackle is figuring out how effective the small ball lineup is in the NBA. Recently with the advance of advanced analytics in basketball, more teams have seen the usefulness of having a smaller lineup (hence the name small ball) that can shoot 3s and play faster. The main idea being faster teams that shoot more 3s on a high enough % will outscore any team that takes 2s at a slower rate. Many teams in the league have gone with smaller lineups to increase the pace of the game. However, many people, especially former players, believe that small ball lineups are not good enough to win a championship, no matter what the mathematics and analytics says. They point to the Rockets, who have tried to go extremely small, and point to their postseason woes as evidence.
- ✓ What is the topic of your project?
The topic of this project is the small ball lineup in basketball, which is frequently defined by lineups that do not have the traditional center, instead moving the power forward over to the center position and putting another strong shooter and defender at the 4 (power forward position).
- ✓ What do you want to learn about it?
We want to learn about how effective it is compared to other lineups using statistics like offensive rating, defensive rating, team wins, championships, etc.

2. Significance of the Problem

- ✓ Why is it important to tackle this problem in your project?
This problem is important for the game of basketball because it could define how we see the game be played in the future. Teams that adapt to the increasing amount of statistics that we have may have a sizeable advantage when competing for a championship.
- ✓ In what ways could the insights from this project be useful?
The insights from this project will largely be useful for basketball but there are also things that can be taken away from it in a general sense. The debate of the effectiveness of small ball generally comes from difference in what the statistics and analytics say about basketball versus what the eye test says. Most former players and critics watch the game being played today and just don't see how such lineups can win the game (especially if they forgo centers entirely like the Rockets have done). A project looking into this could potentially give insight into why it seems that analytics are saying one thing but the eyes are saying something else.

3. Potential Datasets

- ✓ Look into some potential datasets using the online resources provided on Canvas or other data sources you find online. This is exploratory, so you won't commit to any specific dataset yet.
- ✓ You are not expected to identify a dataset yet. This is intended to get you started.
- ✓ Look through the datasets to identify any potential matches.
- ✓ Provide links to potential datasets here, if any.
<https://stats.nba.com/lineups/advanced/>
<https://www.basketball-reference.com/teams/TOR/2020/lineups> (Or any team by substituting TOR with the nickname of team)

Topic Idea 2: <How NBA Players perform relative to their contracts.>

4. Problem Statement

- ✓ Describe the problem you would like to tackle.
 - We would like to understand how players in the NBA are performing relative to their contract and find the contract that the player is actually worth.
- ✓ What is the topic of your project?
 - NBA players' performance relative to their salary, and what they are actually worth.
- ✓ What do you want to learn about it?
 - What players are outperforming their contract values, and what players are getting overpaid for their performance.

5. Significance of the Problem

- ✓ Why is it important to tackle this problem in your project?
 - In the NBA, especially with a salary cap, it is important to utilize the money that you give to players wisely. In order to win the championship, teams must not give out large contracts to players that do not actually deserve it. You must have players that perform at or above their contract value, in order to be successful in the league.
- ✓ In what ways could the insights from this project be useful?
 - It could help NBA GM's and players' agents understand what the player is worth, when negotiating contracts. As well as, show teams what players should be traded, as they are not performing up to the value of their contract.

6. Potential Datasets

- ✓ Look into some potential datasets using the online resources provided on Canvas or other data sources you find online. This is exploratory, so you won't commit to any specific dataset yet.
 - Basketball reference has a lot of data including contract values for all active players, season stats, and advanced mathematical stats for all players. It is very easy to obtain this data in a CSV, and the data is constantly update.
- ✓ You are not expected to identify a dataset yet. This is intended to get you started.
- ✓ Look through the datasets to identify any potential matches.
- ✓ Provide links to potential datasets here, if any.
 - <https://www.basketball-reference.com/contracts/players.html>
 - https://www.basketball-reference.com/leagues/NBA_2020_per_game.html
 - https://www.basketball-reference.com/leagues/NBA_2020_advanced.html
 - https://www.basketball-reference.com/leagues/NBA_2020_play-by-play.html
 - https://www.basketball-reference.com/leagues/NBA_2020_adj_shooting.html

The following popular dataset cannot be used for this project:

- Breast Cancer Wisconsin
- Boston House Prices
- Digits Dataset
- California Housing Dataset
- Any dataset bundled in sklearn (Iris, Diabetes, Wine, etc.)

Topic Idea 3: <Predicting Success of a Movie>

1. Problem Statement

- ✓ Describe the problem you would like to tackle.
 - The problem we are interested in tackling would be predicting the success of a movie based on its underlying factors, where we would look at various factors such as who the producers / directors were, social media presence, genre, actors, and much more and see if any of these does produce statistically significant results in whether or not they influence the success of a movie. To define what is successful or not, we will be basing success on sales and ratings.
- ✓ What is the topic of your project?
 - Predicting the Success of a Movie based on the information available on the movie, and whether or not certain factors does lead to a positive or negative success factor.
- ✓ What do you want to learn about it?
 - What variables in movies impact their success, and whether or not things like a social media presence would lead to better success. We will also be interested in learning about whether or not success can be predicted.

2. Significance of the Problem

- ✓ Why is it important to tackle this problem in your project?
 - Having a way to quantitatively determine the success rate of a certain film in production would directly impact how movies gain investors and a budget, where investors would be able to easily see how risky a certain investment would be. This would also serve a use in the indie film industry as movies with a tight budget would be able to also see what steps they could take to improve the chance of a success for their films.
- ✓ In what ways could the insights from this project be useful?
 - The insights from this project would be useful for both investors in the film industry and the producers themselves. Investors would be able to judge how risky an investment is, or what types of movies do result in good investments. For producers, this project would showcase what made specific movies succeed and would be able to apply this information to subsequent movies they are making. Any factor found to lead to a positive or negative success rate would lead to shifts in the movie industry around that factor, such as certain actors influencing the rate or different directors being a part of the movie. If social media is found to be statistically significant, we could also see marketing strategies changing around how they directly market a film.

3. Potential Datasets

- ✓ Look into some potential datasets using the online resources provided on Canvas or other data sources you find online. This is exploratory, so you won't commit to any specific dataset yet.
- ✓ You are not expected to identify a dataset yet. This is intended to get you started.
- ✓ Look through the datasets to identify any potential matches.
- ✓ Provide links to potential datasets here, if any.
 - <http://ai.stanford.edu/~amaas/data/sentiment/>
 - <https://www.imdb.com/interfaces/>
 - <https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset>