- Include a quick review of what has been done with this data and how this motivates your analysis.

I've been a fan of soccer for a very long time. For Americans, sports is the NFL, for me, it's soccer. Since coming to the United States, I have loved many sports, but soccer is still my most exciting sport. In soccer, specifically, I support Manchester United. United is going through a down period right now with the retirement of the greatest manager in soccer history, Sir Alex Ferguson, so I always worry about the group stage when they play in a competition called the UCL. The UEFA Champions League (historically known as the European Cup and mostly abbreviated worldwide as the UCL) is an annual club association football competition organized by the Union of European Football Associations (UEFA). Simply put, it's a competition between 32 of the best teams in Europe to see who is the best. It's every soccer player's dream, which is why it's so important to get through the group stage and make it to the round of 16.

When I became interested in data science, I suddenly thought, "Does a team's performance on the group stage lead them to the championship?". I have many questions about soccer always because analyzing soccer is my dream. This question specifically is important to all teams that attend UCL.

This raw data is something I've never used before. I'm going to scrape together the data I need from the internet to answer, so I'll detail below what data I need and what I need to do.

- Summarize the central argument/question of your final project - make it compelling

The UCL is a competition where there are always a lot of unpredictable results. Teams you don't think will win the tournament end up winning, and teams you think will win the tournament end up getting knocked out in the group stages, so I wanted to know how much group stage performance helps predict the tournament. The reason for using group stage performance to make analysis is simple if you know the system of the UCL. The UCL is based on the team's performance in the previous season's league ranking, so even if a team does well in the previous season and makes it to the UCL, they may not do well in the current season for various reasons. A lot of changes happen between seasons, mainly player transfers and head-coach changes. This is why a team's group stage performance in the current season is a good indicator of how well the team is doing in the current season, as it can point to the team's current season performance.

But then comes the question: how do you define good in the group stage? A team's success in the group stage can be measured by its position in the group, its winning percentage in the group, and its goals difference. The reason we're thinking about group position and group win percentage separately is because there are a number of different ways a team can advance to the round of 16 without having a good group win percentage: a team can win 6 games and have a 100% win percentage and advance to

the round of 16, or a team can win 3 games, draw 1 game, and have a 50% win percentage and advance to the round of 16.

- Describe the data sets you will use to answer this question and why they match your question (include links to relevant data sources) & Explain how you will retrieve and clean the data or how you will make sense of found data

The data I will be using is the Champions League all matches data from 2017 to 2023. It contains all the game information for a season. The data I want to use includes Round number, Date, Location, Home team, Away team, Group, Result, etc. I will use the data except for the Date and Location, which I will combine with the group stage results (if I can find a .csv of the group stage results, I'll combine that with my data). I will use this data to answer my question because I will get a lot of data just by processing it.

https://fixturedownload.com/results/champions-league-2022

I will drop the ‘Date’ and ‘Location’ from the above data, keep only the rounds (not tournaments), sort by group name, and record each result's winning and losing teams. Also, I will create a winning percentage from the wins, draws, and losses and record the goal difference. I will create a function to process the .csv file and create a new data frame with a .csv file for each season.

- Outline the kind of analysis outputs you'll produce (visualizations, tables, tests, etc.). This should be much more detailed and extensive than the EDA you've done for module assignments. Please include at least two methods you may use to get at your question and explain why they are appropriate for your question.

To answer this question, I think visualization will be the most important thing, and I will choose the right method by varying the visualization. For example, after processing the data, I will create a scatter plot to group the seasons in which a good performance in the group stage led to a UCL win and the seasons in which it did not, and I will use testing to check the correlation and p-value. This is appropriate because my question can be simply rephrased as "Is there a correlation between group stage performance and tournament performance?" If I realize that another method is more appropriate during the analysis project, I can use it. (I will use the method that best answers my question.

- Describe what limitations of your data and analysis approach you can foresee.

If there are any, I would expect them to come out in the visualization, or I might get a completely different result than I expected. But there are no expected limitations at this point.