Team Members:
Vivienne Hughes (ndg9tt)
Corinna Keum (pdx4tw)
Troy S. Meink (ked6na)
Brianna Seekford (bmx4af)

Our goal:
Determine the final winner of the 2022 Men's World Cup (Qatar), we are not going to try and fill out the entirety of the bracket.

Data we will be using:

- Corinna's data: historical world cup data → probably only the last few world cups as players on the team will likely be the same and we can get a better feel of who performs better/usually makes it farther.
- Viv's data: game-by-game statistics → we will take the group stage data and firstly decide who will make it out of the group stages and then we can compare that to who actually did. This cuts our field of teams in half already which, if done right, will hopefully significantly increase our chances of correctly predicting the winner.
- Brianna's data: statistics from individual matches during the 2022 season → this will help us compare the predicted outcomes to the actual outcomes. We will be able to use this to evaluate the strength of the teams over a period of time which can help predict the strongest team and thus the final winner. Probably best dataset we have for regression analysis.
- FIFA rankings of teams going into the world cup → help us know which teams are supposedly stronger going into the world cup

You should address the following questions explicitly:

- What is an observation in your study?
    - The historical data from individual World Cup matches will be the unit of observation for our study. We'll use match results from the 2022 group stage, and historical data from previous world cup years.
- Are you doing supervised or unsupervised learning? Classification or regression?
    - I guess we would be doing classification because we are saying they won they lost? Or maybe regression because we will say how likely we think that one team will win
    - All of our analyses will be supervised because we have match results (y-values) that our data will be fitted too. Our primary analysis will use classification, as the data will be formatted into a win/loss binary to ingest all of the match results. However, we may also use regression on in-game metrics (points score for/against, etc.) if our classification-based model isn't sufficiently accurate.

- What models or algorithms do you plan to use in your analysis? How?
    - We plan to use linear regressions on the data. We believe a linear regression on past match results will be a good way to predict results of matches into the near future (the knockout round of 16)
- How will you know if your approach "works"? What does success mean?
    - We will compare to the actual bracket and see how far teams made it to our estimation of the chance they had at winning, also can easily compare who we expect to make it out of group stages to who did so.
    - Success means that for the most part (70%? Is that too ambitious?) we predicted correctly which teams made it out of the group stages. After our model says who it thinks makes it out of each group then success will be determined by if our model predicts the winner and they were actually (real bracket comparison) a top two or three finisher (Argentina, France, Croatia).
- What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?
    - We anticipate that it will be difficult to predict winners the farther we get in the bracket because if our model has already made a mistake the mistake will follow us to the next step and it won't be possible to correct it. It will also be interesting to see how any upsets that occurred during the time frame of our data will affect our outcomes.

<u>You should address the following topics in the text, as appropriate:</u>

- Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?
    - Most of our data comes from Kaggle so it is very clean and won't require much wrangling. We will likely one-hot encode the win/loss results for each match to be used in a classification fit.
- Results: How will you communicate or present your results? This might be a table of regression coefficients, a confusion matrix, or comparisons of metrics like R^2 and RMSE or accuracy and sensitivity/specificity. This is how you illustrate why your plan succeeded or explain why it failed.
    - We will likely end up with a table of coefficients if we go with regression analysis (Brianna's data) and if we plan on documenting the games guessed correctly that will be calculated as a percentage (classification data).