INFO 432 Project Proposal
Project Group 4: Richardson Chhin, Kevin Shi, Kathryn Swatek
Dataset: FIFA 22 complete player dataset
(https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset?select=players_22.csv)

This dataset is of interest to us since we are fans of the sport of soccer. With the World Cup coming to the US next summer, we thought it would be cool to analyze the stats of the players from four years ago in the year leading up to the last World Cup in 2022. We hope to derive actionable insights into player characteristics, learn more about our favorite players and the top players in the sport, and evaluate our hypotheses and better understand what features define elite versus average players.

Dataset Description:
The dataset provides a comprehensive snapshot of 19,239 male professional football players (samples) across 110 features (variables) for the 2021-2022 season. For the purpose of quantitative modeling, unimportant metadata/identifiers will be excluded as they are not relevant for statistical analysis. The following features can be grouped into:
Core Metrics & Vitals: Primary indicators of a player's quality and status

- overall: The player's current rating (1-99)
- potential: The player's predicted peak rating (1-99)
- value_eur: Estimated market value in Euros
- wage_eur: Weekly wage in Euros
- age: Player's age in years

Physical & Demographic Attributes: Basic information about the player

- height_cm: Player's height in centimeters
- weight_kg: Player's weight in kilograms
- preferred_foot: Player's dominant foot (Ex: Right, Left)
- work_rate: Player's work effort (Ex: High/Medium)
- body_type: Player's physique (Ex: Lean, Stocky)

Positional and Summary Skills:

- player_positions: The player's primary listed position(s)
- pace
- shooting
- passing
- dribbling
- defending
- physic

Detailed Skill Attributes: Skill ratings (1-99) that are important to this analysis

- Attacking: attacking_crossing, attacking_finishing, attacking_heading_accuracy, attacking_short_passing, attacking_volleys
- Skill: skill_dribbling, skill_curve, skill_fk_accuracy, skill_long_passing, skill_ball_control
- Movement: movement_acceleration, movement_sprint_speed, movement_agility, movement_reactions, movement_balance

- Power: power_shot_power, power_jumping, power_stamina, power_strength, power_long_shots
- Mentality: mentality_aggression, mentality_interceptions, mentality_positioning, mentality_vision, mentality_penalties, mentality_composure
- Defending: defending_marking_awareness, defending_standing_tackle, defending_sliding_tackle
- Goalkeeping: goalkeeping_diving, goalkeeping_handling, goalkeeping_kicking, goalkeeping_positioning, goalkeeping_reflexes

Methodology Approach:
For our dimensional reduction analysis, we will utilize PCA to identify maximum variance directions across our 110 features, and Factor Analysis to uncover latent skill constructs that may represent underlying player abilities like technical skill and physical prowess. CCA is not applicable to our single dataset structure.

Initial Hypotheses:
Hypothesis 1:
We expect that players will naturally cluster into groups that correspond to traditional football positions (defenders, midfielders, forwards, goalkeepers), with each cluster showing distinct skill profiles. However, we also anticipate discovering hybrid roles that reflect modern football's tactical evolution.

Hypothesis 2:
A player's market value (value_eur) and wage (wage_eur) is significantly driven by a combination of age, potential, and specific latent skill factors. We predict that for younger players, potential will be a stronger driver of value than overall ability.

Hypothesis 3:
A smaller set of latent dimensions (e.g., technical ability, physical strength, tactical intelligence) can effectively represent player performance across all attributes. We need to discover what the most impactful features are.

- Hypothesis 1 is perfect for clustering analysis
- Hypothesis 2 is excellent for dimensional reduction (exploring latent factors)
- Hypothesis 3 directly addresses the assignment's focus on dimensionality

Potential Statistical Hypotheses:
H0: The average overall rating of players in top tier clubs is equal to that of players in lower tier clubs.
Ha: The average overall rating of players in top tier clubs is not equal to that of players in lower tier clubs.

H0: There is no significant difference in average wage across different player positions (e.g., GK, DEF, MID, FWD).
Ha: At least one player position has a different average wage than the others.

H0: Preferred foot (left vs. right) is independent of player position.
Ha: Preferred foot is dependent on player position.