**Final project proposal**

**Jiyoon (Clover) Jeong - 3032235220, Jin Kweon - 3032235207**

**Data:**

We are inspecting soccer data from Fifa 2017. We downloaded data from “<https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global>.”

The data has 17588 rows (number of players) and 53 columns (name, nationality, rating, height, etc). The data has some missing values. We created a data dictionary to help the audience understand the data and analysis. We have 9 categorical and 44 continuous variables in our data.

**Problems to address (research question):**

\*It is important how we sample the data. Players are selected to play in the national team if they perform well in the club. It is true nowadays in the soccer world, players spend more time playing for the club. So, we are focusing on inspecting players’ club profiles only.

1. We want to know how the ratings are different based on club positions (goalkeeper, defense, midfielder, attacker) and preferred foot of the players.

Hypothesis: The means for each group of position-effect, preferred-foot-effect, and interaction are the same.

2. We are going to find out what quantitative variables are related with the rating.

Hypothesis: There should exist some quantitative variables that have a linear relationship with rating. (multinomial)

3. I want to test how much each variable is correlated with the rating. PCA will also help me find co-linearity issue.

Hypothesis: Different positions have different variables correlated with the rating, and I should find the skills that are important to the position should have high correlation with the rating.

**Major approaches:**

1. Multivariate comparison test: We do the Manova test to answer the first research question. Our data has 28 positions (refer to the data dictionary), and we are going to manually re-categorize positions into four: goalkeeper, defense, midfielder, and attacker.

We have several different populations groups and two factors (four levels for positions and two levels for preferred foot).

2.Linear Regression : We are going to analyze both visual and test-statistics.

3.PCA(Principal Component Analysis) : Each position has different variables they are strong at, so we are going to test out five PCA: a) all the positions on PCs, b) goalkeeper on PCs, c) defense on PCs, d) midfielder on PCs, and e) attacker on Pcs.

**Impact/significance of the research:**

By going through these three research questions (three tests), we aim to find out how the FIFA ratings on the players were decided.