Project Report

# GitHub URL

[*https://github.com/DarranMcNamee/UCDPA\_Darran-McNamee.git*](https://github.com/DarranMcNamee/UCDPA_Darran-McNamee.git)

# Abstract

EA Sports FIFA is a series of soccer video games developed and released annually by EA Sports. The FIFA franchise has been localized into 18 languages and available in 51 countries. Listed in the Guinness World Records as the best-selling sport video game franchise in the world, having sold over 325 million copies as of 2021.

Each year the game uploads the real-life players & teams statistical data to the games database. Given the availability of this data my intention is to put into practice the new skills I have learned from this course into a real-life database and aim to analyze the players positions.

On 10 May 2022, it was announced that EA and FIIFA's partnership is set to come to an end after 30 years from July 2023 onwards; the series will be retitled **EA Sports FC**. FIFA intends to enter a partnership with a new developer to produce “*the real game that has the FIFA name*” which would make the review of the comparable datasets very interesting for the years ahead.

# Introduction

As a father to two boys who play FIFA I was interested to review the data associated with game and share my new data analysis skills. I noticed when they are playing this game that regularly they would argue about the players defined positioning and I was interested to see if I could redefine this into a smaller subset of player positions. While acknowledging that the data set is huge and available from 2015 to date, I choose the most recent years of 2021 & 2022 for review.

# Dataset

* The datasets chosen were players data for the Career Mode from FIFA 15 to FIFA 22.
* This source is an accurate dataset based on the hugely popular EA Games FIFA Videogame.
* The data allows multiple comparisons for the same players across the last 8 version of the videogame.
* I choose FIFA 2021 & 2022 for my data set for python analysis.
* The data files were downloaded from a Kaggle data set expert user, Stefano Leone, based in Dublin.
* <https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset>

# Implementation Process

I utilised an API for downloaded the selected datasets from Kaggle to my hard drive via od.download ("*https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset*")

By using the API function by inputting my Kaggle credentials and Kaggle key token I downloaded the required files. I used this method as I had real difficulties in using the Github API due to system restrictions on my hardware.

The API downloaded the full dataset (2014-2022 player data files), I choose to use the two most recent years data (2021 & 2022) and uploaded the relevant datafiles to my Jupyter Notebook.

1. Uploaded the 2021 (18,944 lines) & 2022 (19,239 lines) datafiles.
2. For the preparation of the data, I then sorted and merged the 2021 & 2022 datasets which gave a combined total of 38,183 lines. The merging of these two datasets was carried out *via “fifa\_complete\_data = pd.DataFrame.merge(fifa22, fifa21, how="outer")”.*
3. Given this would like contain duplicates given the large amount of data (*38,183,110*) I ran a query to find 14,895 duplicate players thereby cleansing my dataset.
4. In order to concentrate my dataset and for my analysis for the top players, I removed all players with a skillset of less than 80 to show the Top Players (537). The reason for this selection of higher-level skill set was for my target audience and their knowledge of the most famous players in the game.
5. I then redefined via reusable code & groupby/looping the player positions. I utilized the regex (regular expression) for the search and replace operation for player positions with multiple positions. While also splitting the data for players with more than one position by taking the primary player position as outlined in the data to make the data accurate and again avoid duplication of positioning.
6. I had to utilizes google search functions to avail of a duplicate id solution due to errors I was receiving with my scripts in relation to index columns.
7. Carried out validation of the data/checkpoint to ensure no data was lost/misplaced with above steps.
8. Sorted and creating a dictionary/list for the mapping of the new player positions “*player\_position\_fifa”* for revised positions. I recategorized the current & extensive player positions of 20 into 4 positions to make it easier for my target audience to know where each player’s best position was on the field.
9. Ran a check to ensure no null values were present in the dataset.
10. Ran new player code check to confirm revised and reduced player positions resulting in the following categorization:

|  |  |
| --- | --- |
| OLD | NEW |
| A picture containing text  Description automatically generated | Graphical user interface, text, application  Description automatically generated |

1. Next step I wished to perform a graphic visualization of my results, by utilizing matplotlib & seaborne to feature a graphic display of the revised player position together with a count of the players in each category.
2. Using SNS, I wanted to them display a visualization of skill rating, player value, wages, movement reaction, mentality composure, skill on the ball control, dribbling and attacking short passing. These results are displayed via histogram graphics.
3. To redefine the datasets used so far, I wanted to show that I could recode the data based on the players laterality (*left or right footed*). With the data contained the preferred foot of each player, I sought to recode for each player and use Pearsons Correlation interpretation to measure the strength of the linear relationship between the two variables.

# Results

1. From refining the player positions to the new defined parameters and using the following visualization. This allowed for user to easily see that Midfield players were the most common player in the game with defender, forward and goalkeeper coming next.

Chart, bar chart

Description automatically generated

1. To review & visual player performance in terms of data available. The following graphic displays the largest number of players (>200) having a skill rating of 80. The player wage shows how most players are receiving between €50k & €100k per week.

|  |  |
| --- | --- |
| Chart, histogram  Description automatically generated |  |

# Insights

1. The EA game program has 15 player positions which can be easily broken down in 4 more regular position.

|  |  |
| --- | --- |
| OLD | NEW |
| A picture containing text  Description automatically generated | Graphical user interface, text, application  Description automatically generated |

1. By recategorizing the positions, I was able to fully rewrite the player positions based on the revised parameters.
2. This allowed for a revised analysis of the data and grouping into the following revised positions.

Graphical user interface, text, application

Description automatically generated

1. The results above show that most players in this data are Midfielders at 193, followed by defenders at 152, forwards at 119 and goalkeepers at 73. Its interesting in that the above data is for players with a skill rating >80 which displays that the midfielder is the most skillful category within the game.
2. By using the Pearson Correlation interpretation to correlate the benefits of left or right footed players versus their monetary value, the results show the players in the output had a moderate movement reaction of 0.56 and moderate wage level of 0.65. However, the players have a correlation coefficient potential of 0.85 & release clause of 0.99.
3. I utilized classification visualization to display the top ranking & most expensive skilled players to their linear relationship to movement, mentality, attacking/short passing, skill on the ball & wages as shown below:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |

Machine Learning

1. I have used classification methods of supervised learning given I have the exact data points for each player and am classifying the outputs based on my requirements.
2. For future data sets and for developing the model, I would recommend adding in the 2023 data set into the training for testing. I would split the data set into two subgroups (*training and testing*) to build and develop and tune the parameters.
3. Next, I would take the generated model to the tested data and run the model.
4. The model could predict the future player skill levels bases on the model parameters.
5. This could be used by EA Sports and the new FIFA game to correlate the players overall dependent and independent variables for end user experience.

# References

* *Stefano Leone, Kaggle Data Set.*