Team Wrigley Field
DSC423 Project

Group Members:
Amanuel Petros
David Guo
Jilei Hao
Kaushik Bandaru

The combined dataset was originally divided into 5 different datasets. The “epl player data.csv” contains 508 rows and 33 columns of player data with information about a players statistics over the course of the 18-19 Premier League season, including information about their appearances and goals scored. The “fpl data.csv” contained 423 rows and 4 columns about a player’s fantasy points accumulated over the 18-19 Premier League season. The “epl goalkeeper.csv” data contains 37 rows and 21 columns about the goalkeepers who played in the Premier League in 18-19 and information about their performance such as goals conceded, and save percentage. The “playerHistory.csv” data contained 685 rows and 43 columns about a player’s historical performance in the Premier League, including information on a player’s appearances and wins. The “transferData.csv” contained 603 rows and 2 columns on a player’s transfer value.

For the “goalkeeper.csv” data, Amanuel cleaned player names using SPSS and converted the file to csv. He also got rid of the accents in the names, so “Martin DÃºbravka\Martin-Dubravka” was changed to Dubravka Martin. He changed the nationality column to be only in three letters (SVK, BRA) when it was previously both in two and three letter names (sk SVK, br BRA). There weren’t any missing data values on the dataset.

For the “epl player data.csv”, David removed the accents on some of the names using a Python script, so the names can be more consistent from file to file. The data had 2 and 3 letter country codes (the ISO code and the FIFA country code) for the nationality of players in the same cell separated by a space (eg a French player had a fr and a FRA in the nation column). He split those two words up as well and used VLOOKUP to get the full country names of the players. Also, if the player had played multiple positions, the player’s “pos” column data contained multiple positions (a player who had played as a defender and midfielder had DFMF), so he separated the data into three different columns with the LEFT and MIDDLE functions. There weren’t any missing data values on this dataset.

Jilei wrote a java program that converts the historical Premier League data (“fpl_data_2018_2019.json”) from the original json format to tabular format “playerhistory.csv”. He stripped accents from players’ names using Apache Common Language library. There was some missing data as players playing certain positions only has a subset of variables populated. For example, goalkeepers don’t have any attacking variables populated. This type of missing values are expected due to differences in player roles.

Kaushik cleaned the data on “transferData.csv” and “fpl data.csv” dataset using Excel. He also removed the accents of the names and converted them to English letters using a VBA Script. There weren’t any missing data values on the two datasets.

Before merging the data, David removed some variables in the goalkeeper data (nationality, position) that were redundant with the EPL player data. He used R to merge the datasets together. He left outer joined “epl player data.csv” with the “goalkeeper.csv” data, and then joined the resulting data with Kaushik’s data. Then he inner joined the resulting data with the historical Premier League data (“playerhistory.csv”).

After processing the dataset, we were left with 413 rows and 95 columns of player information. To further distinguish between the different variables, we added a 90 to the end of a variable if the statistic looked at player information per 90 minutes, and an 18 to the end to the variable if the statistic looked at player information in the 18-19 season, so a player’s goals per 90 in the 18-19 season would be “Gls90.18”. We also deleted some redundant features like the club variable, and so we were left with 90 columns of data.

Appendix:
Variable Dictionary:

Full Name	Categorical	The first and last name of the player
Country code (2 letters)	Categorical	The player’s nationality, either what country the player has citizenship in or what national team the player represents in FIFA (the latter takes priority over the former) This column has the 2 letter country code supplied by iso with the home nations (England, Scotland, Wales, and Northern Ireland) have 3 letter names
Country code (3 letters)	Categorical	3 Letter fifa country code
Country	Categorical	the full country name
Pos1	Categorical	What positions the players have played in GK - Goalkeeper DF - Defender MF - Midfielder FW - Forward FB - Fullback CD - Central Defender DM - Defensive Midfielder CM - Central Midfielder WM - Wide Midfielder AM - Attacking Midfielder

Pos2	Categorical	
Pos3	Categorical	Some players have played in multiple positions, so there are three columns for the positions
Squad	Categorical	The club the player last played for in the 18-19 season
Age	Numeric	Age at the season start (August 1) for league play
Born	Numeric	Birth Year
Apps18	Numeric	Appearances Number of matches the player has appeared in
Starts18	Numeric	Games started by player
Subs18	Numeric	Substitutions: Games the player played in, but did not start
Min18	Numeric	Minutes Played
min/app18	Numeric	Minutes per appearance
gls18	numeric	Goals scored
ast18	numeric	assists
pk18	numeric	Penalty kicks
pkat18	numeric	Penalty kicks attempted
Fls18	numeric	fouls
CrdY18	numeric	Yellow cards
CrdR18	numeric	Red cards
SoT18	numeric	Shots on target
Gls90.18	numeric	Goals per 90 minutes
G+A90.18	numeric	Goals plus assists per 90 minutes
G-PK90.18	numeric	Goals minus penalty goals per 90 minutes
G+A-PK90.18	numeric	Goals plus assists minus penalty goals per 90 minutes
SoT90.18	numeric	Shots on target per 90 minutes
Fls90.18	numeric	Fouls per 90 minutes
Crd90.18	numeric	Cards per 90 minutes This double counts red cards given if a player gets multiple yellow cards in a match

GA18	Numeric	Goals Against
GA90.18	Numeric	Goals Against per 90 minutes
SoTA90	Numeric	Shots on target
Save18	Numeric	(Shots on target against- goals against)/shots on target against
W18	Numeric	wins
D18	Numeric	draws
L18	Numeric	losses
CS_A18	Numeric	Clean sheets against
CS18	Numeric	Percentage of matches that result in clean sheets
Position	Categorical	General Info: Player’s positions: Goalkeeper, Defender, Midfielder, Forward
Club	Categorical	General Info: Player’s latest club
Appearances	Numeric	General Info: Total match appeared
Wins	Numeric	General Info: Total wins
Losses	Numeric	General Info: Total losses
Saves	Numeric	Goalkeeper-specific variable
Penalties Saved	Numeric	Goalkeeper-specific variable
Punches	Numeric	Goalkeeper-specific variable
High Claims	Numeric	Goalkeeper-specific variable
Caches	Numeric	Goalkeeper-specific variable
Sweeper Clearance	Numeric	Goalkeeper-specific variable
Goals Conceded	Numeric	Defence variable
Clean Sheets	Numeric	Defence variable
Tackles	Numeric	Defence variable
Tackle Success	Numeric	Defence variable
Last Man Tackles	Numeric	Defence variable

Blocked Shots	Numeric	Defence variable
Interceptions	Numeric	Defence variable
Clearances	Numeric	Defence variable
Headed Clearance	Numeric	Defence variable
Clearances Off Line	Numeric	Defence variable
Recoveries	Numeric	Defence variable
Duels Won	Numeric	Defence variable
Duels Lost	Numeric	Defence variable
Successful 50/50s	Numeric	Defence variable
Aerial Battles Won	Numeric	Defence variable
Aerial Battles Lost	Numeric	Defence variable
Own Goals	Numeric	Defence variable
Errors Leading to Goal	Numeric	Defence variable
Assists	Numeric	Team-play variable
Passes	Numeric	Team-play variable
Passes Per Match	Numeric	Team-play variable
Big Chances Created	Numeric	Team-play variable
Crosses	Numeric	Team-play variable
Cross Accuracy	Numeric	Team-play variable
Through Balls	Numeric	Team-play variable
Accurate Long Balls	Numeric	Team-play variable
Yellow Cards	Numeric	Discipline variable
Red Cards	Numeric	Discipline variable

Fouls	Numeric	Discipline variable
Offsides	Numeric	Discipline variable
Goals	Numeric	Attack variable
Goals per Match	Numeric	Attack variable
Headed Goals	Numeric	Attack variable
Goals with Right Foot	Numeric	Attack variable
Goals with Left Foot	Numeric	Attack variable
Hit Woodwork	Numeric	Attack variable
Big Chances Missed	Numeric	Attack variable
Value	Numeric	Current transfer Value of the player on FIFA in dollars
Cost	Numeric	Cost of the player the Fantasy Premier League in pounds
Points	Numeric	Number of Fantasy points scored in the 18-19 Premier League Season