CS36 Final Project - Mustafa Hourani

Load packages

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.4     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(modelr)
```

Imports the Fifa 22 player dataset and stores it into a variable

```
fifa22 <- read.csv("players_22.csv")
#fifa22
```

Imports the Fifa 21 player dataset and stores it into a variable

```
fifa21 <- read.csv("players_21.csv")
#fifa21
```

Joins. Progressive joins for each new variable will slowly build up our data tidily.
Creates 2 tibbles, 1 for each Fifa containing id (key) and player name
``r trial_name_2022 <- fifa22 %>% select(sofifa_id, short_name) #trial_name_2022
trial_name_2021 <- fifa21 %>% select(sofifa_id, short_name) #trial_name_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 pla
`r table_name <- inner_join(trial_name_2022, trial_name_2021, by = "sofifa_id") %>% rename("2022" = short_name.x, "2021" = short_name.`
``r trial_dob_2022 <- fifa22 %>% select(sofifa_id, dob) #trial_dob_2022
trial_dob_2021 <- fifa21 %>% select(sofifa_id, dob) #trial_dob_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 dob as 2022 a
`r table_dob <- inner_join(trial_dob_2022, trial_dob_2021, by = "sofifa_id")  %>% rename("2022" = dob.x, "2021" = dob.y) %>% pivot_lon`
`r final <- inner_join(table_name, table_dob, by = c("sofifa_id", "Year")) #final`
Creates 2 tibbles, 1 for each Fifa containing id (key) and nationality
``r trial_nation_2022 <- fifa22 %>% select(sofifa_id, nationality_name) #trial_nation_2022
trial_nation_2021 <- fifa21 %>% select(sofifa_id, nationality_name) #trial_nation_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa
`r table_nation <- inner_join(trial_nation_2022, trial_nation_2021, by = "sofifa_id")  %>% rename("2022" = nationality_name.x, "2021"`
`r final <- inner_join(final, table_nation, by = c("sofifa_id", "Year")) #final`
Creates 2 tibbles, 1 for each Fifa containing id (key) and preferred foot
``r trial_foot_2022 <- fifa22 %>% select(sofifa_id, preferred_foot) #trial_foot_2022
trial_foot_2021 <- fifa21 %>% select(sofifa_id, preferred_foot) #trial_foot_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 pref
`r table_foot <- inner_join(trial_foot_2022, trial_foot_2021, by = "sofifa_id")  %>% rename("2022" = preferred_foot.x, "2021" = prefer`
`r final <- inner_join(final, table_foot, by = c("sofifa_id", "Year")) #final`
Creates 2 tibbles, 1 for each Fifa containing id (key) and age
``r trial_age_2022 <- fifa22 %>% select(sofifa_id, age) #trial_age_2022
trial_age_2021 <- fifa21 %>% select(sofifa_id, age) #trial_age_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 age as 2022 ar
`r table_age <- inner_join(trial_age_2022, trial_age_2021, by = "sofifa_id") %>% rename("2022" = age.x, "2021" = age.y) %>% pivot_long`
`r final <- inner_join(final, table_age, by = c("sofifa_id", "Year")) #final`
Creates 2 tibbles, 1 for each Fifa containing id (key) and player position
``r trial_positions_2022 <- fifa22 %>% select(sofifa_id, player_positions) #trial_positions_2022

trial_positions_2021 <- fifa21 %>% select(sofifa_id, player_positions) #trial_positions_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames th

```
r table_positions <- inner_join(trial_positions_2022, trial_positions_2021, by = "sofifa_id") %>% rename("2022" = player_positions.x,
```

```
r final <- inner_join(final, table_positions, by = c("sofifa_id", "Year")) #final
```

Creates 2 tibbles, 1 for each Fifa containing id (key) and value

```r trial_value_2022 <- fifa22 %>% select(sofifa_id, value_eur) #trial_value_2022

trial_value_2021 <- fifa21 %>% select(sofifa_id, value_eur) #trial_value_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 player

```
r table_value <- inner_join(trial_value_2022, trial_value_2021, by = "sofifa_id") %>% rename("2022" = value_eur.x, "2021" = value_eur
```

```
r final <- inner_join(final, table_value, by = c("sofifa_id", "Year")) #final
~~~

Creates 2 tibbles, 1 for each Fifa containing id (key) and weight

```r trial_weight_2022 <- fifa22 %>% select(sofifa_id, weight_kg) #trial_weight_2022

trial_weight_2021 <- fifa21 %>% select(sofifa_id, weight_kg) #trial_weight_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 we

~~~
r table_weight <- inner_join(trial_weight_2022, trial_weight_2021, by = "sofifa_id") %>% rename("2022" = weight_kg.x, "2021" = weight
```

```
r final <- inner_join(final, table_weight, by = c("sofifa_id", "Year")) #final
~~~

Creates 2 tibbles, 1 for each Fifa containing id (key) and height

```r trial_height_2022 <- fifa22 %>% select(sofifa_id, height_cm) #trial_height_2022

trial_height_2021 <- fifa21 %>% select(sofifa_id, height_cm) #trial_height_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 hei

~~~
r table_height <- inner_join(trial_height_2022, trial_height_2021, by = "sofifa_id") %>% rename("2022" = height_cm.x, "2021" = height
```

```
r final <- inner_join(final, table_height, by = c("sofifa_id", "Year")) #final
~~~

Creates 2 tibbles, 1 for each Fifa containing id (key) and league

```r trial_league_2022 <- fifa22 %>% select(sofifa_id, league_name) #trial_league_2022

trial_league_2021 <- fifa21 %>% select(sofifa_id, league_name) #trial_league_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22

~~~
r table_league <- inner_join(trial_league_2022, trial_league_2021, by = "sofifa_id") %>% rename("2022" = league_name.x, "2021" = leag
```

```
r final <- inner_join(final, table_league, by = c("sofifa_id", "Year")) #final
~~~

Creates 2 tibbles, 1 for each Fifa containing id (key) and club

```r trial_club_2022 <- fifa22 %>% select(sofifa_id, club_name) #trial_club_2022

trial_club_2021 <- fifa21 %>% select(sofifa_id, club_name) #trial_club_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 club as

~~~
r table_club <- inner_join(trial_club_2022, trial_club_2021, by = "sofifa_id") %>% rename("2022" = club_name.x, "2021" = club_name.y)
```

```
r final <- inner_join(final, table_club, by = c("sofifa_id", "Year")) #final
~~~

Creates 2 tibbles, 1 for each Fifa containing id (key) and date joined

```r trial_join_2022 <- fifa22 %>% select(sofifa_id, club_joined) #trial_join_2022

trial_join_2021 <- fifa21 %>% select(sofifa_id, club_joined) #trial_join_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 date joi

~~~
r table_join <- inner_join(trial_join_2022, trial_join_2021, by = "sofifa_id") %>% rename("2022" = club_joined.x, "2021" = club_joine
```

```
r final <- inner_join(final, table_join, by = c("sofifa_id", "Year")) #final
~~~

Creates 2 tibbles, 1 for each Fifa containing id (key) and overall rating

```r trial_overall_2022 <- fifa22 %>% select(sofifa_id, overall) #trial_overall_2022

trial_overall_2021 <- fifa21 %>% select(sofifa_id, overall) #trial_overall_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 overall

~~~
r table_overall <- inner_join(trial_overall_2022, trial_overall_2021, by = "sofifa_id") %>% rename("2022" = overall.x, "2021" = overa
```

```
r final <- inner_join(final, table_overall, by = c("sofifa_id", "Year")) #final
~~~

Creates 2 tibbles, 1 for each Fifa containing id (key) and weak foot rating

```r trial_weak_2022 <- fifa22 %>% select(sofifa_id, weak_foot) #trial_weak_2022

trial_weak_2021 <- fifa21 %>% select(sofifa_id, weak_foot) #trial_weak_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 weak

~~~
r table_weak <- inner_join(trial_weak_2022, trial_weak_2021, by = "sofifa_id") %>% rename("2022" = weak_foot.x, "2021" = weak_foot.y)
```

```
r final <- inner_join(final, table_weak, by = c("sofifa_id", "Year")) #final
~~~

Creates 2 tibbles, 1 for each Fifa containing id (key) and skill moves rating

```r trial_skill_2022 <- fifa22 %>% select(sofifa_id, skill_moves) #trial_skill_2022

trial_skill_2021 <- fifa21 %>% select(sofifa_id, skill_moves) #trial_skill_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 skill mo
~~~

```
r table_skill <- inner_join(trial_skill_2022, trial_skill_2021, by = "sofifa_id") %>% rename("2022" = skill_moves.x, "2021" = skill_m
```

```
r final <- inner_join(final, table_skill, by = c("sofifa_id", "Year")) #final
```

Creates 2 tibbles, 1 for each Fifa containing id (key) and pace rating

~~~
```r trial_pace_2022 <- fifa22 %>% select(sofifa_id, pace) #trial_pace_2022
~~~

~~~
trial_pace_2021 <- fifa21 %>% select(sofifa_id, pace) #trial_pace_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 pace rating
~~~

```
r table_pace <- inner_join(trial_pace_2022, trial_pace_2021, by = "sofifa_id") %>% rename("2022" = pace.x, "2021" = pace.y) %>% pivot
```

```
r final <- inner_join(final, table_pace, by = c("sofifa_id", "Year")) #final
```

Creates 2 tibbles, 1 for each Fifa containing id (key) and shooting rating

~~~
```r trial_shooting_2022 <- fifa22 %>% select(sofifa_id, shooting) #trial_shooting_2022
~~~

~~~
trial_shooting_2021 <- fifa21 %>% select(sofifa_id, shooting) #trial_shooting_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22
~~~

```
r table_shooting <- inner_join(trial_shooting_2022, trial_shooting_2021, by = "sofifa_id") %>% rename("2022" = shooting.x, "2021" = s
```

```
r final <- inner_join(final, table_shooting, by = c("sofifa_id", "Year")) #final
```

Creates 2 tibbles, 1 for each Fifa containing id (key) and passing rating

~~~
```r trial_passing_2022 <- fifa22 %>% select(sofifa_id, passing) #trial_passing_2022
~~~

~~~
trial_passing_2021 <- fifa21 %>% select(sofifa_id, passing) #trial_passing_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 pa
~~~

```
r table_passing <- inner_join(trial_passing_2022, trial_passing_2021, by = "sofifa_id") %>% rename("2022" = passing.x, "2021" = passi
```

```
r final <- inner_join(final, table_passing, by = c("sofifa_id", "Year")) #final
```

Creates 2 tibbles, 1 for each Fifa containing id (key) and dribbling rating

~~~
```r trial_dribbling_2022 <- fifa22 %>% select(sofifa_id, dribbling) #trial_dribbling_2022
~~~

~~~
trial_dribbling_2021 <- fifa21 %>% select(sofifa_id, dribbling) #trial_dribbling_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22
~~~

```
r table_dribbling <- inner_join(trial_dribbling_2022, trial_dribbling_2021, by = "sofifa_id") %>% rename("2022" = dribbling.x, "2021"
```

```
r final <- inner_join(final, table_dribbling, by = c("sofifa_id", "Year")) #final
```

Creates 2 tibbles, 1 for each Fifa containing id (key) and defending rating

~~~
```r trial_defending_2022 <- fifa22 %>% select(sofifa_id, defending) #trial_defending_2022
~~~

~~~
trial_defending_2021 <- fifa21 %>% select(sofifa_id, defending) #trial_defending_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa
~~~

```
r table_defending <- inner_join(trial_defending_2022, trial_defending_2021, by = "sofifa_id") %>% rename("2022" = defending.x, "2021"
```

```
r final <- inner_join(final, table_defending, by = c("sofifa_id", "Year")) #final
```

Creates 2 tibbles, 1 for each Fifa containing id (key) and physical rating

~~~
```r trial_physical_2022 <- fifa22 %>% select(sofifa_id, physic) #trial_physical_2022
~~~

~~~
trial_physical_2021 <- fifa21 %>% select(sofifa_id, physic) #trial_physical_2021 ``` Inner_Joins 2 previous tibbles based on the key (id) and renames the fifa 22 phy
~~~

```
r table_physical <- inner_join(trial_physical_2022, trial_physical_2021, by = "sofifa_id") %>% rename("2022" = physic.x, "2021" = phy
```

```
r final <- inner_join(final, table_physical, by = c("sofifa_id", "Year")) final
```

```
## # A tibble: 26,600 × 23 ##    sofifa_id Year  Name       `Date of Birth` Nationality `Preferred Foot`   Age ##        <int> <chr>
```

Cleaning

Converts value column from Euros to Dollars by using conversion factor of 1.1315

```
final <- final %>% mutate("value_eur" = value_eur * 1.1315) %>% rename("Value in $" = value_eur)
final
```

```
## # A tibble: 26,600 × 23
##    sofifa_id Year  Name       `Date of Birth` Nationality `Preferred Foot`   Age
##        <int> <chr> <chr>      <chr>           <chr>       <chr>            <int>
##  1    158023 2022  L. Messi   1987-06-24      Argentina   Left                34
##  2    158023 2021  L. Messi   1987-06-24      Argentina   Left                33
##  3    188545 2022  R. Lewand… 1988-08-21      Poland      Right               32
##  4    188545 2021  R. Lewand… 1988-08-21      Poland      Right               31
##  5     20801 2022  Cristiano… 1985-02-05      Portugal    Right               36
##  6     20801 2021  Cristiano… 1985-02-05      Portugal    Right               35
##  7    190871 2022  Neymar Jr  1992-02-05      Brazil      Right               29
##  8    190871 2021  Neymar Jr  1992-02-05      Brazil      Right               28
##  9    192985 2022  K. De Bru… 1991-06-28      Belgium     Right               30
## 10    192985 2021  K. De Bru… 1991-06-28      Belgium     Right               29
## # … with 26,590 more rows, and 16 more variables: player_positions <chr>,
## #   Value in $ <dbl>, Kilogram Weight <int>, Centimeter Height <int>,
## #   league_name <chr>, Club <chr>, Date Joined Club <chr>,
## #   Overall Rating <int>, Weak Foot Rating <int>, Skill Move Rating <int>,
## #   Pace Rating <int>, Shooting Rating <int>, Passing Rating <int>,
## #   Dribbling Rating <int>, Defending Rating <int>, Physical Rating <int>
```

Change League names that don't start with League Country to start with League Country in a single word and also ensure consistency in spelling among the nation's other leagues

```
final$league_name[final$league_name == "Campeonato Brasileiro Série A"] <- "Brazilian Campeonato Série A"
final$league_name[final$league_name == "Liga de Fútbol Profesional Boliviano"] <- "Bolivian Liga de Fútbol Profes
ional"
final$league_name[final$league_name == "Rep. Ireland Airtricity League"] <- "Irish Airtricity League"
final$league_name[final$league_name == "South African Premier Division"] <- "South-African Premier Division"
final$league_name[final$league_name == "Spain Primera Division"] <- "Spanish Primera Division"
```

Separates league country from league name in columns section

```
final <- final %>% separate(league_name, into = c("League Country", "League Name"), sep = " ", extra = "merge")
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 117 rows [293,
## 730, 738, 825, 1156, 1377, 1378, 1736, 1812, 1839, 1840, 2269, 2270, 2281, 2282,
## 2318, 2337, 2453, 2454, 2481, ...].
```

```
final
```

```
## # A tibble: 26,600 × 24
##    sofifa_id Year  Name       `Date of Birth` Nationality `Preferred Foot`   Age
##        <int> <chr> <chr>      <chr>           <chr>       <chr>            <int>
##  1    158023 2022  L. Messi   1987-06-24      Argentina   Left                34
##  2    158023 2021  L. Messi   1987-06-24      Argentina   Left                33
##  3    188545 2022  R. Lewand… 1988-08-21      Poland      Right               32
##  4    188545 2021  R. Lewand… 1988-08-21      Poland      Right               31
##  5     20801 2022  Cristiano… 1985-02-05      Portugal    Right               36
##  6     20801 2021  Cristiano… 1985-02-05      Portugal    Right               35
##  7    190871 2022  Neymar Jr  1992-02-05      Brazil      Right               29
##  8    190871 2021  Neymar Jr  1992-02-05      Brazil      Right               28
##  9    192985 2022  K. De Bru… 1991-06-28      Belgium     Right               30
## 10    192985 2021  K. De Bru… 1991-06-28      Belgium     Right               29
## # … with 26,590 more rows, and 17 more variables: player_positions <chr>,
## #   Value in $ <dbl>, Kilogram Weight <int>, Centimeter Height <int>,
## #   League Country <chr>, League Name <chr>, Club <chr>,
## #   Date Joined Club <chr>, Overall Rating <int>, Weak Foot Rating <int>,
## #   Skill Move Rating <int>, Pace Rating <int>, Shooting Rating <int>,
## #   Passing Rating <int>, Dribbling Rating <int>, Defending Rating <int>,
## #   Physical Rating <int>
```

Groups positions under 3 brackets (Defender / Midfielder / Attacker) based on real life classification of each position as either Defence/Midfield/Attack. ^ ensures we look at first appearance of key letters and (.)* let's us know anything can come after our key letters and it won't matter.

```
final <- final %>% rename("Position" = player_positions)

final$Position <- str_replace_all(final$Position, "^CB(.)*", "Defender")
final$Position <- str_replace_all(final$Position, "^LB(.)*", "Defender")
final$Position <- str_replace_all(final$Position, "^LWB(.)*", "Defender")
final$Position <- str_replace_all(final$Position, "^RB(.)*", "Defender")
final$Position <- str_replace_all(final$Position, "^RWB(.)*", "Defender")

final$Position <- str_replace_all(final$Position, "^CAM(.)*", "Midfielder")
final$Position <- str_replace_all(final$Position, "^CDM(.)*", "Midfielder")
final$Position <- str_replace_all(final$Position, "^CM(.)*", "Midfielder")
final$Position <- str_replace_all(final$Position, "^LM(.)*", "Midfielder")
final$Position <- str_replace_all(final$Position, "^RM(.)*", "Midfielder")

final$Position <- str_replace_all(final$Position, "^CF(.)*", "Attacker")
final$Position <- str_replace_all(final$Position, "^LW(.)*", "Attacker")
final$Position <- str_replace_all(final$Position, "^RW(.)*", "Attacker")
final$Position <- str_replace_all(final$Position, "^ST(.)*", "Attacker")

final
```

```
## # A tibble: 26,600 × 24
##    sofifa_id Year  Name       `Date of Birth` Nationality `Preferred Foot`   Age
##        <int> <chr> <chr>      <chr>           <chr>       <chr>            <int>
##  1    158023 2022  L. Messi   1987-06-24      Argentina   Left                34
##  2    158023 2021  L. Messi   1987-06-24      Argentina   Left                33
##  3    188545 2022  R. Lewand… 1988-08-21      Poland      Right               32
##  4    188545 2021  R. Lewand… 1988-08-21      Poland      Right               31
##  5     20801 2022  Cristiano… 1985-02-05      Portugal    Right               36
##  6     20801 2021  Cristiano… 1985-02-05      Portugal    Right               35
##  7    190871 2022  Neymar Jr  1992-02-05      Brazil      Right               29
##  8    190871 2021  Neymar Jr  1992-02-05      Brazil      Right               28
##  9    192985 2022  K. De Bru… 1991-06-28      Belgium     Right               30
## 10    192985 2021  K. De Bru… 1991-06-28      Belgium     Right               29
## # … with 26,590 more rows, and 17 more variables: Position <chr>,
## #   Value in $ <dbl>, Kilogram Weight <int>, Centimeter Height <int>,
## #   League Country <chr>, League Name <chr>, Club <chr>,
## #   Date Joined Club <chr>, Overall Rating <int>, Weak Foot Rating <int>,
## #   Skill Move Rating <int>, Pace Rating <int>, Shooting Rating <int>,
## #   Passing Rating <int>, Dribbling Rating <int>, Defending Rating <int>,
## #   Physical Rating <int>
```

Remove non-outfield players (goalkeepers)

```
final <- final %>% filter(Position != "GK")
final
```

```
## # A tibble: 23,731 × 24
##    sofifa_id Year  Name       `Date of Birth` Nationality `Preferred Foot`   Age
##        <int> <chr> <chr>      <chr>           <chr>       <chr>            <int>
##  1    158023 2022  L. Messi   1987-06-24      Argentina   Left                34
##  2    158023 2021  L. Messi   1987-06-24      Argentina   Left                33
##  3    188545 2022  R. Lewand… 1988-08-21      Poland      Right               32
##  4    188545 2021  R. Lewand… 1988-08-21      Poland      Right               31
##  5     20801 2022  Cristiano… 1985-02-05      Portugal    Right               36
##  6     20801 2021  Cristiano… 1985-02-05      Portugal    Right               35
##  7    190871 2022  Neymar Jr  1992-02-05      Brazil      Right               29
##  8    190871 2021  Neymar Jr  1992-02-05      Brazil      Right               28
##  9    192985 2022  K. De Bru… 1991-06-28      Belgium     Right               30
## 10    192985 2021  K. De Bru… 1991-06-28      Belgium     Right               29
## # … with 23,721 more rows, and 17 more variables: Position <chr>,
## #   Value in $ <dbl>, Kilogram Weight <int>, Centimeter Height <int>,
## #   League Country <chr>, League Name <chr>, Club <chr>,
## #   Date Joined Club <chr>, Overall Rating <int>, Weak Foot Rating <int>,
## #   Skill Move Rating <int>, Pace Rating <int>, Shooting Rating <int>,
## #   Passing Rating <int>, Dribbling Rating <int>, Defending Rating <int>,
## #   Physical Rating <int>
```

Parsing.

```
final$Year <- parse_integer(final$Year)
final$`Date of Birth` <- parse_date(final$`Date of Birth`, "%Y-%m-%d")
final$`Date Joined Club` <- parse_date(final$`Date Joined Club`, "%Y-%m-%d")
final
```

```
## # A tibble: 23,731 × 24
##    sofifa_id  Year Name       `Date of Birth` Nationality `Preferred Foot`   Age
##        <int> <int> <chr>      <date>          <chr>       <chr>            <int>
##  1    158023  2022 L. Messi   1987-06-24      Argentina   Left                34
##  2    158023  2021 L. Messi   1987-06-24      Argentina   Left                33
##  3    188545  2022 R. Lewand… 1988-08-21      Poland      Right               32
##  4    188545  2021 R. Lewand… 1988-08-21      Poland      Right               31
##  5     20801  2022 Cristiano… 1985-02-05      Portugal    Right               36
##  6     20801  2021 Cristiano… 1985-02-05      Portugal    Right               35
##  7    190871  2022 Neymar Jr  1992-02-05      Brazil      Right               29
##  8    190871  2021 Neymar Jr  1992-02-05      Brazil      Right               28
##  9    192985  2022 K. De Bru… 1991-06-28      Belgium     Right               30
## 10    192985  2021 K. De Bru… 1991-06-28      Belgium     Right               29
## # … with 23,721 more rows, and 17 more variables: Position <chr>,
## #   Value in $ <dbl>, Kilogram Weight <int>, Centimeter Height <int>,
## #   League Country <chr>, League Name <chr>, Club <chr>,
## #   Date Joined Club <date>, Overall Rating <int>, Weak Foot Rating <int>,
## #   Skill Move Rating <int>, Pace Rating <int>, Shooting Rating <int>,
## #   Passing Rating <int>, Dribbling Rating <int>, Defending Rating <int>,
## #   Physical Rating <int>
```

Additional column rename, rename to ID for simplicity

```
final <- final %>% rename("ID" = sofifa_id, )
final
```

```
## # A tibble: 23,731 × 24
##        ID  Year Name          `Date of Birth` Nationality `Preferred Foot`   Age
##     <int> <int> <chr>         <date>          <chr>       <chr>            <int>
##  1 158023  2022 L. Messi      1987-06-24      Argentina   Left                34
##  2 158023  2021 L. Messi      1987-06-24      Argentina   Left                33
##  3 188545  2022 R. Lewandows… 1988-08-21      Poland      Right               32
##  4 188545  2021 R. Lewandows… 1988-08-21      Poland      Right               31
##  5  20801  2022 Cristiano Ro… 1985-02-05      Portugal    Right               36
##  6  20801  2021 Cristiano Ro… 1985-02-05      Portugal    Right               35
##  7 190871  2022 Neymar Jr     1992-02-05      Brazil      Right               29
##  8 190871  2021 Neymar Jr     1992-02-05      Brazil      Right               28
##  9 192985  2022 K. De Bruyne  1991-06-28      Belgium     Right               30
## 10 192985  2021 K. De Bruyne  1991-06-28      Belgium     Right               29
## # … with 23,721 more rows, and 17 more variables: Position <chr>,
## #   Value in $ <dbl>, Kilogram Weight <int>, Centimeter Height <int>,
## #   League Country <chr>, League Name <chr>, Club <chr>,
## #   Date Joined Club <date>, Overall Rating <int>, Weak Foot Rating <int>,
## #   Skill Move Rating <int>, Pace Rating <int>, Shooting Rating <int>,
## #   Passing Rating <int>, Dribbling Rating <int>, Defending Rating <int>,
## #   Physical Rating <int>
```

QUESTION 1

After the pandemic (2021), Fifa (the footballing organization) decided to condense a normal football season into a shortened time frame which almost doubled the amount of matches professional footballers played every week. Many managers like Pep Guardiola were furious and said this negatively affected the physical condition of the players who were overworked.

Does the data set in Fifa reflect this real life theory? Is the average Physical Rating lower in Fifa 21 compared to Fifa 22? Was the average Overall Rating also lower in Fifa 21 as a result of a lack of practice? Were the players also slower during the condensed season?

Create a table for average of each of (Physical/Pace/Overall) ratings. Inner join these 3 average tables into a single table with all 3 using Year as the key. Then we pivot in order to tidy the data and create a new column for the Rating Type and another column for the Value. We do this since all these 3 averages are of the same type of objects.

```
avg_physical <- final %>% group_by(Year) %>% summarize("Average Physical" = mean(`Physical Rating`, na.rm = TRUE
))
avg_physical
```

```
## # A tibble: 2 × 2
##    Year `Average Physical`
##   <int>              <dbl>
## 1  2021               65.6
## 2  2022               67.0
```

```
avg_pace <- final %>% group_by(Year) %>% summarize("Average Pace" = mean(`Pace Rating`, na.rm = TRUE))
avg_pace
```

```
## # A tibble: 2 × 2
##    Year `Average Pace`
##   <int>          <dbl>
## 1  2021           68.5
## 2  2022           68.7
```

```
avg_overall <- final %>% group_by(Year) %>% summarize("Average Overall" = mean(`Overall Rating`, na.rm = TRUE))
avg_overall
```

```
## # A tibble: 2 × 2
##    Year `Average Overall`
##   <int>             <dbl>
## 1  2021              67.1
## 2  2022              68.1
```

```
avg_table <- inner_join(avg_physical, avg_pace, by = "Year") %>% inner_join(avg_overall, by = "Year")
avg_table
```

```
## # A tibble: 2 × 4
##    Year `Average Physical` `Average Pace` `Average Overall`
##   <int>              <dbl>          <dbl>             <dbl>
## 1  2021               65.6           68.5              67.1
## 2  2022               67.0           68.7              68.1
```

```
avg_table <- avg_table %>% pivot_longer(c("Average Physical", "Average Pace", "Average Overall"), names_to = "Rat
ing Type", values_to = "Value")
avg_table
```

```
## # A tibble: 6 × 3
##    Year `Rating Type`    Value
##   <int> <chr>            <dbl>
## 1  2021 Average Physical  65.6
## 2  2021 Average Pace      68.5
## 3  2021 Average Overall   67.1
## 4  2022 Average Physical  67.0
## 5  2022 Average Pace      68.7
## 6  2022 Average Overall   68.1
```

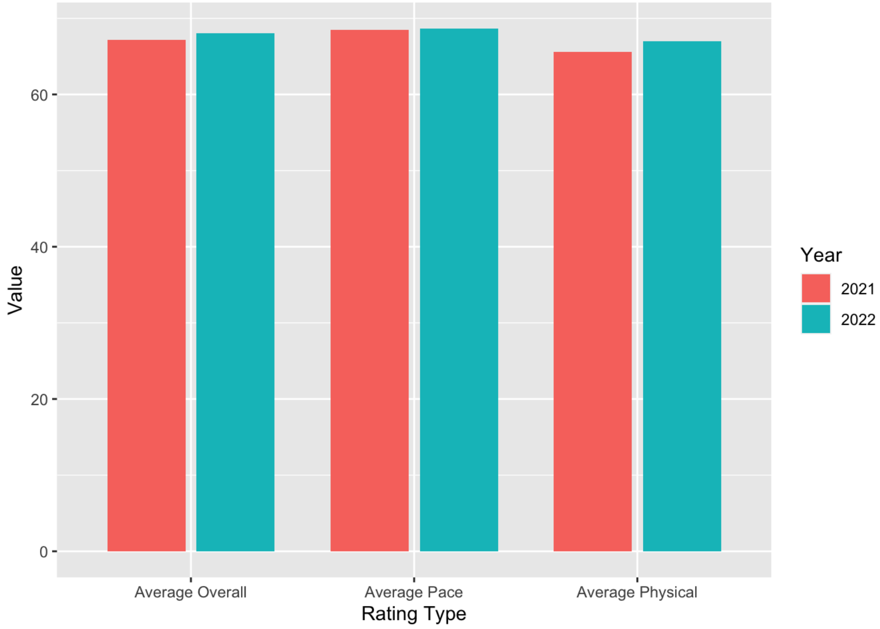
Change column type for year to categorical (char) in order to group bar chart next.

```
avg_table$Year <- as.character(avg_table$Year)
avg_table
```

```
## # A tibble: 6 × 3
##   Year  `Rating Type`    Value
##   <chr> <chr>            <dbl>
## 1 2021  Average Physical  65.6
## 2 2021  Average Pace      68.5
## 3 2021  Average Overall   67.1
## 4 2022  Average Physical  67.0
## 5 2022  Average Pace      68.7
## 6 2022  Average Overall   68.1
```

Visualize bar chart, x represents Rating type and y represents Value, we use Year as a fill in order to create a grouped bar chart.

```
ggplot(data = avg_table, mapping = aes(x = `Rating Type`, y = Value, fill = Year)) +
  geom_bar(stat = "identity", position = position_dodge(width=0.8), width = 0.7)
```



QUESTION 2

There is a theory among soccer fans that the Italian league is more defensive minded than other leagues. Is this theory reflected in the Fifa 22 dataset?

Restrict to Europe’s top 5 leagues to compare defensive stats of Italian league to similar levelled leagues. The reason we specify Seria A and Italian is because there is an Ecuadorian Seria A which we don’t care about. The reason we specify Premier League and English is because there are also non-Enlgish Premier League which we don’t care about.

```
top5 <- final %>%
  filter(`League Name` == "1. Bundesliga" | (`League Name` == "Serie A" & `League Country` == "Italian") | (`League Name` == "Premier League" & `League Country` == "English") | `League Name` == "Ligue 1" | `League Name` == "Primera Division")
top5
```

```
## # A tibble: 4,612 × 24
##        ID  Year Name          `Date of Birth` Nationality `Preferred Foot`   Age
##     <int> <int> <chr>         <date>          <chr>       <chr>            <int>
##  1 158023  2022 L. Messi      1987-06-24      Argentina   Left                34
##  2 158023  2021 L. Messi      1987-06-24      Argentina   Left                33
##  3 188545  2022 R. Lewandows… 1988-08-21      Poland      Right               32
##  4 188545  2021 R. Lewandows… 1988-08-21      Poland      Right               31
##  5  20801  2022 Cristiano Ro… 1985-02-05      Portugal    Right               36
##  6  20801  2021 Cristiano Ro… 1985-02-05      Portugal    Right               35
##  7 190871  2022 Neymar Jr     1992-02-05      Brazil      Right               29
##  8 190871  2021 Neymar Jr     1992-02-05      Brazil      Right               28
##  9 192985  2022 K. De Bruyne  1991-06-28      Belgium     Right               30
## 10 192985  2021 K. De Bruyne  1991-06-28      Belgium     Right               29
## # … with 4,602 more rows, and 17 more variables: Position <chr>,
## #   Value in $ <dbl>, Kilogram Weight <int>, Centimeter Height <int>,
## #   League Country <chr>, League Name <chr>, Club <chr>,
## #   Date Joined Club <date>, Overall Rating <int>, Weak Foot Rating <int>,
## #   Skill Move Rating <int>, Pace Rating <int>, Shooting Rating <int>,
## #   Passing Rating <int>, Dribbling Rating <int>, Defending Rating <int>,
## #   Physical Rating <int>
```

Create a model for all players in top 5 leagues

```
general_model <- lm(`Defending Rating` ~ `League Country`, data = top5)
general_model
```

```
## 
## Call:
## lm(formula = `Defending Rating` ~ `League Country`, data = top5)
## 
## Coefficients:
##            (Intercept)   `League Country`French   `League Country`German  
##                59.9225                  -3.8532                  -2.3012  
## `League Country`Italian  `League Country`Spanish  
##                 0.1724                  -1.1756
```

Creates a data grid with distinct league country options

```
sim1 <- top5 %>% data_grid(`League Country`)
sim1
```

```
## # A tibble: 5 × 1
##   `League Country`
##   <chr>           
## 1 English         
## 2 French          
## 3 German          
## 4 Italian         
## 5 Spanish
```

Data_grid creates a tibble with only distinct leagues which we then add our model predictions to for each league.

```
grid_pred1 <- sim1 %>%
  data_grid(`League Country`) %>%
  add_predictions(general_model) %>%
  rename("General Predicted Defensive Rating" = pred)
grid_pred1
```

```
## # A tibble: 5 × 2
##   `League Country` `General Predicted Defensive Rating`
##   <chr>                                           <dbl>
## 1 English                                          59.9
## 2 French                                           56.1
## 3 German                                           57.6
## 4 Italian                                          60.1
## 5 Spanish                                          58.7
```

Creates a tibble with every possible combination of League Country and Position.

```
top5_grouped <- top5 %>% group_by(`League Country`, Position) %>% summarise(`Defending Rating`)
```

```
## `summarise()` has grouped output by 'League Country', 'Position'. You can override using the `.groups` argumen
t.
```

```
top5_grouped
```

```
## # A tibble: 4,612 × 3
## # Groups:   League Country, Position [15]
##    `League Country` Position `Defending Rating`
##    <chr>            <chr>                 <int>
##  1 English          Attacker                 34
##  2 English          Attacker                 47
##  3 English          Attacker                 47
##  4 English          Attacker                 44
##  5 English          Attacker                 44
##  6 English          Attacker                 45
##  7 English          Attacker                 45
##  8 English          Attacker                 39
##  9 English          Attacker                 45
## 10 English          Attacker                 45
## # … with 4,602 more rows
```

Model for defenders in top 5 leagues, we are considering two x variables without interactions

```
grouped_model <- lm(`Defending Rating` ~ `League Country` + Position, data = top5_grouped)
grouped_model
```

```
## 
## Call:
## lm(formula = `Defending Rating` ~ `League Country` + Position, 
##     data = top5_grouped)
## 
## Coefficients:
##            (Intercept)   `League Country`French   `League Country`German  
##                35.6787                  -2.9091                  -1.8768  
## `League Country`Italian  `League Country`Spanish         PositionDefender  
##                 1.5953                   0.2922                  37.0818  
##      PositionMidfielder  
##                22.9919
```

Creates a data grid with distinct league country options

```
sim2 <- top5_grouped %>% data_grid(`League Country`)
sim2
```

```
## # A tibble: 15 × 2
## # Groups:   League Country, Position [15]
##    Position   `League Country`
##    <chr>      <chr>           
##  1 Attacker   English         
##  2 Defender   English         
##  3 Midfielder English         
##  4 Attacker   French          
##  5 Defender   French          
##  6 Midfielder French          
##  7 Attacker   German          
##  8 Defender   German          
##  9 Midfielder German          
## 10 Attacker   Italian         
## 11 Defender   Italian         
## 12 Midfielder Italian         
## 13 Attacker   Spanish         
## 14 Defender   Spanish         
## 15 Midfielder Spanish
```

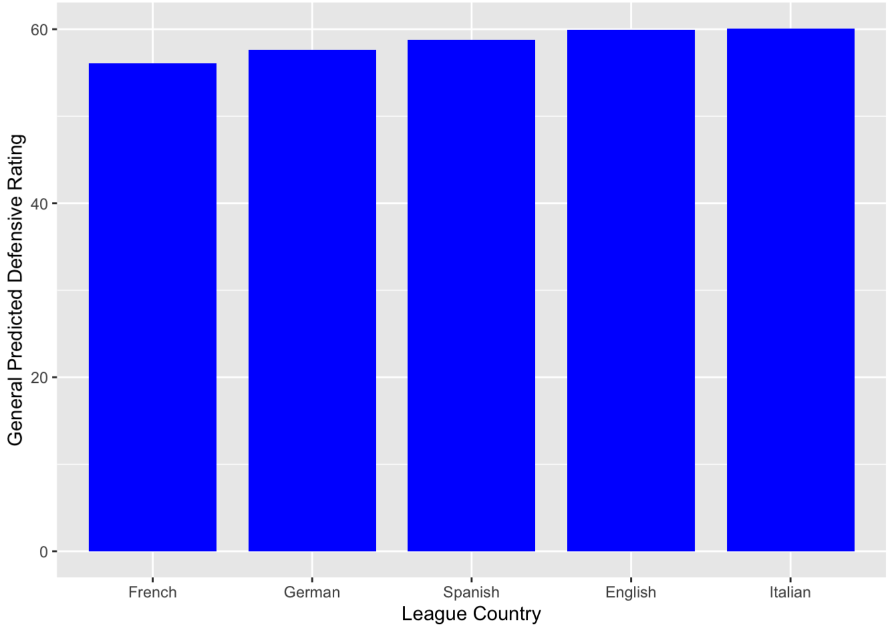
Data_grid creates a tibble with only distinct league and position combinations which we then add our model predictions to for each league/position combination. We then filter for only defensive players since we want to see how defensive players in Italian league compare to defensive players in other top 5 leagues.

```
grid_pred2 <- top5_grouped %>%
  data_grid(`League Country`) %>%
  add_predictions(grouped_model) %>%
  rename("Defender Predicted Defensive Rating" = pred) %>%
  filter(Position == "Defender")
grid_pred2
```

```
## # A tibble: 5 × 3
## # Groups:   League Country, Position [5]
##   Position `League Country` `Defender Predicted Defensive Rating`
##   <chr>    <chr>                                            <dbl>
## 1 Defender English                                           72.8
## 2 Defender French                                            69.9
## 3 Defender German                                            70.9
## 4 Defender Italian                                           74.4
## 5 Defender Spanish                                           73.1
```

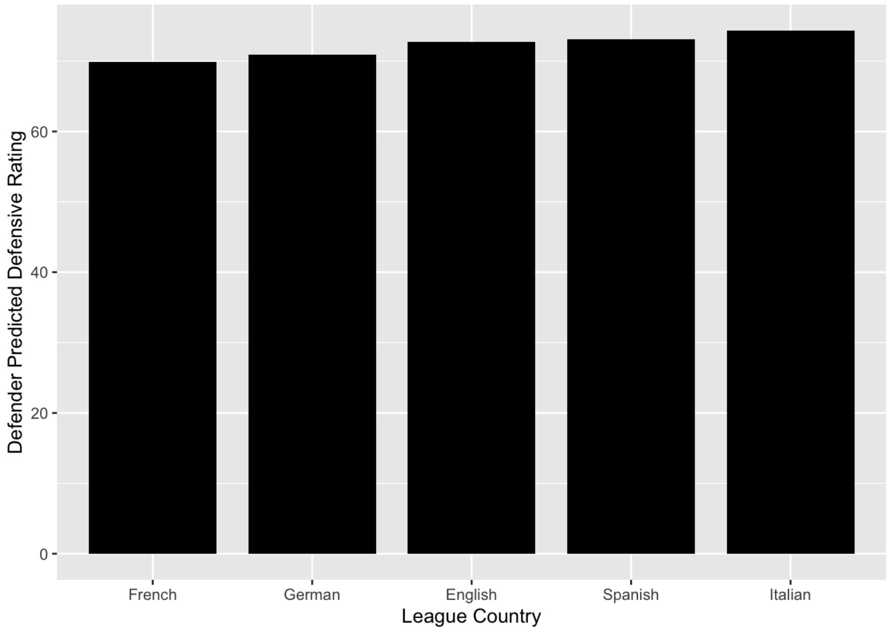
Visualize bar chart for general players. We reorder our x (League Country) bars in ascending order of their General Predicted Defensive Ratings

```
ggplot(data = grid_pred1, mapping = aes(x = fct_reorder(`League Country`, `General Predicted Defensive Rating`),
y = `General Predicted Defensive Rating`)) +
  geom_bar(stat = "identity", fill = "blue", width = 0.8) +
  xlab("League Country")
```



Visualize bar chart for defenders We reorder our x (League Country) bars in ascending order of their Defender Predicted Defensive Ratings

```
ggplot(data = grid_pred2, mapping = aes(x = fct_reorder(`League Country`, `Defender Predicted Defensive Rating`),
y = `Defender Predicted Defensive Rating`)) +
  geom_bar(stat = "identity", fill = "black", width = 0.8) +
  xlab("League Country")
```



QUESTION 3

A widely held theory in soccer is that shorter players have an advantage when it comes to ball control, as they tend to have a lower center of gravity and can therefore control the ball with their feet easier. Is this theory reflected in the Fifa dataset?

Create a model for dribbling rating in terms of height

```
control_model <- lm(`Dribbling Rating` ~ `Centimeter Height`, data = final)
control_model
```

```
##
## Call:
## lm(formula = `Dribbling Rating` ~ `Centimeter Height`, data = final)
##
## Coefficients:
##         (Intercept)  `Centimeter Height`
##            171.5079              -0.5944
```

Creates a data grid with all distinct values for height

```
grid_table <- final %>% data_grid(`Centimeter Height`)
grid_table
```

```
## # A tibble: 45 × 1
##    `Centimeter Height`
##                  <int>
##  1                 156
##  2                 158
##  3                 159
##  4                 160
##  5                 161
##  6                 162
##  7                 163
##  8                 164
##  9                 165
## 10                 166
## # … with 35 more rows
```

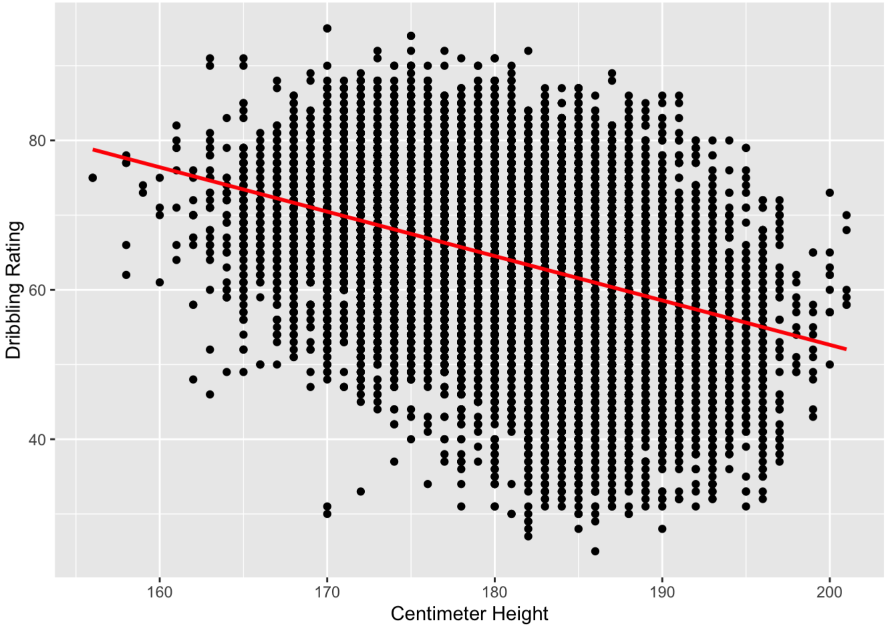
Adds predictions for dribbling rates for each distinct height based on our model.

```
control_pred <- final %>%
  data_grid(`Centimeter Height`) %>%
  add_predictions(control_model)
control_pred
```

```
## # A tibble: 45 × 2
##    `Centimeter Height`  pred
##                  <int> <dbl>
##  1                 156  78.8
##  2                 158  77.6
##  3                 159  77.0
##  4                 160  76.4
##  5                 161  75.8
##  6                 162  75.2
##  7                 163  74.6
##  8                 164  74.0
##  9                 165  73.4
## 10                 166  72.8
## # … with 35 more rows
```

Plots a scatter plot of a relationship between player height and dribbling rating. We then add a line of best fit based on our model that looks at the same x variable (height) but uses predicted values as y instead of actual dribbling ratings.

```
ggplot(data = final) +
  geom_point(mapping = aes(x = `Centimeter Height`, y = `Dribbling Rating`)) +
  geom_line(mapping = aes(x = `Centimeter Height`, y = pred), data = control_pred, color = "red", size = 1)
```



Write Up

Introduction Fifa is company that launches a new video game every year which allows its users to simulate real life soccer by controlling players in-game that play for real life teams. Fifa prides itself in creating a realistic game experience that captures the nuances of real life soccer within its digital boundaries. I decided to test 3 widely agreed upon (leaning towards the un-controversial side of things) real-life theories about the game and see if Fifa reflects these theories within the game in order to help facilitate the real-life experience.

The questions I asked were the following:

1- After the pandemic (2021), Fifa (the footballing organization) decided to condense a normal football season into a shortened time frame which almost doubled the amount of matches professional footballers played every week. Many managers like Pep Guardiola were furious and said this negatively affected the physical condition of the players who were overworked. Does the data set in Fifa reflect this real life theory? Is the average Physical Rating lower in Fifa 21 compared to Fifa 22? Was the average Overall Rating also lower in Fifa 21 as a result of a lack of practice? Were the players also slower during the condensed season?

2- There is a theory among soccer fans that the Italian league is more defensive minded than other leagues. Is this theory reflected in the Fifa 22 dataset?

3- A widely held theory in soccer is that shorter players have an advantage when it comes to ball control, as they tend to have a lower center of gravity and can therefore control the ball with their feet easier. Is this theory reflected in the Fifa dataset?

While I will delve deeper into the results in the conclusion section, I will briefly mention that I discovered that Fifa did a good job reflecting the truths of these theories within their game mechanics, however, for one of the theories, I think they could have done a better job accentuating its truth.

Data Source I obtained my datasets from a complete player Dataset for Fifa 22 and Fifa 21 on the website Kaggle. The information on these datasets were obtained by web scraping from the website “Sofifa”. Web scraping is the process of extracting information from websites using data analytic tools. The person who published this data is Stefano Leone who is an experience data analyst who has multiple certifications for data related skills.

Sofifa obtains its data by directly extracting information from the video game Fifa and publishing it on their website. Fifa has direct legal access to all player names, club names, and player information. As professional footballers, these players agree to provide Fifa with personal information such as height and weight.

Here is a link to the dataset: https://www.kaggle.com/stefanoleone992/fifa-22-complete-player-dataset/version/3?select=players_22.csv (https://www.kaggle.com/stefanoleone992/fifa-22-complete-player-dataset/version/3?select=players_22.csv)

Data Ethics Given the fact that this dataset is obtained entirely from a website based on a video game, which then obtains all of its information from the official body of soccer, I think the ethical considerations of this study are not as pertinent as other studies which might look at more serious topics like medicine or law. However, this is not to say that the ethical considerations do not exist.

While soccer players agree to provide public access to a lot of personal information about themselves like their height or weight, studies which can then go and analyze a lot of this information can cause these footballers (who are major stakeholders) moral degradation. For example, some of the attributes in the dataset listed some footballers as “Injury Prone” which is a negative classification and can cause stress on many of these footballers. Furthermore, studies which then go and compare different ratings of different footballers in order to rank them can make these footballers feel less like humans and more like products which dehumanizes them.

To another degree, I think people can use some of this data to propagate racist ideas. Specifically, there are many nationalities from Asia that have lower overall players ratings (especially in India and China). Someone who wants to promote racism might then go and deduce from this fact that Indians and Chinese people are inferior at soccer and are therefore not as athletic as people from European nations for example. However, this data does not take into account the fact that many of these Asian countries are very athletic, they just excel at other sports like Cricket or Ping Pong for example. As a student of CS36, I would be able to point out the holes in the racist person’s argument as there limitations to how broadly applicable the findings are, however, other people who are not as familiar with data might not realize what this person is doing.

Conclusions 1- After investigating if the physical conditions of the players was negatively affected during the pandemic season was reflected in the game, I noticed that there was in fact a blip in some general player ratings in Fifa 21 as compared to Fifa 22. The average physical rating in Fifa 21 was 65.57 compared to the average physical rating of 66.99 in Fifa 22 showing a clear increase in the new game resulting from a more normal training schedule. The average pace rating in Fifa 21 was 68.52 compared to the average pace rating of 68.68 in Fifa 22. This increase is negligible so it does not appear the speed of players was affected during the pandemic season. Finally, the average overall rating in Fifa 21 was 67.13 compared to the average overall rating in Fifa 22 of 68.06. This increase could very well be from a more regular traning schedule leading to better rating.

It must be noted that these conclusions are not certain. All 3 increases from Fifa 21 to Fifa 22 could may have well resulted from other factors as this was not an isolated study in a lab, and more of a correlation analysis. I will say though, from my experience, I saw that the pandemic season did really take a toll on many players after seeing more frequent injuries, so I would have expected a bigger blip in physical condition ratings for players in Fifa 21 as compared to Fifa 22. Finally, all conclusions that we draw from this study can be used to make inferences about the players within the game, and not in real life. While it is interesting to examine real-life/game differences and similarities, it must be made explicit that these conclusions are applicable only to inferences about the players within the game.

2- After investigating if Fifa reflects the theory that the Italian league is more defensive than other leagues, the data strongly supported that conclusion. Based on the data, our model revealed that the General Predicted Defensive Rating was 60.09 for the Italian league which was the highest out of all top 5 leagues (followed by England with 59.92, Spain with 58.75, Germany with 57.62, and France with 56.07). The average Italian player would generally have a higher defensive rating than a player from another league. Our model also revealed that the Defensive Predicted Defensive Rating was 74.36 for the Italian league which was the highest out of all top 5 leagues (followed by Spain with 73.05, England with 72.76, Germany with 70.88, and France with 69.85). Even for players who were specifically defenders, the average Italian would generally have a higher defensive rating than a player from another league.

It seems pretty fair to say that the Italian league is more defensive than other leagues within the game of Fifa. I must reiterate that conclusions drawn from this study are applicable towards the game rather than real life, even if it seems interesting to apply it to real life. We cannot not yet say that the Italian League in real life is more defensive than other leagues using evidence from this study alone.

3- After investigating if Fifa reflects the theory that shorter players tend to have better dribbling, it appears that the game supports that theory. I plotted the relationship between height (x-axis) and dribbling rating (y-axis) and derived a linear model which gave me a negative slope of 0.6. This suggests that the data shows that as a player's height increases, the dribbling rating tends to decrease.

That being said, we can't be 100% certain and attribute this inverse relationship due to low center of gravity. While it is very likely that is the case, based on the variables we have available to us, we can only infer the existence of an inverse relationship between both variables within the game, and not in real life for the reasons I mentioned for the previous 2 questions.

If a person wanted to go a step further and make deductions about the real life players, the next step would be to acquire more real life information by testing the player's performance in a real life setting and then performing data analysis on that. However, we must abide by the rules of statistics when doing so and ensure we cover basic guidelines like taking random samples of footballers rather than handpicking which players we choose to perform tests on.