Group Project

Hugo Moncada

3/30/2022

This data of international data was collected between February 16th, 2022 and April 23, 2022. Thank you Harry Kim for filling out most of the data and finishing in April. Please not that data is constantly changing through time, and our data should be most representative of data based on the month of April 2022.

Question: How much influence does international caps have on player value based on the best 6 teams in Europe?

We chose only European teams primarily due to regional/cultural similarities and convenience.

```r
MoneyballData = read.csv("Moneyball Soccer Dataset.csv")
```

```r
names(MoneyballData) = c("Players","Nationality","Club","Position","Age","Value","Wage","Caps")
attach(MoneyballData)
MoneyballData[9:13] <- list(NULL)
```

- Players - Player name (First Last)
- Nationality - Country that a player plays for internationally through FIFA
- Club - Team that a player plays for
- Position - The primary position of a player: goalkeeper (GK), defender (DF), midfielder (MF), or forward (FW)
- Age - How old a player is at the time of taking our data
- Position - What a player should theoretically sell for in the transfer market if another team wanted to buy the player based on transfermarkt.us (in € and in millions). Value is primarily based on a player's age, potential, skill, in-game performance, and much more. For more detail on how transfermarkt determines value, read https://www.transfermarkt.co.in/transfermarkt-market-value-explained-how-is-it-determined-/view/news/385100.
- Wage - Amount of money a player earns (in € and in thousands) via salarysport.com.
- Caps - Amount of international games a player has played for their country (only applies if player participates in the game and does not count if they just get called up and sits on the bench)

In order to check whether we can use linear regression for our data, we need to check 4 assumptions (will learn these in STA 141): https://www.godatadrive.com/blog/basic-guide-to-test-assumptions-of-linear-regression-in-r

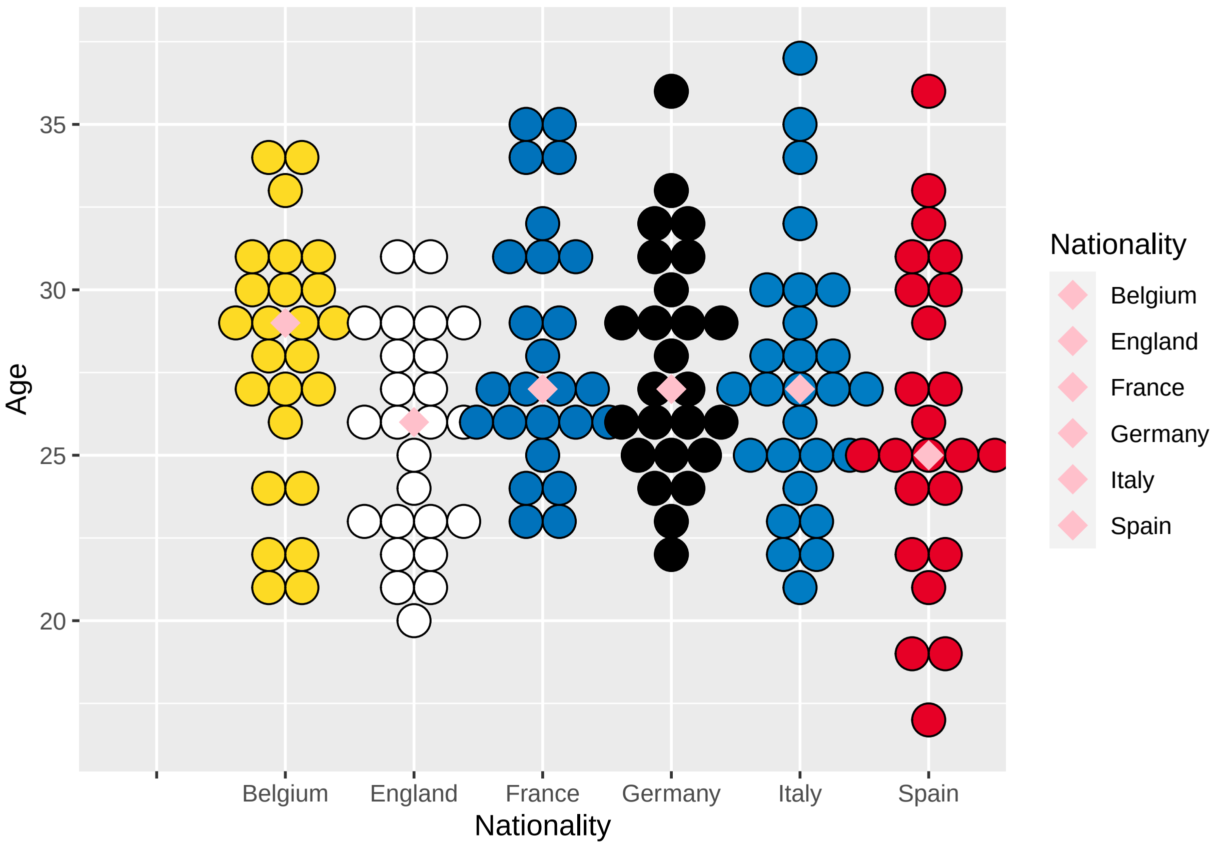
```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1

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

```
library(dplyr)
library(ggplot2)
```

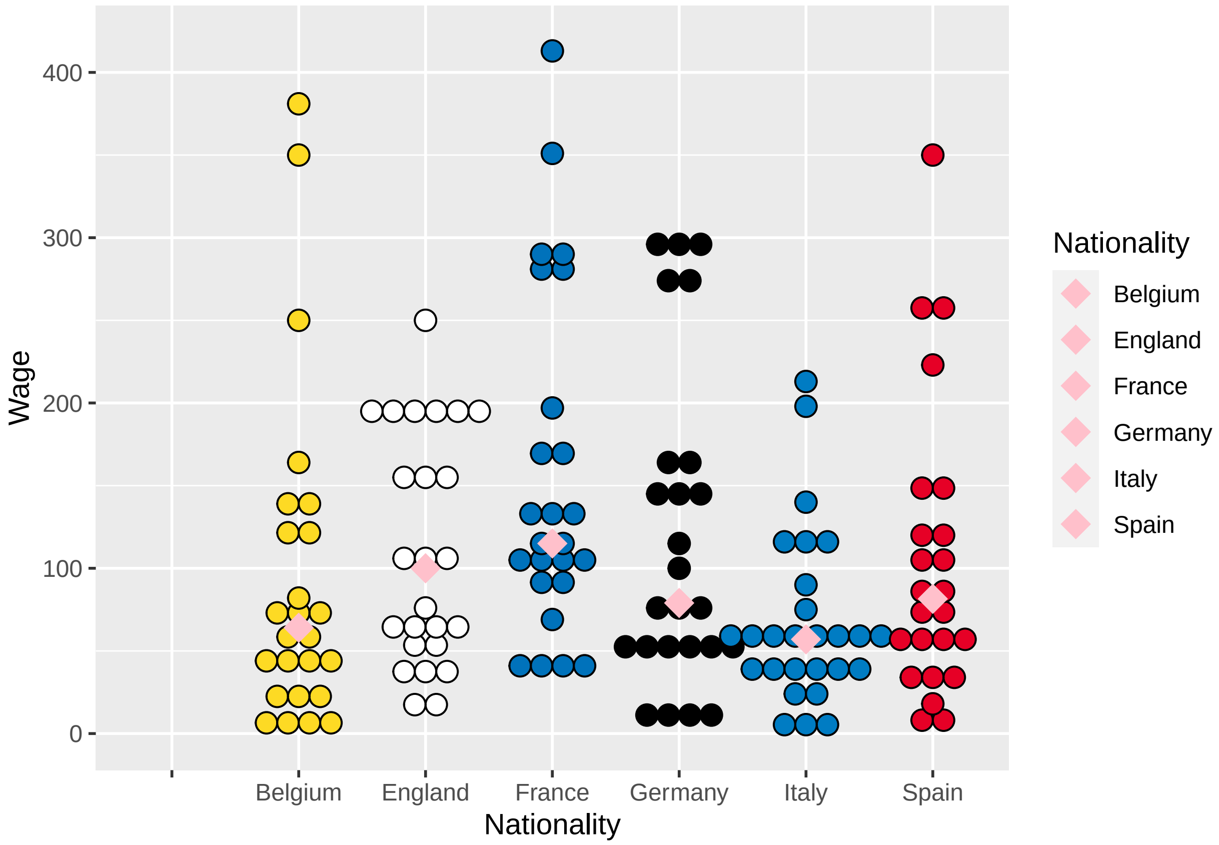
```
# Dot Plot tosee Equal Variance (value, wage and caps)

# age
plot <- ggplot(MoneyballData, aes(x=Nationality, y=Age, fill = Nationality)) + scale_fill_manual(values=
  geom_dotplot(binaxis=’y’, stackdir=’center’, binwidth = 1) + stat_summary(fun=median, geom="point", sh
                 size=5, color="pink")
plot
```

```
## Warning: Removed 32 rows containing non-finite values (stat_bindot).

## Warning: Removed 32 rows containing non-finite values (stat_summary).
```



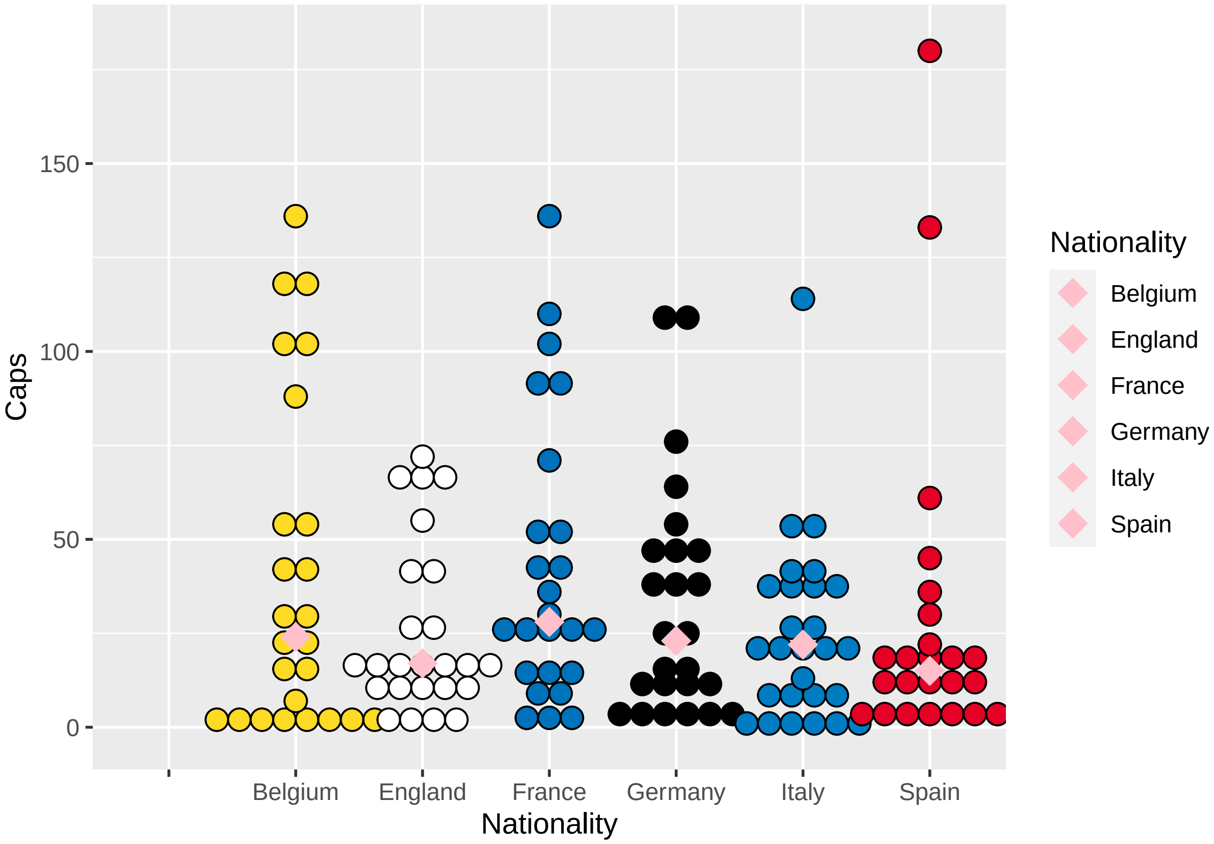
```
# wage
plot <- ggplot(MoneyballData, aes(x=Nationality, y=Wage, fill = Nationality)) + scale_fill_manual(value
  geom_dotplot(binaxis='y', stackdir='center', binwidth = 13) + stat_summary(fun=median, geom="point", 
                  size=5, color="pink")
plot
```

```
## Warning: Removed 32 rows containing non-finite values (stat_bindot).

## Warning: Removed 32 rows containing non-finite values (stat_summary).
```



```
# caps
plot <- ggplot(MoneyballData, aes(x=Nationality, y=Caps, fill = Nationality)) + scale_fill_manual(value
  geom_dotplot(binaxis='y', stackdir='center', binwidth = 6) + stat_summary(fun=median, geom="point", sl
                  size=5, color="pink")
plot
```

```
## Warning: Removed 32 rows containing non-finite values (stat_bindot).

## Warning: Removed 32 rows containing non-finite values (stat_summary).
```



We see that the variances of the factor level means look similar for each variable. While given small number of observations, the variation in the dotplots of each group's observations seems similar (might be randomness).