DASC 1104 Project Proposal

Gerardo Moreno

12-3-20

My blog is available at https://dasc-1104-finalproject-blog.netlify.app

```
library(here)
```

```
## here() starts at /home/gdmoreno/finalproject
```

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## v purrr   0.3.3
```

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

```
library(readxl)
library(tinytex)
library(fivethirtyeight)
knitr::opts_chunk$set(echo = FALSE, tidy = TRUE)
epl1 <- read.csv(file = here::here("epl_data", "epl2020.csv"))
epl2 <- read.csv(file = here::here("epl_data", "EPL_19_20_GW_Timeseries.csv"))
epl3 <- read.csv(file = here::here("epl_data", "epl_players_stats.csv"))
epl4 <- read.csv(file = here::here("epl_data", "players_1920_fin.csv"))
glimpse(epl1)
```

```
## Observations: 576
## Variables: 45
## $ X            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 1
## $ h_a          <fct> h, a, a, h, h, a, h, h, a, a, a, h, a, h, a, h, h, a, a, h, h, a, h, a, h, h, a
## $ xG           <dbl> 2.234560, 0.842407, 3.183770, 1.200300, 1.340990, 0.855516, 0.909241, 0.871590,
## $ xGA          <dbl> 0.842407, 2.234560, 1.200300, 3.183770, 1.598640, 0.670022, 1.087520, 1.224600,
## $ npxG         <dbl> 2.234560, 0.842407, 2.422640, 1.200300, 1.340990, 0.855516, 0.909241, 0.871590,
## $ npxGA        <dbl> 0.842407, 2.234560, 1.200300, 2.422640, 1.598640, 0.670022, 1.087520, 1.224600,
## $ deep         <int> 11, 5, 9, 1, 4, 5, 0, 5, 5, 6, 9, 7, 2, 15, 10, 7, 1, 3, 4, 8, 8, 4, 12, 9, 8, 4
```

```
## $ deep_allowed <int> 5, 11, 1, 9, 6, 7, 9, 5, 5, 4, 0, 5, 15, 2, 1, 3, 10, 7, 8, 4, 4, 8, 9, 12, 4, 3
## $ scored       <int> 4, 1, 5, 0, 1, 3, 3, 0, 0, 1, 0, 0, 1, 3, 1, 0, 0, 0, 0, 4, 2, 1, 1, 2, 1, 1, 2
## $ missed       <int> 1, 4, 0, 5, 1, 0, 0, 0, 0, 1, 3, 3, 3, 1, 0, 0, 1, 0, 4, 0, 1, 2, 2, 1, 1, 0, 1
## $ xpts         <dbl> 2.3863, 0.4405, 2.6312, 0.2522, 1.0172, 1.5458, 1.1422, 1.0512, 1.6257, 1.5905,
## $ result       <fct> w, l, w, l, d, w, w, d, d, d, l, l, l, w, w, d, l, d, l, w, w, l, l, w, d, w, w
## $ date         <fct> 2019-08-09 20:00:00, 2019-08-09 20:00:00, 2019-08-10 12:30:00, 2019-08-10 12:30
## $ wins         <int> 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1
## $ draws        <int> 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0
## $ loses        <int> 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0
## $ pts          <int> 3, 0, 3, 0, 1, 3, 3, 1, 1, 1, 0, 0, 0, 3, 3, 1, 0, 1, 0, 3, 3, 0, 0, 3, 1, 3, 3
## $ npxGD        <dbl> 1.392153, -1.392153, 1.222340, -1.222340, -0.257650, 0.185494, -0.178279, -0.353
## $ teamId       <fct> Liverpool, Norwich, Man City, West Ham, Bournemouth, Brighton, Burnley, Crystal
## $ ppda_cal     <dbl> 5.942857, 21.454545, 6.935484, 16.400000, 13.600000, 11.500000, 9.277778, 11.307
## $ allowed_ppda <dbl> 21.454545, 5.942857, 16.400000, 6.935484, 9.133333, 9.785714, 5.500000, 4.074074
## $ matchtime    <int> 20, 20, 12, 12, 15, 15, 15, 15, 15, 15, 15, 15, 17, 17, 14, 14, 14, 14, 16, 16,
## $ tot_points   <int> 3, 0, 3, 0, 1, 3, 3, 1, 1, 1, 0, 0, 0, 3, 3, 1, 0, 1, 0, 3, 6, 3, 0, 4, 4, 4, 6
## $ round        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2
## $ tot_goal     <int> 4, 1, 5, 0, 1, 3, 3, 0, 0, 1, 0, 0, 1, 3, 1, 0, 0, 0, 0, 4, 3, 4, 2, 3, 4, 1, 6
## $ tot_con      <int> 1, 4, 0, 5, 1, 0, 0, 0, 0, 1, 3, 3, 3, 1, 0, 0, 1, 0, 4, 0, 1, 2, 5, 2, 1, 0, 2
## $ Referee.x    <fct> M Oliver, M Oliver, M Dean, M Dean, K Friend, C Pawson, G Scott, J Moss, J Moss
## $ HS.x         <int> 15, 15, 5, 5, 13, 11, 10, 6, 6, 13, 10, 11, 31, 31, 9, 15, 9, 15, 11, 11, 16, 10
## $ HST.x        <int> 7, 7, 3, 3, 3, 3, 4, 2, 2, 3, 4, 3, 7, 7, 2, 1, 2, 1, 5, 5, 9, 9, 7, 7, 4, 2, 3
## $ HF.x         <int> 9, 9, 6, 6, 10, 15, 6, 16, 16, 10, 6, 15, 13, 13, 12, 3, 12, 3, 15, 15, 13, 13,
## $ HC.x         <int> 11, 11, 1, 1, 3, 5, 2, 6, 6, 3, 2, 5, 14, 14, 5, 12, 5, 12, 3, 3, 10, 10, 10, 10
## $ HY.x         <int> 0, 0, 2, 2, 2, 0, 0, 2, 2, 2, 0, 0, 1, 1, 1, 0, 1, 0, 3, 3, 2, 2, 0, 0, 0, 2, 2
## $ HR.x         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ AS.x         <int> 12, 12, 14, 14, 8, 5, 11, 10, 10, 8, 11, 5, 7, 7, 8, 8, 8, 8, 18, 18, 18, 18, 11
## $ AST.x        <int> 5, 5, 9, 9, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 2, 2, 2, 2, 7, 7, 5, 5, 4, 4, 3, 2, 6
## $ AF.x         <int> 9, 9, 13, 13, 19, 11, 12, 14, 14, 19, 12, 11, 9, 9, 7, 13, 7, 13, 13, 13, 11, 11
## $ AC.x         <int> 2, 2, 1, 1, 4, 2, 7, 2, 2, 4, 7, 2, 0, 0, 3, 3, 3, 3, 5, 5, 7, 7, 5, 5, 6, 7, 9
## $ AY.x         <int> 2, 2, 2, 2, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 3, 2, 3, 2, 4, 4, 1, 1, 2, 2, 2, 3, 1
## $ AR.x         <int> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ B365H.x      <dbl> 1.14, 1.14, 12.00, 12.00, 1.95, 1.90, 2.62, 3.00, 3.00, 1.95, 2.62, 1.90, 1.30,
## $ B365D.x      <dbl> 10.00, 10.00, 6.50, 6.50, 3.60, 3.40, 3.20, 3.25, 3.25, 3.60, 3.20, 3.40, 5.25,
## $ B365A.x      <dbl> 19.00, 19.00, 1.22, 1.22, 3.60, 4.00, 2.75, 2.37, 2.37, 3.60, 2.75, 4.00, 10.00
## $ HtrgPerc     <dbl> 0.46666667, 0.46666667, 0.60000000, 0.60000000, 0.23076923, 0.27272727, 0.40000
## $ AtrgPerc     <dbl> 0.41666667, 0.41666667, 0.64285714, 0.64285714, 0.37500000, 0.60000000, 0.27272
## $ matchDay     <fct> Fri, Fri, Sat, Sat, Sat, Sat, Sat, Sat, Sat, Sat, Sat, Sat, Sat, Sat, Sun, Sun,

## Observations: 1,048,575
## Variables: 77
## $ Teams          <fct>  MCI, MUN, LIV, BHA, BUR, TOT, ARS, BOU, SHU, CRY, EVE, LEI, WOL, NEW, AVL, NO
## $ GW1_Points     <int> 3, 3, 3, 3, 3, 3, 3, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, NA, NA, NA, NA, NA
## $ GW1_Standing   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, NA, NA
## $ GW2_Points     <int> 4, 4, 6, 4, 3, 4, 6, 4, 4, 1, 4, 2, 2, 0, 0, 3, 0, 0, 1, 1, NA, NA, NA, NA, NA
## $ GW2_Standing   <int> 3, 4, 1, 5, 10, 6, 2, 7, 8, 14, 9, 12, 13, 18, 17, 11, 19, 20, 15, 16, NA, NA
## $ GW3_Points     <int> 7, 4, 9, 4, 4, 4, 6, 4, 4, 4, 4, 5, 3, 3, 3, 3, 3, 3, 4, 4, NA, NA, NA, NA, NA
## $ GW3_Standing   <int> 2, 5, 1, 8, 6, 7, 3, 11, 9, 10, 12, 4, 15, 19, 16, 17, 18, 20, 13, 14, NA, NA
## $ GW4_Points     <int> 10, 5, 12, 4, 4, 5, 7, 4, 5, 7, 7, 8, 3, 4, 3, 3, 4, 1, 5, 7, NA, NA, NA, NA,
## $ GW4_Standing   <int> 2, 8, 1, 16, 12, 9, 5, 15, 10, 4, 6, 3, 17, 14, 18, 19, 13, 20, 11, 7, NA, NA
## $ GW5_Points     <int> 10, 8, 15, 5, 5, 8, 8, 7, 5, 7, 7, 8, 3, 4, 4, 6, 7, 2, 8, 8, NA, NA, NA, NA,
## $ GW5_Standing   <int> 2, 4, 1, 16, 14, 3, 7, 9, 15, 12, 11, 5, 19, 18, 17, 13, 10, 20, 6, 8, NA, NA
## $ GW6_Points     <int> 13, 8, 18, 6, 8, 8, 11, 10, 8, 8, 7, 11, 4, 5, 4, 6, 7, 2, 8, 11, NA, NA, NA,
## $ GW6_Standing   <int> 2, 8, 1, 15, 9, 7, 4, 6, 10, 12, 14, 3, 19, 17, 18, 16, 13, 20, 11, 5, NA, NA
```

```
## $ GW7_Points     <int> 16, 9, 21, 6, 9, 11, 12, 11, 8, 11, 7, 14, 7, 5, 5, 6, 7, 2, 11, 12, NA, NA, N
## $ GW7_Standing   <int> 2, 10, 1, 16, 11, 6, 4, 8, 12, 9, 15, 3, 13, 19, 18, 17, 14, 20, 7, 5, NA, NA
## $ GW8_Points     <int> 16, 9, 24, 9, 12, 11, 15, 11, 9, 14, 7, 14, 10, 8, 8, 6, 7, 3, 14, 12, NA, NA
## $ GW8_Standing   <int> 2, 12, 1, 14, 7, 9, 3, 10, 13, 6, 18, 4, 11, 16, 15, 19, 17, 20, 5, 8, NA, NA
## $ GW9_Points     <int> 19, 10, 25, 9, 12, 12, 15, 12, 12, 14, 10, 17, 11, 8, 11, 7, 8, 4, 17, 12, NA
## $ GW9_Standing   <int> 2, 14, 1, 16, 8, 7, 5, 10, 9, 6, 15, 3, 13, 18, 12, 19, 17, 20, 4, 11, NA, NA
## $ GW10_Points    <int> 22, 13, 28, 12, 12, 12, 16, 13, 13, 15, 10, 20, 12, 9, 11, 7, 8, 5, 20, 13, NA
## $ GW10_Standing  <int> 2, 7, 1, 14, 13, 11, 5, 9, 8, 6, 16, 3, 12, 17, 15, 19, 18, 20, 4, 10, NA, NA
## $ GW11_Points    <int> 25, 13, 31, 15, 12, 13, 17, 16, 16, 15, 11, 23, 13, 12, 11, 7, 8, 5, 23, 13, N
## $ GW11_Standing  <int> 2, 10, 1, 8, 14, 11, 5, 7, 6, 9, 17, 3, 12, 15, 16, 19, 18, 20, 4, 13, NA, NA
## $ GW12_Points    <int> 25, 16, 34, 15, 15, 14, 17, 16, 17, 15, 14, 26, 16, 15, 11, 7, 8, 8, 26, 13, N
## $ GW12_Standing  <int> 4, 7, 1, 11, 10, 14, 6, 9, 5, 12, 15, 2, 8, 13, 17, 20, 19, 18, 3, 16, NA, NA
## $ GW13_Points    <int> 28, 17, 37, 15, 18, 17, 18, 16, 18, 15, 14, 29, 19, 15, 14, 10, 9, 8, 26, 13,
## $ GW13_Standing  <int> 3, 9, 1, 12, 7, 10, 8, 11, 6, 13, 16, 2, 5, 14, 15, 18, 19, 20, 4, 17, NA, NA
## $ GW14_Points    <int> 29, 18, 40, 15, 18, 20, 19, 16, 19, 18, 14, 32, 20, 16, 15, 11, 12, 8, 26, 16
## $ GW14_Standing  <int> 3, 9, 1, 16, 10, 5, 8, 12, 7, 11, 17, 2, 6, 14, 15, 19, 18, 20, 4, 13, NA, NA
## $ GW15_Points    <int> 32, 21, 43, 18, 18, 20, 19, 16, 19, 21, 14, 35, 23, 19, 15, 11, 15, 8, 29, 16
## $ GW15_Standings <int> 3, 6, 1, 13, 12, 8, 10, 14, 9, 7, 18, 2, 5, 11, 16, 19, 17, 20, 4, 15, NA, NA
## $ GW16_Points    <int> 32, 24, 46, 19, 18, 23, 22, 16, 22, 22, 17, 38, 24, 22, 15, 11, 15, 9, 29, 16
## $ GW16_Standing  <int> 3, 5, 1, 12, 13, 7, 9, 15, 8, 10, 14, 2, 6, 11, 17, 19, 18, 20, 4, 16, NA, NA
## $ GW17_Points    <int> 35, 25, 49, 20, 21, 26, 22, 19, 25, 23, 18, 39, 24, 22, 15, 12, 15, 9, 29, 19
## $ GW17_Standing  <int> 3, 6, 1, 13, 12, 5, 10, 14, 7, 9, 16, 2, 8, 11, 17, 19, 18, 20, 4, 15, NA, NA
## $ GW18_Points    <int> 38, 25, 49, 20, 24, 26, 23, 19, 28, 23, 19, 39, 27, 25, 15, 12, 18, 12, 32, 19
## $ GW18_Standing  <int> 3, 8, 1, 13, 10, 7, 11, 14, 5, 12, 15, 2, 6, 9, 18, 19, 17, 20, 4, 16, NA, NA
## $ GW19_Points    <int> 38, 28, 52, 20, 24, 29, 24, 20, 29, 26, 22, 39, 30, 25, 18, 12, 21, 13, 32, 19
## $ GW19_Standing  <int> 3, 8, 1, 15, 12, 6, 11, 16, 7, 9, 13, 2, 5, 10, 18, 20, 14, 19, 4, 17, NA, NA
## $ GW20_Points    <int> 41, 31, 55, 23, 24, 30, 24, 20, 29, 27, 25, 42, 30, 25, 18, 13, 22, 16, 35, 19
## $ GW20_Standing  <int> 3, 5, 1, 14, 13, 6, 12, 16, 8, 9, 10, 2, 7, 11, 18, 20, 15, 19, 4, 17, NA, NA
## $ GW21_Points    <int> 44, 31, 58, 24, 24, 30, 27, 20, 29, 28, 25, 45, 30, 25, 21, 14, 25, 19, 36, 22
## $ GW21_Standing  <int> 3, 5, 1, 14, 15, 6, 10, 18, 8, 9, 11, 2, 7, 13, 17, 20, 12, 19, 4, 16, NA, NA
## $ GW22_Points    <int> 47, 34, 61, 24, 24, 30, 28, 20, 32, 29, 28, 45, 31, 26, 21, 14, 28, 22, 39, 23
## $ GW22_Standing  <int> 2, 5, 1, 14, 15, 8, 10, 19, 6, 9, 11, 3, 7, 13, 18, 20, 12, 17, 4, 16, NA, NA
## $ GW23_Points    <int> 48, 34, 64, 25, 27, 31, 29, 20, 33, 30, 29, 45, 34, 29, 22, 17, 28, 23, 39, 23
## $ GW23_Standing  <int> 2, 5, 1, 15, 14, 8, 10, 19, 7, 9, 11, 3, 6, 12, 18, 20, 13, 17, 4, 16, NA, NA
## $ GW24_Points    <int> 51, 34, 70, 25, 30, 34, 30, 23, 33, 30, 30, 48, 34, 30, 25, 17, 31, 23, 40, 23
## $ GW24_Standing  <int> 2, 5, 1, 15, 13, 6, 10, 18, 8, 11, 12, 3, 7, 14, 16, 20, 9, 19, 4, 17, NA, NA
## $ GW25_Points    <int> 51, 35, 73, 26, 31, 37, 31, 26, 36, 30, 33, 49, 35, 31, 25, 18, 31, 23, 41, 24
## $ GW25_Standing  <int> 2, 7, 1, 15, 11, 5, 10, 16, 6, 14, 9, 3, 8, 12, 17, 20, 13, 19, 4, 18, NA, NA
## $ GW26_Points    <int> 54, 38, 76, 27, 34, 40, 34, 26, 39, 30, 36, 50, 36, 31, 25, 18, 31, 24, 41, 24
## $ GW26_Standing  <int> 2, 7, 1, 15, 11, 5, 10, 16, 6, 14, 9, 3, 8, 13, 17, 20, 12, 19, 4, 18, NA, NA
## $ GW27_Points    <int> 57, 41, 79, 28, 37, 40, 37, 26, 40, 33, 36, 50, 39, 31, 25, 18, 34, 24, 44, 24
## $ GW27_Standing  <int> 2, 5, 1, 15, 10, 6, 9, 16, 7, 13, 11, 3, 8, 14, 17, 20, 12, 19, 4, 18, NA, NA
## $ GW28_Points    <int> 57, 42, 79, 28, 38, 40, 37, 27, 40, 36, 37, 50, 42, 32, 25, 21, 34, 27, 45, 27
## $ GW28_Standing  <int> 2, 5, 1, 15, 9, 7, 10, 18, 8, 12, 11, 3, 6, 14, 19, 20, 13, 17, 4, 16, NA, NA
## $ GW29_Points    <int> 57, 45, 82, 29, 39, 41, 40, 27, 43, 39, 37, 53, 43, 35, 25, 21, 34, 27, 48, 27
## $ GW29_Standing  <int> 2, 5, 1, 15, 10, 8, 9, 18, 7, 11, 12, 3, 6, 13, 19, 20, 14, 17, 4, 16, NA, NA
## $ GW30_Points    <int> 63, 46, 83, 32, 39, 42, 40, 27, 44, 42, 38, 54, 46, 38, 26, 21, 37, 28, 51, 27
## $ GW30_Standing  <int> 2, 5, 1, 15, 11, 8, 10, 18, 7, 9, 12, 3, 6, 13, 19, 20, 14, 16, 4, 17, NA, NA
## $ GW31_Points    <int> 63, 49, 86, 33, 42, 45, 43, 27, 44, 42, 41, 55, 49, 39, 27, 21, 37, 28, 54, 27
## $ GW31_Standing  <int> 2, 5, 1, 15, 11, 7, 9, 18, 8, 10, 12, 3, 6, 13, 19, 20, 14, 16, 4, 17, NA, NA
## $ GW32_Points    <int> 66, 52, 86, 33, 45, 45, 46, 27, 47, 42, 44, 55, 52, 42, 27, 21, 40, 28, 54, 30
## $ GW32_Standing  <int> 2, 5, 1, 15, 10, 9, 8, 19, 7, 12, 11, 3, 6, 13, 18, 20, 14, 17, 4, 16, NA, NA
## $ GW33Points     <int> 66, 55, 89, 33, 46, 48, 49, 27, 48, 42, 44, 58, 52, 43, 27, 21, 43, 28, 57, 31
## $ GW33_Standing  <int> 2, 5, 1, 15, 10, 8, 7, 19, 9, 14, 11, 3, 6, 12, 18, 20, 13, 17, 4, 16, NA, NA
```

```
## $ GW34_Points    <int> 69, 58, 92, 36, 49, 49, 50, 28, 51, 42, 45, 59, 52, 43, 27, 21, 44, 31, 60, 3
## $ GW34_Standing  <int> 2, 5, 1, 15, 10, 9, 8, 18, 7, 14, 11, 4, 6, 13, 19, 20, 12, 17, 3, 16, NA, NA
## $ GW35_Points    <int> 72, 59, 93, 36, 50, 52, 50, 31, 54, 42, 45, 59, 55, 43, 30, 21, 45, 34, 60, 3
## $ GW35_Standing  <int> 2, 5, 1, 15, 10, 8, 9, 18, 7, 14, 11, 4, 6, 13, 19, 20, 12, 17, 3, 16, NA, NA
## $ GW36_Points    <int> 75, 62, 93, 37, 51, 55, 53, 31, 54, 42, 46, 62, 56, 43, 31, 21, 46, 34, 63, 3
## $ GW36_Standing  <int> 2, 5, 1, 16, 10, 7, 9, 18, 8, 14, 11, 4, 6, 13, 19, 20, 12, 17, 3, 15, NA, NA
## $ GW37_Points    <int> 78, 63, 96, 38, 54, 58, 53, 31, 54, 42, 49, 62, 59, 44, 34, 21, 49, 34, 63, 3
## $ GW37_Standing  <int> 2, 3, 1, 16, 9, 7, 10, 19, 8, 14, 11, 5, 6, 13, 17, 20, 12, 18, 4, 15, NA, NA
## $ GW38_Points    <int> 81, 66, 99, 41, 54, 59, 56, 34, 54, 43, 49, 62, 59, 44, 35, 21, 52, 34, 66, 3
## $ GW38_Standing  <int> 2, 3, 1, 15, 10, 6, 8, 18, 9, 14, 12, 5, 7, 13, 17, 20, 11, 19, 4, 16, NA, NA

## Observations: 522
## Variables: 41
## $ id                        <int> 816, 622, 812, 287, 158, 576, 506, 491, 988, 177, 217, 857, 764, 1
## $ name                      <fct> Kevin De Bruyne, Bruno Fernandes, Ricardo Pereira, Adama Traoré, R
## $ nationality               <fct> be, pt, pt, es, dz, sn, eg, gb-eng, gb-eng, gb-eng, gb-eng, nl, fr
## $ team                      <fct> Man City, Man Utd, Leicester, Wolves, Man City, Liverpool, Liverpo
## $ team_id                   <int> 875, 314, 652, 518, 875, 339, 339, 875, 314, 652, 512, 339, 314, 6
## $ age                       <int> 29, 26, 26, 24, 29, 28, 28, 25, 22, 23, 27, 29, 24, 32, 29, 30, 23
## $ height                    <int> 181, 179, 175, 178, 179, 175, 175, 170, 180, 175, 188, 193, 181, 1
## $ weight                    <int> 68, 69, 70, 72, 67, 69, 71, 69, 70, 73, 86, 92, 76, 77, 84, 82, 67
## $ position                  <fct> Midfielder, Midfielder, Defender, Forward, Forward, Forward, Forwa
## $ roles                     <fct> "M_C,M_L,M_R,FW", "M_C,M_L,M_R", "D_L,D_R,M_R", "M_L,M_R,FW", "AM_
## $ games                     <int> 35, 14, 28, 37, 33, 35, 34, 33, 31, 31, 29, 38, 32, 36, 18, 24, 26
## $ exit_on_sub               <int> 3, 0, 0, 10, 12, 4, 1, 3, 0, 2, 0, 0, 1, 7, 3, 5, 3, 1, 5, 13, 1, 1
## $ minutes                   <int> 2800, 1223, 2520, 2608, 1942, 2756, 2888, 2661, 2655, 2629, 2590, 3
## $ rating_m                  <dbl> 7.97, 7.73, 7.50, 7.49, 7.48, 7.45, 7.40, 7.35, 7.35, 7.34, 7.33, 7
## $ goals                     <int> 13, 8, 3, 4, 11, 18, 19, 20, 17, 6, 18, 5, 17, 9, 4, 10, 3, 1, 6, 1
## $ assists                   <int> 20, 7, 2, 9, 9, 7, 10, 1, 7, 3, 2, 1, 6, 7, 2, 3, 4, 0, 10, 7, 6, 1
## $ yel_cards                 <int> 3, 2, 1, 1, 0, 3, 1, 5, 3, 4, 4, 1, 1, 4, 4, 3, 0, 0, 0, 3, 3, 0, 8
## $ red_cards                 <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0
## $ shots_m                   <dbl> 2.8, 3.1, 0.6, 1.2, 2.2, 2.2, 3.9, 3.0, 3.1, 2.4, 2.8, 0.8, 2.5, 2
## $ aerials_won_m             <dbl> 0.5, 0.2, 1.5, 1.2, 0.3, 1.2, 0.4, 0.7, 0.7, 0.2, 2.6, 5.0, 0.6, 0
## $ motm                      <int> 10, 3, 2, 6, 3, 7, 6, 1, 7, 4, 4, 5, 4, 4, 1, 2, 3, 0, 3, 2, 3, 4,
## $ successful_passes_pt      <dbl> 81.5, 75.7, 78.9, 74.4, 90.1, 81.6, 76.5, 82.8, 77.3, 83.0, 66.7, 8
## $ key_passes_m              <dbl> 3.9, 2.1, 1.0, 1.3, 1.8, 1.7, 1.8, 1.5, 1.1, 2.5, 0.9, 0.2, 0.9, 2
## $ dribbles_won_m            <dbl> 1.4, 1.1, 2.1, 5.0, 1.6, 2.0, 1.5, 1.7, 1.9, 1.6, 1.0, 0.1, 1.6, 1
## $ fouls_given_m             <dbl> 0.8, 1.0, 1.3, 2.1, 0.9, 1.5, 0.5, 1.2, 1.5, 2.8, 1.5, 0.7, 1.3, 1
## $ offside_m                 <dbl> 0.1, 0.1, 0.0, 0.2, 0.4, 0.4, 0.6, 0.4, 0.4, 0.1, 0.7, 0.0, 0.4, 0
## $ dispossessed_m            <dbl> 0.9, 0.9, 1.1, 1.8, 1.0, 1.6, 2.0, 1.8, 1.1, 0.9, 1.1, 0.1, 1.9, 0
## $ passes_m                  <dbl> 54.5, 55.5, 56.1, 19.2, 33.2, 31.5, 28.8, 35.4, 30.4, 47.5, 19.1, 8
## $ accurate_crosses_m        <dbl> 2.1, 1.0, 0.3, 1.2, 0.6, 0.3, 0.3, 0.2, 0.3, 1.9, 0.1, 0.0, 0.1, 1
## $ accurate_long_passes_m    <dbl> 3.5, 3.3, 2.1, 0.2, 1.7, 1.0, 0.4, 0.3, 1.2, 2.3, 0.8, 5.4, 0.2, 1
## $ accurate_through_passes_m <dbl> 0.4, 0.3, 0.1, 0.1, 0.2, 0.1, 0.1, 0.1, 0.2, 0.2, 0.2, 0.1, 0.1, 0
## $ tackles_m                 <dbl> 1.3, 1.6, 4.2, 0.9, 0.8, 1.3, 0.5, 0.8, 0.4, 1.9, 0.6, 0.6, 0.5, 1
## $ interceptions_m           <dbl> 0.5, 0.8, 1.8, 0.4, 0.5, 0.4, 0.2, 0.5, 0.3, 0.3, 0.2, 1.1, 0.2, 0
## $ fouls_m                   <dbl> 0.7, 0.6, 1.6, 1.1, 0.4, 1.3, 0.5, 0.9, 0.7, 1.2, 1.3, 0.4, 0.9, 0
## $ clearances_m              <dbl> 0.7, 1.0, 2.6, 0.4, 0.2, 0.2, 0.1, 0.1, 0.5, 0.2, 1.0, 4.3, 0.5, 0
## $ dribbled_past_m           <dbl> 1.4, 2.4, 1.9, 0.4, 0.8, 0.8, 0.5, 0.7, 0.5, 1.5, 0.4, 0.2, 0.5, 0
## $ owm_goals                 <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ gk_in_goal_area_saves_m   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ gk_inside_box_saves_m     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ gk_outside_box_saves_m    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ gk_saves_m                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
```

```
## Observations: 22,502
## Variables: 35
## $ X                 <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
## $ assists           <int> 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ bonus             <int> 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ bps               <int> 1, 1, 1, 2, 53, 3, 22, 2, -5, 0, -1, 4, 0, 0, -1, 1, 0, -1, 0, 4, 1, 0, 0,
## $ clean_sheets      <int> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
## $ creativity        <dbl> 0.1, 0.3, 4.8, 0.6, 23.8, 10.6, 0.3, 5.8, 1.3, 0.0, 1.7, 0.3, 0.0, 0.0, 1.0
## $ element           <int> 534, 534, 534, 534, 534, 534, 534, 534, 534, 534, 534, 534, 534, 534, 534,
## $ fixture           <int> 37, 43, 58, 63, 72, 81, 92, 104, 115, 124, 134, 141, 153, 164, 173, 189, 19
## $ goals_conceded    <int> 1, 1, 0, 1, 0, 1, 2, 0, 2, 0, 2, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0,
## $ goals_scored      <int> 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ ict_index         <dbl> 0.0, 2.2, 2.5, 0.1, 20.1, 3.6, 0.2, 4.8, 0.3, 0.0, 2.0, 3.4, 0.0, 0.0, 2.4
## $ influence         <dbl> 0.2, 1.0, 2.0, 0.2, 70.2, 4.4, 0.0, 0.0, 0.0, 0.0, 1.2, 6.2, 0.0, 0.0, 2.2
## $ kickoff_time      <fct> 2019-08-31 14:00:00, 2019-09-14 14:00:00, 2019-09-21 16:30:00, 2019-09-28 1
## $ minutes           <int> 24, 6, 20, 25, 79, 45, 79, 85, 45, 0, 75, 75, 0, 0, 45, 67, 0, 45, 0, 67, 3
## $ opponent_team     <fct> Man City, Burnley, Newcastle United, Chelsea, Tottenham, Aston Villa, Ever
## $ own_goals         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ penalties_missed  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ penalties_saved   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ red_cards         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ round             <int> 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 1
## $ saves             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ selected          <int> 0, 14029, 22804, 32699, 35026, 110944, 146247, 263242, 246255, 208149, 1838
## $ team_a_score      <int> 0, 1, 0, 0, 0, 1, 2, 0, 1, 2, 1, 2, 2, 1, 1, 1, 0, 1, 0, 1, 1, 3, 1, 1, 1,
## $ team_h_score      <int> 4, 1, 0, 2, 3, 2, 3, 2, 3, 0, 2, 1, 2, 1, 0, 2, 2, 1, 1, 1, 3, 3, 1, 1, 0,
## $ threat            <int> 0, 21, 18, 0, 107, 21, 2, 46, 4, 0, 17, 27, 0, 0, 21, 6, 0, 29, 0, 28, 0, 0
## $ total_points      <int> 1, 1, 1, 1, 13, 1, 8, 2, 1, 0, 2, 2, 0, 0, 1, 2, 0, 1, 0, 2, 1, 0, 0, 1, 0
## $ transfers_balance <int> 0, 10589, 8090, 8437, 2156, 70802, 34969, 109148, -10884, -38688, -26463, -
## $ transfers_in      <int> 0, 13500, 11749, 13439, 5952, 81069, 42593, 118970, 18973, 10462, 6386, 967
## $ transfers_out     <int> 0, 2911, 3659, 5002, 3796, 10267, 7624, 9822, 29857, 49150, 32849, 15162, 1
## $ value             <dbl> 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.6, 4.6, 4.7, 4.7, 4.7, 4.7, 4.7, 4.6, 4.6,
## $ was_home          <fct> False, True, False, False, True, False, True, True, False, True, False, Fal
## $ yellow_cards      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ full              <fct> Aaron Connolly, Aaron Connolly, Aaron Connolly, Aaron Connolly, Aaron Conno
## $ team              <fct> Brighton, Brighton, Brighton, Brighton, Brighton, Brighton, Brighton, Brigh
## $ ppm               <dbl> 0.2222222, 0.2222222, 0.2222222, 0.2222222, 2.8888889, 0.2222222, 1.7391304
```

For this project, I am examining the English Premier League statistics by players stats, like height/age compared to their respective position. For example, in the datasets "epl_players_stats.csv" and "players_1920_fin.csv" from FiveThirtyEight, the file goes into full detail about players like height, position, goals per game, goals given up, age, ect, so I will be examining if a player if more likely to succeed in this league based on their height or age. The "players_1920_fin.csv"" there were 22,502 observations and 35 Variables, and "epl_players_stats.csv" has 522 observations and 41 variables. The variable codes is a factor with multiple levels explaing what teams each player plays for and their statistics for the 19/20 season.

The other two datasets "EPL_19_20_GW_Timeseries.csv" and "epl2020.csv" go into detail for the 19/20 statitics for each team in the English Premier League. In these data sets, we can find the results week by week of each team in the EPL and variable like win's, loss's, ties, points gained, goals given up, goals scored, ect.. The "EPL_19_20_GW_Timeseries.csv" has 1,048,575 observations and 77 variable and the "epl2020.csv" data set consits of 576 observations and 45 variables. These files have many levels going into full detail week by week, so i will have many outcomes to analyze.

*Question 1: First, do soccer players score more goals based on their height? I will analyze each players statistics for the 19/20 season based off their heights. Go into detail about if tall players score more goals

than short players or vise versa. The measurement will be based off the overall median height of every player in the league. To test this, I will be putting the data organized in a data visualization like a scatter plot.

*Question 2: Second, do teams tend to preform better with older, experienced players or young, talented players. Getting the average age of the players for each team and basing that with their results. Respectively, comparing if they score more goals, prevent more goals, and their wins, then plotting this into a bar chart.

*Question 3: Third, overall, do teams that spend more money correlate to their performance. Do team that spend more money obtaining and paying more players translate to more wins. I will be comparing the total amount spent by each team and comparing them in a bar chart.

*Question 4: Fourth, To be announced.

```
## Observations: 22,800
## Variables: 10
## $ person      <fct> "Constable Darius Quimby", "Sheriff Cornelius Hogeboom", "Deputy Sheriff Isaac Sm
## $ dept        <fct> "Albany County Constable's Office, NY", "Columbia County Sheriff's Office, NY",
## $ eow         <fct> "EOW: Monday, January 3, 1791", "EOW: Saturday, October 22, 1791", "EOW: Thursday
## $ cause       <fct> Cause of Death: Gunfire, Cause of Death: Gunfire, Cause of Death: Gunfire, Cause
## $ cause_short <fct> Gunfire, Gunfire, Gunfire, Gunfire, Gunfire, Gunfire, Stabbed, Gunfire, Assault,
## $ date        <fct> 1791-01-03, 1791-10-22, 1792-05-17, 1794-01-11, 1797-11-12, 1804-10-16, 1806-12-1
## $ year        <int> 1791, 1791, 1792, 1794, 1797, 1804, 1806, 1807, 1808, 1808, 1808, 1808, 1810, 181
## $ canine      <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS
## $ dept_name   <fct> Albany County Constable's Office, Columbia County Sheriff's Office, Westchester C
## $ state       <fct>  NY,  NY,  NY,  US,  SC,  NC,  NY,  KY,  ME,  MD,  US,  US,  US,  ME,  US,  NC,

## Observations: 467
## Variables: 34
## $ name                 <fct> A'donte Washington, Aaron Rutledge, Aaron Siler, Aaron Valdez, Adam Jovi
## $ age                  <fct> 16, 27, 26, 25, 29, 29, 22, 35, 44, 31, 76, 40, Unknown, 31, 23, 39, 25
## $ gender               <fct> Male, Male, Male, Male, Male, Male, Male, Male, Male, Male, Male, Male,
## $ raceethnicity        <fct> Black, White, White, Hispanic/Latino, White, White, Hispanic/Latino, His
## $ month                <fct> February, April, March, March, March, March, March, March, January, Febr
## $ day                  <int> 23, 2, 14, 11, 19, 7, 27, 26, 28, 7, 26, 12, 20, 25, 6, 30, 30, 19, 17,
## $ year                 <int> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015,
## $ streetaddress        <fct> Clearview Ln, 300 block Iris Park Dr, 22nd Ave and 56th St, 3000 Seminol
## $ city                 <fct> Millbrook, Pineville, Kenosha, South Gate, Munroe Falls, Phoenix, Bakers
## $ state                <fct> AL, LA, WI, CA, OH, AZ, CA, CA, TX, MI, CA, CA, TX, IN, IN, TN, GA, VA,
## $ latitude             <dbl> 32.52958, 31.32174, 42.58356, 33.93930, 41.14857, 33.46938, 35.39570, 33
## $ longitude            <dbl> -86.36283, -92.43486, -87.83571, -118.21946, -81.42988, -112.04332, -119
## $ state_fp             <int> 1, 22, 55, 6, 39, 4, 6, 6, 48, 26, 6, 6, 48, 18, 18, 47, 13, 51, 6, 6, 0
## $ county_fp            <int> 51, 79, 59, 37, 153, 13, 29, 37, 41, 81, 31, 59, 201, 167, 97, 37, 121,
## $ tract_ce             <int> 30902, 11700, 1200, 535607, 530800, 111602, 700, 294200, 603, 14200, 120
## $ geo_id               <dbl> 1051030902, 22079011700, 55059001200, 6037535607, 39153530800, 40131116(
## $ county_id            <int> 1051, 22079, 55059, 6037, 39153, 4013, 6029, 6037, 48041, 26081, 6031, 0
## $ namelsad             <fct> Census Tract 309.02, Census Tract 117, Census Tract 12, Census Tract 535
## $ lawenforcementagency <fct> Millbrook Police Department, Rapides Parish Sheriff's Office, Kenosha Po
## $ cause                <fct> Gunshot, Gunshot, Gunshot, Gunshot, Gunshot, Gunshot, Gunshot, Gunshot,
## $ armed                <fct> No, No, No, Firearm, No, No, Firearm, Non-lethal firearm, Firearm, Other
## $ pop                  <int> 3779, 2769, 4079, 4343, 6809, 4682, 5027, 5238, 4832, 3795, 2786, 9429,
## $ share_white          <fct> 60.5, 53.8, 73.8, 1.2, 92.5, 7, 50.8, 8.6, 14.6, 63.6, 51.4, 67.3, 7.5,
## $ share_black          <fct> 30.5, 36.2, 7.7, 0.6, 1.4, 7.7, 0.3, 0.2, 17.7, 7.7, 0.9, 0.6, 32.1, 7.0
## $ share_hispanic       <fct> 5.6, 0.5, 16.8, 98.8, 1.7, 79, 44.2, 84.1, 66.3, 26.5, 44.9, 12.9, 53.3
## $ p_income             <fct> 28375, 14678, 25286, 17194, 33954, 15523, 25949, 25043, 16778, 22005, 20
## $ h_income             <int> 51367, 27972, 45365, 48295, 68785, 20833, 58068, 66543, 30391, 44553, 50
```

```
## $ county_income        <int> 54766, 40930, 54930, 55909, 49669, 53596, 48552, 55909, 38310, 51667, 48
## $ comp_income          <dbl> 0.9379359, 0.6834107, 0.8258693, 0.8638144, 1.3848678, 0.3887044, 1.1959
## $ county_bucket        <int> 3, 2, 2, 3, 5, 1, 4, 4, 2, 3, 4, 5, 3, 4, 3, 1, 3, 1, 2, 4, 2, 4, 2, 2,
## $ nat_bucket           <int> 3, 1, 3, 3, 4, 1, 4, 4, 1, 2, 3, 5, 3, 2, 2, 1, 3, 2, 4, 5, 4, 3, 1, 2,
## $ pov                  <fct> 14.1, 28.8, 14.6, 11.7, 1.9, 58, 17.2, 12.2, 37.7, 18.4, 9.2, 4.4, 18.1
## $ urate                <dbl> 0.09768638, 0.06572379, 0.16629314, 0.12482727, 0.06354983, 0.07365145,
## $ college              <dbl> 0.16850951, 0.11140236, 0.14731227, 0.05013293, 0.40395421, 0.10295519,

## Observations: 75
## Variables: 8
## $ city              <fct> "New York", "Chicago", "Los Angeles", "Washington", "Houston", "Philadelphi
## $ police_force_size <int> 32300, 12120, 10100, 9340, 7700, 6045, 4475, 4460, 3605, 3265, 3020, 2955,
## $ all               <dbl> 0.61795666, 0.87500000, 0.22821782, 0.11563169, 0.29220779, 0.83540116, 0.3
## $ white             <dbl> 0.44638656, 0.87196262, 0.15277778, 0.05677419, 0.17373461, 0.77689873, 0.1
## $ non.white         <dbl> 0.76441894, 0.87740030, 0.26384840, 0.15736505, 0.39925834, 0.89948007, 0.4
## $ black             <fct> 0.770891365, 0.89740566, 0.387387387, 0.170189099, 0.36637931, 0.924657534
## $ hispanic          <fct> 0.762860728, 0.83982684, 0.217679558, 0.08988764, 0.457142857, 0.817391304
## $ asian             <fct> 0.749235474, 0.966666667, 0.305263158, 0.230769231, 0.408163265, **, **, 0
```

Using data obtained from FiveThirtyEight from multiple sources and the census, the next dataset I am going to investigate is the data of police killings and deaths in the United States through 2015-2016. The variables in the files include the date, name of state, name of deceased, race, and gender. The US population data is from the US Census Bureau website. There are 22,800 observations and 10 variables in the "clean_data.csv" file, which go into depth about police deaths. The "police_killings.csv" goes into depth about the police killing, which has 467 observations, and 34 variables. On the last dataset, the data is a collected from multiple police station from the biggest cities in the United States. With all the different data available, there should be enough data for me to analyze.

*Question 1: First, where do the police killings tend to happen more frequently. Using the "police_killings.csv" data I will analyze where these police killings tend to happen more frequently, like is there a state with notably more killings than the rest. I will preform this analysis with grouping, filtering, and creating a spacial map with ggplot to visualize my analysis.

*Question 2: Second, do a certain race tend to be killed more? Again, using the "police_killings.csv" data, I will group the police killings by race and compare the number of killings by race. Also, I will being doing this for both police killing and police deaths, and I will be visualizing my findings with a bar chart for easy comparison.

*Question 3: Third, Does the size of the task force correspond with the number of police killings? Using the third data set of "police-locals.csv", I will analyze if with more police officers tend to be more police killings. I will create a new variable, setting the new variable with the task force size divided by the amount of police killings and compare that to different cities. This can be answered using a simple bar graph to visualize my analysis.

*Question 4: Fourth, to be decided.