DASC 1104 Project Proposal

Ben Marlow

12/3/2020

1 My blog link

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

```
library(tidyverse)
dat_nfl <- read.csv(file = here::here("data", "nfl_elo_latest.csv"))
dat_mlb <- read.csv(file = here::here("data", "mlb_elo_latest.csv"))
glimpse(dat_nfl)
```

```
## Observations: 269
## Variables: 30
## $ date           <fct> 2020-09-10, 2020-09-13, 2020-09-13, 2020-09-13, 2020-09-13, 2020-09-13, ...
## $ season         <int> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, ...
## $ neutral        <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, ...
## $ playoff        <fct> , , , , , , , , , , , , , , , , , , , , , , , , , , , , , ,
## $ team1          <fct> KC, JAX, CAR, BAL, BUF, MIN, DET, ATL, NE, WSH, CIN, NO, SF, LAR, NYG, D...
## $ team2          <fct> HOU, IND, OAK, CLE, NYJ, GB, CHI, SEA, MIA, PHI, LAC, TB, ARI, DAL, PIT,...
## $ elo1_pre       <dbl> 1664.847, 1438.503, 1417.736, 1638.431, 1511.355, 1571.140, 1404.569, 15...
## $ elo2_pre       <dbl> 1527.930, 1482.655, 1437.326, 1440.533, 1458.063, 1582.459, 1524.565, 15...
## $ elo_prob1      <dbl> 0.7617556, 0.5299664, 0.5649799, 0.8195586, 0.6639485, 0.5766448, 0.4215...
## $ elo_prob2      <dbl> 0.2382444, 0.4700336, 0.4350201, 0.1804414, 0.3360515, 0.4233552, 0.5784...
## $ elo1_post      <dbl> 1676.666, 1457.867, 1399.166, 1649.702, 1526.649, 1543.920, 1391.332, 15...
## $ elo2_post      <dbl> 1516.111, 1463.290, 1455.895, 1429.262, 1442.769, 1609.678, 1537.802, 15...
## $ qbelo1_pre     <dbl> 1651.215, 1392.057, 1416.302, 1628.808, 1532.806, 1544.590, 1437.085, 15...
## $ qbelo2_pre     <dbl> 1497.454, 1518.204, 1461.437, 1499.694, 1451.147, 1555.432, 1527.160, 15...
## $ qb1            <fct> Patrick Mahomes, Gardner Minshew, Teddy Bridgewater, Lamar Jackson, Josh...
## $ qb2            <fct> Deshaun Watson, Philip Rivers, Derek Carr, Baker Mayfield, Sam Darnold, ...
## $ qb1_value_pre  <dbl> 239.69530, 121.75534, 155.17461, 262.05705, 153.17524, 158.20884, 193.20...
## $ qb2_value_pre  <dbl> 195.61581, 155.62454, 181.36357, 130.97056, 134.12095, 176.74031, 140.73...
## $ qb1_adj        <dbl> 6.9428086, -8.6510635, 11.8543782, 17.7484615, 0.6006103, -0.8127627, 15...
## $ qb2_adj        <dbl> 3.6326586, 6.9773865, 4.4939299, -3.5942770, 5.1901076, 0.4903804, -1.57...
## $ qbelo_prob1    <dbl> 0.7519609, 0.3520553, 0.5041818, 0.7432960, 0.6547578, 0.5314521, 0.4441...
## $ qbelo_prob2    <dbl> 0.2480391, 0.6479447, 0.4958182, 0.2567040, 0.3452422, 0.4685479, 0.5558...
## $ qb1_game_value <dbl> 250.96437, 216.97903, 221.33934, 363.49393, 380.75142, 272.75864, 188.31...
## $ qb2_game_value <dbl> 163.113859, 266.985431, 247.800792, 64.243357, 93.647265, 487.232745, 23...
## $ qb1_value_post <dbl> 240.82220, 131.27771, 161.79108, 272.20074, 175.93286, 169.66382, 192.71...
## $ qb2_value_post <dbl> 192.36562, 166.76063, 188.00730, 124.29784, 130.07359, 207.78955, 150.50...
## $ qbelo1_post    <dbl> 1663.567, 1420.368, 1400.051, 1645.369, 1548.567, 1519.870, 1423.036, 14...
## $ qbelo2_post    <dbl> 1485.102, 1489.893, 1477.688, 1483.133, 1435.386, 1580.152, 1541.210, 15...
## $ score1         <int> 34, 27, 30, 38, 27, 34, 23, 25, 21, 27, 13, 34, 20, 20, 16, 14, 35, 28, ...
## $ score2         <int> 20, 20, 34, 6, 17, 43, 27, 38, 11, 17, 16, 23, 24, 17, 26, 16, 30, 11, 3...
```

```
glimpse(dat_mlb)
```

```
## Observations: 951
## Variables: 26
## $ date          <fct> 2020-10-27, 2020-10-25, 2020-10-24, 2020-10-23, 2020-10-21, 2020-10-20, 20...
## $ season        <int> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 20...
## $ neutral       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ playoff       <fct> w, w, w, w, w, w, l, l, l, l, l, l, l, l, l, l, l, l, l, d, d, d, d, d,...
## $ team1         <fct> LAD, TBD, TBD, TBD, LAD, LAD, LAD, TBD, LAD, ATL, TBD, ATL, HOU, HOU, ATL,...
## $ team2         <fct> TBD, LAD, LAD, LAD, TBD, TBD, ATL, HOU, ATL, LAD, HOU, LAD, TBD, TBD, LAD,...
## $ elo1_pre      <dbl> 1607.758, 1566.450, 1563.862, 1567.099, 1608.352, 1604.610, 1602.880, 1564...
## $ elo2_pre      <dbl> 1564.178, 1605.486, 1608.074, 1604.836, 1563.583, 1567.326, 1555.232, 1552...
## $ elo_prob1     <dbl> 0.5828511, 0.4256515, 0.4159707, 0.4280914, 0.5850682, 0.5710566, 0.590421...
## $ elo_prob2     <dbl> 0.4171489, 0.5743485, 0.5840293, 0.5719086, 0.4149318, 0.4289434, 0.409578...
## $ elo1_post     <dbl> 1610.046, 1564.178, 1566.450, 1563.862, 1604.836, 1608.352, 1604.610, 1567...
## $ elo2_post     <dbl> 1561.890, 1607.758, 1605.486, 1608.074, 1567.099, 1563.583, 1553.502, 1550...
## $ rating1_pre   <dbl> 1610.723, 1564.204, 1562.244, 1564.613, 1611.701, 1609.327, 1607.768, 1562...
## $ rating2_pre   <dbl> 1562.805, 1609.323, 1611.284, 1608.914, 1561.827, 1564.200, 1552.173, 1559...
## $ pitcher1      <fct> Tony Gonsolin, Tyler Glasnow, Ryan Yarbrough, Charlie Morton, Tony Gonsoli...
## $ pitcher2      <fct> Blake Snell, Clayton Kershaw, Julio Urias, Walker Buehler, Blake Snell, Ty...
## $ pitcher1_rgs  <dbl> 54.05579, 52.96693, 52.63655, 57.97116, 54.97549, 56.73049, 49.52521, 57.0...
## $ pitcher2_rgs  <dbl> 54.53445, 57.21495, 50.97410, 55.93798, 54.42156, 53.85810, 55.77561, 52.2...
## $ pitcher1_adj  <dbl> 7.9957421, -1.8267342, -3.9595166, 20.5465884, 13.0132222, 21.9364990, -12...
## $ pitcher2_adj  <dbl> 6.0400241, 23.1140964, -6.2048839, 18.0984217, 4.0097101, 0.5290096, 28.11...
## $ rating_prob1  <dbl> 0.5945472, 0.3687158, 0.4111612, 0.4203756, 0.6110917, 0.6208173, 0.528848...
## $ rating_prob2  <dbl> 0.4054528, 0.6312842, 0.5888388, 0.5796244, 0.3889083, 0.3791827, 0.471151...
## $ rating1_post  <dbl> 1612.374, 1562.805, 1564.204, 1562.244, 1608.914, 1611.701, 1609.327, 1564...
## $ rating2_post  <dbl> 1561.153, 1610.723, 1609.323, 1611.284, 1564.613, 1561.827, 1550.614, 1557...
## $ score1        <int> 3, 2, 8, 2, 4, 8, 4, 4, 3, 3, 4, 10, 4, 4, 3, 2, 7, 1, 4, 2, 2, 3, 5, 11, ...
## $ score2        <int> 1, 4, 7, 6, 6, 3, 3, 2, 1, 7, 7, 2, 3, 3, 15, 5, 8, 5, 2, 1, 1, 12, 1, 6, ...
```

2 The Modern Day NFL

For the first half of this project, I'll be examining the NFL Elo dataset contained in the nfl_elo_latest.csv file on the FiveThirtyEight website. The data consist of 269 observations of 30 variables. I obviously won't be focusing on all 30 variables within this dataset but I'll highlight a few that will be key in shaping my analysis. The variables team1 and team2 are the abbreviations for the home and away team. The variables qb1 and qb2 are the names of the home and away starting quarterback. The variables elo_prob1 and elo_prob2 indicate the home and away team's chances of winning according to the predicitve Elo metric. The variables qbelo_prob1 and qbelo_prob2 records the home and away probability of winning based on the quarterback-adjusted Elo rating. Lastly, the variables score_1 and score_2 state the home and away team's final score in each game. Initial exploration shows that the highest rated quarterbacks in the NFL more often than not put more points on the scoreboard and consequently carry their team to victory. There are clearly other factors that have a chance to affect the outcome of a game (weather, injuries), however we will not be including those in this analysis.

3 Is the Starting Pitcher still the King of the Hill?

The second dataset I will be taking a closer look at is the MLB Elo dataset given in the mlb_elo_latest.csv file on the FiveThirtyEight website. The MLB Elo dataset has 951 observations of 26 variables: the variables team1 and team2 are abbreviations for the home and away team, the variables elo_prob1 and elo_prob2

define the home and away team's probability of winning according to Elo ratings, the variables pitcher1 and pitcher2 list the name of the home and away starting pitcher, the variables rating_prob1 and rating_prob2 indicate the home and away team's probability of winning according to team ratings AND starting pitchers, and the variables score1 and score2 provide the number of runs scored for either team in each game.

- Question 1: First, is there a significant gap in a team's win probability with and without the QB? If so, which teams have the biggest margin? To test this, I will generate data visualizations (likely scatterplots) as well as calculate statistics such as the mean, median, and standard deviations of the probability difference.
- Question 2: Second, how has the number of points scored changed over time? To investigate this, I will generate data visualizations of the average amount of points scored in each NFL game over the years. In the visualization I will explore facets and groupings of other variables (like teams) to see if the number of points scored vary with other variables.
- Question 3: Third, has the starting pitcher become more of a detriment to a team's chances of winning rather than of help? First, I will mutate the data to calculate whether the starting pitcher's team won or lost the game they took the mound in. Then, I will produce visualizations combining that outcome with the rating_prob1 and ratingprob_2 variables to see if a correlation exists.
- Question 4: To be determined.