Predictions for 2024-Season Duke Football Attendance

Based on 2011-2023 Home-Game Attendance Records

Overview

This document seeks to utilize attendance records of Duke University home football games from the previous 12 seasons (2011-2023) to predict the number of attendees at Duke football home games during the 2024 season.

In January 2024, initial predictions for 2024 game attendance were created. (These are given in the final section of this document.)

Purpose

The aim of these predictions is to assist in efforts to increase football home-game attendance at Duke University. By using historical game data to predict future game attendance, we can later compare these predictions to actual 2024 attendance figures to determine if a statistically significant improvement in home-game attendance was achieved in the 2024 season.

Packages

```
library(tidyverse)
library(tidymodels)
```

Importing the Dataset

Summary of Duke football opponents at home (Wallace Wade Stadium) from 2011-2023:

```
att_data <- read_csv("data/Duke Stats - DukeAttendanceV3.csv")

att_data <- att_data |>
  mutate(Day = as.factor(Day)) |>
  mutate(Renovated = Rennovated)

home_att_data <- att_data |>
  filter(Site == "Home", Year < 2024)

home_att_data
```

```
# A tibble: 84 x 52
   OppName       OppFPI DukeFPI FPI_diff DukeFPI_NetChange OppFPI_PrevYear
   <chr>          <dbl>   <dbl>    <dbl>              <dbl>           <dbl>
 1 Richmond        NA      -6.1     NA                -2.1            NA
 2 Stanford        24.4    -6.1     30.5              -2.1            24.2
 3 Tulane         -20.3    -6.1    -14.2              -2.1           -17.3
 4 Florida St.     15.3    -6.1     21.4              -2.1            17.2
 5 Wake Forest     -0.2    -6.1      5.9              -2.1            -6
 6 Virginia Tech   11.8    -6.1     17.9              -2.1            18.4
 7 Georgia Tech     5      -6.1     11.1              -2.1             5.3
 8 Florida Int'l   -8      -1.7     -6.3               4.4            -5.1
 9 N.C. Central    NA      -1.7     NA                 4.4            NA
10 Memphis        -13.2    -1.7    -11.5               4.4           -24.6
# i 74 more rows
# i 46 more variables: FPI_Diff_PrevYear <dbl>, Surface <chr>, Month <dbl>,
#   Date <dbl>, Year <dbl>, Day <fct>, Start_Time <dbl>, Site <chr>,
#   Result <chr>, DukePts <dbl>, OppPts <dbl>, PointDiff <dbl>, AttNum <dbl>,
#   AttPct <dbl>, ESPN_WinPred <dbl>, COVID_Limit <lgl>, Rain <lgl>,
#   City <chr>, State <chr>, TV_Coverage <chr>, Bowl <lgl>,
#   DukeRankGametime <dbl>, OppRankGametime <dbl>, OppRankSeasonEnd <dbl>, ...
```

List of Duke football opponents at home (Wallace Wade Stadium) in *2024:*

```
att_data |>
  filter(Site == "Home", Year == 2024) |>
  summarize("Opponent Name" = OppName)
```

```
# A tibble: 6 x 1
  `Opponent Name`
  <chr>
1 Elon
2 Connecticut
3 Florida St.
4 North Carolina
5 SMU
6 Virginia Tech
```

```
home_opp_list <- c("Elon", "Connecticut", "Florida St.",
                   "North Carolina", "SMU", "Virginia Tech")
```

History of At-Home Attendance for 2024 Opponents

Duke faces against 6 opponents at home in 2024. This section shows every game Duke has played *at home* against these 6 opponents from 2011 through 2023.

It is worth noting that Wallace Wade Stadium attendance capacity changed as a result of renovations which completed in 2016:

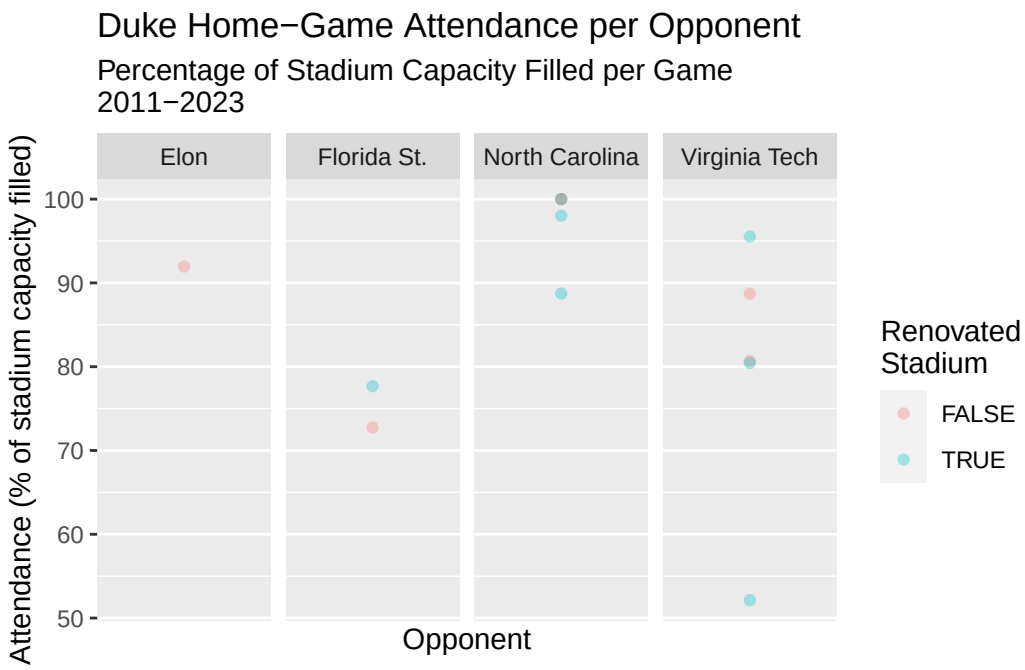
- *Pre-renovation capacity:* 33,941 (1982-2015)
- *Post-renovation capacity:* 40,004 (2016-present)

Whether a game occurred before or after these renovations is often denoted by color (in this section).

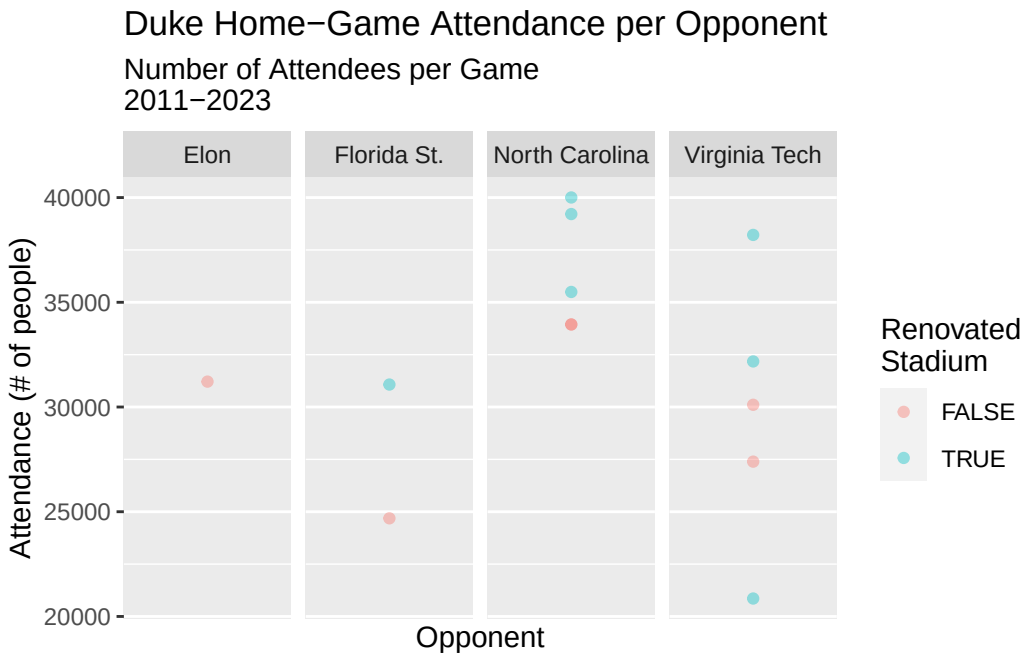
All Teams

```
home_att_data |>
  filter(OppName %in% home_opp_list) |>
  ggplot(
    aes(x = 0, y = AttPct, color = Renovated)
  ) +
  geom_point(alpha = 0.333) +
  facet_wrap(~OppName, strip.position = "top", nrow = 1) +
  scale_x_continuous(labels = NULL, breaks = NULL) +
  labs(title = "Duke Home-Game Attendance per Opponent",
       subtitle = "Percentage of Stadium Capacity Filled per Game\n2011-2023",
```

```
       x = "Opponent",
       y = "Attendance (% of stadium capacity filled)",
       color = "Renovated\nStadium")
```




```
home_att_data |>
  filter(OppName %in% home_opp_list) |>
  ggplot(
    aes(x = 0, y = AttNum, color = Renovated)
  ) +
  geom_point(alpha = 0.4) +
  facet_wrap(~OppName, strip.position = "top", nrow = 1) +
  scale_x_continuous(labels = NULL, breaks = NULL) +
  labs(title = "Duke Home-Game Attendance per Opponent",
       subtitle = "Number of Attendees per Game\n2011-2023",
       x = "Opponent",
       y = "Attendance (# of people)",
       color = "Renovated\nStadium")
```



Elon

```
home_att_data |>
  filter(OppName == "Elon") |>
  summarize("Name" = OppName,
            "End-of-Season FPI" = OppFPI,
            Month,
            Date,
            Year,
            "# of Attendees" = AttNum,
            "% of Stadium Capacity Filled" = AttPct)
```

```
# A tibble: 1 x 7
  Name  `End-of-Season FPI` Month  Date  Year `# of Attendees`
  <chr>               <dbl> <dbl> <dbl> <dbl>            <dbl>
1 Elon                   NA     8    30  2014            31213
# i 1 more variable: `% of Stadium Capacity Filled` <dbl>
```

Connecticut

```
home_att_data |>
  filter(OppName == "Connecticut") |>
  summarize("Name" = OppName,
            "End-of-Season FPI" = OppFPI,
            Month,
            Date,
            Year,
            "# of Attendees" = AttNum,
            "% of Stadium Capacity Filled" = AttPct)
```

```
# A tibble: 0 x 7
# i 7 variables: Name <chr>, End-of-Season FPI <dbl>, Month <dbl>, Date <dbl>,
#   Year <dbl>, # of Attendees <dbl>, % of Stadium Capacity Filled <dbl>
```

UConn never faced against Duke in Wallace Wade Stadium from 2011 to 2023.

Florida St.

```
home_att_data |>
  filter(OppName == "Florida St.") |>
  summarize("Name" = OppName,
            "End-of-Season FPI" = OppFPI,
            Month,
            Date,
            Year,
            "# of Attendees" = AttNum,
            "% of Stadium Capacity Filled" = AttPct)
```

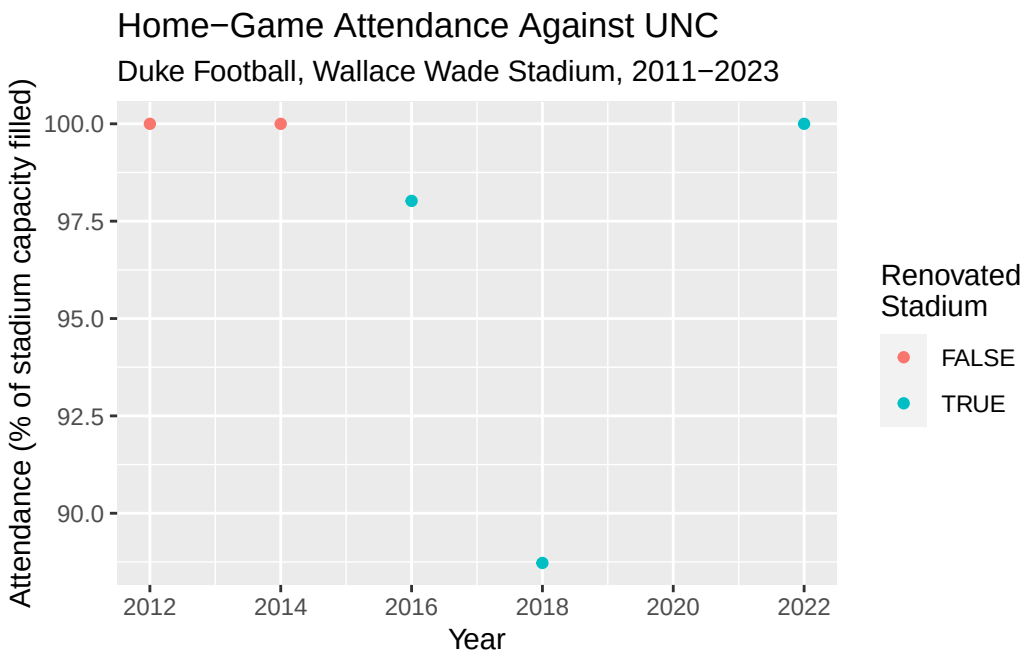
```
# A tibble: 2 x 7
  Name        `End-of-Season FPI` Month  Date  Year `# of Attendees`
  <chr>                     <dbl> <dbl> <dbl> <dbl>            <dbl>
1 Florida St.                15.3    10    15  2011            24687
2 Florida St.                13.3    10    14  2017            31073
# i 1 more variable: `% of Stadium Capacity Filled` <dbl>
```

North Carolina

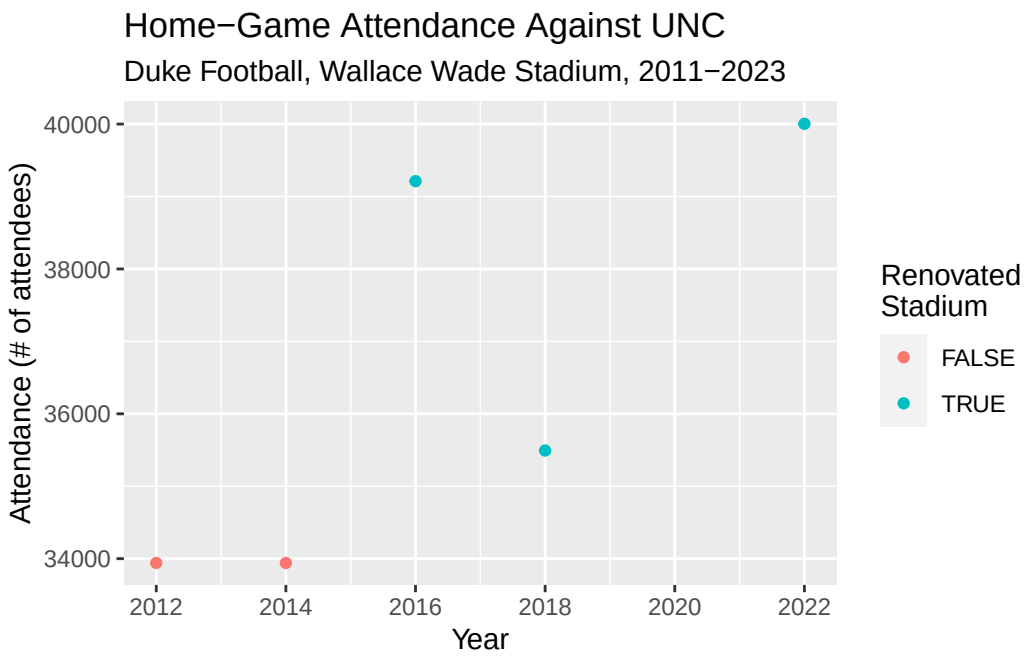
```
home_att_data |>
  filter(OppName == "North Carolina") |>
  summarize("Name" = OppName,
            "End-of-Season FPI" = OppFPI,
            Month,
            Date,
            Year,
            "# of Attendees" = AttNum,
            "% of Stadium Capacity Filled" = AttPct)
```

```
# A tibble: 6 x 7
  Name           `End-of-Season FPI` Month  Date  Year `# of Attendees`
  <chr>                        <dbl> <dbl> <dbl> <dbl>            <dbl>
1 North Carolina                10.6    10    20  2012            33941
2 North Carolina                 4.4    11    20  2014            33941
3 North Carolina                14      11    10  2016            39212
4 North Carolina                -2.6    11    10  2018            35493
5 North Carolina                10.2    11     7  2020               NA
6 North Carolina                 6.2    10    15  2022            40004
# i 1 more variable: `% of Stadium Capacity Filled` <dbl>
```

```
home_att_data |>
  filter(OppName == "North Carolina") |>
  ggplot(
    aes(x = Year, y = AttPct, color = Renovated)
  ) +
  geom_point() +
  scale_x_continuous(breaks = seq(from = 2012, to = 2023, by = 2)) +
  labs(title = "Home-Game Attendance Against UNC",
       subtitle = "Duke Football, Wallace Wade Stadium, 2011-2023",
       x = "Year",
       y = "Attendance (% of stadium capacity filled)",
       color = "Renovated\nStadium")
```




```
home_att_data |>
  filter(OppName == "North Carolina") |>
  ggplot(
    aes(x = Year, y = AttNum, color = Renovated)
  ) +
  geom_point() +
  scale_x_continuous(breaks = seq(from = 2012, to = 2023, by = 2)) +
  labs(title = "Home-Game Attendance Against UNC",
       subtitle = "Duke Football, Wallace Wade Stadium, 2011-2023",
       x = "Year",
       y = "Attendance (# of attendees)",
       color = "Renovated\nStadium")
```



SMU

```
home_att_data |>
  filter(OppName == "SMU") |>
  summarize("Name" = OppName,
            "End-of-Season FPI" = OppFPI,
            Month,
            Date,
            Year,
            "# of Attendees" = AttNum,
            "% of Stadium Capacity Filled" = AttPct)
```

```
# A tibble: 0 x 7
# i 7 variables: Name <chr>, End-of-Season FPI <dbl>, Month <dbl>, Date <dbl>,
#   Year <dbl>, # of Attendees <dbl>, % of Stadium Capacity Filled <dbl>
```

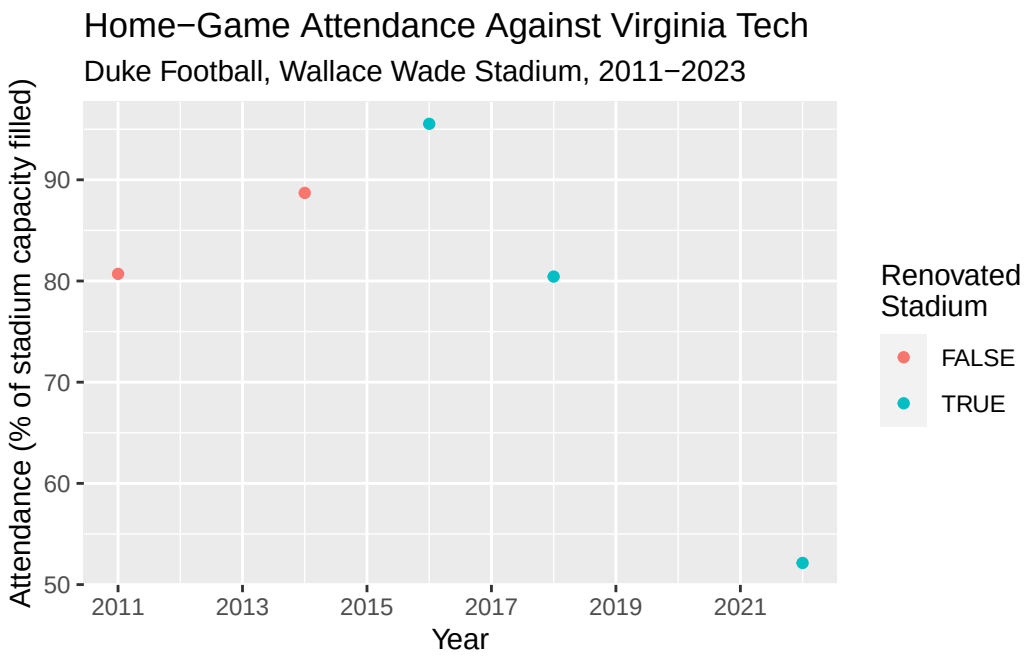
UConn never faced against Duke in Wallace Wade Stadium from 2011 to 2023.

Virginia Tech

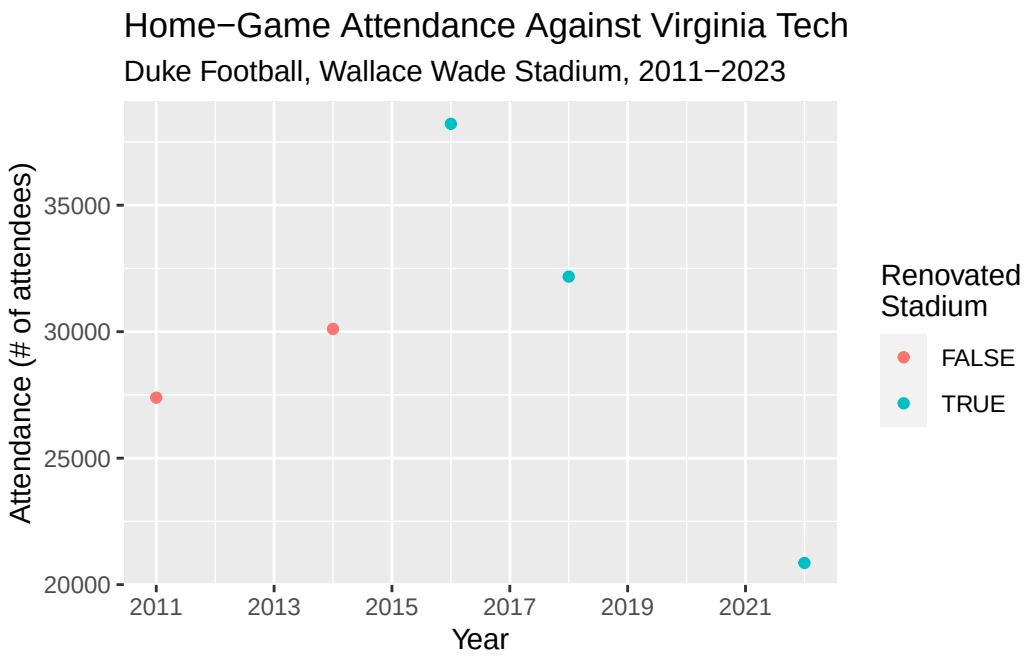
```
home_att_data |>
  filter(OppName == "Virginia Tech") |>
  summarize("Name" = OppName,
            "End-of-Season FPI" = OppFPI,
            Month,
            Date,
            Year,
            "# of Attendees" = AttNum,
            "% of Stadium Capacity" = AttPct)
```

```
# A tibble: 6 x 7
  Name          `End-of-Season FPI` Month  Date  Year `# of Attendees`
  <chr>                       <dbl> <dbl> <dbl> <dbl>            <dbl>
1 Virginia Tech                11.8    10    29  2011            27392
2 Virginia Tech                 7.9    11    15  2014            30107
3 Virginia Tech                13.7    11     5  2016            38217
4 Virginia Tech                 3.4     9    29  2018            32177
5 Virginia Tech                 7.3    10     3  2020               NA
6 Virginia Tech                -6.2    11    12  2022            20857
# i 1 more variable: `% of Stadium Capacity` <dbl>
```

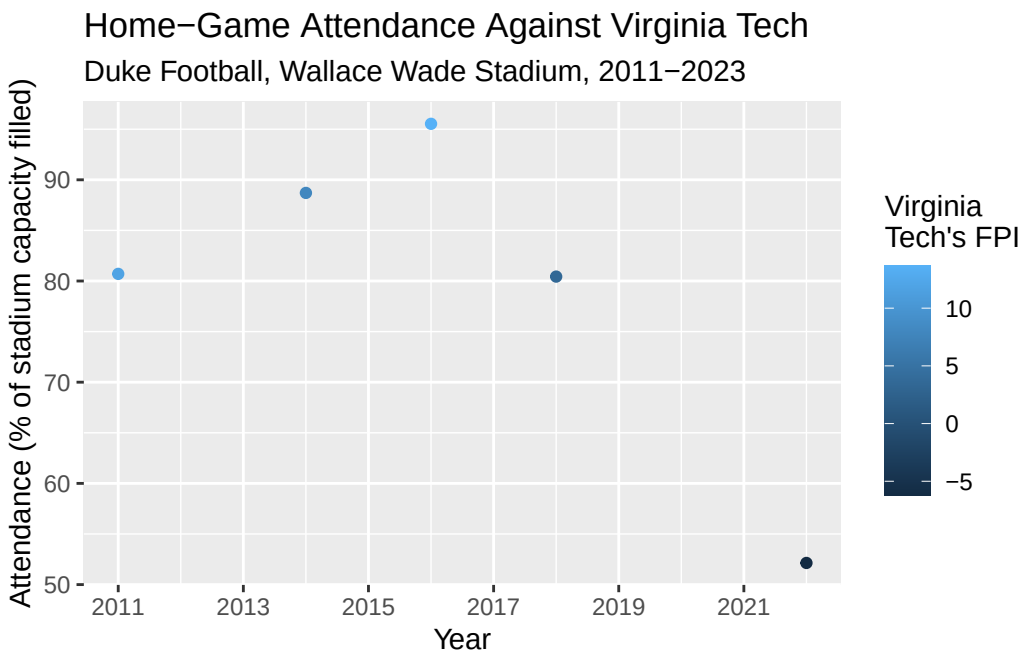
```
home_att_data |>
  filter(OppName == "Virginia Tech") |>
  ggplot(
    aes(x = Year, y = AttPct, color = Renovated)
  ) +
  geom_point() +
  scale_x_continuous(breaks = seq(from = 2011, to = 2023, by = 2)) +
  labs(title = "Home-Game Attendance Against Virginia Tech",
       subtitle = "Duke Football, Wallace Wade Stadium, 2011-2023",
       x = "Year",
       y = "Attendance (% of stadium capacity filled)",
       color = "Renovated\nStadium")
```




```
home_att_data |>
  filter(OppName == "Virginia Tech") |>
  ggplot(
    aes(x = Year, y = AttNum, color = Renovated)
  ) +
  geom_point() +
  scale_x_continuous(breaks = seq(from = 2011, to = 2023, by = 2)) +
  labs(title = "Home-Game Attendance Against Virginia Tech",
       subtitle = "Duke Football, Wallace Wade Stadium, 2011-2023",
       x = "Year",
       y = "Attendance (# of attendees)",
       color = "Renovated\nStadium")
```




```
home_att_data |>
  filter(OppName == "Virginia Tech") |>
  ggplot(
    aes(x = Year, y = AttPct, color = OppFPI)
  ) +
  geom_point() +
  scale_x_continuous(breaks = seq(from = 2011, to = 2023, by = 2)) +
  labs(title = "Home-Game Attendance Against Virginia Tech",
       subtitle = "Duke Football, Wallace Wade Stadium, 2011-2023",
       x = "Year",
       y = "Attendance (% of stadium capacity filled)",
       color = "Virginia\nTech's FPI")
```



Team Performance vs. Attendance

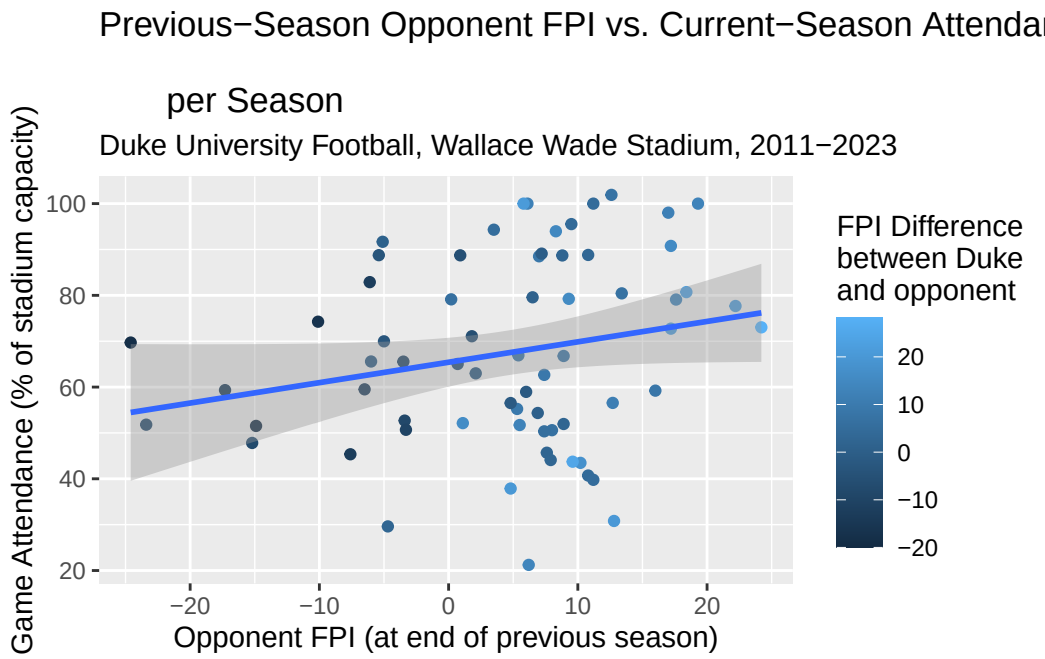
Can football team performance – both of Duke and its opponent – be used to predict the attendance turnout of future Duke home games?

Previous-Season FPI

This section will seek to determine if the Football Power Index (FPI) of an opposing team at the end of one season is a decent predictor of home-game audience turnout in the *following* season.

```
home_att_data_prevFPI <- home_att_data |>
  filter(!is.na(OppFPI_PrevYear)) |>
  mutate(OppFPI_PrevYear = OppFPI_PrevYear,
         FPI_Diff_PrevYear = FPI_Diff_PrevYear)

home_att_data_prevFPI |>
  ggplot(
    aes(x = OppFPI_PrevYear, y = AttPct, color = FPI_Diff_PrevYear)
  ) +
  geom_point() +
```

```
  geom_smooth(method = "lm") +
  labs(title = "Previous-Season Opponent FPI vs. Current-Season Attendance,\n
       per Season",
       subtitle = "Duke University Football, Wallace Wade Stadium, 2011-2023",
       color = "FPI Difference\nbetween Duke\nand opponent",
       x = "Opponent FPI (at end of previous season)",
       y = "Game Attendance (% of stadium capacity)")
```




```
prev_fpi_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear, data = home_att_data_prevFPI)

tidy(prev_fpi_lm)
```

```
# A tibble: 2 x 5
  term            estimate std.error statistic  p.value
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)       65.4       2.67      24.5  1.67e-34
2 OppFPI_PrevYear    0.445     0.242      1.84 7.08e- 2
```

```
glance(prev_fpi_lm)$adj.r.squared
```

```
[1] 0.03473927
```

The scatterplot above shows a fairly weak yet positive correlation between home-game attendance and the FPI of the opponent at the end of the previous season.

The linear model gives the slope of the linear fit depicted in the scatterplot. The model gives a slope of roughly 0.44497, which signifies that for every 1-point increase in the opponent's previous-season FPI, stadium attendance (as a percentage of Wallace Wade's total capacity) is predicted to increase by 0.44497% on average. The model indicates that this slope has a p-value of about 0.071, which is less than 0.1 and is significant given the difficulty of predicting future football attendance.

The adjusted r-squared value of about 0.0347 is very low, indicating that while a positive correlation is likely between attendance and opponent previous-season FPI, attendance is likely to also be based on other factors.

Previous-Season FPI Difference Between Duke & Opponent

```
prev_fpi_diff_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear + FPI_Diff_PrevYear, data = home_att_data_prevFPI)

tidy(prev_fpi_diff_lm)
```

```
# A tibble: 3 x 5
  term              estimate std.error statistic  p.value
  <chr>                <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)         65.5       2.66      24.6  2.52e-34
2 OppFPI_PrevYear      0.843     0.401      2.10 3.94e- 2
3 FPI_Diff_PrevYear   -0.455     0.366     -1.24 2.18e- 1
```

```
glance(prev_fpi_diff_lm)$adj.r.squared
```

```
[1] 0.0427744
```

When additively considering the FPI difference between Duke and its opponent at the end of the season *before* a game, the model gives a slope of roughly 0.843, which signifies that for every increase of 1 in the opponent's previous-season FPI, stadium attendance (as a percentage of Wallace Wade's total capacity) is predicted to increase by 0.843% on average. This is greater than the previous model, and this slope is also more significant ($p = 0.0394$).

Additionally, this model indicates that when the difference in previous-season FPI increases between Duke and its opponent increases (AKA when a matchup is more difficult for Duke based on the previous-season teams), stadium attendance decreases. However, the p-value for this is roughly 0.2183, suggesting that this trend may be due to chance rather than this association truly existing overall.

The adjusted r-squared value of this model is higher than the previous, suggesting that when you consider the FPI difference in addition to the opponent team's FPI, the model better predicts variation in stadium attendance. Thus, we *will* be including the First_Home_Game variable in future models.

Win History

Does the previous recent winning record of a team matter for a game's attendance level?

Duke Undefeated Status

The following models will investigate if whether Duke being undefeated in a season – both undefeated at home and undefeated overall – is related to stadium attendance:

```
prev_fpi_diff_undef_home_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear + FPI_Diff_PrevYear + Undefeated_Home,
      data = home_att_data_prevFPI)

tidy(prev_fpi_diff_undef_home_lm)
```

```
# A tibble: 4 x 5
  term                estimate std.error statistic  p.value
  <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)           62.8       3.56      17.6  4.72e-26
2 OppFPI_PrevYear        0.950     0.411      2.31 2.40e- 2
3 FPI_Diff_PrevYear     -0.489     0.366     -1.33 1.87e- 1
4 Undefeated_HomeTRUE    5.90      5.15       1.15 2.56e- 1
```

```
glance(prev_fpi_diff_undef_home_lm)$adj.r.squared
```

```
[1] 0.04746587
```

```
prev_fpi_diff_undef_overall_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear + FPI_Diff_PrevYear + Undefeated_All,
      data = home_att_data_prevFPI)

tidy(prev_fpi_diff_undef_overall_lm)
```

```
# A tibble: 4 x 5
  term               estimate std.error statistic  p.value
  <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)          64.8       3.08     21.0   3.39e-30
2 OppFPI_PrevYear       0.827     0.405     2.04  4.53e- 2
3 FPI_Diff_PrevYear    -0.419     0.376    -1.11  2.69e- 1
4 Undefeated_AllTRUE    2.88      6.04      0.476 6.35e- 1
```

```
glance(prev_fpi_diff_undef_overall_lm)$adj.r.squared
```

```
[1] 0.03107152
```

When considering whether a team is undefeated overall, the result is not significant and results in a lower adjusted R-squared value for the model. However, whether a team is undefeated *at home* does improve the adjusted R-squared value of the model from 0.04277 to 0.04746. The model estimates that stadium attendance slightly *increases* when Duke is undefeated on its home field in a season, but this result is not statistically significant ($p = 0.2558$).

Duke Win Streak

Does Duke being on a win streak affect stadium attendance?

```
prev_fpi_diff_streak_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear  + FPI_Diff_PrevYear + Win_Streak,
      data = home_att_data_prevFPI)

tidy(prev_fpi_diff_streak_lm)
```

```
# A tibble: 4 x 5
  term              estimate std.error statistic  p.value
  <chr>                <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)         62.4       3.16     19.7   1.22e-28
2 OppFPI_PrevYear      0.704     0.402     1.75  8.49e- 2
3 FPI_Diff_PrevYear   -0.300     0.370    -0.811 4.21e- 1
4 Win_Streak           2.61      1.47      1.77  8.10e- 2
```

```
glance(prev_fpi_diff_streak_lm)$adj.r.squared
```

```
[1] 0.07380137
```

```
prev_fpi_streak_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear  +  Win_Streak,
      data = home_att_data_prevFPI)

tidy(prev_fpi_streak_lm)
```

```
# A tibble: 3 x 5
  term            estimate std.error statistic  p.value
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)       62.0       3.11      19.9  4.18e-29
2 OppFPI_PrevYear    0.441     0.237      1.86 6.71e- 2
3 Win_Streak         2.90      1.43       2.03 4.69e- 2
```

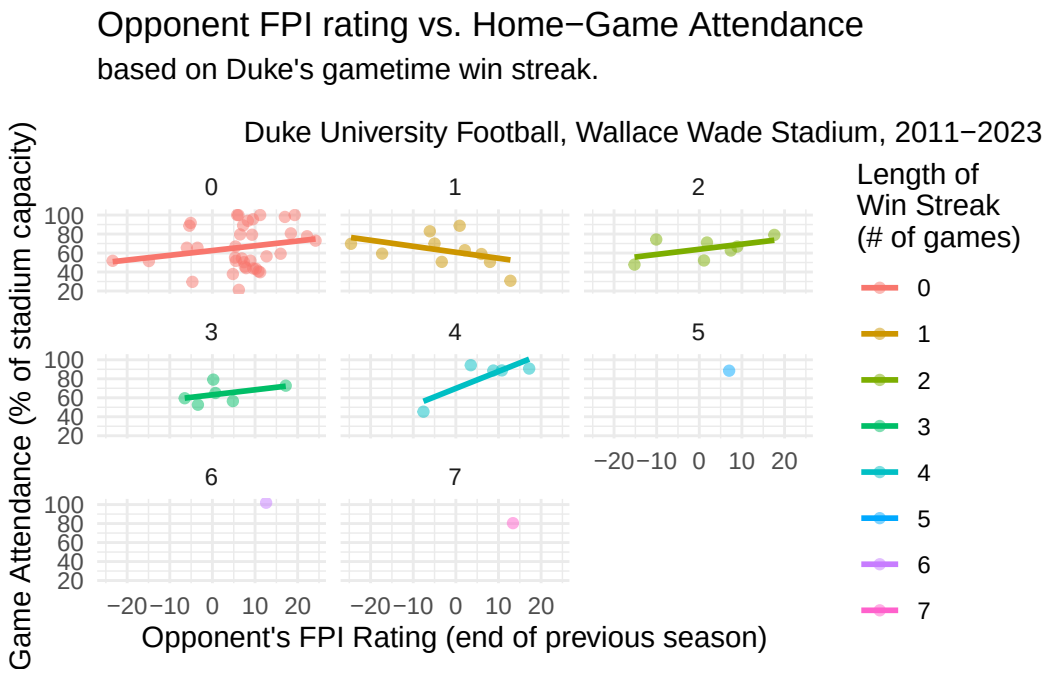
```
glance(prev_fpi_streak_lm)$adj.r.squared
```

```
[1] 0.07876475
```

Factoring in Duke's win streak greatly improves the predictive power of the model. In fact, when the FPI difference between Duke and its opponent is removed, the model becomes even more representative, as the adjusted R-squared value increases to 0.07876 and the p-value of both terms nears 0.05.

This is a strong indication that Duke's win streak performance greatly affects stadium attendance. A visual representation of attendance based on win streak is shown below:

```
home_att_data_prevFPI |>
  mutate(Win_Streak = as.factor(Win_Streak)) |>
  ggplot(aes(x = OppFPI_PrevYear, y = AttPct, color = Win_Streak)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, alpha = 0.5) +
  facet_wrap(~ Win_Streak) +
  theme_minimal() +
  labs(title = "Opponent FPI rating vs. Home-Game Attendance",
       subtitle = "based on Duke's gametime win streak.\n
                   Duke University Football, Wallace Wade Stadium, 2011-2023",
       x = "Opponent's FPI Rating (end of previous season)",
       y = "Game Attendance (% of stadium capacity)",
       color = "Length of\nWin Streak\n(# of games)")
```

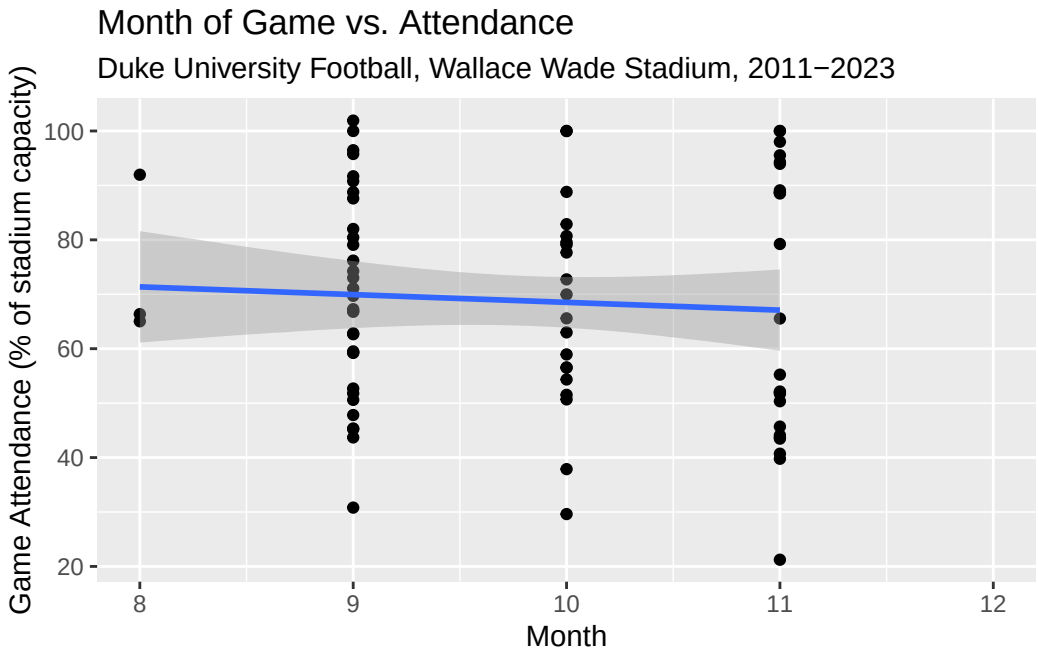


Time

Do factors related to *when* a game takes place – month, day of the week, etc. – affect our ability to predict future games' attendance?

Month

```
home_att_data |>
  ggplot(
    aes(x = Month, y = AttPct)
  ) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Month of Game vs. Attendance",
       subtitle = "Duke University Football, Wallace Wade Stadium, 2011-2023",
       x = "Month",
       y = "Game Attendance (% of stadium capacity)")
```



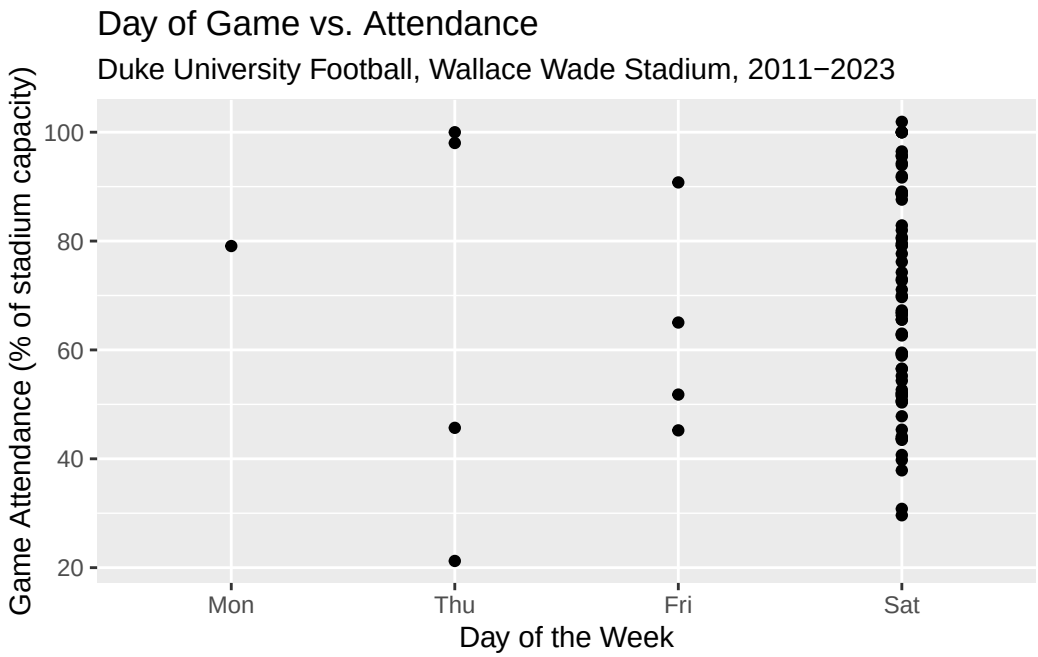
Based on the scatterplot above, no obvious correlation is present between game month and attendance for Duke home games. The spread of attendance percentages for games appears independent of the month on which a game occurs.

Day of Week

```r
home_att_data |>
  mutate('Saturday Game' = if_else(Day == "Sat", TRUE, FALSE)) |>
  filter(!is.na(AttPct)) |>
  group_by(`Saturday Game`) |>
  summarize("Median Attendance %" = median(AttPct),
            "SD of Attendance %" = sd(AttPct))
```

```
# A tibble: 2 x 3
  `Saturday Game` `Median Attendance %` `SD of Attendance %`
  <lgl>                           <dbl>                <dbl>
1 FALSE                            65.0                 27.4
2 TRUE                             67.1                 19.5
```

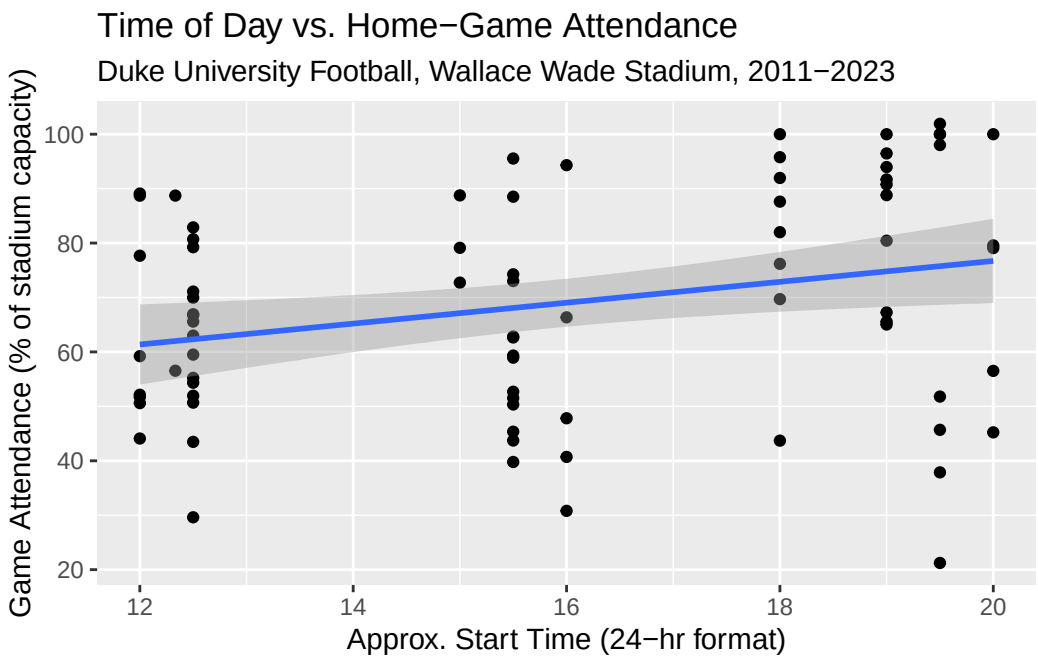
```r
home_att_data |>
  ggplot(
    aes(x = fct_relevel(Day, "Mon", "Thu", "Fri", "Sat"),
        y = AttPct)
  ) +
  geom_point() +
  labs(title = "Day of Game vs. Attendance",
       subtitle = "Duke University Football, Wallace Wade Stadium, 2011-2023",
       x = "Day of the Week",
       y = "Game Attendance (% of stadium capacity)")
```



Based on the metrics and scatterplot above, no obvious correlation is present between game month and attendance for Duke home games. The spread of attendance percentages for games appears independent of whether a game occurs on the usual day (Saturday) or not.

Time of Day

```
home_att_data |>
  ggplot(
    aes(x = Start_Time, y = AttPct)
  ) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Time of Day vs. Home-Game Attendance",
       subtitle = "Duke University Football, Wallace Wade Stadium, 2011-2023",
       x = "Approx. Start Time (24-hr format)",
       y = "Game Attendance (% of stadium capacity)")
```



Based on the scatterplot above, it appears that a slight positive correlation may appear between the start time of a game and the attendance percentage.

This time-of-day variable is added to previous model below:

```
prev_fpi_streak_time_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear  +  Win_Streak + Start_Time,
      data = home_att_data_prevFPI)

tidy(prev_fpi_streak_time_lm)
```

```
# A tibble: 4 x 5
  term            estimate std.error statistic p.value
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)       38.8      12.6        3.09 0.00300
2 OppFPI_PrevYear    0.432     0.232      1.86 0.0672
3 Win_Streak         2.46      1.42       1.73 0.0884
4 Start_Time         1.53      0.807      1.90 0.0625
```

```
glance(prev_fpi_streak_time_lm)$adj.r.squared
```

```
[1] 0.1146884
```

Out of all models tested in this document so far, this model – which includes previous-year opponent FPI, Duke win streak, and time-of-day as predictor variables – **is the best at predicting home game attendance thus far.**

This model has an adjusted R-squared value of approximately 0.11469, which is higher than all previous models. This suggests that including game start time *improves* the model's predictive power. Additionally, all predictor variables had a p-value < 0.1, which is good within this context and suggests a low likelihood that the trends observed in this model occurred by chance alone.

The *beta* of the Start_Time variable was around 1.531, suggesting that for every 1 hour later that the game start time is, the stadium attendance percentage (as a percentage of total stadium capacity) is predicted to increase on average by about 1.531 percentage points.

However, this does not mean that later start times *cause* greater attendance. Often, games are scheduled for a later hour when they are expected to be more popular, such as during prime-time. It is thus unlikely that many games which TV/sporting organizers expect to have large crowds will be during earlier daylight hours.

Other Factors

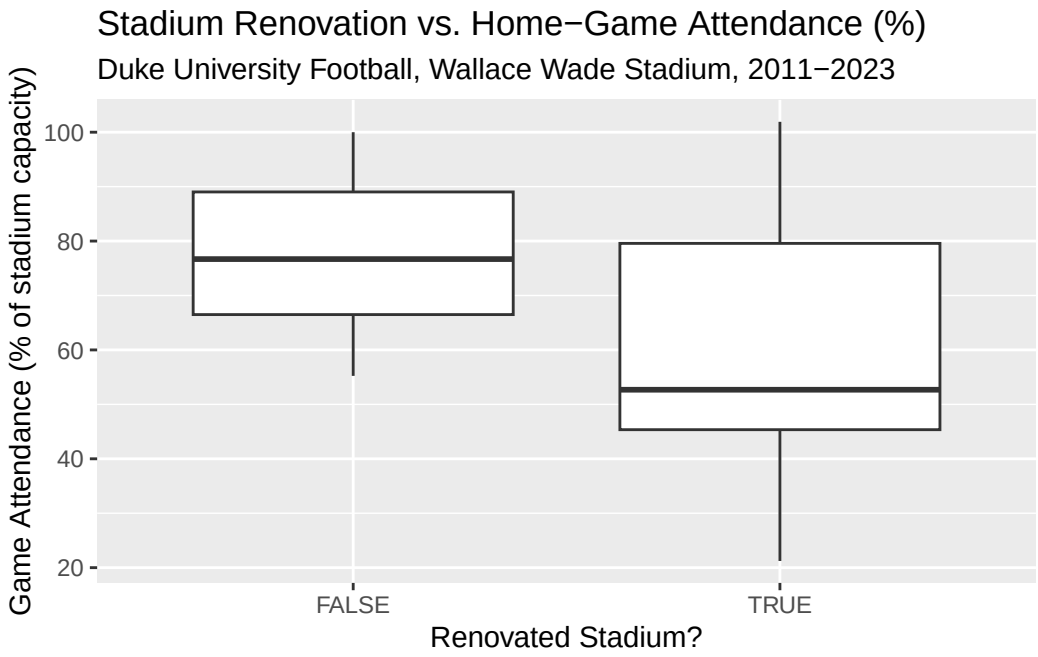
Stadium Renovation

Wallace Wade Stadium capacity:

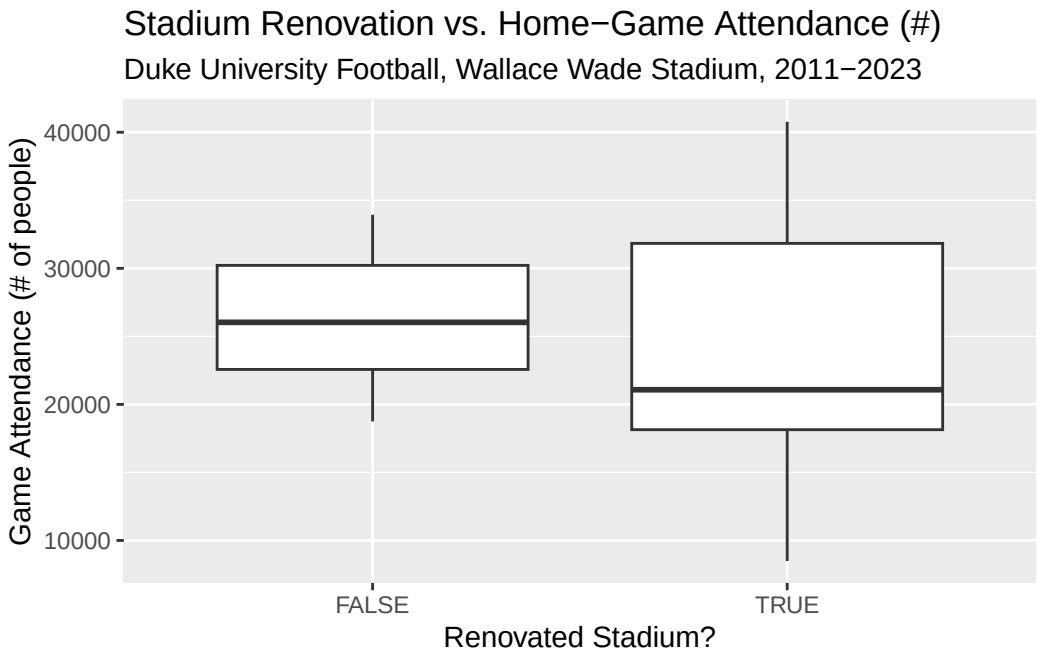
- *Pre-renovation:* 33,941 (1982-2015)
- *Post-renovation:* 40,004 (2016-present)

Does this renovation factor relate to gametime attendance?

```
home_att_data |>
  ggplot(
    aes(x = Renovated, y = AttPct)
  ) +
  geom_boxplot() +
  labs(title = "Stadium Renovation vs. Home-Game Attendance (%)",
       subtitle = "Duke University Football, Wallace Wade Stadium, 2011-2023",
       x = "Renovated Stadium?",
       y = "Game Attendance (% of stadium capacity)")
```




```
home_att_data |>
  ggplot(
    aes(x = Renovated, y = AttNum)
  ) +
  geom_boxplot() +
  labs(title = "Stadium Renovation vs. Home-Game Attendance (#)",
       subtitle = "Duke University Football, Wallace Wade Stadium, 2011-2023",
       x = "Renovated Stadium?",
       y = "Game Attendance (# of people)")
```



Based on the first plot (showing attendance *percentage)*, it is evident that Wallace Wade Stadium tended to reach closer to full capacity before the stadium was renovated than after the renovation. The second plot (showing attendance *count)* indicates that even after stadium renovation, attendance counts did not significantly increase below the 75th percentile of all games. Both plots indicate that in the years after the stadium was renovated, the spread of attendance values increased – stadium attendance varied more greatly from the median.

This does not indicate that stadium renovations directly *caused* a decrease in median attendance (both in terms of percentage-of-capacity and actual count). Other factors may have been at play, such as there being less data included in this research from before the 2016 renovation (2011-16 includes less games/seasons than 2016-2023, as well as a decline in Duke football performance during the period following stadium renovations (2016-2021).

Below, we include the stadium renovation variable within our previous model (opponent previous-year FPI, Duke win streak, game start time):

```
prev_fpi_streak_time_renovated_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear  +  Win_Streak + Start_Time + Renovated,
      data = home_att_data_prevFPI)

tidy(prev_fpi_streak_time_renovated_lm)
```

```
# A tibble: 5 x 5
```

```
  term            estimate std.error statistic   p.value
  <chr>              <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)       42.1      10.9        3.86 0.000269
2 OppFPI_PrevYear    0.564     0.202      2.78 0.00712
3 Win_Streak         2.06      1.23       1.67 0.0997
4 Start_Time         2.03      0.706      2.88 0.00543
5 RenovatedTRUE    -19.7       4.17      -4.72 0.0000137
```

```
glance(prev_fpi_streak_time_renovated_lm)$adj.r.squared
```

```
[1] 0.3384521
```

In this model, the adjusted R-squared value is **greatly**improved from the previous model. The *beta* of the stadium renovation variable is -19.691, suggesting that after the stadium renovation, attendance percentage (as a percent of total stadium capacity) *decreased* on average by about 19.691 percentage points. Additionally, the p-value of this stadium renovation *beta* is less than 0.001. This is a very strong indicator that games after stadium renovations should be predicted to have a lower attendance percentage (out of full stadium capacity) than games before stadium renovations.

Since all games we will be predicting in future seasons (i.e. 2024) will have taken place after the 2016 renovation, this renovation factor will be an important factor to include in any future prediction models that are based on past Wallace Wade Stadium attendance.

New Head Coach

Duke has a new head coach in its 2024 season. Does home-game attendance seem to change during the first season a new head coach is present, based on data from 2011-2023?

```
prev_fpi_diff_coach_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear + FPI_Diff_PrevYear + New_Coach,
      data = home_att_data_prevFPI)

tidy(prev_fpi_diff_coach_lm)
```

```
# A tibble: 4 x 5
  term              estimate std.error statistic  p.value
  <chr>                <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)         65.8       2.87     22.9   3.16e-32
2 OppFPI_PrevYear      0.776     0.508     1.53  1.32e- 1
```

```
3 FPI_Diff_PrevYear   -0.389     0.477    -0.816 4.17e- 1
4 New_CoachTRUE       -2.65     12.2      -0.218 8.28e- 1
```

```
glance(prev_fpi_diff_coach_lm)$adj.r.squared
```

```
[1] 0.02831449
```

The adjusted r-squared value of the model decreases when the coaching variable is introduced, and the p-values become less significant. This suggests that simply having a new head coach does *not* affect home-game attendance. Thus, we will not be including the New_Coach variable in future models.

First Home Game

Does home-game attendance tend to differ when it is the first home game of the season?

```
prev_fpi_diff_first_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear + FPI_Diff_PrevYear + First_Home_Game,
      data = home_att_data_prevFPI)

tidy(prev_fpi_diff_first_lm)
```

```
# A tibble: 4 x 5
  term                estimate std.error statistic  p.value
  <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)           64.8       2.79     23.2   1.35e-32
2 OppFPI_PrevYear        0.900     0.407     2.21  3.05e- 2
3 FPI_Diff_PrevYear     -0.476     0.367    -1.30  1.99e- 1
4 First_Home_GameTRUE    9.26     10.5       0.886 3.79e- 1
```

```
glance(prev_fpi_diff_first_lm)$adj.r.squared
```

```
[1] 0.0395352
```

The adjusted r-squared value of the model decreases when the First_Home_Game variable is introduced, and the p-values become less significant. This suggests that a game being the *first* home game does *not* affect stadium attendance. Thus, we will not be including the First_Home_Game variable in future models.

UNC Game

Since Duke vs. UNC is a historic rivalry, we will investigate: can a model better predict home-game attendance when it accounts for whether or not UNC is the opponent?

```
prev_fpi_diff_unc_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear + FPI_Diff_PrevYear + UNC_Game,
      data = home_att_data_prevFPI)

tidy(prev_fpi_diff_unc_lm)
```

```
# A tibble: 4 x 5
  term              estimate std.error statistic  p.value
  <chr>                <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)         63.6       2.49      25.6  5.71e-35
2 OppFPI_PrevYear      0.820     0.366      2.24 2.86e- 2
3 FPI_Diff_PrevYear   -0.526     0.334     -1.57 1.20e- 1
4 UNC_GameTRUE        31.7       8.50       3.73 4.17e- 4
```

```
glance(prev_fpi_diff_unc_lm)$adj.r.squared
```

```
[1] 0.2032694
```

While the p-values were improved in this model, the adjusted R-squared value decreased, suggesting that the inclusion of the UNC variable is unnecessary. However, this model is still worth noting, since it shows that the filled percentage of total stadium capacity typically increases by around 31.67 when a game is against UNC, and while this exact percentage can vary, this is a strongly statistically significant ($p < 0.001$) trend.

However, since the adjusted R-squared value of the model decreased as a result of adding the UNC variable, we will not be including the UNC variable it in future models.

2024-Season Attendance Predictions

January 2024 Predictions

As of this month (January 2024), some information is not yet available about 2024-season games, such as the time of day and Duke's win-streak standing per game. Many of these factors are to be dynamically determined based on Duke's performance in the 2024 season.

However, it is still possible to loosely predict how attendance may look at future games based on the performance of both teams in the 2023 season.

Out of all the factors we examined within the previous plots and linear regression models in this document, only 3 that were found to have substantial predictive power have the potential to predict 2024-season games at this time:

1. The FPI of an opponent in their *previous* season
2. The difference between Duke's previous-season FPI and the opponent's previous-season FPI
3. The 2016 Wallace Wade Stadium renovation
4. Whether or not the opponent is North Carolina (UNC)

The following model predicts home-game attendance percentage utilizing the 3 variables listed above.

```
pred_pct_lm_jan <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear + FPI_Diff_PrevYear + Renovated + UNC_Game,
      data = home_att_data_prevFPI)

tidy(pred_pct_lm_jan)
```

```
# A tibble: 5 x 5
  term              estimate std.error statistic  p.value
  <chr>                <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)         73.3       3.01      24.3  1.95e-33
2 OppFPI_PrevYear      0.923     0.318      2.90 5.17e- 3
3 FPI_Diff_PrevYear   -0.503     0.290     -1.73 8.83e- 2
4 RenovatedTRUE      -18.2       3.92      -4.65 1.79e- 5
5 UNC_GameTRUE        31.8       7.38       4.31 5.94e- 5
```

```
glance(pred_pct_lm_jan)$adj.r.squared
```

```
[1] 0.3996978
```

MODEL COMMENTARY TO BE ADDED.

Model Prediction

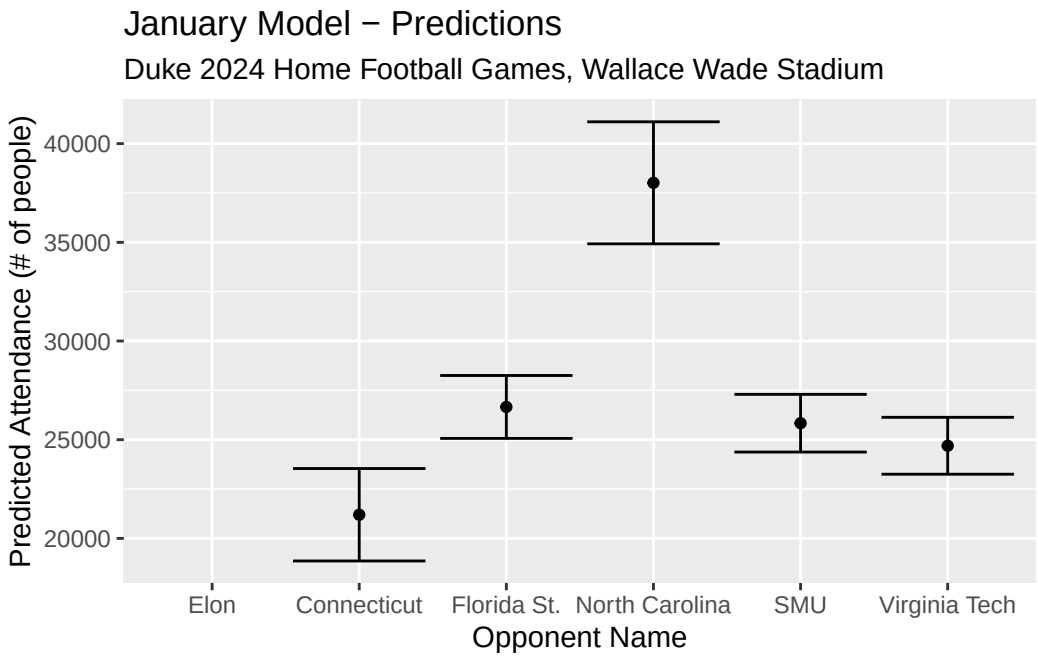
The following table and graph lists attendance predictions for all Duke 2024 home games based on the model above.

```
Elon_2024 <- att_data |>
  filter(Year == 2024, OppName == "Elon")
Elon_2024_pred <- predict(pred_pct_lm_jan$fit,
                          Elon_2024,
                          type = "response",
                          se.fit = TRUE)
Elon_2024_pred_se <- Elon_2024_pred$se.fit
#Elon_2024_pred
#Elon_2024_pred_se

UConn_2024 <- att_data |>
  filter(Year == 2024, OppName == "Connecticut")
UConn_2024_pred <- predict(pred_pct_lm_jan$fit,
                           UConn_2024,
                           type = "response",
                           se.fit = TRUE)
UConn_2024_pred_se <- UConn_2024_pred$se.fit
#UConn_2024_pred
#UConn_2024_pred_se

FSU_2024 <- att_data |>
  filter(Year == 2024, OppName == "Florida St.")
FSU_2024_pred <- predict(pred_pct_lm_jan$fit,
                         FSU_2024,
                         type = "response",
                         se.fit = TRUE)
FSU_2024_pred_se <- FSU_2024_pred$se.fit
#FSU_2024_pred

UNC_2024 <- att_data |>
  filter(Year == 2024, OppName == "North Carolina")
UNC_2024_pred <- predict(pred_pct_lm_jan$fit,
                         UNC_2024,
                         type = "response",
                         se.fit = TRUE)
UNC_2024_pred_se <- UNC_2024_pred$se.fit
#UNC_2024_pred
```

```r
SMU_2024 <- att_data |>
  filter(Year == 2024, OppName == "SMU")
SMU_2024_pred <- predict(pred_pct_lm_jan$fit,
                         SMU_2024,
                         type = "response",
                         se.fit = TRUE)
SMU_2024_pred_se <- SMU_2024_pred$se.fit
#SMU_2024_pred

VT_2024 <- att_data |>
  filter(Year == 2024, OppName == "Virginia Tech")
VT_2024_pred <- predict(pred_pct_lm_jan$fit,
                        VT_2024,
                        type = "response",
                        se.fit = TRUE)
VT_2024_pred_se <- VT_2024_pred$se.fit
#VT_2024_pred

jan_pred_model_output <- tibble(
  Name = c("Elon",
           "Connecticut",
           "Florida St.",
           "North Carolina",
           "SMU",
           "Virginia Tech"),
  "Attendance %" = c(Elon_2024_pred$fit,
                     UConn_2024_pred$fit,
                     FSU_2024_pred$fit,
                     UNC_2024_pred$fit,
                     SMU_2024_pred$fit,
                     VT_2024_pred$fit),
  "Standard Error (Att. %)" = c(Elon_2024_pred$se.fit,
                        UConn_2024_pred$se.fit,
                        FSU_2024_pred$se.fit,
                        UNC_2024_pred$se.fit,
                        SMU_2024_pred$se.fit,
                        VT_2024_pred$se.fit),
  "Estimated # of People" = c(Elon_2024_pred$fit * 40004 / 100,
                          UConn_2024_pred$fit * 40004 / 100,
                          FSU_2024_pred$fit * 40004 / 100,
                          UNC_2024_pred$fit * 40004 / 100,
                          SMU_2024_pred$fit * 40004 / 100,
```

```
                        VT_2024_pred$fit * 40004 / 100)
)

#jan_pred_model_output

jan_pred_model_data <- jan_pred_model_output |>
  mutate(AttNum = `Estimated # of People`,
         AttPct = `Attendance %`,
         SdErr = `Standard Error (Att. %)`,
         SdErrNum = SdErr * 40004 / 100) |>
  data.frame()

jan_pred_model_data |>
  ggplot(
    aes(x = fct_relevel(Name,
                        "Elon",
                        "Connecticut",
                        "Florida St.",
                        "North Carolina",
                        "SMU",
                        "Virginia Tech"),
        y = AttNum)
  ) +
  geom_point() +
  geom_errorbar(aes(ymin = AttNum - SdErrNum, ymax = AttNum + SdErrNum)) +
  labs(title = "January Model - Predictions",
       subtitle = "Duke 2024 Home Football Games, Wallace Wade Stadium",
       x = "Opponent Name",
       y = "Predicted Attendance (# of people)")
```

```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```



Elon

```
attnum_na_fpi <- home_att_data |>
  filter(is.na(OppFPI_PrevYear)) |>
  summarize(median(AttNum),
            sd(AttNum),
            min(AttNum),
            max(AttNum))
#attnum_na_fpi

na_FPI_data <- home_att_data |>
  filter(is.na(OppFPI_PrevYear))
```

Since Elon was not rated on the FPI scale last season, they are not able to produce a value through this model. However, it is possible to loosely estimate the attendance for the Duke vs. Elon game based on other factors:

- The only Duke vs. Elon football game in Wallace Wade Stadium between 2011 and 2023 occurred in 2014. This was before the 2016 stadium renovation.
- The 2014 had 31,213 attendees, which filled the stadium to about 91.963% capacity.

- Elon University is located in NC (like Duke), suggesting that attendance couple be relatively high in the 2024 matchup.
- For all home games in 2011-2023 for which the opponent's previous-season FPI was undefined, the median attendance count was 30,845 people, with a standard deviation of approximately 6,911 people.
- No significant improvement in predictive ability was found when attempting to model attendance count with variables such as First_Game (of the season), Month, Day (of the week), and others.

Thus, based on data available from 2011-2023, we currently estimate an attendance of around 31,000 people at the 2024 Duke v. Elon matchup.

Florida State

While the model predicts an attendance percentage of c(1 = 66.6415207666605), 3.98227388592832, 62, 15.7529595615982 percent, it is likely that the stadium will fill to **100%** capacity. This is because in the 2023 season, FSU achieved 100% attendance at every home game and over 90% attendance (as a percentage of total stadium capacity) in 6 out of 8 away/neutral-location games. At every game, the number of attendees exceeded the total capacity of Wallace Wade Stadium.

Final Prediction & Summary

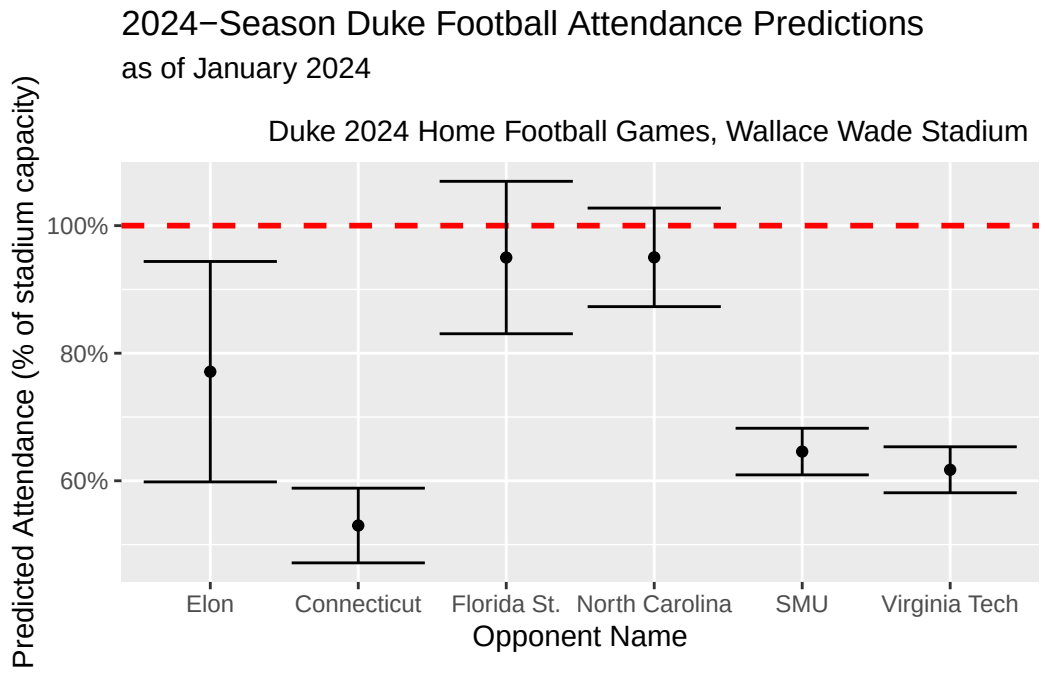
While it is difficult to predict what the home game attendance will be nearly a year in advance, below are my final *January* predictions for the number of attendees at 2024 Duke home football games:

```
janV2_pred_model_output <- tibble(
  Name = c("Elon",
           "Connecticut",
           "Florida St.",
           "North Carolina",
           "SMU",
           "Virginia Tech"),
  "Attendance %" = c(attnum_na_fpi$"median(AttNum)" / 40004 * 100,
                     UConn_2024_pred$fit,
                     95,
                     UNC_2024_pred$fit,
                     SMU_2024_pred$fit,
                     VT_2024_pred$fit),
```

```
  "Standard Error (Att. %)" = c(attnum_na_fpi$"sd(AttNum)" / 40004 * 100,
                                UConn_2024_pred$se.fit,
                                FSU_2024_pred$se.fit * 3, # x3 due to uncertainty
                                UNC_2024_pred$se.fit,
                                SMU_2024_pred$se.fit,
                                VT_2024_pred$se.fit),
  "Estimated # of People" = c(attnum_na_fpi$"median(AttNum)",
                              UConn_2024_pred$fit * 40004 / 100,
                              95 * 40004 / 100,
                              UNC_2024_pred$fit * 40004 / 100,
                              SMU_2024_pred$fit * 40004 / 100,
                              VT_2024_pred$fit * 40004 / 100)
)

janV2_pred_model_data <- janV2_pred_model_output |>
  mutate(AttNum = `Estimated # of People`,
         AttPct = `Attendance %`,
         SdErr = `Standard Error (Att. %)`,
         SdErrNum = SdErr * 40004 / 100) |>
  data.frame()

#janV2_pred_model_data

janV2_pred_model_data |>
  ggplot(
    aes(x = fct_relevel(Name,
                        "Elon",
                        "Connecticut",
                        "Florida St.",
                        "North Carolina",
                        "SMU",
                        "Virginia Tech"),
        y = AttPct)
  ) +
  geom_point() +
  geom_hline(yintercept = 100, color = "red", linetype = "dashed", size = 1) +
  geom_errorbar(aes(ymin = AttPct - SdErr, ymax = AttPct + SdErr)) +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +
  labs(title = "2024-Season Duke Football Attendance Predictions",
       subtitle = "as of January 2024\n
                   Duke 2024 Home Football Games, Wallace Wade Stadium",
       x = "Opponent Name",
```

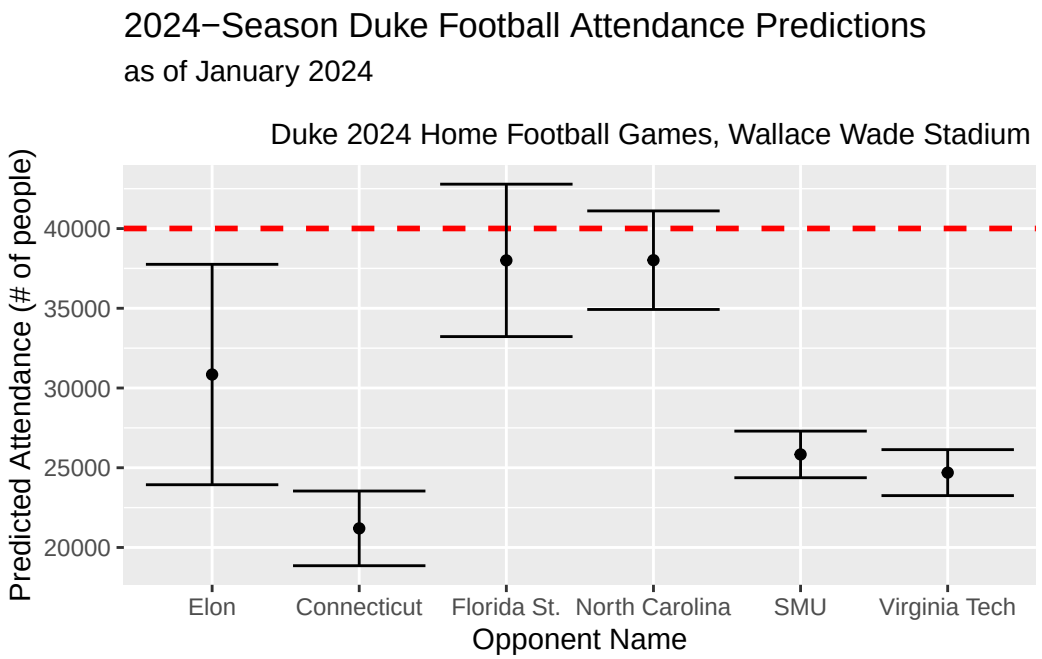
```
       y = "Predicted Attendance (% of stadium capacity)")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```




```
janV2_pred_model_data |>
  ggplot(
    aes(x = fct_relevel(Name,
                        "Elon",
                        "Connecticut",
                        "Florida St.",
                        "North Carolina",
                        "SMU",
                        "Virginia Tech"),
        y = AttNum)
  ) +
  geom_point() +
  geom_hline(yintercept = 40004, color = "red", linetype = "dashed", size = 1) +
  geom_errorbar(aes(ymin = AttNum - SdErrNum, ymax = AttNum + SdErrNum)) +
  labs(title = "2024-Season Duke Football Attendance Predictions",
       subtitle = "as of January 2024\n
```

```
                Duke 2024 Home Football Games, Wallace Wade Stadium",
      x = "Opponent Name",
      y = "Predicted Attendance (# of people)")
```



These predictions are based on:

1. The FPI (strength) of an opponent in their previous season
2. The difference between Duke's previous-season FPI and their opponent's previous-season FPI
3. The 2016 Wallace Wade Stadium renovation
4. Whether or not the opponent is North Carolina (UNC)
5. Other contextual info and/or "common sense" (for Elon and Florida State)

The red dashed line represents the maximum capacity for Wallace Wade Stadium (40,004 attendees).

A linear regression model was created which factored in variables 1-4 listed above. Each point is a prediction provided by this model, while the error bars represent the standard error of these predictions – the range each prediction could plausibly vary by. Notable exceptions to this are:

- Elon. No FPI data was available, so the prediction was the median attendance of all home games from 2011-2023 without FPI values, with the error bars representing the standard deviation of those historical games' attendance counts.

- Florida State. The model is believed to have under-predicted the true 2024 value due to Florida State's outstanding 2023 attendance record. Thus, the estimate was raised, but the error bars were tripled in size as a result of the uncertainty of this manual adjustment.

Future Directions

In the months to come, we hope to improve these predictions based on additional factors, such as:

- Win record. Based on our observations, game attendence varied significantly based on factors related to Duke's season performance, such as the number of consecutive wins achieved before game-time. This factor will change over time throughout the season and thus will dynamically impact attendance estimations.
- Temporal factors. While factors such as month appear to be overall insignificant, the time of day during which a game occurs seems to relate to attendance. However, the time and day on which each game will take place has not yet been assigned (as of January).
- Weather. We have yet to investigate whether game-time weather is related to stadium attendance. This factor is relatively irrelevant for the time being since weather predictions are not available at this time for the 2024 season. However, weather forecasts may prove to be a useful predictor for future attempts.