Predictions for 2024-Season Duke Football Attendance

Based on Previous Home-Game Attendance Records

Calvin Chen

Overview

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

In *Section 1* (early January of 2024), initial predictions for 2024 game attendance were created. These were created based on the previous 12 seasons (2011-2023).

In *Section 2* (late January of 2024), a revised dataset containing data from as early as 2001 and including some weather data was implemented. New models and variables were tested.

This document is continually a work in progress.

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.

Credits

Data was compiled from various sources, including ESPN and cfbstats.com, and analyzed by Calvin Chen. The predictive model and visualizations were developed by Calvin Chen.

Packages

```r
library(tidyverse)
library(tidymodels)
library(cowplot)
```

SECTION 1: EARLY JANUARY, 2024

Importing the Dataset

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

```r
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:*

```r
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
```

```r
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.

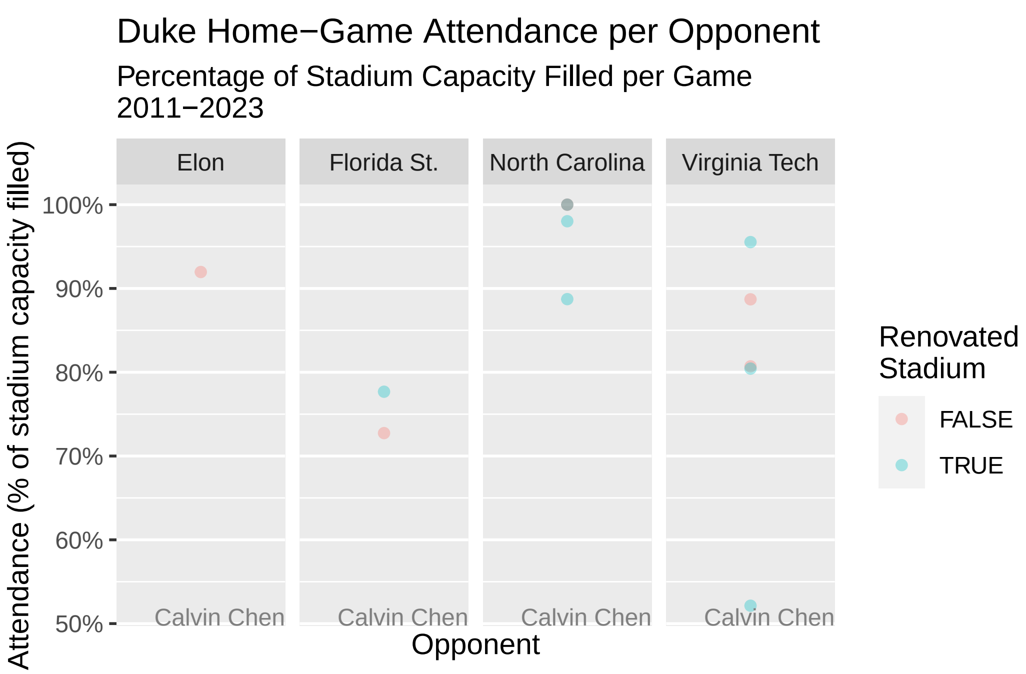
It is worth noting that Wallace Wade Stadium attendance capacity changed as a result of renovations which completed in 2016:

- *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).

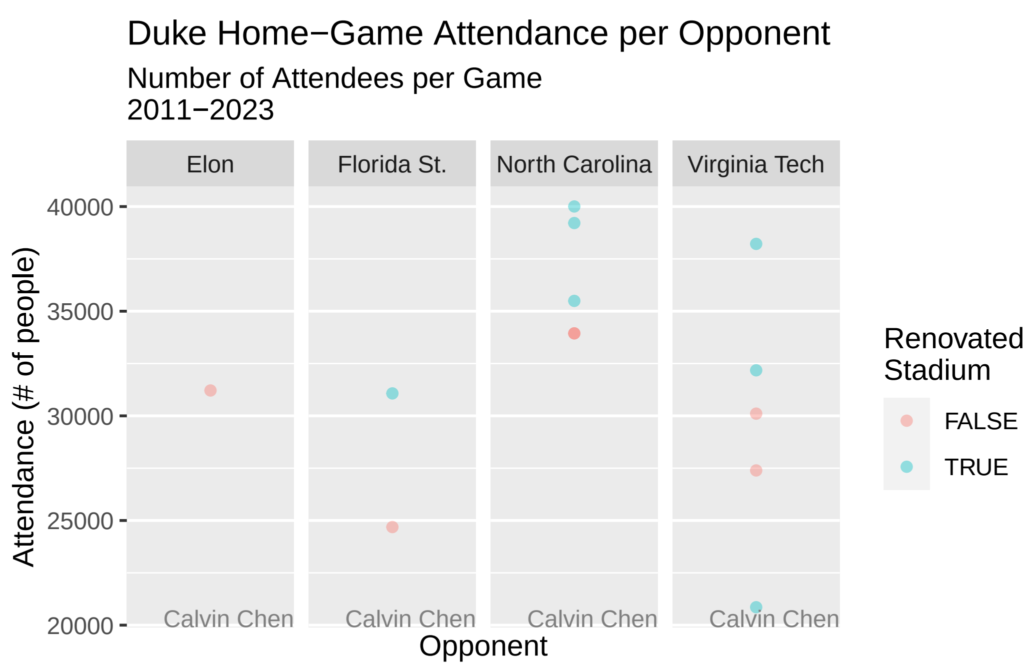
All Teams

```r
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_y_continuous(labels = function(x) paste0(x, "%")) +
  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") +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Calvin Chen",
           size = 3, color = "gray50")
```




```r
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") +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Calvin Chen",
           size = 3, color = "gray50")
```



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>
```

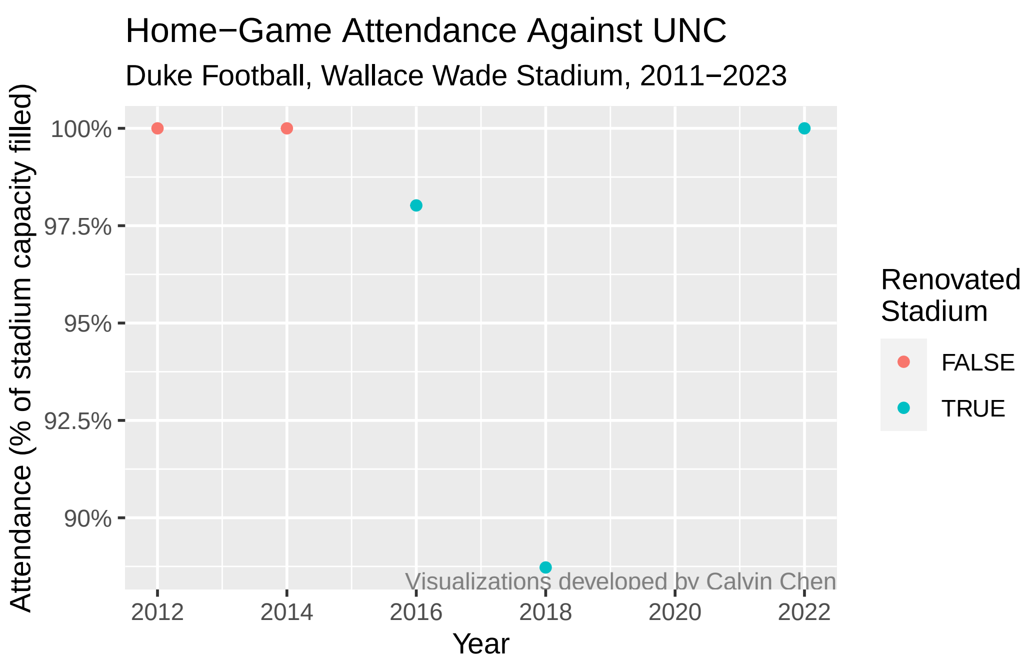
North Carolina

```
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>
```

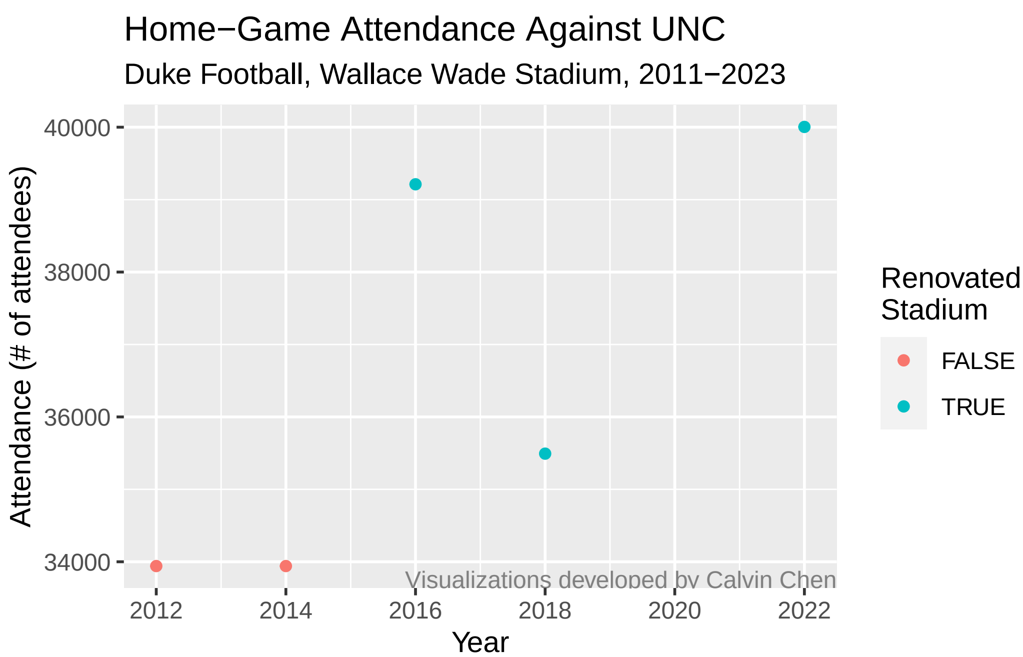
```
home_att_data |>
  filter(OppName == "North Carolina") |>
  ggplot(
    aes(x = Year, y = AttPct, color = Renovated)
  ) +
```

```
  geom_point() +
  scale_x_continuous(breaks = seq(from = 2012, to = 2023, by = 2)) +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +
  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") +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```




```
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") +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```



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.

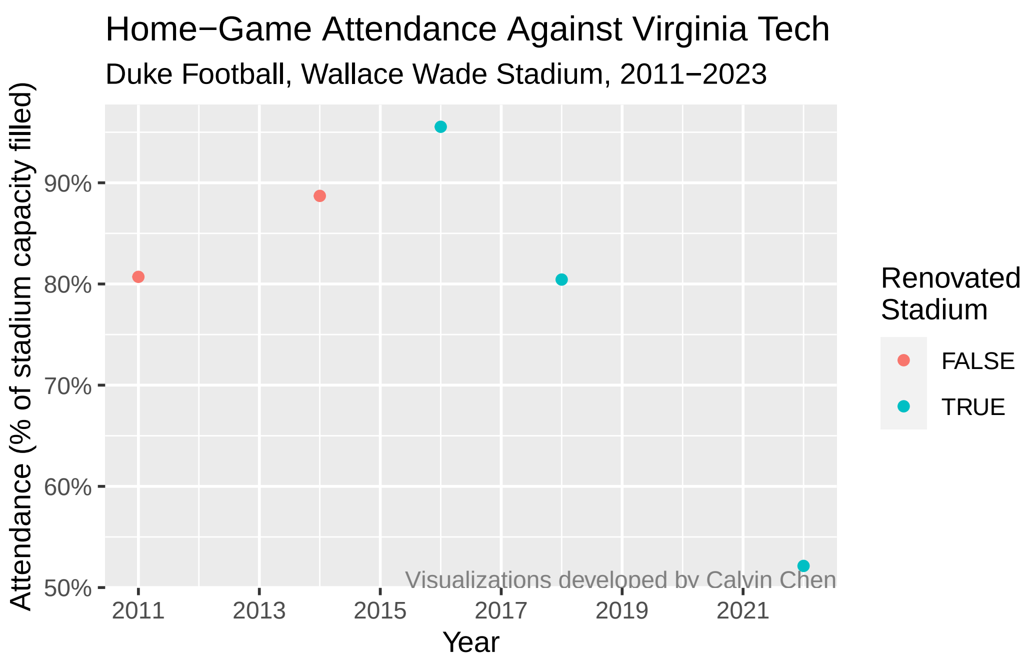
Virginia Tech

```
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)
```

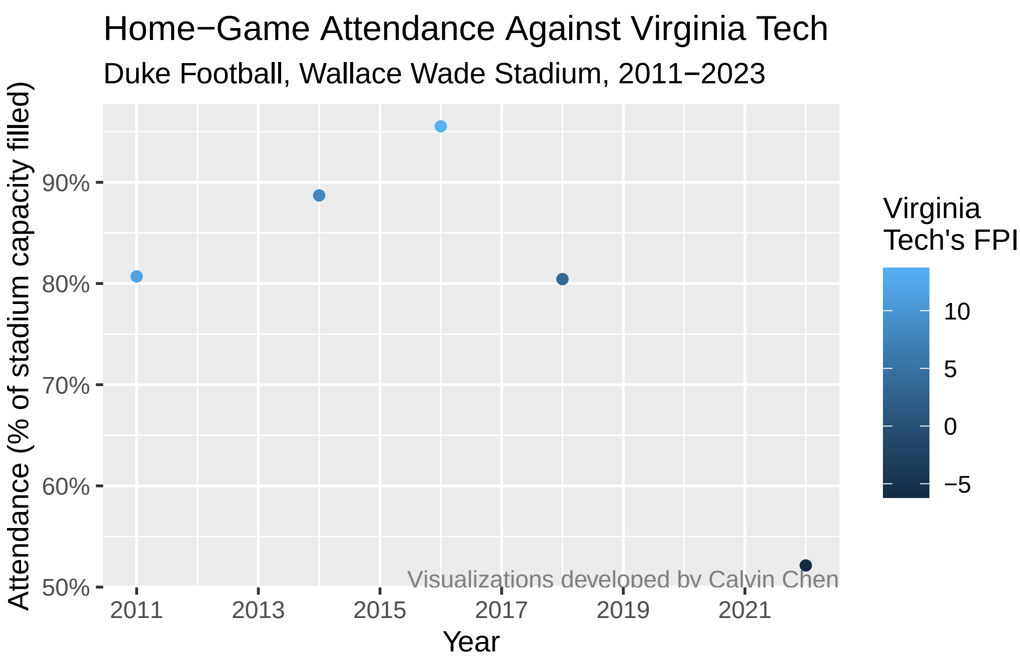
```
# A tibble: 6 x 7
  Name          `End-of-Season FPI` Month  Date  Year `# of Attendees`
  <chr>                       <dbl> <dbl> <dbl> <dbl>            <dbl>
1 Virginia Tech                11.8    10    29  2011            27392
2 Virginia Tech                 7.9    11    15  2014            30107
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)) +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +
  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") +
```

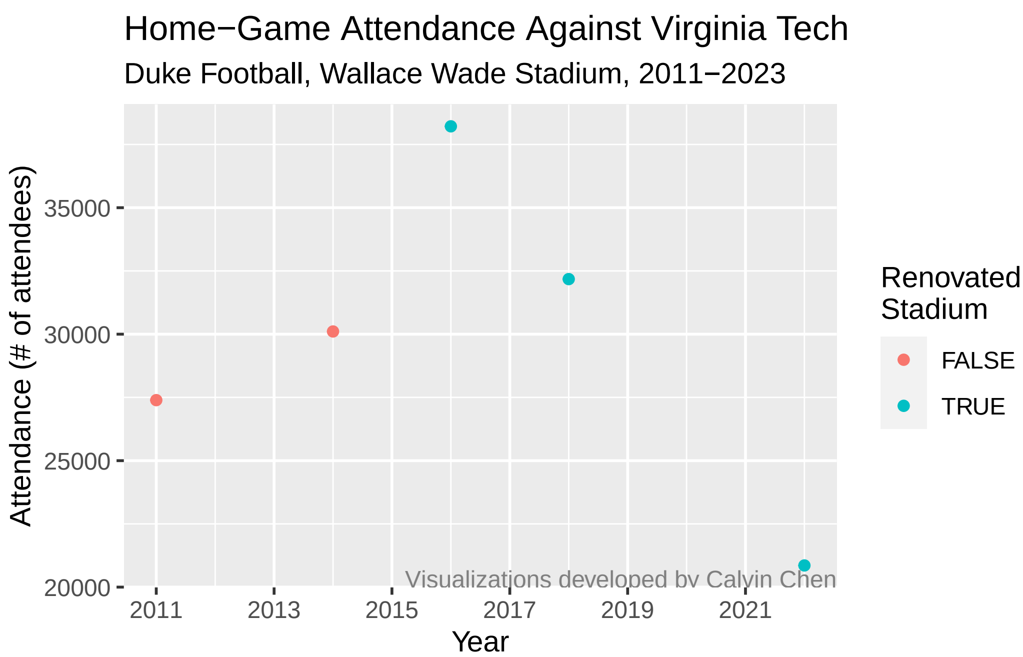
```
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```




```
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)) +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +
  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") +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```




```r
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") +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```



Team Performance vs. Attendance

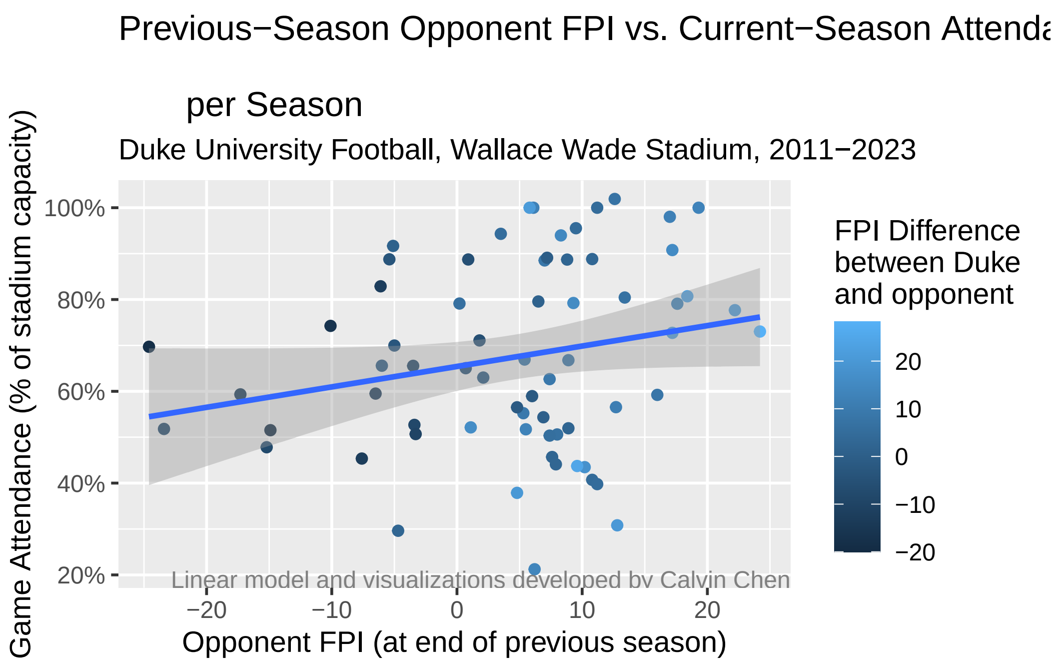
Can football team performance – both of Duke and its opponent – be used to predict the attendance turnout of future Duke home games?

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)") +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Linear model and visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```




```
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
```

```r
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

```r
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
```

```r
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
```

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

```
[1] 0.04746587
```

```r
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
```

```r
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?

```r
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
```

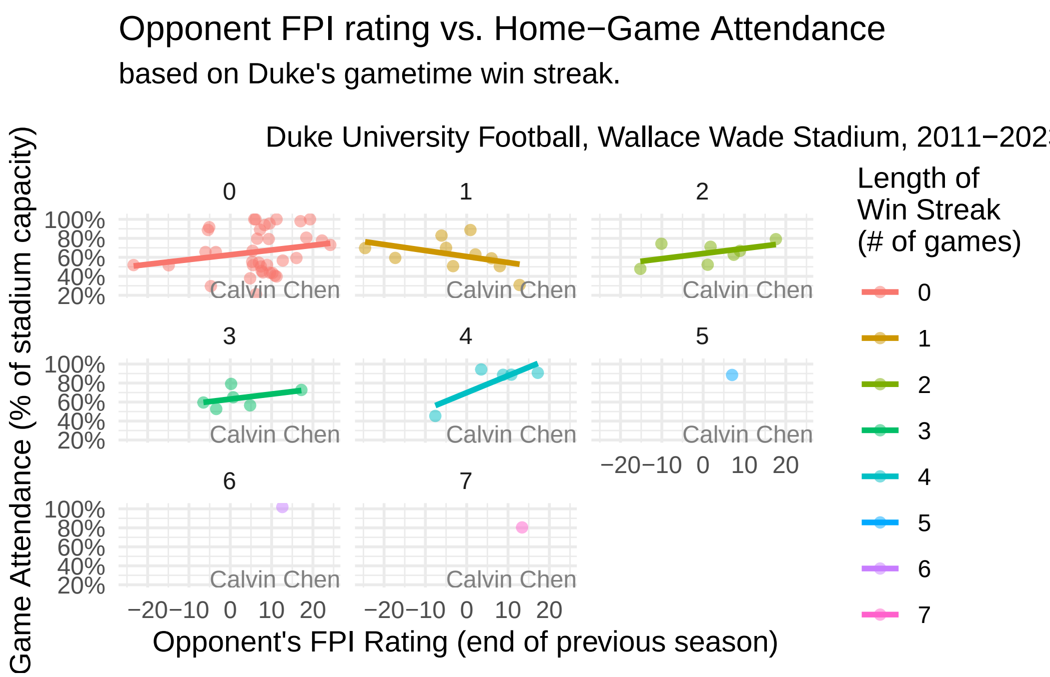
```
glance(prev_fpi_streak_lm)$adj.r.squared
```

```
[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:

```r
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)") +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Calvin Chen",
           size = 3, color = "gray50")
```

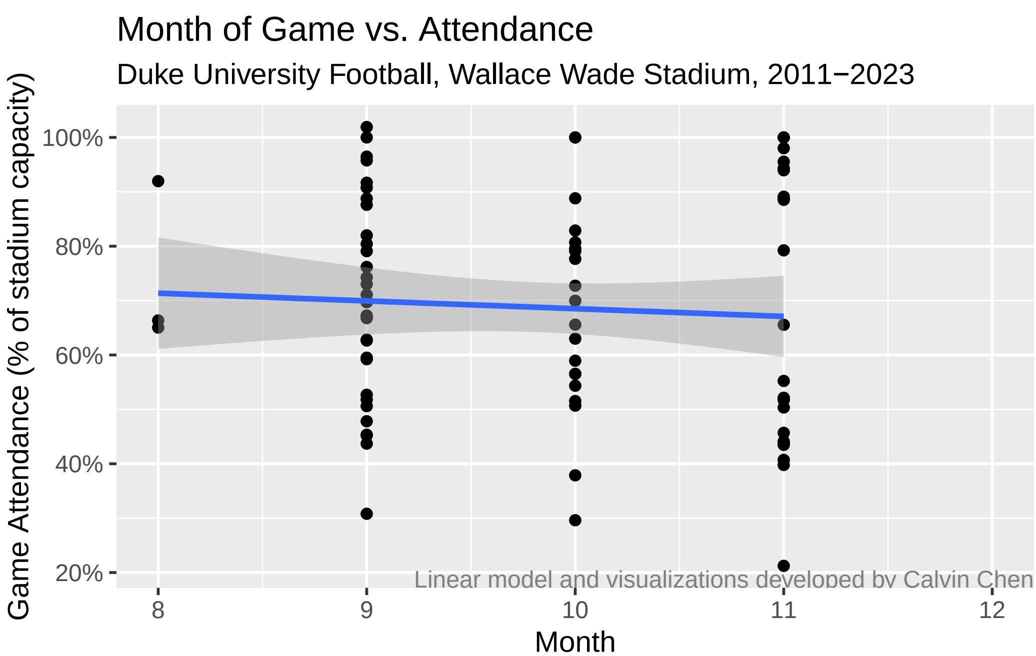


Time

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

Month

```r
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)") +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Linear model and visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```



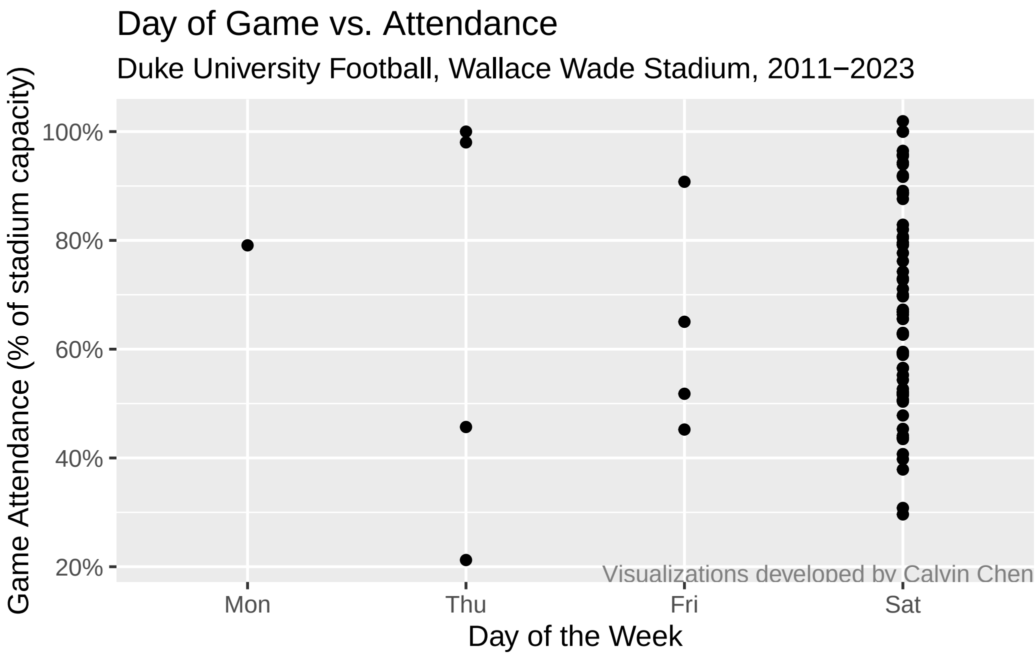
Based on the scatterplot above, no obvious correlation is present between game month and attendance for Duke home games. The spread of attendance percentages for games appears independent of the month on which a game occurs.

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
```

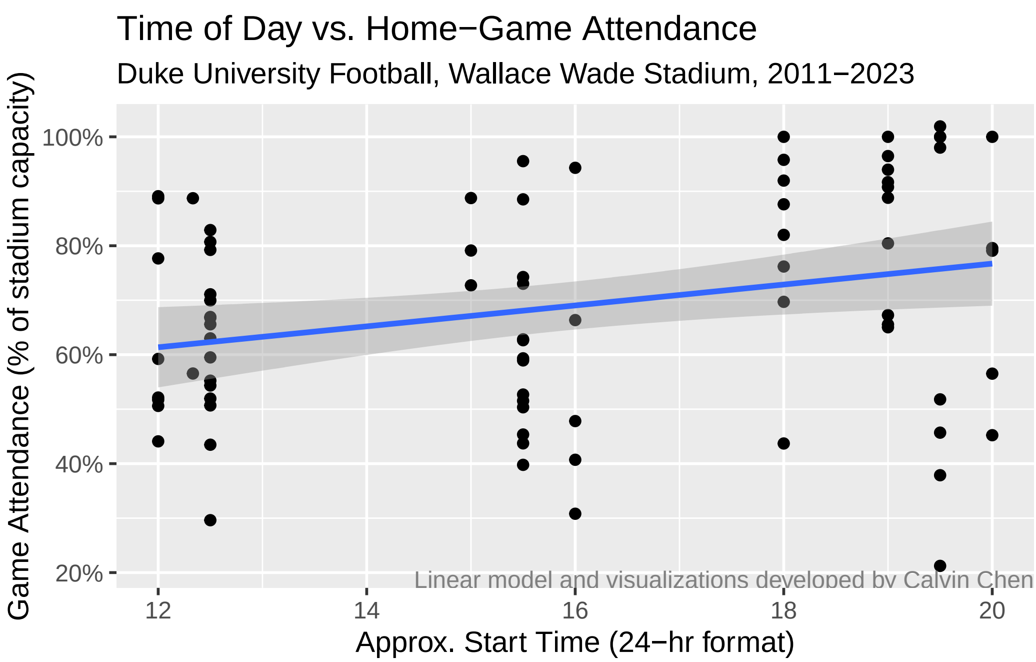
```r
home_att_data |>
  ggplot(
    aes(x = fct_relevel(Day, "Mon", "Thu", "Fri", "Sat"),
        y = AttPct)
  ) +
  geom_point() +
  labs(title = "Day of Game vs. Attendance",
       subtitle = "Duke University Football, Wallace Wade Stadium, 2011-2023",
       x = "Day of the Week",
       y = "Game Attendance (% of stadium capacity)") +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```



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

```r
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)") +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Linear model and visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```



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

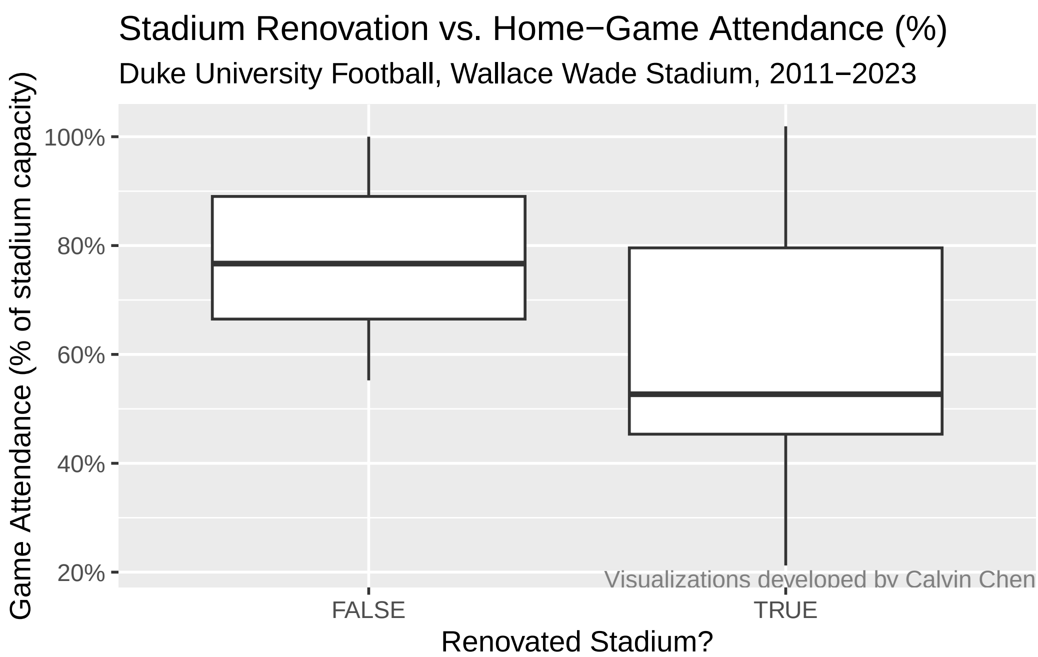
Stadium Renovation

Wallace Wade Stadium capacity:

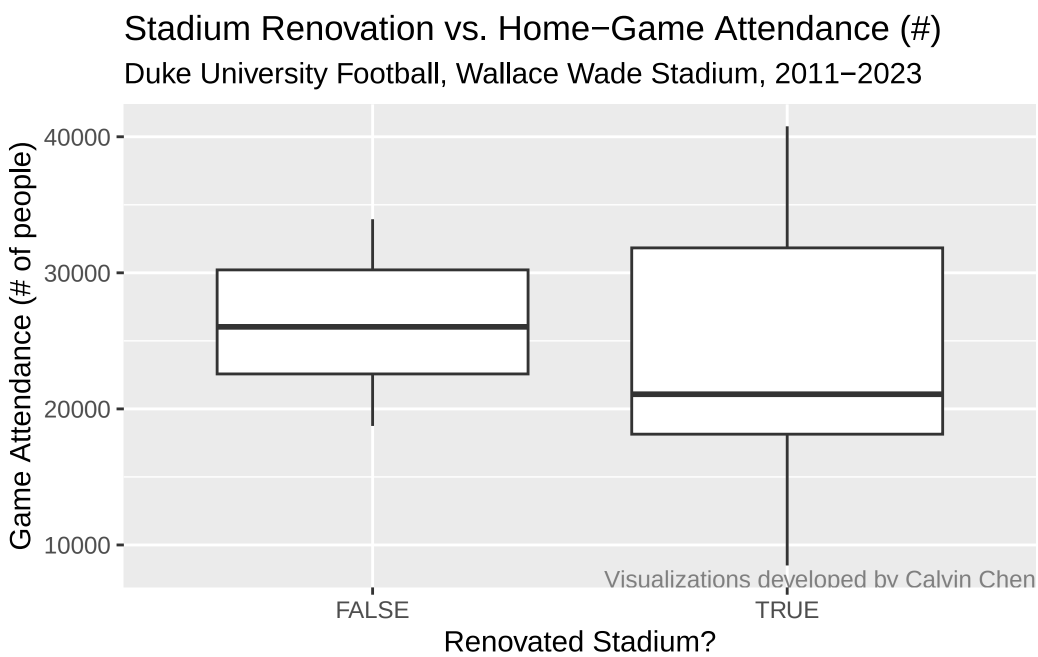
- *Pre-renovation:* 33,941 (1982-2015)
- *Post-renovation:* 40,004 (2016-present)

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)") +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```




```
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)") +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```



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?

```r
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
```

```r
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

As of early 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

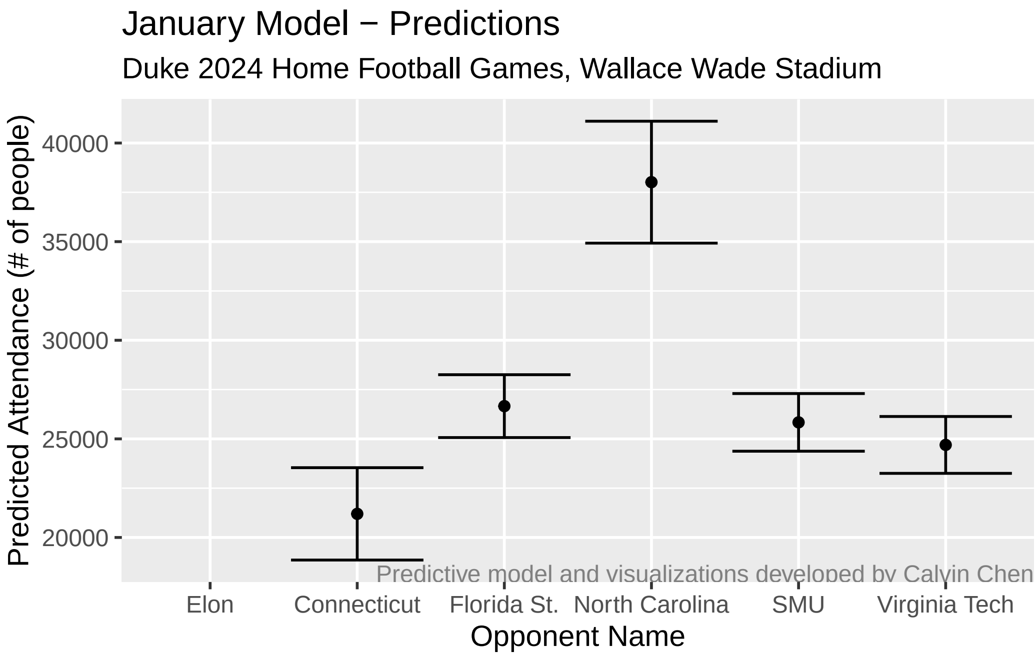
The following table and graph lists attendance predictions for all Duke 2024 home games based on the model above.

```r
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

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
```

```r
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)") +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Predictive model and visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```

```
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

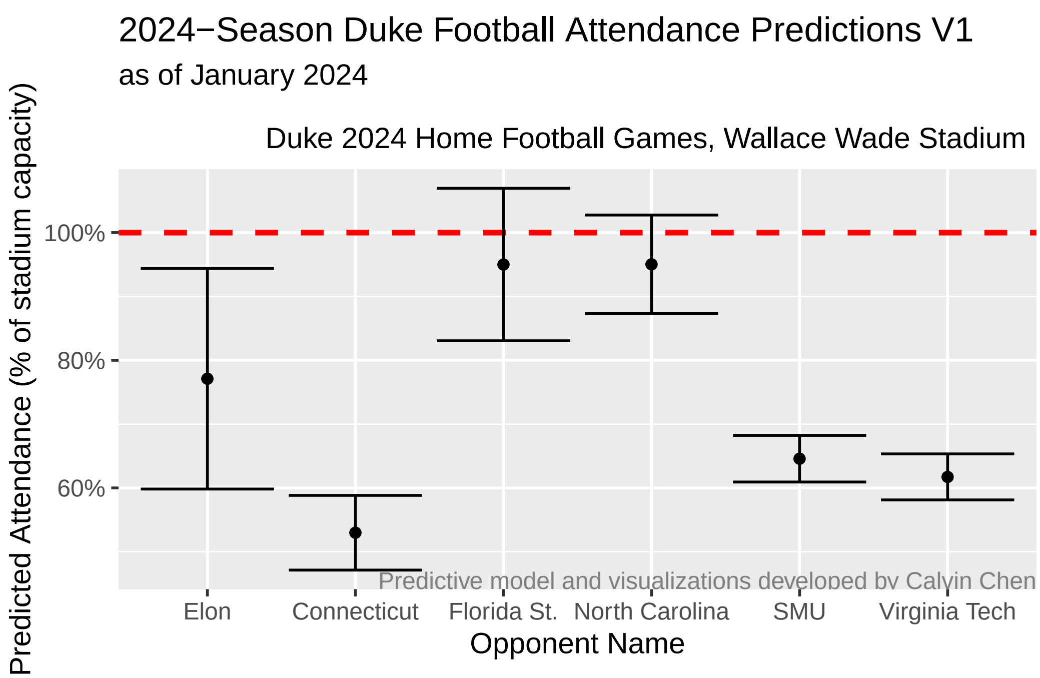
While it is difficult to predict what the home game attendance will be nearly a year in advance, below are my final *January* predictions for the number of attendees at 2024 Duke home football games:

```
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 V1",
       subtitle = "as of January 2024\n
                  Duke 2024 Home Football Games, Wallace Wade Stadium",
       x = "Opponent Name",
       y = "Predicted Attendance (% of stadium capacity)") +
```

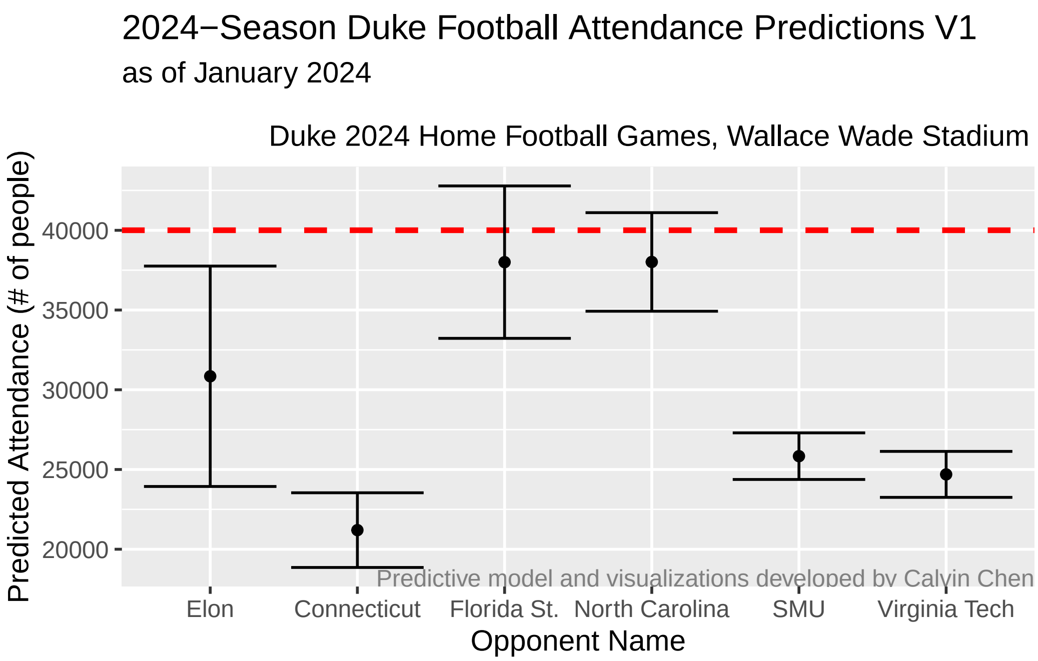
```r
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Predictive model and visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```

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




```r
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)) +
```

```r
labs(title = "2024-Season Duke Football Attendance Predictions V1",
     subtitle = "as of January 2024\n
                 Duke 2024 Home Football Games, Wallace Wade Stadium",
     x = "Opponent Name",
     y = "Predicted Attendance (# of people)") +
annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
         label = "Predictive model and visualizations developed by Calvin Chen",
         size = 3, color = "gray50")
```



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.
- The individual opponents. We have not yet investigated if including some/any of the opponents *other than UNC* as variables would improve the predictive power of the model.
- Data-set time frame. The V3 data-set only contains games as early as 2011, but further investigation could be conducted using data that extends farther back in time.
- New variables. More variables beyond those included in this V3 data-set have the potential to be related to game-time attendance. It could benefit our predictions to investigate the predictive power of additional variables before/during the 2024 season.

SECTION 2: LATE JANUARY, 2024

Duke's home schedule was recently announced, so this section includes updates to the prediction models based on the new available information. Additionally, the dataset used in this section is improved from section 1, since it includes a more extensive range of years (2001-2023 instead of 2011-2023), new data for weather, and other relatively minor updates and additions.

Importing the Dataset

```
att_data_2 <- read_csv("data/Duke Stats - DukeAttendanceV6.csv")

att_data_2 <- att_data_2 |>
  mutate(NumericalDate = (Month-8)*100 + Date/31*100) |>
  mutate(Day = as.factor(Day)) |>
  mutate(Rain = if_else(is.na(Weather), NA,
                        if_else(Weather %in% c("Rain", "Light Rain", "Heavy Rain"),
                                TRUE,
                                FALSE))
  ) |>
  mutate(onSaturday = if_else(Day == "Sat", TRUE, FALSE)) |>
  mutate(FSU_Game = if_else(OppName == "Florida St.", TRUE, FALSE))

home_att_data_2 <- att_data_2 |>
  filter(Site == "Home", Year < 2024)

home_att_data_prevFPI_2 <- home_att_data_2 |>
  filter(!is.na(OppFPI_PrevYear)) |>
  mutate(OppFPI_PrevYear = OppFPI_PrevYear,
         FPI_Diff_PrevYear = FPI_Diff_PrevYear)
```

Testing Prior Model with New Dataset

Since this section's dataset contains more rows, we will examine if the section-one prediction model is any stronger when given the new data.

```
pred_pct_lm_jan2 <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear + FPI_Diff_PrevYear + Renovated + UNC_Game,
```

```r
      data = home_att_data_prevFPI_2)

tidy(pred_pct_lm_jan2)
```

```
# A tibble: 5 x 5
  term               estimate std.error statistic  p.value
  <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)          72.9       2.39      30.5  6.40e-49
2 OppFPI_PrevYear       1.15      0.267      4.30 4.39e- 5
3 FPI_Diff_PrevYear    -0.793     0.222     -3.57 5.80e- 4
4 RenovatedTRUE       -17.1       3.38      -5.06 2.21e- 6
5 UNC_GameTRUE         26.4       5.73       4.61 1.35e- 5
```

```r
glance(pred_pct_lm_jan2)$adj.r.squared
```

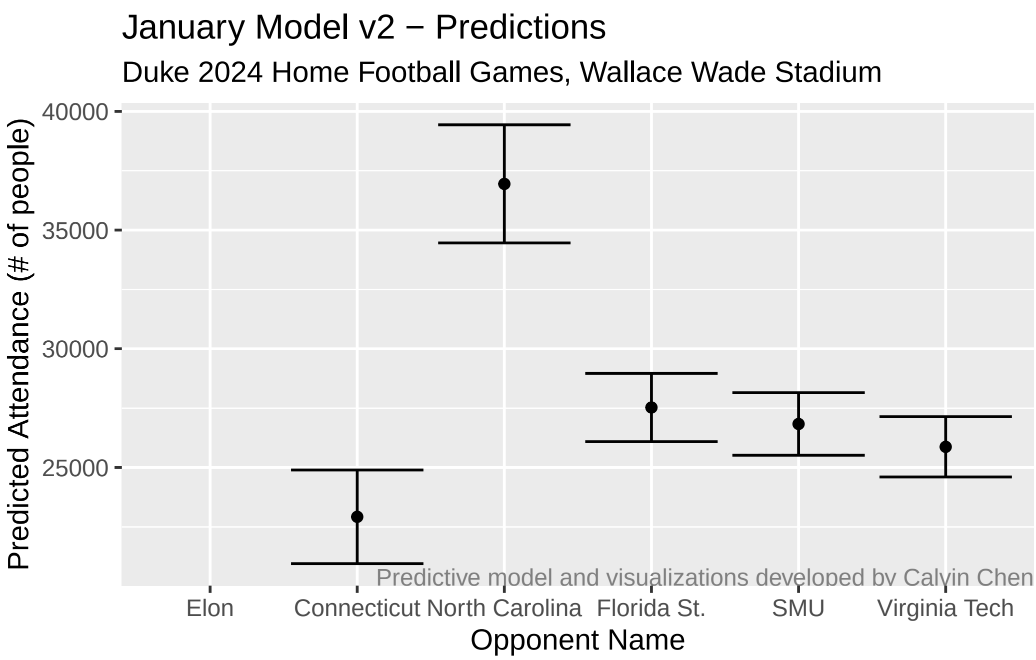
```
[1] 0.3612717
```

```r
Elon_2024 <- att_data_2 |>
  filter(Year == 2024, OppName == "Elon")
Elon_2024_pred <- predict(pred_pct_lm_jan2$fit,
                          Elon_2024,
                          type = "response",
                          se.fit = TRUE)
Elon_2024_pred_se <- Elon_2024_pred$se.fit
#Elon_2024_pred
#Elon_2024_pred_se

UConn_2024 <- att_data_2 |>
  filter(Year == 2024, OppName == "Connecticut")
UConn_2024_pred <- predict(pred_pct_lm_jan2$fit,
                           UConn_2024,
                           type = "response",
                           se.fit = TRUE)
UConn_2024_pred_se <- UConn_2024_pred$se.fit
#UConn_2024_pred
#UConn_2024_pred_se

FSU_2024 <- att_data_2 |>
  filter(Year == 2024, OppName == "Florida St.")
FSU_2024_pred <- predict(pred_pct_lm_jan2$fit,
```

```
                          FSU_2024,
                          type = "response",
                          se.fit = TRUE)
FSU_2024_pred_se <- FSU_2024_pred$se.fit
#FSU_2024_pred

UNC_2024 <- att_data_2 |>
  filter(Year == 2024, OppName == "North Carolina")
UNC_2024_pred <- predict(pred_pct_lm_jan2$fit,
                          UNC_2024,
                          type = "response",
                          se.fit = TRUE)
UNC_2024_pred_se <- UNC_2024_pred$se.fit
#UNC_2024_pred

SMU_2024 <- att_data_2 |>
  filter(Year == 2024, OppName == "SMU")
SMU_2024_pred <- predict(pred_pct_lm_jan2$fit,
                          SMU_2024,
                          type = "response",
                          se.fit = TRUE)
SMU_2024_pred_se <- SMU_2024_pred$se.fit
#SMU_2024_pred

VT_2024 <- att_data_2 |>
  filter(Year == 2024, OppName == "Virginia Tech")
VT_2024_pred <- predict(pred_pct_lm_jan2$fit,
                         VT_2024,
                         type = "response",
                         se.fit = TRUE)
VT_2024_pred_se <- VT_2024_pred$se.fit
#VT_2024_pred

jan_pred_model_output2 <- tibble(
  Name = c("Elon",
           "Connecticut",
           "North Carolina",
           "Florida St.",
           "SMU",
           "Virginia Tech"),
  "Attendance %" = c(Elon_2024_pred$fit,
                     UConn_2024_pred$fit,
```

```r
                     UNC_2024_pred$fit,
                     FSU_2024_pred$fit,
                     SMU_2024_pred$fit,
                     VT_2024_pred$fit),
  "Standard Error (Att. %)" = c(Elon_2024_pred$se.fit,
                       UConn_2024_pred$se.fit,
                       UNC_2024_pred$se.fit,
                       FSU_2024_pred$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,
                         UNC_2024_pred$fit * 40004 / 100,
                         FSU_2024_pred$fit * 40004 / 100,
                         SMU_2024_pred$fit * 40004 / 100,
                         VT_2024_pred$fit * 40004 / 100)
)

#jan_pred_model_output

jan_pred_model_data2 <- jan_pred_model_output2 |>
  mutate(AttNum = `Estimated # of People`,
         AttPct = `Attendance %`,
         SdErr = `Standard Error (Att. %)`,
         SdErrNum = SdErr * 40004 / 100) |>
  data.frame()

jan_pred_model_data2 |>
  ggplot(
    aes(x = fct_relevel(Name,
                        "Elon",
                        "Connecticut",
                        "North Carolina",
                        "Florida St.",
                        "SMU",
                        "Virginia Tech"),
        y = AttNum)
  ) +
  geom_point() +
  geom_errorbar(aes(ymin = AttNum - SdErrNum, ymax = AttNum + SdErrNum)) +
  labs(title = "January Model v2 - Predictions",
       subtitle = "Duke 2024 Home Football Games, Wallace Wade Stadium",
```

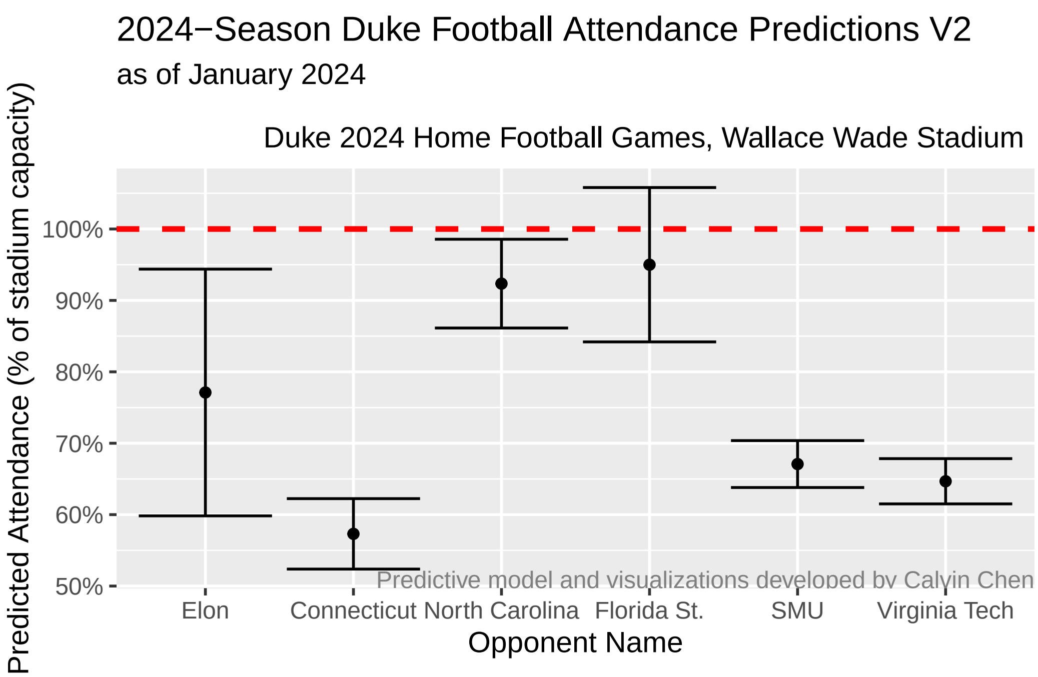
```
      x = "Opponent Name",
      y = "Predicted Attendance (# of people)") +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Predictive model and visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```




```
# below is a new version of the final prediction model:
janV3_pred_model_output <- tibble(
  Name = c("Elon",
           "Connecticut",
           "North Carolina",
           "Florida St.",
           "SMU",
           "Virginia Tech"),
  "Attendance %" = c(attnum_na_fpi$"median(AttNum)" / 40004 * 100,
                     UConn_2024_pred$fit,
                     UNC_2024_pred$fit,
                     95,
                     SMU_2024_pred$fit,
                     VT_2024_pred$fit),
  "Standard Error (Att. %)" = c(attnum_na_fpi$"sd(AttNum)" / 40004 * 100,
                                UConn_2024_pred$se.fit,
```

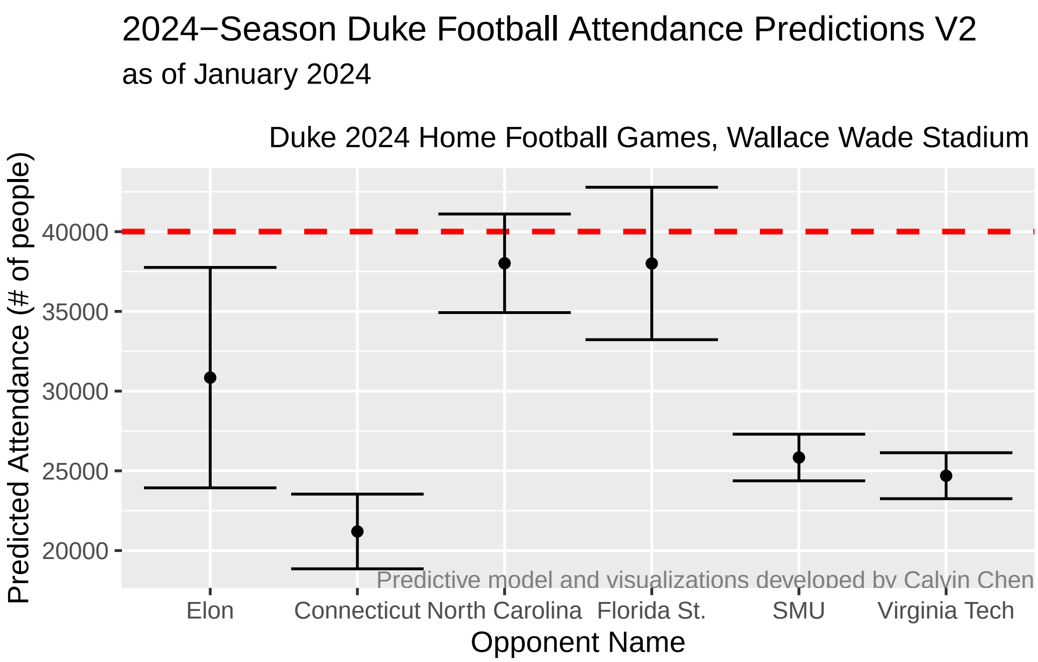
```
                                UNC_2024_pred$se.fit,
                                FSU_2024_pred$se.fit * 3, # x3 due to uncertainty
                                SMU_2024_pred$se.fit,
                                VT_2024_pred$se.fit),
  "Estimated # of People" = c(attnum_na_fpi$"median(AttNum)",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

janV3_pred_model_data <- janV3_pred_model_output |>
  mutate(AttNum = `Estimated # of People`,
         AttPct = `Attendance %`,
         SdErr = `Standard Error (Att. %)`,
         SdErrNum = SdErr * 40004 / 100) |>
  data.frame()

#janV3_pred_model_data

janV3_pred_model_data |>
  ggplot(
    aes(x = fct_relevel(Name,
                        "Elon",
                        "Connecticut",
                        "North Carolina",
                        "Florida St.",
                        "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 V2",
       subtitle = "as of January 2024\n
                  Duke 2024 Home Football Games, Wallace Wade Stadium",
       x = "Opponent Name",
       y = "Predicted Attendance (% of stadium capacity)") +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
```

```
           label = "Predictive model and visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```

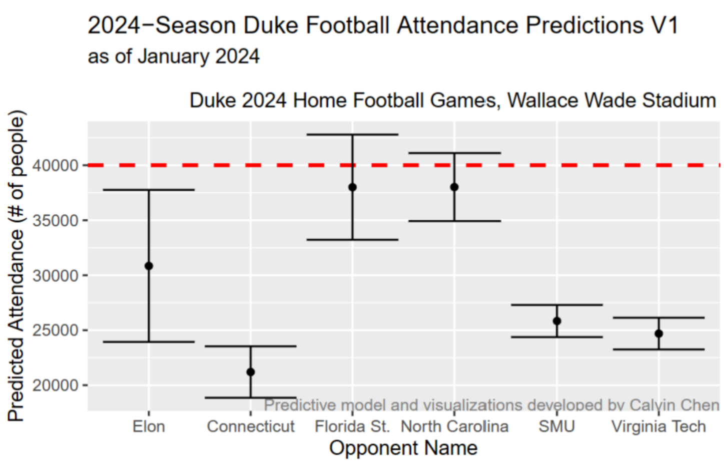



```
janV2_pred_model_data |>
  ggplot(
    aes(x = fct_relevel(Name,
                        "Elon",
                        "Connecticut",
                        "North Carolina",
                        "Florida St.",
                        "SMU",
                        "Virginia Tech"),
        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 V2",
       subtitle = "as of January 2024\n
                  Duke 2024 Home Football Games, Wallace Wade Stadium",
       x = "Opponent Name",
       y = "Predicted Attendance (# of people)") +
```

```r
annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
         label = "Predictive model and visualizations developed by Calvin Chen",
         size = 3, color = "gray50")
```



Comparison



Stadium Renovation

```
renovated_pct_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ Renovated, data = home_att_data_2)

tidy(renovated_pct_lm)
```

```
# A tibble: 2 x 5
  term          estimate std.error statistic  p.value
  <chr>            <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)      70.8       1.97      36.0  2.08e-72
2 RenovatedTRUE    -9.04      3.48      -2.60 1.04e- 2
```

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

```
[1] 0.03946028
```

```
renovated_count_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ Renovated, data = home_att_data_2)

tidy(renovated_count_lm)
```

```
# A tibble: 2 x 5
  term          estimate std.error statistic  p.value
  <chr>            <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     24042.      717.     33.5   1.83e-68
2 RenovatedTRUE     679.     1270.      0.535 5.94e- 1
```

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

```
[1] -0.005125414
```

Based on these models containing data from 2001-2023, evidence can be seen that home-game **attendance head-count remained largely unchanged after the 2016 Wallace Wade Stadium rennovation.** The percentage of the full stadium which was filled per game tended to be *less* after the renovation (by 9% on average), as observable in the first model with a p-value of approximately 0.001. The second model shows that no statistically significant relationship is present between stadium renovation and attendance head-count.

The following models will use both attendance percentage and attendance head-count depending on which yields better predictions (as suggested by the adjusted r-squared value of each model).

Head Coach

Was stadium attendance significantly higher during (either) Elko and Cutcliffe's times as head coach?

```
elko_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ Elko, data = home_att_data_2)

tidy(elko_lm)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    68.5       1.73     39.5  1.82e-77
2 ElkoTRUE       -6.48      5.71     -1.13 2.59e- 1
```

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

```
[1] 0.002033403
```

```
cutcliffe_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ Cutcliffe, data = home_att_data_2)

tidy(cutcliffe_lm)
```

```
# A tibble: 2 x 5
  term            estimate std.error statistic  p.value
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)       21285.      892.      23.9  5.26e-51
2 CutcliffeTRUE      4876.     1142.       4.27 3.61e- 5
```

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

```
[1] 0.1095574
```

While no statistically significant increase in Duke home-game attendance was observed during Elko's time as head coach, **a *statistically significant* increase in attendance was observed during Cutcliffe's seasons as head coach.** On average, an additional 4,876 people attended each home game when Cutcliffe was head coach of Duke football. The p-value of this finding is <0.0001.

Previous-Season FPI

```r
fpi2_pct <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ OppFPI_PrevYear, data = home_att_data_prevFPI_2)

fpi2_num <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ OppFPI_PrevYear, data = home_att_data_prevFPI_2)

tidy(fpi2_pct)
```

```
# A tibble: 2 x 5
  term            estimate std.error statistic  p.value
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)       66.4       2.20      30.2  1.39e-49
2 OppFPI_PrevYear    0.386     0.208      1.86 6.65e- 2
```

```r
glance(fpi2_pct)$adj.r.squared
```

```
[1] 0.02566184
```

```r
tidy(fpi2_num)
```

```
# A tibble: 2 x 5
  term            estimate std.error statistic  p.value
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)       23841.     768.       31.0  1.57e-50
2 OppFPI_PrevYear     169.      72.6       2.33 2.21e- 2
```

```r
glance(fpi2_num)$adj.r.squared
```

```
[1] 0.04537442
```

```r
pred2_fpiDiff <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ FPI_Diff_PrevYear, data = home_att_data_2)

tidy(pred2_fpiDiff)
```

```
# A tibble: 2 x 5
  term              estimate std.error statistic  p.value
  <chr>                <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)        67.8        2.33     29.1   3.75e-48
2 FPI_Diff_PrevYear   0.0573     0.176     0.326 7.45e- 1
```

```r
glance(pred2_fpiDiff)$adj.r.squared
```

```
[1] -0.009700814
```

Even with the new data, the *previous-season* FPI of the opponent continues to be a significant predictor of stadium attendance. For every 1-point increase in FPI of Duke's opponent at the end of the previous season, on average the stadium was filled by an additional 0.37% of its total capacity.

When the model was applied to headcount instead of capacity percentage, it was found that attendance increased by 169 people on average per 1-point increase in the opponent's previous-season FPI.

This method has slightly stronger predictive power when applied to attendance *head-count* rather than attendance percentage, as indicated by a slightly higher adjusted r-squared value and a slightly lower p-value (0.066 vs. 0.022).

Ultimately, this indicates that **when an opponent tended to be more powerful in its previous season, home-game attendance in the current season tends to be increased against that opponent.**

Predictions Based on Weather

Precipitation

```r
precip_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ Rain, data = home_att_data_2)

tidy(precip_lm)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   24677.     1706.     14.5  2.51e-15
2 RainTRUE      -6024.     3465.     -1.74 9.21e- 2
```

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

```
[1] 0.0594363
```

Based on the weather data collected so far, it appears that **rain is potentially related to a decrease in stadium attendance.** When there was rainfall of any strength, home-game attendance *decreased* on average by 6,024 people (which constitutes about *15 percent* of the full stadium capacity). The p-value of this *beta* is about 0.092, which is less than 0.1.

Temperature

```r
temp_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ Temp, data = home_att_data_2)

tidy(temp_lm)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  65.6        8.61      7.62  3.54e-12
2 Temp          0.0337     0.124     0.273 7.86e- 1
```

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

```
[1] -0.006656262
```

Based on the weather data collected so far, it appears that weather temperature is **not** a reliable predictor of stadium attendance for Duke home games.

Humidity

```r
humidity_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ Humidity, data = home_att_data_2)

tidy(humidity_lm)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   76.6      5.15       14.9  1.70e-30
2 Humidity      -0.147    0.0837     -1.76 8.01e- 2
```

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

```
[1] 0.01483338
```

Surprisingly, there is some evidence that **games with greater humidity have slightly lower attendance.** For every 1-percentage increase in humidity, on average the stadium attendance decreased by about -0.15% of the total stadium capacity. In other words, if humidity was decreased by 25, stadium attendance decreased on average by about 3.7% of the total stadium capacity. The p-value for this *beta* is approximately 0.08.

Time

Date

A new variable, *numericDate*, is introduced. It is calculated by the following:

$$(Month - 8) * 100 + (Date/31) * 100$$

This formula means that a football match taking place on August 1 will have a value of approximately 100/31 *(minimum)*, and a match taking place on December 31 will have a value of 500 *(maximum)*.

```
datenum_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ NumericalDate, data = home_att_data_2)

tidy(datenum_lm)
```

```
# A tibble: 2 x 5
  term          estimate std.error statistic  p.value
  <chr>            <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   25208.     1651.       15.3   1.59e-31
2 NumericalDate    -4.04      6.57     -0.616 5.39e- 1
```

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

```
[1] -0.004455857
```

Similar to section 1, we see **no** significant relationship between the date on which a Duke home game takes place and the stadium attendance.

Day of Week

```
day_week_v2_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ Day, data = home_att_data_2)

tidy(day_week_v2_lm)
```

```
# A tibble: 4 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   25286      3527.    7.17   4.27e-11
2 DayMon         6352      7887.    0.805  4.22e- 1
3 DaySat        -1136.     3580.   -0.317  7.52e- 1
4 DayThu         -305.     4988.   -0.0612 9.51e- 1
```

```
glance(day_week_v2_lm)$adj.r.squared
```

```
[1] -0.01264266
```

```
saturday_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ onSaturday, data = home_att_data_2)

tidy(saturday_lm)
```

```
# A tibble: 2 x 5
  term           estimate std.error statistic  p.value
  <chr>             <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)      25856.     2341.    11.0   9.45e-21
2 onSaturdayTRUE   -1706.     2419.    -0.705 4.82e- 1
```

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

```
[1] -0.003605048
```

Similar to section 1, we see **no** significant relationship between the day of the week on which a Duke home game takes place and the stadium attendance. Whether or not a game took place on a Saturday also had no significant relationship with stadium attendance.

Duke Improvement from Previous Year

Does an increase in Duke's FPI from the previous year correlate to increased stadium attendance?

```r
fpi_change_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttPct ~ DukeFPI_NetChange, data = home_att_data_2)

tidy(fpi_change_lm)
```

```
# A tibble: 2 x 5
  term              estimate std.error statistic  p.value
  <chr>                <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)         69.5       1.92      36.1  1.67e-62
2 DukeFPI_NetChange    0.344     0.289      1.19 2.37e- 1
```

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

```
[1] 0.003746385
```

There is very slight evidence that Duke's home-game attendance could increase when they are more improved from the previous season (based on season-end FPI rating). However, this evidence is ***not*** statistically significant.

First Game of Season

```r
first_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ First_Game, data = home_att_data_2)

tidy(first_lm)
```

```
# A tibble: 2 x 5
  term           estimate std.error statistic  p.value
  <chr>             <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)      23819.      616.     38.7  3.05e-76
2 First_GameTRUE    4141.     1889.      2.19 3.01e- 2
```

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

```
[1] 0.02645144
```

There is significant evidence to suggest that **home-game attendance *increases* when a game is the first one of the season.** Attendance increased on average by about 4,141 people when a game was the first game of Duke's football season. The p-value of this *beta* is about 0.030, which is <0.1 and is *statistically significant.*

UNC Game

```
unc_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ UNC_Game, data = home_att_data_2)

tidy(unc_lm)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    23601.      578.     40.8  2.65e-79
2 UNC_GameTRUE    9283.     2170.      4.28 3.48e- 5
```

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

```
[1] 0.1100087
```

As seen in Section 1, **when UNC is the opponent, stadium attendance tends to increase by a greatly significant amount**. On average, home-game attendance increases by about 9,283 people when Duke plays UNC. The p-value of this finding is <0.0001.

Florida State Game

Does stadium attendance differ significantly when the opponent is Florida State?

```
fsu_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ FSU_Game, data = home_att_data_2)

tidy(fsu_lm)
```

```
# A tibble: 2 x 5
  term          estimate std.error statistic  p.value
  <chr>            <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     24280.      605.    40.1   2.78e-78
2 FSU_GameTRUE     -497.     2935.    -0.169 8.66e- 1
```

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

```
[1] -0.006986287
```

It appears that stadium attendance is not significantly related to whether or not Florida State is the opponent. The p-value of the beta in this case is 0.866, which is well above 0.1.

Ranked Opponent

When an opponent is ranked in the top 25 at game-time, does home-game attendance increase? *(This data was only calculated back to 2011, not 2001.)*

```r
ranked_opp_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ OppRankedGametime, data = home_att_data_2)

tidy(ranked_opp_lm)
```

```
# A tibble: 2 x 5
  term                  estimate std.error statistic  p.value
  <chr>                    <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)             24352.      843.     28.9  4.57e-43
2 OppRankedGametimeTRUE    6816.     2078.      3.28 1.56e- 3
```

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

```
[1] 0.1111627
```

There is statistically significant evidence that **home-game attendance increases when Duke's opponent is ranked.** Attendance increased on average by 6,816 people when facing a ranked opponent. The p-value of this *beta* is <0.001.

Summary of Significant Variables

Stadium attendance tended to *increase* when:

- David Cutcliffe was head coach during the season.
- the opponent's previous-season FPI was high.
- the event was the first game of the season.
- UNC was the opponent.
- Duke's opponent is ranked in the top 25.

Stadium attendance tended to *decrease* when:

- precipitation occurred during the game.
- humidity was higher during the game. (likely related to rainfall)

The following variables historically had no statistically significant impact on home-game attendance:

- Mike Elko as head coach during the season.
- difference between Duke and its opponent in terms of previous-season FPI score.
- temperature during gametime.
- month, date, and day of the week.
- whether Duke's FPI rating was improved or decreased from the previous season.
- Florida State was the opponent.

Some factors, such as time of day and win streak, were not reviewed in this section because they are not currently useful to our 2024 home-game predictions.

Updated Model

The model below utilizes variables in this section that were deemed statistically significant: Cutcliffe, previous-season FPI, first-game status, UNC, and ranked-opponent status:

```r
late_jan_pred_lm <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ Cutcliffe + OppFPI_PrevYear +
      First_Game + UNC_Game + OppRankedGametime,
      data = home_att_data_2)

tidy(late_jan_pred_lm)
```

```
# A tibble: 6 x 5
  term                  estimate std.error statistic  p.value
  <chr>                    <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)             21413.     1904.    11.2   1.66e-16
2 CutcliffeTRUE             886.     1974.     0.449 6.55e- 1
3 OppFPI_PrevYear            67.6      80.6    0.840 4.04e- 1
4 First_GameTRUE           3982.     3162.     1.26  2.13e- 1
5 UNC_GameTRUE            12475.     2728.     4.57  2.41e- 5
6 OppRankedGametimeTRUE    6832.     2014.     3.39  1.22e- 3
```

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

```
[1] 0.3564153
```

Below are models that combine the variables from Section 1 with the variables explored in the model above to create models with the highest-possible adjusted r-squared values:

```r
late_jan_pred_lm_merge <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ Cutcliffe + OppFPI_PrevYear +
        First_Game + UNC_Game + OppRankedGametime +
        FPI_Diff_PrevYear + Renovated + Day,
      data = home_att_data_2)

tidy(late_jan_pred_lm_merge)
```

```
# A tibble: 11 x 5
   term                  estimate std.error statistic     p.value
   <chr>                    <dbl>     <dbl>     <dbl>       <dbl>
 1 (Intercept)             28005.     5136.      5.45  0.00000116
 2 CutcliffeTRUE           -2679.     2116.     -1.27  0.211
 3 OppFPI_PrevYear           386.      125.      3.09  0.00314
```

```
 4 First_GameTRUE          4650.     4393.     1.06   0.294
 5 UNC_GameTRUE           15582.     2736.     5.70   0.000000474
 6 OppRankedGametimeTRUE   6456.     1931.     3.34   0.00148
 7 FPI_Diff_PrevYear       -284.      105.    -2.70   0.00920
 8 RenovatedTRUE          -2956.     1507.    -1.96   0.0548
 9 DayMon                 -7816.     6876.    -1.14   0.260
10 DaySat                 -1887.     4372.    -0.432  0.668
11 DayThu                 -9955.     5225.    -1.91   0.0619
```

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

```
[1] 0.4585018
```

```r
late_jan_pred_lm_merge_noFPI <- linear_reg() |>
  set_engine("lm") |>
  fit(AttNum ~ Cutcliffe + First_Game +
        UNC_Game + OppRankedGametime +
        Renovated + Day + Win_Streak,
      data = home_att_data_2)

tidy(late_jan_pred_lm_merge_noFPI)
```

```
# A tibble: 10 x 5
   term                  estimate std.error statistic    p.value
   <chr>                    <dbl>     <dbl>     <dbl>      <dbl>
 1 (Intercept)             18941.     4131.     4.59  0.0000197
 2 CutcliffeTRUE             323.     1960.     0.165 0.870
 3 First_GameTRUE           6996.     2340.     2.99  0.00387
 4 UNC_GameTRUE            15603.     2947.     5.30  0.00000134
 5 OppRankedGametimeTRUE    7215.     1856.     3.89  0.000230
 6 RenovatedTRUE           -1077.     1462.    -0.737 0.464
 7 DayMon                  -2583.     6914.    -0.374 0.710
 8 DaySat                   3174.     3238.     0.980 0.330
 9 DayThu                  -3000.     4541.    -0.661 0.511
10 Win_Streak               1073.      399.     2.69  0.00902
```

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

```
[1] 0.3717644
```

The second model out of the two above does not contain variables related to FPI. This model will be used for Elon only since it has weaker predictive power and because no FPI values are available from ESPN for Elon.

Average attendance figures from 2022-23 (used in the visualizations in the Final Predictions section):

```
home_att_data_2 |>
  filter(!is.na(AttNum)) |>
  filter(Year >= 2022) |>
  summarize("Mean Att. (#)" = mean(AttNum),
            "Mean Att. (%)" = mean(AttPct))
```

```
# A tibble: 1 x 2
  `Mean Att. (#)` `Mean Att. (%)`
            <dbl>           <dbl>
1          24831.            62.1
```

Final Predictions

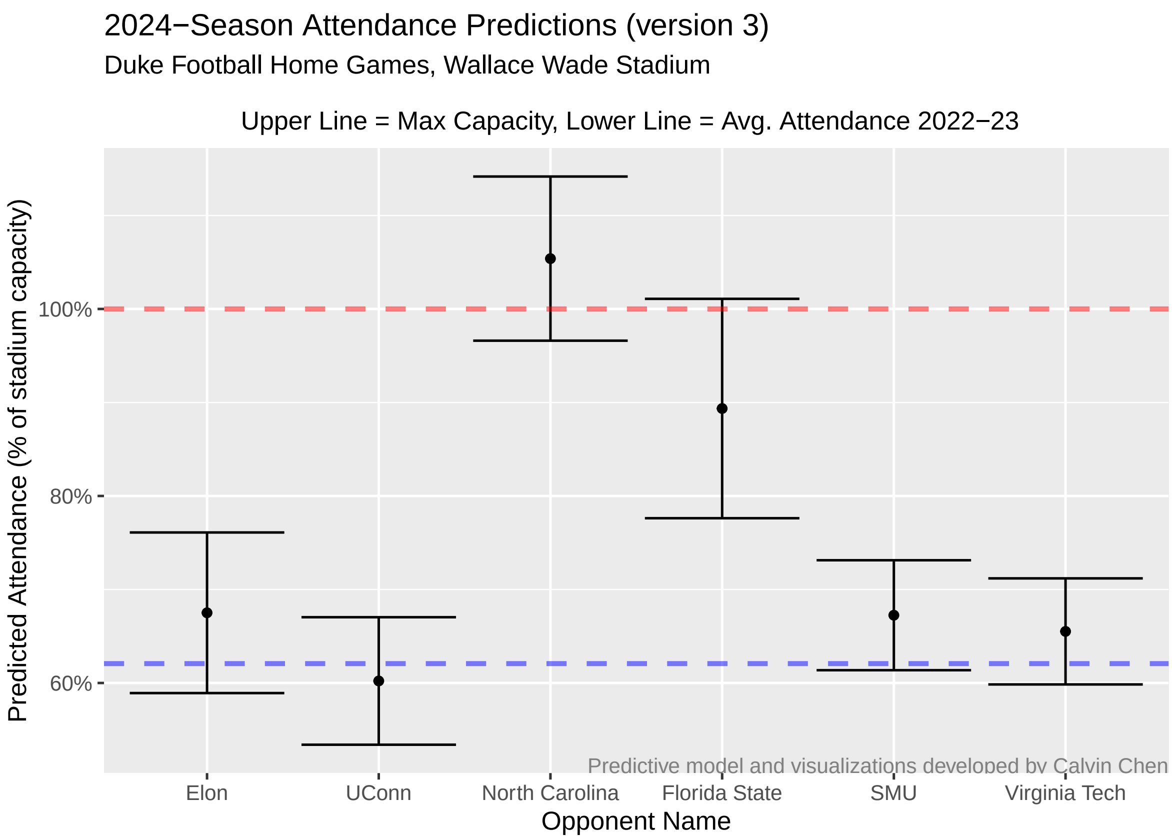
```
Elon_2024 <- att_data_2 |>
  filter(Year == 2024, OppName == "Elon")
Elon_2024_pred <- predict(late_jan_pred_lm_merge_noFPI$fit,
                          Elon_2024,
                          type = "response",
                          se.fit = TRUE)
Elon_2024_pred_se <- Elon_2024_pred$se.fit
#Elon_2024_pred
#Elon_2024_pred_se

UConn_2024 <- att_data_2 |>
  filter(Year == 2024, OppName == "Connecticut")
UConn_2024_pred <- predict(late_jan_pred_lm_merge$fit,
                           UConn_2024,
                           type = "response",
                           se.fit = TRUE)
UConn_2024_pred_se <- UConn_2024_pred$se.fit
#UConn_2024_pred
#UConn_2024_pred_se

UNC_2024 <- att_data_2 |>
```

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

FSU_2024 <- att_data_2 |>
  filter(Year == 2024, OppName == "Florida St.")
FSU_2024_pred <- predict(late_jan_pred_lm_merge$fit,
                         FSU_2024,
                         type = "response",
                         se.fit = TRUE)
FSU_2024_pred_se <- FSU_2024_pred$se.fit
#FSU_2024_pred

SMU_2024 <- att_data_2 |>
  filter(Year == 2024, OppName == "SMU")
SMU_2024_pred <- predict(late_jan_pred_lm_merge$fit,
                         SMU_2024,
                         type = "response",
                         se.fit = TRUE)
SMU_2024_pred_se <- SMU_2024_pred$se.fit
#SMU_2024_pred

VT_2024 <- att_data_2 |>
  filter(Year == 2024, OppName == "Virginia Tech")
VT_2024_pred <- predict(late_jan_pred_lm_merge$fit,
                        VT_2024,
                        type = "response",
                        se.fit = TRUE)
VT_2024_pred_se <- VT_2024_pred$se.fit
#VT_2024_pred

late_jan_pred_model_output <- tibble(
  Name = c("Elon",
           "UConn",
           "North Carolina",
           "Florida State",
           "SMU",
           "Virginia Tech"),
```

```r
  "Attendance %" = c(Elon_2024_pred$fit / 40004 * 100,
                     UConn_2024_pred$fit / 40004 * 100,
                     UNC_2024_pred$fit / 40004 * 100,
                     FSU_2024_pred$fit / 40004 * 100,
                     SMU_2024_pred$fit / 40004 * 100,
                     VT_2024_pred$fit / 40004 * 100),
  "Standard Error (Att. #)" = c(Elon_2024_pred$se.fit,
                               UConn_2024_pred$se.fit,
                               UNC_2024_pred$se.fit,
                               FSU_2024_pred$se.fit,
                               SMU_2024_pred$se.fit,
                               VT_2024_pred$se.fit),
  "Estimated # of People" = c(Elon_2024_pred$fit,
                                UConn_2024_pred$fit,
                                UNC_2024_pred$fit,
                                FSU_2024_pred$fit,
                                SMU_2024_pred$fit,
                                VT_2024_pred$fit)
)

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

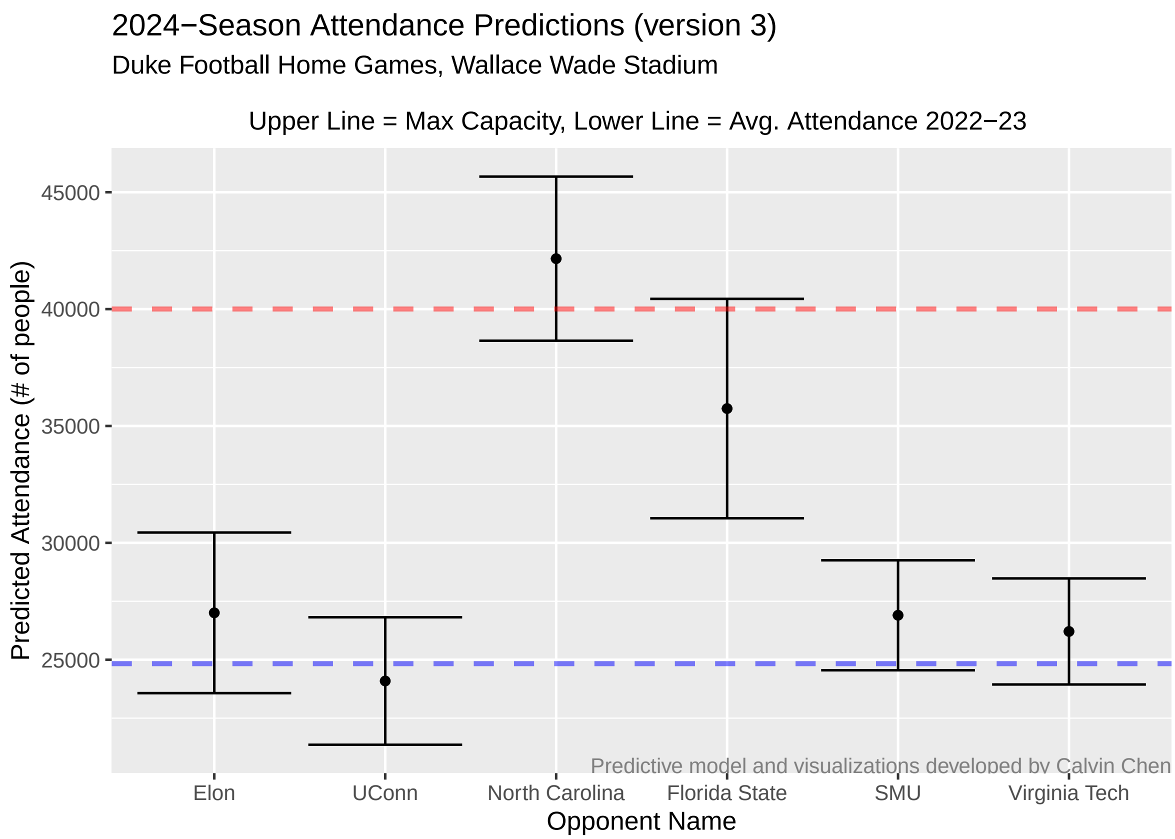
```r
late_jan_pred_model_data |>
  ggplot(
    aes(x = fct_relevel(Name,
                        "Elon",
                        "UConn",
                        "North Carolina",
                        "Florida State",
                        "SMU",
                        "Virginia Tech"),
        y = AttPct)
  ) +
  geom_point() +
  geom_errorbar(aes(ymin = AttPct - SdErrPct, ymax = AttPct + SdErrPct)) +
  labs(title = "2024-Season Attendance Predictions (version 3)",
       subtitle = "Duke Football Home Games, Wallace Wade Stadium\n
```

```
                   Upper Line = Max Capacity, Lower Line = Avg. Attendance 2022-23",
       x = "Opponent Name",
       y = "Predicted Attendance (% of stadium capacity)") +
  geom_hline(yintercept = 100, color = "red",
             linetype = "dashed", size = 1, alpha = 0.5) +
  geom_hline(yintercept = 62.07187, color = "blue",
             linetype = "dashed", size = 1,  alpha = 0.5) +
  scale_y_continuous(labels = function(x) paste0(x, "%")) +
  annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
           label = "Predictive model and visualizations developed by Calvin Chen",
           size = 3, color = "gray50")
```




```
late_jan_pred_model_data |>
  ggplot(
    aes(x = fct_relevel(Name,
                        "Elon",
                        "UConn",
```

```
                        "North Carolina",
                        "Florida State",
                        "SMU",
                        "Virginia Tech"),
      y = AttNum)
) +
geom_point() +
geom_errorbar(aes(ymin = AttNum - SdErr, ymax = AttNum + SdErr)) +
labs(title = "2024-Season Attendance Predictions (version 3)",
     subtitle = "Duke Football Home Games, Wallace Wade Stadium\n
                 Upper Line = Max Capacity, Lower Line = Avg. Attendance 2022-23",
     x = "Opponent Name",
     y = "Predicted Attendance (# of people)") +
geom_hline(yintercept = 40004, color = "red",
           linetype = "dashed", size = 1, alpha = 0.5) +
geom_hline(yintercept = 24831.23, color = "blue",
           linetype = "dashed", size = 1, alpha = 0.5) +
annotate("text", x = Inf, y = -Inf, hjust = 1, vjust = 0,
         label = "Predictive model and visualizations developed by Calvin Chen",
         size = 3, color = "gray50")
```



The graphs above do take predictive liberties by assuming that *Florida State* will be *ranked* (in the top 25) at game-time.

Additional Commentary Coming Soon.