Predictions Based on 2011-2023 Home Attendance Records

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.

Packages

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

Importing the Dataset

List 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))

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

home_att_data
```

```
# A tibble: 84 x 51
   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 45 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")
```

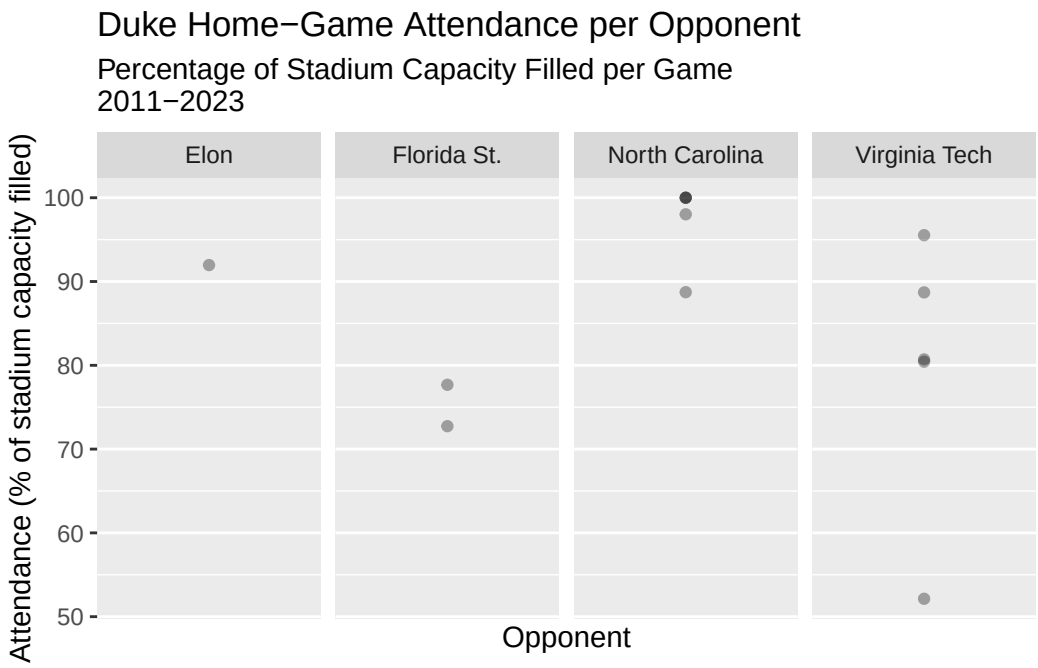
Attendance History for 2024 Home Opponents

Wallace Wade Stadium capacity:

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

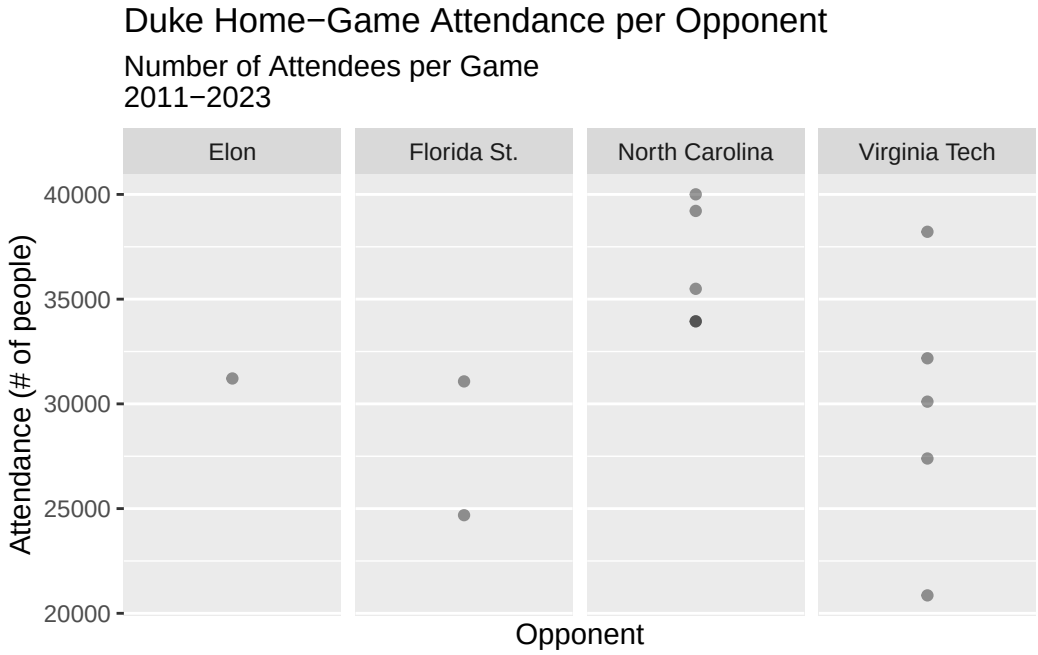
All Teams

```
home_att_data |>
  filter(OppName %in% home_opp_list) |>
  ggplot(
    aes(x = 0, y = AttPct)
  ) +
  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)")
```




```
home_att_data |>
  filter(OppName %in% home_opp_list) |>
```

```
  ggplot(
    aes(x = 0, y = AttNum)
  ) +
  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)")
```



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>
```

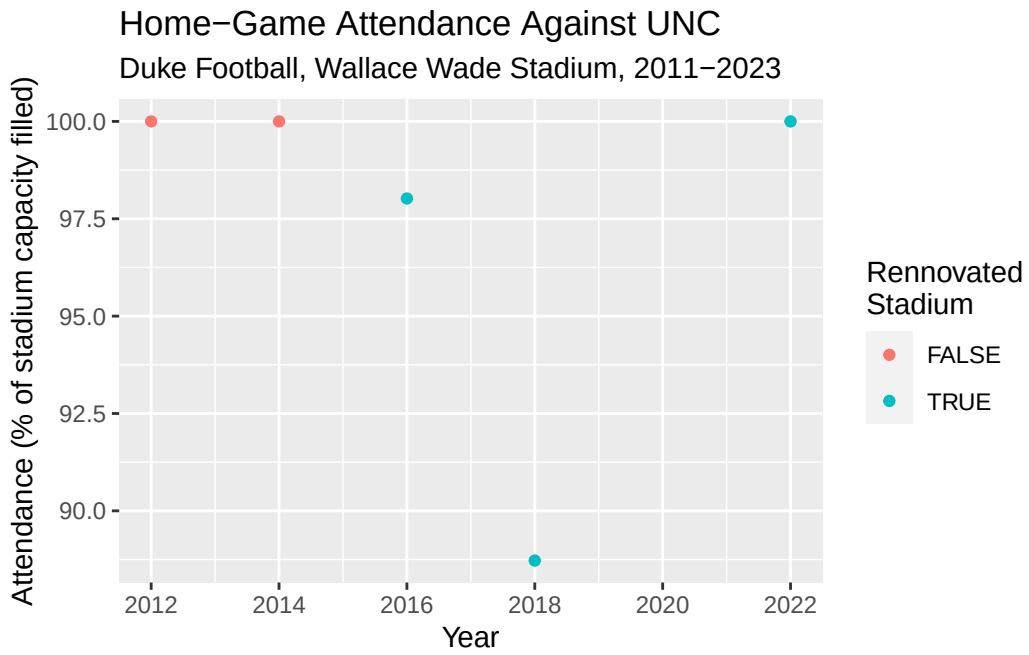
North Carolina

```
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)
```

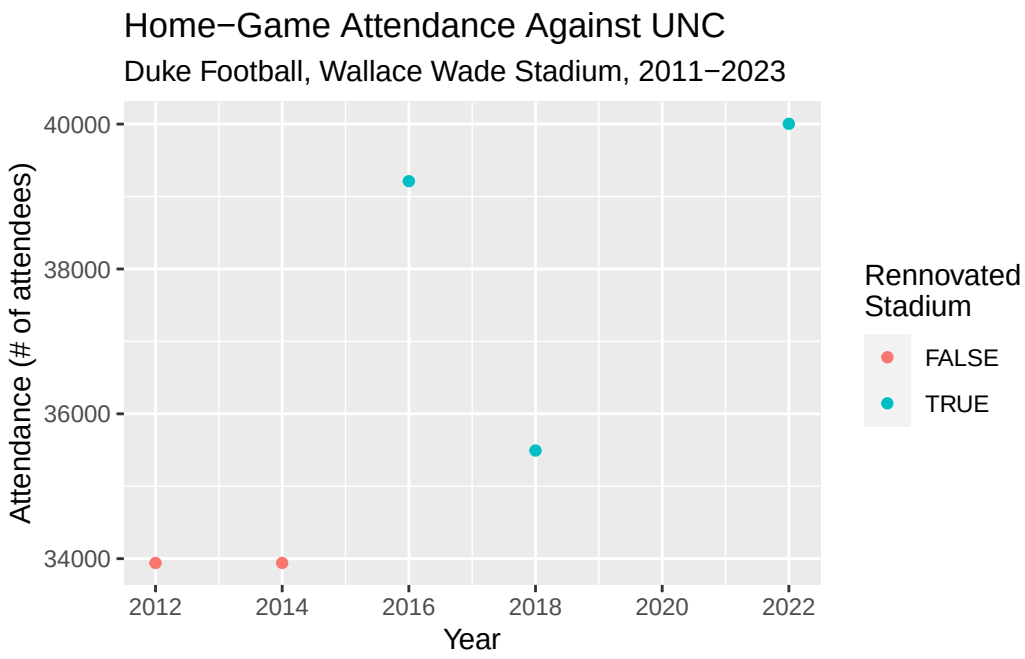
```
# A tibble: 6 x 7
  Name           `End-of-Season FPI` Month  Date  Year `# of Attendees`
  <chr>                        <dbl> <dbl> <dbl> <dbl>            <dbl>
1 North Carolina                10.6    10    20  2012            33941
2 North Carolina                 4.4    11    20  2014            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 = Rennovated)
  ) +
  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 = "Rennovated\nStadium")
```




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

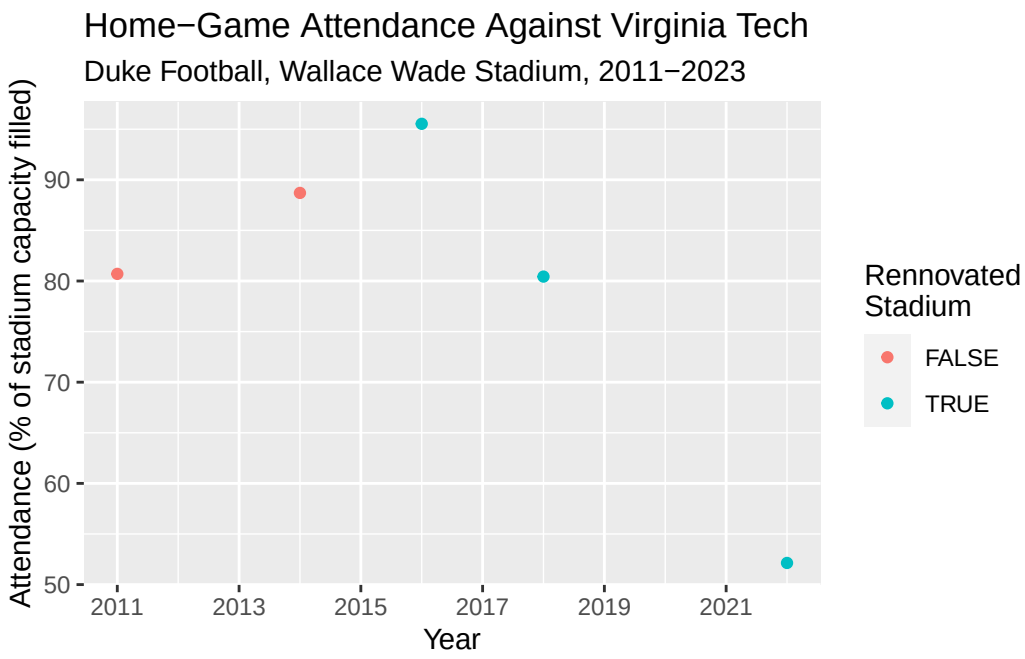
UConn never faced against Duke in Wallace Wade Stadium from 2011 to 2023.

Virginia Tech

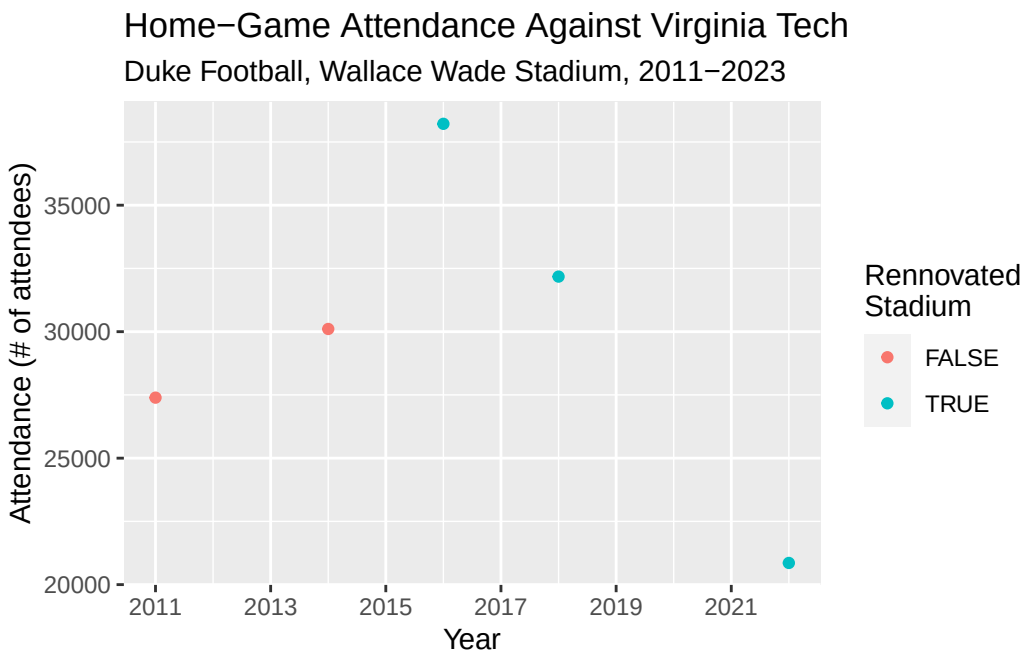
```
home_att_data |>
  filter(OppName == "Virginia Tech") |>
  summarize("Name" = OppName,
            "End-of-Season FPI" = OppFPI,
            Month,
            Date,
            Year,
            "# of Attendees" = AttNum,
            "% of Stadium Capacity" = AttPct)
```

```
# 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>
```

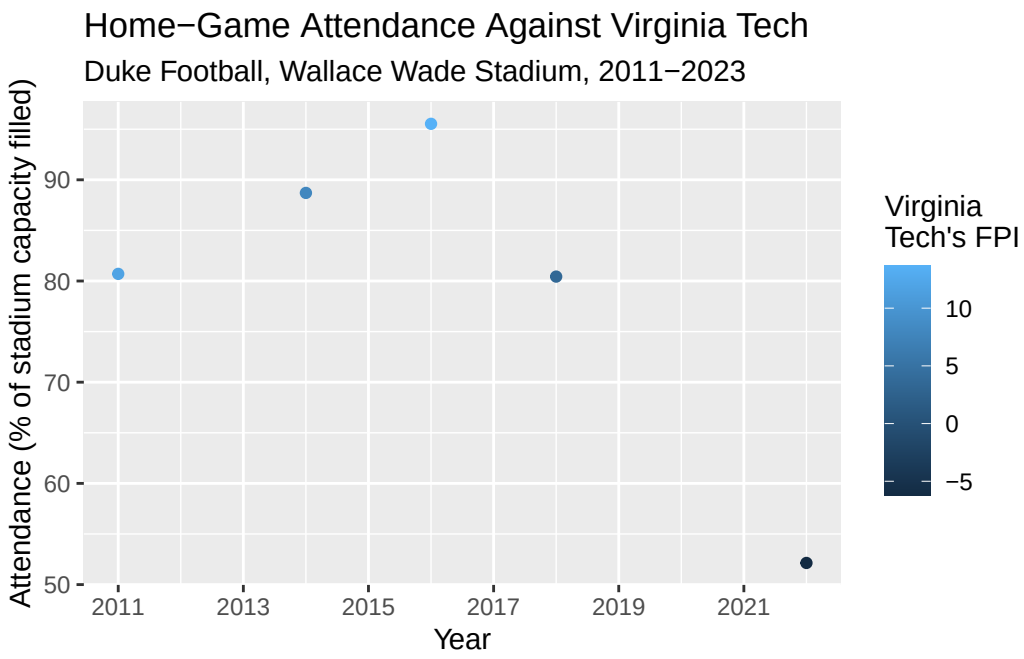
```
home_att_data |>
  filter(OppName == "Virginia Tech") |>
  ggplot(
    aes(x = Year, y = AttPct, color = Rennovated)
  ) +
  geom_point() +
  scale_x_continuous(breaks = seq(from = 2011, to = 2023, by = 2)) +
  labs(title = "Home-Game Attendance Against Virginia Tech",
       subtitle = "Duke Football,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




```
home_att_data |>
  filter(OppName == "Virginia Tech") |>
  ggplot(
    aes(x = Year, y = AttNum, color = Rennovated)
  ) +
  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 = "Rennovated\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?

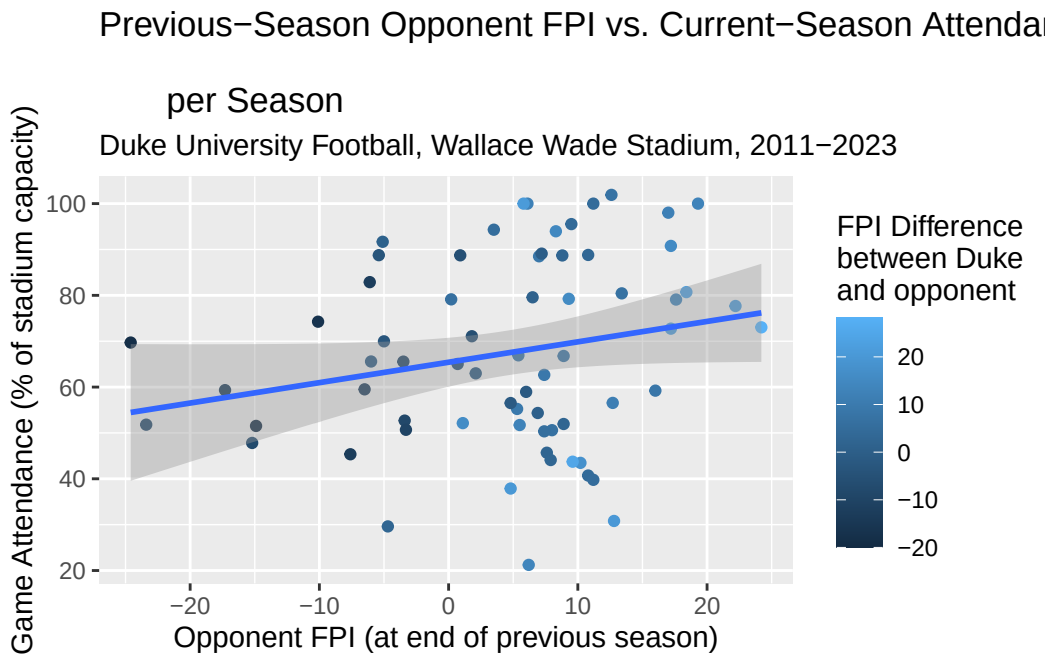
Previous-Season FPI

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
```

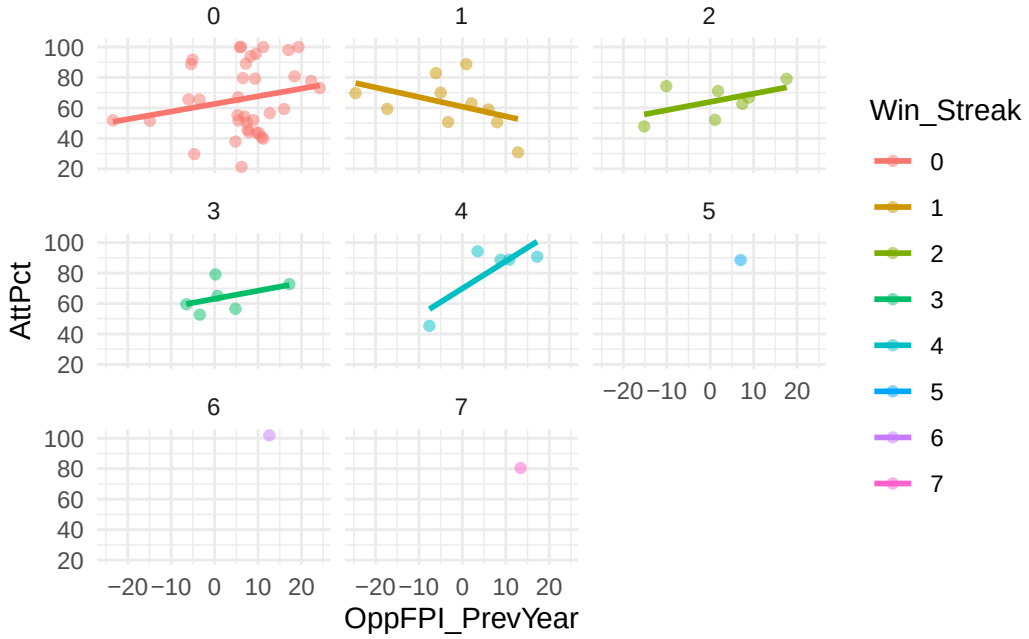
```
glance(prev_fpi_streak_lm)$adj.r.squared
```

```
[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()
```



Other Factors

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

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.