

Stats 107 Project

Omar Sepulveda, Riccardo Gutierrez

2025-11-06

Question: How has the number of NFL players from each SEC school changed over time (e.g., 2011–2015, 2016–2020, and 2021–2025), and which schools have shown the most growth or decline in player representation?

ABSTRACT

The NCAA SEC is one of the premier conferences in college football from recent years, but how much is each program represented in the draft? What we are looking to accomplish is to highlight which schools in the NCAA SEC (South Eastern Conference) have gained representation in the draft and which programs have lost representation in the draft over the last 15 years. We collected our data from Pro-Football-Reference.com and combined each program in the SEC data sets into a singular Excel file and analyzed the data by distinguishing which players came from the which college program, seeing if the player was drafted, then drafted within the years 2011-2025.

Data

The data set used for this analysis, “College_Dataset.xlsx,” contains information on NFL players and the colleges they attended. The data was obtained from Pro-Football-Reference, a reliable source that collects and publishes official NFL statistics, including player backgrounds and school affiliations. Relevant variables include each player’s name, the college they attended, and the range of years they played in the NFL. During the cleaning process, unnecessary columns such as AP1, PB, St, wAV, Ht, and Wt were removed, keeping only the relevant variables. Additionally, since the data set included players from earlier eras, we extracted the starting year from each player’s career range and filtered the data to include only those who began their NFL careers in 2011 or later. The data is entirely real and will not be generated through any randomized simulation.

Visualization

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag
```

```
## The following objects are masked from 'package:base':
##
## intersect, setdiff, setequal, union
```

```
# Read the Excel file
data <- read_excel("College Dataset.xlsx")
```

```
## Warning: Expecting date in H2740 / R2740C8: got '0-0'
```

```
## New names:
## * '' -> '...3'
## * '' -> '...11'
```

```
# Rename columns for clarity (based on the header)
colnames(data) <- c("Pos", "Player", "College_Stats", "AP1", "PB", "St", "wAV", "Ht", "Wt", "Yrs_Num", "School")

# Extract the start year from Years_Range
data <- data %>%
  mutate(start_year = as.integer(sub("-", "", Years_Range)))

# Filter to keep players who started in 2011 or later
filtered_data <- data %>%
  filter(start_year >= 2011)

# Select only the specified columns
fixed_data <- filtered_data %>%
  select(Pos, Player, Yrs_Num, Years_Range, School)

# Rename columns as per user labels if desired
colnames(fixed_data) <- c("Position", "Player", "Years Played in NFL", "Years Played in NFL (From-To)", "School")

# View the selected data (or you can use write_excel_csv to save it)
head(fixed_data)
```

```
## # A tibble: 6 x 5
##   Position Player      'Years Played in NFL' Years Played in NFL ~1 School
##   <chr>    <chr>          <dbl> <chr>                <chr>
## 1 DT      Jonathan Allen      9 2017-2025      Alaba~
## 2 CB      Trey Amos          1 2025-2025      Alaba~
## 3 OLB     Ryan Anderson      5 2017-2022      Alaba~
## 4 LB      Will Anderson      3 2023-2025      Alaba~
## 5 DB      Jalyn Armour-Dav~  4 2022-2025      Alaba~
## 6 DB      Terrion Arnold     2 2024-2025      Alaba~
## # i abbreviated name: 1: 'Years Played in NFL (From-To)'
```

```
# Load necessary libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v forcats 1.0.1    v readr 2.1.5
## v ggplot2 4.0.0    v stringr 1.5.2
## v lubridate 1.9.4  v tibble 3.3.0
```

```

## v purrr      1.1.0      v tidyr      1.3.1
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

# Read the CSV file (update the path if needed)
df <- read.csv("my_updated_dataset.csv")

# Convert 'Years Played (From-To)' into numeric start and end years
df <- df %>%
  separate(`Years.Played.From.To`, into = c("Start", "End"), sep = "-", convert = TRUE)

# Define year ranges
df <- df %>%
  mutate(YearRange = case_when(
    Start >= 2011 & Start <= 2015 ~ "2011-2015",
    Start >= 2016 & Start <= 2020 ~ "2016-2020",
    Start >= 2021 & Start <= 2025 ~ "2021-2025",
    TRUE ~ NA_character_
  ))

# Count players per team and year range
team_summary <- df %>%
  filter(!is.na(YearRange)) %>%
  group_by(School, YearRange) %>%
  summarise(Players = n(), .groups = "drop")

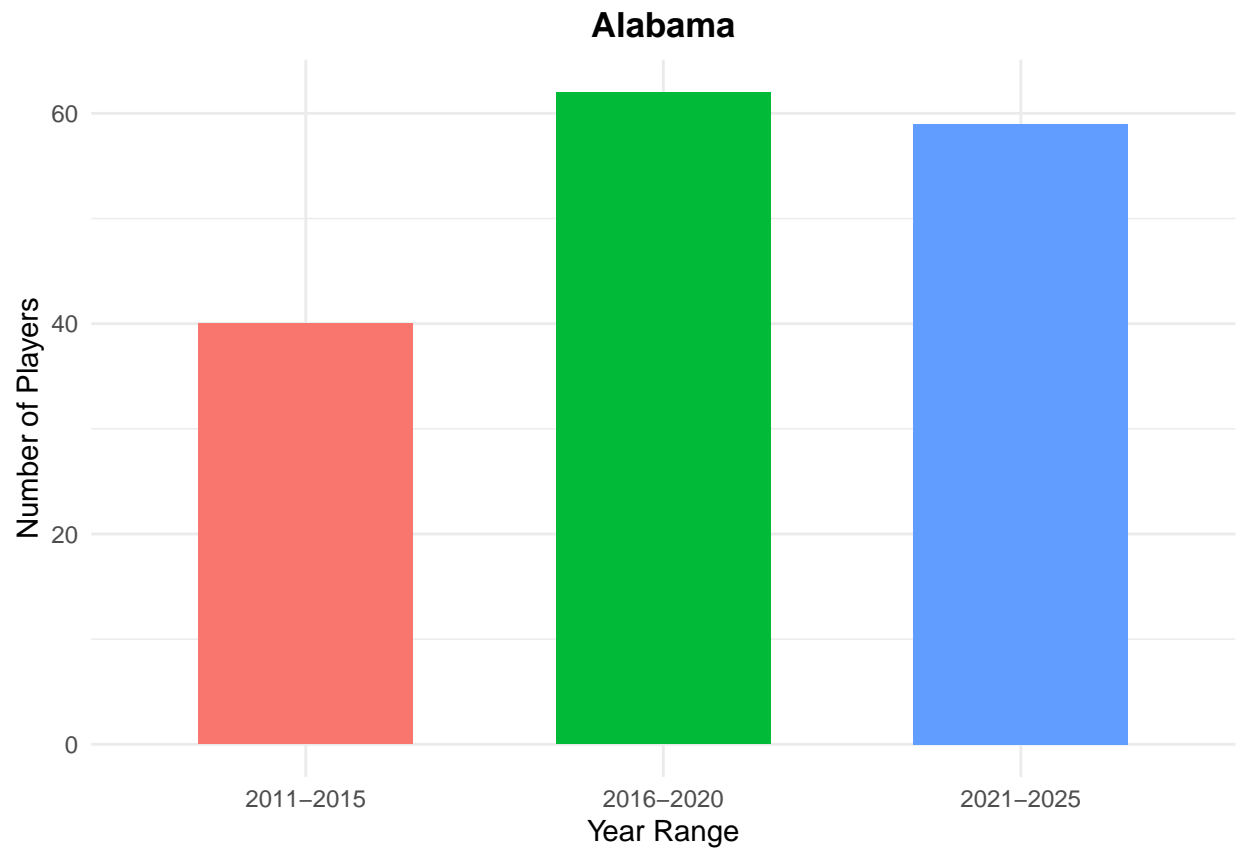
# List of teams
teams <- unique(team_summary$School)

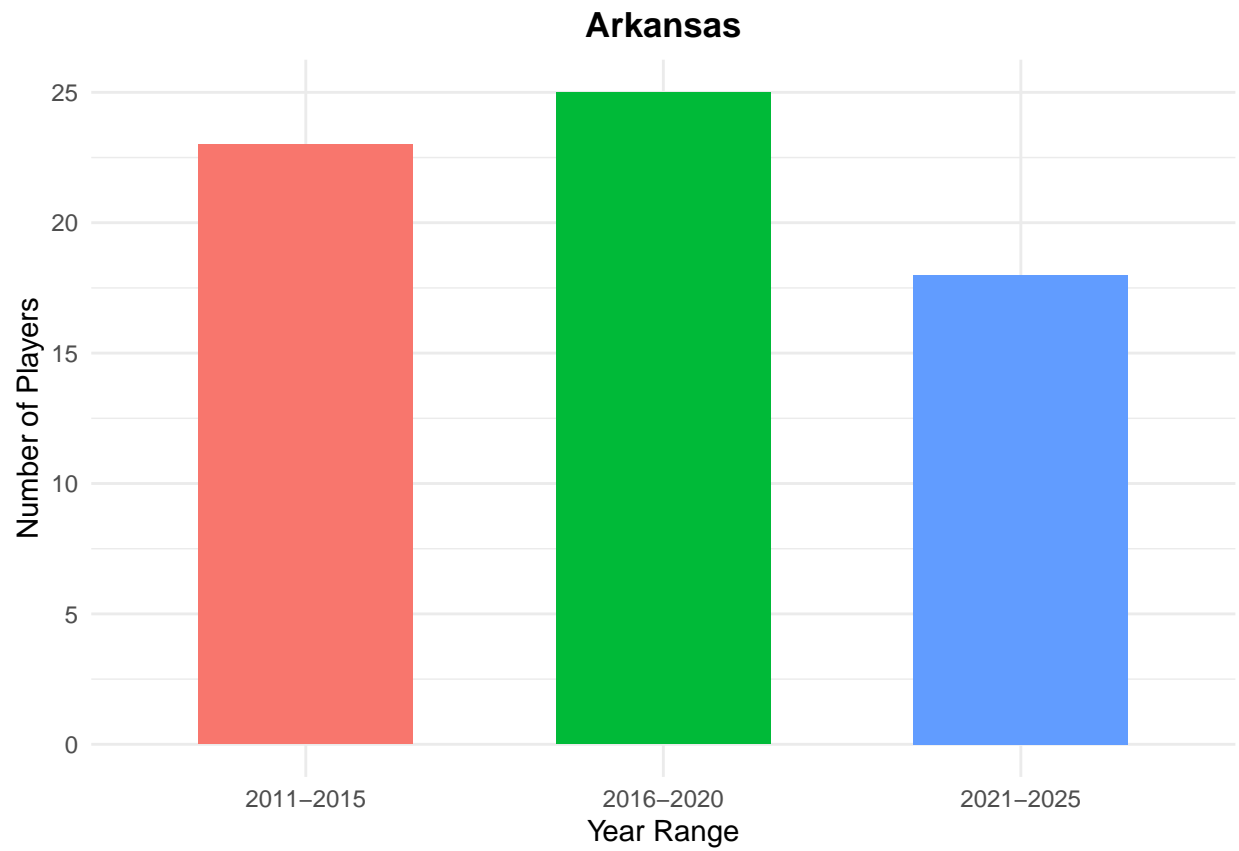
# Create a bar plot for each team
for (team in teams) {
  team_data <- team_summary %>% filter(School == team)

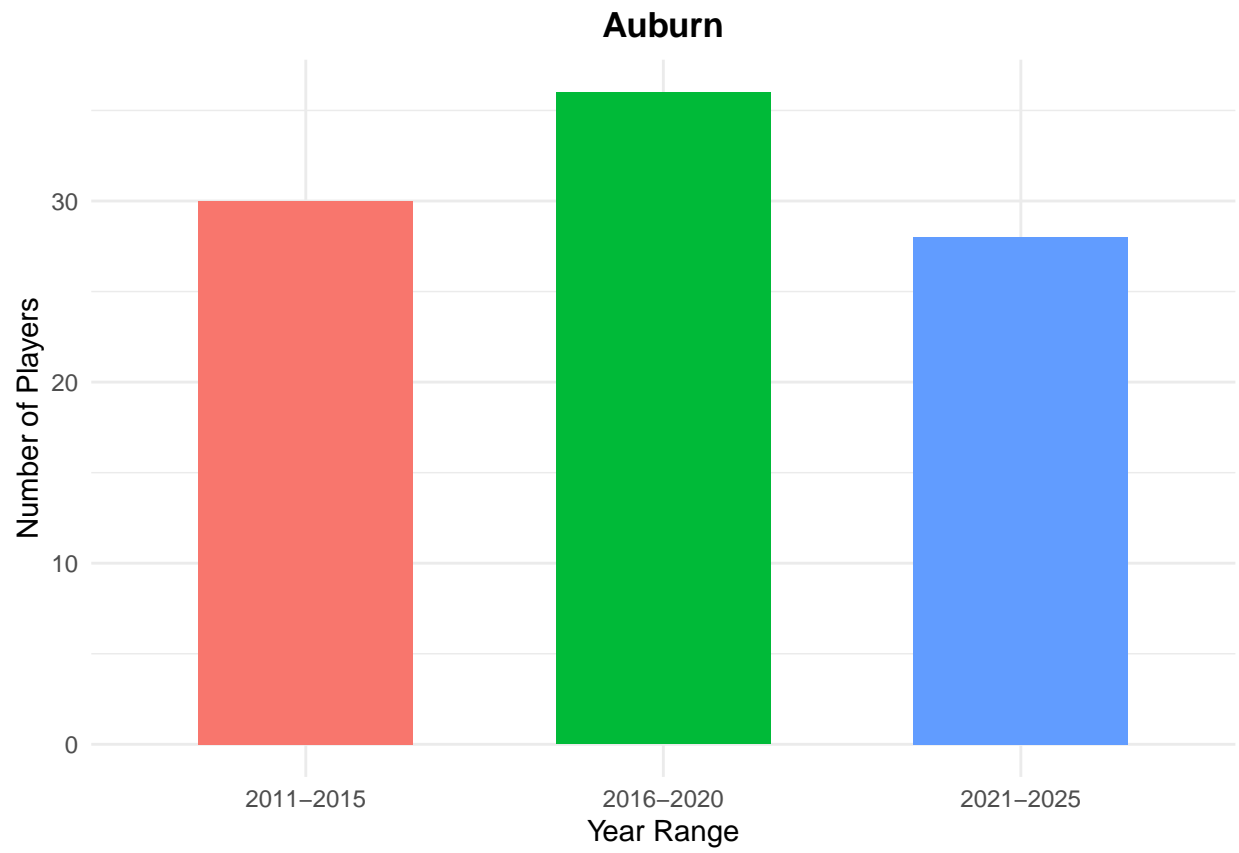
  p <- ggplot(team_data, aes(x = YearRange, y = Players, fill = YearRange)) +
    geom_bar(stat = "identity", width = 0.6) +
    labs(title = team,
         x = "Year Range",
         y = "Number of Players") +
    theme_minimal() +
    theme(legend.position = "none",
          plot.title = element_text(hjust = 0.5, face = "bold"))

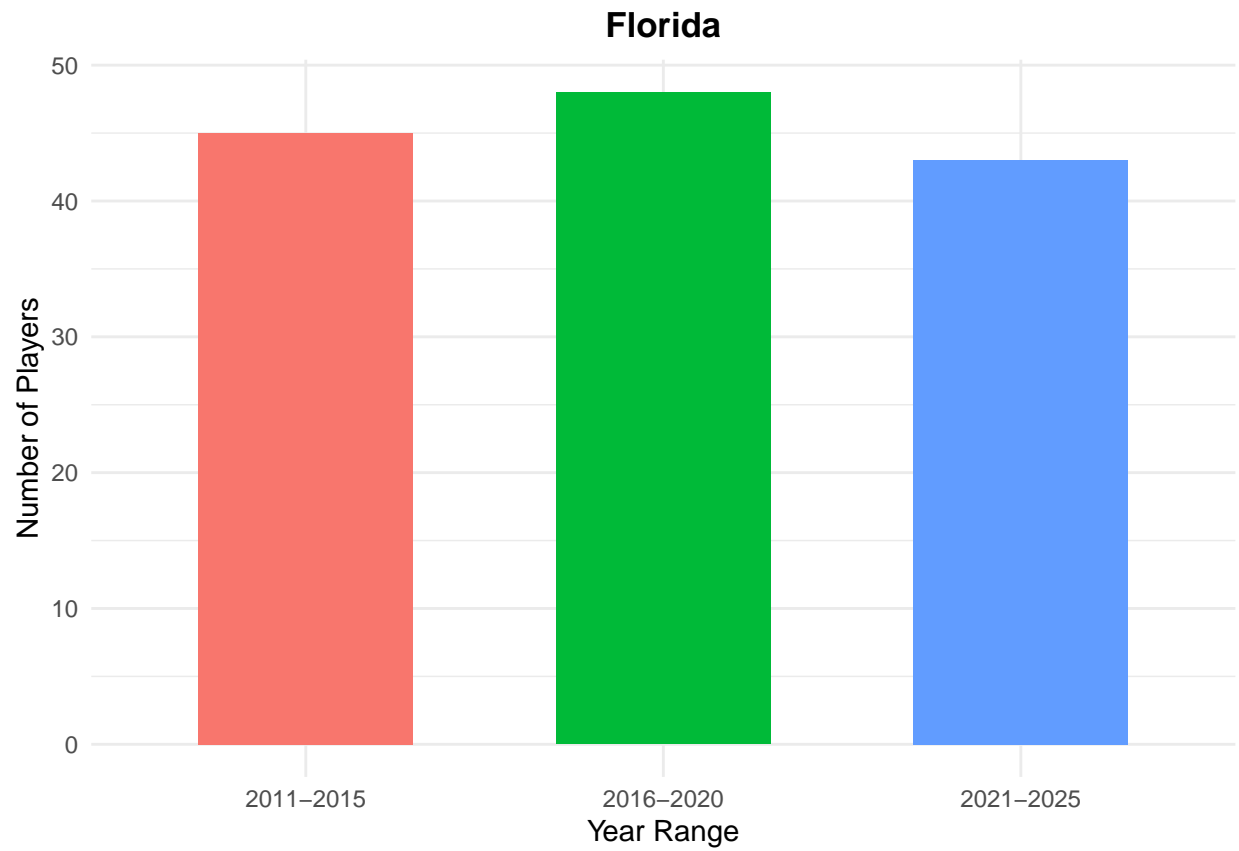
  print(p)
}

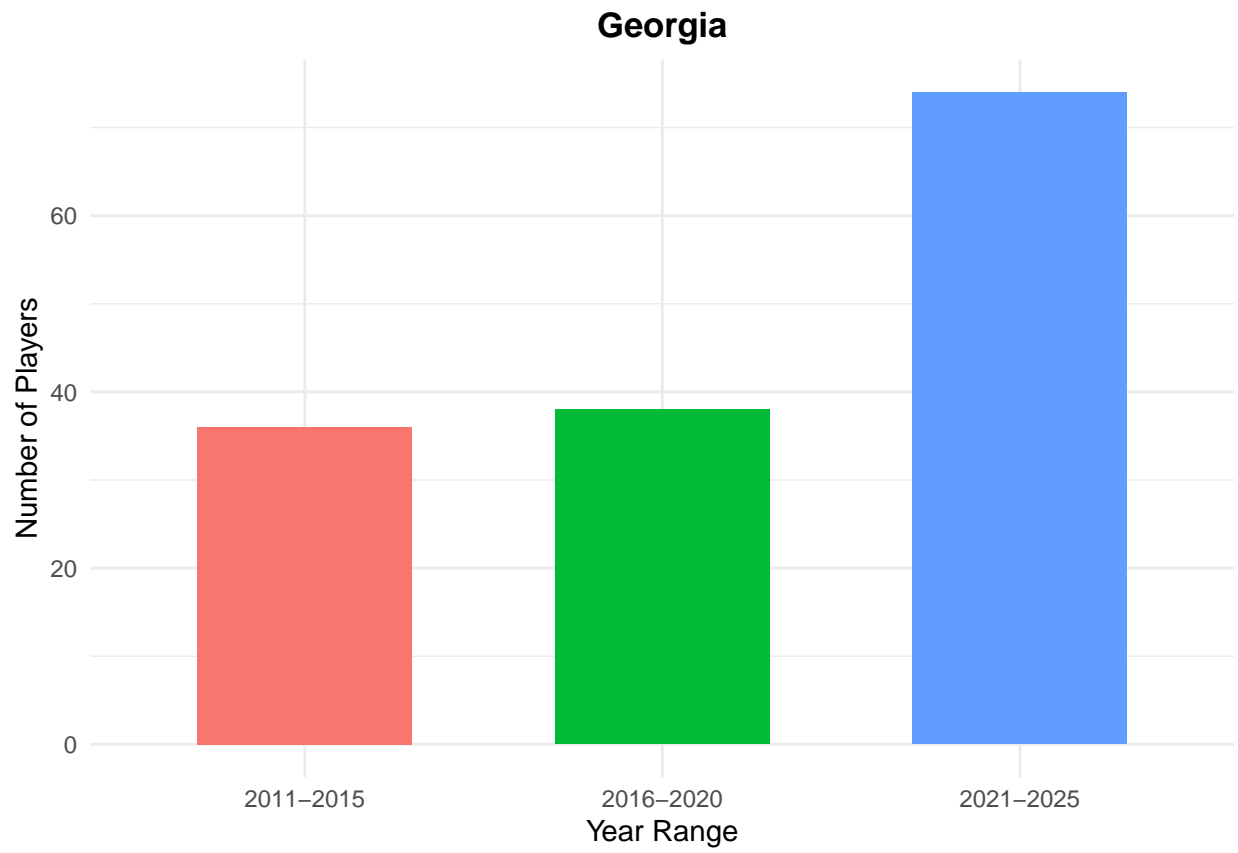
```

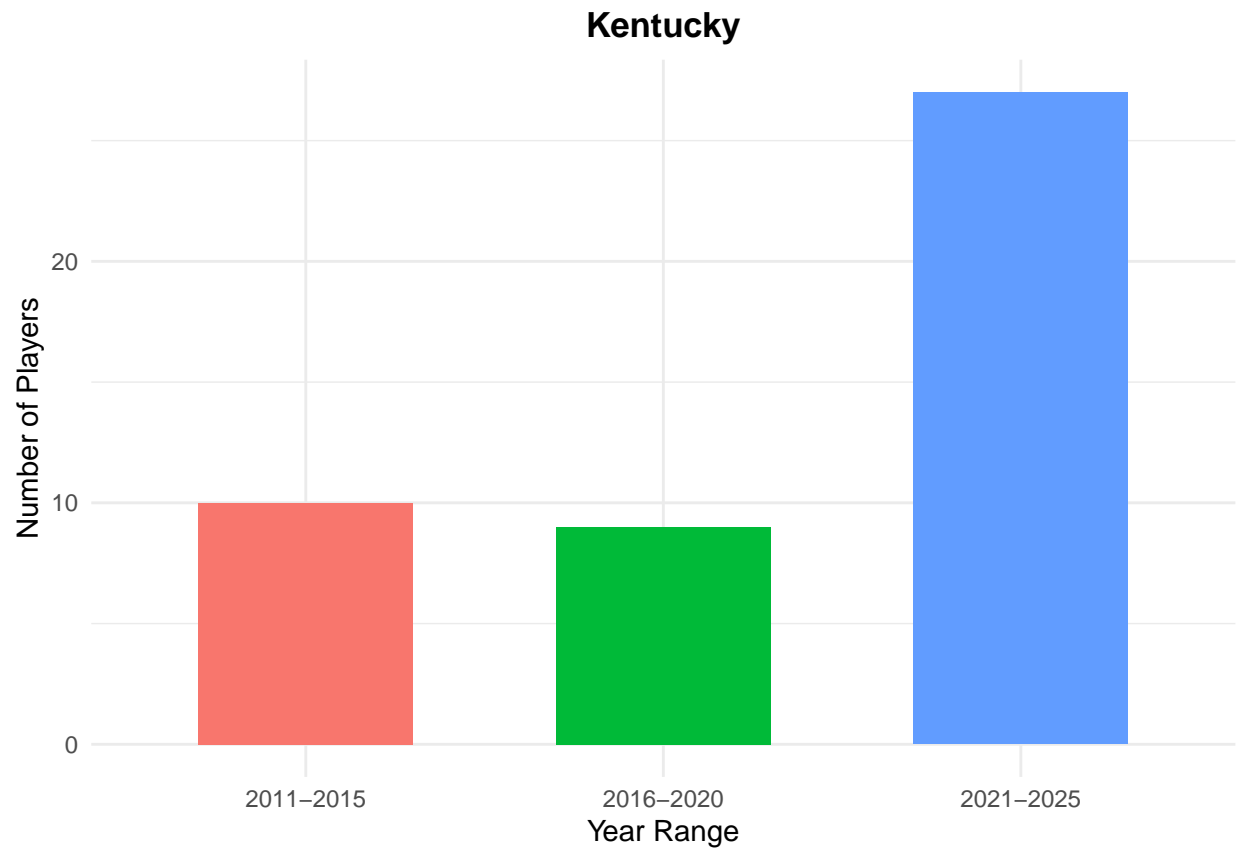


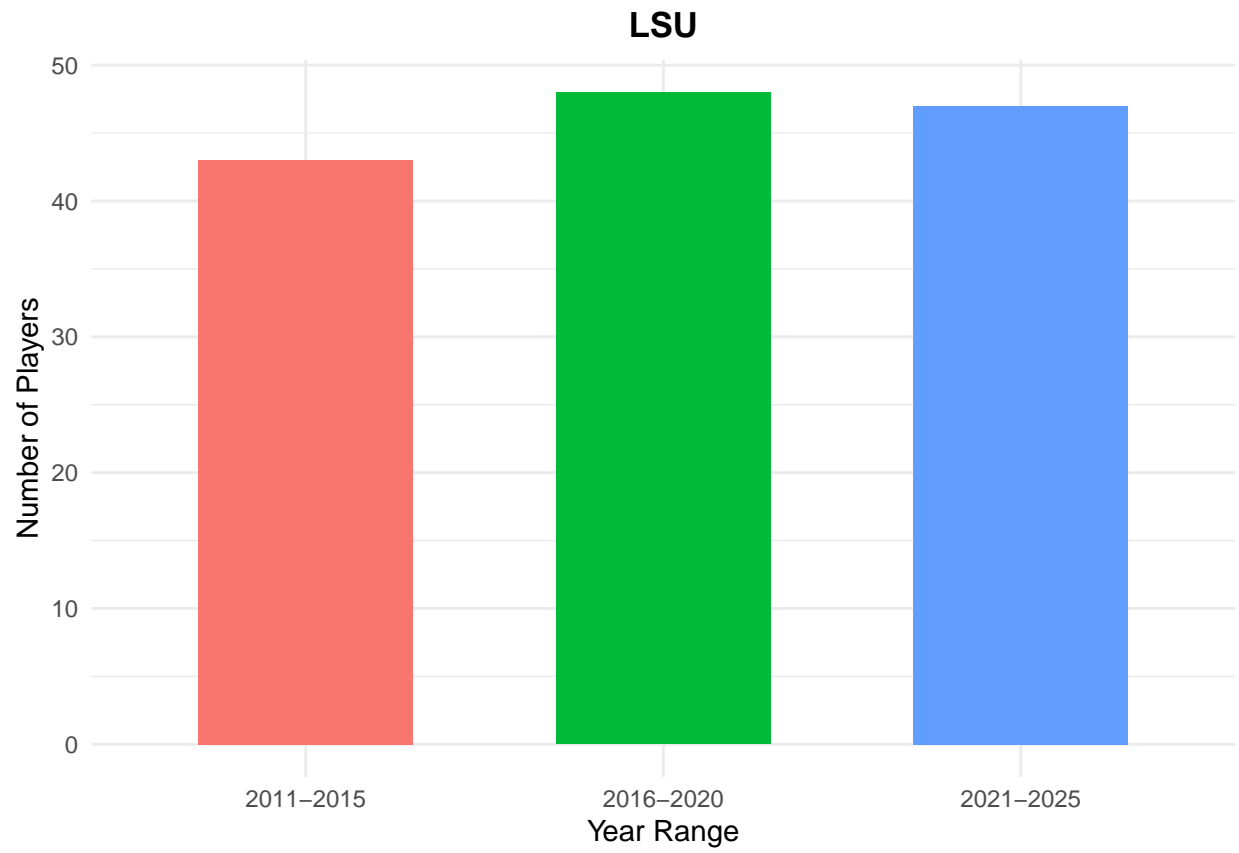


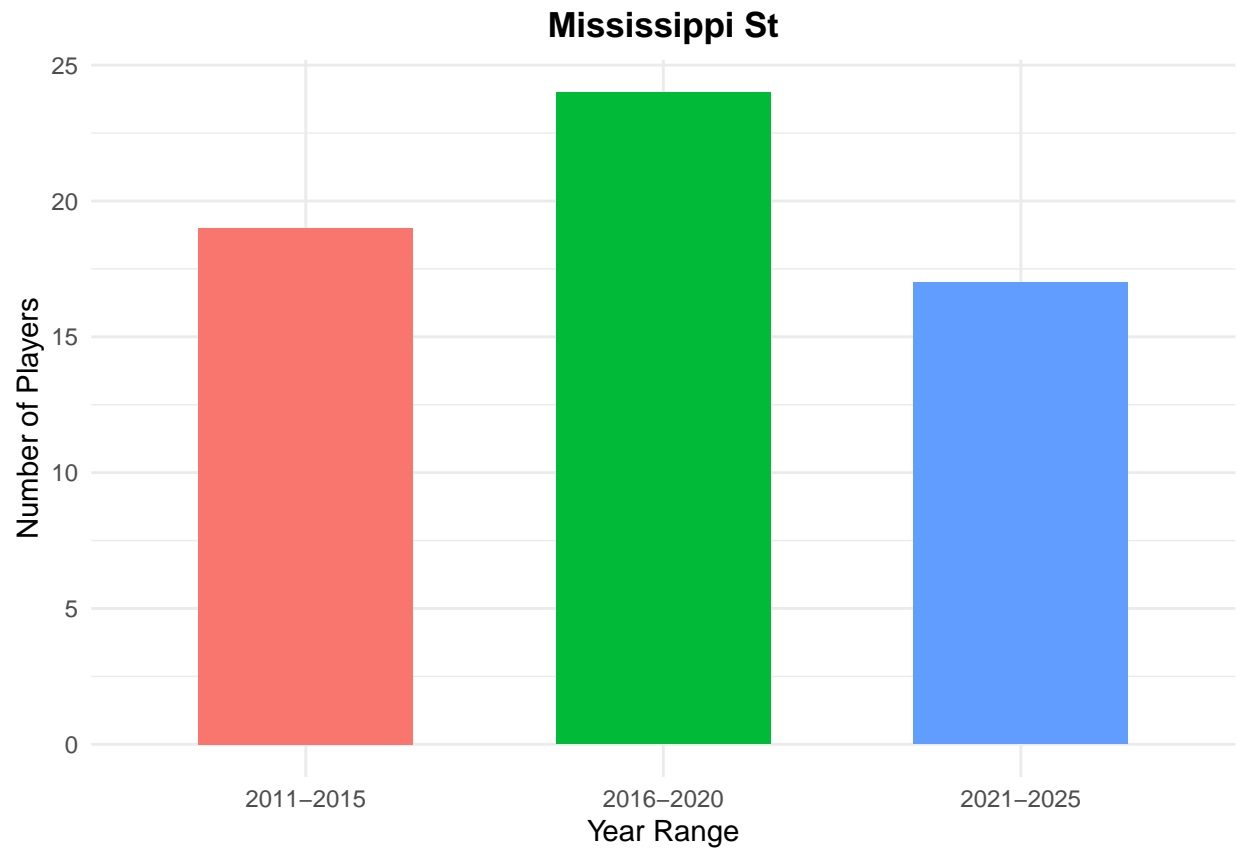


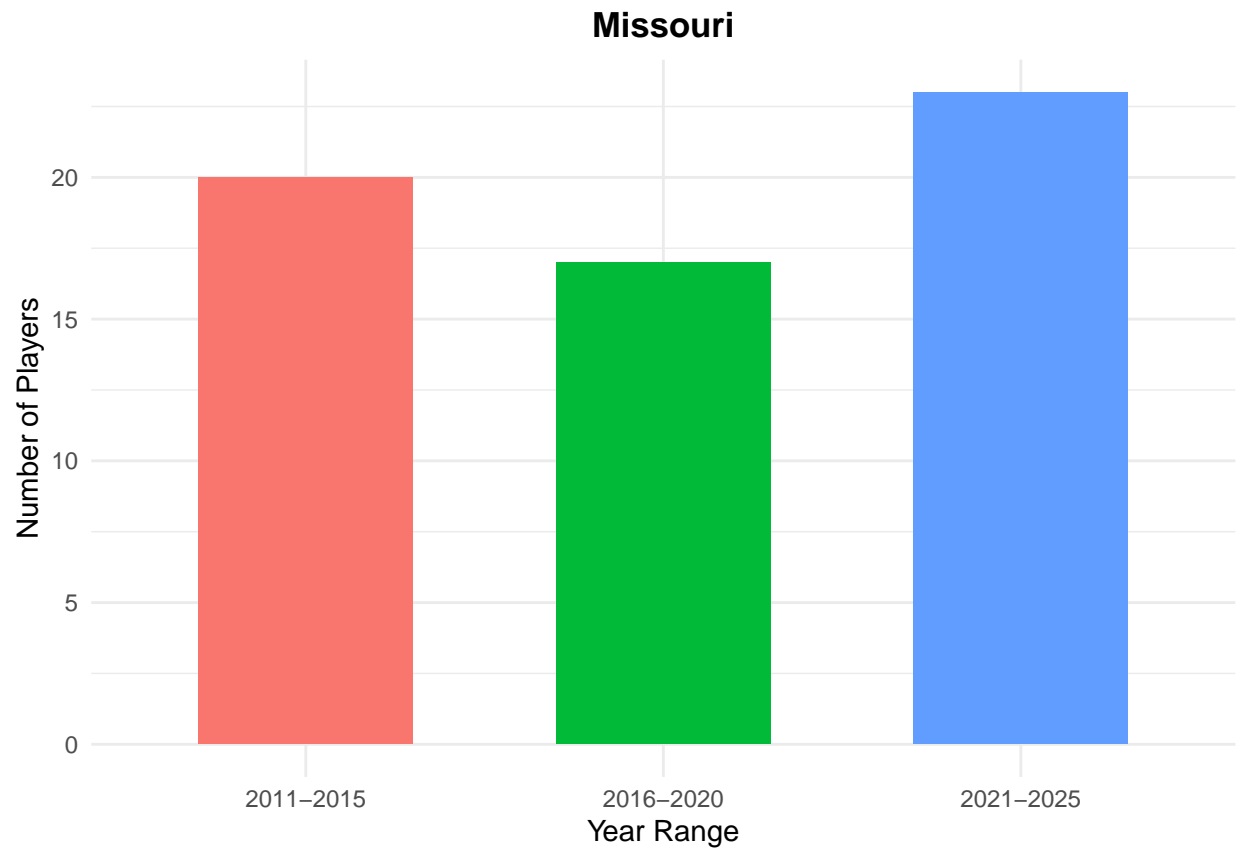


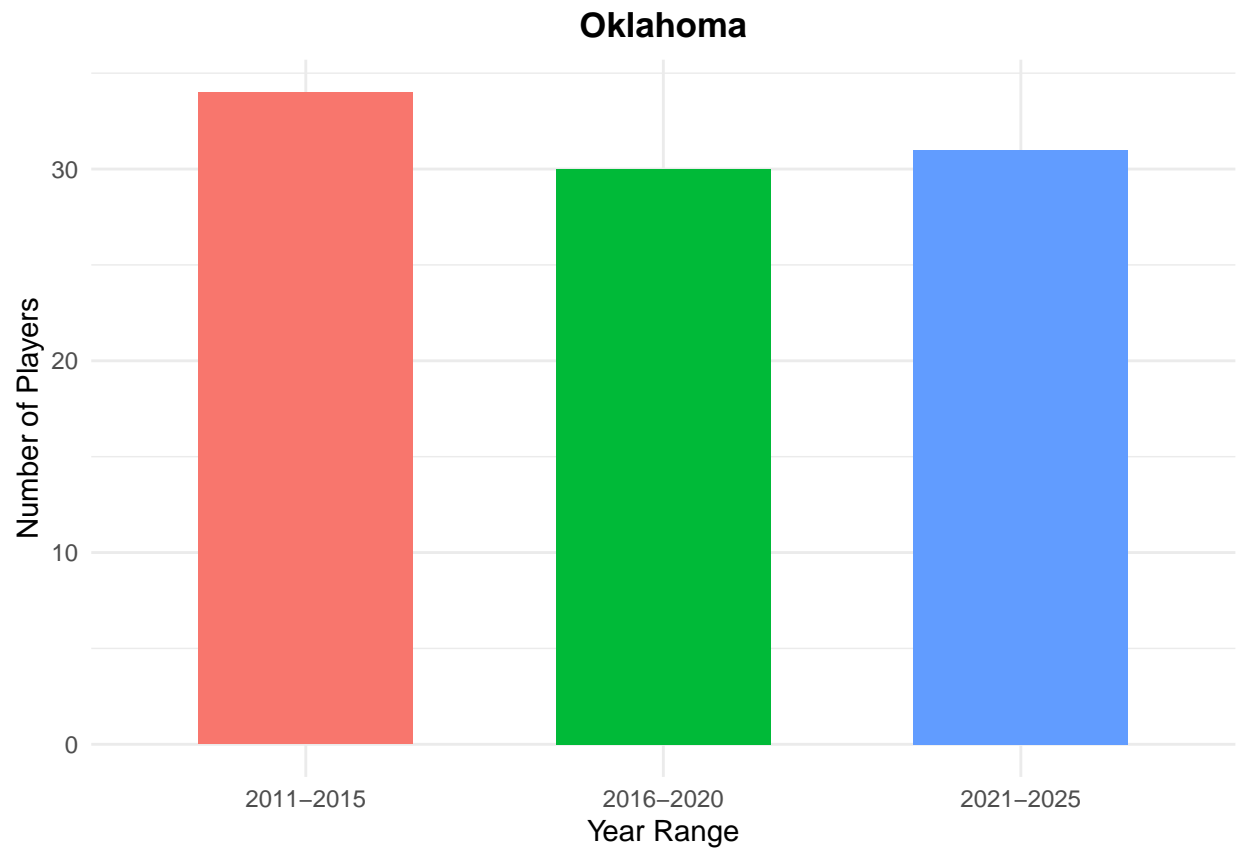


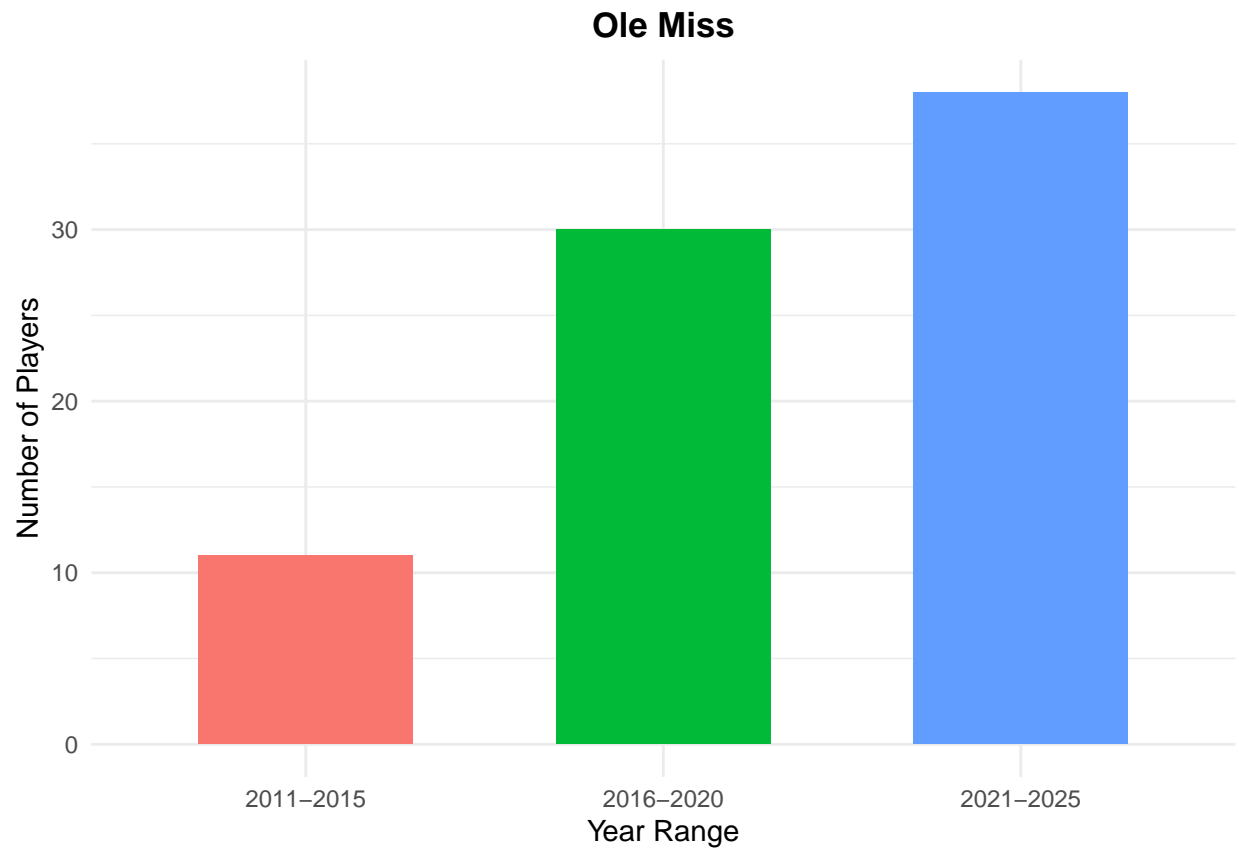


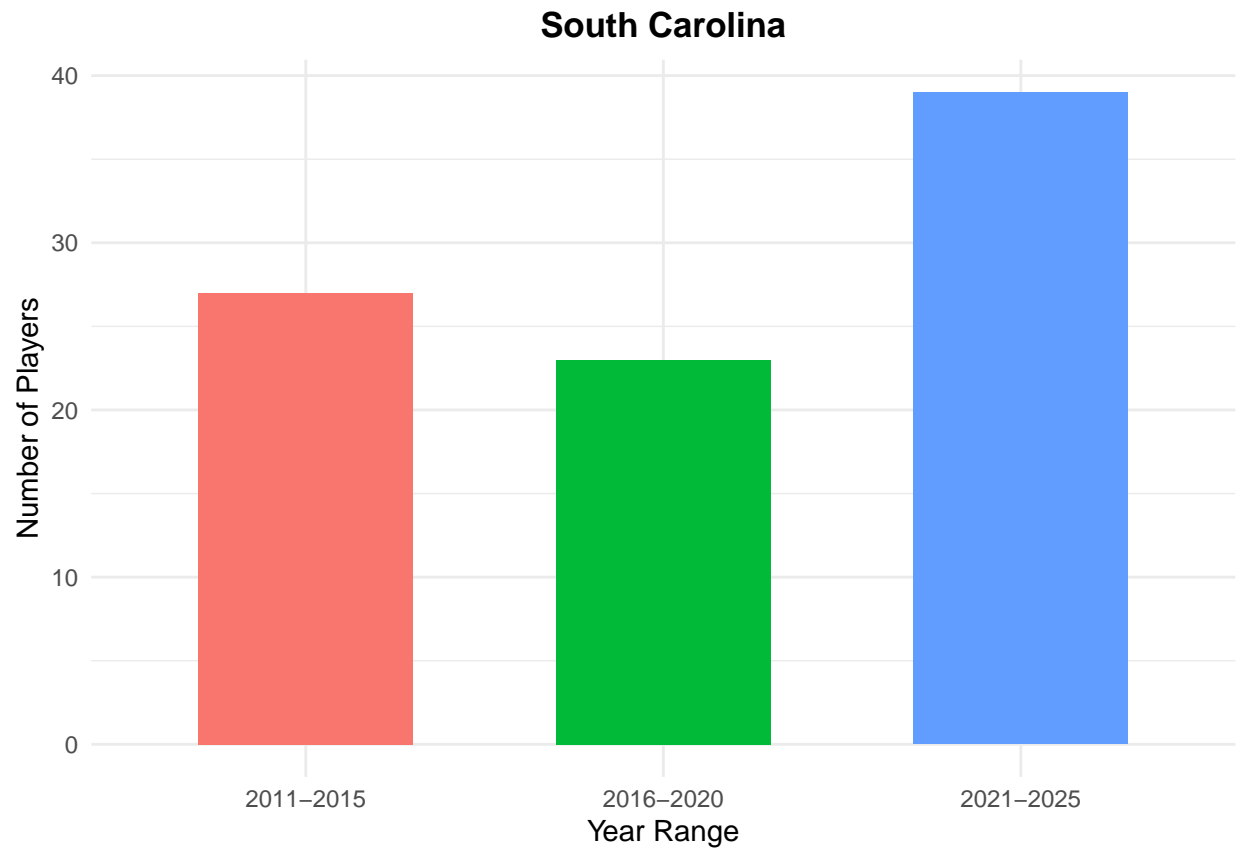


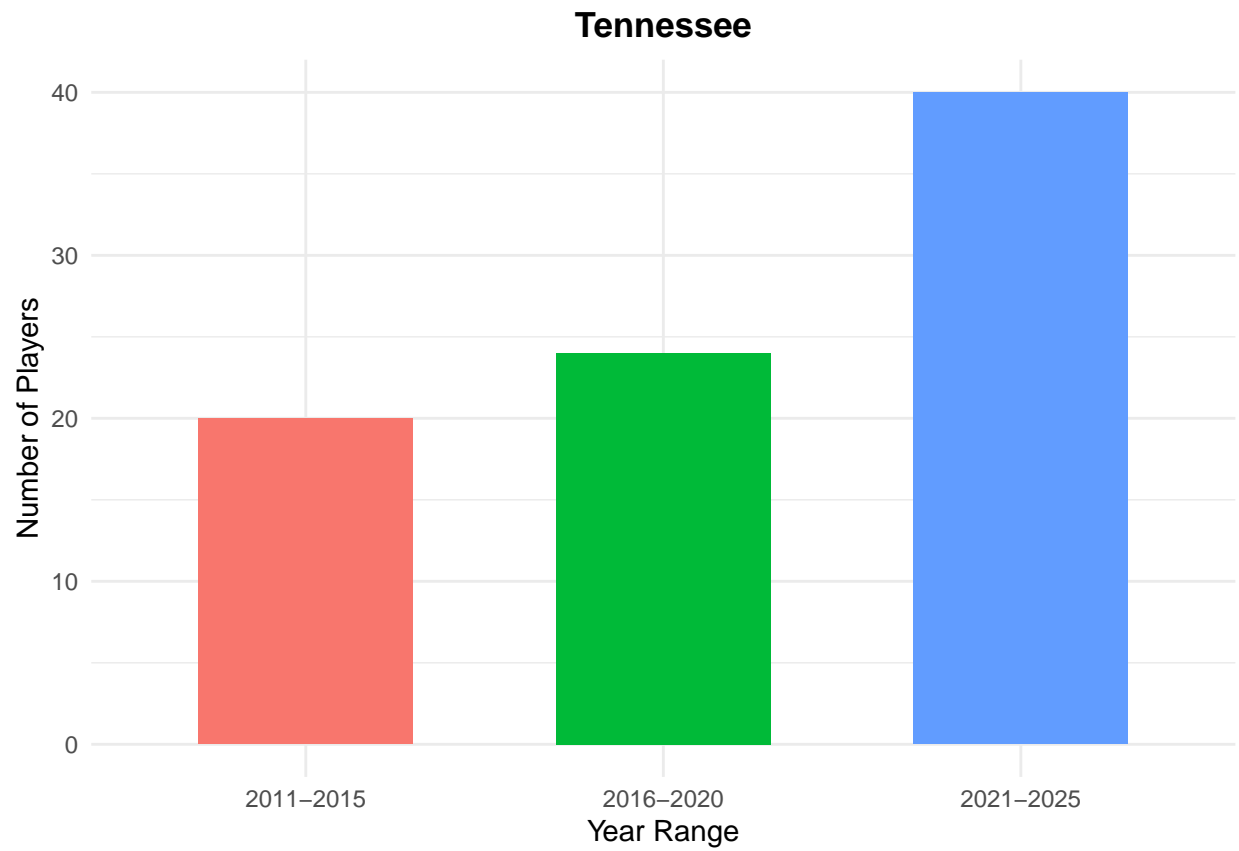


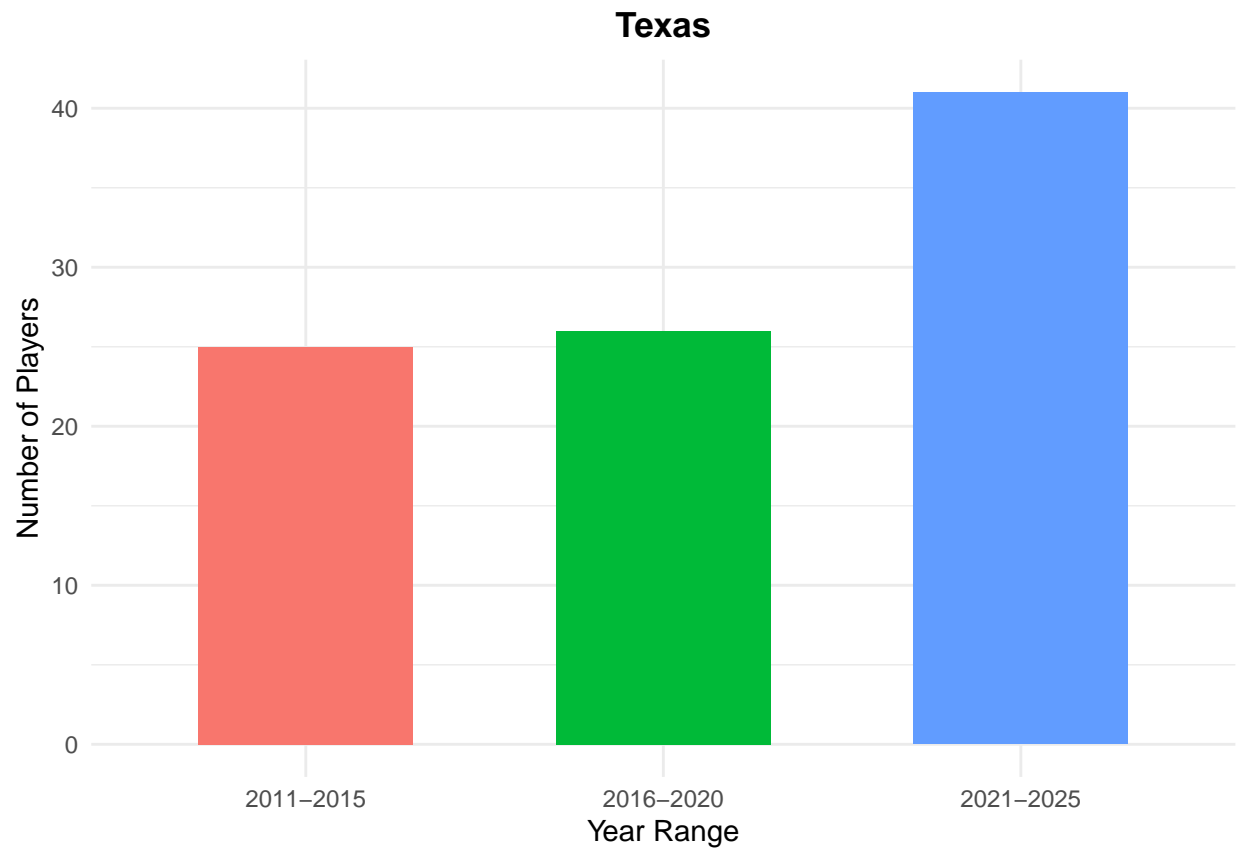


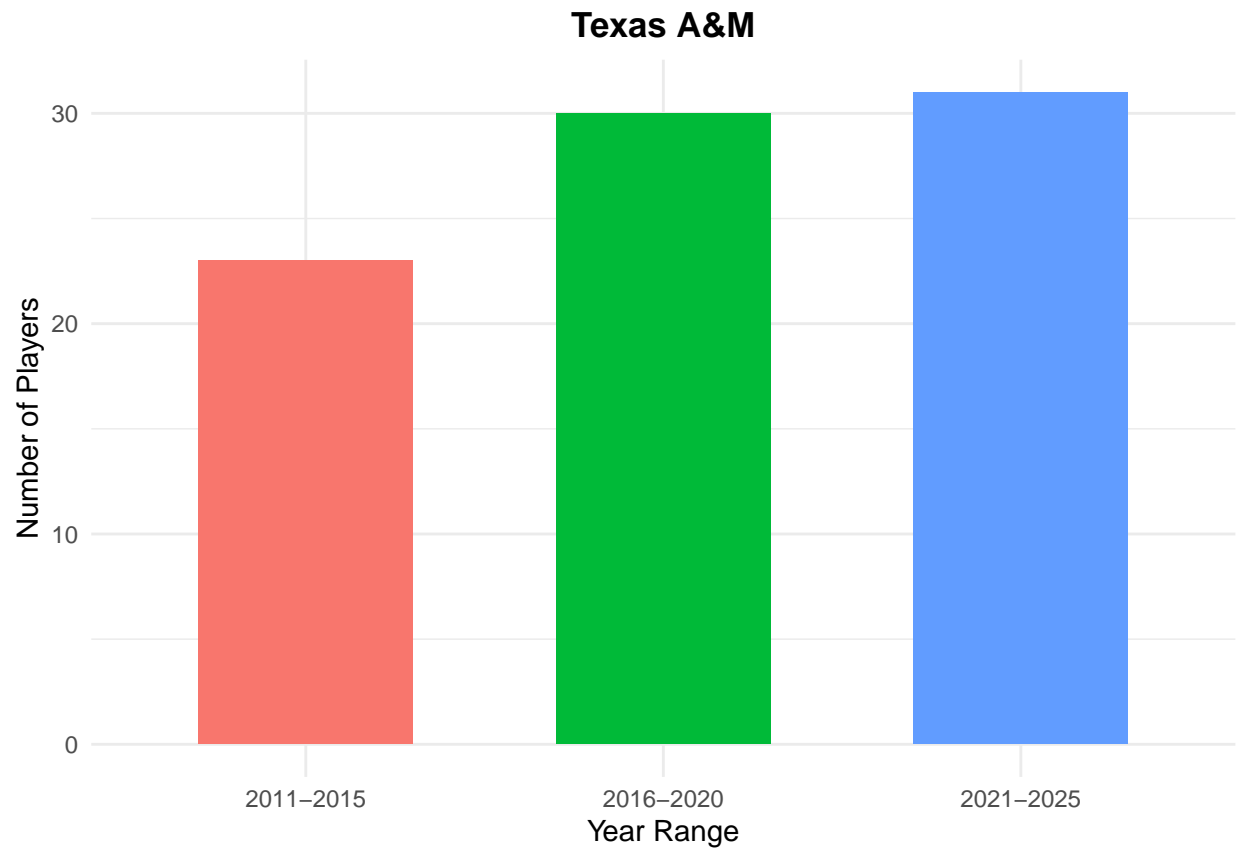


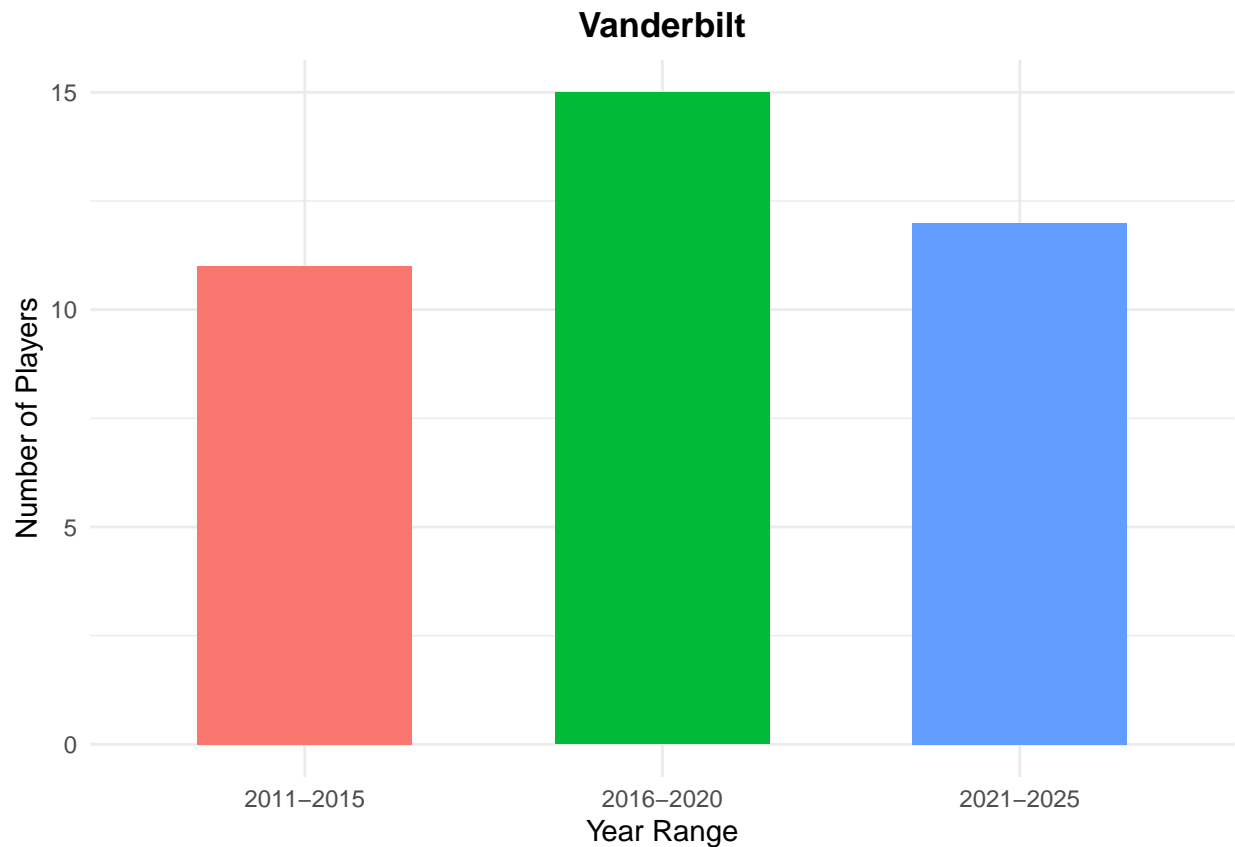












```
# Load necessary library
library(tidyverse)

# Read the CSV file
df <- read.csv("my_updated_dataset.csv")

# Split 'Years Played (From-To)' into Start and End years
df <- df %>%
  separate(`Years Played (From-To)`, into = c("Start", "End"), sep = "-", convert = TRUE)

# Define the year ranges
df <- df %>%
  mutate(YearRange = case_when(
    Start >= 2011 & Start <= 2015 ~ "2011-2015",
    Start >= 2016 & Start <= 2020 ~ "2016-2020",
    Start >= 2021 & Start <= 2025 ~ "2021-2025",
    TRUE ~ NA_character_
  ))

# Summarize player counts per team and year range
team_summary <- df %>%
  filter(!is.na(YearRange)) %>%
  group_by(School, YearRange) %>%
  summarise(Players = n(), .groups = "drop")

# Get list of unique teams
```

```

teams <- unique(team_summary$School)

# --- Plot setup ---
# 2 plots per figure, give more room for axis labels
par(mfrow = c(1, 2), mar = c(5, 4, 3, 1))

# Loop through teams in pairs
for (i in seq(1, length(teams), by = 2)) {
  # Plot first team
  team1 <- teams[i]
  team1_data <- team_summary %>% filter(School == team1)

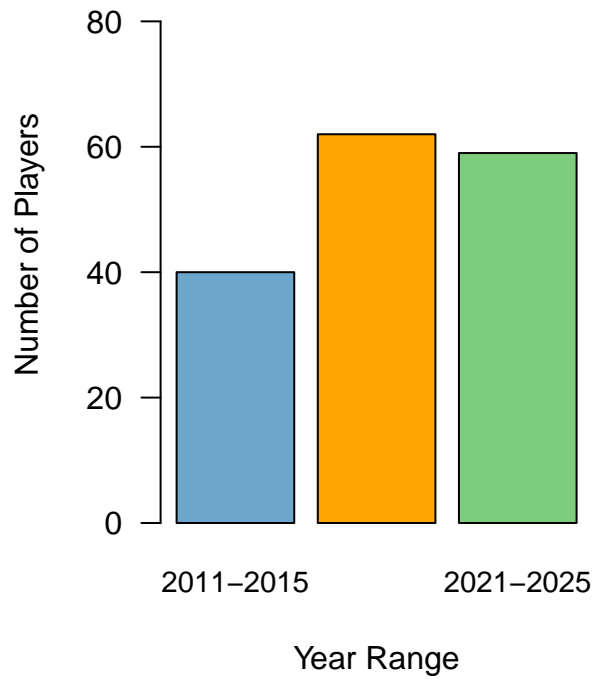
  barplot(
    height = team1_data$Players,
    names.arg = team1_data$YearRange,
    main = team1,
    xlab = "Year Range",
    ylab = "Number of Players",
    col = c("skyblue3", "orange", "palegreen3"),
    ylim = c(0, max(team_summary$Players) * 1.2),
    las = 1,                # horizontal axis labels (always visible)
    cex.names = 0.9        # slightly smaller labels
  )

  # Plot second team (if it exists)
  if (i + 1 <= length(teams)) {
    team2 <- teams[i + 1]
    team2_data <- team_summary %>% filter(School == team2)

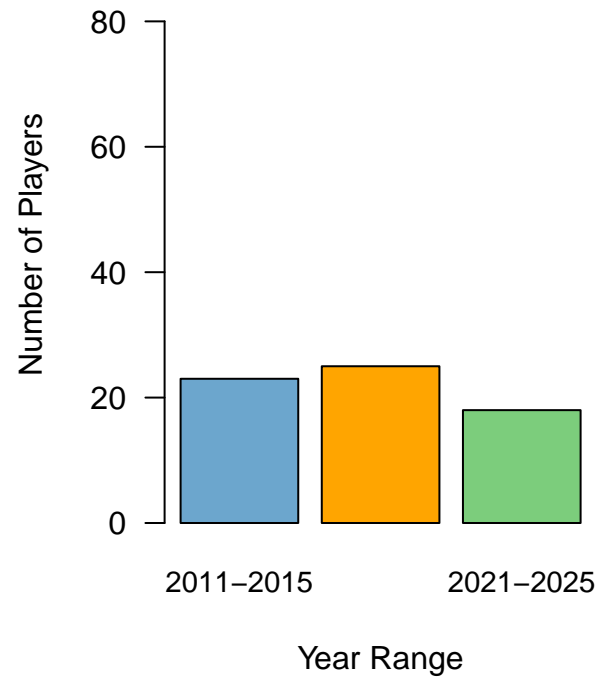
    barplot(
      height = team2_data$Players,
      names.arg = team2_data$YearRange,
      main = team2,
      xlab = "Year Range",
      ylab = "Number of Players",
      col = c("skyblue3", "orange", "palegreen3"),
      ylim = c(0, max(team_summary$Players) * 1.2),
      las = 1,
      cex.names = 0.9
    )
  } else {
    plot.new()
  }
}

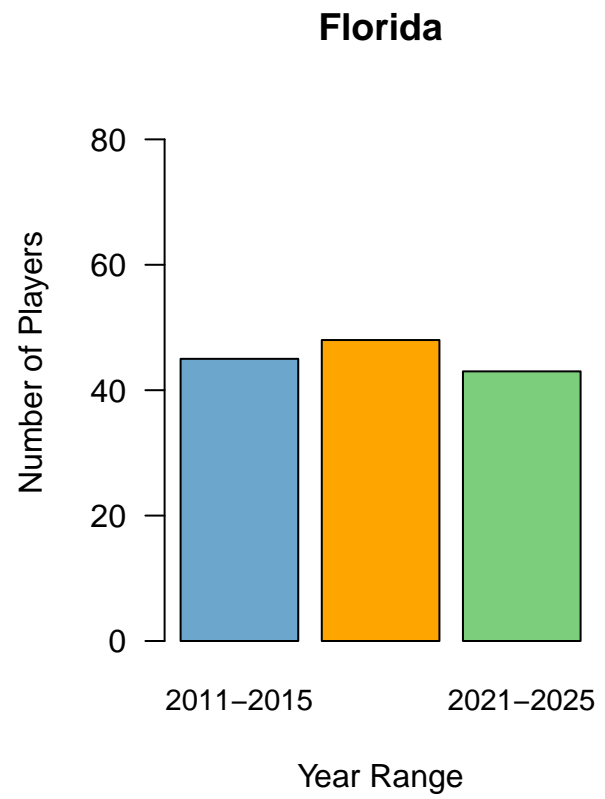
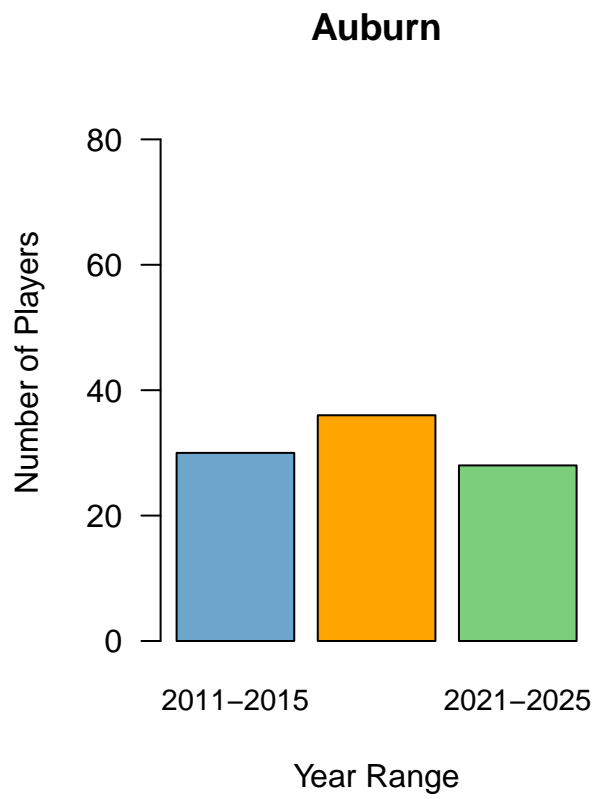
```

Alabama

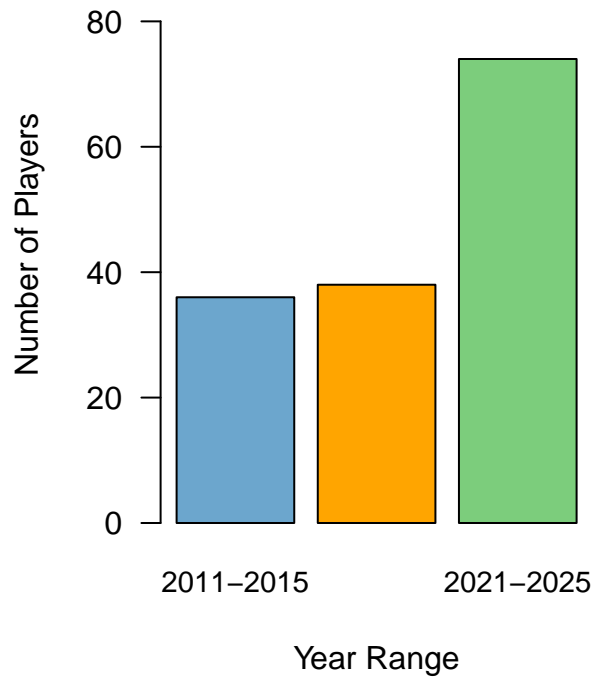


Arkansas

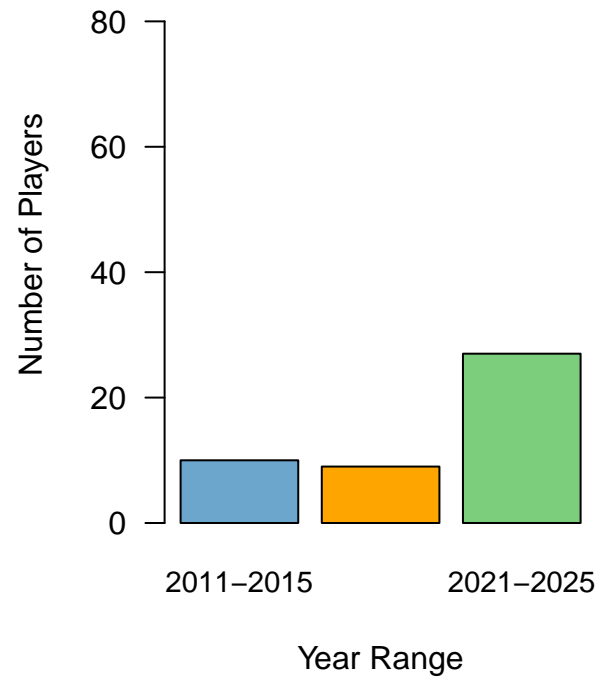




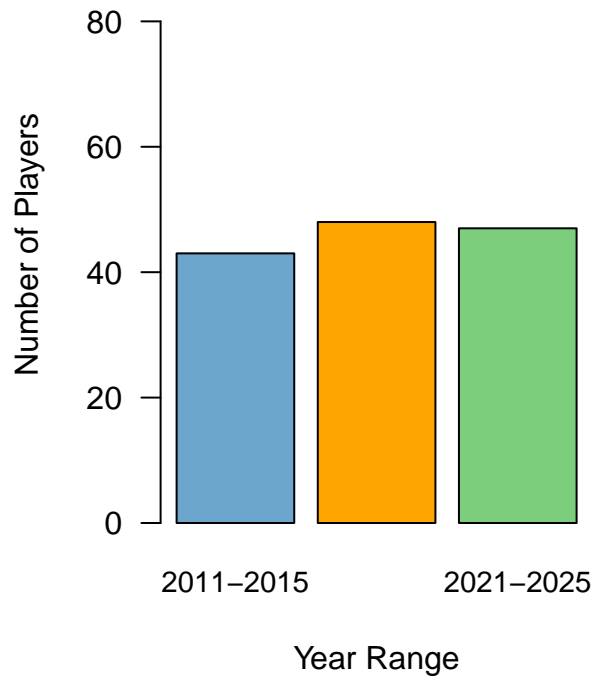
Georgia



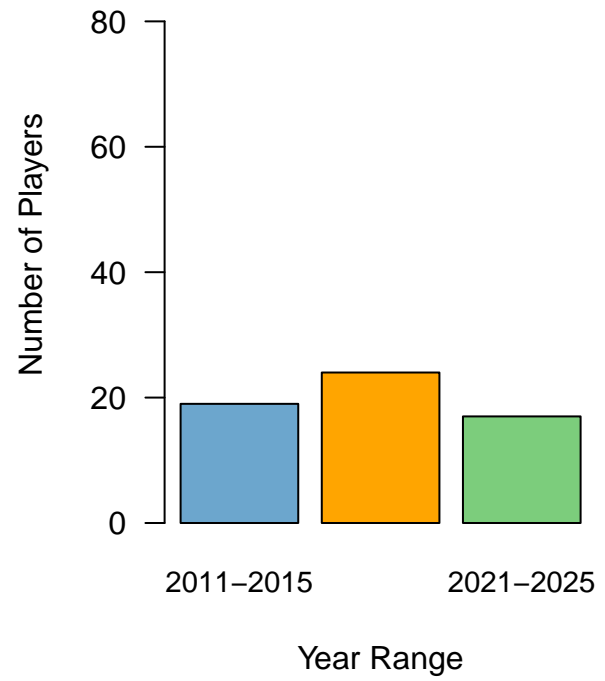
Kentucky



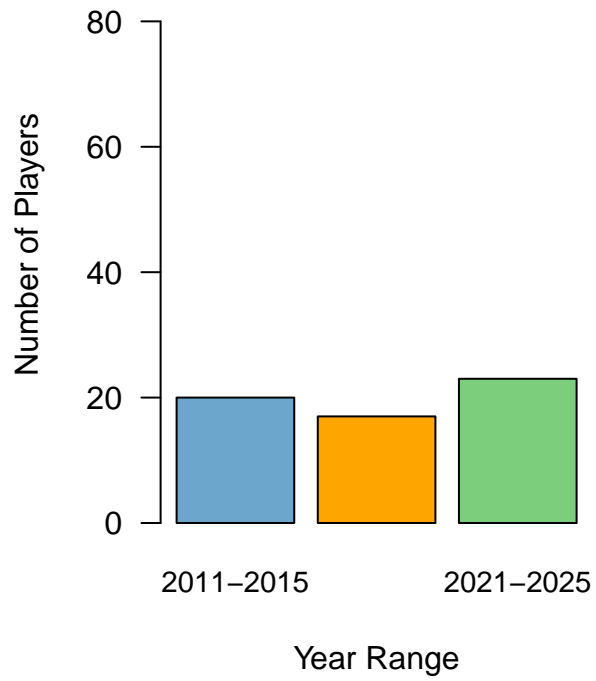
LSU



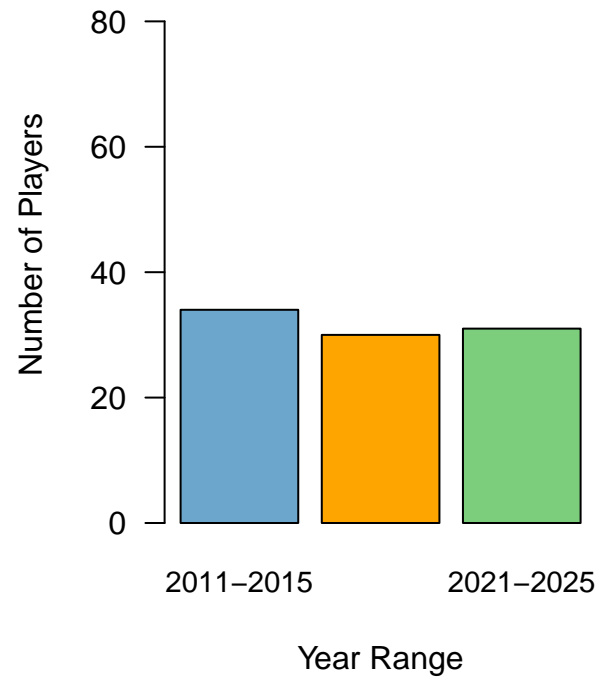
Mississippi St



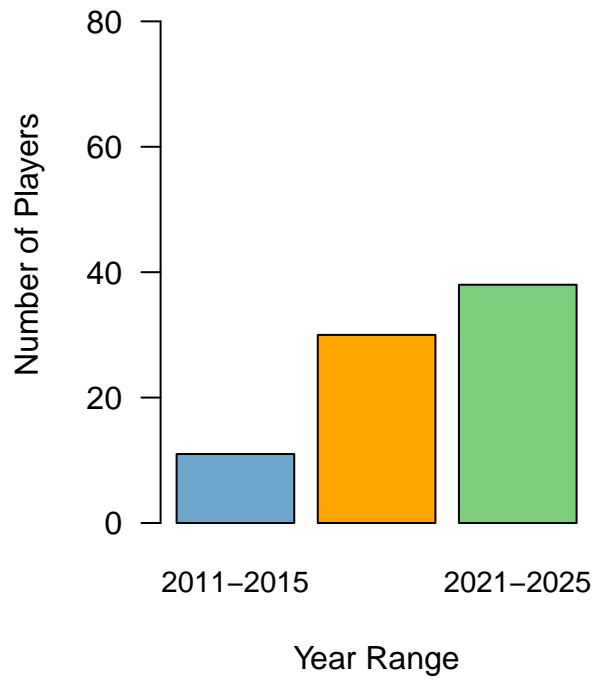
Missouri



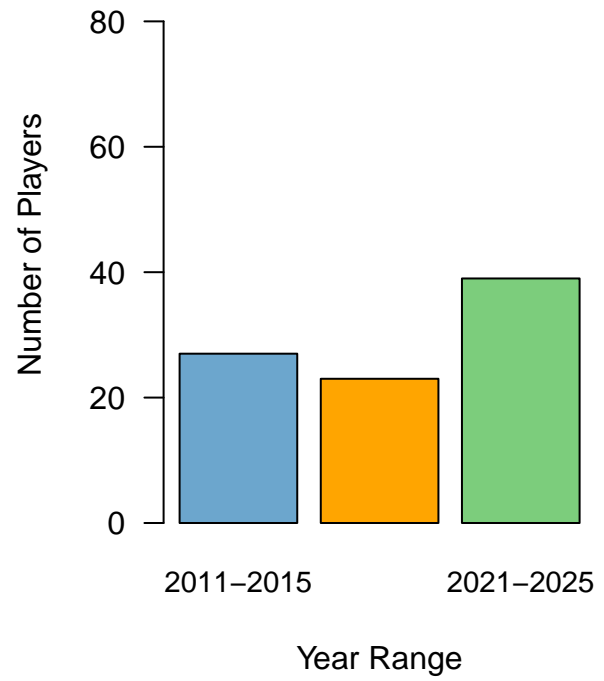
Oklahoma



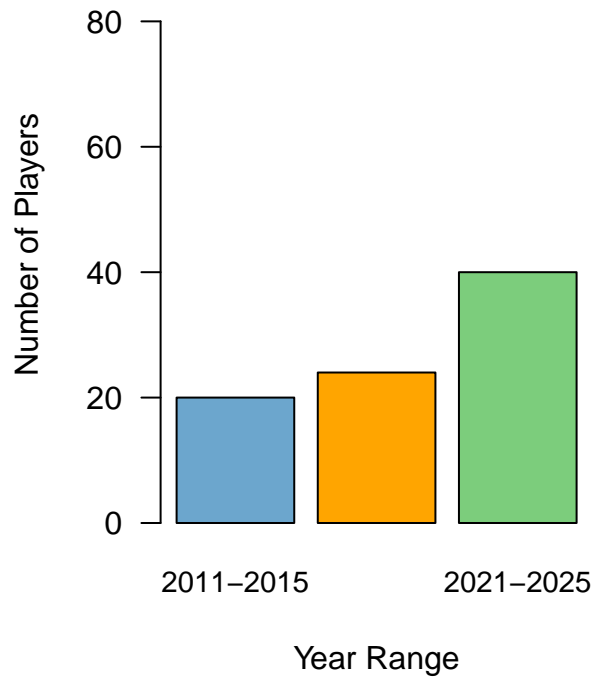
Ole Miss



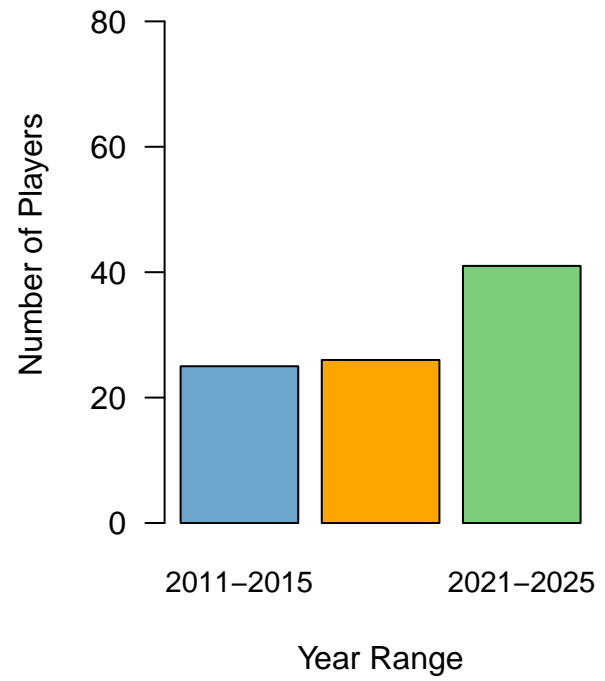
South Carolina



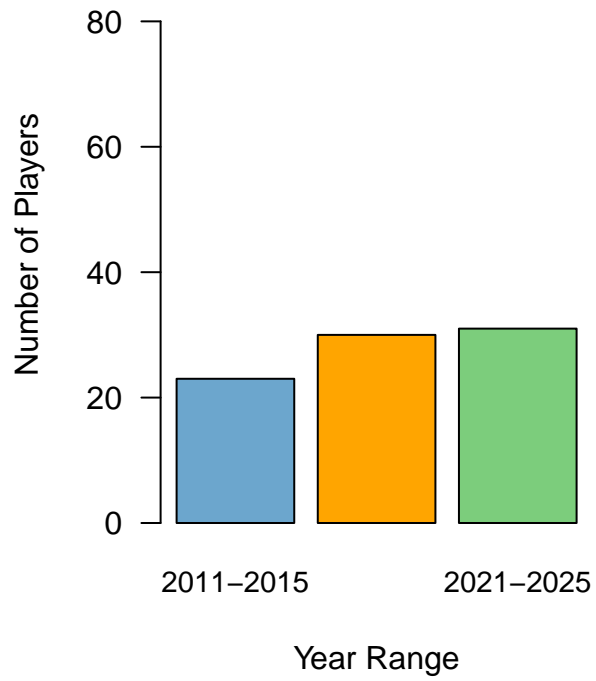
Tennessee



Texas



Texas A&M



Vanderbilt

