NFL Data Exploration

2026-02-01

Background of NFL Data

- ***Prediction of Retirement for a Quarterback***
  - **Event:** Retirement (Last game/season in NFL)
  - **Time Scale:** Number of Seasons/Games since debut
  - **Censoring:** QB still active at end of observation period
- **Source:** NFL Statistics (scraped from official NFL website)
- This data stops at 2016!
- Relevant Files Variables:
  1. Basic_Stats
     1. Name/Player Id (not equal amount)
     2. Age/Birthday (use the one with less NA?), Height, Weight
        1. Take out Age because its calculating from Birthday to Now
        2. ***Maybe Calculate Debut Age?*** older rookies retire sooner
     3. Current Status, Current Team (alot NA), Position (82% NA)
     4. Experience calculated from Years Played, Years Played (18% null)
        1. ***EXPERIENCE IS THE SURVIVAL TIME (the OUTCOME)***
  2. Game_Logs_Quarterback

     Since so many positions are NA, just use Name/Player Id and this file to identify the QBs.

     Filter using Games Played

     1. restrict to just regular season (82%) ???
     2. Year, Season, Week, Game Date

    3. home/away (50/50 split), Outcome
    4. Games Played 0/1, Passed Completed, Passed Attempted, Completion Percentage, Passing Yards
    5. Yards Per Carry (Rushing Yards over Attempts, 62% NA)
    6. Sacks, Sacked Yards Lost, Interceptions, TD Passes (50% NA)
    7. Fumbles (88% NA)

3. Career_Stats_Passing (overlap with variables above) - but this is CAREER STATS for each year - already accumulated compared to game logs
    1. Games Played (Filter if not enough games played that season/year)
    2. Interception Rate

Import Data

```
raw_player <- read_csv(here("nfl", "data", "Basic_Stats.csv"), show_col_types = FALSE)
raw_qb_logs <- read_csv(here("nfl", "data", "Game_Logs_Quarterback.csv"), show_col_types = FA
# very similar data as qb logs, but qb has a bit more
raw_passes <- read_csv(here("nfl", "data", "Career_Stats_Passing.csv"), show_col_types = FALS
#raw_rushing <- read_csv(here("nfl", "data", "Career_Stats_Rushing.csv"))
#raw_fumbles <- read_csv(here("nfl", "data", "Career_Stats_Fumbles.csv"))
```

Read Data

```
# Observe the data
head(raw_player)
```

```
# A tibble: 6 x 16
    Age `Birth Place`      Birthday   College     `Current Status` `Current Team`
  <dbl> <chr>              <chr>      <chr>       <chr>            <chr>
1    NA Grand Rapids , MI 5/23/1921  Notre Dame  Retired          <NA>
2    NA Dayton , OH       12/21/1930 Dayton      Retired          <NA>
3    56 Temple , TX       9/11/1960  Louisiana ~ Retired          <NA>
4    30 New Orleans , LA  9/30/1986  LSU         Retired          <NA>
5    25 Detroit , MI      3/31/1992  Central Mi~ Active           Pittsburgh St~
6    NA Sumner , IL       9/11/1892  Illinois    Retired          <NA>
# i 10 more variables: Experience <chr>, `Height (inches)` <dbl>,
#   `High School` <chr>, `High School Location` <chr>, Name <chr>,
#   Number <dbl>, `Player Id` <chr>, Position <chr>, `Weight (lbs)` <dbl>,
#   `Years Played` <chr>
```

```r
head(raw_qb_logs)
```

```
# A tibble: 6 x 29
  `Player Id`       Name  Position  Year Season  Week `Game Date` `Home or Away`
  <chr>             <chr> <chr>    <dbl> <chr>  <dbl> <chr>       <chr>
1 jaredzabransky/2~ Zabr~ <NA>      2007 Prese~     1 08/11       Home
2 jaredzabransky/2~ Zabr~ <NA>      2007 Prese~     2 08/18       Away
3 jaredzabransky/2~ Zabr~ <NA>      2007 Prese~     3 08/25       Home
4 jaredzabransky/2~ Zabr~ <NA>      2007 Prese~     4 08/30       Away
5 billdemory/25127~ Demo~ <NA>      1974 Regul~     1 09/15       Away
6 billdemory/25127~ Demo~ <NA>      1974 Regul~     2 09/22       Away
# i 21 more variables: Opponent <chr>, Outcome <chr>, Score <chr>,
#   `Games Played` <dbl>, `Games Started` <chr>, `Passes Completed` <chr>,
#   `Passes Attempted` <chr>, `Completion Percentage` <chr>,
#   `Passing Yards` <chr>, `Passing Yards Per Attempt` <chr>,
#   `TD Passes` <chr>, Ints <chr>, Sacks <chr>, `Sacked Yards Lost` <chr>,
#   `Passer Rating` <dbl>, `Rushing Attempts` <chr>, `Rushing Yards` <chr>,
#   `Yards Per Carry` <chr>, `Rushing TDs` <chr>, Fumbles <chr>, ...
```

```r
head(raw_passes)
```

```
# A tibble: 6 x 23
  `Player Id`       Name  Position  Year Team  `Games Played` `Passes Attempted`
  <chr>             <chr> <chr>    <dbl> <chr>          <dbl> <chr>
1 tomfarris/2513861 Farr~ <NA>      1948 Chic~              0 --
2 tomfarris/2513861 Farr~ <NA>      1947 Chic~              9 2
3 tomfarris/2513861 Farr~ <NA>      1946 Chic~             11 21
4 billdemory/25127~ Demo~ <NA>      1974 New ~              1 --
5 billdemory/25127~ Demo~ <NA>      1973 New ~              6 39
6 breezyreid/25239~ Reid~ <NA>      1956 Gree~              7 --
# i 16 more variables: `Passes Completed` <chr>, `Completion Percentage` <chr>,
#   `Pass Attempts Per Game` <dbl>, `Passing Yards` <chr>,
#   `Passing Yards Per Attempt` <chr>, `Passing Yards Per Game` <chr>,
#   `TD Passes` <chr>, `Percentage of TDs per Attempts` <chr>, Ints <chr>,
#   `Int Rate` <chr>, `Longest Pass` <chr>,
#   `Passes Longer than 20 Yards` <chr>, `Passes Longer than 40 Yards` <chr>,
#   Sacks <chr>, `Sacked Yards Lost` <chr>, `Passer Rating` <dbl>
```

Key Limitations:

- Retirement is inferred, not observed
  - No games logged after season t
  - No reappearance in future seasons
- Watch for potential misclassification (like temporary exits for injury, etc)
- Fix using censoring –> censor recent seasons (this data goes up to 2016 so maybe we can disregard 2015 up)

Cleaning the Data

Filter Files for Statistics only on Quarterbacks

Basic_Stats File

```
# get distinct QB player IDs
qb_id <- raw_qb_logs |>
  distinct(`Player Id`)

# Clean player stats - ONLY QUARTERBACKS
player_clean <- raw_player |>
  semi_join(qb_id, by = "Player Id") |>
  select(`Player Id`, Name, Age,`Height (inches)`, `Weight (lbs)`,
         Experience, `Years Played`) |>
  arrange(Name)
player_clean
```

```
# A tibble: 517 x 7
   `Player Id`            Name    Age `Height (inches)` `Weight (lbs)` Experience
   <chr>                  <chr> <dbl>             <dbl>          <dbl> <chr>
 1 tonyadams/2508191      Adam~    67                72            198 5 Seasons
 2 samadkins/2508248      Adki~    62                74            214 6 Seasons
 3 troyaikman/2499369     Aikm~    50                76            220 12 Seasons
 4 erikainge/363          Aing~    30                77            221 3 Seasons
 5 brandonallen/2555365   Alle~    24                74            219 2nd season
 6 derekanderson/2506546  Ande~    33                78            235 13th seas~
 7 kenanderson/2508498    Ande~    68                74            212 16 Seasons
 8 davidarcher/2499447    Arch~    55                74            200 8 Seasons
 9 r.j.archer/2508608     Arch~    29                74            220 3 Seasons
```

```
10 rickarrington/2508672 Arri~    70                 74             200 4 Seasons
# i 507 more rows
# i 1 more variable: `Years Played` <chr>
```

```
# we can possibly calculate debut year using birth year - parse the birthday
# (if the football era is impactful (competition)
```

Game_Logs_Quarterback File

```
# Game logs - regular season only
# if any empty slots, replace with 0 for easy calculation of totals
qb_regular <- raw_qb_logs |>
  filter(Season == "Regular Season") |>
  group_by(`Player Id`) |>
  mutate(across(everything(), ~str_replace_all(., "--", "NA"))) |>
  mutate(Year = as.numeric(Year)) |>
  ungroup()
qb_regular
```

```
# A tibble: 34,657 x 29
   `Player Id`       Name  Position  Year Season Week  `Game Date` `Home or Away`
   <chr>             <chr> <chr>    <dbl> <chr>  <chr> <chr>       <chr>
 1 billdemory/2512~ Demo~ <NA>      1974 Regul~ 1     09/15       Away
 2 billdemory/2512~ Demo~ <NA>      1974 Regul~ 2     09/22       Away
 3 billdemory/2512~ Demo~ <NA>      1974 Regul~ 3     09/29       Away
 4 billdemory/2512~ Demo~ <NA>      1974 Regul~ 4     10/07       Away
 5 billdemory/2512~ Demo~ <NA>      1974 Regul~ 5     10/13       Home
 6 billdemory/2512~ Demo~ <NA>      1974 Regul~ 6     10/20       Home
 7 billdemory/2512~ Demo~ <NA>      1974 Regul~ 7     10/27       Home
 8 billdemory/2512~ Demo~ <NA>      1974 Regul~ 8     11/03       Home
 9 billdemory/2512~ Demo~ <NA>      1974 Regul~ 9     11/10       Away
10 billdemory/2512~ Demo~ <NA>      1974 Regul~ 10    11/17       Away
# i 34,647 more rows
# i 21 more variables: Opponent <chr>, Outcome <chr>, Score <chr>,
#   `Games Played` <chr>, `Games Started` <chr>, `Passes Completed` <chr>,
#   `Passes Attempted` <chr>, `Completion Percentage` <chr>,
#   `Passing Yards` <chr>, `Passing Yards Per Attempt` <chr>,
#   `TD Passes` <chr>, Ints <chr>, Sacks <chr>, `Sacked Yards Lost` <chr>,
#   `Passer Rating` <chr>, `Rushing Attempts` <chr>, `Rushing Yards` <chr>, ...
```

```r
#desired variables
# TD-Int Ratio -> Efficiency
# Sacks -> injury risk -> early retirement
# run-pass-ratio (rushing yards/ passing yards) testing if mobile QBs retire earlier
# run_pass_ratio = rushing_yards / passing_yards
variables <- c( "Passes Completed", "Passes Attempted", "Completion Percentage",
                "Passing Yards", "Sacks", "Ints", "TD Passes", "Rushing Yards")

qb_career_summary <- qb_regular |>
  select(-c(Week, `Passer Rating`)) |>
  mutate(across(all_of(variables), as.numeric)) |>
  group_by(`Player Id`, Name) |>

  summarise(
    # Calculate career timeline for each QB
    First_Year = min(Year, na.rm = TRUE),
    Last_Year  = max(Year, na.rm = TRUE),
    Total_Seasons = length(unique(Year)),
    Total_Games = sum(`Games Played` == 1, na.rm = TRUE),

    # Calculate Totals for Career
    across(all_of(variables[variables != "Completion Percentage"]),
           sum, na.rm = TRUE),
    .groups = "drop") |>

  mutate(
    Time  = Last_Year - First_Year + 1, # survival time in seasons

    # introduce censoring?? the last year is 2016
    # so maybe censor players active post-2015 are censored
    # Adjust this threshold based on your data's latest year
    # if retired or not (0/1)
    Event = if_else(Last_Year >= 2015, 0, 1),

    # efficiency
    TD_INT = `TD Passes` / pmax(Ints, 1), # row wise math

    RUN_PASS = `Rushing Yards` / pmax(`Passing Yards`, 1),

    # categorize career length
    Career_Length = cut(Total_Seasons,
                        breaks = c(0, 2, 5, 10, Inf),
                        labels = c("1-2 seasons", "3-5 seasons", "6-10 seasons", "10+ season
```

```
Warning: There were 8 warnings in `mutate()`.
The first warning was:
i In argument: `across(all_of(variables), as.numeric)`.
Caused by warning:
! NAs introduced by coercion
i Run `dplyr::last_dplyr_warnings()` to see the 7 remaining warnings.

Warning: There was 1 warning in `summarise()`.
i In argument: `across(...)`.
i In group 1: `Player Id = "a.j.feeley/2504566"` `Name = "Feeley, A.J."`.
Caused by warning:
! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
Supply arguments directly to `.fns` through an anonymous function instead.

  # Previously
  across(a:b, mean, na.rm = TRUE)

  # Now
  across(a:b, \(x) mean(x, na.rm = TRUE))
```

```
qb_career_summary
```

```
# A tibble: 466 x 18
   `Player Id`          Name      First_Year Last_Year Total_Seasons Total_Games
   <chr>                <chr>          <dbl>     <dbl>         <int>       <int>
 1 a.j.feeley/2504566   Feeley, ~       2001      2011            11          28
 2 aaronrodgers/2506363 Rodgers,~       2005      2016            12         142
 3 ajmccarron/2543497   McCarron~       2014      2016             3           8
 4 alanrisher/2524210   Risher, ~       1985      1987             2          19
 5 alexespinoza/2513700 Espinoza~       1987      1987             1           1
 6 alexsmith/2506340    Smith, A~       2005      2016            12         141
 7 alextanney/2534870   Tanney, ~       2013      2016             3           1
 8 alexvanpelt/2503454  Van Pelt~       1993      2003            11          31
 9 alpastrana/2522827   Pastrana~       1970      1970             1           4
10 andreware/2503535    Ware, An~       1990      1993             4          14
# i 456 more rows
# i 12 more variables: `Passes Completed` <dbl>, `Passes Attempted` <dbl>,
#   `Passing Yards` <dbl>, Sacks <dbl>, Ints <dbl>, `TD Passes` <dbl>,
#   `Rushing Yards` <dbl>, Time <dbl>, Event <dbl>, TD_INT <dbl>,
#   RUN_PASS <dbl>, Career_Length <fct>
# Note: 18 week seasons usually
```

Notes:

- run-pass-ratio
  - Higher = more mobility, usually more hits from defenders, higher injury risk potentially
    - Indication: earlier retirement
- touchdown-interception-ratio
  - Higher = more TD, better decision-making, protects the ball
    - Indication: longer career

Join Basic_Stats and QB_Logs File

```
# keep everything in QB Logs
qb_combined1 <- qb_career_summary |>
  left_join(player_clean, by = "Player Id")
qb_combined1
```

```
# A tibble: 466 x 24
   `Player Id`          Name.x    First_Year Last_Year Total_Seasons Total_Games
   <chr>                <chr>          <dbl>     <dbl>         <int>       <int>
 1 a.j.feeley/2504566   Feeley, ~       2001      2011            11          28
 2 aaronrodgers/2506363 Rodgers,~       2005      2016            12         142
 3 ajmccarron/2543497   McCarron~       2014      2016             3           8
 4 alanrisher/2524210   Risher, ~       1985      1987             2          19
 5 alexespinoza/2513700 Espinoza~       1987      1987             1           1
 6 alexsmith/2506340    Smith, A~       2005      2016            12         141
 7 alextanney/2534870   Tanney, ~       2013      2016             3           1
 8 alexvanpelt/2503454  Van Pelt~       1993      2003            11          31
 9 alpastrana/2522827   Pastrana~       1970      1970             1           4
10 andreware/2503535    Ware, An~       1990      1993             4          14
# i 456 more rows
# i 18 more variables: `Passes Completed` <dbl>, `Passes Attempted` <dbl>,
#   `Passing Yards` <dbl>, Sacks <dbl>, Ints <dbl>, `TD Passes` <dbl>,
#   `Rushing Yards` <dbl>, Time <dbl>, Event <dbl>, TD_INT <dbl>,
#   RUN_PASS <dbl>, Career_Length <fct>, Name.y <chr>, Age <dbl>,
#   `Height (inches)` <dbl>, `Weight (lbs)` <dbl>, Experience <chr>,
#   `Years Played` <chr>
# all say same thing: "total_seasons", "time", "Experience", "career_length", "Years Played"

# reorder
order <- c("Player Id", "Name.x", "Event", "Age", "Height (inches)", "Weight (lbs)",
           "First_Year", "Last_Year", "Time",
           "Total_Games", "Passes Completed", "Passes Attempted", "Passing Yards",
           "Sacks", "Ints", "TD Passes", "Rushing Yards", "TD_INT", "RUN_PASS")

qb_combined2 <- qb_combined1 |>
  select(all_of(order)) |>
  rename(Name = Name.x,
         Retired = Event,
         Experience = Time) |>
  mutate(Completion_Percentage = `Passes Completed`/`Passes Attempted`,
         Era = case_when(First_Year < 1980 ~ "1970-1980",
                         First_Year < 1990 ~ "1980-1990",
                         First_Year < 1999 ~ "1990-1999",
                         First_Year < 2010 ~ "2000-2010",
                         TRUE ~ "2010+"))

qb_combined2
```

```
# A tibble: 466 x 21
   `Player Id`   Name  Retired   Age `Height (inches)` `Weight (lbs)` First_Year
   <chr>         <chr>   <dbl> <dbl>             <dbl>          <dbl>      <dbl>
 1 a.j.feeley/2~ Feel~       1    40                75            216       2001
 2 aaronrodgers~ Rodg~       0    33                74            225       2005
 3 ajmccarron/2~ McCa~       0    26                75            220       2014
 4 alanrisher/2~ Rish~       1    56                74            190       1985
 5 alexespinoza~ Espi~       1    53                73            193       1987
 6 alexsmith/25~ Smit~       0    33                76            217       2005
 7 alextanney/2~ Tann~       0    29                76            220       2013
 8 alexvanpelt/~ Van ~       1    47                73            220       1993
 9 alpastrana/2~ Past~       1    72                73            190       1970
10 andreware/25~ Ware~       1    48                74            205       1990
# i 456 more rows
# i 14 more variables: Last_Year <dbl>, Experience <dbl>, Total_Games <int>,
#   `Passes Completed` <dbl>, `Passes Attempted` <dbl>, `Passing Yards` <dbl>,
#   Sacks <dbl>, Ints <dbl>, `TD Passes` <dbl>, `Rushing Yards` <dbl>,
#   TD_INT <dbl>, RUN_PASS <dbl>, Completion_Percentage <dbl>, Era <chr>
```

Export Clean Data:

```
safe_write_csv <- function(data, path) {
  dir.create(dirname(path), recursive = TRUE, showWarnings = FALSE)
  readr::write_csv(data, path)}

safe_write_csv(qb_combined2, here("nfl", "data","cleaned", "1_cleaned_nfl_data.csv"))
```

Career_Stats_Passing File (might not use, similar to QB logs))

```
# These are YEARLY career stats

#passes <- raw_passes |>
#  semi_join(qb_id, by = "Player Id") |>
#  arrange(Name)
#passes
```

Explore Data

```
summary(qb_combined2)
```

```
  Player Id             Name              Retired             Age       
 Length:466         Length:466         Min.   :0.0000   Min.   :22.00  
 Class :character   Class :character   1st Qu.:1.0000   1st Qu.:34.00  
 Mode  :character   Mode  :character   Median :1.0000   Median :46.00  
                                       Mean   :0.7897   Mean   :46.77  
                                       3rd Qu.:1.0000   3rd Qu.:58.00  
                                       Max.   :1.0000   Max.   :81.00  
                                                        NA's   :12     
 Height (inches)  Weight (lbs)     First_Year     Last_Year      Experience    
 Min.   :70.00   Min.   :178.0   Min.   :1970   Min.   :1970   Min.   : 1.000  
 1st Qu.:74.00   1st Qu.:205.0   1st Qu.:1981   1st Qu.:1987   1st Qu.: 2.000  
 Median :75.00   Median :215.0   Median :1993   Median :1999   Median : 4.000  
 Mean   :74.64   Mean   :214.4   Mean   :1993   Mean   :1998   Mean   : 5.856  
 3rd Qu.:76.00   3rd Qu.:224.0   3rd Qu.:2006   3rd Qu.:2012   3rd Qu.: 9.000  
 Max.   :80.00   Max.   :250.0   Max.   :2016   Max.   :2016   Max.   :22.000  

  Total_Games     Passes Completed Passes Attempted   Passing Yards    
 Min.   :  0.00   Min.   :   0.0   Min.   :    0.00   Min.   :    0.00  
 1st Qu.:  1.00   1st Qu.:   2.0   1st Qu.:    5.25   1st Qu.:   29.25
 Median : 16.00   Median :  86.5   Median :  177.00   Median : 1029.50
 Mean   : 40.62   Mean   : 537.6   Mean   :  920.66   Mean   : 6395.63
 3rd Qu.: 56.00   3rd Qu.: 539.2   3rd Qu.:  931.25   3rd Qu.: 6253.75
 Max.   :302.00   Max.   :6300.0   Max.   :10169.00   Max.   :71940.00

     Sacks             Ints          TD Passes      Rushing Yards
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : -24.0
 1st Qu.:  0.25   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:   0.0
 Median : 15.00   Median :  8.00   Median :  4.00   Median :  39.0
 Mean   : 66.23   Mean   : 31.69   Mean   : 38.41   Mean   : 295.4
 3rd Qu.: 79.00   3rd Qu.: 35.75   3rd Qu.: 36.50   3rd Qu.: 291.8
 Max.   :525.00   Max.   :336.00   Max.   :539.00   Max.   :6109.0

     TD_INT          RUN_PASS        Completion_Percentage     Era
 Min.   :0.0000   Min.   :-5.00000   Min.   :0.0000        Length:466
 1st Qu.:0.0000   1st Qu.: 0.00000   1st Qu.:0.4987        Class :character
 Median :0.5736   Median : 0.02586   Median :0.5480        Mode  :character
 Mean   :0.6791   Mean   : 0.06850   Mean   :0.5374
 3rd Qu.:1.0556   3rd Qu.: 0.07355   3rd Qu.:0.5964
 Max.   :5.7500   Max.   : 6.85714   Max.   :1.0000
                                     NA's   :88
```

```r
# Check dimensions
print(paste("Number of QBs:", nrow(qb_combined2)))
```

```
[1] "Number of QBs: 466"
```

```r
print(paste("Number of variables:", ncol(qb_combined2)))
```

```
[1] "Number of variables: 21"
```

Notes:

- Age - 81??? Won't be a good variable because its counting their age from birthday until 2016
- Year - 1970 to 2016
- RUN_PASS negative because rushing yards is negative (ex. lose yards if sacked)

Missingness

```r
empty_columns <- colSums(qb_combined2 == 0, na.rm = TRUE)
empty_columns
```

```
            Player Id                  Name               Retired
                    0                     0                    98
                  Age       Height (inches)          Weight (lbs)
                    0                     0                     0
           First_Year             Last_Year            Experience
                    0                     0                     0
          Total_Games      Passes Completed      Passes Attempted
                   78                    93                    88
        Passing Yards                 Sacks                  Ints
                   93                   117                   131
            TD Passes         Rushing Yards                TD_INT
                  164                   114                   164
             RUN_PASS Completion_Percentage                   Era
                  114                     5                     0
```

Notes:

- Some players enter the league for a very short time, and don't play a game at all. We can filter out players who have no Total_Games (Experience is usually 1).
- 98 of 466 NFL players are shown to not be retired (censored to cut off 2015-2016 season). Should we change this cutoff?

Issues:

- Career totals are functions of survival time so they leak outcome into the predictors so we use game stats, not career ones
- Filter Experience > 2 as most empty values come from those with 2 or less years in the league.
- Took Age Out
- Longer Career = Larger Total –> Turn Everything into Ratios

```r
# redued around 150 rows - should we keep or remove?
qb_combined3 <- qb_combined2 |>
  filter(Experience > 2) |>
  mutate(
    Yards_per_game = `Passing Yards` / Total_Games,
    Sacks_per_game = Sacks / Total_Games,
    TD_per_game    = `TD Passes` / Total_Games,
    Ints_per_game  = Ints / Total_Games,
    Rush_per_game = `Rushing Yards`/ Total_Games) |>
  select(-c("Age", "Passing Yards", "Sacks", "TD Passes", "Ints", "Rushing Yards",
            "Passes Completed", "Passes Attempted", "Last_Year"))
            #"Total_Games", "Experience"))

qb_combined3
```

```
# A tibble: 315 x 17
   `Player Id`          Name  Retired `Height (inches)` `Weight (lbs)` First_Year
   <chr>               <chr>   <dbl>             <dbl>          <dbl>      <dbl>
 1 a.j.feeley/2504566  Feel~       1                75            216       2001
 2 aaronrodgers/25063~ Rodg~       0                74            225       2005
 3 ajmccarron/2543497  McCa~       0                75            220       2014
 4 alanrisher/2524210  Rish~       1                74            190       1985
 5 alexsmith/2506340   Smit~       0                76            217       2005
 6 alextanney/2534870  Tann~       0                76            220       2013
 7 alexvanpelt/2503454 Van ~       1                73            220       1993
 8 andreware/2503535   Ware~       1                74            205       1990
 9 andrewluck/2533031  Luck~       0                76            240       2012
10 andrewwalter/25064~ Walt~       1                78            230       2005
# i 305 more rows
# i 11 more variables: Experience <dbl>, Total_Games <int>, TD_INT <dbl>,
#   RUN_PASS <dbl>, Completion_Percentage <dbl>, Era <chr>,
#   Yards_per_game <dbl>, Sacks_per_game <dbl>, TD_per_game <dbl>,
#   Ints_per_game <dbl>, Rush_per_game <dbl>
```

```r
safe_write_csv(qb_combined3, here("nfl", "data","cleaned", "2_cleaned_nfl_data.csv"))

empty_columns1 <- colSums(qb_combined3 == 0, na.rm = TRUE)
empty_columns1
```

```
            Player Id                  Name               Retired
                    0                     0                    72
```

```
      Height (inches)          Weight (lbs)            First_Year 
                    0                     0                     0 
           Experience           Total_Games                TD_INT 
                    0                    10                    50 
             RUN_PASS Completion_Percentage                   Era 
                   25                     1                     0 
       Yards_per_game        Sacks_per_game           TD_per_game 
                    7                    18                    40 
        Ints_per_game         Rush_per_game 
                   27                    15 
```

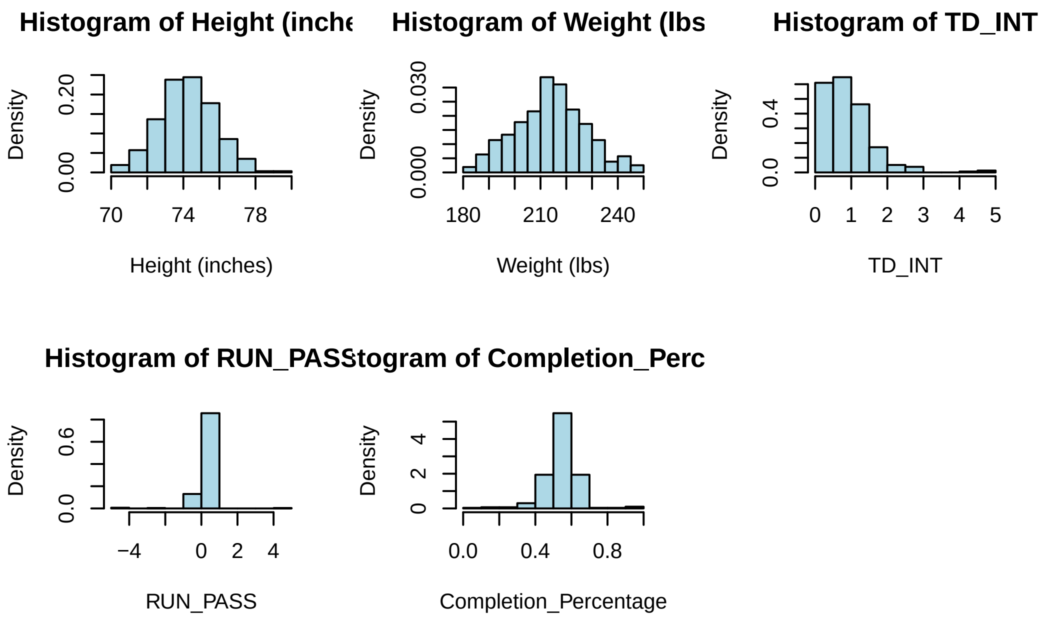
```r
colnames(qb_combined3)
```

```
 [1] "Player Id"             "Name"                  "Retired"              
 [4] "Height (inches)"       "Weight (lbs)"          "First_Year"           
 [7] "Experience"            "Total_Games"           "TD_INT"               
[10] "RUN_PASS"              "Completion_Percentage" "Era"                  
[13] "Yards_per_game"        "Sacks_per_game"        "TD_per_game"          
[16] "Ints_per_game"         "Rush_per_game"        
```

Visuals for Distribution of Variables

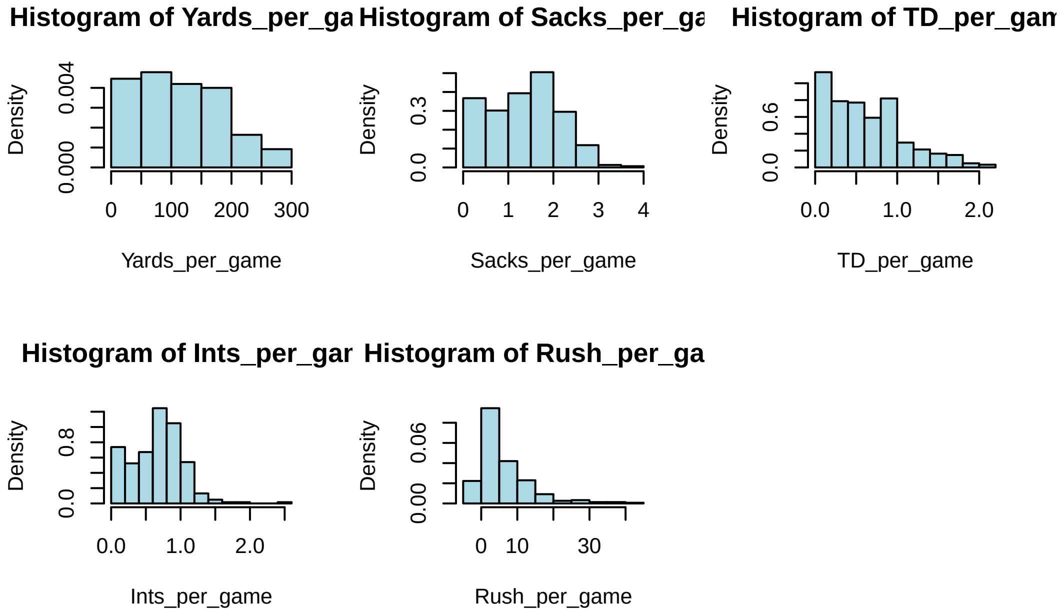
Player Variables

```r
vars <- c("Height (inches)", "Weight (lbs)", "TD_INT", "RUN_PASS", "Completion_Percentage")

# set up a 2 x 3 grid to display all histograms together
par(mfrow = c(2,3))
#par(mfrow = c(3, 3))

for (i in 1:length(vars)) { #game_stats
  current = qb_combined3[[vars[i]]] # get rows of that variable
  hist(current,
       xlab = vars[i],
       main = paste("Histogram of", vars[i]),
       col = "lightblue",
       freq = FALSE)
}

# reset back to default
par(mfrow = c(1, 1))
```



Player Specific Stat Variables

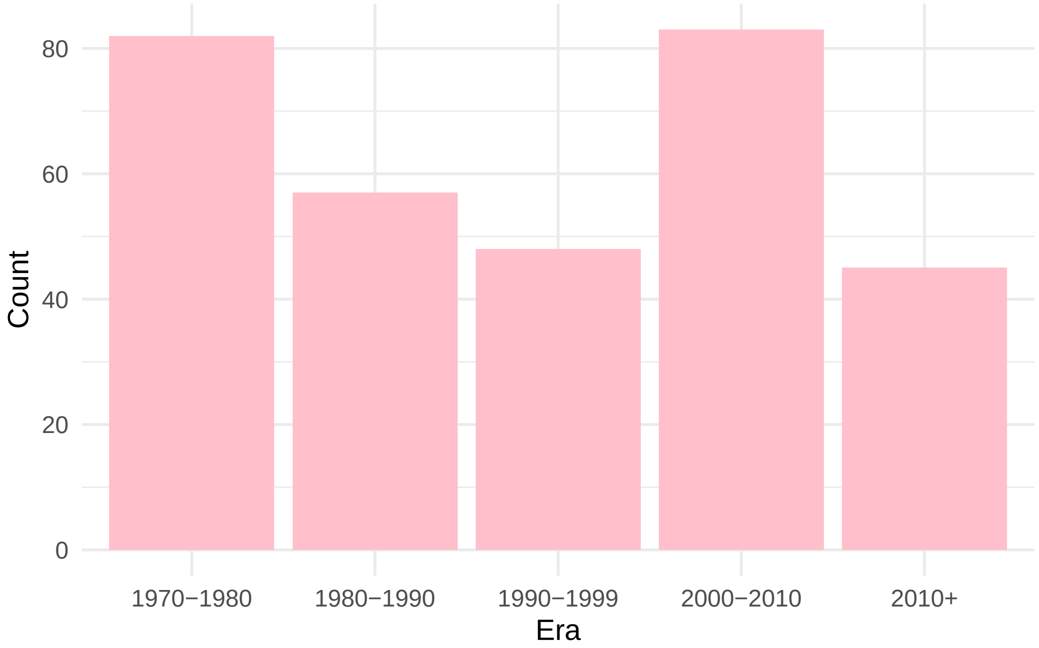
```
other <- c("Yards_per_game", "Sacks_per_game", "TD_per_game", "Ints_per_game", "Rush_per_game

# set up a 3 x 3 grid to display all histograms together
par(mfrow = c(2, 3))

for (i in 1:length(other)) { #game_stats
  current = qb_combined3[[other[i]]] # get rows of that variable
  hist(current,
       xlab = other[i],
       main = paste("Histogram of", other[i]),
       col = "lightblue",
       freq = FALSE)
}

# reset back to default
par(mfrow = c(1, 1))
```



Note: Roughly normally distributed or right-skewed

Era Bar Plot Distribution

```r
ggplot(qb_combined3, aes(x=Era)) +
  geom_bar(fill="pink") +
  labs(x = "Era", y = "Count") +
  theme_minimal()
```



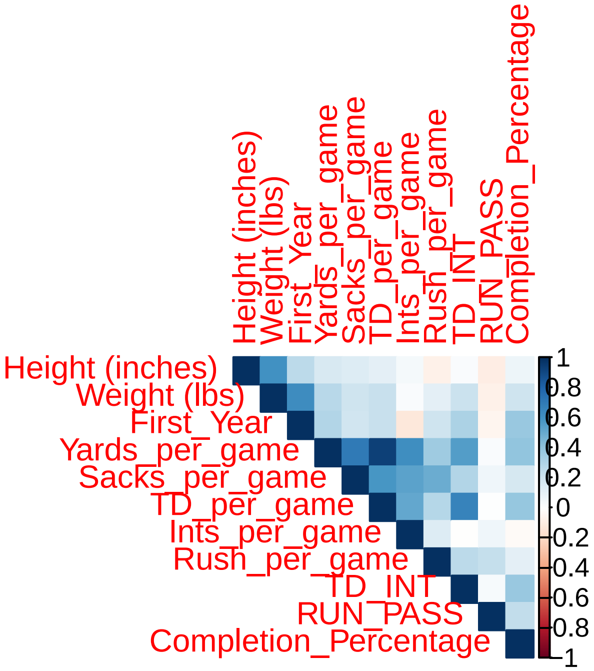
FIXX

Correlation Matrix of Numerical Variables

Taken Out Due to Too High of Correlation –> Multicollinearity (repetition of same info)

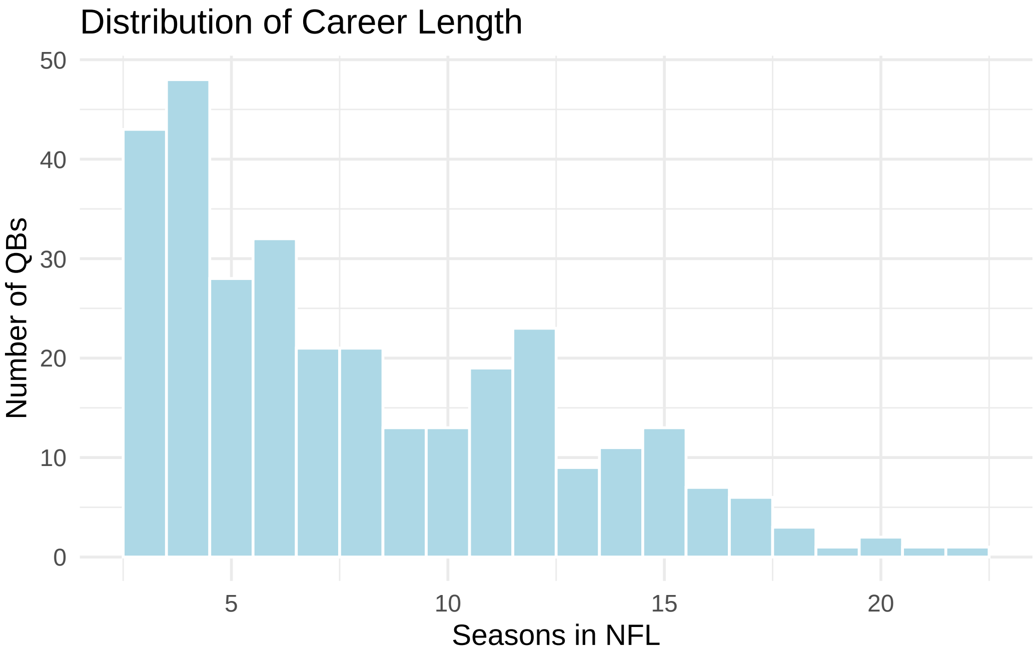
- "Last_Year" –> Already have First_Year and Experience (Years/Seasons)
- "Passes Attempted", "Passing Yards" -> Completion_Percentage
- "Ints", "TD Passes" -> already combined with TD_INT ratio
- Since Total_Games and Sacks are heavily correlated, changed to Sacks_per_game

```
quant_var <- qb_combined3 |>
  select("Height (inches)", "Weight (lbs)", "First_Year",
         "Yards_per_game", "Sacks_per_game", "TD_per_game",
         "Ints_per_game", "Rush_per_game", "TD_INT",
         "RUN_PASS", "Completion_Percentage")
         #"Total_Games", "Experience")
quant_matrix <- cor(quant_var, use = "complete.obs")
corrplot(quant_matrix, method = "color", type = "upper")
```

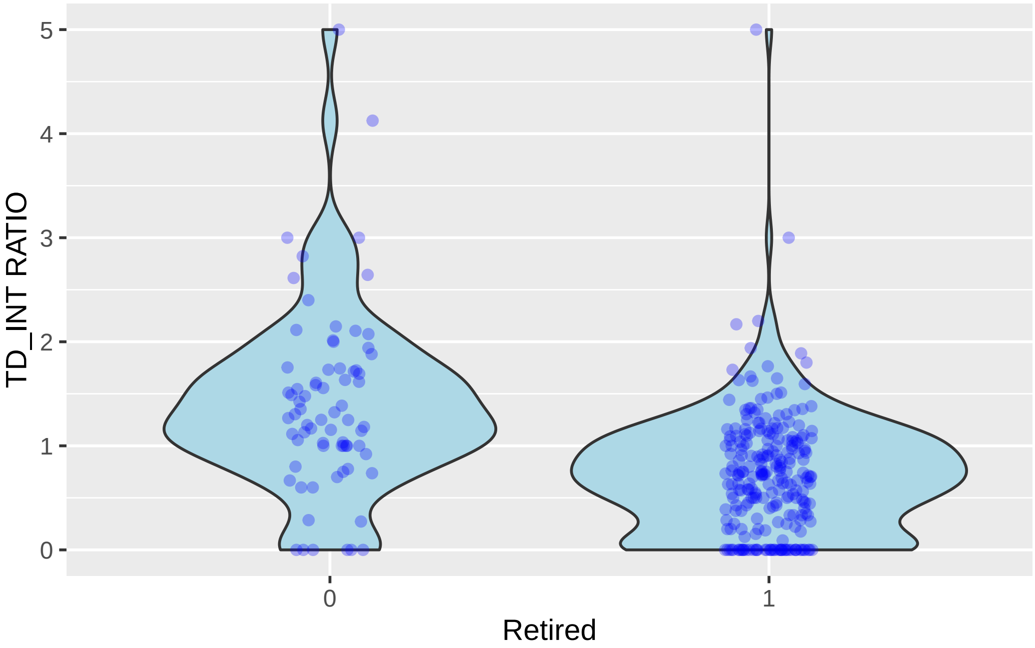


Distribution of Career Length

```
p1 <- ggplot(qb_combined3, aes(x = Experience)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "white") +
  labs(title = "Distribution of Career Length",
       x = "Seasons in NFL",
       y = "Number of QBs") +
  theme_minimal()

print(p1)
```



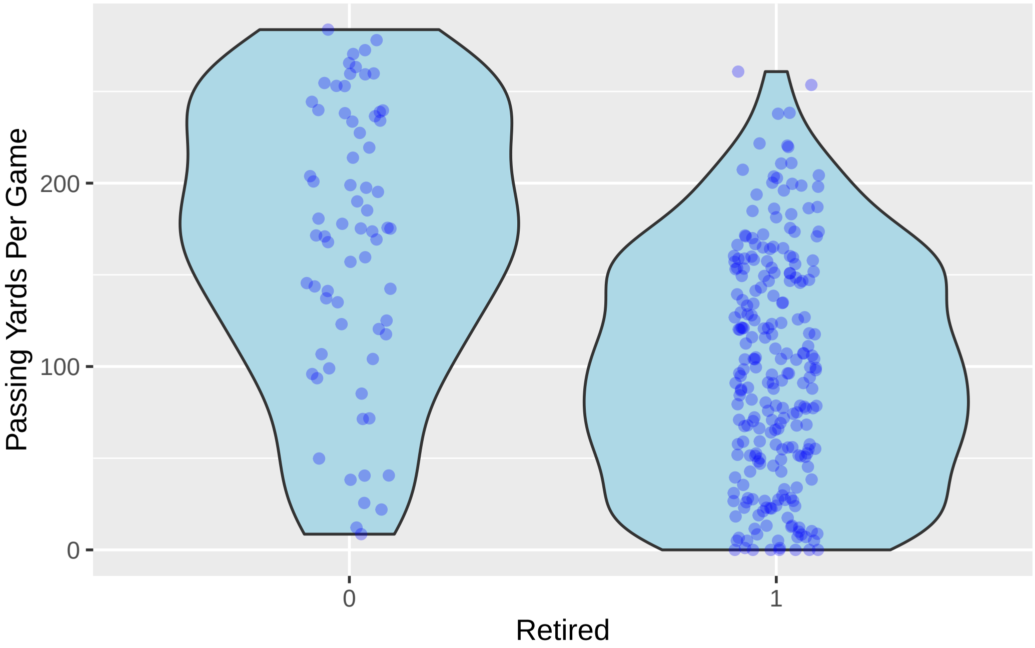
```r
ggplot(qb_combined3, aes(x = factor(Retired), y = TD_INT)) +
  geom_violin(fill = "lightblue") +
  geom_jitter(width = 0.1, alpha = 0.3, color = "blue") +
  labs(x = "Retired", y = "TD_INT RATIO")
```



```
ggplot(qb_combined3, aes(x = factor(Retired), y = Yards_per_game)) +
  geom_violin(fill = "lightblue") +
  geom_jitter(width = 0.1, alpha = 0.3, color = "blue") +
  labs(x = "Retired", y = "Passing Yards Per Game")
```

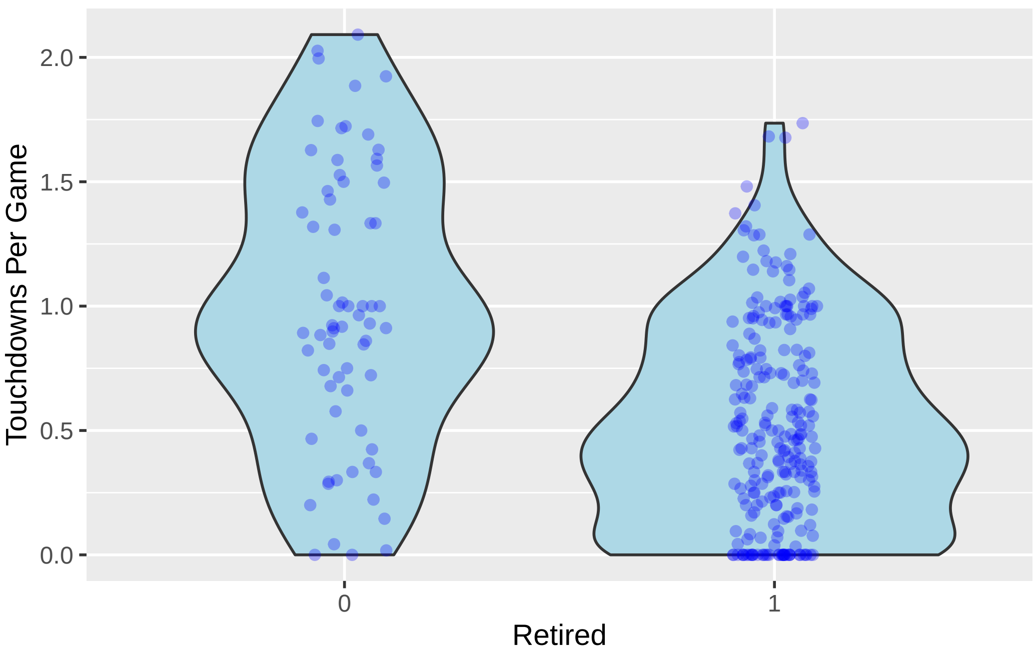
```
Warning: Removed 10 rows containing non-finite outside the scale range
(`stat_ydensity()`).
```

```
Warning: Removed 10 rows containing missing values or values outside the scale range
(`geom_point()`).
```



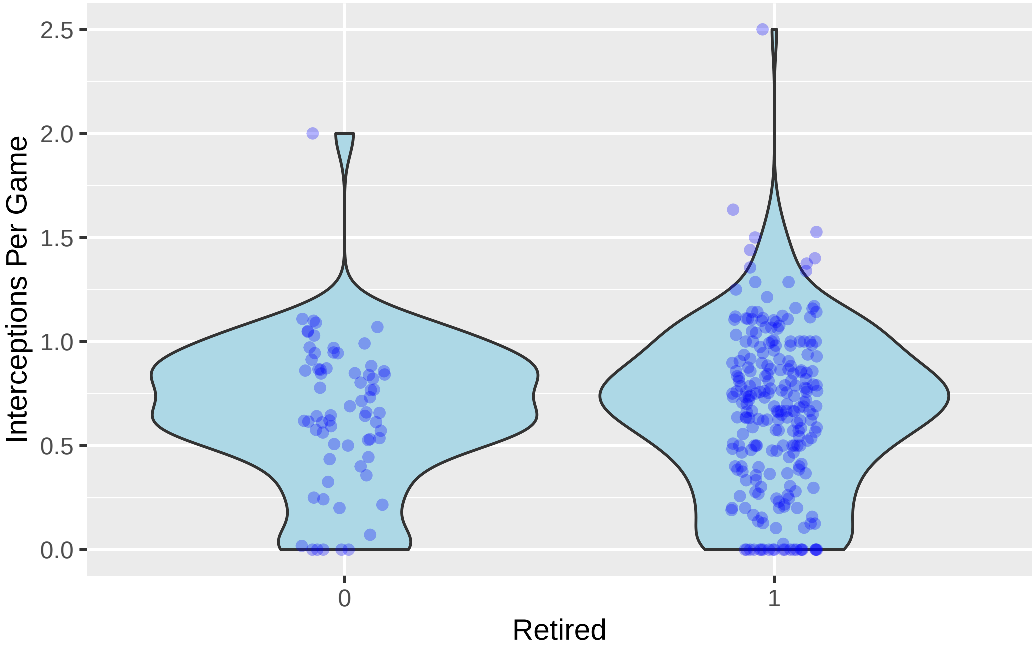
```
ggplot(qb_combined3, aes(x = factor(Retired), y = TD_per_game)) +
  geom_violin(fill = "lightblue") +
  geom_jitter(width = 0.1, alpha = 0.3, color = "blue") +
  labs(x = "Retired", y = "Touchdowns Per Game")
```

```
Warning: Removed 10 rows containing non-finite outside the scale range
(`stat_ydensity()`).
Removed 10 rows containing missing values or values outside the scale range
(`geom_point()`).
```

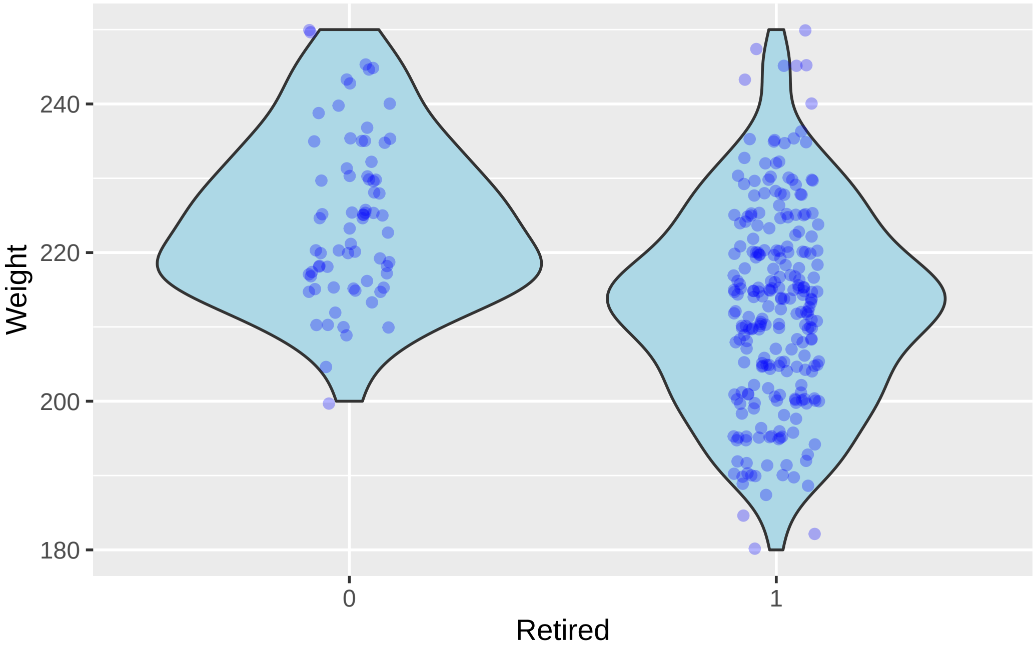


```
ggplot(qb_combined3, aes(x = factor(Retired), y = Ints_per_game)) +
  geom_violin(fill = "lightblue") +
  geom_jitter(width = 0.1, alpha = 0.3, color = "blue") +
  labs(x = "Retired", y = "Interceptions Per Game")
```

```
Warning: Removed 10 rows containing non-finite outside the scale range
(`stat_ydensity()`).
Removed 10 rows containing missing values or values outside the scale range
(`geom_point()`).
```



```
ggplot(qb_combined3, aes(x = factor(Retired), y = `Weight (lbs)`)) +
  geom_violin(fill = "lightblue") +
  geom_jitter(width = 0.1, alpha = 0.3, color = "blue") +
  labs(x = "Retired", y = "Weight")
```


Weight
240
220
200
180
0
1
Retired