11_DataCleaning.Rmd

STAT 107 Team 20

2025-11-08

STEP 1: Load datasets

```
regular <- read_csv("Regular Season.csv")
```

```
## New names:
## Rows: 570 Columns: 15
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): Player, Team dbl (13): ...1, Age, GP, W, L, Min, PTS, FGM, FGA, FG%, 3PM/,
## 3PA, 3P%
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
playoffs <- read_csv("Playoffs.csv")
```

```
## New names:
## Rows: 219 Columns: 15
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): Player, Team dbl (13): ...1, Age, GP, W, L, Min, PTS, FGM, FGA, FG%, 3PM/,
## 3PA, 3P%
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
# Check structure of both datasets
glimpse(regular)
```

```
## Rows: 570
## Columns: 15
## $ ...1   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
## $ Player <chr> "Caris LeVert", "Clint Capela", "Daeqwon Plowden", "Dominick Ba~
## $ Team   <chr> "ATL", "ATL", "ATL", "ATL", "ATL", "ATL", "ATL", "ATL", "ATL", ~
## $ Age    <dbl> 30, 31, 26, 22, 22, 28, 32, 25, 23, 26, 28, 21, 32, 22, 24, 28,~
## $ GP     <dbl> 3, 3, 2, 1, 4, 4, 4, 2, 3, 3, 2, 4, 3, 4, 1, 5, 3, 4, 2, 2, 4, ~
## $ W      <dbl> 0, 1, 0, 1, 1, 1, 0, 2, 1, 1, 0, 1, 1, 1, 0, 4, 1, 1, 1, 1, 1, ~
## $ L      <dbl> 3, 2, 2, 0, 3, 3, 4, 0, 2, 2, 2, 3, 2, 3, 1, 1, 2, 3, 1, 1, 3, ~
```

```
## $ Min    <dbl> 64.0, 47.5, 21.7, 2.9, 91.1, 65.4, 64.8, 14.2, 69.1, 35.9, 34.9~
## $ PTS    <dbl> 26, 24, 3, 2, 43, 29, 30, 2, 50, 11, 22, 60, 21, 15, 9, 34, 37,~
## $ FGM    <dbl> 9, 9, 1, 1, 16, 10, 11, 1, 20, 4, 7, 21, 8, 5, 4, 14, 12, 8, 12~
## $ FGA    <dbl> 27, 16, 9, 1, 38, 32, 26, 5, 29, 10, 18, 56, 15, 16, 7, 20, 27,~
## $ `FG%`  <dbl> 33.3, 56.3, 11.1, 100.0, 42.1, 31.3, 42.3, 20.0, 69.0, 40.0, 38~
## $ `3PM/` <dbl> 3, 0, 0, 0, 8, 9, 4, 0, 7, 2, 0, 9, 2, 1, 1, 3, 7, 4, 4, 0, 6, ~
## $ `3PA`  <dbl> 16, 0, 6, 0, 21, 27, 15, 1, 8, 7, 9, 26, 6, 5, 4, 7, 13, 13, 9,~
## $ `3P%`  <dbl> 18.8, 0.0, 0.0, 0.0, 38.1, 33.3, 26.7, 0.0, 87.5, 28.6, 0.0, 34~
```

```
glimpse(playoffs)
```

```
## Rows: 219
## Columns: 15
## $ ...1   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
## $ Player <chr> "Al Horford", "Baylor Scheierman", "Derrick White", "JD Davison~
## $ Team   <chr> "BOS", "BOS", "BOS", "BOS", "BOS", "BOS", "BOS", "BOS", "BOS", ~
## $ Age    <dbl> 39, 24, 30, 22, 28, 27, 21, 35, 29, 29, 25, 27, 27, 34, 26, 26,~
## $ GP     <dbl> 11, 4, 11, 4, 11, 8, 5, 8, 11, 11, 4, 11, 8, 5, 1, 3, 6, 5, 8, ~
## $ W      <dbl> 6, 3, 6, 3, 6, 4, 4, 4, 6, 6, 3, 6, 5, 4, 0, 2, 4, 3, 5, 5, 5, ~
## $ L      <dbl> 5, 1, 5, 1, 5, 4, 1, 4, 5, 5, 1, 5, 3, 1, 1, 1, 2, 2, 3, 4, 4, ~
## $ Min    <dbl> 347.6, 22.3, 415.3, 13.2, 402.4, 321.6, 15.3, 264.0, 230.8, 180~
## $ PTS    <dbl> 88, 8, 207, 4, 243, 225, 0, 76, 85, 50, 10, 131, 28, 6, 2, 3, 1~
## $ FGM    <dbl> 34, 3, 68, 1, 86, 74, 0, 29, 25, 21, 5, 45, 10, 2, 1, 1, 6, 29,~
## $ FGA    <dbl> 72, 10, 147, 8, 195, 175, 1, 60, 79, 29, 6, 99, 24, 4, 4, 4, 12~
## $ `FG%`  <dbl> 47.2, 30.0, 46.3, 12.5, 44.1, 42.3, 0.0, 48.3, 31.6, 72.4, 83.3~
## $ `3PM/` <dbl> 14, 2, 40, 1, 21, 29, 0, 9, 4, 0, 0, 27, 7, 2, 0, 1, 0, 10, 12,~
## $ `3PA`  <dbl> 35, 5, 104, 2, 63, 78, 1, 26, 26, 0, 0, 67, 21, 3, 2, 4, 3, 35,~
## $ `3P%`  <dbl> 40.0, 40.0, 38.5, 50.0, 33.3, 37.2, 0.0, 34.6, 15.4, 0.0, 0.0, ~
```

Step 2: Keep only relevant columns

```
regular_clean <- regular %>%
  select(Player, Team, GP, Min, PTS, FGM, FGA, `FG%`, `3PM/`, `3PA`, `3P%`)

playoffs_clean <- playoffs %>%
  select(Player, Team, GP, Min, PTS, FGM, FGA, `FG%`, `3PM/`, `3PA`, `3P%`)
```

STEP 3: Rename columns for consistency

```
names(regular_clean) <- c("Player", "Team", "GP", "Minutes", "Points",
                          "FGM", "FGA", "FG_percent", "TPM", "TPA", "TP_percent")

names(playoffs_clean) <- c("Player", "Team", "GP", "Minutes", "Points",
                           "FGM", "FGA", "FG_percent", "TPM", "TPA", "TP_percent")
```

STEP 4: Combine datasets

```
combined <- dplyr::bind_rows(
  dplyr::mutate(regular_clean, type = "Regular"),
  dplyr::mutate(playoffs_clean, type = "Playoffs")
)
```