11_DataCleaning.Rmd

STAT 107 Team 20

2025-11-08

STEP 1: Load datasets

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

```
## New names:
## Rows: 28 Columns: 14
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): Player, Team dbl (12): ...1, 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: 28 Columns: 14
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): Player, Team dbl (12): ...1, 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: 28
## Columns: 14
## $ ...1    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, ~
## $ Player <chr> "Shai Gilgeous-Alexander", "Jalen Williams", "Chet Holmgren", "~
## $ Team   <chr> "OKC", "OKC", "OKC", "OKC", "OKC", "OKC", "OKC", "OKC", "OKC", ~
## $ GP      <dbl> 76, 69, 32, 54, 57, 71, 76, 68, 74, 36, 47, 69, 54, 37, 78, 73,~
## $ W       <dbl> 63, 55, 26, 44, 49, 57, 62, 55, 63, 31, 40, 58, 45, 30, 47, 46,~
## $ L       <dbl> 13, 14, 6, 10, 8, 14, 14, 13, 11, 5, 7, 11, 9, 7, 31, 27, 28, 1~
## $ Min     <dbl> 34.2, 32.4, 27.4, 19.3, 27.9, 29.2, 22.9, 27.6, 21.7, 16.6, 16.~
```

```
## $ PTS    <dbl> 32.7, 21.6, 15.0, 7.1, 11.2, 10.1, 12.0, 8.4, 10.2, 6.5, 5.9, 6~
## $ FGM    <dbl> 11.3, 8.2, 5.2, 2.6, 4.9, 3.6, 4.7, 3.4, 3.5, 2.5, 2.1, 2.5, 0.~
## $ FGA    <dbl> 21.8, 16.9, 10.7, 5.8, 8.4, 8.4, 9.6, 7.2, 7.9, 5.1, 4.7, 5.1, ~
## $ `FG%`  <dbl> 51.9, 48.4, 49.0, 44.6, 58.1, 43.5, 48.8, 47.4, 44.0, 49.5, 43.~
## $ `3PM/` <dbl> 2.1, 1.8, 1.4, 1.1, 0.0, 2.4, 1.7, 1.1, 2.6, 0.6, 1.3, 1.0, 0.3~
## $ `3PA`  <dbl> 5.7, 4.9, 3.6, 3.1, 0.3, 5.8, 4.5, 3.1, 6.3, 1.7, 3.3, 2.5, 1.2~
## $ `3P%`  <dbl> 37.5, 36.5, 37.9, 35.3, 0.0, 41.2, 38.3, 35.6, 41.2, 38.3, 39.9~
```

```
glimpse(playoffs)
```

```
## Rows: 28
## Columns: 14
## $ ...1   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
## $ Player <chr> "Shai Gilgeous-Alexander", "Jalen Williams", "Chet Holmgren", "~
## $ Team   <chr> "OKC", "OKC", "OKC", "OKC", "OKC", "OKC", "OKC", "OKC", "OKC", ~
## $ GP     <dbl> 23, 23, 23, 23, 23, 23, 22, 23, 21, 12, 17, 16, 10, 9, 23, 23, ~
## $ W      <dbl> 16, 16, 16, 16, 16, 16, 15, 16, 14, 8, 11, 12, 7, 6, 15, 15, 15~
## $ L      <dbl> 7, 7, 7, 7, 7, 7, 7, 7, 7, 4, 6, 4, 3, 3, 8, 8, 8, 8, 8, 8, 8, ~
## $ Min    <dbl> 37.0, 34.6, 29.8, 24.4, 22.4, 28.9, 13.8, 22.4, 10.0, 7.0, 8.3,~
## $ PTS    <dbl> 29.9, 21.4, 15.2, 9.2, 8.1, 7.9, 6.0, 5.6, 5.1, 3.4, 2.6, 2.4, ~
## $ FGM    <dbl> 10.1, 7.7, 5.3, 3.1, 3.6, 2.6, 2.1, 2.2, 1.7, 1.3, 0.9, 1.0, 0.~
## $ FGA    <dbl> 21.9, 17.2, 11.6, 7.0, 5.8, 7.1, 5.4, 5.2, 3.5, 2.9, 2.1, 2.5, ~
## $ `FG%`  <dbl> 46.2, 44.9, 46.2, 45.0, 61.9, 36.6, 39.5, 42.9, 49.3, 45.7, 42.~
## $ `3PM/` <dbl> 1.4, 1.5, 1.2, 1.6, 0.0, 2.1, 1.1, 0.9, 1.1, 0.4, 0.5, 0.3, 0.3~
## $ `3PA`  <dbl> 4.9, 5.0, 4.0, 3.9, 0.0, 6.1, 3.1, 2.7, 2.7, 1.1, 1.5, 1.3, 0.4~
## $ `3P%`  <dbl> 28.3, 30.4, 29.7, 41.1, 0.0, 34.3, 36.2, 32.3, 41.1, 38.5, 36.0~
```

Step 2: Keep only relevant columns

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

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

STEP 3: Rename columns for consistency

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

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