Data Wrangling

John Benedict A. Monfero

2025-02-16

Case Study: Major League Baseball

> What is the relationship between payroll and wins among Major League Baseball (MLB) teams? In this homework, we'll find out by wrangling, exploring, and modeling the dataset in `MLPayData_Total.rdata`, which contains the winning records and the payroll data of all 30 MLB teams from 1998 to 2014.
>
> The dataset has the following variables:

• `payroll` : total team payroll (in billions of dollars) over the 17-year period • `avgwin` : the aggregated win percentage over the 17-year period • `Team.name.2014` : the name of the team • `p1998, . . . , p2014` : payroll for each year (in millions of dollars) • `X1998, . . . , X2014` : number of wins for each year • `X1998.pct, . . . , X2014.pct` : win percentage for each year

> We'll need to use the following R packages:

```
library(tidyverse) # tidyverse
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggrepel)    # for scatter plot point labels
library(kableExtra) # for printing tables
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(cowplot)    # for side by side plots
```

```
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

Import

• Import the data into a `tibble` called `mlb_raw` and print it.

```
# Load the dataset
load("ml_pay.rdata")

# Convert the dataset into a tibble
mlb_raw <- as_tibble(ml_pay)

# Print the tibble
print(mlb_raw)
```

```
## # A tibble: 30 × 54
##    payroll avgwin Team.name.2014 p1998 p1999 p2000 p2001 p2002 p2003 p2004 p2005
##      <dbl>  <dbl> <fct>          <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1   1.12   0.490 Arizona Diamo…  31.6  70.5  81.0  81.2 103.   80.6  70.2  63.0
##  2   1.38   0.553 Atlanta Braves  61.7  74.9  84.5  91.9  93.5 106.   88.5  85.1
##  3   1.16   0.454 Baltimore Ori…  71.9  72.2  81.4  72.4  60.5  73.9  51.2  74.6
##  4   1.97   0.549 Boston Red Sox  59.5  71.7  77.9 110.  108.   99.9 125.  121.
##  5   1.46   0.474 Chicago Cubs    49.8  42.1  60.5  64.0  75.7  79.9  91.1  87.2
##  6   1.32   0.511 Chicago White…  35.2  24.5  31.1  62.4  57.1  51.0  65.2  75.2
##  7   1.02   0.486 Cincinnati Re…  20.7  73.3  46.9  45.2  45.1  59.4  43.1  59.7
##  8   0.999  0.496 Cleveland Ind…  59.5  54.4  75.9  92.0  78.9  48.6  34.6  41.8
##  9   1.03   0.463 Colorado Rock…  47.7  55.4  61.1  71.1  56.9  67.2  64.6  47.8
## 10   1.43   0.482 Detroit Tigers  19.2  35.0  58.3  49.8  55.0  49.2  46.4  69.0
## # ℹ 20 more rows
## # ℹ 43 more variables: p2006 <dbl>, p2007 <dbl>, p2008 <dbl>, p2009 <dbl>,
## #   p2010 <dbl>, p2011 <dbl>, p2012 <dbl>, p2013 <dbl>, p2014 <dbl>,
## #   X2014 <int>, X2013 <int>, X2012 <int>, X2011 <int>, X2010 <int>,
## #   X2009 <int>, X2008 <int>, X2007 <int>, X2006 <int>, X2005 <int>,
## #   X2004 <int>, X2003 <int>, X2002 <int>, X2001 <int>, X2000 <int>,
## #   X1999 <int>, X1998 <int>, X2014.pct <dbl>, X2013.pct <dbl>, …
```

• How many rows and columns does the data have? • Does this match up with the data description given above?

```
# Get the dimensions of the dataset
dimensions <- dim(mlb_raw)

# Print the number of rows and columns
cat("Number of rows:", dimensions[1], "\n")
```

```
## Number of rows: 30
```

```
cat("Number of columns:", dimensions[2], "\n")
```

```
## Number of columns: 54
```

Solution: - The dataset contains rows and columns as provided above.

- The description of the case study above, states that the dataset should include *team payroll*, *average win percentage*, *team names*, and *yearly payroll* and *win records from 1998 to 2014*.
- For the next Chapter: we will verify whether these variables match the description by examining the column names in the next steps.

Tidy

> The raw data are in a messy format: Some of the column names are hard to interpret, we have data from different years in the same row, and both year-by-year and aggregate data are present.