Homework 1

Rolando Santos

2023-08-31

```r
library(mdsr)
library(tidyverse)
```

Dataset Reading

```r
cubs <- read.csv("https://raw.githubusercontent.com/gjm112/DSCI401/main/data/cubs_all_time.csv")
```

Question 1: How many total games have the Cubs won and lost between 1876 and 2022?

```r
# Filtering our dataset to only include rows with years between 1876 and 2022
cubs_filtered_years <- filter(select(cubs, Year, G, W, L), Year > 1875 & Year < 2023)
sum(select(cubs_filtered_years, W))
```

```
## [1] 11141
```

After summing the entire wins column, we can see that the total number of cubs wins in the dataset is 11141

Question 2: What year did the Cubs score the most runs? What year did the Cubs score the fewest runs? Do you have any thoughts about the year that the Cubs scored the fewest runs?

```r
# Selecting Data with might find relevant
cubs_sort <- select(cubs, Year, G, W, L, R)
arrange(cubs_sort, desc(R))[1:5, ]
```

```
##   Year   G  W  L    R
## 1 1894 137 57 75 1056
## 2 1930 156 90 64  998
## 3 1929 156 98 54  982
## 4 1886 126 90 34  900
## 5 1889 136 67 65  867
```

The cubs scored their most runs in the year 1894 with a total of 1056 runs.

```r
arrange(cubs_sort, R)[1:5, ]
```

```
##   Year   G  W  L   R
## 1 2020  60 34 26 265
## 2 1877  60 26 33 366
## 3 1981 106 38 65 370
## 4 1878  61 30 30 371
## 5 1879  83 46 33 437
```

The cubs scored their least amount of runs in the year 2020 with a total of 265. A reason for this could be due to the low amount of games played that year, 60, compared to the years they scored a lot of runs, which had more than double the amount of games played.

Question 3: In how many seasons was the Cubs total attendance (i.e. the variable Attendance) over 3 million?

```
cubs_filter_attendance <- filter(select(cubs, Year, W, G, L, Attendance), Attendance > 3e6)
cubs_filter_attendance
```

```
##    Year   W   G  L Attendance
## 1  2019  84 162 78    3094865
## 2  2018  95 163 68    3181089
## 3  2017  92 162 70    3199562
## 4  2016 103 162 58    3232420
## 5  2011  71 162 91    3017966
## 6  2010  75 162 87    3062973
## 7  2009  83 161 78    3168859
## 8  2008  97 161 64    3300200
## 9  2007  85 162 77    3252462
## 10 2006  66 162 96    3123215
## 11 2005  79 162 83    3099992
## 12 2004  89 162 73    3170154
```

```
nrow(cubs_filter_attendance)
```

```
## [1] 12
```

After filtering the data set we end up with a data frame with 12 rows, also confirmed with nrows() returning 12, thus we can conclude that there were 12 years/seasons where the cubs had an attendance of over 3 million.