Homework 1
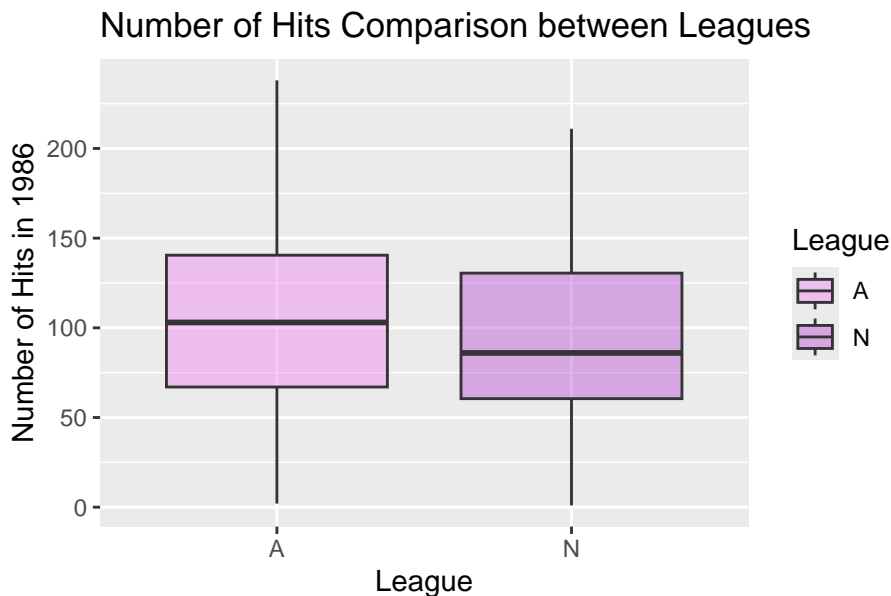
Sarah Butler 24200304

Hitters Dataset

The dataset used in this homework assignment is *Hitters*, a dataset from the *fastliu* package (https://r-packages.io/datasets/Hitters). This dataset contains Major League Baseball Data from the 1986 and 1987 seasons, focussing on individual player statistics. Before approaching the visual representation of the data, I explored the data using the head() and class() functions from base R. The dataset contains 322 obseervations with 20 variables.

Plot Creation

The plot below compares two different leagues in the dataset and the number of hits by players in each league. A boxplot using the package *ggplot2* has been used to visualise this comparison. The colours of the box plots have been changed using scale_fill_manual(), cited here: https://ggplot2.tidyverse.org/reference/scale_manual.html.

Analysis of Graph

The graph above shows the number of hits by players in League A was, on average, larger than the number of hits by players in League N. Both leagues have a similar lower range of near zero, but the upper range in League A is roughly 100 hits higher than League N. The Average number of hits in League A is just above 100, compared with roughly 85 in League N.

Descriptive Statistics of Variables Used

The output below gives a summary of the variable *League*. It shows there are 175 in League A and 147 in League B.

```
  A   N
175 147
```

The summary output below is for the variable *Hits* across both leagues, showing a mean of 101, a median of 96 and a range of 237.

```
 Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
    1      64      96    101     137    238
```