Final Project: Checkpoint 1

STAT 244

Eonbi Choi, Yerim Oh

1. Review the Final Project Description document

2. Describe your data set in 1 - 2 sentences.

The Formula 1 data used in this study are obtained from the f1dataR R package that accesses Formula 1 data via the FastF1 Python library. The dataset includes lap-by-lap session data from the 2024 Miami Grand Prix and comprise 1111 laps and 32 variables. These variables include driver details, lap times, pit in/out times, tire information, and track status.

3. What is the source of your data set? Include a link to where you got it here.

Package 'f1dataR'

- **Description**: https://cran.r-project.org/web/packages/f1dataR/f1dataR.pdf
- **Data sources**: Obtain Formula 1 data via the 'Jolpica API' https://jolpi.caand the unofficial API https://www.formula1.com/en/timing/f1-live via the 'fastf1' 'Python' library https://docs.fastf1.dev/.

4. Load your data set using the code chunk below.

```r
# Install f1dataR package
#install.packages("f1dataR")
#setup_fastf1()     # Run during initial setup
library(f1dataR)
```

```r
#miami2024 <- load_session_laps(season = 2024, session = "R", round = 6)
```

```r
load("lap_dat.Rdata")

# rearrange data
miami2024 <- lap_dat %>%
  select(lap_time, lap_number, compound, tyre_life) %>%
  mutate(compound = as.factor(compound),
         pit_in = ifelse(is.nan(lap_dat$pit_in_time), 0, 1))
head(miami2024)
```

Load packages needed for the data

```
## # A tibble: 6 x 5
##   lap_time lap_number compound tyre_life pit_in
##      <dbl>      <dbl> <fct>        <dbl>  <dbl>
## 1     94.3          1 MEDIUM           1      0
## 2     93.1          2 MEDIUM           2      0
## 3     93.1          3 MEDIUM           3      0
## 4     93.5          4 MEDIUM           4      0
## 5     92.8          5 MEDIUM           5      0
## 6     92.9          6 MEDIUM           6      0
```

- We transformed the variable `pit_in_time` into a binary variable `pit_in`, where `NaN` values in `pit_in_time` were converted to `0` in `pit_in`, indicating no pit stop, and numeric values were converted to `1`, indicating a pit stop occurred.

5. How many rows does your data set have? There should be at least 30 rows.

```r
nrow(miami2024)
```

```
## [1] 1111
```

There are 1111 rows

6. How many quantitative variables are in your data set? There should be at least 3. List them here and briefly describe what they represent and their units.

There are 3 quantitative variables

- `lap_time`: recorded time to complete a lap
- `lap_number`: lap number from which the telemetry data was recorded
- `tyre_life`: number of laps completed on a set of tires

7. How many categorical variables does your data set have? There should be at least 1. List the categorical variable(s) here along with their corresponding categories.

There are two categorical variables

- `compound`: type of tire used
- `pit_in`: whether a driver made a pit stop during a lap (binary: 0 = no pit stop, 1 = pit stop occured)