

Midterm 2 — Sample 2

DATA 412/612

Practice exam, questions only. Cumulative through November 11 (weeks 1–11). The material since Midterm 1 is joins, strings, factors, dates, and statistics (a nested model comparison and numerical summaries).

Put your answers in an R Markdown document and keep the code visible (`knitr::opts_chunk$set(echo = TRUE)` in the setup chunk). You may use your notes, the R help files, and the internet. Do not discuss the exam with anyone else. Load the tidyverse. `mpg` comes with `ggplot2`.

1. **Recode a factor** (8 pts). In `mpg`, collapse `drv` so that "f" and "r" become "two" and "4" stays "four". Drop unused levels. Report how many cars are in each new level.
2. **Periods and zones** (8 pts). `start` is a date-time in `America/New_York`. Build the instant two weeks and three hours later, using a period. Also report the length of the interval from `start` to that later instant in hours. Show `start` again in UTC.

```
start <- ymd_hm("2024-11-03 01:15", tz = "America/New_York")
```

3. **Nested models** (10 pts). Fit two models on `mpg`: `hwy` on `class` alone, and `hwy` on `class` plus `displ`. Compare them with `anova()`. In one sentence, say whether `displ` improves the fit after `class` is already in the model. Use the p-value from that comparison.
4. **Summaries** (8 pts). For each `class` in `mpg`, report the number of cars, the median `hwy`, and the interquartile range of `hwy`. Sort by the median, highest first.
5. **Strings, then a date** (6 pts). The vector below uses a dot between the pieces and a trailing letter that is not part of the date. Remove that letter and parse what remains as a date. Print the dates.

```
c("2024.03.15x", "2023.12.01x", "2022.07.04x")
```