---

title: "Assignment 2"

output: html\_document

date: "2022-11-02"

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

```{r}

library(tidyverse)

library(dplyr)

library(nycflights13)

library(ggplot2)

1. Sort flights to find the fastest flights

```

```{r}

flights %>% mutate(travel\_time = ifelse((arr\_time - dep\_time < 0),

2400+(arr\_time - dep\_time),

arr\_time - dep\_time)) %>%

arrange(travel\_time) %>% select(arr\_time, dep\_time, travel\_time)

arrange(flights, (arr\_time - dep\_time))

```

2. Compare dep\_time, sched\_dep\_time, and dep\_delay. How would you expect those three numbers to be related?

```{r}

flights %>%

mutate (dep\_time = (dep\_time %/% 100) \* 60 + (dep\_time %% 100),

sched\_dep\_time = (sched\_dep\_time %/% 100) \* 60 + (sched\_dep\_time %% 100),

arr\_time = (arr\_time %/% 100) \* 60 + (arr\_time %% 100),

sched\_arr\_time = (sched\_arr\_time %/% 100) \* 60 + (sched\_arr\_time %% 100)) %>%

transmute (near((sched\_dep\_time + dep\_delay) %% (60\*24), dep\_time, tol=1))

```

3. Brainstorm at least 5 different ways to assess the typical delay characteristics of a group of flights. Consider the following scenarios:

a. A flight is 15 minutes early 50% of the time, and 15 minutes late 50% of the time.

b. A flight is always 10 minutes late.

c. A flight is 30 minutes early 50% of the time, and 30 minutes late 50% of the time.

d. 99% of the time a flight is on time. 1% of the time it’s 2 hours late.

Answer:

```{r}

str(flights)

head(flights)

flight\_delay\_summary <- group\_by(flights, flight) %>% summarise(num\_flights = n(),

percentage\_on\_time = sum(arr\_time == sched\_arr\_time)/num\_flights,

percentage\_early = sum(arr\_time < sched\_arr\_time)/num\_flights,

percentage\_15\_mins\_early = sum(sched\_arr\_time - arr\_time == 15)/num\_flights,

percentage\_late = sum(arr\_time > sched\_arr\_time)/num\_flights,

percentage\_15\_mins\_late = sum(arr\_time - sched\_arr\_time == 15)/num\_flights,

percentage\_2\_hours\_late = sum(arr\_time - sched\_arr\_time == 120)/num\_flights)

```

4. Which plane (tailnum) has the worst on-time record?

Answer:

```{r}

flights %>%

group\_by(tailnum) %>%

summarise(prop\_on\_time = sum(arr\_delay <= 30 & !is.na(arr\_delay))/n(),

mean\_arr\_delay = mean(arr\_delay, na.rm=TRUE),

flights = n()) %>%

arrange(prop\_on\_time, desc(mean\_arr\_delay))

flights %>%

group\_by(tailnum) %>%

filter(all(is.na(arr\_delay))) %>%

tally(sort=TRUE)

```

5. Brainstorm as many ways as possible to select dep\_time, dep\_delay, arr\_time, and arr\_delay from flights.

Answer:

```{r}

select(flights, dep\_time, dep\_delay, arr\_time, arr\_delay)

select(flights, c(dep\_time, dep\_delay, arr\_time, arr\_delay))

flights %>% select(dep\_time, dep\_delay, arr\_time, arr\_delay)

flights %>% select\_("dep\_time", "dep\_delay", "arr\_time", "arr\_delay")

flights %>% select\_(.dots=c("dep\_time", "dep\_delay", "arr\_time", "arr\_delay"))

# fancier ways

flights %>% select(dep\_time:arr\_delay, -c(contains("sched")))

flights %>% select(ends\_with("time"), ends\_with("delay")) %>% select(-c(starts\_with("sched"), starts\_with("air")))

flights %>% select(contains("dep"), contains("arr"), -contains("sched"), -carrier)

flights %>% select(matches("^dep|arr\_delay|time$"))

flights %>% select(matches("^dep|^arr"))

flights %>% select(matches("^dep|^arr.\*time$|delay$"))

flights %>% select(matches("^dep|^arr\_time$|delay$"))

```