CPSC 260
Spring 2022

Instructions: This assignment will give you a chance to practice what you have learned about data manipulation using the dplyr package. It will also give you a chance to review previous concepts and pick up some new code that will help you!

For each problem copy and paste your R code from the R script file window into this Word document. Please use a color other than black for your R code. I do not need the code's output, but don't forget to answer any additional questions presented in the problems.

You will upload this Word document to Bb when finished.

Begin by downloading the *nycflights13* package. Then load the package into R along with the dplyr package, which is contained in the tidyverse library. You will be using the *flights* data set from the *nycflights13* package for this assignment. Learn about the data set by looking at the first several rows and by reading the help file. See code below:

```
install.packages("nycflights13")

library(nycflights13)
library(tidyverse)

head(flights)
help(flights)
```

Read Sections 5.1 through 5.6 of Chapter 5: Data Transformation in the R for Data Science book. As you read you will do the exercises specified below. I have provided notes for some problems to help you or to clarify the problem. Please read these carefully.

Section 5.2.4

- #1 parts 1,3,5

```
#Flights that had an arrival delay of two or more hours

filter(flights, arr_delay >= 120)


#Were operated by United, American, or Delta

filter(flights, carrier == 'UA' | carrier == 'AA' | carrier == 'DL')


#Arrived more than two hours late, but didn't leave late

filter(flights, arr_delay >= 120, dep_delay <= 0)
```

- #2 – Use the help file for between() and then apply this function to complete #1 part 7. [The data set uses military time, so midnight=0.]

```
filter(flights, between(dep_time, 0, 6))
```

- #3 –To answer "How many flights have a missing dep_time?" filter out the flights with the specified missing values and then count the number of rows in this subset using

nrow(). When the question asks “What other variables are missing,” they want you to look at the subset you created – Do you see any other columns with NAs?
Don’t forget to answer: “What might these rows represent?”

```
missing_dep_time = filter(flights, is.na(dep_time))
```

```
nrow(missing_dep_time)
```

8255 flights. There is NA values for dep_delay because you don’t know the delay if you don’t know the departure time. Also, NA for arr_delay because you don’t know how long the delay was if you don’t know when the flight took off. Also air time since you don’t know when it took off

Section 5.3.1

- #2 & #4

```
arrange(flights, desc(dep_delay))
arrange(flights, (dep_delay))

arrange(flights, desc(distance))
arrange(flights, (distance))
```

Section 5.4.1

- #2

select(flights, day, day, day)   Nothing changes, it’s like you only use it once

- #4 – Use the help file for contains(). If more than one help file appears, use the one called “Select variables that match a pattern”

select(flights, contains("TIME")). I am maybe a little surprised since it is all capitalized and doesn’t match all lowercase “time”.

Section 5.5.2

- #1 – Do this for just the scheduled departure time.

```
NumOfMins = transmute(flights, sched_dep_time, mins_since_midnight = (sched_dep_time %% 100)
          + ((sched_dep_time %/% 100) * 60))
```

- #2 – As part of your answer create a new variable for arr_time - dep_time, and then display just this column with air_time in order to easily compare them.

The number of minutes in the sky.

```
Question_5_2 = mutate(flights,
          dep_time_mins = (dep_time %/% 100)*60 + dep_time %% 100,
          arr_time_mins = (arr_time %/% 100)*60 + arr_time %% 100,
          subtract_arr_dep = arr_time_mins - dep_time_mins)
nrow(filter(Question_5_2, air_time == subtract_arr_dep))
[1] 196
> nrow(filter(Question_5_2, air_time != subtract_arr_dep))
```

```
[1] 327150
```

So only 196 out of the 327k+ flights met the expected air time.

Section 5.6.7

- #2, #3, & #4

```
not_cancelled %>%
  group_by(dest) %>%
  summarise(n = n())
```

```
not_cancelled %>%
  group_by(tailnum) %>%
  summarize(n = sum(distance))
```

if dep_delay is an NA value then that means the flight didn’t happen because it never took off, this is the most important column for this question because the arrival delay could mean the flight took off still but never made it to its destination

There seems to be a positive correlation between average delay per day and average cancelled flights per day