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)*

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
  + Part 1

flights.1<-flights %>%

filter(arr\_delay>=120)

* + Part 3

flights.2<-flights %>%

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

* + Part 5

flights.3<-flights %>%

filter(dep\_delay<=0,arr\_delay>=120)

* #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.]

help(between)

flights.4<-flights %>%

filter(between(dep\_time,0000,0600))

The between function allows us to select all values between two values. Basically, it allows us to replace x>=0 & x<=0600 for number 7. This simplifies the code we had to write for the problem.

* #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?”

flights.5<-flights %>%

filter(is.na(dep\_time))

nrow(flights.5)

There are 8255 rows with NA dep\_time. Other variables with NA include; dep\_delay, arr\_time, arr\_delay, and air time. I think these rows represent flights that were cancelled.

**Section 5.3.1**

* #2

flights.6<-flights %>% #most delayed – 1301 minutes delayed

arrange(-dep\_delay)

flights.7<-flights %>% #left the earliest – 43 minutes early departure

arrange(dep\_delay)

* #4

flights.8<-flights %>% #farthest distance - 4983

arrange(-distance)

flights.9<-flights %>% #shortest distance - 17

arrange(distance)

**Section 5.3.1 I’m assuming this is meant to be 5.4.1**

* #2

flights.10<-flights %>%

select(dep\_time,dep\_time,dep\_time)

If you select the same variable multiple times it just gives you that variable once in the dataset you create

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

help(contains)

select(flights, contains("TIME"))

no the output does not surprise me after reading the help file. The contains function finds all the variables that have ‘time’ in it. Then the select will exclude all rows except for the ones with time. So it makes sense.

By default, the helper functions of select ignores case by default. However you can change this by setting the argument ignore.case to FALSE.

**Section 5.5.2**

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

flights.11<-flights %>%

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

* #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.

flights.12<-flights %>%

mutate(arr\_dep\_diff=arr\_time-dep\_time) %>%

select(arr\_dep\_diff,air\_time)

I think you are expected to see the two columns have the same result. However, since the times aren’t converted into minutes (example 5:00am to 300 minutes) the arithmetic is flawed. So to fix this, you would have to convert the arr\_time and dep\_time variables to minutes before subtracting dep\_time from arr\_time. The result after these conversions should match air\_time.

**Section 5.6.7**

* #2

not\_cancelled <- flights %>%

filter(!is.na(dep\_delay), !is.na(arr\_delay))

not\_cancelled %>% count(dest)

not\_cancelled %>%

group\_by(dest) %>%

summarize(n=n())

not\_cancelled %>% count(tailnum, wt = distance)

not\_cancelled %>%

group\_by(tailnum) %>%

summarize(n=sum(distance))

* #3

Our definition of canceled flights is suboptimal because we exclude rows without arrival or departure delays. These variables are important, but it would be more important to check the actual departure and arrival times first (arr\_time and dep\_time). These columns would tell us when the plane left and arrived which is more important to know when determining if a flight has been cancelled. Specifically, the departure time because that tells us when a plane leaves.

* #4

flights\_canceled<-flights %>%

filter(is.na(arr\_time),is.na(dep\_time)) %>%

group\_by(year, month, day) %>%

summarize(canceled\_per\_day=mean(n()))

avg\_delay<-flights %>%

filter(!is.na(arr\_time),!is.na(dep\_time)) %>%

group\_by(year, month, day) %>%

summarize(avg\_dep\_delay=mean(dep\_delay))

cancel\_vs\_delay<-left\_join(flights\_canceled,avg\_delay, by=c("year","month","day"))

ggplot(cancel\_vs\_delay)+

geom\_point(mapping=aes(canceled\_per\_day,avg\_dep\_delay))

Looking at the graph I created, it is easy to see that more cancelations tend to occur on days with longer delays. Because of this, I believe there is a pattern and that there is some relation between the two variables (flights canceled per day and avg departure delay).

Chart, scatter chart

Description automatically generated