Data Wrangling with lubridate

Open RStudio and create a new project under your Module 6 folder and call it **Mod6Assignment1.** For this assignment, you will be creating an R Markdown document that will include topics surrounding the ***lubridate*** package as well as packages discussed in prior modules. Once completed, all you need to do is submit the word document that is created.

Create the R Markdown Document

1.) In RStudio, select *File -> New File -> Text File*. This will create a blank text file in the same area that scripts were created in previous assignments (upper left panel). Save this file to your project as **Mod6Assign1Answer.rmd** (it is important to save with the .rmd extension as this saves the text file as an R Markdown file).

2.) Create a Header 1 with the title: **Module 6 - Assignment 1**

3.) Create a Header 2 with the title: **Last Name, First Name** (replace with your name)

4.) Create a Header 3 with the title: **Lubridate**

5.) Click on the dropdown arrow next to the Knit icon 

at the top of the R Markdown Pane in RStudio and select Knit to Word.

6.) Notice that you now have a document in your files for the project named **Mod6Assign1Answer.docx**. This is the file you will be uploading later to Canvas.

7.) For this assignment, you will need to download the **Appointments.csv** file from Canvas. Save this in the same folder that you created the project in.

8.) Create a new chunk of R code that will load both the ***tidyverse*** and ***lubridate***

Part 1: Importing the dataset

9.) Within R Studio, click on the file you just added to the project folder and select "Import Dataset…"

10.) Instead of importing using the "Import Text Data" screen, copy the code in the "Code Preview" and save it in the chunk of code you created to load the libraries needed for this exercise.

Part 2 – Working with Dates

11.) Create a new title labeled: **Working with Dates**

12.) Create a new chunk of R code.

13.) This dataset represents medical appointments for the first 4 months of 2019. However, the file does not include a variable **ReservationYear**. Using the ***dplyr*** package, add a new variable to your Appointments tibble to create a **ReservationYear** that can be used later in the assignment. All you need to do is set the **ReservationYear** to 2019 for all the rows.

14.) Using the lecture and readings for this week, you need to create the following variables (using ***dplyr***) in the Appointments tibble:

a. **ReservationDate** – this variable represents the actual date of their appointment. To create this variable, use the appropriate ***dplyr*** commands to both create the variable and combine the ReservationMonth, ReservationDay and ReservationYear. You can use the make_date ***lubridate*** command to combine these:

   make_date (month=ReservationMonth, day=ReservationDay, year=ReservationYear)

b. **CreationDate** – this variable will represent the time which patient's created the appointment. Using similar code from above, create a new variable that will include the CreationMonth, CreationDay and CreationYear

15.) Create one last variable (**ReservationSpan**) in the tibble that will show how many days in advance the reservation was made (i.e., this is simply a formula to take the difference between the ReservationDate and CreationDate).

Part 3 – Exploratory Analysis

16.) Create a new title: **Exploratory Analysis**

17.) Create a new chunk of code. Using the *summary()* command take a look at the data itself. You may notice that because we used ***lubridate*** to create **ReservationSpan** it is recognized as a *difftime*. However, this does not provide numeric data such as Mean, Median, etc. (see image on next page). We need this as a number to plot later in the assignment.

18.) In the same chunk of code, change the **ReservationSpan** variable to be numeric (hint: we did this on a previous exercise using *as.numeric*). See the image below which shows the summary data before and after the *as.numeric* command.

Before *as.numeric***:**	ReservationSpan Length:61214 Class :difftime Mode :numeric	**After** *as.numeric*:	ReservationSpan Min. : 1st Qu.: Median : Mean : 3rd Qu.: Max. :1

19.) Run the *summary()* command again and answer the following questions below the chunk of code:

a. What is the average time between appointment creation and the actual reservation date?
b. What was the max time?
c. What was the min time?

20.) We are also interested to know if there is a correlation between **ReservationSpan** and **Show** (i.e., does the difference in appointment creation to appointment date impact whether they show). Run a correlation using these two variables and answer the following question:

a. Are ReservationSpan and Show highly correlated?

21.) Create a bar chart using the Show variable in Appointments. Answer the following question:

a. Based on 0 being “no show” and 1 being “show”, did most people make their appointments?

22.) Finally, knit your R Markdown to Word and upload the document to Canvas