Fall 2020 Final Exam

QBS 181: Data Wrangling

Instructions

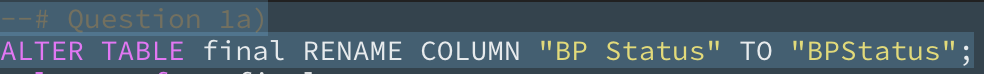
* For this final exam, you MUST work individually, and NO collaboration is allowed. The Dartmouth Honor Principles apply ([https://www.dartmouth.edu/student-handbook/).](https://www.dartmouth.edu/student-handbook)
* For each question, print random 10 rows to show these changes. Please also include the code with comments or descriptions explaining the steps
* For Canvas submission, Please name files which makes it easily understandable. For example: *last\_first\_qbs181\_final.pdf*
* In addition to submitting the final on Canvas, you are required to deposit your code for this exam and previous assignments on GitHub (see question 4 for detail). Please also include a README.md file to help readers understand the contents of your GitHub repository
* Include the link to your GitHub repository in your report while submitting the answers.

Questions

1. Consider the following blood pressure dataset (IC\_BP\_v2.csv). Perform the following operations
2. Convert BP alerts to BP status

--# Question 1a)

ALTER TABLE final RENAME COLUMN "BP Status" TO "BPStatus"



1. Define Hypo-1 & Normal as Controlled blood pressure; Hypo-2, HTN1, HTN2 & HTN3 as Uncontrolled blood pressure: Controlled & Uncontrolled blood pressure as 1 or 0 (Dichotomous Outcomes)

Text

Description automatically generated

A screen shot of a computer

Description automatically generated

Question 1a, 1b combined

1. Merge this table with demographics (SQL table) to obtain their enrollment dates

Text

Description automatically generated

Now we can get the enrollment date

1. Create a 12-week interval of averaged scores of each customer
2. Compare the scores from baseline (first week) to follow-up scores (12 weeks)
3. How many customers were brought from uncontrolled regime to controlled regime after 12 weeks of intervention?
4. Merge the tables Demographics, Conditions and TextMessages.

Obtain the final dataset such that we have 1 Row per ID by choosing on the latest date when the text was sent (if sent on multiple days)

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, application

Description automatically generatedGraphical user interface

Description automatically generated

The screen cannot display all the columns at once, so I took two screenshots for the data now has the latest date and with their contactid by row.

I inner joined all three tables together based on the id name.

1. Repeat Question 2 in R.

Hint: You might want to use tidyr/dplyr packages

Graphical user interface, text, application

Description automatically generated

Graphical user interface, table

Description automatically generated

There are duplicated contacted but they are because of the sendername is different, therefore, they are kept in the table (just want to clarify: I could drop the SenderName column and re-filter the data so no duplicated id is kept even if they are sent by different SenderName, in that way, only one id would should be shown.)

Table

Description automatically generated

(here is the data would like if I truly only keep one id per row, regardless of their sendername or other factors that would make the contact id name shown more than once in the row).

1. Set up a public GitHub repository to share your code. If you didn’t attend the Research Computing’s Git/BASH workshop, or no longer familiar with it, please use online resources (e.g. YouTube tutorials) to re-familiarize yourself with Git/GitHub.
   * Create an account on GitHub, if you haven’t done so. For this final, you are required to use the GitHub and make the repository public.
   * Create a new repository called “Data\_Wrangling\_Project\_and\_Tasks”. This repository should be a public, stand-alone repository.
   * Create sub-directories named “project1”, “project2”, “project3”, “midterm\_project”, “final\_project”, etc.
   * Copy your SQL and R code for each task into their respective repository.
     + Please include R and SQL code only.
     + Important: Do NOT include data, intermediate results, or your screenshots
   * Include a single *README.md* file for the entire repository, containing:
     + Description of the purpose of the repository.
     + A statement (1 sentence) that you have been given the permission to make your work public
     + A description of each sub-directory (you do not need to explain every file; just every task
   * Provide the link of your GitHub repository.

Example: <https://github.com/ydavidchen/data_wrangling_projects>