QMB 6358: Software Tools for Business Analytics
Department of Economics
College of Business
University of Central Florida
Fall 2023

Assignment 3

Due Monday, September 25, 2023 at 11:59 PM
in your GitHub repo.

Instructions:

Complete this assignment within the space on your GitHub repo in a folder called `assignment_03`. In this folder, save your answer to Question 1 in a file called `A3Q1_data.R` and, likewise, save your answer to Question 2 in a file called `A3Q2_data.R`. Samples are available in the `assignment_03` folder within the code repository QMB6358F23.

When you are finished, submit your code by pushing your changes to your GitHub repo through GitHub Desktop.

Question 1:

The repository QMB6358_car_auction_data contains 96 `.csv` files, which contain the records of monthly used car auctions over the years 2010 to 2017. The datasets list the sale prices of cars, the mileage of each car, whether the car has been in an accident, and whether the car has sustained major structural damage, which typically happens only as a result of an accident. Some auctions focus on the sale of damaged cars with "salvage" titles, aimed at the market for used car parts.

Clone the data repository to another location on your computer (outside of any other repo!). Use this location as your working directory but do not push the large dataset to a remote repository; save it somewhere else.

Your job is to collect these files to form one large dataset and print some summary statistics. Use the file `A3Q1_data.R` as a starting point. Complete it in stages by following these steps:

a) Write a loop that prints out the names of all the files. Use this to verify that the file names all match the names of the actual files.

b) Extend the loop to read in each file, assign it to a data frame called `A3Q1_data_sample` and print the function `table(A3Q1_data_sample[, 'accident'], A3Q1_data_sample[, 'damage'])` to show that the files are read in correctly. You should observe the sales of 100 cars in each auction, with some that have been in accidents and some with structural damage.

c) Extend the loop one more time by `bind`ing the files into a full dataset called `A3Q1_data`.

d) Verify that the statistics at the bottom of `A3Q1_data.R` indicate that the dataset has 9,600 rows, 6 columns, with cars 5,400 that have not been in accidents, 4,200 that have, and 600 of those with structural damage.

Note: Do not save a large dataset in a folder within your GitHub repository for your assignments. The GitHub repository is designed to store text files with code but not data; there is a limit on the files size, as well. Please save any large files to a location on your computer outside of the repository.

Question 2:

Your team is responsible for building a model of the value of used airplanes in an online marketplace. The folder `assignment_03` contains three `.csv` files relating to the sales and characteristics of airplanes. Your job is to join these files to form one full dataset and print some output from a regression model. Use the file `A3Q2_data.R` as a starting point. Complete it in stages by following these steps:

a) Read in the `airplane_sales.csv` dataset and store it in a data frame called `airplane_sales` in memory in the R workspace. This file contains a history of the sales of used airplanes obtained from an online marketplace.

b) Read in the `airplane_perf.csv` dataset and store it in a data frame called `airplane_perf` in memory in the R workspace. This file contains records of the performance of used airplanes obtained from the manufacturer.

c) Read in the `airplane_specs.csv` dataset and store it in a data frame called `airplane_specs` in memory in the R workspace. This file contains a list of other specifications of the used airplanes obtained from analyzing the photographs of the airplanes posted on the marketplace Website.

d) Use the `merge` command to join the datasets to form the data frame `airplane_full` in memory in the R workspace. The full dataset should contain all of the valiables from all three files, with matching `Sale_ID` variable in each row.

e) Verify that the dataset is formed correctly by running the block of code at the bottom of `A3Q2_data.R` and checking that the commands `lm` and `summary(lm_model_full)` print output without errors. At this point, you need not know what that output means but you will learn about this in your statistics classes next semester.