**Assignment 2: Data Wrangling in R**

BUAD 5032 – Fall 2021

1. **Objective:**

The purpose of this assignment is to familiarize yourself with the tidyverse package. In particular, you will be using the dplyr and tidyr packages (these are included in the tidyverse package) in this assignment. You will also be asked to make use of other packages not covered in class. This will allow you to understand the process of getting familiarized with new packages in R.

1. **What You Will Need**

Access to a computer with R and RStudio.

1. **What You Will Hand In**

Submit a .R file that includes your code (Assignment2FirstnameLastname.R) via Blackboard - Assignment 2.

1. **Due Date**

10/25/2021 at 11:59PM EST.

1. **Note on Collaboration**

This is a *Category A* assignment. Specifically, you may not receive help from anyone on this assignment except the professor. It must be 100% your own work. All questions concerning this assignment must be addressed to your professor. It is an honor code offense to give or receive any assistance on these assignments.

1. **General Instructions**

Complete all questions by using dplyr and tidyr (these are included in tidyverse). Do not load or use any other package than tidyverse and base R. You should demonstrate when possible your knowledge and understanding of the dplyr and tidyr packages.

**Loading your data set, renaming a columns, selecting variables, and downloading packages in R**

1. Load the tidyverse library. (Make sure you have this installed)
2. Place the avocado.csv file and your Assignment2FirstnameLastname.R in the same folder on your computer. Set the working directory to the source file location (do not include the command in your R file). At the beginning of your file use rm(list=ls()) to clear objects and memory. Load the avocado data set by using the read\_csv() command to load the data set as a tibble (note that you will get a warning message telling you that a column has been named …1). There should be no file directories specified. Only the file name should be included in your command. Name the data frame "Avocado". Take some time to familiarize yourself with the structure of the data set.
3. Rename the …1 column "Week".
4. You will notice that the data set is not sorted by date. Also note that the date column is a character string. A character string in MM/DD/YYYY format will not sort correctly. To solve this problem, install the lubridate package from CRAN using RStudio. Load the package and use the appropriate function to parse the date characters to a date data type. Hint: Download the lubridate cheatsheet from the help menu in RStudio to familiarize yourself with parsing date time functions.
5. Select the regions "West", "Southeast", "SouthCentral", "Northeast", "Midsouth", "Plains" and "GreatLakes" and return a data frame called RegionalAvocado (Make sure the regions are spelled correctly). Hint: You can use the %in% operator or | operator.

**Piping, grouping youar data set, using conditionals, and removing NA’s**

Important: Use only one chain of “pipes” to answer questions 6-9

1. Start by grouping RegionalAvocado by region and type. Next use piping to sort ascending by date. Hint: Use the .by\_group parameter in the arrange() function.
2. Pipe from 6) and create a variable that lags the "AveragePrice" by one period for each region. Name this variable “LagAveragePrice”
3. Pipe from 7) and create a variable that records the change in "AveragePrice". Call this variable "Pchange"
4. Pipe from 8) and create a variable called "Direction" that is "Up" if "Pchange" is strictly greater than zero and "Down" otherwise. Return a tibble called "AvocadoPchange" Hint: Use the if\_else function.
5. Remove the observations (Weeks) that include an NA in the Pchange column of the AvocadoPchange tibble.

**Summarizing grouped data, joining data, and filtering top/bottom observations**

1. Using the AvocadoPchange tibble in 10, create a tibble that contains region and the average “Pchange” (call this variable “MeanPchange”) by avocado type. Return a tibble called AvocadoMean.
2. Using the AvocadoPchange tibble in 10, create a tibble that contains region and the standard deviation of “Pchange” (call this variable “StdevPchange”) by avocado type. Return a tibble called AvocadoStdev.
3. Join the AvocadoMean and AvocadoStdev tibbles so that there is a column for the region, type, MeanPchange, and StdevPchange. Call this tibble SummaryTable Hint: You have to join by region and type.
4. Using the AvocadoPchange tibble in 10, create a tibble that contains each region and the top 3 `Total Volume` values by type. Select only the `Total Volume`, the avocado type and region columns. Call the new tibble AvocadoTop3.