Econometrics 673 Fall 2022 Homework #4

Due by midnight Friday, October 21st

Data Validation

The purpose of this exercise is to perform a data validation of a set of invoice data files and to produce a summary report of average "output."

From the class website, download the "HW5Data.zip" ZIP file. There are four SAS data sets in the compressed zip file, "haul_2012_03.sas7bdat" "haul_2012_04.sas7bdat" "haul_2012_05.sas7bdat" and "haul_2012_06.sas7bdat".

You are told that these data represent monthly invoice records for a trucking company that hauls crude oil. Each observation in the data should represent a "haul" performed by one of the company's drivers. The variables in the data sets relate to the following:

A "Ticket" variable that is a mix of character and numeric values (e.g. SJ00532) and should be unique for each observation. This variable is not particularly important for your analysis.

A "Date" variable in SAS date format (numeric) that represents the transaction date.

A "Driver" variable (mix of character and numeric values e.g. SOFS2389) that is an identification variable for each of the company's drivers.

A "Hauled" variable (numeric) that records the volume of crude oil hauled. You are told that the standard hauling limit of the company's trucks is 190 barrels of crude oil (42 gallons per barrel) though drivers occasionally exceed this limit by as much as 8-10 barrels – an 'overload'. This variable can be recorded as a zero or is missing if the "Driver" was unable to load a particular "Ticket."

Your ultimate task is to produce a summary report that provides various summary statistics of the "hauls" for each "Driver."

- Are the variable names and data formats (character, numeric, character variable lengths) consistent across the four data sets?
- Do the data values of the "Driver" variable appear to be consistent across the four data sets?
- Given the information provided above regarding hauling limits, do the units of measure for the "Hauled" variable appear to be consistent across the four data sets?

Correct any inconsistencies that you find and "stack" the four data sets into a single combined data set.

Create a summary report that shows the following for each unique "Driver"

- The total number of "hauls" (line-items or observations in the data)
- The number of "hauls where the "Driver" was unable to load a particular "Ticket."
- The average number of barrels hauled excluding any observations where the "Hauled" variable is recorded as a zero or missing.
- For "Hauls" over the company haul limit of 190 barrels, the average 'overload.'

Upload a copy of your summary report as an Excel workbook to Canvas using the naming convention "LName Fname UIN HW4".