R Project 3

YOUR NAME

2/26/2020

Note that these exercises should be performed using `dplyr` (do not directly access or manipulate the data frames).

Turn in the html to blackboard with the filename `RProject3_yourlastname.html`. **Points will be taken off for not doing this.**

Part 1

Install and load the "fueleconomy" package and the tidyverse

```
# run these only once
# install.packages("devtools")
# devtools::install_github("hadley/fueleconomy")
```

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(fueleconomy)
```

1. Select the different manufacturers (makes) of the cars in this data set. Save this vector in a variable
2. Use the `distinct()` function to determine how many different car manufacturers are represented by the data set
3. Filter the data set for vehicles manufactured in 1997
4. Arrange the 1997 cars by highway (`hwy`) gas milage
5. Mutate the 1997 cars data frame to add a column `average` that has the average gas milage (between city and highway mpg) for each car
6. Filter the whole vehicles data set for 2-Wheel Drive vehicles that get more than 20 miles/gallon in the city. Save this new data frame in a variable.
7. Of the above vehicles, what is the vehicle ID of the vehicle with the worst hwy mpg? Hint: filter for the worst vehicle, then select its ID.
8. Write a function that takes a `year_choice` and a `make_choice` as parameters, and returns the vehicle model that gets the most hwy miles/gallon of vehicles of that make in that year. You'll need to filter more (and do some selecting)!
9. What was the most efficient Honda model of 1995?

10. Which 2015 Acura model has the best hwy MGH? (Use dplyr, but without method chaining or pipes–use temporary variables!)
11. Which 2015 Acura model has the best hwy MPG? (Use dplyr, nesting functions)
12. Which 2015 Acura model has the best hwy MPG? (Use dplyr and the pipe operator)

Bonus

Write 3 functions, one for each approach in 10, 11 and 12. Then, test how long it takes to perform each one 1000 times

Part 2

Read in the data (from `pulitzer-circulation-data.csv`). Remember to not treat strings as factors!

1. View in the data set. Start to understand what the data set contains
2. Print out the names of the columns for reference
3. Use the 'str()' function to also see what types of values are contained in each column (you're looking at the second column after the `:`) Did any value type surprise you? Why do you think they are that type?
4. Add a column to the data frame called 'Pulitzer.Prize.Change' that contains the difference in the number of times each paper was a winner or finalist (hereafter we'll call this group "winner") during 2004-2014 and during 1990-2003
5. What was the name of the publication that has the most winners between 2004-2014?
6. Which publication with at least 5 winners between 2004-2014 had the biggest decrease(negative) in daily circulation numbers?
7. An important part about being a data scientist is asking questions. Write a question you may be interested in about this data set, and then use dplyr to figure out the answer!