Tidyverse and Tidy Data Programming Concepts

Author: Melina Padron

Part 1:

The dataset `mtcars` was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and other aspects of automobile design and performance for different cars (1973-74 models). Look up the documentation for this data frame with a description of the variables by typing **`?mtcars` in the console** pane.

Question 1:

Take a look at the first 6 rows of the dataset by using an `R` function in the code chunk below. Do you know about any (or all) of these cars?

```
# Used head() to look at the first 6 rows of the dataset 'mtcars'
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Answer: Using the 'head()' function, I was able to see several car names/brands. I know about some of the Mazda cars listed, however, I am not familiar with the other car names or brands.

Question 2:

How many rows and columns are there in this data frame in total?

```
# Loaded in the tidyverse package to utilize the glimpse function
library(tidyverse)

# Used glimpse() to get the total number of rows and columns in the dataset
glimpse(mtcars)
```

```
## Rows: 32
## Columns: 11
## $ mpg  <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8,~
## $ cyl  <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, 4, 4, 8,~
## $ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 140.8, 16~
## $ hp   <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 180, 180~
## $ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92, 3.92,~
## $ wt   <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3.150, 3.~
## $ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 22.90, 18~
## $ vs   <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0,~
## $ am   <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0,~
## $ gear <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3,~
## $ carb <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1, 1, 2,~
```

Answer: After loading the tidyverse package using the 'library()' fucntion I was able to use the 'glimpse()' function. I then used the 'glimpse()' function to count the total number of rows and columns. My conclusion is that there are a total of 32 rows and 11 columns in the 'mtcars' dataset.

Question 3:

Save `mtcars` in your environment and name it as your `eid`. From now on, use this new object instead of the built-in dataset.

```r
# Created an object with my EID as its name and assigned the dataset to it

mdp3327 <- mtcars
```

Answer: I utilized the '<-' operator to create a new object named after my eid, mdp3327, and assigned it to the 'mtcars' dataset. I will now utilize this object instead of the original built in dataset.

Question 4:

When is your birthday? Using indexing, grab the value of `mpg` that corresponds to the day of your birthday (should be a number between 1 and 31).

```r
# Used indexing to return the value of mpg that corresponds to the day of my birthday
mdp3327[25,1]
```

```
## [1] 19.2
```

Answer: My birthday is May 25,2002. Thus, I used the indexing brackets to grab the value of mpg that corresponds to 25 by placing it in the following order: [25,1]. The value of mpg that I received as an output was 19.2 miles/gallon.

Question 5:

Using logical indexing, count the number of rows in the dataset where the variable `mpg` takes on values greater than 30.

```
# Used logical indexing and  sum() to count the number of rows in the dataset where the variable `mpg`

sum(mdp3327$mpg > 30)
```

```
## [1] 4
```

Answer: By using the 'sum()' function and logical indexing I was able to calculate the total number of rows in the dataset where the variable 'mpg' takes on values greater than 30 which is 4.

Question 6:

Let's create a new variable called `kpl` which converts the fuel efficiency `mpg` in kilometers per liter. Knowing that 1 mpg corresponds to 0.425 kpl, complete the following code and calculate the max kpl:

```
# Used the pipe operator to add a new variable called 'kpl' and did the appropriate conversion from mpg

mdp3327 = mdp3327 |>
        mutate(kpl = mpg * 0.425)

# Used summary() to find the max value of kpl
summary(mdp3327$kpl)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.420   6.556   8.160   8.539   9.690  14.408
```

Answer: I utilized the pipe operator to create a new variable called `kpl` and then calculated the max kpl of 14.408 kilometers/liter using the summary() function.

Part 2:

Let's quickly explore another built-in dataset: `airquality` which contains information about daily air quality measruements in New York, May to September 1973.

Question 7:

Calculate the mean `Ozone` (in ppb). Why does it make sense to get this answer? *Hint: take a look at the column `Ozone` in the dataset.*

```
# Attempted to calculate the mean of the 'Ozone' variable with mean()
mean(airquality$Ozone)
```

```
## [1] NA
```

```
# Looked at the 'Ozone' column using the head function
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

I attempted to calculate the mean for the ‘Ozone’ layer using the ‘mean()’ function, however, I received an output of ‘NA.’After looking at the dataset using the ’head()’ function, I noticed that there were a lot of ‘NA’ values within the ‘Ozone’ variable. This makes sense why the mean resulted in ‘NA’

Question 8:

Look at the documentation for the function `mean()` by running **`?mean` in the console**. What argument should be used to find the mean value that we were not able to get in the previous question? What type of values does that argument take?

Answer: We should be able to find the mean value that we were not able to get in the previous attempt using na.rm = TRUE inside the mean function. This argument takes boolean values such as TRUE and FALSE.

Question 9:

Sometimes the R documentation does not feel complete. We wish we had more information or more examples. Find a post online (include the link) that can help you use that argument in the `mean()` function. Then finally find the mean ozone!

```
# Utilized the na.rm argument inside the mean() function to find the mean ozone
mean(airquality$Ozone, na.rm = TRUE)
```

```
## [1] 42.12931
```

Answer: Below is a link to a post online that can help with understanding how to use na.rm in the ‘mean()’ function: After using the na.rm argument inside the ‘mean()’ function, I found the mean ozone which was 42.12931 ppb.

Part 3:

The Internet clothing retailer Stitch Fix has developed a new model for selling clothes to people online. Their basic approach is to send people a box of 5–6 items of clothing and allow them to try the clothes on. Customers keep (and pay for) what they like while mailing back the remaining clothes. Stitch Fix then sends customers a new box of clothes typically a month later.

A critical question for Stitch Fix to consider is “Which clothes should the send to each customer?” Since customers do not request specific clothes, Stitch Fix has to come up with 5–6 items on its own that it thinks the customers will like (and therefore buy). In order to learn something about each customer, they administer an **intake survey** when a customer first signs up for the service. The survey has about 20 questions and the data is then used to predict what kinds of clothes customers will like. In order to use the data from the intake survey, a statistical algorithm must be built in order to process the customer data and make clothing selections.

Suppose you are in charge of building the intake survey and the algorithm for choosing clothes based on the intake survey data.

Question 10:

What kinds of questions do you think might be useful to ask of a customer in an intake survey in order to better choose clothes for them? What kinds of data would be most valuable? See if you can come up with at least 5 items.

Answer: I think questions regarding both qualitative and quantitative data should be utilized in the intake survey. Five examples that I would include in the intake survey are the following: what clothes do they typically wear, what colors do they dislike, what region do they live in (weather), what fabrics do they prefer, and what size they prefer to wear.

Question 11:

In addition to the technical challenges of collecting the data and building this algorithm, you must also consider the impact the algorithm may have on the people involved. What potential negative impact might the algorithm have on the customers who are submitting their data? Consider both the data being submitted as well as the way in which the algorithm will be used when answering this question.

Answer: Some potential negative impacts that the algorithm might have on the customers would be that their data could be sold to third-party companies and they could receive many unwanted ads. Moreover, the algorithm itself could send clothes that may be offensive to the customer regarding the customers’ body image or other personal issues.