IMT 573: Module 5 Lab

Data Analysis

Jenny Skytta

Due: May 01, 2022

Collaborators: ***independent work*** List collaborators here.

Objectives

In this lab exercise you will drive your own data analysis and data science report. Your job is to develop a question to pursue and use data to support some preliminary conclusions. You should also find time to reflect on your results and identify possible errors or concerns you have about the data and analysis.

Instructions

Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Cloud.

1. Open the `05_lab_dataanalysis.Rmd` and save a copy to your local directory. Supply your solutions to the assignment by editing `05_lab_dataanalysis.Rmd`.
2. First, replace the "YOUR NAME HERE" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and I encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`. When the PDF report is generated rename the knitted PDF file to `lab5_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

In this lab you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(openintro)
library(knitr) # this will keep code on the page!
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Data

In this lab we will be working with data from fast food restaurants – you saw this same data in the module demonstration. This dataset contains nutritional information for 515 menu items from some of the most popular fast food restaurants worldwide. You can use the follow code to load and inspect this data.

```
# Load data and inspect it
data(fastfood)
ls()  #show my column and row totals
```

```
## [1] "fastfood"
```

```
glimpse(fastfood)
```

```
## Rows: 515
## Columns: 17
## $ restaurant  <chr> "Mcdonalds", "Mcdonalds", "Mcdonalds", "Mcdonalds", "Mcdon~
## $ item        <chr> "Artisan Grilled Chicken Sandwich", "Single Bacon Smokehou~
## $ calories    <dbl> 380, 840, 1130, 750, 920, 540, 300, 510, 430, 770, 380, 62~
## $ cal_fat     <dbl> 60, 410, 600, 280, 410, 250, 100, 210, 190, 400, 170, 300,~
## $ total_fat   <dbl> 7, 45, 67, 31, 45, 28, 12, 24, 21, 45, 18, 34, 20, 34, 8, ~
## $ sat_fat     <dbl> 2.0, 17.0, 27.0, 10.0, 12.0, 10.0, 5.0, 4.0, 11.0, 21.0, 4~
## $ trans_fat   <dbl> 0.0, 1.5, 3.0, 0.5, 0.5, 1.0, 0.5, 0.0, 1.0, 2.5, 0.0, 1.5~
## $ cholesterol <dbl> 95, 130, 220, 155, 120, 80, 40, 65, 85, 175, 40, 95, 125, ~
## $ sodium      <dbl> 1110, 1580, 1920, 1940, 1980, 950, 680, 1040, 1040, 1290, ~
## $ total_carb  <dbl> 44, 62, 63, 62, 81, 46, 33, 49, 35, 42, 38, 48, 48, 67, 31~
## $ fiber       <dbl> 3, 2, 3, 2, 4, 3, 2, 3, 2, 3, 2, 3, 3, 5, 2, 2, 3, 3, 5, 2~
## $ sugar       <dbl> 11, 18, 18, 18, 18, 9, 7, 6, 7, 10, 5, 11, 11, 11, 6, 3, 1~
## $ protein     <dbl> 37, 46, 70, 55, 46, 25, 15, 25, 25, 51, 15, 32, 42, 33, 13~
## $ vit_a       <dbl> 4, 6, 10, 6, 6, 10, 10, 0, 20, 20, 2, 10, 10, 10, 2, 4, 6,~
## $ vit_c       <dbl> 20, 20, 20, 25, 20, 2, 2, 4, 4, 6, 0, 10, 20, 15, 2, 6, 15~
## $ calcium     <dbl> 20, 20, 50, 20, 20, 15, 10, 2, 15, 20, 15, 35, 35, 35, 4, ~
## $ salad       <chr> "Other", "Other", "Other", "Other", "Other", "Other", "Oth~
```

```
# view a short grouping that displays variables and data
```

Problem 1: Formulate a Question

First, formulate one data science question of interest that can be answered with this dataset. Be sure to comment on why this question in interesting and what you could learn from finding an answer to it.

Question:

For someone following a low carb diet, which restaurant between Subway and Mcdonalds offers more menu items with carbs lower than 20 g items in their menu?

Problem 2: Data Analysis

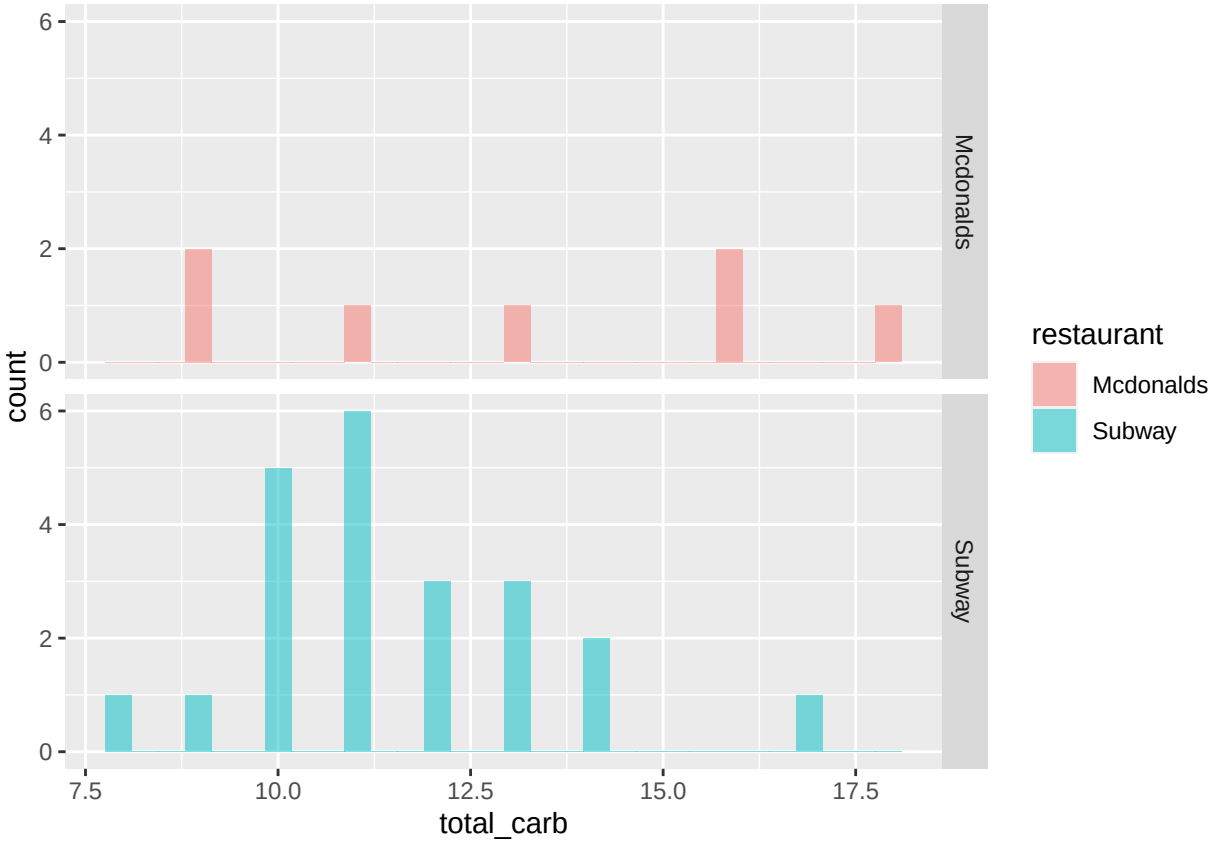
Next, practice using your data science skills to answer your question. Follow the steps below in your data science process. In your analysis, use at least one data visualization to help you communicate your findings.

```
keto_menu <- fastfood %>% #load dataframe and assign variable
  filter(restaurant == "Mcdonalds" | restaurant == "Subway") %>% #filter for restaurants
         filter(total_carb < 20) %>% #filter under 20 carbs
```

```
            select(restaurant, total_carb) #select my variables of choice


ggplot(keto_menu, aes(total_carb, fill = restaurant)) + #plot my tibble
  geom_histogram(alpha = 0.5, position = "identity") + #histogram to view distribution
  facet_grid(rows = vars(restaurant)) #split to view difference between restaurants
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From this dataset, Subway offers the most (n=22) under 20 carb options in their menu items when compared to Mcdonalds (n=7).

(a) Try the Easy Solution First

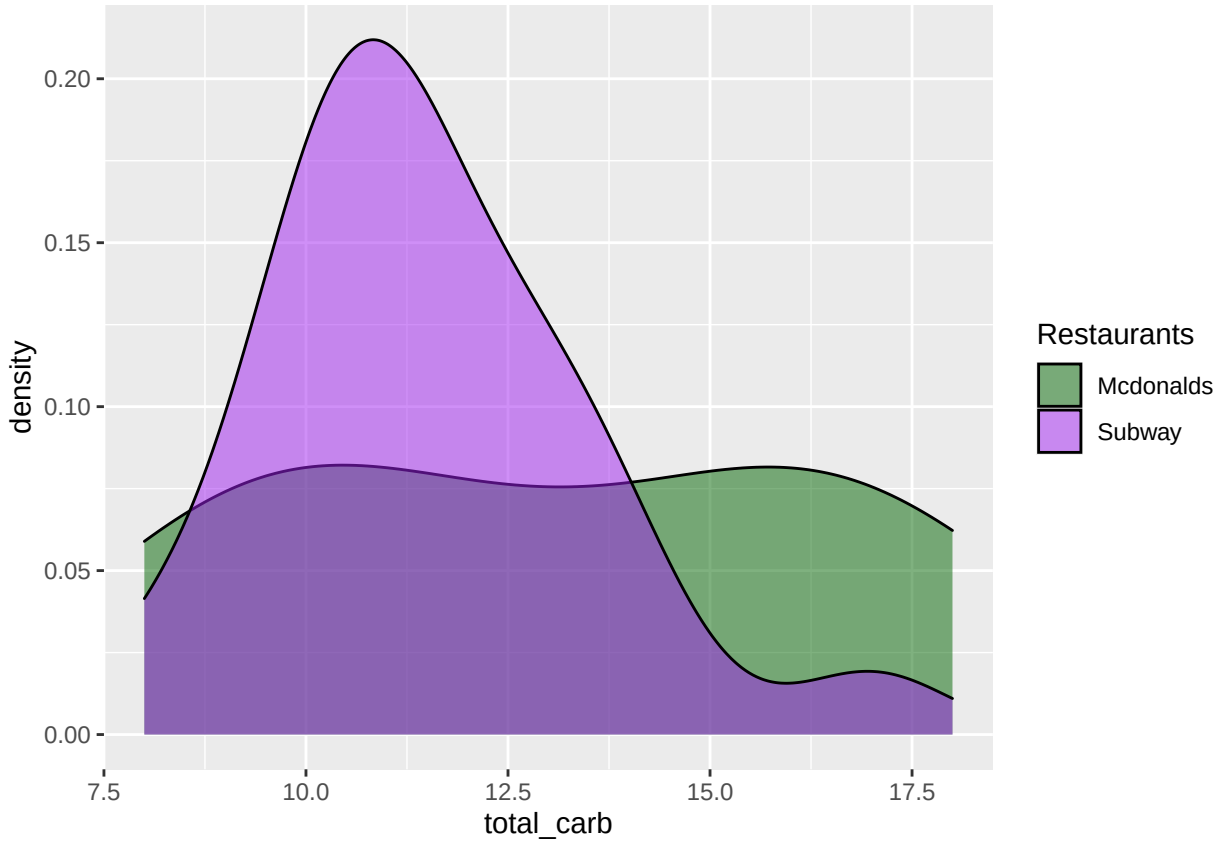
```
# look at desdriptive stats
stats <- keto_menu %>%
  group_by(restaurant) %>%  #group by restaurant
  summarise(mean = mean(total_carb, na.rm = TRUE), #summarize average total carbs
            sd = sd(total_carb, na.rm = TRUE)) #show the standard deviation

stats
```

```
## # A tibble: 2 x 3
##   restaurant  mean    sd
##   <chr>      <dbl> <dbl>
## 1 Mcdonalds   13.1  3.63
```

```
## 2 Subway      11.5  1.97
```

```
ggplot(keto_menu, aes(x=total_carb, fill = restaurant)) + #create density vizualization
  geom_density(alpha=0.5) +
  scale_fill_manual(values = c("darkgreen","purple"), labels = c("Mcdonalds","Subway"), name = "Restaura
```

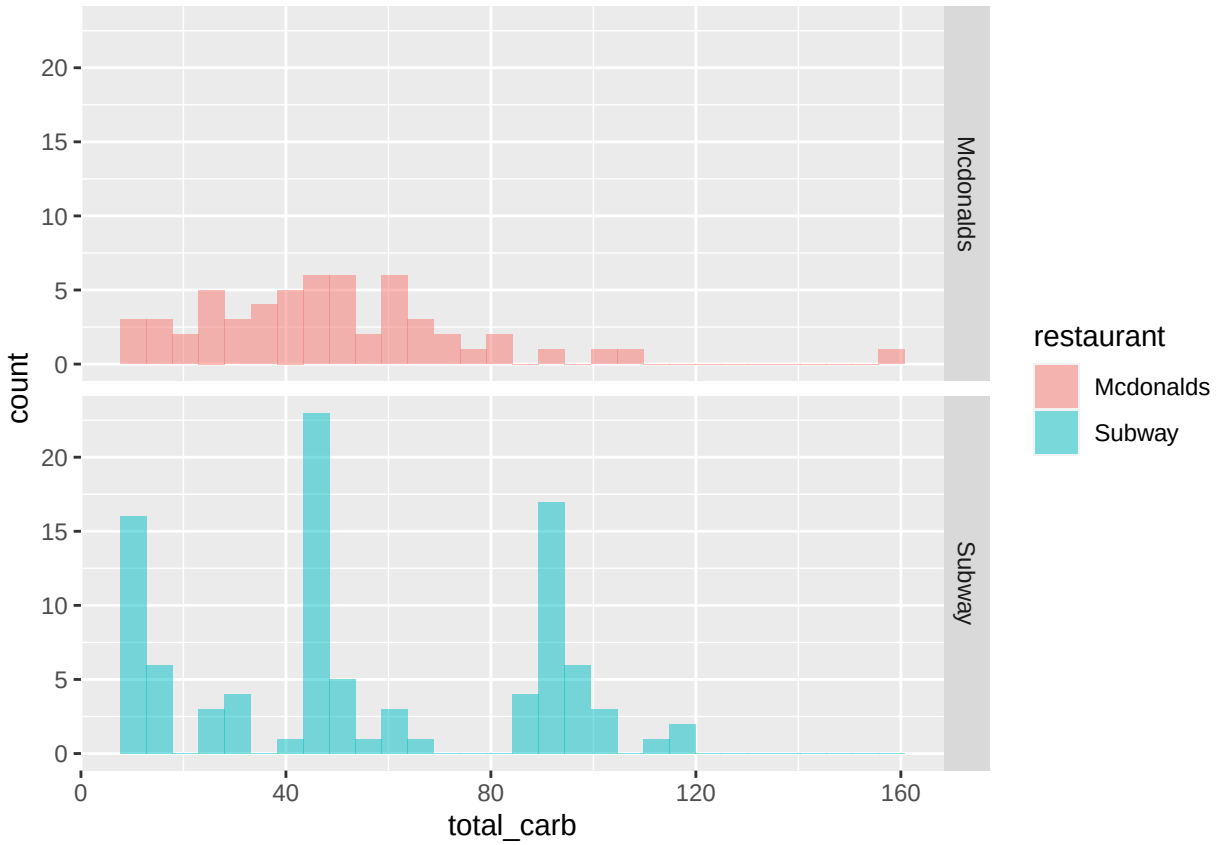


(b) Challenge Your Solution

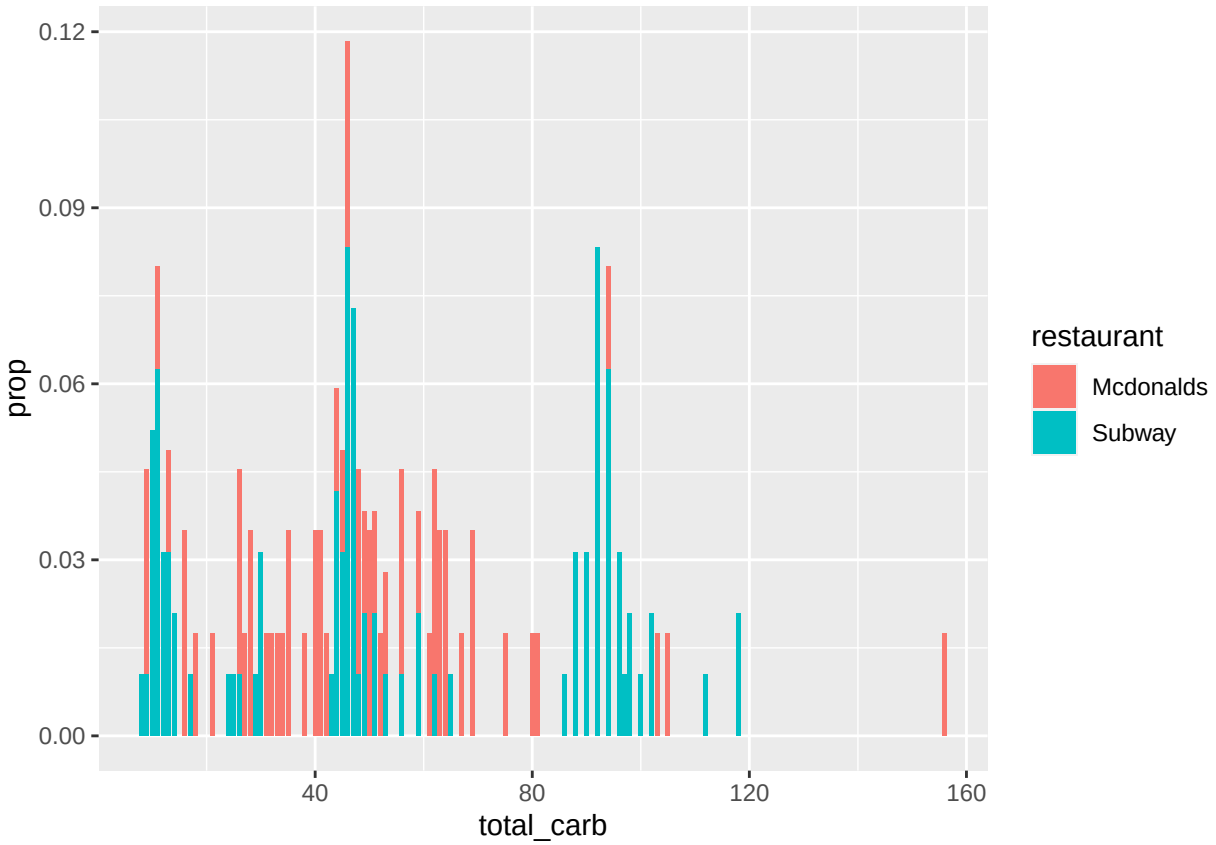
```
total_menu <- fastfood %>% # look at total menus of both restaurants
  filter(restaurant == "Mcdonalds" | restaurant == "Subway") %>%
       select(restaurant, total_carb)

#plot the menu items broken down by restaurant with visual of carb breakdown
ggplot(total_menu, aes(total_carb, fill = restaurant)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  facet_grid(rows = vars(restaurant))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#look at average of under 20 carb menu items per restaurant
avg_menu <- fastfood %>%
  filter(restaurant == "Mcdonalds" | restaurant == "Subway") %>%
    group_by(restaurant) %>%
       summarise(N = n(), total_carb = mean(total_carb))
#plot the proportion of total carb distribution by restaurant
ggplot(data = total_menu) +
  geom_bar(mapping = aes(x = total_carb, y = stat(prop), fill = restaurant))
```



Is it possible that Subway offers more menu options overall? Subway does have a different model for menu items than Mcdonalds as they specifically ask for orders to be customized at the counter.

It does appear that overall Subway offers more menu items overall (n=96) compared to Mcdonalds (n=57). In the barchart, we can see that there is a greater proportion of the subway menu that is over 20 total carbs.

(c) What Next?

Another approach, if provided with more specific data, would be to break down the menus by ingrendient offerings to assess if overall, one menu had more readily available customaiztaions that when combined were below 20 total carbs.

Citations

Code written above is from the previous course IMT 511 which used the below text to support class scripts. https://www.google.com/books/edition/Programming_Skills_for_Data_Science/BnB6DwAAQBAJ?hl=en&gbpv=1&printsec=frontcover —