Dplyr Homework

The data for this homework comes from an Open Data Science website called Kaggle. Kaggle has many open source datasets for you to use and most of them come with code uploaded by other users showing different ways to explore the data. It is a great way to learn about data-wrangling and analysis and if you are interested then set up your own account and get going.

For this task we will make use of their **Starbucks Calorie** dataset. You can find out more information about each dataset and what each column represents , but we have put the version you need in your data folder.

title: "R Notebook"

output: html\_notebook

---

library (tidyverse)

library (devtools)

library (CodeClanData)

library (readxl)

hw\_dplyr <- read.csv("data/starbucks\_drinkMenu\_expanded.csv")

hw\_dplyr

str(hw\_dplyr)

1 Question 1

Load in the dataset and have a look at it. What are the dimensions, variable types, variable names, etc.?

2 Question 2

Let’s first investigate the calories of different drinks. Select the variables Beverage\_category, Beverage, Beverage prep and Calories from your data. Since we are interested in the calorie content, check if there are any NA values in the data, and drop them if there are.

**hw\_dplyr %>%**

**select (Beverage\_category,Beverage,Beverage\_prep,Calories) %>%**

**filter(!is.na(Calories))**

*# let's check if there are any missing values*

drinks\_content %>%

select(Beverage\_category, Beverage, Beverage\_prep, Calories) %>%

filter(is.na(Calories)) %>%

summarise(count\_of\_missing\_calories = n())

|  |
| --- |
|  |

| **count\_of\_missing\_calories**  <int> |
| --- |
| 0 |

1 row

*#there isn't any drinks with missing Calories so we are all good!*

3 Question 3

Create a new variable (column) called calorie\_diff, which stores the difference between 135 calories (135 calories = 10 cubes of sugar!) and the calories in each drink. (hint: you’ll want to subtract the calories from 135 to see which drink have more or less than 10 cups of sugar).

**mutate(hw\_dplyr, calorie\_diff = (Calories - 135), .after = Calories )**

4 Question 4

Summarise the mean number of calories in each beverage\_category. Which 3 drinks have the most calories? Which 3 drinks have the least? Write a small summary of your findings.

**bv\_cat <- hw\_dplyr %>%**

**group\_by(Beverage\_category) %>%**

**summarise (sum\_cat = mean(Calories)) %>%**

**arrange(desc(sum\_cat))**

**bv\_cat1 <- head(bv\_cat,3)**

**bv\_cat2 <- tail (bv\_cat,3)**

**summary(bv\_cat)**

most\_calories <- drink\_calories %>%

group\_by(Beverage\_category) %>%

summarise(mean\_calories = mean(Calories)) %>%

arrange(desc(mean\_calories)) %>%

head(3)

## `summarise()` ungrouping output (override with `.groups` argument)

most\_calories

|  |
| --- |
|  |

| **Beverage\_category**  <chr> | **mean\_calories**  <dbl> |
| --- | --- |
| Smoothies | 282.2222 |
| Frappuccino® Blended Coffee | 276.9444 |
| Signature Espresso Drinks | 250.0000 |

3 rows

least\_calories <- drink\_calories %>%

group\_by(Beverage\_category) %>%

summarise(mean\_calories = mean(Calories)) %>%

arrange(mean\_calories)%>%

head(3)

## `summarise()` ungrouping output (override with `.groups` argument)

least\_calories

5 Question 5

Let’s look at this a different way. What is the average number of calories in each Beverage\_prep type?

**hw\_dplyr %>%**

**group\_by(Beverage\_prep) %>%**

**summarise(Average\_calorie = mean(Calories))**

*# work out the mean calories for each beverage prep*

average\_bev\_prep <- drink\_calories %>%

group\_by(Beverage\_prep) %>%

summarise(mean\_cals = mean(Calories)) %>%

arrange(desc(mean\_cals))

6 Question 6

Which Beverage\_prep type contains more than the average calories of all drinks?  
*Hint: to answer this, you’ll have to first figure out what the average calories across all drinks are, and then use that as a filter for the grouped Beverage\_prep data.*

**cal\_mean <- mean(hw\_dplyr$Calories)**

**cal\_mean**

**filter(hw\_dplyr,Calories> cal\_mean)**

7 Question 7

Which is the best type of **coffee** drink to get if you’re worried about consuming too many calories?

**hw\_dplyr %>%**

**group\_by(Beverage) %>%**

**filter(Calories <=5)**

drink\_calories %>%

group\_by(Beverage\_category, Beverage, Beverage\_prep) %>%

summarise(min\_calories = min(Calories)) %>%

arrange(min\_calories) %>%

head(1)

## `summarise()` regrouping output by 'Beverage\_category', 'Beverage' (override with `.groups` argument)

|  |
| --- |
|  |

| **Beverage\_category**  <chr> | **Beverage**  <chr> | **Beverage\_prep**  <chr> | **min\_calories**  <dbl> |
| --- | --- | --- | --- |
| Tazo® Tea Drinks | Tazo® Tea | Grande | 0 |

1 row

*# But this is a Tea and the answer specifically asked for coffee drink! So we will filter out this tea drink to get the lowest calorie coffee drink.*

drink\_calories %>%

filter(Beverage\_category != "Tazo® Tea Drinks") %>%

group\_by(Beverage\_category, Beverage, Beverage\_prep) %>%

summarise(min\_calories = min(Calories)) %>%

arrange(min\_calories) %>%

head(1)