Homework 4: Introduction To Tidyverse

Enter Your Name and UNI Here

June 1, 2021

Instructions

Please submit both pdf and Rmd files (or html and Rmd files).

Part II: Split/Apply/Combine and tidyverse warm-up

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

Consider the following **loop** that computes the mean of each quantitative variable split by species and stores the computed means in a matrix named **MeanFlowers**.

```
# define a matrix of zeros
MeanFlowers <- matrix(0,nrow=4,ncol=3)

# define a character vector corresponding to the numeric variable names
measurements <- c("Sepal.Length","Sepal.Width","Petal.Length","Petal.Width")

# name the rows and columns of the matrix MeanFlowers
rownames(MeanFlowers)  <- measurements
colnames(MeanFlowers)  <- c("setosa","versicolor","virginica")

# Loop
for (j in measurements) {
  #-- R code goes here ----
  MeanFlowers[j,] <- round(tapply(iris[,j],iris[,"Species"],mean),4)
}
MeanFlowers
```

```
##              setosa versicolor virginica
## Sepal.Length  5.006      5.936     6.588
## Sepal.Width   3.428      2.770     2.974
## Petal.Length  1.462      4.260     5.552
## Petal.Width   0.246      1.326     2.026
```

Problem 1

Replicate the above loop using the **Split/Apply/Combine** model with base R commands.

Solution goes below

```
## solution goes here --
```

Problem 2

Repeat question 1 by constructing a pipe, including the **split()** function from base R and **map_df()** from the **purrr** package.

Solution goes below

```
## solution goes here --
```

Part II: More tidyverse with CDC cancer data

Consider the Center of Disease Control data set **BYSITE_new.csv**, which describes the incidence and mortality counts of several types of cancer over time. The variables of interest are: **YEAR**, **RACE**, **SITE**, **EVENT_TYPE**, **COUNT** and **POPULATION**.

Problem 3

Load in the dataset **BYSITE_new.csv** using the appropriate function from the **readr** package. Display the dimension of the cancer tibble.

Solution goes below

```
## solution goes here --
```

Base R code for reference.

```
# Base R code for reference
cancer <- read.csv("BYSITE_new.csv",header=T)
dim(cancer)
```

```
## [1] 44982     7
```

Problem 4

Using Base R or tydyverse functions, identify any strange symbols that are recorded in the **COUNT** variable. Once you have identified the symbols, use functions from the **dplyr** package to remove any rows in the cancer tibble containing these symbols and then convert **COUNT** to a numeric mode.

```
## solution goes here --
```

Problem 5

For a specific tumor and population, a crude rate is calculated by dividing the number of new cancers observed during a given time period by the corresponding number of people in the population at risk. For cancer, the result is usually expressed as an annual rate per 100,000 persons at risk. https://ci5.iarc.fr/ci5plus/pages/glossary.aspx

In reference to our data, this quantity can be calculated by:

$$\text{CRUDE RATE} = 100000 * \frac{\text{COUNT}}{\text{POPULATION}}$$

Using relevant functions from the **dplyr** package, create a new variable in your dataframe (or tibble) called **CRUDE_RATE**. Then using base R graphics or ggplot, create a histogram of **CRUDE_RATE**. Note that the crude rates are not bounded between [0,1] because they are calculated per 100,000 persons at risk.

```
## solution goes here --
```

Problem 6

Compute the average incidence rate of prostate cancer for each level of **RACE**. To solve this problem, students must build a pipe (**magrittr** package) and utilize the appropriate functions from the **dplyr** package. Also compare your results to a base R solution. Include both the tidyverse and base R solutions in your final write-up. **Note:** before computing the average incidence rates, students should filter the data as follows:

i. Extract the rows corresponding to **EVENT_TYPE** level **Incidence**

ii. Extract the rows corresponding to **SITE** level **Prostate**

iii. Extract the rows corresponding to **SEX** level **Male**

iv. Remove the rows corresponding to **YEAR** level **2010-2014**

v. Remove the rows corresponding to **RACE** level **All Races**

Solution goes below

First filter the dataset:

```
## solution goes here --
```

Compute the average incidence rate of prostate cancer for each level of **RACE**.

```
## solution goes here --
```

Problem 7

Create a plot in base R or ggplot that shows the incidence rate (**CRUDE_RATE**) as a function of time (**YEAR**), split by the levels of **RACE**. Make sure to include a legend and label the graphic appropriately. Before constructing the graphic, perform the data wrangling tasks using a pipe and functions from the **dplyr** package, i.e., the same filtering tasks from problem 6. Students can use some base R functions in the pipe if needed and the plotting code can be included inside or outside the pipe.

Solution goes below

```
## solution goes here --
```

Problem 8

Fit five simple linear regression models, one for each level of **RACE**, relating the incidence rate (**CRUDE_RATE**) as a function of time (**YEAR**). Collect the estimated slopes, t-statistics and p-values of your estimated models. The collection of slopes describe whether cancer has increased or decreased over the selected time period and the p-values describe if the increase or decrease is statistically significant. Solve this problem using a pipe and functions from the **dplyr** and **purrr** packages. **Note:** use the same filtered data from problem 4 and problem 4 in this analysis.

Some hints: (i) this exercise is a natural extension of problem 7; (ii) if needed, students can also define their own functions used in the pipe; (iii) students are not required to use a single pipe to solve this question but it's a fun challenge if interested.

Solution goes below

```
## solution goes here --
```