Lab 4: Wrangling and Visualizations

This worksheet is due by 8pm one day after your lab. You can find the submission dropbox in Brightspace by clicking on Content $->$ Lab Content.

Download the files "Lab4_Worksheet.pdf", "Stat123_Lab4_YourLastName.Rmd" and "Superstores.csv" from Brightspace and save them in your working directory.

Open the Rmarkdown file and write your code in the file.

1. Download the data sets Superstores.csv and save them to whatever directory you are using for this course.

   (a) Load the tidyverse package. This will load both dplyr and ggplot2 packages.

   (b) Read in the file and save it as a data frame *super*. This data set contents information about customer orders. Take a look at the data frame to familiarize yourself with its contents. It is a typical data set with lots of variables. Note that the column "Row ID" is not a variable. It's an identifier, or an index.

   (c) Use a dplyr command to show the total Sales for the Category "Office Supplies".

   (d) Use a dplyr command to create a new data frame from *super* that contains only order information for the California state. Name it superCali and show a few observations.

   (e) Use the original *super* data frame to produce a bar graph of order counts by Regions. Set each bar to a different colour and give an appropriate title. Once you have the basic bar graph, you can experiment with different layouts.

2. Install and load the package **palmerpenguins**. If you write the install command in the chunk, make sure you comment it out after the package is installed. You don't need to install the package multiple times in a session.

   (a) Display a few observations of the data frame *penguins* and get familiar with the variables. If you want to know what the data frame describes. You can type *?penguins* in the console.

   (b) Produce a ggplot2 histogram showing the distribution of penguins body mass in the data set. Add a vertical line showing the average body mass. Then add a title.

   Note: When you calculate the mean body mass, add a parameter $na.rm = TRUE$ in the mean() command. This will exclude any missing values in the data set.

3. Using the penguin data set, produce basic a line plot putting penguins' body flipper length in the x-axis and their body mass in the y-axis.

   What do you think of the plot? Do you think a line plot is appropriate? Why or why not?

4. Once you make sure all the code works in the R markdown file, knit it to an HTML. Make sure the file contains all answers to the questions. Then open the knitted file and print it as a PDF file. The name of the file should be **Stat123_Lab4_YourLastName.pdf**. Then submit the pdf file to the appropriate Brightspace folder.