HW2 - Basic GGPlots

2024-09-06

Instructions

1. Store all your answers in the HW2.R file. Use the variable names given in the instructions(this pdf) and in the HW2.R file. Do not change the variable names for the output of each question as these will be used to check the correctness of your answers by Gradescope. Do not change the name of the file HW2.R. Changing the name will mess up the running of the autograder.

2. Write your name in the "YOUR NAME HERE" comment in the HW2.R file.

3. You can re-do and resubmit your assignment to gradescope as many times as you'd like before the deadline. If you're sure your answer is correct but gradescope marks it as wrong, send Ajay (ak834@bu.edu) an email or a teams message with a screenshot of the relevant section of your code and the Autograder will be modified if it needs to be.

4. Due Date: This assignment is due 11:59pm September 13 2024

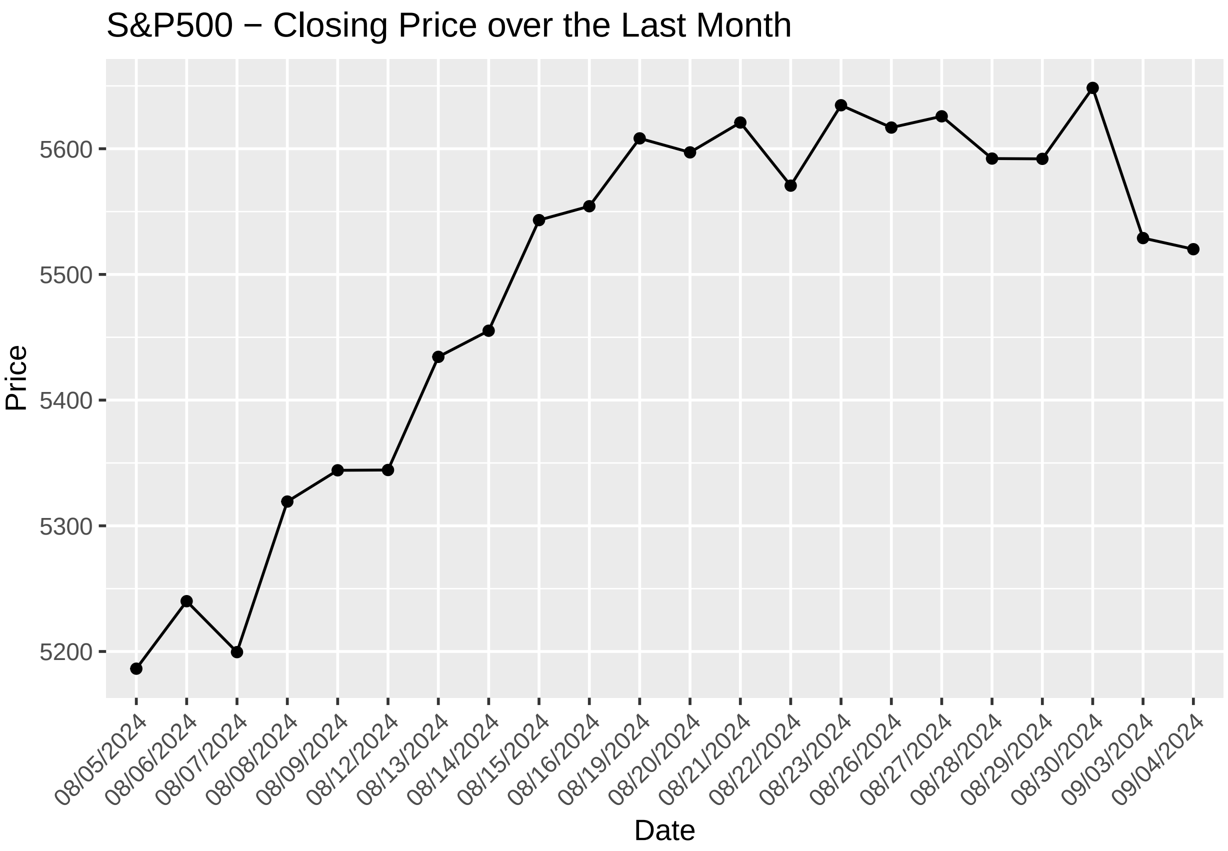
5. It is recommended that you create a github to hold all your homework files. You can then submit this github link to Gradescope.

6. If you are having any trouble with this assignment, please feel free to come to us for help or to turn to your classmates. If you do take help from classmates, you must credit them in a comment at the bottom of the file you turn in. As long as you credit classmates for help, seeking help from them will not affect your grade adversely at all.

Exercise 1

We'll start by creating a simple plot. There are a couple of subtleties to this, so for this first exercise, we'll walk through the steps involving those.

Your job is to recreate the following plot of S&P500 prices for the last month.



Step 1: Read in the csv file as a dataframe. You can use the 'read.csv' function or import the "readr" library and use the 'read_csv' function. This is an entirely arbitrary choice.
(If you are unfamiliar with what a dataframe is or need a refresher since Bootcamp, please let us know and we'll happily run you through it. Alternatively this is a good resource to consult:https://www.w3schools.com/r/r_data_frames.asp).

Step 2: Create a basic plot that looks like the one above. Do not worry about the labels and axes and title yet. Define the data as the variable you read the S&P data into. Look at the data. What will the mappings in the aes argument be for ggplot- What is x and what is y?

Next, what geoms do we want to use? Note that we have the line plotted and also the points.

For line graphs ggplot needs to know which data points need a line drawn through them. This is typically read from groupings in the data. In the absence of that (as in this case) you need to supply 'group=1' to ggplot. You can do this in the aes mapping for ggplot or for the appropriate geom.

Assign this plot to the variable spx_plot1. You will need to call spx_plot1 to see what the plot looks like. Do you notice anything wrong with this plot?

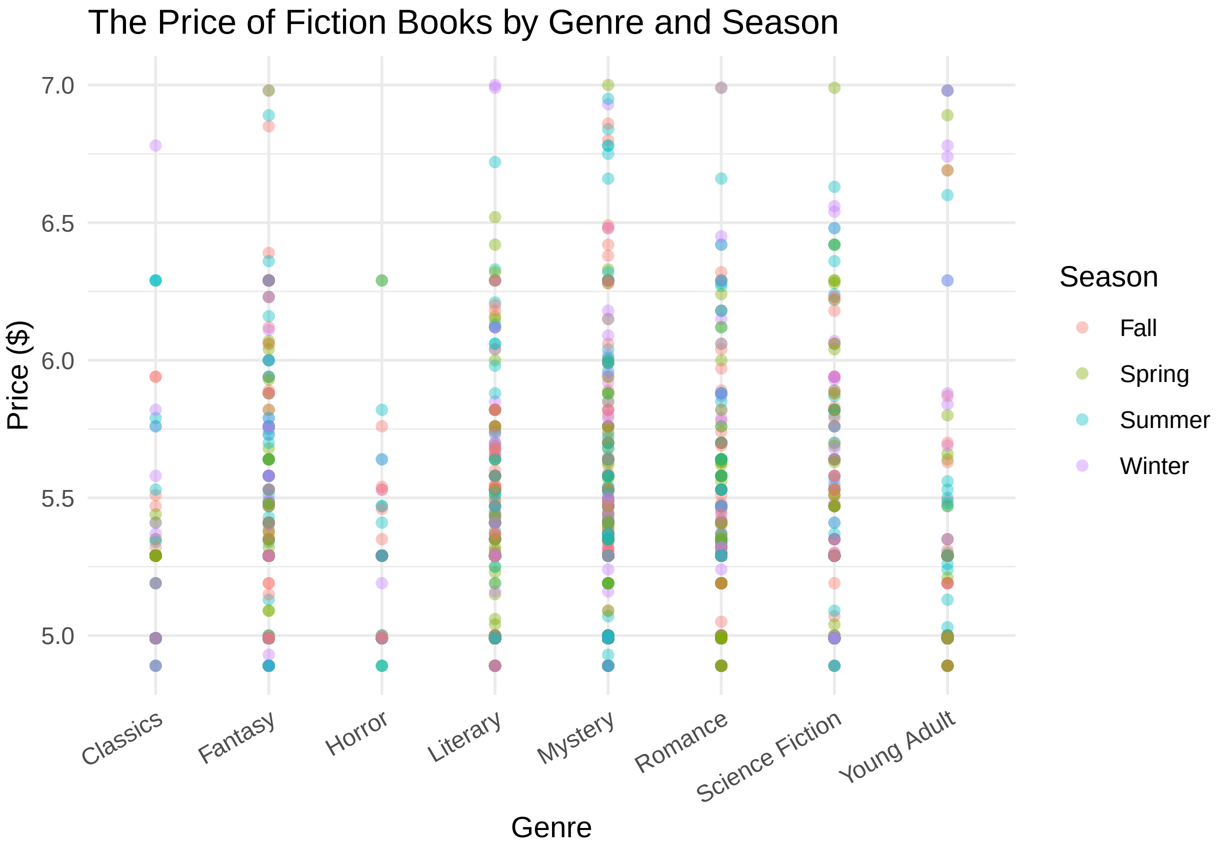
Step 3: There are three things wrong with the plot - the y axis label, the fact that there is no title and the fact that you cannot read the dates clearly. You can solve the first two problems by adding a labs() layer to your plot. Use '?' to see what this does and how to use it. The third issue is solved using theme(). We will cover theme in class and there will be more detailed exercises on its use in future problem sets. For now, know that you can change the angle of the text and horizontal adjustment by adding this layer to your ggplot 'theme(axis.text.x=element_text(angle=45,hjust=1))'.

Assign this plot to spx_plot2.

Exercise 2

For this exercise we'll use a dataset on Books which I've cleaned to include only a few genres. Fair warning, this graph is not the best way to represent this data. There are things we can do this graph that can make it easier to read and which will convey its information better. We will make those changes and improvements in our next ggplot assignment.

For now, write ggplot code to recreate the following plot:



- Use the BookGenres.csv file for the data
- Some of the points look transparent. Use alpha for this. Look up how to use alpha in R4DS or using '?'. Feel free to toggle the value and see what the plot looks like at different alphas. When you submit your assignment though, be sure it's set to 0.40.
- Set the text angle to 30 and use the theme_minimal to recreate this background and these grid_lines.

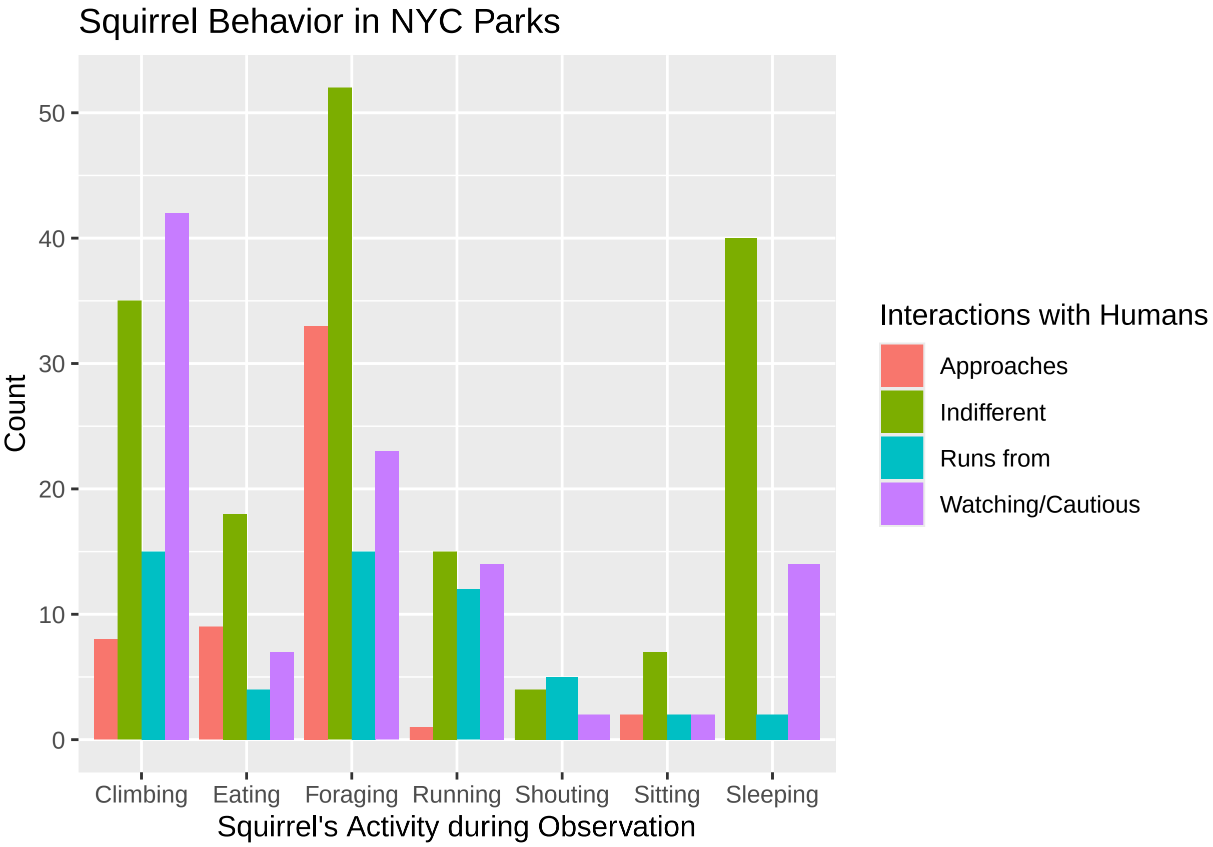
Assign this plot to the variable 'bookplot'

Exercise 3

You may or may not know this, but in 2020, New York City conducted a census of the squirrels in its parks. Why they did this in the first place is unclear. The fact that they haven't seen fit to conduct a squirrel census since is an enduring tragedy.

I have taken a subset of this dataset and stored it in squirrel.csv. Recreate the plot below. Again, there are improvements we can make to this plot and we will work on those in the next assignment.

Assign this plot to 'squirrelPlot'. Use theme_grey().



Exercise 4

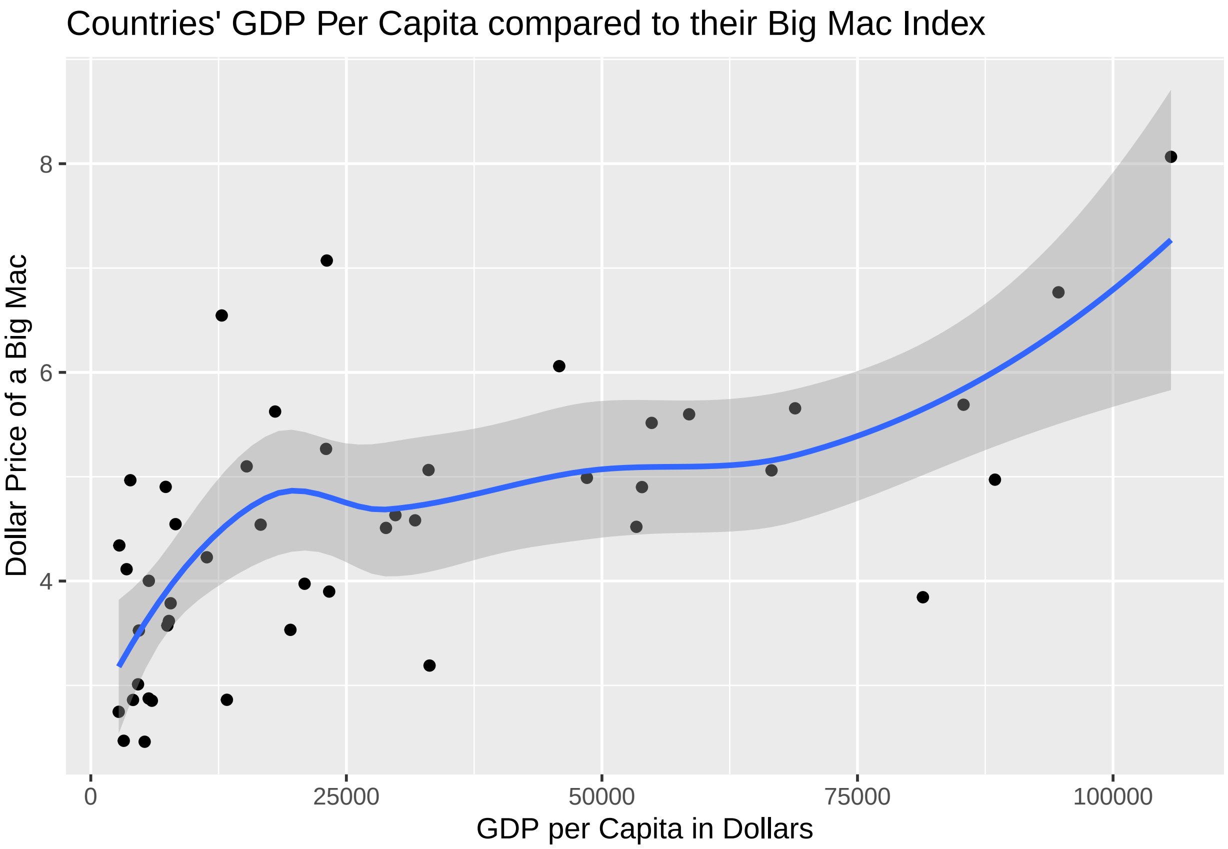
Some of you may have heard of The Economist Magazine's Big Mac Index. For those who haven't, loosely, it is an indicator of the relative exchange rate between countries based on the difference in the price of a Big Mac in each country.

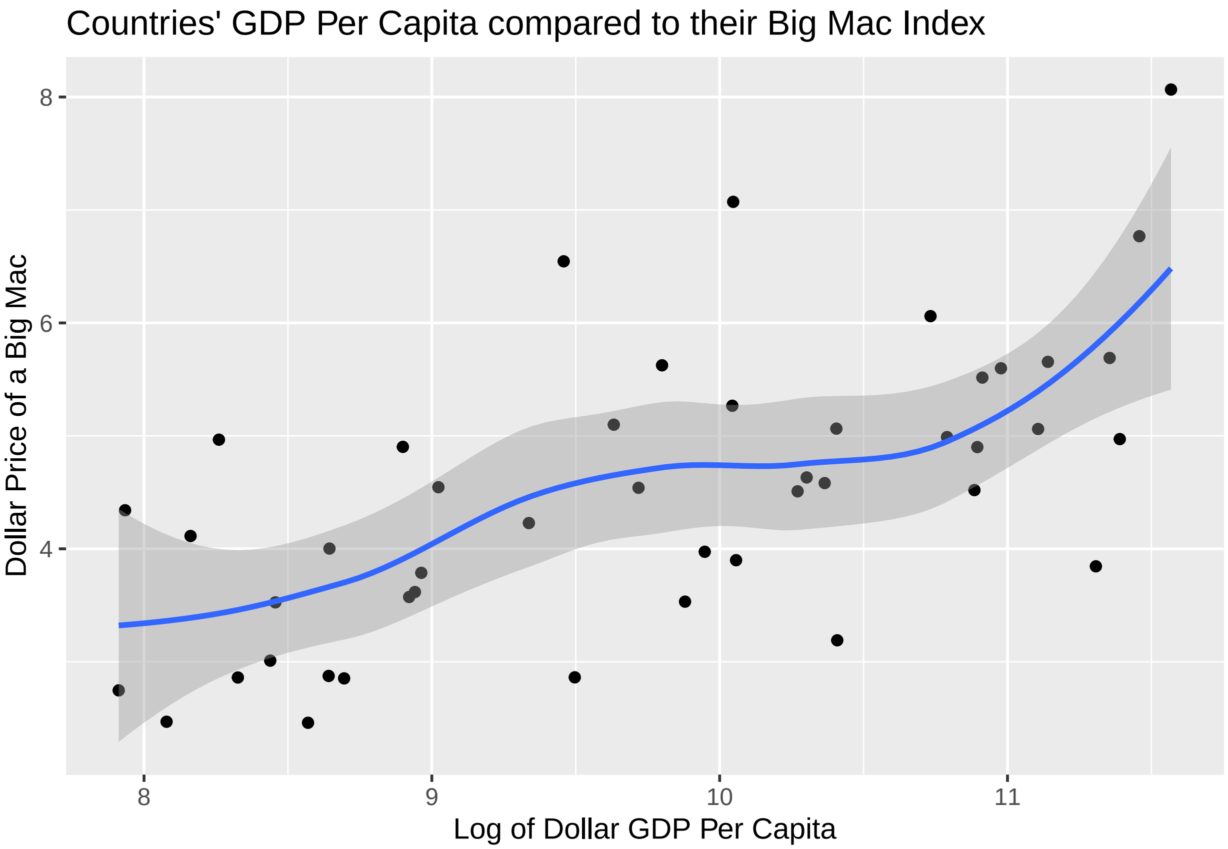
There's a lot of interesting things that we can do with the Big Mac Index, like comparing a country's reported inflation rate with the change in its Big Mac Index (in fact some governments have had their official figures brought into question due to massive discrepancies between their numbers and the Big Mac Index). We'll look at some interesting visualizations in the next problem set. Here, we are making some simple plots of GDP Per Capita against the Big Mac Index.

Note the difference between the two plots. A log transformation of a variable is something you will encounter frequently. In addition to modifying the GDP Per Capita for the second plot so that it is log-transformed, you will also need to make a change to the GDP Per Capita variable to make the plot work. To figure out what that change may be, use the str() function on the columns of the dataframe.

Hint: If you're struggling to figure out what geoms you need to use for this one, read the R 4 Data Science book's first chapter. The relevant geoms are introduced very early on.

Call the first plot 'bigMac' and the second plot 'logBigMac'





A few questions to ponder, the answers to which will be discussed in coming weeks in this class and the practicum.

- What is the difference between the line on these plots and the line in the first plot?
- What is the shaded area indicating? Why is this important to us as statisticians?