Instructions: This assignment will give you a chance to practice what you have learned about graphing using the ggplot2 package. You will also get more practice with extracting items from a data frame.

For each problem copy and paste your R code from the R script file window AND your output – either from the Console window or the Plots tab – into this Word document. Please use a color other than black for your R code. Graphs MUST be made using ggplot2 functions, unless otherwise specified.

You will also be asked to interpret/describe what you learned from many of the graphs that you make. You will upload this Word document to Bb when finished.

Use Beginning R by Gardener as a resource. See HW #2 for a list of sections in the text that may be helpful. Also see chapter 3 of R for Data Science by Wickham and Grolemund.

**PART ONE:**

1. Download the .csv file posted with this assignment on Bb. Then read the *house-and-land-values.csv* file into R. View the first few rows of the data set using the *head()* function. [Use the link provided on Bb to learn about the data.]
2. In preparation for this assignment, create several subsets from the *house-and-land-values* data set, as described below. You will create a new object in R, with a descriptive name, for each. After the subset of data is created, view the first and last few rows using the *head()* and *tail()* functions.

* Subset #1: Just the data for Iowa
* Subset #2: Just the data from California, starting at 1985.
* Subset #3: Data for just the following states: Iowa, California, Massachusetts, and Texas.
* Subset #4: Just the data for the States that have a specified region, i.e. remove any rows that have *NA* values for region
* Subset #5: Data for just the first quarter of 2013 (this is the most recent data available in the data set)

1. Make a histogram displaying the distribution of home values in Iowa between 1975 and the beginning of 2013. [Use the first subset you created.] Set the *binwidth to $50,000*, and to make readability easier, add the argument *color= “green”* outside of *aes().* You may choose a color other than green, if you wish.

What have you learned about home values in Iowa? Write at least 2 sentences.

1. Make a scatter plot showing how home values have changed over time in Iowa. Make all points in the scatter plot a *color* other than black (your choice) and add an appropriate title to the graph using *labs().* You may use either *Year* or *Date* for the x-axis.

Refer to problem 1 ~ What additional information have you learned about home values in Iowa through your scatterplot? Write at least 2 sentences.

1. Add another layer to the scatterplot that you made in problem 4; include how land values have changed over time in Iowa. You may need to adjust your title and y-axis label.

What have you learned about land values in Iowa? Write at least 2 sentences. Also write a couple of sentences relating home values to land values.

1. Make another adjustment to your graph from problems 4 & 5. Add a legend to the graph by doing the following:
   * Create a vector containing the two colors that you would like to use for home and land values. Call the vector: *Graph.Colors*
   * Rather than specify a color for the home value scatterplot outside of the *aes()* function, put *color= “Home.Value”* inside *aes().* The quotes are important!
   * Do the same for the land value layer: put *color= “Land.Value”* inside the *aes()* function.
   * Add *color= “Legend”* to the *labs()* layer.
   * Add a layer to the graph that allows you to manually select the colors that you use: *scale\_color\_manual(values = Graph.Colors)*

No graph interpretation required. Note that a legend is only generated when you map a variable to an aesthetic, and we did NOT do that in problem 5.

1. Morris A. Davis is an assistant professor of Real Estate and Urban Land Economics at the Wisconsin School of Business and a fellow at the Lincoln Institute of Land Policy, who created the data sets. Dr. Davis wrote,

*“Price indices and values of land inform the analysis of trends and cycles in house prices. If housing were simply a manufactured good, and location or land had no value, then the price of housing would be determined by construction costs, and housing prices would increase at roughly the same rate as the price of other goods in the US economy.*

*But housing is on land with a specific location, and good locations are often scarce and valuable. If construction costs rise slowly over time and desirable locations are in limited supply, increases in the demand for housing can translate directly to increases in the price of good locations – the land – and in house prices.”*

For the most part, home values in Iowa are based on the structural cost of the home. To see this, add a layer containing the structure cost to your graph from problem 6. Make sure the legend on the graph works appropriately. No graph interpretation required.

1. A better example of Dr. Davis’ statement (see problem 7) comes from California. Recreate your graph from problem 7 using data from just California, starting at the Year 1985. [Use the second subset you created.]

Explain how Dr. Davis’ statement is exemplified by your graph. Write at least three sentences.

**PART TWO:**

1. You will now make comparisons across multiple states: Iowa, California, Massachusetts, and Texas. [Use the third subset you created.] Create one graph showing side-by-side boxplots comparing home values for the four states. Each boxplot should be a different color. You can do this by mapping a color to each state. Remove the automatically generated legend by adding *show.legend = FALSE* in your *geom\_boxplot().*

Describe how home values between 1975 and the beginning of 2013 compare across the four states. Write at least three sentences.

1. Now compare structure costs for Iowa, California, Massachusetts, and Texas. Create one graph showing side-by-side boxplots, but this time do the following to make a more customized graph:
   * Make all boxplots the same color(s).
   * Specify both the color for the outline of the boxplot and the filled in color. They can be the same color, or you can use different ones for this.
   * Adjust the transparency of the fill color using: *alpha=0.5*.
   * Also add *notch=TRUE*

Describe how structure costs compare across the four states, and also compare this with home values from problem 9. Do you see the same pattern in the two graphs? Write at least four sentences.

1. Last, compare land values for Iowa, California, Massachusetts, and Texas by making a violin plot. This is very similar to a box plot, but you will use *geom\_violin().* You may look at the example and code provided at the R graph gallery (link on Bb), if you need help.

Describe how land values compare across the four states, and also compare this with home values and structure costs from problems 9 and 10. Do you see the same pattern in the three graphs? Write at least four sentences.

1. Let’s compare home values based on region. Using the full data set, create a plot showing a frequency polygon for each region: *geom\_freqpoly().* Use a bin width of 15,000$. No graph interpretation required.
2. To make the graph from problem 12 easier to interpret, we can fill in the colors on the graph. Do this by:
   * Replacing *geom\_freqpoly()* with *geom\_area()*

* + Mapping region to fill
  + And adding the argument: *stat="bin"*

No graph interpretation required.

1. You can see on your graphs from problems 12 and 13 that some regions are labeled as NA. Recreate your graph from problem 13 after removing the rows containing NA values. [Use your fourth subset.] Improve the look of the graph by specifying *color= “black”* which will outline the area for each region.

How do home values between 1975 and the beginning of 2013 compare for the different regions of the U.S.? Write at least 2 sentences.

1. Since there is so much overlap on the graph (see problems 13 and 14), a better choice to compare home values by region might be a *ridgeline* chart. Make this using the *ggridges* package. You may look at the example and code provided at the R graph gallery (link on Bb), if you need help. [We did look at this together in-class!]

Did you learn anything new about home values by looking at your ridgeline chart? Explain why this graph is easier to interpret. Write at least two sentences.

1. Focusing on just one time point, the first quarter of 2013, make a scatter plot that compares land value (x) and structure cost (y). [Use your fifth subset] Then find a way to also include the variables:
   * *Land.Share..Pct.* ~ Land share percent = proportion of the home value attributed to the value of the land.
   * *Home.Price.Index* ~ A home price index measures the price changes of residential housing as a percentage change from some specific reference date.

[I do not know the reference date for our data.]

For example, you can do this by mapping them to various aesthetics. Recall: We’ve used color, shape, size, and alpha. Your final graph should be easy to read and have appropriate, descriptive axis labels and title. No graph interpretation required.

1. Adjust your graph from problem 16 by:
   * Using the log of the land values: *x= log(Land.Value)*
   * Labeling the points with their corresponding state names. [I suggest using *geom\_text\_repel()* from the *ggrepel* library.]
   * Fitting a smooth line to the scatter plot. [Make sure the line is easy to see. You may need to change colors.]
   * And mapping the line type of the smooth line to region.

You may need to adjust your axis labels and/or graph title.

Provide a complete interpretation of the graph, describing how the five variables are (or are not) related to each other. Write at least 6 sentences.