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.]

HLValues<-read.csv(file='house-and-land-values.csv')

head(HLValues)

State region Date Home.Value Structure.Cost Land.Value Land.Share..Pct.

1 AK West 2010.25 224952 160599 64352 28.6

2 AK West 2010.50 225511 160252 65259 28.9

3 AK West 2009.75 225820 163791 62029 27.5

4 AK West 2010.00 224994 161787 63207 28.1

5 AK West 2008.00 234590 155400 79190 33.8

6 AK West 2008.25 233714 157458 76256 32.6

Home.Price.Index Land.Price.Index Year Qrtr

1 1.481 1.552 2010 1

2 1.484 1.576 2010 2

3 1.486 1.494 2009 3

4 1.481 1.524 2009 4

5 1.544 1.885 2007 4

6 1.538 1.817 2008 1

1. 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

HLValues.IA<-HLValues[(HLValues$State)=="IA",]

head(HLValues.IA)

State region Date Home.Value Structure.Cost Land.Value Land.Share..Pct.

1684 IA Midwest 1975.25 25285 21948 3337 13.2

1685 IA Midwest 1975.50 26335 22277 4058 15.4

1686 IA Midwest 1975.75 27304 22568 4736 17.3

1687 IA Midwest 1976.00 28210 22831 5379 19.1

1688 IA Midwest 1976.25 29080 23119 5961 20.5

1689 IA Midwest 1976.50 29957 23404 6553 21.9

Home.Price.Index Land.Price.Index Year Qrtr

1684 0.255 0 1975 1

1685 0.265 0 1975 2

1686 0.275 0 1975 3

1687 0.284 0 1975 4

1688 0.293 0 1976 1

1689 0.302 0 1976 2

tail(HLValues.IA)

State region Date Home.Value Structure.Cost Land.Value Land.Share..Pct.

1831 IA Midwest 2012.00 145969 137439 8530 5.8

1832 IA Midwest 2012.25 147404 137443 9961 6.8

1833 IA Midwest 2012.50 148820 137448 11372 7.6

1834 IA Midwest 2012.75 150213 137454 12759 8.5

1835 IA Midwest 2013.00 151571 137462 14109 9.3

1836 IA Midwest 2013.25 152897 137471 15427 10.1

Home.Price.Index Land.Price.Index Year Qrtr

1831 1.471 0.082 2011 4

1832 1.486 0.096 2012 1

1833 1.500 0.110 2012 2

1834 1.514 0.123 2012 3

1835 1.528 0.136 2012 4

1836 1.541 0.149 2013 1

* Subset #2: Just the data from California, starting at 1985.

ii<-(HLValues$State == "CA") & (HLValues$Date >= 1985)

HLValues.CA1985<-HLValues[ii,]

head(HLValues.CA1985)

State region Date Home.Value Structure.Cost Land.Value Land.Share..Pct.

613 CA West 2005.75 727221 158500 568721 78.2

614 CA West 2006.25 755981 163639 592343 78.4

615 CA West 2006.50 758233 166587 591646 78.0

616 CA West 2006.75 752963 169478 583485 77.5

617 CA West 2006.00 745594 161194 584400 78.4

620 CA West 1985.00 130005 62393 67612 52.0

Home.Price.Index Land.Price.Index Year Qrtr

613 2.423 3.189 2005 3

614 2.519 3.324 2006 1

615 2.526 3.323 2006 2

616 2.509 3.281 2006 3

617 2.484 3.279 2005 4

620 0.433 0.299 1984 4

tail(HLValues.CA1985)

State region Date Home.Value Structure.Cost Land.Value Land.Share..Pct.

728 CA West 2012.00 422384 215423 206961 49.0

729 CA West 2012.25 428610 215800 212810 49.7

730 CA West 2012.50 438735 216195 222540 50.7

731 CA West 2012.75 452037 216611 235426 52.1

732 CA West 2013.00 467720 217053 250667 53.6

733 CA West 2013.25 484516 217521 266995 55.1

Home.Price.Index Land.Price.Index Year Qrtr

728 1.407 1.195 2011 4

729 1.428 1.231 2012 1

730 1.462 1.290 2012 2

731 1.506 1.367 2012 3

732 1.558 1.458 2012 4

733 1.614 1.555 2013 1

* Subset #3: Data for just the following states: Iowa, California, Massachusetts, and Texas.

states<-c("IA","CA","MA","TX")

jj<-HLValues$State %in% states

HLValues.ICMT<-HLValues[jj,]

head(HLValues.ICMT)

State region Date Home.Value Structure.Cost Land.Value Land.Share..Pct.

613 CA West 2005.75 727221 158500 568721 78.2

614 CA West 2006.25 755981 163639 592343 78.4

615 CA West 2006.50 758233 166587 591646 78.0

616 CA West 2006.75 752963 169478 583485 77.5

617 CA West 2006.00 745594 161194 584400 78.4

618 CA West 1984.50 126847 60879 65968 52.0

Home.Price.Index Land.Price.Index Year Qrtr

613 2.423 3.189 2005 3

614 2.519 3.324 2006 1

615 2.526 3.323 2006 2

616 2.509 3.281 2006 3

617 2.484 3.279 2005 4

618 0.423 0.289 1984 2

tail(HLValues.ICMT)

State region Date Home.Value Structure.Cost Land.Value Land.Share..Pct.

6574 TX South 2012.00 164926 155156 9769 5.9

6575 TX South 2012.25 167094 155180 11914 7.1

6576 TX South 2012.50 169654 155213 14441 8.5

6577 TX South 2012.75 172431 155258 17173 10.0

6578 TX South 2013.00 175302 155314 19988 11.4

6579 TX South 2013.25 178180 155380 22800 12.8

Home.Price.Index Land.Price.Index Year Qrtr

6574 1.457 0.058 2011 4

6575 1.476 0.071 2012 1

6576 1.498 0.086 2012 2

6577 1.523 0.102 2012 3

6578 1.548 0.119 2012 4

6579 1.574 0.136 2013 1

* Subset #4: Just the data for the States that have a specified region, i.e. remove any rows that have *NA* values for region

remove.ii<-which(is.na(HLValues$region))

HLValues.noNA<-HLValues[-remove.ii,]

head(HLValues.noNA)

State region Date Home.Value Structure.Cost Land.Value Land.Share..Pct.

1 AK West 2010.25 224952 160599 64352 28.6

2 AK West 2010.50 225511 160252 65259 28.9

3 AK West 2009.75 225820 163791 62029 27.5

4 AK West 2010.00 224994 161787 63207 28.1

5 AK West 2008.00 234590 155400 79190 33.8

6 AK West 2008.25 233714 157458 76256 32.6

Home.Price.Index Land.Price.Index Year Qrtr

1 1.481 1.552 2010 1

2 1.484 1.576 2010 2

3 1.486 1.494 2009 3

4 1.481 1.524 2009 4

5 1.544 1.885 2007 4

6 1.538 1.817 2008 1

tail(HLValues.noNa)

State region Date Home.Value Structure.Cost Land.Value Land.Share..Pct.

7645 WY West 2012.00 243521 207681 35839 14.7

7646 WY West 2012.25 245650 207803 37846 15.4

7647 WY West 2012.50 248189 207936 40253 16.2

7648 WY West 2012.75 250696 208078 42618 17.0

7649 WY West 2013.00 252776 208228 44547 17.6

7650 WY West 2013.25 254408 208384 46024 18.1

Home.Price.Index Land.Price.Index Year Qrtr

7645 1.821 3.602 2011 4

7646 1.837 3.816 2012 1

7647 1.856 4.072 2012 2

7648 1.875 4.326 2012 3

7649 1.891 4.537 2012 4

7650 1.903 4.703 2013 1

* Subset #5: Data for just the first quarter of 2013 (this is the most recent data available in the data set)

HLValues.2013.25<-HLValues[(HLValues$Date == 2013.25),]

head(HLValues.2013.25)

State region Date Home.Value Structure.Cost Land.Value Land.Share..Pct.

40 AK West 2013.25 231891 181592 50299 21.7

304 AL South 2013.25 172864 146343 26521 15.3

428 AR South 2013.25 164279 140514 23765 14.5

580 AZ West 2013.25 238430 179759 58671 24.6

733 CA West 2013.25 484516 217521 266995 55.1

887 CO West 2013.25 297928 167566 130362 43.8

Home.Price.Index Land.Price.Index Year Qrtr

40 1.526 1.248 2013 1

304 1.547 1.054 2013 1

428 1.701 1.661 2013 1

580 1.482 1.000 2013 1

733 1.614 1.555 2013 1

887 1.411 1.221 2013 1

tail(HLValues.2013.25)

State region Date Home.Value Structure.Cost Land.Value Land.Share..Pct.

7038 VT N. East 2013.25 234747 220074 14673 6.3

7191 WA West 2013.25 338871 270429 68442 20.2

7344 WI Midwest 2013.25 184707 172796 11912 6.4

7497 WV South 2013.25 151494 134263 17232 11.4

7650 WY West 2013.25 254408 208384 46024 18.1

7773 DC <NA> 2013.25 757154 172015 585140 77.3

Home.Price.Index Land.Price.Index Year Qrtr

7038 1.604 0.781 2013 1

7191 1.582 1.332 2013 1

7344 1.380 0.029 2013 1

7497 1.705 1.769 2013 1

7650 1.903 4.703 2013 1

7773 2.893 3.505 2013 1

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.

ggplot(data=HLValues.IA)+

geom\_histogram(mapping=aes(x=Home.Value),color='green',binwidth = 50000)

Chart, histogram

Description automatically generated

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

I have learned that over 80 homes were valued in the range 0 to 50,000 dollars from the data set. I also learned that there are less than 40 homes valued 50,001 to 100,000 dollars and there are less than 35 homes valued greater than 100,000 dollars. From these points I have also learned that there are more homes that cost less than 50,000 than ones that cost more during the time frame 1975 to 2013.

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.

ggplot(data = HLValues.IA)+

geom\_point(mapping=aes(x=Date,y=Home.Value),color='green')+

labs(x="Date (in years)",y="Home Value (in US Dollars)",

title="Home Value in Iowa from 1975 to 2013")

Chart, line chart

Description automatically generated

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

I have learned that home values in Iowa have increased significantly from 1975 to 2013. I have also learned that from roughly 2007 to 2012 there was very little growth in home values. I also learned that from roughly 1980 to 1982 there was a decrease in home value.

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.

ggplot(data = HLValues.IA)+

geom\_point(mapping=aes(x=Date,y=Home.Value),color='green')+

geom\_point(mapping=aes(x=Date,y=Land.Value),color='blue')+

labs(x="Date (in years)",y="Home and Land Values (in US Dollars)",

title="Home and Land Values in Iowa from 1975 to 2013")

Chart, line chart, histogram

Description automatically generated

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.

I’ve learned that land values have remained relatively the same from 1975 to 2013. Though, I can also see how much they increased from 1975 to 1980 before returning to a somewhat consistent value. Land value also looks like it began to quickly rise during 2012 and continued to do so to 2013.

When comparing to home values, the growth in land values is very low. Though both do increase over the time span. Also, both follow similar trends from 1975 to 1982. They both increase from 1975 to 1980 then decrease from 1980 to 1982

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*

Graph.Colors<-c('green','blue')

* + 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!

ggplot(data = HLValues.IA)+

geom\_point(mapping=aes(x=Date,y=Home.Value,color='Home.Value'))+

geom\_point(mapping=aes(x=Date,y=Land.Value),color='blue')+

labs(x="Date (in years)",y="Home and Land Values (in US Dollars)",

title="Home and Land Values in Iowa from 1975 to 2013")

* + Do the same for the land value layer: put *color= “Land.Value”* inside the *aes()* function.

ggplot(data = HLValues.IA)+

geom\_point(mapping=aes(x=Date,y=Home.Value,color='Home.Value'))+

geom\_point(mapping=aes(x=Date,y=Land.Value,color='Land.Value'))+

labs(x="Date (in years)",y="Home and Land Values (in US Dollars)",

title="Home and Land Values in Iowa from 1975 to 2013")

* + Add *color= “Legend”* to the *labs()* layer.

ggplot(data = HLValues.IA)+

geom\_point(mapping=aes(x=Date,y=Home.Value,color='Home.Value'))+

geom\_point(mapping=aes(x=Date,y=Land.Value,color='Land.Value'))+

labs(x="Date (in years)",y="Home and Land Values (in US Dollars)",

title="Home and Land Values in Iowa from 1975 to 2013", color="legend")

* + Add a layer to the graph that allows you to manually select the colors that you use: *scale\_color\_manual(values = Graph.Colors)*

ggplot(data = HLValues.IA)+

geom\_point(mapping=aes(x=Date,y=Home.Value,color='Home.Value'))+

geom\_point(mapping=aes(x=Date,y=Land.Value,color='Land.Value'))+

labs(x="Date (in years)",y="Home and Land Values (in US Dollars)",

title="Home and Land Values in Iowa from 1975 to 2013", color="Legend")+

scale\_color\_manual(values=Graph.Colors)

Chart, line chart

Description automatically generated

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.

Graph.Colors2<-c('green','blue','red')

ggplot(data = HLValues.IA)+

geom\_point(mapping=aes(x=Date,y=Home.Value,color='Home.Value'))+

geom\_point(mapping=aes(x=Date,y=Land.Value,color='Land.Value'))+

geom\_point(mapping=aes(x=Date,y=Structure.Cost,color='Structure.Cost'))+

labs(x="Date (in years)",y="Structure Cost, Home, and Land Values (in US Dollars)",

title="Structure Cost, Home, and Land Values in Iowa from 1975 to 2013", color="Legend")+

scale\_color\_manual(values=Graph.Colors2)

Chart, histogram

Description automatically generated

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.]

ggplot(data = HLValues.CA1985)+

geom\_point(mapping=aes(x=Date,y=Home.Value,color='Home.Value'))+

geom\_point(mapping=aes(x=Date,y=Land.Value,color='Land.Value'))+

geom\_point(mapping=aes(x=Date,y=Structure.Cost,color='Structure.Cost'))+

labs(x="Date (in years)",y="Structure Cost, Home, and Land Values (in US Dollars)",

title="Structure Cost, Home, and Land Values in California from 1985 to 2013", color="Legend")+

scale\_color\_manual(values=Graph.Colors2)

Chart, histogram

Description automatically generated

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

Dr. Davis’s statement is exemplified by the graph because of the way home and land values increase and decrease at the same time. In comparison, when you compare structure cost with home cost the relationship is not obvious. For example, the structure cost rises pretty steady for most of the period while the home and land values rise and fall with each other. The biggest example of their relationship is from 2000 to 2010. Both rapidly increase until roughly 2007 before decreasing just as fast.

**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.