Post 1

Zhiyuan He

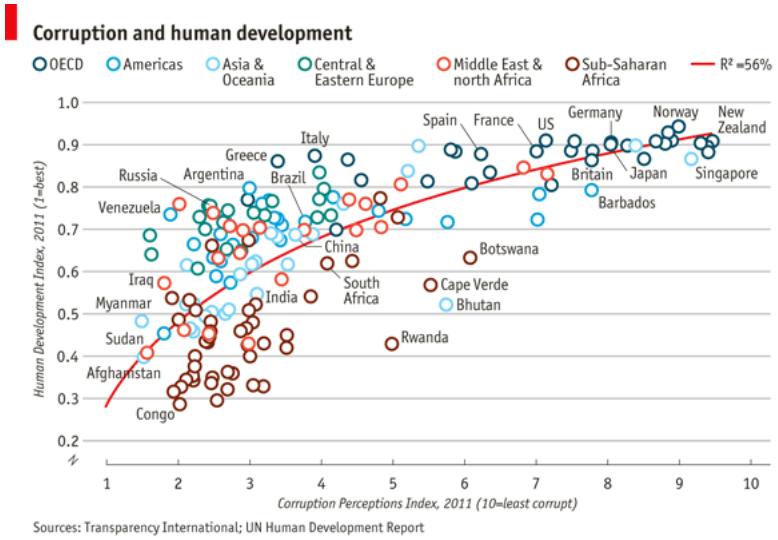
October 28, 2017

Post 1 visualizing data with ggplot

Introduction

When deciding on a topic to write about for this post I wanted to choose a topic that I thought was very important and relevant in the future and I was relatively weak on and did not understand well. So I decided that I did not know ggplot very well and thus that became my motivation to focus the lab on ggplot. ggplot is a powerful visualization tool that can be used in R. It is a plotting system that can produce graphics in R.

Here is a sample of a graphic you can make



Basics of ggplot

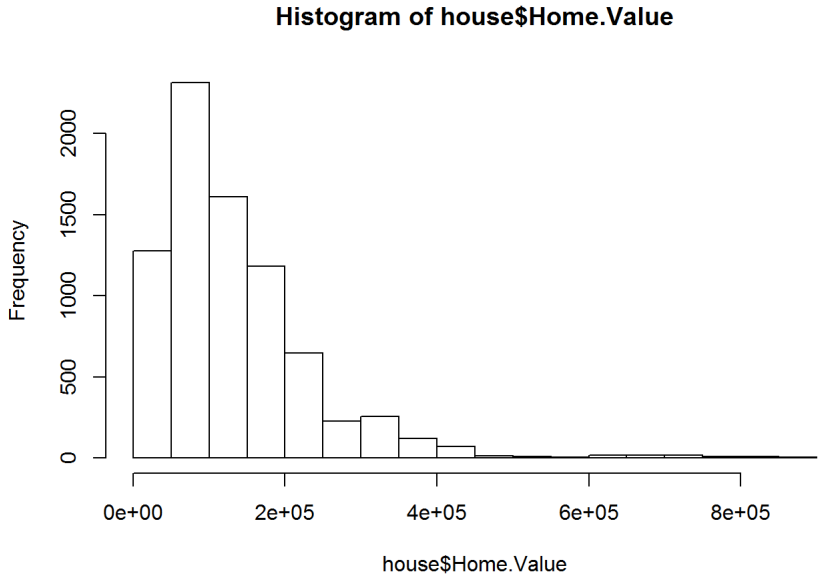
ggplot is very flexible and can have a pretty high degree of flexibility compared to the basic graphics that can be produced by R. However, ggplot can not create interactive graphics and three dimensional graphics.

Using ggplot

ggplot is a package in and to install one needs to use the code install.packages("ggplot2") and library(ggplot2) ggplot is used like building block you can add things to your graph by simply adding them such as the scales, axis, faceting, and most importantly the data.

Let's take a look at a sample graphic using only R's function

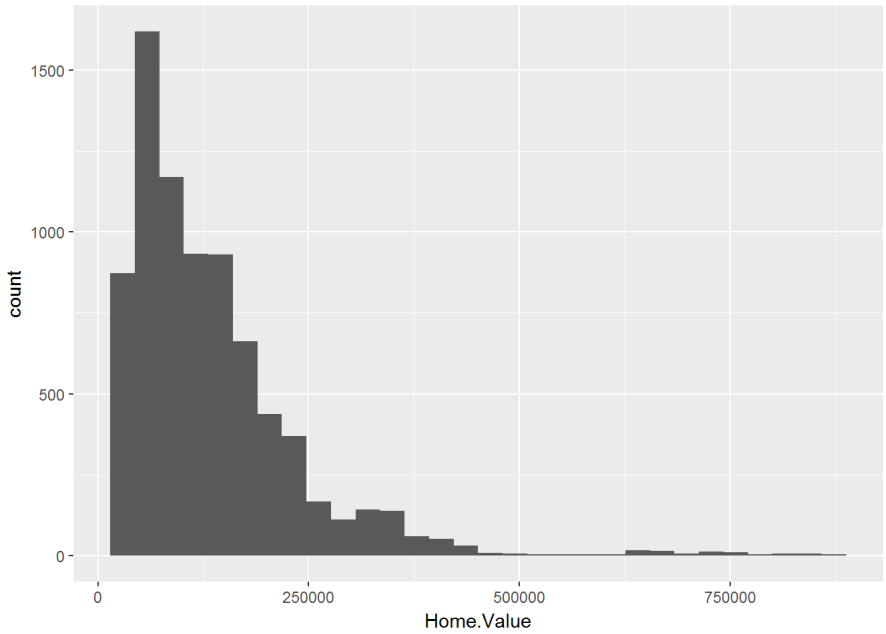
```
house <- read.csv("C:\\Users\\iiyua\\Documents\\stat133\\stat133-hws-fall17\\Post 1\\Rgraphics\\Rgraphics\\dataSets\\landdata-states.csv") #code to load the data house
hist(house$Home.Value)#to make histogram
```



Looks very basic and simple but using ggplot and not even adding much code it can look slightly better

```
library(ggplot2) #loading ggplot2
ggplot(house, aes(x= Home.Value)) + geom_histogram() #Using ggplot to make histogram
```

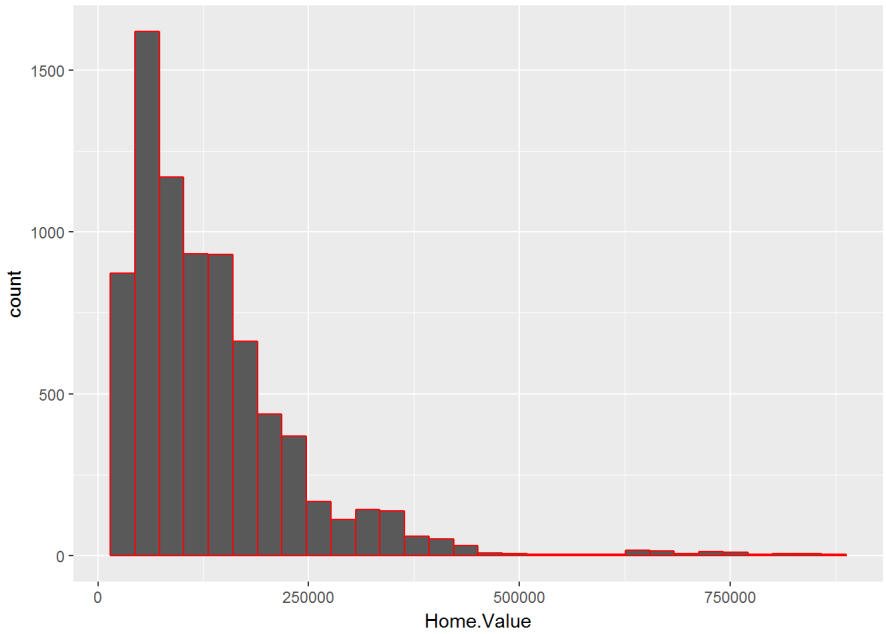
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



As you can see geom hist produces a bar chart as well but it looks better with axis and is shaded in and the number values are more readable for the reader. With a few adjustments we can add color

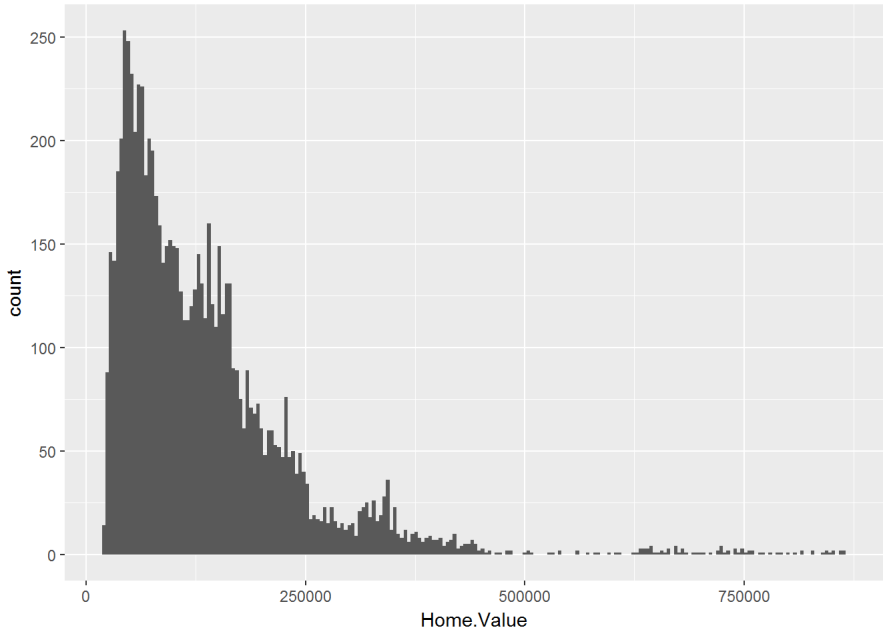
```
ggplot(house, aes(x= Home.Value)) + geom_histogram(color="red") #adding color
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Looking at the red highlighted part we can see that there are small points of data that is above 500,000 in home value. We can clarify this by changing the scale of of the graph

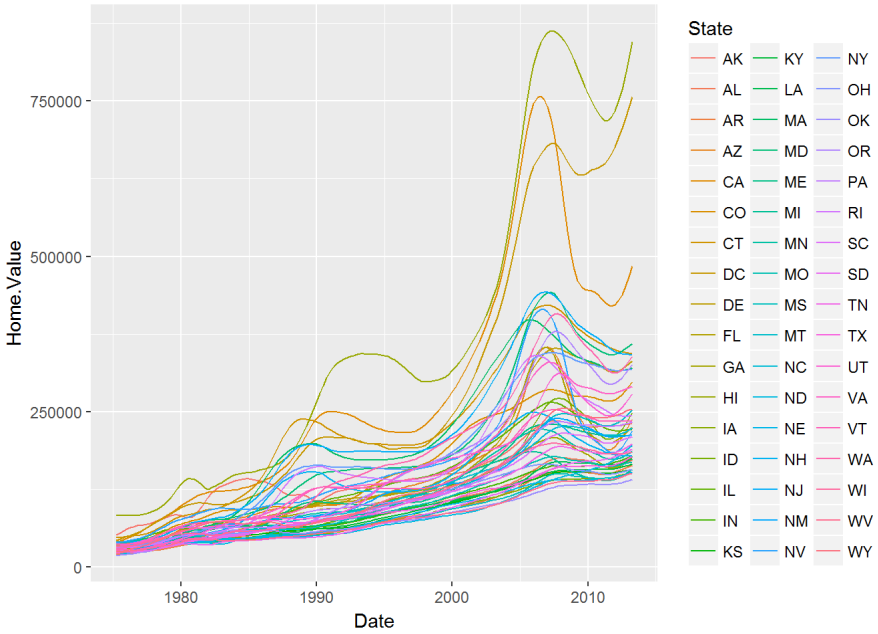
```
ggplot(house, aes(x= Home.Value)) + geom_histogram(binwidth = 4000) #changing the width of each bar
```



Changing the width of each bar gives a more detailed graph that is skinnier and show the values of each home value more accurately. The values above 50000 is more visible than before and people can tell that it is not zero at those values.

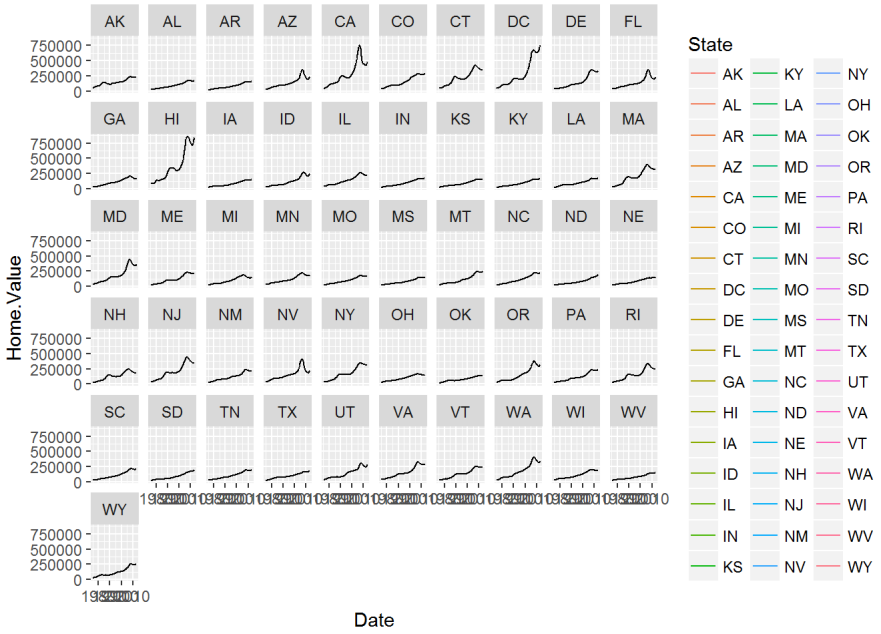
ggplot also has the ability to create a line graph so that you can notice trends. Lets take a look at the price of housing for each state over the past years. In doing so we can define the color of each line to be each state.

```
value_state <- ggplot(house, aes(x=Date, y=Home.Value)) + geom_line(aes(color=State)) #creating line graph looking at home value with different color by state
value_state
```



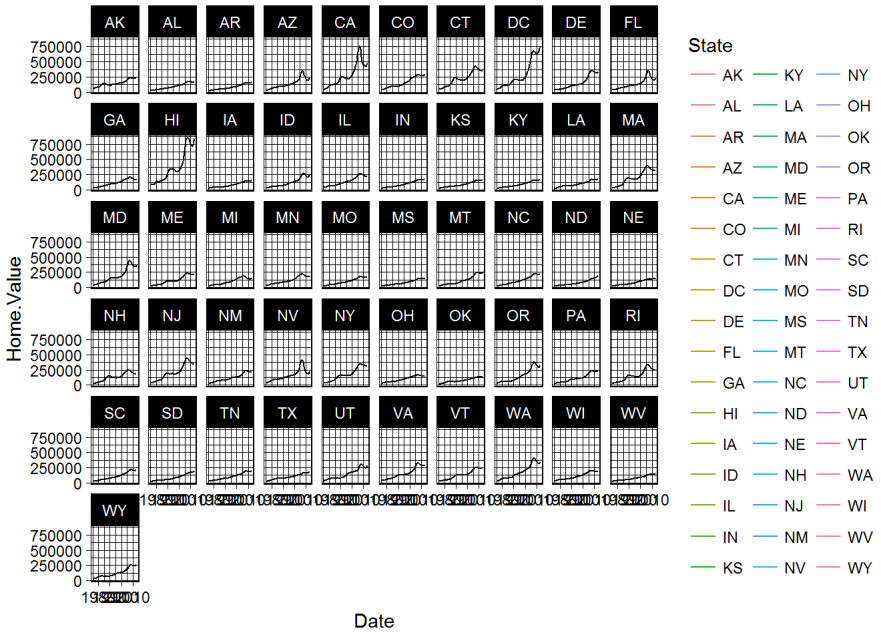
Looking at this graph one can not tell too much info because there are so many states. The information is simply too crowded for anyone to get information out of. This is where an another great tool from ggplot can be used called faceting. faceting allows you to seperate the info in the graph by a certain variable. In this case it would be best to facet by each state.

```
faceted <- value_state + geom_line()+facet_wrap(~State, ncol = 10) #defining what parameter to facet and to seperate by state
faceted
```

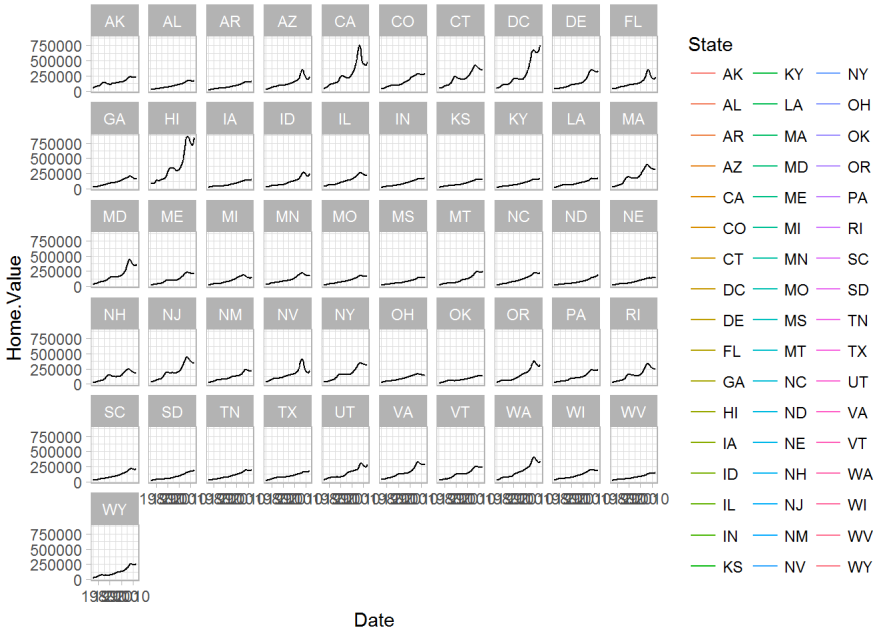


The data can be interperted more easily in this format. As it breaks the graph down for each state. One thing that we did not learn in lecture is themes. The key things that themes change are *axis labels* plot background *facet label background* legend appearance Let's take a look at what themes will do to our previous example

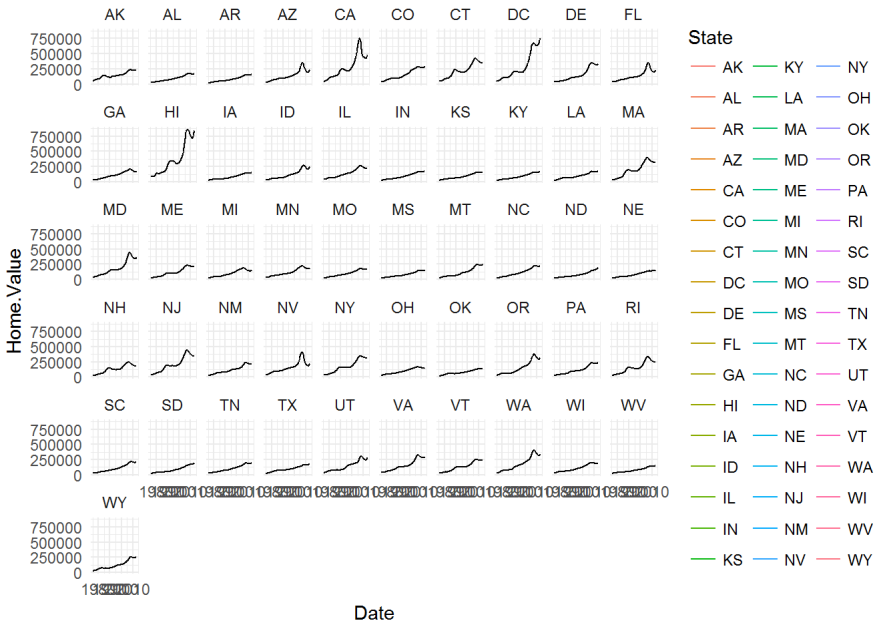
```
faceted + theme_linedraw() #adding themes to graphs
```




```
faceted + theme_light()
```



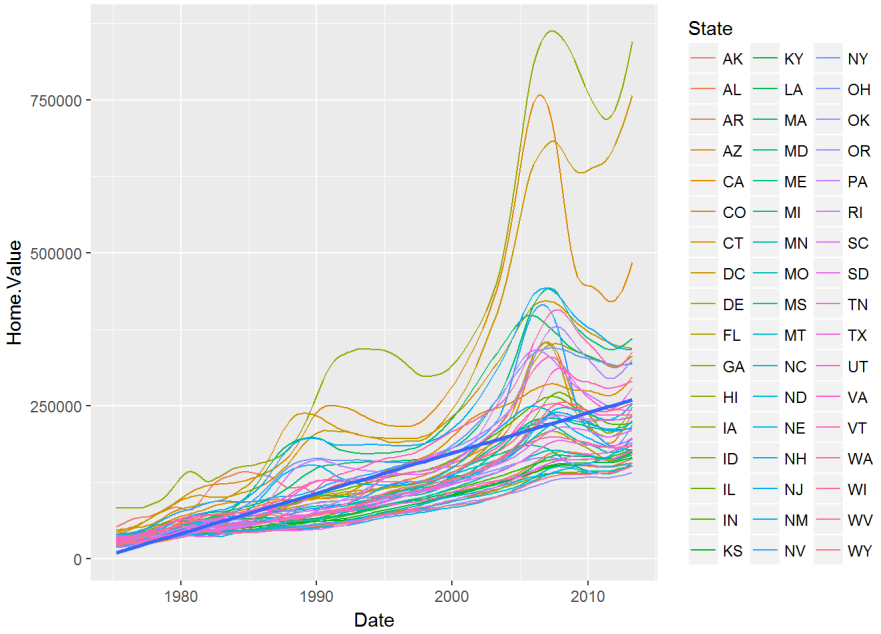

```
faceted + theme_minimal()
```



These are just three themes that can be done to graphs.

Let's take a look at the home value for each state trend graph again. It would be best if we could find a linear best fit line to go through the points and ggplot allows us to do such thing!

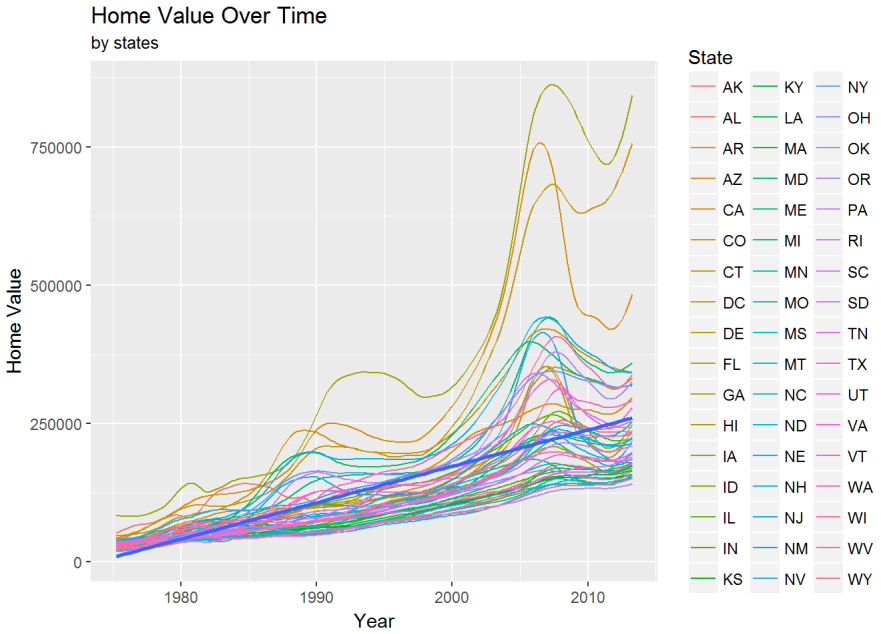
```
library(ggplot2)
g = value_state + geom_smooth(method = "lm") #adding a linear best fit line
g
```



As you can see the blue line is a linear best fit line that goes through what is closest to all the points. This function can be useful in showing the best fit.

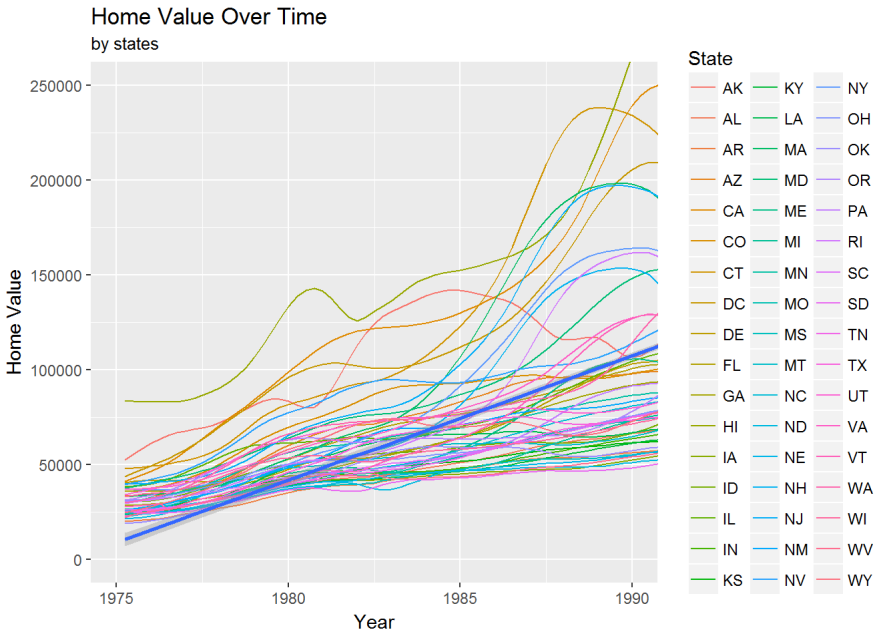
Another useful thing to add to graphs would be titles and axis lets look at how one would add a title. To add a title you can use the function ggtitile and can also include a subtitle in the category. To add axis you can use xlab and ylab

```
g + ggtitle("Home Value Over Time", subtitle = "by states") + xlab("Year") +ylab("Home Value") #adding titles and axis
```



Another feature that can be useful that we did not learn in class is how to adjust the x and y axis limits. The best way to do so without deleting values is to essentially zoom in the graph. Let's zoom in to the time before 1990 where the home values changed drastically for some states

```
g + ggtitle("Home Value Over Time", subtitle = "by states") + xlab("Year") +ylab("Home Value") + coord_cartesian(xlim = c(1975,1990), ylim = c(0,250000)) #Changing scale of paramters
```



Limiting the scale of the axis gives us a better view at what began to happen as the housing prices for different states began to vary drastically.

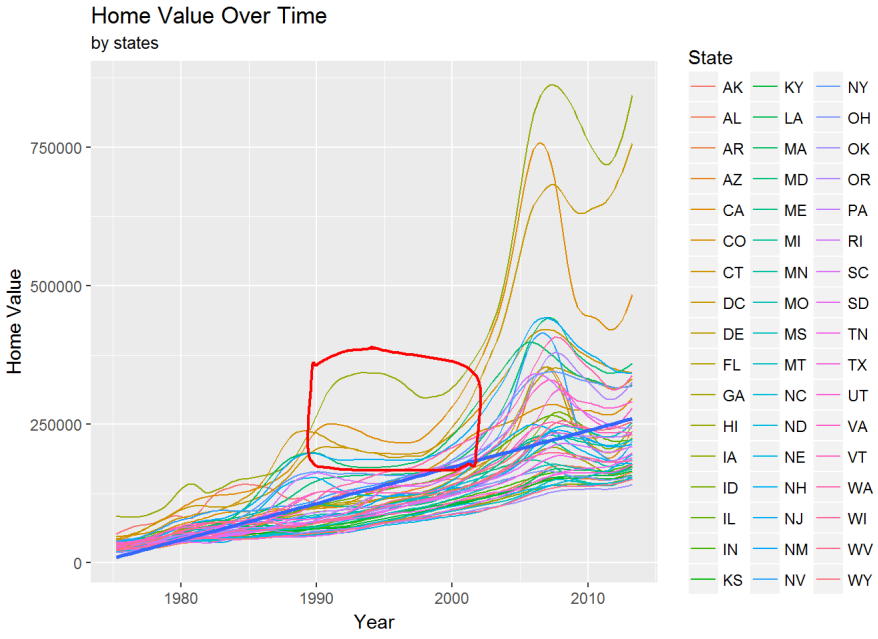
Insteading of zooming in to point out a specific part of the graph you might want to circle or draw out part of a graph but to do so we will need to install a new package called ggalt
Let us try to highlight the point where a few states' home value really jumped off

```
#install.packages("ggalt")
library(ggalt)
```

```
## Warning: package 'ggalt' was built under R version 3.4.2
```

```
circle = house[house$Year> 1990 & house$Year <2000 & house$Home.Value>200000 & house$Home.Value<400000,] #Installi
ng package and defining what the circle will enclose

g + ggtitle("Home Value Over Time", subtitle = "by states") + xlab("Year") +ylab("Home Value")+ geom_encircle(aes(
x=Date, y= Home.Value), data = circle, color="red", size=2)
```




```
#adding the circle to the graph
```

As you can see using this new package we can show the point where some states started to deviate from the normal part and this is something new that we did not in class and could be useful in the future

Conclusion

In conclusion ggplot is a very powerful graphical tool that conveys statistical data that one may want to convey to their audience. The general idea of using ggplot again us to treat it as builing blacks that you can keep adding info and data to as if a building block. To add more paramaters and options one can just add parts using + and whatever one wishes to add.

References used

http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html *http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/* https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf *https://opr.princeton.edu/workshops/Downloads/2015Jan_ggplot2Koffman.pdf* https://github.com/hrbrmstr/ggalt *https://www.r-bloggers.com/installing-r-packages/