Grammer of graphics-ggplot2

Swapnil Awasthi

10 November 2017

## Why ggplot2?

Very robust, useful and flexible for descriptive statistics. Advantages of ggplot2:

1-consistent, plot specification at a high level of abstraction.

2-very flexible.

3-theme system for polishing plot appearance.

4-mature and complete graphics system.

5-Great community support and active mailing list.

NOTE: Data for this exercise can be downloaded from-<http://tutorials.iq.harvard.edu/R/Rgraphics.zip>

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.3

library(ggrepel)

## Warning: package 'ggrepel' was built under R version 3.3.3

housing <- read.csv('C:/Users/awasthi/Desktop/landdata-states.csv', header = TRUE, sep = ",",na.strings = c("NA", "#N/A","" ))  
#checking null  
colSums(is.na(housing))

## State region Date Home.Value   
## 0 153 0 0   
## Structure.Cost Land.Value Land.Share..Pct. Home.Price.Index   
## 0 0 0 0   
## Land.Price.Index Year Qrtr   
## 0 0 0

#saving nulls in the df  
nullhousing <- subset(housing, is.na(housing$region))  
head(nullhousing)

## State region Date Home.Value Structure.Cost Land.Value  
## 7651 DC <NA> 2003.00 384443 93922 290522  
## 7652 DC <NA> 2003.25 399633 93961 305673  
## 7653 DC <NA> 2003.50 417110 94032 323078  
## 7654 DC <NA> 2003.75 436496 94486 342010  
## 7655 DC <NA> 2004.00 457806 95807 361999  
## 7656 DC <NA> 2004.25 481171 98379 382792  
## Land.Share..Pct. Home.Price.Index Land.Price.Index Year Qrtr  
## 7651 75.6 1.469 1.654 2002 4  
## 7652 76.5 1.527 1.740 2003 1  
## 7653 77.5 1.594 1.839 2003 2  
## 7654 78.4 1.668 1.947 2003 3  
## 7655 79.1 1.749 2.062 2003 4  
## 7656 79.6 1.839 2.182 2004 1

#checking the State for which region is null  
unique(nullhousing$State)

## [1] DC  
## 51 Levels: AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA ... WY

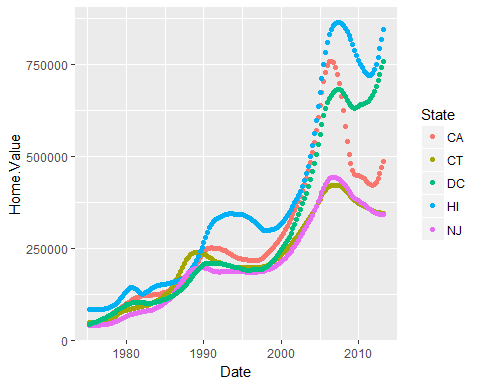
#all the NAs are from State = DC  
subset(housing, housing$State == 'DC')[,2]

## [1] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>  
## [15] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>  
## [29] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>  
## [43] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>  
## [57] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>  
## [71] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>  
## [85] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>  
## [99] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>  
## [113] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>  
## [127] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>  
## [141] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>  
## Levels: Midwest N. East South West

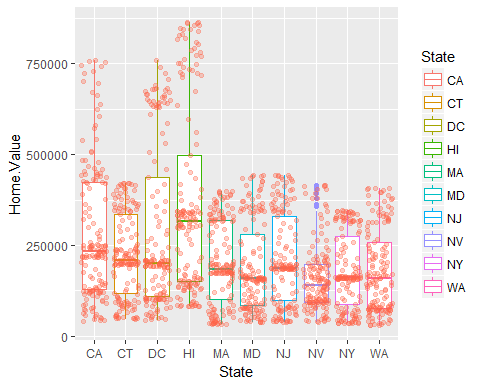
#setting DC to south region as per https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us\_regdiv.pdf  
housing[housing$State=='DC',2] <- 'South'  
#aggregating the data based home value - by state  
Hvalue\_by\_state <- aggregate(Home.Value ~ State, housing, FUN = sum)  
#creating a sorted dataset for further use  
sortedd <- Hvalue\_by\_state[order(Hvalue\_by\_state$Home.Value, decreasing = TRUE),]  
#top 5 states with highest home value  
top5 <- as.list(sortedd[1:5,1])  
#top 10 states with highest home value  
top10states <- as.list(sortedd[1:10,1])

## Including Plots

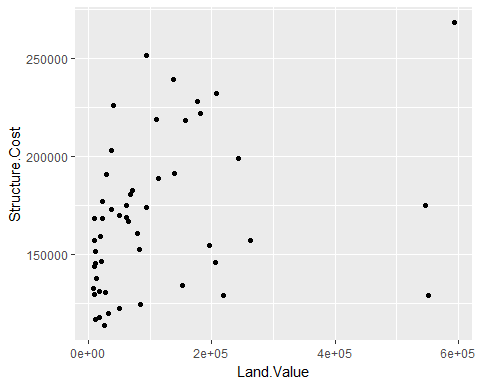
#scatter plot  
ggplot(subset(housing, State %in% (unlist(top5))), aes(x=Date, y=Home.Value, color=State))+geom\_point()



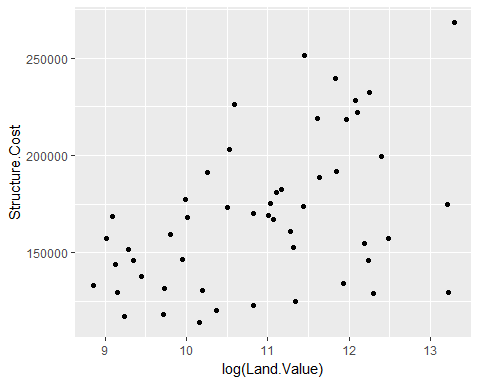
#box plot for top 10 State vs HomeValue #Box plots provide graphic display of five-number summary  
ggplot(subset(housing, State %in% (unlist(top10states))), aes(x=State, y=Home.Value, color=State))+geom\_boxplot()+geom\_jitter(alpha=0.3, color='tomato')



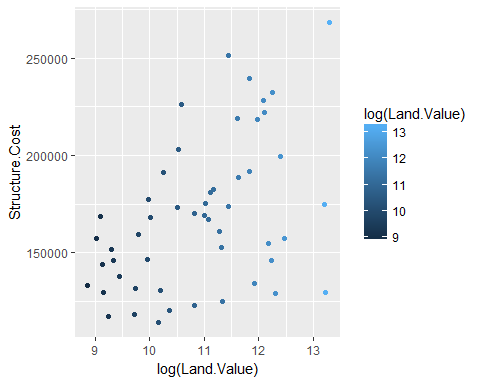
#creating a dataframe based on a date/time-frame  
hp.2007Q1 <- subset(housing, Date == '2007.25')  
  
#let us take a closer look on scatter plot for 2007 1st Quarter - Land value vs Structure Cost  
ggplot(hp.2007Q1, aes(x= Land.Value, y = Structure.Cost))+geom\_point()



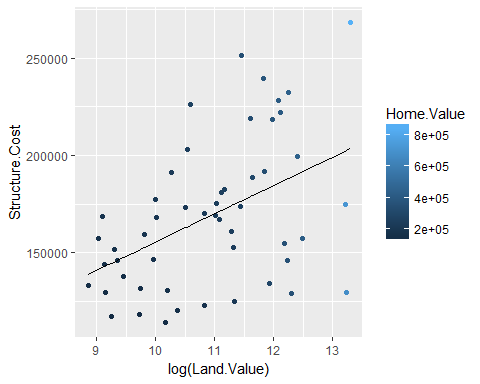
#since our x axis has a high variance - lets try to fit against a log transformation of Land Value  
ggplot(hp.2007Q1, aes(x= log(Land.Value), y = Structure.Cost))+geom\_point()



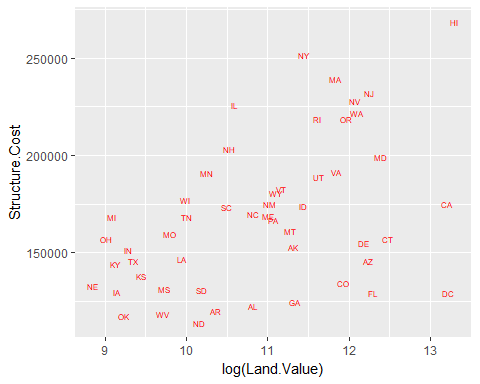
#color coding  
ggplot(hp.2007Q1, aes(x= log(Land.Value), y = Structure.Cost, color = log(Land.Value)))+geom\_point()



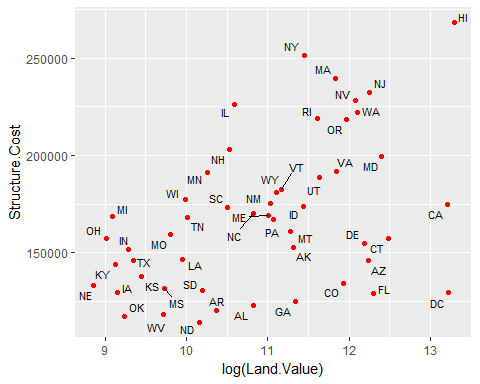
#Let us try to fit a line based on our X and Y and add it to your plot  
hp.2007Q1$pred.SC <- predict(lm(Structure.Cost ~ log(Land.Value), data = hp.2007Q1))  
p1 <- ggplot(hp.2007Q1, aes(x=log(Land.Value), y = Structure.Cost ))  
#adding points and line  
p1 + geom\_point(aes(color = Home.Value))+geom\_line(aes(y=pred.SC))



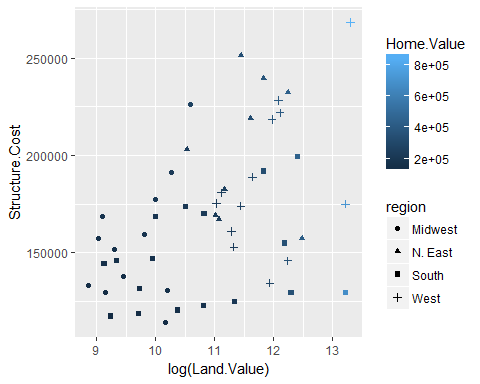
#mapping states on the plot  
p1 + geom\_text(aes(label = State), size = 2, color = 'red')



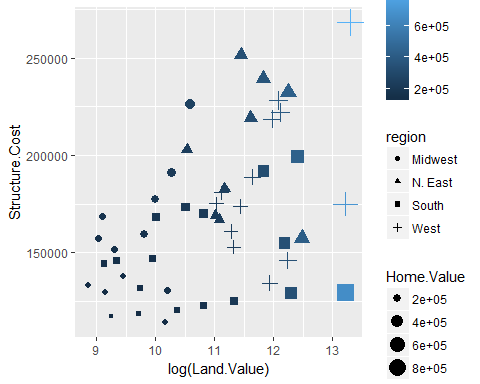
#install.packages("ggrepel") #library("ggrepel") required!!  
#this is to map points with text  
p1 + geom\_point(color='red') + geom\_text\_repel(aes(label=State), size = 3)



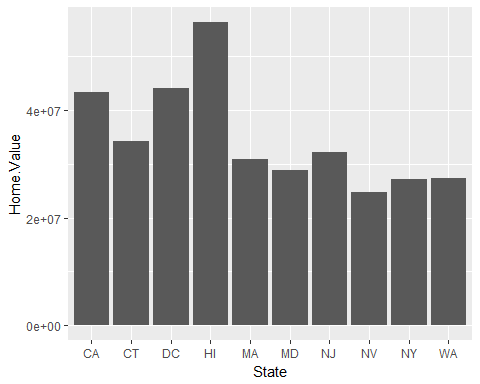
#mapping variable to other aesthetics  
p1 + geom\_point(aes(color=Home.Value, shape = region))



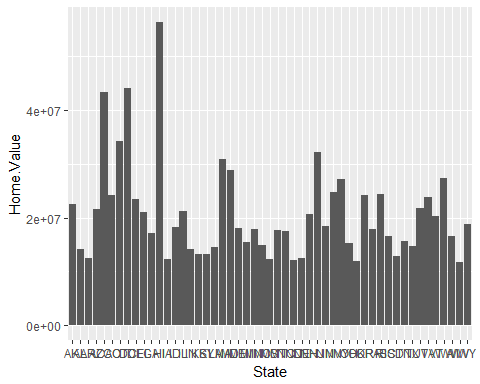
#mapping variable to other aesthetics, adding one more dimension(size) to our plot  
#here size is based on the home value - can be seen in the legends  
p1 + geom\_point(aes(size=Home.Value, shape = region, color = Home.Value))



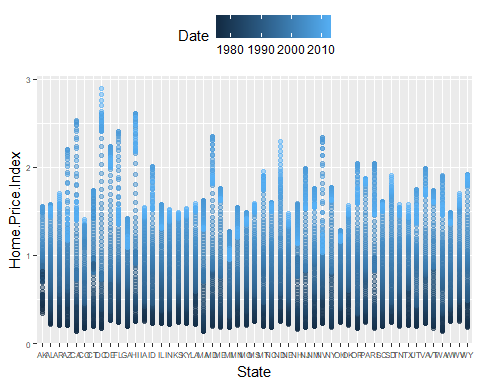
#Statistical transformations  
p2 <- ggplot(housing, aes(x=Home.Value))+geom\_histogram(stat = 'bin', binwidth = 4000)  
#top10states  
top10 <- (sortedd[1:10,])  
ggplot(top10, aes(x=State, y=Home.Value)) + geom\_bar(stat = 'identity')



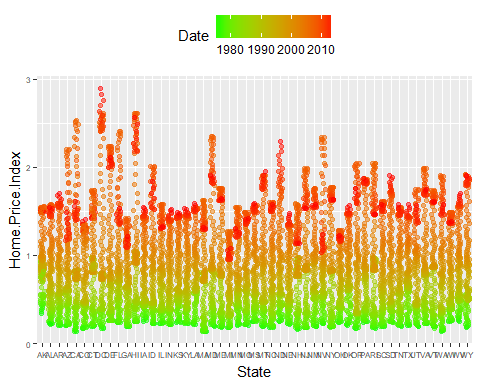
#all states included  
ggplot(Hvalue\_by\_state, aes(x=State, y=Home.Value)) + geom\_bar(stat = 'identity')



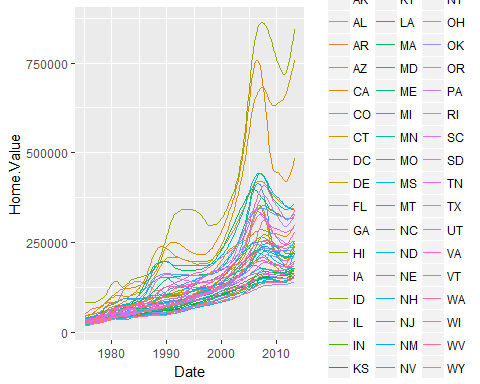
#Home value by Date and State  
p3 <- ggplot(housing, aes(x=State, y=Home.Price.Index)) + theme(legend.position = 'top', axis.text = element\_text(size = 6))  
p3+geom\_point(aes(color=Date), alpha=0.5,size=1.5)



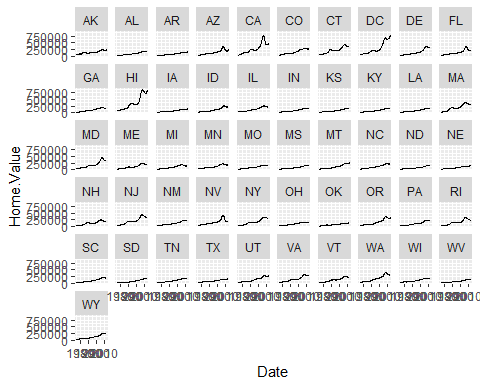
#To avoid overlapping - we can use jitter  
p4 <- p3+geom\_point(aes(color=Date), alpha=0.5,size=1.5, position = position\_jitter(width = 0.25, height = 0))  
#color coding to indicate HIGHs and LOWs  
p4+scale\_color\_continuous(low = 'green', high = 'red')



#Facets  
p5 <- ggplot(housing, aes(x=Date, y=Home.Value))  
p5 + geom\_line(aes(color=State)) #Difficult to distinguish??? Facets can help



#Facets can help here #These can be used to see the correlation as well!  
p5 + geom\_line()+facet\_wrap(~State, ncol = 10)



End of document  
 Feedback/Suggestions are welcome, Thank you!