Resources
Creating Graphical Representations of Data

Applied Data Science

Introduction

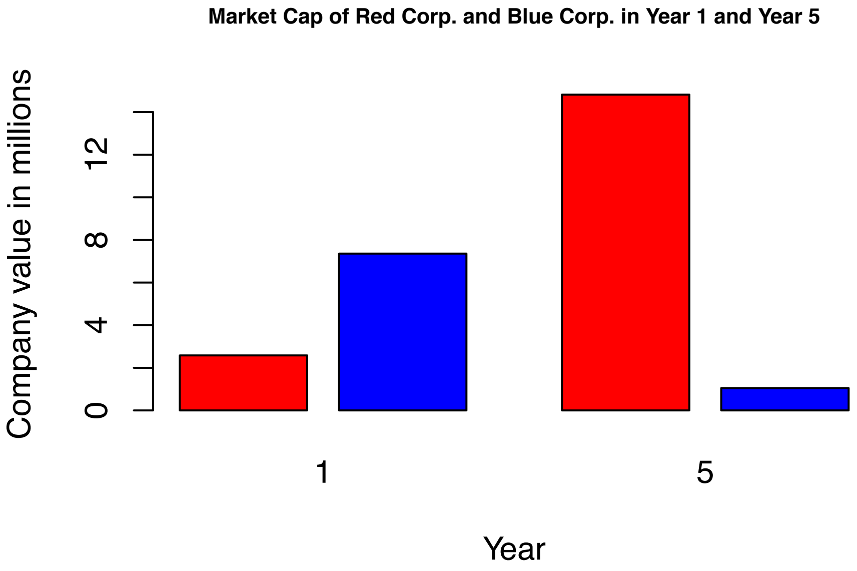
When creating a visualization, it is critical that you match the type of data you have and the question you are answering to the visualization. Mismatches between these factors confuses the point you are trying to make at best and at worst can mislead your audience. Of all the forms of visualization available to you, understanding the bar graph, the line graph and the scatter plot are most important.

Note: the charts in this document were created using R. R is not required for this course, but the code to create these charts is provided in the event you're curious.

Bar graph

Bar graphs are used to display differences between categories or type. These charts are excellent at showing differences in value or average and and answering questions like "how much?" Bar graphs should also be used to demonstrate proportion (as opposed to pie charts, which are misleading) and the stacked bar graph can be used to demonstrate proportion between groups.

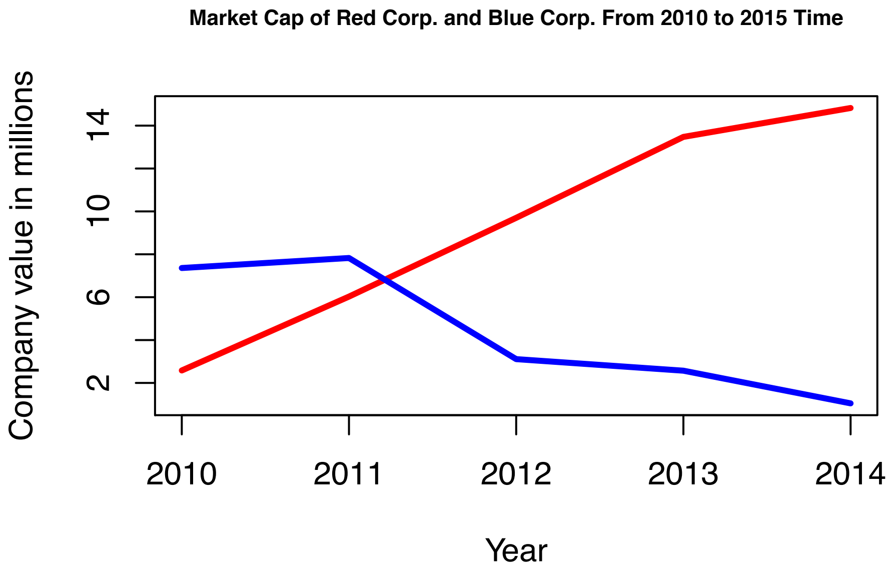
```r
barplot(t(as.matrix(yrs1and5)), space=c(.25,.75), col=c("red","blue"),beside=T,
        ylab="Company value in millions",xlab="Year")
title(main=list("Market Cap of Red Corp. and Blue Corp. in Year 1 and Year 5",cex=.65))
```



Line graph

Line graphs are used to display two changing continuous variables (especially something and time). Lines graphs are excellent at demonstrating the trends between variables by category (by placing multiple lines on the same chart).

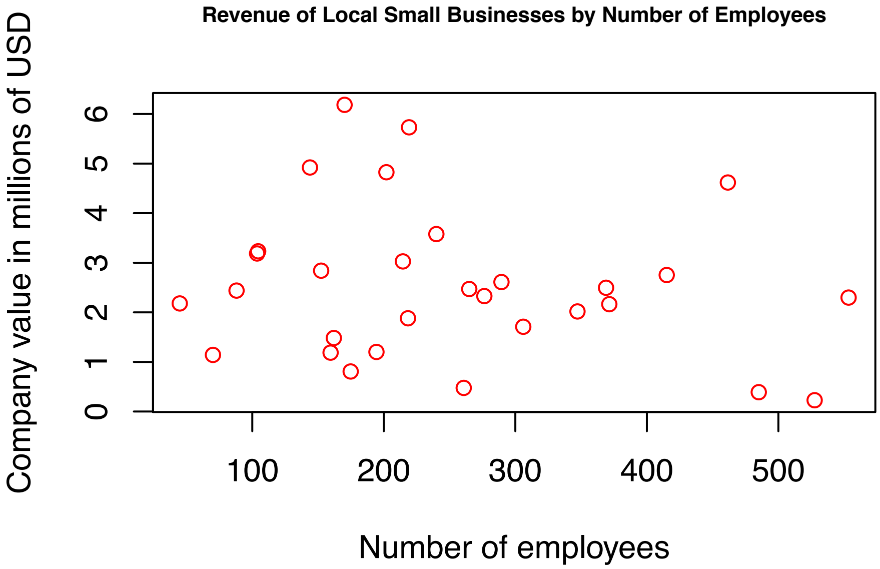
```
plot(years,a,type='l',ylim=c(min(a,b),max(a,b)), col='red',lwd=3,
     ylab="Company value in millions",xlab="Year")
points(years,b,type='l',col='blue',lwd=3)
title(main=list("Market Cap of Red Corp. and Blue Corp. From 2010 to 2015 Time",cex=.65))
```



Scatter plot

Scatter plots are used to display the trend between two variables across (many) distinct observations. These plots may be further combined with size changes to represent the relationship between 3 variables (x, y and size).

```
plot(employees,revenue,type='p',
     col='red',
     ylab="Company value in millions of USD",xlab="Number of employees")
points(years,b,type='l',col='blue',lwd=3)
title(main=list("Revenue of Local Small Businesses by Number of Employees",cex=.65))
```



Recommended reading

This document was an introduction to the most common types of charts. If you only ever used these three types of charts, you can effectively communicate any point worth making; however, more types of charts are available. Tableau and Hubspot both offer good explanations of a wider array of chart types. You can find their resources here (Tableau) and here (Hubspot). Tableau also has an excellent white paper on Data Vizualization best practices.

For *what not to do* with data vizualization, a good resource is Viz.wtf. Their Twitter page is also a good resource; folks regularly post to discuss issues with data visualizations.

For R, Quick-R has two sections Graphs and Advanced Graphs, both of which provide code and examples of common charts. Though not required for this course, R is widely considered the gold standard of data visualization, and is common both in industry (e.g., NY Times) and in academia. Tutorials for for data vizualization in R can be found at FlowingData.com.