Week 4 blackboard

http://www.forbes.com/sites/joshbersin/2013/10/07/big-data-in-human-resources-a-world-of-haves-and-have-nots/

This article discusses how some companies are using data analytics to improve operational performances. It discusses how companies that are investing in data analytics are reaping benefits in market returns and are 4 times more likely to be respected in their industry. Their transformation was to use their research technology for 2 years on 480 large organizations.

In one example, they have found that by using analytics, top performing companies can retain top performers by giving these employees higher raises. They infer that retaining top performers is a key to a top performing company. Without any additional information the article states that top performers out deliver mid-level performers by a wide margins. This statement is unreliable since a wide margin means different things to different people. If I were making this argument I would back it up with real numbers, specific performance metrics in different industries and compare top level performers to mid-level performers. This data will help justify the evidence and counter this specific rebuttal.

Week 6

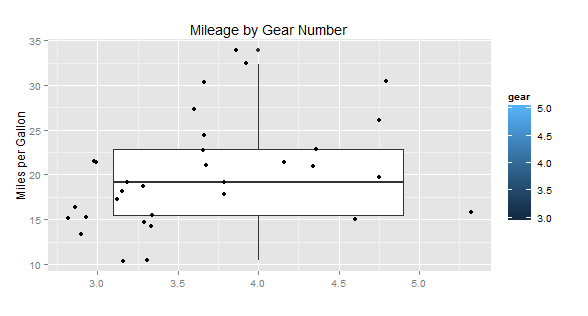
With many great options for graphs in R, having one simple plotting function for testing will be helpful in saving time to review data quickly. Qplot is an alternative ggplot when you just want to run through the command line. Both are part of the GGPLOT2 package. One difference is GGPlot uses data.frame and qplot uses vector, so you must convert to the a vector data type to run the code.

Reference: <http://www.statmethods.net/advgraphs/ggplot2.html>

qplot(gear, mpg, data=mtcars, geom=c("boxplot", "jitter"),

+ fill=gear, main="Mileage by Gear Number",

+ xlab="", ylab="Miles per Gallon")



References: <http://stackoverflow.com/questions/5322836/choosing-between-qplot-and-ggplot-in-ggplot2>

Week 7

The link below describes a process built with R to gather market sentiment and correlate to stock prices.

<http://www.rinfinance.com/agenda/2012/talk/Nagar+Hahsler.pdf>

The idea is to pull data from various sources via several R programming packages, merge the data, and run analysis. It takes pieces of sentences from Google, Yahoo, Twitter and rates the sentiment in the tags. It then tries to correlate market sentiment to the price of a particular stock.

The question is how can we differentiate sentiment from the market vs sentiment specific to the company? Many times stocks will go down based on world events that are out of the scope to the company itself. Can one model necessarily work on different companies that are in different industries? If not, why or how can we build a model that is universal to all sentiments

Week 8

Reproducible markdown tools can help in many areas. In my area we work with several systems tools. One area where we have been adding more resource to is a central database that brings all global system data sources together. I think a reproducible statistical markdown tool can help at the presentation layer by bringing everything together to build reports to track and forecast performance bottlenecks before they occur. The markdown’s ability to include scripts and graphs can help visualize a global environment from a birds-eye view. With its multiple document formatting options, we may also be able to include some drill down capabilities to get more detailed information.

QUIZ 8

1)

Beamer is greater math support it is built for statistics

You can use LaTex without knowing to much about it

2) Packrat does project isolation by recording and packaging other package versions that it depends on into one package. That way when you install your project it also installs any dependencies and correct package versions.

3) To keep code from displaying in R Markdown you use the following code:

{r, echo=FALSE}

4) To keep the results from displaying in R Markdown you use the following code:

{r, eval=FALSE)

5) You can use python as the language by using the following code:

{r, engine=’python’}

Quiz 9

Your task is to:

Identify a web site that has an API that you find interesting.[Please provide the URLs for both the web site home page, and the API main/ instruction page]

This url below has worldbank data that includes time series indicators, projects for worldbank operations and financial market data.

<http://data.worldbank.org/node/9>

The url for instructions and to build RestFul queries is:

<http://data.worldbank.org/node/9>

Using basic drop downs I was able to create this RestFul url query:

http://api.worldbank.org/countries/USA/indicators/9.0.Employee.T60?per\_page=100&date=2000:2014

How might you use some or all of the data to create value in an application?[One or two sentences]

This data would be good for financial applications for market trends. It can also be a good source to keep a pulse on world event.

What problems might you encounter building an application that used this web site’s data? [One or two sentences]

One problem may be that it has a plethora of data. Unless you know exactly what you are looking for and have a specific scope you may get lost in the data.

Write and share out a small R program that successfully pulls data from the web site using the provided API.A very short “beachhead” program will receive 100% credit here!

dats <- html("http://api.worldbank.org/v2/en/country/usa")

require(XML)

xmlChildren(dats)

Include thoughts on your initial experience pulling data from this site.  If you were in charge of the site’s API, what would you have done differently.

The data may need more categorization. Seems to be a wide variety of data and takes more time to decide on data then to use the API.

Blackboard 10

A relational database is used mostly for distributed systems where there are multiple datasets that require access from multiple locations and users. Relational databases are used for large data sets, have a defined schema, allows for remote user access and most relational database systems have a way to access data with internal scripting languages or programming.

Spreadsheets are more user client based way to view data and allow users to run quick calculations. A spreadsheet allows a user to access other relational database remotely and import data to review locally and create visualizations with graphs and charts.

Blackboard 11

|  |  |  |  |
| --- | --- | --- | --- |
| **Airport** | **Destination** | **On Time** | **Delayed** |
| Alaska | Phoenix | 221 | 12 |
| AM West | Phoenix | 4840 | 415 |
| Alaska | San Diego | 212 | 20 |
| Alaska | Los Angeles | 497 | 62 |
| Alaska | Seattle | 1841 | 305 |
| AM West | Los Angeles | 694 | 117 |
| AM West | San Diego | 383 | 65 |
| Alaska | San Francisco | 503 | 102 |
| AM West | Seattle | 201 | 61 |
| AM West | San Francisco | 320 | 129 |

select airport, sum(cast(delayed/ontime as double precision)) as delay\_rate from delayrate group by airport;

|  |  |
| --- | --- |
| **Airport** | **Delay\_Rate** |
| Alaska | 0.641840887 |
| AM West | 1.130652079 |

Based on data the Alaska airport has a lower delay rate.

Summarizing data helps group information collectively. Getting more answers requires additional queries, but may end up increasing the number of questions. Having specific scope and requirements in the beginning of a project, may help data analyst avoid the pitfalls of getting lost in the data.