DataScraping

Statistics 4868/6610 Data Visualization

Prof. Eric A. Suess

1/11/2016

Introduction

Chapter 2 gives an idea about different ways data can be stored. There are examples beyond .csv It also gives some ideas about how to get data from websites.

- websites
- scraping - R, python
- data formats - .xlsx, .txt, .json, .xml

This is related to ETL or Extract, Transform, Load

Data Provided by Others

Most of the time data is not collected by you!

You need to check the data to make sure there are no obvious errors in the data file.

You should understand the **context** of the data

- where it came from
- how it was collected
- what it's about

page 22

Data Sources

Seach Engines:
- google
- wolframalpha

Unversities: - DASL CM - DLAB UCB - Library Data Lab UCB - Machine Learning Repository UCI

Data Sources

Sports: - pro-football

Health: - Global Health Facts - Ebola Health Map

Data Sources

Government: - US census - US data.gov - NYC OpenData - DataSF

Data Sources

Other:

- quandl

Data Scraping

Often you can find the exact data that you need, except there's one problem. It's not all in one place or in one file. Instead it's in a bunch of HTML pages or on multiple websites. What should you do?

Scrape the data

page 27

Data Scraping

The author discusses the use of python and beautifulsoup.

He looks at the Weather Underground website to collected maximum temperature data for Buffalo, NY.

I encourage you to read this section in the book. We will return to this later in the class when we introduce python.

Data Scraping

To download the code and data files for the book, you can go to the book's website

Download the code for Chapter 2.

Examine

- wunderdata.txt
- wunderdata.xml
- wunderdata.json

Data Scraping

A lot of people like to keep everything within a safe click interface, but trust me. Pick up just a little bit of programming skills, and you can open up a whole bag of possibilities for what you can do with data.

page 30

Data Scraping

The reason we will return to this is that the code does not work.

bs4 has changed some of the syntax. The following lines need to be updated.
(I am still working to get it to work on Windows.)

- from bs4 import BeautifulSoup
- dayTemp = soup.find_all(attrs={“class”:“wx-value”})[2].string

page 33-37

R and RStudio

At this point we will introduce R and RStudio to do some loading of data from websites.

r-project.org

R Manuals

R Journal

RStudio

Alternative to consider

BlueSky Statistics

R and RStudio

The google for R

rseek.org

Website for learning R

Quick-R

Datamining with R

RDM

R and RStudio

Online book for doing time series analysis in R.

A little book of R for time series

Forecasting Principles and Practice (fpp)

Histograms

To start with R, try the following FlowingData Tutorial.

How to Read and Use Histograms

Read the blog post and run the code.

quantmod

The quantmod library for R can be used to to perform visualiztions for stock trading.

Try it with your stock from Homework 1.

Use the stocks03.R code.

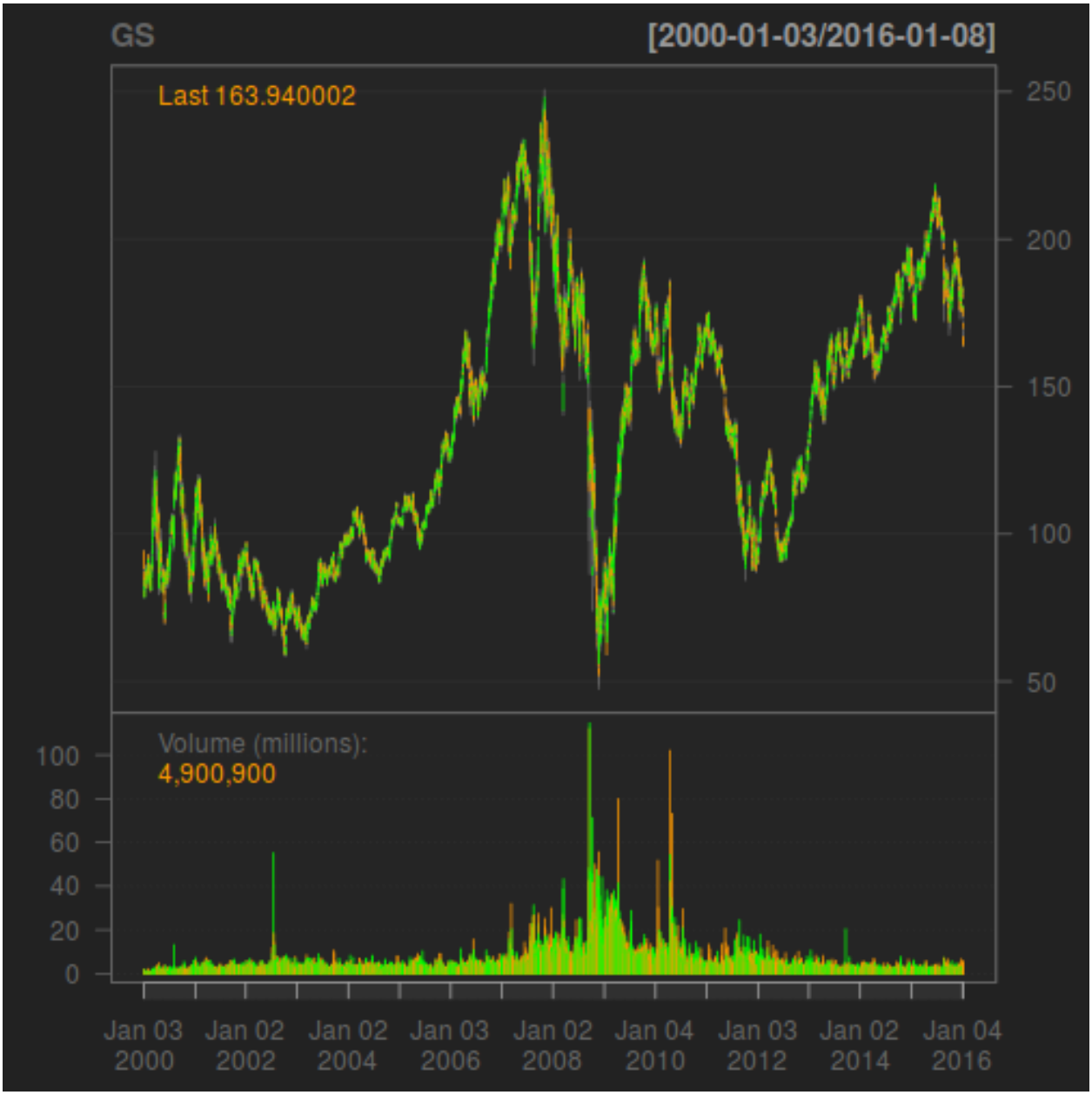
stocks03.R

```
library(quantmod)

getSymbols("GS", src="yahoo",from="2000-01-01")
```

```
[1] "GS"
```

```
chartSeries(GS)
```



Slide With R Code

From the Quick-R website. Creating a Graph

```
attach(mtcars)
summary(mtcars$mpg)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.40   15.42   19.20   20.09   22.80   33.90
```

Slide With Plot

```
plot(wt, mpg)
abline(lm(mpg~wt))
title("Regression of MPG on Weight")
```