USstocks

by Duncan Cummings, David Kane and Andy Yu Zhu Yao

Abstract

The **USstocks** package compiles daily stocks observation of 3081 top performing US companies from 1997-2008, with indictors such as daily returns, trading price, volume, and market capitalization. The data originated from Kane Capital Management's quantitative strategies department. The **USstocks** package cleaned the original data, and merged all valuable information into a single data frame that is user-friendly and easy to work with. In this vignette, we demonstrate some simple techniques and strategies one could employ to explore this rich data set. Specifically, we explore the performance of stocks across time, the market share of different industries, difference in stock volatility across sectors/industries, and the companies that were part of the Dot-com bubble phenomena.

Background

The original data for **USstocks** comes from the **ws.data** package gathered by Kane Capital Management, a hedge fund in operation from 2004 to 2010. The fund maintained a research database of daily equity data for their quantitative strategies, which was determined by a methodical selection process performed at the end of every year. Starting on December 31, 1998, Kane Capital observed all the U.S. stocks over the past year and selected the top 1,500 performing stocks measured by market capitalization. For the following 9 years (1998 - 2007) the process was repeated at the end of each year, culminating in the **ws.data** package of daily stock information for 3,083 companies that were on the top 1500 list for atleast one year.

In **USstocks**, we improved the **ws.data** datasets in two ways. First, we merged all the information into one single data.frame. In the original package, there were multiple data.frames each containing separate information about stocks, such as daily information, yearly information, and sector/industry information. The merged dataframe is called `stocks` in our **USstocks** package, and could be loaded by the simple command `data(stocks)`.

Second, in our examination of the original dataset from Kane Capital, we found certain questionable outliers and cleaned those up. For all intents and purposes of this package, we decided that it was best for the user to perform stock analysis without having to worry about these questionable outliers.

Cleaning Up the Original Stocks Dataset

Specifically, certain questionable outliers in **ws.data** were removed. In the original data set, Kane Capital's methodology was that once a company was recorded as top 1500 for one year, the company's information in the subsequent years would be collected regardless of performance. This led to some suspicious observations as certain companies still had stock information (with unreasonable figures, of course) after the company declared bankruptcy. After cross-verifying with historical data, we found that most of these companies were part of the DotCom bubble. Thus, we removed these statistically insignificant observations from our dataset. A detailed audit trail of the steps we've taken to remove these suspicious observations could be found on our GitHub. Here, we simply summarize the list of stocks that were removed:

Name of Company	Observations Removed	Reason for Removal
CHATHAM CORP-DE	every observation	unreasonably low volume, high return
STRATOSPHERE CORP	every observation	meaningless observation (company filed for bankruptcy)
CYCLELOGIC INC	2 outlying observations removed	inconsistent with data trend, unreasonably high returns
METRICOM INC	1 outlying observations removed	inconsistent with data trend, unreasonably high returns
MARCHFIRST INC	every observation after 2002	DotCom bubble, meaningless observation after bankruptcy
RHYTHMS NETCONNECTIONS INC	every observation after 2002	DotCom bubble, meaningless observation after bankruptcy
CLARENT CORP	every observation after 2002	DotCom bubble, meaningless observation after bankruptcy
LUMINANT WORLDWIDE CORP	1 outlying observations removed	inconsistent with data trend, unreasonably high returns
ACCRUE SOFTWARE INC	every observation	security fraud, insignificant observation
WEBVAN GROUP INC	every observation after 2001	DotCom bubble, meaningless observation after bankruptcy
SCIENT INC	every observation after 2002	DotCom bubble, meaningless observation after bankruptcy
ENGAGE INC	every observation after 2003	DotCom bubble, meaningless observation after bankruptcy

Getting Started

```
install_github("yuzhuyao/USstocks")
library(USstocks)
data(stocks)
```

The USstocks Data

Variables and Meaning	
id	unique security identifier (randomly generated?)
symbol	stock exchange ticker
v.date	date of observation
price.unadj	unadjusted price
price	price (adjusted)
volume.unadj	unadjusted volume
volume	volume (adjusted)
tret	total returns (how is it defined?)
m.sec	sector to which the stock belongs
m.ind	industry to which the stock belongs
name	company name
year	year
cap.usd	market capitalization (of the year)
top.1500	boolean indicating whether the stock is part of the top 1500 performing in that year

Exploring the Data

Before we get started, note that there are two packages that are extremely helpful for working with our `stocks` data.frame. The package **dplyr** enables us to perform basic SQL operations on our data.frame extremely efficiently, and **ggplot2** helps with data trend visualization. Import them by calling: `library(dplyr)` and `library(ggplot2)`. Now, let's get started with looking at IBM's stock from 1998 to 2007.

Let's get started with what the data frame looks like for a specific stock (e.g. IBM) (FIX THE VISUALIZATION):

```
> stocks %>%
+  filter(symbol == "IBM") %>%
+  arrange(v.date) %>% head
```

	id	symbol	v.date	price.unadj	price	volume.unadj	volume	tret	m.sec	m.ind	name	year	cap.usd	top.1500
1	00606601	IBM	1998-01-02	105.625	52.8125	2635000	5270000	0.0095579450	TEC	COMPT INTL BUSINESS MACHINES CORP		1998	170150837500	TRUE
2	00606601	IBM	1998-01-05	106.438	53.2190	5017000	10034000	0.0076970414	TEC	COMPT INTL BUSINESS MACHINES CORP		1998	170150837500	TRUE
3	00606601	IBM	1998-01-06	105.250	52.6250	3556000	7112000	-0.0111614273	TEC	COMPT INTL BUSINESS MACHINES CORP		1998	170150837500	TRUE
4	00606601	IBM	1998-01-07	104.250	52.1250	4308000	8616000	-0.0095011876	TEC	COMPT INTL BUSINESS MACHINES CORP		1998	170150837500	TRUE
5	00606601	IBM	1998-01-08	104.188	52.0940	4059000	8118000	-0.0005947242	TEC	COMPT INTL BUSINESS MACHINES CORP		1998	170150837500	TRUE
6	00606601	IBM	1998-01-09	100.063	50.0315	7183000	14366000	-0.0395918916	TEC	COMPT INTL BUSINESS MACHINES CORP		1998	170150837500	TRUE

We can also visualize the price of the IBM stock over the years (**Figure 1**):

```
> stocks %>%
+  filter(symbol == "IBM") %>%
+  arrange(v.date) %>%
+  ggplot() + geom_point(aes(x = v.date, y = price))
```

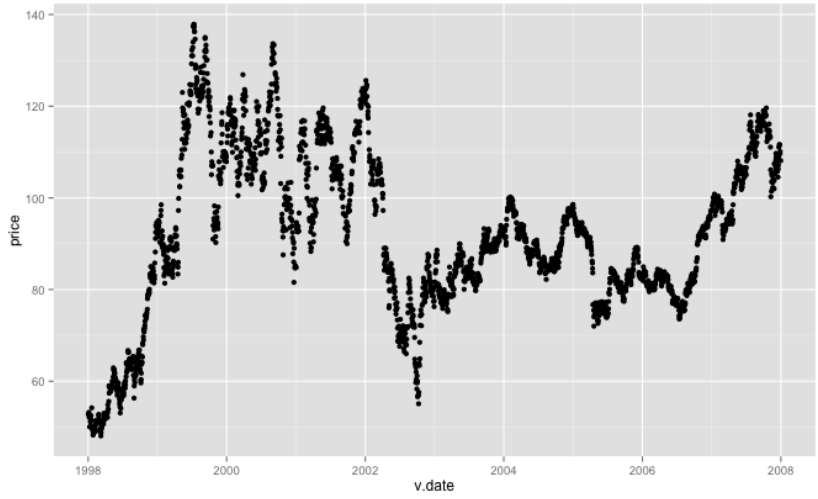

Figure 1: Price of IBM's stock from 1998 to 2007

We could also take a look at some of the key statistics for, say Microsoft:

```
> stocks %>%
+  filter(symbol == "MSFT") %>%
+  summarise_each(funs(mean, median), volume, volume.unadj, price, price.unadj)
```

We could also take a look at market share across sectors

```
> x %>% select(name, m.sec, year, cap.usd) %>% unique %>% filter(year == 1998) -> y
> y %>% filter (!is.na(cap.usd)) %>% group_by(m.sec)
+   %>% summarise(sector_total_1998 = sum(cap.usd)) -> y
> y %>% ggplot(aes(x = "", y = sector_total_1998, fill = m.sec), label = m.sec) +
+   geom_bar(width = 1, stat = "identity") + coord_polar(theta="y")
```

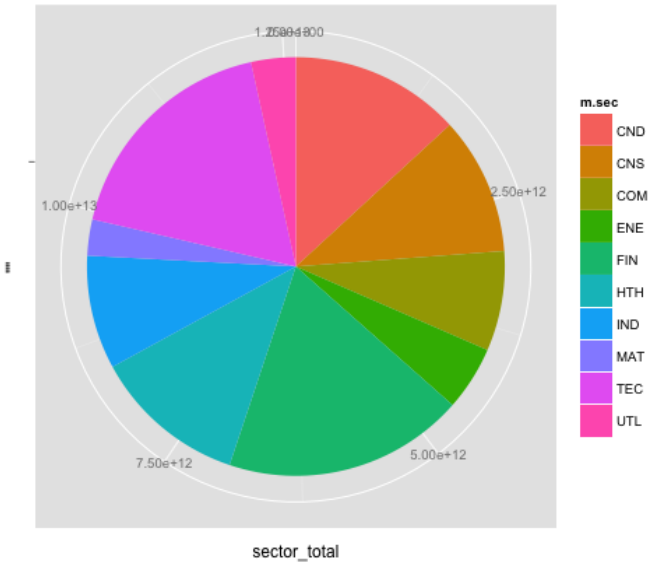

Figure 2: Market share of different sectors for the top 1500 companies in 1998

Cross Sector/Industry Volatility Comparison

We could look at the largest positive and negative daily return across different sectors. As we could see below, the Technology sector contains both the largest and negative daily returns, whereas the Energy sector is much more stable(**Figure 3, 4**).

```
> stocks  %>%
+  group_by(v.date, m.sec)  %>%
+  filter(row_number(desc(tret)) == 1)  %>%
+  ggplot(aes(x = v.date, y = tret)) + geom_point() + facet_wrap(~m.sec)

+ stocks  %>%
+  group_by(v.date, m.sec)  %>%
+  filter(row_number(desc(tret)) == n())  %>%
+  ggplot(aes(x = v.date, y = tret)) + geom_point() + facet_wrap(~m.sec)
```

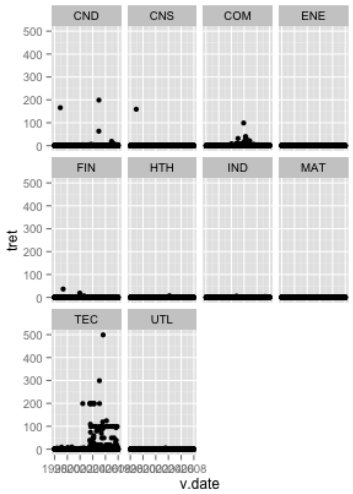

Figure 3: largest positive returns across sectors

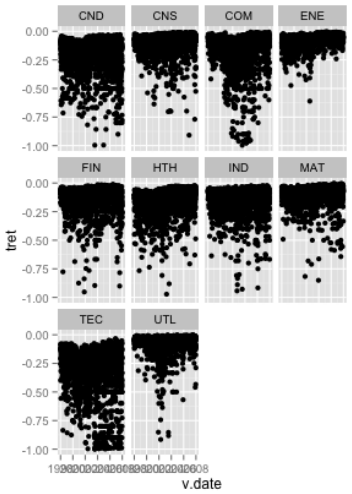

Figure 4: largest negative returns across sectors

First, let's take a glance into the different standard deviation in returns across different sectors(**Figure 5**):

```
> x  %>%
+   group_by(v.date, m.sec)  %>%
+   mutate(sd_ret = sd(tret, na.rm = TRUE)) %>%
+   distinct(v.date, m.sec) %>%
+   ggplot(aes(x = v.date, y = sd_ret)) + geom_point() + facet_wrap(~m.sec)
```

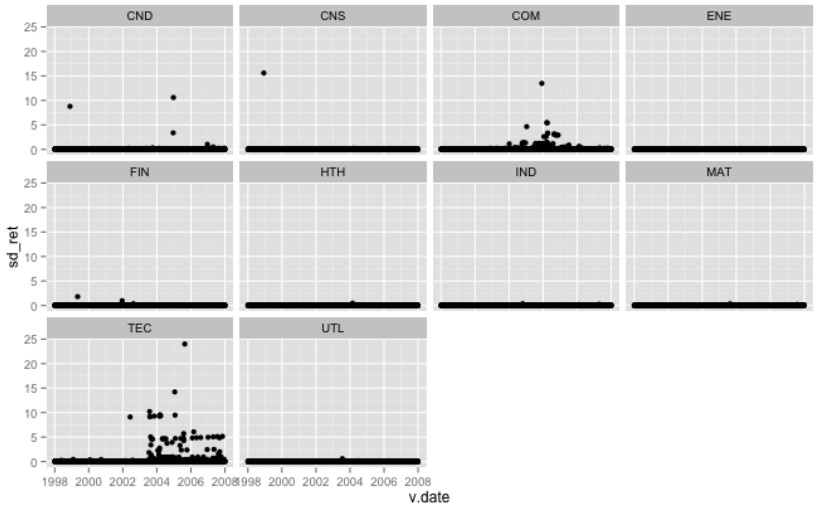

Figure 5: Standard Deviation in returns across sectors

From the plots above, it seems like certain sectors (e.g. Technology) are much more volatile than others. But how can we look into that phenomenon and understand more about what is going on?

A Glimpse into the Dotcom Bubble!

Webvan was a online delivery grocery business that declared bankruptcy in 2001. It was named the largest dot-com flop in history by CNET in 2008. Wouldn't it be interesting to take a look at Webvan's stock across time?

```
> x %>% filter(name == "WEBVAN GROUP INC", !is.na(price), year < 2002) %>%
+   ggplot(aes(x=v.date, y = price)) + geom_point()
```

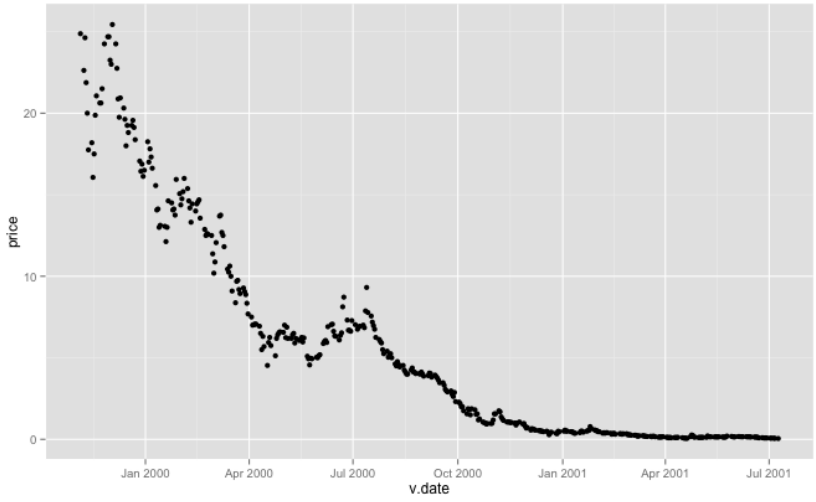

Figure 6: Webvan's Stock Price during the Dot-com bubble burst

As we could see from the figure above, as the Dot-com bubble bursted, the price of Webvan's stock dropped exponentially within a year, until it declared bankruptcy in July 2001.

Summary

Bibliography

D. Kane. ws.data, 2008. [p]

Authors

Duncan Cummings
Williams College
Williamstown, MA
USA
dmc3@williams.edu

David Kane
Hutchin Hill Capital
Address
Country
Dave.Kane@gmail.com

Andy Yu Zhu Yao
Williams College
Williamstown, MA
USA
andy.yu.zhu.yao@williams.edu