IBS intro to R

Shannon B. Hagerty

8/26/2019

Incredibly Important Business Update

What is R - R is an open source programming language that is very popular for data analysis / data scien. - Its popularity means a lot of code is already written and available for you to do common data analysis tasks.

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts -------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggthemes)
```

Why should you learn R - It is open source - It enables you to create reproducible analyses, which can save you both time and some frustration - It gives you a lot of great options for communicating your analyses out (i.e. ggplot for graphing,Rmarkdown for reports, shiny for apps)

Introduction to Rstudio, Rmarkdown, and R - Rstudio is the IDE we like to use to write our R code - Rmarkdown is a document that allows you to put code, graphs, and text in one document.

Load Data

```
forbes<-read_csv('Forbes2000.csv')
```

```
## Parsed with column specification:
## cols(
##   rank = col_double(),
##   name = col_character(),
##   country = col_character(),
##   category = col_character(),
##   sales = col_double(),
##   profits = col_double(),
##   assets = col_double(),
##   marketvalue = col_double()
## )
```

Explore the data set

```
glimpse(forbes)
```

```
## Observations: 2,000
## Variables: 8
## $ rank        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...
```

```
## $ name        <chr> "Citigroup", "General Electric", "American Intl Gr...
## $ country     <chr> "United States", "United States", "United States",...
## $ category    <chr> "Banking", "Conglomerates", "Insurance", "Oil & ga...
## $ sales       <dbl> 94.71, 134.19, 76.66, 222.88, 232.57, 49.01, 44.33...
## $ profits     <dbl> 17.85, 15.59, 6.46, 20.96, 10.27, 10.81, 6.66, 7.9...
## $ assets      <dbl> 1264.03, 626.93, 647.66, 166.99, 177.57, 736.45, 7...
## $ marketvalue <dbl> 255.30, 328.54, 194.87, 277.02, 173.54, 117.55, 17...
```

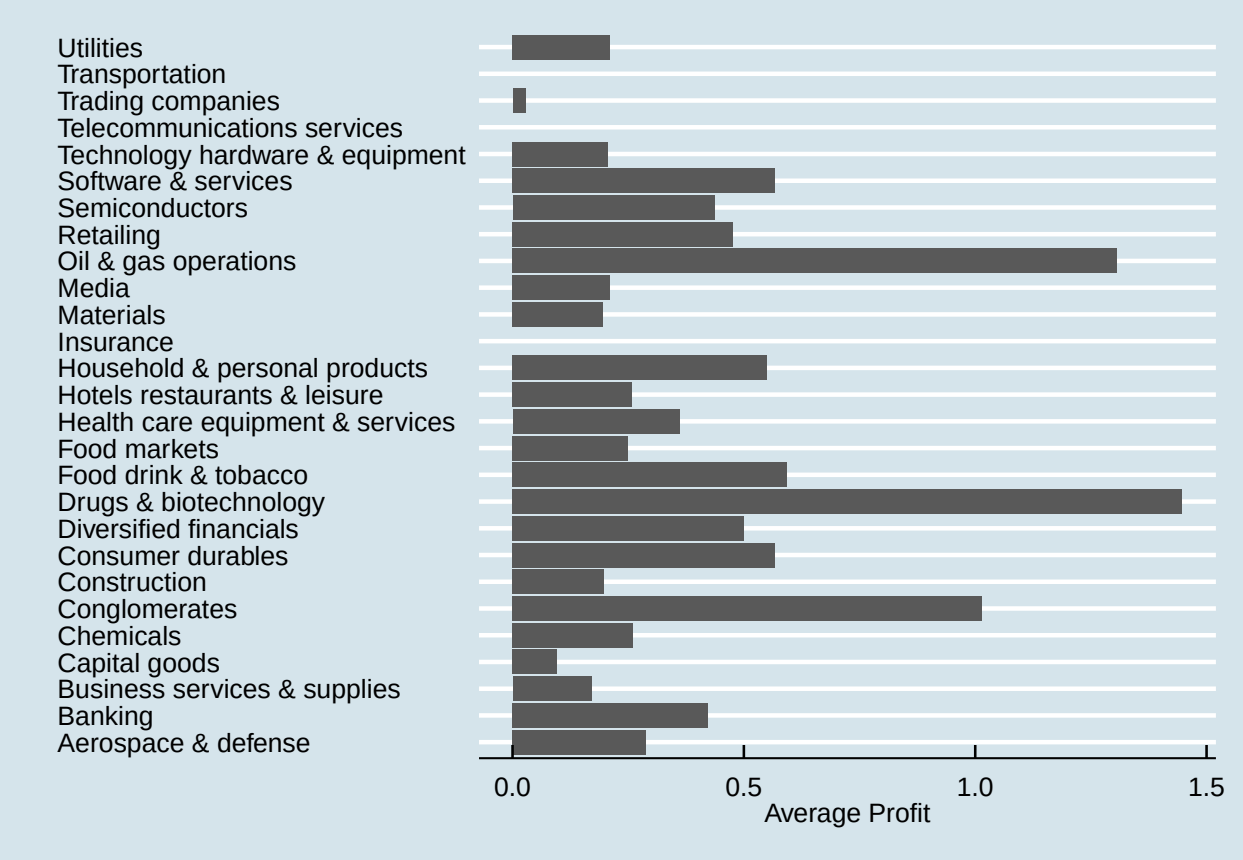
Summarize the Data by category

```
category_summary<- forbes %>% group_by(category) %>% summarize(mean_sales = mean(sales), mean_profit= m
```

Plot the Summary

```
ggplot(category_summary) + geom_bar(aes(x=category, y=mean_profit), stat="identity")+theme_economist()+
```

```
## Warning: Removed 3 rows containing missing values (position_stack).
```



Now I can tell you everything that is important about this graph in a really nice report.

Can you make your own graph?

Try to adjust the code below to summarize by the country column.

Summarize the Data by country

```
country_summary<- forbes %>% group_by(PUT_SOMETHING_HERE) %>% summarize(mean_sales = mean(sales), mean_
```

```
## Error: Column `PUT_SOMETHING_HERE` is unknown
```

Okay, now there are a lot of countries. Take a look at the data set and lets select out only the countries with an average market value greater than 20.

Get only countries with highest market value

```
country_summary<- country_summary %>% filter(mean_marketvalue > 20)
```

```
## Error in eval(lhs, parent, parent): object 'country_summary' not found
```

ADAPT THE CODE BELOW TO Plot the country summary Try changing the metric we're plotting (i.e. instead of mean_marketvalue try mean_profit)

```
ggplot(category_summary) + geom_bar(aes(x=category, y=mean_profit), stat="identity")+theme_economist()+
```

```
## Warning: Removed 3 rows containing missing values (position_stack).
```