hw6

Yan Lin

October 7, 2017

```
## dataset1 from my post - carpinteria

require(tidyr)
```

```
## Loading required package: tidyr

## Warning: package 'tidyr' was built under R version 3.4.2
```

```
require(dplyr)
```

```
## Loading required package: dplyr

## Warning: package 'dplyr' was built under R version 3.4.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
require(stringr)
```

```
## Loading required package: stringr
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(corrr)
```

```
## Loading required package: corrr

## Warning: package 'corrr' was built under R version 3.4.2
```

```
## read the main table
a <- read.csv("C:/Users/Yan/Documents/carpinteria.csv",header=TRUE, sep=",")
View(a)

## give the column name to the first two columns
colnames(a)[1] <- "type"
colnames(a)[2] <- "prey_host"


##locate the NA column and cell
na <- which(is.na(a),TRUE)

## exact non NA columns and rows
a.1 <- a[3:130,1:130]
```

```
## reshape the table
a.2 <- gather(a.1, "consumer","link",3:130)
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
## sign the value
a.2$link <- ifelse(a.2$link == "0","no link", ifelse(a.2$link == "1","1st intermediat host", ifelse(a.2

link <- a.2$link
y <- as.data.frame(table(link))
sum(y$Freq)
```

```
## [1] 16384
```

```
y <- y %>% mutate(ratio = Freq/16384)
y$ratio <- paste(round(100*y$ratio,2),"%",sep="")

## rank
y %>%
  arrange(desc(ratio))
```

```
##                                  link  Freq  ratio
## 1                             no link 14094 86.02%
## 2                 parasite -  parasite   923  5.63%
## 3     predation on parasite in a host   572  3.49%
## 4     predation on parasite, possible   338  2.06%
## 5                       predator-prey   239  1.46%
## 6                          final host    88  0.54%
## 7                2nd intermediate host    44  0.27%
## 8             predation on free-living    33   0.2%
## 9  2ndintermediate host external cyst    17   0.1%
## 10                                4.1    16   0.1%
## 11               1st intermediat host    15  0.09%
## 12          1st & 2nd intermediate hst     2  0.01%
## 13                       egg predator     1  0.01%
## 14                      micropredation     2  0.01%
```

```
## conclusion
## There is 86.02% of no link between prey/hosts and consumers


## dataset 2 - Funding Summary
m <- read.csv("C:/Users/Yan/Documents/FundingSummary.csv",header=TRUE, sep=",")
View(m)

##revise some columns' name
colnames(m)[2] <- "five_digit_NTDID"
colnames(m)[3] <- "four_digit_NTDID"
colnames(m)[14] <- "2015_status"

## locate the NA columns and rows
na.1 <- which(is.na(m),TRUE)

## extract non NA columns and rows, excluding the sum row at the bottom
```

```r
m.1 <- m[1:1114,1:39]

##funding by year
m.2 <- gather(m.1, "year","funding",15:39)
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```r
## extract year in column year
m.2$year <- str_sub(string=m.2$year,start=2,end=5)

## convert the dollar sign value to numeric
m.2$funding <- as.numeric(gsub('[$,]', '', m.2$funding))

## funding by state
m.2[is.na(m.2)] <- 0
state_funding <- aggregate(m.2$funding,by=list(state=m.2$State),FUN=sum)
state_funding
```

```
##    state            x
## 1              50678568
## 2     AK   1247230642
## 3     AL   1159902994
## 4     AR    523419835
## 5     AZ  10343386390
## 6     CA 148606085920
## 7     CO  16339179048
## 8     CT   5403263844
## 9     DC  40813467811
## 10    DE   1904528721
## 11    FL  27657115435
## 12    GA  16177624742
## 13    HI   5849064512
## 14    IA   1467786166
## 15    ID    286659117
## 16    IL  62378307681
## 17    IN   4257832884
## 18    KS    775760535
## 19    KY   2376424439
## 20    LA   4244137031
## 21    MA  43003527186
## 22    MD  17725154574
## 23    ME    640130744
## 24    MI  11468550112
## 25    MN  11394262095
## 26    MO   8690916213
## 27    MS    305334540
## 28    MT    258032387
## 29    NC   6159816269
## 30    ND    214765101
## 31    NE    800713163
## 32    NH    295388363
## 33    NJ  68041037461
## 34    NM   1475908647
## 35    NV   4534959197
## 36    NY 277893491470
## 37    OH  16064926574
## 38    OK    973627753
## 39    OR  12832357469
## 40    PA  45209480726
## 41    PR   4171602784
## 42    RI   2039469666
## 43    SC   1151467731
## 44    SD    186736804
## 45    TN   3673990614
## 46    TX  44487250948
## 47    UT   8195862632
## 48    VA   7553456119
## 49    VI     38612625
## 50    VT    215807746
## 51    WA  41421013435
## 52    WI   6289024421
## 53    WV    645135392
## 54    WY     55151853
```

```
## funding by year
year_funding <- aggregate(m.2$funding,by=list(year=m.2$year),FUN=sum)
year_funding
```

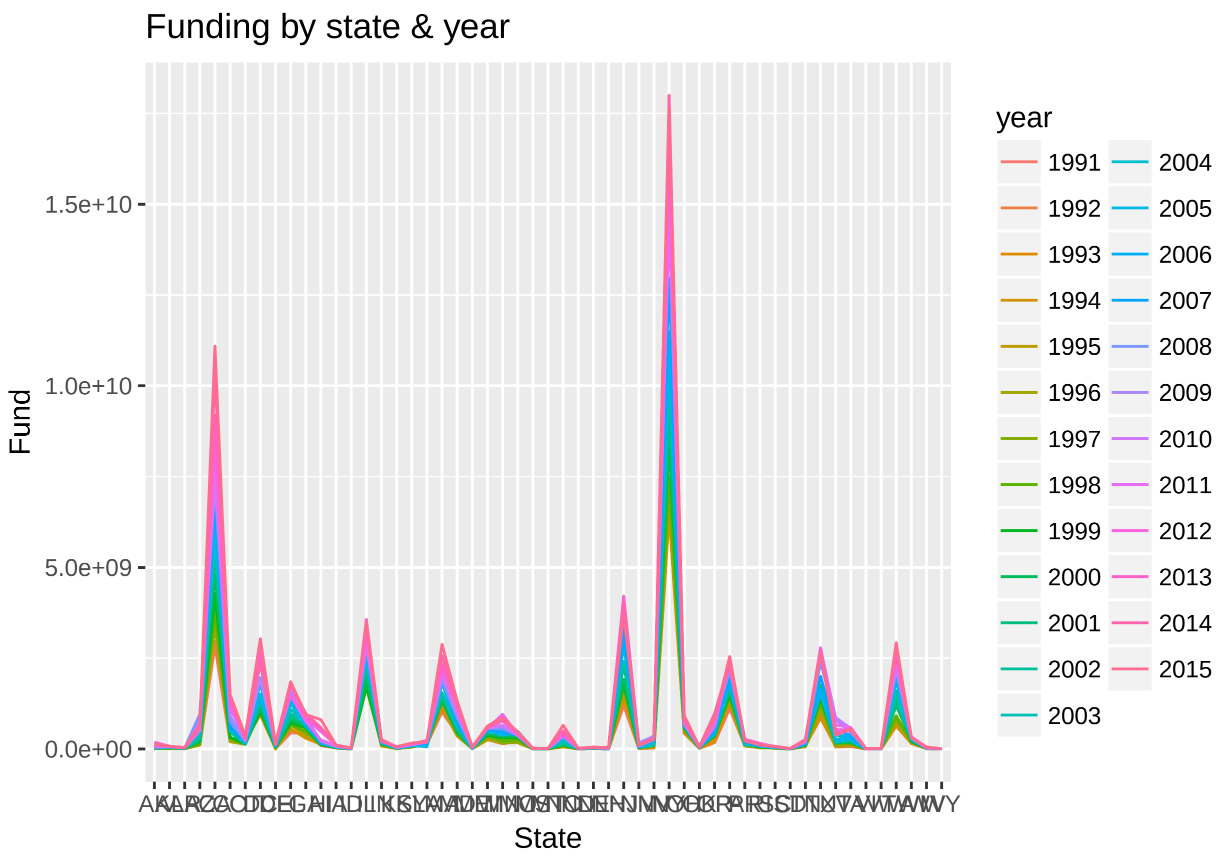
```
##    year           x
## 1  1991 22046218705
## 2  1992 22328262190
## 3  1993 22521107231
## 4  1994 22965224298
## 5  1995 24260887515
## 6  1996 25050153902
## 7  1997 26108472603
## 8  1998 26649378658
## 9  1999 29128822553
## 10 2000 31027267090
## 11 2001 34549276538
## 12 2002 37096627731
## 13 2003 38764669696
## 14 2004 39980023555
## 15 2005 40924317277
## 16 2006 43493139290
## 17 2007 47305205161
## 18 2008 52565656846
## 19 2009 54287636546
## 20 2010 54354844811
## 21 2011 55412791386
## 22 2012 58463758794
## 23 2013 61259936263
## 24 2014 63741620103
## 25 2015 65683520387
```

```
## funding by state and year
state_year_funding <- aggregate(m.2$funding,by=list(state=m.2$State,year=m.2$year),FUN=sum)
```

```
## remove the blank row
state_year_funding[state_year_funding==""] <- NA
state_year_fund <- na.omit(state_year_funding)
colnames(state_year_fund)[3] <- "funding"
state_year_fund %>%
  filter(funding >= 10000000000) %>%
  arrange(desc(funding))
```

```
##    state year     funding
## 1     NY 2014 17999654736
## 2     NY 2015 17675561757
## 3     NY 2013 17321401560
## 4     NY 2012 16351495949
## 5     NY 2009 15340488761
## 6     NY 2011 14842962808
## 7     NY 2008 14648123518
## 8     NY 2010 14615650308
## 9     NY 2007 12967938176
## 10    NY 2005 11489330420
## 11    NY 2006 11273531039
## 12    CA 2015 11092836678
## 13    NY 2004 11010895079
## 14    NY 2003 10542081142
## 15    CA 2014 10352524572
## 16    NY 2002 10303223307
```

```
## graph
g <- ggplot(state_year_fund, aes( x= state, y = funding ))
g <- g + geom_line(aes(color=year,group = year))
g <- g + labs(title = "Funding by state & year", x = "State", y = "Fund")
g
```



```
## conclusion
## out of 54 states, NY and CA have the funding over 10 billions
## New York has the most



## datatset 3 - 2017 Index of Economic Freedom
n <- read.csv("C:/Users/Yan/Documents/index2017.csv",header=TRUE, sep=",")
View(n)

## identify NA columns and extract the non NA ones
na.2 <- which(is.na(n),TRUE)
n.1 <- n[1:186,1:34]

## remove the identical columns
identical(n.1[['Country.Name']],n.1[['Country']])
```

```
## [1] TRUE
```

```
n.1$Country <- NULL

## convert the dollar column to numerics
colnames(n.1)[26] <- "GDP_PPP"
n.1$GDP_PPP <- as.numeric(gsub('[$,]', '', n.1$GDP_PPP))
```

```
## Warning: NAs introduced by coercion
```

```r
n.1$GDP.per.Capita..PPP. <- as.numeric(gsub('[$,]', '', n.1$GDP.per.Capita..PPP.))
```

```
## Warning: NAs introduced by coercion
```

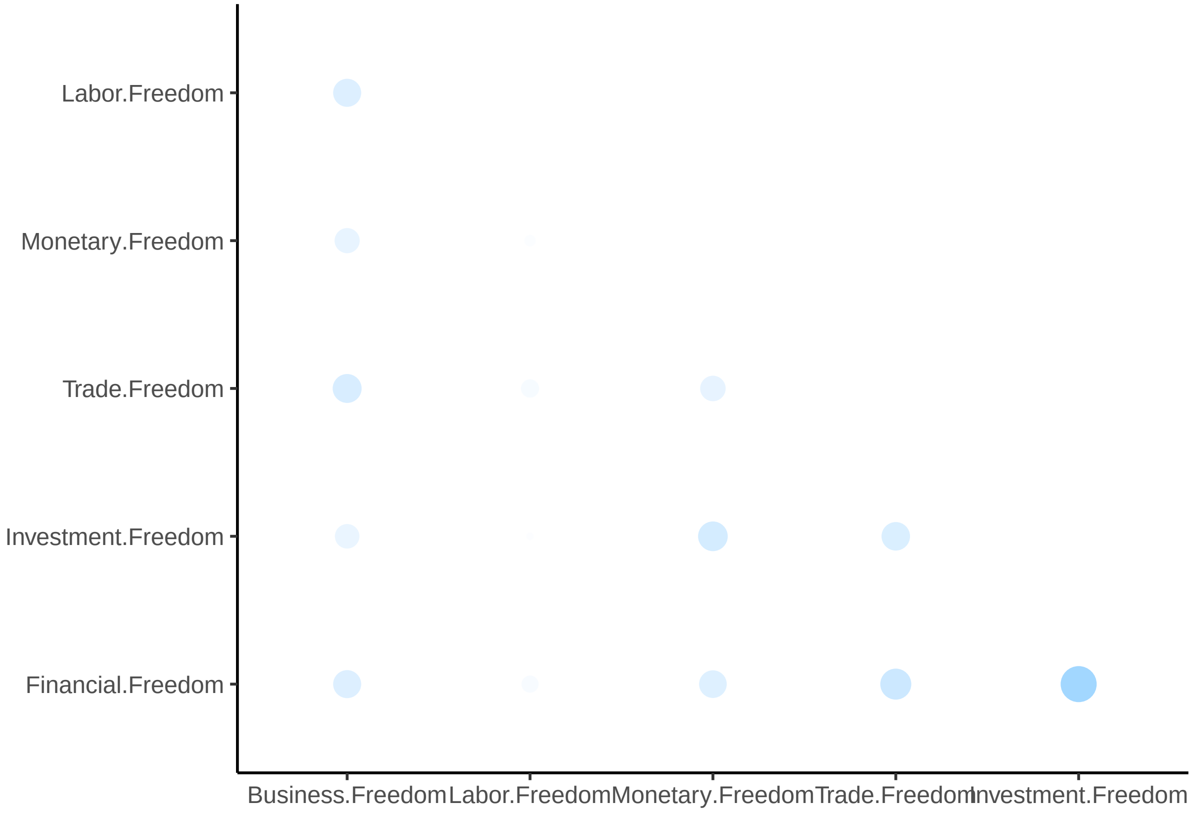
```r
## correlation
n.1$Unemployment.... <- as.numeric(n.1$Unemployment....)
n.1$Business.Freedom <- as.numeric(n.1$Business.Freedom)
n.1$Labor.Freedom <- as.numeric(n.1$Labor.Freedom)
n.1$Monetary.Freedom <- as.numeric(n.1$Monetary.Freedom)
n.1$Trade.Freedom <- as.numeric(n.1$Trade.Freedom)
n.1$Investment.Freedom <- as.numeric(n.1$Investment.Freedom)
n.1$Financial.Freedom <- as.numeric(n.1$Financial.Freedom)

n.2 <- select(n.1,Business.Freedom:Financial.Freedom)

b <- n.2 %>%
  corrr::correlate() %>%
  corrr::shave()
corrr::fashion(b)
```

```
##              rowname Business.Freedom Labor.Freedom Monetary.Freedom
## 1   Business.Freedom
## 2      Labor.Freedom              .54
## 3   Monetary.Freedom              .47           .27
## 4      Trade.Freedom              .57           .33              .48
## 5 Investment.Freedom              .45           .26              .59
## 6  Financial.Freedom              .54           .32              .53
##   Trade.Freedom Investment.Freedom Financial.Freedom
## 1
## 2
## 3
## 4
## 5           .55
## 6           .63                .80
```

```r
corrr::rplot(b)
```



```
## conclusion
## Financial freedom and investment freedom has the strongest effect
```