Trends for startup industries

by Zhihui Xie

==========

Setting and library for this analysis

```
# global setting for this analysis
library(knitr)
opts_chunk$set(fig.width=12, fig.height=8,
               warning=FALSE, message=FALSE)
```

```
# Load all of the packages that you end up using
# in your analysis in this code chunk.
suppressMessages(library(ggplot2))
suppressMessages(library(dplyr))
```

1. Overview

This data set includes information about startups worldwide collected by CruchBase (https://info.crunchbase.com/about/crunchbase-data-exports/ (https://info.crunchbase.com/about/crunchbase-data-exports/)). The exploration here will focus on the global trends of startup industry occupied percentage over time from 1990 to 2013.

2. Exploration

2.1 Load the Data

```
setwd("~/Downloads")
startups <- read.csv("crunchbase_monthly_export_companies_2014.csv")
```

2.2 Summary of the Data Set

```
dim(startups)
```

```
## [1] 54294    18
```

```
summary(startups)
```

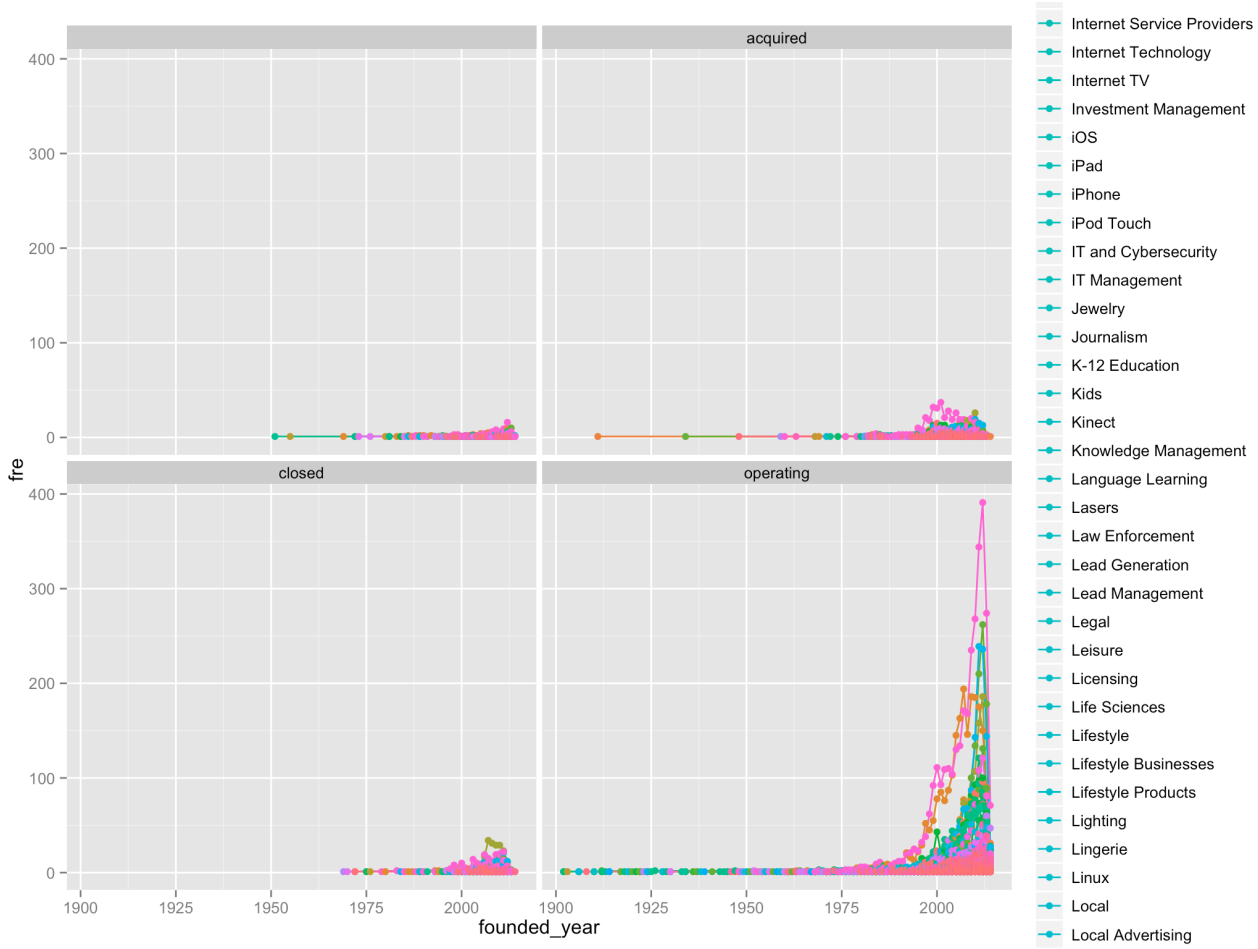
```
##                                          permalink          name
##                                               : 4856          : 4856
##  /organization/prysm                          :    2   Roost  :    4
##  /organization/treasure-valley-urology-services:    2   Spire  :    4
##  /organization/-qounter                       :    1   Compass:    3
##  /organization/0-6-com                        :    1   Cue    :    3
##  /organization/004-technologies               :    1   Hubbub :    3
##  (Other)                                      :49431   (Other):49421
##                          homepage_url           category_list
##                                 : 8305                  : 8817
##  http://app.thotz.co/           :    2   |Software|     : 3650
##  http://attunelive.com          :    2   |Biotechnology|: 3597
##  http://ayibang.com             :    2   |E-Commerce|   : 1263
##  http://bubbly.net              :    2   |Mobile|       : 1211
##  http://clevelandfoundation.org:    2   |Curated Web|  : 1120
##  (Other)                        :45979   (Other)        :34636
##              market        funding_total_usd       status
##                 : 8824    -         : 8460            : 6170
##   Software      : 4620              : 4856   acquired : 3692
##   Biotechnology : 3688    1,000,000 :  928   closed   : 2603
##   Mobile        : 1983    500,000   :  765   operating:41829
##   E-Commerce    : 1805    100,000   :  750
##   Curated Web   : 1655    40,000    :  680
##  (Other)        :31719   (Other)    :37855
##   country_code     state_code              region
##  USA    :28793          :24133                :10129
##         :10129   CA     : 9917   SF Bay Area  : 6804
##  GBR    : 2642   NY     : 2914   New York City: 2577
##  CAN    : 1405   MA     : 1969   Boston       : 1837
##  CHN    : 1239   TX     : 1466   London       : 1588
##  DEU    :  968   WA     :  974   Los Angeles  : 1389
##  (Other): 9118   (Other):12921   (Other)      :29970
##             city       funding_rounds        founded_at    founded_month
##               :10972   Min.   : 1.000             :15740          :15812
##  San Francisco: 2615   1st Qu.: 1.000   2012-01-01: 2181   2012-01: 2327
##  New York     : 2334   Median : 1.000   2011-01-01: 2161   2011-01: 2286
##  London       : 1257   Mean   : 1.696   2010-01-01: 1855   2010-01: 1952
##  Palo Alto    :  597   3rd Qu.: 2.000   2009-01-01: 1603   2013-01: 1722
##  Austin       :  583   Max.   :18.000   2013-01-01: 1575   2009-01: 1655
##  (Other)      :35936   NA's   :4856     (Other)   :29179   (Other):28540
##  founded_quarter  founded_year     first_funding_at   last_funding_at
##         :15812   Min.   :1902              : 4856             : 4856
##  2012-Q1: 2904   1st Qu.:2006    2012-01-01:  468   2013-01-01:  387
##  2011-Q1: 2768   Median :2010    2013-01-01:  463   2014-01-01:  364
##  2010-Q1: 2259   Mean   :2007    2008-01-01:  422   2012-01-01:  348
##  2013-Q1: 2206   3rd Qu.:2012    2011-01-01:  392   2008-01-01:  302
##  2009-Q1: 1852   Max.   :2014    2007-01-01:  342   2011-01-01:  272
##  (Other):26493   NA's   :15812   (Other)   :47351   (Other)   :47765
```

2.3 Calculate number of startups by industry

```
# clean data: remove invalid data point in "market" and "founded_year" column
startups <- subset(startups, market != ""&!is.na(founded_year))
# calculate number of new founded startups in each maket in each year
startups.by.year.market <- startups %>%
  group_by(founded_year, market, status) %>%
  summarise(
            fre = n(),
            funding = sum(as.numeric(funding_total_usd))
            )
```

2.4 First Exploration - overview the changes of number of startups by industry

```
p0 <- ggplot(aes(x = founded_year, y = fre), data = startups.by.year.market) +
  geom_line(aes(color = market)) + geom_point(aes(color = market)) + facet_wrap(~s
tatus)
p0
```



There is too much information there and it's hard to get the effective comparisons. Therefore, the data were truncated and focused on hot industries from 1990 to 2013.

2.5 Get a Subset of Data (founded year from 1990 to 2013)

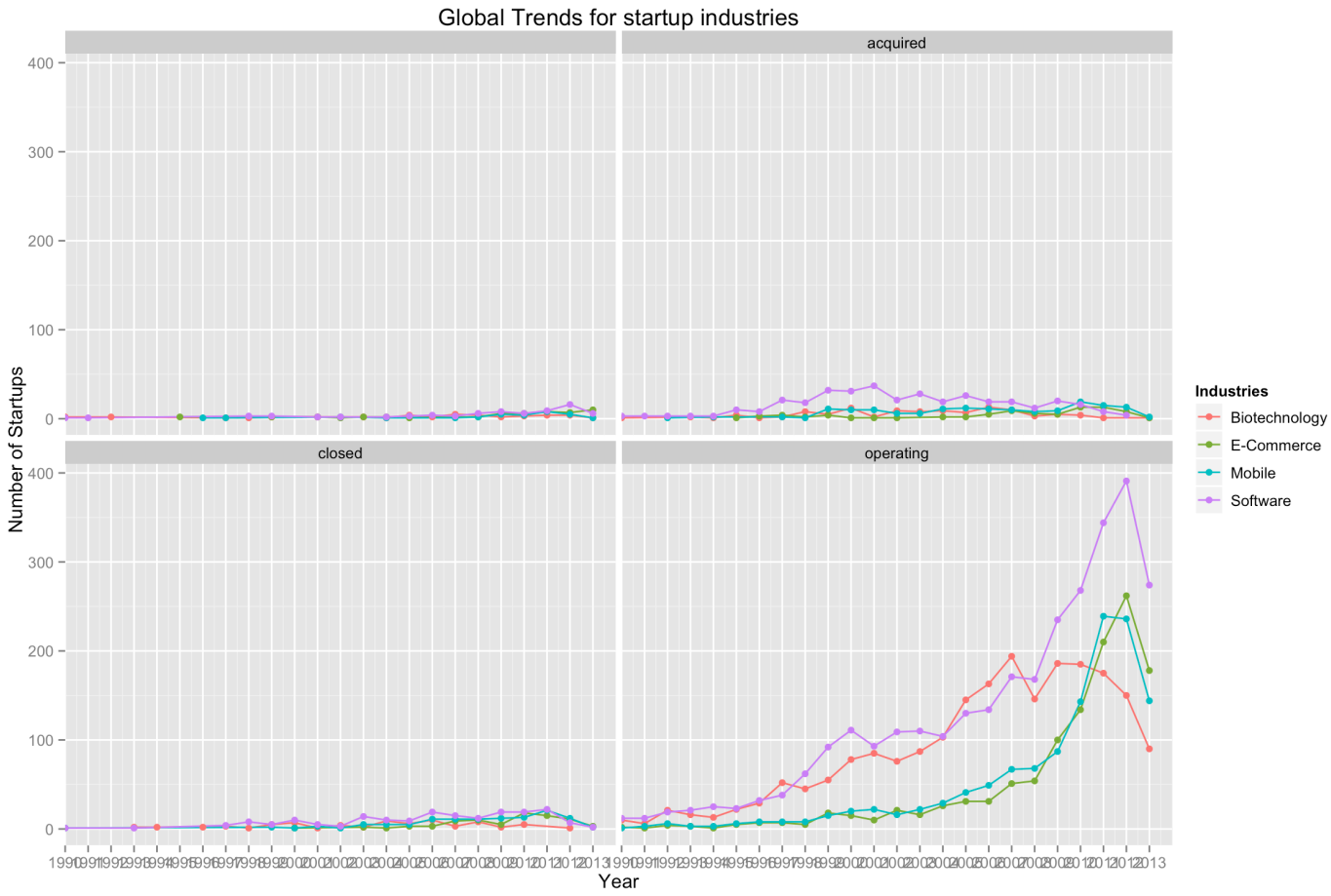
```
# limit data from 1990 to 2013
startups.by.year.market <- subset(startups.by.year.market, (1989 < founded_year) &
(founded_year < 2014))
# find hot industries with total number of startups greater than 1400
markets <- as.data.frame(table(startups$market))
sub_markets <- subset(markets, Freq > 1400 & !Var1 == "")
c_market <- c(as.character(sub_markets$Var1))
# list hot industrie for startups
c_market
```

```
## [1] " Biotechnology " " E-Commerce "    " Mobile "        " Software "
```

```
hot_markets <- c(" Biotechnology ", " E-Commerce ", " Mobile ", " Software ")
# make table of hot industries and founded year
hot_startups.by.year.market <- subset(startups.by.year.market, market %in% hot_mar
kets)
```

2.6 Second Exploration - how does the number of hot startups change by industry?

```
p1 <- ggplot(aes(x = founded_year, y = fre), data = hot_startups.by.year.market) +
  geom_line(aes(color = market)) + geom_point(aes(color = market)) + facet_wrap(~s
tatus) +
  scale_x_continuous(breaks = seq(1990, 2013, 1)) +
  coord_cartesian(xlim = c(1990, 2014)) +
  labs(x = "Year", y = "Number of Startups", color = "Industries") +
  ggtitle("Global Trends for startup industries")
p1
```



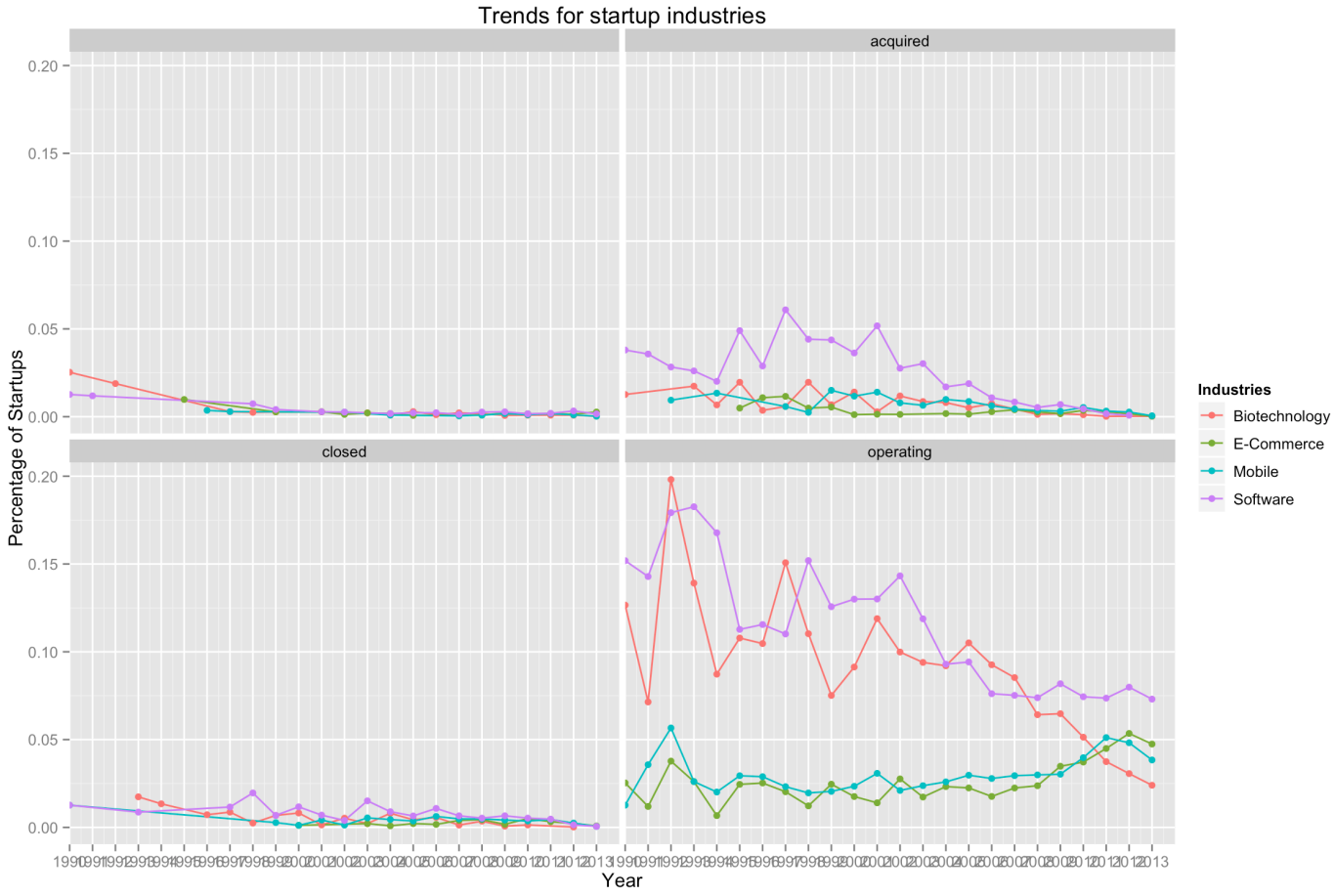
2.7 Data Normalization

```
# calculate total new founded startups each year
df0 <- startups %>%
  group_by(founded_year) %>%
  summarise(
            tol = n()
            )
# calculate new founded startups from 1990 to 2013 in each industry
df1 <- subset(df0, (1989 < founded_year) & (founded_year < 2014))

# merge hot startups and df1
df2 <- merge(hot_startups.by.year.market, df1, by = "founded_year", all = TRUE)
# calculate percentage of startups in each industry each year
df3 <- df2 %>%
  group_by(founded_year, market, status) %>%
  summarise(
    percentage = fre/tol)
#head(df3)
# merge to obtain final subset of data
df.final <- merge(df2, df3, by = c("founded_year", "market", "status"), all = TRU
E)
```

2.8 Third Explorarion - how does the percentage of hot startups change by industry?

```
p2 <- ggplot(aes(x = founded_year, y = percentage), data = df.final) +
  geom_line(aes(color = market)) + geom_point(aes(color = market)) + facet_wrap(~s
tatus) +
  scale_x_continuous(breaks = seq(1990, 2013, 1)) +
  coord_cartesian(xlim = c(1990, 2014)) +
  labs(x = "Year", y = "Percentage of Startups", color = "Industries") +
  ggtitle("Trends for startup industries")
p2
```

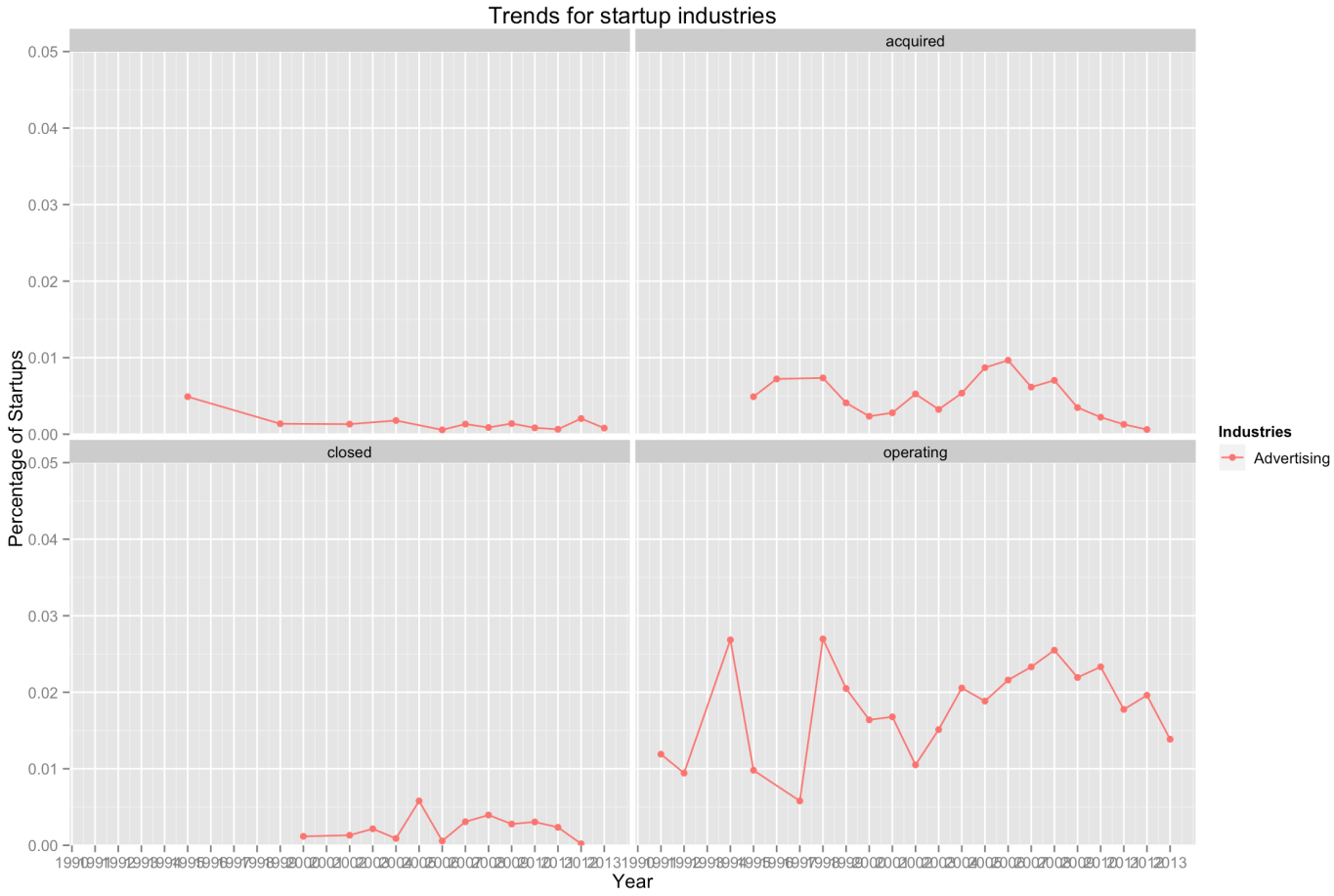


It seems that all the selected hot induries for startups tend to decline. It's interesting to explore which industry for startups has a increased trend. To do that, number of startups greater than 300 in each industry were selected and visulized to find industry with increased trends.

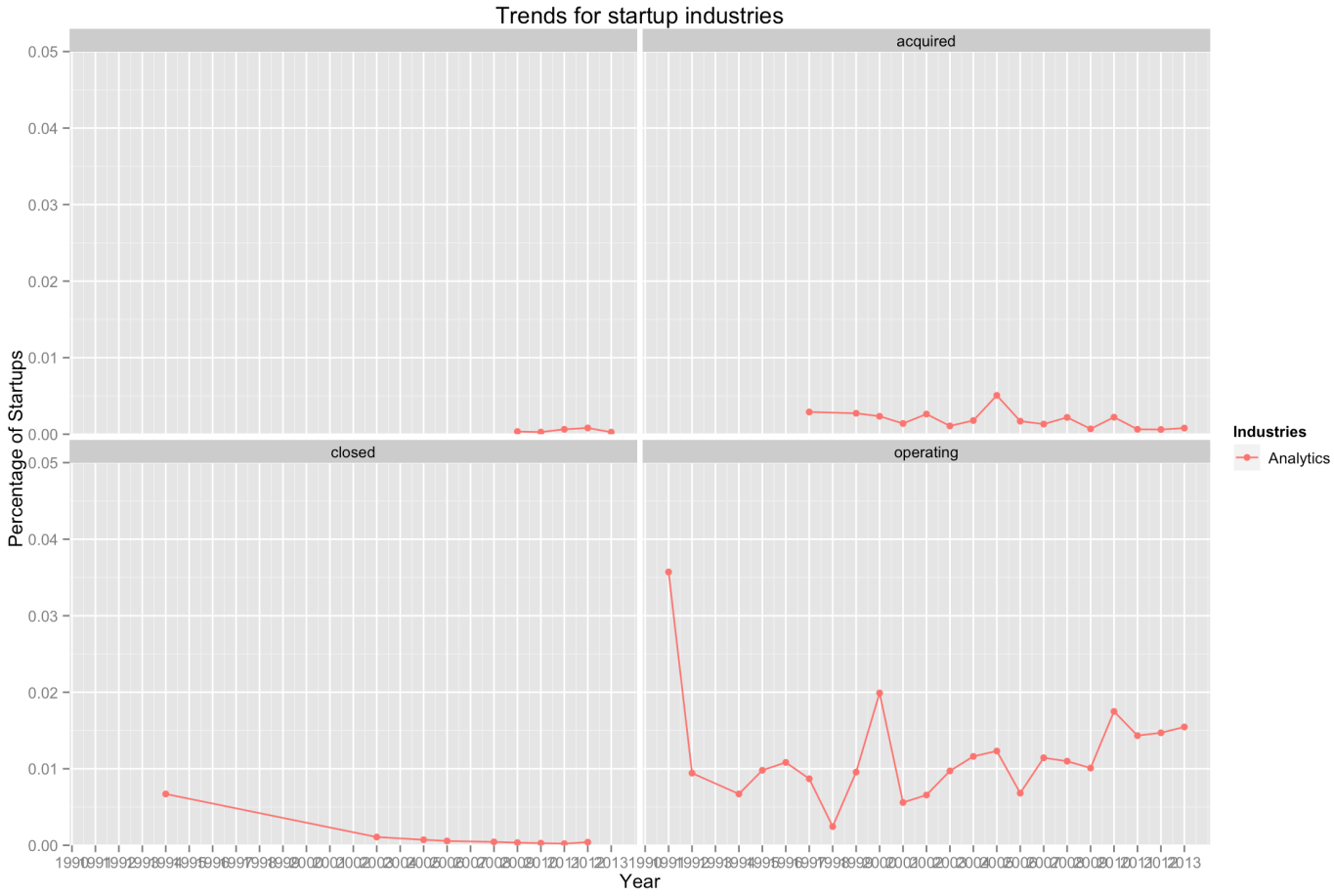
2.9 Explore industries with increased trends

```
# add increased industries
sub_markets <- subset(markets, Freq > 300 & !Var1 == "")
c_market <- c(as.character(sub_markets$Var1))
startups.othermarket <- subset(startups.by.year.market, market %in% c_market)
# limit data by founded year
df2.other <- merge(startups.othermarket, df1, by = "founded_year", all = TRUE)
# calculate percentage of startups in each industry each year
df3.other <- df2.other %>%
  group_by(founded_year, market, status) %>%
  summarise(
    percentage = fre/tol)
# merge to obtain final subset of data
df.other <- merge(df2.other, df3.other, by = c("founded_year", "market", "statu
s"), all = TRUE)
# plot each industry to find the one with increased trends.
for (i in c_market){
      print (i)
      p.other <- ggplot(aes(x = founded_year, y = percentage), data = subset(df.ot
her,
                  market == i)) + geom_line(aes(color = market)) +
                  geom_point(aes(color= market)) + facet_wrap(~status) +
                  scale_x_continuous(breaks = seq(1990, 2013, 1)) +
                  coord_cartesian(xlim = c(1990, 2014)) +
                  coord_cartesian(ylim = c(0, 0.05)) +
                  labs(x = "Year", y = "Percentage of Startups", color = "Industr
ies") +
                  ggtitle("Trends for startup industries")
      print (p.other)
}
```

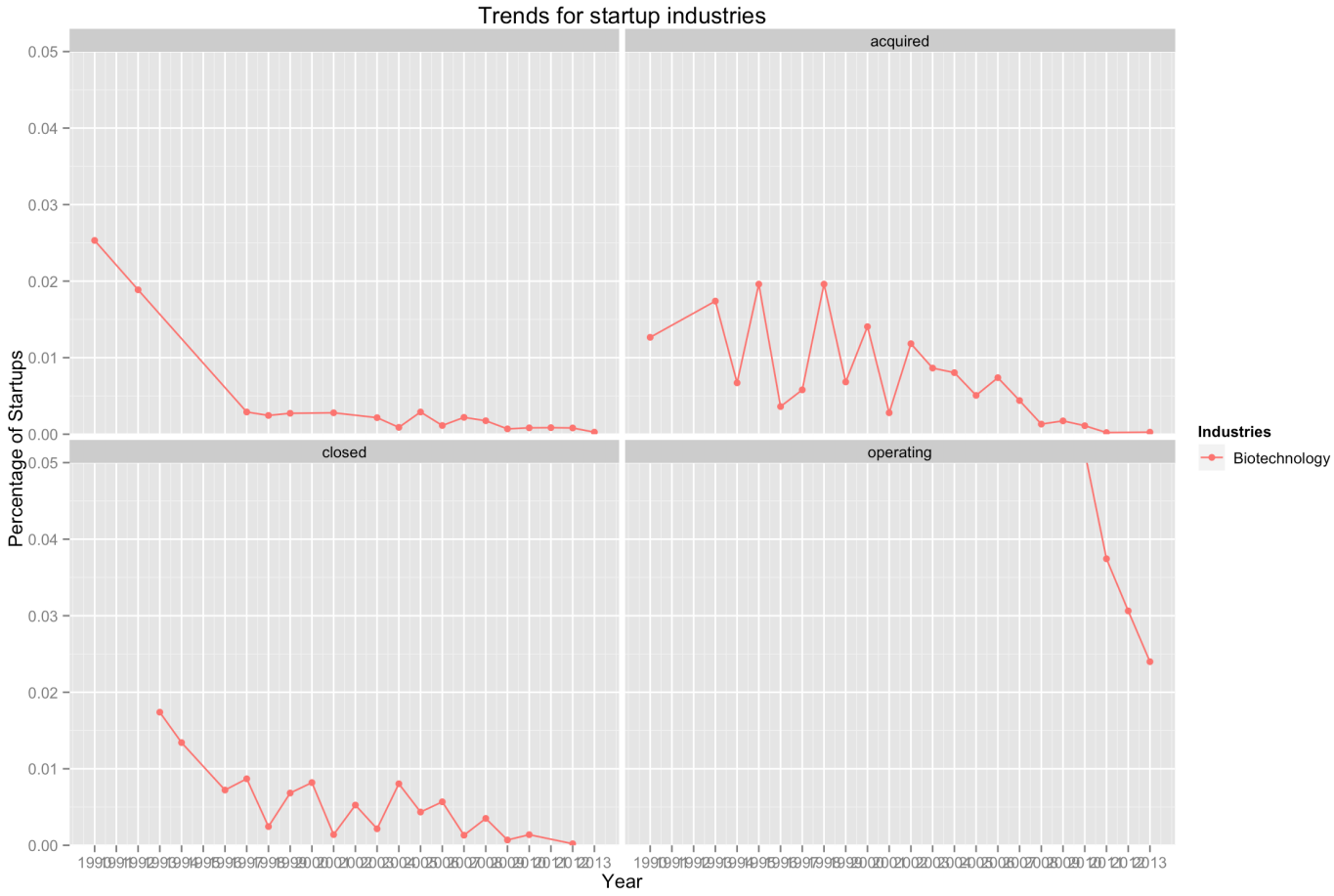
```
## [1] " Advertising "
```


Trends for startup industries
acquired
closed
operating
Percentage of Startups
Year
Industries
Advertising

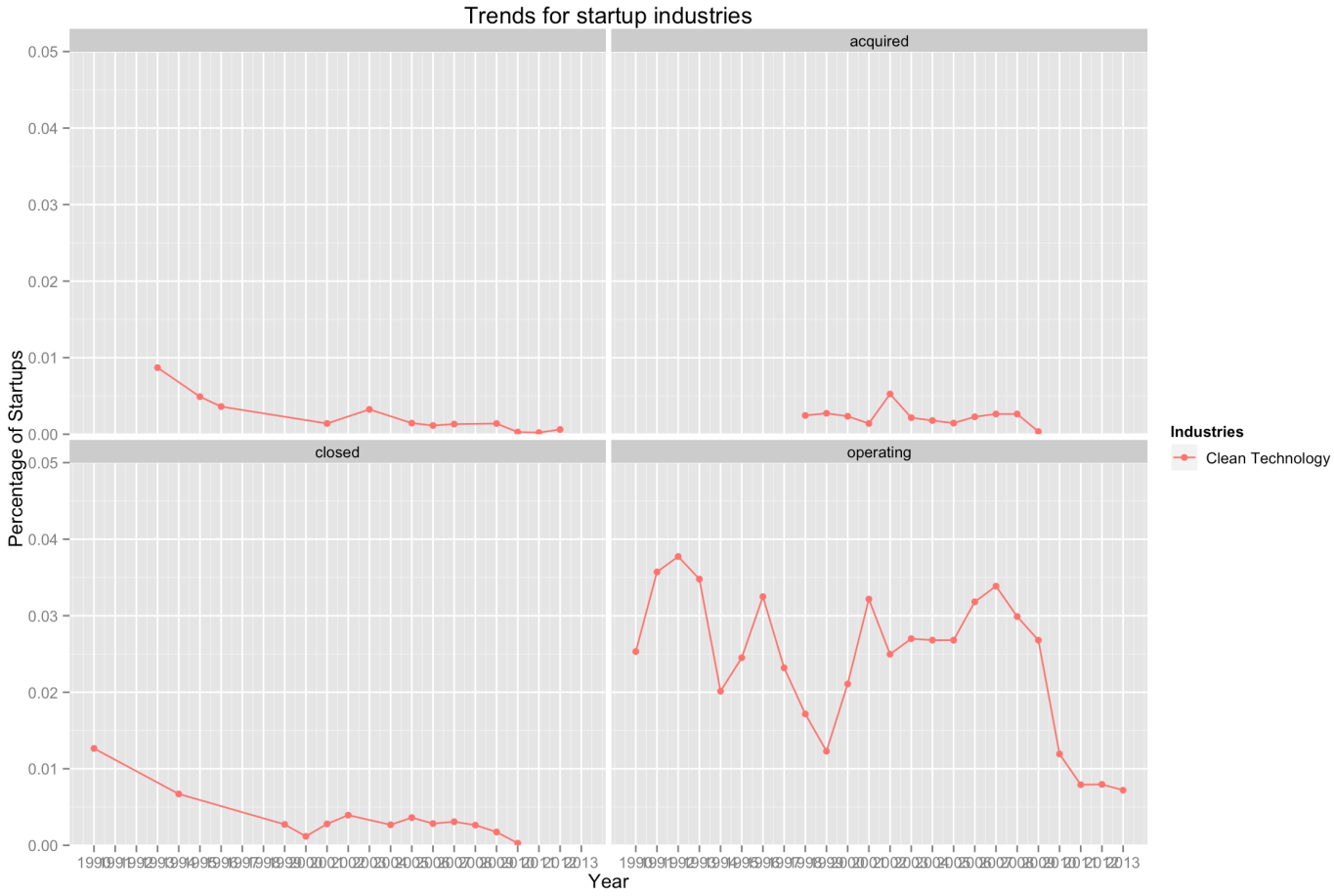
```
## [1] " Analytics "
```



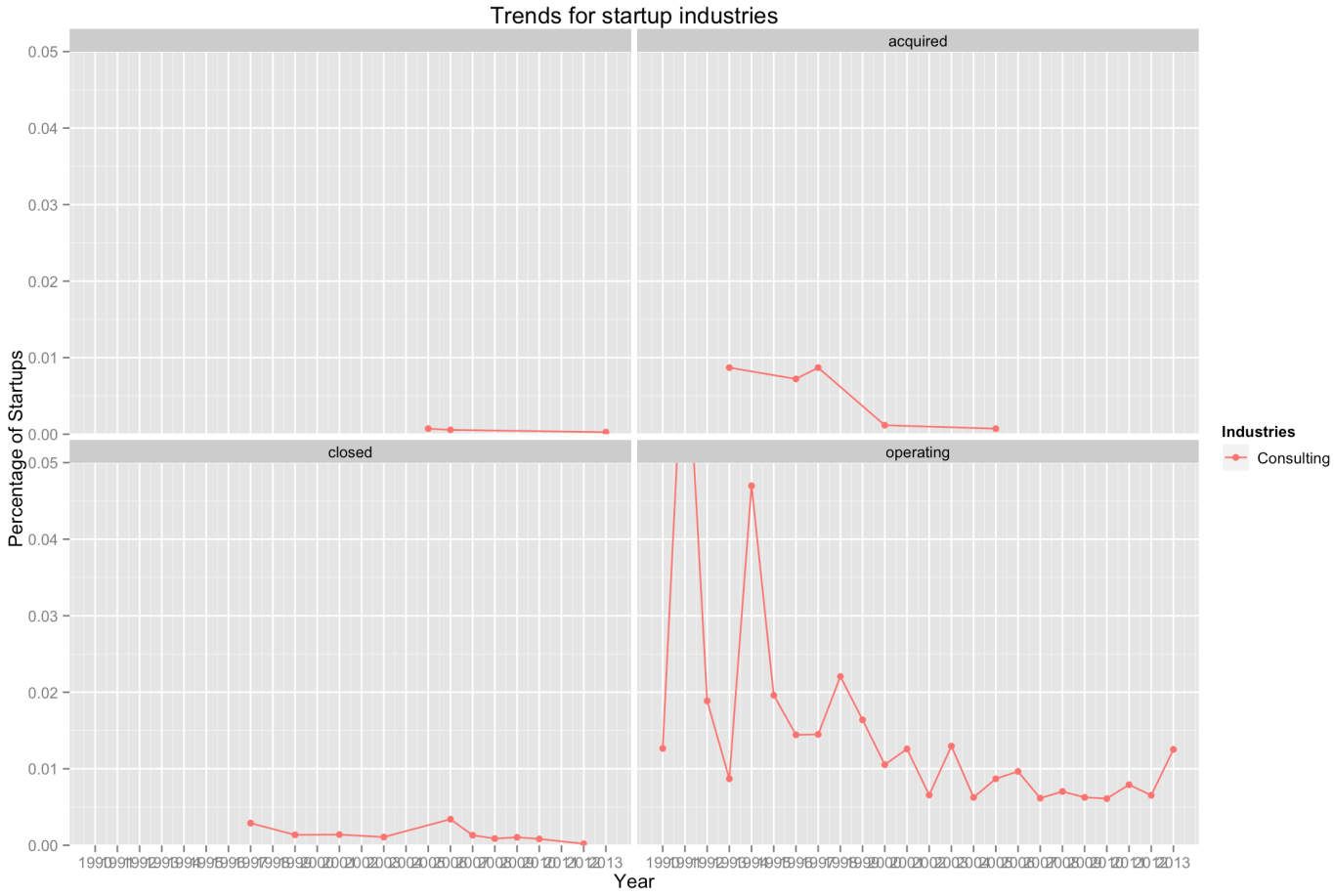

```
## [1] " Biotechnology "
```



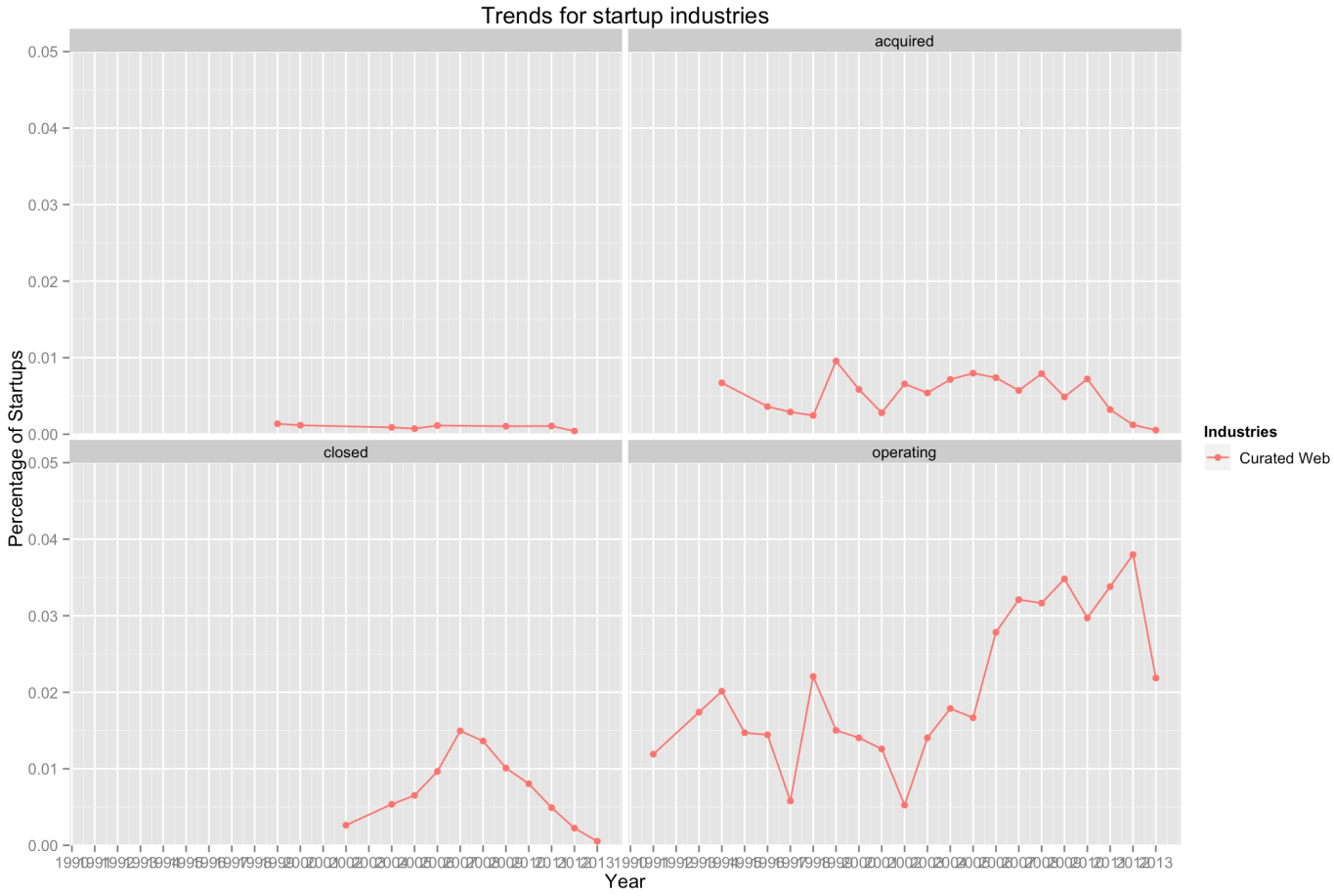

```
## [1] " Clean Technology "
```



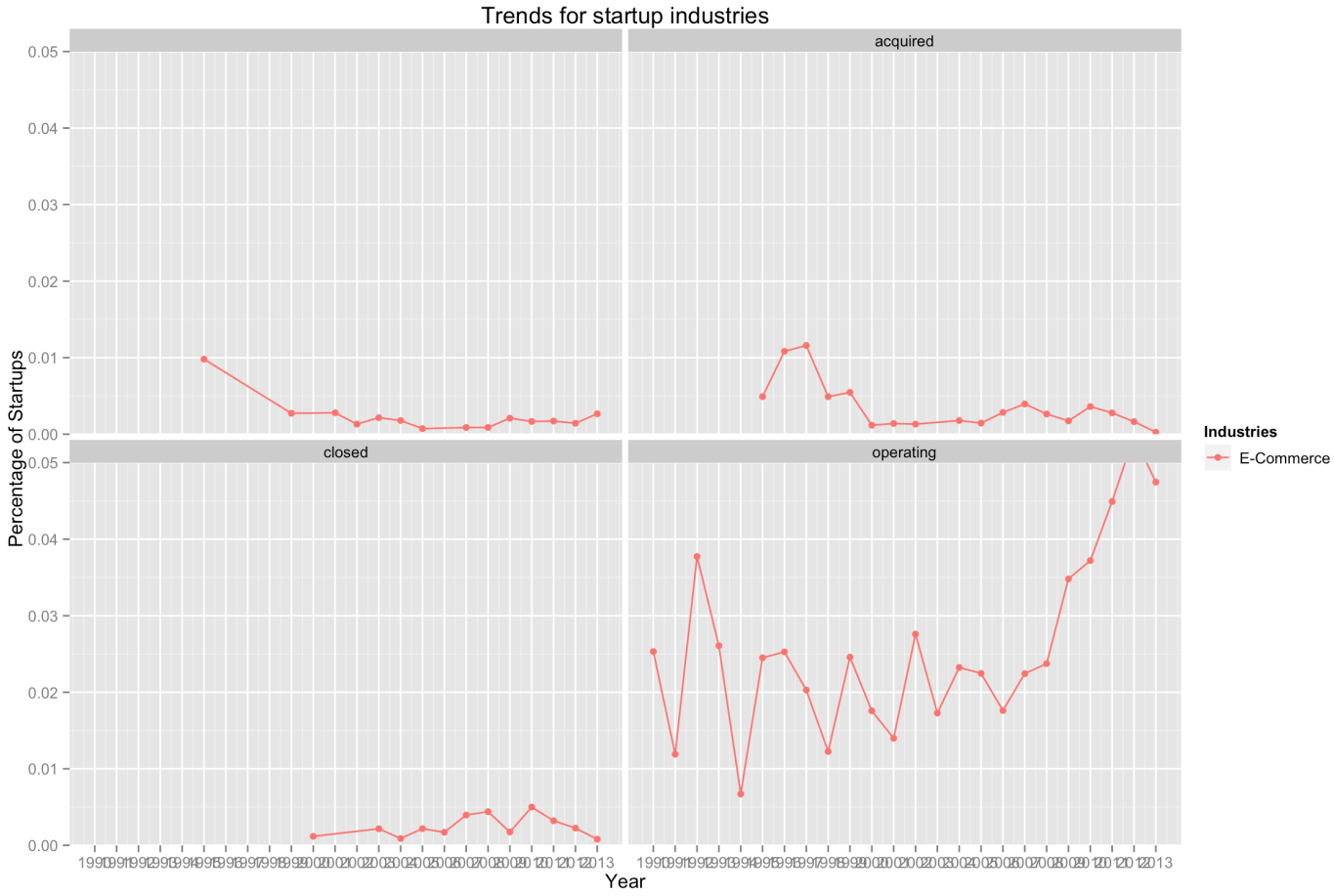

```
## [1] " Consulting "
```



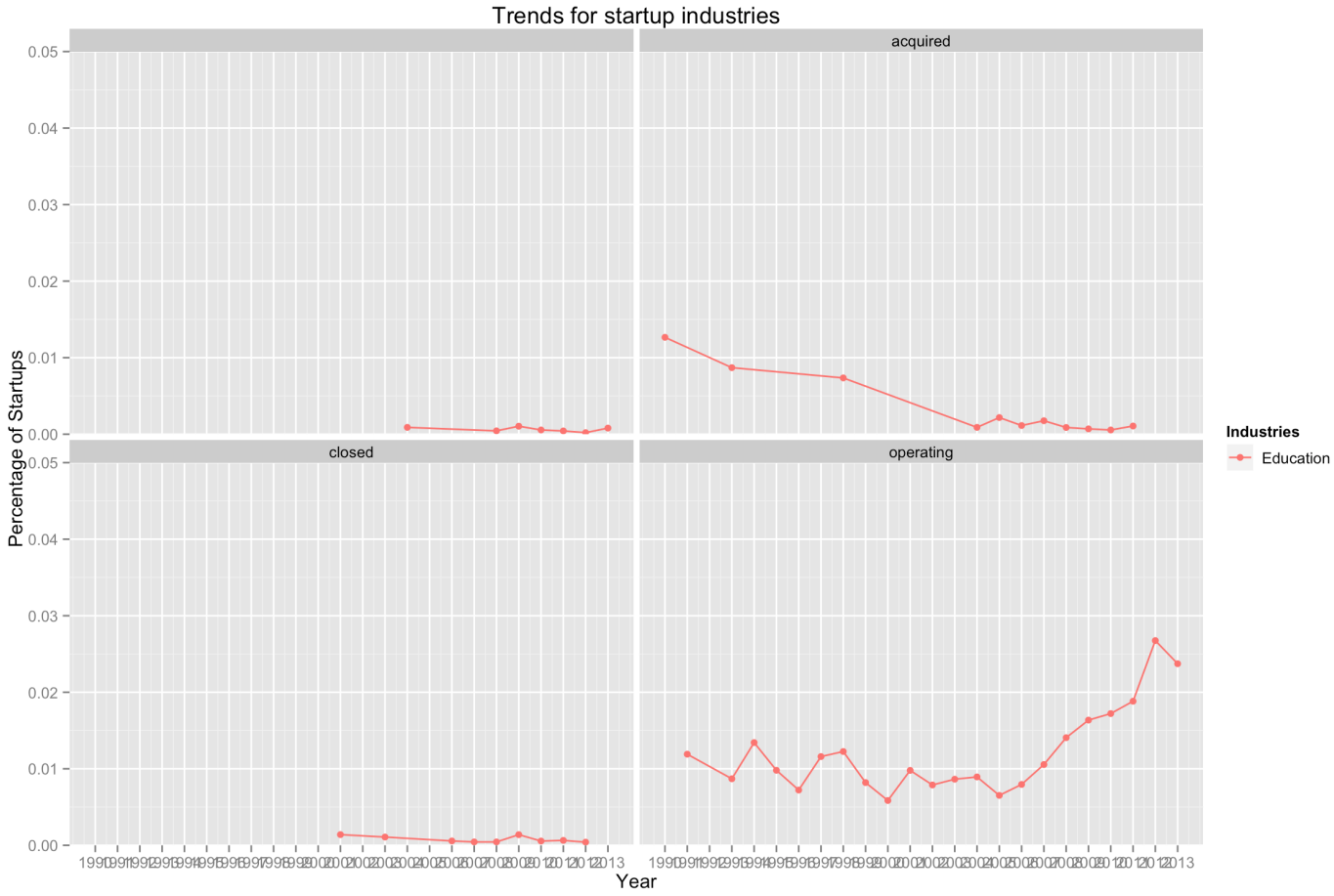

```
## [1] " Curated Web "
```


Trends for startup industries
acquired
closed
operating
Percentage of Startups
Year
Industries
Curated Web

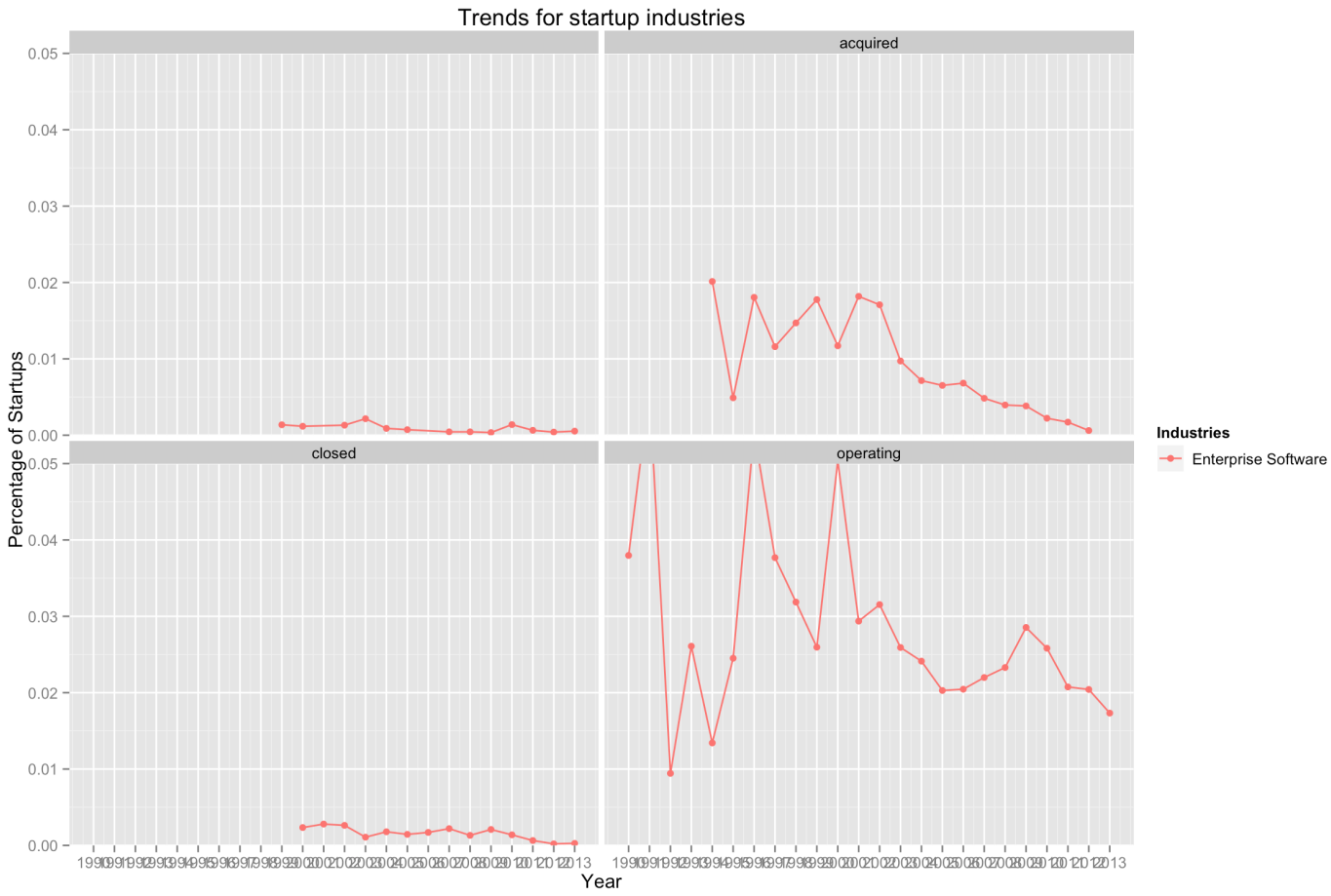
```
## [1] " E-Commerce "
```


Trends for startup industries
acquired
closed
operating
Percentage of Startups
Year
Industries
E-Commerce

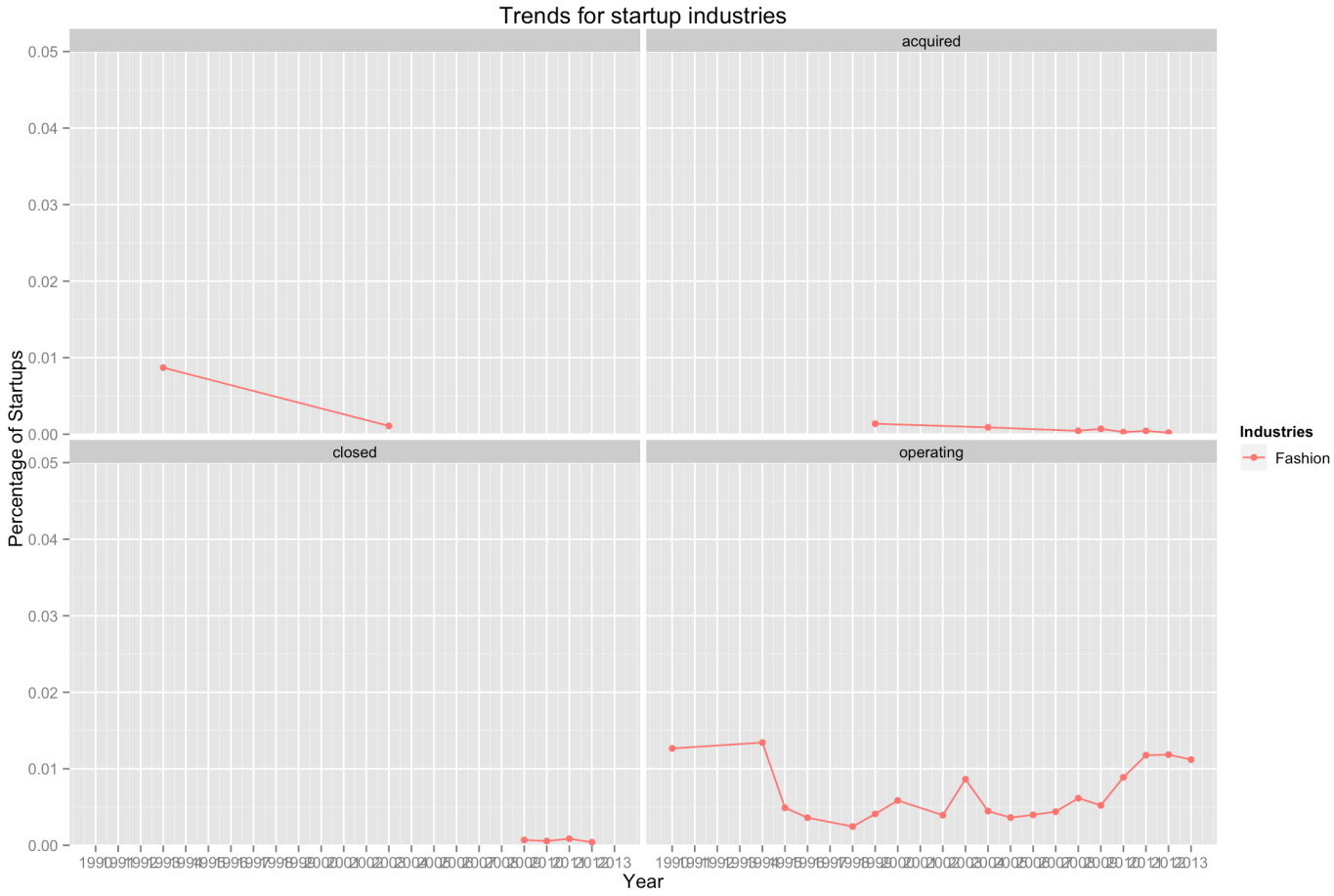
```
## [1] " Education "
```



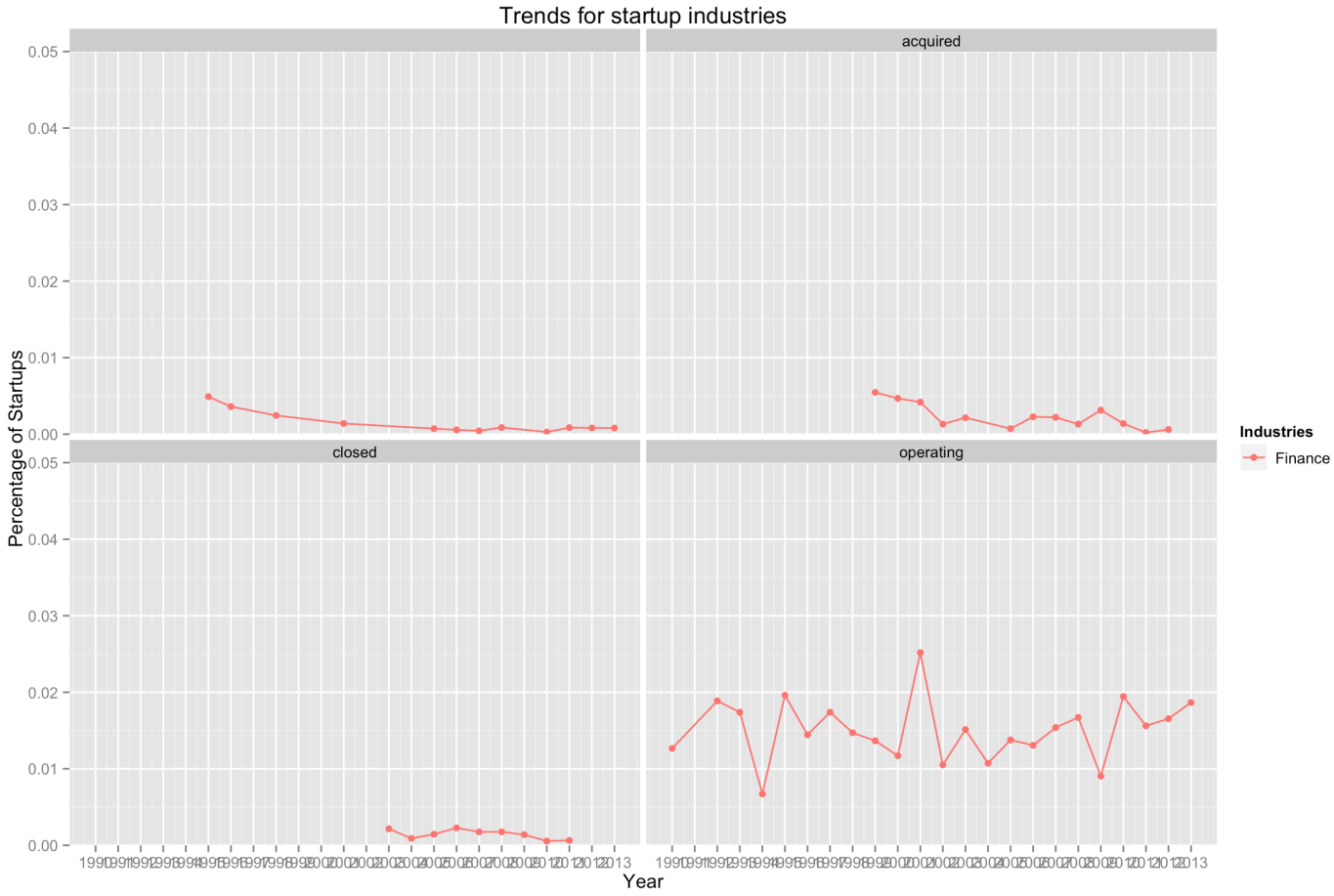

```
## [1] " Enterprise Software "
```


Trends for startup industries
acquired
closed
operating
Percentage of Startups
Year
Industries
Enterprise Software

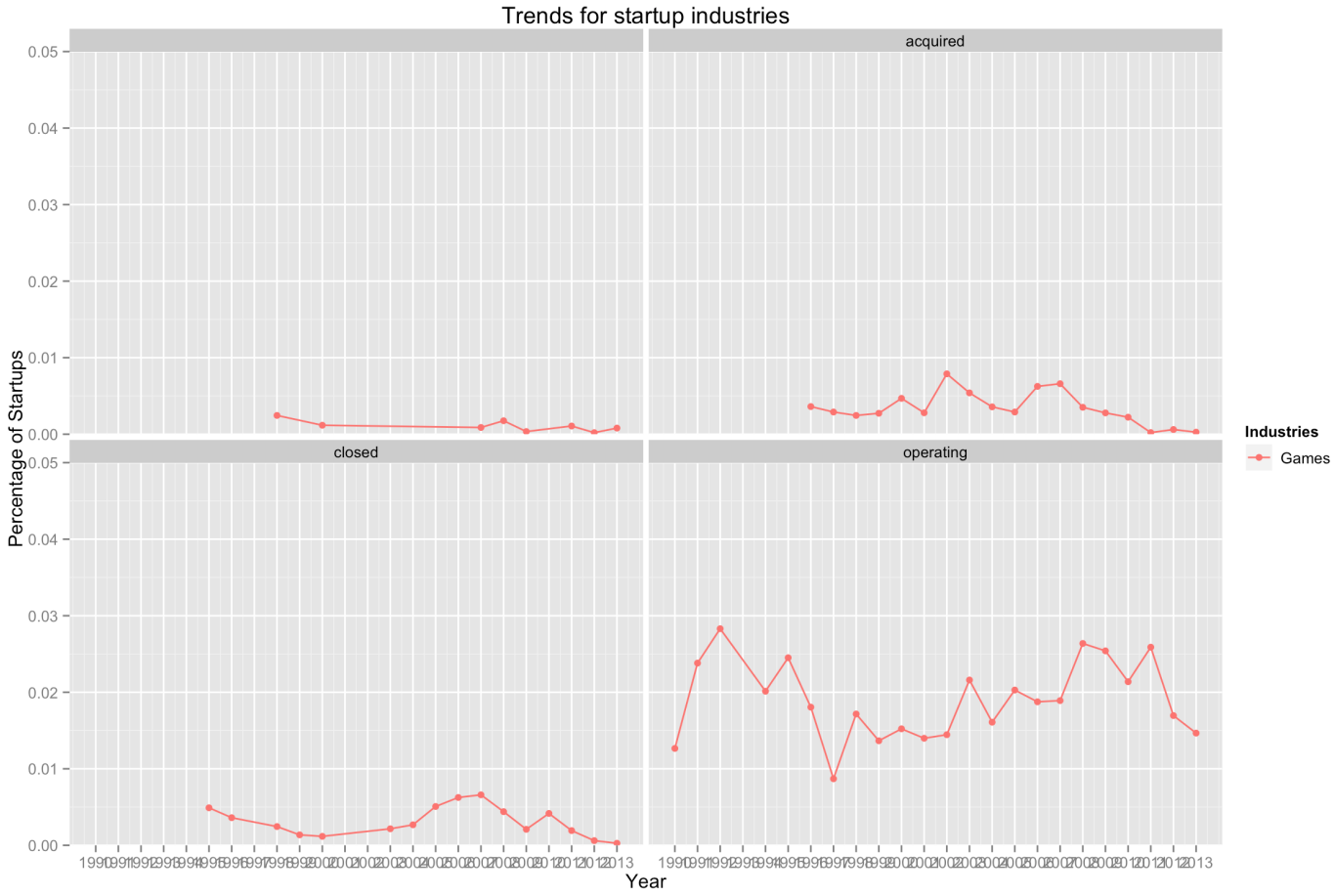
```
## [1] " Fashion "
```


Trends for startup industries
acquired
closed
operating
Percentage of Startups
Year
Industries
Fashion

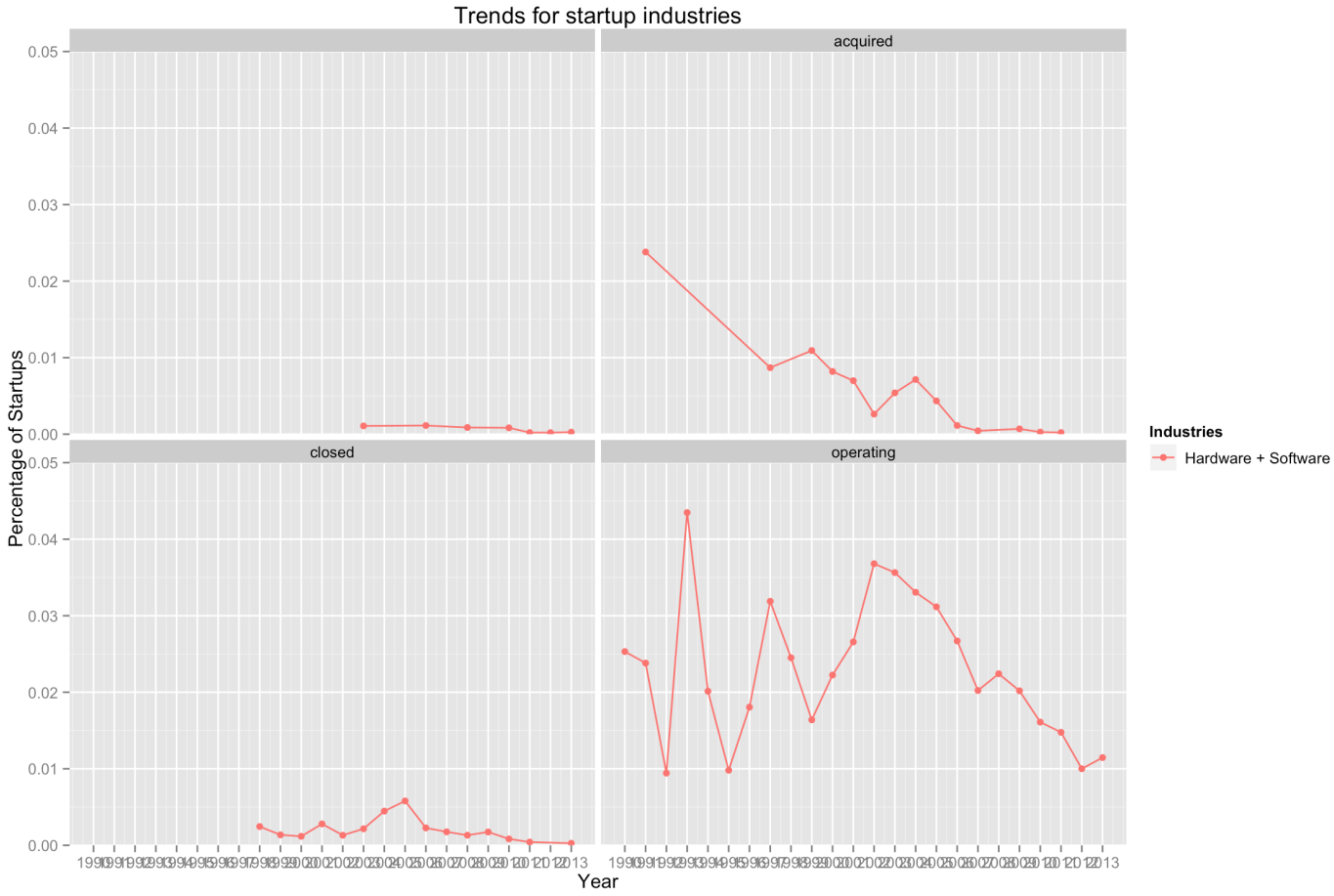
```
## [1] " Finance "
```


Trends for startup industries
acquired
closed
operating
Percentage of Startups
0.05
0.04
0.03
0.02
0.01
0.00
Year
Industries
Finance

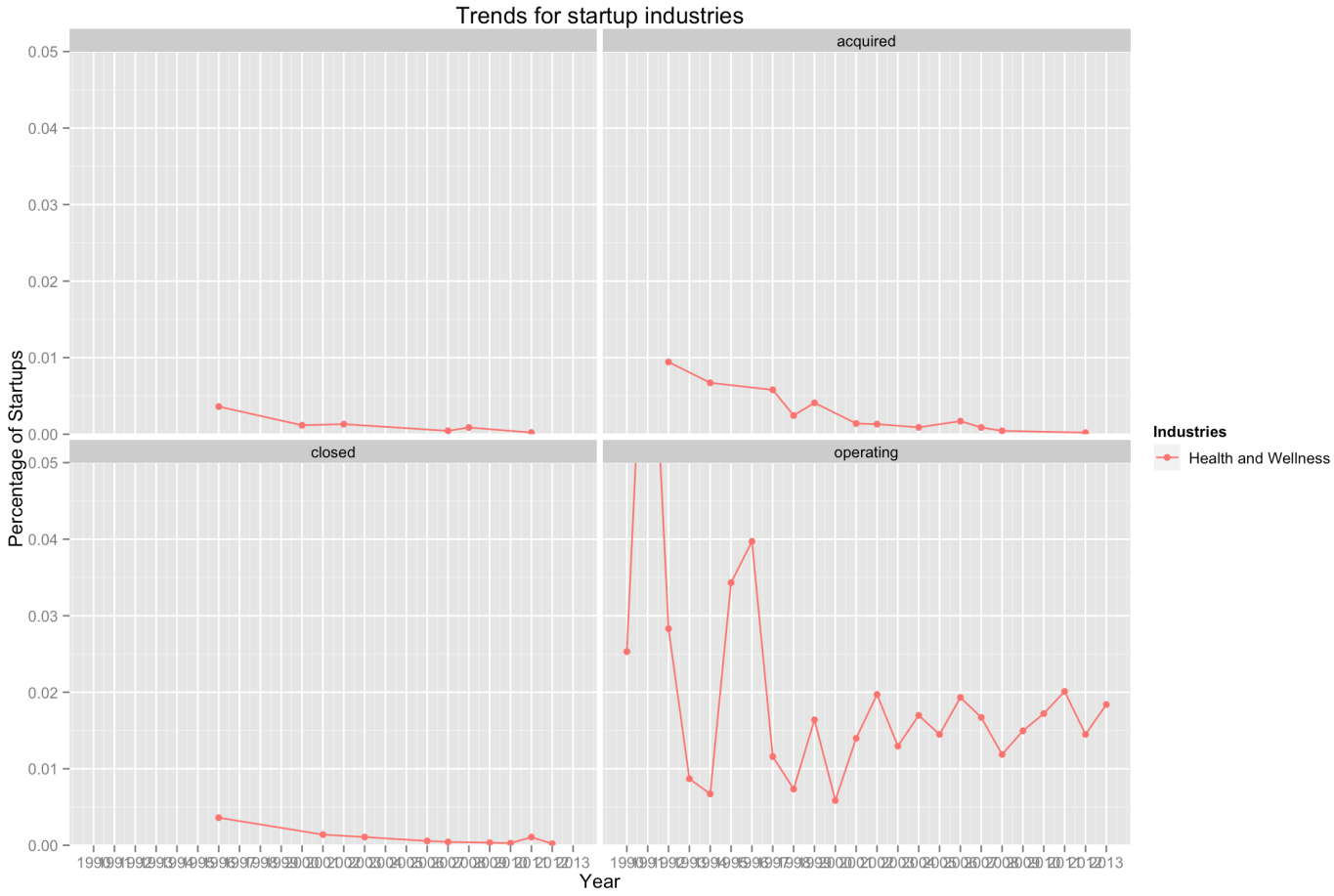
```
## [1] " Games "
```



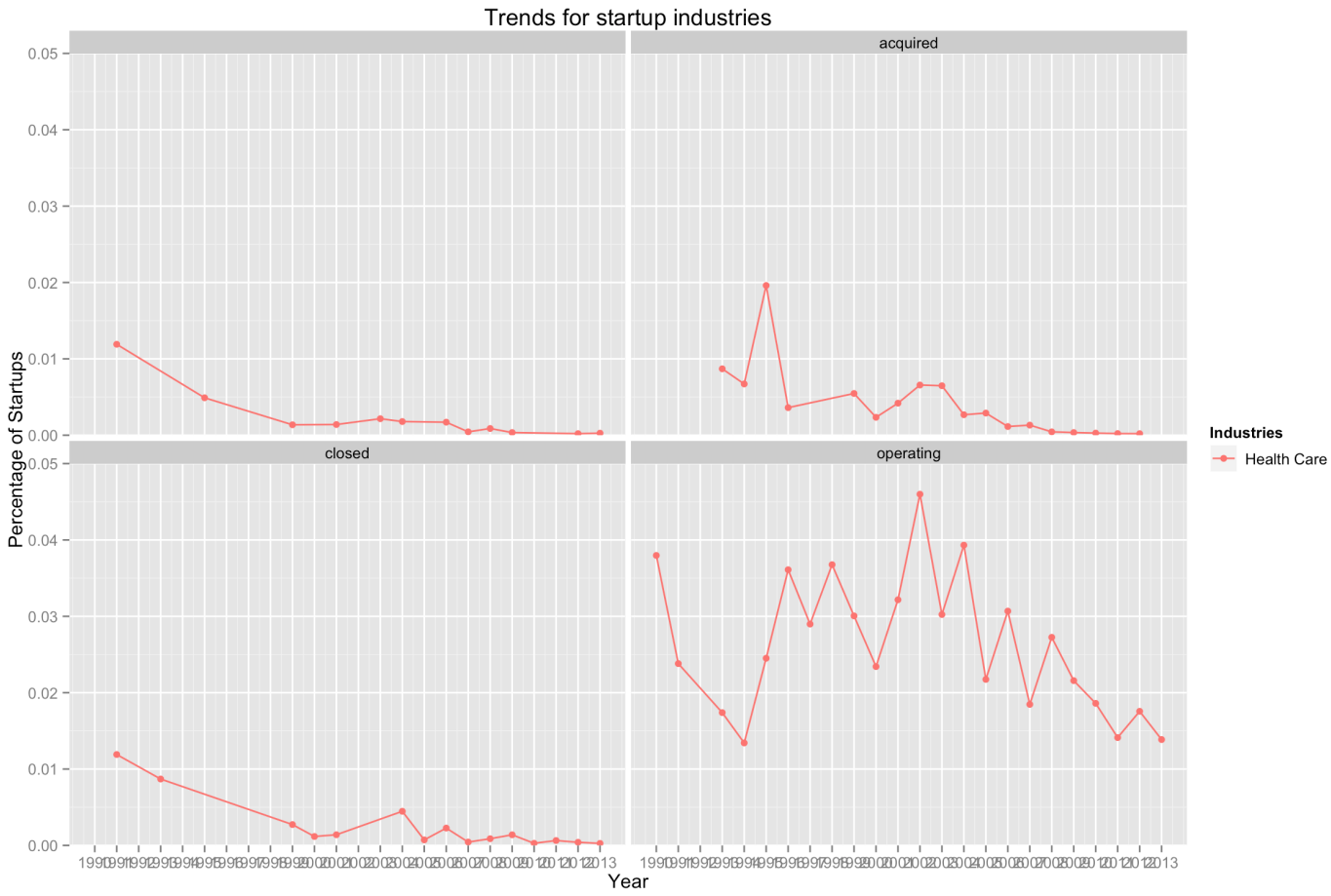

```
## [1] " Hardware + Software "
```



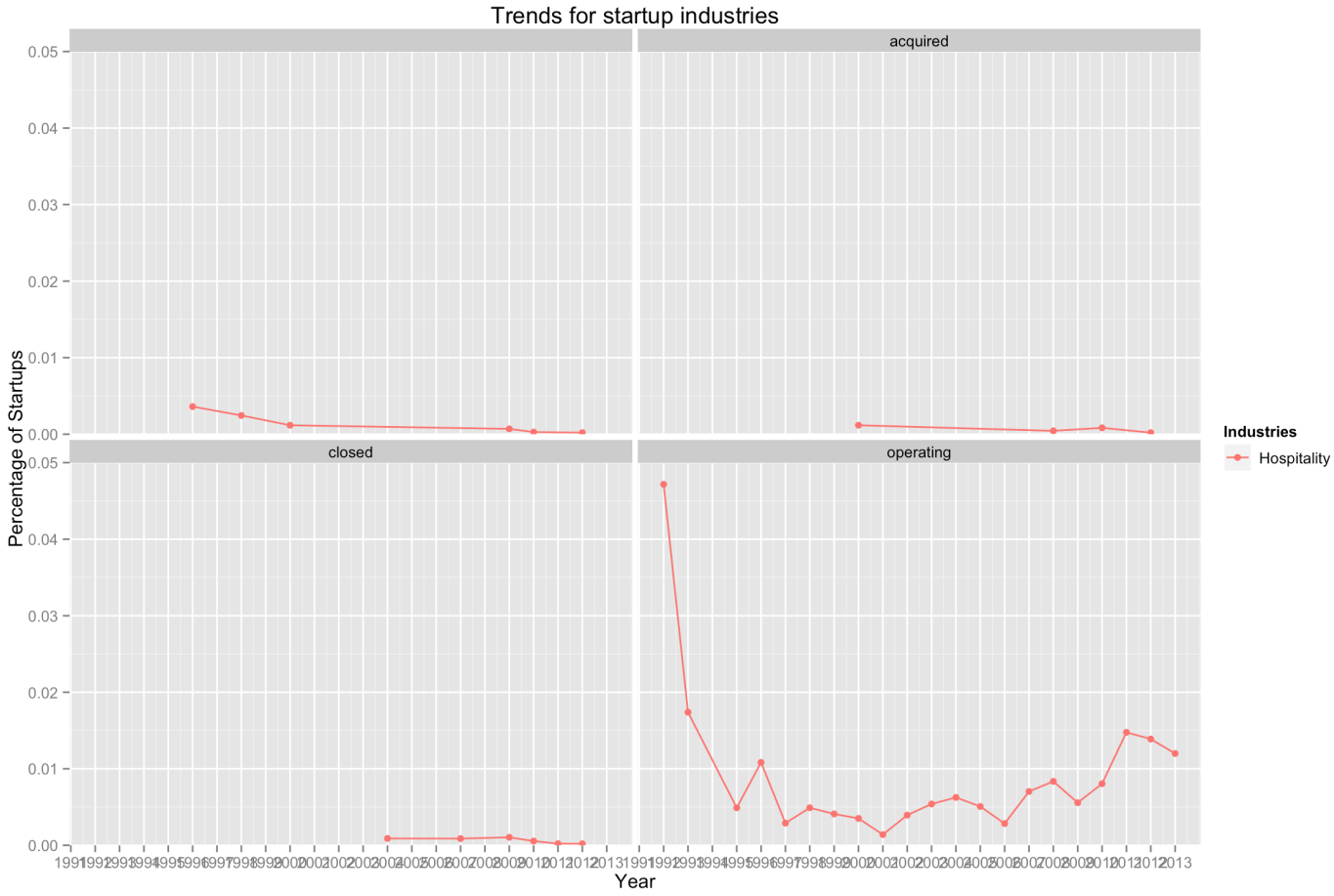

```
## [1] " Health and Wellness "
```



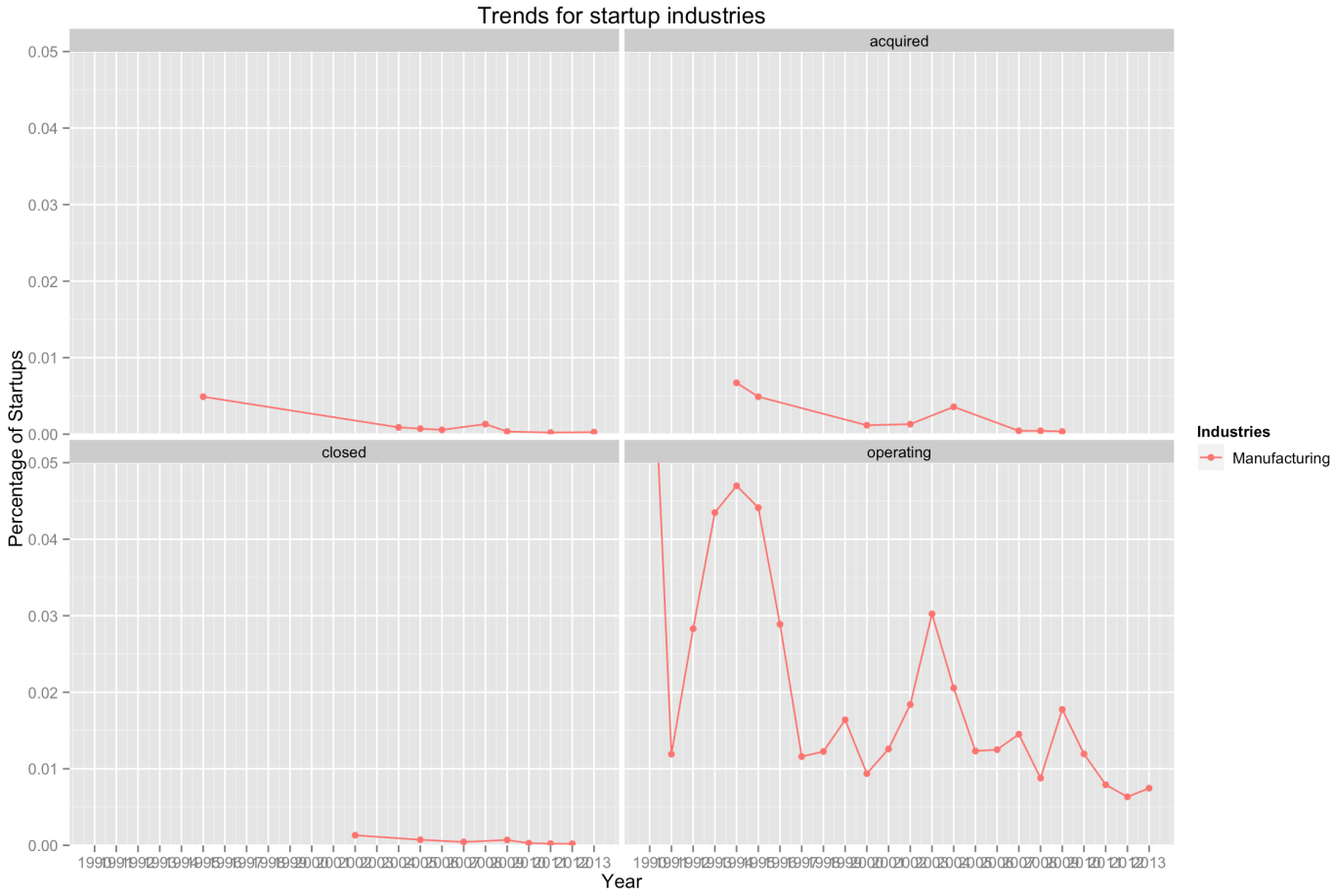

```
## [1] " Health Care "
```



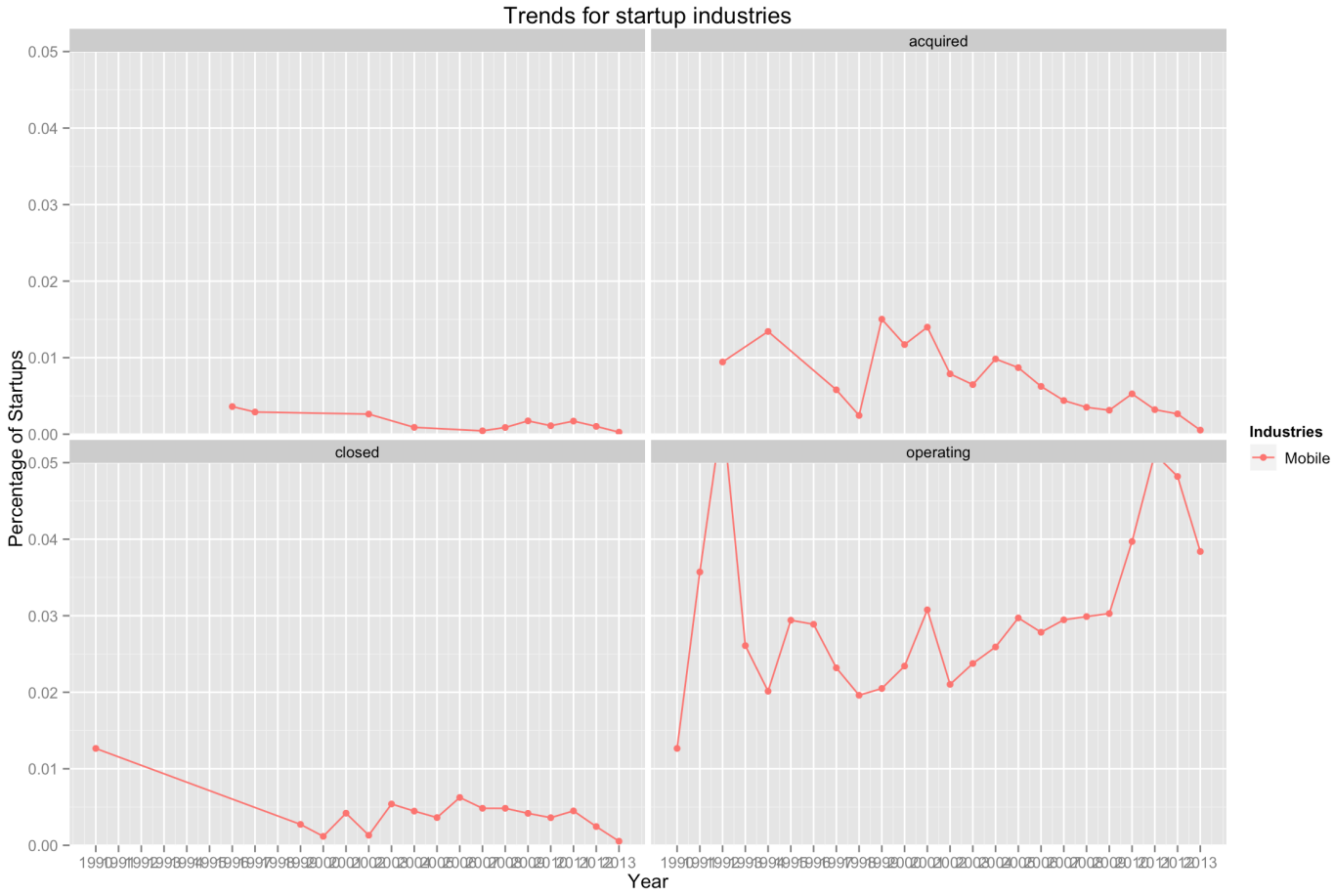

```
## [1] " Hospitality "
```



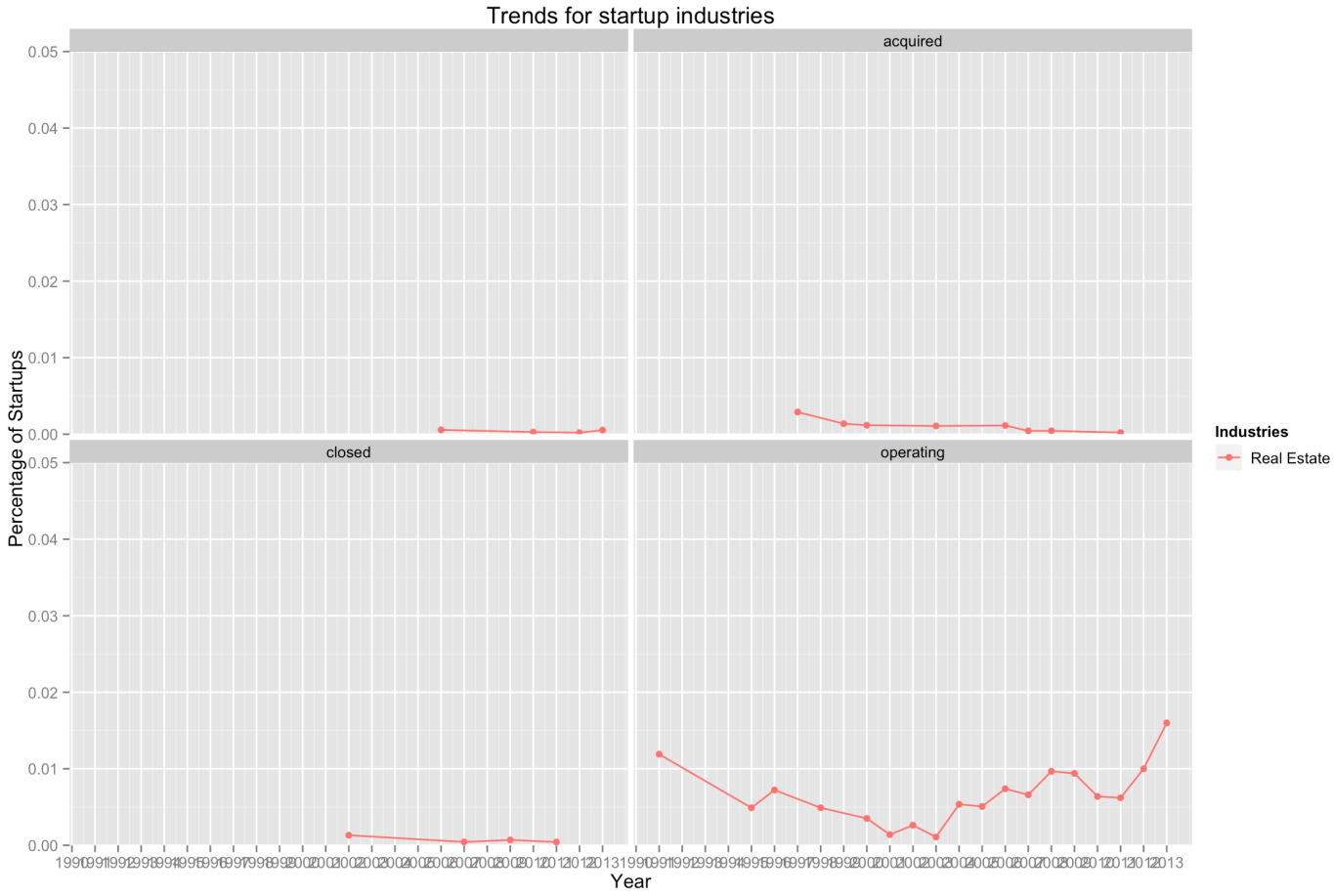

```
## [1] " Manufacturing "
```


Trends for startup industries
acquired
closed
operating
Percentage of Startups
Year
Industries
Manufacturing

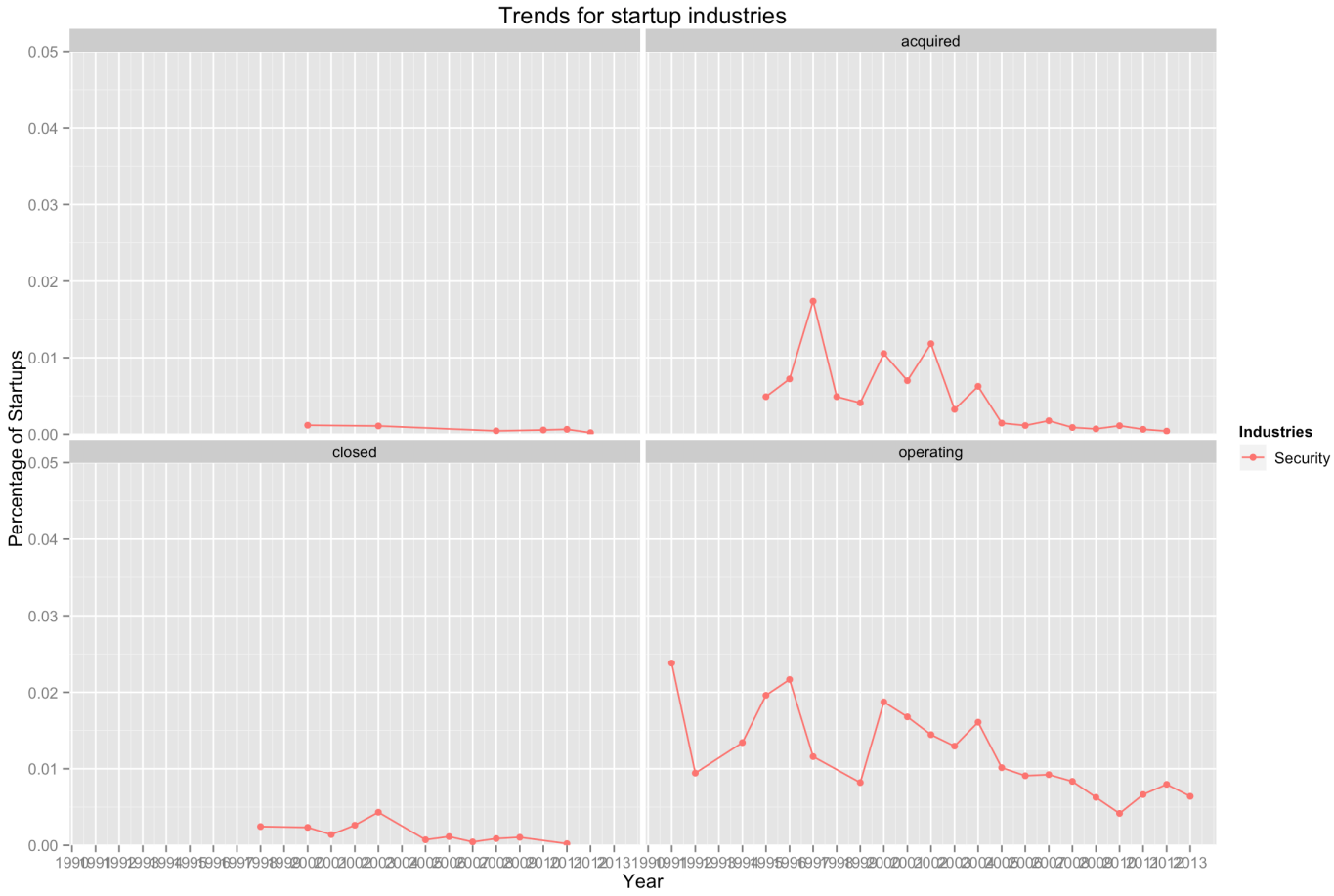
```
## [1] " Mobile "
```



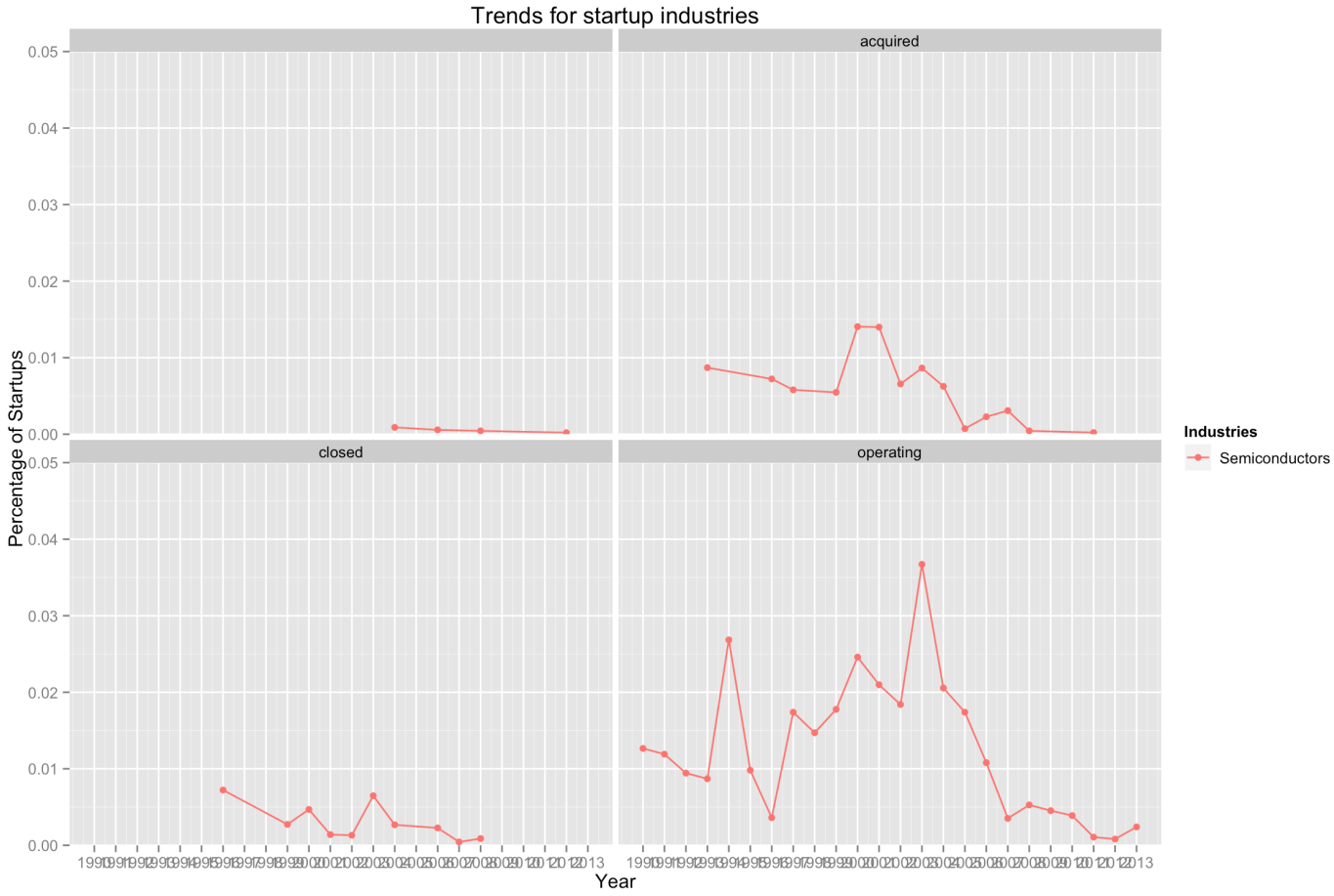

```
## [1] " Real Estate "
```


Trends for startup industries
acquired
closed
operating
Percentage of Startups
Year
Industries
Real Estate

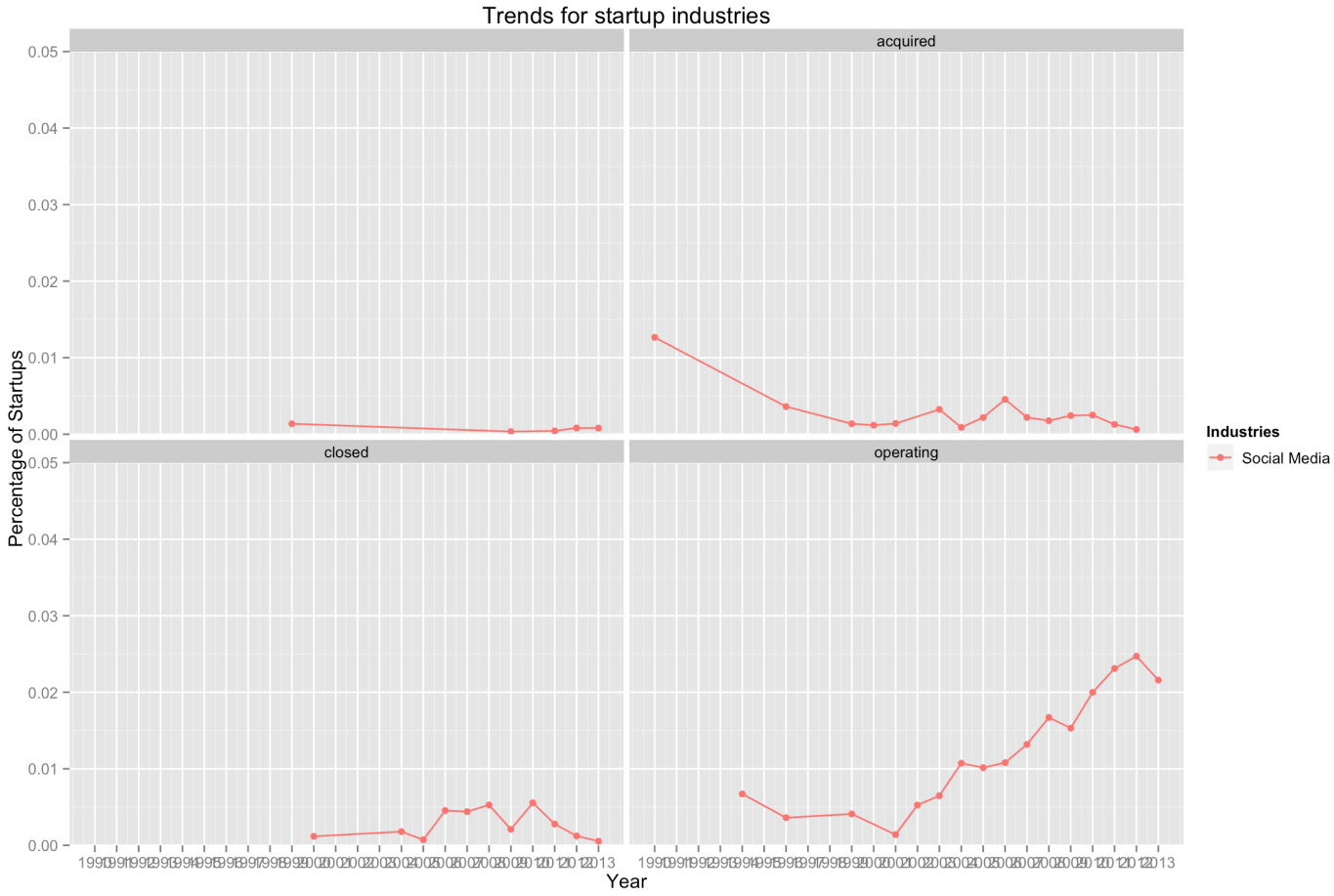
```
## [1] " Security "
```



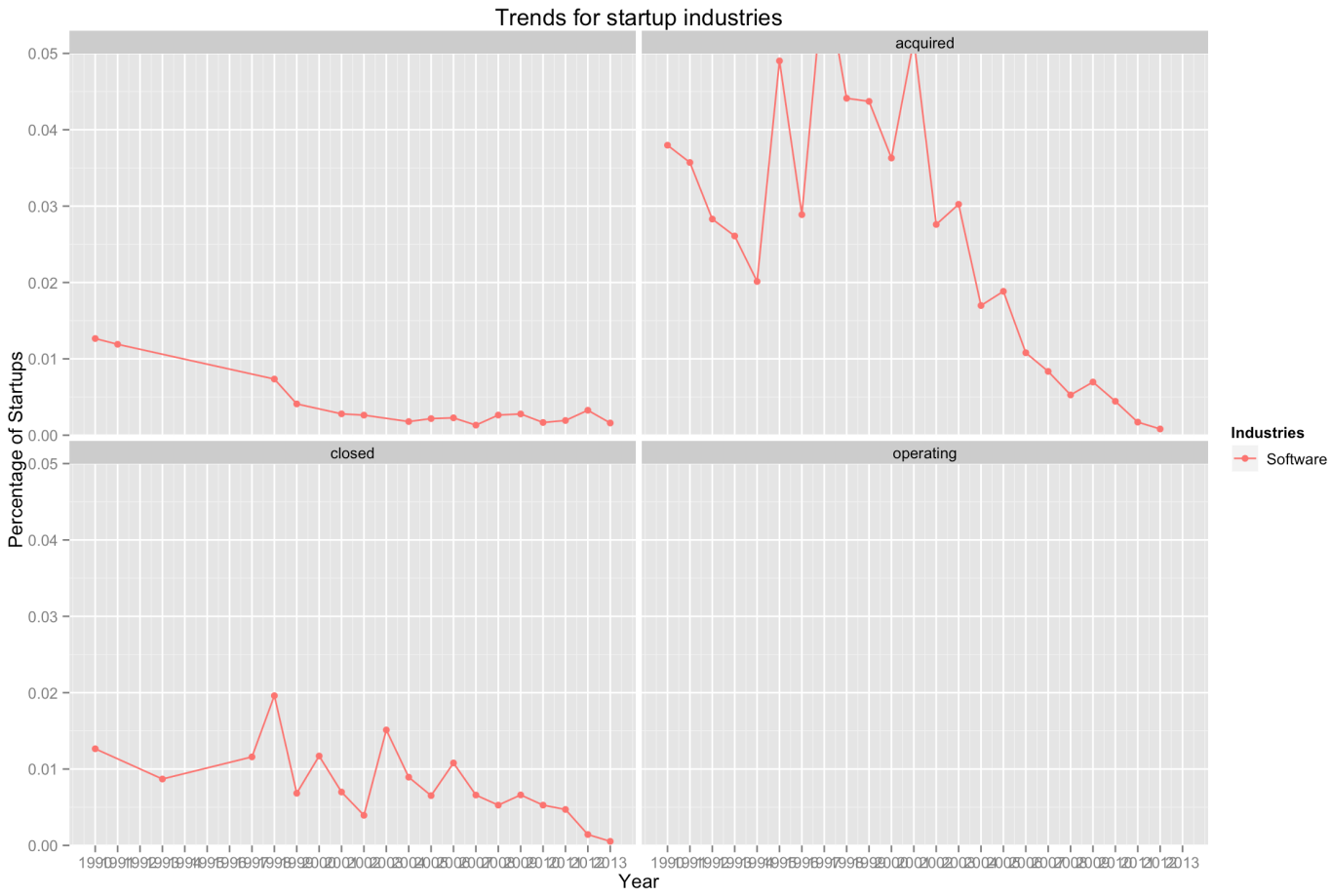

```
## [1] " Semiconductors "
```



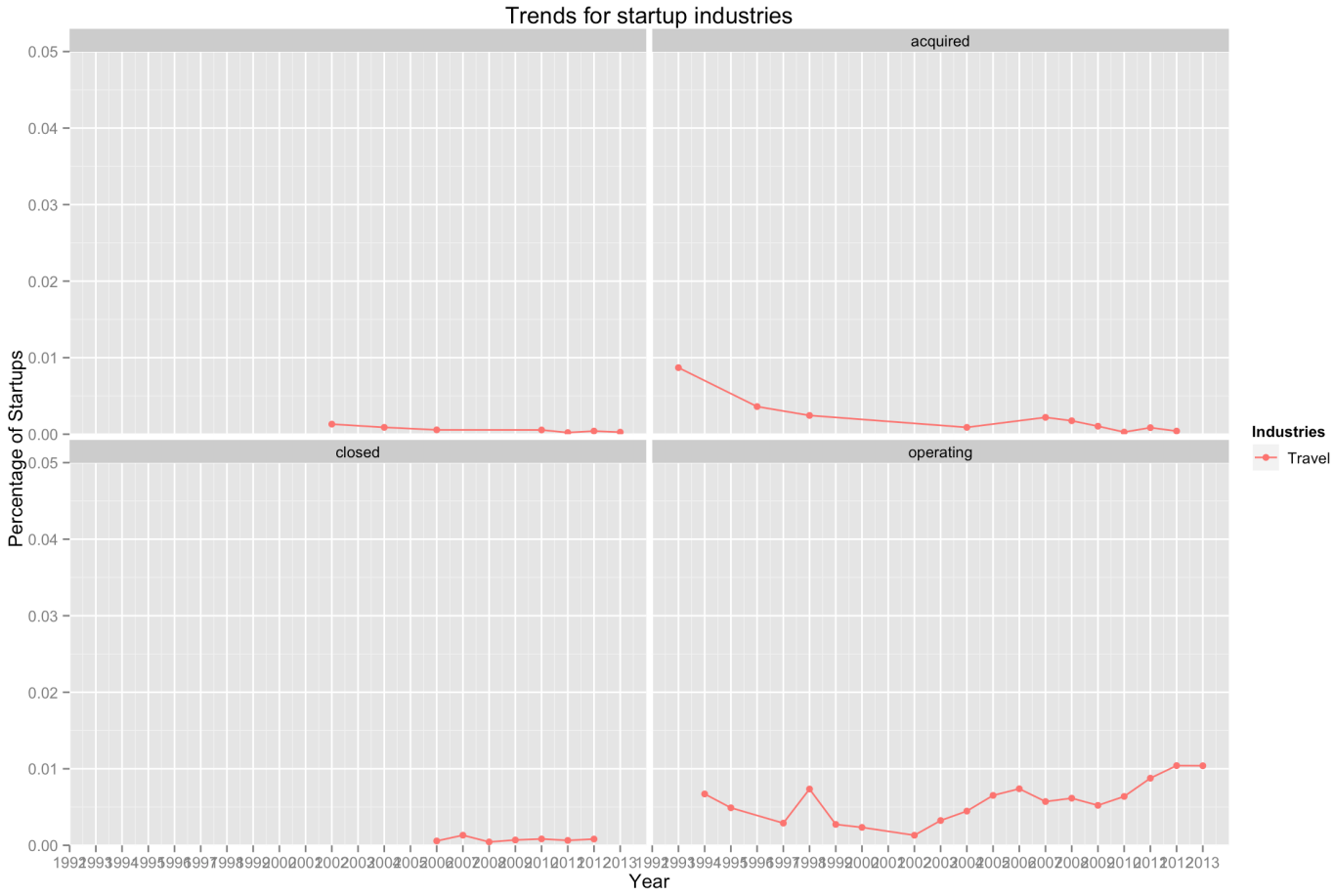

```
## [1] " Social Media "
```


Trends for startup industries
acquired
closed
operating
Percentage of Startups
0.05
0.04
0.03
0.02
0.01
0.00
Year
Industries
Social Media

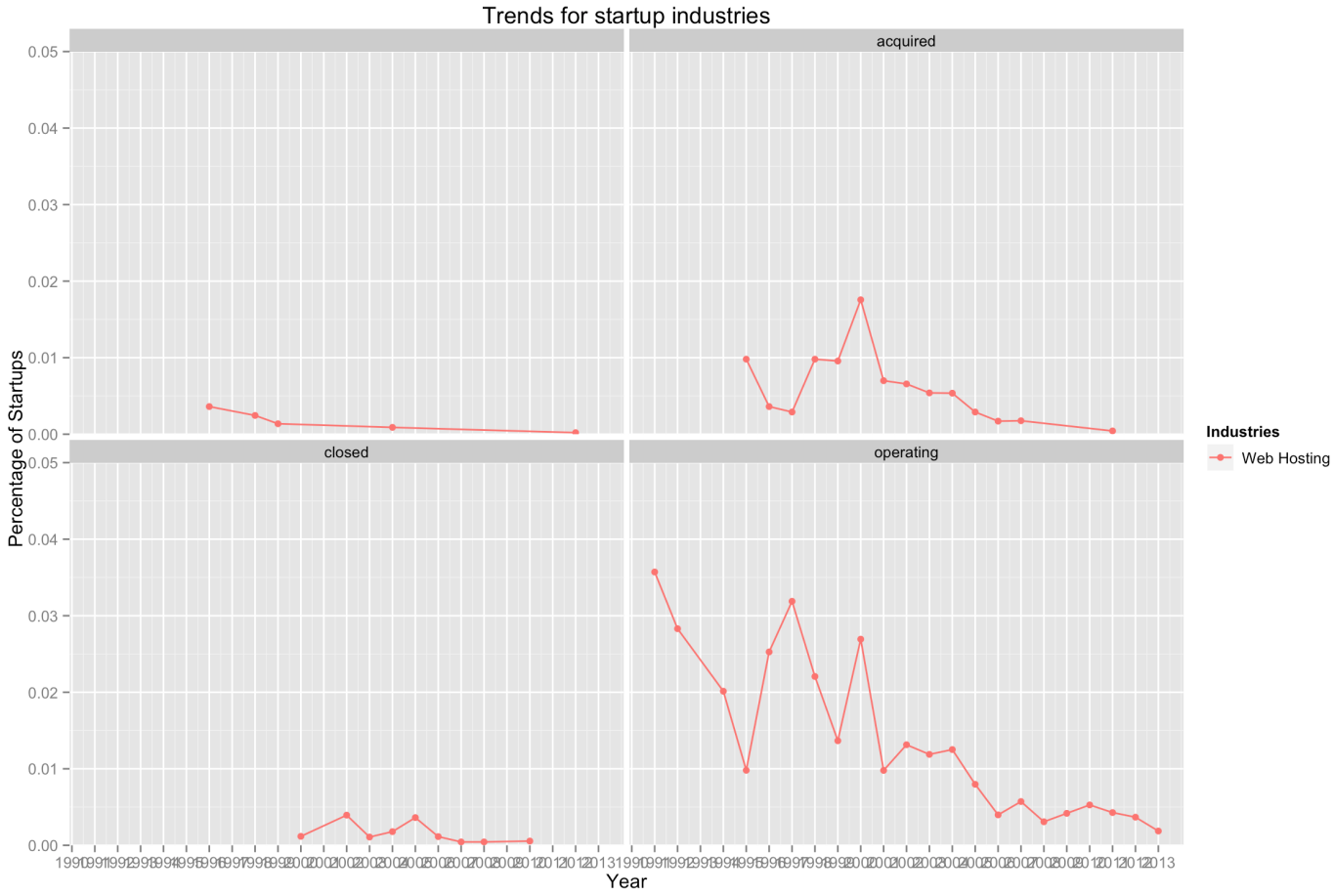
```
## [1] " Software "
```


Trends for startup industries
acquired
closed
operating
Percentage of Startups
0.05
0.04
0.03
0.02
0.01
0.00
Year
Industries
Software

```
## [1] " Travel "
```



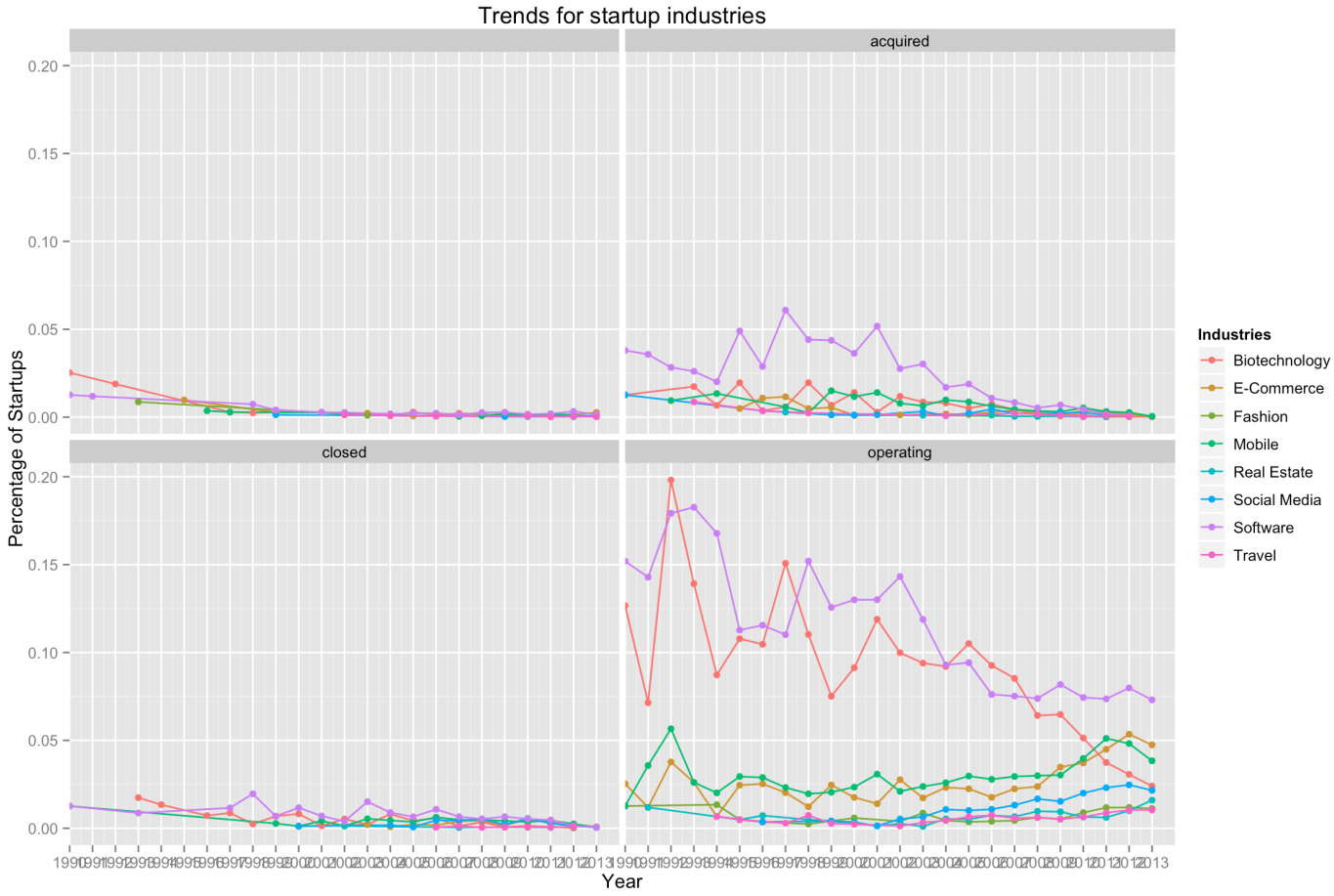

```
## [1] " Web Hosting "
```




```
# list hot industrie for startups with increased trends
other_markets <- c(" Real Estate ", " Travel ", " Fashion ", " Social Media ")
interesting_markets <- c(hot_markets, other_markets)
df.other.final <- subset(df.other, market %in% interesting_markets)
```

2.10 Fourth Explorarion - how does the percentage of startups change by industry across 1990 to 2013?

```
p3 <- ggplot(aes(x = founded_year, y = percentage), data = df.other.final) +
  geom_line(aes(color = market)) + geom_point(aes(color = market)) + facet_wrap(~s
tatus) +
  scale_x_continuous(breaks = seq(1990, 2013, 1)) +
  coord_cartesian(xlim = c(1990, 2014)) +
  labs(x = "Year", y = "Percentage of Startups", color = "Industries") +
  ggtitle("Trends for startup industries")
p3
```

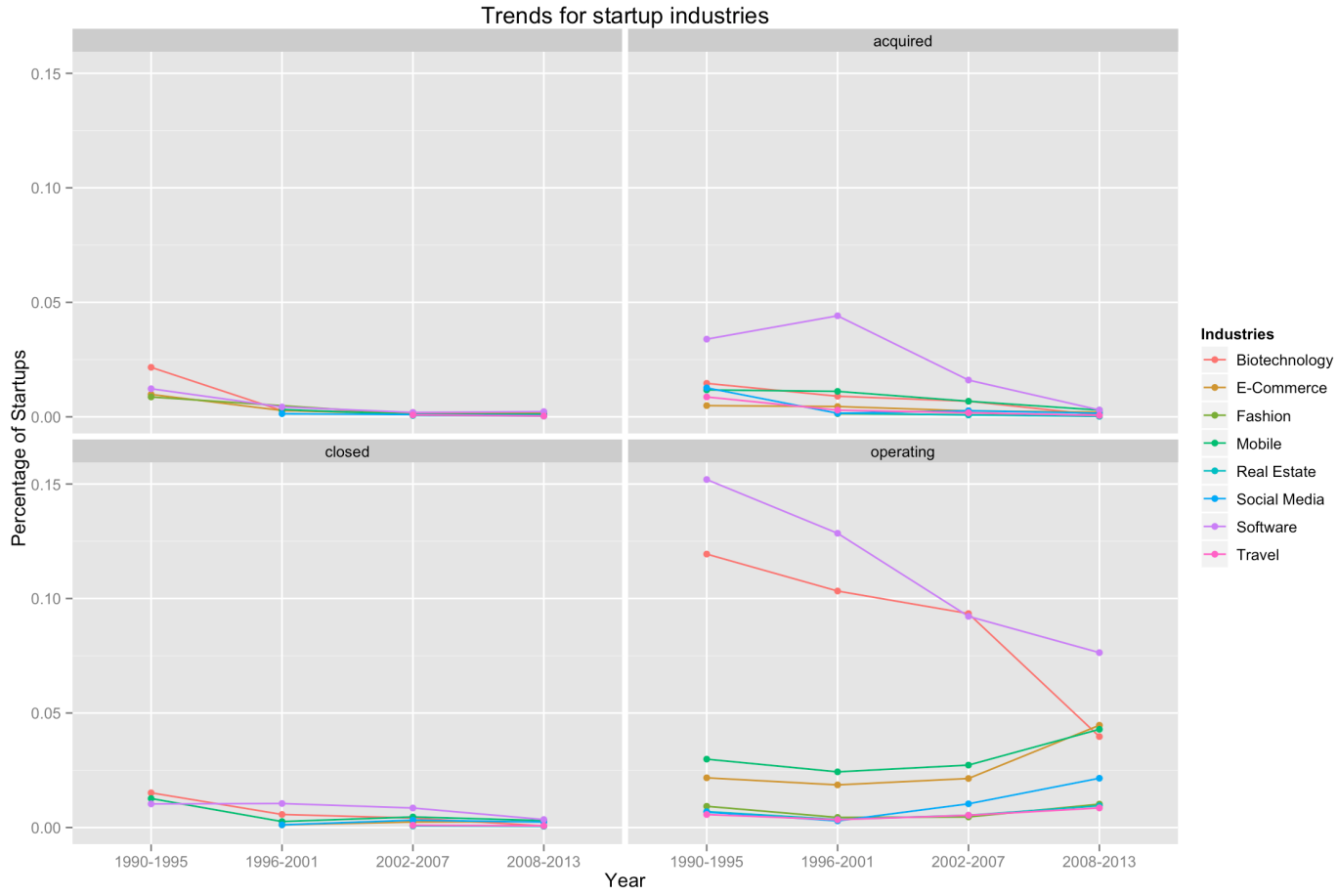


2.11 Group data by time period

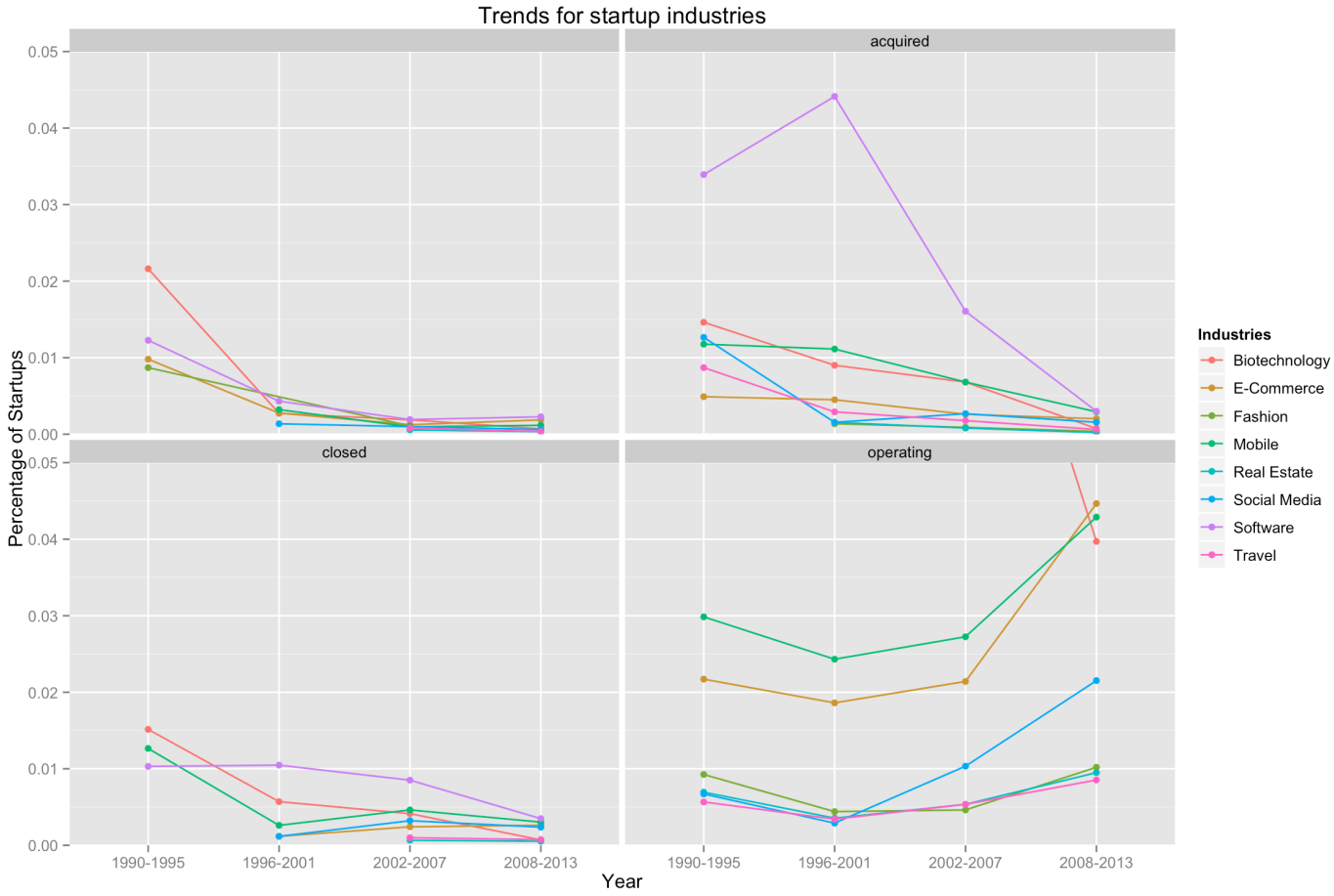
```
df.f1 <-subset(df.other.final, founded_year < 1996)
df.f1 <- df.f1 %>%
  group_by(market, status) %>%
  summarise (
    founded_year = "1990-1995",
    fre = sum(fre),
    tol = sum(tol),
    percentage = fre/tol,
    funding = sum(funding)
    )
df.f2 <- subset(df.other.final, founded_year > 1995 & founded_year < 2002)
df.f2 <- df.f2 %>%
  group_by(market, status) %>%
  summarise (
    founded_year = "1996-2001",
    fre = sum(fre),
    tol = sum(tol),
    percentage = fre/tol,
    funding = sum(funding)
    )
df.f3 <- subset(df.other.final, founded_year > 2001 & founded_year < 2008)
df.f3 <- df.f3 %>%
  group_by(market, status) %>%
  summarise (
    founded_year = "2002-2007",
    fre = sum(fre),
    tol = sum(tol),
    percentage = fre/tol,
    funding = sum(funding)
    )
df.f4 <- subset(df.other.final, founded_year > 2008)
df.f4 <- df.f4 %>%
  group_by(market, status) %>%
  summarise (
    founded_year = "2008-2013",
    fre = sum(fre),
    tol = sum(tol),
    percentage = fre/tol,
    funding = sum(funding)
    )
match.by <- c("founded_year", "market", "status", "funding","fre", "tol", "percent
age")
df.list <- list(df.f4, df.f3, df.f2, df.f1)
df <- Reduce(function(...) merge(..., by=match.by, all=T), df.list)
```

2.12 Fifth Exploration - how does the percentage of startups change by industry in different time period?

```
p4 <- ggplot(aes(x = founded_year, y = percentage), data = df) +
  geom_line(aes(color = market, group = market)) + geom_point(aes(color = market))
+
  facet_wrap(~status) +
  labs(x = "Year", y = "Percentage of Startups", color = "Industries") +
  ggtitle("Trends for startup industries")
p4
```

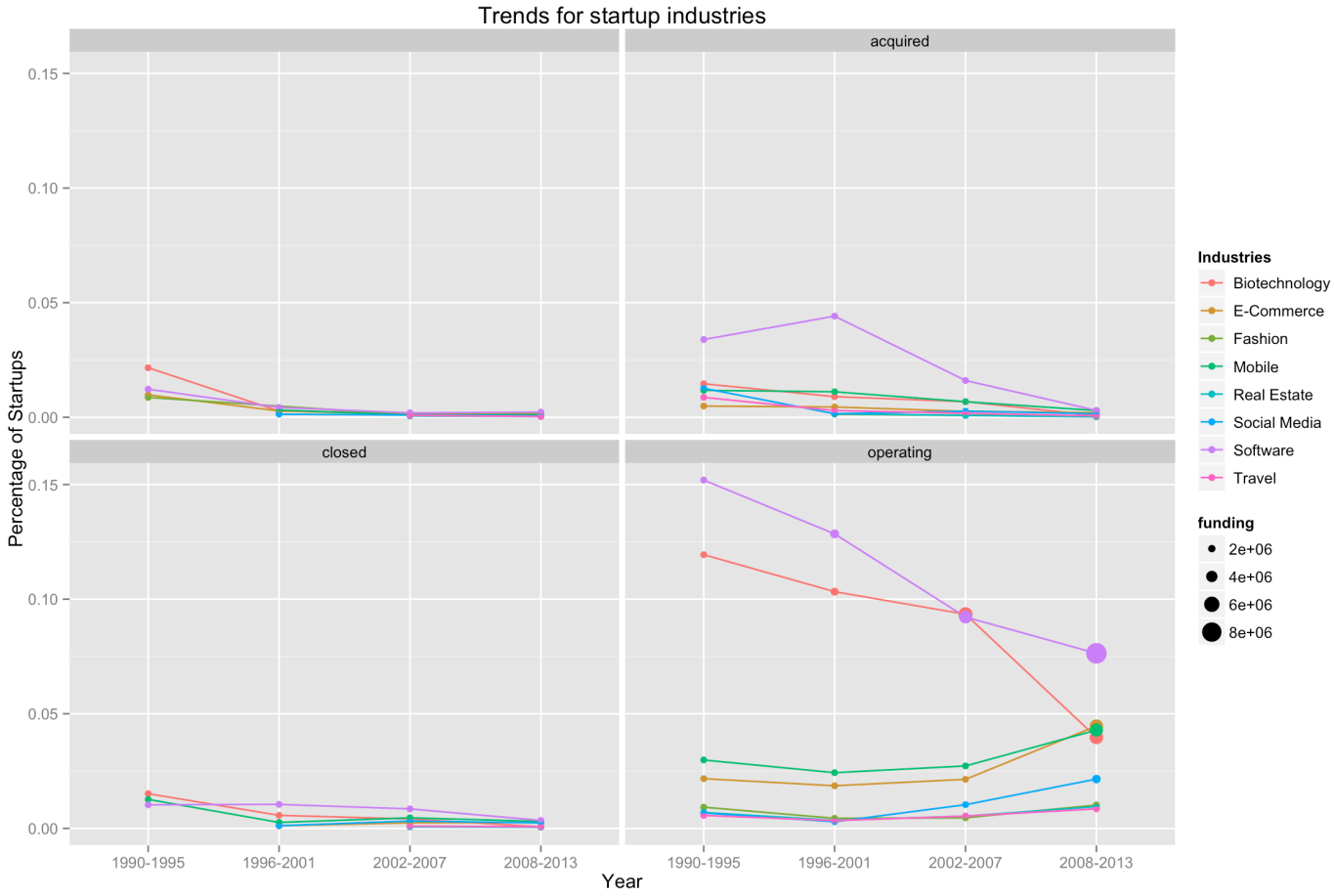



```
# adjust y axis to look at the trends closely
p4 + coord_cartesian(ylim = c(0, 0.05))
```



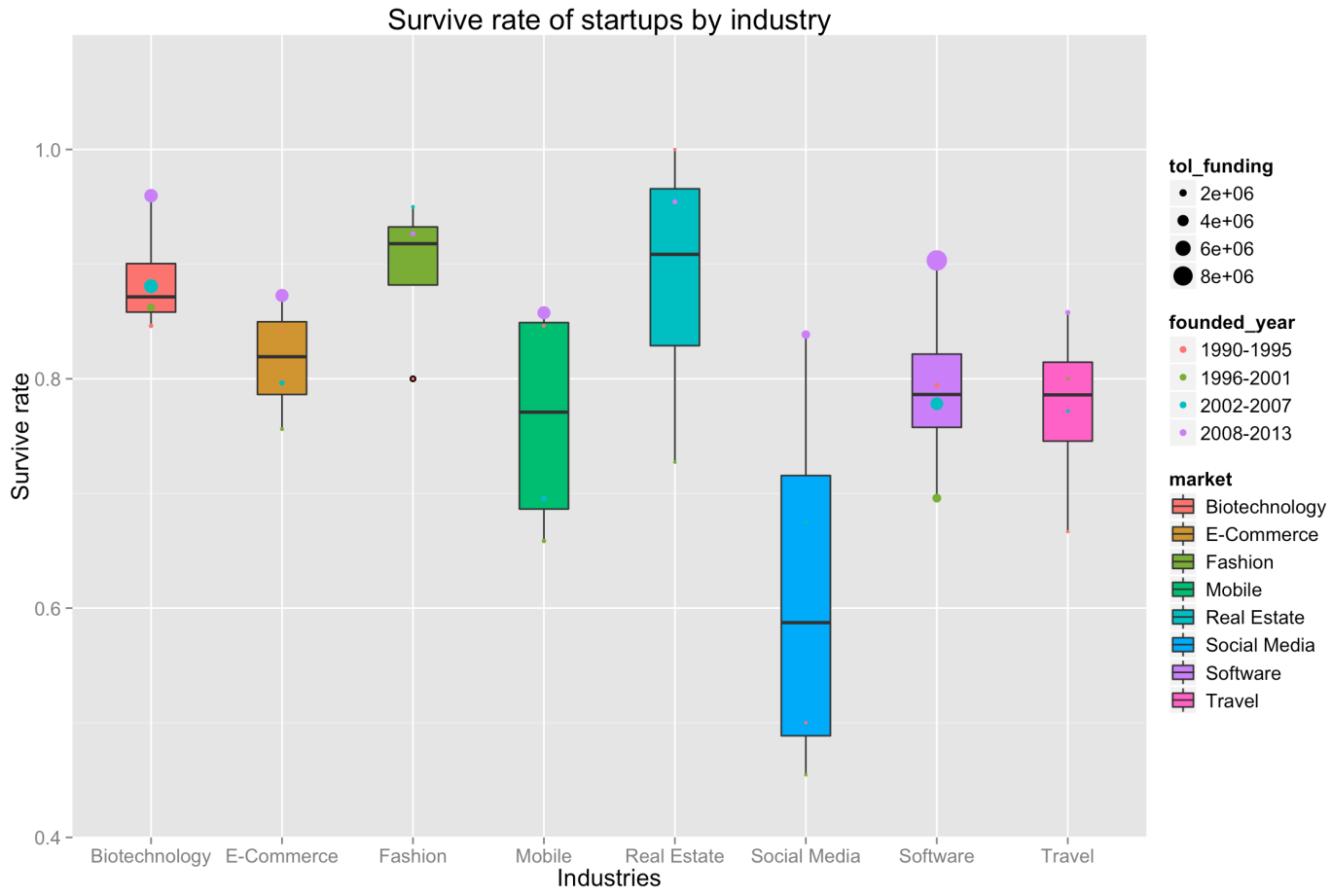
2.13 Sixth Exploration - how does funding change in different founded time periods by industry?

```
p4 + geom_point(aes(size = funding, color = market))
```



2.14 Seventh Exploration - survive rate of startups by industry?

```
# subset dataset with status as operating
df.operating <- subset(df, status == "operating")
# count the total startups in each time period by industry
df.by.market <- df %>%
  group_by(founded_year, market) %>%
  summarise(
    tol.fre = sum(fre)
  )
# merge above two dataset
df.operating.by.market <- merge(df.operating, df.by.market, by = c("founded_year",
"market"))
# calculate survive rate (number of operating / total startups) in each time perio
d by industry
df.operating.by.market <- df.operating.by.market %>%
  group_by(founded_year, market) %>%
  summarise(
    survive.rate = fre/tol.fre,
    tol_funding = funding
  )
# plot graphs
p5 <- ggplot(aes(x = market, y = survive.rate), data = df.operating.by.market) +
      geom_boxplot(aes(fill = market), width = 0.5) +
      geom_point(aes(color = founded_year, size = tol_funding)) +
      coord_cartesian(ylim = c(0.4, 1.1)) +
      theme(text = element_text(size = 15)) +
      ggtitle("Survive rate of startups by industry") +
      labs(x = "Industries", y = "Survive rate")
p5
```




```
# statistics
fit <- aov(survive.rate ~ market, data = df.operating.by.market)
summary(fit)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## market        7 0.2397 0.03425   3.469 0.0104 *
## Residuals    24 0.2369 0.00987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = survive.rate ~ market, data = df.operating.by.market)
##
## $market
##                                       diff         lwr         upr
##  E-Commerce - Biotechnology    -0.070377936 -0.30306398  0.16230811
##  Fashion - Biotechnology        0.009236569 -0.22344948  0.24192262
##  Mobile - Biotechnology        -0.122707944 -0.35539399  0.10997810
##  Real Estate - Biotechnology   -0.001104398 -0.23379045  0.23158165
##  Social Media - Biotechnology  -0.270254708 -0.50294076 -0.03756866
##  Software - Biotechnology      -0.094257262 -0.32694331  0.13842879
##  Travel - Biotechnology        -0.113071407 -0.34575745  0.11961464
##  Fashion - E-Commerce           0.079614505 -0.15307154  0.31230055
##  Mobile - E-Commerce           -0.052330008 -0.28501605  0.18035604
##  Real Estate - E-Commerce       0.069273538 -0.16341251  0.30195959
##  Social Media - E-Commerce     -0.199876772 -0.43256282  0.03280928
##  Software - E-Commerce         -0.023879325 -0.25656537  0.20880672
##  Travel - E-Commerce           -0.042693470 -0.27537952  0.18999258
##  Mobile - Fashion              -0.131944513 -0.36463056  0.10074153
##  Real Estate - Fashion         -0.010340967 -0.24302701  0.22234508
##  Social Media - Fashion        -0.279491277 -0.51217732 -0.04680523
##  Software - Fashion            -0.103493830 -0.33617988  0.12919222
##  Travel - Fashion              -0.122307976 -0.35499402  0.11037807
##  Real Estate - Mobile           0.121603546 -0.11108250  0.35428959
##  Social Media - Mobile         -0.147546764 -0.38023281  0.08513928
##  Software - Mobile              0.028450683 -0.20423536  0.26113673
##  Travel - Mobile                0.009636537 -0.22304951  0.24232258
##  Social Media - Real Estate    -0.269150310 -0.50183636 -0.03646426
##  Software - Real Estate        -0.093152863 -0.32583891  0.13953318
##  Travel - Real Estate          -0.111967009 -0.34465306  0.12071904
##  Software - Social Media        0.175997447 -0.05668860  0.40868349
##  Travel - Social Media          0.157183301 -0.07550275  0.38986935
##  Travel - Software             -0.018814145 -0.25150019  0.21387190
##                                     p adj
##  E-Commerce - Biotechnology    0.9697040
##  Fashion - Biotechnology       1.0000000
##  Mobile - Biotechnology        0.6589544
##  Real Estate - Biotechnology   1.0000000
##  Social Media - Biotechnology  0.0150137
##  Software - Biotechnology      0.8736626
##  Travel - Biotechnology        0.7404658
##  Fashion - E-Commerce          0.9426369
##  Mobile - E-Commerce           0.9944458
##  Real Estate - E-Commerce      0.9721772
##  Social Media - E-Commerce     0.1300784
##  Software - E-Commerce         0.9999661
##  Travel - E-Commerce           0.9984268
##  Mobile - Fashion              0.5774338
```

```
##  Real Estate - Fashion         0.9999999
##  Social Media - Fashion        0.0110367
##  Software - Fashion            0.8137172
##  Travel - Fashion              0.6624385
##  Real Estate - Mobile          0.6685590
##  Social Media - Mobile         0.4424544
##  Software - Mobile             0.9998894
##  Travel - Mobile               0.9999999
##  Social Media - Real Estate    0.0155724
##  Software - Real Estate        0.8800126
##  Travel - Real Estate          0.7493839
##  Software - Social Media       0.2403871
##  Travel - Social Media         0.3659719
##  Travel - Software             0.9999933
```

2.15 Export Data

```
df.final <- merge (df, df.operating.by.market, by = c("founded_year", "market"))
colnames(df.final) <- c("Founded year of startups", "Startup industry", "Status",
"Funding of startups in this industry by status","Number of startups in this indus
try by status", "Total number of founded startups", "Percentage of startups by ind
ustry", "Survive rate of startups in this industry by status", "Total funding by i
ndustry")
write.csv(df.final, file="data.csv", row.names=FALSE)
```

Final question: can we predict the status of a startup in future?

Setup a machine learning model (use python) to answer this question