Financial Data 전처리 및 시각화

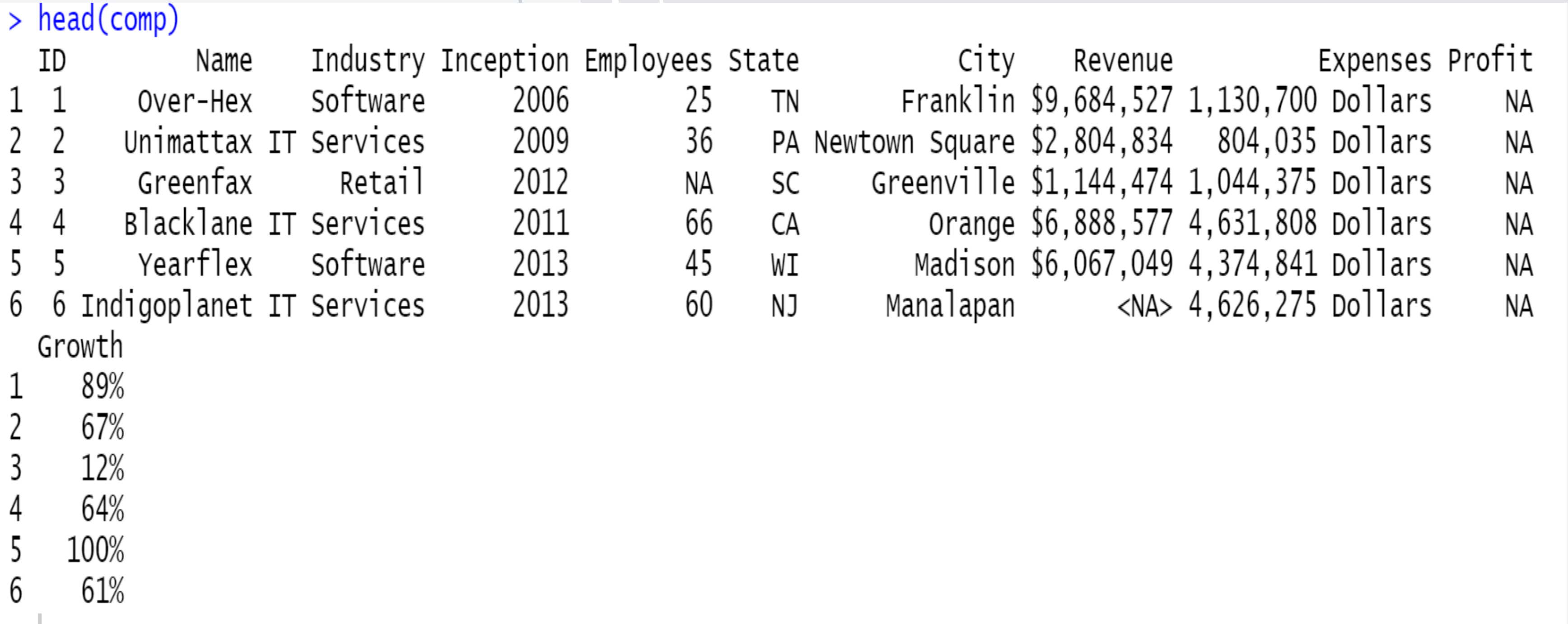
목차

1. 데이터 설명
2. 데이터 속성
3. 속성의 변화
4. 각 변수별 결측치 확인
5. 종업원 결측치 대체
6. State 결측치 대체
7. City 결측치 대체
8. Revenue 결측치 대체
9. Expenses 결측치 대체
10.Growth 결측치 대체
11.시각화

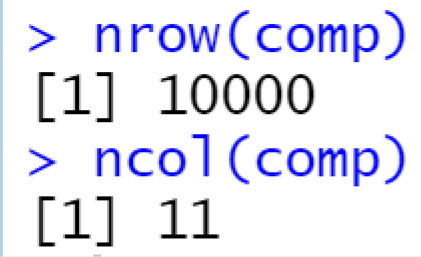
데이터 불러오기

빈 값을 NA값으로 대체
Na를 찾는 것에 걸리지
않아서 사용

```
comp<-read.csv("C:₩₩Users₩₩82109₩₩Desktop/만들어진 Financial Dataset.csv",na.strings=c("","NA"))
```

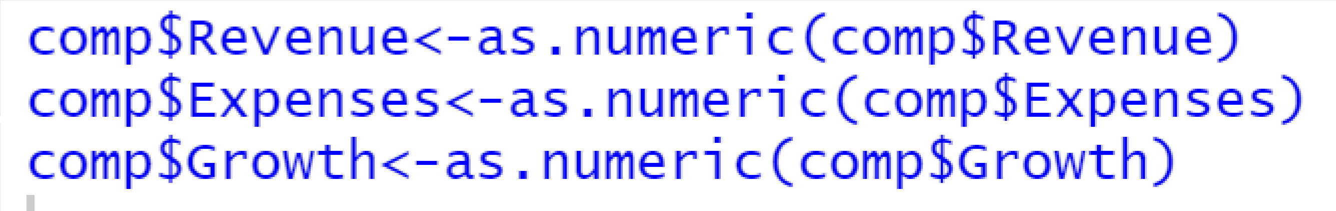


```
> head(comp)
  ID        Name    Industry Inception Employees State           City    Revenue           Expenses Profit
1  1    Over-Hex    Software      2006        25    TN       Franklin $9,684,527 1,130,700 Dollars     NA
2  2   Unimattax IT Services      2009        36    PA Newtown Square $2,804,834   804,035 Dollars     NA
3  3    Greenfax      Retail      2012        NA    SC     Greenville $1,144,474 1,044,375 Dollars     NA
4  4   Blacklane IT Services      2011        66    CA         Orange $6,888,577 4,631,808 Dollars     NA
5  5    Yearflex    Software      2013        45    WI        Madison $6,067,049 4,374,841 Dollars     NA
6  6 Indigoplanet IT Services     2013        60    NJ      Manalapan       <NA> 4,626,275 Dollars     NA
  Growth
1    89%
2    67%
3    12%
4    64%
5   100%
6    61%
```

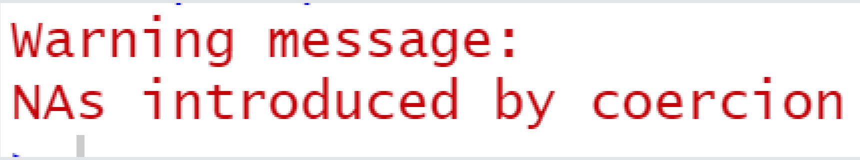


```
> nrow(comp)
[1] 10000
> ncol(comp)
[1] 11
```

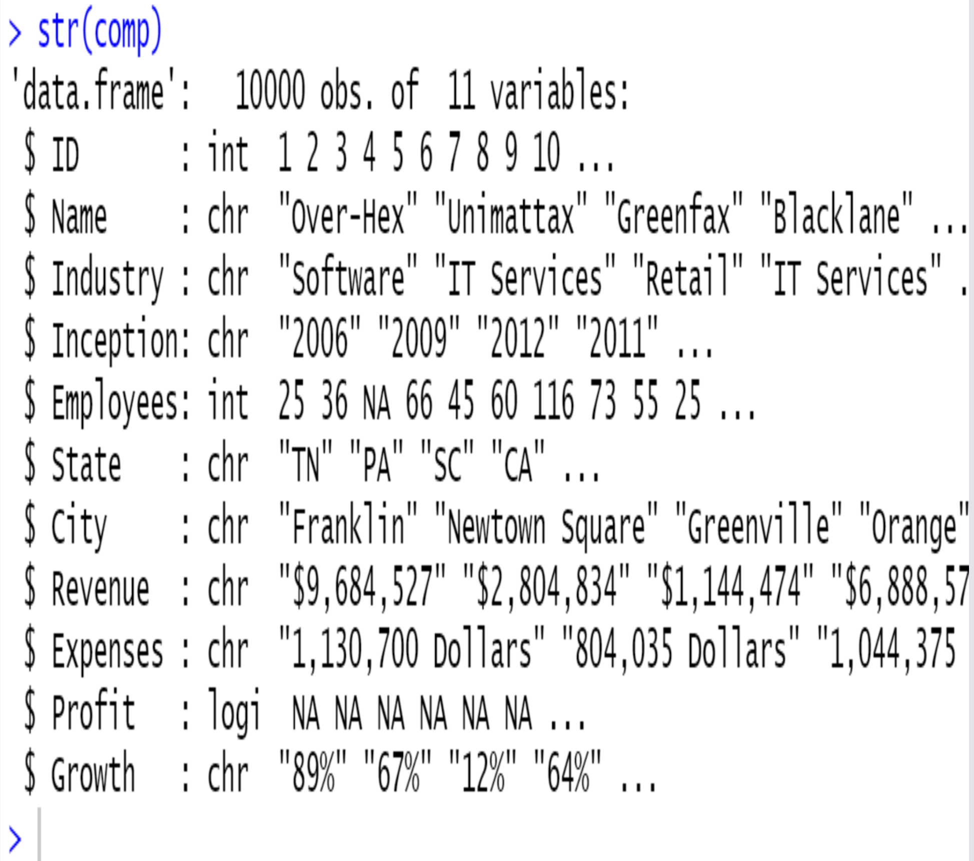
Profit 계산을 위한데이터 속성의 변화



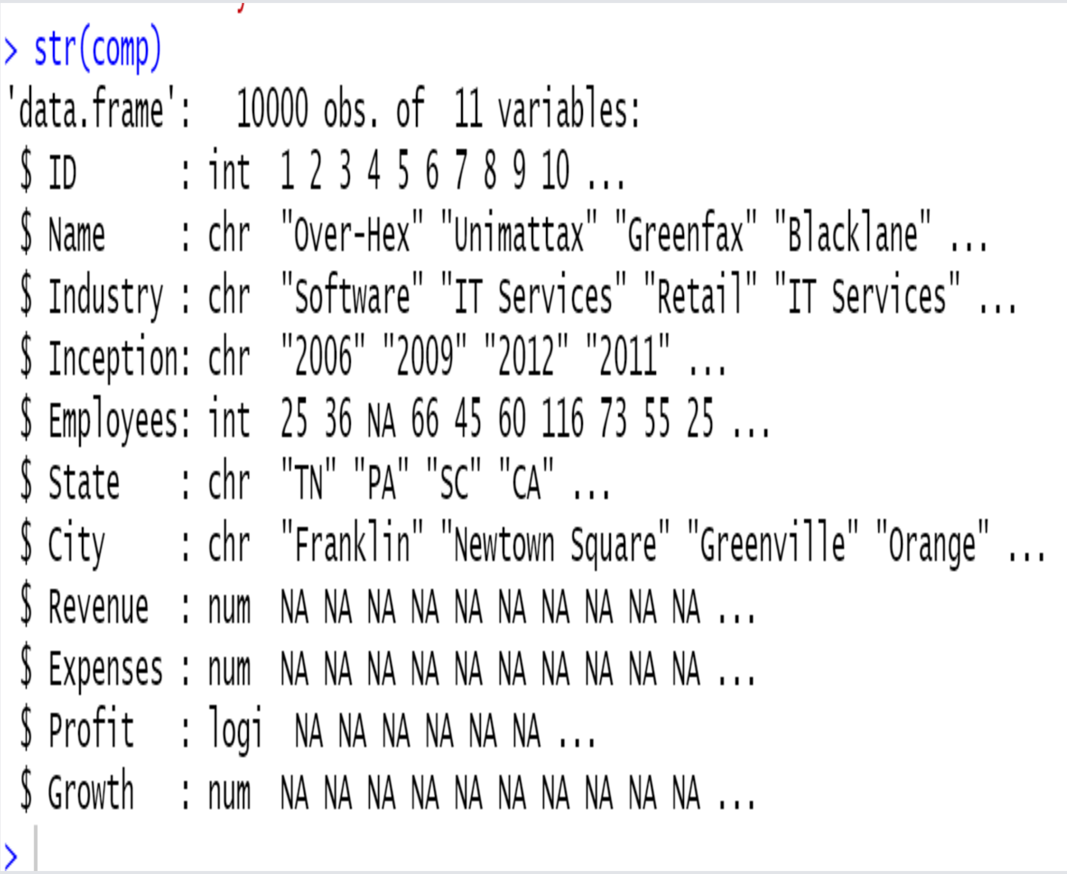
```
comp$Revenue<-as.numeric(comp$Revenue)
comp$Expenses<-as.numeric(comp$Expenses)
comp$Growth<-as.numeric(comp$Growth)
```



```
Warning message:
NAs introduced by coercion
```

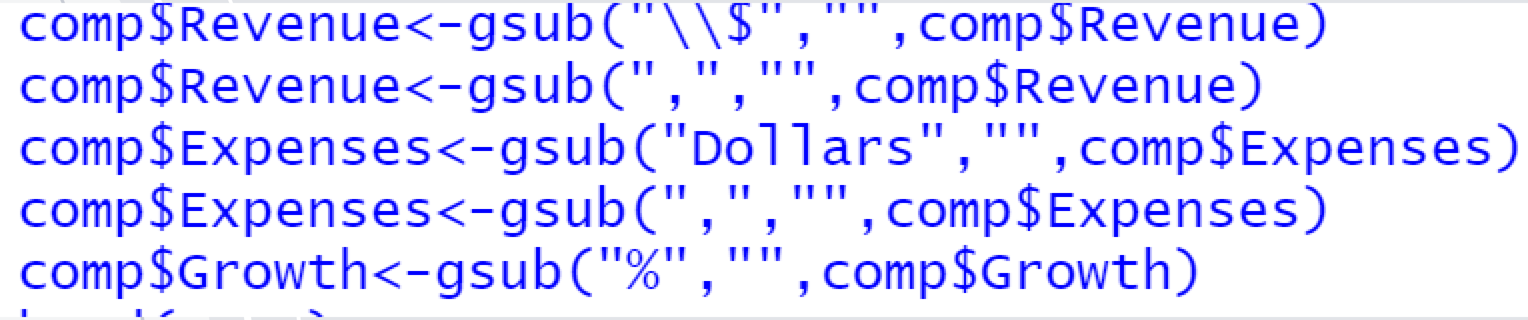


```
> str(comp)
'data.frame':   10000 obs. of  11 variables:
 $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Name     : chr  "Over-Hex" "Unimattax" "Greenfax" "Blacklane" ...
 $ Industry : chr  "Software" "IT Services" "Retail" "IT Services" .
 $ Inception: chr  "2006" "2009" "2012" "2011" ...
 $ Employees: int  25 36 NA 66 45 60 116 73 55 25 ...
 $ State    : chr  "TN" "PA" "SC" "CA" ...
 $ City     : chr  "Franklin" "Newtown Square" "Greenville" "Orange"
 $ Revenue  : chr  "$9,684,527" "$2,804,834" "$1,144,474" "$6,888,57
 $ Expenses : chr  "1,130,700 Dollars" "804,035 Dollars" "1,044,375
 $ Profit   : logi  NA NA NA NA NA NA ...
 $ Growth   : chr  "89%" "67%" "12%" "64%" ...
>
```

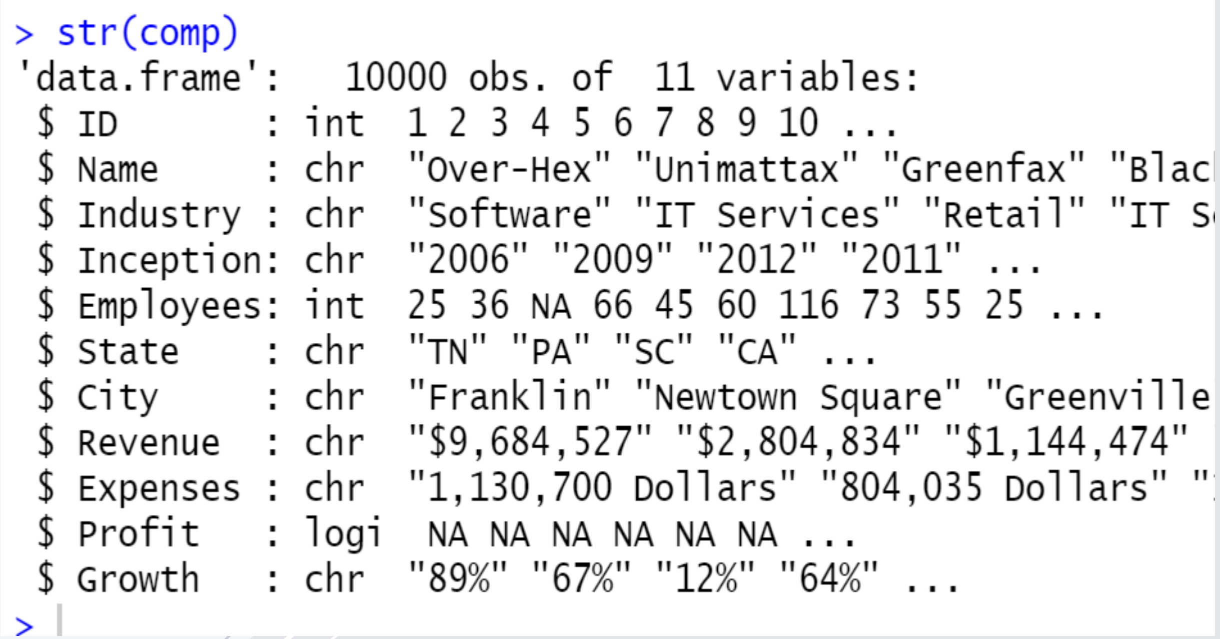


```
> str(comp)
'data.frame':   10000 obs. of  11 variables:
 $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Name     : chr  "Over-Hex" "Unimattax" "Greenfax" "Blacklane" ...
 $ Industry : chr  "Software" "IT Services" "Retail" "IT Services" ...
 $ Inception: chr  "2006" "2009" "2012" "2011" ...
 $ Employees: int  25 36 NA 66 45 60 116 73 55 25 ...
 $ State    : chr  "TN" "PA" "SC" "CA" ...
 $ City     : chr  "Franklin" "Newtown Square" "Greenville" "Orange" ...
 $ Revenue  : num  NA NA NA NA NA NA NA NA NA NA ...
 $ Expenses : num  NA NA NA NA NA NA NA NA NA NA ...
 $ Profit   : logi  NA NA NA NA NA NA ...
 $ Growth   : num  NA NA NA NA NA NA NA NA NA NA ...
>
```

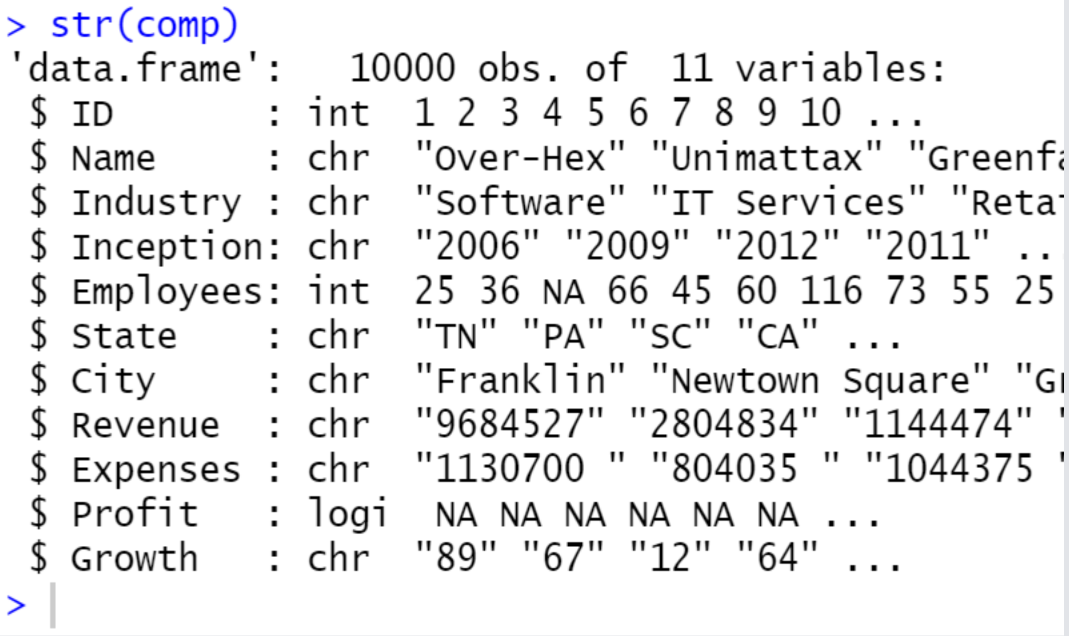
gsub을 통한 Expenses, Revenu, Growth의 부호빼기



```
comp$Revenue<-gsub("\\$","",comp$Revenue)
comp$Revenue<-gsub(",","",comp$Revenue)
comp$Expenses<-gsub("Dollars","",comp$Expenses)
comp$Expenses<-gsub(",","",comp$Expenses)
comp$Growth<-gsub("%","",comp$Growth)
```

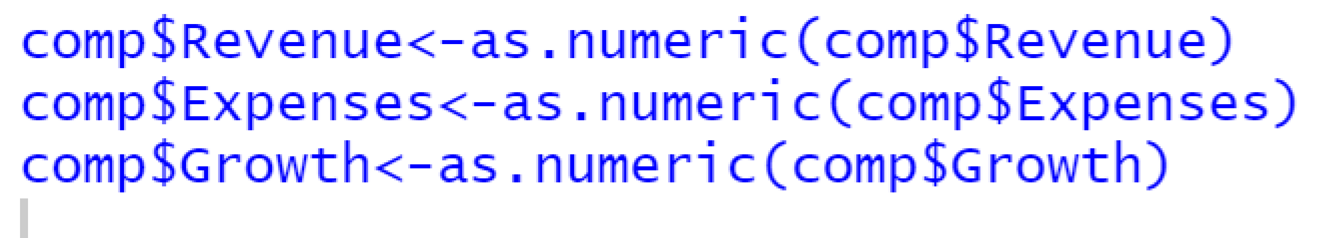


```
> str(comp)
'data.frame':	10000 obs. of  11 variables:
 $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Name     : chr  "Over-Hex" "Unimattax" "Greenfax" "Blac
 $ Industry : chr  "Software" "IT Services" "Retail" "IT S
 $ Inception: chr  "2006" "2009" "2012" "2011" ...
 $ Employees: int  25 36 NA 66 45 60 116 73 55 25 ...
 $ State    : chr  "TN" "PA" "SC" "CA" ...
 $ City     : chr  "Franklin" "Newtown Square" "Greenville
 $ Revenue  : chr  "$9,684,527" "$2,804,834" "$1,144,474"
 $ Expenses : chr  "1,130,700 Dollars" "804,035 Dollars" "
 $ Profit   : logi  NA NA NA NA NA NA ...
 $ Growth   : chr  "89%" "67%" "12%" "64%" ...
> 
```

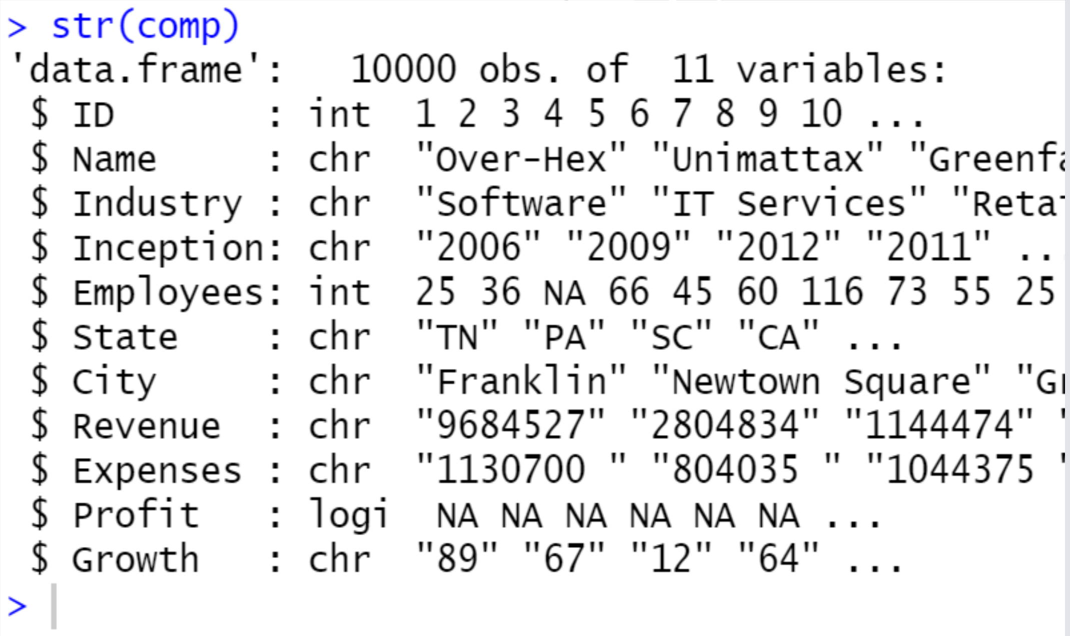


```
> str(comp)
'data.frame':	10000 obs. of  11 variables:
 $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Name     : chr  "Over-Hex" "Unimattax" "Greenfa
 $ Industry : chr  "Software" "IT Services" "Reta
 $ Inception: chr  "2006" "2009" "2012" "2011" ..
 $ Employees: int  25 36 NA 66 45 60 116 73 55 25
 $ State    : chr  "TN" "PA" "SC" "CA" ...
 $ City     : chr  "Franklin" "Newtown Square" "Gr
 $ Revenue  : chr  "9684527" "2804834" "1144474" "
 $ Expenses : chr  "1130700 " "804035 " "1044375 "
 $ Profit   : logi  NA NA NA NA NA NA ...
 $ Growth   : chr  "89" "67" "12" "64" ...
> 
```

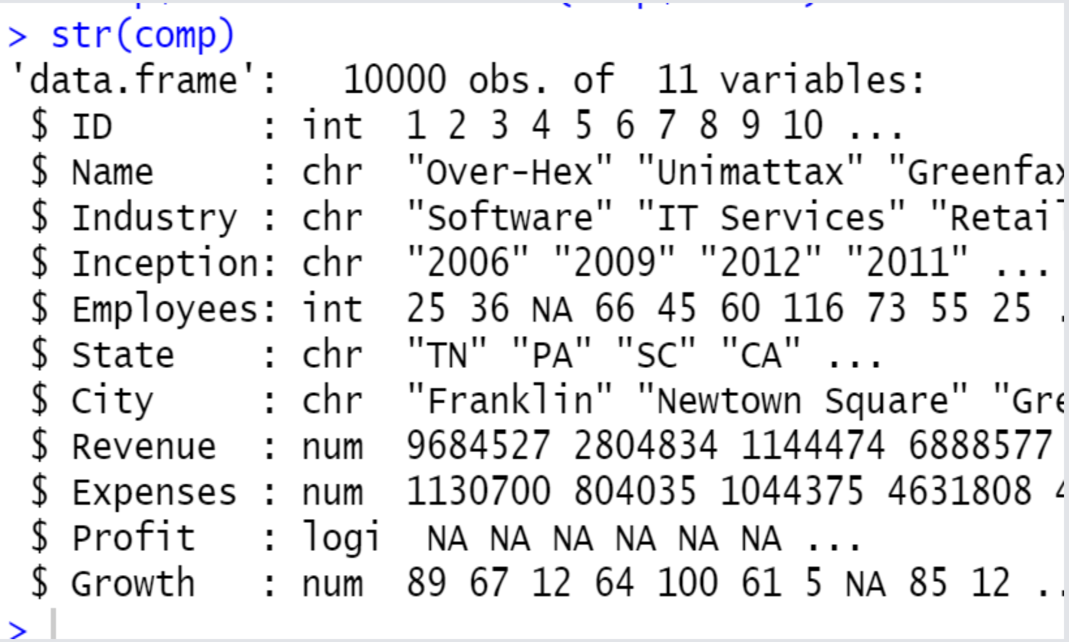
Expenses, Revenue,Growth의 수치화



```
comp$Revenue<-as.numeric(comp$Revenue)
comp$Expenses<-as.numeric(comp$Expenses)
comp$Growth<-as.numeric(comp$Growth)
```

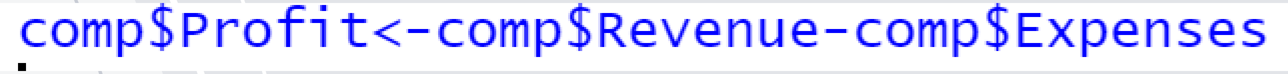


```
> str(comp)
'data.frame':	10000 obs. of  11 variables:
 $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Name     : chr  "Over-Hex" "Unimattax" "Greenfa
 $ Industry : chr  "Software" "IT Services" "Retai
 $ Inception: chr  "2006" "2009" "2012" "2011" ..
 $ Employees: int  25 36 NA 66 45 60 116 73 55 25
 $ State    : chr  "TN" "PA" "SC" "CA" ...
 $ City     : chr  "Franklin" "Newtown Square" "Gr
 $ Revenue  : chr  "9684527" "2804834" "1144474" '
 $ Expenses : chr  "1130700 " "804035 " "1044375 '
 $ Profit   : logi  NA NA NA NA NA NA ...
 $ Growth   : chr  "89" "67" "12" "64" ...
>
```

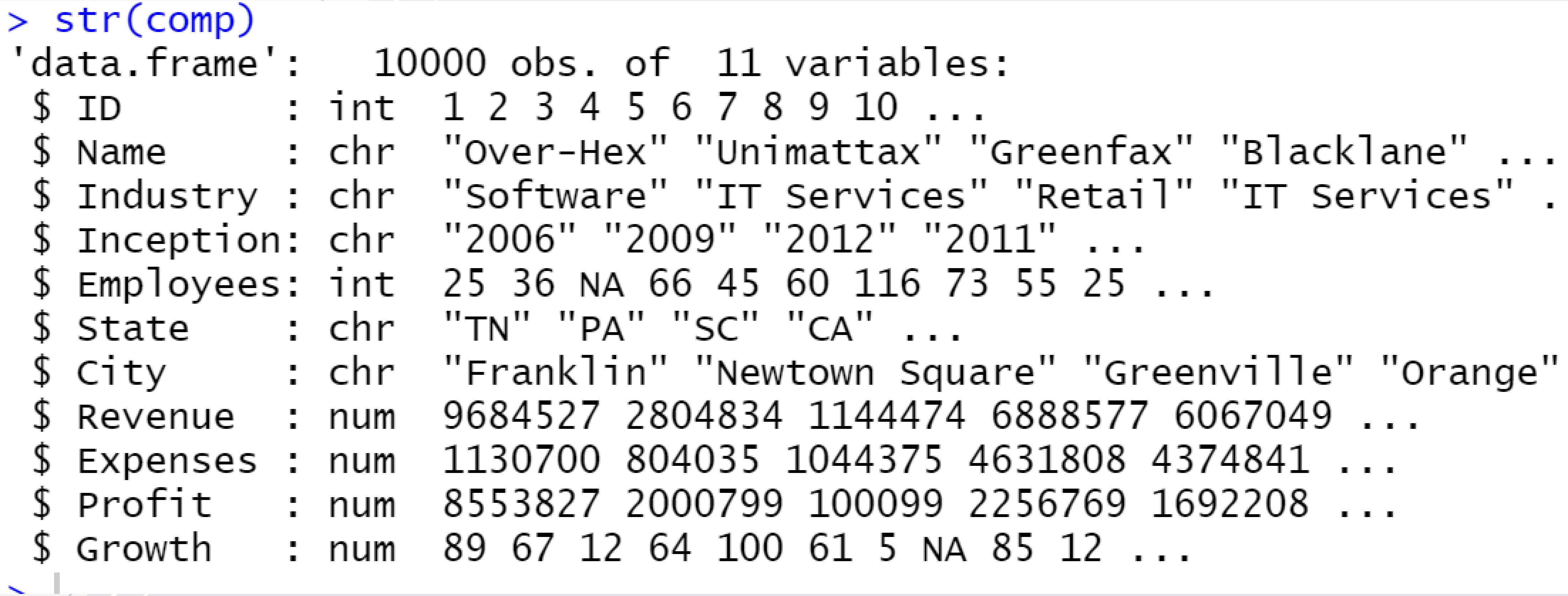


```
> str(comp)
'data.frame':	10000 obs. of  11 variables:
 $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Name     : chr  "Over-Hex" "Unimattax" "Greenfa
 $ Industry : chr  "Software" "IT Services" "Retai
 $ Inception: chr  "2006" "2009" "2012" "2011" ...
 $ Employees: int  25 36 NA 66 45 60 116 73 55 25 .
 $ State    : chr  "TN" "PA" "SC" "CA" ...
 $ City     : chr  "Franklin" "Newtown Square" "Gre
 $ Revenue  : num  9684527 2804834 1144474 6888577
 $ Expenses : num  1130700 804035 1044375 4631808 4
 $ Profit   : logi  NA NA NA NA NA NA ...
 $ Growth   : num  89 67 12 64 100 61 5 NA 85 12 ..
>
```

Profit 계산하기



```
comp$Profit<-comp$Revenue-comp$Expenses
```

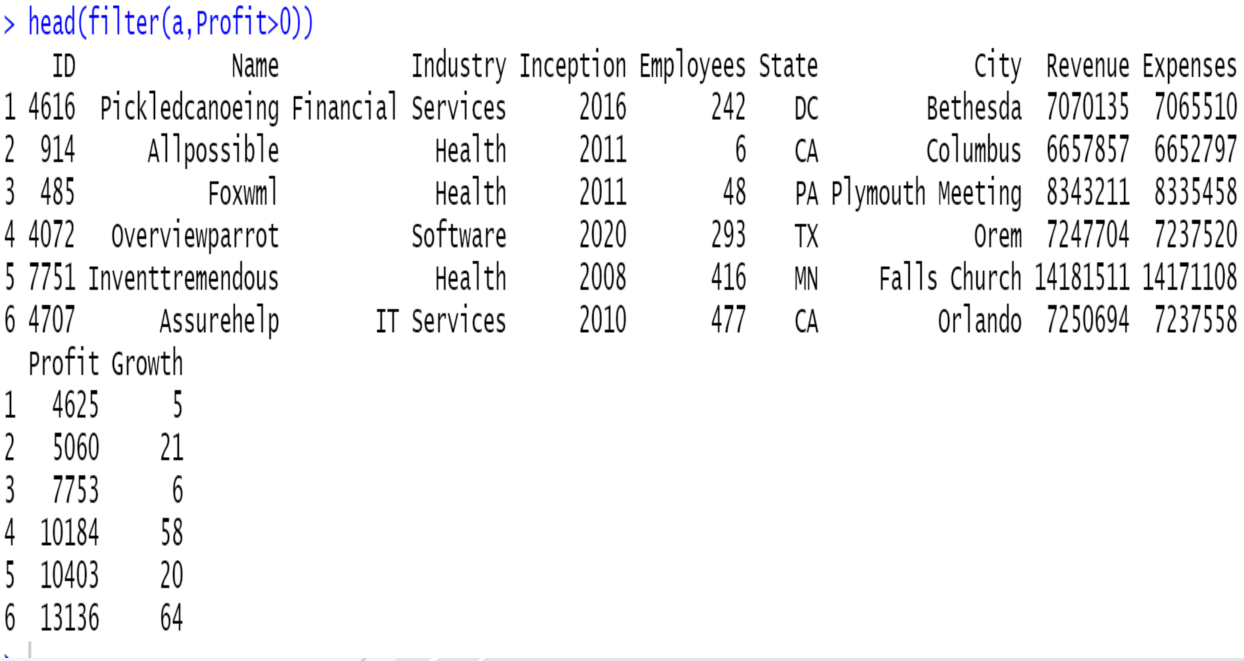


```
> str(comp)
'data.frame':   10000 obs. of  11 variables:
 $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Name     : chr  "Over-Hex" "Unimattax" "Greenfax" "Blacklane" ...
 $ Industry : chr  "Software" "IT Services" "Retail" "IT Services" .
 $ Inception: chr  "2006" "2009" "2012" "2011" ...
 $ Employees: int  25 36 NA 66 45 60 116 73 55 25 ...
 $ State    : chr  "TN" "PA" "SC" "CA" ...
 $ City     : chr  "Franklin" "Newtown Square" "Greenville" "Orange"
 $ Revenue  : num  9684527 2804834 1144474 6888577 6067049 ...
 $ Expenses : num  1130700 804035 1044375 4631808 4374841 ...
 $ Profit   : num  8553827 2000799 100099 2256769 1692208 ...
 $ Growth   : num  89 67 12 64 100 61 5 NA 85 12 ...
>
```

Profit으로 얻을 수 있는 정보들

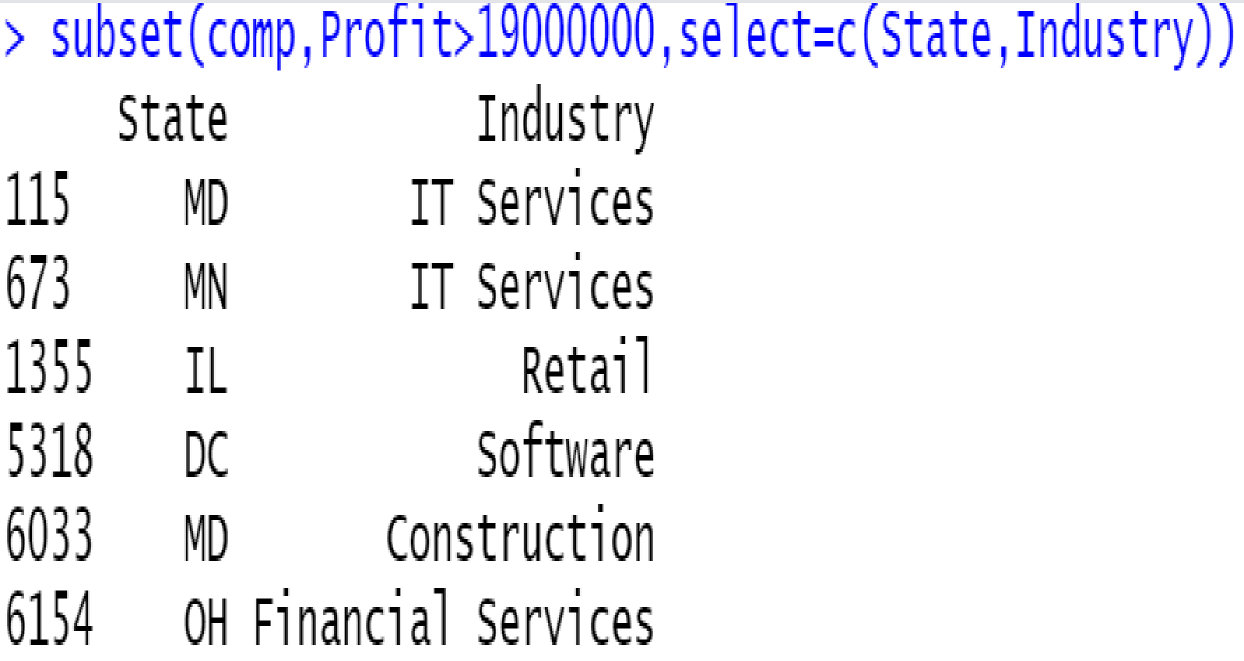
0초과인 Profit 순서대로

```
a<-arrange(comp,Profit)
```



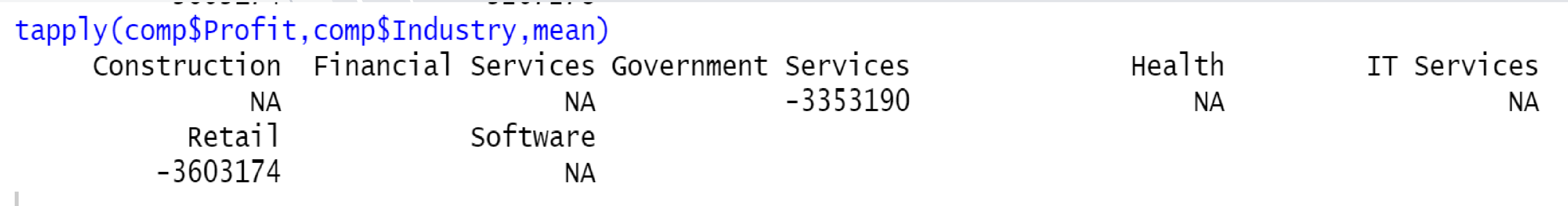
```
> head(filter(a,Profit>0))
    ID             Name           Industry Inception Employees State           City  Revenue Expenses
1 4616  Pickledcanoeing Financial Services      2016       242    DC       Bethesda  7070135  7065510
2  914      Allpossible             Health      2011         6    CA       Columbus  6657857  6652797
3  485           Foxwml             Health      2011        48    PA Plymouth Meeting  8343211  8335458
4 4072   Overviewparrot           Software      2020       293    TX           Orem  7247704  7237520
5 7751 Inventtremendous             Health      2008       416    MN   Falls Church 14181511 14171108
6 4707       Assurehelp        IT Services      2010       477    CA        Orlando  7250694  7237558
  Profit Growth
1   4625      5
2   5060     21
3   7753      6
4  10184     58
5  10403     20
6  13136     64
```

Profit 19000,000 이상인 State와 Industry 추출



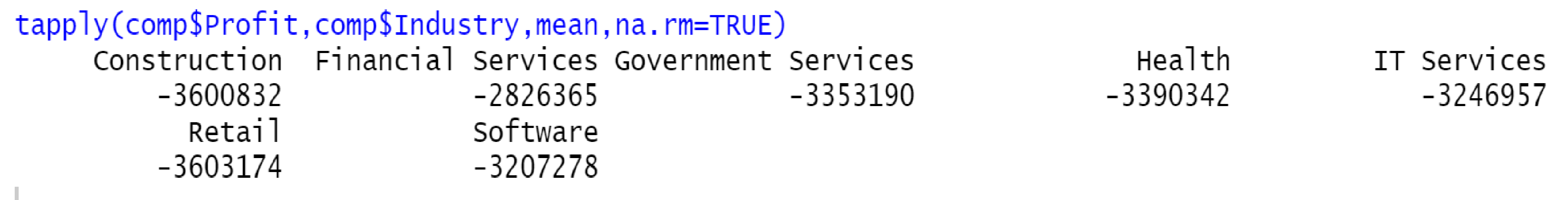
```
> subset(comp,Profit>19000000,select=c(State,Industry))
     State           Industry
115     MD        IT Services
673     MN        IT Services
1355    IL             Retail
5318    DC           Software
6033    MD       Construction
6154    OH Financial Services
```

산업별로 Profit의 평균값



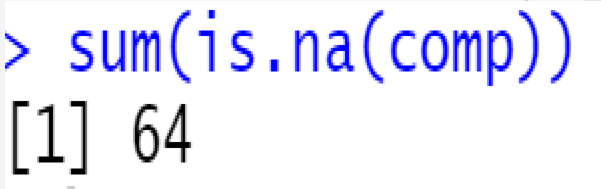
```
tapply(comp$Profit,comp$Industry,mean)
     Construction  Financial Services Government Services              Health         IT Services 
               NA                  NA            -3353190                  NA                  NA 
           Retail            Software 
         -3603174                  NA 
```

Mean값이 NA 값이 있다는 건 Government Services랑
Reatil을 제외하고는 결측값이 있다는 것을 의미



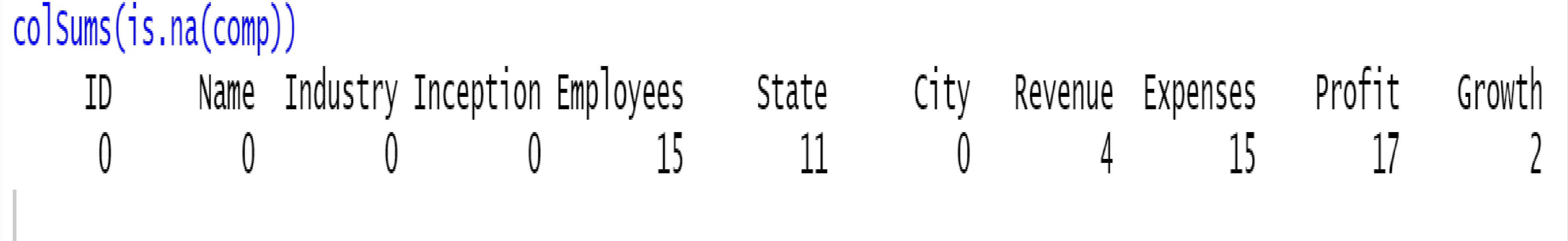
```
tapply(comp$Profit,comp$Industry,mean,na.rm=TRUE)
     Construction  Financial Services Government Services              Health         IT Services 
         -3600832            -2826365            -3353190            -3390342            -3246957 
           Retail            Software 
         -3603174            -3207278 
```

NA값



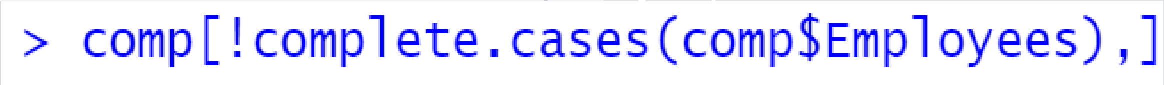
```
> sum(is.na(comp))
[1] 64
```

변수별 NA값

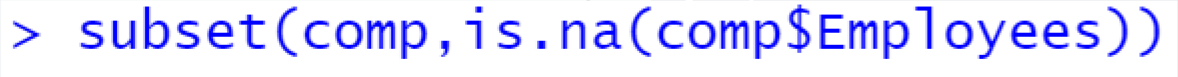


```
colSums(is.na(comp))
       ID      Name  Industry Inception Employees     State      City   Revenue  Expenses    Profit    Growth
        0         0         0         0        15        11         0         4        15        17         2
```

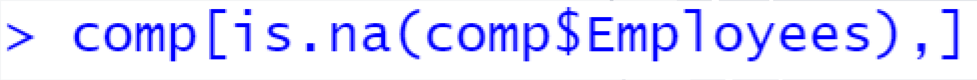
Employees NA값



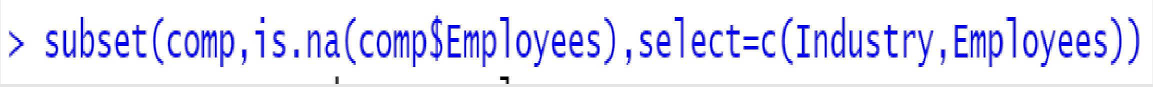
```
> comp[!complete.cases(comp$Employees),]
```



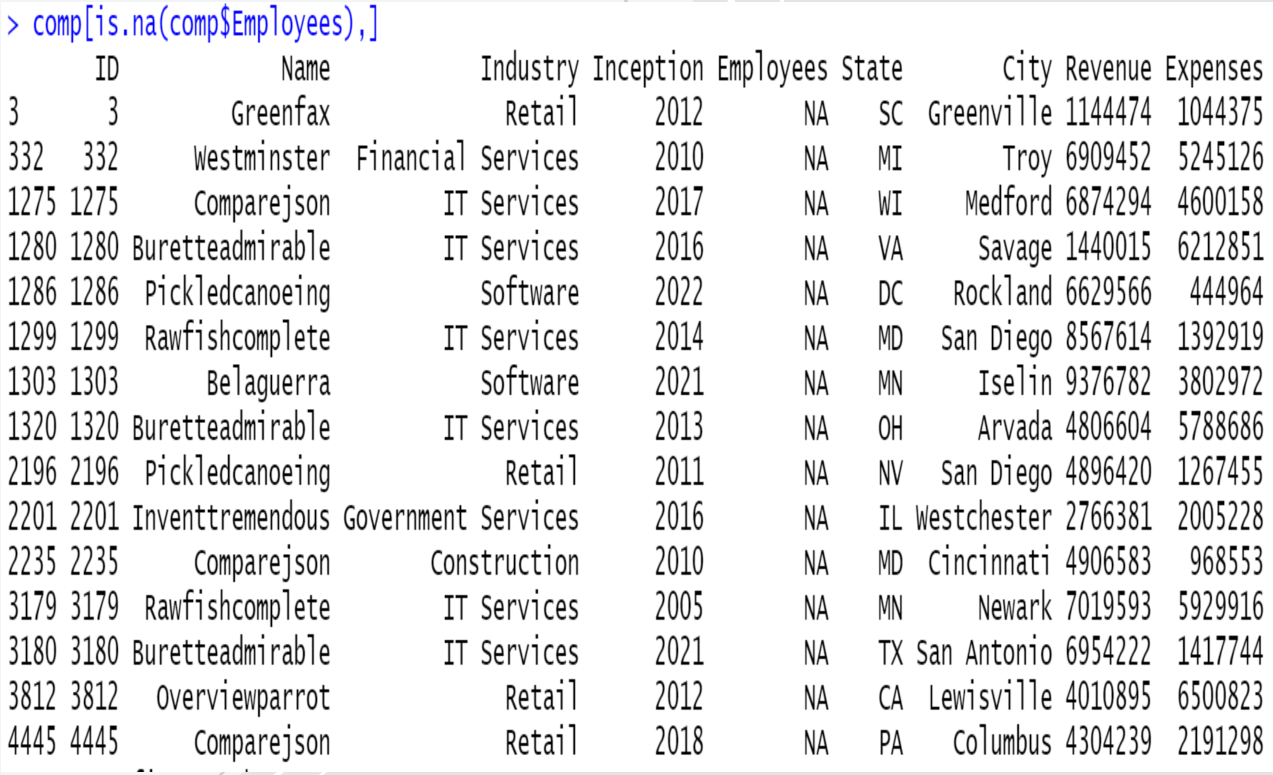
```
> subset(comp,is.na(comp$Employees))
```



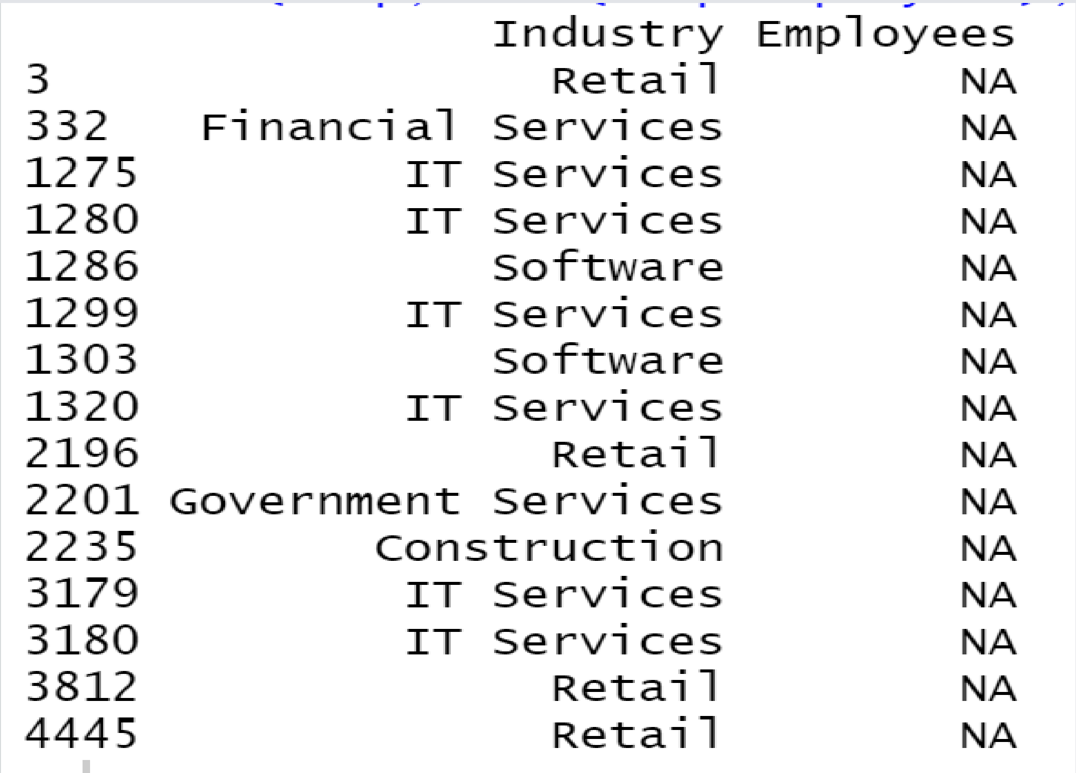
```
> comp[is.na(comp$Employees),]
```



```
> subset(comp,is.na(comp$Employees),select=c(Industry,Employees))
```



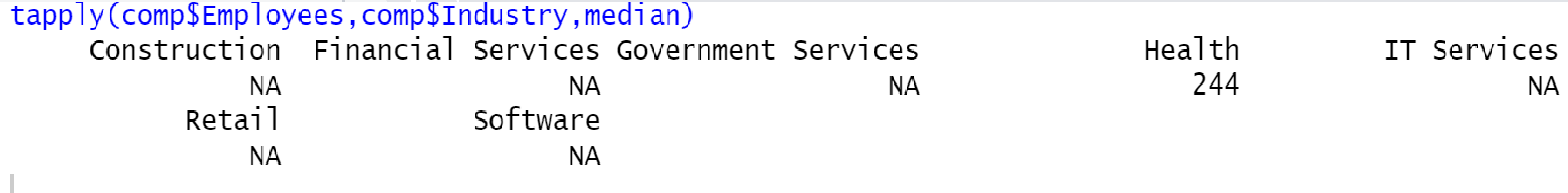
```
> comp[is.na(comp$Employees),]
       ID             Name           Industry Inception Employees State        City Revenue Expenses
3       3         Greenfax             Retail      2012        NA    SC  Greenville 1144474  1044375
332   332      Westminster Financial Services      2010        NA    MI        Troy 6909452  5245126
1275 1275      Comparejson        IT Services      2017        NA    WI     Medford 6874294  4600158
1280 1280 Buretteadmirable        IT Services      2016        NA    VA      Savage 1440015  6212851
1286 1286  Pickledcanoeing           Software      2022        NA    DC    Rockland 6629566   444964
1299 1299  Rawfishcomplete        IT Services      2014        NA    MD   San Diego 8567614  1392919
1303 1303       Belaguerra           Software      2021        NA    MN      Iselin 9376782  3802972
1320 1320 Buretteadmirable        IT Services      2013        NA    OH      Arvada 4806604  5788686
2196 2196  Pickledcanoeing             Retail      2011        NA    NV   San Diego 4896420  1267455
2201 2201 Inventtremendous Government Services     2016        NA    IL Westchester 2766381  2005228
2235 2235      Comparejson       Construction      2010        NA    MD  Cincinnati 4906583   968553
3179 3179  Rawfishcomplete        IT Services      2005        NA    MN      Newark 7019593  5929916
3180 3180 Buretteadmirable        IT Services      2021        NA    TX San Antonio 6954222  1417744
3812 3812   Overviewparrot             Retail      2012        NA    CA  Lewisville 4010895  6500823
4445 4445      Comparejson             Retail      2018        NA    PA    Columbus 4304239  2191298
```



```
                Industry Employees
3                 Retail        NA
332   Financial Services        NA
1275         IT Services        NA
1280         IT Services        NA
1286            Software        NA
1299         IT Services        NA
1303            Software        NA
1320         IT Services        NA
2196              Retail        NA
2201 Government Services        NA
2235        Construction        NA
3179         IT Services        NA
3180         IT Services        NA
3812              Retail        NA
4445              Retail        NA
```

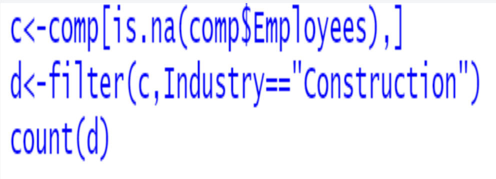
산업별Employees NA값이 많을때

Employees 에 NA값이 하나라도 있으면 NA값으로 반환한다는 것에 착안



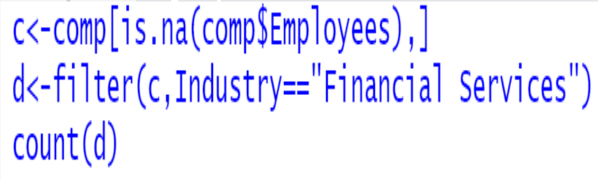
```
tapply(comp$Employees,comp$Industry,median)
      Construction  Financial Services Government Services              Health        IT Services
                NA                  NA                  NA                 244                 NA
            Retail            Software
                NA                  NA
```

Employees의 NA값중에 Industry가 Construction인것 의 수



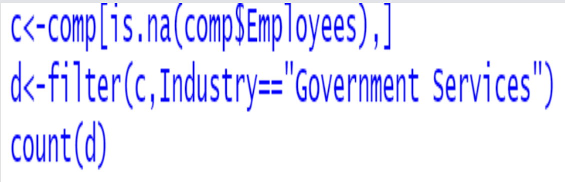
```
c<-comp[is.na(comp$Employees),]
d<-filter(c,Industry=="Construction")
count(d)
```

1



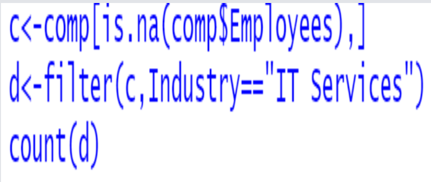
```
c<-comp[is.na(comp$Employees),]
d<-filter(c,Industry=="Financial Services")
count(d)
```

1



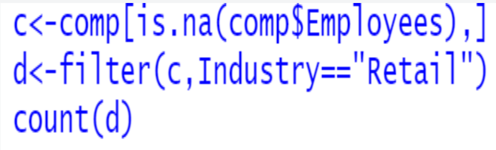
```
c<-comp[is.na(comp$Employees),]
d<-filter(c,Industry=="Government Services")
count(d)
```

1



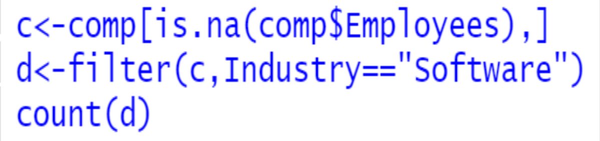
```
c<-comp[is.na(comp$Employees),]
d<-filter(c,Industry=="IT Services")
count(d)
```

6



```
c<-comp[is.na(comp$Employees),]
d<-filter(c,Industry=="Retail")
count(d)
```

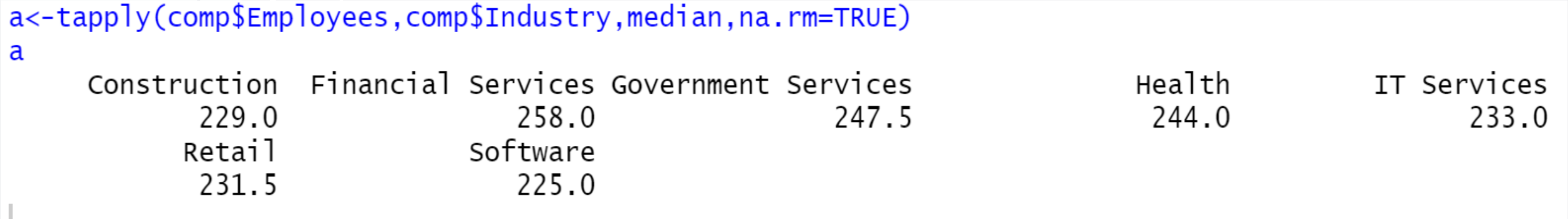
4



```
c<-comp[is.na(comp$Employees),]
d<-filter(c,Industry=="Software")
count(d)
```

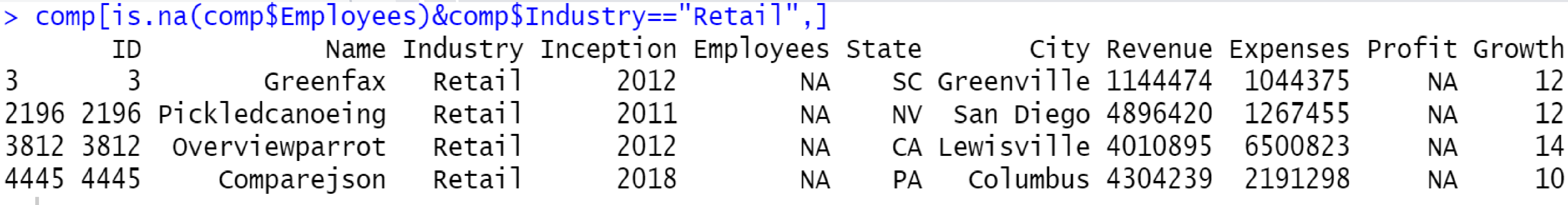
2

Employees 산업별 Median값으로 대체



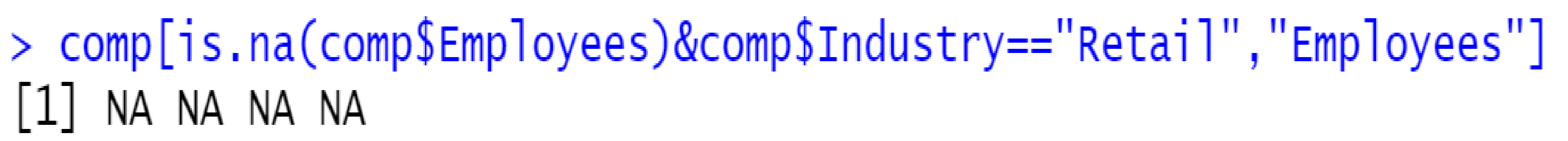
```
a<-tapply(comp$Employees,comp$Industry,median,na.rm=TRUE)
a
     Construction  Financial Services Government Services              Health         IT Services 
            229.0               258.0               247.5               244.0               233.0 
           Retail            Software 
            231.5               225.0 
```

Employees의 결측값에 Industry가 Retail 일때의 전체 값



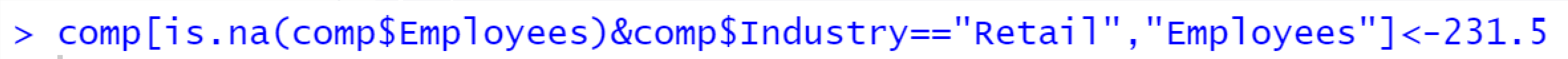
```
> comp[is.na(comp$Employees)&comp$Industry=="Retail",]
       ID             Name Industry Inception Employees State        City Revenue Expenses Profit Growth
3       3         Greenfax   Retail      2012        NA    SC Greenville 1144474  1044375     NA     12
2196 2196 Pickledcanoeing    Retail      2011        NA    NV  San Diego 4896420  1267455     NA     12
3812 3812  Overviewparrot    Retail      2012        NA    CA Lewisville 4010895  6500823     NA     14
4445 4445      Comparejson   Retail      2018        NA    PA   Columbus 4304239  2191298     NA     10
```

Employees의 결측값에 Industry가 Retail 일때의 Employees 값

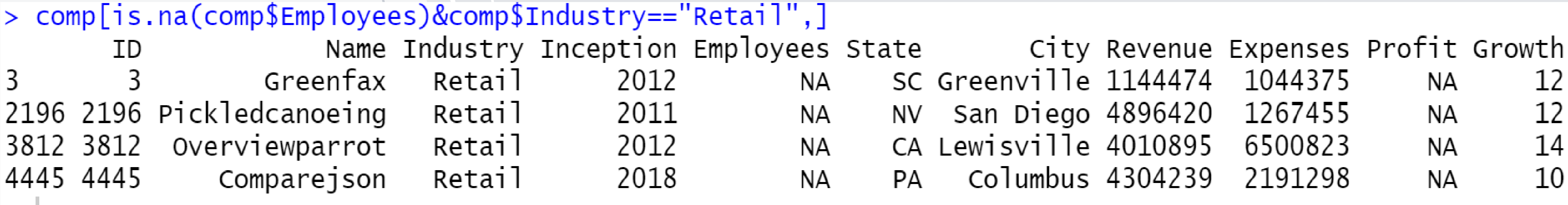


```
> comp[is.na(comp$Employees)&comp$Industry=="Retail","Employees"]
[1] NA NA NA NA
```

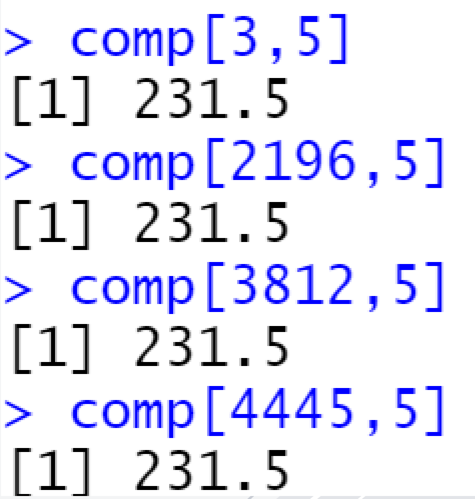
대체 후 확인



```
> comp[is.na(comp$Employees)&comp$Industry=="Retail","Employees"]<-231.5
```

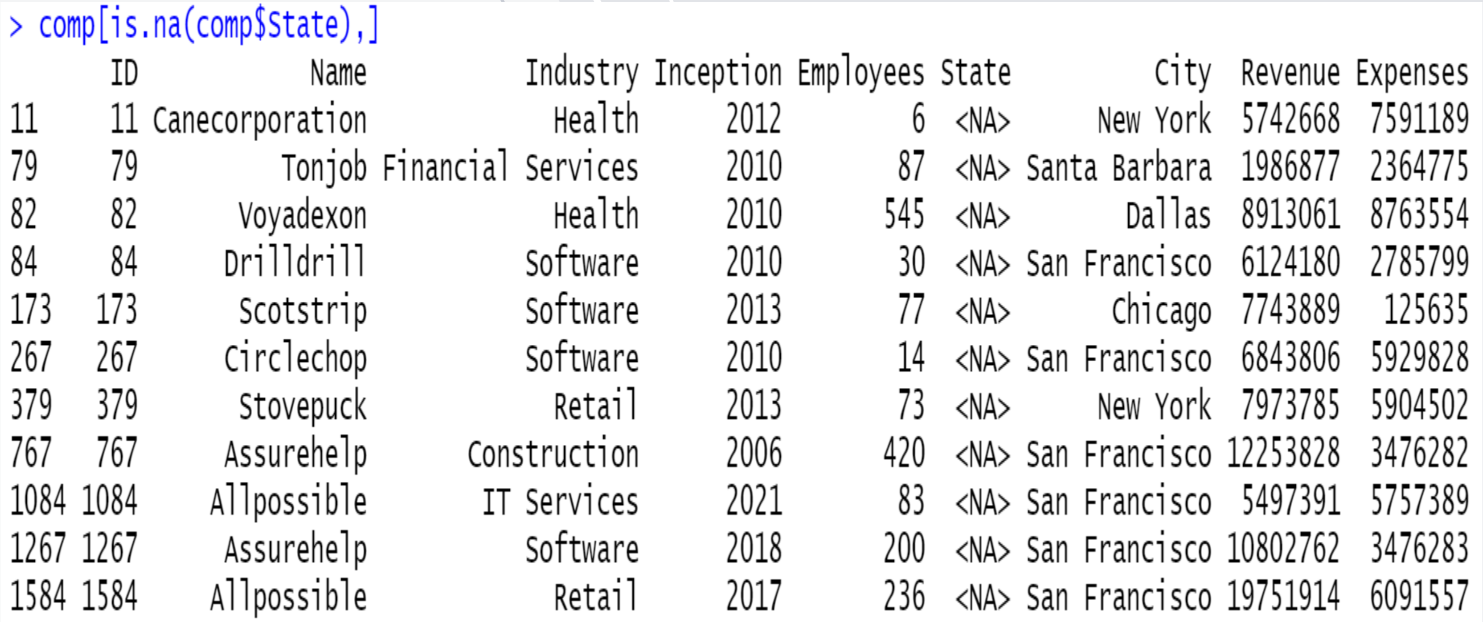


```
> comp[is.na(comp$Employees)&comp$Industry=="Retail",]
       ID             Name Industry Inception Employees State        City Revenue Expenses Profit Growth
3       3         Greenfax   Retail      2012        NA    SC  Greenville 1144474  1044375     NA     12
2196 2196 Pickledcanoeing   Retail      2011        NA    NV   San Diego 4896420  1267455     NA     12
3812 3812  Overviewparrot   Retail      2012        NA    CA  Lewisville 4010895  6500823     NA     14
4445 4445      Comparejson   Retail      2018        NA    PA    Columbus 4304239  2191298     NA     10
```

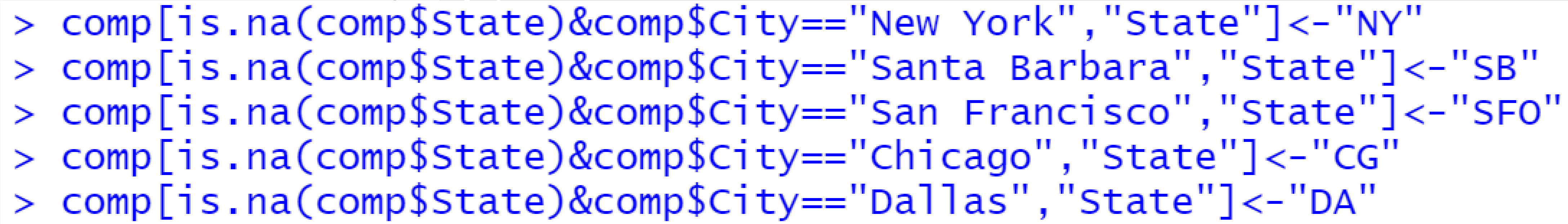


```
> comp[3,5]
[1] 231.5
> comp[2196,5]
[1] 231.5
> comp[3812,5]
[1] 231.5
> comp[4445,5]
[1] 231.5
```

State 채워넣기



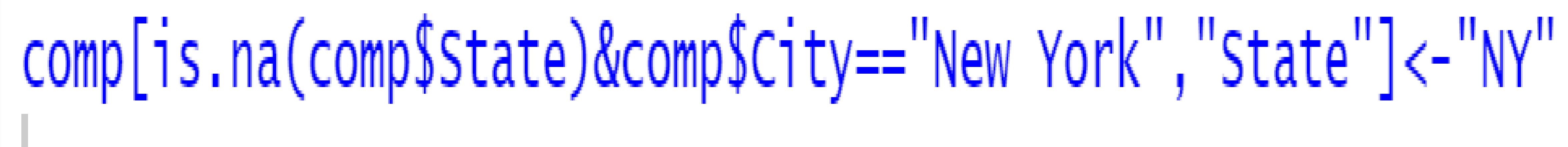
```
> comp[is.na(comp$State),]
       ID            Name           Industry Inception Employees State          City  Revenue Expenses
11     11 Canecorporation             Health      2012         6  <NA>      New York  5742668  7591189
79     79          Tonjob Financial Services      2010        87  <NA> Santa Barbara  1986877  2364775
82     82        Voyadexon             Health      2010       545  <NA>        Dallas  8913061  8763554
84     84        Drilldrill          Software      2010        30  <NA> San Francisco  6124180  2785799
173   173        Scotstrip           Software      2013        77  <NA>       Chicago  7743889   125635
267   267       Circlechop           Software      2010        14  <NA> San Francisco  6843806  5929828
379   379        Stovepuck             Retail      2013        73  <NA>      New York  7973785  5904502
767   767       Assurehelp       Construction      2006       420  <NA> San Francisco 12253828  3476282
1084 1084      Allpossible        IT Services      2021        83  <NA> San Francisco  5497391  5757389
1267 1267       Assurehelp           Software      2018       200  <NA> San Francisco 10802762  3476283
1584 1584      Allpossible             Retail      2017       236  <NA> San Francisco 19751914  6091557
```



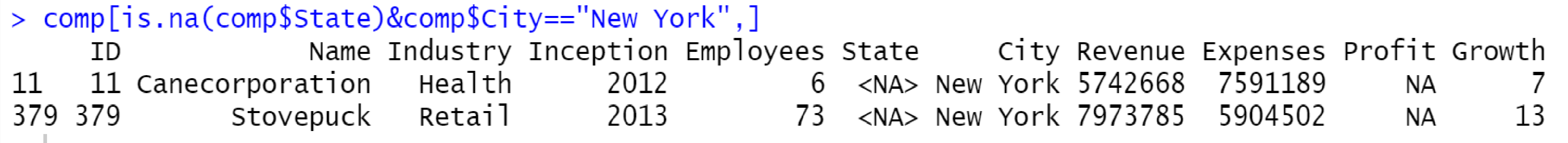
```
> comp[is.na(comp$State)&comp$City=="New York","State"]<-"NY"
> comp[is.na(comp$State)&comp$City=="Santa Barbara","State"]<-"SB"
> comp[is.na(comp$State)&comp$City=="San Francisco","State"]<-"SFO"
> comp[is.na(comp$State)&comp$City=="Chicago","State"]<-"CG"
> comp[is.na(comp$State)&comp$City=="Dallas","State"]<-"DA"
```

City가 New York일때 State NY로 만들기

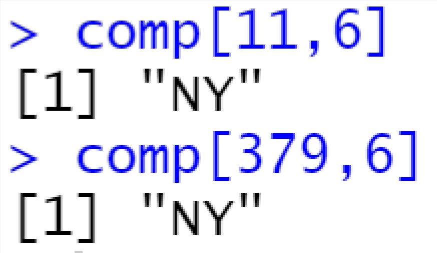
State의 결측치에 City가 New York인 것의 State를 NY로 하겠다



```
comp[is.na(comp$State)&comp$City=="New York","State"]<-"NY"
```

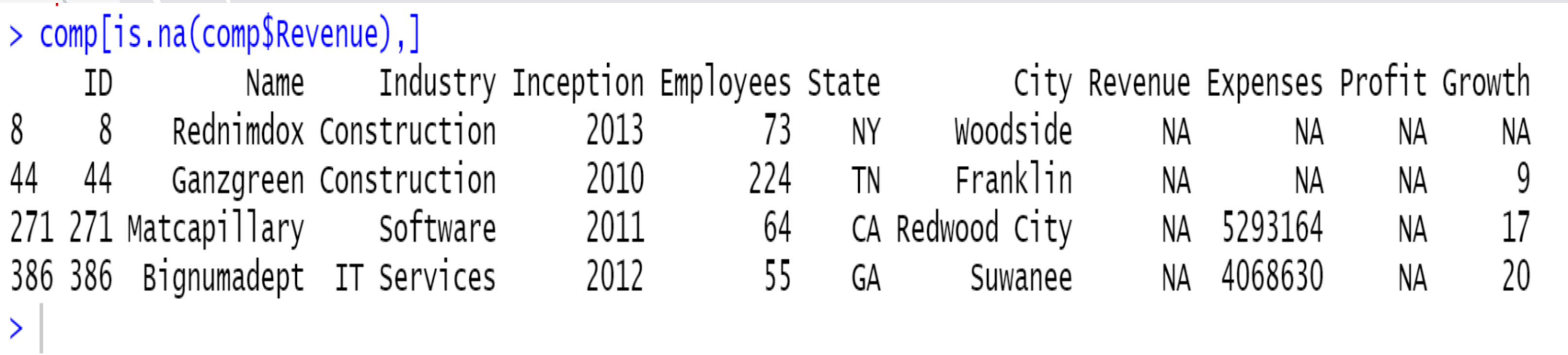


```
> comp[is.na(comp$State)&comp$City=="New York",]
     ID            Name Industry Inception Employees State     City Revenue Expenses Profit Growth
11   11 Canecorporation   Health      2012         6  <NA> New York 5742668  7591189     NA      7
379 379       Stovepuck   Retail      2013        73  <NA> New York 7973785  5904502     NA     13
```

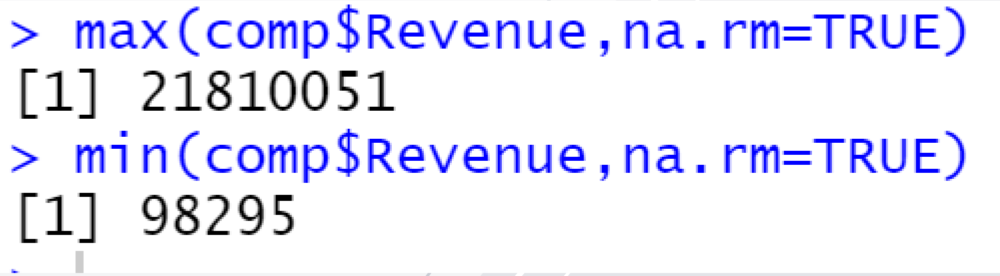


```
> comp[11,6]
[1] "NY"
> comp[379,6]
[1] "NY"
```

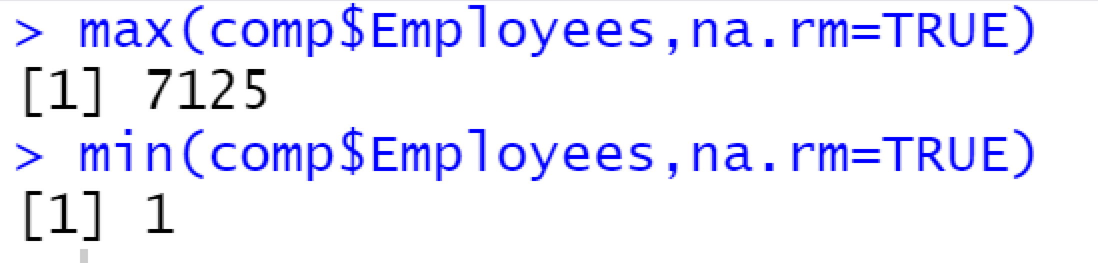
Revenue NA값 채우기



```
> comp[is.na(comp$Revenue),]
     ID        Name     Industry Inception Employees State         City Revenue Expenses Profit Growth
8     8    Rednimdox Construction      2013        73    NY     Woodside      NA       NA     NA     NA
44   44    Ganzgreen Construction      2010       224    TN     Franklin      NA       NA     NA      9
271 271 Matcapillary     Software      2011        64    CA Redwood City      NA  5293164     NA     17
386 386  Bignumadept  IT Services      2012        55    GA      Suwanee      NA  4068630     NA     20
> 
```



```
> max(comp$Revenue,na.rm=TRUE)
[1] 21810051
> min(comp$Revenue,na.rm=TRUE)
[1] 98295
```

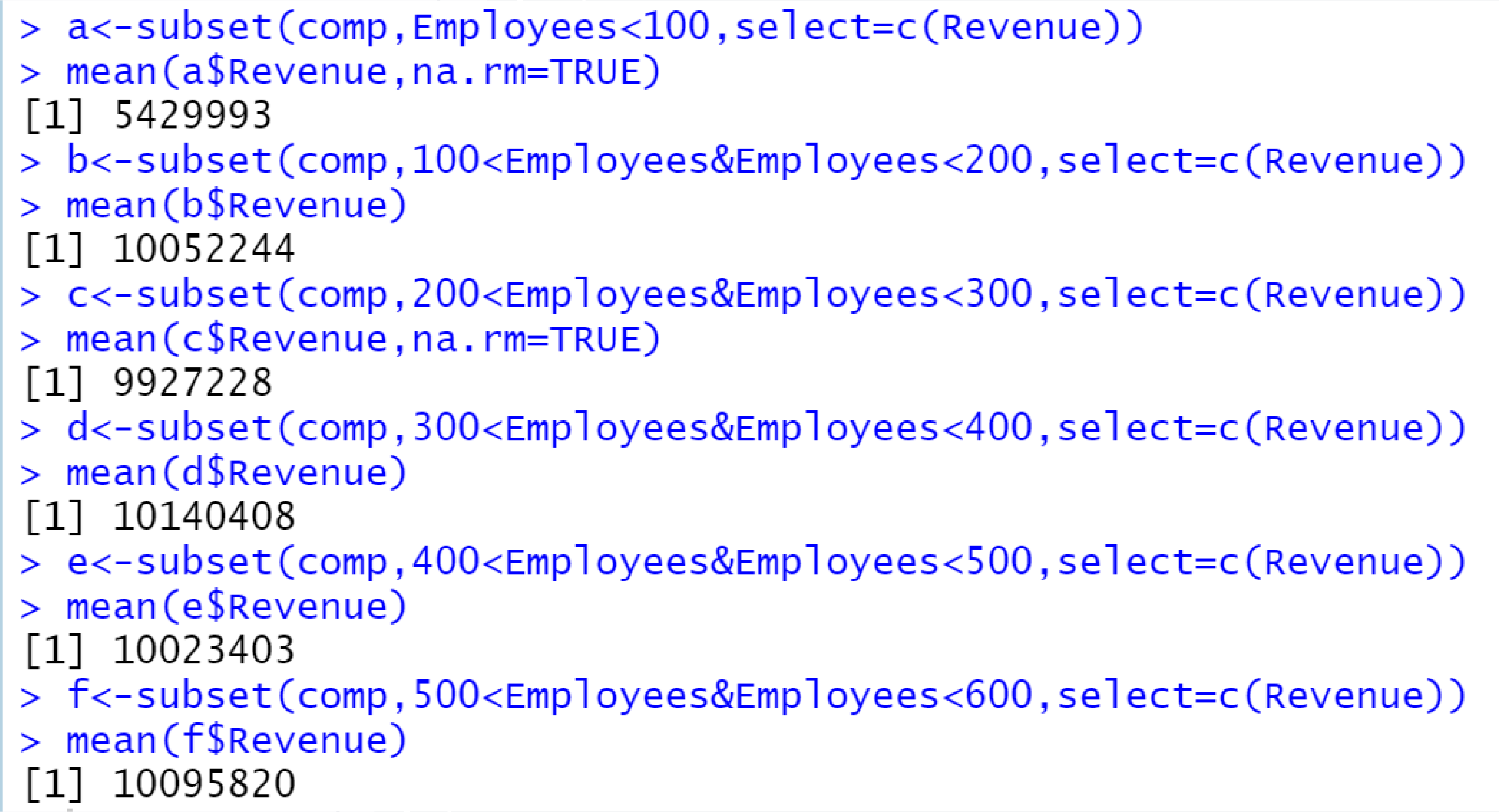


```
> max(comp$Employees,na.rm=TRUE)
[1] 7125
> min(comp$Employees,na.rm=TRUE)
[1] 1
```

큰 수익이 평균에 반영되는 것은 적절하지 않기 때문에 Mean을 사용 하는 것을 적절하지 않음

보통 수익이 많으면 종업원 수가 많기 때문에 종업원 수에 따라 NA을 채운다.

Employees수에 따른 Revenue 평균값



```
> a<-subset(comp,Employees<100,select=c(Revenue))
> mean(a$Revenue,na.rm=TRUE)
[1] 5429993
> b<-subset(comp,100<Employees&Employees<200,select=c(Revenue))
> mean(b$Revenue)
[1] 10052244
> c<-subset(comp,200<Employees&Employees<300,select=c(Revenue))
> mean(c$Revenue,na.rm=TRUE)
[1] 9927228
> d<-subset(comp,300<Employees&Employees<400,select=c(Revenue))
> mean(d$Revenue)
[1] 10140408
> e<-subset(comp,400<Employees&Employees<500,select=c(Revenue))
> mean(e$Revenue)
[1] 10023403
> f<-subset(comp,500<Employees&Employees<600,select=c(Revenue))
> mean(f$Revenue)
[1] 10095820
```

종업원이 100 미만일때 Revenue의 평균

종업원이 100 이상 200미만일때 Revenue의 평균

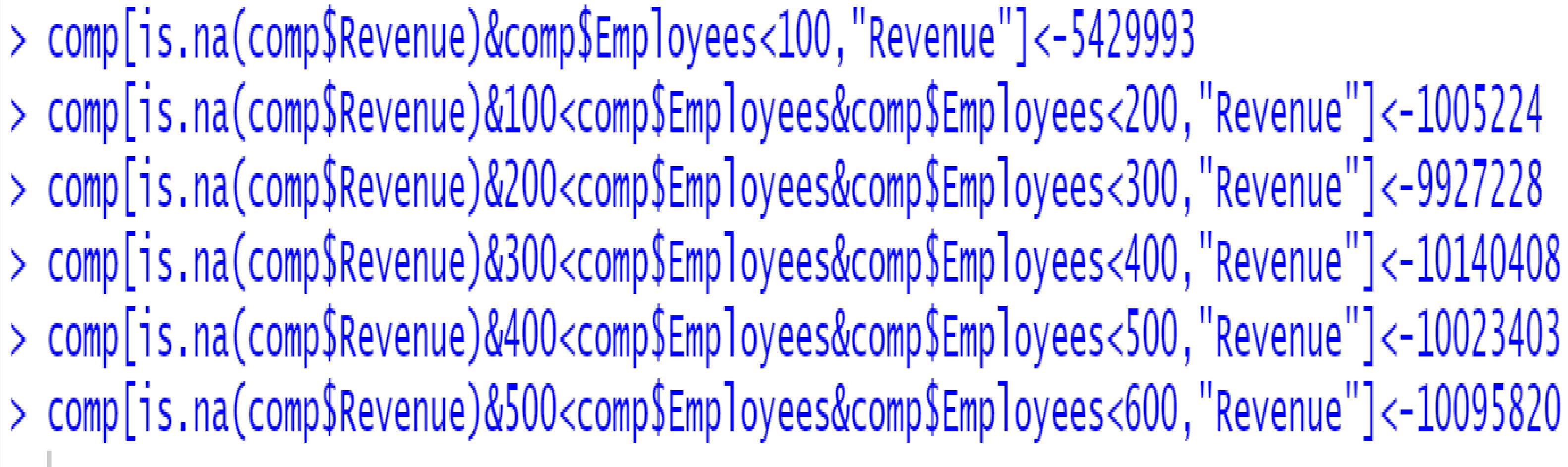
종업원이 200이상 300 미만일때 Revenue의 평균

종업원이 300이상 400 미만일때 Revenue의 평균

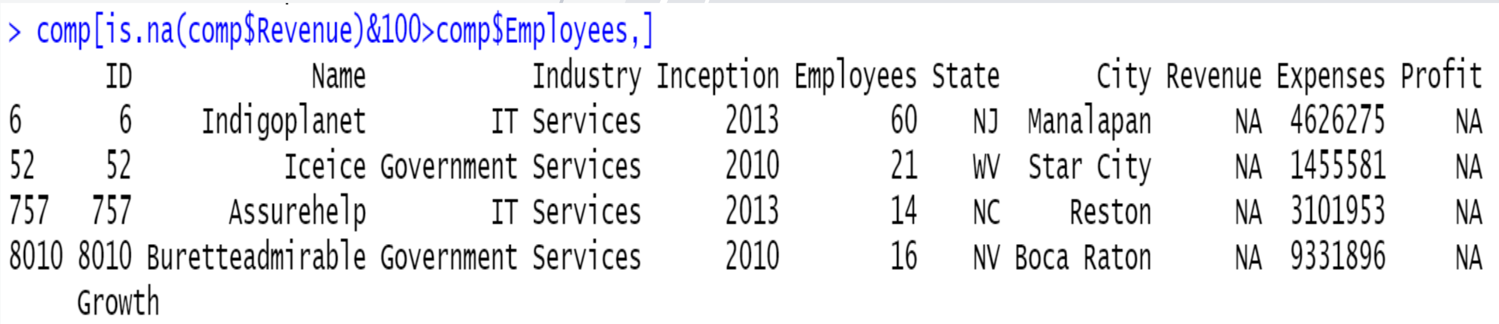
종업원이 400이상 500 미만일때 Revenue의 평균

종업원이 500이상 600 미만일때 Revenue의 평균

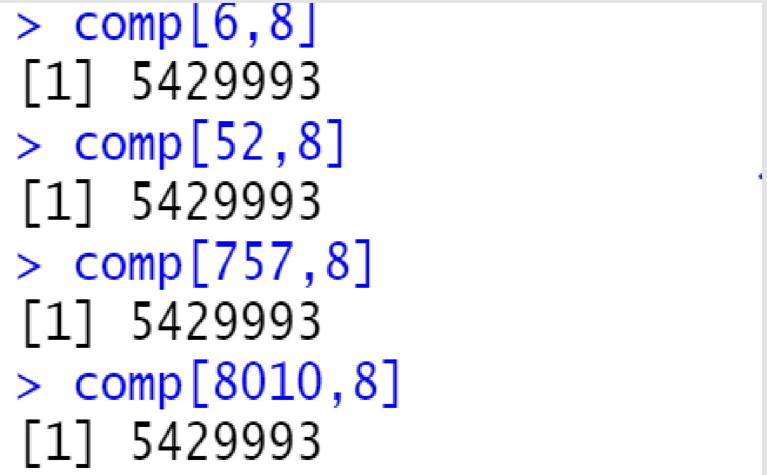
Revenue결측치 대체 및 확인



```
> comp[is.na(comp$Revenue)&comp$Employees<100,"Revenue"]<-5429993
> comp[is.na(comp$Revenue)&100<comp$Employees&comp$Employees<200,"Revenue"]<-1005224
> comp[is.na(comp$Revenue)&200<comp$Employees&comp$Employees<300,"Revenue"]<-9927228
> comp[is.na(comp$Revenue)&300<comp$Employees&comp$Employees<400,"Revenue"]<-10140408
> comp[is.na(comp$Revenue)&400<comp$Employees&comp$Employees<500,"Revenue"]<-10023403
> comp[is.na(comp$Revenue)&500<comp$Employees&comp$Employees<600,"Revenue"]<-10095820
```

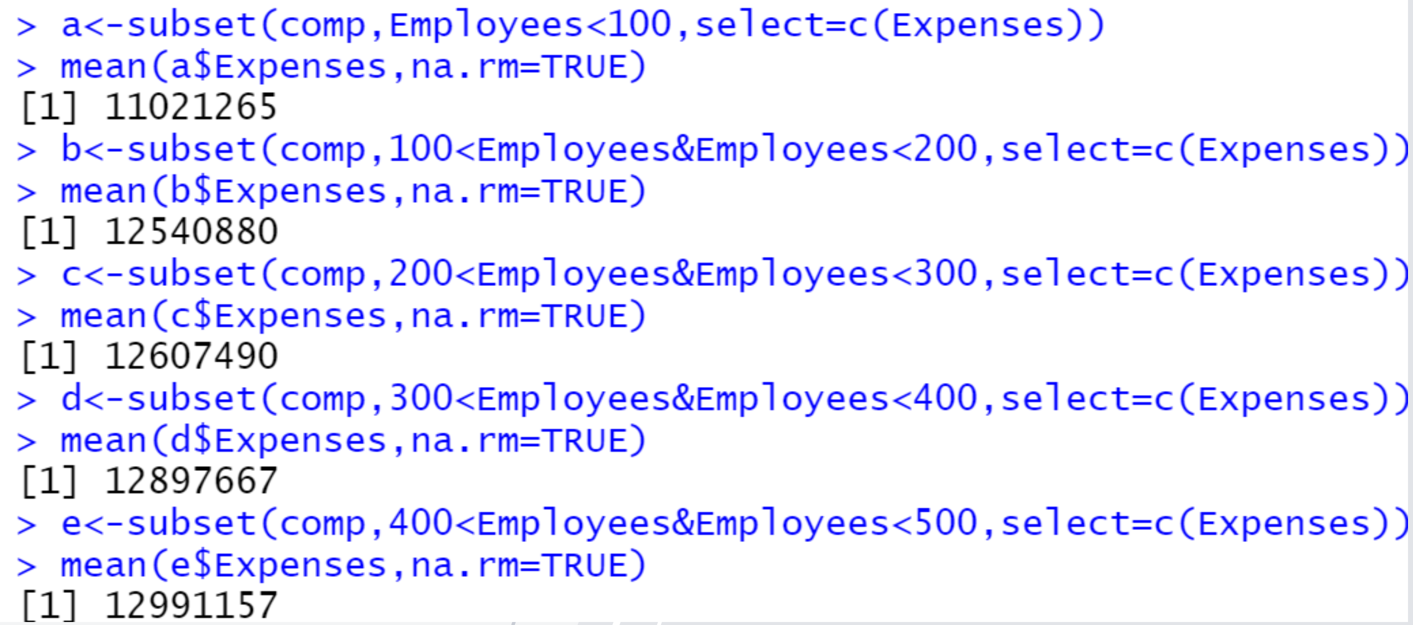


```
> comp[is.na(comp$Revenue)&100>comp$Employees,]
       ID             Name           Industry Inception Employees State       City Revenue Expenses Profit
6       6     Indigoplanet        IT Services      2013        60    NJ  Manalapan      NA  4626275     NA
52     52           Iceice Government Services      2010        21    WV  Star City      NA  1455581     NA
757   757       Assurehelp        IT Services      2013        14    NC     Reston      NA  3101953     NA
8010 8010 Buretteadmirable Government Services      2010        16    NV Boca Raton      NA  9331896     NA
     Growth
```

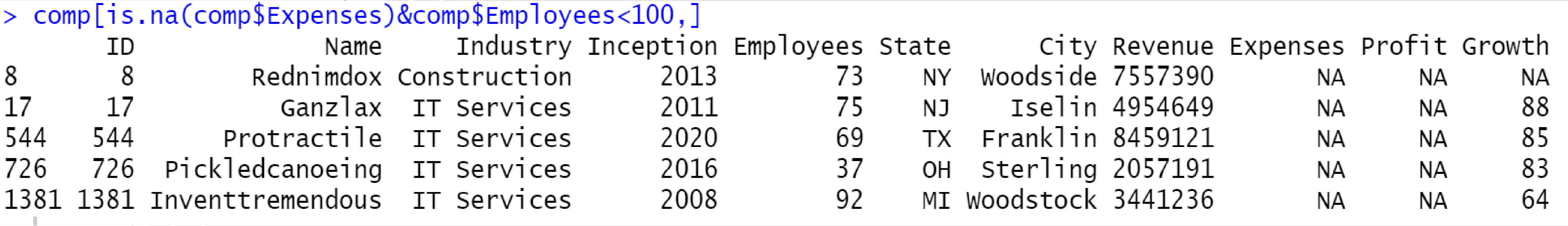


```
> comp[6,8]
[1] 5429993
> comp[52,8]
[1] 5429993
> comp[757,8]
[1] 5429993
> comp[8010,8]
[1] 5429993
```

Employees구간에 따른 Expenses의 평균값

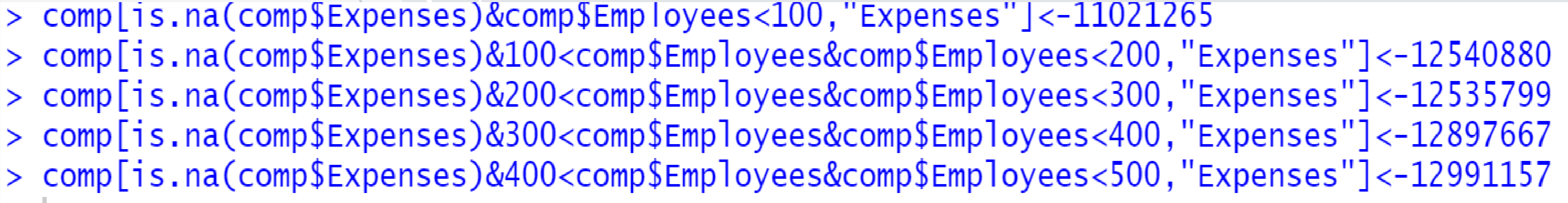


```
> a<-subset(comp,Employees<100,select=c(Expenses))
> mean(a$Expenses,na.rm=TRUE)
[1] 11021265
> b<-subset(comp,100<Employees&Employees<200,select=c(Expenses))
> mean(b$Expenses,na.rm=TRUE)
[1] 12540880
> c<-subset(comp,200<Employees&Employees<300,select=c(Expenses))
> mean(c$Expenses,na.rm=TRUE)
[1] 12607490
> d<-subset(comp,300<Employees&Employees<400,select=c(Expenses))
> mean(d$Expenses,na.rm=TRUE)
[1] 12897667
> e<-subset(comp,400<Employees&Employees<500,select=c(Expenses))
> mean(e$Expenses,na.rm=TRUE)
[1] 12991157
```

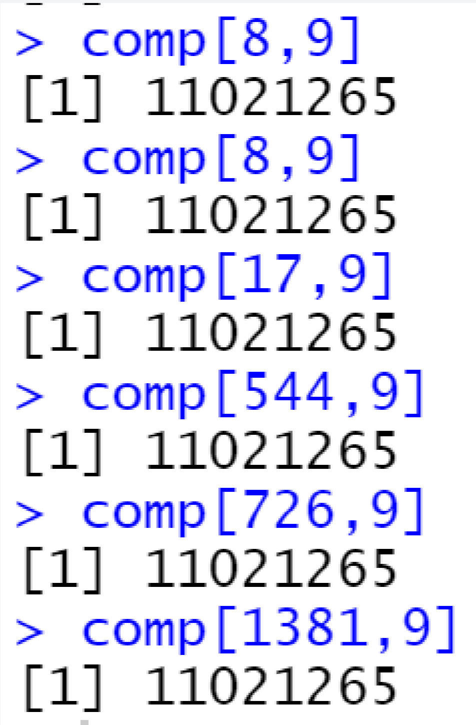


```
> comp[is.na(comp$Expenses)&comp$Employees<100,]
       ID               Name     Industry Inception Employees State      City Revenue Expenses Profit Growth
8       8          Rednimdox Construction      2013        73    NY  Woodside 7557390       NA     NA     NA
17     17            Ganzlax  IT Services      2011        75    NJ    Iselin 4954649       NA     NA     88
544   544         Protractile  IT Services      2020        69    TX  Franklin 8459121       NA     NA     85
726   726     Pickledcanoeing  IT Services      2016        37    OH  Sterling 2057191       NA     NA     83
1381 1381 Inventtremendous  IT Services      2008        92    MI Woodstock 3441236       NA     NA     64
```

Employees구간에 따른 Expenses의 평균값의 대체



```
> comp[is.na(comp$Expenses)&comp$Employees<100,"Expenses"]<-11021265
> comp[is.na(comp$Expenses)&100<comp$Employees&comp$Employees<200,"Expenses"]<-12540880
> comp[is.na(comp$Expenses)&200<comp$Employees&comp$Employees<300,"Expenses"]<-12535799
> comp[is.na(comp$Expenses)&300<comp$Employees&comp$Employees<400,"Expenses"]<-12897667
> comp[is.na(comp$Expenses)&400<comp$Employees&comp$Employees<500,"Expenses"]<-12991157
```

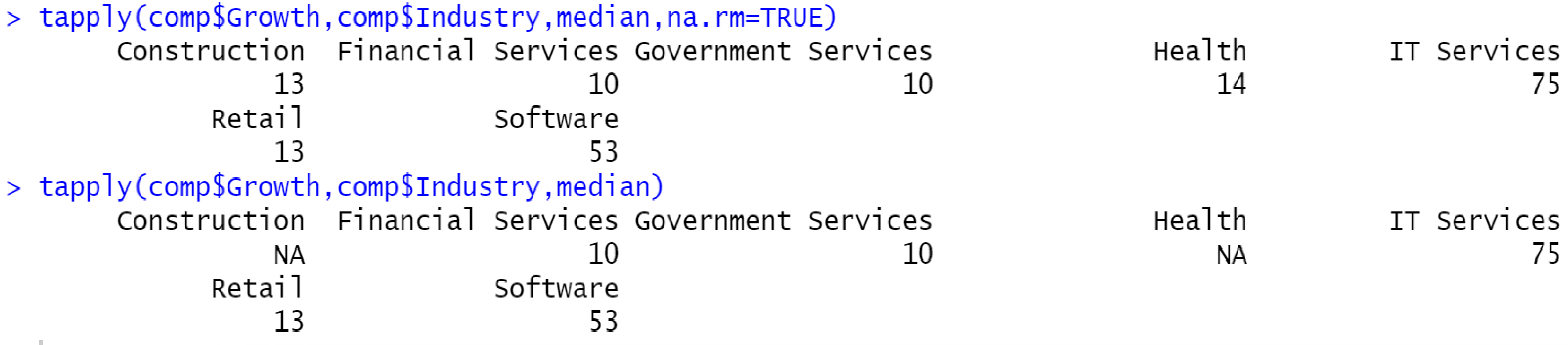


```
> comp[8,9]
[1] 11021265
> comp[8,9]
[1] 11021265
> comp[17,9]
[1] 11021265
> comp[544,9]
[1] 11021265
> comp[726,9]
[1] 11021265
> comp[1381,9]
[1] 11021265
```

Growth 변수 확인

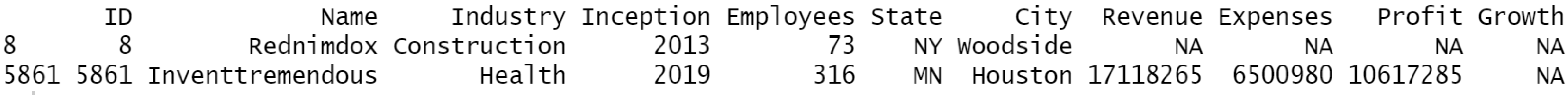
기업 성장률은 업종별 성장률과 밀접한 관련이 있다. 그래서 업종 성장률의 Median을 대입하려고 한다.

Health산업의 Growth의 NA는 14로 대체

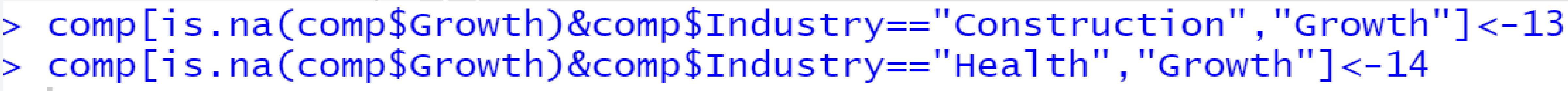


```
> tapply(comp$Growth,comp$Industry,median,na.rm=TRUE)
       Construction  Financial Services Government Services              Health         IT Services 
                 13                  10                  10                  14                  75 
             Retail            Software 
                 13                  53 
> tapply(comp$Growth,comp$Industry,median)
       Construction  Financial Services Government Services              Health         IT Services 
                 NA                  10                  10                  NA                  75 
             Retail            Software 
                 13                  53 
```

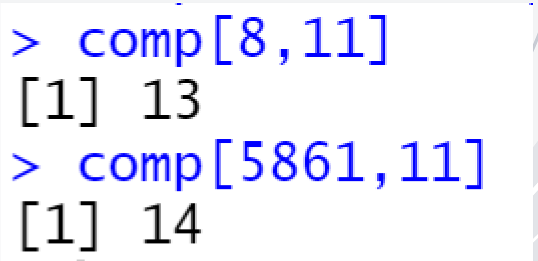
Growth 대체 및 대체 여부 확인



```
       ID             Name     Industry Inception Employees State     City  Revenue Expenses   Profit Growth
8       8        Rednimdox Construction      2013        73    NY Woodside       NA       NA       NA     NA
5861 5861 Inventtremendous       Health      2019       316    MN  Houston 17118265  6500980 10617285     NA
```

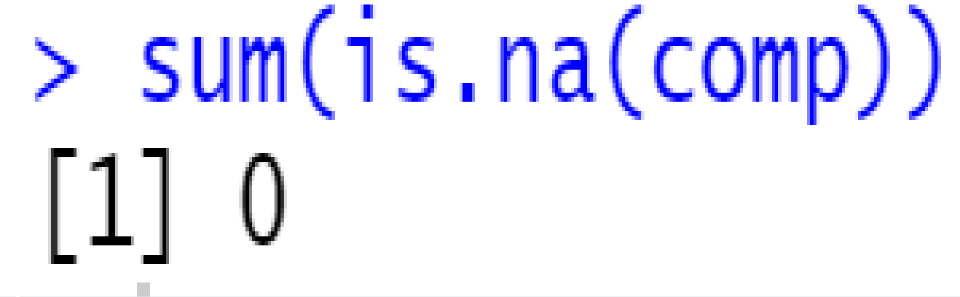


```
> comp[is.na(comp$Growth)&comp$Industry=="Construction","Growth"]<-13
> comp[is.na(comp$Growth)&comp$Industry=="Health","Growth"]<-14
```

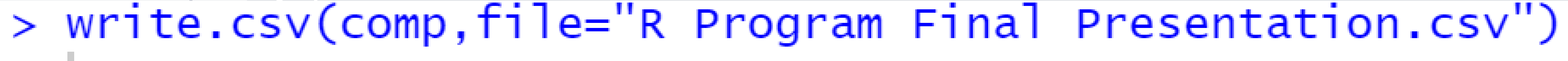


```
> comp[8,11]
[1] 13
> comp[5861,11]
[1] 14
```

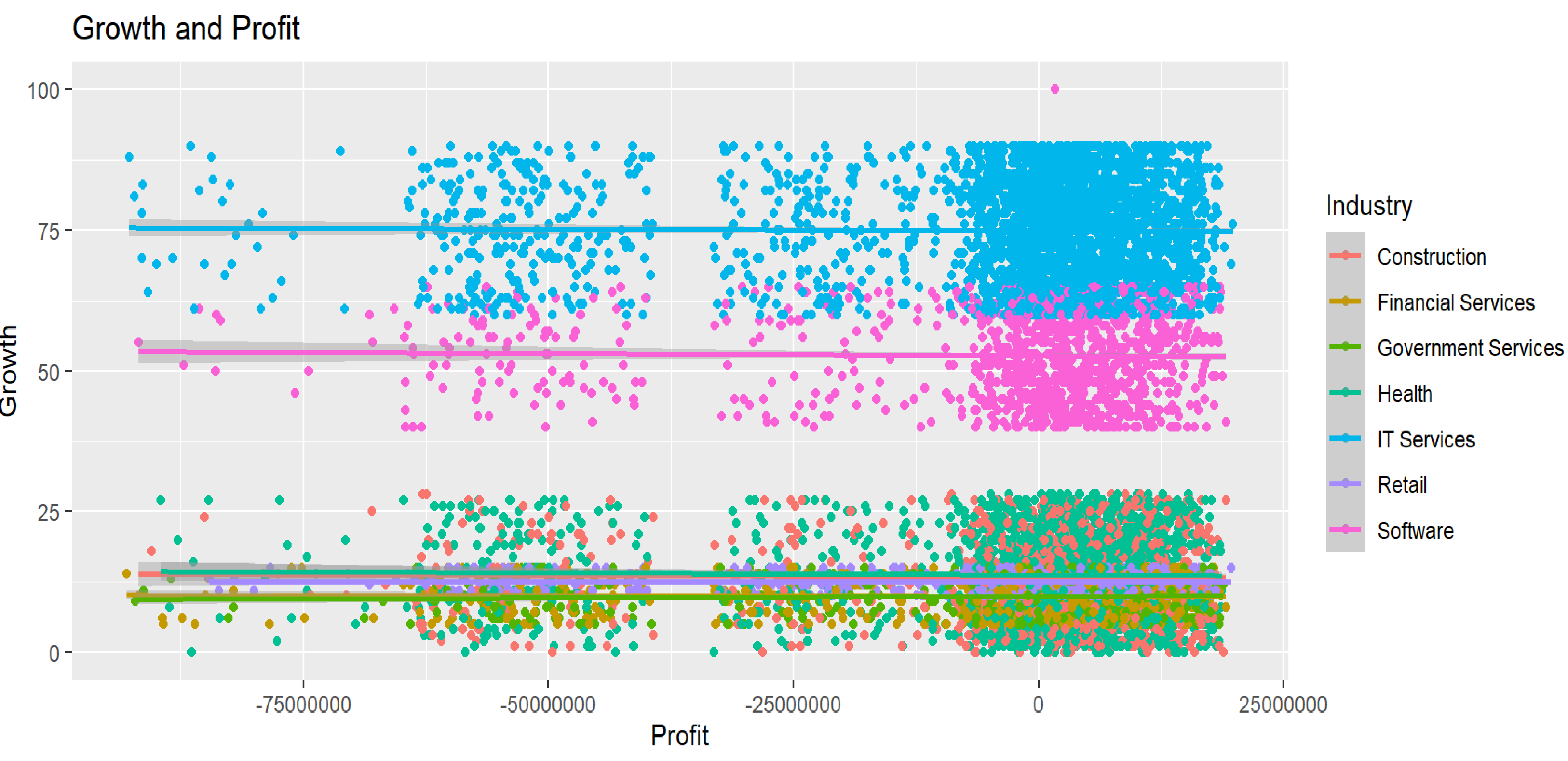
결측치 최종 확인 및 CSV파일 추출



```
> sum(is.na(comp))
[1] 0
```

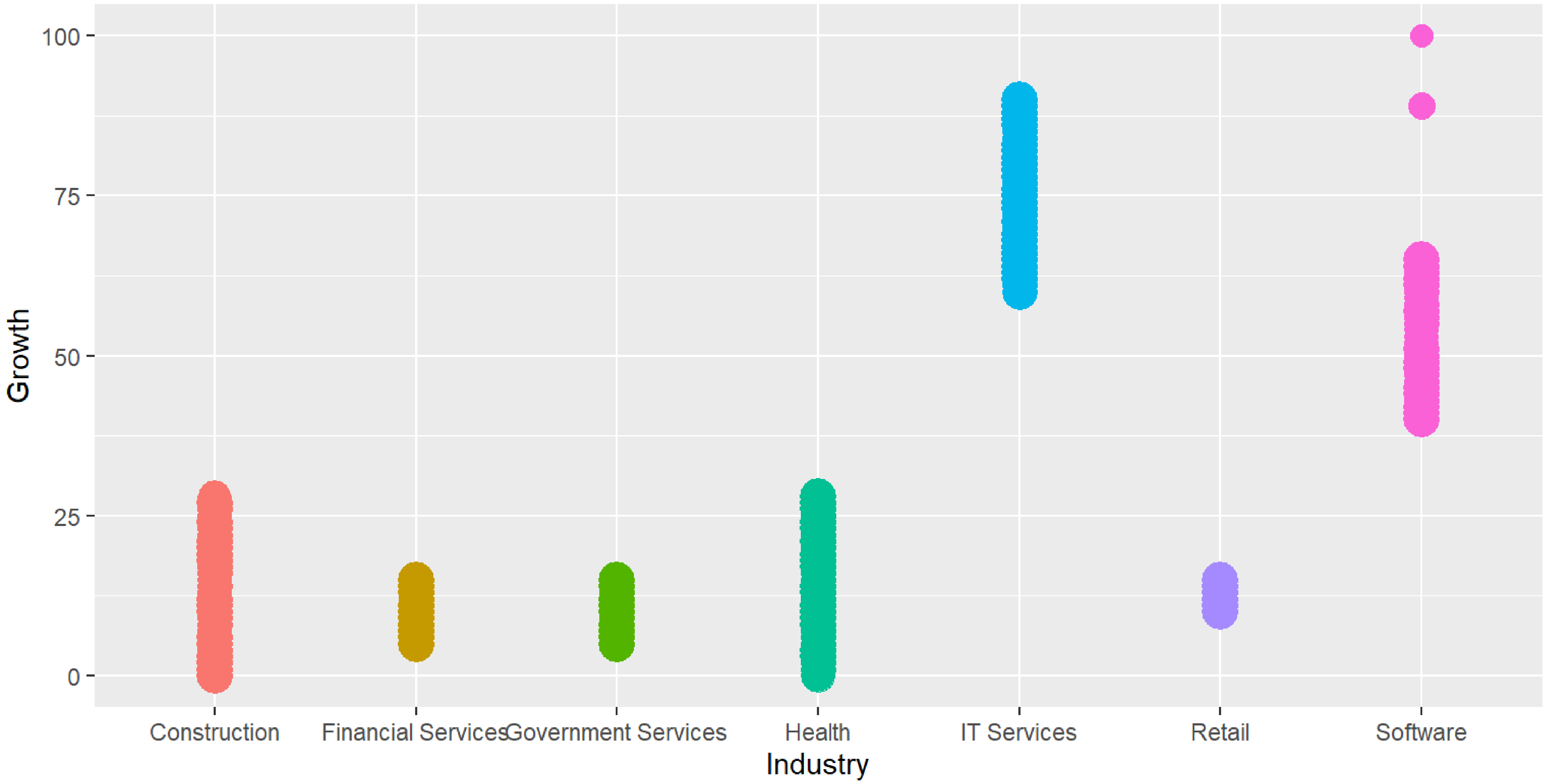


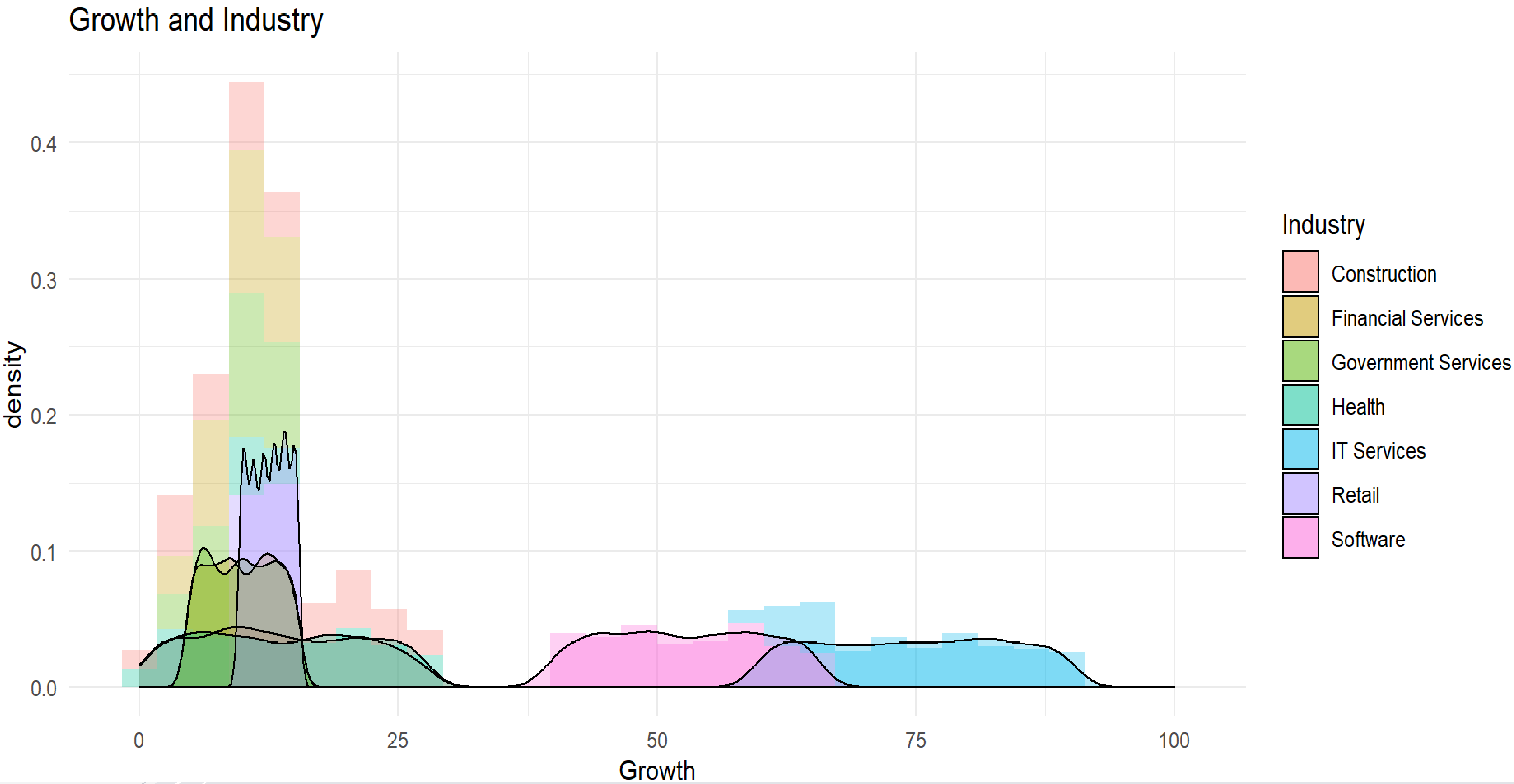
```
> write.csv(comp,file="R Program Final Presentation.csv")
```



Growth and Profit

Growth of an Industry




Growth and Industry
density
0.0
0.1
0.2
0.3
0.4
0
25
50
75
100
Growth
Industry
Construction
Financial Services
Government Services
Health
IT Services
Retail
Software

감사합니다