Future 500

For the “Future 500” magazine*. The stakeholders have supplied you a list of 500 companies and would like you to create some draft visualisations for their upcoming online publication. They have requested the following charts: . A scatterplot classified by industry showing revenue, expenses, profit . A scatterplot that includes industry trends for the expenses~revenue relationship . BoxPlots showing growth by industry Note that the dataset has numerous discrepancies that need to be addressed before analysis can be performed.

```
data <- read.csv("file:///C:/Users/badal/Desktop/datset_/P3-Future-500-The-
Dataset.csv", na.strings = c(""))
head(data,10)
```

```
##    ID         Name           Industry Inception Employees State
## 1   1     Over-Hex           Software      2006        25    TN
## 2   2    Unimattax        IT Services      2009        36    PA
## 3   3     Greenfax             Retail      2012        NA    SC
## 4   4    Blacklane        IT Services      2011        66    CA
## 5   5     Yearflex           Software      2013        45    WI
## 6   6 Indigoplanet        IT Services      2013        60    NJ
## 7   7      Treslam Financial Services      2009       116    MO
## 8   8    Rednimdox       Construction      2013        73    NY
## 9   9      Lamtone        IT Services      2009        55    CA
## 10 10    Stripfind Financial Services      2010        25    FL
##              City     Revenue          Expenses   Profit Growth
## 1        Franklin  $9,684,527 1,130,700 Dollars  8553827    19%
## 2  Newtown Square $14,016,543   804,035 Dollars 13212508    20%
## 3      Greenville  $9,746,272 1,044,375 Dollars  8701897    16%
## 4          Orange $15,359,369 4,631,808 Dollars 10727561    19%
## 5         Madison  $8,567,910 4,374,841 Dollars  4193069    19%
## 6       Manalapan $12,805,452 4,626,275 Dollars  8179177    22%
## 7         Clayton  $5,387,469 2,127,984 Dollars  3259485    17%
## 8        Woodside        <NA>              <NA>       NA   <NA>
## 9       San Ramon $11,757,018 6,482,465 Dollars  5274553    30%
## 10     Boca Raton $12,329,371   916,455 Dollars 11412916    20%
```

```
any(is.na(data))
```

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

```
str(data)
```

```
## 'data.frame':    500 obs. of  11 variables:
##  $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name     : Factor w/ 500 levels "Abstractedchocolat",..: 297 451 168 40
485 199 435 339 242 395 ...
##  $ Industry : Factor w/ 7 levels "Construction",..: 7 5 6 5 7 5 2 1 5 2
```

```
...
##  $ Inception: int  2006 2009 2012 2011 2013 2013 2009 2013 2009 2010 ...
##  $ Employees: int  25 36 NA 66 45 60 116 73 55 25 ...
##  $ State    : Factor w/ 42 levels "AL","AZ","CA",..: 36 33 35 3 41 27 22
29 3 8 ...
##  $ City     : Factor w/ 297 levels "Addison","Alexandria",..: 94 181 105
195 151 154 53 295 232 26 ...
##  $ Revenue  : Factor w/ 498 levels "$1,614,585","$1,835,717",..: 479 194
485 246 402 141 308 NA 96 117 ...
##  $ Expenses : Factor w/ 497 levels "1,026,548 Dollars",..: 6 485 3 248 227
247 57 NA 402 495 ...
##  $ Profit   : int  8553827 13212508 8701897 10727561 4193069 8179177
3259485 NA 5274553 11412916 ...
##  $ Growth   : Factor w/ 32 levels "-2%","-3%","0%",..: 14 16 11 14 14 18
12 NA 26 16 ...
```

```
summary(data)
```

```
##        ID                         Name                   Industry
##  Min.   :  1.0   Abstractedchocolat:  1   IT Services       :146
##  1st Qu.:125.8   Abusivebong       :  1   Health            : 86
##  Median :250.5   Acclaimedcirl     :  1   Software          : 64
##  Mean   :250.5   Admitruppell      :  1   Financial Services: 54
##  3rd Qu.:375.2   Admonishbadelynge :  1   Construction      : 50
##  Max.   :500.0   Ahemparticular    :  1   (Other)           : 98
##                  (Other)           :494   NA's              :  2
##    Inception      Employees          State               City
##  Min.   :1999   Min.   :   1.00   CA     : 57   San Diego  : 13
##  1st Qu.:2009   1st Qu.:  27.25   VA     : 50   New York   : 11
##  Median :2011   Median :  56.00   TX     : 47   Reston     : 10
##  Mean   :2010   Mean   : 148.61   FL     : 34   Houston    :  9
##  3rd Qu.:2012   3rd Qu.: 126.00   MD     : 25   Austin     :  8
##  Max.   :2014   Max.   :7125.00   (Other):283   Minneapolis:  8
##  NA's   :1      NA's   :2         NA's   :  4   (Other)    :441
##          Revenue                 Expenses        Profit
##  $1,614,585 :  1   1,026,548 Dollars:  1   Min.   :   12434
##  $1,835,717 :  1   1,040,662 Dollars:  1   1st Qu.: 3272074
##  $10,064,297:  1   1,044,375 Dollars:  1   Median : 6513366
##  $10,067,223:  1   1,097,353 Dollars:  1   Mean   : 6539474
##  $10,072,452:  1   1,117,206 Dollars:  1   3rd Qu.: 9303951
##  (Other)    :493   (Other)          :492   Max.   :19624534
##  NA's       :  2   NA's             :  3   NA's   :2
##      Growth
##  20%    : 39
##  19%    : 35
##  17%    : 27
##  6%     : 25
##  12%    : 24
##  (Other):349
##  NA's   :  1
```

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

```
## 'data.frame':    500 obs. of  11 variables:
##  $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name     : Factor w/ 500 levels "Abstractedchocolat",..: 297 451 168 40
485 199 435 339 242 395 ...
##  $ Industry : Factor w/ 7 levels "Construction",..: 7 5 6 5 7 5 2 1 5 2
...
##  $ Inception: int  2006 2009 2012 2011 2013 2013 2009 2013 2009 2010 ...
##  $ Employees: int  25 36 NA 66 45 60 116 73 55 25 ...
##  $ State    : Factor w/ 42 levels "AL","AZ","CA",..: 36 33 35 3 41 27 22
29 3 8 ...
##  $ City     : Factor w/ 297 levels "Addison","Alexandria",..: 94 181 105
195 151 154 53 295 232 26 ...
##  $ Revenue  : chr  "9684527" "14016543" "9746272" "15359369" ...
##  $ Expenses : chr  "1130700" "804035" "1044375" "4631808" ...
##  $ Profit   : int  8553827 13212508 8701897 10727561 4193069 8179177
3259485 NA 5274553 11412916 ...
##  $ Growth   : chr  "19" "20" "16" "19" ...
```

```
data$Inception <- factor(data$Inception)
data$Expenses <- as.numeric(data$Expenses)
data$Revenue<- as.numeric(data$Revenue)
data$Growth<-as.numeric(data$Revenue)
```

```
summary(data)
```

```
##        ID                        Name                   Industry
##  Min.   :  1.0   Abstractedchocolat:  1   IT Services       :146
##  1st Qu.:125.8   Abusivebong       :  1   Health            : 86
##  Median :250.5   Acclaimedcirl     :  1   Software          : 64
##  Mean   :250.5   Admitruppell      :  1   Financial Services: 54
##  3rd Qu.:375.2   Admonishbadelynge :  1   Construction      : 50
##  Max.   :500.0   Ahemparticular    :  1   (Other)           : 98
##                  (Other)           :494   NA's              :  2
##    Inception     Employees         State               City
##  2011   : 93   Min.   :   1.00   CA     : 57   San Diego  : 13
##  2010   : 83   1st Qu.:  27.25   VA     : 50   New York   : 11
##  2012   : 80   Median :  56.00   TX     : 47   Reston     : 10
##  2013   : 69   Mean   : 148.61   FL     : 34   Houston    :  9
##  2009   : 60   3rd Qu.: 126.00   MD     : 25   Austin     :  8
##  (Other):114   Max.   :7125.00   (Other):283   Minneapolis:  8
##  NA's   :  1   NA's   :2         NA's   :  4   (Other)    :441
##     Revenue            Expenses           Profit
##  Min.   : 1614585   Min.   :  71219   Min.   :   12434
##  1st Qu.: 8695702   1st Qu.:2758425   1st Qu.: 3272074
##  Median :10647231   Median :4365512   Median : 6513366
##  Mean   :10845170   Mean   :4310134   Mean   : 6539474
##  3rd Qu.:13106928   3rd Qu.:5832473   3rd Qu.: 9303951
##  Max.   :21810051   Max.   :9860686   Max.   :19624534
##  NA's   :2          NA's   :3         NA's   :2
##      Growth
##  Min.   : 1614585
##  1st Qu.: 8695702
##  Median :10647231
##  Mean   :10845170
##  3rd Qu.:13106928
##  Max.   :21810051
##  NA's   :2
```

```
#x<-impute(data$Employees,)
#x[!complete.cases(x),]
#x
```

```
data[!complete.cases(data),]
```

```
##      ID            Name           Industry Inception Employees State
## 3     3        Greenfax             Retail      2012        NA    SC
## 8     8       Rednimdox       Construction      2013        73    NY
## 11   11 Canecorporation             Health      2012         6  <NA>
## 14   14        Techline               <NA>      2006        65    CA
## 15   15         Cityace               <NA>      2010        25    CO
## 17   17         Ganzlax        IT Services      2011        75    NJ
## 22   22      Lathotline             Health      <NA>       103    VA
## 44   44       Ganzgreen       Construction      2010       224    TN
## 84   84      Drilldrill           Software      2010        30  <NA>
## 267 267      Circlechop           Software      2010        14  <NA>
## 332 332     Westminster Financial Services      2010        NA    MI
## 379 379       Stovepuck             Retail      2013        73  <NA>
##              City  Revenue Expenses   Profit   Growth
## 3      Greenville  9746272  1044375  8701897  9746272
## 8        Woodside       NA       NA       NA       NA
## 11       New York 10597009  7591189  3005820 10597009
## 14      San Ramon 13898119  5470303  8427816 13898119
## 15     Louisville  9254614  6249498  3005116  9254614
## 17         Iselin 14001180       NA 11901180 14001180
## 22         McLean  9418303  7567233  1851070  9418303
## 44       Franklin       NA       NA       NA       NA
## 84  San Francisco  7800620  2785799  5014821  7800620
## 267 San Francisco  9067070  5929828  3137242  9067070
## 332          Troy 11861652  5245126  6616526 11861652
## 379       New York 13814975  5904502  7910473 13814975
```

```
rownames(data)<- NULL
```

```
data <- data[!is.na(data$Inception),]
data <- data[!is.na(data$Industry),]

data[is.na(data$State),]

##      ID            Name Industry Inception Employees State          City
## 11   11 Canecorporation   Health      2012         6  <NA>      New York
## 84   84      Drilldrill Software      2010        30  <NA> San Francisco
## 267 267      Circlechop Software      2010        14  <NA> San Francisco
## 379 379       Stovepuck   Retail      2013        73  <NA>      New York
##      Revenue Expenses  Profit   Growth
## 11  10597009  7591189 3005820 10597009
## 84   7800620  2785799 5014821  7800620
## 267  9067070  5929828 3137242  9067070
## 379 13814975  5904502 7910473 13814975

data[is.na(data$State) & data$City == "New York",]

##      ID            Name Industry Inception Employees State     City
## 11   11 Canecorporation   Health      2012         6  <NA> New York
## 379 379       Stovepuck   Retail      2013        73  <NA> New York
##      Revenue Expenses  Profit   Growth
## 11  10597009  7591189 3005820 10597009
## 379 13814975  5904502 7910473 13814975

data[is.na(data$State) & data$City == "New York", "State"] <- "NY"
data[is.na(data$State) & data$City == "San Francisco", "State"] <- "CA"
data[c(11,83,266,378),]

##      ID            Name            Industry Inception Employees State
## 11   11 Canecorporation              Health      2012         6    NY
## 86   86   Treeelectrics Government Services      2013       485    VA
## 269 269  Sparrowchorizo         IT Services      2011       425    CO
## 381 381  Rattlemurrelet              Health      2013        12    GA
##            City  Revenue Expenses  Profit   Growth
## 11     New York 10597009  7591189 3005820 10597009
## 86  Clarksville 12208678  5536775 6671903 12208678
## 269  Broomfield 13163038  3439228 9723810 13163038
## 381   Woodstock  7496334  6186394 1309940  7496334

data[is.na(data$Employees),]

##      ID        Name           Industry Inception Employees State
## 3     3    Greenfax             Retail      2012        NA    SC
## 332 332 Westminster Financial Services      2010        NA    MI
##           City  Revenue Expenses  Profit   Growth
## 3   Greenville  9746272  1044375 8701897  9746272
## 332       Troy 11861652  5245126 6616526 11861652

median(data[,"Employees"], na.rm = T)

## [1] 56
```

```
med_retail <- median(data[data$Industry=="Retail","Employees"], na.rm = T)
med_retail
```

```
## [1] 28
```

```
data[is.na(data$Employees) & data$Industry == "Retail", "Employees"] <-
med_retail
data[c(3,331),]
```

```
##      ID      Name           Industry Inception Employees State
## 3     3  Greenfax             Retail      2012        28    SC
## 334 334 Tocantins Financial Services      2009       290    VA
##               City  Revenue Expenses   Profit   Growth
## 3      Greenville  9746272  1044375  8701897  9746272
## 334 Tysons Corner 14330107  2296074 12034033 14330107
```

```
med_fin <- median(data[data$Industry == "Financial Services","Employees"],
na.rm = T)
med_fin
```

```
## [1] 80
```

```
data[is.na(data$Employees) & data$Industry == "Financial Services",
"Employees"] <-med_fin
data[c(3,331),]
```

```
##      ID      Name           Industry Inception Employees State
## 3     3  Greenfax             Retail      2012        28    SC
## 334 334 Tocantins Financial Services      2009       290    VA
##               City  Revenue Expenses   Profit   Growth
## 3      Greenville  9746272  1044375  8701897  9746272
## 334 Tysons Corner 14330107  2296074 12034033 14330107
```

```
data[is.na(data$Growth),]
```

```
##    ID      Name     Industry Inception Employees State     City Revenue
## 8   8 Rednimdox Construction      2013        73    NY Woodside      NA
## 44 44 Ganzgreen Construction      2010       224    TN Franklin      NA
##    Expenses Profit Growth
## 8        NA     NA     NA
## 44       NA     NA     NA
```

```
med_con<- median(data[data$Industry =="Construction","Growth"],na.rm=T)
data[is.na(data$Growth)& data$Industry == "Construction", "Growth"] <-
med_con
data[c(8,43),]
```

```
##    ID      Name     Industry Inception Employees State       City  Revenue
## 8   8 Rednimdox Construction      2013        73    NY   Woodside       NA
## 46 46 Openjocon Construction      2013        75    IL Midlothian 11374343
##    Expenses  Profit   Growth
## 8        NA      NA  9055059
## 46  4273207 7101136 11374343
```

```
data[!complete.cases(data),]
```

```
##    ID      Name     Industry Inception Employees State     City  Revenue
## 8   8 Rednimdox Construction      2013        73    NY Woodside       NA
## 17 17   Ganzlax  IT Services      2011        75    NJ   Iselin 14001180
## 44 44 Ganzgreen Construction      2010       224    TN Franklin       NA
##    Expenses   Profit   Growth
## 8        NA       NA  9055059
## 17       NA 11901180 14001180
## 44       NA       NA  9055059
```

```
data[is.na(data$Revenue),]
```

```
##    ID      Name     Industry Inception Employees State     City Revenue
## 8   8 Rednimdox Construction      2013        73    NY Woodside      NA
## 44 44 Ganzgreen Construction      2010       224    TN Franklin      NA
##    Expenses Profit  Growth
## 8        NA     NA 9055059
## 44       NA     NA 9055059
```

```
mean_rev <- mean(data[data$Industry == "Construction","Revenue"],na.rm = T)
data[is.na(data$Revenue) & data$Industry=="Construction", "Revenue"] <-
mean_rev
data[c(8,43),]
```

```
##    ID      Name     Industry Inception Employees State       City  Revenue
## 8   8 Rednimdox Construction      2013        73    NY   Woodside  9158737
## 46 46 Openjocon Construction      2013        75    IL Midlothian 11374343
##    Expenses  Profit   Growth
## 8        NA      NA  9055059
## 46  4273207 7101136 11374343
```

```
data[is.na(data$Expenses),]
```

```
##    ID      Name     Industry Inception Employees State     City  Revenue
## 8   8 Rednimdox Construction      2013        73    NY Woodside  9158737
## 17 17   Ganzlax  IT Services      2011        75    NJ   Iselin 14001180
## 44 44 Ganzgreen Construction      2010       224    TN Franklin  9158737
##    Expenses   Profit   Growth
## 8        NA       NA  9055059
## 17       NA 11901180 14001180
## 44       NA       NA  9055059
```

```
med_exp <- median(data[,"Expenses"],na.rm = T)
data[is.na(data$Expenses) & data$Industry=="Construction", "Expenses"] <-
med_exp
data[c(8,17,43),]
```

```
##    ID      Name     Industry Inception Employees State       City  Revenue
## 8   8 Rednimdox Construction      2013        73    NY   Woodside  9158737
## 19 19     E-Zim       Retail      2008       320    OH     Monroe 10746451
## 46 46 Openjocon Construction      2013        75    IL Midlothian 11374343
##    Expenses  Profit   Growth
## 8   4307867      NA  9055059
## 19  4762319 5984132 10746451
## 46  4273207 7101136 11374343
```

```
data[is.na(data$Profit),]
```

```
##    ID      Name     Industry Inception Employees State     City Revenue
## 8   8 Rednimdox Construction      2013        73    NY Woodside 9158737
## 44 44 Ganzgreen Construction      2010       224    TN Franklin 9158737
##    Expenses Profit  Growth
## 8   4307867     NA 9055059
## 44  4307867     NA 9055059
```

```
data[is.na(data$Profit), "Profit"] <- data[is.na(data$Profit), "Revenue"] -
  data[is.na(data$Profit), "Expenses"]
data[c(8,41),]
```

```
##    ID      Name     Industry Inception Employees State     City Revenue
## 8   8 Rednimdox Construction      2013        73    NY Woodside 9158737
## 44 44 Ganzgreen Construction      2010       224    TN Franklin 9158737
##    Expenses  Profit  Growth
## 8   4307867 4850870 9055059
## 44  4307867 4850870 9055059
```

```
data[is.na(data$Expenses), "Expenses"] <- data[is.na(data$Expenses),
"Revenue"] -           data[is.na(data$Expenses), "Profit"]
data[c(17),]
```

```
##    ID  Name Industry Inception Employees State   City  Revenue Expenses
## 19 19 E-Zim   Retail      2008       320    OH Monroe 10746451  4762319
##     Profit   Growth
## 19 5984132 10746451
```

```
data[!complete.cases(data),]
```

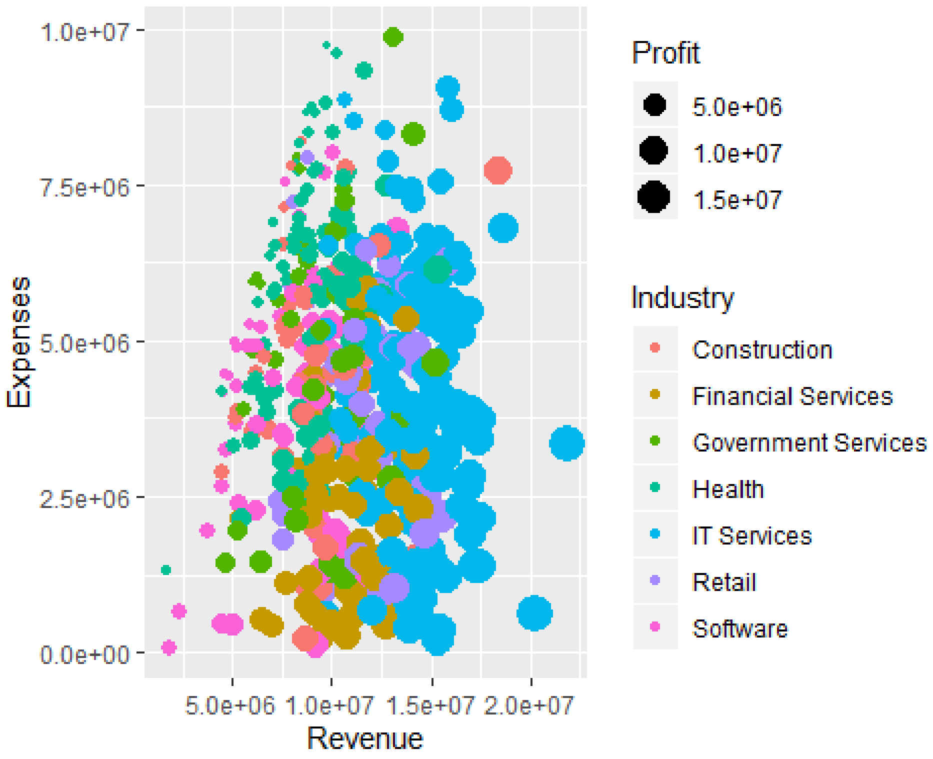
```
##  [1] ID        Name      Industry  Inception Employees State     City
##  [8] Revenue   Expenses  Profit    Growth
## <0 rows> (or 0-length row.names)
```

VIZ

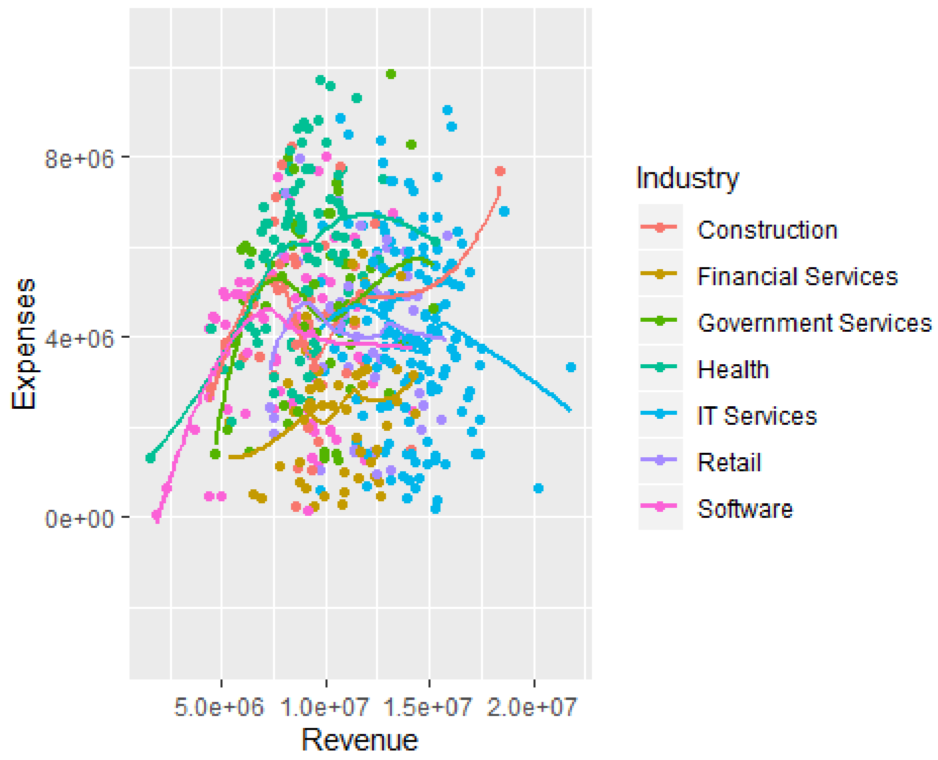
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.1
```

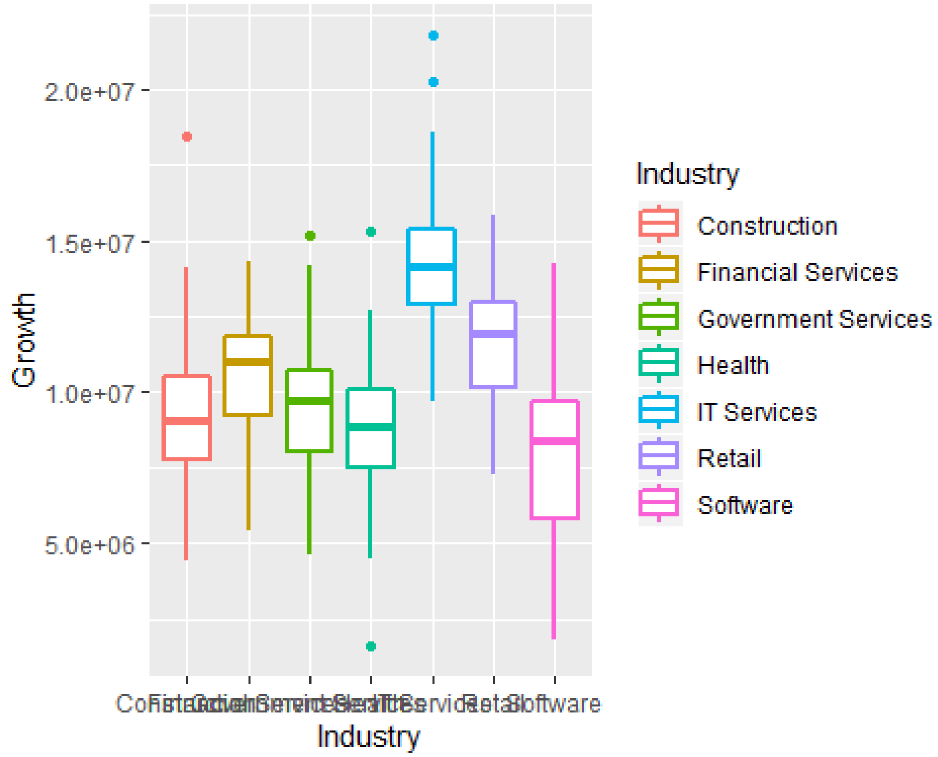
```
plot <- ggplot(data)
plot + geom_point(aes(x=Revenue, y = Expenses,
                  color = Industry, size = Profit))
```



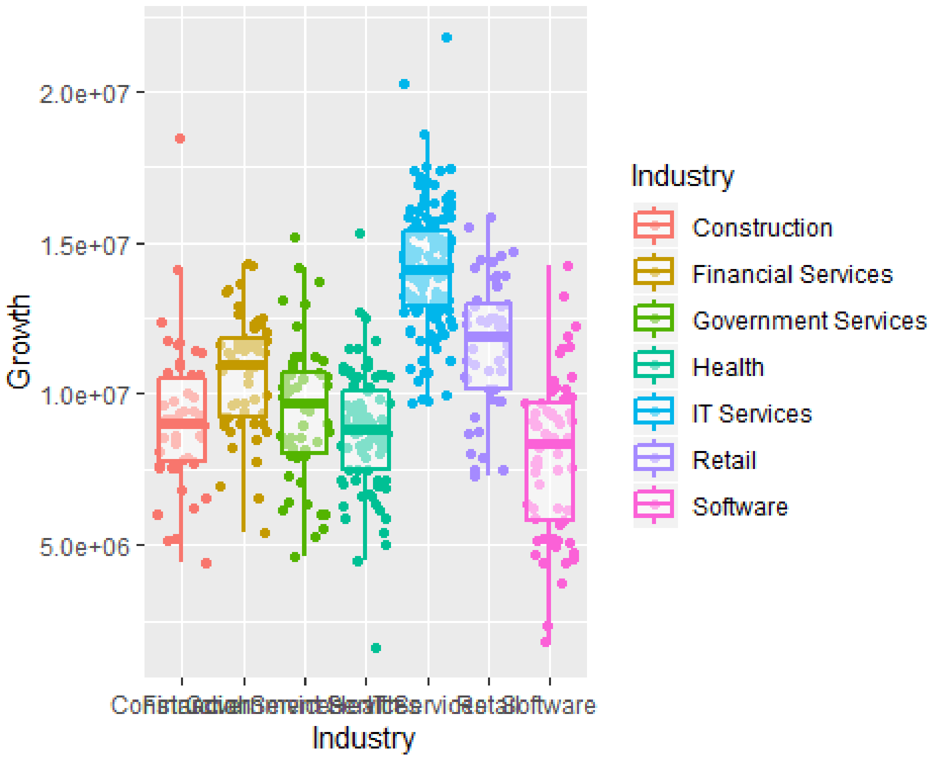
```
t <- ggplot(data, aes(x=Revenue, y= Expenses, color = Industry,
                  ))
t + geom_point()+ geom_smooth(fill = NA, size = 1.0)

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
b <- ggplot(data, aes(x = Industry, y= Growth,
                      color = Industry))
b + geom_boxplot(size = 0.8)
```



```
b + geom_jitter()+
  geom_boxplot(size= 1, alpha= 0.5, outlier.colour = NA)
```



```
summary(data[(data$Growth) & data$Industry == 'IT Services', "Growth"])

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##  9691133 12885778 14087386 14154748 15374274 21810051
```