Companies Database Visualization

Rishab Ravi

This project aims to clean an input excel files which contains a list of companies and provide some insights/patterns of their distribution.

Initial preprocessing:

```
##import necessary libraries

library(openxlsx)

mainDF <- read.xlsx("DATABASE of Director and CEO.xlsx", sheet = 1, colNames = TRUE, rowNames = FALSE,sh
#remove duplicate companies
workDF <- subset(mainDF,!duplicated(mainDF$Company.Name))
#remove duplicate mail IDs
workDF2 <- subset(workDF,!duplicated(workDF$Email.ID,incomparables = NA))

workDF2 <- workDF2[,colSums(is.na(workDF2))<nrow(workDF2)] # remove empty cols
```

Further preprocessing:

```
#assign serial no.
x <- c(1:23543)
workDF2$Sr..No<-x

workDF3 <- workDF2
```

Cleaning Data frame:

```
#region
workDF3$Region<-tolower(workDF3$Region)
workDF3$Region <- gsub('\\s+', '', workDF3$Region)

colnames(workDF3)<-tolower(colnames(workDF3))

workDF3$country<-tolower(workDF3$country)
workDF3$country <- gsub('\\s+', '', workDF3$country)
unique(workDF3$country)
```

```
## [1] "india" NA
```

```
##industry
workDF3$industry<-tolower(workDF3$industry)
workDF3$industry <- gsub('\\s+', '', workDF3$industry)

workDF3$industry[workDF3$industry=="+91(040)27819327"|
                   workDF3$industry=="-"|
                   workDF3$industry=="ignore"|
                   workDF3$industry=="yes"] <- NA

workDF3$industry[workDF3$industry=="networking&telecommunication"]<-"networking&telecommunications"
```

```
workDF3$industry[workDF3$industry=="it&ites"]<-"it/ites"

workDF3$industry[workDF3$industry=="lifescience"]<-"lifesciences"
workDF3$industry[workDF3$industry=="e-comerce"]<-"e-commerce"

workDF3$industry[workDF3$industry=="aerospace&defence"]<-"aerospace&defense"
workDF3$industry[workDF3$industry=="automobile,autoancillaries"]<-"automobile&autoancillaries"
workDF3$industry[workDF3$industry=="chemical"]<-"chemicals"
workDF3$industry[workDF3$industry=="electricalandelectronics"]<-"electrical&electronics"
workDF3$industry[workDF3$industry=="gems&jwellery"]<-"gems&jewellery"

workDF3$industry[workDF3$industry=="logistics&tranportation"|
                   workDF3$industry=="logisticsandtransportation"]<-"logistics&transportation"


workDF3$industry[workDF3$industry=="telecommunication&networking"|
                   workDF3$industry=="telecommunications&networking"]<-"networking&telecommunications"


workDF3$industry[workDF3$industry=="textileandgarments"|
                   workDF3$industry=="textile&garments"]<-"textiles&garments"

workDF3$industry[workDF3$industry=="retailandtrading"]<-"retail&trading"

workDF3$industry[workDF3$industry=="energyandutility"]<-"energy&utilities"


##cities

workDF3$city<-tolower(workDF3$city)
workDF3$city <- gsub('\\s+', '', workDF3$city)

##states
workDF3$state<-tolower(workDF3$state)
workDF3$state <- gsub('\\s+', '', workDF3$state)


workDF3$state[workDF3$state=="chhattisgarh"] <- "chattisgarh"

workDF3$state[workDF3$state=="chhattisgarh"] <- "chattisgarh"

workDF3$state[workDF3$state=="gujrat"] <- "gujarat"

workDF3$state[workDF3$state=="hyderabad"|workDF3$state=="teleangna"|
                workDF3$state=="telengna"] <- "telangana"

workDF3$state[workDF3$state=="newdelhi"] <- "delhi"

workDF3$state[workDF3$state=="kerela"] <- "kerala"

workDF3$state[workDF3$state=="kakinada"|workDF3$state=="karanataka"|
                workDF3$state=="karnatka"|
                workDF3$state=="kranataka"] <- "karnataka"
```

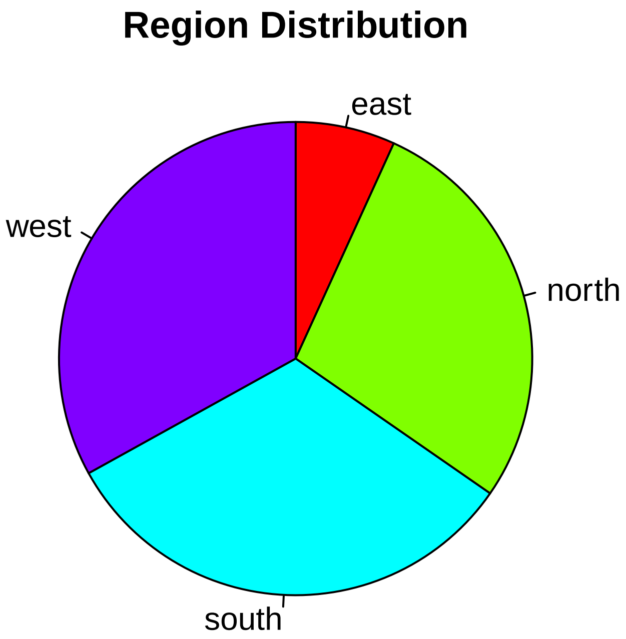
```
workDF3$state[workDF3$state=="odisha"] <- "orissa"

workDF3$state[workDF3$state=="uttaranchal"] <- "uttarakhand"

workDF3$state[workDF3$state=="maharastra"|workDF3$state=="maharshtra"|
                workDF3$state=="maharstra"|workDF3$state=="mahrashtra"|
                workDF3$state=="mahrashtra400007"|workDF3$state=="mahrastra"|
                workDF3$state=="mahrshtra"|workDF3$state=="kolhapur"|
                workDF3$state=="mumbai"] <- "maharashtra"

workDF3$state[workDF3$state=="andaman&nicobar"] <- "unionterritory"
workDF3$state[workDF3$state=="chennai"] <- "tamilnadu"
workDF3$state[workDF3$state=="-"] <- NA
```

Plotting distributions:

```
library(ggplot2)
```

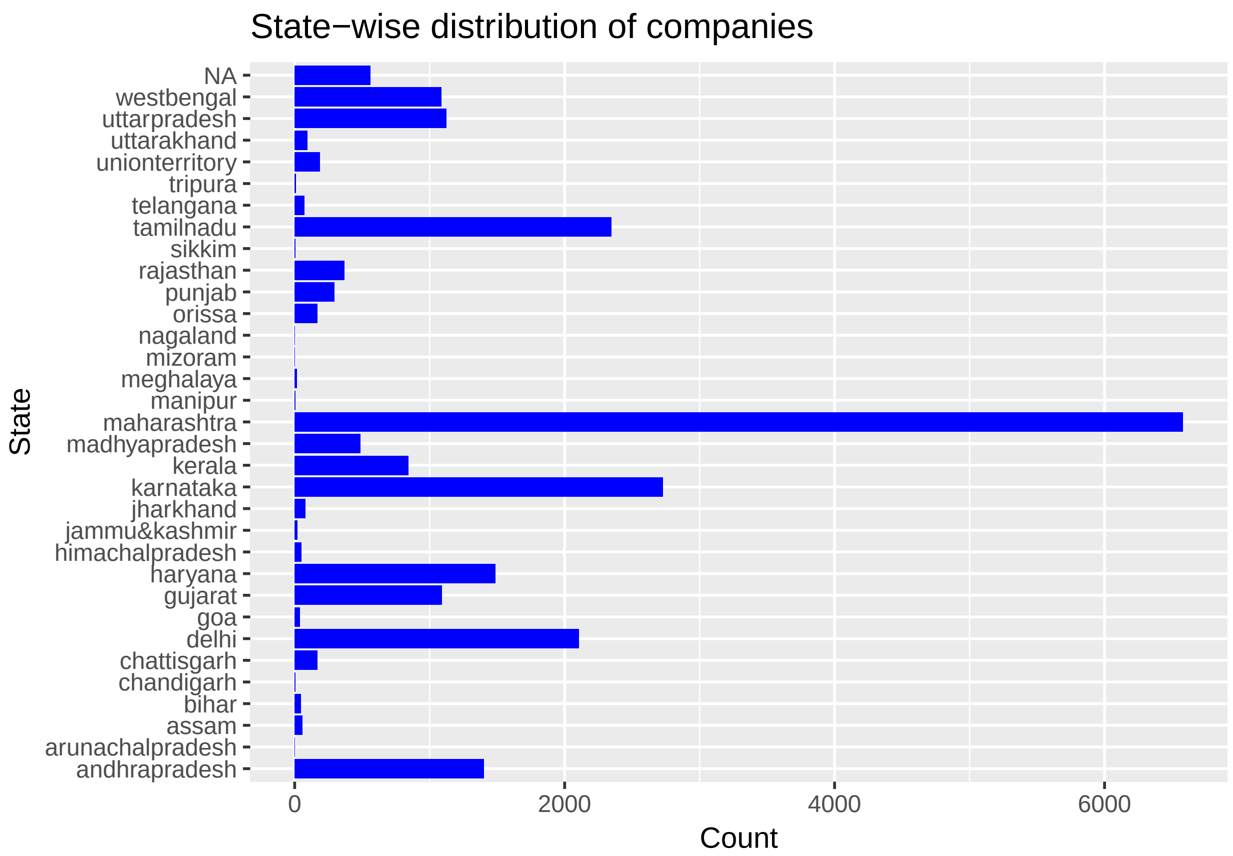
piechart of regions

```
pie(table(workDF3$region),
    clockwise = TRUE, main="Region Distribution",
    radius = 1,col=rainbow(4))
```



barplot of states

```r
ggplot(workDF3, aes(state)) + geom_bar(fill = "blue") + theme_bw() +
  xlab("State") + ylab("Count") + coord_flip() +
  labs(title = "State-wise distribution of companies") + theme_gray()
```



industry type information

```
industryType<-workDF3$industry
as.data.frame(table(industryType))
```

```
##                                            industryType Freq
## 1                                     aerospace&defense   18
## 2                                   agro&foodprocessing    5
## 3                          apparel&othertextileproducts    1
## 4                            automobile&autoancillaries  550
## 5                                           automobiles    1
## 6                                            automotive  300
## 7                                        automotive-oem   36
## 8                              automotive&autoacilaries   63
## 9                                                  bfsi 2623
## 10                                    braverageandgoods    4
## 11                                     businessservices  917
## 12                                            chemicals  194
## 13                             computersoftware&it/ites    2
## 14                              construction&realestate    2
## 15                                consumerpackagedgoods  315
## 16                                          diversified  115
## 17                                           e-commerce    4
## 18                                            education 2760
## 19                               electrical&electronics  298
## 20                                 energy&utilities  310
## 21                                      engineering  461
## 22                                              epc    8
## 23                                 epc-construction    1
## 24                               fashionaccessories    8
## 25                                food&agroindustry    1
## 26                                   gems&jewellery   39
## 27                                   government&ngo  739
## 28      hightech&communications-consumerelectronics    2
## 29     hightech&communications-electroniccomponents   15
## 30  hightech&communications-electronicmanufacturing   12
## 31                                     hometextiles    3
## 32                                      hospitality  510
## 33      industrialmanufacturing-consumerelectronics    1
## 34      industrialmanufacturing-industrialequipment    7
## 35      industrialmanufacturing-industrialmachinery    7
## 36 industrialmanufacturing-metalcastingandfoundries    7
## 37                                   infrastructure  652
## 38                                          it/ites 4594
## 39                            leatherandsportsgoods    1
## 40                                     lifesciences  879
## 41                         logistics&transportation  306
## 42                                    manufacturing   26
## 43                         manufacturing&production 1750
## 44                       manufacturingandproduction    3
## 45                              media&entertainment  648
## 46                                     metal&mining  268
## 47                    networking&telecommunications  259
## 48                                           others    9
## 49                                        packaging   31
## 50                                   retail&trading  154
## 51                                textiles&garments  509
```

mean employee size (approx.)

```
a <- as.integer(workDF3$employee.range)
```

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

```
b <- a[!is.na(a)]
mean(b)
```

```
## [1] 1338.5
```

So, this report has given some insights of the distribution of several companies in the country.