Project 3

Alain T Kuiete

10/17/2019

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

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

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

```
library(tidyr)
```

Data Science Skills

Data Science General Skills

Downloading data

```
dsg <- read.csv("https://raw.githubusercontent.com/AlainKuiete/DATA607/master/ds_general_skills_revised
str(dsg)
```

```
## 'data.frame':    30 obs. of  5 variables:
##  $ Keyword    : Factor w/ 27 levels "","\"data scientist\" \"[keyword]\"",..: 14 8 25 11 10 15 27 7 1
##  $ LinkedIn   : Factor w/ 23 levels "","1,212","1,310",..: 16 15 12 11 8 6 5 4 3 2 ...
##  $ Indeed     : Factor w/ 23 levels "","1,125","1,413",..: 11 12 8 7 6 4 3 2 23 22 ...
##  $ SimplyHired: Factor w/ 23 levels "","1,153","1,497",..: 10 11 9 8 4 3 2 23 22 21 ...
##  $ Monster    : Factor w/ 23 levels "","1,207","1,815",..: 7 10 8 4 6 3 2 22 18 17 ...
```

Subsetting the Data Science soft Skills

```
dskg <-dsg[1:15,]
```

reshaping my dataframe

```
colnames(dskg) <- c("D.Skills", "LinkedIn", "Indeed", "SimplyHired", "Monster")
dskg$LinkedIn <- as.numeric(gsub(",","",dskg$LinkedIn))
dskg$Indeed <- as.numeric(gsub(",","",dskg$Indeed))
dskg$SimplyHired <- as.numeric(gsub(",","",dskg$SimplyHired))
dskg$Monster <- as.numeric(gsub(",","",dskg$Monster))
```

Computation

```
s.dskg <- summarise(dskg, sL=sum(LinkedIn, na.rm=TRUE), sI= sum(Indeed, na.rm=TRUE), sS=sum(SimplyHired
           sM=sum(Monster, na.rm=TRUE))

tsg <- sum(s.dskg)

g.skills <- dskg%>% mutate(D.Skills, pct=(LinkedIn+Indeed+SimplyHired+Monster)/tsg)%>%
  select(D.Skills,pct)
g.skills
```

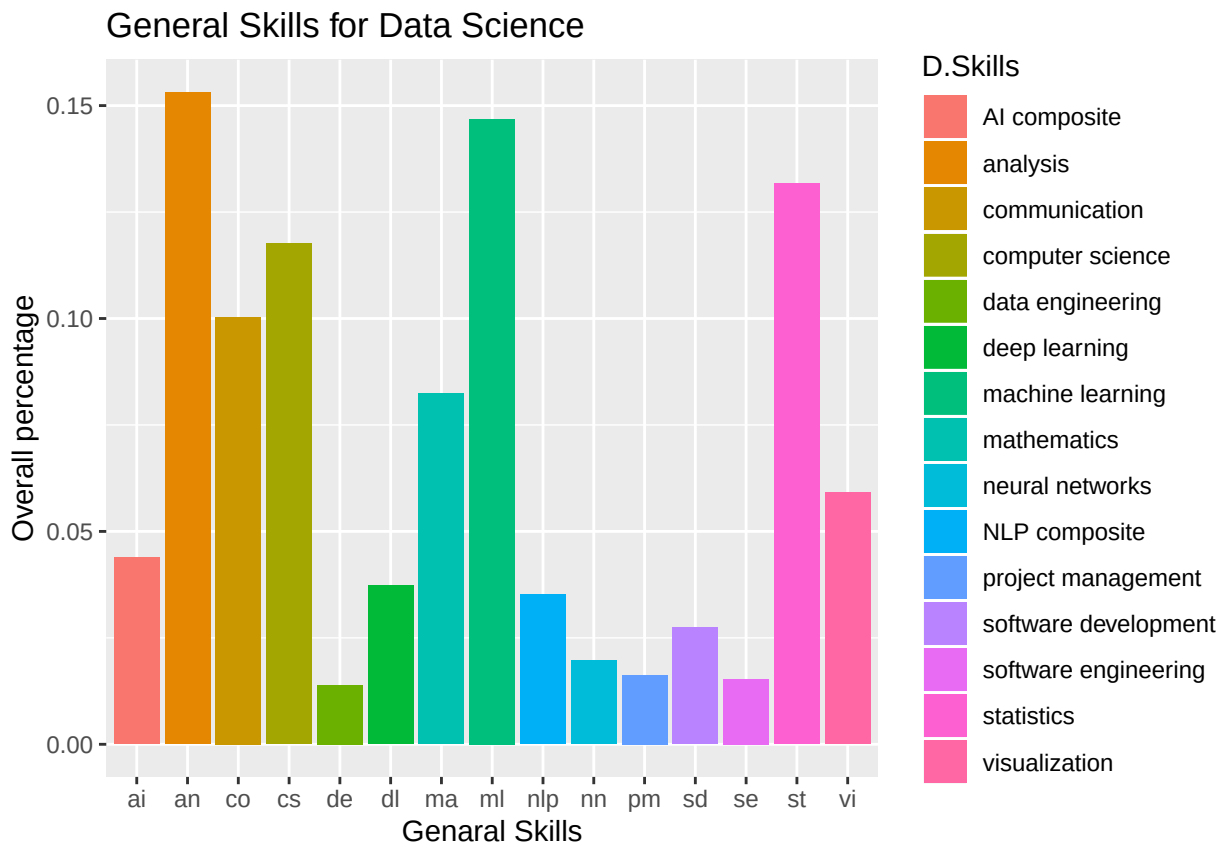
```
##                D.Skills        pct
## 1      machine learning 0.14683092
## 2              analysis 0.15311575
## 3            statistics 0.13167829
## 4      computer science 0.11763414
## 5         communication 0.10030640
## 6           mathematics 0.08238259
## 7         visualization 0.05910465
## 8          AI composite 0.04382653
## 9         deep learning 0.03733255
## 10        NLP composite 0.03517835
## 11 software development 0.02743995
## 12      neural networks 0.01968063
## 13     data engineering 0.01389775
## 14   project management 0.01621927
## 15 software engineering 0.01537223
```

Changing the values of variable D.skill

```
D.Skills.abv <-c("ml", "an", "st", "cs", "co", "ma", "vi","ai",
                     "dl","nlp", "sd", "nn", "de", "pm","se")
g.skild <- data.frame(g.skills,
                      ds.abv = D.Skills.abv,
                      D.Skills = g.skills$D.Skills,
                      pct=g.skills$pct)
```

Visualisation

```
ggplot(g.skild,
       aes(x=ds.abv, y = pct))+
  geom_col(aes(fill=D.Skills), position = "dodge")+
  xlab("Genaral Skills")+ylab("Overall percentage")+
  ggtitle("General Skills for Data Science")
```



Analysis, Machine Learning, Statistics, Computer Science and Communication are general skill required for Data Scientists.

Data Science Soft Skills

Downloading data

```
dss <- read.csv("https://raw.githubusercontent.com/AlainKuiete/DATA607/master/ds_job_listing_software.cs
```

Subsetting the Data Science soft Skills

```
dsk <- dss[1:30,1:5]
str(dsk)
```

```
## 'data.frame':    30 obs. of  5 variables:
##  $ Keyword    : Factor w/ 42 levels "","\"data scientist\" \"[keyword]\"",..: 31 33 39 37 15 18 34 40
##  $ LinkedIn   : Factor w/ 40 levels "","1,024","1,040",..: 33 27 17 9 8 7 6 5 4 3 ...
##  $ Indeed     : Factor w/ 37 levels "","1,012","1,134",..: 23 22 11 5 6 4 3 2 37 34 ...
##  $ SimplyHired: Factor w/ 39 levels "","1,059","1,164",..: 16 15 14 4 3 2 39 38 37 35 ...
##  $ Monster    : Factor w/ 40 levels "","1,002","1,062",..: 18 17 5 3 4 2 39 36 35 33 ...
```

reshaping my dataframe

```
colnames(dsk) <- c("D.Skills", "LinkedIn", "Indeed", "SimplyHired", "Monster")
dsk$LinkedIn <- as.numeric(gsub(",","",dsk$LinkedIn))
dsk$Indeed <- as.numeric(gsub(",","",dsk$Indeed))
dsk$SimplyHired <- as.numeric(gsub(",","",dsk$SimplyHired))
dsk$Monster <- as.numeric(gsub(",","",dsk$Monster))
```

Computation

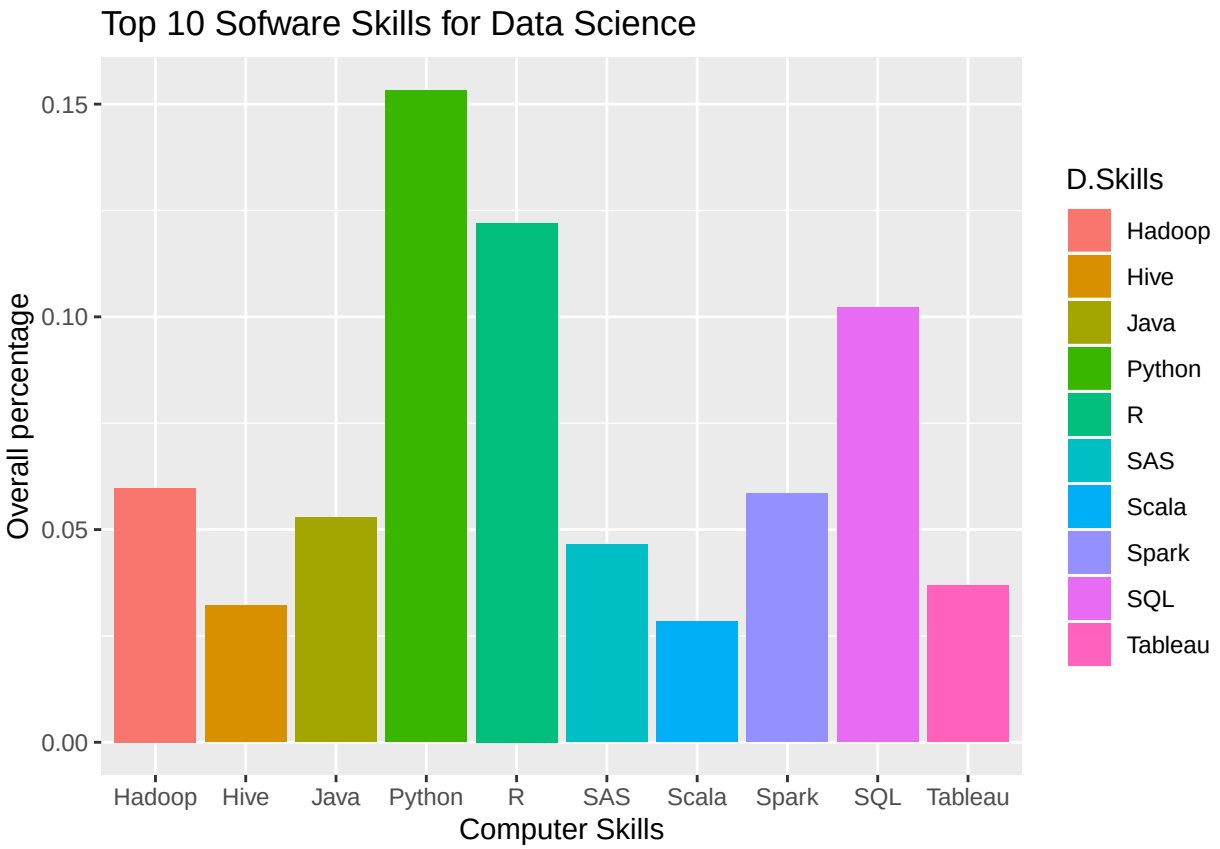
```
s.dsks <- summarise(dsk, sL=sum(LinkedIn), sI= sum(Indeed), sS=sum(SimplyHired),
           sM=sum(Monster))
ts <- sum(s.dsks)

skills <- dsk%>% mutate(D.Skills, pct=(LinkedIn+Indeed+SimplyHired+Monster)/ts)%>%select(D.Skills,pct)
skill <-skills[1:10,]
skill
```

```
##    D.Skills        pct
## 1    Python 0.15324530
## 2         R 0.12200083
## 3       SQL 0.10222248
## 4     Spark 0.05845075
## 5    Hadoop 0.05977716
## 6      Java 0.05287980
## 7       SAS 0.04652282
## 8   Tableau 0.03686455
## 9      Hive 0.03210910
## 10    Scala 0.02837548
```

Visualisation

```
ggplot(skill,
       aes(x=D.Skills, y = pct))+
  geom_col(aes(fill=D.Skills), position = "dodge")+
  xlab("Computer Skills")+ylab("Overall percentage")+
  ggtitle("Top 10 Sofware Skills for Data Science")
```



Python and R are the most software computer skills recommanded for Data Scientist.

Reference: The Most in Demand Skills for Data Scientists by Jeff Hale. Toward Data Science