Untitled

Read in data

```
library(readr)
rachel <- read_csv("C:/Users/Flora Huang/Desktop/Rachel Email/rachel.csv")
```

```
## Parsed with column specification:
## cols(
##   Company = col_character(),
##   Day = col_character(),
##   Date = col_character(),
##   Title = col_character(),
##   Apply = col_character(),
##   Location = col_character(),
##   Industry = col_character(),
##   Type = col_character(),
##   Description = col_character()
## )
```

Analyze day

```
rachel$Day <- factor(rachel$Day, levels= c("Monday",
    "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday","Sunday"))
```

```
library(highcharter)
```

```
## Highcharts (www.highcharts.com) is a Highsoft software product which is
```

```
## not free for commercial and Governmental use
```

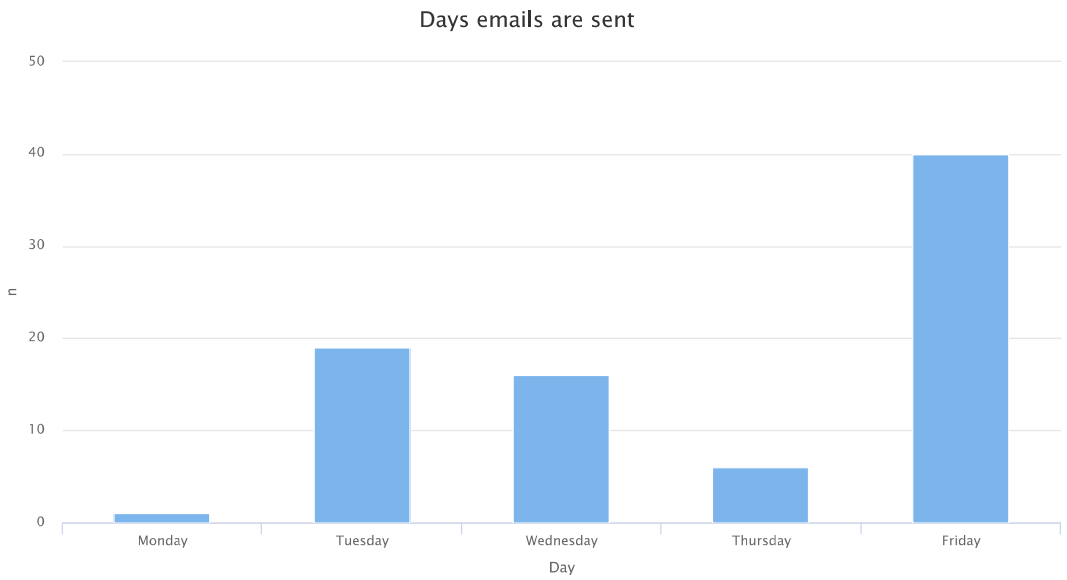
```
library(dplyr)
```

```
##
## 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
```

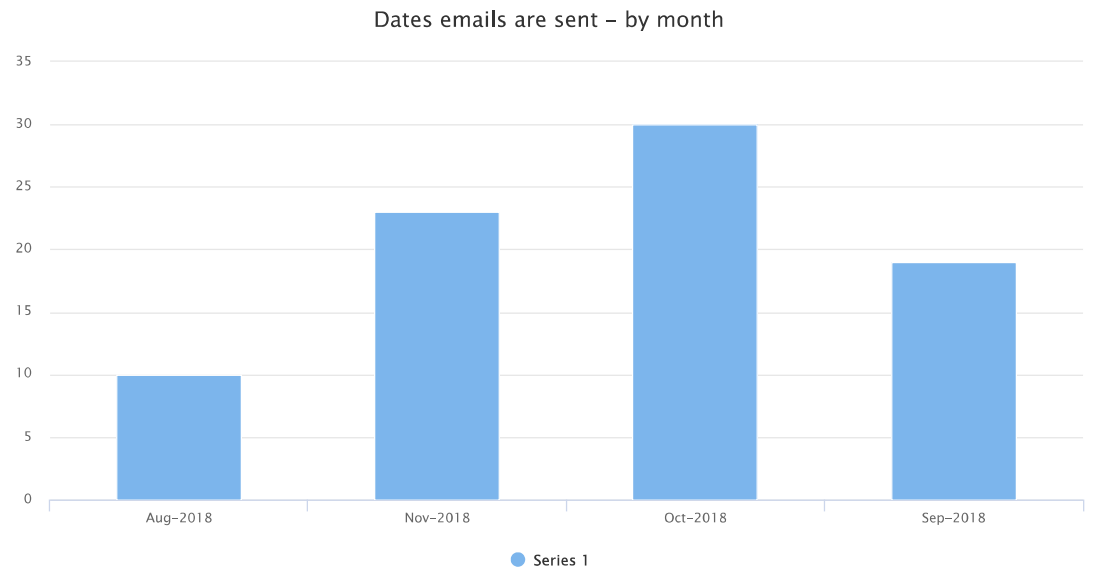
```
library(stringr)
rachel %>% count(Day) %>% hchart(type = "column", hcaes(x = Day, y = n)) %>% hc_title(text="Days emails are sent")
```



Analyze date

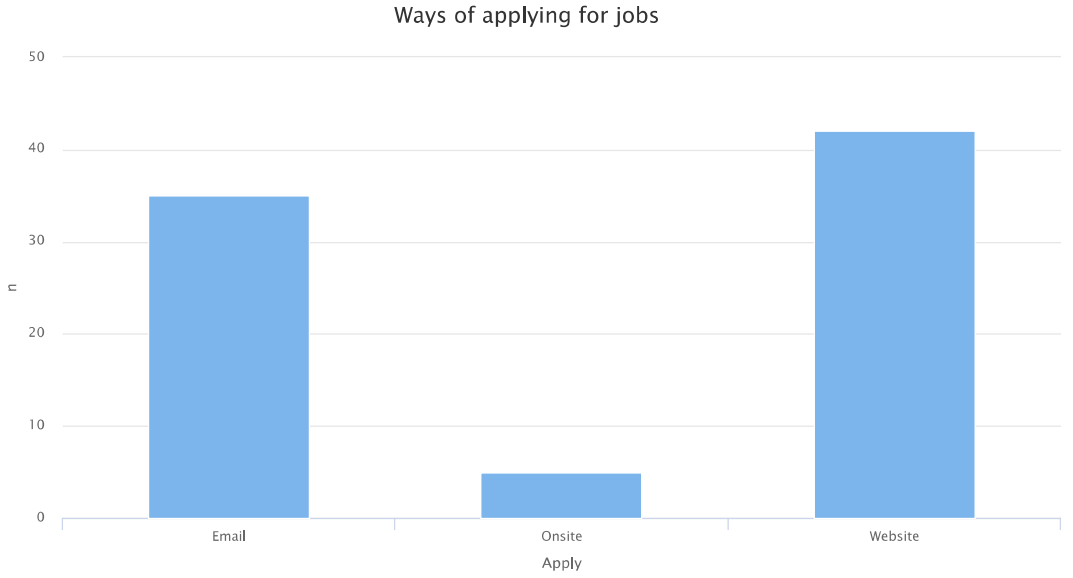
```
rachel$Date <- as.Date(rachel$Date, "%m/%d/%Y")
rachel %>% count(Date) %>% hchart(type = "column", hcaes(x = Date, y = n)) %>% hc_title(text="Dates emails are sent - by dat
e")
```

```
email_months <- as.data.frame(table(format(rachel$Date,"%b-%Y")))
highchart() %>% hc_xAxis(type = 'category') %>% hc_add_series(email_months, "column", hcaes(x = Var1, y = Freq)) %>% hc_titl
e(text="Dates emails are sent - by month")
```



Analyze ways to apply

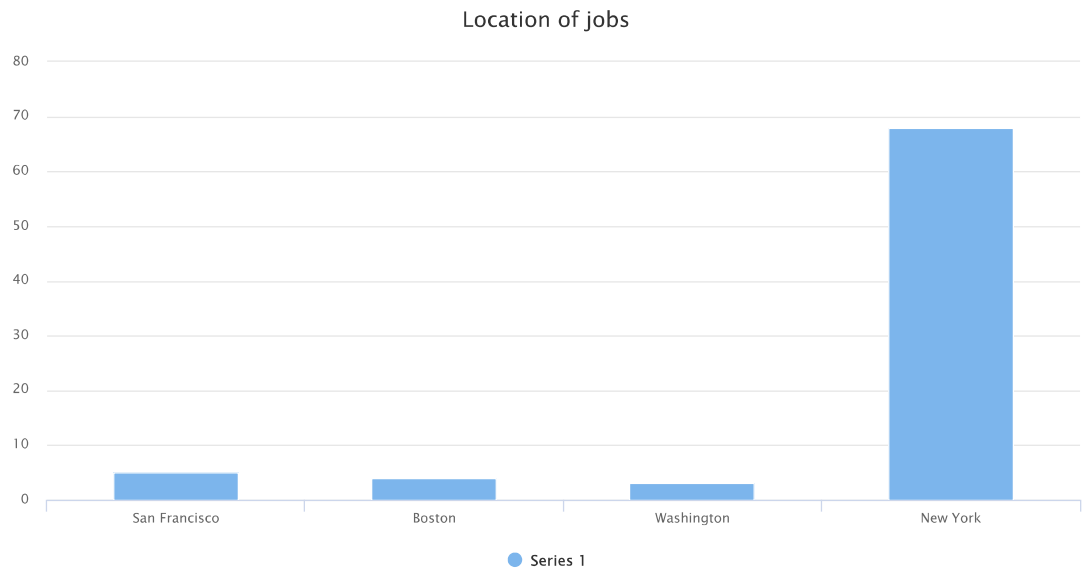
```
rachel$Apply <- factor(rachel$Apply)
rachel %>% count(Apply) %>% hchart(type = "column", hcaes(x = Apply, y = n)) %>% hc_title(text="Ways of applying for jobs")
```



Analyze locations

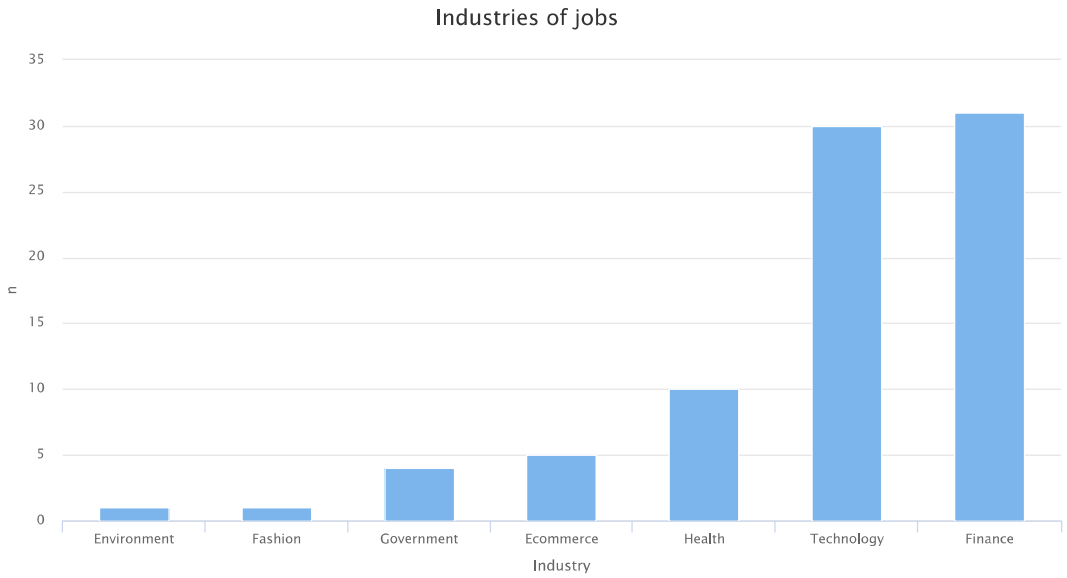
```
job_location <- data.frame("Location" = c("San Francisco", "Boston", "Washington", "New York"), "Jobs" = c(sum(str_count(rachel$Location, "San Francisco")), sum(str_count(rachel$Location, "Boston")), sum(str_count(rachel$Location, "Washington")), sum(str_count(rachel$Location, "New York"))))
```

```
highchart() %>% hc_xAxis(type = 'category') %>% hc_add_series(job_location, "column", hcaes(x = Location, y = Jobs)) %>% hc_title(text="Location of jobs")
```



Analyze industries

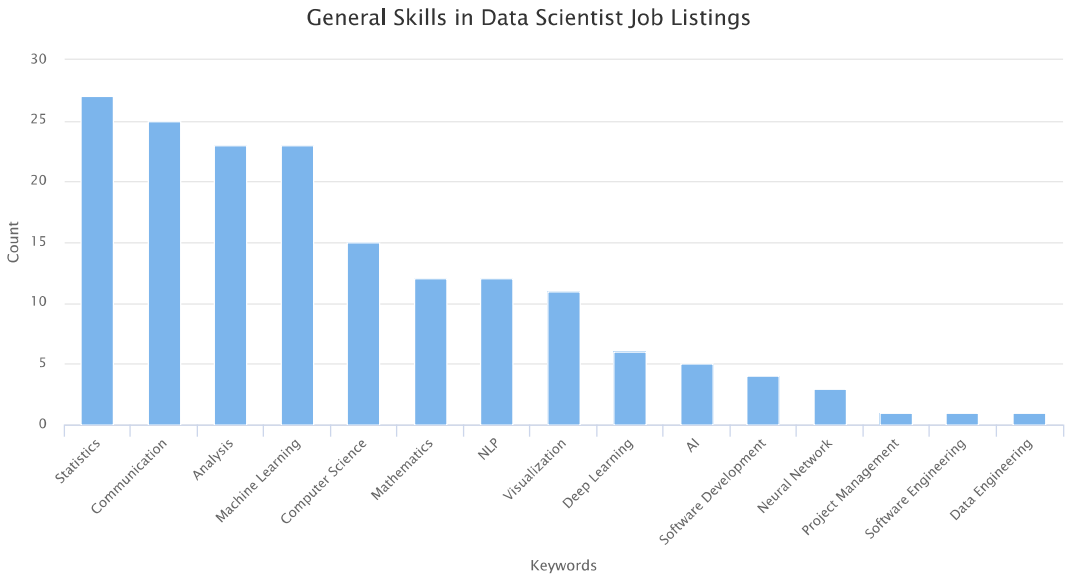
```
rachel %>% count(Industry) %>% arrange(n) %>% hchart(type = "column", hcaes(x = Industry, y = n)) %>% hc_title(text="Industries of jobs")
```



Analyze keywords

```
keywords <- data.frame("Keywords" = c("Analysis", "Machine Learning", "Statistics", "Computer Science", "Communication", "Mathematics", "Visualization", "AI", "Deep Learning", "NLP", "Software Development", "Neural Network", "Project Management", "Software Engineering", "Data Engineering"), "Count" = c(length(grep("analysis", tolower(rachel$Description))), length(grep("machine learning", tolower(rachel$Description))), length(grep("statistics", tolower(rachel$Description))), length(grep("computer science", tolower(rachel$Description))), length(grep("communication", tolower(rachel$Description))),  length(grep("mathematics", tolower(rachel$Description))), length(grep("visualization", tolower(rachel$Description))), length(grep("AI", rachel$Description))+length(grep("artificial intelligence", tolower(rachel$Description))), length(grep("deep learning", tolower(rachel$Description))), length(grep("NLP", rachel$Description))+length(grep("natural language processing", tolower(rachel$Description))), length(grep("software development", tolower(rachel$Description))), length(grep("neural network", tolower(rachel$Description))), length(grep("project management", tolower(rachel$Description))), length(grep("software engineering", tolower(rachel$Description))), length(grep("data engineering", tolower(rachel$Description)))))
```

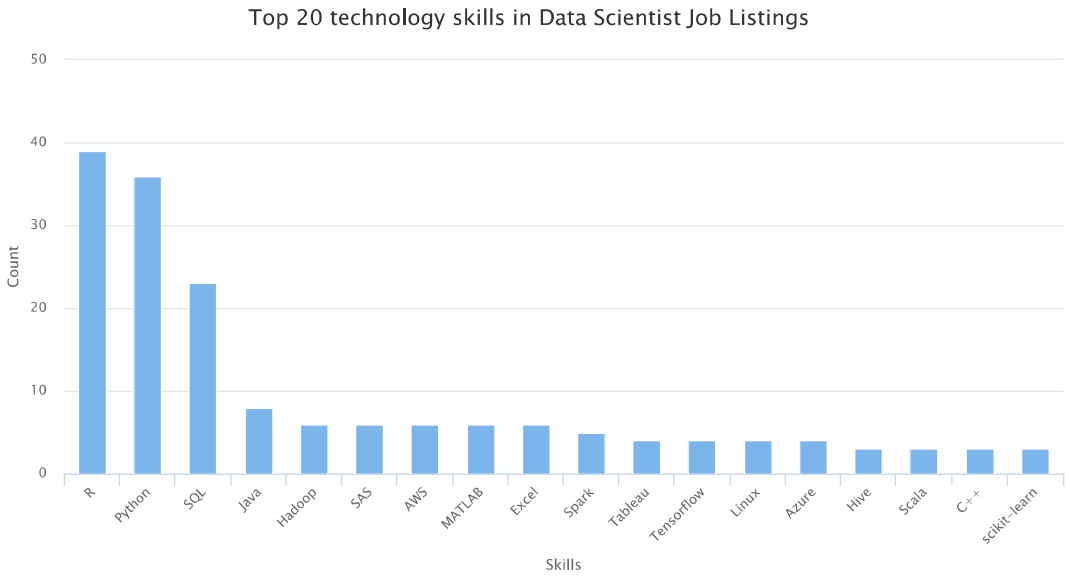
```
keywords %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = Keywords, y = Count)) %>% hc_xAxis(type = 'category') %>% hc_title(text="General Skills in Data Scientist Job Listings")
```



Technology skills

```
skills <- data.frame("Skills" = c("Python", "R", "SQL", "Hadoop", "Spark", "Java", "SAS", "Tableau", "Hive", "Scala", "AWS",
 "C++", "MATLAB", "Tensorflow", "Excel", "Linux", "Azure", "scikit-learn"), "Count" = c(length(grep("python", tolower(rachel
$Description))), length(grep("R", rachel$Description)), length(grep("SQL", rachel$Description)), length(grep("hadoop", tolow
er(rachel$Description))), length(grep("spark", tolower(rachel$Description))), length(grep("java", tolower(rachel$Descriptio
n))), length(grep("SAS", rachel$Description)), length(grep("tableau", tolower(rachel$Description))), length(grep("hive", tol
ower(rachel$Description))), length(grep("scala", tolower(rachel$Description))), length(grep("AWS", rachel$Description)), len
gth(grep("C\\+\\+", rachel$Description)), length(grep("matlab", tolower(rachel$Description))), length(grep("tensorflow", tol
ower(rachel$Description))), sum(str_count( tolower(rachel$Description), "\\bexcel\\b")), length(grep("linux", tolower(rachel
$Description))), sum(str_count( tolower(rachel$Description), "\\bazure\\b")), length(grep("scikit-learn", tolower(rachel$Des
cription))) ))
```

```
skills %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = Skills, y = Count)) %>% hc_xAxis(type = 'category') %
>% hc_title(text="Top 20 technology skills in Data Scientist Job Listings")
```



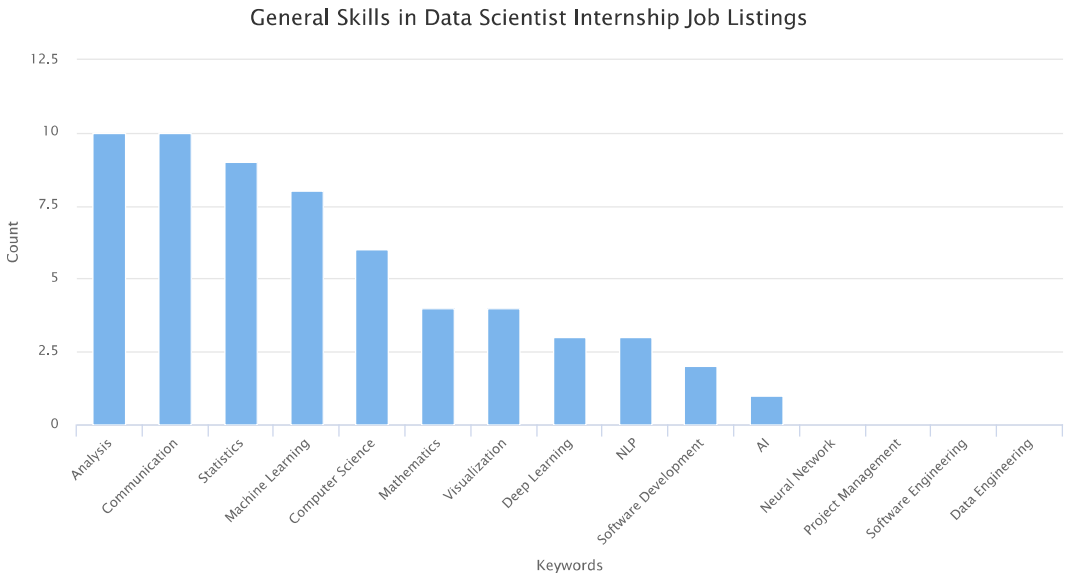
Analyze keywords - by job type

```
rachel_intern <- rachel[rachel$Type == "Internship",]
rachel_full <- rachel[rachel$Type == "Full Time",]
```

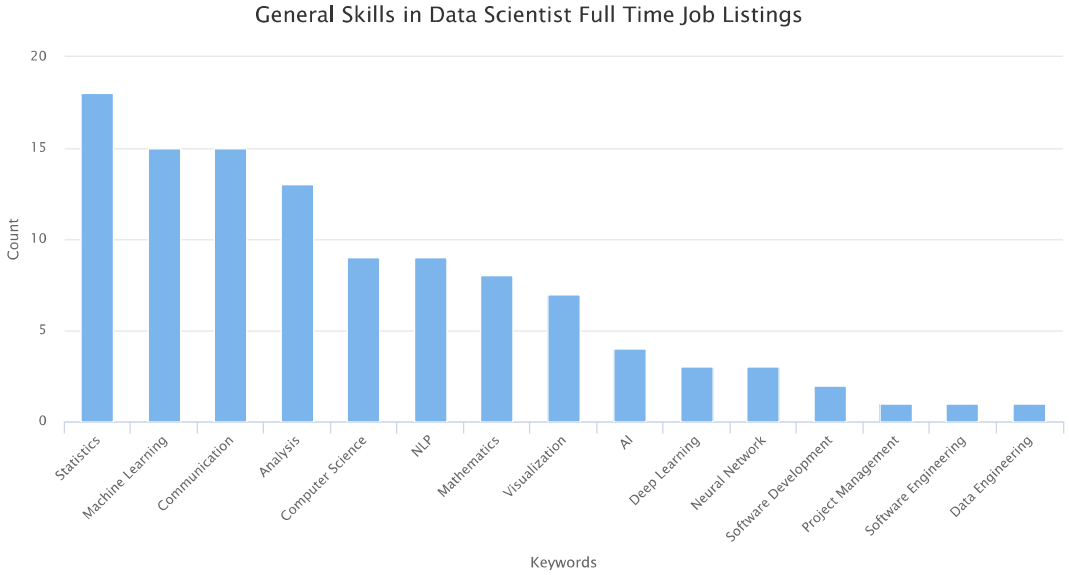
```
#intern
keywords_intern <- data.frame("Keywords" = c("Analysis", "Machine Learning", "Statistics", "Computer Science", "Communicatio
n", "Mathematics", "Visualization", "AI", "Deep Learning", "NLP", "Software Development", "Neural Network", "Project Managem
ent", "Software Engineering", "Data Engineering"), "Count" = c(length(grep("analysis", tolower(rachel_intern$Description))),
 length(grep("machine learning", tolower(rachel_intern$Description))), length(grep("statistics", tolower(rachel_intern$Descr
iption))), length(grep("computer science", tolower(rachel_intern$Description))), length(grep("communication", tolower(rachel
_intern$Description))),  length(grep("mathematics", tolower(rachel_intern$Description))), length(grep("visualization", tolow
er(rachel_intern$Description))), length(grep("AI", rachel_intern$Description))+length(grep("artificial intelligence", tolowe
r(rachel_intern$Description))), length(grep("deep learning", tolower(rachel_intern$Description))), length(grep("NLP", rachel
_intern$Description))+length(grep("natural language processing", tolower(rachel_intern$Description))), length(grep("software
 development", tolower(rachel_intern$Description))), length(grep("neural network", tolower(rachel_intern$Description))), len
gth(grep("project management", tolower(rachel_intern$Description))), length(grep("software engineering", tolower(rachel_inte
rn$Description))), length(grep("data engineering", tolower(rachel_intern$Description)))))
```

```
#full time
keywords_full <- data.frame("Keywords" = c("Analysis", "Machine Learning", "Statistics", "Computer Science", "Communication"
, "Mathematics", "Visualization", "AI", "Deep Learning", "NLP", "Software Development", "Neural Network", "Project Managemen
t", "Software Engineering", "Data Engineering"), "Count" = c(length(grep("analysis", tolower(rachel_full$Description))), len
gth(grep("machine learning", tolower(rachel_full$Description))), length(grep("statistics", tolower(rachel_full$Descriptio
n))), length(grep("computer science", tolower(rachel_full$Description))), length(grep("communication", tolower(rachel_full$D
escription))),  length(grep("mathematics", tolower(rachel_full$Description))), length(grep("visualization", tolower(rachel_f
ull$Description))), length(grep("AI", rachel_full$Description))+length(grep("artificial intelligence", tolower(rachel_full$D
escription))), length(grep("deep learning", tolower(rachel_full$Description))), length(grep("NLP", rachel_full$Description))
+length(grep("natural language processing", tolower(rachel_full$Description))), length(grep("software development", tolower
(rachel_full$Description))), length(grep("neural network", tolower(rachel_full$Description))), length(grep("project manageme
nt", tolower(rachel_full$Description))), length(grep("software engineering", tolower(rachel_full$Description))), length(grep
("data engineering", tolower(rachel_full$Description)))))
```

```
keywords_intern %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = Keywords, y = Count)) %>% hc_xAxis(type = 'ca
tegory') %>% hc_title(text="General Skills in Data Scientist Internship Job Listings")
```




```
keywords_full %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = Keywords, y = Count)) %>% hc_xAxis(type = 'cate
gory') %>% hc_title(text="General Skills in Data Scientist Full Time Job Listings")
```

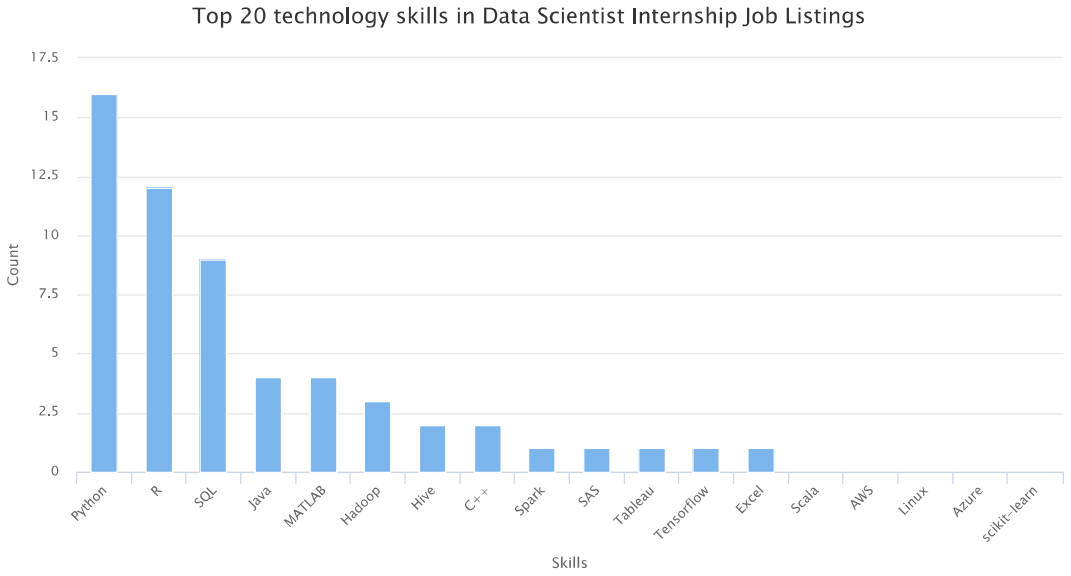


Analyze skills - by job type

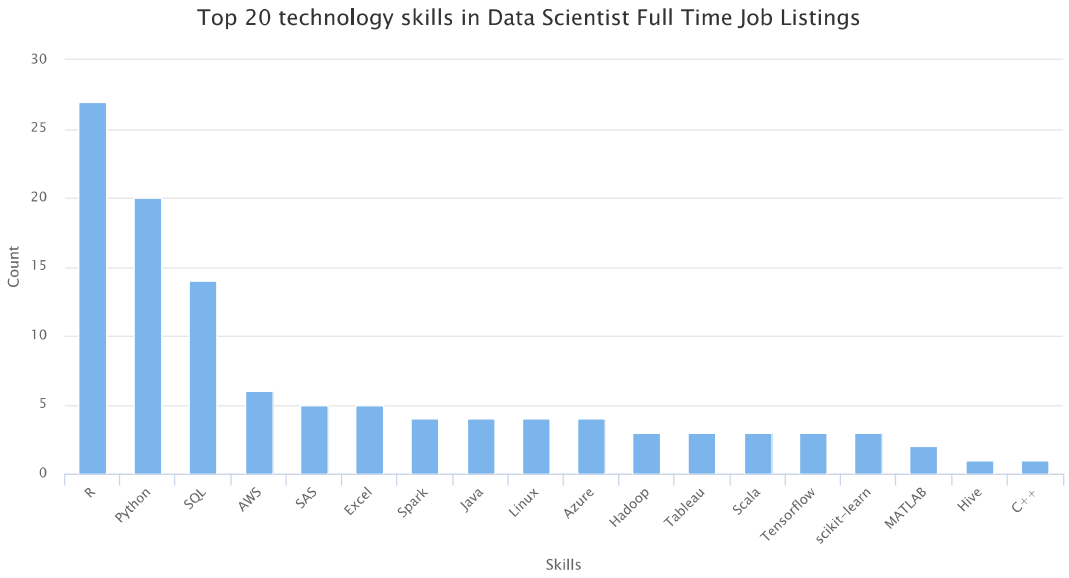
```
#intern
skills_intern <- data.frame("Skills" = c("Python", "R", "SQL", "Hadoop", "Spark", "Java", "SAS", "Tableau", "Hive", "Scala",
 "AWS", "C++", "MATLAB", "Tensorflow", "Excel", "Linux", "Azure", "scikit-learn"), "Count" = c(length(grep("python", tolower
(rachel_intern$Description))), length(grep("R", rachel_intern$Description)), length(grep("SQL", rachel_intern$Description)),
 length(grep("hadoop", tolower(rachel_intern$Description))), length(grep("spark", tolower(rachel_intern$Description))), leng
th(grep("java", tolower(rachel_intern$Description))), length(grep("SAS", rachel_intern$Description)), length(grep("tableau",
 tolower(rachel_intern$Description))), length(grep("hive", tolower(rachel_intern$Description))), length(grep("scala", tolowe
r(rachel_intern$Description))), length(grep("AWS", rachel_intern$Description)), length(grep("C\\+\\+", rachel_intern$Descrip
tion)), length(grep("matlab", tolower(rachel_intern$Description))), length(grep("tensorflow", tolower(rachel_intern$Descript
ion))), sum(str_count( tolower(rachel_intern$Description), "\\bexcel\\b")), length(grep("linux", tolower(rachel_intern$Descr
iption))), sum(str_count( tolower(rachel_intern$Description), "\\bazure\\b")), length(grep("scikit-learn", tolower(rachel_in
tern$Description))) ))
```

```
#full time
skills_full <- data.frame("Skills" = c("Python", "R", "SQL", "Hadoop", "Spark", "Java", "SAS", "Tableau", "Hive", "Scala",
"AWS", "C++", "MATLAB", "Tensorflow", "Excel", "Linux", "Azure", "scikit-learn"), "Count" = c(length(grep("python", tolower
(rachel_full$Description))), length(grep("R", rachel_full$Description)), length(grep("SQL", rachel_full$Description)), lengt
h(grep("hadoop", tolower(rachel_full$Description))), length(grep("spark", tolower(rachel_full$Description))), length(grep("j
ava", tolower(rachel_full$Description))), length(grep("SAS", rachel_full$Description)), length(grep("tableau", tolower(rache
l_full$Description))), length(grep("hive", tolower(rachel_full$Description))), length(grep("scala", tolower(rachel_full$Desc
ription))), length(grep("AWS", rachel_full$Description)), length(grep("C\\+\\+", rachel_full$Description)), length(grep("mat
lab", tolower(rachel_full$Description))), length(grep("tensorflow", tolower(rachel_full$Description))), sum(str_count( tolow
er(rachel_full$Description), "\\bexcel\\b")), length(grep("linux", tolower(rachel_full$Description))), sum(str_count( tolowe
r(rachel_full$Description), "\\bazure\\b")), length(grep("scikit-learn", tolower(rachel_full$Description))) ))
```

```
skills_intern %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = Skills, y = Count)) %>% hc_xAxis(type = 'catego
ry') %>% hc_title(text="Top 20 technology skills in Data Scientist Internship Job Listings")
```




```
skills_full %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = Skills, y = Count)) %>% hc_xAxis(type = 'categor
y') %>% hc_title(text="Top 20 technology skills in Data Scientist Full Time Job Listings")
```



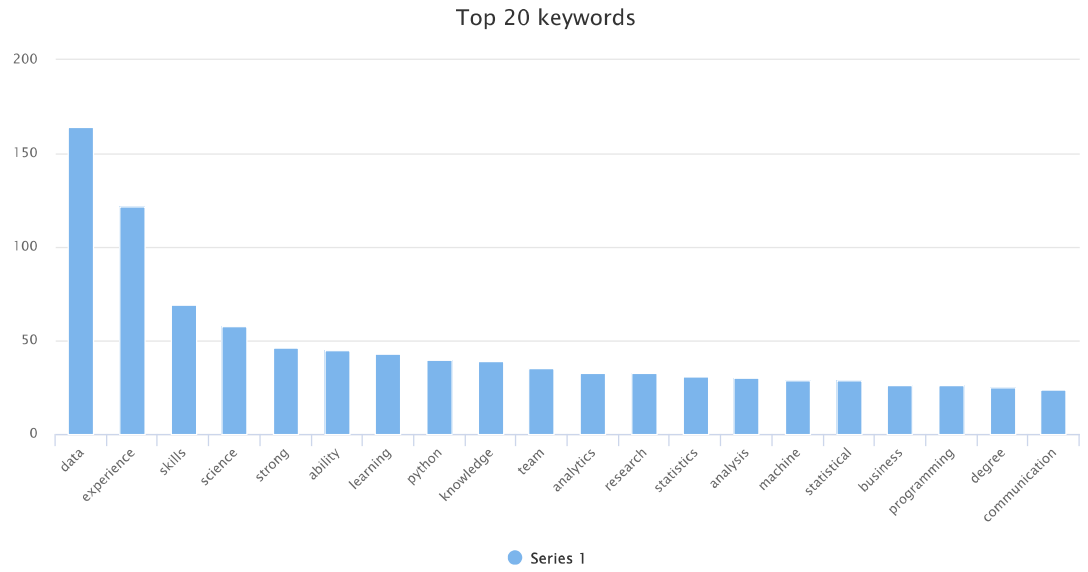
Text mining

```
library(tidytext)
text_df <- data_frame(line = 1:nrow(rachel), text = rachel$Description)
tidy_decp <- text_df %>% unnest_tokens(word, text) %>% anti_join(stop_words)
```

```
## Joining, by = "word"
```

Top 20 keywords

```
top20words <- tidy_decp %>% count(word, sort = TRUE) %>% head(20)
highchart() %>% hc_xAxis(type = 'category') %>% hc_add_series(top20words, "column", hcaes(x = word, y = n)) %>% hc_title(tex
t="Top 20 keywords")
```


Top 20 keywords
200
150
100
50
0
data
experience
skills
science
strong
ability
learning
python
knowledge
team
analytics
research
statistics
analysis
machine
statistical
business
programming
degree
communication
Series 1