Project 3: Best Skills for a Data Scientist

Folorunsho Atanda, Xhulia Turkaj, Ron Balaban

2023-10-22

Our project is guided by a pivotal question: *"What are the most crucial data science skills?"* To answer this question effectively, we have chosen a method that offers unique insights. We believe that examining the curricula of top online Master of Data Science programs at 23 universities is an ideal approach. These programs are meticulously designed to equip students with the skills and knowledge needed to excel in the field. By scrutinizing what these respected institutions prioritize, we gain a deep understanding of the skills that form the backbone of data science success.

We employed web scraping techniques to automate the extraction of data, capturing general information about the top 23 online Masters in Data Science programs. This data forms the foundation of the first table in our relational database.

From FORTUNE EDUCATION's website, we extracted the university's name, location, credit cost, and GRE requirements. While examining the URLs associated with the rankings, we discovered that some led to different programs within the same university. To address this, we manually retrieved the URLs for each program's overview website and then integrated them with the remaining information gathered through web scraping.

#————————————————————————- # libraries

```r
library(data.table)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::between()     masks data.table::between()
## x dplyr::filter()      masks stats::filter()
## x dplyr::first()       masks data.table::first()
## x lubridate::hour()    masks data.table::hour()
## x lubridate::isoweek() masks data.table::isoweek()
## x dplyr::lag()         masks stats::lag()
## x dplyr::last()        masks data.table::last()
## x lubridate::mday()    masks data.table::mday()
## x lubridate::minute()  masks data.table::minute()
## x lubridate::month()   masks data.table::month()
## x lubridate::quarter() masks data.table::quarter()
## x lubridate::second()  masks data.table::second()
## x purrr::transpose()   masks data.table::transpose()
## x lubridate::wday()    masks data.table::wday()
## x lubridate::week()    masks data.table::week()
## x lubridate::yday()    masks data.table::yday()
## x lubridate::year()    masks data.table::year()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(rvest) #HTML
```

```
##
## Attaching package: 'rvest'
##
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```
library(jsonlite) #JSON
```

```
##
## Attaching package: 'jsonlite'
##
## The following object is masked from 'package:purrr':
##
##     flatten
```

```
library(xml2) #XML
library(tm)
```

```
## Loading required package: NLP
##
## Attaching package: 'NLP'
##
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(knitr)
library(stringr)
library(RCurl)
```

```
##
## Attaching package: 'RCurl'
##
## The following object is masked from 'package:tidyr':
##
##     complete
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(RMySQL)
```

```
## Loading required package: DBI
```

#——————————————————————————- # Functions used

```
# function to scrap info from the website used
scrapped_data <- function(data, css_string, regex_string = ".*"){
  data %>%
    html_elements(css_string) %>%
    html_text() %>%
    str_extract(regex_string)
}

# Define a function to scrape skills from a program data entry
scrape_skills <- function(program_data_entry) {
webpage <- read_html(program_data_entry$url)
skills <- webpage %>%
html_nodes(program_data_entry$selector) %>%
html_text(trim = TRUE)
return(skills)
}
```

#——————————————————————————- # Scraping information on schools

```
url <- "https://fortune.com/education/information-technology/best-online-masters-in-data-science/"
html <- read_html(url)

school_names <- html %>%
  scrapped_data("fully-clickable , .card-row:nth-child(1) h2", "([a-zA-Z]+\ ?){1,}")

school_location <- html %>%
  scrapped_data(".card-row:nth-child(1) .text-medium", ".*")

acceptance_rate <- html %>%
  scrapped_data(".footer:nth-child(3) .col-xxs-6:nth-child(1) .value", ".*")

gre_required <- html %>%
  scrapped_data(".col-xxs-6:nth-child(4) .value", ".*")

cost_per_credit <- html %>%
  scrapped_data(".col-xxs-6:nth-child(3) .value", ".*")

avg_work_experience_required <- html %>%
  scrapped_data(".footer:nth-child(3) .col-xxs-6:nth-child(2) .value", ".*") %>%
  str_replace_all("DNP", "Did not state")

school_info <- tibble(
  school_names,
  school_location,
  acceptance_rate,
  gre_required,
  cost_per_credit,
  avg_work_experience_required,
  )

school_info
```

```
## # A tibble: 23 x 6
##    school_names     school_location acceptance_rate gre_required cost_per_credit
##    <chr>            <chr>           <chr>           <chr>        <chr>
##  1 University of S~ Los Angeles, CA 13.00%          No           $2,309.00
##  2 University of C~ Berkeley, CA    36.00%          No           $2,780.00
##  3 Bay Path Univer~ Longmeadow, MA  49.00%          No           $880.00
##  4 New Jersey Inst~ Newark, NJ      70.00%          Flexible     $1,112.00
##  5 Clemson Univers~ Clemson, SC     86.00%          Yes          $1,264.00
##  6 Illinois Instit~ Chicago, IL     66.00%          No           $1,646.00
##  7 Oklahoma State ~ Stillwater, OK  98.00%          Flexible     $611.35
##  8 Texas Tech Univ~ Lubbock, TX     81.00%          No           $1,097.00
##  9 University of M~ Columbia, MO    70.00%          No           $1,212.30
## 10 University of C~ Los Angeles, CA 74.00%          Yes          $1,050.00
## # i 13 more rows
## # i 1 more variable: avg_work_experience_required <chr>
```

#——————————————————————————-

Scraping data on skills you gain by the programs by creating a function that scrapes the text data from each URL and given CSS selector

Here we created a list called program data where each element is a list containing a URL and selector

```
program_data <- list(
  list(
    name = "Program 1",
    url = "https://online.usc.edu/programs/master-of-science-in-applied-data-science/",
    selector = "p:nth-child(2)"
  ),
  list(
    name = "Program 2",
    url = "https://ischoolonline.berkeley.edu/cybersecurity/curriculum/data-science-certificate/",
    selector = "p.u--margin-bottom-2:nth-child(2)"
  ),
  list(
    name = "Program 3",
    url = "https://www.baypath.edu/academics/graduate-programs/applied-data-science-ms/",
    selector = "li:nth-child(2)"
  ),
  list(
    name = "Program 4",
    url = "https://catalog.njit.edu/graduate/computing-sciences/computer-science/ms/index.html",
    selector = "td:nth-child(2)"
  ),
  list(
    name = "Program 5",
    url = "https://www.clemson.edu/graduate/academics/ms-dsa/academics.html",
    selector = c("td:nth-child(4) , td:nth-child(3), tr:nth-child(2) td:nth-child(2)")
  ),
  list(
    name = "Program 6",
    url = "https://bulletin.iit.edu/graduate/colleges/computing/applied-mathematics/master-data-science,
    selector = "td:nth-child(2)"
  ),
  list(
    name = "Program 7",
    url = "https://osuonline.okstate.edu/programs/graduate/business-analytics-master-of-science.html",
    selector = ".intro-with-media__caption"
  ),
  list(
    name = "Program 8",
    url = "https://www.depts.ttu.edu/rawlsbusiness/graduate/ms/datascience/",
    selector = ".curr-info"
  ),
  list(
    name = "Program 9",
    url = "https://dsa.missouri.edu/masters-program/",
    selector = ".shortcode_pb"
  ),
  list(
    name = "Program 10",
    url = "https://www.msol.ucla.edu/data-science-engineering/",
    selector = c(".order-lg-1", ".mb-0")
  ),
  list(
    name = "Program 11",
    url = "https://www.cdm.depaul.edu/academics/Pages/MS-in-Data-Science.aspx",
    selector = c(".tabbody p , .splashIntro")
  ),
  list(
    name = "Program 12",
    url = "https://engineeringonline.ucr.edu/data-science/data-science-curriculum/",
    selector = c("#content__wrap--1 .col-sm-12 , .col-sm-6 div")
  ),
  list(
    name = "Program 13",
    url = "https://onlinestemprograms.wpi.edu/programs/online-master-science-data-science-v3",
    selector = c("#block-pageblock-pb12694 p , #block-pageblock-pb12993 p, li, #block-pageblock-pb12989
  ),
  list(
    name = "Program 14",
    url = "https://lewisu.smartcatalogiq.com/en/graduate-2023-2024/graduate-catalog/new-college-of-avia
    selector = ".programTables"
```

```r
  ),
  list(
    name = "Program 15",
    url = "https://cs.illinois.edu/academics/graduate/professional-mcs/online-master-computer-science-da
    selector = c("#tile5326 , #tile5324, #tile5323")
  ),
  list(
    name = "Program 16",
    url = "https://sps.cuny.edu/academics/graduate/master-science-data-science-ms",
    selector = c(".row , .degree-detail")
  ),
  list(
    name = "Program 17",
    url = "https://umdearborn.edu/cecs/departments/computer-and-information-science/graduate-programs/ms
    selector = c(".accordion-content , .article div :nth-child(3), .text")
  ),
  list(
    name = "Program 18",
    url = "https://www.regis.edu/_documents/academics/fact-sheets/graduate/ms-data-science.pdf",
    selector = c("#5-threecolumntextcallout , .columns--align-left, #1-contentwithsidebar")
  ),
  list(
    name = "Program 19",
    url = "https://csweb.rice.edu/academics/graduate-programs/online-mcs/curriculum#corerequiredcourses
    selector = c(".mb-2-3 , .discover_small, .styled, .discover, .item-mcs p, .tac:nth-child(2) .grid-mw
  ),
  list(
    name = "Program 20",
    url = "https://www.eastern.edu/academics/graduate-programs/ms-data-science",
    selector = c(".txt , #content :nth-child(3), .field_p_intro_text span")
  ),
  list(
    name = "Program 21",
    url = "https://onlinegrad.syracuse.edu/information-science/academics/#applied-data-science",
    selector = c("#section-6012c09d-d581-42b5-ac54-caa876af686c .section--mobileText , #section-6012c09
  ),
  list(
    name = "Program 22",
    url = "https://stevens.smartcatalogiq.com/en/2023-2024/academic-catalog/department-of-mathematical-
    selector = "#rightpanel"
  ),
  list(
    name = "Program 23",
    url = "https://online.pace.edu/graduate-programs/ms-in-data-science/curriculum/",
    selector = c(".mr-lg-5 p:nth-child(2) , p:nth-child(1), .collapsed")
  )
)

# Scrape skills for each program in program_data
scraped_skills <- lapply(program_data, scrape_skills)

# The raw text for the programs
text_for_program_1 <- paste(scraped_skills[[1]], collapse = " ")
```

```
text_for_program_2 <- paste(scraped_skills[[2]], collapse = " ")
text_for_program_3 <- paste(scraped_skills[[3]], collapse = " ")
text_for_program_4 <- paste(scraped_skills[[4]], collapse = " ")
text_for_program_5 <- paste(scraped_skills[[5]], collapse = " ")
text_for_program_6 <- paste(scraped_skills[[6]], collapse = " ")
text_for_program_7 <- paste(scraped_skills[[7]], collapse = " ")
text_for_program_8 <- paste(scraped_skills[[8]], collapse = " ")
text_for_program_9 <- paste(scraped_skills[[9]], collapse = " ")
text_for_program_10 <-paste(scraped_skills[[10]], collapse = " ")
text_for_program_11 <- paste(scraped_skills[[11]], collapse = " ")
text_for_program_12 <- paste(scraped_skills[[12]], collapse = " ")
text_for_program_13 <- paste(scraped_skills[[13]], collapse = " ")
text_for_program_14 <- paste(scraped_skills[[14]], collapse = " ")
text_for_program_15 <- paste(scraped_skills[[15]], collapse = " ")
text_for_program_16 <- paste(scraped_skills[[16]], collapse = " ")
text_for_program_17 <- paste(scraped_skills[[17]], collapse = " ")
text_for_program_18 <- paste(scraped_skills[[18]], collapse = " ")
text_for_program_19 <- paste(scraped_skills[[19]], collapse = " ")
text_for_program_20 <-paste(scraped_skills[[20]], collapse = " ")
text_for_program_21 <- paste(scraped_skills[[21]], collapse = " ")
text_for_program_22 <- paste(scraped_skills[[22]], collapse = " ")
text_for_program_23 <- paste(scraped_skills[[23]], collapse = " ")


#Create one object that stores the combined text
combined_text <- paste(text_for_program_1,
                       text_for_program_2,
                       text_for_program_3,
                       text_for_program_4,
                       text_for_program_5,
                       text_for_program_6,
                       text_for_program_7,
                       text_for_program_8,
                       text_for_program_9,
                       text_for_program_10,
                       text_for_program_11,
                       text_for_program_12,
                       text_for_program_13,
                       text_for_program_14,
                       text_for_program_15,
                       text_for_program_16,
                       text_for_program_17,
                       text_for_program_18,
                       text_for_program_19,
                       text_for_program_20,
                       text_for_program_21,
                       text_for_program_22,
                       text_for_program_23,
                       collapse = " ")
```

#————————————————————————-

Using DataCamp's recommended data scientist skills we came up with a list of what we believe to be the most important skills needed by a data scientist

```
skill_set_url <- "https://www.datacamp.com/blog/top-15-data-scientist-skills"
skill_set_html <- read_html(skill_set_url)
skill_set <- skill_set_html %>%
  scrapped_data(".css-1j302im-RichText h3", ".*")
skill_set <- gsub("Skills", "", skill_set)
skill_set <- gsub("Statistics and Math", "Statistics", skill_set)
skill_set <- c(skill_set, "Math")

skill_set
```

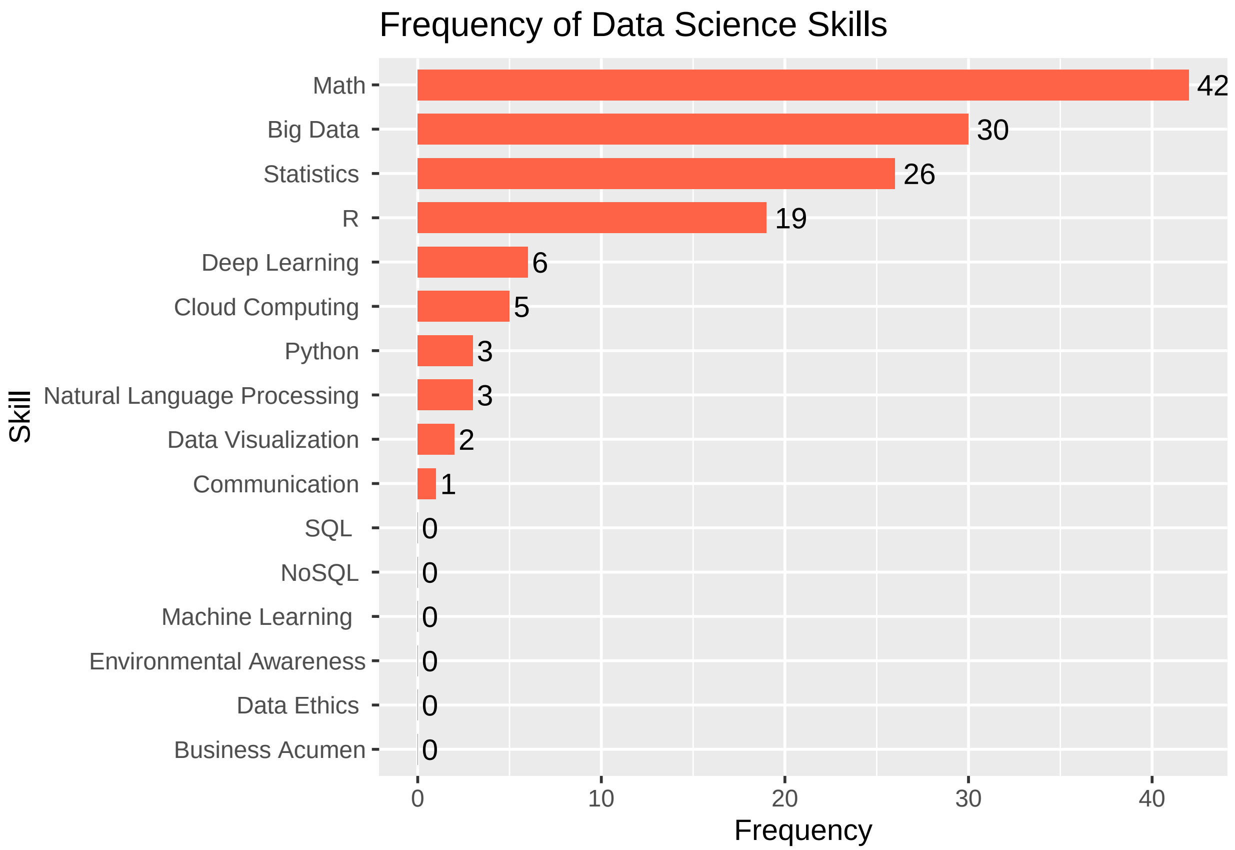
```
##  [1] "Python "                      "R "
##  [3] "Statistics "                  "SQL  "
##  [5] "NoSQL "                       "Data Visualization "
##  [7] "Machine Learning  "           "Deep Learning "
##  [9] "Natural Language Processing " "Big Data "
## [11] "Cloud Computing "             "Business Acumen"
## [13] "Communication "               "Data Ethics "
## [15] "Environmental Awareness"      "Math"
```

#——————————————————————————-

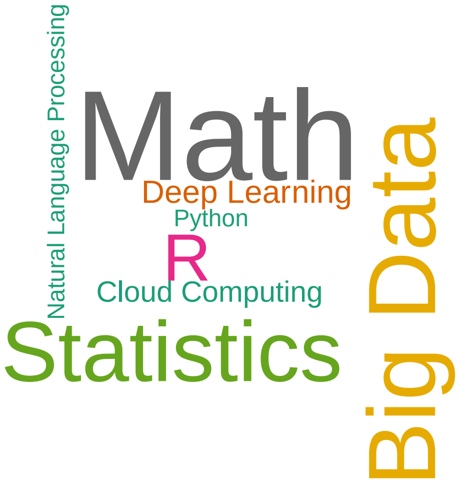
We compare the recommended data scientist skills with the skills taught by the 23 programs we picked

Graph to show the frequency of desired data science skills from the various programs

```
element_count <- str_extract_all(combined_text, skill_set) %>%
lapply(function(x) length(x))
element_count_table <- as_tibble(do.call(rbind, lapply(element_count, function(x) c(Count = x))))

skill_set_table <- as_tibble(do.call(rbind, lapply(skill_set, function(x) c(Key_Skills = x))))

frequency_table <- tibble(skill_set_table, element_count_table)

# Remove those with no appearances
cleaned_frequency_table2 <- frequency_table #%>%
  #filter(Count > 0)

cleaned_frequency_table2 %>%
  ggplot(aes(x = reorder(Key_Skills, +Count), y = Count)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7, fill = 'tomato') +
  coord_flip() +
  xlab('Skill') + ylab('Frequency') + labs(title = "Frequency of Data Science Skills") + theme(legend.p
  geom_text(aes(label = Count), hjust = -0.25, color = "black")
```



```
wordcloud(cleaned_frequency_table2$Key_Skills[1:16],
          as.numeric(cleaned_frequency_table2$Count[1:16]),
          colors = brewer.pal(8, "Dark2"))
```



Conclusion:

The graphical representation reveals that universities predominantly emphasize programming skills, with the exception of SQL (for reasons that are not immediately clear). It's important to note that the data we extracted was from the information pages of the 23 selected master's programs. Further investigation might uncover that these schools do indeed offer a comprehensive range of technical skills, including SQL.

A noteworthy observation is that the information pages primarily highlight technical skills. In our view, it is advisable for schools to also place emphasis on soft skills and explicitly mention their commitment to developing these skills on their information pages.