DATA607_Project3_EDA_V2

Christian Uriostegui

2022-10-21

Load Library

```
library(DBI)
library(RMariaDB)
library(wordcloud)
```

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

```
library(RColorBrewer)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.6      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
user <- 'guest'
pw <- 'guestpass'
hostname <- 'cunyspsds.c5iiratvieki.us-east-1.rds.amazonaws.com'
```

```
projectDb <- dbConnect(MariaDB(), user='guest', password=pw, dbname='Project3', host=hostname)
```

```
dbListTables(projectDb)
```

```
## [1] "EDUCATION"           "EDUCATION_IN_DEMAND" "SKILL"
## [4] "SKILL_IN_DEMAND"     "SOURCE"
```

Import Data

```
# qry import skill_in_demand table
qry <- "SELECT * FROM SKILL_IN_DEMAND;"

# store the results as a dataframe
rs <- dbSendQuery(projectDb, qry)

skills <- dbFetch(rs)

dbClearResult(rs) # clear the result
```

```
# query1: import education_in_demand table
query1 <- "SELECT * FROM EDUCATION_IN_DEMAND;"

# store the results as a dataframe
results1 <- dbSendQuery(projectDb,query1)

education <- dbFetch(results1)

dbClearResult(results1) # clear the result
```

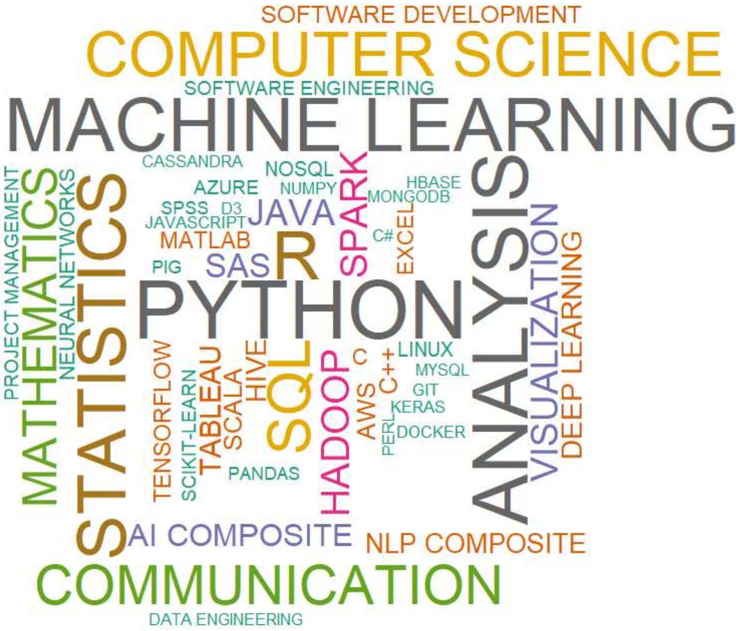
EXPLORATORY DATA ANALYSIS

Summary skill counts

```
# Summary skill counts
skills_summary <- skills %>%
                    group_by(SKILL_KEYWORD) %>%
                    summarise(TOTAL = sum(COUNT))
```

Wordcloud

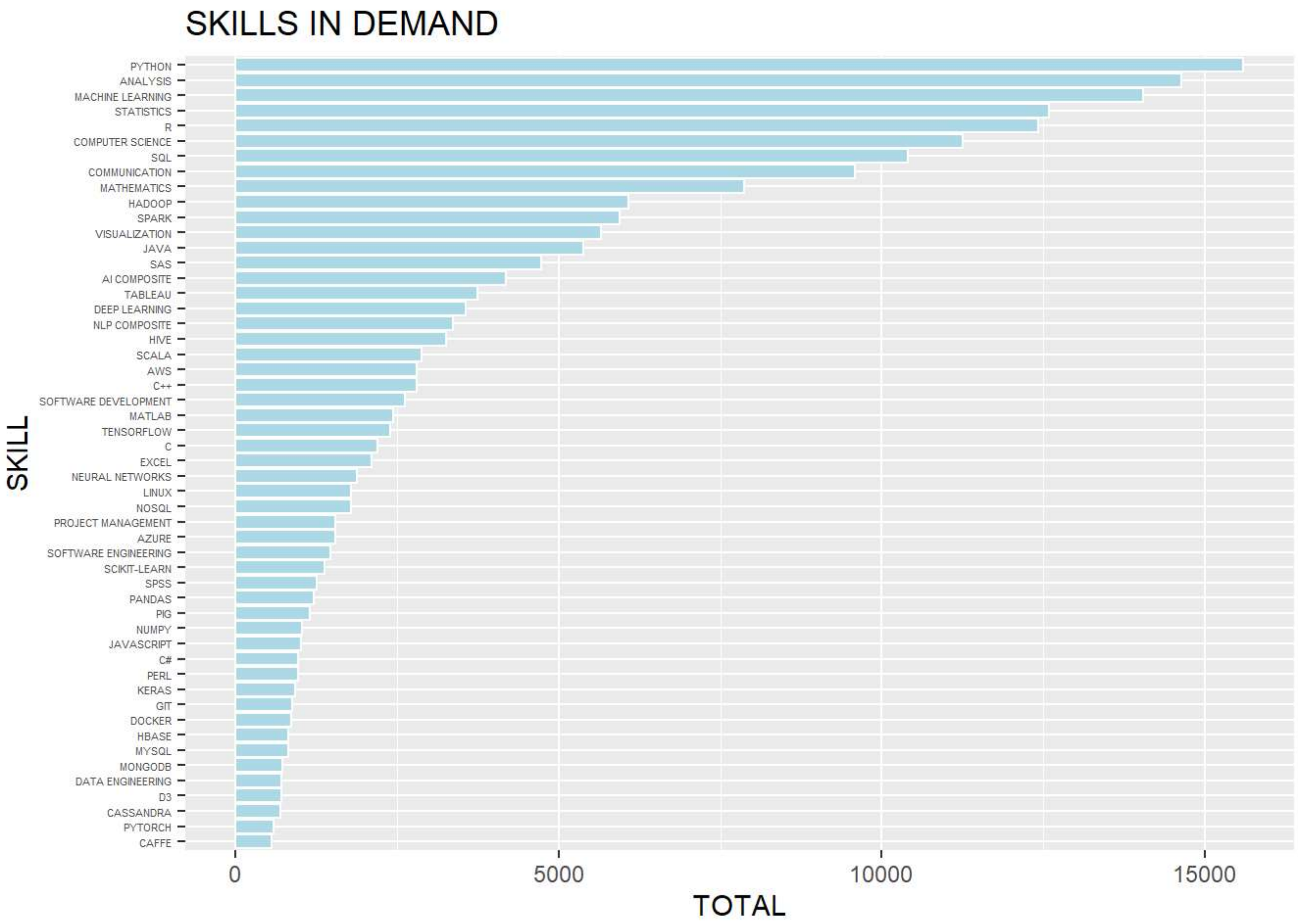
```
set.seed(1234)
wordcloud::wordcloud(words = skills_summary$SKILL_KEYWORD,
                     freq = skills_summary$TOTAL,
                     min.freq = 100,
                     max.words = 50,
                     random.order = FALSE,
                     random.color = FALSE,
                     rot.per = 0.25,
                     colors = brewer.pal(8, "Dark2"),
                     scale = c(2.5, 0.40))
```



Skills count by keyword graphic

```
skills_count <- skills |>
  group_by(SKILL = SKILL_KEYWORD) |>
  summarize(TOTAL=sum(COUNT)) |>
  arrange(desc(TOTAL))
```

```
ggplot(skills_count, aes(x=reorder(SKILL, TOTAL),
                         y=TOTAL)) +
  geom_col(fill="lightblue", color="white") +
  coord_flip() +
  theme(axis.text.y = element_text(size = 4)) +
  labels(x = "SKILL", title="SKILLS IN DEMAND")
```



Skills count by keyword

```
skills |>
  group_by(SKILL = SKILL_KEYWORD) |>
  summarize(TOTAL=sum(COUNT)) |>
  arrange(desc(TOTAL))
```

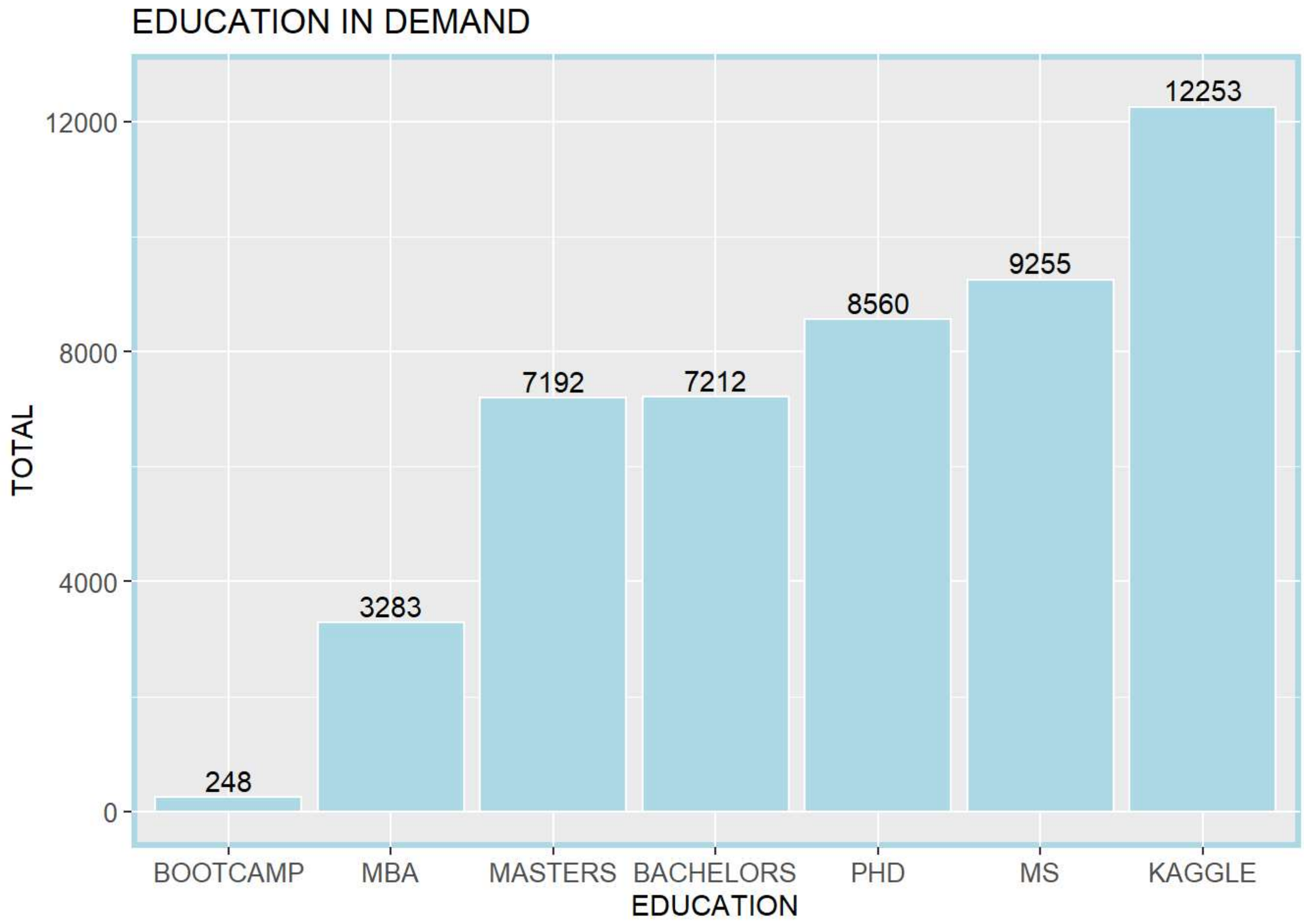
```
## # A tibble: 52 x 2
##    SKILL            TOTAL
##    <chr>            <int>
##  1 PYTHON           15597
##  2 ANALYSIS         14642
##  3 MACHINE LEARNING 14041
##  4 STATISTICS       12592
##  5 R                12417
##  6 COMPUTER SCIENCE 11249
##  7 SQL              10404
##  8 COMMUNICATION     9592
##  9 MATHEMATICS       7878
## 10 HADOOP            6084
## # ... with 42 more rows
```

Education count by keyword

```
education_count <- education |>
  group_by(EDUCATION = EDUCATION_KEYWORD) |>
  summarize(TOTAL=sum(COUNT)) |>
  arrange(desc(TOTAL))
```

Plot of Degrees of Education Count

```
ggplot(education_count, aes(x=reorder(EDUCATION, TOTAL),
                            y=TOTAL)) +
  geom_col(fill="lightblue", color="white") +
  geom_text(aes(label = signif(TOTAL)), nudge_y = 300) +
  theme(axis.text = element_text(size = 10)) +
  theme(panel.background=element_rect(size=2,colour="lightblue")) +
  labs(x = "EDUCATION", title="EDUCATION IN DEMAND")
```


EDUCATION IN DEMAND
TOTAL
BOOTCAMP
MBA
MASTERS
BACHELORS
PHD
MS
KAGGLE
EDUCATION