Data 607 Project 3

Tyler Baker, Noah Collin, Shane Hylton

10/17/2021

Project: What are the most important data science skills?

Goals

In this project, we seek to answer this question by collecting all of the words among a wide number of job descriptions. We will take all of the words and create word cloud plots to show which words occur most frequently in job descriptions. This should provide us with sufficient information to determine the most highly sought after data science skills.

Loading the Data

We used a relational database in Postgres to help determine which columns were consistent across the datasets. The data was downloaded from Kaggle and it was uploaded to github for easy downloading into R. Links to the Kaggle pages can be found at the bottom of the RMarkdown document. The datasets consist of wide dataframes with column information on various job listings for data analysts and data scientists.

```
linka <- 'https://raw.githubusercontent.com/st3vejobs/607Project3/main/DataAnalyst.csv'
DataAnaylstsRaw <- read.csv(url(linka), na.strings = "")

linkb <- 'https://raw.githubusercontent.com/st3vejobs/607Project3/main/DataScientist.csv'
DataScientistRaw <- read.csv(url(linkb), na.strings = "")
```

```
allData <- read_csv("alldata.csv", show_col_types = FALSE)

#allData was not used for analysis into keywords because there were too few columns that could relate b
```

Combining the Data

Here, we combined and analyzed the two datasets. We performed character manipulation and applied tidying techniques to collect data for minimum and maximum salary brackets.

```
Analysts <- subset(DataAnaylstsRaw, select =  (c("Job.Title", "Salary.Estimate","Job.Description")) )
Scientists <- subset(DataScientistRaw, select = c("Job.Title","Salary.Estimate","Job.Description"))
#colnames(Scientists) <- (c("Job.Title", "Salary.Estimate","Job.Description"))
combinedData <- rbind(Analysts, Scientists)
```

```
Sals <- (combinedData$Salary.Estimate)
minsals <- str_extract(Sals,"\\d\\d+")
minsals <-  as.numeric(minsals)

maxSals <- str_extract(Sals, "\\-\\$(\\d\\d+)")
maxSals <- str_extract(maxSals, "\\d\\d+")
maxSals <- as.numeric(maxSals)

combinedData$MinSalaries <-  minsals
combinedData$MaxSalaries <-  maxSals
combinedData <- drop_na(combinedData)


combinedData$MeanSalary <- rowMeans(combinedData[,c("MinSalaries", "MaxSalaries")], na.rm = TRUE)

#See ranges of mean Salaries
range(combinedData$MeanSalary)
```

```
## [1]  18 225
```

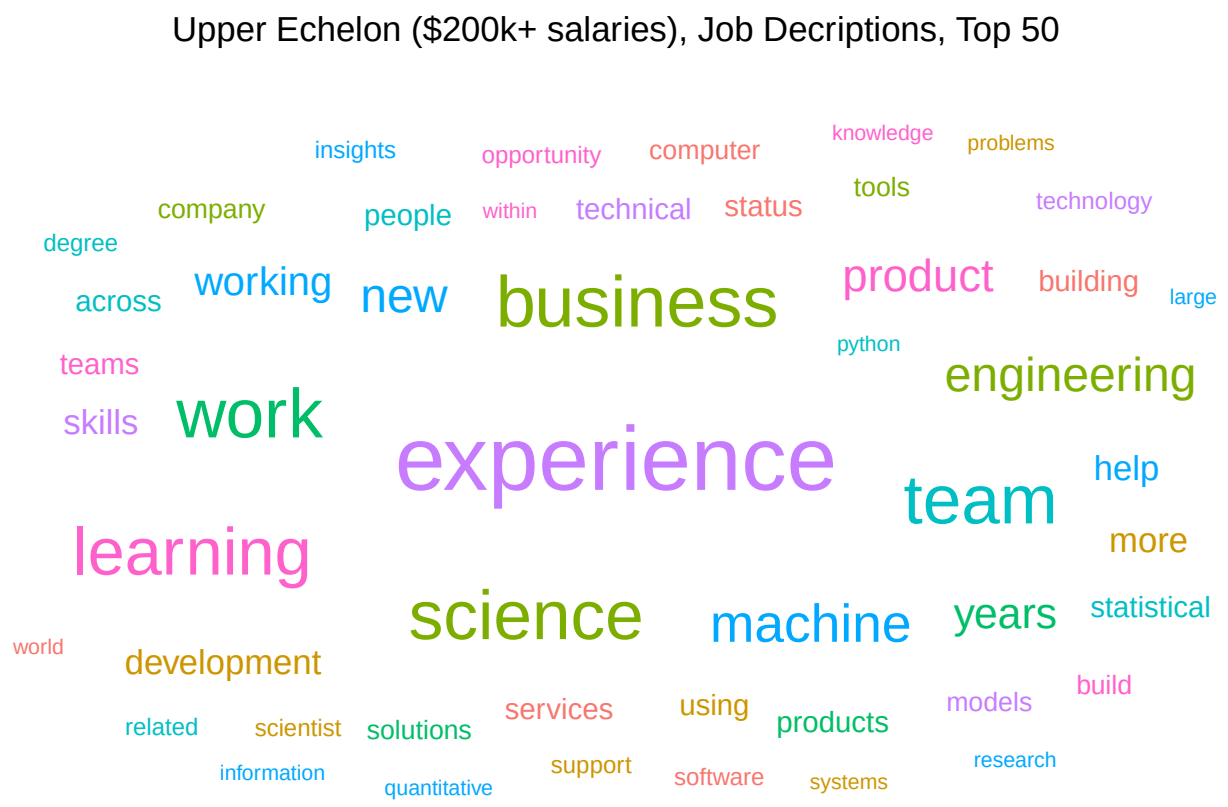
```
OneHundredToTwoHundredK <- combinedData %>% filter(combinedData$MeanSalary>=100 & combinedData$MeanSala

# per https://policyadvice.net/insurance/insights/average-american-income/
LessThanUSMeanSalary <- combinedData %>% filter(combinedData$MeanSalary < 51.9)
UpperEchelon <- combinedData %>% filter(combinedData$MeanSalary > 200)
```

Upper Echelon Key Characteristics

In this section, we explored the key characteristics for the data scientist and data analyst jobs that fall in the highest salary tier. We selected the top 50 keywords for our plots because they are the most relevant keywords and they fit conveniently in the word cloud.

```
description <- data.frame(UpperEchelon$Job.Description)
desclist <- unlist(description)
desclist <- str_remove_all(desclist, '[[:punct:]]')
desclist <- str_remove_all(desclist, '[\r\n]')
desclist <- str_remove_all(desclist, '[0-9]')
desclist <- str_remove_all(desclist, '[\r\n]')
desclist <- tolower(desclist)
descsplit <- strsplit(desclist, " ")


frequenciesUE <- table(unlist(descsplit))
frequenciesUE <- data.frame(frequenciesUE)
colnames(frequenciesUE) <- c('word', 'count')
omit <- c(" ", "and", "with", "from", "for", "the", "our", "your", "are", "will", "with", "that", "othe
#get rid of infrequently used words
frequenciesUE <- subset(frequenciesUE, as.numeric(count) >= 3)
# Get rid of prepositions, etc
frequenciesUE_relevant <- frequenciesUE[!frequenciesUE$word %in% omit, ]
#Sorting most frequent to least frequent, hyphen to sort DESCENDING
```

```
frequenciesUE_relevant <- frequenciesUE_relevant[order(-frequenciesUE_relevant$count), ]

#Plotting
overweighted <- c("data", "analyst", "analysis", "analytics")
freqplot <- frequenciesUE_relevant[!frequenciesUE_relevant$word %in% overweighted, ]
freqplot <- freqplot[1:50, ]
ggplot(freqplot, aes(label = word, size = count, color = factor(sample.int(8, nrow(freqplot), replace =
  geom_text_wordcloud()+
  scale_radius(range = c(0, 12), limits = c(0, NA)) +
  theme_minimal()+
  ggtitle("Upper Echelon ($200k+ salaries), Job Decriptions, Top 50") +
  theme(plot.title = element_text(hjust = 0.5))
```



Lowest Income Key Characteristics

```
description <- data.frame(LessThanUSMeanSalary$Job.Description)
desclist <- unlist(description)
desclist <- str_remove_all(desclist, '[[:punct:]]')
desclist <- str_remove_all(desclist, '[\r\n]')
desclist <- str_remove_all(desclist, '[0-9]')
desclist <- str_remove_all(desclist, '[\r\n]')
desclist <- tolower(desclist)
descsplit <- strsplit(desclist, " ")
```

```
frequenciesUSMean <- table(unlist(descsplit))
frequenciesUSMean <- data.frame(frequenciesUSMean)
colnames(frequenciesUSMean) <- c('word', 'count')
#get rid of infrequently used words
frequenciesUSMean <- subset(frequenciesUSMean, as.numeric(count) >= 3)
# Get rid of prepositions, etc
frequenciesUSMean_relevant <- frequenciesUSMean[!frequenciesUSMean$word %in% omit, ]
#Sorting most frequent to least frequent, hyphen to sort DESCENDING
frequenciesUSMean_relevant <- frequenciesUSMean_relevant[order(-frequenciesUSMean_relevant$count), ]


#Plotting
freqplot <- frequenciesUSMean_relevant[!frequenciesUSMean_relevant$word %in% overweighted, ]
freqplot <- freqplot[1:50, ]
ggplot(freqplot, aes(label = word, size = count, color = factor(sample.int(8, nrow(freqplot), replace =
  geom_text_wordcloud()+
  scale_radius(range = c(0, 12), limits = c(0, NA)) +
  theme_minimal()+
  ggtitle("Less Than US Mean (~$51K) Salaries, Job Descriptions, Top 50") +
  theme(plot.title = element_text(hjust = 0.5))
```

Less Than US Mean (~$51K) Salaries, Job Descriptions, Top 50

One Hundred to Two Hundred Thousand Dollar Salary Characteristics

```
description <- data.frame(OneHundredToTwoHundredK$Job.Description)
desclist <- unlist(description)
desclist <- str_remove_all(desclist, '[[:punct:]]')
desclist <- str_remove_all(desclist, '[\r\n]')
desclist <- str_remove_all(desclist, '[0-9]')
desclist <- str_remove_all(desclist, '[\r\n]')
desclist <- tolower(desclist)
descsplit <- strsplit(desclist, " ")


frequencies100K200K <- table(unlist(descsplit))
frequencies100K200K <- data.frame(frequencies100K200K)
colnames(frequencies100K200K) <- c('word', 'count')
#get rid of infrequently used words
frequencies100K200K <- subset(frequencies100K200K, as.numeric(count) >= 3)
# Get rid of prepositions, etc
frequencies100K200K_relevant <- frequencies100K200K[!frequencies100K200K$word %in% omit, ]
#Sorting most frequent to least frequent, hyphen to sort DESCENDING
frequencies100K200K_relevant <- frequencies100K200K_relevant[order(-frequencies100K200K_relevant$count)
#Plotting
freqplot <- frequencies100K200K_relevant[!frequencies100K200K_relevant$word %in% overweighted, ]
freqplot <- freqplot[1:50, ]
ggplot(freqplot, aes(label = word, size = count, color = factor(sample.int(8, nrow(freqplot), replace =
  geom_text_wordcloud()+
  scale_radius(range = c(0, 12), limits = c(0, NA)) +
  theme_minimal()+
  ggtitle("100K to 200K Salaries, Job Descriptions, Top 50") +
  theme(plot.title = element_text(hjust = 0.5))
```

100K to 200K Salaries, Job Descriptions, Top 50

All Job Descriptions

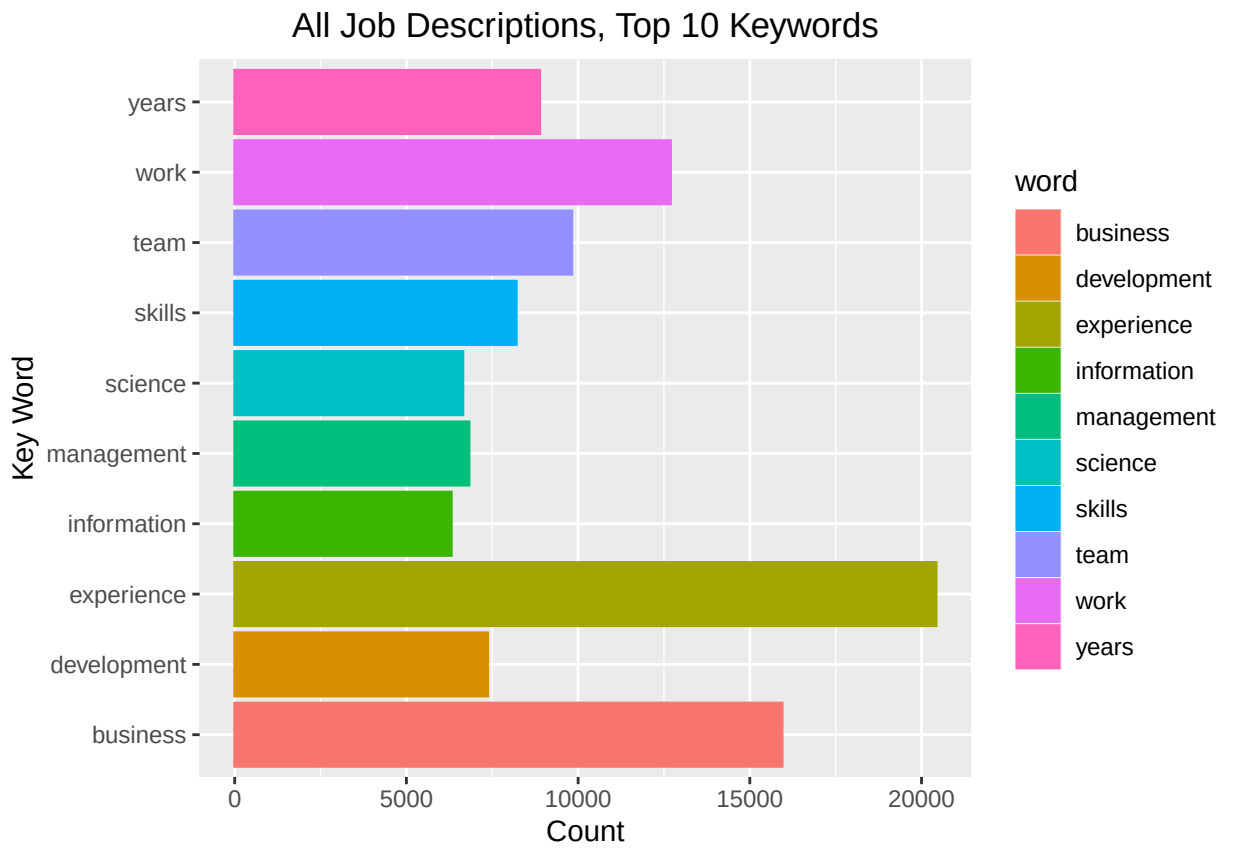
```
description <- data.frame(combinedData$Job.Description)
desclist <- unlist(description)
desclist <- str_remove_all(desclist, '[[:punct:]]')
desclist <- str_remove_all(desclist, '[\r\n]')
desclist <- str_remove_all(desclist, '[0-9]')
desclist <- str_remove_all(desclist, '[\r\n]')
desclist <- tolower(desclist)
descsplit <- strsplit(desclist, " ")


frequenciesall <- table(unlist(descsplit))
frequenciesall <- data.frame(frequenciesall)
colnames(frequenciesall) <- c('word', 'count')
#get rid of infrequently used words
frequenciesall <- subset(frequenciesall, as.numeric(count) >= 3)
# Get rid of prepositions, etc
frequenciesall_relevant <- frequenciesall[!frequenciesall$word %in% omit, ]
#Sorting most frequent to least frequent, hyphen to sort DESCENDING
frequenciesall_relevant <- frequenciesall_relevant[order(-frequenciesall_relevant$count), ]
#Plotting
freqplot <- frequenciesall_relevant[!frequenciesall_relevant$word %in% overweighted, ]
freqplot <- freqplot[1:50, ]
```

```
ggplot(freqplot, aes(label = word, size = count, color = factor(sample.int(8, nrow(freqplot), replace =
  geom_text_wordcloud()+
  scale_radius(range = c(0, 12), limits = c(0, NA)) +
  theme_minimal()+
  ggtitle("All Job Descriptions, Top 50") +
  theme(plot.title = element_text(hjust = 0.5))
```


```
head(freqplot)
```

```
##             word count
## 30966 experience 20425
## 10770   business 15941
## 96395       work 12691
## 87111       team  9817
## 97548      years  8882
## 80060     skills  8197
```

```
freqbar <- freqplot[1:10, ]
ggplot(freqbar, aes(x = word, y = count, color = word, fill = word))+
  geom_col()+
  ggtitle("All Job Descriptions, Top 10 Keywords") +
  xlab("Key Word")+
  ylab("Count")+
  theme(plot.title = element_text(hjust = 0.5))+
  coord_flip()
```



Comments

The biggest in demand skill for data science is experience. Business is another commonly used keyword. Team, development, learning, and work showed up many times as well.

A few ideas for further analysis: Combine plurals and singulars, for example, team and teams. We could use the same techniques to gather all instances of programming languages that are listed in job descriptions to help determine which languages are more valuable than others. We could find a way to loop through the data and make a dictionary for strings of more than one word in order to provide a clearer picture of important skills, like "interpersonal skills."

Team Notes Our team met and worked together concurrently for a combined 8 hours. Individual contributions were also made outside of the team meetings. We used Google Meet for synchronous meetings and Slack for daily communication. Nearly all of the work was performed through collaborative discussion.

Data Sources

The data was pulled from kaggle.

We used two datasets for our final analysis in R.

https://www.kaggle.com/andrewmvd/data-analyst-jobs

https://www.kaggle.com/andrewmvd/data-scientist-jobs