Data 607 Project 3

Teams 5

null

```
#Load Libraries
library(tidyr)
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(textdata)
library(knitr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(wordcloud)
```

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

```
library(RColorBrewer)
library(ggplot2)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(ggthemes)
library(png)
```

Load the dataset

```
job_df<- read.csv('https://raw.githubusercontent.com/quaere1verum/sps_public/master/data_scientist_42928

glimpse(job_df)
```

```
## Rows: 200
## Columns: 42
## $ crawl_timestamp            <chr> "2021-10-03 04:25:55 +0000", "2021-10-03 0~
## $ url                        <chr> "https://www.simplyhired.com/job/mGqvZsPML~
## $ job_title                  <chr> "Data Scientist, GBM Analytics", "Sr. Data~
## $ category                   <chr> "", "Computer/internet", "Full-time", "Ful~
## $ company_name               <chr> "Facebook", "LOVEFOODIES INC", "SEMCON GRO~
## $ city                       <chr> "Chicago", "Fremont", "Atlantic City", "At~
## $ state                      <chr> "IL", "CA", "NJ", "GA", "NY", "TX", "NY", ~
## $ country                    <chr> "United States", "US", "United States", "U~
## $ inferred_city              <chr> "Chicago", "Fremont", "Atlantic city", "At~
## $ inferred_state             <chr> "Illinois", "California", "New jersey", "G~
## $ inferred_country           <chr> "United states", "United states", "United ~
## $ post_date                  <chr> "2021-10-03", "2021-10-03", "2021-10-03", ~
## $ job_description            <chr> "We are seeking an experienced Data Scient~
## $ job_type                   <chr> "Undefined", "Undefined", "Full Time", "Fu~
## $ salary_offered             <chr> "$110,000 - $150,000 a year", "Pay: $3,000~
## $ job_board                  <chr> "simplyhired", "indeed", "simplyhired", "s~
## $ geo                        <chr> "United States", "usa", "United States", "~
## $ cursor                     <dbl> 1.633266e+15, 1.633277e+15, 1.633281e+15, ~
## $ contact_email              <chr> "accommodations-ext@fb.com", "", "", "", "~
## $ contact_phone_number       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ uniq_id                    <chr> "d578cbc1ebd47ee77eba9e981f3c2582", "acc80~
## $ html_job_description       <chr> "<!DOCTYPE html PUBLIC \"-//W3C//DTD HTML ~
## $ valid_through              <chr> "", "", "", "", "", "", "", "", "", "", ""~
## $ has_expired                <chr> "false", "false", "false", "false", "false~
## $ inferred_iso3_lang_code    <chr> "eng", "eng", "eng", "eng", "eng", "eng", ~
## $ latest_expiry_check_date   <chr> "2021-10-03", "2021-10-03", "2021-10-03", ~
## $ duplicate_status           <chr> "no", "no", "no", "no", "no", "no", "no", ~
## $ duplicate_of               <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ inferred_department_name   <chr> "IT", "IT", "IT", "IT", "IT", "IT", "IT", ~
## $ inferred_department_score  <int> 91, 99, 100, 97, 91, 97, 97, 97, 97, 96, 9~
## $ inferred_job_title         <chr> "Data scientist", "Data scientist", "Tool ~
## $ is_remote                  <chr> "false", "true", "false", "false", "false"~
## $ inferred_salary_currency   <chr> "USD", "USD", "USD", "USD", "USD", "USD", ~
## $ inferred_salary_time_unit  <chr> "yearly", "monthly", "yearly", "yearly", "~
## $ inferred_salary_from       <int> 110000, 3000, 92000, 74000, 130000, 150000~
## $ inferred_salary_to         <dbl> 150000, 10000, 120000, 99000, 170000, 1500~
```

```
## $ inferred_skills             <chr> "Quantitative Analysis|Marketing Analytics~
## $ inferred_company_type       <chr> "company", "company", "company", "agency",~
## $ inferred_company_type_score <int> 97, 100, 100, 100, 97, 98, 98, 98, 98, 100~
## $ inferred_seniority_level    <chr> "Mid Level", "Mid Level", "Mid Level", "Mi~
## $ apply_url                   <chr> "", "https://www.indeed.com/job/sr-data-sc~
## $ logo_url                    <chr> "https://www.simplyhired.com/serp/imgkibqy~
```

```
job_df$state[toupper(job_df$state)=="CALIFORNIA"]<-"CA"
job_df$state[toupper(job_df$state)=="NEW YORK"]<-"NY"
job_df$state[toupper(job_df$state)=="COLORADO"]<-"CO"
job_df$state[toupper(job_df$state)=="MARYLAND"]<-"MD"
job_df$state[toupper(job_df$state)=="ILLINOIS"]<-"IL"
job_df$state[is.na(job_df$state)|job_df$state==""]<-"unknown"

job_df$inferred_salary_time_unit[job_df$inferred_salary_time_unit==""]<-"unknown"
```

Skill Set/Skill Company should be in Database

```
test<-strsplit(job_df$inferred_skills, split = "\\|")

skillset <-data.frame(skill=character())
s <-length(test)

for (i in 1:s){
  for (j in 1:lengths(test[i])){
    rows<-data.frame(skill=test[[i]][j])
    skillset <-rbind(skillset,rows)
  }
}

skill_company <-data.frame(skillid=character(),companyid=character(),state=character())

test <-job_df %>% select (company_name,inferred_skills,state)%>%
    filter(inferred_skills!="")

datarow<-nrow(test)

for (i in 1:datarow){
  infer_byrow <-c(skillid=strsplit(test[[i,2]], split = "\\|"))
  rows<-data.frame(skillid=infer_byrow,companyid=test[[i,1]],state=test[[i,3]])
  skill_company <-rbind(skill_company,rows)
}
```

lm function

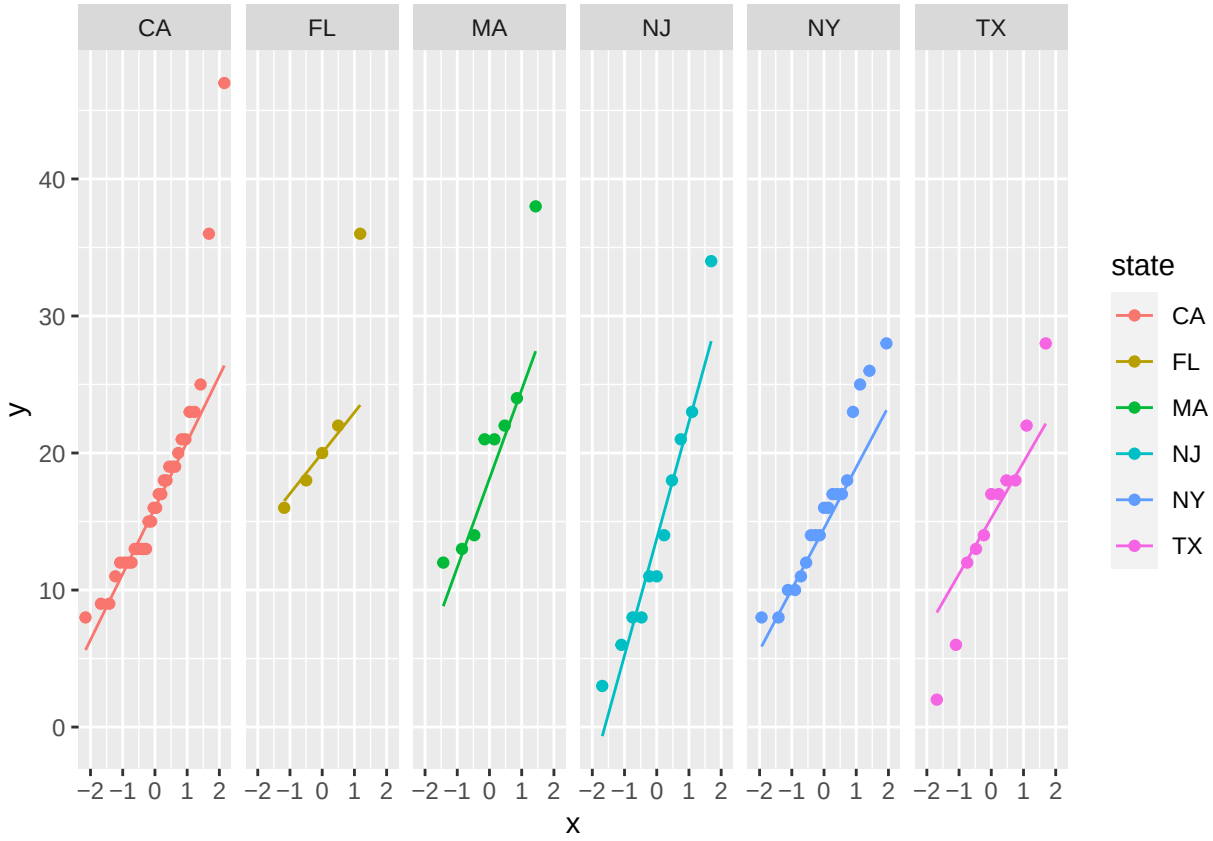
```
##### lm Function

company_state <- skill_company %>% select(companyid,state)%>%
  group_by(companyid,state) %>%
  summarise(skill_count=n())
```

```
## `summarise()` has grouped output by 'companyid'. You can override using the `.groups` argument.
```

```
company_state <- company_state %>% mutate(ratio= skill_count/sum(skill_count))

ggplot(company_state %>% filter (state %in% c("CA","NY","TX","NE","FL","NJ","MA")) , aes(sample=skill_c
  stat_qq(aes(color =state))+
  stat_qq_line(aes(color = state))+
  facet_grid(~state)
```




```
lm_company_state <- lm(ratio~state,data=company_state)
summary(lm_company_state)
```

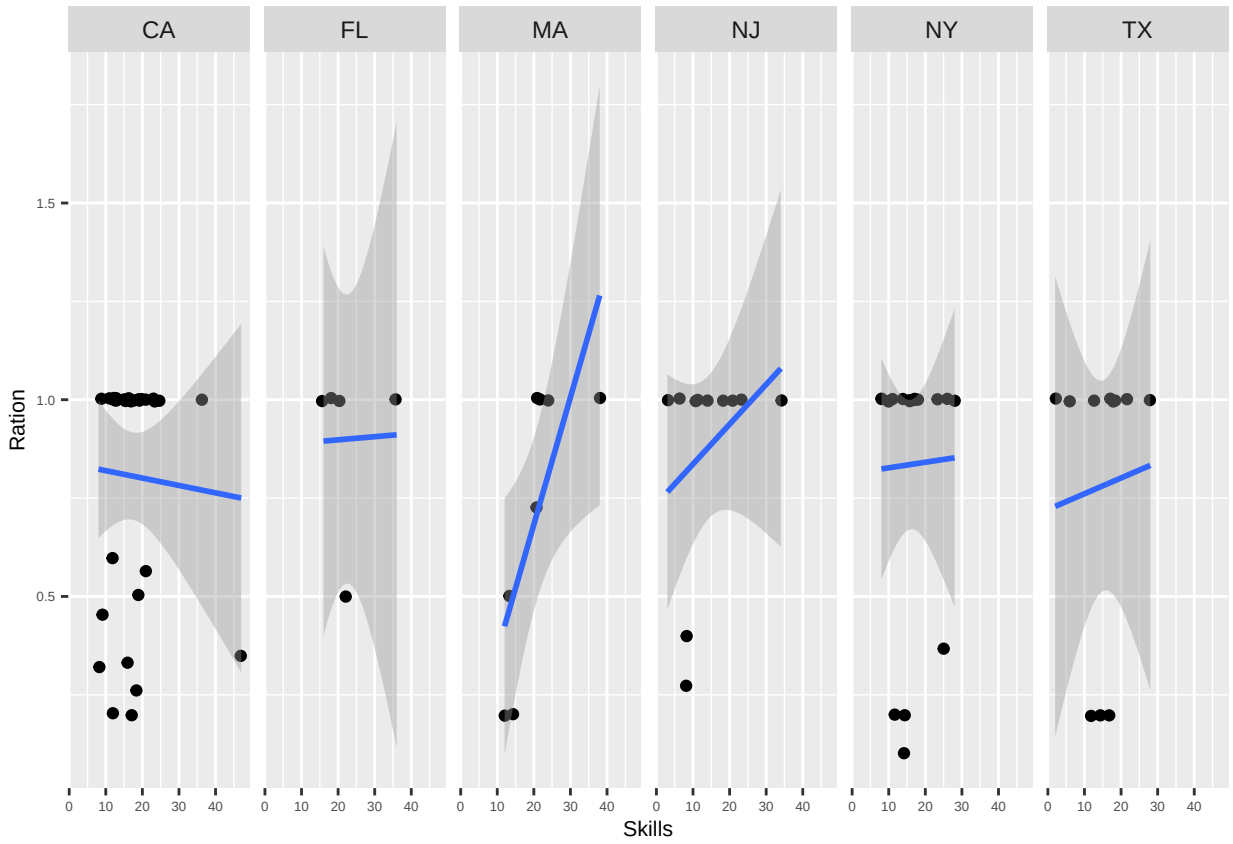
```
##
## Call:
## lm(formula = ratio ~ state, data = company_state)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8225  0.0000  0.1173  0.1941  0.4398
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7777778  0.1768412   4.398 2.07e-05 ***
## stateCA       0.0281282  0.1849450   0.152    0.879
## stateCO      -0.2277778  0.3536824  -0.644    0.521
## stateCT       0.2222222  0.2339389   0.950    0.344
```

```
## stateDC       -0.1150327  0.2339389  -0.492    0.624
## stateDE        0.2222222  0.2796105   0.795    0.428
## stateFL        0.1222222  0.2236884   0.546    0.586
## stateGA        0.1597222  0.2073647   0.770    0.442
## stateIL        0.0001268  0.1961877   0.001    0.999
## stateIN        0.2222222  0.2796105   0.795    0.428
## stateKS        0.2222222  0.3536824   0.628    0.531
## stateMA       -0.0747605  0.2073647  -0.361    0.719
## stateMD        0.2222222  0.2113657   1.051    0.295
## stateMN        0.2222222  0.2796105   0.795    0.428
## stateMO        0.2222222  0.3536824   0.628    0.531
## stateMULTIPLE -0.2287582  0.3536824  -0.647    0.519
## stateNC       -0.2175935  0.2500912  -0.870    0.386
## stateNJ        0.1018460  0.1995038   0.510    0.610
## stateNV        0.2222222  0.3536824   0.628    0.531
## stateNY        0.0576385  0.1902909   0.303    0.762
## stateOH        0.2222222  0.2500912   0.889    0.376
## stateOR        0.2222222  0.3536824   0.628    0.531
## statePA        0.2222222  0.2796105   0.795    0.428
## stateRemote   -0.1777778  0.2796105  -0.636    0.526
## stateTN       -0.4444444  0.3536824  -1.257    0.211
## stateTX        0.0040404  0.1995038   0.020    0.984
## stateunknown   0.2222222  0.2339389   0.950    0.344
## stateUT       -0.0484545  0.2500912  -0.194    0.847
## stateVA        0.1048989  0.1937198   0.541    0.589
## stateWA       -0.0777778  0.2073647  -0.375    0.708
## stateWI        0.2222222  0.3536824   0.628    0.531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3063 on 148 degrees of freedom
## Multiple R-squared:  0.1423, Adjusted R-squared:  -0.03158
## F-statistic: 0.8184 on 30 and 148 DF,  p-value: 0.7346
```

```
lmplot<- company_state %>% filter (state %in% c("CA","NY","TX","NE","FL","NJ","MA"))

ggplot(data = lmplot, aes(x = skill_count, y = ratio)) +
  geom_jitter() +
  geom_smooth(method = "lm")+
  facet_grid(~state)+
  xlab("Skills")+
  ylab("Ration")+
  theme(axis.text.x = element_text(size=5),
        axis.text.y = element_text(size=5),
        axis.title.x = element_text(size = 8),
        axis.title.y = element_text(size = 8))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

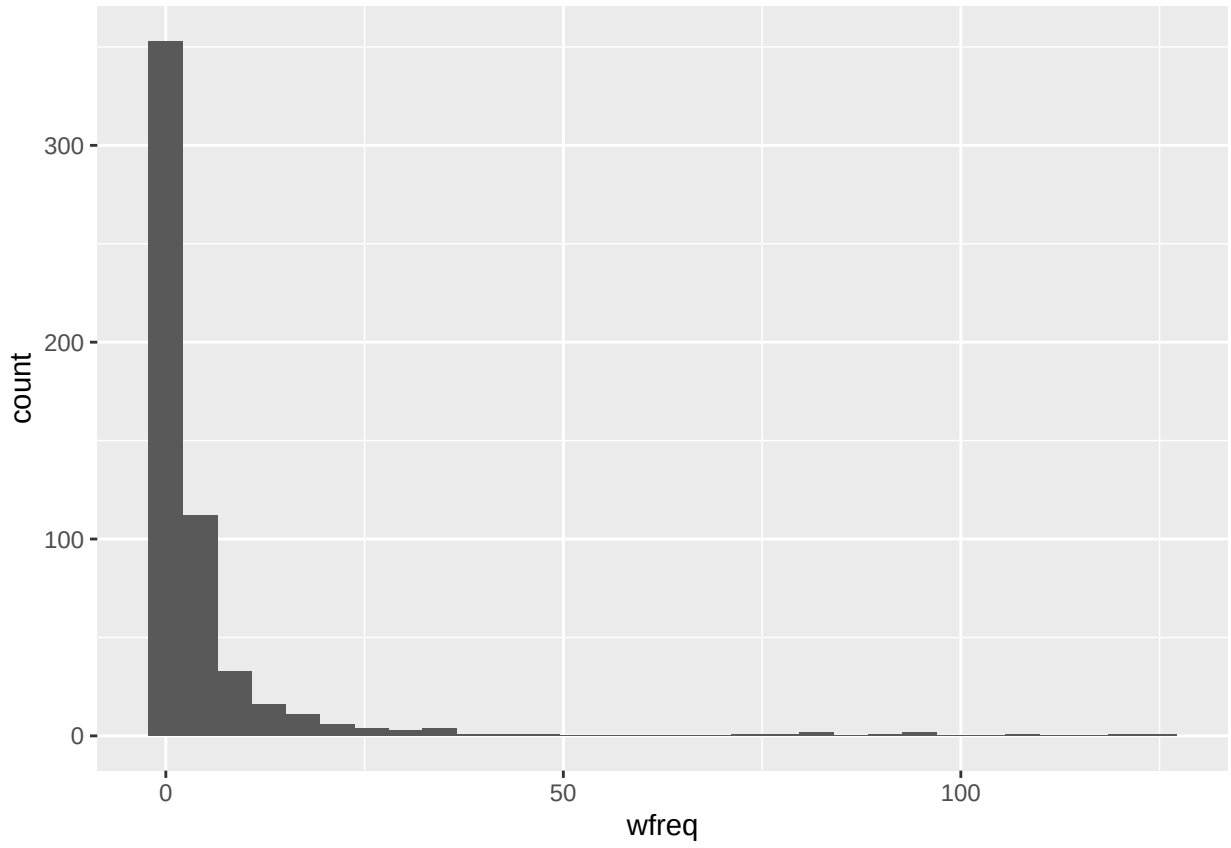


Plots

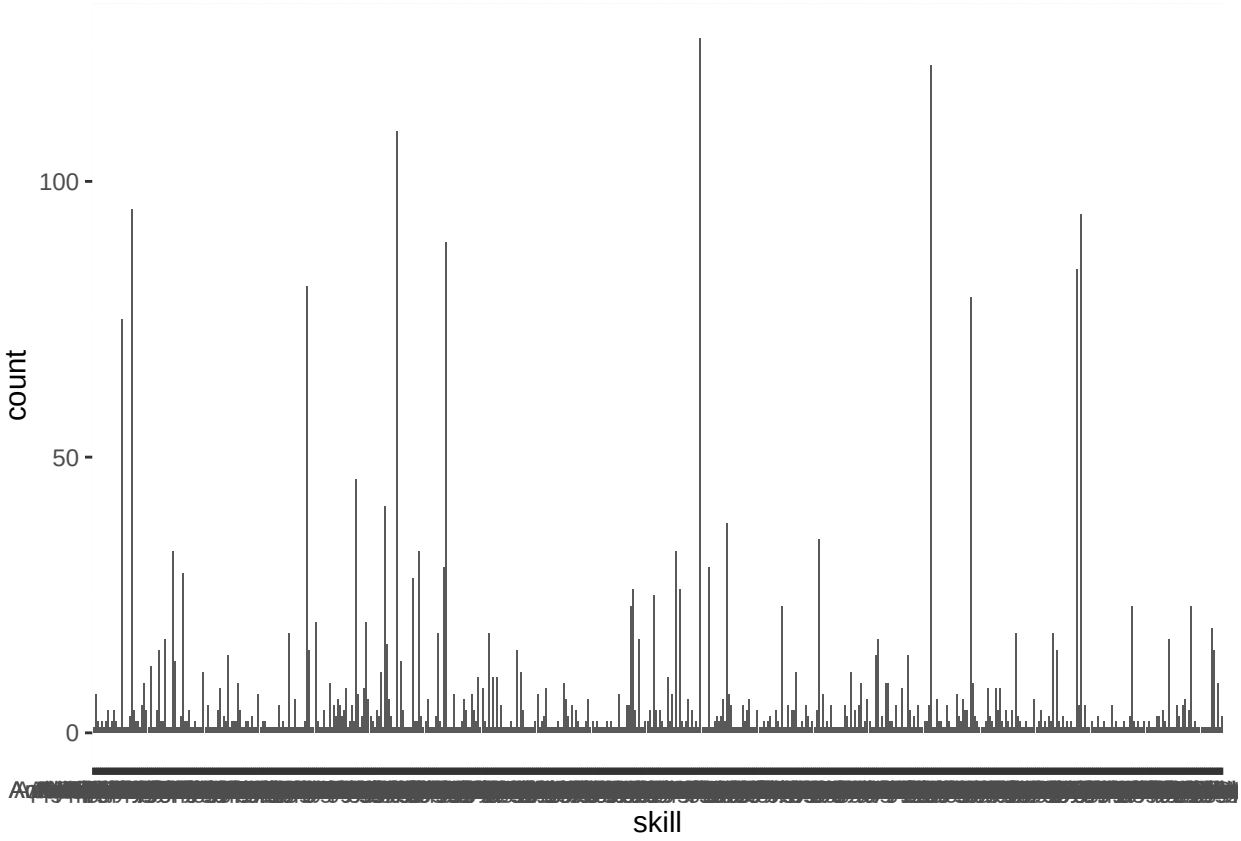
```
word_freq<- skillset %>% group_by(skill)%>%
  summarise(wfreq=n())

# Plots #
ggplot(word_freq, aes(wfreq),horizontal = TRUE) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

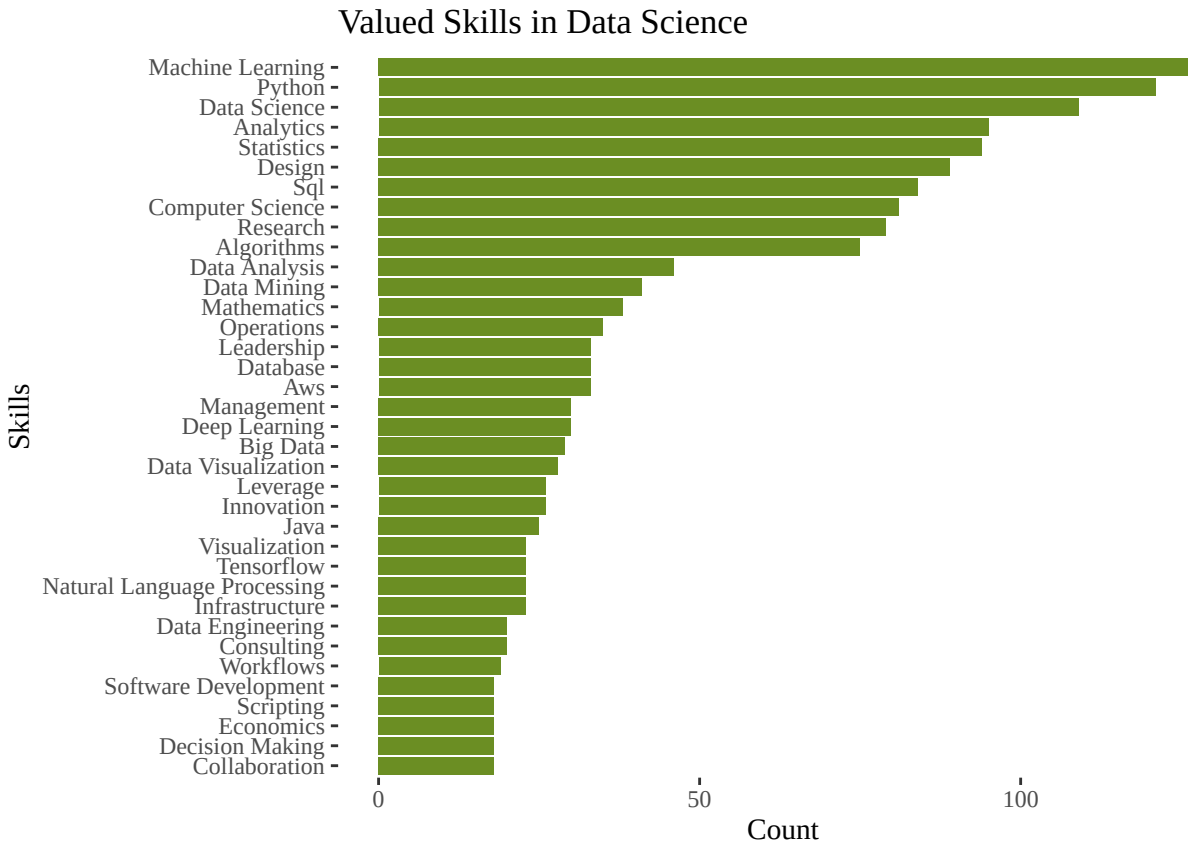



```
ggplot(skillset,aes(skill))+
  geom_bar()
```




```
ggplot(top_n(word_freq,35), aes(x=reorder(skill,wfreq),y = wfreq)) +
  geom_bar(stat='identity',fill="olivedrab")+
  coord_flip()+
  ylab("Count")+
  xlab("Skills")+
  theme_tufte()+
  ggtitle("Valued Skills in Data Science")
```

```
## Selecting by wfreq
```

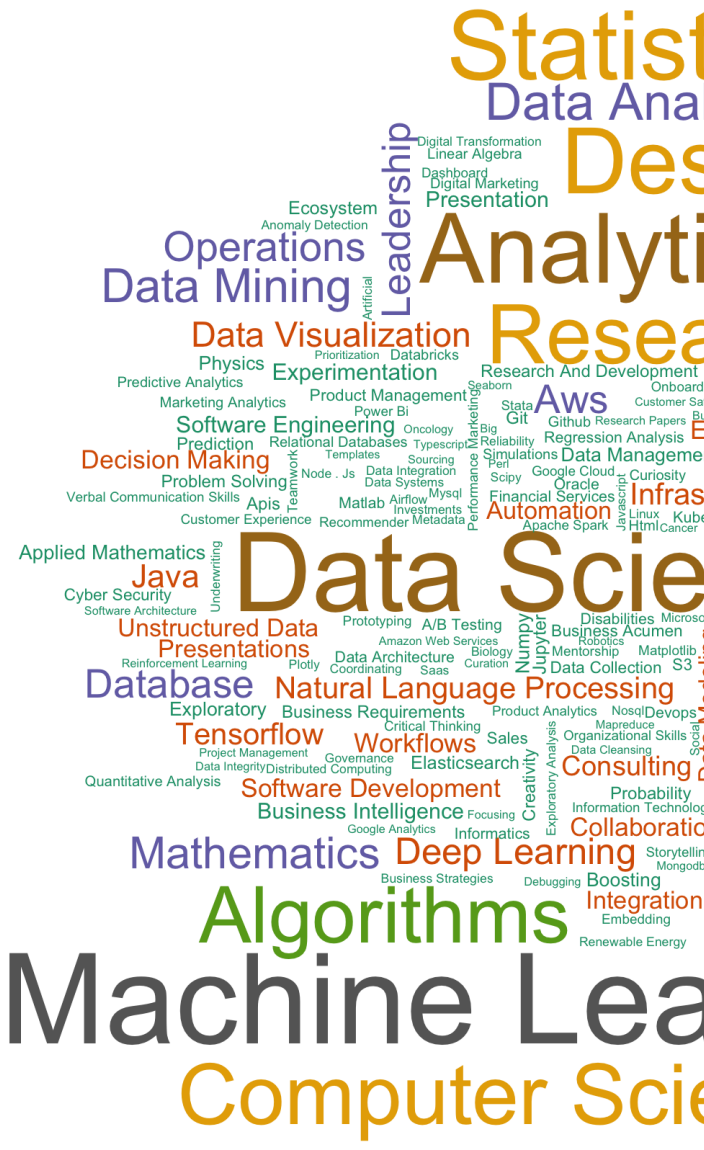



```
# Word Cloud #
png("wordcloud.png",width = 12, height = 8,units = "in", res=300)
par(mar=rep(0,4))
set.seed(10142021)
word_freq <- word_freq %>% arrange(desc(wfreq))

wordcloud(word_freq$skill,freq = word_freq$wfreq,scale=c(3.5,0.25),
          colors=brewer.pal(8,"Dark2"))

wordcloud_pic <- '/Users/admin/Downloads/wordcloud.png'

include_graphics(wordcloud_pic)
```


Statist
Data Ana
Des
Digital Transformation
Linear Algebra
Dashboard
Digital Marketing
Presentation
Ecosystem
Anomaly Detection
Analyti
Operations
Data Mining
Leadership
Artificial
Resea
Data Visualization
Prioritization
Databricks
Physics
Experimentation
Research And Development
Predictive Analytics
Seaborn
Onboard
Product Management
Marketing Analytics
Aws
Stata
Customer Sa
Power Bi
Git
Github
Research Papers
Software Engineering
Oncology
Big
Regression Analysis
Prediction
Relational Databases
Typescript
Reliability
Simulations
Data Manageme
Decision Making
Templates
Sourcing
Perl
Performance Marketing
Teamwork
Node . Js
Data Integration
Scipy
Google Cloud
Curiosity
Problem Solving
Data Systems
Oracle
Infras
Verbal Communication Skills
Apis
Mysql
Financial Services
Matlab
Airflow
Investments
Automation
Javascript
Linux
Kube
Customer Experience
Recommender
Metadata
Apache Spark
Html
Cancer
Applied Mathematics
Data Scie
Java
Underwriting
Cyber Security
Software Architecture
Prototyping
A/B Testing
Disabilities
Microsc
Unstructured Data
Numpy
Jupyter
Business Acumen
Amazon Web Services
Robotics
Presentations
Mentorship
Matplotlib
Reinforcement Learning
Plotly
Coordinating
Saas
Data Architecture
Biology
Curation
Data Collection
S3
Database
Natural Language Processing
Exploratory
Business Requirements
Product Analytics
Nosql
Devops
Critical Thinking
Mapreduce
Tensorflow
Workflows
Sales
Organizational Skills
Project Management
Governance
Data Cleansing
Data Integrity
Distributed Computing
Elasticsearch
Consulting
Quantitative Analysis
Creativity
Exploratory Analysis
Software Development
Probability
Business Intelligence
Focusing
Information Technolog
Google Analytics
Informatics
Collaboratio
Mathematics
Deep Learning
Storytellin
Mongodb
Business Strategies
Debugging
Boosting
Integration
Algorithms
Embedding
Renewable Energy
Machine Lea
Computer Scie