PROJECT3

Team Sloth

3/21/2022

```
library(waterfalls)
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

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

```
library(ggplot2)
library(dplyr)
library(stringr)
library(rvest)
```

```
##
## Attaching package: 'rvest'

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

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

read the data

```
indeed_url<- "https://raw.githubusercontent.com/Benson90/Project3/main/indeed_job_dataset.csv"
indeeddata <-read.csv(indeed_url)

college_url<- "https://raw.githubusercontent.com/jayleecunysps/607data/main/tabn322.csv"
collegedata <-read.csv(college_url, skip = 1)


collegedata<-collegedata %>%
slice(-1,-36,-37,-38,-39,-40)

collegedata <-collegedata[,1:19]
```

```
jobtable <- indeeddata %>%
select("X","Job_Title","Queried_Salary","Job_Type","Skill","No_of_Skills","Company","Company_Industry")


#remove X,Other from the skilltable
skilltable <- indeeddata %>%
select("python","sql","machine.learning","r","hadoop","tableau","sas","spark","java")


#remove X, add QUeried_salary from the companytable
companytable <- indeeddata %>%
select("Company","Queried_Salary","Location","Company_Revenue","Company_Employees","Company_Industry","
```

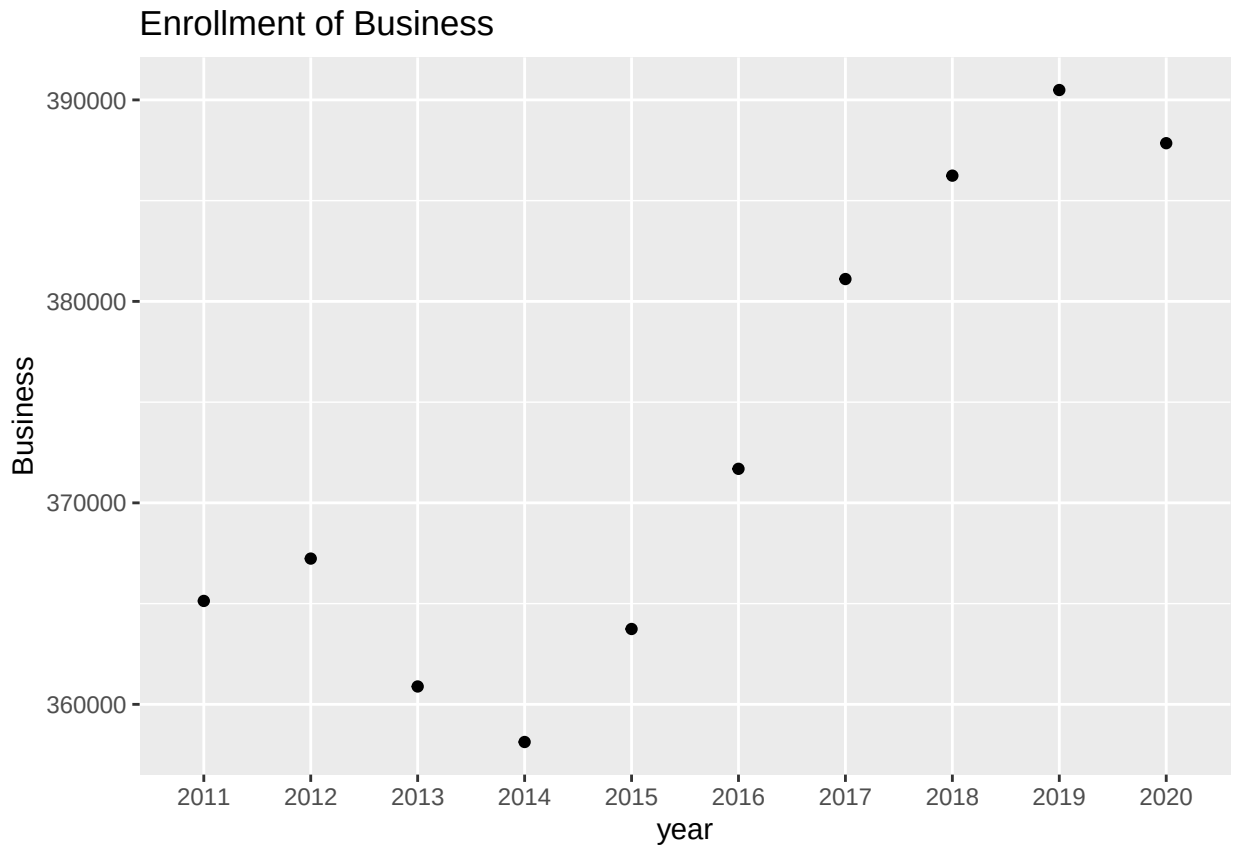
Introduction:

My group consisted of me(Al),Benson and Jay Lee. Our group had found many data sets but ultimately we decided upon an Indeed data set that contains information about jobs posting from Indeed that contained information regarding which company were hiring for what roles,the salaries,and skills they are required to know.My role was to compare salaries by job-type and possibly job-title.

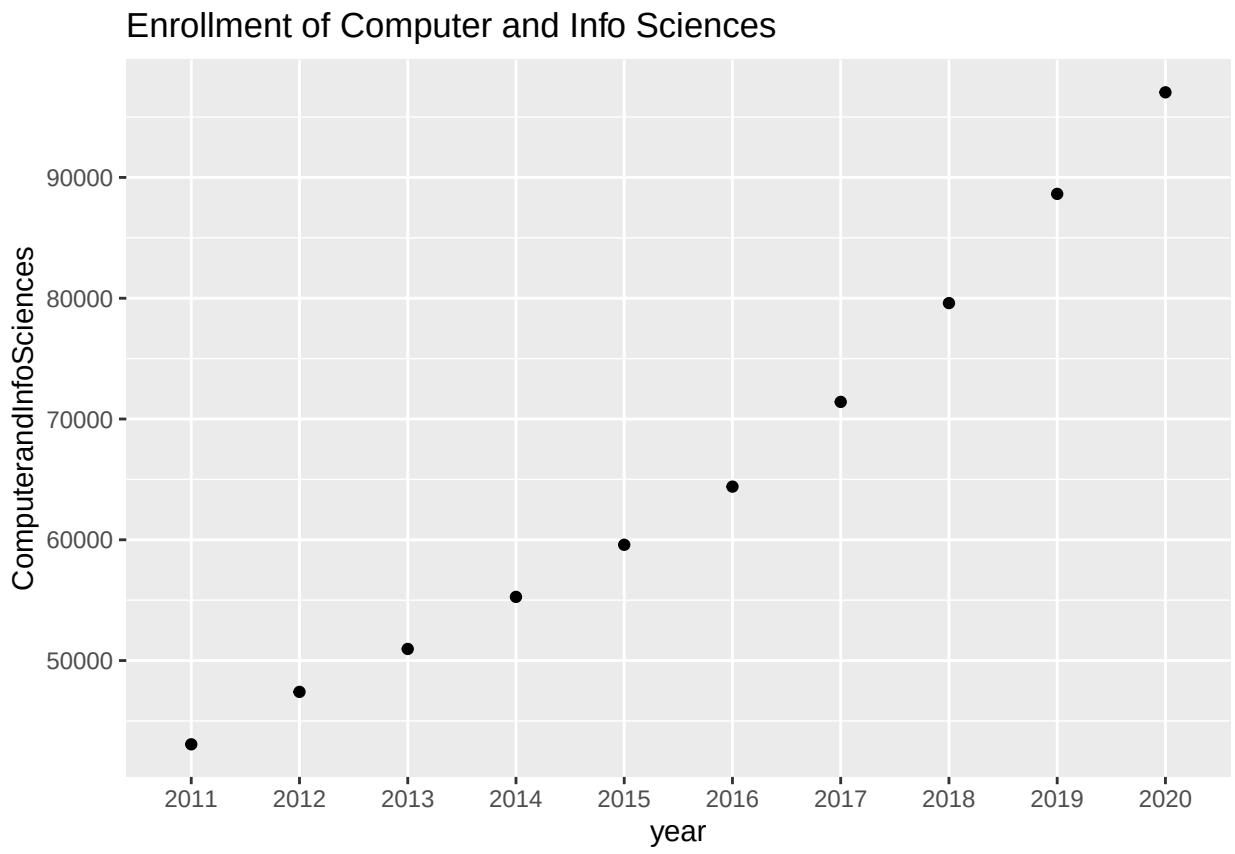
Including Plots

```
collegedata$Field.of.study <- as.character(str_replace_all(collegedata$Field.of.study,"[1-9]",""))
collegedata$Field.of.study <- as.character(str_replace_all(collegedata$Field.of.study,"[/]",""))
collegedata$Field.of.study <- gsub("   ","",as.character(collegedata$Field.of.study))
collegedataplot <- as.data.frame(t(collegedata))
names(collegedataplot) <- collegedataplot[1,]
collegedataplot = collegedataplot[-1,]
collegedataplot2 <- collegedataplot %>%
  select(c('Total','Business\\\\', 'Computer and information sciences and support \nservices','Mathemat
collegedataplot2<-collegedataplot2 %>%
slice(-1,-2,-3,-4,-5,-6,-7,-8)
year <- c(2011:2020)
collegedataplot2 <- cbind(collegedataplot2, year)
colnames(collegedataplot2) <- c('Total','Business','ComputerandInfoSciences','Mathandstatistics','year')
collegedataplot2$Total <- gsub(",", "", collegedataplot2$Total)
collegedataplot2$Business <- gsub(",", "", collegedataplot2$Business)
collegedataplot2$Mathandstatistics <- gsub(",", "", collegedataplot2$Mathandstatistics)
```

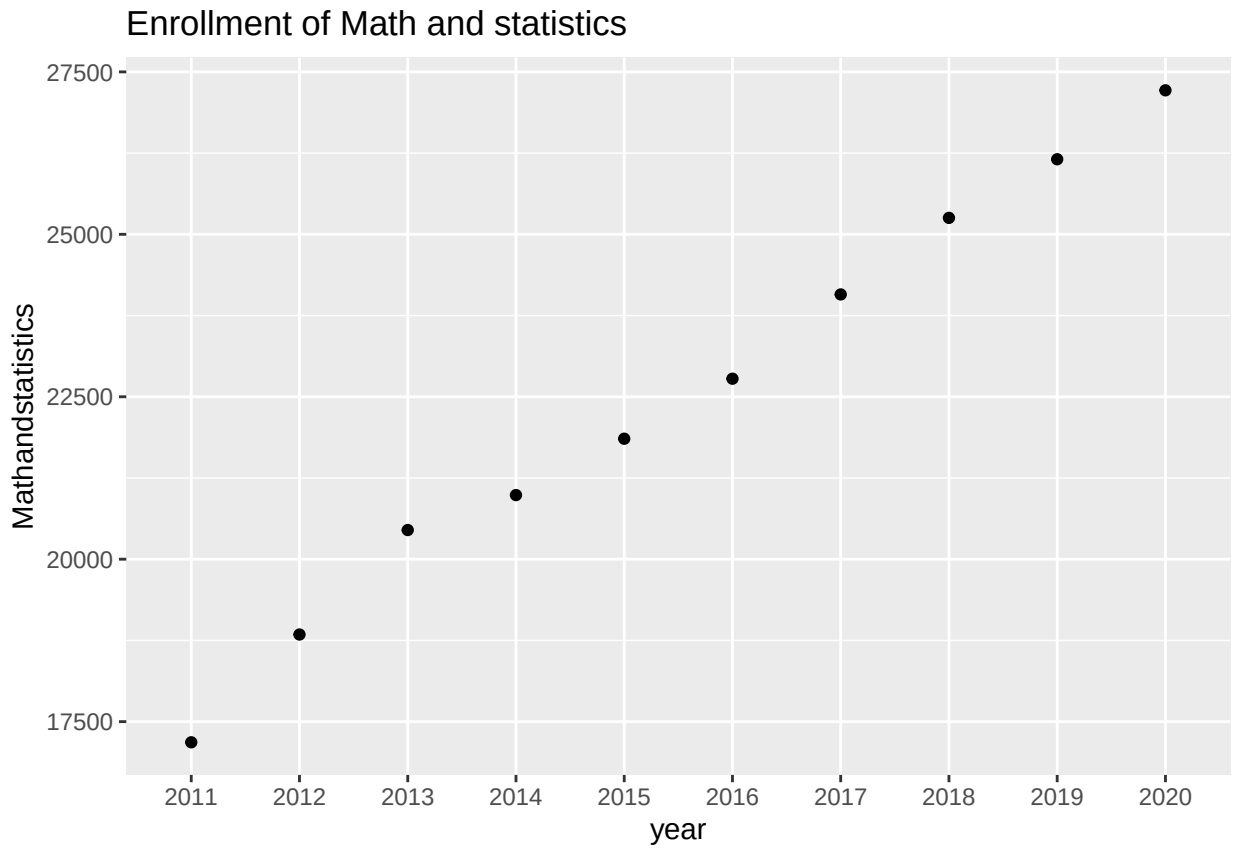
```
collegedataplot2$ComputerandInfoSciences <- gsub(",", "", collegedataplot2$ComputerandInfoSciences)
collegedataplot2$Total <-as.numeric(collegedataplot2$Total)
collegedataplot2$Business <-as.numeric(collegedataplot2$Business)
collegedataplot2$ComputerandInfoSciences <-as.numeric(collegedataplot2$ComputerandInfoSciences)
collegedataplot2$Mathandstatistics <-as.numeric(collegedataplot2$Mathandstatistics)
collegedataplot2$year <-as.factor(collegedataplot2$year)
pl1<-ggplot(data=collegedataplot2) +
  geom_point(aes(x=year,y=Business)) +
  ggtitle("Enrollment of Business")
pl1
```




```
pl2<-ggplot(data=collegedataplot2) +
  geom_point(aes(x=year,y=ComputerandInfoSciences)) +
  ggtitle("Enrollment of Computer and Info Sciences")
pl2
```




```
pl3<-ggplot(data=collegedataplot2) +
  geom_point(aes(x=year,y=Mathandstatistics)) +
  ggtitle("Enrollment of Math and statistics")
pl3
```




```
ComputerandInfoSciencesrate <- collegedataplot2$ComputerandInfoSciences/collegedataplot2$Total*100
Mathandstatisticsrate <- collegedataplot2$Mathandstatistics/collegedataplot2$Total*100
Businessrate <- collegedataplot2$Business/collegedataplot2$Total*100
rateofchange <- cbind(year,ComputerandInfoSciencesrate, Mathandstatisticsrate,Businessrate)

rateofchange
```

```
##       year ComputerandInfoSciencesrate Mathandstatisticsrate Businessrate
##  [1,] 2011                    2.509596              1.001251     21.27749
##  [2,] 2012                    2.645184              1.051299     20.49116
##  [3,] 2013                    2.769046              1.111129     19.60936
##  [4,] 2014                    2.955431              1.122209     19.14991
##  [5,] 2015                    3.144431              1.153264     19.19509
##  [6,] 2016                    3.352961              1.185891     19.35130
##  [7,] 2017                    3.650912              1.230756     19.48296
##  [8,] 2018                    4.018701              1.274976     19.50052
##  [9,] 2019                    4.403091              1.299249     19.39758
## [10,] 2020                    4.760868              1.335145     19.02694
```

You can also embed plots, for example:

Skills Requirement

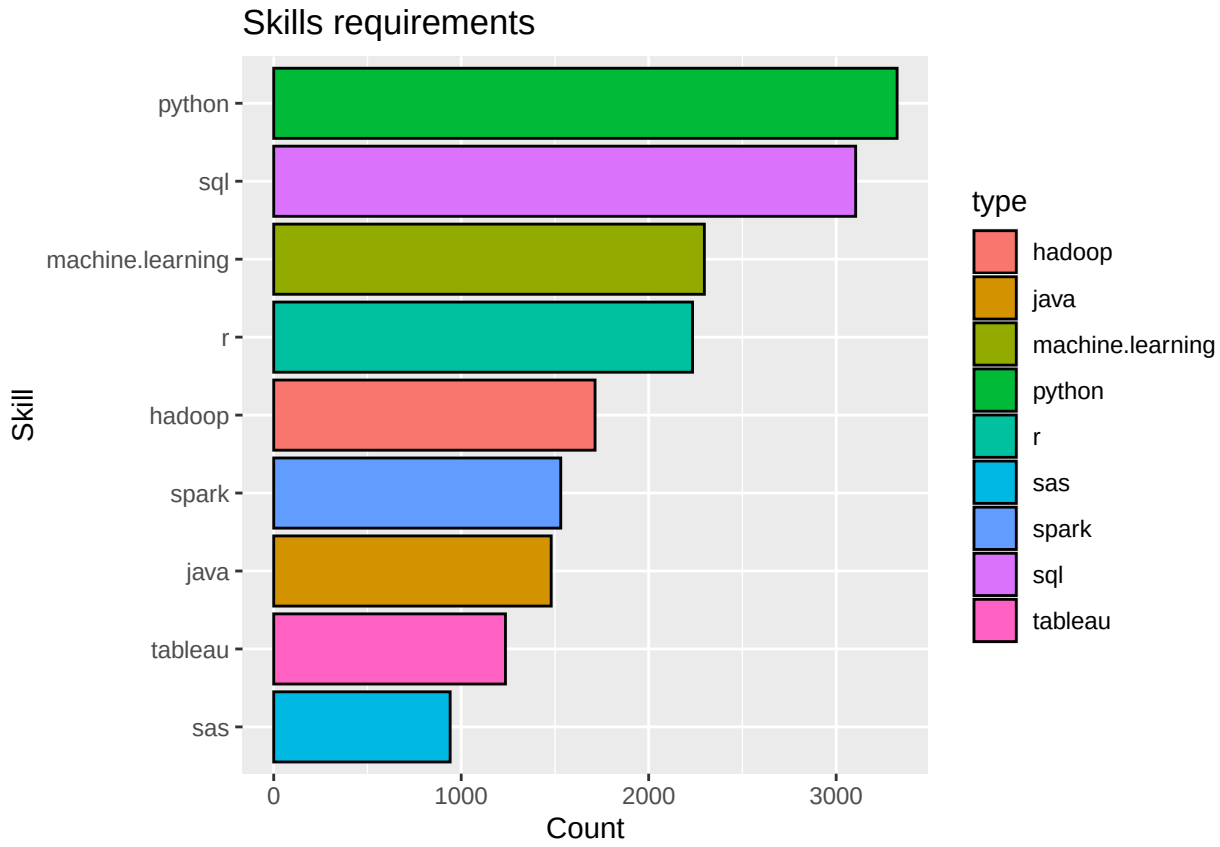
Based on the Indeed job posting from the Employer, the Employer is looking for an average of 3 related Data science skills for their open position. Also, Python and SQL are the most preferred for data science-related

jobs.

```
num_skills <- skilltable %>%
  mutate(total = rowSums(across(where(is.numeric))))

summary(num_skills$total)
```

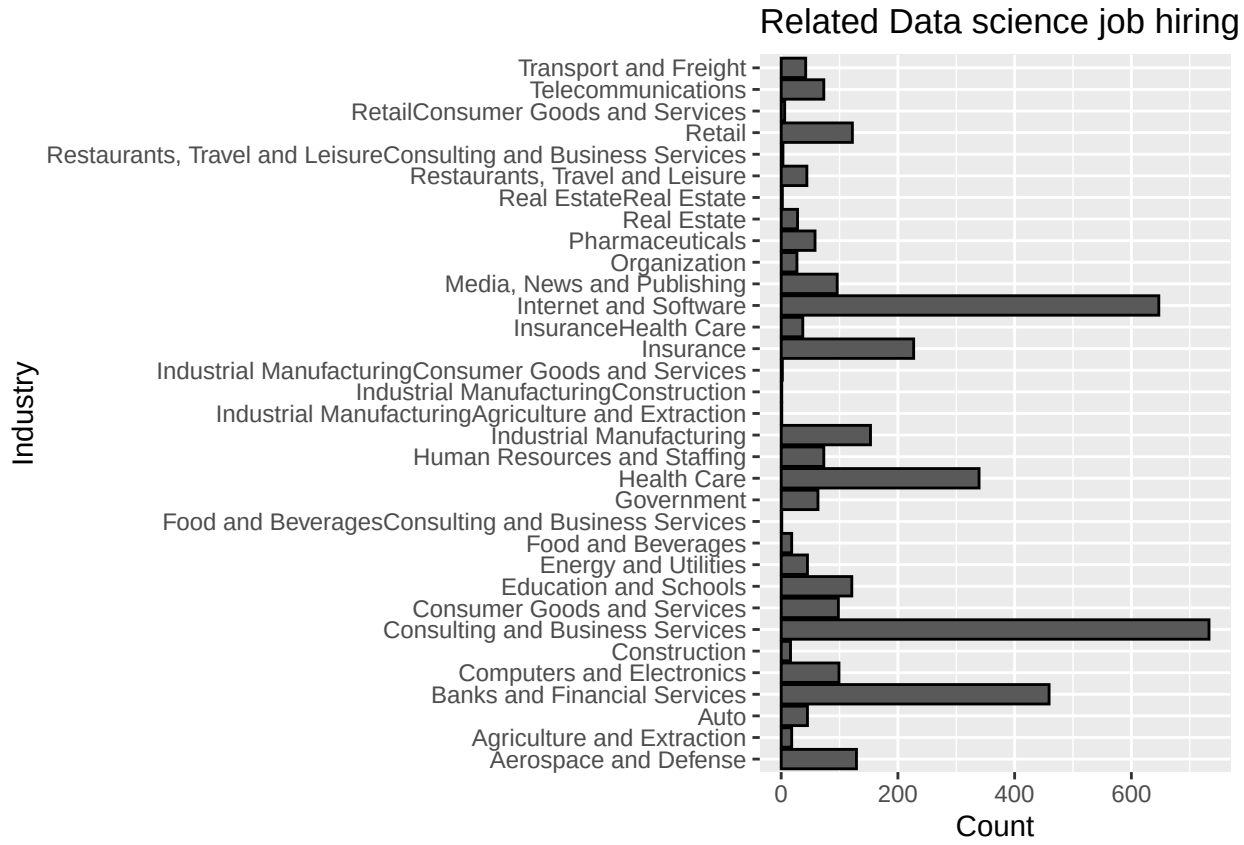
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   3.000   3.125   5.000   9.000
```

```
skills <- data.frame(value = apply(skilltable,2,sum))
skills$type = rownames(skills)


ggplot(data=skills, aes(x=value, y=reorder(type,value), fill=type))+
  geom_bar(colour="black", stat="identity") +
  ggtitle("Skills requirements") +
  xlab("Count") + ylab("Skill")
```

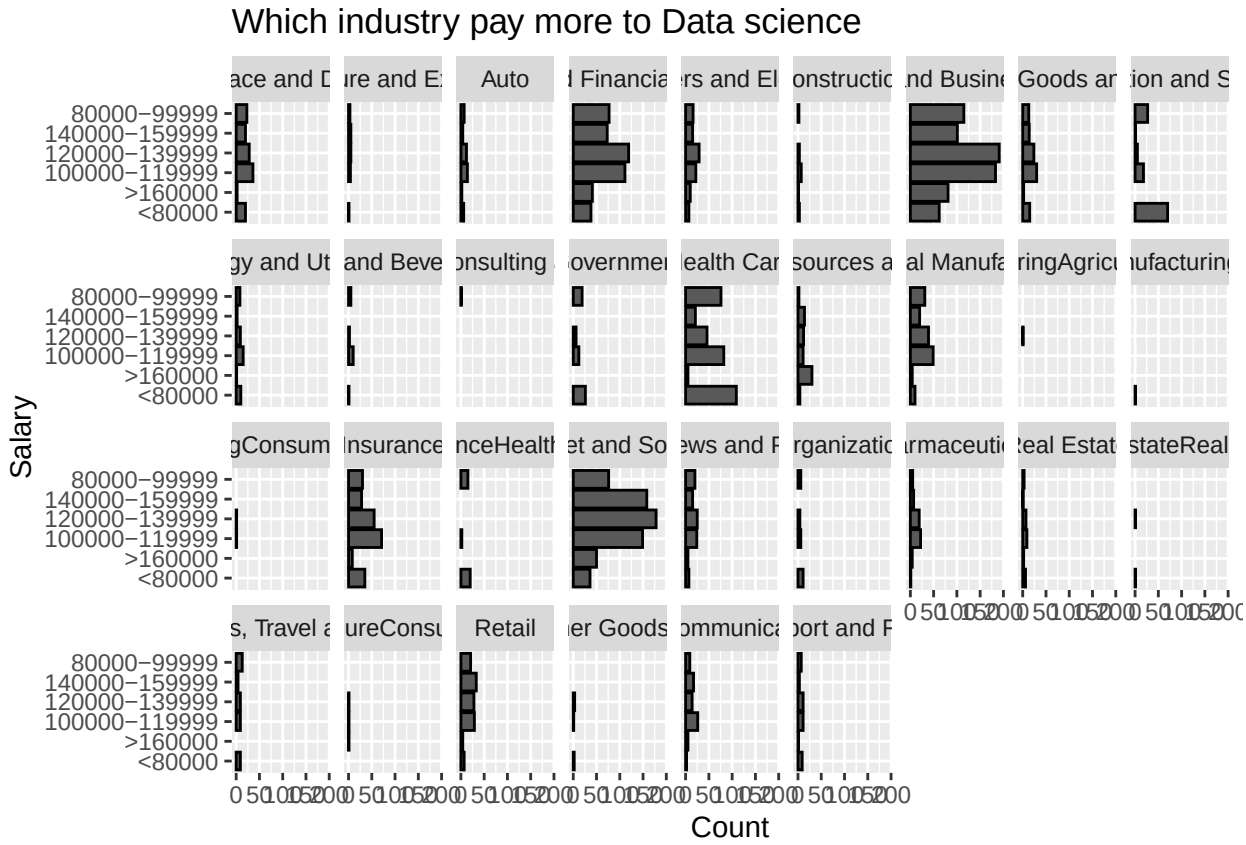


From the Related Data science job hiring by Industry chart, we can see the industry with high demand in data science positions: Consulting and Business Service, Internet and Software, Bank & Financial Service, and Health Care. While High competitive in Consulting and Business Service, Internet and Software, and Bank & Financial Service with the highest pay, Health Care and Education industries seem a better option for an entry-level position. In addition and not surprised, the Data science position is for high revenue company.

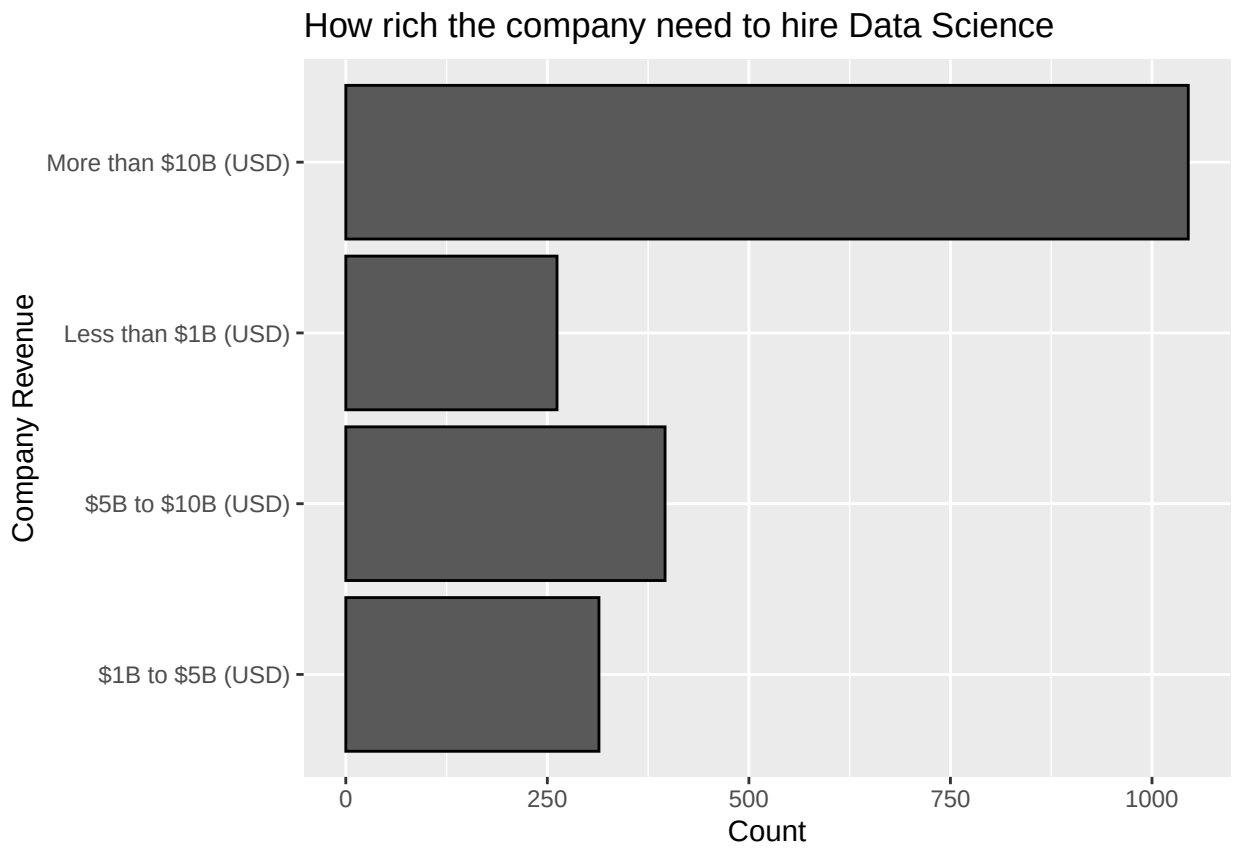
```
# What industry is hiring data science
industry_count <- indeeddata %>%
  select("Company_Industry")

industry_count <- filter(industry_count,Company_Industry!="")

ggplot(data=industry_count, aes(y=Company_Industry))+
  geom_bar(colour="black") +
  ggtitle("Related Data science job hiring by Industry") +
  xlab("Count") + ylab("Industry")
```



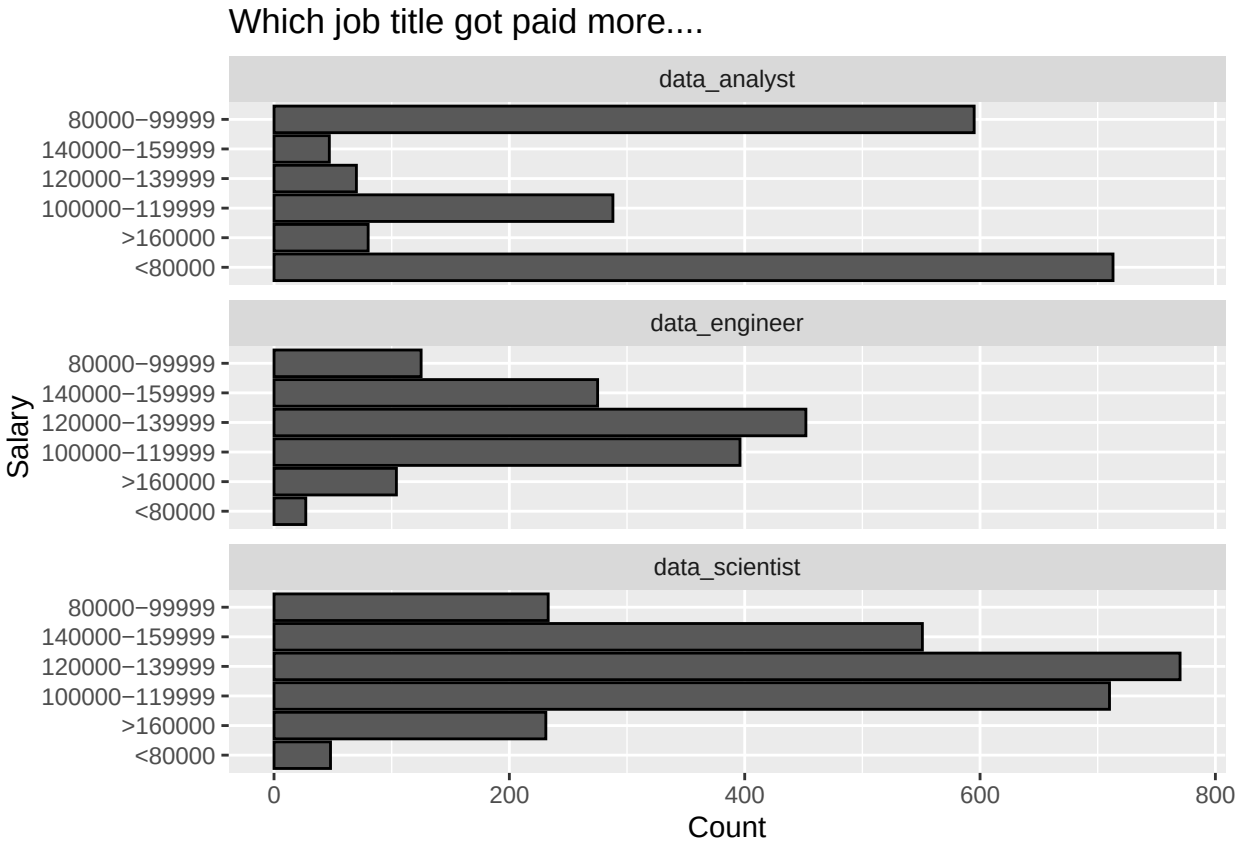

```
#industry with pay
industry_Salary <- indeeddata %>%
  select("Company_Industry","Queried_Salary")

industry_Salary <- filter(industry_Salary,Company_Industry!="")
industry_Salary <-filter(industry_Salary,Queried_Salary!="")

ggplot(data=industry_Salary, aes(y=Queried_Salary))+
  geom_bar(colour="black") +
  ggtitle("Which industry pay more to Data science") +
  xlab("Count") + ylab("Salary") +
  facet_wrap(~Company_Industry, nrow =4)
```




```
# company revunue
industry_revenue <- indeeddata %>%
  select("Company_Revenue")

industry_revenue <- filter(industry_revenue,Company_Revenue!="")

ggplot(data=industry_revenue, aes(y=Company_Revenue))+
  geom_bar(colour="black") +
  ggtitle("How rich the company need to hire Data Science") +
  xlab("Count") + ylab("Company Revenue")
```




```
ggplot(data=jobtable, aes(y=Queried_Salary))+
  geom_bar(colour="black") +
  ggtitle("Which job title got paid more....") +
  xlab("Count") + ylab("Salary") +
  facet_wrap(~Job_Type, nrow =4)
```



My approach: After examining the data many times I had ultimately decided to compare salary ranges of the "Data Scientist" job-type within the data since there were so many entries that consisted of three job types which were data engineer, data analyst and data scientists. I decided upon narrowing the category to data scientists since it was relevant to the question of "which skills are important for data scientists to Have". I first broke down the data scientists job role into 3 salaries ranges which were less than 80K (low-range), Mid-Range and High-End. For each category I broke each job posting to looking at the number of programming language (# of skills as it is called in the data) one should know for each job posting and then I grouped each posting under a industry. This approach was used to help understand how many skills a data scientist should know in each industry.

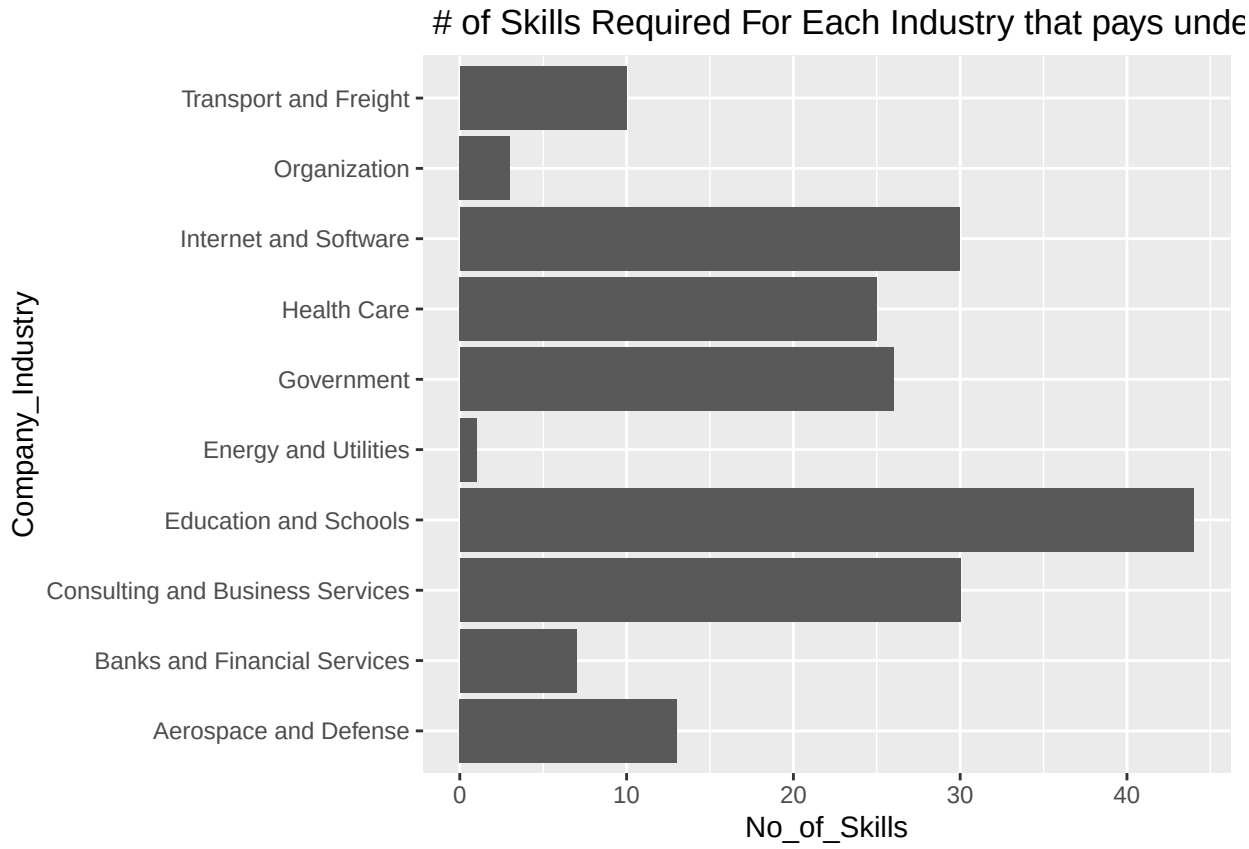
Understanding the data

To better understand the graph, it had added the numbers of skills from each job-posting and added each # of skills its relevant industry. So for example if a job posting under govt industry required u know to 2 programing languages and another govt industry job posting required you to know 3 programming languages then the number of skills under govt would be 5.I used this approach to better understand how many programming "skills" are required for each industry and for which specific salary range.

Low-Range (Less than 80K)

see what job-type are assigned in this range and skills that are desired here I first compared salaries that are less than 80k and I filtered out the jobs that don't require any skills and cleaned out jobs postings that aren't in any industry.

```
## I filtered out the table to less than 80K range salary and I omitted the empty data sets.
LEssthan80k <-jobtable %>%
  filter(Job_Type=="data_scientist",Queried_Salary =="<80000",No_of_Skills != "0",Company_Industry !="")
  group_by(Company_Industry)
## Here I graphed the data to better understand it.. .
ggplot(LEssthan80k,aes(x=Company_Industry,y=No_of_Skills)) +
  geom_bar(stat="identity") +
  coord_flip() + labs(
    title= " # of Skills Required For Each Industry that pays under 80k"
  )
```



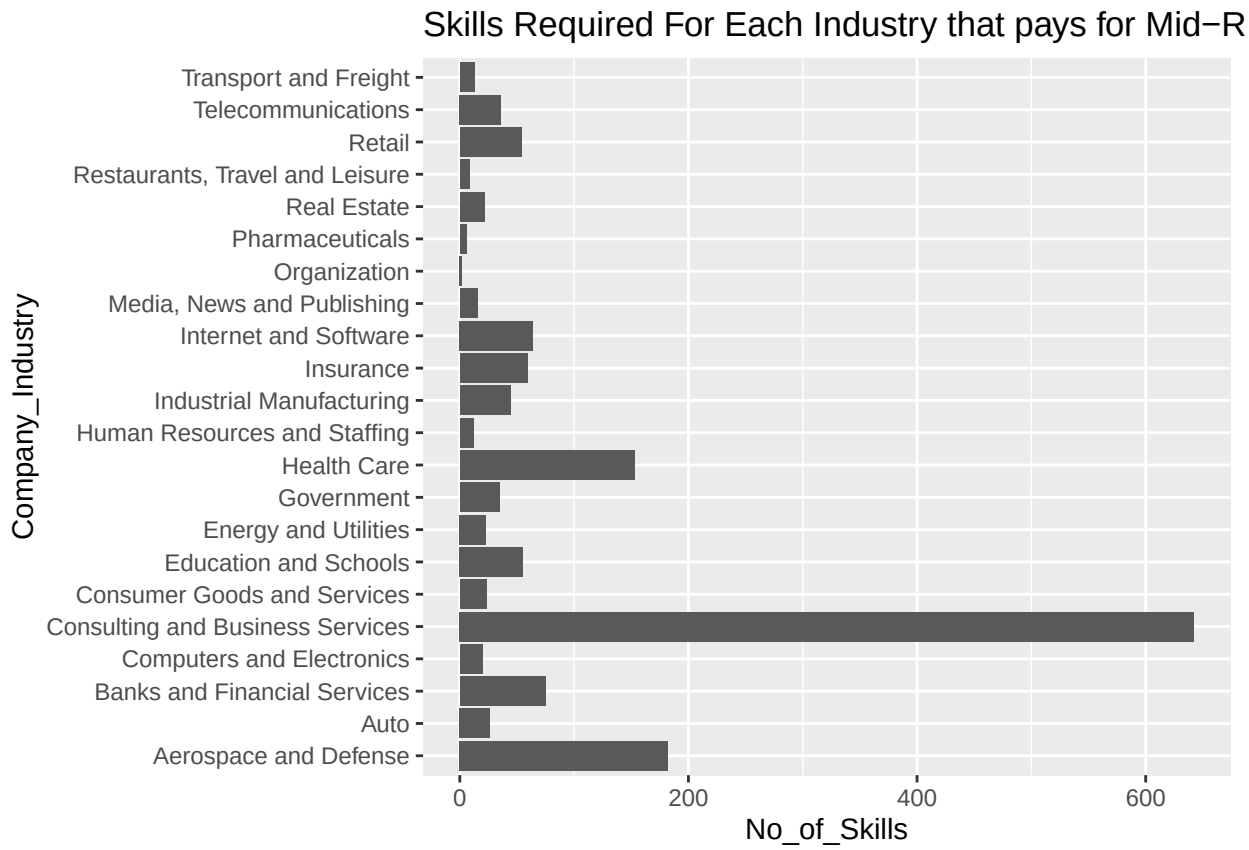
Industries like Education and Schools seems to demand the most out of data scientists while local organizations demand the least amount of skills. This make sense since the education sector has a lot of data regarding students'scores,location,race, and etc and they need all the best qualified candidates they can get to make sense of the data and to help students.Generally schools and education industry are usually underfunded from the city and thus it makes sense that they would pay the least amount of salary.

Low-Mid-Range (80-99K)

For Mid-Range I had split the salary ranges of 80-100K,100K-119K and 120K-139K into 3 bar graphs which I classified as mid-range. To fully view the data just click on the show in new windows to see the whole bar chart.

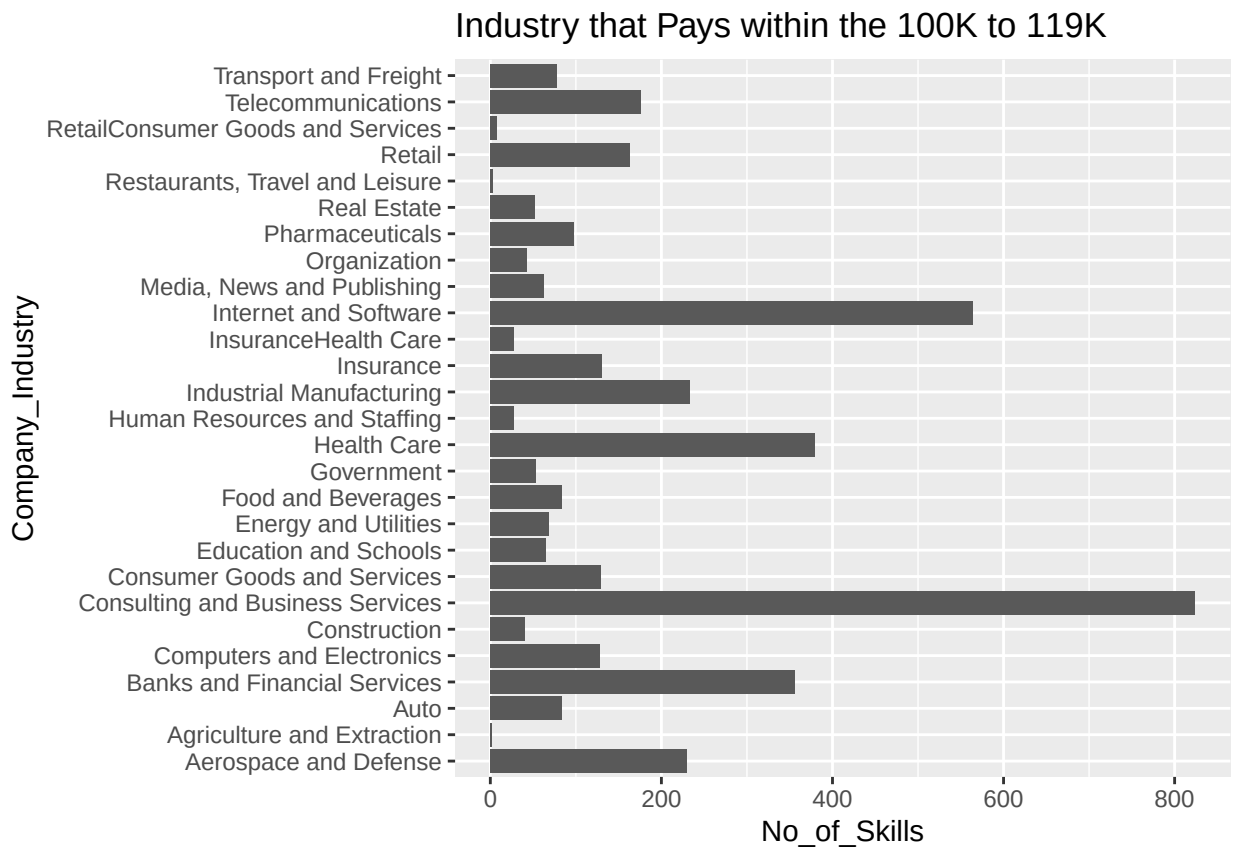
```
## Here I filtered the salary by mid-range which is:
Midrange_1 <-jobtable %>%
```

```
  filter(Job_Type=="data_scientist",Queried_Salary == "80000-99999",No_of_Skills != "0",Company_Industry
  group_by(Company_Industry)
ggplot(Midrange_1,aes(x=Company_Industry,y=No_of_Skills)) +
  geom_bar(stat="identity") +
  coord_flip() +
   labs(
    title= "Skills Required For Each Industry that pays for Mid-Range-Industry"
  )
```



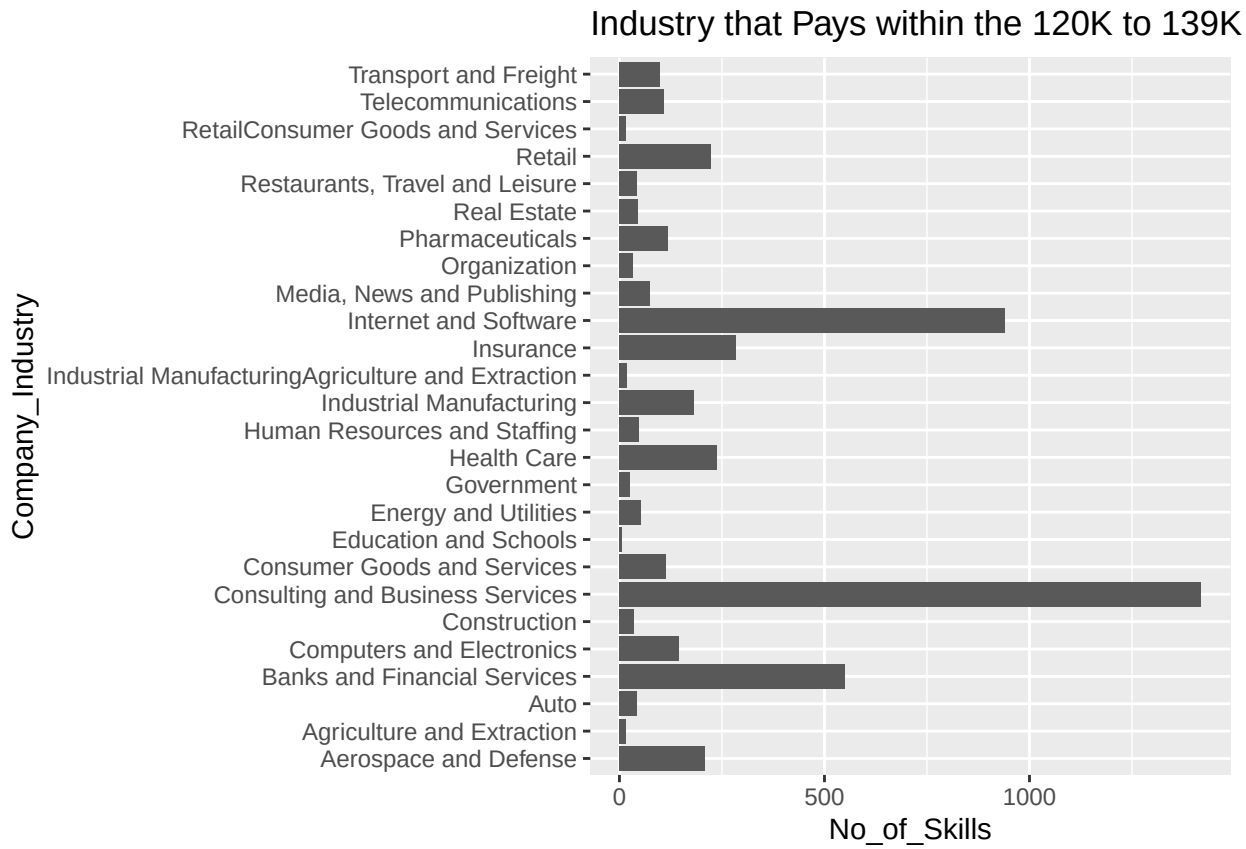
Mid-Range 100K to 119K:

```
Midrange2 <-jobtable %>%
  filter(Job_Type=="data_scientist",Queried_Salary == "100000-119999",No_of_Skills != "0",Company_Indus
  group_by(Company_Industry)
ggplot(Midrange2,aes(x=Company_Industry,y=No_of_Skills)) +
  geom_bar(stat="identity") +
  coord_flip() + labs(
    title= "Industry that Pays within the 100K to 119K"
  )
```



High Mid-Range (120-139K)

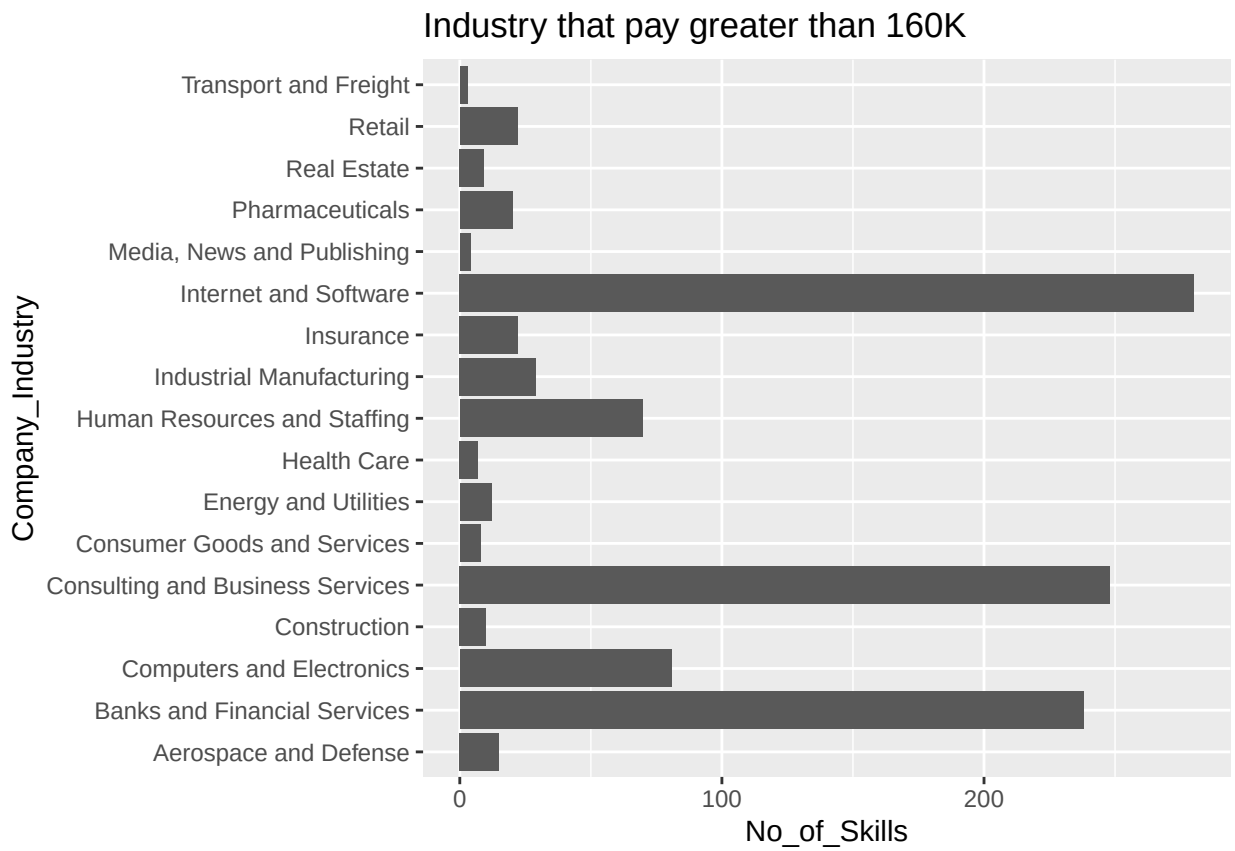
```
## Here I filtered it by 120-139K range
Midrange3 <-jobtable %>%
  filter(Job_Type=="data_scientist",Queried_Salary == "120000-139999",No_of_Skills != "0",Company_Indus
  group_by(Company_Industry)
ggplot(Midrange3,aes(x=Company_Industry,y=No_of_Skills)) +
  geom_bar(stat="identity") +
  coord_flip() + labs(
    title= "Industry that Pays within the 120K to 139K"
  )
```



Interpreting each salary range in this mid-range it seems that the industry that demands the most programming skills for the "mid-range" salary is the consulting and business services industry.

High-Range:

```
HighRange <-jobtable %>%
  filter(Job_Type=="data_scientist",Queried_Salary ==">160000",No_of_Skills != "0",Company_Industry !="
  group_by(Company_Industry)
ggplot(HighRange,aes(x=Company_Industry,y=No_of_Skills)) +
  geom_bar(stat="identity") +
  coord_flip() +
  labs(
    title = "Industry that pay greater than 160K"
  )
```



In the high-range Salary range which is greater than 160K it seems the top industry to pay and demand the greatest skills to work in these industry is the Internet and Software industry followed by consulting and business services and banks and financial services. Makes sense since tech industries like Google,Facebook and other tech companies are able to hire more workers and earn a lot of money to be able to pay their employees in this high salary range.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.