indeedoor:
a better way to find a job

Team Tufte

Jade Bailey-Assam

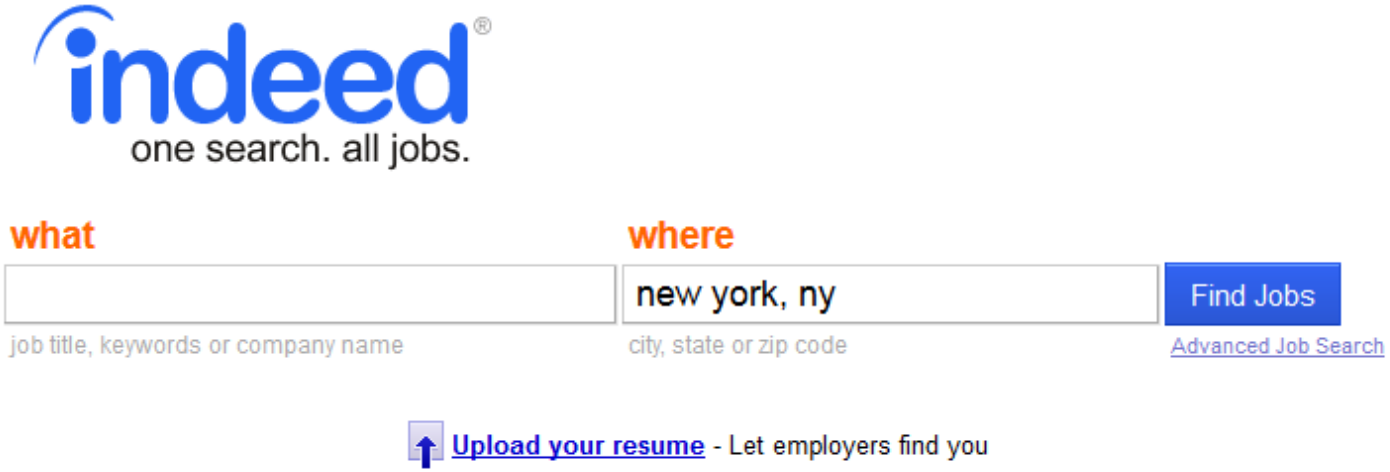
Lucy Drotning

Christine Lee

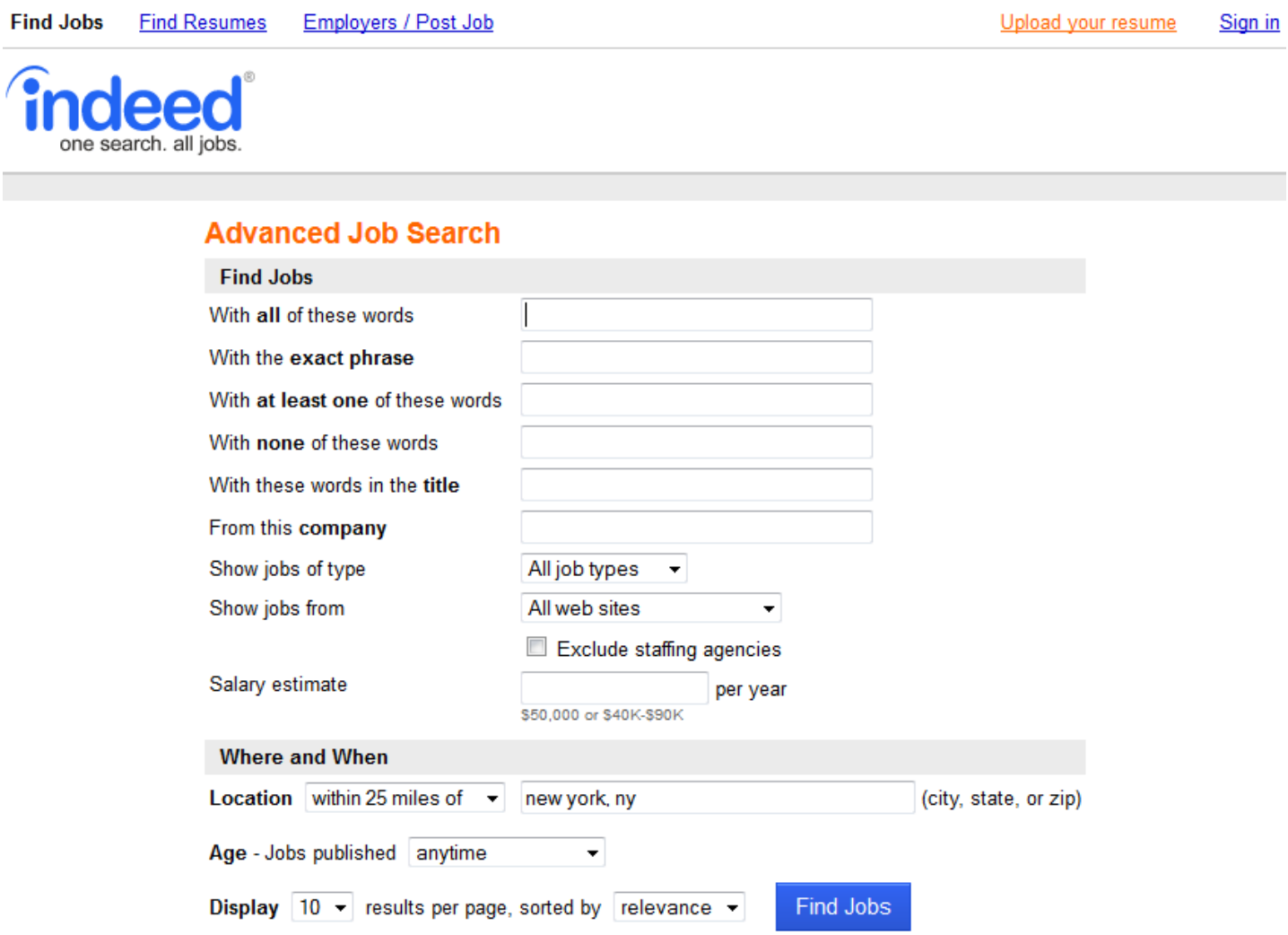
Shruti Pandey

Janet Prumachuk

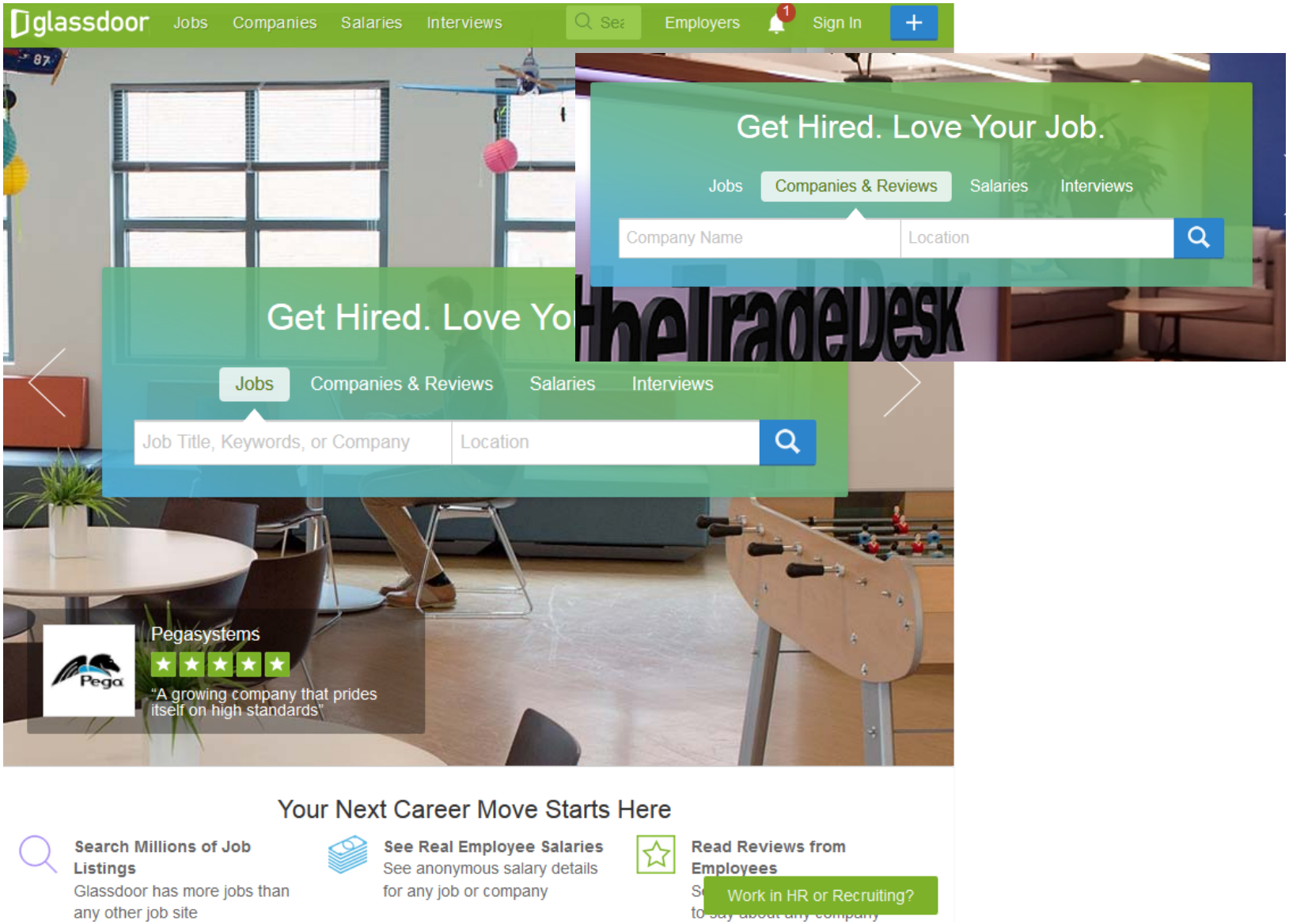
Motivation


indeed
one search. all jobs.
what
where
new york, ny
Find Jobs
job title, keywords or company name
city, state or zip code
Advanced Job Search
Upload your resume - Let employers find you

Motivation


Find Jobs
Find Resumes
Employers / Post Job
Upload your resume
Sign in
indeed
one search. all jobs.
Advanced Job Search
Find Jobs
With all of these words
With the exact phrase
With at least one of these words
With none of these words
With these words in the title
From this company
Show jobs of type
All job types
Show jobs from
All web sites
Exclude staffing agencies
Salary estimate
per year
$50,000 or $40K-$90K
Where and When
Location
within 25 miles of
new york, ny
(city, state, or zip)
Age - Jobs published
anytime
Display
10
results per page, sorted by
relevance
Find Jobs

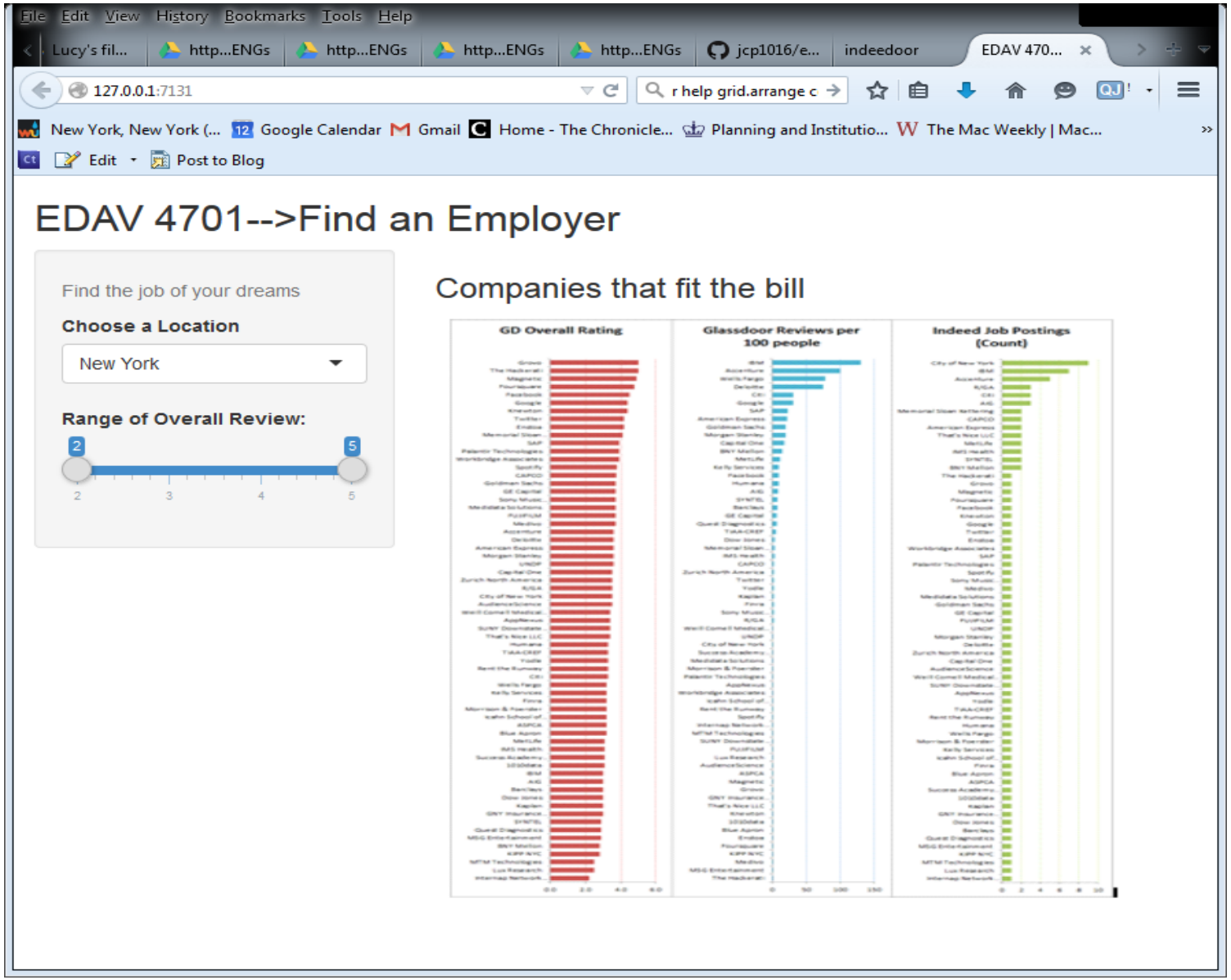
Motivation


glassdoor
Jobs
Companies
Salaries
Interviews
Employers
Sign In
Get Hired. Love Your Job.
Jobs
Companies & Reviews
Salaries
Interviews
Company Name
Location
Get Hired. Love Yo
Jobs
Companies & Reviews
Salaries
Interviews
Job Title, Keywords, or Company
Location
Pegasystems
"A growing company that prides itself on high standards"
Your Next Career Move Starts Here
Search Millions of Job Listings
Glassdoor has more jobs than any other job site
See Real Employee Salaries
See anonymous salary details for any job or company
Read Reviews from Employees
Work in HR or Recruiting?

Our contribution

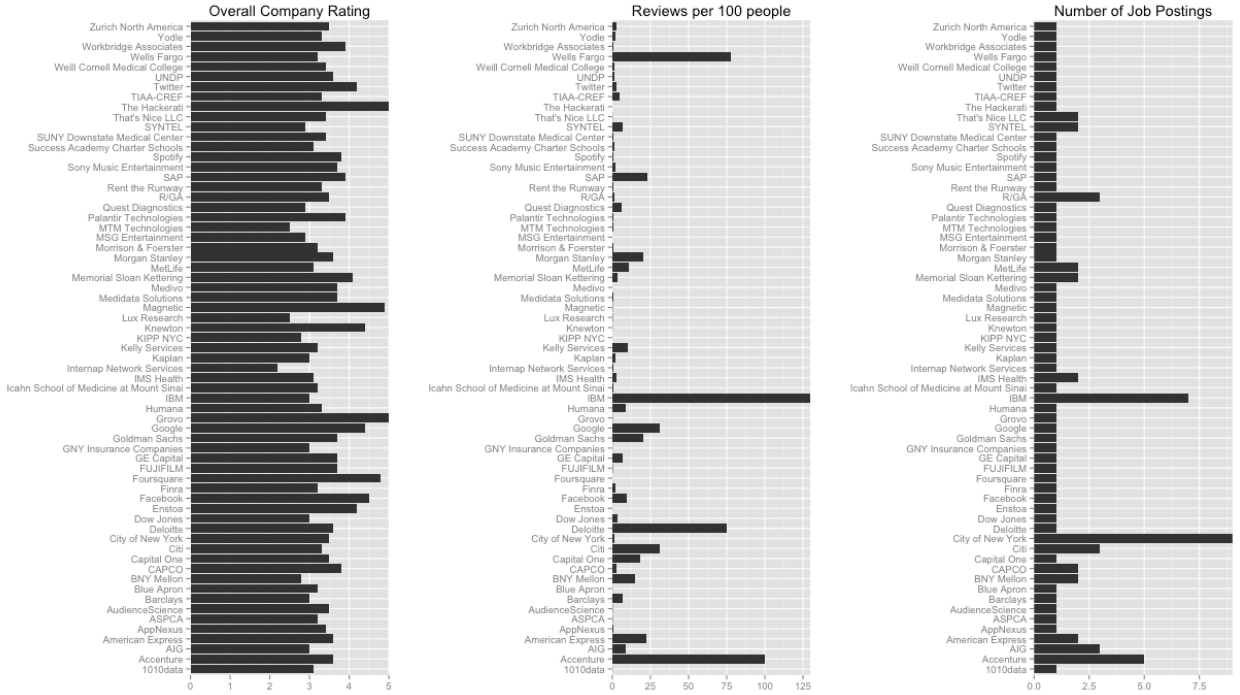
- Both sites contain valuable, under-leveraged data
- Our goal is to give the job seeker more information as part of their job search process:
  - How are the companies rated by industry, overall and by category (Janet, Lucy)
    - Learn more about a particular set of companies (based on some user inputs), including number of reviews, number of positions open…
  - Where are the jobs located (Christine)
  - Differences in jobs by region & regression (Shruti)
  - Leverage job postings to tailor your resume (Jade)

Starting medium

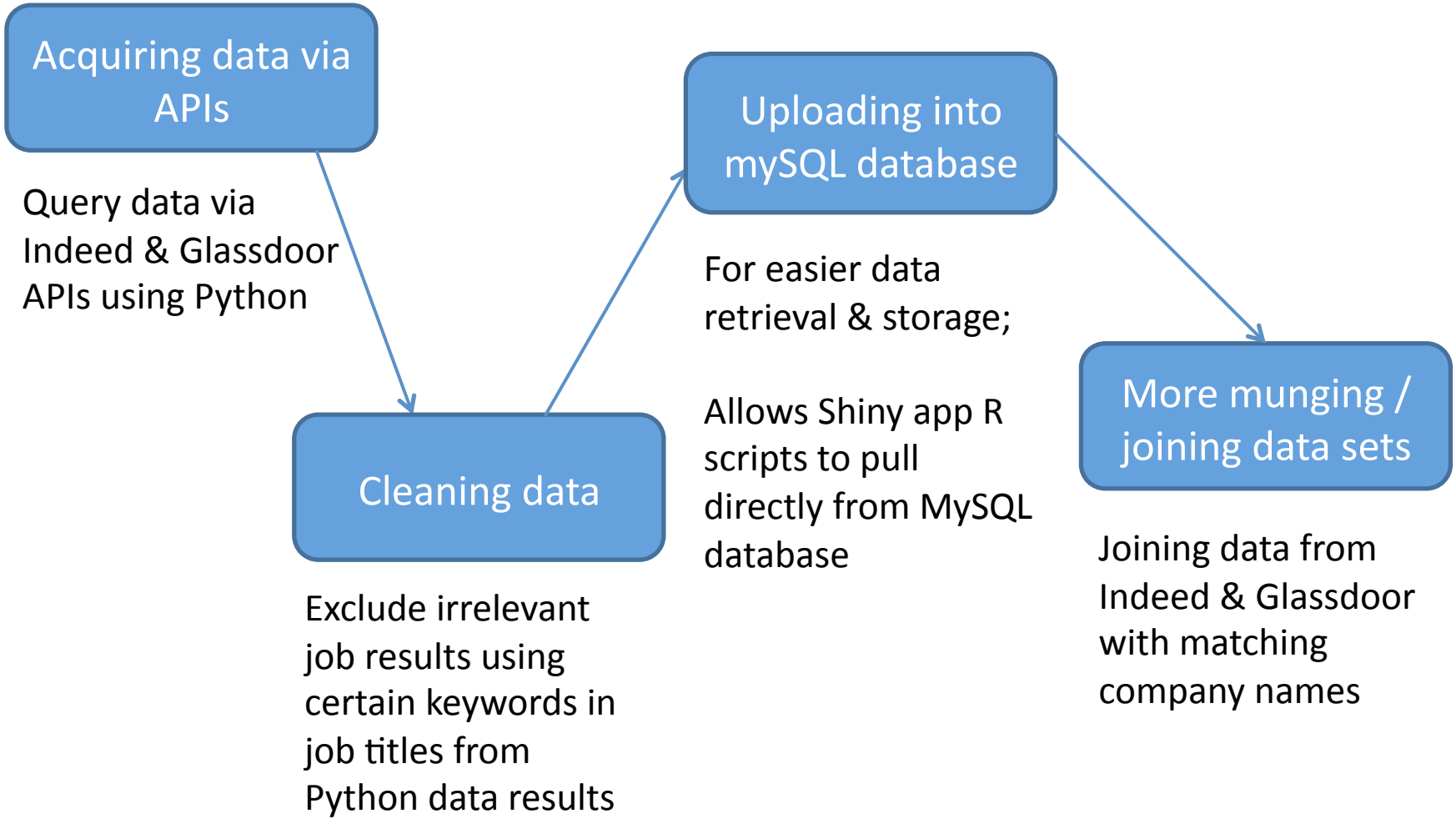


R Code to generate comparison charts (all black bars – very New York data scientist)

```
dscomprev <- read.csv("Workbook1.csv")

library(ggplot2)
library(gridExtra)

x <-ggplot(dscomprev, aes(x=Company.Name, y=Indeed.Postings)) +
                geom_bar(stat="identity") +
                coord_flip() +
                ggtitle("Number of Job Postings") +
                theme(axis.title.x = element_blank(),axis.title.y = element_blank())
x1 <- x + scale_y_continuous(expand = c(0, 0))

y <- ggplot(dscomprev, aes(x=Com
                geom_bar(stat="ide
                coord_flip() +
                ggtitle("Reviews pe
                theme(axis.title.x =
y1 <- y + scale_y_continuous(expa

z <-ggplot(dscomprev, aes(x=Comp
                geom_bar(stat="ide
                coord_flip() +
                ggtitle("Overall Co
                theme(axis.title.x =
z1 <- z + scale_y_continuous(expar

grid.arrange(z1, y1, x1, ncol=3)
```



Our data pipeline



Data acquisition & preparation

Indeed.com jobs data

Scope:

- Jobs with query term “data + scientist”
- Job results within 40 miles or less of 12 cities in the U.S.

Challenges in cleaning data:

- Created conditions in Python script that filtered out irrelevant job titles
- But positions that have term “data scientist” in description still show up
- Meanwhile, we also cannot simply eliminate seemingly “irrelevant” job titles as some of these jobs are actually data science positions with a different name

Glassdoor company data

Scope:

- Company Name
- Location, Ratings

Challenges in cleaning data:

- Minimal data cleaning required
- Ignored rating text fields

The API call only shows us the job snippet, which isn't enough to mine. To get enough data for a POC, we'll extract the URL from the indeed job posting, scrape this text and mine it.

This XML file does not appear to have any style information associated with it. The document tree is shown below.

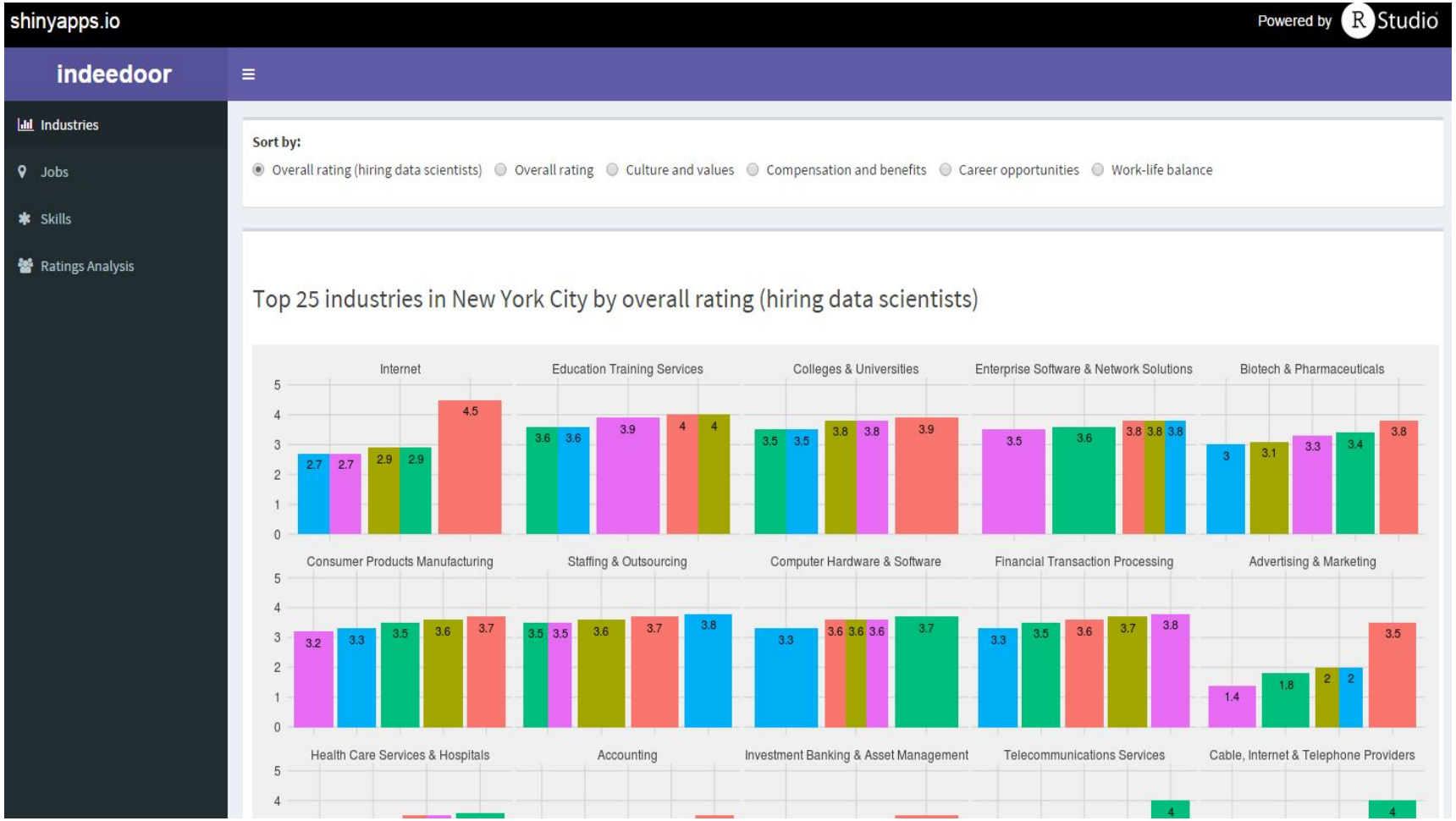

```
▼<response version="2">
  <query>data scientist</query>
  <location>new, york</location>
  <dupefilter>true</dupefilter>
  <highlight>false</highlight>
  <totalresults>1147</totalresults>
  <start>1</start>
  <end>10</end>
  <pageNumber>0</pageNumber>
 ▼<results>
  ▼<result>
    <jobtitle>Data Scientist</jobtitle>
    <company>The College Board</company>
    <city>New York</city>
    <state>NY</state>
    <country>US</country>
    <formattedLocation>New York, NY</formattedLocation>
    <source>The College Board</source>
    <date>Mon, 27 Apr 2015 15:55:37 GMT</date>
   ▼<snippet>
      Experience with descriptive statistics and data visualization in R and Tableau. Create insights from existing data, and drive the collection of new data through...
    </snippet>
   ▼<url>
      http://www.indeed.com/viewjob?jk=aa482c194e06d589&qd=P5t9H8wdm9_p3y0mSnUcn3G_LeWVwgh1sg0a_coxuzer1EAbV7HkKpskzdWnzVRwCNFapReE6dZrtlp5SIe6ou9BeklLMgTs8p5EmYWwc5ULB2pHYt-
      YP2NYagtBg9Lp&indpubnum=4751269202013823&atk=19k10ebv75ucse6f
    </url>
    <onmousedown>indeed_clk(this, '4983');</onmousedown>
    <latitude>40.71154</latitude>
    <longitude>-74.00549</longitude>
    <jobkey>aa482c194e06d589</jobkey>
    <sponsored>false</sponsored>
    <expired>false</expired>
    <indeedApply>false</indeedApply>
    <formattedLocationFull>New York, NY</formattedLocationFull>
    <formattedRelativeTime>1 day ago</formattedRelativeTime>
  </result>
```


Trim URL, add to list to be scraped.

API call gave 10 URLs to work with for POC

URL from API to grab for full job description

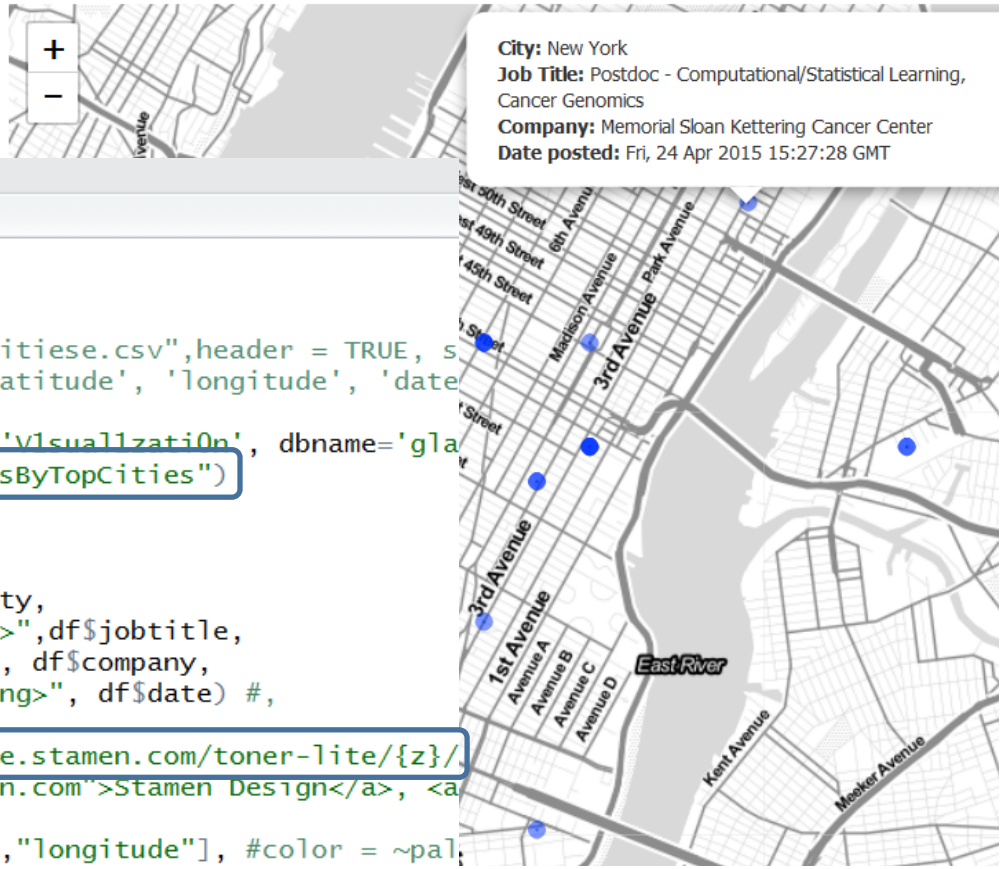
The indeedoor app



Shiny code highlight – jobs map app

Dots on the map represent job postings and their locations from a search Indeed using query term "data scientist":

Please click on job location circle for more information for each job posting

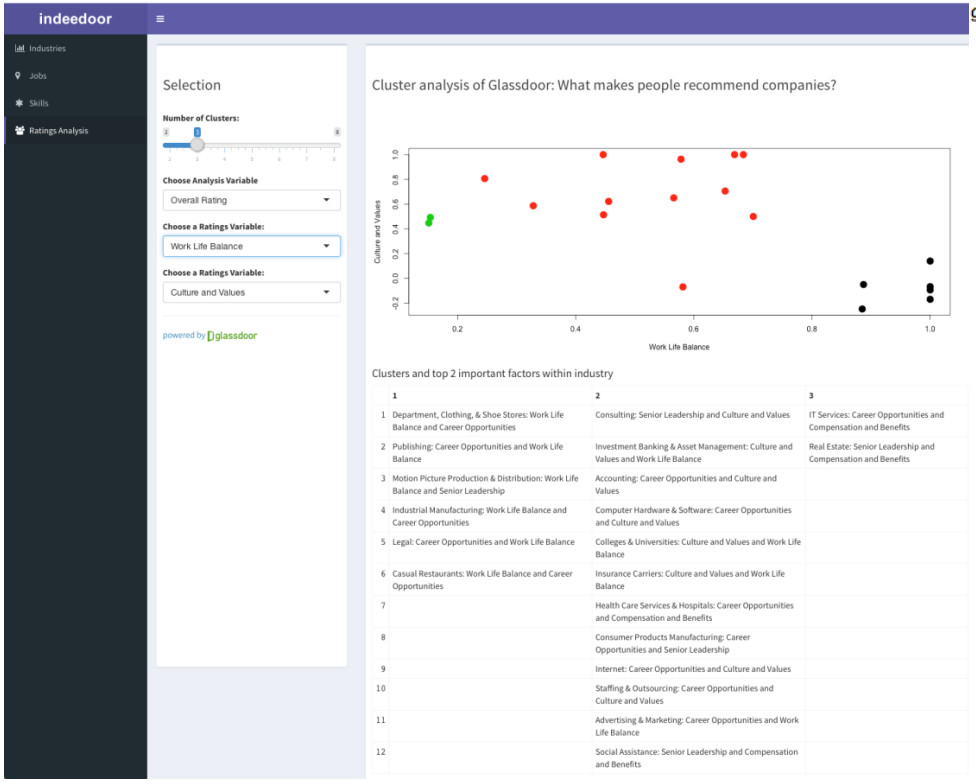


eed_leaflet_penciltile.R × server.r* × ui.r ×

```
shinyServer(function(input, output) {
  library(rgdal)
#indeeddata<- read.csv("./data/Newindeed_datasci_topcitiese.csv",header = TRUE, s
#            col.names=c('city','state','snippet', 'latitude', 'longitude', 'date

mydb = dbConnect(MySQL(), user='STATW4701', password='V1sual1zati0n', dbname='gla
rs <- dbSendQuery(mydb, "select * from IndeedAnalyticsByTopCities")
indeeddata <- fetch(rs, n=-1)
df<-data.frame(indeeddata)

scrape_pop <- paste0("<strong>City: </strong>", df$city,
                     "<br><strong>Job Title: </strong>",df$jobtitle,
                     "<br><strong>Company: </strong>", df$company,
                     "<br><strong>Date posted: </strong>", df$date) #,

m2 <- leaflet(data = df) %>% addTiles('http://{s}.tile.stamen.com/toner-lite/{z}/
    attribution = 'Map tiles by <a href="http://stamen.com">Stamen Design</a>, <a
    setView(-73.9983273, 40.7471983, zoom = 12) %>%
    addCircles(lat = ~ df[ ,"latitude"], lng = ~ df[ ,"longitude"], #color = ~pal,
               fillOpacity = .5,
               radius = 8,
               popup= scrape_pop,
               weight = 9)

output$ClistMap = renderLeaflet(m2)

})
```

```
library(leaflet)
library(ggplot2)
library(maps)
library(RMySQL)
```

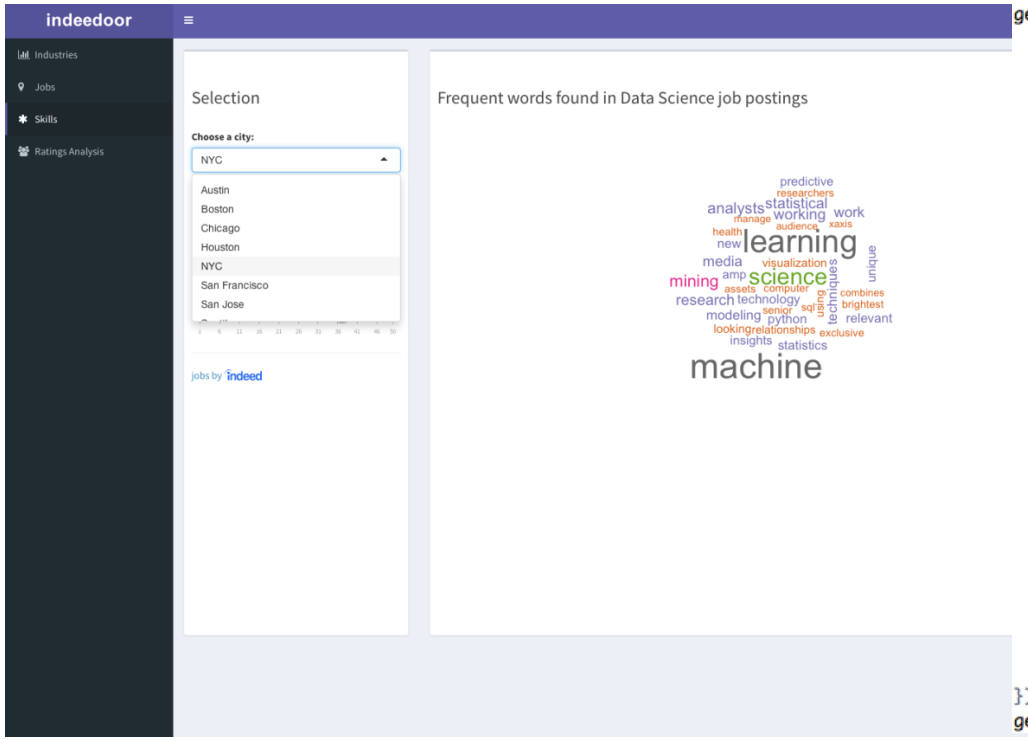
Regression: Screenshot and Code Snippet



	1	2	3
1	Department, Clothing, & Shoe Stores: Work Life Balance and Career Opportunities	Consulting: Senior Leadership and Culture and Values	IT Services: Career Opportunities and Compensation and Benefits
2	Publishing: Career Opportunities and Work Life Balance	Investment Banking & Asset Management: Culture and Values and Work Life Balance	Real Estate: Senior Leadership and Compensation and Benefits
3	Motion Picture Production & Distribution: Work Life Balance and Senior Leadership	Accounting: Career Opportunities and Culture and Values	
4	Industrial Manufacturing: Work Life Balance and Career Opportunities	Computer Hardware & Software: Career Opportunities and Culture and Values	
5	Legal: Career Opportunities and Work Life Balance	Colleges & Universities: Culture and Values and Work Life Balance	
6	Casual Restaurants: Work Life Balance and Career Opportunities	Insurance Carriers: Culture and Values and Work Life Balance	
7		Health Care Services & Hospitals: Career Opportunities and Compensation and Benefits	
8		Consumer Products Manufacturing: Career Opportunities and Senior Leadership	
9		Internet: Career Opportunities and Culture and Values	
10		Staffing & Outsourcing: Career Opportunities and Culture and Values	
11		Advertising & Marketing: Career Opportunities and Work Life Balance	
12		Social Assistance: Senior Leadership and Compensation and Benefits	

```
getRegressionAnalysis <- memoise(function(variable1, variable2, k, clustertype){
        #con <- getConnection()
        #rs = dbSendQuery(mydb, "select * from CompanyRatings")
        #data = fetch(rs, n=-1)
        data <- master_gd_data  ## already have this in a data frame
        print(variable1)
        print(variable2)
        data <- subset(data, data$employers.industry != "")
        industries <-unique(data$employers.industry)
        largedata<-subset(data, data$employers.numberOfRatings > 40)
        clusteringdata<-array()
        nameofindustry<-vector()
        mostimportant<-array()
        j <- 0
        for (i in industries)
        {
                fdata <- subset(largedata, data$employers.industry == i)
                if (nrow(fdata) > 40)
                {
                        print (paste0("Industry is ", i, " and the number of comp
                        id <- fdata$employers.id
                        name <- fdata$employers.name
                        website <- fdata$employers.website
                        industry <- fdata$employers.industry

                        overallRating <- fdata$employers.overallRating
                        ratingDescription <- fdata$employers.ratingDescription
                        cultureAndValuesRating <- fdata$employers.cultureAndValue
                        seniorLeadershipRating <- fdata$employers.seniorLeadershi
                        compensationAndBenefitsRating <- fdata$employers.compensa
                        careerOpportunitiesRating <- fdata$employers.careerOpport
                        workLifeBalanceRating <- fdata$employers.workLifeBalanceR
                        recommendToFriendRating <- fdata$employers.recommendToFri
                        pctApprove <- fdata$employers.ceo.pctApprove

                        employer <- data.frame(id, name, website, industry, overa
```

Skills: Screenshot and Code Snippet




```r
getTermMatrix <- memoise(function(dbCity) {
        ##con <- getConnection()  ## this was giving errors and corrupting the co
        rs = dbSendQuery(mydb, paste0("select * from IndeedAnalyticsByTopCities w
        jobData = fetch(rs, n=-1)
        #dbDisconnect(con)

        # Read in data as dataFrame
        # reading only the text snippets
        x<-data.frame(v1=jobData$snippet)
        x<-subset(x,x$v1 != "")
        docs<- Corpus(DataframeSource(x))

        # now for some cleanups
        docs <- tm_map(docs, content_transformer(tolower))
        docs <- tm_map(docs, removeNumbers)
        docs <- tm_map(docs, removePunctuation)
        docs <- tm_map(docs, removeWords, stopwords("english"))
        docs <- tm_map(docs, removeWords, c("experience", "will", "data", "analyt
        docs <- tm_map(docs, stripWhitespace)
        #  docs <- tm_map(docs, stemDocument)

        # create the Document Term Matrix and compute the frequencies
        dtm <- DocumentTermMatrix(docs)
        m <- as.matrix(dtm)
        sort(colSums(m), decreasing = TRUE)
        #freqs <- colSums(as.matrix(dtm))

})
getVariables <-(function(choice)
```

Text mining code

-Try using rvest() to scrape URLs from R. This is not successful, use import.io to scrape. Merge and organize in old, faithful friend Excel.

-Use text mining tm() package in R to compile stop words and do word stemming. Code below

filterlistnew <- removeWords(filterlist, stopwords, c ("data", "hour", "federal", "within", "hagan", "yeeldr","bank","get","we're","ricci","top", "ability","like","also","skill","working", "job","trend", "new","york“, "ny", "company", "days", "work", "ago", "one", "opportunity", "experience", "scientist","education","city","program","state","use","capital","apply","etc","contact","equal","public","find","ll","employment","post","action","title","resume","inc","save","times","well","forum","require","well","terms","cookie","smart","upload","protect","resumesemployer","2015", "search","keyword","high","responsibility","zip","jobsfind","indeed","privacy","help","sign","indeed.com","affirmative","veteran","review","policy","question",”email","world","disability","including","friend","demonstrate","student","candidate","us","real","knewton","include","view","price","employee","background","enstoa","benefit","status","environment"))

filterlistnew <- wordStem(c ("machine", "learning", 'machine learning',"statistic", "statistics", 'statistics',"communicate","communication",'communication'))

Leverage job postings to tailor your resume

These ten "specific skills" that might be helpful to have on your resume

Ex: Proficiency in *Python, SQL, Hadoop, Tableau, MapReduce*

```
## Source: local data frame [10 x 3]
##
##          Word Frequency           Group
## 1      python        11 specific skills
## 2         sql         8 specific skills
## 3    language         7 specific skills
## 4      hadoop         7 specific skills
## 5  production         7 specific skills
## 6   technique         7 specific skills
## 7        java         6 specific skills
## 8   mapreduce         5 specific skills
## 9     cluster         4 specific skills
## 10    tableau         4 specific skills
```

These six "coursework" words that might be helpful to include on your resume

Ex: Relevant *Engineering* coursework in: *statistics, mathematics, machine learning, algorithms, computer science*

```
## Source: local data frame [6 x 3]
##
##               Word Frequency               Group
## 1       statistics        18 relevant coursework
## 2      mathematics        17 relevant coursework
## 3 machine learning        15 relevant coursework
## 4        algorithm        13 relevant coursework
## 5         engineer        12 relevant coursework
## 6 computer science         6 relevant coursework
```

Continued….
Rest of groups published on RPubs

These are the top "buzzwords" that might be helpful to have on your resume

Ex: *Identify insights* to *benchmark digital knowledge* through use of *big data and advanced analytics*

```
## Source: local data frame [16 x 3]
##
##            Word Frequency     Group
## 1     analytics        16 buzzwords
## 2      advanced        11 buzzwords
## 3         build         9 buzzwords
## 4     knowledge         9 buzzwords
## 5       digital         8 buzzwords
## 6  quantitative         7 buzzwords
## 7      research         6 buzzwords
## 8      big data         6 buzzwords
## 9   investigate         5 buzzwords
## 10      insight         5 buzzwords
## 11       expert         5 buzzwords
## 12     identify         5 buzzwords
## 13      predict         5 buzzwords
## 14       impact         4 buzzwords
## 15    benchmark         4 buzzwords
## 16   understand         4 buzzwords
```

Challenges and Next Steps

- Disparate, incomplete data
- Difficulty merging across jobs and companies datasets
  - Enhance company name matching
- Is Shiny the best tool?
- Generalize and scale application to support all types of jobs, not necessarily data science
- Add more data sources such as the BLS JOLT database for monthly job statistics
- Knock on the Glassdoor with our app....Indeed!

Where to find our work

- http://www.indeedoor.com
- https://github.com/jcp1016/edav-team-project/indeedoor
- http://rpubs.com/jadeemily/textmining
- http://rpubs.com/lucy/78463