Capstone Project Report

Deepali Bhatt

November 12, 2016

# Capstone Project Overview

This capstone project is about analysis into data scientist roles in India.This project was taken up by the need of a dear friend wanting to move to India in search of data scientist roles in India.When people want to move from overseas back to India, they do not do proper research into the job scene in regards with the role they are seeking a career in. Hundreds of job sites in India, with results running into hundreds of pages adds to the daunting taks of getting a holistic picture of a paricular job related market. People do not do enough research into, which companies are hiring for a data scientist position; what locations have the highest density of data scientists roles; what companies are hiring; what cities will likely hire in the future etc.? When the above questions are not researched properly, immigrants are usually stuck at locations that does not hire for the respective roles, or does not need data scientists.  
The aim of this project is to be able to make informed decisions about taking up data scientist roles and take a scientific approach towards it. To aid in this decision-making, I will be exploring data through descriptive statistics and visualisations making use of various statistical tools including but not limited to histograms, bar plots etc. For data visualisations, I will be using ggplot 2 to help understand the significance of data by placing it in visual context. **The goal of this project is to apply the principles of data science and develop insights from the analysis of structured or unstructured data to aid in better decision-making.**

The questions to be answered through this project are as follows:

1. Which states in India have the highest density of “Data Scientist” jobs?
2. Which cities in India have the highest density of “Data Scientist” jobs?
3. Which companies/consultants in India are hiring the most for the “Data Scientist” jobs?
4. What non-technical skills are required mostly for the “Data Scientist” jobs?
5. What technical skills are required mostly for the “Data Scientist” jobs?

# Overview of the Data Files

The data set was extracted using an open API for the website [Indeed] ([www.indeed.com](http://www.indeed.com)). There is one master csv that was used, which is as following: finalindeed1\_trynote.csv The data set is acquired by getting an open API from indeed (job search website), searching for data scientist jobs from all over India. The data set has 10 variables and has over 1200 search results. The data description is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Serial Number | Variable Name | Variable Description | Value Stored | Variable Type |
| 1 | Row Number | Unique Row Number | Unique value identifying the row | Identifier |
| 2 | Job Title | Name of the job role | 1-1104 | Categorical (qualitative variable) |
| 3 | Company | Company offering the job role | 1-1104 | Categorical (qualitative variable) |
| 4 | State | State in which company is located | 1-1104 | Categorical (qualitative variable) |
| 5 | City | City in which company is located | 1-1104 | Categorical (qualitative variable) |
| 6 | Country | India in which company is located | 1-1104 | Categorical (qualitative variable) |
| 7 | Formatted Location | Containing city and respective state in one cell | 1-1104 | Categorical (qualitative variable) |
| 8 | Source | Source through which, job role is extracted on the web | 1-1104 | Categorical (qualitative variable) |
| 9 | Date | Date the job role was posted | 1-1104 | Categorical (qualitative variable) |
| 10 | Snippet | Brief truncated description of the job role | 1-1104 | Categorical (qualitative variable) |
| 11 | Expired | Date when the job posting expires | True, False | Categorical (binary) |
| 12 | Python | Unique character identified as a technical skill extracted from snippet column | True, False | Categorical (binary) |
| 13 | Machine | Unique character identified as a technical skill extracted from snippet column | True, False | Categorical (binary) |
| 14 | Sql | Unique character identified as a technical skill extracted from snippet column | True, False | Categorical (binary) |

# Methodology

**Approach-Prelinimary Preparation**  
Since the above mentioned questions were related to India, I was looking for an Indian job site. After identifying Naukri (www.naurkri.com) as the job site and emailing them, we had to drop it since it did not have an open API.During research of open API's and asking various forums, we identified Indeed (www.indeed.com) has an open API which required me to open an account with them.  
After opening the account, a publisher key was identified which then had to be inserted in a URL given as an example in the Indeed website. Since Indeed has an Indian website (www.indeed.co.in) and the open API was for US (www.indeed.com) website, the output of the API with Indeed's Indian website was giving 0 results. The way I fixed it is by reading the API documentation and changing the parameters and extending the limit of 10 results to an extreme number. In the API link URL string, various parameters like co had to be changed to 'in' and limit had to changed to an extreme number to accommodate the full results. Next steps was to write a program to extract all the data available and put in a format like excel to be mined further.  
After the Indian job site results was saved in an XML site, I toyed with two codes to parse the XML data into the R as a dataframe. I used "xmlTreeParse" which gave a "NULL" and "NA" output. The second approach was to use "xmlToDataframe", which gave the error message as follows:

Error in [<-.data.frame(\*tmp\*, i, names(nodes[[i]]), value = c("Indian Council Of Medical Research (ICMR) Needs ScientistIndian Council of Medical Research (ICMR)INIndiaEmployment SamacharThu, 18 Aug 2016 16:16:15 GMTIndian Council Of Medical Research (ICMR) Needs Scientist. Indian Council of Medical Research (ICMR) invites applications to recruit on vacant posts of...<http://www.indeed.co.in/viewjob?jk=20d1db3c7d973199&qd=704PFtVAS6xUi0-OukCaEmfxgGzxqabhMKv0iphFlwZvghJwQWAysomG7BsaL67IpeRHLNudzQ_v_UGEGMFYq0JvivwR6g0dNKs-MyZMxww&indpubnum=8693092939388569&atk=1arpjr78d5upddvtindeed_clk(this>,'6618');20d1db3c7d973199falsefalsefalseIndia16 days ago", : duplicate subscripts for columns

**Change Of Approach**  
After the above error message, I changed my approach.I had to extract data from the nodes by using "getNodetSet" and then had to combine them.After the above was done, the results would not come in respective columns and would come as a list and the heads would come as "Text 1" and so forth.The data is now in a dataframe, which means any column can be extracted, for example indeed\_df['text'] gave me the first column.  
To get all the 1197 results into these column heads, we had to run a "while" loop.There were 25 searches displayed on a single page. So for 1197 searches ,I would have to press "NEXT" (1197/25 = 48) times; therefore run the "while" loop 48 times. I had to define various other parameters like i and run it as a loop 48 times to cover the full job results and define i=1 and later on add a string 2 that would run 48 times (look at the code).Once the string was divided into small parts, I combined the all the three strings to get the full oputput.After this step, I exported the dataframe into a .csv file.

*Libraries Used In Prepration*  
XML, methods, max.print

**First Step:Data Import**  
The first step was to import data from the open API which was in XML format and is read properly in the R Studio. For that I first created the file in R. After the data is read in the R studio. It is important to convert it in the data frame for analysis. Due to the HTML tagging, separate nodes were extracted using "getnodes". Then the XMl file is parsed and a csv file is created.The structure of the indeedfile csv is below:

indeedfile <- read.csv("C:/Users/deepa/OneDrive/Documents/capstone csv/indeed1.csv")  
  
str(indeedfile)

## 'data.frame': 1016 obs. of 11 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ jobtitle : Factor w/ 476 levels "3rd Party Data Matching Expert / Data Scientist / Product Management",..: 328 102 429 73 202 73 374 323 430 88 ...  
## $ company : Factor w/ 466 levels "24 7 Customer Private Ltd",..: 443 443 334 446 431 217 75 144 159 75 ...  
## $ city : Factor w/ 21 levels "Adoni","Ahmedabad",..: NA NA 4 NA 10 4 NA 4 4 4 ...  
## $ state : Factor w/ 13 levels "AP","AS","CH",..: NA NA 8 NA 1 8 NA 8 8 8 ...  
## $ country : Factor w/ 1 level "IN": 1 1 1 1 1 1 1 1 1 1 ...  
## $ formattedLocationFull: Factor w/ 24 levels "Adoni, Andhra Pradesh",..: 11 11 4 11 10 4 11 4 4 4 ...  
## $ source : Factor w/ 117 levels "Accenture","Adobe",..: 111 111 84 112 108 47 13 27 31 13 ...  
## $ date : Factor w/ 549 levels "Fri, 02 Sep 2016 02:40:34 GMT",..: 459 458 457 79 321 400 396 473 131 134 ...  
## $ snippet : Factor w/ 602 levels "?The core objective of this role is to work closely with the solution partners identified and with Business teams as Data Scien"| \_\_truncated\_\_,..: 318 553 508 242 585 317 349 360 587 31 ...  
## $ expired : logi FALSE FALSE FALSE FALSE FALSE FALSE ...

**Second Step:Data Preparation**  
*Adding the logic*- Now once the XML file is converted into dataframe and then to the csv, it is important to look for companies that have the three top skills "Python", "machine learning", "SQL". This was achieved by putting the logic of TRUE if the respective skills were found otherwise put FALSE and create separate columns with respective logical answer. Then these three columns were merged with the existing data file indeed1.csv and new csv file was named finalindeed1\_trynote.csv.The finalindeed1\_trynote had the following structure:

trynoteindeedfile <- read.csv("C:/Users/deepa/OneDrive/Documents/capstone csv/finalindeed1\_trynote.csv")  
  
str(trynoteindeedfile)

## 'data.frame': 1104 obs. of 14 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ jobtitle : Factor w/ 488 levels "3rd Party Data Matching Expert / Data Scientist / Product Management",..: 87 80 131 165 80 207 80 80 82 80 ...  
## $ company : Factor w/ 475 levels "24 7 Customer Private Ltd",..: 445 122 236 260 455 440 8 196 291 197 ...  
## $ city : Factor w/ 22 levels "Adoni","Ahmedabad",..: 10 NA 8 8 NA 10 4 4 20 4 ...  
## $ state : Factor w/ 13 levels "AP","AS","CH",..: 1 NA 6 6 NA 1 8 8 10 8 ...  
## $ country : Factor w/ 1 level "IN": 1 1 1 1 1 1 1 1 1 1 ...  
## $ formattedLocationFull: Factor w/ 23 levels "600034","Adoni, Andhra Pradesh",..: 11 12 9 9 12 11 5 5 21 5 ...  
## $ source : Factor w/ 114 levels "Accenture","Aconex",..: 111 22 51 60 109 106 2 36 68 37 ...  
## $ date : Factor w/ 555 levels "Fri, 02 Sep 2016 02:40:34 GMT",..: 8 29 337 538 73 353 530 460 356 123 ...  
## $ snippet : Factor w/ 614 levels "? Design and maintain data ingest, cleaning, storage, and query systems. Hinge is growing our data science and analytics team i"| \_\_truncated\_\_,..: 218 483 570 41 241 598 137 47 579 227 ...  
## $ expired : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ python : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ machine : logi FALSE FALSE FALSE FALSE FALSE TRUE ...  
## $ sql : logi FALSE FALSE FALSE FALSE FALSE FALSE ...

For consistency purposes and to simplify analysis, I created another file that would lower the cases of every item in the table. The file structure all in lowercase is below:

dflowerfinal1 <- read.csv("C:/Users/deepa/OneDrive/Documents/capstone csv/finaldflower.csv")  
str(dflowerfinal1)

## 'data.frame': 1104 obs. of 14 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ jobtitle : Factor w/ 477 levels "3rd party data matching expert / data scientist / product management",..: 84 79 128 159 79 201 79 79 79 79 ...  
## $ company : Factor w/ 473 levels "24 7 customer private ltd",..: 443 121 235 259 453 438 8 195 290 196 ...  
## $ city : Factor w/ 22 levels "adoni","ahmedabad",..: 10 NA 8 8 NA 10 4 4 20 4 ...  
## $ state : Factor w/ 13 levels "ap","as","ch",..: 1 NA 6 6 NA 1 8 8 10 8 ...  
## $ country : Factor w/ 1 level "in": 1 1 1 1 1 1 1 1 1 1 ...  
## $ formattedLocationFull: Factor w/ 23 levels "600034","adoni, andhra pradesh",..: 11 12 9 9 12 11 5 5 21 5 ...  
## $ source : Factor w/ 114 levels "accenture","aconex",..: 111 22 51 60 109 106 2 36 68 37 ...  
## $ date : Factor w/ 555 levels "fri, 02 sep 2016 02:40:34 gmt",..: 8 29 337 538 73 353 530 460 356 123 ...  
## $ snippet : Factor w/ 614 levels "?the core objective of this role is to work closely with the solution partners identified and with business teams as data scien"| \_\_truncated\_\_,..: 219 483 570 42 242 598 138 48 579 228 ...  
## $ expired : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ python : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ machine : logi FALSE FALSE FALSE FALSE FALSE TRUE ...  
## $ sql : logi FALSE FALSE FALSE FALSE FALSE FALSE ...

head(dflowerfinal1)

## X jobtitle  
## 1 1 data scientist- valuelabs  
## 2 2 data scientist  
## 3 3 data scientist - intern  
## 4 4 data scientist - orgsolutions, mckinsey solutions  
## 5 5 data scientist  
## 6 6 data scientist (intern)  
## company city state country  
## 1 valuelabs llp hyderabad ap in  
## 2 educational initiatives <NA> <NA> in  
## 3 kvantum gurgaon hr in  
## 4 mckinsey & company gurgaon hr in  
## 5 vserv digital services pvt. ltd <NA> <NA> in  
## 6 unwired hyderabad ap in  
## formattedLocationFull source  
## 1 hyderabad, andhra pradesh wisdomjobs.com  
## 2 india educational initiatives  
## 3 gurgaon, haryana kvantum  
## 4 gurgaon, haryana mckinsey & company  
## 5 india vserv digital services pvt. ltd  
## 6 hyderabad, andhra pradesh unwired  
## date  
## 1 fri, 02 sep 2016 02:49:57 gmt  
## 2 fri, 02 sep 2016 06:20:38 gmt  
## 3 sun, 11 sep 2016 11:47:42 gmt  
## 4 wed, 17 aug 2016 16:39:57 gmt  
## 5 fri, 22 jul 2016 09:47:15 gmt  
## 6 thu, 04 aug 2016 15:33:10 gmt  
## snippet  
## 1 hope you are doing well we have openings for data science analyst for our company job location hyderabad experience 3- 10years notice period imm-15 days ctc...  
## 2 strong knowledge in reporting tools, data manipulation, statistical models and visual presentation. synthesize diverse, complex information to develop a...  
## 3 we are looking for candidates who can be an intern for 6 months and then join us as full-time data scientist after successful completion of internship....  
## 4 are you a data scientist? data scientists are creative problem solvers who love all things data. the data scientist we are looking for is not just someone who...  
## 5 jd - data scientist. data scientist will be responsible to design and implement efficient, adaptable, &amp; reusable code and algorithms....  
## 6 work on a small data science team in charge of data mining, predictive modeling and analytics design. strong technical skills regarding data analysis, machine...  
## expired python machine sql  
## 1 FALSE FALSE FALSE FALSE  
## 2 FALSE FALSE FALSE FALSE  
## 3 FALSE FALSE FALSE FALSE  
## 4 FALSE FALSE FALSE FALSE  
## 5 FALSE FALSE FALSE FALSE  
## 6 FALSE FALSE TRUE FALSE

**Third Step: Data Wrangling**

*Libraries Used In Data Wrangling*  
plyr, dplyr, magrittr, ggplot2

Now that the data sheet is cleaned, it is time to start slicing the data and develop insights from it. To conduct analysis, we would now answer various questions that would give an indepth analysis of the data scientist job situation in India.Below are the questions:

***Which states are hiring for data scientist roles?***  
To answer the above question, I looked at the data in the states column and looked for unique values. Once the unique values were identified, I found the frequency of each unique value and find out it's frequency.The file structure for state frequency is below:

statesarranged <- read.csv("C:/Users/deepa/OneDrive/Documents/capstone csv/finalarrangedfrequencystate1.csv")  
str(statesarranged)

## 'data.frame': 13 obs. of 3 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ states : Factor w/ 13 levels "ap","as","ch",..: 8 10 4 1 11 12 6 3 13 5 ...  
## $ Frequencystates: int 548 228 89 71 65 33 27 15 7 6 ...

head(statesarranged)

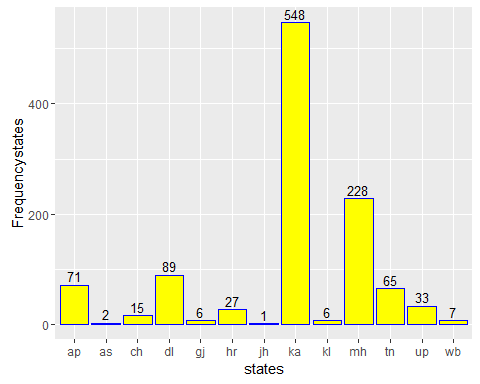
## X states Frequencystates  
## 1 1 ka 548  
## 2 2 mh 228  
## 3 3 dl 89  
## 4 4 ap 71  
## 5 5 tn 65  
## 6 6 up 33

library(ggplot2)

Following is the barplot displaying the number of data scientist jobs in various states in India.

statebarplot<- ggplot(data=statesarranged, aes(x=states, y=Frequencystates)) +  
 geom\_bar(stat = "identity",fill= "yellow", colour= "blue")+  
 geom\_text(aes(label=Frequencystates), vjust=-0.3, size=3.5)  
 statebarplot

**State Frequency Table**



As you can see in the above that the states of ka(Karnataka), mh(Maharashtra), dl(Delhi) have the highest data scientist jobs in India.

***Which cities are hiring for data scientist roles?***  
To answer the above question, I looked at the data in the cities column and looked for unique values. Once the unique values were identified, I found the frequency of each unique value.The file structure for state frequency is below:

citiesarranged <- read.csv("C:/Users/deepa/OneDrive/Documents/finalfreqcity.csv")  
str(citiesarranged)

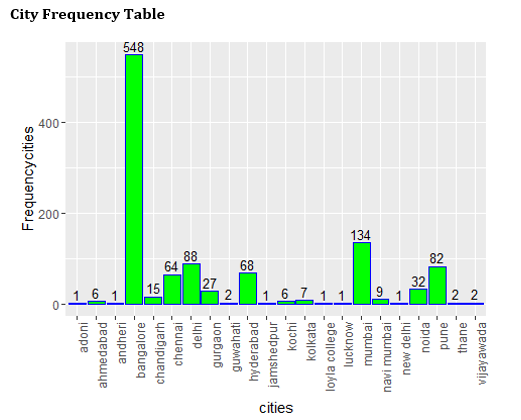
## 'data.frame': 22 obs. of 3 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ cities : Factor w/ 22 levels "adoni","ahmedabad",..: 4 16 7 20 10 6 19 8 5 17 ...  
## $ Frequencycities: int 548 134 88 82 68 64 32 27 15 9 ...

head(citiesarranged)

## X cities Frequencycities  
## 1 1 bangalore 548  
## 2 2 mumbai 134  
## 3 3 delhi 88  
## 4 4 pune 82  
## 5 5 hyderabad 68  
## 6 6 chennai 64

Following is the barplot displaying the number of data scientist jobs in various states in India.

citybarplot<- ggplot(data=citiesarranged, aes(x=cities, y=Frequencycities)) +  
 geom\_bar(stat = "identity",fill= "green", colour= "blue")+  
 geom\_text(aes(label=Frequencycities),vjust=-0.3, size=3.5) +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))  
   
 citybarplot



As you can see in the above that the cities of Bangalore, Delhi, Mumbai have the highest data scientist jobs in India. The findings of this data is consistent with the states data which recognised the respective states corresponsing to these cities as highest concentrated areas for data scientist jobs.

***Which companies/consultants are hiring for data scientist roles?***  
To answer the above question, I looked at the data in the companies column and looked for unique values. Once the unique values were identified, I found the frequency of each unique value. The consultants are also referred as a company for simplicity purposes.

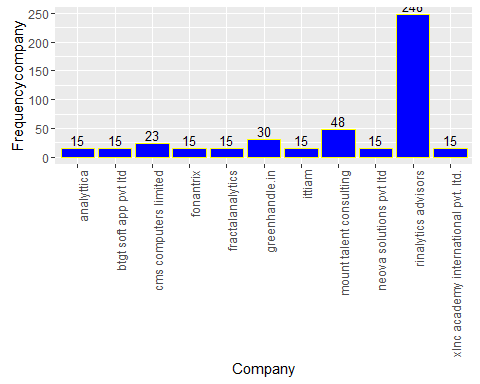
companiesarranged <- read.csv("C:/Users/deepa/OneDrive/Documents/capstone csv/finalarrangedfrequencycompany.csv")  
head(companiesarranged)

## X Company Frequencycompany  
## 1 1 rinalytics advisors 248  
## 2 2 mount talent consulting 48  
## 3 3 greenhandle.in 30  
## 4 4 cms computers limited 23  
## 5 5 analyttica 15  
## 6 6 btgt soft app pvt ltd 15

Following is the barplot displaying the number of data scientist jobs according to jobs posted in India on the Indeed website.

companybarplot<- ggplot(data=companiesarranged, aes(x=Company, y=Frequencycompany)) +  
 geom\_bar(stat = "identity",fill= "blue", colour= "yellow")+  
 geom\_text(aes(label=Frequencycompany),vjust=-0.3, size=3.5)+  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))  
 companybarplot

**Company/Consultant Frequency Table**



**Data Exploration**

**Text Analysis**  
To answer the questions about non-technical skills needed for the data scientist role, I extracted the "snippet" column from the dataframe which consisted the job description including non-technical and technical skills.Then I chose Cran Mirror 50 which was closest to my location.Then I used "Bag of Words" approach to turn text into numeric.Preprocessing consisted of removing punctuations, stop words like "scienti","work", "are", "looking" etc. which were general to the use of the analysis.

*Libraries & Package Used in Text Analysis*  
tm,SnowballC, caTools, Rpart, RcolorBrewer, wordcloud, biclust, cluster,igraph, fpc and used package "Tcampdf".

***What non-technical skills does one need for a Data Scientist role in India?***  
To answer the above question, I resorted to mine the text available in the "snippet" column, which was originally the "job description" column. After pre-processing the data by removing numbers, punctuations and redundant words, I was able to generate a wordcloud. The wordcloud was coded (the code of which is available in the attached code).As you would notice that the words are truncated, I checked for this in the master file and the words are in the same format as in the original file. The wordcloud was saved as a png and is as follows:

**Non-technical Skill Word Cloud**



***What technical skills does one need for a Data Scientist role in India?***

To answer the above questions, we created the code using "if else" , "logical condition" and dplyr "regex". We identified python, sql and machine learning as the three main technical skills. Using the 'if else' conditional statement and dplyr regular expression, we created the condition that if in the snippet column we see "python", "sql", "machine" make a respective column name python, sql and machine and print the logic "true" or "false". Then to identify the the companies who require all the technical skills, used the pipe function to filter the columns that have "true" in all the columns python, sql and machine.The following is the structure of the technical skills file.

alltechnicalskills <-read.csv("C:/Users/deepa/OneDrive/Documents/technicalskills1.csv")  
str(alltechnicalskills)

## 'data.frame': 2 obs. of 15 variables:  
## $ X.1 : int 1 2  
## $ X : int 1 2  
## $ jobtitle : Factor w/ 1 level "data scientist - machine learning nlp java": 1 1  
## $ company : Factor w/ 2 levels "premium-jobs",..: 2 1  
## $ city : Factor w/ 1 level "delhi": 1 1  
## $ state : Factor w/ 1 level "dl": 1 1  
## $ country : Factor w/ 1 level "in": 1 1  
## $ formattedLocationFull: Factor w/ 1 level "delhi, delhi": 1 1  
## $ source : Factor w/ 1 level "wisdomjobs.com": 1 1  
## $ date : Factor w/ 2 levels "fri, 02 sep 2016 02:50:16 gmt",..: 1 2  
## $ snippet : Factor w/ 2 levels "ky skills machine learning sas hadoop octave big data weka nlp nosql mahout mathematica mongodb spss matlab python nlp programm"| \_\_truncated\_\_,..: 1 2  
## $ expired : logi FALSE FALSE  
## $ python : logi TRUE TRUE  
## $ machine : logi TRUE TRUE  
## $ sql : logi TRUE TRUE

head(alltechnicalskills)

## X.1 X jobtitle company city  
## 1 1 1 data scientist - machine learning nlp java staffio delhi  
## 2 2 2 data scientist - machine learning nlp java premium-jobs delhi  
## state country formattedLocationFull source  
## 1 dl in delhi, delhi wisdomjobs.com  
## 2 dl in delhi, delhi wisdomjobs.com  
## date  
## 1 fri, 02 sep 2016 02:50:16 gmt  
## 2 fri, 26 aug 2016 03:13:45 gmt  
## snippet  
## 1 ky skills machine learning sas hadoop octave big data weka nlp nosql mahout mathematica mongodb spss matlab python nlp programming languages- minimum 2 yrs...  
## 2 ky skills machine learning sas hadoop octave big data weka nlp nosql mahout mathematica mongodb spss matlab python nlp programming languages - minimum 2 yrs...  
## expired python machine sql  
## 1 FALSE TRUE TRUE TRUE  
## 2 FALSE TRUE TRUE TRUE

Following is the table of the companies looking for three important technical skills python, sql, machine learning.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Title | Company | City | State | Expired | Python | Machine | SQL |
| data scientist - machine learning nlp java | staffio | Delhi | Delhi | False | True | True | True |
| data scientist - machine learning nlp java | premium-jobs | Delhi | Delhi | False | True | True | True |

# Findings:

Overall from the analysis above, I could conclusively say:

1. States of Maharashtra, Karnataka and Delhi have the highest job concentration of data scientist jobs in India.

2. Cities of Delhi (Delhi state), Banaglore (Karnataka state) and Mumbai (Maharshtra state) are the cities producing data scientist jobs in India.

3. Rinalytics Advisors, Mount Talent Consulting and Greenhandle have more openings for data scientist roles in India.

4. Non-technical Skills like requirement for such jobs are "analyze, responsible, develop, build, predict". So these words should be included in the resume when applying for the data scientist position.

5. Consultants like Staffio and Premium-jobs look for all three most important technical skills namely Python, SQL and Machine learning. If anyone is applying to an agency in Delhi city, must make sure is well versed with these technical skills.

# Lessons Learnt:

1. Opportunities in using a second country like Canada to draw comparisons from my friend’s current country to aid in answering a more compelling question “Is it worth immigrating from Canada to India in search of Data Scientist job?”

2. Opportunities in finding data on the Data Scientist job numbers in India and Canada for conducting predictive analysis.

3. Could have chosen a data set that had more numerical values over various periods to be able to use predictive modeling and data modeling.