Exploring the set of Skills Required for Data Scientists, Data Analysts and Data Engineers

Sahibzada Ali Mahmud

July 29, 2019

Executive Summary

The idea behind undertaking this project was mainly pertaining to the dataset which solicited my interest in exploring the data for further analysis. The dataset is on job postings for data scientists, data analysts and data engineers on the webiste of Indeed. It can be downloaded from Kaggle by visiting this link. The dataset contains the number of skills that are mentioned against each type of job while also separately specifying columns for each of the common core skills that are required for data scientists, data engineers, and data analysts. Due to time constraints, the analysis was only limited to exploratory data analysis while also checking the correlation between certain variables to determine the strength of their linear relationship. Since the relationship turned out to be weak, as future work, other regression and classification techniques can be used to extract meaningful relationshiops between variables and also make predictions. Two interesting plots are also presented that can solicit the interest of those who intend to have a better idea about what this dataset has to offer.

1. Prelimenary Data Exploration

Initially the data was extracted from the csv file through read_csv function which is a part of readr package in R. It is comparatively faster than read.csv and returns a tibble compared to a data.frame. The size of the dataset was 18.5 MB. The reason for using a small dataset is because of the limited computational resources and memory available at disposal as well as the nature of dataset.

2. Loading Libraries and Extracting Data

```
#Load Required Libraries
library(dplyr)
library(tidyverse)
library(caret)
library(readr)

#Extract Data
# Read from the .csv file
ds_jobs_data <- read_csv("F:\\TDI Data Sets\\Indeed Dataset\\indeed_job_dataset.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Job_Title = col_character(),
##   Link = col_character(),
##   Queried_Salary = col_character(),
```

```
##   Job_Type = col_character(),
##   Skill = col_character(),
##   Company = col_character(),
##   No_of_Reviews = col_double(),
##   No_of_Stars = col_double(),
##   Description = col_character(),
##   Location = col_character(),
##   Company_Revenue = col_character(),
##   Company_Employees = col_character(),
##   Company_Industry = col_character()
## )

## See spec(...) for full column specifications.
```

3. Oberving Data

```
str(ds_jobs_data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    5715 obs. of  43 variables:
##  $ X1                            : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Job_Title                     : chr  "Data Scientist" "Data Scientist" "Data Scientist" "Graduat
##  $ Link                          : chr  "https://www.indeed.com/rc/clk?jk=6a105f495c36afe3&fccid=28
##  $ Queried_Salary                : chr  "<80000" "<80000" "<80000" "<80000" ...
##  $ Job_Type                      : chr  "data_scientist" "data_scientist" "data_scientist" "data_sc
##  $ Skill                         : chr  "['SAP', 'SQL']" "['Machine Learning', 'R', 'SAS', 'SQL',
##  $ No_of_Skills                  : int  2 5 9 1 7 6 10 3 4 6 ...
##  $ Company                       : chr  "Express Scripts" "Money Mart Financial Services" "comScor
##  $ No_of_Reviews                 : num  3301 NA 62 158 495 ...
##  $ No_of_Stars                   : num  3.3 NA 3.5 4.3 4.1 ...
##  $ Date_Since_Posted             : int  1 15 1 30 30 30 5 10 1 22 ...
##  $ Description                   : chr  "[<p><b>POSITION SUMMARY</b></p>, <p>\r\r\nThe Business Ana
##  $ Location                      : chr  "MO" "TX" "OR" "DC" ...
##  $ Company_Revenue               : chr  "More than $10B (USD)" NA NA NA ...
##  $ Company_Employees             : chr  "10,000+" NA NA NA ...
##  $ Company_Industry              : chr  "Health Care" NA NA "Government" ...
##  $ python                        : int  0 1 1 0 0 0 1 0 1 1 ...
##  $ sql                           : int  1 1 1 0 0 0 1 1 0 0 ...
##  $ machine learning              : int  0 1 0 0 0 1 1 1 0 0 ...
##  $ r                             : int  0 1 1 0 1 0 1 1 1 1 ...
##  $ hadoop                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ tableau                       : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ sas                           : int  0 1 1 0 0 0 0 0 0 0 ...
##  $ spark                         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ java                          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Others                        : int  1 0 1 1 1 1 1 0 1 1 ...
##  $ CA                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ NY                            : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ VA                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ TX                            : int  0 1 0 0 1 0 0 0 0 0 ...
##  $ MA                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ IL                            : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ WA                              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ MD                              : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ DC                              : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ NC                              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Other_states                    : int  1 0 1 0 0 0 0 1 1 1 ...
##  $ Consulting and Business Services: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Internet and Software           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Banks and Financial Services    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Health Care                     : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ Insurance                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Other_industries                : int  0 0 0 1 0 0 1 0 1 1 ...
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 43
##   .. ..$ X1                              : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Job_Title                       : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ Link                            : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ Queried_Salary                  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ Job_Type                        : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ Skill                           : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ No_of_Skills                    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Company                         : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ No_of_Reviews                   : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ No_of_Stars                     : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ Date_Since_Posted               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Description                     : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ Location                        : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ Company_Revenue                 : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ Company_Employees               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ Company_Industry                : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ python                          : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ sql                             : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ machine learning                : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ r                               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ hadoop                          : list()
```

```
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ tableau                          : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ sas                              : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ spark                            : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ java                             : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Others                           : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ CA                               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ NY                               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ VA                               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ TX                               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ MA                               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ IL                               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ WA                               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ MD                               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ DC                               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ NC                               : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Other_states                     : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Consulting and Business Services: list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Internet and Software            : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Banks and Financial Services     : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Health Care                      : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Insurance                        : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Other_industries                 : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

```
head(ds_jobs_data)
```

```
## # A tibble: 6 x 43
##      X1 Job_Title Link  Queried_Salary Job_Type Skill No_of_Skills Company
##   <int> <chr>     <chr> <chr>          <chr>    <chr>        <int> <chr>
```

```
## 1     0 Data Sci~ http~ <80000         data_sc~ ['SA~            2 Expres~
## 2     1 Data Sci~ http~ <80000         data_sc~ ['Ma~            5 Money ~
## 3     2 Data Sci~ http~ <80000         data_sc~ ['Da~            9 comSco~
## 4     3 Graduate~ http~ <80000         data_sc~ ['Ce~            1 Centra~
## 5     4 Data Sci~ http~ <80000         data_sc~ ['St~            7 Federa~
## 6     5 Data Sci~ http~ <80000         data_sc~ ['AI~            6 Nation~
## # ... with 35 more variables: No_of_Reviews <dbl>, No_of_Stars <dbl>,
## #   Date_Since_Posted <int>, Description <chr>, Location <chr>,
## #   Company_Revenue <chr>, Company_Employees <chr>,
## #   Company_Industry <chr>, python <int>, sql <int>, `machine
## #   learning` <int>, r <int>, hadoop <int>, tableau <int>, sas <int>,
## #   spark <int>, java <int>, Others <int>, CA <int>, NY <int>, VA <int>,
## #   TX <int>, MA <int>, IL <int>, WA <int>, MD <int>, DC <int>, NC <int>,
## #   Other_states <int>, `Consulting and Business Services` <int>,
## #   `Internet and Software` <int>, `Banks and Financial Services` <int>,
## #   `Health Care` <int>, Insurance <int>, Other_industries <int>
```

4. Checking Basic Trends and Creating Subsets based on Job Types

```
# Check Statewise number of opportunities for data scientists: CA highest with 723 followed by NY
ds_state_trend <- ds_jobs_data %>% filter(Job_Type == "data_scientist") %>% group_by(Location) %>% tally
ds_state_trend
```

```
## # A tibble: 46 x 2
##    Location     n
##    <chr>    <int>
##  1 <NA>         2
##  2 AL           9
##  3 AR          18
##  4 AZ          24
##  5 CA         723
##  6 CO          47
##  7 CT          32
##  8 DC          68
##  9 DE           7
## 10 FL          43
## # ... with 36 more rows
```

```
print(ds_state_trend, n=46)
```

```
## # A tibble: 46 x 2
##    Location     n
##    <chr>    <int>
##  1 <NA>         2
##  2 AL           9
##  3 AR          18
##  4 AZ          24
##  5 CA         723
##  6 CO          47
##  7 CT          32
##  8 DC          68
```

```
##  9 DE           7
## 10 FL          43
## 11 GA          56
## 12 HI           4
## 13 IA           9
## 14 ID           5
## 15 IL         106
## 16 IN          18
## 17 KS           5
## 18 KY           6
## 19 LA           4
## 20 MA         133
## 21 MD          94
## 22 ME           2
## 23 MI          30
## 24 MN          26
## 25 MO          33
## 26 NC          55
## 27 NE           7
## 28 NH           3
## 29 NJ          54
## 30 NM           3
## 31 NV           8
## 32 NY         253
## 33 OH          49
## 34 OK           1
## 35 OR          18
## 36 PA          55
## 37 REMOTE       5
## 38 RI           3
## 39 SC           6
## 40 TN          16
## 41 TX         136
## 42 USA         39
## 43 UT          10
## 44 VA         177
## 45 WA         122
## 46 WI          19
```

```
# Check Statewise number of opportunities for data analysts: CA highest with 376 followed by Ny
da_state_trend <- ds_jobs_data %>% filter(Job_Type == "data_analyst") %>% group_by(Location) %>% tally()
da_state_trend
```

```
## # A tibble: 50 x 2
##    Location     n
##    <chr>    <int>
##  1 <NA>         2
##  2 AL           6
##  3 AR           8
##  4 AZ          15
##  5 CA         376
##  6 CO          25
##  7 CT          29
##  8 DC          48
```

```
##  9 DE           6
## 10 FL          44
## # ... with 40 more rows
```

```
print(da_state_trend, n=50)
```

```
## # A tibble: 50 x 2
##    Location     n
##    <chr>    <int>
##  1 <NA>         2
##  2 AL           6
##  3 AR           8
##  4 AZ          15
##  5 CA         376
##  6 CO          25
##  7 CT          29
##  8 DC          48
##  9 DE           6
## 10 FL          44
## 11 GA          59
## 12 HI           2
## 13 IA          10
## 14 ID           1
## 15 IL          66
## 16 IN          20
## 17 KS           1
## 18 KY           6
## 19 LA           1
## 20 MA          86
## 21 MD          55
## 22 ME           6
## 23 MI          39
## 24 MN          27
## 25 MO          37
## 26 NC          46
## 27 NE           1
## 28 NH           2
## 29 NJ          54
## 30 NM           2
## 31 NV           7
## 32 NY         230
## 33 OH          33
## 34 OK           2
## 35 OR          26
## 36 PA          56
## 37 REMOTE       4
## 38 RI           2
## 39 SC          15
## 40 SD           1
## 41 TN          21
## 42 TX         117
## 43 USA         24
## 44 UT          17
## 45 VA          85
```

```
## 46 VT           2
## 47 WA          57
## 48 WI          11
## 49 WV           2
## 50 WY           1
```

```
# Dividing the data set into three subsets i.e. for data scientist jobs,
# for data analyst jobs and data engineer jobs
ds_subset <- ds_jobs_data %>% filter(Job_Type == "data_scientist")

da_subset <- ds_jobs_data %>% filter(Job_Type == "data_analyst")

de_subset <- ds_jobs_data %>% filter(Job_Type == "data_engineer")

# Checking the number of records for each discipline
nrow(ds_subset)
```

```
## [1] 2543
```

```
nrow(da_subset)
```

```
## [1] 1793
```

```
nrow(de_subset)
```

```
## [1] 1379
```

5.Checking the Percentage of Skills in Demand For each Discipline

```
# Cheking the mean of No. of Skills required for each discipline
mean(ds_subset$No_of_Skills)
```

```
## [1] 8.493118
```

```
mean(da_subset$No_of_Skills)
```

```
## [1] 4.490798
```

```
mean(de_subset$No_of_Skills)
```

```
## [1] 10.83974
```

```
# % Requirement of r for each discipline
((ds_subset %>% filter(r == 1) %>% count())/ nrow(ds_subset)) * 100 # 60.9%
```

```
##          n
## 1 60.95163
```

```r
((da_subset %>% filter(r == 1) %>% count())/ nrow(da_subset)) * 100 # 25.4%
```

```
##          n
## 1 25.43224
```

```r
((de_subset %>% filter(r == 1) %>% count())/ nrow(de_subset)) * 100 # 16.5%
```

```
##          n
## 1 16.53372
```

```r
# % Requirement of Python for each discipline
# for data engineers, python is more preferred instead of r
((ds_subset %>% filter(python == 1) %>% count())/ nrow(ds_subset)) * 100 # 75.1%
```

```
##          n
## 1 75.18679
```

```r
((da_subset %>% filter(python == 1) %>% count())/ nrow(da_subset)) * 100 # 28.5%
```

```
##          n
## 1 28.55549
```

```r
((de_subset %>% filter(python == 1) %>% count())/ nrow(de_subset)) * 100 # 65.3%
```

```
##         n
## 1 65.3372
```

```r
# % Requirement of r and python for each discipline
((ds_subset %>% filter(r == 1 & python == 1) %>% count())/ nrow(ds_subset)) * 100 # 54.4%
```

```
##          n
## 1 54.46323
```

```r
((da_subset %>% filter(r == 1 & python == 1) %>% count())/ nrow(da_subset)) * 100 # 19.4%
```

```
##          n
## 1 19.46458
```

```r
((de_subset %>% filter(r == 1 & python == 1) %>% count())/ nrow(de_subset)) * 100 # 13.8%
```

```
##          n
## 1 13.85062
```

```r
# % Requirement of r and sql for each discipline
((ds_subset %>% filter(r == 1 & sql == 1) %>% count())/ nrow(ds_subset)) * 100 # 36.2%
```

```
##          n
## 1 36.29571
```

```r
((da_subset %>% filter(r == 1 & sql == 1) %>% count())/ nrow(da_subset)) * 100 # 19.6%
```

```
##         n
## 1 19.6319
```

```r
((de_subset %>% filter(r == 1 & sql == 1) %>% count())/ nrow(de_subset)) * 100 # 13.3%
```

```
##        n
## 1 13.343
```

```r
# % Requirement of python and sql for each discipline
((ds_subset %>% filter(python == 1 & sql == 1) %>% count())/ nrow(ds_subset)) * 100 # 41.2%
```

```
##          n
## 1 41.21117
```

```r
((da_subset %>% filter(python == 1 & sql == 1) %>% count())/ nrow(da_subset)) * 100 # 22.8%
```

```
##         n
## 1 22.8667
```

```r
((de_subset %>% filter(python == 1 & sql == 1) %>% count())/ nrow(de_subset)) * 100 # 44.7%
```

```
##          n
## 1 44.74257
```

```r
# % Requirement of hadoop and spark for each discipline
((ds_subset %>% filter(hadoop == 1 & spark == 1) %>% count())/ nrow(ds_subset)) * 100 # 20.0%
```

```
##          n
## 1 20.05505
```

```r
((da_subset %>% filter(hadoop == 1 & spark == 1) %>% count())/ nrow(da_subset)) * 100 # 2.0%
```

```
##          n
## 1 2.063581
```

```r
((de_subset %>% filter(hadoop == 1 & spark == 1) %>% count())/ nrow(de_subset)) * 100 # 42.2%
```

```
##         n
## 1 42.2045
```

```r
# % Requirement of all skills for each discipline
((ds_subset %>% filter(hadoop == 1 & spark == 1 & r == 1 & python == 1 & sql == 1 & java == 1 & tableau
```

```
##           n
## 1 0.1966182
```

```
((da_subset %>% filter(hadoop == 1 & spark == 1 & r == 1 & python == 1 & sql == 1 & java == 1 & tableau
```

```
##   n
## 1 0
```

```
((de_subset %>% filter(hadoop == 1 & spark == 1 & r == 1 & python == 1 & sql == 1 & java == 1 & tableau
```

```
##           n
## 1 0.1450326
```

6. Checking Correlation betwen Variables and Making Plots

```
# Setting the grounds for Regression Analysis as a baseline
# Predict number of skills as accurately as possible
# Check correlation between variables
cor(ds_subset$No_of_Reviews, ds_subset$No_of_Skills, use="complete.obs")
```

```
## [1] 0.05154409
```

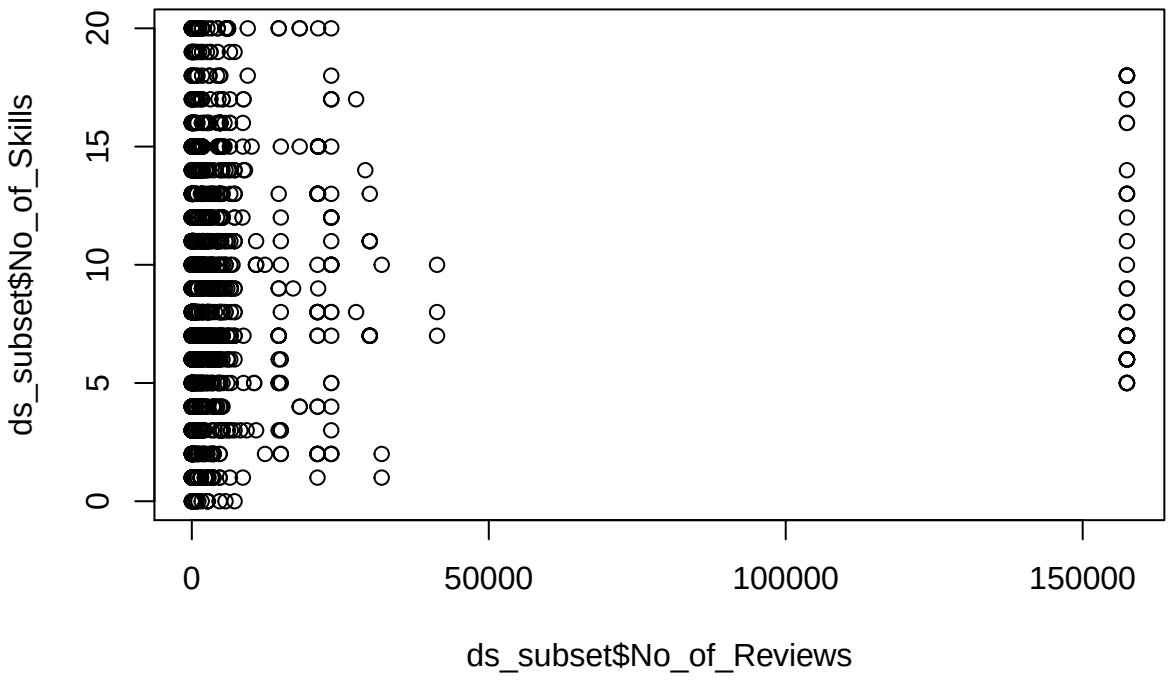
```
cor(ds_subset$No_of_Skills, ds_subset$No_of_Stars, use="complete.obs")
```

```
## [1] -0.0421902
```

```
cor(ds_subset$No_of_Skills, ds_subset$python, use="complete.obs")
```

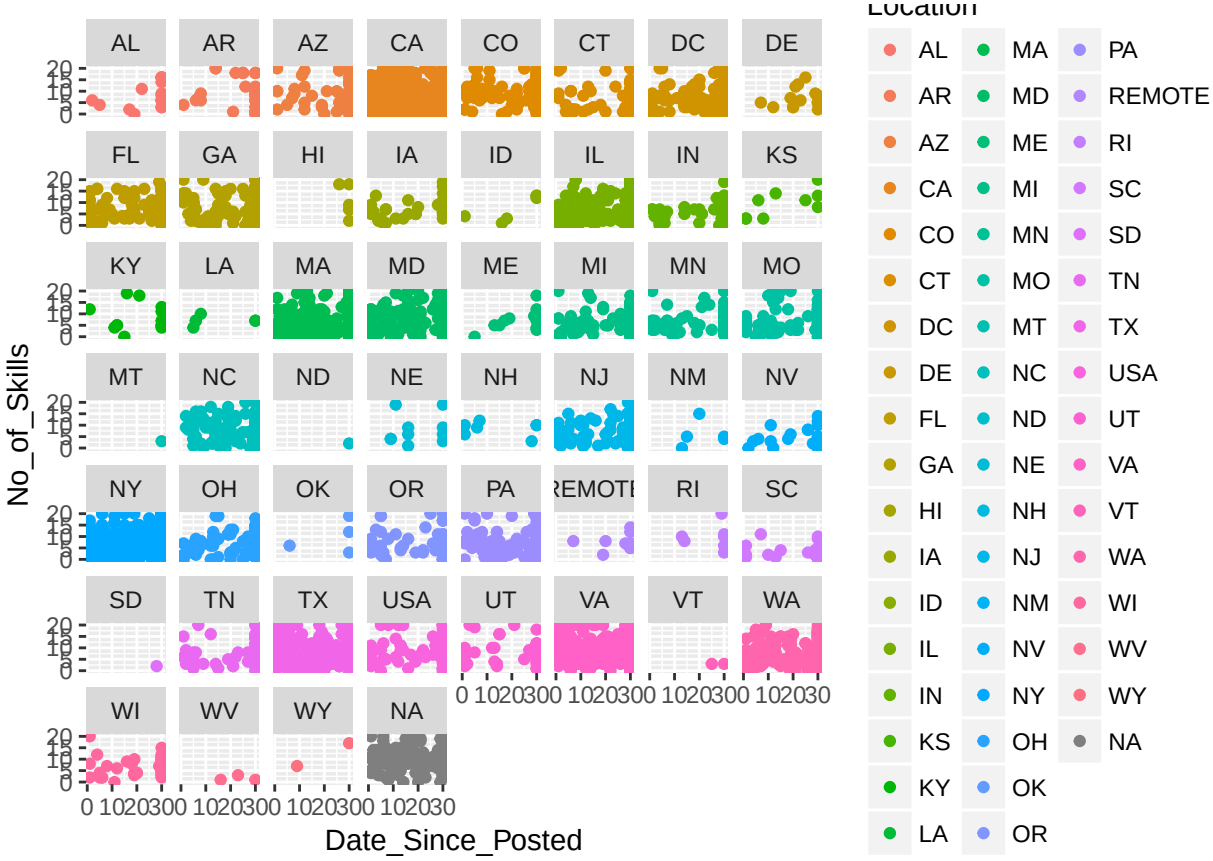
```
## [1] 0.3237191
```

```
plot(ds_subset$No_of_Reviews, ds_subset$No_of_Skills)
```

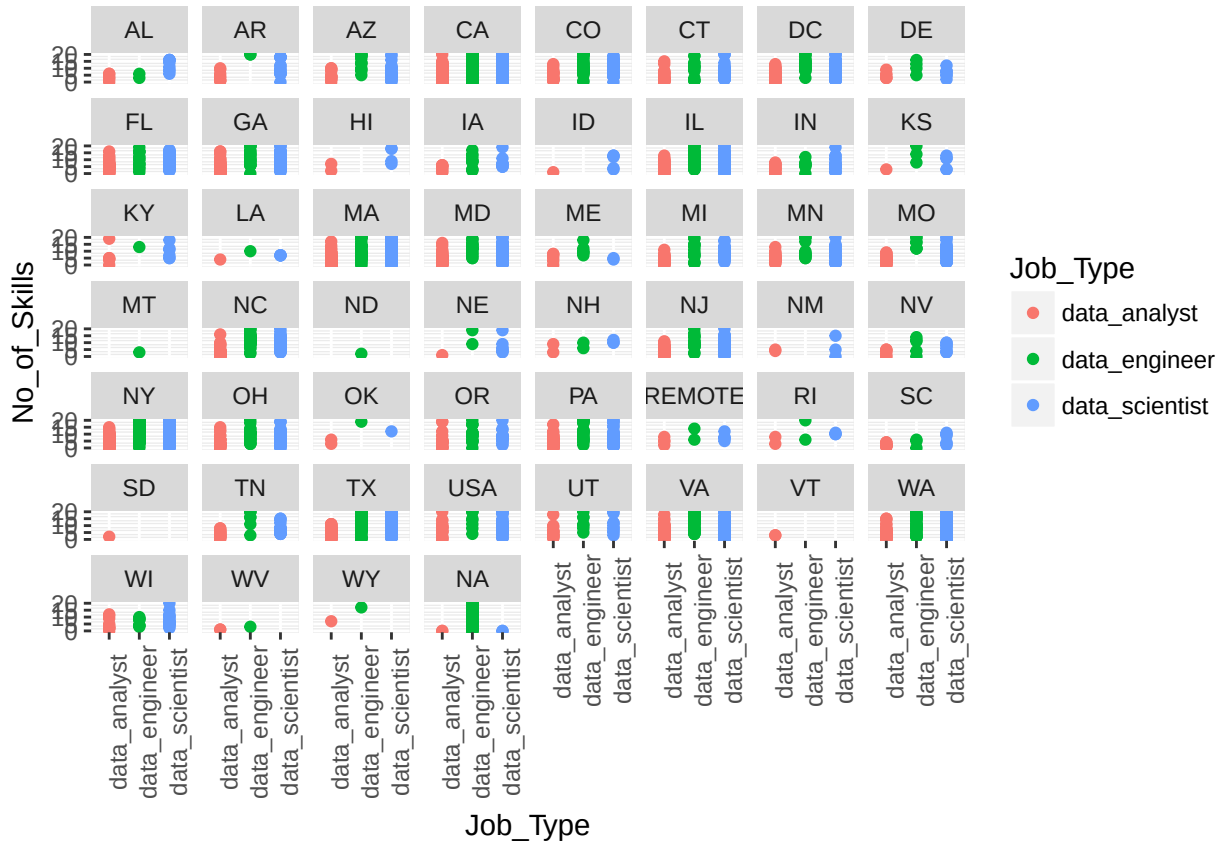



```
# The linear relationship between the variables is weak i.e. weak correlation

# Get some better insight through relevant plots

# This plot shows the No of Skills required per Location
ds_jobs_data %>% ggplot(aes(Date_Since_Posted, No_of_Skills , col = Location)) + geom_point() + facet_wr
```

```
## Warning: Removed 104 rows containing missing values (geom_point).
```




```
# This plot shows the Job Type in demand per Location
ds_jobs_data %>% ggplot(aes(Job_Type, No_of_Skills , col = Job_Type)) + geom_point() + facet_wrap(~Loca
```



7. Conclusion

The dataset from indeed regarding job postings for data scientists, data analysts and data engineers is an interesting one. However, it requires some further clearning and data wrangling before many interesting aspects can be extracted from it. At present, through the exploratory data analysis, some interesting insights regarding the distribution of Number of Skills required per location, Job Type in demand per location, and percentage of skills in demand per Job Type have been provided. As future work, regression analysis and classification can be done for getting meaningful predictions.