Exploratory Data Analysis of Kaggle ML and Data Science Survey, 2017

Hafid Pradipta

April 20, 2018

Introduction

Kaggle conducted an industry-wide survey that receives 16,716 usable respondents from 171 countries and territories. The study was live from August 7th to August 25th. The dataset contains 228 variables. However, some of the variables are the extension of one information. For example, all variables started with WorkToolsFrequency refers to what kind of the work tools that the employee often used for their job.

In general, this dataset contains several groups of variables:

Demographic: Gender, Age, Country, Status as student, Title, Salary, Type of employment, and Education.

Learning platform used: what kind of platform that they use to learn data science. Language recommendation: Which programming language that they recommend.

Job Skill importance: which skills are essential for their career. Learning Category: How do they learn to be data scientist.

Work Tools Frequency: What kind of tools that they use the most.

Work Method Frequency: What method that they use the most.

Time: How do they spend most of their time during their job

Work Challenge Frequency: what are the challenges to be a data scientist. Job Factor: what are the factors that make them a data scientist.

This analysis is the extension of the Jack Cook's code (fromhttps://www.kaggle.com/jackcook/how-to-become-a-data-scientist) who has created an efficient code to extract the information in this dataset. The analysis begins by setting the theme of the visualization and tries to wrangle the data by rename several outcomes to make it simpler.

In the first section, I will compare 16 job titles to figure out what are the characteristics of each job title based on other variables. The list of job titles are: "Business Analyst" "Computer Scientist" "Data Analyst" "Data Miner" "Data Scientist" "DBA/Database Engineer" "Engineer" "Machine Learning Engineer" "Operations Research Practitioner" "Other" "Predictive Modeler" "Programmer" "Researcher" "Scientist/Researcher" "Software Developer/Software Engineer" "Statistician

```
theme1 <- theme(
 plot.background = element_rect(fill = "#eeeeee"),
 panel.background = element_rect(fill = "#eeeeee"),
 legend.background = element_rect(fill = "#eeeeee"),
 legend.title = element_text(size = 11, family  = "Helvetica", face = "bold"),
 legend.text = element_text(size = 9, family = "Helvetica"),
 plot.title = element_text (size = 17, family = "Helvetica", face = "bold", hjust = 0.5),
 panel.grid.major = element_line(size = 0.4, linetype = "solid", color = "#cccccc"),
 panel.grid.minor = element_line(size = 0),
 axis.title = element_text(size = 14, family= "Helvetica", face = "bold"),
 axis.title.x = element_text(margin = margin(t = 20)),
 axis.title.y  = element_text(margin = margin(r = 20)),
 axis.ticks = element_blank()
```

```
)
```

```
## Import the data
results <- read_csv("multipleChoiceResponses.csv")

results$CurrentJobTitleSelect[results$CurrentJobTitleSelect=="Software Developer / Software Engineer"] <
results$MajorSelect[results$MajorSelect == "Engineering (non-computer focused)"] <- "Engineering"

results$MajorSelect[results$MajorSelect == "Information technology, networking, or system administratio

results$MajorSelect[results$MajorSelect == "Management information systems"] <- "Information systems"

results$MajorSelect[results$MajorSelect == "A health science"] <- "Health Science"

results$MajorSelect[results$MajorSelect == "A social science"] <- "Social Science"

results$MajorSelect[results$MajorSelect == "A humanities discipline"] <- "Humanities Discipline"

results_names <- names(results)
results_names[results_names == "WorkMethodsFrequencyA/B"] <- "WorkMethodsFrequencyABTesting"
results_names[results_names == "WorkMethodsFrequencyCross-Validation"] <- "WorkMethodsFrequencyCrossVal

names(results) <- results_names
clean <- results[,c(1:15,35,36,50:59,66:68,74:75,132,167:173,197:228)]
levels(results$CurrentJobTitleSelect)
```

```
## NULL
```

Visualization of Job title and other variables

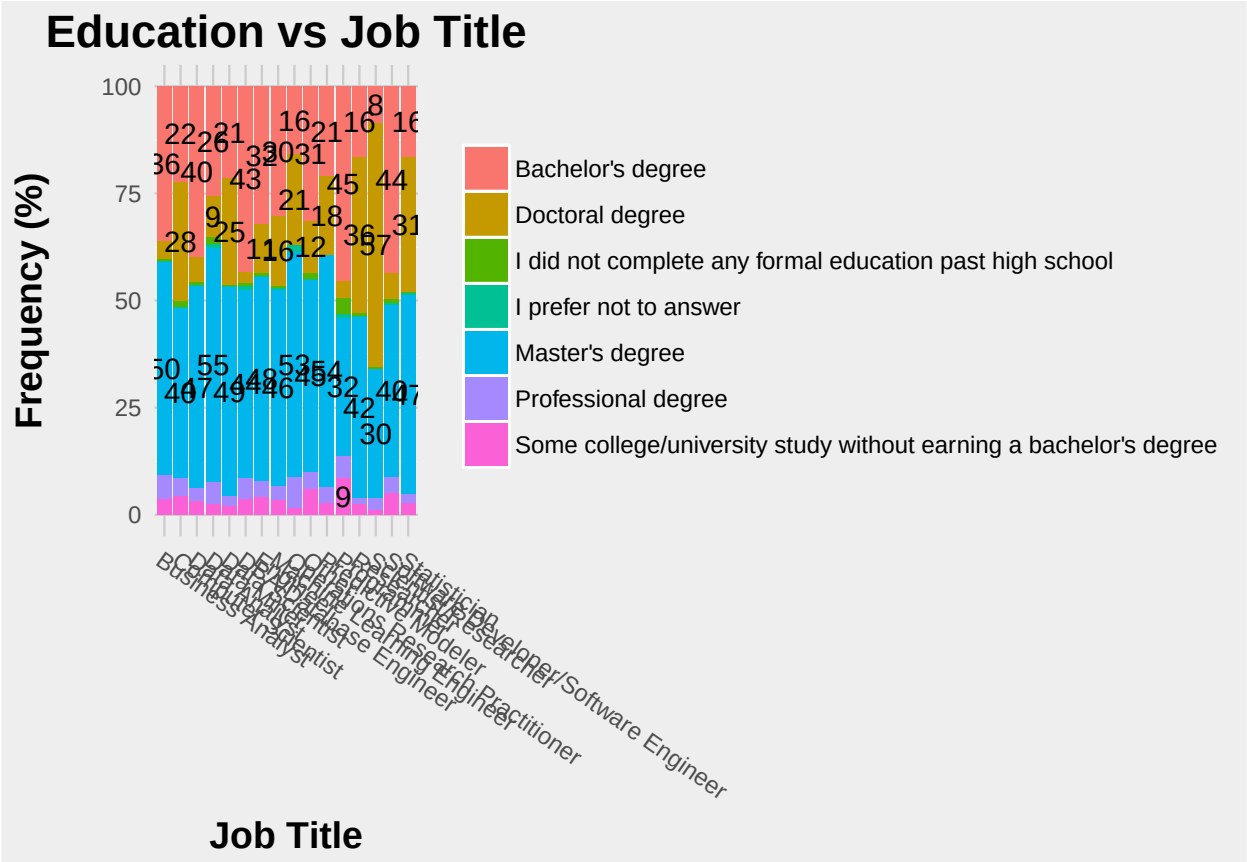
In the first section, I will compare 16 job titles to figure out what are the characteristics of each job title based on other variables. The list of job titles are:

"Business Analyst" "Computer Scientist"
"Data Analyst" "Data Miner"
"Data Scientist" "DBA/Database Engineer"
"Engineer" "Machine Learning Engineer"
"Operations Research Practitioner" "Other"
"Predictive Modeler" "Programmer"
"Researcher" "Scientist/Researcher"
"Software Developer/Software Engineer" "Statistician

Visualization of the Job title and Formal Education

```
results %>%
        rename(title = CurrentJobTitleSelect, Education =FormalEducation ) %>%
        filter(title !="", Education != "") %>%
        group_by(title, Education) %>%
        summarise(n = n()) %>%
        mutate(freq = n/sum(n)*100) %>%
        ggplot(aes(x = title, y = freq, fill = Education, label = ifelse(freq>8, round(freq),"")))+
```

```
        ggtitle("Education vs Job Title")+
        labs(x = "Job Title", y= "Frequency (%)")+
        geom_bar(stat = "identity", position = position_stack())+
        geom_text(position = position_stack(vjust = 0.5))+
        theme1+
        theme(legend.title=element_blank())+
        theme(legend.position="right")+
        theme(axis.title.x = element_text(margin = margin(t=8)),
              axis.text.x = element_text(angle = 325, hjust = 0))
```

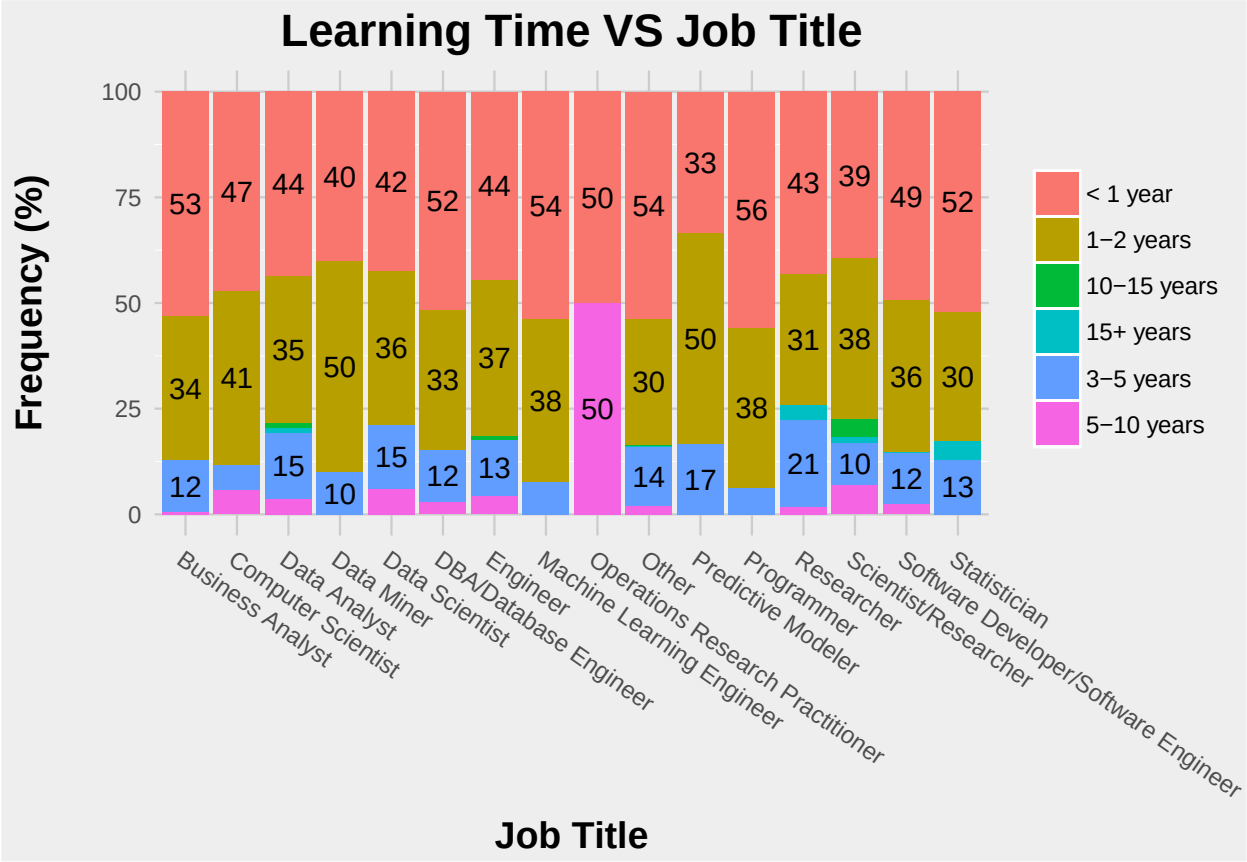


Based on the graph, most of the title in the data scientist fields are dominated by Master's degree with the proportion around 45%. Software developer, Data Analyst, and Database Engineer are the title that employs most of the Bachelor's degree. However, around 57% of the scientist and researchers are Doctoral Degree. It seems that this position requires a high level of education.

Visualization of the Job title and learning time

```
results %>%
        rename(title = CurrentJobTitleSelect, time =LearningDataScienceTime ) %>%
        filter(title !="", time != "") %>%
        group_by(title, time) %>%
        summarise(n = n()) %>%
        mutate(freq = n/sum(n)*100) %>%
        ggplot(aes(x = title, y = freq, fill = time, label = ifelse(freq>8, round(freq),"")))+
        ggtitle("Learning Time VS Job Title")+
```

```
        labs(x = "Job Title", y= "Frequency (%)")+
        geom_bar(stat = "identity", position = position_stack())+
        geom_text(position = position_stack(vjust = 0.5))+
        theme1+
        theme(legend.title=element_blank())+
        theme(axis.title.x = element_text(margin = margin(t=8)),
              axis.text.x = element_text(angle = 325, hjust = 0))
```

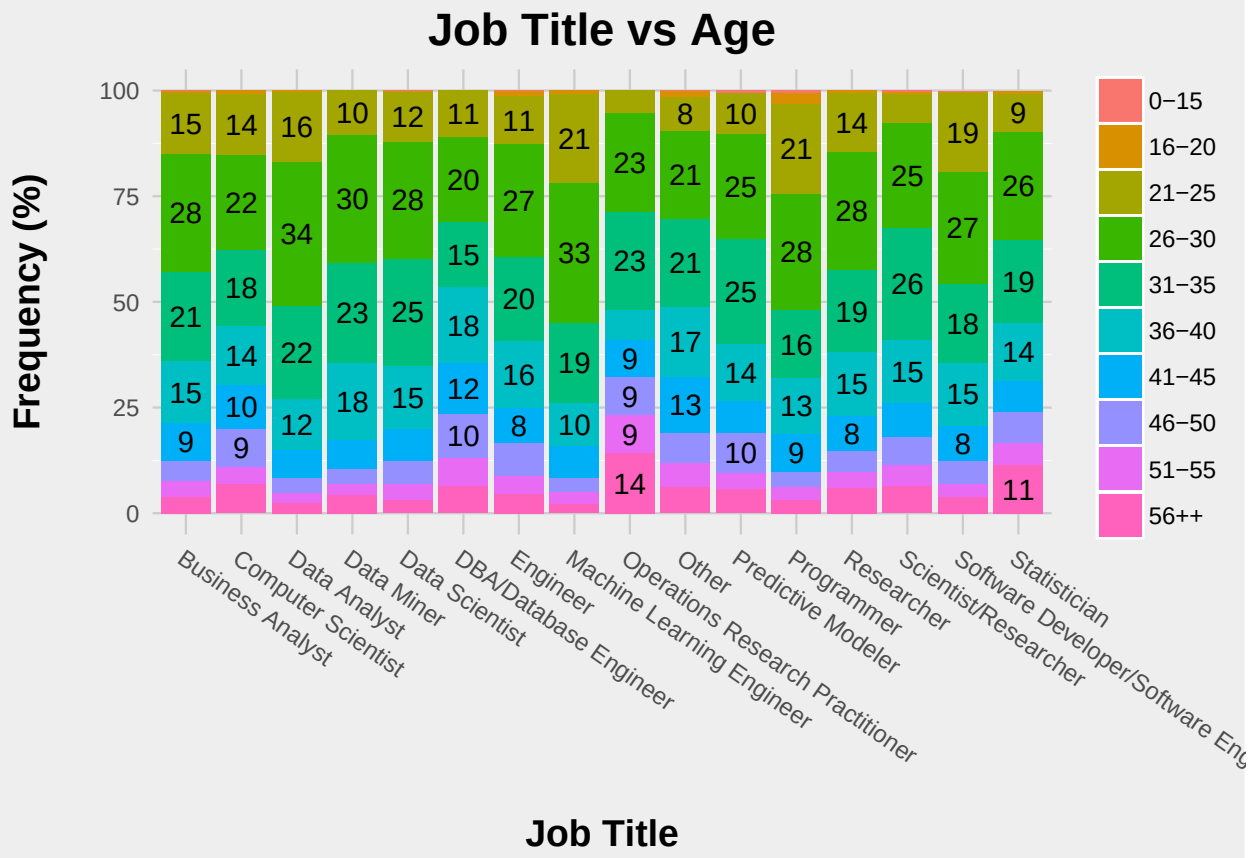


The graph of job title and learning time seems not very informative for me. Most of the job title requires less than one year to be this position. However, most of the data scientists have Master's Degree a who needs at least one year of education. There are a little portion of the employees in scientist/researcher that require 10-15 years of learning time and it kind of make sense to me. I would not take a look more in-depth at this graph and move on to other variables.

Visualization of the job title and age

```
results %>%
        mutate(catage = cut(Age, c(0,15,20,25,30,35,40,45,50,55,100), right = FALSE,
                            labels = c("0-15","16-20","21-25","26-30","31-35","36-40","41-45","46-50","5
        rename(title = CurrentJobTitleSelect ) %>%
        filter(title !="", catage != "") %>%
        group_by(title, catage) %>%
        summarise(n = n()) %>%
        mutate(freq = n/sum(n)*100) %>%
        ggplot(aes(x = title, y = freq, fill = catage, label = ifelse(freq>8, round(freq),"")))+
```

```
        ggtitle("Job Title vs Age")+
        labs(x = "Job Title", y= "Frequency (%)")+
        geom_bar(stat = "identity", position = position_stack())+
        geom_text(position = position_stack(vjust = 0.5))+
        theme1+
        theme(legend.title=element_blank())+
        theme(axis.title.x = element_text(margin = margin(t=8)),
              axis.text.x = element_text(angle = 325, hjust = 0))
```

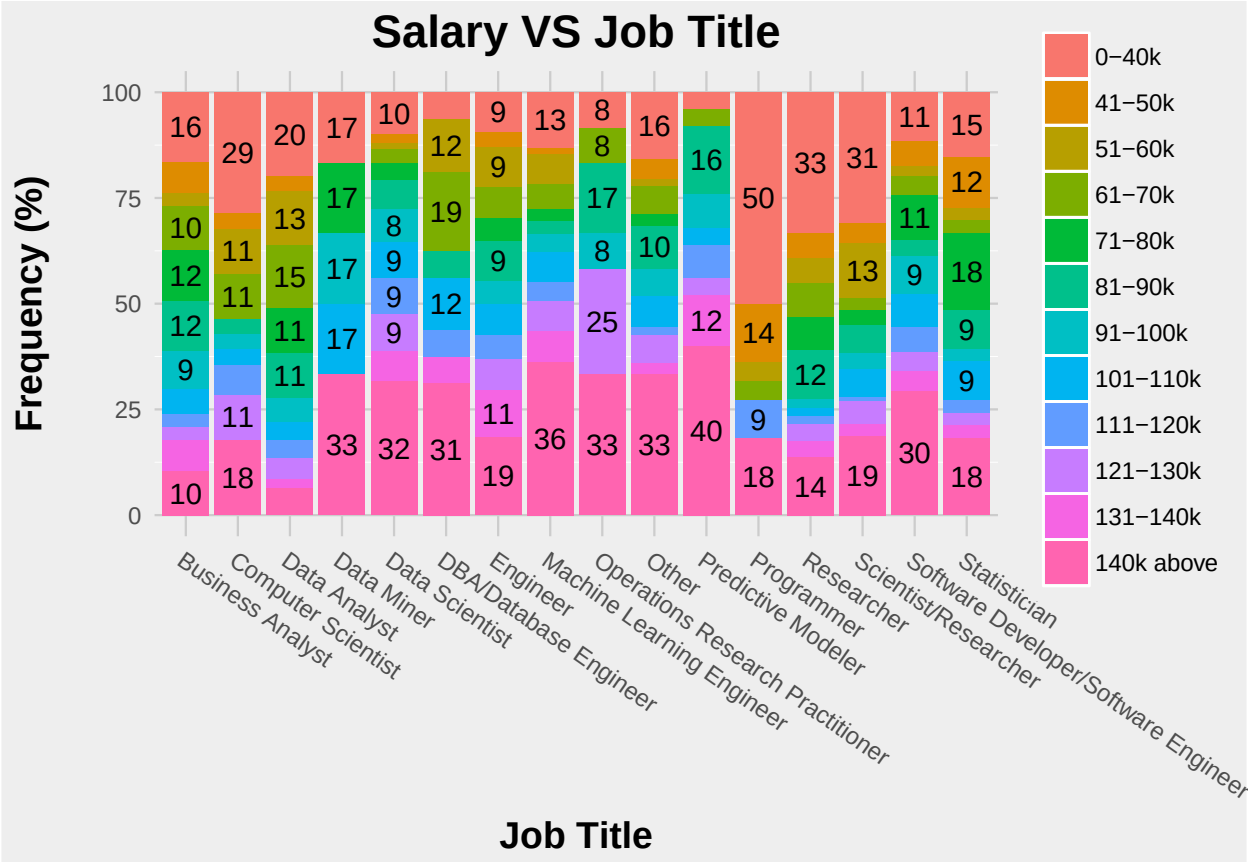


Most of the people in data scientist field are between 26-35 years old. This group age is a productive age, and the starting point of 26 is exciting because most of the Bachelor's Degree student will graduate around 22-23 years old. If they continue their Master's Degree for 1 or 2 years, they will be employed at 24-25 years old. I am wondering whether it is their first job or they do something else first and become data scientist later.

Visualization of the Job title and salary

```
results$CompensationAmount <- as.numeric(results$CompensationAmount)
results %>%
        mutate(catcomp = cut(CompensationAmount,c(0,40000,50000,60000,70000,80000,90000,100000,110000,1
        filter(CompensationCurrency=="USD") %>%
        rename(title = CurrentJobTitleSelect, salary = catcomp ) %>%
        filter(title !="", salary != "") %>%
        group_by(title, salary) %>%
        summarise(n = n()) %>%
        mutate(freq = n/sum(n)*100) %>%
```

```
        ggplot(aes(x = title, y = freq, fill = salary, label = ifelse(freq>8, round(freq),"")))+
        ggtitle("Salary VS Job Title")+
        labs(x = "Job Title", y= "Frequency (%)")+
        geom_bar(stat = "identity", position = position_stack())+
        geom_text(position = position_stack(vjust = 0.5))+
        theme1+
        theme(legend.title=element_blank())+
        theme(axis.title.x = element_text(margin = margin(t=8)),
              axis.text.x = element_text(angle = 325, hjust = 0))
```

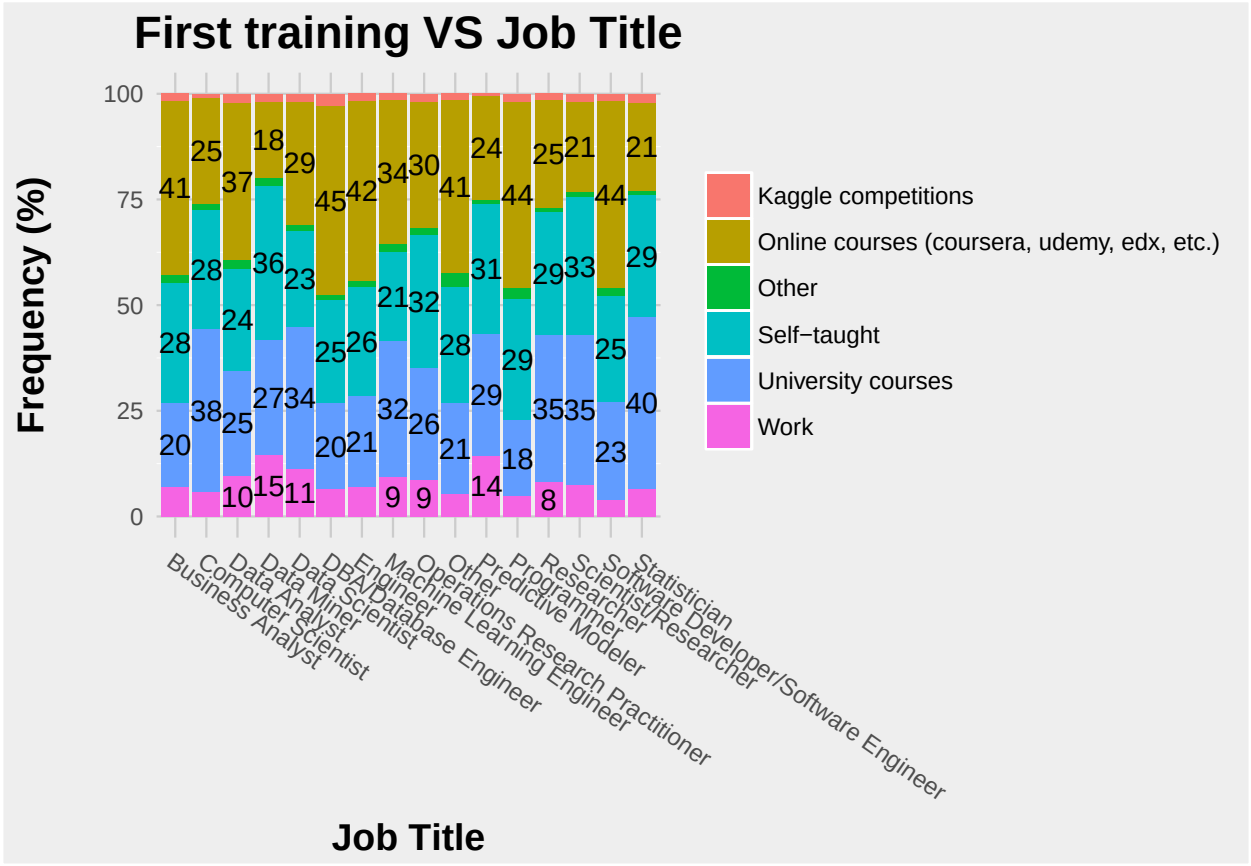


The visualization of Job Title and Salary shows that on average, data scientists get high compensation. On average 40% of the salary for all title are six digits salary. For a programmer, half of the wage is below median household salary in the US which is 55k. I am curious what type of programmers they are.

Visualization of the Job Title and First Method of Training

```
results %>%
        rename(title = CurrentJobTitleSelect, firsttrain =FirstTrainingSelect ) %>%
        filter(title !="", firsttrain != "") %>%
        group_by(title, firsttrain) %>%
        summarise(n = n()) %>%
        mutate(freq = n/sum(n)*100) %>%
        ggplot(aes(x = title, y = freq, fill = firsttrain, label = ifelse(freq>8, round(freq),"")))+
        ggtitle("First training VS Job Title")+
        labs(x = "Job Title", y= "Frequency (%)")+
```

```
        geom_bar(stat = "identity", position = position_stack())+
        geom_text(position = position_stack(vjust = 0.5))+
        theme1+
        theme(legend.title=element_blank())+
        theme(axis.title.x = element_text(margin = margin(t=8)),
              axis.text.x = element_text(angle = 325, hjust = 0))
```

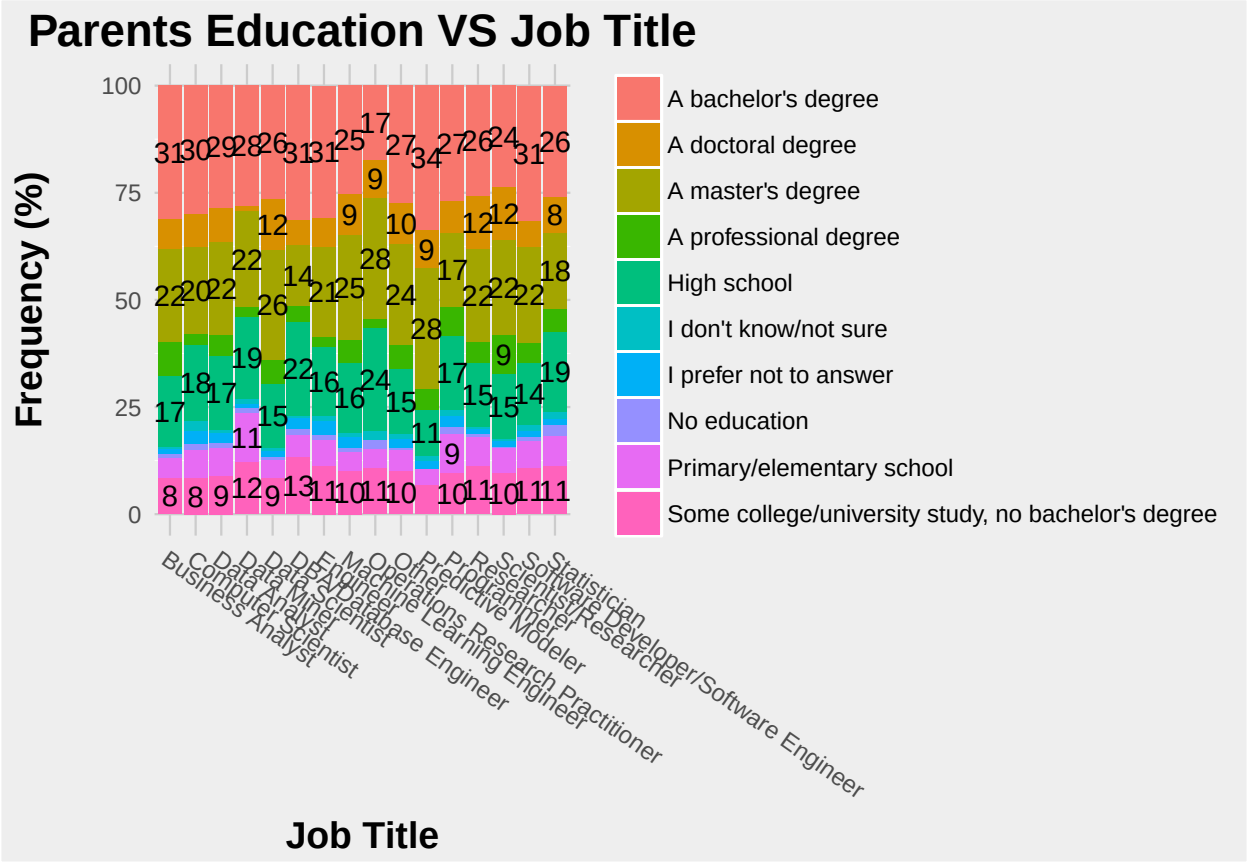


This visualization shows what kind of style of learning that they adopt when they first learn to be a data scientist. Interesting information in here is that on average less than 10% of all title learn the materials from work. It means that most of the data scientist already know what they are doing their job. In other words, what they learn to do is their job. Furthermore, most of them learn from Online courses with a small difference from different categories which are self-taught and university courses. The good news in here is that at least the materials to be a data scientist are on the internet

Visualization of the Job title and Parents Education

```
results %>%
        rename(title = CurrentJobTitleSelect, pedu =ParentsEducation ) %>%
        filter(title !="", pedu != "") %>%
        group_by(title, pedu) %>%
        summarise(n = n()) %>%
        mutate(freq = n/sum(n)*100) %>%
        ggplot(aes(x = title, y = freq, fill = pedu, label = ifelse(freq>8, round(freq),"")))+
        ggtitle("Parents Education VS Job Title")+
        labs(x = "Job Title", y= "Frequency (%)")+
```

```
        geom_bar(stat = "identity", position = position_stack())+
        geom_text(position = position_stack(vjust = 0.5))+
        theme1+
        theme(legend.title=element_blank())+
        theme(axis.title.x = element_text(margin = margin(t=8)),
              axis.text.x = element_text(angle = 325, hjust = 0))
```



I am curious to see whether there is an inclination of parent's education that makes them a data scientist. I can say that most of their parents are entirely educated that has bachelor and master's degree. The graph is too broad to conclude any information.

Visualization of the Job title and Job Satisfaction

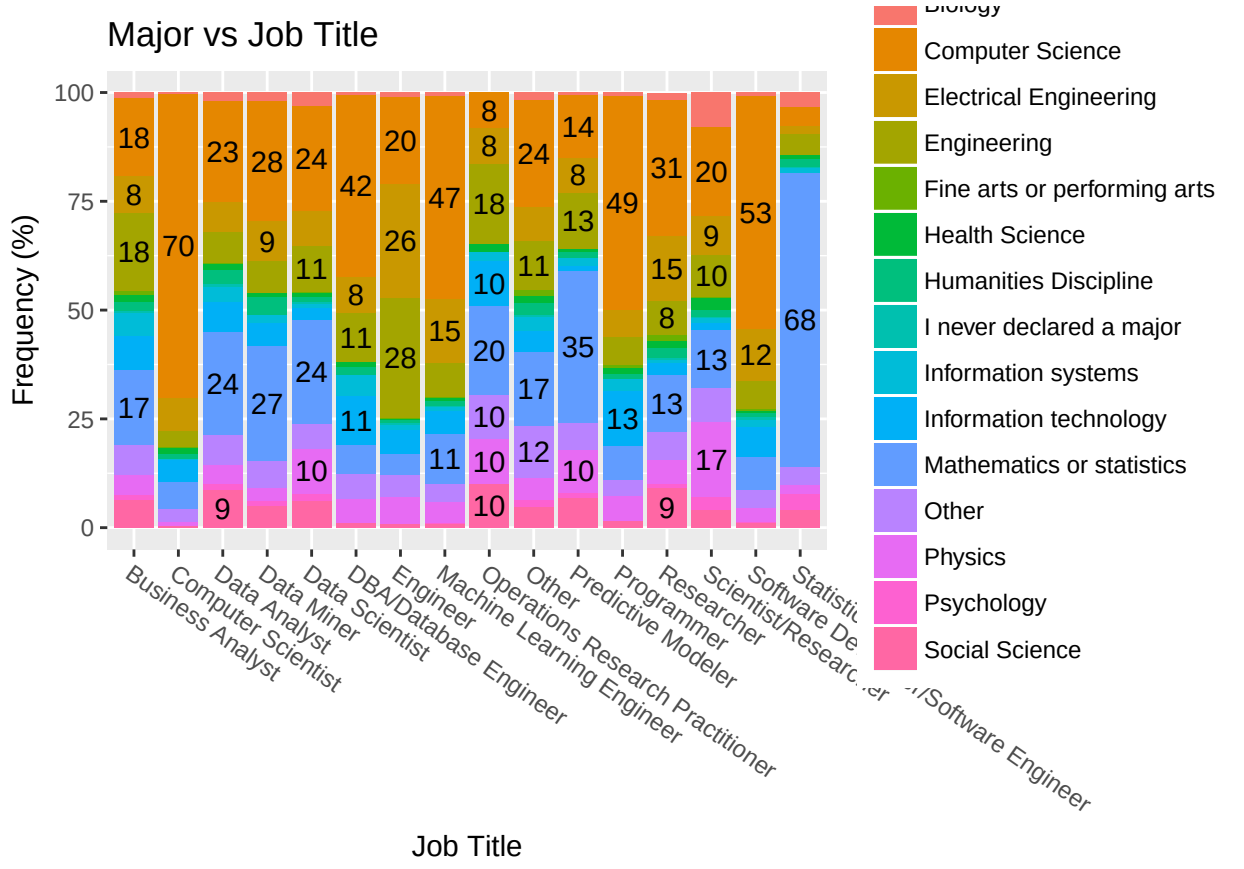
```
rename1 <- which(results$JobSatisfaction=="10 - Highly Satisfied")
     results$JobSatisfaction[rename1] <-  "9.9 - Highly Satisfied"
     results$JobSatisfaction <- as.factor(results$JobSatisfaction)
     a <- results %>%
        rename(JobTitle = CurrentJobTitleSelect ) %>%
        filter(JobTitle !="", JobSatisfaction != "") %>%
        group_by(JobTitle, JobSatisfaction) %>%
        summarise(n = n()) %>%
        mutate(Frequency = n/sum(n)*100) %>%
        ggplot(aes(x = JobTitle, y = Frequency, fill = JobSatisfaction, label = ifelse(Frequency>8, rou
        ggtitle("Job Satisfaction VS Job Title")+
        labs(x = "Job Title", y= "Frequency (%)")+
```

```
        geom_bar(stat = "identity", position = position_stack())+
        geom_text(position = position_stack(vjust = 0.5))+
        theme1+
        theme(legend.title=element_blank())+
        theme(axis.title.x = element_text(margin = margin(t=8)),
              axis.text.x = element_text(angle = 325, hjust = 0))
```

Based on the graph of Job Satisfaction, most of the data scientists are satisfied with the score around 6.5 out of 9.9.

Visualization of the Job title and Major

```
results %>%
        rename(Major = MajorSelect, title = CurrentJobTitleSelect) %>%
        filter(title !="", Major !="") %>%
        group_by(title,Major) %>%
        summarise(n = n()) %>%
        mutate(freq = n/sum(n)*100) %>%
        ggplot(aes(x = title, y = freq, fill = Major, label = ifelse(freq>8, round(freq),"")))+
        ggtitle("Major vs Job Title")+
        labs(x = "Job Title", y= "Frequency (%)")+
        geom_bar(stat = "identity", position = position_stack())+
        geom_text(position = position_stack(vjust = 0.5))+
        theme(axis.title.x = element_text(margin = margin(t=8)),
              axis.text.x = element_text(angle = 325, hjust = 0))
```



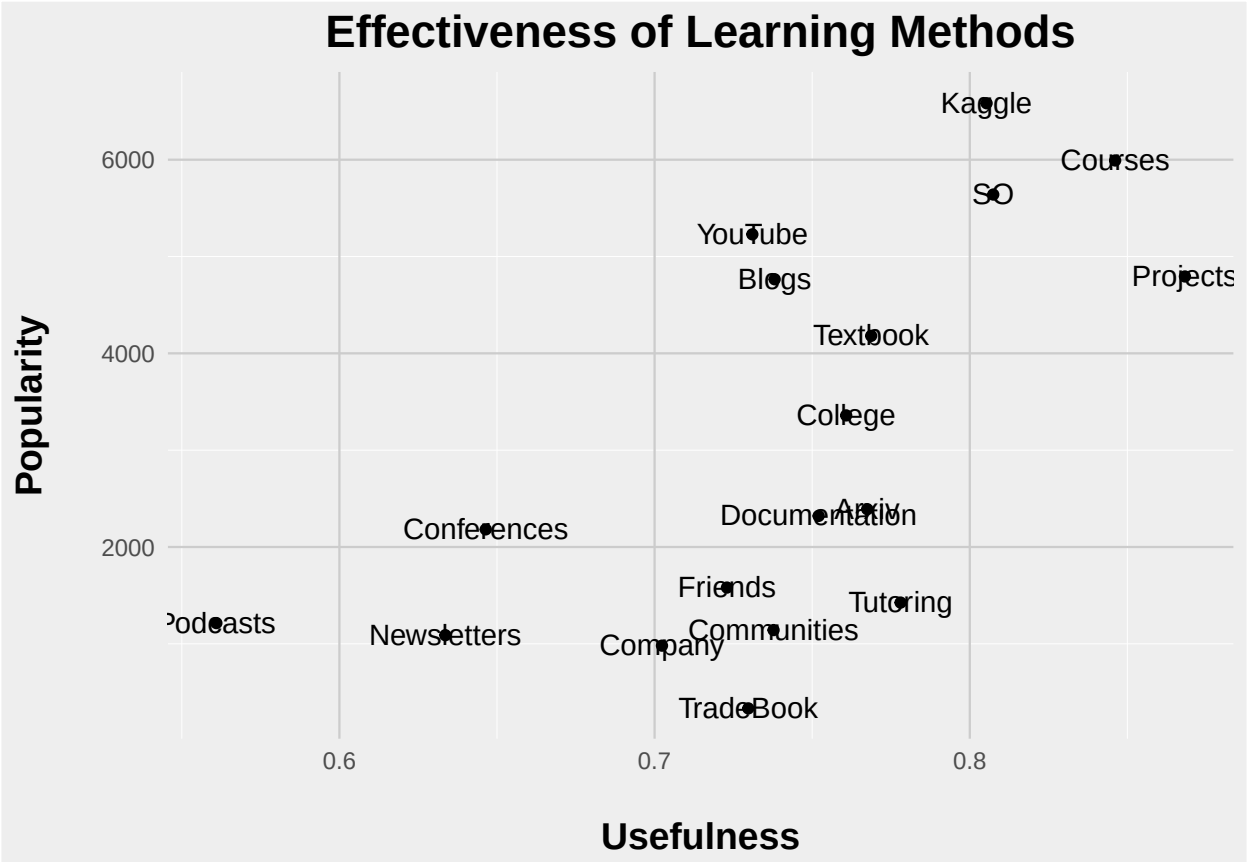
Most of the people in data scientist field come from computer science or information technology major except for those who work as a statistician. They are from math or statistics major which are quite make sense. Even though most of them are form computer science, there are around 10% people from psychology or social science that become a data scientist.

In the second section, I will still adopt the code from Jack Cook to see a more technical aspect of data science field. I create some difference in here because Jack Cook put different weight on the answer to the poll. I put not useful weight as 0, somewhat useful weight as 0.5 and very Useful weight as 1.

Visualization of Platform Usefulness

```
platforms <- grep("LearningPlatformUsefulness", names(results), value = T)

      names <- c()
      popularities <- c()
      scores <- c()

      for(k in platforms){
        usefulness <- results %>%
          group_by_(k) %>%
          count()

        popularity <- usefulness[[2]][1]+usefulness[[2]][2]+usefulness[[2]][3]

        score <- (usefulness[[2]][1]*0+usefulness[[2]][2]*0.5+usefulness[[2]][3]*1)/popularity
```

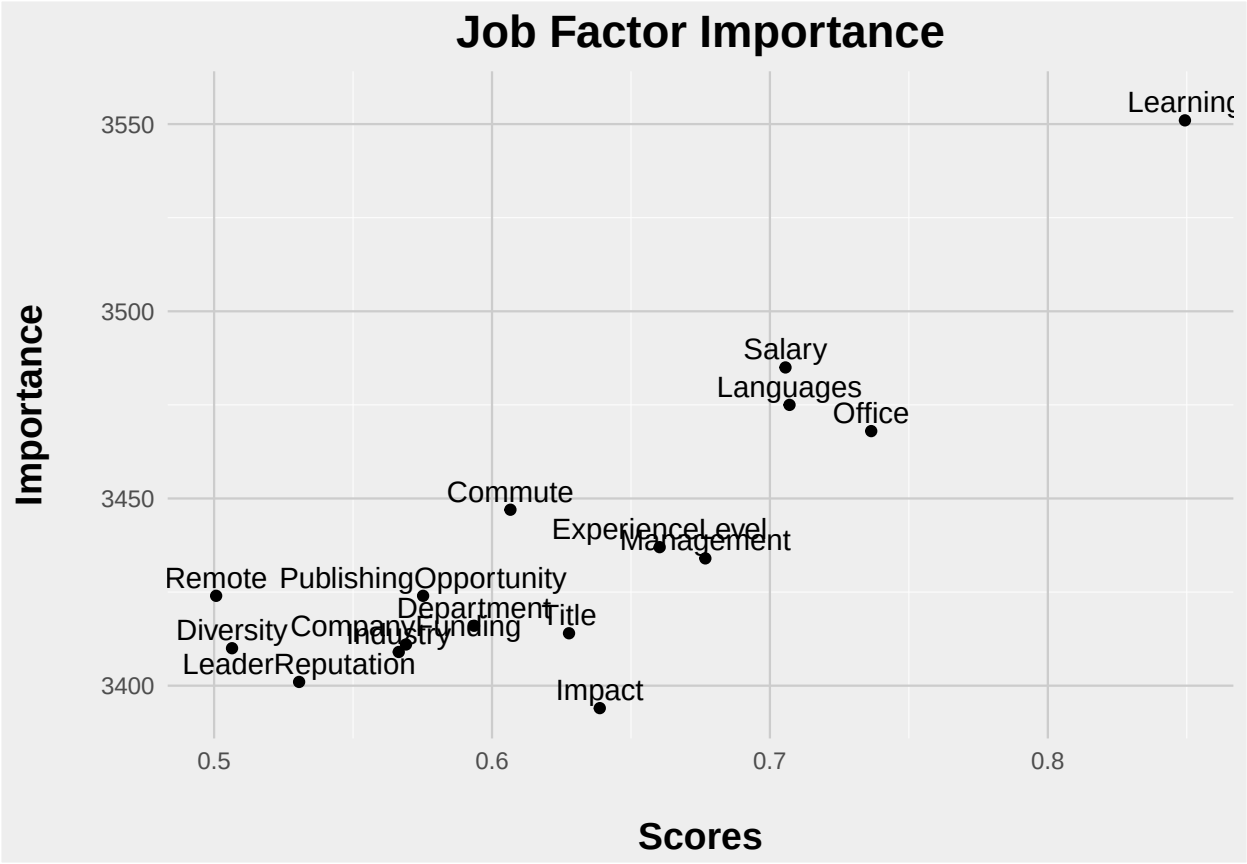
```
    names <- c(names, gsub("LearningPlatformUsefulness","",k))
    popularities <- c(popularities, popularity)
    scores <- c(scores, score)
  }
  scores_df <- data.frame(
    Popularity = popularities,
    Usefulness = scores,
    Name = names
  )

result2a <-  ggplot(scores_df, aes(x = Usefulness, y = Popularity)) +
    ggtitle("Effectiveness of Learning Methods") +
    geom_point() +
    geom_text(aes(label = Name, family = "Helvetica"), nudge_y = 10)+
    theme1
result2a
```



Based on the visualization above, Projects are very useful and very popular to learn about data scientist. Personally, I agree with this poll because I learn a lot when I do something. Furthermore, it's followed by Courses, Kaggle and SO. It seems that online community and projects are the best combinations to learn to be a data scientist.

Visualization of Job Factor Importance

```
JobFactor <- grep("JobFactor", names(results), value = T)
      JobSkillImportance <- grep("JobSkillImportance", names(results), value = T)
      WorkChallengeFrequency<- grep("WorkChallengeFrequency", names(results), value = T)
      names.j <- c()
      importances <- c()
      scores.j <- c()

      for (k in JobFactor){
        weighted <- results %>%
          group_by_(k) %>%
          count()
        importance <- weighted[[2]][1]+weighted[[2]][2]+weighted[[2]][3]
        score <- (weighted[[2]][1]*0+weighted[[2]][2]*0.5+weighted[[2]][3]*1)/importance

        names.j <- c(names.j, gsub("JobFactor","",k))
        importances <- c(importances, importance)
        scores.j <- c(scores.j,score)
      }

      scores.df.j <- data.frame(
        Importance = importances,
        Scores = scores.j,
        Name = names.j
      )

     result2a <-  scores.df.j %>%
        ggplot(aes(x= Scores, y= Importance ))+
        ggtitle("Job Factor Importance") +
        geom_point()+
        geom_text(aes(label = Name), nudge_y  = 5)+
        theme1

     result2a
```

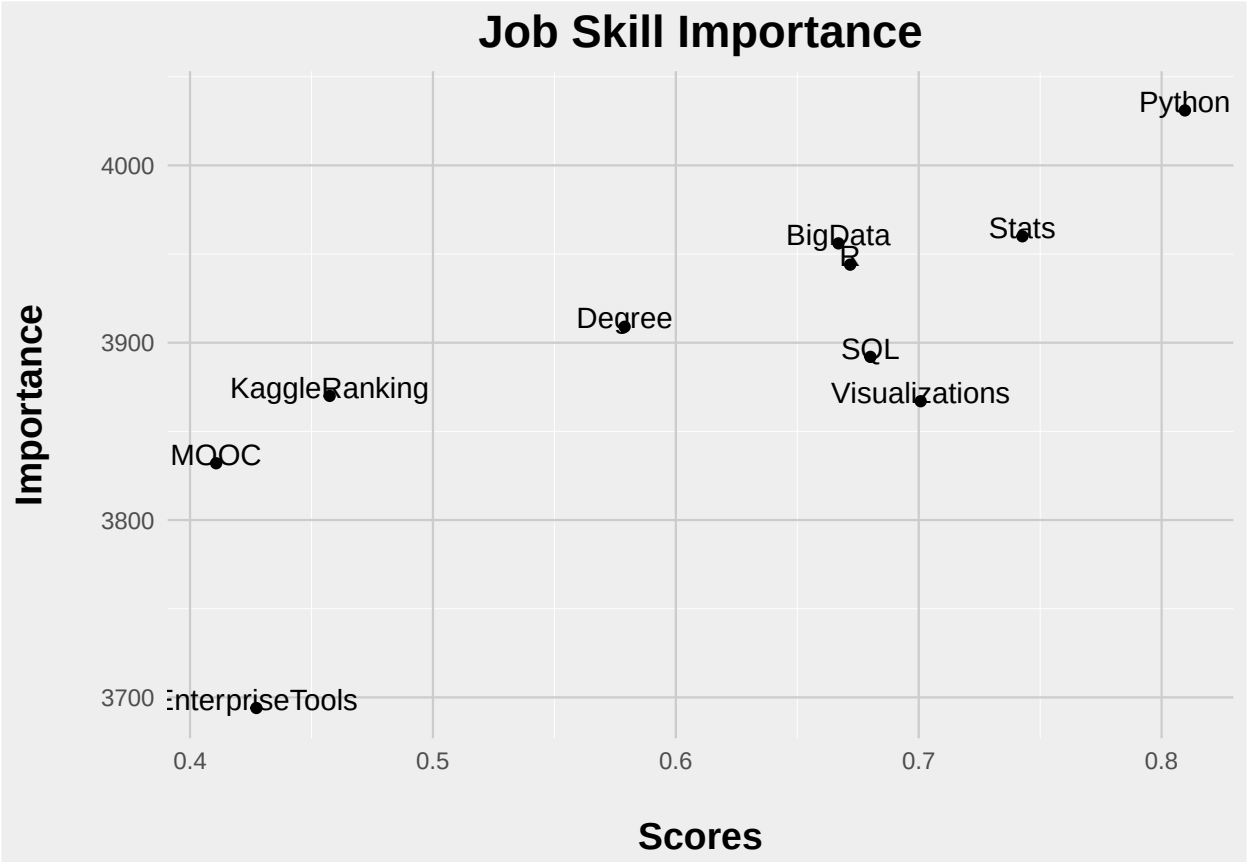


Surprisingly, learning becomes a significant factor for data scientists to remain on their job. It follows by Salary, languages, and Office. I don't get about the language, but from what I extract, it seems that data scientist loves to learn and get paid well.

Visualization of Job Skill Importance

```
JobSkillImportance <- grep("JobSkillImportance", names(results), value = T)
      names.js <- c()
      importances.js <- c()
      scores.js <- c()

      for (k in JobSkillImportance){
        weightedjs <- results %>%
          group_by_(k) %>%
          count()
        importance.js <- weightedjs[[2]][1]+weightedjs[[2]][2]+weightedjs[[2]][3]
        score.js <- (weightedjs[[2]][1]*1+weightedjs[[2]][2]*0.5+weightedjs[[2]][3]*0)/importance.js
        names.js <- c(names.js, gsub("JobSkillImportance","",k))
        importances.js <- c(importances.js, importance.js)
        scores.js <- c(scores.js, score.js)
      }

      scores.df.js <- tibble(
        Importance = importances.js,
        Scores = scores.js,
```

```
    Name = names.js
  )
  scores.df.js <- scores.df.js[-c(11:13),]

  result2a <- scores.df.js %>%
    ggplot(aes(x= Scores, y= Importance ))+
    ggtitle("Job Skill Importance") +
    geom_point()+
    geom_text(aes(label = Name), nudge_y  = 5)+
    theme1
  result2a
```

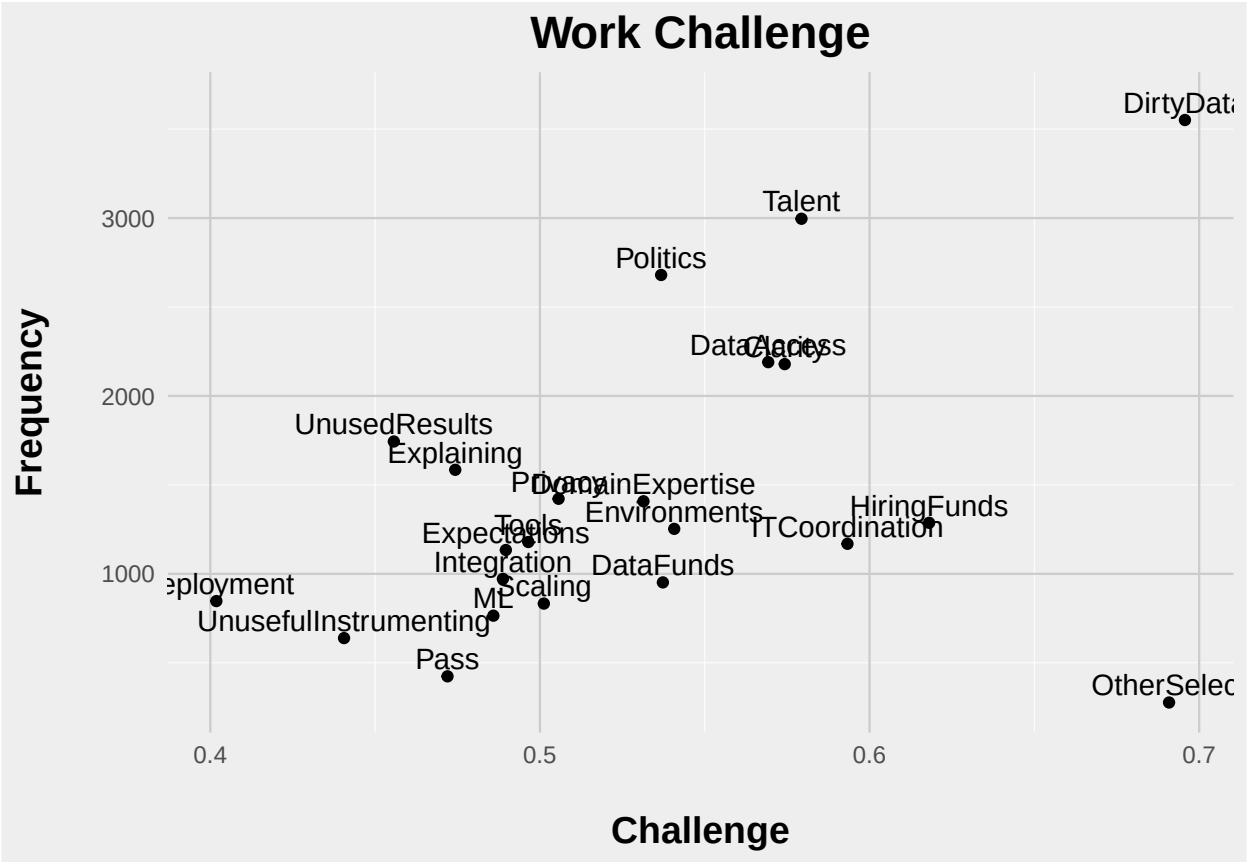


According to the survey, it is important to have skill in python as a data scientist followed by statistical skill. Then it is good to know about R, SQL, BigData, and Visualization. Technically a data scientist has to know everything. However, from this graph we can choose which one to start and how are we going to proceed to the next skill. In contrast, the degree is not important for this field with the score 0.58.

Visualization of Work Challenge Frequency

```
WorkChallengeFrequency<- grep("WorkChallengeFrequency", names(results), value = T)
  names.wc <- c()
  importances.wc <- c()
  scores.wc <- c()

  for (k in WorkChallengeFrequency){
```

```
    weighted <- results %>%
      group_by_(k) %>%
      count()
    importance.wc <- weighted[[2]][1]+weighted[[2]][2]+weighted[[2]][3]+weighted[[2]][4]
    score.wc <- (weighted[[2]][1]*1+weighted[[2]][2]*0.66+weighted[[2]][3]*0.33+weighted[[2]][4]*0)/
    names.wc <- c(names.wc, gsub("WorkChallengeFrequency","",k))
    importances.wc <- c(importances.wc, importance.wc)
    scores.wc <- c(scores.wc, score.wc)
  }

  scores.df.wc <- tibble(
    Frequency = importances.wc,
    Challenge = scores.wc,
    Name = names.wc
  )

  result2a <- scores.df.wc %>%
    ggplot(aes(x= Challenge, y= Frequency ))+
    ggtitle("Work Challenge") +
    geom_point()+
    geom_text(aes(label = Name), nudge_y  = 100)+
    #geom_point(aes(x= Scores, y= Frequency , color="blue"), data= scores.df.wm)+
    #geom_text(aes(label = Name), data= scores.df.wm, nudge_y  = 100, color = "blue")+
    theme1
result2a
```



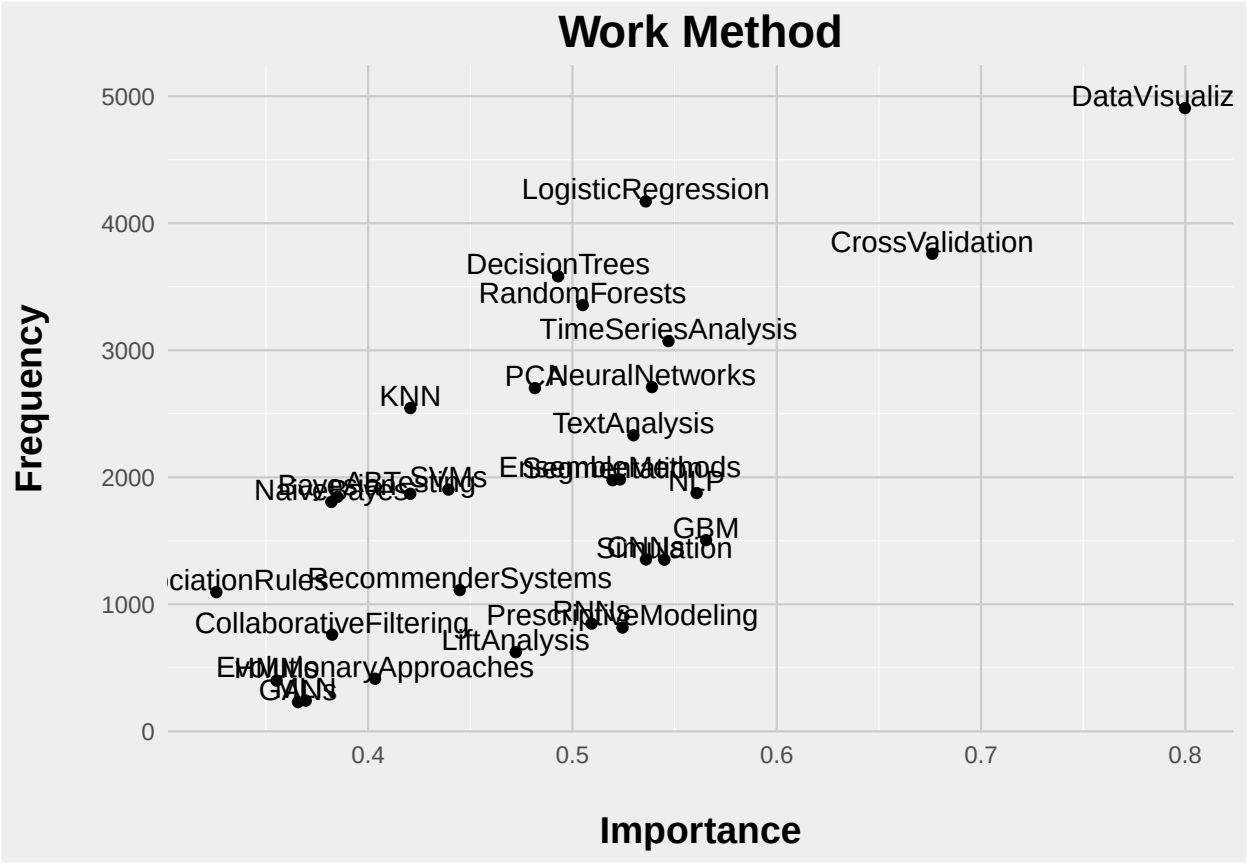
When it comes to the challenge as a data scientist, dirty data is the most frequent and most challenging in this position. Talent is less challenging but quite frequent and similar to the Hiring funds. I am curious for those who answer Other Select. This category is quite challenging even though it is not often.

Visualization of Work Methods Frequency

```
      WorkMethodsFrequency<- grep("WorkMethodsFrequency", names(results), value = T)

      names.wm <- c()
      importances.wm <- c()
      scores.wm <- c()

      for (k in WorkMethodsFrequency){
        weighted <- results %>%
          group_by_(k) %>%
          count()
        importance.wm <- weighted[[2]][1]+weighted[[2]][2]+weighted[[2]][3]+weighted[[2]][4]
        score.wm <- (weighted[[2]][1]*1+weighted[[2]][2]*0.66+weighted[[2]][3]*0.33+weighted[[2]][4]*0)
        names.wm <- c(names.wm, gsub("WorkMethodsFrequency","",k))
        importances.wm <- c(importances.wm, importance.wm)
        scores.wm <- c(scores.wm, score.wm)
      }
      weighted[[1]]
```

```
## [1] "Most of the time" "Often"            "Rarely"
## [4] "Sometimes"        NA
```

```
      scores.df.wm <- tibble(
        Frequency = importances.wm,
        Importance = scores.wm,
        Name = names.wm
      )
      scores.df.wm <- scores.df.wm[-c(31:33),]

      result2a <- scores.df.wm %>%
        ggplot(aes(x= Importance, y= Frequency ))+
        ggtitle("Work Method") +
        geom_point()+
        geom_text(aes(label = Name), nudge_y  = 100)+
        theme1
    result2a
```

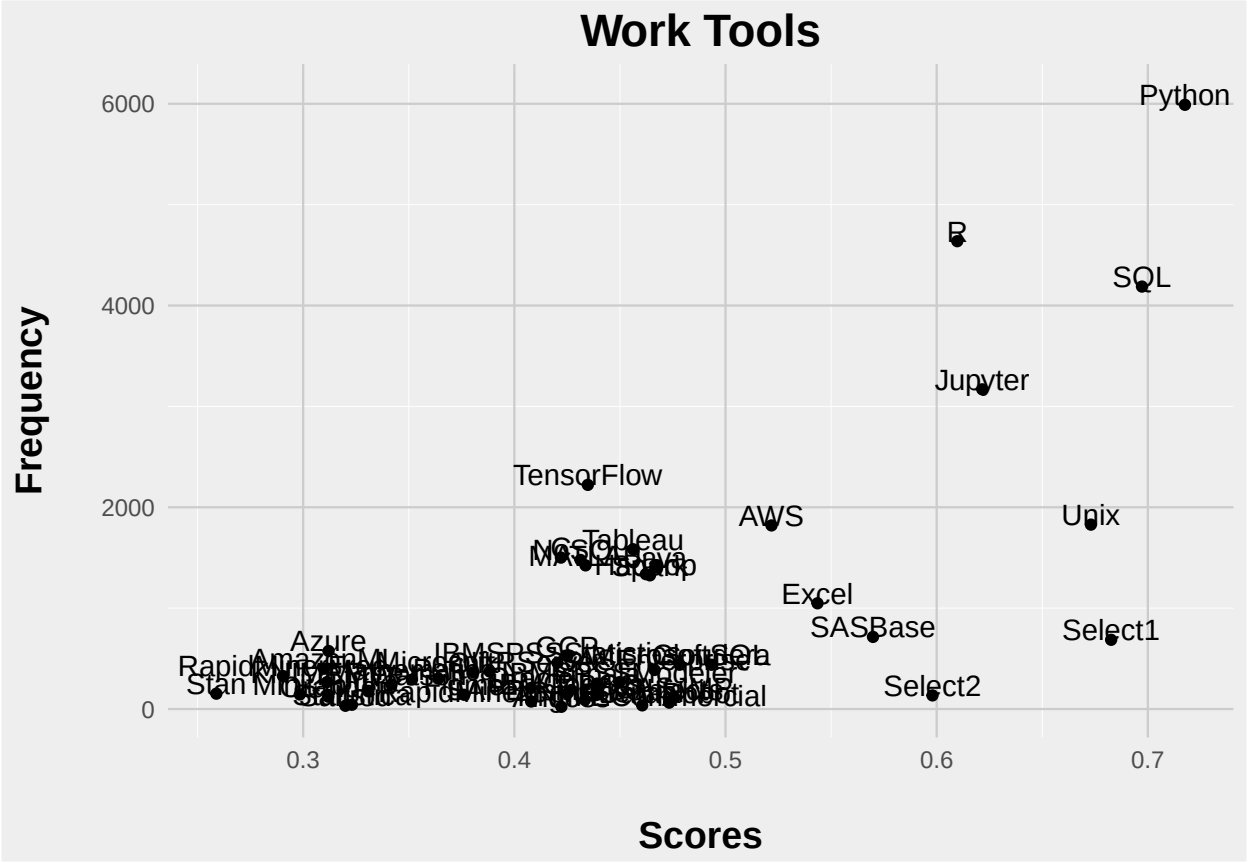


Based on the work method importance and frequency. Data visualization is the most important and most frequently used by data scientist followed by Cross-Validation. It kind of make sense because data visualization and cross-validation are easy to understand. A graph is easy to read, and the cross-validation method compares between prediction rate or means squared error. Interesting information in this graph is most of the method used by data scientist are classification methods such as logistic regression, decision tree, and random forest. I don't know whether Kaggle didn't ask them about the quantitative method or they don't use a quantitative method such as regression on their job.

Visualization of Work Tools Frequency

```
WorkToolsFrequency<- grep("WorkToolsFrequency", names(results), value = T)

      names.wt <- c()
      importances.wt <- c()
      scores.wt <- c()

      for (k in WorkToolsFrequency){
        weighted <- results %>%
          group_by_(k) %>%
          count()
        importance.wt <- weighted[[2]][1]+weighted[[2]][2]+weighted[[2]][3]+weighted[[2]][4]
        score.wt <- (weighted[[2]][1]*1+weighted[[2]][2]*0.66+weighted[[2]][3]*0.33+weighted[[2]][4]*0)
        names.wt <- c(names.wt, gsub("WorkToolsFrequency","",k))
        importances.wt <- c(importances.wt, importance.wt)
        scores.wt <- c(scores.wt, score.wt)
```

```
      }

      scores.df.wt <- tibble(
        Frequency = importances.wt,
        Scores = scores.wt,
        Name = names.wt
      )

     result2a <-  scores.df.wt %>%
        ggplot(aes(x= Scores, y= Frequency ))+
        ggtitle("Work Tools") +
        geom_point()+
        geom_text(aes(label = Name), nudge_y  = 100)+
        theme1
result2a
```



This graph goes hand in hand with the previous graph about Job Skill Importance. It is obvious of python is an important skill, they will use python to work.

Conclusion

Data scientist are the field where most of the employee has a higher degree such as Master of Doctoral degree. They are mostly 26-45 years old. This field pays well with most of the employee get six digits salary. Most of them are from computer science major, but some employees come from social science. They learn to be data scientist by courses both online and university. They said that doing projects are the most useful and popular

method to master this field. They choose this field because they love to learn, and they get paid well. Four important skills in this field are Python, Statistical, R, and Big Data. Their most significant challenge is Dirty Data. They mostly used Data Visualization, Cross-Validation, and Classification method on their job.