Insights Of The Current Data Science Job Market

Yu Han Huang (yh3093), Deepak Ravishankar (dr2998), Jong Hyuk Lee (jl5261)

- 1 Introduction
  - 1.1 General Info
  - 1.2 Study Objective
  - 1.3 Link to repository:
  - 1.4 Team Contribution
  - 1.5 Special note : Colorblind-friendly
  - 1.6 Packages used
- 2 Description of Data
  - 2.1 Supply - Stack Overflow 2018 Developer Survey
  - 2.2 Supply - Kaggle ML and Data Science Survey, 2018
  - 2.3 Demand - Rachel's Mail - Columbia University Data Science Career Opportunities
- 3 Analysis of Data Quality
- 4 Main analysis (Exploratory Data Analysis)
  - 4.1 Who's in the data science job market
    - 4.1.1 Age
    - 4.1.2 Gender
    - 4.1.3 Age by gender
    - 4.1.4 Represented countries
    - 4.1.5 Education and industry
  - 4.2 What skills are in the data science job market (both demand and supply)
    - 4.2.1 General Skills
    - 4.2.2 Technical Skills
    - 4.2.3 Section Conclusion
  - 4.3 What prospects can we expect from the market - Salary analysis
    - 4.3.1 Median salary by location
    - 4.3.2 Median salary by programming language used
    - 4.3.3 Median salary by database used
    - 4.3.4 Median salary by platforms used
    - 4.3.5 Section Conclusion
  - 4.4 Review: Challenges Faced in Exploratory Data Analysis
- 5 Executive summary (Presentation-style)
- 6 Interactive component
- 7 Conclusion
  - 7.1 Key findings
  - 7.2 Summary of findings
  - 7.3 Limitations and future directions

1 Introduction

1 Introduction

1.1 General Info

Data Scientist has been named **The Best Job In America** from Glassdoor's 50 Best Jobs In America For 2018 Survey, and in fact, it has been ranked as the number one profession by Glassdoor three years in a row.

Being one of the hottest careers of this century, there is a huge demand for data scientists in almost every job sectors. Annual demand for the fast-growing new roles of data scientist, data developers, and data engineers is projected to reach nearly 700,000 openings by 2020. IBM predicts demand for data scientists to soar 28% by 2020, that the number of jobs for all US data professionals will increase by 364,000 openings to 2,720,000.

However, data science is more than just a buzz word but a field of research and an area of expertise. The shortage of qualified and experienced professionals has created a unique opportunity for those looking to break into a data scientist career. While five years ago there was no such thing as a data science degree, the list of schools opening up data science programs now runs for pages and pages. There are also a lot of online data science courses offered to people who want to switch careers.

1.2 Study Objective

As students current majoring in Data Science at Columbia, we are interested in the swelling demand for data scientists coupled with the evident skills gap. We want to understand the demographic of the current data science job market: who is in the market, what skills are needed in the market, and what prospects can we expect from the market. For each variable, we would evaluate it from both demand and supply, that is, we would not only evaluate the demanded requirements from job listings, but how people in the industry are meeting the requirements.

The questions we are interested in are:
- Who's in the data science job market: Age, Gender, Country, Education, Industry Backgrouds
- What skills are in the data science job market (both demand and supply): General skills, Technical skills
- What prospects can we expect from the market: Prospective salary

1.3 Link to repository:

Main repository
Interactive Component repository
Website

1.4 Team Contribution

The respected contribution from each team members are:

- *Yu Han Huang (yh3093):* Data collection and analysis on Rachel's Mail - Columbia University Data Science Career Opportunities, Report Making, Creating presentation, Presenting the presentation, Being a good team member
- *Deepak Ravishankar (dr2998):* Data collection and analysis on Kaggle ML and Data Science Survey, 2018, Designing Interactive Component, Creating presentation, Presenting the presentation, Being a good team member
- *Jong Hyuk Lee (jl5261):* Data collection and analysis on Stack Overflow 2018 Developer Survey, Assitant in designing Interactive Component, Creating presentation, Presenting the presentation, Being a good team member

1.5 Special note : Colorblind-friendly

This study is also colorblind-friendly. For ggplot2 plots, we use a colorblind-friendly palette from Color Universal Design (CUD) - How to make figures and presentations that are friendly to Colorblind people. For highchart plots, the default palette of Highcharts is designed with accessibility in mind, so that any two neighbor colors are tested for different types of color blindness. Therefore, we do not alter the default color of highchart.

Highchart Color Blindness Accessibility

```
cbPalette <- c("#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00",
"#CC79A7")
```

1.6 Packages used

Loading packages:

```
library(tidyverse)
library(highcharter)
library(readr)
library(dplyr)
library(stringr)
library(extracat)
library(tidytext)
library(tidyr)
library(wordcloud)
library(tidyverse)
library(ggplot2)
library(plotly)
library(viridis)
library(gridExtra)
library(dygraphs)
library(lubridate)
library(countrycode)
library(leaflet)
library(xts)
library(htmltools)
library(data.table)
library(forcats)
library(corrplot)
#for the maps
library(geosphere) # geospatial locations
library(leaflet.extras) # maps
library(maps) # maps
library(geofacet)
```

2 Description of Data

Three datasets are used in this study, two representing the supply side and one representing the demand side. In below we will go through the source of each dataset and some noteworthy features of the data.

2.1 Supply - Stack Overflow 2018 Developer Survey

This is a survey conducted by Stack Overflow in 2018 (data can be found here) on 98,855 developers about how they learn, build their careers, which tools they're using, and what they want in a job. For our study preliminary on the field of data science, we filtered the data by the developers with the title **Data or business analyst** or **Data scientist or machine learning specialist**.

We use this data as a representative of data scientists in **U.S job market.**

```
#read the file
stackoverflow <-  read_csv("C:/Users/Flora Huang/Downloads/developer_survey_2018/survey_results_public.csv")
```

```
stackoverflow <-  stackoverflow %>% select(`LanguageWorkedWith`,`DevType`,`DatabaseWorkedWith`,`PlatformWorkedWith`,`Country`,`ConvertedSalary`,`Employment`)
```

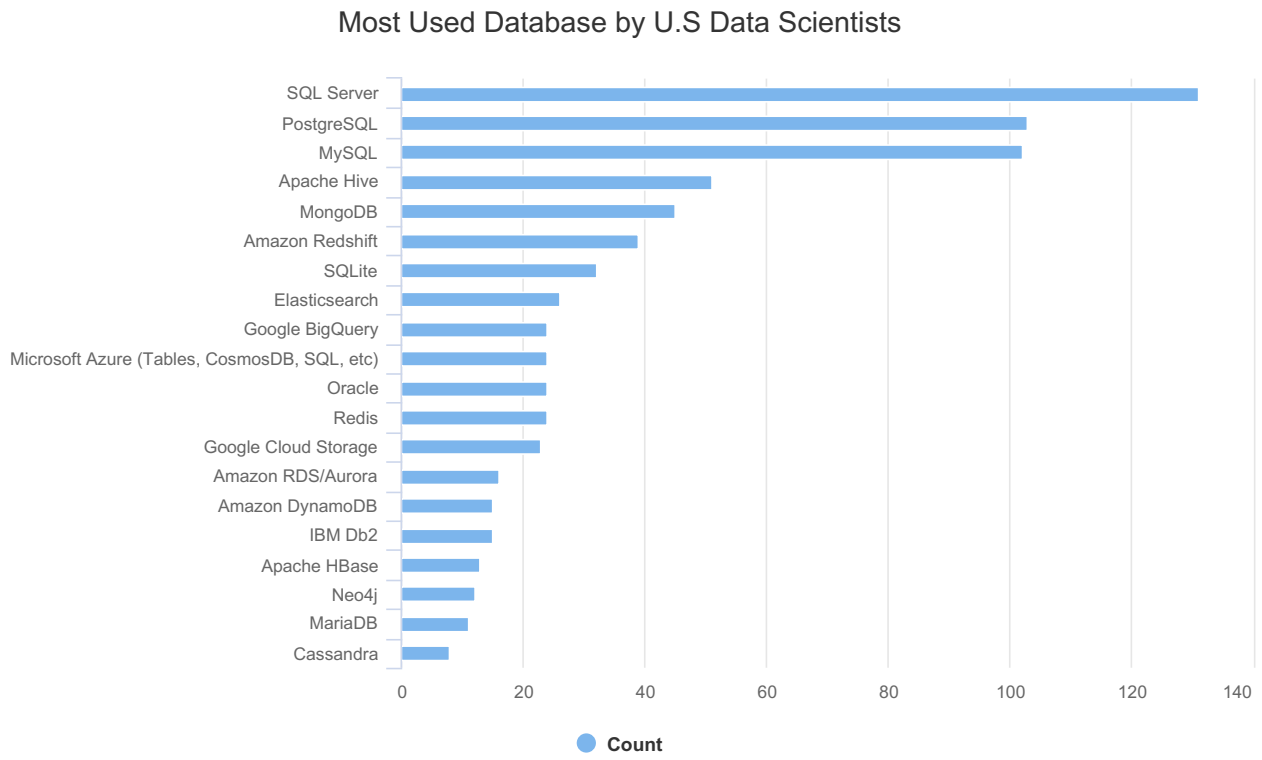
Only the columns in the above table will be used in the analysis of this data set. The LanguageWorkedWith, DatabaseWorkedWith, and PlatformWorkedWith have data exactly corresponding to their column names and are seperated by semi-colons. DevType have developer type data, where we used it to filter data scientis by finding phrases, 'Data or business analyst' and 'Data scientist or machine learning specialist'. Country column contains country names, ConvertedSalary column contains salary data all converted to USD, and Employement column shows whether a person is a full-time or a part-time employee. We will only focus on the full-time employees.

Some noteworthy features of Stack Overflow 2018 Developer Survey:

In a basic analysis of Stack Overflow's survey, we found some noteworthy and interesting features of the data.

- **A. Most Used Database by U.S Data Scientists**

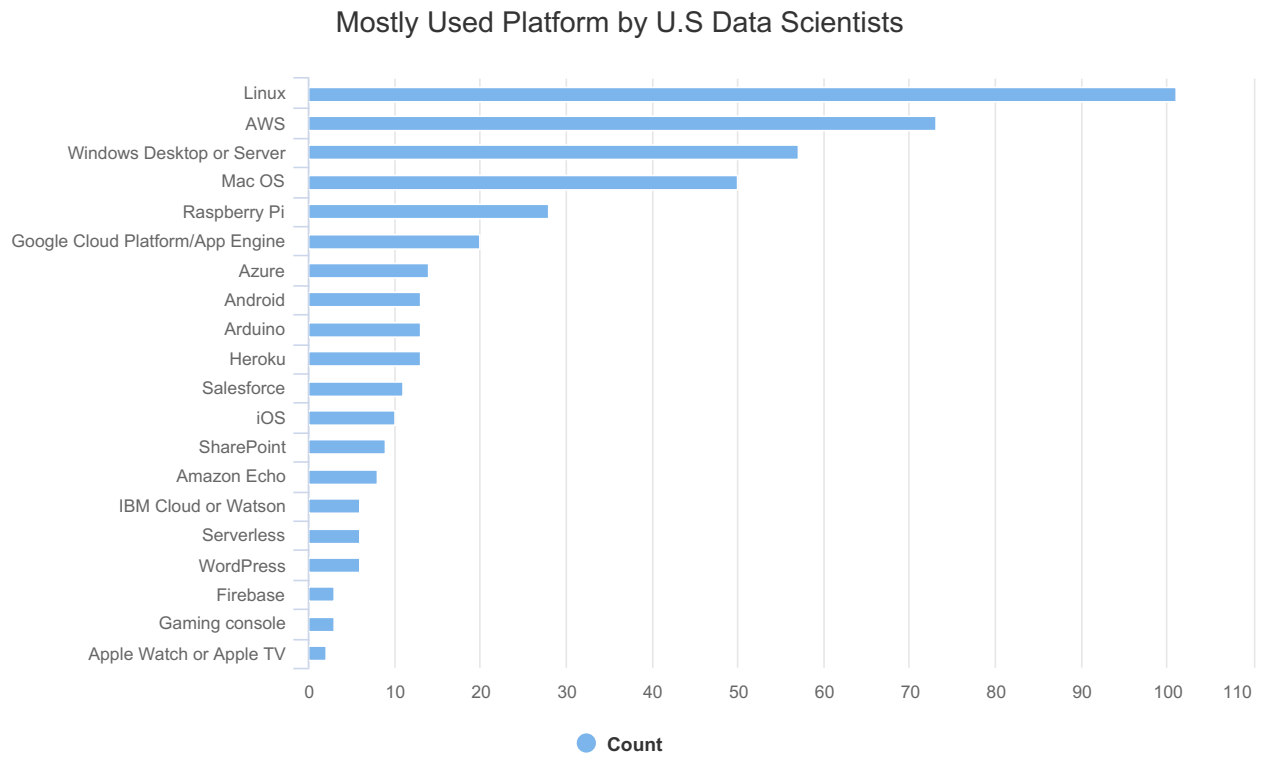
```r
#This code graphs a histogram by mostly used database in decreasing order.
#Filters full-time workers, Data Scientists, US workers, and their missing data

languae_hist = stackoverflow%>%
  filter(Employment %in% 'Employed full-time') %>%  #filter full-time employees
  filter(!is.na(DatabaseWorkedWith))%>% #filter missing data
  filter(!is.na(DevType)) %>%
  filter(DevType %in% c('Data or business analyst','Data scientist or machine learning specialist'))
%>% #filter data scientists
  filter(Country %in% 'United States') %>% #filter US workers
  select(DatabaseWorkedWith)%>%
  mutate(DatabaseWorkedWith = str_split(DatabaseWorkedWith, pattern = ";"))%>%
  unnest(DatabaseWorkedWith)%>%
  group_by(DatabaseWorkedWith)%>%
  summarise(Count = n())%>% #count by database
  arrange(desc(Count))%>% #reorder in descending order
  ungroup()%>%
  mutate(DatabaseWorkedWith = reorder(DatabaseWorkedWith, Count))


#slice only top 20 data
languae_hist = slice(languae_hist, 0:20)

#plot in histogram
highchart()%>% #
  hc_title(text = paste("Most Used Database by U.S Data Scientists"))%>% #title
  hc_xAxis(categories = languae_hist$DatabaseWorkedWith)%>% #xaxis
  hc_add_series(data = languae_hist$Count, name = "Count", type = "bar") #plot
```



This histogram tries to find which database language data scientists use the most. This data filtered full-time employees, data scientists, and US, which consists of 259 rows. The X-axis is count, and the Y-axis denotes each database languages sorted in decreasing order of the counts. It is important to note that the filtered data pertain to data scientists only and does not include database managers. The results shows data scientists tend to prefer the most popular database languages, and SQL is very dominant in this term.

- **B. Most Used Platform by U.S Data Scientists**

```r
#This code graphs a histogram by mostly used platforms in decreasing order.
#Filters full-time workers, Data Scientists, US workers, and their missing data

languae_hist = stackoverflow %>%
  filter(Employment %in% 'Employed full-time') %>%  #filter full-time employees
  filter(!is.na(PlatformWorkedWith))%>% #filter missing data
  filter(!is.na(DevType)) %>%
  filter(DevType %in% c('Data or business analyst','Data scientist or machine learning specialist'))
%>% #filter data scientists
  filter(Country %in% 'United States') %>% #filter US workers
  select(PlatformWorkedWith)%>%
  mutate(PlatformWorkedWith = str_split(PlatformWorkedWith, pattern = ";"))%>%
  unnest(PlatformWorkedWith)%>%
  group_by(PlatformWorkedWith)%>%
  summarise(Count = n())%>% #count by languages
  arrange(desc(Count))%>% #reorder in descending order
  ungroup()%>%
  mutate(PlatformWorkedWith = reorder(PlatformWorkedWith, Count))


#slice only top 20 data
languae_hist = slice(languae_hist, 0:20)

#plot
highchart()%>% #
  hc_title(text = paste("Mostly Used Platform by U.S Data Scientists"))%>% #title
  hc_xAxis(categories = languae_hist$PlatformWorkedWith)%>% #xaxis
  hc_add_series(data = languae_hist$Count, name = "Count", type = "bar") #plot
```



This is a histogram of platform usage by data scientists. Again, it filters platform, full-time, and US with 200 rows. The X-axis is count, and the Y-axis denotes each platforms sorted in decreasing order of the counts. There is nothing particular interesting about this result, since the three widely used platforms, Linux, mac, and Windows are also most popular platforms for data scientists.

2.2 Supply - Kaggle ML and Data Science Survey, 2018

This is an industry-wide survey on 23,859 data scientists and machine learning engineers conducted by Kaggle. (data can be found here) The survey looks into who is working with data, what's happening at the cutting edge of machine learning across industries, and how new data scientists can best break into the field, which matches our objective of this study.

We use this data as a representative of data scientists in **international job market.**

```
#read data
df_survey_mcq <-  as.tibble(fread(str_c("C:/Users/Flora Huang/Desktop/Exploratory Data Analysis and Visualization/Final project/code/data/multipleChoiceResponses.csv"),na.strings = "-1"))
df_survey_free <- as.tibble(fread(str_c("C:/Users/Flora Huang/Desktop/Exploratory Data Analysis and Visualization/Final project/code/data/freeFormResponses.csv"),na.strings = "-1"))
```

While the survey looks into 228 features of respondents' job, we only select the features that we are interested in, which are:

- In which country do you currently reside?
- What is your gender?
- What is your gender? - Prefer to self-describe - Text",
- What is your age (# years)?
- What is the highest level of formal education that you have attained or plan to attain within the next 2 years?
- Which best describes your undergraduate major? - Selected Choice
- Select the title most similar to your current role (or most recent title if retired): - Selected Choice
- Select the title most similar to your current role (or most recent title if retired): - Other - Text
- What is the type of data that you currently interact with most often at work or school? - Selected Choice
- What specific programming language do you use most often? - Selected Choice
- What is your current yearly compensation (approximate $USD)?
- How long have you been writing code to analyze data?
- For how many years have you used machine learning methods (at work or in school)?

Some noteworthy features of Kaggle ML and Data Science Survey, 2018:

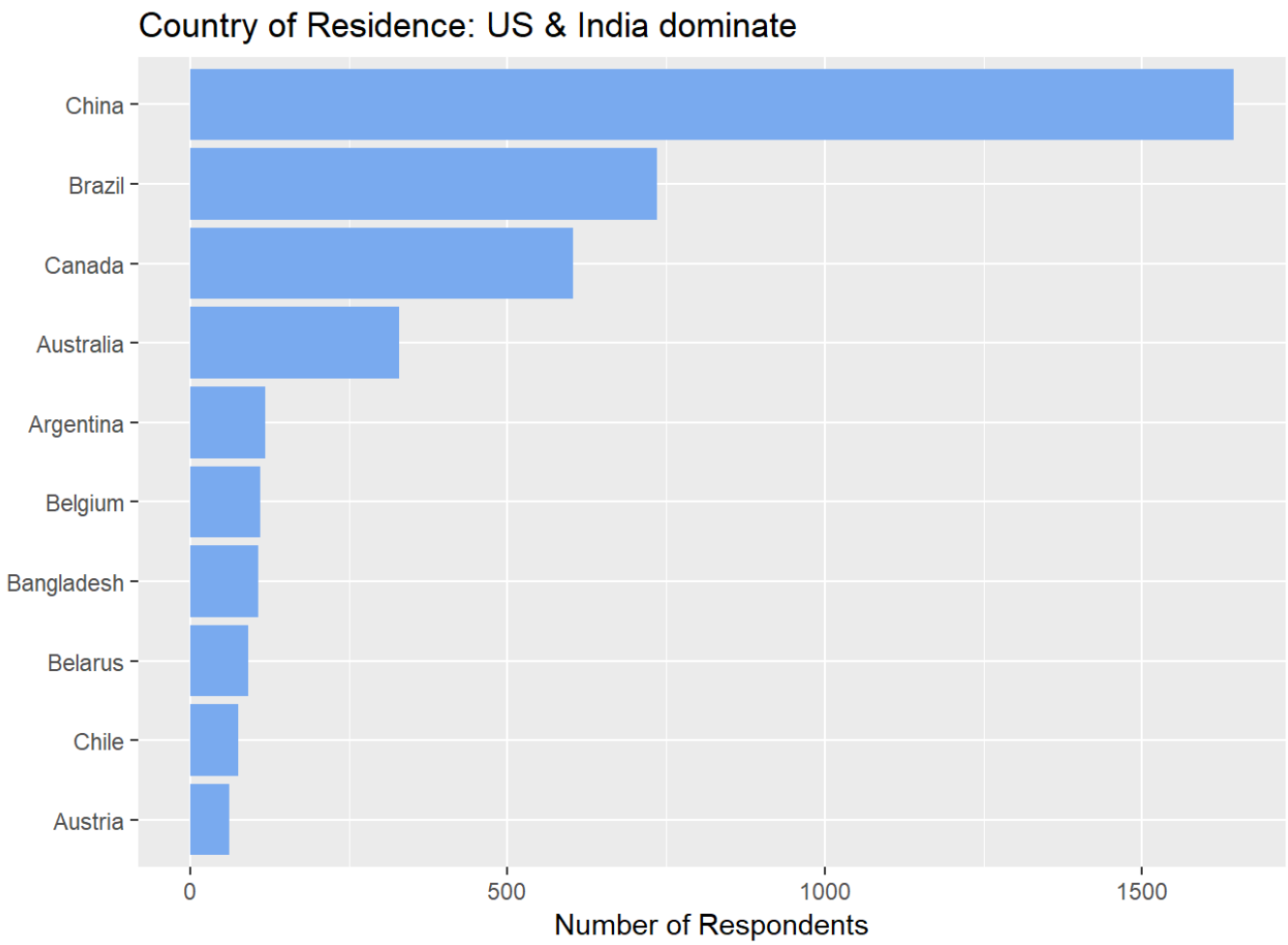
In a basic analysis of Kaggle ML and Data Science Survey, 2018, we found some noteworthy and interesting features of the data.

- **A. Data Cleaning**

```
df_survey_mcq <-  as.tibble(fread(str_c("C:/Users/Flora Huang/Desktop/Exploratory Data Analysis and Visualization/Final project/code/data/multipleChoiceResponses.csv"),skip=1))
df_survey_free <- as.tibble(fread(str_c("C:/Users/Flora Huang/Desktop/Exploratory Data Analysis and Visualization/Final project/code/data/freeFormResponses.csv"), skip = 1))

#clean the country variable to analyze it
df_survey_mcq <- df_survey_mcq %>%
  mutate(country = `In which country do you currently reside?`)
pop <-   df_survey_mcq %>%
  count(country) %>%
  filter(!(country %in% c("Other", "I do not wish to disclose my location"))) %>%
    mutate(iso3 = countrycode(country, origin = "country.name", destination = "iso3c"))
```

- **B. Where do the survey's respondents reside in?**

```
#plot the country variable
df_survey_mcq %>%
  group_by(country) %>%
  count() %>%
  ungroup() %>%
  head(10) %>%
  ggplot(aes(reorder(country, n, FUN = min), n)) +
  geom_col(fill="#79AAEF") +
  labs(x = "", y = "Number of Respondents") +
  theme(legend.position = "none") +
  ggtitle("Country of Residence: US & India dominate") +
  coord_flip()
```




```
df_survey_mcq <- df_survey_mcq %>%
  mutate(country = `In which country do you currently reside?`)
pop <-   df_survey_mcq %>%
  count(country) %>%
  filter(!(country %in% c("Other", "I do not wish to disclose my location"))) %>%
    mutate(iso3 = countrycode(country, origin = "country.name", destination = "iso3c"))
```

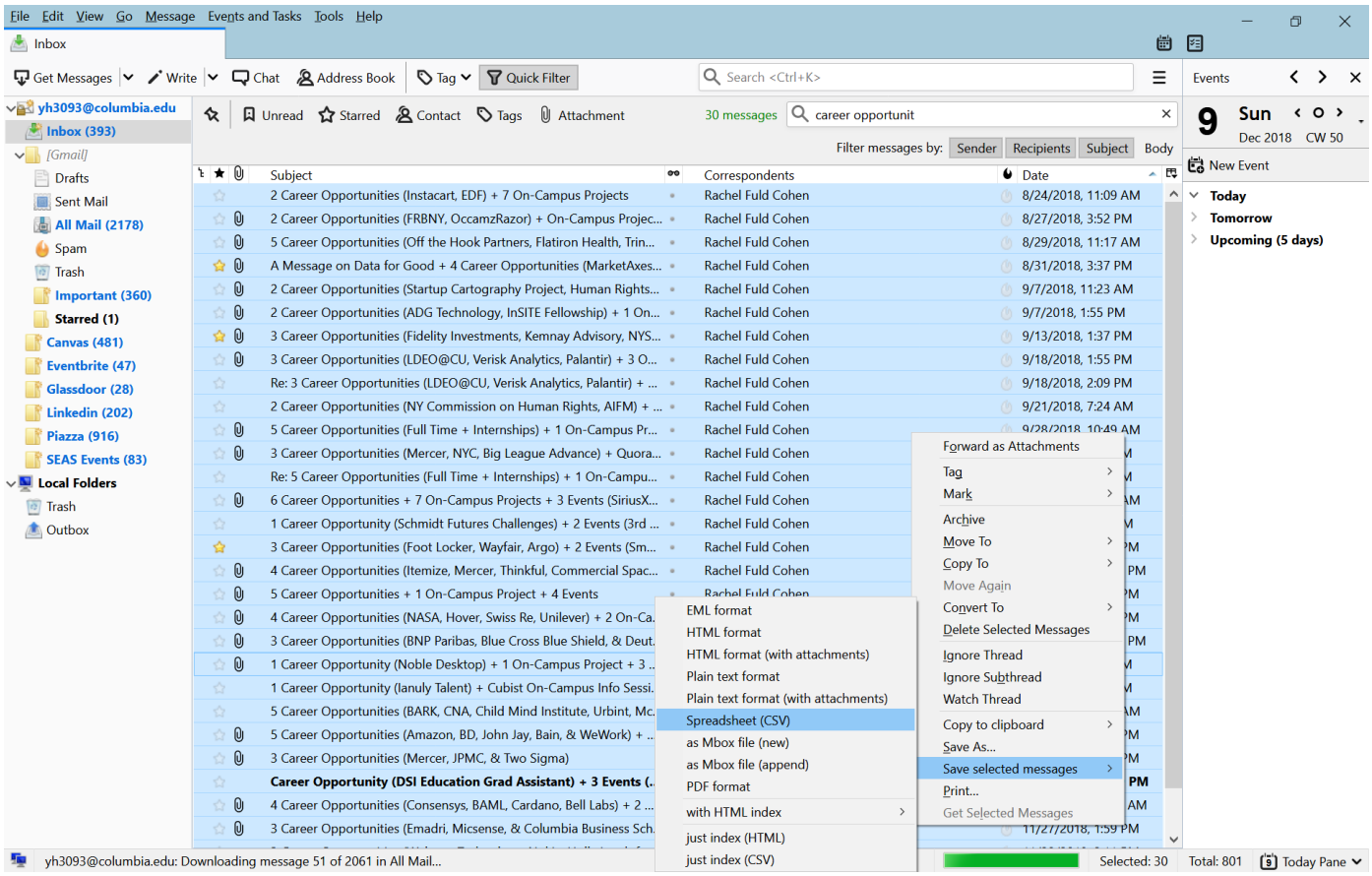
Insights:

We can see that the survey data is mainly from people from U.S and India. Africa, as a country, is very undererepresented, so we might not be able to draw a lot of conclusions for the african continent.

2.3 Demand - Rachel's Mail - Columbia University Data Science Career Opportunities

Students studying in Columbia University's Data Science Master are no strangers to "Rachel's Mail", a mail send out from our Assistant Director of Student Services & Career Development, Ms.Rachel Fuld Cohen. Ms.Cohen sends out emails to every DSI students when new data science jobs are listed and is one of the primary way of students in DSI to apply for summer internships or full time jobs. We used this data as our demand side data that helps us understand what the job market is looking for in prospective job candidates.

We scrapped all Rachel's email tagged with "Career Opportunities" by Mozilla Thinderbird's software (cridentials such as Columbia student Uni accounts are needed), which can be done pretty easily in:



In our first semester (2018/8/24 - 2018/11/29), **30 career opportunities emails were sent out by Rachel, which in total listed 82 job offers.**

The features we studied from the data were:
- General info: Company's name, Job title, Location, Industry
- Job listing time: Date and days the emails were sent to students
- How to apply: Email to HR, Apply on company's websites, On-campus interviews
- Job type: Internships, Full-time jobs
- Description: The requirements of applicants listed on the job posting

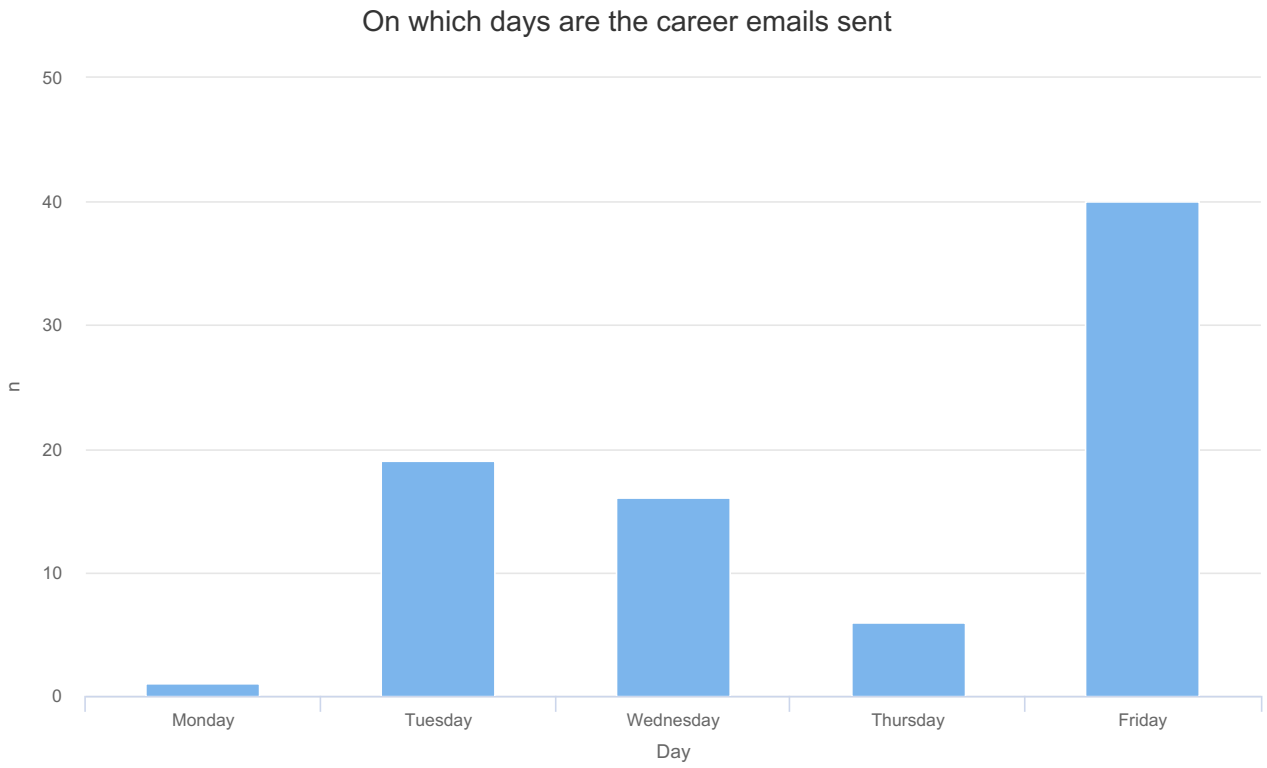
```
rachel <- read_csv("C:/Users/Flora Huang/Desktop/Rachel Email/rachel.csv")
```

Some noteworthy features of Rachel's career opportunity email:

In a basic analysis of Rachel's email, we found some noteworthy and interesting features of the data.

- **A. On which days are the career opportunity emails sent?**

```
rachel$Day <- factor(rachel$Day, levels= c("Monday",
    "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday","Sunday"))
rachel %>% count(Day) %>%  hchart(type = "column", hcaes(x = Day, y = n)) %>% hc_title(text="On which
days are the career emails sent")
```



We can see from the graph that most of Rachel's career opportunity emails are sent out on Friday. Therefore, if a student is actively seeking for job opportunities, they should pay more attention of Friday to be one of the first applicants.

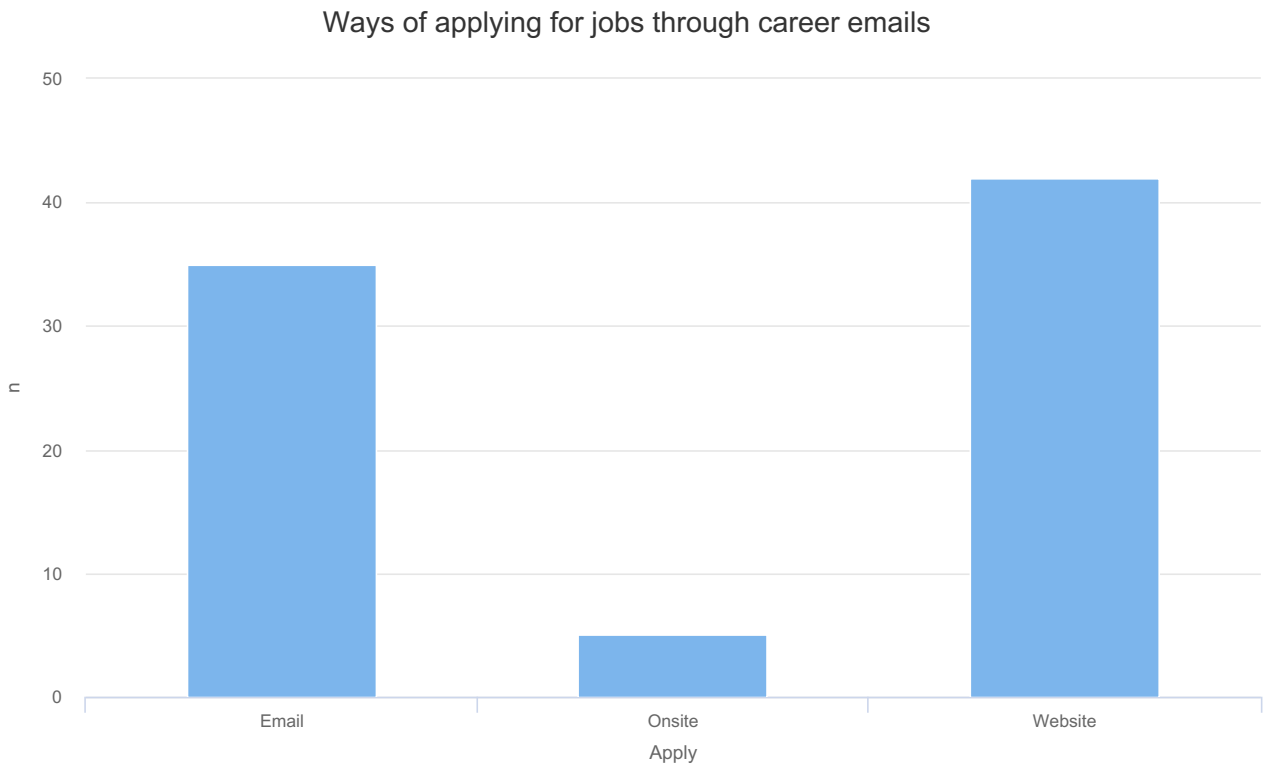
- **B. On which dates are the career opportunity emails sent?**

```
rachel$Date <- as.Date(rachel$Date, "%m/%d/%Y")
rachel %>% count(Date) %>% hchart(type = "column", hcaes(x = Date, y = n)) %>% hc_title(text="On whic
h dates are the career emails sent")
```

We can see that the highest frequency of recieving career opportunity emails are between early October and mid November, which was the time of fall recruitment season and when Data Science Institute held our own career fair. The emails were then pretty fairly distributed within the other times during our research period.

- **C. How can we submit the job applications through career opportunities emails?**

```
rachel$Apply <- factor(rachel$Apply)
rachel %>% count(Apply) %>% hchart(type = "column", hcaes(x = Apply, y = n)) %>% hc_title(text="Ways of applying for jobs through career emails")
```



We can see that the main ways to apply for jobs through the career opportunity emails are emailing the HR and applying on the company's website. There are also some chances that the company would come to Columbia's campus to have on campus interviews with the students.

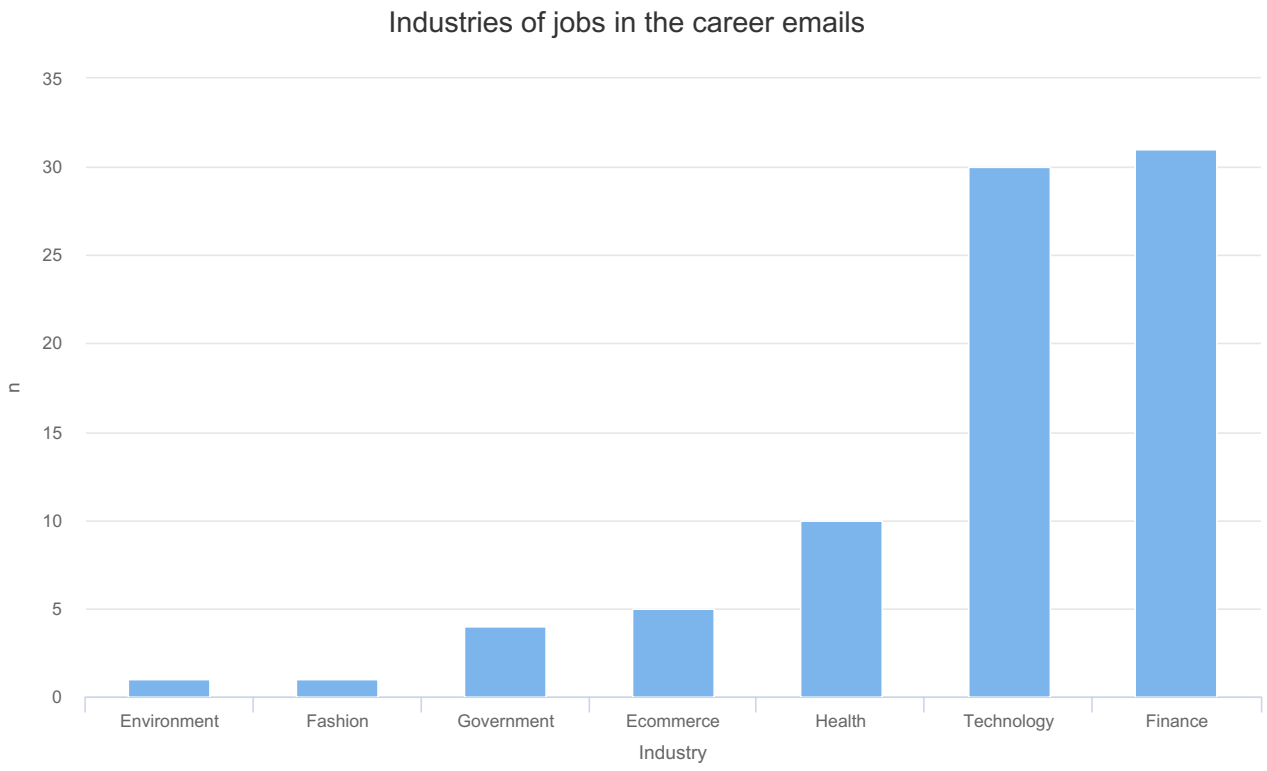
- **D. Where are the jobs based at in the career opportunities emails?**

```
#Analyze the locations
job_location <- data.frame("Location" = c("San Francisco", "Boston", "Washington", "New York"), "Jobs" = c(sum(str_count(rachel$Location, "San Francisco")), sum(str_count(rachel$Location, "Boston")), sum(str_count(rachel$Location, "Washington")), sum(str_count(rachel$Location, "New York"))))

#Plot
highchart() %>% hc_xAxis(type = 'category') %>% hc_add_series(job_location, "column", hcaes(x = Location, y = Jobs)) %>% hc_title(text="Location of jobs in the career emails")
```

We can see from the graph that almost all the jobs in the career opportunities emails are based in New York, which is reasonable due to Columbia University's location.

- **E. What industries are the jobs in the career opportunities emails in?**

```
rachel %>% count(Industry) %>% arrange(n) %>% hchart(type = "column", hcaes(x = Industry, y = n)) %>%
hc_title(text="Industries of jobs in the career emails")
```

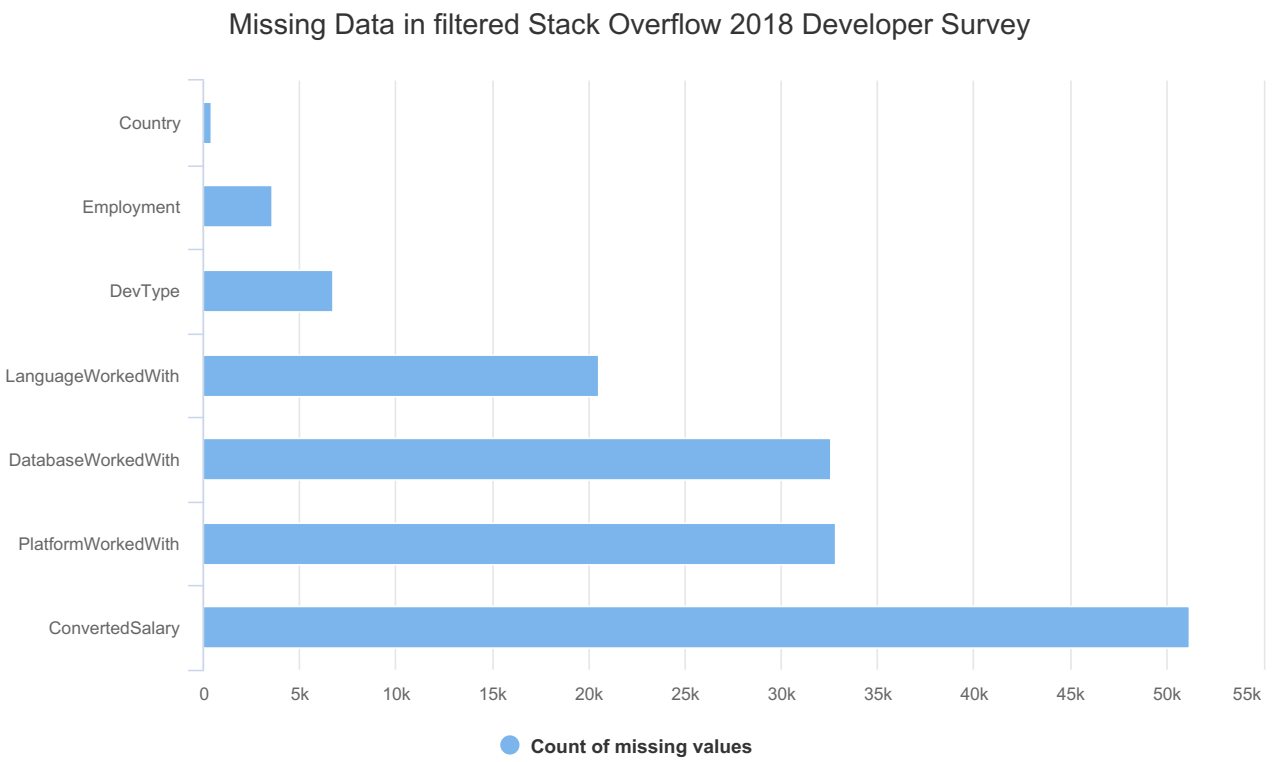


We can see from the graph that most jobs listed in the career opportunities emails are from Finance and Technology sectors, with health sectors following up as the third major industry. There are clear examples of how these industries have been implementing data science in their work, and the great demand for data scientists in these industries can be a good sign for our fellow classmates who are interested in developing a career in them.

3 Analysis of Data Quality

- **Stack Overflow 2018 Developer Survey:**

```
#missing data
missing_d <- (colSums(is.na(stackoverflow))) #count missing data
d <- data.frame(names = colnames(stackoverflow),data = missing_d) #change to data frame
d = d[order(d$data),] #order by count


highchart()%>%
  hc_title(text = paste("Missing Data in filtered Stack Overflow 2018 Developer Survey"))%>% #title
  hc_xAxis(categories = d$names)%>% #xaxis
  hc_add_series(data = d$data, name = "Count of missing values", type = "bar") #plot
```

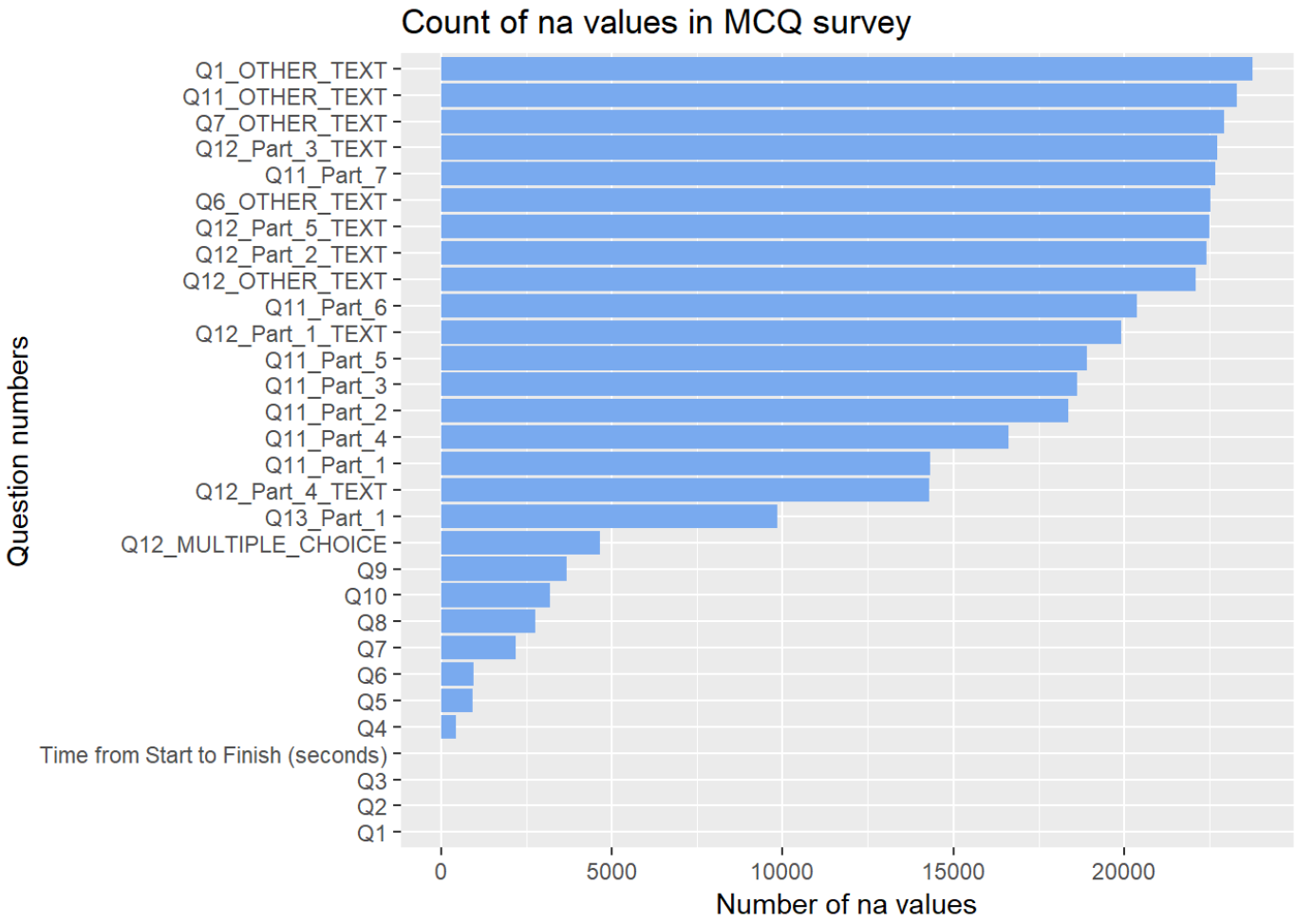


Missing Data histogram describes missing data in Stack Overflow 2018 Developer Survey. ConvertedSalary has the most missing data having more than 50000 NA values, whereas Country as almost no NA values.

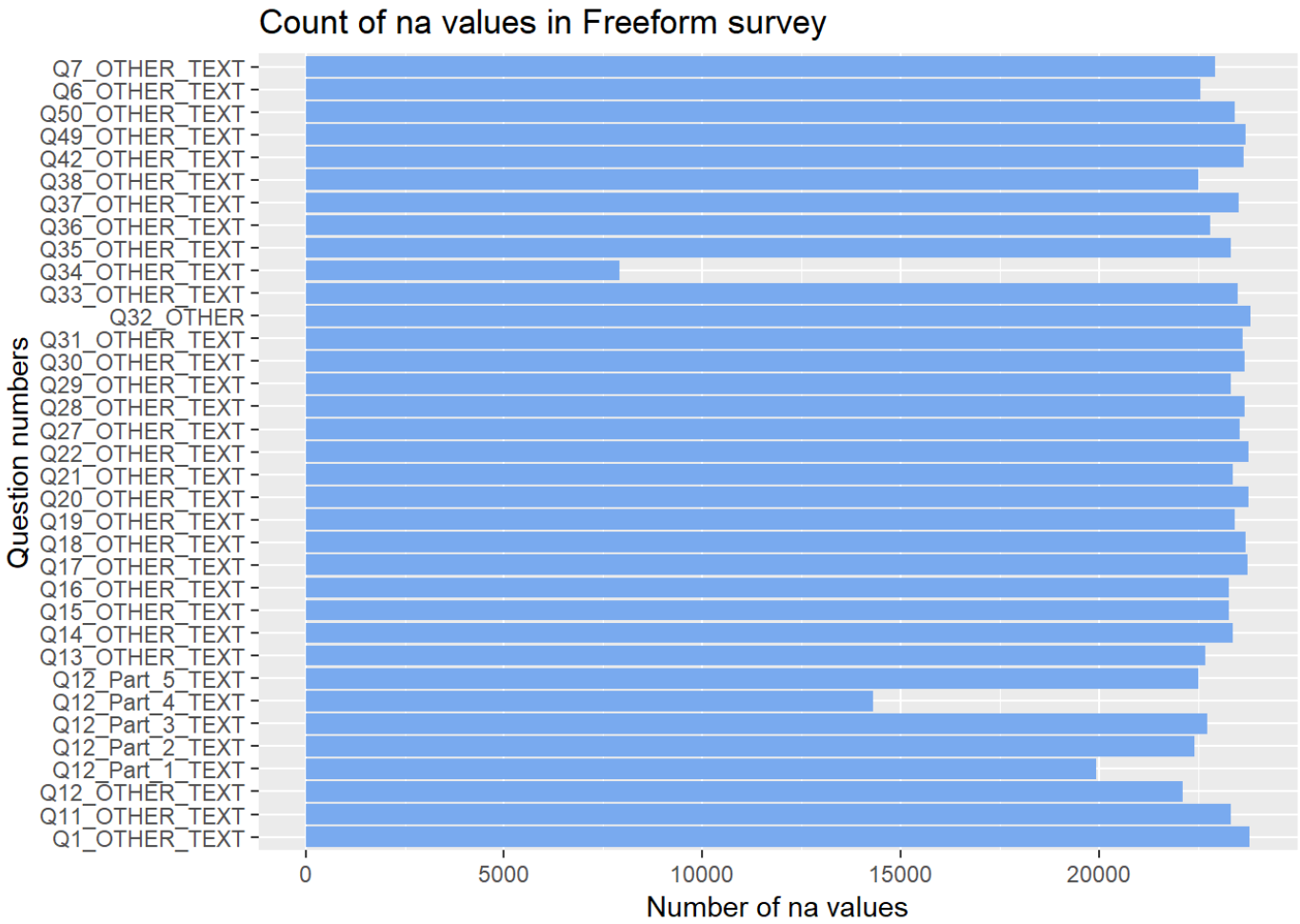
- **Kaggle ML and Data Science Survey, 2018:**

```
#read data
df_survey_mcq_o <-  as.tibble(fread(str_c("C:/Users/Flora Huang/Desktop/Exploratory Data Analysis an
d Visualization/Final project/code/data/multipleChoiceResponses.csv"),na.strings = "-1"))
df_survey_free_o <- as.tibble(fread(str_c("C:/Users/Flora Huang/Desktop/Exploratory Data Analysis an
d Visualization/Final project/code/data/freeFormResponses.csv"),na.strings = "-1"))
```

```
foo <- colSums(is.na(df_survey_mcq_o)|df_survey_mcq_o == '')
foo <- melt(foo,value.name="na_count")
foo <- rownames_to_column(foo, "col_names")

plt <-  ggplot(head(foo,30),aes(x=fct_reorder(col_names,na_count),y=na_count))+
            geom_col(fill="#79AAEF")+
            coord_flip()+
            labs(x = "Question numbers", y = "Number of na values") +
            ggtitle("Count of na values in MCQ survey")
plt
```




```
foo <- colSums(is.na(df_survey_free_o)|df_survey_free_o == '')
foo <- melt(foo,value.name="na_count")
foo <- rownames_to_column(foo, "col_names")

plt <-  ggplot(foo,aes(x=col_names,y=na_count))+
            geom_col(fill="#79AAEF")+
            coord_flip() +
            labs(x = "Question numbers", y = "Number of na values") +
            ggtitle("Count of na values in Freeform survey")
plt
```



Insights: Here we can see that the number of na values of the starting questions are low or zero, but as they get higher, the questions have very high number of NA values. This makes sense as people start leaving the parts of the form when the questions were not required.

We decided not to use the free form part of the survey as it is mostly empty and unstructured, and thus we might not be able to draw proper insightful insights from it.

- **Rachel's Mail - Columbia University Data Science Career Opportunities:**

```
print(paste("The data has",nrow(rachel),"rows and",ncol(rachel), "columns.", sep=" "))
```

```
## [1] "The data has 82 rows and 9 columns."
```

Row / column missing patterns - Missing values by column

```
colSums(is.na(rachel)) %>% sort(decreasing = TRUE)
```

```
##     Company         Day        Date       Title       Apply    Location
##           0           0           0           0           0           0
##    Industry        Type Description
##           0           0           0
```

The dataset scrapped from Rachel's career opportunity emails has no missing data.

4 Main analysis (Exploratory Data Analysis)

4.1 Who's in the data science job market

In 4.1, we want to understand the composition of the current job market in the field of data science.

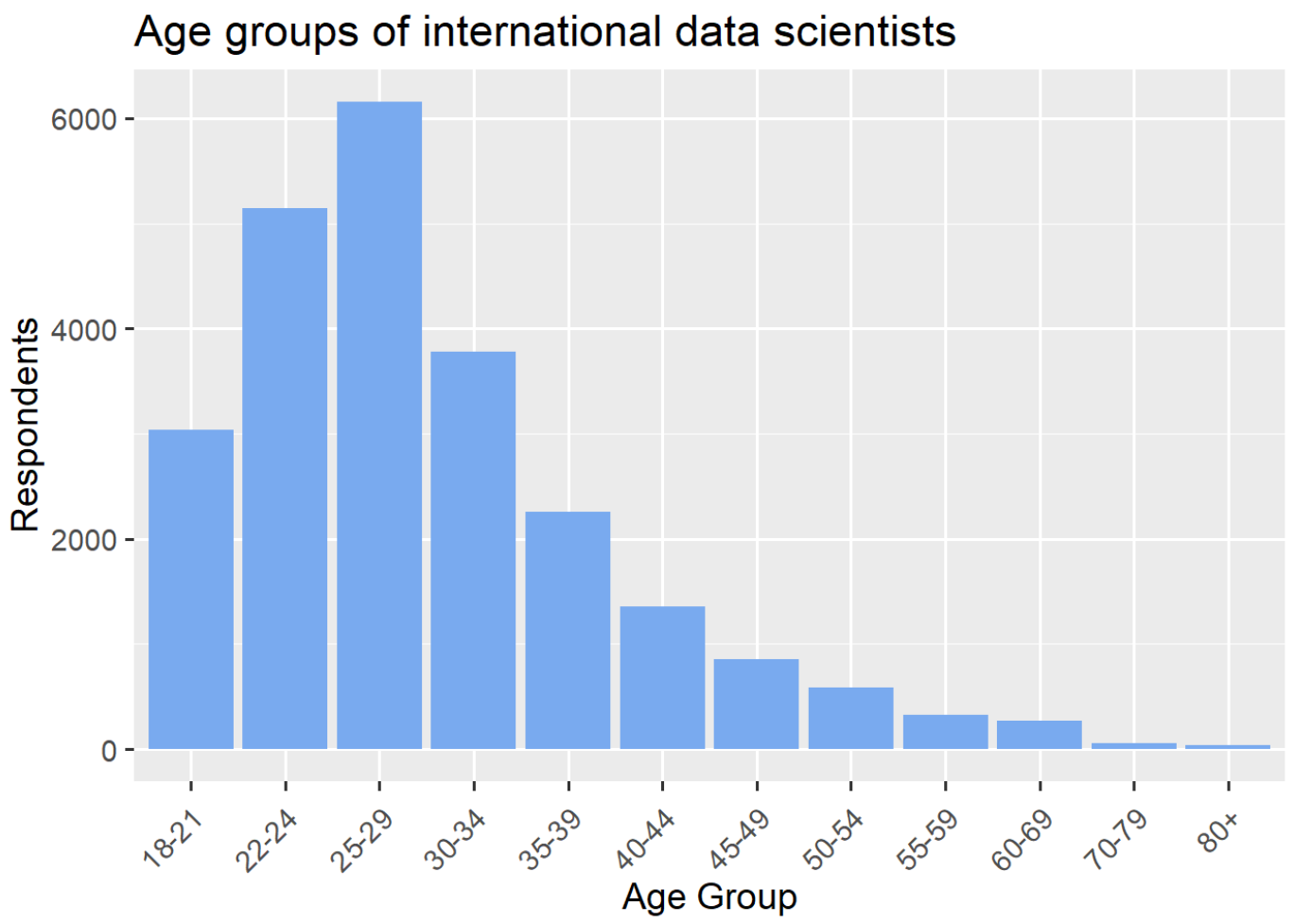
Data Cleaning

```
#clean the age varaible to make is usable
vars <- c(gender = "What is your gender? - Selected Choice",
          gen_txt = "What is your gender? - Prefer to self-describe - Text",
          age = "What is your age (# years)?")

flvl <- c("18-21", "22-24","25-29","30-34","35-39","40-44","45-49","50-54","55-59","60-69","70-
79","80+")

df_survey_mcq <- df_survey_mcq %>%
  rename(!!vars) %>%
  mutate(age = fct_relevel(age, flvl))
```

4.1.1 Age

We are interested in knowing the age distribution of data scientists in the job market.

```
foo <- df_survey_mcq %>%
  group_by(age) %>%
  count()

p2 <- foo %>%
  mutate(percentage = str_c(as.character(round(n/sum(foo$n)*100,1)), "%")) %>%
  ggplot(aes(age, n)) +
  geom_col(fill="#79AAEF") +
  labs(x = "Age Group", y = "Respondents") + theme_grey(14) +
  theme(legend.position = "none", axis.text.x  = element_text(angle=45, hjust=1, vjust=0.9)) +
  ggtitle("Age groups of international data scientists")
p2
```

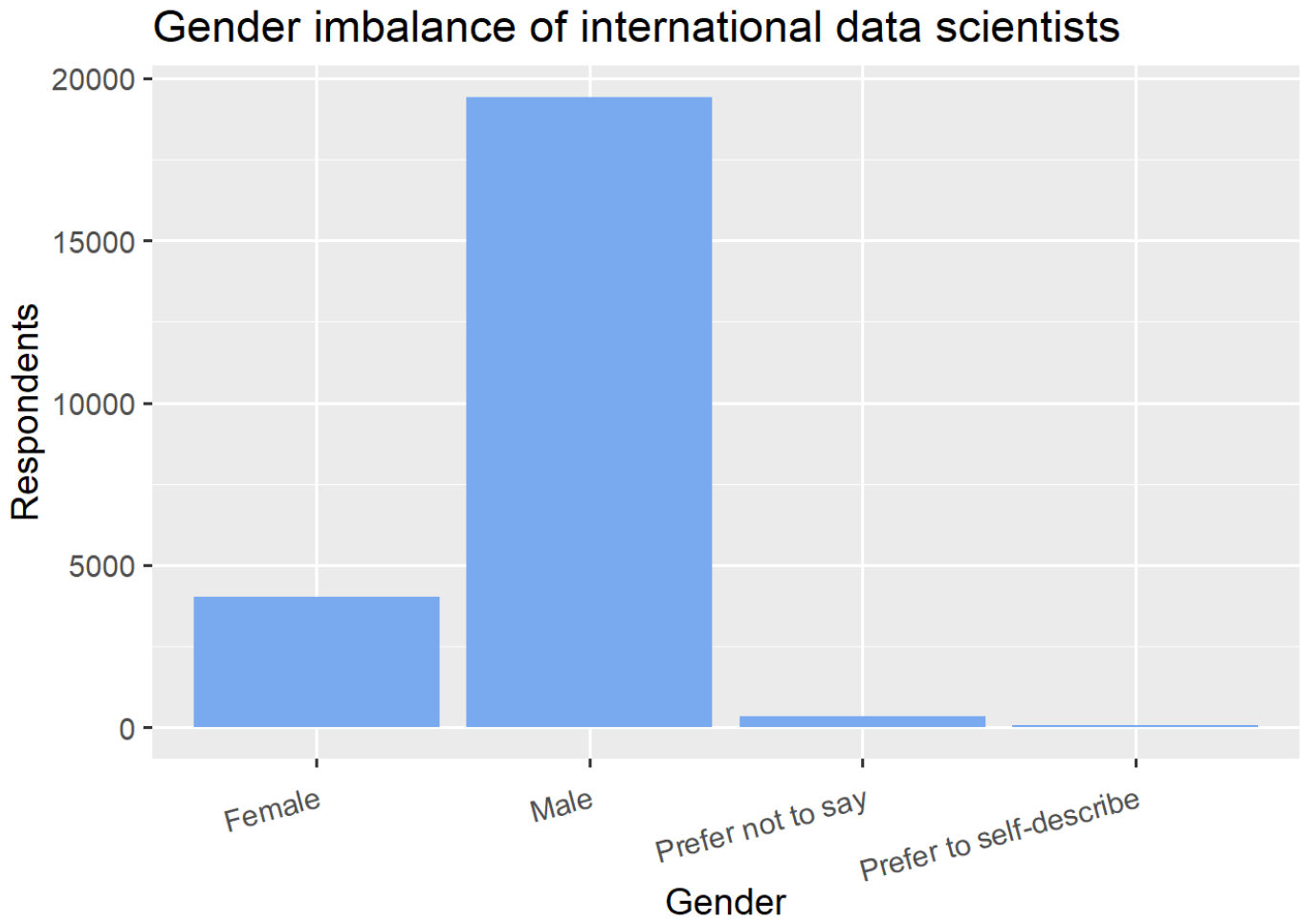


Insights: We find that most of the people are in the age group of 20-35, indicating that the field of data science is mainly dominated by the younger generation. This makes sense as the field is relatively new, and thus people just graduating would be more interested and capable of entering the field.

4.1.2 Gender

We are interested in knowing the gender imbalance of data scientists in the job market.

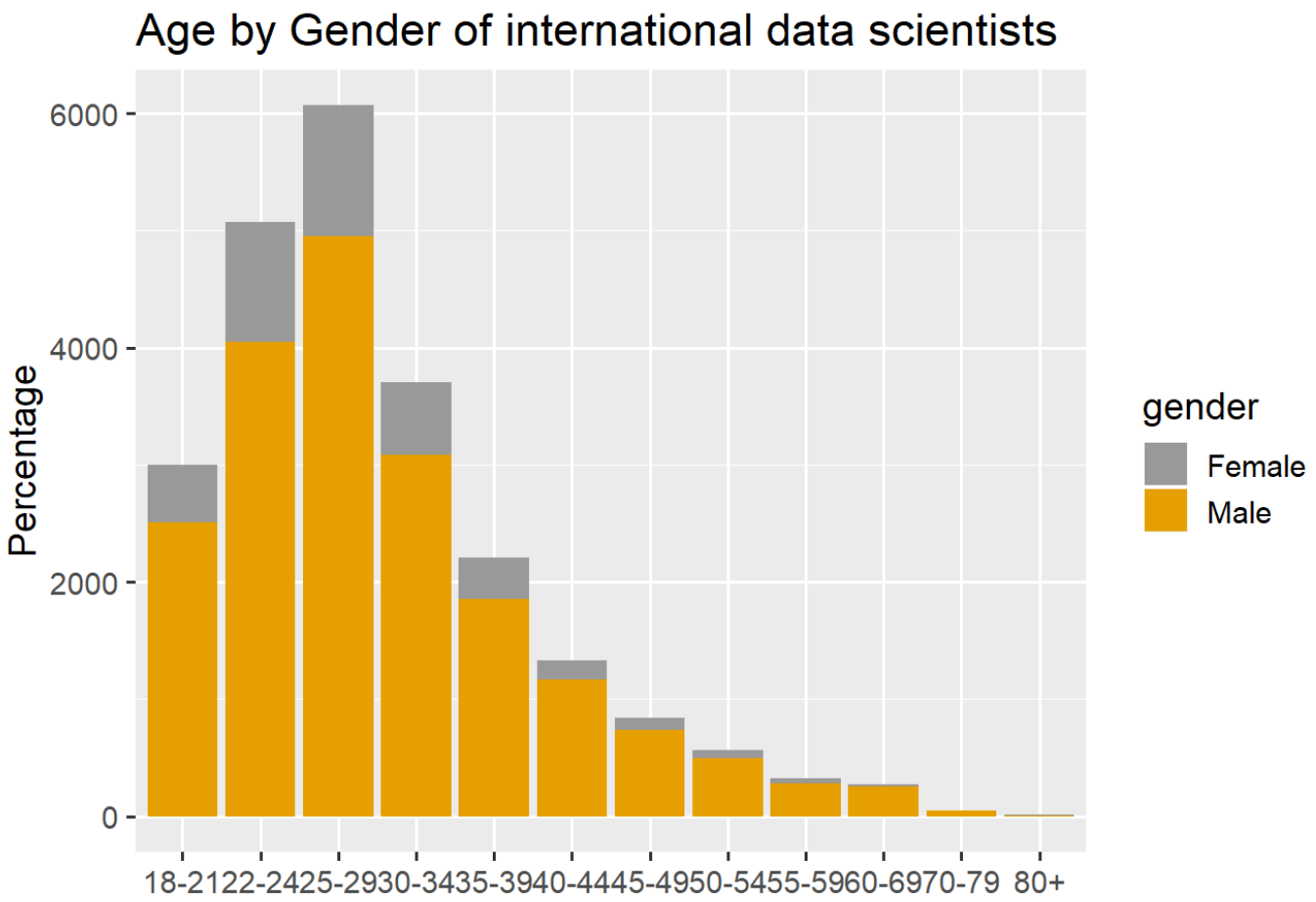
```
foo <- df_survey_mcq %>%
  group_by(gender) %>%
  count()

p1 <- foo %>%
  mutate(percentage = str_c(as.character(round(n/sum(foo$n)*100,1)), "%")) %>%
  ggplot(aes(gender, n)) +
  geom_col(fill="#79AAEF") +
  labs(x = "Gender", y = "Respondents") + theme_grey(14) +
  theme(legend.position = "none", axis.text.x  = element_text(angle=15, hjust=1, vjust=0.9)) +
  ggtitle("Gender imbalance of international data scientists")
p1
```



Insights: We can see that there is a huge gender imbalance among the male and the female and other genders, which is normally the case in any STEM field.

4.1.3 Age by gender

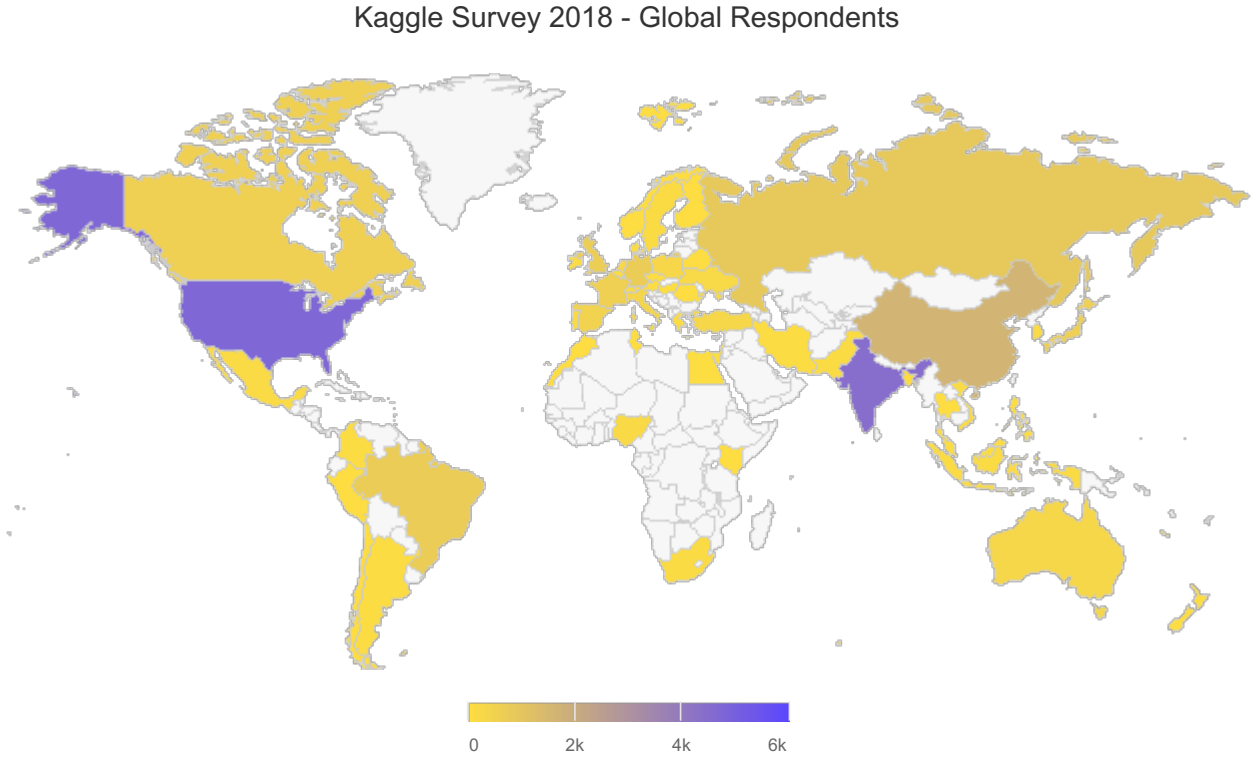
```
#plot and analyze the age and gender variable
p3 <- df_survey_mcq %>%
  filter(gender %in% c("Male", "Female")) %>%
  ggplot(aes(age, fill = gender)) +
  geom_bar() + theme_grey(14) +
  labs(x = "", y = "Percentage") +
  ggtitle("Age by Gender of international data scientists") + scale_fill_manual(values=cbPalette)


p3
```



Comparing age by gender, we see that the younger generations are doing better compared to the older generations in terms of the sex ratio. There are also more women opting for data science as a career path.

4.1.4 Represented countries

```
highchart() %>%
  hc_add_series_map(worldgeojson, pop, value = 'n', joinBy = 'iso3') %>%
  hc_title(text = 'Kaggle Survey 2018 - Global Respondents') %>%
  hc_colorAxis(minColor = "#ffdf3f", maxColor = "#5c46ff") %>%
  hc_tooltip(useHTML = TRUE, headerFormat = "", pointFormat = "{point.country}: {point.n} users")
```



Insights: As seen previously from the bar chart, the map is clear that the countries which dominate the field of data science are USA, India, China, Russia, Brazil and UK. This might be due to their larger population or the more interest in data science from these countries.

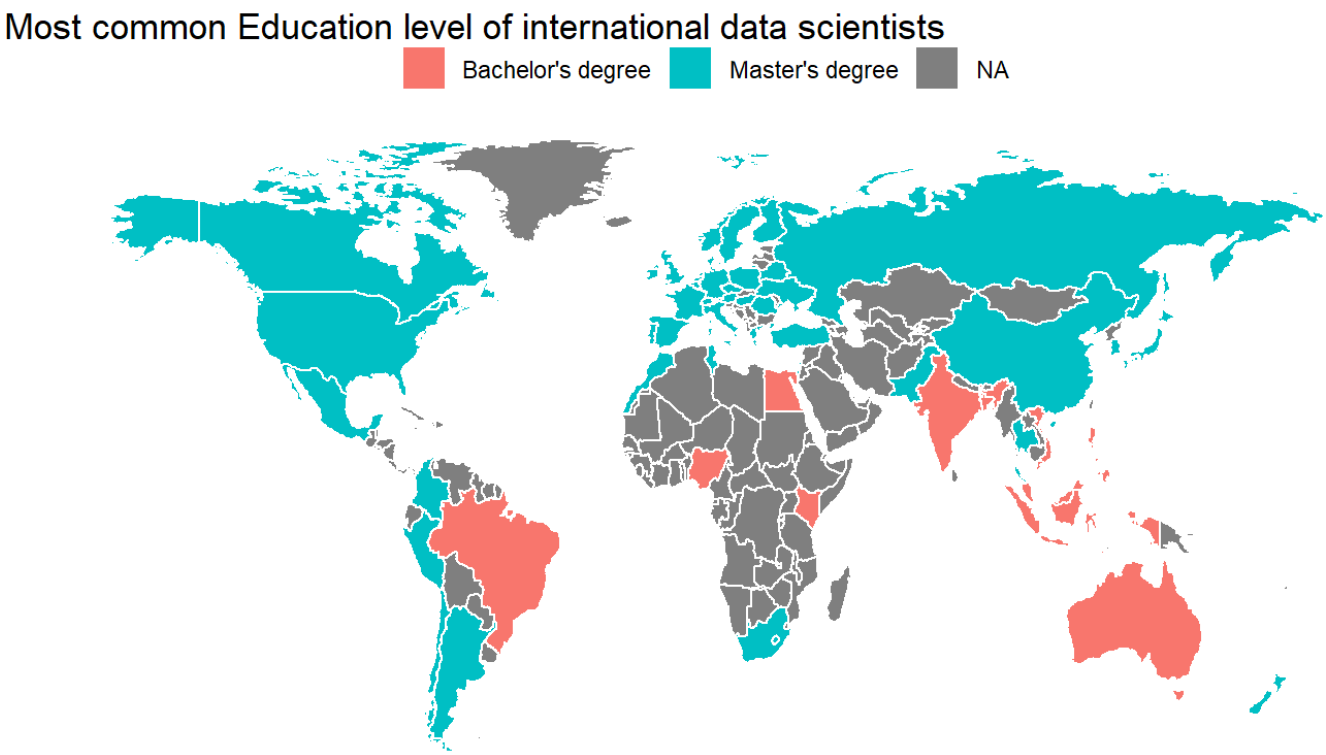
4.1.5 Education and industry

Data cleaning

```
##clean the questions which are related to the education level and degree of the respondents
vars <- c(edu = "What is the highest level of formal education that you have attained or plan to att
ain within the next 2 years?",
          major = "Which best describes your undergraduate major? - Selected Choice",
          role = "Select the title most similar to your current role (or most recent title if
retired): - Selected Choice",
          role_txt = "Select the title most similar to your current role (or most recent title if reti
red): - Other - Text",
          industry = "In what industry is your current employer/contract (or your most recent
employer if retired)? - Selected Choice",
          industry_txt = "In what industry is your current employer/contract (or your most recent
employer if retired)? - Other - Text")

edu_lvl <- c("Doctoral degree","Professional degree", "Master's degree", "Bachelor's degree",
             "Some college/uni",
             "High school",
             "No answer")

df_survey_mcq <- df_survey_mcq %>%
  rename(!!vars) %>%
  mutate(edu = if_else(edu == "Some college/university study without earning a bachelor's degree", "S
ome college/uni", as.character(edu))) %>%
  mutate(edu = if_else(edu == "No formal education past high school", "High school", edu)) %>%
  mutate(edu = if_else(edu == "I prefer not to answer", "No answer", edu)) %>%
  mutate(edu = na_if(edu, "")) %>%
  mutate(edu = fct_relevel(edu, edu_lvl)) %>%
  mutate(major = na_if(major, "")) %>%
  mutate(role = na_if(role, "")) %>%
  mutate(industry = na_if(industry, ""))
```

4.1.5.1 Most common education level

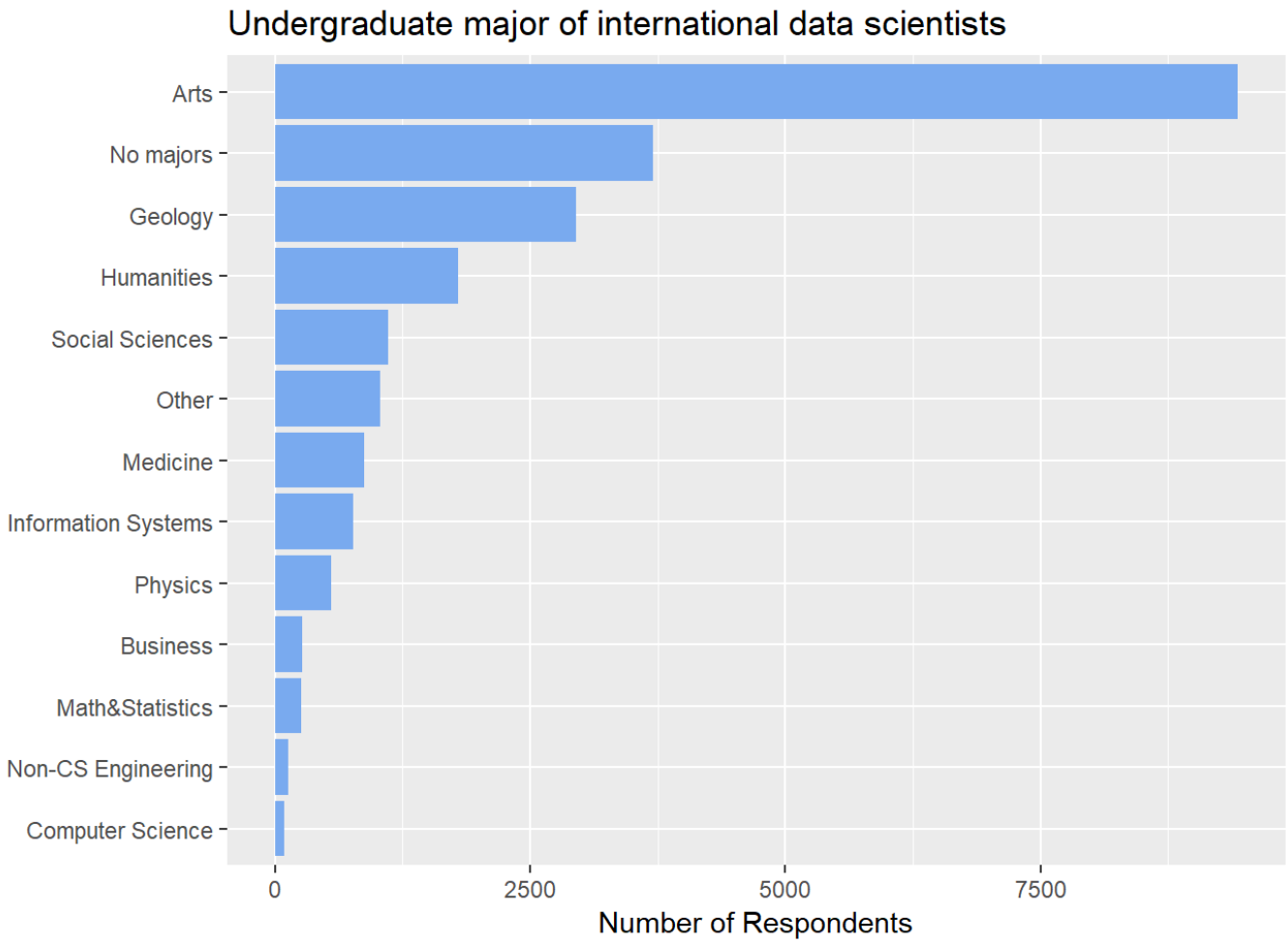
```
#visualize the country of respondents
foo <- df_survey_mcq %>%
  filter(!(country %in% c("Other", "I do not wish to disclose my location"))) %>%
   mutate(country = as.character(case_when(
    country == "United Kingdom of Great Britain and Northern Ireland" ~ "UK",
    country == "United States of America" ~ "USA",
    country == "Viet Nam" ~ "Vietnam",
    TRUE ~ as.character(country)
  ))) %>%
  group_by(edu, country) %>%
  filter(!is.na(major)) %>%
  count() %>%
  arrange(desc(n)) %>%
  group_by(country) %>%
  slice(c(1))
world <- map_data("world")
world %>%
  filter(region != "Antarctica") %>%
  left_join(foo, by = c("region" = "country")) %>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, fill = edu, group = group), color = "white") +
  coord_fixed(1.3) +
  labs(fill = "") + theme_grey(16) +
  theme(legend.position = "top") +
  theme_void() +
  theme(legend.position = "top") +
  ggtitle("Most common Education level of international data scientists")
```



Insights: We can see that the most people in the current data science job market have a masters' degree, but in some countries like India , Brazil and Australia, the barrier of entry is lower as more people with just a bachelors degree are also able to join the industry.

4.1.5.2 Undergraduate major

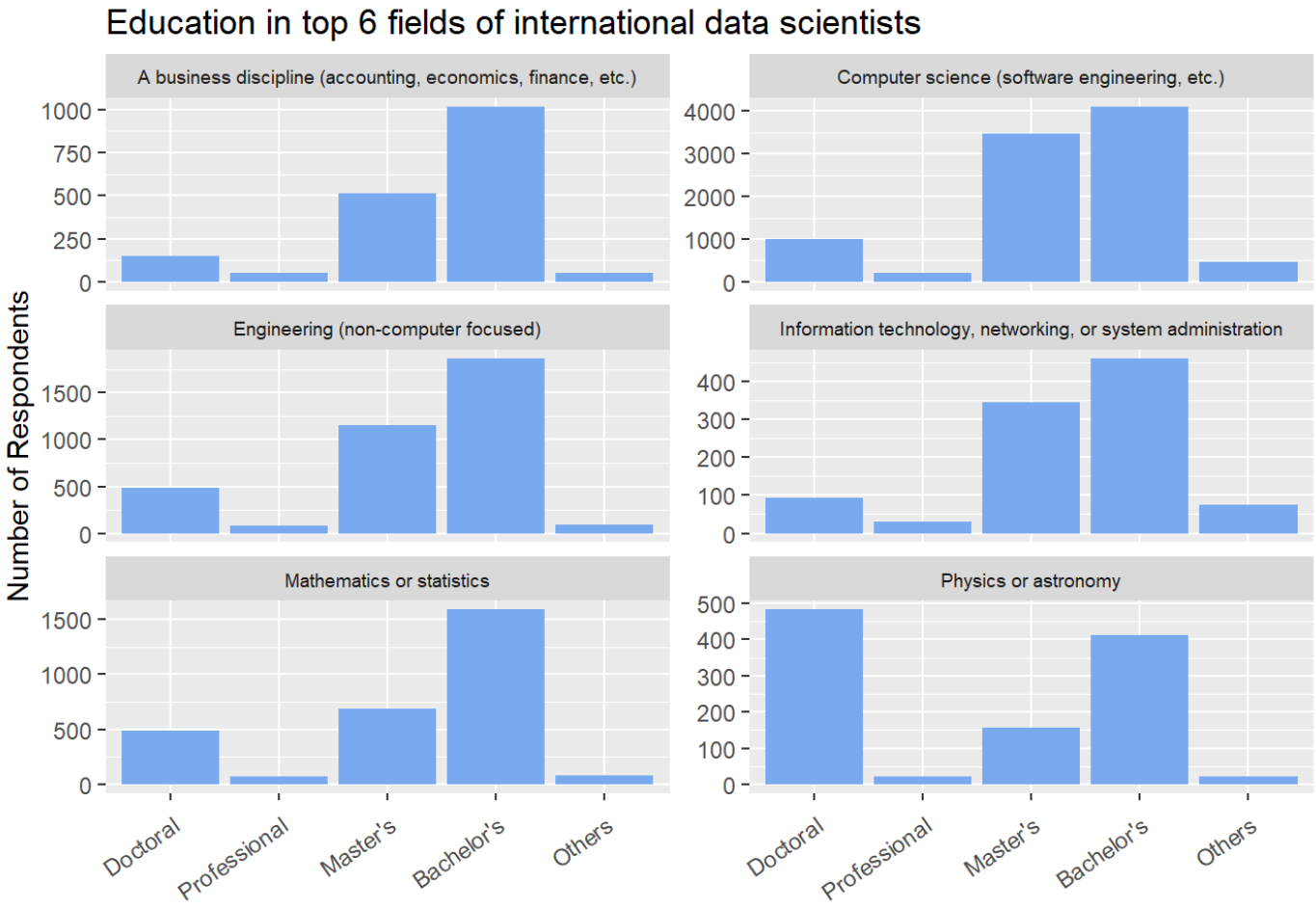
```
labs <- c("Computer Science", "Non-CS Engineering", "Math&Statistics", "Business", "Physics",
"Information Systems", "Medicine", "Other", "Social Sciences", "Humanities", "Geology", "No majors",
"Arts")

#visualize the undergrad major of respondents
df_survey_mcq %>%
  filter(!is.na(major)) %>%
  count(major) %>%
  ggplot(aes(reorder(major, n, FUN = min), n)) +
  geom_col(fill="#79AAEF") +
  coord_flip() +
  theme(legend.position = "none") +
  labs(x = "", y = "Number of Respondents") +
  ggtitle("Undergraduate major of international data scientists") + scale_x_discrete(labels= labs)
```



Insights: We see that as expected, the most undergrad major for data scienctists is Computer Science. However, we also notice that people from lot of the other majors are coming to the field as well, such as Mathematics, Business and physics.

4.1.5.3 Education in top 6 fields

```r
#visulaize the education level of respondents
foo <- df_survey_mcq %>%
  filter(!is.na(major)) %>%
  group_by(major) %>%
  count() %>%
  ungroup() %>%
  top_n(6, n)

labs <- c("Doctoral", "Professional", "Master's", "Bachelor's", "Others")

df_survey_mcq %>%
  filter(!is.na(edu) & edu != "No answer") %>%
  semi_join(foo, by = "major") %>%
  count(edu, major) %>%
  ggplot(aes(edu, n)) +
  geom_col(fill="#79AAEF") +
  theme(legend.position = "none",
        axis.text.x  = element_text(angle=35, hjust=1, vjust=0.9),
        strip.text.x = element_text(size = 7)) +
  guides(fill = guide_legend(ncol = 2)) +
  labs(x = "", y = "Number of Respondents") +
  facet_wrap(~ major, ncol = 2, scales = "free_y") +
  ggtitle("Education in top 6 fields of international data scientists") + scale_x_discrete(labels=
labs)
```



Insights: We see that the most common degree is masters for all the undergrad majors. But a lot of the PhDs are also in this industry. Apart from we notice that the PhD are more common when joining the industry from physics major.

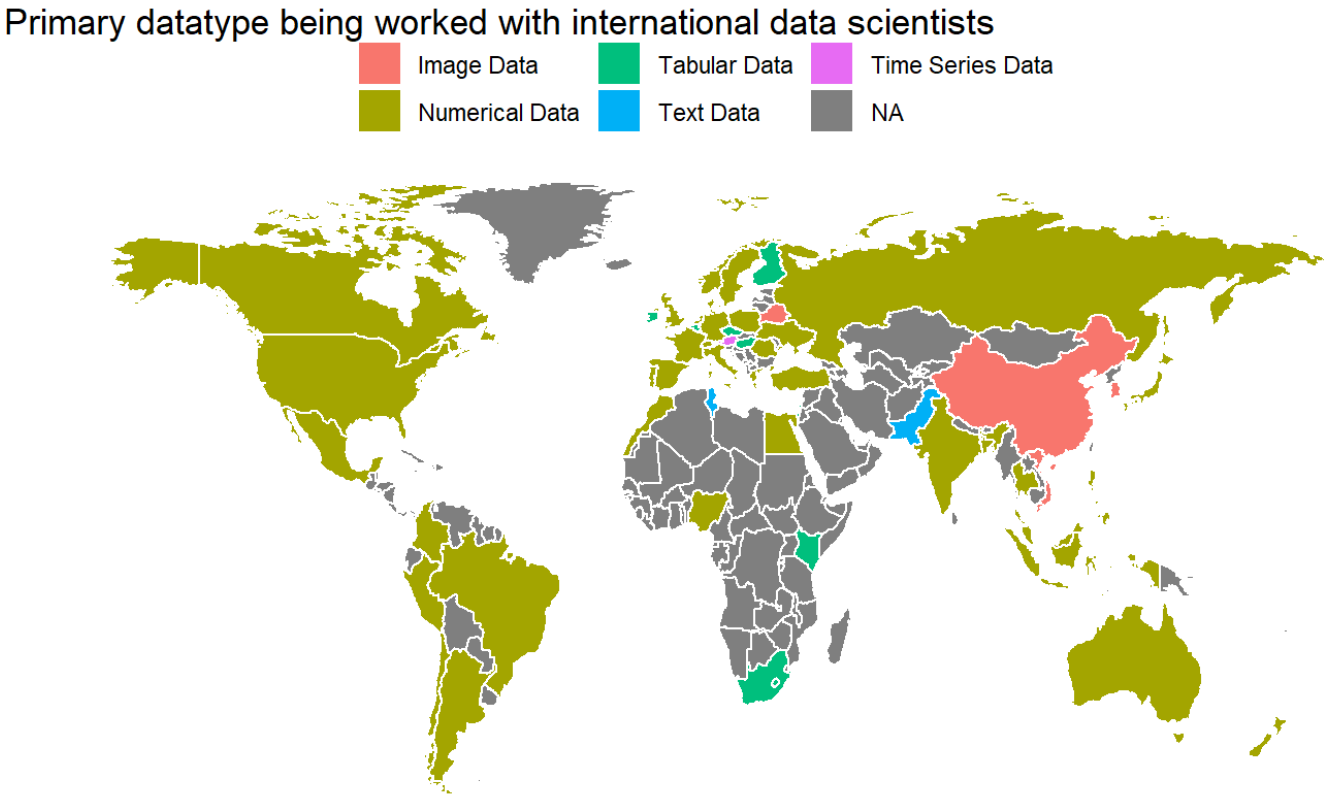
4.1.5.4 Primary Datatype being worked with

Data cleaning

```r
#clean the primary datatype columns
vars <- c(datatype = "What is the type of data that you currently interact with most often at work o
r school? - Selected Choice")

df_survey_mcq <- df_survey_mcq %>%
  rename(datatype = "What is the type of data that you currently interact with most often at work or
school? - Selected Choice") %>%
  mutate(datatype = na_if(datatype, ""))


df_country_data <- df_survey_mcq %>%
  filter(!(country %in% c("Other", "I do not wish to disclose my location"))) %>%
   mutate(country = as.character(case_when(
    country == "United Kingdom of Great Britain and Northern Ireland" ~ "UK",
    country == "United States of America" ~ "USA",
    country == "Viet Nam" ~ "Vietnam",
    TRUE ~ as.character(country)
  ))) %>%
  group_by(datatype, country) %>%
  filter(!is.na(datatype)) %>%
  count() %>%
  arrange(desc(n)) %>%
  group_by(country) %>%
  slice(c(1))
```

Plot

```r
#visualize the primary datatype  attribute
world %>%
  filter(region != "Antarctica") %>%
  left_join(df_country_data, by = c("region" = "country")) %>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, fill = datatype, group = group), color = "white") +
  coord_fixed(1.3) +
  labs(fill = "") +
  theme(legend.position = "top") +
  theme_void() +
  theme(legend.position = "top") +
  ggtitle("Primary datatype being worked with international data scientists")
```



Insights: Looking at the data type being worked on by the people, something interesting that we can notice is that China seems to be mainly working on Image data, while most if the rest of the world seems to be working on numerical data. Also we see that pakistan seems to be interested in textual data more. African countries seems to be working on tabular data but these might not be representative of the complete population as the sample size is very small in african countries.

4.2 What skills are in the data science job market (both demand and supply)

In 4.2, we want to understand what skills are in the current job market. We will break down the skills to two parts, general skills and technical skills. This is in line with the study The Most in Demand Skills for Data Scientists from Jeff Hale, Data Scientist Focused on Machine Learning - CoFounder and COO at E-commerce Firms in Novemeber, 2018. In his study, he searched through the major job listing websites in the United States with “data scientist” “[keyword]” and calculated the exact match search reduced the number of results. In our study, We will use the job listings from Rachel’s Mail - Columbia University Data Science career opportunities emails directly as our source of data science job listings.

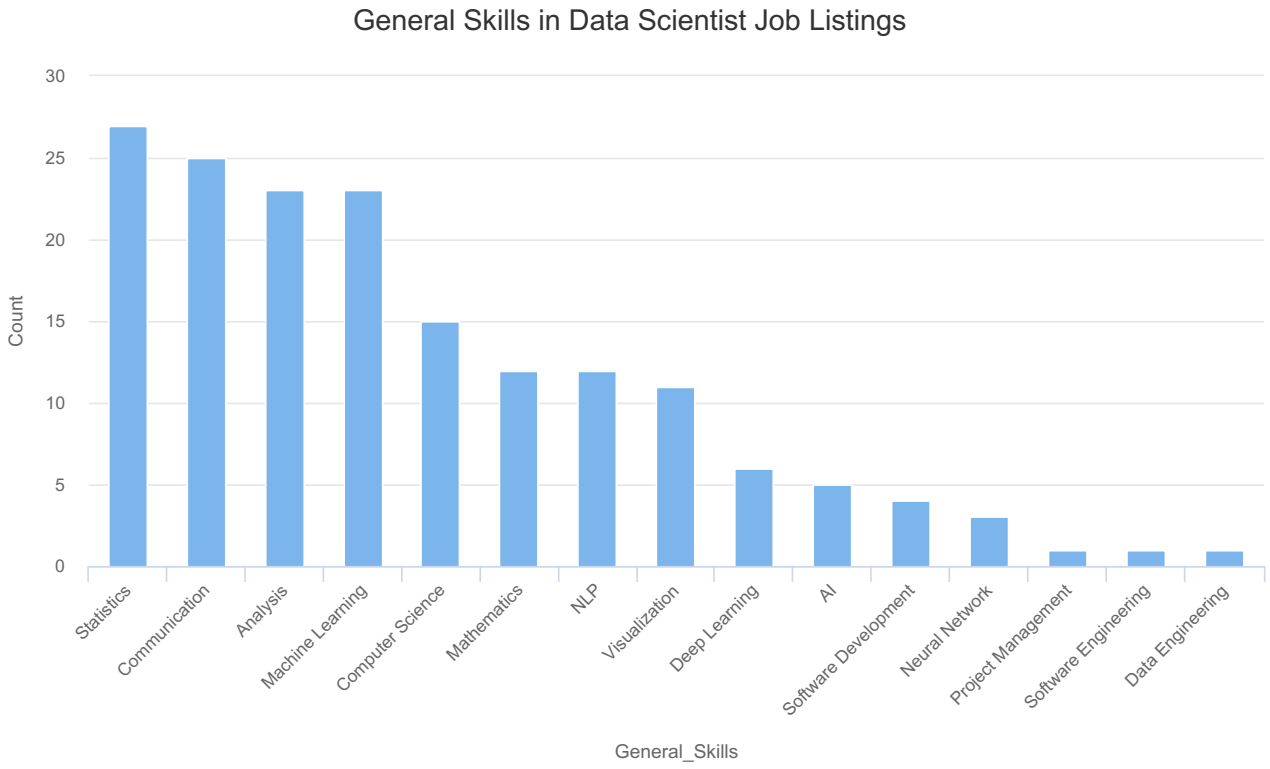
We designed our analysis in line with this study to also compare between major job listing websites and the Columbia University Data Science career opportunities emails. We will also conduct our own keyword search to see what keywords are most listed in Columbia University Data Science career opportunities emails.

4.2.1 General Skills

4.2.1.1 Demand of General Skills – General Skills Wanted in Job Listings

We used the general skills listed in Jeff Hale’s study and searched for their occurences in the job listing requirements in Columbia University Data Science career opportunities emails.

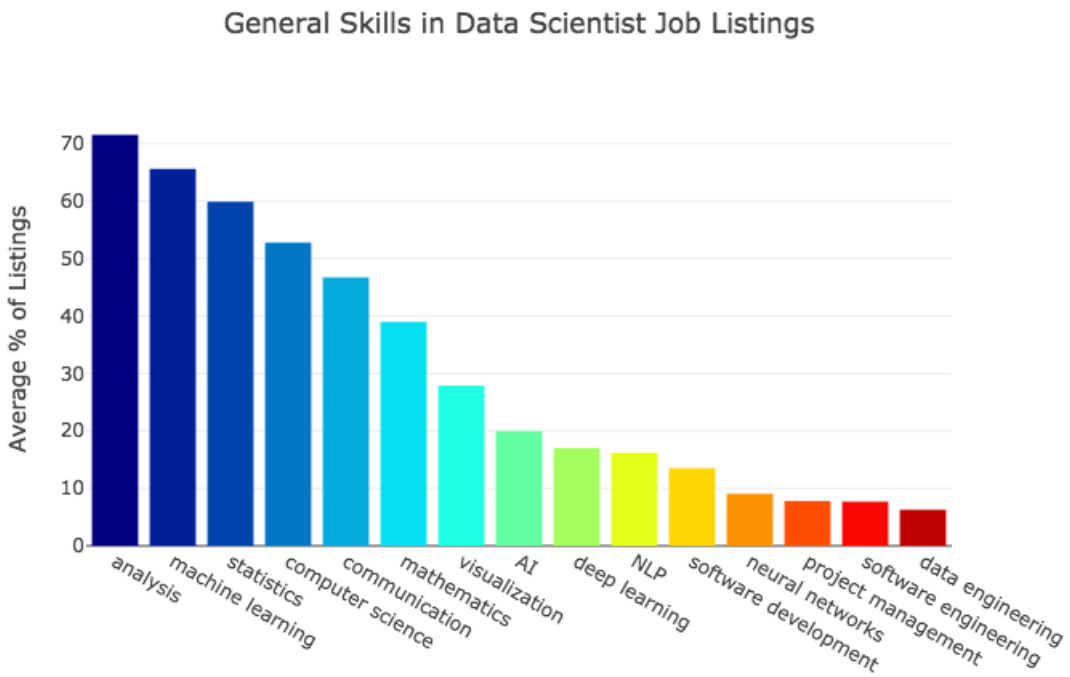
```
#Keyword Analysis
keywords <- data.frame("General_Skills" = c("Analysis", "Machine Learning", "Statistics", "Computer S
cience", "Communication", "Mathematics", "Visualization", "AI", "Deep Learning", "NLP", "Software Dev
elopment", "Neural Network", "Project Management", "Software Engineering", "Data Engineering"),
"Count" = c(length(grep("analysis", tolower(rachel$Description))), length(grep("machine learning", to
lower(rachel$Description))), length(grep("statistics", tolower(rachel$Description))),
length(grep("computer science", tolower(rachel$Description))), length(grep("communication", tolower(
rachel$Description))),  length(grep("mathematics", tolower(rachel$Description))), length(grep("visual
ization", tolower(rachel$Description))), length(grep("AI",
rachel$Description))+length(grep("artificial intelligence", tolower(rachel$Description))), length(gre
p("deep learning", tolower(rachel$Description))), length(grep("NLP", rachel$Description))+length(grep
("natural language processing", tolower(rachel$Description))), length(grep("software development", t
olower(rachel$Description))), length(grep("neural network", tolower(rachel$Description))),
length(grep("project management", tolower(rachel$Description))), length(grep("software engineering",
tolower(rachel$Description))), length(grep("data engineering", tolower(rachel$Description)))))

#Plot
keywords %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = General_Skills, y = Count)) %
>% hc_xAxis(type = 'category') %>% hc_title(text="General Skills in Data Scientist Job Listings")
```



And the result from Jeff Hale's study on major job listing websites is as follow:

(Source: https://www.kdnuggets.com/2018/11/most-demand-skills-data-scientists.html)

Note: We didn't create this plot. This plot was from Jeff Hale's study on major job listing websites listed just for reference and comparison



Our results show that statistical analysis, communication and machine learning are at the heart of data scientist jobs, and that matches with Jeff Hale's study on major job listing websites. The result matches with our expectations, since the primary function of data science is to use statistical analysis to draw useful insights from data. Machine learning and its subsets - AI and deep learning, also show up frequently since these are the major techniques in the field of data science to create systems to predict performance and are very in demand.

It is also noteworthy that communication tops the rank in both studies' job listing descriptions. It tells us that it is very important for data scientists to be able communicate insights and work with others.

4.2.1.2 Demand of General Skills – Top 20 keywords in Job Listings

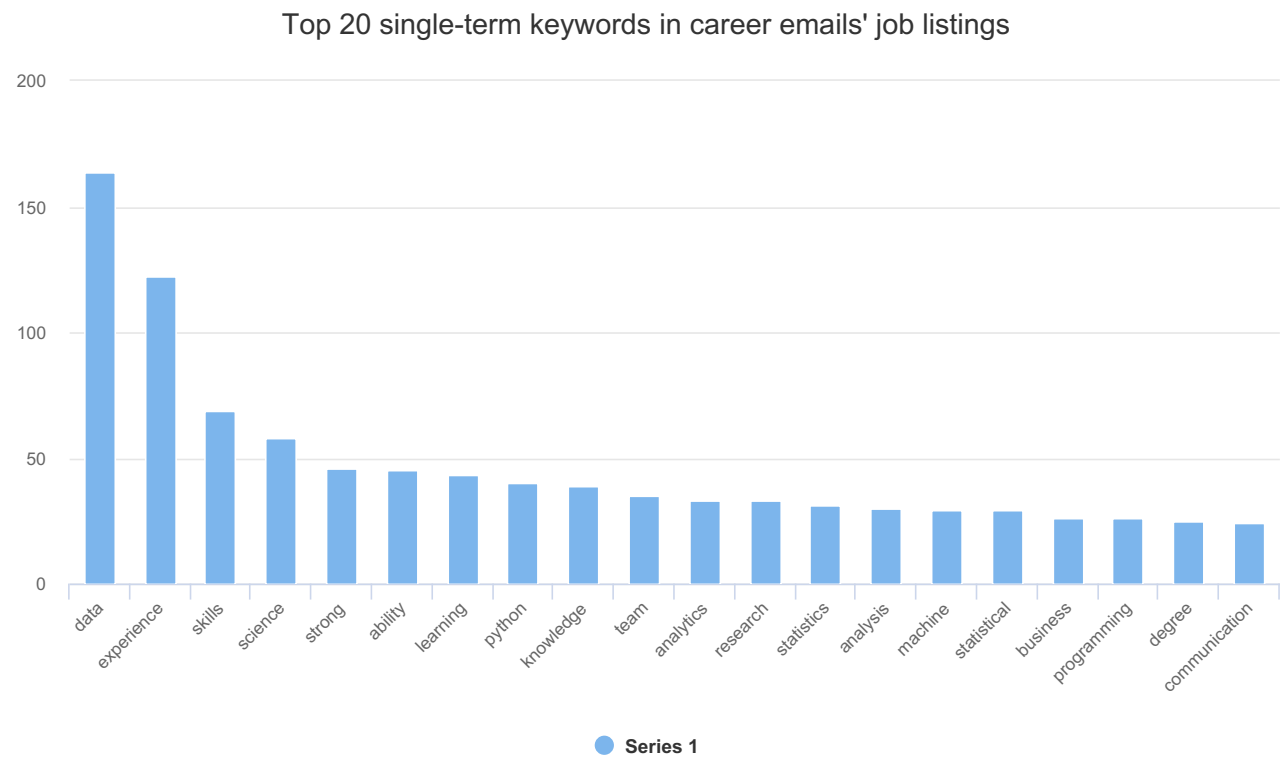
In this part, we used some text mining techniques to look for the top 20 keywords in Columbia University Data Science career opportunities emails. The difference to the previous part is that we did not use a pre-determined set of skills but directly studied the words' frequency in job listings. Therefore, the results may be messier and not as easy to interpret as the previous part.

We use R's tidytext library in this part of analysis.

```
#text mining
text_df <- data_frame(line = 1:nrow(rachel), text = rachel$Description)
#single term
tidy_decp_single <- text_df %>% unnest_tokens(word, text) %>% anti_join(stop_words)
#double term
tidy_decp_double <- text_df %>% unnest_tokens(word, text, token = "ngrams", n = 2)
bigrams_separated <- tidy_decp_double %>% separate(word, c("word1", "word2"), sep = " ")
bigrams_filtered <- bigrams_separated %>% filter(!word1 %in% stop_words$word) %>% filter(!word2 %in%
stop_words$word)
bigrams_united <- bigrams_filtered %>% unite(word, word1, word2, sep = " ")
```

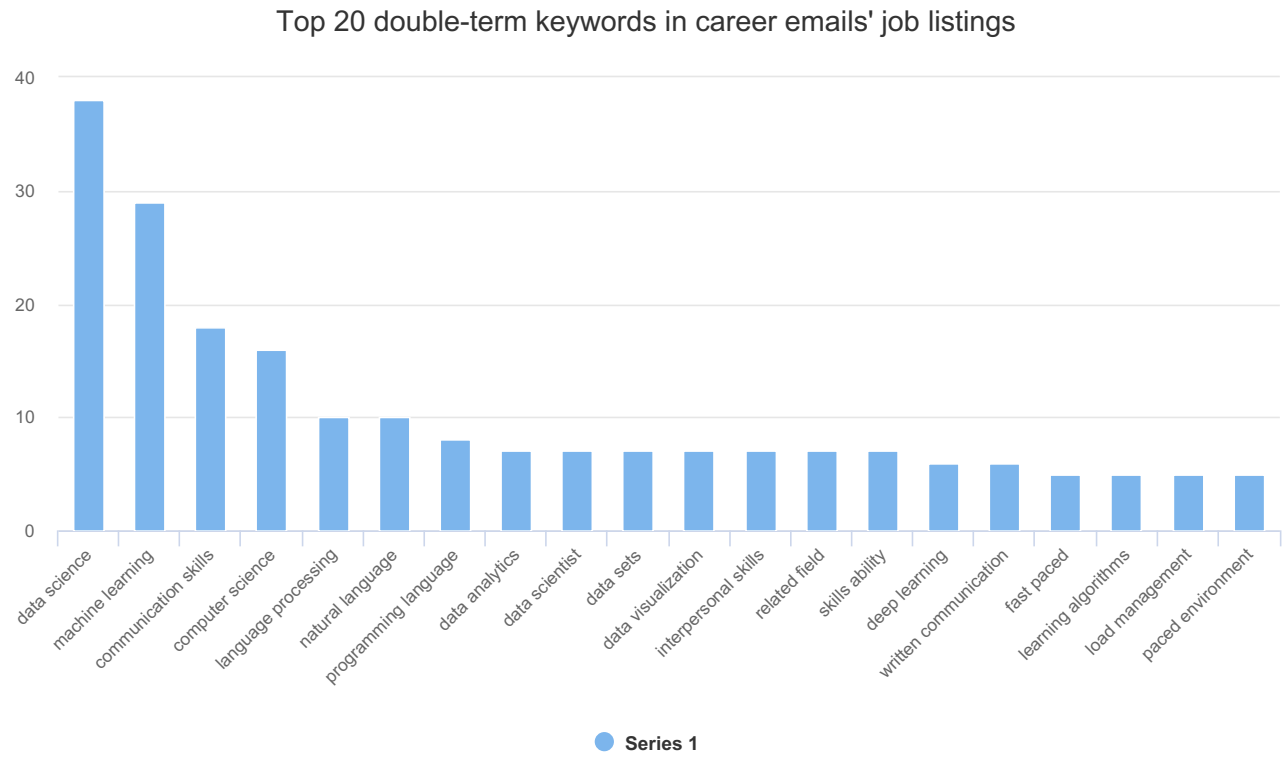
Single term frequency

```
top20words_single <- tidy_decp_single %>% count(word, sort = TRUE) %>% head(20)

highchart() %>% hc_xAxis(type = 'category') %>% hc_add_series(top20words_single, "column", hcaes(x =
word, y = n)) %>% hc_title(text="Top 20 single-term keywords in career emails' job listings")
```



Double term frequency

```
top20words_double <- bigrams_united %>% count(word, sort = TRUE) %>% head(20)

highchart() %>% hc_xAxis(type = 'category') %>% hc_add_series(top20words_double, "column", hcaes(x = word, y = n)) %>% hc_title(text="Top 20 double-term keywords in career emails' job listings")
```



We can see that the analysis of double term frequency is a lot more reasonable than single term analysis in our study, so we would interpret the analysis of double term frequency as our key findings.

The term “Data Science” can doubtlessly be the most mentioned keyword in data science job listings. The second keyword is “Machine Learning”, and its related subsets “language processing” and “natural language” are also topped ranked. Among the top 20 keywords, “communication skills”,“interpersonal skills” and “written communication” can all be categorized as communication skills and are the

third most mentioned keywords. “data analytics” follows up as the fourth with a series of data -related skills such as “data sets” and “data visualization”.

The result matches with our previous findings, that machine learning, communication skills and statistical data analytics are the most demanded general skills in data science job listings.

4.2.1.3 Supply of General Skills – statistical analytics, machine learning and communication

We used the survey question - How long have you been writing code to analyze data? and For how many years have you used machine learning methods (at work or in school)? in Kaggle ML and Data Science Survey, 2018 to analyze people’s experience level in statistical analysis and machine learning.

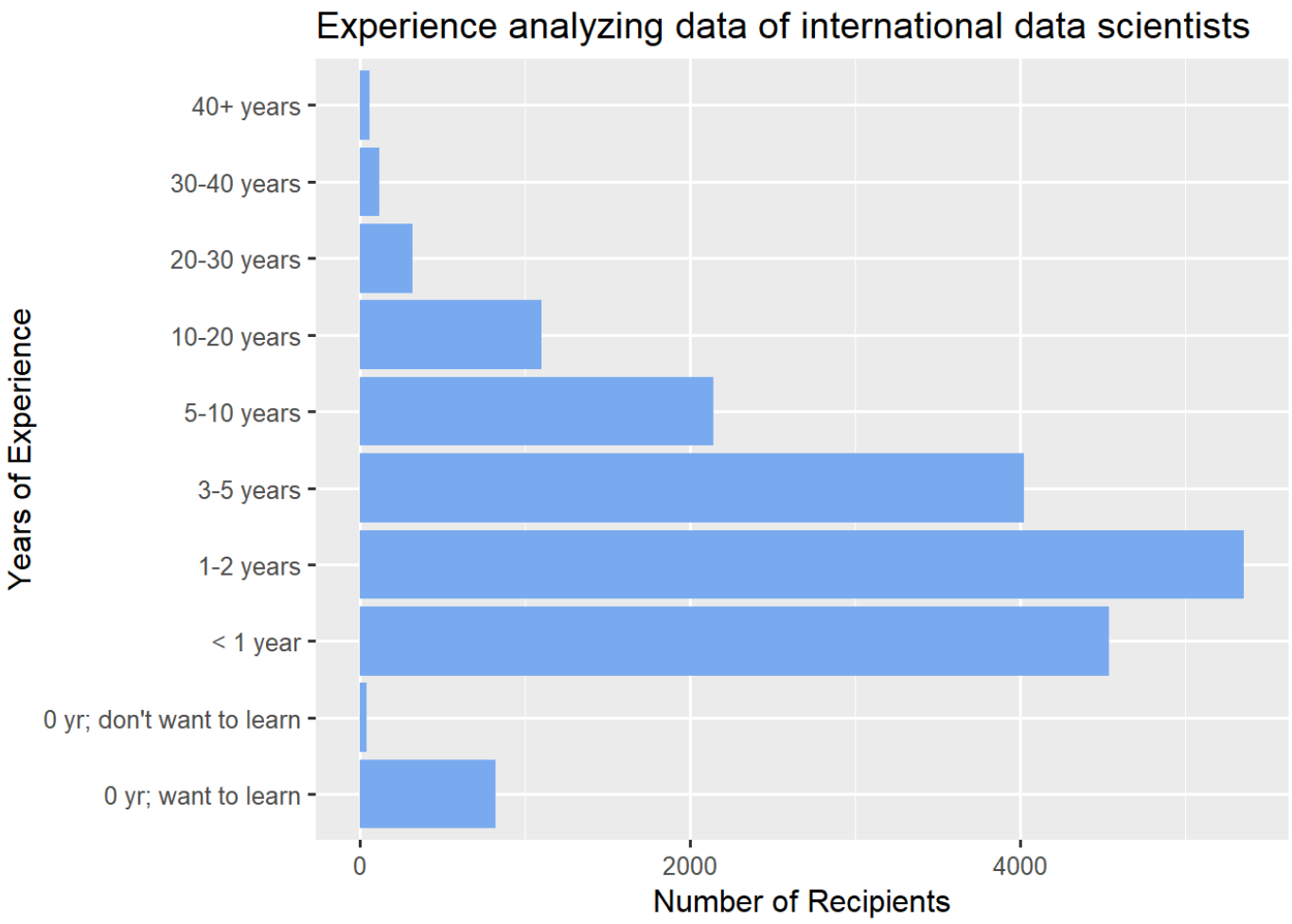
Data Cleaning

```
#clean the variables related to salary
sort_range <- function(df, x) {
 x <- enquo(x)
  df %>%
    mutate(var = !! x) %>%
    distinct(var) %>%
    separate(var, into = c("foo", "bar"), remove = FALSE, fill = "right", extra = "merge") %>%
    arrange(as.numeric(foo)) %>%
    mutate(var = as.character(var)) %>%
    .$var
}
vars <- c(exp_role = "How many years of experience do you have in your current role?",
          salary = "What is your current yearly compensation (approximate $USD)?",
          ml_at_work = "Does your current employer incorporate machine learning methods into their
business?",
          percent_code = "Approximately what percent of your time at work or school is spent actively
coding?",
          exp_code = "How long have you been writing code to analyze data?",
          exp_ml = "For how many years have you used machine learning methods (at work or in
school)?",
          ds = "Do you consider yourself to be a data scientist?")


df_survey_mcq <- df_survey_mcq %>%
  rename(!!vars) %>%
  mutate(salary = if_else(salary == "I do not wish to disclose my approximate yearly compensation",
"Undisclosed", as.character(salary))) %>%
  mutate(exp_code = as.factor(case_when(
    exp_code == "I have never written code but I want to learn" ~ "0 yr; want to learn",
    exp_code == "I have never written code and I do not want to learn" ~ "0 yr; don't want to learn"
,
    TRUE ~ as.character(exp_code)
  ))) %>%
  mutate(exp_ml = as.factor(case_when(
    exp_ml == "I have never studied machine learning but plan to learn in the future" ~ "0 yr; want
to learn",
    exp_ml == "I have never studied machine learning and I do not plan to" ~ "0 yr; don't want to le
arn",
    TRUE ~ as.character(exp_ml)
  )))

exp_role_lvl <- sort_range(df_survey_mcq, exp_role)
salary_lvl <- sort_range(df_survey_mcq, salary)

percent_lvl <- sort_range(df_survey_mcq, percent_code)
exp_code_lvl <- sort_range(df_survey_mcq, exp_code)
exp_ml_lvl <- sort_range(df_survey_mcq, exp_ml)

ds_lvl <- c("Definitely not", "Probably not", "Maybe", "Probably yes", "Definitely yes", "")

df_survey_mcq <- df_survey_mcq %>%
  mutate(exp_role = fct_relevel(exp_role, exp_role_lvl)) %>%
  mutate(salary = fct_relevel(salary, salary_lvl)) %>%
```

```
  mutate(percent_code = fct_relevel(percent_code, percent_lvl)) %>%
  mutate(exp_code = fct_relevel(exp_code, exp_code_lvl)) %>%
  mutate(exp_ml = fct_relevel(exp_ml, exp_ml_lvl)) %>%
  mutate(ds = fct_relevel(ds, ds_lvl)) %>%
  mutate(exp_role = na_if(exp_role, "")) %>%
  mutate(salary = na_if(salary, "")) %>%
  mutate(percent_code = na_if(percent_code, "")) %>%
  mutate(exp_code = fct_relevel(exp_code, "< 1 year", after = 2)) %>%
  mutate(exp_code = na_if(exp_code, "")) %>%
  mutate(exp_ml = fct_relevel(exp_ml, "< 1 year", after = 2)) %>%
  mutate(exp_ml = na_if(exp_ml, "")) %>%
  mutate(ds = na_if(ds, ""))
```

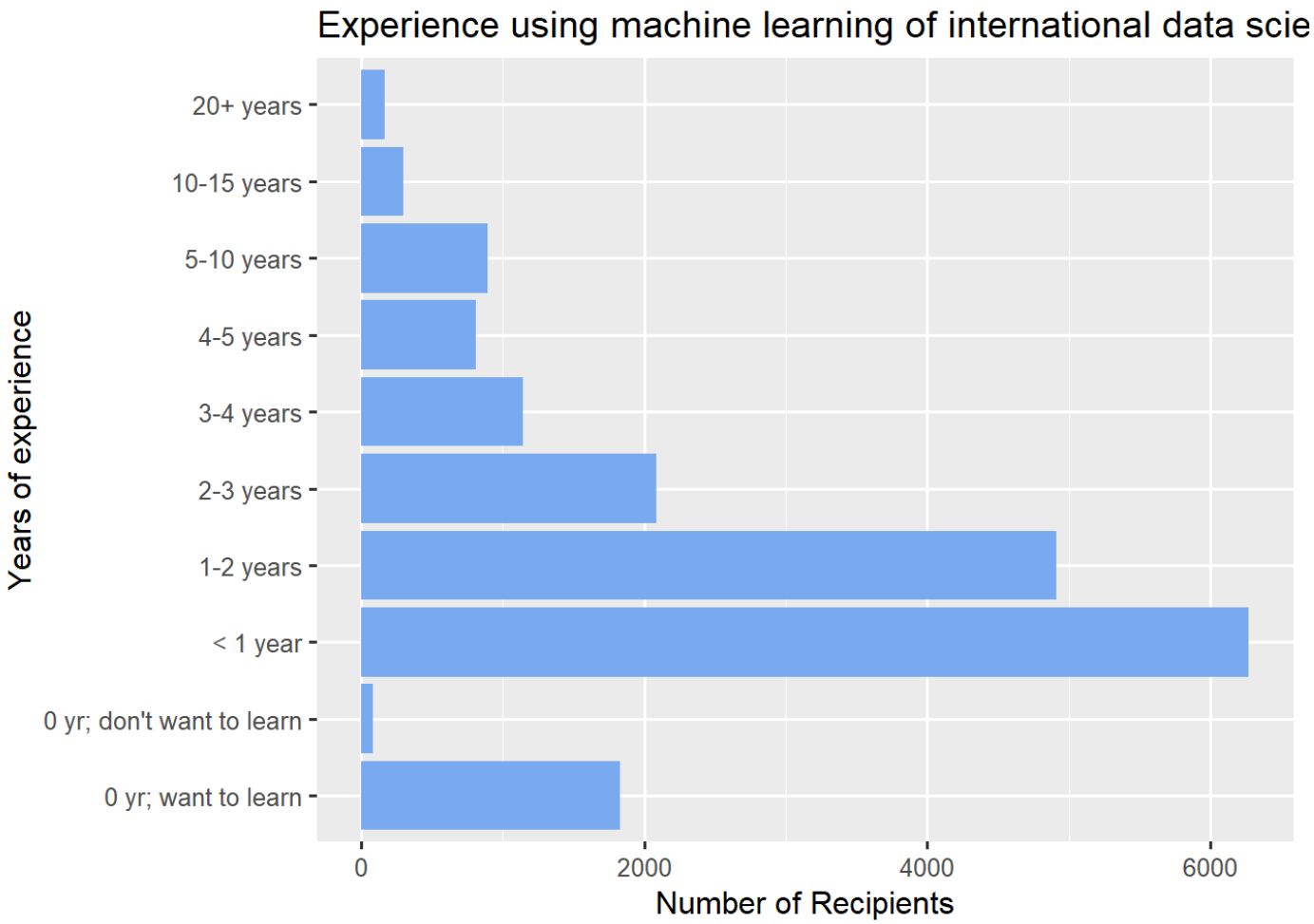
4.2.1.3.1 Statistical analytics

```
#visualize the experience with analyzing data
df_survey_mcq %>%
  filter(!is.na(exp_code)) %>%
  ggplot(aes(exp_code)) +
  geom_bar(fill="#79AAEF") +
  theme(legend.position = "none", axis.text.x  = element_text(angle=35, hjust=1, vjust=0.9)) +
  labs(x = "Years of Experience", y = "Number of Recipients") +
  ggtitle("Experience analyzing data of international data scientists")+
  theme_grey(12)+
  coord_flip()
```



4.2.1.3.2 Machine Learning

```r
#visualize the experience with machine learning
df_survey_mcq %>%
  filter(!is.na(exp_ml)) %>%
  ggplot(aes(exp_ml)) +
  geom_bar(fill="#79AAEF") +
  theme(legend.position = "none", axis.text.x  = element_text(angle=35, hjust=1, vjust=0.9)) +
  labs(x = "Years of experience", y = "Number of Recipients") +
  ggtitle("Experience using machine learning of international data scientists")+
  theme_grey(12)+
  coord_flip()
```



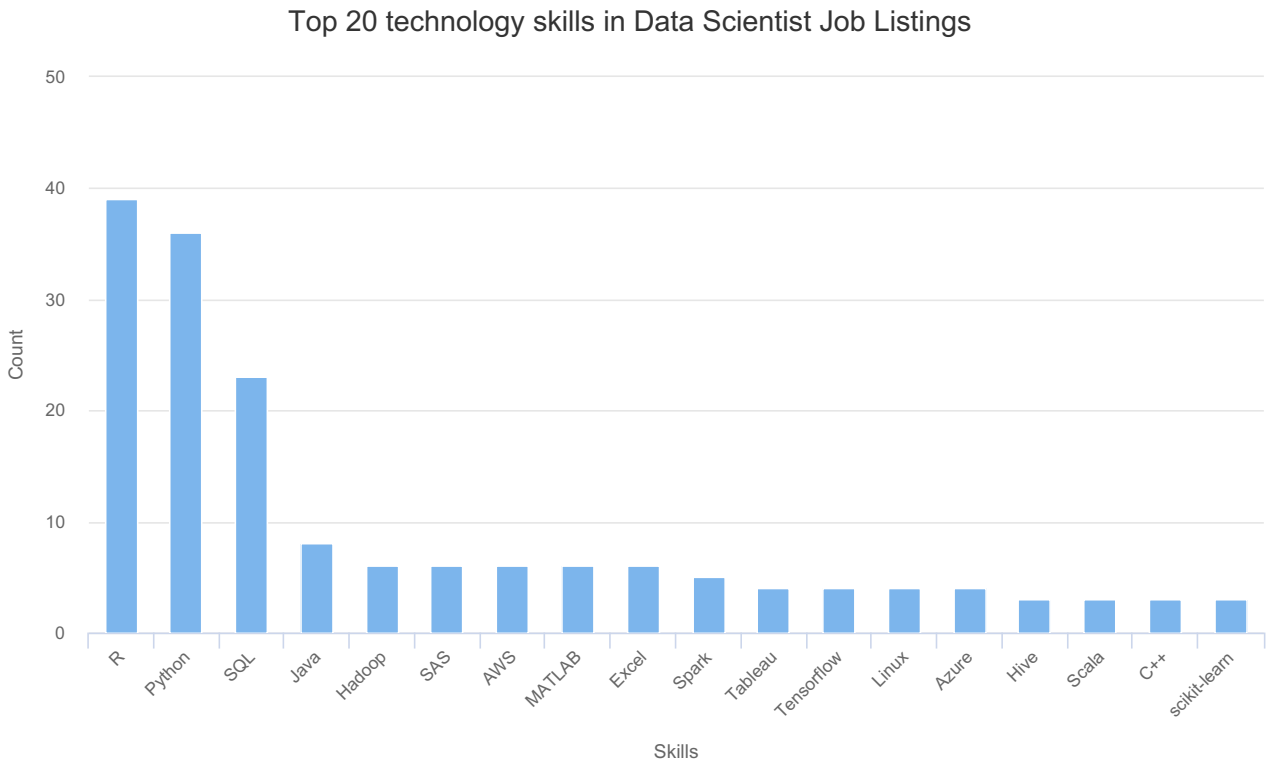
Insights: From the data it is clear that there are more people who are new to using machine learning, but most of the people in the job market have been analyzing data for at least 1-2 years.

4.2.2 Technical Skills

4.2.2.1 Demand – Technical Skills Wanted in Job Listings

In this part, we also used the technical skills listed in Jeff Hale's study and searched for their occurences in the job listing requirements in Columbia University Data Science career opportunities emails.

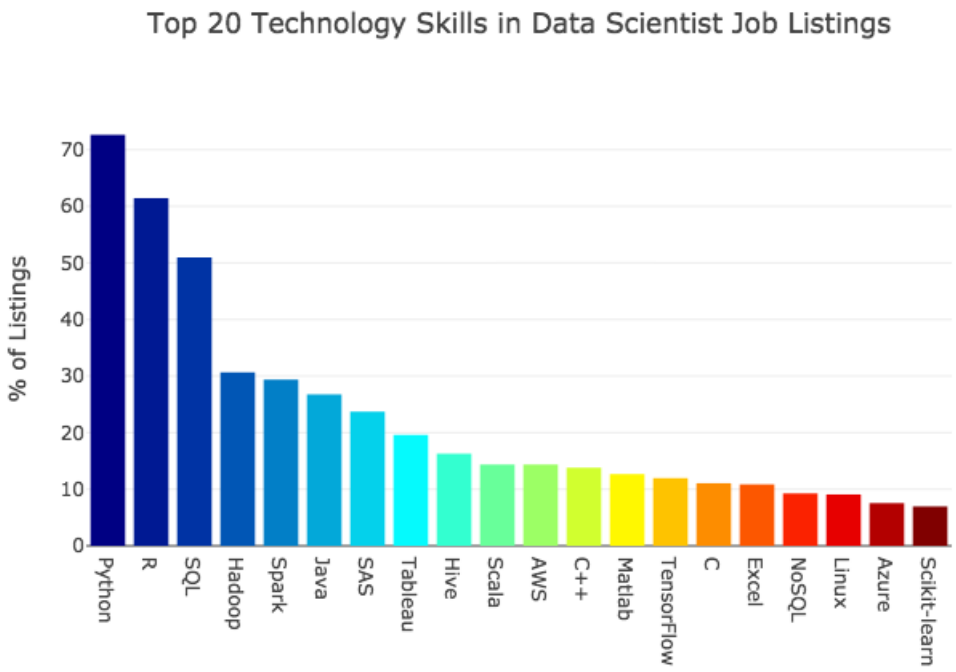
```
#keywords
skills <- data.frame("Skills" = c("Python", "R", "SQL", "Hadoop", "Spark", "Java", "SAS", "Tableau",
"Hive", "Scala", "AWS", "C++", "MATLAB", "Tensorflow", "Excel", "Linux", "Azure", "scikit-learn"), "C
ount" = c(length(grep("python", tolower(rachel$Description))), length(grep("R", rachel$Description)),
length(grep("SQL", rachel$Description)), length(grep("hadoop", tolower(rachel$Description))), length(
grep("spark", tolower(rachel$Description))), length(grep("java", tolower(rachel$Description))),
length(grep("SAS", rachel$Description)), length(grep("tableau", tolower(rachel$Description))), length
(grep("hive", tolower(rachel$Description))), length(grep("scala", tolower(rachel$Description))),
length(grep("AWS", rachel$Description)), length(grep("C\\+\\+", rachel$Description)), length(grep("ma
tlab", tolower(rachel$Description))), length(grep("tensorflow", tolower(rachel$Description))),
sum(str_count( tolower(rachel$Description), "\\bexcel\\b")), length(grep("linux",
tolower(rachel$Description))), sum(str_count( tolower(rachel$Description), "\\bazure\\b")),
length(grep("scikit-learn", tolower(rachel$Description))) ))

#plot
skills %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = Skills, y = Count)) %>% hc_xAxi
s(type = 'category') %>% hc_title(text="Top 20 technology skills in Data Scientist Job Listings")
```



And the result from Jeff Hale's study on major job listing websites is as follow:

(Source: https://www.kdnuggets.com/2018/11/most-demand-skills-data-scientists.html)

Note: We didn't create this plot. This plot was from Jeff Hale's study on major job listing websites listed just for reference and comparison



Our results show that R, Python and SQL are the most demanded technical skills in data scientist jobs, and that matches with Jeff Hale's study on major job listing websites. Python and R are not very far off from each other and are dominant in frequency, which makes the two languages a must for virtually every data scientist position. SQL is also in high demand. SQL stands for Structured Query Language and is the primary way to interact with relational database.

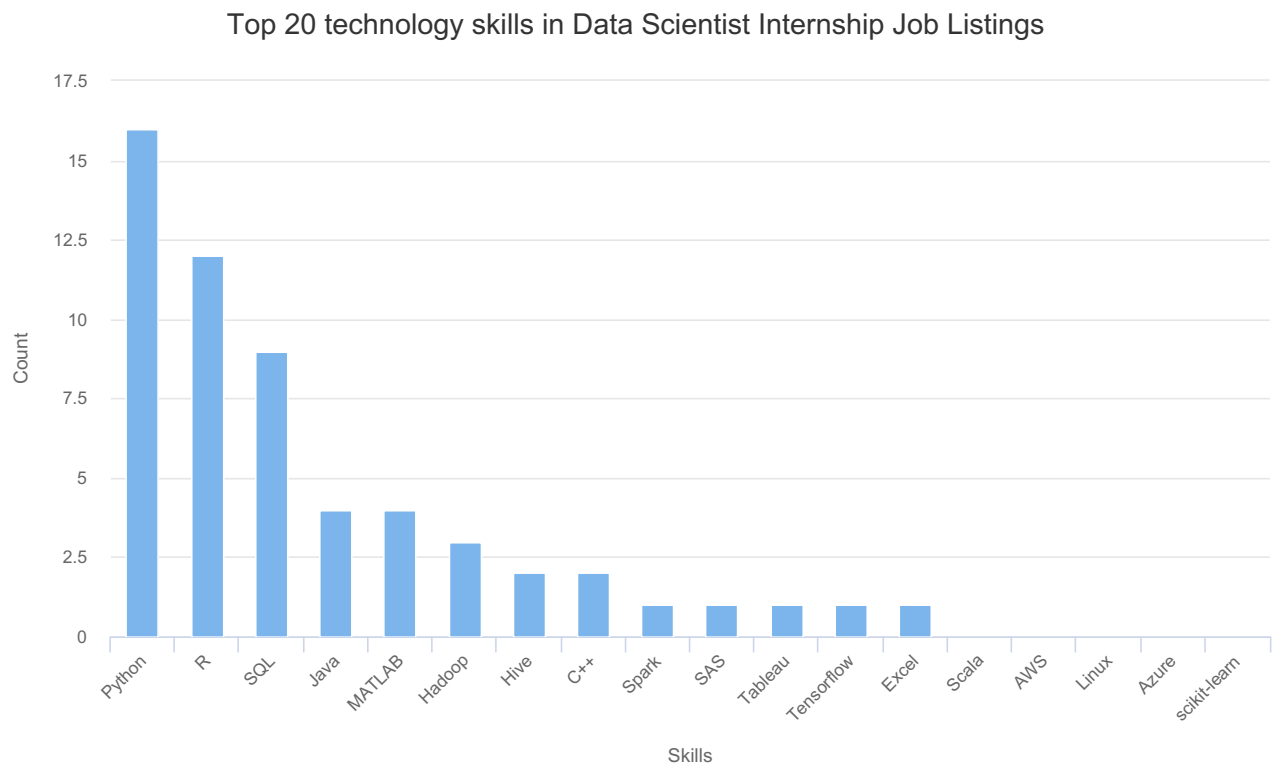
4.2.2.2 Demand – Technical Skills Wanted in Job Listings By Job Type

We are also interested to know if the technical skills wanted in job listings would vary by job types. In Columbia University Data Science career opportunities emails, both internship positions and full-time jobs positions are listed together. We want to know if the industry would be looking for different skills in the two types of applicants.

```
#Filter data set
rachel_intern <- rachel[rachel$Type == "Internship",]
rachel_full <- rachel[rachel$Type == "Full Time",]

#Internship
skills_intern <- data.frame("Skills" = c("Python", "R", "SQL", "Hadoop", "Spark", "Java", "SAS", "Tableau", "Hive", "Scala", "AWS", "C++", "MATLAB", "Tensorflow", "Excel", "Linux", "Azure", "scikit-learn"), "Count" = c(length(grep("python", tolower(rachel_intern$Description))), length(grep("R", rachel_intern$Description)), length(grep("SQL", rachel_intern$Description)), length(grep("hadoop", tolower(rachel_intern$Description))), length(grep("spark", tolower(rachel_intern$Description))), length(grep("java", tolower(rachel_intern$Description))), length(grep("SAS", rachel_intern$Description)), length(grep("tableau", tolower(rachel_intern$Description))), length(grep("hive", tolower(rachel_intern$Description))), length(grep("scala", tolower(rachel_intern$Description))), length(grep("AWS", rachel_intern$Description)), length(grep("C\\+\\+", rachel_intern$Description)), length(grep("matlab", tolower(rachel_intern$Description))), length(grep("tensorflow", tolower(rachel_intern$Description))), sum(str_count( tolower(rachel_intern$Description), "\\bexcel\\b")), length(grep("linux", tolower(rachel_intern$Description))), sum(str_count( tolower(rachel_intern$Description), "\\bazure\\b")), length(grep("scikit-learn", tolower(rachel_intern$Description))) ))

#Full-Time
skills_full <- data.frame("Skills" = c("Python", "R", "SQL", "Hadoop", "Spark", "Java", "SAS", "Tableau", "Hive", "Scala", "AWS", "C++", "MATLAB", "Tensorflow", "Excel", "Linux", "Azure", "scikit-learn"), "Count" = c(length(grep("python", tolower(rachel_full$Description))), length(grep("R", rachel_full$Description)), length(grep("SQL", rachel_full$Description)), length(grep("hadoop", tolower(rachel_full$Description))), length(grep("spark", tolower(rachel_full$Description))), length(grep("java", tolower(rachel_full$Description))), length(grep("SAS", rachel_full$Description)), length(grep("tableau", tolower(rachel_full$Description))), length(grep("hive", tolower(rachel_full$Description))), length(grep("scala", tolower(rachel_full$Description))), length(grep("AWS", rachel_full$Description)), length(grep("C\\+\\+", rachel_full$Description)), length(grep("matlab", tolower(rachel_full$Description))), length(grep("tensorflow", tolower(rachel_full$Description))), sum(str_count(
tolower(rachel_full$Description), "\\bexcel\\b")), length(grep("linux",
tolower(rachel_full$Description))), sum(str_count( tolower(rachel_full$Description), "\\bazure\\b")), length(grep("scikit-learn", tolower(rachel_full$Description))) ))
```

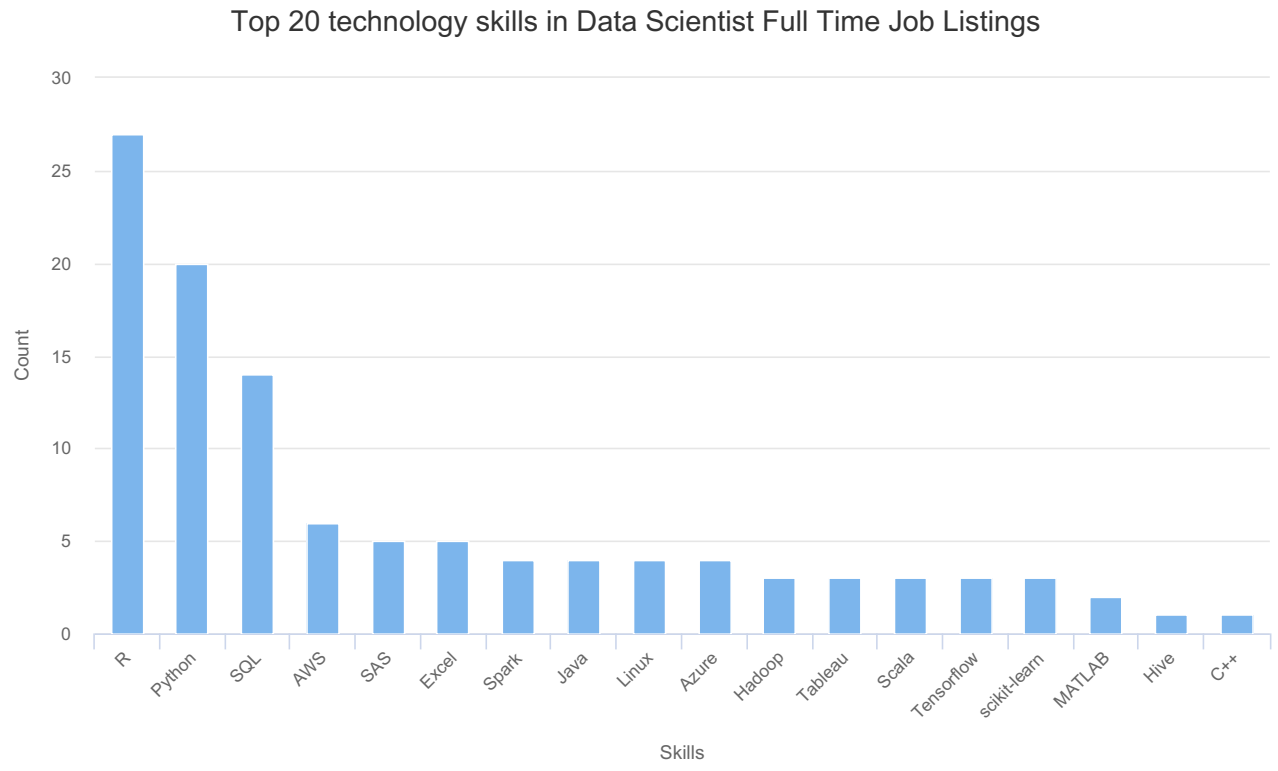
Top 20 technology skills in Data Scientist Internship Job Listings

```
skills_intern %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = Skills, y = Count)) %>% hc_xAxis(type = 'category') %>% hc_title(text="Top 20 technology skills in Data Scientist Internship Job Listings")
```



Top 20 technology skills in Data Scientist Full Time Job Listings

```
skills_full %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = Skills, y = Count)) %>% hc_xAxis(type = 'category') %>% hc_title(text="Top 20 technology skills in Data Scientist Full Time Job Listings")
```



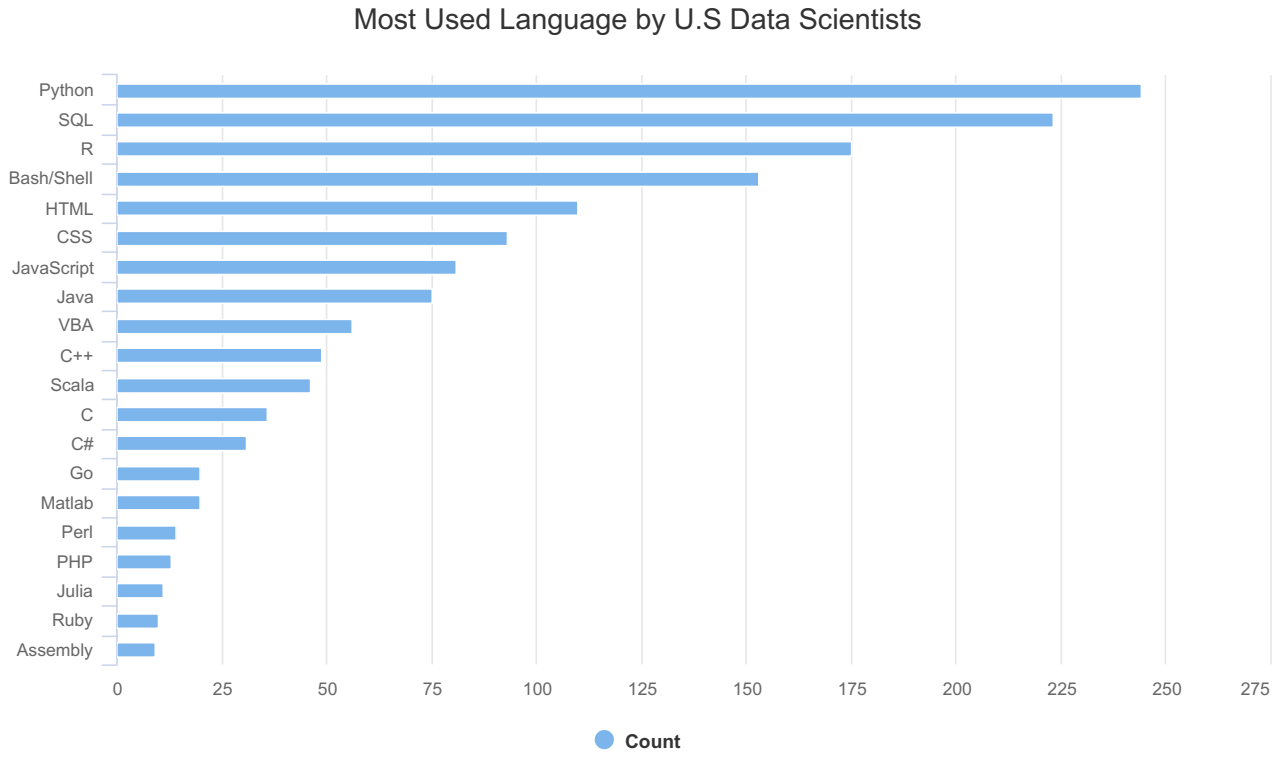
We can see that R, Python and SQL are still the top three demanded technical skill, but the rank is a little different by job types. The rank in full-time job listings matches with our previos findings, whereas R has a slight edge over Python in internship jobs. The technical skills required in internship positions are also less than full-time job positions.

4.2.2.3 Supply – Technical Skills People Have (USA data)

In this part, we used Stack Overflow 2018 Developer Survey to analyze the technical skills U.S based data scientists have. The analysis

filters full-time workers, Data Scientists, US workers, and their missing data.

```
#This code graphs a histogram by mostly used languages in decreasing order.
#Filters full-time workers, Data Scientists, US workers, and their missing data

languae_hist = stackoverflow %>%
  filter(Employment %in% 'Employed full-time') %>%  #filter full-time employees
  filter(!is.na(LanguageWorkedWith))%>% #filter missing data
  filter(!is.na(DevType)) %>%
  filter(DevType %in% c('Data or business analyst','Data scientist or machine learning specialist'))
%>% #filter data scientists
  filter(Country %in% 'United States') %>% #filter US workers
  select(LanguageWorkedWith)%>%
  mutate(LanguageWorkedWith = str_split(LanguageWorkedWith, pattern = ";"))%>%
  unnest(LanguageWorkedWith)%>%
  group_by(LanguageWorkedWith)%>%
  summarise(Count = n())%>% #count by languages
  arrange(desc(Count))%>% #reorder in descending order
  ungroup()%>%
  mutate(LanguageWorkedWith = reorder(LanguageWorkedWith, Count))


#slice only top 20 data
languae_hist = slice(languae_hist, 0:20)

highchart()%>% #
  hc_title(text = paste("Most Used Language by U.S Data Scientists"))%>% #title
  hc_xAxis(categories = languae_hist$LanguageWorkedWith)%>% #xaxis
  hc_add_series(data = languae_hist$Count, name = "Count", type = "bar") #plot
```



The purpose of this histogram is to explore which language data scientists in U.S. use. This data filters full-time data scientists, working in the US, which conists of 317 rows of data. The X-axis is count, and the Y-axis denotes each languages sorted in decreasing order of the counts. Our result shows that most data scientists in U.S use Python, SQL, and R.

4.2.2.4 Supply – Technical Skills People Have (World Data)

In this part, we used Kaggle ML and Data Science Survey, 2018 to analyze the technical skills international data scientists have. The data is collected by two questions in the survey: What specific programming language do you use most often? - Selected Choice

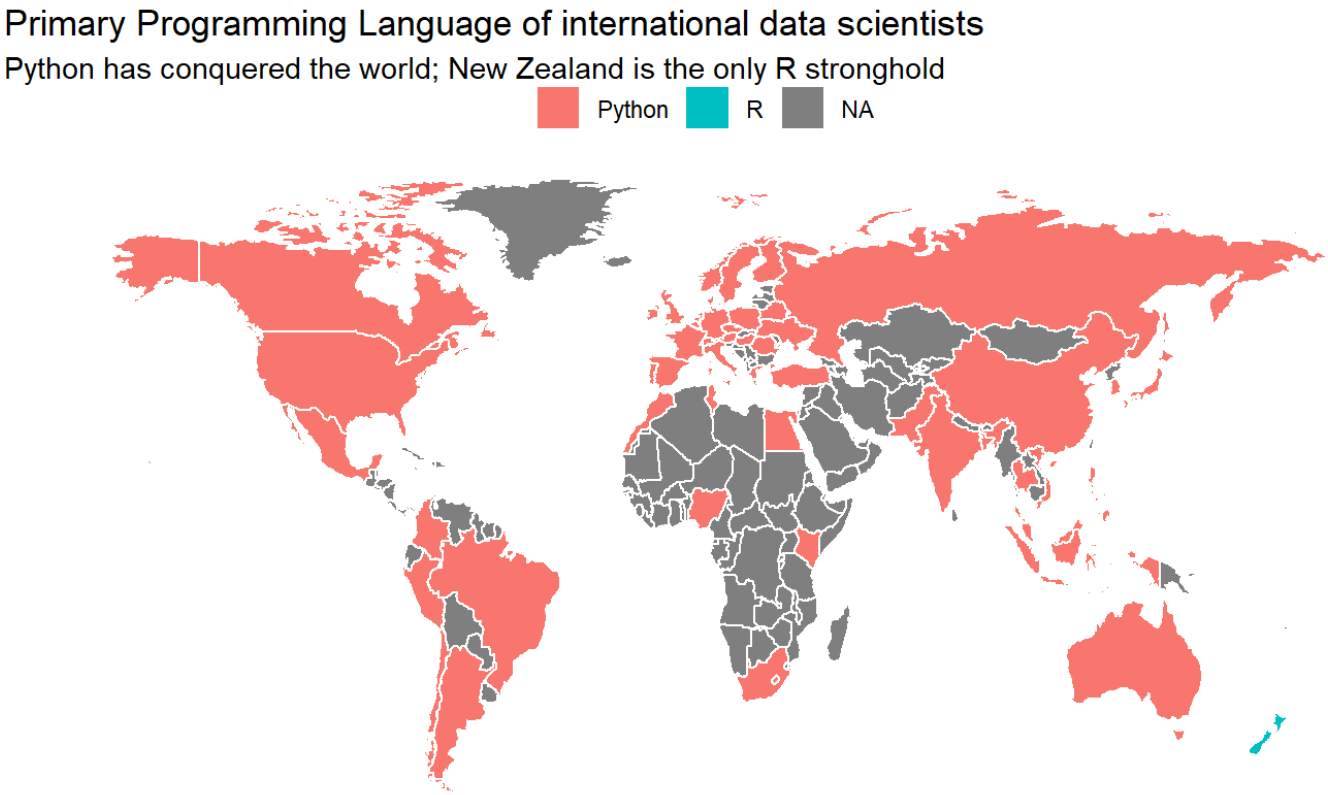
Data cleaning

```r
#clean the programming language attribute
vars <- c(lang =  "What specific programming language do you use most often? - Selected Choice",
          first_lang = "What programming language would you recommend an aspiring data scientist to
learn first? - Selected Choice",
          lang_py = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - Python",
          lang_r = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - R",
          lang_sql = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - SQL",
          lang_jula = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - Julia",
          lang_bash = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - Bash",
          lang_java = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - Java",
          lang_js = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - Javascript/Typescript",
          lang_vs = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - Visual Basic/VBA",
          lang_c = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - C/C++",
          lang_matlab = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - MATLAB",
          lang_scala = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - Scala",
          lang_go = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - Go",
          lang_sharp = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - C#/.NET",
          lang_php = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - PHP",
          lang_ruby = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - Ruby",
          lang_sas = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - SAS/STATA",
          lang_other = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - Other",
          no_lang = "What programming languages do you use on a regular basis? (Select all that
apply) - Selected Choice - None")

df_survey_mcq <- df_survey_mcq %>%
  rename(!!vars) %>%
  mutate(lang = na_if(lang, ""),
         first_lang = na_if(first_lang, ""),
         no_lang = na_if(no_lang, "")) %>%
  mutate_at(vars(starts_with("lang_")), as.integer) %>%
  mutate_at(vars(starts_with("lang_")), log) %>%
  mutate_at(vars(starts_with("lang_")), as.logical)
```

```r
lang_lvl <- df_survey_mcq %>%
  filter(!is.na(lang)) %>%
  mutate(lang = as.character(lang)) %>%
  count(lang) %>%
  arrange(desc(n)) %>%
  .$lang

first_lang_lvl <- df_survey_mcq %>%
  filter(!is.na(first_lang)) %>%
  mutate(first_lang = as.character(first_lang)) %>%
  count(first_lang) %>%
  arrange(n) %>%
  .$first_lang
```

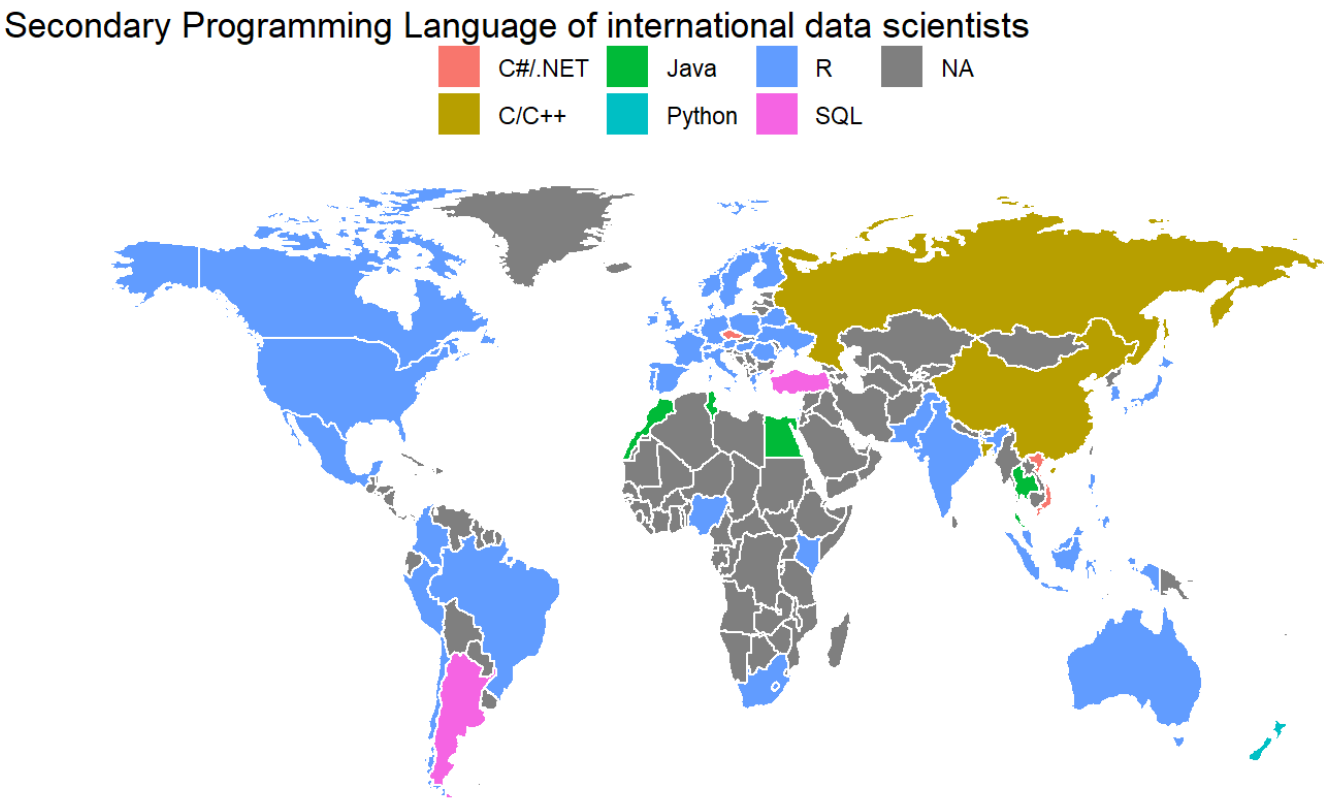
First popular programming language

```
#visualize the first popular language
foo <- df_survey_mcq %>%
  filter(!(country %in% c("Other", "I do not wish to disclose my location"))) %>%
   mutate(country = as.character(case_when(
    country == "United Kingdom of Great Britain and Northern Ireland" ~ "UK",
    country == "United States of America" ~ "USA",
    country == "Viet Nam" ~ "Vietnam",
    TRUE ~ as.character(country)
  ))) %>%
  group_by(lang, country) %>%
  filter(!is.na(lang)) %>%
  count() %>%
  arrange(desc(n)) %>%
  group_by(country) %>%
  slice(c(1))

world %>%
  filter(region != "Antarctica") %>%
  left_join(foo, by = c("region" = "country")) %>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, fill = lang, group = group), color = "white") +
  coord_fixed(1.3) +
  labs(fill = "") +
  theme(legend.position = "top") +
  theme_void() +
  theme(legend.position = "top") +
  ggtitle("Primary Programming Language of international data scientists",
          subtitle = "Python has conquered the world; New Zealand is the only R stronghold")
```



Second popular programming language

```r
#visualize 2nd popular language
foo <- df_survey_mcq %>%
  filter(!(country %in% c("Other", "I do not wish to disclose my location"))) %>%
   mutate(country = as.character(case_when(
     country == "United Kingdom of Great Britain and Northern Ireland" ~ "UK",
     country == "United States of America" ~ "USA",
     country == "Viet Nam" ~ "Vietnam",
     TRUE ~ as.character(country)
  ))) %>%
  group_by(lang, country) %>%
  filter(!is.na(lang)) %>%
  count() %>%
  arrange(desc(n)) %>%
  group_by(country) %>%
  slice(c(2))

world %>%
  filter(region != "Antarctica") %>%
  left_join(foo, by = c("region" = "country")) %>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, fill = lang, group = group), color = "white") +
  coord_fixed(1.3) +
  labs(fill = "") +
  theme(legend.position = "top") +
  theme_void() +
  theme(legend.position = "top") +
  ggtitle("Secondary Programming Language of international data scientists")
```



Insights: We can see that most of the world the promary language used for programming in data science is python except New Zealand. The second most popular programming language though seems to fluctuate between R which is common among most of the world. But China and Russia seems to prefer C++ as their language of choice as their secondary programming language. This might be arising from their preference to working with image data as opencv is mainly in C++.

4.2.3 Section Conclusion

In this section, we studied what skills are in the current job market. We broke down the skills to two parts, general skills and technical skills. We also evaluated from both demand and supply to understand what skills are demanded by the industry and what skills do people in the data science job market possess.

We discovered that the job market in data science is in demand of people who are good at Statistical Analysis, Communication and Machine Learning. The demand is both seen in Columbia University data science career opportunities emails and major job listing websites in the United States.

We found that most people on the job market are still relatively new to machine learning, but already have some years of experiences in statistical analysis. There is also a great percentage of people who don't have any experience in machine learning yet but are eager to learn. Therefore, we see a positive relationship between the supply and demand of the job market, and we are confident that the supply can soon meet the demand.

As for technical skills, the job market in data science is in demand of people who are able to code in Python, R and SQL. The three programming langauges' order may differ due to job types, but are still the top three dominant programming languages required in job listings.

We were delighted to find that the supply and demand of the market match with each other, that people in the job market are also most comfortable working with Python and R. This is both the case for data scientists in U.S and in the world.

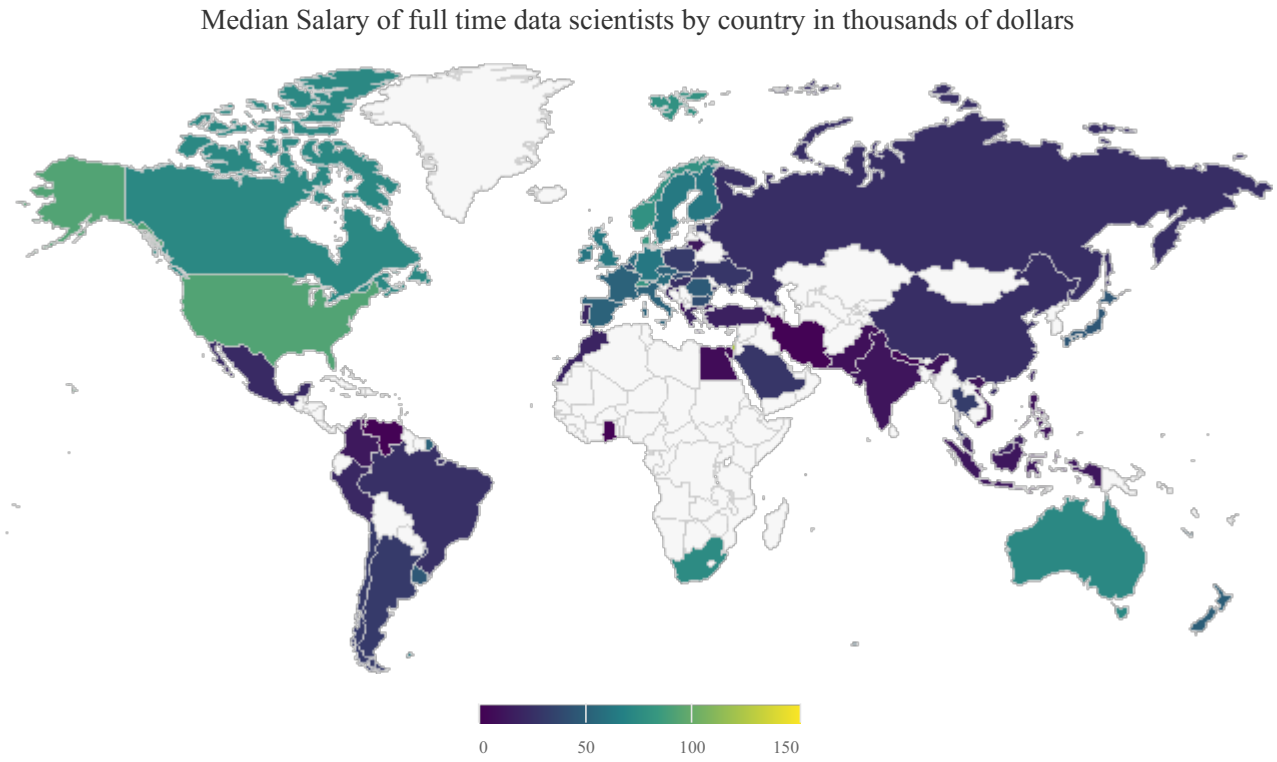
4.3 What prospects can we expect from the market - Salary analysis

Our next question is: What prospects can we expect from the market? As people are preparing themselves with the skills demanded to enter the job market, we want to know how these skills can transform into salary for people. We used Stack Overflow 2018 Developer Survey's data for this part of analysis.

We will look into how salary is affected by the following features:
(1) Location
(2) Programming languages people use
(3) Databases people work with
(4) Platforms people work with

4.3.1 Median salary by location

```
by_country_salary <- stackoverflow %>% select(Country, ConvertedSalary, DevType) %>%
filter(!is.na(DevType)) %>%
  filter(DevType %in% c('Data or business analyst','Data scientist or machine learning specialist'))
%>% #filter data scientists
  mutate(ConvertedSalary=as.numeric(ConvertedSalary))  %>% filter(!is.na(Country)) %>% filter(!is.na(
ConvertedSalary)) %>%
  group_by(Country) %>% summarize(MedSalary = median(ConvertedSalary, na.rm=TRUE))

data(worldgeojson, package = "highcharter") #using highcharter
code <- countrycode(by_country_salary$Country, 'country.name', 'iso3c') #get country code

by_country_salary$iso3 <- code
by_country_salary$MedSalary <- round(by_country_salary$MedSalary/1000) #round

#plot

highchart() %>%
  hc_add_series_map(worldgeojson, by_country_salary, value = "MedSalary", joinBy = "iso3") %>%
  hc_colorAxis(stops = color_stops()) %>%
  hc_legend(enabled = TRUE) %>%
  hc_title(text = "Median Salary of full time data scientists by country in thousands of dollars")
%>%
  hc_tooltip(useHTML = TRUE, headerFormat = "",
             pointFormat = "Country: {point.Country} / Median Salary: ${point.MedSalary}K") %>% hc_a
dd_theme(hc_theme_google())
```

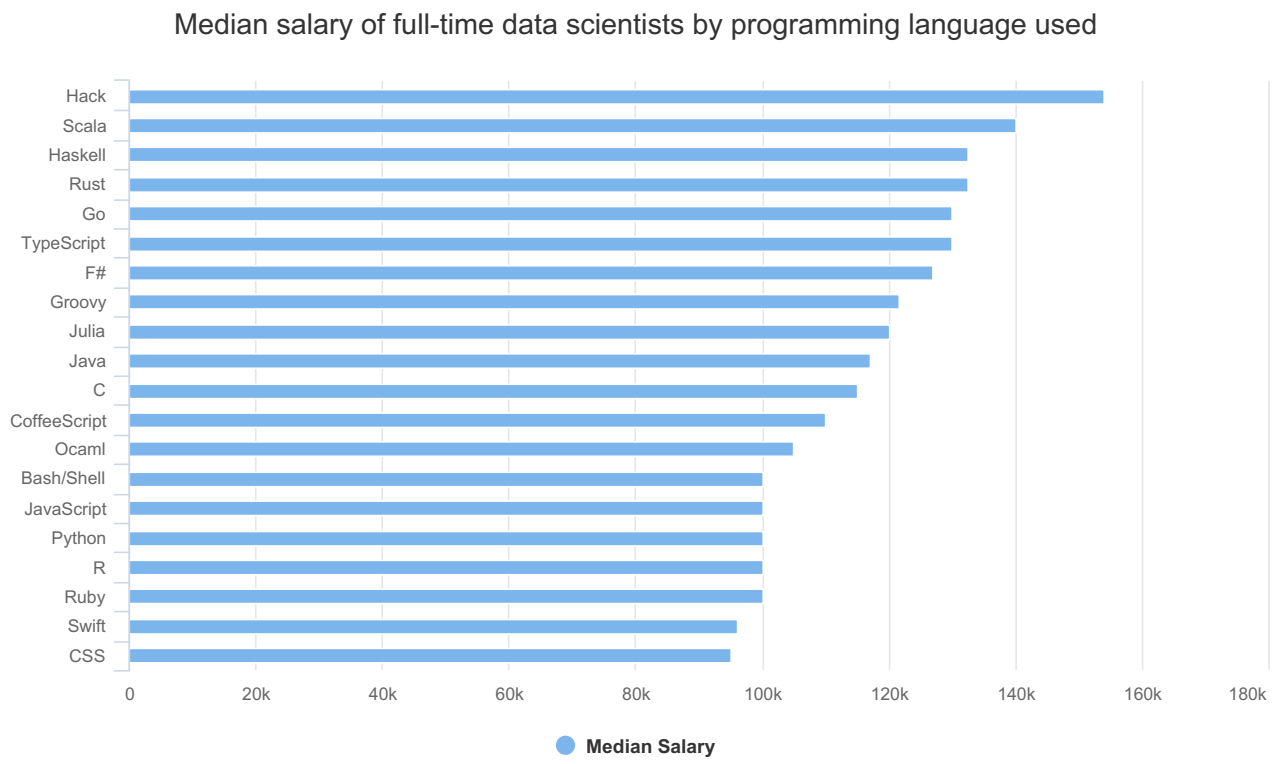


This world map shows median salary for full time data scientists by country. Brighter colored countries have higher median salary, and the darkered colored countries vice versa. The countries without a color are countries with no data.

We can see from the chart that USA has the highest median salary for data scientists, which is around $92k per year.

4.3.2 Median salary by programming language used

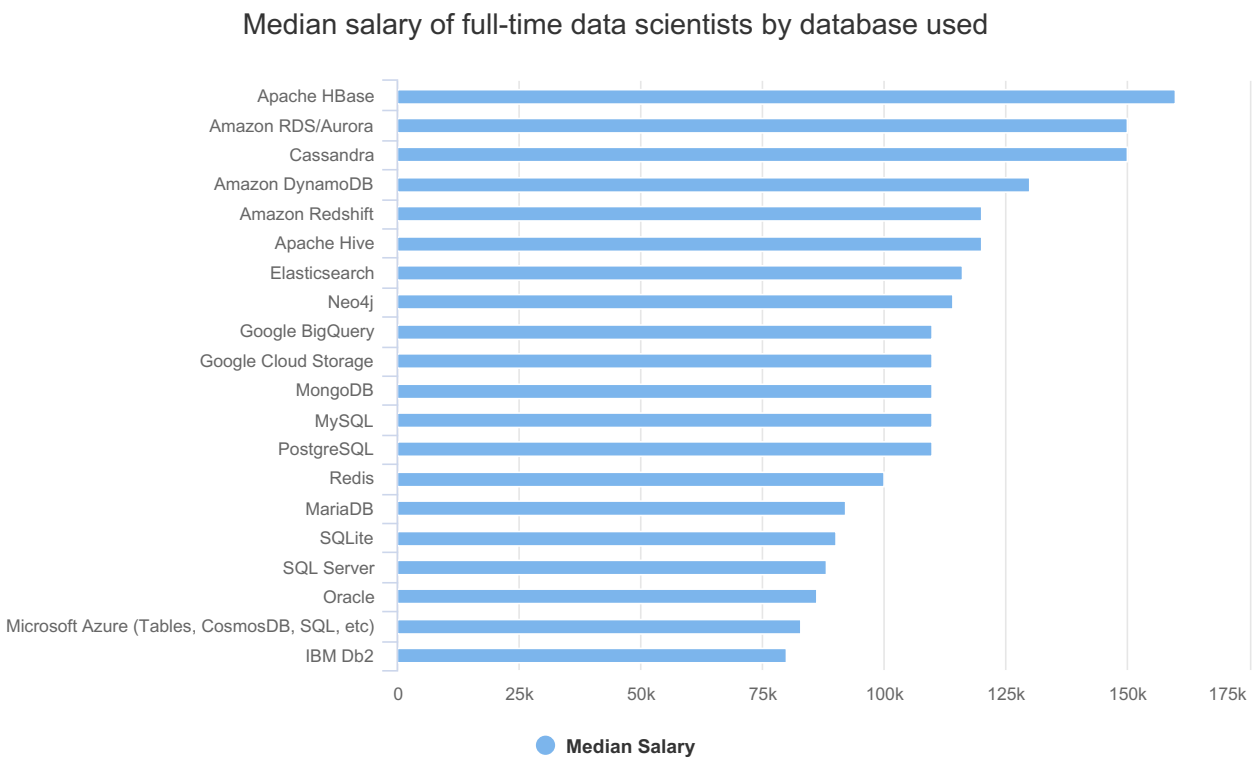
```
#This code graphs a histogram of median salary by most used languages by decreasing order.
#Filters full-time workers, Data Scientists, US workers, and their missing data

median_sal_lang = stackoverflow %>%
  filter(Employment %in% 'Employed full-time') %>%
  filter(!is.na(LanguageWorkedWith))%>%
  filter(!is.na(DevType)) %>%
  filter(DevType %in% c('Data or business analyst','Data scientist or machine learning specialist'))
%>%     filter(Country %in% 'United States') %>%
  select(LanguageWorkedWith,ConvertedSalary) %>% #The data already converted salaries to USD
  mutate(LanguageWorkedWith = str_split(LanguageWorkedWith, pattern = ";")) %>%
  unnest(LanguageWorkedWith) %>%
  group_by(LanguageWorkedWith) %>%
  summarise(Median_Salary = median(ConvertedSalary,na.rm = TRUE)) %>% #summarise each language by sa
lary
  arrange(desc(Median_Salary))%>% #descending order
  ungroup()%>%
  mutate(LanguageWorkedWith = reorder(LanguageWorkedWith, Median_Salary))

#slice top 20 data
median_sal_lang = slice(median_sal_lang, 0:20)

#plot
highchart()%>%
  hc_title(text = paste("Median salary of full-time data scientists by programming language used"))%
>%
  hc_xAxis(categories = median_sal_lang$LanguageWorkedWith)%>%
  hc_add_series(data = median_sal_lang$Median_Salary, name = "Median Salary", type = "bar")
```



This shows a median salary by language data scientists use, and the data consists of 317 rows. It is interesting to see that the newly introduced languages such as Hack or Scala have higher median salary compared to the well known languages such as Python or R. However, median salary for Python and R are both 100k which corresponds to the overall data scientists' median salary in the US.

4.3.3 Median salary by database used

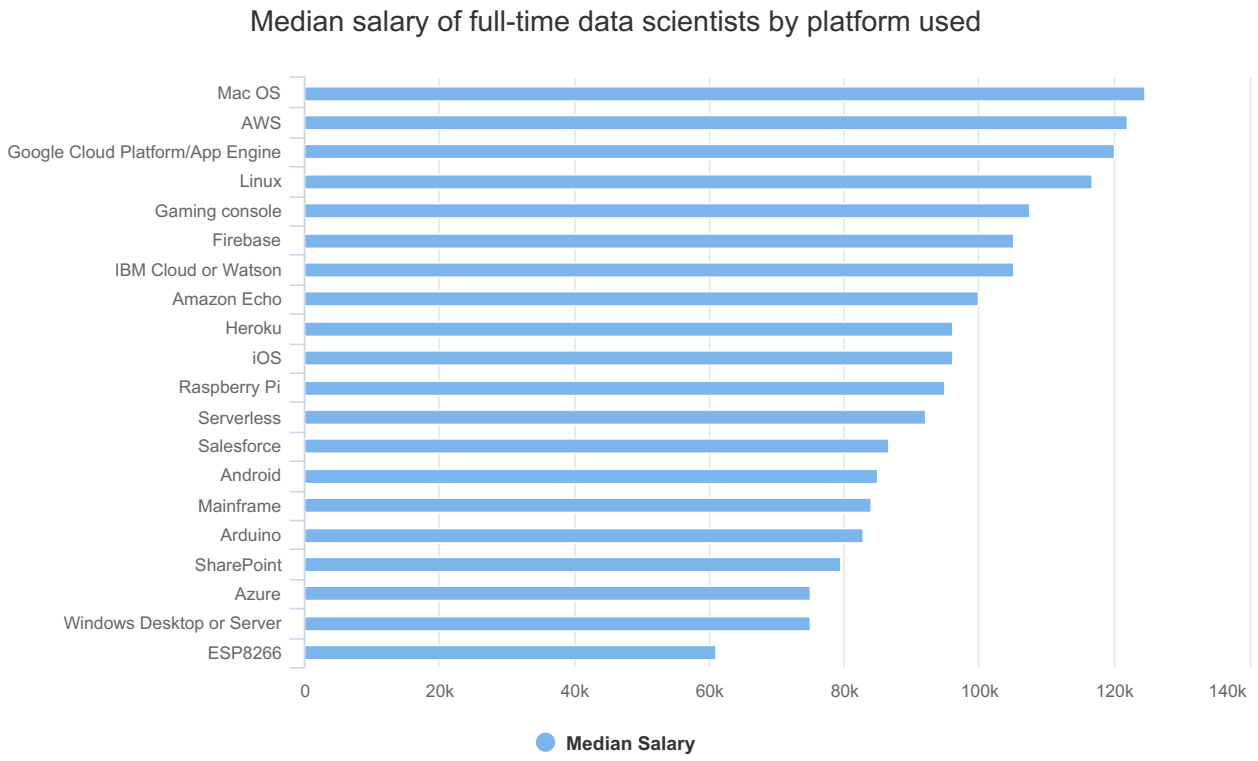
```
#This code graphs a histogram of median salary by most used database by decreasing order.
#Filters full-time workers, Data Scientists, US workers, and their missing data

median_sal_db = stackoverflow %>%
  filter(Employment %in% 'Employed full-time') %>%
  filter(!is.na(DatabaseWorkedWith))%>%
  filter(!is.na(DevType)) %>%
  filter(DevType %in% c('Data or business analyst','Data scientist or machine learning specialist'))
%>%     filter(Country %in% 'United States') %>%
  select(DatabaseWorkedWith,ConvertedSalary) %>% #The data already converted salaries to USD
  mutate(DatabaseWorkedWith = str_split(DatabaseWorkedWith, pattern = ";")) %>%
  unnest(DatabaseWorkedWith) %>%
  group_by(DatabaseWorkedWith) %>%
  summarise(Median_Salary = median(ConvertedSalary,na.rm = TRUE)) %>% #summarise each language by sa
lary
  arrange(desc(Median_Salary))%>% #descending order
  ungroup()%>%
  mutate(DatabaseWorkedWith = reorder(DatabaseWorkedWith, Median_Salary))
#slice top 20 data
median_sal_db = slice(median_sal_db, 0:20)
#plot
highchart()%>%
  hc_title(text = paste("Median salary of full-time data scientists by database used"))%>%
  hc_xAxis(categories = median_sal_db$DatabaseWorkedWith)%>%
  hc_add_series(data = median_sal_db$Median_Salary, name = "Median Salary", type = "bar")
```



This data consists of 259 rows. From the graph above, people who use most recently released database languages seem to earn more than the ones which have existed for long time or are more popular. Working with the most used databases in SQL will also give you the overall data scientists' median salary in the US.

4.3.4 Median salary by platforms used

```
#This code graphs a histogram of median salary by most used platform by decreasing order.
#Filters full-time workers, Data Scientists, US workers, and their missing data
median_sal_plat = stackoverflow %>%
  filter(Employment %in% 'Employed full-time') %>%
  filter(!is.na(PlatformWorkedWith))%>%
  filter(!is.na(DevType)) %>%
  filter(DevType %in% c('Data or business analyst','Data scientist or machine learning specialist'))
%>%     filter(Country %in% 'United States') %>%
  select(PlatformWorkedWith,ConvertedSalary) %>% #The data already converted salaries to USD
  mutate(PlatformWorkedWith = str_split(PlatformWorkedWith, pattern = ";")) %>%
  unnest(PlatformWorkedWith) %>%
  group_by(PlatformWorkedWith) %>%
  summarise(Median_Salary = median(ConvertedSalary,na.rm = TRUE)) %>% #summarise each language by sa
lary
  arrange(desc(Median_Salary))%>% #descending order
  ungroup()%>%
  mutate(PlatformWorkedWith = reorder(PlatformWorkedWith, Median_Salary))

#slice top 20
median_sal_plat = slice(median_sal_plat, 0:20)

highchart()%>%
  hc_title(text = paste("Median salary of full-time data scientists by platform used"))%>%
  hc_xAxis(categories = median_sal_plat$PlatformWorkedWith)%>%
  hc_add_series(data = median_sal_plat$Median_Salary, name = "Median Salary", type = "bar")
```



One interesting information found from this data, which consists of 200 rows, is that although Windows is one of the preferred platform by data scientists (ranked number three in Mostly Used Platform histogram), the medain salary of its users are comparatively lower than other platforms.

4.3.5 Section Conclusion

In this section, we analyzed how salary is affected by location, programming languages, databases and platforms. We found that among all the countries, USA has the highest median salary for full-time data scientists. As for technical skills, if a person uses mainstream programming languages such as Python and R or mainstream databases in SQL, he/she is likely to earn around $100k/year, which is the overall data scientists' median salary in the US. If a person has more new and uncommon technical skills, he/shes is likely to earn more than the median salary.

4.4 Review: Challenges Faced in Exploratory Data Analysis

Some challenges that we faced while carrying out the analysis and our solutions to them:

(1) Scraping data from mail

We had some difficulties using third party APIs to scrape emails from Columbia's LionMail as it requires an extra authentication. Luckily, we found that Columbia's LionMail can be used in Thunderbird API, which also has a function to download all the emails to a csv file.

(2) Finding relevant datasets

It took us some time to find relevant datasets to our studies which are also up to date and representative enough. After some effort, we found the two datasets from Kaggle and Stack Overflow. The two datasets are both survey in 2018 with a great number of respondents from all over the world, and the questions match with our objective of this study.

(3) Combining insights

Since we are using three datasets, it took us some time to combine the interesting features of each dataset to a comprehensive and organized analysis. Teamwork was the key - all three members in our group contributed in their part of analysis, and we spent a lot of time discussing and sharing our findings in each dataset.

5 Executive summary (Presentation-style)

A DATA SCIENTIST IS BORN

JONGHYUK LEE
YU HAN HUANG
DEEPAK RAVISHANKAR

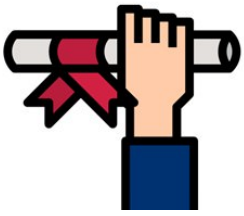
INSIGHTS OF THE CURRENT DATA SCIENCE JOB MARKET

First Why did we choose this topic?

For our data science classmates...

Who will be graduating next year and need jobs

GET PREPARED FOR THE JOB MARKET AND GET HIRED !

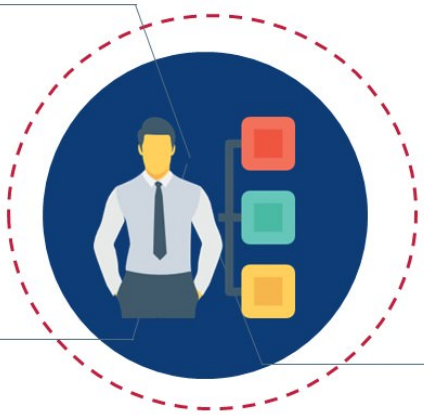


Second

What were we trying to know?

Age
Gender
Education

WHO



HOW

General Skills
Technical Skills

WHAT

Expected Salary

Demographic Of Current Data Science Job Market
Supply v.s Demand

Second

Data Source

Demand: Job Listing

Rachel's Mail:
Data Science Career Opportunities Email
Sample size: 82

Supply: USA Data

Stack Overflow 2018 Developer Survey
Sample size: 100,000

Supply: World Data

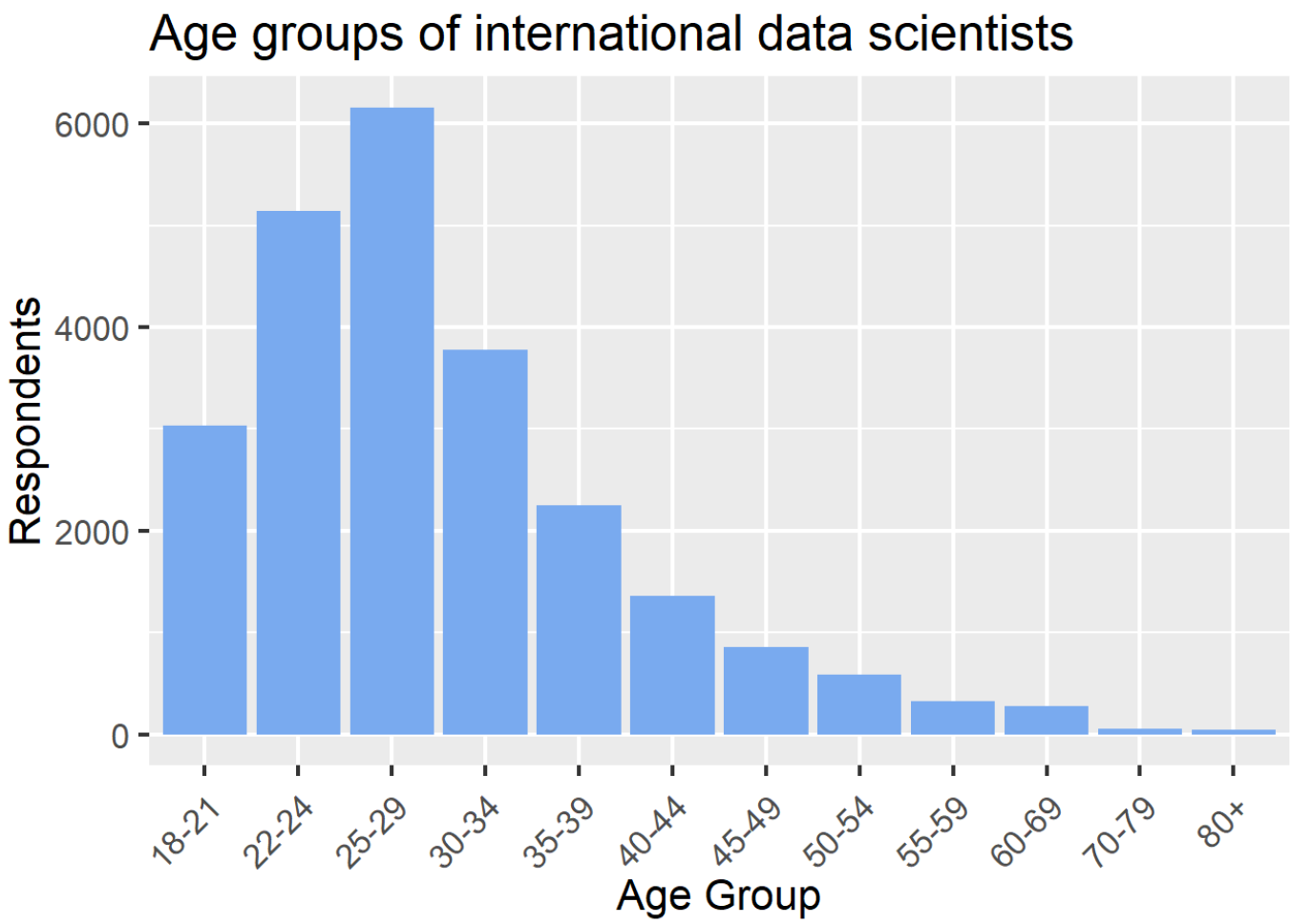
Kaggle ML and Data Science Survey, 2018
Sample size: 23,859

WHO'S IN THE JOB MARKET

Age

We are interested in knowing the age distribution of data scientists in the job market.

```
foo <- df_survey_mcq %>%
  group_by(age) %>%
  count()

p2 <- foo %>%
  mutate(percentage = str_c(as.character(round(n/sum(foo$n)*100,1)), "%")) %>%
  ggplot(aes(age, n)) +
  geom_col(fill="#79AAEF") +
  labs(x = "Age Group", y = "Respondents") + theme_grey(16) +
  theme(legend.position = "none", axis.text.x  = element_text(angle=45, hjust=1, vjust=0.9)) +
  ggtitle("Age groups of international data scientists")
p2
```

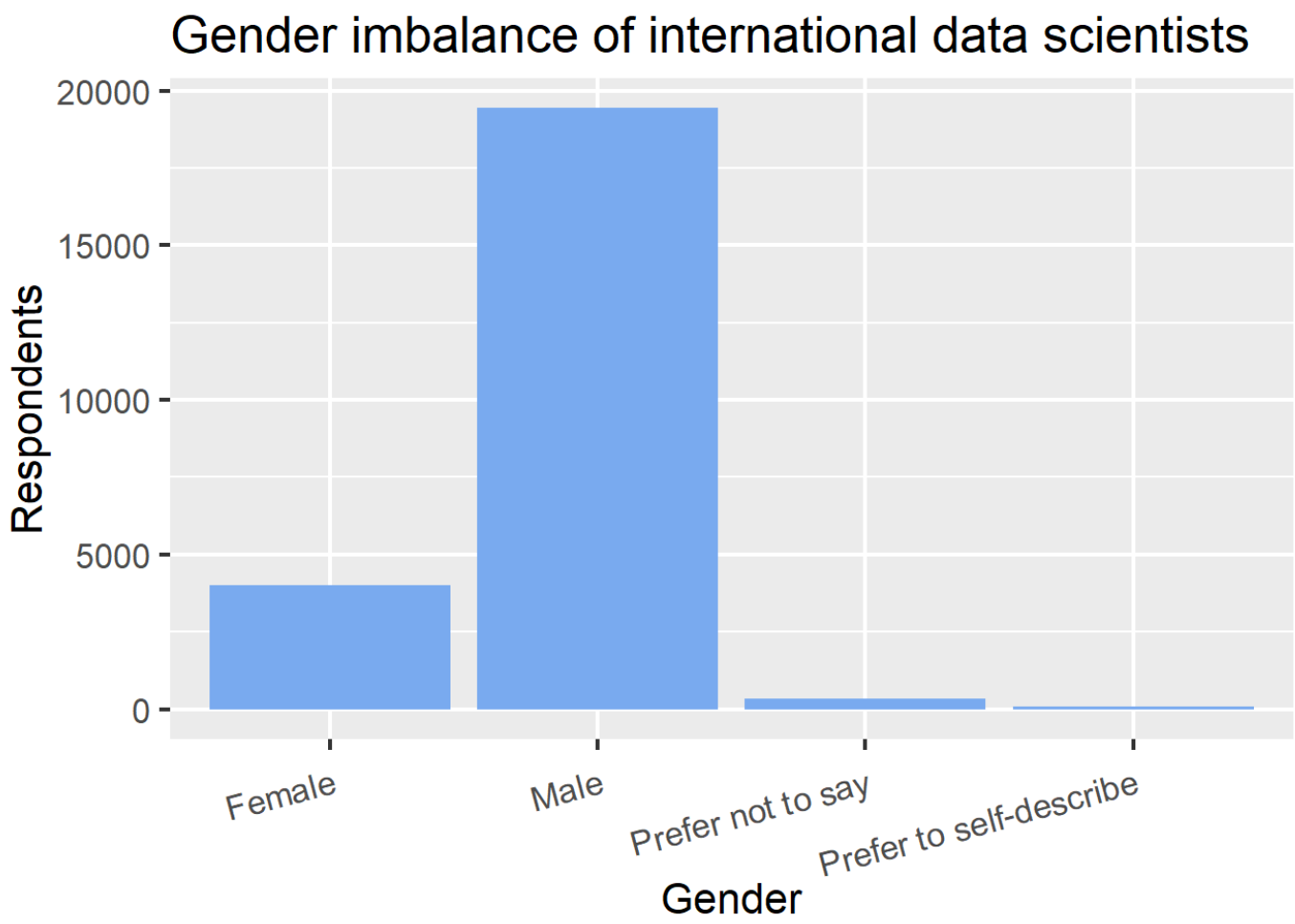


We find that most of the people are in the age group of 20-35 and thus showing that the field is mainly dominated by the younger generation. This makes sens as the field is relatively new and thus people just graduating would be more interested in the field.

Gender

We are interested in knowing the gender imbalance of data scientists in the job market.

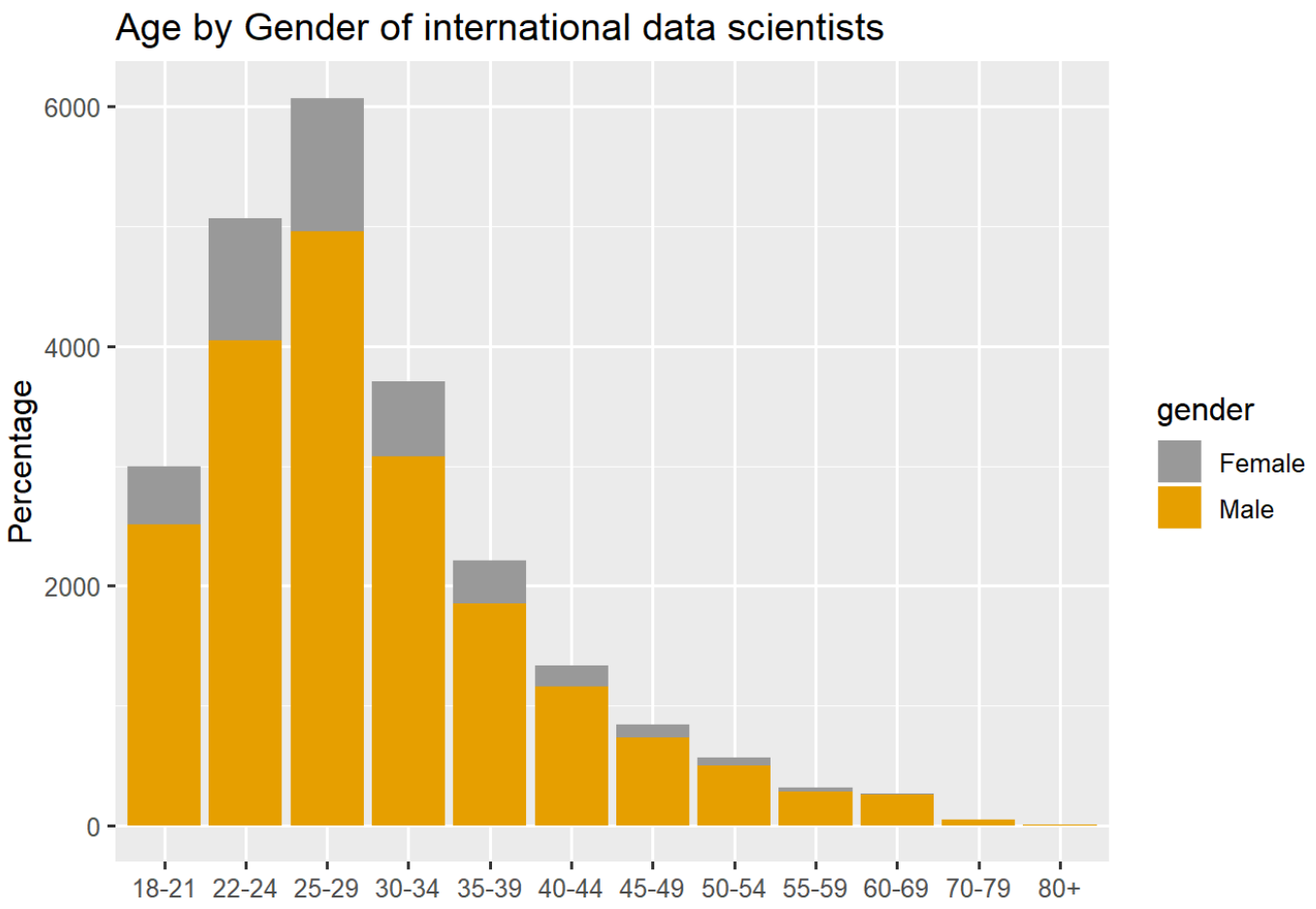
```
foo <- df_survey_mcq %>%
  group_by(gender) %>%
  count()

p1 <- foo %>%
  mutate(percentage = str_c(as.character(round(n/sum(foo$n)*100,1)), "%")) %>%
  ggplot(aes(gender, n)) +
  geom_col(fill="#79AAEF") +
  labs(x = "Gender", y = "Respondents") + theme_grey(16) +
  theme(legend.position = "none", axis.text.x  = element_text(angle=15, hjust=1, vjust=0.9)) +
  ggtitle("Gender imbalance of international data scientists")
p1
```



We can see that there is a huge gender imbalance among the male and the female and other genders, which is normally the case in any STEM field.

Age by gender

```
#plot and analyze the age and gender variable
p3 <- df_survey_mcq %>%
  filter(gender %in% c("Male", "Female")) %>%
  ggplot(aes(age, fill = gender)) +
  geom_bar() + theme_grey(12) +
  labs(x = "", y = "Percentage") +
  ggtitle("Age by Gender of international data scientists") + scale_fill_manual(values=cbPalette)


p3
```

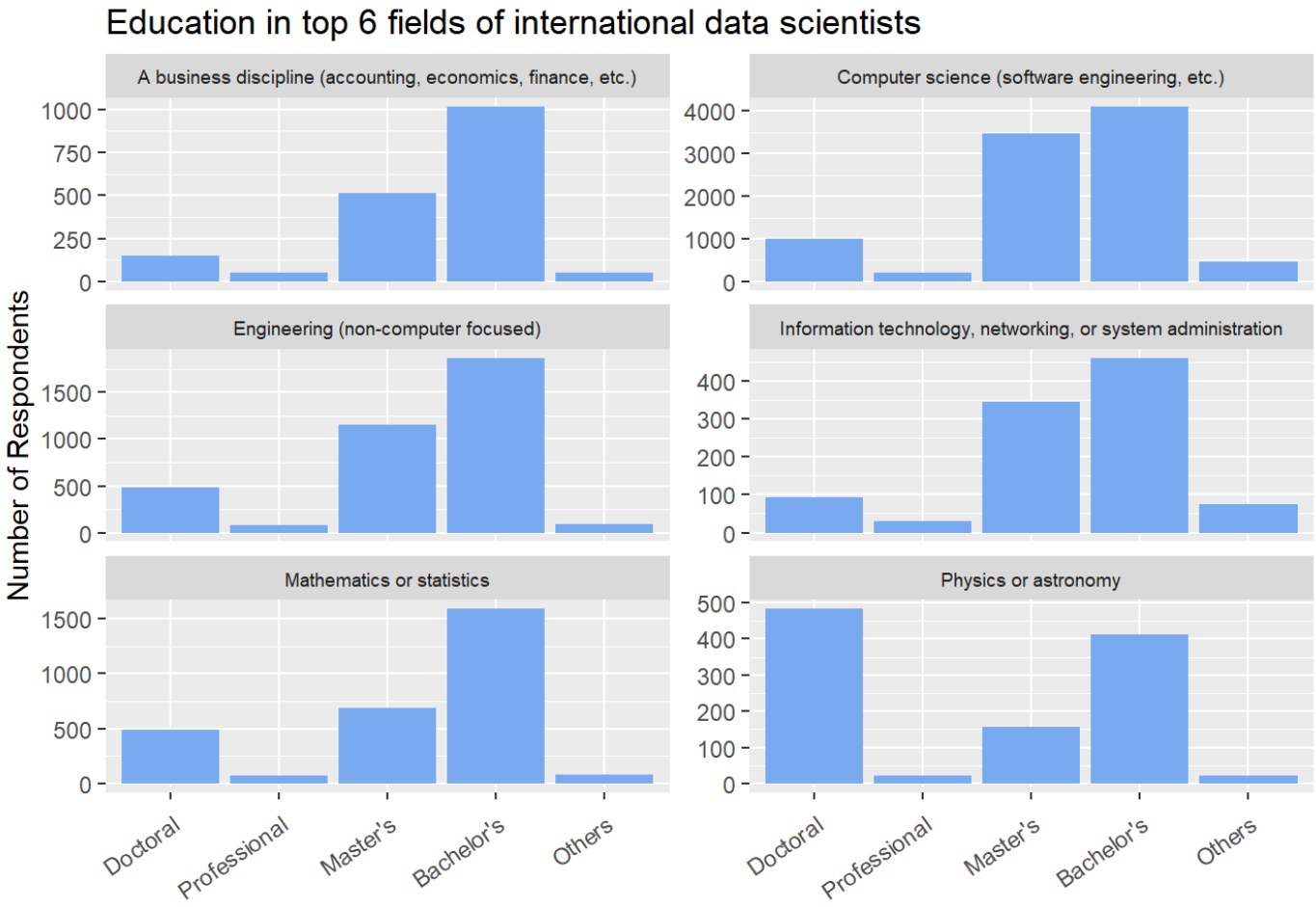


Comparing age by gender, we see that the younger generations are doing better compared to the older generations in terms of the sex ratio and thus there are more women opting for data science as a career path.

Education

Education in top 6 fields

```
#visulaize the education level of respondents
foo <- df_survey_mcq %>%
  filter(!is.na(major)) %>%
  group_by(major) %>%
  count() %>%
  ungroup() %>%
  top_n(6, n)

labs <- c("Doctoral", "Professional", "Master's", "Bachelor's", "Others")

df_survey_mcq %>%
  filter(!is.na(edu) & edu != "No answer") %>%
  semi_join(foo, by = "major") %>%
  count(edu, major) %>%
  ggplot(aes(edu, n)) +
  geom_col(fill="#79AAEF") +
  theme(legend.position = "none",
        axis.text.x  = element_text(angle=35, hjust=1, vjust=0.9),
        strip.text.x = element_text(size = 7)) +
  guides(fill = guide_legend(ncol = 2)) +
  labs(x = "", y = "Number of Respondents") +
  facet_wrap(~ major, ncol = 2, scales = "free_y") +
  ggtitle("Education in top 6 fields of international data scientists") + scale_x_discrete(labels=
labs)
```



We see that the most common degree is masters for all the undergrad majors. But a lot of the PhDs are also in this industry. Apart from we notice that the PhD are more common when joining the industry from physics major.

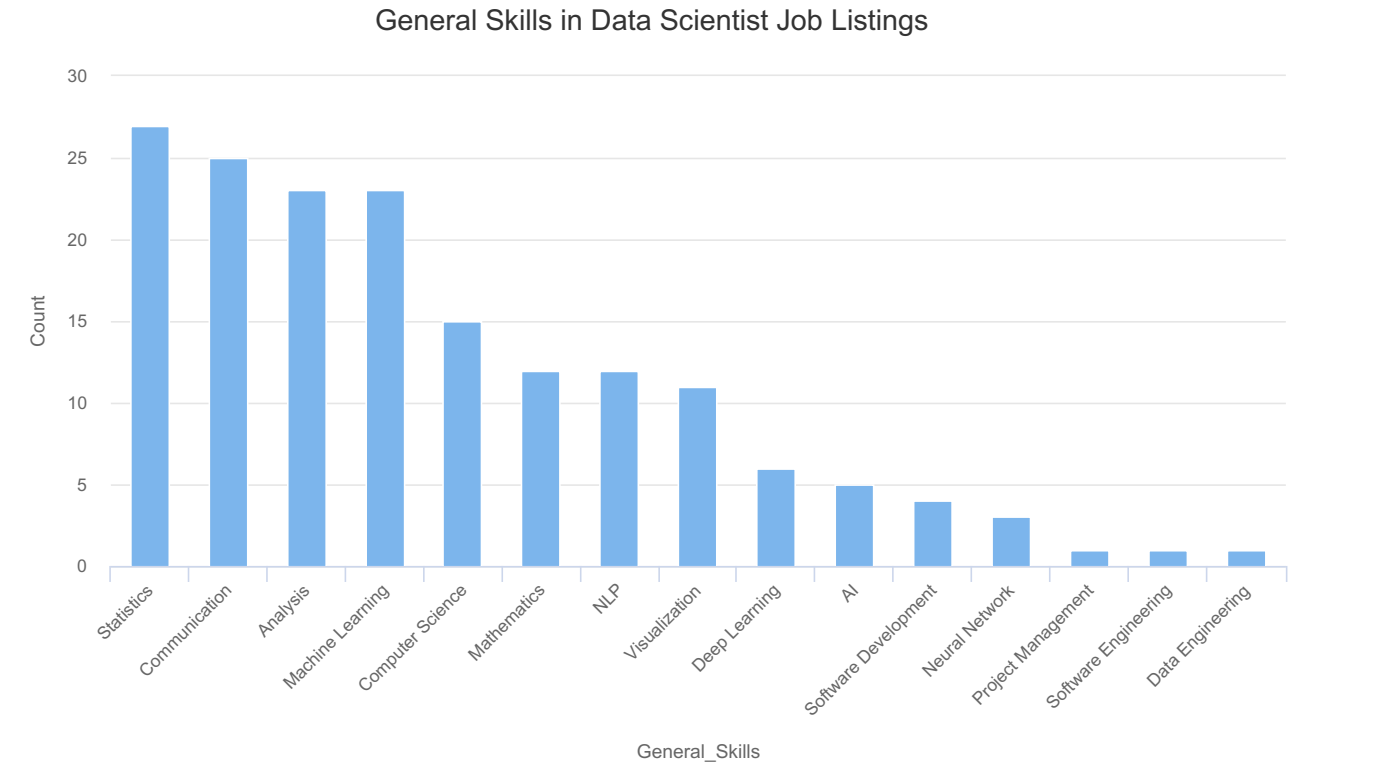
Third What did we find?

WHAT SKILLS ARE IN THE JOB MARKET

Demand – General Skills Wanted in Job Listings

We used the general skills listed in Jeff Hale's study and searched for their occurences in the job listing requirements in Columbia University Data Science career opportunities emails.

```r
#Keyword Analysis
keywords <- data.frame("General_Skills" = c("Analysis", "Machine Learning", "Statistics", "Computer S
cience", "Communication", "Mathematics", "Visualization", "AI", "Deep Learning", "NLP", "Software Dev
elopment", "Neural Network", "Project Management", "Software Engineering", "Data Engineering"),
"Count" = c(length(grep("analysis", tolower(rachel$Description))), length(grep("machine learning", to
lower(rachel$Description))), length(grep("statistics", tolower(rachel$Description))),
length(grep("computer science", tolower(rachel$Description))), length(grep("communication", tolower(
rachel$Description))),  length(grep("mathematics", tolower(rachel$Description))), length(grep("visual
ization", tolower(rachel$Description))), length(grep("AI",
rachel$Description))+length(grep("artificial intelligence", tolower(rachel$Description))), length(gre
p("deep learning", tolower(rachel$Description))), length(grep("NLP", rachel$Description))+length(grep
("natural language processing", tolower(rachel$Description))), length(grep("software development", t
olower(rachel$Description))), length(grep("neural network", tolower(rachel$Description))),
length(grep("project management", tolower(rachel$Description))), length(grep("software engineering",
tolower(rachel$Description))), length(grep("data engineering", tolower(rachel$Description)))))

#Plot
keywords %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = General_Skills, y = Count)) %
>% hc_xAxis(type = 'category') %>% hc_title(text="General Skills in Data Scientist Job Listings")
```



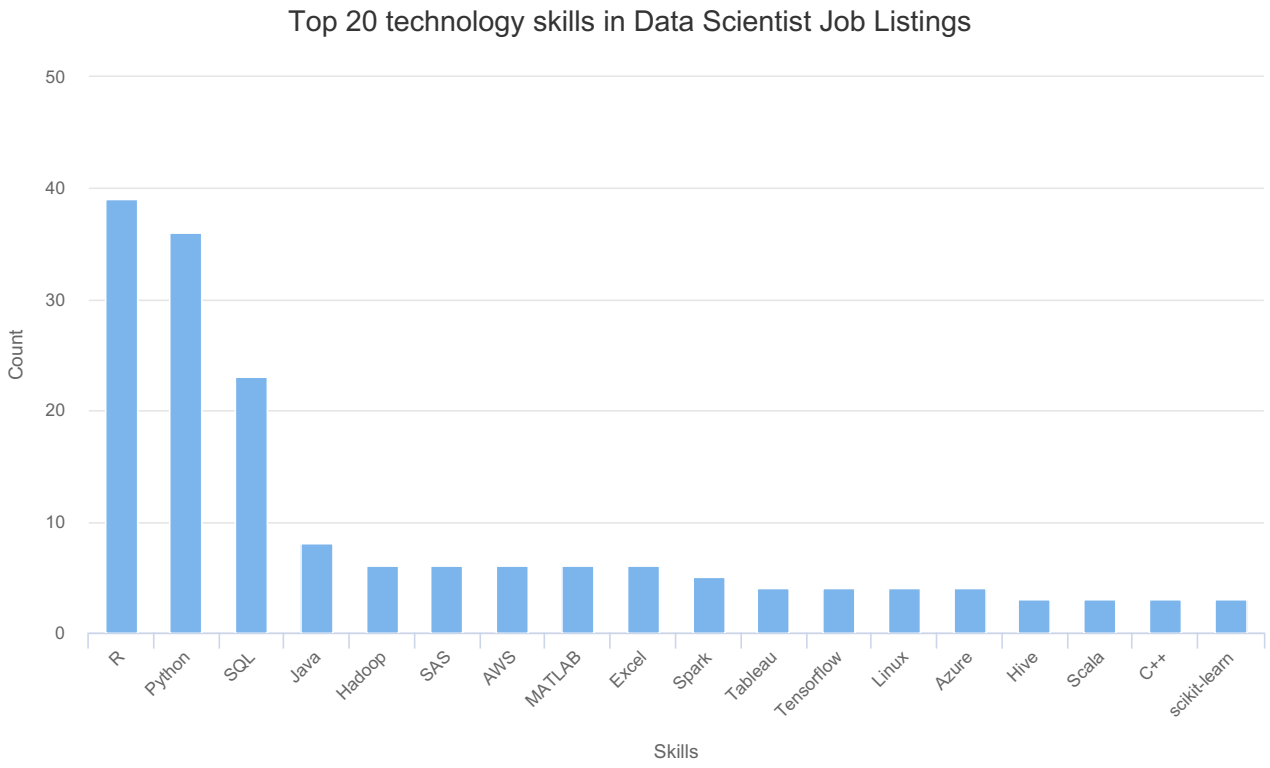
Our results show that statistical analysis, communication and machine learning are at the heart of data scientist jobs, and that matches with Jeff Hale's study on major job listing websites. The result matches with our expectations, since the primary function of data science is to use statistical analysis to draw useful insights from data. Machine learning and its subsets - AI and deep learning, also show up frequently since these are the major techniques in the field of data science to create systems to predict performance and are very in demand.

It is also noteworthy that communication tops the rank in both studies' job listing descriptions. It tells us that it is very important for data scientists to be able communicate insights and work with others.

Demand – Technical Skills Wanted in Job Listings

We used the technical skills listed in Jeff Hale's study and searched for their occurences in the job listing requirements in Columbia University Data Science career opportunities emails.

```
#keywords
skills <- data.frame("Skills" = c("Python", "R", "SQL", "Hadoop", "Spark", "Java", "SAS", "Tableau",
"Hive", "Scala", "AWS", "C++", "MATLAB", "Tensorflow", "Excel", "Linux", "Azure", "scikit-learn"), "C
ount" = c(length(grep("python", tolower(rachel$Description))), length(grep("R", rachel$Description)),
length(grep("SQL", rachel$Description)), length(grep("hadoop", tolower(rachel$Description))), length(
grep("spark", tolower(rachel$Description))), length(grep("java", tolower(rachel$Description))),
length(grep("SAS", rachel$Description)), length(grep("tableau", tolower(rachel$Description))), length
(grep("hive", tolower(rachel$Description))), length(grep("scala", tolower(rachel$Description))),
length(grep("AWS", rachel$Description)), length(grep("C\\+\\+", rachel$Description)), length(grep("ma
tlab", tolower(rachel$Description))), length(grep("tensorflow", tolower(rachel$Description))),
sum(str_count( tolower(rachel$Description), "\\bexcel\\b")), length(grep("linux",
tolower(rachel$Description))), sum(str_count( tolower(rachel$Description), "\\bazure\\b")),
length(grep("scikit-learn", tolower(rachel$Description))) ))

#plot
skills %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = Skills, y = Count)) %>% hc_xAxi
s(type = 'category') %>% hc_title(text="Top 20 technology skills in Data Scientist Job Listings")
```



Our results show that R, Python and SQL are the most demanded technical skills in data scientist jobs, and that matches with Jeff Hale's study on major job listing websites. Python and R are not very far off from each other and are dominant in frequency, which makes the two languages a must for virtually every data scientist position. SQL is also in high demand. SQL stands for Structured Query Language and is the primary way to interact with relational database.

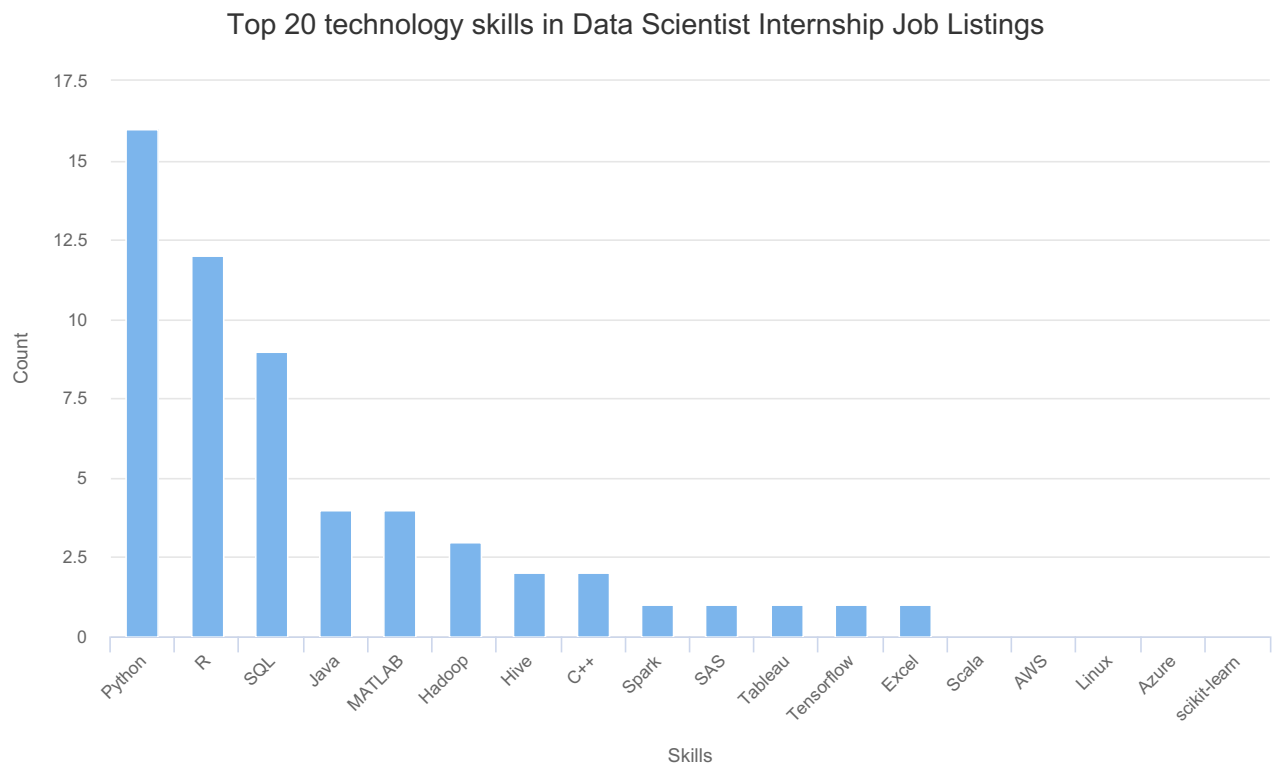
Demanded Technical Skills:Internship vs Full-Time

We are also interested to know if the technical skills wanted in job listings would vary by job types. In Columbia University Data Science career opportunities emails, both internship positions and full-time jobs positions are listed together. We want to know if the industry would be looking for different skills in the two types of applicants.

```
#Filter data set
rachel_intern <- rachel[rachel$Type == "Internship",]
rachel_full <- rachel[rachel$Type == "Full Time",]

#Internship
skills_intern <- data.frame("Skills" = c("Python", "R", "SQL", "Hadoop", "Spark", "Java", "SAS", "Tableau", "Hive", "Scala", "AWS", "C++", "MATLAB", "Tensorflow", "Excel", "Linux", "Azure", "scikit-learn"), "Count" = c(length(grep("python", tolower(rachel_intern$Description))), length(grep("R", rachel_intern$Description)), length(grep("SQL", rachel_intern$Description)), length(grep("hadoop", tolower(rachel_intern$Description))), length(grep("spark", tolower(rachel_intern$Description))), length(grep("java", tolower(rachel_intern$Description))), length(grep("SAS", rachel_intern$Description)), length(grep("tableau", tolower(rachel_intern$Description))), length(grep("hive", tolower(rachel_intern$Description))), length(grep("scala", tolower(rachel_intern$Description))), length(grep("AWS", rachel_intern$Description)), length(grep("C\\+\\+", rachel_intern$Description)), length(grep("matlab", tolower(rachel_intern$Description))), length(grep("tensorflow", tolower(rachel_intern$Description))), sum(str_count( tolower(rachel_intern$Description), "\\bexcel\\b")), length(grep("linux", tolower(rachel_intern$Description))), sum(str_count( tolower(rachel_intern$Description), "\\bazure\\b")), length(grep("scikit-learn", tolower(rachel_intern$Description))) ))

#Full-Time
skills_full <- data.frame("Skills" = c("Python", "R", "SQL", "Hadoop", "Spark", "Java", "SAS", "Tableau", "Hive", "Scala", "AWS", "C++", "MATLAB", "Tensorflow", "Excel", "Linux", "Azure", "scikit-learn"), "Count" = c(length(grep("python", tolower(rachel_full$Description))), length(grep("R", rachel_full$Description)), length(grep("SQL", rachel_full$Description)), length(grep("hadoop", tolower(rachel_full$Description))), length(grep("spark", tolower(rachel_full$Description))), length(grep("java", tolower(rachel_full$Description))), length(grep("SAS", rachel_full$Description)), length(grep("tableau", tolower(rachel_full$Description))), length(grep("hive", tolower(rachel_full$Description))), length(grep("scala", tolower(rachel_full$Description))), length(grep("AWS", rachel_full$Description)), length(grep("C\\+\\+", rachel_full$Description)), length(grep("matlab", tolower(rachel_full$Description))), length(grep("tensorflow", tolower(rachel_full$Description))), sum(str_count(
tolower(rachel_full$Description), "\\bexcel\\b")), length(grep("linux",
tolower(rachel_full$Description))), sum(str_count( tolower(rachel_full$Description), "\\bazure\\b")), length(grep("scikit-learn", tolower(rachel_full$Description))) ))
```

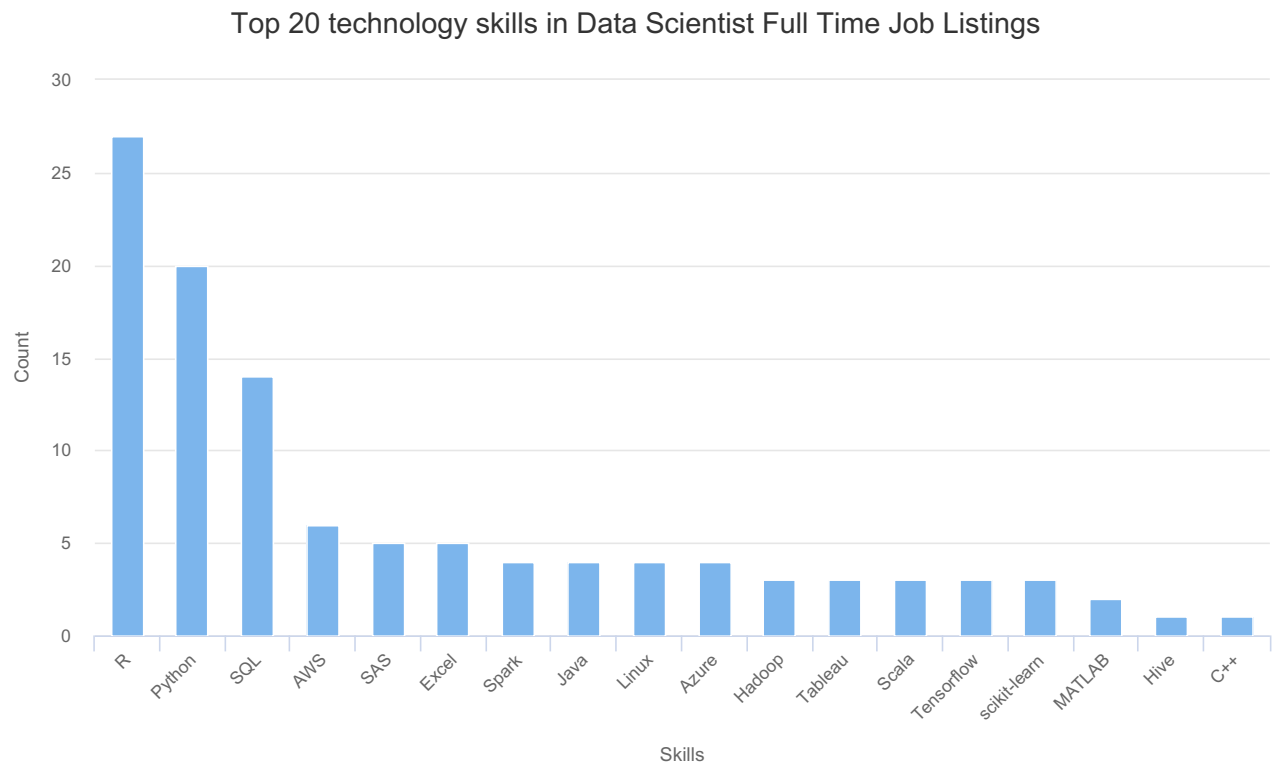
Top 20 technology skills in Data Scientist Internship Job Listings

```
skills_intern %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = Skills, y = Count)) %>% hc_xAxis(type = 'category') %>% hc_title(text="Top 20 technology skills in Data Scientist Internship Job Listings")
```



Top 20 technology skills in Data Scientist Full Time Job Listings

```
skills_full %>% arrange(desc(Count)) %>% hchart(type = "column", hcaes(x = Skills, y = Count)) %>% hc_xAxis(type = 'category') %>% hc_title(text="Top 20 technology skills in Data Scientist Full Time Job Listings")
```



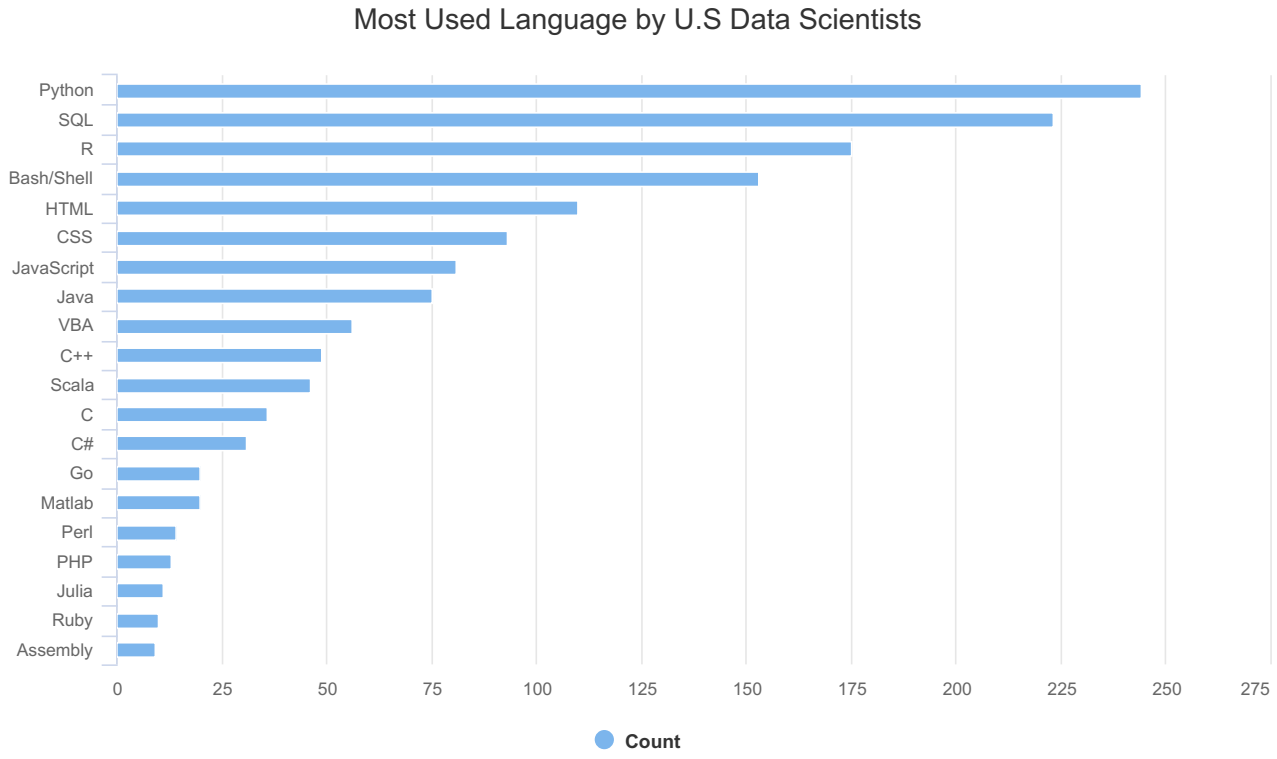
We can see that R, Python and SQL are still the top three demanded technical skill, but the rank is a little different by job types. The rank in full-time job listings matches with our previos findings, whereas R has a slight edge over Python in internship jobs. The technical skills required in internship positions are also less than full-time job positions.

Supply – Technical Skills People Have (USA data)

We used Stack Overflow 2018 Developer Survey to analyze the technical skills U.S based data scientists have. The analysis filters full-time

workers, Data Scientists, US workers, and their missing data.

```
#This code graphs a histogram by mostly used languages in decreasing order.
#Filters full-time workers, Data Scientists, US workers, and their missing data

languae_hist = stackoverflow %>%
  filter(Employment %in% 'Employed full-time') %>%  #filter full-time employees
  filter(!is.na(LanguageWorkedWith))%>% #filter missing data
  filter(!is.na(DevType)) %>%
  filter(DevType %in% c('Data or business analyst','Data scientist or machine learning specialist'))
%>% #filter data scientists
  filter(Country %in% 'United States') %>% #filter US workers
  select(LanguageWorkedWith)%>%
  mutate(LanguageWorkedWith = str_split(LanguageWorkedWith, pattern = ";"))%>%
  unnest(LanguageWorkedWith)%>%
  group_by(LanguageWorkedWith)%>%
  summarise(Count = n())%>% #count by languages
  arrange(desc(Count))%>% #reorder in descending order
  ungroup()%>%
  mutate(LanguageWorkedWith = reorder(LanguageWorkedWith, Count))


#slice only top 20 data
languae_hist = slice(languae_hist, 0:20)

highchart()%>% #
  hc_title(text = paste("Most Used Language by U.S Data Scientists"))%>% #title
  hc_xAxis(categories = languae_hist$LanguageWorkedWith)%>% #xaxis
  hc_add_series(data = languae_hist$Count, name = "Count", type = "bar") #plot
```

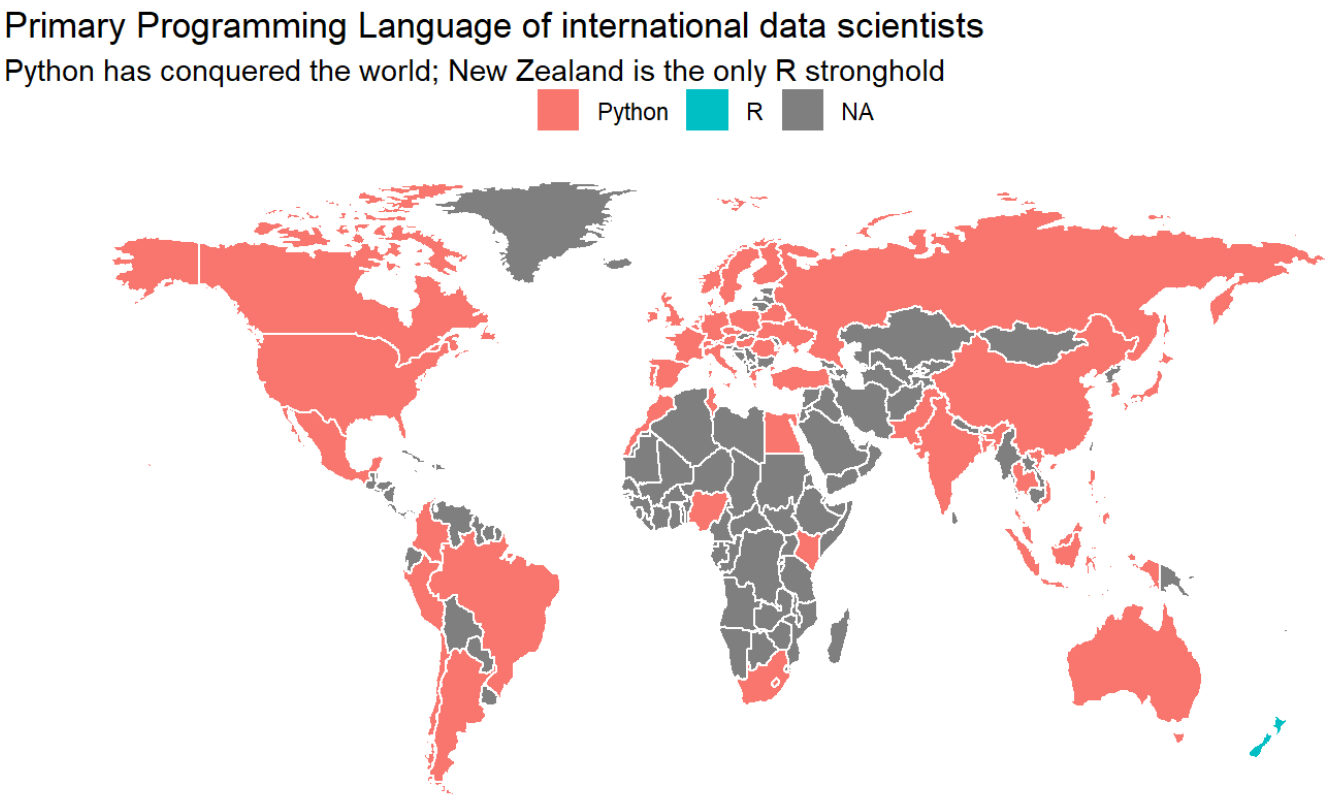


The purpose of this histogram is to explore which language data scientists in U.S. use. This data filters full-time data scientists, working in the US, which conists of 317 rows of data. The X-axis is count, and the Y-axis denotes each languages sorted in decreasing order of the counts. Our result shows that most data scientists in U.S use Python, SQL, and R.

Supply – Technical Skills People Have (World Data)

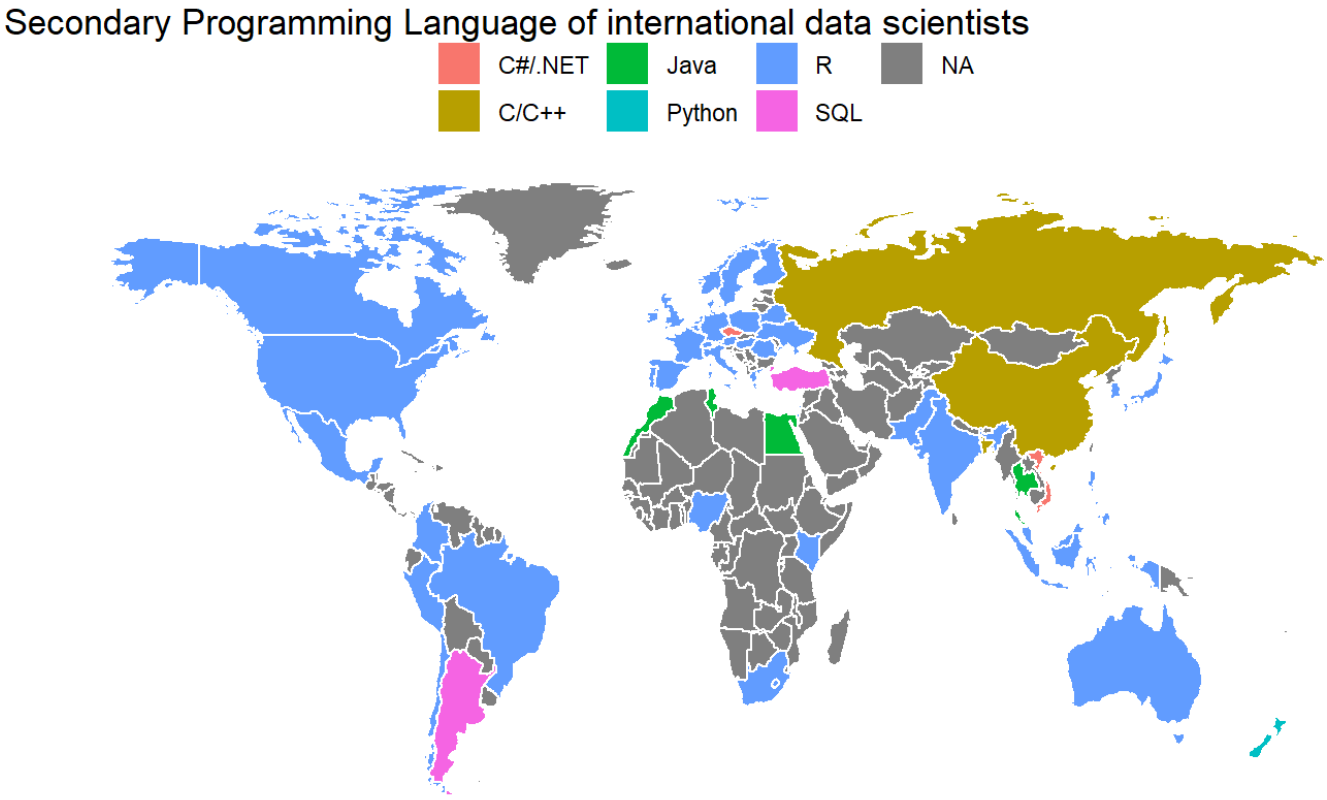
First popular programming language

```r
#visualize the first popular language
foo <- df_survey_mcq %>%
  filter(!(country %in% c("Other", "I do not wish to disclose my location"))) %>%
   mutate(country = as.character(case_when(
     country == "United Kingdom of Great Britain and Northern Ireland" ~ "UK",
     country == "United States of America" ~ "USA",
     country == "Viet Nam" ~ "Vietnam",
     TRUE ~ as.character(country)
  ))) %>%
  group_by(lang, country) %>%
  filter(!is.na(lang)) %>%
  count() %>%
  arrange(desc(n)) %>%
  group_by(country) %>%
  slice(c(1))

world %>%
  filter(region != "Antarctica") %>%
  left_join(foo, by = c("region" = "country")) %>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, fill = lang, group = group), color = "white") +
  coord_fixed(1.3) +
  labs(fill = "") +
  theme(legend.position = "top") +
  theme_void() +
  theme(legend.position = "top") +
  ggtitle("Primary Programming Language of international data scientists",
          subtitle = "Python has conquered the world; New Zealand is the only R stronghold")
```



Second popular programming language

```r
#visualize 2nd popular language
foo <- df_survey_mcq %>%
  filter(!(country %in% c("Other", "I do not wish to disclose my location"))) %>%
   mutate(country = as.character(case_when(
     country == "United Kingdom of Great Britain and Northern Ireland" ~ "UK",
     country == "United States of America" ~ "USA",
     country == "Viet Nam" ~ "Vietnam",
     TRUE ~ as.character(country)
  ))) %>%
  group_by(lang, country) %>%
  filter(!is.na(lang)) %>%
  count() %>%
  arrange(desc(n)) %>%
  group_by(country) %>%
  slice(c(2))

world %>%
  filter(region != "Antarctica") %>%
  left_join(foo, by = c("region" = "country")) %>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, fill = lang, group = group), color = "white") +
  coord_fixed(1.3) +
  labs(fill = "") +
  theme(legend.position = "top") +
  theme_void() +
  theme(legend.position = "top") +
  ggtitle("Secondary Programming Language of international data scientists")
```



Brief Conclusion

1. The job market is in demand of people who are good at **Statistical Analysis, Communication and Machine Learning**
2. The job market is in demand of people who are able to code in **Python, R and SQL, though the order may differ by job type**
3. Supply and demand of the market matches each other: **People are also comfortable working with Python and R**

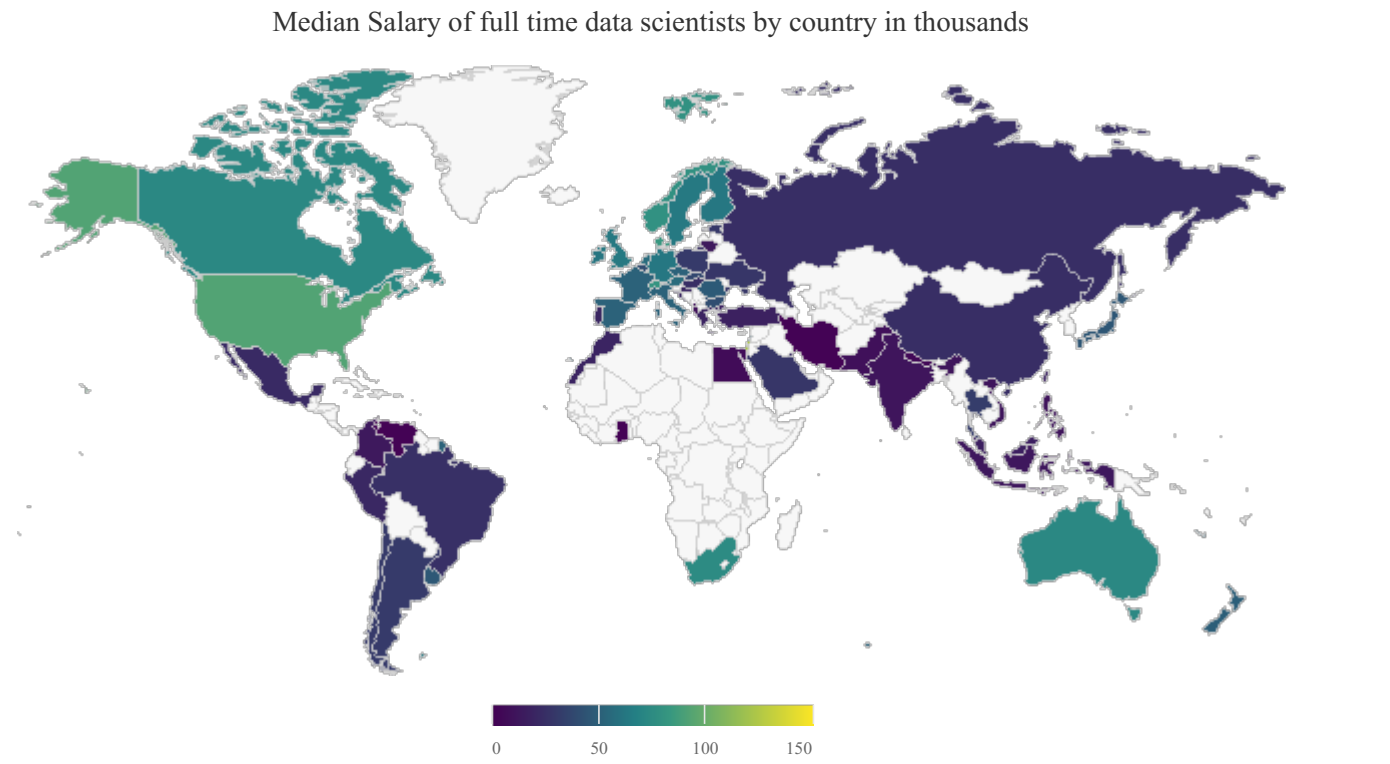
Third What did we find?

Now we have the skills...

WHAT IS THE EXPECTED SALARY IN MARKET?

Median Salary of Data Scientists by Country in thousands

```
by_country_salary <- stackoverflow %>% select(Country, ConvertedSalary, DevType) %>%
filter(!is.na(DevType)) %>%
  filter(DevType %in% c('Data or business analyst','Data scientist or machine learning specialist'))
%>% #filter data scientists
  mutate(ConvertedSalary=as.numeric(ConvertedSalary))  %>% filter(!is.na(Country)) %>% filter(!is.na(
ConvertedSalary)) %>%
  group_by(Country) %>% summarize(MedSalary = median(ConvertedSalary, na.rm=TRUE))

data(worldgeojson, package = "highcharter") #using highcharter
code <- countrycode(by_country_salary$Country, 'country.name', 'iso3c') #get country code

by_country_salary$iso3 <- code
by_country_salary$MedSalary <- round(by_country_salary$MedSalary/1000) #round

#plot

highchart() %>%
  hc_add_series_map(worldgeojson, by_country_salary, value = "MedSalary", joinBy = "iso3") %>%
  hc_colorAxis(stops = color_stops()) %>%
  hc_legend(enabled = TRUE) %>%
  hc_title(text = "Median Salary of full time data scientists by country in thousands")  %>%
  hc_tooltip(useHTML = TRUE, headerFormat = "",
             pointFormat = "Country: {point.Country} / Median Salary: ${point.MedSalary}K") %>% hc_a
dd_theme(hc_theme_google())
```

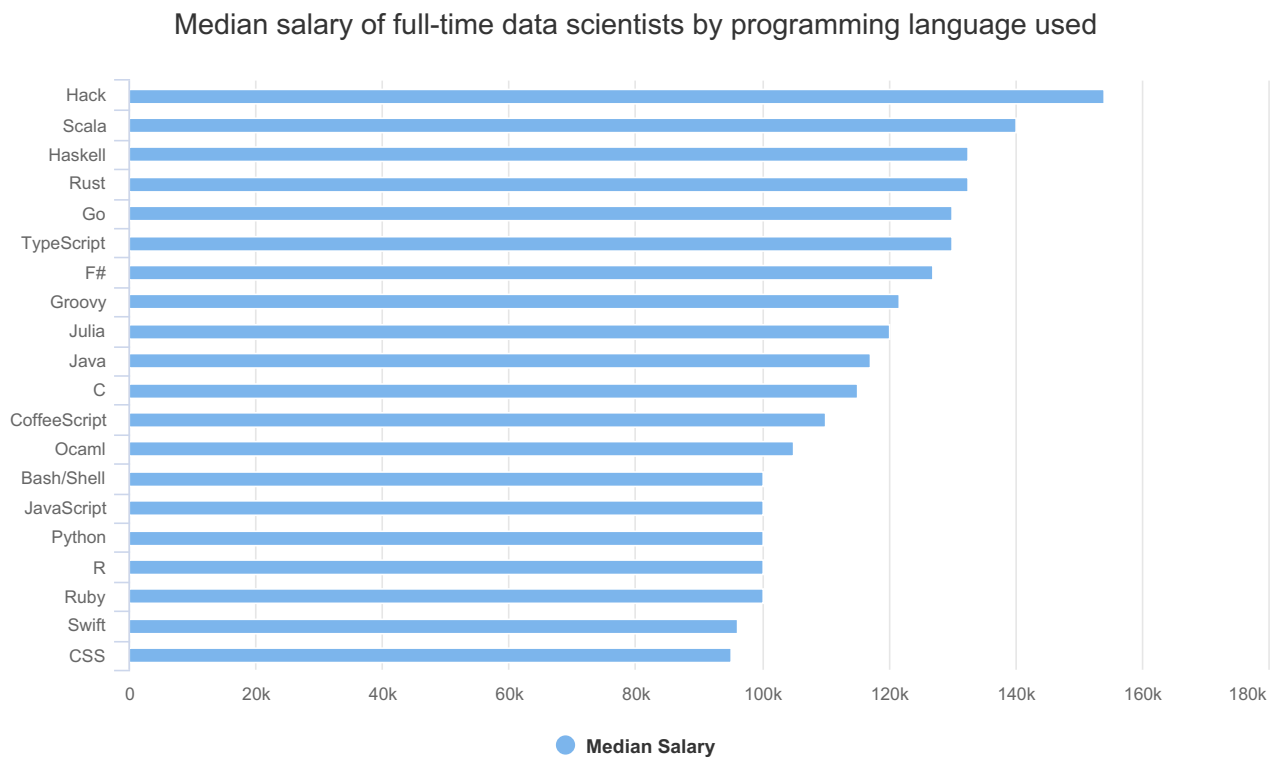


This world map shows median salary for full time data scientists by country. Brighter colored countries have higher median salary, and the darkered colored countries vice versa. The countries without a color are countries with no data.

We can see from the chart that USA has the highest median salary for data scientists, which is around $92k per year.

Median Salary of by Programming Language You Know

```r
#This code graphs a histogram of median salary by most used languages by decreasing order.
#Filters full-time workers, Data Scientists, US workers, and their missing data

median_sal_lang = stackoverflow %>%
  filter(Employment %in% 'Employed full-time') %>%
  filter(!is.na(LanguageWorkedWith))%>%
  filter(!is.na(DevType)) %>%
  filter(DevType %in% c('Data or business analyst','Data scientist or machine learning specialist'))
%>%     filter(Country %in% 'United States') %>%
  select(LanguageWorkedWith,ConvertedSalary) %>% #The data already converted salaries to USD
  mutate(LanguageWorkedWith = str_split(LanguageWorkedWith, pattern = ";")) %>%
  unnest(LanguageWorkedWith) %>%
  group_by(LanguageWorkedWith) %>%
  summarise(Median_Salary = median(ConvertedSalary,na.rm = TRUE)) %>% #summarise each language by sa
lary
  arrange(desc(Median_Salary))%>% #descending order
  ungroup()%>%
  mutate(LanguageWorkedWith = reorder(LanguageWorkedWith, Median_Salary))

#slice top 20 data
median_sal_lang = slice(median_sal_lang, 0:20)

#plot
highchart()%>%
  hc_title(text = paste("Median salary of full-time data scientists by programming language used"))%
>%
  hc_xAxis(categories = median_sal_lang$LanguageWorkedWith)%>%
  hc_add_series(data = median_sal_lang$Median_Salary, name = "Median Salary", type = "bar")
```



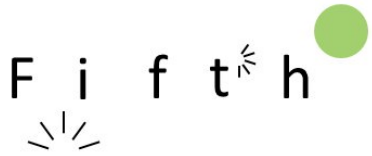
This shows a median salary by language data scientists use, and the data consists of 317 rows. It is interesting to see that the newly introduced languages such as Hack or Scala have higher median salary compared to the well known languages such as Python or R. However, median salary for Python and R are both 100k which corresponds to the overall data scientists' median salary in the US.

Brief Conclusion

1. USA has the highest median salary for data scientists
2. If you use Python and R, you will get US's median salary ~100k
   If you use more uncommon languages, you get paid more.
3. There isn't a big difference between using Python and R

Fourth — Why challenges did we face?

- **Scraping Data From Mail** — Solution: Use Thunderbird API
- **Find Relevant Datasets** — Solution: Use survey from Stack Overflow and Kaggle
- **Combine Insights** — Solution: Teamwork discussion



LARGER DATA

Scrape data from larger job listing websites like LinkedIn

BETTER DATA

Create our survey on Columbia's Data Science students

FURTHER USE

Collaborate with department to prepare students

6 Interactive component

You can access our interactive component here at:
https://deepakshankar94.github.io/edav-interactive-component/

- **Choice of data and plot types to present:**

We use the Stack Overflow 2018 Developer Survey as our interactive component's data. We created a map graph that displays all the countries, and four histograms that display the following features of that country's data scientists: Gender Ratio, Top 3 Databases Used, Top 3 Programming Languages Used and Top 3 Programming Platforms Used.

- **Clear relevance to question(s), project in general:**

We decided to visualize these variables to answer our question in section 4.1 Who's in the data science job market and 4.2 What skills are in the data science job market (both demand and supply). Since our project aims to help people who want to enter the data science job market understand the demographic of current job market, we felt it would be helpful if people can understand each countries' general working environment directly through our interactive gadget.

- **Design of interactive component(s):**

In our interactive gadget, there is a map and four charts. Users can click on the map to select locations, and the charts would display that country's data scientists' gender ratio, what databases do they work with, what programming languages do they use and what platforms do they work on.

- **Technical execution (include a description of what you would work on in the future, what you've attempted, etc. so we know it's on your radar):**

Currently, there are still a lot of countries with missing data in their data scientists' job market. Therefore, we would love to work on more comprehensive datasets in the future to generate better insights for our users.

7 Conclusion

In this study, we wanted to understand the demographic of the current data science job market: who is in the market, what skills are needed in the market, and what prospects can we expect from the market. For each variable, we evaluated it from both demand and supply.

We used three datasets, which respectively are Stack Overflow 2018 Developer Survey, Kaggle ML and Data Science Survey, 2018 and Rachel's Mail - Columbia University Data Science Career Opportunities. The first two datasets that represent the supply side of the job market were directly downloaded from the survey websites, and the last data that represents the demand side of market was scrapped from Columbia's LionMail system.

7.1 Key findings

Our key findings are:

Three major key takeaways:

1. A master degree is important to become a data scientists.

2. Exceling in Python, R and SQL is the basic entry to the job market and will land you with the median salary.

3. If you have time, you should try to learn some latest programming languages as it could lead to a good salary bump.

7.2 Summary of findings

Composition of the current data scientists job market

- Age: We found that most of the people in the current job market are in the age group of 20-35, indicating that the field of data science is mainly dominated by the younger generation.
- Gender: We found that there is a huge gender imbalance among the male and the female and other genders.
- Countries: We found that the countries which dominate the field of data science are USA, India, China, Russia, Brazil and UK.
- Education: We found that most people in the current data science job market have a masters' degree, but in some countries like India , Brazil and Australia, the barrier of entry is lower as more people with just a bachelors degree are also able to join the industry. While most people owns a bachelor degree in Computer Science, people graduating from other majors such as Mathematics, Business and Physics are also joining in the market.
- Primary data types: While most of the countries in the world work with numerical data, China seems to be mainly working on Image data.

Skills demanded and supplied in the data scientists job market

We discovered that the job market in data science is in demand of people who are good at Statistical Analysis, Communication and Machine Learning. The demand is both seen in Columbia University data science career opportunities emails and major job listing websites in the United States.

We found that most people on the job market are still relatively new to machine learning, but already have some years of experiences in statistical analysis. There is also a great percentage of people who don't have any experience in machine learning yet but are eager to learn. Therefore, we see a positive relationship between the supply and demand of the job market, and we are confident that the supply can soon meet the demand.

As for technical skills, the job market in data science is in demand of people who are able to code in Python, R and SQL. The three programming langauges' order may differ due to job types, but are still the top three dominant programming languages required in job listings.

We were delighted to find that the supply and demand of the market match with each other, that people in the job market are also most comfortable working with Python and R. This is both the case for data scientists in U.S and in the world.

Prospects of the data scientists job market- Salary analysis

We found that among all the countries, USA has the highest median salary for full-time data scientists. As for technical skills, if a person uses mainstream programming languages such as Python and R or mainstream databases in SQL, he/she is likely to earn around $100k/year, which is the overall data scientists' median salary in the US. If a person has more new and uncommon technical skills, he/shes is likely to earn more than the median salary.

7.3 Limitations and future directions

Limitations and future improvements on this project:

- **Relative few data samples in demand side compared to supply side:**

Since we used one semester's of Columbia University Data Science Career Opportunities emails as our demand data, the data is relatively few compared to Kaggle and Stack Overflow's worldwide survey. In the future, we may try to scrape data from larger job listing websites like LinkedIn.

- **Better contribution:**

To make better contribution to the data science community in Columbia University, it might be more helpful to create our survey on Columbia's Data Science students to see what skills are owned by the students and what skills are missing. We can also use this study to collaborate with the department to prepare students better for the job market in the program.