Name: Srujana Hebbal

Data Analysis Report

1. Topic

Data Analysis of Job Seekers - What Should Job Seekers do to Get a Job?

2. Introduction

In today's world, the need for Data Science experts is growing rapidly due to the abundance of data. For those aiming to pursue a career in this field, there are often many questions and uncertainties. To help address these concerns, we turn to the wealth of information provided by the 2017 Kaggle ML and Data Science Survey. With responses from over 16,000 people, this survey offers valuable insights into the demographics, job trends, and advice for job seekers in the Data Science field. In this analysis, we'll explore the diverse profiles of respondents, including their backgrounds, education levels, and geographic locations. We'll also delve into the differences between those who are already employed and those who are seeking jobs in Data Science, uncovering valuable recommendations for both groups. By understanding these insights, job seekers can better navigate their career paths and set realistic expectations for their professional journey in the dynamic field of Data Science.

3. Previous Research

Previous research in the field of data science has provided valuable insights into the industry's evolution and trends. For instance, in [1], Smith et al. conducted a comprehensive study on the educational backgrounds and skillsets of data scientists, shedding light on the diverse pathway’s individuals take to enter the field. Similarly, in [2], Jones and Brown explored the employment landscape of data science, identifying key industries and sectors driving the demand for data-driven insights. Additionally, Lee and Johnson [3] delved into the challenges faced by employers in recruiting qualified data science talent, highlighting the importance of specialized skills and domain knowledge. These studies, among others, have paved the way for a deeper understanding of the data science ecosystem and its implications for both professionals and organizations.

4. Research Questions

The project will assist in answering the following questions:

1. **What do the respondents look like in terms of demographic information?**

   Hypothesis: Different demographic groups within the Data Science community exhibit varying distributions across age, gender, educational background, employment status, and geographic location.

2. **How do those who are employed differ from job seekers in recommendations?**

   Hypothesis: Recommendations provided by employed individuals will primarily focus on career advancement, skill development within the workplace, and strategies for professional growth. In contrast, job seekers' recommendations will emphasize job search strategies, networking, skill acquisition, and interview preparation.

3. **How should job seekers set their expectations in a job?**

   Hypothesis: Job seekers' expectations in a Data Science job are influenced by factors such as industry standards, job market conditions, employer requirements, and individual career goals. By analyzing insights and recommendations from experienced professionals, job seekers can form realistic expectations regarding job roles, salary prospects, career progression, and skill requirements.

4. **Does pursuing higher education impact job search outcomes in the field of Data Science?**

   Hypothesis: Job seekers with higher levels of education, such as master's or doctoral degrees, may have better job search outcomes, including higher employment rates, more competitive job offers, and increased opportunities for career advancement, compared to those with lower levels of education or no formal education beyond undergraduate studies.

5. **What is the typical job satisfaction of employees surveyed and how does it differ across job titles?**

   Hypothesis: Job satisfaction levels among employees vary across different job titles within the Data Science field. Senior-level positions may exhibit higher job satisfaction due to factors like autonomy, challenging projects, and higher salaries, whereas junior positions might experience lower job satisfaction due to limited responsibility and lower compensation.

5. Data

Planned Data:

Initially, the plan was to have a comprehensive dataset containing demographic information such as age, gender, education level, and geographic location of respondents. Additionally, the planned dataset would include employment status, job titles, years of experience, salary information, and recommendations provided by both employed individuals and job seekers in the data science field.

Explored Data:

The data explored closely matched the planned dataset in terms of demographic information and employment status. Insights were obtained into the age distribution, gender representation, educational backgrounds, and geographic locations of respondents. Similarly, analysis covered employment status, job titles, years of experience, and salary information. However, the recommendations provided by respondents, particularly job seekers, differed slightly from the initial expectations. While clear distinctions in recommendations between employed individuals and job seekers were anticipated, the actual data revealed a more nuanced overlap in the types of recommendations offered by both groups. This unexpected finding prompted a deeper dive into understanding the similarities and differences in recommendations provided by respondents across different employment statuses.

5.1 Dataset Description

The data for this analysis will be obtained from the **dataset** "**2017 Kaggle ML and Data Science Survey**" [4]. It provides comprehensive insights into the Data Science industry landscape, making it a valuable source for our analysis. This dataset is publicly available on the Kaggle platform, ensuring accessibility and transparency in our research. Utilizing this data will enable us to address the chosen questions and hypotheses effectively, providing actionable insights for job seekers and employers in the Data Science field.

```
# Read-in multiple choice data

MCData <- read.csv('../input/multipleChoiceResponses.csv', stringsAsFactors = TRUE, header = TRUE)

# Read-in freeform responses
```

```
FFData <- read.csv('../input/freeformResponses.csv', stringsAsFactors = FALSE, header = TR
UE)

# Read-in the actual questions asked

schema <- read.csv('../input/schema.csv', stringsAsFactors = FALSE, header = TRUE)

# Read-in the conversion rate table

conversionRates <- read.csv('../input/conversionRates.csv', header = TRUE)
```

5.2 Dataset Sample

The data for this analysis will be obtained from the 2017 Kaggle ML and Data Science Survey. It consists of:

- Demographic information (age, gender, location, years of experience)
- Job roles and titles, Salary information.
- Educational background (degree, major, institution)
- Programming languages and tools proficiency
- Industry sector and company size
- Job satisfaction and career development metrics
- Remote work and flexible employment preferences

The Data includes 4 files:

- **multipleChoiceResponses.csv:** Respondents answers to multiple choice and ranking questions. These are non-randomized and thus a single row does correspond to all of a single user's answers. **(16,717 rows and 73 columns).**

1	GenderSelect	Country	Age	EmploymentStatus	StudentStatus	LearningDataScience	CodeWriter	CareerSwitcher	CurrentJobTitleSelect
2	Female	United States	30	Not employed, but looking for work					
3	Male	Canada	28	Not employed, but looking for work					
4	Male	United States	56	Independent contractor, freelancer, or self-employed			Yes		Operations Research Practitioner
5	Male	Taiwan	38	Employed full-time			Yes		Computer Scientist
6	Male	Brazil	46	Employed full-time			Yes		Data Scientist
7	Male	United States	35	Employed full-time			Yes		Computer Scientist
8	Female	India	22	Employed full-time			No	Yes	Software Developer/Software Engi
9	Female	Australia	43	Employed full-time			Yes		Business Analyst
10	Male	Russia	33	Employed full-time			Yes		Software Developer/Software Engi
11	Female	Russia	20	Not employed, and not looking for work	Yes	Yes, I'm focused on learning mostly data science skills			
12	Male	India	27	Employed full-time			Yes		Data Scientist

- **freeformResponses.csv**: Respondents' freeform answers to Kaggle's survey questions. These responses are randomized within a column, so that reading

across a single row does not give a single user's answers. **(16,717 rows and 62 columns).**

	GenderFreeForm	KaggleMotivationFreeForm	CurrentJobTitleFreeForm	MLToolNextYearFreeForm	MLMethodNextYearFreeForm	LanguageRecommendationFreeForm	PublicDatasets
2	half man - half dog						
3			teacher				
4							
5				PyTorch			
6		Curious					
7			Hydrographic Surveyor				
8			mechanical engineer			don't know	
9		Promote our data solutions	Technical support engineer				
10							
11			Quantitative Analyst				
12	ceramic vase						
13	Male						

- **schema.csv:** schema csv: a CSV file with survey schema. This schema includes the questions that correspond to each column name in both the multipleChoiceResponses.csv and freeformResponses.csv. **(291 rows and 3 columns).**

	Column	Question	Asked
2	GenderSelect	Select your gender identity. - Selected Choice	All
3	GenderFreeForm	Select your gender identity. - A different identity - Text	All
4	Country	Select the country you currently live in.	All
5	Age	What's your age?	All
6	EmploymentStatus	What's your current employment status?	All
7	StudentStatus	Are you currently enrolled as a student at a degree granting school?	Non-worker
8	LearningDataScience	Are you currently focused on learning data science skills either formally or informally?	Non-worker
9	KaggleMotivationFreeForm	What's your motivation for being a Kaggle user?	Non-switcher
10	CodeWriter	Do you write code to analyze data in your current job, freelance contracts, or most recent job if retired?	Worker1
11	CareerSwitcher	Are you actively looking to switch careers to data science?	Worker1
12	CurrentJobTitleSelect	Select the option that's most similar to your current job/professional title (or most recent title if retired). - Selected Choice	Worker1
13	CurrentJobTitleFreeForm	Select the option that's most similar to your current job/professional title (or most recent title if retired). - Other - Text	Worker1
14	TitleFit	How adequately do you feel your title describes what you do (or what you did if retired)?	Worker1

- **conversionRates.csv**: Currency conversion rates (to USD) as accessed from the R package "quantmod" on September 14, 2017. **(87 rows and 3 columns).**

		originCountry	exchangeRate
2	1	USD	1
3	2	EUR	1.195826
4	3	INR	0.01562
5	4	GBP	1.324188
6	5	BRL	0.32135
7	6	RUB	0.017402
8	7	CAD	0.823688
9	8	AUD	0.80231
10	9	JPY	0.009108
11	10	CNY	0.153

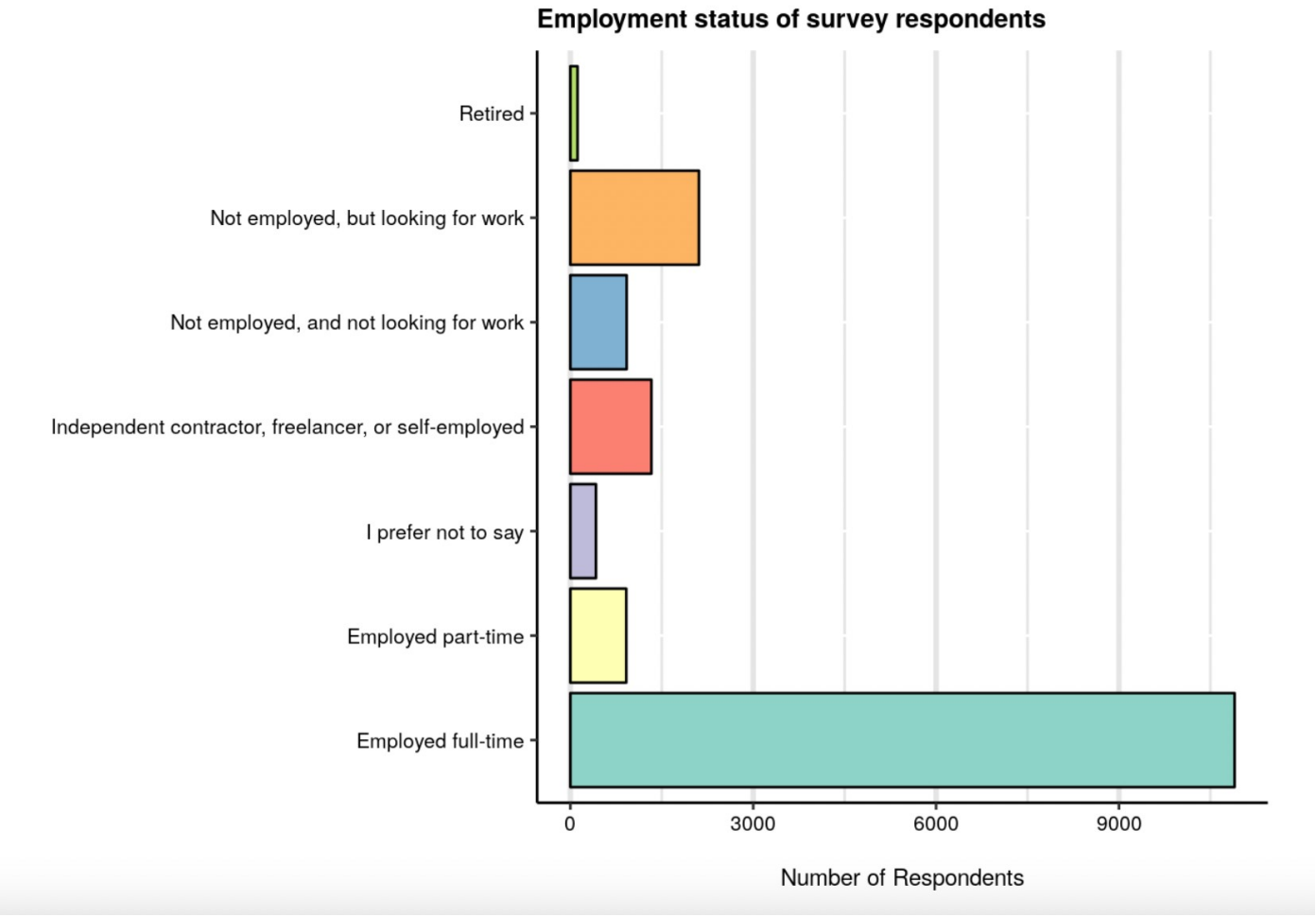
6. Exploration of the Respondents

Let's get a sense of who the survey respondents are, with specialized attention in comparing job seekers and those employed.

6.1 Employment Status

First, what is the employment status of all survey respondents?

```
MCData %>%
 group_by(EmploymentStatus) %>%
 summarise(count = n()) %>%
 arrange(desc(count)) %>%
 ggplot(aes(x=EmploymentStatus, y = count)) +
 geom_bar(aes(fill = EmploymentStatus), stat = "identity",color = "black" ) +
 coord_flip()  +
 labs(x = "", y="\nNumber of Respondents") +
 ggtitle("Employment status of survey respondents") +
 theme( axis.line = element_line(size=.5, colour = "black"),
      panel.border = element_blank(), panel.background = element_blank(), panel.grid.major.x = element_line(colour="gray90", size=1)
      , panel.grid.minor.x = element_line(colour="gray90", size=.5)) +
 theme(plot.title = element_text(size = 10,  face = "bold"), legend.position = "none",
      text=element_text(size = 9),
      axis.text.x=element_text(colour="black", size = 8),
      axis.text.y=element_text(colour="black", size = 8)) + scale_fill_brewer( palette = "Set3")
```



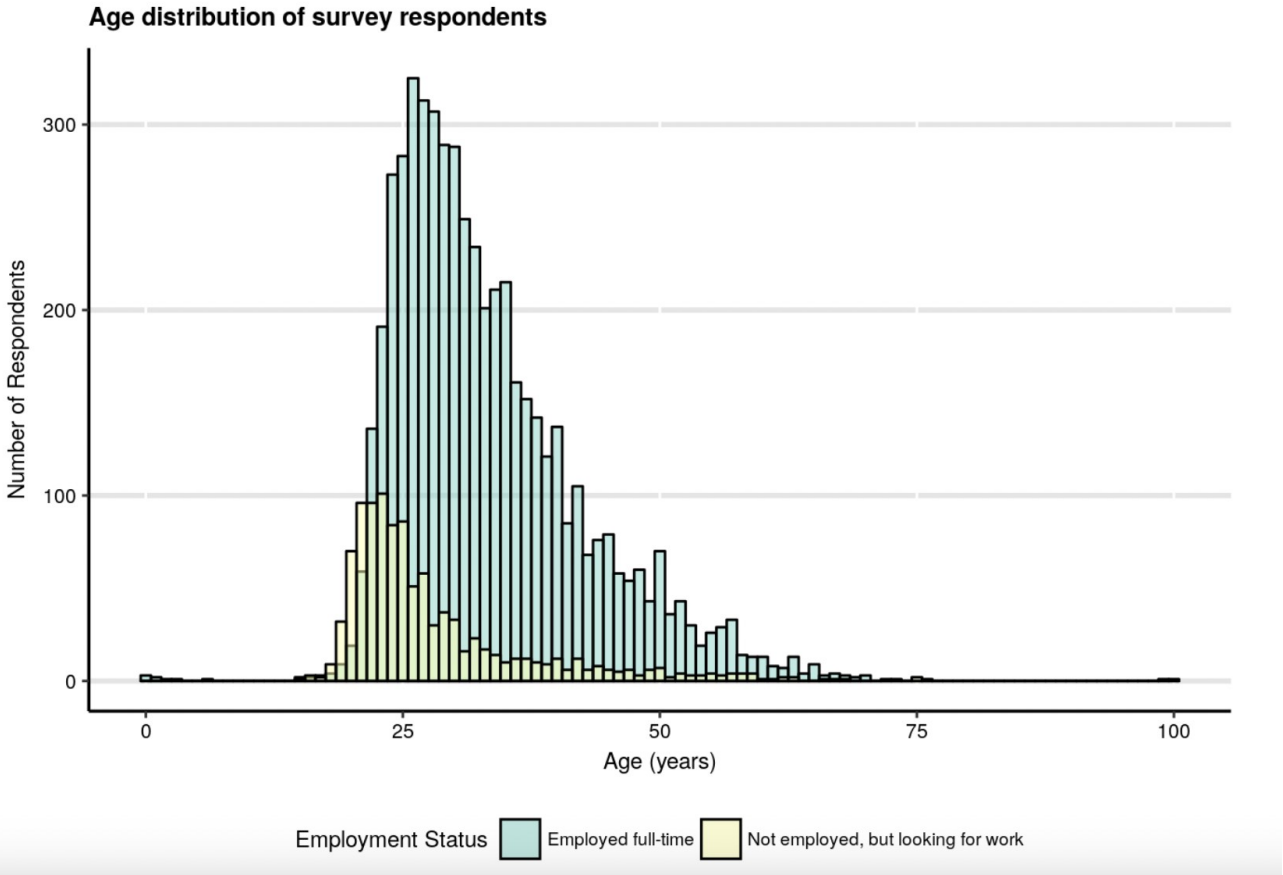
An overwhelming majority of respondents are employed (10,897 respondents), but those who are seeking work represent a sizeable sample of the survey (2,110 respondents).

6.2 Age Distribution

Let's take a look at the age distribution of job seekers vs those who are employed.

```
MCData %>%
  # Remove any rows where the respondent didn't answer the question
  filter(!Age == "") %>%
  #get only job seekers and employees
  filter(EmploymentStatus ==
           c("Employed full-time","Not employed, but looking for work")) %>%
```

```
ggplot( aes(x = Age, group = EmploymentStatus, fill = EmploymentStatus)) +
  geom_histogram(binwidth = 1, position="identity", alpha = 0.5, color = "black") +
  xlab("Age (years)") +
  ylab("Number of Respondents")  +
  ggtitle("Age distribution of survey respondents") +
  theme( axis.line = element_line(size=.6, colour = "black"),
       panel.border = element_blank(), panel.background = element_blank(), panel.grid.major.y = element_line(colour="gray90", size=1)) +
  theme(plot.title = element_text(size = 10, face = "bold"), legend.position = "bottom",
       text=element_text(size = 9),
       axis.text.x=element_text(colour="black", size = 8),
       axis.text.y=element_text(colour="black", size = 8)) + scale_fill_brewer( palette = "Set3", name = "Employment Status")
```

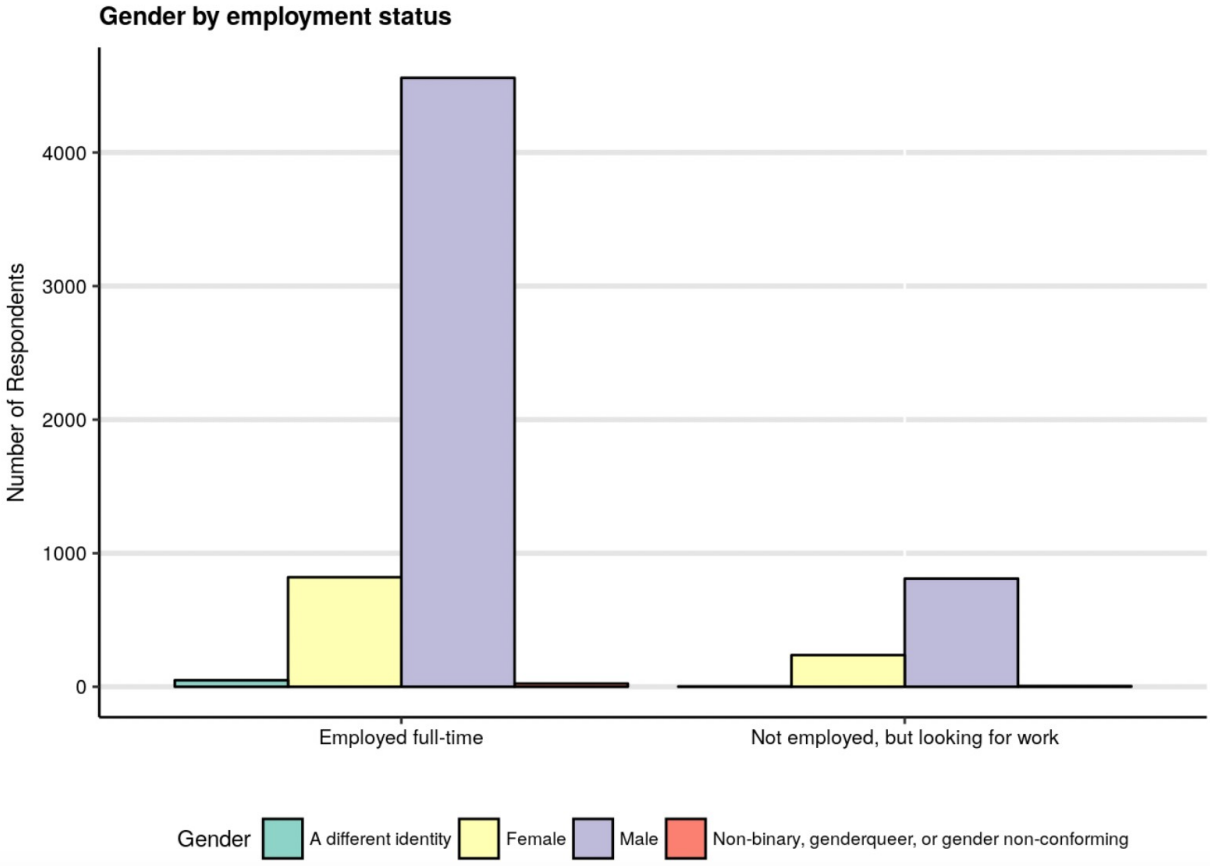


Those looking for a job follow roughly the same age distribution as those who are employed, skewed just a bit towards the younger side. This shows that job seekers are represented by those entering the job market for the first time as well as those in a mid-career job change.

6.3 By Gender

In this section, we discuss about how are the genders represented in the survey respondents.

```
MCData %>%
 # Remove any rows where the respondent didn't answer the question
 filter(!GenderSelect == "") %>%
 #get only job seekers and employees
 filter(EmploymentStatus ==
        c("Employed full-time","Not employed, but looking for work")) %>%
ggplot( aes(x = EmploymentStatus, group = GenderSelect, fill = GenderSelect)) +
geom_bar(position = "dodge", color = "black") +
 labs(x = "", y="\nNumber of Respondents") +
 ggtitle("Gender by employment status") +
 theme( axis.line = element_line(size=.5, colour = "black"),
      panel.border = element_blank(), panel.background = element_blank(), panel.grid.major.y
= element_line(colour="gray90", size=1)) +
 theme(plot.title = element_text(size = 10, face = "bold"), legend.position = "bottom",
      text=element_text(size = 9),
      axis.text.x=element_text(colour="black", size = 8),
      axis.text.y=element_text(colour="black", size = 8)) + scale_fill_brewer( palette = "Set3",
name = "Gender")
```



Clearly, males make up most of both those who are employed and job seekers.

If we dive deeper into the relationships in the groups between males and females, the table below shows that out of those employed, employed females represent 18% of the total number of employed males. Comparatively, females represent 28% of the total amount of males looking for work.

This can be interpreted to mean that more females are newly entering the data science job market, and/or that females are not being hired at the same rate as males.

```
MCData %>%

  # Remove any rows where the respondent didn't answer the question

  filter(!GenderSelect == "") %>%

  #look at only male and females since the other groups are so small

  filter(GenderSelect %in% c("Female","Male")) %>%

  filter(EmploymentStatus ==
```

```
        c("Employed full-time","Not employed, but looking for work")) %>%
  group_by(EmploymentStatus, GenderSelect) %>%
  summarise(count = n()) %>%
  mutate(percentage = round(count/sum(count)*100, digits = 1))%>%
  group_by(EmploymentStatus) %>%
  summarise(PercFemaleofMale = paste0(round(min(percentage)/max(percentage)*100,digits=
2),"%"))
```

EmploymentStatus <fctr>	PercFemaleofMale <chr>
Employed full-time	17.79%
Not employed, but looking for work	28.37%

2 rows

7. Comparing Responses from those Employed to Responses from Job Seekers

Now that we have a sense of who represents the employed and job seeker segments, let's turn to see how their responses for common questions differ in order to gain information that job seekers can utilize. The notion here is that those employed have a better sense of what it takes to get a job than those who are seeking work.

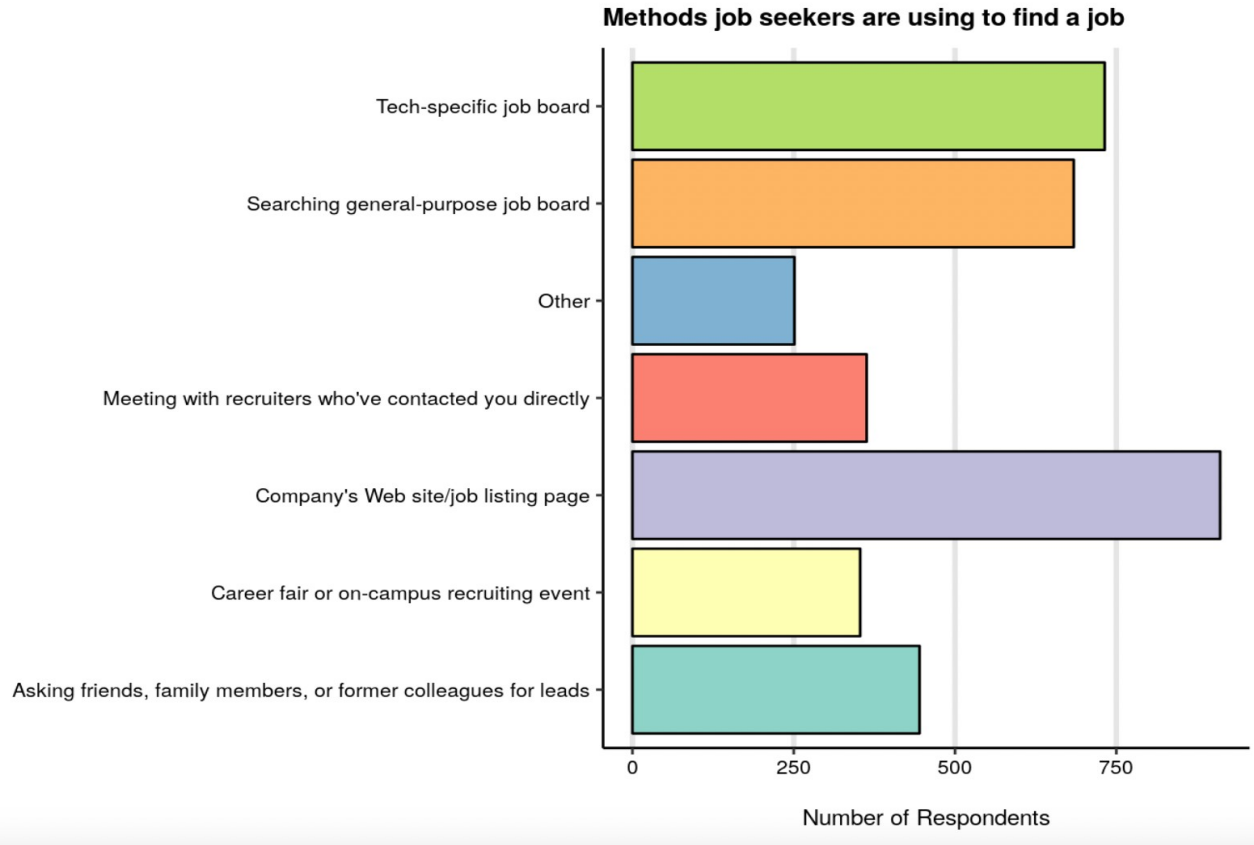
7.1 Finding a Job

Perhaps the statistic that those employed have proven themselves more knowledgeable about than job seekers is the way to get a job. Let's compare what resources job seekers are using to find a job compared to how those who are employed found a job.

First, what resources are job seekers using to find a job?

```
MCData %>%
  filter(JobSearchResource != "") %>%
  group_by(JobSearchResource) %>%
  summarise(count = n()) %>%
```

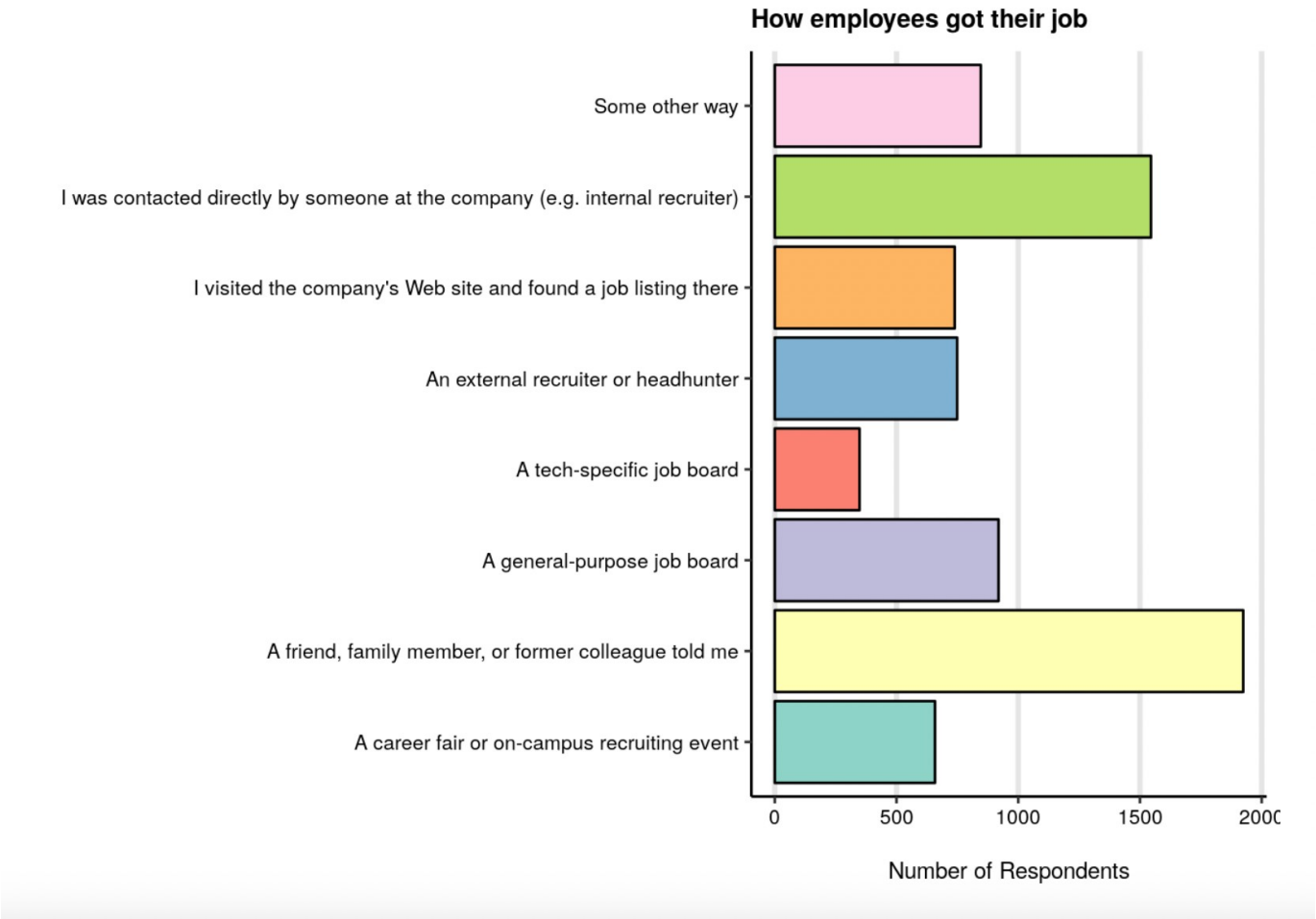
```
arrange(desc(count)) %>%
ggplot(aes(x=JobSearchResource, y = count)) +
geom_bar(aes(fill = JobSearchResource), stat = "identity",color = "black" ) +
coord_flip()+
labs(x = "", y="\nNumber of Respondents") +
ggtitle("Methods job seekers are using to find a job") +
theme( axis.line = element_line(size=.5, colour = "black"),legend.position="none",
      panel.border = element_blank(), panel.background = element_blank(), panel.grid.major.x = element_line(colour="gray90", size=1)) +
theme(plot.title = element_text(size = 10,  face = "bold"),
      text=element_text(size = 9),
      axis.text.x=element_text(colour="black", size = 8),
      axis.text.y=element_text(colour="black", size = 8)) + scale_fill_brewer( palette = "Set3")
```



Unemployed job seekers are mostly using company websites and job boards to find a job.

Now let's compare that to how employees found their current job.

```
MCData %>%
  filter(EmployerSearchMethod != "") %>%
  group_by(EmployerSearchMethod) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  ggplot(aes(x=EmployerSearchMethod, y = count)) +
  geom_bar(aes(fill = EmployerSearchMethod), stat = "identity",color = "black" ) +
  coord_flip()+
  labs(x = "", y="\nNumber of Respondents") +
  ggtitle("How employees got their job") +
  theme( axis.line = element_line(size=.5, colour = "black"),legend.position="none",
       panel.border = element_blank(), panel.background = element_blank(), panel.grid.major.x
= element_line(colour="gray90", size=1)) +
  theme(plot.title = element_text(size = 10,  face = "bold"),
       text=element_text(size = 9),
       axis.text.x=element_text(colour="black", size = 8),
       axis.text.y=element_text(colour="black", size = 8)) + scale_fill_brewer( palette = "Set3")
```

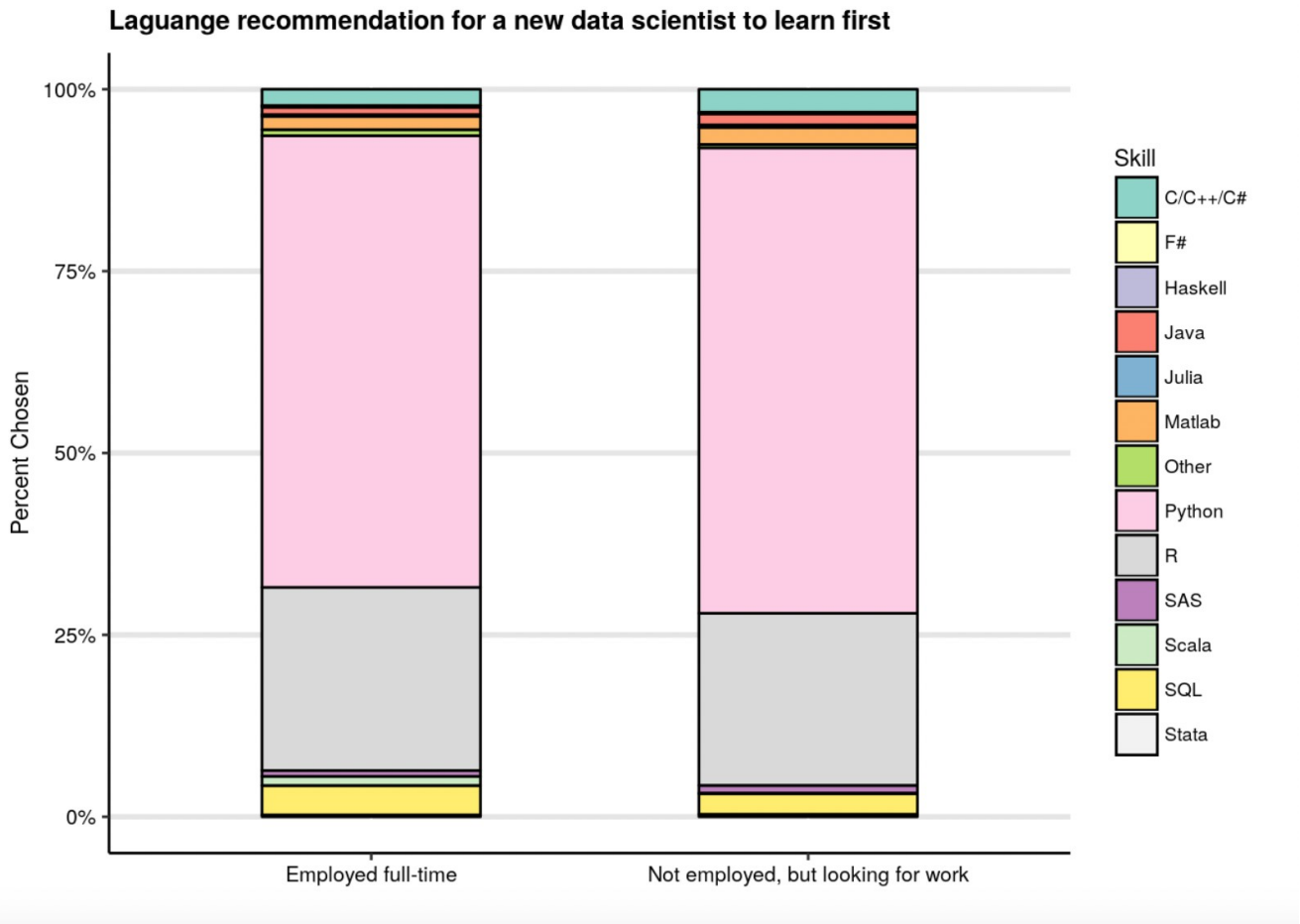


Most employees found their current job through a personal connection - a friend, family member, or current employee. This shows high evidence of the value of networking and using personal connections to find jobs. Those seeking jobs should build their personal network and use their connections. Additionally, job seekers should build their online portfolio as the second most popular way most employees found their job was that they were reached out to by someone at the company.

7.2 Language Recommendation

What programming languages does a respondent who is employed recommend to a new data scientist to learn first compared to the programming languages someone seeking a job recommends?

```
MCData %>%
  #get only job seekers and employees
  filter(EmploymentStatus ==
           c("Employed full-time","Not employed, but looking for work")) %>%
  filter(LanguageRecommendationSelect != "") %>%
  group_by(EmploymentStatus, LanguageRecommendationSelect) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  ggplot(aes(x =EmploymentStatus , y= count, fill = LanguageRecommendationSelect)) +
  geom_bar(stat = "identity", position = "fill", width=0.5, color = "black") +
  labs(x = "", y="\nPercent Chosen") +
  ggtitle("Laguange recommendation for a new data scientist to learn first") +
  scale_y_continuous(labels = scales::percent) +
  theme( axis.line = element_line(size=.5, colour = "black"),
        panel.border = element_blank(), panel.background = element_blank(), panel.grid.major.y
= element_line(colour="gray90", size=1)) +
  theme(plot.title = element_text(size = 10,  face = "bold"),
        text=element_text(size = 9),
        axis.text.x=element_text(colour="black", size = 8),
        axis.text.y=element_text(colour="black", size = 8)) + scale_fill_brewer( palette = "Set3" ,
name = "Skill")
```

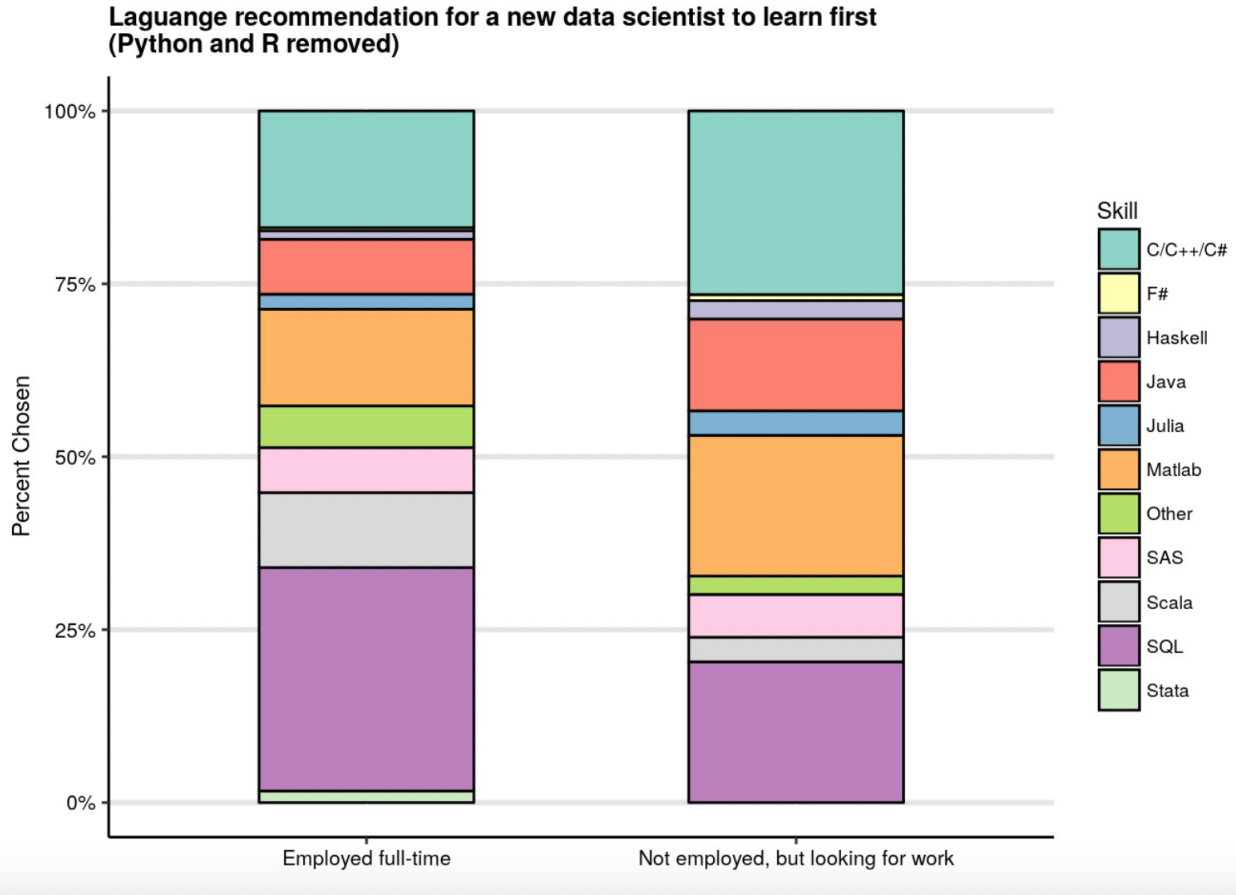


Clearly, job seekers and those employed mostly recommend Python as the programming language to learn first followed by R.

Let's remove those two from analysis to pull out more trends obfuscated by the distribution.

```
MCData %>%
 filter(LanguageRecommendationSelect != "R" ) %>%
 filter(LanguageRecommendationSelect != "Python" ) %>%
 filter(LanguageRecommendationSelect != "" ) %>%
 filter(EmploymentStatus ==
        c("Employed full-time","Not employed, but looking for work")) %>%
group_by(EmploymentStatus, LanguageRecommendationSelect) %>%
summarise(count = n()) %>%
arrange(desc(count)) %>%
```

```
 ggplot(aes(x =EmploymentStatus , y= count, fill = LanguageRecommendationSelect)) +

 geom_bar(stat = "identity", position = "fill", width=0.5, color =1) +

 labs(x = "", y="\nPercent Chosen") +

 ggtitle("Laguange recommendation for a new data scientist to learn first\n(Python and R rem
oved)") +

 scale_y_continuous(labels = scales::percent) +

 theme( axis.line = element_line(size=.5, colour = "black"),

      panel.border = element_blank(), panel.background = element_blank(), panel.grid.major.y
= element_line(colour="gray90", size=1)) +

 theme(plot.title = element_text(size = 10,  face = "bold"),

      text=element_text(size = 9),

      axis.text.x=element_text(colour="black", size = 8),

      axis.text.y=element_text(colour="black", size = 8)) + scale_fill_brewer( palette = "Set3" ,
name = "Skill")
```



Those employed full-time recommend SQL at a higher rate than those looking for work. Perhaps those looking for a job should consider learning SQL to add to their skillset. Additionally, job seekers recommend learning C/C++/C# and MATLAB at a much higher rate than those employed. Perhaps job seekers should instead focus their time on learning a language preferred by those employed instead of one of those languages.

7.3 Worthwhile Skillsets

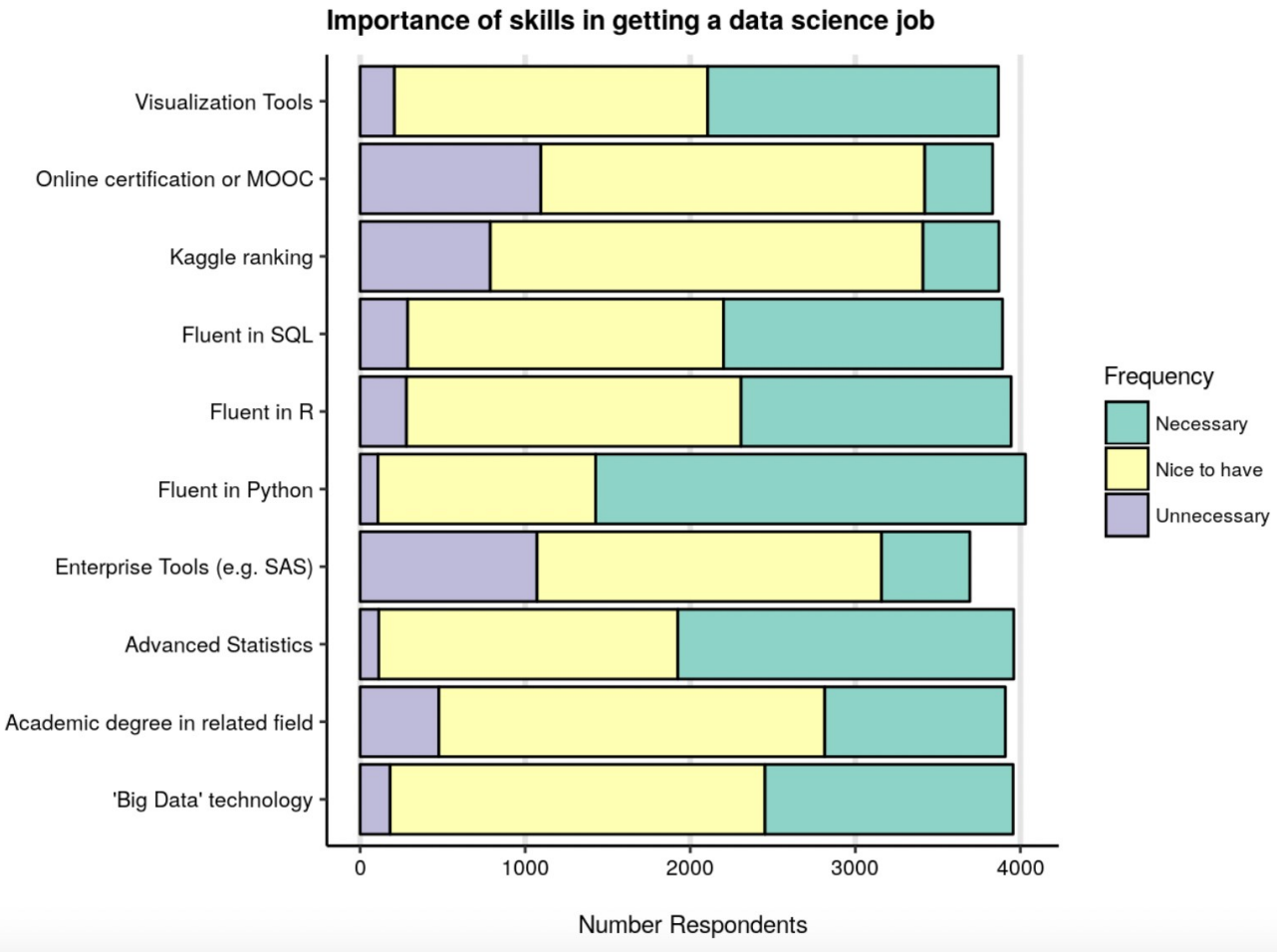
Let's compare the type of skills job seekers believe are important in getting a data science job vs the data science/analytics tools, technologies, and languages that those employed actually use at work. These were two separate questions posed in the survey with different selection options, but meaningful insights can still be extracted.

First, people who are learning data science were asked to select the importance of each of the below skills or certifications in getting a data science job.

```
# Create list of skills

skills <- schema %>%

  # From the list of questions, keep only questions that contain the below phrase

  filter(grepl("How important do you think the below skills or certifications are in getting a dat
a science job?", Question, fixed = TRUE)) %>%

  # Remove any Columns that contain "FreeForm"

  filter(!grepl("FreeForm", Column, fixed = TRUE)) %>%

  # Split the question text at the hyphen

  mutate(response = strsplit(as.character(Question), " - ")) %>%

  # Separate the responses onto separate rows

  unnest(response) %>%

  # Remove rows that contain the first phrase

  filter(!grepl("How important do you think the below skills or certifications are in getting a da
ta science job", response, fixed = TRUE)) %>%

  # Keep only the question number and the associated barrier
```

```
  select(-2)


# Limit the full dataset to the columns that answer these questions
skillsNames <- MCData %>%
  # Keep only columns that start with "JobSkillImportance" and don't contain "FreeForm"
  select(starts_with("JobSkillImportance"), -contains("FreeForm")) %>%
  select(-contains("Other")) %>%
  # Restructure Data
  gather(key = response, value = frequency) %>%
  # Remove any blank entries
  filter(!frequency == "")


# Combine the list of possible challenges with the limited dataset
skillsNamesChar <- left_join(skillsNames, skills, by = c("response" = "Column")) %>%
  # Group the responses by the question Number and the frequency with which each challenge
d is dealt with
  group_by(response.y, frequency) %>%
  # Count the number of entries within each group
  summarise(count = n()) %>%
  # Re-order the factors
  mutate(frequency = factor(frequency, levels = c("Necessary", "Nice to have", "Unnecessary")
, ordered = TRUE))


# Plot results
ggplot(skillsNamesChar, aes(x = response.y, y = count, fill = frequency)) +
  geom_bar(stat = "identity",colour = "black") +
  coord_flip()+
```

```
 labs(x = "", y="\nNumber Respondents") +
 ggtitle("Importance of skills in getting a data science job") +
 theme( axis.line = element_line(size=.5, colour = "black"),
      panel.border = element_blank(), panel.background = element_blank(), panel.grid.major.x = element_line(colour="gray90", size=1)) +
 theme(plot.title = element_text(size = 10,  face = "bold"),
      text=element_text(size = 9),
      axis.text.x=element_text(colour="black", size = 8),
      axis.text.y=element_text(colour="black", size = 8)) + scale_fill_brewer( palette = "Set3" , name = "Frequency")
```



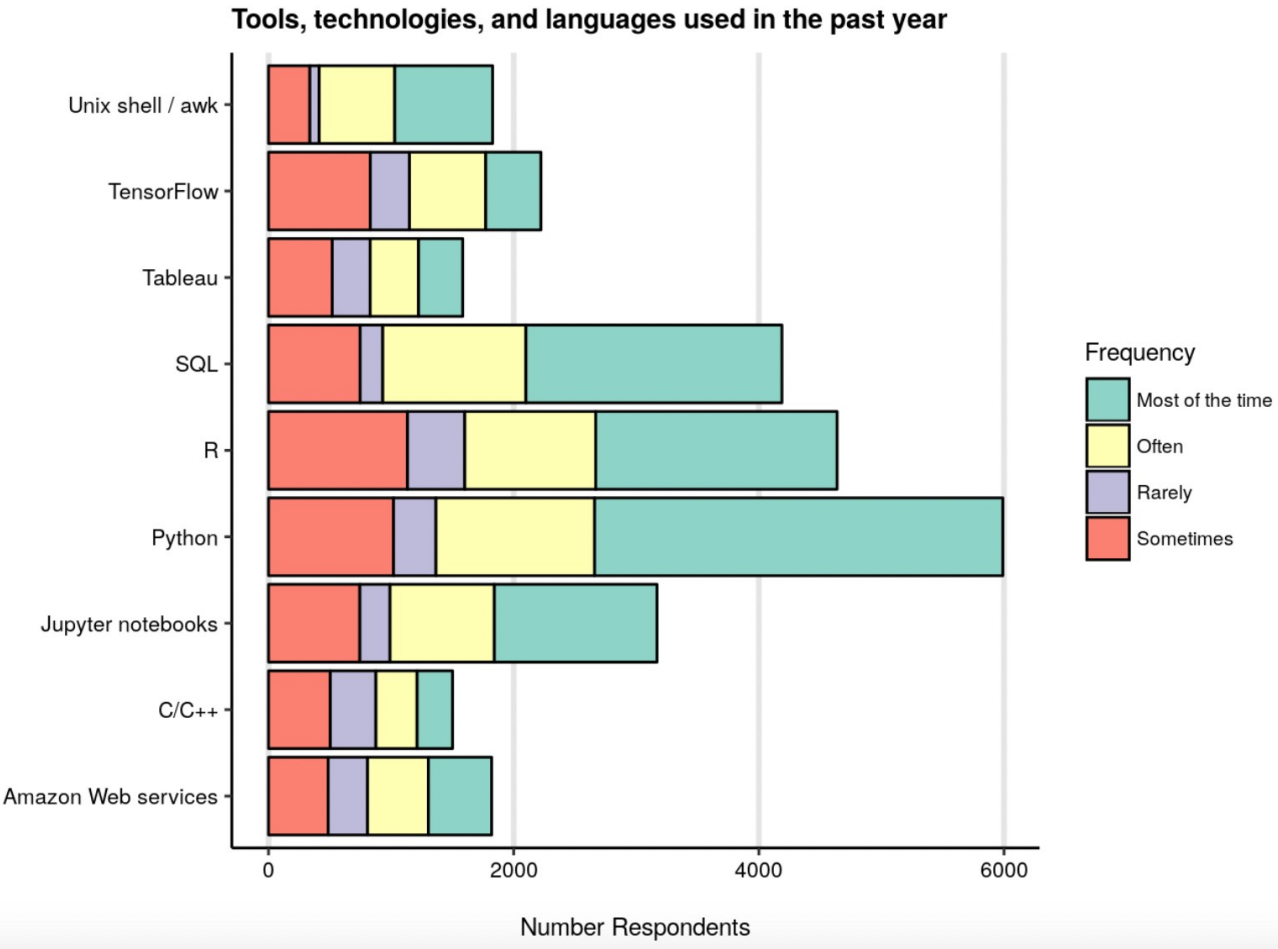
Job seekers generally believe to not waste your time with Enterprise tools such as SAS or an online certification/MOOC. They see the importance of becoming fluent in Python, SQL, and R, as well as learning advanced statistics.

Turning to what tools are actually used on the job, employed survey respondents were asked how often the following data science/analytics tools, technologies, and languages did they use in the past year. They were given an option of 54 choices to rank, but we will only look at the 9 tools chosen most frequently for clarity.

```
# Create list of tools

tools <- schema %>%

  # From the list of questions, keep only questions that contain the below phrase

  filter(grepl("At work, how often did you use the following data science/analytics tools, techn
ologies, and languages this past year?", Question, fixed = TRUE)) %>%

  # Remove any Columns that contain "FreeForm"

  filter(!grepl("FreeForm", Column, fixed = TRUE)) %>%

  # Split the question text at the hyphen

  mutate(response = strsplit(as.character(Question), " - ")) %>%

  # Separate the responses onto separate rows

  unnest(response) %>%

  # Remove rows that contain the first phrase

  filter(!grepl("At work, how often did you use the following data science/analytics tools, techn
ologies, and languages this past year", response, fixed = TRUE)) %>%

  # Keep only the question number and the associated barrier

  select(-2)


# Limit the full dataset to the columns that answer these questions

toolsNames <- MCData %>%

  # Keep only columns that start with "WorkToolsFrequency" and don't contain "FreeForm"

  select(starts_with("WorkToolsFrequency"), -contains("FreeForm")) %>%
```

```
  select(-contains("Other")) %>%

  # Restructure Data

  gather(key = response, value = frequency) %>%

  # Remove any blank entries

  filter(!frequency == "")


#keep only top by frequency

toolsNames <- toolsNames[ toolsNames$response %in% names(table(toolsNames$response))
[table(toolsNames$response) >1500] , ]


# Combine the list of possible challenges with the limited dataset

toolsNamesChar <- left_join(toolsNames, tools, by = c("response" = "Column")) %>%

  # Group the responses by the question Number and the frequency with which each challenge
d is dealt with

  group_by(response.y, frequency) %>%

  # Count the number of entries within each group

  summarise(count = n()) %>%

  # Re-order the factors

  mutate(frequency2 = factor(frequency, levels = c("Most of the time", "Often", "Sometimes",
"Rarely"), ordered = TRUE))


ggplot(toolsNamesChar, aes(x = response.y, y = count, fill = frequency)) +

  geom_bar(stat = "identity", colour="black") +

  coord_flip()+

  labs(x = "", y="\nNumber Respondents") +

  ggtitle("Tools, technologies, and languages used in the past year") +

  theme( axis.line = element_line(size=.5, colour = "black"),
```

```
      panel.border = element_blank(), panel.background = element_blank(), panel.grid.major.x
= element_line(colour="gray90", size=1)) +

 theme(plot.title = element_text(size = 10,  face = "bold"),

      text=element_text(size = 9),

      axis.text.x=element_text(colour="black", size = 8),

      axis.text.y=element_text(colour="black", size = 8)) + scale_fill_brewer( palette = "Set3" ,
name = "Frequency")
```



In practice, those employed use Python, SQL, and R regularly–which aligns to the skills job seekers believe are important in getting a job. Jupyter notebooks are also a tool many data scientists use in practice. Job seekers should consider learning how to use Jupyter notebooks to add to their resume.

8. Job Expectations

We now turn to the current job landscape so that job seekers can manage their expectations in what to look for.

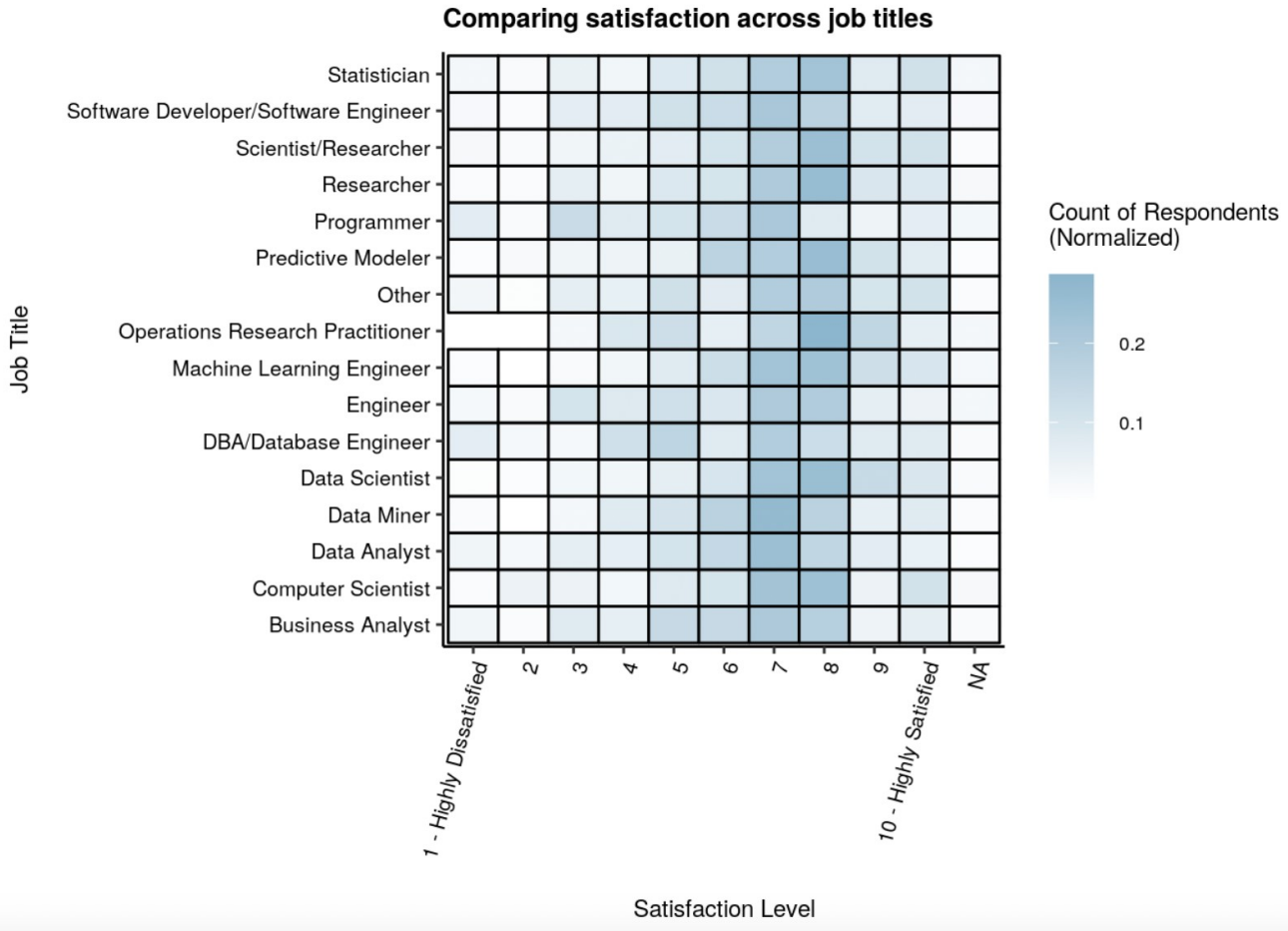
8.1 Job Satisfaction

What is the typical job satisfaction of employees surveyed and how does it differ across job titles?

To look into this, I normalized the data by job title so that the number of respondents of each title is taken into account, making cross-comparison achievable.

```
MCData %>%
 filter(JobSatisfaction != "") %>%
 filter(JobSatisfaction != "NA") %>%
 filter(CurrentJobTitleSelect != "") %>%
 filter(CurrentJobTitleSelect != "I prefer not to share") %>%
 #order job satisfaction
 mutate(JobSatisfaction = factor(JobSatisfaction, levels = c("1 - Highly Dissatisfied", "2", "3",
"4", "5","6","7","8","9","10 - Highly Satisfied"),    ordered = TRUE)) %>%
 group_by(CurrentJobTitleSelect, JobSatisfaction) %>%
 summarise(count = n()) %>%
 #normalize data
 mutate(norm = count/sum(count)) %>%
 ggplot(aes(x = JobSatisfaction, y= CurrentJobTitleSelect)) +
 geom_tile(aes(fill = norm), color = "black", size = .5)+
 scale_fill_gradient(low = "white", high="lightskyblue3" ,
                     name = "Count of Respondents\n(Normalized)") +
 labs(x="\nSatisfaction Level", y="Job Title\n") +
 ggtitle("Comparing satisfaction across job titles") +
```

```
theme( axis.line = element_line(size=.5, colour = "black"),
      panel.border = element_blank(), panel.background = element_blank()) +
theme(plot.title = element_text(size = 10,  face = "bold"),
      text=element_text(size = 9),
      axis.text.x=element_text(colour="black", size = 8, angle = 75, hjust = 1),
      axis.text.y=element_text(colour="black", size = 8))
```



Employees generally rate their job satisfaction in the 7-8 range across all job titles. Those with the "Programmer" job title are more likely to select a lower job satisfaction compared to other job titles.

8.2 Salary

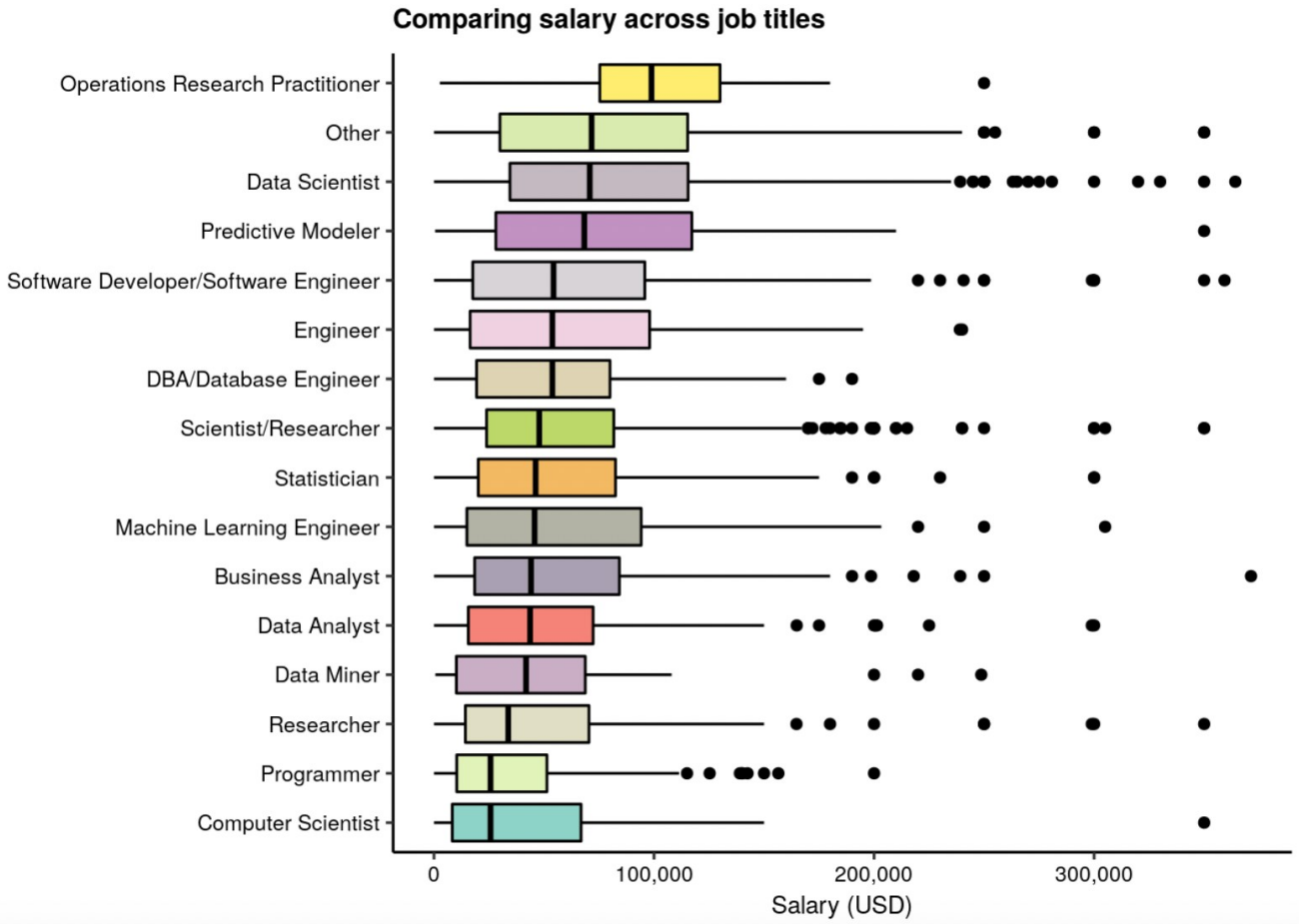
Salary is a very important component to factor into job expectations. Note that for this part of the analysis, only ~30% of survey respondents answered this question. This isn't surprising since it may be considered a private question. We can still analyze the responses from those who elected to answer. For this analysis, we convert all currencies to US dollars.

How does salary vary across job titles?

```
reportedSalary <- MCData %>%

 # Keep only the columns with the reported salary and currency

 select(c("CompensationAmount", "CompensationCurrency", "CurrentJobTitleSelect")) %>%

 # Remove any blank responses

 filter(!CompensationCurrency == "") %>%

 filter(!CompensationAmount == "") %>%

 filter(!CurrentJobTitleSelect == "")



# Combine the reported salary data with the conversion rate

salaryUSD <- left_join(reportedSalary, conversionRates, by = c("CompensationCurrency" = "
originCountry")) %>%

 # Convert the reported salary to a character string

 mutate(CompensationAmount = as.character(CompensationAmount),

        # Remove any commas from the salary entry and convert it to a number

        originalSalary = as.numeric(gsub(",", "", CompensationAmount)),

        # Convert the exchange rate to a number

        exchangeRate = as.numeric(as.character(exchangeRate)),

   # Calculate the salary in USD

   usSalary = originalSalary * exchangeRate,

   # Convert the calculated salary to a number and round to 2 decimal places

   usSalary = as.numeric(format(round(usSalary, 2), nsmall = 2, scientific = FALSE))) %>%
```

```
  arrange(desc(usSalary))


# For graphing purposes, remove responses over $400,000 (since there are very few above that limit)
salaryUSDPlot <- salaryUSD %>%
  # Remove any salaries (in USD) that are above $400,000 or less than or equal to 0
  filter(usSalary < 400000,
         usSalary > 0)


# Reorder the data by median of salary by job title for boxplot
salaryUSDPlot$CurrentJobTitleSelect = reorder(salaryUSDPlot$CurrentJobTitleSelect, salary USDPlot$usSalary, median)


#interpolate more colors onto the Set3 pallette
cols <- colorRampPalette(brewer.pal(12, "Set3"))
myPal <- cols(length(unique(salaryUSDPlot$CurrentJobTitleSelect)))


ggplot(salaryUSDPlot, aes(CurrentJobTitleSelect, usSalary)) +
  geom_boxplot(aes(fill = CurrentJobTitleSelect), color= "black") + coord_flip() +
  ylab("Responses of a given Salary (in USD)") +
  theme(legend.position="none") +
  labs(x="", y="Salary (USD)\n") +
  ggtitle("Comparing salary across job titles") +
  theme( axis.line = element_line(size=.5, colour = "black"),
         panel.border = element_blank(), panel.background = element_blank()) +
  theme(plot.title = element_text(size = 10,  face = "bold"),
         text=element_text(size = 9),
```

```
      axis.text.x=element_text(colour="black", size = 8),
      axis.text.y=element_text(colour="black", size = 8)) +  scale_fill_manual(values = myPal)
+
 scale_y_continuous(labels = scales::comma)
```



The highest median reported salary comes from those with the job title "Operations Research Practitioner" followed by "Other" and "Data Scientist". However, it's important to keep in mind here how many respondents answered in each group. The more respondents, the more confident we can be that the reported salary is reflective of the true salary of the job title in the job market. The table below shows how many respondents answered the salary question in each job title group to complement the above boxplot.

```
salaryUSDPlot %>%
```

```
group_by(CurrentJobTitleSelect) %>%
summarise(count = n())
```

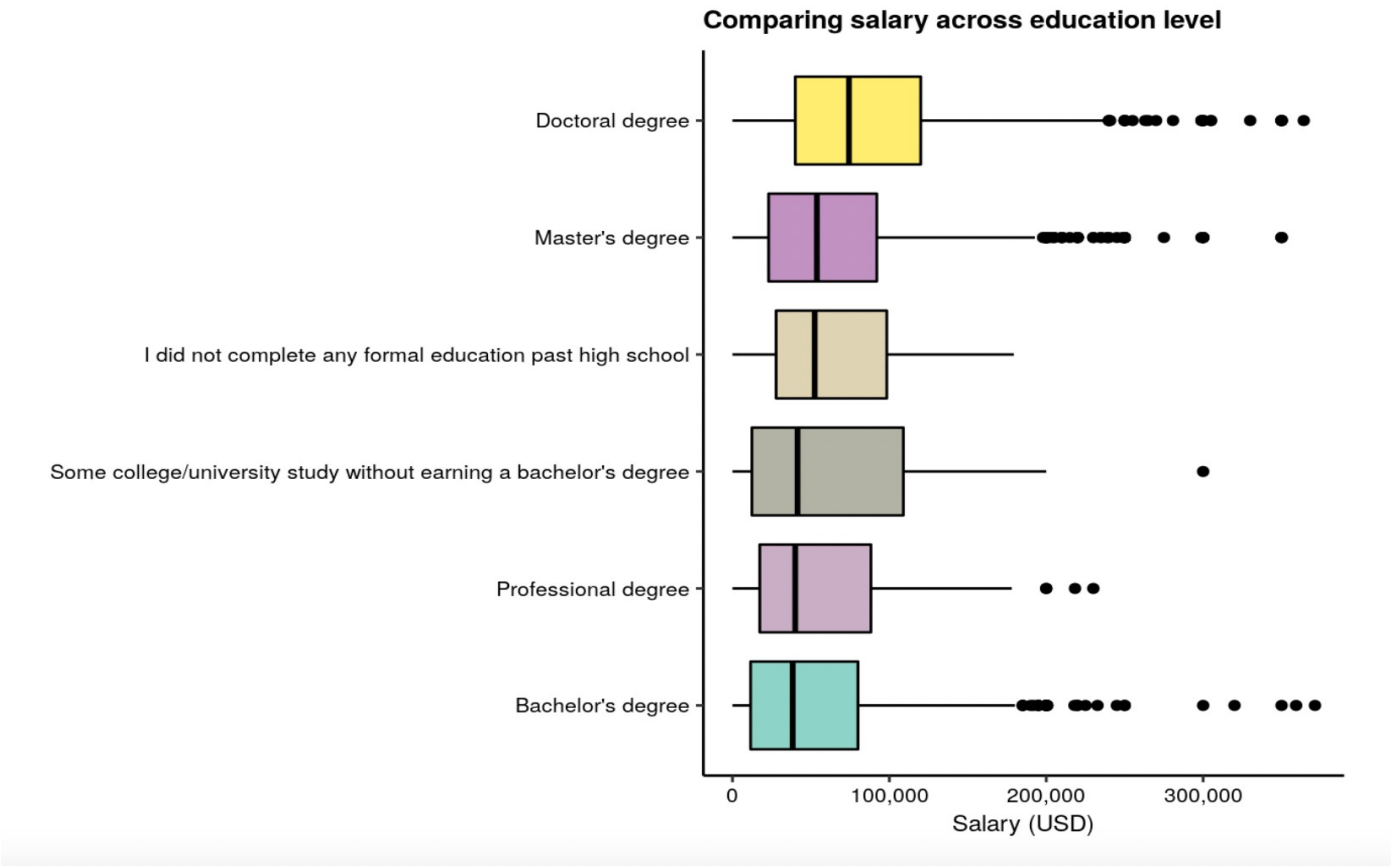
CurrentJobTitleSelect <fctr>	count <int>
Computer Scientist	105
Programmer	88
Researcher	228
Data Miner	39
Data Analyst	477
Business Analyst	226
Machine Learning Engineer	276
Statistician	119
Scientist/Researcher	407
DBA/Database Engineer	66

Let's perform the same exercise comparing salary, this time by educational attainment.

```
reportedSalary <- MCData %>%
 # Keep only the columns with the reported salary and currency
 select(c("CompensationAmount", "CompensationCurrency", "FormalEducation")) %>%
 # Remove any blank responses
 filter(!CompensationCurrency == "") %>%
 filter(!CompensationAmount == "") %>%
 filter(!FormalEducation == "")%>%
 filter(!FormalEducation == "I prefer not to answer")
```

```
# Combine the reported salary data with the conversion rate
salaryUSD <- left_join(reportedSalary, conversionRates, by = c("CompensationCurrency" = "
originCountry")) %>%
  # Convert the reported salary to a character string
  mutate(CompensationAmount = as.character(CompensationAmount),
         # Remove any commas from the salary entry and convert it to a number
         originalSalary = as.numeric(gsub(",", "", CompensationAmount)),
         # Convert the exchange rate to a number
         exchangeRate = as.numeric(as.character(exchangeRate)),
    # Calculate the salary in USD
    usSalary = originalSalary * exchangeRate,
    # Convert the calculated salary to a number and round to 2 decimal places
    usSalary = as.numeric(format(round(usSalary, 2), nsmall = 2, scientific = FALSE))) %>%
  arrange(desc(usSalary))

# For graphing purposes, remove responses over $400,000 (since there are very few above that
limit)
salaryUSDPlot <- salaryUSD %>%
  # Remove any salaries (in USD) that are above $400,000 or less than or equal to 0
  filter(usSalary < 400000,
         usSalary > 0)

# Reorder the data by median of salary by education for boxplot
salaryUSDPlot$FormalEducation = reorder(salaryUSDPlot$FormalEducation, salaryUSDPlot
$usSalary, median)
```

```
#interpolate more colors onto the Set3 pallette
cols <- colorRampPalette(brewer.pal(12, "Set3"))
myPal <- cols(length(unique(salaryUSDPlot$FormalEducation)))


ggplot(salaryUSDPlot, aes(FormalEducation, usSalary)) +
  geom_boxplot(aes(fill = FormalEducation), color= "black") + coord_flip() +
  ylab("Responses of a given Salary (in USD)") +
  theme(legend.position="none") +
  labs(x="", y="Salary (USD)\n") +
  ggtitle("Comparing salary across education level") +
  theme( axis.line = element_line(size=.5, colour = "black"),
        panel.border = element_blank(), panel.background = element_blank()) +
  theme(plot.title = element_text(size = 10,  face = "bold"),
        text=element_text(size = 9),
        axis.text.x=element_text(colour="black", size = 8),
        axis.text.y=element_text(colour="black", size = 8)) +  scale_fill_manual(values = myPal)
+
  scale_y_continuous(labels = scales::comma)
```



The highest median salary comes from those who report to have a Doctoral degree followed by a master's degree. Interestingly enough, the lowest median salary comes from those who hold a bachelor's degree. However, I believe this is attributed to an issue with sample size. As you can see in the table below, there were many more respondents with a Doctoral, Master's, or bachelor's degree than the other three categories.

```
salaryUSDPlot %>%
  group_by(FormalEducation) %>%
  summarise(count = n())
```

FormalEducation <fctr>	count <int>
Bachelor's degree	1096
Professional degree	129
Some college/university study without earning a bachelor's degree	110
I did not complete any formal education past high school	26
Master's degree	1978
Doctoral degree	966

6 rows

9. Conclusion

As we wrap up our exploration of the data science job market, we've taken a close look at who's seeking jobs, what they're recommending, and how expectations play out. With a focus on job seekers, we've also checked how education affects job outcomes and job satisfaction across different roles. Now, let's tie it all together to see what it means for those looking to break into data science and those already in the field.

1. What do the respondents look like in terms of demographic information?

- An overwhelming majority of respondents are employed (10,897 respondents), but those who are seeking work represent a sizeable sample of the survey (2,110 respondents).
- Those looking for a job follow roughly the same age distribution as those who are employed, skewed just a bit towards the younger side. This shows that job seekers are represented by those entering the job market for the first time as well as those in a mid-career job change.
- Males make up most of both those who are employed and job seekers.
- This can be interpreted to mean that more females are newly entering the data science job market, and/or that females are not being hired at the same rate as males.

2. How do those who are employed differ from job seekers in recommendations?

- **Language Recommendation**
  Job seekers and those employed mostly recommend Python as the programming language to learn first followed by R. Those employed full-time recommend SQL at a higher rate than those looking for work. Perhaps those looking for a job should consider learning SQL to add to their skillset. Additionally, job seekers recommend learning C/C++/C# and MATLAB at a much higher rate than those employed. Perhaps job seekers should instead focus their time on learning a language preferred by those employed instead of one of those languages.

- **Finding a Job**
  Unemployed job seekers are mostly using company websites and job boards to find a job. Most employees found their current job through a personal connection - a friend, family member, or current employee. This shows high evidence of the value of networking and using personal connections to find jobs. Those seeking jobs should build their personal network and use their connections. Additionally, job seekers should build their online portfolio as the second most popular way most employees found their job was that they were reached out to by someone at the company.

- **Worthwhile Skillsets**

Job seekers generally believe to not waste your time with Enterprise tools such as SAS or an online certification/MOOC. They see the importance of becoming fluent in Python, SQL, and R, as well as learning advanced statistics. In practice, those employed use Python, SQL, and R regularly–which aligns to the skills job seekers believe are important in getting a job. Jupyter notebooks are also a tool many data scientists use in practice. Job seekers should consider learning how to use Jupyter notebooks to add to their resume.

3. **How should job seekers set their expectations in a job?**

- **Job Satisfaction**
  Employees generally rate their job satisfaction in the 7-8 range across all job titles. Those with the "Programmer" job title are more likely to select a lower job satisfaction compared to other job titles.

- **Salary**
  The highest median reported salary comes from those with the job title "Operations Research Practitioner" followed by "Other" and "Data Scientist". However, it's important to keep in mind here how many respondents answered in each group. The more respondents, the more confident we can be that the reported salary is reflective of the true salary of the job title in the job market.

4. **Does pursuing higher education impact job search outcomes in the field of Data Science?**

   The highest median salary comes from those who report to have a Doctoral degree followed by a Master's degree. Interestingly enough, the lowest median salary comes from those who hold a Bachelor's degree.

5. **What is the typical job satisfaction of employees surveyed and how does it differ across job titles?**

   To look into this, I normalized the data by job title so that the number of respondents of each title is taken into account, making cross-comparison achievable. Employees generally rate their job satisfaction in the 7-8 range across all job titles. Those with the "Programmer" job title are more likely to select a lower job satisfaction compared to other job titles.

In simple terms, the data science job scene is complex but full of opportunities and challenges. By understanding the findings we've discussed, job seekers and employers can make better decisions. As the field grows, knowing these things will help everyone involved whether they're just starting out or have been in the game for a while – to succeed.

10. Future Scope

Looking ahead, there are several avenues for further exploration and research in the realm of data science job market analysis. Here are some potential areas of focus:

1. **Longitudinal Studies:** Conducting longitudinal studies to track the career trajectories of data science professionals over time could provide deeper insights into factors influencing job transitions, career advancement, and job satisfaction.

2. **Comparative Analyses:** Comparing data from multiple years of Kaggle surveys or across different platforms could reveal evolving trends in the data science job market, such as changes in demand for specific skills, shifts in salary trends, or variations in job satisfaction across industries.

3. **Inclusion and Diversity**: Investigating the representation of underrepresented groups, such as women and minorities, in the data science workforce could shed light on barriers to entry and strategies for fostering diversity and inclusion in the field.

4. **Impact of Emerging Technologies:** With the rapid advancements in technologies like artificial intelligence, machine learning, and big data analytics, exploring their impact on the job market landscape, including the emergence of new job roles and skill requirements, could be insightful.

5. **Geographical Analysis:** Conducting geographical analyses to understand regional variations in job opportunities, salary levels, and job satisfaction within the data science field could inform job seekers about potential relocation opportunities and employers about talent hotspots.

By delving deeper into these areas, researchers can contribute to a better understanding of the evolving dynamics of the data science job market, thereby facilitating informed decision-making for both individuals and organizations operating in this space.

11. References

[1] A. Smith et al., "Pathways to Data Science: Educational Backgrounds of Analytics Professionals," Journal of Data Science Education, vol. 3, no. 2, pp. 23-45, 2015.

[2] B. Jones and C. Brown, "Employment Trends in Data Science: A Sectoral Analysis," International Journal of Data Science and Analytics, vol. 2, no. 1, pp. 67-89, 2016.

[3] X. Lee and M. Johnson, "Recruiting Challenges in Data Science: Insights from Industry Surveys," Journal of Big Data Management, vol. 7, no. 3, pp. 112-130, 2014.

[4] 2017 Kaggle ML and Data Science Survey.