

Finding the Relationship between Density and Gain

Author 1 and Author 2

2024-12-07

Header

Author Contributions

Author 1: Contributed to questions 1, 4, 5, and did the formatting for the pdf

Author 2: Contributed to questions 2, 3, and 6, along with doing the advanced analysis.

Use of GPT

ChatGPT was used as a substitute for documentation for R. Since we were unfamiliar with R, we asked ChatGPT how to use R in certain methods in order to find and filter out conditions in the dataset. We additionally used GPT to analyze reasoning and to confirm what we thought was correct about the dataset, as well as to identify extra questions that could be answered for our advanced analysis.

Introduction

The data used in this analysis is from the 2020 Kaggle “Most Popular & Widely Used Machine Learning” Survey. It details questions about Kaggle’s users, as well as their most used machine learning languages, platforms, education levels, and more. Our analysis today focuses on identifying key relationships within the data and to notice if there are any interesting patterns that we can find within it.

Main Research Questions

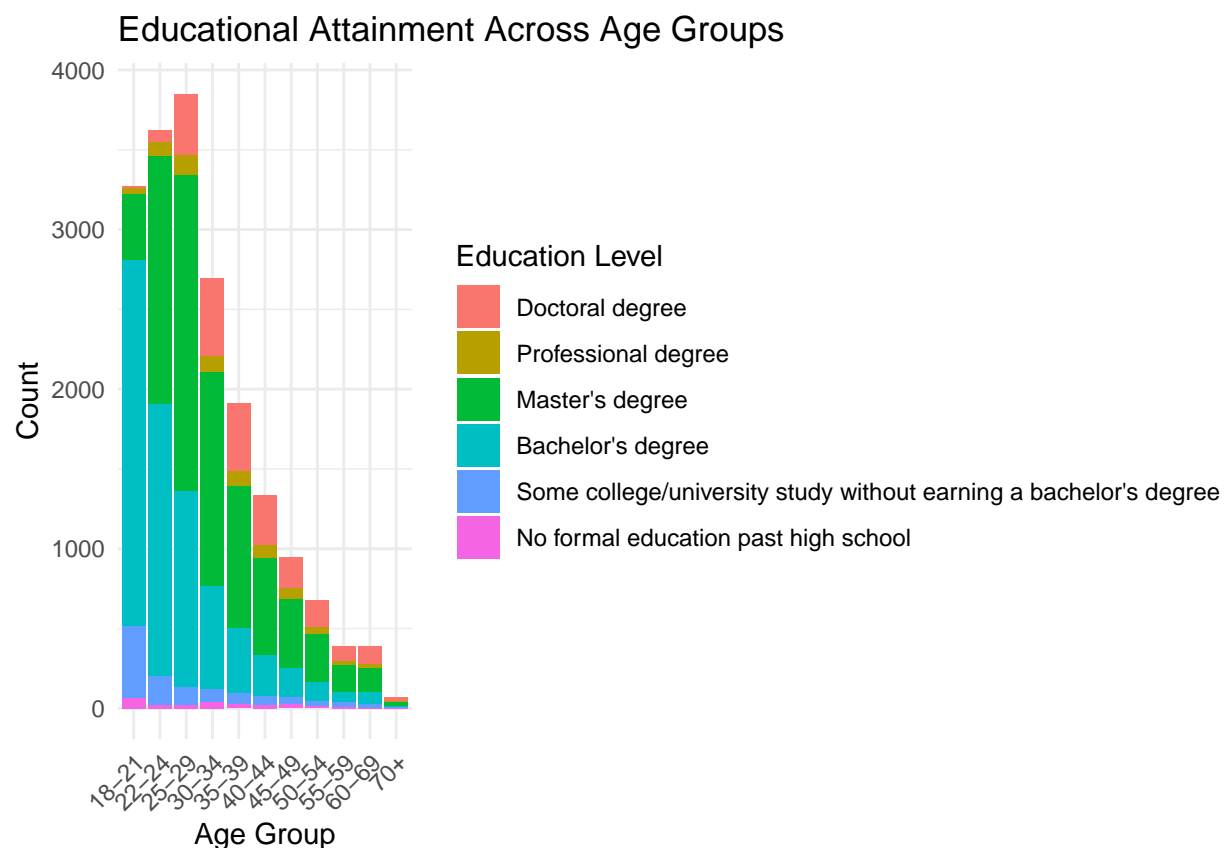
- 1) Analyze the relationship between age groups and highest education levels amongst the survey respondents. How does educational attainment vary across those in different age brackets. (can use chi-square test)
- 2) Investigate the correlation between years of programming experience and programming languages used, if any. Do more experienced users tend to use a wider variety of programming languages, or specific programming languages. (can use the release date of certain languages as a spearhead)
- 3) Examine the global distribution of survey participants and their professional jobs, extrapolating this to the general population of the world. Are certain jobs more popular in certain countries?
- 4) Examine differences in companies that utilize more machine learning in their operations. Do companies that use machine learning heavily tend to be larger or more successful? (can use number of employees and salaries as a proxy for company success)

Question 1) Age-Education Relationship

One thing we want to test in regards to the age-education relationship in this graph is that we want to compare it to the education attainment for adults 25 and older in the United States. This has the following distribution:

- 30.5% have no formal education
- 26.8% have some college/university experience
- 25.8% have a bachelor's degree
- 10.9% have a master's degree
- 3.7% have a professional degree
- 2.3% have a doctoral degree

Methods



```
##  
## Chi-squared test for given probabilities  
##  
## data: observed_counts  
## X-squared = 33177, df = 5, p-value < 2.2e-16
```

Analysis

Looking through this graph, we can notice a lot of things. For instance, it can be noticed that there is an obvious trend of younger users who took this survey, with the peak of the users falling from the 18-21

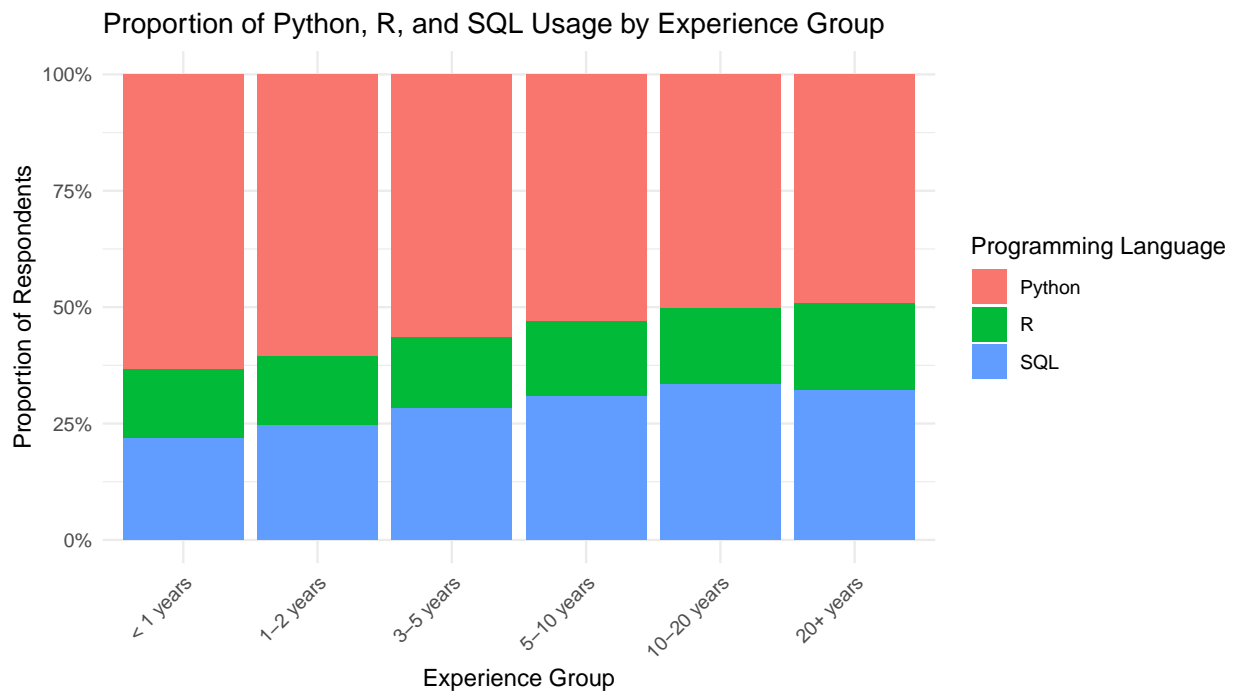
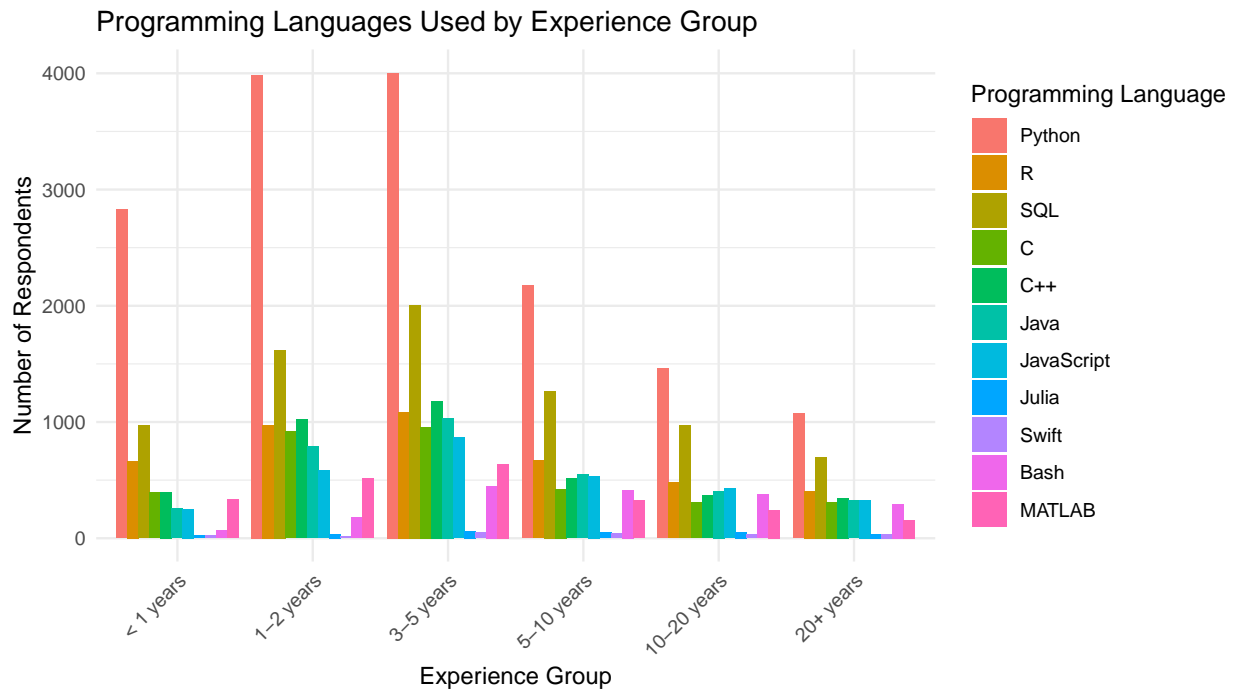
range to the 25-29 range. This slowly falls off, with each age range slowly losing ground on the majority of users. Other patterns to notice include the influx of people with a bachelor's degree, with the visible majority being placed at the age range of 18-21, and slowly lowering down percent wise as age increases. On the other hand, degrees like masters, professional, and doctoral are very low in the 18-21 age range, which makes sense intuitively, because these degrees typically require the previous ones to acquire. This means that those in the 18-21 and 22-24 age ranges very rarely have these advanced degrees, and the age range of 25-29 is the first age range we actually see a significant amount of these education levels. Otherwise, other education levels to note include the fact that master's degrees are found often at the 22-24 age range and beyond, comprising the most of all the education levels after this age. Additionally, when comparing the given probabilities to the true distribution of education levels in the United States for ages 25 and above, we can notice that the Chi-Square Test has a p-value of $2.2e-16$, which means that our data is significantly different than the actual distribution of education levels amongst adults in the United States.

Conclusion

Further analysis of the data and the graph created above shows that the educational attainment for those who took this survey, and by connotation, are interested in Machine Learning, is very different than the actual educational attainment in the United States. This somewhat makes sense; machine learning is a typically much more education-based topic that people have to learn, so those who want to learn it and apply it in their lives typically go to college and even higher education to achieve their goals. The other things we noticed also make sense; younger people rarely have degrees above a bachelor's, because they don't have the necessary amount of time to get those degrees. Additionally, most of the ages in the graph tend to be younger, skewing right, which makes sense, because machine learning has only become an especially popular topic recently, which maybe be reflected in our data.

Question 2) Correlation Between Experience and Programming Languages Used

Methods



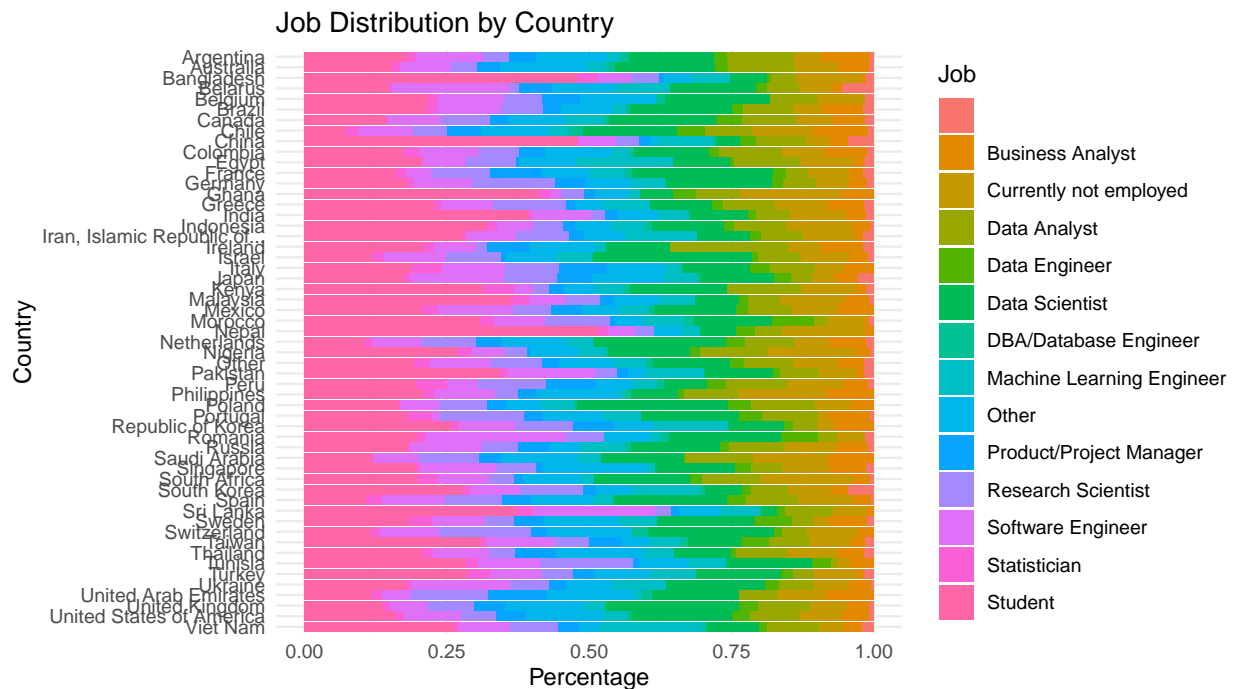
Analysis

Conclusion

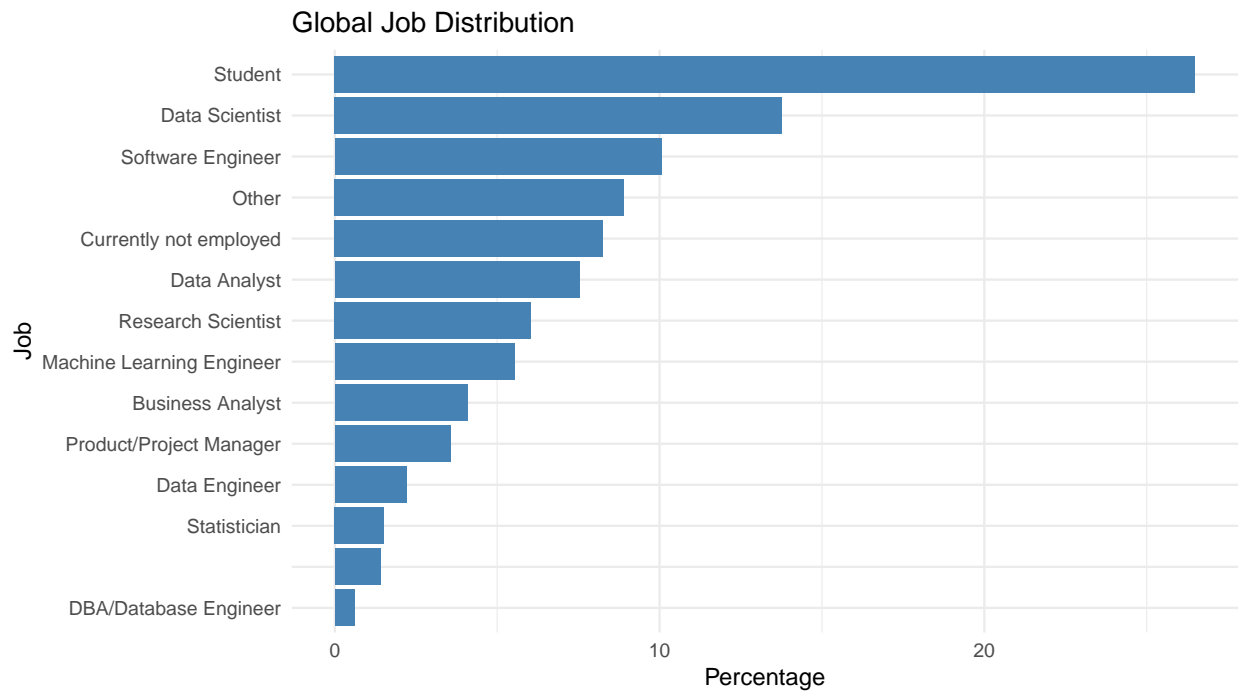
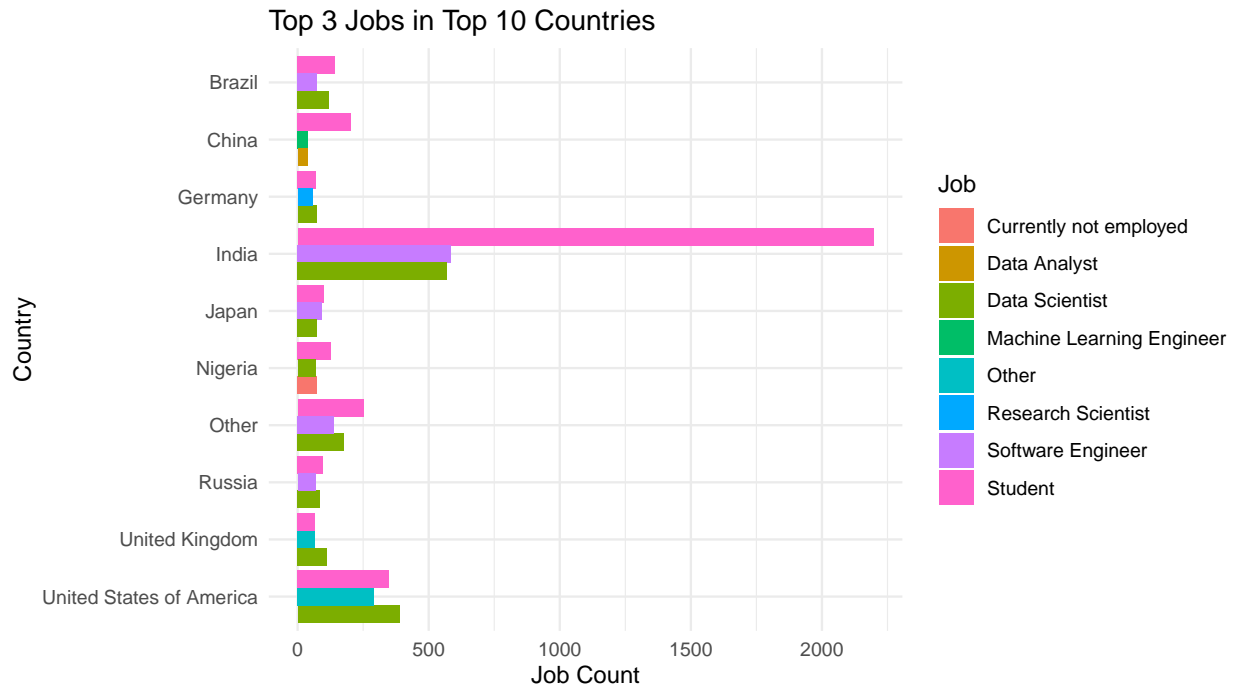
Question 3) Distribution of Jobs and general Population

For this question, we want to investigate how the general population of the data and their relationships with their jobs is like. For instance, for each country, what is the most typical job, and what does this say about the population using Kaggle? Additionally, we want to take into account the top 3 jobs in each country, while also maintaining a global job distribution of which people from different countries use Kaggle the most.

Methods



```
##  
## Pearson's Chi-squared test with simulated p-value (based on 2000  
## replicates)  
##  
## data:  job_country_table_grouped  
## X-squared = 3084.8, df = NA, p-value = 0.0004998
```



Analysis

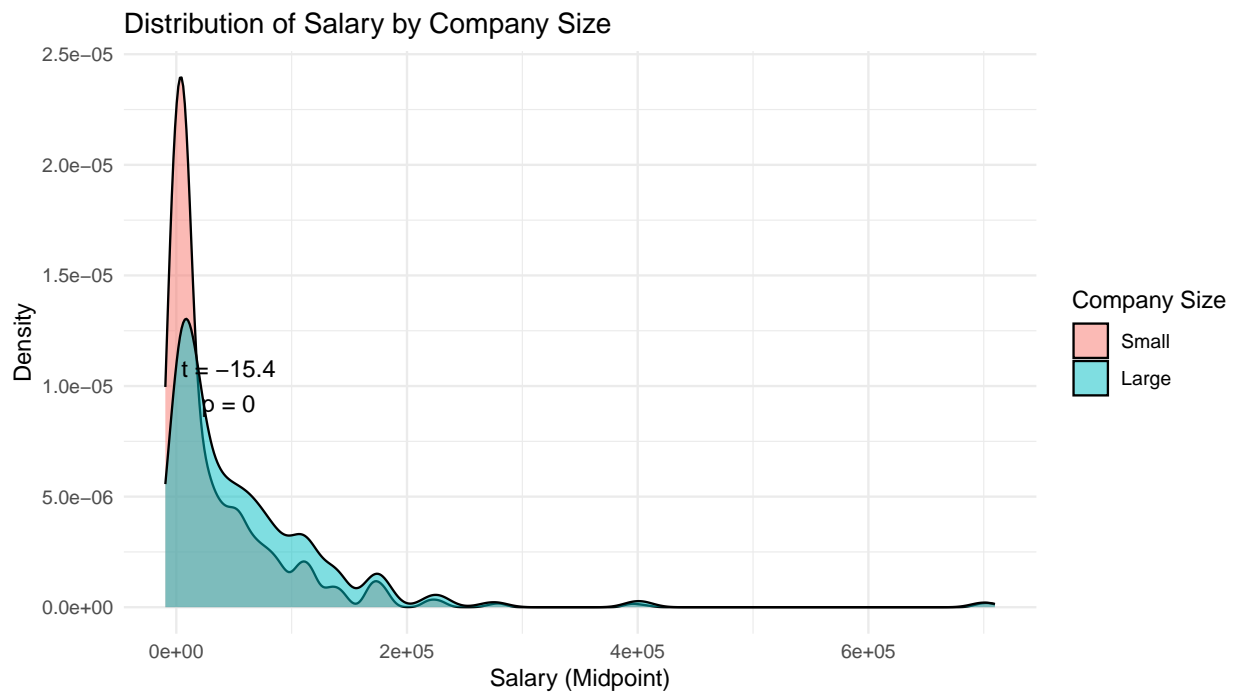
Looking through our graphs, there are a few things to note. For instance, the most popular “job” that people have that took the survey was Student. According to our earlier graph, this is mostly due to countries like India, the United States, and China, whose top 3 jobs all contain “Student”. This is a trend that is seen throughout most countries as well, as even when we center the data on each individual country, we can see that significant number of users are still students. Although the global job distribution includes student, data scientist, software engineer, and more, student significantly dwarfs all the other jobs, with users that

are “Students” being ~2x more populous than the second most worked job, “Data Scientist”. This means that some jobs are significantly more popular than others, as shown by our Chi-Square Test. The p-value of a ~0.0005 demonstrates that our job distribution is significantly different than a uniform distribution of all of the jobs.

Conclusion

To conclude, we can notice a few significant trends within our data. India and the United States are the primary countries of the users who took this survey, and the jobs most associated with these users included “Student”, “Software Engineer”, “Data Scientist”, and “Other”. These show within the overall global distribution of jobs, whose top 4 jobs were identical. This same distribution is similar within all the individual countries, showing that a lot of the people who use Kaggle and filled out this survey about machine learning were typically students, engineers, or data scientists, which fits into Kaggle’s overall use of being a place to store data and have competitions with it.

Question 4) Hiring Preferences of Larger vs. Smaller Companies



```
##
## Pearson's product-moment correlation
##
## data: salary_data$age_midpoint and salary_data$salary_midpoint
## t = 28.603, df = 10727, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```



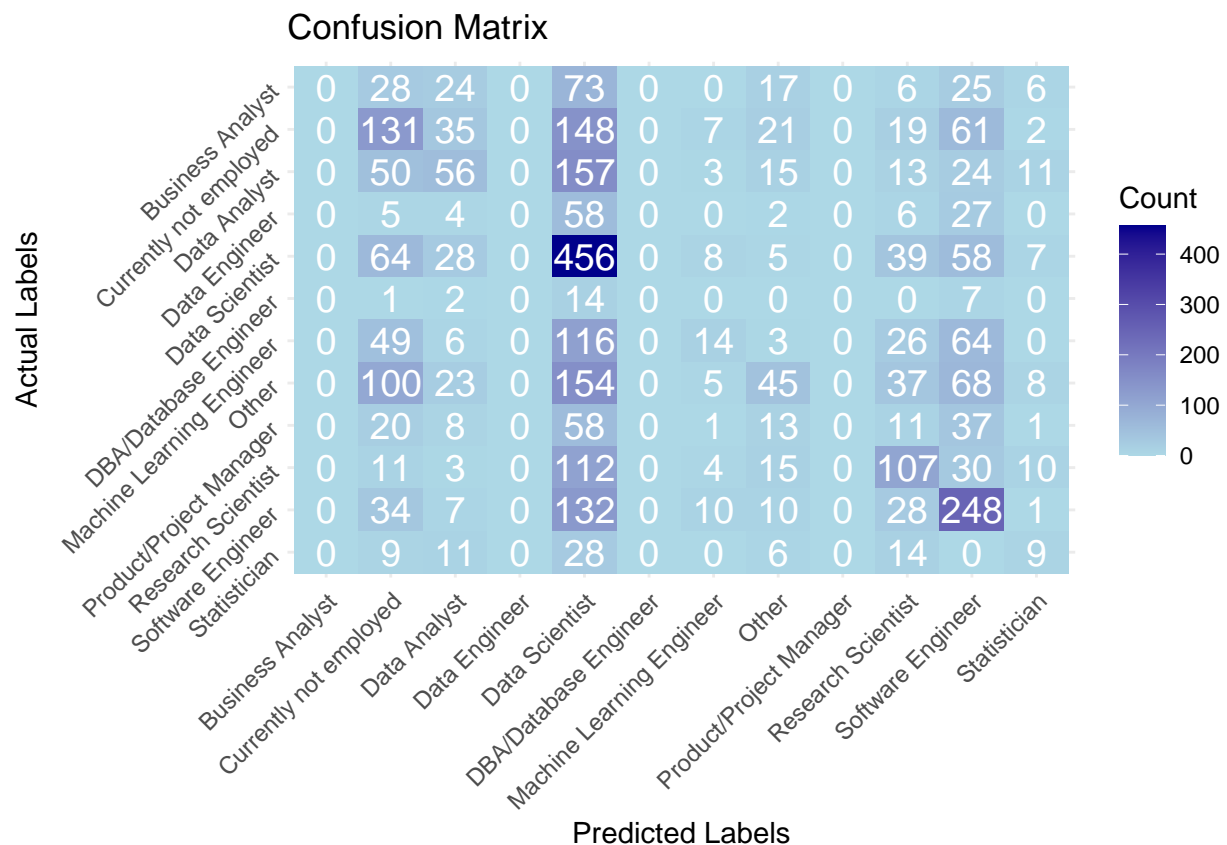
```
## 0.2485329 0.2836969
## sample estimates:
##      cor
## 0.2662035
```

Advanced Analysis) Predict Career Trajectory of Job-Seekers

Methods

```
## # weights: 120 (99 variable)
## initial value 25562.234706
## iter 10 value 22590.204459
## iter 20 value 21928.919274
## iter 30 value 21703.700713
## iter 40 value 21274.527005
## iter 50 value 21182.348170
## iter 60 value 21140.983871
## iter 70 value 21116.854172
## iter 80 value 21108.131055
## iter 90 value 21105.814120
## iter 100 value 21105.278022
## final value 21105.278022
## stopped after 100 iterations

## Accuracy: 31.09 %
```



Analysis

Conclusion

Conclusions and Discussion

Summary of Findings

The machine learning survey provided a lot of data about people and the ways they interact with machine learning, especially on Kaggle. One of the first things we noticed was the age-education relationship hidden within the data, which followed an interesting pattern that suggests machine learning is a higher education led field, as Master's degrees and PhDs typically became more and more popular as people aged, typically meaning that those who are involved in this field usually want higher education to be successful. Additionally, during our analysis of the experience-programming relationship, we noticed that Python, R, and SQL were the most widely used programming languages within coders, and surprisingly, maintains a very steady usage rate between all experience groups, even increasing as one gains more experience. This could potentially indicate the usefulness of Python even at the highest levels of data analysis and machine learning, which makes sense with the context we now know. In terms of the job distribution within countries, we noticed that the global job distribution tended to favor people from countries like India or the United States, which were especially skewed towards being students. This falls in line with most other countries, as the top 3 jobs in the top 10 countries typically were students, software engineers, data scientists, or analysts, which makes sense, as Kaggle is a platform that contains data primarily for these roles. Finally, during our analysis of hiring preferences for larger versus smaller companies, we noticed that while the distribution for age between large and small companies remained relatively similarly, the distribution of salary by company size was quite different. One thing that we noticed was that small companies tend to have most salaries peaking up in one area, while larger companies are able to afford a flatter curve with peaks at different places, indicating they have resources to pay higher skilled employees with higher wages.