315 final project

Introduction

This project explores a dataset of STEM salary records from top companies, such as FAANG, Oracle, Ebay, etc. The dataset contains 22429 data instances and contains total of 29 variables.

The quantitative variables of interest are: 1) totalyearlycompensation (total yearly compensation) 2) yearsofexperience (years of experience) 3) yearsatcompany (years at a company)

The categorical variables of interest are: 1) title 2) gender (male, female, or other) 3) race (as binary variables Race_Asian = 1 or 0, Race_White = 1 or 0, etc.)

The first few lines of data appear as follows:

```
##            timestamp company level             title totalyearlycompensation
## 1  6/7/2017 11:33:27  Oracle    L3   Product Manager                  127000
## 2 6/10/2017 17:11:29    eBay  SE 2 Software Engineer                  100000
##            location yearsofexperience yearsatcompany  tag basesalary
## 1  Redwood City, CA               1.5            1.5 <NA>     107000
## 2 San Francisco, CA               5.0            3.0 <NA>          0
##   stockgrantvalue bonus gender otherdetails cityid dmaid rowNumber
## 1           20000 10000   <NA>         <NA>   7392   807         1
## 2               0     0   <NA>         <NA>   7419   807         2
##   Masters_Degree Bachelors_Degree Doctorate_Degree Highschool Some_College
## 1              0                0                0          0            0
## 2              0                0                0          0            0
##   Race_Asian Race_White Race_Two_Or_More Race_Black Race_Hispanic Race
## 1          0          0                0          0             0 <NA>
## 2          0          0                0          0             0 <NA>
##   Education
## 1      <NA>
## 2      <NA>
```

This data was obtained from Kaggle. According to its description, the data were collected on a range of times from June 7th of 2017 to August 17th of 2021, and 14% of the job locations are in Seattle, WA, 11% in San Francisco, CA, and 75% being other cities in the United States.

As students hoping to pursue a career in STEM, we hope to better understand what qualities impact salary the most across the STEM field. Therefore, the goal for this project is to explore the following hypotheses:

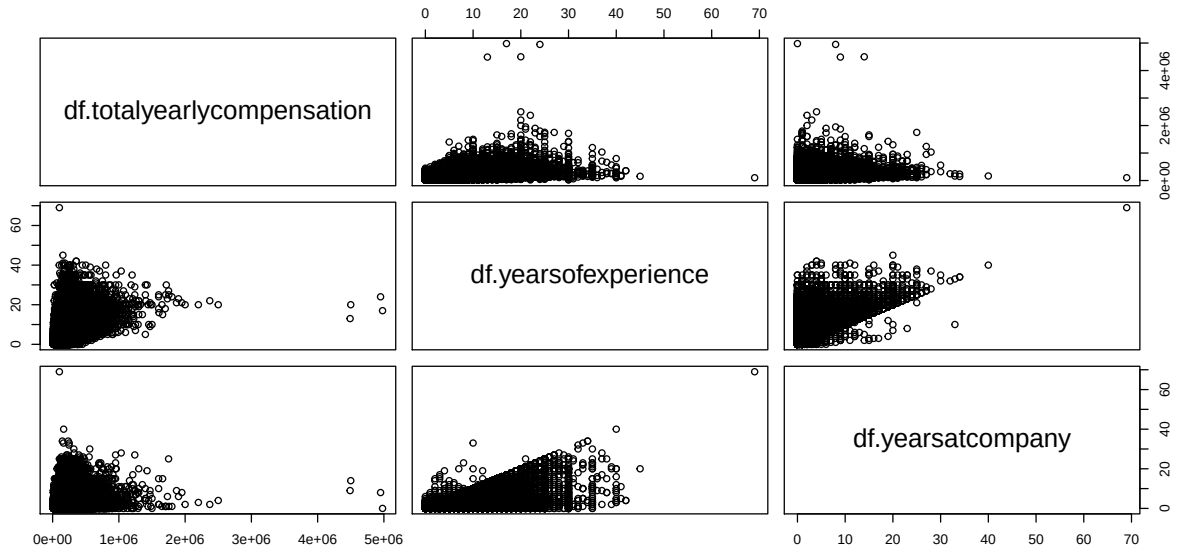
1) How does gender affect company and title in STEM?

2) How do education and experience affect compensation?

3) How does race affect salary and title in STEM?

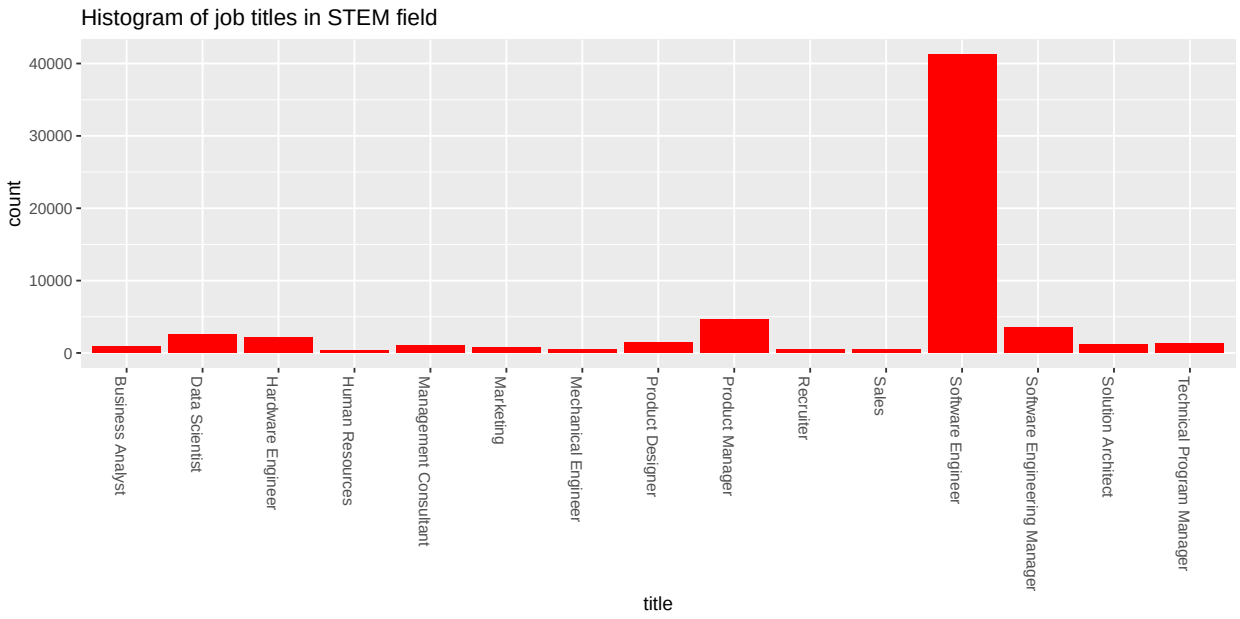
Exploratory Data Analysis

First, we want to explore the relationship between each pair of quantitative variables:

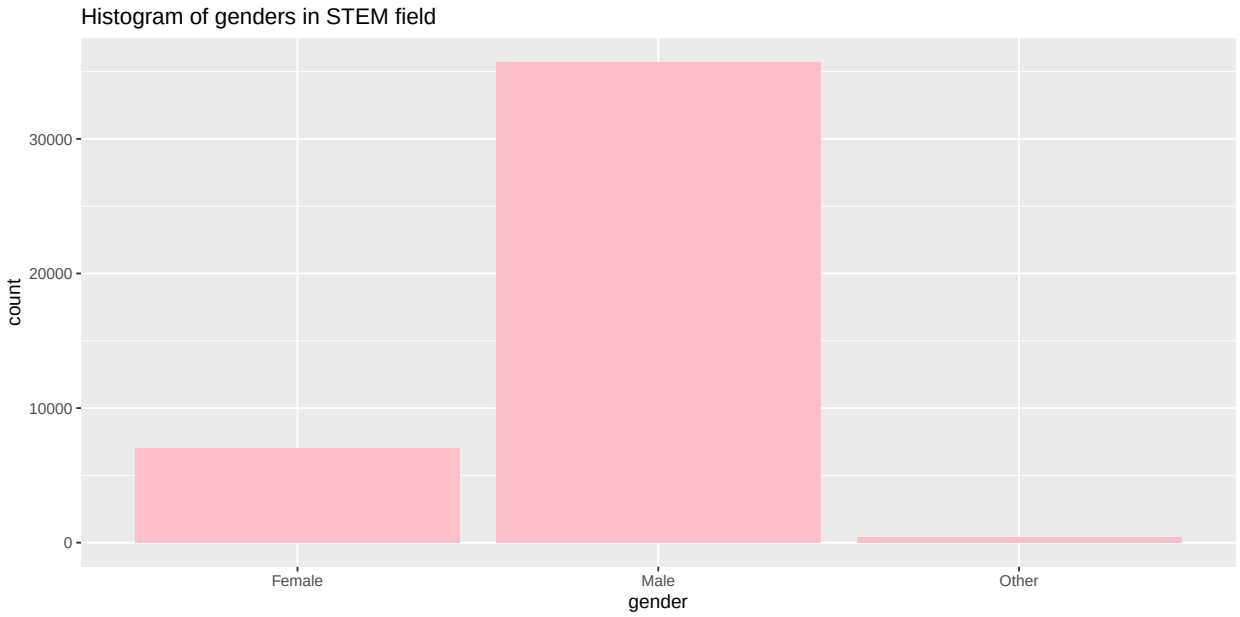


From the pairs plot, we observe interesting trends between years of experience and years at company: while there appears to be a linear correlation, many instances with ample years of experience seem to have very low years at company. However, this makes sense given the context of the data, if someone switch jobs their years at company would reset while years of experience continues to grow. We also observe that years of experience does not have a linear correlation with total yearly compensation as one might suspect. But this also makes sense because someone who has 30 years of industry experience is likely retired or close to retiring, therefore, may lean toward doing more casual, less paying but fun work rather than progressing their career and compensation.

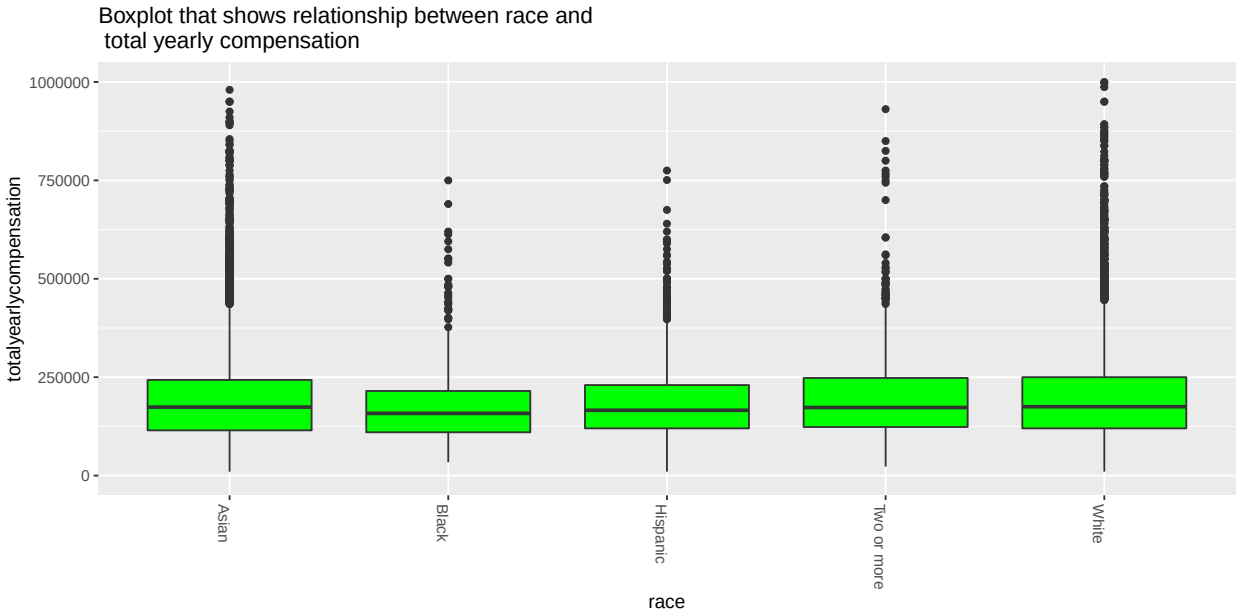
Next we want to explore our categorical variables using bar plots.



Here, we see that software engineer has significantly more count than all other titles, meaning it is the most popular title overall, followed by product manager, software engieer manager, and data scientist.



For the gender variable, we see that there are significantly more male than female, and other gender is virtually non-existent in the STEM work field of top companies.



Lastly, we take a look at race vs salary represented in total yearly compensation. We see that the median salary across all race are roughly the same, with black people being slightly lower than the rest, and Asian people having the highest median salary. The boxes overlap, meaning there is not a significant difference in total yearly compensation across races.

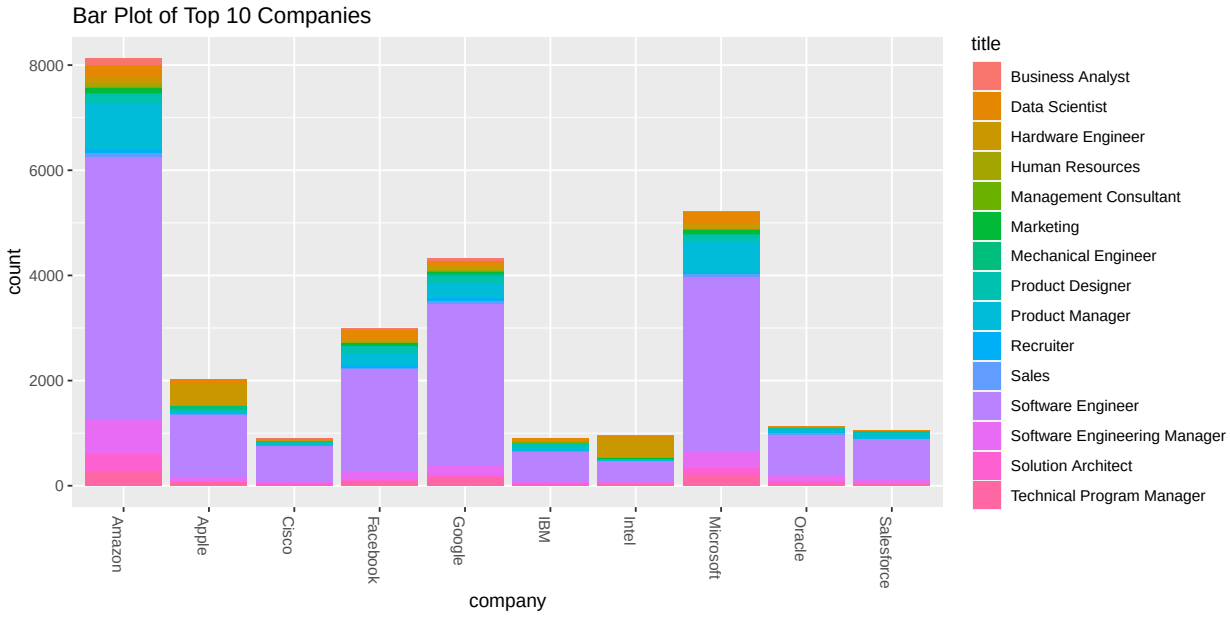
Hypothesis 1: How does gender affect salary and title in STEM?

A comparison word cloud was made using the subset of the data separated by male and female gender on the different companies present in the dataset. The comparison word cloud shows that Amazon hires more males than females by a significant amount and Microsoft hires more females than males. This concludes that there is a correlation between the gender of the applicants and the companies that hire them.

A mosaic plot of the different positions based on gender is made. The different colors show the difference (positive or negative) in the quantity of the specified gender and position compared to if gender and position were actually independent. The plot shows that the position with the largest count, Software Engineer, has a much larger number of males than is expected. Similarly, we can also conclude that the number of females is also much smaller than expected if gender and position were independent. Overall, we can see that females actually have more positions that are more than expected than males, but these positions are less STEM-intensive such as Recruiter. We also note that since most squares are colored, we can safely assume that gender and position are dependent. We will perform a chi-square test of proportions to make prove our assumption.

```
##
##  Pearson's Chi-squared test
##
## data:  table(data.mf$gender, data.mf$title)
## X-squared = 2779.6, df = 14, p-value < 2.2e-16
```

Since the p-value is less than 0.05, there is enough evidence to suggest that in the subset of data of male and female gender, gender and title (position) are dependent. However, we wish to determine if companies really look at gender or if the qualifications of applicants based on the different positions.



From this bar plot, we note that most of the companies actually recruit mostly for the software engineering positions. Since we have noted that gender and position are dependent, we cannot easily come with the conclusion that companies hire based on gender specifically. It could be the case that companies hire based on positions, and as a result companies unintentionally hire more males than females.

Hypothesis 2: How do education and experience affect compensation?

We wanted to learn about how education and experience affect compensation. The question suggests we should examine Bachelors_Degree, Masters_Degree, total yearly compensation, years of experience, years at company, base salary, stock grant value, bonus.

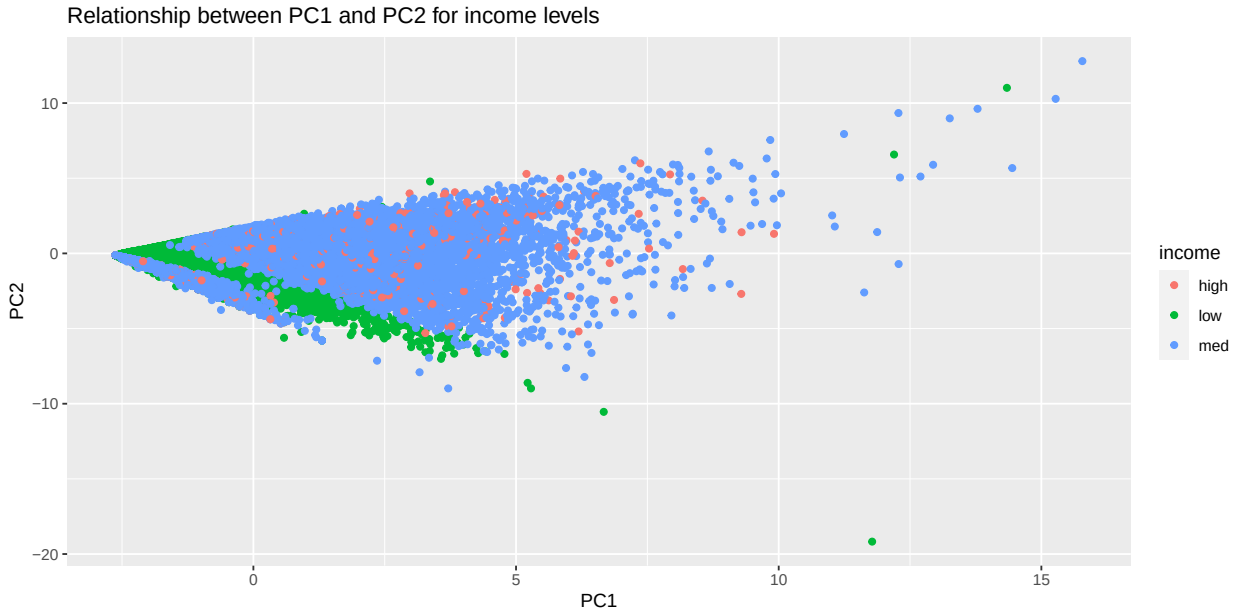
We first inspected the relationship between experience and compensation. To do so, we created a variable 'income' in a way such that it is -low if total yearly compensation is lower than 188000 -medium if the total yearly compensation is between 188,000 and 400,000 -high if the total yearly compensation is higher than 400,000.

For the threshold for low income group, I chose 188000 as it is the median value. Then, for the threshold value that distinguishes medium and high income, I contemplated whether I should go with 90th percentile or 95th percentile; thus, I chose on 400000,which is 92th percentile(nice in-between). As a result, we get following numbers for the three income groups

```
##
##  high   low   med
##  2432 31159 29051
```

Thus, we performed PCA on years of experience, years at company, base salary, stock grant value, bonus. We decided to use the four principal components, as they covered 90% of the variance in the data. As a result, we have the following visualization:

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.4801 1.0658 0.8705 0.70418 0.64797
## Proportion of Variance 0.4381 0.2272 0.1516 0.09918 0.08397
## Cumulative Proportion  0.4381 0.6653 0.8168 0.91603 1.00000
```



Out of all visualization combinations between four principal components, the clustering seemed most clear when using PC1 and PC2. It is difficult to distinguish between medium and high income groups, because there is huge overlap between the two groups. Medium income group seems to be generally between -2.5 and 10 for PC1 and -7 and 5 for PC2; high income group seems to be generally between -2.5 and 7 for PC1 and -5 and 2.5 for PC2.

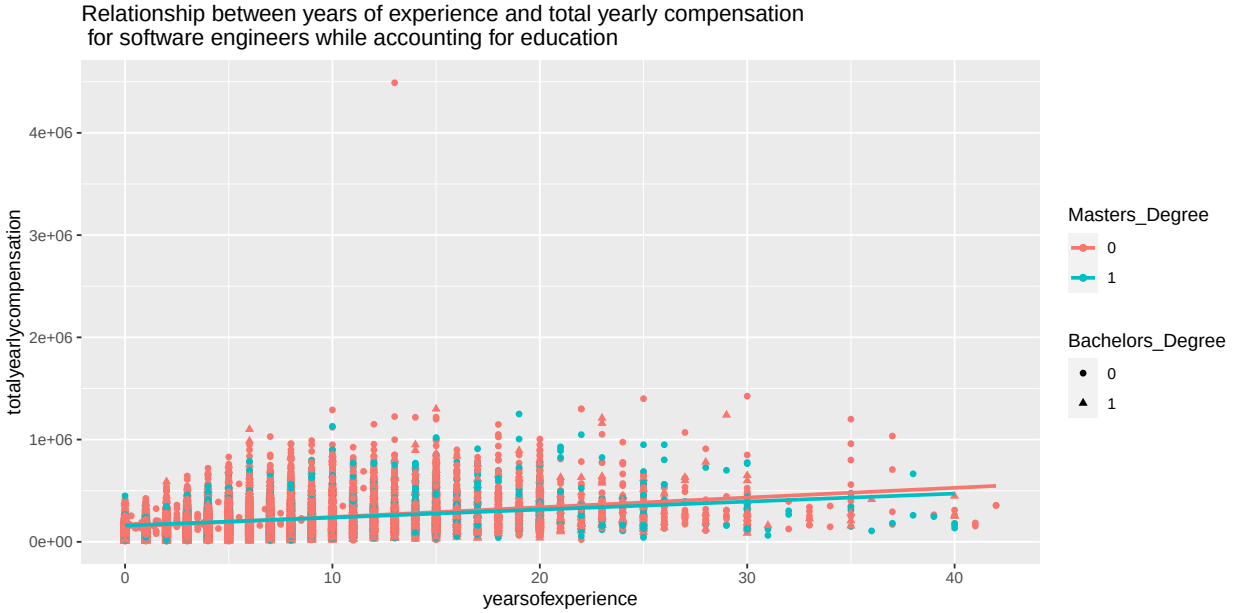
On th other hand, it is easy to distinguish low income group from the other two income groups, as low income group seems to be on the bottom left edge of the overall cluster. The low income group seems to be between -5 and 5 for PC1 and 0 and -7.5 for PC2.

```
##                              PC1        PC2
## yearsofexperience      0.5109546 -0.4024155
## yearsatcompany         0.3559895 -0.7025443
## basesalary             0.5093094  0.3287676
## stockgrantvalue        0.4355541  0.3758565
## bonus                  0.4038483  0.3084429
```

Basically, for low income group, as PC1 increases, PC2 value decreases. When we look at the the first two principal component's rotation matrix, we see that years of experience and years at company both decrease as PC1 increases and PC2 decreases.

Overall, this graph seems to suggest that low income group can clearly be clustered and distinguished from the other two clusters. This means there is certain pattern in the data regarding the variables, which were previously mentioned for creating principal components, that is unique/particular

for the low income group. Since all the variables used to create the two used principal components are relevant experience, this graph shows that greater the experience, more likely one is to have greater compensation. Then, in the graph below, we inspected education's relationship with compensation as the main interest and experience's relationship with compensation as the secondary interest.



It is believed in the STEM field that one needs to attain Masters Degree to earn higher compensation. We decided to analyse whether this is true for software engineers, as that is the most popular job title. However, interestingly, as you can see from the great overlap between the regression lines for with and without masters degree, there is no significant difference in total yearly compensation between those who have or do not have Masters Degree.

In addition, the graph above shows that majority of employees in STEM field do not have Masters Degree; in fact, many of them do not even have Bachelors Degree as well. The top four points of total yearly compensation correspond to no bachelor's degree. Therefore, the visualization suggests the belief of needing to attain Masters Degree to earn higher compensation is likely not true.

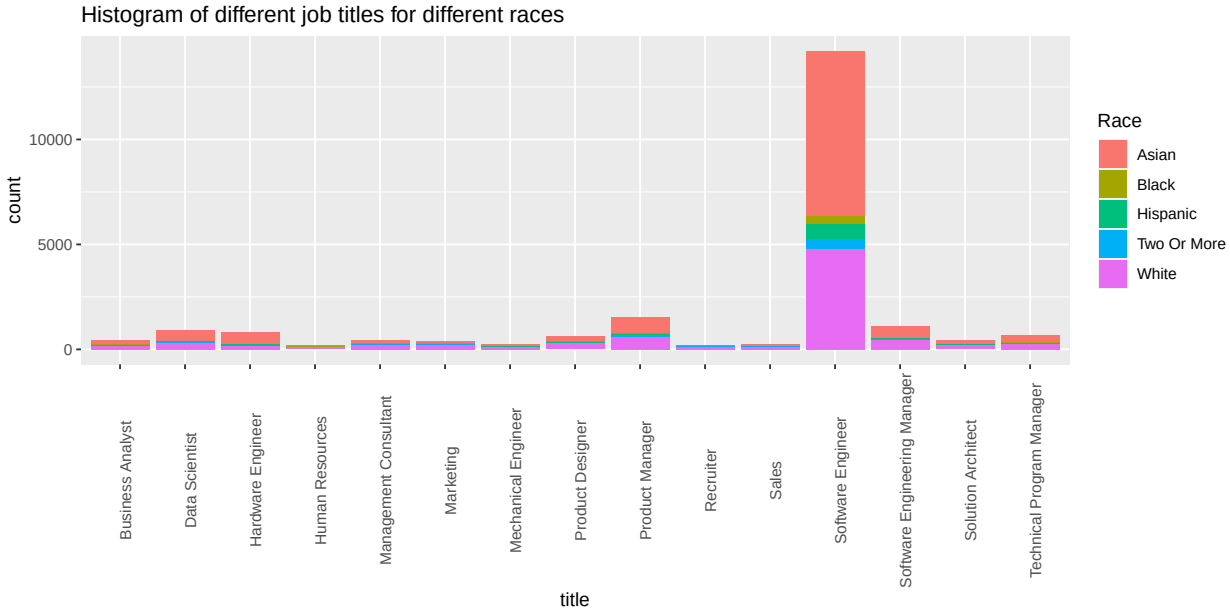
However, out of 15391 instances that have masters degree, there are only 4 instances with both bachelors and masters degree in the original data, and the rest do not have masters degree. While it is certainly possible There are many master's programs that don't always require a related bachelors; however, for stem field, this is very unlikely. It is possible that this was caused by the survey participants, who believed they did not need to click on 1 for Bachelor since they already clicked on 1 for Master's. Thus, we should definitely consider the limitation of this data and subsequently the reliability of this finding.

Also, it seems the starting salary range is from 10000 to 0.5e+06. As the years of experience increase, the range becomes wider. However, after 15 years, the general increase in total yearly compensation seems to stop. Therefore, the education doesn't seem to play a significant role in influencing the compensation, whereas the experience does. This substantiates our previous finding from the PCA

Hypothesis 3: How does race affect salary and title in STEM?

Before we explore this hypothesis, we must clean the race component of the dataset, as it has a lot of NA values for the Race variable. Thus, we make a NA free dataset called df.

First, let's explore the relation between race and title.

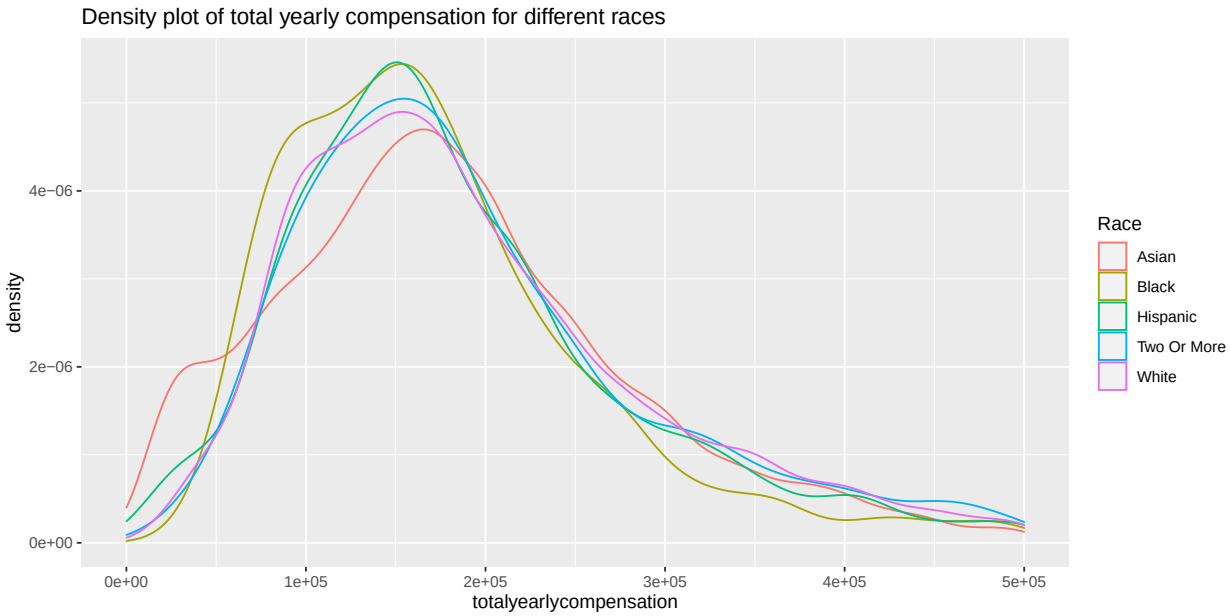


From our stacked bar plot, we can make some observations about race and title as a pair of variables. Firstly, it is clear from our dataset that the most frequent title is as a Software Engineer

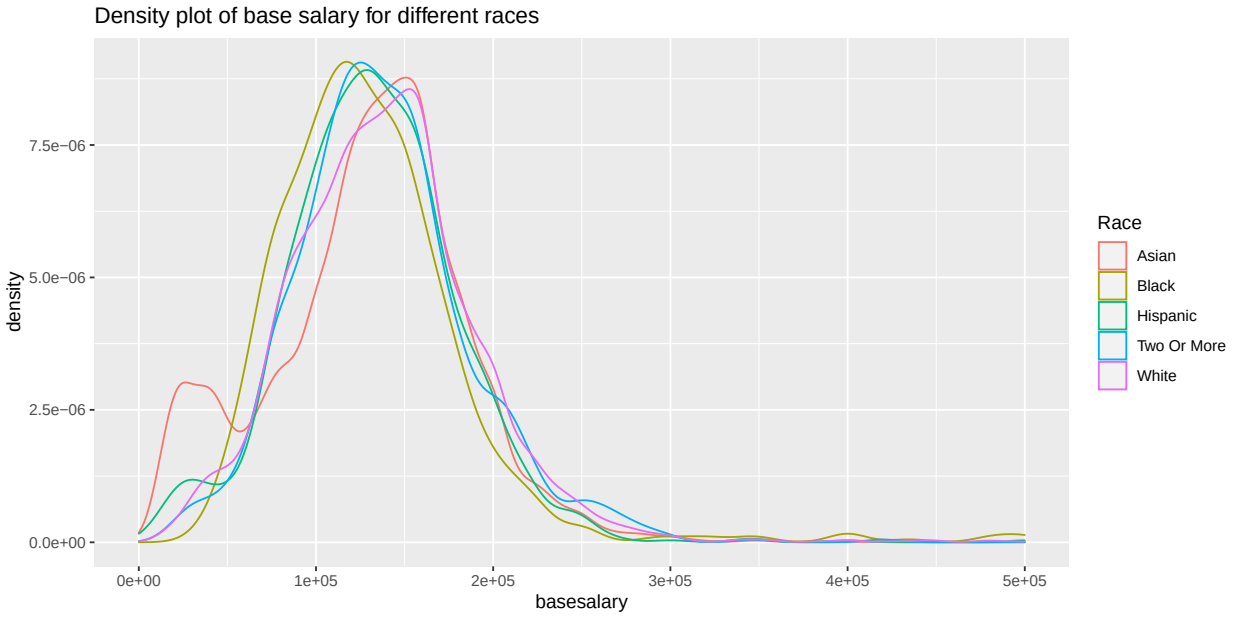
Further, we can see across all the bars that the two most frequently hired races for these titles in STEM are Asian, followed by White as the 2 clear denominators. Then Hispanic has a lot less, and Black and 2 or more are clearly the least. In fact across all the titles the most frequent races are White and Asian (we can even see the smaller bars are mostly orange and/or pink

Thus, we can conclude that there is somewhat a relation. If you are Asian/White, there is a much higer frequency of you in the industry, suggesting that race is more commonly hired, however we cannot make any conclusions on if race affects your specific title, as all titles have a similar distribution of mostly asian and white.

```
## Warning: Removed 613 rows containing non-finite values (stat_density).
```



```
## Warning: Removed 34 rows containing non-finite values (stat_density).
```



Our second graph is a conditional density plot of the total yearly compensation, colored by title. As we can see, all the lines are very similar on the plot, with no clear differences, thus we can say that your race should not affect your total yearly compensation

Going one step further however, you can look into the base salary as well, and here we see that lines colored for Asian and White Races are a bit more to the right, implying that their base salary on average is a bit more than that of the other races, which could be because they are hired more primarily for the Software Engineering and other high paying jobs.

Conclusion

In conclusion, our answers to the 3 research questions are:

1) How does gender affect company and title in STEM? Assuming that this data was gathered at random from different STEM positions from across the country, it is evident from the several graphs and chi-square tests that gender has an effect on the title and company. We find that there are more males than females in the position with the largest number, but there are more females in positions that are not very STEM-intensive such as the recruiter position. We also note that there are certain companies that hire more males than females in general such as Amazon, but at the same time there are companies that hire more females than males such as Microsoft. Overall, for software engineering positions, males are more commonly hired compared to females.

2) How do education and experience affect compensation? It is commonly believed that one must come from prestigious institutions and gain a lot of experience to attain high salary. However, it seems that education does not necessarily play a great role in influencing one's salary, at least for software engineers. It is worth noting that this finding regarding education should be taken with caution and doubt, as the data had extremely few points that contained both Bachelor's and Master's; this is most likely due to human error caused by miscommunication and misassumption in the data collection process. On the other hand, ample experience seem to correlate to higher salary. However, after 15 years, the general increase in compensation seems to gradually stop and plateau.

3) How does race affect salary and title in STEM? It is clear that those with Asian and White backgrounds are hired a lot more frequently in the STEM marketplace, especially in technical fields like Software Engineering, but really across all the titles we've looked at. As a result, they also are payed a little

more on average, and that is probably because they are simply hired for the more technical (and so higher paying) positions. But this difference is slight. That said, your race does not directly have a relationship to a specific title. The trend of Asian and White being the main races is for all titles, and so it is an industry wide feature. Therefore, this suggests that if you are not Asian and White, you may find yourself a minority in the STEM work field. Thus, YES: Race does have an affect on salary and title in STEM. Overall, if you are White or Asian, you are more commonly hired as compared to the other races, and such benefits reflect a bit on the salary and type of titles as well.

Overall, these 3 hypothesis overlap to provide an interesting narrative and some take-away points. We can say that the STEM industry is still dominated by Men of Asian and White backgrounds, as they are the most heavily employed in the industry, especially the more technical/skilled positions. A factor you can control however in the industry is your education and years of experience, and while the former didn't have grounded conclusions, it was clear that the more years of experience in a field will translate to more pay for the most parts

To wrap up the report, it is important we disucss some limitations and the future potential from our report. Firstly, we saw a big flaw in the data collection for the education data, which is that there was no 'Bachelors and Masters' option, so people probably picked their highest level, thus does with Masters were seen as not having a Bachelors in some cases. Also, another limitation is that this data collected seems to be heavily for Software Engineering, with the data points clustering heavily for that tile. This would skew our results to reflect more on race, gender and education trends in software jobs than in the industry as a whole. Further, this data needed a lot of cleaning, with a lot of NAs and weird inputs (like some job titles falling under gender categories), so there was clearly some imperfections in the data collection. One final limitation we saw was in the gender for the companys, with companies like Microsoft looking like they hired a relatively higher proportion of females but it could just be that their data was more for administrative jobs which are currently female dominated. We looked into it a bit, but more exploration could yield some more specific title/company results.

Speaking of which, some more future potential would be to look at this across the variable of time too, to see how these trends are changing and what direction they are going. We can also look into some more variables, as ours were around identity (gender, race and education), but other factors like what city they apply into could be interesting (like Amazon SA vs Amazon CA) to explore. As STEM Majors in a STEM heavy school, this data is full of interesting possibilities to expand on from this report.