Title: Stem Salaries
Group Name: Women Who Code
Group Members:

- Khevna Parikh (kp2936)
- Doma Ghale (dg2491)
- Priya Dhond (pnd220)
- Huyen Nguyen (hn2231)
- Jennifer Trujillo (jr5951)

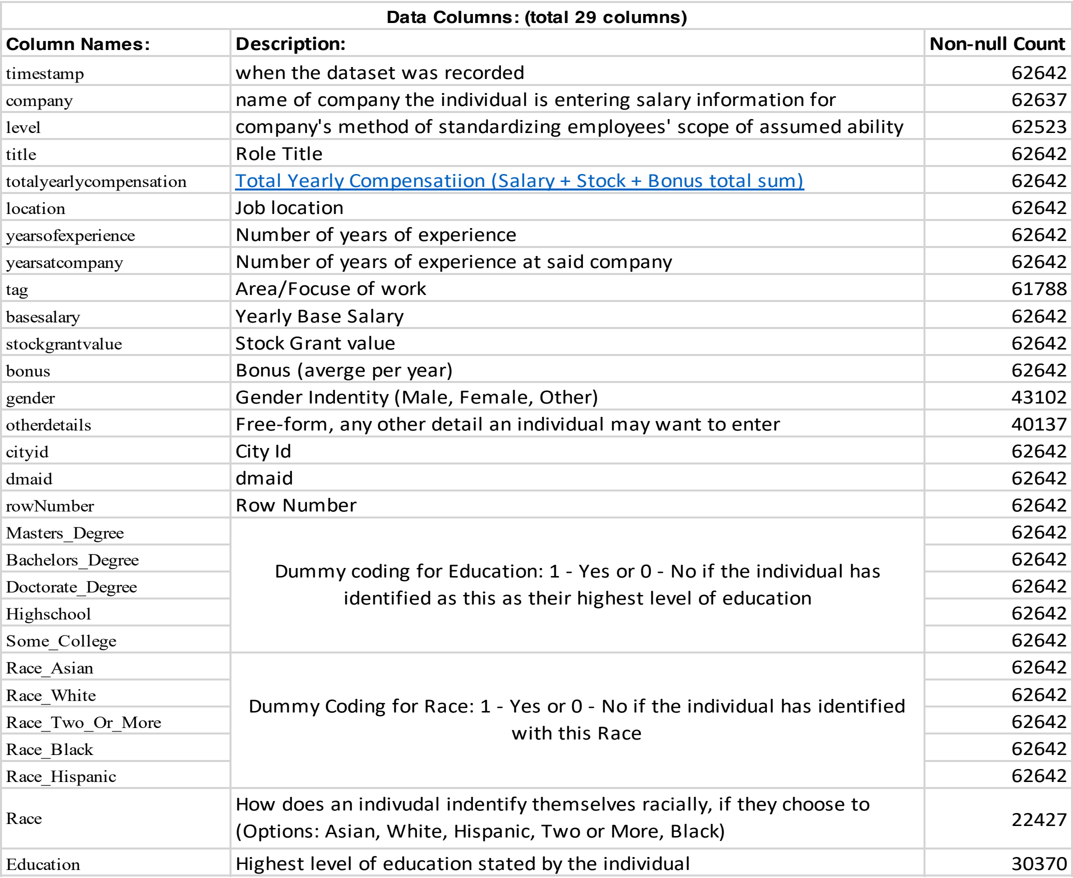
Introduction:

Using hypothesis testing, regression, and various machine learning tools, this project aims to examine the salaries of individuals in the Science, Technology, Engineering, and Mathematics (STEM) fields. All necessary information was obtained from the Kaggle website in a single csv data file, which includes salary records and other personal attributes of 62,642 individuals. This dataset is initially originated from the site level.fyi, where individuals can self-provide salary information to help others make better career decisions.

This dataset was particularly intriguing to us not only as students who are currently working towards a master's degree in Data Science, but also as students with aspirations to enter the workforce upon graduation. First, the significance in correlation between all our variables was carefully analyzed. From which, certain variables were further examined using regression and hypothesis testing. Furthermore, Principal Component Analysis was performed to reduce the dimensionality of the dataset into fewer components. Lastly, clustering techniques were used to determine the optimal number of groups the dataset could be divided into.

Structure of the dataset:

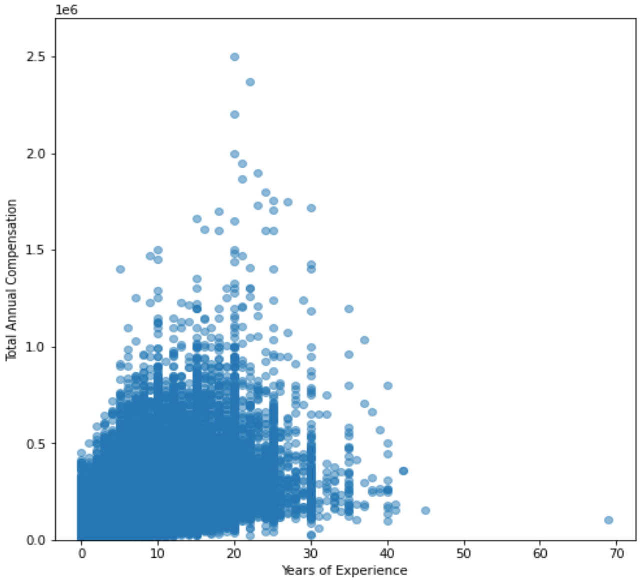
This data set contains 29 columns, specifically:



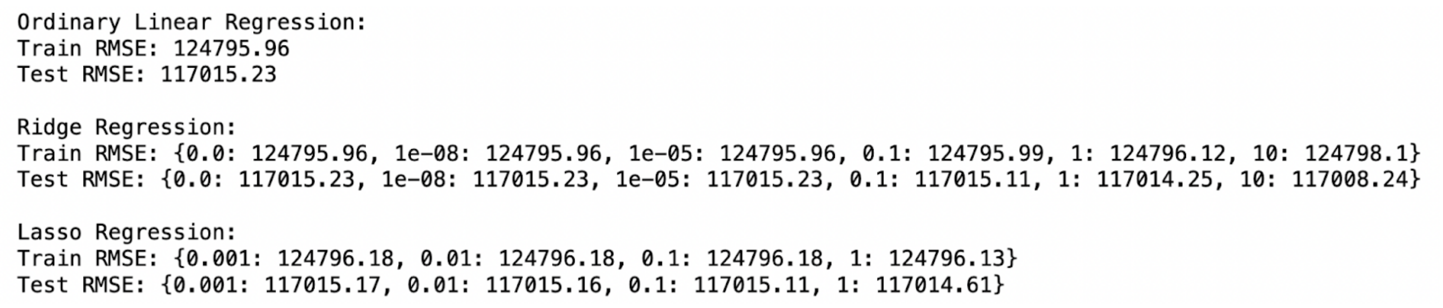
Data Columns: (total 29 columns)		
Column Names:	**Description:**	**Non-null Count**
timestamp	when the dataset was recorded	62642
company	name of company the individual is entering salary information for	62637
level	company's method of standardizing employees' scope of assumed ability	62523
title	Role Title	62642
totalyearlycompensation	Total Yearly Compensatiion (Salary + Stock + Bonus total sum)	62642
location	Job location	62642
yearsofexperience	Number of years of experience	62642
yearsatcompany	Number of years of experience at said company	62642
tag	Area/Focuse of work	61788
basesalary	Yearly Base Salary	62642
stockgrantvalue	Stock Grant value	62642
bonus	Bonus (averge per year)	62642
gender	Gender Indentity (Male, Female, Other)	43102
otherdetails	Free-form, any other detail an individual may want to enter	40137
cityid	City Id	62642
dmaid	dmaid	62642
rowNumber	Row Number	62642
Masters_Degree	Dummy coding for Education: 1 - Yes or 0 - No if the individual has identified as this as their highest level of education	62642
Bachelors_Degree		62642
Doctorate_Degree		62642
Highschool		62642
Some_College		62642
Race_Asian	Dummy Coding for Race: 1 - Yes or 0 - No if the individual has identified with this Race	62642
Race_White		62642
Race_Two_Or_More		62642
Race_Black		62642
Race_Hispanic		62642
Race	How does an indivudal indentify themselves racially, if they choose to (Options: Asian, White, Hispanic, Two or More, Black)	22427
Education	Highest level of education stated by the individual	30370

Correlation and Regression:

To identify plausible relationships between our variables, we first computed a correlation matrix between the following variables in our dataset: total yearly compensation, years of experience, years at company, base salary, stock grant value, bonus, education level (master’s degree, bachelor's degree, Doctorate degree, high school, some college), and race (Asian, White, two or more, Black, Hispanic). For the selected categorical data (education, gender and race) we replace nans with "unknown" so that we can include them in our analysis. In order to compute the correlation matrix, we need to convert the categorical data into one hot encoding. However, some of the categorical data has too many distinct values, including company (1631), level (2923) and location (1050). Consequently, converting them to numerical values would result in an absurd number of columns. Hence, we removed those along with irrelevant columns such as timestamp, tag and rowNumber.



As shown in Figure 1 (Appendix), most correlation coefficients are relatively low. Furthermore, most correlation coefficients are uncorrelated with the exception of the correlation between **total yearly compensation** which is highly correlated with **stock grant value** (0.77), **base salary** (0.67) and **bonus** (0.49). This result is not surprising since total yearly compensation is the sum of the three variables. It should be noted that the sum of these three elements do not always add up to the total yearly compensation column. The form in Figure 2 shows that an individual can choose not to report the breakdown of total yearly compensation in terms of stock grant value, base salary and bonus. In addition to this, years of experience and years at companies have a correlation of 0.52. This is plausible, given there exists a relationship between these two variables. The correlation between total yearly compensation and years of experience was 0.40– plot shown to the right. This reasonable outcome implies that the more experienced an individual is, the higher that individual will be paid. The longer an individual stays at a company, the more years of experience increases alongside. However, it is only moderately correlated (0.52) possibly due to promotion or job title changes within a company. Education unknown and race unknown were also highly correlated (0.72), which tells us that individuals who did not report their education might have also not reported their race. Next, we determined the best-fitting regression model for our dependent variable, total yearly compensation based on years of experience, type of highest-level of education reported (Master’s, Bachelor’s, High School, Some College) and self-reported race (Asian, White, Two or More, Black and Hispanic). To do this, we build the regression models (Ordinary Linear Regression, Ridge Regression and Lasso Regression) by training on 80% of the data and testing on the remaining 20%. Lastly, we used Root Mean Squared Error (RMSE) to check our model performance and found the results below.



```
Ordinary Linear Regression:
Train RMSE: 124795.96
Test RMSE: 117015.23

Ridge Regression:
Train RMSE: {0.0: 124795.96, 1e-08: 124795.96, 1e-05: 124795.96, 0.1: 124795.99, 1: 124796.12, 10: 124798.1}
Test RMSE: {0.0: 117015.23, 1e-08: 117015.23, 1e-05: 117015.23, 0.1: 117015.11, 1: 117014.25, 10: 117008.24}

Lasso Regression:
Train RMSE: {0.001: 124796.18, 0.01: 124796.18, 0.1: 124796.18, 1: 124796.13}
Test RMSE: {0.001: 117015.17, 0.01: 117015.16, 0.1: 117015.11, 1: 117014.61}
```

Given the huge differences between the total yearly compensation and other variables, these numbers are expected. The coefficient of determination for the model is 0.23 and the R^2 score for training and testing are 0.20 and 0.22, respectively. They indicate that the model that we created does not fit the data very well. This could potentially mean that it is difficult to predict salary given the parameters we used, as salaries of individuals can vary drastically from individual to individual. We have observed there does not exist a strong correlation between total compensation and other variables, so this result aligns with the previous findings.

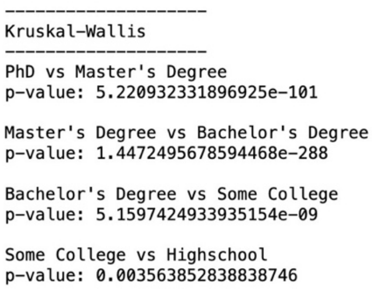
Hypothesis Testing

Note that, we initially omitted gender in our regression analysis since it pre-existed in our data as a categorical variable (i.e., participants entered "Male" or "Female", rather than as a numerical response). To examine this variable further, we conducted several hypothesis tests to specifically address the following questions: a) Do females get paid differently than men? b) Is there a relationship between participants' gender and their job titles? c) Is there a difference in salary between participants with a bachelor's degree and participants with only some college experience?

Our first hypothesis test was to see if there is a difference between male and female salaries. To do this, we iterated through our data-frame and checked whether each user identified as male or female had any missing salary and found none. We created two different lists– one with the salaries of females and one with the salaries of males– and conducted a two-sided Mann-Whitney U test to test our null hypothesis that the distribution of underlying females is the same as the distribution of underlying male. In this hypothesis test, we found a p-value of 1.3564129055940676e-14. From this p-value, we conclude to reject the null hypothesis. Furthermore, there is evidence of a statistically significant difference between the salaries of males and females, as seen in Figure 3. However, note that the female to male ratio in our data is 1:4.

The second hypothesis test was conducted to see the difference between gender and job titles. To start off, the original dataset was first extracted with two categories of interest: gender and title. Note that while initially exploring the data, one record in the gender column was found to have the value 'Title: Senior Software Engineer', which we had replaced with nan. We then replaced all the nans with "Unknown" because there was a significant number of nans in the "gender" column (31%). We split the job titles data by gender into 4 separate groups: "female", "male", "other", and "unknown". Since both our variables are categorical, a Chi-square test of independence of variables in a contingency table was performed. Given that all of our observed and expected frequencies in each cell were above five, this test was valid to use since. In this hypothesis test, we found our Chi-square statistic to be 2964.439088370925, with a corresponding p-value of 0.0 and 42 degrees of freedom. Therefore, we can reject the null hypothesis as the p-value is less than 0.05. We conclude there is evidence of a statistically significant difference between gender and job titles.

In our last hypothesis test, we look to see if there is a difference in salary and various levels of education. Specifically, we compared salaries between individuals with a PhD vs. Master's Degree, Master's Degree vs Bachelor's Degree, Bachelor's Degree vs Some College and Some College vs Highschool. In this hypothesis, we first used a Kruskal Wallis test to best see if the median salary of all of all our groups (different levels of education) are equal. We found the following p-values as shown in the figure to the right. Furthermore, we also conducted a one-way ANOVA to test if the mean salary of all our groups (different levels of education) are equal. Our test-statistics and p-values are shown in Figure 4. We found all groups compared to have p-values less than 0.05. Hence, all the p-values suffice to reject the null hypothesis and conclude there is evidence of a statistically significant difference between salary and level of education.

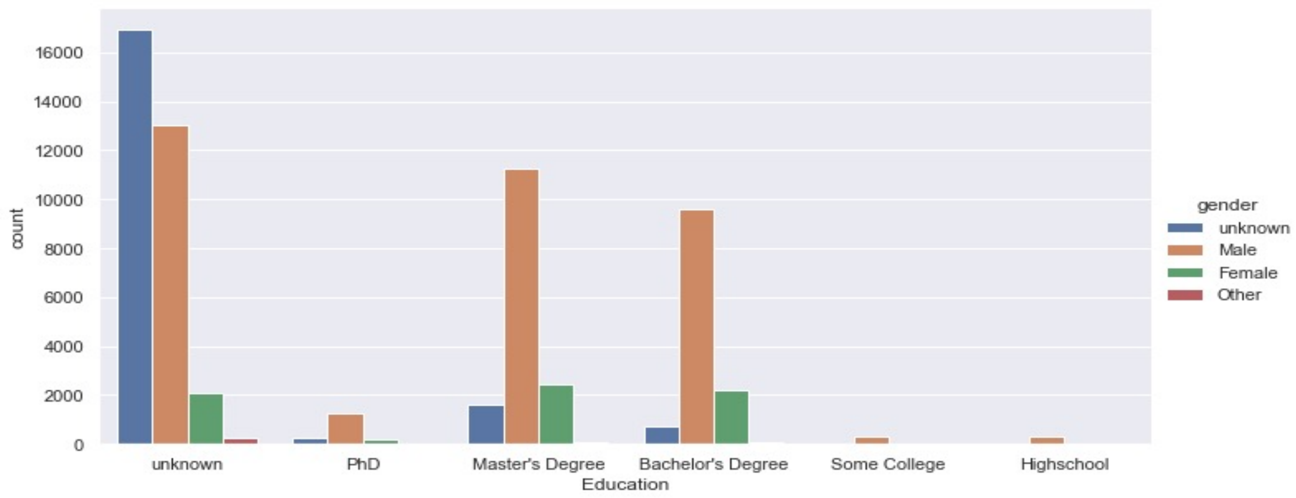


```
--------------------
Kruskal-Wallis
--------------------
PhD vs Master's Degree
p-value: 5.220932331896925e-101

Master's Degree vs Bachelor's Degree
p-value: 1.4472495678594468e-288

Bachelor's Degree vs Some College
p-value: 5.1597424933935154e-09

Some College vs Highschool
p-value: 0.003563852838838746
```

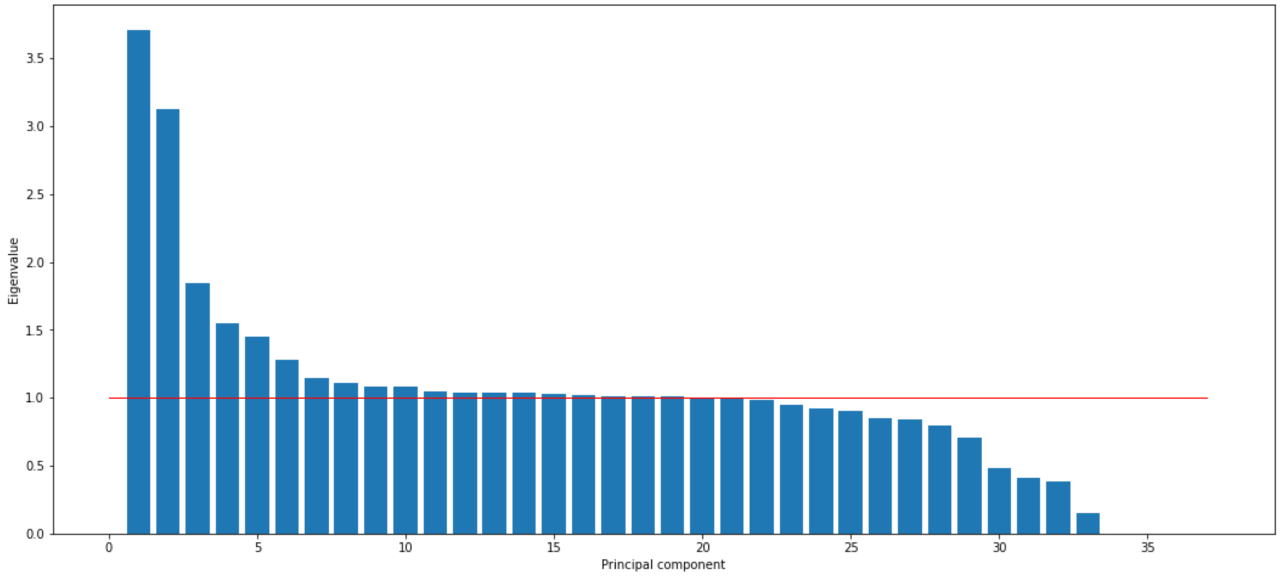


Unsupervised Learning:

Undeniably, smaller datasets are easier to visualize and computationally simpler to work with, especially when it comes to other machine learning techniques. We now perform a Principal Component Analysis (PCA) to see if we can reduce the number of variables in our data, while preserving as much information as we can. We do this by finding directions of maximum variance. Our aim is to find components that are linear combinations of the original one, are uncorrelated with one another, and orthogonal to each other.

Note that some of the categorical data contains too many distinct values such as company (1631), level (2923) and location (1050). This leads us to having over 2,000 eigenvalues greater than 1 while applying one hot encoding to our categorical data. Additionally, the pca.explained_variance_ratio is very, very low. For reference, the highest pca.explained_variance_ratio was 0.0008. So, we decided to take these columns out. We removed those along with irrelevant columns such as timestamp, tag and rowNumber. Since we already have one hot encoding of race and education, we removed those columns as well. Lastly, we do apply one hot encoding to convert variables such as title and gender, leaving us with a dataset consisting of 37 columns and 62,642 rows.

The PCA consisted of 5 key steps which we will explore in detail one by one. First, we standardized our data to avoid large differences in data to be merely caused by differences in scale, so we computed the z-scores such that each variable has mean of zero and a standard deviation of one. Next, we computed the covariance matrix of the standardized data. Variables that are highly correlated with each other contain redundant information and



therefore are unnecessary. From this, we conduct the eigen-decomposition of our covariance matrix, obtaining our eigenvalues and eigenvectors. Luckily, Python's scikit-learn library offers built-in functions that make these computations much easier. As discussed before, eigenvalues represent the magnitude of the eigenvectors. Eigenvectors give us the direction of maximum variance. The largest eigenvalue gives us the first principal component in the direction of the eigenvector. In order to determine the number of significant factors, or principal components, the dataset should be reduced to, we produced a scree plot as shown below.

There were 20 eigenvalues greater than 1, so the Kaiser criterion was not feasible for us. Also, it would take 26 eigenvalues for 90 percent of the variance to be explained. In the end, we settled for the "elbow" method, picking the left of the biggest/sharpest drop. This yields 2 principal components. Lastly, we recast our data using those two principal components. And, we used the transformed data to find clusters among different gender categories, and surprisingly they varied. Using the silhouette method, we obtained the following optimal number of clusters: 3 for all gender and female, 10 for male and other, and 2 for unknown. The silhouette scores and clusters are shown in Figure 5 and 6.

Conclusion:

Through our analysis of STEM salaries, we came across a few surprising and few not so surprising conclusions. From examining the correlation between total yearly compensation and other variables, we did not find a single factor that determines all of how much we get paid. The moderate correlation between compensation and years of experience suggests that to expect a higher salary in STEM careers, one must be more experienced. Furthermore, in our attempt to regress total yearly compensation on variables like education level and race, we discovered a low coefficient of determination for the model, suggesting that a linear model is not the most appropriate for this data. Through hypothesis testing, we discovered that there is evidence of a statistically significant difference in salaries between males and females. Moreover, we found evidence of a statistically significant difference in the job titles held by males and females, as well as a difference in the salaries of individuals with different levels of education. Finally, through PCA, we selected the first two principal components using the elbow method, and transformed our original data in the dimension space of those two components. We then performed kMeans clustering and used the silhouette method to find the optimal number of clusters for all genders, and further examined for different categories of gender and found that the optimal number of clusters varies as such: both all genders and females have 3 clusters, while male and others have 10 clusters and unknown has 2 clusters. This is supported by our previous finding that there are significant differences between male and female salaries and their job titles.

While our analysis was done with the 44,000+ data points that were available to us, our results do come with certain limitations. As we reflect back to the count of our data, we can see our sample sizes for each group are drastically skewed towards males. This outcome can be explained through gender norms associated with the workforce, where there is a statistical significance difference between STEM jobs and gender. There were three times as many males in the dataset as females, thereby skewing the average salaries for both genders. We observed that the median salaries for both genders did not differ significantly in this dataset, however, we wonder if we had more samples, it would give us more clarity about whether the observed ratio between males and females in the sample were representative of the STEM population. On the other hand, we have an unexplained group of individuals that either did not report their gender or classified as "Other'. We have insufficient information provided to determine whether the "Other" group consists of gender non-conformists, non-binary, or potentially other identifications within the LGBTQIA+. This lack of transparency led to the final decision of excluding them from this hypothesis.

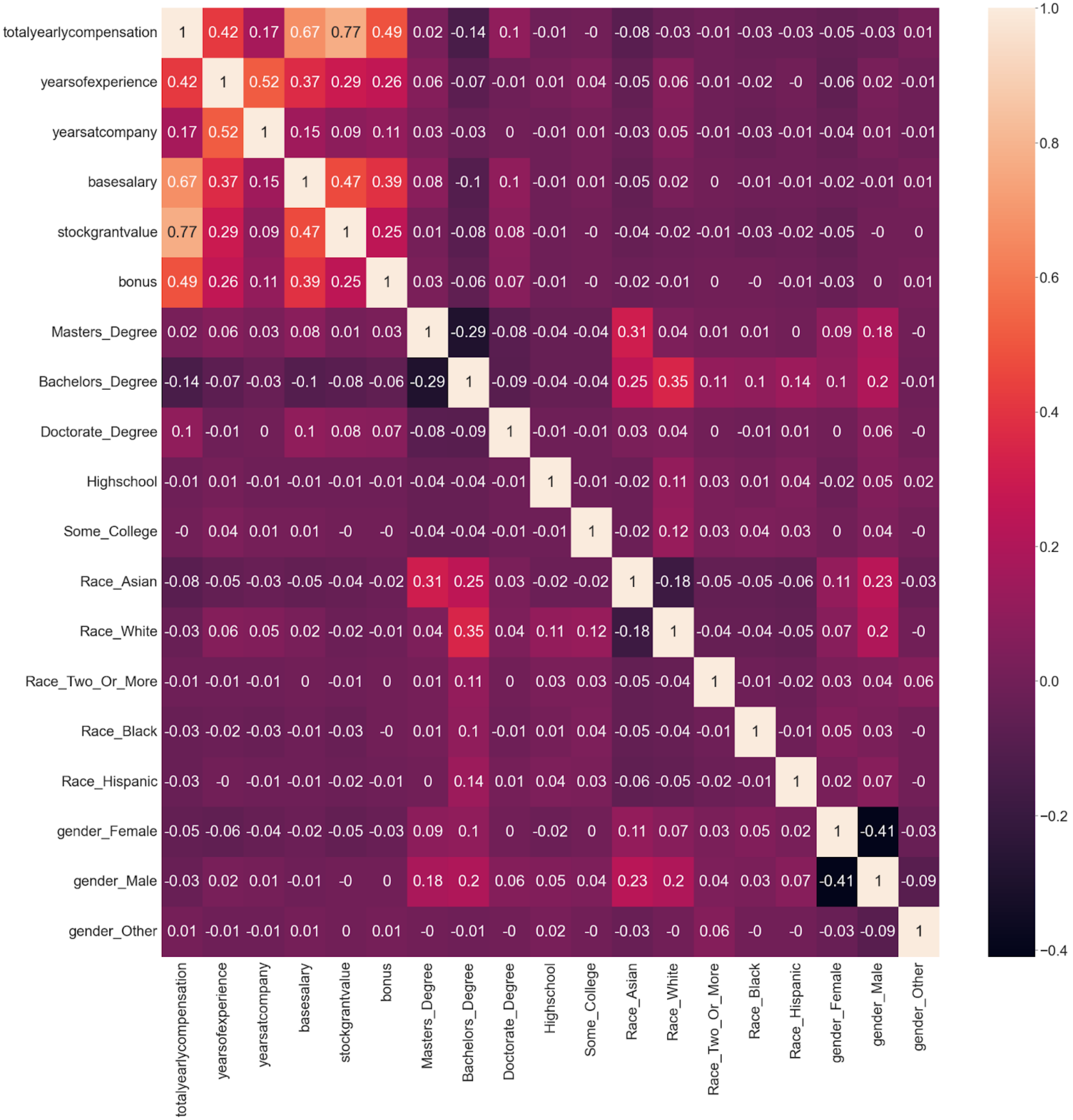
Furthermore, if more data was available, we would have liked to explore the effects of location on compensation. For instance, a person in California is likely paid more than an individual in Texas due to the costof-living adjustments. Through more data and analysis, we would like to analyze if such differences in salaries exist between locations, and if so, if there is a pattern in the salary differences between locations. Another aspect that we would like to explore with more data is the effect of race on job titles and salaries, specifically, we would be interested to learn if there are certain racial groups that “dominate” certain job titles. Our questions could be tested with more hypothesis testing as well as non-regression models. Finally, through our analysis, we noticed that our data is from 2017-2018. With more data from 2020 to present, we would be eager to explore a possible effect on job titles and salaries due to the ongoing COVID-19 pandemic. Moreover, in the last 2 years, the field of data science specifically has grown, and we would be interested to see if this has had an effect on the most popular job titles in the last few years.

Overall, through our analysis of STEM salaries, we developed a deeper understanding of the factors that contribute to salaries of individuals in STEM fields. As we continue our studies at NYU, we are eager to use this information to help guide us as we enter the job market in the upcoming years.

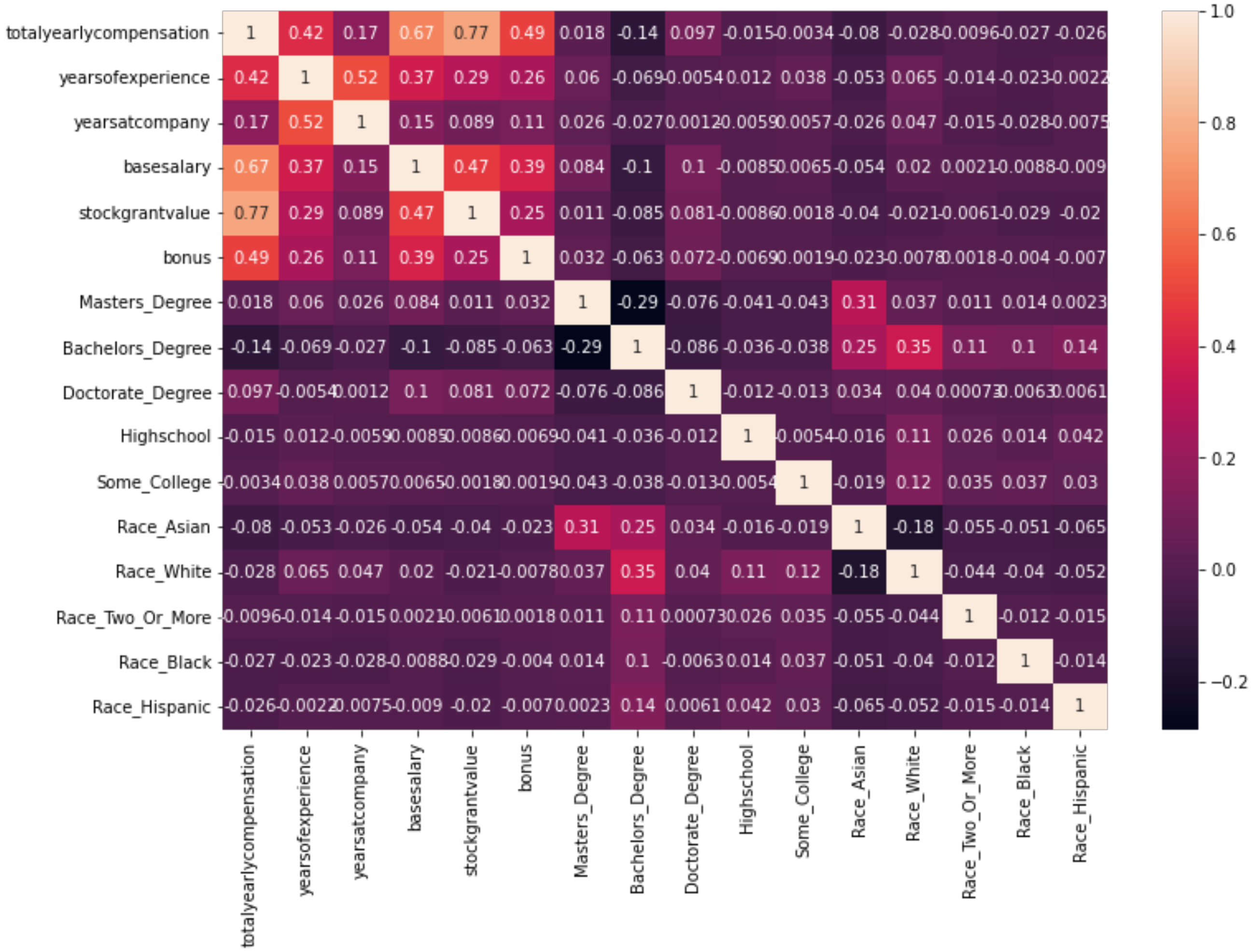
Please note that all Figures mentioned are posted in the Appendix on the next page.

Appendix:

Figure 1: Correlation Matrix:

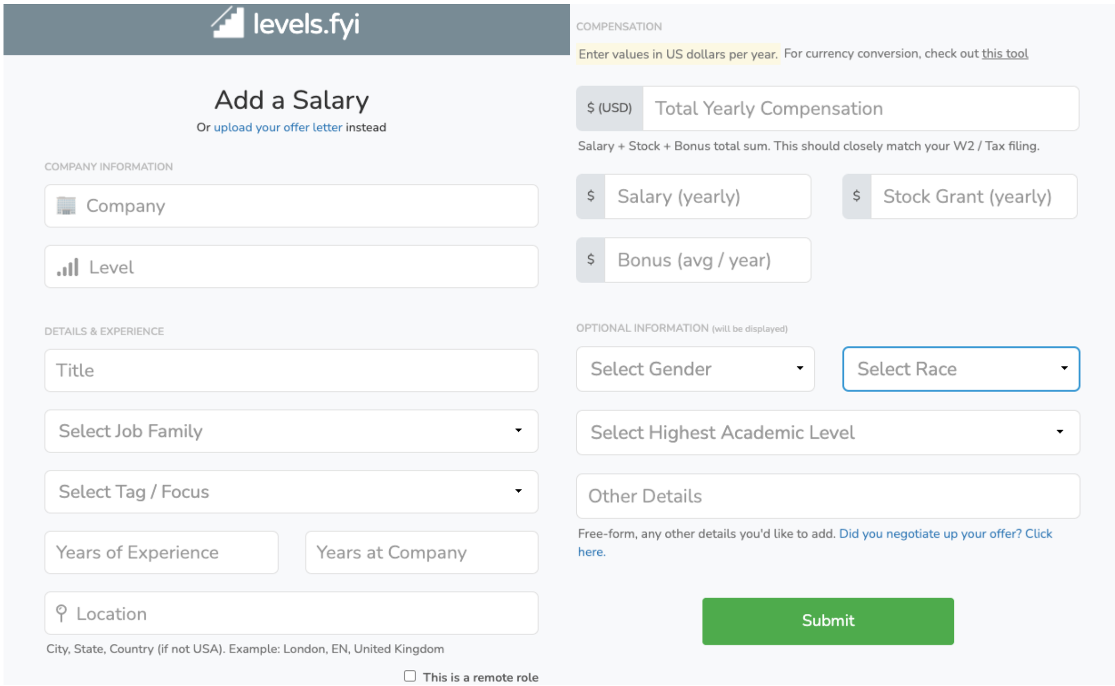


	totalyearlycompensation	yearsofexperience	yearsatcompany	basesalary	stockgrantvalue	bonus	Masters_Degree	Bachelors_Degree	Doctorate_Degree	Highschool	Some_College	Race_Asian	Race_White	Race_Two_Or_More	Race_Black	Race_Hispanic	gender_Female	gender_Male	gender_Other
totalyearlycompensation	1	0.42	0.17	0.67	0.77	0.49	0.02	-0.14	0.1	-0.01	-0	-0.08	-0.03	-0.01	-0.03	-0.03	-0.05	-0.03	0.01
yearsofexperience	0.42	1	0.52	0.37	0.29	0.26	0.06	-0.07	-0.01	0.01	0.04	-0.05	0.06	-0.01	-0.02	-0	-0.06	0.02	-0.01
yearsatcompany	0.17	0.52	1	0.15	0.09	0.11	0.03	-0.03	0	-0.01	0.01	-0.03	0.05	-0.01	-0.03	-0.01	-0.04	0.01	-0.01
basesalary	0.67	0.37	0.15	1	0.47	0.39	0.08	-0.1	0.1	-0.01	0.01	-0.05	0.02	0	-0.01	-0.01	-0.02	-0.01	0.01
stockgrantvalue	0.77	0.29	0.09	0.47	1	0.25	0.01	-0.08	0.08	-0.01	-0	-0.04	-0.02	-0.01	-0.03	-0.02	-0.05	-0	0
bonus	0.49	0.26	0.11	0.39	0.25	1	0.03	-0.06	0.07	-0.01	-0	-0.02	-0.01	0	-0	-0.01	-0.03	0	0.01
Masters_Degree	0.02	0.06	0.03	0.08	0.01	0.03	1	-0.29	-0.08	-0.04	-0.04	0.31	0.04	0.01	0.01	0	0.09	0.18	-0
Bachelors_Degree	-0.14	-0.07	-0.03	-0.1	-0.08	-0.06	-0.29	1	-0.09	-0.04	-0.04	0.25	0.35	0.11	0.1	0.14	0.1	0.2	-0.01
Doctorate_Degree	0.1	-0.01	0	0.1	0.08	0.07	-0.08	-0.09	1	-0.01	-0.01	0.03	0.04	0	-0.01	0.01	0	0.06	-0
Highschool	-0.01	0.01	-0.01	-0.01	-0.01	-0.01	-0.04	-0.04	-0.01	1	-0.01	-0.02	0.11	0.03	0.01	0.04	-0.02	0.05	0.02
Some_College	-0	0.04	0.01	0.01	-0	-0	-0.04	-0.04	-0.01	-0.01	1	-0.02	0.12	0.03	0.04	0.03	0	0.04	-0
Race_Asian	-0.08	-0.05	-0.03	-0.05	-0.04	-0.02	0.31	0.25	0.03	-0.02	-0.02	1	-0.18	-0.05	-0.05	-0.06	0.11	0.23	-0.03
Race_White	-0.03	0.06	0.05	0.02	-0.02	-0.01	0.04	0.35	0.04	0.11	0.12	-0.18	1	-0.04	-0.04	-0.05	0.07	0.2	-0
Race_Two_Or_More	-0.01	-0.01	-0.01	0	-0.01	0	0.01	0.11	0	0.03	0.03	-0.05	-0.04	1	-0.01	-0.02	0.03	0.04	0.06
Race_Black	-0.03	-0.02	-0.03	-0.01	-0.03	-0	0.01	0.1	-0.01	0.01	0.04	-0.05	-0.04	-0.01	1	-0.01	0.05	0.03	-0
Race_Hispanic	-0.03	-0	-0.01	-0.01	-0.02	-0.01	0	0.14	0.01	0.04	0.03	-0.06	-0.05	-0.02	-0.01	1	0.02	0.07	-0
gender_Female	-0.05	-0.06	-0.04	-0.02	-0.05	-0.03	0.09	0.1	0	-0.02	0	0.11	0.07	0.03	0.05	0.02	1	-0.41	-0.03
gender_Male	-0.03	0.02	0.01	-0.01	-0	0	0.18	0.2	0.06	0.05	0.04	0.23	0.2	0.04	0.03	0.07	-0.41	1	-0.09
gender_Other	0.01	-0.01	-0.01	0.01	0	0.01	-0	-0.01	-0	0.02	-0	-0.03	-0	0.06	-0	-0	-0.03	-0.09	1



	totalyearlycompensation	yearsofexperience	yearsatcompany	basesalary	stockgrantvalue	bonus	Masters_Degree	Bachelors_Degree	Doctorate_Degree	Highschool	Some_College	Race_Asian	Race_White	Race_Two_Or_More	Race_Black	Race_Hispanic
totalyearlycompensation	1	0.42	0.17	0.67	0.77	0.49	0.018	-0.14	0.097	-0.015	-0.0034	-0.08	-0.028	-0.0096	-0.027	-0.026
yearsofexperience	0.42	1	0.52	0.37	0.29	0.26	0.06	-0.069	-0.0054	0.012	0.038	-0.053	0.065	-0.014	-0.023	-0.0022
yearsatcompany	0.17	0.52	1	0.15	0.089	0.11	0.026	-0.027	0.0012	-0.0059	0.0057	-0.026	0.047	-0.015	-0.028	-0.0075
basesalary	0.67	0.37	0.15	1	0.47	0.39	0.084	-0.1	0.1	-0.0085	0.0065	-0.054	0.02	0.0021	-0.0088	-0.009
stockgrantvalue	0.77	0.29	0.089	0.47	1	0.25	0.011	-0.085	0.081	-0.0086	0.0018	-0.04	-0.021	-0.0061	-0.029	-0.02
bonus	0.49	0.26	0.11	0.39	0.25	1	0.032	-0.063	0.072	-0.0069	0.0019	-0.023	-0.0078	0.0018	-0.004	-0.007
Masters_Degree	0.018	0.06	0.026	0.084	0.011	0.032	1	-0.29	-0.076	-0.041	-0.043	0.31	0.037	0.011	0.014	0.0023
Bachelors_Degree	-0.14	-0.069	-0.027	-0.1	-0.085	-0.063	-0.29	1	-0.086	-0.036	-0.038	0.25	0.35	0.11	0.1	0.14
Doctorate_Degree	0.097	-0.0054	0.0012	0.1	0.081	0.072	-0.076	-0.086	1	-0.012	-0.013	0.034	0.04	0.00073	-0.0063	0.0061
Highschool	-0.015	0.012	-0.0059	-0.0085	-0.0086	-0.0069	-0.041	-0.036	-0.012	1	-0.0054	-0.016	0.11	0.026	0.014	0.042
Some_College	-0.0034	0.038	0.0057	0.0065	0.0018	0.0019	-0.043	-0.038	-0.013	-0.0054	1	-0.019	0.12	0.035	0.037	0.03
Race_Asian	-0.08	-0.053	-0.026	-0.054	-0.04	-0.023	0.31	0.25	0.034	-0.016	-0.019	1	-0.18	-0.055	-0.051	-0.065
Race_White	-0.028	0.065	0.047	0.02	-0.021	-0.0078	0.037	0.35	0.04	0.11	0.12	-0.18	1	-0.044	-0.04	-0.052
Race_Two_Or_More	-0.0096	-0.014	-0.015	0.0021	-0.0061	0.0018	0.011	0.11	0.00073	0.026	0.035	-0.055	-0.044	1	-0.012	-0.015
Race_Black	-0.027	-0.023	-0.028	-0.0088	-0.029	-0.004	0.014	0.1	-0.0063	0.014	0.037	-0.051	-0.04	-0.012	1	-0.014
Race_Hispanic	-0.026	-0.0022	0.0075	-0.009	-0.02	-0.007	0.0023	0.14	0.0061	0.042	0.03	-0.065	-0.052	-0.015	-0.014	1

Figure 2: Survey Form (take from levels.fyi):



levels.fyi

Add a Salary

Or upload your offer letter instead

COMPANY INFORMATION

Company

Level

DETAILS & EXPERIENCE

Title

Select Job Family

Select Tag / Focus

Years of Experience

Years at Company

Location

City, State, Country (if not USA). Example: London, EN, United Kingdom

□ This is a remote role

COMPENSATION

Enter values in US dollars per year. For currency conversion, check out this tool

$ (USD) Total Yearly Compensation

Salary + Stock + Bonus total sum. This should closely match your W2 / Tax filing.

$ Salary (yearly)

$ Stock Grant (yearly)

$ Bonus (avg / year)

OPTIONAL INFORMATION (will be displayed)

Select Gender

Select Race

Select Highest Academic Level

Other Details

Free-form, any other details you'd like to add. Did you negotiate up your offer? Click here.

Submit

Figure 3: Total Yearly Compensation vs. Gender:

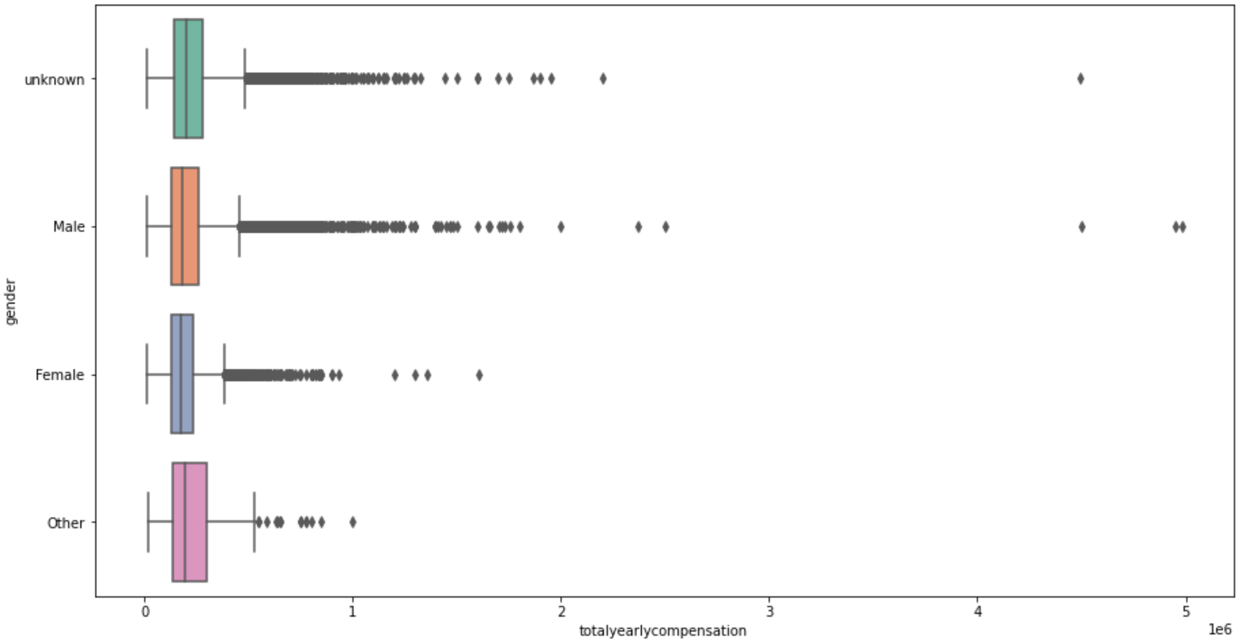
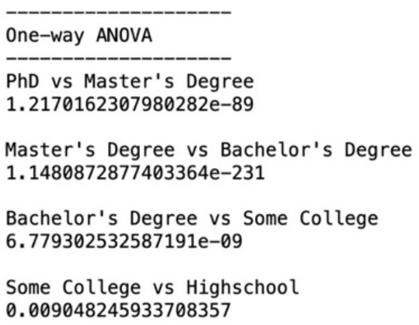


Figure 4: One – way ANOVA Results for Gender and Level of Education:



```
--------------------
One-way ANOVA
--------------------
PhD vs Master's Degree
1.2170162307980282e-89

Master's Degree vs Bachelor's Degree
1.1480872877403364e-231

Bachelor's Degree vs Some College
6.779302532587191e-09

Some College vs Highschool
0.009048245933708357
```

Figure 5: Sum of Silhouette Scores across Genders:

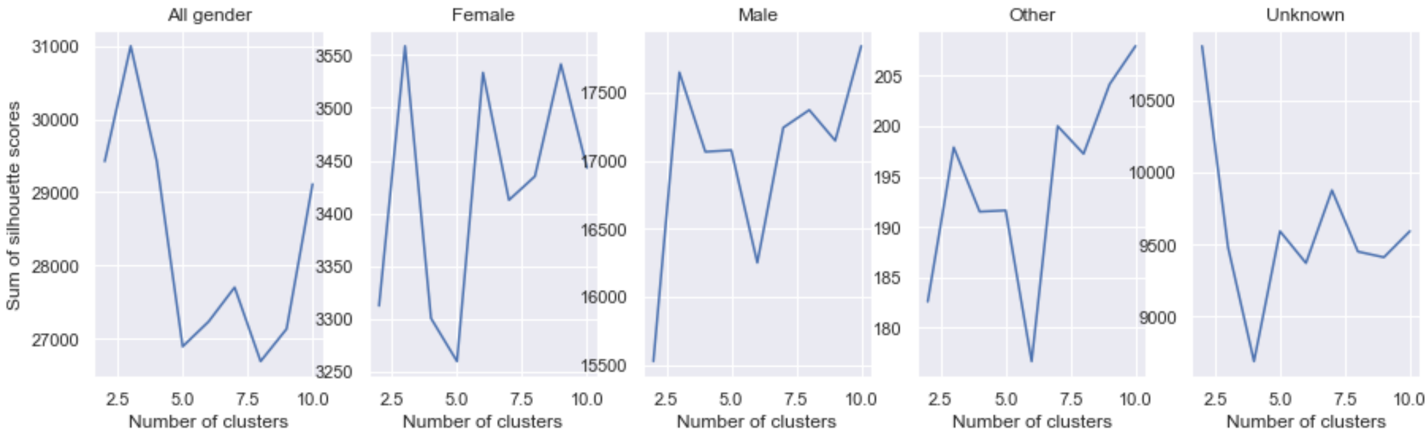
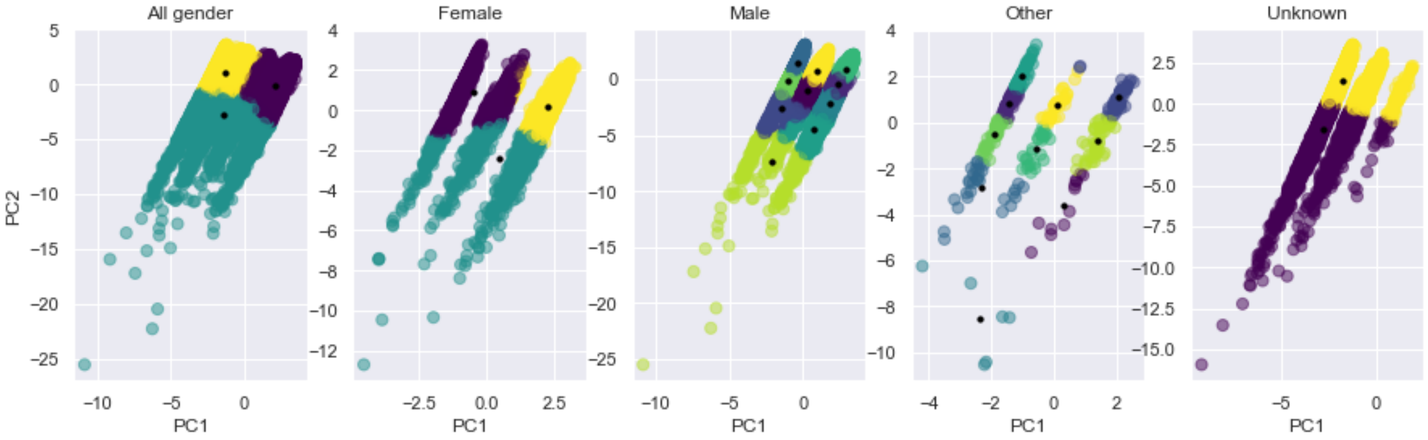


Figure 6: Principal Component 1 vs Principal Component 2 with kMeans Clustering per Gender:


All gender
PC2
PC1
Female
PC1
Male
PC1
Other
PC1
Unknown
PC1