The Minority Report

Christopher Yi, Jim Banya, Shafiullah Hashimi, Rohit Barua

CS 484-002, GMU Spring 2023

1 Project Motivation

Our selected problem is to study how to detect and correct against discriminatory hiring and recruiting practices. As a cohort of senior CS students from various minority backgrounds, this is of particular relevance to us. As we are all on the cusp of entering a very competitive job market, the ”what-ifs” surrounding the eventual outcomes of our search for employment looms over us. In a socioeconomic climate where racial issues in particular have taken center stage in the public consciousness, we feel it would be beneficial to examine discriminatory practices on a broader scale and see if data mining techniques can help examine these lingering issues. To do so, we endeavored to examine existing data sets with various demographic information and to examine any relationships between them, be it whether job performance metrics are a good predictor for a listed racial group, whether non-subjective metrics can be a good predictor for compensation, and so on.

2 Contributions

Christopher compiled the results of the entire group's individual efforts into this report, as well as adding the cursory historical contextual information to the Jupyter notebook. He also examined a simulated HR data set using KNN classification, as well as as much larger data set of salaries across the largest U.S. tech companies through multiple linear regression. Rohit helped gather various data sets for the group and create the skeleton of the final PowerPoint presentation. Additionally, he analyzed a smaller but practical data set involving campus recruitment of candidates based on their gender and attributes, such as education level, using the K-nearest neighbors classification. Jim helped to set up and organize the GitHub repository and analyzed two of the data sets to see if we could draw additional insights. This included utilizing some basic analysis such as mean values for key job-related features, or more complex analysis like linear regression for annual compensation comparison among key groups: gender and race. Shafiullah analyzed two data sets in order to discover new findings in his analysis. His methods included the utilization of KNN and linear regression models. Shafiullah also helped make the video and consolidated the four video into one video.

3 Method

3.1 Chris

Christopher chose to examine two separate data sets through two lenses: classification and regression. The first data set was examined using the K-nearest neighbors algorithm, or KNN:

$$\hat{y} = f(x) = y_{nn}(x)$$

$$nn(x) = \arg\min_{n=1,\ldots,N} ||x - x_n||_2^2 = \arg\min_{n=1,\ldots,N} \sum_{d=1}^{D} (x_d - x_{nd})^2$$

The theory is that any chosen features will prove to be a poor predictor for an output categorical feature (in this case race or sex). The general algorithm would be:

1. Preprocess data and choose relevant input features.
2. Separate features and categorical labels into two matrices.
3. Further separate these features and labels into test and validation submatrices.
4. Calculate Euclidean distance row by row between validation sample and training feature matrix.
5. Select arbitrary K lowest distances. Their indices (arg min in above formula) correspond to the closest labels.
6. Evaluate results.

The second data set was examined through the lens of a multiple linear regression model.

$$f(x) = w_0 + \sum_{d=1}^{D} w_d x_d = w^T x$$

Each of the weights beyond the base term constant w_0 correspond to one of the input features used to predict the output feature y. This function will be evaluated with the goal of minimizing the difference between the predicted labels (the projected linear equation) and actual labels. These differences, or residuals, can be squared and summed to become a residual sum of squares, or RSS. The equivalent expression using matrices is:

$$RSS(w) = (y - Xw)^T (y - Xw)$$

The list of weights w can be evaluated and selected by minimizing this loss function:

$$w^* = \arg\min_w RSS(w)$$

Further, the ordinary least squares (OLS) method can be used to estimate the actual weights themselves. The expansion and derivation (with respect to w) of the RSS function yields this expression: $w = (X^T X)^{-1} X^T y$. All told, the general algorithm using all of this information is:

1. Preprocess data and choose relevant input features.
2. Separate features and categorical labels into two matrices.
3. Further separate these features and labels into 5 submatrices for 5-fold cross validation.
4. Calculate list of weights w for each fold.
5. Calculate RSS for each list of weights.
6. Choose minimum RSS and its corresponding list of weights w.
7. Calculate predicted labels by dot product of weights and augmented feature matrix.
8. Evaluate results.

3.2 Rohit

Rohit analyzed one data set through classification. The data set was examined using the KNN classification, also known as K-nearest neighbors classification:

$$\hat{y} = f(x) = y_{nn}(x)$$

$$nn(x) = \arg\min_{n=1,\ldots,N} ||x - x_n||_2^2 = \arg\min_{n=1,\ldots,N} \sum_{d=1}^{D} (x_d - x_{nd})^2$$

The motivation behind this analysis was to spot any bias during campus recruiting between the two genders, male and female. The theory was that the chosen features will be a clear prediction of apparent bias in campus recruiting between the two genders. The algorithm for this analysis is as follows:

1. Load, preprocess, and separate the data into relevant input features.
2. Apply label encoding to relevant features.
3. Create two matrices of separate features and categorical labels.
4. Create test and validation matrices by further separating the features from the previous matrices.

5. Evaluate the Euclidean distance between training feature matrix and validation sample.

6. Determine random K lowest distances based on their indices equal to the nearest labels.

7. Compute the outcomes.

3.3 Shafiullah

Shafiullah examined two different data sets and chose classification and linear regression to assess and analyze the problem at hand. The first data set used a linear regression model.

$$f(x) = w_0 + \sum_{d=1}^{D} w_d x_d = w^T x$$

A linear regression model can we used to predict and analyze the relationship between the features (inputs) and the target (output). In this example, it is used to predict whether different predictor variables such as education, employment, maternity, gender can influence the likelihood of acquiring a job. This is a consolidation of several factors used to calculate a metric known as the Gender Inequality Index, or GII. This model was further built using ordinary least squares (OLS) regression, and tries to find the best-fit line to help minimize the residual sum of squares errors (RSS) from the predicted values and actual values. $w = (X^T X)^{-1} X^T y$.

The second data set was examined using the K-nearest neighbors algorithm, or KNN:

$$\hat{y} = f(x) = y_{nn}(x)$$

$$nn(x) = \arg \min_{n=1,\ldots,N} ||x - x_n||_2^2 = \arg \min_{n=1,\ldots,N} \sum_{d=1}^{D} (x_d - x_{nd})^2$$

This approach was useful for the second data set, as it was focused on wage gaps in several European countries. This allowed for countries to be used as the ”buckets”, or neighbors, for the classification process. If the Euclidean distance between other more objective input features could accurately sort employees into countries, the model would be accurate.

1. Load the data set and preprocess it and split the data set to input features that you seem are relevant to the issue at hand.

2. Apply encoding to the features and labels to the created categories.

3. When that is done create a test and validation set and then create matrices for them.

4. Calculate Euclidean distance row by row between validation sample and training feature matrix.

5. Select a value of K for your KNN but chose a KNN value that is not to simple where it overloads and a value that is too high where it under loads.

6. Analyze the results.

3.4 Jim

Jim analyzed two data sets. The Human Resources Diversity Analysis data set was analyzed by using the numpy mean calculation. The idea behind that analysis was to try and find trends that would show favorable and seemingly discriminatory, outcomes for workers based on race. The second data set was analyzed using the sklearn library for linear regression. This was a similar approach to Chris's, but slightly different in both the implementation and the method. In this rendition, we attempt to remove outliers from the population by limiting the incoming ceiling to a fixed value. We build our model from that modified set and then attempt to predict the annual compensation first by gender, then by race.

4 Experiment

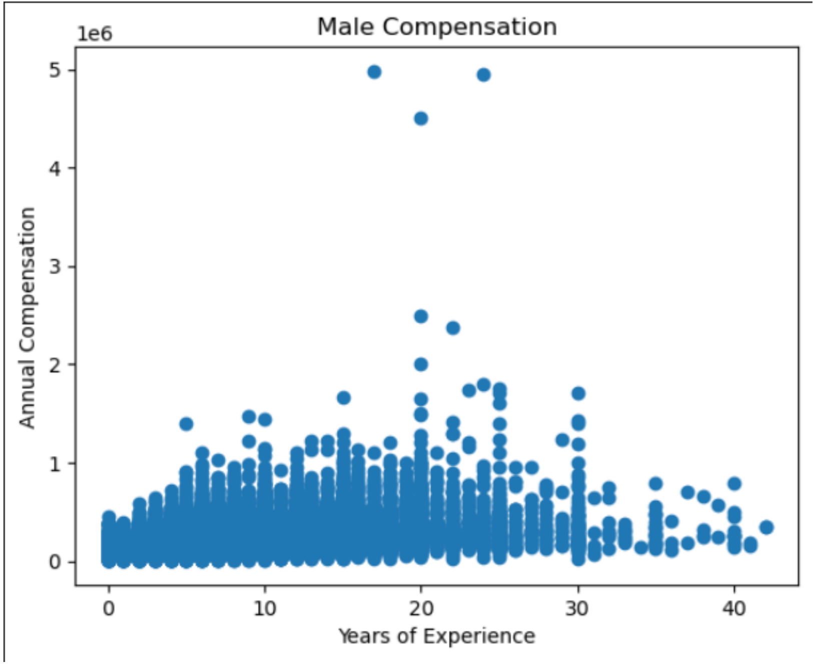
4.1 Chris

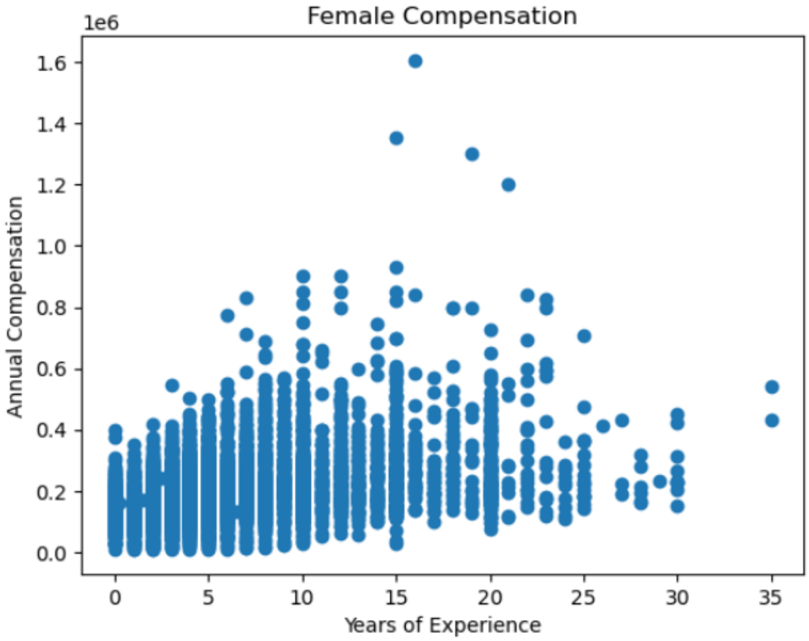
Christopher's first experiment was on a simulated HR data set with multiple columns with little significance, so a number of post-processing steps were necessary. First, only purely quantitative features were saved. In order to preserve discriminatory identifiers, gender and race were preserved by converting their string identifiers to integers. This left the employee's performance scores, engagement, race, and gender. Race was chosen to be the categorical ground truth label, and served as the buckets to which the closest neighbors in the KNN algorithm would be classified as. Then, the KNN algorithm described in "Methods" was applied to the remaining features after consolidating them into a 1:4 split into validation and training sets.

The metrics used to analyze the neighbor classifications were accuracy and micro/macro F1. An accuracy/micro F1 of 38.9% and a macro F1 of 19.2% seemed to indicate that this data set demonstrated little to no relationship between the features and race. While this had little explanatory power, particularly given it was a simulated data set, it was more a proof of concept / test run for the real data to follow.

Christopher's second experiment used a real data set: tech sector salaries of the largest U.S. companies collected from 2017 to 2021. This time, a linear regression modeling technique was used. The preprocessing of the data was more involved since there were so many more samples, and many of the responses included null responses. Culling the data meant losing a significant portion of the data, but unfortunately there was little that could be done given many of the unknown responses involved identifying features, i.e. race and gender. In addition, some non-tech occupations were culled from the data as well.

Before proceeding with the linear regression modeling, scatter plots were created to see if any relationship between years of experience and compensation could be seen visually. After creating a separate data frame centered around binary gender identifiers, the following plots were created:

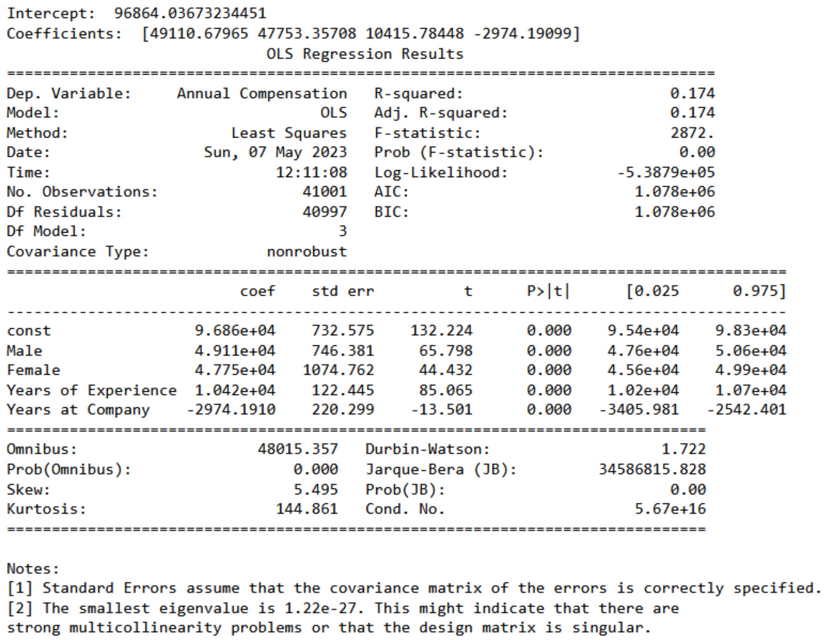




At first glance, there seems to be little to no correlation between experience and compensation across both genders. And if anything, female employees seem to have slightly higher salaries overall. However, given that female employees are very under-sampled compared to male employees (6,252 to 34,749), it is possible that the fewer female employees are comparatively more adept.

This gender-based data frame was then split, using compensation as labels and a list of male/female identifiers, years of experience, and years at the current company as features. The weights yielded by the fold with the lowest RSS (in the last recorded run) were 50023.82812, 46428.78704, 10320.46719 and -3172.74514. It seems that years at the company are slightly negatively correlated, while male employees make slightly more across several different runs (approximately $5000).

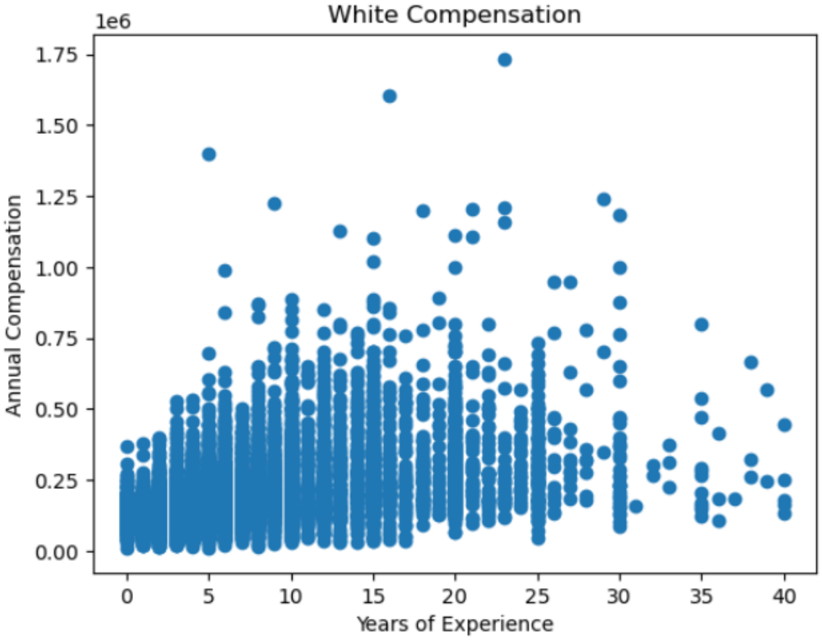
In order to verify these results, the statsmodel API was used to create an OLS model, and a modified version of the self-made OLS function (without using 5-fold OLS) was created to compare the weights from each model:

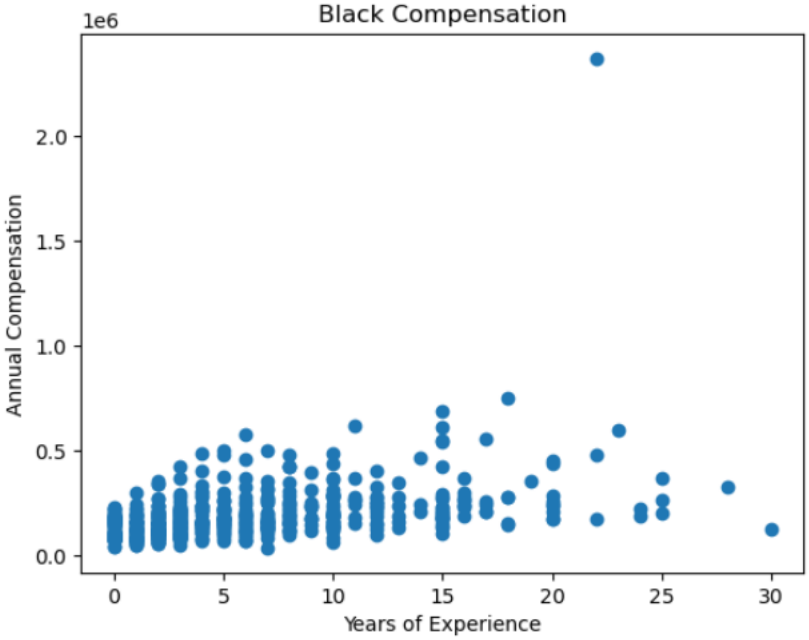


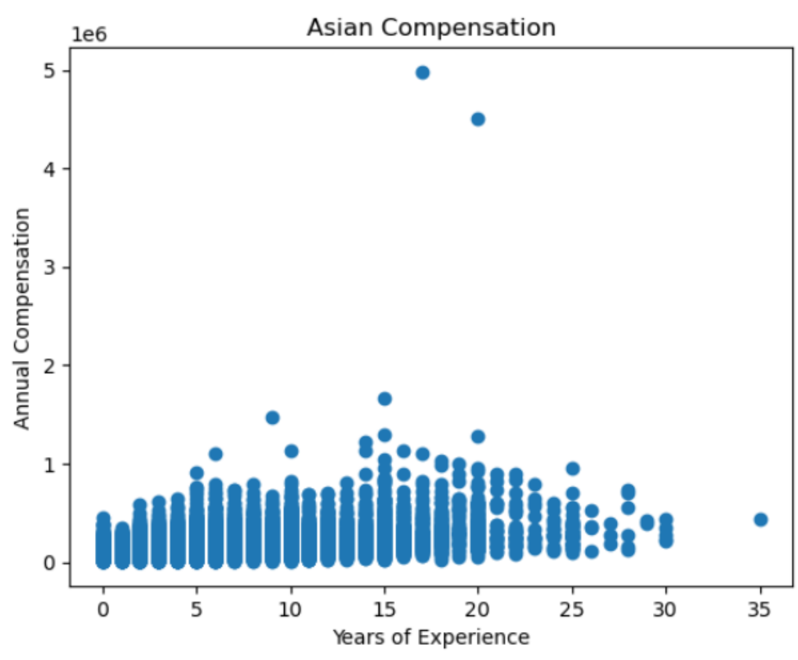
```
Intercept:  96864.03673234451
Coefficients:  [49110.67965 47753.35708 10415.78448 -2974.19099]
                            OLS Regression Results
==============================================================================
Dep. Variable:     Annual Compensation   R-squared:                       0.174
Model:                             OLS   Adj. R-squared:                  0.174
Method:                  Least Squares   F-statistic:                     2872.
Date:                 Sun, 07 May 2023   Prob (F-statistic):               0.00
Time:                         12:11:08   Log-Likelihood:            -5.3879e+05
No. Observations:                41001   AIC:                         1.078e+06
Df Residuals:                    40997   BIC:                         1.078e+06
Df Model:                            3
Covariance Type:             nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                9.686e+04    732.575    132.224      0.000    9.54e+04    9.83e+04
Male                 4.911e+04    746.381     65.798      0.000    4.76e+04    5.06e+04
Female               4.775e+04   1074.762     44.432      0.000    4.56e+04    4.99e+04
Years of Experience  1.042e+04    122.445     85.065      0.000    1.02e+04    1.07e+04
Years at Company    -2974.1910    220.299    -13.501      0.000   -3405.981   -2542.401
==============================================================================
Omnibus:                    48015.357   Durbin-Watson:                   1.722
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         34586815.828
Skew:                           5.495   Prob(JB):                         0.00
Kurtosis:                     144.861   Cond. No.                     5.67e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.22e-27. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

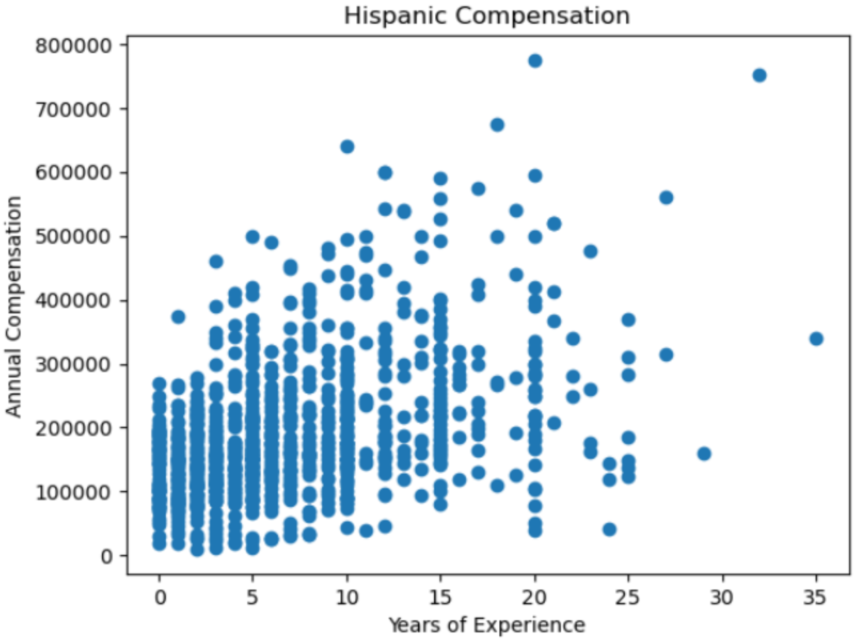
This verification seems to indicate that males seem to make on average $2000 more than female employees in this dataset.

A second run with linear regression was attempted, this time on an additional data frame created around race identifiers instead. Again, scatter plots were created to model the relationship between experience and compensation separately across single-race identifiers:


White Compensation
1e6
Annual Compensation
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
30
35
40
Years of Experience


Black Compensation
1e6
Annual Compensation
2.0
1.5
1.0
0.5
0.0
0
5
10
15
20
25
30
Years of Experience

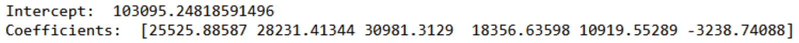




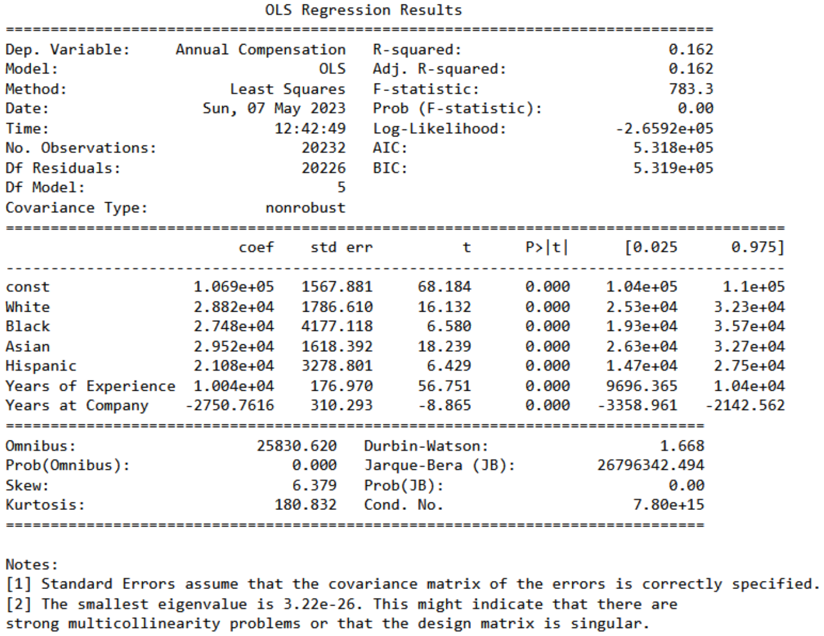
Again, there seemed to be little to no correlation that can be obviously discerned. While 5-fold OLS was applied again, the weights for the races varied so wildly that there was little point including those metrics. The only reliable weights were that years of experience positively correlated at approximately

$10,000 per year while years at the company were consistently negatively correlated.

Once again, the statsmodel API was used to verify the results:



```
Intercept:  103095.24818591496
Coefficients:  [25525.88587 28231.41344 30981.3129  18356.63598 10919.55289 -3238.74088]
```



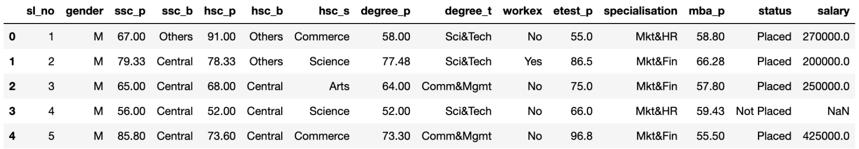
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Annual Compensation   R-squared:                       0.162
Model:                             OLS   Adj. R-squared:                  0.162
Method:                  Least Squares   F-statistic:                     783.3
Date:                 Sun, 07 May 2023   Prob (F-statistic):               0.00
Time:                         12:42:49   Log-Likelihood:            -2.6592e+05
No. Observations:                20232   AIC:                         5.318e+05
Df Residuals:                    20226   BIC:                         5.319e+05
Df Model:                            5
Covariance Type:             nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                1.069e+05   1567.881     68.184      0.000    1.04e+05     1.1e+05
White                2.882e+04   1786.610     16.132      0.000    2.53e+04    3.23e+04
Black                2.748e+04   4177.118      6.580      0.000    1.93e+04    3.57e+04
Asian                2.952e+04   1618.392     18.239      0.000    2.63e+04    3.27e+04
Hispanic             2.108e+04   3278.801      6.429      0.000    1.47e+04    2.75e+04
Years of Experience  1.004e+04    176.970     56.751      0.000    9696.365    1.04e+04
Years at Company    -2750.7616    310.293     -8.865      0.000   -3358.961   -2142.562
==============================================================================
Omnibus:                    25830.620   Durbin-Watson:                   1.668
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         26796342.494
Skew:                           6.379   Prob(JB):                         0.00
Kurtosis:                     180.832   Cond. No.                     7.80e+15
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.22e-26. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Across white, black, and Asian employees, Asian employees seemed to have the highest compensation, with black employees closely following. However, this could be from the oversampling and undersampling of Asian and black employees, respectively (black employees in particular made up only 585 of the positively identified respondents). Hispanic employees seemed to make noticeably less than the other groups.

Overall, the explanatory power of this data is muddied. There aren't any significant indicators of a gap in compensation across races that could not also be potentially be explained by dramatically undersampled groups, particularly given much of the data was culled since their identifying characteristics were not collected. There *does* seem to be a reproducable gap in average compensation that favors men over women, but beyond that it is difficult to grasp other patterns.

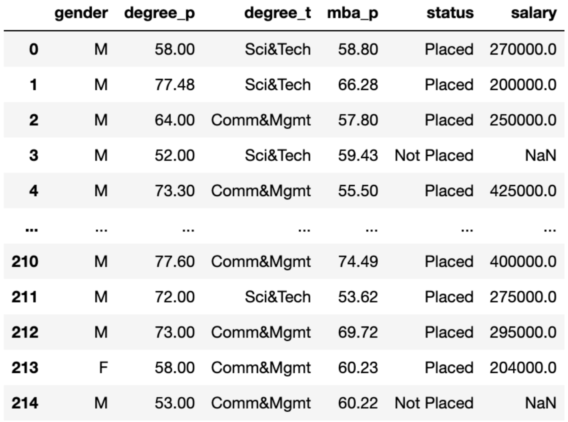
4.2 Rohit

Rohit's experiment was on a small campus recruitment data set consisting of attributes relevant to the job market, they are, different gender candidates, level of education, hiring status, and salary:



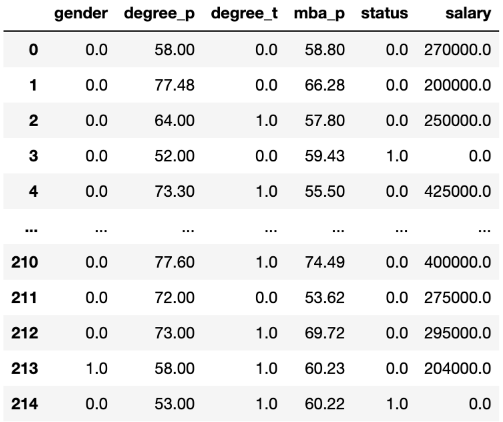
	sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
0	1	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	55.0	Mkt&HR	58.80	Placed	270000.0
1	2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Placed	200000.0
2	3	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	75.0	Mkt&Fin	57.80	Placed	250000.0
3	4	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	66.0	Mkt&HR	59.43	Not Placed	NaN
4	5	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No	96.8	Mkt&Fin	55.50	Placed	425000.0

However, the data set also consisted of many unnecessary attributes, such as serial number of the candidate, secondary education percentage, board of education, etc. Thus, the data set was cleaned and below is the outcome:



	gender	degree_p	degree_t	mba_p	status	salary
0	M	58.00	Sci&Tech	58.80	Placed	270000.0
1	M	77.48	Sci&Tech	66.28	Placed	200000.0
2	M	64.00	Comm&Mgmt	57.80	Placed	250000.0
3	M	52.00	Sci&Tech	59.43	Not Placed	NaN
4	M	73.30	Comm&Mgmt	55.50	Placed	425000.0
...	...	...	...	...	...	...
210	M	77.60	Comm&Mgmt	74.49	Placed	400000.0
211	M	72.00	Sci&Tech	53.62	Placed	275000.0
212	M	73.00	Comm&Mgmt	69.72	Placed	295000.0
213	F	58.00	Comm&Mgmt	60.23	Placed	204000.0
214	M	53.00	Comm&Mgmt	60.22	Not Placed	NaN

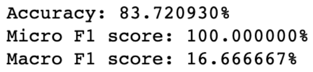
Additionally, there were attributes that needed to be converted to integers from their string identifiers, attributes like gender, status, and ”NaN” values of salary. The integers assigned to these identifiers acted as label encoding for the KNN classification:



	gender	degree_p	degree_t	mba_p	status	salary
0	0.0	58.00	0.0	58.80	0.0	270000.0
1	0.0	77.48	0.0	66.28	0.0	200000.0
2	0.0	64.00	1.0	57.80	0.0	250000.0
3	0.0	52.00	0.0	59.43	1.0	0.0
4	0.0	73.30	1.0	55.50	0.0	425000.0
...	...	...	...	...	...	...
210	0.0	77.60	1.0	74.49	0.0	400000.0
211	0.0	72.00	0.0	53.62	0.0	275000.0
212	0.0	73.00	1.0	69.72	0.0	295000.0
213	1.0	58.00	1.0	60.23	0.0	204000.0
214	0.0	53.00	1.0	60.22	1.0	0.0

For KNN classification, gender was chosen as the categorical ground truth label, being the buckets to the closest neighbors. The step by step process outlined in the methods section was followed to create the validation and training sets.

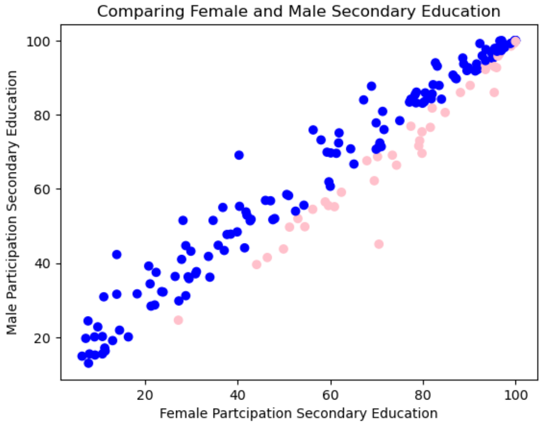
Lastly, the metrics used to evaluate the KNN classifications were micro F1 score, macro F1 score, and accuracy. An accuracy of 83.72% and micro F1 of 100.00% might indicate that there is some sort of relationship between gender and the other features. However, an macro of 16.67% also indicates that there could be some discrepancies. Overall, within the data set, it is possible that some sort of bias is present, however, due to the low macro F1 score, it is not far fetched to conclude that the data set is imbalanced and may require more samples as it was quite a small sample, but theoretically it is a good concept to explore with a much larger data set.



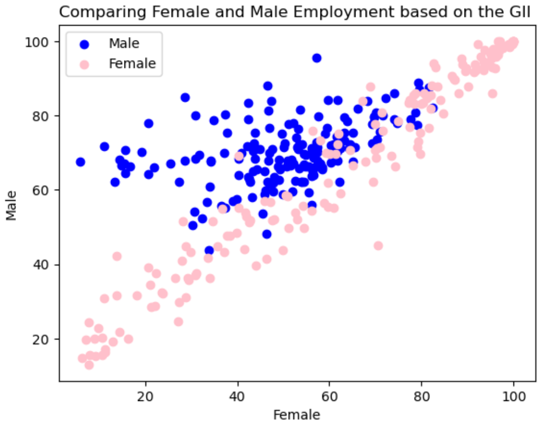
```
Accuracy: 83.720930%
Micro F1 score: 100.000000%
Macro F1 score: 16.666667%
```

4.3 Shafiullah

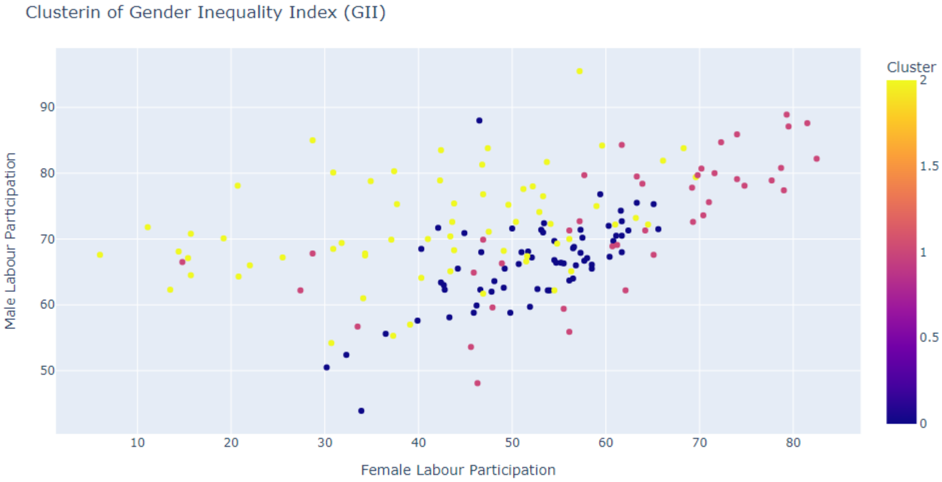
Shafiullah's first experiment was used on the Gender Inequality Index data set as a way to determine a goal or a prediction of a continuous variable (in this case a GII score). e use the continuous variable to represent data in a certain range where it allows us to get a more precise and detailed analysis of the data set. If accurate, the input features will be able to accurately predict the GII of any given sample.



Based on the data from the Female and Male secondary Education, data points between males and females are close together and show that females and males have participated in secondary education almost equally.

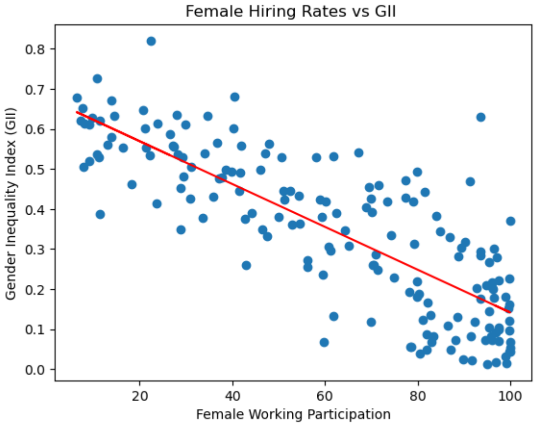


Yet when we look at the graph of Female and Male employment based on the GII, we see a significant discrepancy between male and female participation in the workforce. Even though the data points are close in the female and male secondary education fields, we can see a massive gap between females and males participating in employment. This could factor down to many reasons. For example, unconscious biases towards females, recruitment strategies, and salary gaps between males and females could make female candidates not interested in the job title.

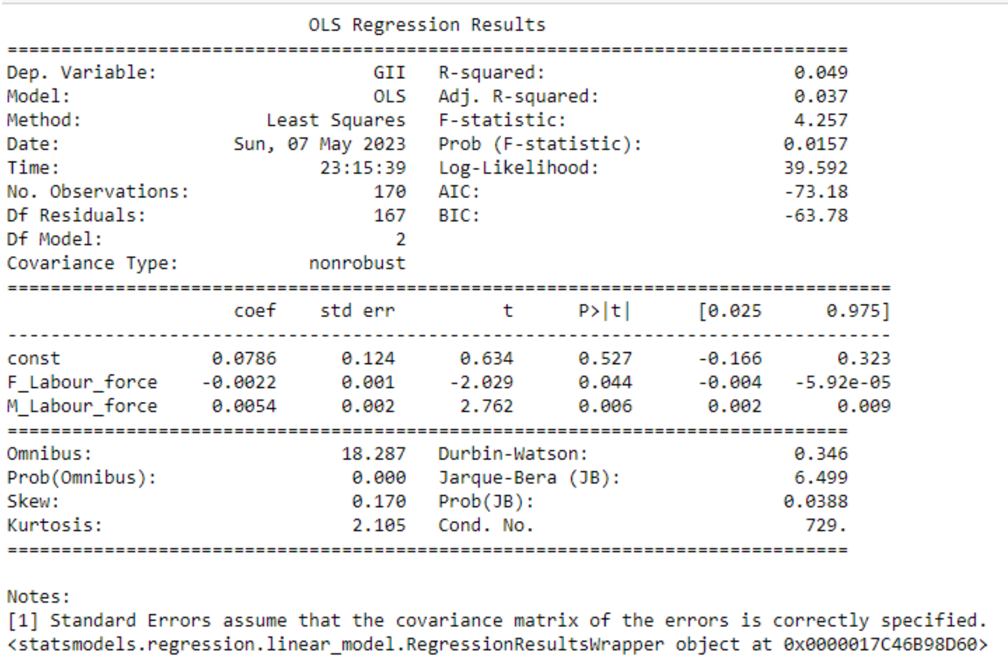


In the image above, I use clustering and separate them into 3 groups to help identify any patterns between the GII and other variables.

I used K means clustering to group countries based on variables such as GII, GDP, per capita, and others. The objective was to group different data points based on similar variables and GII in categories such as low levels of GII, medium levels of GII, and high levels of GII. Based on these groupings, we can analyze that some groups may be targeted by certain factors that can harm their employment, and other groups can benefit from other factors

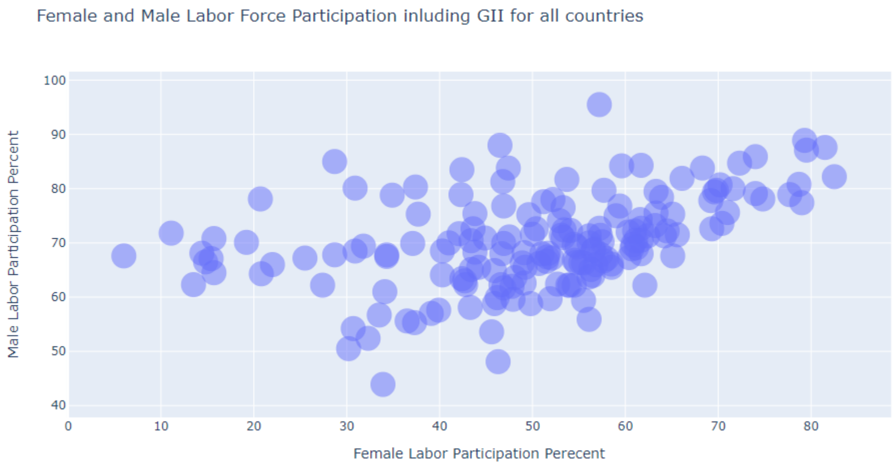


We have used linear regression to fit the model to explore the relationship between the female working participation and gender inequality.

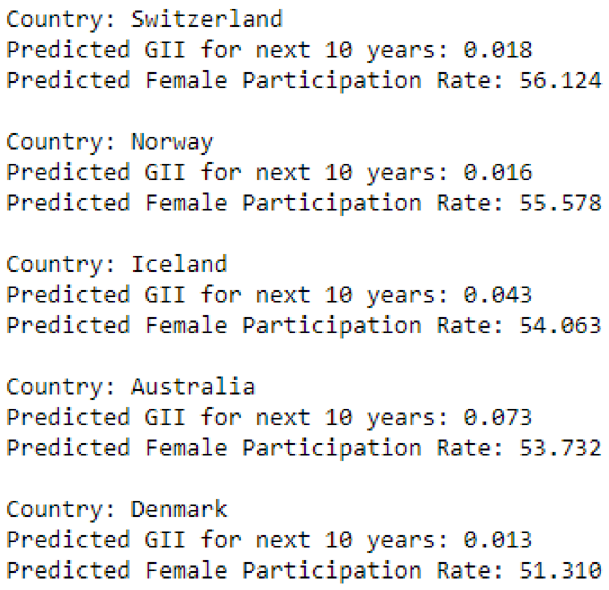


```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    GII   R-squared:                       0.049
Model:                            OLS   Adj. R-squared:                  0.037
Method:                 Least Squares   F-statistic:                     4.257
Date:                Sun, 07 May 2023   Prob (F-statistic):             0.0157
Time:                        23:15:39   Log-Likelihood:                 39.592
No. Observations:                 170   AIC:                            -73.18
Df Residuals:                     167   BIC:                            -63.78
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const              0.0786      0.124      0.634      0.527      -0.166       0.323
F_Labour_force    -0.0022      0.001     -2.029      0.044      -0.004   -5.92e-05
M_Labour_force     0.0054      0.002      2.762      0.006       0.002       0.009
==============================================================================
Omnibus:                       18.287   Durbin-Watson:                   0.346
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                6.499
Skew:                           0.170   Prob(JB):                       0.0388
Kurtosis:                       2.105   Cond. No.                         729.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
<statsmodels.regression.linear_model.RegressionResultsWrapper object at 0x0000017C46B98D60>
```

Above, we used OLS to help us understand the relationship between GII and the country's GDP and make certain assumptions such as, ”countries with high GDP have low GII and countries with low GDP have high GII.” The OLS summary helps us analyze which variables are most likely to influence a country's GII.

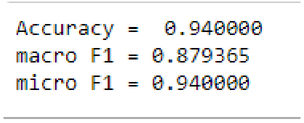

Female and Male Labor Force Participation inluding GII for all countries

Here we have a plot that predicts future GII for all the countries in the data set. Based on the linear regression model we can make an educated prediction based on the factors such as GDP, literacy rates, political factors, cultural factors, and many more.


```
Country: Switzerland
Predicted GII for next 10 years: 0.018
Predicted Female Participation Rate: 56.124

Country: Norway
Predicted GII for next 10 years: 0.016
Predicted Female Participation Rate: 55.578

Country: Iceland
Predicted GII for next 10 years: 0.043
Predicted Female Participation Rate: 54.063

Country: Australia
Predicted GII for next 10 years: 0.073
Predicted Female Participation Rate: 53.732

Country: Denmark
Predicted GII for next 10 years: 0.013
Predicted Female Participation Rate: 51.310
```

The sample output above shows the country, Predicted GII, and the predicted female participation rate in the next ten years. Of course, whether it increases or decreases depends on many external factors of that particular country.

Shafiullah's second experiment was on a data set regarding potential pay gaps in several European countries. The goal was to get some more insight about pay gap discrepancies between women and men based on feature selection, and to identify the more important variables that would affect the wage gap. In order to do this I had to make sure I chose meaningful features and split them from the features that are not meaningful. Using feature selection can help reduce dimensionality and help improve the performance of the model. By using KNN on feature selection we can select the best k features of that data set and then use the target variable to pick the best features. This can be done by using the euclidean distance formula and determine the distance between each data point and the k value. Finally we use the information from the k value and its distance to find the highest data points that have the highest similarities to the target variable based the distance.

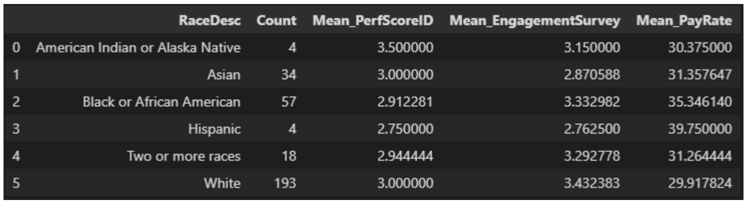


```
Accuracy =  0.940000
macro F1 = 0.879365
micro F1 = 0.940000
```

Above is the KNN result with its Macro and Miro and overall accuracy.

4.4 Jim

Jim's first experiment used the same HR data set to see if insights could be drawn from the mean values for certain key features: job performance, job engagement, and pay rate relative to race. We initially wanted to see if we could perform regression analysis on the data set, but we quickly realized that there were not enough samples for any version of regression analysis to work reliably well, so a decision was made to use the mean for each of those features. The results of that analysis were underwhelming and no patterns could be drawn.

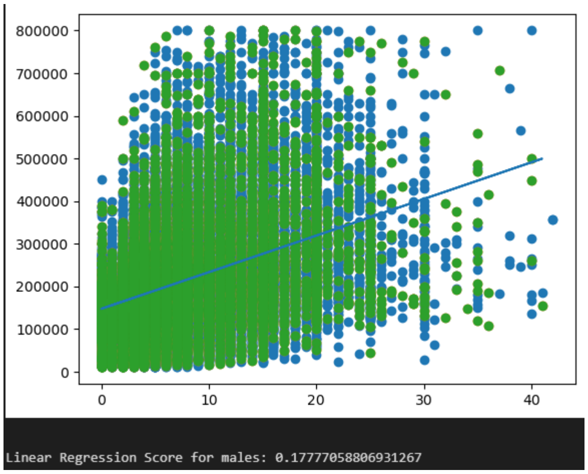


	RaceDesc	Count	Mean_PerfScoreID	Mean_EngagementSurvey	Mean_PayRate
0	American Indian or Alaska Native	4	3.500000	3.150000	30.375000
1	Asian	34	3.000000	2.870588	31.357647
2	Black or African American	57	2.912281	3.332982	35.346140
3	Hispanic	4	2.750000	2.762500	39.750000
4	Two or more races	18	2.944444	3.292778	31.264444
5	White	193	3.000000	3.432383	29.917824

The x-axis = the years of experience while the y-axis = the annual compensation. We can see that there are not any strong correlations between the mean values of PerfScoreID(performance score), Engagement Survey Values, or the mean pay rate. For context, the higher the performance score, the better the performance rating. Knowing this, we see some odd comparisons if we were to look at American Indian or Alaska Native to Hispanic. We see that they have similar representation in our data set, however, the performance score does not translate to higher pay. This could also be due to features not included in the data set like years of experience or location (important for cost of living adjusted pay).

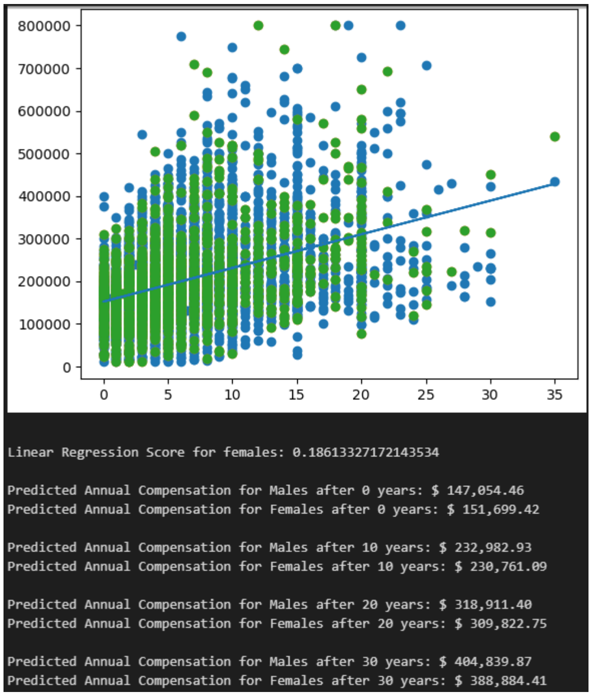
In the second experiment, we analyze a data set from Levels FYI. This is the same data set that Chris had performed a version of linear regression on earlier but we make some slight modifications to massage the data and we make use of Sklearn's library. For the first part of the second experiment, we focus on comparing annual income based on gender and years of experience in STEM roles.

Men:



```
Linear Regression Score for males: 0.17777058806931267
```

Women:

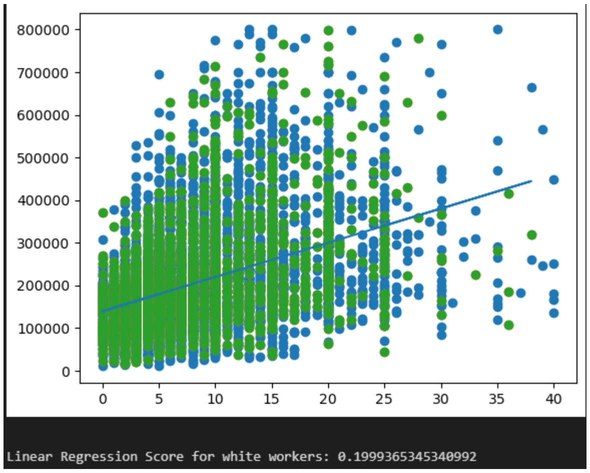


```
Linear Regression Score for females: 0.18613327172143534

Predicted Annual Compensation for Males after 0 years: $ 147,054.46
Predicted Annual Compensation for Females after 0 years: $ 151,699.42

Predicted Annual Compensation for Males after 10 years: $ 232,982.93
Predicted Annual Compensation for Females after 10 years: $ 230,761.09

Predicted Annual Compensation for Males after 20 years: $ 318,911.40
Predicted Annual Compensation for Females after 20 years: $ 309,822.75

Predicted Annual Compensation for Males after 30 years: $ 404,839.87
Predicted Annual Compensation for Females after 30 years: $ 388,884.41
```

Here we can see that our prediction at 0 years of experience shows that men are paid less than women. However, 10 years and beyond show men overtaking women in annual compensation. The difference between gender over time is pretty striking as the gap continues to widen for the equivalent years of experience.

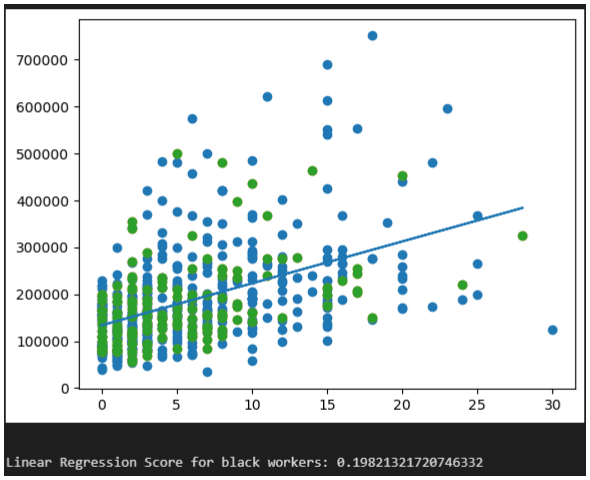
Next, we'll perform the same kind of analysis based on race.

White:



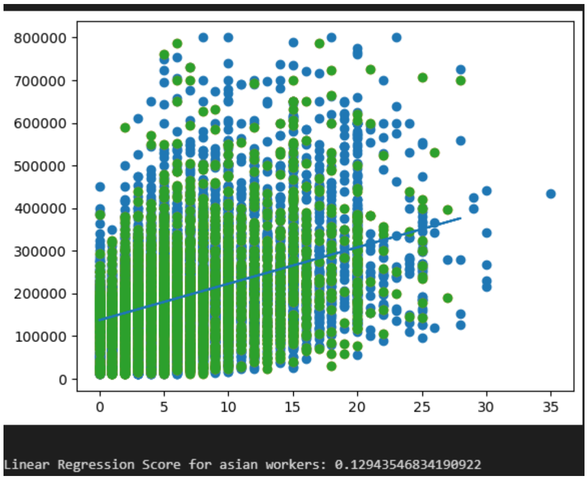
```
Linear Regression Score for white workers: 0.1999365345340992
```

Black:



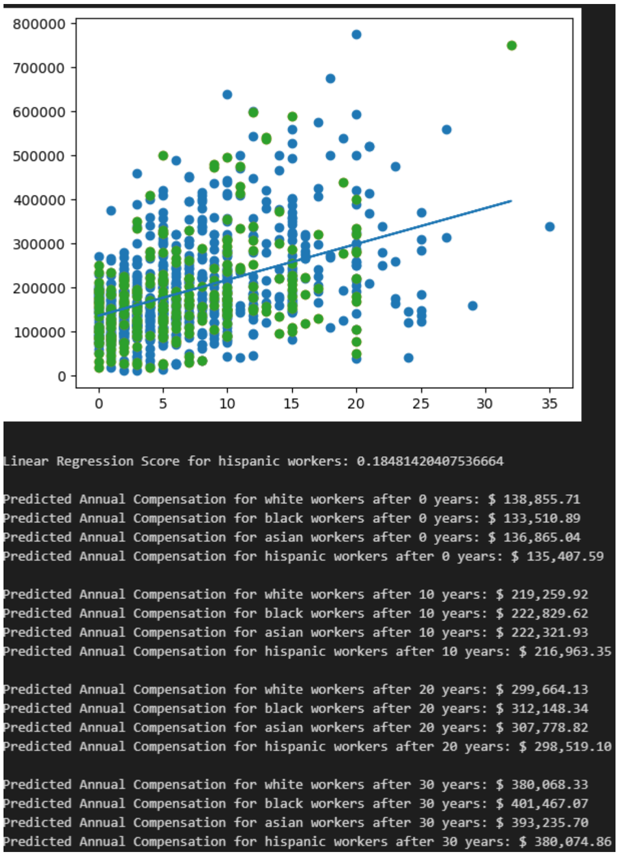
```
Linear Regression Score for black workers: 0.19821321720746332
```

Asian:



```
Linear Regression Score for asian workers: 0.12943546834190922
```

Hispanic:



```
Linear Regression Score for hispanic workers: 0.18481420407536664

Predicted Annual Compensation for white workers after 0 years: $ 138,855.71
Predicted Annual Compensation for black workers after 0 years: $ 133,510.89
Predicted Annual Compensation for asian workers after 0 years: $ 136,865.04
Predicted Annual Compensation for hispanic workers after 0 years: $ 135,407.59

Predicted Annual Compensation for white workers after 10 years: $ 219,259.92
Predicted Annual Compensation for black workers after 10 years: $ 222,829.62
Predicted Annual Compensation for asian workers after 10 years: $ 222,321.93
Predicted Annual Compensation for hispanic workers after 10 years: $ 216,963.35

Predicted Annual Compensation for white workers after 20 years: $ 299,664.13
Predicted Annual Compensation for black workers after 20 years: $ 312,148.34
Predicted Annual Compensation for asian workers after 20 years: $ 307,778.82
Predicted Annual Compensation for hispanic workers after 20 years: $ 298,519.10

Predicted Annual Compensation for white workers after 30 years: $ 380,068.33
Predicted Annual Compensation for black workers after 30 years: $ 401,467.07
Predicted Annual Compensation for asian workers after 30 years: $ 393,235.70
Predicted Annual Compensation for hispanic workers after 30 years: $ 380,074.86
```

Here we see a prediction that white and Asian workers at year 0 are paid more than black and Hispanic. However, with 10 years of experience, we see that black workers are paid the most followed by Asian workers and this trend continues into year 30. Interesting to note that white and Hispanic workers are paid relatively the same amount in our prediction by 30 years of experience.

5 Conclusions and future work

Overall, the results of our analyses are quite mixed. While the nationwide underrepresentation of certain groups in college-educated professions seems evident enough, compensation in STEM-related fields seems on average equitable (or at least equitable enough to not cause any statistical abnormalities in our analyses). While there does seem to exist a gender-based pay gap in Chris's and Jim's regression analysis, it also comes from a data set that required substantial preprocessing which culled a significant number of respondents who did not confirm their gender identity.

Our Github repository can be found through this URL: `https://github.com/GMU-MinorityReport`