How Race, Education Level, and Years at Company Impact Salary

Elijah Ford

Ahamad Jebril

Aaron Johnson

Mark Lukmskiy

Evan Smith

School of Data Science, University of North Carolina at Charlotte

DTSC 1302: Data and Society

Dr. Marco Scipioni and Dr. Ilieva Ageenko

December 8, 2024

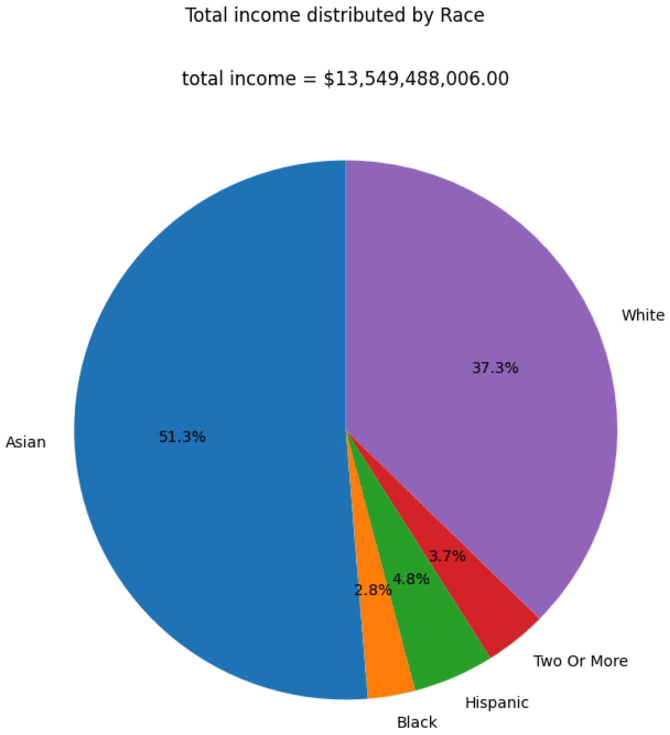
Our project involves studying a dataset that includes over 66,000 workers at various tech companies around the world. This dataset includes several categories that divide each of these workers, which are race, education, years of experience, level of education, and salary. The goal of our project was to find a relationship if any such relationship does exist, between these categories. To investigate these relationships, the technological lead of our project concocted several histograms and pie charts to help explain our findings. Each histogram represents a different aspect of the data that needs to be taken into account. For example, one histogram deals with the total population of each race that can be found within the dataset. Another shows the full breakdown of what percentage of the combined salary of all workers goes to each race. However, there was an issue with the dataset. Approximately half of the entries for workers were missing categories that were necessary for our research question. Therefore, these data entries needed to be removed so that our data would be as accurate and unbiased as possible and fit the research question properly. This brings us to the research question. This went through several brainstorming sessions, as with so many potential categories to work with, it was difficult to find one that could be researched thoroughly as well as be relevant to the data we were observing. We did finally arrive at a question that can be measured and investigated after much deliberation: Does the race of a tech worker, as well as their level of education and years of experience, have a meaningful impact on their total salary?

The findings in our dataset are not only specific to this data set but also applicable to the real world. For example, in our data set, we found that on average blacks and Hispanics earn less on the dollar then compared to whites. An article written by Morela Hernandez explained Black workers seeking jobs. Hernandez said that Black job seekers are expected by those interviewing them to negotiate less than their counterparts. It also found that Black job seekers tend to get penalized when they violate these expectations across three studies. This was not the only instance this information came up in our search for sources, as our next source showed that In 2019, the typical (median) black worker earned 24.4% less per hour than the typical white worker (Valerie Wilson). We also found that in 1970, the median annual earnings among Black men were 59% of those among White men; by 2014, this ratio had worsened to 50% (Xiang Zhou). Adding on to this, an article written by SHRM continued to support our findings by showing that black workers earn less than their white counterparts.

On the other hand, the next few sources we gathered showed how different education levels impacted different races and salaries. Texas A&M University created a graph showing earnings by different Data Scientists based on race/ethnicity and education levels. It showed that among those with an HS degree to a Bachelor's, whites earned the most, while from Master's and up Asians earned the most. We then check what North Carolina's earnings by race and ethnicity were based on the US Department of Labor. This showed that the races on the dollar, in terms of White workers, going highest to lowest, Asian, White, Black, Indian, and then Hispanic. The DOL also put out a chart talking about earnings by race, ethnicity, and educational attainment. This showed that Blacks are earning the most at almost every educational level, with Asians at a close second and Whites at a third. Even though these sources do not prove our overall conclusion of White workers earning the most out of the races, our data set had more Asians than any other race in it, taking up 51% of all the earnings (Figure 1). Therefore, our sample, of thirty thousand, which was only half of the data given to us since we took the other thirty thousand out so our computers could handle the data, could have been too small to show that Asians earn more than whites on the dollar. Since most of the statistics on the sources we used were a population of the entire United States and not a sample that we used.

Figure 1

Total income distributed by Race



Note: Asians and Whites take up the majority of earnings in our data set

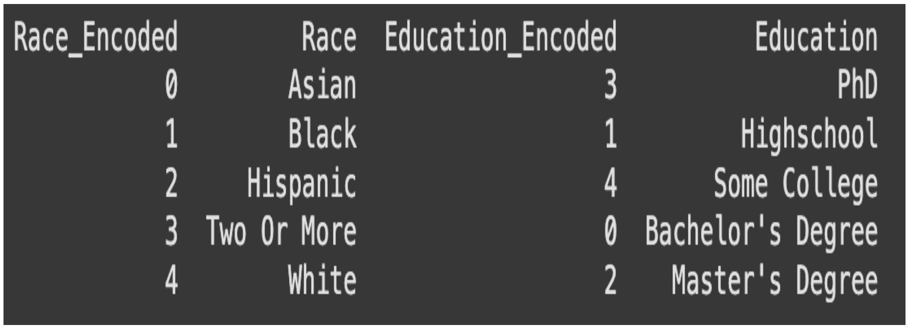
In addition to our first findings, an article written by Ariane Hegewisch took data from all genders and races comparing their wages. Hegewisch found that White men are earning the most per week out of those who work full-time jobs out of all men and women, and all races and ethnicities. This was true based on another study done by Karen Webber because despite women and minorities getting more educated and higher degrees, the gaps continue to increase in salary when compared to their white counterparts. This is mostly because as these minorities are getting paid more, Whites are getting paid even more. This is once more proved true in another study done by AAUW showing a chart on how bachelor's degrees drop the percentage of earnings compared to White non-Hispanics, only because the earnings for White people go up significantly. Overall, these sources prove and support our data set and findings by showing that the majority of the time Whites earn more on the dollar than compared to Asians, Blacks, Hispanics, and others. However, some sources found that Asians earn more than Whites in some instances. This still supports our finding because in our collected data over 50% of the total earnings went to Asians. Therefore, they realistically could be earning more on the dollar due to our sample size not being large enough when compared to the sample size of the sources. For example, our sample size was around thirty thousand, while those sources came from the United States Department of Labor, which is data from the entire US, which has a population of three hundred million, almost ten thousand times greater than our data set.

To be able to answer our research question, we will need to quantify the mean salary of tech workers of each race. This would then be ranked by highest to lowest mean salary to determine which race makes the most money while working in technology. To that end, we have to define and set aside the variables that will be used to accomplish this. Firstly, we need to establish the total count of members of each race in the context of the dataset. To do this, the races must be properly sorted, so no mistakes are made. In the dataset, each race is assigned a number (0-4), and sorted according to said number. After all the sorting is completed, two races are shown to be the most common: Asians and Whites. The second variable is total salary. This one is relatively simple, and the gathering of it is accomplished by simply adding up the salaries contained in all full rows within the dataset. After this is done, the total salary of our dataset participants is calculated as $13,549,488,006.00. Thirdly, although it is not mentioned in our research question, it is helpful to also have a grasp of the education level of the dataset participants. This way, we would have a better chance of explaining potential outliers in the data, if the education levels are different. Similarly to the sorting of races, each education level is assigned a number (0-4), and sorted by those numbers. Once both the total population proportion of each race and the total amount of income generated by the people in the dataset are defined and established as usable numbers, it is time to put them together to properly describe the results we have found after observation of the dataset and thorough manipulation of the data.

The first manipulation of the data was a drop of all rows in the dataset which contained null values. This was done to prevent outside interference by different, unrelated categories. It was also not feasible to use imputation as approximately half of the 66,000 rows contained null values, and even without the removed rows, there was still plenty of usable data for the research question. After this removal took place, the specific columns of race and education level were encoded. This assigned a numerical value to each different category or level within those columns, making it much easier to sort those values into charts and visualization that could explain what they mean. Next, the data needed to be split into two different sets, training, and testing, to properly evaluate the data and gain information from it. The split we used was that 80% of the data was used for training, and 20% of the data was used for testing. Two models were created to help answer the research question, the first one being the multiple linear regression model. The training data was used to fit the model and the testing data was used to make predictions using the model. Using the level of education, race, and years of experience as predictors, the multiple linear regression model was used to predict the salaries. The predicted salaries for different races, levels of education, and years of experience are shown in Figure 2 and Figure 3 below with their encoded values.

Figure 2

Race Encoded

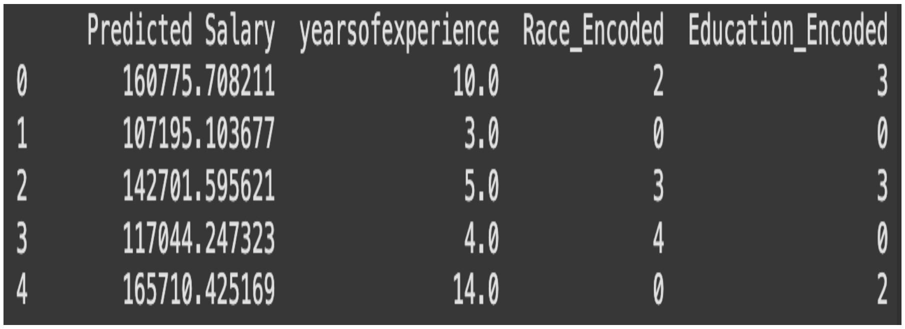


```
Race_Encoded         Race  Education_Encoded          Education
           0        Asian                  3                PhD
           1        Black                  1         Highschool
           2     Hispanic                  4       Some College
           3  Two Or More                  0  Bachelor's Degree
           4        White                  2    Master's Degree
```

Note. Values for race and education encoded

Figure 3

Predicted Salaries



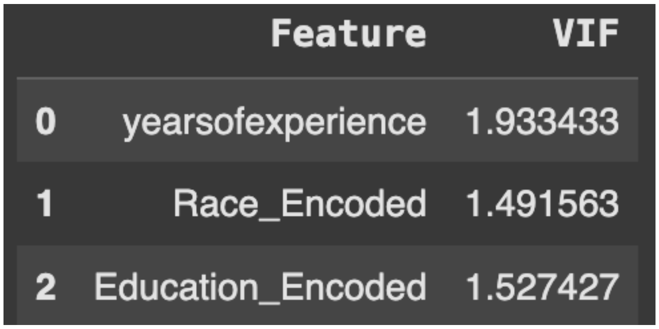
```
   Predicted Salary  yearsofexperience  Race_Encoded  Education_Encoded
0     160775.708211               10.0             2                  3
1     107195.103677                3.0             0                  0
2     142701.595621                5.0             3                  3
3     117044.247323                4.0             4                  0
4     165710.425169               14.0             0                  2
```

Note. Predicted salaries based on years of experience, race, and level of education

Based on the predicted salaries using the multiple linear regression, the assumption is that Asians with a master degree have the highest salaries with an average of 165,710.43 dollars. Calculating the variance inflation factor, R squared, and the mean squared error helped us validate our models performance. Based on the VIF values, the assumption that multicollinearity is not an issue between the predictors could be made. The VIF for the predictors is shown in figure 4 and a heatmap to show correlation between predictors is also shown in figure 5 below.

Figure 4

VIF values

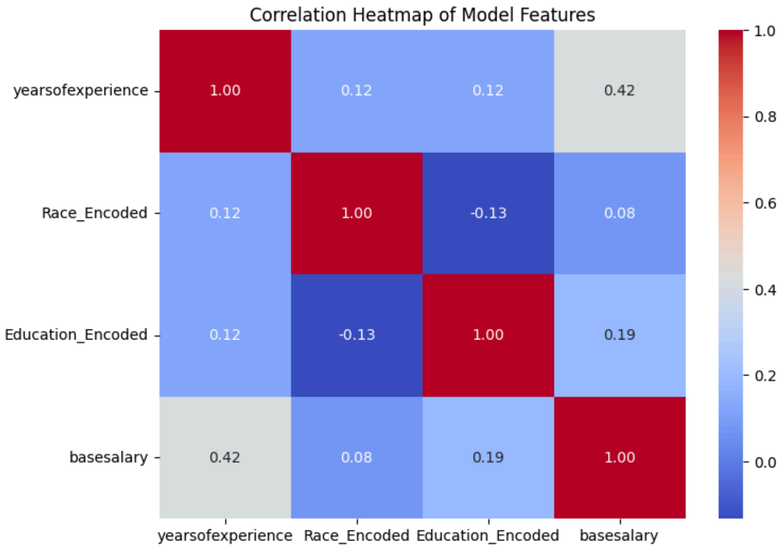


	Feature	VIF
0	yearsofexperience	1.933433
1	Race_Encoded	1.491563
2	Education_Encoded	1.527427

Note. The Variance inflation factor of the predictors

Figure 5

Correlation heatmap

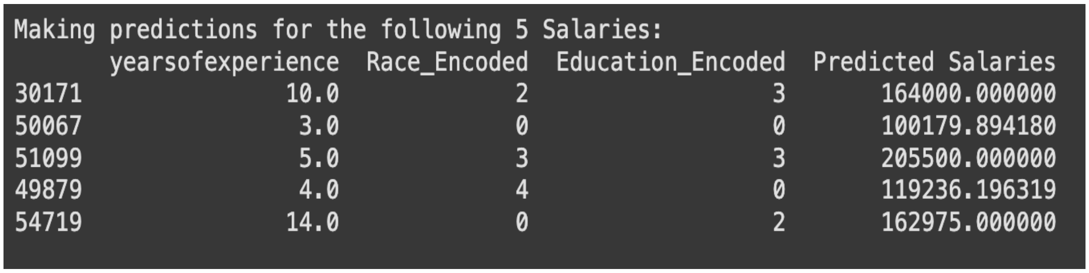


Note. Correlation between predictors and base salary

A Decision Tree Regressor model was also created to predict the salaries and to compare the outcomes with the multiple linear regression model. The model was trained using the same training set as the multiple linear regression model and was used to predict the salaries as shown in figure 6 below. According to the decision tree regressor model, mixed races with a PhD degree had the highest salaries followed by Hispanic people with PhD degrees indicating that having the highest level of education increases your base salary. The model was optimized to have 50 leaf nodes using a for loop that returned the lowest mean absolute error of $35,766. The for loop helped with optimizing the model and choosing the best max-leaf nodes to be used for the model with the lowest mean absolute error.

Figure 6

The predictions of the decision tree regressor model



```
Making predictions for the following 5 Salaries:
       yearsofexperience  Race_Encoded  Education_Encoded  Predicted Salaries
30171               10.0             2                  3       164000.000000
50067                3.0             0                  0       100179.894180
51099                5.0             3                  3       205500.000000
49879                4.0             4                  0       119236.196319
54719               14.0             0                  2       162975.000000
```

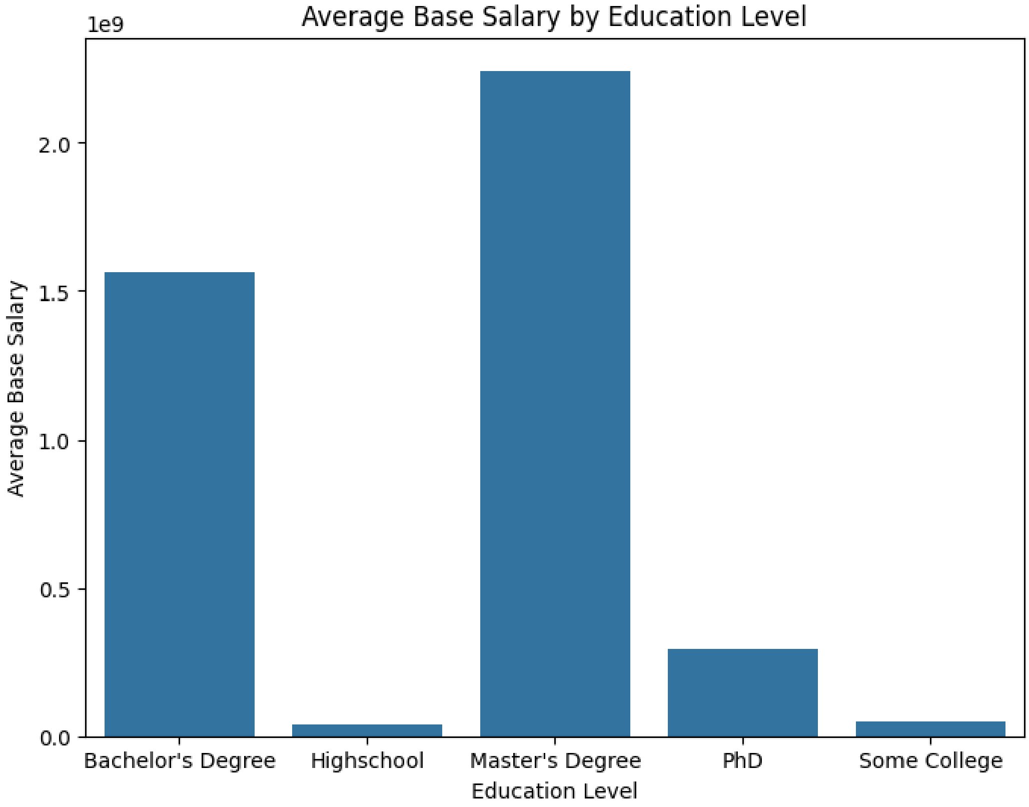
Note. Notice how the highest salary was different from the multiple linear regression model

According to the decision tree regressor model, employees with a PhD degree had the highest salary while employees with a master’s degree had the highest salaries according to the multiple linear regression model. This could be because there are many more employees with a master's degree in the tech salaries dataset as shown in figure 7 below. Since 50% of the employees are Asians in the dataset as shown in the visual earlier, this could explain why the multiple linear regression model predicts that Asians with master degree earn the highest base salaries.

While the decision tree model predicted differently, we could still conclude that the level of education, years of experience, and race could potentially determine the base salary of employees as shown in the predictions of both models. Since the multiple linear regression model predicted the base salaries better than the decision tree regression model based on the dataset, we could use it to make predictions rather than using the decision tree regression model.

Figure 7

Average Base Salary by Education Level



According to both models, the conclusion that the level of education, years of experience, and race have an impact on the salary could be made as shown in the summary statistics above. Therefore, we can safely assume that there is at least some level of bias contained in the minds of employers when they go to hire new employees and decide on raises, bonuses, and promotions for their employees who are already with the company. However, our data is only a small sample of American workers, and any extrapolation may be limited by its scope. 30,000 employees is less than 10% of the workforce, so this sample could not properly be used, as it is too small to be statistically significant. This information can still be useful, though. Whether it is to confirm these findings or disprove them, one could conduct an international sample on the same categories as our data.

Sources

AAUW. (n.d.). *Race and the Pay Gap*. AAUW : Empowering Women since 1881. https://www.aauw.org/resources/research/race-and-the-pay-gap/

Decker, B. (2023, September 15). *Financial Impact of Educational Attainment by Race/Ethnicity*. Data Science. https://www.tamus.edu/data-science/2023/09/15/financial-impact-of-educational-attainment-by-race-ethnicity/

Earnings ratios by race, ethnicity, and educational attainment. (2024). DOL. https://www.dol.gov/agencies/wb/data/earnings/earnings-ratio-race-ethnicity-education-annual

Miller, S. (2020, June 11). *Black Workers Still Earn Less than Their White Counterparts*. Www.shrm.org. https://www.shrm.org/topics-tools/news/benefits-compensation/black-workers-still-earn-less-white-counterparts

National Center for Education Statistics. (2023, May). *COE - Annual Earnings by Educational Attainment*. Nces.ed.gov. https://nces.ed.gov/programs/coe/indicator/cba/annual-earnings

U.S. Department of Labor. (2020). *Earnings Disparities by Race and Ethnicity | U.S. Department of Labor*. Www.dol.gov. https://www.dol.gov/agencies/ofccp/about/data/earnings/race-and-ethnicity

Webber, K. L., & Canché, M. G. (2015). Not Equal for All: Gender and Race Differences in Salary for Doctoral Degree Recipients. *Research in Higher Education*, *56*(7), 645–672. JSTOR. https://doi.org/10.2307/24572049

Wilson, V., & Darity Jr., W. (2022, March 25). *Understanding black-white disparities in labor market outcomes requires models that account for persistent discrimination and unequal bargaining power*. Economic Policy Institute. https://www.epi.org/unequalpower/publications/understanding-black-white-disparities-in-labor-market-outcomes/

Zhou, X., & Pan, G. (2022). *Higher Education and the Black-White Earnings Gap **. https://scholar.harvard.edu/files/xzhou/files/zhou-pan_gap.pdf