APA Citation

Analyzing Unconscious Bias in Indeed's Employee Resume Search

Team DeBIAS:

Rachel Antony, Benjamin Bral, Seth Gleason,

Joanna Ihm, Aarushi Malhotra, Philip Mathew,

Soham Nagaokar, Johnny Rajala,

Kyle Truong, & Daniel Zhu

University of Maryland, College Park

GEMS 497: Team Dynamics and Research Methodology

Dr. Steve Sin

May 1, 2023

Abstract

This project analyzes if artificial intelligence (AI) hiring systems demonstrate racial bias as measured by prestige bias against graduates of historically black colleges and universities (HBCUs), measured in a number of different metrics, and how that bias can be mitigated. The metrics we used to measure prestige were: size, university rankings, public vs. private universities, and attendance of HBCUs vs. attendance of non-HBCUs. We examined how the hiring site Indeed utilizes AI to list candidate resumes by relevance and measured the relationship between candidates' resume rankings and the universities they attended. While we found no significant difference between the overall average rankings of applicants from HBCUs and applicants from non-HBCUs, we did find significant differences between these applicants when we made comparisons based on variables such as major, experience level, and most recent company size.

Future research on this topic includes training an AI model on the collected resumes to see if the same results are generated and adjusting the model to mitigate said biases. This research will shed light on the bias embedded in human hiring departments. With businesses considering AI as a tool for hiring, companies must understand that AI hiring systems can perpetuate the same biases found in human hiring on a larger scale.

Chapter 1: Introduction

Theoretically, the goal of hiring is to select individuals who will most benefit a company. For this reason, the process of hiring is inherently selective to find top-performing individuals with the right skill set. However, it is not uncommon for other factors unrelated to a candidate's potential to influence hiring decisions like gender, race, or religion. These factors are indicative of unwanted biases within hiring, which is important since companies may miss out on a "best candidate" if they use a factor unrelated to fit to select candidates, such as race. There are numerous ways humans introduce unwanted biases into the hiring process. Throughout their daily lives, humans have unconscious biases which constantly influence their decisions. For example, a hiring manager may unknowingly make a decision using a person's accent or facial expression, either of which may not give any insight into how well an applicant will perform.

To reduce the impact of human biases or errors, several industries are turning to artificial intelligence (AI) to make decisions, making it almost ubiquitous in today's society. Artificial intelligence refers to a computer algorithm that is "trained" to complete a specific task and can improve itself for more accurate decisions. In healthcare, AI is on the verge of outperforming professionals in the accuracy of diagnosing (Davenport & Kalakota, 2019). AI is also seen in autonomous vehicles, voice-initiated personal assistants, and other sectors of society. The hiring industry is following suit and using algorithms with the intent of making hiring more efficient and objective.

However, even AI systems are susceptible to biases despite their reputation for objectivity (Derous & Ryan, 2018; Mayson, 2019; Borgesius, 2018). One example is Amazon's automated hiring system. Dastin (2018) found in 2017, Amazon disabled its AI hiring system

after realizing the algorithm discriminated against women. The machine trained on Amazon's previous hiring data, and since the majority (60%) of Amazon's workforce were men, the system falsely assumed women were less desirable to the company.

The Amazon example demonstrates a critical interaction between technological advancements and social impact. As we increasingly rely on AI to make up for human shortcomings, it is crucial we program the technology thoughtfully and carefully; otherwise, it will serve to perpetuate, and in most cases even exacerbate, the discriminatory practices that have existed in society for years.

The goal of our research was to identify and measure potential biases in AI to ensure its implementation in the hiring process is equitable and only considers relevant variables. Specifically, we investigated the hiring platform, Indeed, and whether racial bias is present in its candidate ranking algorithm. This study focused on racial bias as it affects populations that have historically been discriminated against and is prevalent in American society today.

Chapter 2: Literature Review

To inform our study, we examined the history of racial bias in hiring outcomes. Then, we reviewed the use of algorithms in hiring practices and how they may perpetuate existing biases. Finally, we summarized the current solutions to combat these biases.

2.1 Human Bias in Hiring

Previous research provides evidence of implicit human bias in the hiring process that disadvantages racial and/or ethnic minorities. Applicants with "African-American sounding names" are 50% less likely to receive a positive callback than identical applications that have "white-sounding names," demonstrating what researchers call the *Ethnic Name Bias* (Derous & Ryan, 2018). In another study, Kline et al. (2021) sent 83,000 job applications to 108 Fortune 500 companies to test hiring discrepancies. The team found "At least 7% of all jobs in [the] experiment discriminate against distinctively Black names," showing the great discrepancy that exists in hiring practices (Kline et al., 2021).

This pattern can persist regardless of the hiring manager's identity. Derous and Ryan (2018) reported that light-skinned applicants were consistently preferred over dark-skinned applicants. This trend remained even if the hiring manager was dark-skinned. This preference against darker skin, ethnic names, and other markers of racial/ethnic identity are indicators of learned, unconscious racial bias. Ultimately, unconscious human biases have an outsized impact on hiring decisions, and on a population-level it can severely limit entire demographic groups from obtaining employment.

2.2 AI-Driven Hiring Practice

Today, many hiring processes use AI in an attempt to make hiring more efficient and unbiased. According to Sánchez-Monedero et al. (2020), 98% of Fortune 500 companies are using some form of Applicant Tracking System in their hiring practices. Raub (2018) noted how AI companies such as ARYA and MYA offer recruiting software products to companies, including IBM, Lenovo, Parsons, and more. These products automatically find or even interview ideal candidates by matching keywords from job descriptions with candidate profiles. Other services such as HireVue go a step further and apply AI to analyze video interviews and identify which candidates resemble the "ideal employee" based on diction, facial movements, etc. Pymetrics forgoes resumes entirely; instead, it asks applicants to play a series of brain games and compares their results with those of the top company employees.

Ideally, these algorithms avoid human biases that may affect hiring decisions, but it is possible for AI to integrate implicit biases of the people who programmed them. Data journalists from Germany conducted a series of experiments on Retorio, an AI startup that determines the personality of a job applicant based on a video interview. The study revealed that simple changes to appearances like wearing a headscarf can drastically change the results (Harlan & Schnuck, 2021). Even when the audio track was removed or replaced, the reporters found virtually no change in the results, suggesting that the AI was entirely focused on visual cues and ignored audio cues. This poses a concern since appearances are often influenced by identity factors such as religion or culture that are unrelated to candidate quality, and it demonstrates how AI can propagate biases if not developed properly.

2.3 Causes of Bias in AI

There are multiple ways bias can intentionally or unintentionally be embedded within a system. Even in cases where AI is blind to factors such as race or gender, there still exists the possibility of unintentional, inequitable results. The two most common ways unintentional bias can arise in an AI are biased training data and the selection of class labels and target variables (Borgesius, 2018).

Developing most AI systems starts by inputting training data. The AI learns from this training data through "levels" (LeCun et al., 2015). The levels begin with simple tasks and get progressively more complex. For example, when training an AI with an image, the first level may locate where the edges of objects in the photo are while the second level may identify the shapes of objects. These levels get more granular and ultimately enable AI to make decisions like a human. Beyond this basic overview of training, there are many types of machine learning (ML) methods.

However, if the training data is taken from a biased sample, the AI's future decisions will inevitably be biased as well. According to a 2018 paper by Buolamwini and Gebru, three commercial gender classifiers created by Microsoft, IBM, and Face++ were most inaccurate at classifying darker faces. The dataset used for facial recognition had a high concentration of male (77.5%) and/or white (83.5%) faces which skew the training material (Buolamwini & Gebru, 2018). If an overall population is unbiased, it is essential to take a sample that is representative of the population to develop an unbiased AI. Bias in the training data can lead to biased outputs.

Bias also manifests in AI when selecting class labels and target variables. Class labels are the categories into which data are classified, and target variables are the factors used to classify the data. One common problem that results in bias is choosing a target variable that inadvertently

serves as a proxy for another factor. Borgesius (2018) posed a hypothetical example of a hiring algorithm that uses punctuality as a target variable to classify whether or not an employee is a "good" employee or a "bad" employee. Many factors may affect an employee's punctuality such as proximity to the office, parental responsibilities, and transportation needs. Depending on the weight that the company puts on the target variable of punctuality, the algorithm could be more biased against those who cannot afford to buy their own car or those with parental responsibilities outside of work.

2.4 Potential Solutions to Combat AI Bias

Once the bias is identified, one can start mitigating it through a variety of techniques. So far, there are three main approaches to mitigating bias: data-driven approaches, analyzing latent structure, and algorithmic modifications. The details of these approaches will be further discussed below.

Data-Driven Approaches

Dixon et al. (2018) discuss applications of data-driven approaches to reducing bias in a text classification model. Their objective was to discourage their model from classifying comments containing terms used to describe individual identity (such as "queer," "gay," "black," etc.) as "toxic" (p. 68). However, in their pre-existing data, there was a heavy correlation between a comment containing such terms and said comment being deemed as toxic, so they had to manually introduce examples of non-toxic comments that contained these identity-describing terms, which proved successful in removing the bias from the model without negatively affecting its decision-making ability. Overall, while this approach is effective, it is not feasible for all

tasks, as finding more data can be difficult in cases where datasets are limited (Dixon et al., 2018).

Analyzing Latent Structure

One can break a dataset down into a list of variables, with some considered as the input/independent variables and others considered as the output/dependent variables. The goal of artificial intelligence is to use the underlying relationships between these variables and predict what output variables would come from a set of input variables. However, beneath all of these, there exist *latent variables*. Latent variables, unlike the above input and output variables, are inferred instead of observed. We can use a class of machine learning methods called *unsupervised learning* to infer and obtain these variables and how they correlate to the other variables in the dataset. Unsupervised machine learning extracts the structure and relationships within a given dataset when they are not already known. Instead of trying to predict outcomes from a pre-existing structure (which is the objective of *supervised learning*), unsupervised learning fits a structure to a dataset, allowing one to see how the different aspects of the dataset connect. If we apply these techniques to a dataset in a certain context, we would be able to extract the latent variables from the dataset.

Such is the approach of Amini et al. (2019), who used unsupervised learning (specifically, a variational auto-encoder, or VAE) to extract the latent variables from a dataset and evaluate how these latent variables affect the output of their model. From this, they can balance their training dataset by resampling their training batches by removing data that encourages bias within the model. Therefore, they can obtain a less biased model. They validated this theory by applying their algorithm to a facial recognition problem which became more accurate in identifying faces after removing biases caused by race and gender.

Algorithmic Modifications

The rest of the approaches to mitigating bias usually modify the model somehow. Bellamy et al. (2018) and Kamishima et al. (2012) discuss using a given list of sensitive variables to compute a regularization term that affects the model's parameters and updates them in such a way that it aims to minimize the effects of these sensitive variables on the outcomes. Bellamy et al. (2018) also detail other approaches such as encoding the features into a latent space, reweighting training examples, using an adversarial training procedure, and many others. Of note is their use of adversarial machine learning which presents an interesting solution whereby one model is trained to complete a task normally while another model is trained to detect bias in the first model's output. Ultimately, training finishes when the second model fails to detect bias in the outputs of the first model, implying that the first model is less biased in its predictions. Additionally, they present techniques for directly modifying the outputs of the model, allowing for the odds of each outcome to be equalized in some way.

2.5 Educational Prestige Bias

As previously discussed, unconscious racial bias in humans can present itself in the algorithms used in hiring systems. Theoretically, people with unconscious bias against people of color will have negative perceptions of institutions with high populations of people of color, so racial bias may appear in many ways. In the higher education environment, this may present as educational prestige bias. Educational prestige bias refers to unequal treatment based on where an individual received their degree, regardless of the quality of education. This phenomenon can occur when recruiters prefer candidates from schools perceived as more distinguished, such as

Ivy League colleges or long-established universities, without considering the academic achievements of the candidates. Rivera (2011) found that some recruiters for elite positions restrict their search to Ivy-League students and neglect students from other schools who do not necessarily lack the same skills. This bias can manifest in many ways; for example, a liberal arts school versus an R1 research university, or Harvard University versus University of Maryland. There is a dearth of information regarding educational prestige bias as an indirect form of racial bias which this project aimed to address.

2.6 Conclusion

Humans have implicit biases that indirectly influence their decision-making every day. Racial bias is prominent in many American institutions, and as companies increasingly rely on AI for seemingly objective decisions, it is crucial to understand that algorithms are vulnerable to these biases as well. If the goal is to use AI to eliminate racial disparities, it is important to measure both direct and indirect forms of racial bias. Previous literature explored more direct manifestations of racial bias, such as ethnic name bias or skin color. This project focused on an indirect form of algorithmic racial bias by analyzing the correlations between an applicant's university education and visibility on a popular employment site.

Chapter 3: Methodology

In this section, we detail our methodology for measuring prestige bias in a hiring algorithm. First, we give an overview of our procedure and explain the choices behind our data collection and our variables. Next, to give context to the procedure, we present our research questions and hypotheses before detailing the different variables we are collecting and why. Finally, further details about the data collection method and explanation for our choice of variables are discussed.

For this study, we decided to analyze the resume ranking algorithm for Indeed, where they list the resumes of job seekers to employers based on a job position and location search. We chose Indeed's algorithm instead of another hiring website's, such as Glassdoor or Monster, because Indeed's interface is easier to understand and use. This made it simpler to obtain the data for the hypothesis and allowed us to gather more data to be used in a potential AI model. As with any other company in this field, Indeed would avoid displaying conscious bias in its resume rankings to follow federal law, promote talented candidates, and avoid paying costly fines. As a result, any potential bias in the resume ranking would most likely be the result of unconscious bias, giving us a greater incentive to look through Indeed's resume rankings.

For our independent variable, we decided to look at whether the resume holder went to a HBCU to measure the prestige bias of Indeed. Since we could not directly measure racial bias since people do not list their race on their resumes, we used the university that they attended, specifically whether they went to an HBCU or not, as a proxy for race. For our dependent variable, we looked at the rank of the resume in Indeed, which we defined as the position of the resume relative to the first resume on the first page, with rank 1 being the highest and most desirable resume rank. Further details regarding our control variables are discussed below.

3.1 Research Questions

- *How are a hiring algorithm's decisions influenced by educational prestige bias?*
- *What technical solutions can be implemented to reduce bias in AI hiring systems?*

3.2 Hypotheses

The following hypotheses were developed based on our literature review findings and proposed research questions:

Hypothesis 1

If algorithms take on the bias of their creators, then the hiring algorithm results should reflect a certain level of prestige bias.

Hypothesis 2

If prestige bias is detected in Indeed's algorithm, then utilizing adversarial training to create a neural network should mitigate the bias.

3.3 Variables

We organize our variables into the following categories. A more detailed description of how our variables are represented can be found in Appendix B.

Independent Variables

- **College/University:** Using the college a person went to, we used a number of metrics to understand how 'prestigious' this university may be, and therefore we assessed how the prestige level could affect the resume's ranking on Indeed. Once we recorded what college/university the applicant went to, we can measured how prestigious the university

is. The metrics we use are public vs private universities, the US News National University Rankings, university size, and status as a HBCU vs. non-HBCU.

- **Public Vs. Private University**: Most people associate private universities with more prestige than public universities (Volkwein & Sweitzer, 2006). We chose these variables to see if the bias appears in the AI.
- **US News National University Rankings**: We used this as a metric because this list is commonly used to assess university value and prestige; the higher a university is on these rankings, the more prestigious it is considered. Most hiring managers are more likely to hire an identical applicant who came from a university ranked higher on these rankings.
- **Size of University**: This was measured as the number of undergraduate student body. We used this as a metric because we believe that a hiring manager is more likely to hire an applicant coming from a university they have heard of, and the larger the university is, the more likely they are to have heard of it.
- **Historically Black College Vs. Non-Historically Black College**: HBCUs have been shown to be effective in attracting, retaining, and graduating students. Commodore and Njoku (2020) noted that HBCUs, specifically regional public universities, perform well and "are doing so at a disproportionately higher rate compared to their PWI [predominantly white institutions] counterparts" (p. 102). Although the previous statement highlights how HBCUs are equivalent to non-HBCUs in terms of quality of education, they face varying levels of discrimination from potential employers, showing how the bias against the HBCUs affects the individuals that attend them. In a study by Fogle (2011), 20

hiring managers from corporations within Atlanta, Georgia were interviewed regarding their perception of graduates from HBCUs compared to graduates from non-HBCUs. The participants tended to believe HBCUs do not prepare graduates for the real world and lack the necessary skills to succeed. Furthermore, a study by Jackson (2007) found that while the tech employers recognized HBCU graduates received the necessary training and education to succeed, they still did not provide them with the same employment opportunities as their non-HBCU peers.

Dependent Variable

- **Ranking:** Indeed's search engine returns results for a particular search string by order of "relevancy" of resumes, which is determined by a proprietary algorithm. We expect biases in this model will be reflected in the order that the applicants are returned. The order of applicants determines their rank in the search results, with the more "relevant" resumes appearing earlier. Should the model contain bias, we expect it will be reflected through significantly different average ranks between populations. In our study, a discrepancy in average rank between HBCU and non-HBCU graduates will be used to determine whether there is bias in the model.

Control Variables

The following attributes are commonly considered when hiring new applicants and are used to determine an applicant's quality. We want to ensure that there is little correlation between the applicant's college/university and the following variables.

- **Degree:** The highest postsecondary education obtained: Bachelor's Degree, Master's Degree, Professional Degree, or Doctoral Degree. This variable needs to be controlled

since companies would be biased towards applicants with higher qualifications since they tend to be more capable and knowledgeable in their fields.

- **Degree Type:** Type of highest postsecondary degree obtained: Arts, Business/Finance, Science, Engineering, Medical Field, or Legal. These fields may impact the rankings since companies may prefer certain fields over others when it comes to hiring an applicant.
- **Experience Type:** A combination of an applicant's experience in their field (entry level, first line, experienced, and mid-level) and their management experience (non-supervisory, or supervisory). Companies will prefer individuals with more experience since these applicants would be more impactful in improving the workplace and its productivity.
- **Experience Length:** Length of applicant experience in the most recent position (measured in months). Companies would prefer applicants who have more experience since they are likely to possess more relevant skills for the position.
- **Career Length:** Length of overall applicant experience in the field (measured in months). Companies would prefer applicants who have more experience since they would be able to contribute more to the company.
- **Employment:** Is the applicant currently employed: Unemployed, Part-time, Full-time. We want to control this variable since employers would most likely be biased toward full-time employees, especially in the context of consultants.
- **Employer:** The approximate size of the employer that the applicant was most recently employed at. We will differentiate based on the categories of self-employed, small private offices (2-19 employees), small businesses (20-200 employees), mid-size

businesses (201-1000 employees), large businesses (1000+ employees), and well-known large corporations. Employers will prefer applicants who work in similar-sized companies since this experience can be more easily applied if the applicant is to be hired.

- **Most Recent Update:** Latest update of the resume. Calculated by taking the date the resume was last updated and subtracting the date the resume is downloaded (measured in days). Since relevance may be an important factor in evaluating the skills of an applicant, it is important to control this variable so that we know if the resume rankings are affected by HBCU attendance instead of updated time.

3.4 Procedure

To collect the data, we created and used an employer account on Indeed with a standard subscription to be able to view and download the resumes available. We used a standard subscription rather than a professional subscription since the only benefits of the professional subscription are additional chances to contact prospective employees which we did not need to do. Next, we created a project, which is an organizational tool provided by Indeed that allows employers to track resumes of interest. We used the project to keep track of downloaded resumes and avoid collecting redundant data.

Afterward, we searched for resumes in the search tab. Specifically, we searched for the job “consultant” within the city “District of Columbia" and limited the radius to 25 miles. We decided to search for consultants since many consultants graduated from college and can have a range of degree types, from bachelors to PhDs, which we are interested in looking at since there could be different trends based on the degree type. We selected Washington D.C. for our target location since a preliminary search produced a considerable amount of results from Indeed and

the city had a higher concentration of surrounding HBCUs compared to other major cities like New York City or Los Angeles. This allowed us to compare the rankings of HBCUs and non-HBCUs more accurately. The 25-mile radius was selected to incorporate as much of D.C. and the surrounding suburbs, focusing on the demographics of Indeed resumes from the region.

After Indeed returned a list of resumes from the search, we started at the beginning of the list and checked if the resume had been previously collected. If we hadn't collected the resume, we downloaded the resume and collected information on the variables as described in Appendix B. It was important to download the resumes once it was included in the data since the resume could be modified or even removed by the user in the future, which can cause confusion later. The resumes were stored in a secure location and were renamed according to the "resume_id" variable to protect the identity of the resumes involved. We repeated this process for however many resumes we wanted to collect, incrementing the ranking counter as we went down the list of resumes from the top of the search page, with the rank starting at number 1. The exact procedure can be found in Appendix A.

The variables degree type, experience type, and employer were subject to personal opinion since it was impractical for every decision to be reviewed by the team. To mitigate potential issues with any data collector having significantly different opinions from the team, each individual reviewed 10% of the resumes collected by another individual for any egregious differences of opinions on those variables. By reviewing these decisions, we determined whether a set of resumes warranted more review and aligned with the majority of the team's interpretation. Otherwise, we can say that the decisions were within reason and acceptable. In addition, each data collector noted down any confusion or uncertainty they had when categorizing a resume, with the team reviewing and voting on how to interpret the information,

providing another buffer against issues involving personal opinion. If there was more time, we would have conducted a more comprehensive review of the resumes and had more cross-checking, but we believed that the reviews of the samples were sufficient to uncover any significant differences in the interpretation of data within our team.

Chapter 4: Results

4.1 Initial Findings

To evaluate our first hypothesis, we carried out a weighted least-squares regression (WLS). We could not perform ordinary least-squares (OLS) regression because we found that the data exhibited a heteroskedasticity problem, which violates one of the OLS assumptions (Long, 2018). To correct for this heteroskedasticity, we use a method of weighting our data called weighted least-squares (WLS) regression. WLS corrects for heteroskedasticity by weighting the variables by the inverse of variance for the given inputs, and these weights were found by modeling the variance (Ryan, 2008). After completing our regression, we computed the t-value of each coefficient, and we were able to determine the significance of our results with N=438. The results for the WLS can be seen in table 1. As we can see, only one of the categories significantly ($p = 0.05$) contributes to an individual's average ranking in Indeed's search results, the 'other' category for degree type. Further, our regression had an R-squared=0.064.

Variable	coef	p-value
Update	-0.1063	0.977
Career Length	-0.4330	0.895
Currently Employed	3.5150	0.267
University Size	-1.4802	0.741
University Rank	1.2764	0.738
Master's Degree	-1.3491	0.696
Professional Degree	3.3387	0.381
Doctoral Degree	0.3537	0.931
Business/Finance Degree Type	-1.1601	0.796
Science Degree Type	-5.1722	0.178
Engineering Degree Type	-3.3055	0.354
Medical Field Degree Type	-0.8296	0.809
Legal Degree Type	-0.9958	0.798
Other Degree Type	-6.9139*	0.009
HBCU Public University	1.7074	0.684
HBCU Private University	6.3480	0.116
Non-HBCU Private University	-4.3695	0.297
Supervisory, first line experience	-6.6207	0.235
Non-Supervisory, experienced	-2.7964	0.660
Supervisory, mid-level	-8.7858	0.130
Small Private Office	3.2069	0.491
Small Business	6.8560	0.160
Mid-Size Business	0.9630	0.815
Large Business	-2.5532	0.537
Well-Known Large Corporation	3.0826	0.520

Table 1: P-values for coefficients, with stars by the coefficients of variables which are statistically significant ($p < .05$). Note we use one-hot encoding for regression, meaning Bachelor's Degree, Art Degree Type, Non-HBCU Public University, Non-Supervisory entry experience, and self-employed are not shown.

We also ran a two-sample Kolmorgorov-Smirnov (KS) test on the distributions of average rank for HBCUs and non-HBCUs. The KS test measures the difference between two distributions, so comparing the average rank distributions of HBCUs and non-HBCUs with the KS test allows us to tell whether there is any significant difference between how the ranks of HBCU applicants and non-HBCU applicants are distributed. As seen in Figure 1, below, our distributions were not significantly different at the p = 0.05 level. The KS statistic we found was 0.120, which equates to a p-value of 0.136. This indicates that Indeed assigns ranks relatively

similarly for HBCU and non-HBCU applicants. As a result, we could not conclude that Indeed's algorithm has any significant overall biases against applicants from HBCUs.

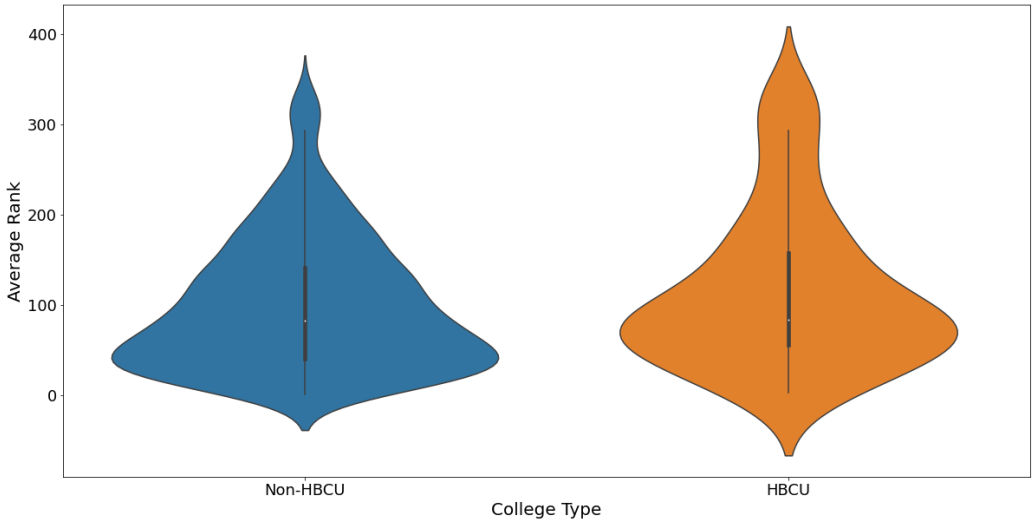

Figure 1: Violin plot for the distribution of average rank for HBCU vs. non-HBCU applicants

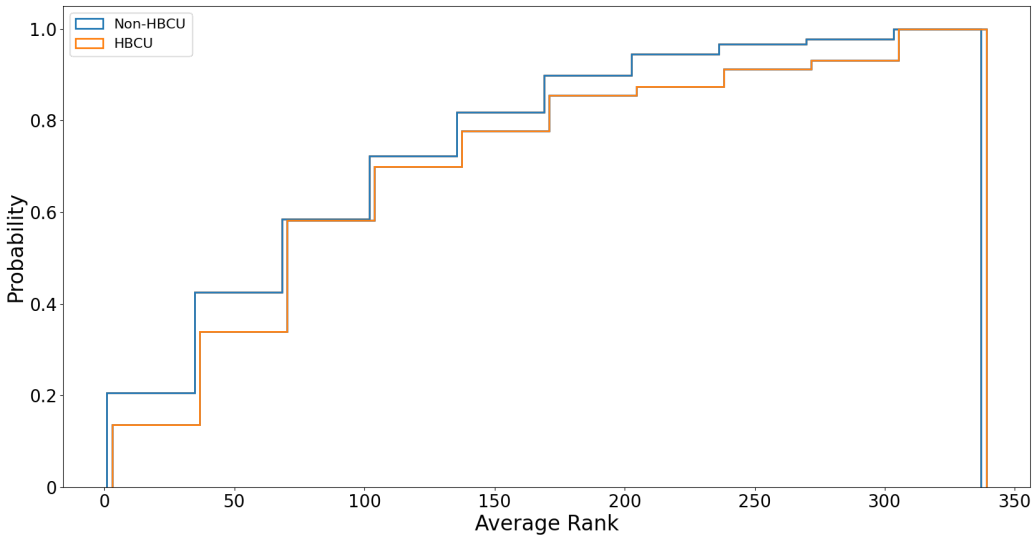

Figure 2: Cumulative distribution (CDF) plot of average rank for HBCU vs. non-HBCU applicants

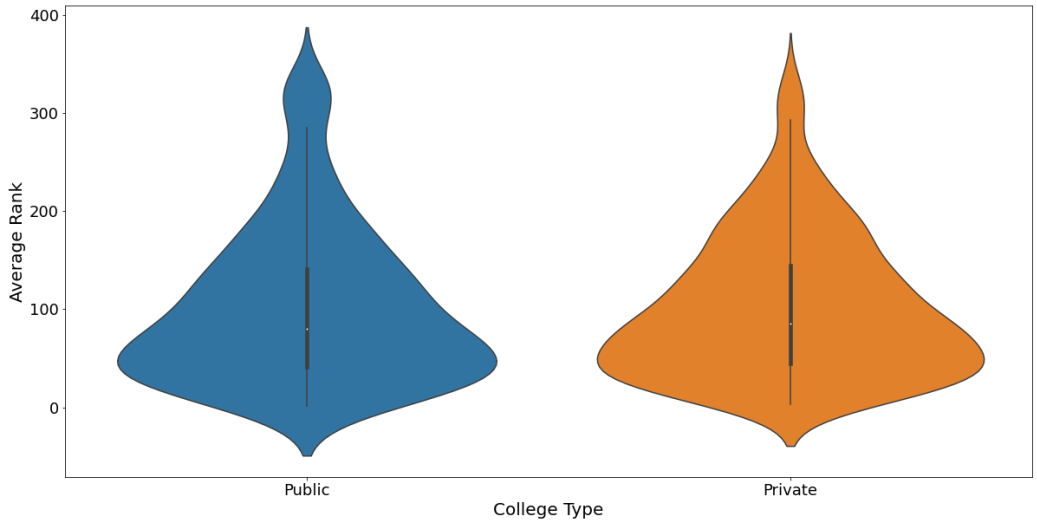

Figure 3: Violin plot for the distribution of average rank for public vs. private college applicants

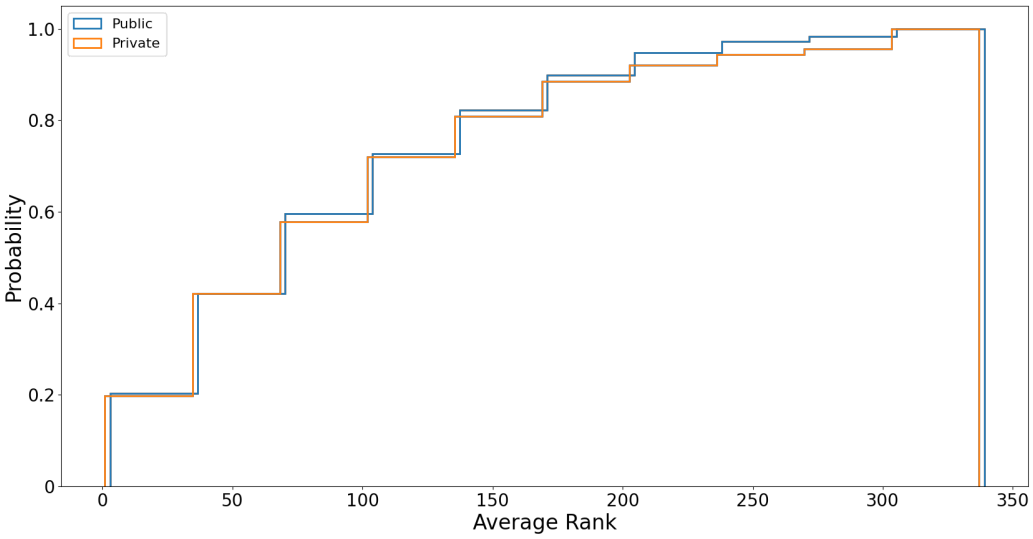

Figure 4: CDF plot of average rank for HBCU vs. non-HBCU applicants

4.2 Secondary Findings

Interestingly, when we compared the average rankings of HBCU and non-HBCU resumes based on specific variables (e.g. major, experience level, experience length, etc.), some significant differences (i.e. outside of a 5% margin of error) in applicant ranking were observed. Specifically, when comparing the average rankings of HBCU and non-HBCU resumes based on the following variables, we observed the following differences:

- **(Major)** Applicants from HBCUs performed significantly worse in the medical field, however, they were significantly better performers in the legal field. See Figure 5, below, for results comparing all majors.

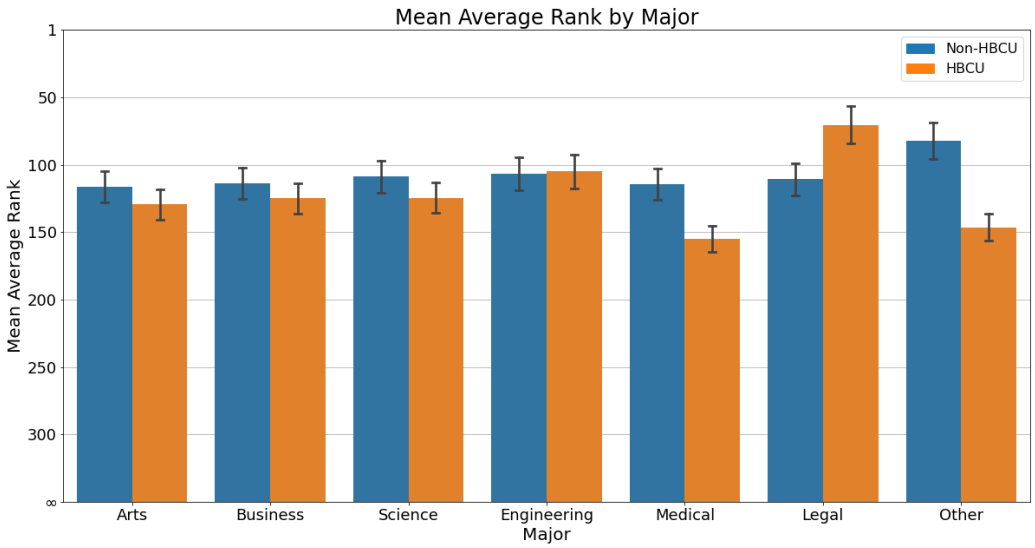

Figure 5: Mean average rank for each major type

- **(Experience Level)** Applicants from HBCUs with the experience level of "experienced supervisory" performed significantly worse than those from non-HBCUs. See Figure 6, below, for results comparing all experience levels.

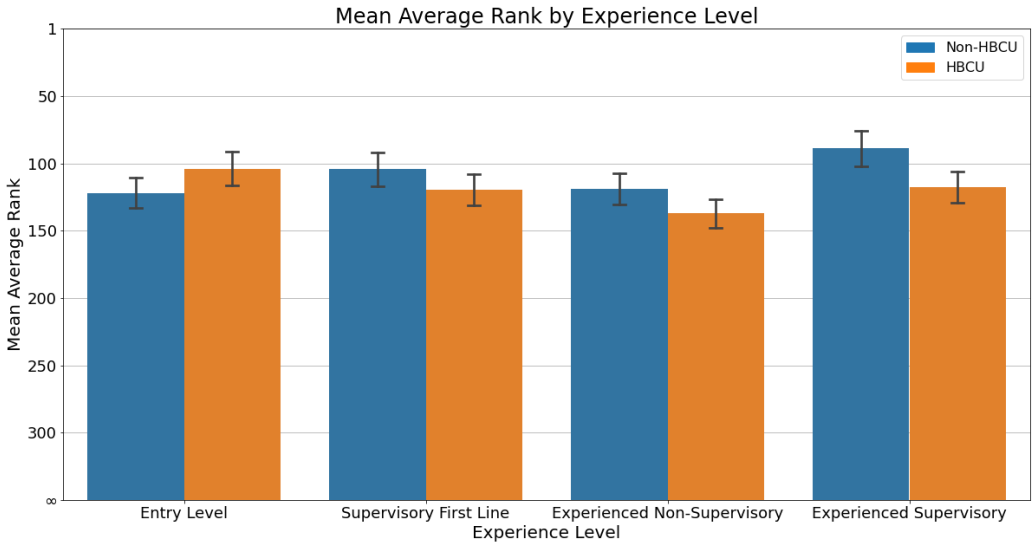

Figure 6: Mean average rank for each experience level

- **(Experience Length)** Applicants from HBCUs in the 3-4 year and 4-5 year ranges for experience length performed significantly worse than their contemporaries from non-HBCUs. See Figure 7, below, for results comparing all experience lengths.

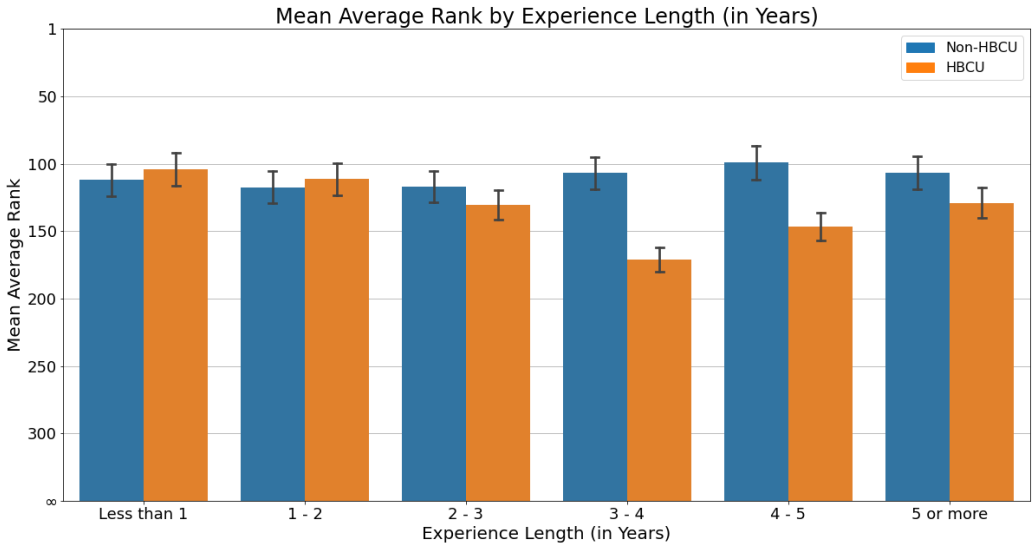

Figure 7: Mean average rank for different experience lengths

- **(Career Length)** Applicants from HBCUs with careers between 100 and 199 months long performed significantly worse than their contemporaries from non-HBCUs, however applicants from non-HBCUs with careers longer than 400 months performed worse than their HBCU-educated counterparts. See Figure 8, below, for results comparing all career lengths.

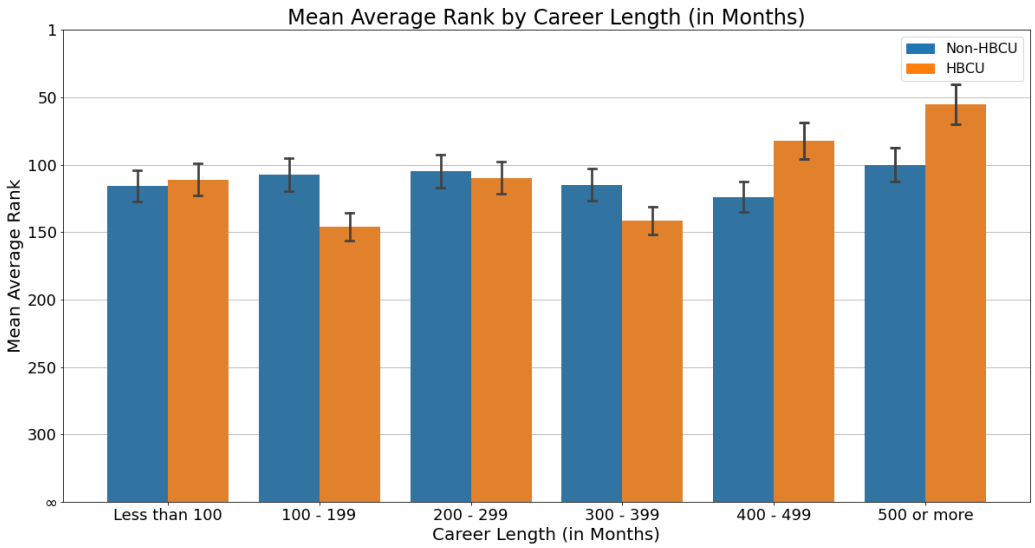

Figure 8: Mean average rank for different career lengths

- **(Company Size)** For the categories “Self-Employed”, “Private Office”, “Mid-Size Business” and “Large Corporation”, applicants from HBCUs performed significantly worse than applicants from non-HBCUs. See Figure 9, below, for results comparing all company sizes.

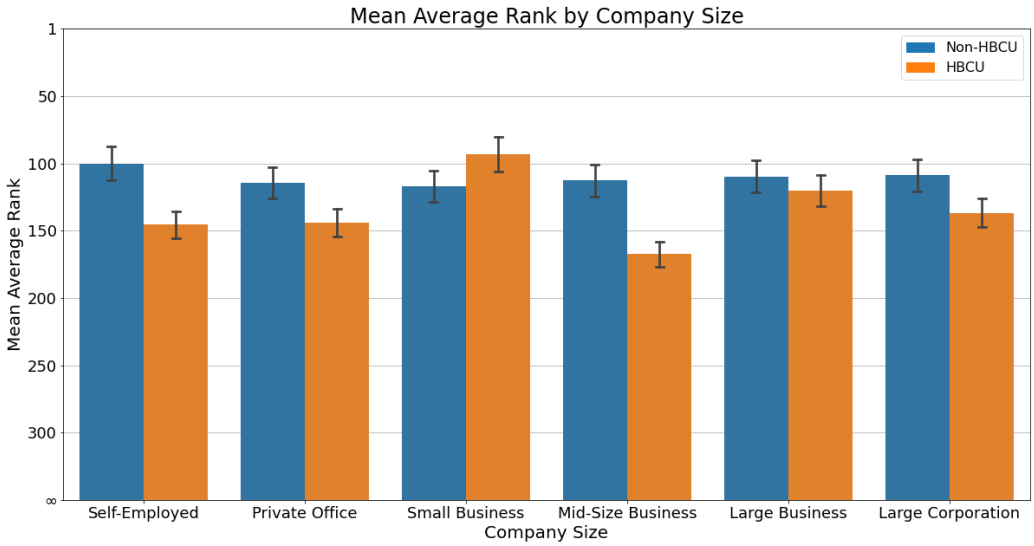

Figure 9: Mean average rank for each company size

Chapter 5: Discussion

Overall, the data analysis did not find a significant difference in the rankings of applicants from HBCUs versus their non-HBCU counterparts. While this is a seemingly positive conclusion, our second analysis uncovered findings that demonstrated preferences for one applicant group in certain situations.

Data Analysis

The data we collected displayed heteroskedastic features such as non-constant variance, detected through a White test. We used weighted least squares (WLS) for our analysis as it involves modeling variance to correct for this heteroskedasticity. However, our statistical analyses using WLS did not yield any clear results that would allow us to reject the null hypothesis for Hypothesis I.

On the other hand, using inference yielded significant differences between HBCU and non-HBCU resumes based on select variables. A comparison of the mean average ranking in these categories between HBCU applicants and applicants not from HBCUs hints that there could be underlying prestige biases manifested in an unexpected way. Notably, when we compared HBCU and non-HBCU applicants based on the experience type variable, within each category, non-HBCU candidates significantly outranked their HBCU counterparts in every category but the entry-level candidates (see **Figure 6**) . This observation contradicts previous work which states that HBCUs graduates perform better than non-HBCU graduates (Hardy et. al 2019) and warrants further research. Should our observations be confirmed and there is a statistically significant difference between candidates who are in supervisory positions, this would indicate biases in the algorithm with respect to specific variables.

Similar observations were made for the major, experience length, and company size variables. For example, when looking at just the medical field, non-HBCU applicants had significantly better rankings than HBCU applicants (see **Figure 5**). Of note, HBCU applicants outperformed their non-HBCU counterparts in the legal field (see **Figure 5**). We believe this is due to geographical context since Howard University is an HBCU in the Washington D.C. area with a renowned law program.

Our findings indicate that measuring and removing bias from AI systems may not be as straightforward as first imagined. As companies continue to implement AI into their hiring protocols, it is important to consider relationships like these that may only appear after a thorough analysis. People of color have a long history of facing racism in America, and if the American future is increasingly reliant on AI, it is our collective responsibility to develop these systems fairly. We encourage future researchers to continue studying bias in AI systems and ways to make them more equitable.

5.1 Limitations & Challenges

As mentioned above, there was no significant difference found between the distribution of HBCU rankings and non-HBCU rankings. A possible explanation for this is that Indeed bases its rankings on a metric that our research did not consider. Our research considered metrics such as degree type, employment length, and other factors which we believed would indicate more desirable applicants. Another possibility is that Indeed intentionally used metrics that are not influenced by systemic biases to avoid potential discrimination lawsuits. For example, if they based their ranking on the number of fields that were completed by an applicant, this metric would be independent of prestige biases which we believed would influence the algorithm.

Data Collection Limitations

It is possible our data collection methodology impacted our findings. Since data collection always began on the first page of search results, our data are skewed towards ranks 1-100. This may have impacted our ability to judge significant differences in applicant rankings as the data representing ranks 101+, and especially 200+, were scarce. This may have contributed to the heteroskedasticity present in the data, as the resumes within ranks 1-100 may have contained some common attributes which skewed the distribution of specific variables, and could have resulted in us not finding any significant results. A solution to this issue would be to use a random number generator to select which pages data would be collected from.

Additionally, we faced challenges with manual data collection. Ideally, we would have collected thousands of resumes in samples of consistent sizes, but due to the amount of individual effort to record one sample, it was difficult to gather thousands of resumes in one attempt without technological assistance such as web scraping. Thus, the manual data collection procedure limited our ability to collect mass amounts of data. We collected ~900 resumes when we had planned to collect at least 3000 resumes.

As a result of these findings, we could not move on with Hypothesis II due to the aforementioned inability to reject the null hypothesis for Hypothesis I and the lack of training data. The data required to train a robust AI model requires thousands of data points, while the data that we had did not even exceed one thousand. Future research using more extensive and robust data collection methods may enable exploration into Hypothesis II, since we would have more data to train an AI model. However, our current data does not allow for further investigation of this task.

It was also difficult to gather accurate information on company size. We relied on LinkedIn and ZoomInfo for the number of employees, and occasionally the two sites gave different values for the employer size, causing confusion and time spent cross-checking data.

In regards to Indeed's interface, we had to pay close attention to resume rankings since the site does not formally indicate numerical rankings in search results. The number of resumes on each page on Indeed is also inconsistent. Finally, Indeed did not respond to our inquiries regarding the way it ranks resumes, so we could only make inferences on how the algorithm operates based on the metrics that we collected.

Other limitations of our research includes the methods used to address heteroskedasticity in our data. We attempted to account for heteroskedasticity by using a WLS model, but we could have explored using other models such as OLS with heteroskedasticity-consistent standard errors, which would provide a simpler and more understandable analysis. Next, we have issues with post-treatment bias resulting from bad control variables. Some of the control variables that we used are somewhat linked to our dependent variable, which makes it unreasonable to control for them. Changing them to be dependent variables or utilizing other variables as the control would be the more effective method. Finally, we were limited in the method that we used to verify our data, which was lacking due to time constraints. To improve this method, a more rigorous inter-coder reliability score system could be adopted and used to ensure better quality work was being produced among all data encoders.

Another limitation presented by our methodology was that only applicants who had completed a secondary education were recorded in the study. If any applicants had not completed at least a bachelor's degree and had only completed a high school level education, we made the decision to disregard the resumes. This was our method of filtering out minors from our data

pool whose personal identifiable information (PII) would pose a complication with the Institutional Review Board (IRB). However, this limited our sample population. This narrowing of our data sample could have adverse effects on the results we found. For example, there is the possibility that many applicants on Indeed were not considered because they could not afford secondary education. In addition to wealth, there are many other factors which may prevent individuals from pursuing or completing a secondary education, many of which are also strongly correlated with race, the primary variable we wanted to investigate.

5.2 Future Areas of Research

There are many areas of our research that can be explored further. One is the analysis of companies similar to Indeed, such as LinkedIn, Glassdoor, Monster, etc. Conducting similar studies on other platforms would allow comparisons between the ways that resumes are presented and how bias can present across multiple platforms. Preferably, the research is conducted on a site whose resume ranking process is known or is obtainable so that the data collection can be tailored to test it.

Additionally, our research focuses on applicants with a post-secondary degree which naturally excludes a large portion of the population and particularly people of color or other marginalized communities. This issue is further compounded by the fact that people of color tend to be employed by small companies due to fewer barriers of entry compared to larger companies. By gathering data on these populations through techniques such as surveys, we can produce a more accurate representation of bias that may be occurring.

Furthermore, our resumes were collected from a relatively small range of resumes, with most of them within the ranks 1 through 100. Expanding the range to include resumes as far as

rank 500 would allow for more data to be gathered as well as different trends to be revealed, which can give a more holistic picture of the distribution of resumes in Indeed or any other sites that lists resumes. One aspect of the analysis that can be expanded further is measuring the interaction between HBCUs and other variables by adding interaction terms to the model. This would give us a way to see whether HBCUs have a direct correlation with any other variable, allowing for a more precise method to narrow down variables that are significant and simplifying the analysis.

If given more time to analyze professions other than consultants, such as doctors, lawyers, teachers, and more, we would expect to find varying results. This is due to the distinct degree requirements necessary for the respective professions; doctors require a Medical Doctor degree (M.D.), lawyers require a Juris Doctor (J.D.) degree, and teachers require at least a bachelor's degree. Emphasis on educational prestige may vary in the hiring process, depending on the type of profession, which would require an exploration of other variables. Additionally, since our research focuses on consultants within a 25-mile radius in Washington DC, this may cause the results in our data to be dependent on the context of the city. If the same approach was applied to another location, such as Atlanta, Georgia, our ranking results may have produced an entirely different outcome since there are at least 10 HBCUs in the state with potentially different program strengths. Subsequently, a potential research area of interest could be how HBCUs are compared to non-HBCUs with high minority populations. Within HBCUs, there exists variability in student demographics, but a majority of students identify as people of color. However, there also exist non-HBCUs, such as Georgia State University, that have a high minority population which may influence results.

It is important to note that the actual racial demographics of HBCUs are not necessarily majority Black Americans, and their non-HBCU counterparts may in reality have higher rates of Black students. That being said, what we are observing in this study is not a comparison of actual demographics but rather a comparison of the perceived demographics of each college. HBCU colleges were originally established with the intention of serving Black Americans and as such are still associated with the Black population. Hence, we observed how *this* association affects hiring decisions, and not the actual demographics of colleges since they are not commonly known statistics. A future area of research would be examining how the actual racial demographics of colleges influence perceptions and how these perceptions might inform hiring trends.

Chapter 6: Conclusion

Biases exist in a plethora of ways in the hiring process. These biases can ultimately distort hiring outcomes and lead to unintended consequences for those involved. This study explored the ever-present biases that exist in human decision-making that inevitably impact AI hiring systems. We examined potential discrimination in employment hiring site Indeed to determine if the site's applicant ranking algorithm reflects potential prestige bias from its creators. By comparing applicants who attended HBCUs versus non-HBCUs, utilizing variables mentioned in Appendix B, we concluded there was no discernable overall bias in the collected data. However, when we examined individual variables for HBCUs, we found that plausible bias seemed to exist and should be further investigated in the future as it could indicate an overall bias against job applicants who attended HBCUs. Should the existence of these biases be confirmed with a much larger sample size, identifying the root of this bias (whether it is a fault of Indeed's algorithm or a systemic difference in applicant suitability) would be another possible step in making the hiring process equitable for all applicants.

There are a number of possible factors that might produce discriminatory bias and should be further studied in future. As we discussed previously, several studies show that an underlying bias that HBCUs provide a lower quality education compared to non-HBCUs still exists. This field of study and documentation of the related bias remains an unexplored area of research that should be further investigated to ensure that the foundations of AI in hiring are unbiased and equitable.

References

Ajunwa, I. (2018). *Age Discrimination by Platforms* (SSRN Scholarly Paper No. 3142979). Social Science Research Network. https://papers.ssrn.com/abstract=3142979

Allen, Walter. R. (1986). Gender and Campus Race Differences in Black Student Academic Performance, Racial Attitudes, and College Satisfaction. Atlanta, GA: Southern Education Foundation. https://education.stateuniversity.com/pages/2046/Historically-Black-Colleges-Universities.html

Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., & Rus, D. (2019). Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society,* 289–295. https://doi.org/10.1145/3306618.3314243

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang Y. (2019). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1-4:15. https://doi.org/10.1147/jrd.2019.2942287

Bevins, F., Fox, K., & Pinder, D. (2021, July 30). *How HBCUs can accelerate Black economic mobility*. McKinsey & Company. Retrieved April 4, 2022, from https://www.mckinsey.com/industries/education/our-insights/how-hbcus-can-accelerate-black-economic-mobility

Borgesius, F. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. Directorate General of Democracy, Council of Europe, 51.

Buolamwini J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research, 81*.

Commodore, F., & Njoku, N. R. (2020). Outpacing Expectations: Battling the Misconceptions of Regional Public Historically Black Colleges and Universities. *New Directions for Higher Education*, *2020*(190), 99–117. https://doi.org/10.1002/he.20370

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved September 11, 2021, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

Derous, E., & Ryan, A.. (2018). When your resume is (not) turning you down: Modeling ethnic bias in resume screening. *Human Resource Management Journal, 29(*2), 113-130. https://doi.org/10.1111/1748-8583.12217

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and Mitigating Unintended Bias in Text Classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society,* 67–73. https://doi.org/10.1145/3278721.3278729

Fogle, C. D. (2011). *Employers' Perceptions of Business Graduates From Historically Black Colleges and Universities* [Ph.D., Walden University]. Retrieved March 9, 2022, from https://www.proquest.com/docview/862090154/abstract/DA5E5CFB36A540CAPQ/1

Hardy, Precious M., Elizabeth J. Kaganda, and Mara S. Aruguete. (2019). Below the Surface: HBCU Performance, Social Mobility, and College Ranking. Journal of Black Studies 50 (5): 468–83. https://doi.org/10.1177/0021934719847910

Harlan, E., & Schnuck, O. (2021, February 16). Objective or biased. Bayerischer Rundfunk. Retrieved March 27, 2023, from https://interaktiv.br.de/ki-bewerbung/en/

Jackson, T. (2007). *Employer perceptions of technology graduates from historically Black colleges and universities: A Q methodological study* [Ph.D., Walden University]. Retrieved March 9, 2022, from https://www.proquest.com/docview/304767319/abstract/46AE2D8C9ED419BPQ/1

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. In P. A. Flach, T. De Bie, & N. Cristianini (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 35–50). Springer Berlin Heidelberg.

LeCun, Y., Bengio, Y., & Hinton, G. (2015, May 27). Deep Learning. *Nature News*. Retrieved September 11, 2021, from https://www.nature.com/articles/nature14539

Long, R. G. (2008). The crux of the method: assumptions in ordinary least squares and logistic regression. *Psychological Reports*, *103*(2), 431-434.

Mayson, S. G. (2019). Bias In, Bias Out. *Yale Law Journal, 128*(8): 2218–2300. https://www.yalelawjournal.org/article/bias-in-bias-out

McDuff, D., Cheng, R., & Kapoor, A. (2018). Identifying Bias in AI using Simulation, 1810.00471, 11. Retrieved from https://arxiv.org/pdf/1810.00471.pdf

Raub, M. (2018). Bots, Bias, And Big Data: Artificial Intelligence, Algorithmic Bias And Disparate Impact Liability In Hiring Practices. *Arkansas Law Journal, 17*(2), 529-570. https://scholarworks.uark.edu/alr/vol71/iss2/7

Rivera, L. A. (2011). Ivies, extracurriculars, and exclusion: Elite employers' use of educational credentials. Research in Social Stratification and Mobility, 29(1), 71–90. https://doi.org/10.1016/j.rssm.2010.12.001

Ryan, T. P. (2008). *Modern regression methods* (Vol. 655). John Wiley & Sons.

Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020, January 1). What does it mean to 'solve' the problem of discrimination in hiring?: Proceedings of the 2020 conference on fairness, accountability, and transparency. ACM Conferences. Retrieved March 28, 2022, from https://dl.acm.org/doi/abs/10.1145/3351095.3372849

Volkwein, J.F., Sweitzer, K.V. Institutional Prestige and Reputation among Research Universities and Liberal Arts Colleges*. *Res High Educ* 47, 129–148 (2006). https://doi.org/10.1007/s11162-005-8883-5.

Appendices

Appendix A: Data Collection Procedure

1. Prepare a spreadsheet to store the data.
2. Create an account on Indeed.com.
3. Select “employer account” before purchasing a standard subscription.
4. After making an account, navigate to the “search resumes” section and create a new project under the “Project” tab. This is where we collect resumes for future reference.
5. Go to the “Search” tab and conduct a search with “consultant” for the job title and “District of Columbia” for the city. In addition, toggle the option to limit the search to the job title.
6. After running the search, be sure to limit the search radius to 25 miles in the left sidebar.
7. Look at the first resume to see if it has already been added to your project.
    a. If it has not been added, add it to the project and record the values of the variables specified in Appendix B in a new row in the spreadsheet. Download the resume in a secure location and rename it to match the assigned resume_id as specified in Appendix B.
        i. Refer to https://sites.ed.gov/whhbcu/one-hundred-and-five-historically-black-colleges-and-universities/ to see whether a college is an HBCU or not. Further research may need to be done online since some HBCUs may have closed and do not show up on the website.
            1. Please note, the original HBCU reference site in which we made comparisons is now host to a scam site. The link was changed, so the results may not align perfectly.

ii. Ignore any resumes with the following features:

1. Highest education is a high school diploma or GED.
2. No education listed.
3. Only has certifications or partial degrees.
4. Only has a degree listed without a major specified (bachelor, master, etc).

iii. To determine the value for the "employer" variable, use LinkedIn and ZoomInfo to get the number of employees and categorize the resumes.

1. Look up the company on both sites. If the company only shows up on one site, use that site to categorize the company according to the ranges listed for the "employer" variable
    a. If the company does not show up on either site, locate the company's website to see if they have listed a figure for the number of employees.
    b. If the company website does not provide any conclusive numbers, record down "1" if the company seems to be composed of only one individual and "2" otherwise.
2. If both sites have the company listed, verify that they have similar ranges for the companies. Otherwise, give precedence to the site with a more narrow range.
3. If any ranges listed would overlap the ranges listed under the "employer" variable, defer to the smaller range.

iv. If the resume is missing any variable from Appendix B, make a note in the appropriate spreadsheet cell. The resume will be removed from the final data set if the variable cannot be found.

v. The rank starts from 1, which represents the best rank. As you go down the page, increment the rank by 1 for each resume that has been processed before in the same session.

b. If the resume has been added, check if the update date has changed from the latest entry.

i. If the update date has changed, add an additional row beneath the previous entry and repeat step 6a in recording values and downloading the resume.

ii. Otherwise, add an abridged resume_id, without the 4-digit sequential number, behind the original entry. For example, if the original entry was "*20220404Smith0001*" and another instance of the resume was found with no change in the update date, the new entry should be "*20220404Smith001, 20220405Smith*." Note that the YYYYMMDD date should be whichever date the resume was reencountered.

8. Repeat step 6 for 100 resumes per session per day.
9. Repeat steps 7 for 30 days for a total of about 3000 resumes.

Appendix B: Variables Collected

1. **resume_id** Reference number assigned to each resume by researcher. The reference ID is constructed in the following format: YYYYMMDD[Last Name of Researcher][4-digit sequential number]. (*EX: 20220404Smith0001*)

2. **update** Latest update of the resume. Enter the difference of days between the date resume was last updated compared to the date resume is downloaded. (*EX: date of update = January 1, 2022; date of download = January 2, 2022; value entered for variable* = 1)

3. **rank** Rank of the resume. Enter a numerical value that represents the rank of the resume

4. **hbcu** Applicant attended HBCU.

Yes = 1

No = 0

5. **college_type** College type applicant attended.

HBCU public university = 1

HBCU private university = 2

Non-HBCU public university = 3

Non-HBCU private university = 4

6. **degree** Highest post-secondary degree obtained

Bachelor's Degree = 1

Master's Degree = 2

Professional Degree (e.g., MBA, JD, MD) = 3

Doctoral Degree = 4

7. degree_type Type of highest post-secondary degree obtained

Arts = 1

Business/Finance = 2

Science = 3

Engineering = 4

Medical Field = 5

Legal = 6

Other = 7

8. experience_type Highest type of applicant experience

Non-supervisory, entry = 1

Supervisory, first line = 2

Non-supervisory, experienced = 3

Supervisory, mid-level = 4

9. experience_length Length of applicant experience in most recent position

Enter length in months (*EX: 24 = 2 years; 18 = 1 ½ year, etc.*)

10. career_length Length of applicant's overall experience in the field

Enter length in months (*EX: 24 = 2 years; 18 = 1 ½ year, etc.*)

11. employment Is the applicant currently employed

Unemployed = 0

Employed = 1

12. employer To the best ability, identify the size of the employer applicant has mostly been employed with

Self-employed = 1

Small private office (e.g., a doctor's office) = 2

Small business (with 20 – 200 employees) = 3

Mid-size business (with 201 – 1,000 employees) = 4

Large business (1,000-10,000 employees) = 5

Well known large corporations (e.g., Google, Ford, Boeing, etc. - 10,000+) = 6