Report on the Impact of Ethnicity in Hiring Practices: A Data-Driven Analysis

Introduction and Aims

Discrimination in hiring practices, particularly with respect to perceived ethnicity, remains a pervasive concern across many industries. Numerous studies have shown that names or other identifying attributes on resumes can influence employers' perceptions, often resulting in unfair treatment. The aim of this research project is to investigate whether the ethnicity signaled by a candidate's name, along with other relevant factors (such as experience level, education, and quality of resume content), correlates with their likelihood of receiving a callback. The overarching objectives include:

1. Defining and clarifying how perceived ethnicity might influence hiring practices in multiple industries.
2. Acquiring and processing a suitable dataset that simulates real-world resumes and callback outcomes.
3. Conducting an exploratory data analysis (EDA) in a Jupyter Notebook environment to identify patterns, distributions, and potential biases.
4. Evaluating whether the dataset supports conclusions about discriminatory practices and how these might vary by industry.
5. Proposing recommendations for mitigating implicit biases in hiring, while demonstrating a thorough data processing pipeline, from cleaning to final analysis.

This document provides a comprehensive overview of how the research was designed and executed, following a systematic approach. By demonstrating each step—acquiring the data, cleaning it, exploring potential biases, and presenting the results—this report confirms the viability of analyzing hiring discrimination through a data-driven lens. In particular, we seek to show that ethnicity, as implied by candidate names, is a substantial factor influencing callback rates, alongside commonly expected predictors such as experience and qualifications.

The dataset used for this research stems from a **l**andmark field experiment conducted by Bertrand and Mullainathan (2004). In this randomized controlled experiment, 4,870 fictitious résumé**s** were sent in response to real employment advertisements in Chicago and Boston during 2001. Because ethnicity is not typically disclosed on a résumé, the researchers differentiated between "Caucasian-sounding names" (such as Emily Walsh or Gregory Baker) and "African American-sounding names" (such as Lakisha Washington or Jamal Jones). A large collection of fictitious résumés were thereby created, and the pre-supposed ethnicity of each résumé—based on the sound of the name—was randomly assigned. These résumés were then sent to prospective employers to see which résumés generated a phone call or callback.

1. **Why This Data Is Appropriate**

- **Relevance**: It directly pertains to our research question regarding perceived ethnicity and callback rates. This is more suitable than general employment surveys lacking precise control of résumé attributes.
- **Experimental Control**: By randomly assigning ethnicity-based names, the experiment minimizes confounding factors, thus enabling clearer insights into name-based hiring discrimination.
- **Structure**: Provided in a **CSV** format, each row corresponds to a single fictitious résumé. Columns capture attributes like *experience*, *gender*, *industry*, *ethnicity*, and *callback status*. This tabular structure is ideal for both numeric and categorical analysis in Python's pandas DataFrame.

2. **Comparison to Other Potential Datasets**
   - **General Social Survey (GSS)**: While broad in capturing public attitudes, it lacks the controlled résumé-based approach that is crucial for isolating name-based discrimination.
   - **LinkedIn Hiring Insights**: Potentially large-scale, but confounded by user-driven profile differences, personal network effects, and lack of random assignment. Data access is also restrictive.

By contrast, the Bertrand and Mullainathan (2004) dataset is profoundly suitable for pinpointing how name-based signals affect callbacks. The carefully controlled nature of fictitious résumés helps researchers infer causal relationships more confidently.

Research Significance

Hiring discrimination is a domain of both social and economic importance. Biased hiring decisions can perpetuate inequality, stifling opportunities for marginalized groups and diminishing the overall pool of talent. This project addresses a notable gap: while anecdotal evidence abounds, robust, large-sample data from a controlled experiment is rarer. By focusing on how perceived ethnicity—signaled through first and last names—shapes callbacks in multiple industries, the research highlights a mechanism of discrimination that often remains hidden.

This dataset has not been exhaustively examined for multi-industry differences. Thus, investigating whether certain industries exhibit stronger or weaker ethnic biases extends the conversation beyond a single aggregated result. We have, however, bounded our scope to focus on a core set of variables: *ethnicity, gender, experience, résumé quality, college attendance,* and *callback outcome*. While additional columns exist (such as volunteer experience or military background), the chosen scope keeps the analysis both tractable and aligned with the primary research question.

Data Processing and Exploration Pipeline

1. **Data Loading**
   The CSV file was read into Python using `pandas`. Each row represents a single fictitious résumé sent to an employer, with columns including **experience** (years of experience), `ethnicity` (Caucasian vs. African American sounding), `gender`, `call` (yes or no), `quality` (high or low résumé), and **industry**. Given the controlled design of the original experiment, the data was already relatively clean.
2. **Data Cleaning and Validation**
   - **Missing Values**: The dataset was large enough (4,870 résumés) to handle minimal missing data without distorting results. Rows with critical missing fields—such as **call_numeric**—were dropped or verified.
   - **Label Encoding**: Converting text fields (e.g., "male/female," "caucasian/non-caucasian," "low/high quality") to numeric values (0/1) allowed logistic regression to be performed.
   - **Regex Checks (if needed)**: Numeric fields (like **experience**) were verified to ensure no stray characters existed, ensuring correct data types (**int64**).
3. **Feature Engineering**
   We created binary columns—**gender_male** (1 for male, 0 for female) and **ethnicity_cauc** (1 for Caucasian, 0 otherwise)—from the original textual variables. We also used **quality** as a numeric indicator (0 for low, 1 for high résumé quality). In some advanced steps, an interaction term (**gender_ethnicity_interaction**) was generated to see if the combination of being male and having a Caucasian-sounding name yielded distinct outcomes.
4. **Exploratory Data Analysis**
   - **Descriptive Statistics**: For numeric columns (**experience**, **call_numeric**), we computed means and standard deviations. The data revealed that overall callback rates hovered around certain percentages, with initial signs that résumés signaled as Caucasian had higher callback rates.
   - **Pivot Tables**: We grouped or pivoted data by **industry** and **ethnicity_cauc**, computing mean callback rates (**call_numeric**). This approach revealed consistent differences favoring Caucasian-labeled résumés across industries like finance, trade, manufacturing, business/personal services, and more.
   - **Visualizations**: Simple bar plots and grouped bar charts helped highlight differences across categories, while logistic regression offered deeper insight into the magnitude of those differences.
5. **Logistic Regression Modeling**
   - **Model Specification**: A simple logistic regression model was chosen as given the data it was sure to have a high accuracy rate and it was not overly complicated and difficult to implement.
   - **Results**: The coefficient for **ethnicity_cauc** was a significant positive predictor of receiving a callback, suggesting substantial name-based differences. Meanwhile, **experience** also showed a strongly significant positive effect, but **gender_male** and **college** were not statistically significant in this model. The pseudo R-squared was modest (~0.0134), common for social-science datasets

Ethical, Licensing, and Provenance Considerations

Because the names in the dataset are fictitious, there is no risk of revealing personally identifiable information for real individuals. The data is openly licensed for academic research, which is appropriate for the goals of this project. However, it is paramount for researchers to be aware that highlighting ethnic disparities could have reputational impacts on certain industries if interpreted without caution. The dataset is explicitly anonymized with respect to employers, so no single company can be identified. This anonymity is valuable for reducing potential harm.

Further, by pointing out differences in callback rates for “Caucasian-sounding” versus “African American-sounding” names, the research underscores the existence of potential discriminatory biases. Yet it does not, by itself, detail the internal decision-making processes of each employer. Therefore, *nuanced interpretation* is critical: the data suggests bias, but the exact roots in any particular company’s HR process remain unknown.

Data Modification and Added Value

- **Encoding**: Mapping “male/female” and “caucasian/non-caucasian” to binary variables was essential for logistic regression and other numeric-based analyses.
- **Dropping Incomplete Rows**: This step ensured the modeling approach would not be skewed by missing data in key fields, like **call_numeric** or **experience**.
- **Data Integration**: While the dataset came from a single experiment, we ensured that it could be extended to a machine-learning pipeline if desired, since it is now neatly arranged with numeric columns for each relevant variable.

These transformations significantly improved the dataset’s compatibility with standard analytical and plotting libraries, allowing for a consistent pipeline from raw CSV to final regression results.

Key Findings and Implications

1. **Ethnicity as a Significant Predictor**
   Logistic regression results show that “Caucasian-sounding” names (encoded as **ethnicity_cauc = 1**) have a **statistically significant** and **positive** coefficient, suggesting greater odds of receiving a callback compared to “non-Caucasian-sounding” names. This effect, observed across thousands of résumés, indicates the presence of an ethnic bias in initial hiring decisions.
2. **Experience**
   Another significantly positive predictor, experience indicates that each additional unit of

experience (e.g., year) increases the probability of a callback, aligning with general expectations that more qualified candidates typically stand out.

3. **Gender and College**
   Neither “male vs. female” nor “college vs. no college” emerged as significant factors in this specific logistic model. This could be due to the design of the experiment (where only name changes were systematically manipulated for bias detection) or broader norms in the dataset. While interesting, it warrants further research to parse out if other aspects of the résumé overshadow these variables.
4. **Industry Comparisons**
   Pivot tables split by industry and ethnicity underscore that the Caucasian group consistently shows higher callback rates across multiple fields. This uniform pattern suggests that name-based discrimination is not confined to one sector. Some industries, like finance/insurance/real estate, appear to exhibit larger disparities, while others show smaller (but still present) gaps.
5. **Limitations**
   - The data captures conditions in Chicago and Boston during 2001, so generalizations to other time periods or cities may require caution.
   - The modest pseudo R-squared indicates many aspects of hiring decisions remain unexplained (interview performance, personal networks, intangible qualities, etc.).
   - While random assignment of names strongly isolates name-based discrimination, it also means some typical real-world variations (like different job fields) may not be fully controlled for if the experiment had partial coverage of industries.

Conclusion and Future Directions

Summarizing the evidence gathered:

- The logistic regression strongly supports the presence of name-based bias, with résumés perceived as Caucasian more likely to receive callbacks.
- Experience also matters significantly, but gender and college do not demonstrate large or statistically significant effects in this dataset.
- Industry-wide pivot tables suggest the bias is consistently present but varies in magnitude.

For future work, researchers could incorporate more sophisticated analyses such as multilevel modeling (treating industry as a random effect) or augment the dataset with additional real-time sources to confirm whether these patterns hold up beyond Chicago and Boston. Moreover, investigating whether volunteer experience or specialized skill sets moderate these biases would deepen our understanding of the interplay between résumé attributes and discriminatory practices.

From an applied standpoint, these findings highlight the need for bias mitigation in recruitment. Techniques such as name-blind recruiting, standardized interview questions, and active training against unconscious bias may help. Policymakers, companies, and advocacy groups can utilize these results to design interventions that promote fair and inclusive hiring processes.

In closing, the well-structured pipeline—covering data acquisition, cleaning, exploratory analysis, modeling, and interpretative discussion—showcases the power of data science in exposing subtle but pervasive issues of discrimination in labor markets. While more work remains, especially in refining models and broadening the data scope, the conclusion is clear: **perceived ethnicity, indicated merely by a résumé's name, exerts a real and significant influence on hiring callbacks.**