DNSC 6317 - Business Analytics Practicum

Project Title: Using ACS Data to Understand the Demographic and Socioeconomic Makeup of Biomedical Research Workers in the United States

Group Members:

- Abdul Haleem Abdul Salam
- Sai Nityamani Sahith Matsa
- Yiying Niu

Executive Summary:

The National Institute of General Medical Sciences (NIGMS) seeks to enhance its understanding of the demographic and socioeconomic characteristics of the U.S. biomedical research workforce to inform programmatic decisions and foster diversity. This report presents findings from our analysis of American Community Survey (ACS) data, aiming to provide insights into the makeup of biomedical research workers across different demographics, industries, and geographic locations. Our methodology involved descriptive analysis using weighted survey responses and, time permitting, regression analysis to identify trends. The results offer valuable information for NIGMS to refine its strategies and initiatives to support a vibrant and inclusive biomedical research community.

Problem Understanding:

NIGMS administers numerous programs to support the biomedical research workforce but lacks detailed demographic and socioeconomic data on this population. This knowledge gap impedes the evaluation of program effectiveness and the development of targeted interventions. Our project addresses this challenge by leveraging ACS data to estimate descriptive statistics, enabling NIGMS to better understand the characteristics and needs of the biomedical research workforce.

Methodology:

Data Analyzed:

We utilized the Census Bureau's American Community Survey (ACS) Public Use Microdata Sample (PUMS), focusing on occupation and industry codes relevant to biomedical research workers.

Analytics Techniques Used:

- Descriptive Analysis: We employed weighted survey responses to estimate demographic and socioeconomic statistics, stratified by industry and state.
- Regression Analysis (Time Permitting): Regression models were utilized to identify potential trends or significant changes in descriptive statistics over time.

Detailed Methodology:

1) **Data Acquisition: Downloading and Dataset Compilation**

- The dataset acquisition phase commenced with retrieving pertinent data from the American Community Survey (ACS) and the associated Public Use Microdata Sample (PUMS).

- Specifically, the 2022 5-year data files were obtained, consisting of five distinct files: psam_pusa, psam_pusb, psam_pusc, psam_pusd, and psam_puse.
- These files were meticulously appended to compile a comprehensive dataset for subsequent analysis.

2) Data Familiarization and Variable Selection

- A critical initial step involved thorough immersion in the ACS questionnaire to grasp the intricacies of the data collection process and comprehensively understand the variables therein.
- Subsequently, a judicious selection process was undertaken to identify and retain variables deemed essential for our analysis. This process entailed dropping extraneous variables to streamline the dataset and focus solely on those relevant to our research objectives.

3) Occupation Identification and Data Preparation

- Five distinct occupations pertinent to our analysis were meticulously identified and isolated:
    - 1610: Biological Scientists (All Others)
    - 1650: Medical Scientists (Except Epidemiologists)
    - 1720: Chemists
    - 1910: Biological Technicians
    - 1920: Chemical Technicians
- Following occupation delineation, the dataset underwent meticulous preparation. This involved the addition of a 'year' column, ingeniously derived from the serial number, facilitating temporal analysis and trend identification.

Incorporation of Person Weights to Rectify Sampling Biases:

- Person weights were meticulously incorporated into the dataset to address potential sampling biases and enhance data representativeness.
- This process involved assigning a weight to each individual observation based on their demographic characteristics and survey response patterns.
- For instance, if an individual's person weight was 3, it indicated that their responses were representative of three individuals in the population. As a result, the corresponding row in the dataset was replicated three times to reflect this weighting.
- By accounting for variations in survey response rates and demographic distributions, the inclusion of person weights ensured that the analysis accurately represented the diversity of the biomedical research workforce.

Adjustment of Wage Data to Account for Inflationary Factors:

- Wage data underwent meticulous adjustment to mitigate the impact of inflationary factors and provide a more accurate depiction of income disparities over time.
- This adjustment process entailed multiplying the wage data by an inflation adjustment factor, derived from relevant economic indicators or inflation indices.
- The resulting 'income adjusted' column was created to capture the inflation-adjusted income levels for each observation in the dataset, thereby accounting for changes in purchasing power over time.

- By incorporating inflation-adjusted income values, our analysis offered a more nuanced understanding of income trends and disparities within the biomedical research workforce, facilitating informed decision-making and policy formulation.
- To maintain analytical rigor, only individuals classified as 'civilian employed' were retained, with other employment statuses filtered out under the 'esr' variable.
- A comprehensive correlation matrix was then deployed to identify highly correlated variables. This facilitated the selection of key variables essential for subsequent analysis.

4) Variable Selection and Refinement

- Post-data wrangling, the dataset underwent a meticulous variable selection process. This entailed retaining variables deemed critical for our analysis objectives while discarding redundant or extraneous variables.
- The final list of variables selected for analysis encompassed a diverse array of dimensions, including demographic, socioeconomic, and temporal factors. These variables included:
    - ST (State)
    - Agep (Age)
    - Cit (Citizenship)
    - Cow (Class of Worker)
    - schl (Educational Attainment)
    - sex (Gender)
    - dis (Disability Status)
    - esr (Employment Status Recode)
    - hisp (Hispanic Origin)
    - occp (Occupation)
    - rac1p (Race)
    - sciengp (Science and Engineering Degree)
    - year (Year)
    - Income_adj (Adjusted Income)

5) Analysis and Visualization Strategies

- With the dataset meticulously prepared and refined, the analysis phase commenced in earnest. A plethora of analytical techniques were judiciously employed to derive meaningful insights and uncover hidden trends.
- Graphs and visualizations served as indispensable tools in illustrating key findings. A diverse array of visual outputs was generated, focusing on critical dimensions such as gender distribution, income disparities, educational attainment, and racial representation.
- Central to our analytical approach was a comprehensive descriptive analysis, which provided a robust foundation for understanding the demographic and socioeconomic characteristics of the biomedical research workforce. The descriptive statistics obtained served as valuable inputs for subsequent interpretation and decision-making.
- To augment the interpretability and accessibility of our analysis results, table shells were meticulously developed to organize the descriptive statistics for each occupation. This facilitated a more structured and systematic approach to data interpretation.

6) **Dashboard Development and Regression Analysis**

- Leveraging the insights gleaned from our analysis, comprehensive Tableau dashboards were meticulously crafted to encapsulate key findings and trends. These interactive dashboards provided stakeholders with an intuitive interface for exploring and interrogating the data.
- In addition to descriptive analysis, regression models were judiciously employed to identify statistically significant variables and elucidate their impact on the demographic and socioeconomic characteristics of the biomedical research workforce. This regression analysis provided valuable insights into the underlying drivers of workforce dynamics and facilitated informed decision-making.

7) **Observations and Comparative Analysis**

- The culmination of our analysis efforts entailed a meticulous examination of analysis results at the state level. These observations were juxtaposed against national averages, with a particular focus on mean income, gender-based income disparities, and gender distribution. This comparative analysis provided valuable insights into regional variations and disparities within the biomedical research workforce.

Observations from Descriptive Analysis:

1. **Gender and Income Disparities by State:**

    - Females exhibit a higher mean income than males in several states, notably Georgia, Mississippi, Maryland, Nebraska, New Mexico, and South Dakota.

    - Upon examining technician predictions, females tend to have higher incomes than males in the majority of states. However, in the field of biochemical sciences, males generally outperform females across most states.

2. **Occupational Analysis:**

    - Occupation 1720 (Chemists) presents an anomaly with a zero male count in Washington D.C. (DC). Consequently, females in states like Idaho and Kentucky earn higher incomes, despite lower counts than males. Similar situations arise in Nebraska, where females earn 1.5 times more than males but represent half the count.

    - In occupation 1910 (Biological Technicians), females and males in Florida receive comparable incomes, although female counts are lower overall. Females outnumber males by 2000 in total count.

    - For occupation 1920 (Chemical Technicians), males outnumber females twofold, yet females earn higher incomes in the majority of states due to their experience in the field.

    - In occupation 1610 (Biological Scientists), both genders hold bachelor's degrees or higher. However, males tend to dominate with experience, leading

to higher average incomes compared to females. With increasing age, males surpass females in income after age 41.

3. **Race and Employment Dynamics:**

    - Among Alaska Native individuals, females are more prevalent in employment and often hold scientist roles.

    - Asians exhibit higher incomes compared to other racial groups, followed by White individuals. Conversely, American Indians and Native Alaska tribes have lower performance relative to other races.

4. **Educational Attainment and Income Disparities:**

    - The majority of individuals, both males and females, hold bachelor's degrees or higher across all occupations. Exceptions exist in occupations other than 1610, where lower educational attainment correlates with lower income, albeit with smaller sample sizes.

    - Notably, individuals with professional degrees earn higher incomes than those with other degrees, including doctorates. This trend is attributed to their experience in the field, as individuals with higher age tend to have greater experience.

Results, Conclusions, and/or Recommendations:

Our analysis revealed significant insights into the demographic and socioeconomic makeup of the U.S. biomedical research workforce. Key findings include variations in income distribution, industry representation, and geographic concentration. These insights can inform NIGMS's decision-making processes and facilitate the development of targeted interventions to address disparities and promote diversity in the biomedical research field.

Potential Next Steps:

To build upon our findings and further enhance understanding, we recommend the following next steps:

- Longitudinal Analysis: Conducting longitudinal analyses to track trends and changes in the demographic and socioeconomic characteristics of the biomedical research workforce.
- Data Integration: Exploring additional datasets or integrating qualitative research to gain deeper insights into the drivers of workforce dynamics.
- Stakeholder Collaboration: Collaborating with relevant stakeholders, including academic institutions and industry partners, to tailor interventions and initiatives to specific workforce needs and challenges.

Appendices:

Code Documentation:

Attached are the codes developed for data analysis using Python programming language, ensuring transparency and reproducibility of our methodology. The code documentation provides detailed explanations of the data preprocessing, analysis steps, and visualization techniques employed in the project.

Presentation Slides:

Included in this appendix are the PowerPoint slides prepared for the presentation of our findings to NIGMS leadership and other stakeholders. The slides summarize key insights, methodologies, and recommendations derived from our analysis of ACS data on the demographic and socioeconomic characteristics of the U.S. biomedical research workforce.

Tableau Dashboard:

We have created a Tableau dashboard to visualize the key findings and trends uncovered in our analysis. The interactive dashboard allows users to explore various demographic and socioeconomic metrics, stratified by industry and state, providing a user-friendly interface for accessing and interpreting the data.

Additional Visualizations:

Supplementary tables and figures are provided to offer further detail on the analysis results. These visualizations enhance understanding of the distribution and trends in the demographic and socioeconomic attributes of biomedical research workers in the United States.

Conclusion:

Our analysis of American Community Survey (ACS) data has provided valuable insights into the demographic and socioeconomic characteristics of the U.S. biomedical research workforce. Through meticulous data analysis using Python programming language, we were able to stratify the workforce by industry and state, offering a nuanced understanding of its composition.

Key findings include disparities in income distribution, variations in industry representation, and geographic concentrations of biomedical research workers. These insights underscore the importance of targeted interventions to address disparities and promote diversity within the biomedical research field.

Our recommendations for future action include conducting longitudinal analyses to track trends over time, integrating additional datasets to gain deeper insights, and collaborating with stakeholders to tailor interventions to specific workforce needs. Furthermore, the development of interactive Tableau dashboards and supplementary visualizations enhances the accessibility and utility of our findings for NIGMS and other stakeholders.

In conclusion, our project not only contributes to the understanding of the biomedical research workforce but also provides actionable insights to inform strategic decision-making and program development aimed at fostering a vibrant and inclusive research community. By leveraging data-driven approaches, we can work towards a more equitable and innovative biomedical research landscape.