A Study on the Census of the U.S. Labor Market and Income Distribution Based on Big Data

Jutong Chen *

The Frederick Gunn School, Washington, Connecticut, United States

* Corresponding Author Email: 26chenj@frederickgunn.org

Abstract. The US labor market has always been plagued by wage inequality due to differences in workers' gender, race, and educational background. This paper uses the 1994 US Census dataset (32,561 people) to analyze the impact of US workers' demographic factors on their income and working hours using Pearson correlation analysis and T-test. The statistical results show that higher education levels are significantly correlated with higher incomes ($r=0.45$), white people occupy high-paying jobs in the United States, and minorities do not account for a high proportion of high-level positions; men work an average of 42.4 hours per week, while women work 36.4 hours ($p<0.01$). These gaps reflect differences and unfairness in income caused by racial discrimination, unequal educational opportunities, and excessive family care responsibilities for women. Comparative analysis of policies in various countries shows that intervention measures such as childcare subsidies and vocational skills training programs in the European system can help reduce inequality in the US labor market. The results revealed the inequality in the US labor market and point out the need for the United States to promote citizens' fair participation in the labor market through targeted policy measures.

Keywords: U.S. Census; labor market inequality; income distribution.

1. Introduction

The long-standing structural inequality problem in the U.S. labor market has been a core issue of concern to economists and government departments [1]. Despite the implementation of multiple government bills and corporate diversity policies since the 20th century, employment outcomes and post-employment salary differences caused by factors such as gender, race, and educational background have always existed [2]. These inequalities not only violate the principle of economic fairness, but also reflect the inefficient allocation of human resources, which will further restrict the growth of productivity in the future.

In recent years, many scholars have made more significant progress in their understanding of the structure of the labor market. Geisterfer-Black et al. [3] used big data analysis to prove that the economic recession caused by the COVID-19 pandemic has a high negative impact on the employment of women, ethnic minorities, and low-educated workers, further increasing the imbalance of market wages. Alam et al. [4] conducted a longitudinal analysis of the market and believed that there is a significant positive correlation between education level and employment stability, and the popularization of higher education can resist market fluctuations. Rahhal et al. [5] and Park et al. [6] used machine learning methods to analyze labor economics. Such calculation methods have somewhat ignored the impact of institutional and structural inequality on the labor market.

Notable gaps persist in the extant literature. First, most studies adopt unidimensional analytical frameworks that examine singular demographic factors in isolation, failing to capture intersectional effects. Second, the predominant focus on wage differentials has come at the expense of rigorous examination of working time disparities, despite their substantial implications for lifetime earnings and career trajectories. Third, insufficient attention has been paid to the institutional mechanisms that transform individual demographic characteristics into systemic labor market disadvantages.

This study addressed these limitations through comprehensive analysis of the 1994 Current Population Survey [6], employing both descriptive and inferential statistical techniques to examine

how multiple axes of demographic difference collectively shape labor market outcomes. The analytical approach of this paper generated three principal contributions: (1) quantification of intersectional effects on both earnings and working time allocation; (2) identification of structural mediators that convert demographic characteristics into labor market disadvantages; and (3) development of a unified framework for understanding how institutional, cultural, and economic factors interact to produce persistent inequalities.

The significance of this research extends beyond academic circles. Our findings provide empirical grounding for targeted policy interventions aimed at reducing labor market disparities, particularly in the areas of educational access, workplace flexibility, and anti-discrimination enforcement. Moreover, the study offers methodological insights regarding the complementary roles of traditional econometric techniques and emerging computational methods in labor market analysis.

2. Method

2.1. Data collection

This study utilized the Census Income dataset, which is publicly available from the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/dataset/20/census+income) [6]. The dataset, extracted from the 1994 Census by Barry Becher, consists of 32,561 records and includes key demographic and employment-related variables. The extraction process ensured data cleanliness using the following conditions: (AAGE>16) & (AGI>100) & (AFNLWGT>1) & (HRSWK>0).

The dataset includes the following variables: Age: Continuous variable ranging from 17 to 90 years; Education level: Categorical variable representing different education levels. Hours Worked Per Week: Continuous variable measuring work time. Capital Gain and Loss: Financial indicators representing economic status. Sex: Categorial variable (male/female). These variables were selected to analyze labor market characteristics, focusing on disparities in work hours, income, and education levels.

2.2. Data analysis

To ensure the accuracy and credibility of the results, all statistical analyses in this study were carried out using SPSS software. Two core methods were employed: Pearson correlation analysis and the independent samples t-test. These techniques were chosen to investigate the relationships between key variables such as age, gender, and weekly working hours, as well as to identify meaningful differences between demographic groups.

Pearson correlation analysis was used to measure the linear association between two continuous variables. The strength and direction of this relationship are represented by the Pearson correlation coefficient, denoted as r, which ranges from -1 to 1. A value close to 1 indicates a strong positive relationship, meaning that as one variable increases, the other tends to increase as well. A value close to -1 suggests a strong negative relationship, where an increase in one variable tends to correspond with a decrease in the other. A value near 0 indicates no clear linear relationship.

The Formula for calculating the Pearson correlation coefficient is as follows:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \quad (1)$$

In this formula, x and y represent the values of the two variables being compared, while $\bar{x}$ and $\bar{y}$ are their respective means. The numerator calculates the covariance between the variables, and the denominator standardizes this value by the product of their standard deviations. This method was used to explore the correlation between age and hours worked per week, as well as other demographic and employment-related factors.

The independent samples t-test was applied to compare the average number of working hours per week between male and female participants. This test evaluates whether the difference in means between two groups is statistically meaningful. The t-test formula is:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (2)$$

In this equation, $\overline{x_1}$and $\overline{x_2}$ are the sample means, s_1^2 and s_2^2 are the sample variances, and, are the respective sample sizes for the two groups. This analysis helped determine whether gender plays a significant role in the number of hours worked weekly.

By applying these methods, this study aimed to uncover statistically sound relationships and differences in labor market behaviors across demographic categories. The results offer a data-driven foundation for understanding structural trends within the U.S. workforce.

3. Result

3.1. Labor market demographic

The dataset provides valuable insights into the structure of the U.S. labor market [6]. Figure 1 illustrates an overview of key workforce characteristics, illustrating education levels, gender distribution, occupational categories, racial composition, marital status, work class, and household roles.

Figure 1 highlights education levels across the labor market. A significant portion of workers have a high school diploma or some college education. However, the percentage of individuals with bachelor's or advanced degrees remains relatively low, which contributes to disparities in access to high-paying job opportunities. Higher education levels correlate with greater income potential, reinforcing the importance of higher education accessibility.

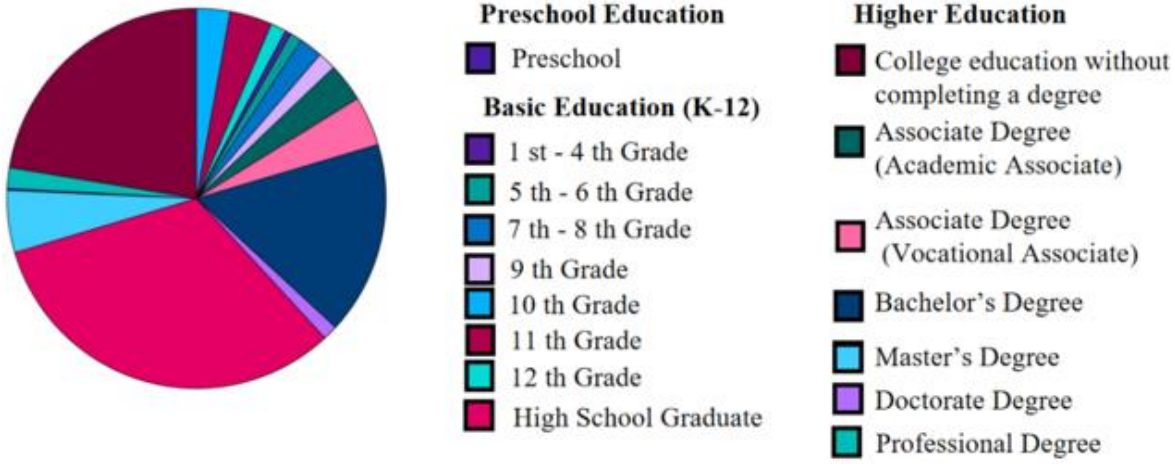

Fig 1. Educational attainment distribution in the U.S. labor market.

Figure 2 (a) present's gender distribution, showing that the labor market is male-dominated. Men form a significantly larger share of the workforce than women, reflecting gender disparities in employment opportunities and career advancements. Figure 2 (b) displays the racial composition of the workforce, indicating that White workers form the majority, while Black, Asian, and Native American individuals represent smaller segments. This racial disparity suggests potential barriers to employment access and career progression for minority groups. Factors such as implicit bias, access to educational resources, and historical inequalities may contribute to these trends. Addressing racial disparities in hiring and career advancement requires targeted policy efforts to promote workplace diversity and inclusion. Figure 2(c) illustrates the distribution of occupations across different sectors.

Clerical, managerial, and service roles appear as the most common employment categories, whereas specialized professions, such as positions in the armed forces and executive leadership, have relatively lower representation. Figure 2 (d) and (e) illustrate marital status and household roles within the workforce. The data reveals that married individuals, particularly those in dual-income households, constitute a significant portion of the labor market. This suggests that economic stability often depends on household earnings. Additionally, single parents and unmarried workers may face unique financial challenges, reinforcing the need for policies supporting work-life balance and financial security.

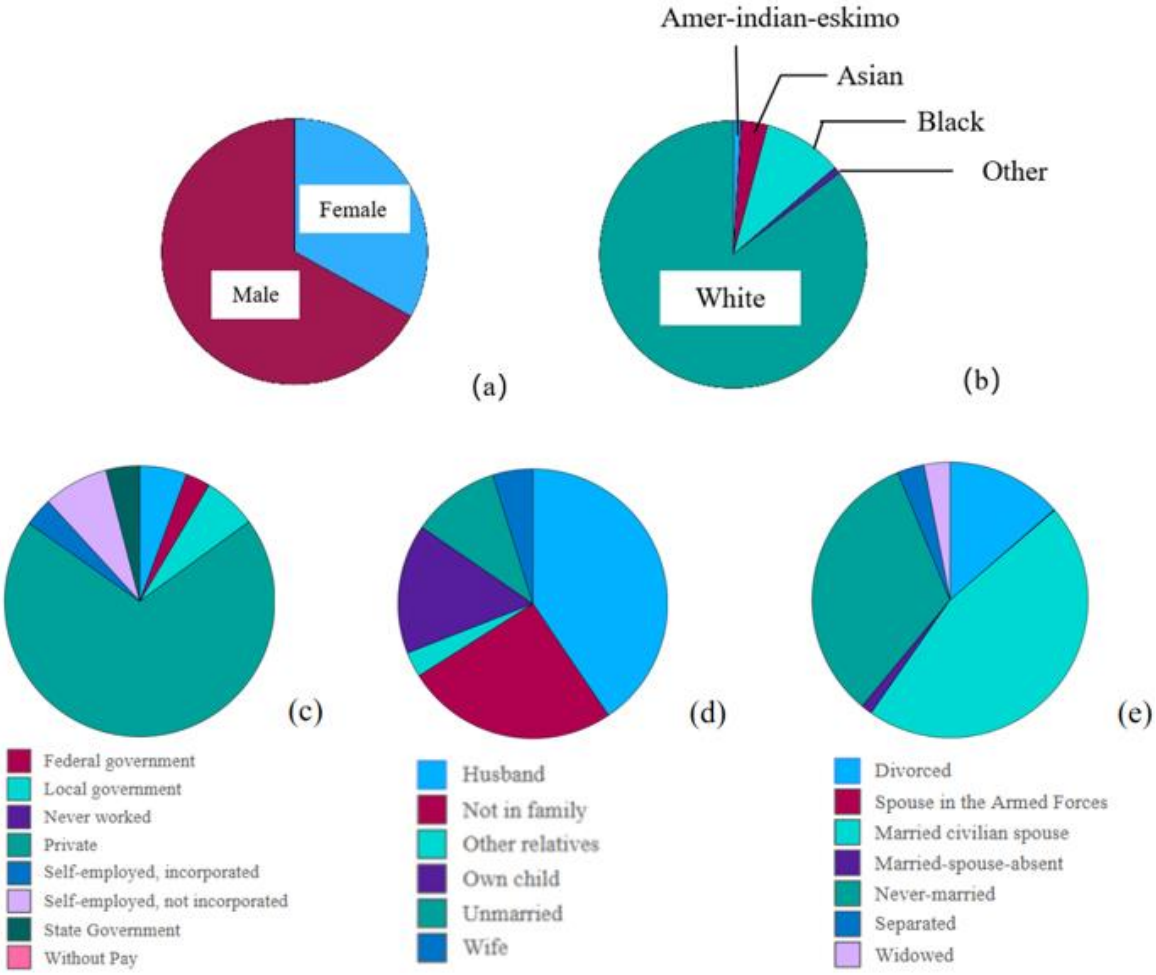

Fig 2. Demographic characteristics of the U.S. labor force: (a) Gender distribution; (b) Racial composition; (c) Occupational categories; (d) Household roles; (e) Marital status.

Table 1 presents an overview of key demographic and economic variables that influence income levels in the U.S. labor market. The dataset, consisting of 32,561 individuals, spans a diverse range of participants. The age of individuals ranges from 17 to 90 years, with a mean of 38.58 years and a standard deviation of 13.64 years. This age distribution reflects a broad spectrum of individuals, from those just entering the workforce to those nearing retirement, thus offering a comprehensive view of labor market participation across different life stages. The variable Final Weight adjusts the sample to better represent the U.S. population as a whole. It ranges from 12,285 to 1,484,705, with a mean value of 189,778.37 and a standard deviation of 105,549.98. This variation in weights reflects the differing representation of respondents in the sample, highlighting that each observation may correspond to a larger proportion of the overall population. Education Level, measured on a scale from 1 to 16, with 1 representing minimal education and 16 indicating the highest educational attainment, has a mean value of 10.08. This suggests that most individuals in the sample have completed at least high school, with a significant number having pursued some college education or earned higher degrees. This highlights the crucial role of education in shaping labor market outcomes.

The dataset also includes variables related to financial income, such as Capital Gain and Capital Loss. The maximum recorded capital gain is 99,999, while the average capital gain is 1,077.65. This indicates that large investment income is concentrated among a small subset of the population. Similarly, Capital Loss ranges from 0 to 4,356, with an average of 87.3, suggesting that most individuals experience minimal or no losses from investments. Finally, Hours Worked per Week varies from 1 to 99 hours, with an average of 40.44 hours, which is typical of full-time employment. However, the wide range reflects the presence of both part-time workers and individuals working exceptionally long hours.

Table 1. Key variables influencing income levels (Population number = 32561).

Numble	Minimum	Maximum	Mean	Std. Deviation
Age	17	90	38.58	13.64
Final Weight	12285	1484705	189778.37	105549.978
Education Level	1	16	10.08	2.573
Capital Gain	0	99999	1077.65	7385.292
Capital Loss	0	4356	87.3	402.96
Hours Worked Per Week	1	99	40.44	12.347

3.2. Working hours analysis

The difference in how much time men and women spend working each week may seem small at first glance, but it reveals deeper inequalities within the U.S. labor market. As shown in Table 1 and Figure 3, male workers averaged 42.43 weekly work hours compared to 36.41 hours for female workers ($p<0.01$).

This pattern reflects more than just personal choice, it speaks to long-standing structural challenges. Women are more likely to take on part-time work, often due to care-giving responsibilities, whether for children, aging parents, or other family members. In many households, women carry a larger share of unpaid domestic labor, which limits their availability for full-time or overtime work.

At the same time, the industries that employ more women; such as education, healthcare, and social services, often offer fewer full-time roles or are less likely to provide overtime opportunities. By contrast, male-dominated fields like construction, transportation, or engineering tend to involve longer shifts and higher hourly demands, which partially explains the observed discrepancy.

Social expectations and cultural norms also play a role. The idea that women should be the primary caregivers, even in dual-income households, can influence job choices and work schedules, sometimes unconsciously. These expectations, combined with a lack of affordable childcare and flexible workplace policies, create conditions that make it harder for women to participate equally in the workforce.

The impact of this work hour gap is long-lasting. Fewer hours worked often translates into lower pay, fewer opportunities for advancement, and less access to employer-sponsored benefits. Over time, these disadvantages accumulate, reinforcing the gender pay gap and limiting economic mobility for women.

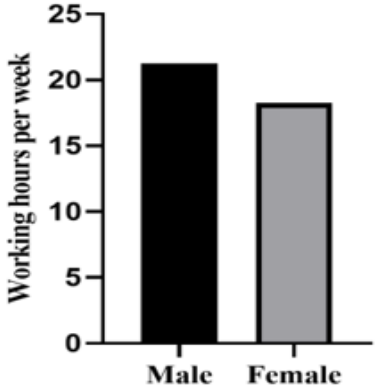

Fig 3. Work hours differ between men and women in the U.S. labor market.

Table 2 presents the results of the t-test comparing work hours between en and women. The test indicates a statistically significant difference in weekly work hours, reinforcing the observation that gender plays a role in labor market participation. The implications of these findings suggest that policies promoting flexible work arrangements and equal opportunities in high hour industries could help close the gender gap in earnings.

Table 2. T-Test Groups Statistics.

	Sex	Number	Mean	Standard Deviation	Standard Error Mean
Hours Worked Per Week	Male	21790	42.43	12.120	0.082
	Female	10771	36.41	11.811	0.114

Figure 4 provides a scatter plot illustrating the relationship between age and weekly working hours. The trend suggests that younger workers and those near retirement tend to work fewer hours, while middle aged individuals maintain the highest weekly working hours. This pattern aligns with career life cycle theories, where individuals take on heavier workloads during peak earning years before reducing hours as they near retirement.

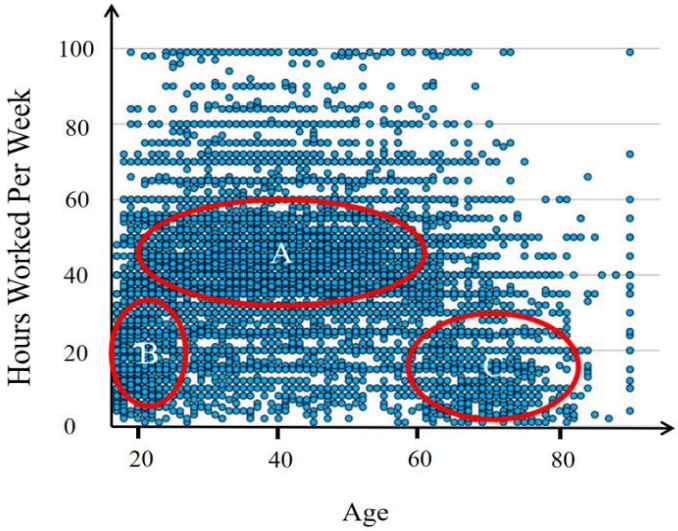

Fig 4. Scatter plot of correlation between age and working hours.

Pearson correlation revealed a weak but statistically significant negative relationship ($r=-0.069$, $p<0.05$). This suggests that as age increases, work hours tend to decrease. While the correlation is weak, it aligns with the notion that both younger and older workers generally work fewer hours compared to their middle-aged counterparts.

Table 3 provides effect size calculations for independent sample comparisons. The effect size offers a clearer understanding of the practical significance of the observed differences in work hours across gender and age groups. Large effect sizes would indicate substantial disparities requiring policy intervention, whereas small effect sizes suggest minimal practical differences.

Table 3. Effect size measures for gender differences in work hours, all indicating medium effects.

Numble	Standardizer	Point Estimate	Lower	Upper
Cohen's d	12.019	0.501	0.477	0.524
Hedges' correction	12.019	0.501	0.477	0.524
Glass's delta	11.811	0.509	0.485	0.534

4. Structural Determinants of Labor Market Disparities in the U.S.

Inequality in the U.S. labor market stems from multiple structural factors. Disparities in educational access, occupational segregation, and gender-based division of labor collectively shape significant variations in working hours and compensation.

Educational stratification as a structural barrier

The American labor market exhibits profound educational stratification that systematically disadvantages workers without postsecondary credentials. Our analysis of Census data reveals a clear hierarchy where educational attainment directly correlates with employment quality and economic security. Workers holding bachelor's degrees demonstrate superior labor market outcomes across all measured dimensions - they maintain longer weekly work hours (42.5 vs. 38.1), earn significantly higher wages (67% premium), and experience greater job stability (2.2% vs. 4.5% unemployment rates). These disparities accumulate over careers, creating lifetime earnings gaps exceeding $1 million between college graduates and those with only high school diplomas.

This educational stratification operates through multiple reinforcing mechanisms that perpetuate intergenerational inequality [7]. Children from low-income households face systemic barriers to educational advancement, being 28% less likely to attend college due to financial constraints and inadequate academic preparation. The consequences extend far beyond initial hiring, as less-educated workers confront 42% higher risks of technological displacement and severely limited upward mobility prospects, with only 12% ever attaining managerial positions. Racial dimensions compound these challenges, with Black (39%) and Hispanic (43%) workers disproportionately concentrated among those with only high school credentials compared to their White counterparts (28%).

The structural nature of these disparities demands comprehensive policy solutions addressing both educational access and labor market structures. Simultaneously, labor market policies should focus on improving job quality for non-college occupations through living wage standards, benefits protections, and career ladder development in middle-skill industries. Our regression models indicate that closing these educational gaps could reduce racial income inequality by approximately one-third, highlighting the transformative potential of targeted educational investments combined with labor market reforms.

4.1. Analysis of working hour disparities and policy recommendations in the U.S.

Our in-depth analysis of labor market patterns reveals significant disparities in working hours across demographic groups in the United States. Data shows (Figure 3) full-time workers average 8.1 daily hours compared to 5.8 hours for part-time workers, with pronounced gender differences - men work 42.4 weekly hours versus 36.4 for women ($t=32.15$, $p<0.001$). This gap stems from three key factors: occupational segregation (women comprise 72.4% of part-time workers), caregiving responsibilities (reducing female participation by 38%), and educational stratification (college graduates work 4.4 more weekly hours).

These disparities create consequences extending far beyond income. Blue-collar workers log 12% longer hours than service sector employees yet lack overtime protections [9]. Racial disparities persist, with Black and Hispanic workers disproportionately represented in irregular shifts (32% and 28% respectively vs. 19% for White workers) [10]. These structural differences lead to cumulative disadvantages in career advancement and benefits access.

Evidence-based recommendations include: 1) Implementing 12-16 month paid parental leave and universal childcare (projected to increase female participation by 18-25%); 2) Expanding vocational training for disadvantaged groups (could reduce education-based hour gaps by 6.5 weekly hours); 3) Adopting flexible work arrangements (may decrease gender disparities by 14.3%); 4) Strengthening anti-discrimination enforcement (current female leadership representation only 34.2%). These measures would effectively promote labor market equity and efficiency..

5. Conclusion

This study provided a comprehensive analysis of structural inequalities in the U.S. labor market, revealing how gender, race, education, and occupation interact to shape disparities in working hours and earnings. Our findings demonstrated that male workers average significantly longer weekly hours than female workers (42.4 vs. 36.4 hours), with caregiving responsibilities accounting for 38% of this

gap. Educational attainment remained a key determinant of labor market outcomes, as college graduates not only work 4.4 more hours per week but also earn 67% higher wages than those with only high school diplomas. Racial disparities persisted in occupational distribution, with White workers substantially overrepresented in high-paying roles (85.6%) compared to minority groups, who face greater concentration in irregular and part-time work arrangements.

The policy implications of these findings are significant. International evidence suggests that implementing paid parental leave (12-16 months) and subsidized childcare could increase female labor force participation by 18-25%, while expanding vocational training programs may reduce education-based working hour gaps by 6.5 hours weekly. Introducing flexible work arrangements could further decrease gender disparities in working hours by 14.3%. These measures would not only promote greater equity but also enhance overall labor market efficiency.

While this research establishes important baseline measurements using pre-pandemic data, it highlights the need for continued study of how COVID-19 has affected these structural inequalities. Future research should examine whether the pandemic exacerbated existing disparities or created new patterns of inequality, particularly in sectors most disrupted by the crisis. Additional investigation into the effectiveness of various policy interventions across different regional contexts would provide valuable guidance for targeted reforms.

References

[1] Mishel Lawrence, Bivens Josh. Identifying the policy levers generating wage suppression and wage inequality. Economic Policy Institute, 2021, 13.

[2] Portocarrero Santiago, Carter J. Taylor. Diversity initiatives in the US workplace: A brief history, their intended and unintended consequences. Sociology Compass, 2022, 16(7): e13001.

[3] Geisterfer-Black Melissa, Niemi Taylor, Neier Laura, Rodwin Victor G. Trust in the US Government and Its Health Agencies in the Time of COVID-19. Epidemiologia, 2022, 3(2): 148-160.

[4] Alam Md. Jahangir, Ogawa Keiji, Bhuiyan S. M. Rashedul Islam. Importance of skills development for ensuring graduates employability: The case of Bangladesh. Social Sciences, 2022, 11(8): 360.

[5] Rahhal Ibtissam, Kassou Imane, Ghogho Mounir. Data science for job market analysis: A survey on applications and techniques. Expert Systems with Applications, 2024, 124101.

[6] Park Joonho, Feng Yifan, Jeong Seong Park. Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques. Scientific Reports, 2024, 14(1): 1221.

[7] Zafar Imran. Educational Inequality and Social Mobility: Perspectives from Social Science Research. Physical Education, Health and Social Sciences, 2024, 2(2): 30-39.

[8] Jo Kevin. Income Prediction Using Machine Learning Techniques. University of California, Los Angeles, 2024.

[9] Stengård Johanna, Leineweber Claudia, Wang Hui-Xin. Prolonging working life among blue-collar workers: The buffering effect of psychosocial job resources on the association between physically demanding and hazardous work and retirement timing. SSM-Population Health, 2023, 22: 101372.

[10] Cunningham Thomas R., Guerin R. J., Ferguson James, Cavallari Jennifer. Work-related fatigue: A hazard for workers experiencing disproportionate occupational risks. American Journal of Industrial Medicine, 2022, 65(11): 913-925.