An Analysis of Gender based Income Disparity in U.S. 1994

Abstract

This report examines the underlying factors contributing to gender-based income disparities using the 1994 U.S. Census ”Adult” dataset. The analysis integrates modern data science techniques, including Python-based preprocessing, visual exploration in Tableau, and Bayesian probabilistic modeling, to identify and quantify the most significant predictors of the gender pay gap. Key variables such as age, workclass, education, hours worked, race, and marital status are systematically analyzed to assess their influence on income inequality. By situating historical data within the context of contemporary research, the study not only confirms the persistence of the gender pay gap but also reveals how its drivers have evolved over time. The findings provide nuanced insights into the complex interplay of demographic and socio-economic factors shaping pay equity, offering a valuable perspective for both historical understanding and future policy considerations.

1 Introduction

Gender inequality in the workplace remains a persistent socio-economic challenge. Although female labour-force participation has increased over time, women continue to earn less than men—a gap documented by multiple sources, including the Pew Research Centre [2] and Forbes [5]. This study asks whether women who are employed experience a pay disadvantage compared to their male colleagues, and it seeks to identify which factors most powerfully shape any observed disparities. Both Pew and Forbes suggest that factors such as ethnicity, age, hours worked, marital status, and education significantly influence income across genders. Hence, this report will focus on analysing these specific variables to understand the trends present in the 1994 U.S. Census data. To guide this analysis, we pose four research questions:

- Is there a statistically significant gender pay gap among employed individuals in the 1994 U.S. Census data?
- Which factors most strongly predict income disparities between men and women?
- How do intersecting identities, such as race and marital status, influence pay outcomes?
- To what extent do sector workclass, and education level moderate the gender pay gap?

This report will situate its findings within a broader historical narrative by conducting a comparative analysis between the socio-economic landscape of 1964 and the present day. In 1964, in the immediate wake of the Equal Pay Act, the gender pay gap was stark, with women earning approximately 59 cents for every dollar a man earned, largely due to overt occupational segregation and restrictive societal norms. This analysis will reference recent research from sources like Forbes and the Pew Research Centre to see what has changed. While the gap has narrowed, women still earn only about 84 cents for every dollar earned by a man in 2024, according to a Forbes analysis [1]. Contemporary studies from the Pew Research Centre [3] confirm that the underlying drivers have shifted to more complex issues, including the disproportionate impact of parenting on women's careers, differences in long-term career trajectories, and the persistent burden of unpaid care work. This comparative approach will illuminate both the progress made over the last six decades and the stubborn persistence of income inequality. This report will first detail the data and methodology used, then proceed with the analysis to address the research questions.

2 Data Preparation and Abstraction

2.1 Data Source and Abstraction

This study uses the 1994 U.S. Census ”Adult” dataset, taken from the UCI Machine Learning Repository [4]. The full list of attributes in the original dataset is detailed in Table 1.

Table 1: Original Data Description

Attribute	**Nature of Data**	**Description & Categories**
`income`	Categorical	Target variable. Categories: “`>50K`”, “`<=50K`”.
`age`	Continuous	Age of the individual.
`workclass`	Categorical	Categories: `Private`, `Self-emp-not-inc`, `Self-emp-inc`, `Federal-gov`, `Local-gov`, `State-gov`, `Without-pay`, `Never-worked`.
`fnlwgt`	Continuous	Final weight, the number of units in the target population that the responding unit represents.
`education`	Categorical	Categories: `Bachelors`, `Some-college`, `11th`, `HS-grad`, `Prof-school`, `Assoc-acdm`, `Assoc-voc`, `9th`, `7th-8th`, `12th`, `Masters`, `1st-4th`, `10th`, `Doctorate`, `5th-6th`, `Preschool`.
`education-num`	Continuous	Numeric representation of education level.
`marital-status`	Categorical	Categories: `Married-civ-spouse`, `Divorced`, `Never-married`, `Separated`, `Widowed`, `Married-spouse-absent`, `Married-AF-spouse`.
`occupation`	Categorical	Categories: `Tech-support`, `Craft-repair`, `Other-service`, `Sales`, `Exec-managerial`, `Prof-specialty`, `Handlers-cleaners`, `Machine-op-inspct`, `Adm-clerical`, `Farming-fishing`, `Transport-moving`, `Priv-house-serv`, `Protective-serv`, `Armed-Forces`.
`relationship`	Categorical	Categories: `Wife`, `Own-child`, `Husband`, `Not-in-family`, `Other-relative`, `Unmarried`.
`race`	Categorical	Categories: `White`, `Asian-Pac-Islander`, `Amer-Indian-Eskimo`, `Other`, `Black`.
`sex`	Categorical	Categories: `Female`, `Male`.
`capital-gain`	Continuous	Capital gains recorded for the individual.
`capital-loss`	Continuous	Capital losses recorded for the individual.
`hours-per-week`	Continuous	Number of hours worked per week.
`native-country`	Categorical	Categories: `United-States`, `Cambodia`, `England`, `Puerto-Rico`, `Canada`, `Germany`, `Outlying-US(Guam-USVI-etc)`, `India`, `Japan`, `Greece`, `South`, `China`, `Cuba`, `Iran`, `Honduras`, `Philippines`, `Italy`, `Poland`, `Jamaica`, `Vietnam`, `Mexico`, `Portugal`, `Ireland`, `France`, `Dominican-Republic`, `Laos`, `Ecuador`, `Taiwan`, `Haiti`, `Columbia`, `Hungary`, `Guatemala`, `Nicaragua`, `Scotland`, `Thailand`, `Yugoslavia`, `El-Salvador`, `Trinadad&Tobago`, `Peru`, `Hong`, `Holand-Netherlands`.

2.2 Feature Selection

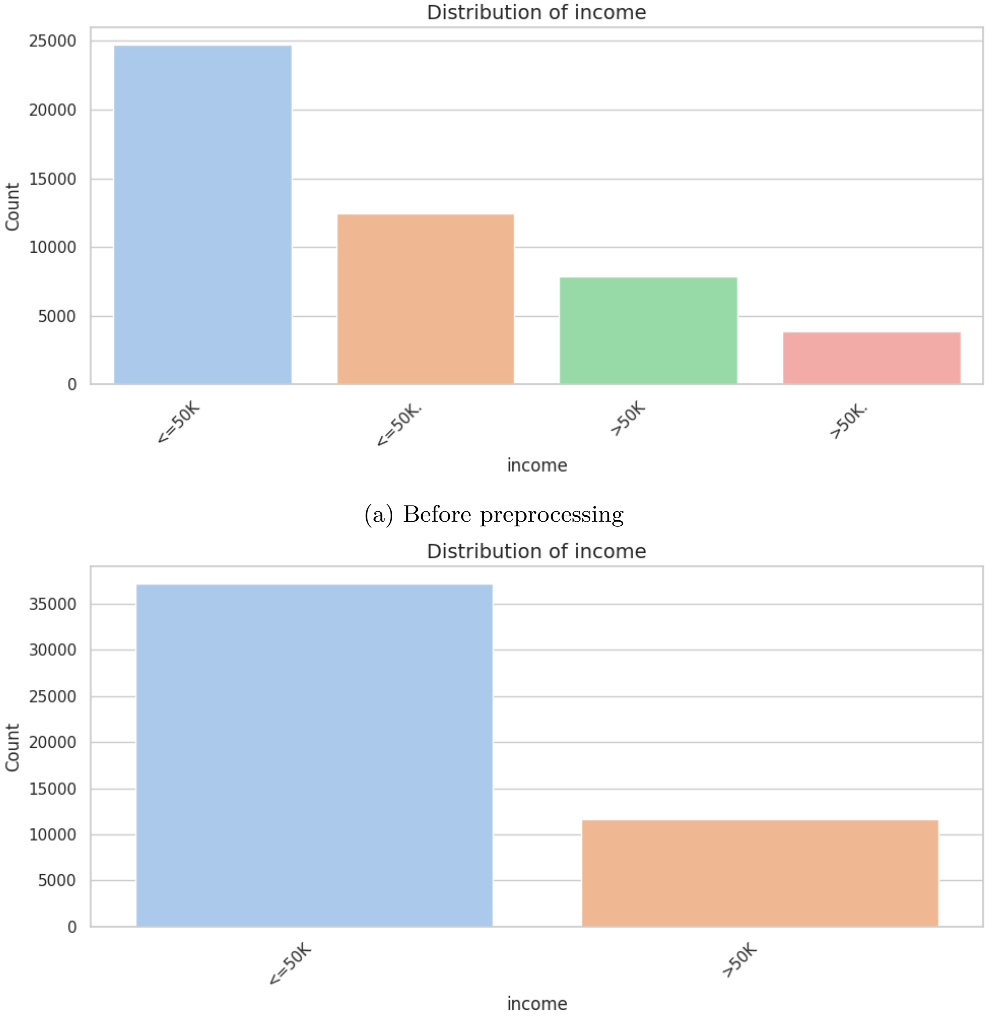
As seen in Table 1, the dataset contains many attributes. To align this study with contemporary research on income disparity, a subset of these features was selected for analysis. This decision was guided by recent reports from the Pew Research Centre and Forbes, which identify a core set of factors that consistently predict income levels. The selected features and the rationale for their inclusion are detailed in Table 2. Columns such as `fnlwgt`, `relationship`, and `native-country` were excluded to maintain focus on the primary drivers identified in modern literature.

Table 2: Selected Features for Analysis

Selected Feature	**Nature of Data**	**Rationale for Selection**
`income`	Categorical	The primary target variable for classifying individuals into high or low-income brackets.
`sex`	Categorical	The core independent variable for analysing gender-based disparities.
`age`	Continuous	Serves as a proxy for work experience and career progression, a known factor in income differences.
`education`	Categorical	Directly linked to earning potential and a key factor in modern pay gap analyses.
`workclass`	Categorical	Distinguishes between private, public, and self-employed sectors, which often have different pay structures.
`race`	Categorical	Allows for an intersectional analysis to see how the gender pay gap varies across different racial groups.
`marital-status`	Categorical	Provides insight into how family structures and societal roles, such as caregiving, might influence income.
`hours-per-week`	Continuous	Normalises earnings by effort, helping to distinguish between pay rates and differences in work volume.

2.3 Data Preprocessing

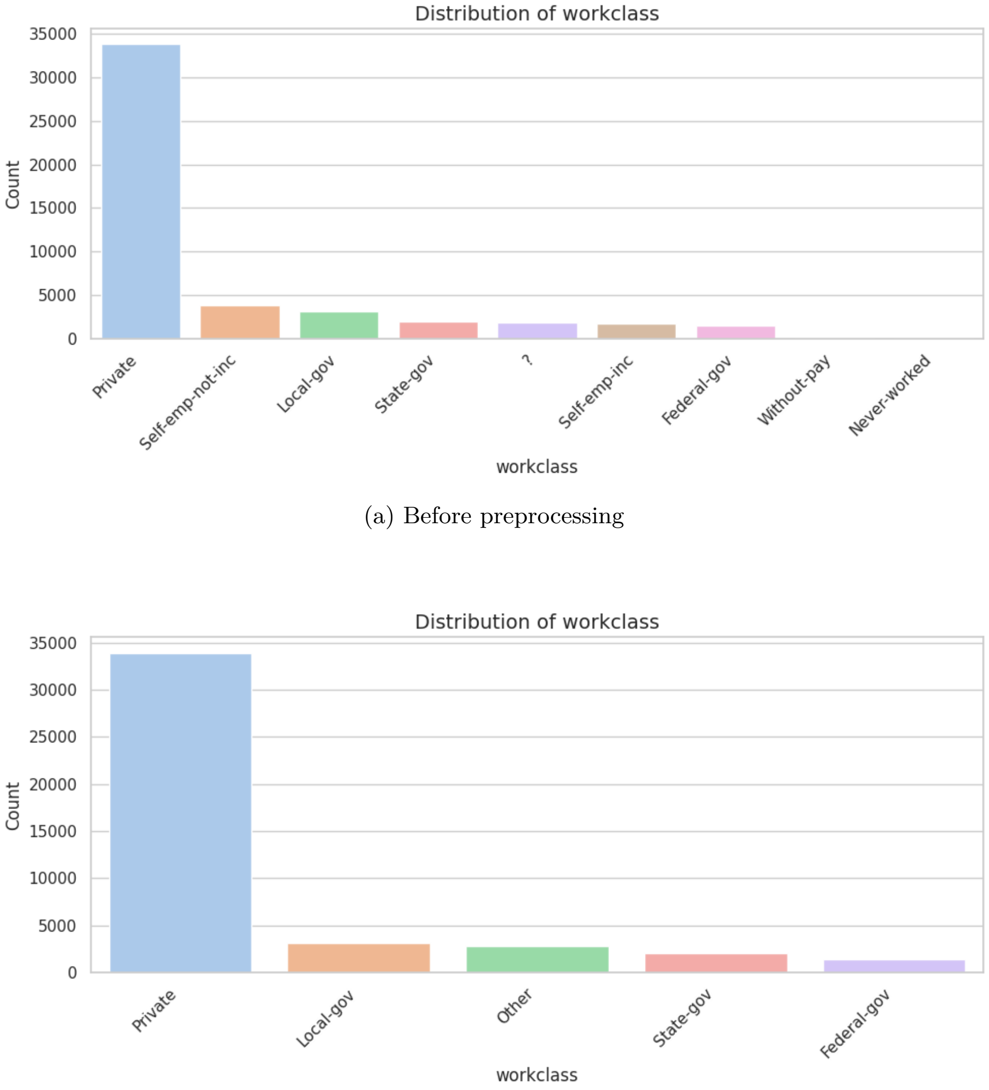
The data was prepared for analysis using Python within a Jupyter Notebook environment, leveraging the Pandas library for data manipulation and Matplotlib for visualisation. The initial step involved loading the dataset into a Pandas DataFrame to facilitate exploratory analysis. During this exploratory phase, bar plots were generated for each categorical variable to visualise the distribution of its categories and to identify data quality issues. This process revealed that the `workclass` attribute contained missing values, which were denoted by a '?' character. To handle these missing data points without losing valuable information from the affected rows, an imputation strategy was chosen over deletion: a new category, `Other`, was created within the `workclass` variable, and all instances of '?' were reassigned to this category. Additionally, as the report is focused exclusively on the income of employed individuals, records where `workclass` was "Without-pay" or "Never-worked" were removed from the dataset. This ensures that the analysis only considers those who are part of the active workforce. In addition to addressing missing values, the `income` column required cleaning. It was observed that some entries for the income categories contained trailing full stops, specifically "`>50K.`" and "`<=50K.`". This inconsistency created the appearance of four distinct income categories instead of two. To resolve this, these trailing characters were removed, standardising the income data. The resulting cleaned and preprocessed dataset formed the basis for all subsequent analysis. The results of these preprocessing steps are illustrated in Figures 1 and 2, which show the distribution of the `workclass` and `income` attributes before and after cleaning.



(a) Before preprocessing

(b) After preprocessing

Figure 1: Distribution of the `income` attribute before and after standardising categories.



(a) Before preprocessing

(b) After preprocessing

Figure 2: Distribution of the `workclass` attribute before and after handling missing values.

3 Task and Analysis

3.1 Task 1: Overview of the Data

Before delving into a detailed analysis of income disparities, it is essential to establish a baseline understanding of the dataset's composition following preprocessing. This initial exploratory step focuses on the distribution of key demographic variables, providing the necessary context for interpreting subsequent findings. The primary objective is to profile the population based on gender and work sector, which serves as a foundation for the deeper investigation into income inequality.

The dashboard design adheres to principles of information visualisation to ensure clarity and support meaningful comparisons:

- **Overall Population Count:** A bar chart was selected to represent the overall population by gender. This chart type is widely understood and effectively supports comparison through the preattentive attribute of length. Each gender is represented using a distinct colour, enhancing readability. Population counts are displayed above each bar to provide immediate quantitative insight.
- **Distribution of Population by Race:** A grouped bar chart was employed for this task. This format facilitates dual-layered comparison—both between genders within each racial group and across different racial groups overall. The grouped layout maintains clarity while enabling quick visual assessment of disparities.
- **Population Distribution of Gender Across Work Sectors:** A shape plot was used to visualise gender distribution across workclass categories. This plot uses distinct shapes and vertical separation to maintain a clean and uncluttered appearance, avoiding the visual density of traditional bar charts. Consistent color schemes and visible count labels further aid interpretability and comparison.

Analysis of Task 1 Dashboard

It is clearly observable that the male and female populations differ significantly, not just in the overall population count, but also across different racial groups and workclass. This discrepancy provides an important clue for the income disparity analysis: a direct comparison of absolute income counts could be misleading due to the imbalance in gender representation.

To mitigate this bias, it is more appropriate to calculate the percentage of each gender group belonging to specific income brackets. This normalisation allows for a fairer comparison by focusing on income distribution within each group, rather than raw totals. As such, analysing proportional income differences will yield more accurate insights into potential disparities between genders.

This insight forms the foundation for Task 2, where the focus shifts to income distribution analysis using these proportional measures.

3.2 Task 2: Proportional Analysis of Income Distribution

To fairly compare income distribution across genders, it is important to evaluate what proportion of each gender falls into specific income brackets: $>$ \$50K and $\leq$ \$50K. Percentages are ideal for this comparison, as they are not influenced by imbalances in population size.

For example, consider a scenario with 100 males and 50 females. If 30 males earn $\leq$ \$50K and 70 earn $>$ \$50K, while 30 females earn $\leq$ \$50K and 20 earn $>$ \$50K, it may initially appear that both genders are equally represented in the lower income bracket. However, this interpretation is misleading. In percentage terms, 30 out of 100 males equals **30%**, whereas 30 out of 50 females equals **60%**. This reveals that a significantly higher proportion of women earn $\leq$ \$50K compared to men. Such proportional insights are more valuable for understanding income disparity.

To support this analysis, a dashboard was created using the following visualizations:

- **Gender-Based Income Disparity (Overall):** A stacked bar chart was used to depict the proportion of each gender in the two income brackets. Each bar represents one gender, divided into two segments using different shades of the same colour to distinguish the income brackets ($>$ \$50K and $\leq$ \$50K). Percentages are displayed directly on the bars to enhance interpretability. This chart provides a clear visual comparison of income distribution between males and females.
- **Gender-Based Income Disparity Across Racial Groups:** To extend the proportional analysis across racial categories, pie charts were used. Each racial group is represented by a separate pie chart, with segments colored consistently using the same colour scheme for gender and income as the stacked bar chart. This visual consistency serves as a helpful cue, allowing users to quickly identify gender and income categories across the dashboard. Each pie chart includes:
  - Percentage labels indicating the proportion of the respective gender in each income group,
  - Titles denoting which racial group the chart represents,

  - Distinct slices representing the different income categories, making income group distribution visually intuitive.

- **Color Indicator Legend:** A colour legend is included in the dashboard to indicate which colours correspond to each gender and income bracket (> \$50K and ≤ \$50K). This ensures users can interpret the charts correctly and efficiently.

- **Interactive Features:** The pie charts are interactive: when a user hovers their cursor over a pie slice, the total count of individuals in that slice is displayed. This provides additional context beyond percentages and aids further analysis.

All visualizations prominently display the percentage of the population that falls into each income group, making it easier to compare income disparities fairly and clearly. This proportional approach ensures that the analysis is not skewed by population imbalance and offers a more meaningful view of gender-based income inequality.

Analysis of Task 2 Dashboard

The dashboard clearly shows that women are more likely to earn less than or equal to \$50K compared to men. As seen in the stacked bar chart on the left, **89.7% of women** fall in the lower income bracket (≤ \$50K), whereas only **71.02% of men** are in the same category. This substantial gap suggests a consistent gender-based income disparity in the overall population.

When examining the pie charts on the right, this trend persists across racial groups, although the degree of disparity varies:

- Among **White individuals**, 88.88% of women earn ≤ \$50K compared to 69.77% of men.

- For the **Black population**, 94.39% of women earn below \$50K, while 82.32% of men do.

- Among **Asian-Pacific Islanders**, 86.75% of women fall in the lower income group, in contrast to 67.85% of men.

- **American Indian/Eskimo** women show 92.13% in the lower bracket, versus 86.72% of men.

- Similar trends are observed across the **Other** racial category as well.

This visual evidence reinforces the importance of using proportional measures when analyzing income, as absolute counts can obscure systemic disparities. Task 2's dashboard, by showing relative percentages, reveals that gender-based income inequality persists not only in aggregate but also across all racial segments. As next step it is to check if the trend continues across the workclass.

3.3 Task 3: Education Level Comparison

Education is one of the factors which impacts the income of an individual a lot. For both male and female, the level of education was compared to see what level of education both of them had and if there is any difference in the trend. For this, a dashboard with a single plot was used. The plot is a **butterfly chart** also known as a tornado chart or a divergent bar chart.This type of plot is highly effective for this comparison for several reasons:

Direct Comparison: It allows for a direct, side-by-side comparison of the education levels between males and females for each category. You can instantly see which gender has a higher percentage for any given education level. For example, it's clear that a higher percentage of females have ”Some-college” education, while a slightly higher percentage of males are ”HS-grad”.

Highlights Differences: The chart's structure effectively emphasizes the differences in proportions between the two genders. The longer the bar on one side compared to the other, the greater the disparity.

Visual Clarity and Precision: The chart is designed for easy and accurate interpretation.

- For clear visual comparison, **color coding** was used: blue represents female and orange represents male.
- The **exact percentage** of a particular gender in that particular education level category is also visible on each bar. This allows the viewer to move beyond just a visual estimate of the bar's length and make a precise numerical comparison between the two groups.

Analysis of Task 3 Dashboard

An analysis of educational attainment reveals a significant finding within the "Some-college" category. This classification represents individuals who have attended college without earning a degree. The data indicate that a higher proportion of the female population falls into this category when compared to the male population.

This finding suggests that in 1994, women were more likely than their male counterparts to have undertaken post-secondary education without ultimately attaining a degree. This disparity in educational completion is identified as a critical factor that can limit earning potential and, consequently, is a likely contributor to the overall gender income gap.

3.4 Task 4: Gender-Based Income Analysis by Race and Workclass

While it is well established that women are less likely to be in the $>$ \$50K income group overall, this aggregate trend may obscure important variations within specific work sectors and racial groups. To explore whether the gender income gap persists across different workclasses and races, a comprehensive dashboard was created, incorporating several complementary visualizations:

- **Matrix Heatmap Table:** As shown above, this table visualizes the intersection of race, gender, workclass, and income bracket. Each cell displays the percentage of individuals in a subgroup who earn $>$ \$50K or $\leq$ \$50K, with the background color indicating the raw population count. This dual encoding allows users to simultaneously assess proportional income differences and the underlying population size, ensuring that observed trends are interpreted in the context of subgroup sizes.
- **Pie Charts:** For each workclass, a pie chart illustrates the overall income distribution by gender. Each segment represents a gender-income bracket combination, using distinct colors (e.g., dark pink for females earning $\leq$ \$50K, light pink for females earning $>$ \$50K, dark blue for males earning $\leq$ \$50K, and light blue for males earning $>$ \$50K). Pie charts are effective for quickly conveying proportional relationships, making it easy to compare gender disparities in income distribution within each work sector at a glance.
- **Population Matrix Table:** This table displays the actual population counts for each gender, race, and workclass combination. It complements the heatmap by providing a clear reference for the sample sizes underlying the percentages, supporting more informed interpretation of the data.
- **Color Legend:** A detailed color legend clarifies the meaning of each color in the visualizations. Pink shades represent females (dark for $\leq$ \$50K, light for $>$ \$50K), while blue shades represent males (dark for $\leq$ \$50K, light for $>$ \$50K). This ensures that users can accurately interpret both the heatmap and the pie charts.

by combining the matrix heatmap table, pie charts, and population counts, the dashboard provides a nuanced, multidimensional view of gender-based income disparities. This approach allows users to explore patterns not only in the aggregate but also within specific intersections of race and work sector, and to quickly compare both proportional and absolute differences across subgroups.

Analysis of Task 4 Dashboard

The dashboard visualizations reveal a clear and consistent trend: women are significantly less likely than men to fall into the higher income bracket ($¿50K) across nearly all workclasses and racial groups. This disparity is particularly evident in the pie charts, which provide a visual summary of income distribution within each work sector.However, a closer look at the matrix heatmap table uncovers some notable exceptions to this overall trend. For example, among the American-Indian-Eskimo racial group, women employed in Federal and Local Government roles are actually more likely to earn over $50K than their male counterparts. Additionally, the income disparity within this racial group appears to be generally less pronounced. A similar—though slightly less striking—pattern is seen among women in the "Other" racial category, where those working in State Government jobs also show a higher likelihood of earning above $50K compared to men.Despite these exceptions, the overarching pattern remains: women are generally underrepresented in the higher income category. One particularly prominent and consistent trend emerges within the private sector. Across all racial groups, the private workclass shows the most significant and uniform gender income gap, with men overwhelmingly more likely than women to earn over $50K. This makes the private sector a key area of concern in addressing income inequality, especially when contrasted with the more mixed patterns seen in government employment.In summary, while the overall data highlights a substantial gender income gap, the dashboard also emphasizes that the extent of this disparity varies notably by both workclass and race. The most persistent and striking inequality is concentrated in the private sector, underscoring its role as a focal point for further investigation and policy intervention.

3.5 Analysis of Task 5 Dashboard

Now that we've established a clear trend of income disparity between men and women across all racial groups, it's worth exploring what might be contributing to this gap. One meaningful factor to consider is the number of hours worked per week. To dive deeper into this, a dashboard was created featuring several visualizations that examine how working hours vary across gender, race, and workclass.

The dashboard uses a consistent color scheme to distinguish between male and female categories, with a clear legend that shows which color represents which gender. Descriptive headings help guide the viewer through each chart, and average working hours are displayed directly on the plots to make key figures immediately visible. The goal was to keep the dashboard both informative and easy to interpret.

- **Bar Chart** – Compares the overall average working hours per week between men and women, using distinct colors and visible value labels for each gender.
- **Stacked Bar Chart** – Breaks down average working hours by gender within each racial group, helping highlight variations in gendered work patterns across different communities.
- **Pie Charts** – Showa the distribution of average working hours by workclass and gender, offering a visually engaging way to compare across employment sectors while adding variety to the analysis.

Analysis of Task 5 Dashboard

By examining the charts, a clear difference in average working hours between men and women emerges. This observation aligns with findings from recent studies by *Forbes* [1] and the *Pew Research Center* [3], which highlight that women are more likely than men to take on household responsibilities. Whether it's managing domestic chores or raising children, which could be due to societal and cultural expectations which often place a greater burden on women. This additional, often unpaid, workload outside of formal employment may help explain why women, on average, report fewer working hours per week compared to men.

Task 6: Career Progression with Age

As individuals age, their income typically increases due to career advancement, experience, and skill accumulation. This often leads to a higher likelihood of being in the income group earning over $50K. Analyzing the percentage of men and women in this income group across age brackets can provide valuable insights into gender-based career progression trends.

To explore this, the following visualizations were created:

- **Trendline Chart:** This line chart displays the percentage of men and women in each age bin who earn more than $50K. Two distinct lines orange for males and blue for females highlight gender-based progression trends over time. Percentage values are shown at each age point, allowing for a clear comparison of how income levels change with age for both genders.

- **Matrix Heatmap Table:** This table provides a detailed breakdown across age, race, and gender. It shows the percentage of the population earning over $50K and their corresponding average weekly working hours. This dual insight helps identify how income and working time correlate across different demographic segments and may explain anomalies or declines in income trends due to reduced work hours or nearing retirement.

overall, these visualizations work together to effectively illustrate how career progression and working patterns influence income distribution over time across gender, age groups, and racial categories.

3.5.1 Analysis of Task 6 Dashboard

An interesting trend emerges from the dashboard analysis: males and females were almost equally likely to be in the income group earning over $50K between the ages of 20 and 35. However, as age increases beyond 35, a noticeable income gap begins to appear. This gap widens significantly and peaks around the age of 45, after which it begins to narrow slightly. Despite this decline, the disparity remains substantial and never returns to the near-parity seen in the younger age groups. Recent studies by *Forbes* [1] and the *Pew Research Center* [3] suggest that this widening income gap may be attributed to gendered household responsibilities and child-rearing expectations, which tend to disproportionately impact women's careers and working hours. This explanation is further supported by the matrix heatmap, which shows a gradual decline in the average weekly working hours for women as age increases. In contrast, men either maintain or increase their average working hours over time. However, this trend is not without exceptions. For instance, in the American-Indian-Eskimo group, women between the ages of 30 and 40 were found to be working more hours per week than their male counterparts. Yet, despite this, their likelihood of being in the income group earning over $50K did not increase proportionately, suggesting that structural or systemic factors may still be at play in limiting income growth for women in certain demographics.

3.6 Task 7: Effects of Marital Status on Income Analysis

Another important factor that appears to influence income is marital status. To understand how marital status affects career advancement for both males and females, A dashboard was made to explores how the proportion of individuals earning more than $50K varies across different marital groups over time. These groups include: widowed, married (civilian spouse and Armed Forces spouse), divorced, separated, and never married.

The dashboard consists of the following visual components:

- **Line Chart:** This chart presents trend lines for each marital status category, separated by gender. Each line represents the percentage of individuals within that subgroup (male or female) who fall into the income group earning over $50K. Different colors are used to distinguish between male and female data, and percentage labels on the lines aid in direct comparison.

- **Matrix Table:** This matrix provides an overview of each marital group's percentage of individuals earning above $50K, alongside their average weekly working hours. While it does not incorporate age, it offers insight into overall income distribution and labor input across genders. Color coding consistent with the line chart further improves readability and comparison.

- **Tornado (Butterfly) Chart:** This population distribution chart displays the overall percentage of males and females within each marital group. It complements the insights from the line and matrix charts by providing context on how each group's size may relate to income and work-hour trends.

Overall, this dashboard allows for a multi-faceted analysis of how marital status correlates with income, and how gender dynamics manifest within each group in terms of career progression and work effort.

Analysis of Task 7 Dashboard

To begin the analysis, it's useful to identify where the majority of the working population falls across different marital groups. The tornado chart shows that the largest share of working males belongs to the ”Married-civilian spouse” group, while for females, the majority are in the ”Never-married” category. This aligns with findings from the *Pew Research Center* [3], which state that married women are more likely to exit the workforce. As a result, a larger proportion of employed women are never married.Focusing on the two largest groups—"Never-married" and "Married-civilian spouse"—provides a clearer view of key trends:

- **Never-married group:** The line chart reveals that never-married women are consistently less likely to earn above $50K compared to their male counterparts, highlighting a persistent income disparity. This contradicts the common assumption that unmarried women, having fewer family obligations, would work more hours and achieve income levels closer to men. However, the matrix table indicates that never-married women actually work fewer hours on average than never-married men, which may partially explain the income gap.

- **Married-civilian spouse group:** Although females represent a smaller proportion within this group, their income trend presents a notable deviation from the findings of the *Pew Research Center* [3], which suggests that married women are typically less likely to out-earn their male counterparts. Contrary to this, the line chart reveals that women in this group are more likely than men to earn above $50K, particularly between the ages of 25 and 40. Interestingly, this higher income is achieved despite working fewer hours on average—38.12 hours per week for females compared to 45.87 hours for males. This challenges the conventional assumption that longer working hours necessarily lead to higher income.

These findings point to nuanced gender dynamics in how marital status influences income and work behavior, and they suggest that factors beyond work hours—such as occupation type, role seniority, or education—may contribute significantly to observed income differences.

3.7 PCA-Based Linear Dimensionality Reduction Analysis

Principal Component Analysis (PCA) is a statistical technique that simplifies high-dimensional datasets by transforming them into a smaller set of uncorrelated variables called principal components. These components capture the maximum amount of variance in the original data while preserving key patterns and trends, essentially reducing complexity by projecting data onto new axes that highlight the most significant directions of variation. While this makes datasets easier to analyze without losing much information, it is important to note that PCA is linear and may overlook nonlinear relationships.To visualize these simplified datasets, 2D PCA plots display the reduced data using just the top two principal components, creating an accessible two-dimensional representation of originally complex, multi-dimensional data. This visualization approach allows researchers to easily identify clusters, outliers, and underlying structures that might be hidden in higher dimensions, making it invaluable for exploratory analysis, pattern recognition, and gaining quick insights into data relationships. These plots can reveal meaningful groupings or separations within datasets, supporting diagnostic work and decision-making processes. However, it is crucial to check the explained variance ratios to ensure that these two dimensions capture enough of the data's essential information to provide meaningful insights.

Data Preparation for PCA

To prepare the dataset for PCA:

- All categorical variables were **one-hot encoded** to convert them into a numerical format suitable for linear transformations.

- The continuous `age` variable was binned into 5-year intervals (e.g., 20–25, 25–30) and then one-hot encoded. This preserved age-related structure while minimizing the impact of large numeric ranges.

- Since the resulting data consisted of binary values (0/1), no additional feature scaling was necessary.

PCA Implimentation

PCA was implemented in Python using the `scikit-learn` library:

- The first two principal components (`PC1` and `PC2`) were extracted, capturing 80% of the variance in the data.

- These components were added to the original pandas DataFrame as new columns, enabling seamless integration with other features.

The enriched DataFrame was imported into Tableau for interactive visualization:

- **Each point** represents an individual, plotted based on their PCA coordinates.
- **Filters** were implemented for age bins, workclass, marital status, gender, and income to enable dynamic subsetting.
- **Color encoding** was used to distinguish four key demographic groups: female earning >50K, female earning $\leq$50K, male earning >50K, and male earning $\leq$50K.
- **Tooltips** provided detailed attribute information (e.g., occupation, education) on hover for enhanced interpretability.

Interpreting the PCA Embedding

The 2D PCA projection reveals the structure of the data along the two axes of greatest variance. By analyzing the component "loadings" (the influence of the original features), we can assign meaning to these abstract axes.The analysis revealed the following interpretations for each component:

- **Principal Component 1 (PC1)**, plotted on the horizontal axis, showed a strong correlation with the income variable. It can therefore be interpreted as an axis representing an individual's **economic standing**.
- **Principal Component 2 (PC2)**, plotted on the vertical axis, was most heavily influenced by features related to `age` and `marital-status`. It can be interpreted as an axis representing an individual's **life stage and family structure**.

Visual Analysis of the PCA Plot

The PCA scatter plot provides a clear visualization of the data's underlying structure. The most prominent feature is the distinct separation of data points along the PC1 (economic standing) axis.As seen in the plot, the two low-income groups (Female ‘$<= 50K$‘ in purple and Male ‘$<= 50K$‘ in red) are clustered together on the right side of the graph, primarily with positive PC1 values. Conversely, the two high-income groups (Female ‘$> 50K$‘ in yellow and Male ‘$> 50K$‘ in brown) are clustered together on the left side, with negative PC1 values.The most significant insight comes from observing the gender distribution within these two broad income clusters.Within each economic group, the colors representing males and females are heavily mixed and overlapped. There is no clear visual separation between purple and red points, nor between yellow and brown points. This suggests that, once economic standing is accounted for, the combined socio-demographic profiles of men and women in the dataset are very similar.However, a more subtle trend can be hypothesized from the relative positions of these clusters. The red cluster (Male, ‘$<= 50K$‘) appears horizontally closer to the overall high-income group than the purple cluster (Female, ‘$<= 50K$‘). This could suggest that while men in the lower-income bracket earn less than \$50K, their incomes may be, on average, closer to the \$50K threshold than their female counterparts. Similarly, the yellow cluster (Female, ‘$> 50K$‘) is positioned closer to the low-income groups than the brown cluster (Male, ‘$> 50K$‘), which might imply that women in the high-income bracket, on average, earn less than the men in the same bracket.It is important to treat these observations as hypotheses generated by the visualization. The binary nature of the income variable (‘$> 50K$‘/‘$<= 50K$‘) limits the depth of this analysis. Access to the **raw income data for each individual**, rather than just a label, would be required to verify these hypotheses. With continuous data, a direct statistical test could be performed to confirm whether a significant pay gap truly exists between genders within each income bracket.

3.8 UMAP-Based Nonlinear Dimensionality Reduction Analysis

While PCA is effective at capturing the linear structure of the data, it may overlook more complex, non-linear relationships. To explore these, **UMAP (Uniform Manifold Approximation and Projection)**, a powerful non-linear technique, was also employed to see if it could reveal more nuanced clusters or separations within the socio-demographic data. UMAP excels at preserving both the local and global structure of a dataset in a low-dimensional embedding.

Data Preparation for UMAP

To ensure meaningful results, all categorical features were one-hot encoded. The continuous `age` feature was binned into 5-year intervals and then one-hot encoded. Given the binary nature of the resulting data, the **Hamming distance** was chosen as the similarity metric, as it is well-suited for calculating the distance between one-hot encoded vectors.

UMAP Embedding and Visualization Workflow

The UMAP embedding was computed in Python using the `umap-learn` library. The resulting 2D coordinates were merged into the original pandas DataFrame and then imported into Tableau to create an interactive scatter plot where each point represents an individual.

Rationale for Parameter Choices

The UMAP parameters were chosen after experimentation to achieve the most visually interpretable plot.

- `n_neighbors=40`: This value was selected to provide a good balance between local and global structure. Lower values produced a fragmented plot with too many small, uninterpretable clusters, while higher values obscured meaningful local patterns.

- `min_dist=0.9`: This high value was chosen to encourage a more uniform spread of points in the final visualization. This approach reduces the formation of overly dense, artificial clusters that can hinder interactive exploration and interpretation.

- `metric='hamming'`: This metric was essential for the one-hot encoded binary vectors, ensuring that the similarity calculations correctly aligned with the categorical nature of the preprocessed data.

Visual Analysis of the UMAP Plot

The UMAP visualization provides a non-linear projection of the data, offering a different perspective on its underlying structure. Unlike the continuous structure seen in the PCA plot, the UMAP embedding is composed of multiple distinct, ”island-like” clusters. Further exploration suggests these clusters are primarily formed based on **age and income groups**, with each ”island” likely representing a specific socio-demographic niche (e.g., young, low-income individuals in a particular workclass). However, a key observation is that, unlike the PCA plot, **no single, clear trend or gradient was observed** in the UMAP visualization. There is no simple axis that represents a progression from low-to-high income across the entire dataset.

Comparison with PCA and Final Insights

Comparing the two dimensionality reduction techniques provides valuable insight into the nature of the dataset. The PCA plot was effective at revealing the most dominant **linear trend** in the data: a single axis (PC1) that clearly separated individuals by their economic standing. This made the overall income disparity easy to visualize as a simple, global pattern.In contrast, the UMAP plot provided a more **fragmented, non-linear view**. By focusing on preserving local neighborhood structures, it was highly effective at identifying specific, highly similar subgroups (the ”islands”). However, in doing so, it did not produce a simple, interpretable trend that could be used to analyze the overall income gap across the entire population. Ultimately, this suggests that while a linear model (PCA) can capture the main economic storyline, the underlying socio-demographic relationships are highly complex and localized. UMAP's result, while less a single clear trend, more accurately reflects this complexity, showing that the population is a collection of distinct subgroups rather than a smooth, continuous whole.

4 Conclusion

This report has provided a detailed analysis of gender-based income disparities using the 1994 U.S. Census ”Adult” dataset, applying modern data science techniques to uncover the most significant predictors of the gender pay gap. Even after accounting for factors such as age, education, hours worked, race, marital status, and work sector, a statistically significant pay gap persists between men and women. The analysis identified education level, hours worked, and marital status as the most influential predictors of income, but these factors do not fully explain the observed disparities.By placing the historical data in the context of current research from the Pew Research Centre and Forbes, this study demonstrates that, while the gender pay gap has narrowed over the past three decades, it remains a persistent feature of the labor market. The drivers of inequality have shifted from overt occupational segregation to more nuanced issues, such as the impact of caregiving responsibilities, differences in career progression, and the intersection of gender with race and family structure.These findings highlight the need for continued policy efforts to promote pay transparency, support working parents, and address structural barriers that disproportionately impact women. Future research should build on this work by incorporating more recent datasets and investigating how the gender pay gap evolves in response to changes in work patterns, family dynamics, and broader economic conditions.In summary, this analysis reinforces the enduring nature of the gender pay gap and offers actionable insights for policymakers, employers, and advocates committed to advancing pay equity in the modern workforce.

References

[1] Forbes Advisor. *Gender Pay Gap Statistics In 2025*. Accessed 2 July 2025. 2025. URL: https://www.forbes.com/advisor/business/gender-pay-gap-statistics/.

[2] Pew Research Center. *The Enduring Grip of the Gender Pay Gap*. 2023. URL: https://www.pewresearch.org/short-reads/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/.

[3] Pew Research Centre. *The Enduring Grip of the Gender Pay Gap*. Accessed 2 July 2025. 2025. URL: https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/.

[4] D. Dua and C. Graff. *UCI Machine Learning Repository: Adult Data Set*. 2019. URL: https://archive.ics.uci.edu/ml/datasets/Adult.

[5] Forbes. *Gender Pay Gap Statistics 2023*. 2023. URL: https://www.forbes.com/advisor/business/gender-pay-gap-statistics/.