DSA210-Term-Project: "Analysis of the Impact of Education Level, Demographic Factors and Inflation on Wage Disparities (2000-2022)"

This project was prepared by **Sıla Kara** as a term project for the **DSA 210** course at **Sabancı University** in the **2025 Spring semester**. It examines the impact of education level and demographic factors on real wage disparities over the years from **2000 to 2022**.

In accordance with the goals of the project, two hypotheses will be tested:

Hypothesis 1: The real wage gap between men and women is significantly larger at higher education levels (Bachelor’s + Advanced) than at lower education levels (Some College + High School + Less than High School) over the period 2000–2022.

Hypothesis 2: For each education level, the real wage gap between White and Black individuals has not significantly changed between the two time periods: 2000–2010 and 2011–2022.

Contents

Important Note on Terminology:
Throughout this project, the term **“wage gap”** is used to refer to **differences in average real wages** between demographic groups (e.g., men and women, Black and White individuals) across education levels. However, it is important to clarify that this analysis does **not control for job-specific variables** such as occupation, industry, experience, or working hours. As such, these differences represent **economic disparities across groups**, not necessarily **"unequal pay for equal work"**. A true wage gap analysis would require comparing individuals with **similar roles and qualifications** within the same occupational categories.

Motivation

As a university student, I have always been interested in the relationship between education, socioeconomic factors, and financial stability. Understanding how wages vary based on education level, gender, and ethnicity, and how inflation further affects these disparities, is not only important for academic research but also crucial for gaining a better understanding of the dynamics of professional life as a future employee. Through this analysis, I will have the opportunity to apply what I have learned in class and gain hands-on experience in data analysis, while also developing a deeper understanding of the socioeconomic factors underlying wage inequality.

Project Goal

This project aims to shed light on the underlying factors of wage inequality over the years and examine the impact of inflation on this disparity. In this context, comparisons will be made regarding the wages of individuals from different groups over the years by considering factors such as education level, gender, and ethnicity. As a conclusion of this project it is aimed to identify long-term wage inequality trends and examine how education, demographics, and inflation impact economic mobility. The findings may help better understand income disparities and support future policies to reduce wage gaps.

Data Collection

For this project, data was collected from multiple publicly available sources to ensure a comprehensive and reliable dataset. The main steps in the data collection process were as follows:

1. **Researching Available Datasets:** I explored various sources, including Kaggle, government economic databases, and academic repositories, to find relevant datasets on wages and inflation. Keywords such as "historical wage data," "inflation rates dataset," and "wage inequality statistics" were used to search for structured datasets.
2. **Selecting the Most Relevant Data:** After reviewing multiple datasets, I prioritized those that covered the years 2000-2022 and included necessary details such as wage levels by education, gender, and ethnicity, along with annual inflation rates. Datasets with missing critical information or inconsistencies were excluded.

3. **Downloading and Storing Data:** The chosen datasets were downloaded in CSV format from Kaggle. To keep the data organized, all files were stored in a structured directory, ensuring easy access during the preprocessing phase.

Data Sources and Preprocessing

Data Sources

1. **Wages by Education**
    - This dataset contains **average wages by education level and gender** for the years **2000-2022**. The data includes:
        - Education levels: Less Than High School, High School, Some College, Bachelor's Degree, Advanced Degree.
        - Gender-based wages for each education level.
    - Information on **average wages by education level** was provided, which served as the foundation for the wage gap analysis.
2. **Inflation Data**
    - The inflation dataset includes **annual inflation rates** for the years **2000-2022**.
    - Inflation rates were used to **adjust nominal wages to real wages**, making the data comparable across years.

Preprocessing

Objective:

The preprocessing aimed to merge the wage data with inflation data in order to calculate **real wages**. Real wages, adjusted for inflation, provide a more accurate representation of the actual purchasing power over time.

Steps:

1. **Inflation Adjustment:**
    - **Nominal wages** were adjusted using the **annual inflation rates** to obtain the **real wages**.
    - The real wage calculation formula used was:
    - Real Wage = Nominal Wage / (1 + Inflation Rate)

    - This adjustment allowed for a **comparison of wages over time**, eliminating the influence of inflation and providing the **true value** of wages in terms of purchasing power.

2. **Data Merging:**
    - The **wages_by_education** dataset was merged with the **inflation** dataset.
    - The merging step aligned the **yearly wage data** (by education level and gender) with the corresponding **inflation rates** for each year, ensuring that wages were adjusted correctly.

3. **Real Wages Cleaned Dataset:**
    - The **real_wages_cleaned** dataset was created using the merged data, which contained **real wages** adjusted for inflation.
    - This dataset includes:
        - Real wages for each year, education level, and gender.
        - The data is now **comparable across years**, providing a true picture of wage changes over time.

Why Was the Data Merged?

The merging process was essential to calculate the **real wages**, adjusted for inflation. Without this adjustment, nominal wage differences across years would not provide meaningful insights. By incorporating inflation rates, it was ensured that **wage comparisons** reflected the **true economic impact** over time, providing more accurate insights into the wage gap across different education levels and genders.

The final **real_wages_cleaned** dataset facilitated accurate comparisons of **wage gaps** by **education level** and **gender**, offering a reliable foundation for hypothesis testing and further analysis.

Data Analysis

1. Data Preprocessing

As outlined in the **Data Sources and Preprocessing** section, two primary datasets were used:

- **Wages by Education**: This dataset contained wage information segmented by **education level** and **gender** for the years **2000–2022**.

- **Inflation Data**: This dataset included **annual inflation rates** for the same period.

These datasets were merged to adjust the nominal wages for inflation, resulting in the **real_wages_cleaned** dataset. This dataset provided **real wages** for each year, education level, and gender, enabling accurate comparisons over time.

---

2. Exploratory Data Analysis (EDA)

The **wage gap** between **men and women** for each education level was analyzed over two time periods: **2000–2010** and **2011–2022**. Findings indicated that **high education levels** experienced a **larger wage gap** compared to **low education levels**. Over time, the gap increased for high education levels, while low education levels remained relatively stable.

---

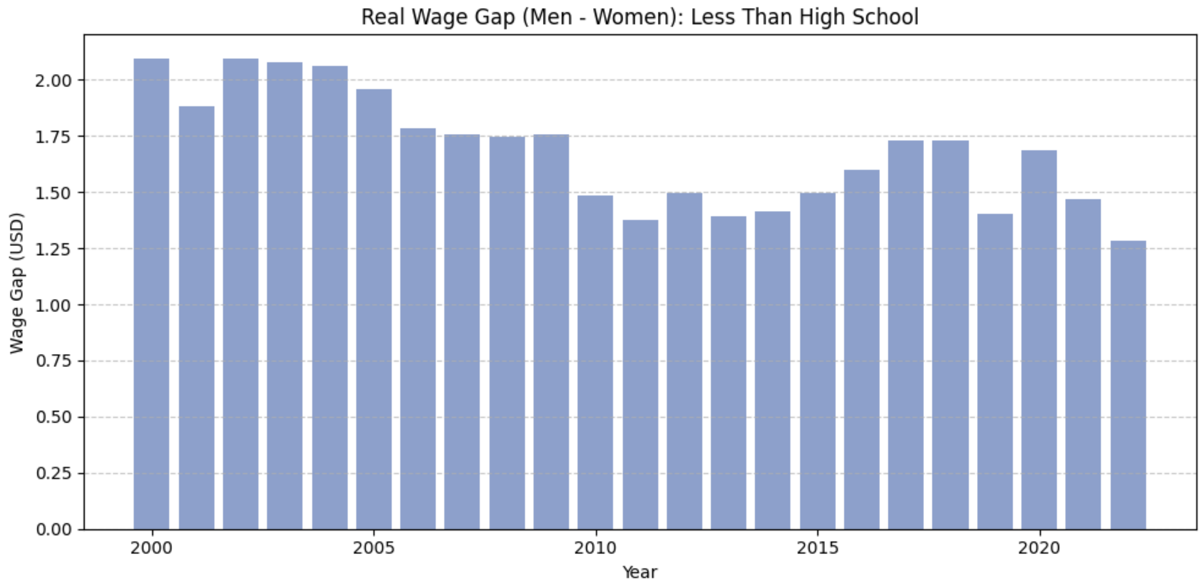
3. Visualization

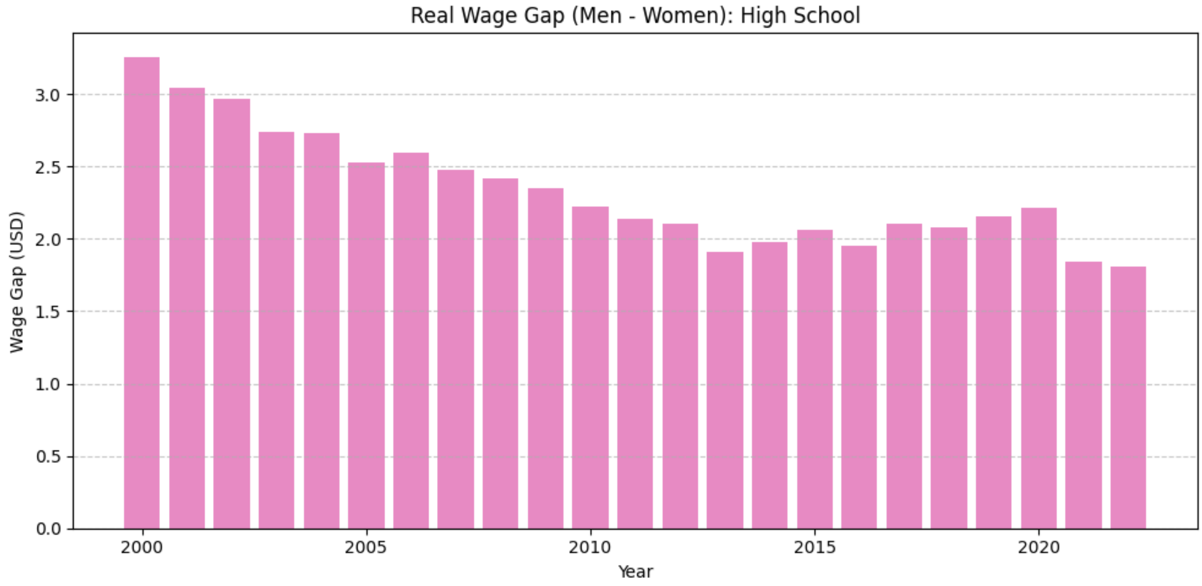
Bar Charts, **Boxplots**, and **Line Charts** were utilized to visualize the **wage gap trends** and **distributions** across education levels and time periods. These visualizations helped in understanding the patterns and changes in the wage gap between men and women over the years.

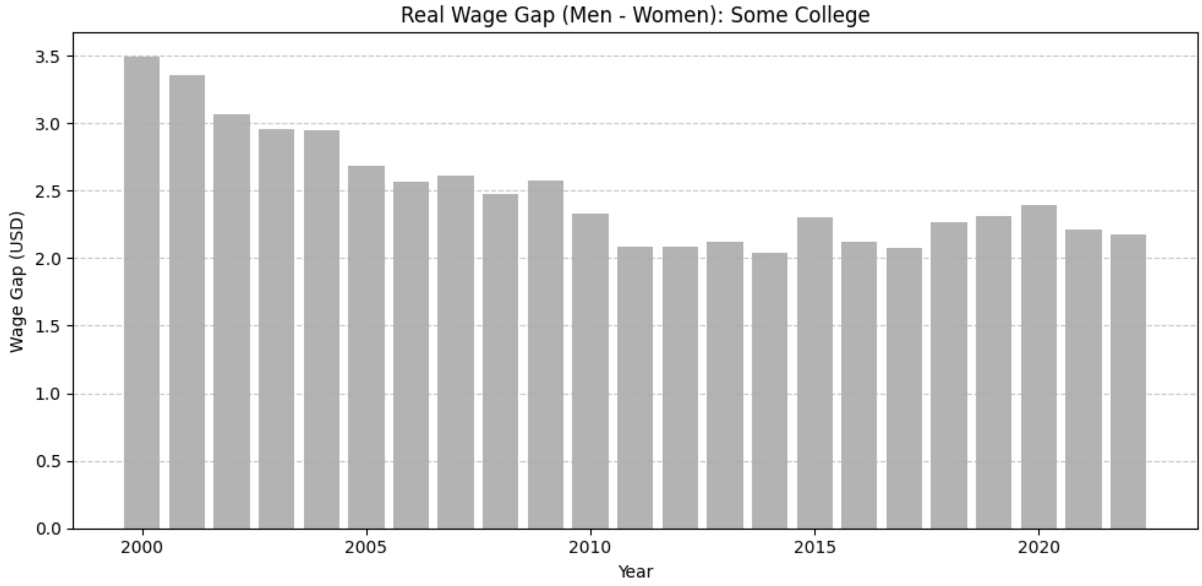
Findings

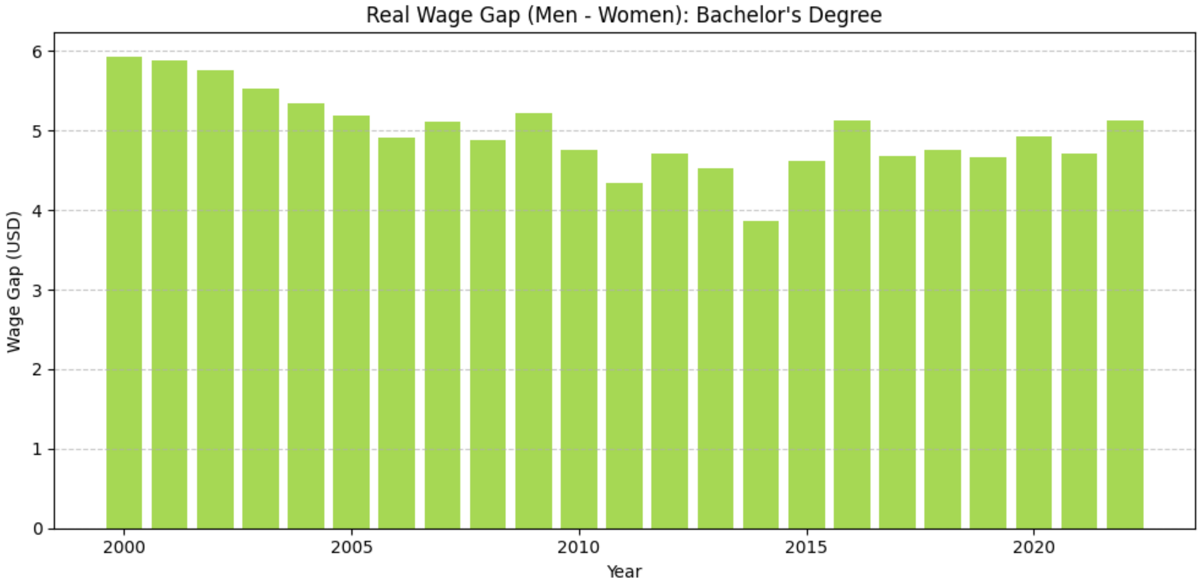
EDA (Visual Summary of The Data)

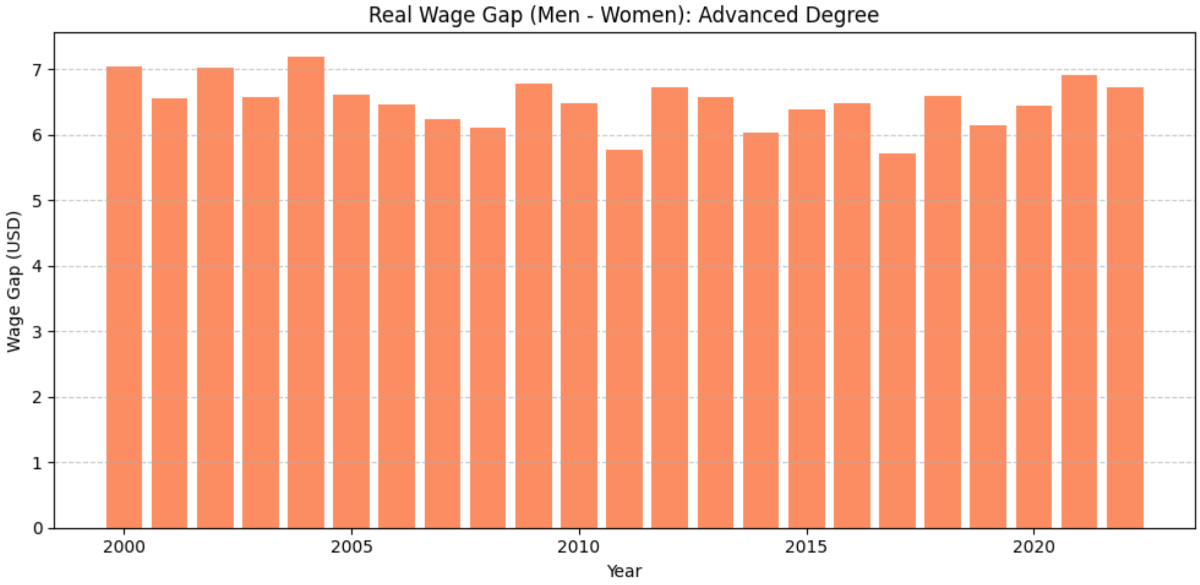
Real Wage Difference Between Men and Women Across Different Educational Background Over The Years (2000-2022)


Real Wage Gap (Men - Women): Less Than High School
Wage Gap (USD)
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2000
2005
2010
2015
2020
Year


Real Wage Gap (Men - Women): High School
Wage Gap (USD)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2000
2005
2010
2015
2020
Year

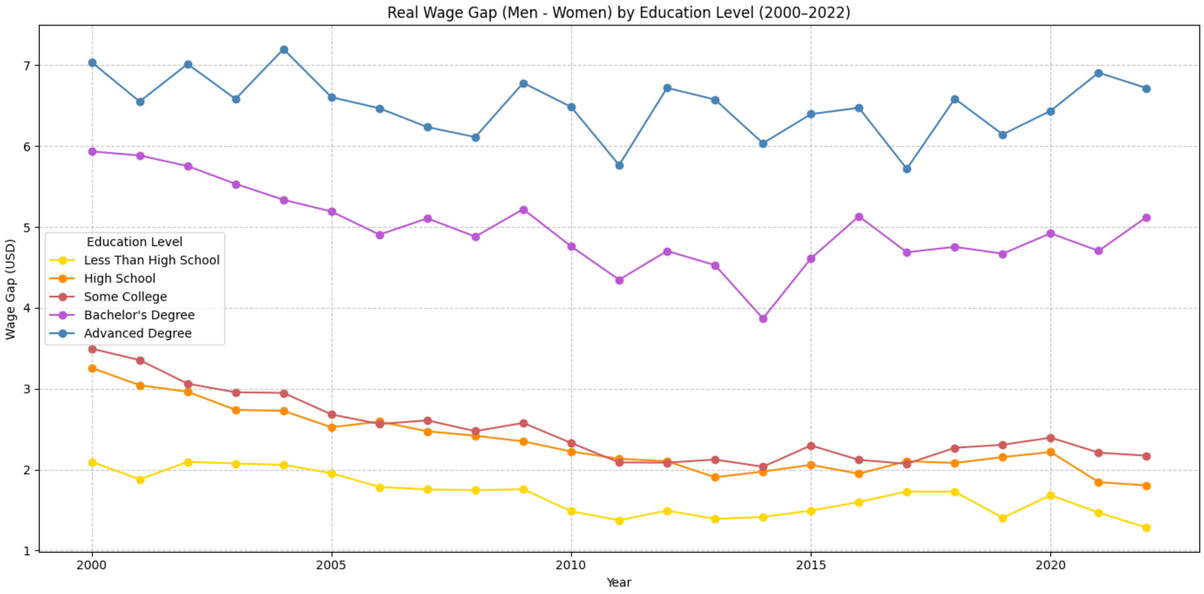

Real Wage Gap (Men - Women): Some College
Wage Gap (USD)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
2000
2005
2010
2015
2020
Year





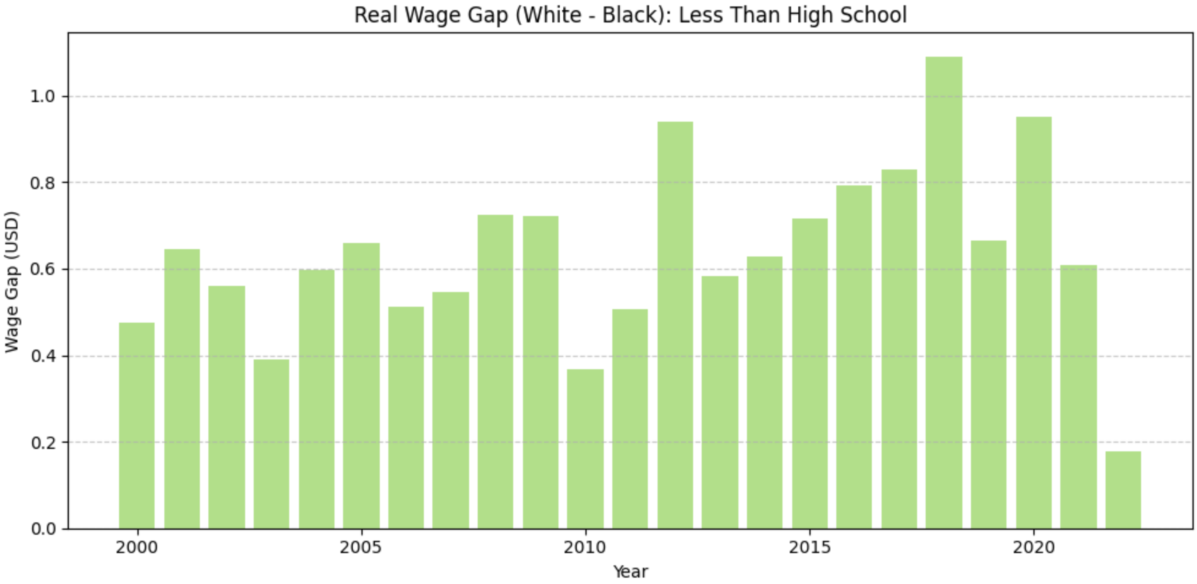
Overview: These charts display the real hourly wage gap between men and women from 2000 to 2022, segmented by education level. Across all educational categories, men consistently earn more than women. The gap is especially prominent at the “Bachelor’s Degree” and “Advanced Degree” levels, hovering around 6–7 USD. While the wage gap has slightly declined over time in the “High School” and “Some College” categories, it has never closed. Notably, as the level of education increases, the gender wage gap also widens. This pattern indicates that even highly educated women continue to face systematic wage disadvantages compared to men, reflecting the ongoing influence of structural inequality and the glass ceiling effect.

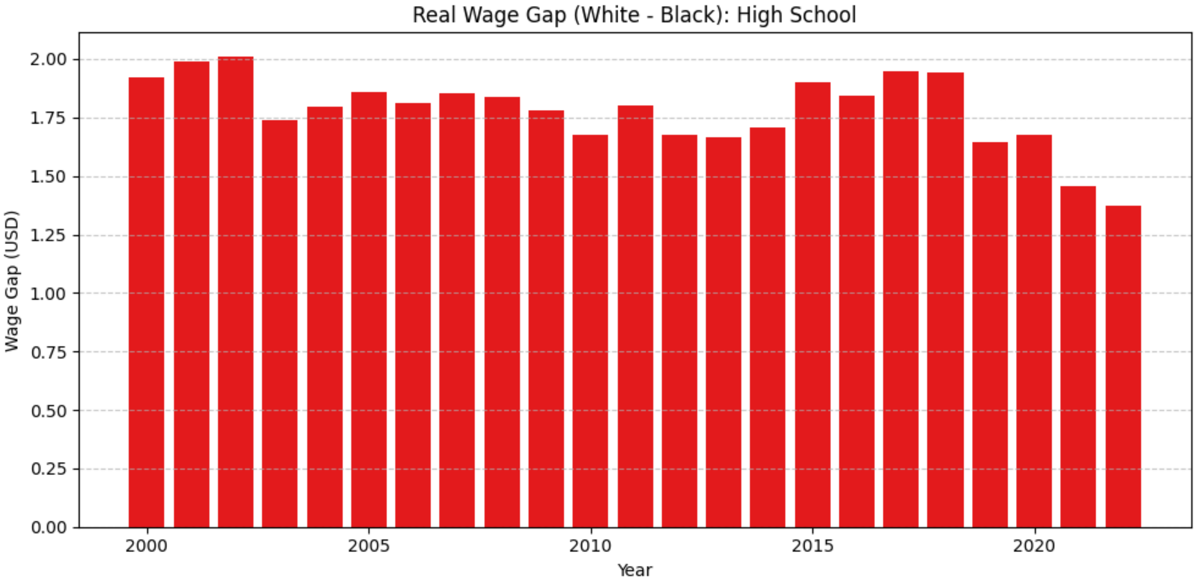
Line chart that shows the real wage gap difference between men and women from all educational backgrounds in the same graph

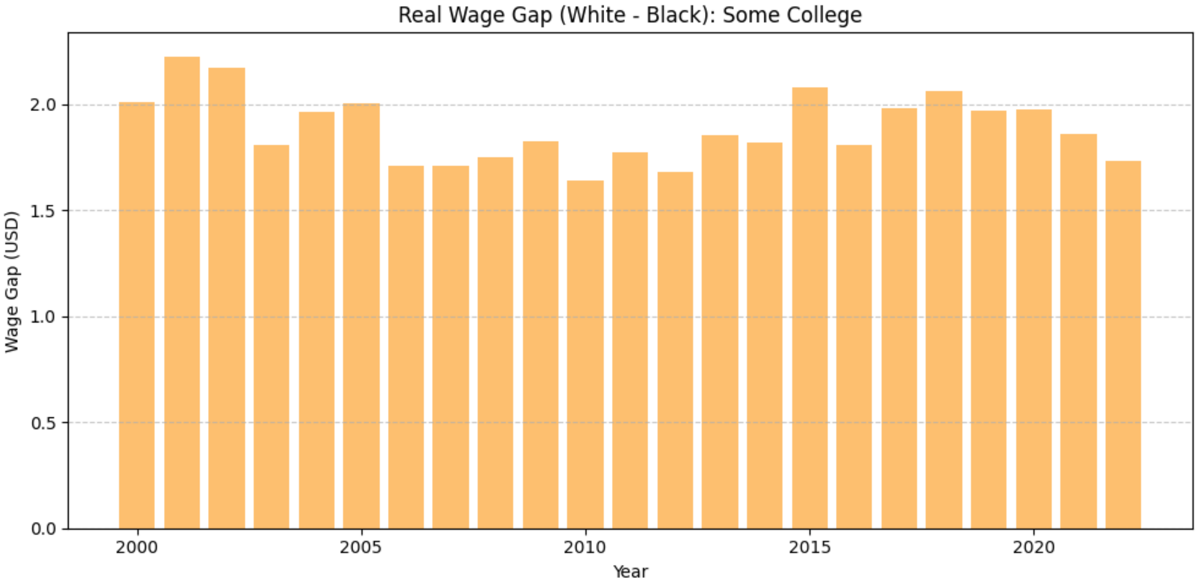


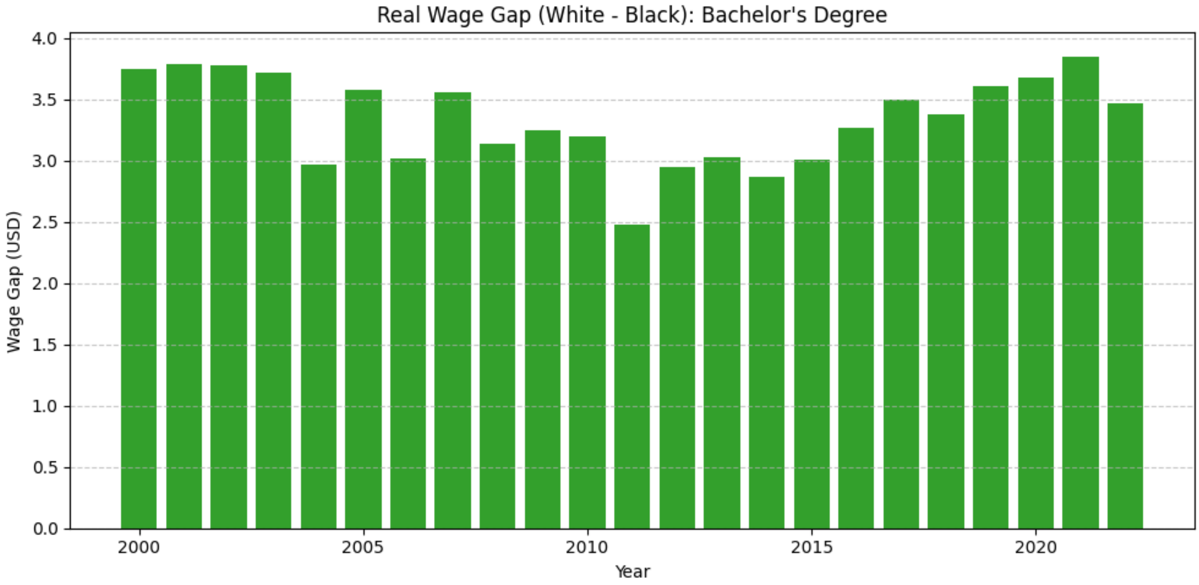
Overview: The chart shows the real wage gap between men and women from 2000 to 2022 by education level. As education increases, the wage gap widens, with the largest differences seen at the advanced degree level. While the gap has slightly narrowed for lower education levels, gender-based wage inequality remains persistent and pronounced among higher-educated individuals.

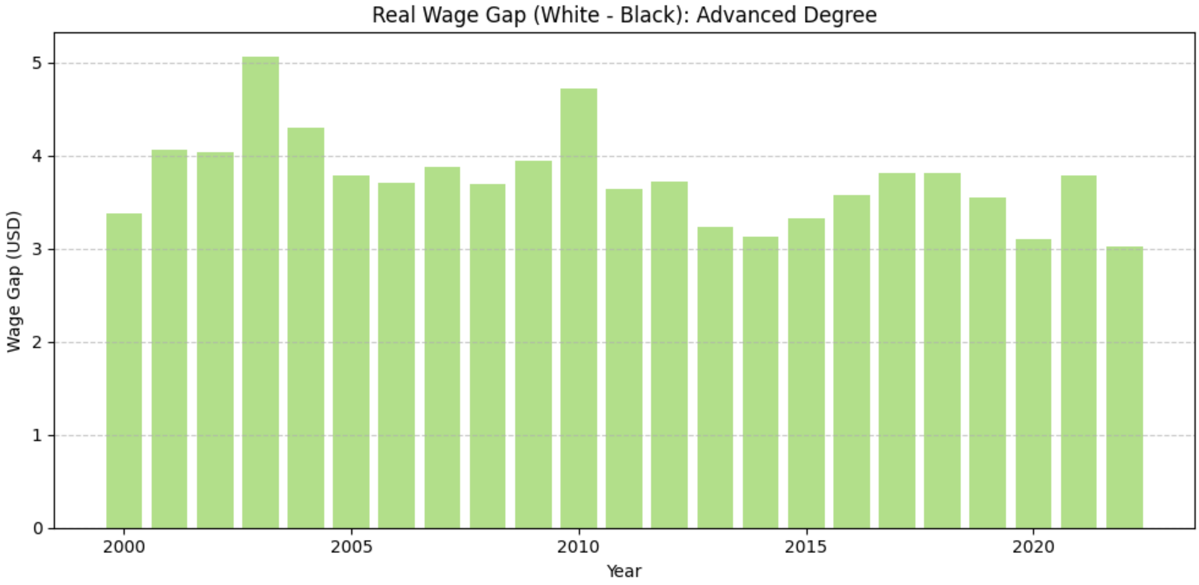
Real Wage Difference Between Black and White Individuals Across Different Educational Background Over The Years (2000-2022)


Real Wage Gap (White - Black): Less Than High School
Wage Gap (USD)
0.0
0.2
0.4
0.6
0.8
1.0
2000
2005
2010
2015
2020
Year


Real Wage Gap (White - Black): High School
Wage Gap (USD)
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2000
2005
2010
2015
2020
Year


Real Wage Gap (White - Black): Some College
Wage Gap (USD)
0.0
0.5
1.0
1.5
2.0
2000
2005
2010
2015
2020
Year

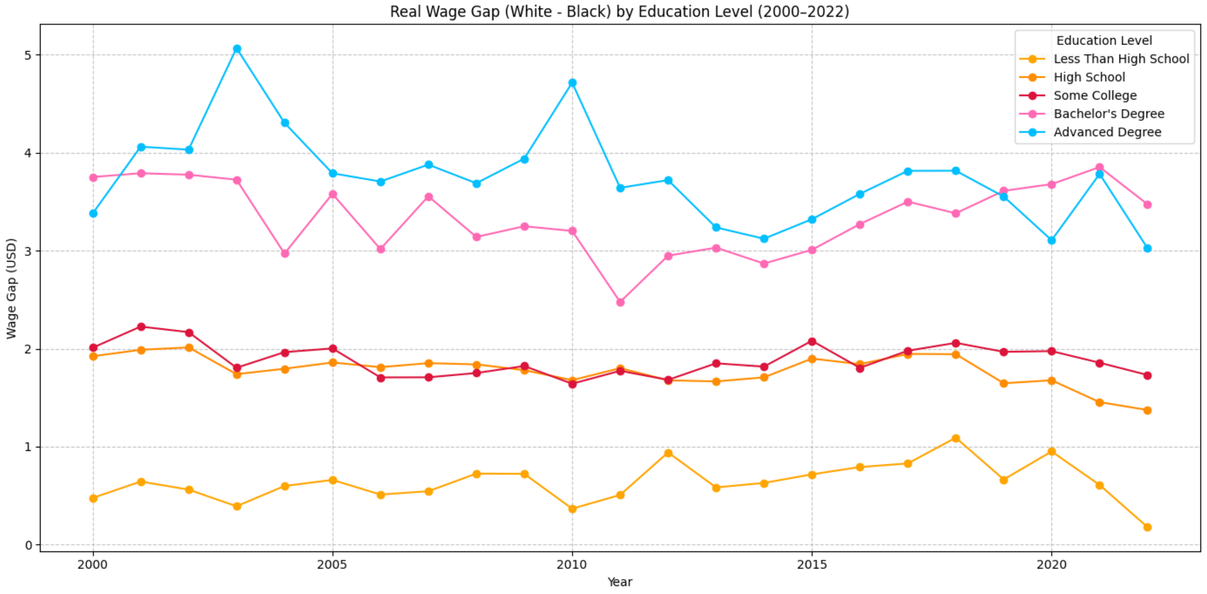




Overview:

These graphs illustrate the real hourly wage gap between White and Black individuals from 2000 to 2022, segmented by education level (Less than High School, High School, Some College, Bachelor's Degree, Advanced Degree). Overall, the wage gap tends to widen with higher education levels. In the "Less than High School" category, the wage gap is low and more variable, whereas in the "Advanced Degree" category, it consistently ranges between 3–5 USD. This trend indicates that racial wage inequality becomes more pronounced as educational attainment increases. Especially in Bachelor's and Advanced Degrees, the gap remains persistently high, revealing that even highly educated Black individuals continue to face significant income disparities compared to their White counterparts.

Line chart that shows the real wage gap difference between black and white individuals from all educational backgrounds in the same graph



Hypothesis Testing

Hypothesis 1: Wage Gap Between Men and Women

For the first hypothesis, **the real wage gap between men and women at different education levels** has been visualized using **five separate bar charts**. Each bar chart represents the wage gap for a specific education level. The bar charts clearly showcase the differences in wages between men and women across various education categories.

Additionally, to provide a more comprehensive comparison, the data from all five education levels was combined into a **single line chart**. This chart illustrates the changes in wage gaps across the years for both men and women, allowing for an easier visual comparison of trends.

Hypothesis 2: Wage Gap Between White and Black Individuals

For the second hypothesis, **the real wage gap between Black and White individuals at different education levels** has been visualized using a **bar chart**. This bar chart highlights the wage disparities between the two groups across each education level.

Similarly to the first hypothesis, the wage gaps for each education level were combined into a **single line chart**. This line chart depicts the wage gap trends over time for Black and White individuals, offering a clear visual comparison of how these gaps evolved across the five education levels.

General Insights

These visualizations provide an overall understanding of the wage gaps by education level and allow for meaningful observations about the trends in the wage disparities between different demographic groups. The detailed visualizations of wage differences between men and women, as well as Black and White individuals, present valuable insights into the nature of wage inequality over time.

The **key insights** derived from these visualizations can be found in the corresponding sections of the .ipynb file, which include specific commentary and overview regarding the charts.

Hypothesis 1: Wage Gap Between Men and Women

The first hypothesis proposed that the **real wage gap** between men and women is significantly larger at **higher education levels** (Bachelor's + Advanced) compared to **lower education levels** (Some College + High School + Less than High School) over the period 2000–2022.

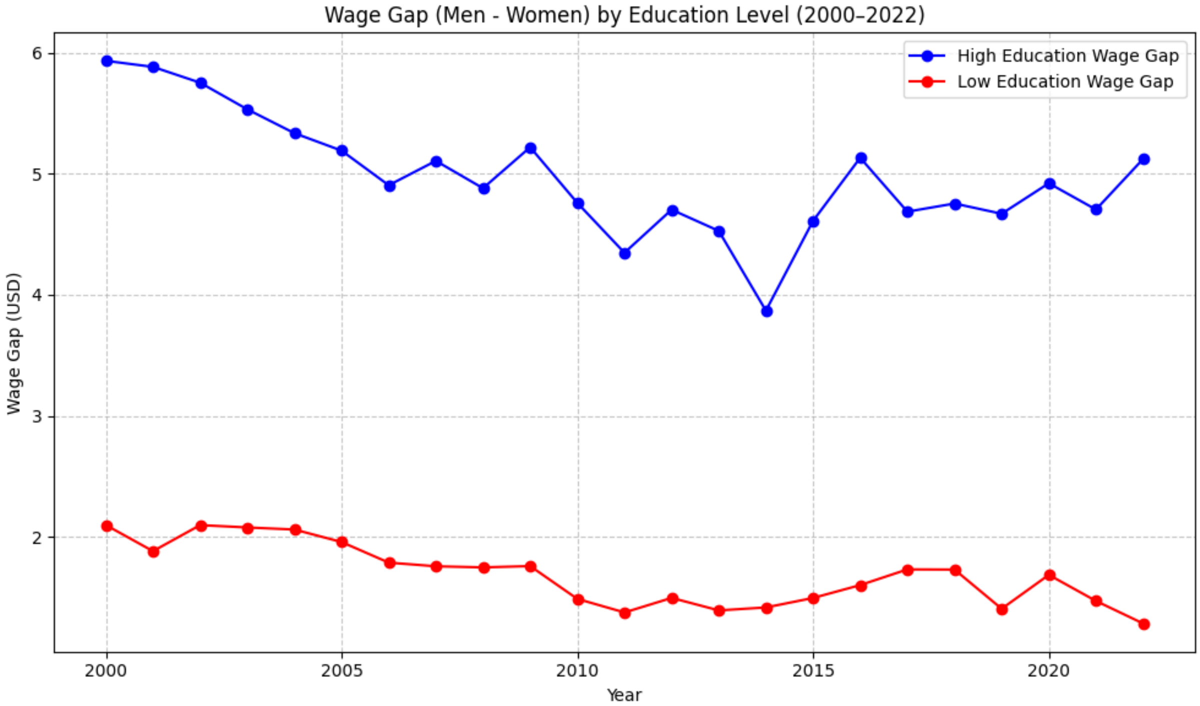
Null Hypothesis (H_0):

"There is no significant difference in the real wage gap between men and women across higher education levels (Bachelor's + Advanced) and lower education levels (Some College + High School + Less than High School) over the period 2000–2022."

Alternative Hypothesis (H_1):

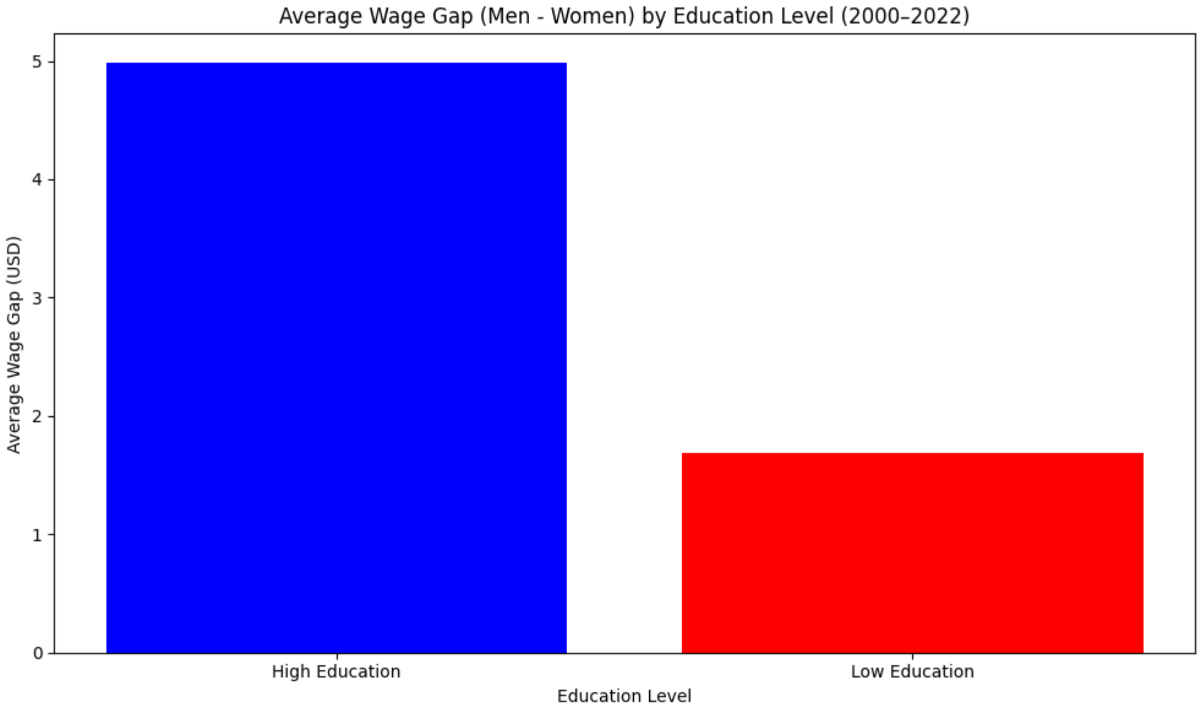
"The real wage gap between men and women is significantly larger at higher education levels (Bachelor's + Advanced) than at lower education levels (Some College + High School + Less than High School) over the period 2000–2022."

Line Chart: Men vs Women Wage Gap by High/Low Education Level (2000–2022)



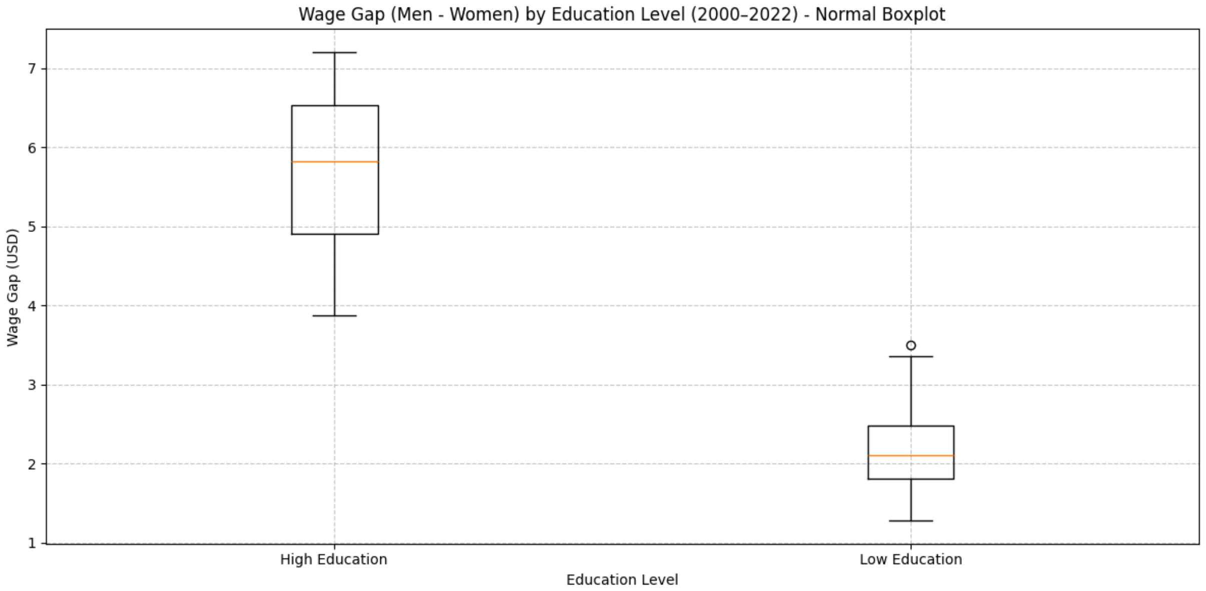
The line chart shows that the **wage gap between men and women is consistently larger at higher education levels** (blue line) compared to lower education levels (red line) from 2000 to 2022. While the gap for low education levels remains relatively stable and small, the high education wage gap remains significantly wider and fluctuates more over time.

Bar Chart: Men vs Women Average Wage Gap by High/Low Education Level (2000–2022)



This bar chart shows that the **average wage gap between men and women is significantly larger for individuals with higher education** levels (around 5 USD) compared to those with lower education levels (around 1.7 USD) over the period 2000–2022, highlighting a stronger gender disparity among the more educated population.

Boxplot: Men vs Women Average Wage Gap by High/Low Education Level (2000–2022)



This boxplot shows that the wage gap between men and women is not only larger but also more variable at higher education levels. The median wage gap for highly educated individuals is significantly greater, and the spread is wider, indicating greater disparity. In contrast, the gap for lower education levels is smaller and more consistent over time.Step-by-Step Explanation of Hypothesis Testing:

Step 1: Data Preparation

We begin by loading the dataset containing real wage information for different education levels and time periods. The data spans 2000–2022, with information on men and women in various education categories:

- **Lower Education Levels**: Less than High School, High School, Some College
- **Higher Education Levels**: Bachelor’s Degree, Advanced Degree

Step 2: Defining Education Levels

To test the hypothesis, we divide the education levels into two groups:

- **Higher Education Levels**: Bachelor’s Degree and Advanced Degree
- **Lower Education Levels**: Some College, High School, and Less than High School

Step 3: Wage Gap Calculation

For each education level, we calculate the **real wage gap** between men and women. The wage gap is defined as: 'Wage Gap = Average Wage of Men - Average Wage of Women'

We calculate the wage gap for both higher and lower education levels over the entire period (2000–2022).

Step 4: Performing the Two-Sample T-Test

We perform a **two-sample t-test** to compare the wage gaps between higher education levels and lower education levels:

- **Null Hypothesis (H_0)**: There is no significant difference in the wage gap between men and women across the higher and lower education levels.
- **Alternative Hypothesis (H_1)**: The wage gap between men and women is significantly larger at higher education levels than at lower education levels.

We perform the t-test to compare the wage gaps for each group (higher and lower education levels). The t-test compares the means of the two groups and tells us if the difference is statistically significant.

Step 5: Interpreting the Results

The t-test returns two main values:

- **T-statistic**: A measure of the difference between the two groups.
- **P-value**: The probability that the observed difference in means is due to chance.

If the p-value is less than the significance level (usually 0.05), we reject the null hypothesis (H_0). This indicates that the wage gap between men and women is significantly different across the two groups (higher vs lower education levels).

If the p-value is greater than 0.05, we fail to reject the null hypothesis (H_0), meaning there is no significant difference in the wage gap between the two education levels.

Results from the Hypothesis Test:

After performing the hypothesis test, we obtained the following results:

- **T-statistic**: 32.73
- **P-value**: 3.32e-32

The p-value is extremely small (much less than 0.05), which means that we reject the null hypothesis (H_0).

Interpretation:

Since the p-value is much smaller than the significance level (0.05), we reject the null hypothesis (H_0). This means that the real wage gap between men and women is significantly larger at higher education levels (Bachelor’s + Advanced) compared to lower education levels (Some College + High School + Less than High School) over the period 2000–2022.

Thus, the alternative hypothesis (H_1) is supported by the data: the wage gap between men and women is indeed significantly larger at higher education levels.

Conclusion:

There is strong evidence to suggest that the real wage gap between men and women has been more pronounced at higher education levels (Bachelor’s and Advanced Degrees) compared to lower education levels (Some College, High School, and Less than High School) over the period 2000–2022.

Hypothesis 2: Wage Gap Between White and Black Individuals Across Time Periods

The second hypothesis proposed that the **real wage gap** between White and Black individuals has significantly changed across two time periods: **2000–2010** and **2011–2022**, for each education level.

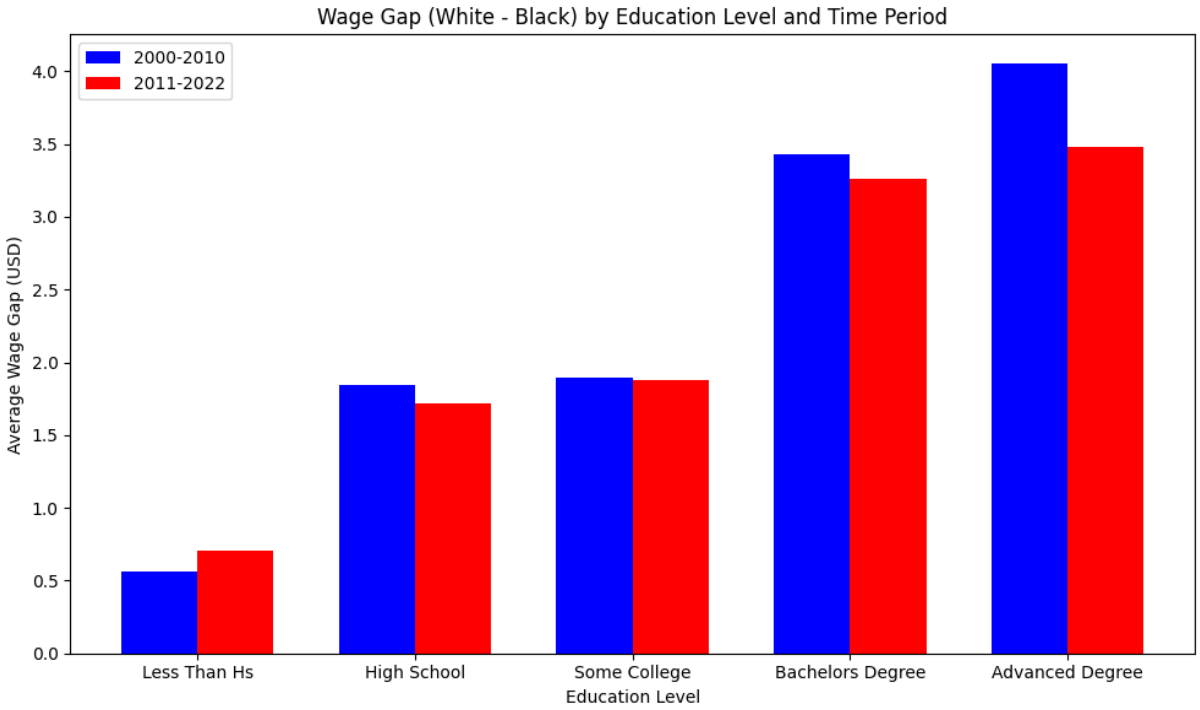
Null Hypothesis (H_0):

"There is no significant difference in the real wage gap between White and Black individuals across the two time periods (2000–2010 and 2011–2022) for each education level."

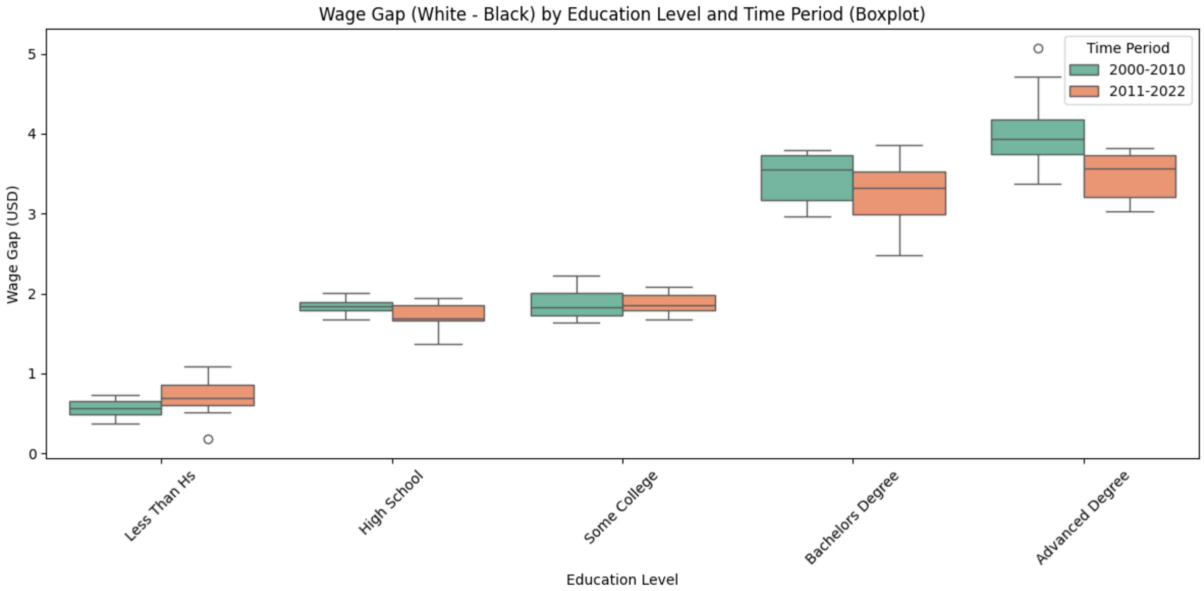
Alternative Hypothesis (H_1):

"There is a significant difference in the real wage gap between White and Black individuals across the two time periods (2000–2010 and 2011–2022) for each education level."

Bar Chart



This bar chart shows that the **racial wage gap between White and Black individuals increases with education level**, and the gap is consistently **higher at advanced education levels** in both periods (2000–2010 and 2011–2022). While a slight decrease is observed in the 2011–2022 period for Bachelor’s and Advanced Degrees, the wage gap remains substantial, indicating persistent inequality.



This boxplot shows that the **wage gap between White and Black individuals widens with higher education levels** in both time periods. While the gap slightly decreased in the 2011–2022 period for Bachelor's and Advanced Degrees, it remains substantial. The variability is also greater at higher education levels, indicating persistent and uneven wage disparities across time.

Step-by-Step Explanation of Hypothesis Testing:

Step 1: Data Preparation

We begin by loading the dataset containing **real wage information** for different education levels and time periods. The data spans **2000–2022**, with information on White and Black individuals in various education categories:

- **Lower Education Levels**: Less than High School, High School, Some College
- **Higher Education Levels**: Bachelor's Degree, Advanced Degree

Step 2: Defining Education Levels

To test the hypothesis, we define two groups of education levels:

- **Higher Education Levels**: Bachelor's Degree and Advanced Degree
- **Lower Education Levels**: Some College, High School, and Less than High School

Step 3: Wage Gap Calculation

For each education level, we calculate the **real wage gap** between **White** and **Black** individuals for both time periods (2000–2010 and 2011–2022). The wage gap is calculated as: Wage Gap = Average Wage of White - Average Wage of Black

We calculate the wage gap for both higher and lower education levels across the two periods.

Step 4: Performing the Two-Sample T-Test

We perform a **two-sample t-test** to compare the wage gaps between 2000–2010 and 2011–2022 for each education level:

- **Null Hypothesis (H_0)**: There is no significant difference in the wage gap between White and Black individuals across the two periods.
- **Alternative Hypothesis (H_1)**: The wage gap between White and Black individuals is significantly different across the two periods.

The t-test compares the means of the wage gaps in both periods to determine if the difference is statistically significant.

Step 5: Interpreting the Results

The t-test produces two key values:

- **T-statistic**: A measure of the difference between the two groups.
- **P-value**: The probability of observing the data if the null hypothesis were true.

If the p-value is less than the significance level (usually 0.05), we reject the null hypothesis (H_0), indicating that there is a significant difference in the wage gap between the two periods. If the p-value is greater than 0.05, we fail to reject the null hypothesis (H_0), meaning there is no significant difference in the wage gap between the two periods.

Results from the Hypothesis Test:

After performing the hypothesis test, we obtained the following results for each education level:

1. **Less than High School:**
   - **T-statistic**: -1.83
   - **P-value**: 0.085
   - **Result**: Fail to reject H_0. The wage gap between White and Black individuals for this education level has not changed significantly between 2000–2010 and 2011–2022.
2. **High School:**
   - **T-statistic**: 2.06
   - **P-value**: 0.054
   - **Result**: Fail to reject H_0. The wage gap between White and Black individuals for this education level has not changed significantly, though it is close to significance.
3. **Some College:**
   - **T-statistic**: 0.15
   - **P-value**: 0.886
   - **Result**: Fail to reject H_0. There is no significant change in the wage gap between White and Black individuals for this education level across the two periods.
4. **Bachelor's Degree:**
   - **T-statistic**: 1.15
   - **P-value**: 0.261
   - **Result**: Fail to reject H_0. There is no significant change in the wage gap between White and Black individuals for this education level.
5. **Advanced Degree:**

- **T-statistic**: 3.39
- **P-value**: 0.0037
- **Result**: Reject H_0. The wage gap between White and Black individuals for this education level may be changed significantly between the two periods.

Interpretation and Conclusion:

Interpretation:

The results show that the wage gap between White and Black individuals has **not changed significantly** across most education levels (Less than High School, High School, Some College, and Bachelor's Degree) between the two time periods (2000–2010 and 2011–2022). Therefore, for these education levels, we **fail to reject the null hypothesis (H_0)**.

However, for individuals with an **Advanced Degree**, there might be a significant difference in the wage gap between White and Black individuals across the two periods. Thus, we **reject the null hypothesis (H_0)** for this group and conclude that the wage gap for individuals with an Advanced Degree has significantly changed over time.

Conclusion:

Even though we reject the null hypothesis for Advanced Degree individuals, the general hypothesis (which applies to all education levels) is affected because there is one exception where the wage gap significantly changed. This exception leads us to conclude that overall, we cannot confidently say there was a consistent change in the wage gap across all education levels. The results for the Advanced Degree group suggest that for this higher education level, there has been a change in the wage gap between White and Black individuals over time. In conclusion, the hyppthesis is not true for all education levels.,

Machine Learning Methods

To predict the real wages of women with a bachelor's degree, three machine learning regression models were developed and evaluated: k-Nearest Neighbors (kNN), Random Forest, and XGBoost. These models were trained using historical wage data, where the target variable was the real wage of women with a bachelor's degree, and the input features included real wages of men, and individuals from different racial and educational backgrounds.

Each model was tested in two stages:

1. **With default hyperparameters, to establish baseline performance.**

2. With optimized hyperparameters, selected through manual tuning, to improve model accuracy and generalization.

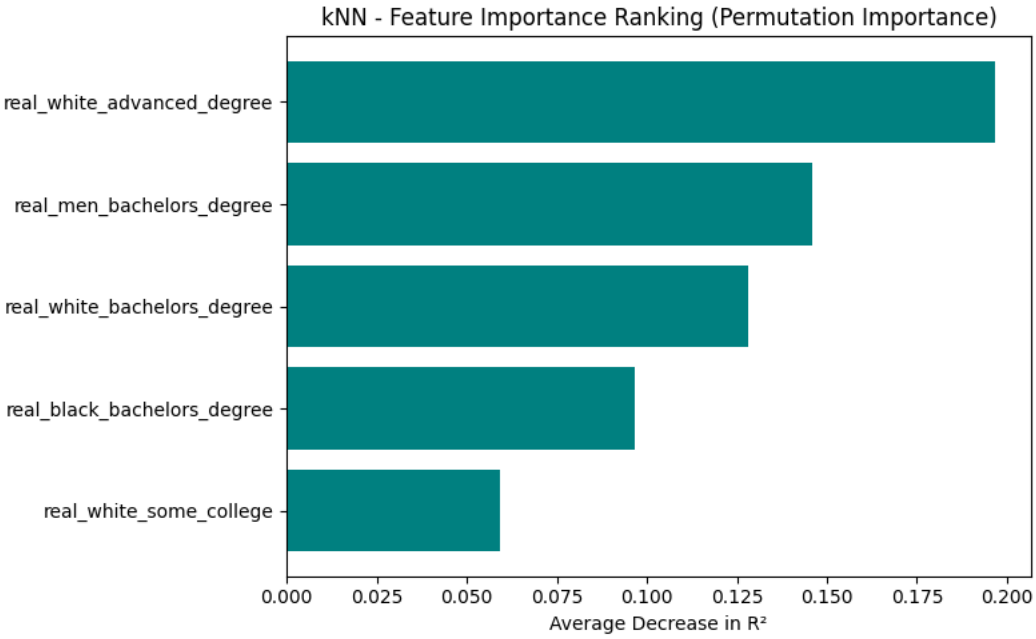
Evaluation Metrics:

- Root Mean Squared Error (RMSE): Reflects the average magnitude of prediction error. Lower RMSE indicates better prediction accuracy.
- R^2 Score: Represents how well the model explains the variance in the target variable. Higher R^2 indicates better explanatory power.

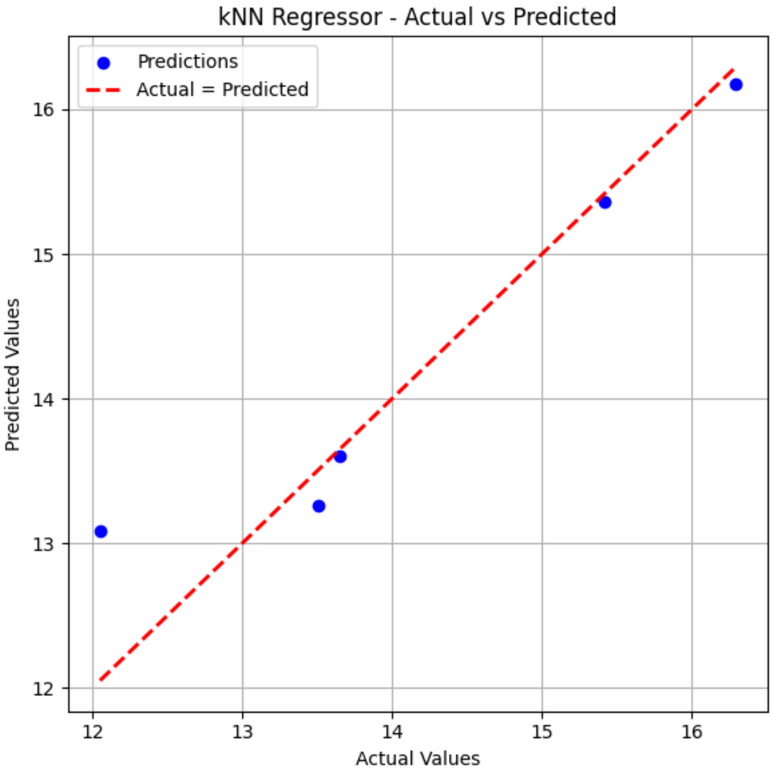
Model Performance & Observations

k-Nearest Neighbors (kNN)

- Default Parameters (k=5): RMSE = 0.634, R^2 = 0.821
- Best Parameters (k=1): RMSE = 0.522, R^2 = 0.879
  Although kNN achieved relatively good results with k=1, the model becomes highly sensitive to noise and individual fluctuations in the data. While simple and interpretable, it tends to overfit and lacks scalability.



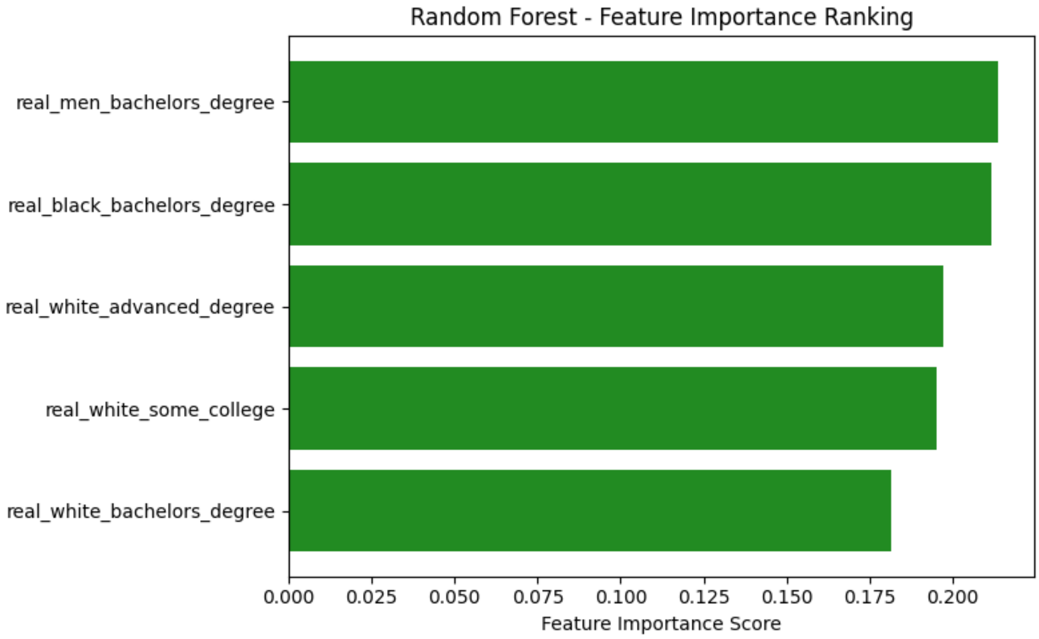
This chart shows the permutation-based feature importance for the kNN model. The real wage of White individuals with an advanced degree had the greatest impact on predicting women's bachelor-level wages, followed by the wages of men and White individuals with a bachelor's degree. Features related to lower education levels had less influence on the model's predictive power.



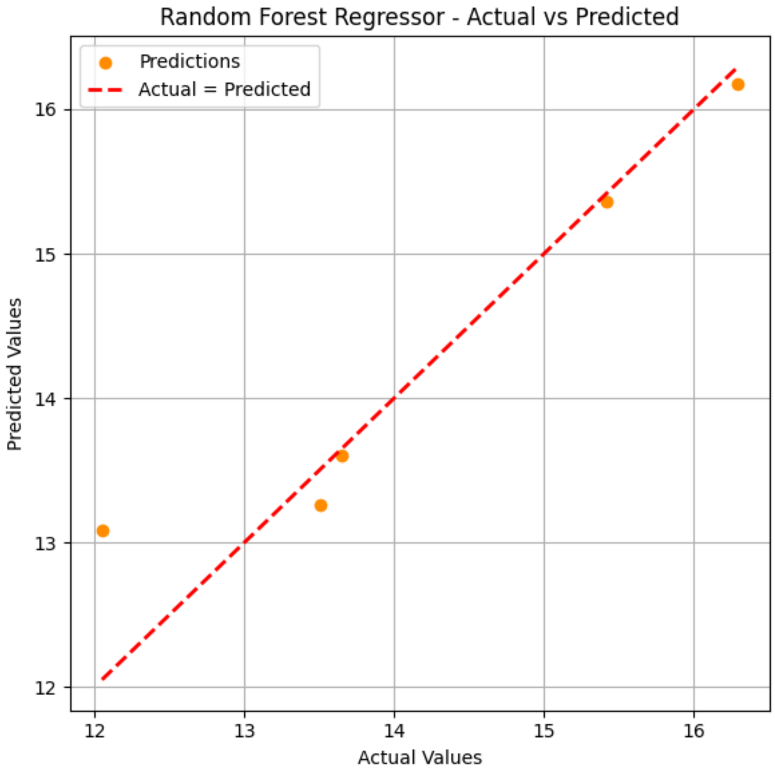
This scatter plot compares actual versus predicted values using the kNN Regressor. Most points lie close to the red dashed line (where prediction equals actual), indicating a good fit. However, slight deviations suggest that while the model performs well, it may still have minor prediction errors on certain data points.

Random Forest Regressor

- Default Parameters: RMSE = 0.479, R^2 = 0.898
- Best Parameters (n_estimators=200, max_depth=10): RMSE = 0.474, R^2 = 0.900
  Random Forest achieved the lowest prediction error and the highest explained variance. It combines predictions from multiple decision trees to reduce overfitting and is robust to noise. It also provides built-in feature importance metrics, enhancing model interpretability.



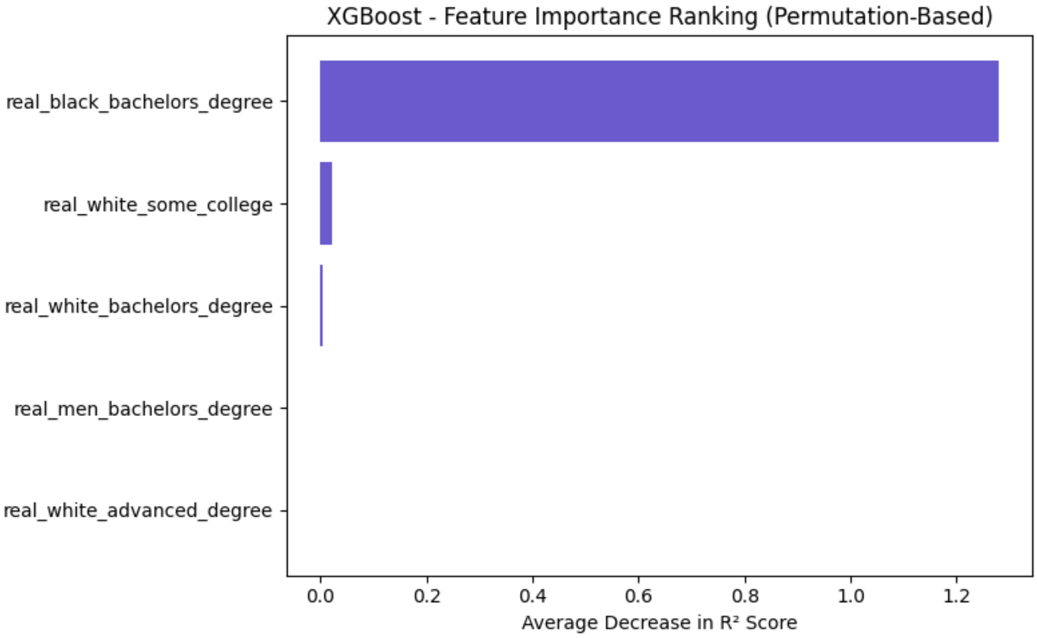
This bar chart shows that in the Random Forest model, all five features contribute significantly to predicting women's real wages with a bachelor's degree. The most influential variables are the real wages of men and Black individuals with the same degree, followed closely by advanced degree and some college wage data, suggesting a well-distributed importance across educational and demographic factors.



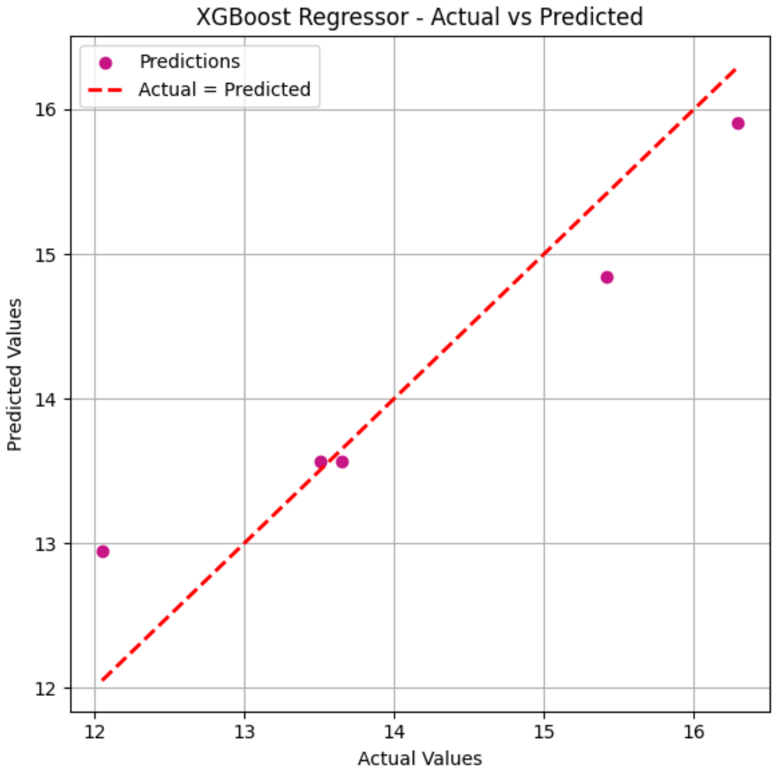
This scatter plot shows that the Random Forest model's predictions closely align with the actual wage values. Most points lie near the red dashed line (ideal prediction line), indicating high accuracy and a strong fit between predicted and true values.

XGBoost Regressor

- **Default Parameters: RMSE = 0.507, R^2 = 0.886**
- **Best Parameters (n_estimators=150, max_depth=4, learning_rate=0.2): RMSE = 0.508, R^2 = 0.885**
  XGBoost offered high predictive accuracy and strong performance on non-linear relationships, but did not outperform Random Forest. It also requires advanced tools (e.g., SHAP) for interpretability and involves greater computational complexity.



This permutation importance chart for the XGBoost model shows that real_black_bachelors_degree is by far the most influential feature in predicting women's real wages with a bachelor's degree. The other variables contribute very little to the model's predictive power, as indicated by their near-zero R^2 decreases when permuted.

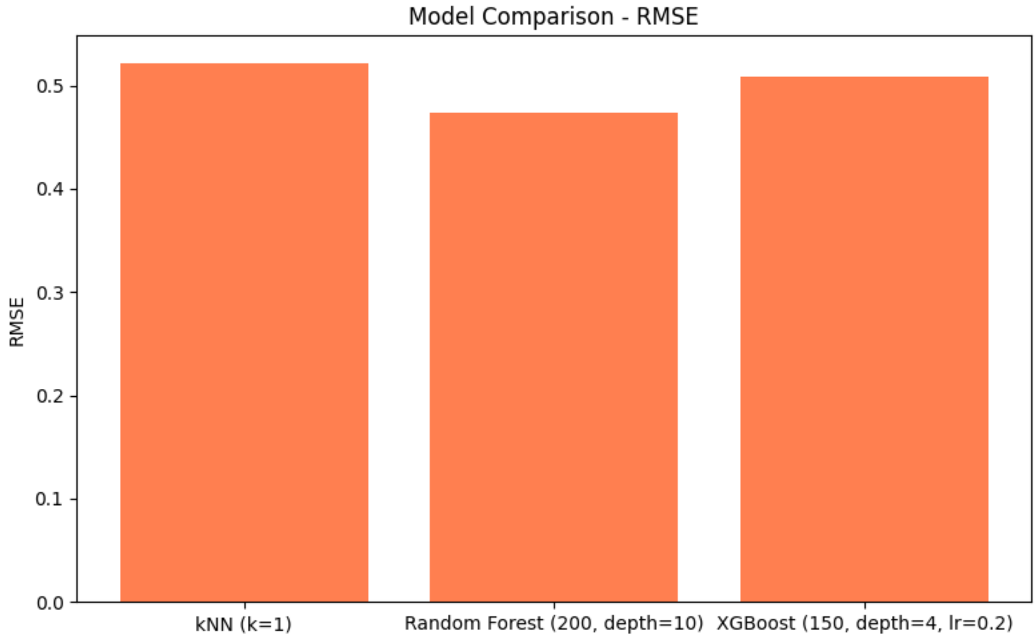


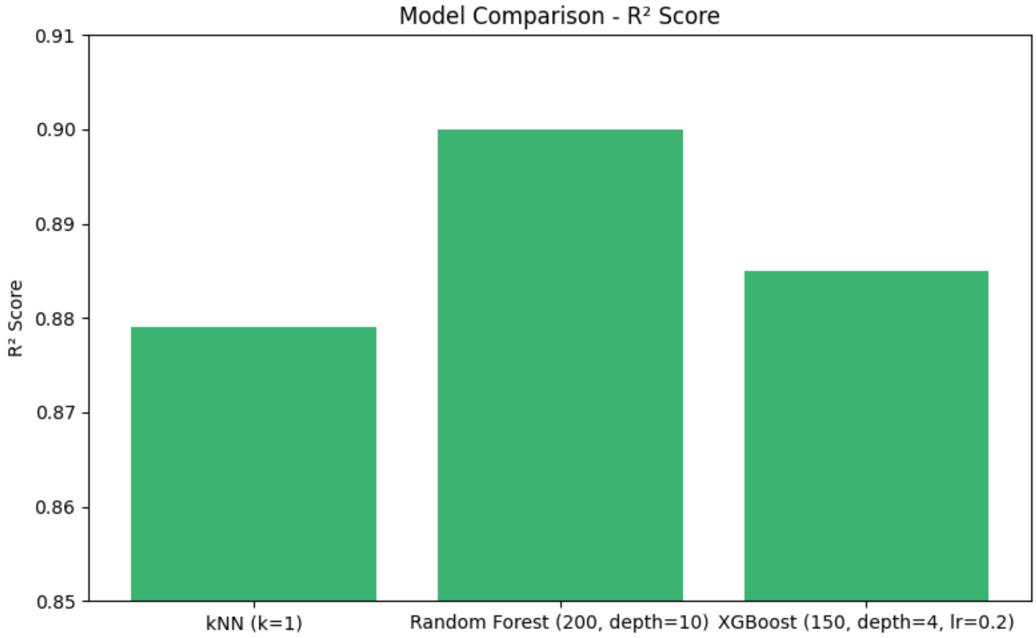
This actual vs. predicted plot for the XGBoost Regressor shows that the model performs well, with predictions closely following the red diagonal line (ideal prediction). However, minor deviations, particularly for higher wage values, suggest some underestimation by the model in those cases. Overall, it demonstrates strong alignment between actual and predicted values.

Model Comparison: Complexity & Interpretability

- **kNN (k=1):**
  - ***Complexity:* Low**
  - ***Interpretability:* High – intuitive, but lacks feature transparency**
  - ***Limitation:* Overfitting risk and poor scalability**
- **Random Forest (n=200, depth=10):**
  - ***Complexity:* Medium**
  - ***Interpretability:* Medium – clear feature importance**
  - ***Strength:* Balanced accuracy, robustness, and explanation power**
- **XGBoost (n=150, depth=4, lr=0.2):**

- ***Complexity:* High**
- ***Interpretability:* Moderate – requires external tools**
- ***Strength:* Handles complex patterns, but less transparent**





Conclusion

While all three models were able to predict real wages with reasonable accuracy, Random Forest Regressor clearly outperformed the others when evaluated under both default and tuned parameters. It combined the lowest RMSE (0.474) with the highest R^2 score (0.900), and offered a practical level of interpretability.

Therefore, Random Forest with n_estimators = 200 and max_depth = 10 is recommended as the final model for deployment. It presents the best trade-off between accuracy, robustness, and usability in practical analysis settings.

Limitations and Future Work

Limitations

One key limitation of this study is the reliance on publicly available datasets, which may have inherent limitations in terms of data completeness and consistency. For example, wage data across different demographic groups may not fully capture regional differences, industry-specific trends, or part-time versus full-time employment distinctions. These unobserved heterogeneities may affect the precision of wage gap estimations.

Second, although the analysis covers a comprehensive time period (2000–2022), the use of annual averages may obscure short-term fluctuations or economic shocks (e.g., recessions or policy reforms) that influence wage dynamics in the short run.

Additionally, the study focuses only on selected demographic variables (gender and race) and does not include other potentially influential factors such as age, marital status, job sector, geographic location, or experience level. This could limit the scope of conclusions regarding the drivers of wage inequality.

Another limitation relates to model complexity versus interpretability. While ensemble models like Random Forest and XGBoost provided strong predictive performance, they function as black-box models and require additional tools (e.g., SHAP) to extract deeper insights into the decision-making process.

Finally, although the dataset was adjusted for inflation, the inflation rates used were national averages. Regional differences in cost of living were not considered, which could lead to an over- or underestimation of "real" wage disparities for specific subpopulations.

Future Work

Future studies could benefit from incorporating more granular demographic and employment information, such as age cohorts, employment sectors, or regional wage indicators. This would help uncover more nuanced patterns in wage inequality across different population segments.

Moreover, integrating additional social and economic variables—such as household income, education quality, job type (public vs. private), or access to

childcare—could offer a more holistic understanding of wage dynamics and their underlying causes.

The application of advanced machine learning methods, such as SHAP analysis or causal inference techniques, could provide deeper explanations of the factors influencing wage disparities and help identify potential intervention points.

Expanding the scope of prediction to include other wage-related outcomes (e.g., wage growth, job mobility, or underemployment rates) may also offer valuable insights into long-term career trajectories and economic mobility.

Lastly, future work could explore policy simulation models to examine how hypothetical changes in legislation (e.g., equal pay laws, affirmative action) might affect wage distributions across gender and race over time.

Important Note

While preparing this assignment, assistance was received from artificial intelligence (ChatGPT and Gemini) in the preprocessing, visualization and hypothesis testing sections.