Impact of Women's Education Levels on Wages and Birth Rates

DSC 630 T303

Allison Biediger
Kathryn Rominger

01 - Introduction

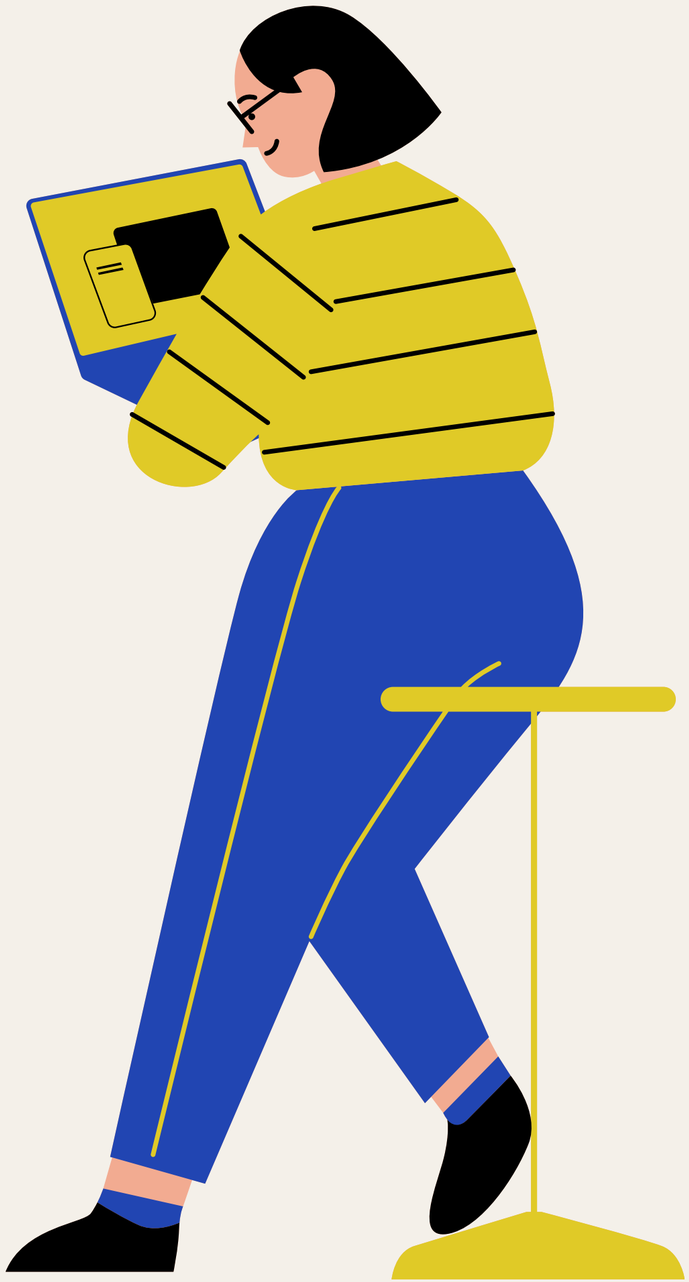
02 - Data Description

03 - Data Cleaning

04 - Data Exploration

05 - Modeling

06 - Conclusions



01 - Introduction

For this analysis, we collected data on U.S. births by education level and wage trends over several years. Our approach included four main steps: data cleaning and preparation, exploring trends, applying models and evaluating results for actionable insights. This structure allowed us to uncover patterns and make informed predictions for future trends.

Objective:
Understand how education impacts birth rates and wages over time and inform policies for economic planning.

02 - Data Description

Data Set 1: "US Births by Year, State, and Education Level"

Provided a detailed view of birth rates by year, state and education level. This data enabled us to assess how education levels impacted birth trends across different time periods.

Data Set 2: "Wages by Education in the USA (1973-2022)"

Provided historical wage information across various education levels. This dataset allowed for the examination of the correlation between education level and wages over time, providing a basis for analyzing economic trends.

03 - Data Cleaning

***Objective**: Ensure data consistency and reliability for analysis*

Wage Data:

- Kept only years that aligned with the birth data (2016-2021).
- Removed all race data as it was outside of the project scope.
- Decided to narrow the scope to only include data on women.

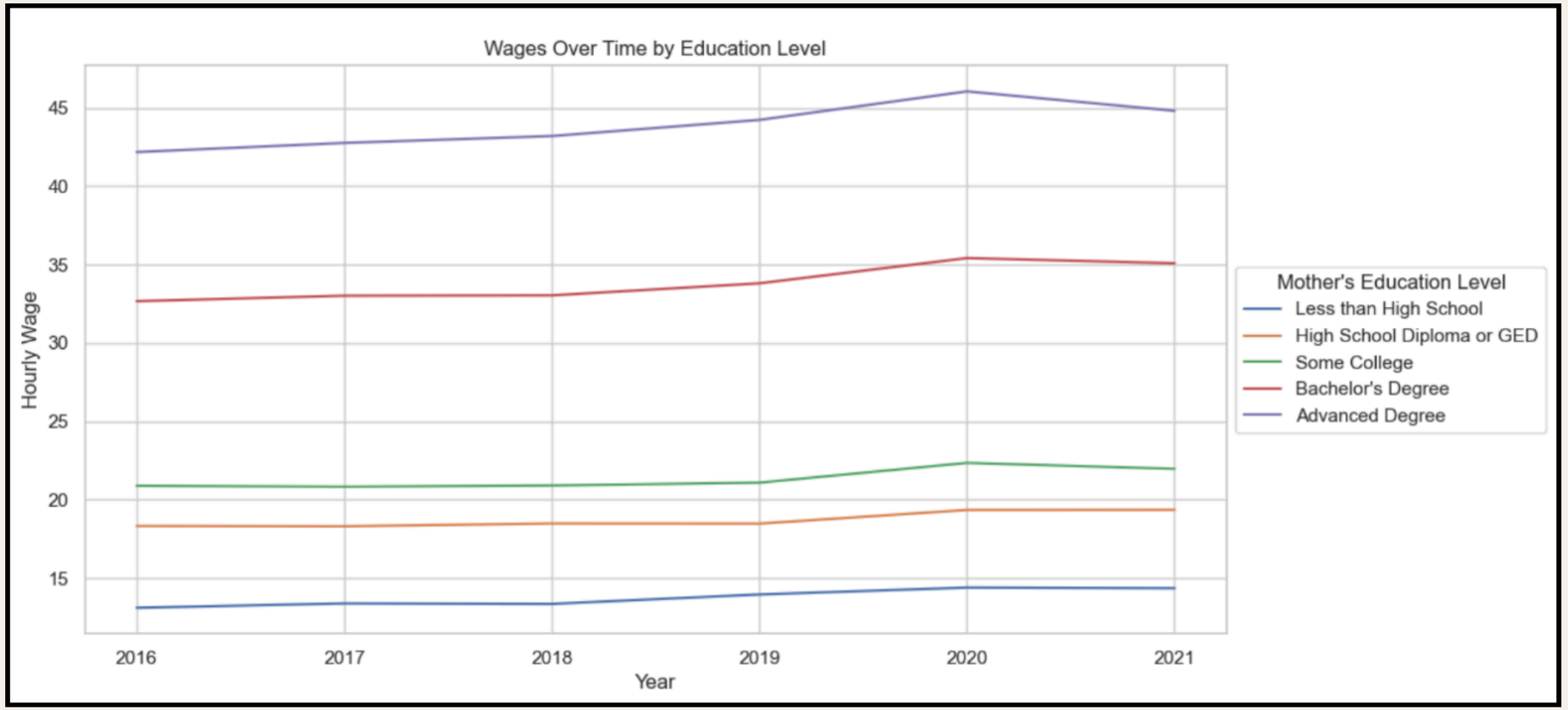
Birth Data:

- Reduced 'Education Level of Mother' column to fit the existing categories in the wage dataset by mapping.

Merging the Data Frames:

- Merged on "*Year*" and "*Mother's Education Level*" by an inner merge.
- Ensured "*Year*" and "*Mother's Education Level*" would appear only once.

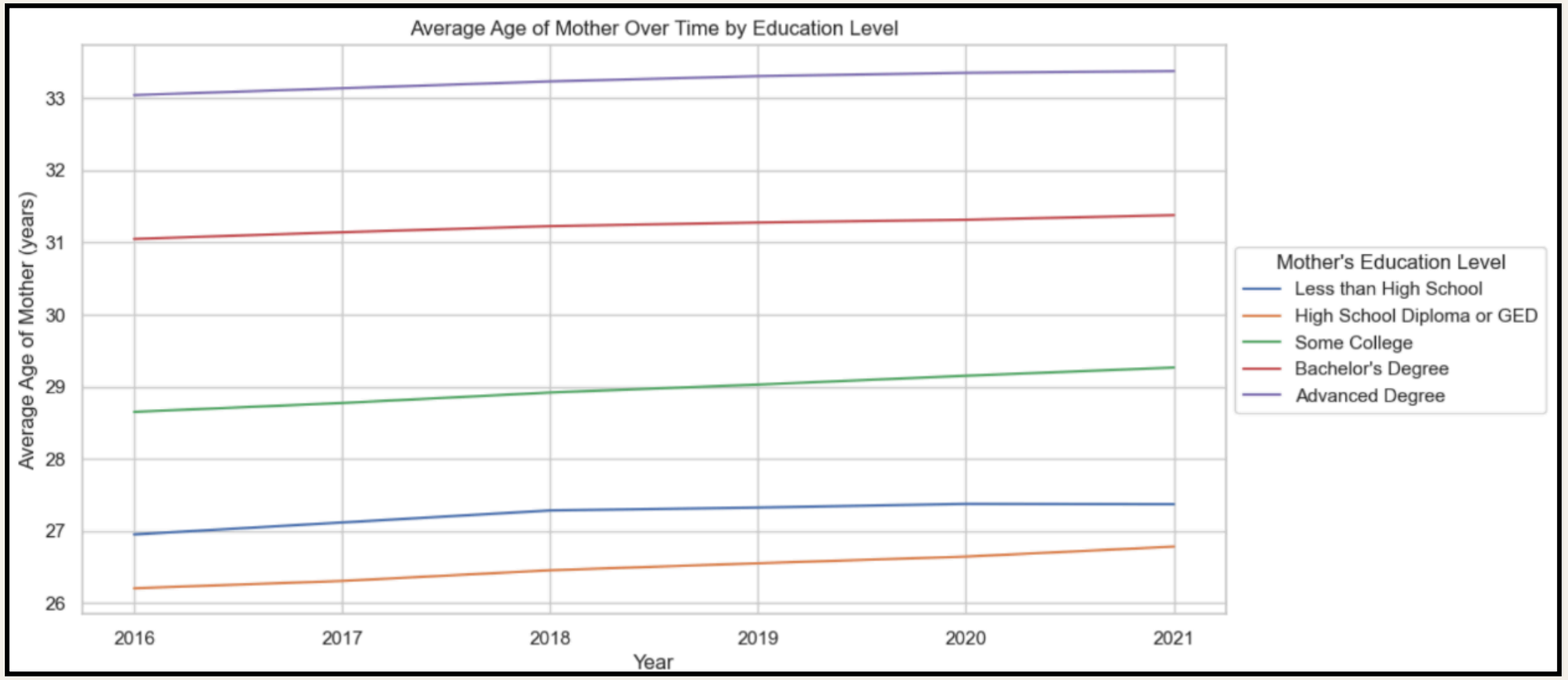
0 4- Data Exploration



Wages over time for women followed the expected trend:

Women with higher levels of education earned higher wages

0 4- Data Exploration



Age of mother at birth followed the expected trend:

Women with higher levels of education wait to have children

05 - Modeling - Linear Regression

Target Variable:

Wage

Performance metrics:

- Mean Squared Error (MSE): 0.338
  - This low MSE shows the model's predictions are highly accurate, with small average squared differences between predicted and actual values.
- R-Squared: 0.9969
  - This value shows that the model explains 99.69% of the variance in the target variable, demonstrating the effectiveness in capturing the relationship between education levels and wages.

Target Variable:

Birth Rate

Performance metrics:

- Mean Squared Error (MSE): 8.46e8
  - The large MSE indicates a substantial difference between predicted and actual birth rates, likely due to the scale and variability of the target values.
- R-Squared: 0.9887
  - Despite the high MSE the R-squared value shows that 98.87% of the variance in birth rates is explained by the model, suggesting it effectively identifies overarching trends.

05 - Modeling - Random Forest

Target Variable:

Wage

Performance metrics:

- Mean Squared Error (MSE): 0.3447
  - While low the MSE indicates the model's predictions are slightly less accurate than those from the Linear Regression model.
- R-Squared: 0.9967
  - This value shows that the model explains 99.67% of the variance in the target variable, demonstrating the effectiveness in capturing the relationship between education levels and wages.

Target Variable:

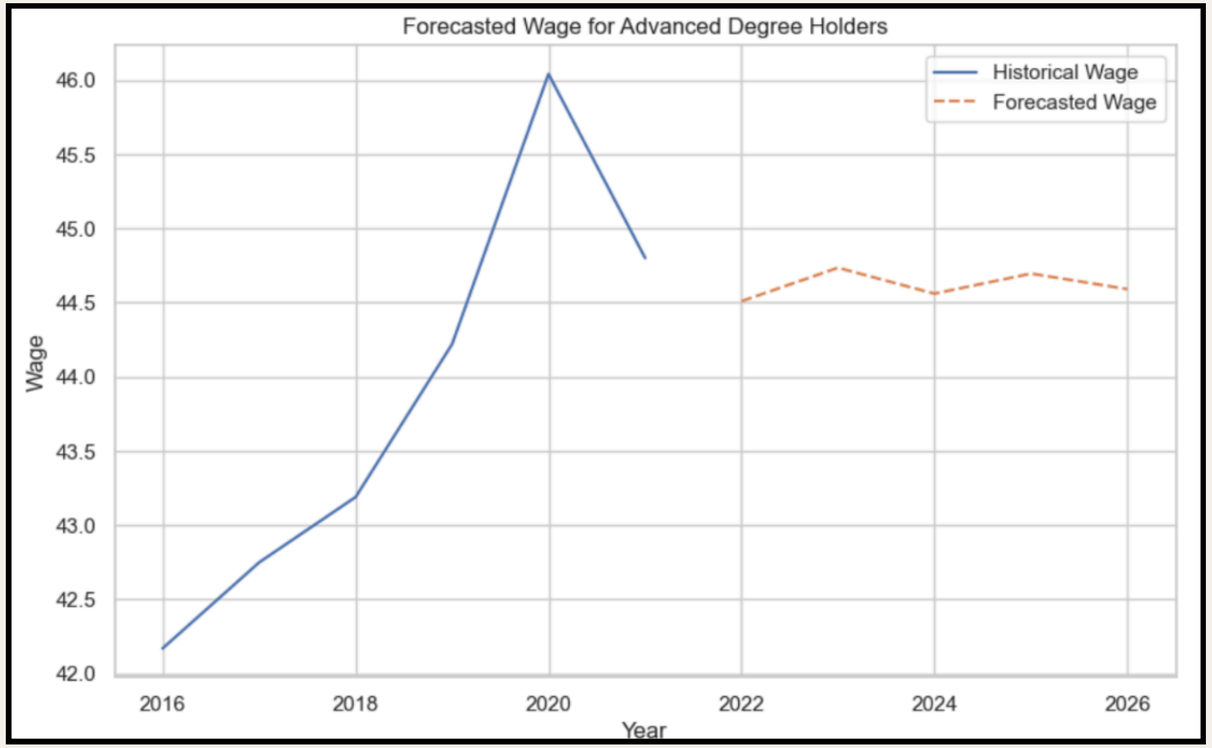
Birth Rate

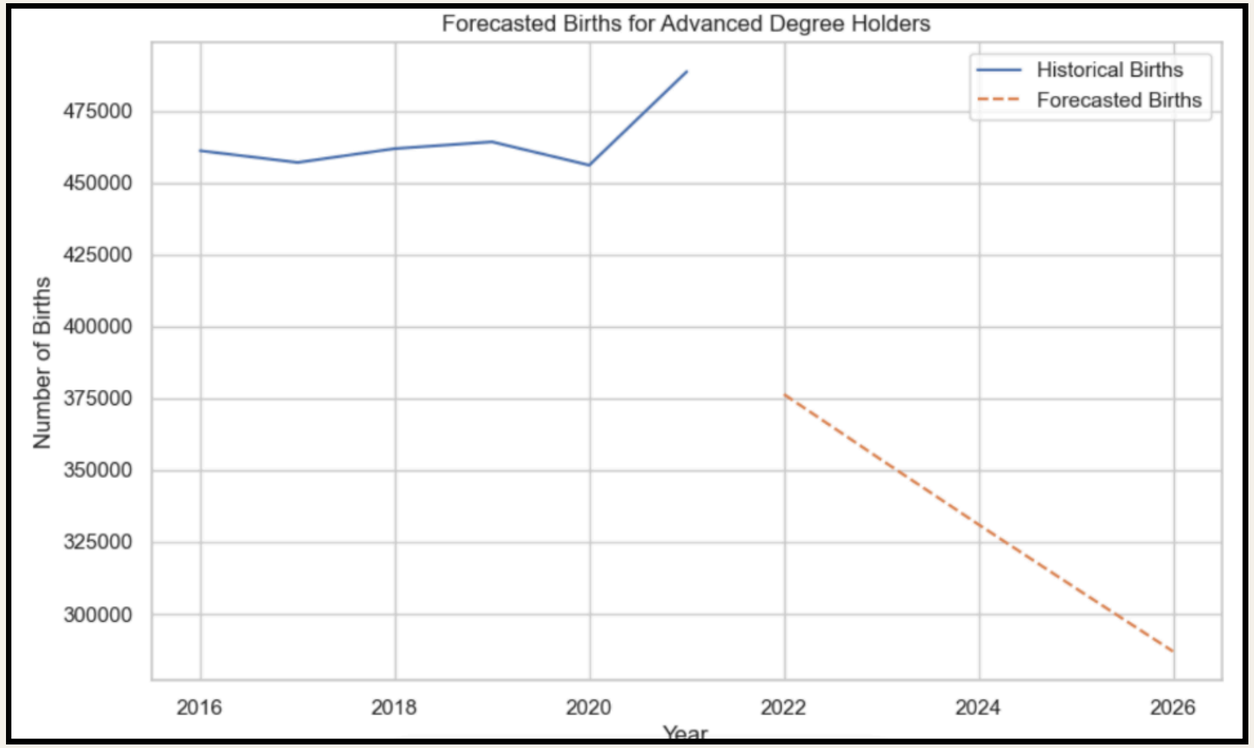
Performance metrics:

- Mean Squared Error (MSE): 2.18e9
  - The high MSE indicates a significant average prediction error, suggesting that the model struggled with this dataset.
- R-Squared: 0.9710
  - Despite the high MSE the R-squared value shows that 97.10% of the variance in birth rates is explained by the model, which is lower than the Linear Regression model's performance.

05 - Modeling - Time Series Analysis

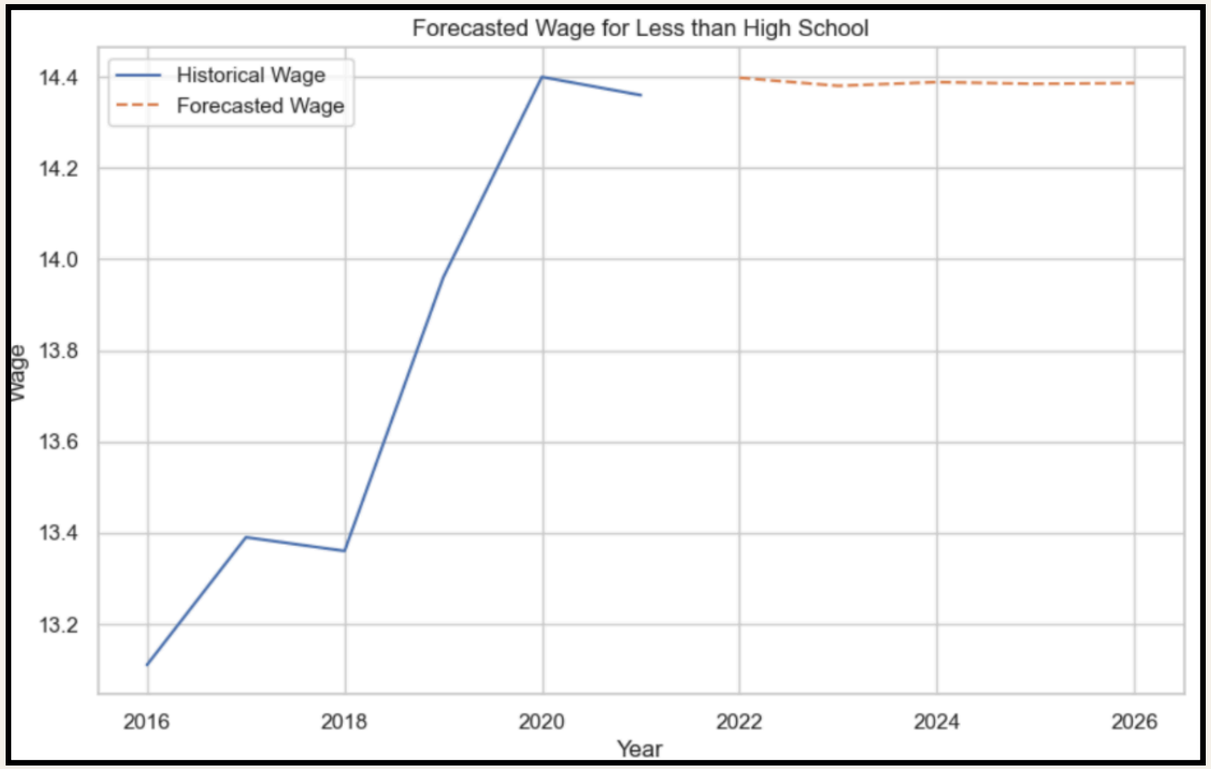
Education Level: Advanced Degree Holder

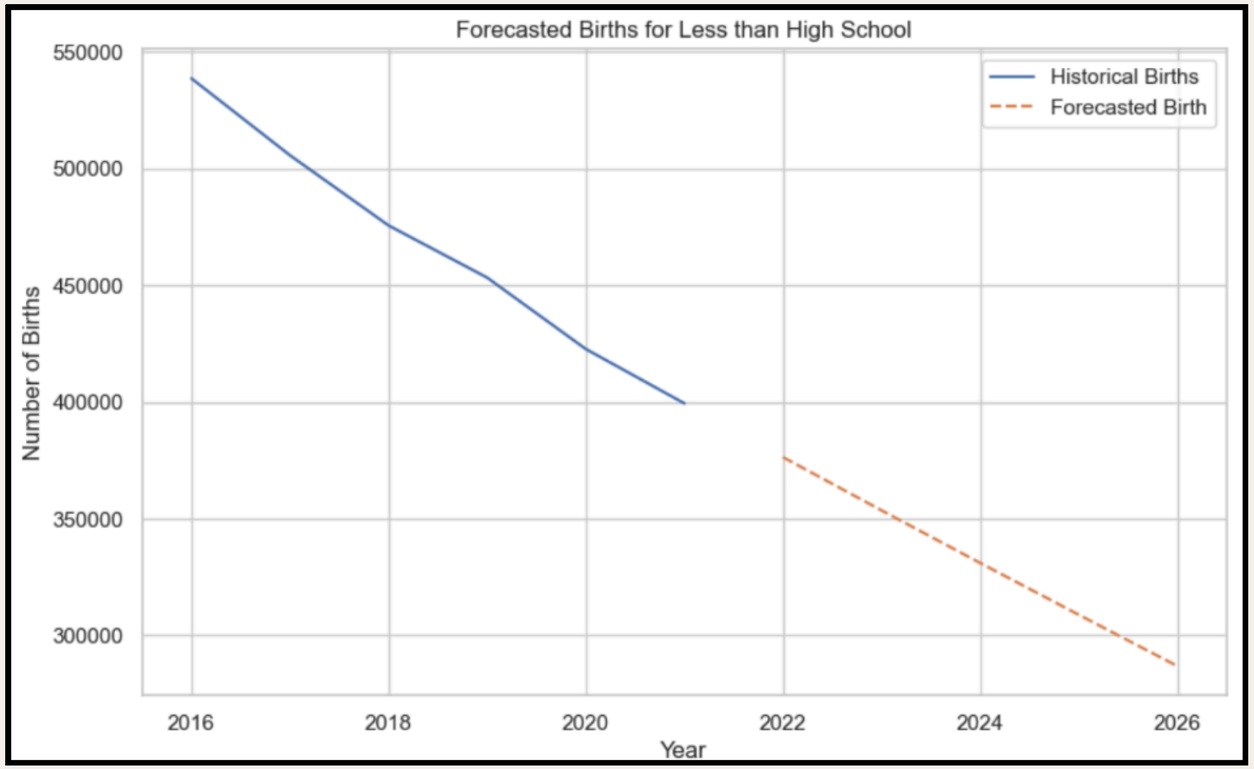




05 - Modeling - Time Series Analysis

Education Level: Less Than High School





04 - Conclusions: Findings

- Wages and Education
  - Higher education levels consistently correlated with higher wages.
  - The Linear Regression model demonstrated the strongest performance in predicting wages with an MSE of 0.3338 and an R-Squared of 0.9989
- Birth Rates and Education
  - Birth rates were generally lower for individuals with higher education levels.
  - Both models struggled to accurately predict birth rates as evidenced by high MSE values with Linear Regression slightly outperforming Random Forest
- Model Comparisons
  - Linear Regression showed greater accuracy for both target variables, indicating that simpler models can sometimes outperform more complex methods when data trends are relatively linear.
  - Time series predictions for both wages and birth rates were affected by 2020 as an outlier due to the COVID-19 pandemic, introducing variability that the model struggled to accomedate

Implications:

These factors introduced challenges in isolating true relationships within the data, highlighting the need for cautious interpretation and consideration of external influences

04 - Conclusions: Limitations and Challenges

- Limited Time Scope
  - The data only covered the years 2016 through 2021, which restricted the ability to observe long-term trends.
- COVID-19 Pandemic Impact
  - *2020 as an Outlier:*
    - The global pandemic led to lockdowns which significantly altered economic and social behaviors, making 2020 a unique and less generalizable year.
  - *Shifts in Education and Employment:*
    - COVID-19 prompted many individuals to return to school, potentially contributing to a temporary decrease in wages for advanced degrees.
  - L*ower Birth Rates for Less-Educated Groups:*
    - Pandemic related economic issues may have influenced family planning decisions, particularly for those with lower educational attainment.

Implications:

Our findings reinforce the importance of education in economic and social outcomes, while also illustrating the challenges of modeling nuanced relationships like birth rates.

04 - Conclusions: Ethical Considerations

- Bias in Data
  - The dataset may have inherent bias due to the underrepresentation and overrepresentation of some groups
    - Careful consideration of these biases should be made when applying results to real-world decisions
- Data Privacy
  - The data used potential sensitive demographic information (e.g. education level and income). It is essential to ensure this data is used responsibly and in compliance with privacy regulations
- Impact of Predictive Models on Policy
  - The models developed have the potential to influence policy decisions related to education and economic development. It is essential to avoid making large decisions without considering broader social, economic and political contexts.

04 - Conclusions: Recommendations and Next Steps

- Incorporate Additional Data Source
  - Expand the dataset to include more years to provide a more comprehensive view of trends.
  - Integrate demographic data, such as population size and employment rates to enhance the models and account for shifts in group sizes.
- Handle Outliers More Effectively
  - Develop strategies for mitigating the impact of outliers, particularly the 2020 data to improve model accuracy and predictions
  - Consider different methods for time series analysis that can better handle anomalies like pandemics or economic issues.
- Investigate External Factors:
  - Do further research into the impact of global events on economic and social outcomes, especially wages and birth rate.
  - Look into other socio-economic variables that could influence trends such as healthcare access and job availability.

Next Steps:

- Validate findings with updated data incorporating external factors that could provide a clearer picture of trends in wages and birth rates
- Further refine models and evaluation metrics for more precise decision making

Thank You For Your Time