**Project Goal:**

In order to study the relationship between gender inequality and crimes against women in the United States for the last ten years, we aim to use traditional statistical analysis to explore how economic gender gap, education gender gap, health gender gap, political empowerment, and state’s abortion legality impact crimes against women. Furthermore, we plan to use machine learning to predict crimes against women using the above five indicators.

**Research Questions:**

1. How does gender inequality, as measured by five key indicators (economic gender gap, education gender gap, health gender gap, political power, and state-level abortion laws), relate to the prevalence of crimes against women in the United States?
2. To what extent can the five key indicators be used as predictors of the prevalence of crimes against women in the United States?

**Hypothesis:**

1. We hypothesize that gender gaps in education and health both correlate to rape incident rates in across 50 states.
2. From the political lens, we hypothesize that female being the Governor/Lieutenant Governor negatively correlates to rape rate. We also hypothesize that the proportion of women in state legislature is negatively correlated to rape incident rates, because we think more female politicians in the state legislature leads to more female involvement and higher likelihood of female-related discourse.
3. We hypothesize that economic gender gap is correlated to rape incident rates. We believe a higher economic gap reflects a more stratified society, which might lead to higher gender inequality and higher crime rate against women.

**Literature Review:**

[1] Yodanis, C. L. (2004). Gender Inequality, Violence Against Women, and Fear: A Cross-National Test of the Feminist Theory of Violence Against Women. Journal of Interpersonal Violence, 19(6), 655–675. <https://doi.org/10.1177/0886260504263868>

[2] Henry, P.J., Steiger, R.L. & Bellovary, A. The Contribution of Gender Equality to the Coexistence of Progressive Abortion and Sexual Orientation Laws. Sex Roles 86, 263–281 (2022). <https://doi.org/10.1007/s11199-021-01263-0>

**Statistical Analysis:**

* + - 1. (Figure 1) Education has a statistically significantly negative relationships with rape rate; year and life expectancy female over male rate have statistically significantly positive relationships with rape rate.
      2. (Figure 2) Controlling for governing by gender, the significance of labor force participation and abortion legality appear when only considering states governed by female, which implies that rape might not be well explained by gender inequality in the states governed by men.
      3. (Figure 3) Controlling for abortion legality alone, female proportion in the state legislature, life expectancy ratio, and gender of state governors have significantly negative relationships with rape rate in states with illegal abortion, where we assume to be more conservative and less equal in general.
      4. (Figure 4) The effect of year and education is consistent; only in states with female governor and lieutenant governor and legal abortion, female proportion in state legislature, bachelor-degree wage gap, and life expectancy ratio have significantly negative relationship with rape rate.
      5. (Figure 5) States with female governor/lieutenant governor and legal abortion have the highest index, indicating the highest level of gender equality. Male-legal and male-illegal are the second and third. Surprisingly, female-illegal has the lowest level of gender inequality. It may be the case that if a woman can be elected as governor or lieutenant governor in such states, she is likely to be more conservative than average and may not bring attention to gender-related issues to canvass for votes. One possible explanation of the interesting finding from the comparison of the four models is that gender inequality sub-indicators (e.g. state\_legislature, bachelor\_wage\_gap) have more impact on rape when states have a relatively low level of gender equality. However, when the states have a high level of gender equality, rape is less relevant to gender inequality, and future investigation on what factors affect the rape rate may be needed.

Chart

Description automatically generated

Figure 1

Chart, scatter chart

Description automatically generatedChart

Description automatically generated with medium confidence

Figure 2 Figure 3

Scatter chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure 4 Figure 5

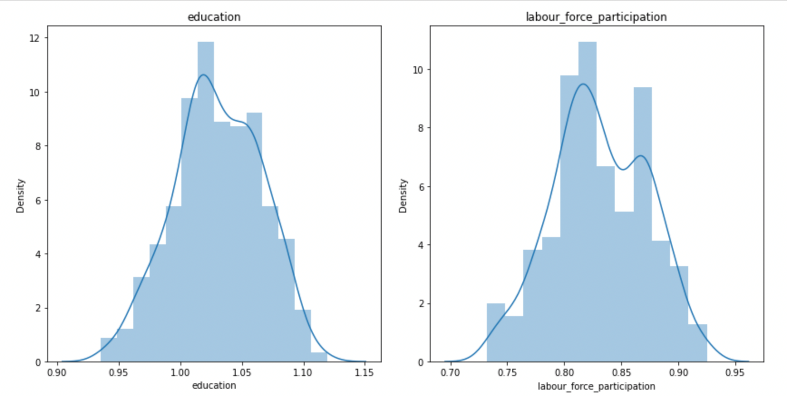
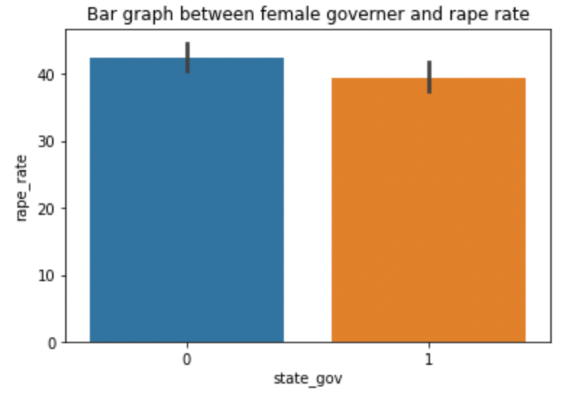
**ML Modeling:**

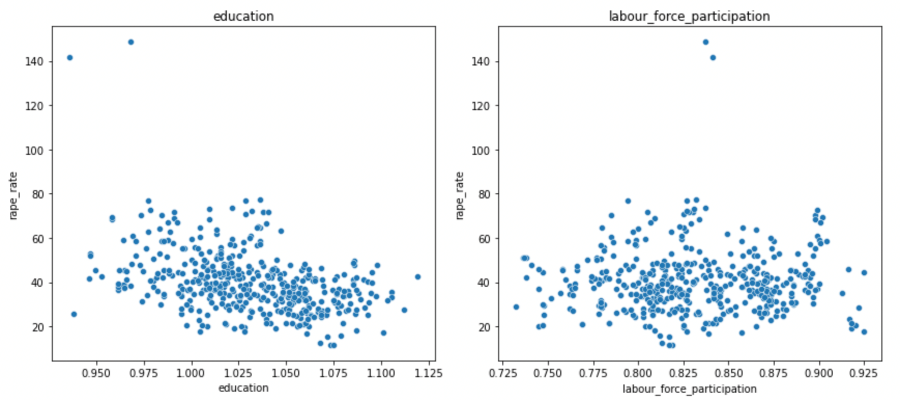
1. Dimension Reduction -- PCA
   1. Originally, we have 15 features as gender inequality indicators. To prevent or limit overfitting, we used PCA to reduce dimensionality on linear regression.
   2. Then we draw the PCA curve and checked cumulative explained variance to determine the principal component number. With 6 principal components, we can explain over 90% of the variance in the dataset, so we choose 6 components.
2. Models
   1. Linear regression with PCA
   2. Linear regression without PCA
   3. Random forest regressor
   4. Lasso regression
   5. Ridge regression
   6. Support vector regression
3. Using the above models, we predicted the rape rates, compared them with our ground truth rape rate, and evaluated each model’s performance with R squared and adjusted R squared.
   1. Predictions on Human behaviors tend to have low R squared.
   2. Support vector regression and random forest models have the best performance on non-linear multivariate regression task.

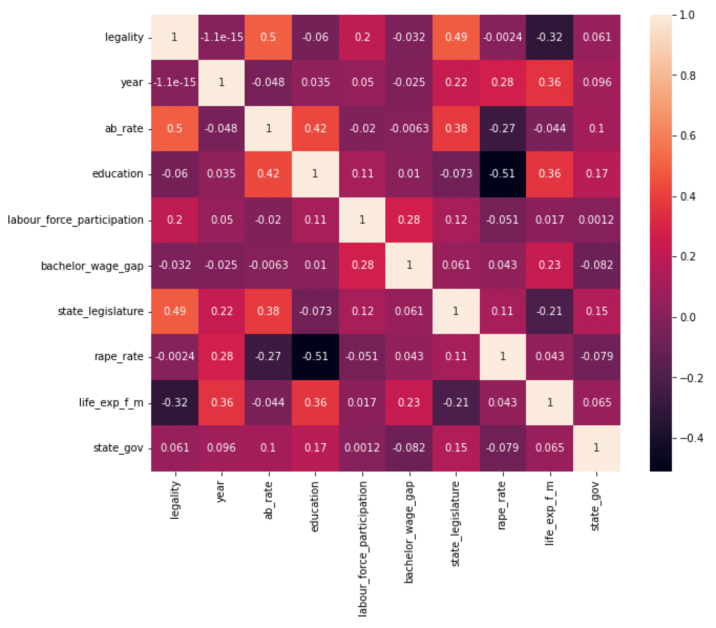
**Exploratory Data Analysis and Visualization:**

1. Using bar graph to compare binary features, including abortion legality and female state governor / lieutenant governor.
2. Using bar graph to represent each numerical feature’s distribution.
3. Using scatter plot to demonstrate the relationship between features (economic gender gap, education gender gap, health gender gap, political empowerment) with abortion rate and rape rate.
4. Using a heatmap to check the correlation and collinearity between each features pair to avoid redundant features.
5. We use line plot to represent the trends of state\_legislature, bachlor\_wage\_gap, life\_expectancy\_f\_m, abortion\_rate, and rape\_rate to gain a longitudinal and comprehensive overview.

**Visualization Examples:**





 Chart

Description automatically generated

**Findings：**

1. Overall, higher gender equality predicts higher rape rate, which confirms our hypothesis.
2. Specifically, large education, economic gap, lower level of political empowerment, or more restrictions on abortion is associated with higher rape rate, which also confirms our hypothesis.
3. Specifically, higher proportion of women receiving degrees of or above high school, working, or having political power is associated with lower rape rate.
4. We are not sure about why life expectancy of women positively predicts rape rate. We hypothesize that this positive correlation may due to a high mortality of young men caused by lower public safety. Future work is needed to substantiate this hypothesis.
5. Machine Learning Model Performance：We used linear regression with PCA, linear regression without PCA, random forest regression, lasso regression, ridge regression and SVR to perform our prediction task, and we found that the random forest model and SVR model performed the best with R-squared scores around 0.60, as they are best for non-linear multivariate regression task. For the other models, although the R-squared is approximately 0.38, it is common that the R-squared is low for human behavior related research.

**Points to note：**

1. Missing data - check for unusual pattern and caution mismatching data and label.
2. Web scraping - cautious about HTML layout when scraping multiple pages at a time.
3. R is much powerful in statistical analysis task and has better visualization packages than Python.

**Responsibility:**

Yingzi Jin: Collected and preprocessed political empowerment data and abortion rate data, conducted traditional statistical analysis, and developed interpretations and possible explanations.

April Wang: Collected and preprocessed education gender gap data and state crime data, conducted exploratory data analysis, and produced the presentation slides.

Ruoyi Wu: Collected and preprocessed health gender gap, web scraped each state’s abortion legality, conducted data visualization and ML modeling, and composed README and final report.

Guangjie Xu: Collected and preprocessed economic gender gap data, conducted data exploration and visualization, and conducted ML modeling.