**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 economic, education, and health all 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:**

Yingzi

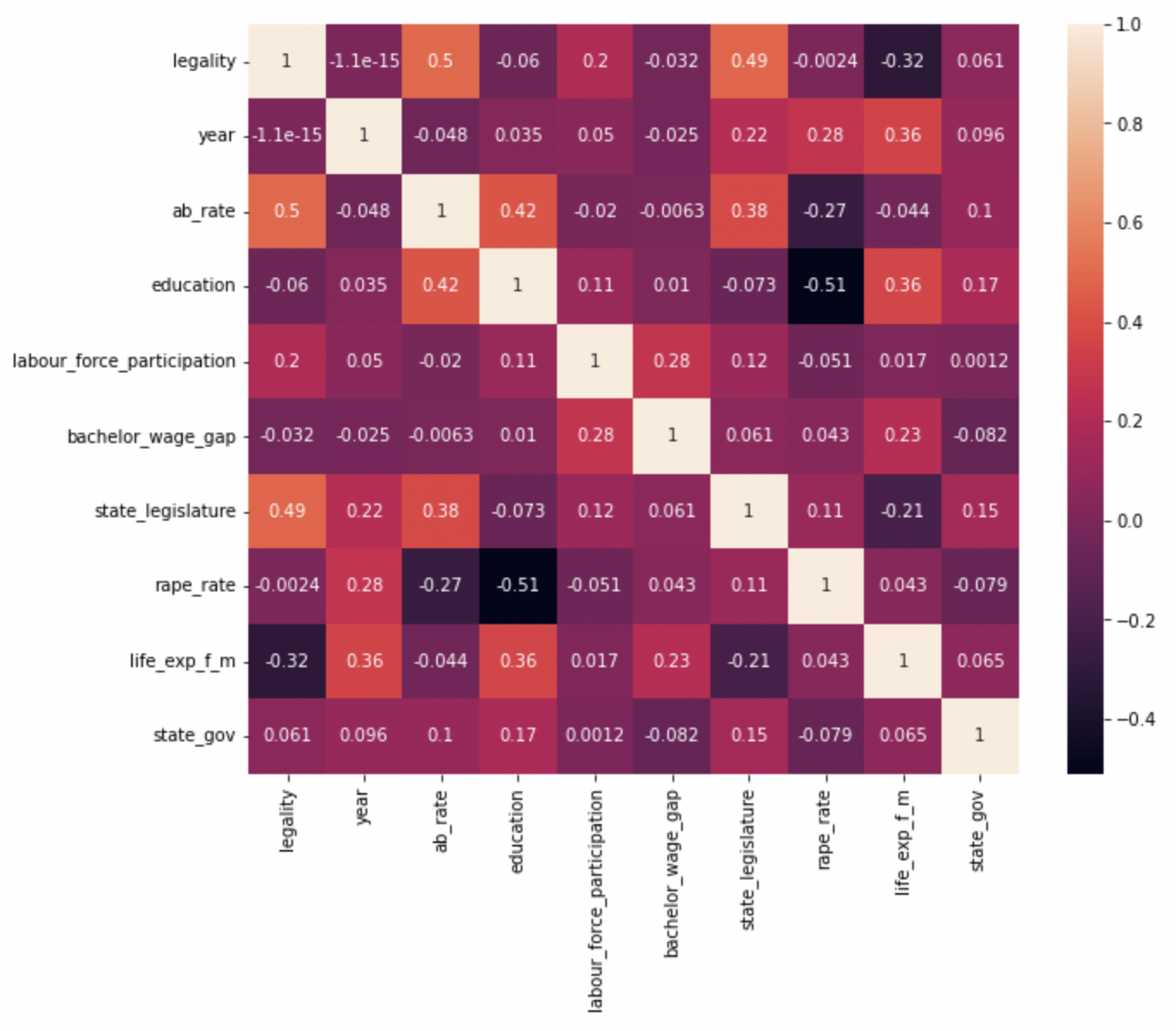
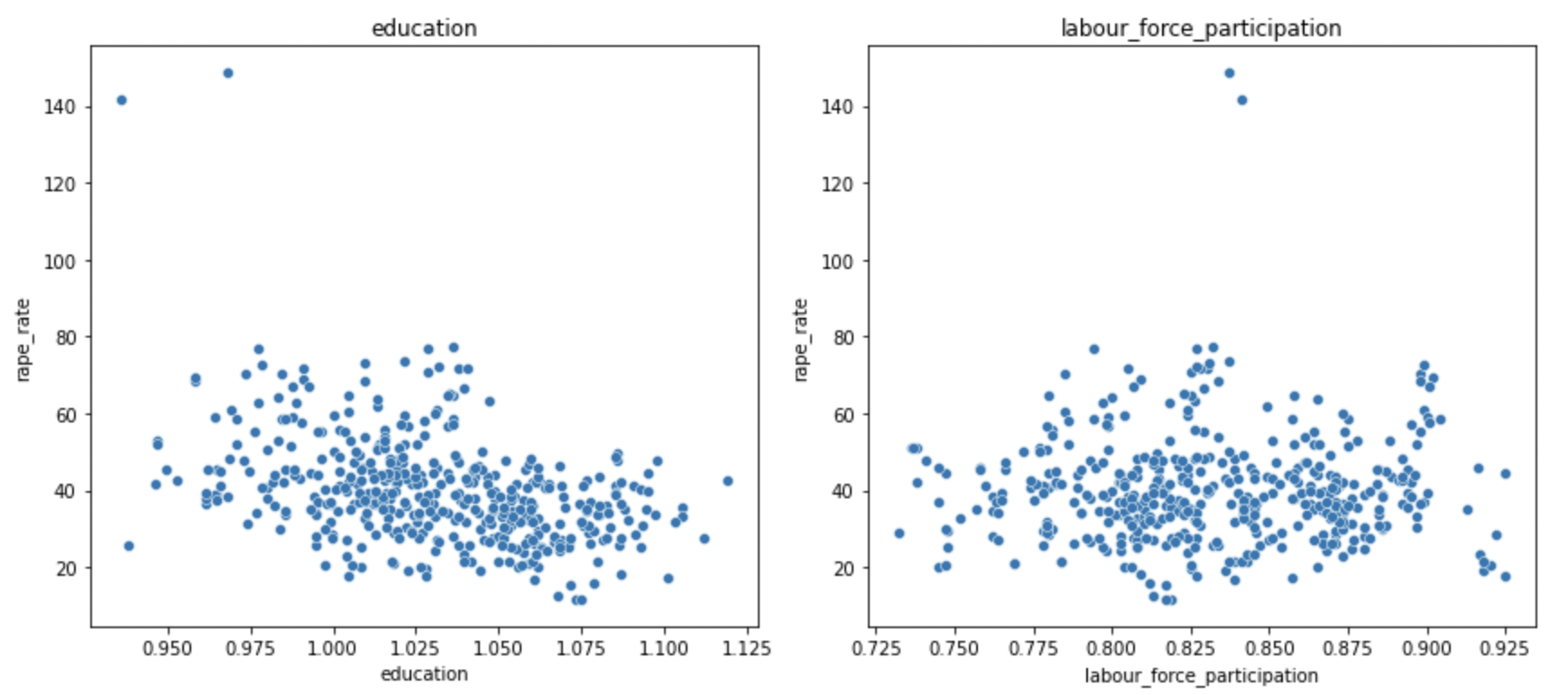
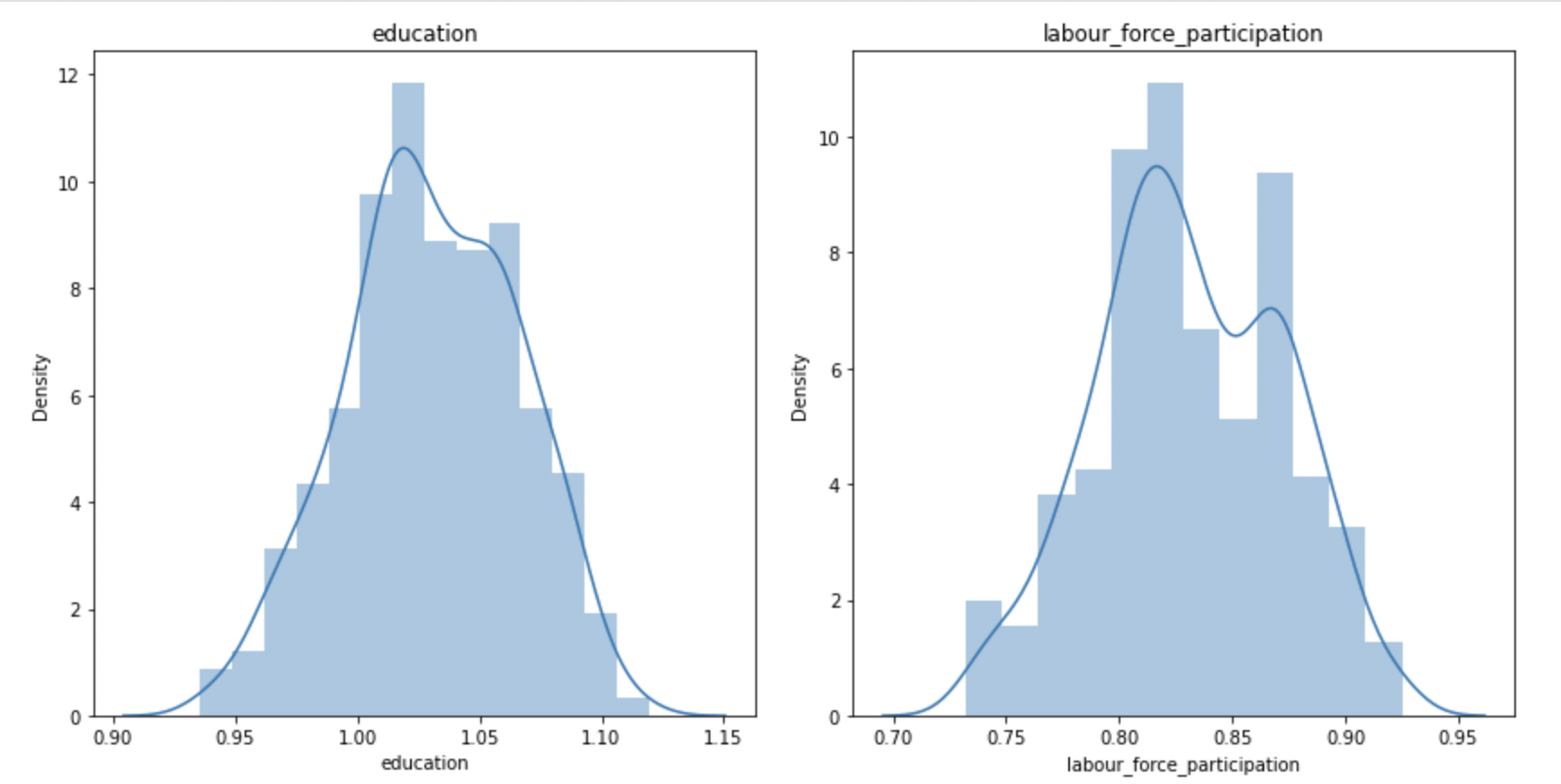
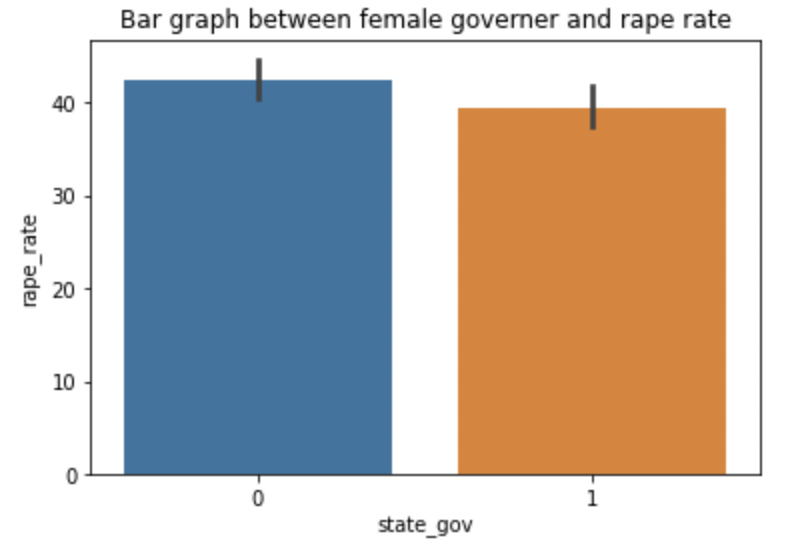
**Modeling:**

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. Model
   1. Linear regression with PCA
   2. Linear regression without PCA
   3. Random forest regressor
   4. Lasso regression
   5. Ridge regression
   6. SVR regression
3. Using the above models, we predicted the rape rate, compared them with our ground truth rape rate, and evaluated each model’s performance with R squared and adjusted R squared

**Exploratory Data Analysis and Visualization:**

1. Using bar graph to compare binary features, including abortion legality and female state 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.
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:**



**Findings：**

1. Overall, higher gender inequality predicts lower 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. Support vector regression and random forest models have the best performance on non-linear multivariate regression task

**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, and conducted data visualization and ML modeling

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