Pre-Analysis Plan (Vanessa Pasquarelli, Edie Thors, Gwen Thompson)

Research Question:

Our goal is to explore the relationships between demographic, economic, social, health, and political variables and their impact on subjective well-being, including happiness and stress levels. Specifically, we aim to examine how factors such as marital status, income, political views, job satisfaction, and family dynamics influence overall happiness and stress. We will also explore how these variables reflect broader societal issues, such as inequality and political polarization.

Observations:

The dataset consists of survey responses from participants, where each observation represents an individual respondent with multiple attributes across several categories. The relevant variables include:

- Demographic Variables: Marital status, gender, education, race, and sexual orientation.
- Economic Variables: Income, socioeconomic status, social class, and employment status.
- Family and Household: Number of children, household population, and number of sexual partners.
- Health-Related Variables: Self-reported health, stress levels, physical activity, and smoking habits.
- Social and Political Views: Political views, religious affiliation, perceptions of gun laws, trust in government, and fear.
- Job and Work-Related Variables: Job satisfaction, job loss, and spouse or coworker work status.
- Well-being and Happiness: Overall happiness, marital happiness, and satisfaction with financial situation.
- Safety and Security: Perceptions of vaccine safety, COVID-19 impacts, and societal safety.
- Other Variables: Health insurance type, contraception use, and arrest history.

Learning Approach:

We are focusing on supervised learning, where we will analyze how independent variables (e.g., education, job satisfaction, stress levels) predict dependent variables related to subjective well-being, specifically happiness and stress. We will use regression analysis for continuous outcomes (e.g., stress level, happiness) and classification for categorical outcomes (e.g., political views, job satisfaction levels).

We will first attempt basic linear regression and logistic regression for predicting well-being and happiness. Additionally, clustering techniques, like k-means clustering, will be used to group respondents based on shared characteristics (e.g., income level, political views, or stress levels), helping us uncover patterns and interactions that might be otherwise missed.

Models and Algorithms:

- Linear Regression: This model will help us understand the relationship between continuous dependent variables (e.g., happiness, stress) and independent predictors (e.g., income, marital status, health).

- Logistic Regression: For categorical outcomes (e.g., political views or job satisfaction), we will use logistic regression to model the likelihood of belonging to a certain category.

- Clustering (K-Means): We will perform k-means clustering to identify subgroups of respondents that share similar characteristics, particularly in terms of political views, health status, and economic conditions. These cluster labels can then be incorporated into the regression models.

- Regularization: We may use Lasso or Ridge regression if we encounter multicollinearity or overfitting issues.

Model Validation and Success Criteria:

- R-squared (R^2): For regression models, R^2 will tell us how much of the variance in happiness and stress can be explained by our predictors.

- RMSE (Root Mean Squared Error): For continuous outcomes like happiness, we will evaluate the RMSE to measure prediction accuracy.

- Accuracy, Sensitivity, Specificity: For classification models (e.g., predicting political views or job satisfaction), we will use these metrics to assess the effectiveness of the model.

- Cross-validation: We will use k-fold cross-validation to ensure our models generalize well to unseen data.

Anticipated Weaknesses and Mitigation Strategies:

1. Missing Data: The dataset contains missing values, particularly in variables such as income and health status. We will use imputation methods where possible (e.g., mean imputation for continuous variables or mode imputation for categorical ones), or we may drop rows with excessive missing data if necessary.

2. Multicollinearity: Many variables in the dataset are likely to be correlated (e.g., income and social class, health status and stress levels). We will assess multicollinearity using the Variance Inflation Factor (VIF) and, if needed, use PCA (Principal Component Analysis) or regularization techniques (e.g., Lasso regression) to address it.

3. Data Standardization: Variables such as income, stress, and political views may be recorded in different formats. We will standardize continuous variables for consistency and to avoid skewing the results.

4. Non-linear Relationships: Some relationships between variables (e.g., income and happiness, job satisfaction and stress) might be non-linear. We will explore non-linear models or apply transformations (e.g., logarithmic) to better capture these relationships.

Feature Engineering:

- Categorical Encoding: We will one-hot encode categorical variables such as political views, marital status, and religious affiliation.

- Interaction Terms: We will explore potential interactions between key variables (e.g., income and education, stress and health) to identify more complex relationships.

- Temporal Variables: If applicable, we may transform date variables (e.g., when the survey was taken) into time-based features.

- Cluster Variables: After performing k-means clustering, we will create new features representing the cluster each respondent belongs to and use these in the regression models.

Results Presentation:

- Regression Coefficients Table: For linear regression, we will present the coefficients, p-values, and confidence intervals for each predictor to understand their impact on well-being.

- Cluster Analysis: We will visualize the results of k-means clustering using a scatter plot or heat map, showing how respondents cluster based on key variables like income and stress.

- Model Performance Metrics: We will compare the R^2, RMSE, and accuracy scores for each model and present them in a summary table.

- Residual Plots: For regression models, we will generate residual plots to check for any patterns that may suggest model inadequacies.

Conclusion: This pre-analysis plan outlines our approach to studying the relationships between various social, economic, and health factors and subjective well-being, focusing on factors such as income, stress, and political views. By using regression and clustering techniques, we aim to uncover meaningful patterns and interactions that can inform social science theories and guide policy interventions to improve well-being and reduce societal inequality.

Ack: We used ChatGPT to help us with parts of this assignment.