Statistic Analysis Plan | Aarya Desai | Jeremy Tan | Cassie Kang | Osama Ahmed

Data Overview:

This dataset used is a combination of two datasets, namely, the World Happiness Dataset, 2022 retrieved from the official World Happiness Report website which conducts a yearly survey and analysis of world happiness metrics across the globe (i.e., each country), including different variables such as, life ladder, log GDP per capita, social support, life expectancy at birth, freedom, perception of corruption, generosity, positive affect and negative affect. The second dataset used was a World Mental Health Disorders dataset retrieved from the University of Oxford website containing datasets related to mental health metrics across the globe (i.e., each country) spanning over multiple years. The latter dataset included metrics for mental health disorders recorded and analyzed on a Likert Scale (1 = Least Like to have given mental health disorder to 5 = Most likely to have given mental health disorder) and included this metric for the following disorders: Alcohol Use Disorders, Drug Use Disorders, Schizophrenia, Bipolar Disorders, Anxiety Disorders, Eating Disorders, and Depression. The resultant dataset had 1462 observations and 16 unique variables. Each row represents a specific country's happiness and mental health data for a given year.

RQ1: To what extent do happiness metrics (predictor variables: log GDP per capita, social support, life expectancy, freedom, perception of corrpution, generosity) predict mental health disorders, specifically, depress ion (outcome variable: continuous) in countries?

RQ2: To what extent do mental health disorders and happiness metrics (predictor variables: schizophrenia, bipolar disorder, eating disorders, anxiety disorders, depression, log GDP per capita, social support, freedom, perception of corrpution, generosity) predict life ladder (outcome variable: categorical)?

Modeling:

Research Question 1:

- Model type: Linear Regression
- Question type: Prediction
- Variable selection will be done via Backwards Stepwise Selection where we start with variables we believe have an impact on influencing depression, and seeing how the p-values changes
- Model assessment will be done using the classic linear regression techniques, such as, RMSE, MSE, CV, and etc.
- List of variables:
  - Outcome: Depression (continuous)
  - Predictors: log GDP per capita, social support, life expectancy, freedom, perception of corruption, generosity (all are continuous)
  - Confounds: Country, Year
  - Interaction term: Year*Log GDP per capita

Research Question 2:

- Model type: Logistic Regression
- Question type: Prediction
- Variable selection will be done in an a priori manner, where all the variables showing some relationship with Lifestyle Category in the EDA will be used in the model.
- Model assessment will be done using a confusion matrix, ROC curve, VIF, and AIC values.
- List of variables:
  - Outcome: Lifestyle Category (categorical)
  - Predictors: schizophrenia, bipolar disorder, eating disorders, anxiety disorders, log GDP per capita, social support, freedom
  - Interaction term: log GDP per capita*freedom

Potential Challenges:

1. Assumptions about Data Homogeneity: It is assumed that the data is uniform across countries and years, disregarding potential influences from these variables. This is crucial because factors like GDP and cultural nuances can significantly impact happiness metrics.
   1. Solution: We can factor in country and year as confounding variables in the model to account for this influence.
2. Addressing NAs: There's a need to understand why there are missing values (NAs). This could be due to merging processes or a specific pattern (e.g., related to certain years, countries, or occurring randomly). It's important to address this issue, as it could affect the accuracy of the results or model.
   1. Solution: After merging the datasets, there are some rows with NA values. We will need to look carefully whether we should replace them with values or drop them.
3. Binary Variable Creation: Creating a binary variable based on the median may not capture nuanced differences between groups. This might limit the depth of insights derived from the analysis.
   1. Solution: This transformation seems viable, and the EDA showed promising results for the same, therefore, this transformation will be continued and carried forth.
4. Limited Insights from EDA for RQ1: Initial exploratory data analysis (EDA) for Research Question 1 suggests weak relationships. This could potentially lead to a model with limited predictive power and insights.
   1. Solution: Add more interaction terms and look at outliers to see if these points are causing weak relationships
5. Multicollinearity Concerns: Given the nature of predictor variables in the World Happiness Dataset, there's a possibility of high correlation among them. This can lead to multicollinearity in regression models, complicating the interpretation of individual predictor effects on the outcome.
   1. Solution: Use VIF to identify variables with high collinearity with each other and either remove or combine them. We can also attempt to use L1 Regularization and essentially take a step back into variable selection to see what variables we want to use in the model.