# STAT 390 Weekly Report 3
## Oct 15-20

Group 2: Cindy Ha, Willie Xie, Erica Zhang

## Table of contents

## 1 Progress/Accomplishments

- Finishing randomness preprocessing for various model types including linear `Preprocessing Code/preprocess_lm`, tree based `Preprocessing Code/preprocess_tree`, and neural network `Preprocessing Code/preprocess_nn`.
- Removed countries in dataset with numerous predictors with large missingness.
- Determined dates to split data into training and testing sets. Training will be data before 2023 and testing will starting 2023.
- Will use clustering with `life_expectancy` to make predictions for some random missing values and then impute predictor variables like vaccination, extreme poverty and icu units with cluster median/mode.

## 2 Challenges

- Deciding how to address missingness for predictors with large missingness for different models (i.e. remove predictors in linear, impute large value for trees, create boolean indicators for neural networks).
- Deciding how to address missingness for new models like ARIMA.

# 3 Next Steps

- Generate validation sets from training set.
- Determining how to deal with missingness in the testing set for different models.
- EDA should be mostly done - can start with training models.