STAT 390 Weekly Report 4

Oct 30- Nov 2

Group 2: Cindy Ha, Willie Xie, Erica Zhang

Table of contents

1 Progress/Accomplishments

- Added 2 new variables: `month` and `day_of_week`
- Made training and testing sets for each model (located in `data/finalized_data` folder
  - Clustered with `life_expectancy`, `female_smokers`, `male_smokers` to impute missingness by cluster median
- Looked at outliers again using Cook's distance > 0.5
- Set up tuning scripts for xgboost and arima
  - Using rolling origin forecast resampling: validation sets for every 3 months train + 1 month test with month increments

2 Challenges

- There was still missingness in the testing sets because some data was completely missing (no median could be computed) so we had to impute by training median
- We are removing China from list of countries we are using due to different Covid reporting policies there since the data may be unreliable (irreducible error)

3 Next Steps

- Looking into lagged features over the weekend to complete feature engineering
- Each member will run the 6 models by Nov 10 to provide starting model performance