Safe Gergis
24 Sept 2025

1. The 80/20 split reflects the importance on data quality and quantity. As they say, garbage in results in garbage out. While modeling, evaution, and deployment are all important, it will all result in nothing if the data the model is built on is garbage. I feel like an 20% split for understanding the data is adequate, as most time consuming activity should be preparing the data at roughly 60%. The last 20% can be split equally between evaluation, modeling, and deployment.
2. I believe this could be a form of bias. Many marginalized groups may be hesitant to answer the census.
3.
    a. Model 1 generalizes better to cases not in the dataset
    b. Simplicity bias
    c. Model 2 overfits with too many features
    d.