1. Code target variable into binary (0 and 1) and make sure there are enough data points in each outcome 🡪 Ansul
2. Deal with missing values

* Drop columns:

V18q1 – number of tablets owned

Rez\_esc – years behind in schooling

Dependency – dependency rate 🡪 Juan

* Change columns:

Edjefe – change all “no” to 0

Edjefa – change all “no” to 0 🡪 Ansul

* Other columns with missing Data:

Meaneduc -- fill in 0 for the missing data 🡪 Aneesha

SQBsquared -- fill in 0 for missing data 🡪 Aneesha

\* The above values are correlated and only very few are missing so we can fill in zero

V2A1 – show the distribution of people who own houses vs. rent and impute 0 for values 🡪 Aneesha

1. Exploratory Analysis w/ visuals 🡪 Everyone

* Correlation matrix 🡪 Katie
* Distribution of outcomes – Katie/Juan
* More visuals 🡪 Katie

1. Feature selection 🡪 Juan
2. Split test/train data 🡪 Aneesha
3. Run Statistical Model on training data (maybe poisson, logistic) 🡪 we can all explore this, Ansul start prepping the code
4. Run statistical model on test data
5. Graph of predictions vs. actual
6. Run machine Learning Models on training data 🡪 we can all explore this as well
7. Cross validation
8. Predict on actual test data
9. Graph of predictions vs. actual