* Data Exploration
  + Finding a Link
    - Census Tract Number
      * Unique code for each location’s health and infrastructure data
      * State FIPS code + County FIPS code + Tract Number = Census Tract Number
      * Able to link datasets together
  + Cleaning
    - Obesity
      * 500 cities by CDC narrowed down to obesity rate
      * Had census tract number built in
    - Life Expectancy
      * Had census tract number built in
    - Infrastructure
      * Food Accessibility
        + USDA dataset
        + Split up by Census Tract Number
        + Narrowed down to food accessibility scores based on distance
      * How people move around a city
        + Walkability, car ownership, transit frequency, etc
        + Only contained truncated Census Tract Number

To fix it:

Merged State and County FIPS codes

Concat FIPS + Census Tract = Census Tract Number

Drop duplicates/NA

* Modeling
  + Data Preprocessing
    - Merged infrastructure and health data on Census Tract Number
  + Feature Engineering
    - Dropped location ids
      * County, state, census tract number
    - X
      * population density, employment density, walkability, car ownership, accessibility of food, frequency of transit usage, and connectivity of pedestrian walkways
    - Y
      * Two models
        + Life expectancy
        + Obesity
  + Modeling
    - Random Forest Regressor
      * Predict continuous values
      * No PCA
      * Feature importance
      * Reduces overfitting/ no normalization needed
  + Model Training
    - Minimal
    - Dropped % Low Income
      * Skewed feature importance heavily
      * Didn’t fit thesis
  + Accuracy
    - No balanced accuracy score
    - MAPE = 100 \* ((predicted – actual)/actual)
    - Accuracy = 100 – mean(MAPE)
    - Life Expectancy Model
      * 96.67% accuracy
    - Obesity Model
      * 88.13%