New York Mental Health

Group 9
Naman Bareja (NB3236)
Arvind Nagabhirava (AN3317)
Pranav Bidve (PB2972)
Nilaa Raghunathan (NR3005)
Ke Wang (KW3173)

Introduction

Research Question

- Can demographic, socioeconomic, and environmental factors accurately predict the prevalence of poor mental health at the neighborhood level?

Project Objectives

- Analyze how community-level factors (e.g., income, education, healthcare access, environmental conditions) influence mental health rates across census tracts.
- Target Variable (Mental Health): **MHLTH**
  - Represents % of adults reporting 14+ days of poor mental health per month.
  - A key indicator of frequent mental health challenges in the community.

Goals

- Identify high-risk areas for poor mental health.

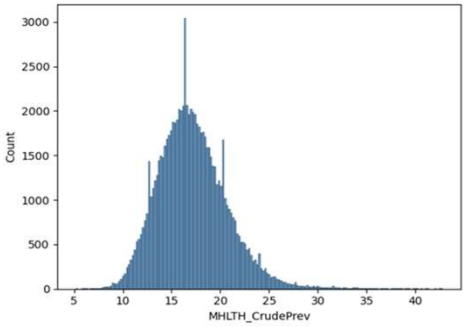
Datasets

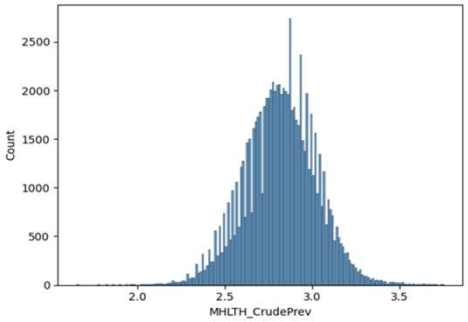
Following geographical datasets were joined based on latitude and longitude:

- Places Data - https://www.cdc.gov/places/about/index.html
  - 40 chronic disease and other measures from US
  - Reports data at county, place and census tracts
- Health facility Information - https://health.data.ny.gov/Health/Health-Facility-General-Information/vn5v-hh5r/about_data
  - Locations of healthcare facilities like hospitals, nursing homes, diagnostic treatment centres, certified home healthcare agencies, licensed adult care facilities etc.
- Smart Location Database - https://www.epa.gov/smartgrowth/smart-location-mapping#SLD
  - Measures location efficiency with variables like housing density, diversity of land use, neighborhood design etc.
- ~~Yelp dataset~~ - not used - NY state not available, explained in further slides.

CDC Dataset EDA/Visualizations

- **Data Wrangling**
  - Data present in two forms - absolute values and 95% Confidence Intervals - retained only the absolute values
  - Removed columns with over 5% missing values
- **Pre-processing**
  - Distribution of target variable shown below - as-is and log-transformed (on the right)
  - Log transformed - appears more symmetric and closer to normal distribution - could improve model stability and reduce impact of extreme values

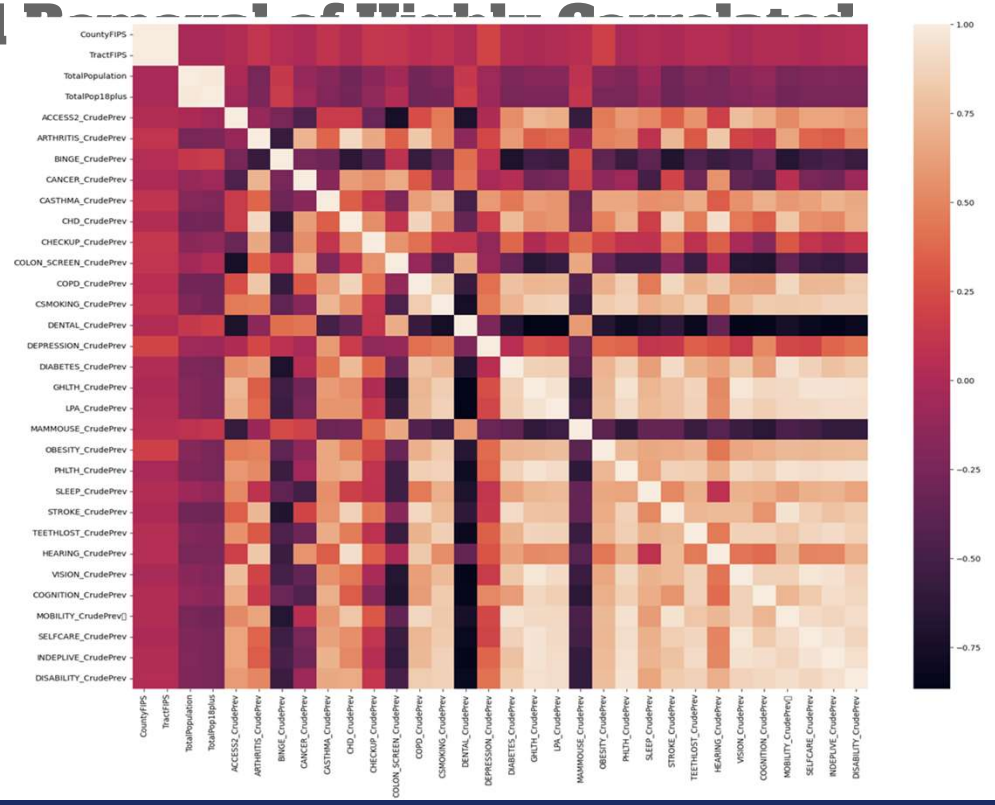




CDC Dataset EDA/Visualizations

Feature Correlation Analysis and Removal of Highly Correlated Features (>80%)

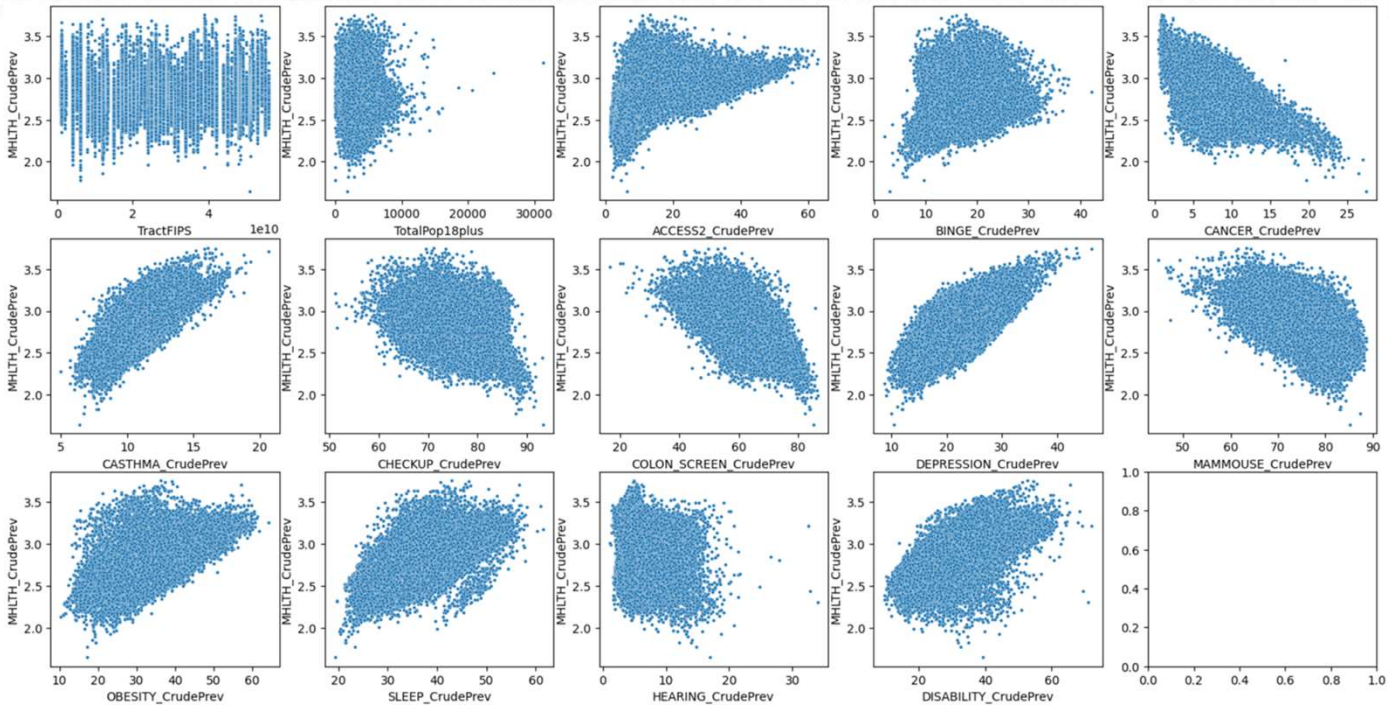
- Dropped highly correlated features - 18 features remain
  - 14 features - numerical
  - Rest - categorical features
- Investigating the correlation of variables with target variable : `MHLTH_CrudePrev`



Positively Correlated	Negatively Correlated	Mixed / Ambiguous Correlation
CASTHMA	CANCER	BINGE
OBESITY	DEPRESSION	(Binge drinking)
	COLON_SCREEN	

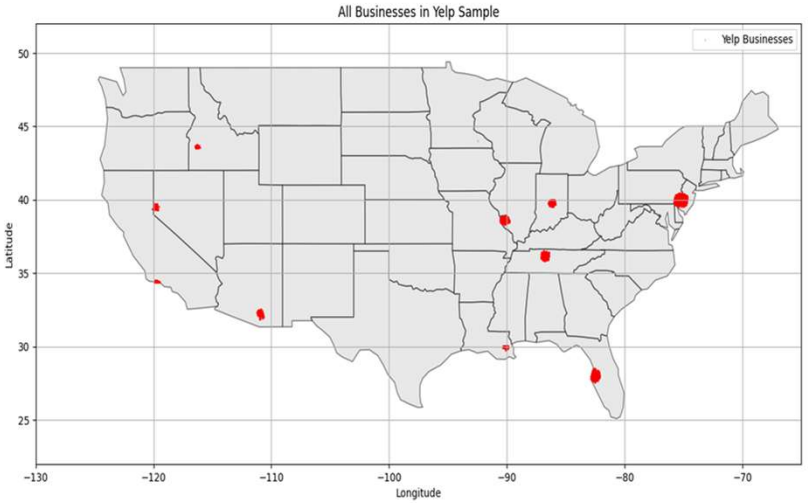
CDC Dataset EDA/Visualizations

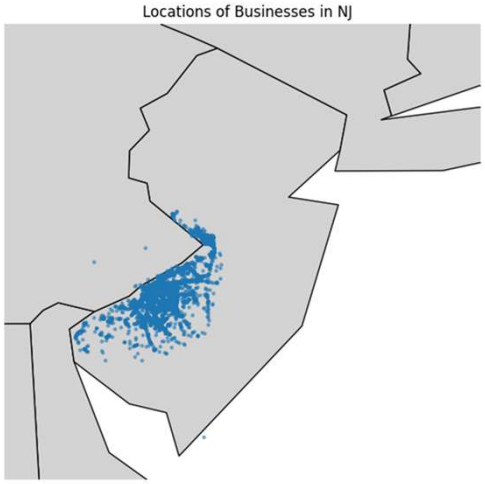
Distribution Analysis of Numerical Features vs. Target Variable



Yelp Dataset Missing Data

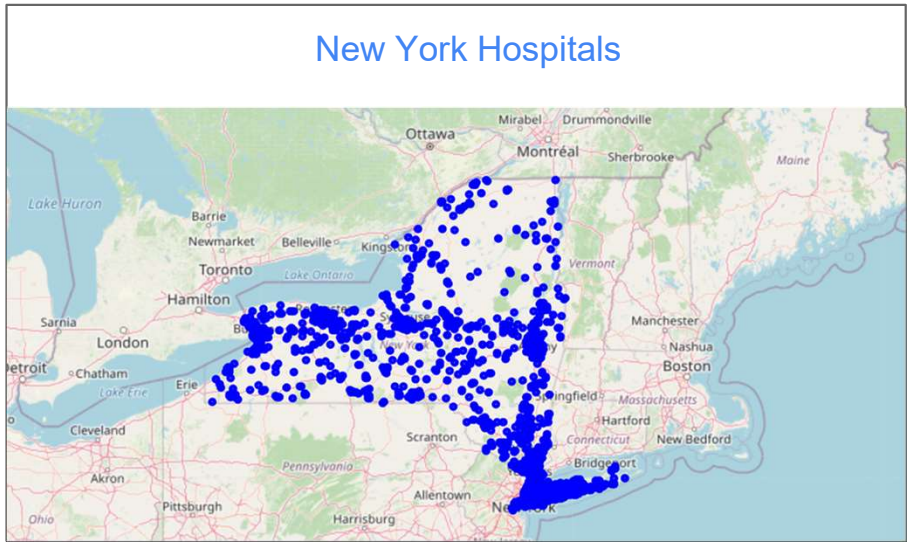
- The Yelp dataset we intended to use does not have New York data—in fact it only contains a small sample of data even in the states present in the dataset *(left)*
- The best sampled states were found to be Indiana *(middle)* and New Jersey *(right)*, but there is insufficient data across even these states and so we abandoned the Yelp dataset

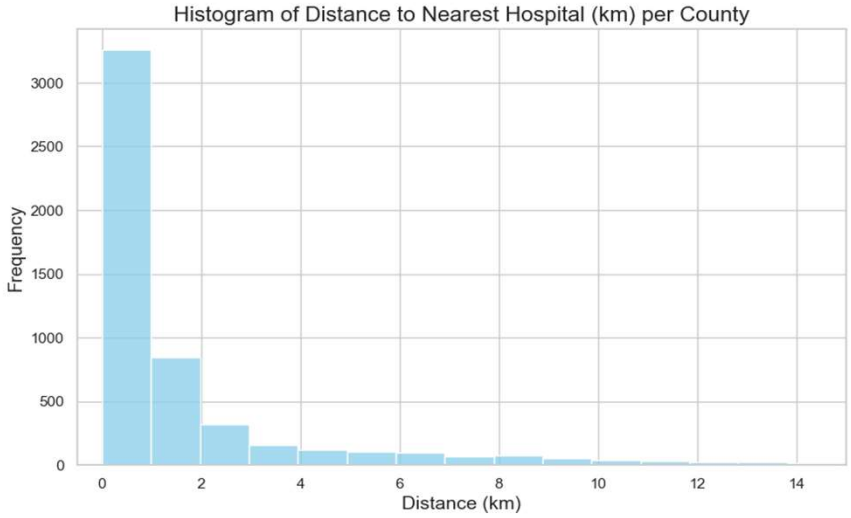




New York Hospitals Dataset

- To replace the Yelp dataset, we decided to use government data on New York Hospitals, this data is not a sample and covers all of New York *(left)*
- We then used the Haversine formula to engineer a feature: distance to nearest hospital *(right)*
- This feature may simply be a proxy for indicating whether the county is urban or rural

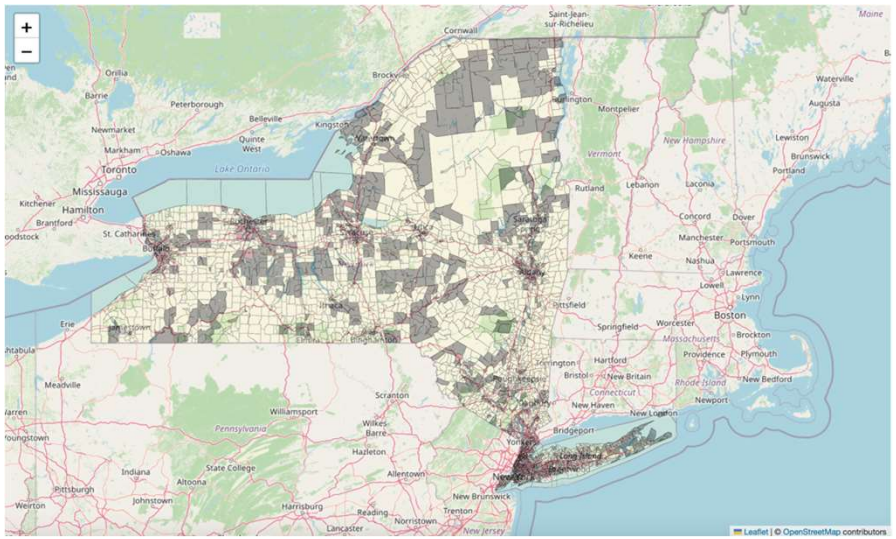




Smart Location Database Visualizations

Data Exploration

- Filtered out the rows for ‘STATEEP=36’, data for NY state.
- Merged the SLD data with Census Block Group boundaries shapefile.
- Explored the data correlation.
- Analyzed the density distribution both in heatmap and density plot.



The missing values in the grey polygons

Data Preprocessing

- Checked missing value and filled them with the interpolation.
- Initialize the scaler of MinMaxScaler and fit transform the data.
- Visualized the density distribution both in heatmap and

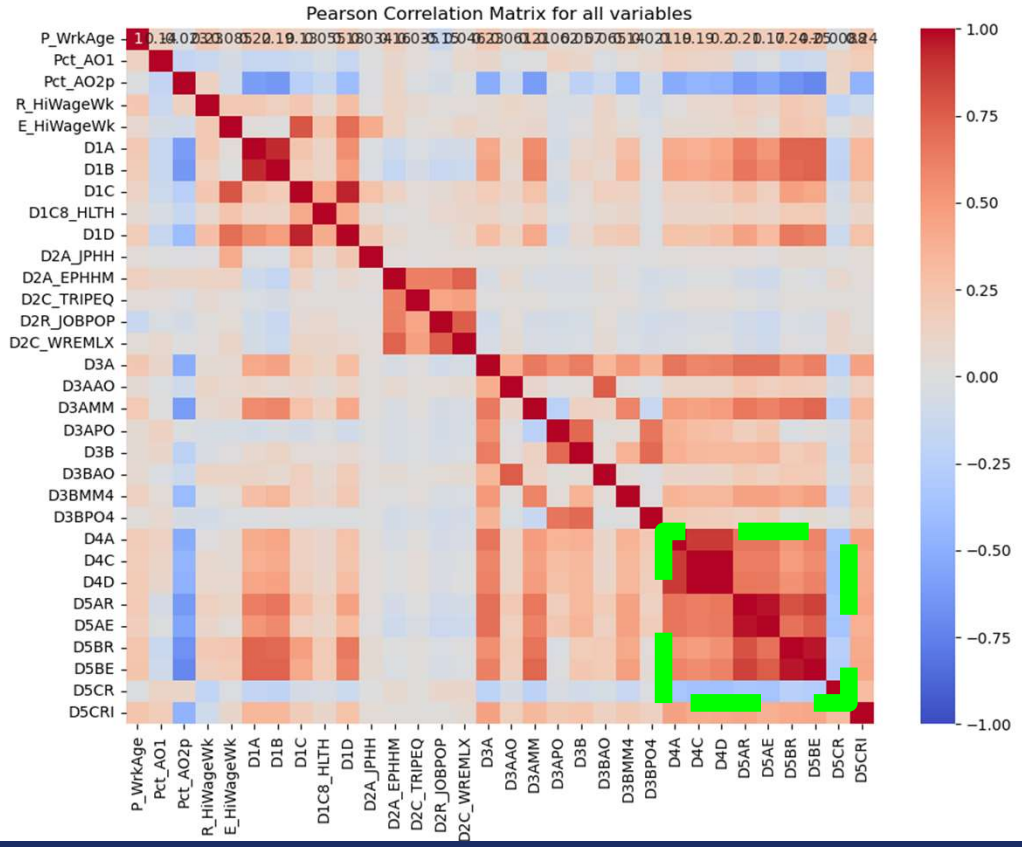
Smart Location Database Visualizations

Explored the data correlation

High correlation between ‘D1A’ and ‘D1B’; ‘D4A’ ,‘D4C’ and ‘D4D’; ‘D5AR’and ‘D5AE’; ‘D5BR’and ‘D5BE’.

D1a: Gross residential density (HU/acre) on unprotected land
D1b: Gross population density (people/acre) on unprotected land
D4a, Distance from the population-weighted centroid to nearest transit stop (meters) ,2020
D4c, Aggregate frequency of transit service within 0.25 miles of CBG boundary per hour during evening peak period, 2020
D4d, Aggregate frequency of transit service [D4c] per square mile
D5ar, Jobs within 45 minutes auto travel time, time- decay (network travel time) weighted, 2020
D5ae, Working age population within 45 minutes auto travel time, time-decay (network travel time) weighted
D5br , Jobs within 45-minute transit commute, distance decay (walk network travel time, GTFS schedules) weighted
D5be, Working age population within 45-minute transit commute, time decay (walk network travel time, GTFS schedules) weighted

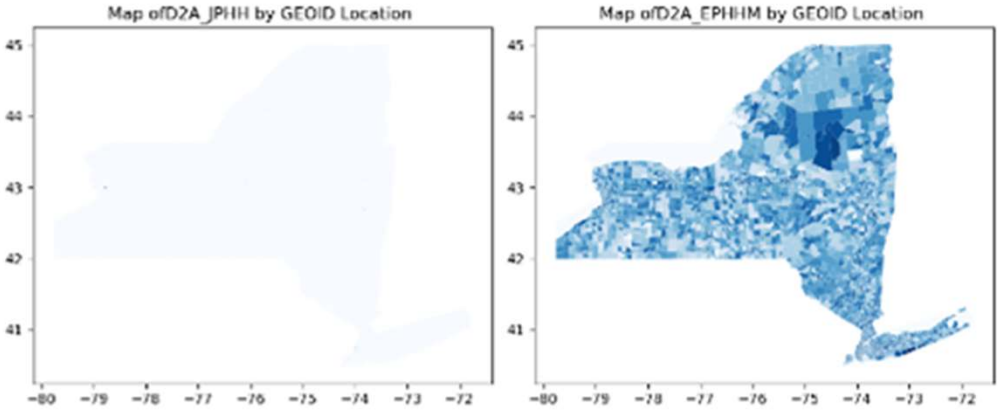
We keep D1a, D4a, D4c, D5AR, D5BE from the above



Smart Location Database Visualizations

Analyzed the density distribution both in heatmap and density plot

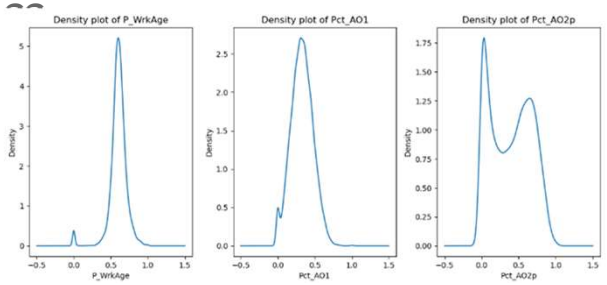
From the density map and plot, there are several features that are imbalanced, and with extremely large outliers. We remove the imbalanced features with most of zeros (looks empty in the density plot and signifies missing values). We finally filtered 18 features out of



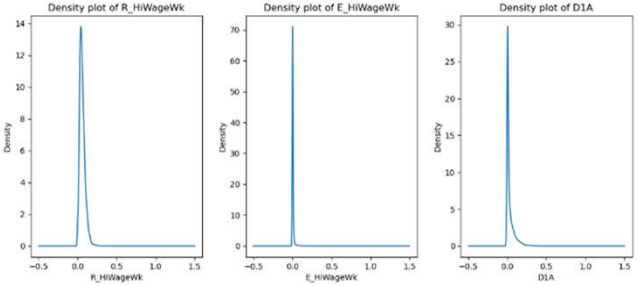
Feature removed

Feature kept

Features kept



Features removed



Proposed Models

Random Forest Regressor

- Purpose: Baseline model to explore interactions among variables.
- Strength: Interpretable feature importance, robust with complex datasets.

XGBoost (Gradient Boosting)

- Purpose: Provides accurate predictions by focusing on difficult cases.
- Strength: High performance on structured data with nonlinear relationships.

Geographically Weighted Regression (GWR)

- Purpose: Models spatial variations by allowing local adjustments to predictors.
- Strength: Reveals geographic trends; captures community-specific influences.

Feedforward Neural Network

- Purpose: Captures complex, nonlinear patterns across all predictors.
- Strength: Flexible in detecting hidden relationships across high-dimensional data.