Urban Food Production

CS 896 Capstone Project - Lokesh Das

Week 2 – Data Understanding & Initial Exploration

Team 9:
Venkata Subba Rao Are
Navya Boddupalli
Mallikarjun Maguluri

Project Objective

- Identify suitable areas for urban food production to enhance community food security.
- Analyze health, socioeconomic, and built-environment factors influencing site suitability.
- Understand spatial disparities before suitability modeling.
- Perform initial exploratory data analysis (EDA) to validate data quality and relationships.

Transition: Data exploration serves as the foundation for future suitability modeling.

Datasets Used

1. Health Dataset – CENSUS TRACT LEVEL

- Obesity prevalence
- Diabetes prevalence
- Disability prevalence

2. Census Dataset – TRACT LEVEL

- Poverty rate
- Median income
- Demographic breakdowns

3. ZCTA Dataset – ZIP LEVEL

- ZIP-level health indicators
- Cross-regional health trends

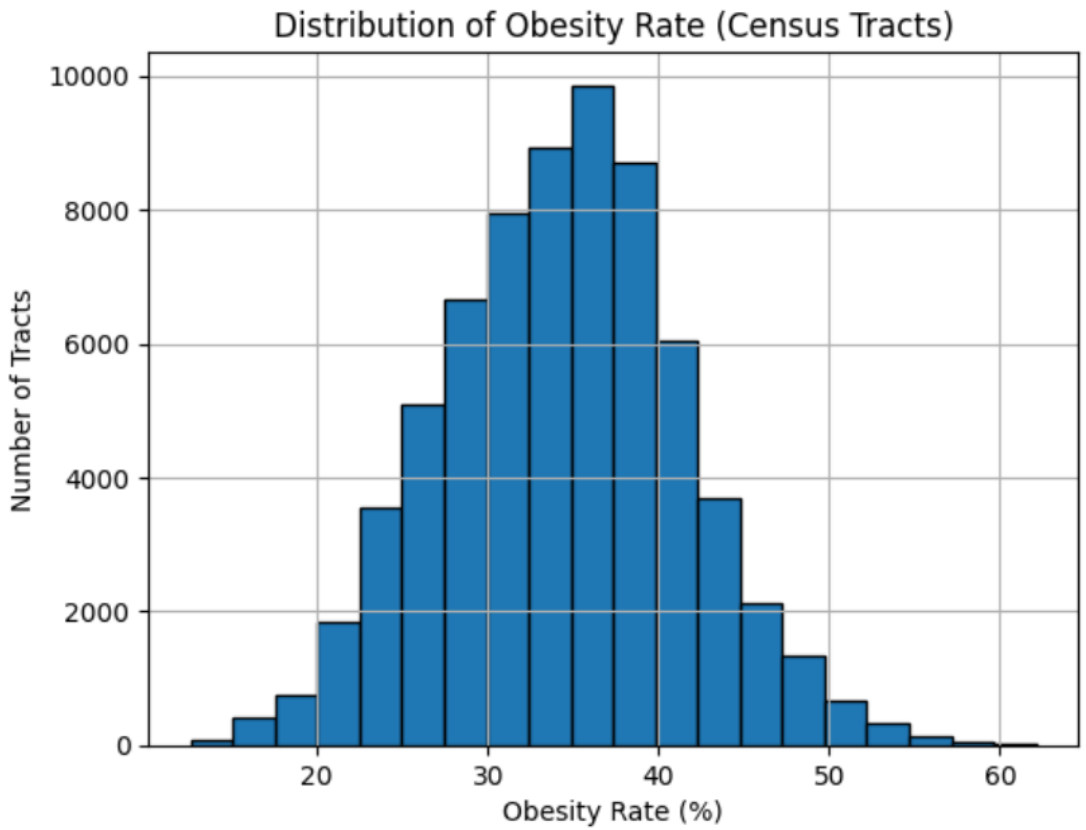
4. Parcels Dataset – PARCEL LEVEL

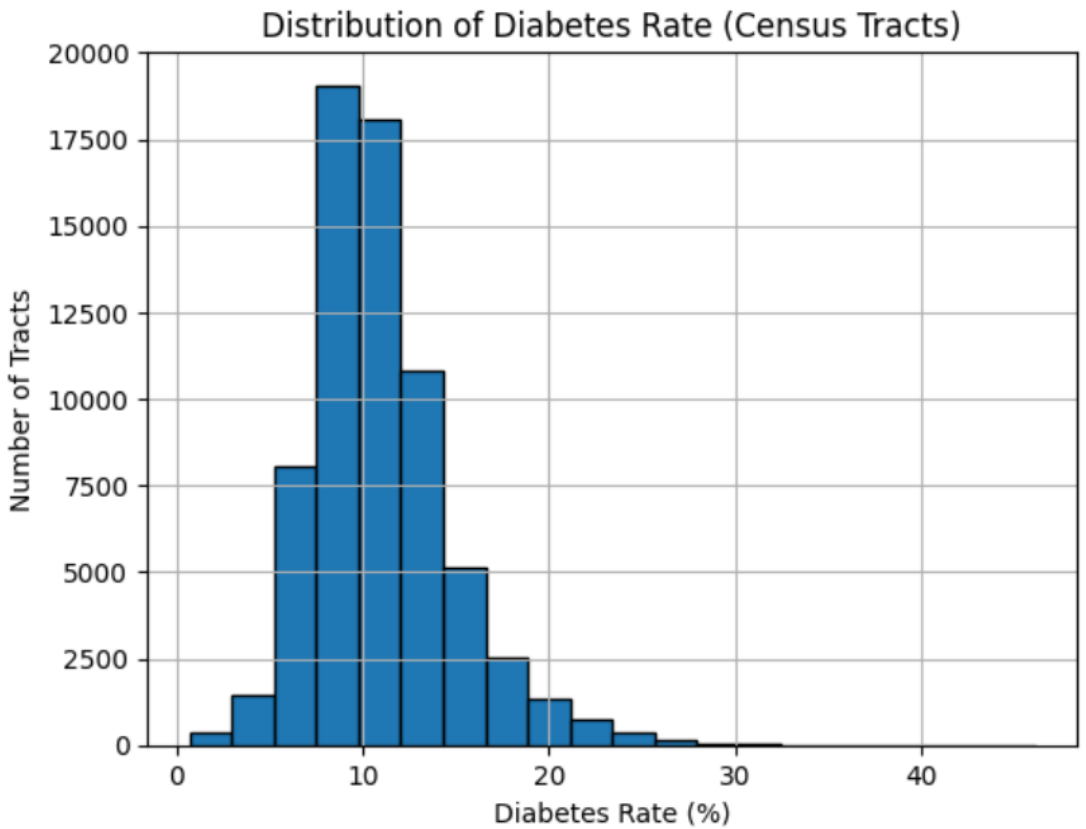
- Vacant land parcels
- Parcel geometry and area
- Built-environment constraints

Dataset Rows and Columns

Dataset Name	Geographic Level	Rows	Columns
Census Tract GIS-friendly	Tract-level	72,337	81
PLACES Local Health	Tract-level	2,555,113	23
ZCTA GIS-friendly	ZIP/ZCTA-level	32,409	77
PLACES Place Data	Place-level	2,054,768	21

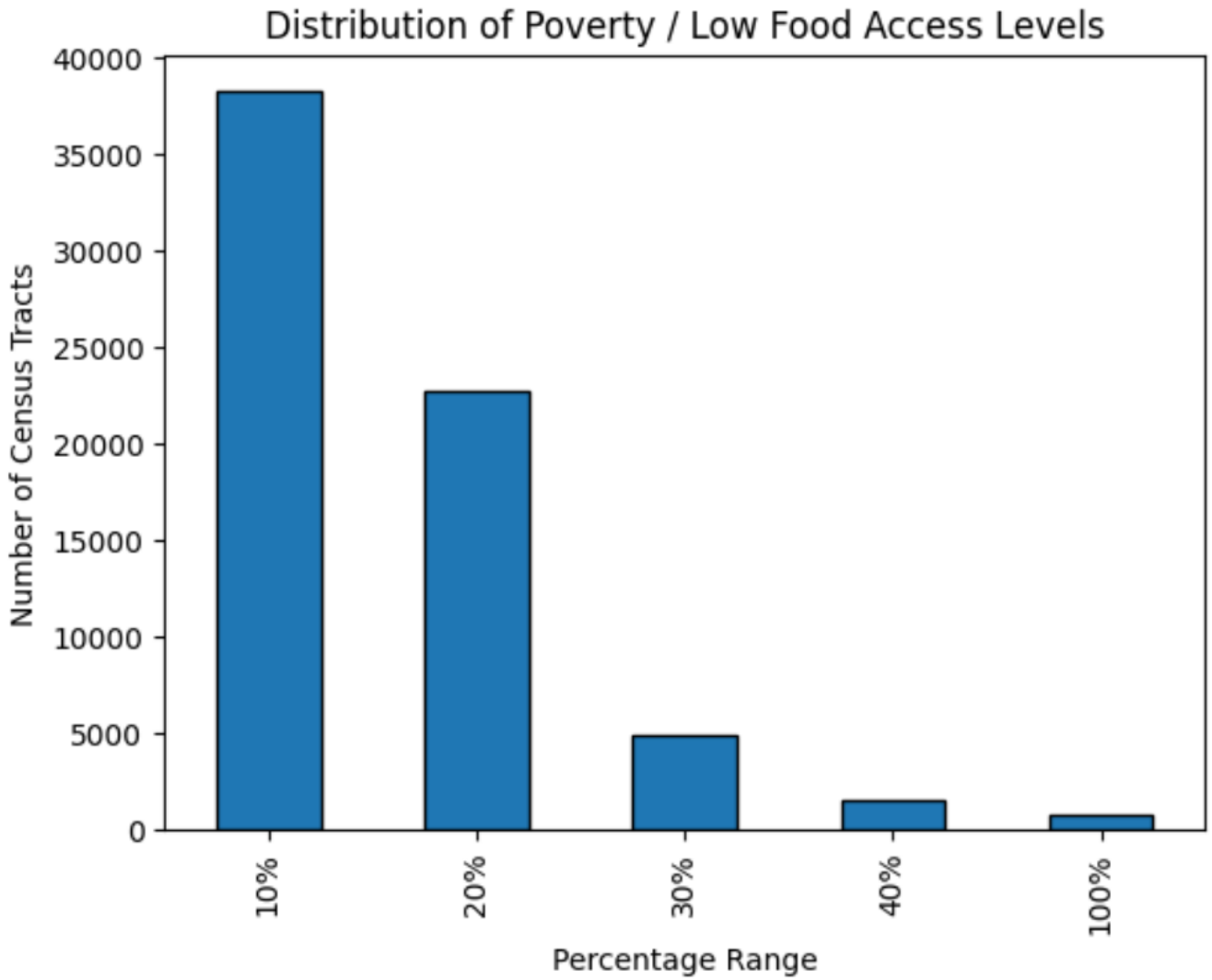
Health Indicators – Initial EDA





- Obesity rates show a clear central tendency around 30-40%, with a relatively symmetric bell-shaped distribution across tracts.
- Diabetes rates exhibit a right-skewed distribution, with most tracts clustering at lower prevalence levels but having a long tail of high-need areas.
- The spread of both indicators suggests significant spatial variability, confirming the need for localized suitability modeling.

Socioeconomic Variables – Initial EDA



- The categorical distribution highlights that a vast majority of tracts fall into the lower percentage ranges of poverty/low food access.
- This distribution will be used to prioritize "food deserts" where high poverty coincides with minimal access to fresh produce.
- Identifying these priority areas is a critical step before layering built-environment constraints in the suitability model.

Demographics – Initial EDA

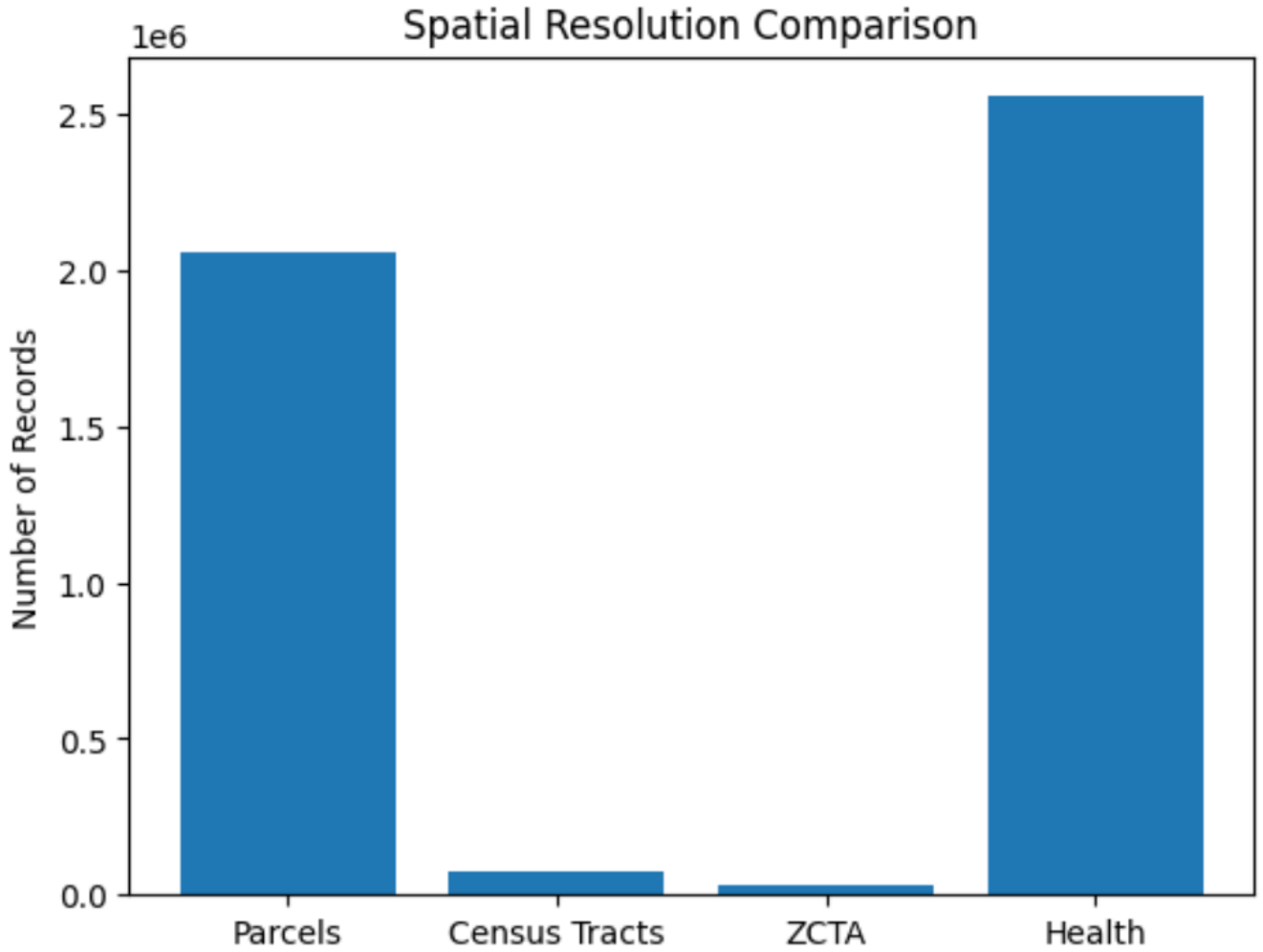
Variables for Exploration:

- Total population & density
- Age structure (Elderly/Youth)
- Race/Ethnicity composition
- Household composition
- Education attainment levels
- Vehicle access / Mobility

Planned EDA Checks:

- Assess data missingness and identify spatial outliers in demographic counts.
- Analyze correlations between demographics and health indicators (Obesity/Diabetes).
- Map spatial patterns to visualize demographic clusters across census tracts.

Geographic Scale Differences



- The vast differences in record counts between Parcels/Health and Tract/ZCTA levels highlight the Modifiable Areal Unit Problem (MAUP).
- Mixed geographic scales necessitate precise alignment and spatial joins to avoid data loss or misrepresentation during aggregation.
- Week 2 focuses on documenting a consistent integration strategy to ensure all layers contribute accurately to the suitability model.

Week 3 Plan

- Gather raw datasets (tract, ZCTA, parcels)
- Clean + standardize fields/IDs (TractFIPS, ZCTA5)
- Validate joins + fix missing/duplicates
- Harmonize scales (aggregation + spatial joins)
- Output analysis-ready dataset + data dictionary

Weeks 3–4: Data Prep (cleaning & integration)

Thank You