Project 2 Proposal

Team Members
Eliot Stein, Dan Manzano, Ivan Peteh

Github Repo
https://github.com/UC-Berkeley-I-School/datasci200_project2_stein_peteh_manzano

Primary Dataset
https://catalog.data.gov/dataset/walkability-index8
https://www.epa.gov/system/files/documents/2023-10/epa_sld_3.0_technicaldocumentationuserguide_may2021_0.pdf
https://catalog.data.gov/dataset/walkability-index8/resource/356986ec-b9ab-4fdf-8838-7262a08502e3

This data set contains 1 table broken into 11 sections

- The Primary Key for this whole data set is the OBJECTID
- Administrative

  Description:
  This section contains geographic and jurisdictional identifiers for Census Block Groups (CBGs), enabling linkage to other datasets (e.g., Census, LEHD) and regional analyses.

  Important Variables:

  1. GEOID20: Unique 12-digit FIPS code for the CBG (2018 boundaries). Primary key for joins.
  2. STATEFP/COUNTYFP/TRACTCE/BLKGRPCE: Hierarchical codes for state, county, tract, and block group.
  3. CSA/CBSA: Combined Statistical Area and Core-Based Statistical Area codes (for metro/micro regions).
  4. CSA_Name/CBSA_Name: Text labels for CSA/CBSA (e.g., "New York-Newark, NY-NJ-PA").

  Use Cases

  - Use GEOID20 (not GEOID10) for consistency with 2018+ datasets.
  - CBSA fields help aggregate block groups into metropolitan/micropolitan areas.

- Core-Based Statistical Area Measures

  Description:

Provides regional context by summarizing population, jobs, and workers for the broader CBSA (metropolitan/micropolitan area) containing each block group.

Important Variables:

CBSA_Pop: Total population in the CBSA (2018 ACS). Use to calculate % of regional population in a CBG.

CBSA_Emp: Total jobs in the CBSA (2017 LEHD). *Helps assess job concentration (e.g., "Does this CBG hold 1% or 10% of regional jobs?").

CBSA_Wrk: Total workers residing in the CBSA (2017 LEHD). Use with CBSA_Emp to analyze jobs-housing balance regionally.

Use Cases

- CBSA-level metrics enable comparisons like:
- *"Is this CBG’s job density (D1c) above/below the CBSA average?"*
- "Do low-wage workers (R_LowWageWk) cluster in high-job CBSAs?"

- Area

Description:
Quantifies land acreage, distinguishing between total, water, and unprotected land (developable area). Critical for density calculations.

Important Variables:

1. Ac_Total: Total CBG area (acres). Includes water and protected land.
2. Ac_Land: Land area excluding water (acres).
3. Ac_Unpr: Land not protected from development (excludes parks, conservation areas). Used as denominator in density metrics (D1).
4. Ac_Water: Water area (acres).

Use Cases:

- Key nuance: Density variables (D1a–D1d) use Ac_Unpr, not Ac_Total (except when D1_Flag=1).
- Example analysis:
  - "Do CBGs with >20% protected land (Ac_Total - Ac_Unpr) have lower residential density (D1a)?"

- Demographics

  Description:
  This section captures population and household characteristics, including age distribution, auto ownership, and worker wage stratification. It helps analyze socioeconomic patterns and transportation behavior at the block group level.

  Important Variables:

  1. TotPop: Total population (2018 ACS estimates).
  2. CountHU: Total housing units (occupied + vacant).
  3. HH: Occupied households (used for density/entropy calculations).
  4. P_WrkAge: % of the working-age population (18–64 years).
  5. AutoOwn0/AutoOwn1/AutoOwn2p: Households with 0, 1, or 2+ cars.
  6. Workers: Total workers residing in the block group (from LEHD RAC data).
  7. R_LowWageWk/R_MedWageWk/R_HiWageWk: Workers stratified by income (≤$1,250/mo, $1,251–$3,333/mo, ≥$3,333/mo).

  Use Case:
  Study car dependency (e.g., zero-car households vs. transit access) or income disparities in job accessibility.

- Employment

  Description:
  This section categorizes jobs by sector (using NAICS codes) and wage levels at workplace locations (from LEHD WAC data). It supports land use diversity and economic analyses.

  Important Variables:
  1. TotEmp: Total jobs in the block group.
  2. E5_Ret/E5_Off/E5_Ind/E5_Svc/E5_Ent: Jobs grouped into 5 tiers (retail, office, industrial, service, entertainment).
  3. E8_Ret/E8_Off/E8_Ind/E8_Svc/E8_Ent/E8_Ed/E8_Hlth/E8_Pub: More granular 8-tier classification (adds education, healthcare, public admin).

  More information for each category respectively:

  - Retail jobs within a 5-tier employment classification scheme (LEHD: CNS07), 2017
  - Office jobs within a 5-tier employment classification scheme (LEHD: CNS09 + CNS10 + CNS11 + CNS13 +CNS20) , 2017

    - Industrial jobs within a 5-tier employment classification scheme (LEHD: CNS01 + CNS02 + CNS03 + CNS04 +CNS05 + CNS06 + CNS08) , 2017
    - Service jobs within a 5-tier employment classification scheme (LEHD: CNS12 + CNS14 + CNS15 + CNS16 +CNS19) , 2017
    - Entertainment jobs within a 5-tier employment classification scheme (LEHD: CNS17 + CNS18), 2017
    - Retail jobs within an 8-tier employment classification scheme (LEHD: CNS07), 2017
    - Office jobs within an 8-tier employment classification scheme (LEHD: CNS09 + CNS10 + CNS11 + CNS13) ,2017
    - Industrial jobs within an 8-tier employment classification scheme (LEHD: CNS01 + CNS02 + CNS03 + CNS04 + CNS05 + CNS06 + CNS08) , 2017
    - Service jobs within an 8-tier employment classification scheme (LEHD: CNS12 + CNS14 + CNS19) , 2017
    - Entertainment jobs within an 8-tier employment classification scheme (LEHD: CNS17 + CNS18), 2017
    - Education jobs within an 8-tier employment classification scheme (LEHD: CNS15), 2017
    - Health care jobs within an 8-tier employment classification scheme (LEHD: CNS16), 2017
    - Public administration jobs within an 8-tier employment classification scheme (LEHD: CNS20), 2017

4. E_LowWageWk/E_MedWageWk/E_HiWageWk: Jobs by wage level at workplace locations.

Use Case:
Compare job type distribution (e.g., retail vs. office) with walkability or transit access.

- Density (D1)

  Description:
  Measures activity intensity per acre of unprotected land (excluding parks/water). Critical for assessing urban compactness.

  Important Variables:

  1. D1a: Residential density (housing units/acre).
  2. D1b: Population density (people/acre).
  3. D1c: Employment density (jobs/acre).
  4. D1d: Combined activity density (jobs + housing units/acre).
  5. D1c5_Ret/D1c8_Ed/etc.: Sector-specific job densities.
  6. D1_Flag: Indicates if density uses total land area (if unprotected area <0.5%).

Use Case:
Identify high-density areas for transit prioritization or sprawl analysis.

- Diversity (D2)

  Description:
  Quantifies land use mix via entropy metrics and job-housing balance. Higher values indicate more diverse neighborhoods.

  Important Variables:

  1. D2a_JpHH: Jobs per household.
  2. D2b_E5Mix/D2b_E8Mix: Employment entropy (5- or 8-tier; higher = more balanced mix).
  3. D2a_EpHHm: Combined jobs/household entropy.
  4. D2r_JobPop: Jobs/population ratio relative to regional average.
  5. D2a_WrkEmp: Workers per job (local balance).

  Use Case:
  Test if mixed-use areas (high entropy) correlate with reduced car dependency.

- Design (D3)

  Description:
  Measures street network connectivity and pedestrian/auto orientation using intersection and road densities.

  Important Variables:

  1. D3a: Total road network density (miles/sq. mile).
  2. D3aao/D3amm/D3apo: Auto-oriented, multimodal, or pedestrian-oriented road densities.
  3. D3b: Weighted intersection density (prioritizes pedestrian-friendly 4-way intersections).
  4. D3bmm4/D3bpo4: Multimodal/pedestrian 4-way intersections per sq. mile.

  Use Case:
  Compare grid-like vs. suburban street designs and their impact on walkability.

- Transit Access (D4)

  Description:
  Evaluates proximity and service quality of transit (GTFS data) for each block group.

  Important Variables:

  1. D4a: Walk distance (meters) to nearest transit stop.
  2. D4b025/D4b050: % of jobs within ¼-mile or ½-mile of fixed-guideway transit (e.g., rail).
  3. D4c: Peak-hour transit frequency (trips/hour) near CBG.
  4. D4d/D4e: Transit frequency per sq. mile or capita.

  Use Case:
  Identify "transit deserts" or assess equity in service distribution.

- Walkability Index (NatWalkInd)

  Description:
  Composite score (1–20) combining intersection density (D3b), land use mix (D2b_E8MixA, D2a_EpHHm), and transit proximity (D4a).

  Important Variables:

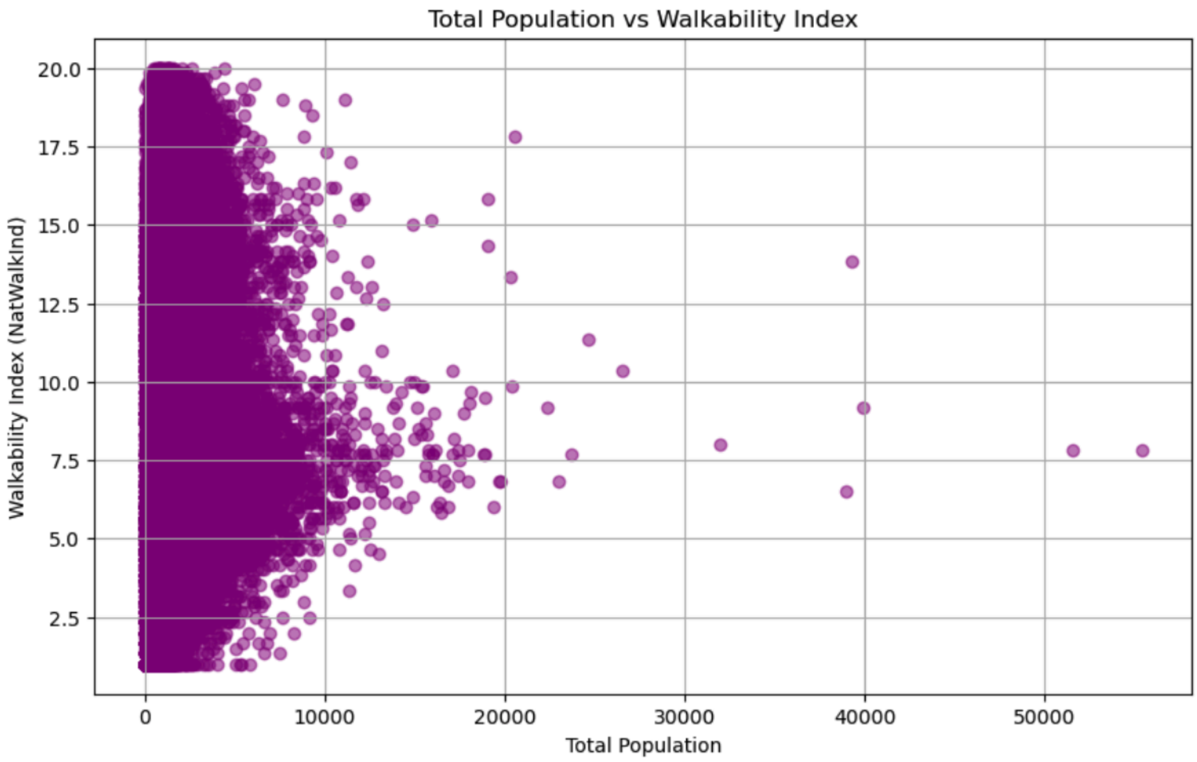
  1. NatWalkInd: Walkability score (higher = more walkable).
  2. D2A_Ranked/D3B_Ranked/etc.: Quantile-ranked components (1–20).

  Use Case:
  Rank neighborhoods for urban planning or public health studies.

Initial Exploration:

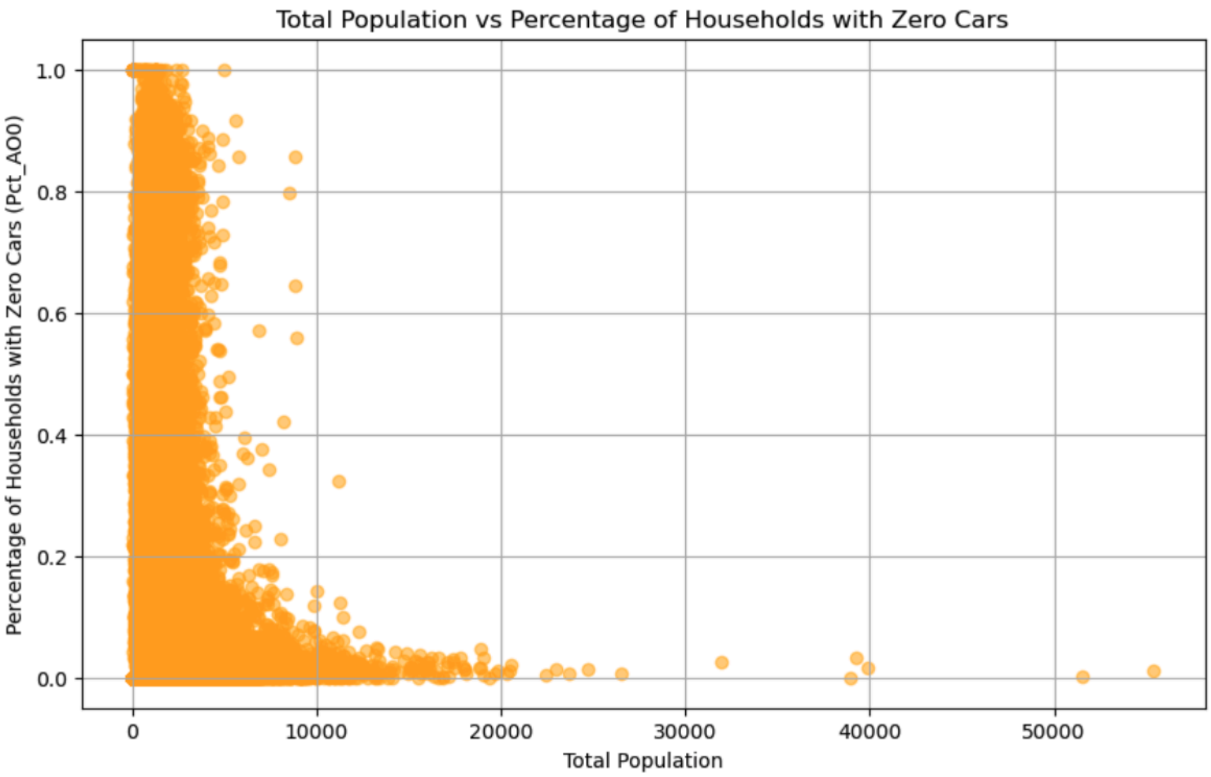
Total Population vs. Walkability Index (`NatWalkInd`)

- See if more populous areas tend to have higher walkability



Total Population vs. Zero Car Households (`AutoOwn0` or `Pct_AO0`)

- Understand how population size relates to car ownership patterns.
- Larger populations may or may not have proportionally more zero-car households.



What we plan to cover in the final report:

What type of Employment Categories have employees that walk to work?
What is the average wage of employees that walk to work?
Compare total employment to total number of workers based on walkability. (how many jobs there are vs. how many workers there are to fulfil the jobs)
What is the average number of household workers per job that walk to work?
What metropolitan areas have a high road network density and transit density with low amounts of people walking to work?
Which Core Based Statistical Areas (CBSAs)(e.g., New York, Los Angeles, Chicago) have the highest composite walkability scores (NatWalkInd
How do the number of households per acre vary based on walkability?