





Best counties to live in the U.S.

SECTION 1:

This project—Best Places to Live in the U.S.—is designed to help identify the most suitable places for four distinct household types: young families, first-generation immigrants, recent graduates, and empty nesters. These groups often face different constraints and have unique needs related to affordability, safety, education, job access, healthcare, and lifestyle quality.

With rising housing costs, uneven access to services, and growing regional disparities, a data-driven approach that helps individuals make more informed decisions about where to live is needed. This project curates and analyzes a broad set of county-level metrics across the U.S., including cost of living, healthcare access, school quality, diversity, and more. The outcome will be a set of tailored insights and geospatial visualizations to support decision-making for these key demographic groups.

Stakeholder Group	Household Type	Goals / Interests	Constraints	Key Metrics to Include
 Low- to Middle-Income Young Families	2 adults + 1 child (under 18)	Safe, affordable living with good schools & healthcare	Housing costs, job access, safety	Cost of living, rent, crime rate, school rating
 First-Gen Immigrants	2 adults + 1 child (under 18)	Diverse, welcoming places with support services & job opportunities	Language, cultural fit, affordability	Diversity index, ESL access, job availability
 Recent Grads / Young Professionals	1 person	Career growth + lifestyle balance, affordability, social life	Limited income, student debt, job market, commute	Youth population %, income vs. cost ratio, leisure score, job access
 Empty Nesters / Pre-Retirees	1–2 adults, age 50–65	Calm, affordable, health-focused lifestyle with culture/recreation	Fixed income, healthcare needs, walkability	Healthcare quality, cost of living, climate comfort, crime rate

Data Source

The final dataset was generated by integrating seven distinct datasets sourced from various U.S. federal agencies. These datasets, detailed in the attached data grid, include information on source descriptions, formats, reference links, and coverage periods. Most of the data spans a five-year range from 2019 to 2023, ensuring a consistent and comprehensive basis for analysis.

Data was sourced from publicly available government and non-profit datasets, including HUD, Census/ACS, and County Health Rankings. County-level granularity was prioritized to enable localized insights, and standard geographies (like FIPS codes) were used to integrate multiple sources into a unified dataset.

Data Source Summary

The following county-level datasets were merged to create one unified file covering U.S. counties.

- **file_1** – Geographic boundaries for population estimates
Period: 2020–2024 | **Source:** U.S. Census Bureau
- **file_2** – Social, economic, and health-related indicators
Period: 2020–2024 | **Source:** County Health Rankings (UW + RWJF)
- **file_3** – 2024 Cost of Living Index
Period: 2024 | **Source:** World Population Review
- **file_4** – Median household income estimates
Period: 2018–2022 | **Source:** U.S. Census Bureau
- **file_5** – County population by age & race/ethnicity
Period: 2019–2023 | **Source:** U.S. Census Bureau (ACS)
- **file_6** – State & local sales tax rates
Period: 2025 | **Source:** Tax Foundation
- **file_7** – Fair Market Rent for 1-bedroom units
Period: 2020–2024 | **Source:** U.S. HUD

Why This Dataset(s)

This dataset was built to support a multi-dimensional analysis of **the best places to live** in the U.S., tailored to different population groups. It includes metrics aligned with key quality-of-life factors such as affordability, health, safety, opportunity, and community diversity. Each data source was selected to reflect real-life decision-making criteria (e.g., families with children, young professionals, and pre-retirees).

Suitability for geospatial, regression, and clustering:

The data is standardized at the **county level** and includes **FIPS codes**, making it ideal for geospatial mapping and analysis. With hundreds of numeric indicators, the dataset is well-suited for:

- **Clustering** counties into typologies

- **Regression analysis** to find drivers of housing cost or life expectancy.
- **Geospatial modeling** to visualize trends and disparities across the U.S.

Diversity of topics:

This dataset brings together variables across **multiple domains**:

- **Economy:** median income, cost of living, rent burden
- **Health:** access to providers, physical distress, life expectancy
- **Safety:** violent crime, injury deaths
- **Education:** high school graduation, some college rates
- **Demographics:** age groups, race/ethnicity breakdown, diversity index
- **Housing & Environment:** Fair Market Rent, severe housing problems, pollution

SECTION 2:

Data Profile

General note about limitations & ethics regarding all input datasets

All data this analysis uses comes from open and publicly available federal or national sources. There are no legal restrictions or limitations on its use for research, policy, or planning purposes.

The primary ethical consideration is interpreting specific statistics — especially sensitive indicators such as violent crime rates. These are based on FBI data, the federally designated source for such statistics. However, reporting practices vary by jurisdiction; not all local agencies submit complete data. As a result, figures may be incomplete or inconsistently reported across counties.

It is recommended not to make direct community-level recommendations or comparisons solely based on this data point. Instead, it should be contextualized within a broader set of indicators and interpreted cautiously to avoid misrepresenting public safety conditions.

File 1

This dataset was from a federal source — the U.S. Census Bureau's Annual Population Estimates (CO-EST2024-POP). It provides county-level geographic boundaries and population estimates for all U.S. counties as of January 1, 2024. The data came clean and required minimal preprocessing. Standard quality checks were performed, including validation for duplicate rows and null values. No issues were found, confirming its readiness for analysis and merging using FIPS codes.

File 2

This foundational dataset, compiled by the University of Wisconsin Population Health Institute in collaboration with the Robert Wood Johnson Foundation, includes 45 key county-level metrics. It aggregates national data on health outcomes and social determinants from trusted sources like the CDC, CMS, and the U.S. Census.

This dataset served as the **backbone** of the entire analysis — all other county-level datasets were joined to this base using FIPS codes. Due to its breadth, it required substantial preprocessing, including de-duplication, renaming, and null value auditing. For each metric, missing data was documented in the Python workflow using markdown cells to track completeness and imputation levels.

For all variables included in the analysis, missing values were first assessed on a **yearly basis** across five years (2020–2024). For example, the *flu vaccinations* variable had a missing rate of approximately **0.69%–0.72%** per year before imputationmulti_data.

To handle these gaps, we implemented a **state-level mean imputation strategy**. Specifically, for each year and variable, missing county-level values were replaced with the average of that variable calculated **within the same state**. This approach preserves geographic context and avoids distorting regional distributions

File 3

This dataset from *World Population Review* provides a breakdown of cost of living (COL) scores for U.S. states in 2024. It includes an overall index and category-specific subindexes (e.g., housing, transportation, groceries). All values are scaled to a national average of 100.

Standard data cleaning procedures—such as renaming columns, adjusting casing, and checking for null or duplicate values—were applied. The data was mostly clean.

The main limitation is granularity: the dataset provides only **state-level** observations, so county-level variation in the cost of living cannot be captured.

File 4

This dataset, sourced from the U.S. Census Bureau, provides 5-year estimates of county-level median household income, adjusted for inflation as of 2023.

The dataset is well-documented and reliable as a federal source. Standard data cleaning procedures (e.g., column validation, checking for nulls and duplicates) were applied, and the dataset was clean and ready for analysis. No major preprocessing was required.

File 5

This dataset, sourced from the U.S. Census Bureau's American Community Survey (ACS 5-Year Estimates, published 2024), provides granular county-level demographic data. It includes population counts across 378 columns, covering detailed age groups, race/ethnicity, and Hispanic origin.

All values are based on 5-year estimates (not single-year snapshots), making it suitable for small-area trend analysis. Margins of error were included in the original dataset but were removed to streamline the file.

The Diversity Index was calculated using Simpson's Index ($D = 1 - \sum(p_i^2)$, where p_i is the proportion of each racial/ethnic group in the population). Age groups were standardized into eight federally aligned brackets. Standard cleaning and column renaming were applied. The data was clean, consistent, and ready for use in diversity and demographic segmentation analyses.

File 6

This dataset provides a snapshot of base state-level and population-weighted local sales tax rates across U.S. states as of January 1, 2025. It was sourced from the Tax Foundation and official state revenue sites.

The dataset came as clean, well-structured raw data that required no additional preprocessing or cleaning. All variables were ready for direct use in comparison and ranking analyses.

File 7

This dataset represents the average Fair Market Rent (FMR) for 1-bedroom units across all U.S. counties and county equivalents. It was generated by calculating the mean rent values across 5 years (2020 to 2024) using annual data from HUD.

The dataset was derived programmatically, including standardized county identifiers (FIPS) for merging. It was clean and ready for use without additional preprocessing beyond averaging.

SECTION 3:

Questions to Explore

Focused Research Questions

The core of this project lies in asking the *right questions*. These guide both the design and application of the analysis:

- **For young families:**
Which counties offer the best combination of income, school quality, and healthcare?
- **For first-gen immigrants living in LA County:**
Is it still affordable to remain in California? What are the viable alternatives with cultural and support infrastructure?
- **For future homeowners in California:**
What counties offer better housing affordability while maintaining similar services to LA?
- **For anyone considering leaving California:**
Which states or counties offer a better socio-economic and climate profile, with access to nature or the ocean?
- **For general exploration:**
What hidden regional opportunities can be uncovered through visual and statistical pattern analysis?