Distribution of Amenities Across Neighborhoods in SF

Group 15 // MTC Project 5
Megan Hu and Neha Suresh

Track A (Hoping to do Track B for final if feasible)

GitHub Repo Link

November 10th, 2025

1. Problem & Motivation

Defining Amenities:

For this project, we define amenities as desirable features and services that enhance the quality of life, comfort, and convenience for residents in neighborhoods. To manage this broad definition, we categorize amenities as follows:

1. Public services: Schools, libraries, public transportation, and healthcare facilities like hospitals and clinics
2. Recreational: Parks, playgrounds, community centers, swimming pools, gyms, and walking or biking trails
3. Commercial: Shops (grocery), restaurants, coffee shops, and other local businesses
4. Natural features: natural areas like lakes, rivers, mountains, or open green spaces
5. Infrastructure: Basic needs like roads, electricity, and water supply

This is not an exhaustive list, as there are more amenities like museums and tourist attractions. We'll focus on categories 1-3, given the time constraints for this project and the lack of prior research for these categories.

Motivating Problem: *Why is **equitable** access to amenities important in neighborhoods?*

Analyzing the equitable distribution of neighborhood amenities is important because it directly impacts two core groups of stakeholders: residents and the community at large. For residents, equitable access leads to immediate improvements in their health and well-being through features like walkability and access to essential services (grocery, healthcare, etc.), and also results in a higher overall quality of life, ensuring convenience and satisfaction. The positive impact extends broadly to the community, boosting social cohesion by offering shared spaces where people can gather, build trust, and form relationships. Ultimately, this research promotes the potential for reduced isolation and discrimination, while stimulating economic opportunity by helping create more vibrant, inclusive neighborhoods that support both development and job creation.

Updated Research Question and Hypothesis:

How do the size and distribution of neighborhood amenities in San Francisco relate to the socioeconomic and racial demographics of neighborhoods at the census-tract level for the most recent available data (most are 2023 or updated as needed), and what are the quantifiable disparities in access?

Success Metrics: *How are we **quantifying** the disparities in access?*

1. Composite Amenity-Access Indices: For each type of amenity (parks and recreation or hospitals), we'll form an index from standardizing per-capita availability, proximity (travel-time or walkability), and amenity quality, and potentially adding weights.
2. Gini Coefficient: For each tract, use the Census API (B19082) to get the Gini Coefficient for each tract as a measure of income inequality (0 close to equality, 1 is inequality).

3. Proportion of Non-Hispanic White Residents: Use the US Census API to calculate this by doing B03002_003E (count white, non Hispanic) / B03002_001E (total population count)
4. OLS Regression Coefficients: Fit an OLS model with 3 and 4 as features and the target variable is 1. Interpret the coefficients to understand how they 3 and 4 individually affect amenity scores, being aware of the effects of collinearity.
5. Predictors and Variables

Variable Name	ACS Code	Description	Why Needed
population_total	B03002_001E	Total population	Normalization denominator for all per-capita amenity metrics
white_nonhisp	B03002_003E	White Non-Hispanic (count)	Demographic composition (race)
percent_white_nonhisp	derived	White Non-Hispanic proportion	Key racial equity variable
gini	B19083_001E	Income inequality (0–1)	Income inequality metric for regressions
median_income	B19013_001E	Median household income	Economic stratification and control variable
poverty_denom	B17001_001E	Poverty denominator	Required for poverty rate
poverty_count	B17001_002E	Poverty count	Required for poverty rate
poverty_rate	derived	Poverty rate	Socioeconomic covariate (SES)
transit_commuters	B08301_010E	Workers commuting by public transit	Commute access proxy (accessibility indicator)

Unit & Level of Analysis: The analysis for this project is focused on the census tract-level data, a decision that is motivated as a means to balance spatial detail with high-coverage ACS data. We recognize that analysis at the neighborhood level would be inefficient, and thus census-tract level is the smallest unit for which we can reliably work with detailed demographic, economic (Gini), and amenities data.

2. Data & Cleaning

2.1 Data Description

American Community Survey (ACS 2015-2024): This is data accessed via the Census API, specifically targeting the area of San Francisco county. Given the timeline of interest (2015-2024), the Census API was queried for intervals across that time period at: 2015, 2018, 2020, 2022, and 2023.

TIGER/Line Tract Polygons (2023): This was queried to access the census tract geometries for San Francisco. Since the fluctuations for geometry lines were expected to be little to none across 2015-2024, the geometry was queried from 2023.

Walk Score APIs: Walk Score is the platform that hosts APIs for WalkScore, TransitScore, and BikeScore. Analysis was focused on the WalkScore and TransitScore metrics, where the centroids of tracts were calculated and used as input when querying.

Parks and Recreation (csv) & City_Facilities_20251109 (geojson): These files are from the Open Data Portal and include a facility ID identifier column as well as point geometries for each facility point. For both these sets, the data is updated very recently to match the current facilities in SF.

2.2 EDA and Cleaning:

American Community Survey (ACS 2015-2024): ACS data was queried by year and standardized across GEOID, tract formats, and column names. Some variables were derived from existing ACS predictors, these include the percent of white (non-hispanic) residents within a given tract, as well as poverty rate utilizing a given predictor variable and dividing by population. For the ACS dataframe, data cleaning included filtering out zero population tracts (some are special facilities; non-residential areas) as well as pseudo tracts.

TIGER/Line Tract Polygons (2023): Then, the TIGER geometry dataframe was merged to build a geometry table which then allowed for the calculation of tract centroids and their given latitudes and longitudes.

Walk Score APIs: Lastly, the WalkScore API was queried once per GEOID then left-joined into the comprehensive ACS Dataframe across selected years.

Parks and Recreation (csv) & City_Facilities_20251109 (geojson): The Facilities geojson file has a department_name column that provides more granular information about the type of amenity it is (ex. MTA, Public Health) while both of these datasets have “address” and “common_name” categories which provide their names and searchable locations. The Parks and Recreation dataset includes all the Parks and Recreation data in the City Facilities data, as well as others not in SF, but are within the discretion of SF. These 2 sets were inner (spacially) joined to sf_tracts data which is extracted from the TIGER/Line Tract Polygons (2023) data using SF FIPs code. Through the respective notebooks for each of these files, I followed a rigorous data cleaning process, starting with ensuring the data types were encoded with the right types, exploring

inconsistencies in the data like Groveland data, and exploring individual addresses to ensure that the interpretations make sense.

2.3 Bottlenecks:

ACS 2015-2024, TIGER/Line Tract Polygons (2023), Walk Score APIs:

Some of the tract centroids fell into *invalid* zones that triggered the WalkScore API status 30, which just means that the given latitude and longitude coordinates were not in bound within the WalkScore database. This portion of data was retained and flagged as missing as a means to better understand and visualize the distribution of missing WalkScore data across San Francisco in the year of 2018. Another roadblock was when GeoPandas flagged the centroid calculations within a geographic CRS, but since the WalkScore API lookup only needs a *general* coordinate it was treated as acceptable and any spatial kernels were reprojected to EPSG:26910.

Parks and Recreation (csv), TIGER/Line Tract Polygons (2023), City_Facilities_20251109

My biggest bottleneck (besides addressing missing data, which is detailed below) is that there wasn't tract-level data in the Parks and Recreation and City Facilities data, but I was able to spacially join these GeoDataFrames with SF tract data from the TIGER/Line Tract Polygons shp file and get the GEOID column, which is at tract level.

2.3.1 Missing Data Analysis:

WalkScore's were missing for 78 rows across the ACS San Francisco Panel, largely clustered between the years 2015 and 2018. In Hand's framing within *Dark Data*, it is understood that these data were systematically neglected and not collected rather than random error. This type of missing data usually reflects coverage limits, and were visualized in the analysis after sensitivity checks found that missing San Francisco tracts have a slightly lower median income in the year 2018.

The relevant missing data in the Parks and Recreation and City Facilities are from the gross_sq_ft columns. This data falls under the "not data dependent" category because this data is missing because some of the facilities are either too new, too small, mobile, changing, or difficult to quantify in some way, but the FRRM is updating the data in real time. In the notebooks, I've outlined how and why I handled the missing data in detail. For a quick summary, I removed data points not in SF and of amenities that aren't related to our definition of amenities. I interpolated data using medians or 25th quantiles after exploring individual addresses with missing data (and without), and I've provided more details for this process in the notebooks.

3. Initial Results & Visualizations

3.1 WalkScore vs. Median Income (2018): This scatter plot represents WalkScore (y-axis) against median income (x-axis) per tract within San Francisco for 2018. Each individual plot represents a tract, and is assigned a color based on percent of population being White (non-hispanic). The gray X-markers represent tracts with unavailable WalkScores, sitting slightly

below 0 on the y-axis within the plot. From these markers, we see that missing WalkScore tends to cluster at the low-middle median income level, with essentially no clustering within the high median income population.

3.2 Transit Score Choropleth (2022): This choropleth maps the TransitScore across San Francisco in 2022. At first glance, it seems that there is a higher TransitScore across the northeast region of San Francisco. The hover tooltip allows for closer analysis into the median socio-economic status and distribution of race within each tract, which viewers analyze patterns within demographics.

3.3: City Facilities Map (html): This is an interactive map that explores the point geometry of city facilities overlaid on a map of SF tracts. The points are categorized by their overarching amenity types, as in our definition of amenities above.

3.4: Parks and Recreation Top 20 Gross Sq Ft Boxplot: For the top 20 tracts with the greatest total parks and recreation gross square footage across all facilities, I plotted the distribution of the gross square footage.

3.5: Gross Sq Ft by Amenity Category Box Plot: For each of the different types of amenities, I plotted a box plot describing the distribution of the gross square footage.

3.6: Tracts Gross Sq Ft Amenity Types: This is a choropleth map with the gross square footage quantiles colored in each census tract and with the points of each facility color-coded by amenity type plotted over.

4. Progress Since Proposal & Plan to Completion

4.1 Progress since proposal (with MTC guidance): Scope narrowed to census tracts, and a clean pipeline now pulls ACS 2015–2023 data and standardizes GEOIDs. Poverty rate and percent White non-Hispanic were derived, then joined to TIGER 2023 tracts for consistent geometry. Following MTC guidance, the Walk Score stack was prototyped by scoring tract centroids and merging WalkScore, TransitScore, and BikeScore into the panel. “Status 30” returns were logged as informative missingness rather than discarded. An interactive 2022 TransitScore choropleth with hover diagnostics now mirrors the Tulsa City Index example (MTC provided example), which is complemented by initial equity plots and a concise variable codebook that is generated when running the San Francisco ACS notebook.

4.2 Changes from proposal: Instead of jumping to multi-amenity per-capita indices, we prioritized WalkScore and TransitScore as an access proxy to validate our joins and reveal early equity patterns while matching MTC’s quick-win mapping guidance. Additional layers are queued for the next sprint.

4.3 Plan to completion:

November 12 - 15:	Modularize code from midterm submission; Clean up github page; Consolidate feedback from MTC mentor
November 16 - 21:	Add amenity and data layers; Implement feedback; Refine tract-level analysis
November 22 - 27:	Implement appropriate statistical analysis and frameworks depending on status of implemented data
November 28 - December 5:	Ensuring code is modularized and published on Github pages; Begin final report writing
December 6 - 12:	Refine and submission

5. References

City and County of San Francisco. San Francisco Open Data. https://data.sfgov.org/. (Accessed November 10, 2025).

Ferreira, Fernando. "Location Choice." The Wharton School, University of Pennsylvania, https://real-faculty.wharton.upenn.edu/fferreir/wp-content/uploads/~fferreir/documents/LocationChoice_65FF.pdf. (Accessed November 10, 2025).

Hand, D. J. (2020). *Dark data: Why what you don't know matters*. Princeton University Press.

"14 Neighborhood Amenities That Can Increase Home Value." Discover Big Sky, https://discoverbigsky.com/blog/14-neighborhood-amenities-that-can-increase-home-value. (Accessed November 10, 2025).

U.S. Census Bureau. (2025). *American Community Survey (ACS) 5-year estimates* (Tables B03002, B19013, B19083, B17001, B08301). Census Data API. https://www.census.gov/data/developers/data-sets/acs-5year.html (Accessed November 10, 2025).

U.S. Census Bureau. (2023). *TIGER/Line® Shapefiles: 2023, Census tracts, California*. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html. (Accessed November 10, 2025).

Walk Score. (2025). *Walk Score API documentation*. https://www.walkscore.com/professional/api.php (Accessed November 10, 2025).

Walk Score. (2025). *Public Transit API documentation*. https://transit.walkscore.com, (Accessed November 10, 2025).