Optimal Store Location Analysis Using Geospatial and Demographic Data

By: Aditya Prabhu

Table of Contents

Objective **3**
Data Understanding **4**
Data Cleaning and Preparation **5**
Geospatial Analysis **6**
4.1 Mapping Demographics 6
Business Recommendations **8**
Conclusion **9**

Objective

Problem:
Starbucks aims to expand its store network strategically to maximise revenue, market penetration, and customer accessibility while avoiding cannibalisation of existing stores.

Goal:
In this project, I analysed Starbucks' current US store locations alongside county-level population and income data to:

- Identify counties with high market potential and low store density
- Recommend optimal expansion strategies based on demographic and geographic insights

Business Impact:
Data-driven location decisions can increase revenue, strengthen brand presence, and improve operational efficiency by targeting high-opportunity areas.

Data Understanding

Datasets Used:

1. **Starbucks Locations Dataset**
    - Source: Kaggle
    - Key Variables: Store Name, City, State, Latitude, Longitude
2. **US County Demographics Dataset (ACS 2017)**
    - Source: US Census / Kaggle
    - Key Variables: County, State, Total Population, Median Household Income
3. **US Counties Latitude-Longitude Dataset**
    - Source: Simplemaps
    - Key Variables: County, State, Latitude, Longitude, County Population

Data Cleaning and Preparation

To ensure the datasets were ready for accurate geospatial analysis, I performed the following cleaning and preparation steps:

1. **Filtered Starbucks Dataset for US Stores Only**
   I began by filtering the global Starbucks dataset to include only stores located within the United States. This allowed me to focus the analysis on the US market and maintain consistency when integrating with US-specific demographic data. The dataset was reduced to over 14,000 US store records, providing a comprehensive view of Starbucks' national footprint.

2. **Removed Records with Missing Coordinates**
   Accurate latitude and longitude data is essential for any geospatial analysis. I identified and removed records with missing or null geographic coordinates to avoid inaccuracies in mapping and spatial calculations. This ensured that all stores plotted on the map represented valid physical locations.

3. **Cleaned County Naming Conventions to Enable Merging Across Datasets**
   When preparing to merge the county demographic dataset with the Simplemaps latitude-longitude dataset, I observed inconsistencies in county naming formats. For example, the ACS demographic dataset included the word "County" (e.g. "Autauga County"), whereas the Simplemaps dataset used only the county name (e.g. "Autauga"). I standardised the county names by removing the word "County" in the demographic dataset to ensure a seamless and accurate merge.

4. **Merged Demographic Data with County Latitude-Longitude Data**
   After standardising naming conventions, I merged the ACS demographic dataset with the Simplemaps US counties dataset using both county and state as keys. This integration brought together critical information such as total population, median household income, latitude, and longitude into a single dataframe. The resulting merged dataset provided a comprehensive view of each county’s demographics along with geographic coordinates necessary for mapping.

5. **Validated Merged Data and Assessed Coverage**
   Finally, I assessed the merged dataset for completeness. Out of 3220 US counties, the merge successfully integrated approximately 3000 counties, with the remaining unmatched counties largely due to unique naming formats (e.g. boroughs in Alaska or parishes in Louisiana). For this analysis, the successfully merged data covered over 93% of US counties, sufficient for national-level strategic insights.

Geospatial Analysis

4.1 Mapping Demographics

To begin the geospatial analysis, I created an interactive Folium map visualising US counties by their demographic attributes:

- **Circle Size:** Represented the total population of each county. Larger circles indicated counties with higher population counts, highlighting market size and potential customer base.

- **Colour:** Represented median household income. I used a dual-colour scale:

    - **Blue** for counties with median household incomes greater than or equal to $60,000, signifying higher purchasing power and potentially greater demand for premium coffee products.

    - **Green** for counties with median household incomes below $60,000, indicating moderate to lower purchasing power, which may influence product pricing, store design, and menu offerings.

Insights:

Through this demographic mapping, several important patterns emerged:

1. **Concentration in Metropolitan Areas:**
   Counties with the largest populations and highest median incomes were predominantly located around major metropolitan areas such as New York City, Los Angeles, Chicago, Houston, and San Francisco. This aligns with broader economic and urbanisation trends where income levels and population density are highest in city centres and their surrounding counties.

2. **Regional Variations:**
   The Midwest and parts of the South showed counties with substantial populations but generally lower median incomes, suggesting a need for regionally tailored pricing or store format strategies.

3. **Sparse High-Income Counties:**
   Some counties with relatively small populations but high median incomes, such as certain suburban counties in California, New Jersey, and Connecticut, represent niche opportunities for premium or reserve store formats.

4.2 Overlaying Starbucks Stores

Building upon the demographic map, I overlaid Starbucks store locations as **red circle markers** to analyse store density and market coverage across counties.

Observations:

1. **High Store Density in Major Cities:**
   Starbucks store locations are heavily concentrated in and around large urban centres such as Seattle, San Francisco, Los Angeles, Chicago, New York City, and Washington D.C. This reflects Starbucks' core strategy of dense urban penetration to ensure brand visibility and customer convenience.

2. **Underpenetrated High-Potential Counties:**
   Despite significant urban coverage, the map revealed noticeable gaps in several counties with **high population and high median income** where Starbucks stores were either sparse or absent. For example:

   - Certain suburban counties in Texas, Georgia, and Florida showed high population counts with relatively few Starbucks stores, indicating expansion potential.

   - Some affluent counties in the Midwest and Northeast presented similar opportunities for targeted growth.

3. **Opportunities in Growing Regions:**
   Emerging suburban counties near fast-growing metropolitan areas showed increasing populations and incomes but limited Starbucks presence, suggesting strategic long-term expansion opportunities as these regions develop further.

Overall, this integrated geospatial analysis combining demographic mapping with existing store locations provided actionable insights to guide Starbucks' expansion strategy, ensuring that new stores are opened in counties with the highest demand potential and minimal internal competition.

Business Recommendations

1. **Target High-Income, High-Population Counties with Low Store Density**
   Why: These counties present the best opportunity for new stores with a higher likelihood of profitability due to existing purchasing power and market size.
   Recommendation: Prioritise counties with populations over 200,000 and median household incomes above $60,000 that currently have limited or no Starbucks presence.

2. **Conduct Competitive and Site Feasibility Analysis**
   Why: Although demographic potential is high, physical feasibility and local competition must be considered.
   Recommendation: Perform on-ground assessments and competitive analysis in shortlisted counties before final site selection.

3. **Integrate Customer Preference Data in Future Analysis**
   Why: Demographics provide potential, but incorporating lifestyle, consumer behaviour, and traffic data will improve decision precision.
   Recommendation: Enrich future models with loyalty data, spending patterns, and mobility data for granular targeting.

Conclusion

In this project, I conducted a geospatial analysis of Starbucks store locations integrated with US county demographic data to identify optimal expansion opportunities. The analysis revealed counties with high populations and incomes but low Starbucks density, suggesting strategic opportunities for growth.

This project demonstrates how geospatial intelligence combined with demographic insights can drive effective market expansion strategies for retail businesses.