Analytics for a New Coffee Brand Entering the USA Market

SENG425

MUSA YÜKSEL
21050911018

SENA DİLAN ÇAKIR
21050911027

CONTENT

1. RESEARSCH QUESTION
2. PIPELINE OVERVIEW
3. DATASET OVERVIEW
4. COMPETITION ANALYSIS
5. VISUALIZATION
6. MACHINE LEARNING MODEL

Research Question

In which US cities is demand high but competition relatively manageable, and is there purchasing power for premium coffee?

Pipeline Overview

Demographic filtering (Census)

Geographic matching (City → County)

Consumer interest (Yelp)

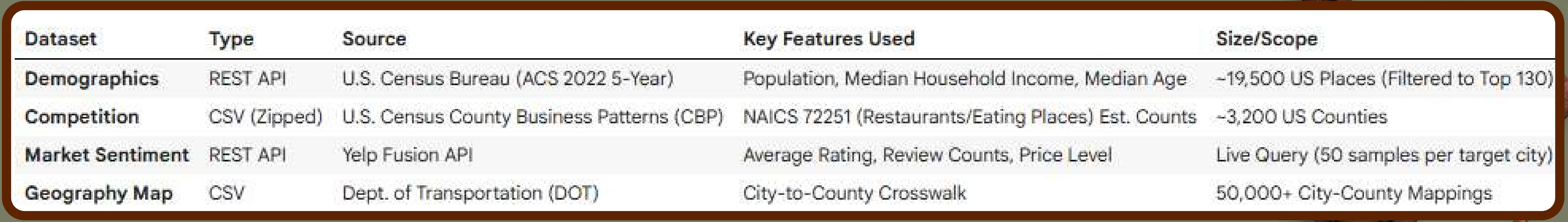
Competitive analysis (NAICS)

Opportunity Index

Visualization

Machine Learning (Clustering)

Datasets Overview



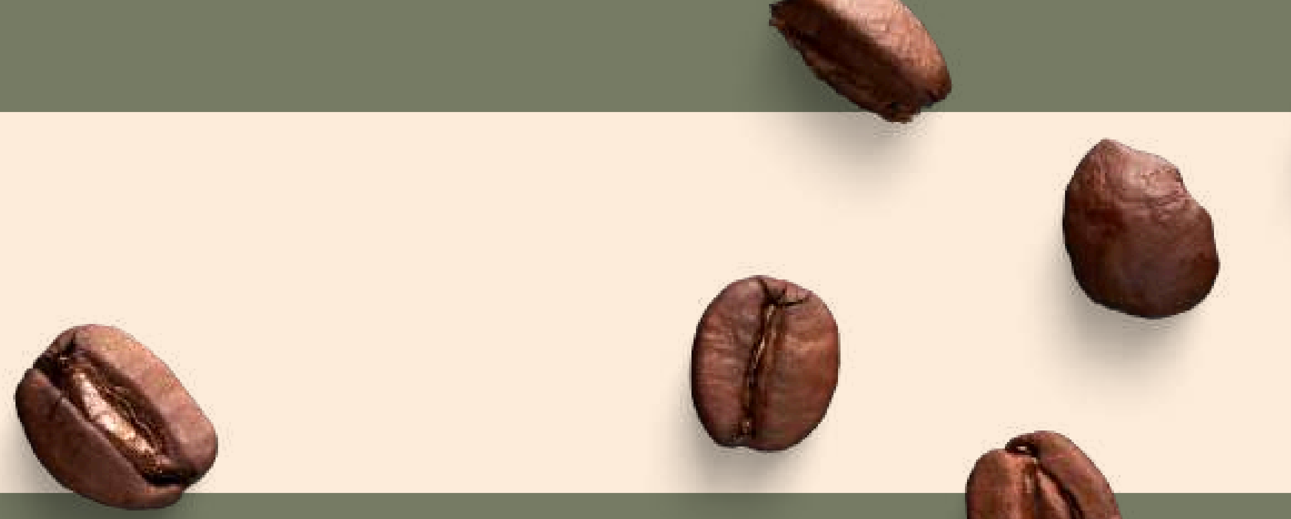
Dataset	Type	Source	Key Features Used	Size/Scope
Demographics	REST API	U.S. Census Bureau (ACS 2022 5-Year)	Population, Median Household Income, Median Age	~19,500 US Places (Filtered to Top 130)
Competition	CSV (Zipped)	U.S. Census County Business Patterns (CBP)	NAICS 72251 (Restaurants/Eating Places) Est. Counts	~3,200 US Counties
Market Sentiment	REST API	Yelp Fusion API	Average Rating, Review Counts, Price Level	Live Query (50 samples per target city)
Geography Map	CSV	Dept. of Transportation (DOT)	City-to-County Crosswalk	50,000+ City-County Mappings

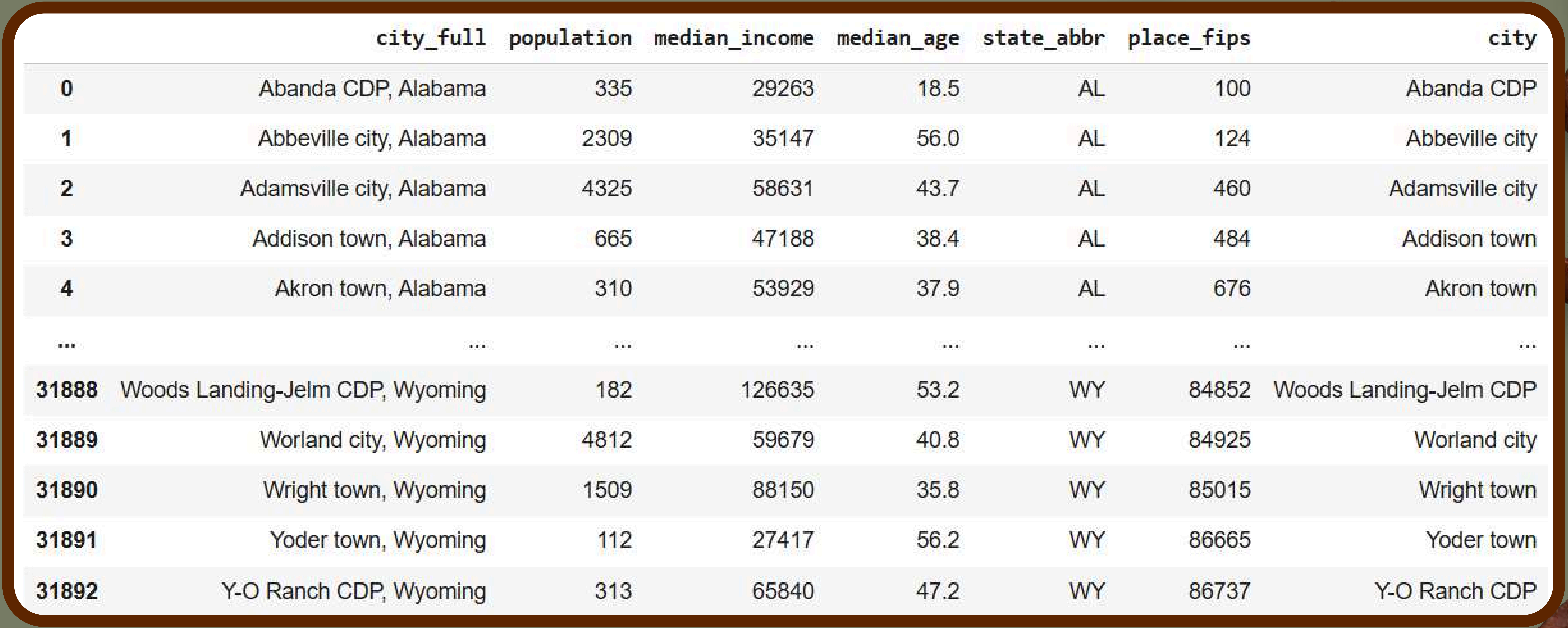
U.S. Census Bureau ACS 2022

Provides detailed demographic and socioeconomic information at the city level. The ACS dataset is comprehensive, standardized, and publicly available, making it a reliable source for cross-city comparisons.

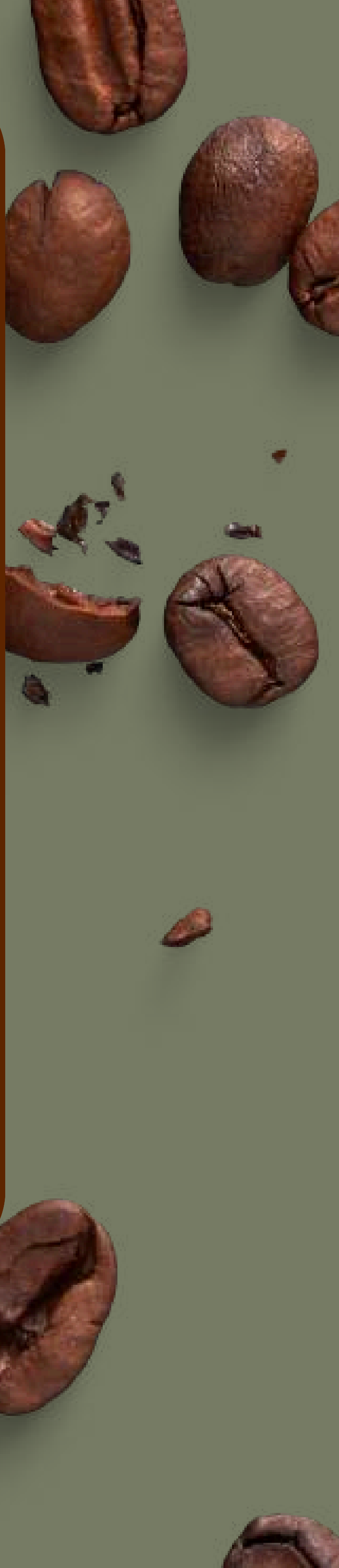
In this project, ACS data is used to capture core demand-side characteristics of U.S. cities, including **population size**, **median household income**, and **median age**. These variables form the **foundation of the demand analysis** by indicating **market size**, **purchasing power**, and **demographic suitability** for premium coffee consumption.

CENSUS DATA





	city_full	population	median_income	median_age	state_abbr	place_fips	city
0	Abanda CDP, Alabama	335	29263	18.5	AL	100	Abanda CDP
1	Abbeville city, Alabama	2309	35147	56.0	AL	124	Abbeville city
2	Adamsville city, Alabama	4325	58631	43.7	AL	460	Adamsville city
3	Addison town, Alabama	665	47188	38.4	AL	484	Addison town
4	Akron town, Alabama	310	53929	37.9	AL	676	Akron town
...	...	...	...	...	...	...	...
31888	Woods Landing-Jelm CDP, Wyoming	182	126635	53.2	WY	84852	Woods Landing-Jelm CDP
31889	Worland city, Wyoming	4812	59679	40.8	WY	84925	Worland city
31890	Wright town, Wyoming	1509	88150	35.8	WY	85015	Wright town
31891	Yoder town, Wyoming	112	27417	56.2	WY	86665	Yoder town
31892	Y-O Ranch CDP, Wyoming	313	65840	47.2	WY	86737	Y-O Ranch CDP



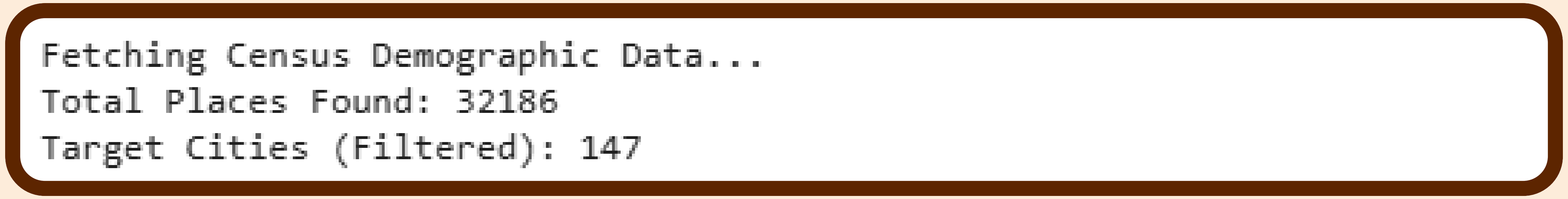
31893 rows × 7 columns

Demographic Filtering (The "Target" List)

US Census Bureau (ACS 5-Year Estimates, 2022)

Selection Criteria:

- Population: > 150,000 (Ensures sufficient foot traffic volume)
- Median Household Income: > $55,000 (Ensures purchasing power for premium coffee)
- Median Age: 22-45 years old (The core coffee-consuming demographic)



```
Fetching Census Demographic Data...
Total Places Found: 32186
Target Cities (Filtered): 147
```

Department of Transportation

This dataset maps each city to its corresponding county and Federal Information Processing Standards (FIPS) codes.

Additional cleaning and normalization steps are applied to correct malformed or inconsistent FIPS identifiers, ensuring accurate geographic alignment across datasets.

Geographic Normalization (City-to-County Crosswalk)

Problem: Census city, CBP county-based

- Demographic data is available at the City level (e.g., "Seattle").
- Reliable business competition data (NAICS) is only available at the County level (e.g., "King County").

Solution: DOT Crosswalk + FIPS cleansing

We use a Department of Transportation (DOT) Crosswalk file to map every city to its corresponding county FIPS code.

Yelp Fusion API

The Yelp Fusion API is used to collect sampled café-level data for each target city. While Yelp does not provide a complete census of all businesses, it offers valuable proxies for consumer engagement and market sophistication.

Consumer Engagement Signals (Yelp Fusion API)

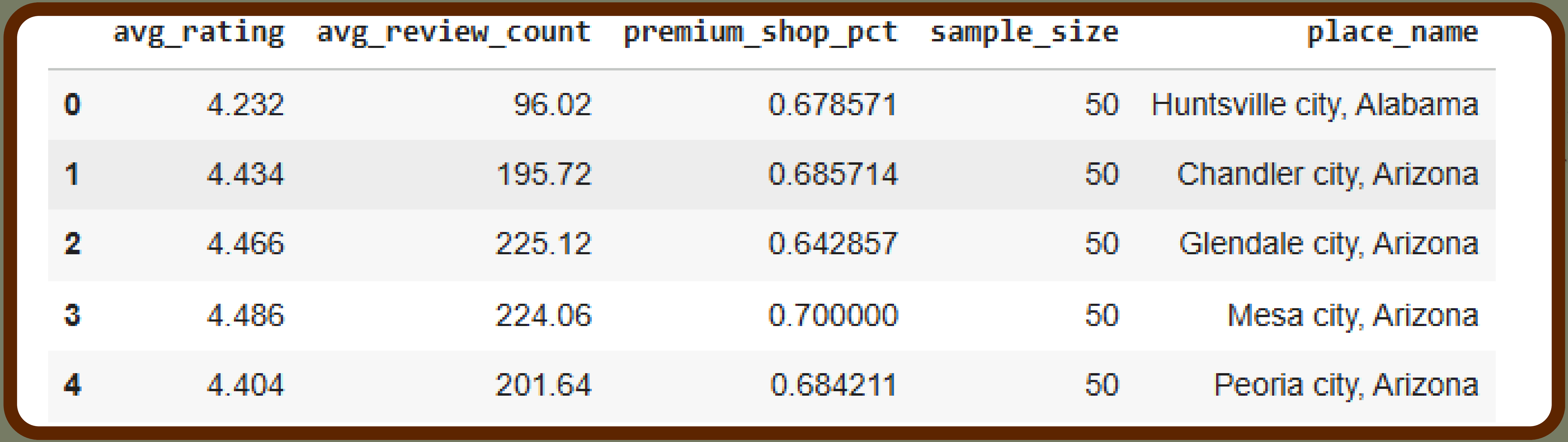
Metrics extracted

- Average rating
- Average review count
- Share of premium-priced cafés

Purpose

- Proxy for coffee culture sophistication
- Measures consumer engagement beyond population size

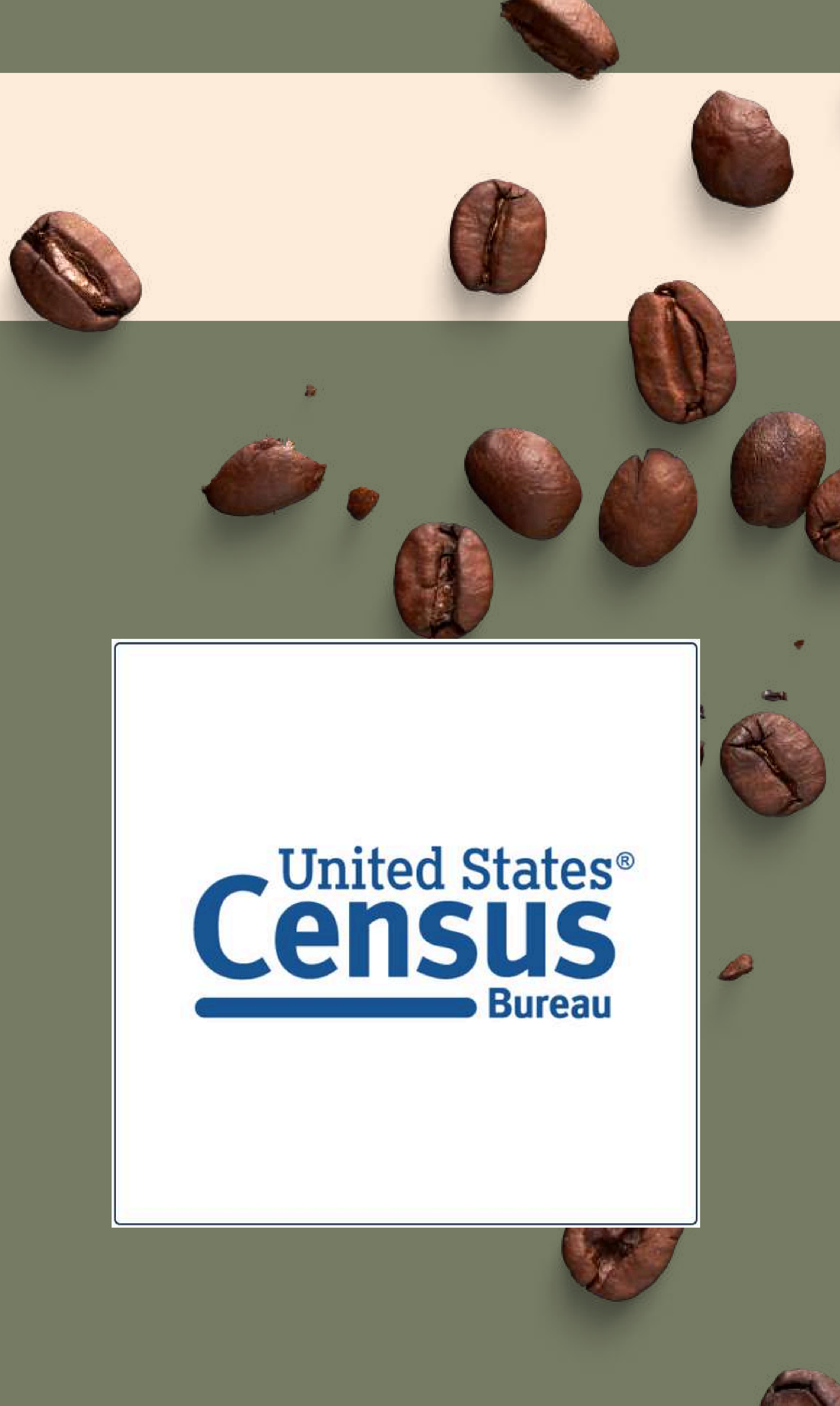
Querying Yelp API



	avg_rating	avg_review_count	premium_shop_pct	sample_size	place_name
0	4.232	96.02	0.678571	50	Huntsville city, Alabama
1	4.434	195.72	0.685714	50	Chandler city, Arizona
2	4.466	225.12	0.642857	50	Glendale city, Arizona
3	4.486	224.06	0.700000	50	Mesa city, Arizona
4	4.404	201.64	0.684211	50	Peoria city, Arizona

U.S. Census Bureau CBP

Competition intensity is measured using County Business Patterns (CBP) data published by the U.S. Census Bureau. CBP provides establishment counts by industry at the county level using NAICS classifications. This project uses NAICS group 72251 (Restaurants and Similar Eating Places) as a proxy for coffee shop competition, as consistent 6-digit café-specific counts are not available at the county level. CBP data offers a comprehensive and unbiased view of market saturation, avoiding the sampling limitations inherent in online platforms.



Competition Analysis

Competition Dataset

- Source: Census County Business Patterns (CBP)
- NAICS Code: 72251 (Restaurants & Eating Places)

Metrics:

- Total competitor count per county
- Density per 10,000 residents

Why CBP?

- More reliable and unbiased than Yelp listings

The Opportunity Index (Scoring Model)

All datasets are consolidated into a single analytical dataframe, from which three normalized indices (scaled between 0 and 1) are constructed to evaluate market entry potential.

- Demand Index
- Competition Index
- Opportunity = D - 0.6 × C

Demand Index (D)

Demand Index Composition

- Median income (40%)
- Population size (30%)
- Yelp review activity (30%)

Interpretation

- High values indicate wealthy, active, and large consumer markets

Competition Index (C)

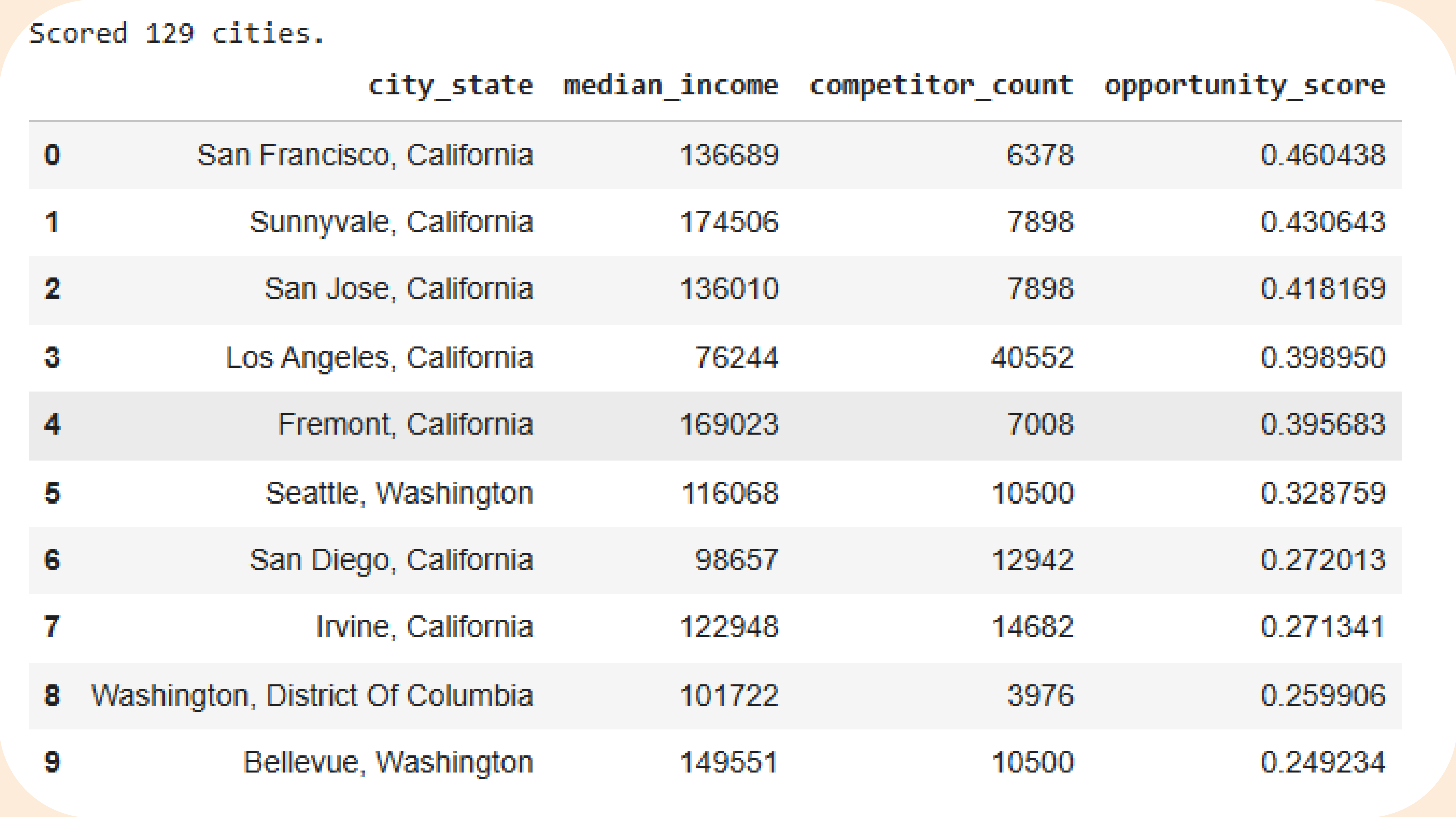
Competition Index Composition

- Café density per capita (70%)
- Average rating proxy (30%)

Interpretation

- High values indicate saturated, competitive markets

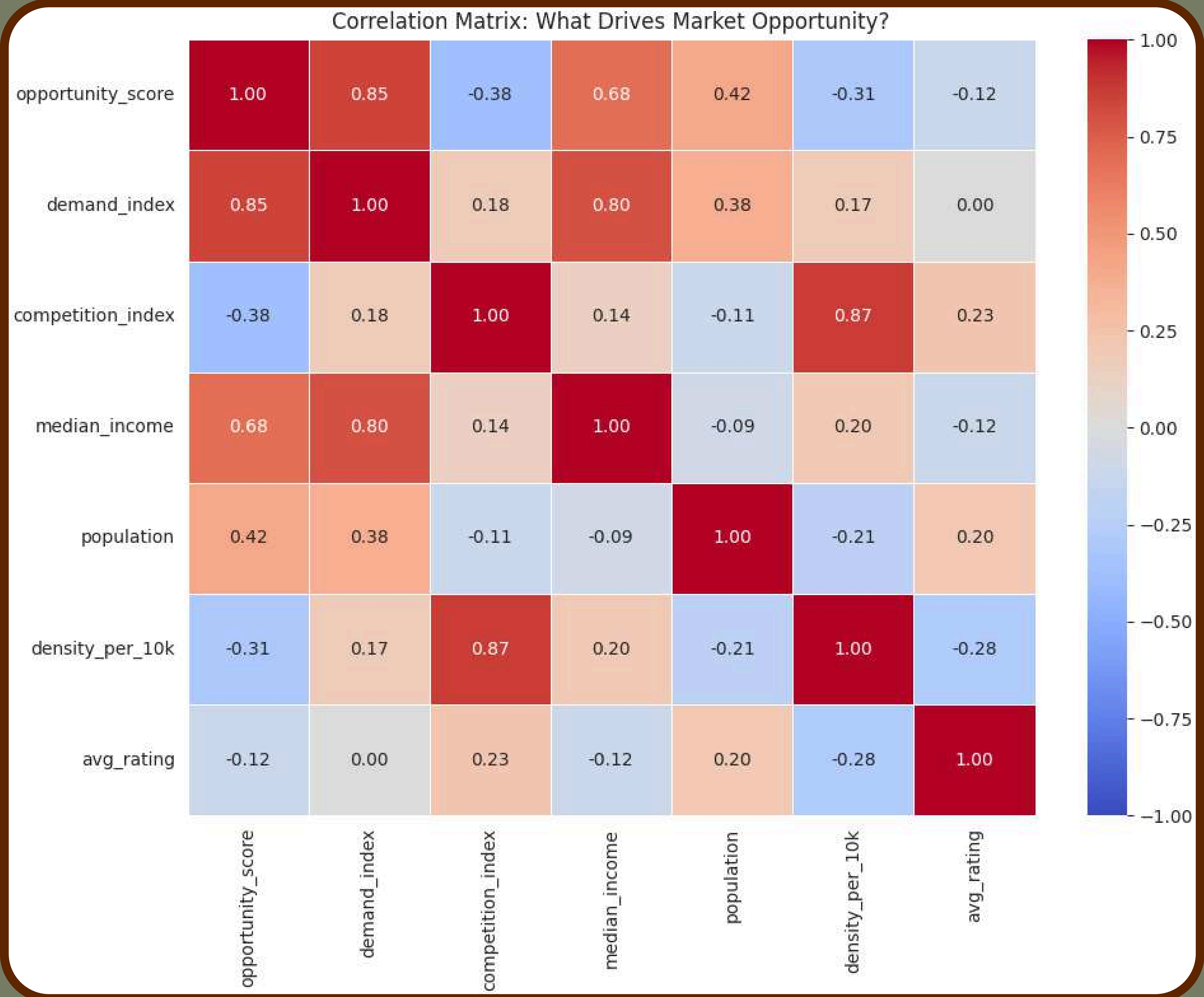
City Ranking Results



```
Scored 129 cities.
```

	city_state	median_income	competitor_count	opportunity_score
0	San Francisco, California	136689	6378	0.460438
1	Sunnyvale, California	174506	7898	0.430643
2	San Jose, California	136010	7898	0.418169
3	Los Angeles, California	76244	40552	0.398950
4	Fremont, California	169023	7008	0.395683
5	Seattle, Washington	116068	10500	0.328759
6	San Diego, California	98657	12942	0.272013
7	Irvine, California	122948	14682	0.271341
8	Washington, District Of Columbia	101722	3976	0.259906
9	Bellevue, Washington	149551	10500	0.249234

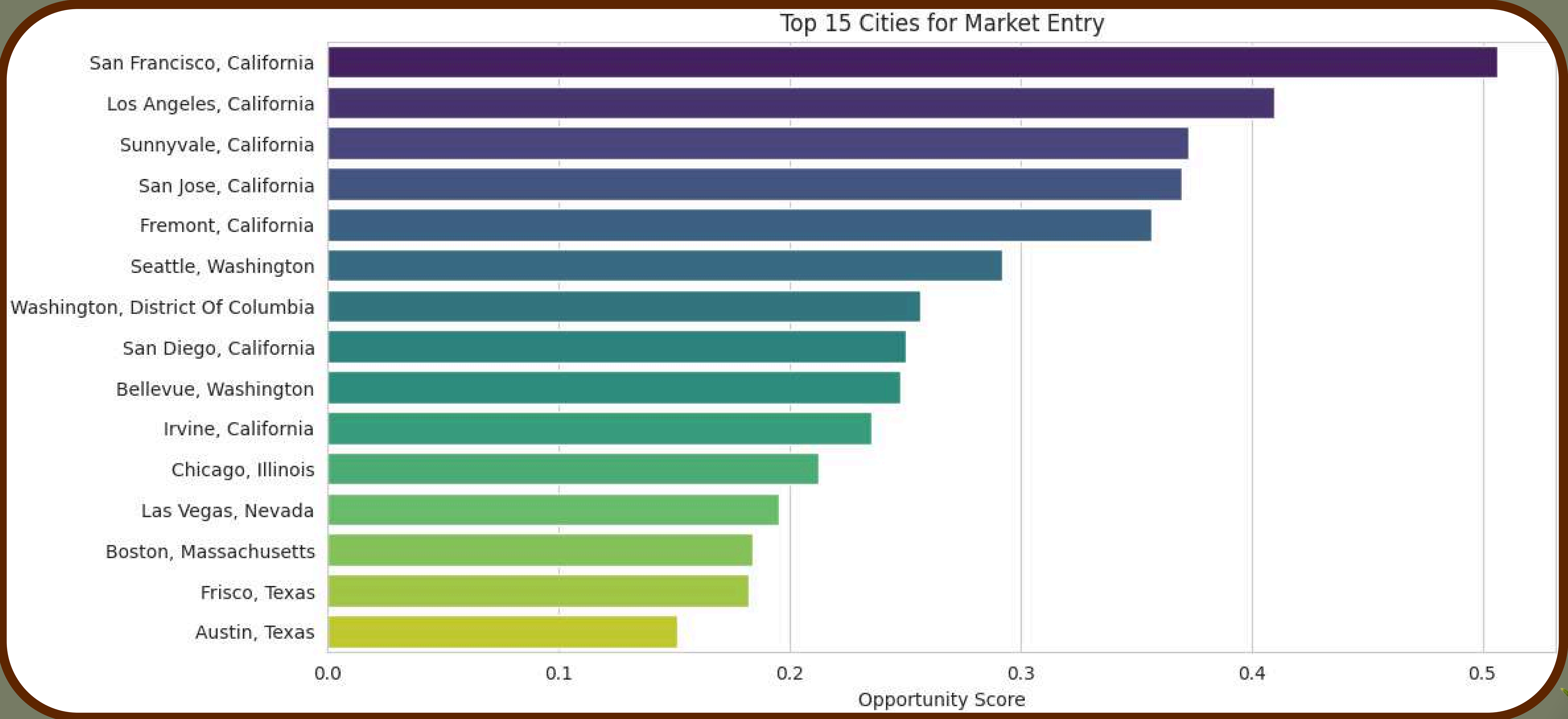
Correlation Heatmap



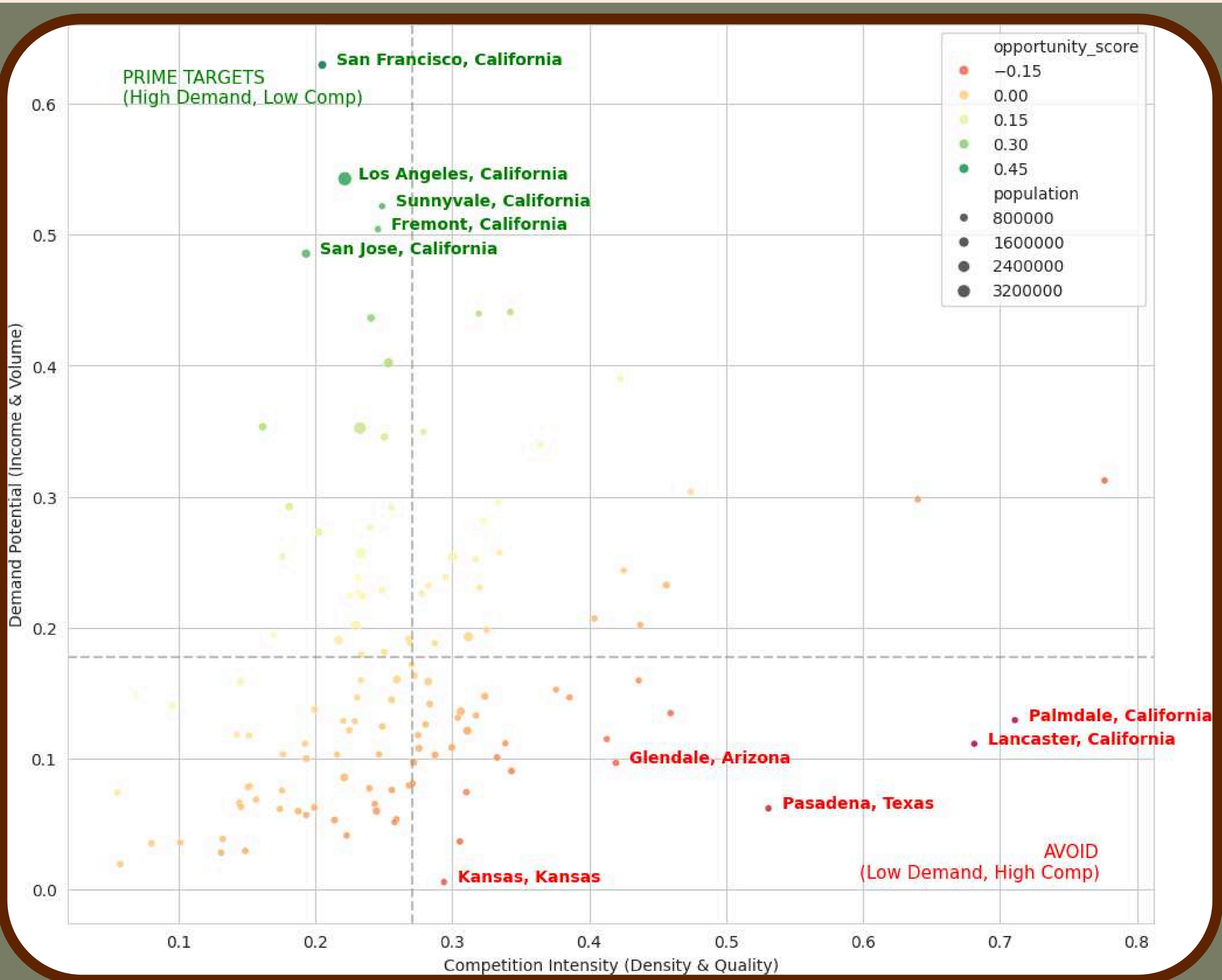
Correlation Matrix: What Drives Market Opportunity?

	opportunity_score	demand_index	competition_index	median_income	population	density_per_10k	avg_rating
opportunity_score	1.00	0.85	-0.38	0.68	0.42	-0.31	-0.12
demand_index	0.85	1.00	0.18	0.80	0.38	0.17	0.00
competition_index	-0.38	0.18	1.00	0.14	-0.11	0.87	0.23
median_income	0.68	0.80	0.14	1.00	-0.09	0.20	-0.12
population	0.42	0.38	-0.11	-0.09	1.00	-0.21	0.20
density_per_10k	-0.31	0.17	0.87	0.20	-0.21	1.00	-0.28
avg_rating	-0.12	0.00	0.23	-0.12	0.20	-0.28	1.00

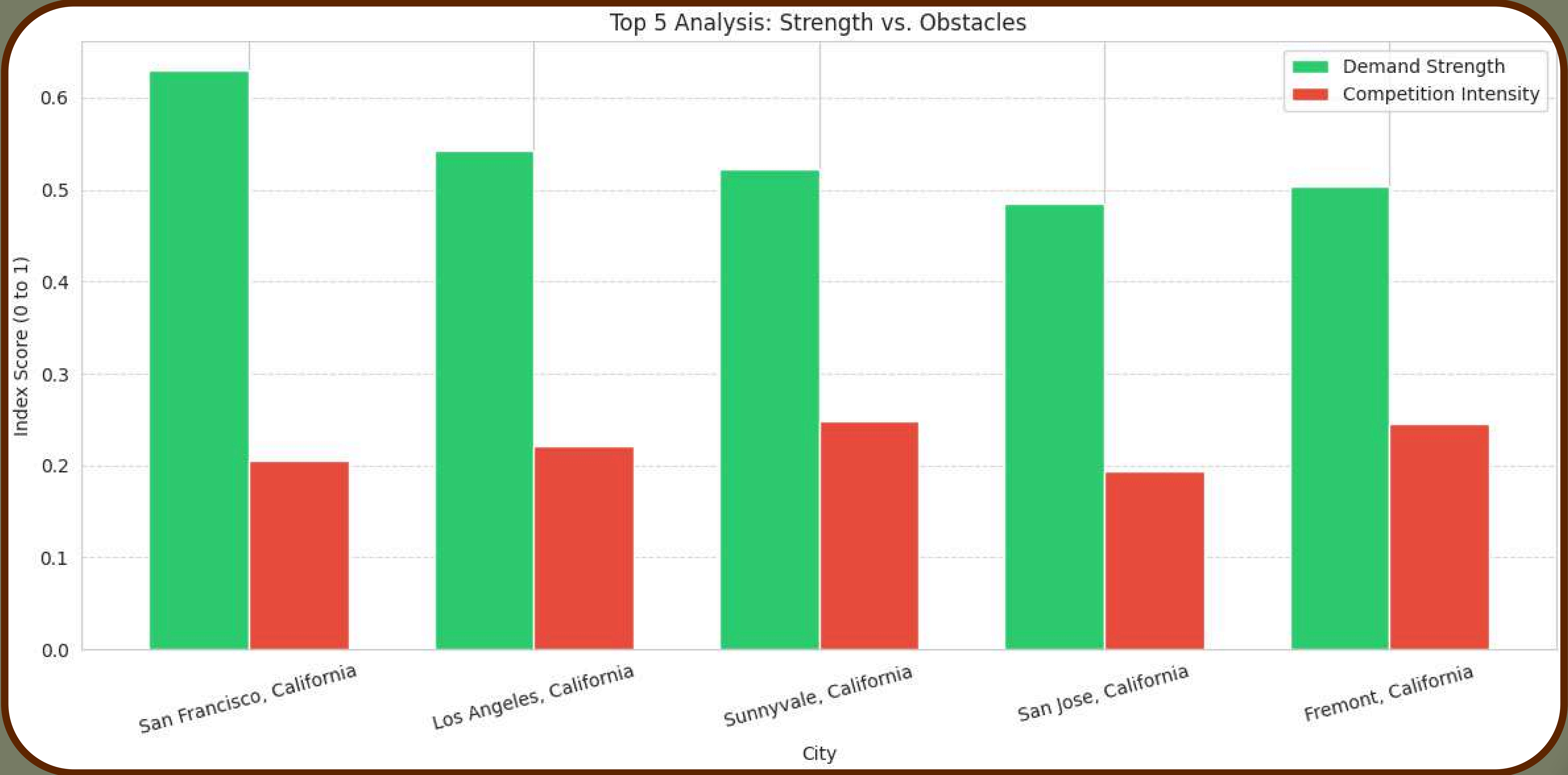
Top 15 Cities for Market Entry



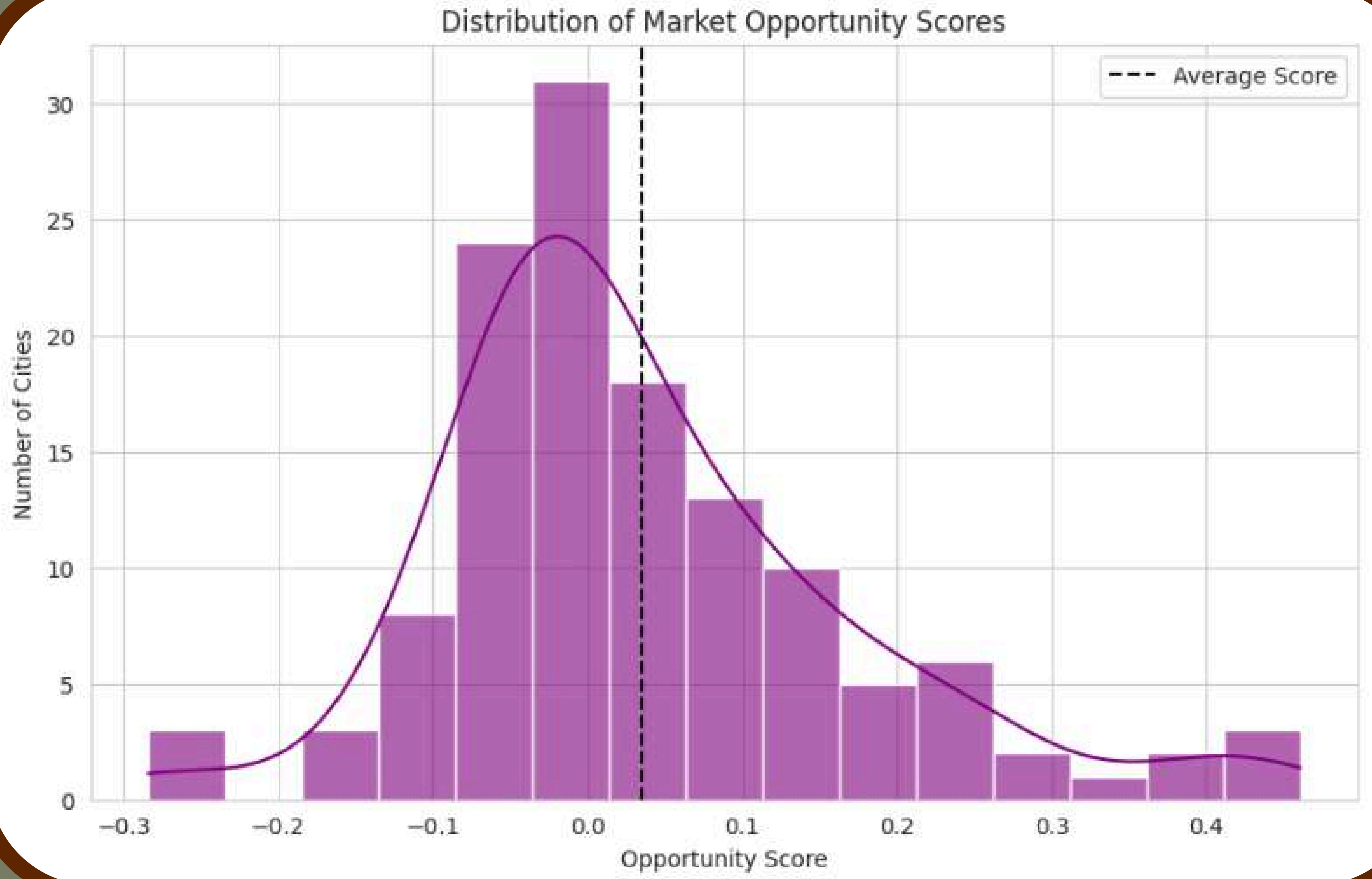
The Strategy Matrix



Demand vs. Competition for Top 5 Cities



Distribution of Oppurtunity

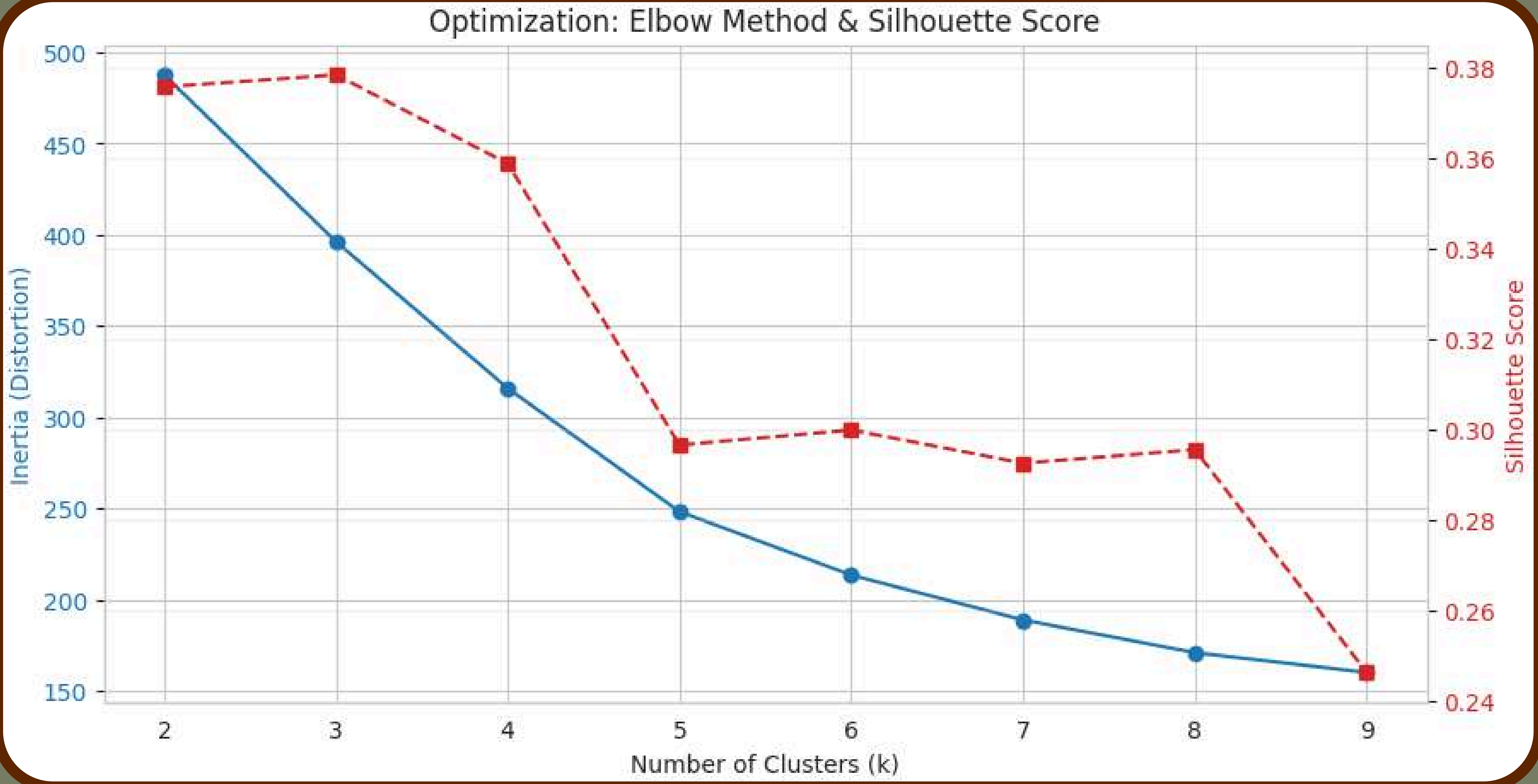


Unsupervised Market Segmentation with Machine Learning

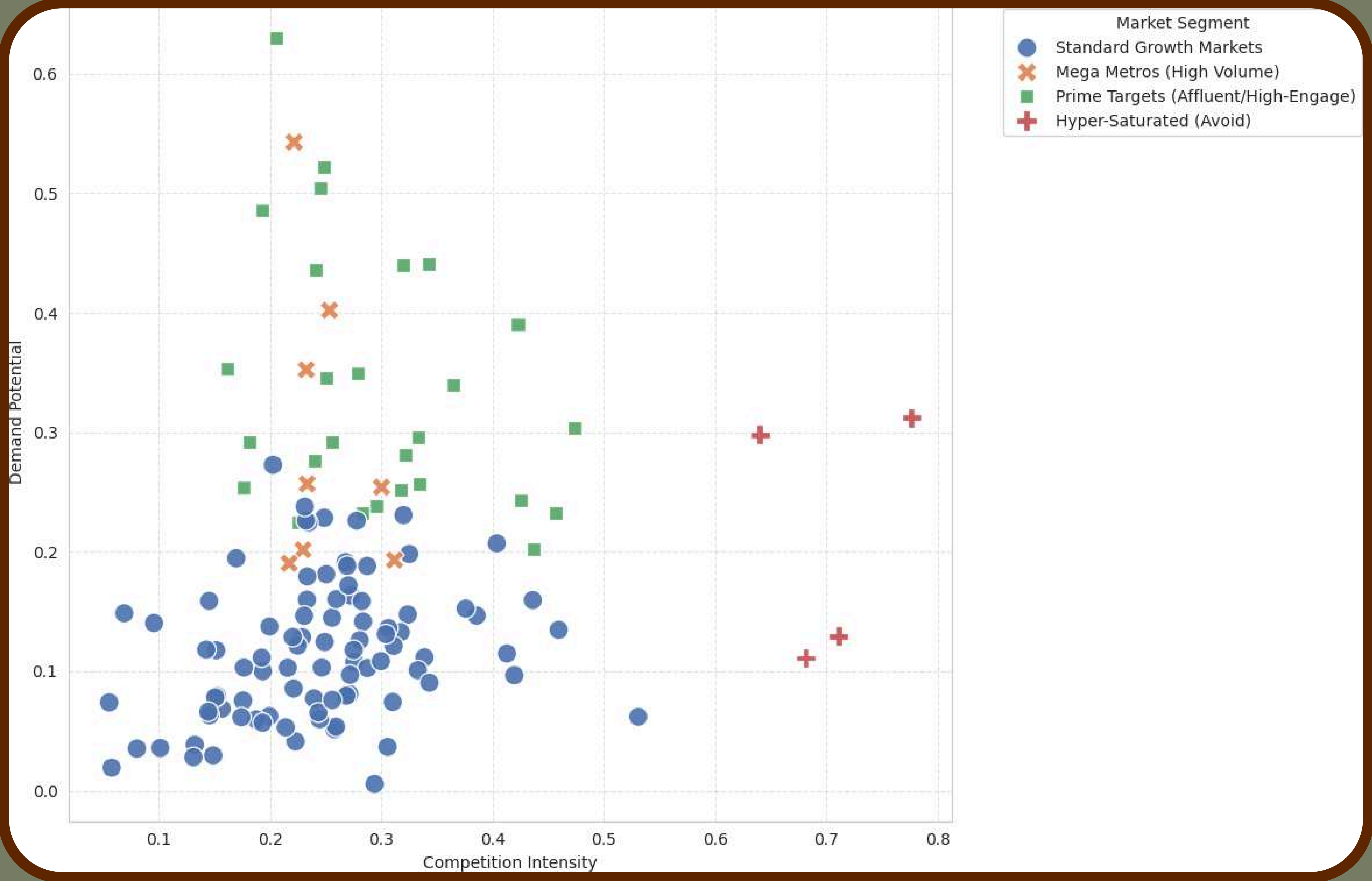
- Unsupervised
- K-Means

Methodology

- Features
- Preprocessing
- Optimal K: Elbow Method & Silhout Score


Optimization: Elbow Method & Silhouette Score
Inertia (Distortion)
500
450
400
350
300
250
200
150
Silhouette Score
0.38
0.36
0.34
0.32
0.30
0.28
0.26
0.24
2
3
4
5
6
7
8
9
Number of Clusters (k)

Machine Learning Segmentation: 4 Market Archetypes



Thank You For Listening

MUSA YÜKSEL
21050911018

SENA DİLAN ÇAKIR
21050911027