Capstone Project: Coffee Shop Locator

Amanda Khairunnisa

Problem Statement

When opening an offline retail store, one of the most important things of highest capital expense to consider is location. Choosing the right location can be detrimental towards the survival of the business.

This project aims to explore the use of machine learning in determining where to open a coffee shop in the San Francisco Bay Area.

Source: Flaticon

Methodology

Data Collection and Cleaning

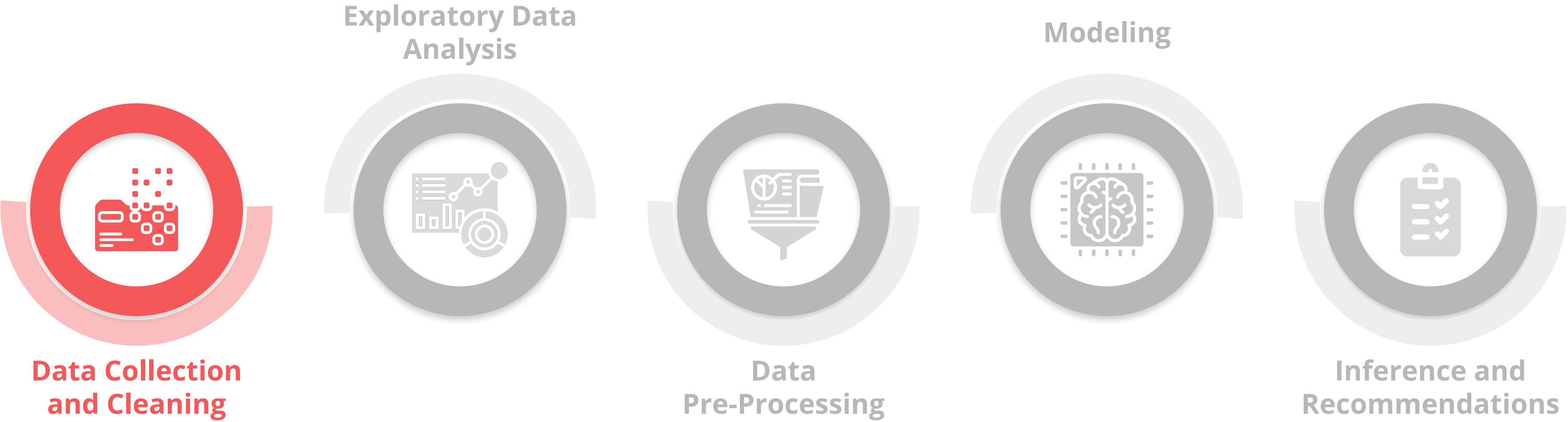
Exploratory Data Analysis

Data Pre-Processing

Modeling

Inference and Recommendations

Source: Flaticon



Data Collection and Cleaning

Exploratory Data Analysis

Data Pre-Processing

Modeling

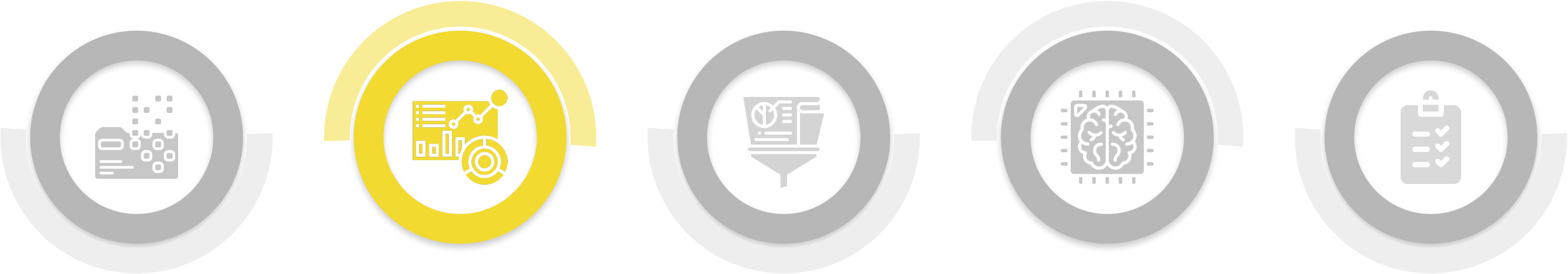
Inference and Recommendations

Source: Flaticon

Data Collection and Cleaning

- Demographics
- Recommended Venues + Category
- Geographic Data
- User Generated Data
    - Handling Null Values

Source: Flaticon



Data Collection and Cleaning

Exploratory Data Analysis

Data Pre-Processing

Modeling

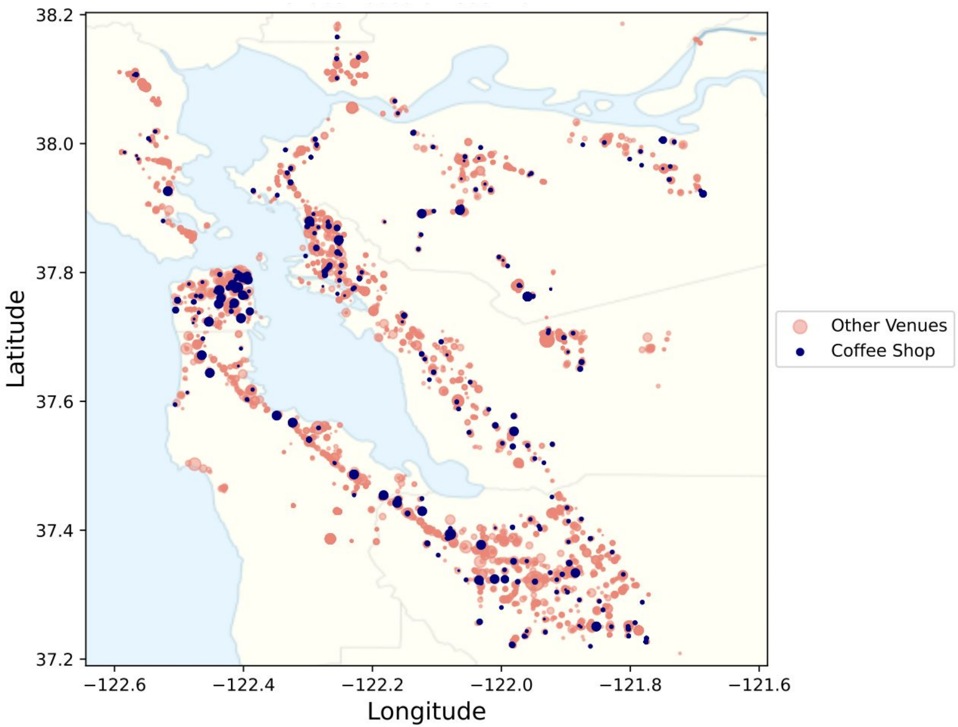
Inference and Recommendations

Source: Flaticon

Venues Collected

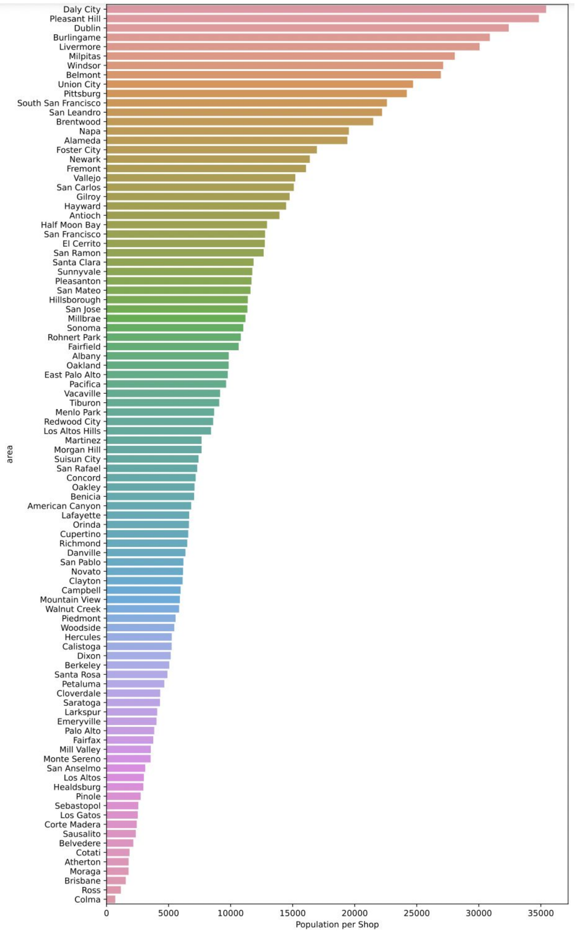
Weighted by User Ratings

Total Venues Collected: 14,939
Coffee Shops: 740



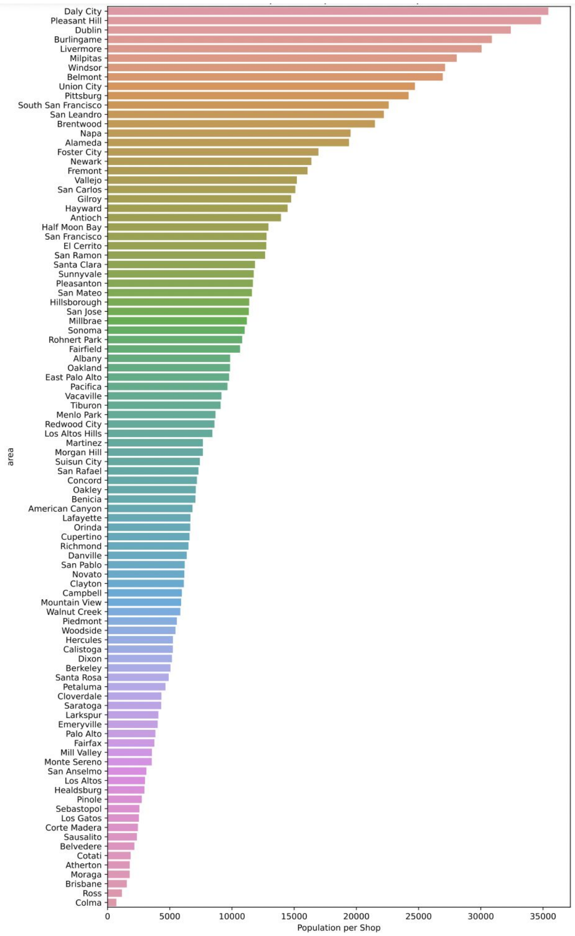
Source: Flaticon

Coffee Shop per Area Population



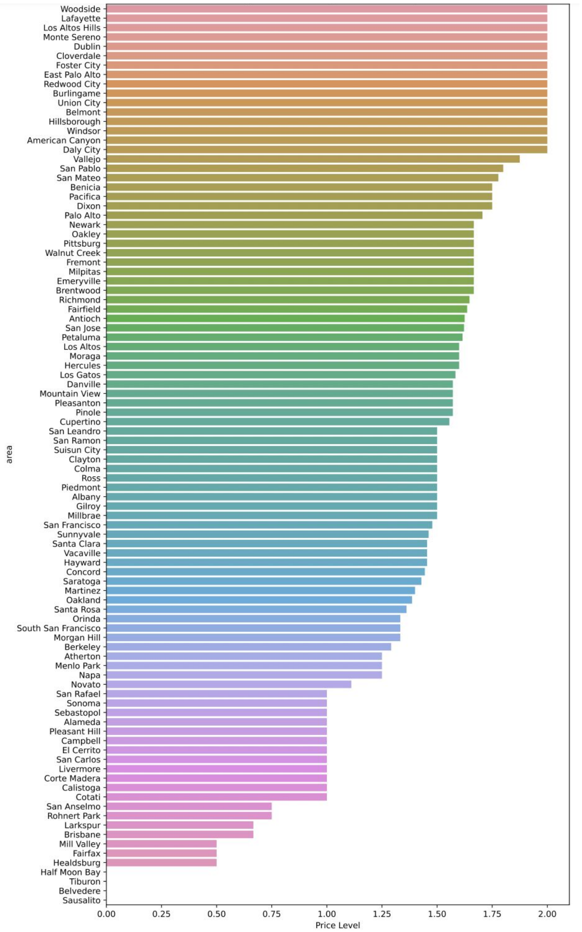
Source: Flaticon

Average Coffee Shop Ratings per Area

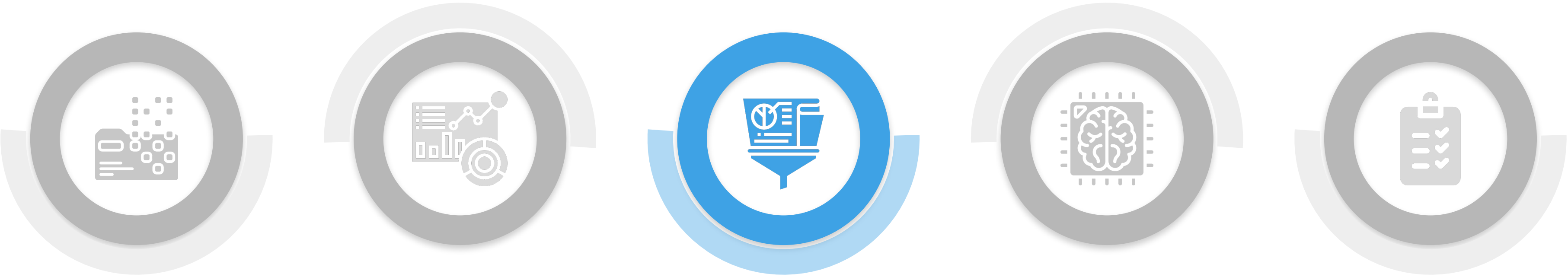


Source: Flaticon

Average Coffee Shop Price Level per Area



Source: Flaticon



Data Collection and Cleaning

Exploratory Data Analysis

Data Pre-Processing

Modeling

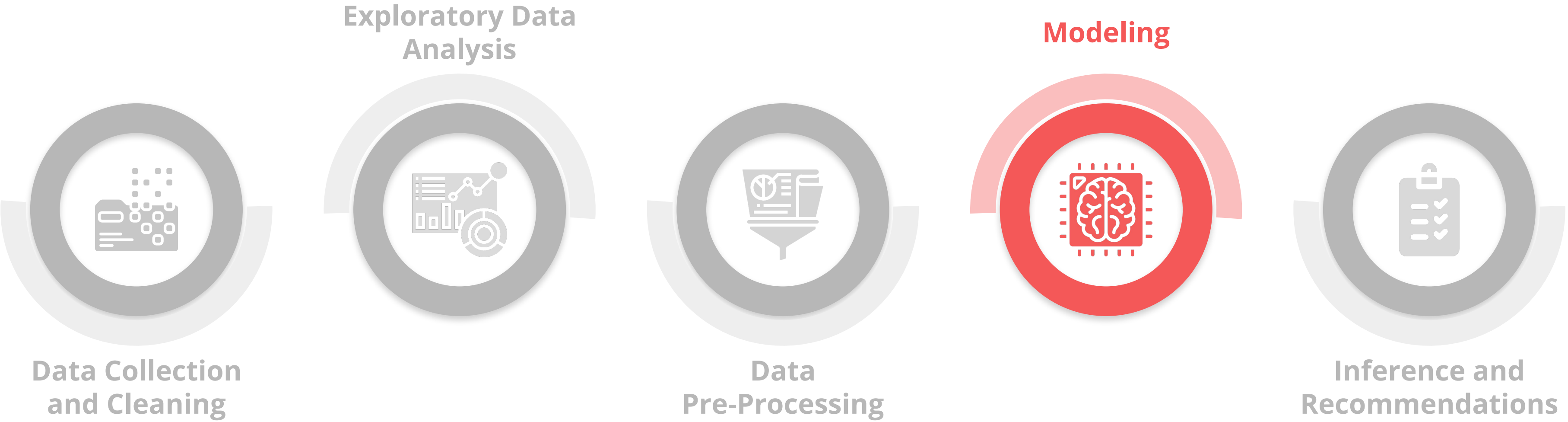
Inference and Recommendations

Source: Flaticon

Data Processing

- Distance Matrix
- Population and Density Metrics
- Venue Frequency per Category

Source: Flaticon



Data Collection and Cleaning

Exploratory Data Analysis

Data Pre-Processing

Modeling

Inference and Recommendations

Source: Flaticon

Classification Modeling: Key Performance Indicator

Baseline: 94.3%

Metrics	Logistic Regression	Random Forest Classifier	Extra Trees Classifier
Train Accuracy	89.3%	97.0%	93.6%
Test Accuracy	82.4%	94.6%	89.4%
Test Sensitivity	94.0%	99.3%	97.9%
Test Specificity	81.7%	94.3%	88.9%
Test ROC AUC Score	92.5%	99.7%	98.1%

Note: Imbalance data was dealt with using SMOTE

Source: Flaticon

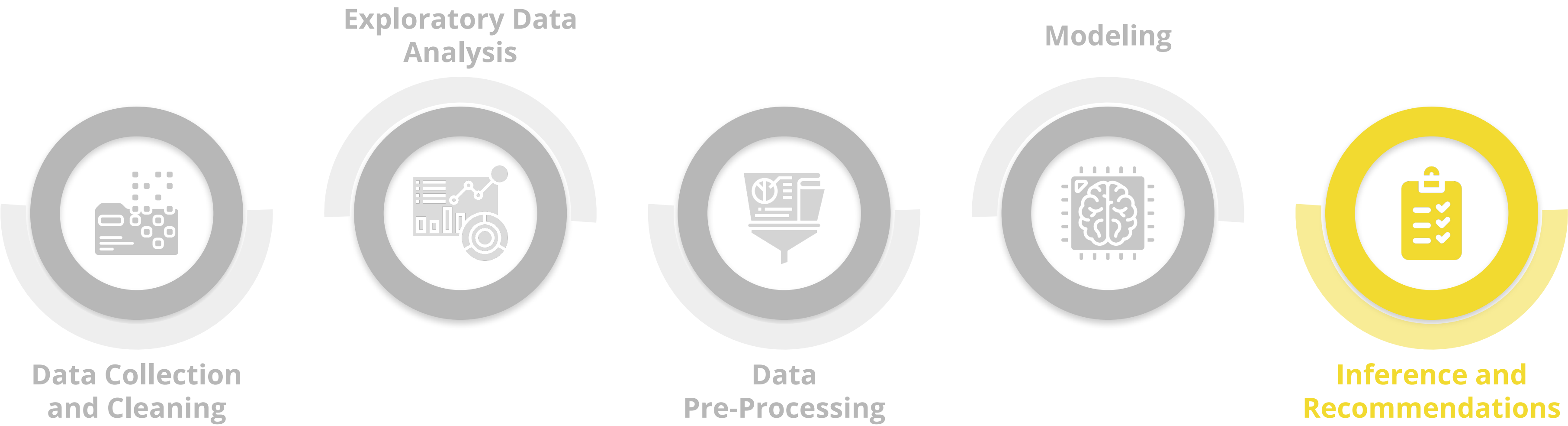
Classification Modeling: Key Performance Indicator

Baseline: 94.3%

Metrics	Logistic Regression	Random Forest Classifier	Extra Trees Classifier
Train Accuracy	89.3%	97.0%	93.6%
Test Accuracy	82.4%	94.6%	89.4%
Test Sensitivity	94.0%	99.3%	97.9%
Test Specificity	81.7%	94.3%	88.9%
Test ROC AUC Score	92.5%	99.7%	98.1%

Note: Imbalance data was dealt with using SMOTE

Source: Flaticon



Data Collection and Cleaning

Exploratory Data Analysis

Data Pre-Processing

Modeling

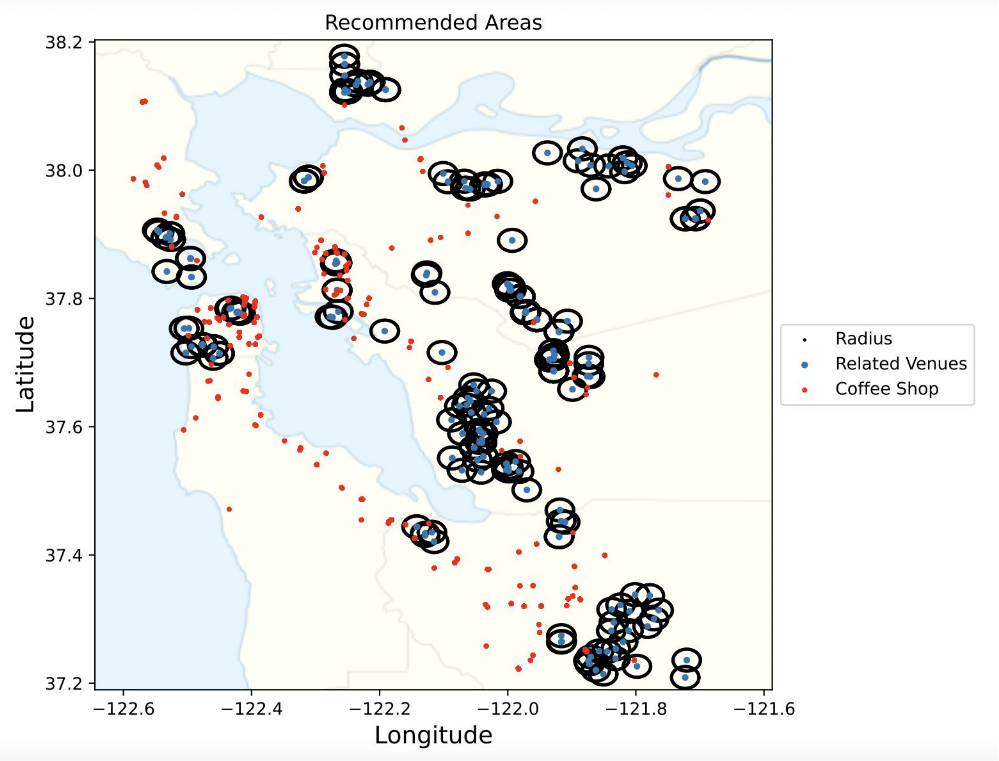
Inference and Recommendations

Source: Flaticon

Inference

Park, Pizza Place, Ice Cream Shop, Mexican Restaurant, Bakery

Areas Recommended: 265



Source: Flaticon

Recommendations

Understanding Limitations

- Simplification of real life
- Visualization limitations
- Data limitations

Recommendations

- Narrow down location using engineered / collected features
- Gather more data
- Model Application

Source: Flaticon