Popular Venue Modeling

How to predict popular venues for development based on city data

Michael Harreld
Jan 29, 2019

The problem

Investor

An investor is looking to develop new venues within a city. There are many choices of cities and venue types to consider. The investor has limited funds and wishes to choose the most profitable option.

Context

Venue data and city demographic data is widely available. Combining this data and analyzing relationships may enable popular venues to be predicted for a particular city as well as venue gaps.

Problem statement

How can we use city data to predict which venues could be most profitably developed within a city? How can we find cities with the most potential for new development?

Approach Overview

Bring in and clean data

- Data from wikipedia largest US city data
- Venue data from Foursquare

Visualize relationships

- Map of cities clustered by similar venues
- Venue popularity by population and density

Decision tree

Using a validated decision tree, predict popular venues given city data.

Solution

Prediction of successful venue categories for city development

Use a Decision tree model to predict what venue categories should be popular in a particular city. Compare to actual venues within the city to identify profitable development projects.

Methodology

Data Sources

- Wikipedia City Data
  - Downloaded directly from website
  - Must be cleaned and structured
  - Contains Population, Population Density, Latitude, Longitude
- Foursquare venue search API
  - Returns venues near to a location
  - Includes venue category such as Mexican Restaurant or Museum

Data Exploration

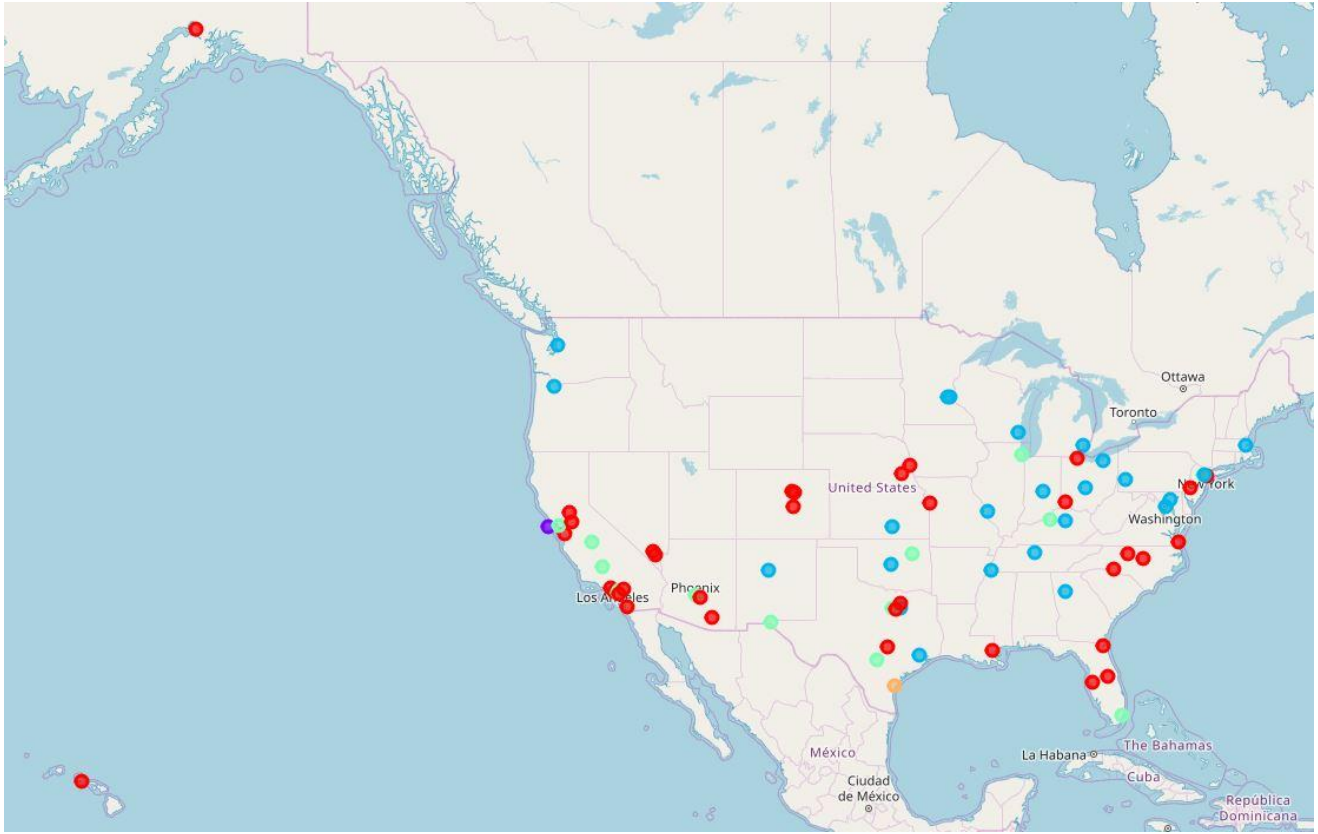
- Clustering Cities by Top Venues
  - Cities are clustered by similarity in top ten venue categories
  - Clusters displayed on map of United States
  - Visualizes geographic relationships in the data
- Venue Popularity by Population and Density
  - Categories are scatter plotted based on mean population and density
  - Suggests relationships between population, density, and venue popularity

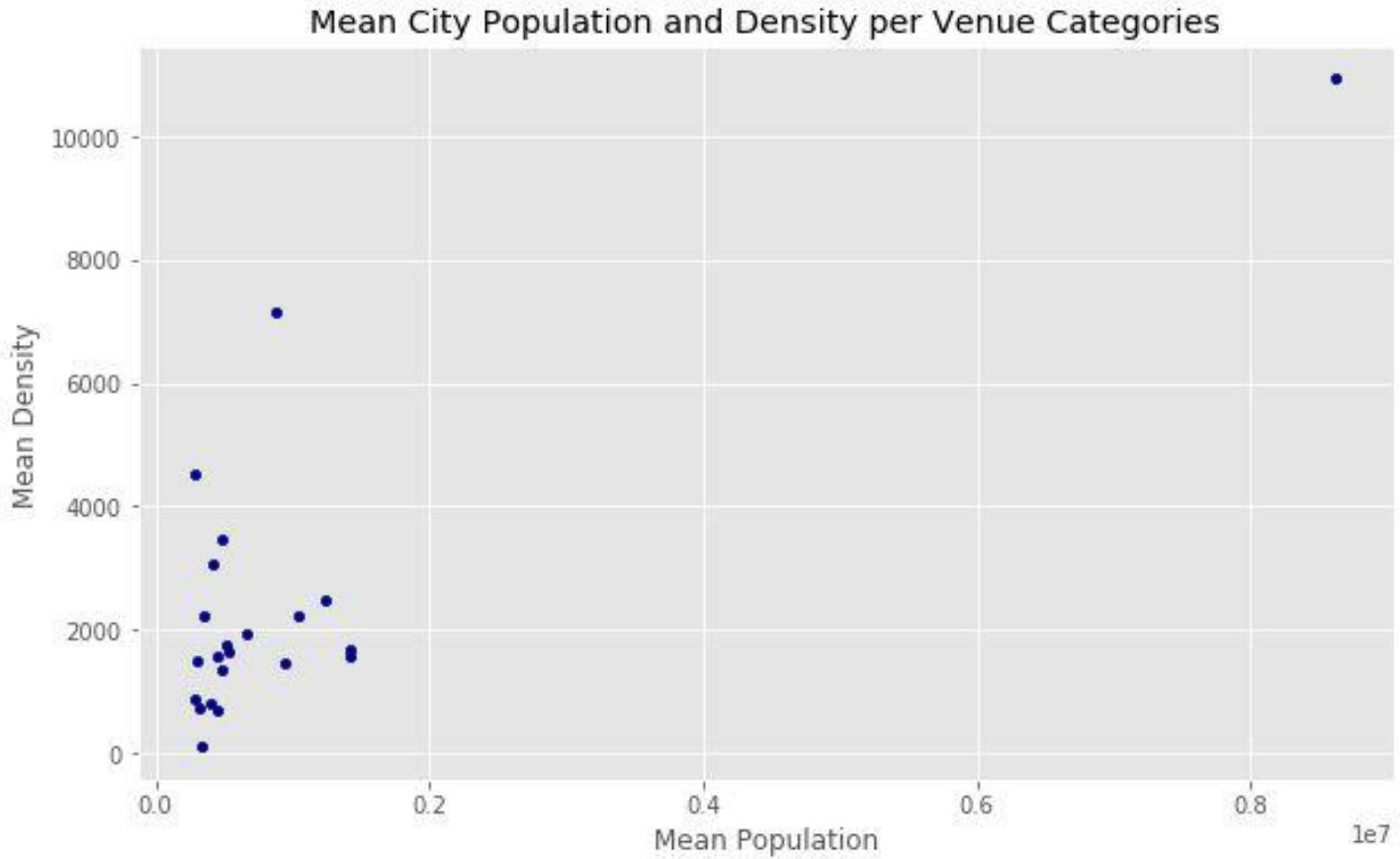
Data Modeling

- Training of decision tree model
  - Inputs - numeric population, population density, latitude, longitude
  - Output - categorical top venue
  - 80% of data used for training, remainder kept for validation
- Validation of model
  - 20% of held back city data used to predict likely venues
  - Predicted values compared to actuals

Results

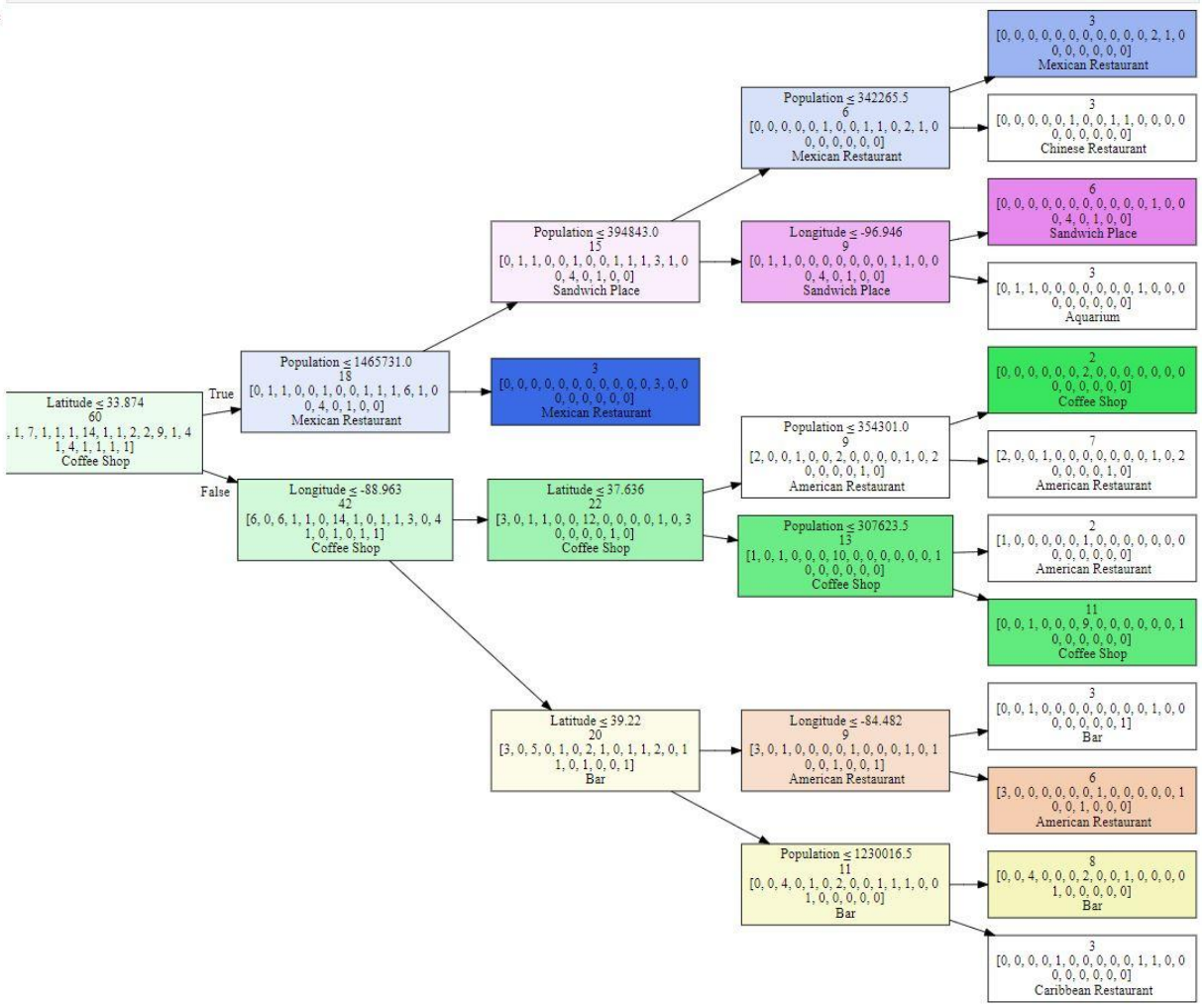
Clustering of similar venue cities




Mean City Population and Density per Venue Categories
Mean Density
10000
8000
6000
4000
2000
0
0.0
0.2
0.4
0.6
0.8
1e7
Mean Population

Decision tree predicting likely popular venue by testing:

- Population
- Population Density
- Latitude
- Longitude



Summary

- Decision tree modeling is a viable technique for predicting popular venue categories based on a city’s population, population density, latitude, and longitude
- Popular venue prediction is potentially a powerful tool to choose development venues and locations to maximize return on investment.