Predict Crowded San Francisco Areas

Dorris Wong

Objective

Predict what time certain areas of San Francisco will be crowded

Acquiring and Processing Data

- Cron job calling Foursquare API for trending places in SF
  - Every hour
  - Civic Center as the center with a 2 mile radius
  - 1 week
- Stored data in S3 and used EC2 to run the Cron job

Problems

1. Small geographical area
2. What is considered Foursquare's Trending algorithm?

Not very many places considered trending

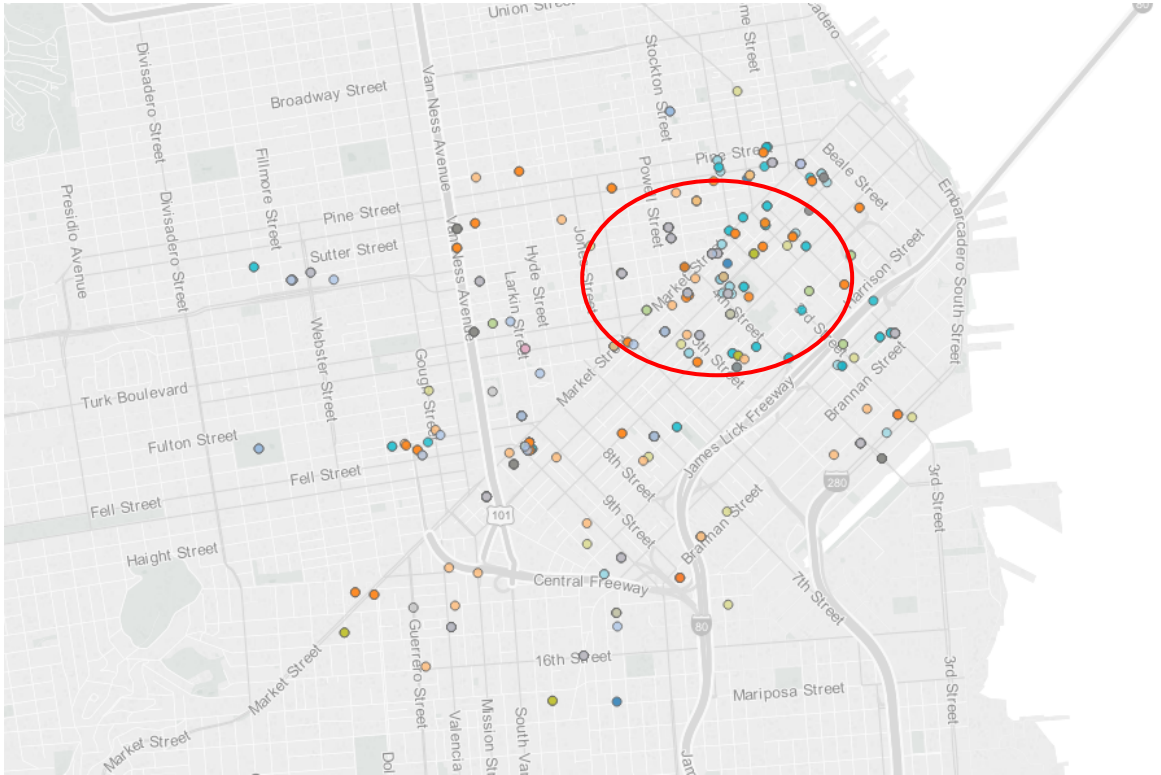
3. People checking into **silly** places like "San Francisco"
4. Do people regularly use Foursquare?
5. What type of people check-in and the type of places they like to check into

DATA EXPLORATION

Weekdays

Downtown is the most popular area, but geographically everything is spread out

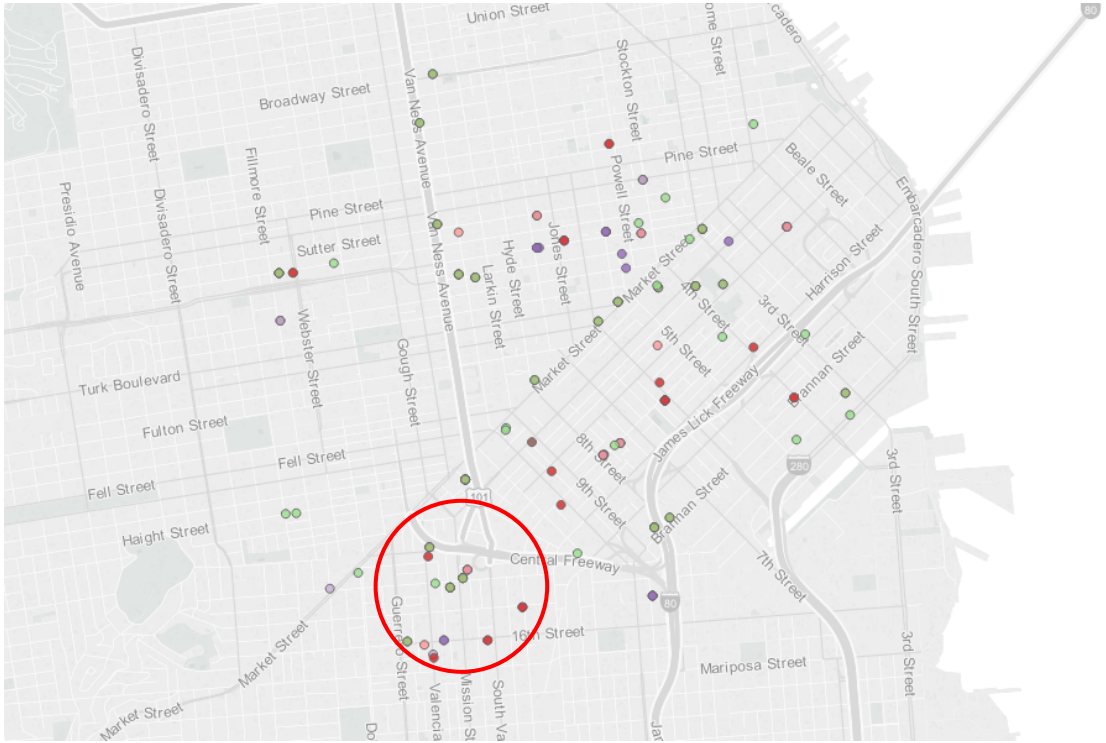
Assumption:
Offices and lunch places can be anywhere, but most offices are located downtown



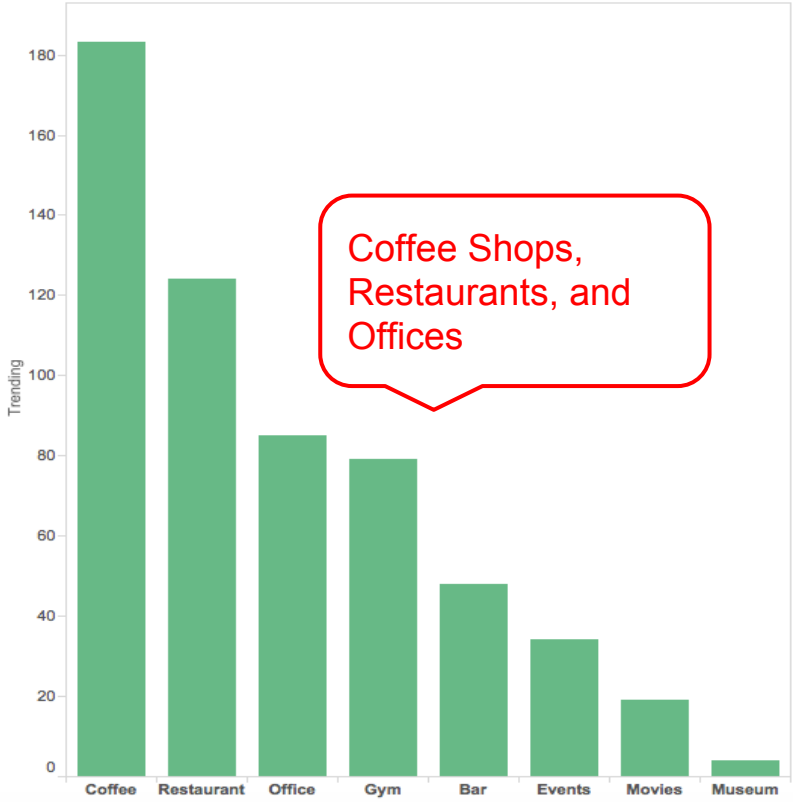
Weekend at Night

Mission is the most popular area

Assumption:
The popular bars and restaurants are located in the Mission



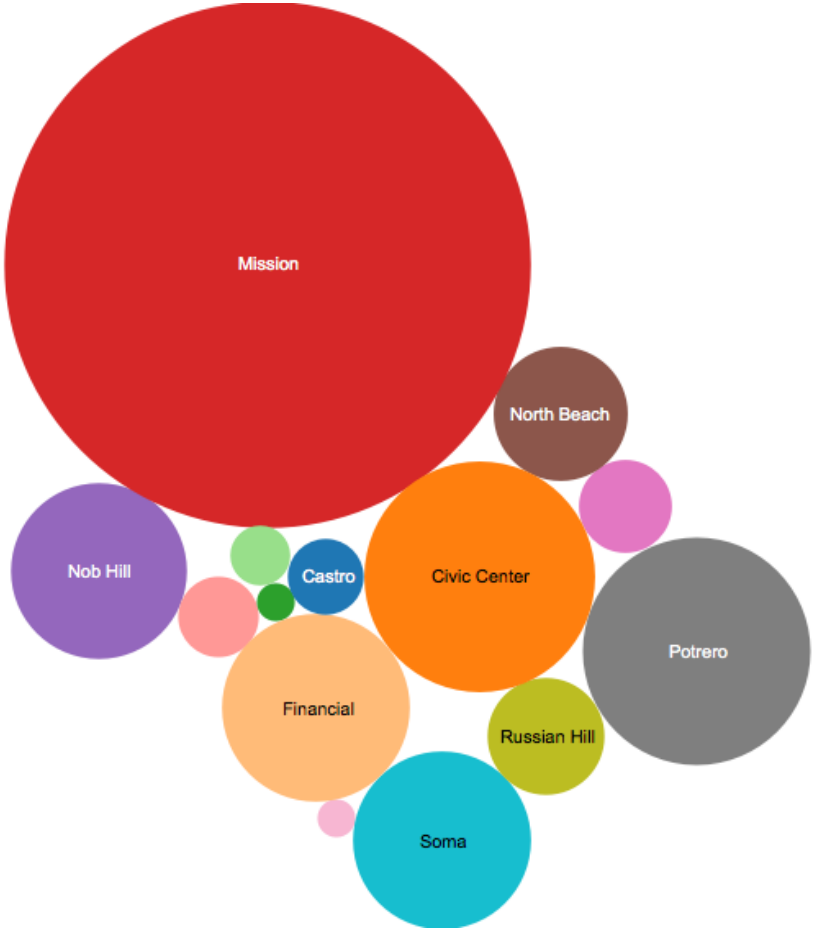
Weekdays



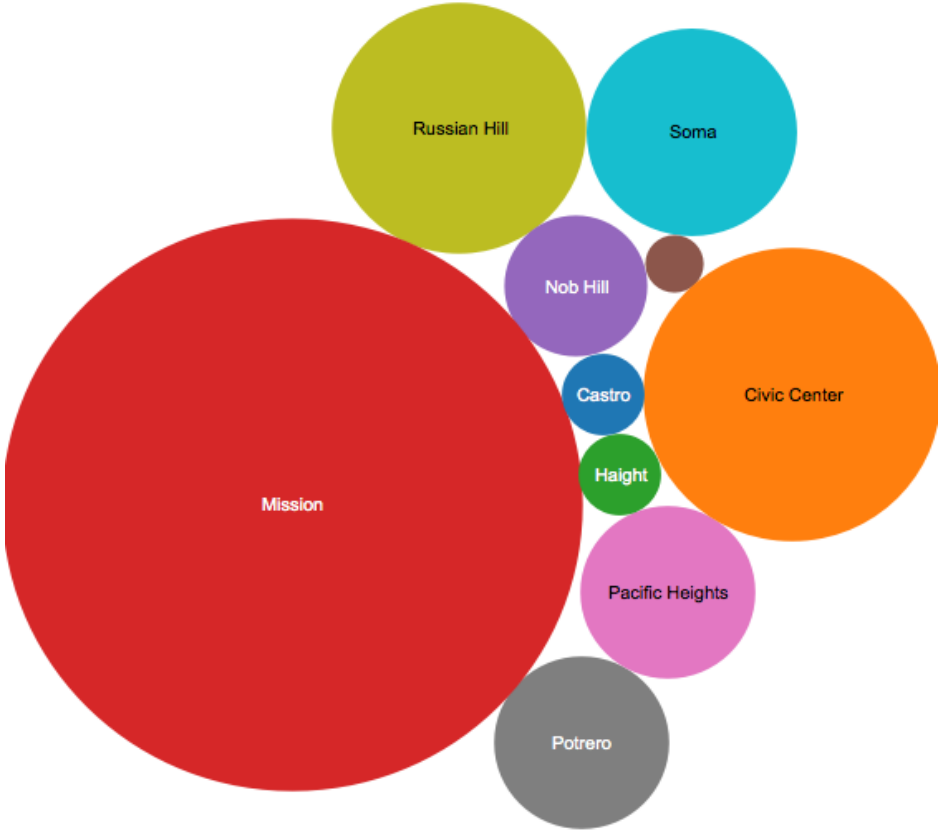
Weekends

Types of Establishments Considered Trendy

Weekdays



Weekends

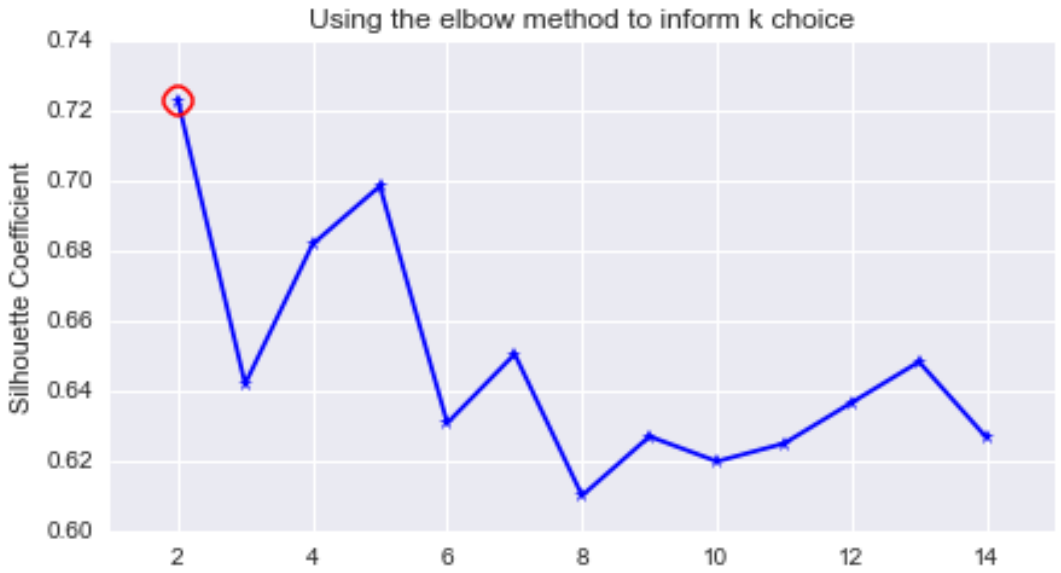


DATA SCIENCE

Classification Problem

- Supervised Classification
  - Decision Tree
  - Ensembling
  - SVM
- Unsupervised Classification
  - K-Means Clustering

Deciding K


Using the elbow method to inform k choice
Silhouette Coefficient
0.74
0.72
0.70
0.68
0.66
0.64
0.62
0.60
2
4
6
8
10
12
14

Features and Response

- Features: Time and Week
  - Dummy Variable: Combination of time and week
- Response: District

Results - Accuracy

- Clustering
  - Silhouette Score 72%
    - Decent separation and closeness
- Decision Tree (Tree Depth 2)
  - Accuracy Score 15%
- Ensembling
  - Accuracy Score 16%
- SVM
  - Accuracy Score 33%

Conclusion

- Most trendy places are the same and tend to be in certain clustered area (street with all bars or restaurants)
- Hard to predict with models and better to use business intelligence instead (possibly need more data)
- Mission is always popular

Next Steps

Use a wider radius range