Python Data Science Project Presentation

Battle of the Neighbourhoods

Author: Pieter Joan van Voorst Vader

Date: 17 Januari 2019

Agenda

- Introduction Business Case
- Data Sources and Usage
- Methodology
- Results
- Discussion
- Conclusion

Introduction Business Case

- Gain insight into *venue* and *crime-incident climates,*
  of the top 20 *education locations* in San Francisco.

- Intended audience :

  - Prospective students

  - Current security conscious students

- Who want to know interesting neighbourhoods with venues of their choice,
  and the crime-climate in those neighbourhoods

- The neighbourhood climates will be determined by
  semi-steered unsupervised clustering algorithms,
  in the domain of Data Science algorithms.

Data Sources and Usage

- Folium map rendering for OpenStreetMap data,
- FourSquare API for venue related data,
- https://data.sf.gov for crime-incident related data,
- http://www.city-data.com/city/San-Francisco-California.html for top 20 education location listing,
- Python reverse address to geolocation packages for, geolocation determination based on address only.http://www.city-data.com/city/San-Francisco-California.html

Methodology

- Python Anaconda Suite with SciKit-Learn, SciPy and Pandas for :
  - Data Cleaning and Selection,
  - Applying Machine Learning algorithms and evaluation metrics
  - Statistical Correlations

Results

- For usability reasons for the intended audience,
  only DayTime events are clustered.

- Applied Unsupervised Machine Learning Algorithms :

  - K-Means, in multiple applications:
    - For initial neighbourhood determination,
    - Neighbourhood climates
    - Concave Hull Polygon nearby points determination.

  - Mean Shift

  - Density-Based Spatial Cluster of Applications with Noise (DBSCAN)

  - Gaussian Mixture Modelling (GMM)

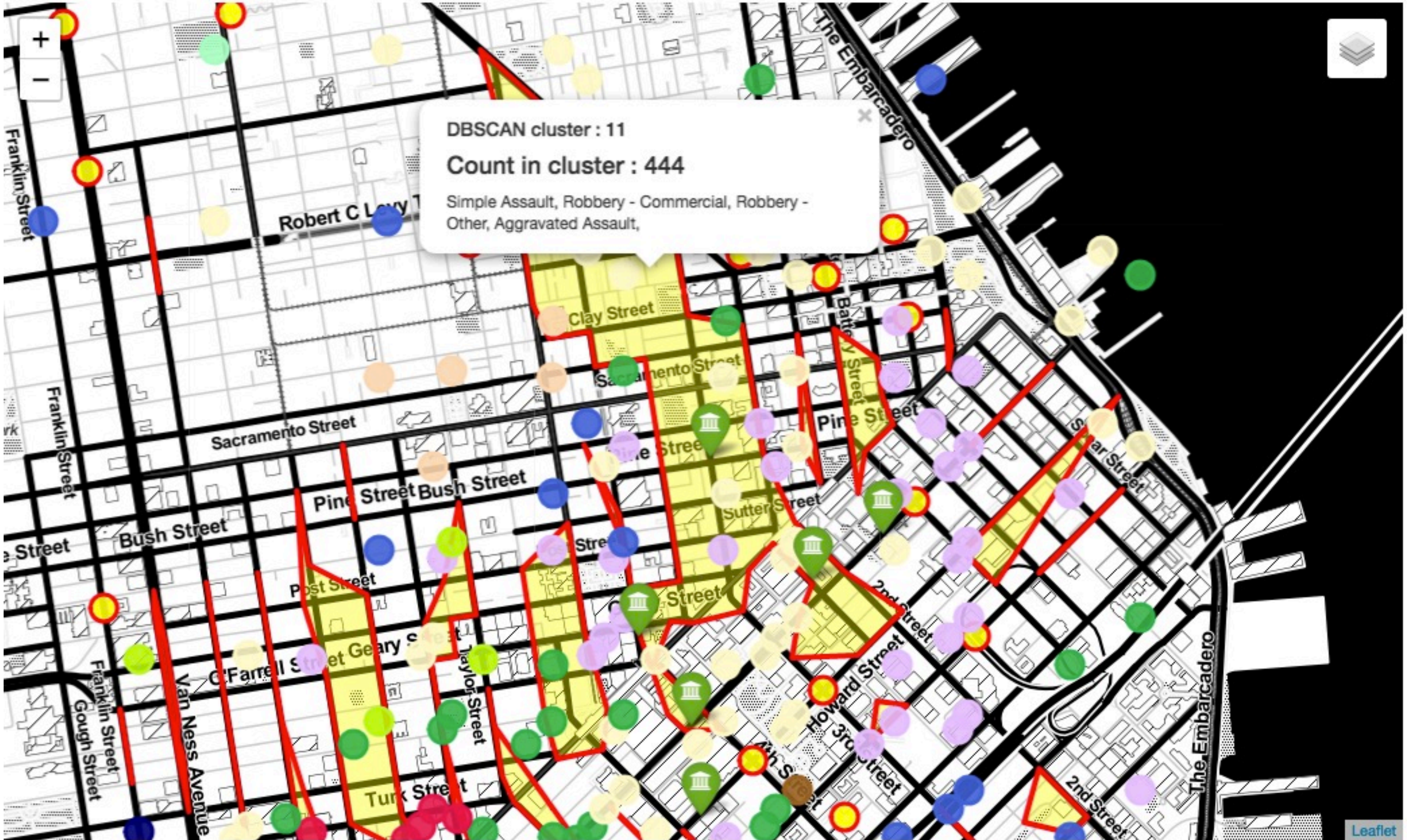
  - Agglomerative Hierarchical Clustering (AHC)

Results (continued)

- The most useable combination of algorithms after evaluation of the produced clusters :
  - Per education location, neighbourhoods with a 4x4 quadrant specified, within the Bounding Box of an approximate 2 km radius for *venues, with 20 defined cluster climates*
  - DBSCAN with 151 clusters, that minimised the *noise-count for crime related incidents, within the boundaries of the education location neighbourhoods.*
- Resulting in quite homogeneous clusters, automatically named, and plotted, with labelling, selectable per education location.

Results (continued)

- The resulting plot with active cluster selection.
- Crimes are with a red-outline and translucent yellow filling.
- Venue climates are coloured per cluster.
- Layer selection is available for exploration purposes.



Results (continued)

- Metric evaluation and Correlations
- The K-Means elbow method did provide insight into a right amount of clusters.
  - Cluster examination *tuned* the choice of K-clusters, while evaluating the homogeneity of the produced clusters
- For completeness reasons, a DBSCAN *Epsilon* and *minimum sample size per cluster* was chosen on the configuration with the least labelled *noise.*
- Automatic naming was applied with the names of the most occurring top4 venues/crime-categories in each clustered neighbourhood.
- No significant correlations were found after clustering, and exploratory data plotting efforts.

Discussion

- Multiple cluster represenations could be deemed suitable,
  with about the same cluster sizes.

- Suitable would be:
  - K-Means, in quadrant-pre-determined neighbourhoods,
  - K-Means in K-Means unsupervised determined neighbourhoods,
  - DBSCAN,
  - Agglomerative Hierarchical Clusters,
  - GMM Clusters

- Each algorithm has advantages and disadvantages.

Discussion

- Disadvantages K-Means:
  - Sperical cluster shapes,
  - Pre-Set Cluster amounts

- Disadvantages GMM:
  - Overlapping clusters can create user-confusion issues

- Disadvantages AHC:
  - Includes all crime-data points,
    and could provide a too intense experiences of the more gravely criminal incidents.
- *Advantage* AHC*:* irregular cluster shapes, that provide insightful details of crime neighbourhoods.

- Advantage of DBSCAN:
  - Ignores noise, and therefore creates a more realistic view of neighbourhood incidents that occurred during a *whole year period.*
  - Creates irregular cluster shapes, that provide insightful details in higher crime neighbourhoods.

Conclusion

- Due to the undeterministic nature of unsupervised learning,

  - metrics can only guide,
  - a *best* cluster determination algorithm is always arbitrary and should suit the project goal and intended audience,
  - a final product with the well working crime algorithms could be presented in one view

- Future Research:
  - Perform field survey amongst a corpus of potential users for cluster type selection
  - Explore K-Means in K-Means in further detail and compare with other methods
  - Apply Mean Shift clustering iteratively until *desired* amount of detail/clusters is reached
  - Improve automatic naming algorithm