**Galvanize Data science immersive capstone project**

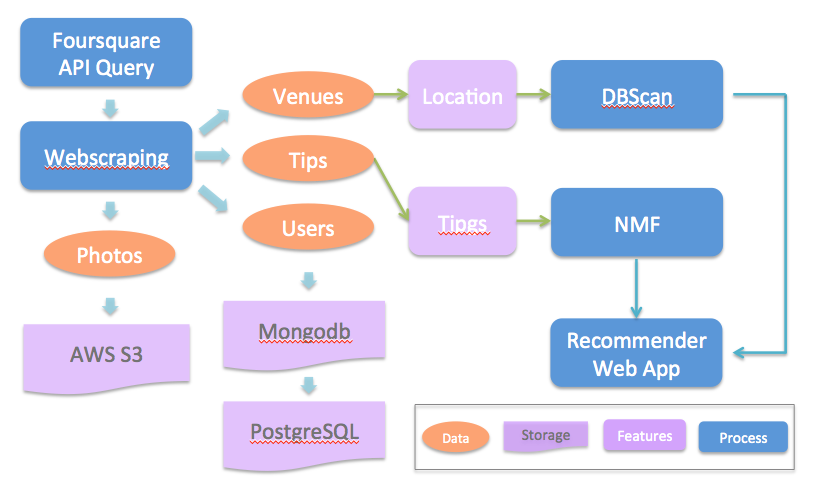
**Introduction & Data Source:**

The 2-week long capstone project was developed to fulfill the requirement of Galvanize data science immersive program. The object of this project was to explore city attractions in Tokyo and make recommendations for travellers, Data were obtained using [Foursquare](https://foursquare.com/) API and web scraping tools (Selenium package with python).

**Goal:**

Build a city attraction recommender system based on web-scraping data; Test is we can summarize tourists' intention toward an attraction using the review and photos they posted on phone/web app.

**Method:**

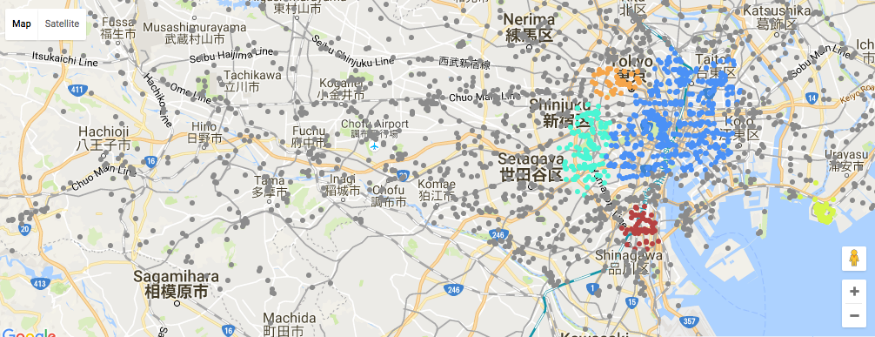


**Figure 1: Workflow**

1. Explore the data with mongodb query to understand the content, range, and quality of data.

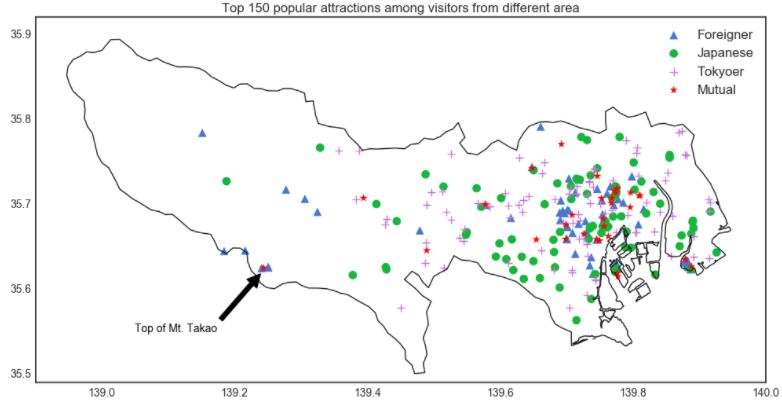
* terminology:
  + venues: city attractions;
  + tip: user review on the venue or suggestion made to other visitors;
* Venue statistics indicates older tips were removed from Foursquare's database
* Mixed language could be challenging in NLP processing

1. Filter out venues which are restaurants or train stations (except for some famous travel attractions)
2. Use DBScan algorithm to group locations with different "spatial density": downtown area has highest density and can be explore in short-term trip, and often with limited types of activities.

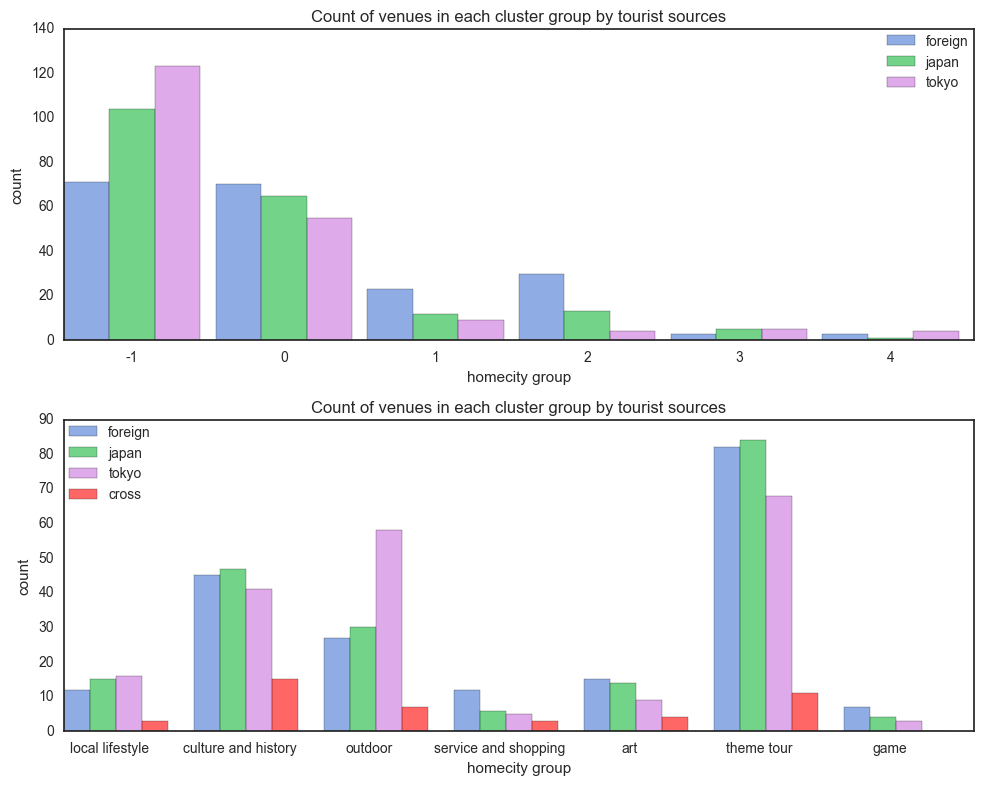


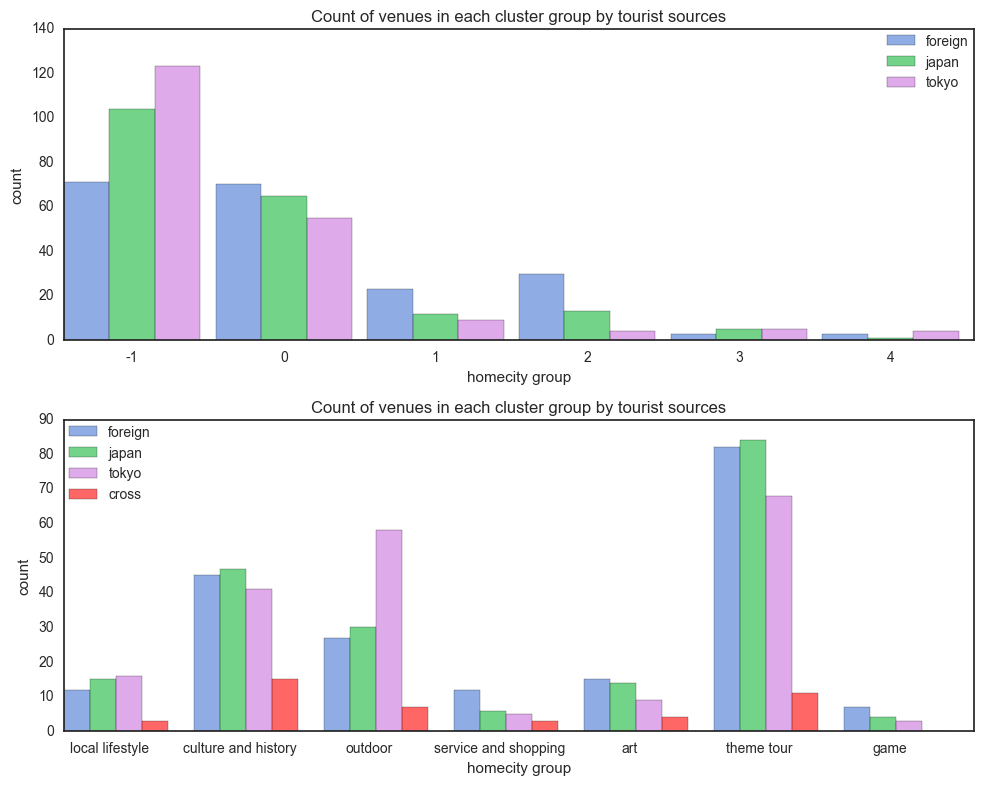
**Figure 2: DBScan result: downtown area has highest density of venues and mixed categories, except for large theme park (e.g. Disneyland, Sea Park, and etc.) and mountain trails**

1. A baseline model was built with features related to popularity (how many visitors, percentage "liked", percentage check-ins with photos and/or tips posted) of all venues.
2. Use google geocoding api and translate api to group user home city data to tokyoer, Japanese , and foreigner.



**Figure 3: Location of top 200 popular venues of each tourist source groups**





**Figure 4: Counts of top 200 popular venues of each tourist source groups**

1. Comparison among reviews from different visitor sources were performed to refine the recommendation:
2. popular keywords were all translated to English,
3. For each venue, the keywords were formatted into a vector with length equal to total counts of keywords in the entire dataset. If a keyword appears in a venue's document, the value was assigned to one; otherwise 0
4. Non-negative matrix factorization was applied on the aforementioned keywords vectors to group the venues to 5 different tops
5. Manually adjusted topics of the venues.
6. A pseudo user-item evaluation score matrix was generated with each visitor's action on each venue's page (namely "likes", "photos posted", and "tips posted"), and use graph lab's factorization recommender model to make a list of recommended venues.
7. Weighted the score predicted by the aforementioned model by the topics and homecity group, and return the top few venues.
8. UI development: Let user pick photos they are interested and also their homecity group, plot recommendation attractions on map.

**Limitations:**

* Used a small portion of the text data due to translation cost and time limit;
* Tested tensor flow on Spark EMR for image content tagging, the result was not relevant to this recommender system, but could be a side project afterward.
* The project relied on geo-tagged data of foursquare website, which is not a popular app in the most countries. Therefore, Tokyo was chosen as my target area.

**Discussion:**

* The method applied was able to distinguish (to a certain extent) preferences of different groups (local, visitors from other areas in Japan, and foreign travelers).
* My recommender system product of this project will include only the top 200 venues of each visitor source group (sum up to 442 venues) as a toy example that can be deployed on a small amazon instance. The framework can be extended when more data available, and he business features and A/B testing evaluation can be added.
* The NMF analysis indicates visitors to all the venues tend to mention some food, which also indicates that food is an important element that shared among all city attractions! Restaurant recommender is not the topic of this project, but I am expecting to see interesting patterns among different tourist sources in Tokyo.