Jalen Wang

GitHub Repository: https://github.com/wangjalen7/CS4501-Locations-Project2

Data Sources and Hypotheses:

This report examines my Google Maps location data over a one-month period, from mid-February to late March, with the goal of inferring the significant locations in my life. The analysis relies on two primary data sources: the Google Takeout Map Timeline data and the Foursquare Places API. The Google Takeout export, provided as a JSON file, offers a comprehensive record of my location history, including timestamps, GPS coordinates, and activity types. This detailed dataset serves as the foundation for identifying clusters of repeated visits, which are then enriched with contextual information from Foursquare, such as venue names, addresses, and category labels.

Before analyzing the data, I hypothesized that the GPS data will naturally form distinct clusters corresponding to frequently visited locations, and that clusters with higher visit frequencies and longer dwell times will represent places of greater personal significance, such as my home, classroom, or favorite restaurants. Furthermore, by cross-referencing these clusters with Foursquare data, it is expected that the auto-assigned labels will accurately reflect the true nature of these locations.

Methods:

The analysis began with data cleaning and filtering. From the raw Google Takeout Map Timeline JSON, only those entries containing a “visit” field were retained, as these entries indicate significant stops rather than transient movements. The raw data was parsed to extract GPS coordinates and timestamps, and any noisy or irrelevant data (such as quick transit events) was discarded. This preprocessing ensured that only meaningful location data remained.

Once the data was cleaned, a clustering algorithm was applied to group together nearby location points. Specifically, the DBSCAN algorithm was employed using a haversine distance metric to account for the curvature of the Earth, ensuring that spatial distances were measured accurately in meters. DBSCAN was chosen because it does not require a predefined number of clusters and is effective at identifying clusters of arbitrary shapes while filtering out noise. The clustering process produced several clusters, each representing a significant location, characterized by a centroid, total visit duration, and visit frequency. These clusters provided a preliminary mapping of the places most frequently visited.

To enrich these clusters with semantic context, the Foursquare Places API was used. For each cluster’s centroid, the API was used to obtain nearby venue candidates along with details such as the venue’s name, address, and categories. The candidate with the smallest geographic distance to the cluster centroid was selected as the representative for that cluster. In parallel, manual mapping was also performed to assign known location names to clusters, enabling a direct comparison between the automated Foursquare results and ground truth based on personal knowledge. Finally, the results were visualized through bar plots showing both the frequency of clusters by auto-assigned labels and the total number of visits per location. These visualizations provided clear insights into the distribution and relative significance of the various locations captured in the data.

Jalen Wang

Results:

I identified 12 significant clusters from my Google Maps location data using DBSCAN, and the results show a detailed picture of the places I frequent over the past month. The clusters represent key areas in my life—from where I live and study to where I socialize and even travel. Notably, one cluster, manually labeled as my college apartment at 1707 Jefferson Park Avenue, stands out with 47 visits and over 570 cumulative hours. This cluster confirms my expectation that my college apartment is the main part of my daily routine, as I spend the majority of my time there. In contrast, another "Home" cluster, which I manually mapped to my family home in Richmond, shows 18 visits and nearly 160 hours of dwell time, reflecting my occasional visits back to my hometown.

My academic life also emerges prominently from the data. Clusters labeled as Thornton Hall and Rice Hall indicate that I spend a significant portion of my time on campus, where I attend classes. Although Foursquare's automated labeling sometimes misclassifies these clusters under a generic "Other" label—likely because it returns broader departmental names—the manual labels clearly identify these buildings as central to my academic routine. The moderate number of visits and durations here align with my daily schedule on campus, reinforcing the idea that these spaces are critical for my education and professional development.

Social and leisure activities also feature in the analysis. For instance, Cluster 2, manually labeled as Cook Out, and Cluster 10, labeled as McDonald's, each exhibit brief yet frequent visits. Both locations are correctly identified by Foursquare as restaurants, which corresponds well with grabbing quick meals. Meanwhile, Cluster 3, which I label as The Flats, shows 9 visits and a total duration of about 35 hours. This longer duration at The Flats makes sense as it reflects my habit of hanging out with friends in their apartment complex.

Recreational and retail spaces also emerge in the data. For example, Cluster 4, manually labeled as Harris Teeter / Barracks Road Shopping Center, and Cluster 8, labeled as Short Pump Town Center, capture locations where I run errands or enjoy occasional shopping trips. Foursquare's results for these clusters generally align with the manual mapping, even if some candidate details—like the exact store name—differ slightly. Similarly, Cluster 9, which I identified as Pouncey Tract Park Pickleball Courts, is correctly matched by Foursquare, reflecting my regular visits to play pickleball near my home in Richmond.

Travel and vacations are also evident in the clustering. Cluster 11, manually labeled as The Bellagio in Las Vegas, shows 8 visits and a cumulative duration of nearly 49 hours. While this might seem high for a typical location, the extended duration is explained by my spring break vacation in Las Vegas, during which I spent a considerable amount of time there. This travel-related cluster stands apart from my daily routines, yet it provides valuable insight into how vacation patterns manifest in my location data.

Overall, comparing manual labels with Foursquare's nearest-candidate approach reveals strong alignment in certain clusters (e.g., Cook Out, McDonald's, Short Pump Town Center) but differences in others where Foursquare defaults to a broader or slightly off-target place name. The largest clusters in terms of duration—Clusters 0 and 7—are each identified as "Home," which underscores the intuitive correlation between time spent there and personal significance. Mid-

range clusters (e.g., Rice Hall, Thornton Hall, Aquatic Fitness Center) correspond to campus locations, while smaller clusters with fewer visits or shorter durations typically represent quick errands, recreational stops, or fast-food restaurants. From a frequency standpoint, Cluster 0 (47 visits) and Cluster 7 (18 visits) confirm that residences dominate the user's daily life. In total duration, Cluster 0 (572.40 hours) and Cluster 7 (159.62 hours) dwarf all others, reflecting the essential role of "home" in daily routines. For smaller clusters such as Cook Out or McDonald's, the visit count is modest, and total hours remain low, consistent with brief stops. The Foursquare bar charts highlight the distribution of auto-assigned labels, showing categories such as "Restaurant," "Shopping," "Gym," "Park," "Home," and "Other." While these categories capture general usage patterns, the manual mapping helps refine the understanding of each location, especially for academic or residential buildings that Foursquare might mislabel.

In summary, the results demonstrate that DBSCAN successfully identified meaningful clusters, each corresponding to a significant place in my life, and that a combination of manual knowledge and Foursquare data can provide a well-rounded perspective on the nature of each cluster. Large durations and high visit frequencies consistently indicate residential locations, mid-range values capture academic or recreational spots, and lower durations reflect quick errands or meal stops. This analysis offers clear evidence that location data, enriched by external APIs and personal knowledge, can yield valuable insights into an individual's daily patterns.

Insights:

The analysis provided deep insights into how raw location data can be transformed into a meaningful narrative about daily routines and personal habits. By clustering GPS points using DBSCAN and enriching these clusters with venue information from Foursquare, I was able to distinguish between my primary college apartment, my family home in Richmond, academic locations on campus, favorite dining spots, and even vacation destinations. The process highlighted that significant locations can be inferred not only by frequency and duration of visits but also by integrating contextual data from external sources, offering a richer and more nuanced understanding of my behaviors.

Looking ahead, there are several opportunities for improvement. Future work could focus on refining the automated labeling process through machine learning techniques that incorporate additional contextual features such as temporal patterns, weather, or calendar events. Incorporating other data sources like social media check-ins could also provide even greater detail, allowing for more robust information and location significance. Additionally, expanding the timeframe beyond a single month would help distinguish more consistent, long-term habits. Overall, this analysis demonstrates the potential of using location data to generate actionable insights, which could be applied to fields such as urban planning, personalized marketing, or even individual well-being studies.