GPS data and dwell pattern analysis

1. Data Overview

The dataset under study is a comprehensive collection of GPS location data from anonymized users, encapsulated in a file named 'gps.csv'. It comprises 1,770,011 records spread across four columns: 'user_id', 'datetime', 'lat' (latitude), and 'lon' (longitude). These columns represent the anonymized ID of the user, the timestamp of the location data in UTC format, and the geographical coordinates where the user was located at the time, respectively. A notable aspect of this dataset is its cleanliness and consistency, characterized by the absence of missing values or discrepancies in data formatting. There are a few location outliers but should not be removed as they might not be outliers in this context. It encompasses data from 31,606 unique users, offering a substantial basis for analysis.

2. Data Insights

The distribution patterns within the dataset reveal critical insights. The latitude data follows a normal distribution with a pronounced peak at 51.490, albeit with a leftward skew. This skewness indicates a concentration of data points below the mean, suggesting a geographic concentration of users in specific areas. Conversely, the longitude data does not conform to a normal distribution, hinting at a more dispersed or varied geographic distribution of users longitudinally. An interesting temporal pattern emerges in the distribution of hours, seeming normal with noticeable spikes on the 8th, 17th, and 18th hours indicating periods of increased activity or data collection (see Fig 2.1).

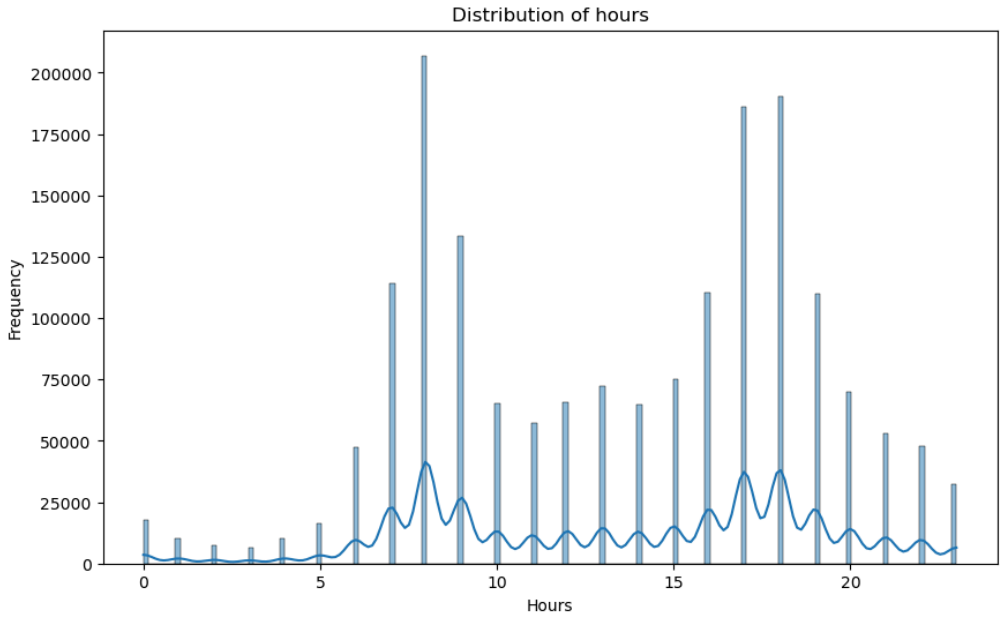
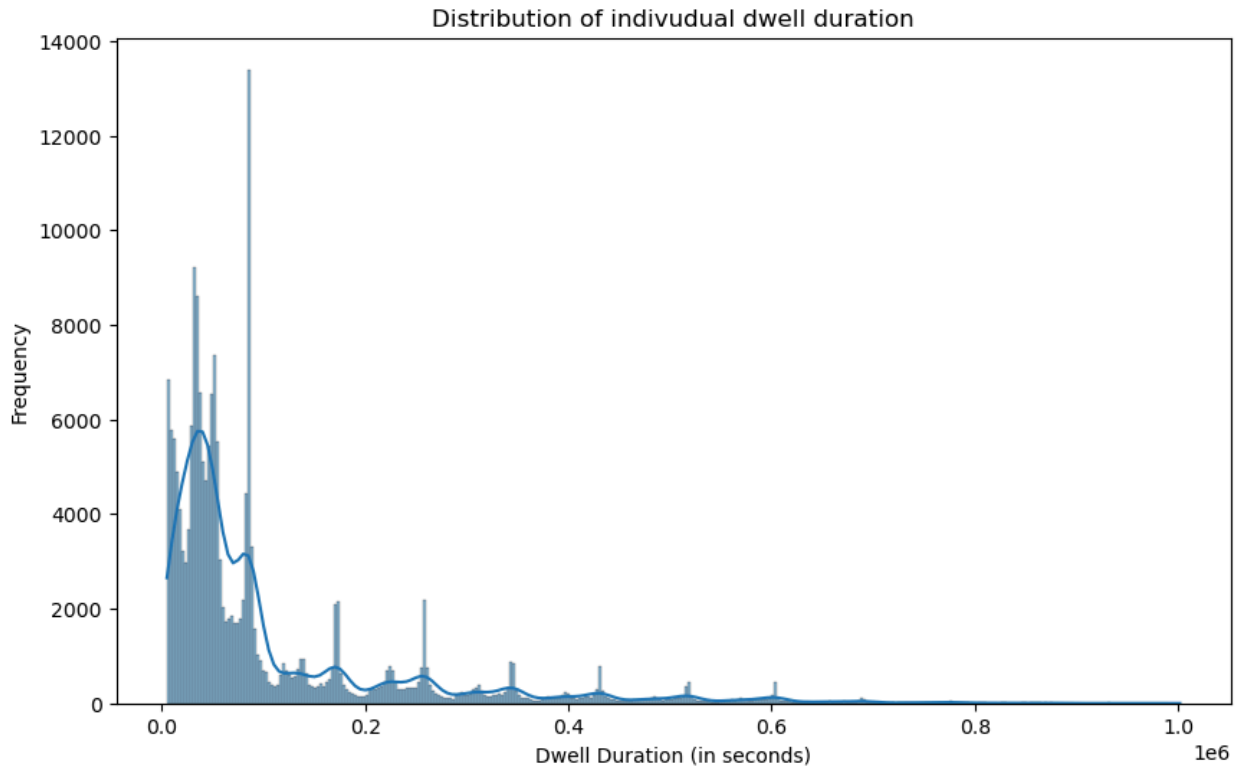


Fig 2.1: Distribution of hours in a day in the dataset ‘gps.csv’

These insights were augmented through feature extraction performed on the 'datetime' column, facilitating a more nuanced analysis of temporal distributions.

The distribution of individual dwell durations (in seconds) within 10 meter radius is also depicted in Fig 2.2.



Fing 2.2: Distribution of individual duration for dwelling more than 5000 seconds.

3. Algorithm Implementation

To extract meaningful patterns from the GPS data, particularly focusing on identifying user dwell times, a specific algorithm was developed with the following steps:

Define Dwell Periods: A dwell is identified based on a minimum duration of 5 minutes and a spatial threshold of 10 meters to differentiate between significant stops and negligible movements.

Time-Based Clustering: The algorithm proceeds by sorting the data by 'user_id' and 'datetime', then iteratively processes each user's data to identify dwells. It employs a combination of time difference and spatial distance criteria to detect when a user has remained within a predefined area for a significant duration. This methodology effectively segments the data into meaningful periods of user activity and inactivity.

Furthermore, to predict user locations, Random Forest and Decision Tree Regressors were trained on features extracted from 'datetime', such as 'day', 'hour', 'minute', and 'second'. These models aim to forecast the latitude and longitude of users independently.

4. Architecture

The system architecture integrates predictive modeling with time-based clustering to ascertain whether a predicted location should be flagged as a dwell. This dual approach leverages the strengths of machine learning for location prediction and algorithmic processing for dwell detection, offering a robust framework for analyzing GPS data.

5. Evaluation

The performance of the predictive models is quantitatively assessed using the Root Mean Square Error (RMSE) metric, which provides a measure of the prediction accuracy. The latitude prediction models achieved an RMSE between 0.00677127 and 0.00735561, while the longitude prediction models recorded an RMSE between 0.0133835 and 0.0145708. These scores reflect a high degree of precision in the models' ability to forecast user locations based on user’s time data.

6. Future works

Plotting standardized location data of users and performing extensive analysis could uncover more user pattern insights. The Time-Based Clustering algorithm used has a time complexity of O(n) and the volume of data in this situation is considerably high. Hence, time complexity of the algorithm can be further improved to a considerably lower time complexity algorithm of O(log(n)). Tuning parameters of Regression models could also improve user location prediction accuracy considerably.

7. Conclusion

The analysis of the GPS dataset, grounded in a rigorous methodological approach, offers significant insights into user mobility patterns and demonstrates the potential of machine learning algorithms in enhancing our understanding of geographical data. The implemented algorithm for identifying dwells, coupled with the predictive accuracy of the latitude and longitude models, sets a foundation for further exploration and application of GPS data in various domains, from urban planning to personalized location-based services. This report not only showcases the depth of analysis possible with such datasets but also highlights the intricate relationship between data patterns, algorithmic processing, and model evaluation in the realm of data science.