Insights Report: GPS Data Analysis

Introduction

This report presents a comprehensive analysis of GPS data, focusing on estimating individual dwells and extracting meaningful insights. The primary objective was to implement algorithms, scrutinize patterns, durations, and frequencies, and identify potential shortcomings and intriguing findings within dwell periods.

Methodology

Data Preprocessing

Commencing with the import of GPS data, the analysis initiated with meticulous data cleaning and preprocessing to ensure the highest data quality and consistency standards.

Identifying Dwell Periods

Two clustering algorithms, DBSCAN and K-Means, were strategically employed to estimate individual dwells. DBSCAN, known for its spatial density-based approach and noise-handling capabilities, was selected alongside K-Means for its simplicity and efficiency in identifying distinct clusters during initial data exploration.

Criteria for Identifying Dwell Periods

The criteria for identifying dwell periods were meticulously established to capture instances where users exhibit stationary behaviour for a defined minimum duration. The criteria included:

1. **Sorting by User and Timestamp**: Sorting the GPS data by user_id and datetime ensures that each user's subsequent analysis is performed chronologically, a critical step for accurate time difference calculations and dwell period identification.

2. **Calculating Time Differences**: Time differences between consecutive GPS points were calculated to quantify the temporal gap, forming the foundational basis for identifying periods of stationary behaviour.

3. **Setting Dwell Threshold**: The dwell threshold, set at 10 minutes, defines the minimum duration for a period to be considered a dwell. Users must remain close for at least 10 minutes to qualify as part of a dwell period.

4. **Filtering Out Dwell Data Points**: Filtering based on the dwell threshold ensured that only significant time intervals were considered dwell periods, eliminating short movements or data noise and focusing on periods where users likely remained stationary for a specified duration.

User-Specific Insights

The analysis is intricately structured to provide insights on a per-user basis, offering a granular understanding of GPS patterns. Key user-specific insights include:

Dwells Estimation: Each algorithm identifies each unique dwell duration for each user based on the number of clusters formed. Each cluster formed has a dwell duration showing the date, dwell duration, Minimum and maximum durations to understand the range of time the user spent at specific locations. For example, User fffd15f4-bbd5-4bac-9d9f-a9568bb701fc exhibited a dwell duration of 0 days 13:45:28 to 10 days 22:28:13 days using DBSCAN, and using the KMeans Dwell durations ranged from 3 days 00:50:42 to 9 days 14:34:18. The wider range of durations suggests a more diverse set of locations, including both short stops and longer dwells.

2. Individual Dwells Clusters Visualization: Examining individual user behaviors through scatter plots, KMeans and DBSCAN showcased diverse clusters, capturing distinct patterns in user behavior, modes of movement, or stationary behavior. These visualizations revealed variations in density-based clusters specific to each user and identified outliers.

DBSCAN: This incremental algorithm exhibited different clusters for each user, emphasizing individual behaviours without assuming predefined shapes.

K-means Clustering: Assumed roughly similarly sized clusters like a spherical shape, KMeans displayed about 10 clusters for each user with diverse densities, including both low-density and high-density clusters.

Visualization Analysis

1. Dwell Duration vs. Day of the Week: Analyzing dwell durations based on the day of the week provided insights into temporal patterns and variations in user behavior for both algorithms.

2. Dwell Duration vs. Time of Day: Analyzing dwell duration in relation to the time of day identified patterns throughout the day, indicating longer dwells during certain hours or shorter dwells during peak travel times. Notably, some users displayed no correlation between dwell durations and the time of day.

3. Daily Variation of GPS Points: Comprehensive plots showcased patterns in user movements or stationary behavior during different times of the day and each cluster. The plots effectively captured irregularities in movement, particularly in regions with varying point densities.

4. Clusters Frequencies: Analyzing the frequency of clusters identified by both DBSCAN and K-Means provided insights into the distribution of dwell periods among the clusters. This analysis revealed whether certain clusters were more common or if dwell periods were evenly distributed across clusters.

5. Dwell Durations: Examining dwell durations helped understand the variability in the lengths of individual dwells. The analysis provided insights into whether users tended to stay in one location for consistent durations or if dwell lengths varied significantly. Analyzing dwell durations for K-Means clusters demonstrated variations in the time spent in different locations, capturing potential different types of activities. Similarly, for DBSCAN clusters, irregular patterns of stay were detected, highlighting areas with varying densities.

6. Dwell Durations Distribution: This visualization identified the spread of dwell periods, including peaks and outliers.

7. Dwell Duration vs. Time of Day: Analyzing dwell duration in relation to the time of day revealed patterns throughout the day. This analysis provided insights into whether dwell durations varied based on the time of day, such as longer dwells during certain hours or shorter dwells during peak travel times.

8. Dwell Frequencies by User: Understanding dwell frequencies by user for both algorithms highlighted individual preferences and routines. This analysis provided insights into whether certain users had more frequent or longer dwells than others. Dwell frequencies were plotted by users to understand how dwell periods were distributed among users.

9. Dwell Frequencies by Cluster: Analyzing dwell frequencies by cluster revealed which locations or regions were more frequently visited, offering insights into the importance of different clusters.

10. Distribution of Data Among Different Clusters

K-Means: its data distribution showed a more balanced distribution if the data naturally forms well-separated clusters.

DBSCAN revealed uneven cluster sizes, as it is density-based and may merge regions with lower densities.

11. Boxplot Analysis: the two algorithm plots visualized the distribution of dwell durations within each cluster. This analysis helped identify outliers and compare the variability of dwell durations across different clusters.

Limitations:

The analysis relies on accurate GPS data and may overlook outliers or inaccuracies in the dataset. Moreover, selecting a minimum duration for dwell periods is subjective and can influence the results.

Conclusion

Analyzing GPS data using the defined criteria provides valuable insights into user behaviours, dwell periods, and temporal variations. The methodology's flexibility allows for customization based on specific use cases, enhancing its applicability across diverse scenarios. This insights report is a foundation for further exploration and refinement, encouraging ongoing analysis to extract meaningful intelligence from GPS data.