Assumptions for All Questions

Question 1: Estimating Vehicle Residence Location

- Residence location is defined as the most frequently visited GPS location during the night hours (12:00 AM to 5:00 AM).
- If GPS drift caused many close points, DBSCAN clustering could have been used to group them, and consequently the most frequently occurring cluster would have been selected as the likely residence but that was not the case here.
- It is assumed that vehicles are generally stationary and parked at the residence location during night hours and only most common location is considered; no secondary residences analyzed.
- Nevertheless I did use DBSCAN and it still got similar results for residence location.

Question 2: Detecting Vehicle Charging Locations

- Each detected charging start was associated with its corresponding GPS coordinate as the charging location.
- Records with missing or invalid GPS or battery data were excluded prior to analysis.

Question 3: Identifying At-Risk Drivers Based on Distance

- Earning is assumed to be directly proportional to the daily distance driven.
- Distances were calculated between consecutive GPS coordinates using the Haversine formula.
- The daily distance for each vehicle was summed, and the average was computed across all days.
- Vehicles were flagged as "likely to default" if:
  - Their average daily distance was less than 15 km (fixed threshold), or
  - They fell in the bottom 20% of average daily distances (percentile-based).
- Invalid GPS points were excluded before distance computation.

Question 4: High and Low Density Area Detection

- Duplicate location entries for each vehicle (`Vin`, `Avg_lat`, `Avg_long`) were removed to reduce overweighting of stationary points.
- DBSCAN was used with parameters: $\varepsilon = 0.001$ (approx. 100 meters), `min_samples` $= 5$.
- Clusters with point count greater than the median were labeled as "High Density"; the rest as "Low Density".

- DBSCAN cluster label -1 was treated as noise, representing very low-density or outlier regions.
- K-Means clustering was also explored using the Elbow Method to determine the optimal number of clusters based on inertia.