Public Route Optimization Roadmap

1. Data Formats for Mobility Pattern Data

- **CSV (Comma-Separated Values)**: A simple text format easily read and manipulated using various tools and programming languages.
- **JSON (JavaScript Object Notation)**: A lightweight format for data exchange, easy for humans to read and write, and machines to process.
- **XML (eXtensible Markup Language)**: A flexible text format commonly used for structured information.
- **Parquet**: A columnar file format optimized for data processing frameworks like Apache Spark and Hadoop.
- **GeoJSON**: A format for encoding geographic data structures using JSON.

2. Data Cleaning Methods

Handling Missing Data

- **Imputation**: Replace missing values with reasonable estimates (e.g., mean, median, mode, or advanced methods like KNN imputation).
- **Removal**: Remove records with missing data if the proportion is small.

Dealing with Outliers

- **Statistical Methods**: Identify outliers using techniques like Z-score or IQR and decide to remove or transform them.
- **Domain Knowledge**: Apply specific knowledge to determine if an outlier is truly unusual or a valid edge case.

Noise Reduction

- **Smoothing Techniques**: Apply algorithms like moving averages or Gaussian smoothing to reduce noise.
- **Aggregation**: Aggregate data over larger periods or regions to minimize noise impact.

Data Consistency

- **Standardization**: Ensure uniform data formats (e.g., date and time, coordinates).
- **Deduplication**: Identify and remove duplicate records.

3. Implementation Roadmap

Initial Data Analysis

1. **Data Ingestion**: Load mobility data into a suitable environment (e.g., Python, R, Apache Spark).
2. **Preliminary Analysis**: Conduct exploratory data analysis to understand structure, characteristics, and initial insights. Generate summary statistics and visualizations.
3. **Data Cleaning**: Apply data cleaning techniques as discussed above.

Integration with Existing Transportation Routes

1. **Data Matching**: Match mobility data with public transport stops using coordinates.
2. **Temporal Analysis**: Analyze time patterns to understand peak times and daily trends.

Identification of Gaps and Optimization Opportunities

1. **Gap Analysis**: Identify areas with high mobility but low transport coverage, and redundant routes with low usage but high mobility.
2. **Optimization Modeling**: Develop models to propose new or modified routes using techniques like integer programming, genetic algorithms, or machine learning.

Visualization of Optimized Routes

1. **Visualization Tools**: Use tools like Folium, Kepler.gl, or custom applications to visualize routes.
2. **Interactive Maps**: Create maps for stakeholders to explore optimized routes, compare with current routes, zoom, filter by time, and view statistics.