Project Assignment: Unsupervised Exploration of the Electric School Bus Adoption Dataset

1. Introduction

Electric school buses (ESBs) are an emerging technology aimed at reducing students' exposure to diesel exhaust and cutting greenhouse gas emissions in school transportation. Adoption of ESBs is accelerating in the United States, but still represents a small fraction of the total school bus fleet. As of mid-2024, over 12,000 electric school buses have been committed (funded or ordered) across the country. This project will examine a dataset tracking ESB adoption nationwide. The dataset, compiled by the World Resources Institute (WRI), provides detailed information on which school districts are adopting electric buses and how the adoption process is unfolding. Through this project, students will leverage unsupervised machine learning techniques to uncover patterns and insights from the data, without a predefined target variable. The focus is on exploration and knowledge discovery, using methods such as clustering, anomaly detection, dimensionality reduction, and association rule mining, to better understand the landscape of electric school bus adoption.

2. Dataset Overview

The Electric School Bus Adoption dataset *[1]* contains comprehensive data on school districts that have begun transitioning to electric buses. Key features of the dataset include:

- **District-Level Information:** Demographic and environmental indicators for each school district, such as poverty rates, racial composition, and local air pollution levels. The dataset also notes the district's locale classification (urban, suburban, town, or rural) and other socio-economic characteristics. These features enable analysis of equity in ESB adoption (e.g., are disadvantaged communities being reached?).
- **Bus Adoption Details:** The number of committed electric school buses in each district, where a "committed" ESB means the district has secured funding or made a purchase agreement for the bus. For each electric bus, details are provided such as the bus manufacturer and the funding source(s) that enabled its purchase. This allows analysis of which manufacturers are most common and how buses are being funded (federal grants, state programs, utility incentives, etc.).
- **Adoption Timeline and Phases:** Each electric bus's progress through the adoption process is tracked across four phases – "awarded", "ordered", "delivered", and "operating". Dates or statuses indicate when a bus reached each phase, and the current phase of each bus is recorded. This timeline data makes it possible to study how quickly buses move from funding award to active operation, and to identify any bottlenecks or delays in deployment.
- **Additional Characteristics:** The dataset may include other relevant attributes, such as the size of the school district (e.g., number of students), geographic identifiers (state or region), and possibly the total fleet size of buses per district for context. These can be used to normalize or contextualize adoption (for instance, percentage of a district's buses that are electric).

In summary, the dataset provides a rich, multi-faceted view of electric school bus adoption across U.S. school districts. It combines technical adoption metrics with socio-economic context, enabling a wide range of analyses.

For more detailed information about the dataset structure, definitions, and data sources, students are encouraged to consult the official technical note published by the World Resources Institute *[2]*.

3. Objective and Scope of the Assignment

The objective of this assignment is to define a specific analytical question or problem related to the electric school bus adoption data, and to explore that question using appropriate unsupervised learning techniques. Rather than focusing on prediction, this project emphasizes pattern discovery, grouping, and relationship mining in the data. You are expected to:

- Formulate a **clear problem statement** or research question based on the dataset. For example, you might be interested in finding clusters of similar school districts, detecting outlier districts with unusual adoption patterns, or uncovering hidden associations between funding sources and adoption success. This problem definition should be explicitly stated in your report/notebook.
- Apply **unsupervised machine learning** methods to investigate the question. Relevant techniques include:
    - Clustering algorithms (such as K-Means, DBSCAN, or Gaussian Mixture Models) to identify groups of districts or buses with similar characteristics.
    - Anomaly detection methods (such as Isolation Forest) to find districts or individual buses that have unusual or noteworthy properties (for instance, extremely fast or slow adoption timelines, or atypical funding situations).
    - Dimensionality reduction techniques (such as Principal Component Analysis (PCA) to reduce high-dimensional data into a few principal components) for visualization and to possibly improve clustering or detection by denoising the data. PCA or similar methods can help reveal underlying structure and highlight which factors contribute most to variance in adoption.
    - Association rule mining (e.g., using the Apriori algorithm or others) to discover frequent patterns or co-occurrences in categorical data. This could reveal, for example, rules like "Districts of type urban with high pollution and high poverty are often those that received federal funding for ESBs," or other interesting associations between demographic factors and funding or manufacturer choices.
- **Combine multiple techniques** when appropriate. Many insights may emerge from using techniques in tandem. For instance, you might use PCA to project the data into two dimensions and then run K-Means clustering on those principal components to see clearer group separation. Or you might cluster districts first and then apply anomaly detection within each cluster to find outliers relative to their peer group. Likewise, association rule mining might be applied to each cluster separately to find distinct patterns in different clusters. You are encouraged to be creative and use more than one method if it helps analyze your chosen problem from different angles.
- **Interpret and analyze** the results of your unsupervised learning. Since unsupervised methods do not yield simple “right or wrong” answers, a crucial part of the project is making

sense of what the algorithms find. For each technique you apply, discuss what the results mean in the context of electric school bus adoption. For example, if clusters are found, describe the profile of each cluster (what characterizes the districts in each group?). If outliers are detected, investigate why those cases might be unusual (perhaps a district has an exceptionally large number of buses or an atypical funding source). If association rules are mined, assess their interestingness and what they imply (are they obvious correlations or surprising discoveries? do they suggest any causal hypotheses or important factors?).

By exploring the data with unsupervised methods, you might uncover patterns such as clusters of districts that are leading in adoption, relationships between district demographics and adoption rates, common factors among the most delayed bus deployments, or typical combinations of funding sources and manufacturers. Your analysis should aim to highlight such findings and provide a deeper understanding of how electric school bus adoption is playing out across different contexts.

4. Guiding Questions and Suggested Directions

To help you get started, here are some guiding questions and directions you might consider. These are suggestions – you do not have to cover all of them, and you may formulate a different question entirely – but they illustrate the kind of exploration expected:

- **Clustering of Districts:** How do school districts group together based on their characteristics and adoption status? For example, using clustering algorithms (KMeans, DBSCAN, GMM), can you identify clusters of districts that have similar demographics and electric bus adoption metrics? What features define each cluster – do some clusters represent wealthy, urban districts with early adoption, while others represent rural or high-poverty districts? Analyzing cluster composition could reveal if there are archetypes of ESB-adopting districts.
- **Outlier Detection:** Can we detect any districts or buses with unusual adoption patterns or characteristics? Apply an anomaly detection technique (e.g., Isolation Forest) to identify outliers. This might highlight, for instance, a district that adopted a surprisingly large number of electric buses relative to its size, or a bus whose timeline from funding to operation is exceptionally long or short. Investigate these anomalies – are they data errors, or do they point to interesting cases (such as a pilot program that rapidly deployed buses, or conversely a district that secured funding but has faced delays)?
- **Dimensionality Reduction & Visualization:** What patterns emerge when we reduce the dataset's complexity and visualize it? Use PCA (or a similar method) to compress the dataset's features into principal components. By plotting districts or buses in the space of the first two or three principal components, you may observe natural groupings or gradients. Visualizing clusters on such a plot can validate if the clustering makes intuitive sense. You could also examine the loadings of PCA to understand which variables contribute most to variations in the data.
- **Association Rules in Categorical Data:** Can we extract meaningful association rules from the data's categorical attributes? Identify categorical features in the dataset (e.g., types of funding sources – federal grant, state grant, utility rebate, etc.; bus manufacturer; region or state; district locale type). Using association rule mining (Apriori or similar), find patterns like "IF [High poverty] AND [Received EPA grant] THEN [Locale = rural] with support X and confidence Y" (as a hypothetical example). Focus on rules that have strong support and

confidence, and interpret their significance. Do these rules reveal anything notable, such as certain funding programs being predominantly used by certain kinds of districts, or certain manufacturers being favored in specific regions?

In formulating your analysis, you might choose one of the above questions as your primary focus or combine elements from multiple questions. You are not limited to these examples – any well-defined exploration that uses unsupervised learning on the given dataset is welcome. However, ensure that your analysis remains **coherent**: it should have a central question or theme, rather than a disjointed set of unrelated mini-analyses. The guiding questions above can often be related (for instance, you might cluster districts and then look for association rules within each cluster to characterize them). Consider how the different techniques can complement each other to tell a fuller story.

5. Project Work Expectations

While working on this project, keep the following expectations and best practices in mind:

- **Data Preprocessing:** Thoroughly examine the dataset and perform any necessary preprocessing. This may include handling missing values, encoding categorical variables (for clustering or PCA), scaling features (especially important for distance-based methods like KMeans or PCA), and aggregating or deriving new features (e.g., computing the percentage of buses electrified in a district, or time differences between adoption phases). Document any cleaning or transformation steps in your notebook.
- **Justification of Methods:** When you choose a particular method or algorithm, provide a brief justification. For example, if you decide to use DBSCAN for clustering, explain that it can find non-linear cluster shapes and can determine the number of clusters automatically, which might be suitable if you suspect non-globular groupings. If you use KMeans, discuss how you selected the number of clusters (perhaps by using the elbow method or silhouette score). For anomaly detection, explain how you set any parameters (like contamination rate for Isolation Forest). This shows your understanding of the methods and their application.
- **Validation and Sensitivity:** Since unsupervised learning lacks a simple accuracy metric, validate your findings through alternative approaches. For instance, compare results from different clustering algorithms to see if similar groupings occur, or test different values of k in KMeans. If using PCA, check how much variance the principal components explain. For association rules, experiment with different support/confidence thresholds to find a balance between too many trivial rules and too few high-quality rules. Discuss what remains consistent versus what changes under these variations – this will strengthen the credibility of your insights.
- **Use of Visualization:** Leverage visualizations to support your analysis at every stage. Plot demographic variables, create bar charts of funding sources, draw timelines for adoption phases, or map out clusters in two dimensions (perhaps using PCA results). Visual exploration can often reveal patterns that quantitative metrics miss, and it will make your findings more interpretable. Include relevant charts or graphs in your notebook with clear labels and captions, and refer to them in your discussion.
- **Interpretation and Domain Context:** Always tie your technical findings back to the context of electric school bus adoption. For example, if you find that one cluster of districts is

characterized by high poverty and low adoption counts, discuss what that might mean (e.g., perhaps those districts face financial or infrastructural barriers). If an outlier bus took unusually long to go from "delivered" to "operating," consider external factors (maybe infrastructure delays like charging station installation). Show that you understand the real-world implications of patterns in the data. Where appropriate, you may reference external knowledge or reports on electric bus adoption to support your interpretations (ensure to cite any such sources).

- **Clarity and Organization:** Present your work in a well-organized manner. Introduce the problem you're addressing, then document your analysis steps logically. Use markdown headings, bullet points, and narrative explanations in your Jupyter notebook to make it readable as a report. Each section of your analysis (e.g., data understanding, clustering analysis, anomaly analysis, etc.) should be clearly separated and titled. This will help readers follow your thought process.

6. Deliverables

By the end of the project, you are expected to submit the following:

Jupyter Notebook – This notebook should contain all your code, analysis, and results. It should be executed in sequence so that all outputs (figures, tables, etc.) are visible. The notebook must include written explanations and commentary interwoven with the code to explain what you are doing and why. It should read like an analysis report, not just a collection of code cells. Make sure to save the notebook with outputs (e.g., do not clear the cells before submission).
Ensure that your notebook can be run on a standard environment to reproduce the results (i.e., include any necessary instructions for environment setup or mention if any special libraries are used aside from common ones like pandas, numpy, scikit-learn, etc.). If you performed any steps that are not obvious (such as filtering out certain data or creating new variables), make sure these steps are coded or explained.

7. Conclusion

This project is an opportunity to apply **unsupervised machine learning techniques to a real-world dataset** with significant societal relevance. Electric school bus adoption is a topic at the intersection of technology, environment, and equity. By analyzing this dataset, you will gain experience in extracting meaning from unlabeled data and communicating data-driven insights. There is **no single "right answer" to the questions posed** – the quality of your work will be judged on the soundness of your approach, the depth of your exploration, and the clarity with which you derive and explain insights. Approach the assignment systematically: start with understanding the data, then gradually apply techniques, and iteratively refine your analysis based on what you discover. We look forward to seeing how you define a problem and uncover patterns in the electric school bus adoption landscape using unsupervised learning.

- *[1]* https://electricschoolbusinitiative.org/dataset-us-electric-school-bus-adoption
- *[2]* https://www.wri.org/research/technical-note-dataset-electric-school-bus-adoption-united-states