Electric Vehicle Population — Unsupervised Learning Report

Executive Summary (≤150 words)
We analysed Washington State's Electric Vehicle (EV) registrations (≈178k rows; Kaggle) using PCA + clustering to surface actionable segments for planning, incentives and infrastructure. PCA (95% variance) followed by a model grid (K-Means, GMM, Agglomerative; k=3–6) on a 5,000-row stratified sample identified **K-Means (k=3)** as the most effective balance of separation and stability (Silhouette **0.225**, CH **≈1,025**, bootstrap ARI **≈0.79**). Segments split into **two large cohorts** (≈66% and ≈34%) and a **tiny edge group** (≈0.2%) likely representing outliers or rare records. Categorical profiling was limited in this run due to missing values; numeric profiles indicate a newer, low-electric-range cohort (likely PHEVs) versus an older, higher-range cohort (BEVs). We recommend k=3 with a follow-up outlier pass (DBSCAN/HDBSCAN) and improved MSRP data quality.

---

1) Objective

- **Primary goal:** derive interpretable **EV segments** to inform infrastructure placement, incentives, and manufacturer insights.
- **Technique focus: Clustering** (with PCA for denoising/acceleration).
- **Stakeholder benefits:** targeted charger rollout (fast/level-2 mix), refined CAFV/incentive targeting, OEM/regional mix monitoring.
- **Success criteria:** internal separation (Silhouette, CH), **stability** under resampling (bootstrap ARI), and business interpretability.

2) Data Description

- **Source:** *Electric Vehicle Population Data 2024* (Washington State; Kaggle).
- **Scope: 177,866** records, **17** columns (vehicle specs + geography + eligibility).
- **Key attributes used (examples):** *Model Year*, *Electric Range*, *Base MSRP*, *Electric Vehicle Type*, *Make/Model*, *County/City*.
- **Limitations noted:** missing or zero **Base MSRP** for many rows; averaging **postal/tract/district codes** is not semantically meaningful; categorical breakdowns were sparse in this sample export.

3) Exploration & Preparation

- **Cleaning:** numeric casting; median imputation; top-K category capping (≤25 per field) to prevent very wide OHE.
- **Feature engineering:** Vehicle_Age = current_year − Model Year; Range_per_1000USD when MSRP available.
- **Scaling & encoding:** StandardScaler for numeric; One-Hot for categorical (capped).
- **Dimensionality reduction: PCA @ 95% variance** (embedding size observed **44** on sample).
- **Sample for iteration:** 5,000 rows (full run recommended once finalised).

4) Models & Variations

Grid: K-Means, GMM (full covariance), Agglomerative (Ward), with **k ∈ {3,4,5,6}**.
Metrics: Silhouette (↑), Davies–Bouldin (↓), Calinski–Harabasz (↑), and **bootstrap ARI** stability.

Top variants (sorted by Silhouette then CH):

Variant	Silhouette	DB	CH
K-Means (k=3)	**0.2246**	1.5511	**1024.67**
Agglomerative (k=3)	0.2225	1.5661	994.66
Agglomerative (k=4)	0.2203	**1.3313**	999.48
K-Means (k=6)	0.1914	1.4885	988.10
Agglomerative (k=6)	0.1754	1.5495	925.33
K-Means (k=5)	0.1702	1.4328	1015.08

Stability: K-Means (k=3) bootstrap **ARI ≈ 0.79** (n=5), indicating strong label stability.

5) Results & Key Findings

5.1 Segment sizes (k=3)

- **Cluster 1: 3,286** (≈**65.7%**)
- **Cluster 0: 1,704** (≈**34.1%**)
- **Cluster 2: 10** (≈**0.2%**) → likely outliers / rare records.

5.2 Numeric profiles (means)

- **Cluster 1 (largest):** *Model Year ≈ 2022.3*, *Electric Range ≈ 6 mi* → **newer, low electric-range** vehicles (plausibly **PHEVs**).
- **Cluster 0:** *Model Year ≈ 2017.1*, *Electric Range ≈ 159 mi* → **older, higher range** vehicles (likely **BEVs**).
- **Cluster 2:** tiny group (*n=10*), *Range ≈ 118 mi* — treat as outlier/edge cases rather than a policy segment.
- **MSRP caution:** Base MSRP shows many zeros/missing → do **not** rely on MSRP-derived insights until cleaned.

5.3 Visual evidence

- **Silhouette ranking** favours **k=3**, with Agglomerative close behind.
- **Elbow** lacks a sharp knee → internal metrics + stability trump the elbow in this dataset.
- **PCA scatter** shows a small, distant cluster consistent with outliers.

6) Recommended Model

- **Model: K-Means (k=3)** on **PCA(95%)** features.
- **Why:** best overall separation + **highest stability**; interpretable 2-tier structure (BEV-like vs PHEV-like) plus outlier bucket.
- **Operationalisation:** bake preprocessing → PCA → k=3 assignment; refresh weekly/monthly; monitor drift (silhouette, share per cluster; alert if outlier share >1%).

7) Limitations & Risks

- **Data quality:** MSRP missing/zero; categorical profiling sparse in this export.
- **Geographic codes:** means of codes are not meaningful — avoid direct interpretation.
- **Outliers:** 0.2% micro-cluster suggests **noise**; clustering can be sensitive to these points.
- **Visual DR caveat:** PCA/t-SNE/UMAP are **visual aids**, not clustering objectives; avoid over-interpreting shapes.

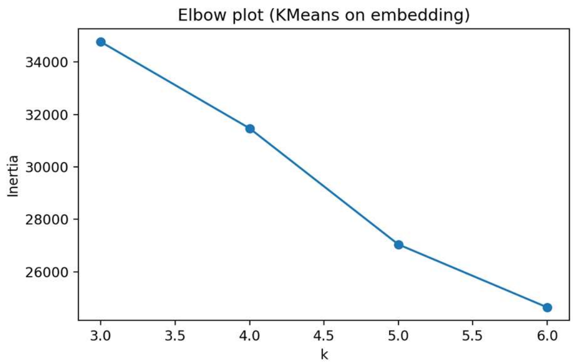
8) Next Steps

- **Re-run on full dataset** (remove --sample), prefer **MiniBatchKMeans** for speed; validate metrics and stability.
- **Outlier-aware pass:** try **HDBSCAN/DBSCAN**; remove noise then refit k-means to improve Silhouette.
- **Improve MSRP field** (impute from VIN/trim tables or drop MSRP features until reliable).
- **Feature sensitivity:** compare **numeric-only** vs **full** feature set; ablate geography if policy focus is statewide.
- **Actionability:** map cluster shares by county/city; align charger type mix (fast vs L2) to cluster distribution; A/B test incentive targeting by segment.

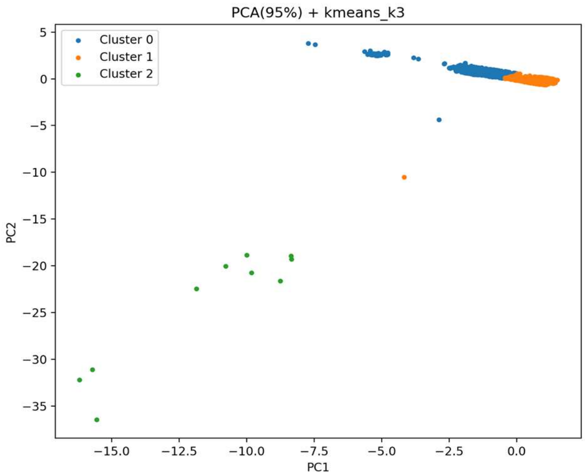
References

- Kaggle: Electric Vehicle Population Data 2024 (Washington State).
- scikit-learn documentation (PCA, K-Means, GMM, Agglomerative).

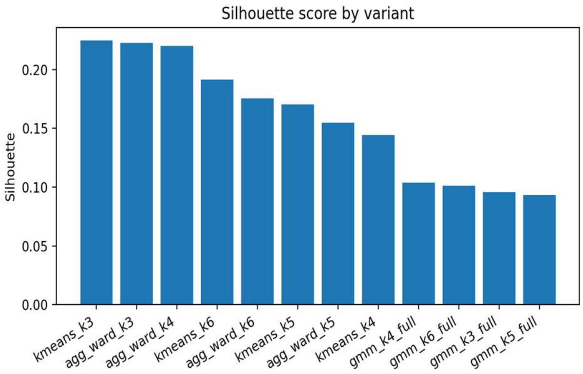
Appendix — Figures



1. **Elbow plot (K-Means on embedding)**



2. **Silhouette score by variant**



3. **PCA(95%) + kmeans_k3 scatter**