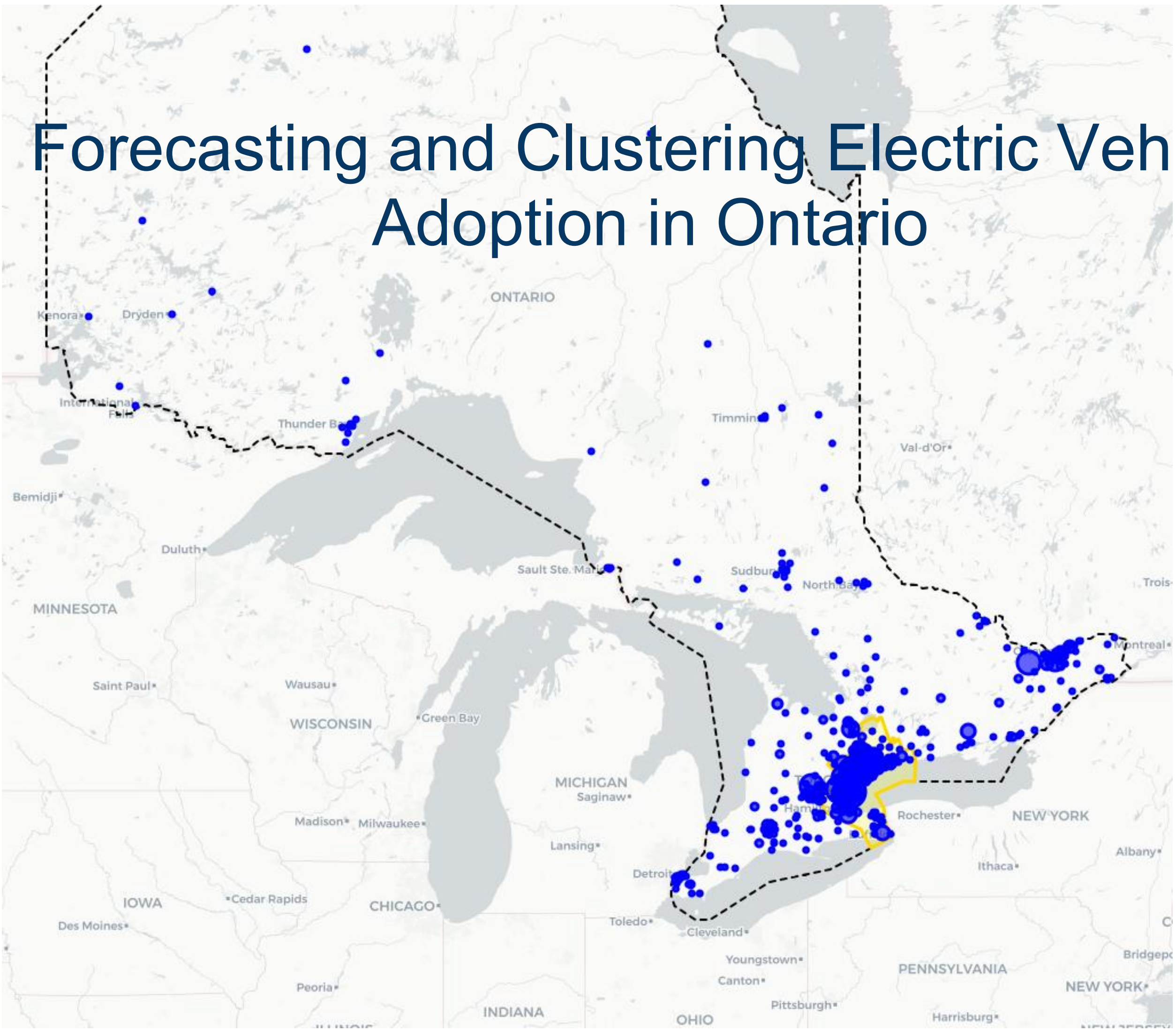
Machine Learning

GROUP ASSIGNMENT: Predictive Model

Forecasting and Clustering Electric Vehicle Adoption in Ontario



Instructor: Delina Ivanova
Date: Aug 11, 2025

Group 4 Members
Oleg Byelyayev
Myunghee Jung
Noah Levin
Priyanshu Verma
Ning Liu
Megan Da Costa

Table of Contents

1. Objectives

1.1 Problem Statement

This project examines electric vehicle (EV) adoption trends across Ontario using quarterly registration data organized by Forward Sortation Area (FSA). The overarching goal is to apply data science techniques to uncover meaningful spatial and temporal patterns in EV adoption. To achieve this, the analysis focuses on **two specific objectives**:

1. **Clustering:** Identify groups of FSAs that exhibit similar EV adoption patterns, highlighting regional similarities and differences in growth trends.
2. **Prediction:** Develop forecasting models and select the best model to estimate future EV uptake in different regions, providing insights into likely adoption trajectories and potential areas of rapid growth.

1.2 Hypotheses to test

Hypothesis 1 -- Overall Growth Hypothesis: EV adoption in Ontario FSAs has increased consistently over recent quarters, with measurable growth detectable across most regions.

Hypothesis 2 -- Regional Disparity Hypothesis: Certain FSAs experience significantly higher adoption rates than others, due to favorable socioeconomic conditions and greater charging infrastructure availability.

Hypothesis 3 -- Technology Preference Hypothesis: BEV and PHEV adoption patterns differ across clusters, with urban/suburban FSAs showing stronger BEV dominance and rural FSAs exhibiting a greater reliance on PHEVs.

Hypothesis 4 -- Predictive Modelling Hypothesis: Machine learning models incorporating prior-quarter adoption metrics can forecast near-term EV growth at the FSA level, enabling identification of future high-growth areas.

2. Data Preparation

2.1 Dataset Description

The **EV dataset**, sourced from the Ontario Open Data Catalogue, provides quarterly counts of EVs-- both battery-electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs)--registered in Ontario, broken by FSA. Each record includes:

- Total EVs: The total number of registered EVs.
- BEVs: The number of registered battery-electric vehicles.
- PHEVs: The number of registered plug-in hybrid electric vehicles.

Data from six consecutive quarters—from Q2 2023 to Q1 2025—were downloaded and merged, enabling both temporal and spatial analyses of EV adoption trends across Ontario.

In addition, the "ontario_fsa_coordinates_with_city_region.csv" dataset contains **geographic and administrative information** for FSAs across Ontario, Canada. Each row represents a unique Ontario FSA and includes:

- FSA: The first three characters of a Canadian postal code identify specific geographic areas within Ontario.
- City: The main city or community associated with the FSA.
- Region: A broader classification grouping FSAs into major regions or metropolitan areas (e.g., Toronto, Ottawa East, Rural Central Ontario).
- Latitude and Longitude: Decimal coordinates representing the approximate center point of the FSA, used for mapping and spatial analysis.

This dataset was compiled and cleaned from multiple open-data and public-domain sources, as comprehensive geospatial datasets from Canada Post or commercial providers often require paid access. Leveraging freely available sources avoids these costs while ensuring the data is fully shareable and usable within the course setting.

2.2 Data Cleaning & Feature Engineering

- Filtering for Ontario FSAs: The dataset was limited to Forward Sortation Areas (FSAs) in Ontario by selecting codes starting with K, L, M, N, or P.
- Handling Missing Geographic Data: Missing geographic coordinates or administrative labels were filled using publicly available sources. When exact coordinates were unavailable, approximate centroids from neighboring FSAs or known postal regions were used, following standard geospatial imputation practices.

2.3 Tools and Technologies Used

Table 1. *Toolkit for Python-Based Data Analysis in Jupyter/Colab*

Category	Technology/Tool	Primary Function
Data Manipulation	pandas, geopandas	Data cleaning, aggregation, and geospatial analysis
Plotting (Static)	matplotlib, contextily	Line charts, geospatial base maps
Plotting (Interactive)	folium, plotly	Interactive spatial and time series visualization
Scripting Language	Python 3.12.3	All programmatic operations
Environment	Jupyter/Colab	Interactive analysis workspace
Data Storage	CSV, Google Drive	Easy data access and results sharing

2.4 Challenges Encountered

FSA is not present in reliable geographic sources and is typically not officially assigned as a postal region by Canada Post or mapping authorities, leading to potential gaps in geospatial plotting. Forecasting was hindered by limited and volatile data, resulting in poor LSTM performance with large errors and negative R^2. While ARIMA performed better, it still did not reliably beat simple baselines. The ensemble modestly improved results over LSTM but did not surpass ARIMA. These issues highlight the need for more historical data, richer features, simpler models, and robust validation methods to improve forecast reliability.

2.5 Exploratory Data Analysis (EDA)

- **Quarterly Growth Rate in ON**

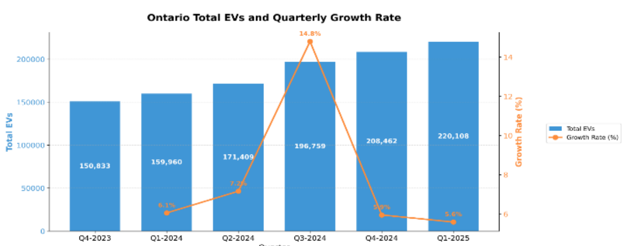


Figure 1. *Quarterly Growth Rate.* Created in Jupyter Lab.

Overall growth rates ranged from 5.6% to 14.8% across 6 Quarters. The most significant increase occurred between Q2 and Q3 2024, with a 14.8% rise in total EVs, as the province surpassed 196,000 registered vehicles. This sharp spike may reflect policy changes, new model releases, or infrastructure improvements. Growth remained positive in subsequent quarters, albeit at a slower pace.

- **BEV/PHEV Ratios**

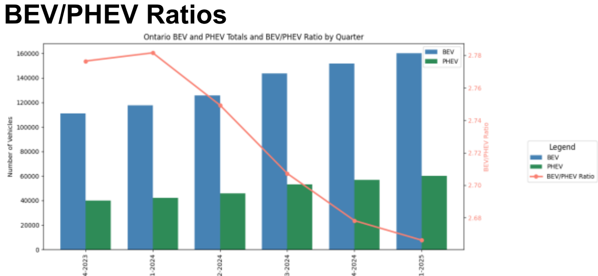


Figure 2. *BEV/PHEV Ratios.* Created in Jupyter Lab.

BEVs dominate the Ontario EV landscape, making up nearly three-quarters of all registrations.
Both BEV and PHEV totals rose steadily, with BEVs consistently outnumbering PHEVs. However, the BEV/PHEV ratio declined from over 2.78 to below 2.68 across 6 quarters, suggesting PHEVs are growing slightly faster—possibly due to changing costs or charging infrastructure.

- **Visualization of Data Locations**

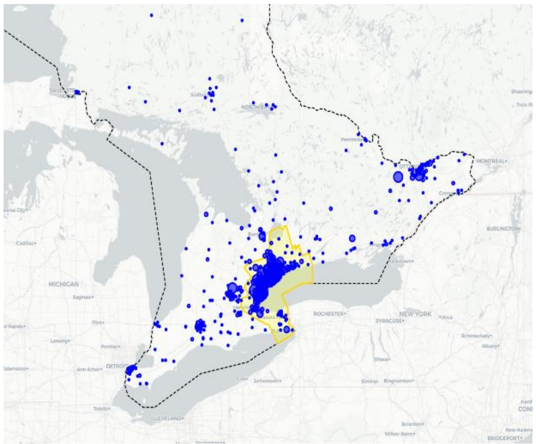


Figure 3. *Spatial Density of Data Points in Southern Ontario—Greater Golden Horseshoe Boundary*. Created in Jupyter Lab.
(View downloadable interactive map at https://drive.google.com/file/d/16N6x4WS4ZBa566XPs3ppLJEh0L_-WlcM/view?usp=drive_link and in Appendix)

The yellow boundary marks the officially designated Greater Golden Horseshoe (GGH) region—a key urban and economic area that stretches from Niagara and Hamilton through Toronto and north to Barrie, encompassing much of Ontario's population and economic activity. The map illustrates how data points are densely clustered within and around the GGH, particularly around the Greater Toronto Area (GTA), reflecting significant activity, infrastructure, or population focus in this region. This visual is pivotal for understanding spatial trends, regional planning, and resource allocation within Ontario's most dynamic urban corridor.

- **Top and bottom FSAs by adoption**

Figure 4 and Figure 5 show cumulative EV registrations from Q4 2023 to Q1 2025 for Ontario's top and bottom ten FSAs, highlighting differences in regional adoption trends in the fastest-growing and slow-growing areas. Each line represents an FSA, and the legend lists FSAs with their primary city or region. Flat, overlapping lines reveal minimal growth in certain areas, emphasizing ongoing disparities and stagnant EV adoption within Ontario.

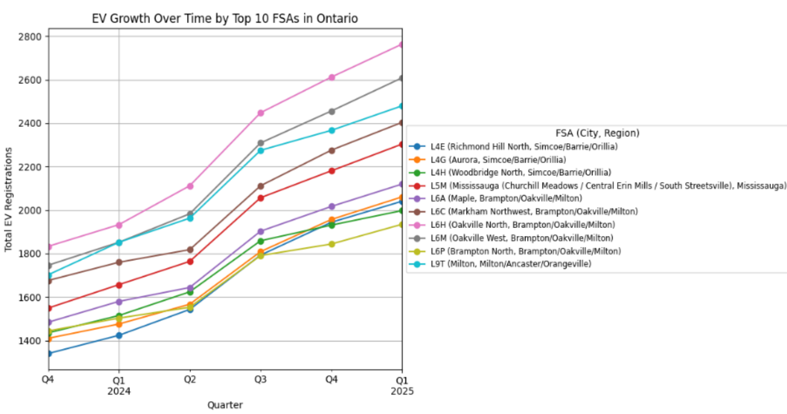


Figure 4. *EV Growth Over Time in the Top 10 Ontario FSAs.*
Created in Jupyter Lab..

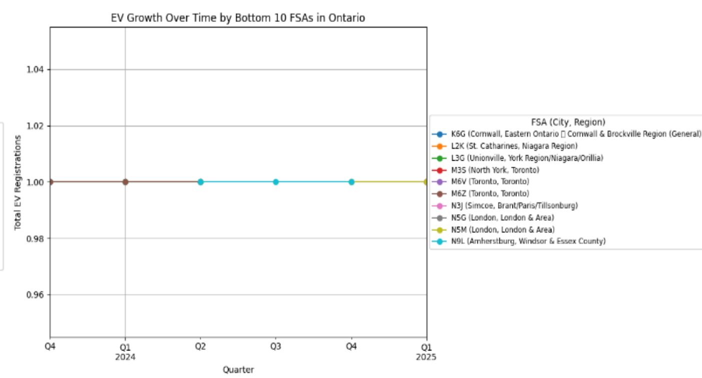


Figure 5. *EV Growth Over Time in the Bottom 10 Ontario FSAs.* Created in Jupyter Lab.

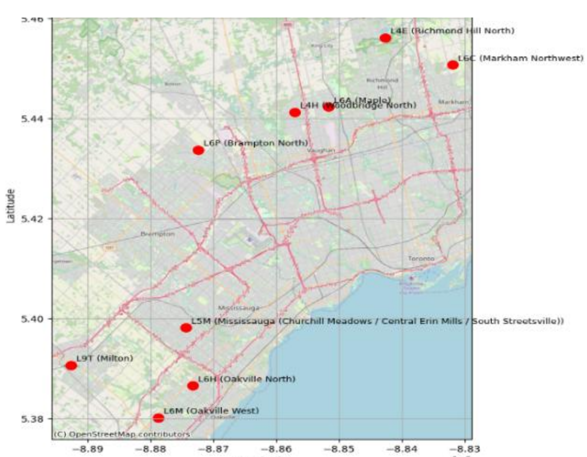


Figure 6. *Top 10 Ontario FSAs by Electric Vehicle (EV) Registrations*
Created by Myunghee in Jupyter Lab.

More rapid growth is in high-income or suburban regions, but strong adoption is evident province-wide.

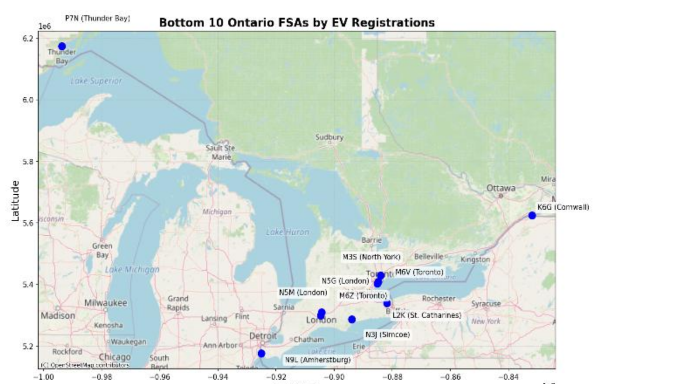


Figure 7. *Bottom 10 Ontario FSAs by Electric Vehicle (EV) Registrations.*
Created in Jupyter Lab.

Most slow growth lie outside the GHH, where EV adoption is higher. The clustering of low-registration FSAs in rural areas highlights ongoing regional disparities likely driven by factors like population density, infrastructure, and economic development.

3. Clustering Analysis

3.1 Model Design and Architectures

K-Means clustering was used to group FSAs with similar EV adoption patterns.
Using the engineered quarterly dataset, each FSA was assessed on three key metrics:

1. **Adoption Level:** Total registered EVs in Q1 2025, indicating market maturity.
2. **Growth Rate:** Percentage increase in EVs from Q2 2023 to Q1 2025, showing adoption speed.
3. **BEV Preference:** Average share of BEVs in total EVs, reflecting preference for BEVs over PHEVs.

After standardizing features for equal weighting, K-Means segmented the FSAs into four clusters.

3.2 Justification for Model Choice

K-Means clustering was chosen for its efficiency and ease of interpretation when grouping FSAs based on standardized adoption level, growth rate, and BEV preference. Its use of Euclidean distance works well with standardized data, enabling clear identification of distinct regional patterns. Alternative methods like hierarchical clustering or DBSCAN were considered but dismissed due to higher complexity and less suitable clustering goals. Overall, K-Means offered a good balance of simplicity, speed, and meaningful segmentation aligned with the project objectives.

3.3 Interpretation of Clusters

The analysis revealed four unique profiles of EV adoption across the province. The characteristics of each cluster are summarized in the table below, followed by a detailed interpretation and supporting visualizations.

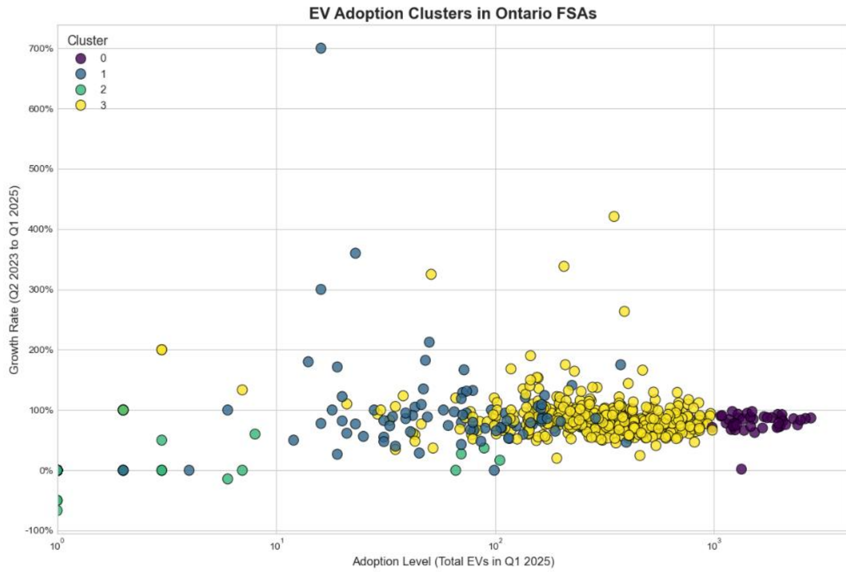
Table 2. Adoption, Growth, and BEV Preference Profiles Across Market Clusters in Q1 2025

Cluster	Profile	Avg. Adoption Level (Q1 2025)	Avg. Growth Rate	Avg. BEV Preference	Number of FSAs
0	Lagging & Rural Regions	89	89.42%	46.01%	101
1	High-Growth, BEV-Dominant Suburbs	1637	79.39%	78.11%	41
2	New/Stagnant Micro-Markets	9	2.39%	94.24%	46
3	Established Urban & Commuter Zones	365	88.00%	70.88%	394

- ***Cluster 1 – High-Growth, BEV-Dominant Suburbs:*** Highest adoption levels, strong growth, and a clear BEV preference (78%), representing Ontario's most mature EV markets.
  Concentrated in affluent GGH suburbs such as Oakville (L6H), Milton (L9T), and parts of Vaughan and Richmond Hill, these areas likely benefit from higher incomes, at-home charging access, and established public infrastructure—supporting H2 that certain FSAs adopt EVs faster due to favorable conditions.
- ***Cluster 3 – Established Urban & Commuter Zones:*** Largest group, with moderate-to-high adoption, the highest growth rate, and BEV preference matching provincial trends.
  This cluster Includes major urban and commuter hubs such as parts of Toronto, Mississauga, and Ottawa. Strong, steady growth suggests EV adoption is firmly in the mainstream across these major population centers.
- ***Cluster 0 – Lagging & Rural Regions:*** Low adoption, moderate growth, and lower BEV preference (46%), indicating greater reliance on PHEVs.
  Predominantly rural and northern Ontario FSAs, where long travel distances, limited charging infrastructure, and differing vehicle needs hinder adoption—supporting H3 on the impact of regional disparities.
- ***Cluster 2 – New/Stagnant Micro-Markets:*** Very low EV counts and little to no growth, yet the highest BEV preference—likely due to a few early adopters.
  These micro-markets have strong future potential but currently lack catalysts such as infrastructure, incentives, or local demand.

3.4 Visual Analysis of Clusters

To explore how each feature shapes the clusters, several visualizations were created.



Scatter Plot of Adoption vs. Growth

This plot visualizes the clusters based on their adoption levels and growth rates. It clearly distinguishes the four groups: Cluster 1 occupies the top-right (high adoption, high growth), while Cluster 2 is situated at the bottom-left (low adoption, low growth).

Figure 8. *EV Adoption Clusters in ON FSAs*

Feature Distribution Box Plots

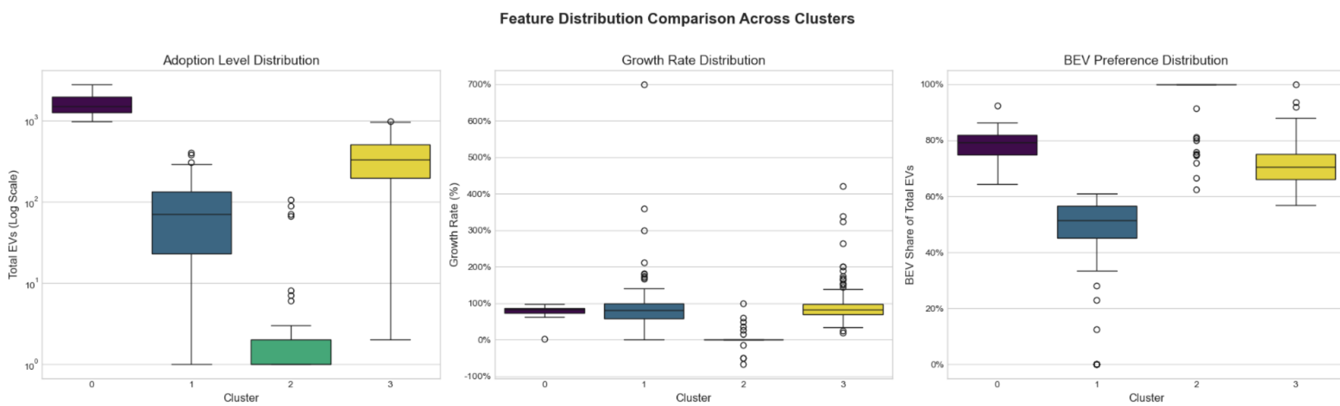
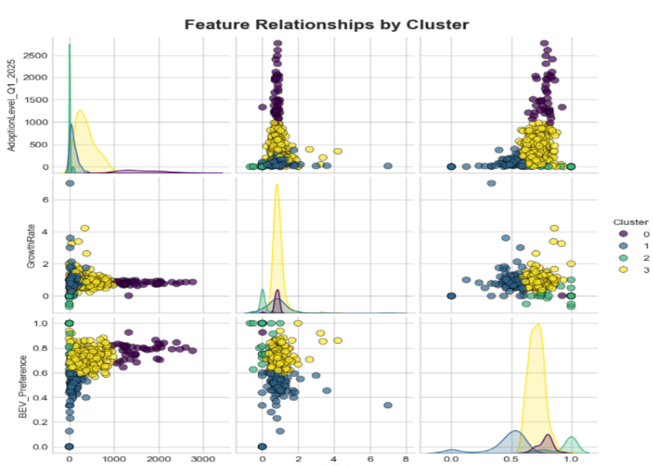


Figure 9. *Feature Distribution Comparison Across Clusters*

Plots in Figure 9 compare the three key features across clusters, confirming their profiles. For example, the “Adoption Level” plot highlights Cluster 1’s lead, while the “Growth Rate” plot shows Cluster 3’s strong performance.



Pair Plot of All Features

The pair plot provides a comprehensive view of how all features interact. The distinct groupings visible in the scatter plots confirm that the K-Means algorithm successfully identified FSAs with genuinely different adoption profiles.

Figure 10. *Feature Relationships by Cluster*

4. Predictive Modeling

4.1 Model Design and Architectures

We explored two predictive models from scikit-learn to forecast quarterly EV adoption by FSA in Ontario: **Linear Regression** and **Random Forest**.

We considered **more complex models** (e.g., ARIMA, LSTM neural networks, and the ensemble of both) but found the limited temporal data (only 8 quarters) insufficient for effective sequential learning.

Features used in all models included:

- Previous quarter's total EVs (Prev_TotalEV)
- Previous quarter's BEVs (Prev_BEV)
- Previous quarter's PHEVs (Prev_PHEV)
- Previous quarter's BEV share (Prev_BEV_Share)
- Quarter indicator (Quarter_Encoded)

These features reflect recent momentum and regional trends. Models predicted the current quarter's total EV registrations based on prior quarter data.

4.2 Justification for Final Model Choices

We chose these models because Linear Regression offers a clear benchmark, while Random Forest balances flexibility and interpretability.

- **Linear Regression** was used as a baseline due to its simplicity, interpretability, and efficiency. However, it assumes a linear relationship between features and target, which is likely too restrictive given EV adoption is influenced by complex, nonlinear factors such as policy changes and infrastructure.
- **Random Forest**, a tree-based ensemble model, was selected as the primary model for its ability to capture nonlinear relationships and feature interactions, robustness to outliers, and strong performance with tabular data.

4.3 Model Training and Hyperparameter Tuning

The Random Forest model was optimized using GridSearchCV with 5-fold cross-validation. Tuning involved an initial broad search followed by fine-grained parameter adjustments. To avoid overfitting, we used cross-validation with negative RMSE as the scoring metric. The final parameters of the tuned model used n_estimators = 100, max_depth = None, and min_samples_split = 2.

The mean cross-validated R^2 was approximately 0.97, with minimal training-test score gaps, indicating balanced model fit without overfitting.

4.4 Model Evaluation and Results Comparison

We used the following metrics to evaluate our models:

- Root Mean Squared Error (RMSE): penalizes large errors more heavily
- Mean Absolute Error (MAE): gives the average prediction error
- R^2 (coefficient of determination): measures the proportion of variance explained

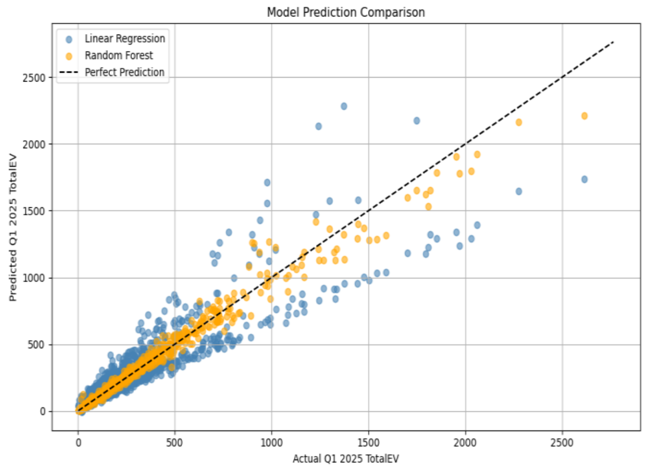


Figure 11 shows actual vs. predicted values for both models. While both perform well, the Random Forest's predictions are more tightly clustered around the 1:1 line, especially at higher EV counts, indicating better performance overall.

Figure 11*: Prediction of Linear Regression vs. Random Forest*

Evaluation metrics in Table 3 further confirmed that the Random Forest model dramatically outperformed Linear Regression, reducing RMSE by over 26,000 and MAE by 85, with R^2 improving from 0.77 to nearly 0.98. This highlights the importance of capturing nonlinear relationships in EV adoption data. (Table 3)

Although tuning slightly reduced the Random Forest's performance, the change was negligible. Its main benefit was improved generalizability, ensuring the model avoids overfitting and performs consistently across different subsets.

Table 3. Model Performance Metrics: RMSE, MAE, and R^2 for Regression Models

Model	RMSE	MAE	R^2
Linear Regression	29,211.10	114.44	0.7722
Random Forest	2,931.24	28.77	0.9771
Tuned RF (final)	2,963.94	29.02	0.9769

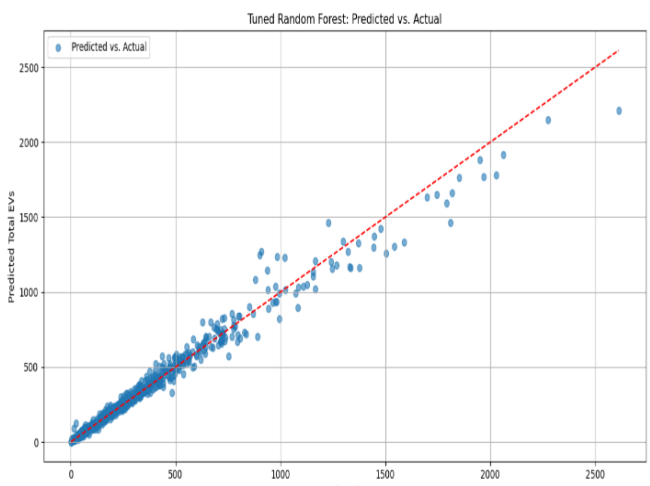


Figure 12 shows that the tuned model's predictions closely follow the actual values, particularly in the mid and upper range of EV counts.

Figure 12*: Tuned Random Forest: Predicted vs. Actual*

4.5 Additional Model Experiments: LSTM, ARIMA, and Ensemble

We also tested time-series forecasting models. All three models performed poorly on the test set, with negative R^2 values indicating worse performance than a mean baseline. ARIMA was the least inadequate, though still unreliable. LSTM showed the worst performance, likely due to limited data and model tuning challenges. The ensemble improved over LSTM alone but did not surpass ARIMA, reflecting data quantity and quality limitations rather than model architecture. (Table 4)

Given the poor performance, simpler models or baselines may provide more reliable forecasts initially. Employing rolling validation can offer better insights into model generalization. Advanced ensembling or weighted model combinations might yield marginal improvements, but cannot fully compensate for core data issues. Enhancing data quality, simplifying models, and using robust validation are essential next steps.

Table 4. Model Performance of forecasting models

Model	RMSE	MAE	R^2
LSTM	77,482.88	76,974.79	-65.07
ARIMA	45,373.89	43,784.34	-21.66
Ensemble of LSTM and ARIMA ARIMA	61,265.58	60,379.56	-40.31

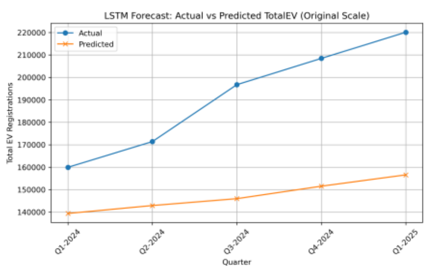


Figure 13*: LSTM Forecast: Actual vs Predicted TotalEV (Original Scale)*
Created in Jupyter Lab.

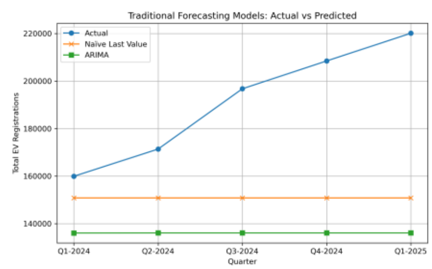


Figure 14*: Traditional Forecasting Models: Actual vs Predicted*
Created in Jupyter Lab.

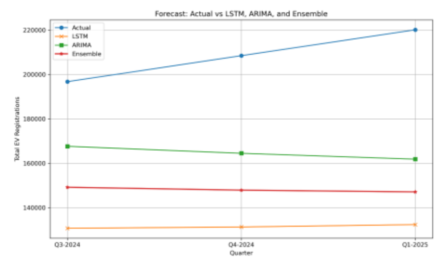


Figure 15*: Forecast: Actual vs LSTM, ARIMA, and Ensemble*
Created in Jupyter Lab.

4.6 Future EV Adoption Forecast

Using the tuned Random Forest model, we forecasted EV adoption for Q2 2025 across Ontario FSAs. The top predicted FSAs are concentrated in suburban and urban regions of the Greater Toronto Area, including Brampton (L7A, L6P, L6R, L6T, L6S, and L6Z), Oakville (L6M and L6H), and Mississauga (L5M), reflecting dense populations, affluence, and supportive charging infrastructure.

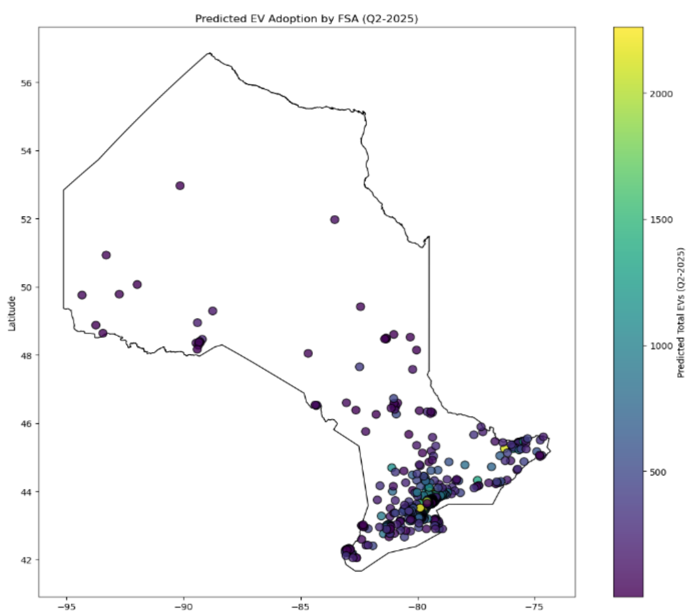


Figure 16. Map of Ontario by EV Adoption Prediction

Table 5. Top 10 Predicted Total EV Counts by Postal Code, City, and Region

Postal Code	City	Region	Predicted Total EVs
L6M	Oakville	Halton	12760
L6H	Oakville	Halton	11946
L7A	Brampton	Peel	11692
L6P	Brampton	Peel	11479
L7A	Brampton	Peel	11471
L6R	Brampton	Peel	11365
L6T	Brampton	Peel	11322
L6S	Brampton	Peel	11307
L5M	Mississauga	Peel	11295
L6Z	Brampton	Peel	11284

5. Conclusions

5.1 Key Findings

This project combined spatial clustering and predictive modelling to EV adoption patterns across Ontario's FSAs.

- **Distinct Regional Profiles:** K-Means clustering revealed four clear EV adoption profiles:
    - High-growth, BEV-dominant suburbs
    - Established urban commuter zones
    - Lagging rural regions
    - New or stagnant micro-markets
- **Geographic Disparities:** Affluent suburban and urban FSAs, particularly in the Greater Golden Horseshoe, show higher adoption levels and faster growth than rural areas.
- **Technology Preference Differences:** Urban/suburban FSAs tend toward BEV dominance, while rural areas rely more heavily on PHEVs.
- **Strong Predictive Performance:** A tuned Random Forest model achieved $R^2 \approx 0.98$, outperforming both Linear Regression and complex time-series models, enabling accurate short-term adoption forecasts.
- **Future Hotspots:** Forecasts for Q2 2025 place several Brampton, Oakville, and Mississauga FSAs among the top projected growth areas.

5.2 Hypothesis Validation

- **H1: Overall Growth Hypothesis — Supported.** EV adoption increased consistently across most FSAs between Q2 2023 and Q1 2025, with growth rates ranging from 5.6% to 14.8% per quarter.
- **H2: Regional Disparity Hypothesis — Supported.** Affluent suburban and urban commuter FSAs, especially in the Greater Golden Horseshoe, displayed markedly higher adoption levels and faster growth, supported by income, infrastructure, and policy advantages.
- **H3: Technology Preference Hypothesis — Supported.** Urban and suburban clusters showed stronger BEV dominance, while rural and lagging regions had higher PHEV reliance, likely reflecting differences in driving range needs and charging access.
- **H4: Predictive Modelling Hypothesis — Supported.** The Random Forest model accurately forecasted short-term EV adoption trends at the FSA level, enabling targeted identification of high-growth areas.

5.3 Insights from Data

The findings underscore that EV adoption in Ontario is uneven, influenced by socioeconomic, infrastructural, and geographic factors. High-growth urban and suburban clusters may require accelerated investment in charging infrastructure and grid capacity, while slower-growth rural areas could benefit from tailored incentive programs and hybrid-friendly policies. The gradual decline in the BEV/PHEV ratio suggests shifting consumer preferences and potential constraints in charging accessibility.

5.4 Recommendations for Future Work

- **Expand Data Horizon:** Collect more historical data to improve model reliability, particularly for sequential forecasting methods.
- **Integrate Additional Variables:** Include socioeconomic, demographic, and infrastructure data to enhance model interpretability.
- **Refine Validation Strategies:** Use rolling cross-validation and strong baselines to ensure robust model generalization.
- **Policy Tailoring:** Develop differentiated incentives—accelerated infrastructure in urban/suburban hotspots and hybrid-friendly programs in rural areas.

References

Innovation, Science and Economic Development Canada. (n.d.). *Forward sortation area (FSA) definition*. Government of Canada. https://ised-isde.canada.ca/site/office-superintendent-bankruptcy/en/statistics-and-research/forward-sortation-area-fsa-and-north-american-industry-classification-naics-reports/forward-sortation-area-definition

Esri. (n.d.). *Ontario municipal boundaries* [ArcGIS Online item]. ArcGIS Online. https://www.arcgis.com/home/item.html?id=f8919b3b0ac74039a0ddb1657128962d

Statistics Canada. (2023, March 29). *Population and dwelling counts: Canada and forward sortation areas (Table 98-10-0019-01)*. Statistics Canada. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=981000

Wikipedia. (n.d.). *List of postal codes of Canada: K*. In *Wikipedia*. Retrieved July 15, 2025, from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_K

Wikipedia. (n.d.). *List of postal codes of Canada: L*. In *Wikipedia*. Retrieved July 15, 2025, from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_L

Wikipedia. (n.d.). *List of postal codes of Canada: M*. In *Wikipedia*. Retrieved July 15, 2025, from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Wikipedia. (n.d.). *List of postal codes of Canada: N*. In *Wikipedia*. Retrieved July 15, 2025, from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_N

Wikipedia. (n.d.). *List of postal codes of Canada: P*. In *Wikipedia*. Retrieved July 15, 2025, from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_P

Appendices

1. Data:
   - Q2-2023 - Q1-2025.csv
   - ontario_fsa_coordinates_with_city_region.csv

2. Jupyter Notebook:
   - [Group_4_Final_Notebook.ipynb]

3. Visualization Map:
   - ev_golden_horseshoe_q1_2025.html

4. Shapefile/GeoJSON Formats:
   - Ontario Simple Boundary.shp
   - Lfsa000b21a_e.shp
   - Ontario_boundary.geojson
   - Golden Horseshoe.geojson

5. Figure:
   - 1. ontario_ev_growth_rate_plot.png
   - 2. 2-Quarter Rolling Avg.png
   - 3. ev_golden_horseshoe_q1_2025.html
   - 4. Top10 FSA.png
   - 5. Bottom10_FSA.png
   - 6. Top10_map.png
   - 7. Bottom10_map.png
   - 8. Cluster Adoption Levels.png
   - 9. distribution_of_EV_clusters.png
   - 10. Pair_Plot_all_EV_Features.png
   - 11. Model_Prediction_Comparison.png
   - 12. Tuned_RF_vs.png
   - 13. lstm_forecast.png
   - 14. Lstm_tfm.png
   - 15. Ensemble.png
   - 16. Predicted Total EVs (Q2-2025).png