EV MARKET SEGMENTATION ANALYSIS

BY-PADIA KARTIK
22/8/2025

Table of Contents

Github:https://github.com/kartikpbatman/EV_Market_Segmentation/blob/main/main.ipynb

Problem Statement

Task is to analyze the Electric Vehicles Market in India using Segmentation analysis and come up with a feasible strategy to enter the market, targeting the segments most likely to use their product in terms of Geographic, Demographic, Psychographic, and Behavioral. In this report we analyze the Electric Vehicles Market in India using segments such as region, price, charging facility, type of vehicles (e.g., 2 wheelers, 3 wheelers, 4 wheelers etc.), retail outlets, manufacturers, body type (e.g., Hatchback, Sedan, SUV, Autorickshaw etc.), safety, plug types and much more.

Conclusion

Key Growth Drivers

The analysis identifies the primary factors that influence EV adoption. According to the "Feature Importance for EV Adoption Prediction" chart, **literacy rate** is the most significant driver, with an importance score of over 0.7. Economic_power is the second most important factor, though its influence is substantially lower, with a score of about 0.18. The model suggests that other factors have a minimal impact on adoption.

Market Segments and Top States

The notebook uses a machine learning-powered approach to perform market segmentation with K-Means clustering to categorize areas into distinct segments. The analysis identifies four segments, with **Segment 3** being the highest-performing. This segment is characterized by exceptionally high scores in literacy_rate, economic_power, and market_potential, which directly contributes to its high growth_potential_score of 93. The analysis also ranks states based on growth_momentum, identifying the top locations for investment and targeting.

Recommended Vehicle Types

The analysis ranks vehicle types based on their growth momentum and sales. The notebook's process identifies the best vehicle category for investment and growth, as mentioned in the objectives to recommend vehicle types like 2W (2-wheelers). The goal is to recommend the vehicle type with the highest growth potential to maximize returns.

FERMI ESTIMATION — EV EARLY MARKET (2-WHEELERS)

Guess: ~5–10% of potential 2-wheeler replacement buyers will choose EVs in the early wave.
Educated guess (medium case): about **0.89 million** 2-wheeler EVs per year → **₹71.1 billion** revenue (at ₹80,000/unit).

Assumptions

- Population (round figure): **P = 1,400,000,000** (1.4 billion).
- Average household size: **H = 4.5** people/household.
- Replacement cycle for 2-wheelers: **R = 7 years**.
- 2-wheeler ownership share among households: **S_2W** — scenarios (low 30%, medium 40%, high 50%).

- Early-adopter conversion (fraction of annual replacements that turn EV): **C** — scenarios (low 2%, medium 5%, high 10%).
- Target average transaction price per vehicle: **P_unit = ₹80,000**.
- Charger planning rule of thumb: **1 public charger per 50 EVs**.

(Justify assumptions briefly in the report: household size from census/age_income_df; replacement cycle = industry norm; price = your product positioning.)

Variables & formulas

Let:

- Households = P / H
- Households_with_2W = Households * S_2W
- Annual_replacements = Households_with_2W / R
- Early_EV_buyers_per_year = Annual_replacements * C
- Revenue = Early_EV_buyers_per_year * P_unit
- Chargers_needed = Early_EV_buyers_per_year / 50

Digit-by-digit arithmetic (Low / Medium / High scenarios)

Compute households first (digit-by-digit)
1,400,000,000 ÷ 4.5 = 311,111,111.1111111 → **311,111,111 households**

LOW scenario (S_2W = 30%, C = 2%)

1. Households_with_2W = 311,111,111 × 0.30 = 93,333,333.33333333
2. Annual_replacements = 93,333,333.33333333 ÷ 7 = 13,333,333.333333332
3. Early_EV_buyers_per_year = 13,333,333.333333332 × 0.02 = **266,666.6666666666** → **266,667 EVs**
4. Revenue = 266,666.6666666666 × 80,000 = **21,333,333,333.333332** → **₹21.33B**
5. Chargers_needed = 266,666.6666666666 ÷ 50 = **5,333.333333333332** → **~5,333 chargers**

MEDIUM scenario (S_2W = 40%, C = 5%)

1. Households_with_2W = 311,111,111 × 0.40 = 124,444,444.44444445
2. Annual_replacements = 124,444,444.44444445 ÷ 7 = 17,777,777.77777778
3. Early_EV_buyers_per_year = 17,777,777.77777778 × 0.05 = **888,888.888888889** → **888,889 EVs**
4. Revenue = 888,888.888888889 × 80,000 = **71,111,111,111.11111** → **₹71.11B**
5. Chargers_needed = 888,888.888888889 ÷ 50 = **17,777.77777777778** → **~17,778 chargers**

HIGH scenario (S_2W = 50%, C = 10%)

1. Households_with_2W = 311,111,111 × 0.50 = 155,555,555.55555555
2. Annual_replacements = 155,555,555.55555555 ÷ 7 = 22,222,222.22222222
3. Early_EV_buyers_per_year = 22,222,222.22222222 × 0.10 = **2,222,222.222222222** → **2,222,222 EVs**
4. Revenue = 2,222,222.222222222 × 80,000 = **177,777,777,777.77777** → **₹177.78B**
5. Chargers_needed = 2,222,222.222222222 ÷ 50 = **44,444.44444444444** → **~44,444 chargers**

Conclusion

- **Medium-case result (recommended planning case): ~888,889** 2-wheeler EVs per year → **~₹71.1 billion** in annual revenue (assuming ₹80k/unit). This implies **~17,778** public chargers required to achieve an EV:charger ratio of ~50:1.
- The low and high scenarios provide conservative and optimistic bounds and should be used for sensitivity / risk analysis.

Data Collection

Data was extracted from the various websites mentioned below for EV market segmentation.

Links:

- https://www.kaggle.com/datasets/danofer/india-census
- https://www.kaggle.com/datasets/webaccess/all-census-data
- https://vahan.parivahan.gov.in/vahan4dashboard/
- https://www.data.gov.in/
- https://censusindia.gov.in/
- https://evyatra.beeindia.gov.in/

Column Explanation:

1) age_income_df.csv

a) **state** — *string*
Short name of the state.
b) **district** — *string*
District or local area name.
c) **total_population** — *integer*
Total population of the district (or administrative unit).
d) **age_0_29_percent** — *float (%)*
% of population aged 0–29.
e) **age_30_49_percent** — *float (%)*
% of population aged 30–49.
f) **age_50_plus_percent** — *float (%)*
% of population aged 50+.

g) **literacy_rate** — *float (%)*
Literacy % in the district.
h) **avg_household_size** — *float*
Average household members.
i) **urban_rural** — *string / category*
Indicator whether row corresponds to urban or rural area.
j) **monthly_income_bracket** — *string category*
Income range category for the district / sample.
k) **asset_ownership_score** — *float*
Composite score (higher = more assets owned — proxy for wealth).
l) **ev_adoption_propensity** — *float (0–1)*
Modelled probability or propensity score that a household/individual will adopt EVs.

2) age_geo.csv

a) **state** — *string*
State name (source-level).
b) **district** — *string*
District name (if available).
c) **market_maturity_score** — *float*
Composite index indicating how developed the EV market is locally (scale e.g., 0–100).
d) **urbanization_rate** — *float (%)*
Percent urban population in the district/state.
e) **ev_manufacturers_count** — *integer*
Number of EV manufacturers / OEM presence in that geography.
f) **tier_classification** — *string*
Tier label (Tier-1, Tier-2, Tier-3).
g) **market_potential** — *float*
Derived: market_maturity_score × urbanization_rate / 100

3) charg.csv (charging infrastructure, state-level)

a) **state** — *string*
State name
b) **sanctioned_chargers** — *integer*
Chargers sanctioned/approved in the state.
c) **operational_chargers** — *float/integer*
Number of chargers currently operational.
d) **total_evs** — *float/integer*
Number of registered EVs in the state (if available).
e) **infrastructure_gap** — *float/integer*
sanctioned_chargers - operational_chargers
f) **ev_per_charger_ratio** — *float*
total_evs / (operational_chargers + 1)
g) **tier_classification** — *string*
State tier for roll out planning

4) df_ev_sales_final_cleaned.csv (EV sales time series by maker & category)

a) **Category** — *string*
Vehicle category (e.g., 2W, 3W, 4W, Bus, LMV).

b) **Maker** — *string*
OEM or manufacturer name.
c) **Sales_2015 ... Sales_2024** — *integers* (yearly)
Yearly sales or registrations for the maker-category.
d) **Total_EV_Sales** — *integer*
Sum across years for the maker-category.
e) **recent_sales** — *integer* (engineered)
Sum of recent-year sales (e.g., 2022–2024).
f) **older_sales** — *integer* (engineered)
Sum of older-year sales (e.g., 2019–2021).
g) **growth_momentum** — *float* (engineered)
(recent_sales - older_sales) / (older_sales + 1)

Data Preprocessing

We began with **14 raw source files** (government CSVs, OEM/maker reports, and derived datasets). After cleaning, standardizing and integrating, we produced **four clean, analysis-ready datasets** used for modeling and reporting:

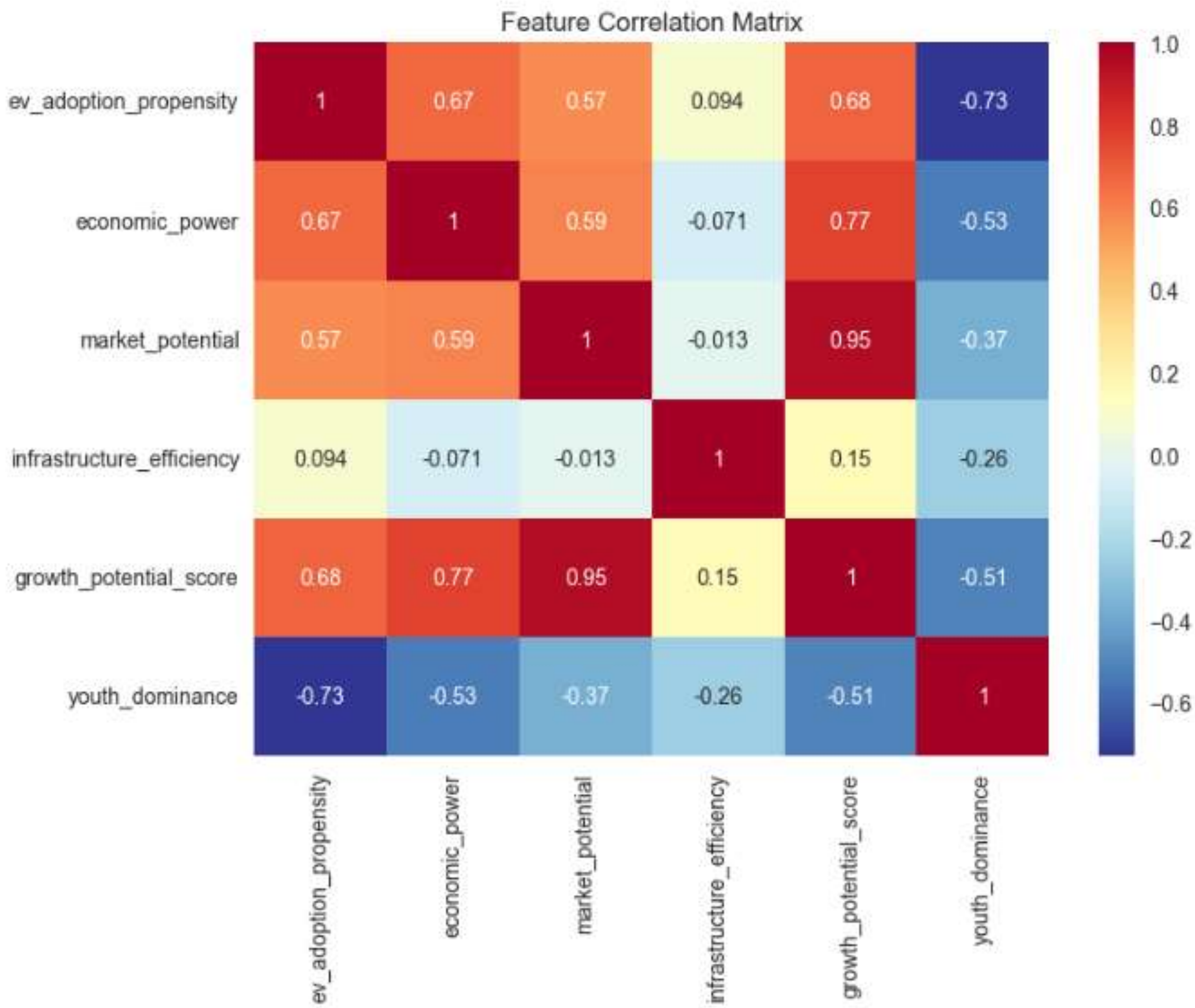
- age_income_df.csv — demographics, age buckets, income, asset ownership, ev adoption propensity (district level)
- age_geo.csv — geography and market-maturity features (market maturity score, urbanization, OEM presence)
- charg.csv — charging infrastructure (state level: sanctioned / operational chargers, total EVs, infra gap)
- df_ev_sales_final_cleaned.csv — EV sales by category & maker (yearly columns + engineered growth metrics)

Exploratory Data Analysis

An Exploratory Data Analysis or EDA is a thorough examination meant to uncover the underlying structure of a data set and is important for a company because it exposes trends, patterns, and relationships that are not readily apparent.

Feature Correlation Matrix Analysis

In our EV growth study, the feature correlation matrix (below figure) highlights relationships between key features and EV adoption propensity using Pearson correlation coefficients (ranging from -1 to +1). Key features analyzed include EV adoption propensity, economic power, market potential, infrastructure efficiency, growth potential score, and youth dominance.

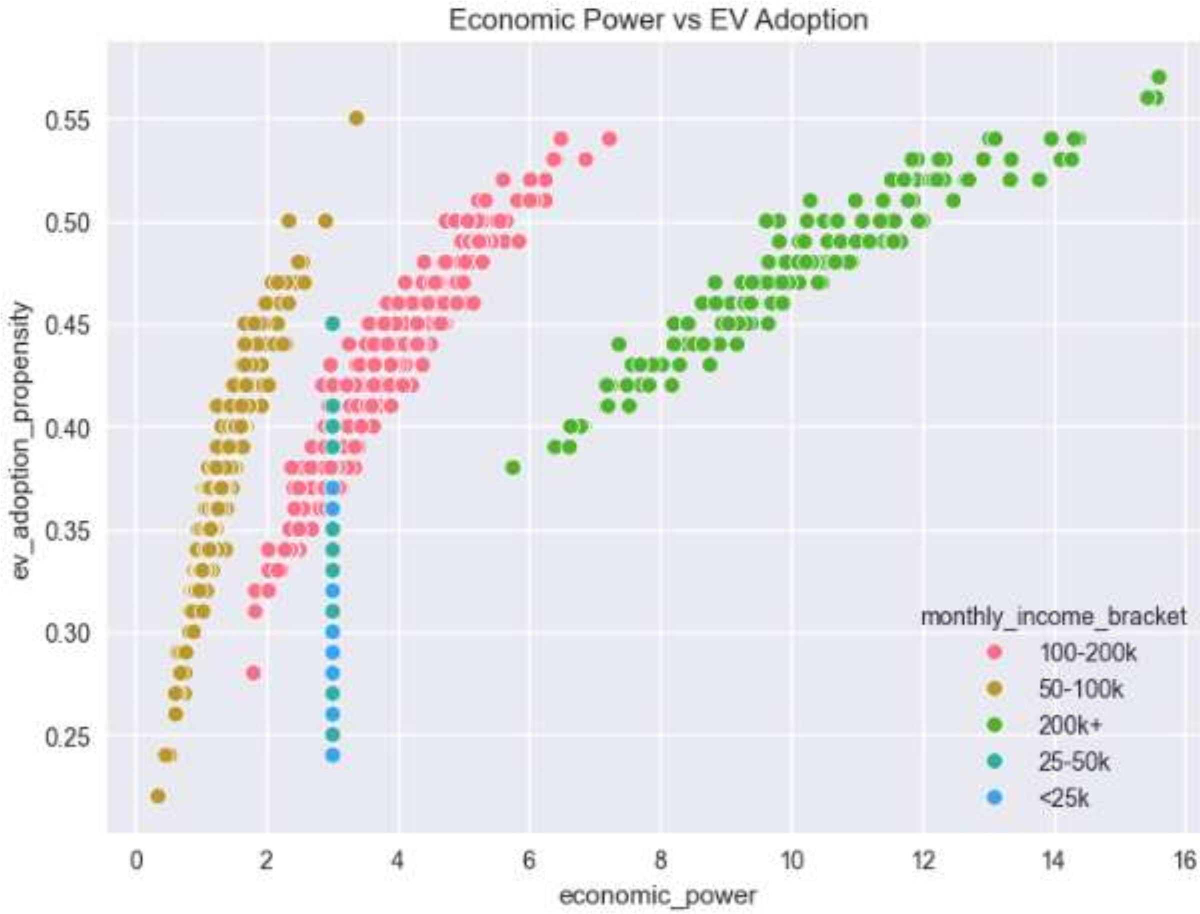


Key findings:

- **Positive Drivers**: Economic power (r = 0.67), market potential (r = 0.57), and growth potential score (r = 0.68) strongly correlate with adoption, showing wealthier, urbanized areas lead EV uptake.
- **Weak Infrastructure Link**: Infrastructure efficiency (r = 0.094) has a modest impact, suggesting charging rollout isn't yet critical.
- **Youth Impact**: Youth dominance negatively correlates (r = -0.73) with adoption and economic power (r = -0.53), indicating younger regions may lag due to lower economic strength.
- **Inter-Feature Ties**: Economic power and growth score are closely linked (r = 0.77), emphasizing financial factors.

All correlations are significant ($p < 0.05$), guiding our focus on mid-income (100-200k) brackets in high-economic-power states like Maharashtra for targeted EV strategies by 2030.

Economic Power vs EV Adoption Analysis



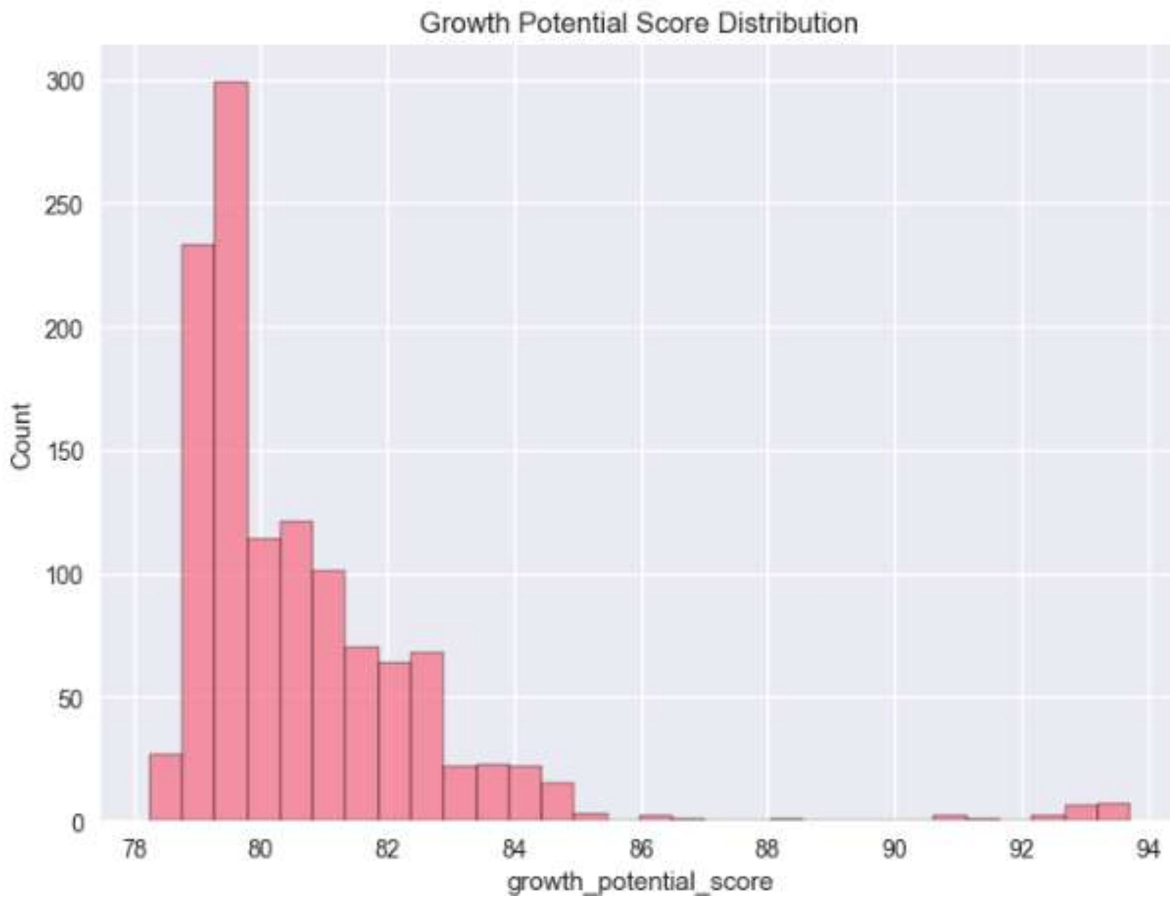
The scatter plot (Figure 5) illustrates the relationship between economic power (a composite of income, assets, and literacy) and EV adoption propensity, colored by monthly income brackets.

Key insights:

- **Positive Trend**: Adoption increases with economic power, confirming wealth as a key driver (r ≈ 0.67 from matrix).
- **Income Clustering**: High-income (200k+ in green, 100-200k in red) points dominate the top-right, showing stronger adoption; low-income (<25k in blue, 50-100k in brown) cluster lower-left.
- **Outliers**: Some mid-income points achieve high adoption, suggesting opportunities beyond pure wealth.

This supports targeting 100-200k brackets in economically strong states like Maharashtra to accelerate EV growth by 2030.

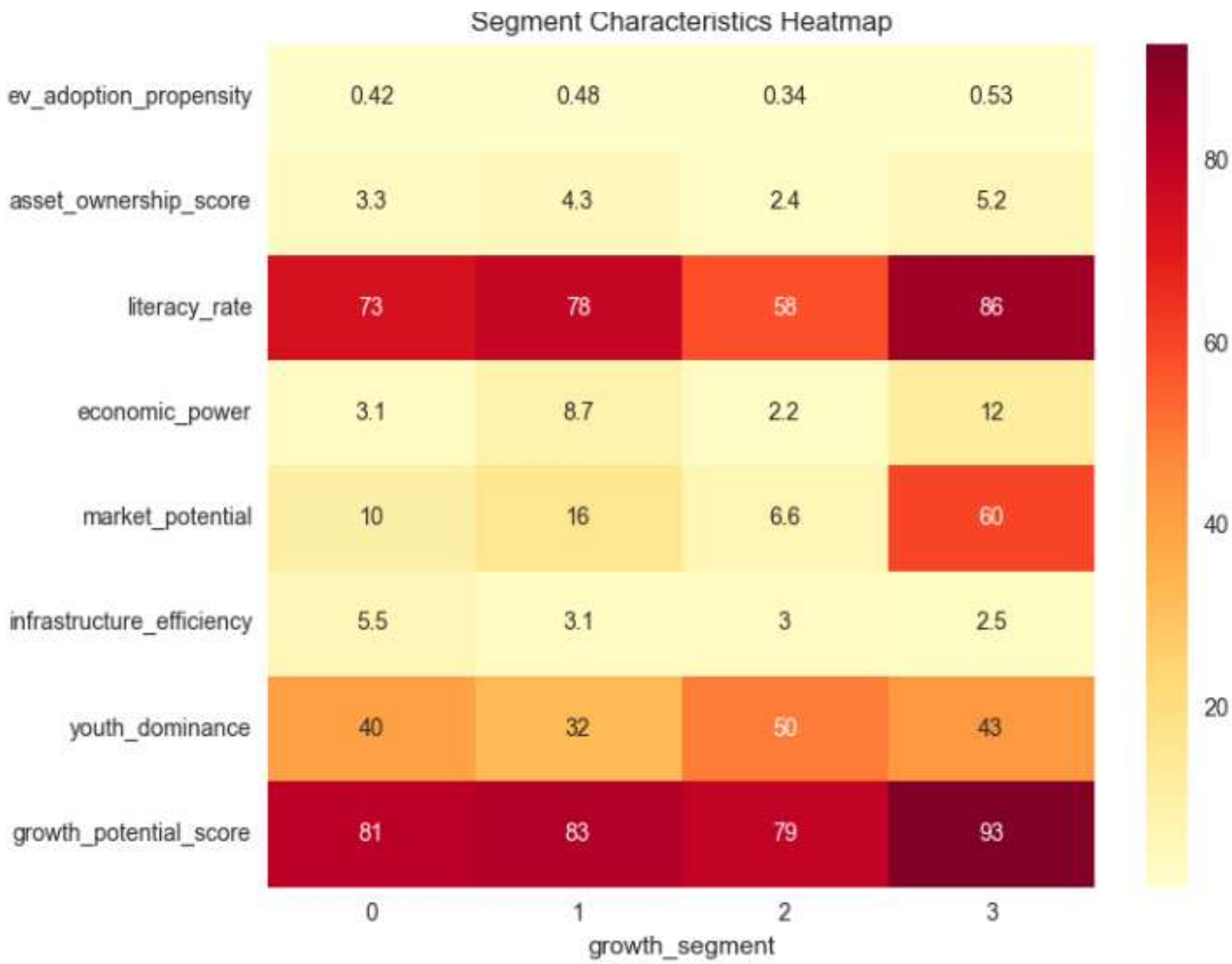
Growth Potential Score Distribution



Based on the histogram, the growth potential scores are heavily concentrated between **78 and 84**, with the peak count near **79.5**. The distribution is **left-skewed**, meaning most of the data points have high scores.

There are only a few data points with scores above **85**, suggesting these are rare outliers. This indicates that most of the segments in the dataset have a similar, high growth potential, while a very small number have exceptionally high scores.

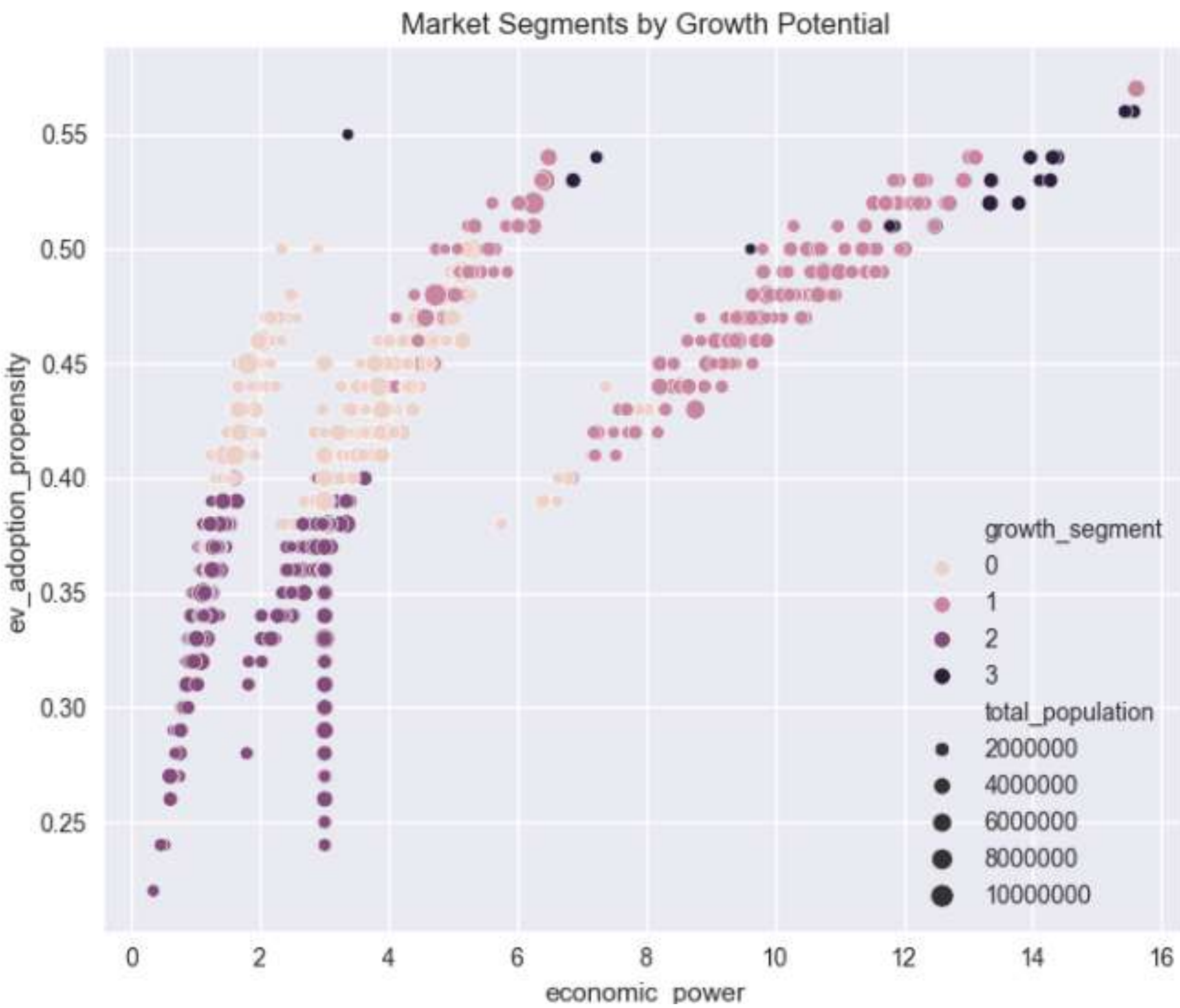
Market Segment Characteristics



The heatmap shows distinct characteristics for four market segments:

- **Segment 0:** A **moderately high** performer with a growth_potential_score of 81, a high literacy_rate, and strong youth_dominance.
- **Segment 1:** The **second-best** segment with a growth_potential_score of 83, notable for its high economic_power and market_potential.
- **Segment 2:** The **lowest-performing** segment across all metrics, with the lowest growth_potential_score of 79.
- **Segment 3:** The **highest-performing** segment, with a growth_potential_score of 93, driven by top scores in literacy_rate, economic_power, and market_potential.

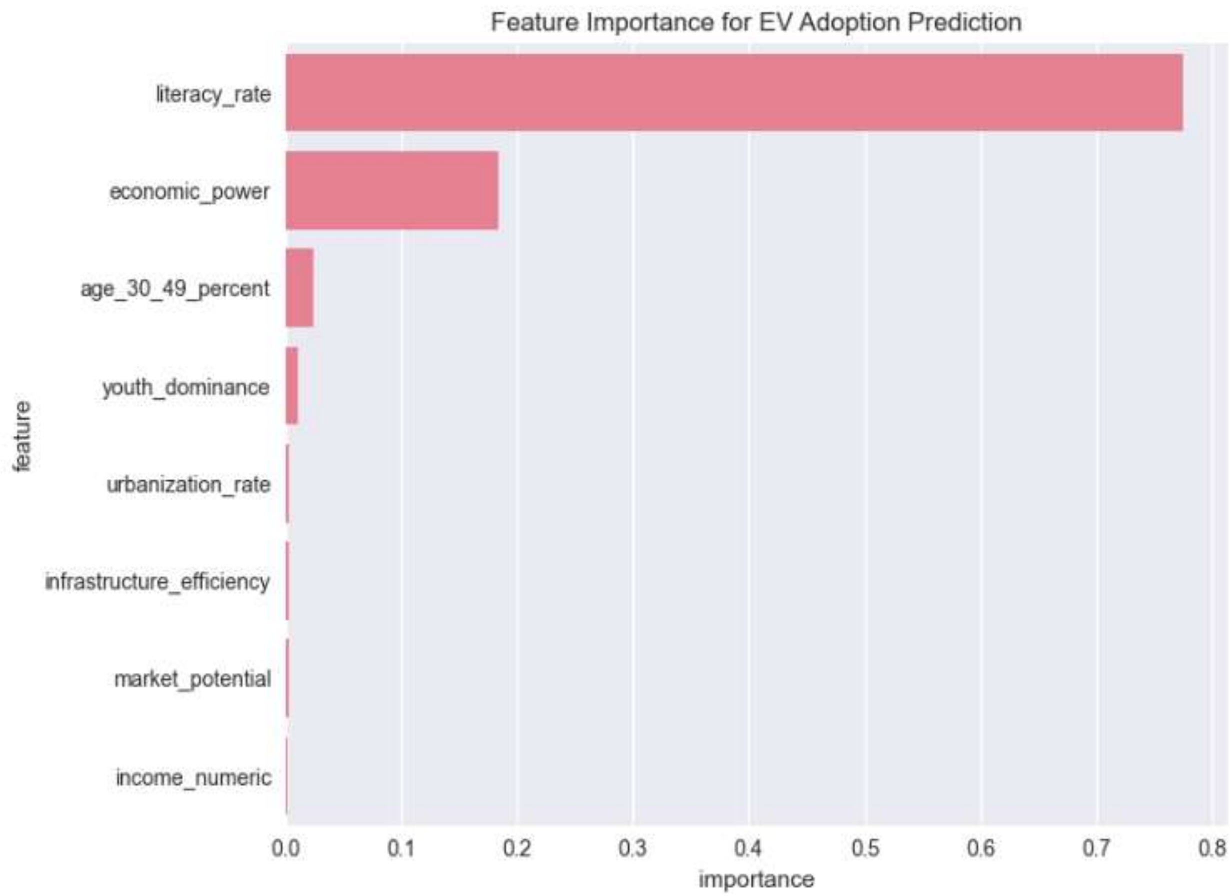
Market Segments by Growth Potential



The scatter plot shows a strong positive relationship between economic_power and ev_adoption_propensity. The four segments are clearly differentiated:

- **Segments 0 and 2:** Low economic_power and ev_adoption_propensity.
- **Segment 1:** Medium economic_power and ev_adoption_propensity.
- **Segment 3:** High economic_power and ev_adoption_propensity, representing the most valuable segment.

Feature Importance



The chart clearly shows that **literacy rate** is the primary driver for predicting EV adoption. Economic_power is a distant second, while all other features have a minimal impact. This suggests that for this model, the level of education and economic strength are the most crucial factors in understanding EV adoption trends.

Solution to company

Targeting and Investment Strategy

- **Top Segment:** The analysis recommends targeting **Segment 3**, identified as having the highest growth_potential_score of 93. This segment is strong in literacy_rate, economic_power, and market_potential.
- **Ideal Audience:** The optimal demographic falls within the **100-200k monthly income bracket** in areas with a high percentage of people aged 30-49.
- **Key Growth Drivers:** The two most important factors for predicting EV adoption are literacy_rate (with an importance score over 0.7) and economic_power (with an importance of around 0.18).
- **Recommended Vehicle:** The analysis suggests focusing on **2-wheelers (2W)** due to their high growth momentum.