How Demographics and Urbanization Affect EV Adoption in New York

Data Report – Mario Zacherl (23464824)

1. Main Question

The adoption of electric vehicles (EVs) has been increasing globally, yet the rate of progress across regions has varied considerably. This project aims to identify the factors influencing the adoption of EVs by correlating the demographic characteristics of counties in New York with their respective EV penetration rates.

New York was selected for its extensive demographic diversity, including wealthy, highly urbanized areas such as New York City and sparsely populated rural regions within the Adirondack Park.

2. Data Sources

The data utilized in this analysis is derived from two principal sources. The State of New York supplies the number of registered vehicles categorized by engine type, while the United States Census Bureau provides demographic data through the American Community Survey..

2.1 New York State Open Data Portal

The Department of Motor Vehicles for New York State provides its anonymized registration data for all types of vehicles as open data on the New York State Data Portal. Based on this dataset, multiple views exist, one of which shows the **vehicle and boat registrations by fuel type per county**:

The data may be downloaded in CSV format via the API, which is accessible at the following link:

https://data.ny.gov/resource/vw9z-y4t7.json

The data is structured as follows:

Column	Meaning
record_type	Type of Vehicle (**BOAT**, **TR**(ai)**L**(er), **SNOW**, (other) **VEH**(icles))
fuel_type	Fuel Type used (**ELECTRIC**, **GAS**, **DIESEL**, **PROPANE**, ...)
county	County of Residence or use (**ALBANY**, **ALLEGANY**, **BRONX**, ...)
registration_class	Amount of vehicles in this class

License:

The data is part of New York's Open Data Portal, which forms a component of the Open NY program. Any data published as part of this program can be used freely, without any requirement for attribution. (see Terms of Service)

Data Quality:

As all vehicles are required to be registered with the Department of Motor Vehicles, it may be assumed that the data is complete and accurate. No missing values were identified in the

dataset. The primary limitation of this dataset is that it is updated multiple times a year, yet no historic data is available. This may introduce uncertainty into the pipeline.

2.2 US Census Bureau

The United States Census Bureau administers the American Community Survey on an annual basis. This survey entails the collection of data from a representative sample of the American population, which is then compiled and published on the Census Bureau Data Portal.

A variety of tables ([1], [2], [3], [4]) were employed to filter out particular characteristics of the population residing in a single county. All data were derived from the 2022 5-Year Estimates, as the most recent 2023 1-Year Estimates did not include data for the smallest counties, as the minimum population threshold for publication is 65,000.

All tables can be downloaded via the API as JSON, but the headers are cryptic. These were manually matched based on the web view. Another metadata API exists, but it is not within the scope of this project..

The following represents an incomplete list of the columns that were selected for further processing:

Original column name	**New column name**
Name	county
H2_002N	Urban Households
K201501_007E	Bachelor's Degree
S0101_C01_032E	Median Age
B02001_005E	Asian (%)
S1901_C01_012E	Median Household Income

License:

The United States Census Bureau makes data available via its data portal as open data, subject to the terms of service, which require attribution. This has been incorporated into the pipeline and readme file.

Data Quality:

The US Census Bureau serves as the principal data source for government officials in support of their decision-making processes. In accordance with these standards, the Census Bureau only publishes data that represent the population observed with a low margin of error. This is also the reason why data for the small counties is not published every year. As a result, the data can be relied upon to be accurate. Furthermore, the cryptic column names do not present an issue, given that this is static data that will not be altered in the future. Matching them by hand is an adequate solution. However, there may be a need for adjustments to be made in order to run the project again on future data, should the format change.

3. Pipeline

The pipeline is implemented in Python and follows the ETL model. Initially, data is extracted from web sources and transformed using the Pandas library. Subsequently, the transformed data is joined together, and the result is saved to disk.

Each step also has built-in error handling that tells the user if and where something went wrong.

3.1. Extraction

The data is downloaded from the APIs via a GET request. The raw files are then saved to disk, thus enabling the data to be accessed without the necessity of downloading it again. As previously described, the data is either received as a CSV or as a JSON file. Both file types can be directly loaded using the Pandas library.

3.2 Transformation

Each table contains the county as one column, but there are differences in the exact formatting between the NY Data Portal and the Census Data Portal. To address this discrepancy, the county names are processed by a shared function, which removes spaces, points, and an optional appendix, and transforms the result to uppercase. The following example illustrates this transformation:

St. Lawrence County, New York → STLAWRENCE

From the remaining columns, only the relevant ones are selected. They are then converted to percentages for further analysis. In the majority of cases, this is achieved by dividing by the total. However, there are some exceptions to this rule:

- In order to calculate the mean education label, a value is assigned to each degree
- In the case of electric vehicle registrations, only those rows with the record type "VEH" are being counted, as including all types of vehicles, including boats and snowmobiles, would result in a distortion of the results

In a final step, all the tables are merged using the county name as a matching criterion for the entries.

3.3 Loading

The results are stored in a CSV file using the Pandas library for subsequent analysis.

4. Results and limitations

Data Output:

The pipeline generates a CSV file for each county, which contains the following metrics:

EV_Percent, Rural Percent, Average Education Level, Median Age, Females per 100 Males, White (%), Black or African American (%), Asian (%) ,Other (%), Mixed (%), Median Household Income

The data has already been normalized in order to facilitate subsequent processing.

Limitations:

The two data sources in question are subject to different updating procedures. The NY Data Portal consistently provides the most recent data, with no option to select data from a specific date. In contrast, the Census Data Portal returns data for a fixed date. At the time of writing, the NY Data Portal is displaying data from November 2024, while the Census Data Portal is displaying data from 2022. This discrepancy is likely to widen in the future, potentially necessitating manual adjustments. However, the impact of this gap should be minimal, given that the Census Data is changing at a relatively slow pace and is standardized before analysis, which mitigates the influence of discrepancies in absolute values (e.g., the median income).