In this analysis project I am looking at predictors of Batter Electric Vehicle (BEV or EV) registration.

EVs are a major emerging market in the US at time of writing. Mass BEV adoption is a climate and sustainability goal for numerous countries and a few US states. The US is purchasing EVs slower than much smaller European countries and its market is dwarfed by China's EV industry. Despite this, EV sales are steadily rising and infrastructure is growing to meet the demands.

Unfortunately, EV adoption and usage is dependent on many regional factors, since their range and charging requirements limits their use relative to gasoline vehicles. This is why I chose to look at the future of EV adoption in the US by analyzing current trends based on geographic factors. The core question of this analysis is "what regional factors correlate with EV registrations?"

This analysis works with multiple datasets from various sources. Atlas EV Hub (https://www.atlasevhub.com/materials/state-ev-registration-data/#data-format) has compiled several state reports of EV registrations, though none of the states standardized their reporting meaning I had to aggregate to the county level. I also used US census data for average commute times, average individual income, and population by county. The last source I used for my data was the Google Geocoding API (https://developers.google.com/maps/documentation/geocoding/overview), which was applied in a separate script with a 0.1 sleep timer to respect rate limits.

The gaps between the reporting of these datasets ultimately created more holes than I would have liked, as can be seen on the geographic maps I generated. Still, I managed to retain enough counties once I cleaned the data to properly graph trends and train and test a model once I used the Geocoding API to create a file that mapped ZIP codes and other identifiers to standardized county names.

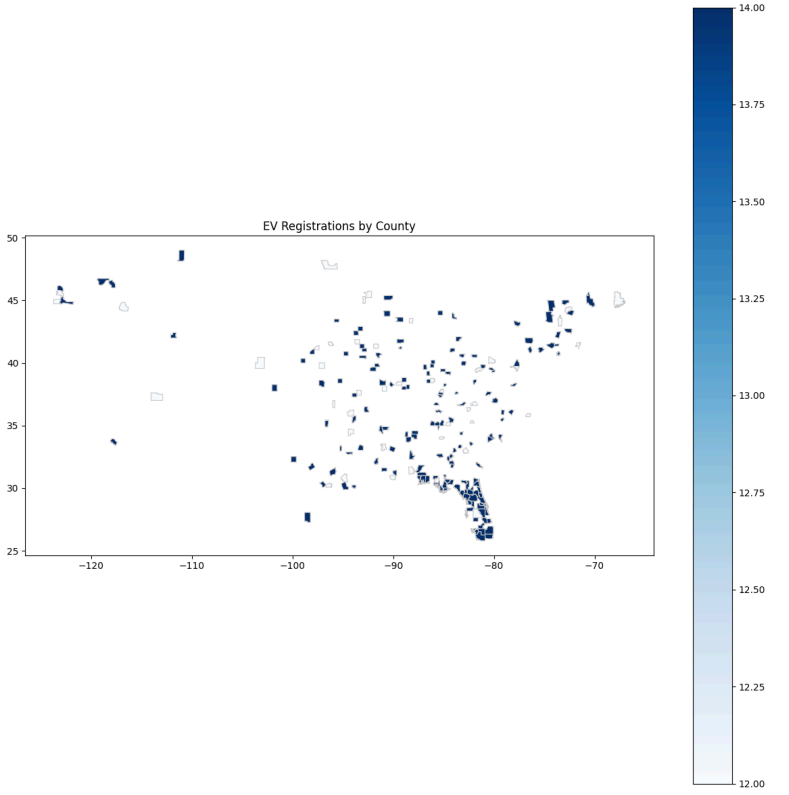
The data cleaning process began with retrieving the data for each state. California's data had to be acquired separately since Atlas EV had not updated their database since 2020. All state data was concatenated into a single dataframe. Census data for income (https://www.census.gov/library/visualizations/interactive/median-household-income.htm), commute (https://www.kaggle.com/datasets/benwhite/average-commute-time-by-city-us-census-bureau/code), and population (https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-total.html) was next to be loaded and standardized into a consistent format of the county, 2021 data, 2022 data, 2023 data, and an average of the years I calculated. Commute data required additional cleaning to reach this point since it was separated into frequencies of ranges, which is respectable but inconvenient for this analysis since I had to assume the mean or minimum for each range then calculate the average from there. It should also be noted that the CSV files used in this analysis had to be manually edited in a text editor to remove notes that broke the formatting, so I have included the modified files in the

DropBox page attached to this project (https://www.dropbox.com/scl/fo/3lfpv92dhodru4rlsignq/AFuVhFkW-GXTeWsQNFLr1NE?rlkey=rua0zjh4z6ke3cl1ge0o484v4&st=cu4chz5s&dl=0).

In a separate script the Google Geocoding API was used to create a csv file that maps ZIP codes to county names. After running overnight and loading the conversions into the project, the data of states that reported only ZIP code registrations was aggregated to the county level. With this standardized level the counts of each county in the registrations dataset was converted into its own dataframe of the total EVs registered within county limits.

At this point multiple optimization issues were encountered as the device hosting this project lacked sufficient RAM to allocate for the size of the datasets while running the scripts and cells. To fix this, all non-standardized columns were dropped from analysis and county names were standardized to the name of the county with no words following it in uppercase. Any rows that were missing a variable for the upcoming analysis were also dropped to save space and reduce the visualization loading times from 30 to ~9 minutes on average. Finally, all numeric columns were downcast to floats for additional efficiency.

Once all datasets were standardized, they were merged together, validated, and all duplicate or NaN columns were dropped.

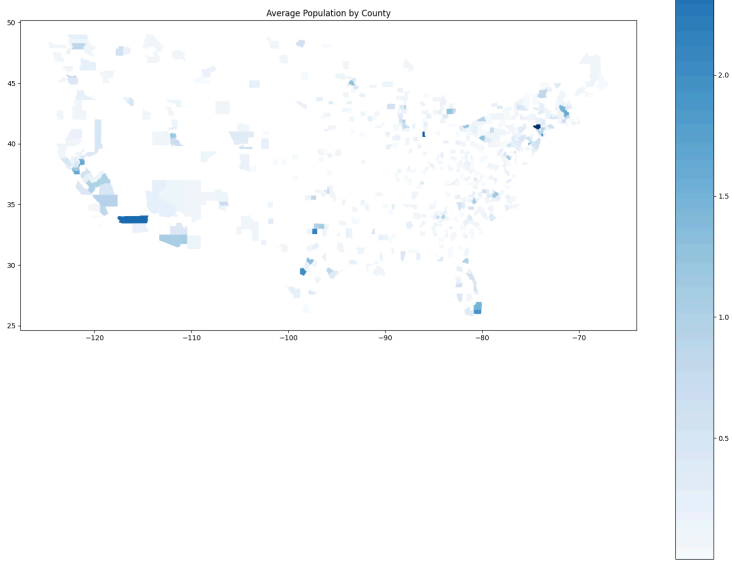


The visualizations relied on to identify correlations were geographic maps of the mainland US (Alaska and Hawaii were omitted due to their climates, geographies, and populations not being generalizable or significant as a part of the sample or the visualization analysis this project requires).

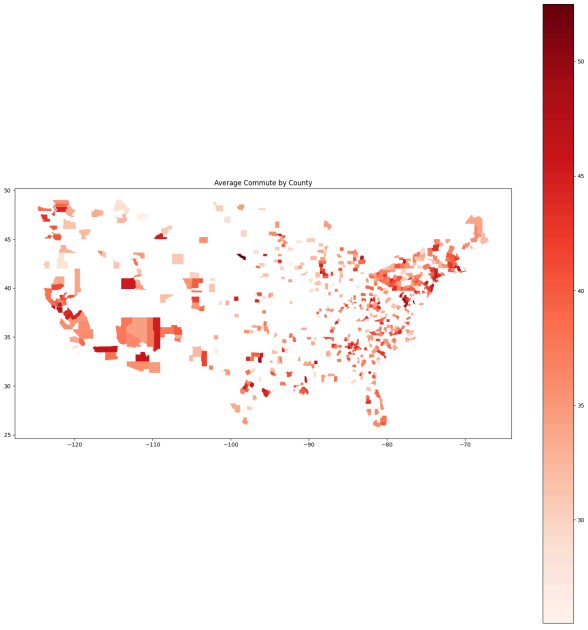
To the left (blue) is a graph of the EV registrations per county. California could not be displayed due to reporting inconsistencies with the rest of the states, but its values in the data indicate it is one of the most EV populated regions of

the US rivaled only by florida and new england.

To the right (blue) is a graph of the average population by county in this analysis. Below (green) is a map of average income in those same counties.



Below (red) is a map of the average commute by county of interest in the US. Noticeably this does not take into account the drain on the charge of an EV from various conditions including miles driven, elevation changes, temperature, traffic, etc.

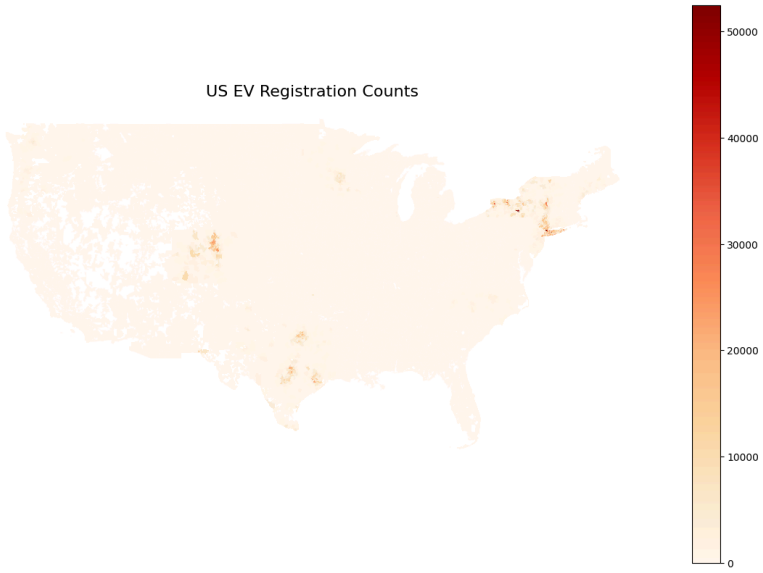


Each of these factors was included in the exploratory visualization analysis stage for their theoretical effect on EV registrations. Population is correlated because of course the more people in an area, the more people there are to own and use EVs. Commute was included because EV's are best used as commuter vehicles and drivers may gravitate towards EVs if they have a shorter or longer commute. Finally, income was included because historically EVs have had larger price tags than gasoline vehicles. Latitude was initially investigated as an additional variable included in the US Census shapefiles used to build the maps, but the density in New England, Florida, and the West Coast ultimately meant that no correlation was observed. This also indicates that temperature and snowfall is not a significant factor in EV adoption for the US, which bodes well for the EV market's expansion into the northern midwest in the coming years. EV charging station frequencies were also assessed but due to inconsistent and dated reporting as well as a "chicken and egg" problem it was ultimately omitted from the analysis. All that can be said for that statistic is that there is a heavy surface-level correlation between EV ownership and EV chargers in a region.

The most popular EV in the data was the Tesla Model 3, which has a range of 341 miles in ideal conditions. Other popular vehicles, like the Chevrolet Bolt (which I use), only have a range of ~200 miles on a good day with a battery upgrade (the new batteries also don't explode, which is nice). This variation in vehicles may have had an effect on the registrations/commuting correlation if the stats of different models were included in the analysis, but since very few states reported the models of registered EVs it could not be assessed.

There are several concerns that could be put forward relating to the scope of this analysis, namely the choice to look at the county level instead of state or ZIP code and limiting the analysis to only the US instead of including more EV-developed nations like European states and China.

Firstly, the county was chosen as the level of analysis for practical reasons relating to the choices of some states to protect their citizens privacy by not providing detailed registration info to Atlas EV. This made it impossible to get any more granular with the analysis, not that it would have necessarily been a good idea too anyways since most Americans who own vehicles don't travel exclusively within their ZIP code. That said I did begin this analysis by looking at the ZIP code level (map shown below) to identify the factors that correlated with EV registrations for a few states that reported ZIP code registrations, which led to identifying population and income as relevant factors.



Secondly, this project began with the idea to include Europe and China in the analysis. I can read Chinese and most European states publish data in English so the language barrier wasn't a problem. Despite the excellent studies and datasets (much better than any of the US ones used in this analysis) in these regions, further research indicated that the US's slow adoption of EVs is not an accident, but stemming from the much larger relationship between Americans, American infrastructure, and personal vehicles that makes comparing the US to those other countries difficult since the needs and resources available to US drivers differ drastically from the EU and China. Since the focus of this analysis is on US EV adoption, those other markets don't serve as accurate models for the US's immediate future in EVs.

In terms of the analysis process, there are still plenty of critiques to be leveled. Firstly is the very obvious gaps in the datasets, as shown in the spaces where counties were punished for failing to report any EV registrations by being dragged into the gaping white void. Although some of those spaces are just locations where no area codes exist and thus the registrations couldn't be mapped to a county, there is still a clear issue with the sample data.

This county-level analysis also avoids the impact States can have on their EV population, such as policies and incentives that may improve infrastructure, subsidise the industry, or otherwise encourage the purchase and thus registration of EVs within their borders. However, EV incentives take a long time to go into effect and their impacts are sometimes lost in the rapidly-evolving industry. Most of the data used in this project doesn't range over a long enough period to create a year-by-year timeline, making temporal analysis to identify the effects of this factor difficult.

There were also assumptions made in the data cleaning that may have compromised analysis. For instance, the mean and minimum were assumed for the

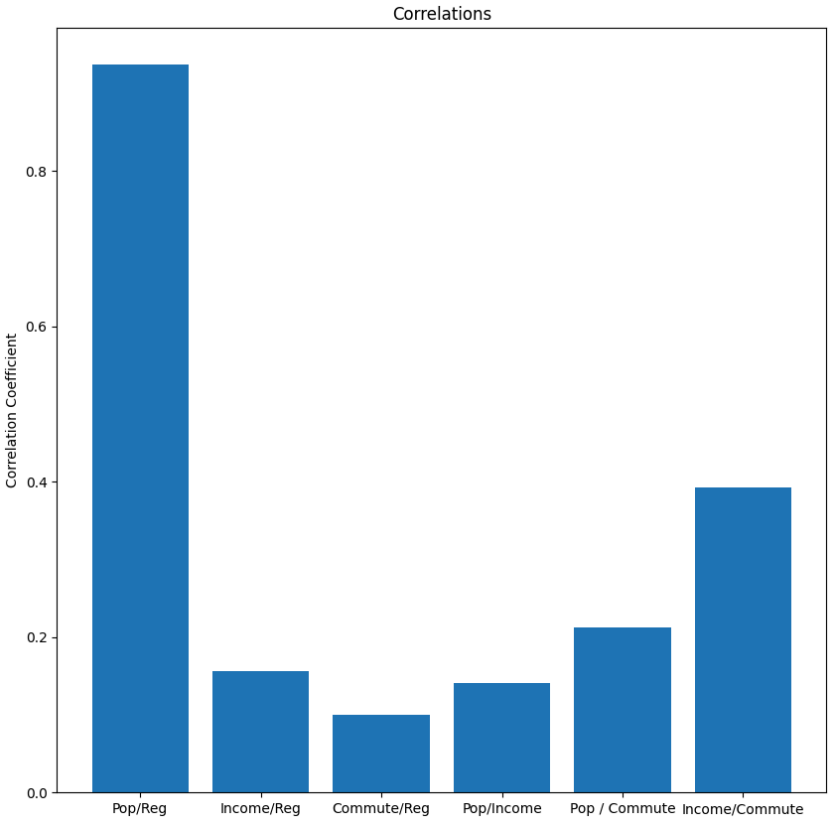
range of ~10 to 60+ of commute times from census data. This may have introduced an inaccuracy or oversimplified a more complex variable.

The selected variables for the analysis are also flawed. Unfortunately the data cleaning absorbed most of the time allotted to this investigation so the scope had to be limited to population, income, and commute. In the future I plan to modify this piece to the original scope and include weather data from the Open Weather Service, geographic data, road density data. Weather data would relate to issues of maintaining and using EV chargers as well as the difficulties many EVs have in snow and cold temperatures since it weakens their efficiency. By obtaining weather normals, expected annual weather averages, the analysis can account for these factors.

Road density data, when combined with commute data, could give a picture of how far and for how long the average vehicle-owning resident drives on a daily basis, which is something that would need to be achieved by any and every EV owned in that region. Counties with a low Density/Commute factor would likely be less likely to have EV's given that many EVs simply aren't ideal for traveling long stretches of road daily without long charge times. I look forward to testing this hypothesis when I update this analysis.

With the target parameters identified (population, income and commute) the statistical analysis began with examining the correlation between the factors, which yielded:

```
Population vs Registrations: 0.9372983234395142
Income vs Registrations: 0.15584718069004122
Commute vs Registrations: 0.09967393557945381
Population vs Income: 0.14066632146764022
Population vs Commute: 0.2119171678620782
Income vs Commute: 0.3925285317319831
```



The registration correlations were then assessed using a t-test to test the relationship hypothesis further. The results are shown below:

Population vs Registrations: t-stat: **2.741666**, p-value: **0.0079**
Income vs Registrations: t-stat: **14.8668**, p-value: **1.9591**
Commute vs Registrations: t-stat: **-2.8374**, p-value: **0.0060**

With correlations identified, these variables were placed into a linear regression model in different combinations. Due it high collinearity with registrations, average population was omitted from the first model, which produced the following results:

Mean Squared Error: 76727.4
R^2: 0.115
Intercept: 0.717
Average Commute Coefficient: 4.44628864
Average Income Coefficient: 0.00187789380

The MSE is high relative to the data, so the model doesn't fit the data well and it's predictions are likely far from the true values and mean registrations of the counties.

The R^2 is relatively low, also indicating a poor fit of the model to the data.

The coefficients tell most of the story of this model. It seems that income doesn't have much of an effect on the EV registrations of a county, meaning that EVs, as of the 2020s, are no longer considered a luxury item or out of reach for a majority of Americans. Meanwhile the commute coefficient suggests that people are more likely to own cars the longer they commute, at least within the defined range of the census dataset of ~10 to 60+ minutes for the reported counties, which may be more likely to fall within this range.

The next model included population as a parameter despite high collinearity with registrations since it is possible that people may be more incentivised to own EVs if they live in a more populated area due to the increased charging infrastructure and eliminated emissions.

Mean Squared Error: 6712.4
R^2: 0.922
Intercept: 736.29
Average Commute Coefficient: -1.93339563e+01
Average Income Coefficient: 1.68634934e-03
Average Population Coefficient: 2.74905381e-04

The EU has stated that by 2030 EVs should account for 80% of new car sales, and 100% in 2035. The EU already has a head start with 23% as of 2023, while the US is significantly behind at 9.4% that same year. As the US struggles to meet emission standards and remains determined never to break from its reliance on consumer passenger vehicles, the EV market will have to identify what different Americans want from their beloved cars and adjust their designs, charging infrastructure, and business strategies to meet those needs and provide a compelling alternative to gasoline engines. The results of this study show that the US EV industry has already overcome one hurdle, that of price, with cheaper models of EV like the Tesla Model 3 alongside even cheaper and more recent non-Tesla EVs leading in popularity even in lower-income counties.

The analysis suffers from several issues of procedure and scope, but still provides insight in the geographic and socioeconomic factors that affect the popularity of EV's in a given county. Further research and analysis is needed in order to produce a comprehensive guide to this evolving market, but once it is fully understood these datasets and variables will be vital for accurately predicting and facilitating widespread EV adoption.

Link to Colab:
https://colab.research.google.com/drive/18WbLhdvv595yR91qJuhO8T8YqWciL3qw?usp=sharing

Link to Dropbox: https://www.dropbox.com/home/EV_Analysis_Files#

California Light EV Dataset Link: https://www.energy.ca.gov/data-reports/energy-almanac/zero-emission-vehicle-and-infrastructure-statistics-collection/light