We are performing an exploratory data analysis (EDA) of electric vehicles (EVs), specifically of what factors (e.g. income) are associated with electric vehicle purchase, how existing infrastructure in an area (e.g. charging stations) provides insight into EV ownership in that area. Other variables we have followed include EV Make and Model, EV Type (BEV vs. PHEV), EV Price and Year, and overall geographic location that may correlate to an increase in EVs.

*Other variables we might want to mention summarily? model year trends(mostly new bought), popular makes, electric ranges, pricing (in Ben’s visualizations)*

Our project is comprised of these major parts:

* Exploratory data analysis of an electric vehicle population in a certain geographic area. We have chosen, in part due to the region’s particular affinity to electric technological innovation and alternative energy options, to refine our analysis to Washington state.
* We have utilized datasets from two sources: a 2023 EV dataset published on Kaggle.com that shows the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through Washington State Department of Licensing (DOL); and U.S. Census Bureau data (agsc5) for 2022, the latest available. The former gives us a dataset of 8340 unique electric vehicles with which to work.
  + Details of each dataset to highlight in brief explanation:
* In merging these two datasets, we created a thorough dataframe that provides us vehicular and population data necessary to test our hypotheses.

The Role of Income

Firstly, we predicted that income would have a positive correlation with the number of EVs in a given area. To delve into this hypothesis further, we methodologically organized all data by zip codes in Washington and started with the median household income and population of each zip code. In order to circumvent external biases, we arrived at the per capita variables for income and EV ownership, converted into percentages, which we could then plot and analyze. In other words, we tried to level the proverbial playing field across zip codes as much as possible, as we acknowledge that zip codes in metropolitan areas, like Seattle, near the coast differ in demographics than those in zip codes in more sparsely populated rural areas inland.

With the majority of our charts breaking data down by zip codes or by county, there is a precipitous drop in EVs outside of these former regions (e.g. King County around Seattle comprises 52% of total EV registrations in the entire Washington state). We must admit, though, that fully understanding this dynamic requires external research to fill in the holes our datasets cannot. For example, Redmond, WA (in King County), zip code 98052, by sheer volume contains the most EVs by an exponential margin. The 2916 unique EVs registered there is numerically the most of all zip codes and translates to 3.77% of Population Percentage/EVs Per Capita, which is in the top three in the state. When one researches the brick-and-mortar make-up of Redmond, one easily realizes that it is the corporate headquarters of Microsoft—the workers of which would tend to be more technologically “savvy” and likely to be more amenable to EV ownership.

### To better assess the prevalence of EVs, we identified the ten zip codes with the highest EVs per capita. (See [Top 10 Zip Codes EV Per Capita.png](https://github.com/caelwillis/group-4-proj-1/blob/main/Graphs/Top%2010%20Zip%20Codes%20EV%20Per%20Capita.png)) In short, zip code 98134 leads the state with 18.45% of the population owning EVs, nearly twice that of the second zip code. The general trend showed a rather bifurcated EV per capita count overall, with there being a sizeable gap (of ~6-7%) between the mean top EV per capita zip codes and that of the bottom, all of which reported at less than 0.1% of the population. Thus, we see just how spread out our EV ownership data is across the state, with the urban areas near the coast far outweighing the rest of the state inland.

The Make of EVs

### Since we set out to analyze the relationship of income and EV ownerships, we organized all EVs in our original dataset by Make. This allowed us to get a better sense of how luxury brands (i.e. Lexus, Mercedes, Land Rover, etc.) are faring in the EV market. We found a negative correlation between Per Capita Median Income and EV Ownership per Population (r-value = ?). After filtering the outliers, we saw that, while higher income certainly allows for more spending power, the prevalence of expensive luxury EVs seems to be in its seed stage. In other words, because the higher Per Capita Income in an area does not denote that there will be more EVs in that area, there remains speculation as to why the luxury, more costly EV market is so outnumbered compared to cheaper counterparts. This negative slope in “[EV-make-pc-Income-line.png](https://github.com/caelwillis/group-4-proj-1/blob/main/Graphs/EV-make-pc-Income-line.png" \o "EV-make-pc-Income-line.png)”could also be indicative of general disinterest to buy and/or switch to EV vehicles amongst wealthier populations with traditional gas cars. At the very least, we can conclude that EVs Per Capita are not indicative of the Per Capita Income; rather, we see similar purchasing rates across socioeconomic areas with EVs of different Makes and prices being accessible to all.

### What we also did find, as to be expected, was that, when using the Make of EV as the independent variable, its price/brand is strongly positively correlated to the Median Income Per Capita of the buyer. (see i[ncome\_make\_bar.png](https://github.com/caelwillis/group-4-proj-1/blob/main/Graphs/income_make_bar.png)) We used a bar graph to visualize the correlation between Median Per Capita Income and the specific EV Makes. As one might imagine, we found that Porsche, Land Rover, Volvo, and Audi purchases were the most numerous among those with the highest incomes (all $60,000+ Median Income Per Capita). On the other end of the spectrum, Mitsubishi, Cadillac, and Ford were the EV Makes most bought by those with lowest Median Per Capita Income. We concede that we have not introduced other variables, such as MSRP and Model and Age, here into a multivariable correlation but we feel that our charts illustrate our overall analytical hypotheses.

The Type of EVs to Support Our Hypothesis

### Two types of EVs we have differentiated are Battery Electric Vehicle (BEV) and Plug-in Hybrid Electric Vehicle (PHEV). In general, independent of the make, PHEVs are cheaper relatively to their BEV counterparts. In our charts measuring this relationship, we see that more PHEVs are purchased by a larger range of Median Incomes Per Capita, with the \*average income being $50,000 per year versus nearly $60,000 for BEVs. (See [PHEV-box.png](https://github.com/caelwillis/group-4-proj-1/blob/main/Graphs/PHEV-box.png) and [BEV-box.png](https://github.com/caelwillis/group-4-proj-1/blob/main/Graphs/BEV-box.png)) Also notable in the BEV box chart, we notice an outlier for BEVs at the $80,000 income level. Causes for this could be anything from spontaneous purchases or (…..?).

We can also postulate that BEVs usually have longer electric ranges than PHEVs (see dataframe “clean\_merged\_df” in EV-ownership-visuals.ipynb). This disparity also has implications on the need for more densely located EV Charging Stations in areas where PHEVs are considerably more popular than BEVs. Our analysis did not delve into these relationships in detail; however, perhaps with a country-wide EDA with more diverse datasets and larger distances (e.g. EVs traveling on interstates), we would be able to show these relationships with numerical data not bound by confounding variables.

The Role of EV Charging Stations

Just as gas stations are essential to the infrastructural capacity of traditional vehicles, charging stations play an important role in providing access to EV ownership. As in the prior section, we hypothesized a positive correlation between EV Charging Stations and the Number of EVs in that area, as well as between EV Charging Stations and area Income. Again, we parsed our data points so that our metrics are all per capita and by zip codes. For example, we calculated the EV Charging Stations per capita by presenting the number of them in each zipcode as a percentage of that zip code’s population. Thereby, our results would not be skewed by the reality that larger and more populated areas able to purchase EVs would most likely necessitate more stations.

### We noticed a couple of important relationships. Our bar graph [bar\_top\_charging\_stations.png](https://github.com/caelwillis/group-4-proj-1/blob/dev-michael/Graphs/bar_top_charging_stations.png) shows the 10 zip codes with the largest Charging Station Percentage of Population. Most notably, any strong correlation between income and charging stations is absent; thus, the initial presumption that “richer” areas would necessarily have more “modern” infrastructure features, in this case charging stations, falls short. Zip code 98757 has the highest Per Capita Charging Stations and a quick look into this area shows that it is on the edge of Olympic National Park. This leads one to surmise that EV Charging Stations are perhaps better studied in terms of strategic location in conjunction with public funding abilities.

### To support this, we graphed the a bar chart showing highest Per Capita Income Zip Codes vs their Per Capita Charging Stations (%) (see [bar\_wealthy\_charging\_stations.png](https://github.com/caelwillis/group-4-proj-1/blob/dev-michael/Graphs/bar_wealthy_charging_stations.png)). There was no particular correlation noticeable. In fact, when we regressed the two variables, we arrived at a near zero r-value of 0.12, denoting a very weak positive correlation. See SCATTER

What other considerations do we want to discuss??? Tell me what I’ve missed and I’ll add it and make it sound pretty. Then will write a concise Conclusion.

Conclusion:

**Analysis and Conclusion (20 points)**

* Write-up summarizes major findings and implications at a professional level (5 points)
* Each question in the project proposal is answered with precise descriptions and findings (5 points)
* Findings are strongly supported with numbers and visualizations (5 points)
* Each question response is supported with a well-discerned statistical analysis from lessons (e.g., aggregation, correlation, comparison, summary statistics, sentiment analysis, and time series analysis) (5 points)