Example Manuscript Template for a Data Analysis Project

Sri Lakshmi Chundru, Collin Real, Joaquin Ramirez, Seth Harris

2024-07-25

Data Sources for the work:

* https://www.google.com/publicdata/directory
* https://www.data.gov/
* https://github.com/awesomedata/awesome-public-datasets
* https://registry.opendata.aws/
* https://www.kaggle.com/datasets/jainaru/electric-vehicle-population

# 1. Summary/Abstract

*Write a summary of your project.*

Electric vehicle market penetration is an important topic of interest for a variety of government agencies and automotive companies in terms of planning future charging infrastructure and inventory management. This project aims to aid this process by creation of a data driven model that can predict the number of electric vehicles in a given zip code based on a variety of factors. The data used in this project was sourced from the Washington State Department of Licensing (DOL) and includes information on Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) registered in Washington State. The data set contains 181,458 observations and 17 variables, including Vehicle Identification Number (VIN), county, city, model year, electric range, and more. The project will involve data cleaning, exploratory data analysis, and statistical modeling to identify key factors influencing electric vehicle adoption and predict the number of electric vehicles in a given zip code. The results of this project will provide valuable insights for policymakers, urban planners, and automotive companies to support the transition to sustainable transportation.

# 2. Introduction

The dataset obtained from Kaggle provides detailed information about Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) registered through the Washington State Department of Licensing (DOL). It comprises 181,458 observations and 17 variables, encompassing key details such as Vehicle Identification Number (VIN), county, city, model year, electric range, and more.This document is an exploratory data analysis of electric vehicle data.

A second set of data was obtained from the IRS and contains information on electric vehicles, their owners, and the tax benefits they receive. The data was cleaned and processed in R and saved as a CSV file. This data was then combined with thwe data from kaggle to conduct the analysis of data and model creation that predicts number of electric vehicles in a given zip code based on various input varaibles.

# 3. Data set Analysis

data\_location <- here::here("data","raw-data","Electric\_Vehicle\_Population\_Data.csv")  
data <- read.csv(data\_location)  
output <- capture.output(str(data))  
cat(output[1:3], sep="\n")

'data.frame': 181458 obs. of 17 variables:  
 $ VIN..1.10. : chr "WAUTPBFF4H" "WAUUPBFF2J" "5YJSA1E22H" "1C4JJXP62M" ...  
 $ County : chr "King" "Thurston" "Thurston" "Thurston" ...

head(data)

VIN..1.10. County City State Postal.Code Model.Year Make Model  
1 WAUTPBFF4H King Seattle WA 98126 2017 AUDI A3  
2 WAUUPBFF2J Thurston Olympia WA 98502 2018 AUDI A3  
3 5YJSA1E22H Thurston Lacey WA 98516 2017 TESLA MODEL S  
4 1C4JJXP62M Thurston Tenino WA 98589 2021 JEEP WRANGLER  
5 5YJ3E1EC9L Yakima Yakima WA 98902 2020 TESLA MODEL 3  
6 1C4JJXP66P Thurston Olympia WA 98501 2023 JEEP WRANGLER  
 Electric.Vehicle.Type  
1 Plug-in Hybrid Electric Vehicle (PHEV)  
2 Plug-in Hybrid Electric Vehicle (PHEV)  
3 Battery Electric Vehicle (BEV)  
4 Plug-in Hybrid Electric Vehicle (PHEV)  
5 Battery Electric Vehicle (BEV)  
6 Plug-in Hybrid Electric Vehicle (PHEV)  
 Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility Electric.Range Base.MSRP  
1 Not eligible due to low battery range 16 0  
2 Not eligible due to low battery range 16 0  
3 Clean Alternative Fuel Vehicle Eligible 210 0  
4 Not eligible due to low battery range 25 0  
5 Clean Alternative Fuel Vehicle Eligible 308 0  
6 Not eligible due to low battery range 21 0  
 Legislative.District DOL.Vehicle.ID Vehicle.Location  
1 34 235085336 POINT (-122.374105 47.54468)  
2 22 237896795 POINT (-122.943445 47.059252)  
3 22 154498865 POINT (-122.78083 47.083975)  
4 20 154525493 POINT (-122.85403 46.856085)  
5 14 225996361 POINT (-120.524012 46.5973939)  
6 22 220675367 POINT (-122.89692 47.043535)  
 Electric.Utility X2020.Census.Tract  
1 CITY OF SEATTLE - (WA)|CITY OF TACOMA - (WA) 53033011500  
2 PUGET SOUND ENERGY INC 53067011100  
3 PUGET SOUND ENERGY INC 53067012226  
4 PUGET SOUND ENERGY INC 53067012620  
5 PACIFICORP 53077000800  
6 PUGET SOUND ENERGY INC 53067010802

summary(data)

VIN..1.10. County City State   
 Length:181458 Length:181458 Length:181458 Length:181458   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
   
 Postal.Code Model.Year Make Model   
 Min. : 1545 Min. :1997 Length:181458 Length:181458   
 1st Qu.:98052 1st Qu.:2019 Class :character Class :character   
 Median :98122 Median :2022 Mode :character Mode :character   
 Mean :98174 Mean :2021   
 3rd Qu.:98370 3rd Qu.:2023   
 Max. :99577 Max. :2024   
 NA's :3   
 Electric.Vehicle.Type Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility  
 Length:181458 Length:181458   
 Class :character Class :character   
 Mode :character Mode :character   
   
   
   
   
 Electric.Range Base.MSRP Legislative.District DOL.Vehicle.ID   
 Min. : 0.00 Min. : 0 Min. : 1.00 Min. : 4385   
 1st Qu.: 0.00 1st Qu.: 0 1st Qu.:18.00 1st Qu.:183068667   
 Median : 0.00 Median : 0 Median :33.00 Median :228915522   
 Mean : 57.83 Mean : 1040 Mean :29.11 Mean :221412778   
 3rd Qu.: 75.00 3rd Qu.: 0 3rd Qu.:42.00 3rd Qu.:256131982   
 Max. :337.00 Max. :845000 Max. :49.00 Max. :479254772   
 NA's :398   
 Vehicle.Location Electric.Utility X2020.Census.Tract   
 Length:181458 Length:181458 Min. :1.001e+09   
 Class :character Class :character 1st Qu.:5.303e+10   
 Mode :character Mode :character Median :5.303e+10   
 Mean :5.298e+10   
 3rd Qu.:5.305e+10   
 Max. :5.603e+10   
 NA's :3

colSums(is.na(data))

VIN..1.10.   
 0   
 County   
 0   
 City   
 0   
 State   
 0   
 Postal.Code   
 3   
 Model.Year   
 0   
 Make   
 0   
 Model   
 0   
 Electric.Vehicle.Type   
 0   
Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility   
 0   
 Electric.Range   
 0   
 Base.MSRP   
 0   
 Legislative.District   
 398   
 DOL.Vehicle.ID   
 0   
 Vehicle.Location   
 0   
 Electric.Utility   
 0   
 X2020.Census.Tract   
 3

**- (c) Explain the question you want to answer using the data. What will be your outcome(s) of interest (if any)? What (if any) specific predictors will you focus on? What relations/patterns are you looking for in the data?**

Question: What insights can be gleaned from the dataset regarding the electric vehicle landscape ?

## 3.1 Objectives:

* Understand Geographic Distribution: Determine the distribution of electric vehicles across different counties and cities in Washington State.
* Identify Popular Models: Identify the most commonly registered electric vehicle makes and models.
* Analyze Electric Range: Investigate the distribution of electric range among registered electric vehicles.-
* Explore Clean Alternative Fuel Vehicle Eligibility: Examine the eligibility for clean alternative fuel vehicle incentives among different vehicle types.

## 3.2 Outcomes of Interest

* Geographic Distribution: The distribution of electric vehicles across counties and cities in Washington State.
* Popular Models: The most frequently registered electric vehicle makes and models.
* Electric Range Analysis: The range distribution and average electric range of registered electric vehicles.
* Clean Alternative Fuel Vehicle Eligibility: The proportion of vehicles eligible for clean alternative fuel vehicle incentives.

## 3.3 Predictors

* Geographic Variables: County, City, Postal Code.
* Vehicle Attributes: Model Year, Make, Model, Electric Vehicle Type, Electric Range.

## 3.4 Relations/Patterns

* Geographic Distribution: Analyzing clusters of electric vehicles in specific regions.
* Popular Models: Identifying dominant electric vehicle makes and models.
* Electric Range Analysis: Understanding the distribution of electric ranges and potential implications for consumer preferences.
* Clean Alternative Fuel Vehicle Eligibility: Examining eligibility patterns and potential factors influencing eligibility.
* Correlations: Investigating any associations between electric vehicle characteristics and legislative districts or census tracts, which may offer insights into policy implications and consumer behavior.

**(d) As much as you know, suggest how you will analyze it. At this stage in the course, we haven’t covered analysis approaches yet, so you can keep things vague and non-technical here.**

* Descriptive Analysis: Summarize the dataset to understand its contents, including summary statistics and frequency tables.

-Visualization: Use plots and charts to visualize aspects like geographic distribution, electric range, and vehicle types.

* Comparative Analysis: Compare different groups within the dataset, such as BEVs vs PHEVs or eligibility for incentives by vehicle type.
* Correlation Analysis: Explore potential relationships between variables, such as electric range and base MSRP.
* Geospatial Analysis (optional): Use R packages to analyze spatial patterns in the data.
* Exploratory Data Analysis (EDA): Engage in exploratory analysis to identify interesting patterns or trends.

**(e) You are allowed, but not yet required, to provide background information for the question you plan to answer. For instance you can describe why it’s an interesting question, who else has done similar analyses, how your analysis will be new/different, etc. Similar to what you read in an introduction to a research paper. For the final report, you’ll need these parts. For part 1, they are not required, but you are welcome to already write down some of that.**

## 3.5 Primary Research Question: What insights can be gleaned from the dataset regarding the electric vehicle landscape in Washington State?

Sub-questions:

* How are electric vehicles distributed across different counties, cities, and States?
* Which EV makes and models are most popular?
* What is the range of electric vehicles in terms of electric range and base MSRP?
* How does eligibility for clean alternative fuel vehicle incentives vary among different vehicle types?
* Are there any significant correlations between EV characteristics and geographic or legislative factors?

**(d) Eventually, for your final report, what you write for this part will go into different sections of the full report. Some will go into the introduction, some in the methods section. You can already place these items there, or for now just write them as a single section.**

**Introduction** Electric vehicles (EVs), including Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs), are crucial in the global effort to reduce greenhouse gas emissions and dependence on fossil fuels. This dataset, sourced from the Washington State Department of Licensing (DOL) and obtained from Kaggle, provides a comprehensive overview of registered electric vehicles in Washington State. Analyzing this data offers valuable insights into the adoption patterns and characteristics of EVs, informing policies and infrastructure development to further promote sustainable transportation.

**Objectives:**

**Geographic Distribution:** Determine the distribution of electric vehicles across different counties and cities. **Popular Models:** Identify the most commonly registered EV makes and models. **Electric Range Analysis:** Investigate the distribution of electric range among registered EVs. **CAFV Eligibility:** Examine the eligibility for clean alternative fuel vehicle incentives among different vehicle types. **Correlation Analysis:** Explore potential correlations between EV characteristics and geographic or legislative factors.

**Methods**

**Data Cleaning and Preparation:**

* Address missing values and ensure correct data types.
* Transform variables as necessary for analysis.

To ensure the dataset is ready for analysis, the following steps were undertaken:

**1. Data Import and Initial Inspection:**

* The dataset Electric\_Vehicle\_Population\_Data.csv was imported.
* Summary statistics and data types were displayed to understand the initial structure and content of the data.
* UPDATE (6.17.2024) - Downloaded IRS tax data to join with the EV data by zip code
  + Tax data provides many financial variables that we will include in our models (i.e. can we predict the amount of EVs per zip code based off tax data?)
* Added a geojson file to draw zip codes on a map, will use for visualizations but still figuring out how to elegantly filter the json

**2. Checking for Duplicates:**

* Identified and counted the number of duplicate rows in the dataset to avoid redundant information.
* Result: num\_duplicates duplicates were found.

**3. Counting Rows with Zero MSRP:**

* Counted the number of rows where the Base.MSRP (Manufacturer’s Suggested Retail Price) is zero to identify incomplete pricing data.
* Result: count\_zero\_msrp rows with zero MSRP were identified.

**4. Checking for Null Values:**

* Calculated the total number of null (NA) values in the dataset to address missing data.
* Result: total\_na null values were found.

**5. Removing Null Values and Identifying Outliers:**

* Filtered out rows with null values to ensure data completeness.
* Identified the maximum Base.MSRP for each electric vehicle type to spot potential outliers.
* Result:The maximum MSRP for each vehicle type was summarized.

**6. Filtering Out Outliers and Zero MSRP:**

Removed the row with an extreme outlier Base.MSRP of 845,000 to prevent distortion in the analysis. Also filtered out rows where Base.MSRP is zero to ensure only valid pricing data is considered. By performing these steps, the dataset was cleaned and prepared, ensuring it is more accurate and reliable for subsequent analysis.

**Descriptive Statistics:**

* Compute summary statistics (mean, median, mode, standard deviation) for numerical variables.
* Generate frequency tables for categorical variables.

**Boxplot - Price Distribution vs. Electric Vehicle Type:**

This boxplot compares the price distribution of Battery Electric Vehicles (BEV) and Plug-in Hybrid Electric Vehicles (PHEV). It shows that BEVs have a wider price range with a higher median price compared to PHEVs, which have a more compact price distribution with some outliers.

**Boxplot - Price Distribution vs. Electric Vehicle Type (Clean Alternative Fuel Vehicle Eligible):**

This boxplot illustrates the price distribution based on eligibility for Clean Alternative Fuel Vehicle incentives. Vehicles eligible for the incentive tend to have a higher price range compared to those that are not eligible due to low battery range.

**Boxplot - Price Distribution vs. Electric Vehicle Type (Drive Range):**

This boxplot shows the distribution of drive range for clean alternative fuel vehicle eligible and non-eligible vehicles. Eligible vehicles have a wider and higher range compared to non-eligible vehicles, which have a limited drive range.

**Scatterplot - Electric Vehicle Range Growth over Time:**

This scatterplot depicts the growth in electric vehicle range over time. There is a clear trend showing that the electric range of vehicles has generally increased over the years.

**Boxplot - Price Distribution by Brand:**

This boxplot compares the price distribution of electric vehicles by different brands. Tesla has the highest price range, followed by Porsche. Brands like KIA and Chrysler have lower price distributions.

**Scatterplot - Price vs. Range Correlation:**

This scatterplot shows the correlation between the vehicle price and its electric range. There is a positive correlation where higher-priced vehicles tend to have a longer electric range. These plots provide insights into the price and range variations of electric vehicles based on type, eligibility for incentives, brand, and over time.

**Data Visualization:**

* Create bar charts, histograms, and pie charts to visualize distributions.
* Use scatter plots and box plots to explore relationships between variables.
* Generate geographic maps to visualize EV distribution across the state.

**Comparative Analysis:**

* Compare BEVs and PHEVs in terms of electric range and base MSRP.
* Analyze differences in CAFV eligibility across makes and models.

**Correlation Analysis:**

* Calculate correlation coefficients for numerical variables.
* Use cross-tabulations to explore relationships between categorical and numerical variables.

**Geospatial Analysis (if applicable):**

* Use geographic mapping tools to identify clusters or patterns in vehicle registrations.

**Exploratory Data Analysis (EDA):**

-Perform in-depth analysis to identify trends, patterns, or outliers. -Use advanced visualization techniques for deeper insights.

**Statistical Model** - First iteration of models - Produced a few statistically significant tax metrics for predicting the number of EVs in a given county - Washington demographics datasets could further enhance our analysis

**Conclusion:**

By analyzing the electric vehicle registration data for Washington State, this project aims to uncover significant patterns and trends that can inform policy decisions and infrastructure planning. The insights gained from this analysis will contribute to the broader understanding of EV adoption and support the transition towards a more sustainable transportation future.