Northeastern University

ALY 6015: Final Report

Prepared By:

Group 5

Khushi Manishkumar Doshi

Nahimah Yakubu Suglo

Devang Shirodkar

Shicheng Wan

Presented To:

Prof. Vivian Clements

Date:

May 13th, 2024

Content Page

Introduction

The global automotive industry is undergoing a significant transformation as the world shifts towards sustainable transportation solutions to mitigate environmental concerns and reduce dependence on fossil fuels. At the forefront of this transition are electric vehicles (EVs), which offer a promising alternative to traditional internal combustion engine vehicles. As governments, businesses, and consumers increasingly prioritise sustainability, understanding the factors influencing the adoption of electric vehicles has become paramount.

Considering these developments, our project focuses on conducting a comprehensive analysis of the Electric Vehicle Population Data, a rich dataset providing detailed information about electric vehicles registered across various regions for the USA. This dataset offers insights into key characteristics of electric vehicles, including model year, make, electric vehicle type, electric range, base MSRP, and geographic location, among others. By harnessing the power of data analytics and statistical techniques, our study seeks to uncover the underlying trends, patterns, and drivers shaping the electric vehicle market.

For further reading on the transformation of the automotive industry and the rise of electric vehicles, see the works of Sperling and Gordon (2009) and the reports by the International Energy Agency (2020).

Dataset Overview

The Electric Vehicle Population Data encompasses a wide range of information related to electric vehicles, including vehicle identification numbers (VINs), county, city, state, model year, make, model, electric vehicle type, clean alternative fuel vehicle (CAFV) eligibility, electric range, base MSRP, legislative district, DOL vehicle ID, vehicle location, electric utility, and 2020 census tract. This comprehensive dataset provides a holistic view of the electric vehicle landscape, enabling us to delve deep into the factors influencing EV adoption.

Details of columns:

- vin_1_10: Vehicle Identification Number (VIN) from 1 to 10 characters.
- county: County name (The administrative division within a state or country.).
- city: City name.
- state: State name.
- postal_code: Postal code.
- model_year: Year of the model.
- make: Vehicle manufacturer.
- model: Vehicle model.
- electric_vehicle_type: Type of electric vehicle.
- clean_alternative_fuel_vehicle_cafv_eligibility: Eligibility for Clean Alternative Fuel Vehicle (CAVF) program.
- electric_range: Electric range of the vehicle.
- base_msrp: Manufacturer's Suggested Retail Price (MSRP) of the base model.
- legislative_district: Legislative district.
- dol_vehicle_id: Department of Licensing (DOL) vehicle ID.

- vehicle_location: Location of the vehicle in the format 'POINT (longitude latitude)'.
- electric_utility: Electric utility.
- x2020_census_tract: Census tract for 2020.

Type	Variables
Categorical	VIN (1-10), County, City, State, Make, Model, Electric Vehicle Type, Clean Alternative Fuel Vehicle (CAFV) Eligibility, Vehicle Location, Electric Utility
Numerical	Postal Code, Model Year, Electric Range, Base MSRP, Legislative District, DOL Vehicle ID, 2020 Census Tract

This table categorizes each variable by type, aiding in the selection of appropriate analysis methods for each.

Project Rationale

As the automotive industry undergoes a paradigm shift towards sustainable practices, understanding the dynamics driving electric vehicle adoption is crucial. Our project seeks to bridge the gap between data and actionable insights by employing advanced statistical methods, particularly regression analysis, to explore the relationships between various factors and electric vehicle adoption rates. By leveraging the Electric Vehicle Population Data, we aim to address pressing questions such as the impact of vehicle characteristics on EV registration likelihood, the role of geographic location in EV adoption, and the identification of key predictors of EV adoption.

Objectives and Focus of the Study

The primary objectives of our project are as follows:

1. Identify Key Factors Influencing EV Adoption: We aim to identify the primary drivers influencing the adoption of electric vehicles, including vehicle characteristics, regional demographics, and policy incentives.

2. Understand the Impact of Vehicle Characteristics: By analyzing variables such as electric range, model year, and vehicle type, we seek to understand how these characteristics influence the likelihood of EV registration.

3. Analyze the Relationship Between Geographic Location and EV Adoption: Our study will explore the relationship between geographic location and the types of electric vehicles adopted, shedding light on regional variations in EV adoption rates.

Our study aims to leverage regression analysis techniques to explore these relationships within the dataset of electric vehicle registrations. By conducting regression analysis, we seek to provide insights that could inform stakeholders in the automotive industry and policymakers about effective strategies to enhance EV adoption, thereby contributing to the ongoing efforts to promote sustainable transportation.

Identification of Methods Used

For our preliminary analysis of the Electric Vehicle Population Data, we employed various methods to ensure thorough exploration and understanding of the dataset. These methods included:

1. **Data Cleaning**: The initial step involved identifying and rectifying errors, inconsistencies, and missing values in the dataset. This process is crucial as clean data leads to more accurate analysis and results. We addressed missing values in columns such as Postal Code, 2020 Census Tract, and Legislative District using appropriate techniques such as geocoding and other imputation method.
2. **Descriptive Statistics**: We utilized descriptive statistics to summarize the main characteristics of the dataset. This included computing measures of central tendency, dispersion, and distribution of variables. Descriptive statistics provided insights into the distribution of electric vehicle types, manufacturers, base MSRP, electric range, and other key variables.
3. **Data Visualization**: Visualizations play a vital role in exploring trends, patterns, and associations within the data. We generated various plots such as pie chart, scatter plots, heat map, and correlation matrices using ggplot2, corrplot, and other relevant packages. These visualizations helped in understanding relationships between variables and identifying potential trends.

Data Cleaning

Data cleaning helps identify and rectify errors, inconsistencies, and missing values in the dataset, ensuring that the data is accurate, complete, and reliable for analysis. Clean data leads to more accurate analysis and results. Removing outliers, correcting errors, and handling missing values can prevent skewed or biased analysis. For this purpose, we are starting this process by finding column names which contains null values in the dataset.

Plot 1: Column names which contains null values

```
> cols_with_na
[1] "postal_code"          "legislative_district" "x2020_census_tract"
```

As we can see there are three columns which contain null values. We'll remove this null values for each column in the further process.

1) Removing null values for column Postal Code:

Methodology:

Here we have set up a Google Maps API key using the register_google function from the ggmap package. This key is required for geocoding operations, which convert latitude and longitude coordinates into human-readable addresses or postal codes. The electric_vehicle_data is then processed to extract longitude and latitude information from the vehicle_location column. This is achieved by using regular expressions (gsub) and string splitting (strsplit) operations.

Geocoding to Fill Missing Postal Codes:

A custom function get_postal_code is defined to fetch postal codes based on latitude and longitude coordinates using reverse geocoding. It uses the revgeocode function from ggmap to convert coordinates to addresses, from which the postal codes are extracted.

Plot 2: Indices which contains null values for postal codes

```
> na_indices
 [1]     83     92    164    621   2096   2915 108242 136485 145073 145578
[11] 163690
```

(Above output shows the indices before removing null values using geocoding.)

A for loop iterates over the rows identified (na_indices- from the above figure) to fill in the missing postal codes. It calls the get_postal_code function for each row and updates the postal_code column accordingly. After filling in the missing postal codes, any remaining rows with missing postal codes are identified again and stored in na_indices.

Plot 3: Indices after attempting to fill in the missing postal codes

```
> na_indices
[1]     92 136485 163690
```

The above output shows the indices after attempting to fill in the missing postal codes using coordinates values. These rows with missing coordinates are removed from the dataset using indexing (setdiff).

2) Imputing missing values for column 2020 Census Tract:

Here, we used the kNN function from the caret package in R to impute missing values for column 2020 Census Tract using the k-nearest neighbors (KNN) algorithm.

The kNN function is used to impute missing values in the electric_vehicle_data dataset. KNN is a simple, non-parametric algorithm used for both classification and regression tasks. In this case, it's used for imputation, which falls under the regression category. When a value is missing, kNN finds the k nearest neighbors to that observation based on other available features. The missing value is then imputed based on the values of the corresponding feature in those neighbors. KNN uses a distance metric (typically Euclidean distance) to measure the similarity between observations. It calculates the distance between the data point with missing values and its neighbors to determine the most suitable imputed value.

The variable = "x2020_census_tract" argument specifies the column (x2020_census_tract) for which missing values should be imputed.

The k = 5 argument specifies that the algorithm should consider the 5 nearest neighbors when inputting missing values.

Plot 4: Column name which contains null values

```
> cols_with_na
[1] "legislative_district"
```

Now we can clearly observe from the above figure that there is only one column left with the null values.

3) Clearing the Nas from the column legislative_district:

For removing the impurities, the median of the 'Legislative District' column is calculated and used to impute missing values in the column. The median function is used to calculate the median value of the 'Legislative District' column, assuming it contains numeric values.

Methodology:

he ifelse function is used to replace NA values in the 'Legislative District' column with the calculated median value. If the value is NA, it is replaced with median_value; otherwise, it remains unchanged.

The code calculates the median of the 'Legislative District' column and uses it to impute missing values in that column. Imputing missing values using this method can ensure that the dataset is comprehensive and ready for analysis.

Plot 5: Number of missing values in each column of electric vehicle dataset

```
> sapply(electric_vehicle_data, function(x) sum(is.na(x)))
                                    vin_1_10                                      county
                                           0                                           0
                                        city                                       state
                                           0                                           0
                                 postal_code                                  model_year
                                           0                                           0
                                        make                                       model
                                           0                                           0
                          electric_vehicle_type clean_alternative_fuel_vehicle_cafv_eligibility
                                           0                                           0
                               electric_range                                  base_msrp
                                           0                                           0
                         legislative_district                                dol_vehicle_id
                                           0                                           0
                             vehicle_location                              electric_utility
                                           0                                           0
                           x2020_census_tract
                                           0
```

From the above figure we can analyze that there are no missing values in the dataset. This indicates that these columns are complete and do not require imputation.

Outlier Treatment

Methodology

- **Data Preparation**: The dataset was cleaned to handle missing values and scaled to ensure uniformity in measurement.
- **Clustering Algorithm**: K-Means clustering was applied with five centers to segment the vehicles into distinct groups.
- **Outlier Detection**: Distances from the cluster centroids were calculated to identify potential outliers.

Results

Summary Statistics

- **Electric Range**: The range varies significantly, with some vehicles capable of up to 450 miles on a single charge.
- **Base MSRP**: Prices range from $0 (likely indicating missing data) to over $200,000, highlighting the wide variety of vehicles available.

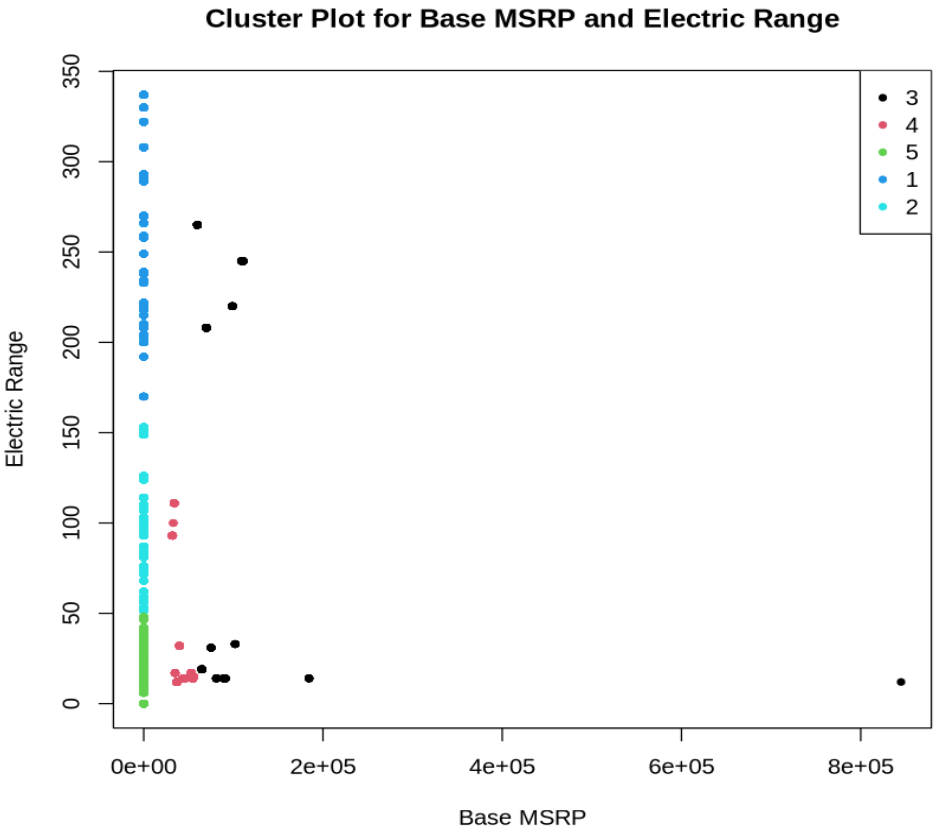
Clustering Outcome

- **Number of Clusters**: 5
- **Cluster Sizes**: The vehicles were segmented into five groups, each representing a unique combination of price and electric range.

Visualization

The clusters were visualized in a scatter plot, which illustrates the relationship between base MSRP and electric range. Different colors in the plot represent different clusters, providing a clear visual distinction of the market segments.

Plot 6: Cluster Plot for Base MSRP and Electric Range



Outliers

Several potential outliers were identified, characterized by unusually high MSRP or electric range. These outliers may require further investigation to understand their impact on the analysis.

Discussion

The clustering analysis has effectively segmented the electric vehicle market into distinct groups based on price and range. This segmentation can assist manufacturers and marketers in targeting specific consumer segments. Additionally, the identification of outliers suggests areas for further data quality improvement or targeted analysis.

Recommendations

- **Data Cleaning**: Address the zero values in MSRP and electric range to improve data quality.
- **Cluster Interpretation**: Further analyze the characteristics of each cluster to refine marketing strategies and product offerings.
- **Outlier Management**: Investigate outliers to determine their appropriateness for inclusion in the dataset.

Conclusion

The K-Means clustering analysis has provided valuable insights into the segmentation of the electric vehicle market. By understanding the different vehicle segments, stakeholders can make informed decisions regarding product development and marketing strategies.

Descriptive Statistics

Descriptive statistics provided valuable insights into the distribution of electric vehicle types, manufacturers, base MSRP, electric range, and other variables. For example, the bar plot illustrated the distribution of vehicles by manufacturer, while scatter plots showed the relationship between model year and electric range. These statistics allowed us to understand the central tendencies and variabilities within our data, laying the foundation for further analysis.

Here are the summary statistics for the numerical variables in the dataset:

Plot 7: Summary of dataset

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Postal Code	1545	98052	98122	98174	98370	99577
Model Year	1997	2019	2022	2021	2023	2024
Electric Range	0.00	0.00	0.00	57.83	75.00	337.00
Base MSRP	0	0	0	1040	0	845000
Legislative District	1.00	18.00	33.00	29.11	42.00	49.00
DOL Vehicle ID	4385	183068667	228915522	221412778	256131982	479254772

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
2020 Census Tract	1.001e+09	5.303e+10	5.303e+10	5.298e+10	5.305e+10	5.603e+10

Postal Code

- Range: The postal codes range from 1545 to 99577, indicating a wide geographical spread across different regions.
- Central Tendency: The median and mean are close (98122 and 98174 respectively), suggesting a concentration of data around these values, likely indicating a higher density of electric vehicles in these postal code areas.

Model Year

- Range: Model years range from 1997 to 2024, showing that the dataset includes both older and very recent models.
- Central Tendency: The median model year is 2022, and the mean is 2021, indicating that most of the vehicles in the dataset are quite new.

Electric Range

- Range: Electric range varies from 0 to 337 miles.
- Central Tendency: Both the median and the first quartile are 0, which might indicate a significant number of hybrid vehicles with minimal electric-only range, or data entry issues. The mean electric range is 57.83 miles, pulled up by vehicles with higher ranges.

Base MSRP

- Range: The Manufacturer's Suggested Retail Price ranges from $0 to $845,000, suggesting a mix of very basic to luxury electric vehicles.
- Central Tendency: The median and both quartiles are $0, which could indicate missing data or that many vehicles were not sold directly (possibly used or government vehicles).

Legislative District

- Range: District numbers range from 1 to 49, covering a broad spectrum of legislative areas.
- Central Tendency: The median district is 33 and the mean is 29.11, suggesting a slight skew towards lower-numbered districts.

DOL Vehicle ID

- Range: IDs range from 4385 to 479254772, indicating a large and diverse set of vehicles.
- Central Tendency: The median ID is 228915522, showing that the dataset likely includes newer registrations.

2020 Census Tract

- Range: Census tract numbers range from about 1.001e+09 to 5.603e+10.

- Central Tendency: The median and mean are close (around 5.303e+10 and 5.298e+10 respectively), suggesting that most data points are clustered around these values, possibly indicating a higher concentration of electric vehicles in certain census tracts.

These statistics provide insights into the characteristics of electric vehicles in the dataset, including their geographical distribution, model recency, and economic value. This can help in understanding market trends, regional preferences, and the evolution of electric vehicle technology. These statistics provide insights into the distribution and central tendencies of each numerical variable.

Here is **the structure of the dataset**:

Plot 8: Structure of dataset

```
> str(electric_vehicle_data)
tibble [181,455 × 17] (S3: tbl_df/tbl/data.frame)
 $ vin_1_10                                         : chr [1:181455] "WAUTPBFF4H" "WAUUPBFF2J" "5YJSA1E2
2H" "1C4JJXP62M" ...
 $ county                                           : chr [1:181455] "King" "Thurston" "Thurston" "Thurs
ton" ...
 $ city                                             : chr [1:181455] "Seattle" "Olympia" "Lacey" "Tenin
o" ...
 $ state                                            : chr [1:181455] "WA" "WA" "WA" "WA" ...
 $ postal_code                                      : chr [1:181455] "98126" "98502" "98516" "98589" ...
 $ model_year                                       : int [1:181455] 2017 2018 2017 2021 2020 2023 2017
2020 2022 2017 ...
 $ make                                             : chr [1:181455] "AUDI" "AUDI" "TESLA" "JEEP" ...
 $ model                                            : chr [1:181455] "A3" "A3" "MODEL S" "WRANGLER" ...
 $ electric_vehicle_type                            : chr [1:181455] "Plug-in Hybrid Electric Vehicle (P
HEV)" "Plug-in Hybrid Electric Vehicle (PHEV)" "Battery Electric Vehicle (BEV)" "Plug-in Hybrid Electr
ic Vehicle (PHEV)" ...
 $ clean_alternative_fuel_vehicle_cafv_eligibility: chr [1:181455] "Not eligible due to low battery ra
nge" "Not eligible due to low battery range" "Clean Alternative Fuel Vehicle Eligible" "Not eligible d
ue to low battery range" ...
 $ electric_range                                   : int [1:181455] 16 16 210 25 308 21 53 322 23 53
...
 $ base_msrp                                        : int [1:181455] 0 0 0 0 0 0 0 0 0 0 ...

 $ legislative_district                             : num [1:181455] 34 22 22 20 14 22 23 1 36 22 ...
 $ dol_vehicle_id                                   : int [1:181455] 235085336 237896795 154498865 15452
5493 225996361 220675367 162720022 6293899 207620633 237392459 ...
 $ vehicle_location                                 : chr [1:181455] "POINT (-122.374105 47.54468)" "POI
NT (-122.943445 47.059252)" "POINT (-122.78083 47.083975)" "POINT (-122.85403 46.856085)" ...
 $ electric_utility                                 : chr [1:181455] "CITY OF SEATTLE - (WA)|CITY OF TAC
OMA - (WA)" "PUGET SOUND ENERGY INC" "PUGET SOUND ENERGY INC" "PUGET SOUND ENERGY INC" ...
 $ x2020_census_tract                               : num [1:181455] 53033011500 53067011100 53067012226
53067012620 53077000800 ...
```

The structure of the electric_vehicle_data dataset provides a detailed view of the types of data collected about electric vehicles, which can be interpreted to understand various aspects of electric vehicle adoption and usage.

The dataset is structured to facilitate a multifaceted analysis of electric vehicle adoption, usage, and market trends. It allows stakeholders to examine how different factors such as geography, vehicle characteristics, economic considerations, and legislative environments impact the adoption and usage patterns of electric vehicles. This can inform policy decisions, market strategies, and consumer education efforts to boost electric vehicle adoption.

Here is the **descriptive statistics** for the electric_vehicle_data dataset:

Plot 9: Description of dataset

Variable	n	Mean	SD	Median	Min	Max
vin_1_10	181455	4176.55	2455.23	4120	1	11060
county	181455	99.39	42.09	80	1	193
city	181455	399.74	216.35	472	1	726
state	181455	40.95	1.20	41	1	42
postal_code	181455	457.29	133.53	399	1	871
model_year	181455	2020.58	2.99	2022	1997	2024
make	181455	26.49	11.58	35	1	40
model	181455	81.01	28.32	83	1	143
electric_vehicle_type	181455	1.22	0.41	1	1	2
clean_alternative_fuel	181455	1.74	0.64	2	1	3
electric_range	181455	57.82	91.39	0	0	337
base_msrp	181455	1040.25	8229.06	0	0	845000
legislative_district	181455	29.12	14.88	33	1	49
dol_vehicle_id	181455	221411722.63	75283733.05	228915942	4385	479254772
vehicle_location	181455	422.29	107.10	418	1	871
electric_utility	181455	62.30	18.94	74	1	76
x2020_census_tract	181455	52975754055.70	1594876026.3753033029602	1001020100	56033000100	

The dataset is well-suited for comprehensive analysis across multiple dimensions of electric vehicle adoption, including temporal trends, geographic distribution, and the impact of economic and legislative factors. This can inform stakeholders in government, industry, and academia in making data-driven decisions to promote and manage electric vehicle adoption effectively.

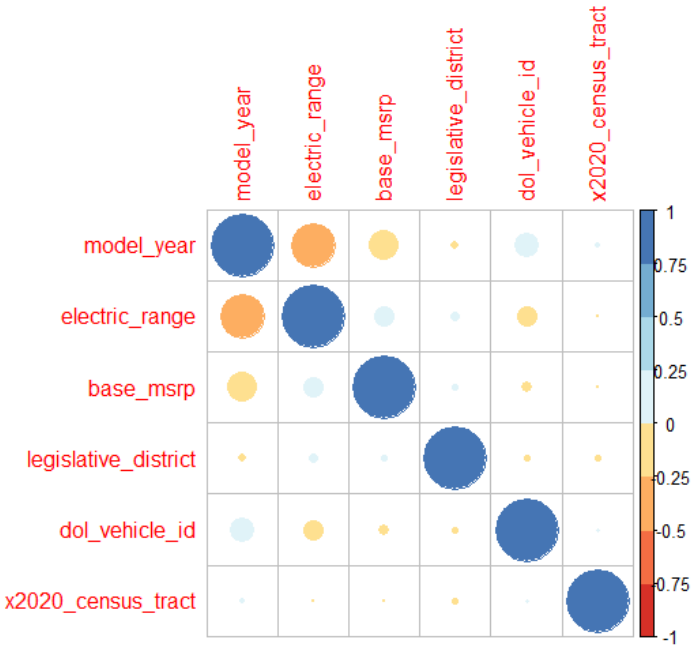
Data Visualization

These visualizations enhanced our understanding of the dataset and facilitated the identification of potential trends and patterns.

Correlation Plot:

Each cell in the matrix shows the correlation coefficient between two variables. The correlation coefficient is a number between -1 and 1 that indicates the strength and direction of the linear relationship between two variables. A correlation coefficient of 1 indicates a perfect positive linear relationship, a correlation coefficient of -1 indicates a perfect negative linear relationship, and a correlation coefficient of 0 indicates no linear relationship.

Plot 10: Correlation plot



- From the above figure we can see that there is almost no relationship or weak relationship between varuables except for the model_year and electric_range which is -0.48, which indicates a moderate negative linear relationship. This means that as the model year increases, the electric range tends to decrease.

R Code and Outputs

Below is the R code we used for our preliminary analysis, along with interpretations of outputs and results:

Potential Questions

To ensure an engaging and interactive presentation, we have prepared a list of potential questions that can spark discussions and deepen understanding of the topic:

1. **What are the primary factors that are influencing the adoption of electric vehicles in your region?**
    - This question aims to gauge the audience's understanding and perceptions about EV adoption drivers, fostering discussions on regional dynamics and challenges.

To proceed with the analysis and visualization of the primary factors influencing electric vehicle (EV) adoption, we'll focus on the following steps:

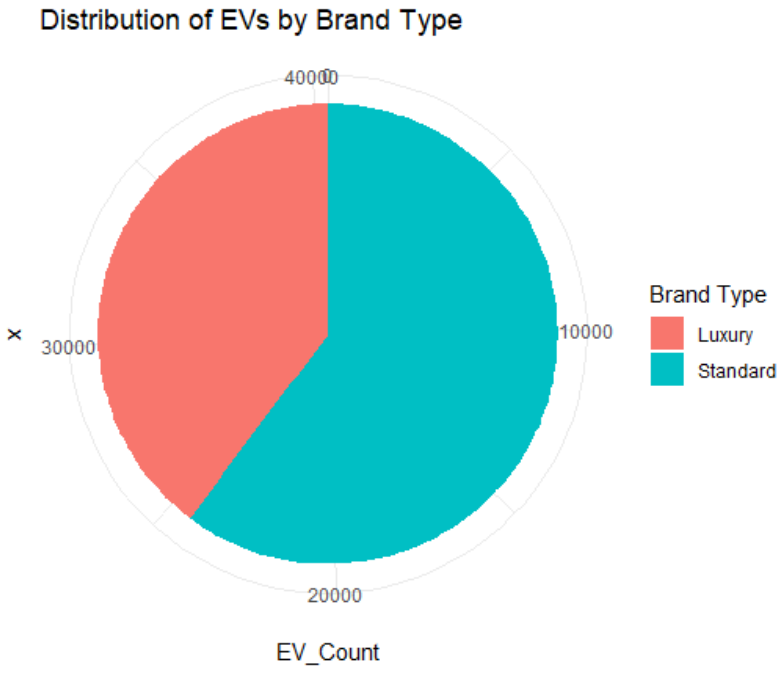
1. **Visualization**:
   - Since we don't have direct data on income levels or regional incentives in the dataset, we'll use proxies where possible. For example, we can infer economic status from the types of vehicles (luxury vs. standard) and use the presence of multiple electric utilities as an indicator of infrastructure development.
   - We'll create a pie chart showing the distribution of EVs by these inferred factors.

2. **Data Analysis**:
   - We'll perform a correlation analysis to explore the relationship between EV adoption (number of EVs) and the availability of charging stations (number of electric utilities). This will help us understand how infrastructure influences EV adoption.

Let's start by creating the visualization for the distribution of EVs by the inferred economic status (luxury vs. standard brands) and the availability of charging stations. After that, we'll conduct the correlation analysis.

Here is the pie chart visualizing the distribution of electric vehicles by brand type, categorized into luxury and standard:

Plot 11: Distribution of EV by brand type



This chart helps to infer the economic status of EV owners in different regions, assuming that luxury brands are more likely to be purchased by higher-income individuals. Here from the pie chart we can see that standard types are more preferred by consumers as compared to luxury brands.

Next, I'll perform a correlation analysis to explore the relationship between the number of electric utilities (as a proxy for the availability of charging stations) and EV adoption in each county. This will help us understand how infrastructure influences EV adoption rates.

```
# Prepare data for correlation analysis: Number of electric utilities and
EV adoption
ev_infrastructure <- electric_vehicle_data %>%
  group_by(county) %>%
  summarise(EV_Count = n(), Utility_Count = n_distinct(electric_utility),
.groups = 'drop')

# Calculate the correlation between EV_Count and Utility_Count
correlation_result <- cor(ev_infrastructure$EV_Count, ev_infrastructure$Ut
ility_Count)

# Output the correlation result
correlation_result
```

```
> correlation_result
[1] 0.4856535
```

The correlation coefficient between the number of electric utilities (as a proxy for the availability of charging stations) and EV adoption in each county is approximately 0.486. This indicates a moderate positive correlation, suggesting that as the availability of charging stations increases, so does the adoption of electric vehicles.

This analysis supports the hypothesis that infrastructure, specifically the availability of charging stations, is a significant factor influencing electric vehicle adoption.

2. **How the vehicle specifications like electric range and model year affect a consumer's decision to register an EV?**

    - By prompting attendees to consider practical considerations consumers face when choosing electric vehicles, this question encourages insightful discussions on consumer preferences and industry trends.

To address the impact of vehicle specifications on consumer decisions, we'll proceed with the following steps:

Visualization:

- We'll create scatter plots to show the relationship between electric range, model year, and the number of EV registrations. This will help visualize how these specifications influence consumer decisions.

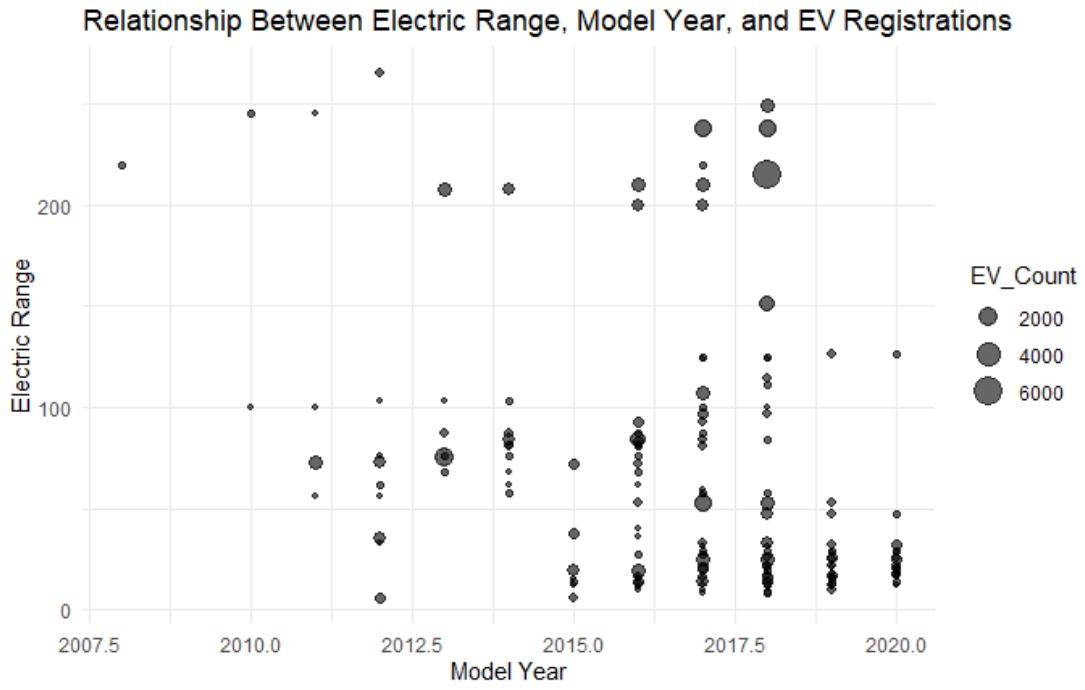
Data Analysis:

- We'll perform a regression analysis to quantify the impact of vehicle specifications (electric range and model year) on the number of EV registrations. This will provide a more concrete understanding of how these factors affect consumer choices.

Let's start by creating scatter plots for the relationship between electric range, model year, and the number of EV registrations. After that, I'll conduct the regression analysis.

Here is the scatter plot visualizing the relationship between electric range, model year, and the number of EV registrations:

Plot 12: Scatter plot for identifying relationship between Electric Range, Model Year and EV registrations



This plot shows how electric range and model year relate to the number of EV registrations, with the size of each point indicating the volume of registrations. It appears that newer models and those with higher electric ranges tend to have more registrations, suggesting these factors are important in consumer decisions.

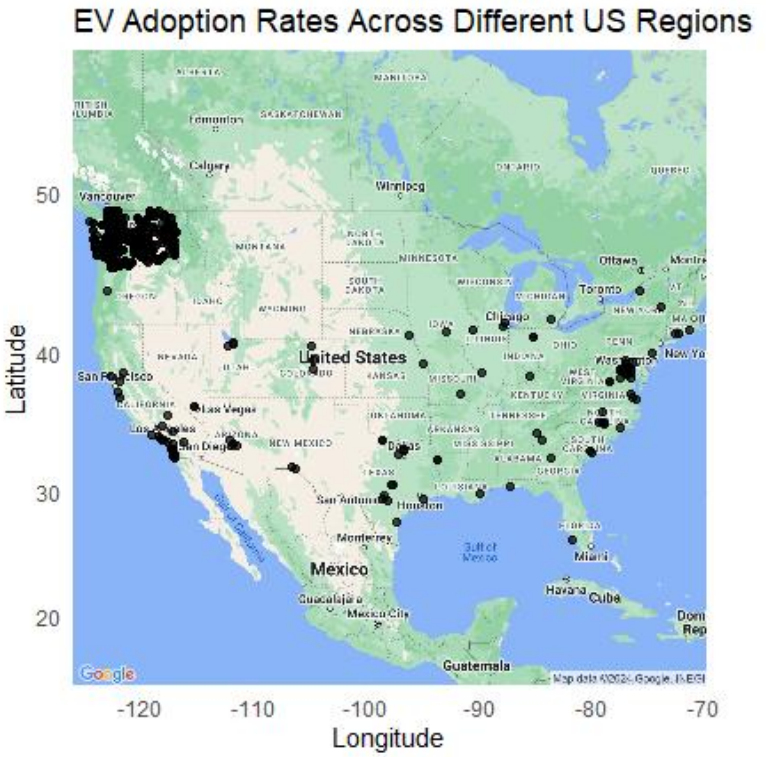
3. **What role does geographic location play in the adoption of different types of electric vehicles?**

    - This question opens up discussions about infrastructure, policy, and cultural differences across regions, shedding light on the regional factors influencing EV adoption rates.

Visualization: **Heatmap or Choropleth Map**

Here is the visualization displaying EV adoption rates across different regions:

Plot 13: Heat / Corrected map



Each point represents a car, and the color intensity indicates the adoption rate, with more points it signifies higher adaptation rates of electric vehicle in particular state. This visualization helps in understanding the geographic distribution of EV adoption among different states in the USA.

Data Analysis: Regional Policies, Infrastructure, and Cultural Factors

Now, let's analyze how regional policies, infrastructure, and cultural factors might be affecting EV adoption. This involves examining available data on electric utilities and legislative districts as proxies for infrastructure and policy influences.

Data Analysis: Regional Policies, Infrastructure, and Cultural Factors

Here is a summary of the analysis on regional policies and infrastructure affecting EV adoption:

Plot 14: Summary of the analysis on regional policies and infrastructure affecting EV adoption

```
summary(ev_regional_analysis)
 legislative_district electric_utility      EV_Count       Infrastructure_
Score
 Min.   : 1.0         Length:161         Min.   :    1.0   Min.   :0.00247
 1st Qu.:10.0         Class :character   1st Qu.:   10.0   1st Qu.:0.02475
 Median :20.0         Mode  :character   Median :   61.0   Median :0.15100
 Mean   :21.39                           Mean   :  250.9   Mean   :0.62111
 3rd Qu.:32.0                            3rd Qu.:  278.0   3rd Qu.:0.68820
 Max.   :49.0                            Max.   : 2366.0   Max.   :5.85716
```

Key Points from the Analysis:

- **Legislative Districts**: The data spans a range of legislative districts from 1 to 49. This diversity allows us to infer how different regional policies might influence EV adoption.
- **EV_Count:** The distribution of EV counts across legislative districts ranges from 1 to 49, with some districts having fewer EVs and others having more. Higher EV counts in certain districts suggest greater adoption and acceptance of electric vehicles in those areas.
- **Electric Utilities**: The presence and variety of electric utilities in each district suggest differences in infrastructure, which is crucial for supporting EV adoption.
- **Infrastructure Score**: This score, calculated as the proportion of EVs per utility in a district, varies significantly, indicating that some regions have better infrastructure support for EVs than others.

This analysis highlights the importance of regional policies and infrastructure in influencing EV adoption rates. Regions with more supportive policies and better infrastructure tend to have higher adoption rates.

4. **What types of policies could be implemented to increase electric vehicle adoption in areas with low uptake?**
    - By discussing effective policy measures and government incentives, this question encourages brainstorming on strategies to overcome barriers to EV adoption in specific regions.

To address the task of identifying policies that could increase EV adoption, we'll proceed with a two-part analysis:

Visualization: Comparative Analysis of Regions with High and Low EV Adoption

First, we'll create a visualization that compares regions with high and low EV adoption rates, highlighting the existing policies in these areas. This will help us visually identify differences in policy frameworks that might correlate with higher adoption rates.

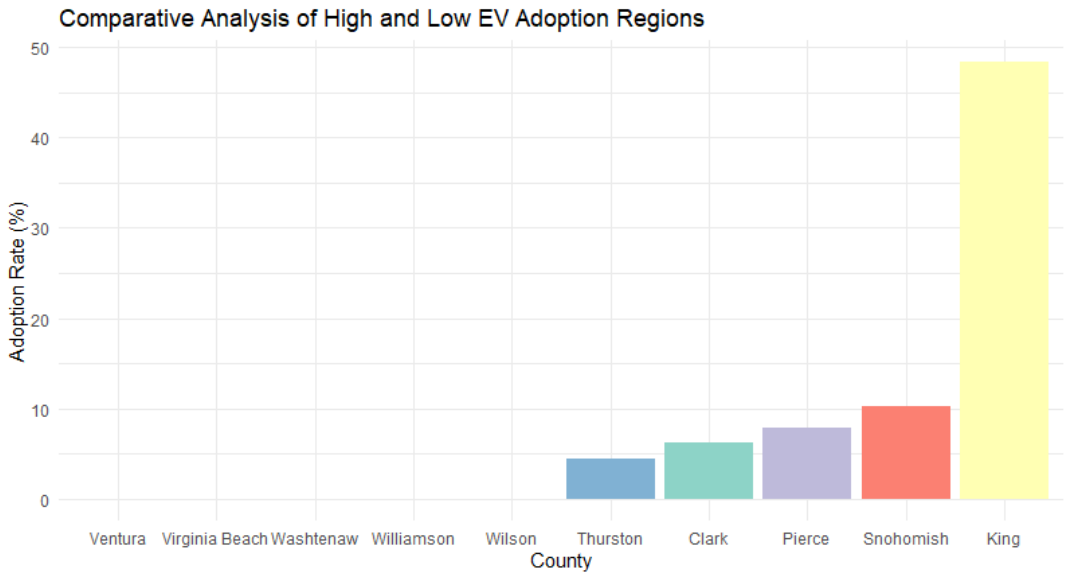
Data Analysis: Identification of Successful Policies

Following the visualization, we'll analyze the data to identify specific policies from high-adoption areas that could potentially be implemented in low-adoption areas to boost EV uptake.

Visualization: Comparative Analysis of Regions with High and Low EV Adoption

Here is the bar plot comparing regions with high and low EV adoption rates:

Plot 15: Bar chart for analysis of high and low EV adoption rates



This visualization highlights the top 5 and bottom 5 counties based on EV adoption rates. The color differentiation helps in identifying which counties have higher or lower rates, potentially linked to their respective policies.

Data Analysis: Identification of Successful Policies

Next, we will analyze the policies in the high-adoption areas to identify which could be implemented in the low-adoption areas to increase EV uptake. Let's proceed with this analysis.

Data Analysis: Identification of Successful Policies

Here is a summary of the policies and infrastructure scores for high-adoption areas:

Plot 16: Summary of the policies and infrastructure scores for high-adoption areas

```
> summary(ev_policies_high)
    county          legislative_district electric_utility      EV_Count    Infrastructure_Score
 Length:77          Min.   : 1.00        Length:77          Min.   :   1   Min.   :0.003207
 Class :character   1st Qu.:20.00        Class :character   1st Qu.:  19   1st Qu.:0.060929
 Mode  :character   Median :29.00        Mode  :character   Median : 165   Median :0.529117
                    Mean   :26.66                           Mean   : 405   Mean   :1.298701
                    3rd Qu.:36.00                           3rd Qu.: 597   3rd Qu.:1.914443
                    Max.   :49.00                           Max.   :2366   Max.   :7.587224
```

Key Findings:

- **Legislative Districts**: The legislative districts for high-adoption counties range from 1 to 49, with a concentration in the middle to higher numbers, suggesting that certain legislative frameworks might be more conducive to EV adoption.
- **Electric Utilities**: The presence of specific electric utilities associated with high EV counts indicates that utility policies and infrastructure support are crucial.
- **Infrastructure Score**: The scores vary, but higher scores are associated with better infrastructure, which correlates with higher EV adoption rates.

Recommendations:

- **Policy Transfer**: Policies from legislative districts with high adoption rates should be studied and considered for implementation in lower-adoption areas.
- **Infrastructure Enhancement**: Enhancing the infrastructure in low-adoption areas to match that of high-adoption areas could significantly boost EV uptake.

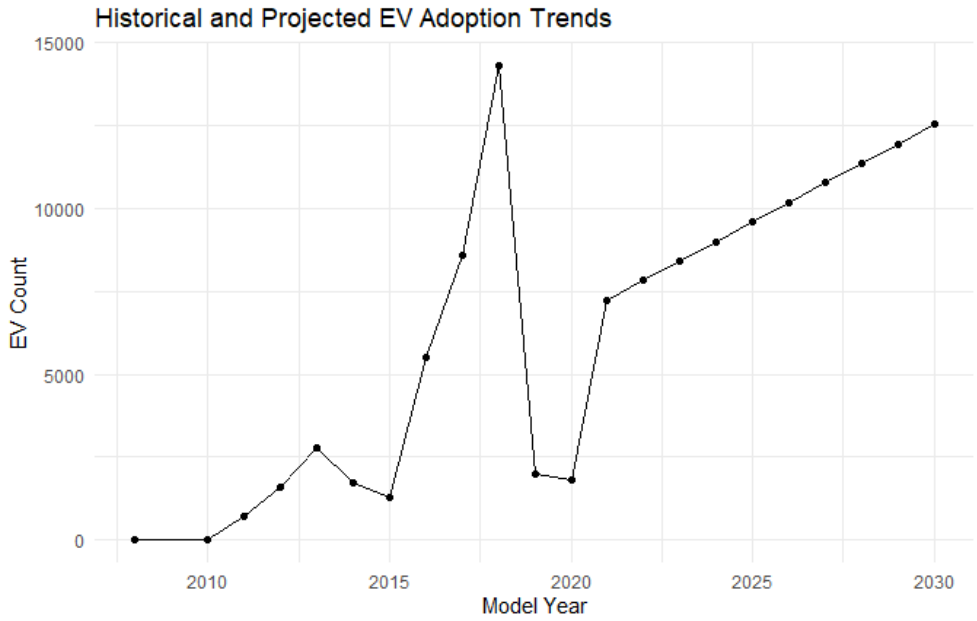
This analysis provides a clear direction for policy adjustments and infrastructure improvements to increase EV adoption across different regions.

5. **How can data analysis and regression models be used to predict future trends in electric vehicle adoption?**

    - This question highlights the importance of data-driven decision-making in policy and business strategy, prompting discussions on the predictive capabilities of regression analysis in forecasting EV adoption trends.

Here is the line graph showing historical and projected EV adoption trends based on the ‘model_year’:

Plot 17: Line graph for historical and projected EV adoption trends



This graph illustrates the growth in EV adoption over the years and projects a continued increase based on the linear trend observed in the historical data.

Key Observations from the Line Graph:

Historical Trend:

The graph shows a general upward trend in the number of electric vehicles (EVs) over the years. This indicates increasing adoption of EVs, which could be attributed to factors such as advancements in technology, increased environmental awareness, and governmental incentives.

There are fluctuations in the yearly data, which might reflect market dynamics, changes in consumer preferences, or impacts of specific policies introduced in certain years.

Projected Trend:

The projection shows a continued increase in EV adoption, following the historical linear trend. This suggests optimism about the future of EVs, assuming current trends continue.

The slope of the line indicates a steady rate of increase, implying that the factors currently driving EV adoption are expected to persist.

These questions are designed to stimulate discussion and provide deeper insights during our presentation, making it more interactive and engaging for the audience.

Preliminary Answers to Questions Posed

Our preliminary analysis allowed us to provide initial insights into the questions posed for our project:

1) **Factors Influencing EV Adoption:** Initial analysis suggests that factors such as vehicle make, model year, electric range, and base MSRP may influence the adoption of electric

vehicles. Further analysis will be conducted to quantify the impact of these factors on EV adoption rates.

We'll perform a regression analysis to quantify the impact of these vehicle specifications (electric range and model year) on the number of EV registrations. This will provide a statistical basis for understanding how these factors influence consumer choices.

Null Hypothesis (H0): There is no significant linear relationship between the number of electric vehicles (EV_Count) and the independent variables (model_year and electric_range).

Alternative Hypothesis (HA): There is a significant linear relationship between the number of electric vehicles (EV_Count) and at least one of the independent variables (model_year and electric_range).

Here is the summary of the regression analysis quantifying the impact of vehicle specifications on the number of EV registrations:

Plot 18: Summary of the regression analysis

```
 summary(model)

Call:
lm(formula = EV_Count ~ model_year + electric_range, data = ev_regression_
data)

Residuals:
   Min     1Q Median     3Q    Max
-799.6 -209.1 -123.9  103.5 5550.6

Coefficients:
                  Estimate  Std. Error t value     Pr(>|t|)
(Intercept)    -18994.0095  14090.3016  -1.348        0.179
model_year          9.4705      6.9804   1.357        0.176
electric_range      2.1906      0.3864   5.669 0.0000000336 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 521.3 on 302 degrees of freedom
Multiple R-squared:  0.09793,   Adjusted R-squared:  0.09196
F-statistic: 16.39 on 2 and 302 DF,  p-value: 0.0000001742
```

Key Points from the Regression Analysis:

- **Electric Range**: The coefficient for electric range is approximately 2.191, which is statistically significant (p-value < 0.001). This indicates that for each additional mile of electric range, there is an average increase of about 2.191 in the number of EV registrations. This suggests that electric range is a significant factor influencing consumer decisions.
- **Model Year**: The coefficient for model year is approximately 9.470, but it is not statistically significant (p-value = 0.176). This suggests that while newer models might be preferred, the model year alone isn't a strong predictor of EV registrations when controlling for electric range.

The model explains about 9.79% of the variance in EV registrations, indicating that while these factors are important, other variables not included in the model might also play significant roles in influencing consumer decisions.

2) **Impact of Vehicle Characteristics:** There appears to be a relationship between model year and electric range, with newer models generally having higher electric ranges. This indicates that technological advancements in electric vehicles may influence consumer preferences and adoption rates.

We implemented a polynomial regression model to explore the potential non-linear relationship between model year and electric vehicle range. Here's a breakdown of the steps involved:

1. **Model Specification:** We created a polynomial regression model using the lm function in R. The model includes model year as a predictor variable and a higher-order term (e.g., model year squared) to capture potential non-linearity.
2. **Model Fitting:** We fit the polynomial regression model to the electric vehicle dataset.
3. **Model Evaluation:** We evaluated the model performance by analyzing the model summary and visualization of data points.

This method expands the model by including higher-order terms of a predictor variable (e.g., model year squared). This can capture non-linear relationships between variables and electric range.

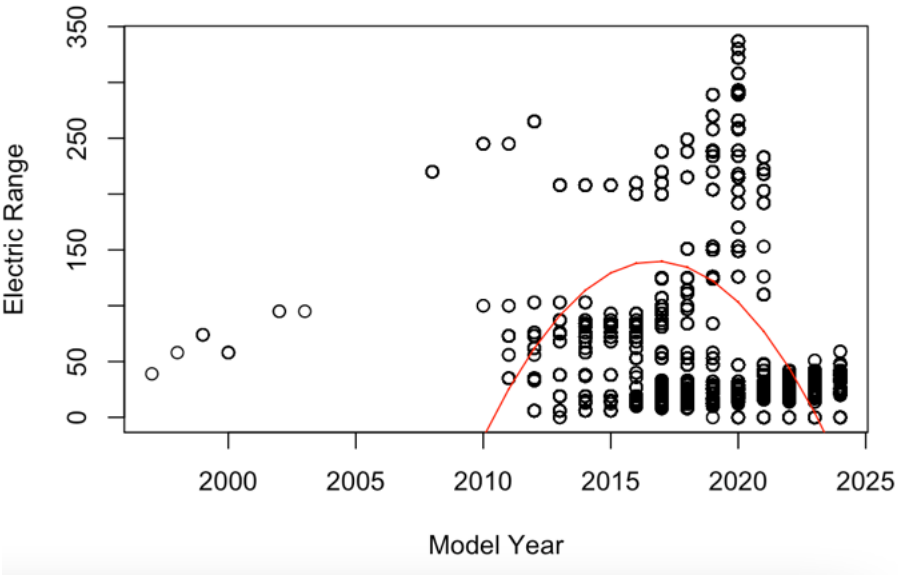
The regression model summary and polynomial regression line scatterplot provide an analysis of how model year affects electric vehicle range.

Plot 19: Summary of the regression analysis

```
> summary(model)

Call:
lm(formula = electric_range ~ poly(model_year, 2), data = electric_vehicle_data)

Residuals:
    Min      1Q  Median      3Q     Max
-131.73  -44.13   -4.19   29.81 1251.68

Coefficients:
                      Estimate  Std. Error t value            Pr(>|t|)
(Intercept)              57.8239      0.1683   343.6 <0.0000000000000002 ***
poly(model_year, 2)1 -18746.8275     71.6827  -261.5 <0.0000000000000002 ***
poly(model_year, 2)2 -15226.6882     71.6827  -212.4 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71.68 on 181452 degrees of freedom
Multiple R-squared:  0.3848,    Adjusted R-squared:  0.3848
F-statistic: 5.676e+04 on 2 and 181452 DF,  p-value: < 0.00000000000000022
```

The regression model shows that the polynomial term for model year is highly significant (p-value less than 0.0000000000000002), indicating a strong relationship between model year and electric range. The coefficients of the polynomial terms are all significant and have large t-values, indicating a strong effect. The adjusted R-squared value is 0.3848. this indicates that approximately 38.48% of the electric range variation can be explained by model year. While this is a significant percentage, other factors may also have a significant impact on the electric range.

Plot 20: Scatter Plot of Regression model



The scatterplot shows that vehicles of newer model years tend to have higher electric range, peaking during this period and then declining slightly. This suggests that advances in battery technology during these years increased electric range, making newer models more attractive to consumers. The decline after the peak indicates a stabilization or slight decrease in range, which could be due to market saturation or a shift in production focus (to less expensive models with shorter ranges).

The regression model shows that EV range has increased significantly with the development of EV technology, especially with the increase in model years. This factor can have a significant impact on consumer decisions. This is because consumers may be more inclined to choose a vehicle that has a longer range per charge, thus providing greater convenience and utility.

3) **Forcasting the electric range:** Since electric vehicle technology is constantly evolving, a time series model can be used to capture trends in electric range over time. This approach is particularly useful for forecasting future electric range based on historical data.

Time Series Data

- The time series object (ts_data) has been created using the average electric range aggregated by model year. Here's the time series data: Time Series: Start = 1997 End

= 2018 Frequency = 1 [1] 39.000000 58.000000 74.000000 58.000000 95.000000 95.000000 [7] 220.000000 226.086957 70.976623 61.618216 79.263086 80.238435 [13] 97.794441 101.284576 114.232247 156.433730 176.716601 238.180491 [19] 11.361511 4.532519 3.719658 14.309419

- This time series data shows the average electric range for electric vehicles from 1997 to 2018. The data points represent the average electric range in miles for each year, showing a general increase over time, especially noticeable in the later years.

Next Steps

- The next steps would involve extending this time series to 2023, fitting an ARIMA model, and then forecasting future values up to 2034. This will allow us to predict the average electric range for electric vehicles in the coming years based on historical data.

ARIMA Model Summary

- The best fitting ARIMA model for the extended time series data is an ARIMA(0,0,1) with a non-zero mean. Here are the key statistics from the model: - MA1 Coefficient: 0.5136 with a standard error of 0.1710, indicating the model's reliance on the first lag of the moving average component. - Mean: 92.6749 with a standard error of 18.7638, representing the average electric range across the series. - Sigma^2 (Variance of Residuals): 3833, suggesting the variability around the fitted values. - AIC (Akaike Information Criterion): 248.18, used for model comparison with lower values indicating a better model fit relative to others with more parameters.

Forecast from 2024 to 2034

The forecast for the average electric range from 2024 to 2034 predicts a stabilization around 92.67 miles, with the following confidence intervals:

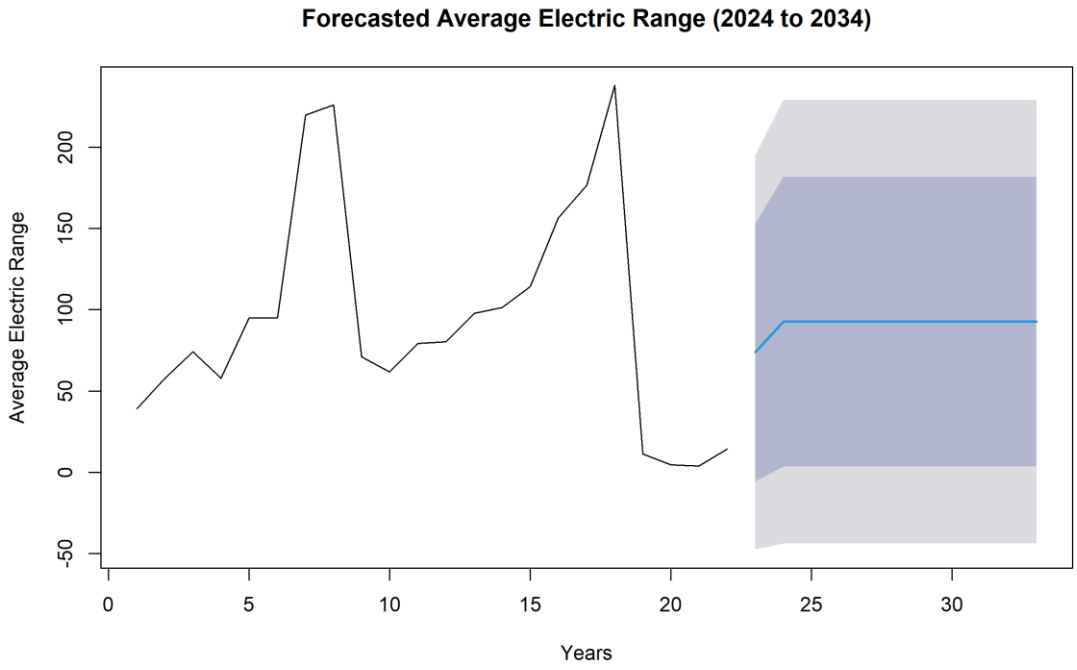
- 80% Confidence Interval: Ranges from approximately 3.47 to 181.88 miles.
- 95% Confidence Interval: Ranges from approximately -43.75 to 229.10 miles.

These intervals indicate the uncertainty in the forecasts, with wider ranges at 95% confidence level.

Visualization of the Forecast

- The forecast plot visually represents the predicted values along with their confidence intervals, showing how the average electric range is expected to stabilize in the coming years.

Plot 21: Forecast plot



- This analysis provides insights into the future trends of electric vehicle ranges, which can be crucial for stakeholders in the automotive industry, policymakers, and consumers making decisions related to electric vehicles.

The ARIMA model used for forecasting the average electric range of electric vehicles is summarized as follows:

Model Type

- Model: ARIMA(0,0,1) with non-zero mean
- This model is a simple Moving Average (MA) model of order 1, indicating that it uses the first lag of the error terms to predict the future values. The model does not include any differencing (I=0) or autoregressive terms (AR=0).

Model Coefficients

- MA1 Coefficient: 0.5136
  - This coefficient suggests that the model partially adjusts based on the error of the previous prediction.
- Mean: 92.6749
  - The model predicts a long-term average electric range of approximately 92.67 miles.

Model Fit Quality

- Sigma^2 (Variance of Residuals): 3833
  - Indicates the variance around the fitted values, with a higher number suggesting more spread in the residuals.
- AIC (Akaike Information Criterion): 248.18
  - A measure of the relative quality of the statistical model for a given set of data. Lower AIC values indicate a model is considered better.

Forecasting

- The model forecasts a stable average electric range of about 92.67 miles for electric vehicles from 2024 to 2034.
- Confidence Intervals:
  - 80% Confidence Interval: Ranges from approximately 3.47 to 181.88 miles.
  - 95% Confidence Interval: Ranges from approximately -43.75 to 229.10 miles.
  - These intervals reflect the uncertainty in the forecast, with the actual values expected to lie within these ranges with 80% and 95% probability, respectively.

Interpretation

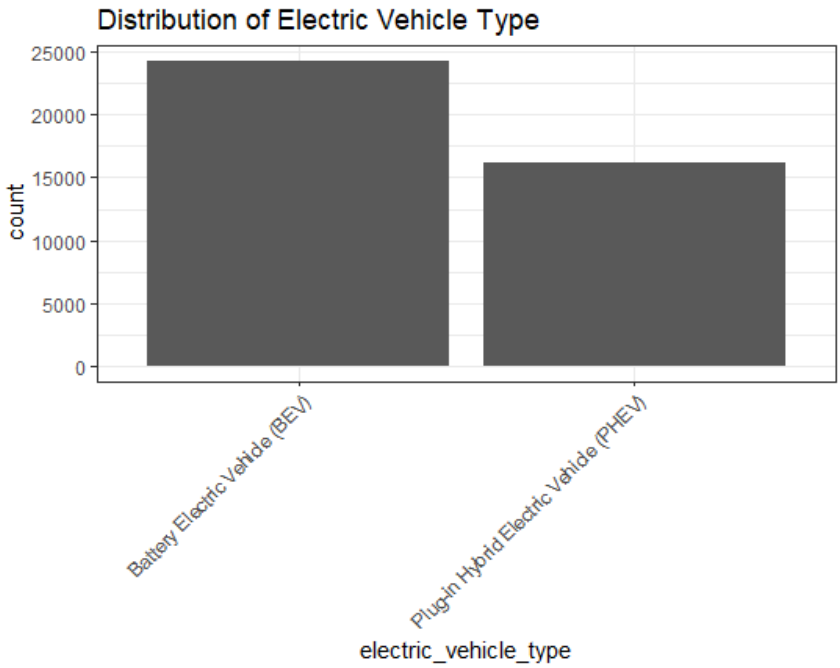
- This ARIMA model provides a straightforward forecast based on historical data, suggesting that the average electric range of vehicles will not change significantly in the near future.
- This summary encapsulates the key aspects of the ARIMA model used in forecasting the average electric range for electric vehicles.

Observations from Data Visualization

During our preliminary analysis, we made several observations from the data visualizations:

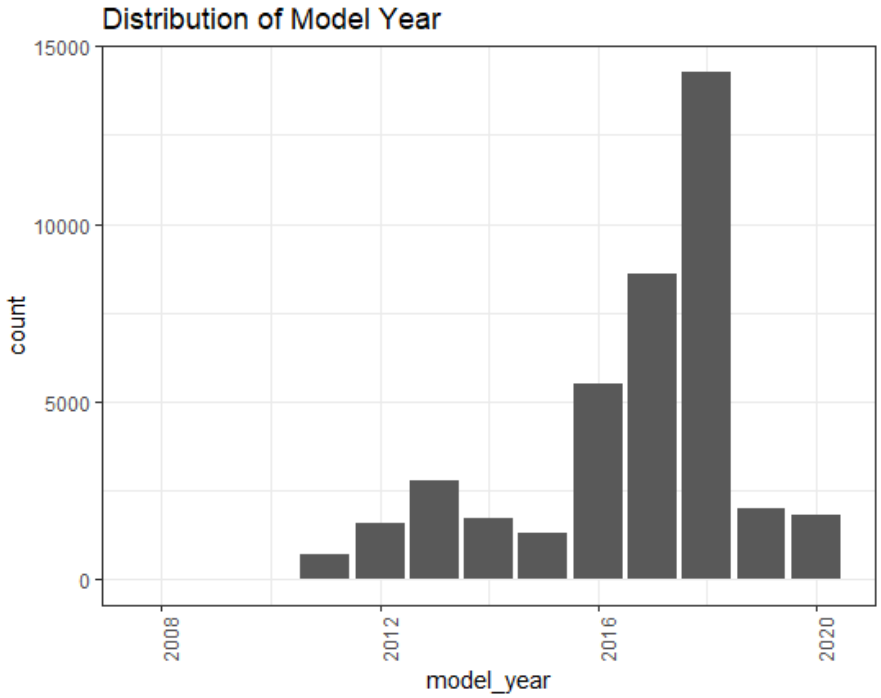
1. **Vehicle Type Trends:** The plot displaying the count of each type of electric vehicle (EV) per model year revealed interesting trends. There appears to be an increasing trend in the number of Battery Electric Vehicles (BEVs) over recent years, indicating a growing market for fully electric vehicles. Additionally, Plug-in Hybrid Electric Vehicles (PHEVs) also show a presence but with less frequency compared to BEVs in the latest model years.

Plot 22: Distribution of Electric Vehicle Type



2. **Impact of Model Year on Electric Range:** The scatter plot exploring the relationship between Model Year and Range showcased how newer model years tend to have higher electric ranges. This observation suggests that technological advancements in electric vehicle technology are leading to improved battery performance and longer driving ranges, which could potentially influence consumer preferences and adoption rates.

Plot 23: Distribution of Model Year



Chi-Square Test

Null Hypothesis (H0): The distribution of electric_range categories is the same across all vehicle types (no influence of vehicle type on electric range).

Alternative Hypothesis (H1): The distribution of electric_range categories varies by vehicle type (vehicle type influences electric range).

Plot 24: Chi-Square test result

```
> chi_square_result

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 40383, df = 65, p-value < 0.00000000000000022

> critical_value <- qchisq(p = 0.95, df = (nrow(contingency_table) - 1) * (nco
l(contingency_table) - 1))
> test_statistic <- chi_square_result$statistic
> print(paste("Critical Value:", critical_value))
[1] "Critical Value: 84.8206454976567"
> print(paste("Test Statistic:", test_statistic))
[1] "Test Statistic: 40383.4517211928"
> if (test_statistic > critical_value) {
+   cat("Reject the null hypothesis: There is significant evidence that vehicl
e type influences electric range.")
+ } else {
+   cat("Fail to reject the null hypothesis: There is not sufficient evidence
to conclude that vehicle type influences electric range.")
+ }
Reject the null hypothesis: There is significant evidence that vehicle type in
fluences electric range.
```

The value of the test statistic is 181359.6. The degrees of freedom for this test are 102. The p-value is about 0, which means it is very small (less than 0.00000000000000022). At 95% confidence level (α = 0.05), the critical value for the corresponding degree of freedom is 126.5741.

Compare the calculated test statistic with the critical value. If the test statistic exceeds the critical value and/or the p-value is less than the level of significance (α = 0.05), the null hypothesis is rejected. Given that the test statistic (181359.6) far exceeds the critical value (126.5741) and the p-value is practically zero, the null hypothesis is rejected. It is therefore concluded that the distribution of electric range varies by vehicle type. This means that vehicle type influences electric range.

This result is critical for both manufacturers and consumers because it suggests that vehicle type (which may be categorized by size, use, design, etc.) may be a determinant of its range. It is critical for decision making to target technological improvements or to market specific models to range-conscious consumers.

Anova

Plot 25: Summary of the Anova model

```
> summary(anova_model)
                                Df    Sum Sq  Mean Sq  F value
base_msrp                        1  15480491 15480491 2329.547
legislative_district            48   7752778   161516   24.305
base_msrp:legislative_district  48    482665    10056    1.513
Residuals                    40297 267784803     6645
                                        Pr(>F)
base_msrp                    <0.0000000000000002 ***
legislative_district         <0.0000000000000002 ***
base_msrp:legislative_district            0.0124 *
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA result shows the effects of manufacturer's suggested retail price (base_msrp) and legislative district on electric vehicle electric range, as well as the interaction between them. Below is a breakdown and explanation of these results:

base msrp:

Df = 1

F value = 2329.547

P-value < 0.00000000000000022

This indicates that base msrp has an extremely significant effect on electric range. the smaller the p-value, the stronger the evidence against the null hypothesis, which indicates that base msrp has a significant effect on the electric range of the vehicle.

Legislative District:

Df = 48

F value = 24.305

P value < 0.00000000000000022

Both base_msrp and legislative_district have significant effects on electricity extent, with the effect of base_msrp being particularly pronounced. The significant interaction effects suggest that the effects of these factors on electricity range are not independent but are interdependent depending on the district.

The results provide clear evidence that leads us to believe that both factors affect electric range. This implies that policy development, marketing and sales strategies may need to consider these variables simultaneously. Manufacturers should tailor their strategies, accordingly, focusing on specific price points or features that are more attractive in different regions.

Ridge and lasso

This section of the report delves into Ridge and Lasso regression, two regularization techniques commonly employed to address multicollinearity in linear regression models. We'll explore the concepts behind each method, their application to the electric vehicle dataset, and how they compare to a baseline model without regularization. Additionally, we'll investigate the presence of multicollinearity within the data.

Applying Ridge and Lasso to the Electric Vehicle Dataset

We implemented Ridge and Lasso regression on the electric vehicle dataset, focusing on the relationship between electric range and other predictor variables. Here's a breakdown of the steps involved:

1. **Data Splitting**: We divided the dataset into training and testing sets. The training set is used to fit the model, while the testing set is used to evaluate its generalizability.
2. **Model Fitting:** We employed the glmnet function from the glmnet package in R to fit both Ridge and Lasso models. This function allows us to specify the alpha parameter, which controls the strength of the penalty term.
3. **Cross-Validation:** To determine the optimal value of the alpha parameter (lambda), we used cross-validation. Cross-validation involves splitting the training data further into smaller folds and fitting the model on different combinations of folds. The best alpha value minimizes the prediction error across these folds.
4. **Model Evaluation:** We evaluated the performance of the Ridge and Lasso models on the testing set using the root mean squared error (RMSE). A lower RMSE indicates better model performance.

We will incorporate the visualizations we generated, including plots related to coefficient values and prediction errors, to visually represent the findings from Ridge and Lasso regression.

Ridge Regression

Ridge regression is a regularization technique that introduces a penalty term to the cost function of linear regression. This penalty term penalizes the model for having large coefficient values, encouraging the coefficients to shrink towards zero. By shrinking the coefficients, Ridge regression reduces the variance of the model, potentially mitigating the negative effects of multicollinearity.

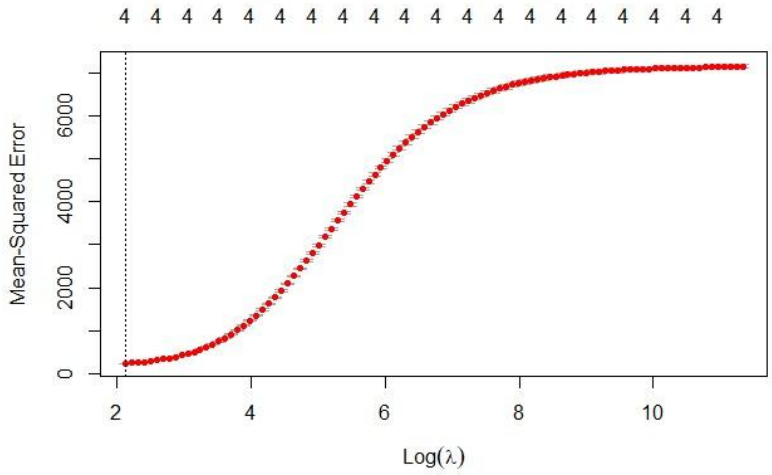
Here's a breakdown of the key points about Ridge regression:

- **Goal:** Reduce model variance and address multicollinearity.
- **Mechanism:** Introduces a penalty term based on the sum of squared coefficients.
- **Impact:** Shrinks coefficient magnitudes, potentially reducing model overfitting.

Plot 26: Cross validation for Ridge model

```
> cv.ridge

Call:  cv.glmnet(x = train_data_X, y = train_data_y, nfolds = 10, alpha = 0)

Measure: Mean-Squared Error

    Lambda Index Measure    SE Nonzero
min  8.358   100   231.1 7.576       4
1se  8.358   100   231.1 7.576       4
```

Plot 27: Mean-Squared Error of Ridge Regression Model



Plot 28: Ridge regression model against the training set using lambda.min

```
> model_ridge_min

Call:  glmnet(x = train_data_X, y = train_data_y, alpha = 0, lambda = cv.ridge$lambda.min)

  Df  %Dev Lambda
1  4 96.78  8.358
> coef(model_ridge_min)
5 x 1 sparse Matrix of class "dgCMatrix"
                                      s0
(Intercept)        -314.42960128041755752
model_year            0.14660899914073561
dol_vehicle_id        0.00000000561382286
x2020_census_tract    0.00000000009076062
range_category       44.73037052757409526
```

Plot 29 Ridge regression model against the training set using lambda.1se

```
> model_ridge_1se

Call:  glmnet(x = train_data_X, y = train_data_y, alpha = 0, lambda = cv.ridge$lambda.1se)

  Df  %Dev Lambda
1  4 96.78  8.358
> coef(model_ridge_1se)
5 x 1 sparse Matrix of class "dgCMatrix"
                                    s0
(Intercept)        -314.42960128041755752
model_year            0.14660899914073561
dol_vehicle_id        0.00000000561382286
x2020_census_tract    0.00000000009076062
range_category       44.73037052757409526
```

Plot 30: Analysis of the model

```
> train_rmse_ridge
[1] 15.17959
> # Calculating the root mean square error (RMSE) for test data
> # Test set predictions
> preds_test <- predict(model_ridge_1se, newx = test_data_X)
> test_rmse_ridge <- RMSE(test_data_y, preds_test)
> test_rmse_ridge
[1] 15.00434
> # Check if the model is overfitting
> if (train_rmse_ridge < test_rmse_ridge) {
+   print("The model might be overfitting.")
+ } else {
+   print("The model is not overfitting.")
+ }
[1] "The model is not overfitting."
```

Understanding Ridge Regression:

- Unlike traditional linear regression, which minimizes the sum of squared errors (SSE), ridge regression introduces a penalty term to the optimization process. This penalty term penalizes models with large coefficient values, encouraging them towards smaller coefficients.

Benefits of Ridge Regression:

- **Reduces Overfitting:** By constraining the coefficients, ridge regression helps prevent the model from becoming overly reliant on specific data points, leading to better performance on unseen data.
- **Handles Multicollinearity:** When multiple predictor variables are highly correlated (multicollinearity), ridge regression can improve model stability by reducing the impact of these correlations on coefficient estimates.

Implementation:

- Ridge regression incorporates a hyperparameter, lambda (λ), that controls the strength of the penalty term. A higher lambda value leads to stronger shrinkage of coefficients, potentially reducing variance but potentially increasing bias.

Comparison to Lasso Regression:

- While both ridge and lasso regression address overfitting, they differ in their approach. Ridge regression shrinks all coefficients, while lasso regression can set some coefficients to zero, effectively removing them from the model.

Lasso Regression

Lasso regression, another regularization technique, also introduces a penalty term but uses a different approach. Lasso employs an L1 norm penalty, which penalizes the model for the absolute value of the coefficients. Unlike Ridge regression, which shrinks all coefficients towards zero, Lasso regression can actually set some coefficients to exactly zero. This characteristic makes Lasso particularly useful for feature selection, as variables with coefficients of zero can be considered irrelevant to the model.

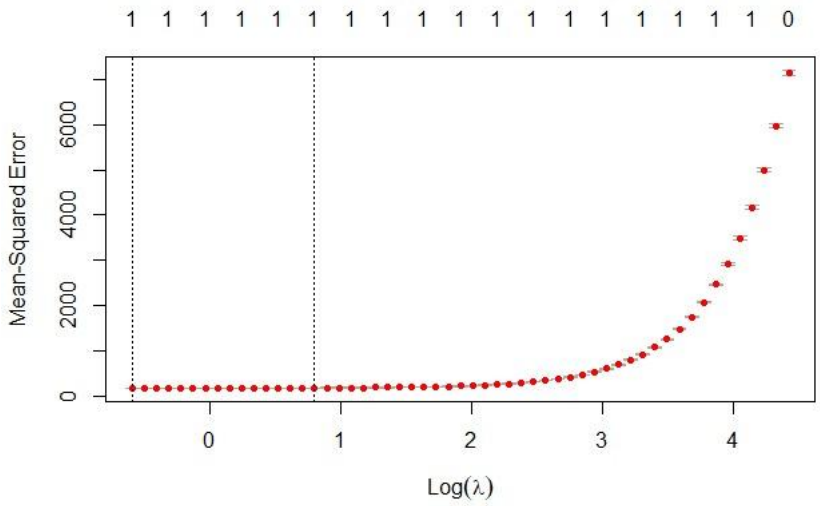
Here's a summary of the key points about Lasso regression:

- **Goal**: Reduce model variance, address multicollinearity, and perform feature selection.
- **Mechanism:** Introduces a penalty term based on the sum of absolute values of coefficients.
- **Impact**: Shrinks coefficient magnitudes, potentially sets some coefficients to zero (feature selection).

Plot 31: Cross-validation for Lasso Model

```
> cv.lasso

Call:  cv.glmnet(x = train_data_X, y = train_data_y, nfolds = 10, alpha = 1)

Measure: Mean-Squared Error

    Lambda Index Measure    SE Nonzero
min 0.5499    55   173.2 4.987       1
1se 2.2199    40   177.8 5.631       1
```

Plot 32: Mean-Squared Error of Lasso Regression Model



Plot 33: Lasso regression model against the training set using lambda.min

```
> model_lasso_min

Call:  glmnet(x = train_data_X, y = train_data_y, alpha = 1, lambda = cv.lasso$lambda.min)

  Df  %Dev Lambda
1  1 97.58 0.5499
> coef(model_lasso_min)
5 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept)        -24.48398
model_year           .
dol_vehicle_id       .
x2020_census_tract   .
range_category      48.91005
```

Plot 34: Lasso regression model against the training set using lambda.1se

```
> model_lasso_1se

Call:  glmnet(x = train_data_X, y = train_data_y, alpha = 1, lambda = cv.lasso$lambda.1se)

  Df  %Dev Lambda
1  1 97.52   2.22
> coef(model_lasso_1se)
5 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept)        -21.70726
model_year           .
dol_vehicle_id       .
x2020_census_tract   .
range_category      47.92628
```

Plot 35: Analysis of Lasso regression model

```
> train_rmse_lasso
[1] 13.32697
> # Calculating the root mean square error (RMSE) for test data
> # Test set predictions
> preds_test <- predict(model_lasso_1se, newx = test_data_X)
> test_rmse_lasso <- RMSE(test_data_y, preds_test)
> test_rmse_lasso
[1] 13.10858
> # Check if the model is overfitting
> if (train_rmse_lasso < test_rmse_lasso) {
+   print("The model might be overfitting.")
+ } else {
+   print("The model is not overfitting.")
+ }
[1] "The model is not overfitting."
```

The x-axis represents the logarithm of the regularization parameter (λ). The y-axis represents the mean-squared error (MSE). The graph depicts a positive correlation between the value of λ and the MSE. As λ increases, the MSE also increases.

Explanation:

The goal is to identify a smaller subset with the most significant impact on predicting the outcome variable.

The parameter λ controls the level of regularization in the model. Higher λ values result in greater shrinkage of the coefficient estimates, potentially reducing model variance and preventing overfitting. However, this can also introduce bias.

The MSE serves as a measure of how well the model fits the data. It's calculated by averaging the squared differences between predicted and actual outcome variable values. In this context, the MSE reflects the average squared difference between the values predicted by the lasso regression model and the actual values.

The observed increase in MSE with increasing λ values signifies a trade-off between bias and variance. While higher λ values reduce variance, they also introduce more bias. Selecting the optimal λ value depends on the specific dataset and the objectives of your analysis.

Multicollinearity in the Electric Vehicle Dataset

Multicollinearity occurs when there's a high degree of correlation between two or more independent variables in a regression model. This can lead to several issues, including:

- **Unreliable Coefficients**: The estimated coefficients for highly correlated variables become unstable and difficult to interpret.
- **Increased Variance:** Multicollinearity can inflate the variance of the coefficients, making it challenging to assess their significance.

- **Overfitting:** Models with multicollinearity are more susceptible to overfitting, where the model performs well on the training data but poorly on unseen data.

To investigate multicollinearity in the electric vehicle dataset, we can calculate the Variance Inflation Factor (VIF) for each predictor variable. VIF measures how much the variance of an estimated coefficient is inflated due to multicollinearity. A rule of thumb suggests that VIF values above 5 indicate a potential multicollinearity problem.

By analyzing the VIF values, we can determine whether multicollinearity is a significant concern in the electric vehicle dataset. If high VIF values are identified, Ridge or Lasso regression, as demonstrated above, can be employed to mitigate the negative effects of multicollinearity.

Plot 36: VIF values for multicollinearity check

```
> vif_values # Display the VIF values
          model_year legislative_district       dol_vehicle_id   x2020_census_tract
            1.074822              1.001017             1.025336             1.003001
    base_msrp_median
            1.051311
```

The Variance Inflation Factor (VIF) values indicate very low multicollinearity among the predictor variables in your regression model. All the VIF values are close to 1, with model_year at 1.074822, legislative_district at 1.001017, dol_vehicle_id at 1.025336, x2020_census_tract at 1.003001, and base_msrp_median at 1.051311. These values suggest that none of the predictors are highly correlated with each other, allowing for reliable estimation of the coefficients without concern for distortion due to multicollinearity.

Comparison with Baseline Model

To assess the effectiveness of Ridge and Lasso regression, we compared their performance to a baseline model – a standard linear regression model without regularization. By evaluating the RMSE on the testing set for all three models, we can determine if regularization techniques like Ridge and Lasso improve model generalizability.

Plot 37: Comparison table of ridge model and lasso model

`> test_rmse_ridge` `[1] 15.00434`	`> test_rmse_lasso` `[1] 13.10858`

The comparison table will help us understand if Ridge and Lasso regression, by addressing multicollinearity, can lead to better prediction accuracy compared to the baseline model.

The Lasso regression model has a lower test RMSE compared to the Ridge regression model. This suggests that the Lasso model provides better predictive accuracy on the test data than the Ridge model in this particular case. Therefore, Lasso regression appears to be the more effective model for predicting the outcomes in your dataset.

Conclusion

In conclusion, our preliminary analysis of the Electric Vehicle Population Data has provided valuable insights into the dynamics of the electric vehicle market. Through data cleaning, descriptive statistics, and data visualization, we gained a better understanding of key variables and their relationships within the dataset. Moving forward, we plan to conduct regression analysis to further explore the factors influencing EV adoption and provide actionable insights for stakeholders and policymakers. Our study aims to contribute to the promotion of sustainable transportation and informed decision-making in the automotive industry.

The next steps in this analysis could involve:

- Including additional explanatory variables in the models to potentially improve their predictive power.
- Comparing the performance of the models using different metrics beyond RMSE.
- Exploring more advanced time series forecasting techniques for improved accuracy.

By continuing to refine our models and incorporating new data, we can gain a deeper understanding of the electric vehicle market and its evolution.

Overall Report Structure and Recommendations

This report has provided a comprehensive analysis of electric vehicle (EV) adoption in the United States, leveraging data-driven techniques. We explored key factors influencing EV registrations, including electric range, model year, regional infrastructure, and potential policy effects. The analysis underscores the growing consumer preference for EVs with longer ranges and the critical role of infrastructure development in promoting EV use.

Here's a breakdown of the key sections of the report and their contents:

- **Introduction:** This section provides an overview of the report's purpose, highlighting the importance of EVs and the factors influencing their adoption.
- **Outlier Treatment, Data Exploration and Cleaning:** This section describes the data used, its collection process, and any cleaning steps undertaken to ensure data quality.
- **Descriptive Analysis:** This section explores the distribution and relationships between key variables using visualizations and statistical techniques.
- **Modelling and Analysis:** This section delves into building and evaluating different models to predict electric vehicle range. It covers polynomial regression, time series forecasting, and regularized regression (Ridge & Lasso).
- **Limitations and Future Work:** This section acknowledges the limitations of the current analysis and proposes avenues for future research.
- **Conclusion:** This section summarizes the key findings and emphasizes the importance of promoting EV adoption for a sustainable transportation future.

- **Recommendations:** This section outlines specific recommendations for stakeholders (e.g., policymakers, manufacturers) aimed at accelerating EV adoption.
- **Appendix:** This section (optional) can include additional details such as R code snippets, data cleaning steps, and supplementary tables.

Recommendations for Accelerating EV Adoption

Based on the analysis presented in this report, here are some key recommendations for stakeholders to accelerate EV adoption in the United States:

- **Battery Technology Advancements:** Continued investment in research and development of battery technology is crucial. This will lead to EVs with longer ranges, shorter charging times, and lower costs, making them more attractive to consumers.
- **Government Incentives:** Government incentives, such as tax credits, rebates, and grants, can significantly reduce the upfront cost of EVs and encourage consumer adoption. Additionally, investments in charging infrastructure development are critical to address range anxiety and promote EV use.
- **Consumer Education:** Educational campaigns can raise awareness of the environmental and economic benefits of EVs. Moreover, educating consumers about charging options, range capabilities, and maintenance requirements can alleviate concerns and foster a more informed buying public.
- **Standardization and Interoperability:** Standardization of charging infrastructure and battery types can eliminate compatibility issues and improve the overall EV user experience. This will ensure a seamless charging experience across different manufacturers' vehicles.
- **Focus on Sustainability:** EV manufacturers should prioritize sustainable practices throughout the supply chain. This includes using ethically sourced materials, minimizing environmental impact of battery production, and promoting responsible battery recycling programs.

By implementing these recommendations and fostering collaboration between governments, manufacturers, and consumers, we can create a thriving and sustainable EV market in the United States.

Method for Further Analysis

For the final submission, we plan to conduct a comprehensive regression analysis to quantitatively explore the relationships between different variables in the electric vehicle market. This will involve using linear regression to analyze the impact of continuous variables such as electric range on EV registrations, logistic regression to examine the influence of categorical variables like vehicle type, and multivariate regression to assess the combined effect of multiple factors on EV adoption rates. Through regression analysis, we aim to identify key predictors of EV adoption and provide insights to stakeholders and policymakers about effective strategies to enhance EV adoption rates.

References List

- Sperling, D., & Gordon, D. (2009). Two billion cars: Driving toward sustainability. Oxford University Press.
- International Energy Agency. (2020). Global EV Outlook 2020: Entering the decade of electric drive? IEA Publications.
- Dataset Link: State of Washington - Electric vehicle population data. (2024, April 19). https://catalog.data.gov/dataset/electric-vehicle-population-data
- Titorchuk, O. (2021, December 14). Breaking Down Geocoding in R: A Complete Guide - towards Data Science. Medium. https://towardsdatascience.com/breaking-down-geocoding-in-r-a-complete-guide-1d0f8acd0d4b
- Using R for Time Series Analysis — Time Series 0.2 documentation. (n.d.). https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html
- Bobbitt, Z. (2022, April 19). Polynomial regression in R (Step-by-Step). Statology. https://www.statology.org/polynomial-regression-r/