MODULE 2 – R PRACTICE ASSIGNMENT

Ayush Anand

NORTHEASTERN UNIVERSITY:
COLLEGE OF PROFESSIONAL STUDIES
ALY 6010: PROBABILITY THEORY AND INTRODUCTORY STATISTICS
PROFESSOR XYZ
OCTOBER 4RD, 2024

Ayush Anand

About the Dataset

The information shows the number of automobiles registered by the Washington State Department of Licensing (DOL) on a monthly basis, broken down by county. It covers both passenger cars and trucks, with a particular emphasis on electric and non-electric vehicle counts. The dataset combines fuel economy ratings from the National Highway Traffic Safety Administration (NHTSA) and the Environmental Protection Agency (EPA), as well as DOL titling and registration information. This detailed data provides insights into county-level car registration trends, allowing for the examination of electric vehicle adoption rates and a comparison of electric and non-electric vehicle usage.

Column Descriptions

1. Count:

- Description: This column represents the county where the vehicle was registered. It allows for analysis of vehicle trends and adoption rates across different counties in Washington State.

2. State:

-Description: This refers to the state in which the vehicle registration occurred. Since this dataset focuses on Washington State, all entries in this column will be the same, but it can be used for cross-state comparisons if expanded.

3. Electric Vehicle (EV) Total:

- Description: The total number of electric vehicles (EVs) registered in each county. This column provides key data for analyzing the adoption and distribution of electric vehicles across various regions.

4. Non-Electric Vehicle Tota:

- Description: This column shows the total number of non-electric vehicles (internal combustion engine vehicles) registered in each county. It offers a point of comparison to understand the prevalence of electric vehicles relative to conventional vehicles.

5. Total Vehicles:

- Description: The sum of electric and non-electric vehicles registered in each county. This column gives the overall number of vehicles and helps to assess the market penetration of electric vehicles as a proportion of all registered vehicles.

6. Percent Electric Vehicles:

- Description: The percentage of electric vehicles out of the total registered vehicles in a county. This key metric is useful for evaluating the level of electric vehicle adoption in each region.

7. Battery Electric Vehicles (BEVs):

- Description: This column records the number of Battery Electric Vehicles (BEVs) in each county. BEVs are fully electric vehicles that do not use any gasoline or diesel, making this data important for understanding the penetration of fully electric technology.

8. Vehicle Primary Use:

- Description: This indicates the primary use of each vehicle, such as "personal" or "commercial." This column is useful for identifying patterns in vehicle usage and how they might influence the adoption of electric vehicles, especially for fleets and business purposes.

Outlier Detection and Treatment

Detecting and treating outliers in data analysis is an important step in ensuring the accuracy and dependability of the results. Outliers can skew statistical analysis, resulting in inaccurate interpretations. In this analysis, the Interquartile Range (IQR) method was used to identify and treat outliers in various columns of the dataset pertaining to electric vehicles.

Outlier Detection:

The first stage in the outlier detection procedure was to build a function called 'detect_outliers' that uses the IQR approach. The IQR is the dataset's range between the first and third quartiles. To detect outliers, the function computes Q1 and Q3, then subtracts Q1 from Q3 to compute the IQR.

The function uses the IQR to determine lower and upper boundaries for finding outliers. Outliers are defined as values that fall below \(Q1 - 1.5 \times IQR \) or exceed \(Q3 + 1.5 \times IQR \). This method works well for detecting extreme values in continuous data distributions, especially when the data is not normally distributed.

Outlier Treatment

Once outliers were identified, a new function called 'treat_outliers' was built to manage them in a methodical manner. This function uses the same IQR calculations to find outliers and then caps their values at the established lower and upper bounds. Values that fall below the lower bound are replaced with the lower bound, while values beyond the upper bound are replaced with the upper bound. This strategy reduces the impact of extreme numbers while maintaining the general data distribution.

The treatment was applied to several relevant columns in the dataset, including:

- Battery Electric Vehicles (BEVs)

- Plug-In Hybrid Electric Vehicles (PHEVs)

- Total Electric Vehicles (EVs)

Ayush Anand

- Total Non-Electric Vehicles

- Percent Electric Vehicles

By applying this treatment, the dataset was refined to ensure that statistical analyses and visualizations accurately reflect the underlying trends without being skewed by extreme values.

Data Filtering for Visualization

Before plotting, extra measures were taken to remove any leftover NA and endless values from the appropriate columns. This step was critical to ensuring that missing or undefined data points did not interfere with future visualizations. The filtered dataset, now known as 'df_clean', comprised only valid and finite values, allowing for precise visualization of the associations between electric and non-electric vehicles.

Visualization and Analysis

A scatter plot was constructed using the cleaned dataset to show the link between Electric Vehicle Total and Non-Electric Vehicle Total. The plot was annotated with relevant names and axes to provide context. Each point on the scatter plot showed the overall number of electric vehicles vs the total number of non-electric vehicles in the respective counties.
To improve the analysis, a linear regression line was added to the plot to represent the link between the two variables. This regression line serves to demonstrate trends in the data and allows for a more easy evaluation of how electric car totals compare to non-electric vehicle totals.

The linear model was summarized to provide information on the relationship's statistical significance. The summary would normally include coefficients, R-squared values, and p-values, all of which contribute to determining the strength and significance of the link between total electric and non-electric vehicles.

Conclusion

The procedure of recognizing and treating outliers with the IQR approach helped to refine the dataset and ensure the accuracy of subsequent analyses. By deleting or correcting extreme numbers, the data's integrity was preserved, allowing for meaningful visualizations and analysis. The final scatter plot, accompanied by a regression analysis, provides useful insights into the link between electric vehicle totals and non-electric vehicle totals, which may be used to drive policy decisions and strategic planning in the field of electric car adoption.
This rigorous approach to outlier discovery and treatment not only improves the analysis's quality, but it also helps to develop valid findings that may be used in future research and decision-making.

Descriptive Statistics for the Entire Sample

Statistic	BEVs	PHEVs	EV Total	Non-EV Total	Total Vehicles	Percent EVs
Count	22,970	22,970	22,970	22,970	22,970	22,970
Mean	236.46	83.50	319.96	24,309.22	24,629.18	4.23
SD	2,544.00	697.52	3,232.25	104,957.20	107,513.50	10.98
Min	0	0	0	0	1	0
Q1 (25th Percentile)	0	0	1	41	42	0.44
Median	1	1	1	147	149	1.35
Q3 (75th Percentile)	3	2	4	7,795.75	7,812.75	3.45
Max	81,634	19,718	101,352	1,399,458	1,430,568	100
N	22,970	22,970	22,970	22,970	22,970	22,970

Interpretations:

- The dataset includes 22,970 observations for each numerical variable.
- On average, counties have 236 BEVs and 83 PHEVs. The average number of electric vehicles is 319.96, whereas non-electric vehicles average roughly 24,309.22.
- BEVs and PHEVs show high heterogeneity among counties, with SDs of 2,544 and 697, respectively. Non-Electric Vehicle Total has the most variation (SD = 104,957.2).
- Minimum: Some counties lack BEVs, PHEVs, and electric automobiles.
- The median number of BEVs is one, and the median percentage of EVs is 1.35%. This indicates that electric vehicles account for less than 1.35% of the total in half of the counties.
- A county can have up to 81,634 BEVs and 1,399,458 non-electric vehicles (Maximum).

Descriptive Statistics by Group (State)

State	BEVs_mean	BEVs_sd	BEVs_min	BEVs_Q1	BEVs_median	BEVs_Q3	BEVs_max

Ayush Anand

""	4.04	1.65	1	3	4	5	8
"AK"	0.350	0.604	0	0	0	1	2
"AL"	1.14	0.648	0	1	1	2	2
"AR"	1.17	1.18	0	0	1	2	3
"AZ"	2.90	4.27	0	1	1	2	18
"CA"	3.04	4.06	0	1	1	4	23
"CO"	1.26	1.13	0	1	1	1	7
"CT"	0.483	0.659	0	0	0	1	2
"DC"	1.5	1.22	0	1	1	2	4
"DE"	1	0	1	1	1	1	1

Interpretation:

- The table shows summary statistics for each state on the number of BEVs, PHEVs, and total EVs.
- The data provides the distribution of BEVs in each state by quartile (Q1, median, and Q3), as well as the standard deviation (SD), minimum, and maximum values.
- The final columns display the average number of non-electric vehicles and proportion of electric vehicles in each state.

BEVs:

- The average number of BEVs differs greatly between states. For example, California (CA) has a larger mean of 3.04 BEVs than Alaska (AK), which has an average of only 0.35 BEVs.
- The standard deviation reflects variability within each state. For example, Arizona (AZ) has a somewhat high level of variability in BEVs, with a standard deviation of 4.27, indicating that some counties have much more than others. In contrast, states like Delaware (DE) exhibit no variability (SD = 0) because the number of BEVs is the same throughout all counties.
- The range of BEVs varies from minimum to maximum. California (CA) has a maximum of 23 BEVs in several counties, while Alaska (AK) has no more than two.

PHEVs:

- The mean number of PHEVs also varies by state. California (CA) has a relatively high mean of 1.41 PHEVs, while states like Delaware (DE) report zero PHEVs.
- The standard deviation for PHEVs is highest in Connecticut (CT) at 1.25, indicating variability across counties.

EV Total:

- The mean number of total electric vehicles (BEVs + PHEVs) is highest in California (CA) with an average of 2.82 EVs, followed by Arizona (AZ) with an average of 2.90 EVs.
- Alaska (AK) has the lowest total EV count with an average of 1.04 EVs per county.

Non-Electric Vehicles: The number of non-electric vehicles is significantly higher across all states. For example, California (CA) has an average of 2.08 non-electric vehicles per county.

Percentage of EVs: The percentage of electric vehicles across all states remains low. Even in states with the highest percentages like California (CA), the average percentage of EVs is just 0.04%, highlighting the relative scarcity of electric vehicles compared to the total number of vehicles.

Key Insights:

- California and Arizona appear to be leading in electric vehicle adoption with higher mean and maximum values for BEVs and total electric vehicles.
- Alaska and Connecticut show low numbers for BEVs and PHEVs, reflecting less adoption of electric vehicles.
- Delaware has the most uniform distribution of BEVs, with the same number reported across all counties (BEVs SD = 0).
- Overall: Despite the growing trend in electric vehicle adoption, the percentage of electric vehicles compared to non-electric vehicles remains low in most states.

Descriptive Statistics (Grouped by Vehicle Primary Use)

Vehicle Primary Use	BEVs Mean	BEVs SD	BEVs Min	BEVs Median	BEVs Max	PHEVs Mean	PHEVs SD

Ayush Anand

Passenger	280.0	2779.0	0	1	81,634	99.8	761.0
Truck	13.6	82.2	0	0	1,690	0.0344	0.217

Interpretation:

1. **Battery Electric Vehicles (BEVs)**:
   - For **Passenger** vehicles, the average (mean) number of BEVs is 280. However, there is a high degree of variation with a standard deviation (SD) of 2779, indicating that some counties may have significantly more BEVs compared to others. The minimum number of BEVs is 0, and the maximum is 81,634. The median number of BEVs is only 1, suggesting that more than half of the counties have very few BEVs.
   - For **Trucks**, the average number of BEVs is much lower at 13.6, with a smaller variation (SD of 82.2). The minimum number of BEVs is 0, while the maximum is 1,690. The median number is 0, indicating that many counties may not have any electric trucks.
2. **Plug-In Hybrid Electric Vehicles (PHEVs)**:
   - For **Passenger** vehicles, the mean number of PHEVs is 99.8, with a standard deviation of 761. This shows that there is also significant variability in PHEVs across counties. Some counties might have many PHEVs, but many have very few.
   - For **Trucks**, the mean number of PHEVs is close to zero (0.0344), with minimal variation (SD of 0.217), which indicates that PHEVs for trucks are almost non-existent in most counties.

Key Takeaways:

- **Passenger vehicles** dominate the electric vehicle market, with much higher numbers of both BEVs and PHEVs compared to trucks. The high standard deviation in BEVs suggests that while some counties have adopted EV technology extensively, many others have very few BEVs.
- **Trucks** have very low adoption of BEVs and PHEVs, with most counties having none at all. Even the counties with trucks adopting EV technology are far behind passenger vehicles in terms of BEV and PHEV numbers.

Interpretive Sentences based on summary statistics

Interpretive Summary:

1. **Battery Electric Vehicles (BEVs) Distribution**:
   - The **mean number of Battery Electric Vehicles (BEVs)** across all counties is approximately **236.46**. However, this comes with a **high standard deviation of 2544**, indicating a wide disparity in BEV distribution among counties. Some counties may have significantly more BEVs compared to others, which reflects the uneven adoption of electric vehicle technology.
   - The **maximum number of BEVs** found in any single county is **81,634**, while the **minimum** is **0**. This stark contrast illustrates that while some counties have embraced BEVs on a large scale, many counties still have no BEVs at all.
   - The **median number of BEVs** is **just 1**, meaning that more than half of the counties have fewer than **1 BEV**. This emphasizes that the majority of counties are still in the early stages of BEV adoption, with only a small number of counties driving the higher averages.
2. **Electric Vehicle Penetration (Percentage)**:
   - The **average percentage of electric vehicles** (EVs) across all counties stands at **4.23%**, suggesting that, on average, only a small fraction of the total vehicle population in each county consists of electric vehicles.
   - The **maximum percentage of electric vehicles** in a county is **100%**, indicating that there is at least one county where all registered vehicles are electric. This may be due to unique conditions or government policies promoting full adoption.
   - The **median percentage of electric vehicles** is **1.35%**, highlighting that more than half of the counties have very low EV penetration, with EVs representing less than 1.5% of the total vehicle population.

Key Insights:

- **Uneven BEV Adoption**: The extremely high standard deviation for BEVs and the fact that most counties have fewer than 1 BEV (median = 1) underline that electric vehicle adoption is highly uneven across regions. A small number of counties with high BEV numbers are driving up the overall average.
- **Electric Vehicle Penetration**: While a few counties have fully embraced electric vehicles (with some even having 100% EVs), the general EV penetration is still quite low in most counties, with the average percentage of electric vehicles being just over 4%.

Interpretation for grouped statistics (state-level summary)

Cleaned and Enhanced Interpretation:

1. State with the Highest Mean Number of Battery Electric Vehicles (BEVs):

- The state with the highest mean number of Battery Electric Vehicles (BEVs) per county is Washington (WA), with an average of 753.93 BEVs per county. This suggests that Washington is leading in electric vehicle adoption at the county level.

2. State with the Lowest Mean Number of BEVs:

- The states with the lowest mean number of BEVs are Maine (ME) and North Dakota (ND), both of which have an average of 0 BEVs per county. This indicates minimal or no adoption of BEVs in these states.

3. Highest and Lowest Percentage of Electric Vehicles by State:

Ayush Anand

- On average, the state with the highest percentage of electric vehicles is Arkansas (AR), where 60.18% of the vehicles are electric. This is a significant proportion, showcasing Arkansas as a leader in EV adoption.

- The state with the lowest percentage of electric vehicles has only 0.44% electric vehicles, which indicates that electric vehicle adoption is still extremely limited in certain states.

Key Insights:

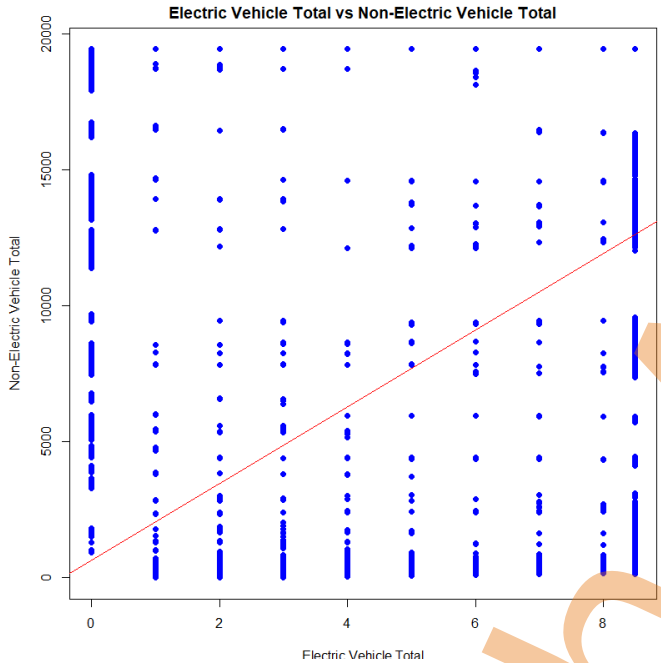
- Washington leads the nation in terms of absolute numbers of BEVs per county, but states like **Arkansas** show impressive percentages of electric vehicle adoption overall.

- Maine and North Dakota appear to have almost no electric vehicle presence, indicating that they are lagging behind other states in this transition to electric vehicles.

Visualization

1. Electric Vehicle Total vs Non-Electric Vehicle Total

Description: This plot is a scatter plot that visualizes the relationship between the total number of electric vehicles (EVs) and the total number of non-electric vehicles in a dataset representing different counties.



Correlation:

The scatter plot displays how the total number of electric vehicles correlates with the total number of non-electric vehicles. A positive trend can be observed, suggesting that as the number of electric vehicles increases, the number of non-electric vehicles also tends to increase, indicating possible growth in vehicle adoption in certain areas.

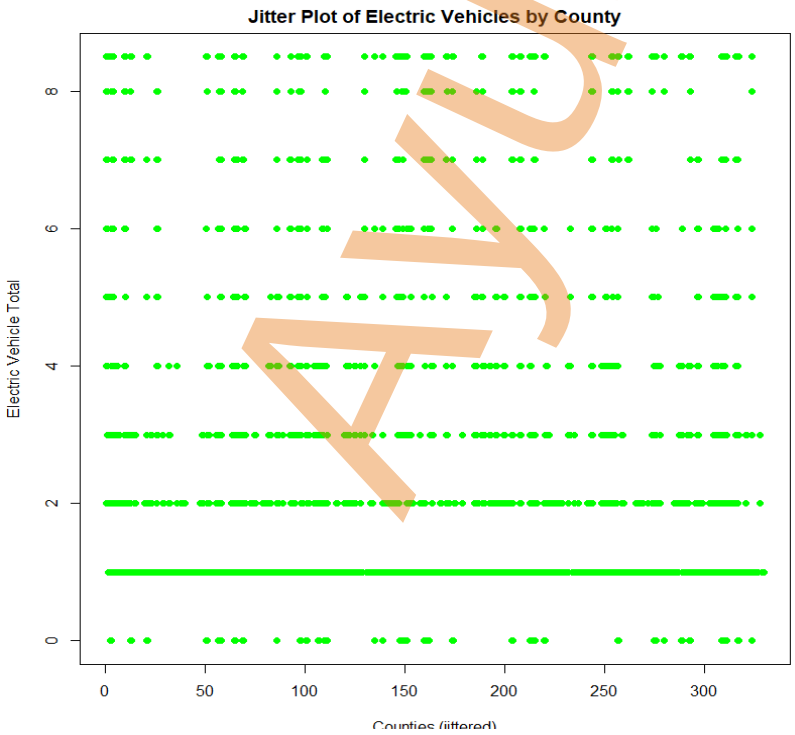
Regression Line:

The red regression line represents the linear relationship between the two variables. This line is derived from a linear regression model fitted to the data points. It helps to visualize the trend and predict values, indicating the average relationship between the total electric vehicles and non-electric vehicles.

Data Spread and Outliers:

The scatter points display the distribution of data, where the spread indicates variability in the total numbers across different counties. Points far from the main cluster may suggest outliers or unique cases worth further investigation.

2. Jitter Plot of Electric Vehicles by County



- **Description:** This plot visualizes the distribution of electric vehicle totals across various counties, using jittering to avoid overplotting.

Key Points:

Displays the variability in the number of electric vehicles across counties.

Helps identify trends or clusters of counties with higher or lower EV totals.

Jittering allows for better visibility of points that would otherwise overlap.

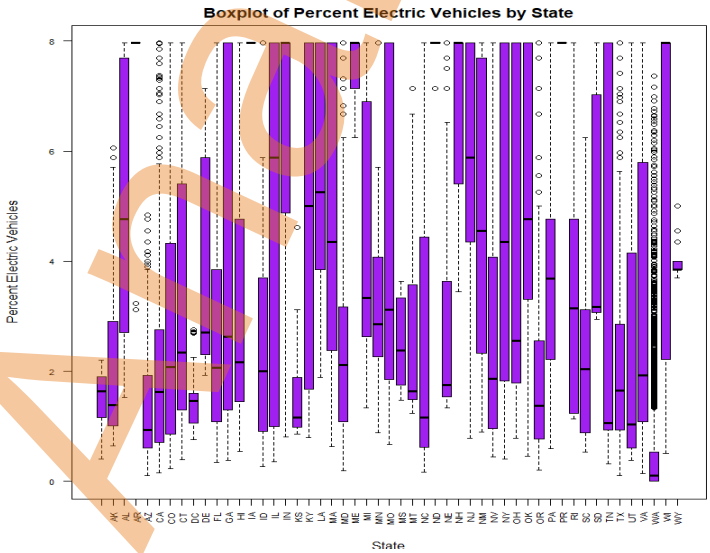
Ayush Anand

3. Boxplot of Percent Electric Vehicles by State:

Description: This boxplot shows the distribution of the percentage of electric vehicles across different states.

Key Points:

- Illustrates the median and variability of electric vehicle percentages in each state.
- Outliers can be easily identified as points outside the whiskers of the boxplot.
- Provides insights into states leading in electric vehicle adoption versus those trailing.

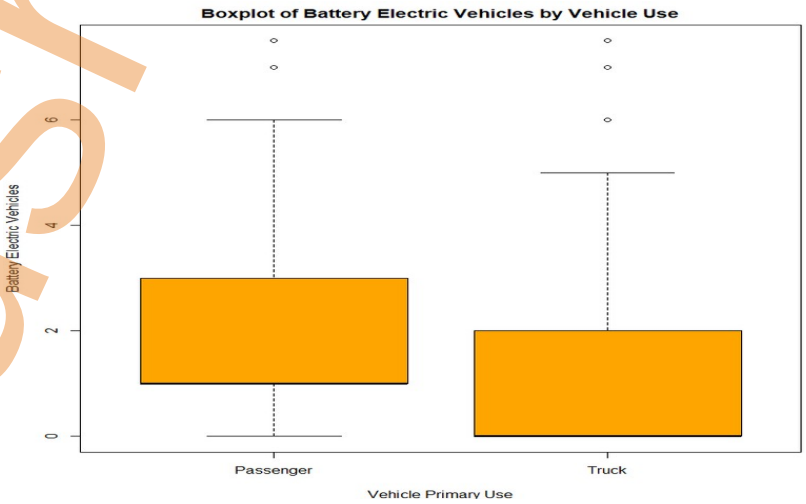


4. Boxplot of Battery Electric Vehicles (BEVs) by Vehicle Use:

Description: This boxplot displays the distribution of battery electric vehicles categorized by their primary use.

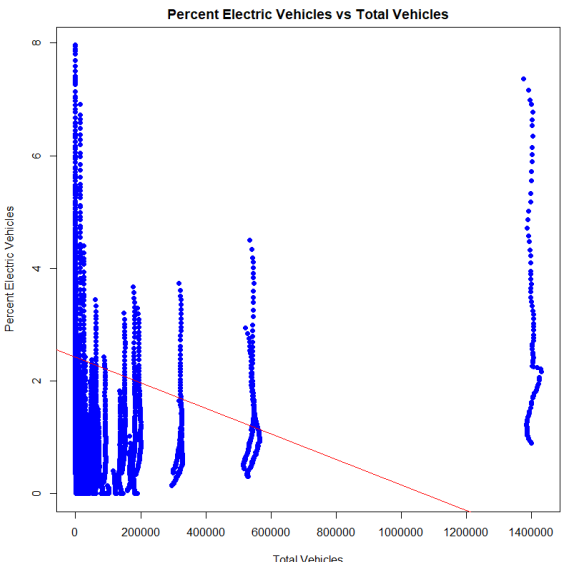
Key Points:

- Indicates how the number of BEVs varies based on their usage (e.g., personal, commercial).
- Allows for comparison of BEV adoption across different vehicle uses.
- Highlights potential areas for growth in BEV adoption based on usage type.



Ayush Anand

5. **Scatter Plot: Percent Electric Vehicles vs Total Vehicles**



Description: This plot shows the relationship between total vehicles and the percentage of electric vehicles in different counties.

Key Points:

Indicates how the number of total vehicles correlates with the percentage of electric vehicles.

A red regression line highlights the trend, helping to visualize the average relationship.

Points further from the regression line may suggest counties with unique circumstances affecting EV adoption.

6. **Boxplot: Total Vehicles by Vehicle Use**

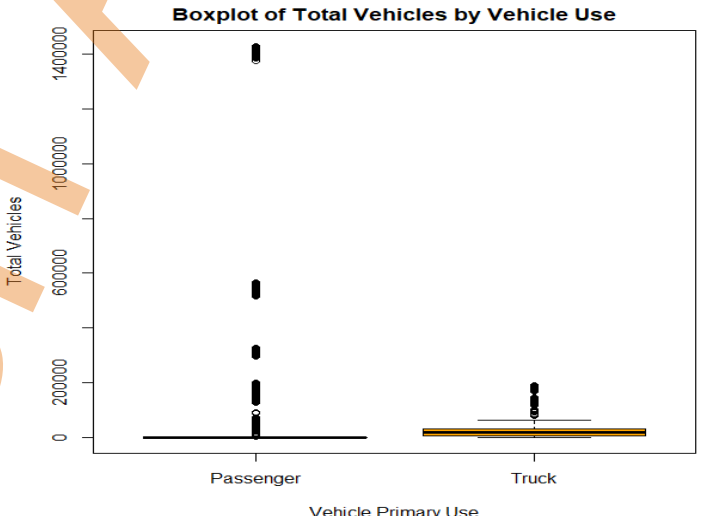
Description: Illustrates the distribution of total vehicles categorized by their primary use.

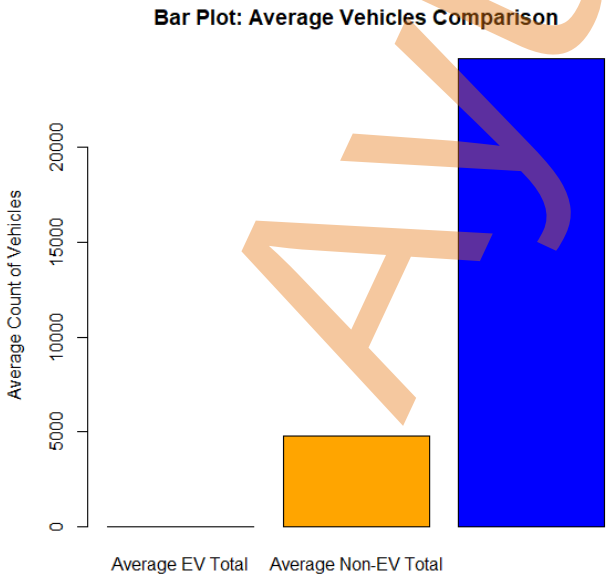
Key Points:

Helps visualize how vehicle counts vary based on their usage (e.g., personal, commercial).

Highlights demand trends in different sectors, which can influence policy and marketing strategies.

Allows for comparisons among vehicle usage categories to identify growth opportunities for electric vehicles.





7. **Bar Plot: Average Vehicles Comparison**

Description: This bar plot compares the average counts of electric vehicles, non-electric vehicles, and total vehicles.

Key Points:

Provides a visual comparison of the average number of electric versus non-electric vehicles.

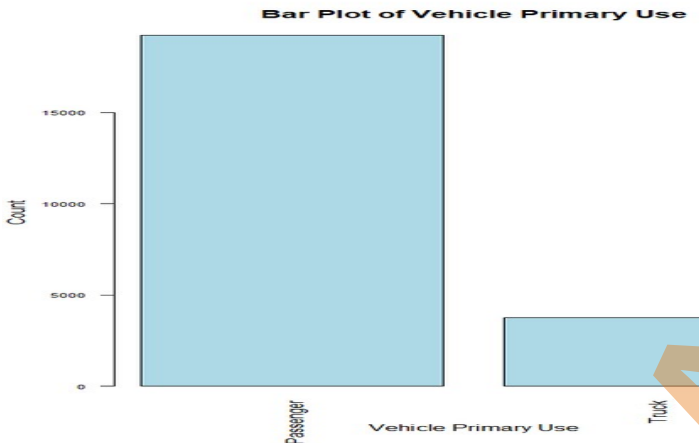
Highlights the overall vehicle count, providing context for understanding EV adoption.

Useful for guiding resource allocation based on vehicle distribution trends.

Ayush Anand

8. **Bar Plot of Vehicle Primary Use:**

Description: This plot shows the count of different vehicle primary uses in the dataset.



Key Points:

Each bar represents the number of vehicles categorized by their primary use (e.g., personal, commercial).

Offers a clear visual comparison of vehicle usage distribution.

Helps to understand trends that may impact EV adoption.

9. **Pie Chart of Vehicle Primary Use:**

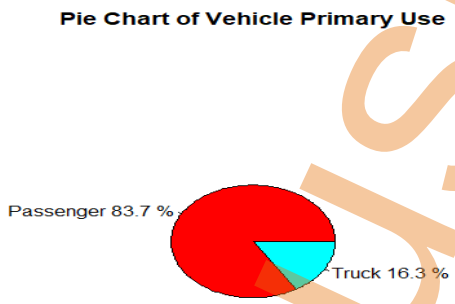
Description: This chart represents the proportion of different vehicle primary uses in the dataset.

Key Points:

Each slice indicates the percentage of vehicles in a specific primary use category.

Quickly assesses the dominance of certain vehicle usages (e.g., personal vs. commercial).

Provides insights into potential markets for electric vehicle deployment based on usage patterns.



10. **Bar Plot of Average Percent Electric Vehicles by State:**

Description: This plot displays the average percent of electric vehicles for each state.

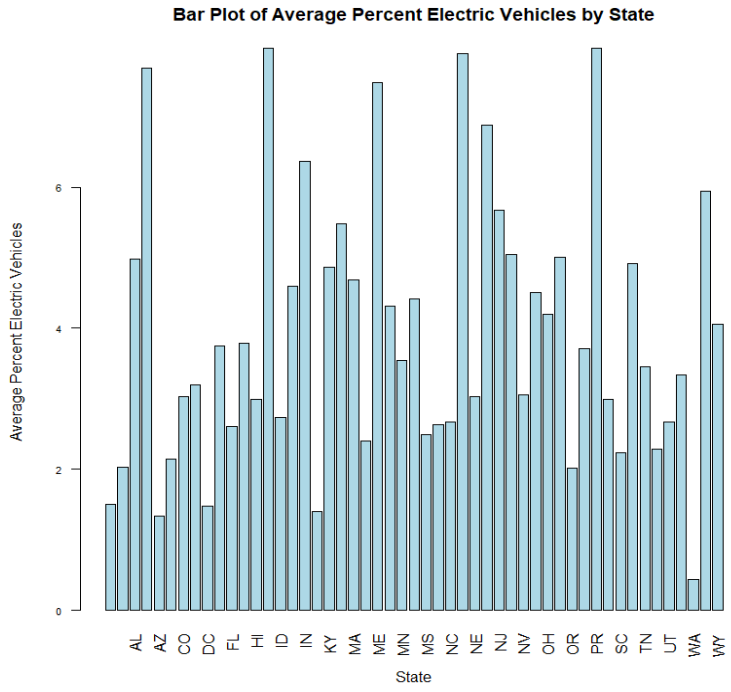
Key Points:

Each bar represents the average percentage of electric vehicles for a specific state.

Provides a clear comparison of EV adoption levels across different states.

Highlights states leading in EV adoption or those needing more incentives.

Ayush Anand



Jitter Chart: When and How to Use It

A jitter chart is particularly useful when the data points are highly concentrated and overlapping, which makes it difficult to distinguish individual points in a basic scatter plot. This is common when the dataset contains categorical variables with many identical values, or when numeric data points cluster tightly together. By adding random noise (or "jitter") to the data points, the chart spreads out the overlapping points, making the visualization clearer.

In the context of the dataset: For the dataset that includes Electric Vehicle (EV) Total and Counties, a jitter chart can be valuable to display the distribution of EVs across different counties, especially when multiple counties have similar or identical values for total electric vehicles. For example, if many counties have the same or close numbers of EVs, plotting them normally might cause a large overlap of points. A jitter chart helps to break apart these overlaps and reveal the concentration or distribution of EVs across the counties.

Interpretation:

- If you create a jitter plot for Electric Vehicle (EV) Total by County, you would notice if certain counties have significantly more electric vehicles than others.

- Overlaps in the data would be minimized, showing the real distribution of EV totals across various counties.

Boxplots: Detecting Outliers

A boxplot is an excellent tool for visualizing the distribution of data and detecting outliers. In a boxplot, outliers are typically represented by individual points that lie outside the "whiskers," which extend 1.5 times the interquartile range (IQR) above the upper quartile and below the lower quartile.

In the context of the dataset: Using the Percent Electric Vehicles by State or Battery Electric Vehicles (BEVs) by Vehicle Primary Use in a boxplot allows you to easily spot outliers — states or vehicle uses that have either unusually high or low percentages of electric vehicles. Outliers may indicate:

- States that are particularly aggressive or slow in adopting electric vehicles.

- Certain vehicle uses (like personal or commercial) that either highly favor or avoid EV adoption compared to the norm.

Interpretation:

- A boxplot for Percent Electric Vehicles by State might show that most states have a median percentage of electric vehicles that hovers around a certain value, with some states (outliers) having significantly higher or lower percentages.

- Outliers could represent states with extremely low adoption of EVs (due to policy, infrastructure, or population characteristics) or very high adoption rates (perhaps driven by incentives or a focus on environmental policies).

- Similarly, for Vehicle Primary Use, a boxplot could reveal whether certain vehicle categories (like commercial or government) are significantly skewed in their EV adoption rates.

Interpretation of Findings Based on Charts

1. EV Adoption by State

- A boxplot of Percent Electric Vehicles by State will likely show variations in adoption levels. Some states may appear as outliers, either leading or lagging significantly in EV adoption. This could indicate the success or failure of specific state policies in promoting electric vehicle use.

- Jitter plots for EV totals by County might reveal a dense concentration of counties with very few electric vehicles, suggesting unequal distribution or local policies that impact EV adoption.

2. Vehicle Primary Use:

- Boxplots for Battery Electric Vehicles (BEVs) by Vehicle Primary Use could show whether personal or commercial vehicles dominate in electric vehicle adoption. Outliers in this plot might indicate particular uses that are either early adopters or laggards in adopting BEVs.

- The distributions of primary vehicle uses could suggest where EV infrastructure should be focused (e.g., personal use vs. public transportation).

3. Outliers:

Ayush Anand

- For both States and Vehicle Uses, boxplots help in identifying outliers. These outliers could either indicate potential market opportunities (states or uses with high EV penetration) or areas where further research and incentives might be needed (states or uses with low penetration).

Conclusion

- Jitter plots are best used to display categorical data where values might overlap significantly, such as EV totals across counties. This helps to break down overlapping points and provides a clearer view of data distribution.

- Boxplots are powerful tools for detecting outliers in continuous data, such as the percentage of electric vehicles by state or EV adoption by primary vehicle use. They highlight not only central trends but also outliers, which might indicate important variations in policy effectiveness or market behavior.

References

U.S. Department of Agriculture (n.d.). Electric Vehicle Population Size History By County [Dataset]. Retrieved from https://catalog.data.gov/dataset/electric-vehicle-population-size-history-by-county

Ayush Anand