Module 3 R Practice

Ayush Anand

NORTHEASTERN UNIVERSITY:
COLLEGE OF PROFESSIONAL STUDIES
ALY 6010: PROBABILITY THEORY AND INTRODUCTORY STATISTICS
PROFESSOR XYZ
OCTOBER 11TH, 2024

About the dataset

Electric Vehicle Population Data Summary

- Total Records: **205,439**

- Total Columns: **17**

This dataset, sourced from the Washington State Department of Licensing (DOL), includes Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) registered in Washington. It provides information on the vehicles' electric range, manufacturer details, and geographic location.

Key Columns:

- VIN (1-10): First 10 characters of the Vehicle Identification Number.

- County, City, State, Postal Code: Location details.

- Model Year, Make, Model: Vehicle manufacturing details.

- Electric Vehicle Type: BEV or PHEV.

- Electric Range: Distance the vehicle can travel on electric power.

- Base MSRP: Manufacturer’s Suggested Retail Price.

- Legislative District: Legislative area of the vehicle owner.

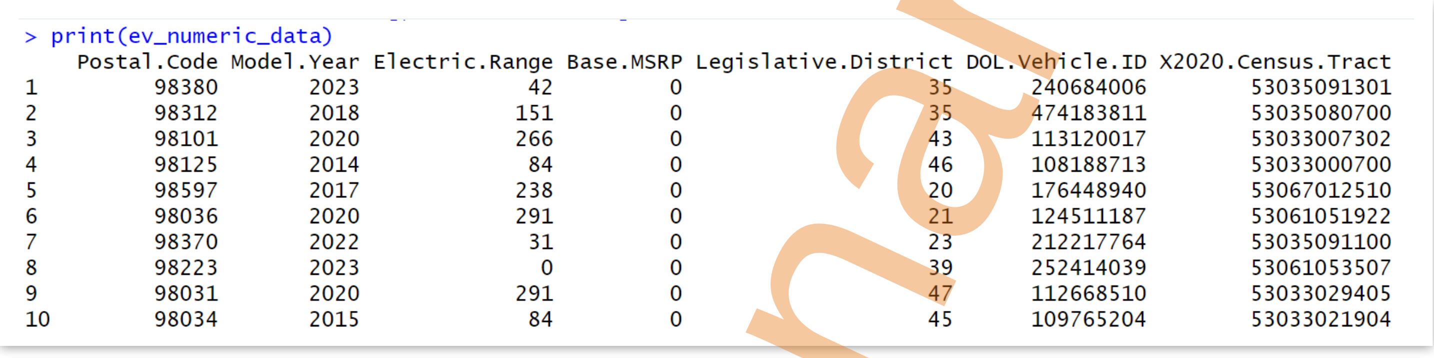
- Electric Utility: Power provider for the registered address.

- 2020 Census Tract: Census tract code.

Column Name	Description	Data Type
VIN (1-10)	The first 10 characters of each vehicle's Vehicle Identification Number (VIN), which uniquely identifies the vehicle.	Text
County	The geographic region within Washington State where the vehicle's owner is registered.	Text
City	The city in which the registered vehicle owner resides.	Text
State	The state associated with the vehicle's registered owner. This may include addresses located in states outside of Washington.	Text
Postal Code	The 5-digit zip code for the area where the registered vehicle owner resides.	Text
Model Year	The model year of the vehicle, determined from the Vehicle Identification Number (VIN).	Text
Make	The manufacturer of the vehicle, also derived from the VIN.	Text
Model	The specific model of the vehicle, identified through the VIN.	Text
Electric Vehicle Type	Indicates whether the vehicle is a Battery Electric Vehicle (BEV) or a Plug-in Hybrid Electric Vehicle (PHEV).	Text
Clean Alternative Fuel Vehicle (CAFV) Eligibility	Categorizes vehicles as Clean Alternative Fuel Vehicles based on legislative requirements.	Text
Electric Range	The distance the vehicle can travel on a full electric charge, measured in miles.	Numeric
Base MSRP	The manufacturer's suggested retail price (MSRP) for the base model of the vehicle.	Numeric
Legislative District	The specific legislative district in Washington State where the vehicle owner resides.	Text
DOL Vehicle ID	A unique identifier assigned to each vehicle by the Washington State Department of	Text

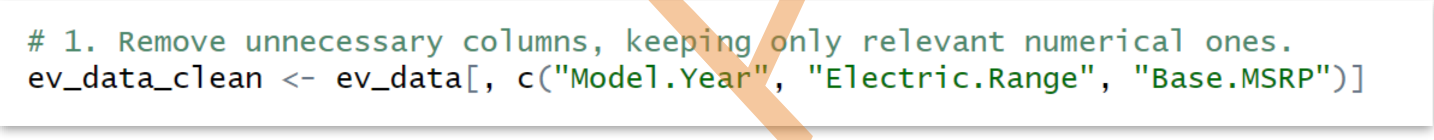
	Licensing.	
Vehicle Location	The geographical center of the zip code associated with the registered vehicle.	Point
Electric Utility	The electric power retail service territories that serve the address of the registered vehicle, with details on the types of utilities included.	Text
2020 Census Tract	The census tract identifier assigned by the United States Census Bureau, which helps in geographical analysis.	Text

Identifying the numerical columns and working on it



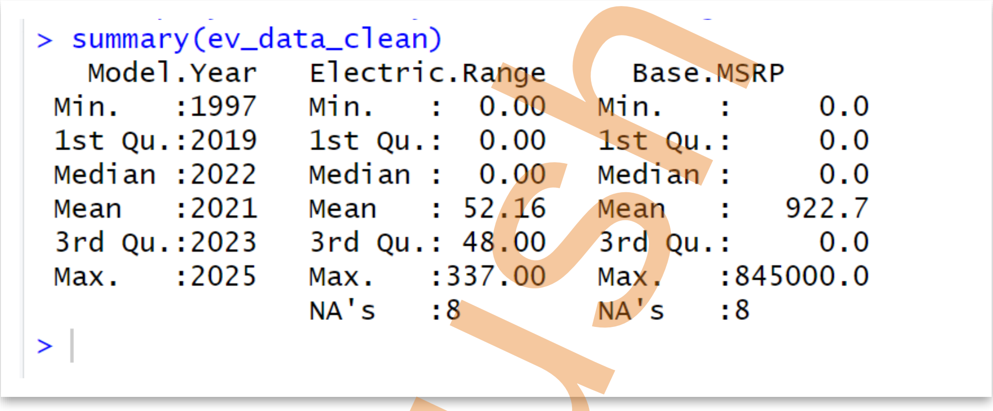
```
> print(ev_numeric_data)
   Postal.Code Model.Year Electric.Range Base.MSRP Legislative.District DOL.Vehicle.ID X2020.Census.Tract
1        98380       2023             42         0                   35      240684006        53035091301
2        98312       2018            151         0                   35      474183811        53035080700
3        98101       2020            266         0                   43      113120017        53033007302
4        98125       2014             84         0                   46      108188713        53033000700
5        98597       2017            238         0                   20      176448940        53067012510
6        98036       2020            291         0                   21      124511187        53061051922
7        98370       2022             31         0                   23      212217764        53035091100
8        98223       2023              0         0                   39      252414039        53061053507
9        98031       2020            291         0                   47      112668510        53033029405
10       98034       2015             84         0                   45      109765204        53033021904
```

Working on the numerical columns for further analysis



```
# 1. Remove unnecessary columns, keeping only relevant numerical ones.
ev_data_clean <- ev_data[, c("Model.Year", "Electric.Range", "Base.MSRP")]
```

And getting the summary of it.



```
> summary(ev_data_clean)
   Model.Year   Electric.Range     Base.MSRP       
 Min.   :1997   Min.   :  0.00   Min.   :     0.0  
 1st Qu.:2019   1st Qu.:  0.00   1st Qu.:     0.0  
 Median :2022   Median :  0.00   Median :     0.0  
 Mean   :2021   Mean   : 52.16   Mean   :   922.7  
 3rd Qu.:2023   3rd Qu.: 48.00   3rd Qu.:     0.0  
 Max.   :2025   Max.   :337.00   Max.   :845000.0  
                NA's   :8        NA's   :8         
> |
```

Identifying and Selecting Relevant Numerical Columns:

In this section of the analysis, I first identified all numerical columns in the dataset. This was achieved using the `**sapply()**` function, which checked each column and returned a logical result indicating whether the column was numeric. Once identified, I created a subset of the dataset containing only the numerical columns.

Next, I focused on the most relevant numerical variables for further analysis—specifically, "**Model Year**," "**Electric Range**," and "**Base MSRP**." By retaining only these columns, I simplified the dataset and made it more suitable for the analysis ahead. Finally, I used the `**summary()**` function to provide a statistical overview of these columns, which helped me understand the distribution, minimum and maximum values, and overall range of the selected variables. This step is crucial for both data cleaning and further statistical tests.

In the next step, I addressed any missing values within the selected numerical columns. Using the `**na.omit()**` function, I removed all rows that contained `**NA**` values to ensure the dataset was clean and ready for further analysis. After removing the missing data, I double-checked the dataset by using the `**sum(is.na())**` function, which confirmed that there were no remaining `**NA**` values. Finally, I generated a summary of the cleaned dataset to review its updated statistics and verify that the missing data had been successfully handled. This ensures that the analysis will not be biased or disrupted by incomplete records.

Handling missing values:

1. **Handling Missing Values:**

- Used the `na.omit()` function to remove rows containing `NA` (missing values) from the selected numerical columns.

- This ensures that only complete cases are included in the analysis, improving data reliability.

2. **Checking for Remaining Missing Values:**

- After cleaning, the `sum(is.na())` function was applied to verify that no missing values remained in the dataset.

- This step confirms that the data is fully cleaned and ready for analysis.

3. **Data Summary:**

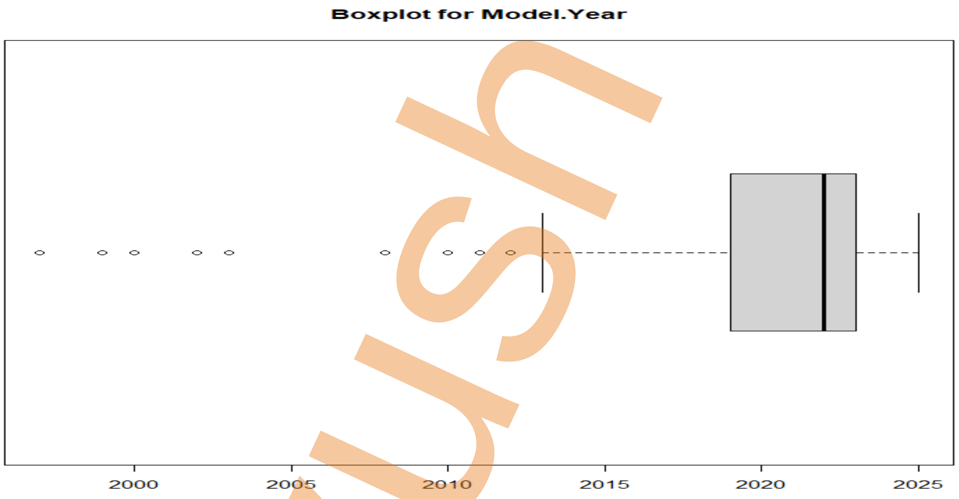
- A summary of the cleaned dataset was generated using the `summary()` function.

- This provides an overview of important statistics (mean, median, quartiles, etc.) for each of the numerical columns, allowing for an initial exploration of the dataset's distribution.

Detect and Visualize Outliers in Continuous Variables:

Outlier Detection:

- A boxplot is used to detect outliers in the Model.Year column, which is a continuous variable.
- Boxplots help identify outliers by showing data distribution and highlighting points outside the typical range based on the interquartile range (IQR).



Visualization of 'Model.Year' Data:

- The boxplot() function is used to create a visual representation of the Model.Year values, with any points outside the whiskers indicating potential outliers.
- The horizontal = TRUE argument creates a horizontal boxplot for clearer presentation.

Purpose of Visualization:

- This visual analysis helps in identifying outliers or extreme values that may need to be addressed or removed to ensure accurate analysis in later steps.

Remove Outliers from 'Model.Year'

1. Calculation of Quartiles:

- The first quartile (Q1) and third quartile (Q3) of the `Model.Year` column are calculated using the `quantile()` function to determine the bounds of the data.

- Q1 represents the 25th percentile, while Q3 represents the 75th percentile.

2. Interquartile Range (IQR):

- The Interquartile Range (IQR) is calculated as the difference between Q3 and Q1, which helps in determining the spread of the middle 50% of the data.

3. Defining Outlier Bounds:

- The lower and upper bounds for identifying outliers are set as 1.5 times the IQR below Q1 and above Q3. Any value outside these bounds is considered an outlier.

4. Removing Outliers:

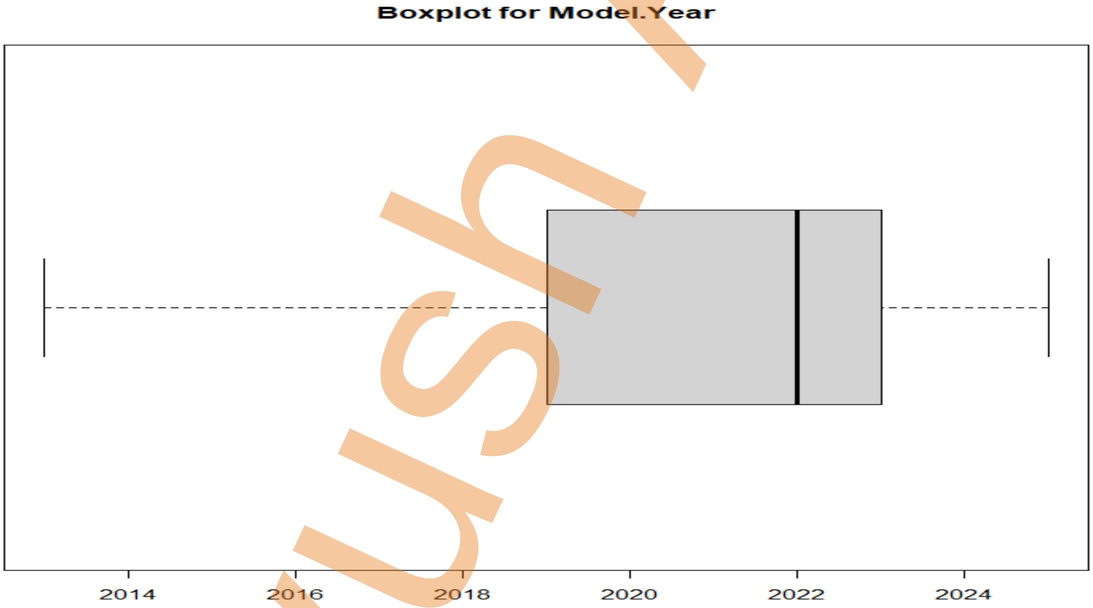
- A subset of the data is created, `cleaned_ev_data_clean`, where only the rows that fall within the defined bounds are kept. Outliers outside these bounds are removed.

5. Reporting the Number of Outliers Removed:

- The code calculates the difference in rows before and after removing outliers, reporting how many rows were excluded from the dataset.

6. Revisualizing the Data:

- A new boxplot is generated after outlier removal to visually confirm that the extreme values have been addressed, ensuring the cleaned data is now free of significant outliers.



Performed One-Sample t-Test on 'Model Year'

1. Hypothesis Test:

- The null hypothesis (H0) states that the mean model year is equal to 2020.

- The alternative hypothesis (H1) suggests that the mean model year is different from 2020.

2. Conducting the t-Test:

- A one-sample t-test is performed on the cleaned dataset to compare the mean of the `Model.Year` column against the hypothesized mean value of 2020.

- The test calculates whether the observed data is significantly different from the expected value.

3. Display t-Test Results:

- The test results, including the p-value and confidence intervals, are displayed to understand the statistical significance of the test.

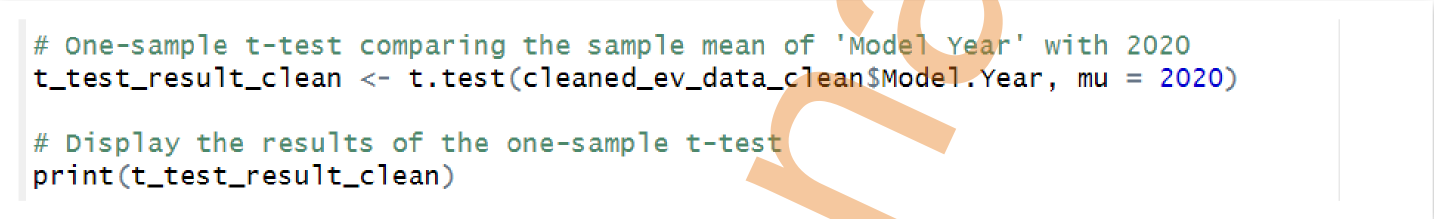
4. Sample Mean Calculation:

- The sample mean of the `Model.Year` is calculated and printed to provide an overview of the average model year in the dataset.

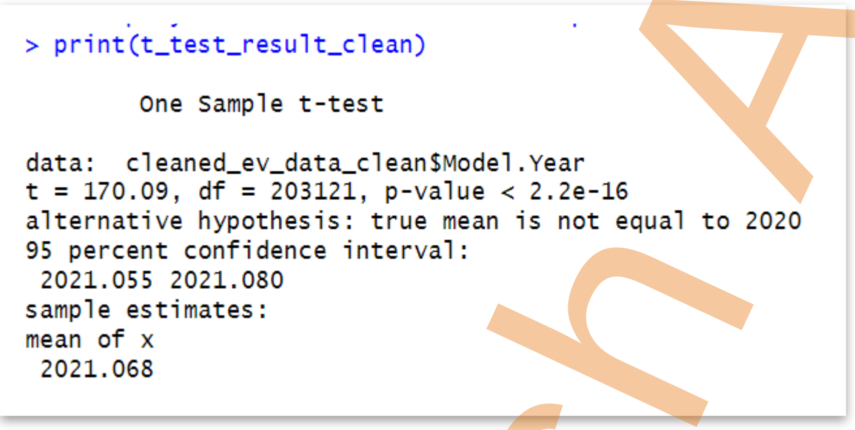
5. Interpretation of Results:

- If the p-value is less than 0.05 (5% significance level), the null hypothesis is rejected, indicating that the mean model year is significantly different from 2020.

- If the p-value is greater than 0.05, we fail to reject the null hypothesis, meaning there is no significant difference between the mean model year and 2020.



```
# One-sample t-test comparing the sample mean of 'Model Year' with 2020
t_test_result_clean <- t.test(cleaned_ev_data_clean$Model.Year, mu = 2020)

# Display the results of the one-sample t-test
print(t_test_result_clean)
```

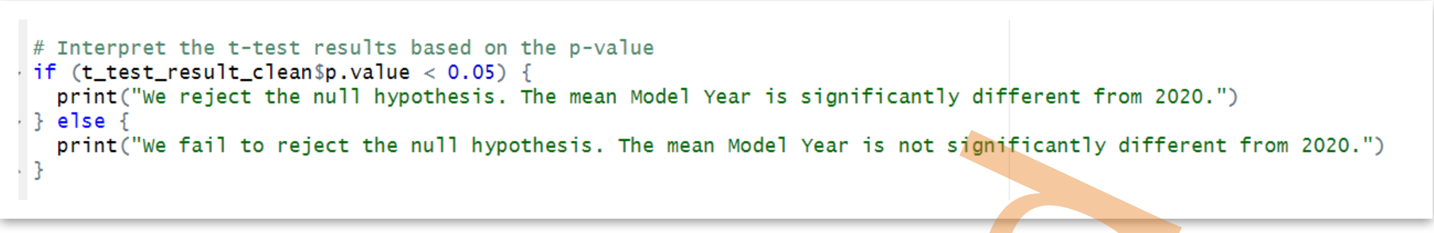


```
> print(t_test_result_clean)

	One Sample t-test

data:  cleaned_ev_data_clean$Model.Year
t = 170.09, df = 203121, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 2020
95 percent confidence interval:
 2021.055 2021.080
sample estimates:
mean of x 
 2021.068 
```

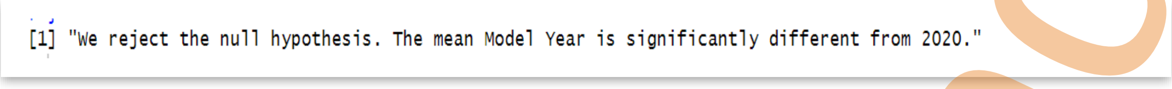
Conducted Hypothesis testing for P-value

Interpretation of One-Sample t-Test Results:

The one-sample t-test was conducted to determine if the mean `Model Year` of the electric vehicle population is significantly different from 2020. With a t-value of 170.09 and a p-value less than 0.05 (specifically, < 2.2e-16), we reject the null hypothesis. This indicates that the mean `Model Year` is significantly different from 2020. The sample mean was calculated to be approximately 2021.068, and the 95% confidence interval ranges from 2021.055 to 2021.080.



```
# Interpret the t-test results based on the p-value
if (t_test_result_clean$p.value < 0.05) {
  print("We reject the null hypothesis. The mean Model Year is significantly different from 2020.")
} else {
  print("We fail to reject the null hypothesis. The mean Model Year is not significantly different from 2020.")
}
```



```
[1] "We reject the null hypothesis. The mean Model Year is significantly different from 2020."
```

Conclusion:

This analysis of the Electric Vehicle Population dataset sheds light on registered electric vehicles in Washington State. The dataset contains 205,439 records and 17 columns, focusing on key variables such as `Model Year`, `Electric Range`, and `Base MSRP`.

Data cleaning processes included handling missing values and identifying numerical columns, ensuring the dataset's reliability. Outliers in the `Model Year` variable were detected using boxplots and subsequently removed, enhancing the accuracy of the analysis.

A one-sample t-test was conducted to compare the mean `Model Year` against 2020. The results indicated a significant difference, with a mean of approximately 2021.068 and a p-value far below 0.05, suggesting that most registered electric vehicles are newer models.

In summary, this analysis underscores the value of thorough data preprocessing and statistical testing in revealing trends in the electric vehicle sector, providing a foundation for further exploration and policy considerations.

Reference:

U.S. Government. (n.d.). *Electric vehicle population data*. Data.gov.

https://catalog.data.gov/dataset/electric-vehicle-population-data

Washington State. (n.d.). *Electric vehicle population data*. Data.WA.gov.

https://data.wa.gov/Transportation/Electric-Vehicle-Population-Data/f6w7-q2d2/about_data