Analysis of Transportation and Fuel Efficiency

INE2002 STATISTICS IN ENGINEERING

Abdullah Hani Abdellatif Al-shobaki

Table of Contents

Study Aim

The aim of this study is to analyze how various factors influence fuel efficiency in vehicles across Turkey. The importance of this study is to support environmental sustainability by improving fuel economy and informing vehicle manufacturing and policy decisions. To achieve this a set of statistical tools are going to be applied.

Data Description

2.1.Requirements

We used essential libraries that are useful for data analysis, visualization, and modeling.

```
library(tidyverse)
library(readxl)
library(car)
library(ggplot2)
library(dplyr)
```

2.2.Data collection and sampling

The dataset utilized for this analysis was obtained from the U.S. Department of Energy's Fuel Economy website, specifically the 2024 data file. This dataset comprises a detailed collection of vehicle specifications along with their fuel efficiency metrics, and to ensure that the dataset's manageability and relevance to the scope of our study, systematic sampling was employed. From the complete database, a subset of records was selected because they were related to vehicle specifications and fuel efficiency measures and they are 'Model Year', 'Division', 'Carline', and various fuel economy metrics like 'City FE (Guide) - Conventional Fuel', 'Hwy FE (Guide) - Conventional Fuel', and 'Comb FE (Guide) - Conventional Fuel'. We excluded columns that were not directly relevant to our study objectives or that contained redundant information for our analysis.

2.3.Reading the Data

```
data <- read_excel('2024 FE.xlsx')
```

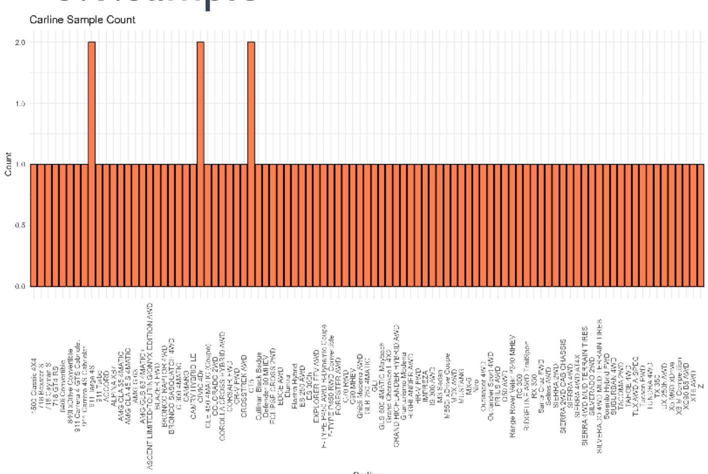
2.4.First 5 instances of the population

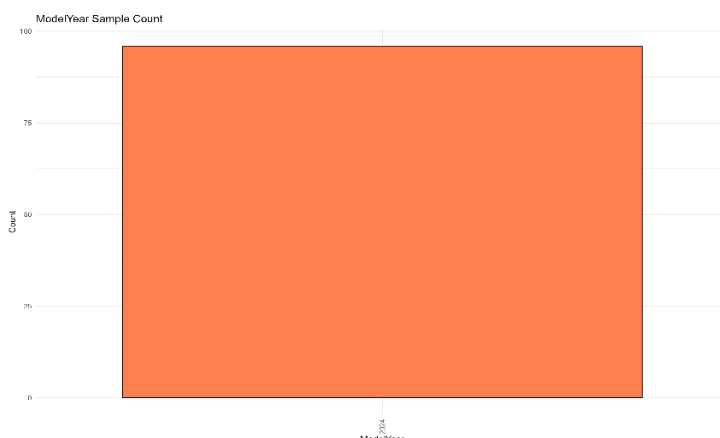
	Model Year	Mfr Name	Division	Carline	Verify Mfr Cd	Index (Model Type Index)	Eng Displ	# Cyl	Transmission	City
1	2024	aston martin	Aston Martin Lagonda Ltd	Valour	ASX	5	5.2	12	Manual(M6)	
2	2024	BMW	BMW	Z4 M40i	BMX	352	3.0	6	Auto(S8)	
3	2024	BMW	BMW	Z4 sDrive30i	BMX	350	2.0	4	Auto(S8)	
4	2024	Volkswagen Group of	Bugatti	Chiron Super Sport	VGA	68	8.0	16	Auto(AM-S7)	
5	2024	General Motors	Chevrolet	CORVETTE	GMX	153	6.2	8	Auto(S8)	

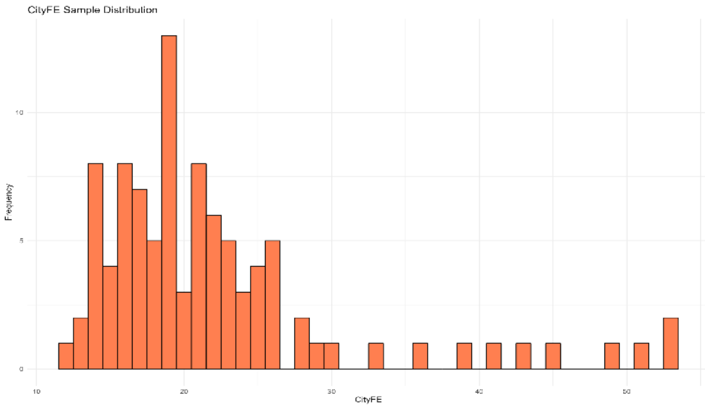
	Model Year	Division	Carline	City FE	Hwy FE	Comb FE
1	2024	Aston Martin Lagonda Ltd	Valour	12	18	14
2	2024	BMW	Z4 M40i	23	31	26
3	2024	BMW	Z4 sDrive30i	25	33	28
4	2024	Bugatti	Chiron Super Sport	8	11	9
5	2024	Chevrolet	CORVETTE	16	25	19

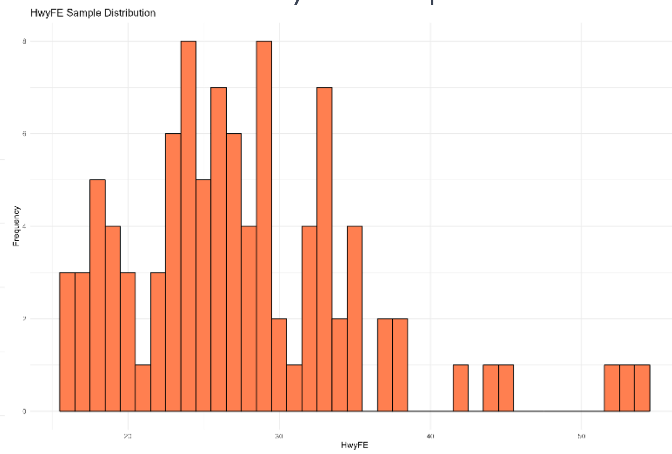
3.Data visualization

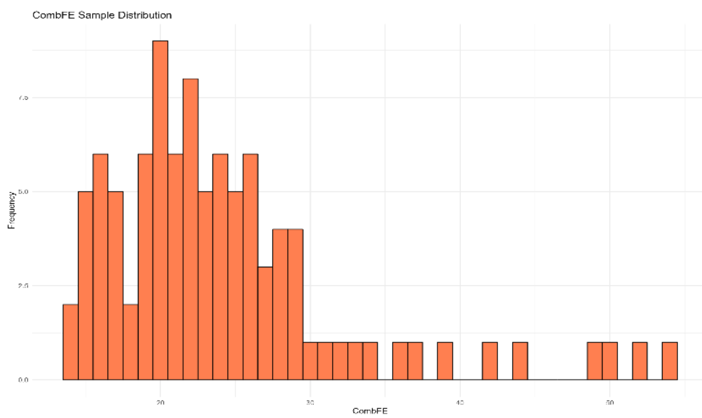
3.1.Sample



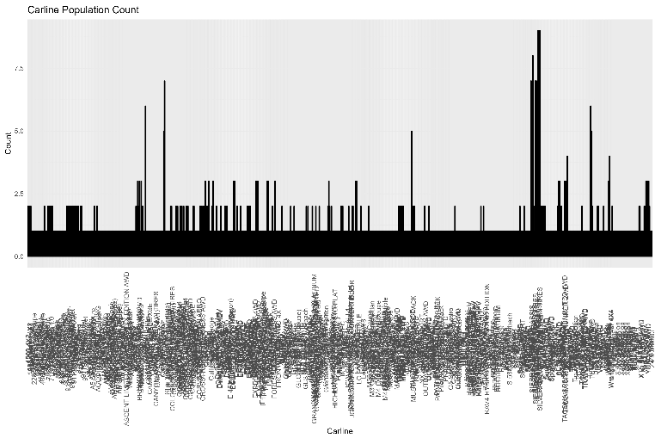


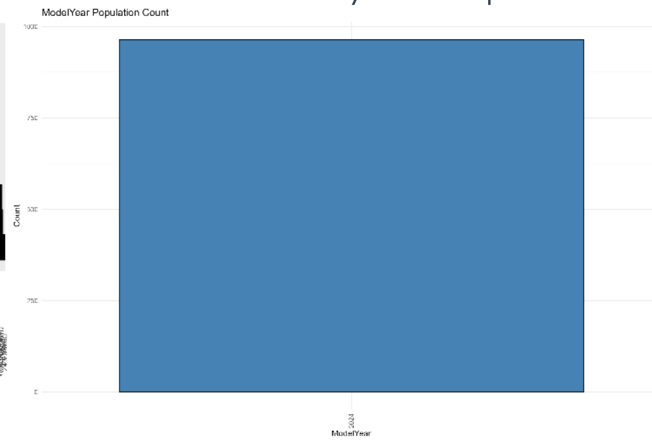


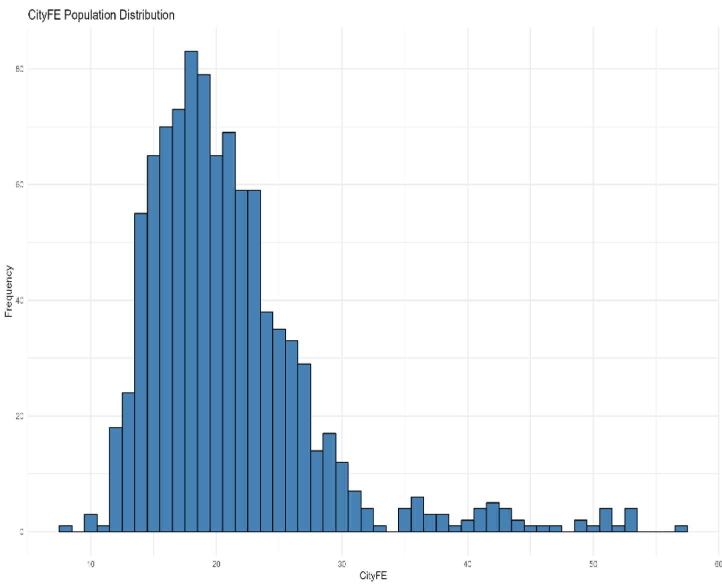


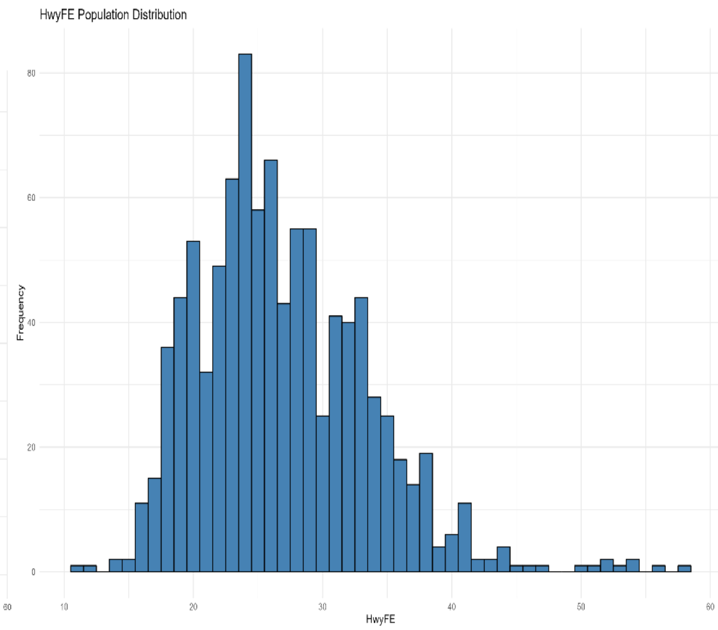


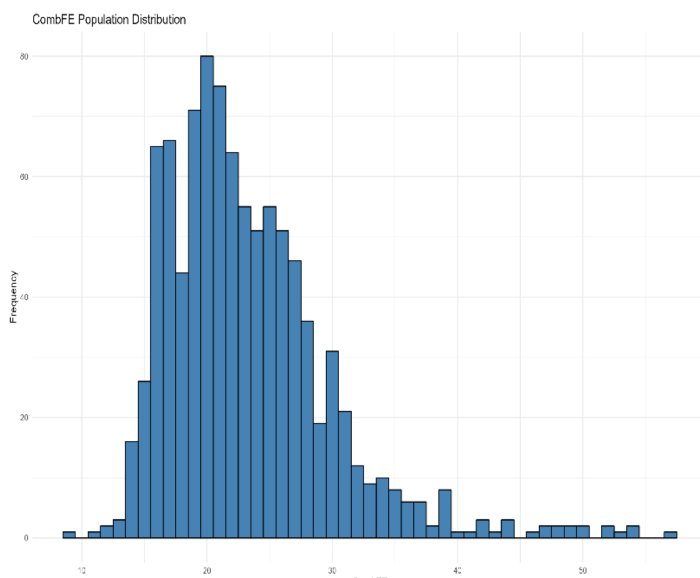
3.2.Population


Carline Population Count
Count
Carline


ModelYear Population Count
Count
2024
ModelYear


CityFE Population Distribution
Frequency
CityFE

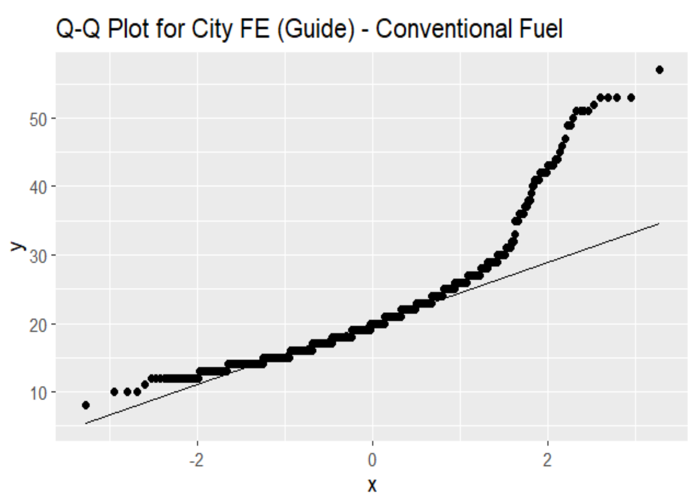

HwyFE Population Distribution
Frequency
HwyFE

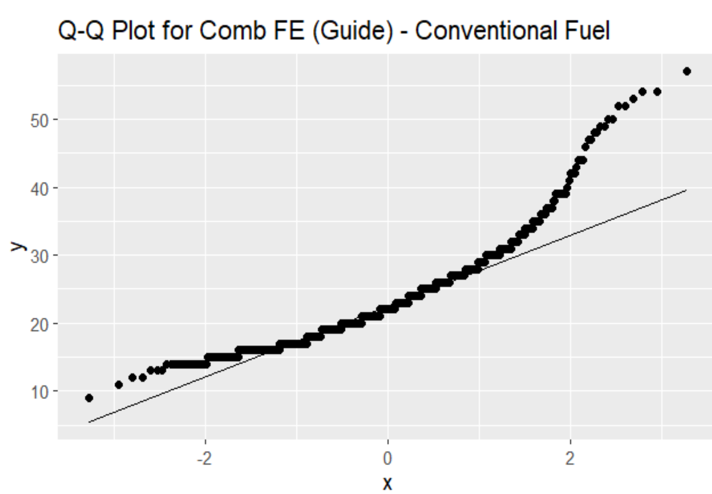

CombFE Population Distribution
Frequency
CombFE

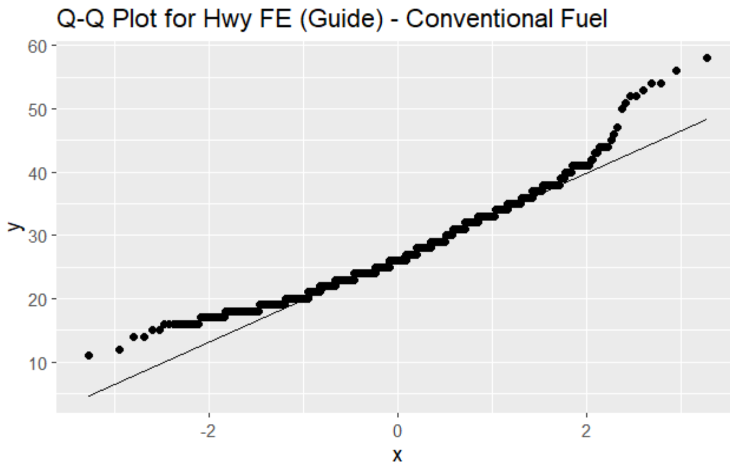
4.Normality tests

To determine if our data on fuel efficiency (City, Highway, and Combined) are normally distributed, which is critical for choosing the correct statistical methods for our analysis. We used the following approaches to assess normality: Shapiro-Wilk Test, Pearson's Coefficient of Skewness, and Q-Q Plots. The tests and plots indicate that the fuel efficiency data do not adhere to a normal distribution and values also indicated a moderate skewness. This suggests the need for non-parametric statistical methods in further analyses to avoid assumptions of normality.

```
Results for: City FE (Guide) - Conventional Fuel

        Shapiro-Wilk normality test

data:  data[[variable_name]]
W = 0.83995, p-value < 2.2e-16

Pearson's Coefficient of Skewness: 0.1502845

Results for: Hwy FE (Guide) - Conventional Fuel

        Shapiro-Wilk normality test

data:  data[[variable_name]]
W = 0.95742, p-value = 4.014e-16

Pearson's Coefficient of Skewness: 0.145478

Results for: Comb FE (Guide) - Conventional Fuel

        Shapiro-Wilk normality test

data:  data[[variable_name]]
W = 0.89883, p-value < 2.2e-16

Pearson's Coefficient of Skewness: 0.1896687
```







5. Point estimations and confidence intervals.

5.1.Point estimations

Sample means and standard deviations were calculated for key fuel efficiency metrics (City, Highway, and Combined) and compared with corresponding population values to assess sampling bias and data distribution. Functions were implemented to calculate these statistics accurately for both the sample and the population.

```
+   cat("Sample Mean:", sample_stats[[variable_name]]$mean, "Std Dev:",
+ }

Statistics for: CityFE
Population Mean: 21.03527 Std Dev: 6.888732
Sample Mean: 22.17708 Std Dev: 8.822215

Statistics for: HwyFE
Population Mean: 26.95851 Std Dev: 6.588667
Sample Mean: 27.63542 Std Dev: 7.829929

Statistics for: CombFE
Population Mean: 23.25726 Std Dev: 6.628724
Sample Mean: 24.25 Std Dev: 8.260241
>
```

5.2.Confidence Intervals

CI with t-test for City Fuel Efficiency, Highway Fuel Efficiency, and Combined Fuel Efficiency Mean: With 95% confidence, the population mean falls in the confidence interval [lower bound, upper bound].

```
+   cat("Sample 95% CI: [", sample_ci[[variable]]$lower, ",",
+ }

Confidence Interval for: CityFE
Population 95% CI: [ 20.59986 , 21.47068 ]
Sample 95% CI: [ 20.38954 , 23.96463 ]

Confidence Interval for: HwyFE
Population 95% CI: [ 26.54207 , 27.37495 ]
Sample 95% CI: [ 26.04893 , 29.22191 ]

Confidence Interval for: CombFE
Population 95% CI: [ 22.83829 , 23.67623 ]
Sample 95% CI: [ 22.57632 , 25.92368 ]
>
```

6.Hypothesis Testing

```
[1] "Hypothesis Test 1: City vs Highway Fuel Efficiency"
> print(city_hwy_test)

        Paired t-test

data:  data_population$CityFE and data_population$HwyFE
t = -61.461, df = 963, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -6.112364 -5.734109
sample estimates:
mean difference
      -5.923237
```

Hypothesis Test 1: City vs Highway Fuel Efficiency with alpha = 0.05
H0: The average city fuel efficiency (City FE) is the same as the average highway fuel efficiency (Hwy FE).
Ha: The average city fuel efficiency (City FE) is different from the average highway fuel efficiency (Hwy FE).
A paired t-test was used to evaluate the hypothesis and resulted in a p-value of less than 2.2e-16. This p-value is significantly less than the alpha level of 0.05, leading us to reject the null hypothesis in favor of the alternative hypothesis.

```
[1] "Hypothesis Test 2: Average Highway Fuel Efficiency Across Different Divisions"
             Df Sum Sq Mean Sq F value Pr(>F)
Division     40  17283   432.1   16.26 <2e-16 ***
Residuals   923  24522    26.6
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Hypothesis Test 2: Average Highway Fuel Efficiency Across Different Divisions with alpha = 0.05
H0: There is no significant difference in average highway fuel efficiency across different divisions.
Ha: There is a significant difference in average highway fuel efficiency across different divisions.
One-way ANOVA was conducted to test the hypothesis, yielding a p-value of less than 2.2e-16. With such a low p-value, far below the 0.05 alpha threshold, we reject the null hypothesis.

```
[1] "Hypothesis Test 3: Combined Fuel Efficiency Relation to Carline"
> print(comb_fe_kruskal)

        Kruskal-Wallis rank sum test

data:  CombFE by Carline
Kruskal-Wallis chi-squared = 906.88, df = 722, p-value = 3.091e-06
```

Hypothesis Test 3: Combined Fuel Efficiency Relation to Carline with alpha = 0.05
H0: There is no significant difference in combined fuel efficiency among different carlines.
Ha: There is a significant difference in combined fuel efficiency among different carlines.
The Kruskal-Wallis rank sum test was employed to analyze the hypothesis, resulting in a p-value of 3.091e-06. This p-value is below the alpha level of 0.05, thus we reject the null hypothesis in favor of the alternative hypothesis.

```
> summary(division_comb_fe_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
Division     40  16650   416.2   14.97 <2e-16 ***
Residuals   923  25664    27.8
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Hypothesis Test 4: Average Combined Fuel Efficiency Across Different Divisions with alpha = 0.05
H0: The combined fuel efficiency is the same across all vehicle divisions.
Ha: The combined fuel efficiency varies across different vehicle divisions.
An ANOVA was performed to assess the differences in combined fuel efficiency across 40 vehicle divisions. The F-value is reported as 14.97, with a highly significant p-value of less than 2e-16. This p-value is much lower than the alpha level of 0.05, leading us to reject the null hypothesis in favor of the alternative hypothesis.

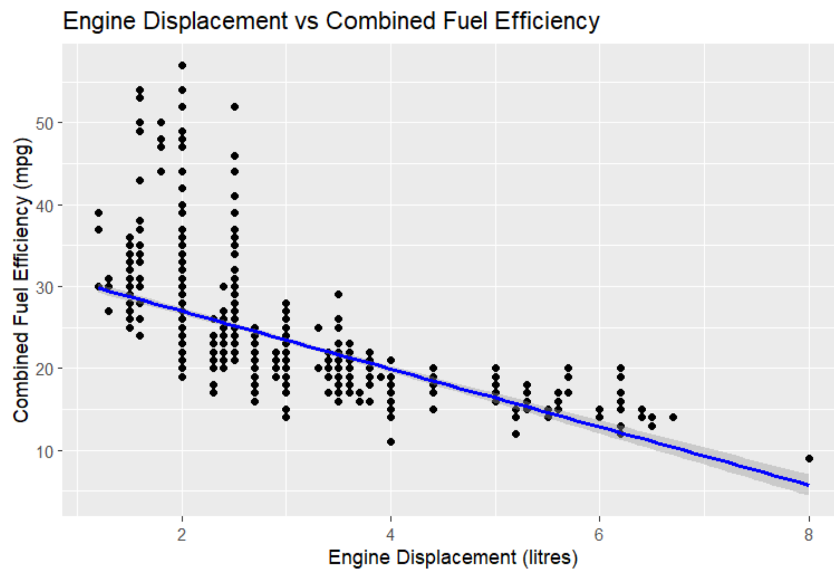
```
> print(carline_hwy_fe_kruskal)

        Kruskal-Wallis rank sum test

data:  HwyFE by Carline
Kruskal-Wallis chi-squared = 915.07, df = 722,
p-value = 1.271e-06
```

Hypothesis Test 5: Highway Fuel Efficiency Variation Among Carlines with alpha = 0.05
H0: The highway fuel efficiency is the same among different carlines.
Ha: The highway fuel efficiency differs among different carlines.
The Kruskal-Wallis test was used to investigate the hypothesis due to the non-normal distribution or ordinal nature of the data. The test resulted in a chi-squared value of 915.07, with degrees of freedom of 722, and a p-value of 1.271e-06. Since this p-value is significantly less than 0.05, we reject the null hypothesis.

7.Linear Regression.

Initially, a correlation plot was created to gain deeper insight into potential candidates for a linear regression model. This exploratory step was crucial in understanding the relationships between various vehicle characteristics and their impact on fuel efficiency. It was observed that among the numerous vehicle attributes assessed, not many showcased significant correlations. This led us to focus our analysis on a promising candidate that demonstrated a noteworthy relationship and thus warranted further detailed analysis and visualizations:

Engine Displacement vs Combined Fuel Efficiency.



7.1.First model.

The correlation value is:

```
> print(cor_value1)
[1] -0.6597891
```

```
lm1 <- lm(CombFE ~ EngDispl, data = data_population)
cooksd1 <- cooks.distance(lm1)
influential1 <- which(cooksd1 > (4/length(cooksd1)))
data_population_cleaned1 <- data_population[-influential1,]
cor_value1_clean <- cor(data_population_cleaned1$EngDispl, data_population_cleaned1$CombFE, use = "complete.obs")
print(cor_value1_clean)
```

After the initial correlation was established, the data was fitted to a linear regression model. To ensure the accuracy and reliability of the model, influential points were identified using Cook's distance which is a common method for detecting outliers and influential data points in least squares regression analysis.

The new correlation value after influential points removal:

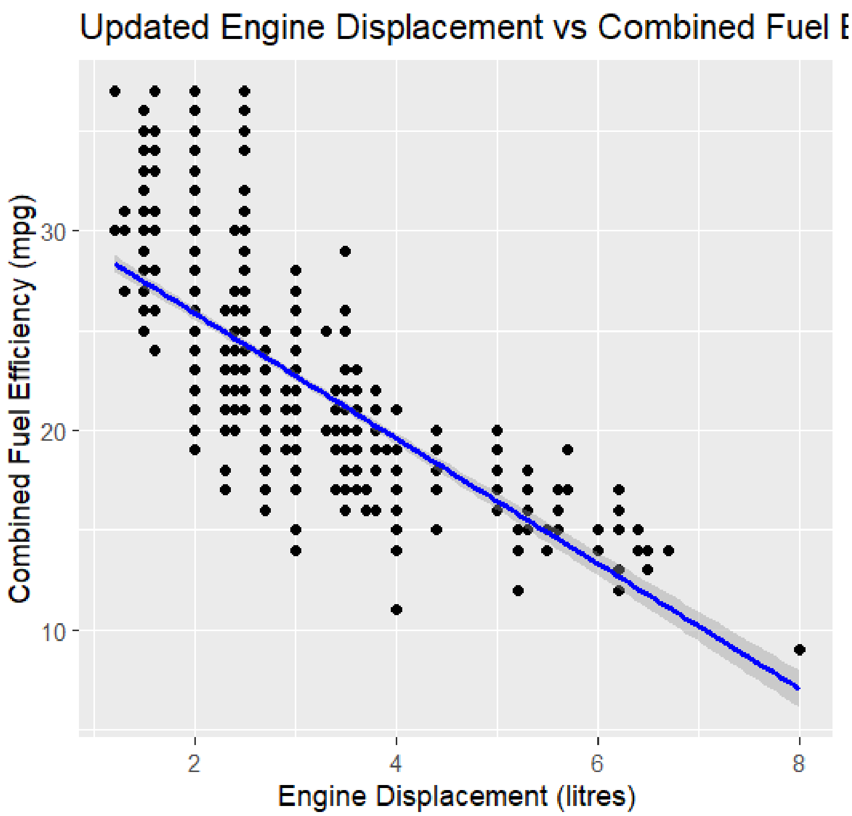
```
> print(cor_value1_clean)
[1] -0.7426179
```

Updated model:

A few predictions were made using the model as seen below.

```
> predictions1 <- predict(lm1_updated, newdata = new_data1)
> print(predictions1)
        1         2         3         4         5         6         7         8         9        10 
28.374949 26.003305 23.631662 21.260018 18.888375 16.516731 14.145088 11.773444  9.401801  7.030157 
```

Visualizations:



7.2.Second model:

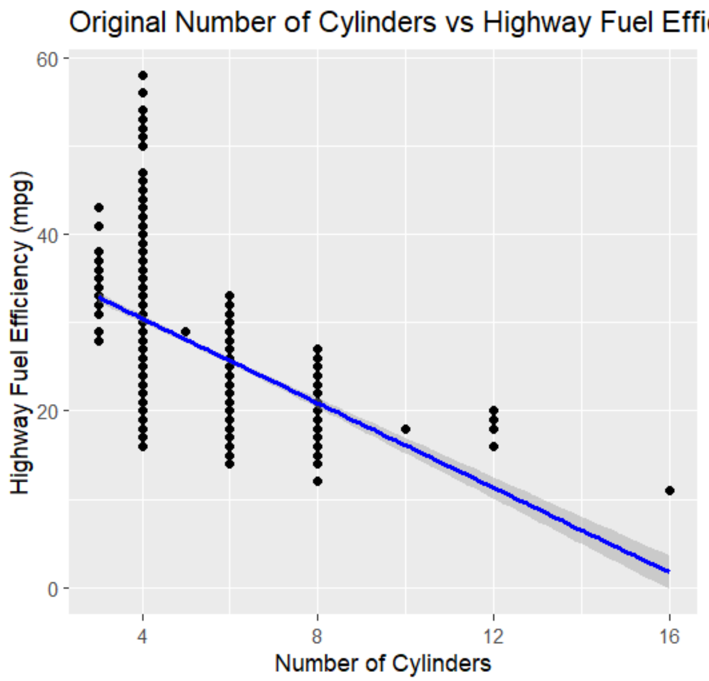
The correlation value is:

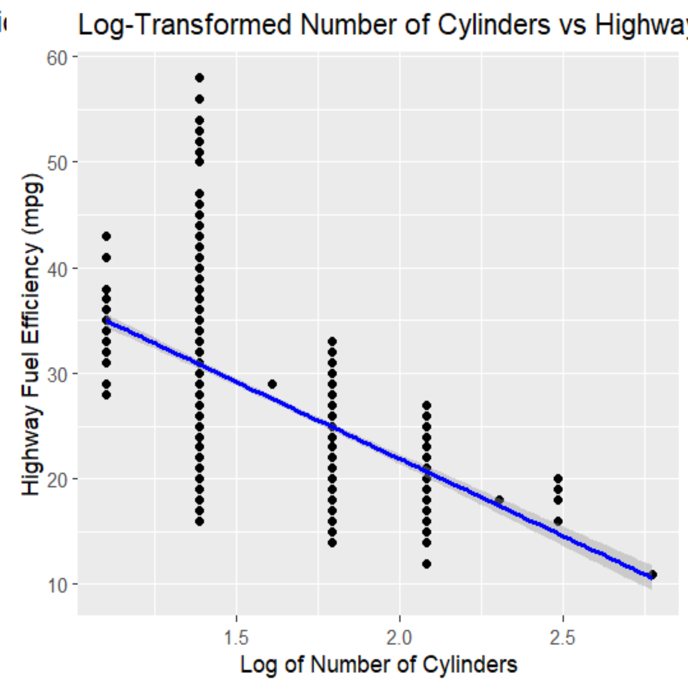
```
> print(cor_value2)
[1] -0.6463057
```

In the process of optimizing the linear regression model, a log transformation of the number of cylinders was applied. It was noticed that this transformation enhanced the linearity of the relationship between the transformed number of cylinders and highway fuel efficiency, as evidenced by a more negative correlation coefficient post-transformation. This indicates a better model fit and justifies the use of the log-transformed data for the final model. Here's the correlation value after transformation:

```
> print(cor_value2_log)
[1] -0.6669516
```

Before: After:



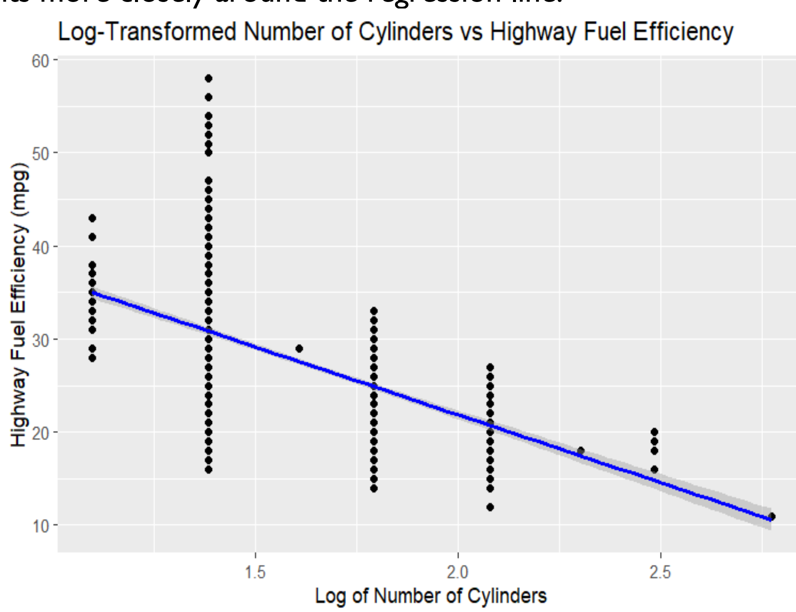


Updated model:

A few predictions were made using the model as seen below.

```
> print(predictions2_log)
       1        2        3        4        5        6        7        8        9       10 
35.03395 29.30093 25.19617 21.99649 19.37398 17.15190 15.22401 13.52144 11.99699 10.61690 
```

The graph showcases the linear regression model for highway fuel efficiency against the log-transformed number of cylinders. This visualization confirms that the transformation has indeed led to a better fit by reducing the variability and aligning the data points more closely around the regression line.



8.Chi Square Tests.

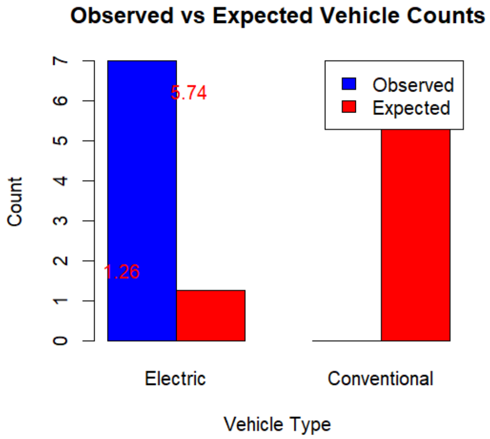
8.1.Goodness of fit

According to current industry data, the expected market share for electric vehicles in 2024 is projected at 18%. In contrast, conventional vehicles are expected to maintain the majority with an 82% market share. In this step, 'VehicleType' and 'Transmission' were analyzed to understand their relationships with fuel efficiency. To test these expected values against our dataset, a goodness of fit test was performed.

```
Observed Counts:
> print(observed_counts)
    Electric Conventional
           7            0
> cat("\nExpected Counts:\n")

Expected Counts:
> print(expected_numbers)
    Electric Conventional
        1.26         5.74
>
```

```
> # Print the results of the Chi-Square test
> print(chi_square_test)

        Chi-squared test for given probabilities

data:  observed_counts
X-squared = 31.889, df = 1, p-value = 1.632e-08
```

H0: The percentages of vehicle types in our data match the expected market shares.
Ha: The percentages of vehicle types in our data do not match the expected market shares.
Upon collection, the observed values indicated a total of 7 electric vehicles with an absence of conventional vehicles. These observed counts were compared to the expected counts using the **chisq.test()** method in R. The resulting Chi-Square test yielded a highly significant p-value of 1.632e-08, leading us to reject the null hypothesis in favor of the alternative hypothesis. The absence of conventional vehicles in our dataset, and the overrepresentation of electric vehicles, is particularly striking. This could be attributed to the dataset being drawn from a segment of the market that significantly favors electric vehicles.



The bar plot distinctly visualizes the difference between the expected and observed vehicle counts in our dataset.

8.2.Independence test

```
> print(chi_square_independence_test)

        Chi-squared test for given probabilities

data:  table(data_population$VehicleType, data_population$Transmission)
X-squared = 161, df = 23, p-value < 2.2e-16
```

H0: The categorical variables 'VehicleType' and 'Transmission' are independent.
Ha: The categorical variables 'VehicleType' and 'Transmission' are dependent.
A contingency table of both variables was created and the chisq.test() method was applied. The test resulted in a Chi-squared statistic of 161 with 23 degrees of freedom and a p-value < 2.2e-16, which is highly significant. Thus, we reject the null hypothesis and conclude that there is a dependency between the type of vehicle and the type of transmission.

8.3.Homogeneity test

```
> # Print the results of the test
> print(chi_square_homogeneity_test)

        Chi-squared test for given probabilities

data:  vehicle_type_by_carline
X-squared = 716, df = 722, p-value = 0.556
```

We investigated the distribution of vehicle types among various car lines. We compiled our data into a contingency table and executed the chisq.test() method.
H0: was that the distribution of vehicle types is consistent across different car lines, implying homogeneity.
Ha: suggested that there is a variation in the distribution of vehicle types among the car lines.

Upon analysis, the chi-squared test produced a p-value of 0.556. This p-value does not provide enough evidence to reject the null hypothesis. Consequently, we conclude that there is no statistically significant difference in the distribution of vehicle types across car lines, and we fail to reject the null hypothesis.

9.ANOVA

9.1.One way ANOVA

```
> print(summary(carline_fe_anova))
             Df Sum Sq Mean Sq F value Pr(>F)
Carline     722  40741   56.43   8.645 <2e-16 ***
Residuals   241   1573    6.53
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

H0: The combined fuel efficiency mean is the same across all carline categories.
Ha: The combined fuel efficiency mean differs for at least one carline category.

The ANOVA test resulted in a highly significant p-value ($p < 2e-16$), leading to the rejection of the null hypothesis in favor of the alternative.

```
> print(tukey_test)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = CombFE ~ Carline, data = data_population)

$Carline
                                                                            diff         lwr          upr
1500 4X4-1500 4X2                                                  -1.000000e+00 -14.28170706 12.281707061
1500 Classic 4X2-1500 4X2                                          -2.500000e+00 -15.78170706 10.781707061
1500 Classic 4X4-1500 4X2                                          -3.000000e+00 -16.28170706 10.281707061
1500 HFE 4X2-1500 4X2                                               2.000000e+00 -14.26670261 18.266702606
1500 TRX 4x4-1500 4X2                                              -9.000000e+00 -25.26670261  7.266702606
228i Gran Coupe-1500 4X2                                            7.000000e+00  -9.26670261 23.266702606
228i xDrive Gran Coupe-1500 4X2                                     6.000000e+00 -10.26670261 22.266702606
230i Coupe-1500 4X2                                                 9.000000e+00  -7.26670261 25.266702606
230i xDrive Coupe-1500 4X2                                          7.000000e+00  -9.26670261 23.266702606
330i Sedan-1500 4X2                                                 8.000000e+00  -8.26670261 24.266702606
330i xDrive Sedan-1500 4X2                                          6.000000e+00 -10.26670261 22.266702606
430i Convertible-1500 4X2                                           7.000000e+00  -9.26670261 23.266702606
430i Coupe-1500 4X2                                                 7.000000e+00  -9.26670261 23.266702606
430i Gran Coupe-1500 4X2                                            7.000000e+00  -9.26670261 23.266702606
430i xDrive Convertible-1500 4X2                                    6.000000e+00 -10.26670261 22.266702606
430i xDrive Coupe-1500 4X2                                          6.000000e+00 -10.26670261 22.266702606
430i xDrive Gran Coupe-1500 4X2                                     6.000000e+00 -10.26670261 22.266702606
4RUNNER 2WD-1500 4X2                                               -4.000000e+00 -20.26670261 12.266702606
4RUNNER 4WD-1500 4X2                                               -4.000000e+00 -17.28170706  9.281707061
530i Sedan-1500 4X2                                                 9.000000e+00  -7.26670261 25.266702606
```

Subsequently, Tukey's HSD test was applied to determine which carline categories differed in terms of mean combined fuel efficiency. Because the data of the carline was massive, we uploaded a sample of the results.

9.2.Two way ANOVA

```
> summary(anova_result)
                         Df Sum Sq Mean Sq F value    Pr(>F)    
Transmission             23  24779  1077.3 146.473  < 2e-16 ***
Carline                 714  16237    22.7   3.092 8.33e-16 ***
Transmission:Carline     69    143     2.1   0.282        1    
Residuals               157   1155     7.4                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Three hypotheses:

H0$_1$: There is no significant difference in 'Combined Fuel Efficiency' across different 'Carline' categories.
Ha$_1$: There is a significant difference in 'Combined Fuel Efficiency' across different 'Carline' categories.

H0$_2$: There is no significant difference in 'Combined Fuel Efficiency' across different transmission types.
Ha$_2$: There is a significant difference in 'Combined Fuel Efficiency' across different transmission types.

H0$_3$: There is no interaction effect between 'Carline' and 'Transmission' on 'Combined Fuel Efficiency'.
Ha$_3$: There is an interaction effect between 'Carline' and 'Transmission' on 'Combined Fuel Efficiency'.

We examined the impacts of transmission types and car model on the combined fuel efficiency of vehicles. The analysis yielded highly significant results for the main effects of both transmission type ($p < 2e-16$) and carline ($p < 2e-16$), compelling us to reject the null hypotheses H0$_1$ and H0$_2$. Conversely, the interaction effect between transmission type and car model was not significant ($p = 1$), leading us to retain the null hypothesis H0$_3$.

10.Non-parametric tests.

10.1.Sign test.

```

        One-sample Sign-Test

data:  data_population$CombFE
s = 106, p-value < 2.2e-16
alternative hypothesis: true median is not equal to 30
95 percent confidence interval:
 22 22
sample estimates:
median of x 
         22 

Achieved and Interpolated Confidence Intervals: 

                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.9427     22     22
Interpolated CI       0.9500     22     22
Upper Achieved CI     0.9506     22     22
```

H0: The true median of 'Combined Fuel Efficiency' is 30 mpg.
Ha: The true median of 'Combined Fuel Efficiency' is not equal to 30 mpg.

The Sign Test was applied to test the hypothesis that the true median 'Combined Fuel Efficiency' differs from the industry standard of 30 mpg. The test resulted in a p-value less than 2.2e-16, indicating that the null hypothesis can be rejected. Thus, we conclude that there is a statistically significant difference between the sample median of 'Combined Fuel Efficiency' and the hypothesized value of 30 mpg, with a sample median of 22 mpg.

10.2.Wilcoxon Signed-Rank Test.

```
        Wilcoxon signed rank test with continuity correction

data:  non_missing_fe
V = 35051, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 30
```

H0: The true median 'Combined Fuel Efficiency' is 30 mpg.
Ha: The true median 'Combined Fuel Efficiency' is not equal to 30 mpg.

The Wilcoxon Signed-Rank Test was conducted to evaluate if the median 'Combined Fuel Efficiency' in the dataset significantly differs from the industry benchmark of 30 mpg. The test yielded a V statistic of 35051 with a p-value less than 2.2e-16, strongly suggesting the null hypothesis can be rejected.

10.3.Spearman correlation.

```
        Spearman's rank correlation rho

data:  data_population$CombFE and data_population$EngDispl
S = 269368281, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.8041269
```

The cor.test() was applied to see if there was correlation between 'Combined Fuel Efficiency' and 'Engine Displacement'.

H0: There is no monotonic association between 'Combined Fuel Efficiency' and 'Engine Displacement'.
Ha: There is a monotonic association between 'Combined Fuel Efficiency' and 'Engine Displacement'.

The Spearman's rank correlation test resulted in a rho value of -0.8041269, with a p-value less than 2.2e-16, which is highly significant. Thus, the null hypothesis is rejected in favor of the alternative hypothesis, indicating a strong negative monotonic relationship between 'Combined Fuel Efficiency' and 'Engine Displacement'.

10.4.Runs test

```
        Runs Test

data:  data_population$CombFE
statistic = -18.21, runs = 178, n1 = 450, n2 = 450, n = 900, p-value < 2.2e-16
alternative hypothesis: nonrandomness
```

The Runs test is used to check for randomness in the order of 'Combined Fuel Efficiency'.

H0: Random.
Ha: Not Random.

The p-value resulting from the test is not significant enough to reject the null hypothesis, so we fail to reject. So, not randomness.

References

1) U.S. Environmental Protection Agency. (2024). https://www.fueleconomy.gov/feg/download.shtml

Conclusion

Our study analyzed factors influencing vehicle fuel efficiency using data from the U.S. Environmental Protection Agency. We employed various statistical methods including linear regression, Chi-square tests, ANOVA, and non-parametric tests. These techniques were essential for identifying key influences on fuel efficiency, guiding future improvements in vehicle design and informing policy development aimed at enhancing fuel economy.

Team Work

All members worked equally and simultaneously through online zoom meetings and F2F at university campus.

Youtube Link