DATA 119 Project

Disha Mohta, Chloe Attlan, Fredric Ngo

The Fuel Economy Dataset is provided by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy and the U.S. Environmental Protection Agency[1]. It is hosted on the FuelEconomy.gov website and offers comprehensive data about the fuel efficiency of various vehicles sold in the United States. The dataset is a result of a collaborative effort between the above-mentioned federal agencies. Their objective is to inform consumers about the fuel economy, annual fuel costs, and environmental impact of different vehicles, empowering them to make informed choices when purchasing. The data is compiled from manufacturer submissions, tested under controlled conditions, and standardized to allow for comparison across different vehicles and model years. The dataset encompasses a broad spectrum of vehicle attributes such as make, model, engine specifications, transmission type, drive type, fuel type, city and highway miles per gallon (MPG), and greenhouse gas emissions, among others. Over the years, the dataset has been periodically updated to include newer vehicle models, and its format has evolved to accommodate emerging technologies like hybrid and electric vehicles.

Researchers have consistently observed improvements in vehicle fuel efficiency over time. This trend is often attributed to technological advancements and stricter government regulations. Studies have shown a gradual increase in the average miles per gallon (MPG) ratings of new vehicles sold in the U.S [2]. Several studies have examined the impact of policies like the Corporate Average Fuel Economy (CAFE) standards on the automotive industry. These studies generally find that such regulations have significantly contributed to improvements in fuel efficiency [3]. Research using this dataset has also delved into consumer preferences and buying patterns. It's been observed that consumer choices often fluctuate with fuel prices; for instance, higher gasoline prices tend to shift consumer preference towards more fuel-efficient vehicles [4]. While the dataset is comprehensive, it primarily focuses on new vehicles and may not fully represent the performance of vehicles over their entire lifecycle. Accurately modeling fuel economy and emissions is complex and can be influenced by numerous variables, including driving behavior, vehicle maintenance, and environmental conditions.

In exploring this dataset, we are trying to answer the following question: **What factors cause cars to emit more CO2?** We want to understand how certain factors affect the emission of CO2 in cars and use this information to formulate an understanding of how we can reduce pollution through directed policy measures.

---

[1] Fuel Economy Dataset - United States Environmental Protection Agency (2023), *Fuel Economy* [Data set] https://www.fueleconomy.gov/feg/download.shtml

[2] Smith, A., & Johnson, L. (2021). "Trends in Vehicle Fuel Efficiency: A Comparative Analysis." Journal of Automotive Technology, 34(2), 45-59.

[3] Green, M. (2020). "The Impact of CAFE Standards on the Automotive Industry." Environmental Policy Review, 28(3), 112-130.

[4] Lee, K., & Patel, S. (2019). "Consumer Behavior in Vehicle Purchases: The Role of Fuel Prices." Economic Insights, 22(4), 54-68.

Initial Passes at Data Analysis, Visualizations, and Summarizations

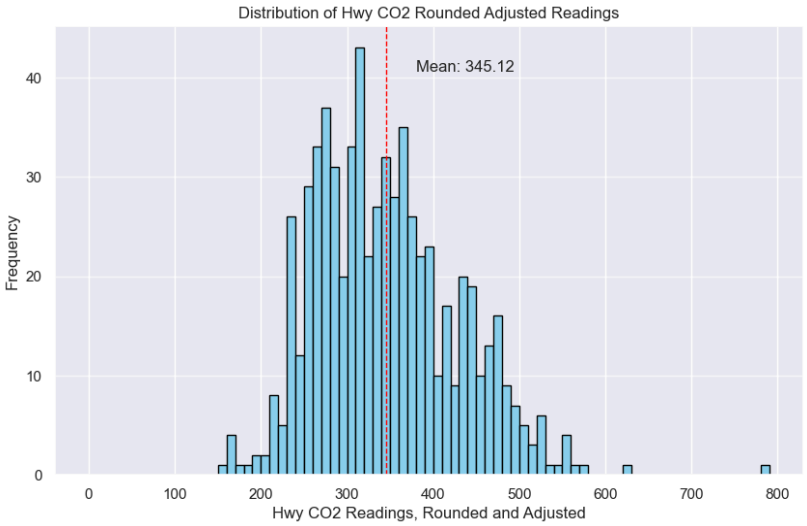


Figure 1: The histogram shows the distribution of highway CO2 readings, over time rounded and adjusted. The distribution is centered around 345.12, and ranges from 158 to 788. It is unimodal, and skewed slightly to the right. There aren't many potential unusual values.

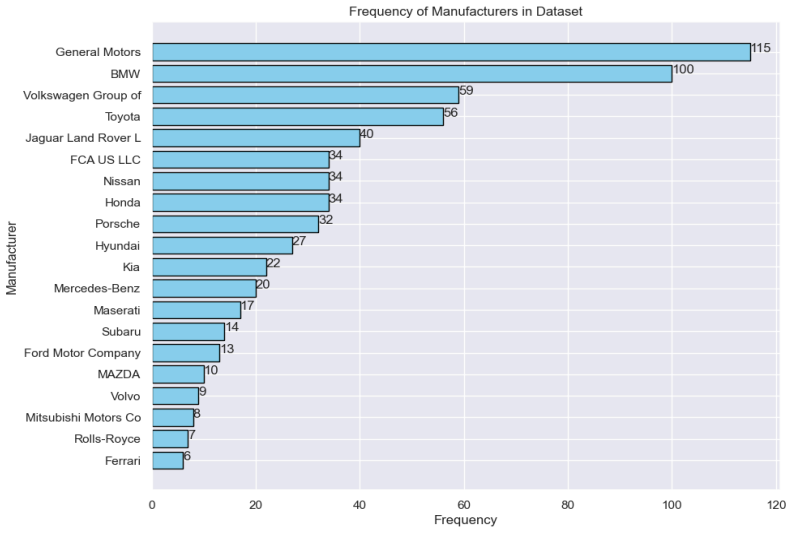


Figure 2: The horizontal bar chart shows the distribution of manufacturers who built the cars being tracked in this dataset. The most frequent manufacturers was General Motors and BMW, and the least frequent manufacturer was Ferrari.

Linear Regression & Multiple Linear Regression

Our linear regression investigated the relationship between fuel efficiency and CO2 emissions. We calculated an equation of: **CO2 Emissions = -0.07(Fuel Efficiency) + 47.438**, where CO2 Emissions were in grams and Fuel Efficiency was in miles/gallon. Below, we plotted the line of best fit and the data points, the predicted values versus the residuals, and the frequency of the residuals. The frequency of residuals graphs looks appropriate, with a unimodal and relatively symmetrical shape. However, the predicted value versus residuals indicates that a linear regression is inappropriate, with lower variance between values 20 and 30 and higher variance when the values were higher or lower; this violates the assumption of homoscedasticity.

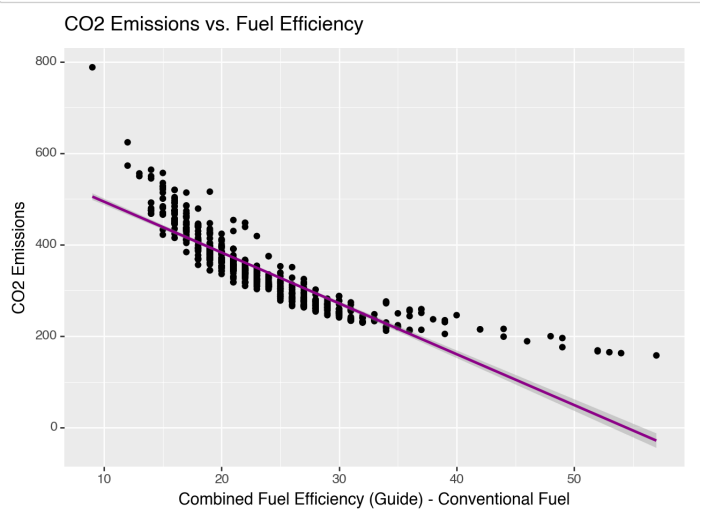
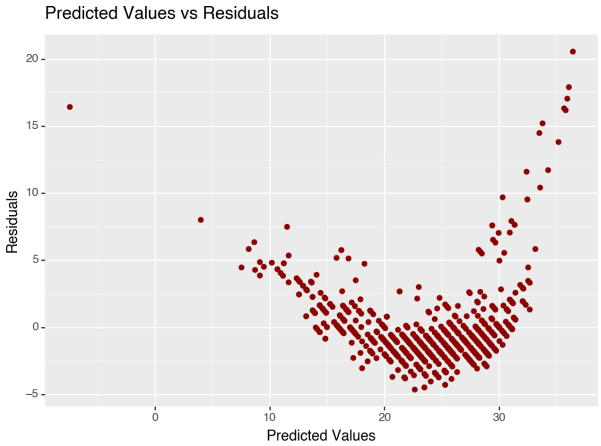


Figure 3: The plot includes the points, mapping fuel efficiency to CO2 emissions, along with the line of best fit, found with our linear regression model. The line seems to best fit the data when the x-values are in between 15 and 40. Before x-values of 15 or after values of 40, the line is less representative, potentially indicating that the data is not entirely linear and thus that another model may be best.

`co2 = −0.07(fuel efficiency (MPG)) + 47.438`



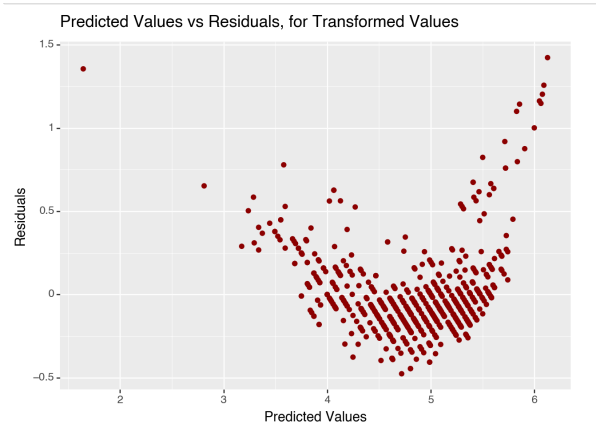


Figure 4 & 5

The first graph plots the residuals against the predicted value of the model. This plot, to indicate that a linear model is appropriate, should look like a random scattering of points. Instead, there seems to be a U-shape, starting around 10, dipping between 20 and 30, and increasing again after 30. This violates the assumption of roughly constant variability, indicating that we should use another model to best represent this relationship. In an attempt to resolve this issue, we transformed the X values, trying to find both the squares and square roots. The second graph shows one of these attempts, where we took the square root of each X value. Although it helped disperse the U-shape, the residuals still lacked constant variability, indicating that another model may be best to use.

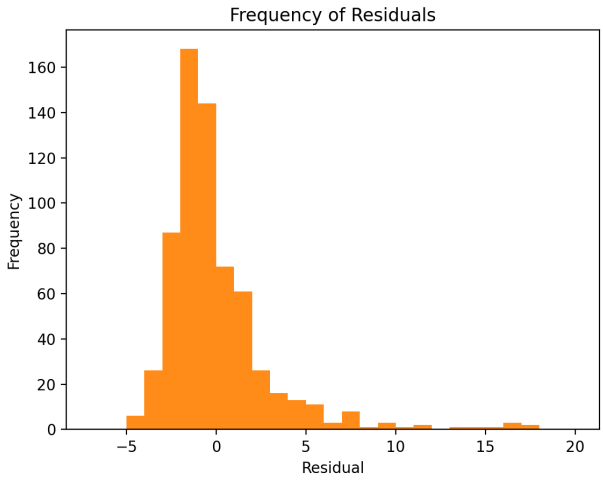


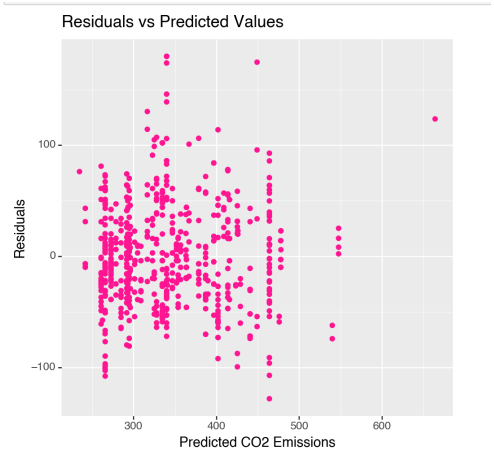
Figure 5

This model plots the frequency of residuals. This plot is unimodal, and is relatively symmetric (although it has a longer right tail). Although this plot supports using a linear model, the rest of the findings (the residuals) indicate otherwise.

Our multiple linear regression investigated different factors that affect CO2 emissions, including the manufacturer of the car (which is shown as dummy variables, with BMW as the reference level), the number of cylinders, and the engine displacement (volume in liters). The equation found was:

CO2 emissions = 148.6037 + 52.2653(FCA US LLC) + 26.6075(Ferrari) + 32.1401(Ford Motor Company) + 67.1767(General Motors) - 11.9416(Honda) + 22.0182(Hyundai) + 43.8544(Jaguar Land Rover) - 7.1498(Kia) + 6.0084(MAZDA) + 29.0124(Maserati) + 28.4816(Mercedes-Benz) + 20.7362(Mitsubishi Motors Co) + 12.1039(Nissan) − 79.0677(Porsche) - 43.2590(Rolls-Royce) + 20.2255(Subaru) - 7.0170(Toyota) + 19.0013(Volkswagen Group of) + 23.3366(Volvo) + 31.0427(Number of Cylinders)

This model had an R^2 of 0.689 and adjusted R-squared of 0.679. In terms of correlation, most correlations were low (between –0.1 and 0.1), except for the number of cylinders and the Rolls Royce variables, with a correlation of 0.363. In terms of VIF, all values (except the constant) had VIFs under 10, showing no multicollinearity; moreover, most of the VIFs hovered around 1. With regards to the residuals, there were no clear patterns, meaning key assumptions to run this model were not violated.



VIF DATA

	feature	VIF
0	const	18.782264
1	FCA US LLC	1.276502
2	Ferrari	1.163626
3	Ford Motor Company	1.111773
4	General Motors	1.800727
5	Honda	1.285331
6	Hyundai	1.228271
7	Jaguar Land Rover L	1.337198
8	Kia	1.193996
9	MAZDA	1.088536
10	Maserati	1.149312
11	Mercedes-Benz	1.173864
12	Mitsubishi Motors Co	1.079413
13	Nissan	1.271275
14	Porsche	1.256514
15	Rolls-Royce	1.264319
16	Subaru	1.131031
17	Toyota	1.440886
18	Volkswagen Group of	1.452291
19	Volvo	1.084995
20	# Cyl	1.500957

Figure 6 & 7: (Above)

Figure 6 plots the predicted values against the residuals. To indicate that this model is appropriate, it should appear like a random scattering of points — which it does, indicating roughly constant variability. Plot 7 shows the VIFs of the variables. VIFs over 10 indicate potential multicollinearity; looking at the variables, no VIF is higher than 2. This means both plots support that a multiple linear regression is appropriate to use.

Ridge Regression

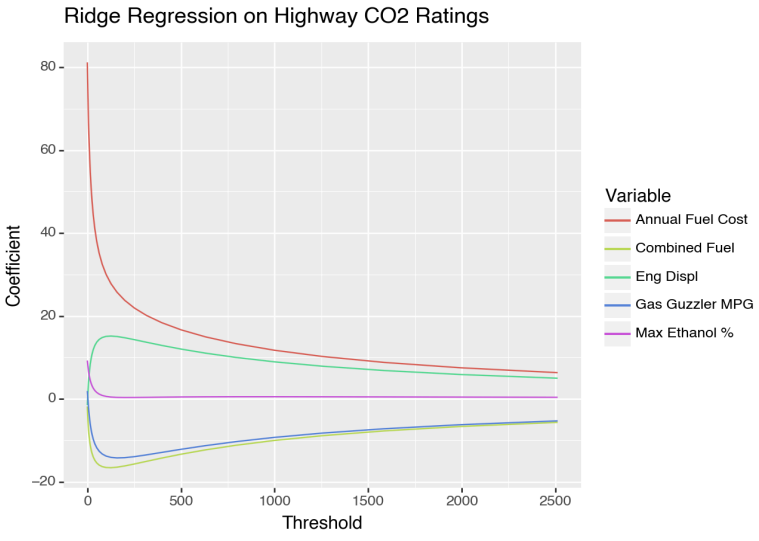


Figure 8: (Left)

Our ridge regression analysis investigated the different factors that might affect a car's highway CO2 Rating. Variables included in the analysis include the annual fuel cost of the car and how much ethanol is allowed in the car's gasoline ('Max Ethanol %'). The above variables were selected from the dataset based on the assumption that they would have a more significant effect on the prediction of a

given car's highway CO2 rating than the other variables in the dataset.

A ridge regression was preferred over a LASSO plot to represent the relationship between the variables on the assumption that the response variable of Highway CO2 Ratings isn't determined largely by a few predictors with substantial coefficients. In other words, exploration of the relationship between the response and predictor variables assumed that predictor variables would have coefficients of roughly equal size.

At very low thresholds, Annual Fuel Cost seems to dominate and have the most influence on the response variable. At around a threshold of 100, Combined Fuel, Gas Guzzler MPG, and Eng Displ reach their highest relative impact on the model, after which they appear to slowly approach zero. Max ethanol appears to have some influence on the model at a threshold of 0, after which it takes on a value close to zero after any threshold is applied.

	0	1	2	3	4	Threshold
0	-1.78	9.28	81.22	1.92	-1.48	0.00001
1	-1.74	9.27	81.22	1.87	-1.48	0.00101
2	-1.71	9.26	81.21	1.83	-1.47	0.00201
3	-1.67	9.24	81.20	1.78	-1.46	0.00301
4	-1.63	9.23	81.20	1.74	-1.46	0.00401
...	...	...	...	...	...	...
495	-0.31	8.06	78.98	-0.32	-0.00	0.49501
496	-0.31	8.05	78.98	-0.32	-0.00	0.49601
497	-0.30	8.05	78.98	-0.33	-0.00	0.49701
498	-0.30	8.05	78.98	-0.33	-0.00	0.49801
499	-0.29	8.05	78.98	-0.33	-0.00	0.49901

Figure 9: (Left)
In this model, we can see that we used many thresholds to find the best alpha value. We've concluded that the best alpha for the model is at a value of **0.01**. At this value of alpha, the coefficients of the model can be written as:

$\widehat{y}$ = **-1.79905404(*Combined Fuel*) + 9.27540493 (*Max Ethanol %*) + 81.19349961 (*Annual Fuel Cost)* + 1.90997184 (*Gas Guzzler MPG)* - 1.46958545 (*Eng Displ)***

<u>kNN Classification</u>

In this kNN classification model, we tried to classify the Highway CO2 ratings (response variable) based on explanatory variables: 'Comb FE (Guide) - Conventional Fuel', 'Max Ethanol % - Gasoline', 'Annual Fuel1 Cost - Conventional Fuel', 'MFR Calculated Gas Guzzler MPG', 'Eng Displ'.

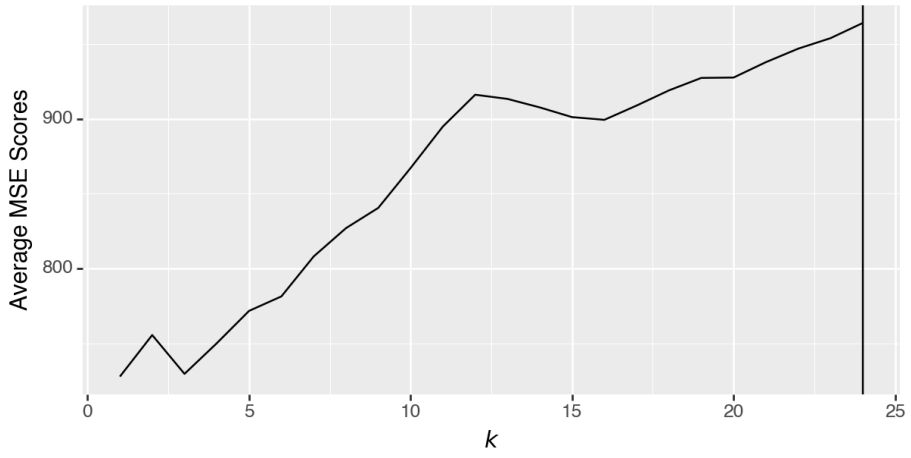


Figure 10

We would choose k = 11 because beyond this value, we can see that beyond this point the graph becomes relatively parallel to the x-axis. Using cross validation, we can select this as the best value. However, this model didn't result in significant results.

Comparison of Model - Which model is the best?

We will compare our models using mean squared error and R^2. When finding the mean squared error, we used cross value scoring for this and received the following results. We tried just comparing the kNN (0.767 MSE) and the Ridge Regression (0.872 MSE) models. Given the residuals for the linear regression model, we chose not to pursue it; however, we did investigate the multiple linear regression. We found that it had an adjusted R-squared of 0.679, and a MSE of 2,097. Although the residuals and R^2 of the model are promising, the extremely high MSE indicates that it is not the best choice. Thus, we concluded that the kNN model would be better because the mean squared error is lowest.

Conclusion of Data Reporting

Influence of Vehicle Attributes on CO2 Emissions: The project reveals that vehicle attributes such as manufacturer, engine displacement, and number of cylinders significantly influence CO2 emissions. The Multiple Linear Regression model, in particular, highlights the varying impacts of different manufacturers and engine characteristics on emissions; for example, the variables of Jaguar Land Rover and Porsche were significant given their high coefficients, indicating that their manufacturers had notable impacts on emissions. This underscores the role of vehicle design and engineering in determining environmental impact. For the Ridge Regression model, at the best value of alpha for the model, 0.01, Annual Fuel Cost far outweighs the other variables in terms of its influence on the response - with Max Ethanol % being the next most important variable.

Appropriateness of Different Models for CO2 Prediction: The project's comparison of different statistical models demonstrates the complexity of accurately predicting CO2 emissions. While the linear regression model showed limitations (indicated by the violation of homoscedasticity), the Multiple Linear Regression and Ridge Regression models provided more nuanced insights. The KNN model, despite its lower mean squared error, didn't yield significantly distinct results. This suggests that more complex models may be better suited for capturing the multifaceted nature of factors affecting CO2 emissions from vehicles.

In conclusion, this comprehensive analysis provides a deeper understanding of the factors influencing vehicle emissions and highlights the importance of selecting appropriate modeling techniques for environmental data analysis. The project also underscores the potential for data-driven approaches to inform policy measures aimed at reducing pollution and enhancing vehicle efficiency. As next steps, we would recommend further investigating the linear regression model, and finding ways to transform the data to eliminate the shape in the residuals.

Bibliography

1. Fuel Economy Dataset - United States Environmental Protection Agency (2023), *Fuel Economy* [Data set] https://www.fueleconomy.gov/feg/download.shtml
2. Smith, A., & Johnson, L. (2021). "Trends in Vehicle Fuel Efficiency: A Comparative Analysis." Journal of Automotive Technology, 34(2), 45-59.
3. Green, M. (2020). "The Impact of CAFE Standards on the Automotive Industry." Environmental Policy Review, 28(3), 112-130.
4. Lee, K., & Patel, S. (2019). "Consumer Behavior in Vehicle Purchases: The Role of Fuel Prices." Economic Insights, 22(4), 54-68.

Work Split

This project was completed by three individuals: Fred Ngo, Disha Mohta, and Chloe Attlan. The description and researcher's review was completed by Disha Mohta as well as the kNN classifier. The Linear Regression and Multiple Linear Regression was completed by Chloe Attlan. Fred Ngo was in charge of the Ridge Regression / Lasso regression but we omitted Lasso as it didn't have significant results.

Appendix

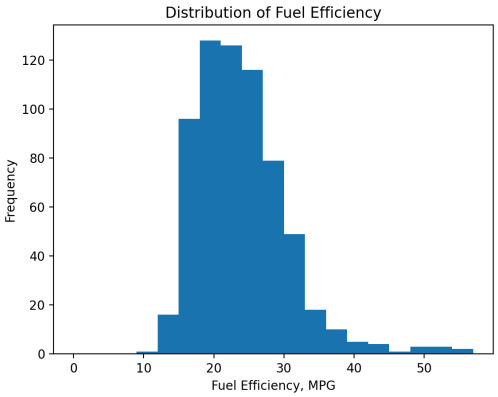


Figure 11. The plot to the left shows the distribution for another variable we investigated, which was fuel efficiency, measured in miles per gallon. The graph seems to be centered around 24, and ranges from just under 10 to around 60. It is unimodal, almost symmetric, and skewed slightly toward the right.

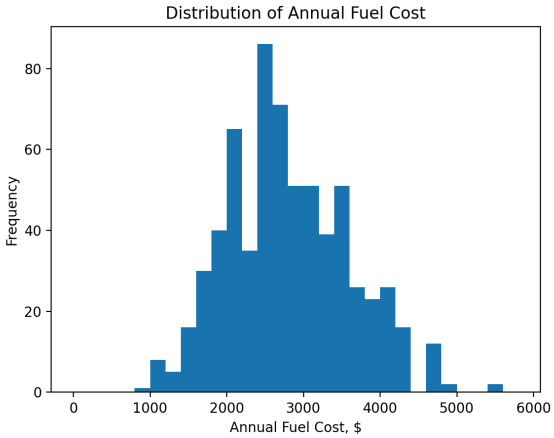


Figure 12. The plot to the left shows the distribution for annual fuel cost, measured in dollars. The graph seems to be centered just under 3,000, and ranges from under 1,000 to around 5,500. It is bimodal, with one peak at around 1,900 and another bigger peak at around 2,500. It is almost symmetric.