Fuel Efficiency through Transmission Design: Hoax or Reality?

Paul J. Motes

Student ID: 001433985

Western Governors University

October 16, 2021

Table of Contents

A. Project Overview

A1. Research Question or Organizational Need

This exploratory data analysis project will help analyze the relationship between vehicle transmission design and vehicle fuel economy. The point of the project is to use Python, statistics, and data visualizations to answer the following question: to *what extent* do the total number of gears in the transmission have on fuel economy for vehicles driven in the United States that were produced from 2020-2021?

The research question comes out of an organizational need for the Legend (fictional) company. Legend has done well in the US market, and they want to increase fuel efficiency and vehicle sales of their most popular model, the Legend Alanty (fictional). Legend needs to see if improving transmission design will improve fuel economy. Legend is looking at ways in which they can meet higher vehicle fuel efficiency standards for their upcoming 2024 Alanty model.

A2. Context and Background

Fuel efficiency has been an important element of vehicle design for many years. Likewise, fuel efficiency numbers have been an important aspect of marketing vehicles to customers. Increasing a vehicle's fuel efficiency does pass significant savings to customers. Most of the discussion about increasing fuel efficiency in the past related to reducing vehicle weight and improving engine efficiency. As fuel efficiency standards continue to increase, all vehicle systems must be explored to reduce fuel consumption and increase fuel efficiency. The main outcome for this exploratory analysis is to explore and describe the relationship(s) between transmission design and vehicle fuel efficiency. Exploring transmission design and fuel efficiency will help Legend in their decision-making process to see if improving transmission design in their most popular model (Alanty) will help meet overall 2024 fuel efficiency standards.

A3. Summary of Published Works

Fuel efficiency in vehicles has been emphasized for many years, especially since the oil embargo in the early 1970's. The United States government implemented the Corporate Average Fuel Economy (CAFE) standards back in 1978. CAFE implemented regulations about how far in miles the vehicle travels on one gallon of fuel (miles per gallon) ("Corporate average fuel economy," n.d.). Just to get an idea, in 1978, the CAFE standard was about 18 miles per gallon(mpg) for all passenger car vehicles. The CAFE standard was set in 2017 to be an average of about 39 mpg for all passenger cars, and about 29.4 mpg for all light trucks ("Corporate average fuel economy," n.d.).

There are many factors that contribute to the understanding of a vehicle's fuel efficiency. Weight is a key factor that has been brought up (Gautam, 2010, p. 2-4). Changes in manufacturing technology and materials engineering have helped reduce weight and improve vehicle fuel efficiency. Gautam wisely points out that fuel efficiency is shaped by a lot of features. Features that are important to fuel efficiency are things like: "[the] type of passenger vehicle, fuel prices, government policies, vehicle weight, and engine power" (Gautam, 2010, p. 10). Features related to fuel efficiency ultimately impact fuel economy (Gautam, 2010, p. 3, 9-10). Gautam clarifies the nuance between fuel economy and fuel efficiency. *Fuel economy* is more about how many miles per gallon your vehicle gets whereas *fuel efficiency* is more related to how well fuel (a specific energy source, such as gasoline) is used to move the vehicle (Gautam 2010, p. 2).

Since fuel efficiency affects the fuel economy of a vehicle, the design of the powertrain system is one of utmost importance. The powertrain system of most vehicles consists of the combustion engine, and transmission/transmission components; although, for at least the past 15 years, there have also been electric and hybrid electric vehicles. The powertrain for full electric or partial electric vehicles will include things like driver motors, electrically powered transmission units, and other carefully engineered mechanical components.

Advancements over the years in engine and transmission design have been valuable in improving overall fuel economy. Within the past ten years, there has been a lot of talk about improving transmissions to improve fuel economy and customer satisfaction (the way the vehicle shifts in different conditions). In a 2018 article from Gears Magazine, the author makes the point that increased fuel economy standards have led to increasing the number of gears in transmission design ("Are More Gears the Answer," 2018). Increasing the number of gears in the transmission decreases the load on the engine. Decreasing the load on the engine helps to optimize engine performance, increase fuel efficiency, and decrease fuel consumption. To prove the point, there is some research data from manufacturers like Toyota, Dodge, and Honda. Data presented in the article indicated that increasing the number of gears can create a change in fuel economy, up to 8% ("Are More Gears the Answer," 2018).

A3a. Relation of Published Works to Project

The Wikipedia article on the "Corporate average fuel economy" gives a background about the importance of fuel economy and how fuel economy standards developed in the late 1970's. Fuel economy standards are an ongoing factor that impacts vehicle produced in the United States. Understanding that fuel economy standards change over time is important for Legend's organizational need.

The article written by Gautam helps illuminate the nuance between "fuel efficiency" and "fuel economy (Gautam, 2010, p. 2-4). Fuel efficiency has many features, one of those being the powertrain system. Since the transmission is a vital piece of the powertrain system, fuel economy and transmission design should be investigated. Understanding that transmission design is a feature that could improve fuel economy connects to both the research question and the organizational need.

The data driven research article "Are More Gears the Answer to Improve MPG?" identifies that car manufacturers are looking to add more gears to their transmission design; data shared in the article shows that many new transmission designs have increased the number of total gears in the transmission, leading up to an 8% increase in fuel economy under certain driving conditions. This article gives data driven reasons to look at the interaction between transmission design and fuel economy.

A4. Summary of Data Analytics Solution

Legend company needs to get a better understanding of transmission design and fuel efficiency in vehicles so that they can meet 2024 fuel efficiency standards. Specifically, they want to see if transmission design improves fuel efficiency in their 2024 Alanty model. I am proposing a solution to this need by performing an exploratory analysis by using Python and a Jupyter Notebook; the analysis will help explore and describe the relationship between transmission design and fuel efficiency. This will provide needed background information so that Legend can make decisions about transmission improvement for their new 2024 Alanty model.

A5. Benefit to Organization and Decision-Making Process

Legend has a research and development (R&D) team that will look at the results of the analysis. Having a detailed analysis about the interaction between transmission design and fuel economy (for all vehicles between 2020-2021) will help the R&D team determine if further plans need to be developed for transmission improvements for the 2024 Alanty. In other words, the analysis provides a baseline to aid their plan for transmission improvements as it relates to Alanty's fuel economy. An improvement plan for transmission development could include planning for capital, supply chain, infrastructure, and manufacturing needs. Proper planning saves the business time, money, and significant resources. Second, this analysis could potentially benefit Legend by minimizing or eliminating fines paid if they do not meet fuel efficiency requirements set out by CAFE and the Environmental Protection Agency (EPA). Even if the analysis does not show any connection between transmission design and fuel economy, Legend will be able to move onto other vehicle features that could help improve overall fuel economy. Legend still benefits from the analysis and saves time in the decision-making process.

B. Data Analytics Plan

B1. Goals, Objectives, and Deliverables

The main goal for this project is to develop an exploratory data analysis in Python using a Jupyter notebook to investigate and describe the relationship between transmission design and vehicle fuel economy. The end game is to give a satisfactory conclusion to the research question. These are the defined objectives:

1. Determine the proper data needed to perform an appropriate exploratory data analysis.
   - Deliverable: produce a cleaned dataset (with the help of Python) from EPA vehicle data from 2020-2021 fuel efficiency testing data. Save the cleaned and merged dataset to a new file. May involve transforming some of the data (as appropriate).
2. Describe the relationship(s) between transmission design features and fuel economy.
   - Deliverable: produce a html (markdown) report in Jupyter notebook that summarizes, illustrates, and give insight into the dataset and makes clear connections to the research question. Use data visualizations as appropriate.
3. Describe how the conclusions for the project report connect to Legend's business needs.
   - Deliverable: connect the conclusions and insights from the project to answer the research question. Answer the research question in the context of Legend's business needs.

B2. Scope of Project

Project scope will include a html (markdown) report developed in Python. The report will provide the process of combing and cleaning the original dataset, organizing, and analyzing the data, and using statistics and visualizations to help analyze, clarify, and explain the relationship(s) between transmission design and vehicle fuel economy. Project scope is *limited* to only look at relevant features and relevant data that will provide insight in exploring transmission design and vehicle fuel economy. The project scope will *not* include investigating other possible features that could help improve fuel economy.

B3. Standard Methodology

The CRISP-DM (Cross Industry Process for Data Mining) methodology for this data analysis project will provide a structured way to complete the project efficiently. This is one of the most popular project methodologies for many data analytic projects. There are six basic steps of the CRISP-DM method.

1. **Business Understanding**: I want to review Legend’s business need, and what parts of the dataset would be good to use as applied to the relationship between transmission design and fuel economy.
2. **Data Understanding**: During this step, I want to perform a comprehensive look at the dataset, and, based off the federal EPA laws (Title 40 CFR Part 600), understand about the meaning behind the column terms in the dataset; for example, when it says FTP or NO_x, what do those terms mean? The meaning of some of the terms is not obvious. I also will investigate and report on data types for each variable (e.g. numerical, discrete, categorical). Understanding the data will help lead to data preparation.
3. **Data preparation**: This step is one of the key parts to the whole project. Here, I will use insights that are taken from business understanding and data understanding. Using these insights, I will be able to clean data, transform data (as appropriate), and get the data wrangled so that an appropriate analysis and modeling can take place.
4. **Modeling**: Here, with the cleaned data, I will spend time reviewing the data, analyzing the data with descriptive statistics, producing relevant visualizations, and using linear regression for modeling data. Emphasis will be given to descriptive statistics to help answer the research question. Modeling using linear regression will be more for additional insight as it pertains to the research question.
5. **Evaluation**: After performing the exploratory data analysis, looking at descriptive statistics / modeling, and making appropriate visualizations, I want to respond to the research question, and make a conclusion on the hypothesis and the null hypothesis.
6. **Deployment**: In the final step, I will take the project report and make a presentation to the board and the R&D team at Legend.

B4. Timeline and Milestones

Milestone	Projected Start Date	Projected End Date	Duration (days/hours)
Business Understanding: Business Needs / Creating strategy plan to address.	10/18/2021	10/18/2021	1 day
Data Understanding and Data Preparation: Cleaning and preparing dataset. Translate appropriate data column abbreviations using EPA federal standards Title 40 CFR Part 600.	10/18/2021	10/19/2021	2 days
Modeling: Analysis and Exploration, modeling as needed.	10/20/2021	10/22/2021	3 days
Evaluation: fine tune visualizations, review report, and make conclusions. Produce final report, connect to business needs.	10/25/2021	10/25/2021	1 day
Deployment: Review report, prepared for presentation; perform 10 minutes presentation to stakeholders/R&D team..	10/26/2021	10/26/2021	2.5 hours.

B5. Resources and Costs

Personnel, technology, and infrastructure needs	Cost in US dollars ($)
Data from EPA for 2020-2021 vehicles	Public Domain - N/A
Anaconda software environment with Python 3.8 and Jupyter Notebook	N/A
Desktop with two 32-inch monitors.	N/A
Data analyst (Paul), 58.5 hours of work	N/A

Since this is a fictional project, the analyst will need 58.5 work hours to: make a project plan, perform the exploratory analysis, make informed conclusions, create a finalized project report, and present the polished report to Legend's stakeholders. The fuel economy testing dataset was from the United States Environmental Protection Agency; the data in the datasets are public domain and not restricted data. All software and hardware needs were provided by the data analyst.

B6. Criteria for Success

For this project, I will use *five* different criteria to measure project success. The first criterion relates to the business understanding and data understanding step in the CRISP-DM project methodology. The second, third, and fourth criteria are related to the evaluation and modeling phase. The fifth criterion is to help finish the project and provide deployment to Legend's stakeholders. See the table below for detailed information on each criterion and the cut score for success.

Criterion/Metric	**Required Data**	**Cut Score for Success**
1. Was the dataset cleaned appropriately? Were any necessary variables established for analysis? Were any missing values in the data addressed? Were any calculated or derived values used, and if so, are they appropriate for the analysis?	Two raw data sets downloaded from EPA website. Output from python showing appropriate changes in the data frame(s).	Success if and only if the data is cleaned, missing values were addressed, and calculated/derived values were included as appropriate.
2. Was the correlation explored between transmission design and fuel economy? Was aggregated and categorical data explored with respect to the research question?	Output of Python data analysis in Jupyter Notebook.	If and only if the notebook explores the correlation between transmission design and fuel economy. Appropriate aggregations and calculate/derived values are explored with respect to the research question.
3. Specifically, was the relationship between the number of gears and fuel economy explored? Was this explored with respect to fuel source type?	Output of python data analysis in Jupyter Notebook.	If and only if there are at least 2 valid points explaining and illustrating the relationship between transmission design and fuel economy. The relationship should show the differences between fuel source categories (combustion engine, full electric, etc.) and fuel economy numbers.

4. Were there at least 3-5 relevant and aesthetically pleasing visualizations that connected with the research question?	Output of python data analysis in Jupyter Notebook.	If and only if there are 3-5 quality visualizations that help illuminate the data as it relates to the research question.
5. Was there a clear conclusion to accept or reject the hypothesis and accept or reject the null hypothesis? Were the conclusions connected to Legend's business need?	Completed exploratory data analysis (html markdown) generated from Jupyter Notebook.	If and only if there are two conclusive points which help stakeholders understand which hypothesis was accepted and which was rejected and the reasons for each conclusion. A clear explanation of the findings of the research was presented. All research points tie to Legend's business needs.

C. Design of Data Analytics Solution

C1. Hypothesis

The *hypothesis* is as follows: the total number of gears in the transmission have a *positive* impact on fuel economy in vehicles driven between 2020 and 2021 in the United States. The *null hypothesis* is as follows: The total number of gears in the transmission have a *neutral* impact on fuel economy in vehicles driven between 2020 and 2021 in the United States.

C2. Analytical Method

Since this project is more about discovering a relationship between transmission design and fuel economy, there is not a strict focus in using predictive methods. I will use descriptive and predictive methods but will mainly focus on descriptive methods due to the inferential nature of the research question. During the exploration and discovery phase, using things like correlation testing (Pearson's) will be helpful. Using descriptive statistics like median, mean, variance, standard deviation, standard error, p-values, confidence interval (95%), minimum values, and maximum values will also be helpful in investigating and comparing data between the 2020 data set and the 2021 data set. Descriptive statistics will provide the weight of the methods that contribute to conclusions made to the research question.

For some additional information and impact, I want to use the linear regression model as it relates to the research question. A linear regression model may help to see if there are any trends in fuel economy. In short, the model will investigate how one discrete predictor variable (number of gears in the transmission) predicts a certain type of change in the continuous response variable (fuel economy). I can also use linear regression modeling with grouped data (data grouped by the number of transmission gears) or ungrouped data (showing all data together without groupings). Using a linear regression model will help to see if there is a positive or negative linear relationship and if there is a strong or weak linear relationship, either between groupings, or overall (without any groupings).

C2a. Justification of Analytical Method

The analytical method is a blend of descriptive statistics and descriptive modeling (linear regression). Descriptive statistics gives evidence to support inferential conclusions related to a hypothesis and null hypothesis. Using descriptive statistics aligns with the point of the research in determining and describing the relationship between transmission design and fuel economy. Using a linear regression model will help stakeholders visually see association between the number of transmission gears and fuel economy.

C3. Tools and Environments of Solution

I will be using Python 3 with a Jupyter notebook to perform: data cleaning, exploratory data analysis, and visualization creation. Using a Jupyter notebook provides a way to track the process of cleaning, exploring, and visualizing the data as it relates to the research question. Overall, this will provide one tool, instead of many, to get the needed results.

C4. Methods and Metrics to Evaluate Statistical Significance

I am going to use two different metrics based on the business needs and the research question. The primary metric to evaluate statistical significance will be by using the p-value with a confidence interval of alpha = 0.05 (95% confidence). Since I have a smaller dataset, there may not be enough data to push the confidence interval to 99%. I will look at descriptive statistics related to the research question, especially correlation values, p-values, mean, median, minimum, maximum, standard deviation, standard error, and variance. If the p-value < alpha, then I can reject the null hypothesis accept the alternative (which is the hypothesis).

The second metric, which is more for description than predictive power, is the following: if the fit of the linear regression model shows that the slope coefficient is very close to zero, then evidence can point towards accepting the null hypothesis. On the other hand, if the linear regression model shows that the slope coefficient is either negative or positive (and significantly above zero), then the null hypothesis will be rejected.

C4a. Justification of Methods and Metrics

Both descriptive statistics and descriptive modeling (linear regression) will give a specific way to provide evidence which supports answering the research question. Using both methods together will help make a reasonable conclusion on the hypothesis and null hypothesis. The two metrics described above (using a 95% confidence interval with p-values and using slope coefficients for linear regression models) help give meaning and validity to the conclusions made about the research question.

C5. Practical Significance

In our scenario, if the hypothesis is validated, then the number of transmission gears positively effects vehicle fuel economy. The practical significance relates to Legend’s business need. They want to know if improving their transmission design will increase fuel economy to meet EPA fuel economy standards in 2024 for the Alanty model. For example, if they only need a 2% increase in fuel economy, and the data shows that overall fuel economy with transmission design goes up by 3% every year between 2020 and 2021, then Legend would move to the next stage of research and development for transmission improvement. If the data analysis shows that, for example, a 12-speed transmission provides a 9% fuel economy improvement over other transmission designs, Legend will consider what it would take to implement a transmission improvement in their 2024 Alanty. Improvements in transmission design would practically yield an increase in fuel economy.

C6. Visual Communication

There will be two ways to communicate the research findings. First, I want to use a distribution curve to illustrate the range of fuel economy values. There will likely be more than one curve to illustrate differences in data due to different categories for things like fuel source and vehicle brands. Any of these distribution curves may need a curve transformation to put the curve into a normal distribution; a normal distribution is needed to perform a proper p-value statistical test. Using a p-value test, and graphically illustrating this with distribution curves will help underscore the statistical significance in relation to the research question. Explaining the p-value test will help give meaning to what type of inferences are being made (e.g., accepting or rejecting null hypothesis). Second, using the linear regression model, I will use scatter plots with a best fit line to show how each linear regression was processed; a regression line can be grouped by each transmission type (e.g., 6 speed, 8 speed, 10 speed) or can be ungrouped (leaving all fuel economy data points ungrouped by transmission type). Visually this will show if transmission gears have an association with fuel economy and which groups may have stronger or weaker associations. Using a visualization here is better than discussing linear equations and coefficients of the independent variable.

D. Description of Datasets

D1. Source of Data

The data source for this project will be two datasets that came from the United States EPA website. The first dataset is 2020 Test Car List Data (U.S. Environmental Protection Agency, 2020). The second dataset is the 2021 Test Car List Data (U.S. Environmental Protection Agency, 2021). Both data sets contain real world data from the EPA testing center in Michigan.

D2. Appropriateness of Dataset

The dataset is functional and gives relevance to explore the research question, which centers around understanding the relationship between transmission design and fuel economy. Since Legend will be working on design upgrades for their 2024 Alanty model, they need recent data to see if they want to put time, money, and research into transmission improvements. To this end, using recent data from the EPA helps give validity to develop a baseline for looking into the research question.

D3. Data Collection Methods

First, the vehicle data was collected by the EPA and was completed at a testing facility in Michigan. There are a few preliminary things to know related to the dataset. First, there are many regulations that are used to run these vehicle fuel efficiency and environmental tests that the EPA data reports. The EPA federal regulations go into depth about many details of testing vehicles for fuel economy and environmental output tests (e.g., measuring carbon dioxide, and other pollutants). Second, the scope of all the different definitions, processes, and rules for EPA fuel economy testing are way outside the scope of this project. However, I will be using the EPA regulations, specifically Title 40 CFR 600, to get a better understanding about what column terms mean in the datasets (“ECFR :: 40 CFR Part 600”, 2021). One example would be to look up what FTP means in the dataset.

The main advantage of using EPA fuel economy data is that vehicle testing is in controlled testing environment. Having a controlled testing environment helps the data to be as complete as possible and as accurate as possible. The data in a controlled testing environment is more accurate because the same set of conditions are applied across every single vehicle that was tested. There are some other factors that could have an impact on vehicle testing, such as ambient outside temperature, oil temperature, transmission oil temperature, and other metrics. However, those metrics are not included in the data set so no conclusions can be made on those features because the data is not included.

The challenge of a controlled environment means that data may not be as fluid. In real world driving, parameters can change many times in inconsistent ways, which will affect data results. So, we really don’t know how vehicles will respond when the environment has real world constraints (as compared to a controlled testing environment). If, for example, there was both controlled testing, and long-term real-world testing, then comparing the two sets of data could be explored.

D4. Data Quality

Even with the technology used for capturing fuel economy/vehicle data by the EPA, I will still need to check data for formatting and quality. There may be missing data, but it may be for valid reasons. For example, some fuel economy training methods may not have certain parameters. If there is missing data that needs to be adjusted (or ignored), that will need to be dealt with in the Jupyter notebook. I will also check if there are any outliers, and if so, make an informed decision if removing them is needed for data quality, accuracy, or modeling purposes. Data from the vehicle testing dataset should be fairly accurate, since the EPA had a controlled environment and assumably had quality checks before publishing data online for the public to view.

D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulatory Compliance

The EPA datasets did not use any private data, such as Vehicle Identification Numbers (VIN) but rather the EPA generated a specific ID number for each vehicle tested (which was probably randomly generated based off the vehicle model). There are no concerns for sensitive data. The EPA published these datasets as public domain and anyone who has access to the internet can view this data. Since this is public data, there are no major concerns related to ethics, law, privacy, or other industry specific regulations.

D5a. Precautions

Data in the datasets are public data and can be freely disseminated. The cleaning, modification, and analysis of data is done for this project only and is not connected to any government agency. Data preparation, analysis, storage, access, and dissemination is for the purpose of this data analysis project only. Data for this project will be stored privately and will be given only to identified parties who work with Legend (e.g., stakeholders, R&D team).

E. Sources

Are More Gears the Answer to Improve MPG? (2018, May 7). Gears Magazine. Retrieved October 13, 2021, from https://gearsmagazine.com/magazine/gears-answer-improve-mpg/

Corporate average fuel economy. (2021, October 2). In Wikipedia. Retrieved October 6, 2021, from https://en.wikipedia.org/w/index.php?title=Corporate_average_fuel_economy&oldid=1047714517

ECFR :: 40 CFR Part 600 -- Fuel economy and greenhouse gas exhaust emissions of motor vehicles. (2021). eCFR :: Home. Retrieved October 16, 2021, from https://www.ecfr.gov/current/title-40/chapter-I/subchapter-Q/part-600

Gautam, Suman (2010). What Factors Affect Average Fuel Economy of US Passenger Vehicles? *Honors Projects, 104*. Retrieved October 6, 2021, from https://digitalcommons.iwu.edu/econ_honproj/104

U.S. Environmental Protection Agency (2021). *2020 Test Car List Data* [Data Set]. U.S. Environmental Protection Agency, Office of Transportation Air Quality, National Vehicle and Fuel Emissions Laboratory. https://www.epa.gov/sites/default/files/2021-03/20tstcar-2021-03-02.xlsx

U.S. Environmental Protection Agency (2021). *2021 Test Car List Data* [Data Set]. U.S. Environmental Protection Agency, Office of Transportation Air Quality, National Vehicle and Fuel Emissions Laboratory. https://www.epa.gov/system/files/documents/2021-08/21tstcar-2021-08-25.xlsx