Fuel Efficiency through Transmission Design: Hoax or Reality?

Paul J. Motes

Western Governors University

Table of Contents

Project Overview

A. Project Highlights

A1. Research Question

The research question for this project is to answer the following question: To *what extent* do the total number of gears in the transmission have on fuel economy (FE) for vehicles driven in the United States that were produced from 2020-2021? I used Python, descriptive statistics, modeling, and data visualizations to answer the research question.

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. Project Scope

The broader scope of the project was to create a report (using Python) to clean, describe and explore data as it relates to transmission design and FE. This report explored and analyzed two EPA vehicle data sets to make conclusions about the research question. Project scope is limited to look at relevant features and relevant data that will provide insight in exploring transmission design and vehicle FE. The project scope *will not* include investigating other possible features that could help improve vehicle fuel economy. Please see additional constraints given in section E3.

A3. Solution Overview – Tools

The main tool that I used in the project was a Jupyter Notebook running Python 3. I loaded two CSV (comma separated value) data files into Jupyter notebook, performed relevant data cleaning, analysis, descriptive statistics, modeling, and visualizations. After the notebook was completed, I exported the notebook into a web page so that the report can be accessed easily.

A4. Solution Overview – Methodologies

The project methodology I used was the CRISP-DM (Cross Industry Process for Data Mining). CRISP-DM is one of the most popular methods in many data analytic projects. The basic steps of the CRISP-DM are: (1) Business Understanding, (2) Data Understanding, (3) Data preparation, (4) Modeling, (5) Evaluation, and (6) Deployment. Each step has an appropriate role as the project goes from the planning stage all the way to the deployment of the project. Please see section B2. to see how the CRISP-DM steps were applied.

Project Plan

B. Project Execution

B1. Project Plan

The main goal for this project is to develop a data analysis to investigate and describe the relationship between transmission design and vehicle FE. The project plan was followed, as per the project proposal. No changes were made. All goals and subsequent deliverables were completed as per the project proposal.

Objectives with pertinent deliverables are as follows:

- 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).
- Describe the relationship(s) between transmission design features and fuel economy.
    - Deliverable: produce a html (markdown) report in Jupyter notebook that summarizes, illustrates, and gives insight into the dataset and makes clear connections to the research question. Use data visualizations as needed.
- 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. Project Planning Methodology

The methodology for this project was the CRISP-DM. The basic steps of the CRISP-DM are: (1) Business Understanding, (2) Data Understanding, (3) Data preparation, (4) Modeling, (5) Evaluation, and (6) Deployment. I chose the CRISP-DM because it provides a structured way to complete this data analysis project.

Business Understanding: I reviewed what parts of the dataset would be good to use as applied to the relationship between transmission design and fuel economy. I analyzed and determined the scope of the project and listed the appropriate resources that were needed.

Data Understanding: I performed a comprehensive look at the dataset visually. I also explored and read some of the EPA laws to understand some of the column abbreviations (Title 40 CFR Part 600) ("ECFR :: 40 CFR Part 600", 2021). I was able to find out the meaning of some of the test abbreviations (such as FTP, SC06). Some of the other terms in the data set I was not able to find the meaning. I also investigated and reported on the data types (e.g., numerical, categorical) of the columns in the combined dataset. Understanding the data helped lead to data preparation. Please note that since the scope of the project focuses on transmission design and fuel economy, data that was not within scope was removed.

Data Preparation: I used insights taken from business needs, and data understanding. I used data cleaning to do things like remove rows that did not show values for adjusted fuel economy or remove records that have outlier concerns related to the adjusted fuel economy

(e.g., values like 4000 mpg). Using those types of insights, I was able to perform data wrangling such that an appropriate analysis can take place.

Modeling: I spent time reviewing the cleaned data, analyzing the data with descriptive statistics and modeling, producing relevant visualizations, and using linear regression for modeling data. I gave emphasis in using descriptive statistics to help answer the research question. I also used linear regression and a one-way ANOVA test to help give additional insights as it pertains to the research question.

Evaluation: After I performed the exploratory data analysis, looked at descriptive statistics, looked at the linear regression modeling, and made appropriate visualizations, I responded to the research question; I also made a conclusion on the hypothesis and the null hypothesis.

Deployment: In the final step, I took the completed html project report and made a presentation to the board and the R&D team at Legend (fictional) on 10/30/21.

B3. Project Timeline and Milestones

The project timeline followed the schedule as outlined in the proposal, but the dates were adjusted. The milestones based off the steps of the CRISP-DM were followed and were completed as per the proposal.

Please note: a bracket like [ ] shows the actual date or duration of the project.

Milestone	Projected Start Date	Projected End Date	Duration (days/hours)
Business Understanding: Business Needs / Creating 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 [Date should be 10/19/21]	10/19/2021 [10/20/21]	2 days
Modeling: Analysis and Exploration, modeling as needed.	10/20/2021 [10/25/21, other project needs postponed start date]	10/22/2021 [10/28/21]	3 days [4 days]
Evaluation: fine tune visualizations, review report, and make conclusions. Produce final report, connect to business needs.	10/25/2021 [10/29/21]	10/25/2021 [10/29/21]	1 day

Deployment: Review report, prepared for presentation; perform 10 minutes presentation to stakeholders/R&D team..	10/26/2021 [10/31/21, presentation had to be done by 11/1/21, presentation done early]	10/26/2021 [10/31/21]	2.5 hours. [Took 4 hours]

The timeframe of the project was *altered* as follows:

- The duration of the modeling phase took 4 days and exceeded the initial expectation of 3 days. This was due to challenges in cleaning and adjusting the real-world EPA dataset. The modeling phase was also postponed due to other project needs.
- The evaluation phase exceeded the 1 day expected and took two days. This was due to the evidence that was collected in the report and some coding difficulties.
- Reviewing the report and making a presentation for Legend took 4 hours instead of allocated 2.5 hours. Needed more time to make presentation slides that incorporated visualizations and some screenshots of python code. Still met goal of presenting the final report to Legend by 11/1/21.

Methodology

C. Data Collection Process

- Actual data selection vs. planned collection process
  The data sets I used had vehicle data that was compiled and collected by the EPA and was completed at a testing facility in Michigan. I am sure that planning all these tests for hundreds of vehicles and collating all the data was quite a challenge. The other challenge with data collection is that the EPA had many rules and regulations that govern these vehicle tests. EPA federal regulations go into great depth about all aspects of testing vehicles for fuel economy and environmental output (e.g., measuring things like carbon dioxide). The scope of all the different definitions, processes, and rules for EPA fuel economy testing are way outside the scope of this project.

- Obstacles to data collection
  I suspect that the EPA had minor obstacles to data collection that were encountered. However, this does assume that the EPA followed all rules / regulation in performing tests and recording data. Some of the data in the dataset did lead to major outliers; perhaps some of the data was inputted by hand. Data entry by hand could be a possible obstacle to accurate data collection.

- Unplanned data governance handling
  Since the EPA vehicle data was public data, there really were not any data governance concerns. Data that is open to the public can be distributed and used freely. For example, the EPA dataset avoided data that could be considered private, such as specific Vehicle Identification Numbers (VIN). Instead of using a VIN number, the EPA generated vehicle specific identification numbers. If no entity is selling this public data to others, there are no major ethical, legal, privacy, or industry specific concerns.

C1. Advantages and Limitations of Data Set

The main advantage of using EPA fuel economy data is that vehicle testing was performed in controlled testing environment. Having a controlled testing environment helps the data to be as complete 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 limit vehicle testing data, such as outside ambient temperature, oil temperature, transmission oil temperature, and other metrics. However, those metrics are *not included* in the EPA data set so no conclusions can be made on those features because the data is not included.

There are two limitations to the data set. First, 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. Second, having an understanding about what all the column abbreviations mean is a challenge. For example, what does a STP or US06 test even mean? I had to use the EPA regulations, specifically Title 40 CFR 600, to get a better understanding about what abbreviations in column terms mean in the datasets ("ECFR :: 40 CFR Part 600", 2021).

D. Data Extraction and Preparation Processes

Data was created by the EPA. The data sets were first converted from Excel files to comma separated files (CSV). First, I performed a visual check of both CSV dataset files. After converting and inspecting the CSV files, they were combined and then imported into a Jupyter Notebook using Python. Next, I removed most of the columns in the datasets and kept only the necessary features. To complete the data preparation process, I renamed 2 columns to clearly identify the two main variables in this study, "Number Gears" and "Adjusted Fuel Economy."

E. Data Analysis Process

E1. Data Analysis Methods

In the project, I used both descriptive and predictive techniques. I used descriptive statistical techniques such as analyzing mean, standard deviation, quartiles to help give meaning to the data as it relates to transmission design and FE. Using descriptive techniques helps understand the data as it relates to the research question. Using the predictive power of linear and polynomial regression modeling helps to give additional insights as it relates to transmission design and FE. Both methods help to investigate and understand the relationship between transmission design and FE.

E2. Advantages and Limitations of Tools/Techniques

I will be using Python 3 with a Jupyter notebook to perform data cleaning, exploratory data analysis, and visualization creation. Using things like data frames (pandas) and NumPy arrays help to perform most of the steps outlined in the CRISP-DM project method. Python has

some helpful ways to perform visualizations using matplotlib and other built in python functions. Using python in a Jupyter notebook provides an all-in-one tool to help work through the data analysis process.

As mentioned above, the primary techniques of analysis were descriptive statistics and regression modeling (linear and polynomial). Using these techniques gave tools to investigate the relationship between transmission design and FE. However, these techniques are limited by the quantity and quality of data that is used. The dataset used followed the normal distribution (at least in general). However, if there was ten times or one hundred times more data, the form and features of the data could be vastly different.

E3. Application of Analytical Methods

This project applied analytical methods in the following ways to help analyze the data:

1. Perform univariate exploration
    a. Visually check all features (variables) in the dataset.
    b. Investigate adjusted fuel economy data and clean data due to:
        i. Constraints:
            1. Data to be analyzed is only from 2020 and 2021 EPA data. Verified that all data pertained to only 2020 and 2021 vehicles.
            2. Any fuel economy (FE) values that are non-numeric will be deleted. Most values that were non-numeric were full electric vehicles and some of those did not get tested for some reason. Verified by using python function dropna as applied to data frame to remove FE values that were non-numeric.
            3. Limit data to vehicles that are run off either gasoline, diesel, or partial hybrid. Partial hybrid can use a combination of combustion engine and high-capacity batteries. Visually inspected data and verified by using boolean masking to keep all records except vehicles that were full electric (vehicle type was “Electricity” in the dataset).
            4. The analysis will have a strict focus on the relationship between transmission design (number of gears in transmission) and adjusted FE. Verified by content of methods in cleaning and univariate/bivariate exploration phases.
            5. After the initial data cleaning, all reasonable outliers will be kept in data set to allow for real-world data analysis. Verified by output of boxplot for gears versus FE.
        ii. Outlier detection: visual inspection of dataset helped to give guidelines to remove extreme outliers or values that were erroneous (e.g., 10,000 miles per gallon (mpg)). Any values that are above 84.5mpg and below 8mpg were removed. Used boolean masking to filter values that were outside of the range [8mpg, 84.5mpg].
    c. Used a histogram as an analysis tool to see FE distribution as right skewed; logarithmic transformation puts FE into normal distribution.
    d. Plot all FE data on scatter plot to see where most of the data lies (between 20 and 50 miles per gallon (mpg).

2. Perform bivariate exploration:
    a. Checked descriptive statistics of mean, standard deviation, and quartiles for gears versus FE data.
    b. Visualized average FE across all transmission designs using a bar plot. Found that 1 speed CVT transmission had the highest average.
    c. Explored any differences between fuel types and FE. Found that there were about 315 vehicles that were diesel based and that the FE values had more variance than the FE values for gasoline/gasoline blend-based vehicles.
    d. Grouped FE data by transmission design (gears). Visualized each category of transmission design by plotting density lines. Found that most distribution lines for each category of transmission design generally followed a normal distribution.
    e. Checked and analyzed the following relationships: Vehicle type versus FE, vehicle make versus FE, fuel types (diesel and gasoline) versus FE.
    f. Created a correlation matrix to get initial impressions. Found that FE was negatively correlated with the number of gears in the transmission design.
    g. Created boxplot to investigate which gears had FE outliers. Found that the 6 speed and 8-speed transmission had more outliers than any of the other speeds. The rest of the speeds had very minimal outliers.
3. Performed Modeling:
    a. Trained a linear regression model and using a scatter plot, plotted the linear equation to represent the trend. Found a general linear equation of approximately $y = -2.01x + 46.22$ (in the form $y = mx + b$ where m is the slope coefficient, and b is the y intercept value when x is equal to zero). The negative slope coefficient shows a general sense that as the number of gears goes up, the average FE goes down.
    b. Using another python model, created an ordinary least squares model. I got the same results. Note also that the p value was 0.0. R squared was about 0.219.
    c. Used the ANOVA Kruskal Wallis Test to check analysis of variance across FE as grouped by transmission design (number of gears). After running this test, p value was 0.0.
    d. Performed polynomial line regression analysis
        i. Used a polynomial line modeling tool in Python to see if there were any other curves that could fit the data better. Found a 4^{th} order polynomial in the form of $y = ax^1 + bx^2 + cx^3 + dx^4$ (where a, b, c, and d are the coefficients of independent variables).
        ii. R squared went up to 0.23, only slightly better than linear regression/OLS. Chose to visualize with a scatter plot and a polynomial line graph.

Results

F. Project Success

F1a. Conclusions about Research

The research question is 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 null hypothesis was given as: 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. The hypothesis (or alternative hypothesis) was given as: the total number of gears in the transmission have a[n] ~~positive~~ impact on fuel economy in vehicles driven between 2020 and 2021 in the United States. The alternative hypothesis needs to be *revised* to state that transmission design has an impact on FE and not give a direction of impact. This allows the research to help illuminate what direction the impact of transmission design may have on FE.

F1b. Statistical Significance

As given in the proposal, I laid out two metrics to evaluate statistical significance. The primary metric given was the p-value with a confidence interval of 95%. In this situation, alpha is 0.05. If the p-value is less than alpha (0.05), then reject the null hypothesis. If the p-value is greater than alpha (0.05), then accept the null hypothesis.

Metric 1: I evaluated the p-value in several different ways. I used the model of linear regression / ordinary least squares and the one-way ANOVA test (Kruskal-Wallis Test). I found that p-value for all tests were 0.0. The conclusion is to reject null hypothesis because the *p-value of FE distribution (0.00) < alpha (0.05).* This confirmed that transmission design does have an impact on FE.

Metric 2: Using the linear regression model in python stats module, the linear equation for the data was -2.01x + 46.22 (in the form y = mx + b where m is the slope coefficient, and b is the y intercept value when x is set to zero). Since the slope coefficient is not zero and a negative number, the metric suggests the conclusion of rejecting the null hypothesis. Thus, transmission design has an impact on FE.

F2. Practical Significance

In this data analysis, I found that increasing gears in transmission design had a negative impact on FE. However, the popularity of 1 speed and 8 speed transmissions does influence the general trend discovered. CVT transmissions (1 speed) are lighter, more efficient, and cost less to produce than most of the other transmission designs. 8 speed designed transmissions seem to be the most popular type of transmission. 8 speeds do yield stable FE results. As FE values need to increase, as per federal regulation, it will be interesting to see how vehicle manufactures react with their transmission designs. In terms of practicality, average FE ratings were approximately 25% higher for the 1 speed CVT transmission than any other transmission design. See chart below:

	Adjusted Fuel Economy		
	count	mean	std
Number Gears			
1	1374.0	45.124090	12.587355
4	2.0	29.600000	8.909545
5	60.0	35.631667	13.724091
6	1474.0	33.538331	12.284610
7	1003.0	28.933699	10.493684
8	2636.0	30.288505	9.453875
9	721.0	30.815811	7.980535
10	969.0	26.922807	8.373606

Practically, 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. Currently, the Alanty has a 6-speed transmission design and gets 29.5 mpg (fictional).

CVT transmission are designed for passenger cars, some SUV, (and very few trucks due to towing needs). Since the Alanty is a lighter passenger car, Legend and their research team should investigate implementing a CVT or 8 speed designed transmission. This research is practical considering the Alanty is their most popular selling passenger vehicle. The engineering of the Alanty needs further refinement to meet increased FE values and customer satisfaction goals.

F3. Overall Success

Overall, the project was successful. All five criteria were met *with some modifications to criterion 3 and 5*.

Criterion 1: The data was cleaned, missing values were addressed, and no calculated values were needed. Data was grouped by the type of transmission design.
Criterion 2: Jupyter notebook did explore the correlation between transmission design and FE. The FE data was aggregated by category (transmission design). Correlation between transmission design and FE was thoroughly explored
Criterion 3: There was at least 2 points explaining the relationship between transmission design and FE. Due to problems with FE values for full electric vehicles, I chose to exclude those from the research. I did explore the correlation between transmission design in relation to both gasoline and diesel fuel types.
Criterion 4: There was at least 3 quality visualizations that illustrate and illuminate the dataset as it relates to the research question. Please see the html markdown file for sufficient evidence.
Criterion 5: There were two clear points which help understand which hypothesis was rejected and which hypothesis was accepted. Didn't quite meet this goal because the alternative hypothesis needed to be modified in the language to say that transmission design impacted FE rather than transmission design had a positive impact on FE.

The data analysis did provide a reasonable solution to the research question and exposes some evidence related to the interaction between transmission design and FE. I am hoping

that Legend corporation will look further into developing a 1 speed CVT or 8 speed transmission for their new Alanty model.

G. Key Takeaways

G1. Summary of Conclusions

This data analysis project creates a Jupyter Notebook to explore the relationship between transmission design and FE. There were two main metrics used to clearly communicate the research findings.

Metric 1: I evaluated the p-value in several different ways. I used the model of linear regression / ordinary least squares. I found that p-value was 0.0. This confirmed that transmission design does have an impact on FE. R squared value using linear regression / OLS was at 0.219. That means that the OLS model explains about 22% of the variability of FE as related to the transmission design (gearing). In other words, the strength of the relationship between FE and gear is about 22%. However, in this model, the mean of residual values was close to zero ($6.78 * 10^{-16}$). That means that the OLS model still has some validity to show that as transmission design (gearing) increases, FE decreases. Second, I used the Kruskal-Wallis Test which analyzes variance across different categorical groups of values. This is a classic one-way ANOVA test. The main assumption for this test is that the samples were taken from the same dataset (or data distribution). Since FE could be grouped by transmission design, this was a good test to run. Again, comparing the different distributions (grouped by design) yielded a p-value of 0.0. Just to verify that p-value wasn't just 0 for all possible groupings (and to validate), I compared 1 speed and 4 speed groups and got a p-value of 0.07. I compared FE with respect to the groups of 4 speed, 7 speed, and 10 speed and got p-value of 0.0009. The last method I used was a polynomial line mapping tool in Python. I found a 4^{th} order polynomial that fit the general idea of the data. R squared was 0.23, which was only slightly better than the linear regression/OLS model. See below for descriptive statistics summary and the polynomial line map to the FE data.

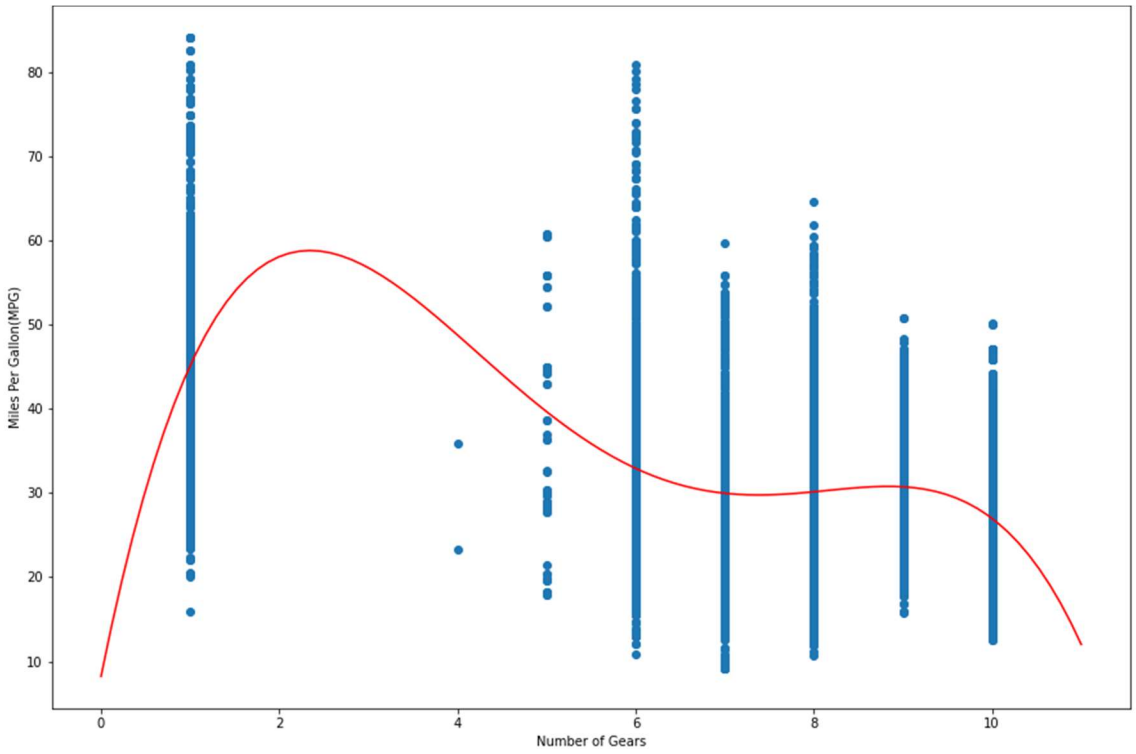
Method	P-Value	Alpha	P-value < Alpha	R squared
Linear Regression (sklearn ML modeling)	0.0	0.05	True	0.219
OLS (stats module)	0.0	0.05	True	0.219
Kruskal-Wallis (ANOVA)	0.0	0.05	True	N/A
4^{th} order polynomial	N/A	0.05	N/A	0.230

OLS Regression Results

Dep. Variable:	y	R-squared:	0.219
Model:	OLS	Adj. R-squared:	0.218
Method:	Least Squares	F-statistic:	2304.
Date:	Wed, 27 Oct 2021	Prob (F-statistic):	0.00
Time:	20:10:00	Log-Likelihood:	-31155.
No. Observations:	8239	AIC:	6.231e+04
Df Residuals:	8237	BIC:	6.233e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	46.2248	0.302	153.128	0.000	45.633	46.817
x1	-2.0076	0.042	-47.997	0.000	-2.090	-1.926

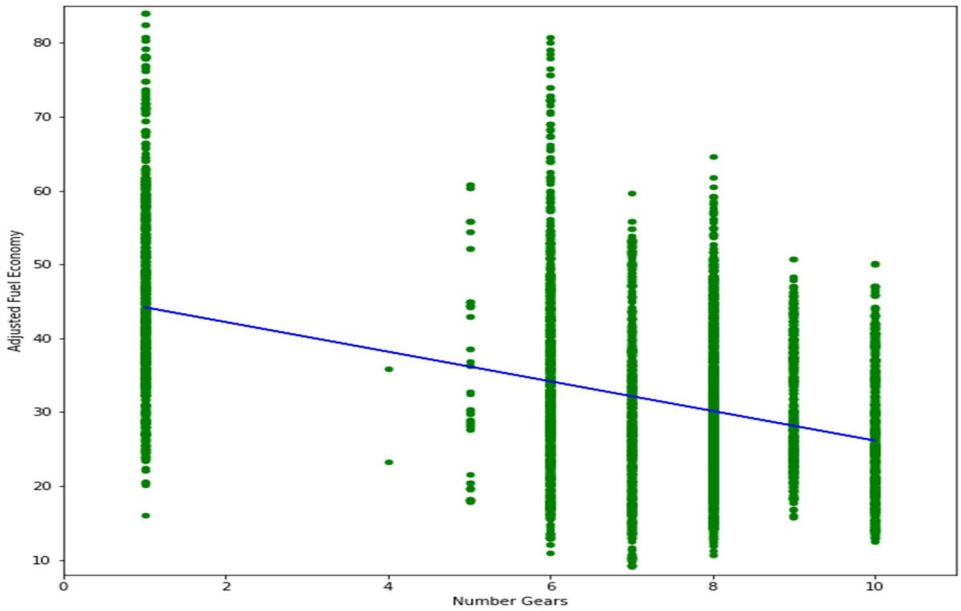
Omnibus:	430.729	Durbin-Watson:	1.256
Prob(Omnibus):	0.000	Jarque-Bera (JB):	501.534
Skew:	0.580	Prob(JB):	1.24e-109
Kurtosis:	3.338	Cond. No.	18.9



Meaning: Conclusion is to reject null hypothesis because the p-value of FE distribution (0.00) < alpha (0.05). Having a p-value below the 95% confidence interval (alpha = 0.05)

means that the dataset does *not* suggest that there is neutral impact between transmission design and FE.

Metric 2: 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.



Using the linear regression model in python stats module, the linear equation for the data was -2.01x + 46.22 (in the form y = mx + b where m is the slope coefficient, and b is the y intercept value when x is equal to zero).

Meaning: Slope coefficient approaches 0 or is close to zero, accept null hypothesis. However, visually, and with the help of linear regression, see that the slope coefficient is approximately -2.01. Since the slope coefficient is not close to zero. As the number of gears increase in transmission design, the association to FE decreases rather than increases. Reject the Null Hypothesis.

G1b. Key Conclusion

Based on the data and the completed metrics, I concluded to accept a modified version of the alternative hypothesis that states: the total number of gears in the transmission has an impact on fuel economy in vehicles driven between 2020 and 2021 in the United States.

G2. Effective Storytelling

The tools and visualizations used in this analysis were an effective way to present conclusions as related to answering the research question. These methods helped achieve all five of the objectives laid out in the project proposal (although a few were slightly modified). Using a chart showing p-values was the best way to present the confidence interval test. Showing a scatter plot with linear regression line is the best visual representation to show how transmission design (gearing) relates to FE.

G3. Findings-based Recommendations

Just to reiterate, the research question was: 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?

From the research presented, the data shows that increasing transmission design (in terms of gearing) does not increase FE. The research did show that overall increasing transmission design (in gearing) reduces FE. There were two other important findings. First, the 1 speed CVT transmission provides some of the best FE numbers, especially for passenger vehicles. Second, the 8-speed transmission was the most popular type of transmission that was tested in the dataset and had the most stable FE numbers.

Since increasing the gearing in the transmission design does not guarantee increased FE numbers, Legend should explore using the CVT (1 speed) in their popular Alanty model. Performing further analysis would be beneficial to see if a CVT transmission would be beneficial for FE.

H. Panopto Presentation

Link for Panopto Presentation:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=0753728c-9dab-4d84-9bc1-add2013610fd

Appendices

I. Evidence of Completion

Compressed file with all following files (.zip)

- Excel Original EPA dataset files. (2 files)
- CSV excel files for EPA dataset for 2020-2021 vehicles. (2 files)
- CSV excel file that has the combined 2020-2021 EPA datasets together. (1 file)
- Jupyter Notebook file (.ipynb). (1 file)
- Polished data analysis notebook as html markdown file. This shows all output of commands and appropriate visualizations. (1 file)
- Powerpoint file for Panopto Presentation (.pptx) (1 file)

Sources

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