Data Science Final Paper

CHARLIE ANGEVINE, Marquette University

A demonstration of rudimentary data science skills including data cleaning, visualisation, and making inferences/drawing conclusions. Work based on public dataset "Vehicle Fuel Economy Estimates, 1984-2017", published to kaggle.com by the US Environmental Protection agency. The process of finding, cleaning, and visualizing the data will be reviewed as well as methods of visualizing and drawing conclusions from the data. I will then discuss my findings and shed light on their significance in the context of past, present and future fuel economy.

ACM Reference Format:
Charlie Angevine. 2018. Data Science Final Paper. 1, 1 (December 2018), 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The dataset examined in this paper is entitled "Vehicle Fuel Economy Estimates, 1984-2017". It contains quantitative and qualitative information about nearly every car manufactured between 1984 and 2017, especially those commonly used in the United States. Some of the more relevant facets of the data included a car's make, model, year, fuel type, information about the engine, and miles per gallon for city, highway, and the two combined. Based on the information provided, visualisations were created in order to examine trends in fuel economy over time and compare the relative performance of different types of cars.

The most important variables in the dataset (for my own purposes) include the Year, Class, Engine Displacement, and combined MPG of each value. The visualisations and conclusions that I made had to do with examining engine displacement in relation to MPG, then evaluating trends in average MPG of common vehicles over the time frame covered by the dataset.

2 METHODS REVIEW/EXPLANATION

One key variable in the dataset that proved to be particularly useful was the engine displacement. I chose to examine this one more closely because it can vary quite a bit in common passenger vehicles and I was interested to see what kinds of affects it could have on fuel economy. This variable, in combination with the class of vehicle, the year, and the combined MPG gave me most of the information necessary to create my visualisations and begin examining them.

2.1 Engine Displacement Background

Engine displacement is the combined volume of all the pistons in every cylinder of a piston engine, not including the volume of combustion chambers. It affects the amount of fuel that an engine is able to take in as well as how much power it is able to put out. Higher displacement engines can

Author's address: Charlie Angevine, charles.angevine@marquette.edu, Marquette University, 1313 W. Wisconsin Ave, Milwaukee, Wisconsin, 53233.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Association for Computing Machinery.

XXXX-XXXX/2018/12-ART $15.00
https://doi.org/10.1145/1122445.1122456

produce significantly more torque and higher horsepower, but tend to burn a lot more fuel in the process. Engines with smaller displacements are simply not able to take in as much fuel, so their power output is lower in exchange for a higher number of miles to the gallon.

Engine displacement is also usually an indicator of how big an engine is overall, and so the displacement of a car's OEM (Original Equipment Manufacturer) engine is sometimes included in the name of the car. For example, the "Audi A4 2.0T" gives the make, model, engine displacement in liters, and turbo indicator, respectively. In some places, the engine displacement will directly correlate with the road tax of owning and operating the vehicle due to the higher amount of emissions and lower efficiency that high displacement engines tend to have.

2.2 Initial Questions

Based on the data that was available to me, I had to choose some general questions that I could try to answer. First, I wanted to know what aspects of the data had the greatest bearing on fuel economy and what would be interesting to explore further. Some of the variables in the dataset that had potential impacts on fuel economy included vehicle class, engine cylinders, engine displacement, fuel type, drive, transmission, and make/model.

After examining the data further, I found that some of the variables were missing a significant amount of necessary values to effectively examine or visualise them. Engine displacement was chosen as the main variable to examine, as the data for it was complete and reasonably related to some of the things I was hoping to show in my visualisations. I also sought to examine some of my findings over time and see trends in how fuel economy has changed over time, both for the dataset in general and for possible small multiples examples, like for specific classes of vehicle or different transmissions.

I set out to find some of the following things:

- Visualization of MPG based on engine displacement
- Overall improvement of MPG over time
- Specific improvements in MPG for different vehicles
- The genesis and continuation of electric cars

These tasks all seemed like things I could easily visualize and draw meaningful conclusions from, so it was time to go about the process of cleaning the data.

3 DATA CLEANING AND ORGANIZATION

The raw dataset I used was found on kaggle.com. It was one of the larger ones, with one single .csv file measuring 11MB. The dataset itself can be found here: https://www.kaggle.com/epa/fuel-economy

3.1 Initial Cleaning

At first, the data was large and contained a great deal of missing or incorrect values. The uncleaned dataframe contained 81 columns and about 38,000 rows. Many of the variables had inconsistent, missing, or seemingly incorrect values. In order to begin cleaning the data into a more usable and sensible format, I removed altogether every column containing a significant amount of missing or erroneous values. Some contained almost exclusively missing or erroneous values, which I knew would skew the data, so those columns were removed. The columns that were retained were either complete with data or mostly complete and still useful.

Next, after further examination of the dataset, I found that cars from the years 1984-1998 had a greater proportion of missing or erroneous values. Within the R code, the partially cleaned data was read in and then immediately cut into a smaller dataset containing only values from 1999 and onward. I felt that this was useful for two reasons. Firstly, it removed a number of missing values that would have skewed the data and made the dataset easier to clean. Secondly, it gave me a smaller scope so that the data would be easier to understand and trends or patterns would be better represented in my visualizations.

3.2 Additional Cleaning and Organization

In order to make one of the visualizations, I needed a smaller subset of the main dataframe. In this subset, most columns were deleted, with the ones retained being the ones that I deemed necessary to complete the visualisation: year, class, engine displacement, and MPG. To make a small multiples visualisation, I selected 9 of the most common vehicle classes and created another subset containing only their information. This allowed me to compare trends in fuel economy over time for some of the most common types of cars, which would give me useful information for drawing my conclusions.

4 VISUALIZATIONS/RELATIONSHIPS EXAMINED

As previously stated, the four main variables that I worked with when examining the data were year, class, engine displacement, and MPG. Three visualizations were created in order to examine 3 different kinds of relationships in the data.

4.1 Relationship: Displacement and MPG

The first visualization that I created was one to show the relationship between an engine's displacement and the average MPG of the vehicle. The purpose of this visualization was to showcase a very simple relationship that is likely already known to anyone who has rudimentary knowledge of cars or engines. The visualization is shown in figure 1.

I thought that this would be a helpful visualization to begin with, as it confirms the relationship between fuel economy and engine displacement. Not everyone is necessarily aware of the relationship between engine displacement and fuel economy, or even what engine displacement is. For that reason, I decided it would be necessary to include. It shows the average fuel economy for each measure of engine displacement, as well as a rough visual of how common each engine displacement is.

4.2 Relationship: Displacement over the years

My next visualization was made to show the relationship between average fuel economy of vehicles and the year the vehicle was manufactured. This shows an important trend in how average fuel economy has improved over the years. Each data point is also colored according to its engine displacement so that the average for each year reflects where each displacement lies. The visualization is shown in figure 2.

The visualization shows a gradual upward trend, indicating that average fuel economy is slowly improving over time. The lighter blue data points indicate higher displacement engines, which are consistently on the lower end of the average. They are also somewhat stagnant, such that the average fuel economy of high displacement engines has not improved much over time. The darker blue data points indicate lower displacement engines, which make a slow but clear rise as time goes on. Near the end of the graph, they begin to increase by a greater margin. The outlier data points in gray far above the rest of the points indicate electric vehicles, which are less common than gasoline powered cars but are much more efficient in terms of fuel consumption. When comparing MPG of a gasoline car to an electric car, the metric is no longer effective as electric cars do not use any gasoline. Electric cars can thusly be measured in MPGe, or Miles per Gallon equivalent. The conversion from MPG to MPGe is as follows:

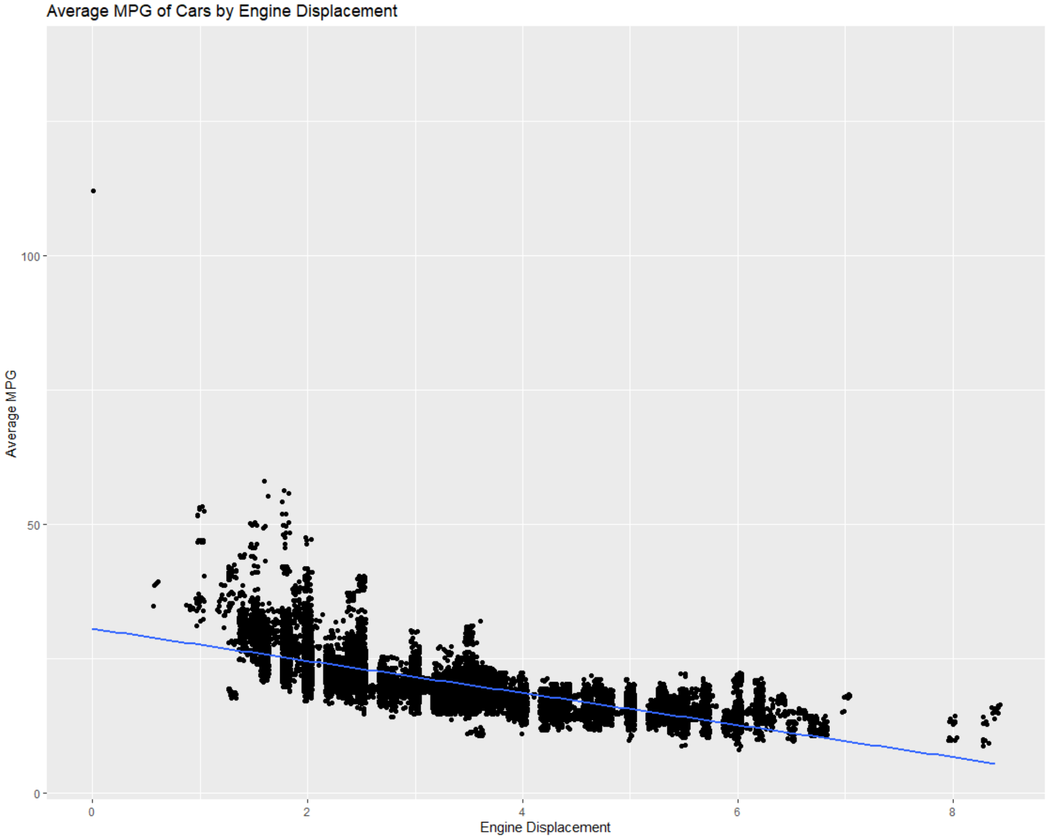


Fig. 1. Average MPG of vehicles based on engine displacement

"According to the definition used by the federal government, MPGe represents the number of miles a vehicle can travel using a quantity of fuel – electricity and gas – that has the same energy content as one gallon of gasoline. A gallon of gasoline is roughly equal to 33 kilowatt-hours of battery power. Roughly speaking, 33 kilowatt hours equals about 102 miles of city driving and 94 MPGe on the highways, give or take a few miles based on driving conditions." [4]

4.3 Relationship: Small Multiples

For my third and final visualization, I intended to compare the average fuel economy over time for 9 specific classes of cars, being some of the most common ones:

- Compact cars, like two door coupes
- Medium cars, like 4 door sedans
- Large cars, like SUVs
- Subcompact cars, like Smart Cars
- Passenger vans
- 4WD Pickup Trucks
- Minivans
- Two seaters
- Midsize station wagons

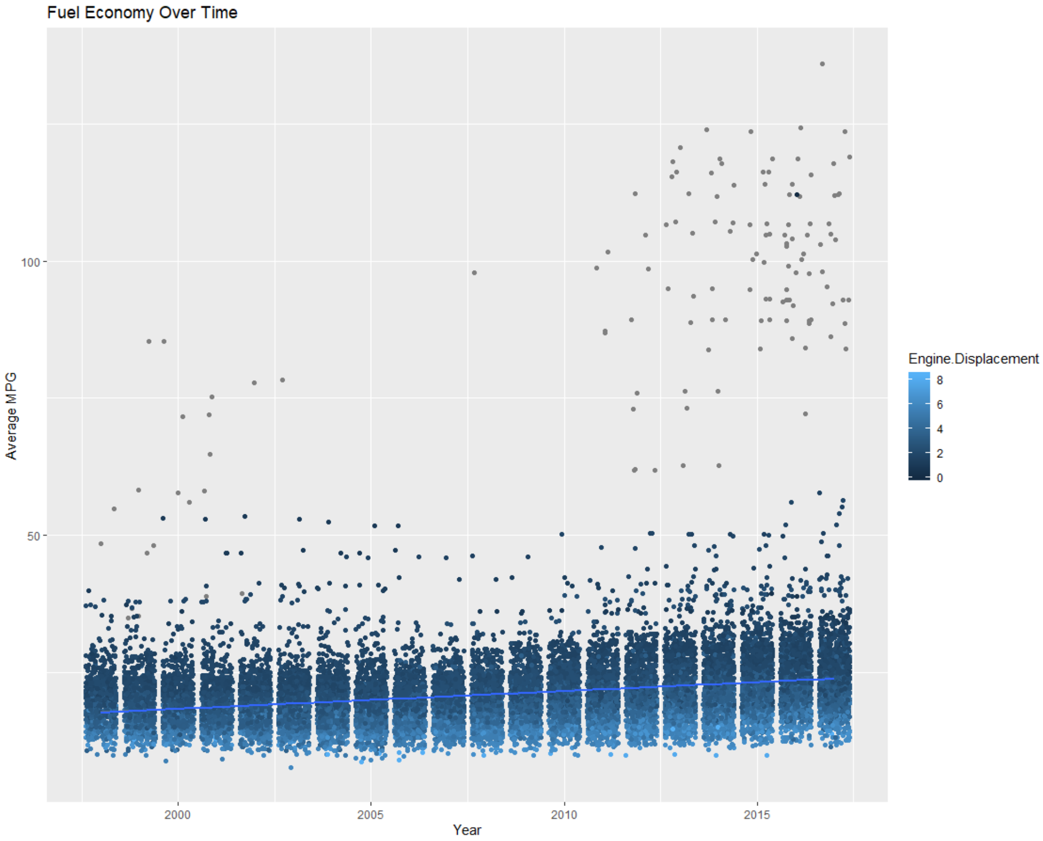


Fig. 2. Average MPG of different engine displacements over time.

With this visualization, I hoped to examine not only how different vehicle classes stacked up against one another in terms of fuel efficiency trends, but also see how each individual class has performed over time. The final visualization is shown in figure 3.

5 RESULTS

The results shown by the visualizations and linear regressions can help us draw conclusions and make inferences, but we also need to show statistical significance. In order to show this significance, I created a linear model in R comparing MPG to year and also comparing MPG to engine displacement. The statistical results are shown in the table in figure 4.

5.1 Engine Displacement vs. MPG

The visualization of Engine Displacement vs. MPG and the statistical results of the linear model constructed both show a statistically significant relationship between MPG and engine displacement. While this relationship is already widely known, it is important to display and examine it regardless of common knowledge. We can see from the visualization that the most fuel efficient cars are between 2 and 4 cubic inches in engine displacement. These data points represent common passenger vehicles, and make up a good portion of daily commuter traffic. We can also see that, for most

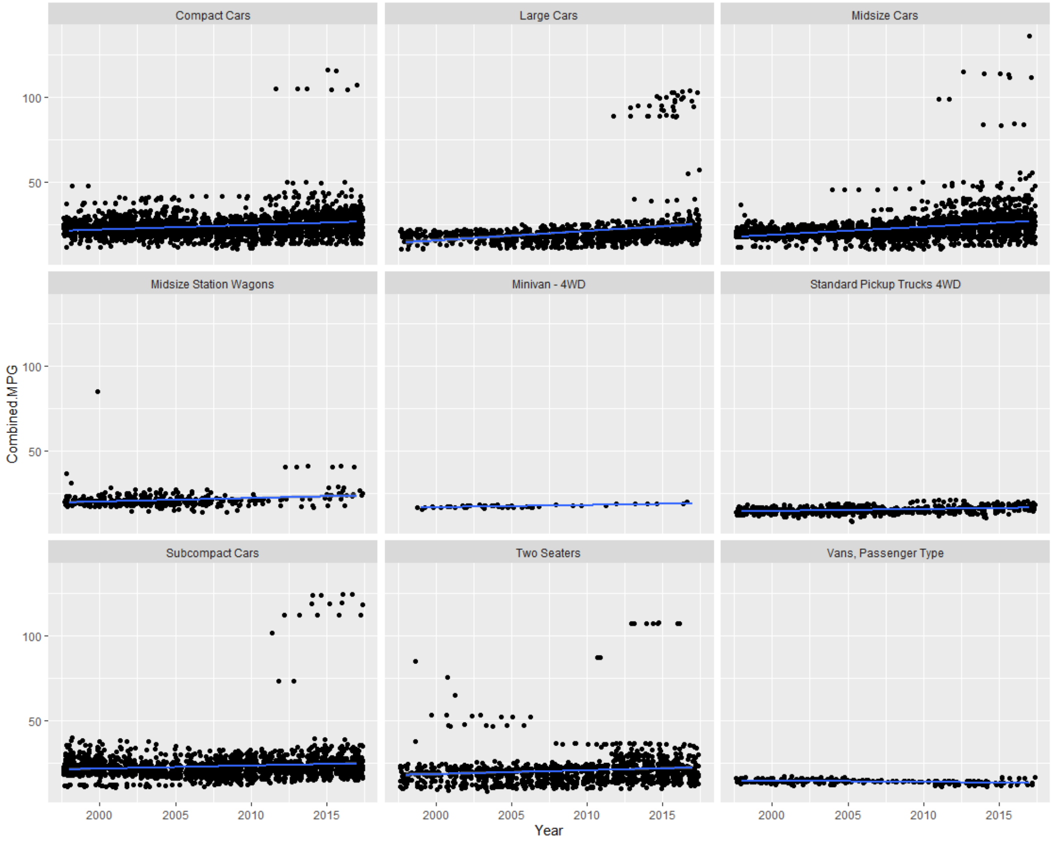
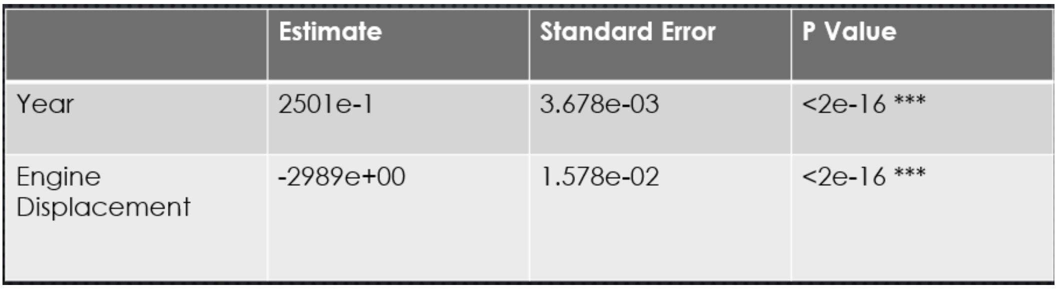


Fig. 3. Comparing average MPG over time of 9 common classes of vehicle.



	Estimate	Standard Error	P Value
Year	2501e-1	3.678e-03	<2e-16 ***
Engine Displacement	-2989e+00	1.578e-02	<2e-16 ***

Fig. 4. Statistical output of a linear model evaluating the relationships between MPG and manufacturing year as well as MPG and engine displacement.

every car where MPG can be measured (this would be in gasoline powered engines), the average MPG seems to stagnate around 50 MPG. Few non-electric cars in the dataset surpassed this.

5.2 MPG Over Time With Engine Displacement

The visualization comparing average MPG over time with respect to engine displacement and the statistical results of the linear model showed a statistically significant relationship between MPG and manufacturing year. The trend is upward, indicating that the average MPG of vehicles on the road has been steadily increasing over time. For higher displacement engines, we can see that the average MPG increases over time, but not by much. This could indicate that higher displacement engines are difficult to make more efficient, especially because they are usually powering larger and heavier vehicles. The lower displacement engines have a wider range of MPG ratings, and the ratings are generally higher. Towards the end of the graph, the outliers of the small displacement engines grow with the trend, possibly indicating that we're still able to make fuel efficiency improvements on smaller displacement vehicles. Additionally, based on the gradient that the color assignment shows, lower displacement cars appear to be more common than higher displacement cars.

As for electric vehicles, we can see that they consistently outperform their gasoline powered counterparts by a large margin. That being said, electric vehicles are shown to be far less common than gas powered vehicles. This could be due to price, inconvenience, lack of access to charging stations, or many other factors. Additionally, we can see a drop in high-MPG electric vehicles from approximately 2003-2011. This could be due to a lack of interest due to some inaccessibility factors in the earlier years of electric cars. The resurgence could possibly be attributed to Tesla Motors going public in 2010, as their vehicles are strictly electric.[3]

5.3 MPG Over Time For Common Vehicle Classes

In the third visualization, we compared the average MPG over time for certain common classes of cars. Different classes of cars are made for a number of different purposes. For example, we know before modeling that a standard sedan probably gets better mileage than a passenger van, but this is because vans are built to consume more gas in order to transport more weight. This is important to keep in mind when trying to decide if any type of car is objectively "better" than another one. There are clear disparities in MPG of different vehicle classes, and each class has trends over time that can tell us more about the vehicle itself. We can draw some conclusions from the visualizations:

- Minivans and passenger vans are exclusively found with ratings below 25 MPG. This is likely due to the nature of the vehicles themselves as their use requires a larger engine. Neither of these vehicle classes showed a significant improvement over time; the average MPG of a passenger van has actually slightly decreased over time.
- Standard pickup trucks and midsize station wagons tend to hover around 25 MPG, and both showed slight improvements in average fuel efficiency over time. However, both classes have fewer data points than some other classes, which could suggest these types of cars are less popular or less practical.
- Subcompact cars, compact cars, and two seaters all display average MPGs around 25, and the average increases over time for all of them. Unlike the classes mentioned thus far, all of these classes have numerous outliers that are around the 100 MPG mark. This could indicate a higher prevalence for electric or hybrid cars being of one of these classes. They are also shown to be more common and with a greater range than the classes previously mentioned.
- Midsize and large cars both have positive trends in average mileage over time, and both begin to show outliers in hybrid and electric range. Midsize cars, like crossover SUVs, show the greatest range in efficiency and a large jump in both popularity and efficiency in later years. This could suggest that there is a higher demand for these cars, and we could infer that more resources are being put into the manufacturing and efficiency of these cars.

The statistical output of these linear models confirms a linear relationship between a vehicle's engine displacement and its average MPG. It also confirms a linear relationship between the average MPG of cars and the year of manufacturing. We have successfully answered some of the questions we sought to answer, and in the process we have found information that allows us to speculate other auxiliary relationships.

5.4 Possible inferences

The automotive industry is a complex and multifaceted one, and a great deal of factors are involved when making observations and inferences about it. Our data has concluded, with statistical significance, that there is a positive relationship between average MPG and year of manufacturing and that there is a negative relationship between average MPG and engine displacement. Our data also shows us that the average MPG trend of a car over varies for different classes of cars. Using this information, we can make a number of inferences, speculations, and other observations. For example, we could infer that electric cars are gaining popularity and becoming more common or more accessible based on the data points shown in figure 2.

Observing figure 3, we can see some trends in hybrid and electric vehicles. Since no gasoline car surpasses 50 MPG, we can assume that all data points indicate hybrid or electric vehicles. In the early days of electric cars, most electric or hybrid vehicles were of the two seater class. In the past decade, electric cars of other classes, like compact cars and large cars, have begun to appear. We can infer that electric cars are becoming more common and more types of electric cars are becoming available.

Perhaps the most important piece of information viewed in this dataset is that partly because of electric vehicles, average fuel economy is improving.[2] In the United States, we do not have a great deal of public transportation in comparison to many other countries. We rely heavily on other cars, with the U.S. cars per capita being at 910 vehicles on the road for every 1,000 people, in comparison to the U.K.'s 526 or Germany's 554.[1] In a time where sustainability is paramount, we should be making efforts to increase efficiency and lower environmental impact, and this visualization shows that we are beginning to do so.

6 CONCLUSION

Based on the data we have examined and the conclusions from our statistical tests, we can successfully conclude a few things. Firstly, we can confirm that the higher your car's engine displacement is, the worse its fuel efficiency will be. Secondly, we can confirm that over time, fuel efficiency has been improving, partly thanks to the relatively new electric vehicle industry. We can also conclude that fuel efficiency varies between different classes of cars. The inferences we made above, while supported by our visualizations, cannot be confirmed to have a linear relationship or correlation until a statistical test is conducted. Thankfully, our data and examination of it give us a promising premise for the future: we're becoming more efficient.

REFERENCES

[1] 2019. *How Many Cars Per Capita In The USA*. Retrieved December 11th, 2019 from https://capitol-tires.com/how-many-cars-per-capita-in-the-us.html

[2] Steven Finlay. 2018. *Auto Analyst Forecasts Industry Trends for 2019*. Retrieved December 11th, 2019 from https://www.wardsauto.com/dealers/auto-analyst-forecasts-industry-trends-2019

[3] Greg Kumparak, Matt Burns, and Anna Escher. 2015. *A brief history of Tesla*. Retrieved December 11th, 2019 from https://techcrunch.com/gallery/a-brief-history-of-tesla/

[4] Kevin Woo. 2016. *What is MPGe: electric car fuel economy ratings explained*. Retrieved December 11th, 2019 from https://www.yourmechanic.com/article/what-is-mpge-electric-car-fuel-economy-ratings-explained-by-kevin-woo