Are Newer Cars More Efficient?

Ethan Hansen - Stella Obeng-Darko - 04/04/21 - Business Analytics Final

Introduction

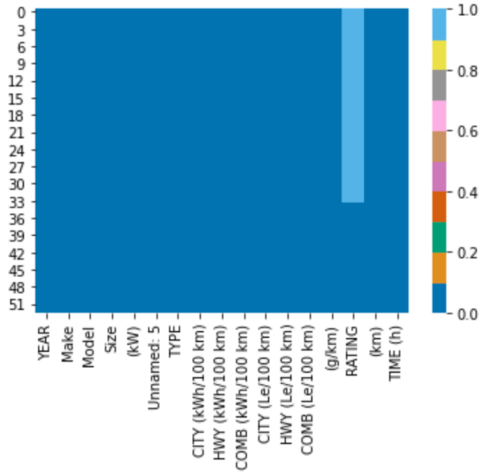
On Feb 9 2021, John DeCicco wrote an article on ***GreenBiz*** where he discussed how an increase in fuel efficiency is required for a greener auto fleet. In the article, he talked about the green deal in the new Biden administration that is meant to encourage more industries to produce products with less emissions. This includes the electric cars in our data set. According to DeCicco, with more cars being produced, we should also expect an increase in their fuel economies.

We'll be looking at a data set that contains the properties of electric cars such as Tesla, Kia, Nissan, Ford, Chevrolet, Smart, and Mitsubishi.

Methodology

Data Cleaning

The first thing we did to clean our data was to create a heatmap that would locate all of our missing values. The image below is the heatmap that was created.



The heatmap is showing that there is one column that has missing values. This was the “Ratings” column, the only rows that had values in that column were the cars that were made in 2016. The light blue color in the heat map is showing where these missing values are. We were also able to check what percent of values were missing in this “Ratings” column. There is about 64% of the values missing in the “Ratings” column, because of this we decided that we should just delete the whole column from the dataset because there wouldn't be enough information to use.

After looking for columns that have missing values, we looked to see if there were any columns that had too many values in that column that were the same. To do this we searched for rows that had higher than 95% of the values in the column being the same. The image below shows the result that we got after this search.

```
Unnamed: 5: 100.00000%
A1    53
Name: Unnamed: 5, dtype: int64

TYPE: 100.00000%
B    53
Name: TYPE, dtype: int64

(g/km): 100.00000%
0    53
Name: (g/km), dtype: int64
```

We can see in this image that the “Unnamed: 5” column, the “Type” column, and the “(g/km)” column have 100% of the values in those columns being the same. This means that in those columns every value for every row is the same. We decided that these columns would not be useful so we dropped all of these columns from our dataset.

One other thing that we did to clean up our columns was to rename the columns so the column names looked nicer and to make it easier for future coding. We renamed 'YEAR' to 'Year', 'CITY (kWh/100 km)' to 'City_kWh', 'HWY (kWh/100 km)' to 'Hwy_kWh', 'COMB (kWh/100 km)' to 'Comb_kWh', 'CITY (Le/100 km)' to 'City_Le', 'HWY (Le/100 km)' to 'Hwy_Le', 'COMB (Le/100 km)' to 'Comb_Le', and 'TIME (h)' to 'Time'. Other than the columns that we deleted the data set was very clean and there were no other null values in any of the columns. The image below shows the top 5 columns in our data set after cleaning.

	Year	Make	Model	Size	(kW)	City_kWh	Hwy_kWh	Comb_kWh	City_Le	Hwy_Le	Comb_Le	(km)	Time	Before_2015
0	2012	MITSUBISHI	i-MiEV	SUBCOMPACT	49	16.9	21.4	18.7	1.9	2.4	2.1	100	7	1.0
1	2012	NISSAN	LEAF	MID-SIZE	80	19.3	23.0	21.1	2.2	2.6	2.4	117	7	1.0
2	2013	FORD	FOCUS ELECTRIC	COMPACT	107	19.0	21.1	20.0	2.1	2.4	2.2	122	4	1.0
3	2013	MITSUBISHI	i-MiEV	SUBCOMPACT	49	16.9	21.4	18.7	1.9	2.4	2.1	100	7	1.0
4	2013	NISSAN	LEAF	MID-SIZE	80	19.3	23.0	21.1	2.2	2.6	2.4	117	7	1.0

Formulate Hypothesis

Based on our literature review, we develop some hypotheses that we can test to either confirm or deny. For this specific research question, we anticipate that as the years go by, there is an increase in the car's efficiency. That is, the kWh/100km it exerts in cities and highways is lesser. This kWh/100km represents the fuel economy. The fuel in these cars is measured in kilowatts because these are electric cars. The lower the kWh/100km the more efficient the car is.
This hypothesis means that if the data is statistically significant, there is a positive correlation between combined efficiency of a car and the year it was manufactured.

Test Hypothesis

The data is then plotted into *jupyterlab* to observe and analyze the results. Based on the regression analysis below, we construct a table showing the coefficient and the probability values(p-value). We then interpret the results if it is statistically significant. Our Null Hypothesis(βj) here is that there is no significant difference between the fuel economy of electric cars as the years go up.

Plot Tables/Graphs

To enhance our conclusions, we plot graphs to give us a visual representation of the data analysis. This data would be more conclusive if we look at histograms and box plots. The histograms help us look at the discrepancy in our data for cars . The box plots will allow us to see which years have more of the data within a certain range of kWh/100km.

Variables of Interest

All the variables that we are going to be working with include “Year” which is the year that the car was made, “Comb_kWh” which is the kilowatts per 100 killometers that each car uses, and the “Size” which is the size of the car. We will also be looking at a dummy variable in order to get a variable that shows if a car was made before 2015 or not. This dummy variable column will be called “Before_2015”. In this column you will see a “1” in every row that has a car that was made before 2015 and the rest of the rows will have a “0”. We also found the mean of this dummy variable column which was about 0.37. This means that approximately 37% of the years represented are from 2012-2014 and approximately 63% of the years are from 2015-2016.

Outcome Variable

Our outcome variable in our equation is going to be the amount of kilowatts used every 100 kilometers by the car (Comb_kWh). We want to see if different factors of the car like the year it was made or the amount of liters that it uses per 100 kilometers affects the amount of kilowatts the car uses.

Main explanatory variables

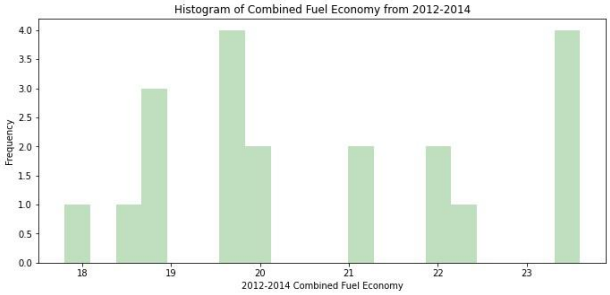
The main variable that we are going to be using is the dummy variables that we created for the year the car was made. We want to see if the year that the car was made will affect the amount of kilowatts that the car uses.

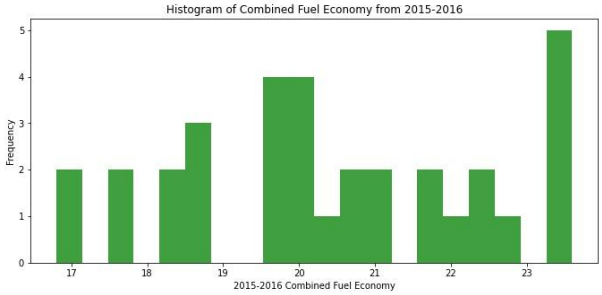
Control Variables

From the data available, we noticed that the model and the make of the car could change the results. The model of a car could have some extra features that made it more efficient no matter the year it was manufactured in. The make of a car could also be very significant because all companies make cars differently and some could be more fuel efficient than others. To best capture both the effect that the make and model have on the efficiency of the car we thought it would be best to use the “Size” column as our control variable because this will explain the shape of the car.

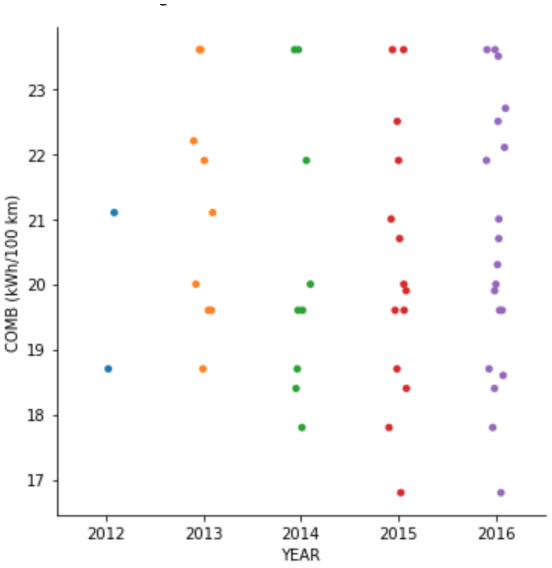
Results

Interpret Graphs/Tables

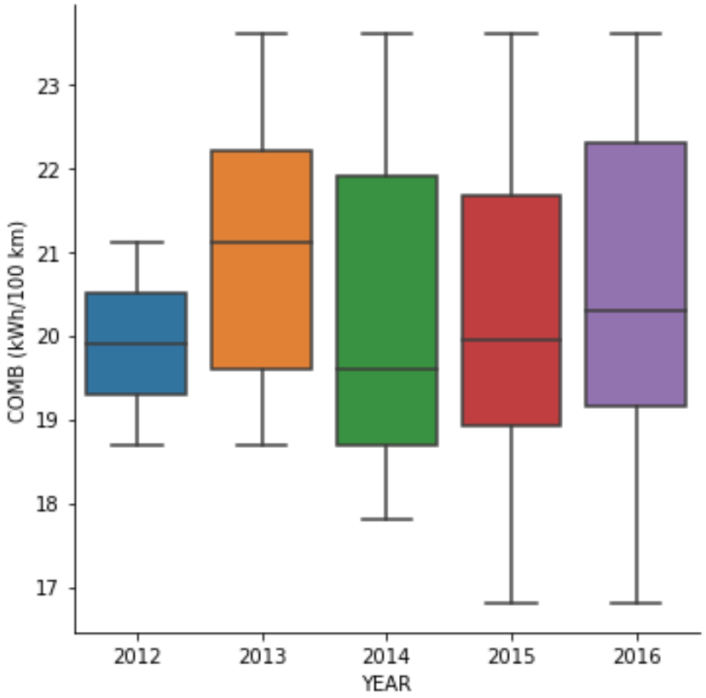




The number of combined kWh/100km below 20 is 8 for cars made between 2012 and 2014 and the number of combined kWh/100km below 20 for cars between 2015 and 2016 is 15. This difference is not really significant because the cars above 2015 and 2016 represent 63% of the data. Based on our hypothesis, we expect to see the majority of the combined efficiencies from 2015 to 2016 to be below 20. Thus, there is not really a difference between the years with regards to having 20 or less combined kwH/100km.



From the scatter plot above, we can observe that there is no correlation between the years and the combined kWh/100km. The points on the plot are scattered randomly and do not actually form a line that slants from left to right or from right to left. Thus, we cannot draw a line of best fit.



We can see that the median for the different years all fall within the same areas around 20.

The data for 2014-2016 is positively skewed. Which means that the majority of the data fall below the median of 20. Which supports our alternative hypothesis of expecting that the majority of the kWh/100km would be lesser for those years.

Regression Model

Our regression equation is as follows:

$Y = \beta_0 + \beta_1 (Before\_2015) + \beta_2 (Size) + error$

Where, Y = Combined Efficiency(Comb_kwH)

β_0 = Y intercept

β_1 = Coefficient of Before_2015

β_2 = Coefficient of Size

error = the difference between the true value and the regression line

We reject βj if p-value < α.

α here is our significance level which is 5%.

Conclusion

Are Newer Cars More Efficient?

	Before vs After 2015
Intercept	19.64***
	(0.47)
Before_2015	0.71***
	(0.26)
Size[T.STATION WAGON - SMALL]	0.26
	(0.79)
Size[T.SUV - STANDARD]	3.46***
	(0.79)
Size[T.TWO-SEATER]	-0.40
	(0.55)
Size[T.SUBCOMPACT]	-1.88***
	(0.53)
Size[T.FULL-SIZE]	2.47***
	(0.49)
Size[T.MID-SIZE]	-0.67
	(0.58)
R-squared	0.82
R-squared Adj.	0.80

The image above shows the regression table that we made after running our regression. This regression table is giving us all the information we need to create our regression equation. The

table gives us all the coefficients for each variable, this is the number across the variable without the parentheses. The table also gives us an r - squared value which tells us how much of the data fits our regression model. One other thing that the table is showing us is whether the variables are significant or not. The variables that have three stars next to them are significant at a 1% p - value and the variables that have no stars are insignificant.

After looking at all the data that the regression table is giving us we are able to make our regression equation which you can see below:

Comb_kWh = 0.71(Before_2015) + 3.46(T.SUV - Standard) - 1.88(T.Subcompact) + 2.47(T.Full - Size) + 19.64

Based on the regression analysis above, all other things being equal, when the year is between 2012 and 2014, the combined efficiency increases by 0.71 kWh/100km, if the car is a standard SUV then the efficiency is increased by 3.46 kWh/100km, if the car is a subcompact car then the efficiency is decreased by 1.88 kWh/100km, and if the car is a full - size car than the efficiency is increased by 2.47 kWh/100km. We can also say that about 82% of our data is representing our regression so our variables have a strong impact on efficiency.

For the hypothesis test, we reject the null.

- H_0: $\beta j = 0$ (no impact)
- H_1: $\beta j \neq 0$ (have an impact)
  - Decision Rule: Reject the null hypothesis if p-value $< \alpha = 0.05$.

Because our alpha value here is 5% and the p-value is less than 0.05 we reject the null hypothesis. We were able to see that there is a difference between efficiency of cars made before 2015 and cars that were made in 2015 or 2016 as we hypothesized but it is only by 0.71 kWh/100km. Therefore, based on our regression, we have sufficient evidence that newer cars are more efficient than older cars but not by a lot.

Limitations

- There is very little variation in the years that we have. Only 5 years are represented here but they are not even represented equally.
- The car brands here are not very diverse as it includes very little cars and includes a lot more car brands than others.
- We observed the combined efficiency of the cars but maybe we could have looked at the individual efficiencies of the cars for cities and highways. This probably might have allowed us to draw a more definitive conclusion.

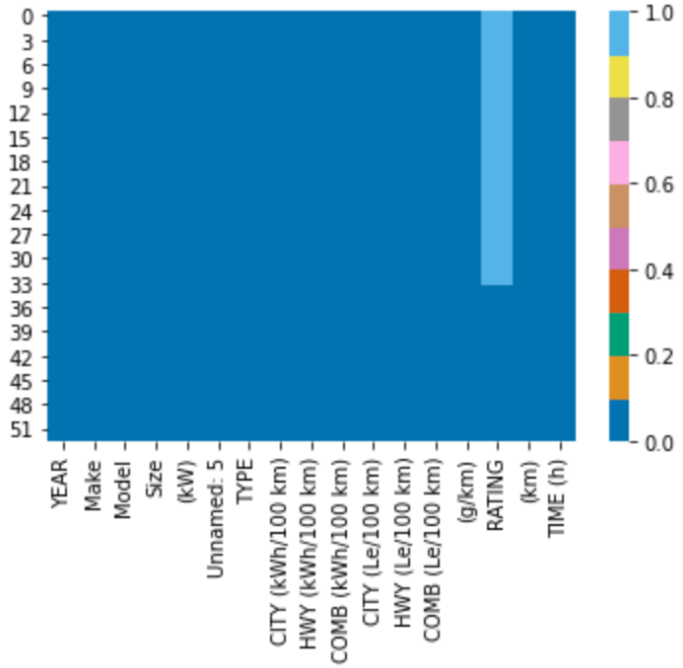
Recommendations

Based on our results, we suggest that car manufacturers increase the efficiency of cars that they make as the years go on. We did observe an increase in cars' efficiency because the newer cars have a dip in their kWh/100km. But, this dip was only by 0.71 kWh/100km. When a new car is released, the expectancy is that the car should be more fuel efficient than the previous version.

Bibliography

DeCicco, John. "A greener auto fleet requires increasing fuel efficiency and selling electric vehicles." *GreenBiz*, 9 Feb 2021, https://www.greenbiz.com/article/greener-auto-fleet-requires-increasing-fuel-efficiency-and-selling-electric-vehicles. Accessed 30 March 2021.

Appendix




```
Unnamed: 5: 100.00000%
A1    53
Name: Unnamed: 5, dtype: int64

TYPE: 100.00000%
B    53
Name: TYPE, dtype: int64

(g/km): 100.00000%
0    53
Name: (g/km), dtype: int64
```

	Year	Make	Model	Size	(kW)	City_kWh	Hwy_kWh	Comb_kWh	City_Le	Hwy_Le	Comb_Le	(km)	Time	Before_2015
0	2012	MITSUBISHI	i-MiEV	SUBCOMPACT	49	16.9	21.4	18.7	1.9	2.4	2.1	100	7	1.0
1	2012	NISSAN	LEAF	MID-SIZE	80	19.3	23.0	21.1	2.2	2.6	2.4	117	7	1.0
2	2013	FORD	FOCUS ELECTRIC	COMPACT	107	19.0	21.1	20.0	2.1	2.4	2.2	122	4	1.0
3	2013	MITSUBISHI	i-MiEV	SUBCOMPACT	49	16.9	21.4	18.7	1.9	2.4	2.1	100	7	1.0
4	2013	NISSAN	LEAF	MID-SIZE	80	19.3	23.0	21.1	2.2	2.6	2.4	117	7	1.0

Are Newer Cars More Efficient?

	Before vs After 2015
Intercept	19.64***
	(0.47)
Before_2015	0.71***
	(0.26)
Size[T.STATION WAGON - SMALL]	0.26
	(0.79)
Size[T.SUV - STANDARD]	3.46***
	(0.79)
Size[T.TWO-SEATER]	−0.40
	(0.55)
Size[T.SUBCOMPACT]	−1.88***
	(0.53)
Size[T.FULL-SIZE]	2.47***
	(0.49)
Size[T.MID-SIZE]	−0.67
	(0.58)
R-squared	0.82
R-squared Adj.	0.80

OLS Regression Results

Dep. Variable:	Comb_kWh	R-squared:	0.823
Model:	OLS	Adj. R-squared:	0.796
Method:	Least Squares	F-statistic:	29.94
Date:	Fri, 02 Apr 2021	Prob (F-statistic):	6.36e-15
Time:	14:19:48	Log-Likelihood:	-64.962
No. Observations:	53	AIC:	145.9
Df Residuals:	45	BIC:	161.7
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	19.6441	0.466	42.136	0.000	18.705	20.583
Size[T.FULL-SIZE]	2.4710	0.490	5.043	0.000	1.484	3.458
Size[T.MID-SIZE]	−0.6667	0.577	−1.155	0.254	−1.830	0.496
Size[T.STATION WAGON - SMALL]	0.2559	0.786	0.326	0.746	−1.327	1.839
Size[T.SUBCOMPACT]	−1.8788	0.530	−3.546	0.001	−2.946	−0.812
Size[T.SUV - STANDARD]	3.4559	0.786	4.398	0.000	1.873	5.039
Size[T.TWO-SEATER]	−0.4000	0.548	−0.730	0.469	−1.503	0.703
Before_2015	0.7118	0.263	2.706	0.010	0.182	1.242

Omnibus:	0.748	Durbin-Watson:	1.900
Prob(Omnibus):	0.688	Jarque-Bera (JB):	0.765
Skew:	0.046	Prob(JB):	0.682
Kurtosis:	2.419	Cond. No.	12.3

```
YEAR – 0%
Make – 0%
Model – 0%
Size – 0%
(kW) – 0%
Unnamed: 5 – 0%
TYPE – 0%
CITY (kWh/100 km) – 0%
HWY (kWh/100 km) – 0%
COMB (kWh/100 km) – 0%
CITY (Le/100 km) – 0%
HWY (Le/100 km) – 0%
COMB (Le/100 km) – 0%
(g/km) – 0%
RATING – 64%
(km) – 0%
TIME (h) – 0%
RATING_ismissing – 0%
num_missing – 0%
0.0
```