LEARN DATA SCIENCE IN A DAY: NO STATISTICS OR MATH REQUIRED!

I am sure the readers are well aware that Data Scientist has been called the "*Sexiest Job of the 21st Century*" by Harvard Business Review. We have all also heard and read about the forecasted talent gap in this field. This seems to have led to a sort of modern-day gold rush, and, unsurprisingly, shops springing up everywhere competing to sell you the lightest, quickest pickaxes. For every new course claiming that it can turn you into a data scientist in one week, there's another one claiming that it can do just as much in just one day!

Based on my experience, not all of these courses are of the same or even similar quality. Like everything else out there, there are some incredibly well designed courses and some others which do a disservice to the field and its students. Specifically, there is this new breed of courses, that focus almost entirely on how to use the tools of data science rather than on the when and why of it.

Understanding the Statistics and Math behind the relevant algorithms is as critical to data science as learning the programming tools. Just learning R, Python, Scala and knowing how to use many machine learning algorithms doesn't make one a data scientist. Being aware of the various machine learning algorithms is important, but that, by itself, doesn't give one an understanding of why a specific algorithm works best for the problem at hand.

The dangers associated with having a peripheral knowledge of these techniques can be demonstrated with even a simple example. I will walk you through the application of a simple Linear Regression model to one of the famous Motor Trends (mtcars) dataset (Download here) that most students of data science are familiar with. You can follow along even if you don't know Python or R, as the dataset is small enough to be modelled with Excel. I encourage readers to reproduce my example. I include screenshots to facilitate that

Let us begin by looking at the dataset. It lists a few cars from the 1970s and several of their attributes.

Model	mpg	cyl	disp	hp	drat	wt	qsec	vs
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0
Valiant	18.1	6	225	105	2.76	3.46	20.22	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18	0

For the sake of keeping this demo simple, let's set out to answer the simple question:

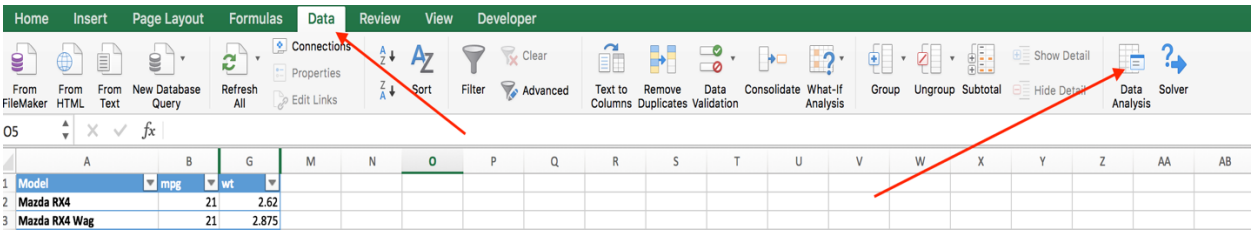
How does the mileage(mpg) of the car vary with respect to its weight (wt)?

Implementing a model for this in Excel is quite easy. Just follow along

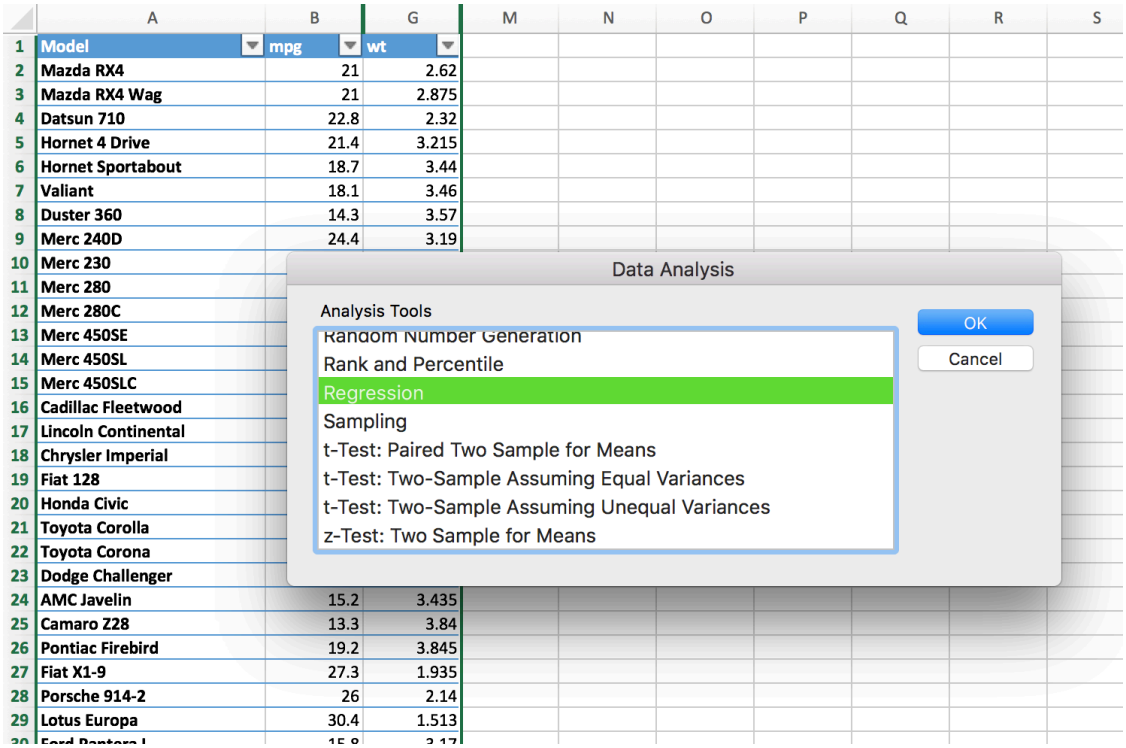
Step1: Let us begin by removing some unnecessary columns and re-arranging the data as shown below. We keep only the "Model," "wt," and "mpg" columns, as we are only interested in predicting gas mileage "mpg" as function of the car's weight "wt."

Model	mpg	wt
Mazda RX4	21	2.62
Mazda RX4 Wag	21	2.875
Datsun 710	22.8	2.32
Hornet 4 Drive	21.4	3.215
Hornet Sportabout	18.7	3.44
Valiant	18.1	3.46
Duster 360	14.3	3.57
Merc 240D	24.4	3.19
Merc 230	22.8	3.15
Merc 280	19.2	3.44
Merc 280C	17.8	3.44
Merc 450SE	16.4	4.07
Merc 450SL	17.3	3.73
Merc 450SLC	15.2	3.78
Cadillac Fleetwood	10.4	5.25
Lincoln Continental	10.4	5.424
Chrysler Imperial	14.7	5.345
Fiat 128	32.4	2.2
Honda Civic	30.4	1.615
Toyota Corolla	33.9	1.835
Toyota Corona	21.5	2.465
Dodge Challenger	15.5	3.52
AMC Javelin	15.2	3.435
Camaro Z28	13.3	3.84
Pontiac Firebird	19.2	3.845
Fiat X1-9	27.3	1.935
Porsche 914-2	26	2.14
Lotus Europa	30.4	1.513
Ford Pantera L	15.8	3.17
Ferrari Dino	19.7	2.77
Maserati Bora	15	3.57
Volvo 142E	21.4	2.78

Step 2: Click the **Data** tab, followed by the **Data Analysis** button in the right corner

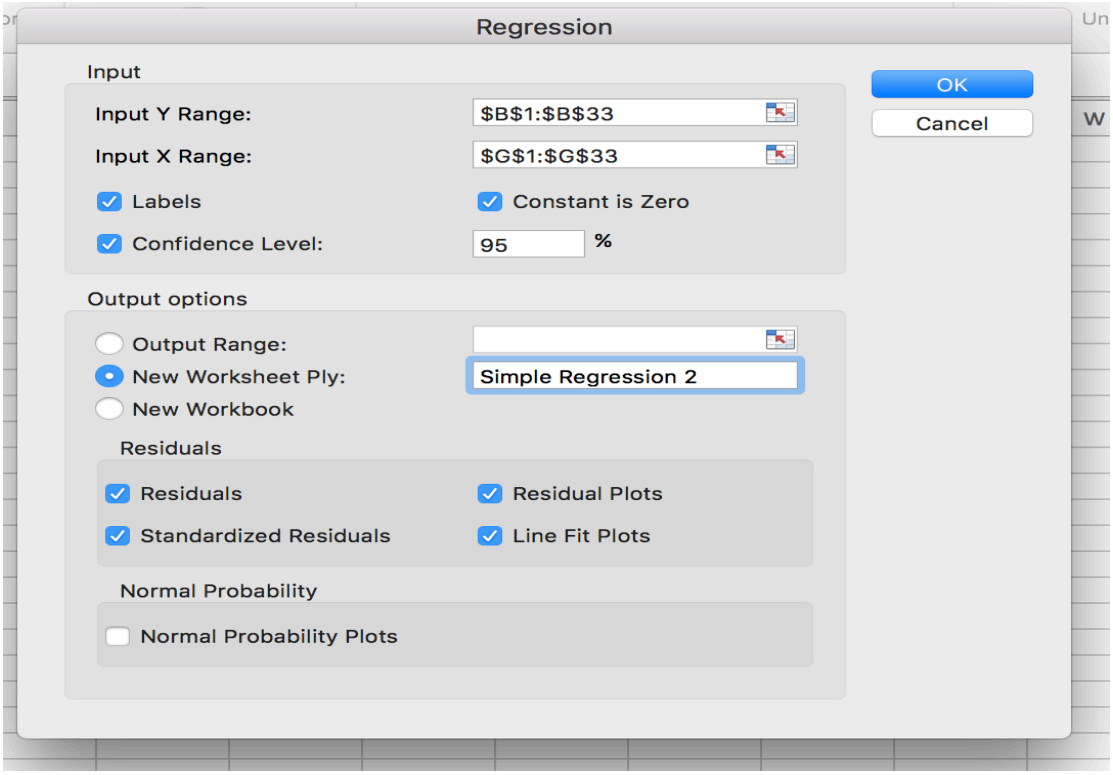


Step 3: In the Dialogue Box that opens, select **"Regression"**



Step 4: Click OK and it should open another dialogue box where you can configure the inputs for the model

- *Input Y Range:* Enter the range for the "mpg" column
- *Input X Range*: Enter the range for the "wt" column
- Check the boxes for *"Labels"*
- Set the *"Confidence Intervals" to 95%*
- Since a weightless car is expected to have zero gas mileage check *"Constant is Zero"*
- Select or give a name to the sheet where you want your output
- Select a few more boxes as shown below to let Excel produce the charts



Step 5: Click OK and Excel should show you the linear regression model and some of its details.

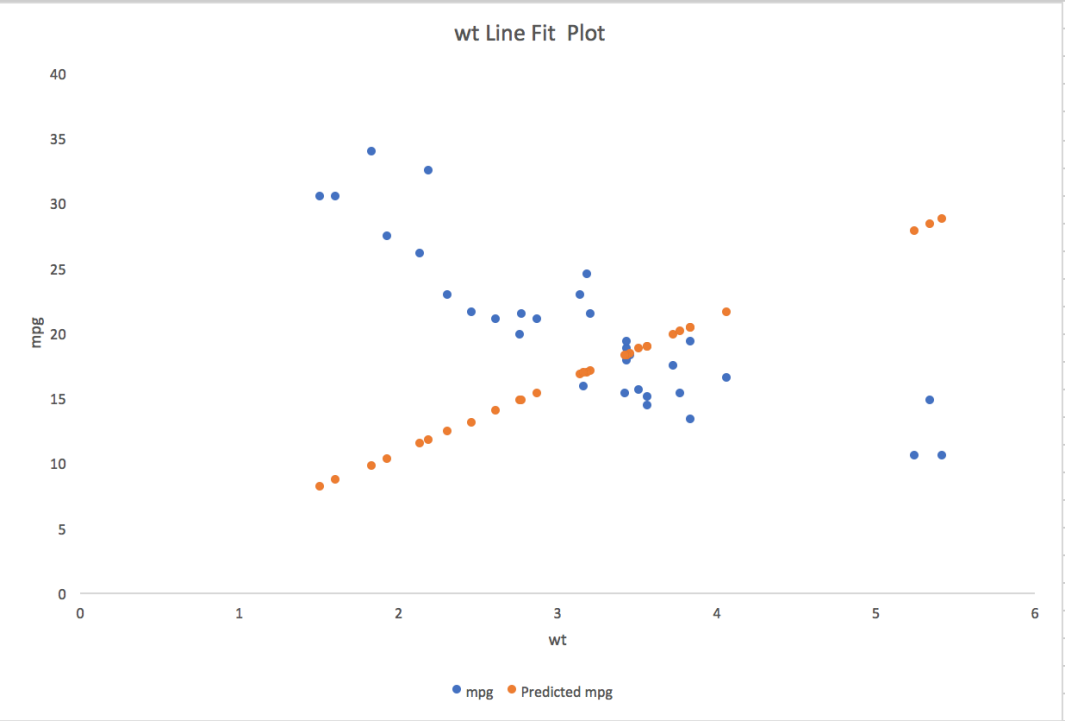
SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.848327982					
R Square	0.719660365					
Adjusted R Square	0.687402301					
Standard Error	11.26887815					
Observations	32					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	10105.69394	10105.69394	79.58015404	6.09652E-10	
Residual	31	3936.616057	126.9876147			
Total	32	14042.31				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
wt	5.291624101	0.593180134	8.920770933	4.55314E-10	4.081825241	6.501422961

This is it! It is really this simple to build a linear regression model in Excel. We now have a model for estimating the gas mileage for any car if we are given its weight.

HOLD ON A SECOND THOUGH!

An experienced data scientist or even an astute observer would stop me right here. They would have identified something absurd with this model. Can you identify what that is?

Don't sweat if you don't see what they see yet. Take a look at the chart Excel generated as part of this regression and see if you can spot the absurdity.



The blue dots in the chart represent the actual data. The orange ones represent the corresponding predictions by our model. We wanted a linear model, so we got that part right. The orange dots are indeed lined up in a straight line. So, what is wrong with our model?

It's the orientation of the line (mathematically the slope) that seems wrong. The supposed ***"best fit line"*** is running across our dataset instead of along it as we would have expected!

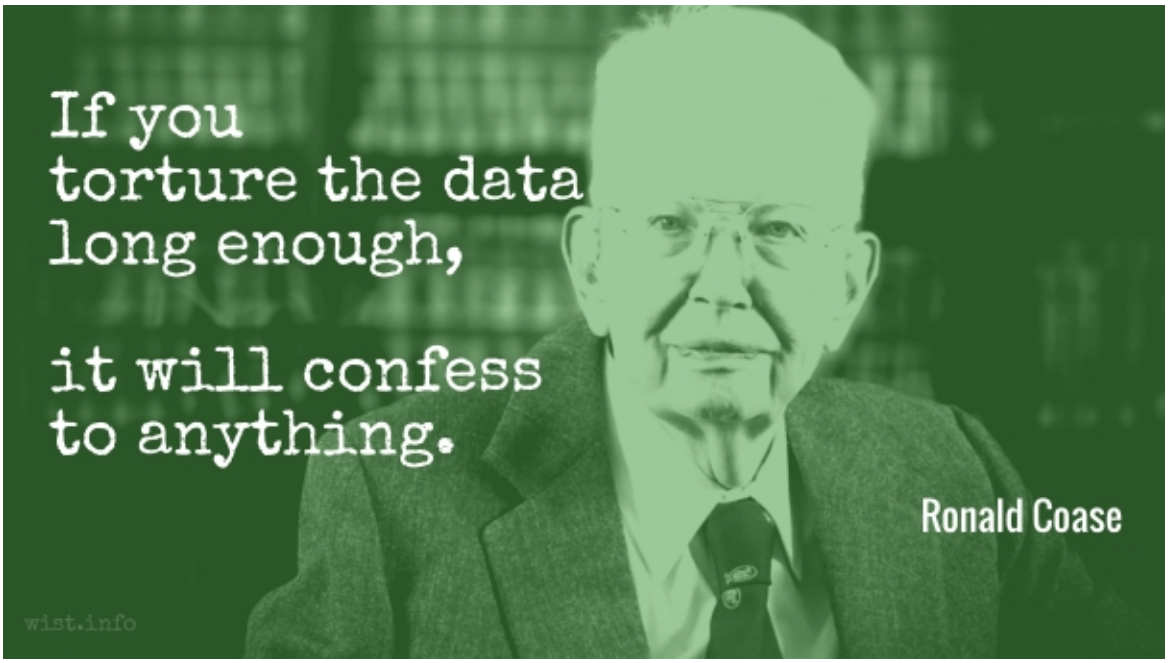
Now go back and take another look at the coefficient for Weight "wt". It is a positive 5.29. What conclusion does this lead to?

Heavier cars are predicted to have higher gas mileage???

I am not claiming to agree with that conclusion. This is not something we generally assume to know to be true in practice. I am with you on the intuition – that conclusion makes no sense!

"***But the data is telling us this – says the Data Scientist to his Boss. If you want better gas mileage make your cars heavy!"***

We can imagine an experience manager offering the common adage

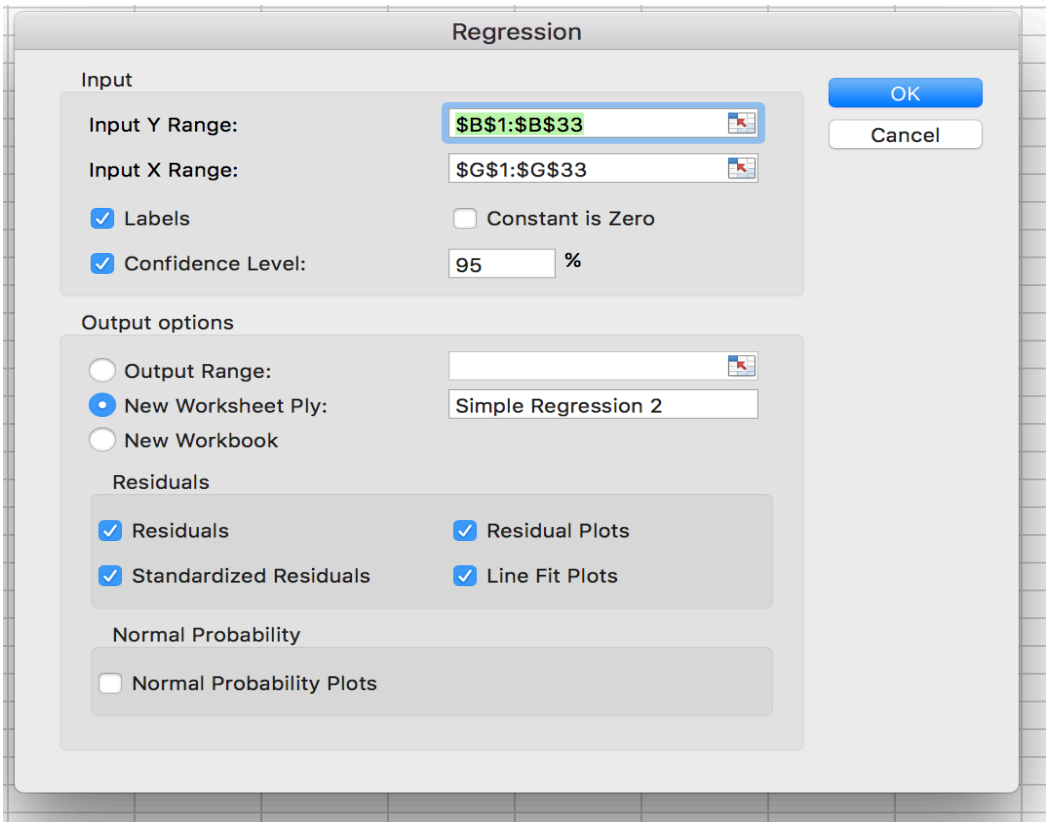


However, you can see here that we barely touched the data, let alone torture it. So what happened? What did we do wrong in our analysis?

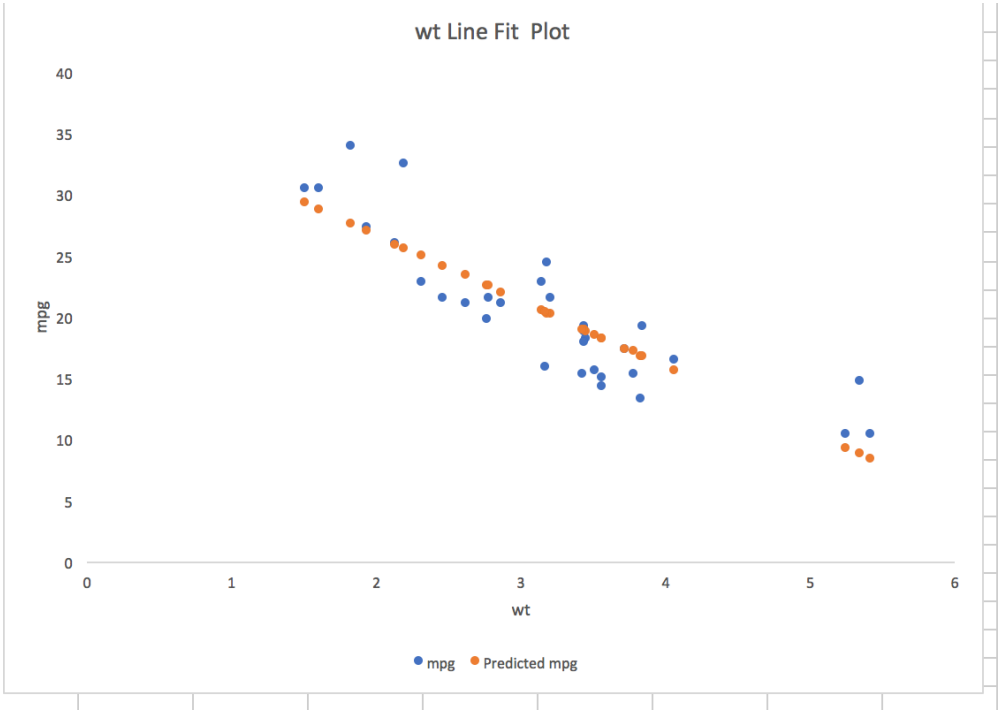
Turns out, we checked a box too many!

We ***forced*** our line to be oriented the way it was, by checking the "Constant is Zero" box. When we did that, Excel's implementation of Linear Regression, found the best fit line given our criteria. (Mathematically, Excel found the line passing through the origin that minimizes the Root Mean Square Error)

If instead, if we run the same process on the same data but with the "Constant is Zero" box unchecked, as shown below



Our best-fit line would have looked like this



And our model description would have looked like this

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.86765938
R Square	0.75283279
Adjusted R S	0.74459389
Standard Err	3.04588212
Observations	32

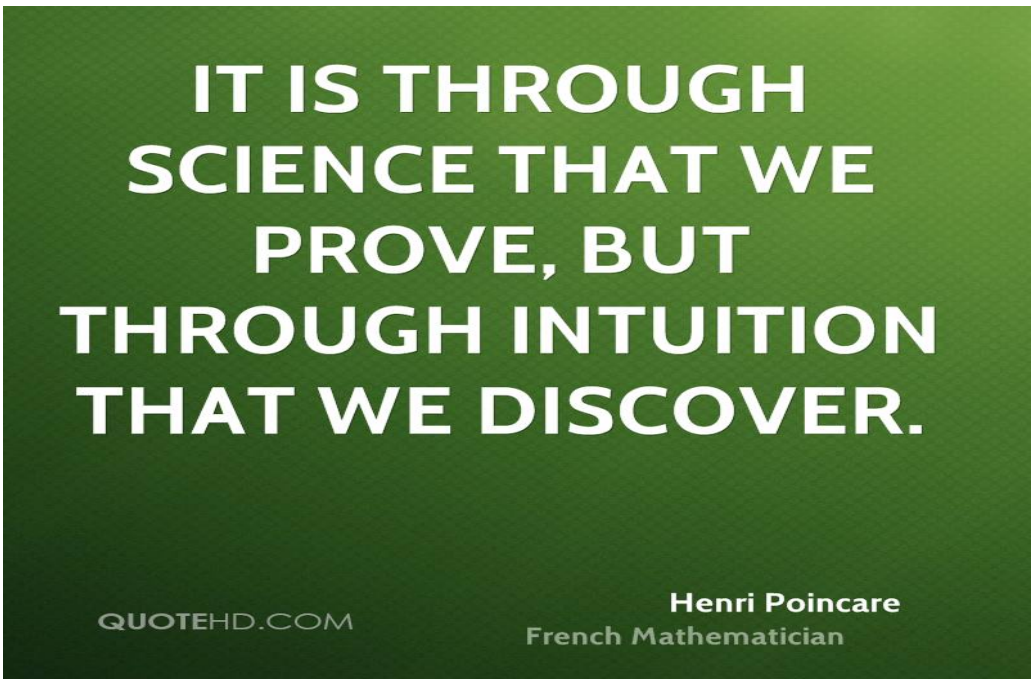
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	847.72525	847.72525	91.375325	1.294E-10
Residual	30	278.321938	9.27739792		
Total	31	1126.04719			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	37.2851262	1.87762734	19.8575753	8.2418E-19	33.4504996	41.1197528
wt	-5.3444716	0.55910105	-9.5590441	1.294E-10	-6.4863082	-4.2026349

As our intuition and experience would agree, the coefficient of Weight "wt" here is now a negative 5.34. The Data Scientist and his boss could now agree that as the weight of the car increases, the predicted gas mileage "mpg" decreases.

The moral of the story is, as French mathematician Henri Poincare once said



I hope through this simple exercise, I demonstrated, the importance of understanding the mathematics behind the data science, the assumptions made by the model and by the data scientist's implementation of the model, to come up with results that are useful and valid.

Yes, I do believe anyone can learn Data Science. All it takes is the hunger and curiosity to learn the how, the why, the when and the will to be a lifelong student. Not everyone can teach Data Science though. Definitely not in a day!

This article was written by Rohit Padebettu with invaluable contributions from Cynthia E. Correa and Dr. Sushil Bhatia