EMPOWERING USED CAR BUYERS TO SNAG A GOOD DEAL

Ted Fitch
DATA610: Fall 2020
Assignment 4
Tedfitch4@gmail.com
Dr. Laila Moretto

Summary of Dataset:

No one wants to leave the car lot with a used car feeling like a salesman pulled the wool over their eyes. It's a game that most people play, but most don't know how to play to win. So how do you know you left with a steal or were left behind in the dust (**Figure 2**)? The best strategy is to go in armed with knowledge: both green flags and reg flags to watch out for. It's been previously found by predictive studies that mileage is the greatest predictor of sales price (Engers et al., 2009). Here, we explore a dataset of used car sales to look for insights and predictors of sales price (**Figure 1**). This is in order to empower buyers to make informed decisions and know how each variable might affect price. This dataset contains 12 columns with 1,437 rows. There are continuous categories (price, age, CC, KM, HP, and weight), binary variables (manual or automatic transmission and whether the car has a metallic color or not), and categorical variables (fuel type and number of doors). Fortunately, there were no missing entries through the whole dataset. It is assumed the unit of the age column is months based on the range of values (1-80). It is unlikely a car would be sold if it were 80 years old. In addition, a new column was calculated by dividing KM by Age to describe how much usage the car has had. This was made to explore the old adage "It's not the years, it's the mileage". It's predicted that KM/Age will be the best predictor of price.

The data was preliminarily explored to determine if there were any cases that needed to be filtered. It was found there were some variables which had very few cases involved. For example, there were only 80 cars with automatic transmission; there were only 17 CNG (fuel type) vehicles; and there were only 4 cars with 2 doors (**Figure 4**). Thus, these were filtered out in order to improve results. Lastly, there were 2 vehicles with only 1 KM of mileage, but they were 50 and 76 months. Their prices were exceptionally low for having such low mileage but

high age. These were also excluded from further analyses. In total, there were 1,309 rows of data remaining. All other rows showed relative consistency and the other variables showed variance without extreme outliers. For example, the variable “metallic coloring” had a ratio of 2:1 (metallic to non-metallic) and this is acceptable since there are plenty of cases in both categories (>400)(**Figure 3**).

Key Findings:

*The first graphics made were around what was expected to have the greatest effect on sales price (***Figure 5**, **Figure 6****Figure 2. Slide 1 of story: introductory question.**

Figure 3. Slide 2 of story: Graphs showing average price separated by transmission type and whether a car has a metallic color (top). Graphs showing proportion of cars per transmission type and per having a metallic color or not (bottom). Transmission: Manual=0 Automatic=1. Metcolor: Non-metallic: 0 Metallic=1.

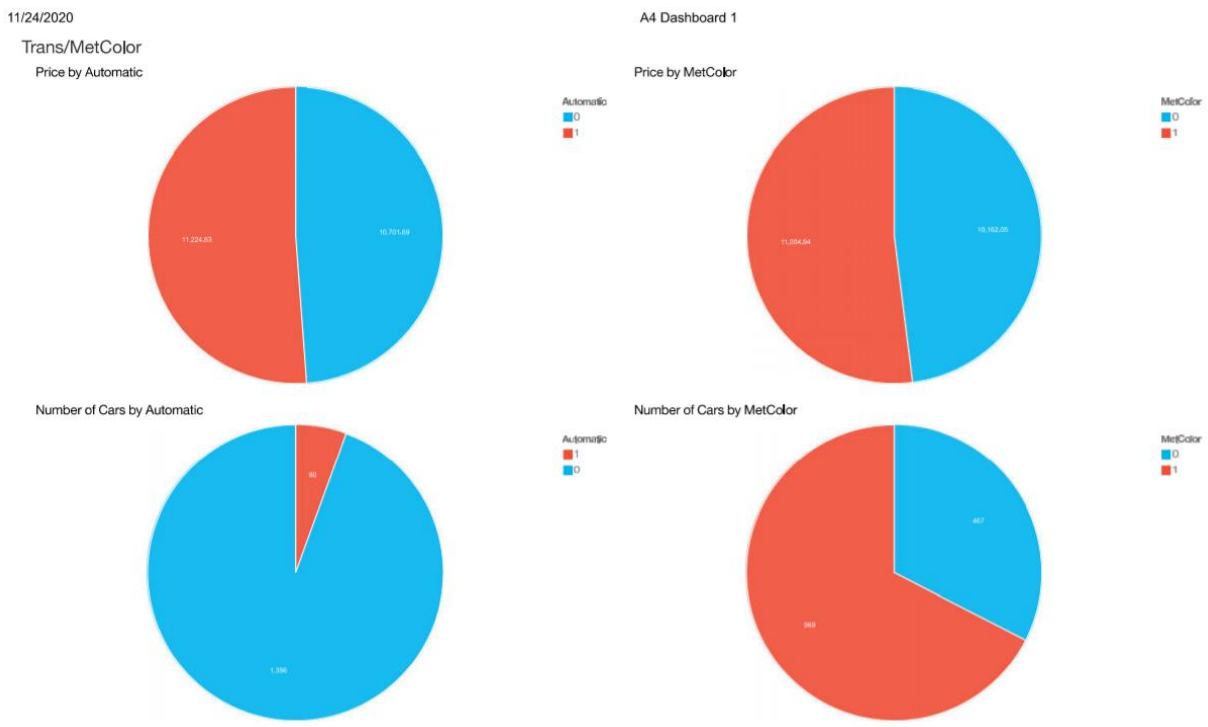

Figure *4*). Age, mileage, and KM/age were all explored. Graphics were made before and after the filters (described above) were applied to observe any increases in predictive strength. There we no major discrepancies observed from the filtered versus non-filtered graphs. There were only small variations observed expected with the filters. Interestingly, age had by far the greatest predictive strength on sales price. It was more than mileage (86% versus 41%)(which is contrary to most current literature). This could be due to the sample size being too small. For example, it's unknown if this data is from a particular region where driving patterns are irregular. It could be that these cars tend to drive less in general and so aging becomes a stronger predictor because the cars aren't driven enough to devalue them. Alternatively, this dataset of vehicles could have a unique durability where they don't senesce much with usage, but they do with time passage. Perhaps the sensors, wiring, or frame wear out faster than the engine, transmission or axles. In any case, it was also discovered that KM/age was not a predictor of price which was a surprise to

find. It's possible that no correlation was found between these because the dataset is nor large enough or because KM/age is not a valuable variable. This was explored and it was found that the distribution of KM/age was fairly skewed. Most cars (75.4%) fell into only 2 categories and the top five categories only contained 38 cars. Thus, this factor isn't able to predict price in this case because the dataset is too small. It's still completely possible that this variable is useful, but we aren't able to make that claim that from this data.

Next, age was used to predict both KM and price (**Figure 7**). It was found age had a predictive strength of 86% when predicting the price but only have a 28% strength for KM. This is surprising to find the latter because intuitively we should think that the older a car is the more it's been driven. Based on what was previously discussed, it seems confirmed this is a dataset of cars which were being driven at a rate below average. Compared to the unfiltered results, age has the same predictive strength, but KM has a 26% strength (a 2% increase occurred with filtering). Likewise, KM had a 42% predictive strength for price but a 41% strength when unfiltered (a 1% increase with filtering). While these are small differences, this would have a much larger impact in bigger datasets. This finding about mileage seems to be at odds with the previous finding about KM/age where most cars (75.4%) are being driven at a similar rate. It could be that the other 24.6% are spread out throughout the data enough that it prevents the software from recognizing the relationship. Alternatively, it could be that the outliers are high enough to seriously skew the data and prevent accurate predictions from occurring.

Subsequently, it was explored why some variables weren't linked with price as strongly. Many of the variables listed intuitively should be linked with price like HP, CC, fuel type, and weight. Fuel type and HP were the first to be investigated (**Figure 8**). It would seem reasonable that car price would vary for different types of fuel used (since diesel is a more efficient fuel) and

with different strengths of the engine (horsepower). A predictive model was made with a surprising 92.2% accuracy with only 1 variable. Fuel type could be predicted with only 4 rules concerning HP. This seemed surprising and it was discovered diesel made up only 11% of total cars. The total variance within HP and fuel type was fairly low as seen in a column graph created (**Figure 8**). In order to make solid claims about sales price for HP and fuel type, the sample size would need to be larger; there would need to be more values for diesel specifically and for each category of HP.

Similarly, weight was explored since it again seems fairly clear by intuition that the heavier a vehicle is, the more it is going to cost. Generally speaking, a heavy-duty Chevy with a turbo 2000 CC diesel engine is going to be heavier and pricier than a classic Ford F150 with a 1600 CC petrol engine. When using weight to predict price for all cars, there was only a 64% predictive strength (**Figure 9**). However, this graph was remade twice: once for all petrol cars and one for all diesel which yielded respective strengths of 75% and 86%. It was previously seen there was large variation between HP and fuel type categories; that is to say the HP for each fuel type varied substantially. So, by splitting up the two major fuel types, we're able to more accurately predict the prices. This is a key factor to keep in mind; the more simplified we can make the data, the more accurately we're able to predict the price. Comparing "apples to apples" is necessary. To recap some predictors discussed thus far: the more a vehicle weighs, the more it will cost. Further, a diesel car will cost more than a petrol car. Age doesn't always predict mileage for this dataset; however, mileage does moderately predict price while age strongly predicts price.

So back to the primary question: how do you know if you're getting a good deal? First, you need to know what characteristics are a "must-have" for you. Normally, people would start

with a budget, make, and model. But since make and model aren't listed here in the dataset, we'll pick a couple other variables. Let's say that you need to have a diesel engine, you're looking for something with around 70-75 HP, and your budget is $8,500 (**Figure *9*. Slide 8 of story: graphics showing prediction of sales price by weight for all vehicles (right), for diesel cars (top left), and for petrol cars (bottom left).**

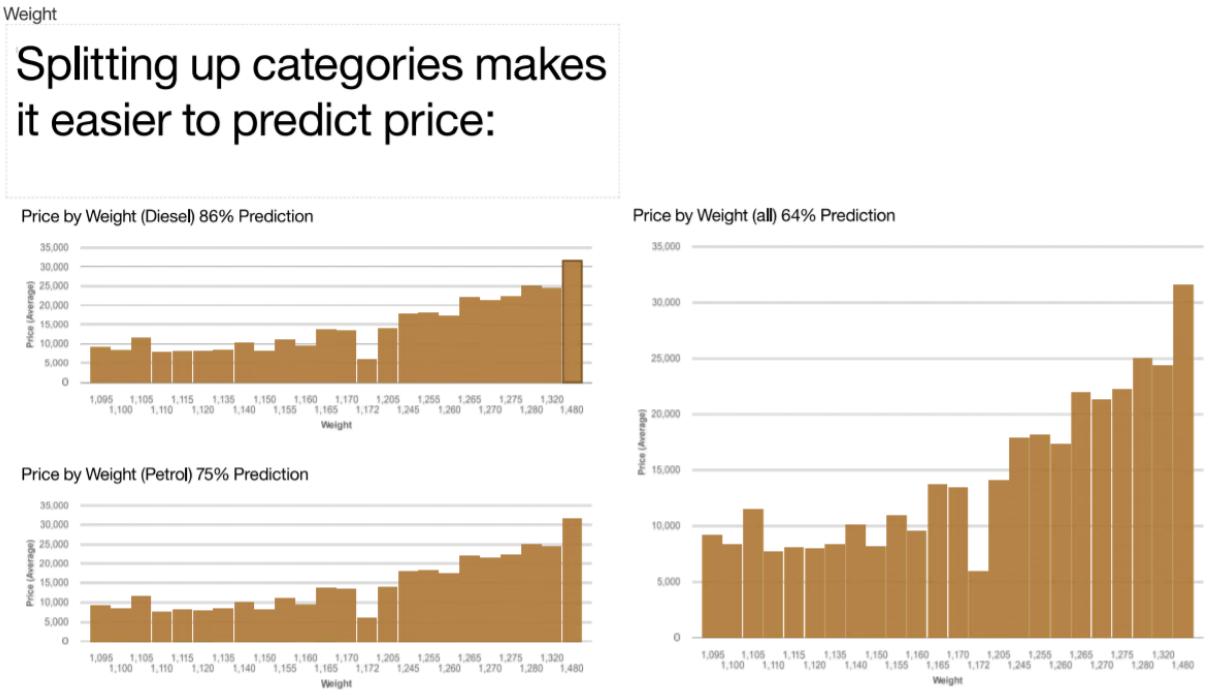

Figure *10*10). These become our filters to pare down the data. Using this dataset as the reference point, one should look at the averages across both HP and fuel type. Individually, the average prices for HP and diesel are $13,300 and $11,300. But once the filters are applied the average price for a diesel engine with 70-75 HP drops to $8,200. This shows the importance of filtering variables in order to determine average price. With these filters applied, there are 44 options to choose from. Most of them are older with the youngest being 54 months and oldest being 80 months (average of 68). This explains why the price is cheaper (because the vehicle is older). There are almost always tradeoffs in desirable characteristics and price for used automobiles.

However, if the vehicle you've found is the right make, model, and aesthetic you want and it's in good condition, then it's worth making the purchase. Sometimes price is lowered in order to make a car sell quickly. That is the type of buy you want to find; where quality is not decreased but price is. Since we don't have how long a car has been listed for, instead we can multiply the age by the price. This factor will tell us how good of a deal this car is for its age. Most of the other variables are all static with the above filters applied: CC is all 2000; most doors and weight are similar; HP is all 72. Thus, most other factors are all now normalized. Once someone has calculated this factor, they can browse the top hits and see which deal sounds best to them. The top hit in this dataset is a car where age is 56, KM is 114k, and price is $5,150. This is likely a good deal for the buyer since it has the lowest price and lowest age for the parameters specified above. In this same manner, anyone could find the best deal in a dataset of used cars. They would need to: 1. Determine the greatest predictor of sales price (mileage or age) 2. Determine the parameters that they need to have (make, model, budget, engine power, etc.) 3. Calculate the variable by multiplying the price by the greatest predictor (age or mileage) 4. Sort so that the rows with the lowest variable value are shown at the top. Then, they can look at the top results and pick which one works best for them (based on aesthetic, transmission, fuel type, etc.)

Business Application:

The findings discussed in this exploration could be useful in a few ways in the everyday business setting. Kelly Blue Book is a car pricing service that relies on everyday used-car data in order to find the averages and rate cars appropriately (Scott et al., 2011). It's a business formally centered on what has been discussed in this paper. They have to normalize for a plethora of variables: make, model, year, mileage, condition, drivetrain, transmission, door number, CC, HP,

and even accessory features like moon roofs and heated seats. Just as we have discussed here, when valuing vehicles, they have to be as specific as possible in order to be accurate. All the above variables must be taken into account in order to determine price. Interestingly, if you don't know some factors about your car, they will allow you to move forward. However, they will put a notice on your results that it may not be as accurate as possible due to that missing information. This is just like what was earlier discussed where the average from an unfiltered category will be less accurate than the average of a known category. While there are definitely some factors which are stronger than others (like age or mileage), each one makes a difference in the final value. As an example of what happens in the background when a car is being priced, 3 decision trees were made: one for all vehicles in the dataset, one for just diesel cars, and one for just petrol cars (**Figure 11**). The respectively have predictive powers of 81%, 75%, and 76%. While it may seem surprising at first that the predictive model for all cars had the greatest strength (considering the other findings), in actuality these models further the point. The "all" model takes into account 1. Age 2. Weight 3. HP 4. Doors. The other two models only take into account 1. Age 2. KM (diesel) or 1. Age 2. KM 3. HP (petrol). As previously mentioned, the more variables that are assessed, the more accurately the model can predict the price. In the case of Kelley Blue Book, they are sure to assess every relevant factor.

When using stories to present data in an organization, there are some major pitfalls to watch out for. Stories puts an emphasis on narrative, presentation, aesthetic, and emotion. They're a tool to make data intriguing and sexy (since talking about raw numbers and R^2 values can easily make people yawn). However, problems can occur when the story becomes the driver instead of the presentation tool. Data science must remain objective at its core with its proprietors focusing primarily on methodology and secondarily on narrative. It's well-recorded

that narrative without rigorous methodology will result in bad business decisions (Armerding, 2013). McNulty (2018) gives data scientists 3 specific warnings around this topic: always be driven by a question not an objective; build the narrative only after the results have been validated; and ensure scientists are putting the objectivity of analysis before aesthetic or storytelling. Data scientists must remain vigilant to ensure business objectives or outcomes are not driving their analyses. Storytelling is a powerful tool making the analysis interesting and making an appeal to the "pathos" of an audience. When you can get your audience to laugh or feel sad, they'll care more about the data and its implications. Our minds are more likely to remember story-form facts (data derived from a story) than raw facts. Stories are engaging and ideally draw audiences not just to consume the information but to be participants.

Next Steps and Closing Thoughts:

There were several important variables not listed in this dataset which would have significantly helped predict prices. One of the most important missing features was "condition". While this variable might be a bit subjective, this is a variable commonly used by Kelley Blue Books one of the foremost used car pricing services and other pricing services (Scott et al., 2011). Make and model of a car can also show interesting insights to peoples' purchasing decisions. These along with title status and drivetrain would certainly have impact on consumers' choices and should be garnered for future datasets. However, with this dataset, we have shown age is by far the primary factor driving price followed by mileage. We have shown how consumers can take the attached data and make comparisons to the averages for each category to know if a used car being offered is a good deal. Lastly, we showed how these principles are regularly used in the business setting where Kelley Blue Books routinely predicts prices based on a score of variables. Fully equipped with this knowledge, buyers have the know-how to explore

datasets, peruse current listings for used cars, and figure out what might be the best deal for them.

References:

Armerding, T. (2013). Big data without good analytics can lead to bad decisions. https://www.infoworld.com/article/2611729/big-data-without-good-analytics-can-lead-to-bad-decisions.html

Engers, M., Hartmann, M., & Stern, S. (2009). Annual miles drive used car prices. Journal of Applied Econometrics, 24(1), 1-33.

McNulty, K. (2018). Beware of 'storytelling' in data and analytics. Towards Data Science. https://towardsdatascience.com/beware-of-storytelling-with-data-1710fea554b0

Scott, D. C., Vickers, S., McBride, T., & Wansolich, B. (2011). U.S. Patent Application No. 13/023,326.

Appendix:

Figure 1. Screenshot of dataset of used cars.

KM/Year	Row Id	Price	Age	KM	FuelType	HP	MetColor	Automatic	CC	Doors	Weight
2042.8695652173913	1	13500	23	46986	Diesel	90	1	0	2000	3	1165
3171.1739130434785	2	13750	23	72937	Diesel	90	1	0	2000	3	1165
1737.9583333333333	3	13950	24	41711	Diesel	90	1	0	2000	3	1165
1846.1538461538462	4	14950	26	48000	Diesel	90	0	0	2000	3	1165
1283.3333333333333	5	13750	30	38500	Diesel	90	0	0	2000	3	1170
1906.25	6	12950	32	61000	Diesel	90	0	0	2000	3	1170
3504.1481481481483	7	16900	27	94612	Diesel	90	1	0	2000	3	1245
2529.633333333333	8	18600	30	75889	Diesel	90	1	0	2000	3	1245
729.6296296296297	9	21500	27	19700	Petrol	192	0	0	1800	3	1185
3092.9565217391305	10	12950	23	71138	Diesel	69	0	0	1900	3	1105
1258.44	11	20950	25	31461	Petrol	192	0	0	1800	3	1185
1982.2727272727273	12	19950	22	43610	Petrol	192	0	0	1800	3	1185
1287.56	13	19600	25	32189	Petrol	192	0	0	1800	3	1185
741.9354838709677	14	21500	31	23000	Petrol	192	1	0	1800	3	1185
1066.59375	15	22500	32	34131	Petrol	192	1	0	1800	3	1185

Figure 2. Slide 1 of story: introductory question.

Figure 3. Slide 2 of story: Graphs showing average price separated by transmission type and whether a car has a metallic color (top). Graphs showing proportion of cars per transmission type and per having a metallic color or not (bottom). Transmission: Manual=0 Automatic=1. Metcolor: Non-metallic: 0 Metallic=1.

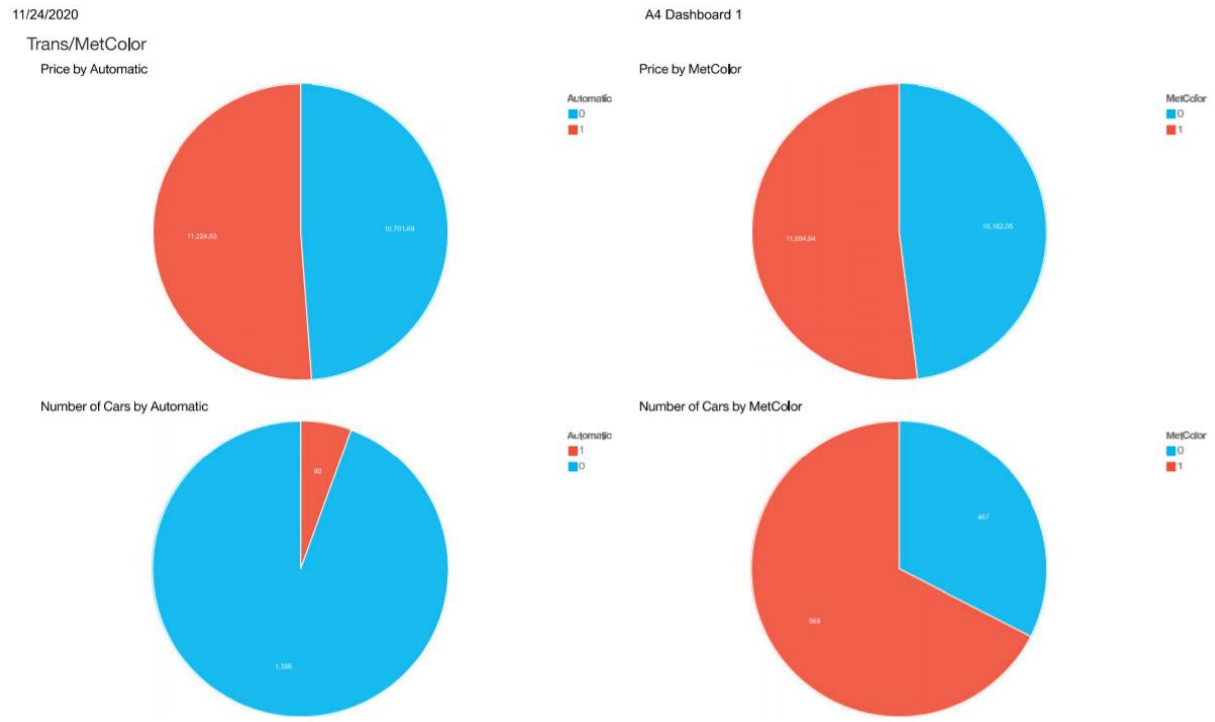

Figure 4. Slide 3 of story. Graphs showing different categories needing filtering. Automatic vehicles: 80; CNG vehicles: 17; 2-Door vehicles: 4.

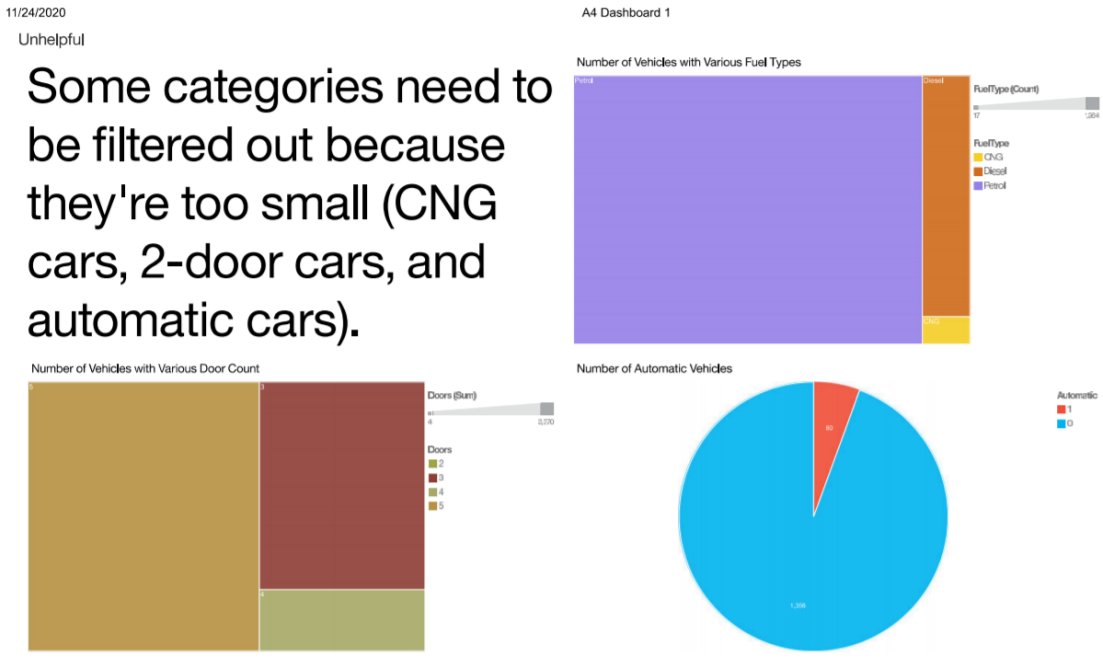

Figure 5. Slide 4 of story: Top left: graph shows price as a function of age. Top right: graph shows price as a function of mileage. Bottom left: graph shows distribution of the variable mileage per year. Bottom right: graph shows price as a function of the variable mileage per year.

Figure 6. Slide 5 of story: Top left: graph shows price as a function of age. Top right: graph shows price as a function of mileage. Bottom left: graph shows distribution of the variable mileage per year. Bottom right: graph shows price as a function of the variable mileage per year. Results have applied filters described in Figure 4.

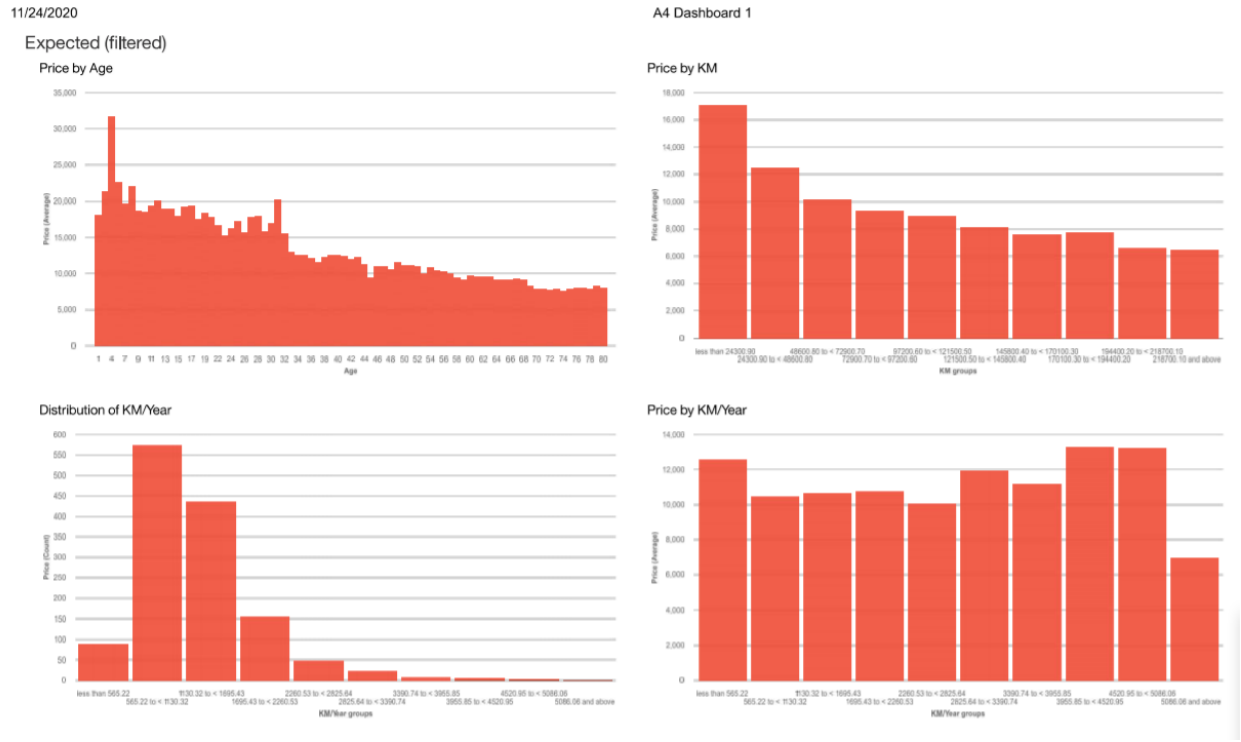

Figure 7. Slide 6 of story: Graphics showing price and mileage as a function of age (left – unfiltered; right - after filters were applied).

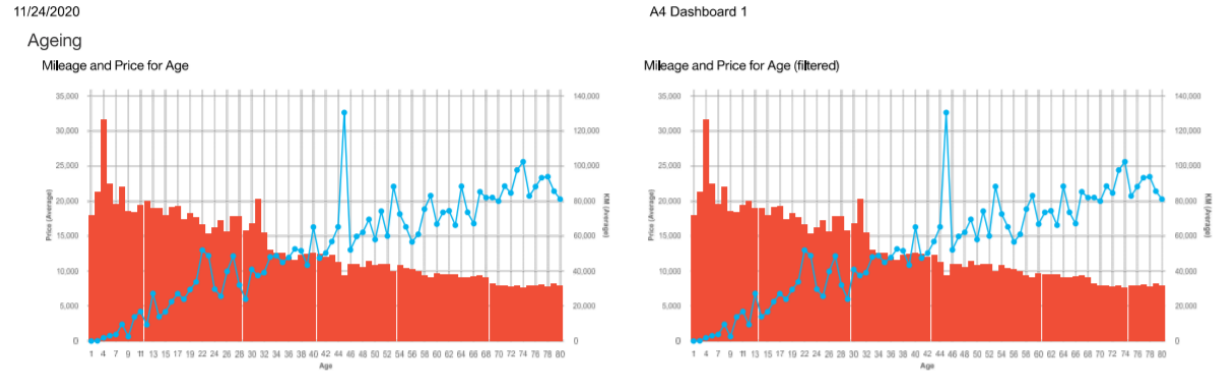

Figure 8. Slide 7 of story: graphics showing prediction of fuel type using HP (below) and distribution of cars by HP and fuel type.

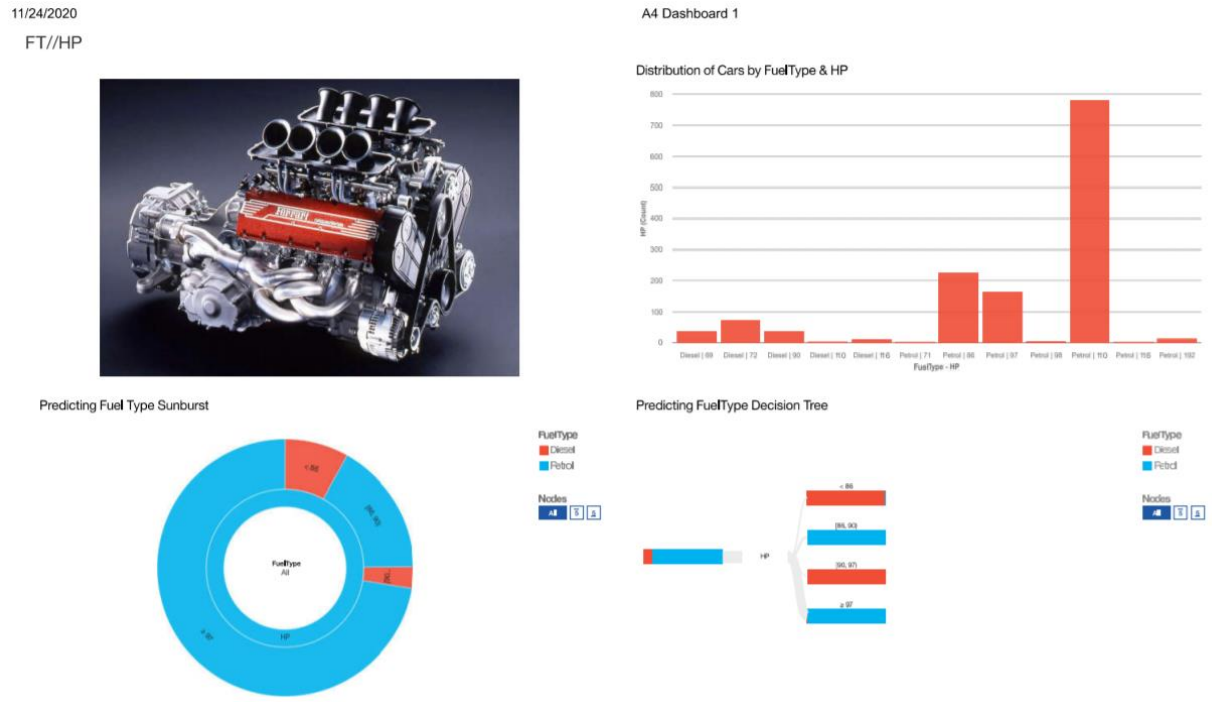

Figure 9. Slide 8 of story: graphics showing prediction of sales price by weight for all vehicles (right), for diesel cars (top left), and for petrol cars (bottom left).

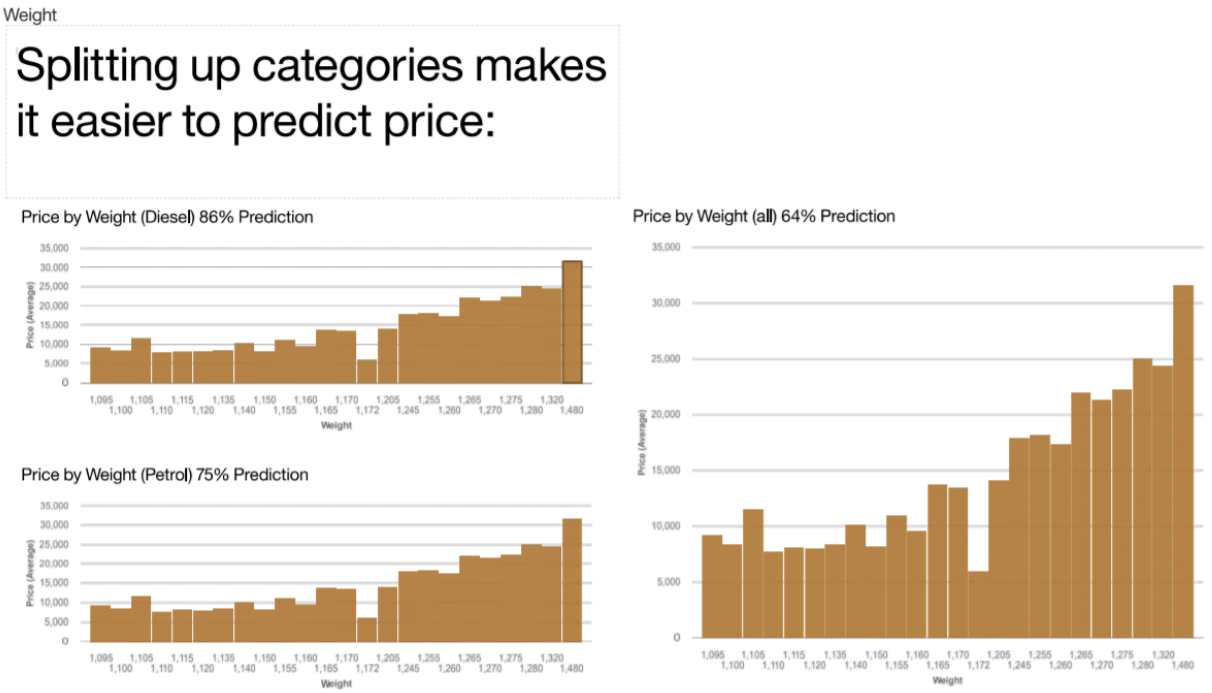

Figure 10. Slide 9 of story: graphics showing average prices of fuel type (top) and HP (below).

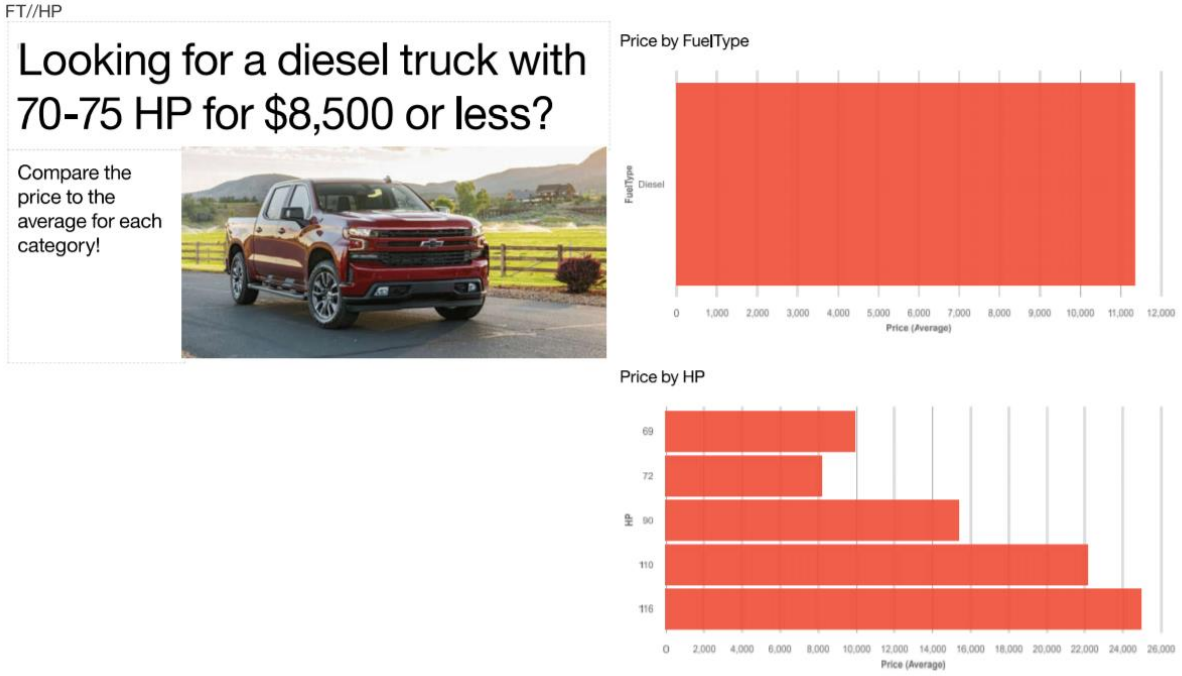

Figure 11. Slide 10 of story: graphics showing prediction models for price (all), for diesel cars (bottom left), and petrol cars (bottom right).