CARS DATASET EXPLORATION

DATA610: Fall 2020
Assignment 5
Nick Arndt-Kohlway, Cynithis Simpson-Booth, Jamie Dalrymple, Ted Fitch, Elaine Labach, Yobell Woldu
Dr. Laila Moretto

Introduction:

For assignment five, groups were given various datasets to choose from to perform predictive analysis on. For group one, we chose the focus of our research on cars, specifically how much fuel they utilize. To begin, fuel efficiency is a type of measurement of how vehicles can use a source of energy to travel specific distances according to the make, model and year of the vehicle. Analyzing fuel efficiency is very important since in 2018, there were over 273 million cars registered in the United States and in 2016, there were 6.3 million cars sold as well ("Number of cars in U.S.," n.d.). Tony Markovich (2020) stated that

> *"Fuel efficiency is important from a personal standpoint, as well as from a global view. The macro perspective sees that a fuel-efficient vehicle benefits the planet by using less gasoline and releasing less harmful gas emissions into the atmosphere. From an individual perspective, a fuel-efficient vehicle benefits its owner by reducing trips to the gas station, which keeps money in your pockets".*

Based on this information, our research will contain the analysis of variables that pertain to the characteristics of American driven car models. This document will address the data contents and the data exploration done to the data set that was chosen. We will then conduct our analysis using IBM Cognos Analytics software features such as visualizations to discover and present our fuel efficiency results. Additionally, we will go over a created decision tree predictive models developed and the results to include the top five rules and how they affect our findings. We will also compare the displays that were created to find any other correlating information to help with this analysis. There will be a video at the end that will present our work

in a story format for viewing. Lastly, we will individually provide examples on how this analysis would assist our respective organizations.

Dataset Exploration:

The dataset being explored contains 18 variables and 5076 rows. The variables include examples of categorical variables (driveline, engine type, fuel type, ID, make, model year), continuous variables (city MPG, highway MPG, height, horsepower, length, number of forward gears, torque, width, year), and binary variables (transmission type, hybrid). There are no missing values in this dataset. When examining the variables, it was noticed only 63 cars were listed as hybrid. Furthermore, there were only 2 entries using compressed natural gas and 27 entries using diesel. In addition, the dataset only contains vehicles from 2009 to 2012 with only 1% of the vehicles being from 2009. It was decided to filter out hybrids, diesel, and compressed natural gas vehicles but keep the 2009 vehicles because it is less likely that the MPG or HP will change dramatically year over year; however, fuel type and a hybrid engine will affect MPG and HP dramatically. It was also noticed one vehicle incorrectly had a highway MPG listed as 223 when it should be 23 (U.S. News Best Cars Staff, 2015); this was corrected. It is unknown what unit the dimension variables are in or what they are describing (whether engine size, car size, etc.). Since these values range from 1 to 255, it is unlikely they are representing feet or inches; the most likely unit is centimeters. It's important to note that whenever a low value appears, the other two dimensions are always much higher. So it's also possible some measurements were listed in different dimensions (e.g. the high dimension could be CM and low could be M). In order to consolidate the dimension values, a new value was made, volume (length * width * height), and will be used in the analysis. It is presumed that the dimensions are gestalt; they matter more together than individually. The driveline category is relatively well-balanced with

no major outliers; all four categories have more than 800 values. With the filters applied, the dataset then contained 4,984 values. Finally, a calculation was performed taking the average of highway MPG and city MPG. This was done for simplicity's sake so that one variable can be used for prediction analysis. This data exploration will focus its efforts on predicting horsepower (HP) and MPG (as average of both city and highway). What factors are correlated with a high MPG? What drives HP? Are there certain factors that matter more to MPG than to HP (and vice versa)? It is presumed that factors like torque and HP will be correlated together and while these will be inversely correlated with MPG. This is predicted since the more power a system has, the less energy efficient it tends to be (thus it's predicted that the more HP a car has, the lower the MPG will be).

Dataset Visualization:

The first visualization is a spiral diagram that utilizes horsepower as the target variable (**Figure 1**). The highest predictor of horsepower at 91% correlation is a combination of torque and make of the vehicle. This is expected since torque is typically highly correlated to horsepower. The top 7 drivers of horsepower all include torque as a predictor. Mathematically, horsepower is equal to torque multiplied by Revolutions per Minute (RPM) then divided by a constant. Therefore, to generate more horsepower it is inherent that torque or RPM would increase as well. It also makes sense that the make of the car would be a high predictor of horsepower as well. Most car manufacturers only use a handful of engines that are usually based on each other. For example, Infiniti used to have a 2.5L, 3.5L, 3.7L, and 4.5L that are mechanically exactly the same, just with different displacements. Furthermore, American manufacturers are more likely to use larger displacement engines to generate a higher horsepower output. On the other hand, foreign manufacturers use smaller engines that might take

advantage of turbochargers to lessen the gap of horsepower output compared to American manufacturers. The second highest predictor of horsepower was torque and transmission at 84% which has the exact same correlation as torque and width. It makes sense that torque and transmission is a high predictor since the transmission will determine horsepower at different RPMs. However, the width of the car having the same strength of correlation as transmission caught me by surprise. Looking deeper into that variable it might make sense that wider vehicles tend to have higher outputs since they're produced for performance. Most supercars/hypercars are wider and lower to create better aerodynamics when racing. And with that increase in width comes a higher performance engine.

The second visualization is a bar and line graph that shows horsepower and MPG based on different drivelines (**Figure 2**). Front-wheel drive had the highest MPG as expected since a front-wheel-drive system is lighter and provides better traction to the front wheels due to the weight of the engine. Front-wheel drive also had the lowest average horsepower which also helps explain the increase in MPG. Rear wheel drive had the highest average horsepower which is also understandable since many performance cars are built-in rear-wheel drive. Four-wheel drive had the second-highest average horsepower possibly explained by the heavier weight of the 4WD system and the many off-road utilities that these larger SUVs encompass. What was most surprising to me was that AWD had the second-highest MPG. I figured RWD would be up there in MPG due to how light the driveline is compared to AWD and 4WD. However, what RWD gains in a weight decrease it loses in poor traction which might explain how poorly it performed in terms of MPG.

The third visualization is another bar and line graph that shows average MPG and average horsepower based on the make of the vehicles (**Figure 3**). Ferrari had the lowest average

MPG at a measly 12.5 MPG, which can be explained by Ferrari's propensity to build high-performance cars that typically utilize RWD. The average horsepower for Ferrari was 540. Combined with the RWD and light weight, this creates a lot of opportunity for shredding wheels due to terrible traction. Mini had the highest average MPG at 30.2 and a very low average horsepower at 160.97. Mini is able to achieve the highest average horsepower due to the light weight and utilization of front-wheel drive. The better traction and usage of small 3 cylinder engines in their cars puts it above and beyond other manufacturers in terms of fuel economy. Make of the car moderately drives MPG (36%).

The fourth visualization is a tree sunburst with the target variable being the number of forward gears (**Figure 4**). The highest average number of forward gears with 6.74 was when the make was high-end car brands such as Audi, BMW, Rolls-Royce, Lexus, Mercedes-AMG, Mercedes-Benz, and Infiniti that had automatic transmissions. These cars accounted for 7% of all cars in the dataset. The second-highest average number of forward gears was 6, which included all the high-end car brands stated above, but instead had a manual transmission. The highest and second-highest number of forward gears accounted for 9% of the cars in the dataset. The third highest average number of forward gears was 5.98 which consisted of Chevrolet, Nissan, Toyota, GMC, Suzuki, Grand Cherokee, and Chrysler that had a horsepower greater than or equal to 328. These cars made up 6% of all cars in the dataset. It makes sense that the highest number of forward gears will be in higher-end car brands designed for performance and/or luxury. It's also understandable that the manual version of the above cars would have fewer gears than the automatic since it would get tiring when flipping through gears constantly.

The fifth visualization is a comparison of two pie graphs (**Figure 5**). One pie graph is a breakdown of horsepower by fuel type and the other pie graph is a breakdown of MPG by fuel

type. The average horsepower for E85 is 306.31 compared to 266.93 for regular gasoline. The MPG for E85 was 12.94 compared to 21.38 for gasoline. This is as expected since E85 is more resistant to detonation and burns much more quickly than gasoline. These two factors account for an increased output of horsepower at the expense of reduced fuel efficiency. To further this analysis, two more pie graphs were created that compared highway and city MPG by all fuel types (**Figure 6**). The high fuel efficiency among CNG vehicles is surprising since CNG typically gets lower MPG than gasoline (Compare, 2018). This could be due to the small sample size of CNG vehicles. However, diesel is the second most fuel efficient when looking at both city and highway due to the slow rate at which it burns and the increase in heat needed to detonate the fuel. This also contributes to the increased fuel efficiency for diesel vehicles. From these models we find that fuel type slightly drives highway MPG by 19%. However, fuel type drives city MPG by 22%, this 3% increase might be from driving in the city more or because there's a larger difference in city MPG among vehicles. When everyone is going 60-70 MPG on the highway, then it closes the gap among differences in MPG among vehicles.

The sixth and final visualization is a tree sunburst with volume as the target variable (**Figure 7**). The highest volume at 6,390,500 was when the make of the vehicle was either a Volvo or Mitsubishi that had an MPG greater than or equal to 29 and horsepower less than 242. This caught me by surprise since one would figure that a higher volume vehicle would have a higher horsepower output to compensate for the increase in size. And due to the increase in size and horsepower the MPG would be lower. These specifications only applied to 43 vehicles which is less than 1% of vehicles in the dataset. The prediction model only had a 30% predictive strength with make being 5 times more important than any other factor.

Predictive Models:

The Model 1 decision tree selected the calculated field "Average MPG" as its target variable which took the average of "City mpg" and "Highway mpg" (**Figure 8**). As the model shows, the diagram is initially split by "Torque" into 4 subgroups, (Torque<172), (172=<Torque<243), (243=<Torque<278), and (Torque>=278). It is quickly noted that "Torque" is considered a strong predictor due to it being the primary splitting variable. The diagram continues to split many more times, but the focus is kept on the lowest 2 groups because the top 5 target value "Average MPG" observations are found there. Subgroup (172≤Torque<243) is then split by "Make" into 3 more subgroups. One of the groups here containing (Audi, Acura, Nissan, Volkswagen, Hyundai, Kia, Scion, MINI) holds the 4th highest average MPG at 26.8. Backtracking a bit, the subgroup (Torque<172) is split by "Driveline" into 2 subgroups then by "Width" into 3 more. All 3 groups are listed very high on the average scale for mpg with 2 groups containing the 1st and 5th highest average MPGobservations at ~31 and ~26, (113≤Width<199) and (Width<113), respectively. Subgroup (Width≥199) is ultimately split by "Make" into 2 final decision nodes which are recorded as the 2nd and 3rd highest average MPG values, (Honda, Hyundai, MINI) and (Chevrolet, Nissan, Volvo, Toyota, Volkswagen, Kia, Mazda, Mitsubishi, Dodge, Jeep, Suzuki, Scion), respectfully.

Model 2 selected "Horsepower" as its target variable to predict (**Figure 9**). Just as Model 1 had started, "Torque" was used to initially split the data demonstrating it being a strong predictor. Five subgroups resulted from this split, (Torque<172), (172≤Torque<243), (243≤Torque<278), (278≤Torque<365), and (Torque≥365). The first group ends its classification while the other 4 continue on, each being split by "Make". It is important to note that the top 5 target values for "Horsepower" fall into one half of the diagram just as Model 1 developed so the

focus is once again directed on that half. Subgroup (278≤Torque<365) is split by "Make" into 3 subgroups. Of the three, group 3 (Audi, Lexus, Hyundai, Mercedes-Benz, Porsche, Aston Martin, Maserati) has the 5th highest average horsepower at 373. It also ends its classification here. Subgroup (Torque≥365) is split by "Make" as well into 4 subgroups. Three out of the four contain the 4th, 2nd, and 1st highest average horsepower with the highest sitting at 582 belonging to (Bentley, Maybach, Lamborghini). The fourth group is split by "Width" into 2 subgroups, one of the two holding the 3rd highest average value.

From the models, we can see a few similarities. Each model begins their initial development by splitting the data with "Torque". This shows that it is the single best predictor of both "Horsepower" and "Average MPG". Another observation worth mentioning is that the same variables are used to construct the diagrams with the exceptions of "Driveline" and "Fuel Type" in the first model. However, if we direct our attention to the rules, it shows that "Fuel Type" serves as a good indicator for MPG. Type 'E85' indicates a lower MPG while 'Gasoline' indicates a higher MPG. This is true for the two occurrences for "Fuel Type". The first model uses "Make" and "Width" to record the majority of the highest target values. This tells us that the width of the car can greatly determine how fuel efficient it is. This is likely because having greater width would open up more space for higher performance parts. Also, cars that are typically more affordable and common amongst middle-class communities can also serve as a fuel-efficient indicator (i.e. the groups for the 2nd and 3rd highest average MPG values; Honda, Hyundai, Chevrolet, Nissan). We can also see from model 1's rules that there is a large contrast of values after the "Torque" subgroup two splits into three "Make" subgroups. The average MPG starts at 22, then reroutes to 17, 21, and 26 depending on the "Make". This shows that the range for torque is based on the make of the car which determines the average MPG. Model 2 uses

only “Torque”, “Make”, and “Width” as the splitting variables with“Make” as the most recurring. The combination of these 3 result in an overwhelming predictive strength of 90% for “Horsepower”. One key insight discovered is that “Torque” is not only the strongest predictor of “Horsepower” but is twice as strong as any other predictor. Another key takeaway derived from the rules for model 2 is given by the “Width” split. Although width only occurs once, the values corresponding to the two groups are noticeably spaced. When Width$\geq$113, it results in a lower average horsepower. When 54$\leq$Width<113, it produces a high average and the 3rd largest target value. This rule suggests that a wider car produces lower average horsepower. From the two models, it can be concluded that “Horsepower” is an easier predictor than “Average MPG”. This is supported by the steep predictive strength for model 2. On the contrary, model 1 allowed us to discover more insights about the data while having a lower peak predictive strength. Through the use of these decision trees, we were able to answer a few of our pre-exploration questions. We determined what factors resulted in a higher MPG alongside the greatest horsepower driver.

Dashboard:

The dashboard (**Figure 10**) was created with the aforementioned visualizations in mind. Most of the visualizations dealt with horsepower and average MPG. It was decided that the dashboard would show a comparison of these variables. The first two visualizations used are the pie charts from **Figure 5**. The first showed horsepower by fuel type and the second showed average MPG by fuel type. They are placed side by side to show the comparison. As you can see, horsepower and average MPG are inversely related. E85 vehicles have a higher horsepower but a lower average MPG. Gasoline vehicles have a higher average MPG but a lower horsepower. The third visualization is a bar line graph of the horsepower and average MPG by driveline type. The fourth visualization is another bar line graph of the horsepower and average MPG by make of the

vehicle. Altogether, these visualizations show how horsepower and average MPG differ across different vehicles. You can quickly see which make, which driveline, and which fuel type has the best horsepower and the best average mile per gallon. You can also see which of those have the worst horsepower and which have the worst average mile per gallon. This information can be useful in a used car dealership. When you have many different types of cars on the lot, you can easily pull up this dashboard and show a customer that if they want the best average MPG for a gasoline car, they can choose a MINI. You can also easily show them the difference between an all-wheel drive MINI and a front-wheel drive MINI (only about 3 miles per gallon and about 22 horsepower). This dashboard can also be used for specific car brands to see how they fare against the competition. It may be well known that Ferraris have high horsepower and low fuel efficiency and that MINIs have low horsepower and high fuel efficiency. But the companies in the middle of the pack aren't as well known. Subaru may want a dashboard such as this to see how they fare against Chrysler or Volvo for instance. Also, since companies don't make one type of vehicle, the fuel type comparisons and the bar line graph of driveline would also be helpful. Similar to the old car dealership, a new car dealership would want to use this data to show potential customers their information. There are many decisions to make when buying a car. The company can show customers how they compare to other companies and additionally they can show them, the differences in the models of cars they sell. This can all be done with this dashboard. There were many attempts made to fit the visualizations in the picture of an actual car dashboard. It proved too difficult to legibly display the graphs and charts while also showing enough of the picture to see that it was a car dashboard. The colors for the dashboard were chosen based on the two colors that occurred most frequently in the car industry (Marshall, 2020).

Organization Discussion:

The original cars.csv dataset was used to better understand what factors influenced fuel efficiency or average MPG. The models that were developed focused on two important predictors: torque and horsepower. Both of these factors were inversely related to fuel efficiency. One example of how this same approach could be used in an actual organization is to better understand voluntary employee turnover in an organization. Employee retention refers to how well an organization is able to keep its performing employees within a certain amount of time, and it is measured by the percentage of employees who voluntarily leave an organization (who are in good standing) compared to total employees. High employee retention rates are important to organizations for two primary reasons: retention improves the organization's financial health (avoids costs associated with hiring and training new employees) and retention improves business relationships (reduces negative impacts that a departing employee may have with suppliers and customers). By analyzing human resource data (such as employee tenure, salary, salary increases, type of position, training received, compensation, work location, commute time/distance, time in current role, experience/education, stress, and others), important factors of turnover rate can be identified (Mizart, 2018). Identifying important drivers can help organizations craft more effective employee retention strategies. For example, if commute time is a major driver for retention, organizations may be able to offer remote working options to its employees. This would be determined in the same method outlined in this paper; first, the data would be explored to filter outliers, assess the variables, and address missing data. Insights would be developed based on visualizations (one example would be to graph average retention percentage over time for this company versus a competitor). A decision tree model could be used to see what variables best predict employee retention. Finally, dashboards and stories would be

an effective way to communicate the insights to anyone regardless of their analytic background. This can be especially useful to explain the technical results to executives so they can effectively make decisions because dashboards and stories show and tell the data in an intuitive and minimalist way that doesn't require an analytic background.

Conclusion:

Through the analysis of this "Cars" dataset, we've learned a few things:

- Torque is correlated with HP and inversely correlated with MPG
- Torque and make are the best predictors for HP and MPG
- The driveline with highest fuel efficiency is FWD
- High-end vehicles (Ferrari, Bently, Maybach, Mercedes) tend to have the highest HP and lowest MPG
- Gasoline has much high fuel efficiency than E85 (but similar HP on average)
- Volume could not be correlated strongly with any the variables

This helps us answer our original questions of what factors are correlated with MPG and HP. People looking to purchase a vehicle with high MPG should look for: a non-high performance, FWD, gasoline powered, and low torque vehicle. On the other hand, people looking to purchase a car with high HP would look for: high width, high performance make (Ferrari, Maybach, Lamborghini), RWD, gasoline powered, and high torque vehicle. Further exploration would include understanding the dimension variables and finding differences between similar "Make" vehicles (i.e. comparing what drives MPG and HP among high-performance vehicles like Ferrari, Lamborghini, Mercedes).

References:

Compare (2018). Natural Gas Vehicles: Why Aren't We Buying Them? *Compare.* Retrieved from: https://www.compare.com/ways-to-save/vehicle/natural-gas-vehicles-guide

Markovich, T. (n.d.). The Most Fuel-Efficient Cars in 2020. Retrieved December 10, 2020, from The Drive website: https://www.thedrive.com/cars-101/36546/most-fuel-efficient-cars

Marshall, A. (2020). Color psychology: The logo color tricks used by top companies—and how to design your own. Retrieved from https://www.canva.com/learn/color-psychology-the-logo-color-tricks-used-by-top-companies/

Mizart, S. (2018, November). Using Predictive Analytics in Employee Retention. *Financial Management.* Retrieved December 12, 2020 from Financial Management website: https://www.fm-magazine.com/issues/2018/dec/using-predictive-analytics-in-employee-retention.html

Number of cars in U.S. (n.d.). Retrieved December 10, 2020, from Statista website: https://www.statista.com/statistics/183505/number-of-vehicles-in-the-united-states-since-1990/#statisticContainer

U.S. News Best Cars Staff (2015). *U.S. News*. https://cars.usnews.com/cars-trucks/chevrolet/silverado-1500-hybrid/2011

Appendix:

Figure 1. Spiral diagram with Horsepower as the target variable.

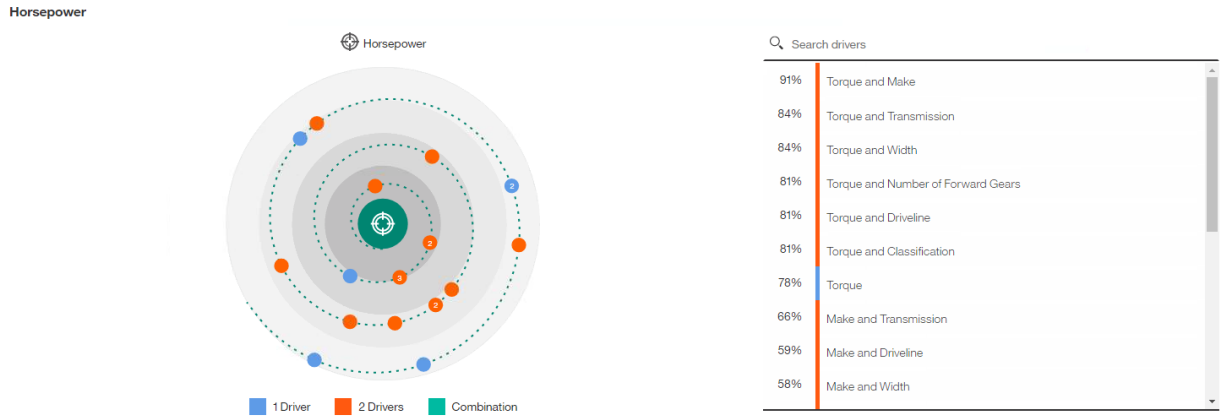


Figure 2. Bar and line graph showing different levels of MPG and Horsepower for each Driveline type.

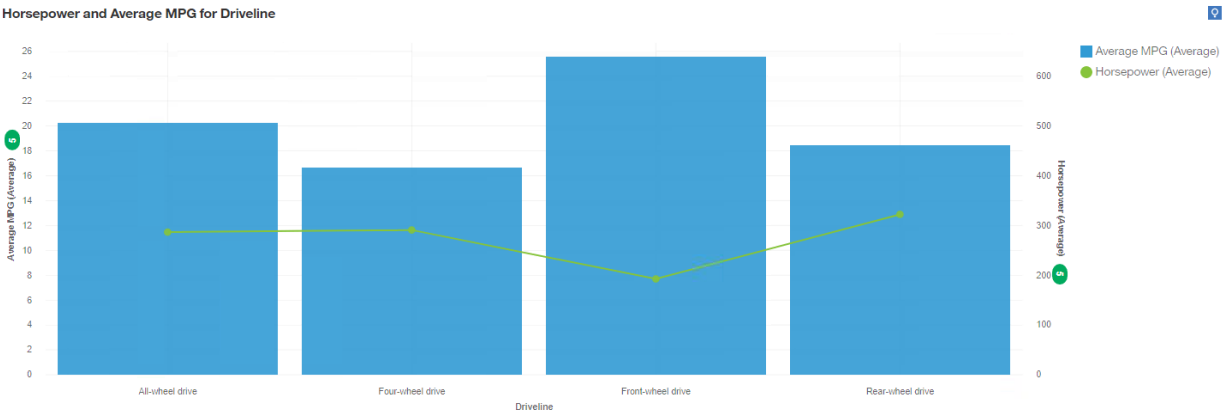


Figure 3. Bar and line graph showing different levels of MPG and Horsepower for each vehicle make type.

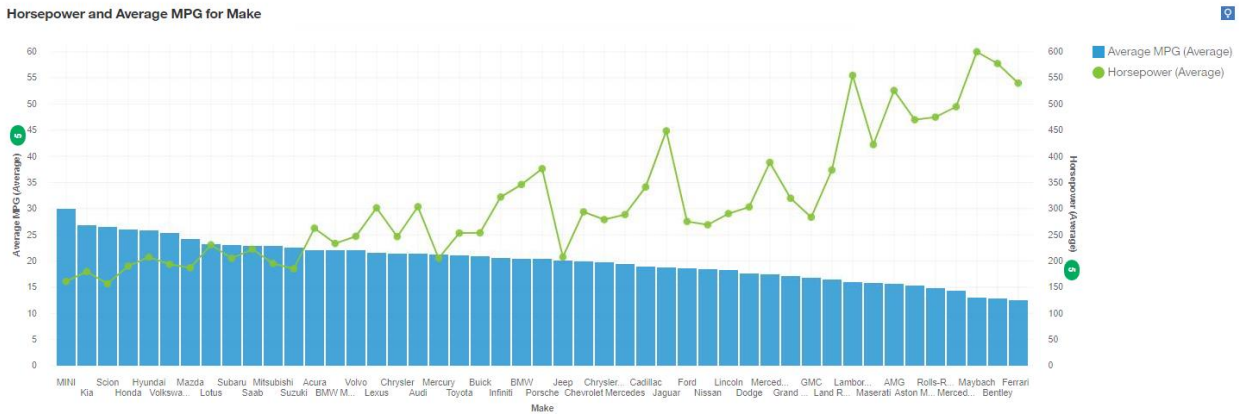


Figure 4. Sunburst diagram with the target variable being the number of forward gears.

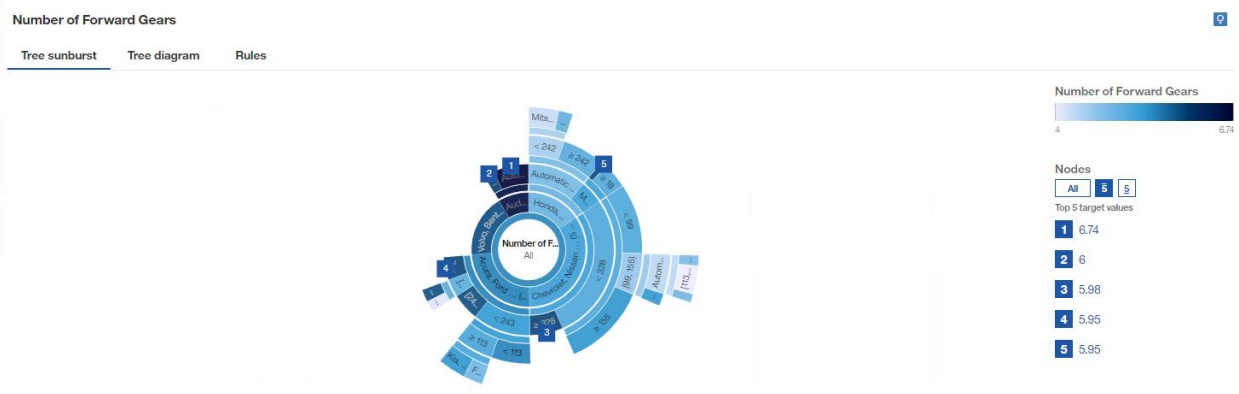
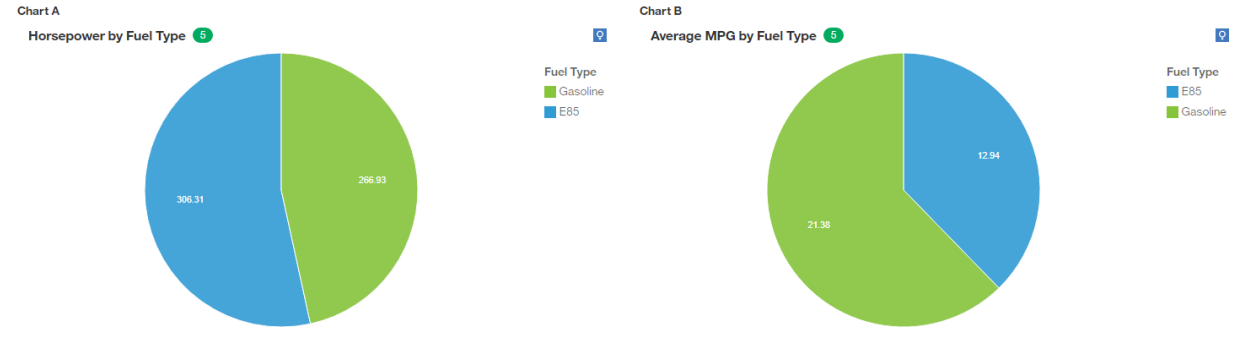
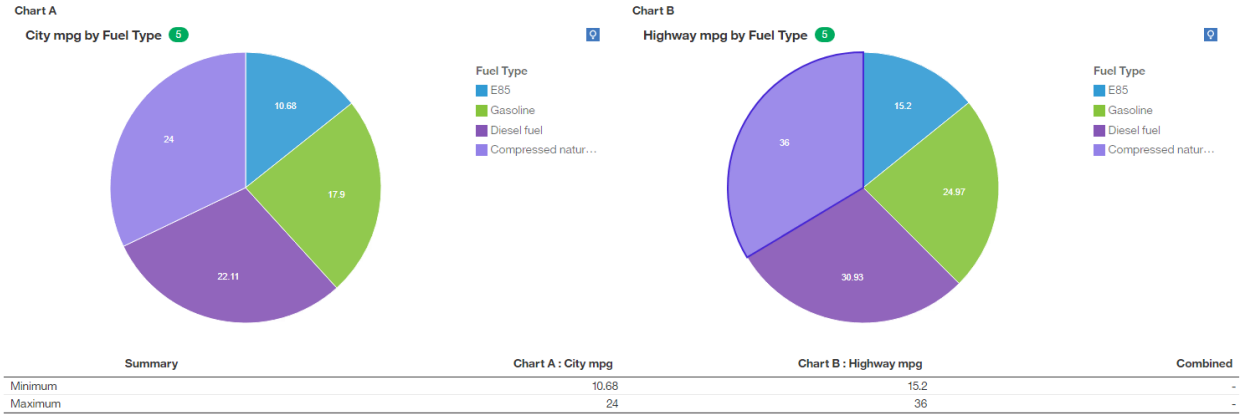


Figure 5. Two pie charts showing: Horsepower by fuel type (left) and MPG by fuel type (right).



Summary	Chart A : Horsepower	Chart B : Average MPG	Combine
Minimum	266.93	12.94	
Maximum	306.31	21.38	

Figure 6. Two pie charts showing: City MPG by fuel type (left) and Highway MPG by fuel type (right). Data was unfiltered to include all fuel types.



Summary	Chart A : City mpg	Chart B : Highway mpg	Combined
Minimum	10.68	15.2	-
Maximum	24	36	-

Figure 7. Sunburst diagram with volume as the target variable.

Figure 8. Decision tree with Average MPG as the target variable.

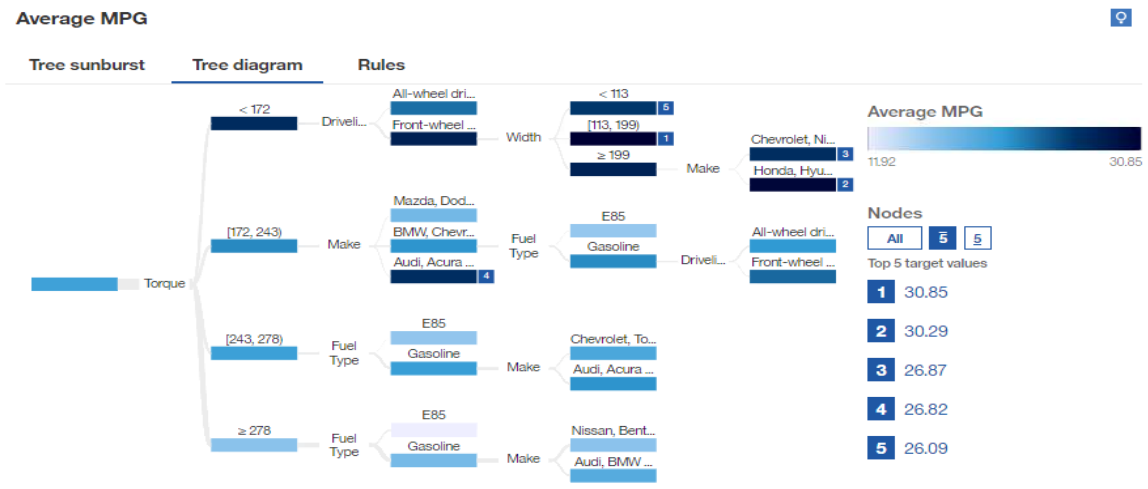


Figure 9. Decision tree with Horsepower as the target variable.

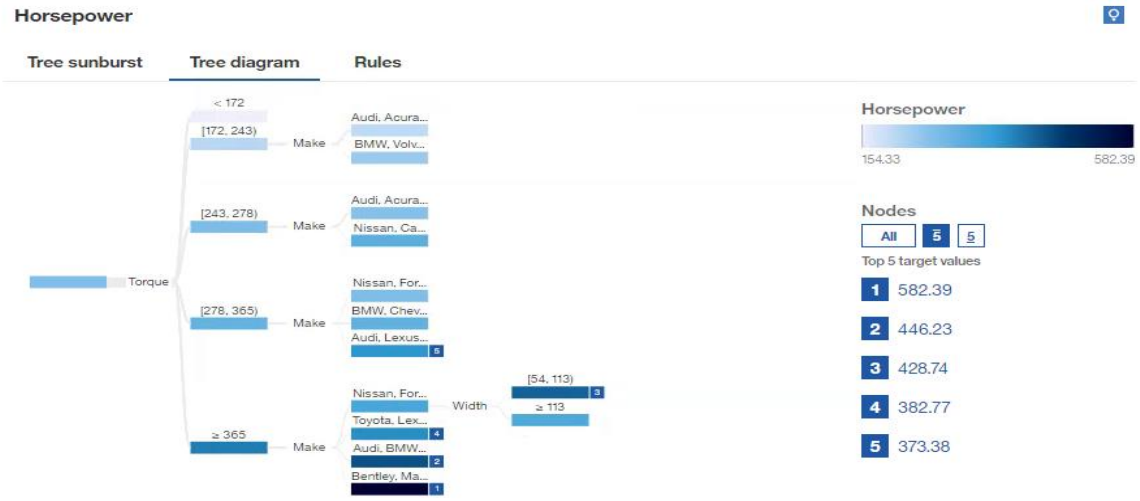


Figure 10. Horsepower and MPG dashboard