

A Network Analysis of Car Shopping

*Stats-R-Us: Nayib Gloria, Michelle Moffa, Daniel Pak, Emelina Vienneau, Wuming Zhang
Duke University, Class of 2018*

As consumers with next to no knowledge about cars ourselves, we approached edmund.com's challenge to make shopping easier with a simple idea: what new feature would make edmunds.com more helpful for us? We concluded it would be helpful to present us with the car options most relevant to what we desire with limited prior car knowledge. By utilizing R and Matlab to identify connections between cars, we identified three ways edmunds.com could offer website visitors valuable information to enhance their shopping experience: links to cars "Viewed by People Who Also Viewed This Car," "Also Viewed by People Who Bought this Car," and "Cars Similar to This Car." Providing links to other cars that might interest visitors when uneducated visitors may not have found those cars themselves is crucial to making the shopping experience easier. To visualize these connections, we created networks of cars based on the given data.

Our first visualization shows a network connecting cars based solely on views. We isolated the visitor keys of people who made transactions, then made a list of all of the make/model combinations that those visitors viewed on edmunds.com. We then constructed a matrix using this data to count how many times every car make/model was connected to every other car make/model in order to generate our visualization. The network shown is based on connections made by over 0.5% of buyers. We observed that clusters form to show car combinations most frequently viewed together, which can show what factors are most common to all of the cars customers are viewing. For example, by color-coding the points in the network according to car make and car body type, we can see that body type is far more important to customers than make is. By looking at the outliers, we can also see what factors are less sought after by customers.

Our second visualization takes transactions into account as well, this time connecting the cars bought by a visitor to the cars that same visitor viewed, using the same process we used to create our first network. The network shown is based on connections made by over 0.05% of buyers. Again, the existence or lack of clusters can be used to show what factors do or do not play a role in the cars a consumer views, as shown by color-coding by make and body type.

Our third visualization is a network of connections based not on viewer data of make/model combinations but on actual individual bought cars themselves. By looking at ten different factors, we were able to generate algorithms to calculate similarity scores between a random subset of 200 of the individual bought cars. The factors we looked at included: was the vehicle a motorcycle, was the vehicle a truck, was it a large car, was it a medium car, was it a small car, the price paid, the transmission, was the car new, the year manufactured, and the mileage. Weighing these factors in various ways leads to vastly different networks. In the future, the weighting system and information factored into this algorithm can be determined based on information gleaned using visualizations of the first two types we displayed. The visualizations shown display all connections with similarity scores of above one standard deviation above our mean similarity score. Color-coding this visualization based on car body type shows how our different weighted similarity score algorithms contributed to clustering. The radius of cars around any particular starting point can be used to pinpoint the most similar cars to the initial.

This project is a proof-of-concept of the information that can be gleaned and the service that edmunds.com can provide using a network analysis of car similarities. A quantitative clustering analysis of visualizations of extrapolated and more detailed versions of our first and second type can be used to make suggestions to consumers and to determine what factors are present within buyers' viewing histories. This information could then be used to generate similarity score algorithms between

individual cars being sold on edmunds.com in order to provide suggestions for customers to make the car shopping process easier and more efficient.