Finding Your Ideal Car

Avani Badugu, Mehar Saini, John Spurrier

Decision and Action Insight/Audience

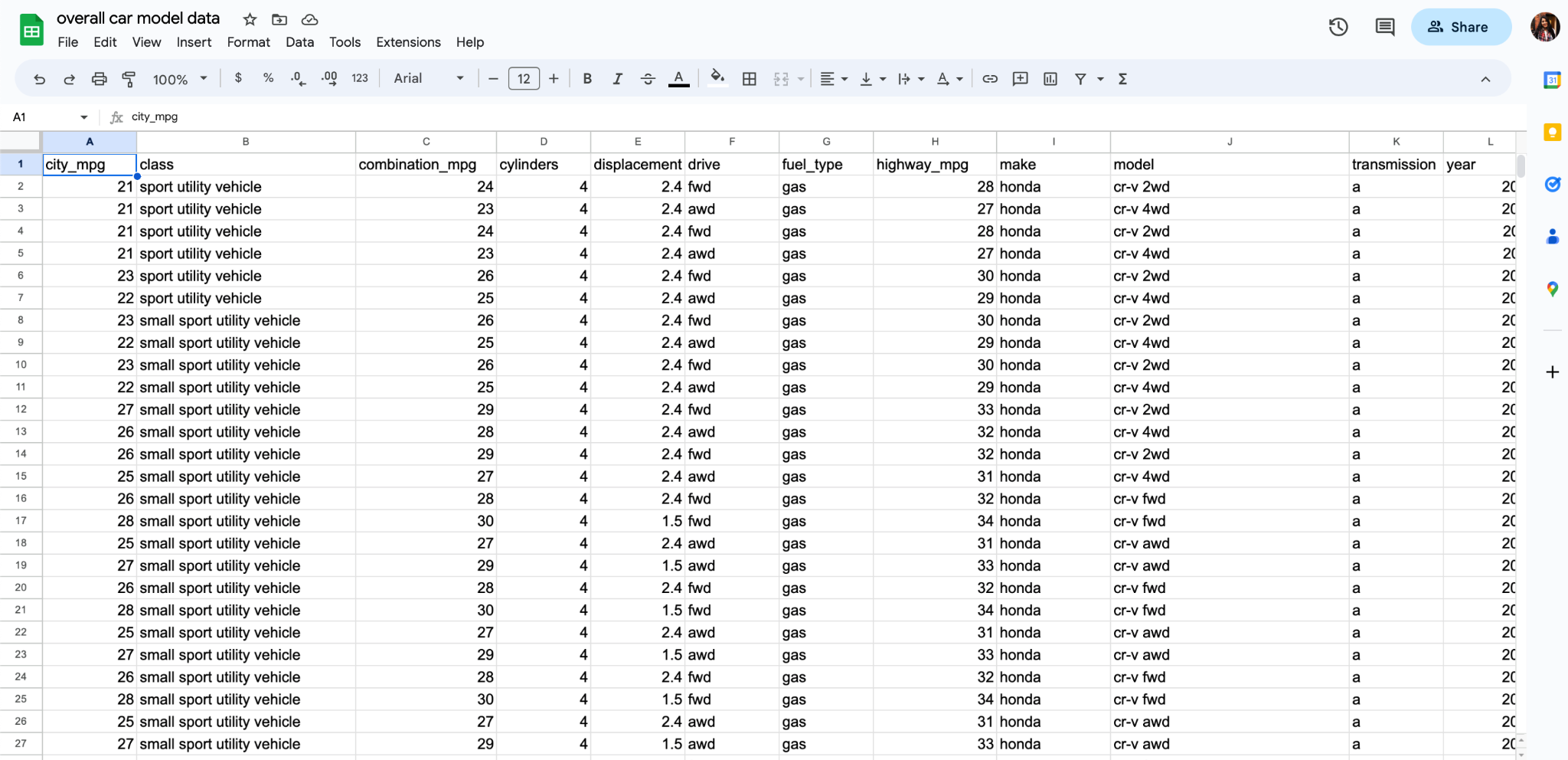
A common problem that people often have is deciding on which car to purchase. Often, they have minimal knowledge about all of the different factors that go into purchasing a car. What sparked this idea is that within our group many of us are moving to different areas for jobs post graduation and we are in the process of looking for cars. We realized that we do not really know much detail about cars and what factors to think about when buying one. In this case we wanted to explore the relationship between efficiency and cost of cars. To measure the efficiency of a car we looked at miles per gallon (mpg). Our target audience is individuals who are looking to buy cars. This can include all people from ages 16 to 65 as everyone needs a car and has certain criteria that they seek out. Our audience needs to know what specific criteria they prefer like their preferred price range, difference between highway mpg and city mpg which is based on their location, and their driving habits.

This core insight that we intended to discover from the data would allow us to make recommendations to the community on which car model is most worth its price. We wanted to focus on a specific class of car, such as an SUV. We decided on SUVs because we feel they are getting more and more popular, and our whole group all drive SUVs. The car models that we have chosen to assess are 10 types of SUV models, which are Toyota RAV-4, Honda CR-V, Toyota Highlander, Ford Explorer, Nissan Rogue, Jeep Wrangler, Chevrolet Equinox, Jeep Grand Cherokee, Ford Escape, and Honda Pilot. From conducting research, we decided on these 10 cars because they were among the most popular/best selling SUVs in the United States. We believed that focusing on one specific class would provide a better analysis than comparing cars from different classes because other classes could give us a different insight.

Data Collection

The data collection process had two parts. The first part was we obtained the data by using a HTTP get request given to us by the [API Website](https://api-ninjas.com/api/cars) itself. We found this API by researching and looking for API’s about cars that contain data about the efficiency of a car such as the make, model, highway mpg, city mpg, transmission type, etc. Using Jupyter Notebook, we were able to use the “requests.get” module to pull the data using the API url and a given API key. Along with this, we imported the json and pandas library to be able to convert the data into a json format and then put the data into a dataframe. With this API dataset, we pulled data for certain car makes and models and added the data into a dataframe everytime we made a request. The API itself has very large amounts of data, but only allows users to pull out 50 entries per request made. For each make and model we identified previously, we extracted 10 years of data for each; 2010-2020. Once we collected the 10 years of data for each car, we converted the data to a CSV. Once we finished collecting data for all of the cars, we combined all of the CSVs into one large overall CSV with all 10 models. The file size for this csv document is 43 KB. Including all of the data we have collected, we have over 500 records of cars. The data itself has 13 columns of information about each car.

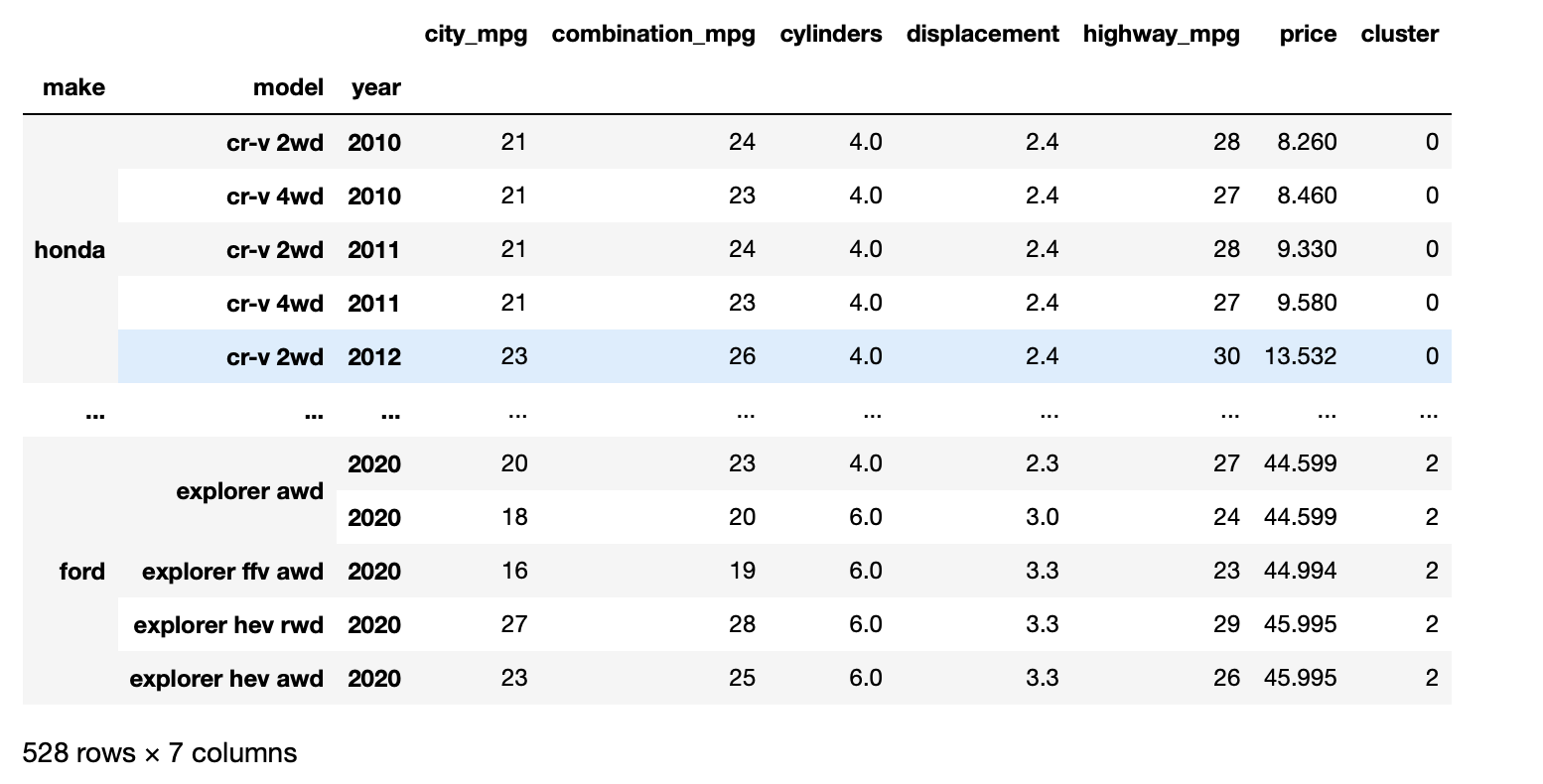
The second part of our data collection involved gathering the prices for all of the models we chose. From a [Kaggle](https://www.kaggle.com/datasets/jpayne/852k-used-car-listings) dataset we found and using another market value [API](https://portal.vehicledatabases.com/api/services/market-value?title=Market+Value&desc=Get+access+to+trade-in,+private-party+%26+retail+values+of+a+vehicle+based+on+its+VIN+or+year,+make,+model.), we were able to extract specific price values for each car we had in our data by filtering the make, model, and year. We had two different sources of the price information as the kaggle dataset had most of the information from the years 2010-2017 and the API had a lot of price information from 2018-2020. We were able to manually put in the price for each of the different cars in the data set which allowed for seamless manipulation of data.



To clean the data, we removed NaN values to remove any rows that had incomplete data. Then we normalized the price column. To do this, we divided the price column by 1000 in order to match the units of the price column with the highway\_mpg and city\_mpg columns because these values were much lower and with normalizing the price it would reduce how large the spread is between these variables. Ultimately, this would allow for clustering that was more meaningful. Accounting for the fact that we had to join many data frames and manually add a lot of the price information, the cleaning of the data was minimal.

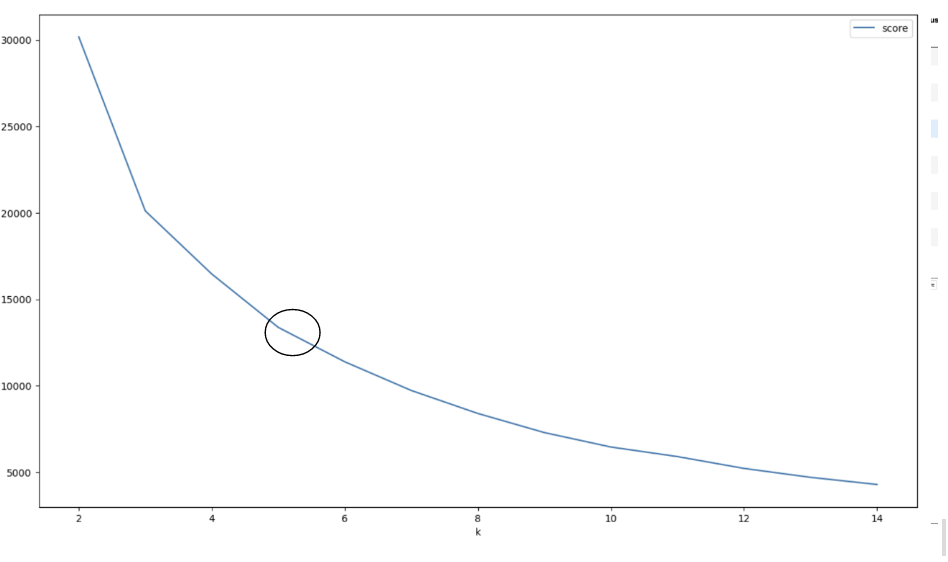
Data Science Techniques

The main data science technique we utilized within this project was k-means clustering. This involved manipulating that data so that it fit the format required for the clustering. The first step we took was to set the make, model, and year of each car in the dataset as the index of the dataframe. This allowed us to group the make of each car with all of its models that we had in our dataset by year. Also, this allowed us to keep the labels for each car while meeting the requirement of only using numeric values for the clustering. Next, we dropped any columns that consisted of non-numeric values such as, drive, fuel type, transmission and class. Dropping these columns will not affect our analysis still.



As mentioned before, we wanted to observe the cluster of cars through the relationship of price and highway mpg and price and city mpg. The K-means clustering allowed us to cluster the cars based on their similarities with the two characteristics we are focusing on. This type of clustering takes into consideration the euclidean distance which finds the distance between the price and mpg of each car and clusters them based on their closeness.

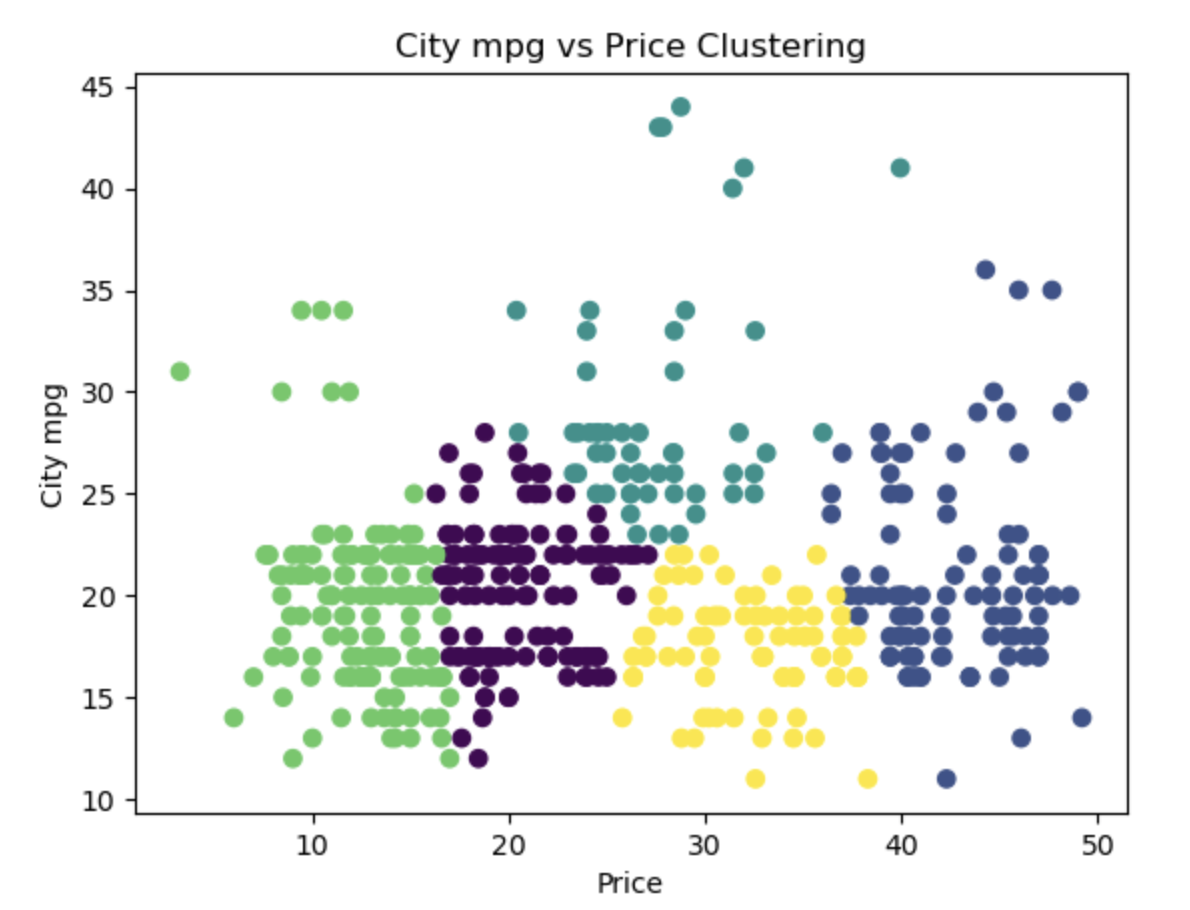
Before clustering, we had to fight the best value for k. In order to calculate the k-value, we used the elbow method. For both of the variables highway\_mpg and city\_mpg, we tested k-values ranging from 2-14. To do this, we fit the k-means model with each k-value and then calculate the inertia score using the sklearn library. Once we calculated the inertia score for each k-value, we added the k-value and inertia score to a dataframe and plotted the data to visualize the elbow. Looking at the image below, we see our elbow graph. Where the graph starts to like more of a straight line is what we should choose for the k-value. Circled in the graph, is where it starts to turn into that straight, making our k-value to be 5.



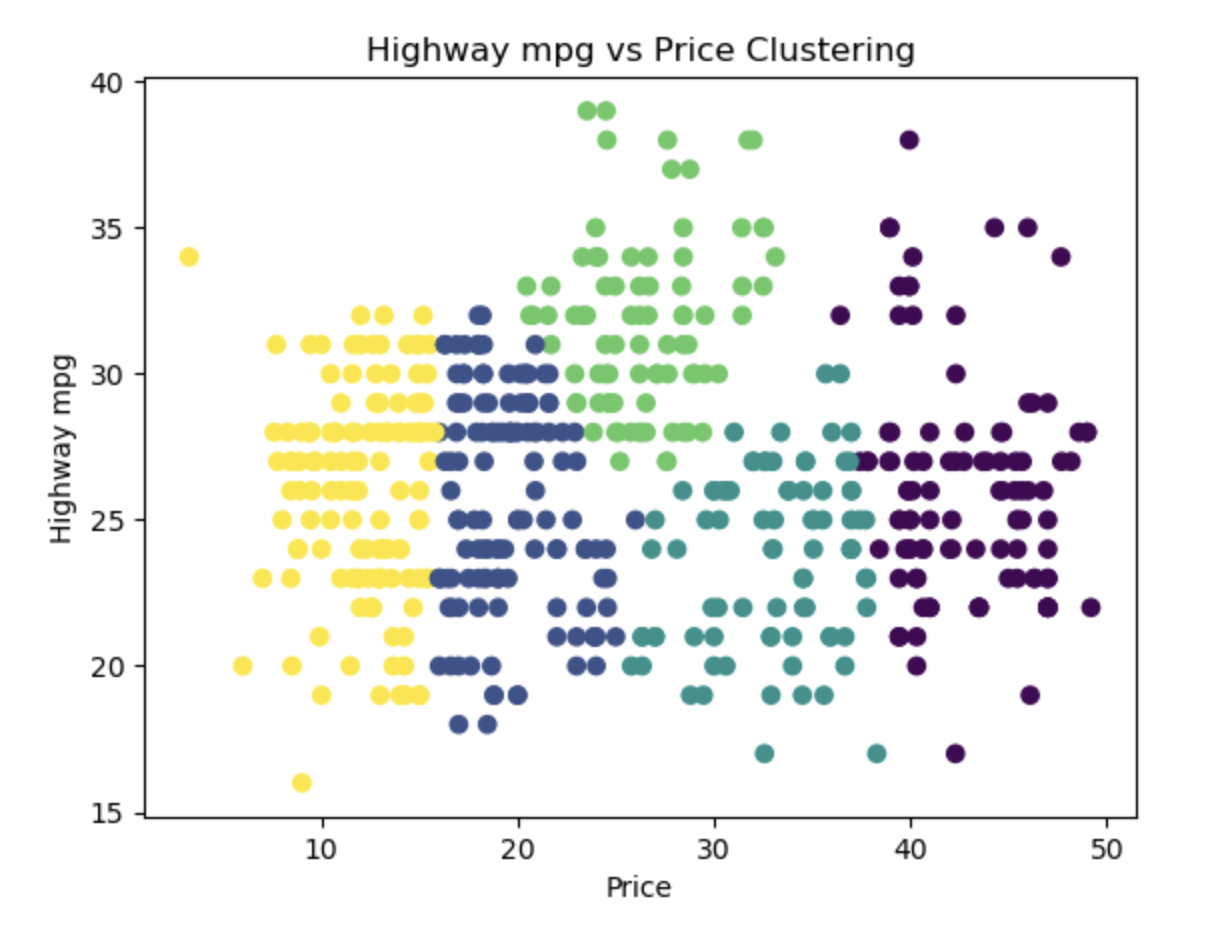
To cluster our data, we fit the k-means model with 5 clusters. To visualize the clustering, graphed the data onto a scatter plot using Matplotlib.pyplot. The clusters were color coded and the different points represented different cars that were in the dataset. We also included original cluster labels back in the dataframe for more efficient analysis. Ultimately, we had to do this process for both the highway\_mpg comparison with price and the city\_mpg variable. This allowed us to find similarities between cars and determine what each cluster represented. Lastly, it would also be important to mention the use of the pandas library for a lot of our data manipulation and cleaning to allow us to complete the k-means clustering.

Data Analysis/Evaluation

To analyze our data, we first wanted to examine each cluster, which is shown in the visual below. This first graph shows us the clusters with price and city mpg.



Looking at the graph above, the overall trend of the clusters seems to go by price. However, in the middle, you can see that clusters are grouped by city mpg. The green cluster (or cluster 3) represents the cars with the lowest prices (less than $10,000 - around $18,000). This cluster has a wide range of city mpg values. Most of the cars in this cluster are from between 2010-2013 and have a dispersed type of makes and models. The purple cluster (or cluster 0), represents cars that have a range of prices from $16,000 to $25,000. This cluster is higher in price but has the same relative range of city mpg. We also found that there are more newer cars within the cluster going up to the year 2018. The turquoise cluster (cluster 2) represents the car price range starting at $20,000 and going up to $40,000. This cluster shows the best performing cars based on city mpg. The yellow cluster (or cluster 4) represents cars in the price range of $26,000 to almost $40,000. This cluster is similar to the turquoise cluster price range wise, but it has lower city mpg ranges. The blue cluster (or cluster 1) represents the highest car price range going from $38,000 to close to $50,000. This cluster has an above average city mpg range, but can still be comparable to the green cluster.



The graph above looks at the highway mpg of the cars in the data. This visualization has a much greater highway mpg spread as compared to the city mpg graph. The yellow cluster (or cluster 4) contains all of the cars with the lowest prices ranging from less than $10,00 to about $16,000. This cluster has a large spread of highway mpg values but the years are consistently around 2010-2014. The blue cluster (or cluster 1), represents cars that have a range of prices from $16,000 to $23,000. This cluster contains some newer cars that correlate with the price increase, however, the overall performance of the cars do not increase respectively. The green cluster (cluster 3) the best performing cluster with the minimum highway mpg being around 26. The prices of cars in the cluster range from $22,000 to $35,000. The turquoise cluster (or cluster 2) represents cars in the price range of $25,000 to $39,000. This cluster is similar to the blue cluster mpg wise, but it has higher prices overall. The purple cluster (or cluster 0) represents the highest car price range going from $38,000 to close to $50,000. This cluster has the largest range of highway mpg amongst all the clusters.

Outcomes/Recommendations

Based on our analysis of the different representations shown in the clusters above, we can recommend different cars to our audience based on certain criteria. The two categories of cars we chose to examine are lowest price and best performing. Most individuals looking to buy cars fall into these two categories of criteria.

For the city mpg analysis, we found that the lowest priced cars are represented in the green cluster. We specifically recommend the 2010-2012 Ford Escape Hybrid, which has a city mpg of 34 and the price is $9,431. These cars have the best prices for their respective city mpg. We understand that our audience will have more preferences so they should be able to find the best, lowest priced, car for them within the green clusters. Considering the best performing category, this essentially looks at the cars which have the best mpg value. These cars were found in the turquoise cluster. The 2020 Ford Escape Plug-In Hybrid, which has a city mpg range of 43-44 and a price of $28,763 and the 2020 Toyota Rav4 Hybrid, which has a mpg of 41 and the price to be $31,986 are the two cars we found fit this category. There are many other cars in this cluster with high mpg but these were the best performing we found.

For the highway analysis, we found that the lowest priced cars are represented in the yellow cluster. The 2010 Chevrolet Equinox, which has a highway mpg of 34 and the price to be $12,000 is what we recommend. These cars have the best prices for their respective highway mpg, but we do understand that our audience can have more preferences.For the best performing cars in this analysis, we found that they were in the green cluster. We recommend the 2019 Chevrolet Equinox which has a 39 mpg value and was priced at $24,500. This car had the highest highway mpg value and a relatively lower price compared to the pricier cars we had had in our dataset.

One interesting takeaway we found was that hybrids were the best performing cars within cities however that did not apply to the highways.

Limitations

While we support the outcomes of our project, there are a few limitations we considered as a result. One limitation is that since we pulled car data from certain sources like the APIs and Kaggle dataset, only the variables that came from those sources are what we could use. We were lacking other factors when it came to certain criteria of cars such as we could not assess the maintenance or car lifespan for each car we had in our data. People might care more about this information when wanting to buy a car that is a little older as these criteria could be important to know.

Another limitation with our project is that we only considered SUVs. We felt that SUVs were a good choice and more personal to us, but we want to acknowledge that many people do not buy SUVs. We have seen how SUVs have been becoming more popular in the last few years and thought that they would be the most appropriate class of car to analyze. However, SUVs tend to be slightly more expensive than sedans so one could argue that our analysis is somewhat biased towards those individuals that have the ability to afford an SUV. As well, SUVs are generally more practical for families as they are large and have more seats and space. This analysis might not be useful for those who do not need the extra room.

One other limitation we had within the project was the scope of the data. Essentially, we focused on SUV models that were most popular/best selling in the U.S. This did not include a broader range of cars, especially luxury cars. In this case, we can say that the data was limited to specific SUV cars that were more affordable and reliable. The cars that were included in our data consisted more of models such as Toyota, Honda, Chevrolet, Jeep. Luxury cars in this case would include models such as Tesla, Mercedes, BMW, etc. Including luxury cars would allow us to have more of an insight about the relationship between price and mpg and if they do have a higher mpg (since they are more expensive). These factors would be important to include in future analysis.

One final issue that we ran into over the course of our project was combining that data from our car API with the price information that we collected. The car API had over 500 rows of specific car models with specific drives and transmissions. Since this data was so specific, it was very difficult to find the price data that matched perfectly. Thus, we had to use multiple sources for our price data. We do not know the specifics about how each source collected their data, so it is possible that the methods are not consistent with each other.

GitHub Code and Data: <https://github.com/meharsaini12/inst414finalproject.git>

Appendix

For the division of work on this project, we split up the responsibilities for data collection. Each team member was responsible for collecting data for 3 or 4 of our 10 car models. Each member gathered their own data from the API and collected the pricing information as well.

1. Avani
   1. Toyota Highlander
   2. Ford Escape
   3. Honda Pilot
2. Mehar
   1. Jeep Grand Cherokee
   2. Honda CR-V
   3. Chevrolet Equinox
3. John
   1. Jeep Wrangler
   2. Toyota Rav 4
   3. Nissan Rogue
   4. Ford Explorer

The remainder of the project, including all of the analysis, we completed together. Every time any of us worked on the project after data collection, we did so together. We held bi-weekly meetings in order to stay up to date on our work. Nobody had individual responsibilities while conducting our data analysis and evaluation.