Dalhousie University

Final Report: Best Automotive Vehicle in 2004

April 10, 2018

Submitted by:

Jin Shi

B00642606

1 Introduction

Staring From the beginning of 21th century, the retail sales number of new automotive vehicle in US has significantly increased compare to the 90th. The car sales number has an average about 16 to 17 million a year expect the year 2008 to 2013 where America was experiencing the Subprime Crisis and the financial market was stroked. During the year 2004, almost 17 million new automotive vehicles was sold in US. [1] With that huge amount of sales number over 380 car models was introduce by car manufactures. [2] Since we have this huge amount of needs in cars and varies choose of car in US, it is very difficult for customer to pick a car model that perfectly match their needs. As a result, in this project a visualization system called *Best Automotive Vehicle in 2004* will be implemented to help potential vehicle customers to choose a car that they desire. A ranking system will be implemented to give the best result

2 Content

In the part of the report a detailed description of this visualization system including: the goals the system wants to achieve, data collection and processing, system design and implementation.

2.1 Goals

In this visualization system the goal I want to achieve is simple. First is to implement some algorithms or functions that can manipulate the data set and create automotive vehicle purchasing suggestions to people that have various kinds of needs.

For example, some people works alone in a very populated city where road are narrow with lots of traffics and they are short on the money. For this type of people, a vehicle with huge dimension and fuel cost is not suitable for them since they don't need a huge car to just sit one person and huge fuel cost will increase during the time that they got stuck in traffic. Instead in this visualization system a car that have low fuel mileage and small dimensions that can easily fit one person will be suggested.

Put all those under consideration I decided 11 types of cars will be ranked on.

2.2 Data Collection and Processing

2.2.1 Data Collection

To create a visualization system, we find need to find a data set that have enough data about new cars in 2004, also the data set should multiple attributes about the cars specifications, so it can be used in functions to determine the best car for the specific needs. From the links in course web site I find a data set that is perfect matches my needs. *2004 Cars and Trucks* is a data set that have 428 new car models that was introduced in 2004 where each model have detailed

specifications on name, type horse power etc. [2] the following table shows all features that the dataset has:

Vehicle Name	Car Type	Drive Train	Retail Price	Dealer Cost	Engine Size (l)	Len
Cylinders	HP	City MPG	Hwy MPG	Weight	Wheel Base	Width

Figure 2.2.1 [2] table of attributes

2.2.2 Understand the Data and Data Cleaning

Since the data set was obtained we can now looks at the data set and understand them for feature analysis. First when we looked at vehicle name we can find that even the model of the cars is the same, but they still have trivial differences in categories like price and weight. As the result we need to keep the format this way instead of making it just one model for multiple sub-models.

In the specification part of the data set there are many car models that have missing values, as the result it will cause our functions unable to calculate the right score for the model. So, I removed all the car models that have incomplete data.

As the result of the data cleaning process on 388 car models was used in the visualization system.

2.2.3 Data processing

Since the data size for this visualization system is quite small. I decided not any data warehousing or database software for the processing of the data set. Excel was the only software I used for data processing in this project.

Considering a car's performance horse power is a very important variable, but if a car has a huge weight high horse power still can not going to be very fast. As the result I put weight into the measurements of a car's performance. A new attribute called *horse power per pound* will be used to. Horse power per pound will measure a cars performance by divide the horse power with weight to find how many horse powers was used to push one pound of weight.

Horse Power Per Pound=HP/Weight(pound)

Money to dealer is another new attributed added to the table where it will find the brand with cheapest dealer fee.

Money to Dealer=Dealer Cost-Retail Cost

Overall fuel was another integration of city MPG and highway MPG it will take the average of those two to give one single value to make the

data easier for feature processing. Higher overall fuel will mean more fuel saving the car will be.

Overall Fuel= (City MPG+Hwy MPG)/2

Dimension is also an attribute that multiplies the length and width of a car to get its dimension. The higher the dimension more sitting and cargo room a car will have.

*Dimension= Length*Width*

2.2.4 Ranking functions

In this part a ranking of all 11 different car needs will be created. The first step is to rank all the useful features from 1 to 387. This ranking can be used for some of the rankings like cheapest car or best performance car. And for those more complicated needs this primary ranking can act as a score. if a weighting system is added to those ranks more difficult problems can be resolved.

power ranking	size ranking	fuel ranking	dealer fee ranking	car price ranking
176	97	280	341	316
180	97	280	346	327
162	73	315	301	274
18	271	300	384	381
72	334	67	159	134
50	171	148	255	243
135	233	89	199	177
263	249	75	207	162
115	249	148	263	227
184	240	163	326	307

Figure 2.2.4.1 screenshots of ranking

For more complicated problems like budget Family car. Three attributes will be put into consideration: 1. How much fuel it will cost? 2. Is the car big enough for the whole family? 3. How much it will cost? Because of those three problems we need to integrate all 3 of those attributes assume that they how equal importance and create a new ranking. My way of approaching this is to add all three attributes together to get a score and give a ranking to the score.
Following is my functions for scores:

Budget Daily Commute= Car Price Ranking + Fuel Ranking
Budget Family Car= Size Ranking + Fuel Ranking +Car Price Ranking
Budget Performance= Performance Ranking + Car Price Ranking

budget daily comute	best budget family car	buget performance
596	693	492
607	704	507
589	662	436
681	952	399
201	535	206
391	562	293
266	499	312
237	486	425
375	624	342

Figure 2.2.4.2 result of the function

As for car rankings that species car type, just rank the cars by rankings and filter out the car that is not the type that we wanted. Also, an overview of car types, drivetrain and country is created so the visualization system will have an overall view. A final table of ranking and overview will be in the appendix

2.2.5 Figure Collection

The picture of cars, brands and country flag was downloaded from website to be used in the system. [3]

2.3 Design and Implementation

When the data is ready to be used a design of the visualization interface was created. The system will have following parts: title, overview of data, 11 Car rankings

For second part I decided to use 3 pie car that will demonstrate the data. Interactions will also be added when moving the mouse on the chart it will display the picture for the type. And in the Ranking section each rank will have a section: one the Top it will display the name of ranking and the right will display the top 3 cars for the ranking. And on the left-hand side of the section a picture of the car or brand will be display. The default picture will be the number one ranking car but if the user clicks on the name or the rank number the picture will switch to the picture of that car.

The last step is to implement the system, I used 3 iterations to complete the project where each iteration has its own tasks. In the first iteration, a basic layout of the system was drawn, and charts added without any data. In the second iteration data was putted into the system and implementation of interactions. In the last iteration a final tune of styling was implemented.

3 Results

In conclusion, a visualization system about recommending new cars in 2004 was created. A screenshot of the system will be in the appendix. For the feature works more interactions could be added to system like let user to choose what type of cars they want to be ranked on. Also, a better scoring function can be implemented by more research on the topic.

References

1. Light vehicle retail sales in the United States from 1977 to 2017 (in 1,000 units). Statista. [online] https://www.statista.com/statistics/199983/us-vehicle-sales-since-1951/

2. 2004 Cars and Trucks. Interactive Data Visualization. [online] http://www.idvbook.com/teaching-aid/data-sets/2004-cars-and-trucks-data/

3. Used Cars. Autobytel. [online] https://www.autobytel.com/used-cars/

Appendix A

Table of overview and ranking data

money saving dealer	cheapest car	most expensive car	best performance car	fuel saver	best budget daily comute
suzuki	Kia Rio 4dr manual	Porsche 911 GT2 2dr	Porsche 911 GT2 2dr	Honda Insight 2dr (gas/electric)	Toyota Echo
kia	Hyundai Accent 2dr hatch	Mercedes-Benz CL600 2dr	Mercedes-Benz SL55 AMG 2dr	Toyota Prius 4dr (gas/electric)	Scion xA 4dr hatch
toyota	Toyota Echo 2dr manual	Mercedes-Benz SL600 convertible 2dr	Audi RS 6 4dr	Honda Civic Hybrid 4dr manual (gas/electric)	Honda Civic DX 2dr

best budget family car	best buget sport car	performance suv	fuel saving mini van	best budget winter AWD
Toyota Camry Solara SE 2dr	Toyota Celica GT-S 2dr	Cadillac SRX V8	Toyota Sienna XLE Limited	Suzuki Aerio SX
Chevrolet Impala 4dr	Nissan 350Z coupe 2dr	Porsche Cayenne S	Dodge Caravan SE	Suzuki Vitara LX
Dodge Intrepid SE 4dr	Subaru Impreza WRX 4dr	BMW X5 4.4i	Nissan Quest S	Honda Element LX

Appendix B Screenshot of the project