IOWA STATE UNIVERSITY SPRING 2013

Data Mining on the Fuel Economy of Vehicles

Statistics 503 Data Mining Class Project

Bo Wang

4/26/2013

1.0 Description

There are three objectives in this fuel economy study. The main objective of this research is to investigate how different vehicle designs (i.e. number of cylinders, vehicle types, transmission types, etc.) affect fuel economy using data mining techniques. The second objective is to come up with a list of vehicle makes that most likely to inflate their fuel economy values. The third objective of this study is to cluster vehicles with similar types into groups, so that we know which types of vehicles are more similar to each other in terms of fuel economy performance. This information is useful for vehicle purchase guidance.

The fuel economy data can be accessed on the Environmental Protection Agency's website (EPA, 2013). The raw dataset has 20 variables with 33,245 observations, and the vehicle models range from 1984 to 2013 year. This dataset contains abundant information about the features of tested vehicles, such as engine type, transmission type, number of cylinders, fuel type, etc. The researcher will first conduct exploratory analysis to understand how different vehicle designs affect fuel economy. Then the researcher will utilize K-means and Hieratical clustering technique to classify vehicles into groups. Since city fuel economy has linearly relationship with highway fuel economy, we only examine the city fuel economy in the data analysis, and the findings should be very similar for highway driving cycles. The description of the dataset is listed in Table 1.

Table 1. Descriptions of Variables

Variables	Description
Year	Model year
Make	Manufacture
Model	Model name
Cylinders	Engine cylinders
Displacement	Engine displacement (liters)
Drive	FWD, RWD, 4WD, FWD, AWD
Transmission	Manual, Auto
Gear	Number of gears
Vehicle type	Car, suv, ptruck, spv, van
Alternative fuel vehicle type	Hybrid, EV, Plug-in Hybrid, CNG, etc.
Turbo charge	Vehicle is turbocharged (Yes, No)
Super charge	Vehicle is supercharged (Yes, No)
Battery type for electric vehicles	Ni-MH, Li-ion, DCPM, etc.
City	City Fuel Economy (mpg)
Highway	Highway Fuel Economy (mpg)
CO2 tail pipe emission	Tested tail pipe CO2 emission (gram per mile)
Fuel Cost	Annual Fuel Cost (dollars)

2.0 Outline of Analysis

This section lists the data analyses that conducted in this study.

Table 2. Outline of Analysis

Approach	Reason	Type of questions addressed
Data Reconstruction	Prepare data for EDA and clustering analysis. ● Impute missing values; ● Create new variables; ● Transform variables;	
Summary Statistics	Extract location and scale information, such as mean, standard deviation, and quantile.	"What is the average fuel economy for cars?" "Does truck have better fuel economy than SUV?"
Histograms	Explore univariate distribution	"What is the pattern of fuel economy over years?" "Do diesel vehicles have better fuel economy than gasoline vehicles?"
Boxplots	Explore distributions; detect any outliers;	"Why the outliers have better or worse fuel economy?"
Pairewise Scatter Plot	Explore bivariate distribution and correlations between variables; Explore any data patterns;	"What is the relationship between city fuel economy and highway fuel economy?"
Clustering Model (k-means & hierarchal clustering)	Cluster vehicles (2000 year or newer) into groups;	"Which vehicles are more similar to each other?" "What are the major differences among groups?"

3.0 Results

This section summarizes the exploratory data analysis through summary statistics and plots. The data is analyzed using R version 2.15.3.

3.1 Summary Statistics

Table 3. Summary statistics of city and highway fuel economy. The fuel economy values range from 5 mpg to 138 mpg. Highway fuel economy is generally higher than city fuel economy.

	City Fuel Economy	**Highway Fuel Economy**
Median	17.00	23.00
Max	138.00	108.00
Min	5.00	9.00
Mean	17.43	23.43
SD	5.21	5.98

Table 4. Counts (and proportions). Overview of the vehicle types in the dataset. Forward drive and rearward drive are equally popular. Passenger cars accounts for 58% of all vehicle types on market.

drive types			**Transmission**			**Vehicle Type**		
FWD	12493	0.376	Automatic	21593	0.65	CAR	19342	0.58
RWD	12185	0.367	Manual	11653	0.35	PTRUCK	5413	0.16
4WD	6222	0.187				SUV	4034	0.12
AWD	2257	0.068				VAN	2165	0.07
P4WD	70	0.002				SPV	2292	0.07
Electric	19	0.001						

Table 5. Counts (and proportions) for alternative fuel vehicles. Diesel fuel and flex-fuel vehicles are the two most popular alternative fuel vehicles on the market. Ni-MH battery types are the most popular battery for electric and hybrid vehicles.

Alternative Vehicle Types			**Battery Types for Electric Vehicles**			**Fuel Type**		
FFV	973	0.43	Ni-MH	186	0.68	Regular Gasoline	24073	0.724
Diesel	892	0.39	Li-ion	41	0.15	Premium Gasoline	8049	0.242
Hybrid	294	0.13	AC	28	0.10	Diesel	992	0.030
CNG	47	0.02	SCPM	11	0.04	Natural Gas	57	0.002
EV	47	0.02	DC	6	0.02	Electricity	47	0.001
Biofuel (CNG	18	0.01				Midgrade (Gasoline)	28	0.001

Biofuel (LPG)	8	0						
Plug-in Hybrid	8	0						

Summary of the basic statistics in this section are shown below:

- The highest fuel economy is 138 mpg for Scion iQ EV, while the lowest fuel economy is Lamborghini Countach with 6 mpg.
- The average fuel economy for city driving is 17.43 mpg, while it is 23.43 mpg for highway fuel economy.
- Passenger vehicles account for 58% of all vehicle types.
- Regular gasoline and premium gasoline accounts for 96% of all vehicle types.
- Forward Drive and Rearward Drive are the two most popular drive types with 74% market share.
- Automatic transmission is two times popular than manual transmission.
- Flexible-fuel vehicles accounts for 43% of all alternative fuel vehicles.
- Ni-MH is the most popular type of batteries for electric and hybrid vehicles.

3.2 Plots

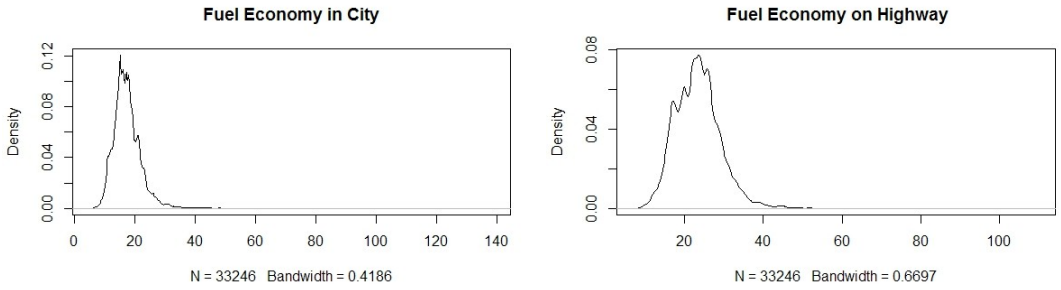

Figure 1. The distribution of fuel economy for city driving (left) and highway driving (right). Both distributions are right skewed with electric vehicles as outliers.

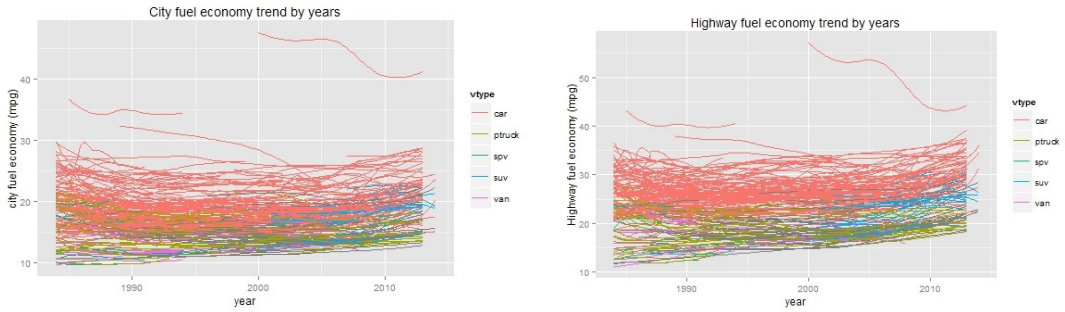

Figure 2. City and Highway Fuel Economy Trend from 1988 to 2013. This plot draws the time-series of fuel economy data by same vehicle model. A slightly improvement is observed over the past two years.

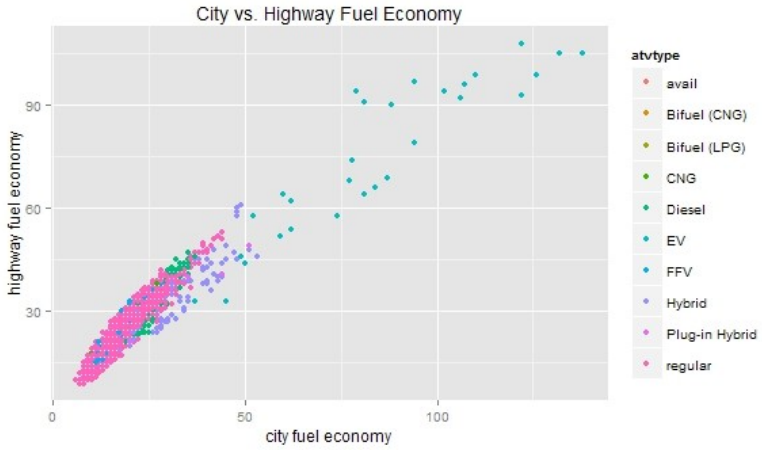

Figure 3. City vs. highway fuel economy by vehicle types. Highway fuel economy and city fuel economy values are in linear relationship. The outliers on the right above region are electric vehicles and flex fuel vehicles. The hybrid vehicle (in dark purple) tends to have higher city fuel economy than highway fuel economy.

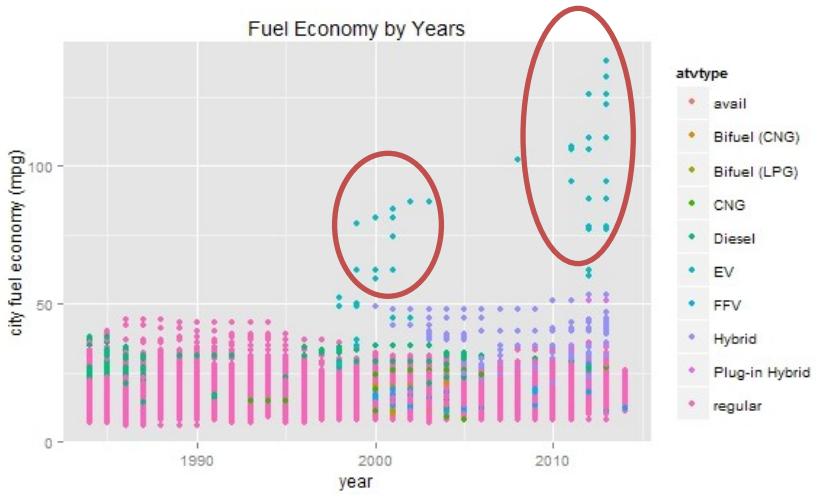

Figure 4. Fuel Economy by years. (Colors indicate vehicle types). Two groups of outliers were observed in the boxplot. Those outliers represent flex-fuel vehicles and electric vehicles.

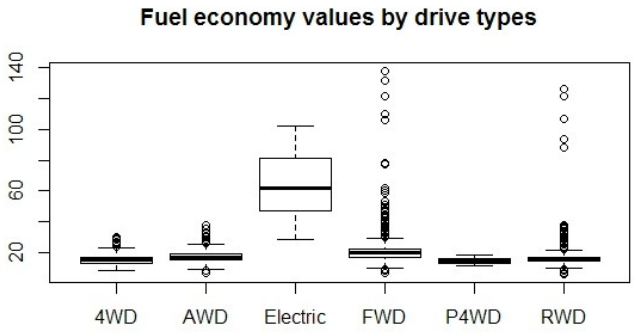

Figure 5. Fuel Economy values by drive types. Electric vehicles has significant higher fuel economy, but with large variation. Both FWD and RWD drive vehicles have a large number of

outliers due to electric vehicles.

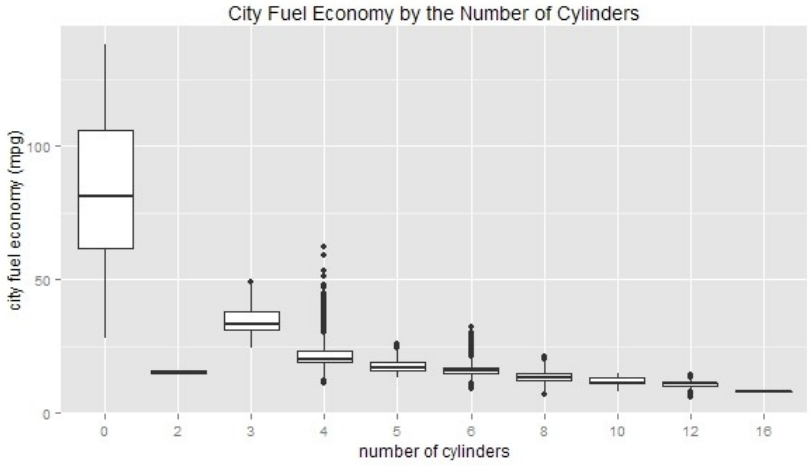

Figure 6. Fuel economy values by the number of cylinders. Zero cylinders indicate electric vehicles. In general, less number of cylinders indicates better fuel economy.

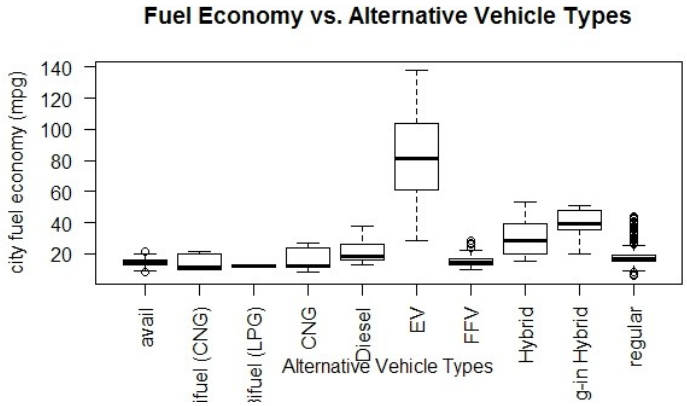

Figure 7. Vehicle fuel economy by alternative fuel types. Electric vehicle has clearly much better fuel economy than any other types of alternative fuel vehicle technologies. Plug-in hybrid vehicles are second to electric vehicles.

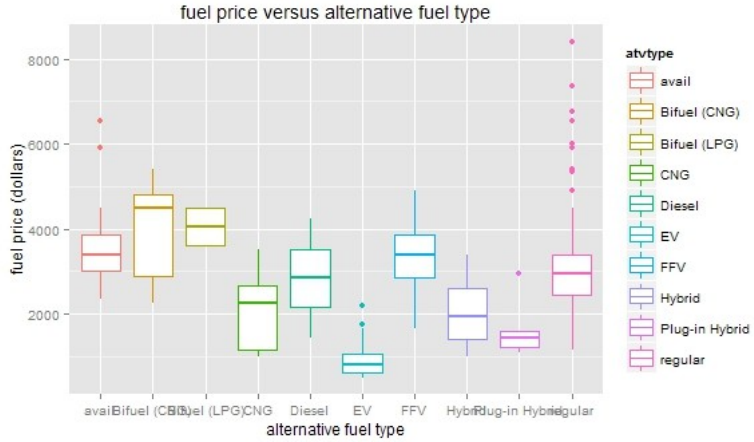

Figure 8.Annual Fuel Cost by Vehicle Types. The electric vehicle has lowest annual fuel cost with 916 dollars, which is one third of regular passenger vehicles.

The findings in this section are summarized below. All those findings come from exploratory data analysis, and not tested for statistical significance.

- Both distributions of city fuel economy and highway fuel economy are skewed to the right.
- The fuel economy has been increasing slightly over the past twenty years.
- Electric vehicle has the highest fuel economy compared to other types of vehicles.
- Highway fuel economy and city fuel economy are in linear relationship.
- Electric vehicle has much higher fuel economy than other vehicles. Electric vehicles are also getting more and more popular in recent years.
- The second best fuel economy vehicles are plug-in hybrid, while the third best vehicle types are hybrid vehicles.
- Although Ni-MH battery is the most popular battery by far, the DCPM battery provides better fuel economy than other battery types for electric vehicles.
- The higher number of combustion engines leads to worse fuel economy.
- Manual drive has slightly better fuel economy than automatic drive.
- Passenger cars have slightly better fuel economy than other types of vehicles.
- The tail pipe emission has log linear relationship with fuel economy. Electric vehicle has zero tail pipe emission.

3.3 Fraud Detection Results

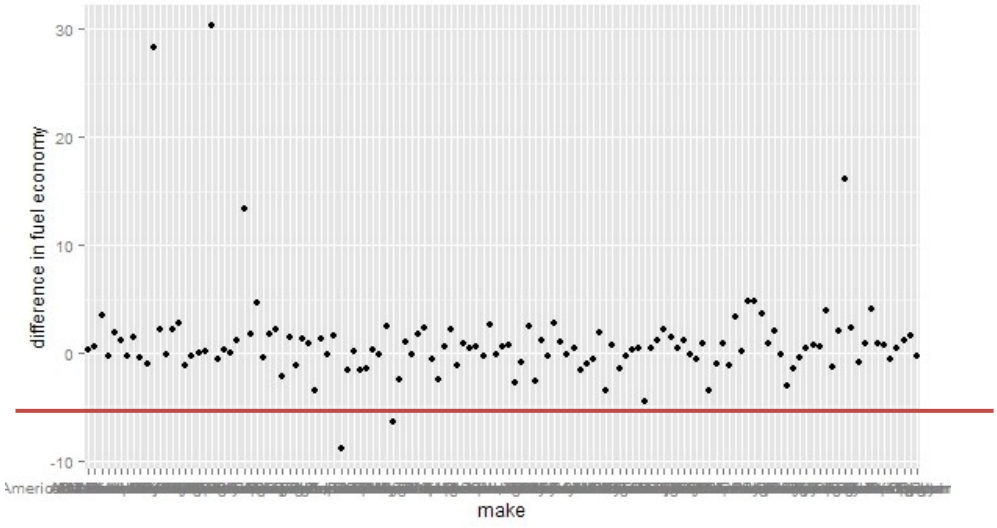

Figure 9. Difference between vehicle make and other similar vehicles. The higher difference indicates either very good fuel economy performance or inflated fuel economy.

Since EPA has limited resources, they can only test a sample of vehicles in each year. The vehicle manufacture actually reports the testing fuel economy by them and EPA only audit a sample of the vehicles in each year. The problem is which vehicles should be selected for auditing. The hypothesis is that vehicle should have similar fuel economy with other similar vehicles. If a manufacture consistently reports higher fuel economy compared with other similar vehicles, it is suspicious that they inflate the fuel economy value (or they did a really good job to improve fuel economy). The vehicles on top of the chart in Figure 9 are mostly

electric vehicles, but some other vehicles also shown up on top of the chart, such as Rolls-Royce, Maserati, Saab, etc. Those vehicles are most suspicious for fuel economy inflation.

4.0 Cluster Models

Before I fit the cluster model, the categorical variables are converted to dummy variables for each variable level. Since we only concern about the vehicles on current market, only the vehicles between 2000 and 2013 are used for clustering. This reduced dataset has 6,515 observations with 40 variables. Both K-means and Hieratical clustering methods are used for clustering.

4.1 K means

K-means clustering method is one of the most robust clustering methods, and has efficient calculation algorithm. K-means clustering is modeled with 2 to 10 groups. The first step is to determine how many groups should be used. Two criterions were used to evaluate the goodness of fit. The first criterion calculates the variation between points and their clusters. The second criterion calculates distance between clusters relative to variation around cluster means. Two graphs are shown below to find the optimum number of clusters. Based on criteria 1, 5 groups is chosen for clustering

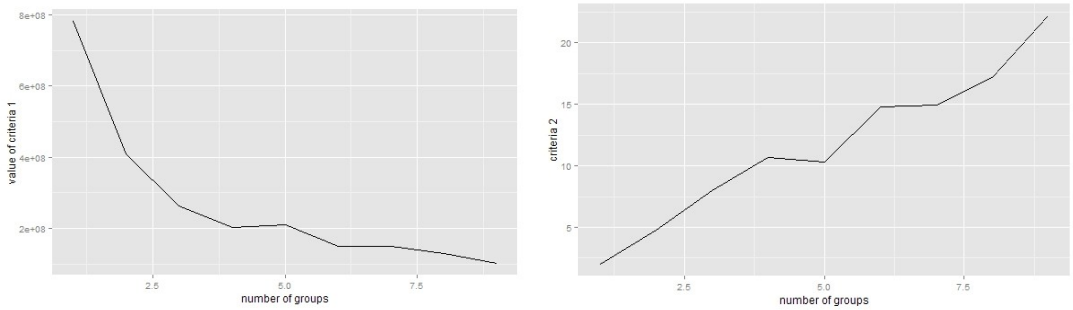

Figure 10. Criteria1 (left) and criteria2 (right) are calculated for k value from 2 to 10. The optimum number of groups is about 5 groups based on criteria 1. Therefore, the 6 groups are used for clustering.

4.2 Hieratical Clustering

Ward, Single, and Complete linkage methods are used for hieratical clustering. However, the Single linkage and complete linkage method are not present here, since they are very sensitive to outliers. Only Ward linkage hieratical clustering is shown here. The Figure 11 draws the dendrogram of Hieratical clustering.

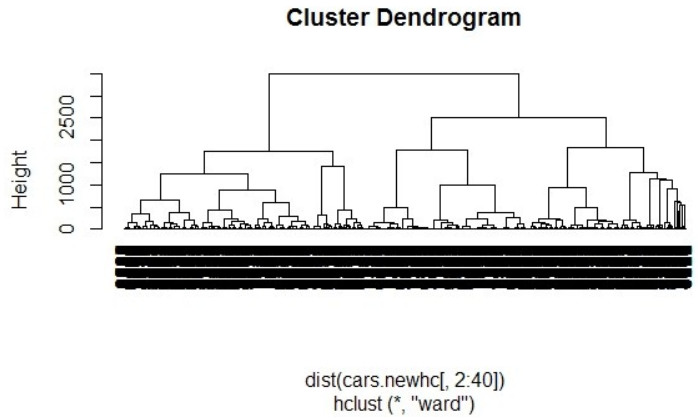

Figure 11. Dendrogram for Ward Linkage Hieratical Clustering

4.3 Model Comparison

Based on Table 6, K-means only agrees with Ward linkage hieratical clustering for 29% of all observations. In order to determine which clustering method is more efficient, the wb.ratio is calculated for both K-means and Ward linkage hieratical clustering. The wb.ratio is 0.22 for k-means and 0.77 for Ward linkage hieratical clustering. Lower wb.ratio indicates better clustering. K-means is clearly better model in this case. Therefore, the K-means method is used for final clustering.

Table 6. Comparison of K-means and Ward Linkage Hieratical Clustering

	K-means				
Ward	1	2	3	4	5
1	568	423	284	552	465
2	62	166	166	361	196
3	866	293	213	223	110
4	890	142	58	0	0
5	396	42	39	0	0

5.0 Summary of Findings

In summary, the vehicles are clustered into five groups using K-means clustering method. The representative vehicles of those five groups are listed below. The summary statistics for each cluster is shown in Table 7.

Group1: Honda Civic, Volkswagen Jetta, Toyota Corolla, Nissan Sentra, and Mazda 3, etc.
Group2: Honda Accord, Nissan Sentra, Toyota Camry, Buick Regal, and Hyundai Sonata, etc.
Group3: Volvo S60, Cadillac CTS, Mitsubishi Eclipse, Nissan Altima, and Ford Taurus, etc.
Group4: Ford Mustang, Audi A4 quattro, Ford Ranger Pickup, Chevrolet Corvette, and Toyota Tacoma, etc.
Group5: Ford F150 Pickup 2WD, Dodge Ram 1500 Pickup, Jeep Grand Cherokee, Dodge Dakota Pickup 4WD, and Chevrolet Silverado, etc.

Table 7. Summary of Final Clustering Results

	Groups				
	1	2	3	4	5
City (mpg)	26.28	20.56	17.4	14.24	12.38
Highway (mpg)	34.67	28.26	24.58	19.33	16.65
Cylinders	4	4.3	5.4	7	7.8
Displacement	1.89	2.32	3.04	4.37	5.13
Drive FWD	98%	80%	51%	1%	0%
Drive RWD	0%	7%	28%	59%	41%
Transmission Manual	49%	44%	31%	23%	13%
Ptruck	0%	1%	13%	44%	44%
SUV	2%	21%	16%	35%	41%
Electric Vehicles	0%	0%	0%	0%	0%
Hybrid	3%	0%	0%	0%	0%
fuel cost	1901	2345	2804	3427	3991
Group Size	1888	1936	131	951	1609

In addition, we also have some interesting findings based on exploratory data analysis. They are summarized below.

- In general, we can see the fuel economies improved slightly over the past 20 years.
- More cylinders would lead to lower fuel economy;
- Higher engine displacement would lead to lower fuel economy;
- FWD Drive type has higher fuel economy;
- Manual transmission would have slightly better fuel economy in general;
- Passenger cars have better fuel economy than any other type of vehicles;
- Van has the worst fuel economy performance.
- Electric vehicle has lowest fuel economy with annual fuel cost of 916 dollars, which is one third of average passenger cars.
- DCPM battery types provide best overall fuel economy among electric vehicles.

References

Environmental Protection Agency. EPA Investigation Prompts Carmaker to Correct Inflated Mileage Claims. Retrieved on April 20, 2013 at http://www.epa.gov/fueleconomy/labelchange.htm

Environmental Protection Agency. Download Fuel Economy Data. Retrieved on April 10, 2013 at http://www.fueleconomy.gov/feg/download.shtml

R version 2.15.3

Statistics 503. Class Handouts. Iowa State University. Spring 2013.

Torgo, L. (2011). Data Mining with R.

Appendix

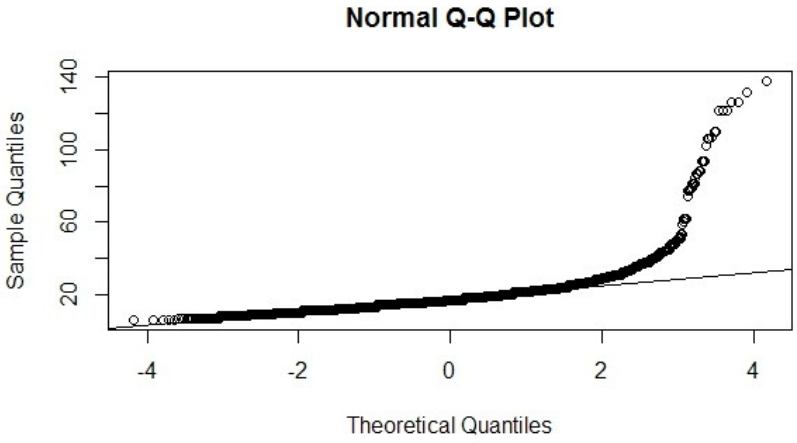

Figure A1. QQ Plot for City Fuel Economy variable

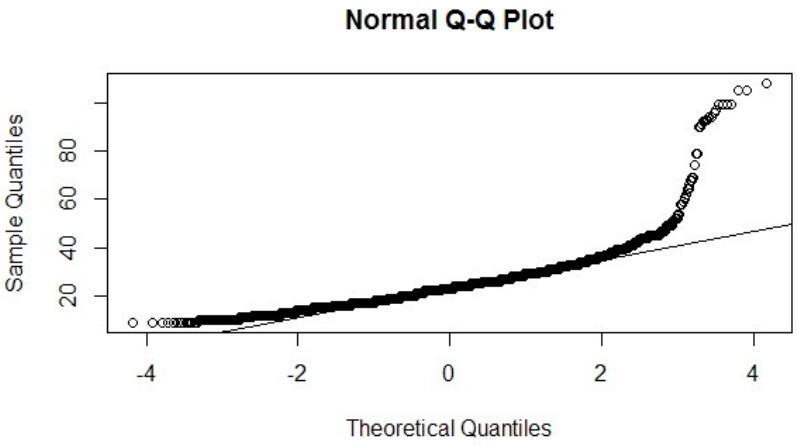

Figure A2. QQ Plot for Highway Fuel Economy Variable

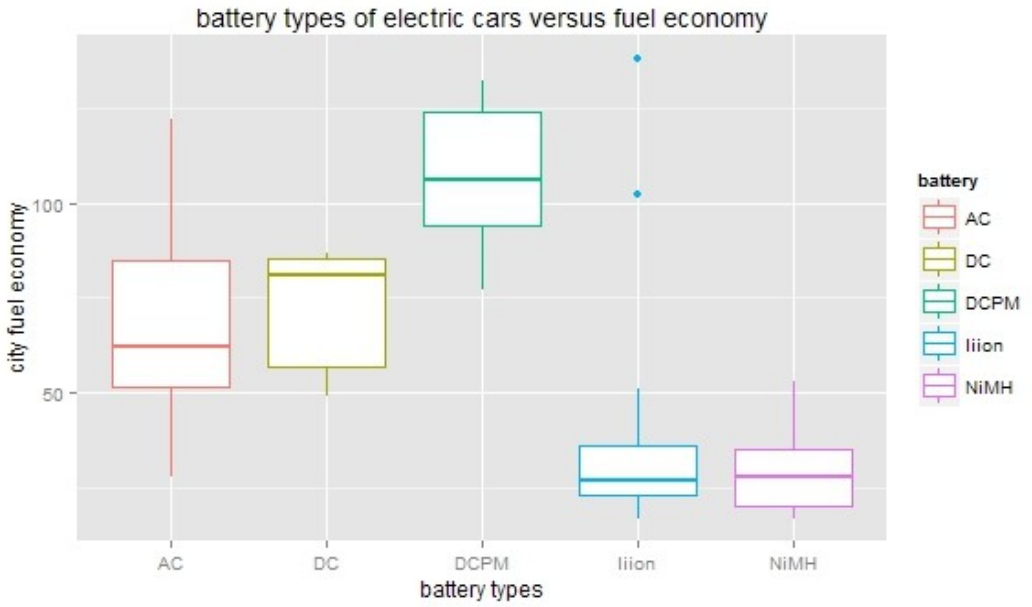

Figure A3. Fuel Economy by Battery Types. Battery type has strong association with fuel economy. DCPM battery type has clearly better fuel economy than other battery types.

Table A1. The top five vehicles with best fuel economy. The vehicles with highest fuel economies are all electric vehicles.

Make	**Model**	**Fuel Economy (Miles per gallon)**
Scion	iQ EV	138
Honda	Fit EV	132
Mitsubishi	i MiEV	126
smart	Fortwo electric drive convertible	122
Fiat	500e	116

Table A2. The top five vehicles with worse fuel economy. The vehicles with worst fuel economies are luxury racing cars.

Make	**Model**	**Fuel Economy (Miles per gallon)**
Aston Martin	Saloon Vantage Volante	7
Rolls-Royce	Corniche	7
Maserati	Quattroporte	7
Rolls-Royce	Silver Spur Limousine	7
Lamborghini	Countach	6