Green Vehicle Guide

CS170A Final Project

Introduction

About the Dataset

The United States Environmental Protection Agency's (EPA) Office of Transportation and Air Quality (OTAQ) evaluates vehicles for environmental performance in the "Green Vehicle Guide" report annually. Every vehicle model is given a Greenhouse Gas Score, Fuel Economy (MPG for city, highway, and combined), and Air Pollution Score. Here is the link to the online dataset: http://www.epa.gov/greenvehicles/Download.do. The Greenhouse Score is based on CO2, methane, and nitrous oxide emissions. The Air Pollution Score is based on vehicle tailpipe emissions that create haze, smog, and health problems.[1] More definitions can be found at the Glossary of Terms: http://www.epa.gov/greenvehicles/Glossary.do

The dataset has the following attributes for about 2,000 car models, in this order:

- Engine Displacement
  - Engine size given in units of liter (e.g. 3.2 liters)
- Cylinders
  - Number of cylinders in the engine
- Transmission Type
  - Auto, Semi-Auto, Auto-Manual, CVT, Manual, Other
- Gears
  - Number of different gear ratios between the engine and the drive wheels including reverse gear
- Drive
  - 4 wheel drive or 2 wheel drive (usually)
- Fuel Type
  - Gasoline, diesel, battery electric, CNG, ethanol\gas, hydrogen
- Vehicle Class
  - small, midsize, large, minivan, van, SUV, pickup, station wagon, special purpose
- Air Pollution Score
  - Scores 1-10, with 10 the highest
- City MPG
  - Estimated miles per gallon in stop and go traffic
- Highway MPG
  - Estimated miles per gallon on traffic-free highways
- Combined MPG
  - Calculated mathematically as 1/[(0.55/city mpg) + (0.45/highway mpg)]
- Greenhouse Gas Score
  - Scores 1-10, with 10 the highest
- Smartway
  - Term given to vehicles with a combined Greenhouse Gas and Air Pollution scores that place them in the top percentile (about top 20%)

[1] *Green Vehicle Guide,* http://www.epa.gov/greenvehicles/Aboutratings.do

Also, note that the dataset does not include vehicles weighing over 8,500 pounds and therefore does not include the heaviest pick-ups and SUVs.

For vehicles that use alternative fuels, such as plug-in hybrid electric vehicles and compressed natural gas vehicles, the MPG rating is actually the miles per gallon of gasoline equivalent (MPGe). This MPGe instead represents the numbers of gallons the vehicle can go corresponding to what amount has the equivalent energy content as gasoline. This supposedly provides a reasonable comparison between gasoline and alternative fuel cars. Therefore, according to the EPA, you can use the MPGe to compare non-gasoline fueled vehicles and gasoline fueled vehicles, even though the 'gasoline replacement' fuel type is not exactly measured in gallongallons.

Modifications to the Dataset

Some attributes in the dataset did not have numerical values, and a numerical conversion was necessary to analyze the information in Matlab. For example, for car types, '1' was used for a small car, '2' for a midsize car, etc. Also, the drive and the fuel type were combined into one variable and converted into numbers. This conversion was done for the different transmission types as well.

The original dataset had 2166 cars. However, due to some of the cars missing information, usually the greenhouse and air pollution scores, I narrowed down the dataset to 1877 vehicles for analysis. Vehicles that were removed included, but were not limited to: Cadillac Escalade, Ford Fusion FFV, Chevrolet Camaro, Chevrolet Malibu, Chevrolet Tahoe, Dodge Ram 2500, Ford E150, Ford F350, Ford Escape, GMC Sierra 15, Lincoln Navigator, Nissan Titan, etc. As far as I could tell, the vehicles which were missing data seemed to be random, and were not all of a certain brand or attribute. This was unfortunate, as some of the few alternative-fuel vehicles that are in the market were deleted.

Objective

By analyzing this data, I hope to gain a better understanding of what factors correspond to environmentally-friendly, fuel-efficient vehicles. And also what attributes correspond to a less fuel-efficient vehicle. Are small cars the most fuel efficient? Do the "smartway" cars also have the highest MPG? Is 2-wheel drive or 4-wheel drive better? What type of fuel is best? And so on.

Log

After getting started analyzing this dataset, I noticed that the largest challenge for this dataset would be working with the massive amount of data (about two thousand vehicles were tested) and dealing with how only whole numbers were used in evaluating vehicles. In other words, when plotting, sometimes it is not apparent whether a dot represents one car or twenty cars, which is problematic.

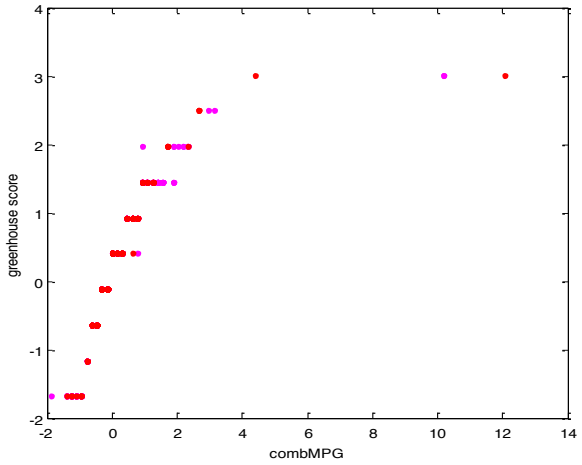
Sometimes I was able to quickly get good results. In those cases, my work is not mentioned in this section but is entirely in the Results section. Other times, I received some strange results that took a couple of tries to comprehend, which can be seen below.

Plotting different vehicle types

At first, I tried to use only the MPG and the greenhouse score to analyze the results. However, the problem with this is that there are multiple cars sharing the same characteristics, since both attributes have whole numbers. As can be seen below, the points superimpose on top of each other, and it shows only a couple of points instead of a couple hundred and we cannot distinguish the different car types.

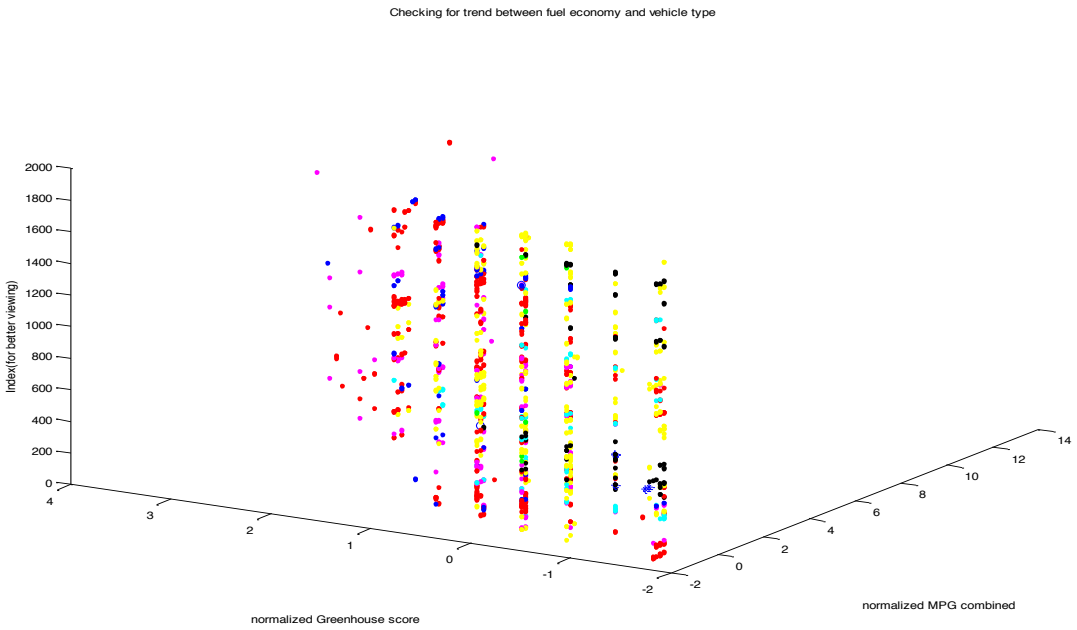
For example:

```
figure; plot(cmbMPG(list_small), …
greenhouse(list_small), '.m');
hold on; plot(cmbMPG(list_mid), …
greenhouse(list_mid), '.r' );
xlabel('combMPG'); ylabel('greenhouse
score');
```



This difficulty occurred in other similar situations.

Next, I tried adding in an index in the z-direction to prevent points for being on top of each other. I got the resulting plot:



This was not helpful because it is cluttered and messy. Please see the Results section for how I sorted the data according to car type to address this issue.

Finding the average car

I hoped that by averaging out the values I would find the average car. However, since whole numbers were used there wasn't quite a car that fit the results. Also, since non-numeric characteristics were coded in as numbers, this was not helpful in any sense.

```
avgCar = mean(Cars);

Results:
3.2620 %Disp
5.6622 %Cyl
2.5301 %Trans
5.4353 %Gears
5.3474 %Drive
3.1129 %VehClass
5.6276 %AirPoll
19.4763 %City MPG
26.4774 %Highway MPG
21.9989 %Combined MPG
4.2206 %greenhouse score
0.1747 %smartway
```

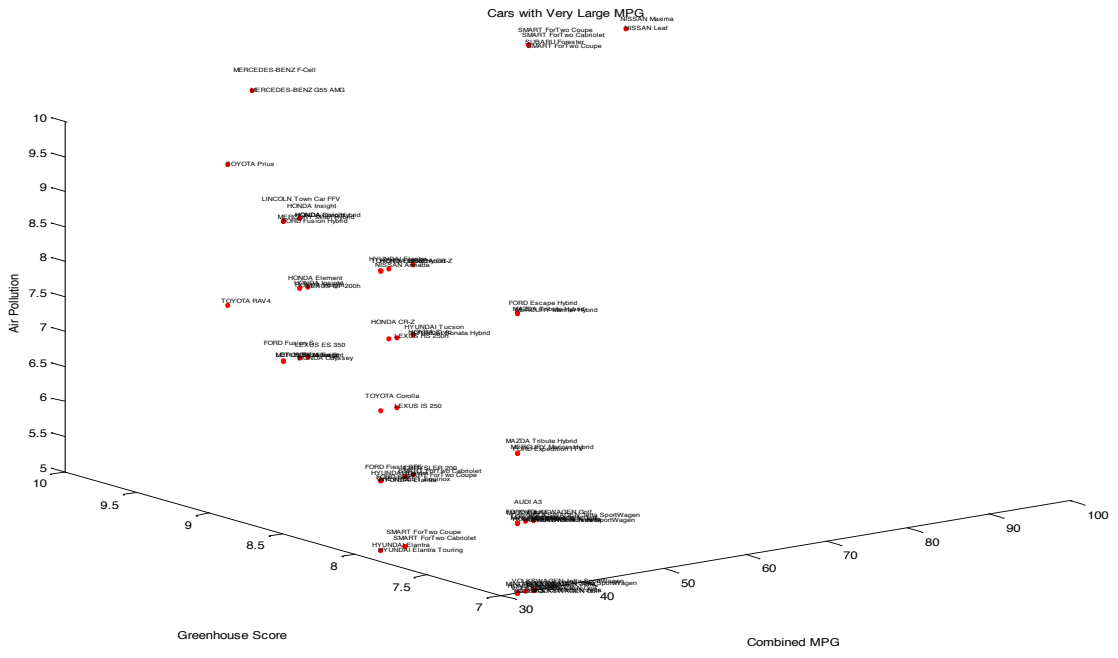
		Displ	Cyl	Trans	Gears	Drive	Veh Cla	Air Poll	City MP	Hwy M	Cmb M	GreenH	SmartW
1877	VOLVO XC 60	3.2	6	2	6	4	5	6	18	25	21	4	0
1885	VOLVO XC 70	3.2	6	2	6	4	5	6	18	25	21	4	0

		Displ	Cyl	Trans	Gears	Drive	Veh Cla	Air Poll	City MP	Hwy M	Cmb M	GreenH	SmartW
1164	MERCEDES-BENZ S400 Hybrid	3.5	6	1	7	4	3	6	19	25	21	4	0

Analysis of Outliers

After noticing that the extremely high MPG scores were skewing my results, I thought it might be helpful to see which cars these were and analyze their results more closely. Using basic statistics, I found the outliers.

```
CmbMPG = Cars(:,10);
interQ = iqr(CmbMPG);
q = quantile(CmbMPG,[.25 .50 .75]); % the quartiles of x

extra_large = find(CmbMPG > (q(3)+1.5*interQ) );
extra_small = find(CmbMPG < (q(1)-1.5*interQ));

plot3(CmbMPG(extra_large), greenhouse_scores(extra_large),
airpollution_scores(extra_large), 'r.');

for j = 1:length(extra_large)
    i = extra_large(j);
    t = 1+abs(randn/100);
    text(CmbMPG(i) * t, greenhouse_scores(i)*t, airpollution_scores(i)*t, Name(i),
'FontSize', 5);
end
rotate3d on;
xlabel('Combined MPG'); ylabel('Greenhouse Score'); zlabel('Air Pollution');
title('Cars with Very Large MPG');
disp(Name(extra_large));
```



I got 79 cars that were 1.5 * Interquartile Range beyond the third quartile. I was able to plot the cars with their scores and see which were of interest. However, I soon realized that these results were not helping me towards what I was trying to achieve. I am not a car expert, and these results do not give me

any quantitative information to allow me to draw conclusions. The end result is a list of cars with extra large combined MPG scores, which can be seen much more cohesively by using a filter in Excel.

'AUDI A3
'AUDI A3
'CHEVROLET Equinox
'CHRYSLER 200
'FORD Escape Hybrid
'FORD Expedition FFV'
'FORD Fiesta
'FORD Fiesta SFE
'FORD Flex
'FORD Fusion Hybrid
'FORD Fusion S
'HONDA CR-Z
'HONDA CR-Z'
'HONDA CR-Z
'HONDA Civic
'HONDA Civic Hybrid
'HONDA Element
'HONDA Fit
'HONDA Insight
'HONDA Insight
'HONDA Insight
'HONDA Insight
'HONDA Insight
'HONDA Insight
'HONDA Odyssey
'HYUNDAI Elantra
'HYUNDAI Elantra
'HYUNDAI Elantra
'HYUNDAI Elantra
'HYUNDAI Elantra Touring
'HYUNDAI Sonata Hybrid
'HYUNDAI Tucson
'LEXUS CT 200h
'LEXUS ES 350
'LEXUS HS 250h
'LEXUS IS 250
'LINCOLN Town Car FFV
'LOTUS Elise/Exige'
'MAZDA 2
MAZDA 3
'MAZDA Tribute Hybrid
'MAZDA Tribute Hybrid
'MERCEDES-BENZ F-Cell
'MERCEDES-BENZ G55 AMG
'MERCURY Mariner Hybrid
'MERCURY Mariner Hybrid
'MERCURY Milan Hybrid
'MERCURY Milan S
'MINI Mini Cooper
'MINI Mini Cooper S
'NISSAN Armada
'NISSAN Lea
'NISSAN Maxima
'SMART ForTwo Cabriolet
'SMART ForTwo Cabriolet
'SMART ForTwo Cabriolet
'SMART ForTwo Coupe
'SMART ForTwo Coupe
'SMART ForTwo Coupe
'SMART ForTwo Coupe
'SUBARU Forester
'TOYOTA Camry Hybrid
'TOYOTA Corolla
'TOYOTA Prius
'TOYOTA RAV4
'TOYOTA Yaris
'VOLKSWAGEN CC
'VOLKSWAGEN Golf
'VOLKSWAGEN Golf
'VOLKSWAGEN Golf
'VOLKSWAGEN Golf
'VOLKSWAGEN Jetta
'VOLKSWAGEN Jetta
'VOLKSWAGEN Jetta
'VOLKSWAGEN Jetta
'VOLKSWAGEN Jetta SportWagen
'VOLKSWAGEN Jetta SportWagen
'VOLKSWAGEN Jetta SportWagen
'VOLKSWAGEN Jetta SportWagen

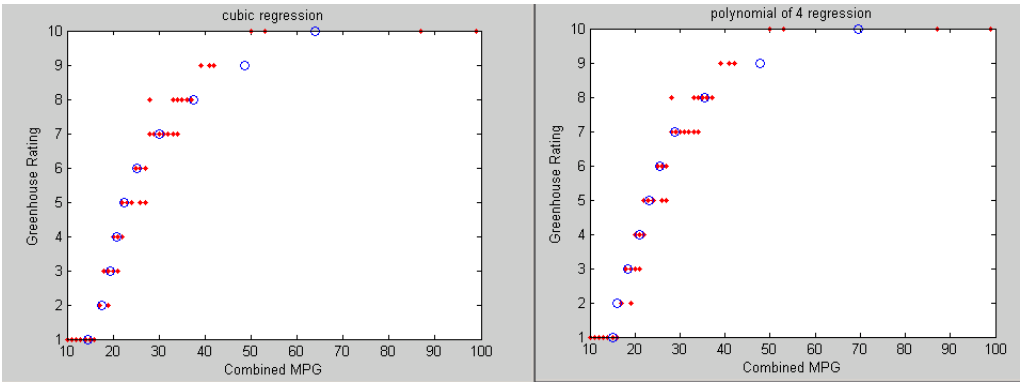
See the results section for more meaningful results.

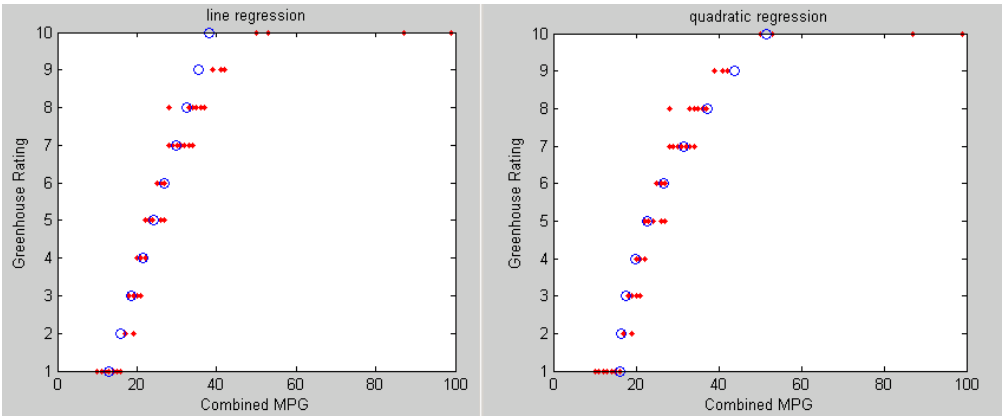
Curve Fitting

I also attempted to fit the best curve between the greenhouse scores and the MPG combined score in order to minimize the error ||Ax – b||.

```
combMPG = Cars(:,10); %u
ghouse = Cars(:,11); %v

%fit to a line
poly = polyfit(ghouse, combMPG, 1)';
pop = polyval(poly, ghouse);
errors1 = pop - combMPG;
norm(errors1) % gets 148.3806
plot(combMPG, ghouse, 'r.');
hold on
plot(pop, ghouse, 'o');
title('line regression');
xlabel('Combined MPG');
ylabel('Greenhouse Rating');

%fit to quad
figure;
poly2 = polyfit(ghouse, combMPG, 2)';
pop2 = polyval(poly2, ghouse);
errors2 = pop2 - combMPG;
norm(errors2) % gets 116.0861
plot(combMPG, ghouse, 'r.');
hold on
plot(pop2, ghouse, 'o');
title('quadratic regression')
xlabel('Combined MPG');
ylabel('Greenhouse Rating');

%fit to cubic
figure
poly3 = polyfit(ghouse, combMPG, 3)';
pop3 = polyval(poly3, ghouse);
errors3 = pop3 - combMPG;
norm(errors3) %gets 93.3082
plot(combMPG, ghouse, 'r.');
hold on
plot(pop3, ghouse, 'o');
title('cubic regression');
xlabel('Combined MPG');
ylabel('Greenhouse Rating');

%fit to power of 4
figure
poly4 = polyfit(ghouse, combMPG, 4)';
```

```
pop4 = polyval(poly4, ghouse);
errors4 = pop4 - combMPG;
norm(errors4) %gets 85.3078
plot(combMPG, ghouse, 'r.');
hold on
plot(pop4, ghouse, 'o');
title('polynomial of 4 regression')
xlabel('Combined MPG');
ylabel('Greenhouse Rating');
```





The red points represent the actual values and the blue circles represent the estimated values of the model. Also, the calculated error ||Ax – b|| gets smaller each iteration as the polynomial power is increased. Visually, the cars all cluster together in a curve, but there are a few cars with greenhouse ratings of 10 and very high MPG scores, that are have too much power over the error. Please see the Results section for how I removed the outliers and the resulting plots.

Fixing the Dataset (again)

Also, at this point, I realized that I was depending on the Combined MPG rating because it is more practical than picking between City MPG and Highway MPG. However, upon looking at my results more closely, I noticed that whoever is responsible for doing the math chose to round to the nearest whole number, which is not what I would do. See for example the table below:

	Model Name	Displ	Cyl	Trans	Gears	Drive	Veh Cl	Air Poll	City MP	Hwy MI	Cmb M
517	FORD Escape	2.5	4	4	5	4	5	6	23	28	25
520	FORD Escape	2.5	4	4	5	4	5	5	23	28	25
559	FORD Fusion	2.5	4	2	6	4	2	6	22	30	25

The Ford Escape has a 23 city MPG rating and a 28 highway MPG rating. In comparison, the Ford Fusion has a 22 city MPG rating and a 30 highway MPG rating. However, they both have the same Combined MPG rating according to the dataset. Even though the MPG ratings are 'estimates only' I do not see why this number also has to be an estimate. One car definitely has a higher MPG than the other car, and I think that the combined MPG score should reflect that.

As a result, I used the EPA's mathematical formula 1/[(0.55/city mpg) + (0.45/highway mpg)] to evaluate the Combined MPG column myself. At first, I thought that using their combined MPG would not be much different than my recalculated MPG. However, I chose to redo almost all my analysis with my calculation of the MPG after realizing the significant difference that occurred with Linear Least Squares (see next section).

I already talked about how I had to eliminate incomplete data from the cars. Now, I had to decide whether to recalculate the combined MPG. I had already done a lot of work in the result section, and I wanted to see if it was necessary to redo.

Here is a snippet of CmbMPG_new – CmbMPG_old:

-1.0512
1.0512
-1.0512
0.6165
0.0290
-0.5943
0.4560
0.7164
-1.1724
0.1724
0.4717
-0.6441
0.1724
0.4717
-0.4129
0.2371

As you can see, it seems the original combined MPG is a poor indication of the other two MPG ratings. Some cars even had a difference greater than 0.5 between the two scores! I concluded that redoing it was worthwhile and decided to email the Environmental Protection agency that their combined MPG ratings are misleading.

**Note: if re-running with the 'corrected' combined MPG did not produce significantly different results, than I simply left the answers in the results section with the original Combined MPG values.

Attempt at Linear Least Squares

```
%Linear Least squares code, using the original combined MPG rating
meanCol = mean(Cars(:,1:11));
stdCol = std(Cars(:,1:11));
Mean = ones(m,1)*meanCol;
Std = ones(m,1)*stdCol;
Z_cars = (Cars(:,1:11) - Mean)./Std;
smartway = Cars(:,12);

c = Z_cars\smartway;

%Results of Linear Least squares:
c =

   -0.0305  %Displ
    0.0965  %Cyl
   -0.0099  %Trans
   -0.0109  %Gears
   -0.0067  %Drive
    0.0003  %VehClass
    0.0736  %AirPoll
   -0.1134  %City MPG
   -0.0472  %Highway MPG
    0.2366  %Combined MPG
    0.2026  %Greenhouse Score
```

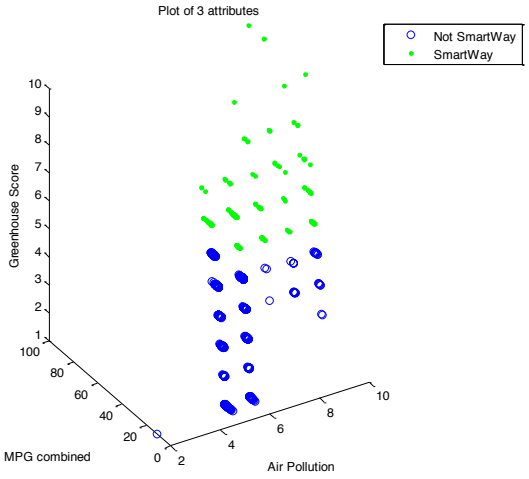
By using the simple principle of x =A\ b, with b containing '1' for SmartWay and '0' for not SmartWay, the best fit can be found. The results show that SmartWay cars, which were computed in the "Green Vehicle Guide" using the Air Pollution and Greenhouse scores, also have a strong correspondence with the Combined MPG of the car.

However, upon closer inspection, the linear least squares solution above was actually incorrect. When the improved combined MPG score is used, significantly different results are obtained. Instead, the Combined MPG score is even a larger indication of Smartway. See the results section for the correct results and note that this showed me that re-running all my code might be a good idea.

Results

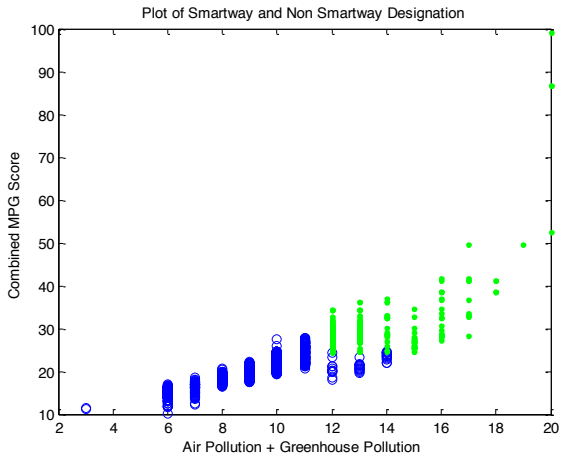
I) Smartway

```
%retrieve smartway cars only
count = 0;
list_smartway = [ ]; list_notsmartway = [ ];
for i = 1:m
    if ( Cars(i, 12)== 1)
        list_smartway = [list_smartway i];
    else
        list_notsmartway = [list_notsmartway i];
    end
end
%//////////////////////////////
airpollution = Cars(:,7);
cmbMPG = Cars(:,10);
greenhouse = Cars(:,11);
plot3(airpollution(list_notsmartway), cmbMPG(list_notsmartway) , greenhouse(list_notsmartway),
'o');
hold on;
plot3(airpollution(list_smartway),
cmbMPG(list_smartway),
greenhouse(list_smartway), 'g.');
rotate3d on
legend('Not SmartWay', 'SmartWay');
xlabel('Air Pollution'); ylabel('MPG
combined'); zlabel('Greenhouse Score');
title('Plot of 3 attributes');

figure
sumS = greenhouse + airpollution;
plot(sumS(list_notsmartway),
cmbMPG(list_notsmartway), 'o')
hold on; plot(sumS(list_smartway),
cmbMPG(list_smartway), 'g.');
xlabel('Air Pollution + Greenhouse Pollution
'); ylabel('Combined MPG Score ');
title('Plot of Smartway and Non Smartway
Designation');
```



In the figure to the top right, the blue circles represent that were labeled as 'Not Smartway' and the green dots represent vehicles that were labeled as 'SmartWay'. It can be observed that the Smartway vehicles have the higher Air Pollution, MPG combined, and Greenhouse score values. Also, there are a few outliers, which are most likely the electric cars. It seems that the divide between whether the car is Smartway or not occurs around where the greenhouse score is 5, and the air pollution score is not very impactful. I investigate this more in a later section.

The plot on the next page represents more accurately the method in which EPA evaluated whether cars were SmartWay. The website says the sum of Air Pollution and Greenhouse scores was calculated and used to determine whether the SmartWay designation was awarded. However, notice that for when the sum is 14, for example, the cars with lower MPG were not awarded the SmartWay



Designation, although cars with a sum of 13 but higher MPG were considered SmartWay. Therefore, their method must be more elaborate than they specified.

Throughout the rest of the report, I sometimes use the SmartWay designation to analyze other characteristics. Upon closer investigation, I discovered that the designation is achieved if a vehicle gets a combined (Air Pollution and Greenhouse Pollution) score of ≥ 12 and also a minimum greenhouse score of six and a minimum air pollution score of five.

2) Linear Least Squares

```
%Linear Least squares code
meanCol = mean(Cars(:,1:11));
stdCol = std(Cars(:,1:11));
Mean = ones(m,1)*meanCol;
Std = ones(m,1)*stdCol;
Z_cars = (Cars(:,1:11) - Mean)./Std;
smartway = Cars(:,12);

c = Z_cars\smartway;

Results of Linear Least Squares:
c =

    0.0035 %Displ
    0.1001 %Cyl
   -0.0106 %Trans
   -0.0046 %Gears
   -0.0048 %Drive
   -0.0208 %Vehicle Class
    0.0852 %Air Poll Rating
   -2.5559 %City MPG
   -1.2981 %Highway MPG
    3.9480 %Combined MPG
    0.1272 %Greenhouse score
```

By using the simple principle of x =A\ b, with b containing '1' for SmartWay and '0' for not smartway, the best fit can be found. The results show that for SmartWay cars, although the Air Pollution and Greenhouse scores were used to compute them, there is also a very strong correspondence with the Combined MPG of the car. Also, we can conclude that although the 'SmartWay' cars were in the top 20% for their Greenhouse and Air Pollution score, due to the nature of the scores over all cars, the Greenhouse score is a much more contributing factor because it has a larger variance. To be more specific, the

standard deviation of the Greenhouse scores is 1.9176 and of the air pollution scores is 1.0076, which is the smallest standard deviation out of all the attributes.

3) Vehicle Type

By plotting the mileage and the greenhouse scores in a way where the different types of the vehicles were apparent, I hoped to find a relationship between them. The vehicle types were coded in as numbers since Matlab requires numerical values.

The following numerical values were used:

- Small car: 1
- Midsize car: 2
- Large car: 3
- Stationwagon: 4
- SUV: 5
- Minivan: 6
- Van: 7
- Pickup: 8
- Special purpose: 9

Lists were created for created with the index of each type:

```
vech_type = Cars(:, 6);
list_small = (find(vech_type == 1))';
list_mid= (find(vech_type == 2))';
list_large = (find(vech_type == 3))';
list_stationwagon = (find(vech_type == 4))';
list_SUV = (find(vech_type == 5))';
list_minivan = (find(vech_type == 6))';
list_van = (find(vech_type == 7))';
list_pickup = (find(vech_type == 8))';
list_other = (find(vech_type == 9))'; %"special purpose"

cmbMPG = Cars(:,10);
greenhouse = Cars(:,11);
```

Also, to calculate the quantitative results:

```
meanMPG = [ mean(Cars(list_small, 10)) mean(Cars(list_mid, 10))
mean(Cars(list_large, 10))....
    mean(Cars(list_stationwagon, 10)) mean(Cars(list_SUV, 10))
mean(Cars(list_minivan,  10))....
    mean(Cars(list_van, 10)) mean(Cars(list_pickup, 10)) mean(Cars(list_other,
10)) ];

stdMPG = [ std(Cars(list_small, 10)) std(Cars(list_mid, 10)) std(Cars(list_large,
10)) ....
    std(Cars(list_stationwagon, 10)) std(Cars(list_SUV, 10))
std(Cars(list_minivan,  10))....
    std(Cars(list_van, 10)) std(Cars(list_pickup, 10)) std(Cars(list_other, 10))];

meanGreen = [ mean(Cars(list_small, 11)) mean(Cars(list_mid, 11))
mean(Cars(list_large, 11)) ....
```

```
    mean(Cars(list_stationwagon, 11)) mean(Cars(list_SUV, 11))
mean(Cars(list_minivan,  11))....
    mean(Cars(list_van, 11)) mean(Cars(list_pickup, 11)) mean(Cars(list_other,
11)) ];

stdGreen = [ std(Cars(list_small, 11)) std(Cars(list_mid, 11))
std(Cars(list_large, 11)) ....
    std(Cars(list_stationwagon, 11)) std(Cars(list_SUV, 11))
std(Cars(list_minivan,  11))....
    std(Cars(list_van, 11)) std(Cars(list_pickup, 11)) std(Cars(list_other, 11))
];

display(meanMPG)
display(stdMPG)
display(meanGreen)
display(stdGreen)
```

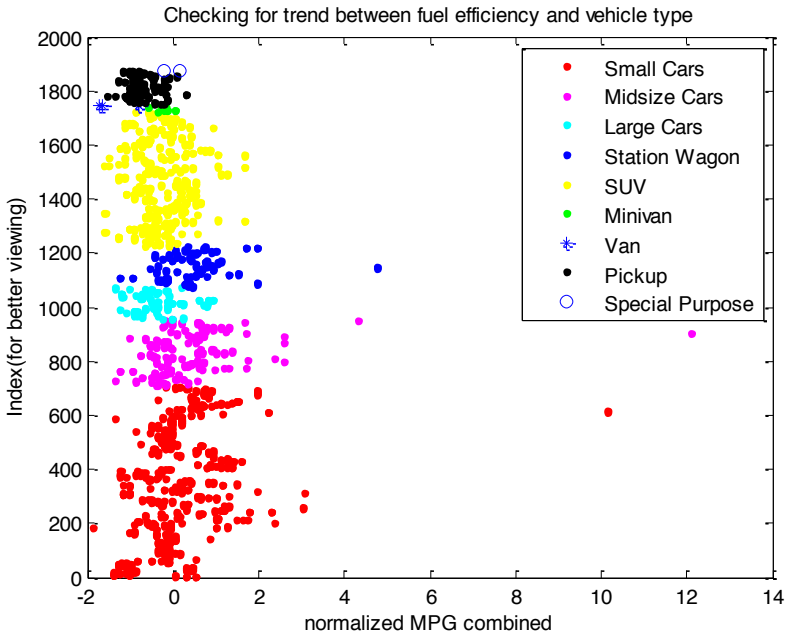
Results	Mean Greenhouse score	Std Deviation greenhouse score	Mean Combined MPG	Std Devation combined MPG	amount
1 small	4.6667	1.9777	23.2200	7.1390	708
2 midsize	5.0122	1.8104	24.3820	8.6854	246
3 large	3.2066	1.6377	19.2877	3.6660	121
4 station wagon	5.3000	1.4225	24.8470	5.0070	150
5 SUV	3.6024	1.6106	20.2140	3.7145	498
6 minivan	4.0000	0.4714	20.5656	1.0265	19
7 van	1.6667	0.5164	15.0400	2.8855	6
8 pickup	2.4646	1.2459	17.8893	2.3649	127
9 special purpose	4.5000	0.7071	21.7922	1.6932	2

From the results in the table, it is apparent that the station wagon has the highest greenhouse and MPG scores. However, surprisingly, the midsize cars (not the small cars) are in second place. This may be due to many sports cars being classified as 'small cars,' and they are usually not green cars. However, these results are a huge generalization, since the standard deviation is very high. There is very wide-spread data for small, midsize, large cars, station wagons, and SUVs. The others have smaller standard deviations since they are smaller in number, although the pickup is an exception. By plotting, we will be able to visually see where each car stands in relation to each other.

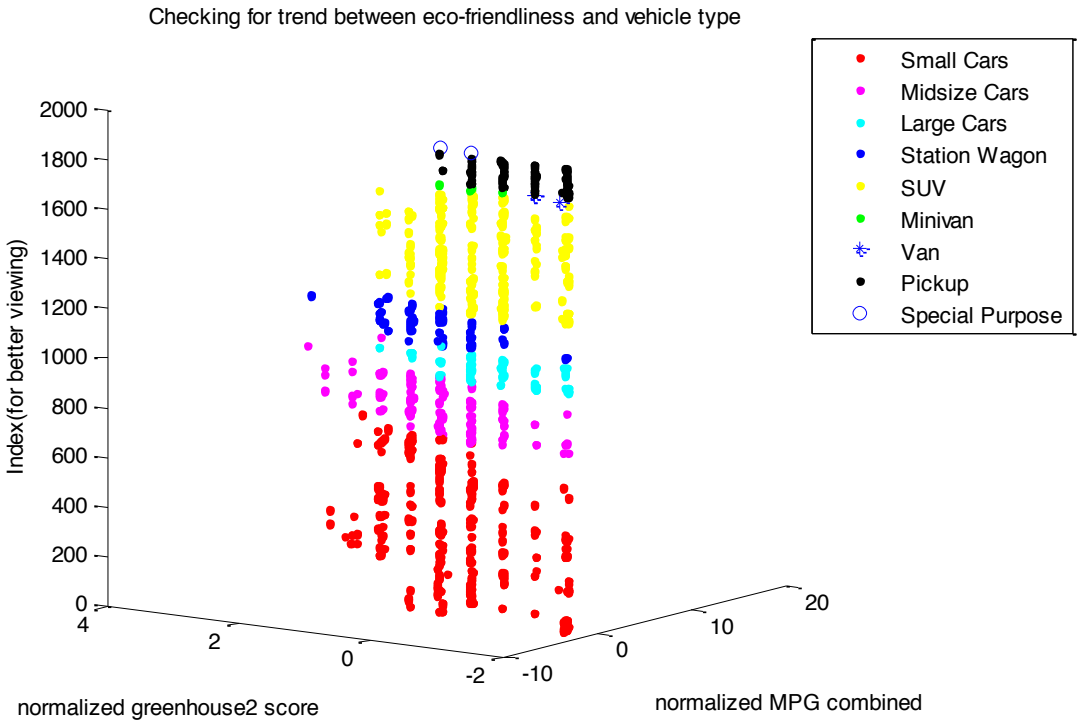
Next, to plot the results in an organized way (unlike in the Log):

```
%normalize the scores for better results
cmbMPG2 = (cmbMPG - mean(cmbMPG))./std(cmbMPG);
greenhouse2 = (greenhouse - mean(greenhouse))./std(greenhouse);

szsmall = length(list_small); szmid = length(list_mid); szlarge =
length(list_large);
szwagon = length(list_stationwagon); szSUV = length(list_SUV); szminivan =
length(list_minivan);
```

```
szvan = length(list_van); szpickup = length(list_pickup); szother =
length(list_other);
one = 1:szsmall;  next = szsmall+1;
two = next:next+szmid-1; next=next+szmid;
three = next:next+szlarge-1; next = next+szlarge;
four = next:next+szwagon-1; next = next+szwagon;
five = next:(next+szSUV-1); next = next+szSUV;
six = next:(next+szminivan-1); next = next+szminivan;
seven = next:(next+szvan-1); next = next+szvan;
eight = next:(next+szpickup-1); next = next+szpickup;
nine = next:(next+szother-1);

plot3(cmbMPG2(list_small), greenhouse2(list_small), one, 'r.');
hold on
plot3(cmbMPG2(list_mid), greenhouse2(list_mid),two , 'm.');
hold on
plot3(cmbMPG2(list_large), greenhouse2(list_large), three, 'c.');
hold on
plot3(cmbMPG2(list_stationwagon), greenhouse2(list_stationwagon), four,  'b.');
hold on
plot3(cmbMPG2(list_SUV), greenhouse2(list_SUV),five, 'y.');
hold on
plot3(cmbMPG2(list_minivan), greenhouse2(list_minivan), six, 'g.');
hold on
plot3(cmbMPG2(list_van), greenhouse2(list_van),seven, '*');
hold on
plot3(cmbMPG2(list_pickup), greenhouse2(list_pickup),eight, 'k.');
hold on
plot3(cmbMPG2(list_other), greenhouse2(list_other),nine, 'o');
rotate3d on;
xlabel('normalized MPG combined'); ylabel('normalized greenhouse2 score');
zlabel('Index(for better viewing)');
title('Checking for trend between eco-friendliness and vehicle type');
legend('Small Cars', 'Midsize Cars', 'Large Cars', 'Station Wagon', 'SUV',
'Minivan', 'Van', 'Pickup', 'Special Purpose');

figure
plot(cmbMPG2(list_small),  one, 'r.'); hold on
plot(cmbMPG2(list_mid),  two, 'm.'); hold on
plot(cmbMPG2(list_large), three, 'c.'); hold on
plot(cmbMPG2(list_stationwagon), four,  'b.'); hold on
plot(cmbMPG2(list_SUV), five, 'y.'); hold on
plot(cmbMPG2(list_minivan),  six, 'g.'); hold on
plot(cmbMPG2(list_van), seven, '*'); hold on
```



```
plot(cmbMPG2(list_pickup), eight, 'k.'); hold on
plot(cmbMPG2(list_other), nine, 'o');
xlabel('normalized MPG combined'); ylabel('Index(for better viewing)');
title('Checking for trend between fuel efficiency and vehicle type');
legend('Small Cars', 'Midsize Cars', 'Large Cars', 'Station Wagon', 'SUV',
'Minivan', 'Van', 'Pickup', 'Special Purpose');
```



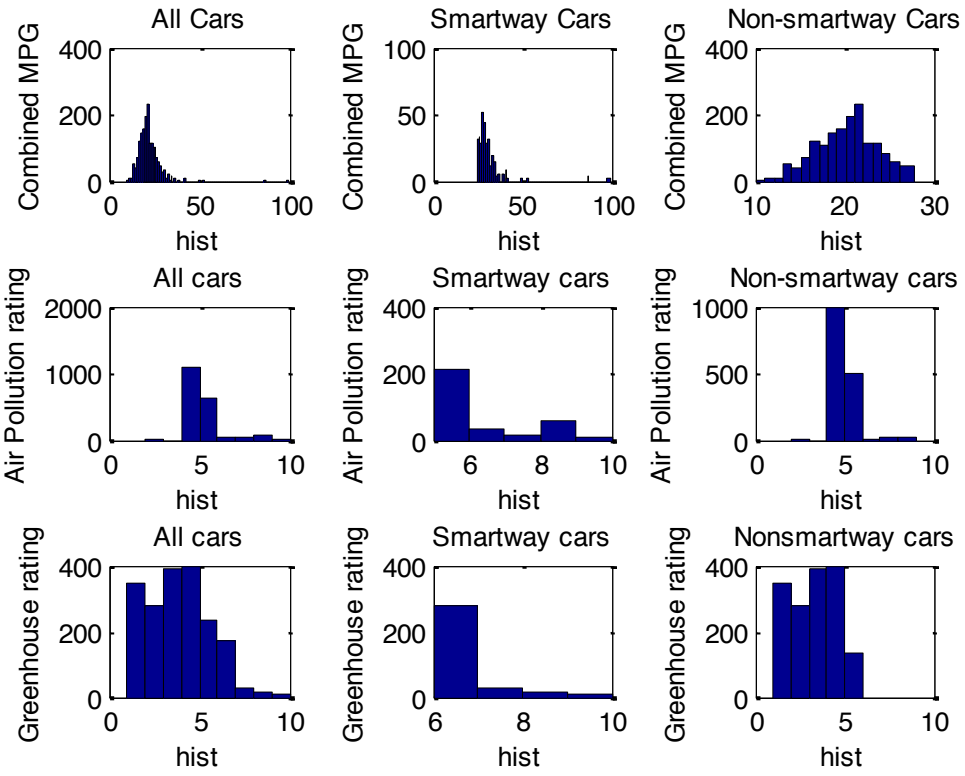
Note that these two plots have an index in the z-direction to prevent data points from being superimposed on top of each other. To understand the data, look only in the horizontal direction.

General summary:

- Small car
  - MPG: some cars have very low MPG, some have very high; huge variance
  - Greenhouse: thoroughly dispersed ratings
- Midsize car
  - MPG: slightly better than small cars
  - Greenhouse: thoroughly dispersed ratings
- Large car:
  - MPG: clustered at lower values
  - Greenhouse: skewed towards poor ratings
- Station wagon
  - MPG: clustered at higher values
  - Greenhouse: skewed towards higher ratings
- SUV
  - MPG: clustered at lower values, worst so far
  - Greenhouse: thoroughly dispersed
- Minivan
  - MPG: near average, Greenhouse: near average

- Van
  - MPG: very poor, Greenhouse: very poor
- Pickup
  - MPG: clustered at below average ratings
  - Greenhouse: skewed to poor ratings
- Special purpose:
  - MPG: slightly above average
  - Greenhouse: slightly below

4) Histograms (Chi-Squared)




```
%create histograms
close all;
subplot(3,3,1);
CmbMPG = Cars(:,10);
maxM = max(CmbMPG);
minM = min(CmbMPG);
diff = maxM - minM;
hist(CmbMPG, diff );
xlabel('hist'); ylabel('Combined MPG'); title('All Cars');

subplot(3,3,2);
list_smartway = find(Cars(:,12) == 1);
list_smartway = list_smartway';
diff2 = max(CmbMPG(list_smartway)) - min(CmbMPG(list_smartway));
hist(CmbMPG(list_smartway), diff2 );
xlabel('hist'); ylabel('Combined MPG'); title('Smartway Cars');

subplot(3,3,3);
list_nonsmartway = find(Cars(:,12) == 0);
```

```
list_nonsmartway = list_nonsmartway';
diff3 = max(CmbMPG(list_nonsmartway)) - min(CmbMPG(list_nonsmartway));
hist(CmbMPG(list_nonsmartway), diff3 );
xlabel('hist'); ylabel('Combined MPG'); title('Non-smartway Cars');

subplot(3,3,4);
airpollution = Cars(:,7);
diffA = max(airpollution) - min(airpollution);
hist(airpollution, diffA);
xlabel('hist'); ylabel('Air Pollution rating'); title('All cars');

subplot(3,3,5)
diffA2 = max(airpollution(list_smartway)) - min(airpollution(list_smartway));
hist(airpollution(list_smartway), diffA2);
xlabel('hist'); ylabel('Air Pollution rating'); title('Smartway cars');

subplot(3,3,6)
diffA3 = max(airpollution(list_nonsmartway)) - min(airpollution(list_nonsmartway));
hist(airpollution(list_nonsmartway), diffA3);
xlabel('hist'); ylabel('Air Pollution rating'); title('Non-smartway cars');

subplot(3,3,7)
greenhouse= Cars(:,11);
diffG = max(greenhouse)-min(greenhouse);
hist(greenhouse, diffG)
xlabel('hist'); ylabel('Greenhouse rating'); title('All cars');

subplot(3,3,8)
diffG2 = max(greenhouse(list_smartway)) - min(greenhouse(list_smartway));
hist(greenhouse(list_smartway), diffG2)
xlabel('hist'); ylabel('Greenhouse rating'); title('Smartway cars');

subplot(3,3,9)
diffG3 = max(greenhouse(list_nonsmartway)) - min(greenhouse(list_nonsmartway));
hist(greenhouse(list_nonsmartway), diffG3);
xlabel('hist'); ylabel('Greenhouse rating'); title('Nonsmartway cars');
```

	Mean	Standard Deviation
Combined MPG (all cars)	22.0361	6.3693
Greenhouse rating (all cars)	4.2206	1.9176
Air pollution rating (all cars)	5.6276	1.0076
Combined MPG (Smartway)	30.8232	9.3483
Greenhouse rating (Smartway)	6.9604	0.9261
Air pollution (Smartway)	6.5640	1.4825

The histogram above helps gain a better understanding of the significance of the ratings in relationship to one another. Also, MPG seems to have a chi-squared distribution, with almost all of the right side corresponding to greener vehicles. Also, as I predicted from the linear least squares estimate, the greenhouse rating more clearly defines a greener vehicle than does the air pollution rating. The air pollution rating is 4 or 5 for most cars, neither of which are acceptable for 'SmartWay' cars. Also, while a few cars have a high air pollution rating but a low greenhouse rating, no cars have a high greenhouse rating and a low air pollution rating. In other words, the greenhouse rating is a more accurate indicator of a greener vehicle.

length(find(greenhouse(list_smartway) == 6)) → 101

length(find(greenhouse(list_nonsmartway) == 6)) → 134
length(find(greenhouse(list_smartway) == 7)) → 175
length(find(greenhouse(list_nonsmartway) == 7)) → 0
length(find(greenhouse(list_nonsmartway) == 8)) → 0
length(find(airpollution(list_nonsmartway) == 7)) → 4
length(find(airpollution(list_nonsmartway) == 8)) → 17
length(find(airpollution(list_nonsmartway) == 9)) → 27

5) Principal Components Analysis

By finding the Singular Value Decomposition of the cars matrix, I used PCA three times: on transmission, gears, and fuel type.

```
%find SVD of covarinace matrix of the data
C = cov(Cars);
[U,S,V] = svd(C);

PrincComp1 = U(:,1);
PrincComp2 = U(:,2);

display(PrincComp1); display(PrincComp2);

X = Cars*PrincComp1;
Y = Cars*PrincComp2;
```

Results	Principal Component 1	Principal Component 2
Displ	-0.0790	-0.0385
Cyl	-0.1101	0.0492
Trans	0.0322	0.0073
Gears	-0.0566	0.2817
Drive	-0.0570	-0.2856
Vehicle Class	-0.0534	-0.6962
Air Pollution Score	0.0375	-0.0417
City MPG	0.5846	-0.3832
Hwy MPG	0.5370	0.4193
Cmb MPG	0.5639	-0.0960
Greenhouse gas	0.1460	0.1293
Smart way	0.0216	0.0132

The first principal component seems to approximately be based on the three kinds MPG. The other values are very small. A large first component mostly means high MPG. Also, the greenhouse gas is the next largest score, also contributing to a larger first component.

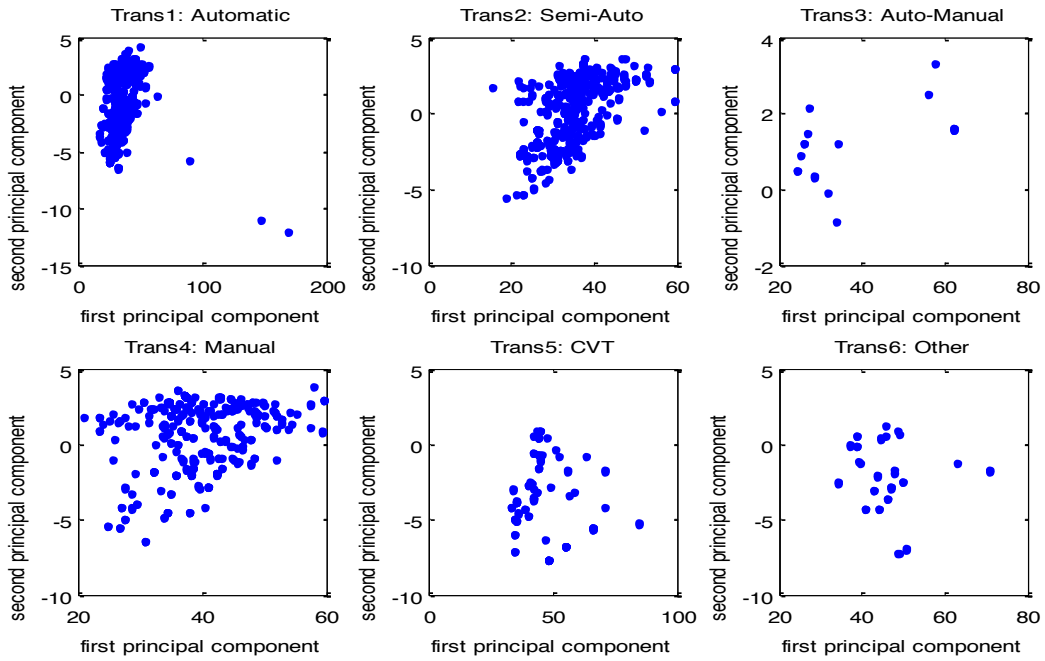
The second principal component has four major components: highway MPG, city MPG, gears, and drive. Notice the Combined MPG score is small even though it is calculated based on the city MPG and the highway MPG. I can then conclude that since one MPG is positive and the other is negative, they

are canceling each other out. Also, the gears and drive have opposite signs. A higher value for the projection onto the second principal component would have a large number of gears and a small number representing the drive, while a smaller value for the projection would have few gears and a large drive.

Transmission

```
trans_type = Cars(:, 3);
list_trans1 = (find(trans_type == 1))';
list_trans2 = (find(trans_type == 2))';
list_trans3 = (find(trans_type == 3))';
list_trans4 = (find(trans_type == 4))';
list_trans5 = (find(trans_type == 5))';
list_trans6 = (find(trans_type == 6))';

meanMPG = [mean(Cars(list_trans1, 10)) mean(Cars(list_trans2, 10)) mean(Cars(list_trans3, 10))....
    mean(Cars(list_trans4, 10)) mean(Cars(list_trans5, 10)) mean(Cars(list_trans6,  10)) ];

stdMPG = [ std(Cars(list_trans1, 10)) std(Cars(list_trans2, 10)) std(Cars(list_trans3, 10))....
    std(Cars(list_trans4, 10)) std(Cars(list_trans5, 10)) std(Cars(list_trans6,  10)) ];

meanGreen = [ mean(Cars(list_trans1, 11)) mean(Cars(list_trans2, 11)) mean(Cars(list_trans3, 11))....
    mean(Cars(list_trans4, 11)) mean(Cars(list_trans5, 11)) mean(Cars(list_trans6,  11)) ];

stdGreen = [ std(Cars(list_trans1, 11)) std(Cars(list_trans2, 11)) std(Cars(list_trans3, 11))....
    std(Cars(list_trans4, 11)) std(Cars(list_trans5, 11)) std(Cars(list_trans6,  11)) ];

subplot(2,3,1)
plot(X(list_trans1), Y(list_trans1), '.');
xlabel('first principal component'); ylabel('second principal component')
title('Trans1: Automatic')

subplot(2,3,2)
plot(X(list_trans2), Y(list_trans2), '.');
xlabel('first principal component'); ylabel('second principal component')
title('Trans2: Semi-Auto')

subplot(2,3,3)
plot(X(list_trans3), Y(list_trans3), '.');
xlabel('first principal component'); ylabel('second principal component')
title('Trans3: Auto-Manual')

subplot(2,3,4)
plot(X(list_trans4), Y(list_trans4), '.');
xlabel('first principal component'); ylabel('second principal component')
title('Trans4: Manual')

subplot(2,3,5)
plot(X(list_trans5), Y(list_trans5), '.');
xlabel('first principal component'); ylabel('second principal component')
title('Trans5: CVT')

subplot(2,3,6)
plot(X(list_trans6), Y(list_trans6), '.');
xlabel('first principal component'); ylabel('second principal component')
title('Trans6: Other')
```

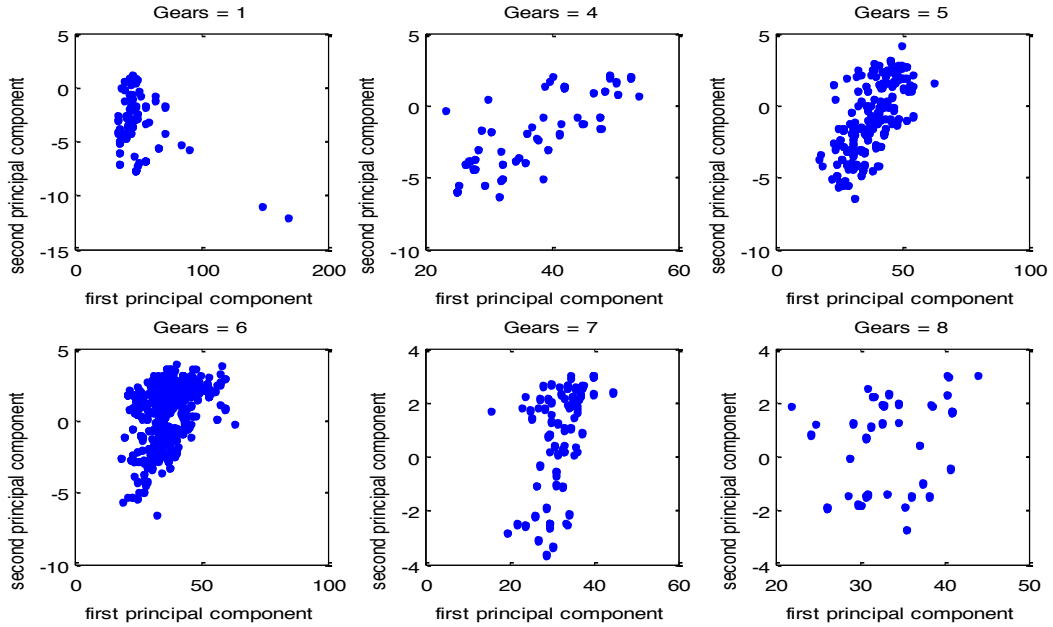
Results	Mean Greenhouse score	Std Deviation greenhouse score	Mean Combined MPG	Std Deviation combined MPG	amount
1 auto	3.7289	1.9699	21.3254	8.6492	568
2 semi-auto	3.8026	1.6063	20.5989	3.8612	618
3 auto-manual	3.0000	2.7543	20.4135	7.6384	30
4 manual	4.7505	1.7308	22.9554	4.4269	501
5 CVT	6.1743	1.6602	27.8113	6.7180	109
6 other	6.0980	1.1359	26.9455	4.8893	51



The automatic vehicles are all close together in the top corner, with mostly x-values between 20 and 60 and y-values between 4 and -6. The semi-automatic vehicles also have a similar PCA result, but closer to the center. The manual vehicles have similar behavior but with no outliers and most of the points concentrates at the highest y-values. The auto-manual vehicles seem to be centered at 50 in the x-direction and 2 in the y-direction. The CVT and 'other' vehicles also seem to be centered at 50 and have lower values in the y-direction.

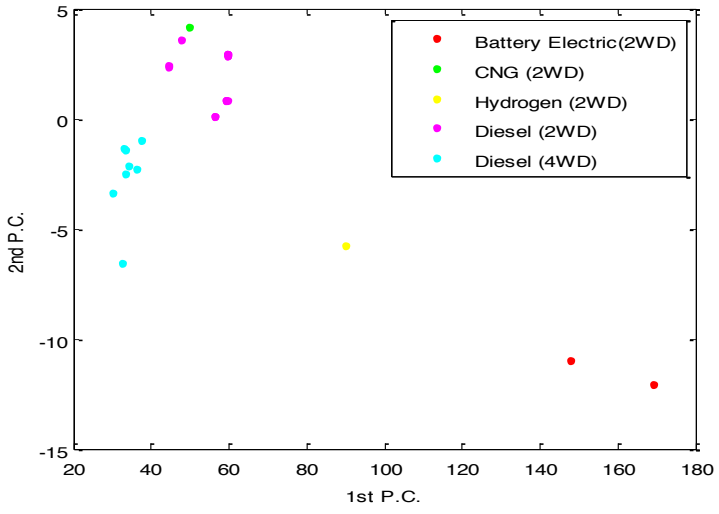
These results generally demonstrate that cars with CVT, Auto-Manual, and other transmissions have higher MPG ratings than do cars with Automatic, Semi-Automatic, Auto-Manual, and Manual transmission (from the x-component). Also, these types of transmissions tend to have a higher number for the Vehicle class and drive numbers that were encoded.

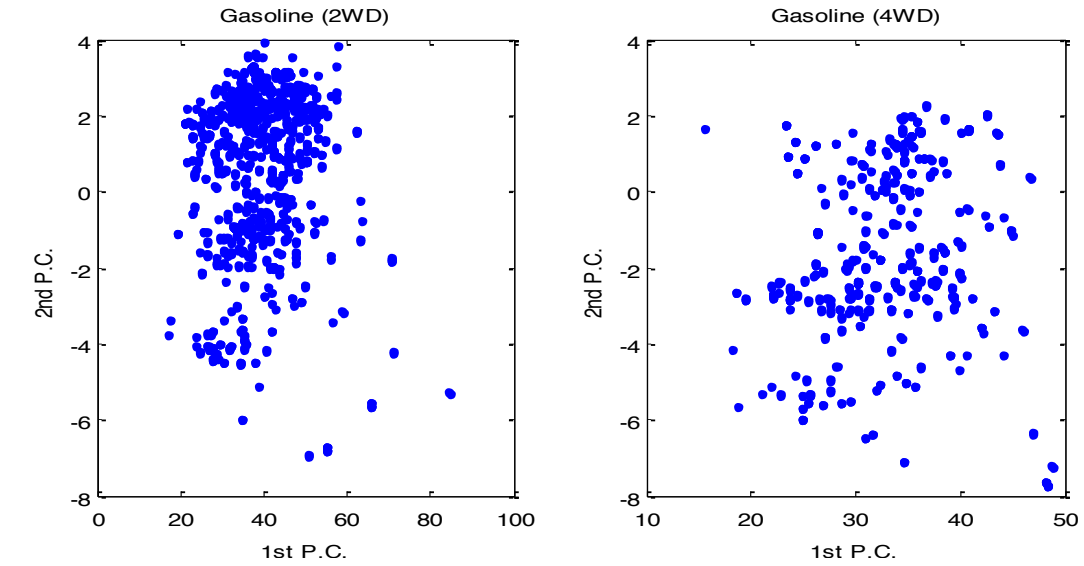
Gears

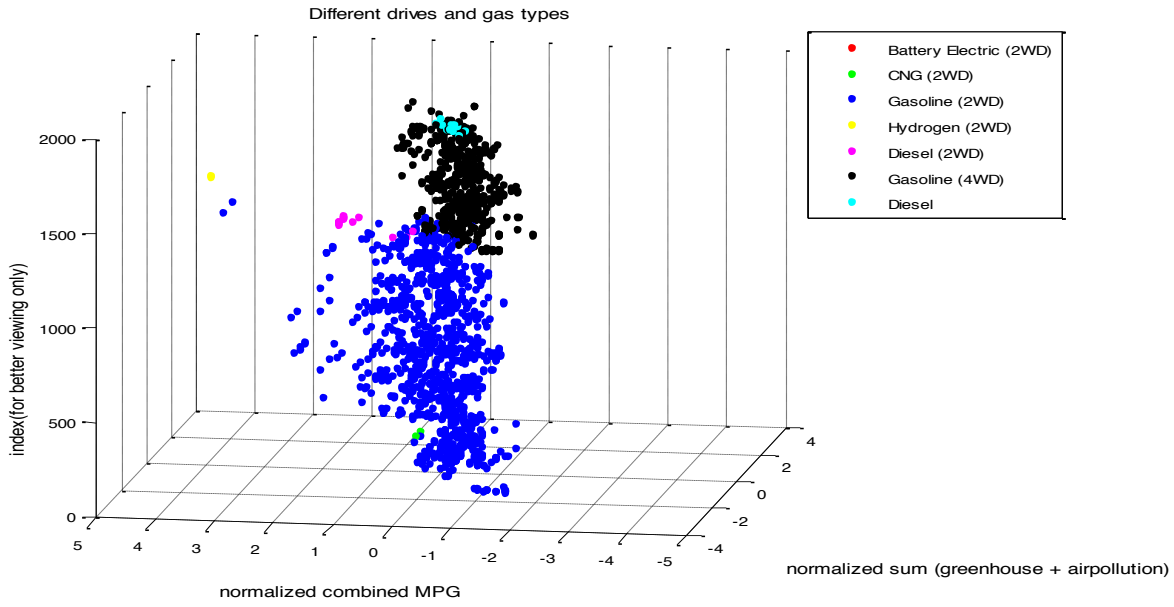


The number of gears seems to have very little correlation to the Combined MPG scores. The first principal component in all of the above plots seems to mostly be concentrated between 20 and 60, with gears = 7 or 8 having slightly worse combined MPG. Also, as expected, gears = 1 has slightly better values in addition to some outliers, since the alternative fuel vehicles tend to have one gear. The second principal component is not much more interesting. The values become more negative with an increase in the number of gears, as expected because gears are a significant part of the component.

Drive/Fuel Type





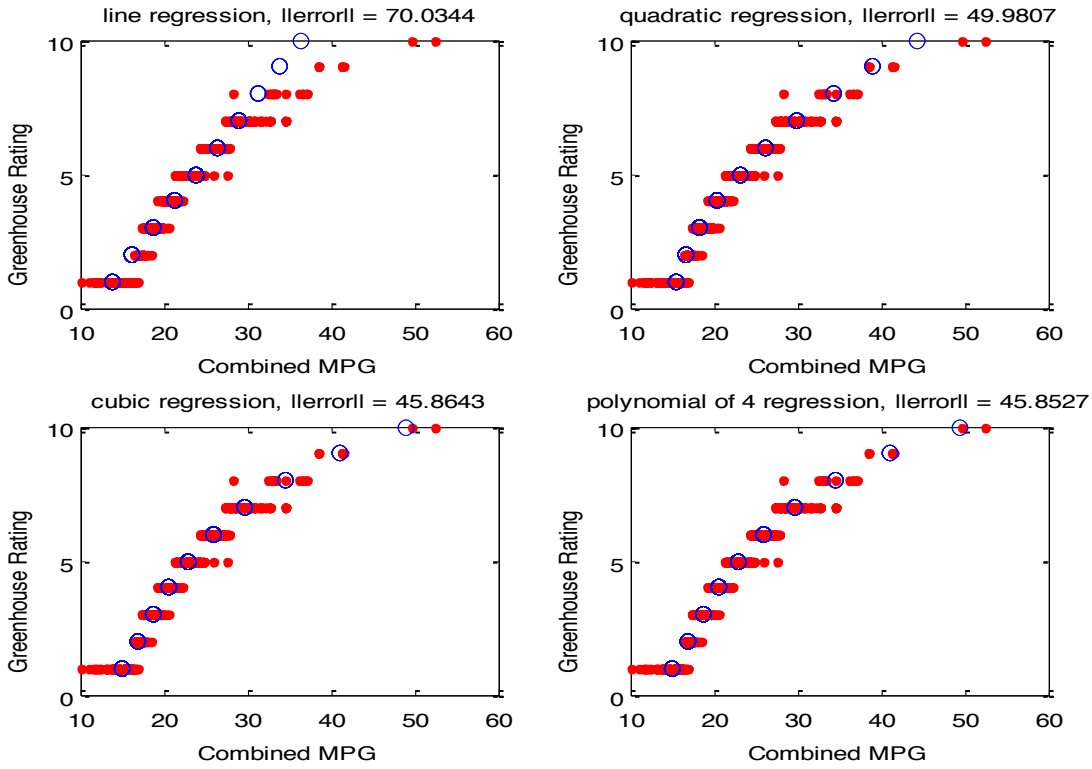


It is clear from the results above that the alternative fuel vehicles are much better than the regular gasoline vehicles, generally speaking. However, there are some 2WD gasoline vehicles that have very good eco-friendly scores. The average diesel vehicle is better than the average gasoline car. Also, it is clear from the results above that cars with 4WD have by far the worst ratings in terms of eco-friendliness.

There are very few non-gasoline vehicles. Notice that the one hydrogen vehicle is the Mercedes Benz F-Cell. The electric vehicle models are Nissan Leaf, SMART ForTwo Cabriolet, and Smart ForTwo Coupe. The CNG vehicle is a Honda Civic model.

6) Linear Least Squares Regression

```
%remove a few outliers before doing linear least squares regression
Cars2 = Cars;
cmbMPG = Cars(:,10);
toErase = find(cmbMPG > 55);
Cars2(toErase,:) = [ ];

combMPG = Cars2(:,10); %u
ghouse = Cars2(:,11); %v

%fit to a line
subplot(2,2,1)
poly = polyfit(ghouse, combMPG, 1)';
pop = polyval(poly, ghouse);
errors1 = pop - combMPG;
norm(errors1)
plot(combMPG, ghouse, 'r.');
hold on
plot(pop, ghouse, 'o');
title(['line regression, ||error|| = ', num2str(norm(errors1))]);
xlabel('Combined MPG'); ylabel('Greenhouse Rating');

%fit to quad
subplot(2,2,2)
poly2 = polyfit(ghouse, combMPG, 2)';
pop2 = polyval(poly2, ghouse);
errors2 = pop2 - combMPG;
plot(combMPG, ghouse, 'r.');
hold on
plot(pop2, ghouse, 'o');
title(['quadratic regression, ||error|| = ', num2str(norm(errors2))])
xlabel('Combined MPG'); ylabel('Greenhouse Rating');


%fit to cubic
subplot(2,2,3)
poly3 = polyfit(ghouse, combMPG, 3)';
pop3 = polyval(poly3, ghouse);
errors3 = pop3 - combMPG;
plot(combMPG, ghouse, 'r.');
hold on
plot(pop3, ghouse, 'o');
title(['cubic regression, ||error|| = ', num2str(norm(errors3))]);
xlabel('Combined MPG'); ylabel('Greenhouse Rating');

%fit to power of 4
subplot(2,2,4)
poly4 = polyfit(ghouse, combMPG, 4)';
pop4 = polyval(poly4, ghouse);
errors4 = pop4 - combMPG;
plot(combMPG, ghouse, 'r.');
hold on
plot(pop4, ghouse, 'o');
title(['polynomial of 4 regression, ||error|| = ', num2str(norm(errors4))])
xlabel('Combined MPG'); ylabel('Greenhouse Rating');

disp(poly)
disp(poly2)
disp(poly3)
disp(poly4)
```



Polyfit results: y = Combined MPG, x = Greenhouse rating

Line: 2.5198 x + 11.2273 = y

Quadratic: 0.2562 x^2 + 0.3779 x + 14.7867 = y

Cubic: 0.0434 x^3 - 0.3341 x^2 + 2.6460 x + 12.6224 = y

Poly of 4: 0.0010 x^4 + 0.0239 x^3 - 0.2059 x^2 + 2.3278 x + 12.8458 = y

As can be seen from the results above, the difference between the observed and expected values became smaller every iteration. However, by the time we reached a quartic function, there was little improvement in the air and the constant in front of the fourth term was quite small.

Therefore, I think that the relationship between the Greenhouse rating and the MPG score is an increasing, cubic function. In other words, generally, a higher greenhouse score corresponds to a much higher combined MPG score.

7) Model comparisons

```
hondas = []; toyotas = []; hybrids = []; convertibles = []; ford = [];
volkswagen = []; chevy = [];
carNames = (char(Name));
for i=1:length(Name)
    str = carNames(i, :);
    pattern = 'HONDA';     k = strfind(str, pattern);
    pattern2 = 'TOYOTA';   k2 = strfind(str, pattern2);
    pattern3 = 'Hybrid';   k3 = strfind(str, pattern3);
    pattern4 = 'Convertible';   k4 = strfind(str, pattern4);
    pattern5 = 'FORD';       k5 = strfind(str, pattern5);
    pattern6 = 'VOLKSWAGEN';     k6 = strfind(str, pattern6);
    pattern7 = 'CHEVROLET';      k7 = strfind(str, pattern7);
    pattern8 = 'AUDI';     k8 = strfind(str, pattern8);
    pattern9 = 'NISSAN';   k9 = strfind(str, pattern9);
    pattern10 = 'HYUNDAI';     k10 = strfind(str, pattern10);
    pattern11 = 'MERCEDES';     k11 = strfind(str, pattern11);
    if (~isempty(k)) %not empty
        hondas = [hondas i];
    end
    if (~isempty(k2))
        toyotas = [toyotas i];
    end
    if (~(isempty(k3)))
        hybrids = [hybrids i];
    end
    if (~(isempty(k4)))
        convertibles = [convertibles i];
    end
    if (~(isempty(k5)))
        ford = [ford i];
    end
    if (~(isempty(k6)))
        volkswagen = [volkswagen i];
        end
    if (~(isempty(k7)))
        chevy = [chevy i];
    end
    if (~(isempty(k8)))
        audi = [audi i];
    end
    if (~(isempty(k9)))
        nissan = [nissan i];
    end
    if (~(isempty(k10)))
        hyundai = [hyundai i];
    end
    if (~(isempty(k11)))
        mercedes = [mercedes i];
    end
end

attrib = [7 10 12 6];
disp('Honda'); disp(mean(Cars(hondas,attrib)));
disp('Toyota');disp(mean(Cars(toyotas,attrib)));
disp('Ford');disp(mean(Cars(ford,attrib)));
disp('Volkswagen');disp(mean(Cars(volkswagen,attrib)));
disp('Chevrolet');disp(mean(Cars(chevy,attrib)));
disp('Audi');disp(mean(Cars(audi,attrib)));
disp('Nissan');disp(mean(Cars(nissan,attrib)));
disp('Hyundai');disp(mean(Cars(hyundai,attrib)));
disp('Mercedes');disp(mean(Cars(mercedes,attrib)));
disp('Convertibles'); disp(mean(Cars(convertibles,attrib)));
disp('Hybrids');disp(mean(Cars(hybrids,attrib)));
```

Average Score Results:

	AirPoll	CmbMPG	Greenhouse	Smartway	Class	Number of cars
Honda	6.1912	26.9289	5.8529	0.4706	3.0147	68
Toyota	5.5000	21.9071	4.1500	0.1917	4.9833	120
Ford	5.6531	21.9077	4.2551	0.2143	4.3367	98
Volkswagen	6.2813	26.6488	5.8281	0.6406	2.1563	64
Chevrolet	5.6393	20.9141	3.8525	0.1967	4.3607	61
Audi	5.4625	21.0050	3.8250	0.1000	2.1750	80
Nissan	5.4018	23.0836	4.2321	0.1786	3.9821	112
Hyundai	5.7083	24.6295	5.3611	0.3194	3.4722	72
Mercedes	5.9314	18.2365	2.6275	0.0294	2.8922	102
Convertibles	5.5517	21.7418	4.2414	0.1379	1.0862	58
Hybrids	6.3488	24.5782	5.0233	0.3488	3.9302	43

Conclusions can easily be found about the eco-friendliness of the different types of vehicles listed above. I chose a few of the best-selling vehicle types and it is easy to do a side-by-side comparison of the attributes. Volkswagen and Honda have by far the best statistical scores and have a large percentage of their models on the 'Smart Way' list. Chevrolet and Audi have poor Greenhouse scores, but Mercedes has the worst.

Many hybrids have a vehicle class 5 (SUV) or vehicle class 2 (midsize), which probably explains the lower MPG. Surprisingly, only around a third of hybrids received the Smartway rating, probably due to their too low Greenhouse average. Also, the majority of Chevrolet cars and Ford cars are larger

vehicles, with classes 5, 7, or 8. It is important to realize that different types of cars are being compared, with perhaps Ford making larger cars than Volkswagen. These results just give a general idea, and plots and a more detailed analysis could be done.

Conclusion

This was a very interesting and large dataset to work with. I could have done a more detailed analysis of the different car models, but instead focused on trying to find characteristics that signify a green vehicle. I discovered that when creating the dataset, the EPA chose to round the Combined miles-per-gallon score to the nearest whole number, where I wouldn't have done so. Also, I found that characteristics that signify an eco-friendly vehicle are:

- Station wagon
- 2 wheel, not 4 wheel drive (Very few eco-friendly cars have 4 wheel drive)
- Battery Electric, then Hydrogen (Mercedes F-Cell), then Diesel, then CNG, then Gasoline
- Transmission: Manual, CVT, or Other
- Volkswagen or Honda

Also, just because a car says 'hybrid' on it, does not mean it is better.

Characteristics that signify a vehicle is NOT eco-friendly

- Van, pick-up, or large
- Gasoline with 4 wheel drive
- Chevrolet, Audi or Mercedes

Overall, I don't think my results are very valuable or useful in choosing a car to buy. I wish the dataset had the MSRP so that I could take price into account. I discovered that having good air pollution is not an indication of a eco-friendly car.

I think that using these results, a more in-depth analysis could be done, focusing more on certain characteristics. In retrospect, I don't think I should have looked at the entire set of 1877 cars the entire time, and could have done a more detailed analysis on just the smaller cars, or just the diesel cars, for example.