

DATA MINING EPA'S GREEN VEHICLE GUIDE:  
PROFILING AND PREDICTION USING  
K-MEANS CLUSTERING AND  
NEURAL NETWORKS

A THESIS  
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF MASTER OF SCIENCE  
IN THE GRADUATE SCHOOL OF THE  
TEXAS WOMAN'S UNIVERSITY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE  
COLLEGE OF ARTS AND SCIENCES

BY  
TERA DAUN SMITH, B.S.

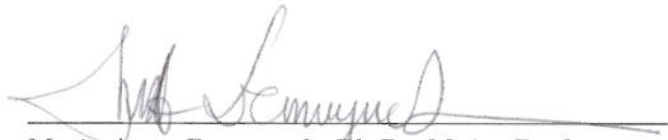
DENTON, TEXAS  
AUGUST 2012

TEXAS WOMAN'S UNIVERSITY  
DENTON, TEXAS

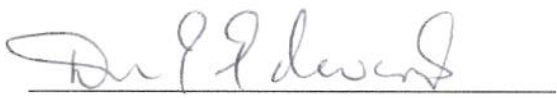
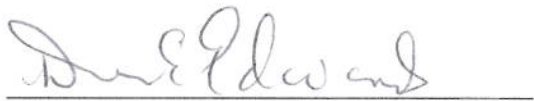
April 20, 2012

To the Dean of the Graduate School:

I am submitting herewith a thesis written by Tera Daun Smith entitled "Data Mining EPA's Green Vehicle Guide: Profiling and Prediction Using k-Means Clustering and Neural Networks." I have examined this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science with a major in Mathematics.

  
Marie-Anne Demuyne, Ph.D., Major Professor

We have read this thesis and recommend its acceptance:

  
Department Chair

Accepted:

  
Dean of the Graduate School

## ACKNOWLEDGEMENTS

I am indebted to the many people who made this thesis conceivable.

I extend my gratitude foremost to my Master's thesis advisor, Dr. Marie-Anne Demuynck. With her enthusiasm, inspiration, and extensive efforts to explain things clearly and simply, she fortified an intense interest in computer science and data mining. Throughout my thesis-writing period, she provided encouragement, sound advice, extensive direction and a plethora of insatiable ideas.

Thanks to the many people who have taught me mathematics and computer science: my undergraduate professors at TWU (especially Dr. Jian Zhang, Dr. Wayne Zimmerman, Dr. Hillary Risser, and Dr. Junalyn Navarra-Madsen), and my graduate professors (especially Dr. David Marshall, Dr. Ellina Grigorieva, Dr. Turner Hogan and Dr. Mark Hamner). For his kind assistance with writing letters, giving wise advice, helping with various applications, and so on, I wish to thank Dr. Don Edwards.

I wish to thank Jennifer Bonder for helping me get through the most difficult times, and for all the emotional support, camaraderie, guidance, and care she provided.

I wish to thank my entire family for providing a loving environment for me; Bill and Perry Smith, Doug and Velma Smith, Delaney Smith and those not mentioned who support, teach, and love unconditionally and to them I dedicate this thesis.

Lastly, but not least, I wish to thank God for the gift of creative intent which made it all possible.

## ABSTRACT

TERA DAUN SMITH

### DATA MINING EPA'S GREEN VEHICLE GUIDE: PROFILING AND PREDICTION USING K-MEANS CLUSTERING AND NEURAL NETWORKS

AUGUST 2012

This thesis is designed to study data mining techniques and explore the predictive value of data from the EPA's Green Vehicle Guide which supplies pertinent information regarding environmental performance for each vehicle sold in the United States from 2000 to 2010. Using IBM® SPSS® Modeler to discover patterns most advantageous to statistical analysis of the data set, each vehicle's various variables and scores in relation to emission, air quality and SmartWay status are modeled using two techniques, *k*-means clustering and artificial neural networks. Predictions based on analysis of this data set are as expected with all models claiming greenhouse gas scores to be the greatest predictor variable for SmartWay status. Therefore, technological focus to improve greenhouse gas scores by reducing emissions is essential if SmartWay status for vehicles and environmental consciousness is a goal.

## TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS .....	iii
ABSTRACT .....	iv
LIST OF TABLES .....	vii
LIST OF FIGURES .....	viii
 Chapter	
I. INTRODUCTION .....	1
Problems in Environmental Research .....	3
Case Study: EPA Green Vehicle Guide Data .....	3
Data Mining Tasks and Proposed Statistical Models .....	4
Summary .....	5
II. DATA PREPROCESSING .....	6
Variables Background .....	6
Scoring .....	9
Raw Data .....	11
Processed Data .....	14
Exploratory Data Analysis .....	15
Summary .....	19
III. FAMILIARIZATION WITH IBM® SPSS® MODELER SOFTWARE .....	20
SmartWay Model Results .....	26
Summary .....	30
IV. <i>K</i> -MEANS CLUSTERING MODEL .....	31
<i>k</i> -Means Clustering Technique .....	31
<i>k</i> -Means Clustering Green Model .....	33
<i>k</i> -Means Clustering Model With Selective Inputs, KMSI Model .....	40
<i>k</i> -Means Summary .....	44

V. ARTIFICIAL NEURAL NETWORK MODEL .....	45
The Neural Network Technique .....	45
Normalizing the Data .....	47
Training the ANN .....	49
ANN Algorithmic Functions .....	50
The Neural Network Model .....	52
<i>k</i> -Means Clustering Output as Input to an ANN .....	57
Summary .....	59
VI. CONCLUSIONS .....	60
REFERENCES .....	62

## LIST OF TABLES

Table	Page
1. Structure Matrix of Discriminate Model for SmartWay Classification .....	27

## LIST OF FIGURES

Figure	Page
1. Vehicle search output .....	8
2. SmartWay elite search output .....	8
3. Air pollution score 2009-2010 .....	10
4. Air pollution score 2008 and earlier .....	10
5. Raw data .....	11
6. Missing values and date errors .....	12
7. Clean data part 1.....	12
8. Clean data part 2 step 1.....	13
9. Clean data part 2 step 2 .....	13
10. Processed data .....	14
11. IBM® SPSS® Statistics 19 correlation matrix scatter plot.....	15
12. IBM® SPSS® Modeler SmartWay proportions .....	16
13. IBM® SPSS® Statistics 19 SmartWay cross tab with cmb mpg .....	17
14. IBM® SPSS® Statistics 19 SmartWay cross tab with aps.....	18
15. IBM® SPSS® Statistics 19 SmartWay cross tab with ggs.....	18
16. IBM® SPSS® Modeler GUI .....	20
17. IBM® SPSS® Modeler sources variable file ep .....	21
18. IBM® SPSS® Modeler opening data file .....	22



19. IBM® SPSS® Modeler field operations auto data prep and type.....	22
20. IBM® SPSS® Modeler field operation: type .....	23
21. IBM® SPSS® Modeler EPA data audit .....	24
22. IBM® SPSS® Modeler SmartWay auto classifier model .....	25
23. IBM® SPSS® Modeler three classifying models .....	26
24. CR Tree .....	28
25. QUEST Tree .....	29
26. IBM® SPSS® Modeler SmartWay prediction .....	30
27. IBM® SPSS® Modeler $k$ -means panel .....	34
28. IBM® SPSS® Modeler $k= 3$ clusters .....	35
29. IBM® SPSS® Modeler $k$ -means pi for initial clusters.....	36
30. IBM® SPSS® Modeler $k$ -means cluster summary for initial 3 clusters .....	37
31. IBM® SPSS® Modeler $k$ -means cluster-1 histogram .....	38
32. IBM® SPSS® Modeler $k$ -means cluster-2 histogram .....	38
33. IBM® SPSS® Modeler $k$ -means cluster-3 histogram .....	39
34. IBM® SPSS® Modeler $k= 3$ clusters with si .....	40
35. IBM® SPSS® Modeler $k$ -means predictor importance for si three clusters .....	41
36. IBM® SPSS® Modeler $k$ -means cluster summary for selected variables .....	42
37. IBM® SPSS® Modeler $k$ -means ggs histogram cluster-1 .....	43
38. IBM® SPSS® Modeler $k$ -means ggs histogram cluster-2 .....	43
39. IBM® SPSS® Modeler $k$ -means ggs histogram cluster-3 .....	44

40. ANN with 3 layers .....	45
41. Biological neuron .....	46
42. Greenhouse gas score for min-max .....	47
43. Sigmoid function graph .....	51
44. IBM® SPSS® Modeler neural net node .....	52
45. IBM® SPSS® Modeler neural net fields .....	53
46. IBM® SPSS® Modeler neural network model summary .....	54
47. IBM® SPSS® Modeler neural network predictor importance .....	55
48. IBM® SPSS® Modeler neural network using all variables.....	56
49. IBM® SPSS® Modeler neural network $k$ -means input model summary .....	57
50. IBM® SPSS® Modeler neural network $k$ -means input predictor importance ..	58
51. IBM® SPSS® Modeler neural network $k$ -means input classification summary	58
52. IBM® SPSS® Modeler neural network $k$ -means input model .....	59

## CHAPTER I

### INTRODUCTION

Data mining is the process of discovering interesting information or patterns in data, formerly unidentified or hidden, and discovering the relationships between them. The process can find potentially useful predictor variables and trend patterns when large, observational data sets or repositories, such as databases and data warehouses, are mined (Larose, 2005).

The Cross-Industry Standard Process for Data Mining, or CRISP-DM, sets the standard process for business or research ventures to incorporate data mining into their overall problem-solving approach. This is a six phase iterative and adaptive process where each phase builds from one phase to the next and also revisits phases as necessary for reassessment and modification. The six phases are Business/Research Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment (Larose, 2005, p.5-6).

The first phase of the process is the Business or Research Understanding phase where objectives, limitations and a preliminary solution strategy are defined. Questions are asked to aid in defining a research objective. What processes will the data undergo? What data mining tasks will be best suited for analysis? Where do statistics and data mining overlap or more specifically, how is statistics used in data mining and vice versa? The Environmental Protection Agency (EPA) *Green Vehicle Guide for Consumers* provides the data set under investigation for this study which is available to the public at <http://www.epa.gov/greenvehicles>.

Understanding the data is the primary task of the second phase of the CRISP-DM process. Data is collected and explored to familiarize oneself with the data set, its quality and any patterns noticeable on the surface. Overall, the data set in this study is a large, complete, high quality set with limited missing values and several variables to consider for analysis.

Some research questions of particular interest regarding this study are as follows: What are the implications of the EPA scores data set and what pattern subsets can be observed in the study?

To use data mining tools to find specific patterns and trends, the data set must be cleaned by deleting duplicate records, handling missing values, and integrating similar variables of records. This process is a part of the Data Preparation phase, the third phase, and can be the most time consuming. Exploratory Data Analysis (EDA) of the variables in this study will be analyzed using IBM® SPSS® Statistics 19. Once the data has undergone this process, different data mining techniques such as association rules, classification, clustering and segmentation can be applied. During this phase, Microsoft®Excel will be used to clean and prepare the data for the next phase in the process, the modeling phase. The modeling phase incorporates the selection and use of one or more modeling techniques for optimized conclusions. Previous phases may be revisited to prepare data as needed for use with a particular technique.

For modeling data, IBM® SPSS® Modeler will be used to discover patterns most advantageous to statistical analysis of the data set. Modeler is a graphical interface that aids in visualizing the data-mining process, accesses predictive capabilities using IBM® SPSS® Statistics data transformation, conducts hypothesis testing and reports capabilities on a single interface. Once transformation and predictive variables are identified, models can be built and employed. In our data set, each vehicle has various variables and scores in relation to emission and air quality. A SmartWay status is awarded to those considered to be clean, green vehicles. In this paper, these data will be mined and modeled for the data-mining problem previously defined in the first phase.

The final stages of processes are the Evaluation and Deployment phases. Effective models are assessed for potential application and in that case the models will be used and simple deployments made such as generating reports and graphs with more complex statistical analysis and prediction. Ineffective models will be reassessed and/or eliminated.

### **Problems in Environmental Research**

A common problem in data mining in general and more specifically in the field of environmental research is the problem of missing values. Many environmental data sets date back to a pre-technological era where a complete data set will have variables from post-technology combined with pre-technological data. These types of data sets contain values and new variables collected more recently integrated with dated collections resulting in missing values where data for variables were not collected or recorded due to lack of technology. Also, human or computer errors in data entry and collection are important factors limiting accurate and effective data analysis and mining pattern deficiencies. Other problems include structural organization of variables within the data set. In particular, the EPA data set under study contains variables on a single line that should be separated into two lines of usable information and scores. These problems will be addressed in the most efficient way by cleaning and preparing the data set in Microsoft® Excel. Variable values not addressed manually in Excel will be taken care of internally with the modeling software using various algorithms and transformations explained in detail by way of a case study of the process.

### **Case Study: EPA Green Vehicle Guide Data**

The proposed data set comes from the EPA's *Green Vehicle Guide* which supplies pertinent information regarding environmental performance for each vehicle sold in the United States from 2000 to 2010. The guide itself was created as a database for users to customize a search to access and compare up to three vehicles at a time based on fuel economy and emission scores and can be found on the web at <http://www.epa.gov/greenvehicles/Index.do>.

Variables include but are not limited to various makes and models and their specific ratings with regards to standardized emissions in relation to air pollution, greenhouse gas, fuel economy and SmartWay scores.

Comparisons of air pollution and greenhouse gases use scores ranging from 0-10 for emission ratings. Major pollutants in vehicle exhaust are monitored by emission standards set up by the EPA. These include types of hydrocarbon and other carbon-containing com-

pounds such as non-methane organic gases (NMOG), non-methane hydrocarbons (NMHC), and total hydrocarbons (THC). Nitrogen oxides ( $\text{NO}_x$ ) combine with hydrocarbons to form smog and particulate matter (PM), which are tiny particles of solid matter that inbed in the lungs, are also measured. Further air contaminants are carbon monoxide (CO) and formaldehyde ( $\text{CH}_2\text{O}$ ), a carcinogenic lung irritant . A vehicle with a score of 10 emits none of the pollutants and is considered the cleanest.

The greenhouse gas score refers to  $\text{CO}_2$  emissions. Vehicles with greater fuel economy, using less gas to travel the same distance than other vehicles, have a higher greenhouse gas score. The fuel economy (mpg) rating variable combines city and highway gas milieage to a single score by regression analysis. The combined score, or average of city and highway gas mileages, references an approximate relation between the fuel economy ratings and greenhouse gas scores. Again, a higher score is cleaner.

The SmartWay variable specifies the vehicles with the best environmental performance measured against other vehicles. The EPA verifies that these vehicles have exceeded environmental thresholds on air pollution and greenhouse gas scores.

### **Data Mining Tasks and Proposed Statistical Models**

Data mining tasks applied in this study are clustering and classification, using the  $k$ -means clustering method, and estimation and prediction using a neural network model. Clustering methods attempt to group records based on similarities in the variables. The  $k$ -means method is a relatively fast method though it does not yield optimal results. In this study, using the results of the  $k$ -means clustering method discussed in detail in Chapter IV, variables from the *Green Vehicle Guide* will be classified by similarity to one another and clustered as a preliminary step, then those results are used as inputs into a neural network. This speeds the network modeling process by limiting the input variables by reducing the number of variables to be smaller than the original variables.

Neural networks are modeled after animal brains' feed forward network of layered levels of neurons or nodes. At a minimum, a neural network consists of an input, output and one or more hidden layers. The nodes in each layer connect to all the nodes of the next layer

by edges. Each edge initially receives a randomly assigned weighted value. Combination functions, which are frequently summations, and activation functions, most commonly sigmoid functions, serve as nonlinear processes similar to the biological neurons in our own brains, signaling and firing other neurons. Outputs from one layer become inputs to the next layer in the network. This permits the model to have a more flexible range of tasks including classification, estimation and prediction. This paper will include two neural network models, one with inputs from the original data set and one with inputs obtained from the  $k$ -means clustering model.

### Summary

In conclusion, in the research understanding phase of this project, applications of knowledge discovery patterns from mining the EPA *Green Vehicle Guide* data are used to create classification and predictive models by means of  $k$ -means clustering and neural network modeling. The process will be mostly hidden due to the proprietary software, however, mathematical analysis will be applied at each step from the modeling phase to the deployment phase. Definitive assumptions may or may not be determined but the general functions processes will be shown. The following chapters will include an in depth analysis of the the Data Preparation Phase complete with an overview of the IBM® SPSS® Modeler software, and creating models using the  $k$ -means clustering technique and artificial neural networks.

## CHAPTER II

### DATA PREPROCESSING

Data preprocessing is a preliminary stage in data mining and is essential to developing sound models. Poor or insufficient data can skew outcomes and yield misleading results, if any at all. Many problems observed in raw data sets are mistakes in data entry, ambiguous and/or contradictory coding of variables, and missing values. A common rule of thumb when planning a modeling strategy is “garbage in, garbage out,” meaning if the data fed into a model is skewed then the output will be as well. Therefore, this preprocessing phase is not only important but necessary if the final output is to be usable. The data set must be as clean, organized and accurate as possible before applying any data mining tasks and modeling techniques.

#### **Variables Background**

The data collected from the EPA was originally downloaded as a text file, and imported to Excel with tab delimited settings and a general data format. The first task was to become familiar with the data set. This meant looking over the variables and further researching each according to the EPA at <http://www.epa.gov/greenvehicles/Aboutratings.do>.

The variables in the data set are as follows, including a few mentioned previously in the first chapter:

- Model: vehicle makes and models from 2000-2010 and sold in the United States
- Displ: engine displacement by liters of air
- Cyl: cylinder values ranging from 4-12 cylinders
- Trans: transmission type (manual, automatic or semi-automatic and number of gears)
- Drive: drivetrain type (2WD or 4WD)
- Fuel: fuel type (gasoline or ethanol)



- Sales Area: CA (California) or FA (any other state in the U.S.)
- Stnd: emissions standard levels defined by the EPA
- Stnd Description: emissions standard description defined by the EPA
- Underhood ID: vehicle model number
- Veh Class: vehicle class (SUV, pickup, small car, etc.)
- Air Pollution Score: calculated by EPA
- City MPG: fuel mileage measured by driving in the city
- Hwy MPG: fuel mileage measured by driving on the highway
- Cmb MPG: average MPG of city and hwy per model
- Greenhouse Gas Score: calculated by EPA
- SmartWay: yes or no status

These variables contain information and measurements representing the causes and effects of greenhouse emissions and pollution as the result of burning fossil fuels by individual transportation (not including busses, vans, commercial delivery, etc.). Providing a database for consumers to search for fuel efficient and clean vehicles is the main purpose for the EPA's formulation of this data set. Via EPA's *Green Vehicle Guide* online at <http://www.epa.gov/greenvehicles/Index.do>, searching for vehicles in general, vehicles by type or the vehicles by state is simple and fast and gives values retrieved from the same data set analyzed in this study.

For example, in searching for a vehicle in general, the year, state, make and model are selected. The results for choosing the year 2009, Texas, Ford, and Explorer are shown in Figure 1. Four types of Ford Explorers are listed with ranges of air pollution, fuel economy

and greenhouse score differences. None of the vehicles are SmartWay, because of low greenhouse scores and poor fuel economy. The best vehicle to choose from the list, however, would be Explorer with a 4.6L, 6 cyl. engine and 2WD since out of the four listed, this model has the best fuel economy value.

Home Basic Search Advanced Search Compare Complete Guide									
Basic Search Criteria									
Year: 2009									
State: Texas									
Make: FORD									
Model: Explorer									
Click ▲ arrow to sort									
Definitions   New Search   Advanced Search									
Model	Engine	Fuel	Trans	Drive	Air Pollution Score 10 = Best	Fuel Economy (mpg) City / Hwy	Greenhouse Gas Score 10 = Best	SmartWay	Compare Choose up to 5
<a href="#">FORD Explorer</a>	4L 6 cyl	Gasoline	Auto 5 speed	2WD		14 / 20			<input type="checkbox"/>
<a href="#">FORD Explorer</a>	4L 6 cyl	Gasoline	Auto 5 speed	4WD		13 / 19			<input type="checkbox"/>
<a href="#">FORD Explorer</a>	4.6L 8 cyl	Gasoline	Auto 6 speed	2WD		15 / 21			<input type="checkbox"/>
<a href="#">FORD Explorer</a>	4.6L 8 cyl	Gasoline	Auto 6 speed	4WD		14 / 19			<input type="checkbox"/>
Definitions   New Search   Advanced Search									
									Compare

Figure 1. Vehicle search output

However, if you wanted to search for the greenest vehicle by state, the database allows selection of the SmartWay status, an indicator of good environmental performance. In particular, SmartWay Elite is selected to narrow the search and displays top performing vehicles in Texas in 2009 and shown in Figure 2.

Home Basic Search Advanced Search Compare Complete Guide									
Basic Search Criteria									
Year: 2009									
State: Texas									
SmartWay Types: Smartway Elite Vehicles Chosen									
Click ▲ arrow to sort									
Definitions   New Search   Advanced Search									
Model	Engine	Fuel	Trans	Drive	Air Pollution Score 10 = Best	Fuel Economy (mpg) City / Hwy	Greenhouse Gas Score 10 = Best	SmartWay	Compare Choose up to 5
<a href="#">HONDA Civic</a>	1.8L 4 cyl	CNG	Auto 5 speed	2WD		24 / 36			<input type="checkbox"/>
<a href="#">HONDA Civic Hybrid</a>	1.3L 4 cyl	Gasoline	Auto Variable	2WD		40 / 45			<input type="checkbox"/>
Definitions   New Search   Advanced Search									
									Compare

Figure 2. SmartWay elite search output

Only two vehicles were top performers in 2009, the Honda Civic, and the Honda Civic Hybrid, each with nearly perfect scores on gas mileage and greenhouse gas scores.

### Scoring

Air pollution and greenhouse gas scores are based on fuel economy and emission levels of each vehicle and range from 0 to 10. Fuel economy uses a combined highway and city gas mileage calculated as:

$$MPG = \frac{1}{\frac{0.55}{CityMPG} + \frac{0.45}{HwyMPG}}$$

SmartWay status is determined by the best performance on both air pollution and greenhouse scores and is limited only to vehicles from the same year and scoring basis. These scoring thresholds are available to the public at the EPA's *Green Vehicle Guide* website, <http://www.epa.gov/greenvehicles/Aboutratings.do>.

Several resources are available to the public. Figure 3 represents air pollution scores from emissions standards for 2009 and 2010. All the vehicles in the guide meet standard emission requirements for the EPA. Air pollution scores and vehicle identification status are shown in Figure 3.

If a vehicle's emission standard in the United States (excluding California) is identified as Bin 1 then an air pollution score of 10 is assigned; such vehicles are therefore on their way to gaining smartway status.

In California, this same vehicle with a ZEV emissions standard description would receive a score of 10. Of importance is that in 2009, Bin's 9,10 and 11 were phased out representing a possible permanent reduction of low emission standards. With the data set ranging from 2000-2009, however, some vehicle models from previous years fit into the 2008 and earlier standards as shown Figure 4.

<b>US EPA - Vehicle Environmental Scoring</b> <b>Air Pollution Score</b> <b>MY 2009 &amp; MY 2010</b>		
Score	US EPA Tier 2 Emissions Standard	California Air Resources Board LEV II Emissions Standard
10	Bin 1	ZEV
9	Bin 2	SULEV II
8	Bin 3	xx
7	Bin 4	ULEV II
6	Bin 5	LEV II
5	Bin 6	LEV II option 1
4	Bin 7	xx
3	Bin 8	SULEV II lg trucks
2	xx	ULEV II lg trucks
1	xx	LEV II lg trucks

\* Bin 9, 10, 11 phased out in MY 2009

Source: <http://www.epa.gov/greenvehicles/Aboutratings.do>

Figure 3. Air pollution score 2009-2010

<b>US EPA - Vehicle Environmental Scoring</b> <b>Air Pollution Score</b> <b>MY 2008 &amp; Earlier</b>		
Score	US EPA Tier 2 Emissions Standard	California Air Resources Board LEV II Emissions Standard
10	Bin 1	ZEV
9	Bin 2	SULEV II
8	Bin 3	xx
7	Bin 4	ULEV II
6	Bin 5	LEV II
5	Bin 6	LEV II option 1
4	Bin 7	xx
3	Bin 8	SULEV II lg trucks
2	Bin 9	ULEV II lg trucks
1	Bin 10	LEV II lg trucks
0	Bin 11	xx

Source: <http://www.epa.gov/greenvehicles/Aboutratings.do>

Figure 4. Air pollution score 2008 and earlier

## Raw Data

After reviewing the different aspects of the database including definitions and purposes, the raw data were assessed. After importing the raw data into Excel, many problems were apparent immediately. Figure 5 shows a portion of the raw data set with both formatting and data entry errors.

Gasoline	FA	B5	Federal Tier 2 Bin AGMXV02.4040	midsize car	6	26	34	29	8
Gasoline	CA	L2	California LEV-II   AGMXV02.4040	midsize car	6	26	34	29	8
Ethanol/Gas	FC	B4	Federal Tier 2 Bin AGMXT05.3381	pickup	7-Jul	14-Oct	14/19	16-Dec	3-Mar
Ethanol/Gas	FA	B5	Federal Tier 2 Bin AGMXT05.3373	pickup	6-Jun	14-Oct	14/19	16-Dec	3-Mar
Ethanol/Gas	FA	B5	Federal Tier 2 Bin AGMXT05.3373	pickup	6-Jun	13-Oct	13/18	15-Nov	2-Feb
Ethanol/Gas	FA	B5	Federal Tier 2 Bin AGMXT05.3373	pickup	6-Jun	15-Nov	16/21	13/17	3-Apr
Ethanol/Gas	FA	B5	Federal Tier 2 Bin AGMXT05.3373	pickup	6-Jun	15-Nov	16/21	13/17	3-Apr
Ethanol/Gas	FA	B5	Federal Tier 2 Bin AGMXT06.2375	pickup	6-Jun	13-Oct	14/19	15-Nov	2-Feb
Ethanol/Gas	FA	B5	Federal Tier 2 Bin AGMXT06.2375	pickup	6-Jun	12-Sep	13/19	14-Nov	1-Feb
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT04.3186	pickup	6	15	20	17	3
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT04.3186	pickup	6	14	18	15	2
Gasoline	CA	L2	California LEV-II   AGMXT04.3186	pickup	6	15	20	17	3
Gasoline	CA	L2	California LEV-II   AGMXT04.3186	pickup	6	14	18	15	2
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT06.0371	pickup	6	21	22	22	6
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT06.0371	pickup	6	21	22	21	5
Ethanol/Gas	FC	B4	Federal Tier 2 Bin AGMXT05.3381	pickup	7-Jul	15-Nov	16/22	13/18	4-Apr
Ethanol/Gas	FA	B5	Federal Tier 2 Bin AGMXT05.3373	pickup	6-Jun	15-Nov	16/22	13/18	4-Apr
Gasoline	FA	HDV1	Federal HDV (85C AGMXD06.0396	pickup	2	N/A	N/A	N/A	N/A
Gasoline	FA	HDV1	Federal HDV (85C AGMXD06.0396	pickup	2	N/A	N/A	N/A	N/A
Gasoline	CA	L2	California LEV-II   AGMXD06.0396	pickup	2	N/A	N/A	N/A	N/A
Gasoline	CA	L2	California LEV-II   AGMXD06.0396	pickup	2	N/A	N/A	N/A	N/A
Gasoline	FA	HDV1	Federal HDV (85C AGMXD06.0396	pickup	2	N/A	N/A	N/A	N/A
Gasoline	FA	HDV1	Federal HDV (85C AGMXD06.0396	pickup	2	N/A	N/A	N/A	N/A
Gasoline	CA	L2	California LEV-II   AGMXD06.0396	pickup	2	N/A	N/A	N/A	N/A
Gasoline	CA	L2	California LEV-II   AGMXD06.0396	pickup	2	N/A	N/A	N/A	N/A
Ethanol/Gas	FA	B5	Federal Tier 2 Bin AGMXT05.3373	SUV	6-Jun	15-Nov	16/21	13/17	3-Apr
Ethanol/Gas	FA	B5	Federal Tier 2 Bin AGMXT06.2375	SUV	6-Jun	12-Sep	14/19	14-Nov	1-Feb
Gasoline	FC	B5	Federal Tier 2 Bin AGMXT06.0395	SUV	6	N/A	N/A	N/A	N/A
Gasoline	FC	B5	Federal Tier 2 Bin AGMXT06.0395	SUV	6	N/A	N/A	N/A	N/A
Ethanol/Gas	FA	B5	Federal Tier 2 Bin AGMXT05.3373	SUV	6-Jun	15-Nov	16/21	13/17	3-Apr
Ethanol/Gas	FA	B5	Federal Tier 2 Bin AGMXT05.3373	SUV	6-Jun	15-Nov	16/21	13/17	3-Apr
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT06.0371	SUV	6	21	22	22	6
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT06.0371	SUV	6	21	22	21	5
Gasoline	CA	U2	California LEV-II   AGMXJ03.6151	SUV	7	17	24	19	4
Gasoline	CA	U2	California LEV-II   AGMXJ03.6151	SUV	7	16	23	19	4
Gasoline	FA	B5	Federal Tier 2 Bin AGMXJ03.6151	SUV	6	17	24	19	4
Gasoline	FA	B5	Federal Tier 2 Bin AGMXJ03.6151	SUV	6	16	23	19	4
Gasoline	FC	B4	Federal Tier 2 Bin ACRXV05.7UP0	large car	7	16	23	18	4
Gasoline	FA	B5	Federal Tier 2 Bin ACRXV03.SRN0	large car	6	17	23	19	4

Figure 5. Raw data

The most noticeable errors are missing values, which are labeled N/A. Another readily apparent problem is the vehicles that are the same in all variable values, except differences in fuel type. These differ in that one make and model will have two fuel types, one being ethanol and the other gasoline. Each vehicle's air pollution, various mpg and greenhouse scores have individual scores for each type respectively. However, if the number value is less than 12, they are being interpreted as dates instead of numbers. (See Figure 6).

2	N/A	N/A	N/A	N/A	no
6-Jun	15-Nov	16/21	13/17	3-Apr	no
6-Jun	12-Sep	14/19	14-Nov	1-Feb	no
6	N/A	N/A	N/A	N/A	no
6	N/A	N/A	N/A	N/A	no
6-Jun	15-Nov	16/21	13/17	3-Apr	no
6-Jun	15-Nov	16/21	13/17	3-Apr	no

Figure 6. Missing values and date errors

Therefore, the first task implemented involved conversion from dates to general number values. Where dates were listed instead of numerical scores, they were formatted to date such that 6/6 would replace the original 6-June. The second task was to delete N/A values, leaving a blank cell because the modeling software used to perform the datamining tasks in this study cannot handle missing values as N/A. The results of this process can be seen in Figure 7.

6/6	11/15	16/21	13/17	4/3	no
6/6	9/12	14/19	11/14	2/1	no
1/7					no
1/7					no
6/6	11/15	16/21	13/17	4/3	no
6/6	11/15	16/21	13/17	4/3	no

Figure 7. Clean data part 1

The next task was to separate the vehicles with combined Ethanol/Gasoline fuel types and their respective data. This step was the most time consuming but simple.

For every vehicle that had both an ethanol and a gasoline fuel choice, a new row would be inserted as shown in Figure 8.

Ethanol/Gas	FC	B4	Federal Tier 2 EAGMXT05.3381pickup	7/7	10/14	14/19	12/16	3/3
Ethanol/Gas	FA	B5	Federal Tier 2 EAGMXT05.3373pickup	6/6	10/14	14/19	12/16	3/3
Ethanol/Gas	FA	B5	Federal Tier 2 EAGMXT05.3373pickup	6/6	10/13	13/18	11/15	2/2
Ethanol/Gas	FA	B5	Federal Tier 2 EAGMXT05.3373pickup	6/6	11/15	16/21	13/17	4/3

Figure 8. Clean data part 2 step 1

After all vehicles with these combined fuel types were separated with a new line, the data was copied and pasted leaving ethanol on the top line due to correlation with the first mpg values and emission scores on the line. The first number of each score stays on the top line and the second number of each score was copied to the second data line so that each score column contains a single fuel type and associated score. Once the numbers were separated per line they were still formatted in date format so each cell manipulated had to be reformatted back to number with zero decimal places to fit with the other data. The result is shown in Figure 9.

Ethanol	FC	B4	Federal Tier 2 EAGMXT05.3381pickup	7	10	14	16	3
Gasoline	FC	B4	Federal Tier 2 EAGMXT05.3381pickup	7	14	19	12	3
Ethanol	FA	B5	Federal Tier 2 EAGMXT05.3373pickup	6	10	14	16	3
Gasoline	FA	B5	Federal Tier 2 EAGMXT05.3373pickup	6	14	19	12	3

Figure 9. Clean data part 2 step 2

## Processed Data

After all the processing was complete, the final cleaned data set was prepared for modeling. Figure 10 is a sample of the processed data and as shown. Fuel types have been separated and listed individually, scores have been entered as appropriate values instead of dates and all missing values are left blank.

Gasoline	FA	B5	Federal Tier 2 Bin AGMXT06.2375	pickup	6	13	19	15	2
Ethanol	FA	B5	Federal Tier 2 Bin AGMXT06.2375	pickup	6	9	13	11	2
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT06.2375	pickup	6	12	19	14	1
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT04.3186	pickup	6	15	20	17	3
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT04.3186	pickup	6	14	18	15	2
Gasoline	CA	L2	California LEV-II L AGMXT04.3186	pickup	6	15	20	17	3
Gasoline	CA	L2	California LEV-II L AGMXT04.3186	pickup	6	14	18	15	2
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT06.0371	pickup	6	21	22	22	6
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT06.0371	pickup	6	21	22	21	5
Ethanol	FC	B4	Federal Tier 2 Bin AGMXT05.3381	pickup	7	11	16	13	4
Gasoline	FC	B4	Federal Tier 2 Bin AGMXT05.3381	pickup	7	15	22	18	4
Ethanol	FA	B5	Federal Tier 2 Bin AGMXT05.3373	pickup	6	11	16	13	4
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT05.3373	pickup	6	15	22	18	4
Gasoline	FA	HDV1	Federal HDV (850 AGMXD06.0396	pickup	2				
Gasoline	FA	HDV1	Federal HDV (850 AGMXD06.0396	pickup	2				
Gasoline	CA	L2	California LEV-II L AGMXD06.0396	pickup	2				
Gasoline	CA	L2	California LEV-II L AGMXD06.0396	pickup	2				
Gasoline	FA	HDV1	Federal HDV (850 AGMXD06.0396	pickup	2				
Gasoline	FA	HDV1	Federal HDV (850 AGMXD06.0396	pickup	2				
Gasoline	CA	L2	California LEV-II L AGMXD06.0396	pickup	2				
Gasoline	CA	L2	California LEV-II L AGMXD06.0396	pickup	2				
Ethanol	FA	B5	Federal Tier 2 Bin AGMXT05.3373	SUV	6	11	16	13	4
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT05.3373	SUV	6	15	21	17	3
Ethanol	FA	B5	Federal Tier 2 Bin AGMXT06.2375	SUV	6	9	14	11	2
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT06.2375	SUV	6	12	19	14	1
Gasoline	FC	B5	Federal Tier 2 Bin AGMXT06.0395	SUV	6				
Gasoline	FC	B5	Federal Tier 2 Bin AGMXT06.0395	SUV	6				
Ethanol	FA	B5	Federal Tier 2 Bin AGMXT05.3373	SUV	6	11	16	13	4
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT05.3373	SUV	6	15	21	17	3
Ethanol	FA	B5	Federal Tier 2 Bin AGMXT05.3373	SUV	6	11	16	13	4
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT05.3373	SUV	6	15	21	17	3
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT06.0371	SUV	6	21	22	22	6
Gasoline	FA	B5	Federal Tier 2 Bin AGMXT06.0371	SUV	6	21	22	21	5
Gasoline	CA	U2	California LEV-II L AGMXJ03.6151	SUV	7	17	24	19	4
Gasoline	CA	U2	California LEV-II L AGMXJ03.6151	SUV	7	16	23	19	4
Gasoline	FA	B5	Federal Tier 2 Bin AGMXJ03.6151	SUV	6	17	24	19	4
Gasoline	FA	B5	Federal Tier 2 Bin AGMXJ03.6151	SUV	6	16	23	19	4
Gasoline	FC	B4	Federal Tier 2 Bin ACRXV05.7UP0	large car	7	16	23	18	4
Gasoline	FA	B5	Federal Tier 2 Bin ACRXV03.5RNO	large car	6	17	23	19	4

Figure 10. Processed data



## Exploratory Data Analysis

Since there is not any clear hypothesis regarding the relationships between the variables in the data set, exploratory data analysis (EDA) was performed using IBM® SPSS® Statistics 19 and IBM® SPSS® Modeler softwares. Several interesting correlations and interrelationships were discovered and a few target variables for modeling appeared as well.

First, correlations between the numerical variables, particularly air pollution and greenhouse gas scores, as well as city, highway and combined gas mileages are compared in a matrix scatter plot. Figure 11 shows this scatter plot for these variables, and strong correlations exist between greenhouse gas scores and the various types of gas mileages. The air pollution scores have no correlation and therefore can be modeled with the others. Some caution may need to be taken when modeling with gas mileages and greenhouse scores. Particularly, city and highway gas mileages could be eliminated because of the combined mpg variable.

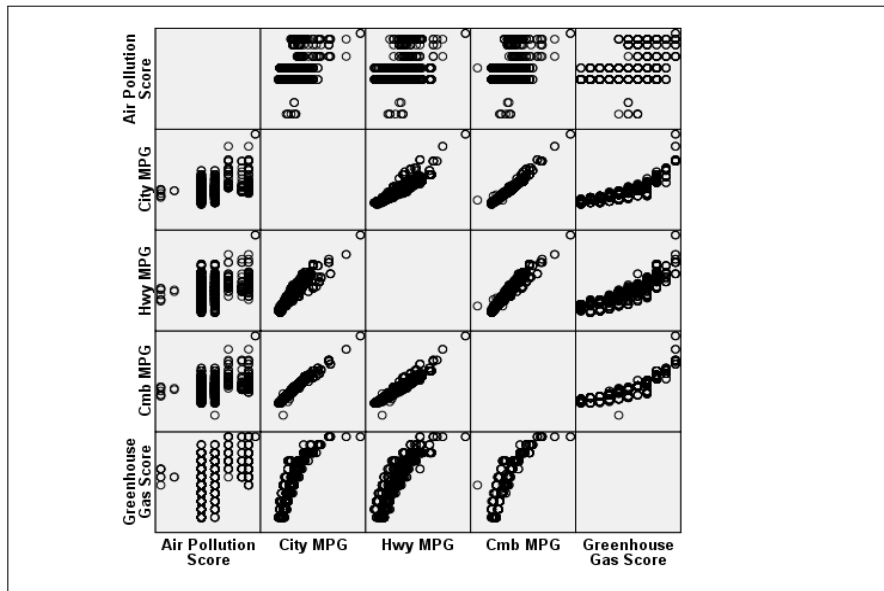


Figure 11. IBM® SPSS® Statistics 19 correlation matrix scatter plot

Another variable to explore is SmartWay status proportions, which is a categorical variable. Each vehicle either has or doesn't have a SmartWay status. A comparison of SmartWay proportions are observed and shown in figure 12.

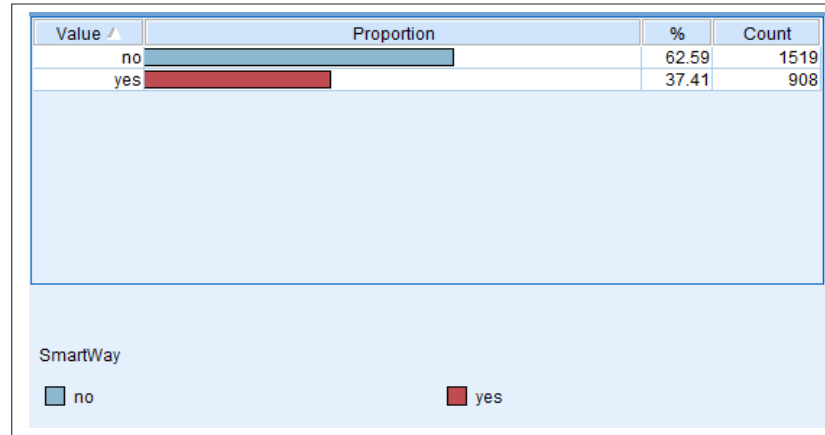


Figure 12. IBM® SPSS® Modeler SmartWay proportions

These proportions are quite different with 37.4% of vehicles having a SmartWay status and 62.5% of vehicles not qualifying. Which variables have a relationship with or influence SmartWay status?

A quick cross tabulation analysis, using IBM® SPSS® Statistics 19, compared SmartWay status with gas mileage, fuel type, air pollution score and greenhouse emission scores and yielded the following results. Figure 13 illustrates the relationship between SmartWay status and combined fuel economy scores.

First of all, as one would expect, higher combined gas mileages had smartway status beginning at 22 mpg, while all the disqualified SmartWay status vehicles had 21 mpg and lower scores. However, a very small number of lower gas mileage vehicles did make SmartWay status, so gas mileage alone is definitely not a predictor variable, although highly influential.

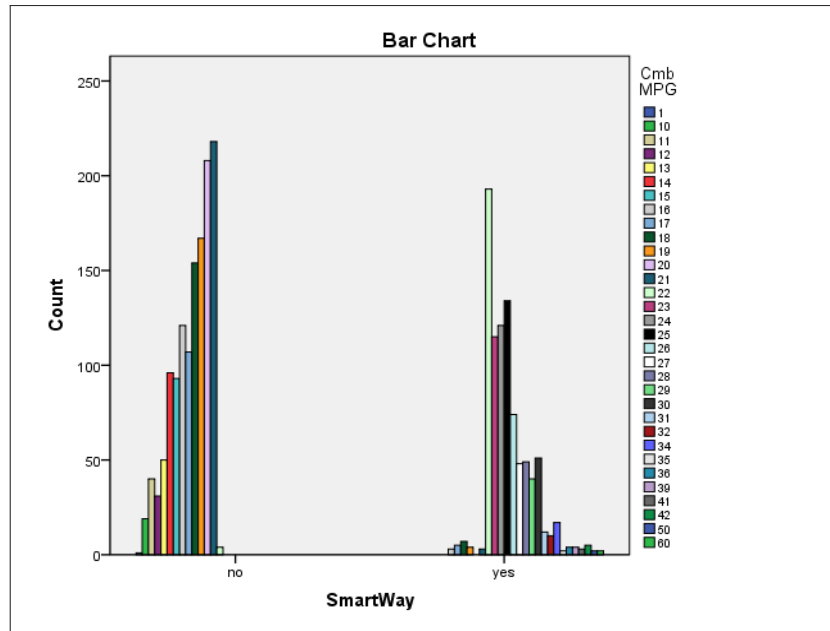


Figure 13. IBM® SPSS® Statistics 19 SmartWay cross tab with cmb mpg

Figure 14 shows air pollution scores compared with SmartWay status. The most common air pollution score was 6.0 for vehicles with and without SmartWay status. Vehicles without SmartWay specification ranged in scores from 1.0-9.5. Those with SmartWay status ranged from 6.0-9.5 with no scores below 6.0. Air pollution scores have an apparent correlation with SmartWay status as lower scores are not considered.

Greenhouse gas scores seem to carry the most weight with regard to gaining SmartWay status, with a more definite split between higher and lower scores dispursed according to SmartWay status as shown in Figure 15. SmartWay vehicles had greenhouse gas scores ranging from approximately 6.0-10 and those without SmartWay status scored in the 0-5.0 range.

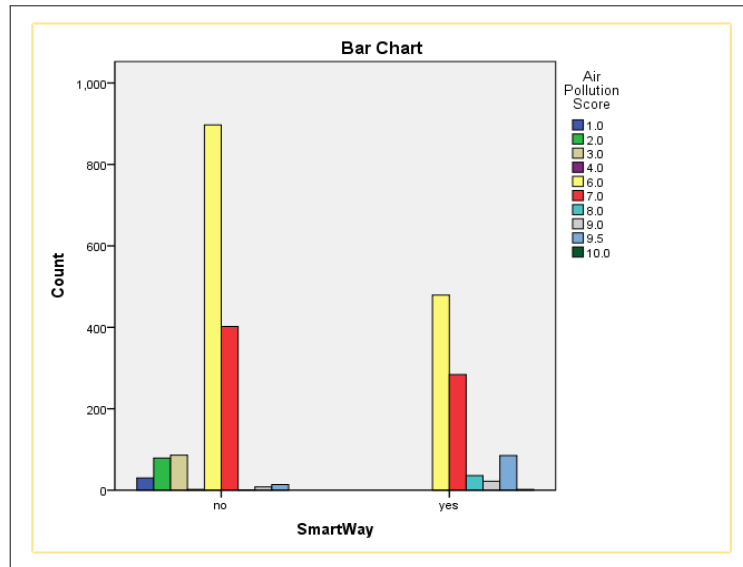


Figure 14. IBM® SPSS® Statistics 19 SmartWay cross tab with aps

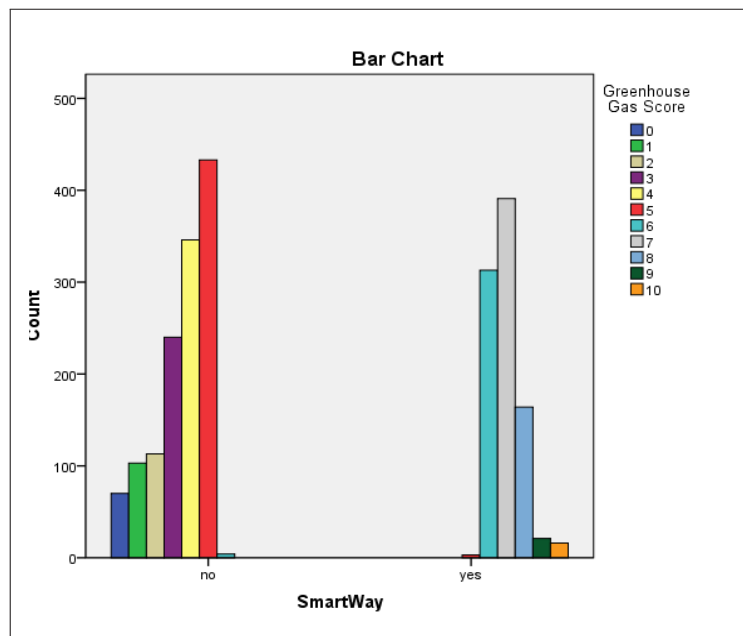


Figure 15. IBM® SPSS® Statistics 19 SmartWay cross tab with ggs

## Summary

In summary, data understanding and preparation phases were performed in this chapter. At this point, the data is much more familiar, of high quality and a potentially useful source for creating a model. Preparations were made by cleaning the raw data set so that it may yield a more effective output. Some variables including fuel, combined mpg, SmartWay, air pollution and greenhouse gas scores may be particularly useful in the next modeling phase using *k*-means clustering and neural network techniques.

### CHAPTER III

#### FAMILIARIZATION WITH IBM® SPSS® MODELER SOFTWARE

Before beginning the modeling process, data preparation had to be completed. The data was saved from Excel as a text file to be brought into IBM® SPSS® Modeler software. Modeler is a graphical user interface(GUI) which allows the user to drag and drop various nodes to stream together in an interactive model. The GUI is depicted in Figure 16.

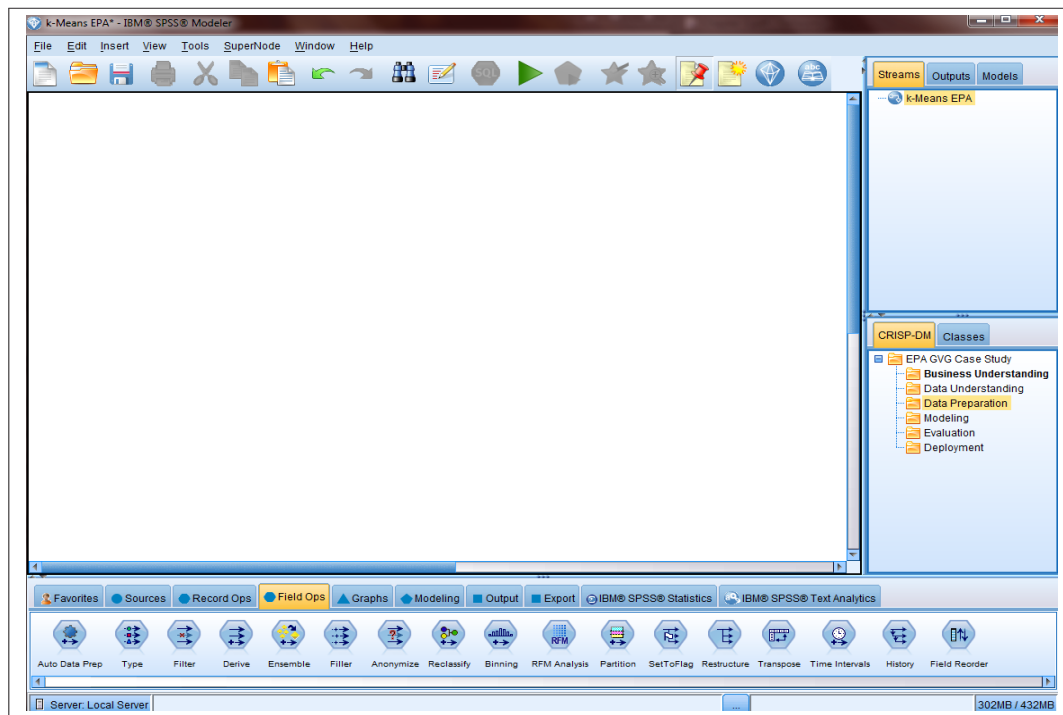


Figure 16. IBM® SPSS® Modeler GUI

The first node necessary for the model imports the data into the software. As shown in Figure 17, the “Sources” tab in the program allows a “Var. File” to be selected and dragged to the workspace.

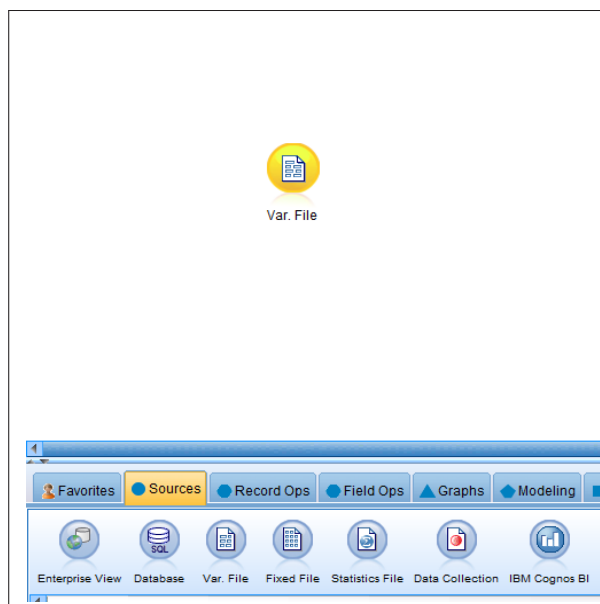


Figure 17. IBM<sup>®</sup> SPSS<sup>®</sup> Modeler sources variable file ep

Double clicking on the node will open it up and allow browsing for the desired file. File extensions include text, SPSS<sup>®</sup> and SAS<sup>®</sup>. Once the file is selected, the text can be previewed and modified for legibility, such as selecting whether columns are tab delimited, comma or space etc. (See Figure 18). When using files other than those accepted by the program, it’s best to save the raw data in a text file for import. When importing, most following operations in the general process will follow this simple drag and drop procedure on the interface and the software will inform of errors or problems that may occur when importing a file. Theses are usually as simple as selecting the correct delimiter.

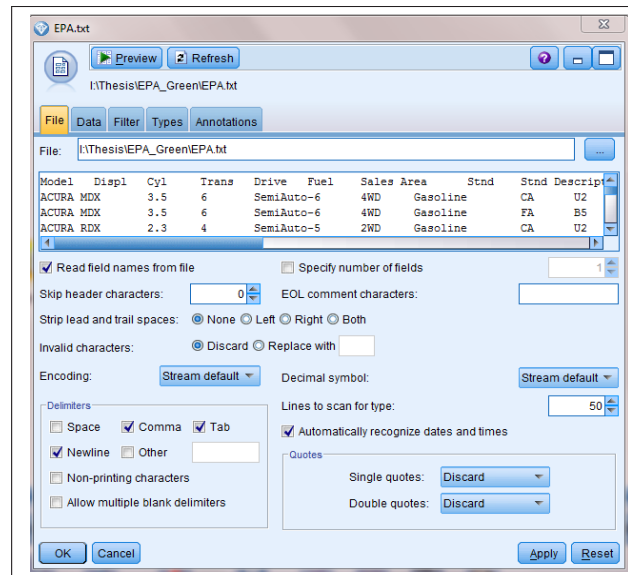


Figure 18. IBM® SPSS® Modeler opening data file

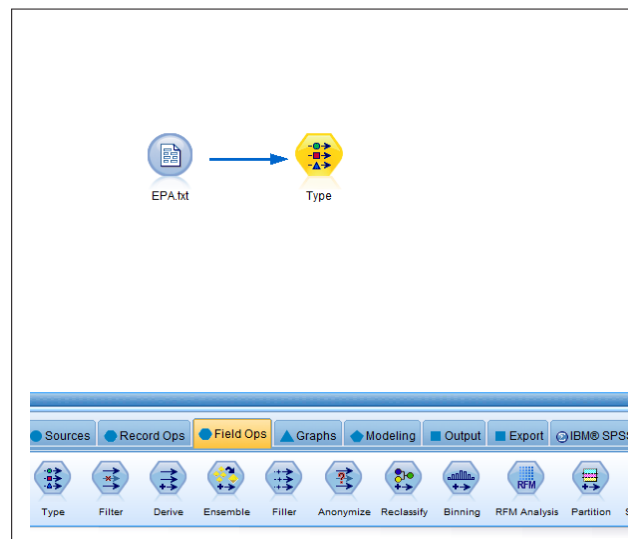


Figure 19. IBM® SPSS® Modeler field operations auto data prep and type



Once the file is uploaded, select the “Field Ops” tab and drag the “Type” node to the stream as shown in Figure 19. Right click on the “EPA.txt” node to connect the nodes. A small symbol with an arrow appears where the arrow is clicked, then dragged to the “Type” node and they are connected. Most nodes can be controlled using the right click feature.

The values are then read into the stream through the node by importing values from the text file and organizing them into fields. Of particular importance, are the measurement and role classes. The measurement category defines variables as continuous, nominal or flag. Field roles are selected as either input, target, both or none manually using the “Type” node. (See Figure 20).

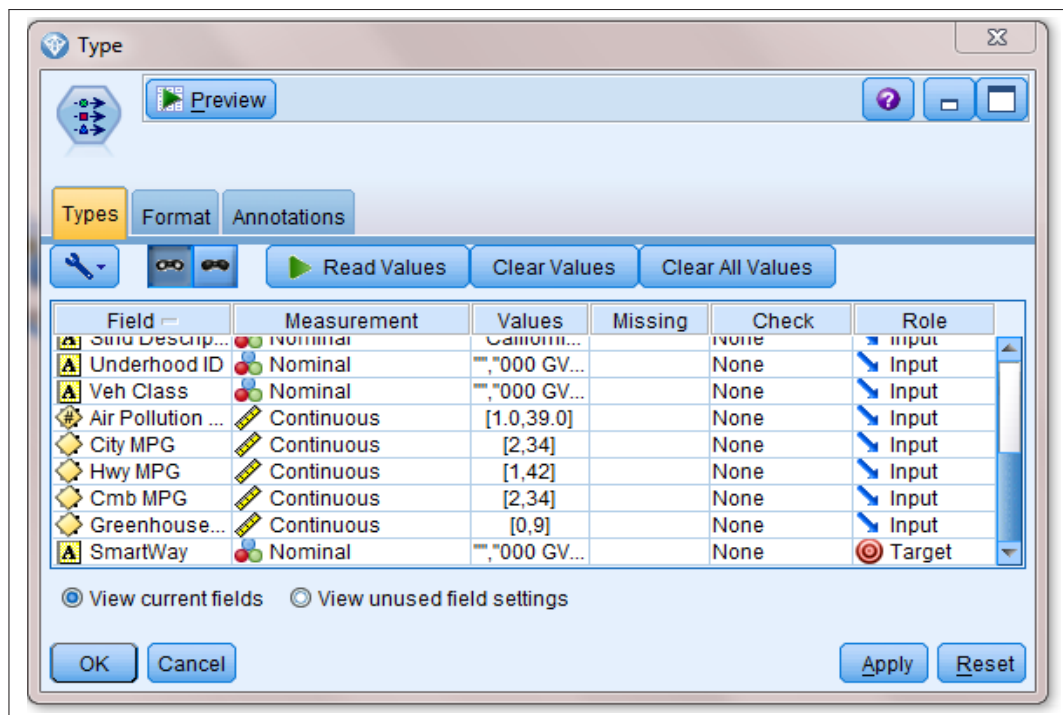


Figure 20. IBM® SPSS® Modeler field operation: type

Streaming the data through “Type” allows different classifications of the variables, as mentioned, including whether they are numerical or categorical or subgroups of either of those. This is where our target variable is defined for supervised learning models. Figure 20 shows the different classes of the node and SmartWay status is selected as a target variable to forecast the field for this example classification model.

It is important that all the data is imported, so another “Field Ops” node will be selected, called “Auto Data Prep,” to ensure the stream is operating smoothly and all the data is read in, as depicted in Figure 21. Continuous and numerical variables are listed as fields giving sample graphs, minimum and maximum values of each variable, mean, standard deviation, and the number of valid entries. Some fields had missing values and therefore had a lower number of valid data values.

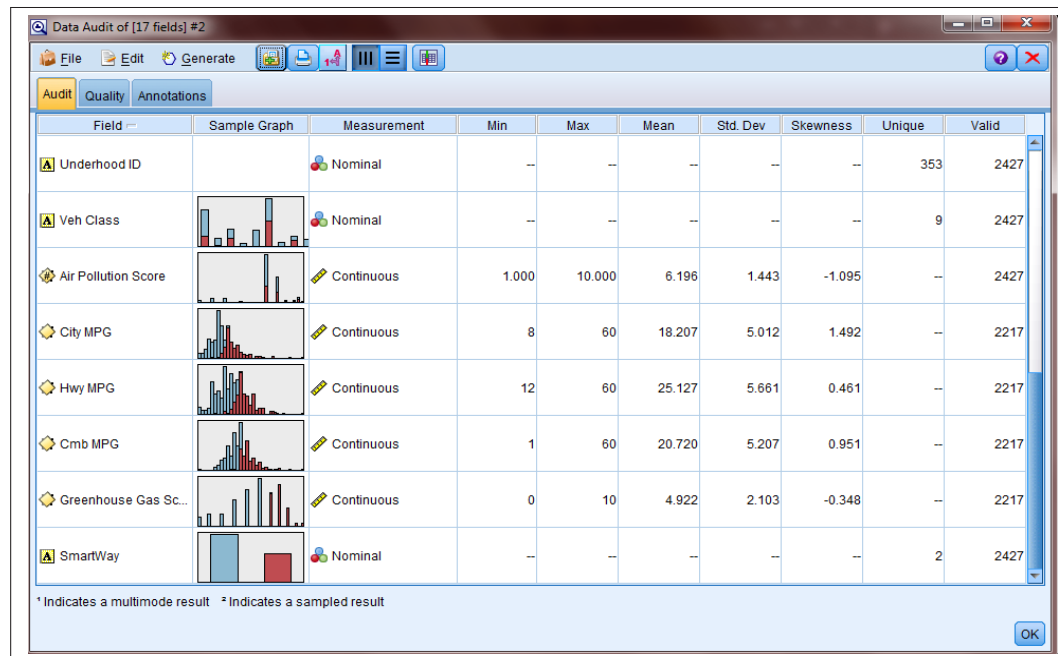


Figure 21. IBM® SPSS® Modeler EPA data audit

The “Auto Data Prep” node also yields another node, shown in Figure 22 marked “17 Fields”. This verifies that all the variables have been accepted into the model. A general “Auto Classifier” model will be created to test the software before creating a  $k$ -means clustering model. Using the target variable SmartWay and dragging the ”Auto Classifier” node to the stream, the stream is run and a SmartWay model is created shown by the dipyramidal node in Figure 22.

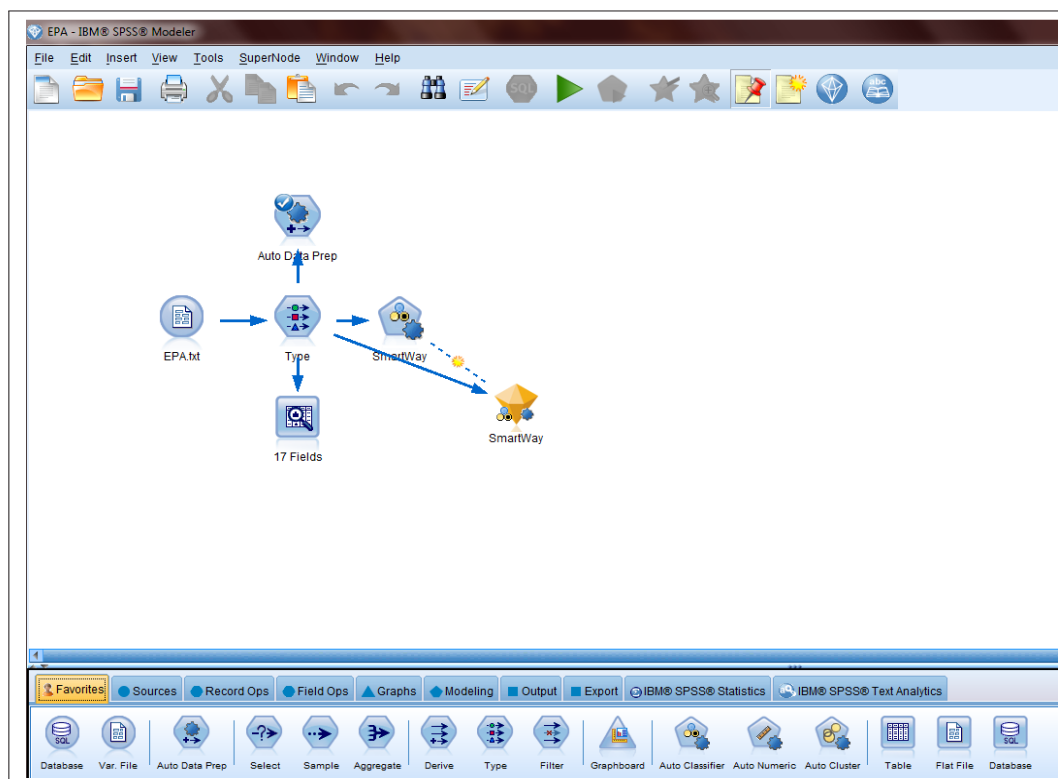
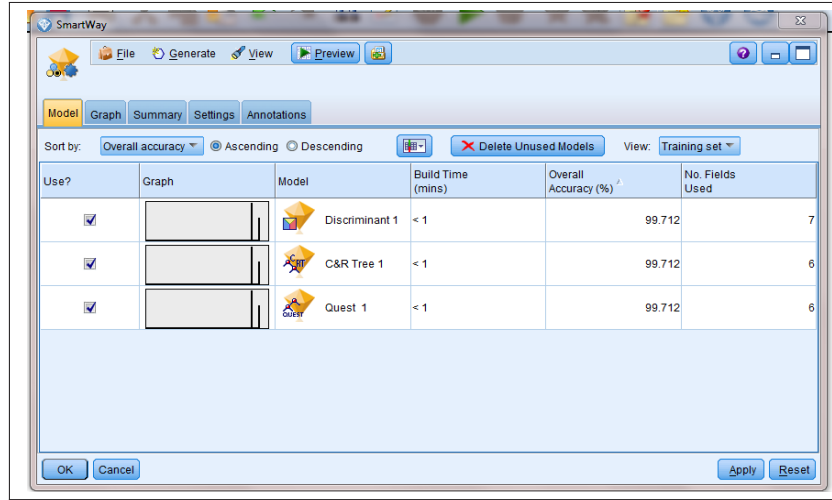


Figure 22. IBM® SPSS® Modeler SmartWay auto classifier model

### SmartWay Model Results

The classification predictive model used a combination of three different models: Discriminate, C & R Tree, and Quest. Each model had an accuracy of 99.7 %and were built in less than one minute. (See Figure 23).




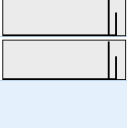
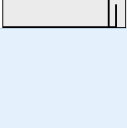
Use?	Graph	Model	Build Time (mins)	Overall Accuracy (%)	No. Fields Used
<input checked="" type="checkbox"/>		Discriminant 1	< 1	99.712	7
<input checked="" type="checkbox"/>		C&R Tree 1	< 1	99.712	6
<input checked="" type="checkbox"/>		Quest 1	< 1	99.712	6

Figure 23. IBM® SPSS® Modeler three classifying models

Discriminate Analysis (DA) is similar to MANOVA (multivariate analysis of variance), computationally by testing discriminant functions for statistical significance using the fundamental equations used in MANOVA. However, where MANOVA's predictors are dependent variables, DA predictors are independent variables. Classifications by variables discriminating between groups are independent and dependent for MANOVA and DA respectively (Poulsen & French, 2012).

Therefore, for the Discriminate model in this study, IBM® SPSS® Modeler determines that seven variables discriminate between two groups, those vehicles with and without SmartWay status. The pooled within-group correlations between each of the seven discriminating variables and standardized canonical discriminate functions were structured in a matrix. The results are shown in Table 1.

Table 1.

Structure Matrix of Discriminate Model for SmartWay Classification.

Variable	Function Correlation
Greenhouse Gas Score	.994
Cmb MPG	.868
Hwy MPG	.834
City MPG	.813
Air Pollution Score	.193
Disp	-.689
Cyl	-.641

C&R Tree (classification and regression tree) and QUEST (quick, unbiased and efficient statistical tree) are both classification tree algorithms (StatSoft, 2011) that determine group membership by dependent variable measurements on the the predictor variables on SmartWay status. Figures 22 and 23, created by the author, represent the summarized results of each algorithm respectively.

Differences appear in the number of input data values ( $n$ ) in the zero node, grouping by different greenhouse gas scores resulting in minor difference in each of the children nodes. The C&R tree in Figure 24 started with 1711 values, with node 1 having 1065, node 2 with 646. The majority of classification for node 1 did not have SmartWay status by 99.8% and node 2 has more values considered SmartWay by 99.4%. The QUEST tree started with 1722 values, with node 1 having 1072, node 2 with 650. The majority of classification for node 1 did not have SmartWay status by 99.7% and node 2 has more values considered SmartWay by 99.5% as shown in Figure 25.

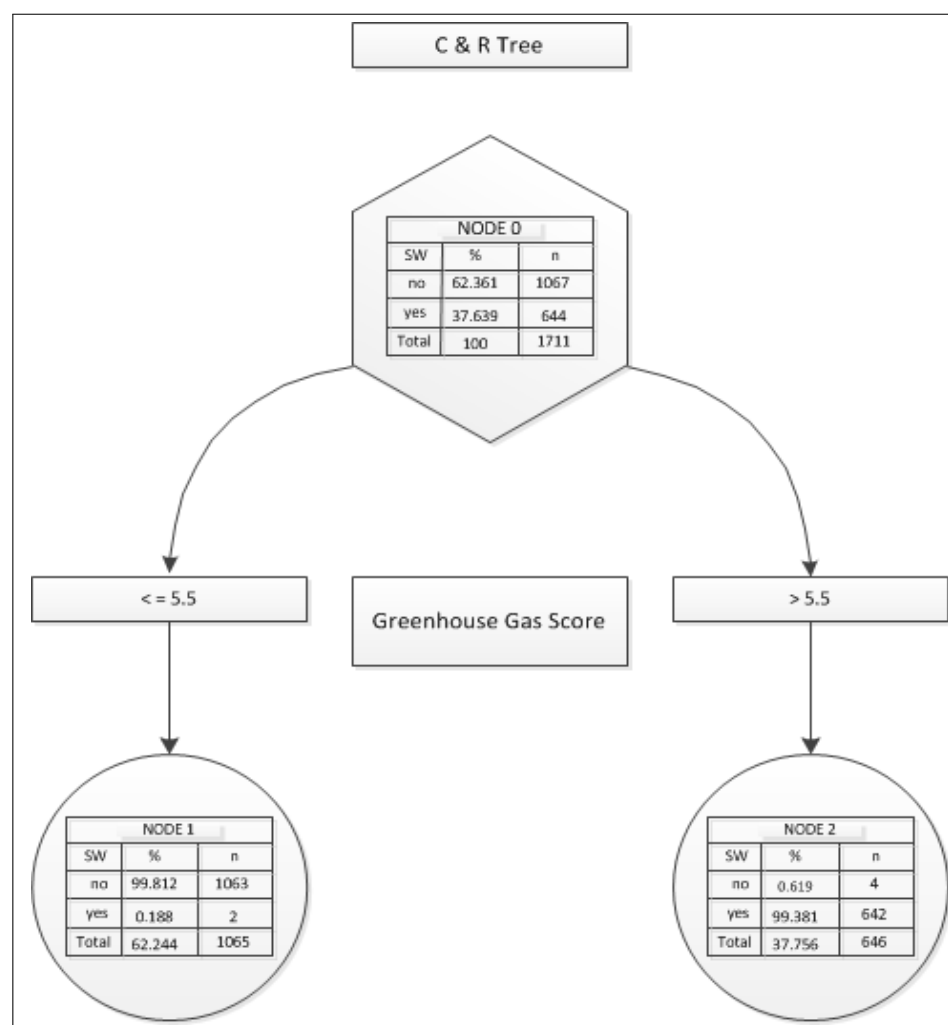


Figure 24. CR Tree

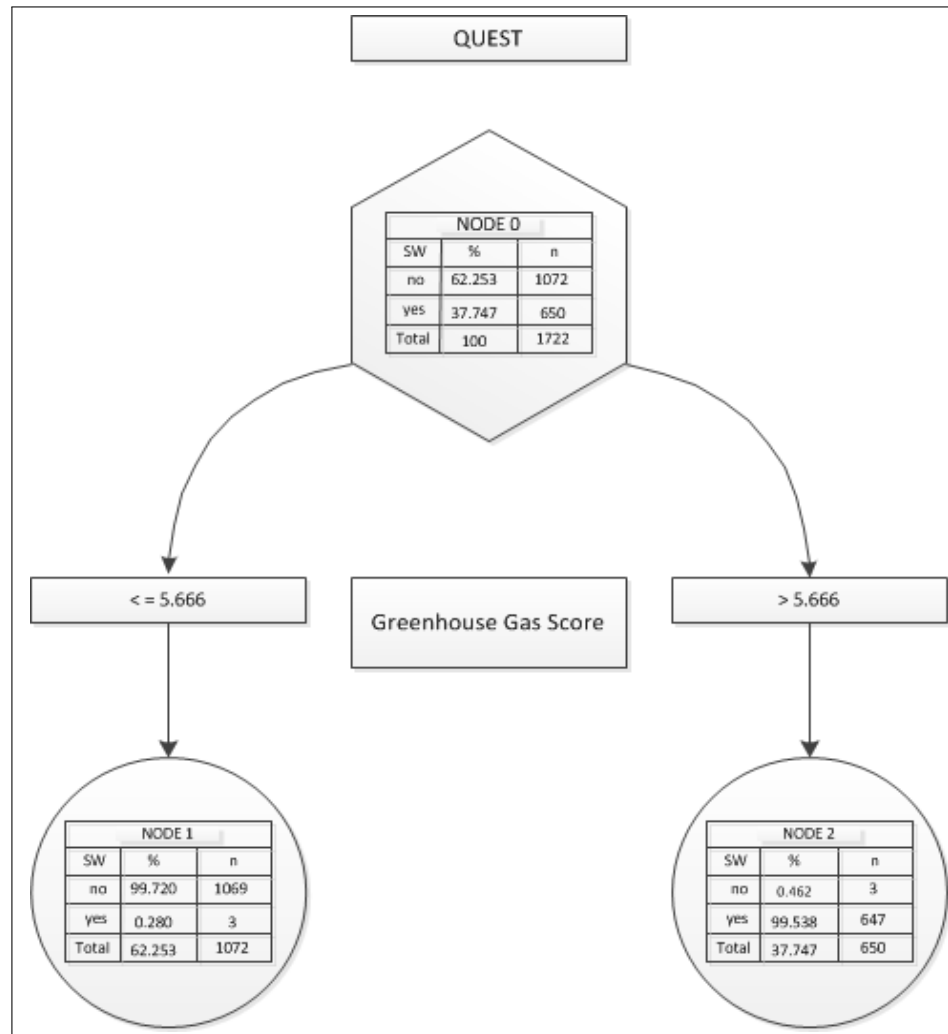


Figure 25. QUEST Tree

The predictions were similar to the assumptions from chapter 2 with respect to greenhouse gas scores and SmartWay status. As depicted in Figure 26, greenhouse gas scores were the most predictive of smartway status, however only by a small margin, and the sales area variable is the least predictive.

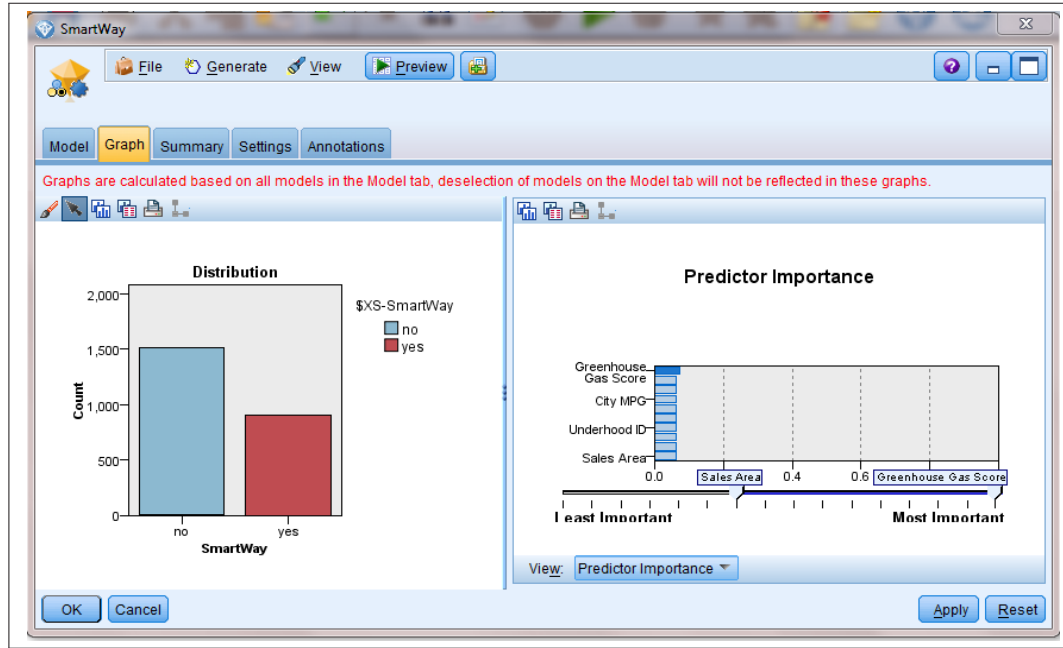


Figure 26. IBM® SPSS® Modeler SmartWay prediction

### Summary

Of particular interest for future models are the narrowed down input fields. These fields, yielding highest accuracy for classifying SmartWay status, are greenhouse gas score, air pollution score, combined gas mileage, cylinder and displacement values. These specific fields will be considered for a  $k$ -means clustering model, where, although a target variable is not valid, the fields can be used to find patterns between themselves specifically. These fields will serve as inputs into neural network models as well.



## CHAPTER IV

### *K*-MEANS CLUSTERING MODEL

#### ***k*-Means Clustering Technique**

Clustering in data mining and statistical analysis refers to the grouping of similar objects, cases or observations together as an unsupervised learning method (Larose, 2005). Unsupervised learning methods do not identify a target variable in the algorithm, rather examine and compare patterns and relations among all the variables and sometimes aid in determining variables needed further down the pipeline for algorithms such as neural networks, as will be discussed in the next chapter.

The *k*-means clustering algorithm is a conventional hierarchical clustering method that uses recursive partitioning. This particular clustering algorithm seeks to find *k* clusters where *k* is an integer defined by the analyst. Partitioning the observations by nearest average distances to *k* cluster centers within the data, the algorithm continues running while modifying the centers as the clusters expand. This is repeated until the centers, or centroids, no longer change or change minimally. The termination process of the algorithm requires convergence of the centers and is determined by minimizing the Sum of Squared Errors (SSE) within each cluster.

The algorithm runs essentially by first assigning how many clusters to partition for the model such that  $k = n$  clusters. Next, *k* clusters are randomly assigned centers. The algorithm then allocates records to a cluster based on nearness to a center and the process of redefining cluster centers and record appropriation repeats until there is no longer a change between the distance within records and centroids as the SSE gets smaller, converges and the algorithm terminates (Wagstaff, Cardie, Rogers, Schroedl, 2001).

The ratio of variance between the clusters with respect to variance within each cluster is calculated and observed to test the efficiency of the algorithm. As the variance between the clusters increases with respect to the variance within each cluster, the algorithm is more efficient and closer to its goal of grouping highly defined clusters.

The fundamental steps and calculations in the algorithm are as follows.

- How many  $k$  clusters do you want to partition? ( $k = n$ )
- Centers for  $k$  initial centers are randomly assigned as  $\mu_1, \mu_2, \dots, \mu_k$ . Alternatively, the first  $k$  records can be designated as the initial clusters.
- Records are allocated with reference to the nearest cluster center using a distance formula. The Euclidean distance formula is represented as:

$$d_E = \sqrt{\sum_{i=1}^k (x_n - \mu_i)^2}.$$

By definition, if the distance,  $d$ , between a vector,  $x_n$ , and a cluster center,  $\mu_i$ , is least, then  $n$  will belong to that group cluster  $S_i$ , represented by  $n \in S_i$ .

- Once a new record is added to a cluster, the new centroid is calculated and becomes the new  $\mu_i$  values.
- The algorithm repeats the process of finding distances and updating new center locations until the convergence of distance ratios, where the ratio no longer changes and the SSE has been minimized. SSE is defined (Weisstein, 2011) as:

$$SSE = \sum_{i=1}^k \sum_{n \in S_i} |x_n - \mu_i|^2.$$

- When this minimization has occurred, the centroids no longer change,  $k$  clusters are defined and the algorithm terminates. The efficiency of the algorithm and each cluster is analyzed by comparing ratios of variance between and within the clusters.

- The ratios of variation can be calculated by a simple estimation. The variance between clusters, or  $S_B^2$ , can be calculated by finding the distance between the centroids.

$$S_B^2 = \sqrt{\sum_{i=1}^k (\mu_1 - \mu_2)^2}.$$

When estimating the variance within clusters,  $S_W^2$ , the SSE can approximate this value so that maximizing the ratio can be calculated by:

$$\frac{S_B^2}{S_W^2} = \frac{\sqrt{\sum_{i=1}^k (\mu_1 - \mu_2)^2}}{\sum_{i=1}^k \sum_{n \in S_i} |x_n - \mu_i|^2} = \frac{d(\mu_1, \mu_2)}{SSE}.$$

### ***k*-Means Clustering Green Model**

Modeler is used in the same way for the  $k$ -means model as the SmartWay classification model processed previously. The only exception to the data preparation is that when reading the data through the “Type” filter, no target variable is selected and all variables are labeled as input values. Once the fields are read in and filtered through the data audit process, the “K-Means” node is selected. As shown in Figure 27, “K-Means” is connected to the prepared data.

Double-clicking the “K-Means” node will allow selection of  $k$  clusters by selecting the “Model” tab and typing in the desired number of clusters. In the first run, “3” is typed in for the initial run of  $k = 3$  clusters. Next, “Run” is selected and the model is created, signified by a dipyramidal node, which we will simply call the model.

By double-clicking the “K-Means” model, the results yield a model summary for 3 clusters as shown in Figure 28. Notice that only 15 variables were included as inputs into the model since “Drive” and “SmartWay” variables were flag variables due to their categorical nature. The  $k$ -Means algorithm works with normalized, numerical values ranging from zero

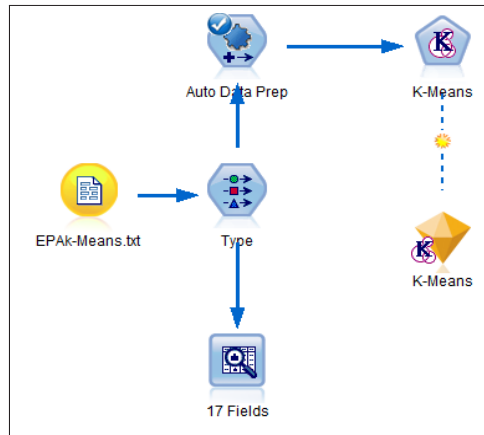


Figure 27. IBM® SPSS® Modeler *k*-means panel

to one. The model has a fair cluster quality measured by the separation between the clusters and cohesiveness within the clusters. A cluster size graph is shown to the right of the model summary. The first cluster contains 38% of the records, the second and smallest cluster contains 482 records, which is approximately 20% of the total records, and the third and largest cluster contains approximately 42% of the data, with 1022 records.

At the bottom of the “Cluster Sizes” window, several views may be selected depending on the contents of the left “Model Summary” window. At the current view of “Model Summary,” we can select the view “Predictor Importance” and the graph in Figure 29 appears.

In this model, the EPA FA and CA level descriptions for emission standards hold the greatest weight of predictive importance and fuel types hold the least from the clustering algorithm. The algorithm considers the SmartWay variable a predictor as well, illustrating reliability in categorical variables. Among the most important predictors, the algorithm identifies clusters by sales area, standard and standard description which describes the sales area and standard variables at 100% of predictor importance. Smartway status ranks approximately 98% of classification by predictor importance. Greenhouse gas score and gas mileages including highway and combined, rank moderately at 50%-70%. The number of engine cylinders, city gas mileage and engine displacement ranged from 45% to 48%

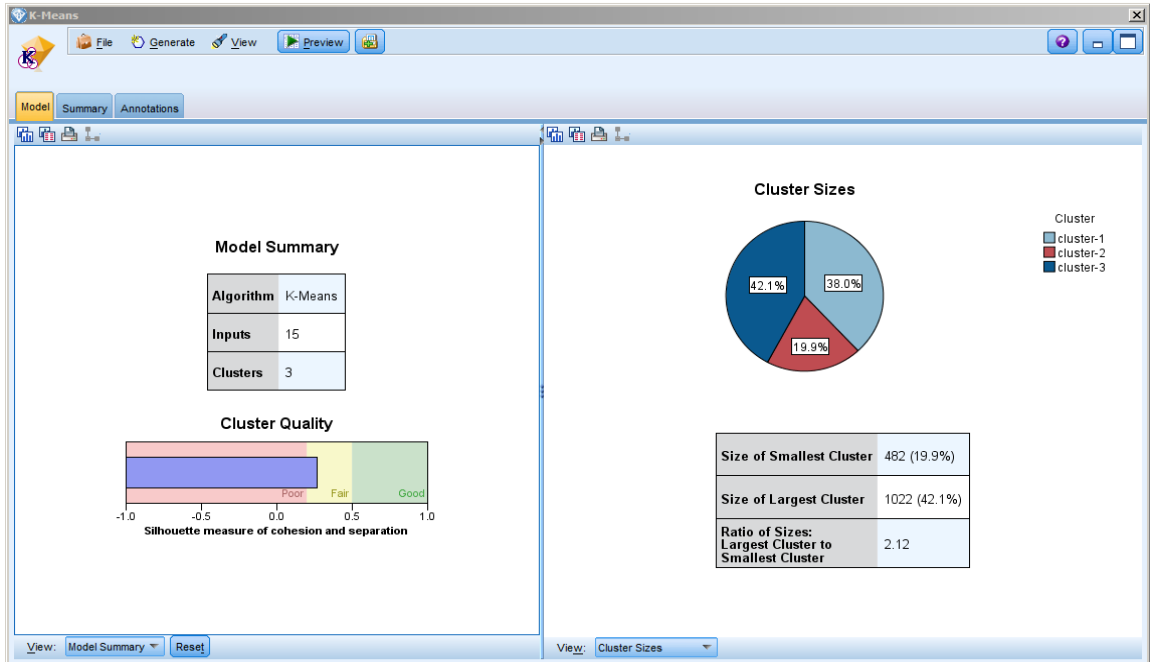


Figure 28. IBM® SPSS® Modeler  $k = 3$  clusters

respectively. Fuel and drive types, air pollution score, transmission type and vehicle class rank lowest from 5%-22% of importance for the initial three clusters.

Switching back to the “Model Summary” side, “View” can be selected and a closer look at the clusters themselves for further analysis can be obtained by selecting “Clusters”. The portion of results which contains the most important variables within each cluster is shown in Figure 30. The largest cluster is cluster-3 containing 42.1% of the vehicles and the smallest is cluster-2 with 19.9

Other variables considered within each cluster correspond with these main predictors. For example, gas mileages are higher in cluster-2 than in cluster-1. Standard and standard descriptions correlate with the sales area, therefore cluster-3 has a standard of U2 for California standards and B5 for clusters 1 and 2 with other federal standards and descriptions.

Another interesting variable that appears in the cluster distributions of this model is the combined gas mileage. As could be expected, the “Cmb MPG” for cluster-2 is highest

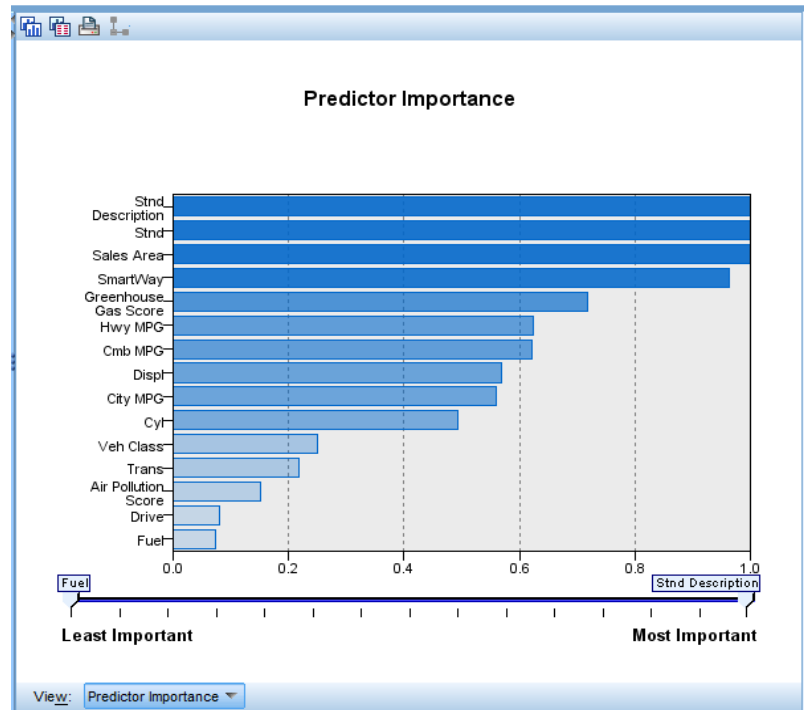


Figure 29. IBM® SPSS® Modeler  $k$ -means pi for initial 3 clusters

with an mean of 25.19 and cluster 1 has the lowest average at 17.21 mpg in correlation to the clusters having and not having SmartWay status vehicles respectively. Naturally, those vehicles with better gas mileages and lower emissions have more potential to having SmartWay status and belong to cluster-2. Cluster-3 has an average “Cmb MPG” of 21.37. It appears that combined gas mileage has a primary effect on SmartWay status and distribution of status within clusters consisting with vehicles sold outside California. Further analysis of these variables by examining the mean distributions in Figures 31 -33 may verify these findings.

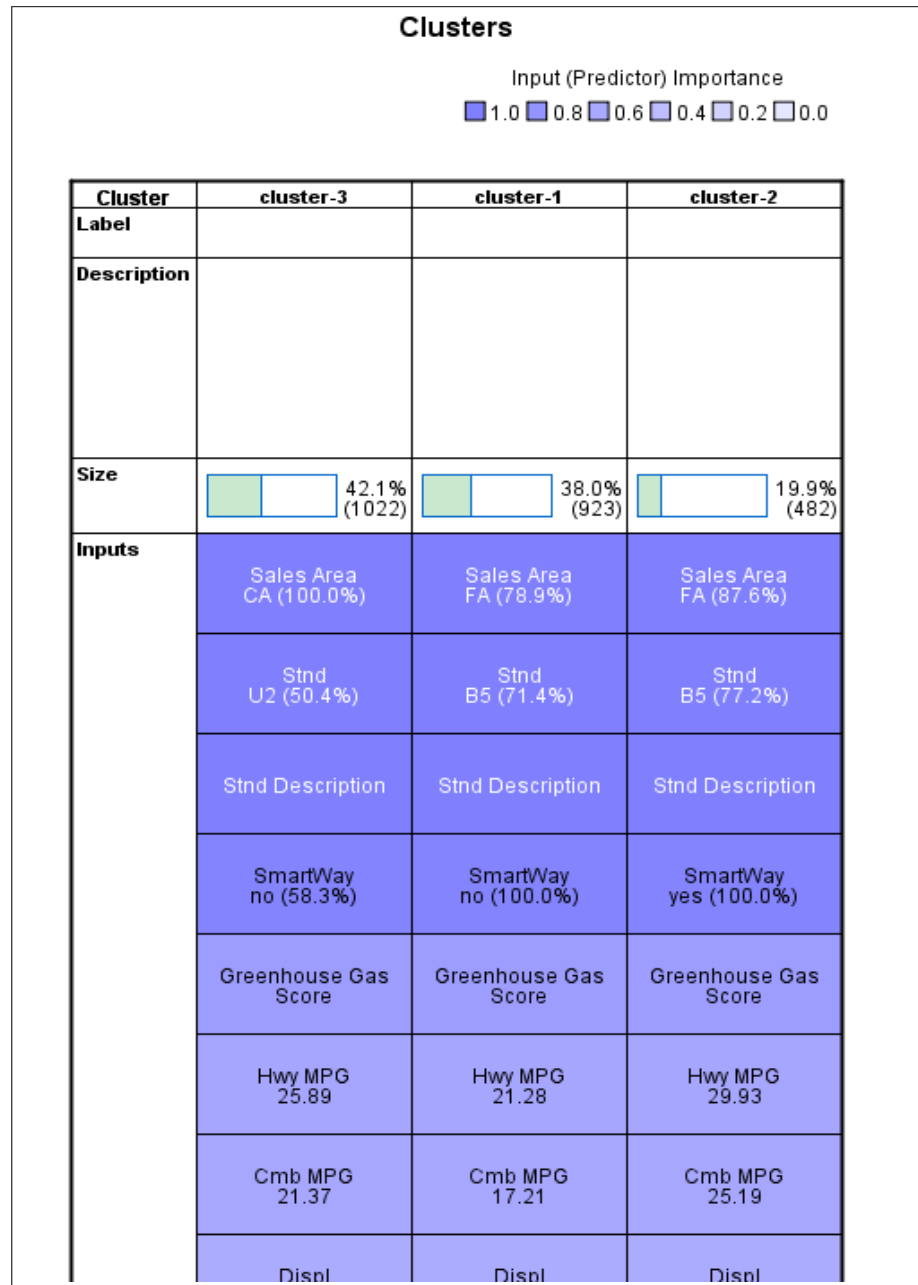


Figure 30. IBM® SPSS® Modeler  $k$ -means cluster summary for initial 3 clusters

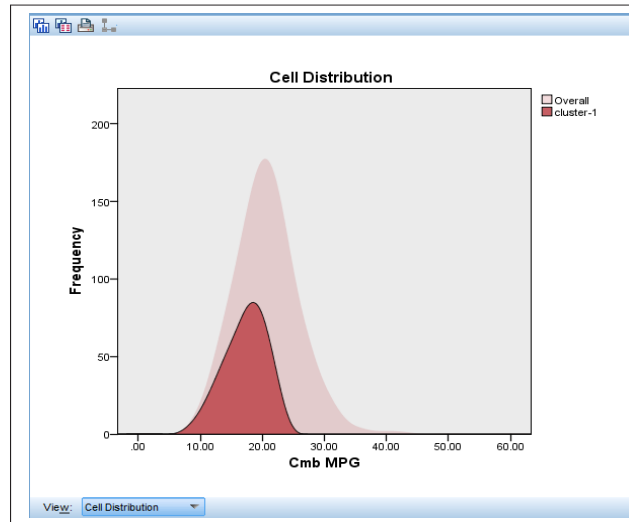


Figure 31. IBM® SPSS® Modeler  $k$ -means cluster-1 histogram

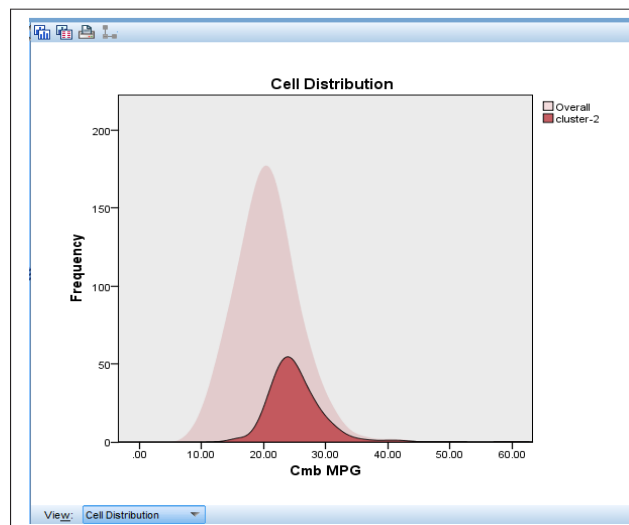


Figure 32. IBM® SPSS® Modeler  $k$ -means cluster-2 histogram



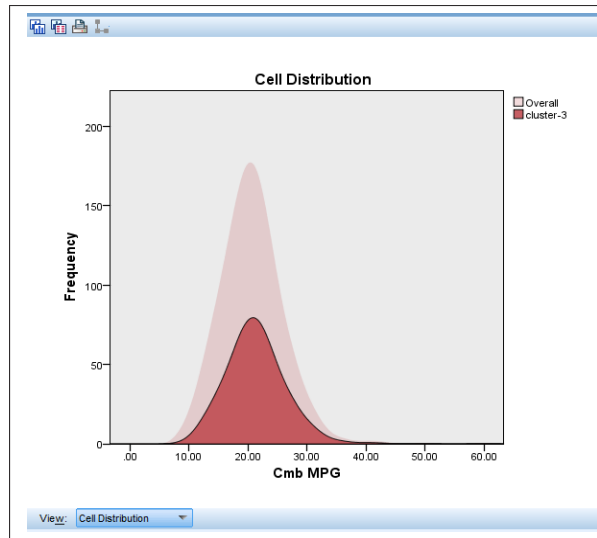


Figure 33. IBM® SPSS® Modeler  $k$ -means cluster-3 histogram

A closer look at the mean distributions of “Cmb MPG” give noticeable comparisons which illustrate how the cluster quality from the model summary is not optimal due to skewness. Figure 31 exhibits the normalized average of combined gas mileages for cluster-1 compared to the overall distribution which is normally distributed with low variation. The distribution of cluster-1 is left-skewed with respect to the overall distribution but moderately right-skewed within the cluster with moderate variation. The majority of gas mileages for cluster-1 is below the overall average of 20 mpg. Within the cluster distribution, the majority of gas mileages are greater than the mean gas mileage for cluster-1 of 17.21.

On the other hand cluster-2 in Figure 32 is right-skewed with respect to the overall distribution, but within the cluster itself is moderately left-skewed with low variation, where the cluster median and mean is greater than the overall average and the median within the cluster is slightly less than the average of 25.19.

Cluster-3, Figure 33, has combined gas mileage records normally distributed with moderate variation and normal with respect to the overall distribution. This cluster’s mean and median are closest to the overall average.

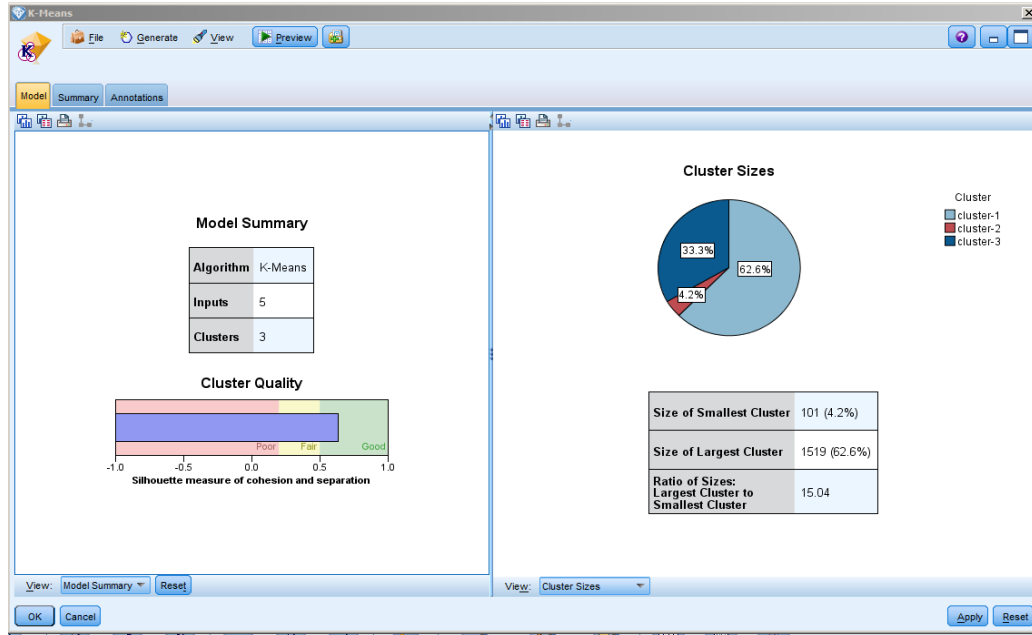


Figure 34. IBM<sup>®</sup> SPSS<sup>®</sup> Modeler  $k = 3$  clusters with si

Modeler illustrates the variation of the mean of every variable in the model with each cluster. In this thesis, another model of  $k = 3$  clusters with distinct inputs from the classification model results of top predictors will be implemented.

#### ***k*-Means Clustering Model With Selective Inputs, KMSI Model**

Closing all windows to the model, going back and double-clicking the “K-Means” node will allow selection of the “Fields” tab. The default selection of fields is called “Use type node settings” which will select all applicable variables in the model. Selecting the other option, “Use custom settings” allows specific selection of variables. A browsing option becomes available and individual variables from the data set can be selected as inputs one at a time. For the new KMSI model, variables SmartWay status, standard description, combined gas mileage, air pollution score and greenhouse score are selected by results from previous model’s important predictors and air pollution score as a variable of interest.

Double-clicking the “K-Means” node will allow selection of  $k$  clusters by selecting the model tab and typing in the desired number of clusters. The second model is run with  $k=3$  also. In the interface, it again shows us the dipyramidal “K-Means” node, creating the new KMSI model.

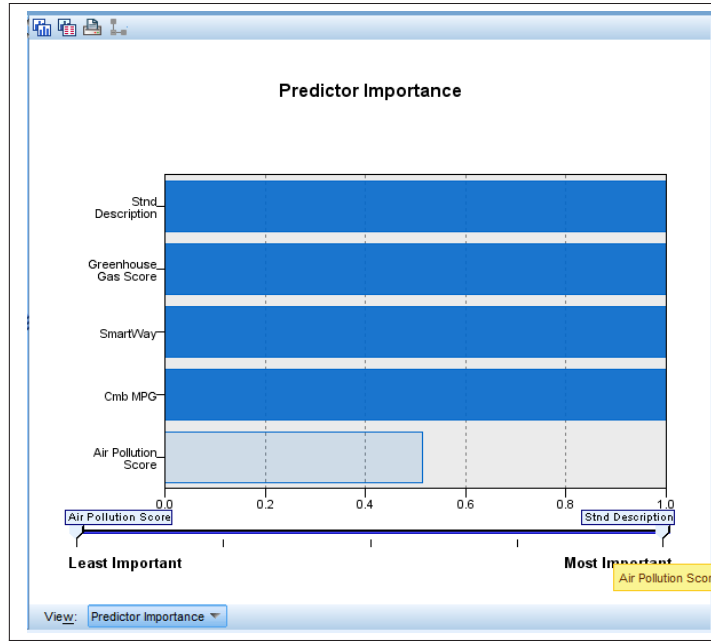


Figure 35. IBM® SPSS® Modeler  $k$ -means predictor importance for si three clusters

By double-clicking the KMSI model, the results yield a model summary with 3 clusters from 5 inputs as shown in Figure 34. The model now has a good cluster quality measured by separation between and cohesiveness within the clusters at 65%. The largest cluster, cluster-1, now contains approximately 63% of the data, with 1519 records and the smallest cluster, cluster-2 contains 101 records, which is approximately 4.2% of the total records. Cluster-3 is 33.3% of all records with 807 records.

Selecting “Predictor Importance” once again yields the output graph in Figure 35. In this model, combined gas mileages hold the greatest weight of predictive importance and air pollution scores hold the least from the clustering algorithm.

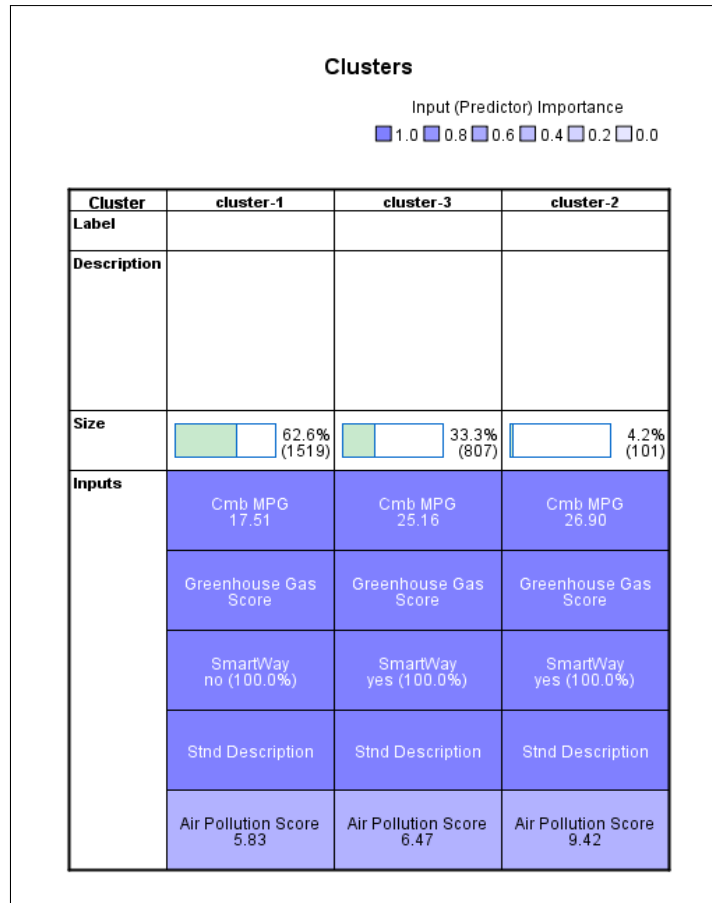


Figure 36. IBM® SPSS® Modeler *k*-means cluster summary for selected variables

Switching back to the “Model Summary” side, “View: Clusters” is selected and the results shown in Figure 36. Cluster-1 does not have Smartway status and clusters 2 and 3 contain records that have Smart Way status status. Vehicle characteristics in cluster-1 are low gas mileages, low greenhouse and air pollution scores and all the vehicles that do not have SmartWay status. Cluster-2 vehicles have SmartWay status, best fuel economy and highest greenhouse and air pollution scores, but the fewest number of vehicles, 101. Characteristics of cluster-3 are similar to cluster-2 but with lower gas mileages, air pollution and greenhouse gas scores and have 807 vehicles with these characteristics in his cluster.

Over half of the records in cluster-1, are distributed relatively between the overall distribution and is seen most apparently by looking at the distribution of greenhouse gas scores, GGS, between the clusters. (See Figures 37). Cluster-1 is right-skewed with moderate variation with the majority of greenhouse gas scores below a 6.0 and containing all of the overall greenhouse gas scores below approximately 6.0 that do not have SmartWay status.

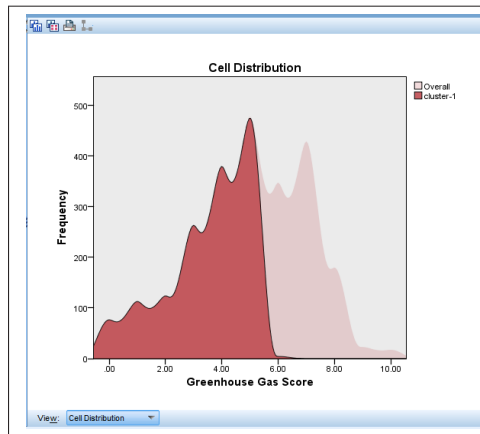


Figure 37. IBM® SPSS® Modeler  $k$ -means ggs histogram cluster 1

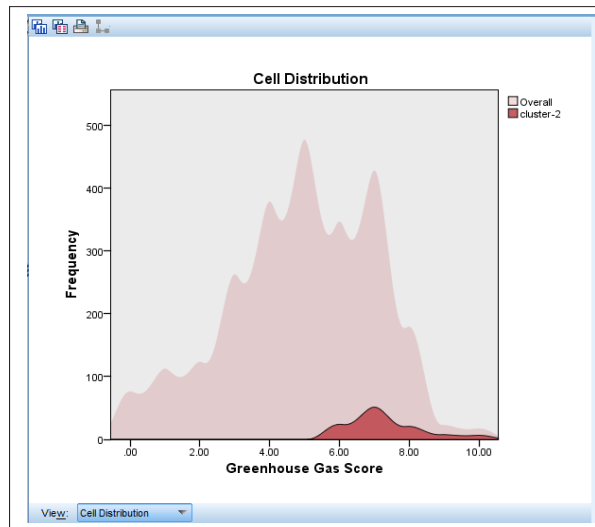


Figure 38. IBM® SPSS® Modeler  $k$ -means ggs histogram cluster 2

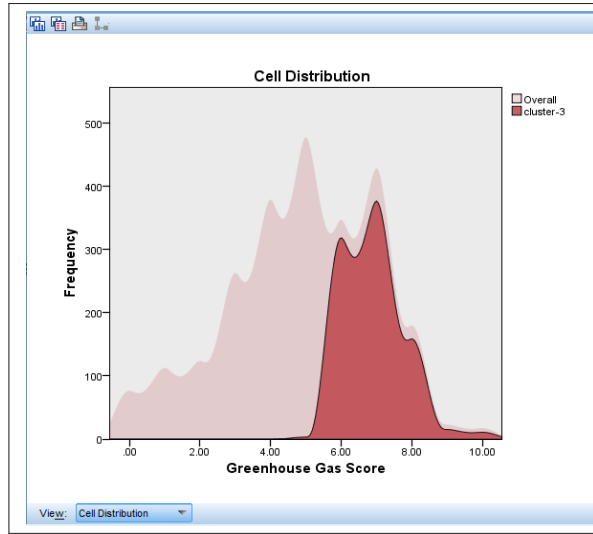


Figure 39. IBM<sup>®</sup> SPSS<sup>®</sup> Modeler  $k$ -means ggs histogram cluster 3

Cluster-2, as shown in Figure 38, is left-skewed with high variation and greenhouse gas scores ranging from 5.5 to 10. The cluster has a very small proportion of scores compared to the overall distributions of scores.

Similar to cluster-1 in distribution though left-skewed in comparison to the overall distribution, Figure 39, shows moderate variation and contains the majority of vehicles with greenhouse gas scores from 5.5-10.

### **$k$ -Means Summary**

In summary, comparing the distributions of various variables, in particular greenhouse scores, to the distributions between the clusters, significant correlation of the distributions support predictor importance of particular variables to determining which vehicles are assigned to each group. Selective inputs, partially derived from the previous classification model and using variable of predictor importance from the first  $k$ -means clustering model yielded more favorable and noticable patterns to grouping within the second model. Further analysis will be performed modeling an artificial neural network to see what variables are important predictors of SmartWay status.

# CHAPTER V

## ARTIFICIAL NEURAL NETWORK MODEL

### The Neural Network Technique

As mentioned briefly in Chapter I, artificial neural networks, or ANN's, similar in structure to an animal brain's neural network, consist of a feed forward, entirely joined and layered network of artificial neurons or nodes (Larose, 2005). The layers consist of an input, output and one or more hidden layers. The connections of nodes in each layer link to all the nodes of the next layer, initially by randomly assigned weighted values and initial bias weights as shown in the author's representation in Figure 40.

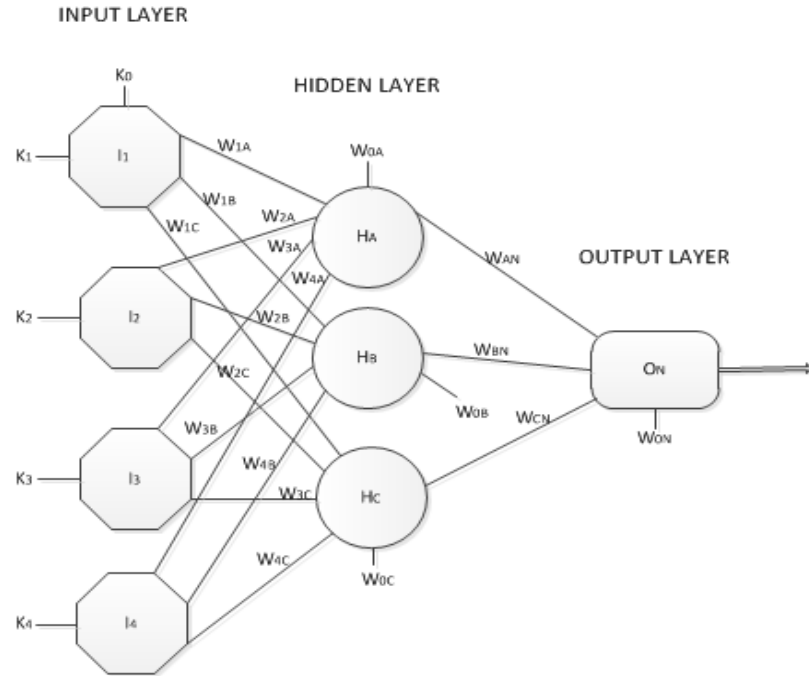


Figure 40. ANN with 3 layers

Combination functions, which are frequently summations, and activation functions, most commonly sigmoid functions, serve as nonlinear processes emulating the biological neurons in our own brains, signaling and firing other neurons. A neuron by definition is a nerve cell consisting of a cell body, axons and dendrites. Comparing artificial to biological neural networks in Figure 41 (Rhode, 2010), nodes act as neurons, inputs like dendrites, outputs are like axons and weighted values of an artificial neural network resemble synapses (not shown), very narrow spaces where signals are transmitted, of a neural network (MIT, 2011).

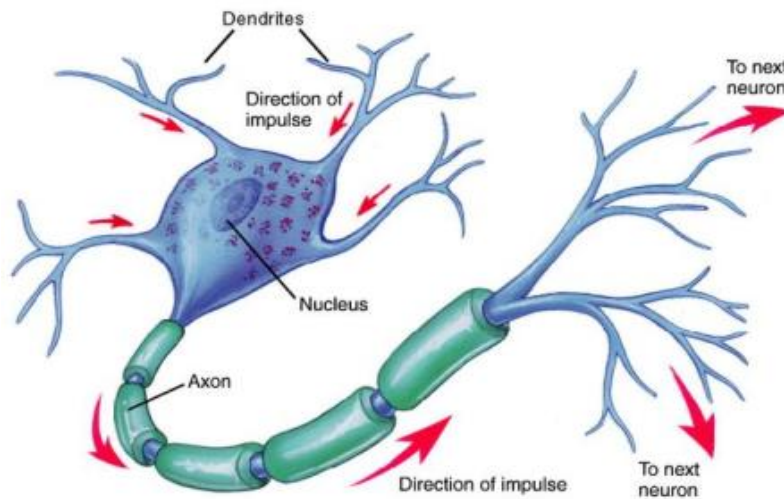


Figure 41. Biological neuron

Like  $k$ -means clustering, artificial neural networks can find patterns and relationships among data. However, unlike  $k$ -means clustering, ANN's serve multiple data mining tasks by representing an imitated approach to recognizing complex learning systems. Since neural networks, NN's, exhibit this learning behavior, they do not have to be reprogrammed and can still run efficiently if there is a problem within the network.

The downside of this learning phenomena, since ANN's are adaptive systems, is that training is a vital element for the network system to advance and work. This can take a considerable amount of time for large artificial neural networks. Outputs of the layers



lend themselves to inputs to other layers in the network and are continuous. This permits the model to have a more flexible range of tasks including classification, estimation and prediction.

### Normalizing the Data

Variable values within an ANN must be standardized using input and output coding represented by values ranging from zero to one. Continuous variables can be coded via min-max normalization. Min-max normalization is a process which transforms numerical values from the original data to a values ranging between zero and one. This is achieved by taking a given value, subtracting the minimum value in the set and dividing by the range (Larose, 2005). The equation for a normalized value from this process is as follows:

$$N = \frac{X - X_{min}}{X_{range}} = \frac{X - X_{min}}{X_{max} - X_{min}}.$$

For example, referring to the EPA data set, an excerpt of vehicles are shown in Table 4 with corresponding greenhouse gas scores and SmartWay status.

Make and Model	Drive	Fuel	Class	GGS	SmartWay
VOLKSWAGEN GTI	2WD	Gasoline	small car	7	yes
VOLKSWAGEN Golf	2WD	Gasoline	small car	7	yes
VOLKSWAGEN Jetta	2WD	Gasoline	small car	7	yes
VOLKSWAGEN New Beetle	2WD	Gasoline	small car	6	yes
VOLKSWAGEN Passat	2WD	Gasoline	midsize car	7	yes
VOLKSWAGEN Touareg	4WD	Diesel	SUV	4	no

Figure 42. Greenhouse gas score for min-max

Since, IBM<sup>®</sup> SPSS<sup>®</sup> Modeler will normalize the entire data set yielding it transformed and shown in later results using the “Data Prep” node as in the *k*-means models, normalization of the variables in the table only will be referenced here.

Each vehicle's greenhouse gas score is normalized from Figure 42. An example of the normalization of the Volkswagen, VW, GTI follows:

$$N_{gti} = \frac{7 - 4}{3} = 1$$

Taking the greenhouse gas score for the GTI, which is 7, and subtracting the minimum score of the six vehicles from Table 4, which is 4, and dividing that difference by the difference between the maximum and minimum values, again and coincidentally, 7 and 4. The normalized greenhouse gas score for the GTI is 1. To repeat the process for another example, the VW New Beetle is normalized:

$$N_{beetle} = \frac{6 - 4}{3} = 0.67$$

Normalized scores for the other vehicles are  $N_{golf} = 1$ ,  $N_{jetta} = 1$ ,  $N_{passat} = 1$ , and  $N_{touareg} = 0$ .

Categorical variables, on the other hand must be taken into careful consideration when normalizing. Variables can be defined, for example, male = 0 and female = 1. Variables such as fuel type in EPA's data set variables may be defined as gasoline = 0, ethanol = 0.2, CNG = 0.4, Diesel = 0.6, and Hydrogen = 0.8. However, this could cause complications as the learning system would learn irrespective of the meaning behind the numerical values. For example, females may be considered more valuable than males since 1 is greater than 0, or diesel fuel could be interpreted as a better fuel than ethanol because 0.6 is greater than 0.2, even though the values were assigned randomly or to distinguish the variables.

Because of the normalization process at the input layer, the output results must be denormalized, so they can be interpreted. For example, if greenhouse gas scores were inputs into the network, and yielded an output,  $O_n$ , of 0.43, then this normalized output value must be denormalized to interpret the corresponding gas score.

Using the previous example from Table 4, prediction for a greenhouse score that ranged from 4-7 would be the difference of values multiplied by the output normalized value and added to the minimum value.

A general denormalization formula is as follows:

$$D_{norm} = O_n(X_{range}) + X_{min}$$

and more specifically from the example:

$$D_{norm} = 0.43(3) + 4 = 5.29$$

such that the predicted greenhouse score would be approximately 5.

### **Training the ANN**

In general, an artificial neural network will, as stated before, consist of an input layer, at least one hidden layer, and an output layer with associated weights as illustrated in Figure 43. Similar to the user defining the number of clusters in  $k$ -means clustering models, the user defines the number of hidden layers and nodes in an ANN. This selection must be carefully arranged as there must be balance within the network for more accurate output. Since the ANN relies on a training set, there is danger of overfitting, where the training set is memorized and the output is generalized. On the other hand, there is also the danger of underfitting where the network cannot detect any predictive pattern.

An ANN is a supervised learning method and uses a training set consisting of a target variable and a large data set. Once the training set is run through the network, outputs are compared to initial values and like the  $k$ -means clustering algorithm, seek to minimize the SSE. Among the challenges of a training set is defining weights on the connections to minimize the SSE. This is solved by using the gradient-descent method and assuming that the training weight values have adjusted throughout the network.

The training set is utilized to minimize the SSE, by minimizing prediction errors, or errors between the output and actual values. Since there is a lack of actual values to compare

weights and weights must be included to minimize the SSE, random weights adjusted using the gradient descent algorithm are assigned to connections in the network. By taking the prediction errors and filtering them back through the network, the back-propagation algorithm runs, updating adjusted weights using the gradient-descent algorithm and a sigmoid activation function.

Simultaneously, ANNs take the weighted connections between the initial values, or input nodes, and the next layer of nodes and use a combination function of summations of multiplied weights and inputs to create a single value. This value feeds into a sigmoid function which then normalizes the value between zero and one. The output is a new value for the next node layer. These processes are outlined as in the next section.

### **ANN Algorithmic Functions**

Working simultaneously are several algorithms.

- Input nodes are based on data attributes.
- How many hidden layers are needed?
- How many nodes in each hidden layer do you want?
- Weights are assigned to each connection from  $i^{th}$  input to node  $j = W_{ij}$ .
- A combination function producing a value,  $V$ , with initial inputs,  $K$ , such as:

$$V = \sum W_{ij}K_{ij}$$

creates a single value on the  $j^{th}$  node.

- This value becomes an input to an activation function, most frequently a sigmoid function:

$$F(x) = \frac{1}{1 + e^{-x}}$$

which takes a linear value and normalizes the value between zero and one from the hidden layers to the output layer, which is important not only to this algorithm, but to the others involved as well.

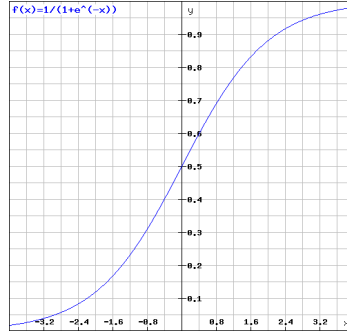


Figure 43. Sigmoid function graph

This function, as shown in Figure 44, is similar to a light switch with “On” and “Off” type outputs dependent on inputs.

- ANNs learn by minimizing the SSE through summing the prediction errors squared, or the the output values ( $V_{out}$ ) subtracted from the actual values ( $V_{act}$ ) squared over the output ( $out$ ) and total records ( $tot$ ) in the data set.

$$SSE = \sum_{tot} \sum_{out} (V_{act} - V_{out})^2$$

For the SSE to be minimized, a few methods need to be implemented as the actual values are unknown. Back-propagation uses the sigmoid function and the gradient-descent method to estimate these values and then filters them back through the network to further minimize prediction errors.

- The gradient-descent method adjusts weights on the connections to minimize SSE by taking the partial derivatives of the SSE with respect to the partial derivatives of each connection weight.
- Once the training weights are defined and those most current and best fitting are determined, the input weights run simultaneously through the network until the current weights supercede the best weights as calculated by SSE.
- The algorithm terminates once multiple stopping criteria are met such as a minimized SSE, accurate training set predicting unknown weights, and designated time the for the network to train.

### The Neural Network Model

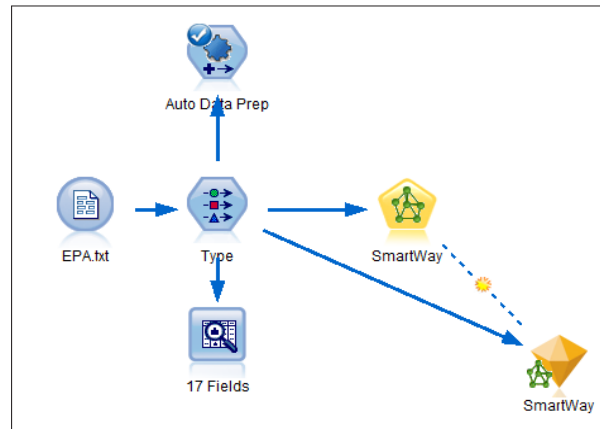


Figure 44. IBM® SPSS® Modeler neural net node

To create the Neural Network, the same process of filtering is used as in other models with the "Neural Net" node selected to set up for the model. Double-clicking the node yields options for defining predictors and the target variable. In this experiment, SmartWay status is the target variable and all other variables in the data set are initially used as predictors. See Figures 45 and 46 for initial settings.

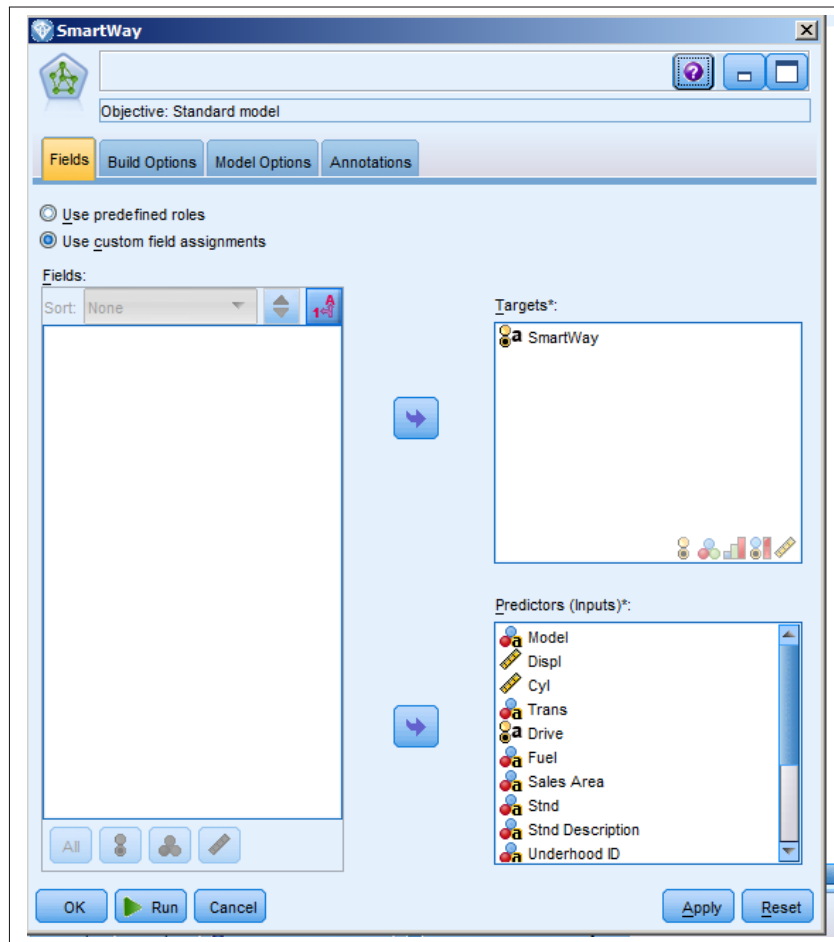


Figure 45. IBM® SPSS® Modeler neural net fields

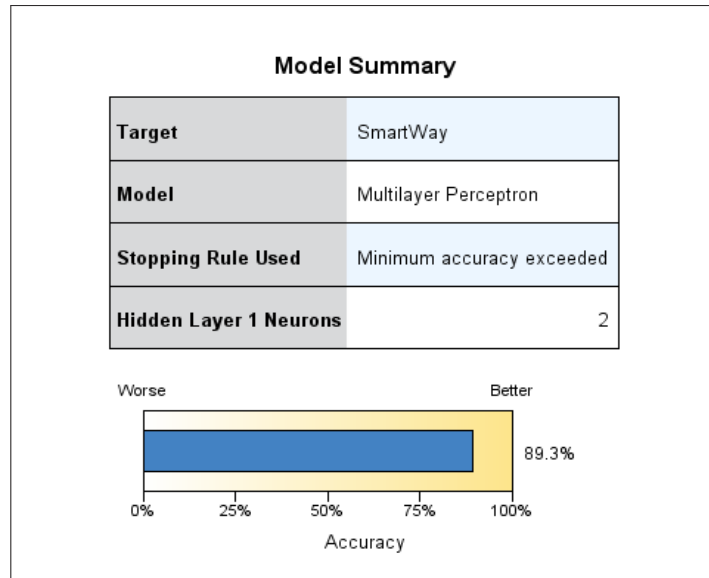


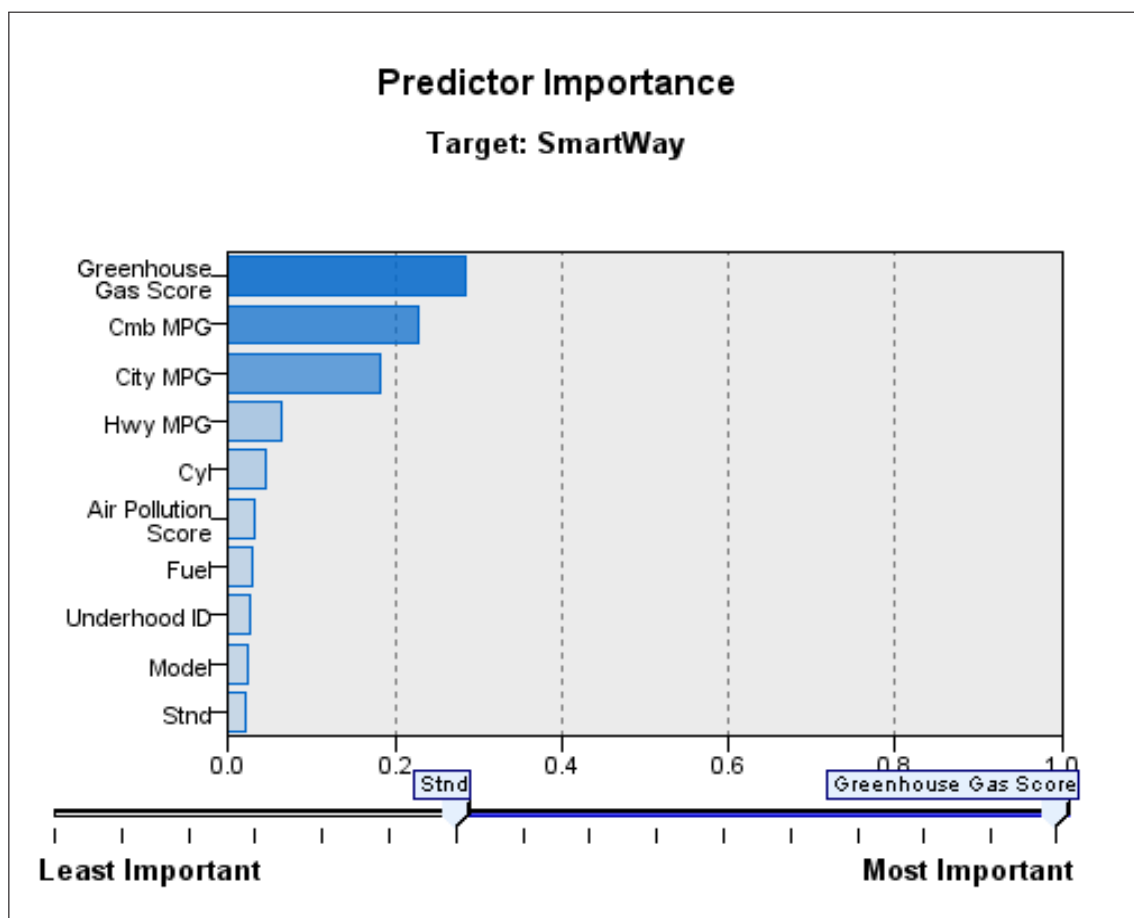
Figure 46. IBM® SPSS® Modeler neural network model summary

Running the model, the network is created and graphed accordingly. The model automatically produces a network with two hidden layers and an 89.3% accuracy of the model and using minimum accuracy for the termination or stopping rule as shown in Figure 47.

Further investigation of the model leads us to Figure 48 where predictor importance is generated and greenhouse gas score is the leading predictor with the greatest importance and the least important is engine displacement, not shown in the figure. Greenhouse Gas Score is the top predictor variable at 30%. Combined Gas mileage follows with 23%, and variables, city gas mileages were 18% important, and highway gas mileages, cylinder sizes, air pollution scores, fuel types, vehicle identifications, models and standard descriptions were under 5% of predictor importance for SmartWay status.

The Neural Network model in figure 49 shows the top ten predictors for Smartway status, the two hidden layers, the SmartWay target output node and a bias node act like initial random weight factors generated by IBM® SPSS® Modeler and assigned to the input and output weight connections for training the network.





*Figure 47.* IBM<sup>®</sup> SPSS<sup>®</sup> Modeler neural network predictor importance

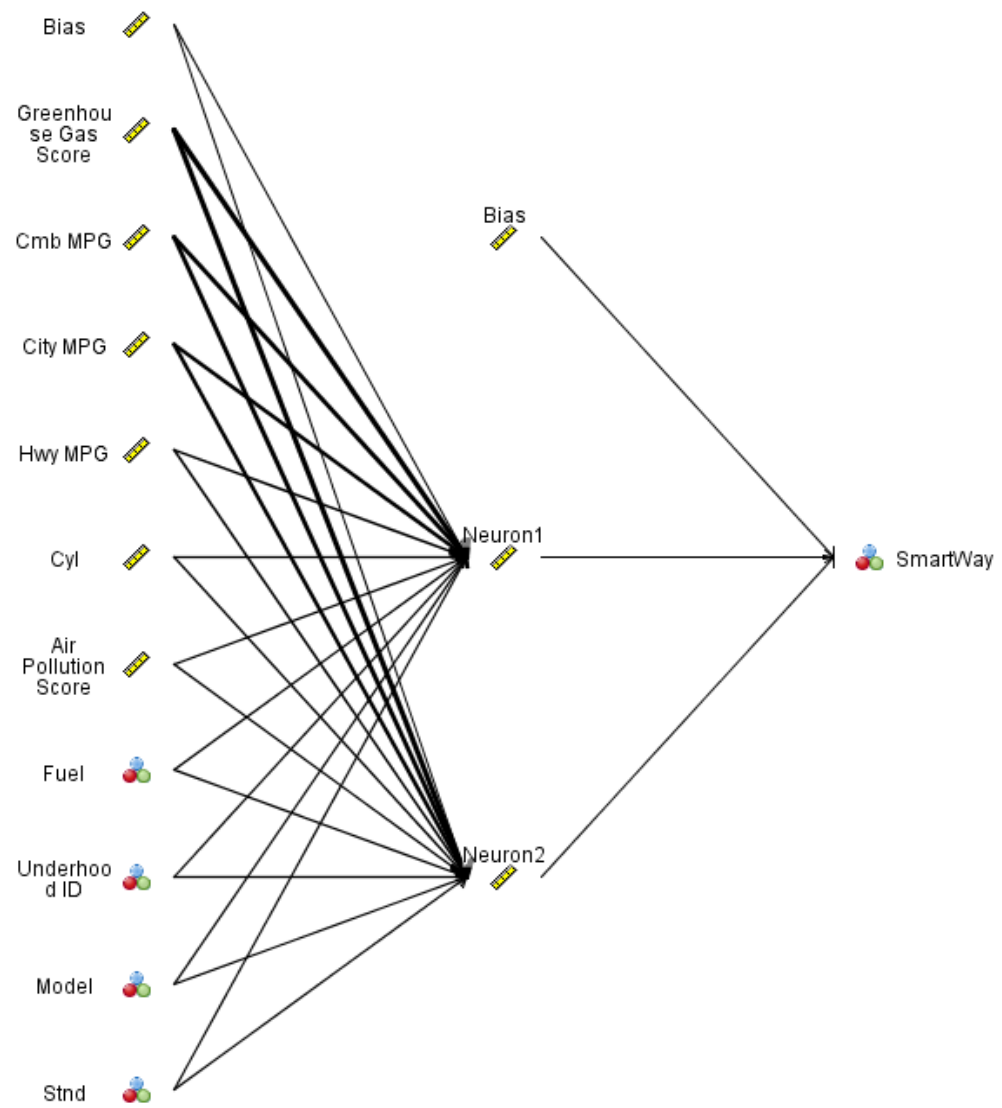


Figure 48. IBM® SPSS® Modeler artificial neural network using all variables

### *k*-means Clustering Output as Input to an ANN

Next, the output variables of the top predictor variables from the *k*-means clustering model (i.e., standard description, greenhouse gas score, air pollution score, and combined gas mileage) are used as inputs to the Neural Network model. This yields a new network with two hidden layers with an increased accuracy to 91.2%. Again, minimum accuracy for the termination or stopping rule is used as shown in Figure 50.

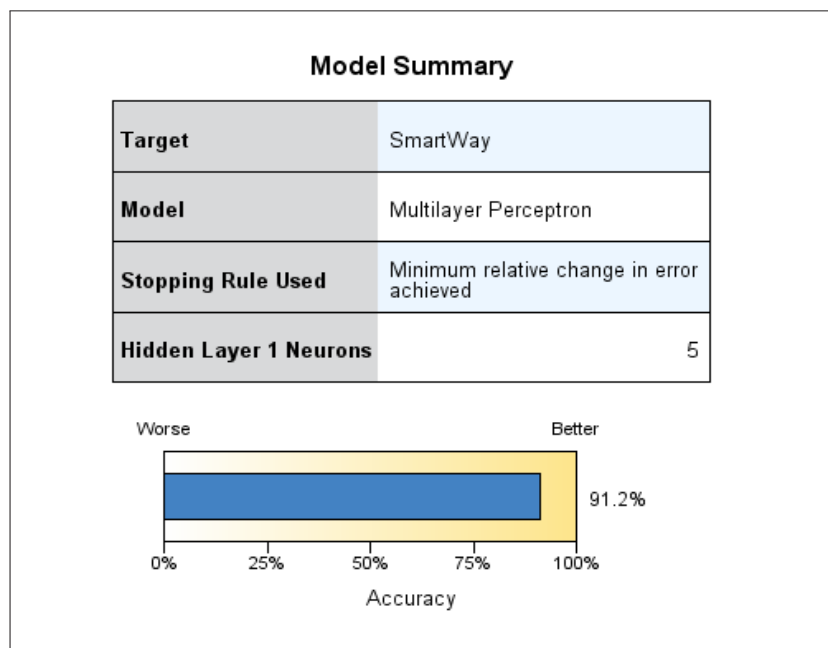


Figure 49. IBM® SPSS® Modeler neural network *k*-means input model summary

Figure 51 reveals greenhouse gas scores as the greatest predictor once again of SmartWay status by 46%, however the least important is standard description at 8%. Combined gas mileages yielded 40% predictor importance and air pollution scores predict SmartWay status by less than 10%.

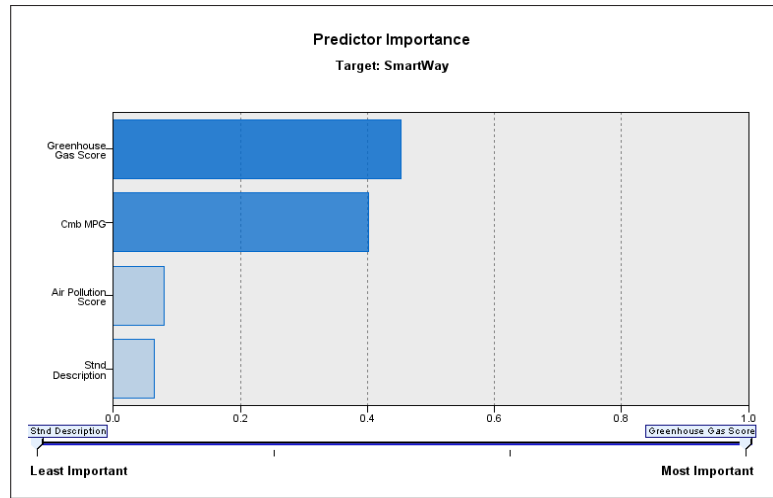


Figure 50. IBM® SPSS® Modeler neural network *k*-means input predictor importance

Of the variables observed in this network, Figure 52 shows that 0% were predicted but not observed, 0.3% were not predicted, but observed, 100% were not predicted or observed and 99.7% were predicted with an overall classification for SmartWay status at 99.9%. Finally, Figure 53 shows the ANN for SmartWay status as a result of using only *k*-means top cluster variable types as inputs.

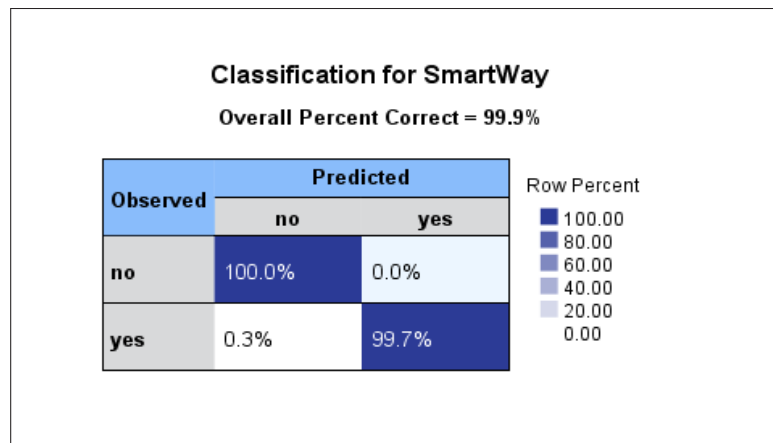


Figure 51. IBM® SPSS® Modeler neural network *k*-means input classification summary

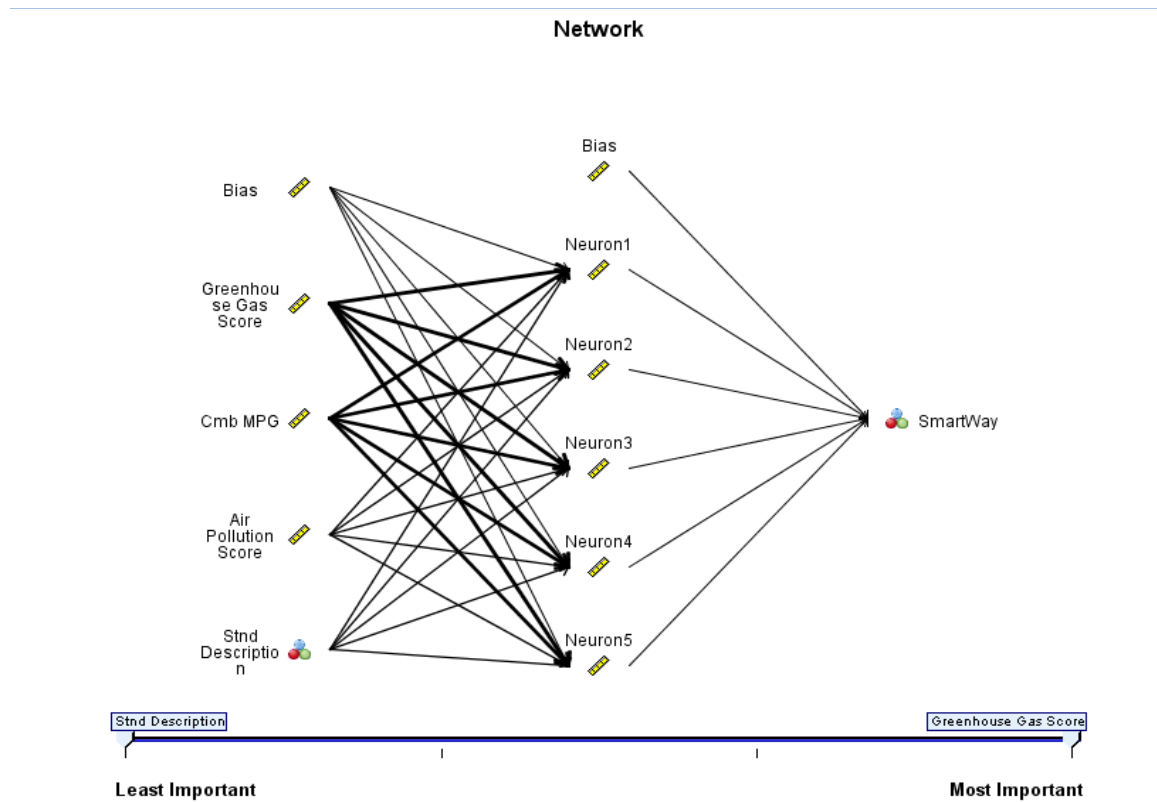


Figure 52. IBM® SPSS® Modeler neural network  $k$ -means input model

### Summary

In summary, the predictive output variables from the  $k$ -means cluster analysis used as input variables to an ANN yielded a more efficient model than the ANN with all variables by 1.9%. Both models accurately reflected greenhouse gas scores and combined gas mileages as the greatest predictors for SmartWay status of vehicles. Greenhouse gas scores were a strong influence on classifying vehicles by groups in the  $k$ -means clustering model as well as an important predictor of SmartWay status in the ANN. Predicting whether a vehicle will classify as SmartWay will be primarily determined by its greenhouse gas score.

## CHAPTER VI

### CONCLUSIONS

The Cross-Industry Standard Process for Data Mining, CRISP-DM, provided a basis for examination and analysis via data mining the EPA's Green Vehicle Guide data set. The first research understanding phase from Chapter 1 questioned many areas which were addressed in this study, such as (i), what processes the data underwent for preparation and (ii), which data mining techniques were used and why. The statistical implementations of data mining techniques as well as the statistical measures involved in the data mining techniques were defined.

Chapter 2 provided an in depth look at both data understanding and data preparation phases. Understanding the variables and definitions of the data set was imperative to later modeling processes. Exploring both the data set itself as well as the providing website assisted with the familiarization process.

The data preparation phase was the most time consuming element of analysis. Data had to be reorganized, filtered and missing values redefined while maintaining data integrity. Excel and IBM® SPSS® Statistics were used in this phase to clean the data. Techniques and useful variables to use as possible target variables, flag variables and inputs began to filter into preparation for the future models. Using IBM® SPSS® Modeler, data preparation analysis transformed variables as needed for the  $k$ -means and artificial neural network models.

Again, IBM® SPSS® Modeler was used for the modeling phase which included classification,  $k$ -means and artificial neural network models in chapters 3, 4 and 5. Adjustments had to be made regarding target variables and input values as well as using several modeling techniques for conclusive findings. Runs were made in both  $k$ -means and ANN models to compare general and selective inputs.

Conclusions regarding this study are that the  $k$ -means KMSI model was a good, quality model that yielded conclusive clusters to use for further research, including the great predictor, greenhouse gas scores. The ANNSI model proved successful as well and yielding greenhouse gas scores and combined city and highway fuel economy as reliable predictors of SmartWay status.

Unfortunately, IBM® SPSS® Modeler software due to the nature of node connectivity would not allow the clusters themselves from the KMSI to serve as inputs into a networking model. As part of the deployment phase of this study, future research could be conducted as far as recreating data sets synonymous to the data records in each cluster from the KMSI model. From there, further modeling techniques, including ANN, could be implemented and examined for any prediction obscurities and findings. Further research could also include an updated data set.

Predictions based on analysis of this data set are as expected with all models claiming greenhouse gas scores to be the greatest predictor variable for SmartWay status. Therefore, engineers and companies should focus on better technology to improve greenhouse gas scores if SmartWay status is a goal. Further analysis with other models may yield a higher weighted predictive variable although the probability is low.

## REFERENCES

- Executive Branch of the Federal Government. Green Vehicle Guide Data Downloads.  
Retrieved February 5, 2011, from [www.data.gov/raw/2004](http://www.data.gov/raw/2004)
- Larose, D. T. *Discovering Knowledge in Data: an Introduction to Data Mining*.  
Hoboken, NJ: Wiley-Interscience, 2005.
- MIT. Artificial Neural Networks. Retrieved November 11, 2011, from  
<http://mitpress.mit.edu/books/chapters/0262062410chap1.pdf>
- Office of Transportation and Air Quality (OTAQ). About the Ratings — Green Vehicle  
Guide — US EPA. Retrieved February 5, 2011, from  
<http://ofmpub.epa.gov/greenvehicles/Aboutratings.do>
- Office of Transportation and Air Quality (OTAQ). Green Vehicle Guide— US EPA.  
Retrieved February 5, 2011, from <http://ofmpub.epa.gov/greenvehicles/Index.do>
- Poulsen, J., & French, A. Discriminant Function Analysis. Retrieved January 14, 2012,  
from <http://userwww.sfsu.edu/efc/classes/biol710/discrim/discrim.pdf>
- Rhode, C. (2010). Intro to Neural Networks. Lower Columbia College.  
Retrieved April 20, 2012, from <http://lowercolumbia.edu/students/academics/facultyPages/rhode-cary/intro-neural-net.htm>
- StatSoft, Inc. Classification and Regression Trees (C&RT). *Electronic Statistics Textbook*.  
Tulsa, OK: StatSoft, 2011. Retrieved January 14, 2012, from  
<http://www.statsoft.com/textbook/classification-and-regression-trees/>
- Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained K-means Clustering  
with Background Knowledge. Retrieved August 8, 2011, from  
<http://web.cse.msu.edu/cse802/notes/ConstrainedKmeans.pdf>
- Weisstein, E. W. K-Means Clustering Algorithm. *MathWorld—A Wolfram Web*.  
Retrieved August 8, 2011, from  
<http://mathworld.wolfram.com/K-MeansClusteringAlgorithm.html>