Statistical Analysis for an Automobile Research Firm

John Wensink

MIS 450 - Data Mining

Colorado State University-Global Campus

Dr. Steve Chung

June 22, 2020

Statistical Analysis for an Automobile Research Firm

Using data mining techniques on the CARS database located within SAS will provide our organization with powerful insights into how the measured fuel economy in mpg city and mpg highway correlates with other analysis variables such as weight, horsepower, MSRP, invoice, number of cylinders, and engine size for the vehicles found in the database. Using SAS University Edition, insights were able to be gleaned from performing statistical analysis on the data such as summary statistics, distribution analysis, as well as clustering variables analysis showing strong evidence of correlation based on the results.

**Statistical Analysis**

Using SAS University Edition’s pre-loaded analysis tasks, statistical analysis was performed on the CARS dataset. Starting out, the Summary Statistics task was able to generate informative statistics for our analysis variables pertaining to all observations within our dataset (SAS, 2020). Although the capability to create histograms and box plots exists in the summary statistics task, the information that they provided did not seem to be useful, and they have been omitted. The summary statistics analysis provided useful information into different manufacturer’s vehicle’s statistics as they pertain to MSRP, invoice, mpg city, and mpg highway. For each manufacturer, the analysis report describes the number of observations in the group, the variable with its label code, mean, standard deviation, minimum values, and maximum values as well as n, which in this case matches the recorded observations. Perhaps unsurprisingly, the lighter vehicles manufactured by Honda, Mini, and Scion all came in with highway mpg > 30. Perhaps we will consider the value of a vehicle’s weight to be accounted for in cluster analysis.

The second statistical analysis task we performed was distribution analysis on MSRP, invoice, mpg city, and mpg highway. Distribution analysis gives us data about distributions of continuous variables (SAS, 2020). Standard histograms are created, with the capabilities to add curves to an easier to digest description of the different variables. To these histograms and curves, one is able to add statistics that measure different statistics such as n number of observations, mean, standard deviation, skewness, and kurtosis seem to be valid inset statistics. From this plot, we can see that the mean mpg highway was 26.8mpg, and mpg city was 20.06mpg. The vast majority of vehicles were priced under $60,000 with a high degree of left skewness (2.79).

**Clustering**

For the purposes of this investigation, we will analyze nine variables in the CARS dataset including engine size, cylinders, horsepower, weight, MSRP, invoice, as well as mpg city and mpg highway. The pre-loaded task Cluster Variables “The Cluster Variables task finds clusters of variables to use in additional clustering or to select non-redundant variables in further clustering” (SAS, 2020). We will use the principal components analysis with a maximum of 10 clusters. When an n-dimensional cluster’s eigenvectors and corresponding eigenvalues reach a maximum second eigenvalue, the algorithm will split the variable off into a different cluster (SaNgasoongsong and Bukkapatnam, 2012). Not surprisingly, the algorithm has placed invoice, MSRP, and horsepower into their own cluster. The other cluster seems to contain the principal components of fuel efficiency; weight, cylinders, and engine size are clustered together with mpg city and mpg highway. If we remove MSRP and invoice from our analysis, and group the remaining variables into two clusters, as we expected, the principal component of fuel efficiency is the weight of the vehicle. The second cluster seems to suggest that horsepower, cylinders, and engine size, while related to fuel efficiency, do not share the same level of relationship that the vehicle’s weight does. One might suggest that a vehicle’s mpg is primarily a function of the vehicle’s weight, with horsepower, cylinders, and engine size having a lesser effect on the fuel efficiency of the vehicle.

**Conclusion**

As it pertains to our company needing to come to a decision on whether or not the data contained within the CARS database shows a strong enough correlation between fuel efficiency and other variables, it is my opinion that this data is highly correlated with a vehicles fuel efficiency. In fact, the R-squared score within the cluster containing mpg highway, mpg city, and weight has an r-squared score of .8843 suggesting that the clusters contained within this group are highly correlated to each other (Haytes, 2020) (SAGE, 2014) as the score is greater than 0.7.

Our Summary Statistics task does a wonderful job of laying out the manufacturer’s vehicle fuel consumption. This study shows that the manufacturers Honda, Mini, and Scion are the manufacturers of vehicles with highway mpg > 30mpg, and from our distribution analysis, we have discovered that vehicles with a highway mpg > 26.8mpg are performing greater than average in terms of fuel economy. From this information, we were able to deduce that perhaps the manufacturers of vehicles with a lighter weight might benefit in terms of fuel efficiency, as less fuel is needed to overcome the inertia of a heavy vehicle. When we test this hypothesis with principal component clustering analysis, we find that the R-squared scores of city mpg and highway mpg are nearly perfectly correlated with scores both greater than 0.9 in their own cluster, and when averaged in with weight, the score of .8843 shows that this clustering seems to be of a good fit, and acceptable to proceed to the next stage of analysis.

References

Hayes, Adam. “R-Squared.” *Investopedia*, Investopedia, 18 Mar. 2020, retrieved from:

[www.investopedia.com/terms/r/r-squared.asp](http://www.investopedia.com/terms/r/r-squared.asp).

SAGE. “R-Squared.” *The SAGE Encyclopedia of Social Science Research Methods*, 2014,

doi:10.4135/9781412950589.n877.

SaNgasoongsong , A, and T.S. Bukkapatnam. “Variable Selection for Multivariate Cointegrated

Time Series Prediction with PROC VARCLUS in SAS® Enterprise Miner™ 7.1.” *SAS Global Forum*, 2012, retrieved from: :<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.648.5869&rep=rep1&type=pdf>.

SAS. (2020). Clustering Variables Task. Retrieved June 22, 2020, from <https://documentation.sas.com/?docsetId=webeditorref>

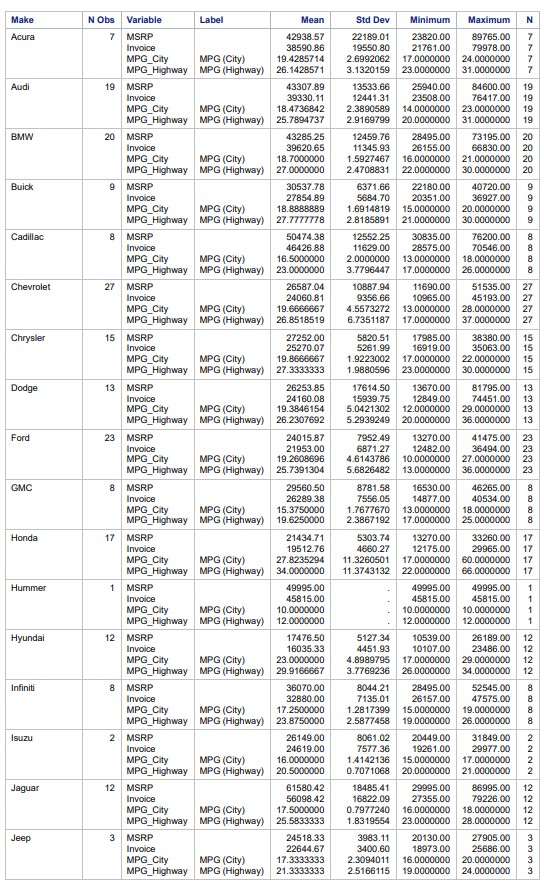
SAS. (2020). Distribution Analysis Task. Retrieved June 14, 2020, from <https://documentation.sas.com/?docsetId=webeditorref&docsetVersion=5.1&docsetTarget=n0uiqaj7iu2zdkn13erg3gl4u83t.htm&locale=en>

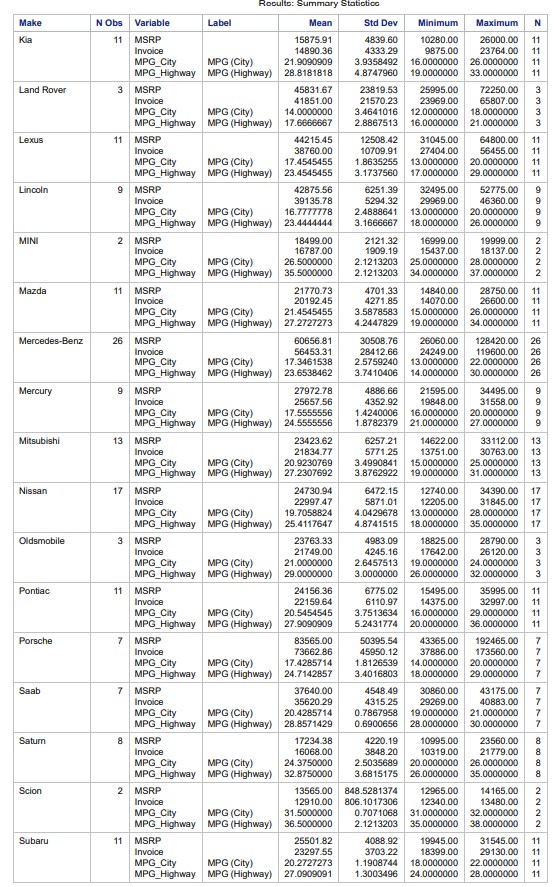
SAS. (2020). Summary Statistics Task. Retrieved June 14, 2020, from <https://documentation.sas.com/?docsetId=webeditorref&docsetVersion=5.1&docsetTarget=p0aheyslqgh8sxn1lfe9pj883r1m.htm&locale=en>

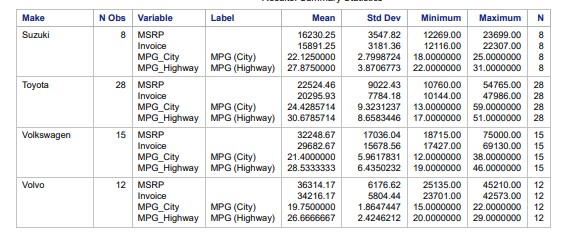
Appendix A

Figure 1

*Summary Statistics*

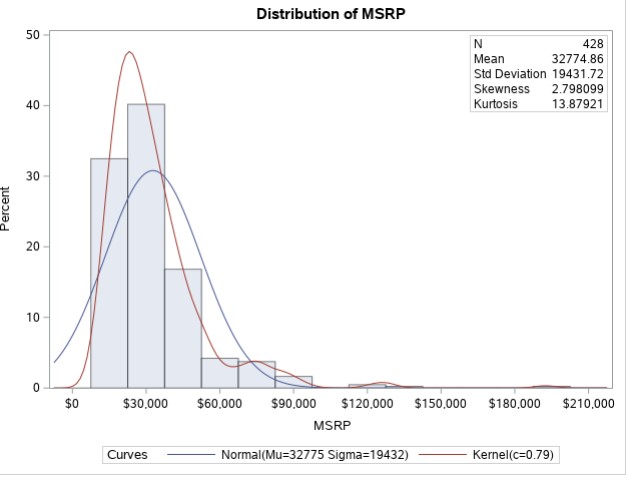


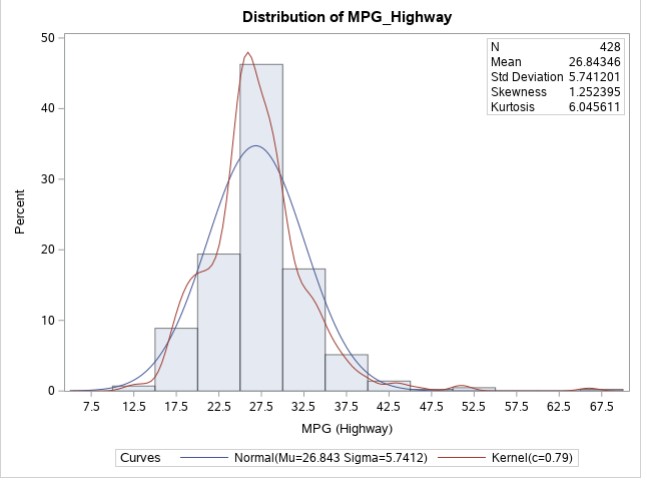
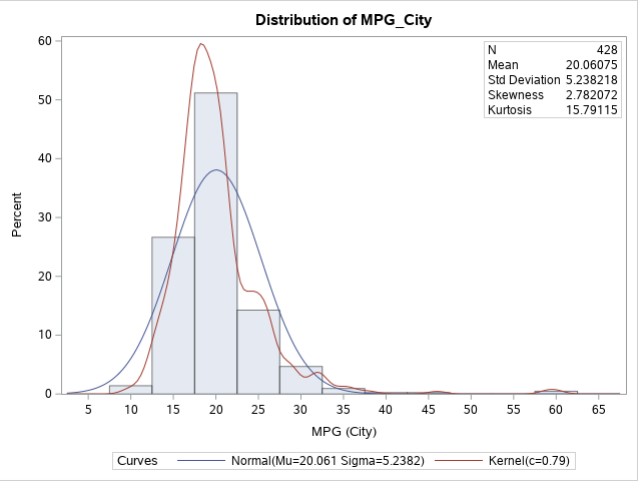
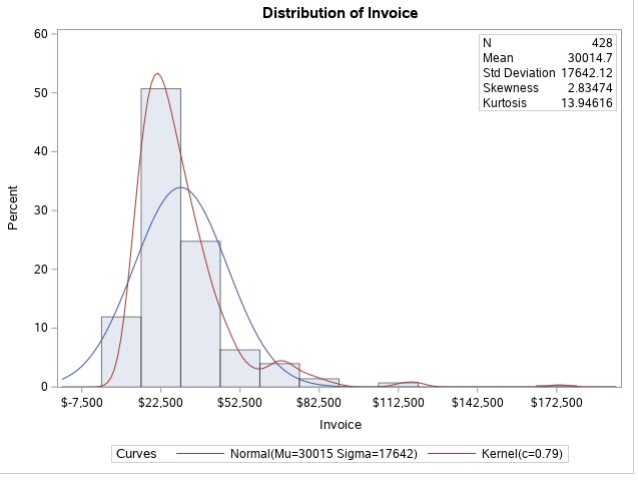




Appendix B

Figure 1

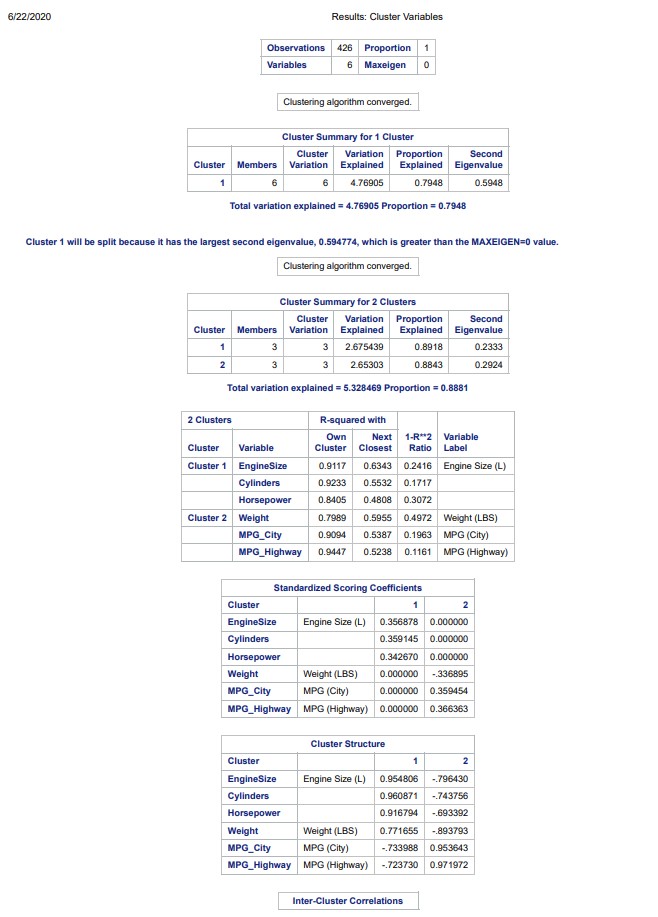
*Distribution Analysis*

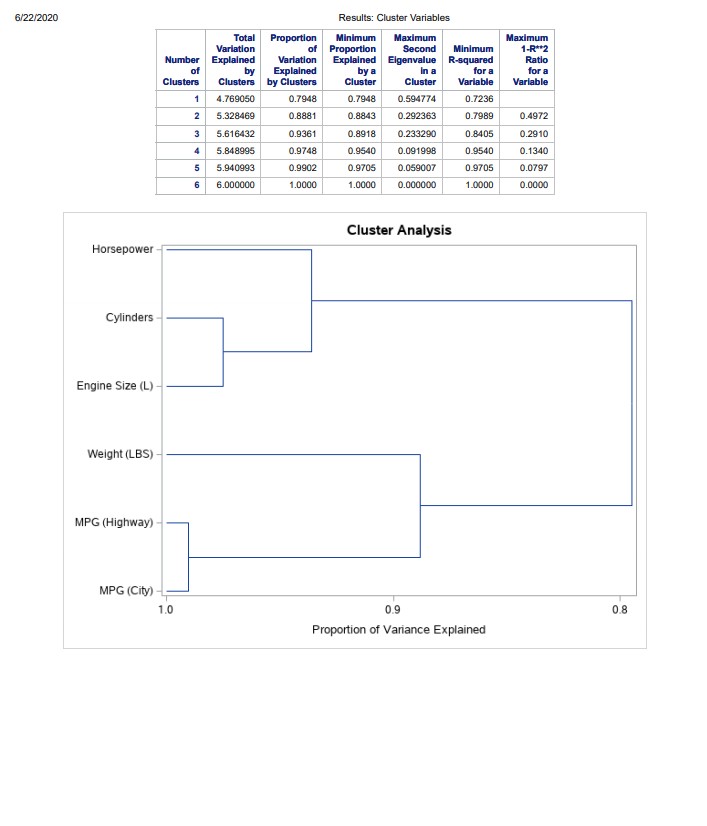
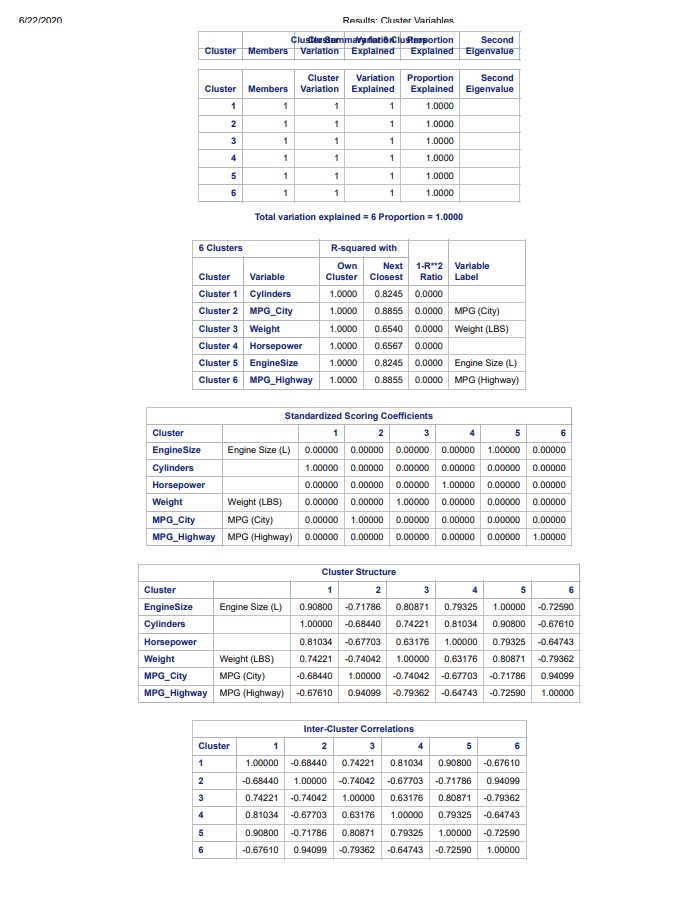
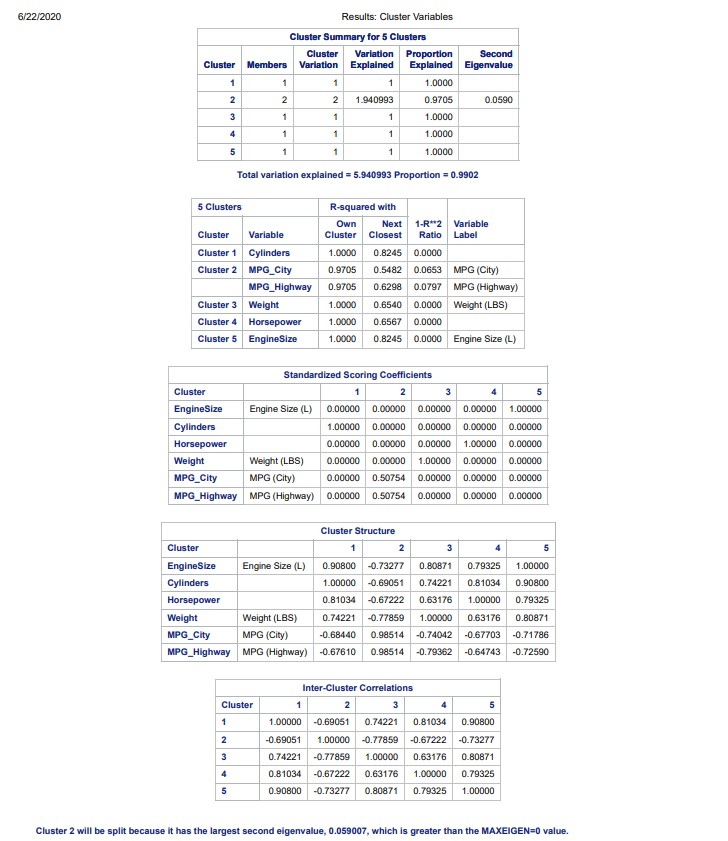
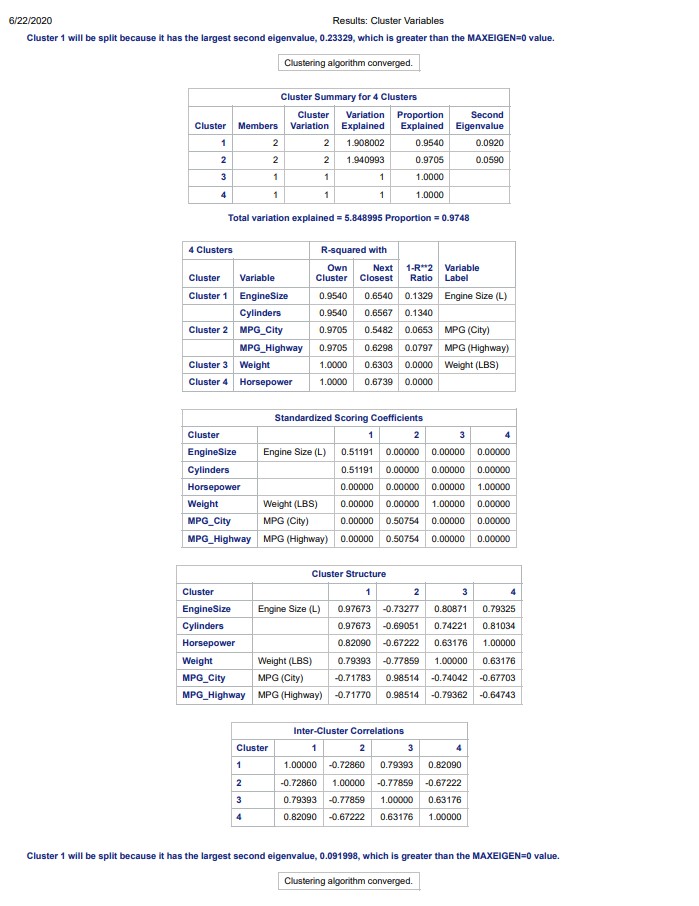
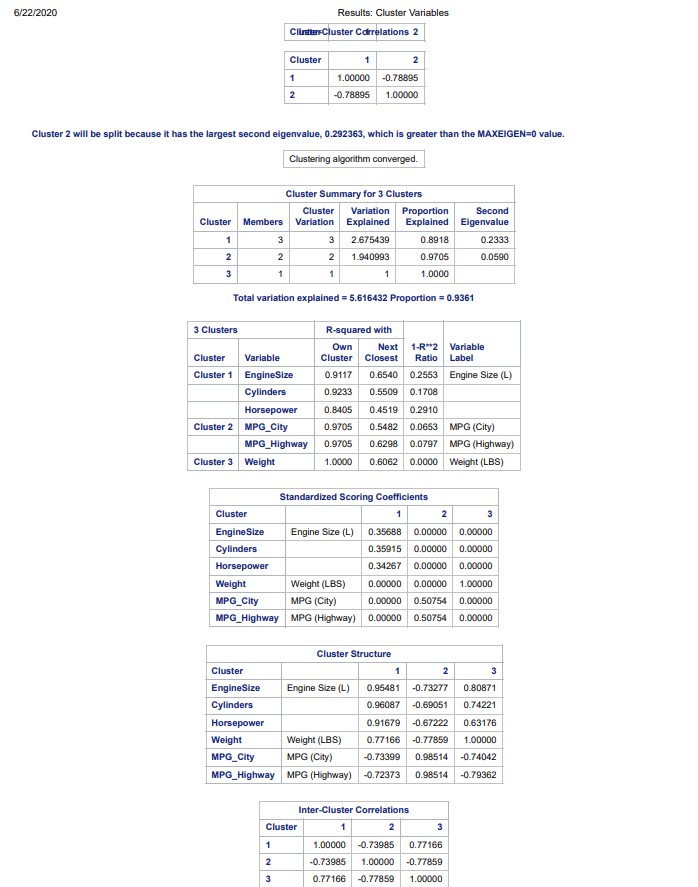
**

Appendix C

Figure 1

*Cluster Variables*

**

**