2.3 Dimension Reduction Methods

Stepwise regression was deemed to be the appropriate method for feature selection for a dataset containing both continuous and categorical variables. Based on the research question in this project, "Price" would be the dependent variable for the stepwise regression. The Akaike Information Criterion (AIC) would be considered as the goodness-of-fit metric in the stepwise regression process.[1] The significant variables on the regression (that were expected to be related to the price variable), coincided also with our clustering analysis.

2.4 Outlier Detection Methods

Mahalanobis distance and density-based clustering were used to detect outliers on the reduced-dimension datasets. Mahalanobis distance considers distance and the variance between each data point. Density-based clustering not only considers the variance, but also considers the number of nearest data points that could form a cluster or not. The performance of these two outlier detection methods were compared in the descriptive analysis section. For the Mahalanobis distance, the significant level of the cut-off point is 0.999. For the density-based clustering, the radius of clustering is 0.1 and the minimum amount of data in each cluster is 20.

2.5 Clustering Methods

The clustering methods considered in this report were hierarchical, K-means and K-medoids. The performance of each method was evaluated using internal evaluation

2.5.1 Training Set

2.5.1.1 Internal Evaluation

Internal evaluation with three metrics was executed to identify the performance of different clustering methods and number of clusters present in the training set. The metrics and clustering methods used in the evaluation were summarized in Table 2. Gower distance was used to calculate the dissimilarity matrix for each clustering method since the data contained mixed data of both numerical and categorical.

[1] Hirotogu Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," *Springer Series in Statistics Selected Papers of Hirotugu Akaike*, 1998, pp. 199-213, https://doi.org/10.1007/978-1-4612-1694-0_15.

Internal Evaluation Metrics on Training Set	**Metric Description**
Dunn Index[2]	Measures ratio between the minimum within-cluster distance and maximum between-cluster distance. Larger values would indicate better clusters as the observations within clusters are more compact and the different clusters are more spread apart.
Connectivity[3]	Measures the closeness of observations with the neighboring observations within the same cluster. Values range from 0 to infinity with smaller values indicating better clusters.
Average Silhouette Width[4]	Measures quality of clusters with value of -1 indicating misclassification and 1 indicating correct classification. Larger values indicate better clusters.

Table 2. Clustering methods and evaluation metrics.

2.5.1.2 External Evaluation

External evaluation using Adjusted Rand Index as the metric was also conducted to evaluate if the optimal number of clusters found for price were the same as those found for the remaining features in the training set.[5] The price was binned using three methods: equal width binning, equal frequency binning and customized binning.

Rand Index has always been used as the metrics in machine learning. However, as Rand Index has no constant value, it makes Rand Index larger than minimum value even the two sets of labels are randomly generated.[6] Due to this disadvantage of Rand Index, the Adjusted Rand Index was used in the part.[7]

[2] J. C. Dunn, “Well-Separated Clusters and Optimal Fuzzy Partitions,” *Journal of Cybernetics* 4, no. 1 (September 1, 1974): pp. 95-104, https://doi.org/10.1080/01969727408546059.

[3] Lukasz Nieweglowski, “Connectivity: Connectivity Index - Internal Measure in Clv: Cluster Validation Techniques,” connectivity: Connectivity Index - Internal Measure in clv: Cluster Validation Techniques, March 17, 2020, https://rdrr.io/cran/clv/man/connectivity.html.

[4] Peter J. Rousseeuw, “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis,” *Journal of Computational and Applied Mathematics* 20 (November 1987): pp. 53-65, https://doi.org/10.1016/0377-0427(87)90125-7.

[5] Dalton, Lori, Virginia Ballarin, and Marcel Brun. “Clustering Algorithms: On Learning, Validation, Performance, and Applications to Genomics.” *Current Genomics* 10, no. 6 (2009): 430–45. https://doi.org/10.2174/138920209789177601.

[6] Ka Yee Yeung and Walter L. Ruzzo, “Details of the Adjusted Rand Index and Clustering Algorithms, Supplement to the Paper an Empirical Study on Principal Component Analysis for Clustering Gene Expression Data.” *Bioinformatics* 17, no. 9 (2001): pp. 763-773.

[7] Jorge M. Santos and Mark Embrechts, “On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification,” *Artificial Neural Networks – ICANN 2009 Lecture Notes in Computer Science*, 2009, pp. 175-184, https://doi.org/10.1007/978-3-642-04277-5_18.

The Adjusted Rand Index was calculated and compared for different optimal number of clusters, as well as clustering and binning methods.[8] With the constant value of 0 present in the Adjusted Rand Index, it provides a more conservative approach when dealing with overestimating the level of agreement due to random labeling.[9] Adjusted Rand Index closer to one would indicate higher agreement of the clusters between the suggested clustering method and the customized binning.[10]

For equal width binning, each width bin was the same. For equal frequency binning method, the count of price in each bin was the same. Customized binning had different width combinations to determine which width would be the most applicable to the data. The range of width was set to be at least 5% quantile and at most [(20 - total number of cluster) * 5%] quantile.

2.5.1.3 Feature Extraction

After performing the clustering analysis, different significant tests were used to extract the important features which can differentiate the cluster effectively. We assume that the population of Cluster 1 and the population of Cluster 2 are two independent groups to perform Two-sample significant test between 2 clusters.[11] The conclusion from the significant tests could indicate either the mean, median, or dependence differences of certain features are significant. If that specific feature has a significant difference between of the population of Cluster 1 and 2, it also implies that this specific feature is the important feature to classify the used car into cluster 1 or 2.

2.5.1.3.1 Continuous Data

Two-Stage significant test procedure was used for continuous data.[12] The first stage used the Shapiro-Wilk test to test the normality of each cluster.[13][14] Depending on the result of the Shapiro-Wilk test, different significant tests would be used in continuous data.

If the continuous data in all clusters are normally distributed, two-sample Student's t-test was implemented to evaluate whether the cluster mean differences are significant. If the continuous data in any cluster is not normally distributed, Wilcoxon Rank Sum Test was used to evaluate whether the cluster median differences are significant.[15]

[8] Yeung and Ruzzo, "Details of the Adjusted Rand Index and Clustering Algorithms."

[9] Yeung and Ruzzo, "Details of the Adjusted Rand Index and Clustering Algorithms."

[10] Yeung and Ruzzo, "Details of the Adjusted Rand Index and Clustering Algorithms."

[11] S. Galbraith, J. A. Daniel, and B. Vissel, "A Study of Clustered Data and Approaches to Its Analysis," *Journal of Neuroscience* 30, no. 32 (November 2010): pp. 10601-10608, https://doi.org/10.1523/jneurosci.0362-10.2010.

[12] Justine Rochon, Matthias Gondan, and Meinhard Kieser , "To Test or Not to Test: Preliminary Assessment of Normality When Comparing Two Independent Samples," BMC Medical Research Methodology 12, no. 81 (2012), https://doi.org/10.1186/1471-2288-12-81.

[13] S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika* 52, no. 3/4 (1965): pp. 591-611, https://doi.org/10.2307/2333709.

[14] Asghar Ghasemi and Saleh Zahediasl, "Normality Tests for Statistical Analysis: A Guide for Non-Statisticians," *International Journal of Endocrinology and Metabolism* 10, no. 2 (January 2012): pp. 486-489, https://doi.org/10.5812/ijem.3505.

[15] Graeme D. Ruxton, "The Unequal Variance t-Test Is an Underused Alternative to Student's t-Test and the Mann–Whitney U Tvest," Behavioral Ecology 17, no. 4 (2006): pp. 688-690, https://doi.org/10.1093/beheco/ark016.

2.5.1.3.2 Categorical Data

If all expected frequency for each level in each cluster is larger than five, Chi-Square Test of Independence was implemented to evaluate the dependence between that categorical data and cluster.[16][17] If either one expected frequency is smaller than five, Fisher Exact Test was used.[18]

2.5.2 Testing Set

The testing set was the subsample of was divided into three parts in which external evaluation and feature extraction were performed for each set.[19] The result obtained would show the consistency and performance of the clustering methods obtained from the training set when applied on the three testing sets.

2.5.2.1 External Evaluation

In each testing data set, the same iteration process in the training set would be performed to find out the optimal custom binning width with the largest Adjusted Rand Index. If the test sets have the same custom binning width as the train set, it implies that the price binning criteria in the train set is not random or coincident case. This also indicates stable implication.

2.5.2.2 Feature Extraction

Same set of significant test procedures would also be performed to see whether the clusters from the testing set also have the same result from the training set. If the same features exist, we could conclude that either the cluster mean, median difference or dependence are not the random or coincident cases.3. Descriptive Analysis

Prior to conducting cluster analysis, data was split into training and testing sets. Training set was plotted to understand the structure, distribution, sparsity, and trends of all variables that were of interest for used cars with year models between 2018 and 2020. After which, dimension reduction using stepwise regression was conducted on the training set to identify variables significant in determining price of used cars. Outliers and noise were then detected in the reduced training set to assess the appropriate clustering method for the data. The remaining analysis in this report used the reduced training set.

[16] Mary L. Mchugh, "The Chi-Square Test of Independence," *Biochemia Medica* 23, no. 2 (January 15, 2013): pp. 143-149, https://doi.org/10.11613/bm.2013.018.

[17] Eleonore ten Thij and Justin de Nooijer, "A Framework for Exploring Relationships Between Online Community Characteristics and Regulation Principles," SAW, 2007.

[18] Thij Nooijer, "A Framework for Exploring Relationships"

[19] Dalton, Lori, Virginia Ballarin, and Marcel Brun. "Clustering Algorithms: On Learning, Validation, Performance, and Applications to Genomics." *Current Genomics* 10, no. 6 (2009): 430–45. https://doi.org/10.2174/138920209789177601.

3. Descriptive Analysis

3.1 Dimension Reduction

Stepwise regression was utilized to get the most significant features in relation to used car prices from the initial train set containing 2,877 observations and 68 variables. The output indicated 46 features that were significant to used car prices. The reduced training and testing set each has 2,877 observations and 46 variables to perform the remaining analysis in this report. The results from stepwise regression are summarized in Appendix D, Table D1.

3.2 Distribution

The reduced training set contained continuous, discrete, binary and nominal data. For a more detailed breakdown of each variable, see Table A2 in Appendix A. The distribution of variables is shown in Appendix B.

The distribution for the quantitative data in the training set were either skewed or multimodal distributions. There was no significant difference in the distribution for each year except for mileage and price. Large portion of the mileage distribution in 2020 indicated lesser miles travelled compared to other years, which was expected for newer used cars.

Consistently across all years, SUV/Crossover dominated body types in the West Coast, accounting for approximately a total of 47% of body types between 2018 and 2020 car models. The second most common body type was sedan which made up approximately 31%. Engine cylinders of type I4 accounted for about 56%, followed by 22% for V6. Gasoline accounts for approximately 90% of fuel type used in the West Coast with other fuel types accounting less than 6% throughout the three years. A few of the popular colors were white, black, silver, and gray accounting for 14% to 24% of the total used cars in the data. 81% of transmission types between 2018 and 2020 car models were type A. Forward-drive (FWD) wheel system was the most common, accounting for 47% of data. Used cars listed in California accounted for around 65% of the data. Overall, the distribution of nominal data across each year stayed the same; there was no drastic change in proportions observed except for car brands. There were slight variations throughout the years for all brands.

3.3 Noise and outliers

After the dimension reduction, the training set had 16 columns with continuous data and 30 columns with categorical data. Mahalanobis distance and Density-based clustering was used to detect outliers in the reduced-dimension dataset. Table 3 summarizes the number of outliers for each method and their respective cutoff points.

Number of Mahalanobis Outliers	172 Cutoff point = 39.25
Number of Density-based Cluster Outliers	141 Min. point: 20 | Radius: 0.1

Table 3. Outliers detected from Mahalanobis distance and Density-based clustering.

The number of outliers by using Mahalanobis distance method is larger than using Density-based clustering method. Below are two possible reasons:

1. Mahalanobis distance only considers continuous data, but Density-based clustering considers continuous and categorical data at the same time. Hence, the presence of categorical data in the latter method would reduce the distance between each data point.

2. Density-based clustering considers the number its of nearest neighbors. If the data point is far from the data cloud but there are 20 data points in proximity, that data point is not an outlier. Some outliers from Mahalanobis Distance method could form a small cluster which explains the smaller number of outliers from Density-based clustering compared to the Mahalanobis distance method.

4. Clustering Analysis

Clustering analysis was first conducted on the training set to develop the optimal number of clusters and clustering methods that would then be applied to the test set. The test set was further split into three equal parts, Tests 1 to 3, each having 959 observations and 46 variables. These variables were from the output of stepwise regression.

4.1 Training Set

4.1.1 Internal Evaluation

Originally, the result from internal evaluation suggested hierarchical clustering with two clusters by using single and average linkages. Using these linkages resulted in one of the clusters having only one record. This result from the internal evaluation was not acceptable and was not considered. Based on Table 4, each index suggested a different clustering method and number of clusters. External evaluation was utilized to finalize the optimal clustering method.

	Clustering Method	Number of Clusters	Score
Dunn Index	Hierarchical - Ward.D2	5	0.058
Connectivity	Hierarchical - Ward.D2	2	140.251
Silhouette Width	K-means	2	0.221

Table 4. Summary of the best clustering methods, number of clusters and their scores obtained from internal evaluation for each metric.

4.1.2 External Evaluation

	Hierarchical - Ward.D2 (5 Clusters)	Hierarchical - Ward.D2 (2 Clusters)	K-means (2 Clusters)
Equal Frequency	0.1534	0.3347	0.3355
Equal Width	0.0396	0.0004	0.0010
Customized Binning	0.2886 (50%, 10%, 15%, 5%, 10%)	0.3672 (40%, 60%)	0.3355 (50%, 50%)

Table 5. Adjusted Rand Index between different clustering methods and different binning criteria.

Based on Table 5, the hierarchical clustering method with Ward.D2 linkage and two clusters has the highest adjusted Rand Index (0.3672) with customized binning. The width of the bin for clusters 1 and 2 are 40% quantile and 60% quantile, respectively. This implied that used car price in cluster 1 is lower or equal to 40% quantile, while cluster 2 is higher than 40% quantile.

	Hierarchical - Ward.D2 Cluster 1	Hierarchical - Ward.D2 Cluster 2	Total
Price Binning Cluster 1	845	306	1151
Price Binning Cluster 2	259	1467	1726
Total	1104	1773	2877

Table 6. Confusion matrix between two clusters using hierarchical clustering method with Ward.D2 linkage and customized binning.

Based on Table 6, the accuracy rate is 80.36%. The overall performance of the clusters is meaningful and reasonably good.

4.1.3 Features

Significance tests were conducted to check whether the features between the two clusters were significantly different from each other in terms of their means or medians.

The following categorical features pass the significance test for the training dataset. The list is separated in two parts: the first part is the most predominant in cluster 1 - cars with a relative economical price; the features marked with red text represent the predominant features in cluster 2 - the higher end vehicles.

Accessory Specifications: Alloy.Wheels, Blind.Spot.Monitoring, Heated.Seats, Navigation.System,
Premium.Package, Quick.Order.Package, Remote.Start, Sunroof.Moonroof.
Body Specifications: Third.Row.Seating, back_legroom, height, length, width.
Engine Specifications: engine_cylinders_I4, fuel_type_Gasoline, city_fuel_economy,
engine_cylinders_V6, engine_displacement, fuel_tank_volume, horsepower, torque, power.
Transmission Specifications: wheel_system_FWD, transmission_CVT, transmission_A,
wheel_system_4WD, wheel_system_AWD, wheel_system_RWD.
Marketing Information: seller_rating.
Manufacturer: make_name_Honda, make_name_Nissan, make_name_Toyota,
make_name_Jeep, make_name_Ford.
Vehicle Usage: mileage.
Color and Body Type: listing_color_SILVER, body_type_Sedan, body_type_Pickup_Truck.
Misc.: franchise_dealer_True.

4.2 Testing Set

The price variable in the three test sets were also binned using a customized binning method. Different binning widths were evaluated for different clustering methods and number of clusters for each test set. The combination of bin widths obtained from the three test sets will be used to form the cluster boundary.

4.2.1 External Evaluation

Based on Table 7, the bin width for hierarchical with two clusters differs for different test sets. Test 1, 2 and 3 have 50th, 35th and 60th percentile, respectively, as the bin points for cluster 1. In Test 1, used car price in cluster 1 and 2 were divided equally, each containing 50% quantile. In Test 2, used car price in cluster 1 is lower or equal to 35% quantile, while cluster 2 is higher than 35% quantile. In Test 3, used car price in cluster 1 is lower or equal to 60% quantile, while cluster 2 is higher than 60% quantile.

	Test 1	Test 2	Test 3
Hierarchical - Ward.D2 (2 Clusters)	0.2384 (50%, 50%)	0.40775 (35%, 65%)	0.2553 (60%, 40%)

Table 7. Summary of adjusted Rand Index values for three test sets.

4.2.2 Feature Extraction

After collecting the significant factors, these were grouped into the following specifications groups: accessory, vehicle body, engine, transmission, marketing information, manufacturer, vehicle usage, color and body type, and miscellaneous variables.

Our training dataset will serve as a baseline to compare and check if the same significant features remain on three other testing datasets. We did this comparison to see if our used car dataset, along with our sampling and subsampling methods would be able to reproduce consistent clustering results. The training dataset gives more details into what sort of features are predominant in cluster 1 and cluster 2.

4.2.2.1 Test Datasets

The following test datasets were compared to the training dataset to observe if the significant features remain constant through different samples. As it is shown below, they are consistent in some categories such as engine specifications, transmission specifications, and for the most part in vehicle body specifications, and color and body type. In these sections, we saw that most of their proportion values and ranges, as well as the medians between cluster 1 and 2 are very similar through the three test and the training sets. This is an indication that our sampling and subsampling methodology can provide consistent results in both feature extraction and clustering results. In categories such as accessory specifications, marketing information, and vehicle usage a larger sample might help improve the consistency among all latent categories among the significant variables.

In the three testing sets, the following variables remained significant compared to the training dataset:

Testing Dataset #1

Accessory Specifications: Navigation.System, Quick.Order.Package, Remote.Start.
Vehicle Body Specifications: Third.Row.Seating, back_legroom, height, length, width.
Engine Specifications: engine_cylinders_I4, fuel_type_Gasoline, city_fuel_economy, engine_cylinders_V6, engine_displacement, fuel_tank_volume, horsepower, torque, power.
Transmission Specifications: transmission_CVT, wheel_system_AWD, wheel_system_FWD, transmission_A, wheel_system_4WD, wheel_system_RWD.
Marketing Information: None.
Manufacturer: make_name_Ford, make_name_Honda, make_name_Nissan.
Color and Body Type: body_type_Pickup_Truck, body_type_Sedan.
Vehicle Usage: None. - **Misc.:** None.

Testing Dataset #2

Accessory Specifications: Alloy.Wheels, Heated.Seats, Navigation.System, Premium.Package, Quick.Order.Package, Remote.Start, Sunroof.Moonroof.
Vehicle Body Specifications: Third.Row.Seating, back_legroom, height, length, width.
Engine Specifications: engine_cylinders_I4, fuel_type_Gasoline, city_fuel_economy, engine_cylinders_V6, engine_displacement, fuel_tank_volume, horsepower, torque, power.
Transmission Specifications: transmission_CVT, wheel_system_FWD, transmission_A, wheel_system_4WD, wheel_system_AWD, wheel_system_RWD.
Manufacturer: make_name_Honda, make_name_Ford, make_name_Jeep, make_name_Nissan, make_name_Toyota.
Vehicle Usage: mileage.
Color and Body Type: body_type_Sedan, body_type_Pickup_Truck.
Marketing Information: None. - **Misc.:** None.

Testing Dataset #3

Accessory Specifications: Heated.Seats, Navigation.System, Quick.Order.Package, Remote.Start.
Vehicle Body Specifications: Third.Row.Seating, back_legroom, height, length, width.
Engine Specifications: engine_cylinders_I4, fuel_type_Gasoline, city_fuel_economy, engine_cylinders_V6, engine_displacement, fuel_tank_volume, horsepower, torque, power.
Transmission Specifications: transmission_CVT, wheel_system_FWD, wheel_system_RWD, transmission_A.
Manufacturer: make_name_Nissan, make_name_Jeep.
Color and Body Type: body_type_Sedan, body_type_Pickup_Truck.
Marketing Information: None. - **Vehicle Usage:** None. - **Misc.:** None.

5. Discussion

Based on the results obtained from the training set, a hierarchical clustering method with Ward.D2 linkage and 2 clusters was selected. It is expected that this clustering method could determine whether the price of used cars is higher or lower than 40% quantile by using other used cars' features.

5.1 External evaluation

Based on the result of external evaluation in the training and test sets, it is concluded that using used cars' features with Ward.D2 linkage hierarchical clustering method could cluster the car into two clusters, one having lower price and the other having higher price. The range of binning points that define the boundary between clusters 1 and 2 is from 35% quantile to 60% quantile.

In other words, cluster 2 would have quantiles greater than 35% and cluster 1 would have quantiles smaller than 60%. Hence, if a new data point sampled from the current data is classified into cluster 1, its price should not be higher than 60% quantile of all used car. On the other hand, if a new data point is classified into cluster 2, its price should not be lower than 35% quantile of all used car.

From the above observation, it is concluded that equal width or equal frequency binning methods do not work well for the suggested clustering.

5.2 Important feature

As shown in Table E1 in Appendix E, some feature groups are not stable across the three test samples. The criteria for judging feature instability are the following:

1. If a feature is not consistently classified in the same cluster across all three testing datasets, then it is deemed unstable, and/or
2. If a feature only appears in ⅓ of the testing sets, then it is deemed unstable.

One such example is in the category of Transmission Specifications: the "wheel_system_AWD" feature has an opposite cluster proportion compared to the training and Test 2. Another example is in the Vehicle Body Type Specification where the "front_legroom" appears only in Test 1. In the first example, the feature's cluster classification is not consistent with all the other datasets. In the second example, the feature only appears once. These features are counted as unstable.

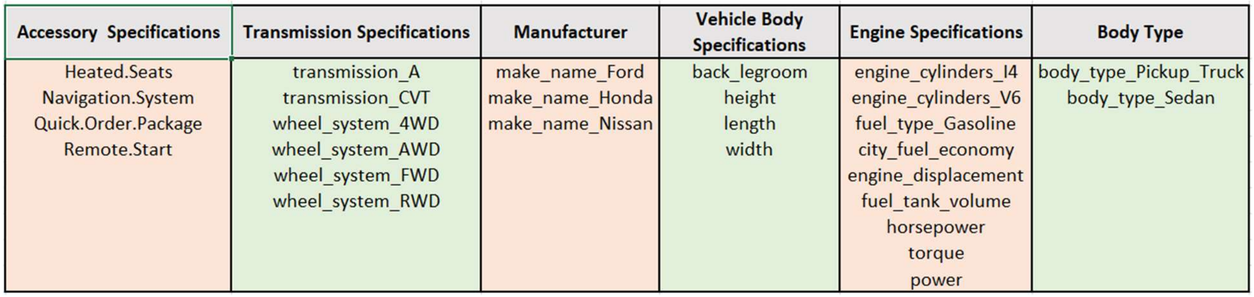
On the opposite end, an example of consistent features, is in the Engine Specification group. This group includes features such as engine cylinder and fuel types, fuel tank volume, and engine power. This group's stability is demonstrated by correctly classifying bigger and more powerful engines in cluster 2 and including all the group's features across the three testing sets. These features are regarded as very stable.

6. Suggestions

A way to improve the feature extraction and external evaluation steps is to increase the number of subsampling. Apart from subsampling, using resampling with replacement is also an option, which could be have a better performance than subsampling.[20] For feature extraction, this would prove if the features that appeared only once in the three testing samples are there because they are truly significant, or because it was by chance (e.g., coincidental sampling). For external evaluation, this improvement would result in many optimal binning points for different sets, producing a distribution of the binning point. From this distribution, we could identify which binning point occurred most often and implement a narrower range of binning points.

Another method to improve external evaluation is to reduce the range of binning points obtained. This can be achieved by reducing the minimum quantile from 0.05, as used in this study, to 0.01. This might produce binning points that are more compact.

7. Conclusion



Accessory Specifications	Transmission Specifications	Manufacturer	Vehicle Body Specifications	Engine Specifications	Body Type
Heated.Seats	transmission_A	make_name_Ford	back_legroom	engine_cylinders_I4	body_type_Pickup_Truck
Navigation.System	transmission_CVT	make_name_Honda	height	engine_cylinders_V6	body_type_Sedan
Quick.Order.Package	wheel_system_4WD	make_name_Nissan	length	fuel_type_Gasoline	
Remote.Start	wheel_system_AWD		width	city_fuel_economy	
	wheel_system_FWD			engine_displacement	
	wheel_system_RWD			fuel_tank_volume	
				horsepower	
				torque	
				power	

Table 8. Summary of significant and stable features.

Overall, we can conclude that the used car grouped into cluster 1 should have price lower than or equal to 60th percentiles of the used car population, and higher than 35th percentiles for cluster 2. The features in the above table could be used as the reference points to decide whether the price of the used car is in the higher or lower price cluster.

Generally, researchers perform multiple resampling with replacement to evaluate the overall significance of the clustering performance or use multiple subsamples to evaluate the clustering method efficiency. However, in practical prospect, the clustering method should also provide clusters' implication and the level of significance of the car's features in the clusters.

This project subsampled the testing data set three times and identified the best binning criteria on the interested variable to give a higher level of generalization power on the clusters' implication. Important features of the suggested clustering method were also identified.

[20] Hans-Joachim Mucha and Hans-Georg Bartel, "Resampling Techniques in Cluster Analysis: Is Subsampling Better Than Bootstrapping?," *Data Science, Learning by Latent Structures, and Knowledge Discovery*, 2015, pp. 113-122, https://doi.org/10.1007/978-3-662-44983-7_10.

Reference

Akaike, Hirotogu. “Information Theory and an Extension of the Maximum Likelihood Principle.” *Springer Series in Statistics Selected Papers of Hirotugu Akaike*, 1998, 199–213. https://doi.org/10.1007/978-1-4612-1694-0_15.

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. “NbClust: AnRPackage for Determining the Relevant Number of Clusters in a Data Set.” *Journal of Statistical Software* 61, no. 6 (October 2014). https://doi.org/10.18637/jss.v061.i06.

Dalton, Lori, Virginia Ballarin, and Marcel Brun. “Clustering Algorithms: On Learning, Validation, Performance, and Applications to Genomics.” *Current Genomics* 10, no. 6 (2009): 430–45. https://doi.org/10.2174/138920209789177601.

Datta, Somnath, and Glen A Satten. “Rank-Sum Tests for Clustered Data.” *Journal of the American Statistical Association* 100, no. 471 (2005): 908–15. https://doi.org/10.1198/016214504000001583.

Dunn, J. C. “Well-Separated Clusters and Optimal Fuzzy Partitions.” *Journal of Cybernetics* 4, no. 1 (September 1, 1974): 95–104. https://doi.org/10.1080/01969727408546059.

Galbraith, S., J. A. Daniel, and B. Vissel. “A Study of Clustered Data and Approaches to Its Analysis.” *Journal of Neuroscience* 30, no. 32 (2010): 10601–8. https://doi.org/10.1523/jneurosci.0362-10.2010.

Ghasemi, Asghar, and Saleh Zahediasl. “Normality Tests for Statistical Analysis: A Guide for Non-Statisticians.” *International Journal of Endocrinology and Metabolism* 10, no. 2 (2012): 486–89. https://doi.org/10.5812/ijem.3505.

Kerr, M. K., and G. A. Churchill. “Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments.” *Proceedings of the National Academy of Sciences* 98, no. 16 (2001): 8961–65. https://doi.org/10.1073/pnas.161273698.

Mchugh, Mary L. “The Chi-Square Test of Independence.” *Biochemia Medica* 23, no. 2 (January 15, 2013): 143–49. https://doi.org/10.11613/bm.2013.018.

Mucha, Hans-Joachim, and Hans-Georg Bartel. “Resampling Techniques in Cluster Analysis: Is Subsampling Better Than Bootstrapping?” *Data Science, Learning by Latent Structures, and Knowledge Discovery*, 2015, 113–22. https://doi.org/10.1007/978-3-662-44983-7_10.

Nemec, A. F. L., and R. O. Brinkhurst. “Using the Bootstrap to Assess Statistical Significance in the Cluster Analysis of Species Abundance Data.” *Canadian Journal of Fisheries and Aquatic Sciences* 45, no. 6 (June 1, 1988): 965–70. https://doi.org/10.1139/f88-118.

Nieweglowski, Lukasz. connectivity: Connectivity Index - Internal Measure in clv: Cluster Validation Techniques, March 17, 2020. https://rdrr.io/cran/clv/man/connectivity.html.

Rochon, Justine, Matthias Gondan, and Meinhard Kieser . "To Test or Not to Test: Preliminary Assessment of Normality When Comparing Two Independent Samples." *BMC Medical Research Methodology* 12, no. 81 (2012). https://doi.org/10.1186/1471-2288-12-81.

Rocke, David M., and Jian Dai. "Sampling and Subsampling for Cluster Analysis in Data Mining, with Applications to Sky Survey Data." *Data Mining and Knowledge Discovery*, April 2003. https://doi.org/10.1023/A:1022497517599.

Rousseeuw, Peter J. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20 (November 1987): 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Ruxton, Graeme D. "The Unequal Variance t-Test Is an Underused Alternative to Student's t-Test and the Mann–Whitney U Test." *Behavioral Ecology* 17, no. 4 (2006): 688–90. https://doi.org/10.1093/beheco/ark016.

Santos, Jorge M., and Mark Embrechts. "On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification." *Artificial Neural Networks – ICANN 2009 Lecture Notes in Computer Science*, 2009, 175–84. https://doi.org/10.1007/978-3-642-04277-5_18.

Shapiro, S. S., and M. B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52, no. 3/4 (1965): 591–611. https://doi.org/10.2307/2333709.

Thij, Eleonore ten, and Justin de Nooijer. "A Framework for Exploring Relationships Between Online Community Characteristics and Regulation Principles." *SAW*, 2007.

Yeung, Ka Yee, and Walter L. Ruzzo. "Details of the Adjusted Rand Index and Clustering Algorithms, Supplement to the Paper an Empirical Study on Principal Component Analysis for Clustering Gene Expression Data." *Bioinformatics* 17, no. 9 (2001): 763–73.

Appendix A: Data Sampling and Cleaning

Variables that were deemed to be insignificant in our hypotheses were excluded, along with cities and states that are not part of the West Coast. The West Coast region consisted of states Alaska, California, Oregon and Washington. As the proportion of each state varied in the dataset, stratified sampling method was used to prevent underrepresentation of states. A total of 6,000 observations were sampled with each year having 2,000 observations.

After sampling, data cleaning was done to check accuracy and presence of missing data. Some variables such as back_legroom had values and units recorded together. Variable major_options also had multiple levels of variables grouped together. These were separated and the data type was converted appropriately. The missing data was Missing At Random (MAR) as no specific patterns were observed. Prior to multiple imputations using classification and regression trees, observations with greater than 5% missing data was excluded.

Dummy variables for nine variables were then created after imputations. These variables were namely body_type, engine_cylinder, fuel_type, listing_color, major_options, make_name, transmission, wheel_system and state. As some of these variables had 2 to 111 levels, the number of columns expanded from 34 to 236. Tables A1 and A2 outline these variables and their levels, respectively. After imputations, additional five variables, city, listed_date, latitude, longitude and franchise_dealer_FALSE were removed. These variables were not significant for the cluster analysis but were initially kept to generate better imputations on missing data. The final dataset was then split into training and testing sets, each having 2,877 observations and 231 variables. Variables containing less than 5% of ones were removed after observing the distribution of the training set. With the removal of 163 categorical variables, the final dataset used for dimension reduction analysis contained 2,877 observations and 68 variables.

Variables	**Descriptions**	**Variables Used**	**Data Type**
vin	Vehicle identification number	No	Nominal
back_legroom	Space for leg for backseat (inch)	Yes	Continuous
bed	Size of bed for pickup trucks	No	Ordinal
bed_height	Height of pickup truck bed (inch)	No	NA (no values in data)
bed_length	Length of pickup truck bed (inch)	No	Continuous
body_type	Type of vehicles	Yes	Nominal
cabin	Type of cabin	No	Nominal
city	Cities in USA	No (removed after imputations)	Nominal
city_fuel_economy	Fuel consumption driving in city	Yes	Discrete

combine_fuel_economy	Fuel consumption when driving in city and highway	No	NA (no values in data)
daysonmarket	Number of days the car was on the market	Yes	Discrete
dealer_zip	Dealer’s zip code	No	Nominal
description	Description of car	No	Text
engine_cylinder	Number and type of cylinders	Yes	Nominal
engine_displacement	Engine displacement	Yes	Discrete
engine_type	Type of engine	No	Nominal
exterior_color	Exterior color of car	No	Nominal
fleet	Was it a fleet vehicle (rental/corporate) before?	No	Binary
frame_damaged	Was the frame damaged?	No	Binary
franchise_dealer	Was the car in a franchise dealership?	Yes	Binary
franchise_make	Franchise manufacturer/brand	No	Nominal
front_legroom	Space for leg for front seat (inch)	Yes	Continuous
fuel_tank_volume	Volume of fuel tank (gal)	Yes	Continuous
fuel_type	Types of fuel	Yes	Nominal
has_accidents	Has the car been into accidents?	No	Binary
height	Height of the car (inch)	Yes	Continuous
highway_fuel_economy	Fuel consumption when driving on highway	Yes	Discrete

horsepower	Horsepower	Yes	Discrete
interior_color	Interior color of car	No	Nominal
isCab	Was the car a cab?	No	Binary
is_certified	Was the car certified?	No	Binary
is_cpo	Was the car certified pre-owned?	No	Binary
is_new	Was the car new?	No (removed after sampling)	Binary
is_oemcpo	Was the car OEM certified pre-owned?	No	Binary
latitude	Latitude (North-South location)	No (removed after imputations)	Continuous
length	Length of car (inch)	Yes	Continuous
listed_date	Date when car was listed for sale	No (removed after imputations)	Nominal
listing_color	Color of car listed	Yes	Nominal
listing_id	Id of car listing	No	Nominal
longitude	Longitude (East-West location)	No (removed after imputations)	Continuous
main_picture_url	Link to the car picture	No	Text
major_options	Additional features of car or packages available	Yes	Nominal
make_name	Car manufacturer/brand name	Yes	Nominal
maximum_seating	Maximum number of seats in the car	Yes	Discrete
mileage	Number of miles traveled	Yes	Discrete
model_name	Car model	No	Nominal

owner_count	Number of owners previously owned the car	No	Discrete
power	Power of car	Yes	Discrete
price	Price of car	Yes	Discrete
salvage	Can the car be salvaged?	No	Binary
savings_amount	Amounts saved	No	Discrete
seller_rating	Seller rating	Yes	Continuous
sp_id	ID of salesperson	No	Nominal
sp_name	Name of salesperson	No	Nominal
theft_title	Any theft history on the car?	No	Binary
torque	Torque	Yes	Discrete
transmission	Transmission type	Yes	Nominal
transmission_display	(same as transmission)	No	Nominal
trimId	ID for trim_name	No	Nominal
trim_name	Shortened description of car used as label names	No	Nominal
vehicle_damage_category	NA (no values in data)	No	NA (no values in data)
wheel_system	Wheel system/mechanism of car	Yes	Nominal
wheel_system_display	(same as wheel_system)	No	Nominal
wheelbase	Horizontal distance between the centers of front and rear wheels (inch).	Yes	Continuous
width	Width of the car (inch)	Yes	Continuous

year	Year of the car model	Yes	Nominal
state	State of the cities	Yes (was not present in original dataset and added to aid in the analysis)	Nominal

Table A1. List of variables and their descriptions

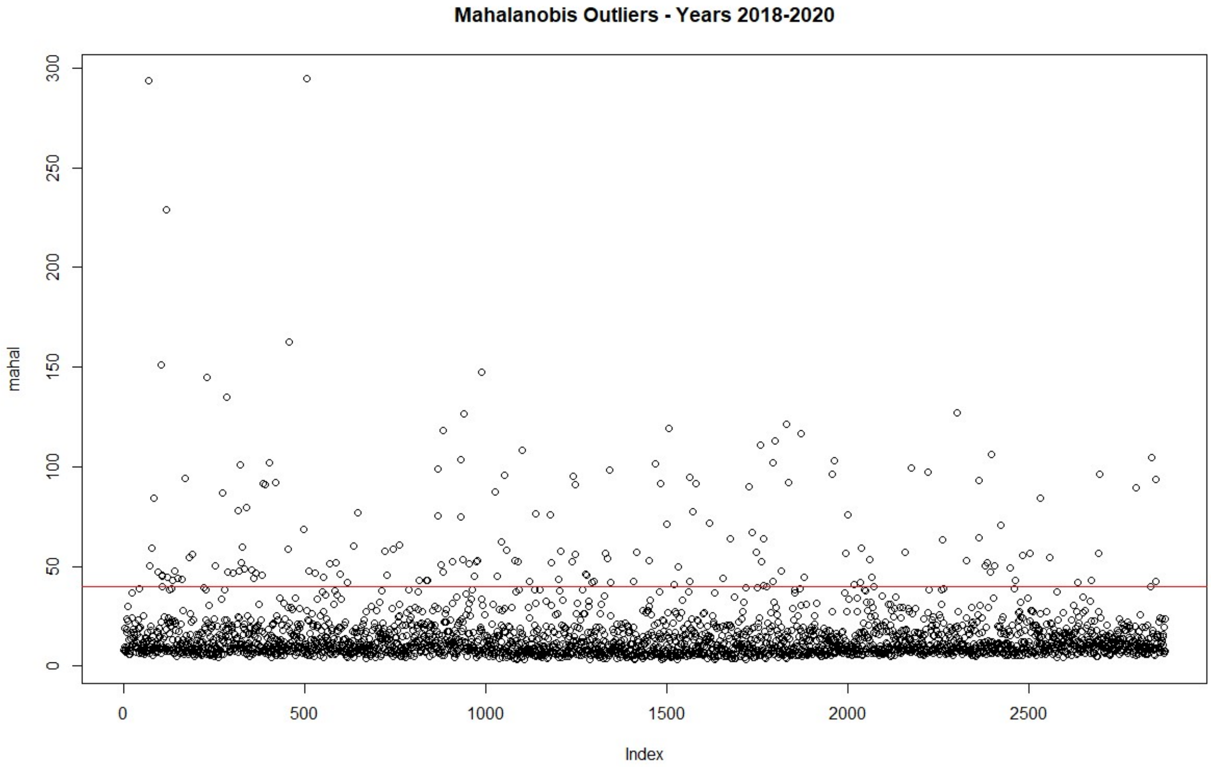
Variables	**Levels**
body_type (9 levels)	Convertible, Coupe, Hatchback, Minivan, Pickup_Truck, Sedan, SUV/Crossover, Van, Wagon
engine_cylinder (21 levels)	H4, H6, I2, I3, I4, I4 Diesel, I4 Flex Fuel Vehicle, I4 Hybrid, I5 Biodiesel, I6, I6 Diesel, V10, V12, V6, V6 Biodiesel, V6 Diesel, V6 Flex Fuel Vehicle, V6 Hybrid, V8, V8 Biodiesel, V8 Flex Fuel Vehicle
franchise_dealer (2 levels)	True, False
fuel_type (5 levels)	Biodiesel, Diesel, Flex Fuel Vehicle, Gasoline, Hybrid
listing_color (15 levels)	BLACK, BLUE, BROWN, GOLD, GRAY, GREEN, ORANGE, PINK, PURPLE, RED, SILVER, TEAL, UNKNOWN, WHITE, YELLOW
make_name (37 levels)	Acura, Alfa Romeo, Audi, Bentley, BMW, Buick, Cadillac, Chevrolet, Chrysler, Dodge, FIAT, Ford, Genesis, GMC, Honda, Hyundai, INFINITI, Jaguar, Jeep, Kia, Lamborghini, Land Rover, Lexus, Lincoln, Lotus, Maserati, Mazda, Mercedes Benz, MINI, Mitsubishi, Nissan, Porsche, RAM, Subaru, Toyota, Volkswagen, Volvo
major_options (111 levels)	X101A.Mid.Equipment.Group, X2lt.Package, X301A.Mid.Equipment.Group, X302A.Luxury.Equipment.Group, X501A.Mid.Equipment.Group, X502A.Luxury.Equipment.Group, X5th.Wheel, X601A.Luxury.Equipment.Group, X701A.Luxury.Equipment.Group, X802A.Luxury.Equipment.Group, AMG.Sport.Package, Adaptive.Cruise.Control, Adaptive.Suspension, Alloy.Wheels, Ambient.Light.Package, Android.Auto, Appearance.Package, Audio.Package, Backup.Camera, Blind.Spot.Monitoring, Bluetooth, Bose.High.End.Sound.Package, CarPlay, Carbon.Ceramic.Brakes, Cargo.Package, Chrome.Wheels,

	Cold.Weather.Package, Comfort.Package, Convenience.Package, Convenience.Plus.Package, Customer.Preferred.Package, DVD.Entertainment.System, Driver.Assistance.Package, Driver.Confidence.Package, Dual.Rear.Wheels, Executive.Package, Extra.Value.Package, Handicap.Accessible, Heat.Package, Heated.Seats, LE.Package, LS.Package, LT.Package, Lariat.Package, Leather.Seats, Light.Package, Limited.Package, Luxury.Package, M.Sport.Package, Memory.Package, Multi.Zone.Climate.Control, Multimedia.Package, Navigation.System, Off.Road.Package, P01.Premium.Package, P1.Package, P2.Package, Parking.Sensors, Performance.Package, Popular.Equipment.Package, Power.Mirror.Package, Power.Package, Preferred.Accessory.Package, Preferred.Equipment.Package, Preferred.Package, Premium.1.Package, Premium.2.Package, Premium.Audio.Package, Premium.Package, Premium.Plus.Package, Premium.Smooth.Ride.Suspension.Package, Premium.Sound.Package, Premium.Wheels, Quick.Order.Package, RS.Package, Rear.Climate.Package, Remote.Start, S.Line.Sport.Package, SE.Package, SL.Package, SLE.Package, SLT.Package, SR5.Package, SXT.Package, Safety.Package, Security.Package, Skid.Plate.Package, Smoker.Package, Sound.Package, Special.Edition.Package, Sport.Chrono.Package, Sport.Package, Standard.Suspension.Package, Steel.Wheels, Storage.Package, Sunroof.Moonroof, Suspension.Package, TRD.Package, Technology.Package, Third.Row.Seating, Tow.Package, Trailer.Package, Trim.Package, Ultimate.Package, Upgrade.Package, Utility.Package, Value.Package, XLE.Package, XLT.Package, Z.71.Package, Z71.Package
transmission (4 levels)	A, CVT, Dual Clutch, M
state (4 levels)	Alaska, California, Oregon, Washington
wheel_system (5 levels)	4WD, 4X2, AWD, FWD, RWD

Table A2. List of variables which dummy variables were created for.

Appendix B: Distribution of Training Dataset

Appendix C: Outlier Detection



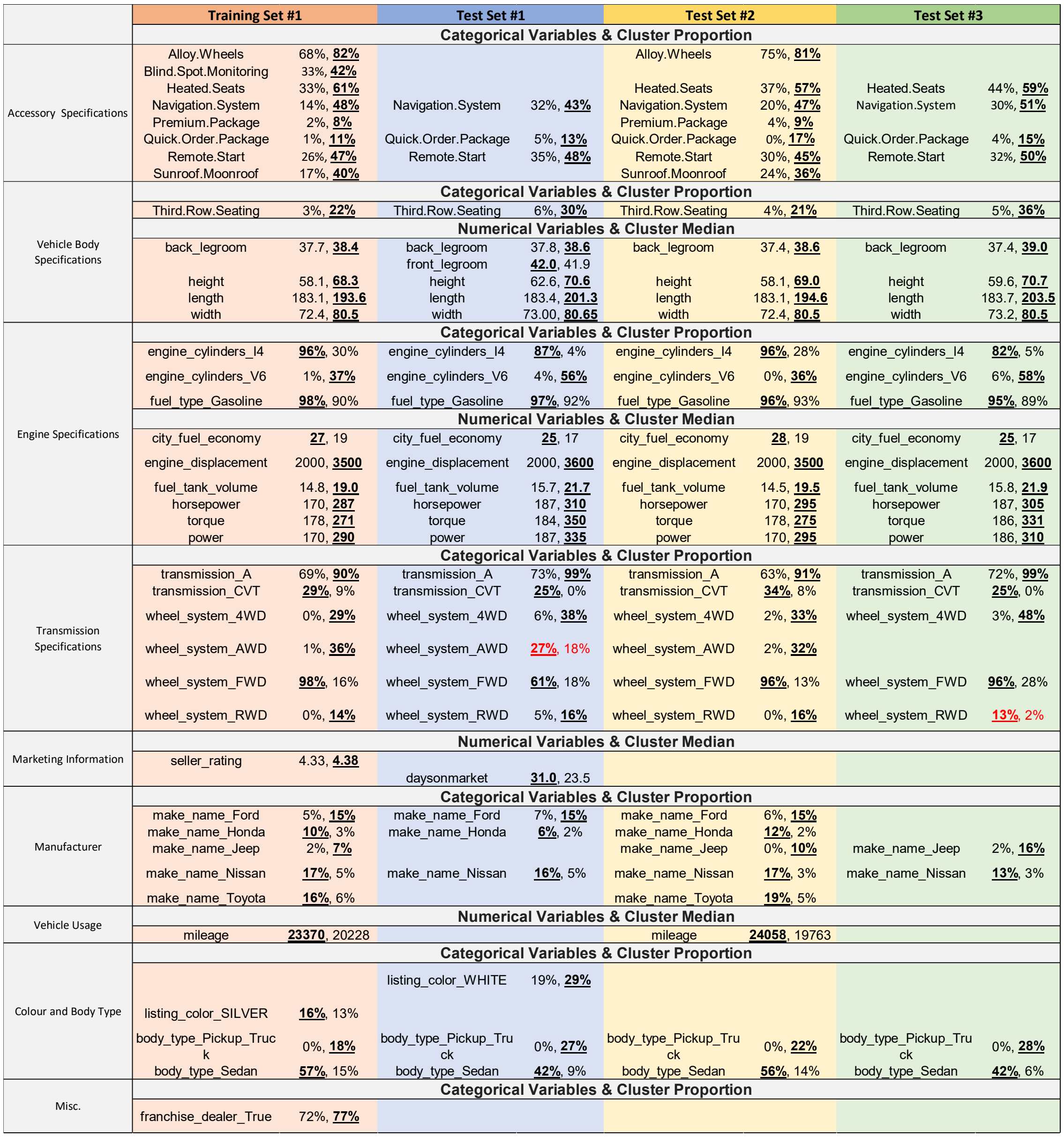
Appendix D: Dimension Reduction

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3620000	414000	-8.745	< 2e-16
back_legroom	-101	54.80	-1.841	0.065738
city_fuel_economy	298	42.40	7.035	2.49E-12
daysonmarket	3.19	1.75	1.829	0.067433
engine_displacement	-4.72	0.32	-14.892	< 2e-16
front_legroom	229	102	2.233	0.025598
fuel_tank_volume	746	79.50	9.379	< 2e-16
height	267	51.70	5.164	2.58E-07
horsepower	133	6.96	19.117	< 2e-16
length	-158	27.40	-5.762	9.22E-09
maximum_seating	-449	185	-2.43	0.015152
mileage	-0.12	0.01	-11.638	< 2e-16
seller_rating	727	268	2.713	0.006716
width	89.50	36.80	2.43	0.015147
year	1780	205	8.694	< 2e-16

torque	38	3.23	11.76	< 2e-16
power	-9.57	6.05	-1.581	0.113932
Alloy.Wheels1	-1230	352	-3.492	0.000487
Backup.Camera1	-2020	761	-2.655	0.007987
Blind.Spot.Monitoring1	564	323	1.749	0.080319
Heated.Seats1	562	352	1.597	0.110275
Navigation.System1	2680	353	7.585	4.48E-14
Premium.Package1	2140	601	3.562	0.000374
Quick.Order.Package1	-3180	618	-5.154	2.72E-07
Remote.Start1	-649	330	-1.964	0.049575
Sunroof.Moonroof1	2330	381	6.111	1.12E-09
Third.Row.Seating1	1340	573	2.334	0.019641
body_type_Pickup_Truck1	2640	989	2.669	0.007643
body_type_Sedan1	1950	574	3.391	0.000706
engine_cylinders_I41	-4510	658	-6.852	8.90E-12
engine_cylinders_V61	-5380	546	-9.843	< 2e-16
franchise_dealer_True1	1220	318	3.835	0.000128
fuel_type_Gasoline1	5510	713	7.721	1.59E-14
listing_color_SILVER1	-571	389	-1.467	0.142594
listing_color_WHITE1	-503	325	-1.548	0.121756
make_name_Ford1	-8280	617	-13.405	< 2e-16
make_name_Honda1	1720	688	2.501	0.012451
make_name_Jeep1	2600	827	3.149	0.001657
make_name_Nissan1	1350	583	2.321	0.020362
make_name_Toyota1	3620	552	6.563	6.26E-11
transmission_A1	-1930	1080	-1.783	0.074761
transmission_CVT1	-1730	1150	-1.508	0.131712
wheel_system_4WD1	5290	760	6.966	4.05E-12
wheel_system_AWD1	6570	878	7.483	9.62E-14
wheel_system_FWD1	3920	882	4.449	8.95E-06
wheel_system_RWD1	3310	975	3.393	0.000702
state_Oregon1	-934	325	-2.875	0.004071

Table D1. Result of stepwise regression.

Appendix E: Feature Extraction



	Training Set #1		Test Set #1		Test Set #2		Test Set #3	
	Categorical Variables & Cluster Proportion							
Accessory Specifications	Alloy.Wheels	68%, **82%**			Alloy.Wheels	75%, **81%**		
	Blind.Spot.Monitoring	33%, **42%**						
	Heated.Seats	33%, **61%**			Heated.Seats	37%, **57%**	Heated.Seats	44%, **59%**
	Navigation.System	14%, **48%**	Navigation.System	32%, **43%**	Navigation.System	20%, **47%**	Navigation.System	30%, **51%**
	Premium.Package	2%, **8%**			Premium.Package	4%, **9%**		
	Quick.Order.Package	1%, **11%**	Quick.Order.Package	5%, **13%**	Quick.Order.Package	0%, **17%**	Quick.Order.Package	4%, **15%**
	Remote.Start	26%, **47%**	Remote.Start	35%, **48%**	Remote.Start	30%, **45%**	Remote.Start	32%, **50%**
	Sunroof.Moonroof	17%, **40%**			Sunroof.Moonroof	24%, **36%**		
Vehicle Body Specifications	Categorical Variables & Cluster Proportion							
	Third.Row.Seating	3%, **22%**	Third.Row.Seating	6%, **30%**	Third.Row.Seating	4%, **21%**	Third.Row.Seating	5%, **36%**
	Numerical Variables & Cluster Median							
	back_legroom	37.7, **38.4**	back_legroom	37.8, **38.6**	back_legroom	37.4, **38.6**	back_legroom	37.4, **39.0**
			front_legroom	**42.0**, 41.9				
	height	58.1, **68.3**	height	62.6, **70.6**	height	58.1, **69.0**	height	59.6, **70.7**
	length	183.1, **193.6**	length	183.4, **201.3**	length	183.1, **194.6**	length	183.7, **203.5**
	width	72.4, **80.5**	width	73.00, **80.65**	width	72.4, **80.5**	width	73.2, **80.5**
Engine Specifications	Categorical Variables & Cluster Proportion							
	engine_cylinders_I4	**96%**, 30%	engine_cylinders_I4	**87%**, 4%	engine_cylinders_I4	**96%**, 28%	engine_cylinders_I4	**82%**, 5%
	engine_cylinders_V6	1%, **37%**	engine_cylinders_V6	4%, **56%**	engine_cylinders_V6	0%, **36%**	engine_cylinders_V6	6%, **58%**
	fuel_type_Gasoline	**98%**, 90%	fuel_type_Gasoline	**97%**, 92%	fuel_type_Gasoline	**96%**, 93%	fuel_type_Gasoline	**95%**, 89%
	Numerical Variables & Cluster Median							
	city_fuel_economy	**27**, 19	city_fuel_economy	**25**, 17	city_fuel_economy	**28**, 19	city_fuel_economy	**25**, 17
	engine_displacement	2000, **3500**	engine_displacement	2000, **3600**	engine_displacement	2000, **3500**	engine_displacement	2000, **3600**
	fuel_tank_volume	14.8, **19.0**	fuel_tank_volume	15.7, **21.7**	fuel_tank_volume	14.5, **19.5**	fuel_tank_volume	15.8, **21.9**
	horsepower	170, **287**	horsepower	187, **310**	horsepower	170, **295**	horsepower	187, **305**
	torque	178, **271**	torque	184, **350**	torque	178, **275**	torque	186, **331**
	power	170, **290**	power	187, **335**	power	170, **295**	power	186, **310**
Transmission Specifications	Categorical Variables & Cluster Proportion							
	transmission_A	69%, **90%**	transmission_A	73%, **99%**	transmission_A	63%, **91%**	transmission_A	72%, **99%**
	transmission_CVT	**29%**, 9%	transmission_CVT	**25%**, 0%	transmission_CVT	**34%**, 8%	transmission_CVT	**25%**, 0%
	wheel_system_4WD	0%, **29%**	wheel_system_4WD	6%, **38%**	wheel_system_4WD	2%, **33%**	wheel_system_4WD	3%, **48%**
	wheel_system_AWD	1%, **36%**	wheel_system_AWD	**27%**, 18%	wheel_system_AWD	2%, **32%**		
	wheel_system_FWD	**98%**, 16%	wheel_system_FWD	**61%**, 18%	wheel_system_FWD	**96%**, 13%	wheel_system_FWD	**96%**, 28%
	wheel_system_RWD	0%, **14%**	wheel_system_RWD	5%, **16%**	wheel_system_RWD	0%, **16%**	wheel_system_RWD	**13%**, 2%
Marketing Information	Numerical Variables & Cluster Median							
	seller_rating	4.33, **4.38**						
			daysonmarket	**31.0**, 23.5				
Manufacturer	Categorical Variables & Cluster Proportion							
	make_name_Ford	5%, **15%**	make_name_Ford	7%, **15%**	make_name_Ford	6%, **15%**		
	make_name_Honda	**10%**, 3%	make_name_Honda	**6%**, 2%	make_name_Honda	**12%**, 2%		
	make_name_Jeep	2%, **7%**			make_name_Jeep	0%, **10%**	make_name_Jeep	2%, **16%**
	make_name_Nissan	**17%**, 5%	make_name_Nissan	**16%**, 5%	make_name_Nissan	**17%**, 3%	make_name_Nissan	**13%**, 3%
	make_name_Toyota	**16%**, 6%			make_name_Toyota	**19%**, 5%		
Vehicle Usage	Numerical Variables & Cluster Median							
	mileage	**23370**, 20228			mileage	**24058**, 19763		
Colour and Body Type	Categorical Variables & Cluster Proportion							
			listing_color_WHITE	19%, **29%**				
	listing_color_SILVER	**16%**, 13%						
	body_type_Pickup_Truck	0%, **18%**	body_type_Pickup_Truck	0%, **27%**	body_type_Pickup_Truck	0%, **22%**	body_type_Pickup_Truck	0%, **28%**
	body_type_Sedan	**57%**, 15%	body_type_Sedan	**42%**, 9%	body_type_Sedan	**56%**, 14%	body_type_Sedan	**42%**, 6%
Misc.	Categorical Variables & Cluster Proportion							
	franchise_dealer_True	72%, **77%**						

Table E1. Summary of significance tests done on training and three test sets for feature extraction.