Clustering on Used Car Dataset

Data set: used car data worksheet.
The data set has details of 1008 used cars along with the following variables: 1. Brand, 2. Car model, 3. Resale price, 4. Mileage, 5. Seat capacity, 6. Vehicle type, 7. Fuel type, 8. Transmission, 9. Parking sensor, 10. Airbag, 11. Cruise Control, 12. Keyless entry, 13. Alloy wheel, 14. ABS, 15. Climate control, 16. Rear AC vent and 17. Power Steering

1. For the given dataset, which distance measure is more appropriate?

   In an unsupervised learning method like cluster analysis, the aim is to combine similar datapoints into a set of groups which are homogenous within and heterogenous across. In order to do that, it is important to define the similarity criterion. More often, this criterion is based on the choice of the distance metric we choose.
   In the current dataset, we have a nice mix of numerical variables (Resale Price,Mileage and Seat Capacity)as well as categorical variables. It is important to note that not all the variables are binary. Moreover, it is difficult to convert the value of mileage and price which have very large ranges making Jaccard an unsuitable metric. On the other hand using a Euclidean Norm in this case wouldn't make sense either since there are categorical variables present. Thus to deal with these data where there are a mixture of variable types, Gower's distance is a suitable metric which can be employed. It can handle each variable type efficiently and eliminates the need to one-hot encode the categorical variables.

$$d(i,j) = \frac{\sum_{k=1}^{p} \delta_{ij}^{k} d_{ij}^{k}}{\sum_{k=1}^{p} \delta_{ij}^{k}} \tag{1}$$

   where d_{ij}^{k} is calculated depending based on attribute type and δ_{ij}^{k} is either 0 if there are missing values or if two values are equal or if the attribute is asymmetric binary. In all other cases, this value is equal to 1.

2. Use the distance measure identified in (a) and cluster the data. Identify the cluster characteristics.

   As discussed in the previous subpart, we use Gower's metric to cluster the data. Since k-means usually works on numerical data only (where we can compute the mean), for the current dataset, we prefer to use Hierarchical clustering. This method is more flexible since it relies on a distance matrix to decide which observation is going to be clustered first.
   The nature of the dataset has already been discussed in the previous subpart. In this question, agglomerative hier-

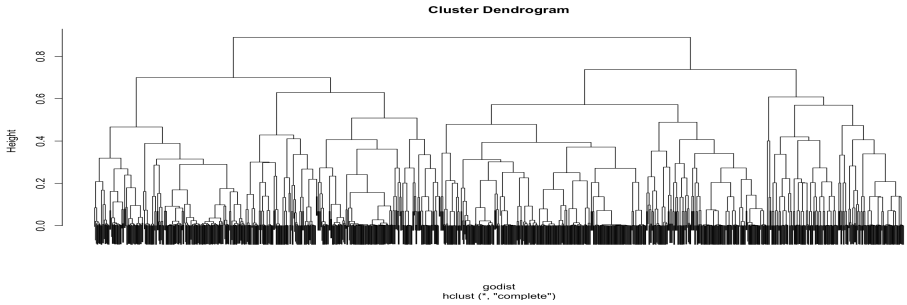
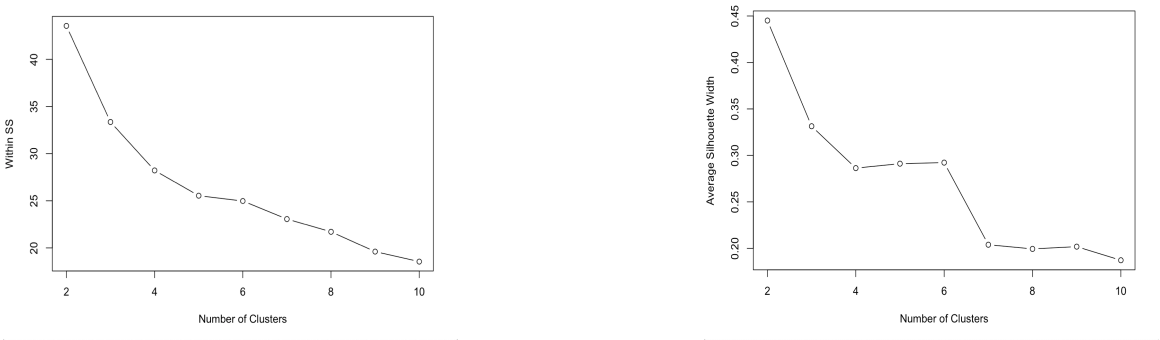


Figure 1: Hierarchical Clustering Based Gower's distance

   archical clustering has been applied with the "complete" linkage method. The linkage method was chosen based on considerations with respect to outliers in the data which might exist in the numerical attributes. Also, while single linkage is also able to handle outliers in the data, complete linkage usually provides more balanced clusters in terms of size. The dendrogram obtained for the cluster is then shown in Figure 1.

   Now that the dendrogram is ready, the next step is to identify how many clusters are needed to capture the patterns in the data. Depending on where the dendrogram is cut, we end up with "k" clusters. The choice of where to cut the dendrogram is usually very subjective but we can rely on certain metrics which can show us how many clusters would be appropriate. In the current scenario, the within sum of squares ("elbow plot"), Average Silhouette width are the metrics chosen to decide the number clusters. Based on Figure 2, we can see that a good choice of clusters would be 5-6 where we have a proper elbow in the first plot. Although Silhouette width suggests 2 clusters should be good enough, however it would lead to a lot of points in one cluster and not really aid us in interpreting the clusters



(a) (b)

Figure 2: Deciding the number of clusters based on (a) Within Sum of Squares and (b) Average Silhouette Width

properly. There is a significant drop in silhouette width after 6 clusters, which also agrees with our elbow plot. Thus we choose 6 as the optimal choice. Accordingly, cutting our dendrogram yields the following clusters: Now that

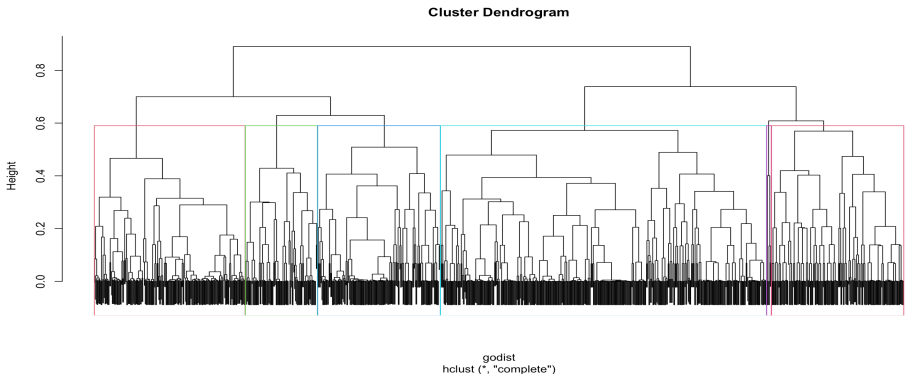


Figure 3: 6 identified clusters

we have our clusters, we can get an idea of what each cluster is talking about or in other profile these clusters. The cluster sizes from 1 through 6 are 406, 165 , 90, 188, 153 and 6 respectively with cluster 1 being the largest and 6 being the smallest. Based on an exploratory analysis of the clusters, the following observations were made:

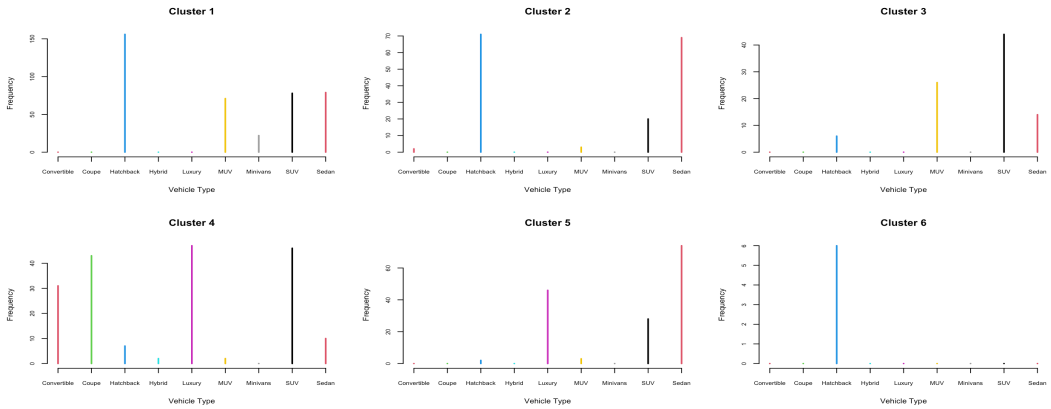


Figure 4: Distribution of Vehicle Types among Clusters

- The analysis of Vehicle Types for all clusters revealed interesting trends.
  (a) Cluster 4 is the only cluster which has Coupe and Hybrid Cars. Cluster 4 boasts a lot of luxury cars and convertibles as well. This is indicative of the fact that most of the cars in cluster 4 lean to the expensive side.

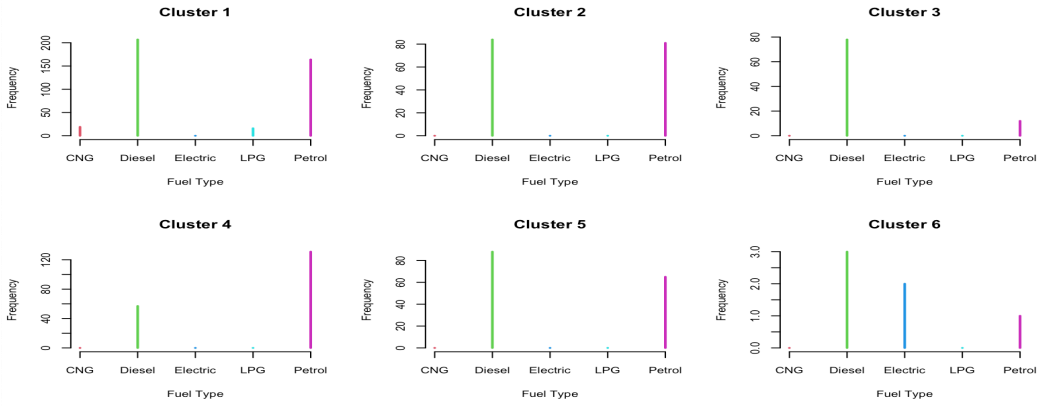


Figure 5: Distribution of Fuel Types among Clusters

Figure 6: Starplot for Price, Mileage, Seating Capacity for all 6 Clusters

(b) Cluster 1 is the only cluster that has minivans and it also has the maximum number of Hatchbacks.
(c) Cluster 2 is characterized by a large number of Hatchbacks and Sedans
(d) Cluster 5 like Cluster 4 has a lot of Luxury Cars and a very large number of Sedans
(e) Cluster 6 is a small cluster and it has only Hatchbakcs.
(f) Cluster 3 has a large number of SUVs

- The clusters were also analysed based on fuel types. Four out of six clusters shared the major proportion of petrol and diesel based cars in the dataset. CNG and LPG fuelled cars were limited and all of them fall in cluster 1. Cluster 6 has unique electric cars.
- Other features were also analysed for each of these clusters. It was found that Cluster 1 was primarily characterised by a very few amenities (Parking Sensors, Airbags, etc.) while the cars in Cluster 4 majorly had all amenities present. All cars in Cluster 6 were automatic, other clusters showed a mixture of automatic and manual based cars with Cluster 4 having the most number of automatic cars. Cluster 1 had manual cars in a larger proportion.
- . An analysis of the Resale Price, Seating Capacity and Mileage showed patterns which agreed with what we were thinking.

(a) Cluster 4 has the highest price on an average which is not surprising given that it has cars which fall under luxury categories and with most of the features present. However they lack in terms of mileage
(b) The cars in Cluster 6 gave the highest average mileage
(c) Cluster 3 had cars which have a reasonable resale price but they had the maximum seating capacity.
(d) Cluster 2 on the other hand was found to have the lowest resale price with a moderate seating capacity and mileage

From the discussion above, we can classify these cars into buckets like:

- Cluster 1 - Basic Models and mini cars (Small cars with only basic amenities)

- Cluster 2 - Cost effective Cars (Cheap Resale)
- Cluster 3 - Family Friendly (High seating Capacity)
- Cluster 4 - Luxurious Cars (Highest Price, Best Amenities)
- Cluster 5 - Best of both worlds (A decent mixture of all factors and couldn't come up with a name)
- Cluster 6 - Long Lasting Cars (Highest Mileage)

3. Use only the numerical variables (resale price, mileage, and capacity) in the data set and build clusters. Compare clusters developed in (b) and (c). Which clusters are better? Justify your answer.

   For this question, the numerical variables are considered alone and hence, it is possible to use k-means algorithm for clsutering the data in this case. Now before proceeding to cluster, two important things are required to be done. First, we scale all the input variables and second, we need to determine the number of clusters before proceeding with the algorithm. For deciding the number of clusters, we resort to the elbow plot as usual and based on that we decide 6 as the optimal number of clusters. On fitting the kmeans model, we get a cluster solution as follows:
   From Figure 7b, we can see there are 6 clusters but there is one major thing to notice. There is heavy influence of

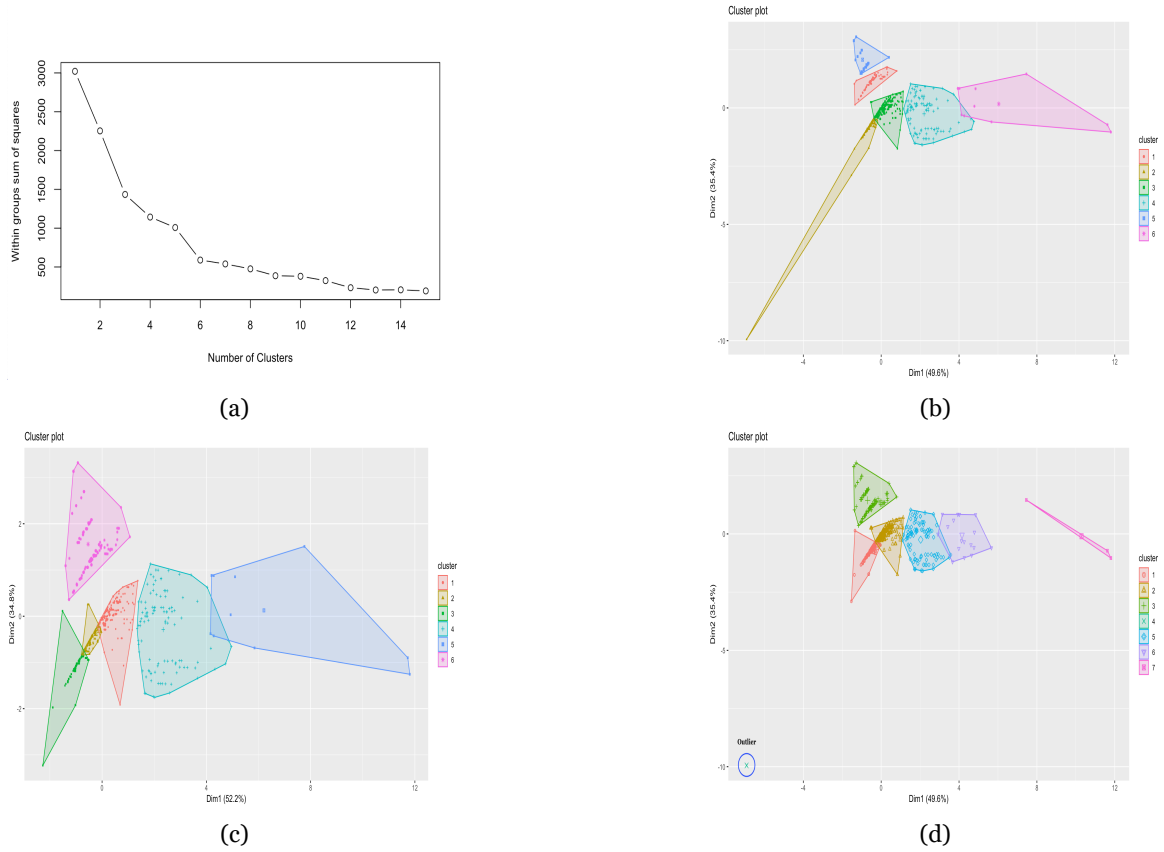


Figure 7: Clustering Results: (a) Elbow Plot;(b)K-means with random cluster means; (c) k-means with random centers and outliers removed and (d) Hierarchical Clustering followed by K-means

   an outlier which leads to a very elongated cluster 2. As we already know, k-means is sensitive to outliers and this may thus prove to be a little problematic. Thus the observation causing this behaviour is removed and the clusters are revisited as shown in Figure 7c. While this isn't a bad way to decide the value of k or the number of clusters, there is a more robust method of doing the same without having to remove the outlier.

   The results of Hierarchical clustering can be used to get the cluster centers for k-means instead of them being decided on random. This is extremely helpful in the presence of outliers in the data, since the linkage methods like "single", "complete" and "average" can be pretty resistant to outlier effects. The algorithm is as follows:

   - Compute the distance matrix based on a suitable distance metric ("Euclidean" in this case)
   - Fit the hierarchical clustering model on the distance matrix

- Cut the dendrogram into k clusters and compute the cluster centres
- Use the cluster centres obtained in the k-means algorithm

The results obtained are shown in Figure 7d and we can see that the outlier is easily identified as a separate cluster while we have 6 dominant clusters.

Comparison of Models: To compare the clustering models, a variety of cluster validation metrics were used and they have been listed in the table below: Now, for a good clustering result, within sum of squares should be

Measure	K-means	K-means with Hierarchical	Hierarchical
Number of Clusters	6	7	6
W/B Ratio	0.5196	0.3226	0.4426
Within Sum of Squares	1642.2	440.966	24.976
Calinski Harabasz Index	168.652	976.118	559.078
Dunn Index	0.1039	0.01354	1.1883

minimum and between sum of squares should be maximum. W/B Ratio is the ratio of Within to Between Sum of Squares. On the other hand, CH index is the ratio of Between to Within distances. Thus ideally, W/B ratio should be lower and CH index should be higher. A higher value of Dunn Index is preferred for a good cluster. Now based on the ideal choices of these metrics, K-means with Hierarchical Clustering performs the best out of all the 3 methods.