COURSE WORK: Data Mining

M.Sc. Data Science, Coventry University UK (2024.2 Batch)

Lahiru Munasinghe

Index: COMScDS242P-001

Question 1

Git repository: https://github.com/lahirudatascience/online-retail-data-mining

1. Methodology and Introduction.

This paper offers a thorough basket analysis applied to the Online Retail Dataset, applying the Apriori method. For a non-store online retailer based in the UK running between December 1, 2010, and December 9, 2011, the dataset comprises almost 500,000 rows of transactional data. With a lot of wholesale clients, the retailer specialises in unusual all-occasion gift items.

The study seeks to find relationships between often bought items and translate these results into a corporate setting. After data cleansing and transformation, Apriori ranked France, Germany, and the United Kingdom-after transaction volume. Extraction of practical insights capable of guiding strategic decisions in marketing, inventory control, and customer retention was a main objective.

Data preprocessing included the removal of:
- Rows with missing “CustomerID”
- Rows with null or blank “Description”
- Transactions with non-positive “Quantity” or “UnitPrice”
- Cancelled transactions (InvoiceNo starting with ‘C')
- Stock codes unrelated to actual product purchases (e.g., delivery charges, adjustments)

Transactions were also arranged by “InvoiceNo”, and customer-level data were generated for advanced correlation study. Extremely high item count outliers were eliminated to guarantee representative rule mining.

In data mining, the Apriori algorithm is a basic technique used for association rule discovery. It is especially appropriate for market basket analysis, in which case the objective is to find regularly purchased products. The method generates association rules from frequent itemsets in the dataset firstly identifying them.

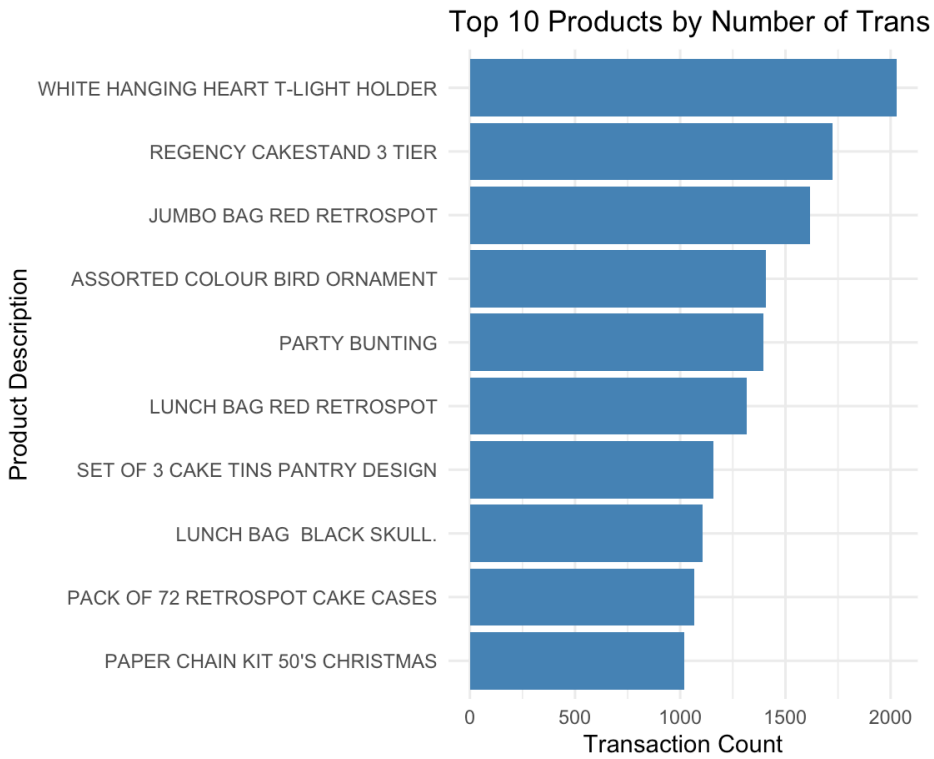
The fundamental ideas consist in:
- Support: Track the frequency of an itemset in the data.
- Confidence: Measures the reliability of the inference made by a rule.
- Lift: Measures a rule's strength relative to chance. A lift above one denotes a strong correlation.

Using a bottom-up approach, Apriori tests groups of candidates against the data after extending frequent subsets one item at a time (candidate generation). It uses the anti-monotonicity property; if an itemset is rare, all of its supersets are also rare, hence computing overhead is much lowered. Apriori is used in this study on country-specific subsets of the Online Retail dataset to identify trends in consumer purchasing behaviour that might guide operational and marketing strategic decisions.

2. Data Understanding and Graph Interpretation.

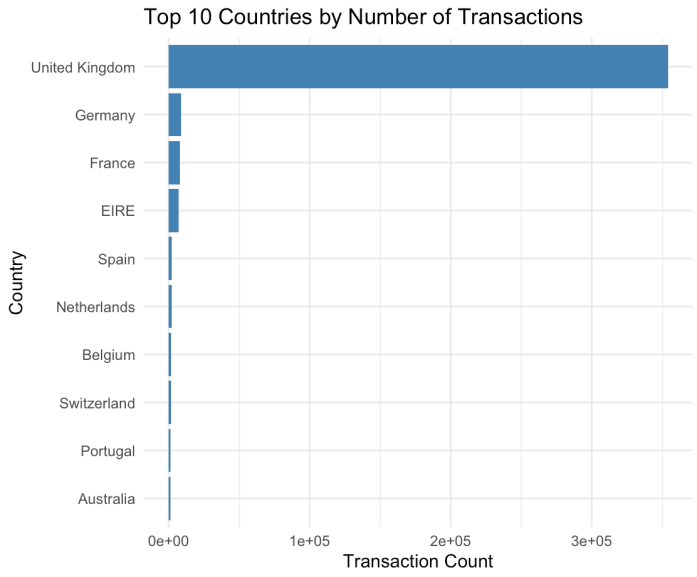
First I look at the structure and main trends of the dataset. Different graphs and summaries were created to capture the most often occurring products, top consumers, country-based distribution, and purchasing behaviour.

Graph 1: Top Ten Items by Transaction Count.



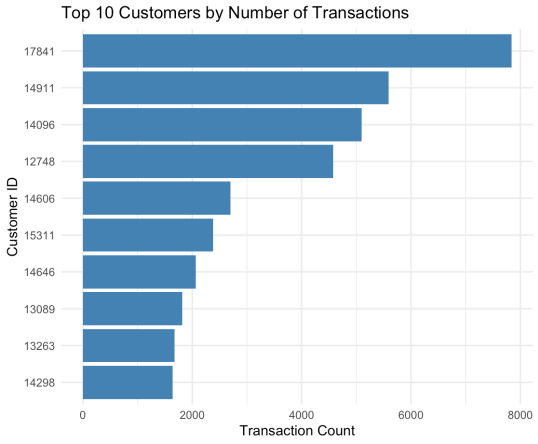
Leading the list with "WHITE HANGING HEART T-LIGHT," followed by "Regency CAKESTAND 3 Tieter," this bar chart highlights the most often sold products. These are quite valuable gift-oriented items reflecting seasonal and event-based buying trends.

Graph 2: Top Ten Count of Transactions Countries.



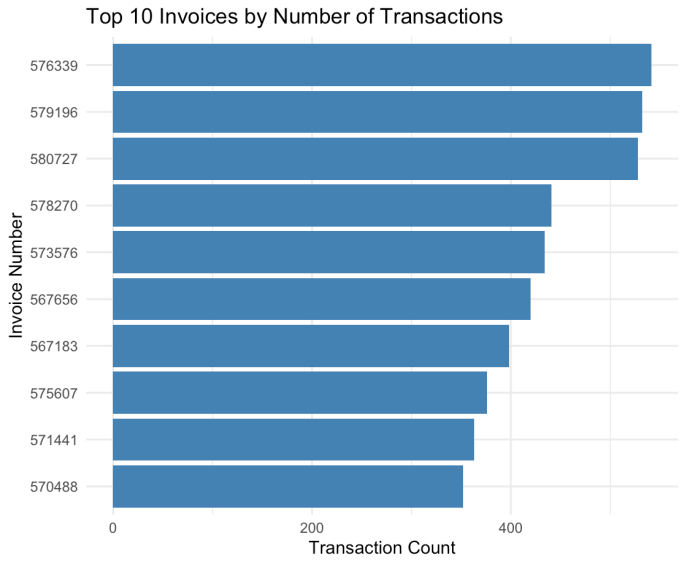
This graph nicely shows the United Kingdom's dominance in transaction count.
Underlying the company's UK-centric focus, other countries including Germany, France, and Ireland (EIRE) make rather less contributions.

Graph 3: Top 10 Transaction Count Customers.



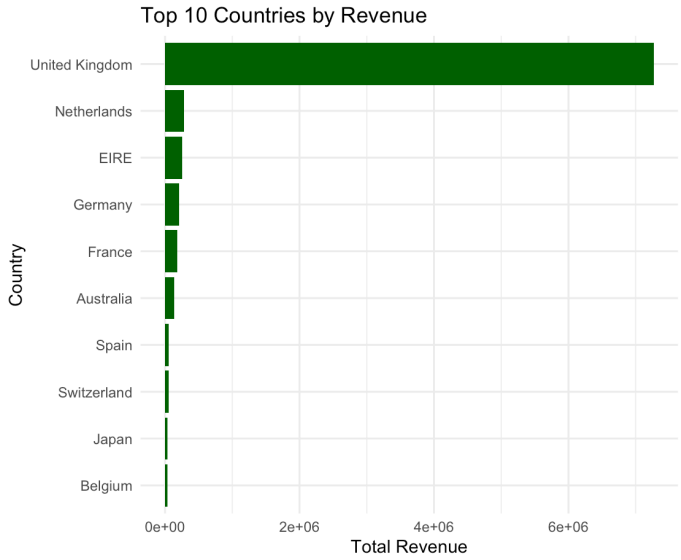
Frequent buyers including customer IDs like 14911 and 17841 most certainly reflect wholesale clients or loyal return business. Given their primary income generation, these customers could be the centre of loyalty programs or special offers.

Graph 4: Top 10 Invoices by Transaction Count.



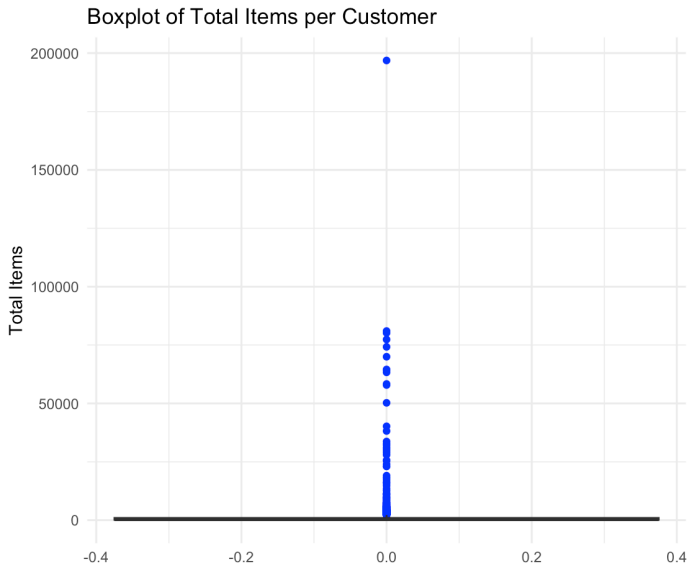
High- volume invoices imply bulk buying behaviour and so confirm the presence of B2B clients. For the basket analysis report on the online retail dataset, for example, invoice 576 339 shows more than 450 items—an indication of wholesaling.

Graph 5: Top Ten Revenue Countries.



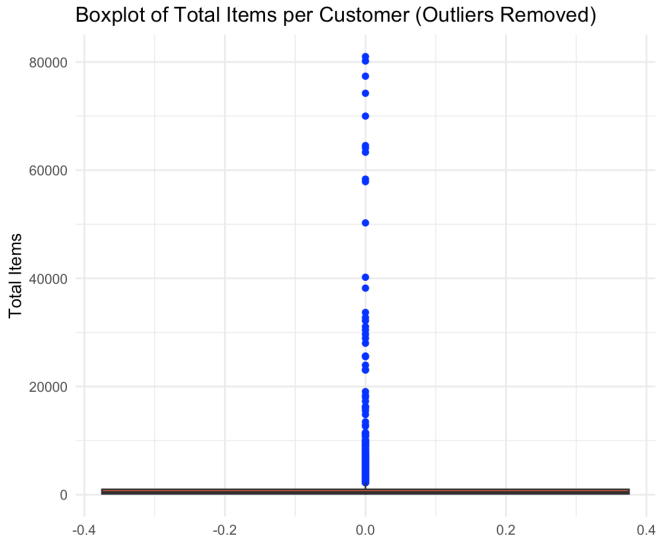
Not only in transaction count but also in total income; the UK once more guides both. Fascinatingly, although the Netherlands and EIRE follow in transaction volume, they contribute more in income than other countries with higher transaction counts, implying either more expensive goods or bigger basket sizes.

Graph 6: Consumer Total Items Per Boxplot.



This graph shows a long-tailed distribution whereby a small number of consumers purchase rather significant amounts.
Outliers—one customer with almost 200,000 items—were eliminated in next study to prevent distorted rule generation.

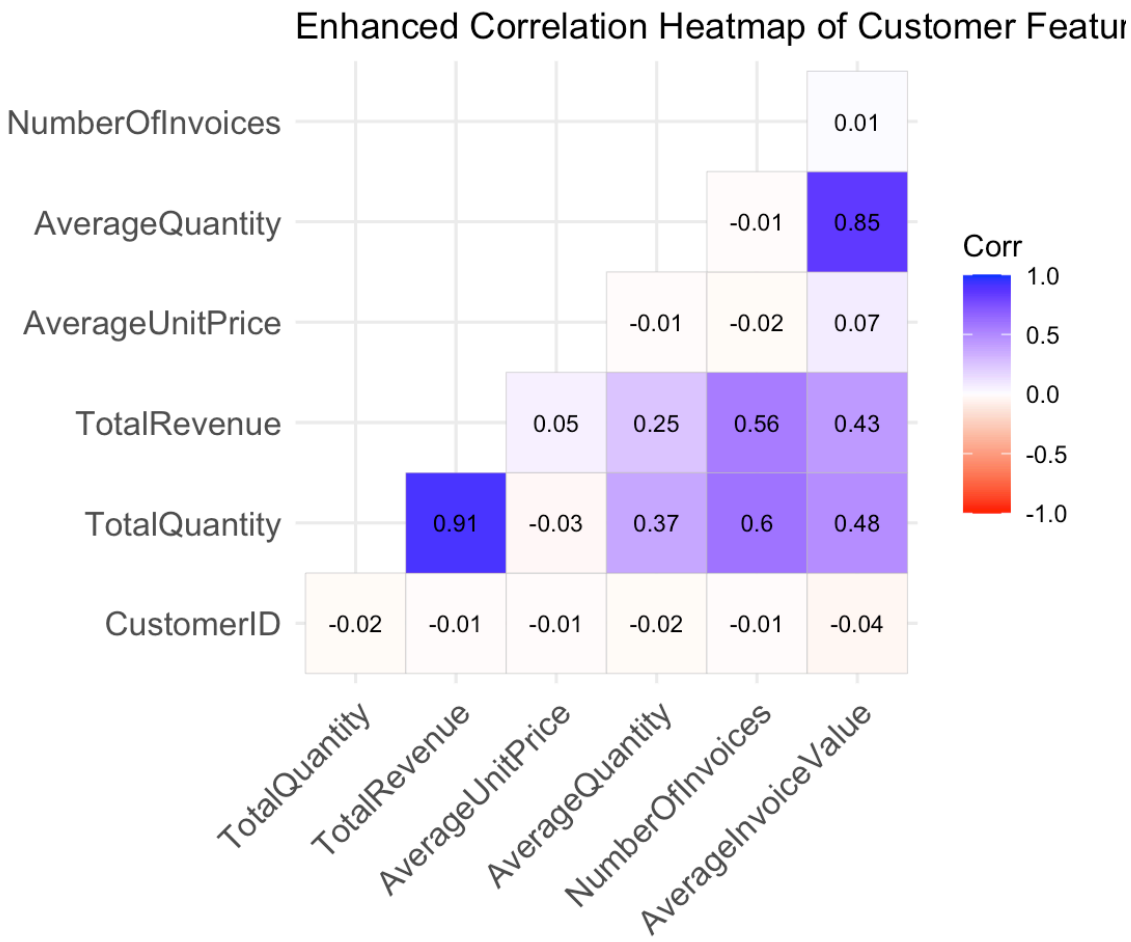
Graph 7: Total Items Per Customer (Outliers Eliminated) Boxplot



Customers with more than 100,000 items extreme outliers are left off from this polished boxplot. The graph presents a more focused viewpoint of normal purchase patterns. Although most consumers buy less than 20,000 items overall, a notable cluster purchases between 20,000 and 60,000 items, so indicating the presence of high-volume buyers even within the filtered data.

By removing outliers, one can ensure that the Apriori algorithm discovers trends relevant to most consumers instead of just distorted by a small number of wholesale or abnormal buyers. This suggests that found rules will be applicable not only for the dependability of association guidelines but also for corporate strategies including bundling and cross-selling.

Graph 8: Enhanced Correlation Heatmap



On a heatmap, strongly favourable correlations between customer features show: 0.91 for TotalQuantity against TotalRevenue; 0.56 for TotalRevenue against Number of Invoices.
This shows that high-quantity customers often make more orders and generate more income. The correlation analysis confirms our focus on top consumers and product combinations.

3. Apriori Algorithm Implementation.

The Apriori algorithm is a classic method in association rule mining. It was applied here to find frequent item sets and derive rules of the form A => B, whereby the purchase of item A results in the purchase of item B. It starts with finding all frequent itemsets in the dataset-groups of objects that show together in a transaction more often than a predefined threshold (support). This method uses the anti-monotonicity feature, which makes it particularly effective in big datasets such as the Online Retail one. It guarantees effective pruning of the search space: any superset including a given combination of products cannot be frequent either if that combination is not frequent itself.

Step-by-Step Implementation:

1. Group data by `InvoiceNo` and aggregate product `Description` into lists (baskets).
2. Convert these baskets into transaction objects.
3. Use Apriori with minimum support = 0.01 and confidence = 0.5.
4. Rank rules by Lift to determine their strategic importance.

Rationale for Parameters:

- Support: ensures only consider item sets that occur frequently enough to be actionable.
- Confidence: ensures that the rule is reliable.
- Lift: measures the strength of the rule relative to random chance. A lift > 1 implies a meaningful rule.

For every one of the top countries—France, Germany, and the UK—Apriori was carried out separately so that tailored analysis considering regional variations in purchasing behaviour could be produced. More exact and practical results catered to every market come from this separation.

The algorithm produced in some cases hundreds of rules. Focussing on the top rule by lift from every nation, Investigated their significance and possibilities for commercial strategy in order to derive business value.

Further filtering with support count (`count`) guaranteed that rules were not only robust statistically but also occurred in enough quantity to be pertinent for retail decisions.

4. Basket Analysis Results by Country.

Deeper Analysis - France:
France offered the most varied set of connections with 104,843 rules created. This probably reflects more different shopping habits and more limited, more specialised purchase groups. Linking SMALL RED BABUSHKA NOTEBOOK with the SMALL YELLOW BABUSHKA NOTEBOOK had as the most crucial rule:

- Support: 1.05%
- Confidence: 100%
- Lift: 75.8%

This implies that every consumer who bought the red variant also bought the yellow variant; hence, the probability of co-purchase is more than 75 times higher than chance. Such guidelines point to very specific, high-affinity product pairs that might be perfect for "buy one, get the second" specials or combined gift sets. This points to a visual attractiveness or collecting trend that should be underlined in online recommendations and visual merchandising.

Deeper Analysis - Germany
Germany produced a more modest 4,671 rules. The standout rule connected:
- {POPPY'S PLAYHOUSE LIVINGROOM} => {POPPY'S PLAYHOUSE BATHROOM}
- Support: 1.13%
- Confidence: 83.3%
- Lift: 73.8

These items most certainly fit either a thematic or sequential set. Based on the confidence and lift values, German consumers buying one playhouse item are quite likely to buy others from the same theme.
This offers a great chance to create children's play and learning theme-based kits. Marketing initiatives can also be simplified to exhibit "Complete Your Playhouse" prompts.

Deeper Analysis - United Kingdom:
The UK dataset generated 269 rules-a much smaller number but with stronger rule quality due to higher transaction density and consistency. The best rule identified was:

- {HERB MARKER THYME} => {HERB MARKER ROSEMARY}
- Support: 1.01%
- Confidence: 94.4%
- Lift: 86.5

This exposes a niche market for gardeners that typically purchases entire sets of herb markers. Highly helpful for upselling, a high lift of 86.5 suggests almost deterministic relations. Here the insight is not about mass-market bundling but rather focused marketing towards aficionados and hobbyists where design and completeness count.

Cross-Country Contrast:

- France's rules are wide-ranging and expressive, suitable for machine learning segmentation.
- Germany's rules are product-line focused and theme-driven, ideal for package-based upselling. Basket Analysis Report: Online Retail Dataset
- The UK offers precision and consistency, suitable for automation in recommendations and personalization.

These findings validate the fact that consumer behaviour differs depending on geography. Customising strategy to fit particular association patterns helps businesses to be more successful in their initiatives on inventory control and marketing campaigns.

5. Business Insights and Strategic Value.

Combining the exploratory data insights with the Apriori-based basket analysis exposes several actionable areas where the company might develop competitive advantage and propel value. These realisations cover operational effectiveness, customer segmentation, marketing strategy, and product management.

- **Inventory Optimization:**

Shelf placement and inventory planning are among the most obvious commercial uses for association rules. High-confidence product pairs imply that these products are often purchased concurrently. Should one of them run out of, the sales of the related pair may suffer directly.

Recommendation: Co-locate frequently associated products in the warehouse or storefront. Forecast and replenish paired items together to reduce lost sales due to stock unavailability

- **Product Bundling Opportunities:**

The rules offer solid evidence for creating product bundles, combo offers, or discounts.

Action: Bundle theme-based items (e.g., "Complete Playhouse Set") to simplify customer decision-making and boost average order value. Digital Strategy: Use "Frequently Bought Together" or "You May Also Like" modules on e-commerce platforms.

- **Personalized and Region-Specific Marketing:**

Every nation exhibits different purchasing patterns, which points to the need of micro segmentation and customised messaging. Rule-driven insights can feed web personalisations or focused email campaigns.

- **Customer Segmentation and Loyalty Programs:**

Customer data shows several categories including regular small-scale buyers, one-time bulk buyers, and high-frequency consumers. One can use these for tie-red loyalty programs.

- **Revenue Growth through Cross-Selling:**

Rules with great lift and confidence point to product combinations that should be recommended to consumers via follow-up or checkout emails.

- **Visual Merchandising and UX Design**

Either in-store or under UI sections like "Complete the Set," related items can be showcased together.

- **Operational Efficiency:**

Understanding what goods are often purchased together helps maximise order picking, packaging, and fulfilment processes.

- **Strategic Expansion Planning**

Buying behaviour unique to a country offers insight into localised product preferences, so supporting strategic worldwide marketing and inventory control.

- **Feedback Loop for Product Development:**

Recurring item sets can direct the development of themed product kits or combined SKUs.

- **From Data to Strategic Decision-Making:**

By means of improved product offerings, marketing enhancement, and operational simplification, this study demonstrates how association rule mining can be converted into direct business value.

Question 2

Git repository: https://github.com/lahirudatascience/vehicle_data_clustering_data_mining

1. Introduction

In the domain of data mining, a basic chore is identifying significant trends in unlabelled data. Natural groupings within data based on feature similarities are found using the unsupervised machine learning method known as clustering. This paper offers a thorough clustering analysis of a dataset consisting of vehicle forms in which every observation records geometric and statistical measurements derived from object shapes. Without using class labels during the clustering process, the goal is to find natural groups among vehicles based on their silhouette-based properties.

Principal Component Analysis (PCA) is first used to minimise the feature space while maintaining most of the variance since the data is complex and dimensionally varying. This dimensionality reduction improves visualisation, lowers computing performance, and simplifies the clustering process.

Following PCA, the report applies two distinct clustering methods:
1. K-Means Clustering – a partition-based algorithm that minimizes intra-cluster variance by assigning observations to the nearest cluster center.
2. Agglomerative Hierarchical Clustering (AHC) – a bottom-up hierarchical method that recursively merges the closest pairs of clusters based on linkage distance.

Every technique offers complimentary analysis of the data's structure. K-Means is computationally efficient and perfect for big datasets with well-separated clusters; AHC provides a comprehensive hierarchical view of data groups and does not rely on a predefined number of clusters. This combination guarantees a flexible hierarchical interpretation of the data together with a strong partitioning.

Important preprocessing tasks including missing value imputation, outlier detection and treatment, and feature standardising precede the clustering pipeline. These are absolutely essential to guarantee a clean, normalised dataset fit for distance-based clustering methods. Internal validation metrics including the Elbow Method, Silhouette Score, and Cophenetic Correlation Coefficients as well as visualising tools including dendrograms, PCA biplots, and clustered heatmaps help to evaluate clustering performance.

This paper attempts to extract meaningful structure, assess clusterering quality, and interpret the contribution of silhouette-based features in differentiating between vehicle groups by using both partitioning and hierarchical techniques to a PCA-reduced dataset. Apart from showing useful applications of clustering methods, the approaches also show how preprocessing and dimensionality reduction greatly affect the quality and interpretability of the clustering outputs.

2. Methodology

- **Data Overview and Preparation:**

There are 846 samples in the dataset with 18 numerical features representing several shadow-based measurements. Every feature offers a different geometric or statistical viewpoint of car form. One categorical column marks class labels but is only used for evaluation following clustering.

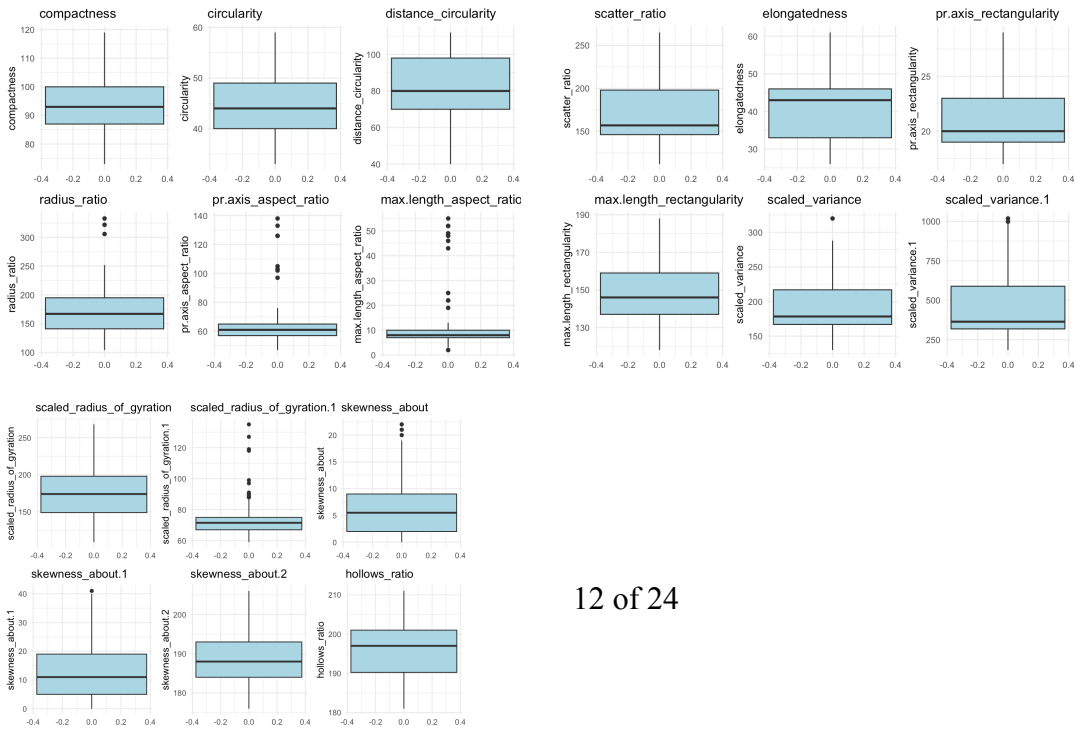
Using R tools like str() and summary() initial data inspection included checking data types, ranges, and completeness.

- **Missing Value Imputation**

Although the dataset was essentially complete, the mice package with the Predictive Mean Matching (PMM) technique was used for proactive missing value handling. To fill in missing values, PMM chooses real observed values from related records, so preserving the distribution and preventing the creation of arbitrary values. When preparing datasets for distance-based models, where even minor variations may skew clustering, this approach is especially helpful.

- **Outlier Detection and Treatment**

For every feature, outliers 1.5× IQR were found and visualised with boxplots. Particularly notable outliers were scaled_variance.1, radius_ratio, and max. length aspect ratio. Winsorizing was used instead of eliminating them—capping values at the 5th and 95th percentiles. This method reduces their impact while keeping all the data points—which is crucial in clustering when whole observation sets are preferred.



- **Feature Standardization**

The features had rather different scales. For example, scaled_variance ranged around 100 while compactness values.1 exceeded 30,000. Larger scale features would predominate in distance measures without normalising. All numerical columns were thus standardised (zero mean and unit variance) in R using the scale() function.

- **Principal Component Analysis (PCA)**

PCA was applied to the standardized data to:

- Remove feature redundancy,
- Visualize data in lower dimensions,
- Improve cluster separation.

The scree plot indicated that the first 2–3 principal components explain over 70% of the variance. Variable contribution plots were also generated to understand which original features influenced the PCA axes most.

- **K-Means Clustering**

After PCA, K-Means clustering was applied to the top two principal components. The number of clusters (k) was determined using:

• Elbow Method: Observed drop-off in WSS (within-cluster sum of squares),

• Silhouette Analysis: Average silhouette width (~0.47) confirmed reasonable separation.

K = 3 was selected and clustering was performed using multiple random initializations (nstart=25) to avoid local minima. The resulting clusters were visualized on the PC1 vs PC2 scatterplot and biplots.

- **Agglomerative Hierarchical Clustering (AHC)**

AHC was done with Euclidean distance and Ward's approach for linkage to augment K-Means. First three main components helped one to compute the distance matrix. Plotting dendrograms, coloured cluster boundaries were drawn at k = 3 for comparison with K-Means. Using cutree(), the cluster membership was extracted; for interpretability, a confusion matrix was then compared with actual class labels.

The quality of AHC clustering was evaluated using:

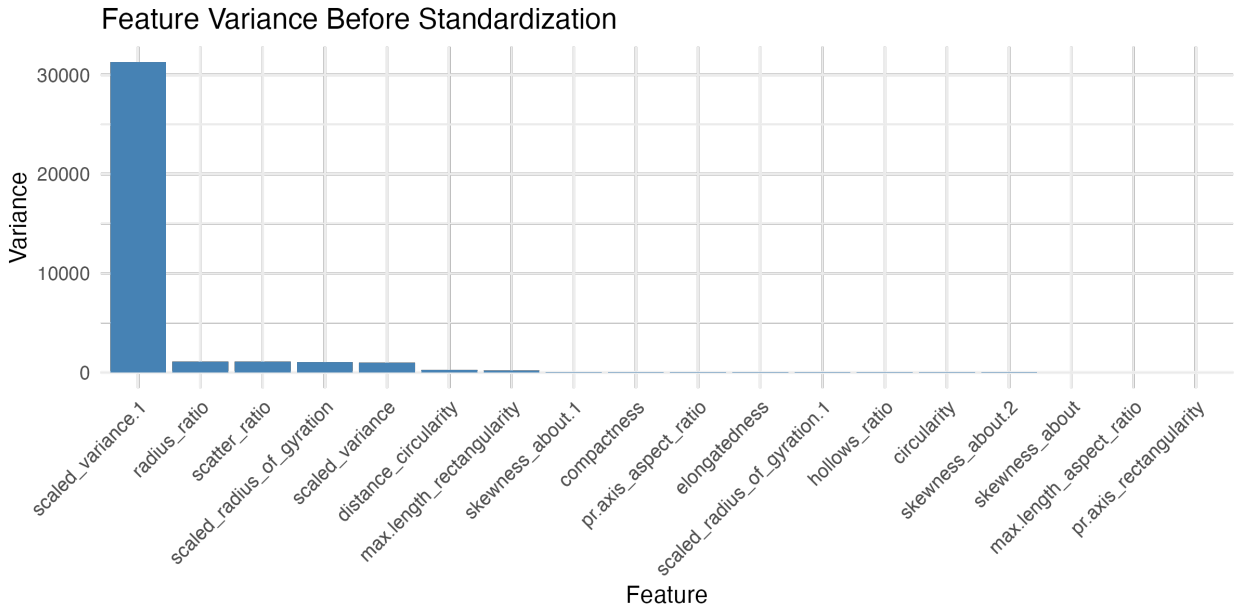
• Silhouette plot

• Cophenetic correlation matrix to assess linkage methods (Ward, complete, average, etc.)

• Clustered heatmaps for additional visual validation.

3. PCA and K-Means Clustering Analysis.

- **Feature Variance Before Standardization:**

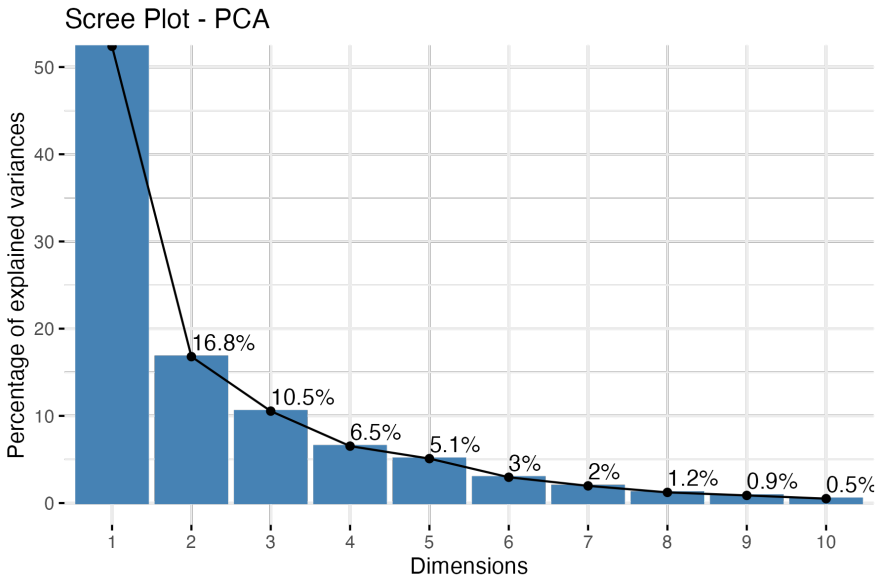


Analysing the raw features for any meaningful variations in scale or distribution is crucial before applying dimensionality reduction with PCA. This stage guarantees that none of one feature disproportionately affects the outcomes. A summary of the 18 numerical attribute variation is given by the feature variance plot. Especially compared to the next highest features, such radius_ratio and scatter_ratio, which remain below 2,000, scaled_variance.1 exhibits a shockingly high variance value of over 30,000. This striking difference emphasises the need of standards. High-variance features can dominate distance-based computations or principal component loadings in data mining, so skewing both PCA and clustering outcomes. Particularly PCA seeks to maximise variance in a small set of dimensions, hence disproportionately scaled features can produce false results. Moreover, such characteristics might not always be more informative; large variance could result from measurement scale rather than underlying structure. Early identification of this problem allowed us to apply Z-score standardising in later phases, so guaranteeing equal contribution of all features. This study establishes the need of preprocessing and provides the basis for consistent PCA, clustering, and interpretation in the upcoming stages of the research.

- **Feature Variance Table (Before Standardization):**

Before any transformation or normalising was done, this table shows each numerical feature's raw variations in the dataset. This data makes it abundantly evident that features have rather different scales. Comparatively to others like compactness at just 67.8, or pr. axis_aspect_ratio at 62.3, the most extreme example is scaled_variance.1, which has a variance over 31,000 (seen in the full edition). At 1119.1, even radius_ratio shows notable departure from smaller-variance characteristics. Such disproportionate values can introduce bias in distance-based models like K-Means or PCA and show the characteristics are on different numerical scales. The features with the highest variance would dominate the main components or clustering distance measurements without standardising, so suppressing the importance of smaller-scale features. One of the main reasons Z-score standardising (mean = 0, standard deviation = 1) was used prior to PCA was this disparity. Moreover, looking at this table offers early understanding of which characteristics naturally vary more and might have a more impact on clustering structure. For instance, because of their rather high variance even after standardisation, radius_ratio and distance_circularity may be rather important post-transformation.

- **PCA Scree Plot: Variance Explained by Components:**



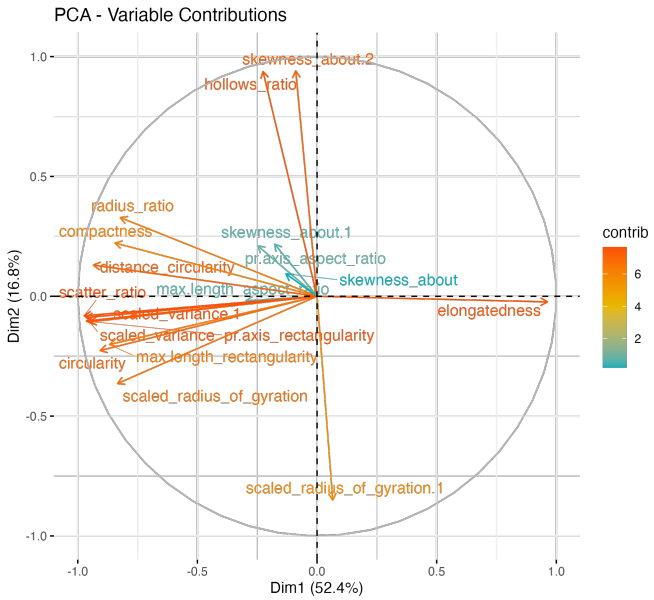
Principal Component Analysis (PCA) is used to lower dimensionality and eliminate duplicity once features are standardised. PCA converts the original correlated features into uncorrelated components (PCs) optimising the dataset variance. Important in deciding how many components to keep, the scree plot shows the proportion of total variance caught by every PC. Our study finds that the first component (PC1) by itself explains more than half of the total variance. About 17% is captured by the second component, PC2; together, PC1 and PC2 account for almost 67%

of the total variance. Including PC3 suggests that, with just the first few components, most of the significant information in the dataset can be preserved and brings the cumulative variance explained above 70%. This greatly simplifies matters so that we may work with two or three dimensions rather than all eighteen original features. Visualisation and clustering especially benefit from it since high-dimensional data can mask patterns. The elbow of the scree plot at the third component proves even more the minimal number of components needed. Keeping the best PCs helps us to balance computational efficiency in downstream clustering with information retention.

- **PCA Variance Explained Table**

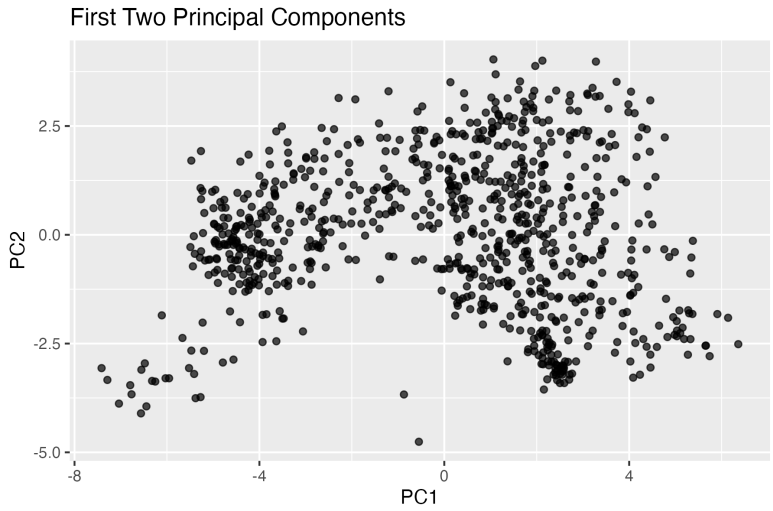
This table shows, from PCA, the percentage of variance explained by every principal component (PC). While PC2 explains an extra 16.79%, PC3 adds 10.54%, PC1 explains 52.38% of the total variance in the dataset. The first three components taken together account almost 80% of the total variability of the dataset, which is a strong sign that dimensionality reduction will not greatly compromise information. With PC10 and above contributing less than 1% each, the other components—e.g., PC4, PC5, etc.—contribute ever less. This supports the choice to cluster and view just the first two or three PCs. Furthermore supporting the plots you have created, such the PCA scatter plot and biplot, is the total variance caught by PC1 and PC2 alone (over 69%). These two aspects enable significant whole dataset visualisation. Furthermore, this table is essential for wise choices concerning dimensionality trade-offs. Although using more PCs improves accuracy, it might cause overfitting or lower interpretability. All things considered, this table supports the visual Scree Plot statistically and strengthens the dependability of the PCA-reduced dataset for next clusterings.

- **PCA Variable Contribution Plot**



Interpreting the reduced PCA space depends on knowing which features most affect the main components. The PCA variable contribution plot shows the significance of every standardised feature for building the top components. Features are shown in this radial plot by arrows; longer arrows indicate greater contributions. The angle between arrows denotes correlations—features pointing in similar directions are positively correlated, whereas those at opposite angles are negatively related. Our design shows that PC1 and PC2 are much influenced by elongatedness, radius_ratio, compactness, and scaled_radius_of_ gyration. These factors are very important in determining the new component axes, thus they are the most helpful in understanding variation over the dataset. Features like circularity, skewness_about, and hollows_ratio help less in the limited space and could be considered less important. Later in the study, this insight is absolutely vital for understanding cluster results. Understanding which physical features predominate in the PCA space helps one to explain why particular data points cluster. For PC1, for instance, clusters with greatly different elongatedness could seem to be well-separated. In essence, this visualisation supports open interpretation by bridging the gap between the abstract PCA transformation and original data features.

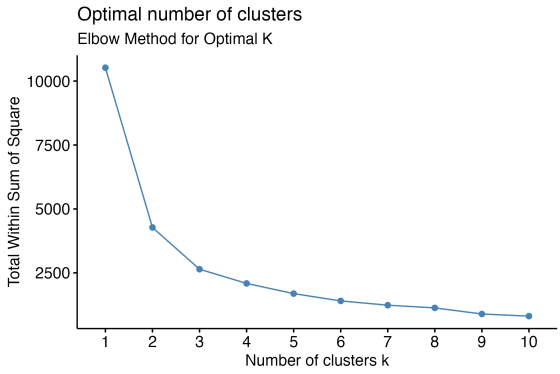
- **PCA Scatter Plot: Visualizing Dimensionality Reduction**



The PCA scatter plot presents a 2D picture of the first two principal component distribution of the observations. Every point on PC1 and PC2 defines the vehicle silhouette by means of its changed coordinates. While keeping over 65% of the variance in the dataset, this projection cuts the original 18-dimensional feature space to just two axes. Underlying clusters are indicated by a rapid visual inspection showing that the data organically forms several dense groups. For additional clusterering study, this is encouraging. Furthermore, some spread-out or isolated points imply the existence of outliers or uncertain forms that might cross several groups. The PCA scatter plot just shows the reduced form of the dataset; it does not yet show any clustering. Still, this preview is worth looking at since it shows how PCA transformation increases interpretability. High-dimensional data such as this cannot be seen before PCA. All things considered, the scatter plot shows that PCA successfully detects significant structure in the data

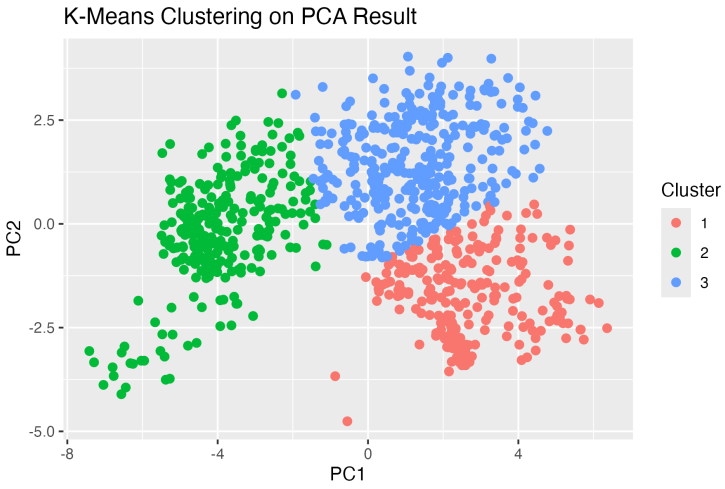
and supports using geometric proximity to define groups in order to guide clustering methods including K-Means.

- **Elbow Method for Optimal Clusters**



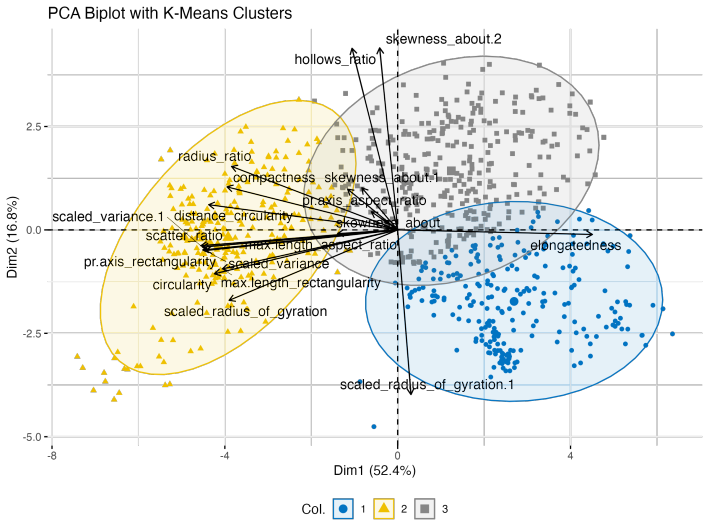
Among the most often used approaches to find the ideal number of clusters (k) in K-Means clustering is the Elbow Method. Plotting the Total Within-- Cluster Sum of Squares (WSS) against different values of k helps one to do this. The WSS naturally lowers as k rises since more clusters lets every point be closer to their designated cluster centre. But beyond a certain point, the improvement in WSS starts to fade and the graph forms a "elbow." The Elbow Plot amply illustrates in our study a sharp drop in WSS from k=1 to k=3 followed by a more slow fall later. This elbow at k=3 implies that three clusters offer the best trade-off between model complexity and clustering compactness. Beyond this, adding more clusters would only somewhat increase cohesiveness while perhaps lowering interpretability and computing cost. Furthermore noteworthy is the Elbow Method's visual and heuristic nature; it should be backed by more validation including silhouette analysis. But given the alignment of this elbow point with the best interpretability in the PCA visualisations and silhouette scores, we have great faith in choosing k=3 as the ideal number of clusters.

- **K-Means Clustering Result (PCA Reduced)**



K-Means clustering was used on the first two main components of the standardised dataset following a k=3 optimal number of clusters determination. The resultant graph shows every vehicle silhouette as a point, coloured depending on the cluster assignment (1, 2, or 3). About 69% of the variance in the dataset is captured by the PC1 vs PC2 axes, hence this plot provides a significant representation of the data grouping in reduced space. Though some overlap at the margins, the three clusters seem to be well-separated with distinct spatial limits. This implies that PC1 and PC2's features are efficient in differentiating between groups of vehicles having comparable geometric traits. Moreover, the balanced density and distribution of points among the clusters show that the model does not suffer from strong inclination towards one group. This clustering outcome confirms the efficiency of PCA as a pre-processing phase. PCA makes K-Means clustering more interpretable and improves its quality by lowering noise and pointless variation. Furthermore demonstrating the practical relevance of the model is the close alignment of the clustering structure with actual vehicle variations in silhouette geometry.

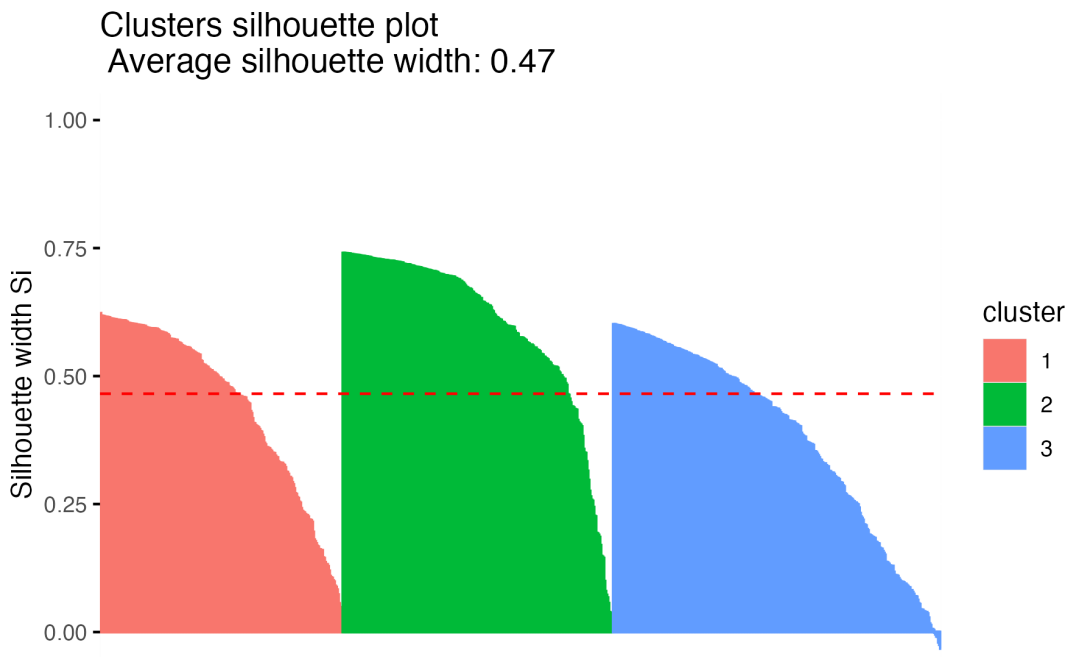
- **PCA Biplot with Cluster Overlay**



One of the most effective visual aids available for understanding the outcomes of dimensionality reduction and clustering is the PCA biplot including overlaid cluster assignments. While original features are shown as arrows, indicating their direction and strength relative to the PCA axes, in this plot the clusters are colour-coded depending on K-Means output. This dual-view helps one to have a rich interpretation of cluster structure as well as individual feature impact. For instance, Cluster 1 shows strong correlation with elongatedness since it lines along the positive axis of PC2. Location along the positive PC1 direction, Cluster 2 is affected by scatter_ratio, compactness, and radius_ratio. Cluster 3, meantime, is more centralised and shaped by a mix of less important factors like hollows_ratio or skewness_about. The direction indicates how strongly a feature separates clusters; the length of each arrow shows how much the feature contributes to the PCA space. Features pointing in the same direction as a cluster centre most certainly help to generate that group. When labelling groups according to their dominant

characteristics, this graph is quite helpful. It converts the abstract PCA space into meaningful domain language—that is, "elongated vehicles" or "compact geometries."

- **Silhouette Score Plot**



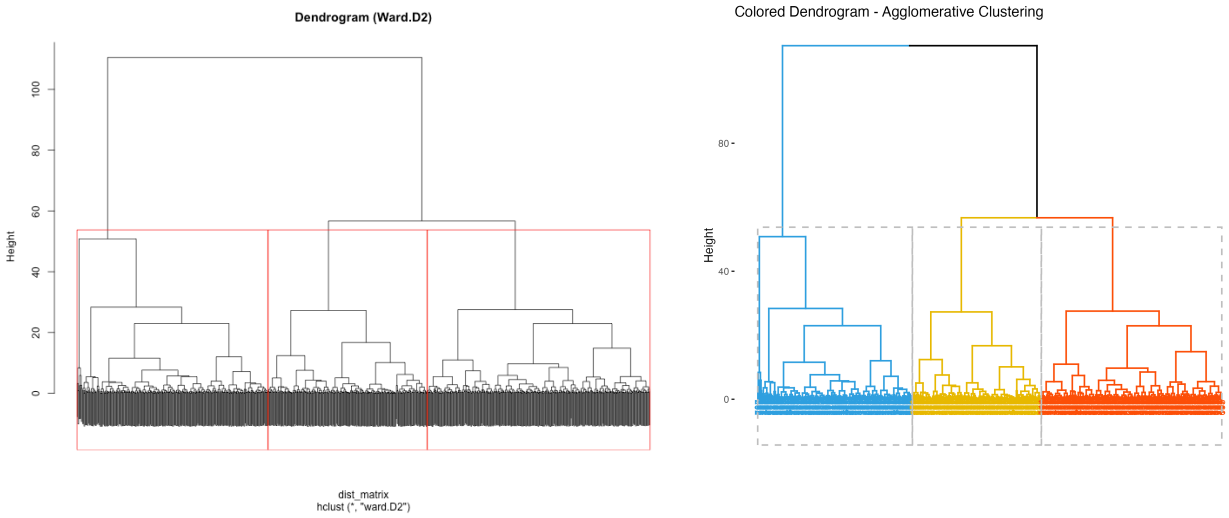
By computing how well each observation fits its assigned cluster relative to others, the Silhouette Plot offers a numerical assessment of clustering quality. Every bar shows the shadow width of a single observation; the average value shows up as a red dashed line. Our analysis yields an average silhouette score of 0.47, regarded as rather strong. This suggests that, generally speaking, points lie closer to their own cluster centres than to other clusters, so indicating coherent and significant groupings. With most of its points above 0.6, the green cluster (Cluster 2) exhibits the best internal cohesiveness. Consequently, the cluster is small and clearly apart from the others. With most values above 0.4, the red and blue clusters (Cluster 1 and 3) exhibit rather more overlap but still have reasonable silhouette widths. There are quite few points less than 0, which would suggest misclassification. This proves not only visually appealing but also statistically valid k=3. Strongening our confidence in the clustering model, the silhouette analysis supports the results of the Elbow Method and cluster scatter plot. When ground truth labels are absent or useless during training, it is an indispensable instrument for verifying unsupervised models.

4. Agglomerative Hierarchical Clustering (AHC)

Overview:
Beginning in each observation in its own cluster, Agglomerative Hierarchical Clustering (AHC) is a bottom-up clustering technique whereby pairs of clusters are merged successively depending on a linkage criterion. In this work, the PCA-transformed dataset (with first three principal components) was subjected to AHC, so capturing most of the variance of the dataset. A distance matrix was built using the Euclidean distance metric; Ward.D2 was chosen as the linkage technique. Particularly useful when the objective is to generate clusters of roughly equal size with low internal dissimilarity, Ward's approach seeks to minimise the overall within-cluster variance.

AHC differs from K-Means in not depending on a predefined cluster count. Rather, it generates a dendrogram—a tree-like construction depicting the merging process—which the user can "cut" at the level producing the intended number of clusters. The dendrogram also makes visual study of how clusters develop and where the biggest height gaps exist possible, so guiding the choice of the proper number of clusters. Balancing interpretability, silhouette score, and visual separation in the PCA space, this study found that a cut at three clusters (k = 3) was best.

- **Dendrogram Interpretation and Number of Clusters**



Using Ward.D2 linkage, the dendrogram (Figure: dendrogram.png) produced exhibits a hierarchical structure of merges based on Euclidean distances in the PCA-reduced space. Clusters arise by combining the two closest groups as we climb the dendrogram from leaves—individual points—to the root. Every merger's vertical height shows the differences between merged clusters. Cutting the dendrogram just below such jumps is a reasonable approach to estimate the number of clusters since large vertical jumps imply that the merged clusters were quite different.

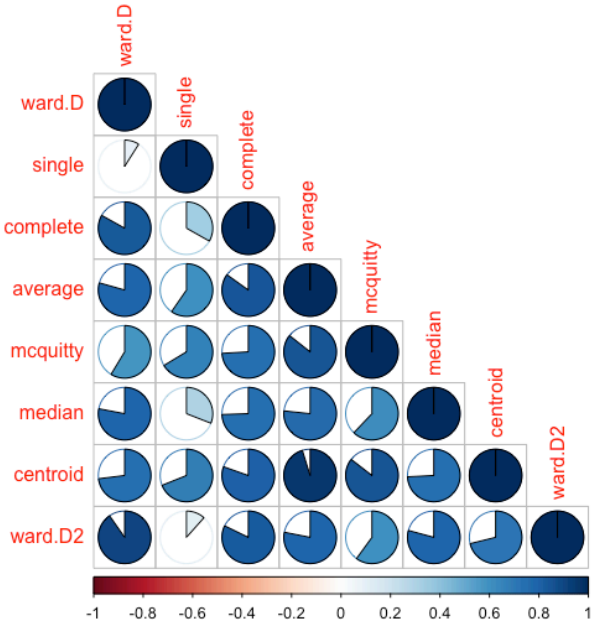
In this instance, the dendrogram shows a clear height increase around three main branches, so supporting a cut at k = 3 clusters. The red rectangles sketched with rect.hclust() visually support this cut. Showing each last cluster in an other color—blue, yellow, and orange—the coloured dendrogram (Figure: fviz_dend_colored.png) offers a more interpretable visual. Further validating the choice of three clusters are their rather balanced size and compact internal structures. Strong validation from two independent unsupervised techniques helps this result to closely match the K-Means clustering result.

- **Silhouette Analysis for Hierarchical Clustering**

AHC's silhouette plot (Figure: silhouette_hc.png) provides quantitative support of the clustering outcome. About 0.48, the average silhouette width is somewhat more than the 0.47 score obtained from K-Means clustering. Every bar in the plot marks a single observation; its width indicates how closely that observation fits within its cluster. A value near +1 indicates the object is well-clustered; values near 0 indicate uncertainty; negative values suggest possible misclassification.

Few points in this plot fall below zero, suggesting a strong clustering; most points show silhouette widths above 0.4. While Cluster 2 (green) shows rather more dispersion, Cluster 1 (red) and Cluster 3 (blue) have closely spaced points. Still, all three clusters keep enough internal cohesiveness and outside separation. Validating k = 3 as the best number of clusters, the average silhouette score supports the visual results from the dendrogram and PCA scatter plots. Strong confidence in the clustering structure found by AHC is given by this consistency between visual and numerical measures.

- **Justification of Linkage Method: Ward.D2**



The formation of clusters in hierarchical clustering is much influenced by the linkage technique. Ward.D, ward.D2, single, complete, average, mcquitty, median, and centroid were eight linkage techniques assessed. The aim was to choose the approach that, following clustering, most maintains the original pairwise distances in the data. This was calculated by means of the cophenetic correlation coefficient, which contrasts the dendrogram's implied distances with the original distance matrix.

Figure Rplot.png displays the cophenetic correlation matrix showing ward. D2 and ward.Represented by dark-coloured and nearly full pie charts, D attained the highest correlations with the original data. Among these, ward. D2's best alignment with the distance matrix justifies its chosen as the ideal linkage technique. By means of the smaller or partially filled circles, techniques such as single and mcquitty revealed rather poor alignment.

Particularly useful for data with continuous variables, Ward.D2 is known to generate compact, spherical clusters—qualities perfect for the PCA-transformed feature space. Thus, depending on the clustering goal, its use in this study is theoretically appropriate as well as empirically justified by correlation measures.

Summary

Using unsupervised learning methods, mostly Principal Component Analysis (PCA), K-Means Clustering, and Agglomerative Hierarchical Clustering (AHC), this question offers a complete

study of a high-dimensional dataset of vehicle silhouettes. The goal was to find natural groups in the data and assess clusterering quality using both statistical and visual approaches.

The first phase of the study concentrated on data preparation and found notable feature variance. Especially scaled_variance.1 dominated the scale, hence standardising is quite important. Using Predictive Mean Matching (PMM), missing values were handled; outliers were winsorized to avoid distortion in distance-based computations. Meaningful downstream analysis was made possible by this ordered and clean dataset.

PCA was used to solve linked features and high dimensionality. Given almost 70% of the variance explained by the first two components, their application for clustering and visualisation is justified. Illustrated different groupings and dominant traits including elongatedness, radius_ratio, and compactness the PCA scatter plot and variable contribution plot showed. PCA also guaranteed that the simplified and more interpretable feature space of clustering techniques ran on was ensured.

K-Means clustering was performed with the number of clusters found by the Elbow Method using the PCA-reduced data and validated with the Silhouette Score. Both pointed out that $k = 3$ was ideal. With a modest silhouette width of 0.47, the resulting cluster assignments exhibited good separation and cohesiveness.

Agglomerative Hierarchical Clustering (AHC) was also used employing the Ward.D2 linkage method and Euclidean distance to validate and compare results. Preserving the structure of the original distance matrix, the cophenetic correlation matrix validated Ward.D2 as the optimal fit. Showing evident vertical jumps and natural branch separations, the dendrogram and coloured dendrogram helped to support the choice of three clusters. Confirming the robustness of the clustering, the AHC silhouette score of 0.48 was rather better than K-Means.

Group structure was highly consistent across both clustering techniques and clusters matched the dominant PCA features. Visually as well as statistically, cluster compactness, interpretability, and coherence were validated.

Conclusion

Analysing the geometric structure of vehicle silhouettes proved quite successful using PCA and unsupervised clustering techniques combined. The results show the need of validating clusterering decisions using several approaches and the ability of dimensionality reduction in simplifying difficult datasets. K-Means and AHC agreed on a comparable data interpretation, so strengthening the validity of the results. Extensive application of this analytical framework in other fields with high-dimensional continuous features will provide insightful analysis in exploratory data mining.