# Car Insurance Customer Segmentation

**Executive Summary:**

This project analyzed data on policyholders of a car insurance company. Exploratory data analysis was performed to analyze the distributions, correlations, and skew of all the features. Data preprocessing included Yeo-Johnson transformations on skewed features, one-hot encoding, normalization, and principal component analysis (PCA) to reduce the number of features from 68 to 15. Various clustering models were fitted and evaluated.

The K-means model was determined to be best, identifying six distinct clusters. Key metrics showed that the clusters were compact and well-separated (silhouette score = 0.564, Davies-Bouldin index = 0.954). Feature importance conducted with PCA showed that where policyholders lived and policy tenure were most important.

The six clusters as well as business strategies with each are summarized below.

| Customer Profile | Business Interpretation |
|---|---|
| Small City Drivers with Modern Cars | Stable, safety-conscious, loyal; retain with loyalty discounts or bundled coverage |
| New Customers with Unsafe Cars | High-risk, prone to churn; apply higher base rates for lower-safety cars |
| Drivers with the Safest and Largest Cars | Premium, safety-conscious; target for loyalty and upselling |
| Drivers with Basic, Manual, Diesel Cars | Middle-market, budget-conscious; good for retention campaigns |
| Reliable Customers in Small Cities with Basic Cars | Stable and steady-income; offer renewal-based loyalty programs |
| Big City Drivers with Basic, Modern Cars | Low-risk, profitable; ideal for cross-selling add-ons |

**Introduction:**

Insurance companies manage diverse customers with varying risk profiles, vehicle types, and claim tendencies. Treating all customers uniformly can lead to suboptimal marketing, retention, and pricing strategies. Customer segmentation helps insurers better understand behavioral and demographic differences within their portfolio and tailor approaches to each group.

The main objectives of this project are:

1. To uncover natural clusters within the insurance customer base using unsupervised learning models.
2. To translate clusters into actionable business insights, such as charging higher base rates for high-risk segments.

**Data Description:**
The "Car Insurance Claim Prediction" dataset was obtained from Kaggle. It contains data on 58,592 policyholders and 44 features. The dataset contains key attributes about the policyholders:

- **Demographics:** Age of policyholder, area cluster, population density
- **Vehicle Information:** Vehicle make, model, segment, fuel type, NCAP rating, and safety features
- **Policy Attributes:** Policy tenure, credit limit, number of claims, claim amount
- **Binary Safety Indicators:** Indicators for brake assist, central locking, power steering, etc.

The dataset also contains a target variable indicating whether the policyholder files a claim in the next six months or not. The target variable was not used to make the clusters but was used in post-hoc validation.

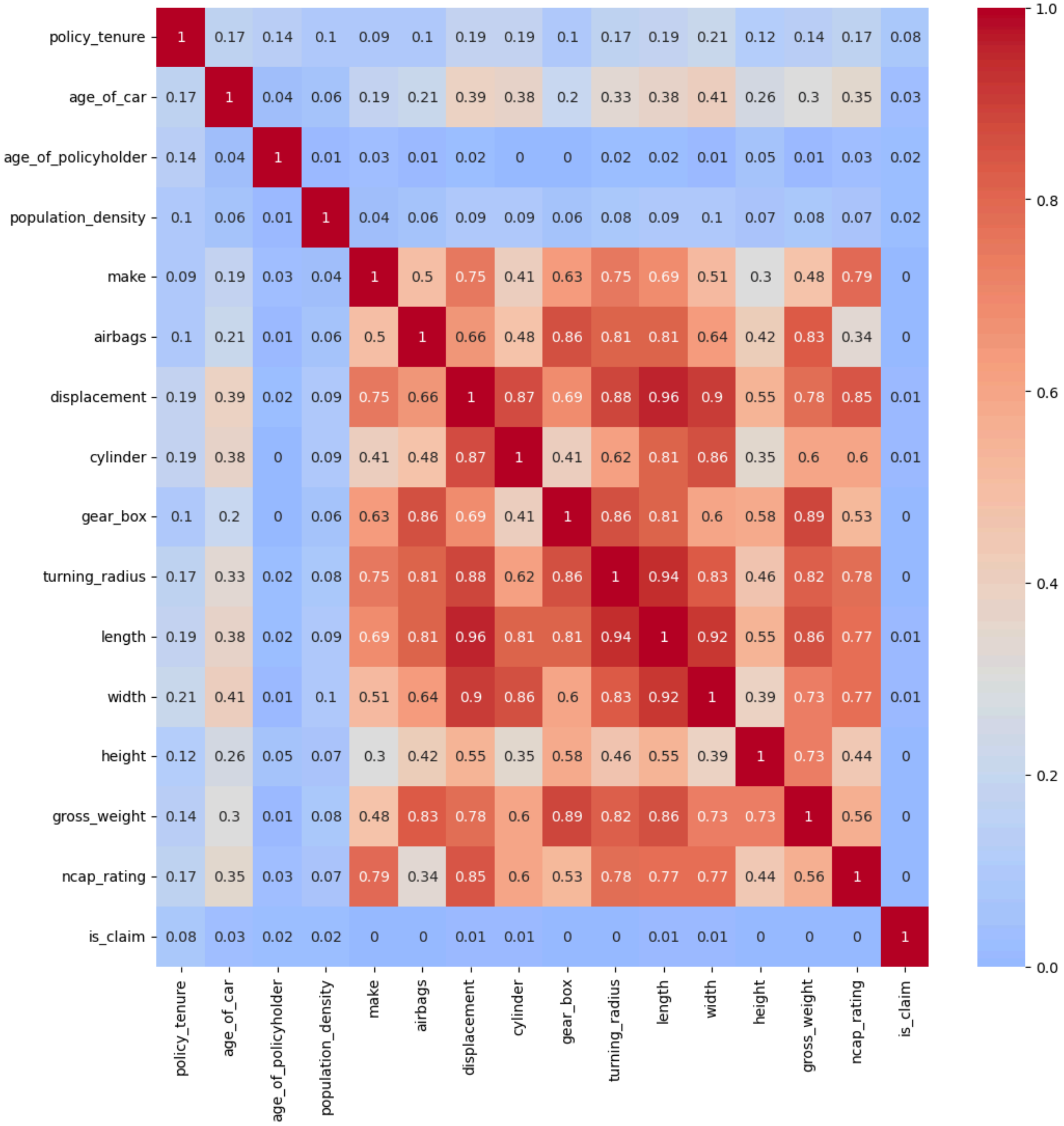
**Methodology:**

1. Exploratory Data Analysis
    - Visualized data distributions of all numeric features with histograms
    - Generated pair plots with all continuous features
    - Constructed a correlation matrix and heatmap between all features
2. Data Preprocessing
    - One-hot encoded nominal categorical features, resulting in a total of 69 features
    - Normalized features using MinMaxScaler
    - Performed dimensionality reduction with PCA to reduce the number of features to 15 while maintaining over 95% of the global variance
    - Identified the most important features with PCA to be where policyholders lived and policy tenure
3. Clustering Models
    - K-means clustering

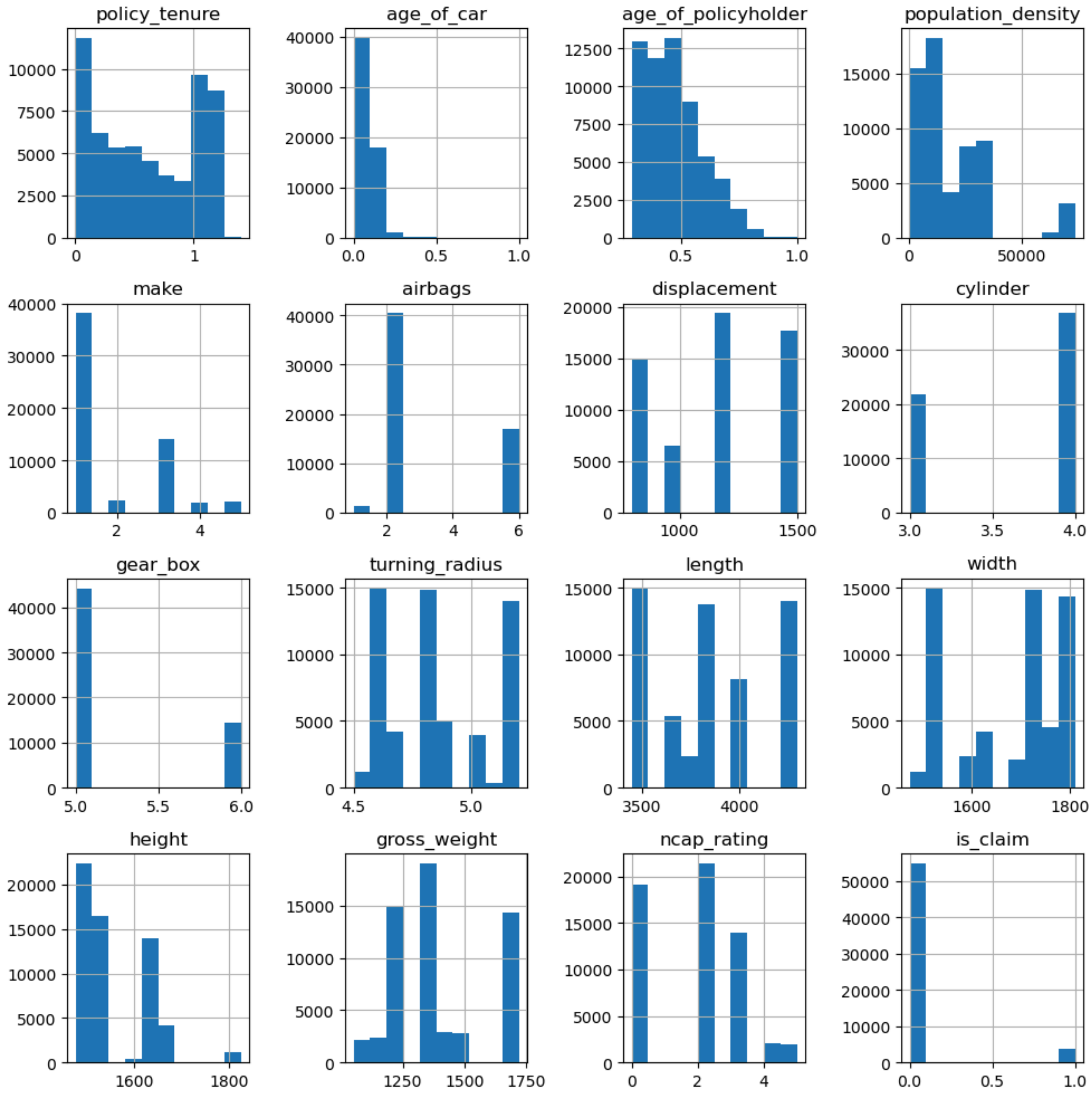
        - Found ideal number of clusters to be 6 by plotting inertia vs. K and using the elbow method
        - The key metrics for the model are a silhouette score of 0.564, Davies-Bouldin index is 0.954, and Calinski-Harabasz score 44,200
    - Hierarchical Agglomerative Clustering
        - Fitted model using only 1,000 random samples for computational efficiency
        - Found ideal number of clusters to be 6 using a dendrogram
        - The key metrics for the model are a silhouette score of 0.605, Davies-Bouldin index is 0.849, and Calinski-Harabasz score 1,060
    - DBSCAN
        - Fitted model using only 1,000 random samples for computational efficiency
        - Found ideal number of clusters to be 5 by testing various epsilon values
        - The key metrics for the model are a silhouette score of 0.543, Davies-Bouldin index is 0.786, and Calinski-Harabasz score 641
    - Determined K-means model to be best based on evaluation metrics, computational efficiency, and scalability

4. Cluster Profiling
    - Computed feature means per cluster for numeric features and feature modes per cluster for categorical features to identify differentiating characteristics
    - Visualized key differences between clusters using bar graphs and boxplots

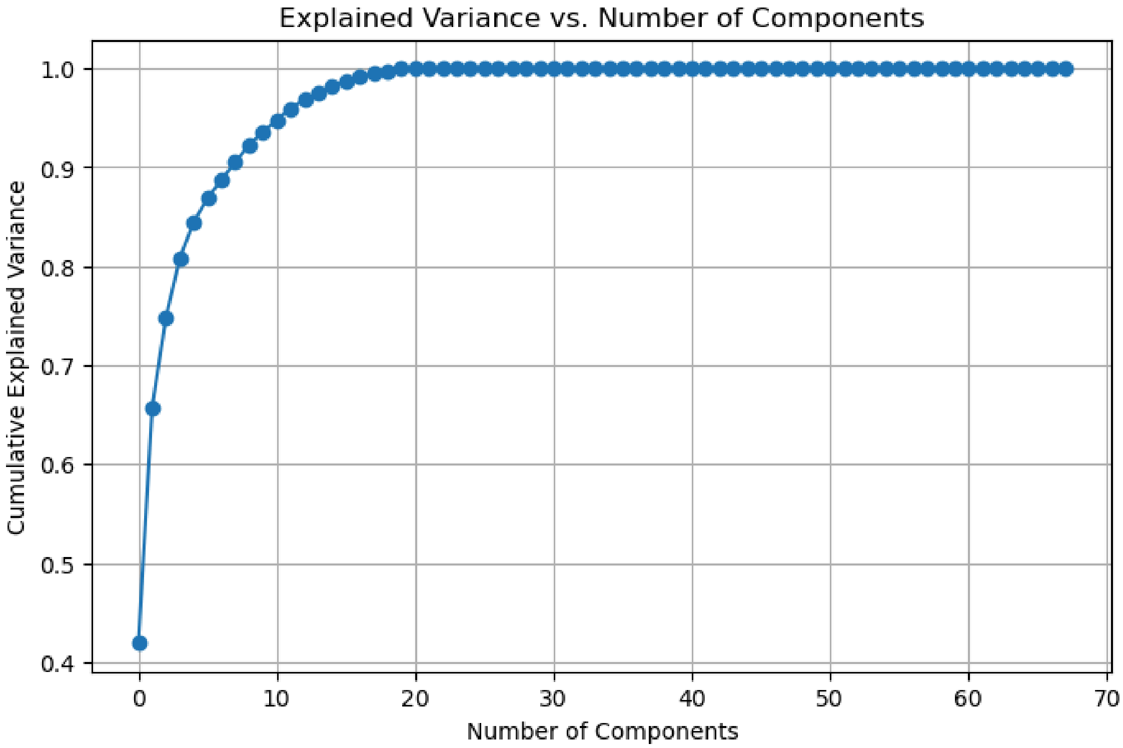
**Results:**
Exploratory data analysis revealed that many features were highly correlated.



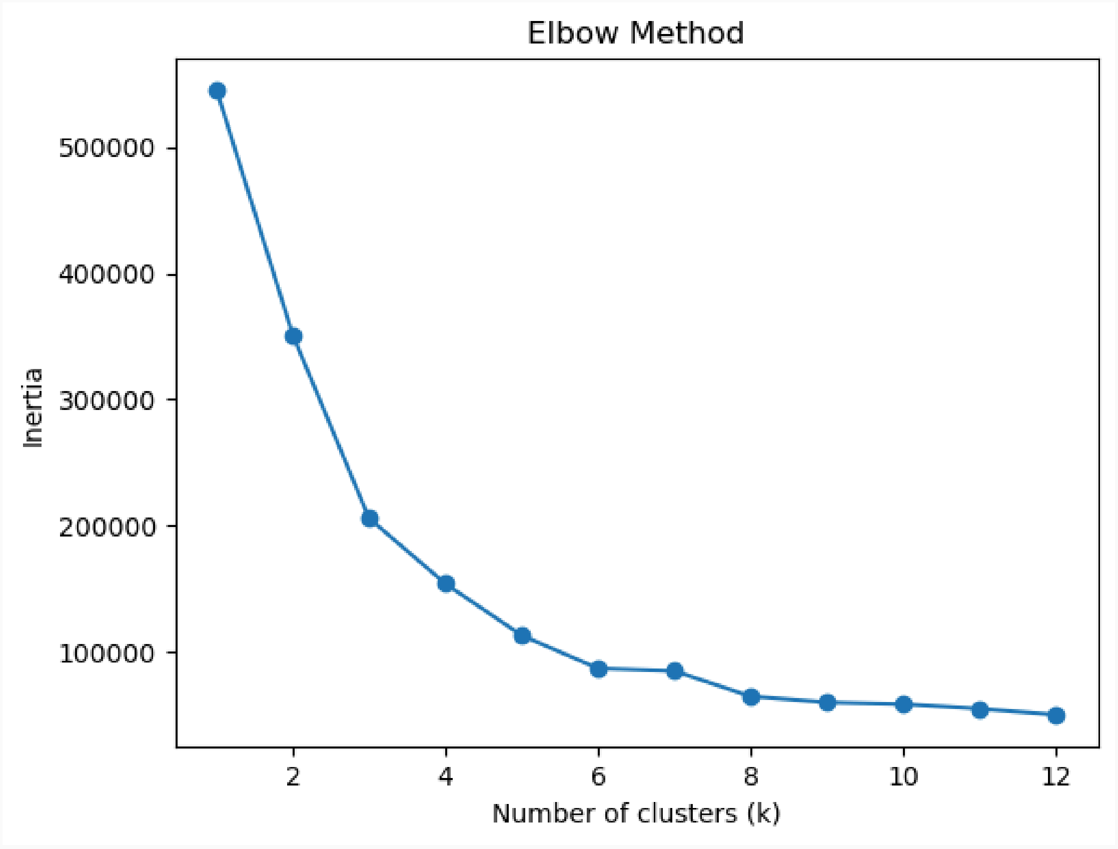
Several features showed skewed distributions.



Dimensionality reduction with PCA was used to reduce the number of features to 15 while maintaining over 95% of the global variance.

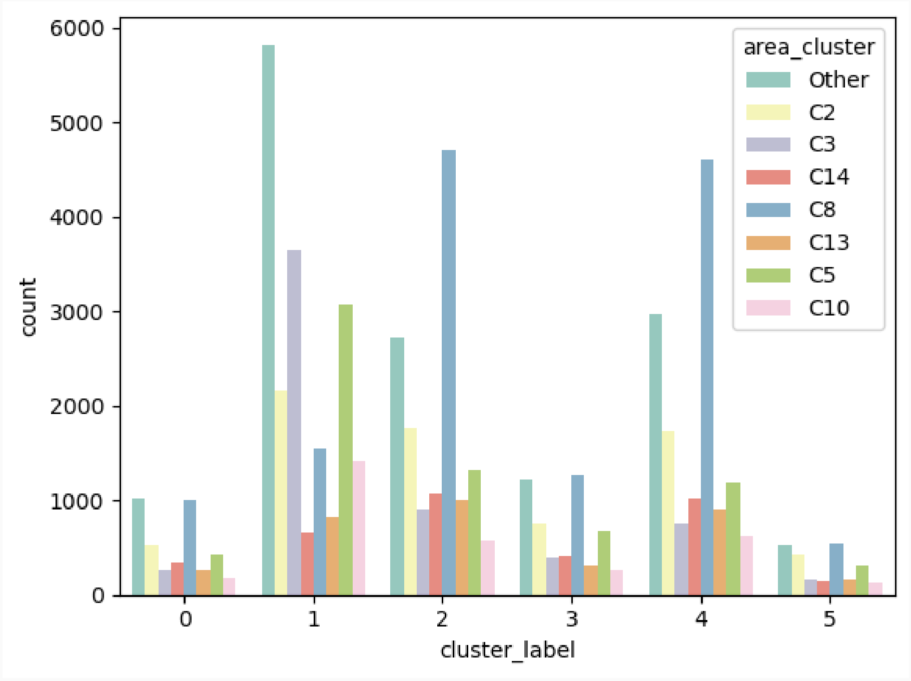


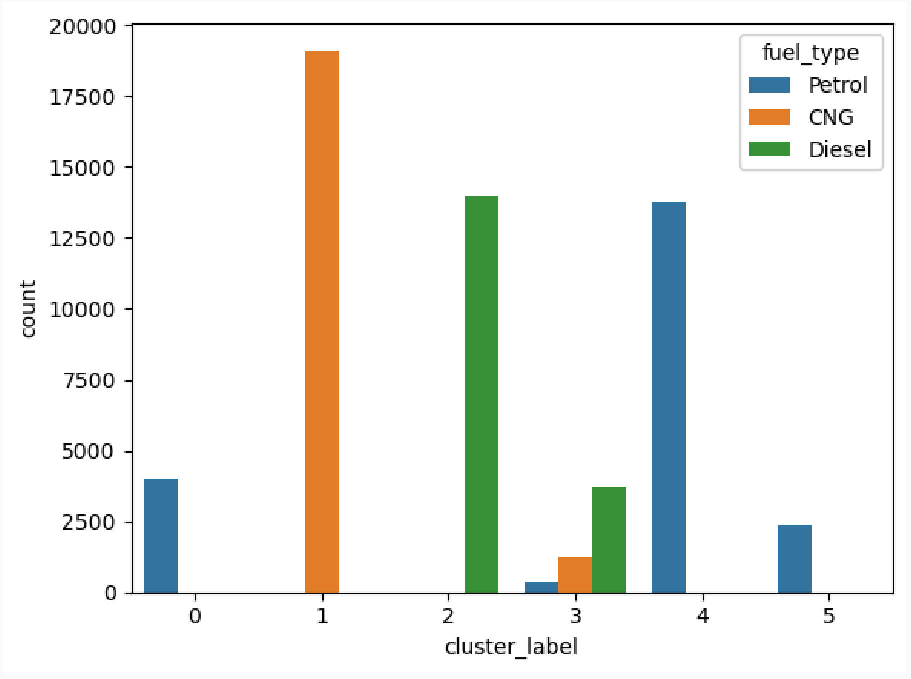
The clustering process with K-means identified six distinct customer segments.

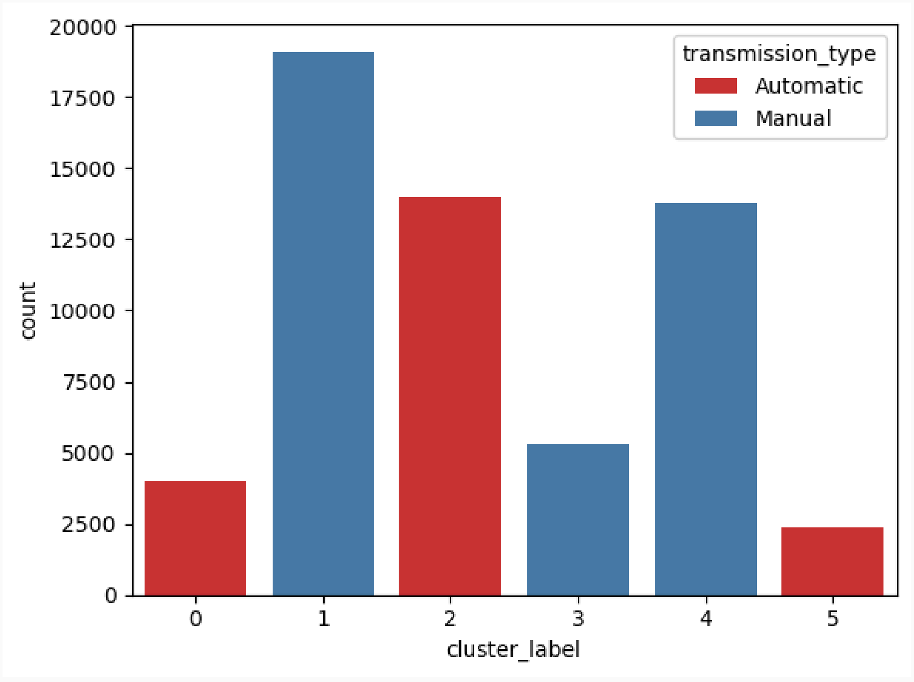


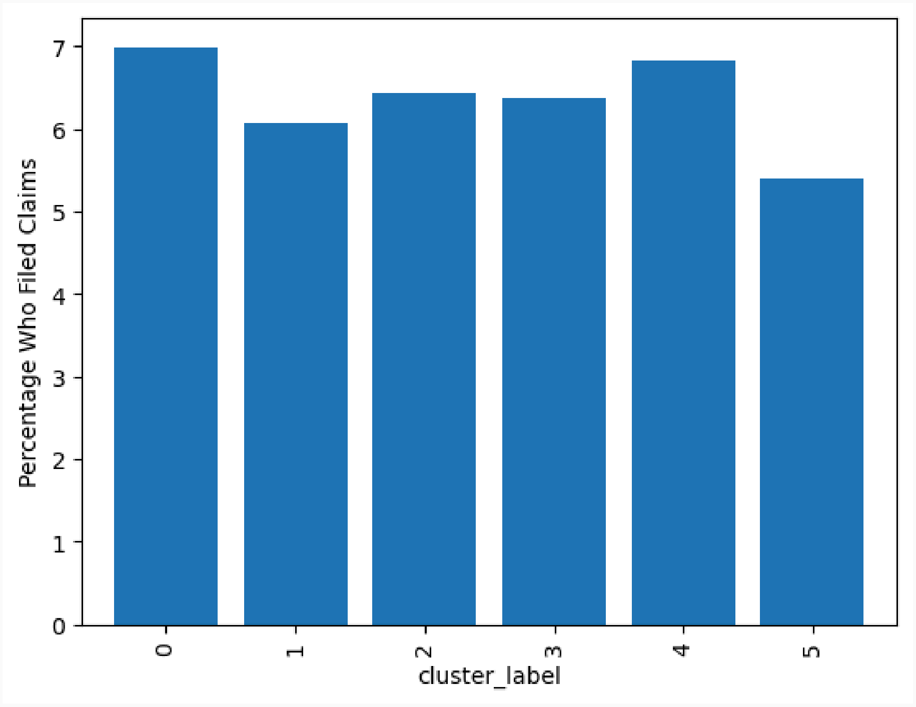
The silhouette score is 0.564, indicating that clusters are reasonably well-separated. The Davies-Bouldin index is 0.954, indicating that the clusters are compact and distinct. The Calinski-Harabasz score is 44,200, indicating high between-cluster variance and low within-cluster variance.

Key differences between clusters of various features were visualized.









6.9% of policyholders belong to cluster 0, 32.6% belong to cluster 1, 23.9% belong to cluster 2, 9.0% belong to cluster 3, 23.5% belong to cluster 4, and 4.1% belong to cluster 5.

**Discussion:**

The six clusters are described below with possible business strategies for each.

### Cluster 0: Small City Drivers with Modern Cars

- Car Safety: High number of airbags, more safety features on the car despite lower NCAP rating
- Vehicle Type: Petrol fuel, automatic, electric steering
- Demographics: Small cities

Business Interpretation: Likely stable, safety-conscious customers with good vehicles and loyalty to the insurer; retain with loyalty discounts or bundled coverage.

**Cluster 1: New Customers with Unsafe Cars**

- Car Safety: Low number of airbags, less safety features on the car, and lower NCAP rating
- Vehicle Type: CNG fuel, manual, power steering, smaller cars, newer cars
- Demographics: Large cities, shortest average policy tenure

Business Interpretation: High-risk customers who are prone to churn; offer higher base rates for lower-safety cars.

**Cluster 2: Drivers with the Safest and Largest Cars**

- Car Safety: High number of airbags, more safety features on the car, and higher NCAP rating
- Vehicle Type: Diesel fuel, automatic, power steering, larger cars
- Demographics: Small cities

Business Interpretation: Premium, safety-conscious customers driving higher-end vehicles; good targets for long-term loyalty and upselling.

**Cluster 3: Drivers with Basic, Manual, Diesel Cars**

- Car Safety: Low number of airbags, high NCAP rating
- Vehicle Type: Diesel fuel, manual, electric steering
- Demographics: Medium-sized cities

Business Interpretation: Middle-market customers, budget-conscious, and moderate risk; good candidates for retention campaigns.

**Cluster 4: Reliable Customers in Small Cities with Basic, Manual Cars**

- Car Safety: Low number of airbags, moderate NCAP rating
- Vehicle Type: Petrol fuel, manual, electric steering
- Demographics: Small cities, highest average policy tenure

Business Interpretation: Stable, steady-income customers if engaged properly; offer renewal-based loyalty programs.

**Cluster 5: Big City Drivers with Basic, Modern Cars**

- Car Safety: Low number of airbags, moderate NCAP rating
- Vehicle Type: Petrol fuel, automatic, electric steering
- Demographics: Large cities

Business Interpretation: Low-risk, low-premium but profitable base segment; good candidates for cross-selling add-ons (zero depreciation, theft protection).

**Conclusion and Next Steps:**
This project successfully segmented insurance customers into six actionable groups using unsupervised learning. The findings demonstrate how clustering can inform targeted marketing, pricing, and risk strategies within the insurance sector.

Possible next steps are:

1. To incorporate additional behavioral data such as payment history or policy renewals to enrich segmentation.
2. Integrate predictive modeling to estimate claim probability per cluster for proactive risk management.