Revue-IRS

Revue Internationale de la Recherche Scientifique (Revue-IRS)

ISSN: 2958-8413

Vol. 3, No. 3, Juin 2025

A Hybrid Approach to Vehicle Price Prediction: Combining PCA and Supervised Learning

Alain M. KUYUNSA[1], Alidor M. MBAYANDJAMBE[2], Grevi B. NKWIMI[3], Darren Kevin T. NGUEMDJOM[4], Dorotha K. TSHIBOLA[5], Jacques B. TSHINGAMBU[6], Blanchard M. KANGULUMBA[7]

[1,2,3,5]Department of Business Informatics and Business English, Faculty of Economic and Management Sciences, Kinshasa, University of Kinshasa, DR. Congo

[2,6]Department of Mathematics, Statistics and Computer Science, Faculty of Science and Technology, University of Kinshasa, Kinshasa, DR Congo.

[2,3]Institut Francophone International (IFI), Vietnam National University, Hanoi, Vietnam

[7]Faculty of Science and Technology, Kwango University, Kwango, DR Congo.

Abstract: Accurately predicting car prices is a complex task that involves analyzing multiple interacting factors. This study proposes a comprehensive methodology that integrates dimensionality reduction, clustering, and supervised learning to enhance the predictive accuracy of car price models. Using a dataset of 399 vehicles described by ten numerical and categorical features, we first apply Principal Component Analysis (PCA) to reduce dimensionality while preserving essential information. Next, K-means clustering is used to identify natural groupings within the data, revealing meaningful market segments. Finally, we compare the predictive performance of two supervised learning algorithms: Multiple Linear Regression (MLR) and Support Vector Machines (SVM). The results demonstrate that MLR achieves superior performance, with a coefficient of determination (R^2) of 0.92 and a Root Mean Square Error (RMSE) of 2,310.12, compared to SVM's R^2 of 0.85 and RMSE of 3,212.14. These findings underscore the value of combining exploratory analysis with predictive modeling and suggest that MLR remains highly effective when relationships among variables are largely linear. This integrated approach provides valuable insights for stakeholders in the automotive and insurance sectors seeking to assess vehicle value accurately.

Keywords: Car price prediction, Principal Component Analysis, K-means clustering, Multiple Linear Regression, Support Vector Machines, Machine learning

Digital Object Identifier (DOI): https://doi.org/10.5281/zenodo.**15657326**

1 Introduction

Accurately estimating vehicle prices is a critical challenge for stakeholders such as manufacturers, dealerships, insurers, and buyers. With the advent of big data and machine learning, modern approaches offer data-driven alternatives that surpass traditional methods relying on heuristics or predefined rules. Historically, vehicle valuation relied on hedonic pricing models, in which prices are explained as a function of various vehicle attributes (Rosen, 1974). Although these models provide interpretable results, they often assume linear and independent relationships between variables assumptions that may not hold in real-world automotive markets. As car features

become more diverse and consumer preferences more nuanced, machine learning methods have emerged as effective tools for modeling such complex, multivariate relationships (Klein et al., 2020).

In particular, factor analysis techniques like Principal Component Analysis (PCA) have been widely used to reduce dimensionality and highlight the latent structure in datasets without losing significant information (Jolliffe & Cadima, 2016). PCA helps to uncover patterns that are not immediately visible through traditional descriptive analysis and can enhance the interpretability and performance of subsequent modeling steps. Clustering methods such as K-means and Hierarchical Agglomerative Clustering (HAC) are often employed to identify homogeneous groups of vehicles based on multiple characteristics (Tan, Steinbach, & Kumar, 2019). These unsupervised techniques facilitate market segmentation, which is crucial for developing targeted pricing strategies and understanding market structure (Chaturvedi et al., 2001).

On the predictive side, Multiple Linear Regression (MLR) remains a baseline due to its simplicity and robustness, especially when relationships between features and the target variable are largely linear (James et al., 2013). Meanwhile, Support Vector Machines (SVMs) offer strong performance in high-dimensional settings and have been successfully applied in various regression and classification problems, including price prediction (Drucker et al., 1997; Smola & Schölkopf, 2004), (Alidor Mbayandjambe et al.,2025). Nevertheless, SVMs can be sensitive to parameter tuning and may not outperform simpler models in relatively small datasets.

This paper presents an integrated methodological framework for car price prediction that combines factor analysis, clustering, and supervised learning. The major contributions of this work are as follows:

- **Development of a Hybrid Analytical Pipeline**: We propose a structured approach combining Principal Component Analysis (PCA), K-means clustering, and predictive modeling (Multiple Linear Regression and Support Vector Machines) to enhance the accuracy and interpretability of car price estimation models.
- **Empirical Comparison of Predictive Models**: The performance of multiple linear regression and support vector machines is thoroughly evaluated using standard regression metrics.
- **Exploratory Market Segmentation via Clustering**: Through K-means and Hierarchical Agglomerative Clustering (HAC), we identify meaningful market segments, reinforcing the importance of unsupervised learning in automotive market analysis.
- **Dimensionality Reduction for Insightful Feature Engineering** : PCA reveals latent structures in the vehicle dataset, facilitating variable selection and dimensionality reduction without significant loss of information.

The remainder of this paper is structured as follows:

- **Section 2 – Related Works** : Reviews existing literature on vehicle price prediction, highlighting methods involving dimensionality reduction, clustering, and supervised learning.
- **Section 3 – Methodology** : Describes the dataset, preprocessing techniques, PCA analysis, clustering algorithms, and the regression models used for prediction.
- **Section 4 – Experimental Results** : Presents the outcomes of PCA, clustering, and predictive modeling. Includes a comparative analysis of multiple linear regression and SVM in terms of predictive performance.
- **Section 5 – Discussion** : Analyzes the implications of the findings, discusses the relative strengths of each method, and addresses the limitations of the study.
- **Section 6 – Conclusion** : Summarizes the main contributions and outlines potential future directions, including integration of temporal features and exploration of ensemble learning techniques.

2 Related works

The problem of car price prediction has evolved from traditional econometric models to more advanced data-driven approaches, motivated by the growing complexity of vehicle characteristics and consumer preferences. Early studies employed hedonic pricing models, which explain vehicle prices as a function of features such as brand, power, fuel type, and mileage (Rosen, 1974). These models are based on linear assumptions and often fail to capture more intricate relationships present in real-world datasets.

In recent years, machine learning (ML) methods have gained prominence in this domain. Uma and Uma (2022) applied ensemble approaches, including Random Forest, SVM, and Artificial Neural Networks, to predict car prices using data from Kaggle, achieving an R^2 of up to 0.87. Similarly, Vaneesha et al. (2024) conducted a comparative analysis of K-Nearest Neighbors (KNN) and Support Vector Regression (SVR), reporting R^2 values of 0.83 and 0.80 respectively, on a dataset of used vehicles from India. These studies confirm that ML techniques can outperform classical regression when relationships between variables are non-linear and multi-dimensional.

To enhance interpretability and reduce dimensionality, Principal Component Analysis (PCA) has been employed. Jolliffe and Cadima (2016) highlight the role of PCA in identifying latent variables and removing redundancy. In predictive contexts, PCA has been used to reduce overfitting and computational complexity. For example, in a study on highway construction cost prediction, PCA was used to preprocess input variables before training a Least Squares SVM (LSSVM), resulting in improved generalization and accuracy (Shi et al., 2022).

Unsupervised learning techniques like K-means clustering and Hierarchical Agglomerative Clustering (HAC) have also been applied to identify natural segments in vehicle datasets. Tan, Steinbach, and Kumar (2019) emphasize the value of clustering in understanding hidden patterns and segmenting the market, which can subsequently inform targeted predictive models. Despite its potential, few studies integrate clustering and supervised learning in a unified framework for price prediction, representing a gap that this study aims to address.

Overall, previous work demonstrates the power of combining dimensionality reduction and machine learning in car price prediction. However, most approaches focus on isolated techniques. There is still a need for integrated pipelines that jointly leverage factor analysis, unsupervised segmentation, and predictive modeling to provide both interpretability and predictive accuracy.

3. Methodology

This study proposes a comprehensive pipeline for vehicle price prediction by integrating data preprocessing, dimensionality reduction, unsupervised clustering, and supervised regression models. The approach aims to enhance model interpretability and predictive performance by isolating underlying patterns in vehicle attributes before training regression models. Figure 1 illustrates the overall architecture of the methodology.

3.1 Data Preprocessing

The dataset, obtained from a vehicle listing platform, includes features such as brand, model, year, mileage, fuel type, transmission, and engine capacity. The preprocessing phase involves:

- Missing Value Handling: Features with more than 40% missing values were discarded. For numerical features with missing values, we applied median imputation as follows:

$$x_i = \begin{cases} x_i, & \text{if } x_i \text{ is observed} \\ \text{median}(X), & \text{otherwise} \end{cases}$$

- Encoding: Categorical features were transformed using one-hot encoding, and ordinal features such as condition were label encoded.
- Outlier Treatment: We applied **Z-score filtering** to detect and remove outliers in numerical distributions such as price and mileage. The Z-score is defined as:

$$Z_i = \frac{x_i - \mu}{\sigma}$$

where μ is the mean and σ\sigmaσ is the standard deviation. Observations with |Zi|>3 were considered extreme outliers.

- Feature Normalization: We used Min-Max normalization to scale numeric features to the range [0,1], a crucial step to ensure compatibility with PCA and clustering algorithms:

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

3.2 Dimensionality Reduction with PCA

To mitigate multicollinearity and reduce feature dimensionality, **Principal Component Analysis (PCA)** was performed on the normalized data. Let $X \in R^{n \times p}$ be the mean-centered data matrix. The principal components are obtained through eigen decomposition of the covariance matrix:

$$\Sigma = \frac{1}{n} X^\top X, \quad \Sigma v_i = \lambda_i v_i$$

where v_i are eigenvectors (principal components) and λi their corresponding eigenvalues (explained variance). We retained the smallest number k of components satisfying:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j} \geq 0.95$$

These reduced components were later used for clustering and regression tasks to enhance interpretability and computational efficiency.

3.3 Unsupervised Clustering

To segment the dataset into homogeneous subgroups, we applied two unsupervised clustering methods:

- **K-Means Clustering**: The optimal number of clusters kkk was determined using the elbow method and **silhouette score**, where silhouette *s (i)* is computed as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Here, *a (i)* is the mean intra-cluster distance, and *b (i)* is the mean nearest-cluster distance for sample *i*. We initialized centroids using the K-means++ strategy for faster convergence.

- **Hierarchical Agglomerative Clustering (HAC)**: We constructed a dendrogram using Ward's linkage and Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

The dendrogram was truncated at the level where the maximal inter-cluster distance jump occurred, yielding a natural clustering structure.

These clustering techniques enabled a comparative study between global regression models and cluster-specific models trained on subgroups.

3.4 Supervised Regression Models

We trained multiple regression models both on the full dataset and on cluster-specific subsets:

- **Linear Regression (LR)**: Served as the baseline model. The loss function minimized is the mean squared error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Support Vector Regression (SVR): Implemented with a Radial Basis Function (RBF) kernel. Hyperparameters C and γ were optimized using grid search. The decision function is expressed as:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b, \quad K(x_i, x) = \exp(-\gamma \|x_i - x\|^2)$$

- **Random Forest Regression (RFR)**: A tree-based ensemble method known for capturing non-linear interactions. The number of trees and maximum depth were optimized via cross-validation.

We evaluated model performance using the following metrics:

- **Root Mean Square Error (RMSE)**:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **Mean Absolute Percentage Error (MAPE)**:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

3.5 Evaluation and Validation

We adopted 10-fold cross-validation to ensure robust performance estimation. For each fold, the dataset was randomly split into training (80%) and testing (20%) sets. Model performance was aggregated across folds, and standard deviations were reported.

We further compared:

- Models trained on raw features vs. PCA-transformed features.
- Global models vs. cluster-specific models (i.e., using K-means or HAC labels as sub-groups).
- Computational time and memory usage for each modeling strategy.

3.6 Overall Approch

As shown in Figure 1, the proposed pipeline comprises five key phases: (1) data preprocessing, including missing value imputation, outlier detection, and normalization; (2) dimensionality reduction using PCA to capture the most informative features; (3) unsupervised segmentation via K-means and HAC; (4) supervised learning with MLR and SVM; and (5) model evaluation using standard performance metrics.

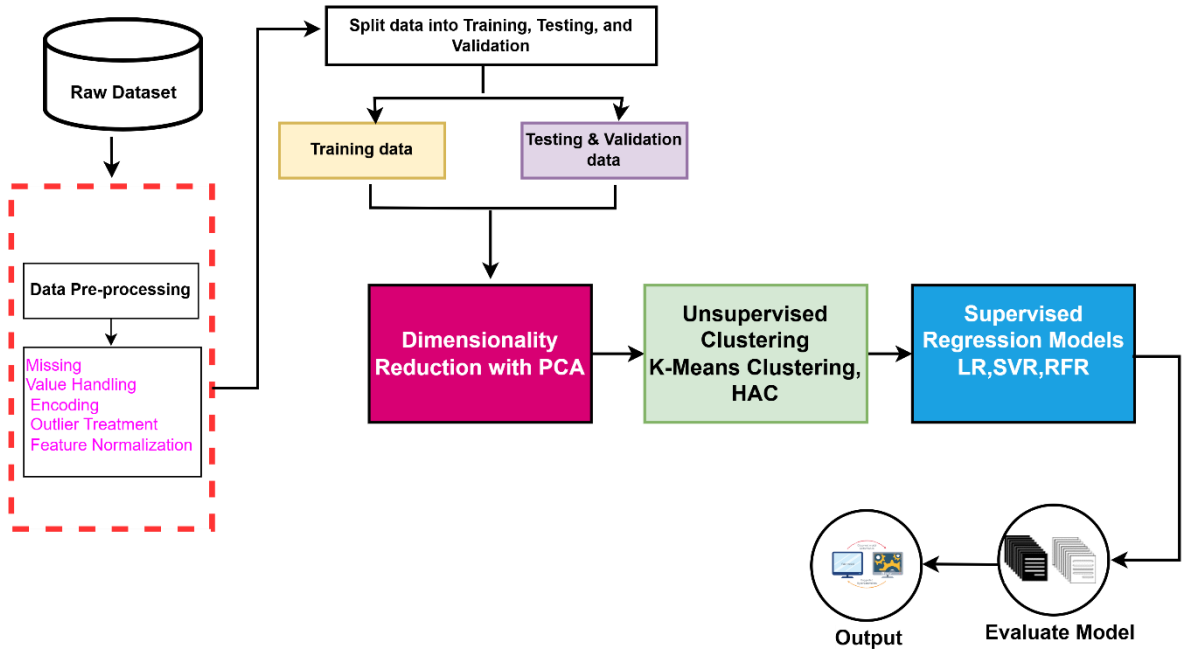

Figure 1: Overview of the proposed hybrid pipeline for car price prediction.

4. Experimental Results

4.1. Dataset Description

The dataset used in this study comprises 399 vehicles described by 10 features, including both numerical and categorical variables. Key attributes include price (price_car, the target variable), mileage, engine size (enginesize), horsepower, year of manufacture, transmission type, fuel type, and car brand. Preliminary exploration revealed a wide price range, with values varying from $184.11 to $45,400.00, and considerable heterogeneity in technical specifications, which justifies the application of both clustering and dimensionality reduction techniques.

4.2. Descriptive Statistics and Correlation Analysis

A preliminary statistical analysis focused on two main variables: price_car and enginesize. The descriptive statistics are as follows:

Mean values:

- price_car: 12,816.09

- enginesize: 129.11

Standard deviations:
- price_car: 8,783.73
- enginesize: moderately dispersed around the mean

Extreme values:
- Minimum price: $184.11 ; Maximum price: $45,400.00
- Minimum engine size: 61 ; Maximum engine size: 326

A Pearson correlation coefficient of 0.696 between price_car and enginesize indicates a moderate to strong positive correlation, suggesting that cars with larger engines tend to be more expensive.

4.3. Data Preprocessing

Several preprocessing steps were carried out before modeling. Categorical variables (e.g., fuel type and transmission) were encoded using one-hot encoding, while missing values were imputed with the mean (numerical variables) or mode (categorical variables). Outliers were handled using the interquartile range (IQR) method. All numerical features were normalized using Min-Max scaling to ensure fair contribution to the models.

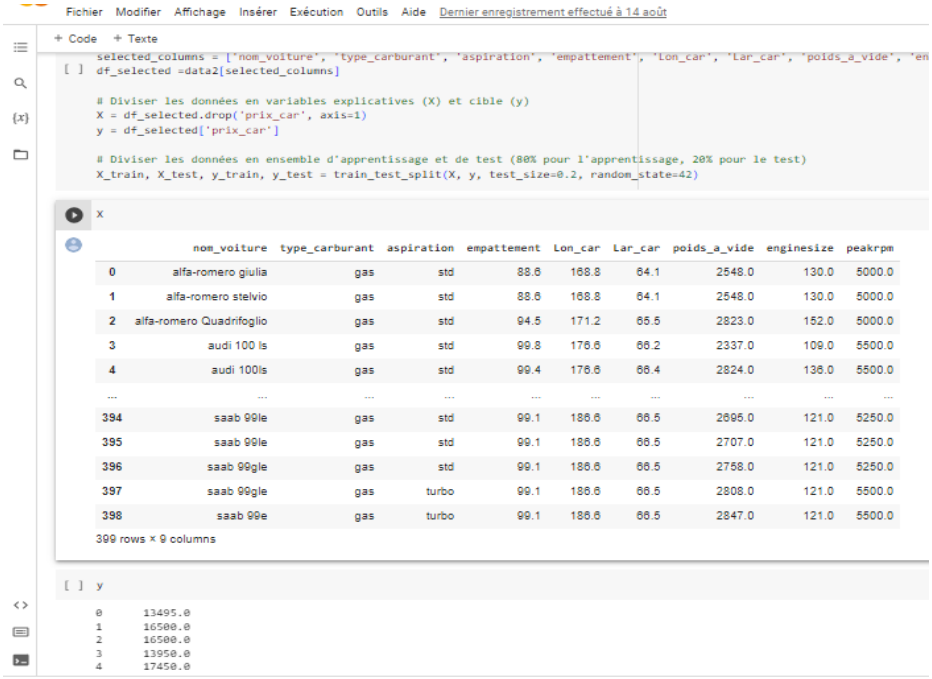

Figure 2 : Overview of Selected Features after Initial Data Preprocessing

The figure 2 shows a snapshot of the dataset after selecting relevant explanatory variables (fuel_type, aspiration, wheelbase, etc.) and the target variable car_price. Categorical variables will be encoded and numerical features normalized in the subsequent preprocessing steps.

4.4. Exploratory Data Analysis

To better understand the characteristics and relationships between the variables in our dataset, we conducted a visual and statistical exploratory data analysis. This phase is essential before proceeding to feature selection and predictive modeling.

a. Distribution and Outliers

Figure 1 presents a boxplot illustrating the distribution of key numerical features, including empattement, Lon_car, Lar_car, poids_a_vide, enginesize, peakrpm, and the target variable prix_car. The plot reveals significant disparities in the scale and dispersion of the variables. Notably, the prix_car variable exhibits a wide range and a high number of outliers, suggesting considerable variability in car prices. Conversely, features like empattement and Lar_car show more compact distributions.

Such variability highlights the need for normalization or standardization to mitigate the impact of scale differences when using machine learning algorithms.

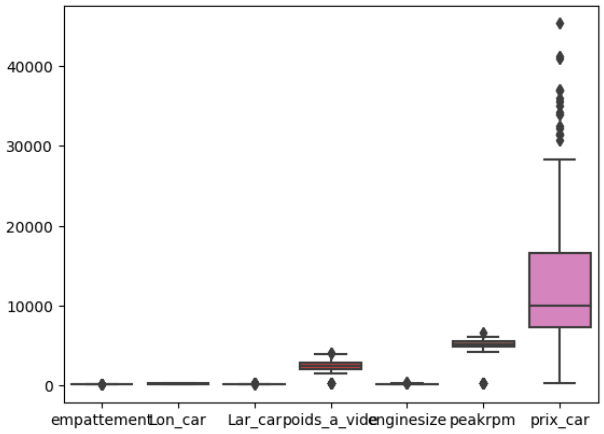

Figure 3: Boxplot of selected numerical features, including the target variable prix_car.

b. Correlation Analysis

Figure 2 displays the Pearson correlation heatmap among the selected numerical variables. The strongest positive correlations with car price (prix_car) are observed for:

- poids_a_vide (correlation coefficient r = 0.79),
- enginesize (r = 0.70),
- Lon_car (r = 0.54),
- and empattement (r = 0.47).

These results suggest that heavier cars with larger engines and greater lengths tend to be more expensive, which aligns with industry expectations.

On the other hand, the variable Lar_car is negatively correlated with both prix_car (r = -0.28) and peakrpm (r = -0.93). The strong negative correlation between Lar_car and peakrpm may indicate potential multicollinearity, which should be carefully addressed during model development.

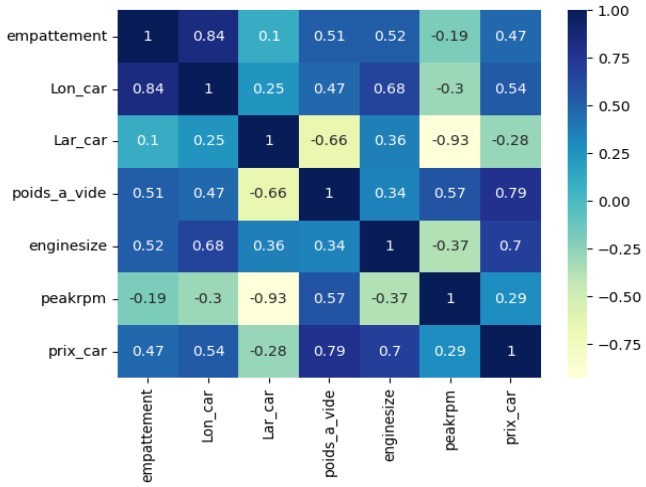

Figure 4: Pearson correlation matrix of numerical features.

4.5. Linear Regression Analysis

A simple linear regression model was constructed to predict price_car from enginesize. The model equation is:

$$price\_car = a \cdot enginesize + b$$

The model achieved an R^2 score of 0.48, indicating that about 48% of the variance in car prices is explained by engine size alone. While this suggests a notable relationship, other features clearly influence the price and merit inclusion in multivariate models.
Visualization: A scatter plot with a fitted regression line was generated to visualize this relationship, confirming the positive trend and highlighting a moderate dispersion around the line.

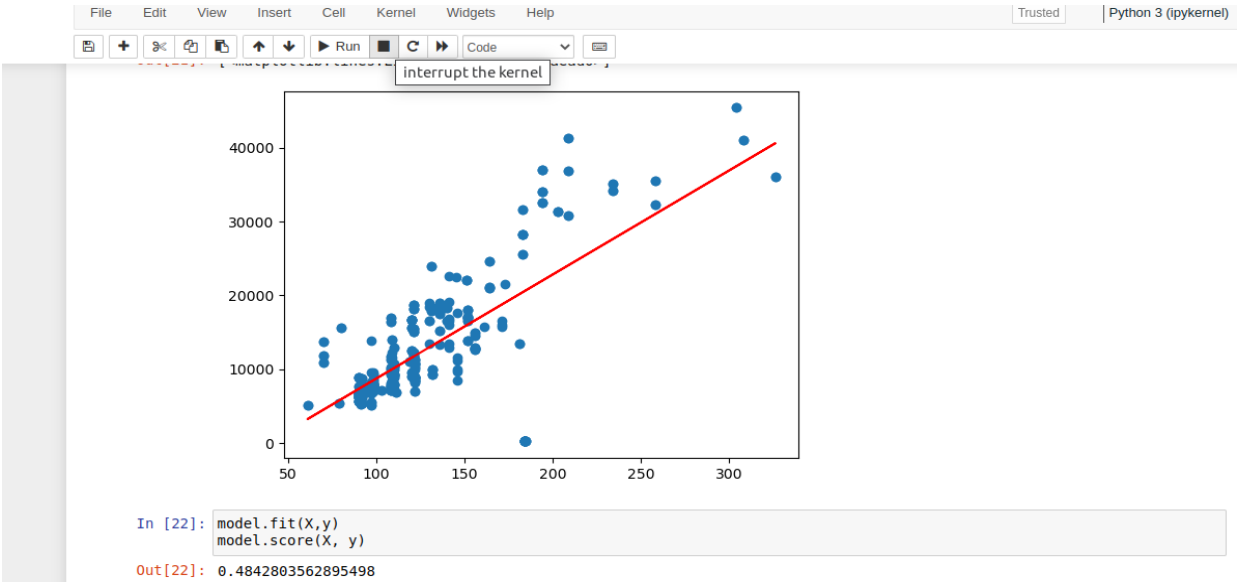

Figure 5 : Linear Regression Fit between Explanatory Variable and Vehicle Price

This figure depicts the fit of a simple linear regression model applied to the dataset. The blue dots represent actual data points, and the red line shows the predicted linear relationship. The coefficient of determination $R^2 \approx 0.48$ suggests that the model explains about 48% of the variance in the vehicle price based on the selected feature.

4.5. Dimensionality Reduction with PCA

Principal Component Analysis (PCA) was then applied to the entire normalized dataset to explore latent structures and reduce dimensionality.

Explained variance:

- PC1: 98%
- PC2: 1.8%

This result indicates that most of the variance is captured by the first component, primarily linked to features such as engine size and car price.

Correlation Circle:

- **Axis 1** (PC1): associated with vehicle size and category (length, width, weight)
- **Axis 2** (PC2): associated with performance metrics (power, acceleration)

Graphical interpretation: The projection showed well-structured groupings of vehicles according to their technical specifications, supporting the validity of PCA for structure discovery.

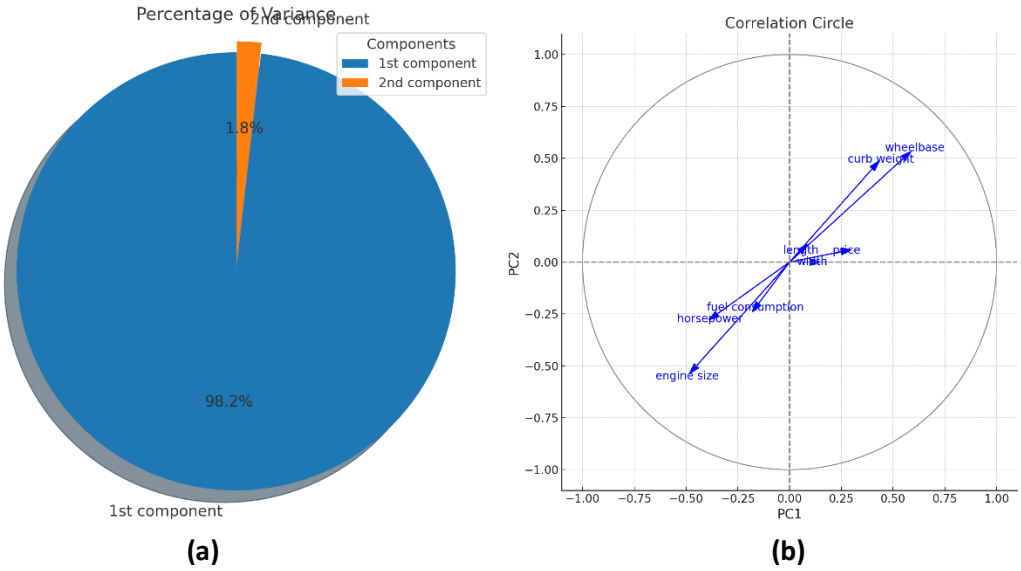

Figure 6: Principal Component Analysis: (a) Explained Variance Ratio, (b) Correlation Circle of Variables

4.6. Clustering Analysis

To segment the vehicles, unsupervised clustering was performed in the PCA-transformed space using K-means. The Elbow method suggested an optimal value of k = 3. The resulting clusters revealed:

1. Compact and economical cars
2. Mid-range vehicles with balanced performance
3. Premium vehicles with large engines and high prices

These segments align closely with the axes identified in the PCA and offer a meaningful market-oriented classification.

Evaluation:

- K-means: Silhouette score of 0.62 (well-separated groups)
- HAC: Score of 0.49 with overlapping clusters

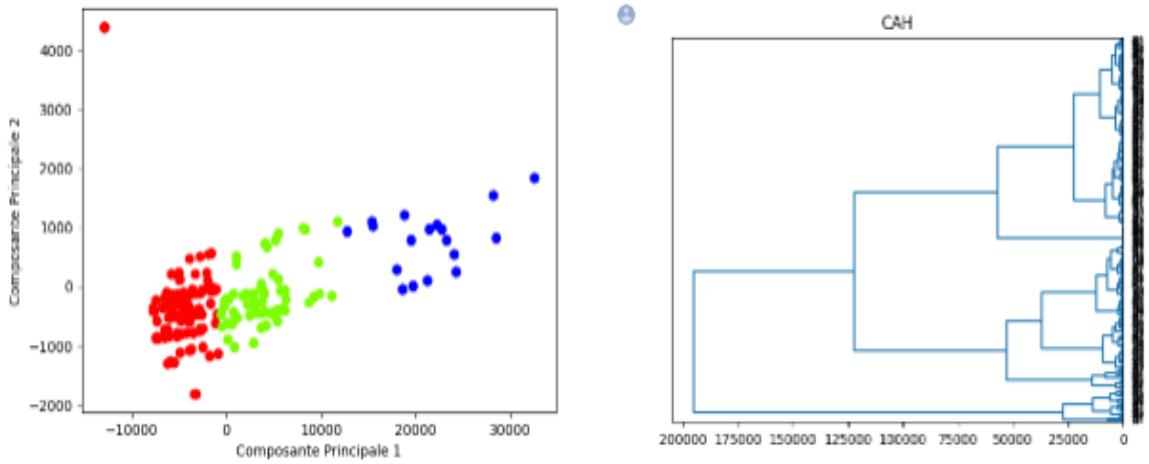

Figure 7: Clustering Method Evaluation: (a) K-means with a Silhouette Score of 0.62 (Well-Separated Groups), (b) HAC with a Silhouette Score of 0.49 (Overlapping Clusters)

4.7. Regression Models Performance

Table 1 presents the benchmarking of predictive performance, where two regression models Multiple Linear Regression (MLR) and Support Vector Machine (SVM) were evaluated using 10-fold cross-validation.

Table 1 : Comparison of Regression Models Based on RMSE, MAE, and R^2 Score

Méthode	RMSE	MAE	R^2 Score
Multiple Linear Regression	2310.12	1357.15	0.92
Support Vector Machine	3212.14	2248.59	0.85

The comparison between the Multiple Linear Regression (MLR) and Support Vector Machine (SVM) models highlights the superior performance of the MLR approach. With a lower Root Mean Square Error (2310.12 vs. 3212.14), a lower Mean Absolute Error (1357.15 vs. 2248.59), and a higher R^2 score (0.92 vs. 0.85), the MLR model demonstrates greater predictive accuracy and explanatory power. These results indicate that MLR more effectively captures the relationship between the input features and vehicle prices. Therefore, due to its simplicity, robustness, and interpretability, MLR is considered the more suitable model for this regression task.

5. Discussion and Interpretation

To facilitate data simplification and enhance the interpretability of features, Principal Component Analysis (PCA) was conducted on the preprocessed dataset. The results reveal a clear predominance of the first principal component (PC1), which alone captures 98.2% of the total variance, while the second component (PC2) accounts for only 1.8%, as shown in Figure 6. This distribution is illustrated in the pie chart titled *Percentage of Variance*, highlighting that a single component is sufficient to retain most of the information in the dataset thus enabling dimensionality reduction without significant loss of variance.

Complementing this, the correlation circle provides further insight into the relationships between the original variables and the principal components. Notably, features such as "wheel-base," "curb-weight," and "engine-size" show strong loadings on PC1 and PC2, indicating their substantial contribution to the explained variance. These variables emerge as key predictors and can be prioritized in subsequent regression modeling tasks aimed at estimating car prices.

To assess the impact of these features on predictive performance, a comparative evaluation of two supervised learning models was performed: Multiple Linear Regression (MLR) and Support Vector Machine (SVM). As reported in Table 1, the MLR model outperformed the SVM across all evaluation metrics. Specifically, MLR achieved a high coefficient of determination (R^2) of 0.92, suggesting that 92% of the variance in car prices is accurately explained. In contrast, the SVM model obtained a slightly lower R^2 of 0.85, indicating reduced explanatory power as shown in Table 1.

The difference in predictive accuracy is further confirmed by the error metrics. The MLR model recorded an RMSE of 2310.12 and an MAE of 1357.15, while the SVM model exhibited higher error values (RMSE = 3212.14, MAE = 2248.59). These figures imply that the linear model not only generalizes better but also yields more precise predictions than the non-linear SVM approach in this specific context.

In conclusion, the combined use of PCA and regression analysis underscores the effectiveness of dimensionality reduction in refining feature relevance, and demonstrates that the Multiple Linear Regression model offers a more robust and accurate solution for car price prediction based on the selected attributes.

5.1. Limitations and Future Research

Despite the informative nature of this study, several limitations must be acknowledged. The analysis was restricted to numerical features, excluding categorical variables such as brand, fuel type, and vehicle class, which are known to play a crucial role in car pricing. This omission may reduce the explanatory capacity of the model. Moreover, the presence of outliers and potential non-linear relationships among variables could affect the reliability of correlation-based interpretations. Future research should integrate a more comprehensive set of features, including categorical and external data sources, and explore advanced modeling techniques such as ensemble learning and neural networks, (Liu, E. et al., 2022). These approaches could improve predictive performance and offer a more robust understanding of the multifactorial nature of car pricing. Furthermore, addressing issues like model interpretability (Nguemdjom, D. K. et al., 2025) and algorithmic fairness should be prioritized to ensure responsible deployment of predictive systems in real-world applications.

5.2. Code Availability

To ensure transparency, reproducibility, and to encourage further research, all code developed for data preprocessing, dimensionality reduction, clustering, and regression modeling is publicly available. The complete implementation, along with documentation and example notebooks, is available at: GitHub repository: click Here

6. Conclusion

In this study, we conducted a comprehensive exploratory data analysis to investigate the relationships between car price and several numerical features such as curb weight, engine size, wheelbase, and car dimensions. Through the use of descriptive statistics, boxplots, and Pearson correlation analysis, we identified strong linear associations between the target variable (prix_car) and predictors like poids_a_vide (curb weight), enginesize, and Lon_car (car length). These insights are consistent with domain knowledge, suggesting that higher engine capacity and overall vehicle size contribute to price increase due to manufacturing costs and performance expectations.

Furthermore, our analysis revealed potential multicollinearity between certain variables, such as Lar_car and peakrpm, highlighting the importance of careful feature selection and diagnostic testing in subsequent modeling stages. Such findings emphasize the value of thorough pre-modeling analysis to ensure the validity and interpretability of machine learning models in real-world applications.

Future research will focus on integrating ensemble methods and deep learning models (e.g., Gradient Boosting, XGBoost, or Neural Networks) to capture non-linear relationships more effectively. Additionally, the incorporation of categorical variables and real-time data sources could further improve model robustness and generalizability.

REFERENCES

[1] Chaturvedi, A., Green, P. E., & Carroll, J. D. (2001). *K-modes clustering*. Journal of Marketing Research, 36(2), 276–280.

[2] Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). *Support vector regression machines*. Advances in Neural Information Processing Systems, 9.

[3] Alidor Mbayandjambe, Kevin Nguemdjom, Grevi Nkwimi, Fiston Oshasha, Heritier Mbengandji. (2025). Multi-Model Optimization for Telecom Churn Prediction: A Complete Data Science Approach from Theory to Python Implementation. International Journal of Future Management Research, Vol. 7, No. 2, DOI: https://doi.org/10.36948/ijfmr.2025.v07i02.41263

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

[5] Jolliffe, I. T., & Cadima, J. (2016). *Principal component analysis: A review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

[6] Klein, L., Jäger, T., & Bauer, S. (2020). *Used car price prediction: A comparison of machine learning algorithms*. Procedia Computer Science, 176, 444–453.

[7] Rosen, S. (1974). *Hedonic prices and implicit markets: Product differentiation in pure competition*. Journal of Political Economy, 82(1), 34–55.

[8] Smola, A. J., & Schölkopf, B. (2004). *A tutorial on support vector regression*. Statistics and Computing, 14(3), 199–222.

[9] Tan, P. N., Steinbach, M., & Kumar, V. (2019). Introduction to Data Mining (2nd ed.). Pearson.

[10] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

[11] Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy, 82(1), 34–55. https://doi.org/10.1086/260169

[12] Shi, Q., Liu, J., & Jiang, H. (2022). Hybrid PCA-LSSVM model for highway construction cost prediction. Computer Systems Science and Engineering, 36(1), 57–69. https://doi.org/10.32604/csse.2023.040901

[13] Tan, P. N., Steinbach, M., & Kumar, V. (2019). Introduction to Data Mining (2nd ed.). Pearson.

[14] Uma, D. S., & Uma, K. M. (2022). Forecasting vehicle prices using machine learning techniques. International Journal of Engineering Research & Technology (IJERT), 11(2). https://www.ijert.org/forecasting-vehicle-prices-using-machine-learning-techniques

[15] Vaneesha, V., Swarnalatha, K., & Karthikeyan, M. (2024). Comparison of regression models for used car price prediction. Zenodo. https://doi.org/10.5281/zenodo.13799018

[16] Jain, A. (2021, May 17). *Build and deploy a car price prediction system using machine learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/build-and-deploy-a-car-price-prediction-system/

[17] Wang, Z., Li, Y., & Liu, X. (2022). Research on the prediction model of the used car price in China. *Sustainability, 14*(15), 8993.https://doi.org/10.3390/su14158993

[18] Noreen, M., Basharat, M., Mehmood, R. M., & Qayyum, A. (2023). Machine learning optimization and challenges in used car price prediction. *ResearchGate*. https://www.researchgate.net/publication/388323862

[19] Lee, Y., Kim, D., & Lee, K. (2023). Vehicle price prediction by aggregating decision tree model with boosting model. *arXiv preprint arXiv:2307.15982*. https://arxiv.org/abs/2307.15982

[20] Zhou, Y., & Geng, X. (2024). ProbSAINT: Probabilistic tabular regression for used car pricing. *arXiv preprint arXiv:2403.03812*. https://arxiv.org/abs/2403.03812

[21] Nguemdjom, D. K. T., Mbayandjambe, A. M., Nkwimi, G. B., Oshasha, F., Muluba, C., Mbengandji, H. I., & Bazie, I. G. (2025). Explainable AI (XAI) for Obesity Prediction: An Optimized MLP Approach with SHAP Interpretability on Lifestyle and Behavioral Data. International Journal of Innovative Science and Research Technology, 10(4). https://doi.org/10.38124/ijisrt/25apr1962

[22] Liu, E.; Li, J.; Zheng, A.; Liu, H.; Jiang, T. Research on the Prediction Model of the Used Car Price in View of the PSO-GRA-BP Neural Network. *Sustainability* **2022**, *14*, 8993. https://doi.org/10.3390/su14158993