Carismatic: Unmasking the True Drivers of Car Sales

Kaifeng Gao
Department of Statistics & Data Science
Yale University
New Haven, CT 06510
kaifeng.gao@yale.edu

Jiayi Chen
Department of Statistics & Data Science
Yale University
New Haven, CT 06510
jiayi.chen@yale.edu

Qianmeng Chen
Department of Statistics & Data Science
Yale University
New Haven, CT 06510
qianmeng.chen@yale.edu

Abstract

Understanding the factors that influence car sales is essential for the automotive industry. In this project, we analyze various variables, including the reputation of the car brand (rate) and the "facial expressions" of cars, to evaluate their impact on sales performance. Leveraging data from Deep Visual Marketing, we applied random forest, linear regression, backpropagation neural network, and recurrent neural network to identify the key determinants of car sales. Contrary to our hypothesis, we found that while variables like fuel consumption, car model, and size significantly influence sales, the perceived facial expressions of cars do not. These findings provide valuable insights for manufacturers aiming to optimize sales strategies.

1 Introduction

Understanding the determinants of car sales is a critical area of research within the automotive industry. As consumer preferences evolve, both functional attributes such as fuel efficiency and aesthetic elements such as design play essential roles in purchasing decisions. Recent advancements in machine learning have enabled researchers to analyze large datasets, uncovering nuanced insights into what drives sales performance. Motivated by these developments, this study aims to investigate the impact of various factors, including car brand reputation, physical dimensions, and perceived "facial expressions," on car sales. While prior research highlights the importance of functional characteristics like fuel consumption and engine performance, the role of visual design elements remains underexplored. This project seeks to address this gap by combining traditional statistical methods with state-of-the-art machine learning models.

The dataset used in this study was obtained from the Deep Visual Marketing platform and consists of 5,269 car records, each containing detailed attributes. These include physical specifications (e.g., engine size, dimensions), performance metrics (e.g., fuel consumption, average sales), and sentiment-based probabilities derived from visual features resembling emotions (e.g., happiness, sadness). Preprocessing steps such as imputation for missing values, feature scaling, and one-hot encoding were applied to prepare the data for modeling.

By examining both practical and aesthetic factors, this study provides a comprehensive analysis of car sales determinants, offering valuable insights for the automotive industry to refine its design, production, and marketing strategies.

2 Data Preparation

2.1 DMV-CAR Dataset

DVM-CAR 2.0 [Huang et al., 2023] is a large-scale automotive dataset for visual marketing research and applications. The dataset contains car images, model specification and sales information about 899 car models that have been sold in the UK market over the last 20 years. It comprises two data parts: the image data and the table data. The former contains 1,451,784 car images that have been deliberately cleaned and organized according to their viewpoints, unified to the same size, and strictly follow the same storage structure. While the latter includes six CSV tables that cover the non-visual attributes such as brand, price, sales etc.

2.1.1 Car Emotion Classification

To classify car emotions, we employed a Vision Transformer (ViT) [Dosovitskiy et al., 2021] model, specifically utilizing the pre-trained model `trpakov/vit-face-expression`. This model is pretrained on Facial Expression Recognition 2013 Dataset (FER2013) [Khaireddin and Chen, 2021] dataset and is capable of distinguishing seven distinct emotional expressions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

The emotion classification process focused on frontal-view car images, which share greater structural similarities with human faces. From DMV-CAR dataset, they provided a frontal view subset with 61,827 front-view images manually verified and selected for analysis. Figure 1 illustrates the predicted car emotions for a representative sample.

Figure 1: Predicted Car Emotions

The ViT architecture processes images through a unique approach, treating image analysis as a sequence processing task. Unlike traditional CNNs, the model:

- Divides images into fixed-size patches (typically 16x16 pixels)
- Applies linear embedding to these patches
- Combines them with position embeddings to preserve spatial information
- Processes the embedded sequence through multiple transformer encoder layers
- Utilizes self-attention mechanisms to capture both local and global image dependencies

Our preprocessing pipeline consists of three main stages:

1. **Image Loading and Conversion**:
   - Loading images via PIL (Python Imaging Library)
   - Converting to grayscale using `convert('L')` to neutralize car color influence
   - Converting back to RGB format for model compatibility
2. **Model Preparation**:
   - Implementing AutoImageProcessor from the Transformers library
   - Automating resizing, normalization, and tensor conversion
3. **Batch Processing**:
   - Processing images recursively from the root directory
   - Recording predictions and probability scores systematically

The results are stored in a structured pandas DataFrame containing:

- `image_path`: Full path to the processed image
- `label`: Predicted emotion label
- `prob_*`: Probability scores for each emotion category

This systematic approach to emotion classification yields both categorical predictions and probability distributions, providing rich data for subsequent analysis. The grayscale conversion ensures that car colors do not influence the emotion classification, while the structured data format facilitates efficient access and analysis of the results across the dataset.

2.2 Edmunds-Consumer Car Ratings and Reviews Dataset

The Edmunds-Consumer Car Ratings and Reviews dataset provides comprehensive insights into consumer feedback on various car manufacturers, models, and types. It includes data from 62 major brands, organized in CSV files named using the format `Scraped_car_review_brand.csv`. Each file contains structured data with columns for Author Name, Vehicle Title, Review Title, Review, and Rating. This dataset serves as a valuable resource for analyzing consumer opinions and ratings, offering a foundation for research into automotive brand performance and market trends.

2.2.1 Emotional Rating Extraction

To analyze and categorize consumer sentiment from car reviews, we implement a transformer-based approach utilizing the `nlptown/bert-base-multilingual-uncased-sentiment` model from the Hugging Face Transformers library. This model is specifically optimized for text classification tasks, allowing it to proficiently evaluate the sentiment across a spectrum of languages. The process involves the following steps:

- **Tokenization and Truncation:** Each review is tokenized using the model's dedicated tokenizer. If a review exceeds the model's maximum input length (512 tokens), it is truncated to fit within this constraint.
- **Sentiment Prediction:** The transformer model produces a sentiment label for each review, such as '1 star', '2 stars', and so forth. These labels are interpreted as an indicator of the review's emotional tone.
- **Quantitative Conversion:** These qualitative sentiment labels are mapped to numeric values ranging from 1 to 5. This transformation allows for quantitative analysis of sentiment trends across different brands and models.

This methodology enhances the understanding of consumer feedback by converting it into structured numerical insights that are usable in further data analysis.

2.2.2 Integration with DMV-CAR

Integrating the consumer reviews with the DMV-CAR dataset presents a challenge, as there is no common identifier available to directly join the two datasets. To address this, we implemented a method using fuzzy string matching to identify probable matches based on available attributes. This process is detailed below:

- **Attribute Extraction:** For each review, the `Vehicle_Title` is parsed into three main components: the year, maker, and model of the vehicle.
- **Composite Identifier:** In the DMV-CAR dataset, a composite string is created by concatenating attributes like the car model (`Genmodel`) and year, e.g., `"Corolla 2020"`.
- **Similarity Scoring:** The similarity between entries is calculated for each attribute using the FuzzyWuzzy library. The formula used to compute a weighted score S for each potential match is:

$$S = 0.45 \times \text{fuzz.ratio(maker)} + 0.45 \times \text{fuzz.ratio(model)} + 0.1 \times \text{fuzz.ratio(year)}$$

- **Threshold and Matching:** A total score S above 70 is considered an acceptable match, which is selected as the best possible match for a review. For instance, a review with a `Vehicle_Title` "2020 Honda Civic" might be matched with the DMV-CAR entry "Civic 2020" based on component similarities and calculated scores.

This detailed methodology enables the effective combination of these two datasets, facilitating a broader and more profound analysis of the automotive landscape by incorporating comprehensive consumer sentiment insights with technical specifications found within DMV-CAR data.

2.3 Data Cleaning

The dataset underwent extensive preprocessing to ensure data quality and suitability for analysis. This section details the steps taken to process the various data sources, including sales data, vehicle characteristics, advertisements, and image-based expression analysis. These steps were crucial to address missing values, handle outliers, and aggregate the data into a unified format.

1. **Sales Data Preprocessing**: Sales data were processed to derive the adjusted average sales for each vehicle generation model (`avg_sales`) using the following steps:
   - Exclusion of 2020 Sales Data: Sales data for 2020 was excluded to account for market anomalies caused by the COVID-19 pandemic.
   - Filtering Non-Zero Sales Values: Years with zero sales were removed to focus on meaningful sales activity.
   - Exclusion of the Last Non-Zero Year: For generation models with multiple non-zero sales years, the most recent year was excluded to mitigate the influence of short-term trends.
   - Outlier Removal: Unusually low sales values were identified and excluded using the interquartile range (IQR) method.
   - Calculation of Adjusted Sales: The adjusted average sales were calculated as the mean of the remaining sales values. If no valid values remained after filtering, the result was set to NA
2. **Vehicle Trim Data Aggregation**: The vehicle trim data was aggregated by generation model ID to compute average values and majority categories for key attributes:
   - Numerical Averages:
     - Average price (`avg_price`)
     - Average gas emissions (`avg_gas_emission`)
     - Average engine size (`avg_engine_size`)
   - Categorical Modes:
     - The most common fuel type (`majority_fuel_type`) was identified for each generation model.
3. **Advertisement Data Preprocessing**: The advertisement data was processed to standardize and aggregate key attributes:
   - **Unit Conversion:**
     - Columns such as `Average_mpg` and `Top_speed` were cleaned by removing units (e.g., "mpg," "mph") and converting values to numeric format.
   - **Attribute Aggregation by Generation Model:**
     - **Numerical attributes:**
       - Average engine power (`avg_engine_power`)
       - Average dimensions, including wheelbase, height, width, and length (`avg_wheelbase`, `avg_height`, `avg_width`, `avg_length`)
       - Average fuel efficiency (`avg_mpg`)
       - Average top speed (`avg_top_speed`)
     - **Categorical attributes:**
       - Majority values for the number of seats (`majority_seat_num`) and doors (`majority_door_num`).

    - Most common engine size (`majority_engine_size`), body type (`majority_bodytype`), and gearbox type (`majority_gearbox`).

4. **Image-Based Expression Data Preprocessing**: The image data and expression analysis results were combined and aggregated to derive probabilistic and categorical features for each generation model:
    - **Average Emotion Probabilities:**
        - The probabilities of different emotional expressions (e.g., angry, happy, neutral) were averaged across all images for each generation model: `avg_prob_angry`, `avg_prob_disgust`, `avg_prob_fear`, `avg_prob_happy`, `avg_prob_sad`, `avg_prob_surprise`, and `avg_prob_neutral`.
    - **Majority Label:**
        - The most common emotional expression label (`majority_label`) was determined based on frequency.

Finally, the processed data from sales, trims, advertisements, and image-based expression analysis were merged using the generation model ID as the common key. This unified dataset served as the foundation for downstream analysis and modeling.

3 Data Exploration and Visualization

3.1 Relationship Between Car Expression and Sales Performance

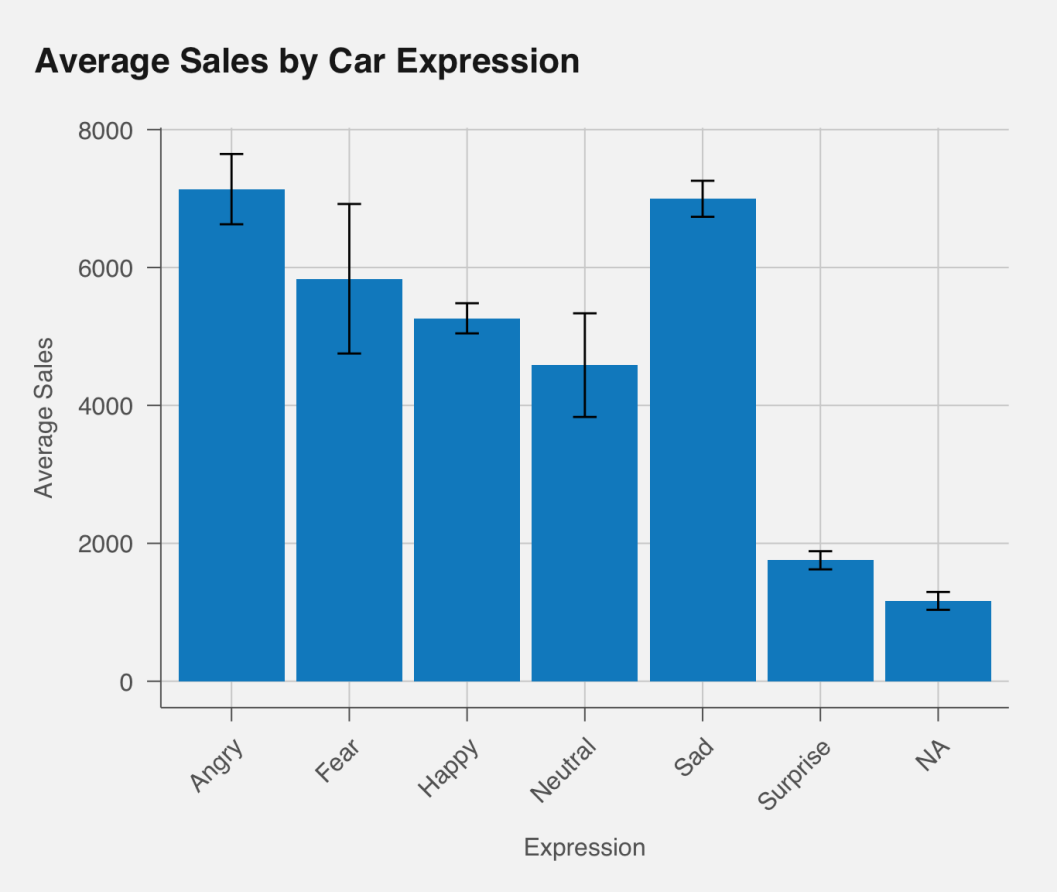


Figure 2: Average Sales by Car Emotion

Figure 2 presents the average sales volumes across different car expressions, revealing distinct patterns in consumer preferences or market performance. Cars characterized with "Angry" and

"Sad" expressions demonstrate the highest average yearly sales, approximately 7,000 and 6,800 units respectively, while those with "Surprise" expressions or undefined expressions (NA) show significantly lower sales volumes, below 2,000 units.

This stark contrast in sales performance across different expressions suggests several important insights:

- The substantial variation in sales across expressions (ranging from approximately 1,000 to 7,000 units) indicates that front-end design and perceived expression may be a significant factor in consumer purchasing decisions.
- The dominance of "Angry" and "Sad" expressions might reflect broader market preferences for aggressive or serious-looking vehicles, possibly associated with luxury or performance car segments.
- The lower sales of cars with "Surprise" expressions could indicate either consumer resistance to unconventional designs or might be confounded with other variables such as price point or vehicle category.
- The presence of error bars in the visualization suggests considerable variation within each expression category, pointing to the likely influence of other factors such as brand, price range, or vehicle type.

These patterns suggest that car expression could be a valuable predictor in our sales modeling approach, though careful consideration should be given to potential confounding variables and interaction effects with other predictors such as brand positioning and vehicle category.

3.2 Expression Distribution Across Major Manufacturers

Figure 3 illustrates the distribution of car expressions among the top 10 manufacturers in the UK market, revealing distinct design philosophies and brand identities. This visualization unveils several noteworthy patterns:

- "Sad" and "Happy" expressions dominate across most manufacturers, suggesting these are generally accepted design approaches in the automotive industry. This could reflect either established design conventions or proven market preferences.
- Japanese manufacturers (Toyota, Honda, and Nissan) display a notably higher proportion of "Angry" expressions in their vehicle designs. This consistent pattern might reflect:
  - A regional design philosophy specific to Asian manufacturers
  - Strategic brand positioning in the sports and performance segments
  - Cultural differences in design interpretation and preference
  - European manufacturers like Volkswagen and Vauxhall show a strong preference for "Sad" expressions, possibly aligning with a more conservative or traditional design approach.
- Luxury brands such as BMW and Mercedes demonstrate more diverse expression distributions, potentially indicating more experimental design approaches or broader model ranges.

These manufacturer-specific patterns suggest a complex relationship between brand identity, market positioning, and front-end design choices. For modeling purposes, this indicates potential interaction effects between manufacturer and expression variables in predicting sales performance. The strong manufacturer-specific patterns also suggest that expression effects on sales might need to be considered within the context of individual brands rather than globally across the entire market.

3.3 Trends in Sales Across Expressions

Figure 4 presents the distribution of normalized sales (z-score standardized) across different car expressions from 2001 to 2019, controlling for the general upward trend in automotive sales. This visualization reveals several significant temporal patterns:

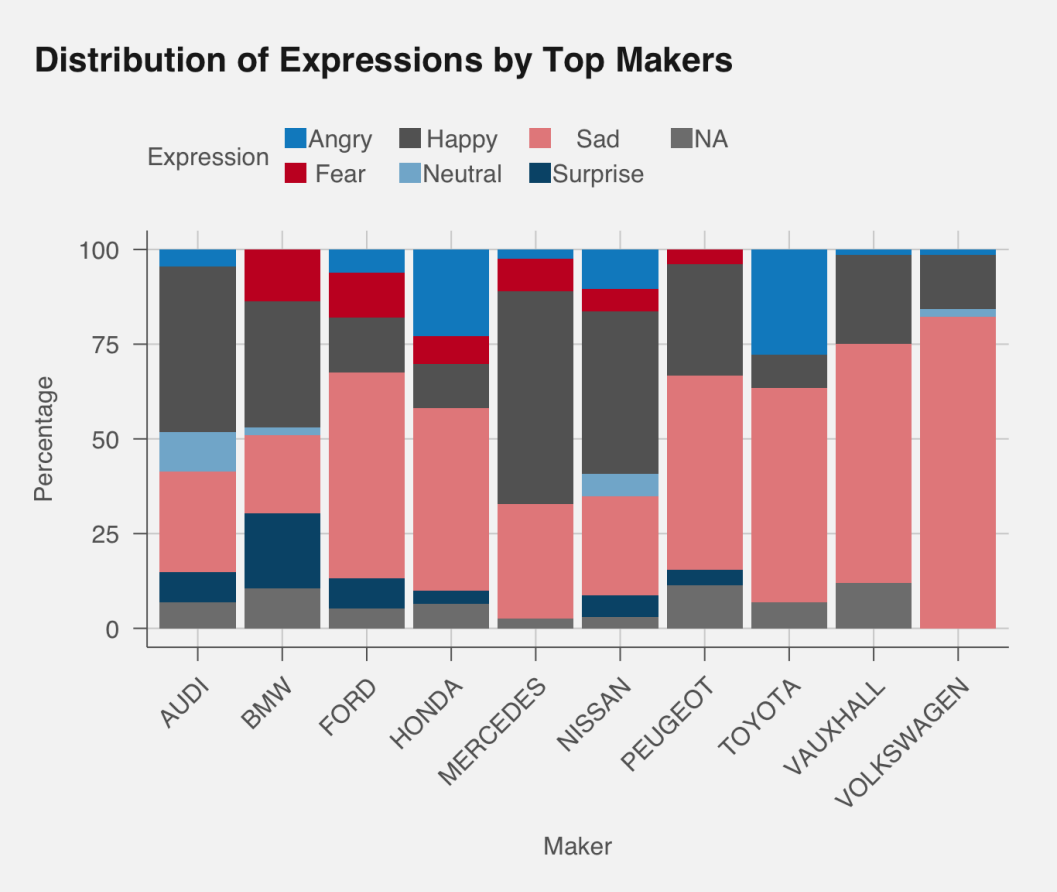


Figure 3: Distribution of Expressions by Top Makers

- Cars with "Surprise" expressions show a consistent decline in normalized sales over the study period, suggesting:
  - A growing consumer preference for more conventional design languages
  - Possible market saturation for unconventional designs
- "Angry" and "Sad" expressions maintain relatively stable or slightly increasing normalized sales distributions, particularly in recent years (2015-2019). This stability suggests:
  - These expressions represent enduring design choices that align with consistent consumer preferences
  - Possible association with successful vehicle categories or market segments
- The distributions for "Neutral" and "Happy" expressions show moderate variability but maintain relatively consistent central tendencies across years, indicating:
  - These expressions represent a "safe" design choice for manufacturers
  - Steady market acceptance without strong positive or negative trends

The temporal analysis reinforces earlier observations about consumer preferences for conventional and serious-looking designs. For modeling purposes, these trends suggest the importance of including temporal components and potentially interaction terms between year and expression variables.

3.4 Evolution of Expression Proportions in Vehicle Design

Figure 5 tracks the relative proportions of different car expressions from 2011 to 2019, revealing subtle but meaningful shifts in automotive design trends. The visualization highlights several key patterns:

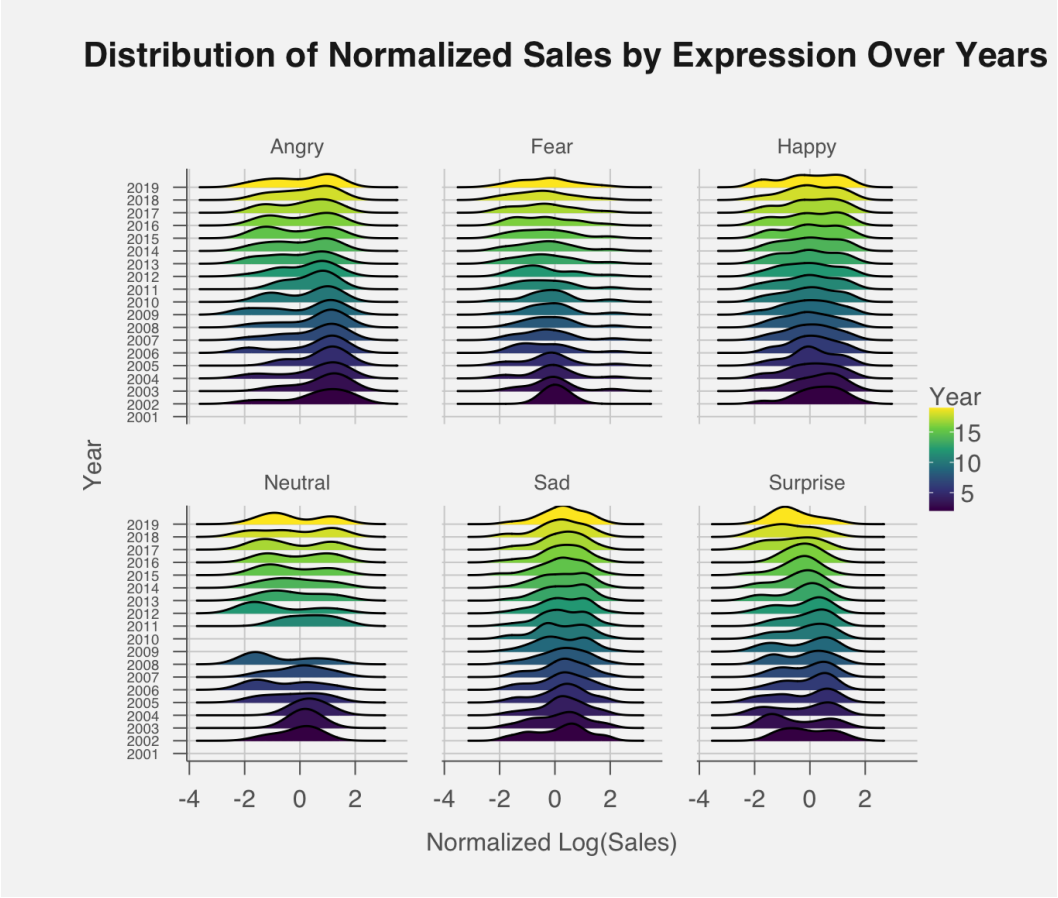


Figure 4: Distribution of Sales by Expression over Years

- "Angry" expressions show a notable upward trend, increasing from approximately 5% to 9% of the market.
- "Sad" and "Happy" expressions, while remaining the two most common designs, show contrasting trends:
  - "Sad" expressions demonstrate a gradual decline from about 52% to 45%
  - "Happy" expressions maintain relative stability around 33%, with minor fluctuations
- Other expressions ("Fear," "Neutral," and "Surprise") maintain relatively stable proportions throughout the period

These evolving proportions suggest a gradual shift in automotive design language, potentially reflecting changing consumer preferences or market dynamics. For modeling purposes, this temporal variation in expression prevalence should be considered when assessing the relationship between expressions and sales.

3.5 Detailed Analysis: Evolution of Honda's Design Language

While this trend toward more aggressive and "angry" expressions appears across the automotive industry as a whole, it is particularly instructive to examine how this evolution has manifested within individual manufacturers' design languages.

Figure 6 presents a detailed examination of Honda's design evolution through three complementary visualizations: a ridge plot showing the distribution of "Angry" expression probabilities over time, and chronological series of Honda CR-V and Civic front-end designs. This case study reveals a clear transformation in Honda's design philosophy:

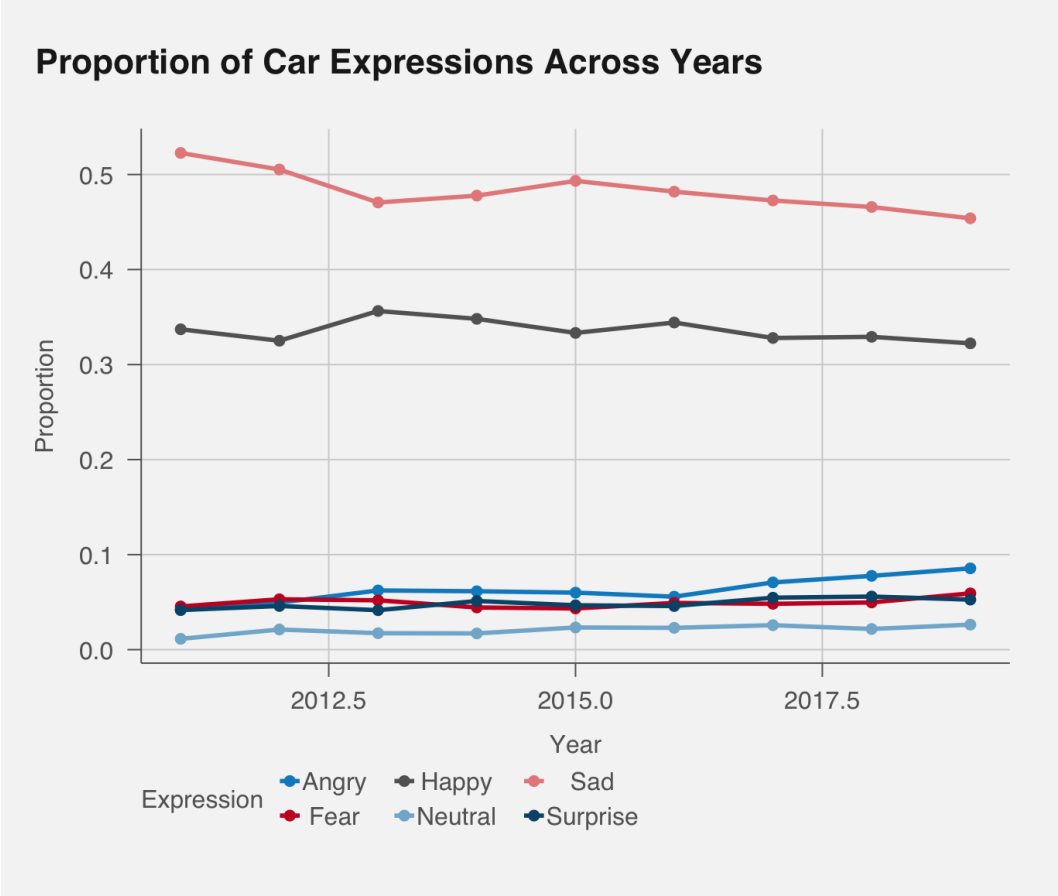
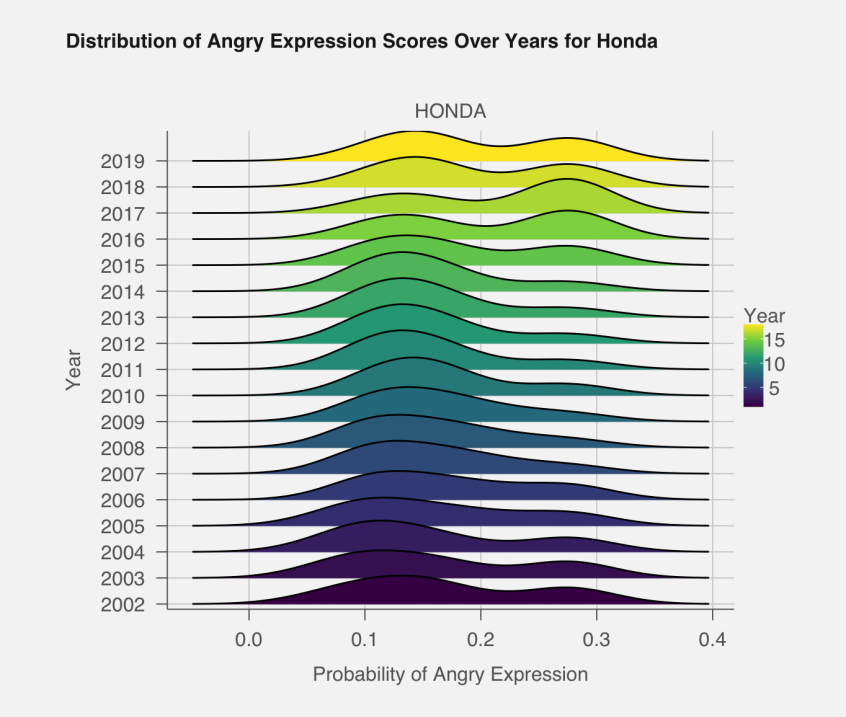


Figure 5: Proportion of Car Expressions across Years

- The ridge plot demonstrates a systematic shift in the probability distribution of "Angry" expressions:
  - Early 2000s: Distributions centered around 0.1-0.15 probability
  - Mid-2010s: Distributions shift rightward, with peaks around 0.2-0.3
  - Recent years: Broader distributions with higher probabilities of "Angry" expressions
- The Honda CR-V example provides visual confirmation of this trend:
  - 2000 model: Relatively neutral front-end design (0.10 anger probability)
  - 2008 model: Softer appearance (0.05 anger probability)
  - 2012-2015 models: Progressive adoption of more aggressive styling elements, with anger probability increasing from 0.10 to 0.16
- The Honda Civic provides another striking example of this design transformation:
  - 2000 model: Conservative, neutral design with minimal aggressive elements (0.07 anger probability)
  - 2009 model: Introduction of sharper angles and more pronounced grille (0.21 anger probability)
  - 2010-2011 models: Dramatic increase in aggressive styling elements, with anger probability reaching 0.33-0.34

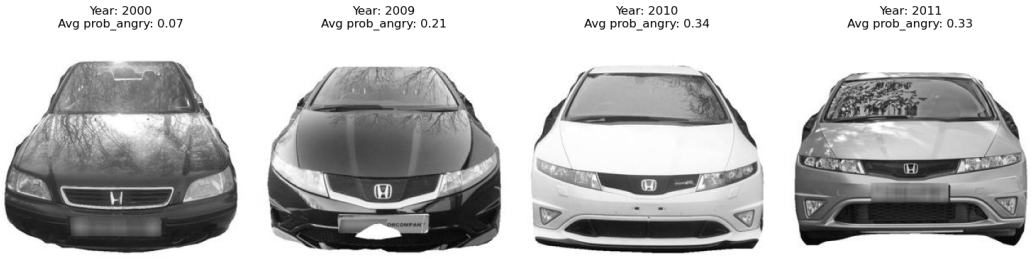
This detailed analysis of Honda's design evolution exemplifies broader industry trends, particularly among Asian manufacturers. The systematic increase in "Angry" expression scores suggests a deliberate strategic shift in design language, possibly responding to:

- Changing consumer preferences for more assertive styling



(a) Distribution of Angry Expression Scores Over Years for Honda

(b) Evolution of Honda CR-V front-end design (2000-2015)



(c) Evolution of Honda Civic front-end design (2000-2011)

Figure 6: Evolution of Honda's Design Language Across Model Lines

- Competitive pressure in the crossover/SUV segment
- Global market dynamics favoring more aggressive design languages

For modeling purposes, this granular analysis suggests that the relationship between expressions and sales should consider not just the categorical expression labels, but also the continuous probability scores that capture subtle variations in design intensity.

4 Modeling/Analysis

4.1 OLS

4.1.1 Assumptions

- **Linearity:** The OLS model assumes a linear relationship between `avg_sales` (the target variable) and the predictors, such as `avg_price`, `avg_engine_power`, and `avg_mpg`. This assumption is critical for unbiased coefficient estimates. However, linearity may not fully capture complex relationships or interactions among predictors, such as those between `majority_bodytype` and `avg_top_speed`. Residual plots for the aggregated dataset indicated minor deviations from linearity, which could limit the interpretability of coefficients.
- **Independence:** The assumption of independence is generally satisfied for the aggregated dataset, as each observation corresponds to a unique car generation model. However, for the yearly dataset, temporal autocorrelation could violate this assumption, as sales figures across years for the same car model may be interdependent.
- **Multicollinearity:** Predictors such as `avg_width`, `avg_height`, and `avg_length` exhibited multicollinearity, which can inflate standard errors and reduce the stability of coefficient estimates. Variance Inflation Factor (VIF) analysis was conducted, and features with high VIF values were removed or consolidated to mitigate this issue.
- **Homoscedasticity:** Residuals should have constant variance across all levels of predictors. Residual plots showed slight heteroscedasticity for variables like `avg_price`, especially in the yearly dataset. While the impact was minor, it suggests that variability in errors may increase for extreme values of predictors.
- **Normality:** The residuals of the aggregated dataset largely followed a normal distribution, as confirmed through Q-Q plots and histograms. For the yearly dataset, deviations from normality were more pronounced, possibly due to temporal trends and outliers.

4.1.2 Model Description

- **Aggregated Dataset**: A linear regression model was applied to predict the average sales (`avg_sales`) for each car generation model using the equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

  where:

  - y: `avg_sales` (target variable)
  - X_i: Predictors such as `avg_price`, `avg_engine_power`, `avg_mpg`, `avg_height`, and `avg_gas_emission`
  - β_0: Intercept
  - β_i: Coefficients of the predictors
  - ϵ: Error term

- **Yearly Dataset**: An additional temporal predictor, `Year`, was incorporated to account for temporal trends in car sales. While the inclusion of `Year` allows the model to capture average annual sales changes, the assumption of independence is likely violated due to interdependencies across years for the same model. The model was still implemented to gain preliminary insights, acknowledging that the results may serve as exploratory findings rather than definitive conclusions.

4.1.3 Evaluation Metrics

- R^2 **(Coefficient of Determination)**: The R^2 value for the aggregated dataset was 0.211, indicating that approximately 21.1% of the variance in `avg_sales` was explained by the predictors. For the yearly dataset, including the `Year` variable improved the R^2 to 0.247. While these values suggest a moderate fit, they highlight that linear regression may not fully capture the complexity of the data.
- **Mean Squared Error (MSE)**: The MSE for the aggregated dataset was approximately 57,576,706, and for the yearly dataset, it was 86,143,616. These metrics reflect the average squared deviation between the observed and predicted sales, with lower values indicating better predictive accuracy. Although the error remains significant, the results provide baseline insights into key predictors of car sales.

4.1.4 Results

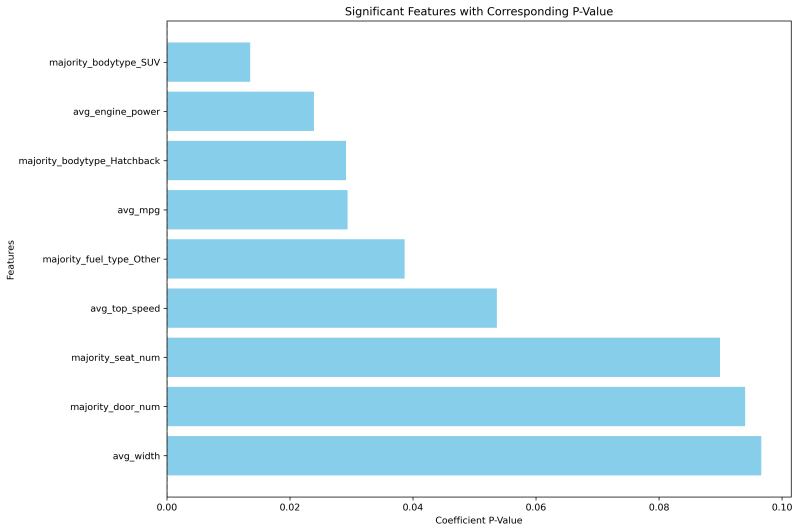


Figure 7: Variable Significance from Linear Regression of Aggregated Data (Significant Variables Only)

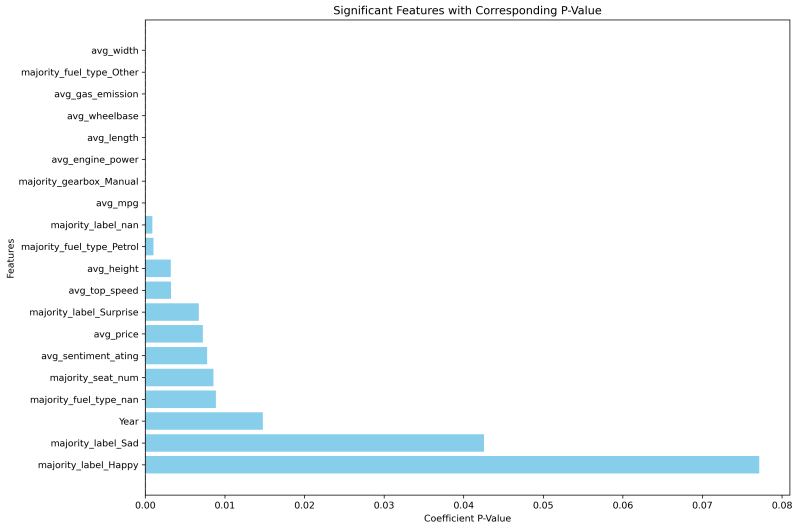


Figure 8: Variable Significance from Linear Regression of By Year Data (Significant Variables Only)

- **Key Predictors:** The linear regression model identified several significant predictors influencing car sales:
  - `avg_price`**:** A negative coefficient indicates that higher car prices are associated with lower sales, likely reflecting affordability constraints for consumers.
  - `majority_bodytype`**:** The majority body type of a vehicle significantly influences sales, reflecting consumer preferences for specific car categories such as SUVs, sedans, or hatchbacks.
  - `avg_engine_power`**:** A positive coefficient suggests that more powerful engines are preferred, particularly in performance-oriented or luxury segments.
  - `avg_mpg`**:** Fuel efficiency positively influenced sales, aligning with consumer preferences for cost-effective and environmentally friendly vehicles.
  - `avg_gas_emission`**:** Higher emissions showed a significant negative association with sales, further reinforcing trends toward sustainable and green technologies.
  - `Year`**:** In the yearly dataset, the positive coefficient for `Year` revealed a modest upward trend in car sales over time, suggesting growing market demand or population-driven growth.
- **Insights and Interpretation:** The findings reveal that car sales are influenced by a combination of functional attributes and temporal trends. Specifically:
  - Fuel efficiency (`avg_mpg`) and engine power (`avg_engine_power`) are significant drivers of consumer demand, highlighting the importance of balancing performance with efficiency.
  - Environmental concerns are increasingly prominent, as evidenced by the negative impact of emissions (`avg_gas_emission`) on sales. This underscores the importance of sustainable engineering.
  - Price (`avg_price`) remains a critical factor, as affordability continues to drive purchasing decisions in most market segments.
- **Model Limitations:** Despite identifying significant predictors, the linear regression model had limited explanatory power, as indicated by the low R^2. Additionally, the yearly dataset's lack of independence and possible temporal autocorrelation may have biased the coefficient estimates. Future models, such as time-series analysis or Random Forests, can better address these challenges by capturing both temporal and non-linear patterns.

4.2 Random Forest

4.2.1 Assumptions

- **Independence of Observations:** Assumes that each car model's data is independent, which was ensured by aggregating the data by model to eliminate repeated entries.
- **No Assumption of Linearity:** Captures non-linear relationships naturally, such as interactions between `avg_engine_power` and `majority_fuel_type`, making it particularly suited for complex datasets.
- **Handling of Missing Data:** Assumes consistent patterns of missingness, which were addressed through reliable imputation methods for any missing values in predictors.

4.2.2 Model Description

- A Random Forest model was implemented using the `ranger` package with the following specifications:
  - **Number of Trees:** 1000
  - **Sample Size:** 488 observations
  - **Number of Predictors:** 25 independent variables
  - **Mtry:** 5 (number of variables randomly selected at each split)
  - **Node Size:** 5 (minimum size of terminal nodes)
  - **Variable Importance Mode:** Permutation-based importance
- The model aimed to predict `avg_sales` using features like `avg_price`, `avg_gas_emission`, and `avg_engine_power`, along with categorical variables like `majority_fuel_type` and `majority_bodytype`.

4.2.3 Evaluation Metrics

- **Out-of-Bag (OOB) Prediction Error (MSE):** 57,793,370
- **R-Squared (OOB):** 0.2079

```
> print(rf_model)
Ranger result

Call:
 ranger(avg_sales ~ Maker + avg_price + avg_gas_emission + avg_engine_size +      majority_fuel_type + avg_engine_power + avg_wheelbase + avg_heigh
t +      avg_width + avg_length + avg_mpg + avg_top_speed + majority_seat_num +      majority_door_num + majority_engine_size + majority_bodytype +
majority_gearbox + avg_prob_angry + avg_prob_disgust + avg_prob_fear +      avg_prob_happy + avg_prob_sad + avg_prob_surprise + avg_prob_neutral +
majority_label, data = cleaned_df, importance = "permutation",      num.trees = 1000, seed = 123)

Type:                             Regression
Number of trees:                  1000
Sample size:                      488
Number of independent variables:  25
Mtry:                             5
Target node size:                 5
Variable importance mode:         permutation
Splitrule:                        variance
OOB prediction error (MSE):       57793370
R squared (OOB):                  0.2079003
```

Figure 9: Random Forest Model Output Summary

4.2.4 Results

- The Random Forest model identified the most influential predictors of car sales:
  - `avg_mpg`, `avg_gas_emission`, and `avg_top_speed` emerged as the top three predictors of `avg_sales`.
    - `avg_mpg`**:** This variable's high importance highlights that fuel efficiency is a critical factor influencing consumer preferences. Vehicles with better mileage are generally more appealing to cost-conscious buyers, especially in markets with high fuel prices.
    - `avg_gas_emission`**:** Lower emissions tend to align with increasing environmental regulations and consumer awareness about sustainability. Cars with lower emissions are often perceived as more environmentally friendly, which can positively impact sales.
    - `avg_top_speed`**:** While not a practical metric for everyday driving, a higher top speed may indicate better performance and power, which could appeal to performance-oriented consumers or reflect overall engineering quality.
  - Other important variables included `avg_engine_size`, `avg_height`, and `majority_engine_size`.
    - `avg_engine_size`**:** Larger engines are often associated with higher performance and power, which could explain their influence on sales. However, they might also reflect higher fuel consumption and emissions, potentially balancing their appeal across different market segments.
    - `avg_height`**:** This variable likely represents consumer preferences for certain vehicle types, such as SUVs or crossovers, which are typically taller than sedans. The growing popularity of SUVs in global markets could explain its relevance.
    - `majority_engine_size`**:** This categorical variable indicates the most common engine size across a car model's configurations. Its importance might reflect consistency in power and performance across trim levels, which can influence overall consumer trust and appeal.
- These findings highlight several important implications for automakers. Fuel efficiency (`avg_mpg`) and emissions (`avg_gas_emission`) emerged as critical predictors of sales, underscoring the importance of investing in technologies that improve these metrics to remain competitive in the market. Consumers are increasingly aware of environmental impacts and fuel costs, making these attributes essential for driving demand. Additionally, the significance of performance-related variables such as `avg_top_speed` and `avg_engine_size` suggests that customers seek a balance between efficiency and performance. This creates an opportunity for automakers to cater to diverse consumer preferences by offering trims that appeal to both performance-oriented buyers and those focused on cost-effectiveness and environmental considerations. To enhance model performance further, incorporating additional data sources such as marketing expenditures, economic indicators, or customer

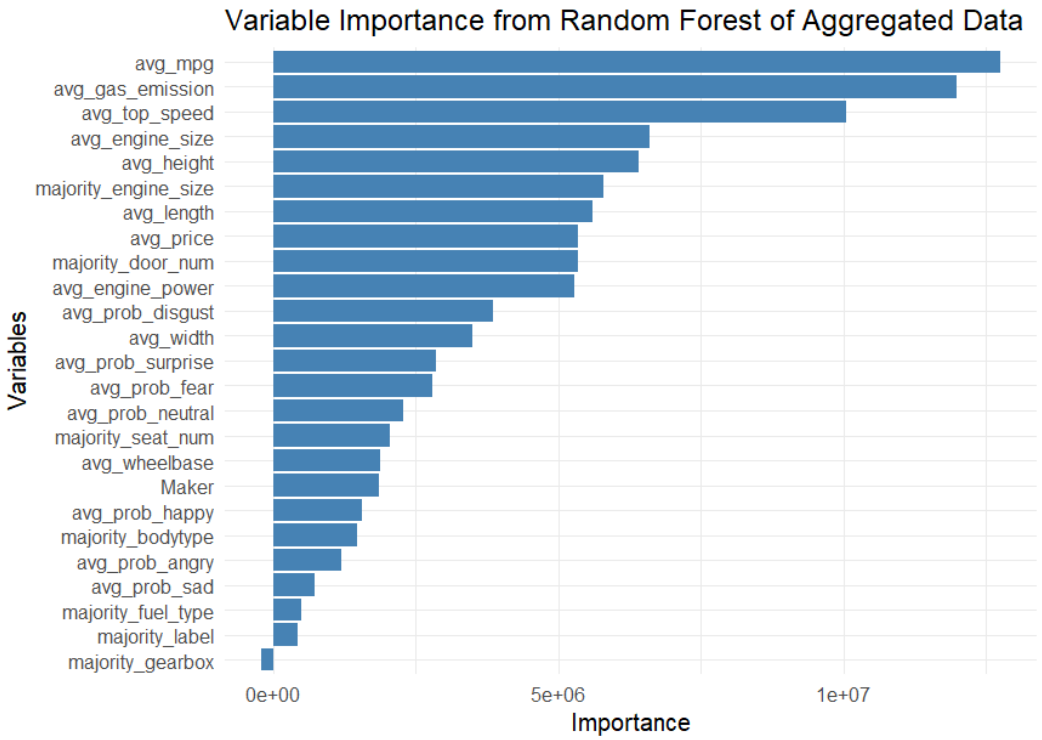


Figure 10: Variable Importance from Random Forest of Aggregated Data

reviews could provide a more comprehensive understanding of the factors influencing car sales.

- Despite identifying key predictors, the Random Forest model achieved a relatively low R^2 value (OOB $R^2 = 0.2079$), indicating that a substantial portion of the variability in `avg_sales` remains unexplained. This limitation may stem from several factors. First, the dataset may lack critical predictors, such as variables capturing marketing efforts, dealer-level incentives, or broader macroeconomic conditions, which can significantly affect sales. Second, the inherent noise in the sales data—driven by unpredictable consumer behavior, seasonal trends, and regional market dynamics—likely reduces the model's explanatory power. Lastly, the aggregated nature of the data, which consolidates sales across years and configurations, might obscure temporal and regional variations that could better capture sales patterns. Addressing these limitations through richer datasets and more granular segmentation could improve the model's ability to explain and predict car sales effectively.

4.3 Neural Networks (BPNN)

4.3.1 Assumptions

- **Non-Linearity:** The BPNN (Backpropagation Neural Network) assumes it can effectively model non-linear relationships among predictors, such as `avg_engine_power`, `avg_prob_angry`, and categorical variables like `majority_gearbox`.
- **Feature Scaling:** Predictors were standardized using z-score normalization, as neural networks require scaled input for stable gradient descent convergence.
- **Data Size:** BPNNs require sufficiently large datasets to avoid overfitting. While the dataset size (487 observations) is moderate, techniques like dropout regularization and early stopping were implemented to mitigate overfitting.

4.3.2 Model Description

- A BPNN was implemented as a multi-layer perceptron with the following architecture:

- **Input Layer:** 40 input features for the aggregated dataset, which included numerical and one-hot encoded categorical predictors.
- **Hidden Layers:** Three hidden layers with 128, 64, and 32 neurons, respectively. Each layer utilized the ReLU activation function to introduce non-linearity.
- **Dropout Layers:** A dropout rate of 0.3 was applied after each hidden layer to prevent overfitting.
- **Output Layer:** A single neuron with a linear activation function to predict `avg_sales`.

- **Optimization:** The model was trained using the Adam optimizer with a learning rate of 0.1, batch size of 64, and for 100 epochs. The MSE loss function was used as the training criterion.
- **Hyperparameter Tuning:** Random search was conducted over hyperparameters, including learning rate, dropout rate, and number of neurons/layers. The best hyperparameters were:
  - Learning rate: 0.1, Dropout rate: 0.3, Hidden Units: 128, Batch size: 64, Epochs: 100, Layers: 3.

4.3.3 Evaluation Metrics

- **Mean Absolute Error (MAE):** Measures the average magnitude of prediction errors.
- **Mean Squared Error (MSE):** Quantifies the average squared differences between predicted and actual `avg_sales`.

4.3.4 Results

- **Aggregated Dataset:**
  - The model achieved a **Test MAE of 3,312** and a **Test MSE of 55,026,580**, showing moderate predictive performance.
  - Training metrics were slightly better, with a **Training MAE of 2,948** and **Training MSE of 26,212,184**, suggesting that overfitting was minimal but still present.

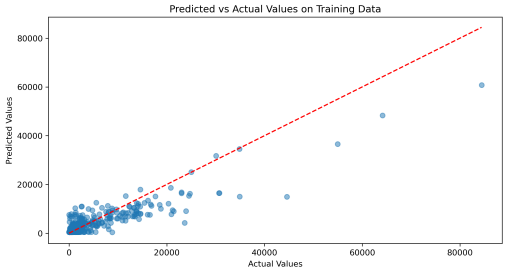


Figure 11: Train and Test Loss Curves

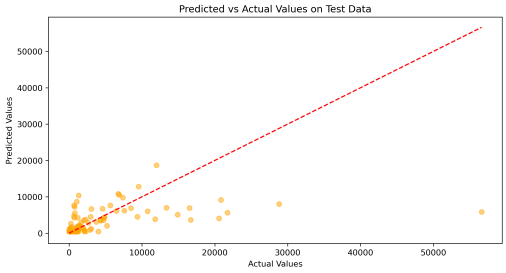


Figure 12: Train and Test Accuracy Curves

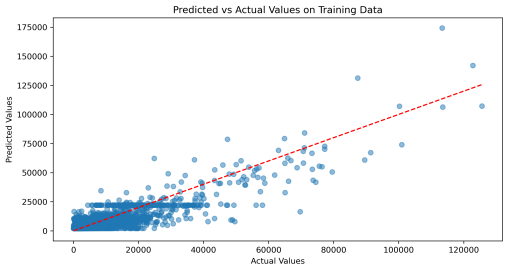


Figure 13: Train and Test Loss Curves

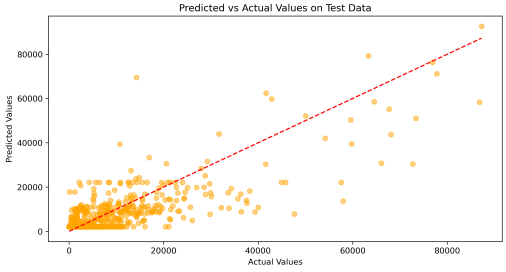


Figure 14: Train and Test Accuracy Curves

- **Insights:**
  - The BPNN successfully captured non-linear relationships among predictors, particularly between numerical features such as `avg_engine_size` and emotion-based probabilities (`avg_prob_sad`, `avg_prob_angry`).
  - However, the results indicate that neural networks may not fully leverage the dataset's structure, likely due to its limited size and complexity.

- **Limitations:**
  - **Dataset Size:** Neural networks generally perform better on larger datasets. With only 487 observations for the aggregated dataset, the model's ability to generalize remains limited.
  - **Model Interpretability:** The black-box nature of BPNNs makes it difficult to derive clear insights into feature importance, unlike simpler models such as linear regression.
  - **Training Complexity:** Training required careful hyperparameter tuning to balance performance and overfitting, further complicating implementation.
  - **Yearly Data Concerns:** Temporal trends in the yearly dataset were not explicitly handled, potentially introducing bias when applying the model to yearly sales.
- **Final Reflections:** Despite its challenges, implementing the BPNN provided an opportunity to explore the potential of neural networks for this task. While the results did not vastly outperform simpler models, the effort reflects an important step toward leveraging advanced machine learning techniques. With larger datasets and further tuning, neural networks could reveal deeper insights into car sales determinants.

4.4 RNN

4.4.1 Assumptions

- **Sequential Data Assumption:** The model expects temporal or sequential dependencies within the dataset, thus aiming to capture year-to-year variations in car sales trends.
- **Feature Scaling:** Scaled predictors are required to ensure stable gradients during training, as implemented using Min-Max scaling in the code.
- **Complete Data Assumption:** The model assumes no missing target values during training, achieved by dropping such instances.

4.4.2 Model Description

- The RNN architecture is implemented using a Simple RNN layer, designed to capture sequential dependencies over time.
- Both year-specific predictors (e.g., `avg_price`, `avg_prob_happy`) along with temporal patterns are utilized to predict annual sales.
- The functional form of the model can be represented as:

$$\mathbf{y}_t = f(\mathbf{X}_t; \Theta) = \text{Dense}(\text{Dropout}(\text{SimpleRNN}(\mathbf{X}_t)))$$

  where $\mathbf{X}_t$ represents the sequence of input features at time t, and $\mathbf{y}_t$ is the output sales prediction.

4.4.3 Evaluation Metrics

- **Mean Absolute Error (MAE):** Measures the absolute difference between predicted and observed sales values.

4.4.4 Results

The RNN model, implemented using a Simple RNN layer, struggles to accurately predict car sales. The Mean Absolute Error (MAE) of 31,766 on the original sales scale indicates significant discrepancies between predicted and actual sales values.

The learning curve 15 and prediction results 16 highlight the model's limitations. The consistent gap between training and validation loss suggests potential overfitting. Moreover, the scatter plot of predicted versus actual values shows substantial deviation from the line of perfect prediction.

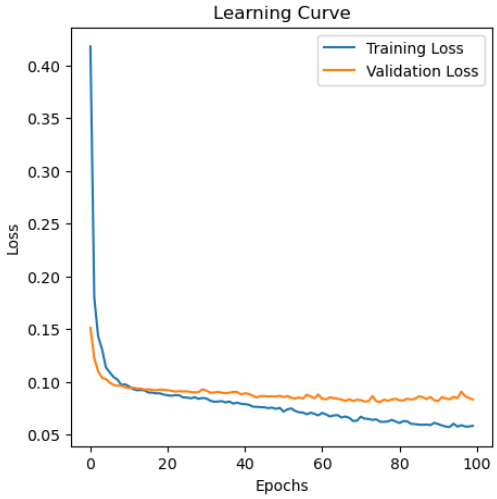


Figure 15: Train and Test Loss Curves

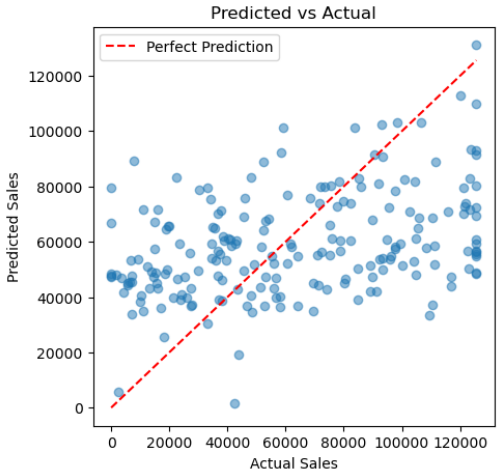


Figure 16: Prediction

4.4.5 Reflections

- **Model Complexity:** The RNN architecture may not capture complex sequential dependencies.
- **Feature Representation:** Additional features or re-engineering might be necessary to improve predictive accuracy.
- **Data Limitations:** Limited or noisy data might impair the model's learning capability. Also, the volume of data might not be enough to train an RNN.

5 Conclusion

In this study, we analyzed the determinants of car sales using a multifaceted approach that combined traditional statistical methods with advanced machine learning techniques. Our investigation focused on a diverse range of factors, including car specifications, consumer sentiment, and the aesthetic "facial expressions" of vehicles. Through extensive data processing, exploration, and modeling, we uncovered valuable insights into the variables that significantly influence sales performance.

Contrary to initial hypotheses, aesthetic features like car "facial expressions" demonstrated limited impact on sales compared to functional attributes such as fuel efficiency, engine power, and price. Our findings emphasize the importance of practical considerations in consumer decision-making, with fuel efficiency and lower emissions emerging as critical drivers of demand in an environmentally conscious market. Additionally, the interplay between design trends and consumer preferences, as reflected in manufacturer-specific styling choices, highlights the nuanced role of aesthetics in shaping brand identity and market appeal.

While our models, including OLS regression, Random Forests, and Backpropagation Neural Networks, revealed key predictors and provided a foundation for understanding sales dynamics, the relatively low R^2 values across models suggest that additional factors—such as marketing strategies, economic indicators, or dealer-level incentives—may play a significant role in influencing car sales. Incorporating these variables into future analyses could enhance the predictive power of the models and provide a more comprehensive understanding of the automotive market.

In conclusion, our study underscores the need for a balanced approach that integrates functional, aesthetic, and contextual factors to optimize sales strategies. By leveraging the insights gained from this analysis, automotive manufacturers can better align their design, production, and marketing efforts with evolving consumer preferences and market demands. Future research with enriched datasets and advanced modeling techniques holds the potential to further refine our understanding of this dynamic and complex industry.

References

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

J. Huang, B. Chen, L. Luo, S. Yue, and I. Ounis. Dvm-car: A large-scale automotive dataset for visual marketing research and applications, 2023. URL https://arxiv.org/abs/2109.00881.

Y. Khaireddin and Z. Chen. Facial emotion recognition: State of the art performance on fer2013, 2021. URL https://arxiv.org/abs/2105.03588.