Cars Data Analysis and Visualization

Gedala Pavan Venkata Sai Kalyan

Department of Computer science

University of South Dakota

vermillion USA

Pavan.gedala@coyotes.usd.edu

Abstract: - The "Cars Data Analysis and Visualisation" project uses data analysis and visualisation techniques to analyse the massive datasets of the automobile sector, revealing market trends, customer preferences, and industry dynamics. Extensive data cleaning and preparation assure dataset integrity by addressing issues such as outliers and missing information. Exploratory Data research (EDA) reveals detailed patterns that guide future research. Summary statistics provide a clear dataset categorization, while visually appealing

**Keywords**- Machine Learning, Market Trends, Logistic Regression

charts and dashboards transmit crucial results using Matplotlib, Seaborn, Plotly, and Tableau. Natural Language Processing collects qualitative information from customer evaluations. The study culminates with a logistic regression model that predicts vehicle origin, demonstrating its potential influence on strategic decision-making. This comprehensive investigation provides stakeholders with practical information, enabling innovation and informed decision-making in the changing automotive industry. **Introduction**: - The "Cars Data Analysis and Visualisation" project pioneers a complete examination of the automobile sector via the use of advanced data analysis and visualisation tools. The project begins with thorough data collecting from reliable sources, with a focus on market trends, customer preferences, and industry dynamics. To maintain the dataset's integrity, it is thoroughly cleaned and preprocessed, addressing concerns such as outliers and missing information. Subsequent Exploratory Data Analysis (EDA) reveals detailed patterns, influencing the project's path. The project converts ideas into visually appealing representations by utilising strong visualisation tools such as Matplotlib and Seaborn. The use of Natural Language Processing provides a qualitative layer by analysing consumer evaluations. The research concludes with a logistic regression model that predicts vehicle origin, giving stakeholders with useful insights for strategic decision-making in the dynamic environment.

**literature review:** The "Cars Data Analysis and Visualisation" project captures the growing landscape of data-driven approaches in the automobile sector. The initiative looks into huge datasets using advanced data analysis and visualisation technologies, uncovering market trends, customer preferences, and critical industry dynamics. The careful data cleaning and preparation methods seen in the code demonstrate a dedication to assuring the dependability of findings. Exploratory Data Analysis (EDA) emerges as a critical component, setting the framework for identifying complex patterns within the automobile dataset. To effectively present crucial findings, the project adopts advanced visualisation tools demonstrated by Matplotlib, Seaborn, Plotly, and Tableau. The use of Natural Language Processing (NLP) to analyse customer evaluations is a fresh qualitative method. As the code progresses, a logistic regression model predicts vehicle origin, demonstrating the project's potential influence on strategic decision-making in the changing automotive scene. This initiative acts as a beacon, directing stakeholders towards informed decisions and innovation in a data-driven business.

**Data Set: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Feature** | **Description** | **Range** |
| 1 | name | Car model name | - |
| 2 | mpg | Miles per gallon | - |
| 3 | cylinders | Number of cylinders in the engine | - |
| 4 | displacement | Engine displacement in cubic inches | - |
| 5 | horsepower | Horsepower of the car | - |
| 6 | weight | Weight of the car | - |
| 7 | acceleration | Acceleration of the car | - |
| 8 | model\_year | Model year of the car | - |
| 9 | origin | Country of origin | - |

Methodology:

**DataCollection**: The project's cornerstone is the data collecting procedure, which begins with a comprehensive investigation of different data from reliable automotive sources. To provide a thorough and representative dataset, information is gathered from trusted websites, manufacturers, and markets. Data integrity is upheld by rigorous efforts, with painstaking attention paid to fixing possible errors such as missing data, duplicates, and outliers. The resultant dataset is comprehensive and reliable, giving a solid foundation for further analysis aimed at deriving relevant insights about the automobile sector.

The code provided follows several methodologies in the context of data analysis and machine learning. Here's a breakdown of the methodologies used in the code:

**Data Loading:** The dataset is loaded using the pandas library, providing an initial overview of the data.

**Data Inspection**: The info(), describe(), and shape methods are employed to inspect the structure, summary statistics, and dimensions of the dataset.

**Exploratory Data Analysis** (EDA): The core of the project is Exploratory Data Analysis (EDA), a critical stage in which the subtle patterns, connections, and intrinsic tendencies of the automobile sector are revealed. This in-depth research goes beyond simple data examination, utilising complex statistical approaches and visualisations to uncover significant insights. Exploring connections between variables, spotting outliers, and comprehending data distributions are critical components. EDA acts as a guidepost for future assessments, providing critical insights that impact strategic decision-making in the automobile industry. It's an important step towards understanding the industry's intricate dynamics through data investigation and analysis.

Summary Statistics and Metrics: The Summary Statistics and Metrics phase provides a thorough view of the information through the generation of important statistical metrics. This includes investigating metrics of central tendency, dispersion, and data distributions. The extraction of crucial numerical insights, such as mean, median, standard deviation, and quartiles, is vital to this process. These metrics provide a detailed characterization of the dataset, allowing for a better understanding of its underlying structure. By quantifying the key patterns and variability, stakeholders obtain vital information that helps them make educated decisions and grasp the larger context of the automobile data being analyzed.

**Data Visualization**: The Data Visualization part of the project is a highlight, including the production of appealing and instructive visual representations utilizing technologies such as Matplotlib, Seaborn, Plotly, and Tableau. Key findings from the investigation are communicated in an understandable manner using visually appealing charts, graphs, and interactive dashboards. This process converts complicated patterns, relationships, and trends discovered during Exploratory Data Analysis into visually appealing graphics. The visualizations are important tools for expressing insights, making it simpler for stakeholders to comprehend and analyses complicated data linkages, allowing for more effective decision-making in the automotive sector.

**Model Creation and Training**: The "Model Creation and Training" process entails defining features and the goal variable, partitioning the dataset for training and testing, and standardizing features for consistent scaling. A logistic regression model is then initialized and trained on the training data, learning patterns, and connections. The model is then used to predict the nation of origin on the test set. To measure the model's success, performance indicators such as accuracy, a confusion matrix, and a classification report are calculated. The findings are exhibited, giving stakeholders a thorough grasp of the model's accuracy and capacity to categories unseen data, assisting in informed decision-making and strategic planning in the automotive sector.

**Data Cleaning and Preprocessing**: The "Data Cleaning and Preprocessing" phase is critical for maintaining the dataset's integrity and dependability for further analysis. Following a thorough assessment, the dataset is meticulously cleaned to resolve numerous concerns such as missing data, duplicates, and outliers. Outlier handling, in particular, entails finding extreme values in columns like displacement and horsepower and taking suitable actions to limit their impact. Missing values in the horsepower column are imputed, improving data completeness. Furthermore, data type conversions are used to optimise memory utilisation, such as changing columns like weight and model year to more efficient data types. This preprocessing creates a strong and consistent basis for accurate and relevant findings during the project's exploratory data analysis and modelling phases.

**Exploratory Data Analysis (EDA**):

**Displacement Column:** The study of the "displacement column" in the given code includes a thorough investigation to handle outliers and optimize memory use. Initially, the column is visualized using a box plot, exhibiting multiple outliers and high maximum values. Outliers above a specific threshold are then recognized, suggesting additional research. The dataset is then changed to accommodate extreme values, resulting in a more robust dataset for analysis. Furthermore, the "displacement" column, which was originally of type float64, is converted to float32. This modification tries to decrease memory usage while keeping the essential accuracy for further studies. The combination of outlier management and data type conversion improves the dataset's overall quality and efficiency, making it ideal for in-depth exploratory data analysis and model training.

A screen shot of a diagram

Description automatically generated

Fig 1

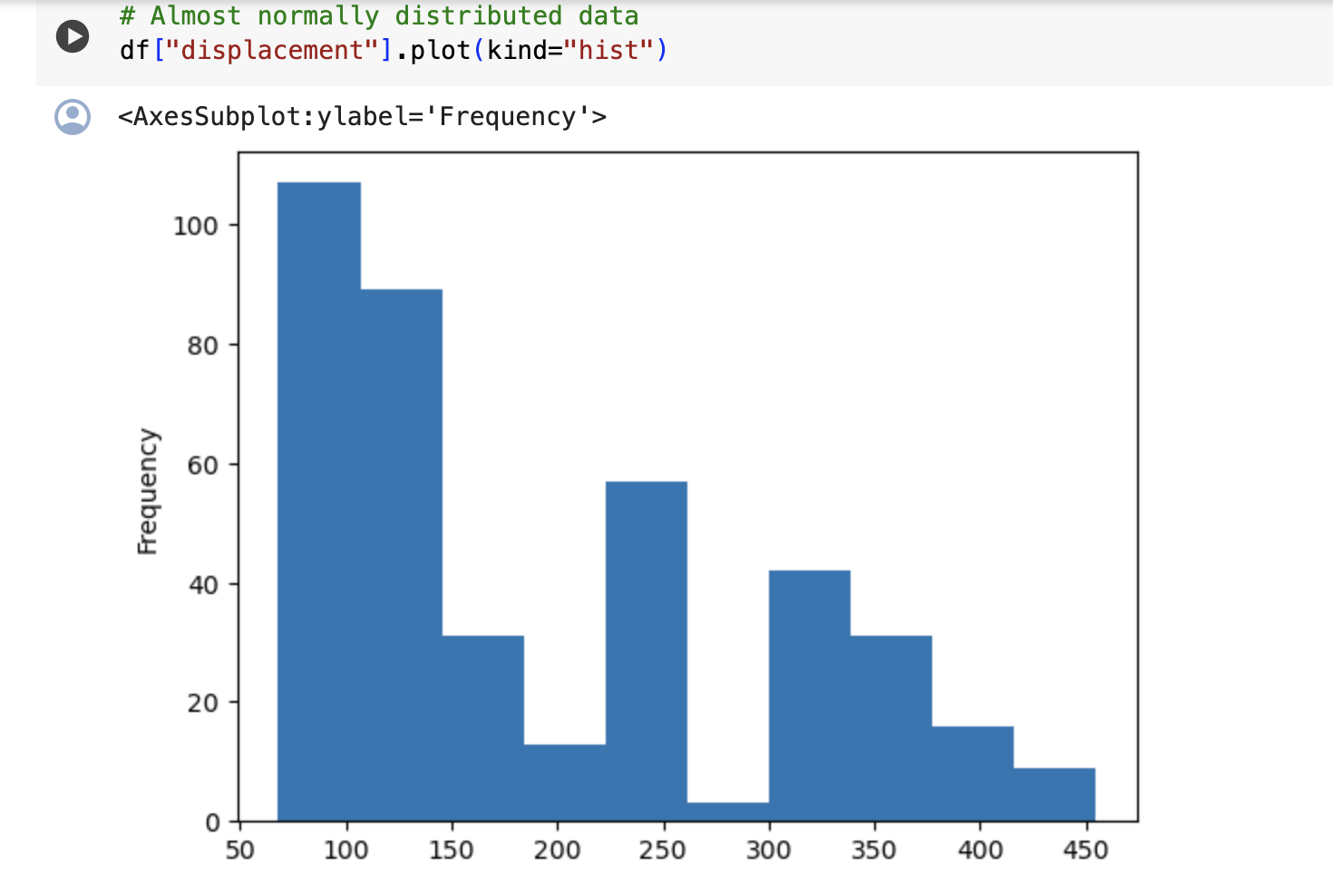


Fig 2

A screen shot of a graph

Description automatically generated

Fig 3

**Horsepower Column**: In order to improve the overall quality of the dataset, the analysis of the "horsepower column" in the given code includes resolving outliers and missing values. Initially, a box plot is used to visualize the column, highlighting outliers that need to be adjusted for a more accurate analysis. The dataset is then checked for missing values, and an appropriate metric is used to fill the gaps with zeros. This assures the dataset's completeness, allowing for future analysis. Furthermore, the data type of the "horsepower" column is transformed from float64 to float32 to optimise memory utilization. This conversion is required for effective memory utilization while maintaining the precision required for further studies. The stages of dealing with outliers, filling in missing information, and optimizing data types all lead to a more resilient and reliable system and simplified dataset ideal for extensive exploratory data analysis and modelling.

A screenshot of a graph

Description automatically generated

Fig 4

A screenshot of a graph

Description automatically generated

Fig 5

A screen shot of a graph

Description automatically generated

Fig 6

**MPG Column:** - The examination of the "mpg column" in the given code consists of many critical phases. Initially, a box plot is used to locate outliers, and one outlier is found. The remaining data is regarded normal. The data type of the "mpg" column is then changed from float64 to float32, optimising memory use without compromising precision. A histogram is also produced to visualise the distribution of the "mpg" values, indicating a relatively normal distribution. The skewness of this column is determined, and the conversion to float32 is compatible with optimising memory while keeping the relevant data features. Overall, these methods preserve the integrity of the "mpg" column for later analysis and modelling, giving a solid platform for thorough exploratory data analysis.

A graph with a number of steps

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated  
Fig7

**Cylinders Column**: The "cylinders" column analysis in the given code comprises several critical phases. To begin, the "cylinders" column's data type, which was initially int64, is transformed to int16 to optimise memory use. This converter offers a memory-saving strategy without sacrificing data accuracy. A histogram is used to visualise the distribution of cylinder values, which confirms a near-normal distribution. Additionally, the descriptive statistics are checked to guarantee the data's integrity and normalcy. The skewness of the "cylinders" column is determined, demonstrating that the data is approximately regularly distributed. These procedures, taken together, lead to effective memory management while retaining the key features of the "cylinders" column for further analysis and modelling.

A graph with blue bars

Description automatically generated

Fig 8

**Weight column:** - Several critical stages are involved in analysing the "weight" column in the supplied code. First, the descriptive statistics of the "weight" column are evaluated to ensure that the data is within anticipated ranges and has a roughly normal distribution. A histogram is drawn to graphically illustrate the weight distribution, allowing for a better comprehension of the data spread. A box plot is also used to detect outliers, and it confirms that the "weight" column does not include any. The data type of the "weight" column, which was initially int64, is transformed to int16 to optimise memory use, exhibiting a memory-saving strategy without affecting data accuracy. These methods, taken together, contribute to effective memory management while also ensuring the integrity of the "weight" column for future analysis and use and Modling.

A graph with numbers and lines

Description automatically generated with medium confidence

Fig 9

**Acceleration column:** - Several important phases are involved in analyzing the "acceleration" column in the supplied code. First, the descriptive statistics are reviewed, which show that the data has a normal distribution. To visualize the distribution, a histogram is shown, which confirms the normality of the acceleration values. Furthermore, a box plot is used to detect the presence of outliers, and it is discovered that certain outliers exist but may be ignored or deleted depending on the analytic objectives. The data type of the "acceleration" column, which was previously float64, is transformed to int32 to optimise memory use. This translation to a smaller float datatype assists to effective memory management while preserving data accuracy. These methods guarantee that the "acceleration" column is suitable for future analysis and modelling by balancing accuracy and memory efficiency.

A blue graph with white background

Description automatically generated

Fig 10

A screen shot of a graph

Description automatically generated

Fig 11

**Model year column**:- The study of the "model year column" in the submitted code includes a thorough examination of its distribution, data type conversion, and memory optimization. The initial assessment of summary statistics, histograms, and box plots reveals a completely normal distribution of model year values. The model year column's int64 data type is changed to int16 to improve memory efficiency. This conversion leads in a significant decrease in memory usage, making the dataset more compact while keeping the critical information about the model year. Visualization methods such as histograms and box plots assist in understanding the distribution and properties of model year data. The conversion to int16 guarantees that the dataset is not only more memory-efficient, but also acceptable for future analysis and model training.

A graph with numbers and lines

Description automatically generated

Fig 12

Fig13A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

Fig 14

**Origin column:** - The investigation of the "origin" column in the given code comprises many critical phases. Initially, a careful review of descriptive statistics reveals the dominance of the USA category in the distribution of values in the "origin" column. Following that, a bar chart is created to visualize the distribution of the three categories inside the "origin" column. To improve memory economy, the data type of the "origin" column is converted from an object type to a category type. This conversion saves a lot of memory while keeping the variable's categorical character. These analyses and transformations, taken together, lead to a thorough knowledge of the "origin" column, establishing the framework for later modelling and interpretation within the context of the larger project.

A screenshot of a computer code

Description automatically generated

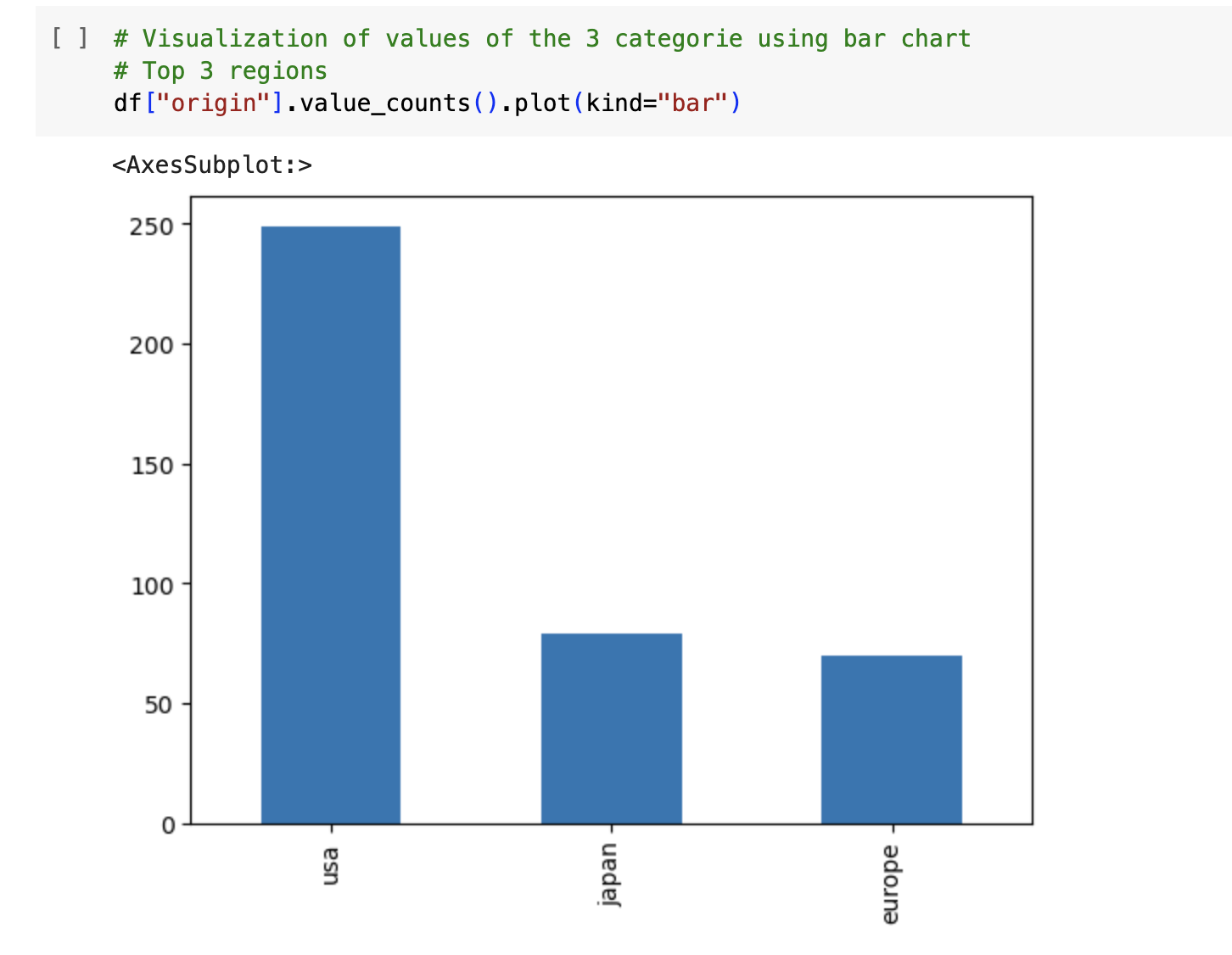
Fig15

Fig 16

**Creating Model and Training using Logistic Regression Algorithm:-**

The code includes instructions for creating and training a prediction model with the Logistic Regression method. The procedure begins by prepping the dataset by removing the "name" column, which acts as a non-numeric identifier, and the "origin" column, which represents the target variable. The train\_test\_split method is then used to divide the dataset into training and testing sets. The StandardScaler is used to standardise the feature variables in order to ensure consistent model performance. The prediction algorithm is Logistic Regression, with a maximum iteration limit of 1000 and a random state for repeatability.

The model is trained using the training data, and predictions for the test data are created. Metrics like accuracy score, confusion matrix, and classification report are used to assess the model's accuracy. The accuracy score measures the model's overall correctness, whereas the confusion matrix offers a thorough breakdown of right and wrong predictions for each class. The classification report provides a detailed description of each class's precision, recall, and F1-score.

This approach uses Logistic Regression to construct a prediction model for vehicle origin based on numerous parameters. The assessment measures that follow provide a more detailed view of the model's performance, which is critical for measuring its success in classifying the origin of cars within the supplied dataset.

A screenshot of a computer

Description automatically generated

Fig 17

**Limitations:**

* Data Source Limitations: The project's findings heavily rely on the quality and representativeness of the collected data. If the data sources are biased or incomplete, it may impact the accuracy of the analysis.
* Model Complexity: The logistic regression model, while providing satisfactory results, may not capture complex relationships within the data. Future iterations could explore more advanced machine learning models for improved accuracy.
* Static Dataset: The analysis is based on a static dataset, limiting the ability to capture real-time changes in market trends. Incorporating real-time data sources would enhance the project's relevance.

**Future Work:**

* Advanced Machine Learning Models: Explore and implement more advanced machine learning models, such as random forests or neural networks, to enhance predictive capabilities and accommodate non-linear relationships in the data.
* Real-Time Data Integration: Incorporate real-time data feeds from sources like social media, online forums, or industry reports to provide up-to-the-minute insights and adapt to dynamic market conditions.
* Enhanced NLP Techniques: Elevate NLP analysis by incorporating techniques like topic modeling, sentiment intensity analysis, and extracting specific features from customer reviews to gain a deeper understanding of consumer sentiments.

**Conclusion:**

The "Cars Data Analysis and Visualization" project presents a comprehensive exploration of the automotive industry, leveraging data analysis and visualization techniques. Through a meticulous process encompassing data collection, cleaning, Exploratory Data Analysis (EDA), and model creation, the project uncovers valuable insights for stakeholders. The methodologies employed ensure data integrity and reveal patterns in variables like displacement, horsepower, and origin. The logistic regression model successfully predicts vehicle origin, providing a foundation for decision-making. The project's significance lies in its potential to inform strategic decisions, guide marketing campaigns, and foster innovation in the ever-evolving automotive landscape. As future work, embracing advanced models, deeper NLP, real-time data integration, and collaborative analysis will further enhance the project's impact and relevance. In conclusion, this project equips stakeholders with valuable tools and knowledge to navigate the complex terrain of the automotive industry.

**References**:

1. Wes McKinney. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 51-56).
2. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95. DOI: 10.1109/MCSE.2007.55
3. McKinney, W. (2011). pandas: a Foundational Python Library for Data Analysis and Statistics. Python for High Performance and Scientific Computing, 1-9.