**Used Cars Price Analysis for Fair Pricing and Decision Making**

Vikram Kumar Reddy Ayaluri  
11711384  
Dept. Computer Science  
University of North Texas  
Denton, TX

Tharun Kuravadi Sathish Babu  
11659825  
Dept. Computer Science   
University of North Texas  
Denton, TX Lekhitanand Bandi  
11727419  
Dept. Computer Science  
University of North Texas  
Denton, TX

Sai Vignesh Rayal Rolla  
11569007  
Dept. Computer Science  
University of North Texas  
Denton, TX Akshadha Reddy Itikela   
11700379  
Dept. Computer Science  
University of North Texas  
Denton, TX

[Github Link for](https://github.com/vikram741/ea-group10/tree/main/Final%20Submission) [Used Cars Price Analysis for Fair Pricing and Decision Making](https://github.com/vikram741/ea-group10/tree/main/Final%20Submission)

[Python Code Link for Used Cars Price Analysis for Fair Pricing and Decision Making](https://colab.research.google.com/drive/1n4_-b_YhRhMOFDG8WVx-9N9SEghWF1zQ#scrollTo=_YgncPwNg0wF)

# **GOALS AND OBJECTIVES**

**Motivation:**

The used car market is a significant industry, and pricing plays a crucial role in attracting buyers and ensuring sellers receive fair value for their vehicles. Determining fair prices for used cars is essential for both buyers and sellers, as it affects financial decisions and market stability. In this digital world, humans are mostly spent time using internet and technologies. Where humans are going through many different online platforms to sell their products. To earn good value for the product, the consumer visits the most well-known platforms such as social media marketing, advertisements, broadcasting or by consumer recommendations. So, there are many types of websites and social media platforms such as OLX, telegram groups, Instagram product promotions and many more. But the products must meet the customers trust with assurance that they are spending their allowance on the right product for the right price.

**Significance**:

Fathalla et al. [1]. proposed the biggest concern is price prediction for used vehicles. Nowadays the automotive industries are rapidly increasing and most of these industries are releasing their products with new features incorporating and tries to attract the customers. So, In this way the previous models are left behind. Also, consumers are mostly focused on the models of the product that even in the future upcoming models are released the price drop for the current model should at least meet their minimum round price.

**Objectives:**

* The goal of this project is to perform an exploratory data analysis on a dataset containing information about cars, with a primary focus on understanding the factors that influence car prices. The objectives include:
* Data Cleaning: Initially, we inspected the dataset for missing values and addressed them. The data was almost clean, with only the 'engine\_capacity' column containing a few null values, which were subsequently dropped.
* Descriptive Statistics: We conducted a descriptive statistical analysis to gain insights into the distribution and central tendencies of the numerical features in the dataset. This provided a preliminary understanding of the data.
* Visualizations: Utilizing various visualization techniques, we explored relationships between car prices and different features such as transmission type, car color, age, engine type, and engine capacity. These visualizations helped us identify trends and patterns in the data.
* Feature Engineering: We created a new feature, 'car\_age,' by calculating the age of each car based on the 'year\_produced' column. This feature was then used to analyze how car age correlates with car prices.
* Statistical Analysis: We conducted statistical tests such as t-tests and ANOVA to assess the significance of differences in car prices based on factors like transmission type and engine fuel. Additionally, a two-way ANOVA model was fitted to examine the combined effect of engine fuel and transmission on car prices.
* Machine Learning Model: We developed a simple linear regression model to predict car prices based on various features. The model was trained, validated using cross-validation, and evaluated using performance metrics such as Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R² score.

**Features:**

* We review the dataset's shape and basic statistics when it's first loaded, handling missing values is part of data cleaning.
* A lot of attention is paid to how certain attributes affect other variables. this is shown by using boxplot.
* For predictive analysis, libraries for machine learning (sklearn) are imported, suggesting Lasso, Ridge, LinearRegression models.
* We evaluate model performance using metrics like mean squared error and R-squared (mean\_squared\_error, r2\_score).
* Using matplotlib.pyplot and seaborn for boxplots indicates a focus on data visualization.
* To find out if there are statistically significant differences between the different car categories, we could do hypothesis testing.
* To determine how a car's popularity or resale value are impacted by particular attributes, use regression analysis.
* Cross-validation (cross\_val\_score) and statistical metrics are used to evaluate linear models (Lasso, Ridge, LinearRegression).
* To validate models, train\_test\_split suggests splitting the dataset into training and testing sets.
* Combining predictive modeling, inferential statistics, and descriptive statistics to present a clear overview of the used car market.
* The used car market may exhibit patterns or segments that can be found using multivariate analysis, time-series analysis, or clustering techniques.

# **RELATED WORK**

**Price Prediction and Valuation:**

* Regression Models: Machine learning algorithms such as regression to forecast used car prices, considering factors like mileage, age, brand, and more are often used by researchers.
* Market Analysis: Price trends, regional variations and seasonal fluctuations can be understood by analysis of historical data.

**Consumer Behavior and Preferences:**

* Feature Importance: Research aimed to identify influential factors (e.g., mileage, brand, color) in buyer decision-making and their relative significance.
* Regional Disparities: Comparative analysis across regions highlighted divergent preferences for specific car types or brands.

**Market Dynamics and Economics:**

* Supply-Demand Evaluation: Utilizing used car datasets contributed to understanding market dynamics, demand and supply and fluctuations.
* Feature Impact: Research explored how various features or conditions impact the resale value of cars.

**Sustainability and Environmental Impact:**

* Vehicle Emissions and Age: Studies leveraged datasets to examine the relationship between a car's age and its emissions, to make environmental policy considerations.

**Business Strategies and Marketing:**

* Management of Inventory: Businesses used these datasets to for optimization of inventory management by identifying fast-selling or high-demand car types.
* Tailored Marketing: Understanding consumer preferences assisted in developing targeted marketing strategies.

**Predictive Maintenance and Risk Assessment:**

* Maintenance Predictions: Analysis of historical data facilitated predictive maintenance models, forecasting maintenance needs and associated costs for different car models at different ages.
* Risk Analysis: Assessing the probability of car issues or repair needs based on previous patterns and data trends.

# **DATASET**

Dataset Title: “Used-cars-catalog”

The used cars catalog dataset consists of car advertisements containing numerous categorical and numerical data attributes. The main aim of this dataset is to explore the market of used cars which can later be used to predict the price of the car.

**Data Characteristics:**

* Feature Variety: This dataset encompasses both categorical and numerical attributes, providing comprehensive details about various aspects of cars in advertisements.
* Categorical Attributes: Details like Manufacturer\_name, Model\_name, Transmission, Color, Engine\_fuel, Engine\_type, Body\_type, State, Drivetrain, Location\_region, among others.
* Numeric Attributes: Include Odometer\_value, Year\_produced, Engine\_capacity, Price\_usd, Number\_of\_photos, Up\_counter, duration\_listed, along with multiple "Feature" attributes.

**Dataset Purpose:**

* Analytical & Predictive Potential: The dataset is well-suited for predictive modeling, aiming to comprehend and forecast factors influencing used car prices or the probability of an exchange.
* Market Analysis: It also serves as a valuable resource for analyzing market trends, exploring regional preferences, and examining aspects influencing the duration cars remain listed for sale.

**Data Granularity:**

* Individual Advertisement Level: The dataset provides specific details about each car advertisement, offering a detailed perspective on numerous attributes.
* Diverse Feature Set: Encompassing a broad spectrum of features, it offers in-depth insights into manufacturer specifics, technical details, pricing, and more.
* Temporal Dimension: Features such as Year\_produced and duration\_listed potentially unveil the age of cars and their average market listing duration, providing temporal insights.

**Data Source:**

The dataset is publicly available on Kaggle, a platform for data science competitions and collaborative projects. It was uploaded by user [lepchenkov], and can be accessed through the following link: [Used Cars Price Analysis for Fair Pricing and Decision Making](https://www.kaggle.com/datasets/lepchenkov/usedcarscatalog)

# **DETAIL DESIGN OF FEATURES**

**Data Cleaning:**

* Initial exploration indicates that the dataset is relatively clean, with only the "engine\_capacity" feature having 10 null values. These missing values are addressed by dropping the respective rows.
* Unnecessary features, such as "feature\_0" to "feature\_9", are removed from the dataset.

**Exploratory Data Analysis (EDA):**

* EDA involves visualizing the data using various plots, such as boxplots and bar graphs, to understand relationships between different features and the target variable (car prices).
* Analysis includes examining price distribution by transmission type, average price versus car color, and the impact of engine type and capacity on prices.

**Feature Engineering:**

* A new feature, "car\_age," is created by calculating the age of each car based on the year it was produced.
* The dataset is then categorized into budget, medium, and high-end cars based on specified price ranges.

**Data Encoding:**

* Categorical variables are one-hot encoded to facilitate machine learning model implementation.

**Statistical Tests:**

* T-tests are conducted to examine if there is a significant difference in prices between automatic and mechanical transmissions.
* ANOVA tests are performed to assess the impact of engine fuel type on car prices.

**Machine Learning Model (Linear Regression):**

* + The dataset is split into training and testing sets.
  + Linear Regression is implemented to predict car prices based on various features.
  + Model performance is evaluated using cross-validation, and validation metrics such as Mean Absolute Error, Mean Squared Error, and R² Score are calculated.

# **ANALYSIS**

In this phase, we dive into the outcomes and interpretations derived from the implemented methods, including data visualizations, statistical tests, and machine learning models.

**Data Visualization:**

Boxplots and Histograms:

* Boxplots were used to visually represent the distribution of car prices. This aids in identifying the central tendency, spread, and potential outliers in the dataset.
* Outliers, if present, could significantly impact the model's performance. Understanding the distribution allows for informed decisions on whether to further preprocess the data or explore advanced modeling techniques.
* Histograms offer insights into the frequency distribution of car prices and ages, aiding in understanding the dataset's overall characteristics.

Scatter Plots:

* Scatter plots were employed to illustrate the distribution of car prices against the age of the vehicles. This visual representation helps identify trends, such as the general decrease in price with increasing age and any potential outliers.
* These plots assist in uncovering nuances in the relationship between age and price, offering additional context to the model's predictive capabilities.

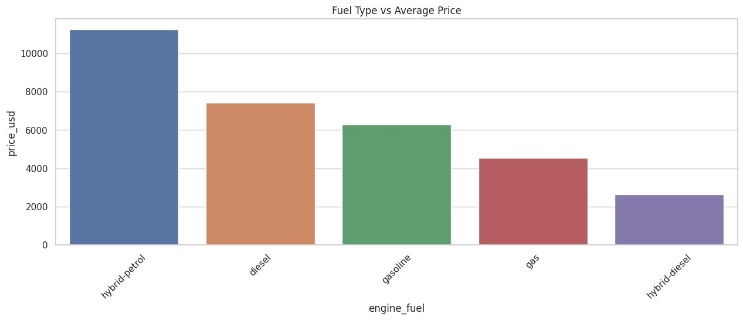
**Statistical Analysis:**

T-Tests - Transmission Type:

* T-tests compare the mean prices between automatic and mechanical transmissions.
* The results guide us in understanding if there is a significant difference in prices based on the car's transmission type.

ANOVA Tests - Engine Fuel Type:

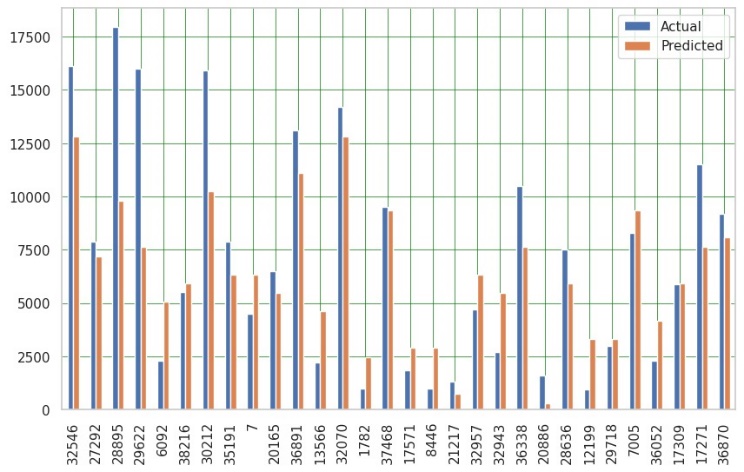
* ANOVA tests explore the impact of engine fuel types on car prices.
* By examining the variation between different fuel categories, we determine if these variations are statistically significant.



**Machine Learning Model - Linear Regression:**

Model Accuracy:

* The Linear Regression model achieved a performance level of approximately 60% accuracy in predicting car prices. This indicates a moderate level of success in capturing the underlying patterns within the dataset.
* Cross-validation scores were used to assess the model's consistency across different subsets of the training data. The average performance across these folds provides confidence in the model's generalization capability.

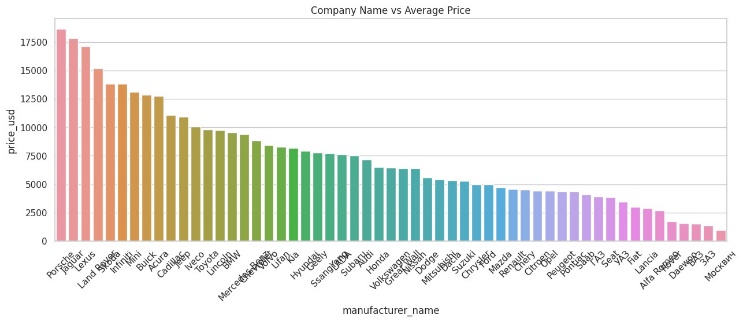


Validation Metrics:

* Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were calculated to quantify the performance of the Linear Regression model.
* These metrics offer insights into the magnitude and direction of errors, providing a clear picture of the model's precision in predicting car prices.

Visualization of Model Results:

* To visually assess the model's predictions, a bar plot comparing actual versus predicted prices for a subset of the dataset was generated. This provides an intuitive understanding of the model's ability to approximate true car prices.
* The visual comparison allows stakeholders to identify trends, patterns, or discrepancies between the model's predictions and the actual prices.



# **IMPLEMENTATION**

The code is implemented in Python using libraries such as Pandas, NumPy, Matplotlib, Seaborn,Scikit-learn and Statsmodels.

**Machine Learning Model (Linear Regression)**

Implementing the Linear Regression model involves several steps, from preparing the data to evaluating the model's performance.

Data Preparation:

* + The dataset is split into two parts: the independent variables (X) and the target variable (y).
  + Features such as "price\_usd," "car\_age," and one-hot encoded categorical variables are included in the independent variables.

Data Splitting:

* The data is divided into training and testing sets using the `train\_test\_split` function from Scikit-learn. This step is crucial to ensure the model's ability to generalize to unseen data.

Model Creation:

* The Linear Regression model is instantiated using the `LinearRegression` class from Scikit-learn.
* The model is then fitted to the training data using the `fit` method, learning the relationships between the independent and target variables.

Model Coefficients and Intercept:

* After training, the coefficients and intercept of the linear equation are accessible through the model's attributes.
* These coefficients represent the weightage assigned to each feature in predicting the target variable.

Cross-Validation:

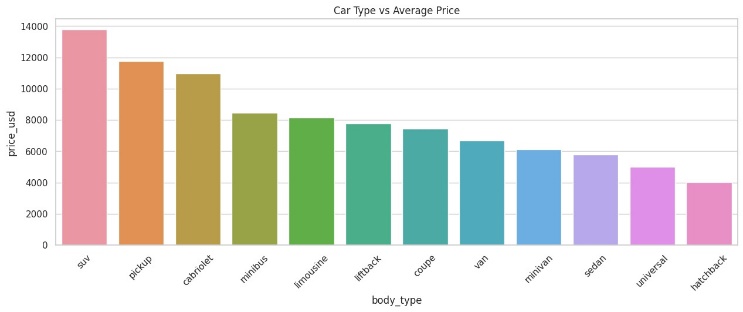
* Cross-validation is performed using the `cross\_val\_score` function to assess the model's performance on different subsets of the training data.
* This step ensures robustness and helps identify potential overfitting or underfitting.

Prediction:

* The trained model is used to predict car prices on the testing set, generating a set of predicted values.

Visualization of Results:

* The actual versus predicted values are visualized using bar plots, providing an intuitive comparison of how well the model captures the pricing patterns.



Model Evaluation:

* Performance metrics such as Mean Absolute Error, Mean Squared Error, and R² Score are calculated to quantify the accuracy of the model.
* These metrics offer insights into how well the model aligns with the true prices.

Extrapolation:

* To showcase the model's usability, a new set of data (representing a car with a specific age) is introduced.
* The model predicts the price for this hypothetical car, demonstrating its applicability beyond the training data.

**Statistical Tests**

* T-tests and ANOVA tests are applied to validate hypotheses and determine the significance of certain variables in influencing car prices.
* The choice of these tests is based on the nature of the variables being compared.

**Additional Insights**

* The code includes visualizations such as boxplots, histograms, and scatter plots to facilitate a better understanding of the dataset's distribution and relationships.

# **PRELIMINARY RESULTS**

**Transmission Type vs. Price:**

T-Test Results:

* We separated the data into two groups based on transmission type: automatic and mechanical.
* A t-test was performed to compare the mean prices between these two groups.
* The obtained t-statistic and p-value were used to assess whether there is a statistically significant difference in car prices based on transmission type.
* The results indicated a significant difference in prices between automatic and mechanical transmissions.

Visualization - Boxplot:

* + To provide a visual representation of the price distribution, we created a boxplot.
  + The boxplot displayed the spread of car prices for both automatic and mechanical transmissions.
  + This visualization helped in understanding the central tendency, spread, and potential outliers in each transmission category.

**Bootstrap Mean and Confidence Interval:**

Bootstrap Procedure:

* + We performed bootstrapping on the 'price\_usd' data, generating multiple samples by resampling with replacement.
  + For each sample, the mean price was calculated.
  + This process was repeated multiple times to create a distribution of bootstrap means.

Confidence Interval:

* + The obtained distribution of bootstrap means allowed us to estimate the mean price more robustly.
  + A 95% confidence interval was then calculated, providing a range within which we can be reasonably confident the true population mean lies.

**Engine Fuel Type and ANOVA:**

ANOVA Test Results:

* + We investigated the impact of engine fuel type on car prices using ANOVA (Analysis of Variance).
  + ANOVA tests whether there are any statistically significant differences in the means of different groups.
  + The F-statistic and p-value from the ANOVA test were used to determine if there are significant differences in car prices based on engine fuel type.
  + The results indicated a statistically significant difference in prices across engine fuel categories.

Two-Way ANOVA Model:

* + We extended the analysis by fitting a two-way ANOVA model considering both engine fuel and transmission type.
  + This model helped assess the combined effects of these two factors on car prices.

**Linear Regression Model:**

Model Training and Coefficients:

* We developed a linear regression model to predict car prices based on various features.
* The coefficients of the model were examined to understand the impact of each feature on the predicted prices.
* The intercept provided the baseline value for car prices.

Cross-Validation:

* + Cross-validation was employed to assess the model's performance on different subsets of the training data.
  + The average cross-validation score gave an indication of the model's generalization ability.

Model Evaluation:

* + The model was evaluated on the test set using standard metrics such as Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R² score.
  + Visualizations - Actual vs. Predicted:
  + To visually inspect the model's predictions, we created a bar plot comparing actual and predicted prices for a subset of the test data.

|  |  |  |  |
| --- | --- | --- | --- |
|  | USAGE OF  LIBRARIES | | SPSS |
| Regression | Coefficient | -432.20 | -560.66 |
| Intercept | 14554.80 | 17873.22 |
| Independent T-test | T-statistic | 106.36 | 106.36 |
| P-value | 0.0 | <0.01 |
| Bootstrap | Mean | 6640.20 | 6476.18 |
| 95% CI [UL]  [LL] | 6703.21 | 6537.46 |
| 6578.46 | 6414.73 |
| 1-way ANOVA | F-statistic | 111.42 | 111.55 |
| P-value | 2.65×10-117 | <0.01 |

\*\*UL – Upper Limit, LL – Lower Limit

**Visualizations:**

Exploratory Visualizations:

* Various visualizations, including boxplots, line plots, and histograms, were generated to explore relationships between car prices and different features.
* For example, age and price were visualized to understand how the price of cars varies with their age.

**Conclusion:**

This project provides valuable insights into the factors influencing car prices. The analysis consists of statistical tests and machine learning models, contributing to a comprehensive understanding of the dataset. The implemented methods showcase a multi-faceted approach to extracting meaningful information from the data. The implementation phase involves a careful and systematic approach to preparing the data, creating, training, and evaluating the Linear Regression model. Visualization aids in interpreting results, and statistical tests provide a rigorous analysis of the dataset. The inclusion of cross-validation ensures the model's robustness, contributing to a comprehensive exploration of car pricing dynamics.

# **PROJECT MANAGEMENT**

**Implementation status report:**

In this current project, the team focused on delivering a fair price to the consumers on the used cars based on the age of the car. To achieve this Exploratory data analysis (EDA), visualization strategies, Machine learning models and tools play a huge role in improving the outcome of the potential insights to solve the problem. Explanatory data analysis, which interprets the current problem into sub-problems through visualization techniques gives the ability to identify intricacies and unseen patterns in the dataset. Finally, the employed techniques in both using Python environment and the SPSS tool, can able to add robustness to the analysis and draw meaningful conclusions from the study. These conclusions can provide effective collaboration and use of data science methodologies.

**Work Completed:**

* + Data gathering, cleaning and preprocessing
    - Description: Finding a dataset that can be used to perform Empirical Analysis. Then cleaning the data set and preparing it for further analysis using preprocessing techniques.
    - Responsibility: Lekhitanand Bandi
    - Contribution: Data gathering, cleaning and preprocessing
  + Exploratory Data Analysis
    - Description: Performing EDA operations to understand the data better.
    - Responsibility: Akshadha Reddy Itikela
    - Contribution: Exploratory Data Analysis
  + Visualization
    - Description: Used appropriate graphs to represent the data, find patterns, draw insights and provide easy understanding.
    - Responsibility: Tharun Kuravadi Sathish Babu
    - Contribution: Performing visualization techniques.
  + Linear Regression using Python and SPSS
    - Description: Build a linear regression model that can predict the price of a used car based on the age.
    - Responsibility: Sai Vignesh Rayal Rolla
    - Contribution: Building and validating the model using python and SPSS
  + One-way Anova using Python
    - Description: Using python in built libraries to perform one-way anova
    - Responsibility: Vikram Kumar Reddy Ayaluri
    - Contribution: Implementing one-way Anova using Python
  + Implementation of independent t-test
    - Description: Performing Independent t-test on several categorical features to find out if there is any significant difference between the data.
    - Responsibility: Vikram Kumar Reddy Ayaluri
    - Contribution: Taking independent t-tests and analyzing the results to determine if there's a difference.
  + Implementation of One-way Anova in SPSS
    - Description: Perform One-way Anova on certain types of engines in SPSS to check if there are any Significant differences among the engine types.
    - Responsibility: Lekhitanand Bandi, Akshadha Reddy Itikela
    - Contribution: Conducting One-way Anova analysis in SPSS after data gathering, cleaning, and preprocessing.
  + Bootstrapping using python and SPSS
    - Description: Performing bootstrapping using SPSS
    - Responsibility: Tharun Kuravadi Sathish Babu, Sai Vignesh Rayal Rolla
    - Contribution: Data analysis using python and SPSS bootstrapping and inferences from the results.

# **REFERENCES**

1. Fathalla, A.; Salah, A.; Li, K.; Li, K.; Francesco, P. Deep end-to-end learning for price prediction of second-hand items. Knowl. Inf. Syst., vol. 62, pp. 4541–4568, 2020.Available [[CrossRef](http://www.cs.newpaltz.edu/~lik/publications/Ahmed-Fathalla-KIS-2020.pdf)]
2. Available <https://www.kbb.com/car-advice/is-now-the-time-to-buy-sell-or-trade-in-a-used-car/>
3. Lucija Bukvić et al. “Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning”. Faculty of Transport and Traffic Science, University of Zagreb, 10000 Zagreb, Croatia,vol 14,issue 24,pp. 17034, Dec-2022. Available [[ref](https://www.mdpi.com/2071-1050/14/24/17034)].
4. Kanwal Noor.; Sadaqat Jan, “Vehicle Price Prediction System using ML Techniques”,CSE dept, Vol. 167,pp .9, June 2017. available [[ref](https://www.ijcaonline.org/archives/volume167/number9/noor-2017-ijca-914373.pdf)]