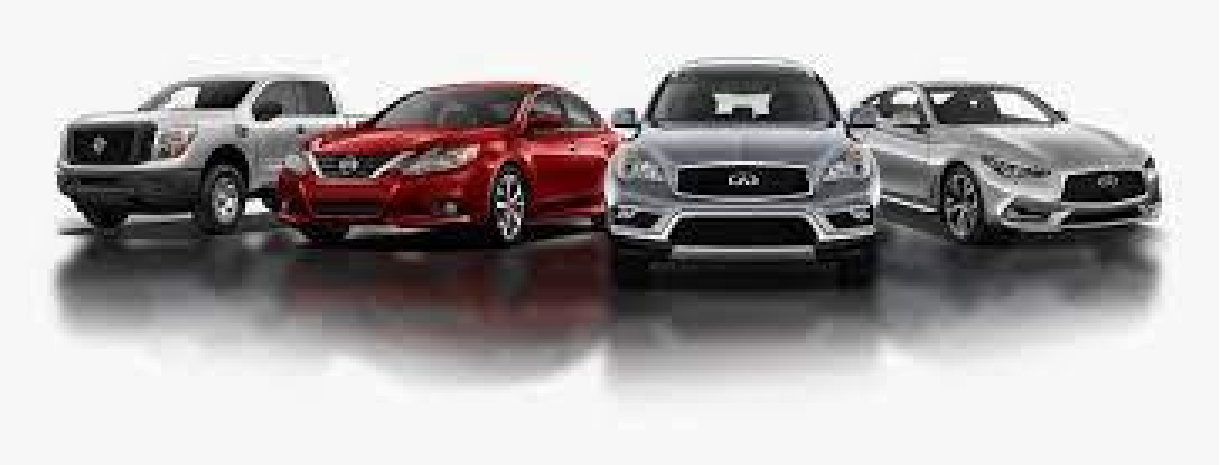
Project Proposal for Used Car Price Prediction Dataset

By: Chase Mueller, Gina Butler, Neelam Prasad, Kriti Khatri, Thai Champagne



Dataset Overview:

The Used Car Price Prediction Dataset is a comprehensive collection of automotive data extracted from the popular marketplace website Cars.com. The dataset consists of 4,009 data points, with each entry representing a unique vehicle listing. It includes nine features providing in-depth insights into the automotive market, which is particularly useful for analyzing trends, predicting vehicle prices, and studying consumer preferences.

Dataset Link: Kaggle Dataset Link

The features in the dataset are as follows:

- **Brand & Model**: Information about the vehicle's brand and specific model.
- **Model Year**: Year of manufacture, essential for understanding depreciation and technological advancements.
- **Mileage**: The number of miles the vehicle has been driven, a critical indicator of its wear and tear.
- **Fuel Type**: The fuel type the vehicle uses, such as gasoline, diesel, electric, or hybrid.
- **Engine Type**: Specifications related to the vehicle's engine, impacting performance and efficiency.
- **Transmission**: Whether the vehicle is automatic, manual, or another type of transmission.
- **Exterior & Interior Colors**: Aesthetic features, which may influence the vehicle's appeal.
- **Accident History**: Whether the vehicle has a history of accidents, important for assessing its condition and resale value.
- **Clean Title**: Indicates if the car has a clean title, which affects its legal status and marketability.

- **Price**: The listing price of the vehicle, which is the target variable we aim to predict.

This dataset is highly valuable for automotive enthusiasts, potential buyers, and researchers in the automotive industry. It provides rich data to explore pricing trends, make informed purchasing decisions, and gain insights into consumer preferences in the used car market.

Why We Chose This Dataset:

Our group chose this dataset due to our shared interest in automotive trends and the potential to create practical predictive models. Many of us are passionate about cars, and this dataset allows us to delve into real-world data that influences purchasing decisions. Additionally, by leveraging machine learning techniques and creating interactive visualizations in Tableau, we can provide valuable insights that could help buyers make smarter decisions and sellers optimize their pricing strategies.

Research Questions for Used Car Price Prediction:

1. **What factors most significantly affect the price of a used car?**
    - **Goal**: Analyze how features like brand, model year, mileage, and fuel type influence car prices.
    - **Visualization**: Correlation heatmap, scatter plots (e.g., Price vs. Mileage, Year).
2. **Can we predict the price of a used car based on its attributes?**
    - **Goal**: Build a predictive model using car features (brand, mileage, engine type, accident history) to estimate prices.
    - **Visualization**: Regression line chart, actual vs. predicted price scatter plot.
3. **How does the model year of a car affect its price?**
    - **Goal**: Explore if newer cars are priced higher and how car age impacts price.
    - **Visualization**: Line chart or bar chart comparing average prices by model year.
4. **Do accident history and clean title status impact the price of a used car?**
    - **Goal**: Investigate how accident history and clean title status influence car prices.
    - **Visualization**: Box plot or bar chart comparing prices based on accident history and clean title status.
5. **Which fuel type (gasoline, diesel, electric, hybrid) has the highest average price?**
    - **Goal**: Determine how fuel type affects the price of used cars.

- **Visualization**: Bar chart comparing average prices by fuel type.

6. **How does engine type influence the price of used cars?**
    - **Goal**: Analyze whether certain engine types (e.g., V6, V8, electric) lead to higher prices.
    - **Visualization**: Box plot or bar chart comparing prices based on engine type.

Inspiration & References:

1. **Kaggle Notebooks**:
    - We will explore Kaggle Notebooks on used car price prediction to guide our preprocessing and model building. Relevant notebooks include:
        - "Used Car Price Prediction" – Focuses on cleaning data, feature engineering, and regression models.
        - "Predicting Car Prices" – Demonstrates machine learning models (like Random Forest and XGBoost) for price prediction.
        - **Key takeaways**: Data cleaning, encoding categorical variables, and model comparison.
2. **Tableau Public**:
    - We will draw inspiration from Tableau Public visualizations of automotive data to design our dashboards. Examples include:
        - Price vs. Mileage: Scatter plots or bar charts to analyze price trends based on mileage.
        - Feature Importance: Visuals showing which features most impact car prices.
    - These visuals will help us create intuitive and clear dashboards.
3. **Kaggle Code Tab**:
    - The "Code" tab in the Kaggle dataset will be explored for efficient data preprocessing techniques, including:
        - Handling missing data and encoding categorical features.
        - Applying machine learning models like Linear Regression and Random Forest.
        - Model evaluation metrics such as RMSE and R-squared.

- These references will help us streamline data processing, feature engineering, and visualization, ensuring we follow best practices and build effective predictive models.

Tableau Dashboards/Stories:

We will develop two Tableau Dashboards to showcase our analysis:

1. **Dashboard 1: Overview of Car Prices**
    - **Price Distribution by Brand**: Bar chart or box plot showing average car prices by brand.
    - **Price vs. Mileage**: Scatter plot to explore how mileage affects price.
    - **Price vs. Accident History**: Box plot comparing prices based on accident history.
    - **Price vs. Model Year**: Bar or line chart showing price trends by model year.
    - **Price vs. Fuel Type**: Bar chart comparing average prices by fuel type.
    - **Interactive Filters**: Filters for brand, mileage, engine type, etc., to allow user exploration.
2. **Dashboard 2: Predictive Model Dashboard**
    - **Actual vs. Predicted Prices**: Scatter plot comparing actual and predicted prices.
    - **Model Performance Metrics**: KPIs showing R-squared, MAE, RMSE.
    - **Feature Importance**: Bar chart showing which features most affect the price.
    - **Price Prediction by Segment**: Bar charts displaying predicted prices for different segments.
    - **Error Analysis**: Heatmap or scatter plot showing model residuals by factors like brand and mileage.

Prediction/Recommendation:

We aim to predict the price of a used car based on its features (e.g., brand, mileage, fuel type, engine type). The prediction will be made using a regression model (such as linear regression or random forest regression), as the target variable (price) is continuous. Additionally, we may experiment with a recommendation system to suggest similar cars based on factors like price range, model, and brand using techniques such as KNN (K-Nearest Neighbors).

- **Target Column**: Price (the variable we are trying to predict).

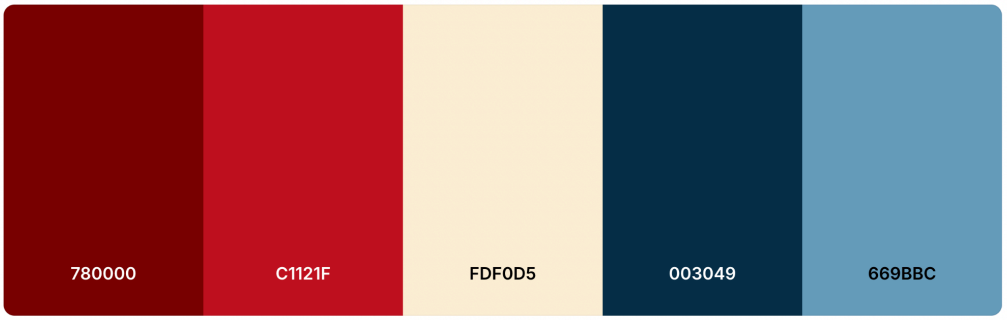
Methodology:

- **Regression vs Classification vs KNN**: The problem we are solving is a regression problem since the target variable (price) is continuous. We will evaluate several regression models (e.g., linear regression, random forest, gradient boosting) to determine the best performer. We may also explore KNN for building a recommendation system.

- **Data Preprocessing**: We will clean and preprocess the data by handling missing values, encoding categorical variables, and normalizing numerical features where necessary.

- **Machine Learning Models**: The team will experiment with different regression models, and we will evaluate them using metrics like Mean Squared Error (MSE) or R-squared to assess model performance.

Color Palette:



Roles and Responsibilities:

1. Chase Mueller: Tableau, report
2. Gina Butler: Tableau, slide
3. Neelam Prasad: flask app, Machine learning
4. Kriti Khatri: Machine learning, proposal,
5. Champagne Thai: Machine learning,

Deliverables:

- **PowerPoint Presentation**: A summary of the project, including key insights, visualizations, and model results.

- **Google Colab**: Jupyter notebooks hosted on Google Colab for data analysis, model building, and testing.

- **PythonAnywhere**: Code hosted and executed on PythonAnywhere for real-time access to the project environment.

- **Project PDF Report**: A professional document containing the project proposal, initial analysis, and visualizations.

- **GitHub Repository**: All project code, documentation, and version control will be uploaded to GitHub for collaboration and sharing. A comprehensive README file in the GitHub repository explaining the project.

This project will provide meaningful insights into the used car market by applying data analysis, machine learning models, and interactive visualizations. We are excited to collaborate on this project and deliver a valuable resource to those interested in the automotive industry.
