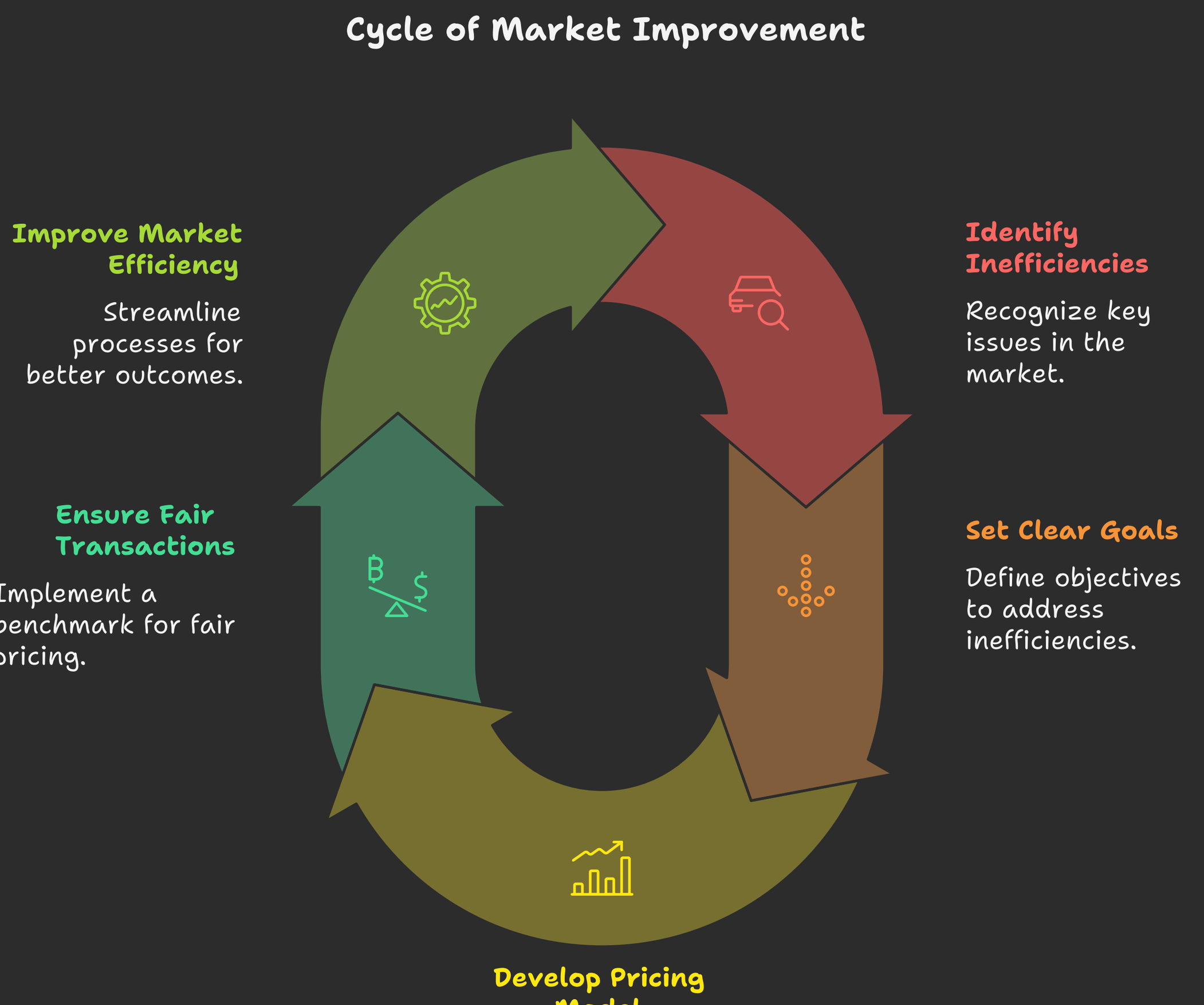
Used Car price prediction

Here is a comprehensive overview of the entire workflow for the used car price prediction project.

1. Business Problems and Goals

The project begins by identifying key inefficiencies in the used car market and setting clear goals to address them.

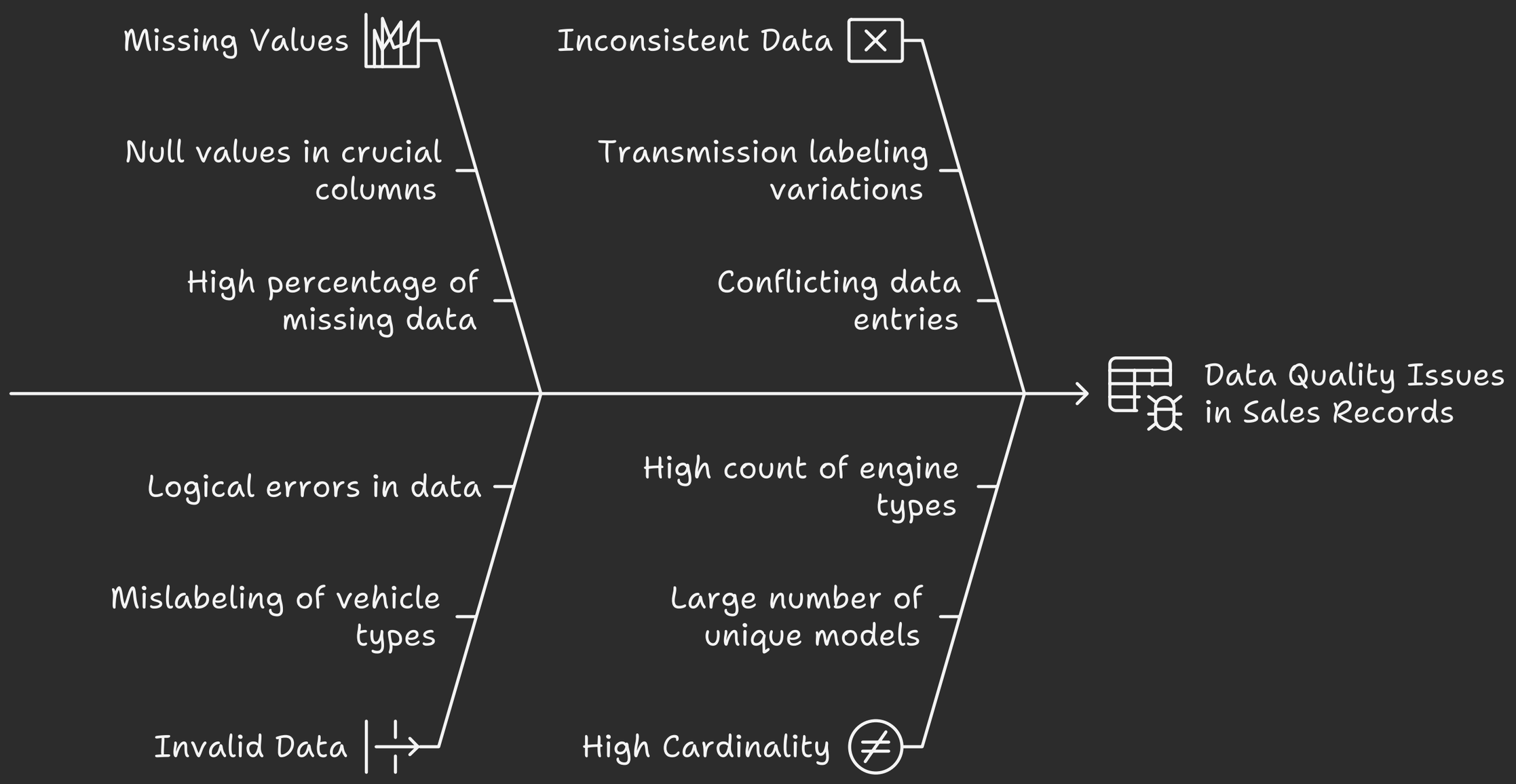
- **Business Problems Identified**:
    - **Information Gap for Buyers**: Buyers often lack sufficient data to determine if a car's price is fair.
    - **Suboptimal Returns for Sellers**: Sellers struggle to price their vehicles accurately to maximize their returns.
    - **Market Inefficiency**: The absence of a dependable pricing standard leads to unfair transactions and overall market inefficiency.
- **Project Goals**:
    - **Accurate Price Prediction**: Develop a machine learning model to precisely predict used car prices.
    - **Ensure Fair Transactions**: Establish a trusted, data-driven pricing benchmark to promote fairness for both buyers and sellers.
    - **Improve Market Efficiency**: Streamline the buying and selling process with reliable insights.



2. Data Collection and Understanding

The next phase involved sourcing and understanding the dataset.

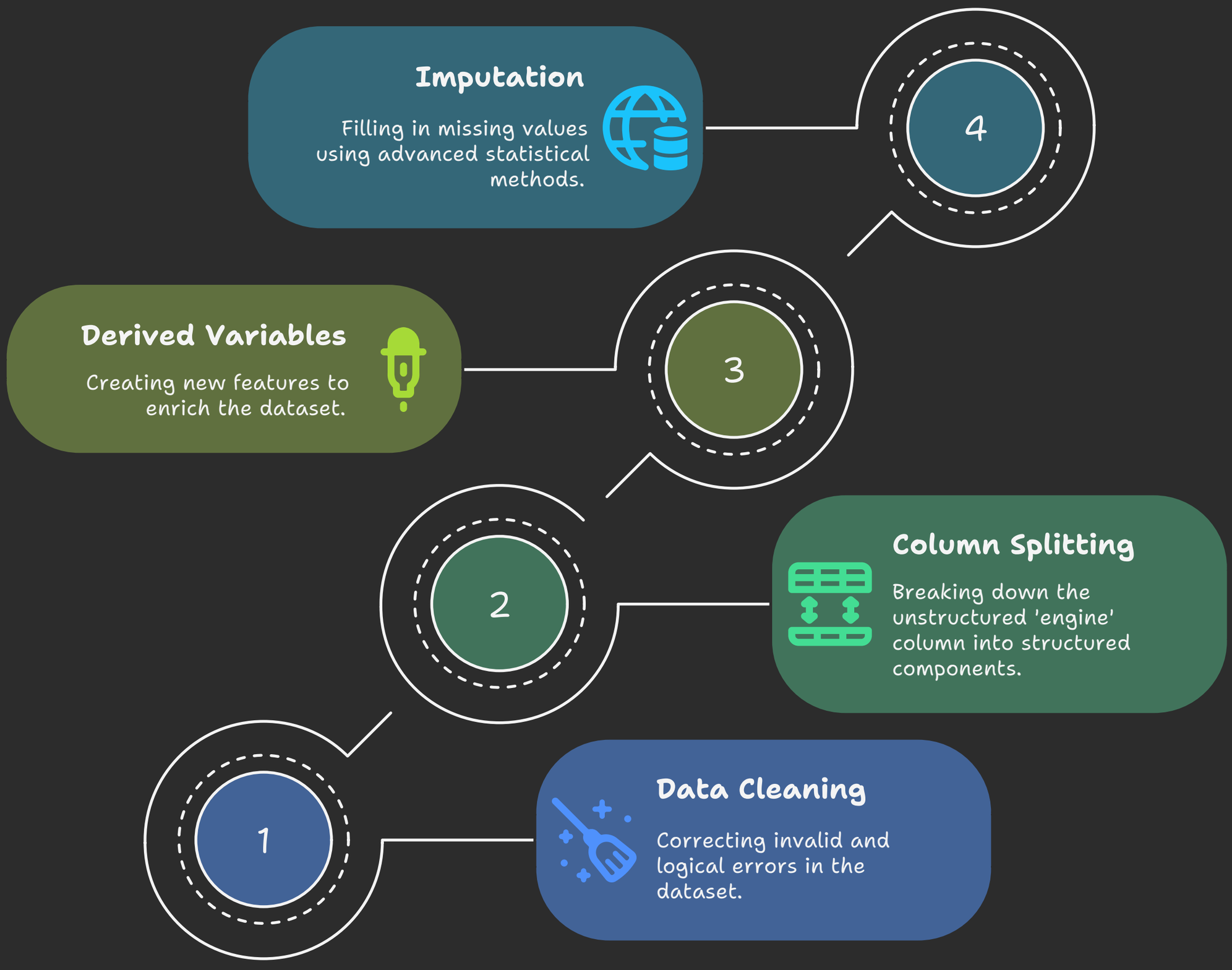
- **Data Source**: The dataset, containing 188,533 sales records with 13 columns, was obtained from a Kaggle competition.
- **Key Features**: Important features included brand, model, model year, mileage, fuel type, engine details, transmission, color, accident history, and clean title status.
- **Initial Data Issues**:
    - **Missing Values**: The dataset had null values in crucial columns like clean title (11.36%), fuel type (2.70%), and accident (1.30%).
    - **Invalid and Inconsistent Data**: Logical errors were found, such as 369 Tesla vehicles being mislabeled as "Gasoline" or "Diesel". Additionally, some data was inconsistent, with transmission being labeled as both “A/T” and “Automatic”.
    - **High Cardinality**: Categorical features like model (1,897 unique classes) and engine (1,117 unique classes) had a very high number of distinct values.



3. Data Cleaning and Feature Engineering

This stage focused on correcting data issues and creating new, more informative features.

- **Data Cleaning**:
    - Invalid values, such as encoding errors in the fuel type column, were corrected.
    - Logical errors, like the incorrect fuel types for Tesla vehicles, were fixed.
    - The CLEAN_TITLE column was dropped entirely because it had a high percentage of missing data and only one non-null value ("Yes"), making it unusable.
- **Feature Engineering**:
    - **Column Splitting**: The unstructured engine column was split into several new, structured columns: HP, Litres, Cylinder, and engine fuel type using Open Refine.
- **Derived Variables**: New features like fuel type new and color category were created to enhance the dataset.
- **Imputation**: **Multiple Imputation by Chained Equations (MICE)** was used to fill in missing values for features like Cylinder, Litres, and HP. This method was chosen because it preserves data relationships better than simpler techniques like mean or median imputation.

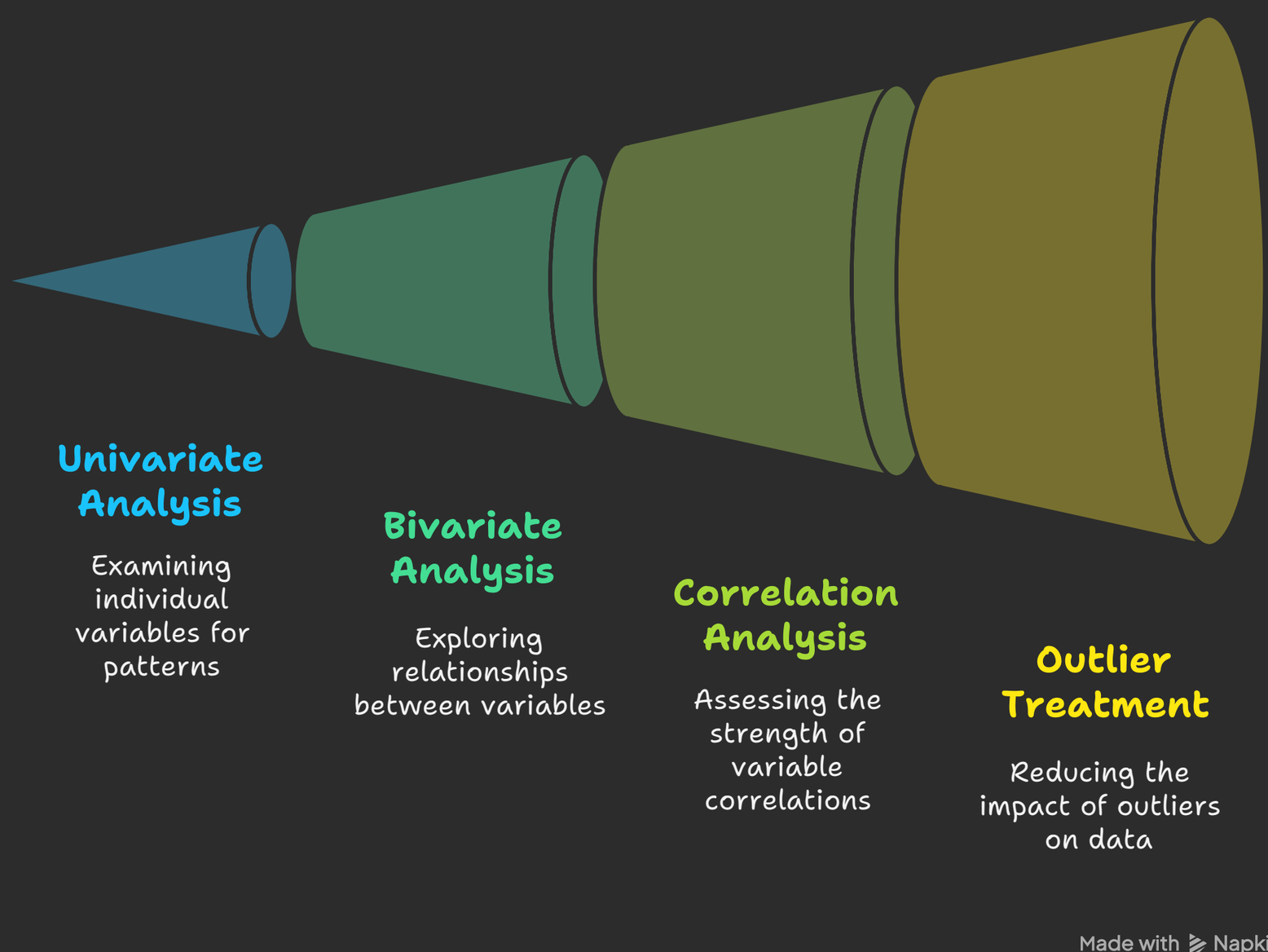


4. Exploratory Data Analysis (EDA) and Visualization

EDA was performed to uncover patterns and relationships within the data.

- **Univariate Analysis**:
    - **Continuous Variables**: Analysis showed that mileage was right-skewed, horsepower (HP) was concentrated in the mid-range (200–400 HP), and popular engine sizes peaked around 2.0, 3.0–3.5, and 5.0 litres.
    - **Discrete Variables**: Ford, Mercedes, and BMW were the most frequent brands. Most vehicles were from recent model years (2018-2022), had automatic transmissions, and were accident-free.
- **Bivariate Analysis**:
    - Higher horsepower, more cylinders, larger engine sizes, and newer model years were all associated with higher prices.
    - **Porsche** was the brand with the highest average price, while electric vehicles commanded the highest prices among fuel types.
- **Correlation and Outlier Analysis**:
    - A strong negative correlation (-0.67) was found between Model Year and Mileage.

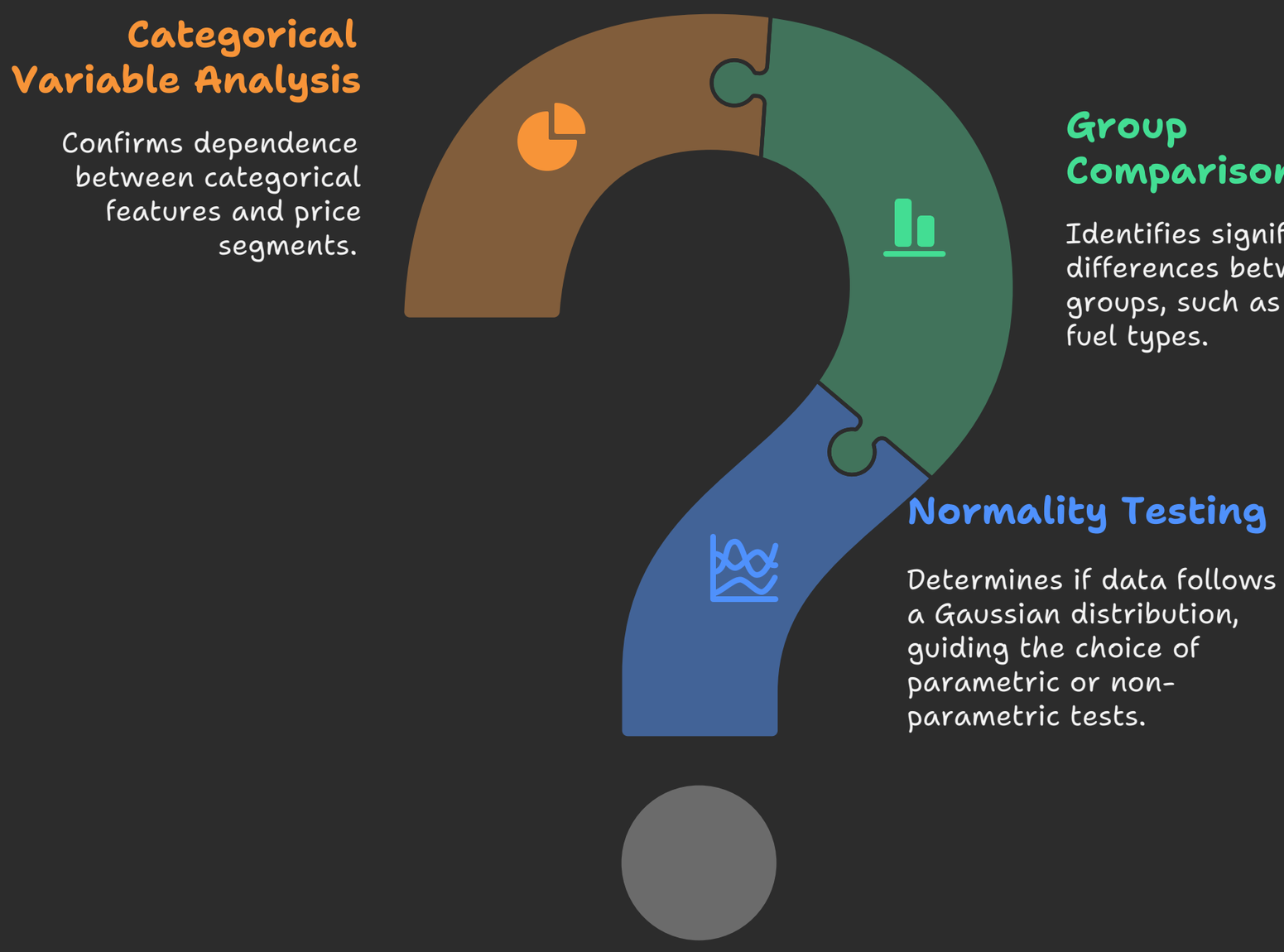
treated using **Winsorization** (capping) to reduce their impact on the model.



5. Statistical Analysis

Formal statistical tests were conducted to validate the observations from EDA.

- **Normality Testing**: The D'Agostino's K^2 Test revealed that all numerical variables had non-Gaussian distributions, indicating that non-parametric tests would be necessary.
- **Group Comparison**:
    - The **Kruskal-Wallis Test** confirmed that there were statistically significant differences in price medians across different categories like brand and fuel type.
    - **Dunn's Post Hoc Test** was then used to identify which specific brand pairs differed significantly. For example, it showed that luxury brands like Porsche and BMW differed sharply from mainstream brands like Ford and Chevrolet.
- **Categorical Variable Analysis**: The **Chi-Square Test** was used to confirm a significant dependence between categorical features (e.g., brand, transmission) and binned price categories (Low, Medium, High). This provided insights into how brand listings are distributed across different price segments.

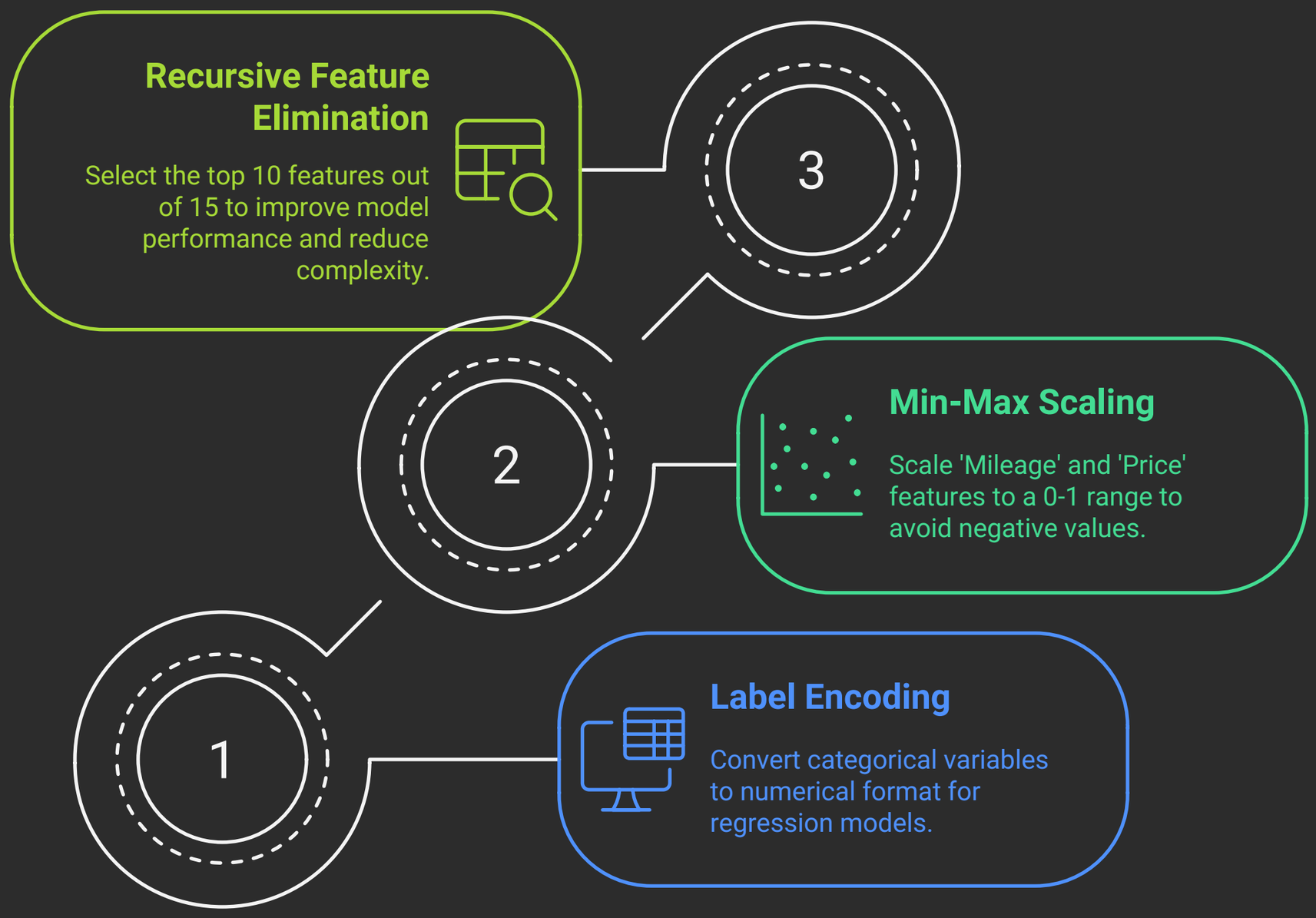


6. Data Preprocessing and Feature Selection

The data was prepared for modeling.

- **Transformation and Scaling**:
    - **Label encoding** was applied to convert categorical variables into a numerical format suitable for regression models.
    - **Min-Max Scaling** was applied to the Mileage and Price features to scale their values between 0 and 1, as other scalers produced unsuitable negative values.

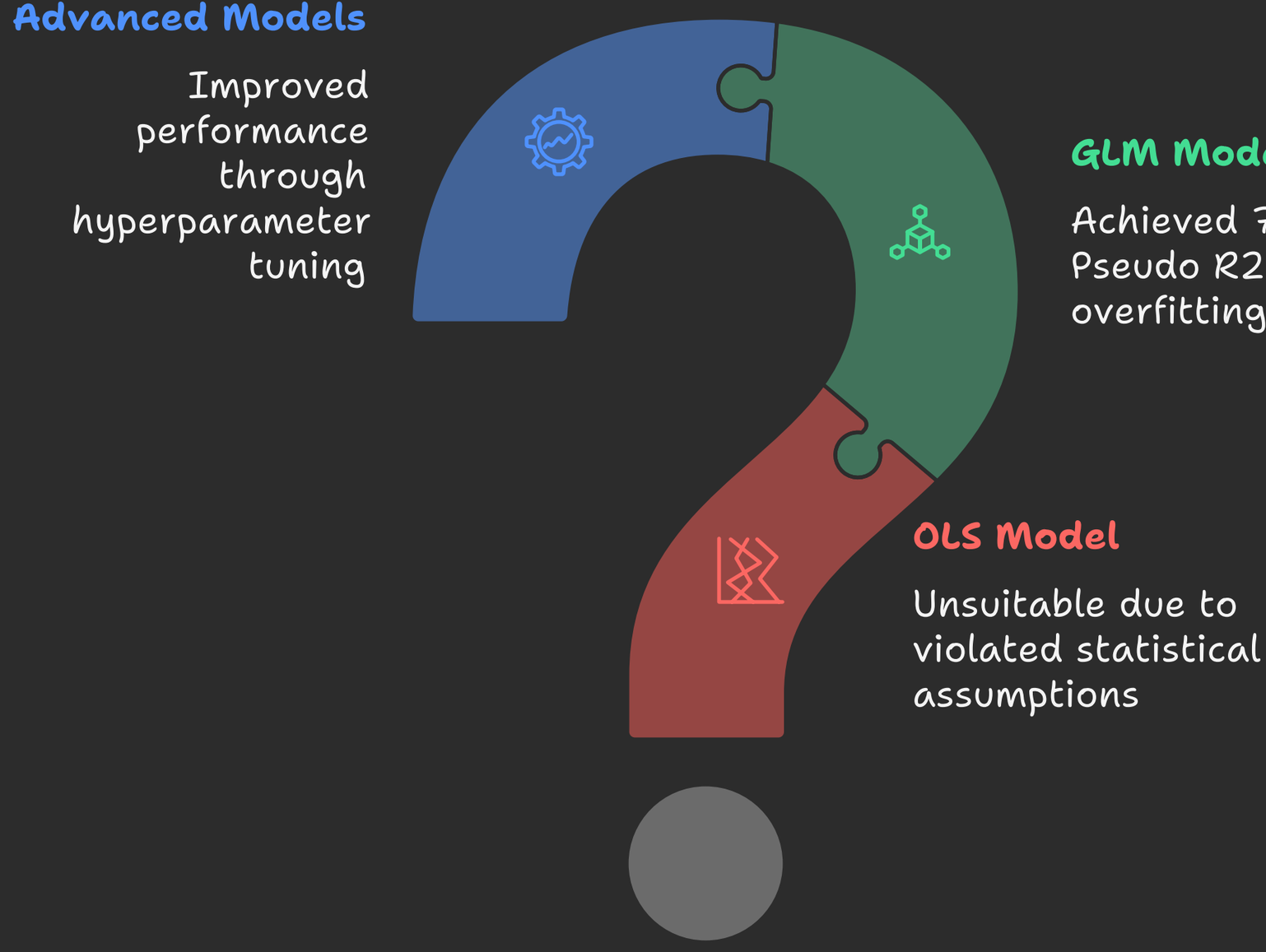
- **Feature Selection**:
    - **Recursive Feature Elimination (RFE)** was used to select the most impactful features. The top 10 out of 15 features were chosen to improve model performance, reduce complexity, and prevent overfitting.



7. Model Building and Evaluation

Multiple regression models were built, tuned, and evaluated to find the best performer.

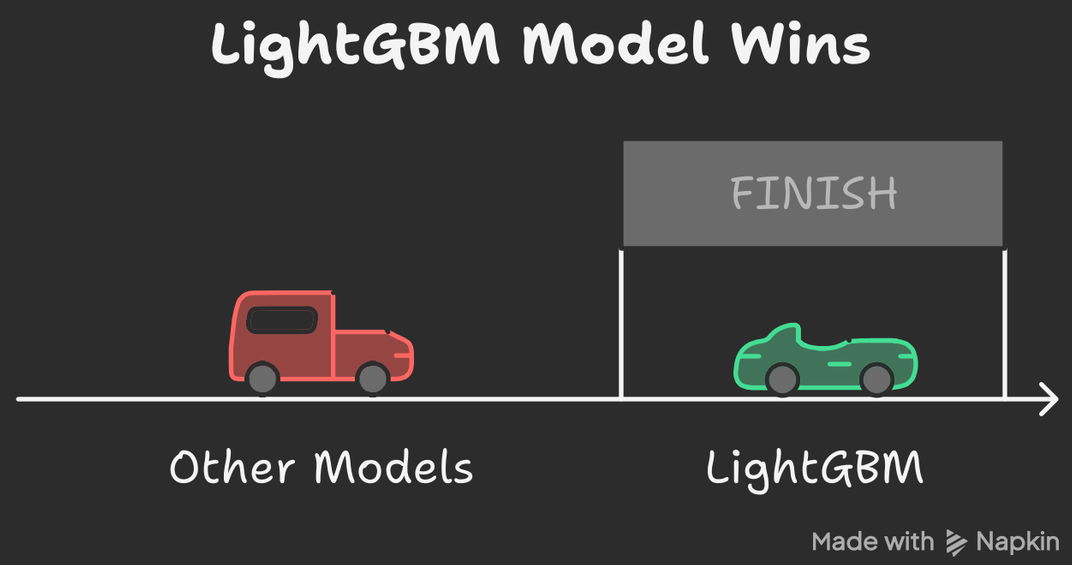
- **Model Selection**: A variety of models were tested, including GLM, Decision Tree, Random Forest, KNN, XGBoost, CatBoost, and LightGBM.
- **Baseline Model**: An initial Ordinary Least Squares (OLS) model was found unsuitable because it violated key statistical assumptions like linearity and normality of residuals. A **Generalized Linear Model (GLM)** was then established as a better baseline, achieving a Pseudo R^2 of 70% with no signs of overfitting.
- **Evaluation and Tuning**:
    - Models were evaluated using **10-fold cross-validation** to ensure robustness.
    - Performance was measured by **RMSE** (Root Mean Squared Error) and **Explained Variance**.
- Hyperparameter tuning using **Bayesian optimization** was applied to advanced models, which significantly improved their performance.



8. Final Model Comparison and Conclusion

The final step involved comparing the tuned models to select the best one.

- **Best Performing Model**: **LightGBM** emerged as the top performer, achieving the highest explained variance (0.6867) and the lowest RMSE (0.1110). It was chosen for its efficiency, accuracy, and ability to handle large datasets and complex interactions.
- **Project Conclusion**: The project successfully developed a reliable price prediction model by using robust preprocessing, feature selection, and advanced machine learning techniques. This outcome enhances market transparency and provides a valuable tool for stakeholders.



9. Future Scope

The project concluded by outlining potential future improvements.

- **Data and Model Enhancement**: Future work could involve incorporating newer datasets and exploring more advanced techniques like deep learning or stacking models.
- **Real-World Implementation**: The team proposed collaborating with dealerships and automotive platforms (e.g., CarDekho, Cars24) to pilot the model for real-world pricing and valuation. This would involve creating a feedback loop to continuously refine the model based on performance metrics and customer trust.