USED CAR PRICE PREDICATION

GROUP - 4

Mentor:

Mr. Subramanian P V

Group Members:

Mr. Anubhav V

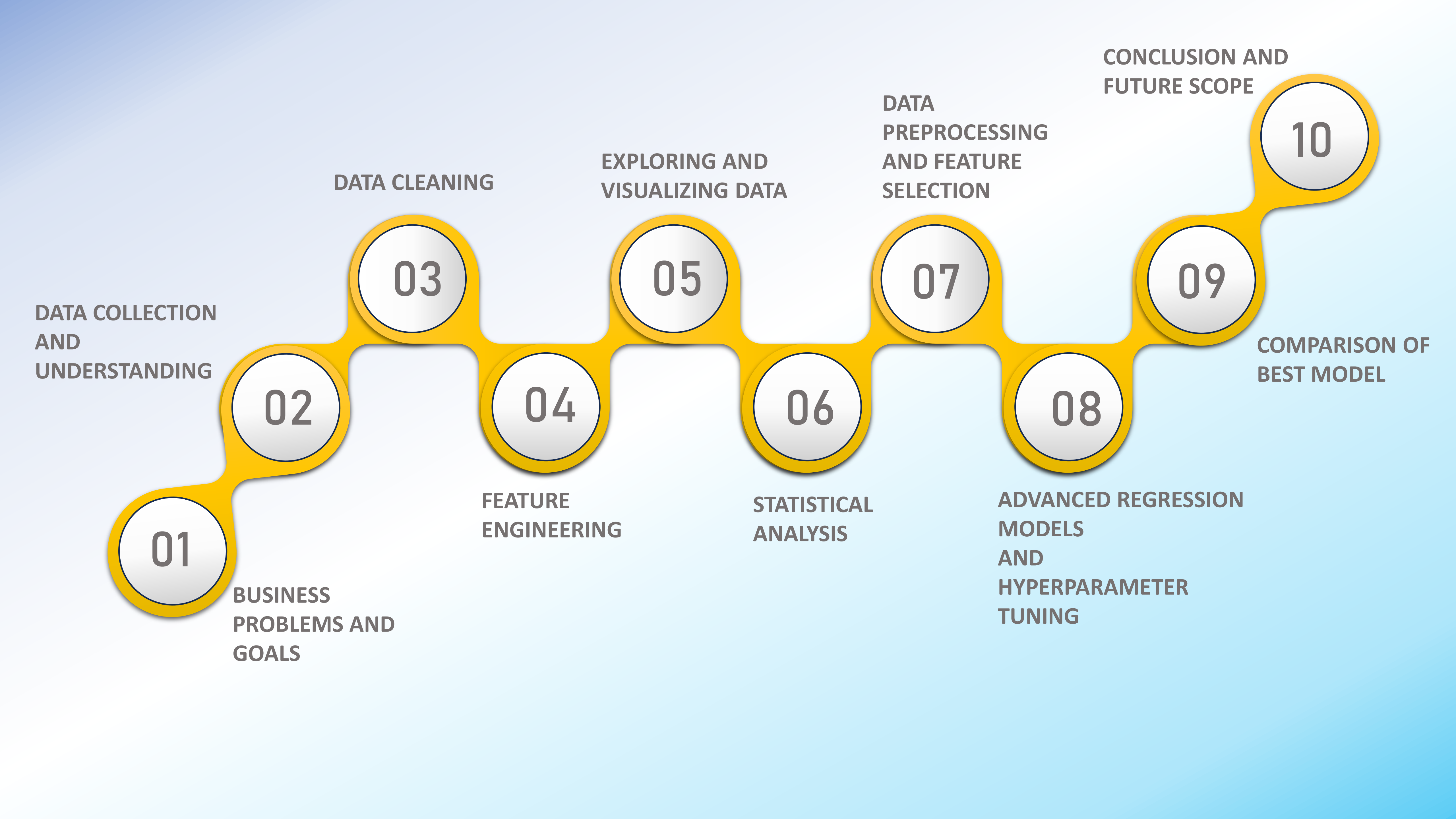
Ms. Agnes Louis

Mr. Afrith H

Mr. Sanjith B

Ms. Pooja P

Mr. Gajarajan V Y


01
BUSINESS PROBLEMS AND GOALS
02
DATA COLLECTION AND UNDERSTANDING
03
DATA CLEANING
04
FEATURE ENGINEERING
05
EXPLORING AND VISUALIZING DATA
06
STATISTICAL ANALYSIS
07
DATA PREPROCESSING AND FEATURE SELECTION
08
ADVANCED REGRESSION MODELS AND HYPERPARAMETER TUNING
09
COMPARISON OF BEST MODEL
10
CONCLUSION AND FUTURE SCOPE

BUSINESS PROBLEMS AND GOALS

PROBLEMS

- **Information Gap for Buyers:** Buyers often lack the data to make truly informed decisions on whether a car's price is fair.

- **Suboptimal Returns for Sellers:** Sellers may not be able to accurately price their vehicles to optimize returns.

- **Market Inefficiency:** The lack of a reliable pricing standard leads to unfair transactions and overall inefficiency in the used car market.

BUSINESS PROBLEMS AND GOALS

GOALS

- **Accurate Price Prediction:** Develop a machine learning model to accurately predict the prices of used cars.

- **Ensure Fair Transactions:** Create a trusted benchmark for pricing that promotes fairness for both parties.

- **Improve Market Efficiency:** Streamline the buying and selling process with reliable, data-driven insights.

DATA COLLECTION AND UNDERSTANDING

DATA COLLECTION

- Source:https://www.kaggle.com/competitions/playground-series-s4e9
- Author: Srinivasa Rao Bittla
- Published: 2024

DATA COLLECTION AND UNDERSTANDING

DATA UNDERSTANDING

- **Dataset Purpose & Size:** The dataset contains 188,533 sales records with 13 columns, intended for predicting used car prices.
- **Key Features:** Includes brand, model, model year, milage, fuel type, engine, transmission, color, accident, and clean title.
- **Missing Values:** The dataset contains null values in key columns:
  - clean title: 11.36%
  - fuel type: 2.70%
  - accident: 1.30%
- **Invalid Values:**
  - **Encoding Errors:** In 781 rows, fuel type has a "-" character represented as "â\x80\x93", which needs to be treated as null.
  - **Logical Errors:** 369 Tesla vehicles are incorrectly labeled as "Gasoline" or "Diesel" and must be corrected.

DATA COLLECTION AND UNDERSTANDING

DATA UNDERSTANDING

- **Inconsistent Data:**
  - Labels for the same category are inconsistent. For example, transmission is labeled as both “A/T” and “Automatic”.
- **Data Integrity:**
  - There are no duplicate rows or zero values in the dataset.
- **Column Types:**
  - **Numerical (4):** id, model year, milage, price
  - **Categorical (9):** brand, model, fuel type, engine, transmission, ext col, int col, accident, clean title
- **Feature Cardinality:** Several categorical columns have a very high number of unique classes, such as model (1,897) and engine (1,117).
- **Problematic Column (clean title):** This feature has only one non-null value ("Yes") and a high percentage of missing data (11.36%), making imputation impossible.

DATA CLEANING

CLEANING

- **Handling Invalid Values:**
  - **Encoding Errors Corrected:** The "-" character (appearing as "â\x80\x93") in the FUEL TYPE column was successfully converted to null.
  - **Logical Errors Fixed:** The incorrect "Gasoline" and "Diesel" labels for Tesla vehicles were corrected to "Electric".

- **Column Removal:**
  - The problematic CLEAN_TITLE column was dropped from the dataset due to its high percentage of missing data and lack of useful variance.

FEATURE ENGINEERING

SPLITTING COLUMNS

- **Primary Technique Used: Column Splitting**
  - The single engine column, which contained multiple pieces of information, was split into several new, distinct columns.
  - This makes the individual pieces of information more accessible and useful for the model.
- **Tool and Process**
  - Open Refine was used to perform the column split.
  - The process involved using the "Split into several columns" function with "HP" as the separator.
- **Outcome of Splitting**
  - Successfully extracted and created new columns for HP, Litres, Cylinder, and engine fuel type.
  - This transformed unstructured engine details into structured, usable features.
- **Additional Technique Used: Derived Variables**
  - Created new variables like fuel type new and color category to further enhance the dataset.

FEATURE ENGINEERING

MICE IMPUTATION

•**MICE Implementation**: Applied Multiple Imputation by Chained Equations

•Why: Superior to mean/median imputation as it uses **similar row patterns**

•**Features Used**: Cylinder, Litres , HP, Brand, Model.

•**Advantage**: Preserves data relationships and reduces bias

FEATURE ENGINEERING

REDUCE LEVEL

- To ensure clearer graph representation, we reduced the levels of categorical variables to only the top 10 values.

- This technique was applied to the following features:

1. model year
2. model,
3. brand,
4. Exterior color
5. interior color
6. transmission

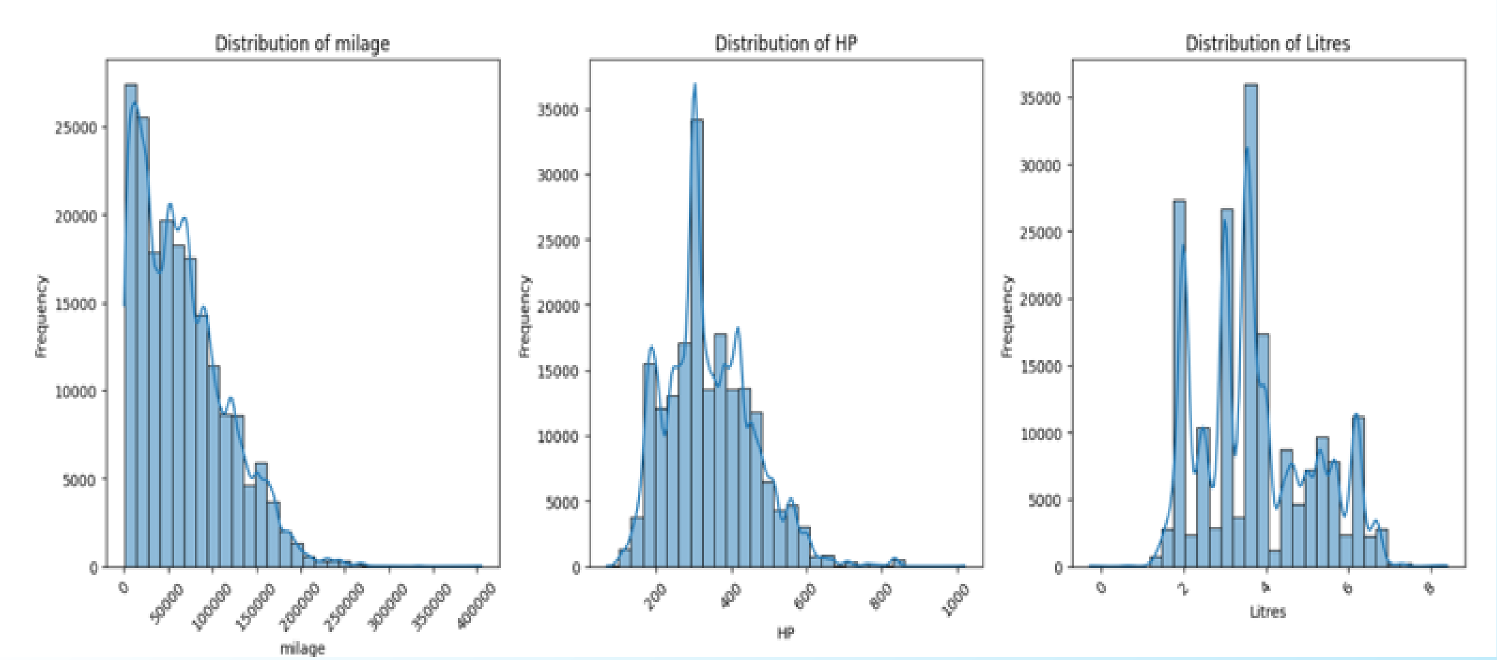
EXPLORING AND VISUALIZING DATA

Univariate Analysis - Continuous Variables

- **Mileage**: Distribution is generally right-skewed, with most vehicles having lower mileage, though some have higher values.
- **Horsepower (HP)**: Primarily concentrated in the mid-range (200–400 HP), with fewer vehicles exhibiting high performance.
- **Litres**: Multiple peaks suggest popular engine sizes around 2.0, 3.0–3.5, and 5.0 litres.

EXPLORING AND VISUALIZING DATA

Univariate Analysis - Continuous Variables



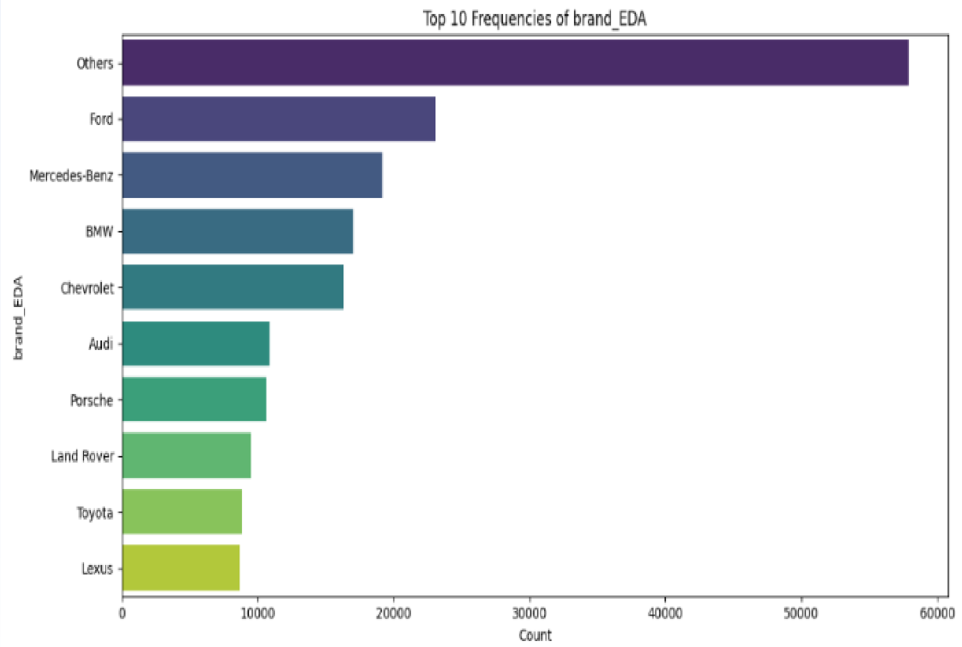
EXPLORING AND VISUALIZING DATA

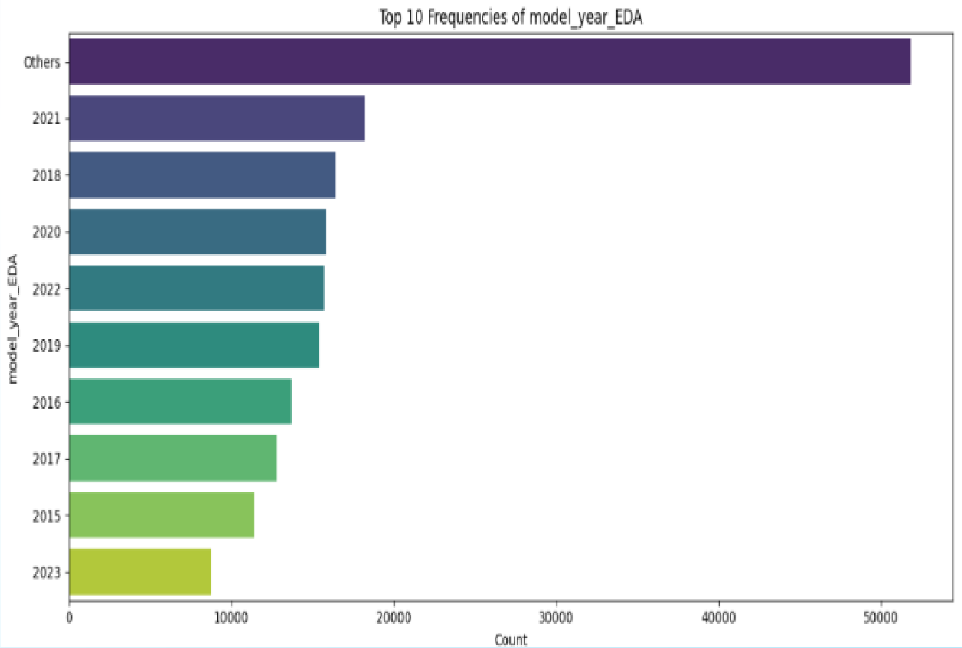
Univariate Analysis - Discrete Variables

- **Brand**: Ford, Mercedes, BMW, Chevrolet frequent; many "Others.“
- **Model Year**: Recent years (2021, 2018, 2020, 2022) common; many "Others.“
- **Color**: "Luxury" leads, then "Premium," "Standard."
- **Cylinder Layout**: V-engines common; many unknown.
- **Model**: Wide variety, mostly "Others.“
- **Accident**: Most vehicles accident-free.
- **Fuel Type**: Gasoline dominant; hybrids, electric less common.
- **Transmission**: Mostly automatic (1, 8-speed); manuals rare.

EXPLORING AND VISUALIZING DATA

Univariate Analysis - Discrete Variables





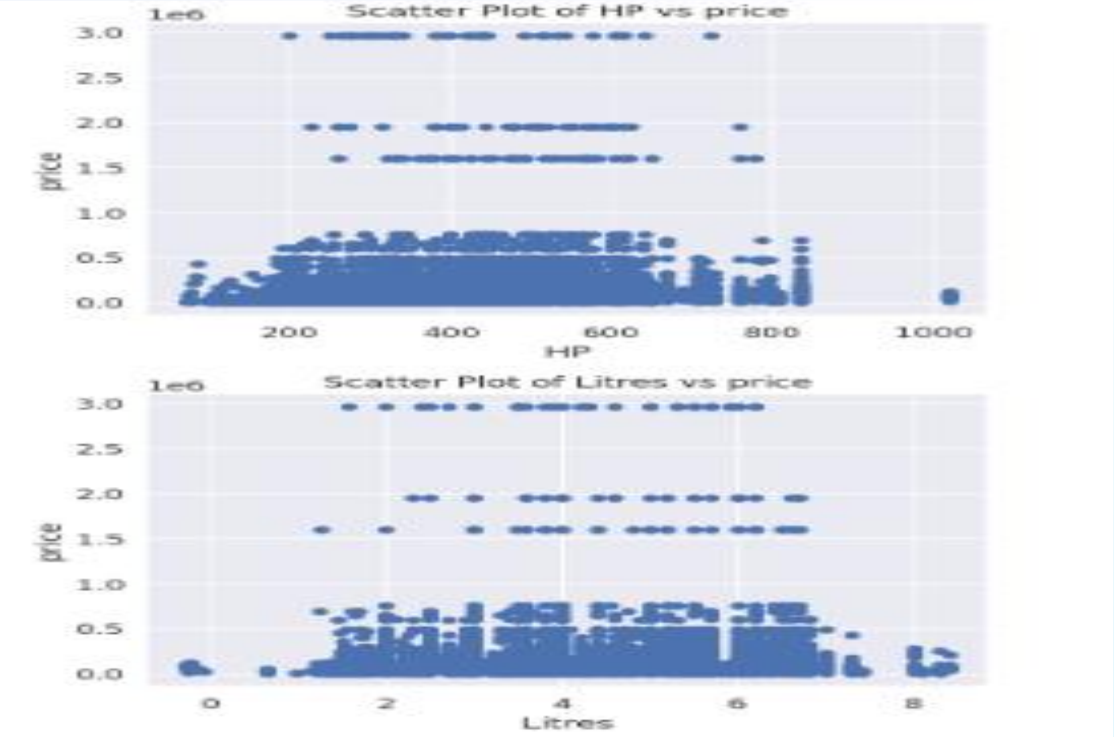
EXPLORING AND VISUALIZING DATA

Bivariate Analysis - Numerical vs. Price

- **Horsepower**: Higher HP → Higher price tiers; Most vehicles → Low price range
- **Engine Size (Litres)**: Top prices → 2–6L engines; Lower prices span all sizes
- **Cylinders**: 8–16 cylinders → Premium priced; Lower prices across all counts
- **Model Year**: Post-2000 → Wider price spread including high-end; Older → Mostly low-priced
- **Mileage**: Low mileage → High price; High mileage → Lower price

EXPLORING AND VISUALIZING DATA

Bivariate Analysis - Numerical vs. Price



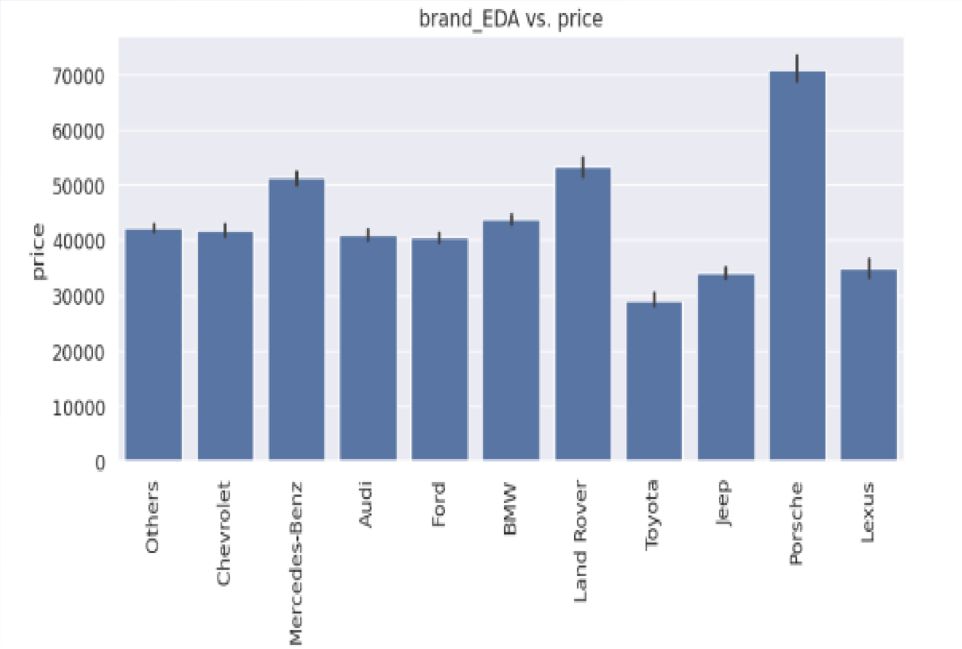
EXPLORING AND VISUALIZING DATA

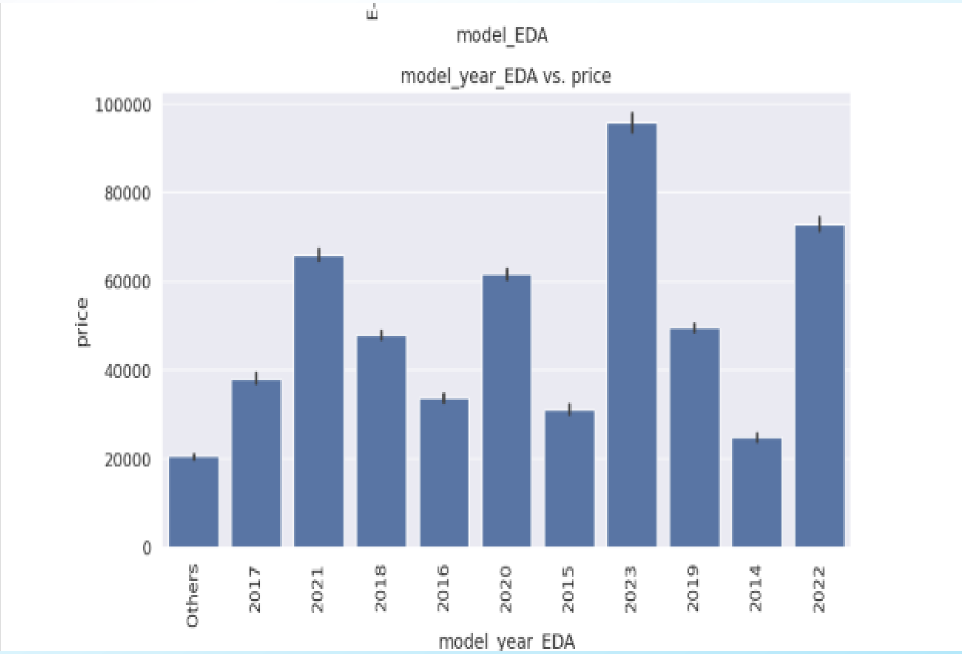
Bivariate Analysis - Categorical vs. Price

- **Brand**: Porsche → Highest average price; Toyota/Jeep → Lower end
- **Model Year**: Newer years (2022–2023) → Command higher prices; Older → Typically lower
- **Color**:
  - **Exterior**: Green/Orange → Higher prices; Silver, Gold, Brown → Lower
  - **Interior**: Orange/Others → Higher prices; Gray/Beige → Lower
- **Cylinder Layout**: W-layout engines → Significantly pricier than alternatives
- **Model**: Premium models (e.g. Porsche 911 Carrera S, 1500 Laramie) → Top-tier pricing
- **Fuel Type**: Electric → Highest prices; Flex-fuel → Lowest
- **Cylinders**: More cylinders → Generally linked to higher prices

EXPLORING AND VISUALIZING DATA

Bivariate Analysis - Categorical vs. Price





EXPLORING AND VISUALIZING DATA

Correlation Analysis

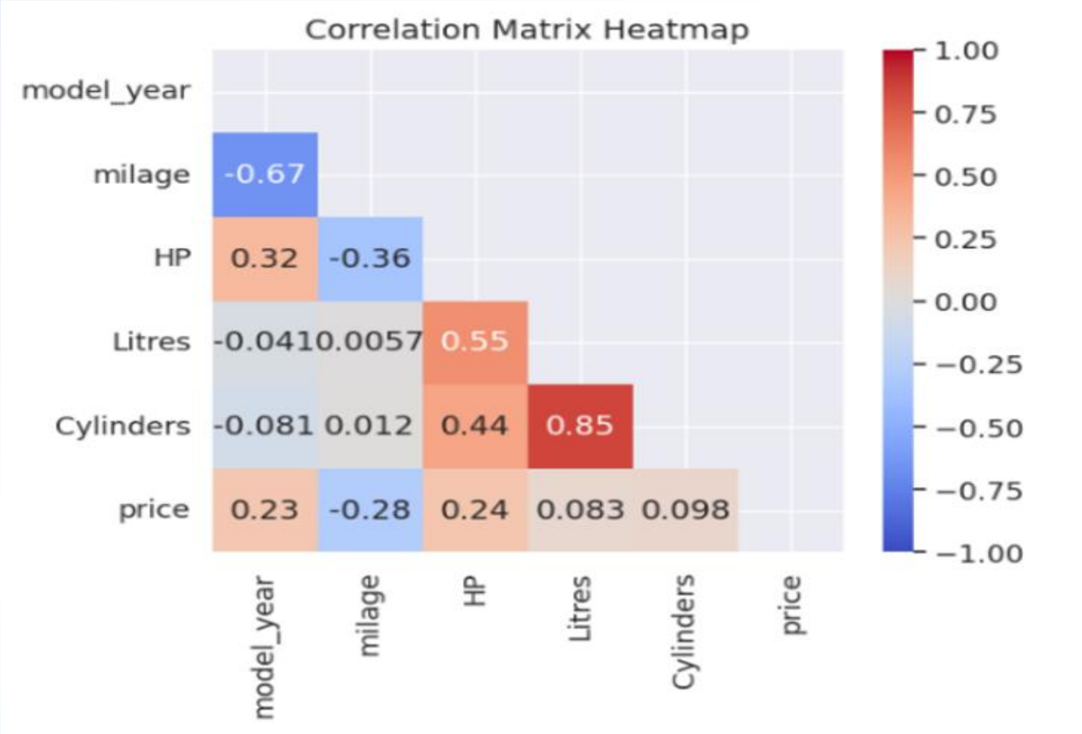
- **Litres & Cylinders**: Strong positive correlation (0.85).

- **HP & Litres**: Moderate positive correlation (0.55).

- **Model Year & Mileage**: Strong negative correlation (-0.67), indicating newer vehicles tend to have lower mileage.

EXPLORING AND VISUALIZING DATA

Correlation Analysis



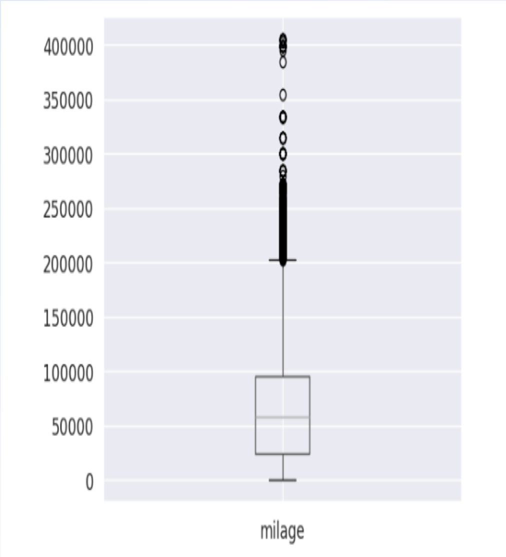
EXPLORING AND VISUALIZING DATA

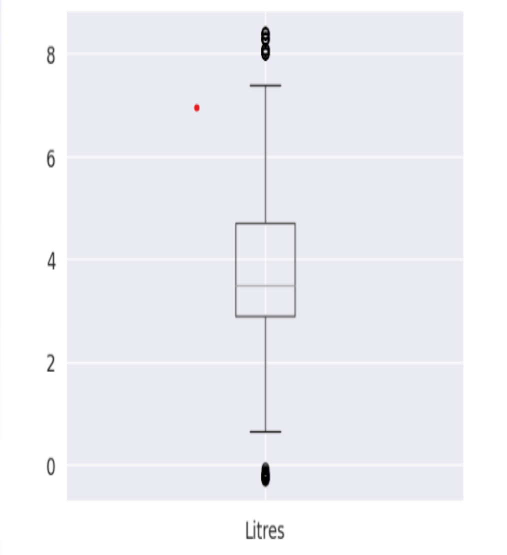
Outlier Analysis

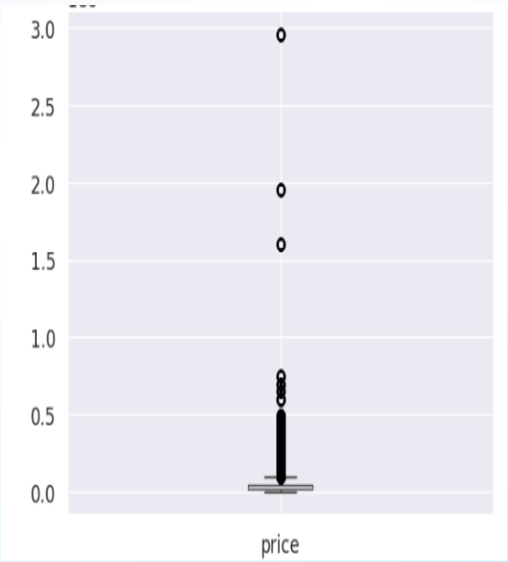
- **Observations**
  - **Litres**: Mostly compact; one standout outlier
  - **Horsepower (HP)**: Skewed with several high-HP outliers
  - **Mileage**: Broad spread; many high-mileage outliers
  - **Price**: Highly skewed; numerous high-price outliers
- **Treatment**
  - **Method**: Winsorization (capping) applied to all four variables
  - **Evidence**:
    - HP, Mileage, Price → Capped at upper quartile
    - Litres → Capped at both ends; minimum now positive
  - **Status**: No remaining outliers (based on IQR)

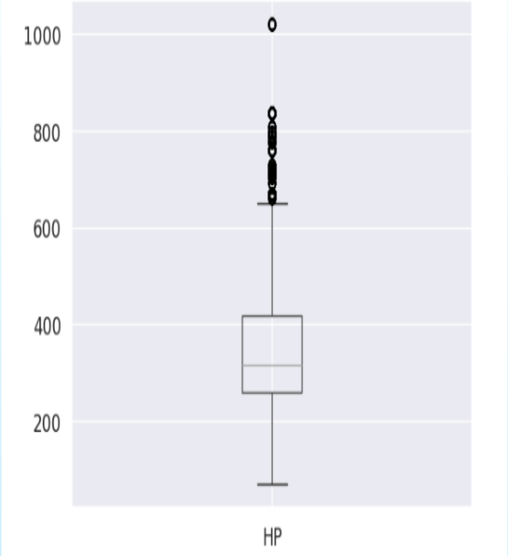
EXPLORING AND VISUALIZING DATA

Outlier Analysis









STATISTICAL ANALYSIS

Normality Testing - Numerical Variables

- **Test Used**: D’Agostino’s K^2 Test (suitable for $n \geq 20$, assesses skewness and kurtosis).
- **Variables**: Model_year, Mileage, HP, Litres, Cylinders, Price.
- **Results**: All p-values = 0.000, rejecting normality (non-Gaussian distributions).
- **Implications**: Non-parametric tests or data transformations (e.g., logarithmic) needed for future analysis.
- **Visualization**: Reference Figure (D’Agostino’s K^2 Test results).

STATISTICAL ANALYSIS

Normality Testing - Numerical Variables

```
Statistics=26194.461, p=0.000
Sample does not look Gaussian (reject H0) col  model_year
Statistics=14064.437, p=0.000
Sample does not look Gaussian (reject H0) col  milage
Statistics=6332.010, p=0.000
Sample does not look Gaussian (reject H0) col  HP
Statistics=2223.565, p=0.000
Sample does not look Gaussian (reject H0) col  Litres
Statistics=15699.246, p=0.000
Sample does not look Gaussian (reject H0) col  Cylinders
Statistics=21737.049, p=0.000
Sample does not look Gaussian (reject H0) col  price
```

STATISTICAL ANALYSIS

Kruskal-Wallis Test

- **Purpose**: Compare **two or more groups** on a continuous or ordinal variable **without assuming normality**.

- **Type**: Non-parametric alternative to one-way ANOVA.

Hypotheses:
- **Null (H_0)**: All group medians are equal.
- **Alternative (H_a)**: At least one group median differs.

Result :
- All Variables like brand_EDA, model_EDA, fuel_type, etc., show **statistically significant differences**—indicating **at least one group stands out**.

STATISTICAL ANALYSIS

Kruskal-Wallis Test

The Kruskal–Wallis test is a statistical test used to compare two or more groups for a continuous or discrete variable. It is a non-parametric test, meaning that it assumes no particular distribution of your data and is analogous to the one-way analysis of variance (ANOVA). The Kruskal Wallis test is sometimes referred to as the one-way ANOVA on ranks or the Kruskal Wallis one-way ANOVA.

The hypotheses of the Kruskal–Wallis test are as follows:

The H_0 - null hypothesis () is that the population medians are equal.

The H_A - alternative hypothesis () is that the population medians are not equal, or that the population median differs from the population median of one of the other groups.

[106]:

```
# Define the numerical and categorical columns for analysis
for var in cat_columns_EDA :
    chk_kruskal(df, var, target)
```

```
At least one group stands out from the rest, brand_EDA

At least one group stands out from the rest, model_EDA

At least one group stands out from the rest, transmission_EDA

At least one group stands out from the rest, cylinder_layout

At least one group stands out from the rest, fuel_type

At least one group stands out from the rest, ext_col_EDA

At least one group stands out from the rest, int_col_EDA

At least one group stands out from the rest, color_category

At least one group stands out from the rest, accident

At least one group stands out from the rest, model_year_EDA
```

STATISTICAL ANALYSIS

Dunn's Post Hoc Test

Purpose:

- Identify **which specific car brands differ** significantly in the target variable (e.g., price) after Kruskal-Wallis detects overall group differences.

Interpretation of Matrix:

- Each cell = **p-value** for brand pair comparison
- **Low p-value (< 0.05)** → significant difference in medians
- **High p-value (~1.0)** → no significant difference

Result:

- BMW vs. Audi, Chevrolet, Ford → significant differences ($p < 0.00001$)
- Luxury brands (e.g., Porsche, Land Rover, Lexus) differ sharply from mainstream brands
- Audi, Chevrolet, Ford → no significant difference among themselves
- Jeep vs. Lexus → borderline significance ($p \approx 0.028$).

STATISTICAL ANALYSIS

Dunn’s Post Hoc Test

```
for var in cat_columns_EDA :
    dunns_post_hoc(df, var, target)
```

	Audi	BMW	Chevrolet	Ford
Audi	1.000000e+00	3.647001e-07	1.000000e+00	1.000000e+00
BMW	3.647001e-07	1.000000e+00	5.375899e-06	2.768910e-06
Chevrolet	1.000000e+00	5.375899e-06	1.000000e+00	1.000000e+00
Ford	1.000000e+00	2.768910e-06	1.000000e+00	1.000000e+00
Jeep	1.302726e-27	8.074801e-63	2.763066e-36	5.610460e-41
Land Rover	4.506395e-107	1.357044e-76	1.425963e-116	1.000092e-127
Lexus	2.123561e-57	4.143991e-116	6.989128e-75	1.455308e-85
Mercedes-Benz	7.789148e-85	6.470090e-54	2.113313e-96	4.497394e-111
Others	3.201042e-26	6.250828e-99	4.451207e-45	1.766150e-61
Porsche	1.774249e-280	2.743059e-249	0.000000e+00	0.000000e+00
Toyota	7.247392e-184	1.352803e-300	1.300893e-229	7.315886e-260

	Jeep	Land Rover	Lexus	Mercedes-Benz
Audi	1.302726e-27	4.506395e-107	2.123561e-57	7.789148e-85
BMW	8.074801e-63	1.357044e-76	4.143991e-116	6.470090e-54
Chevrolet	2.763066e-36	1.425963e-116	6.989128e-75	2.113313e-96
Ford	5.610460e-41	1.000092e-127	1.455308e-85	4.497394e-111

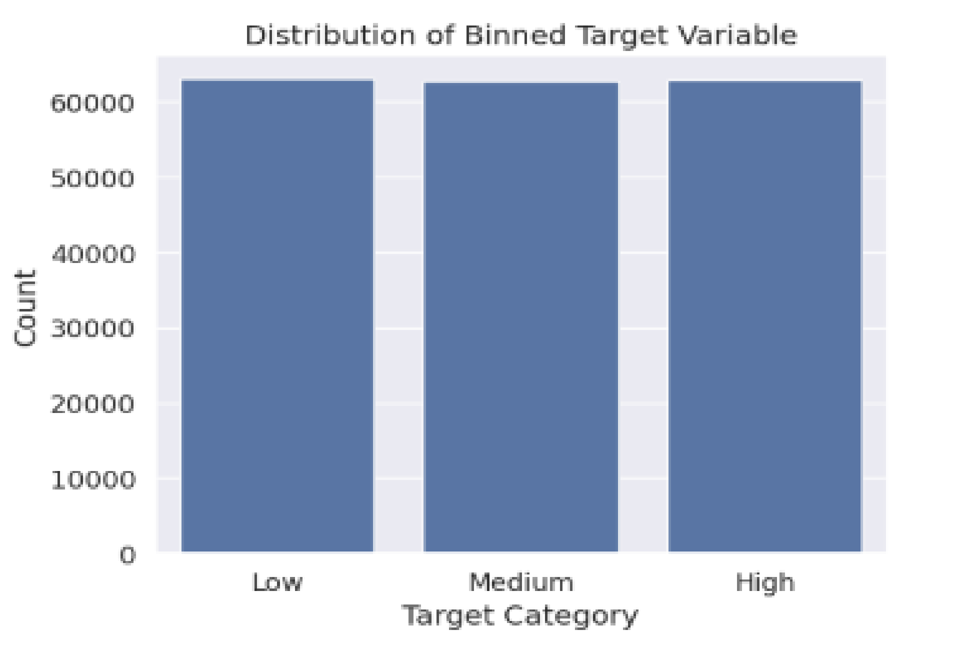
STATISTICAL ANALYSIS

Chi-Square Test - Categorical Variables

- **Test Used**: Chi-square test of independence (requires categorical data).
- **Target Variable**: Numerical target binned into categories (e.g., Low, Medium, High).
- **Variables Tested**: brand_EDA, model_EDA, transmission_EDA, cylinder_layout, fuel_type, ext_col_EDA, int_col_EDA, color_category, accident, model_year_EDA.
- **Assumptions Met**:
  - All variables categorical, observations independent.
  - 0% of expected cell counts < 5, satisfying test requirements.
- **Results**: All p-values = 0.000, indicating significant dependence between each categorical variable and binned target.
- **Visualization**: Reference Figures (Chi-square test validity) and (Chi-square results).

STATISTICAL ANALYSIS

Chi-Square Test - Categorical Variables



#	Variable	P_ value	% of expected frequency cells having cell count < 5
1	brand_EDA	0%	0%
2	model_EDA	0%	0%
3	transmission _EDA	0%	0%
4	cylinder_layo ut	0%	0%
5	fuel_type	0%	0%
6	ext_col_EDA	0%	0%
7	int_col_EDA	0%	0%
8	color_catego ry	0%	0%
9	accident	0%	0%
10	model_year	0%	0%

STATISTICAL ANALYSIS

Chi-square contingency test

- The Chi-square test compares how each brand distributes across **Low**, **Medium**, and **High** price bins.

- **Why Dunn’s Test Can’t Do This**
  Dunn’s test compares **pairwise median prices**, but it doesn’t reveal how a brand’s listings are **distributed across price segments**.
  It lacks the **categorical breakdown** needed for segmentation planning, marketing personas, or inventory tiering.

Example

- **Chi-square tells you**: “80% of Porsche listings fall in the High-price bin—confirming its luxury status.”
- **Dunn’s test tells you**: “Porsche has a significantly higher median price than Ford”—but not how many listings are high-priced

STATISTICAL ANALYSIS

Chi-square contingency test

```
# Define the categorical columns for analysis
categorical_columns    =  [ 'brand_EDA', 'model_EDA', 'transmission_EDA', 'cylinder_layout', 'fuel_type', 'ext_col_EDA', 'int_col_EDA',\
                            'color_category', 'accident','model_year_EDA']
for i, var in enumerate(categorical_columns):
    i += 1
    chk_chisq(df, i, var, 'target_binned')
```

```
1: Variable, brand_EDA

The important assumption: No more than 20% of the cells have and expected cell count < 5

This can be checked by looking at the expected frequency table.
Chi2ContingencyResult(statistic=6845.78382504452, pvalue=0.0, dof=20, expected_freq=array([[ 3639.37500597,  3619.45268468,  3628.17230936],
       [ 5692.22720691,  5661.06735691,  5674.70543618],
       [ 5460.56679732,  5430.67508076,  5443.75812192],
       [ 7718.67078973,  7676.41797457,  7694.9112357 ],
       [ 2165.17239953,  2153.32001825,  2158.50758223],
       [ 3184.74563074,  3167.31199843,  3174.94237083],
       [ 2889.90510945,  2874.08546514,  2881.00942541],
       [ 6408.9370455 ,  6373.85385052,  6389.20910398],
       [19358.51994081, 19252.54936802, 19298.93069118],
       [ 3547.44627201,  3528.02717827,  3536.52654973],
       [ 2958.43380204,  2942.23902447,  2949.32717349]]))


Percentage of cells with expected counts less than 5: 0.00%

Independent Variable,brand_EDA and Target variable are dependent
```

DATA PREPROCESSING AND FEATURE SELECTION

DATA PREPROCESSING

- **TRANSFORMATION**
  - Label encoder is used.
  - In order to build a regression model, we have performed label encoding to categorical variables.

- **SCALING**
  - Applied Min-Max Scaling to Mileage and Price
  - Reason: Standard and Robust Scalers gave negative values, which are not suitable for our data.
  - Min-Max Scaler kept all values between 0 and 1 without turning them negative

DATA PREPROCESSING AND FEATURE SELECTION

FEATURE SELECTION

Method Used:

Applied Recursive Feature Elimination (RFE) to select key features for the regression model.

Why RFE?

- It Improves model performance by removing less relevant features.
- Reduces dimensionality for better interpretability.
- Highlights feature importance.

Selected Features:

- We selected the top 10 features out of 15 based on their importance.
- This helped keep the model accurate and simple. The other 5 features had less impact and were not useful.
- Choosing fewer, important features also helped prevent overfitting.

ADVANCED REGRESSION MODELS AND HYPERPARAMETER TUNING

BASE MODEL

Our initial base model, Ordinary Least Squares (OLS), was found to be unsuitable as it violated key statistical assumptions.

Linearity:

✕ Not met — Scatter plot of actual vs. predicted values did not align well.

Normality of Residuals:

✕ Not met — Residuals failed the normality test (p-value < 0.05)

ADVANCED REGRESSION MODELS AND HYPERPARAMETER TUNING

BASE MODEL

No Autocorrelation:

✓ Met — Durbin-Watson statistic ≈ 1.99 indicates no autocorrelation

Homoscedasticity (Constant Variance of Errors):

✓ Met — Residuals were evenly spread around zero

No Multicollinearity:

✓ Met — VIF values for final features were below threshold

ADVANCED REGRESSION MODELS AND HYPERPARAMETER TUNING

BASE MODEL

- Model Choice: Generalized Linear Models (GLM)
- Why: Relaxes normality and linearity assumptions for linear regression models.
- Performance Metrics:
  - Pseudo R^2 = 70% → Shows that the model fits the data well
  - Train RMSE = 0.1334, Test RMSE= 0.1348 → Very similar values, Which means no overfitting
- Among the models we tried, GLM gave the best performance.

ADVANCED REGRESSION MODELS AND HYPERPARAMETER TUNING

BASE MODEL

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                  price   No. Observations:               150826
Model:                            GLM   Df Residuals:                   150819
Model Family:                Gaussian   Df Model:                            6
Link Function:               Identity   Scale:                        0.017788
Method:                          IRLS   Log-Likelihood:                 89848.
Date:                Thu, 17 Jul 2025   Deviance:                       2682.7
Time:                        13:31:15   Pearson chi2:                 2.68e+03
No. Iterations:                     3   Pseudo R-squ. (CS):             0.7014
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
const              0.8711      0.001    601.595      0.000       0.868       0.874
brand_EDA         -0.0005      0.000     -4.552      0.000      -0.001      -0.000
model_year_EDA    -0.0062      0.000    -58.027      0.000      -0.006      -0.006
int_col_EDA        0.0012      0.000      9.961      0.000       0.001       0.001
color_category     0.0026      0.000      5.839      0.000       0.002       0.003
accident           0.0350      0.001     42.053      0.000       0.033       0.037
milage            -0.5670      0.002   -367.769      0.000      -0.570      -0.564
==================================================================================
Psuedo R square 0.7014
Train Root Mean Squared Error(RMSE): 0.1334
Test Root Mean Squared Error(RMSE): 0.1348
```

ADVANCED REGRESSION MODELS AND HYPERPARAMETER TUNING

Models Built

- **Models**: Linear Regression, GLM (Gaussian, identity link), Decision Tree (CART), Random Forest, K-Nearest Neighbors (KNN), XGBoost, CatBoost, LightGBM.

- **Purpose**: Predict used car prices using diverse approaches (linear, tree-based, ensemble, and distance-based models).

ADVANCED REGRESSION MODELS AND HYPERPARAMETER TUNING

Evaluation Methodology

- **Validation**: 10-fold cross-validation to ensure robustness and prevent overfitting.
- **Metrics**: RMSE (measures prediction error) and Explained Variance (measures variance explained by the model).
- **Tuning**: Hyperparameter tuning applied to Decision Tree, Random Forest, KNN, XGBoost, CatBoost, LightGBM for improved performance.
- **Baseline**: GLM with Pseudo R^2 = 70%, RMSE = 0.1334 (train), 0.1348 (test), indicating no overfitting.

ADVANCED REGRESSION MODELS AND HYPERPARAMETER TUNING

RMSE

- **Improvement Across Models**: Tuning reduced RMSE for all models.
- **Largest Gain**: Decision Tree showed the biggest drop—from 0.1517 to 0.1147.
- **Consistent Leaders**: LightGBM and CatBoost consistently outperformed GLM (0.1300) before and after tuning.
- **Stability**: Low standard deviations (~0.0010) across 10-fold CV indicate highly stable performance.

ADVANCED REGRESSION MODELS AND HYPERPARAMETER TUNING

RMSE

ADVANCED REGRESSION MODELS AND HYPERPARAMETER TUNING

Explained Variance

- **Improvement Across Models**: Tuning increased Explained Variance for all models.
- **Largest Gain**: Decision Tree rose from 0.4146 to 0.6654.
- **Consistent Leaders**: LightGBM and CatBoost consistently outperformed GLM (0.567), both pre- and post-tuning.
- **Top Performers**: Ensemble models (LightGBM, CatBoost, Random Forest) explained the most variance in used car prices.

ADVANCED REGRESSION MODELS AND HYPERPARAMETER TUNING

Explained Variance

COMPARISON OF BEST MODEL

- **Baseline**: GLM used as benchmark; mean explained variance = 0.567, RMSE = 0.130 ± 0.001.
- **GLM Limitations**: Violated key linear regression assumptions; residuals not normally distributed, data non-linear.
- **Tuning Strategy**: Bayesian optimization + K-Fold CV applied across models.
- **Best Performer**:
  - **LightGBM** achieved highest explained variance (0.6867)
  - **Lowest RMSE**: 0.1110 ± 0.0010
  - **Why LightGBM?**: Efficient, accurate, scalable for large datasets and complex feature interactions—ideal for used car price prediction.

CONCLUSION AND FUTURE SCOPE

CONCLUSION

- **Success**: Leveraged robust preprocessing, feature selection, and advanced ML models (e.g., LightGBM, CatBoost) to achieve low RMSE and high Explained Variance, delivering reliable used car price predictions.

- **Impact**: Enhanced market transparency and pricing accuracy, supporting stakeholders in decision-making.

CONCLUSION AND FUTURE SCOPE

Future Scope

- **Data Enhancement**: Incorporate recent datasets for model validation and refinement.
- **Advanced Techniques**: Explore ensemble methods, deep learning, or stacking for better performance.
- **Real-World Collaboration**:
  - Partner with dealerships, platforms (e.g., CarDekho, Cars24), and fleet firms.
  - Pilot: Deploy for pricing, valuation; run A/B tests.
  - Feedback Loop: Collect inputs from teams/customers; track deviations.
  - Metrics: RMSE, Explained Variance, customer trust, sale time, feedback sentiment.
  - Iteration: Refine models; expand to SaaS/licensing for broader adoption.

THANK YOU

Q&A