**Toyota Corolla**

**Advance Project-4**

Predicting Car Prices – Toyota Corolla Dataset

**Objective**

The goal of this project is to develop a Multiple Linear Regression model that can accurately predict the price of a used Toyota Corolla based on various features such as age, mileage (KM), horsepower, fuel type, and other technical specifications.

**Dataset Information**

* **Dataset Name:** Toyota Corolla
* **Source:** [Provide source link or file name]
* **Size:** ~1,400 records, 30+ columns
* **Target Variable:** Price
* **Feature Types:**
  + Numerical: Age\_08\_04, KM, HP, Weight, cc, etc.
  + Categorical: Fuel\_Type, Color, Airco, etc.
  + Binary: ABS, Airbag\_1, CD\_Player, etc.

**Solution Architecture**

Raw Data (.csv) ──► Data Cleaning ──► Feature Encoding ──► EDA &

Correlation

│

▼

Feature Scaling & Transformation ──► Model Building (MLR)

│

▼

Evaluation (R² Score, OLS Summary)

* **Tools Used:** Python, Pandas, Seaborn, Scikit-learn, Statsmodels
* **Modeling Techniques:**
  + Raw Linear Regression
  + Standard Scaled Regression
  + Power Transformed Regression
* **Model Comparison Metric:** R² Score

**Methodology**

| **Step** | **Description** |
| --- | --- |
| **Data Loading** | Imported dataset using pandas.read\_csv() |
| **EDA** | Used Seaborn for pairplots, heatmaps, and boxplots |
| **Feature Selection** | Removed non-informative columns like ID and Model |
| **Encoding** | One-hot encoding for categorical variables |
| **Transformations** | Tried Standard Scaling and Power Transformer |
| **Modeling** | Used LinearRegression() from sklearn |
| **Evaluation** | Compared R² scores across models, used OLS summary for coefficient significance |

**Time Taken**

| **Phase** | **Time Spent** |
| --- | --- |
| Data Understanding & Cleaning | 25 mins |
| Exploratory Data Analysis | 30 mins |
| Encoding & Preprocessing | 20 mins |
| Modeling & Evaluation | 30 mins |
| Documentation & Reporting | 30 mins |
| **Total** | **~2 hours 15 mins** |
| **Challenges Faced**   * Encoding categorical columns without introducing object dtype issues * Dealing with ValueError: Pandas data cast to numpy dtype of object in statsmodels.OLS() * Managing large number of dummy variables * Outliers in features like KM and cc * Ensuring transformations do not distort interpretability |  |

**Complexity**

* **Level:** Intermediate
* The modeling itself was straightforward, but handling a large dataset with many categorical features added complexity.
* Transformation tuning required experimentation to improve accuracy.

**R² Score Comparison Table**

| **Model Variant** | **R² Score** |
| --- | --- |
| Raw Data | 0.86 |
| Standard Scaled | 0.86 |
| Power Transformed | 0.88 |

**Conclusion**

Power Transformations led to the best performance.  
Key drivers of price included **Age**, **KM**, **HP**, and **Weight**.  
The final model provides a reliable, data-driven basis for used car price predictions, which can assist both customers and dealerships.