# Project Introduction

The goal of this machine learning project is to develop a regression model to precisely predict the selling price of a used car based on various features such as year, km driven, fuel type, seller type and transmission. The project involves cleaning and preprocessing real-world car listing data, performing exploratory data analysis, selecting relevant features and training a predictive model to estimate car prices. I chose this problem because buying and selling used cars is a common activity. However, pricing can be inconsistent and subjective. A data-driven pricing tool can help buyers avoid overpaing and sellers price competitively.

# Dataset

Dataset: [Car data](https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?select=Car+details+v3.csv)

I researched and found the car price dataset from Kaggle platform. As I have mentioned, I have used these features: year, km driven, fuel type, seller type and transmission. Feature selection is done by logic and correlations. Based on the correlation matrix, I found the related and non-related features on the selling price of the car.

# Modelling Approach

Modelling method used: Polynomial Regression

## Preprocessing

The data in the dataset is not cleaned and needed to do some cleaning and transformation.

In some categorical data, there are a few missing values and needed to clean these out.

Based on the correlation matrix, I needed to use categorical features. Some data transformation methods such as encoding the categorical data into numeric data.

For example, under fuel column, there are 4 types of unique value: Diesel, Petrol, CNG and LPG. So, they are encoded as Diesel = 0, Petrol = 1, CNG = 2 and LPG = 3.

Likewise, I needed to do for other columns: Seller type and transmission.

To capture potential non-linear relationships between the features and the target variable (car selling price), Polynoial Regression was selected as one of the regression techniques.

Furthermore, to determine the best polynomial degree, a GridSearchCV approach was used in combination with a pipeline that included the following components:

1. StandardScaler – to normalize all features
2. PolynomialFeatures – to generate non-linear features of varying degrees.
3. LinearRegression – to fit the transformed data.

After evaluating a total of 6 polynomial degrees: 2,3,4,5,7,9 across 5 folds each (30 fits total), GridSearchCV selected the best model pipeline based on the lowest mean squared error across the validation folds. The optimal polynomial degree was 2.

# Evaluation

The performance of the model is moderate. From the metrics data,

Training data in the model:

Mean absolute error: 257563.28

Mean squared error: 252684033739.49

R2 score is: 0.62

Testing data in the model:

Mean absolute error: 252969.26

Mean squared error: 241438919730.05

R2 score is: 0.64

The metrics used to validate are mean absolute error (mae), mean squared error (mse) and r2 score.

# Reflection

## Challenges:

The first thing I encountered was to find the dataset. There are a lot of datasets out there, but this is my first project starting from scratch (finding my own dataset). So, it is a little bit challenging.

During the cleaning and transformation processes, there are some burdens such as I do not know how to find the missing values and encode the categorical data into numeric data.

At first, I added all the features available without feature selection and train the model. But the model performance became inaccurate and even the predicted car price value was negative. So later I needed to perform the correlation matrix to do the feature selection.

## Lessons and Future Improvements

I researched and learnt how to clean and manipulate data in pandas.

I finally realized the whole process of supervised learning starting from scratch even though this is a small project.

If I do another project or start over this project next time, I will check the correlation between each feature and decide first before training the model.