



Linear regression Assignment

Problem statement :

- A Chinese automobile company **Geely Auto** aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts.
- They have contracted an **automobile consulting company** to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. The company wants to know:
- Which variables are significant in predicting the price of a car
- How well those variables describe the price of a car
- Based on various market surveys, the consulting firm has gathered a large dataset of different types of cars across the American market.

Business Goal

- You are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

Data Preparation

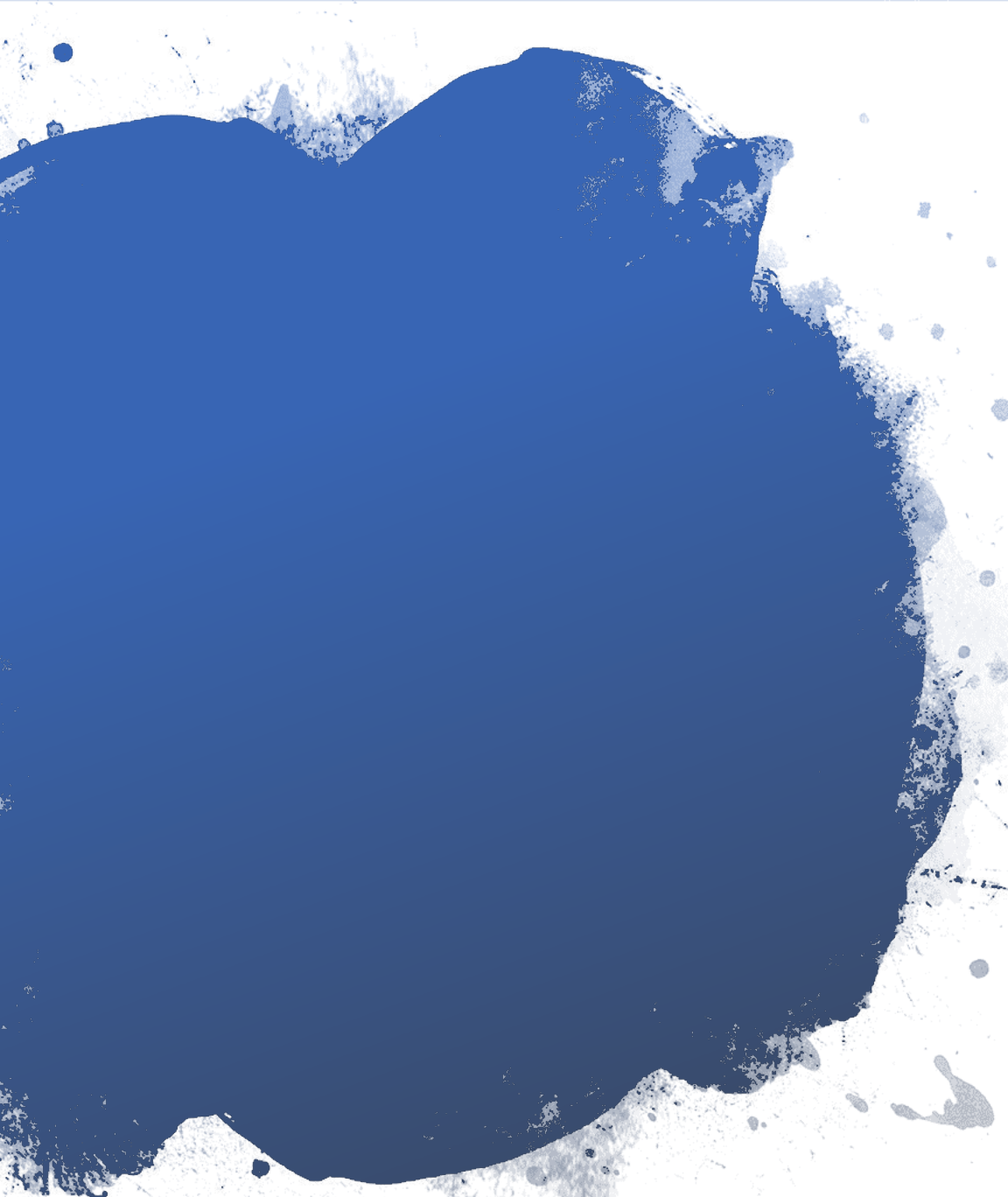
- There is a variable named **CarName** which is comprised of two parts - the first word is the name of 'car company' and the second is the 'car model'. For example, **chevrolet impala** has 'chevrolet' as the car company name and 'impala' as the car model name. You need to consider only company name as the independent variable for model building.

Model Evaluation:

- When you're done with model building and residual analysis, and have made predictions on the test set, just make sure you use the following two lines of code **to calculate the R-squared score on the test set.**

```
from sklearn.metrics import r2_score  
r2_score(y_test, y_pred)
```

- where *y_test* is the test data set for the target variable, and *y_pred* is the variable containing the predicted values of the target variable on the test set.

- 
- **Please don't forget to perform this step** as the R-squared score on the test set holds some marks. The variable names inside the 'r2_score' function can be different based on the variable names you have chosen.



Submission :

- **Submissions required:** A well-commented Jupyter notebook.