Project Report:

Project: regression Analysis.

Dataset: List of 2nd hand used cars.

Variables we have in dataset: Brand, price, body, Mileage, Engine Volume, Type, Year, Registration, Model.

Expected outcome: We want to predict the price of car depending on its specs(regressors)

1st regressor: **Brand**, as some brands are more expensive than others.

2nd regressor: **Mileage**, more driven means cheaper.

3rd regressor: Engine volume, as sports cars have larger engines and economy cars have smaller ones.

4th regressor: Production years, older cars are cheaper than latest ones.

Basic data cleaning done on dataset as we had missing values and 4.4% of the missing values were removed from the dataset.

Regression analysis:

1. **Linearity**

Checked by using the Scatter plot and used log transformation to fix those.

2. **Endogeneity:**

Assumptions was not violated for this dataset so skipping this.

3. **Normality & Homoscedasticity:**

For sufficiently large sample sizes, the distribution of errors is assumed to be normal, a principle underpinned by the Central Limit Theorem. Furthermore, the regression model's inclusion of an intercept term is crucial for achieving a zero mean in the distribution of these errors. And homoscedasticity holds as we can see in the graph. Reason is we already implemented the log transformation.

4. **No Autocorrelation:**

Since data is not time series or panel data. It is simply snapshot of certain time of data of 2nd hand car sales website. No reason for observation to be dependent on each other.

5. **Multicollinearity:**

To check this, we find the correlation of all variables with each other to check. Since correlations are far from strong so we conclude that we are not in presence of multicollinearity. Now creating the Regression.

Regression analysis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.90
R Square	0.80
Adjusted R Square	0.80
Standard Error	0.18
Observations	4003.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9.00	526.97	58.55	1795.03	0.000
Residual	3993.00	130.25	0.03		
Total	4002.00	657.22			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-78.41	1.02	-77.16	0.000
log_mileage	-0.11	0.01	-20.19	0.000
EngineV	0.14	0.00	41.60	0.000
Year	0.04	0.00	81.43	0.000
Audi_D	0.06	0.01	5.29	0.000
BMW_D	0.11	0.01	10.96	0.000
Mercedes_D	0.10	0.01	10.97	0.000
Mitsubishi_D	-0.05	0.01	-4.06	0.000
Renault_D	-0.16	0.01	-15.04	0.000
Toyota_D	0.04	0.01	3.91	0.000

The model and all variables are significant. Adjusted R is 0.80, which means variables together explain 80% of the total variability. Model has very high explanatory power.

We used 6 dummy variables for brands (N-1), as they're categorical. This prevents multicollinearity; including a separate dummy for Volkswagen would create perfect dependency, as it's implied when all other brand dummies are zero.

Interpretation:

log Price = -78.41 - 0.11 x log Mileage + 0.14 x EngineV + 0.04 x Year + 0.06 x Audi_D +0.11 x BMW_D + 0.1 x Mercedes_D-0.05 x Mitsubishi_D-0.16 x Renault_D + 0.04 x Toyota_D

- Intercept is clear.
- For each % change in Mileage price decreases by 0.11% or 11 %.
- For each extra litre of engine volume price increases by 0.14 or 14%.
- For each additional year price will increase by 4%.

- We only kept 6 dummy variables(N-1) as to avoid multicollinearity and the benchmark is Volkswagen.
- So, if the car is Audi then price of that car would be 6% higher than Volkswagen and so on.

Showcasing linearity before and after applying the log transformations.