IE 5318 - APPLIED LINEAR REGRESSION

MULTIPLE LINEAR REGRESSION PROJECT REPORT

CAR RESALE VALUE

<u>TEAM MEMBERS</u>

AKSHATA AGINE - 1002065578 - ana5578@mavs.uta.edu
VINAY HIRVE - 1002122413 - vxh2413@mavs.uta.edu
BHARATH RACHARLA -1001848518 - bxr8519@mavs.uta.edu

We did not provide or receive any help on this report; the report submitted is our work.

Akshata Agine	**Vinay Hirve**	**Bharath Racharla**
	V. V. Hirve	

Date of Submission: 12/13/2023

DATASET BACKGROUND AND DESCRIPTION:

We went forward with the same dataset used in (SLR), sourced from Kaggle. After performing data cleaning in Microsoft Excel, we selected a random sample of 100 observations for our analysis. Our focus is on the *resale price*, which serves as the response variable, while mileage, registered year, and kilometers driven are our three predictor variables for Multiple Linear Regression (MLR).

Response Variable

Resale Price: The monetary value at which a used car is sold in the secondary market.

Predictor Variable

Mileage: The total distance a car has traveled, typically measured in miles or kilometers (for this project), indicating its fuel efficiency and overall wear.

Registered Year: The calendar year in which the car was officially registered, reflecting its age and often influencing its resale value.

Kilometer Driven: The total distance a car has traveled, measured in kilometers, providing insight into its usage and potential impact on its condition and value.

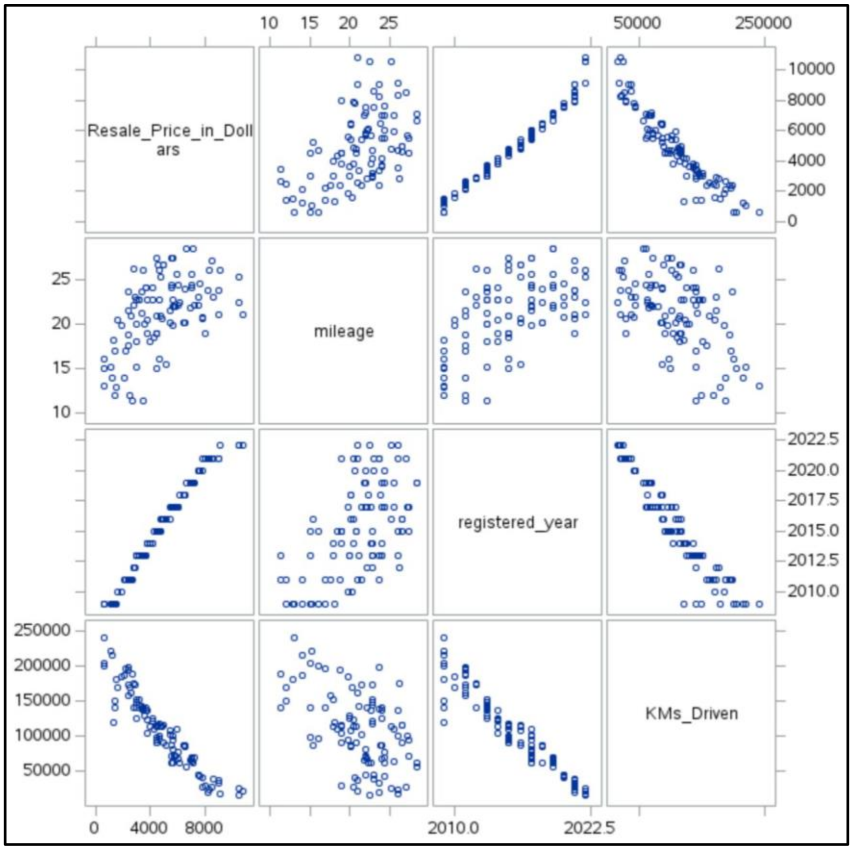

Figure 1. Matrix Scatter Plot for preliminary model

Figure 1 shows the Matrix Scatter Plot analysis for the preliminary model, this shows the correlation between the response variables, which are the Resale_Price_in_Dollars of the car, and the predictor variables are mileage(X_1), registered_year(X_2), and KMs_Driven(X_3). The above plot shows not much of a linear relationship between resale price and mileage, whereas resale price has a positive linear association with registered year and a negative linear relationship with the kilometers driven.

The CORR Procedure

4 Variables:	Resale_Price_in_Dollars mileage registered_year KMs_Driven

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Resale_Price_in_Dollars	100	4823	2368	482340	600.00000	10800	Resale_Price_in_Dollars
mileage	100	21.18220	4.06810	2118	11.30000	28.40000	mileage
registered_year	100	2015	3.74127	201527	2009	2022	registered_year
KMs_Driven	100	109012	53280	10901231	15683	240250	KMs_Driven

Pearson Correlation Coefficients, N = 100
Prob > |r| under H0: Rho=0

	Resale_Price_in_Dollars	mileage	registered_year	KMs_Driven
Resale_Price_in_Dollars Resale_Price_in_Dollars	1.00000	0.52526 <.0001	0.98614 <.0001	-0.93970 <.0001
mileage mileage	0.52526 <.0001	1.00000	0.55453 <.0001	-0.51910 <.0001
registered_year registered_year	0.98614 <.0001	0.55453 <.0001	1.00000	-0.94914 <.0001
KMs_Driven KMs_Driven	-0.93970 <.0001	-0.51910 <.0001	-0.94914 <.0001	1.00000

Table 1. Correlation Analysis

Upon analyzing the correlation coefficients presented in Table 1, it is evident that there exists a combination of positive linear associations among the predictor variables. While some correlation values surpass the threshold of 0.7, indicating a high degree of interrelation between specific variables, others fall below this threshold. This mixed correlation pattern suggests a nuanced situation regarding multicollinearity in the dataset. The variables exhibiting correlations greater than 0.7 may pose challenges in terms of potential multicollinearity, raising concerns about the interpretation of individual predictor effects and the overall stability and reliability of the regression model. However, the variables with correlations below 0.7 suggest that multicollinearity might not be a pervasive issue across all predictors. Careful consideration and further diagnostics are warranted to assess the impact of multicollinearity on the regression model's performance and the reliability of its results.

The REG Procedure
Model: MODEL1
Dependent Variable: Resale_Price_in_Dollars Resale_Price_in_Dollars

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	540247462	180082487	1165.11	<.0001
Error	96	14837982	154562		
Corrected Total	99	555085444			

Root MSE	393.14413	R-Square	0.9733
Dependent Mean	4823.40000	Adj R-Sq	0.9724
Coeff Var	8.15077		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1231640	69642	-17.69	<.0001
mileage	mileage	1	-17.93270	11.67616	-1.54	0.1279
registered_year	registered_year	1	613.82032	34.46651	17.81	<.0001
KMs_Driven	KMs_Driven	1	-0.00156	0.00236	-0.66	0.5086

Table 2. Analysis of Variance

Table 2's ANOVA reveals variable significance, with some below 0.1 and others surpassing one. The model's SSE is 14,837,982, with 96 degrees of freedom, while SSR is 540,247,462 with 3 degrees of freedom. SSTO is 555,085,444. Parameter estimates show an intercept (b_0) of -1,231,640 and slopes (b_1, b_2, b_3) of -17.93270, 613.82032, and -0.00156, respectively. These findings offer insights into variable impacts and model performance.

Model Form: $\boldsymbol{Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i}$
Where, i = 1,.. 2981 observations
Y_i is the response when p-1 regressors are set to $(x_{i1}, x_{i2},.., x_{p-1})^T$
β is the vector of unknown parameter $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^T$
x_i is the vector of p-1 regressors $(x_{i1}, x_{i2},.., x_{p-1})^T$, here in this project p = 4
ϵ is the error term (equation error + measurement error)
idd Normal $(0, \sigma^2)$

Model:
Resale_Price_in_Dollars = -1231640 - 17093270(mileage) + 613.82032(registered_year) - 0.00156(KMs_Driven)

The REG Procedure
Model: MODEL1
Dependent Variable: Resale_Price_in_Dollars Resale_Price_in_Dollars

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	540247462	180082487	1165.11	<.0001
Error	96	14837982	154562		
Corrected Total	99	555085444			

Root MSE	393.14413	R-Square	0.9733
Dependent Mean	4823.40000	Adj R-Sq	0.9724
Coeff Var	8.15077		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-1231640	69642	-17.69	<.0001	0
mileage	mileage	1	-17.93270	11.67616	-1.54	0.1279	1.44516
registered_year	registered_year	1	613.82032	34.46651	17.81	<.0001	10.65034
KMs_Driven	KMs_Driven	1	-0.00156	0.00236	-0.66	0.5086	10.09576

Table 3. Anova table with VIF values of the Predictor Variables

From table 3, VIF values for registered year and Kilometers driven are greater than 5, this indicates that there is serious multicollinearity.

VIF_{avg} = (1.44516+10.65034+10.09576) / 3 = 7.397

The average VIF is 7.397, so we can conclude that there is not much serious multicollinearity.

RESIDUALS VS PREDICTED VALUES:

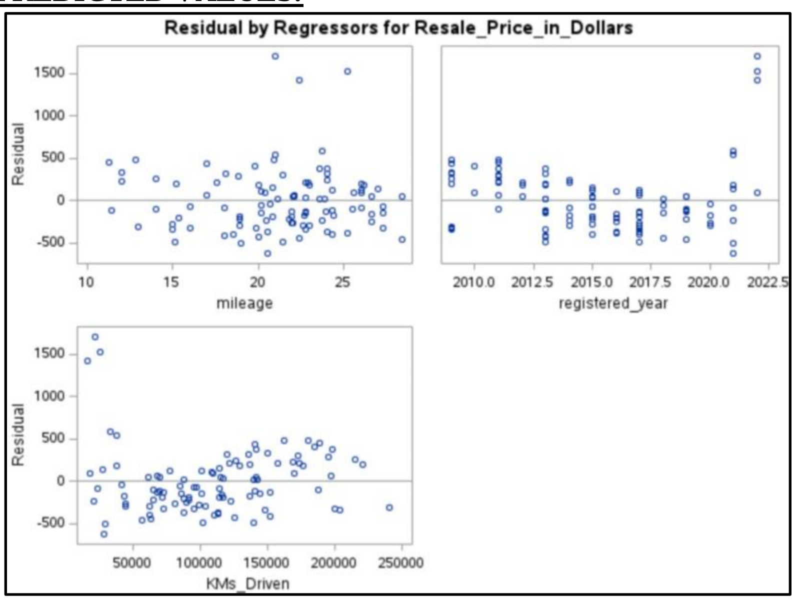

Figure 2. Residual(e) vs Response variable(ŷ)

In figure 2, Residuals vs. Predicted Value of Resale_Price_in_Dollars, it is evident that no curvature is present, and the data points exhibit random dispersion without a funnel shape, suggesting the fulfillment and non-violation of the constant variance assumption. This allows us to reasonably assume that the model adheres to underlying assumptions. Nevertheless, potential outliers in both predictor (X) and response (Y) variables are apparent, prompting a more in-depth investigation through outlier tests to ensure the robustness of the model.

RESIDUALS VS NORMAL SCORES:

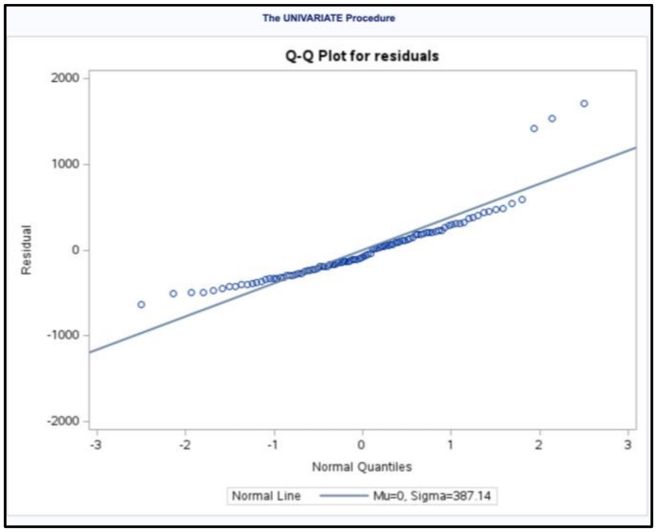

Figure 3. Residuals vs Normal Scores

Figure 3, showcasing the Residuals vs. Normal Scores (Q-Q plot), reveals a deviation from a straight line, indicating potential non-normality in the residuals. The presence of a long-left tail and a short-right tail raises concerns about the normality assumption. To verify this, a formal Normality test is recommended to provide a conclusive assessment of whether the residuals adhere to a normal distribution.

TEST FOR NORMALITY:

The CORR Procedure

2 Variables:	e enrm

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
e	100	8.1031E-11	387.14158	8.10314E-9	-629.95010	1705	e(Y| x1,x2,x3)
enrm	100	0	0.98921	0	-2.49859	2.49859	Normal Scores

Pearson Correlation Coefficients, N = 100

	e	enrm
e e(Y| x1,x2,x3)	1.00000	0.91940
enrm Normal Scores	0.91940	1.00000

Table 4. Test for Normality

Where, α= 0.10
H_0 = Normality is OK
H_1= Normality is violated
Decision rule: Reject H_0 if $\hat{\rho}$< c (α, n). $\hat{\rho}$ = 0.91940
From table B.6 (Kutner et al.) c (α = 0.10, n = 100) = 0.989
$\hat{\rho}$ < c (α, n), Hence, we reject H_0
Therefore, we are 90% confident that normality is violated.
Since we observed that the Normal probability plot does not appear to be a straight line, a long tail appears on both ends. Normality is violated; therefore, we need to conduct a Modified-Levene Test to check for constant variance.

MODIFIED LEVENE TEST:

Obs	group	mede
1	1	49.484
2	2	-128.529

Obs	group	meand
1	1	245.321
2	2	282.494

The TTEST Procedure

Variable: d

group	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		50	245.3	135.2	19.1267	0.3411	540.3
2		50	282.5	388.4	54.9261	0.4669	1833.1
Diff (1-2)	Pooled		-37.1732	290.8	58.1611		
Diff (1-2)	Satterthwaite		-37.1732		58.1611		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		245.3	206.9	283.8	135.2	113.0	168.5
2		282.5	172.1	392.9	388.4	324.4	484.0
Diff (1-2)	Pooled	-37.1732	-152.6	78.2455	290.8	255.2	338.1
Diff (1-2)	Satterthwaite	-37.1732	-153.5	79.1382			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	98	-0.64	0.5242
Satterthwaite	Unequal	60.711	-0.64	0.5251

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	49	49	8.25	<.0001

Table 5. Modified Levene Test

The data is divided into 2 equal halves and take α= 0.10. Initially we check if variance is equal or not.
H_0: Variance is equal.
H_1: Variance is not equal.
From table 5, p-value for t test = <0.0001 which is less than α= 0.10
Therefore, we reject H_0 and conclude that variance is not equal. Now we conduct a test to check if variance is constant or not.
H_0: Variance is constant.
H_1: Variance is not constant.
From table 5, p-value for unequal or Satterwaite variances is < 0.0001 which is less than α= 0.10
Therefore, we reject H_0 and thereby conclude Variance is not Constant.

But this contradicts the earlier assumption from the graph that does not show funnel shape. Hence, we can assume that the variance is constant.

<u>OUTLIERS:</u>

Bonferroni Outlier Test:
The test is conducted for determining Y-outliers. The results parameters are generated from SAS and loaded in the Excel file for further calculation. The test uses a t-distribution to test whether the model's largest studentized residual value's outlier status is statistically different from the other observations in the model. In the table below, the t_i value is calculated with the following Studentized deleted residual formula.

$t_i = e_i\ [(n-p-1)/SSE\ (1 - h_{ii}) - e_i^2]^{1/2}$

Test Condition:
$|t_i| > t\ (1- \alpha/2;\ n-p-1)$

From the table below, it is observed that the condition for the outlier test is not satisfied by any of the data points. Hence, there are no outliers in the data.

						DFBETAS						
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	mileage	registered_year	KMs_Driven	t_i	$t(1-\propto/2n;n-p-1)$	$\lvert ti \rvert > t(1-\propto/2n;n-p-1)2$
1	-325.4195	-0.8442	0.0416	1.056	-0.1759	-0.0063	0.0368	0.0063	-0.0332	-0.8442433	2.629	3.4732433
2	-337.0982	-0.8775	0.0475	1.0599	-0.1959	0.0195	0.0656	-0.0196	-0.0569	-0.8775085	2.629	3.5065085
3	-316.2873	-0.8639	0.1351	1.1685	-0.3414	0.2205	0.1286	-0.2206	-0.2622	-0.8639180	2.629	3.4929180
4	192.2858	0.5049	0.0689	1.108	0.1373	-0.0588	-0.0365	0.0588	0.0848	0.5049000	2.629	2.1241000
5	262.1674	0.6879	0.0655	1.0939	0.1821	-0.0639	-0.0744	0.064	0.0926	0.6879272	2.629	1.9410728
6	307.1603	0.8922	0.2348	1.318	0.4942	0.4771	0.0336	-0.4767	-0.4515	0.8922035	2.629	1.7367965
7	324.6754	0.8847	0.1306	1.1607	0.3429	0.2462	-0.1453	-0.2454	-0.2447	0.8846993	2.629	1.7443007
8	438.7041	1.2077	0.1421	1.1437	0.4916	0.4548	-0.004	-0.4544	-0.4197	1.2076357	2.629	1.4213643
9	485.9262	1.2813	0.0632	1.0394	0.3329	0.129	-0.1779	-0.1283	-0.1042	1.2812729	2.629	1.3477271
10	92.0422	0.2376	0.0389	1.0825	0.0478	0.0282	0.0174	-0.0283	-0.0154	0.2376327	2.629	2.3913673
11	405.6315	1.0501	0.0336	1.0304	0.1959	0.0309	0.0564	-0.0312	0.0289	1.0501072	2.629	1.5788928
12	-108.541	-0.2831	0.0583	1.1035	-0.0704	0.0325	0.0409	-0.0326	-0.0373	-0.2831360	2.629	2.9121360
13	60.0915	0.1567	0.0587	1.1067	0.0391	-0.0233	-0.0085	0.0233	0.0295	0.1567399	2.629	2.4722601
14	301.0763	0.7764	0.0312	1.0494	0.1392	0.0179	0.0654	-0.0182	0.0247	0.7764393	2.629	1.8525607
15	289.2011	0.7537	0.0517	1.0737	0.1759	-0.0935	0.0008	0.0933	0.1306	0.7536957	2.629	1.8753043
16	379.2874	1.0047	0.0778	1.0839	0.2917	-0.1085	0.1495	0.1076	0.1893	1.0046742	2.629	1.6243258
17	208.3351	0.5353	0.0274	1.0593	0.0898	0.045	-0.0154	-0.0449	-0.0313	0.5353327	2.629	2.0936673
18	226.0843	0.5917	0.0619	1.0954	0.152	-0.0119	-0.1201	0.0124	0.0101	0.5917239	2.629	2.0372761
19	475.6891	1.2311	0.0288	1.0078	0.212	0.0965	0.0868	-0.0968	-0.0381	1.2310622	2.629	1.3979378
20	444.2571	1.186	0.0884	1.0786	0.3694	-0.1663	-0.274	0.1671	0.1675	1.1860374	2.629	1.4429626
21	178.3281	0.4704	0.0777	1.1202	0.1365	-0.0258	0.0998	0.0253	0.0638	0.4703956	2.629	2.1586044
22	49.8255	0.1281	0.0306	1.0749	0.0227	0.0148	0.0113	-0.0148	-0.0098	0.1280596	2.629	2.5009404
23	215.0554	0.5573	0.0434	1.076	0.1188	-0.0343	0.0605	0.034	0.0668	0.5572733	2.629	2.0717267
24	-419.8999	-1.0811	0.0223	1.0156	-0.1632	0.0678	0.0611	-0.068	-0.0822	-1.0811169	2.629	3.7101169
25	-341.5202	-0.878	0.0234	1.0338	-0.1358	0.0059	-0.076	-0.0056	-0.0408	-0.8779854	2.629	3.5069854
26	-490.8276	-1.2739	0.0333	1.0081	-0.2365	0.0181	0.1814	-0.0188	-0.0065	-1.2739006	2.629	3.9029006
27	-426.0125	-1.0965	0.0214	1.0133	-0.1621	-0.1121	-0.0088	0.112	0.0969	-1.0965420	2.629	3.7255420
28	-134.7613	-0.3458	0.0264	1.0657	-0.057	0.0108	-0.0297	-0.0106	-0.0252	-0.3457980	2.629	2.9747980
29	-146.6625	-0.3755	0.022	1.0599	-0.0563	-0.0044	-0.0327	0.0046	-0.0099	-0.3755313	2.629	3.0045313
30	10.8129	0.0277	0.0263	1.0709	0.0046	0.0013	0.0032	-0.0013	-0.0001	0.0277272	2.629	2.6012728
31	20.6539	0.0527	0.0154	1.059	0.0066	0.0007	0.002	-0.0007	0.0008	0.0526687	2.629	2.5763313
32	-115.8168	-0.3051	0.0763	1.1245	-0.0877	0.0138	0.0794	-0.0142	-0.0042	-0.3050655	2.629	2.9340655
33	199.8406	0.519	0.048	1.083	0.1165	0.0405	0.098	-0.0409	-0.0119	0.5189845	2.629	2.1100155
34	179.5989	0.4591	0.0181	1.0526	0.0623	0.0376	0.0034	-0.0376	-0.0305	0.4591195	2.629	2.1698805
35	312.3538	0.8052	0.0299	1.0461	0.1413	0.0561	0.1022	-0.0565	-0.022	0.8051740	2.629	1.8238260
36	370.3323	0.9556	0.0292	1.0338	0.1657	0.0389	0.1204	-0.0394	0.0036	0.9556050	2.629	1.6733950
37	-173.4907	-0.4438	0.0194	1.0547	-0.0625	0.0145	-0.0296	-0.0144	-0.0272	-0.4437663	2.629	3.0727663
38	-235.9736	-0.6017	0.0117	1.0391	-0.0654	-0.0191	-0.001	0.0191	0.0136	-0.6017547	2.629	3.2307547
39	-292.9939	-0.7554	0.0309	1.0506	-0.1349	-0.0962	0.0317	0.096	0.1022	-0.7553512	2.629	3.3843512
40	-92.5964	-0.2369	0.0211	1.0627	-0.0348	-0.0145	0.0167	0.0144	0.0165	-0.2368795	2.629	2.8658795
41	233.2254	0.598	0.0225	1.0509	0.0907	0.0215	0.0644	-0.0218	-0.0025	0.5980090	2.629	2.0309910
42	203.3365	0.5196	0.0169	1.0487	0.0681	0.0279	0.0366	-0.028	-0.0162	0.5196456	2.629	2.1093544
43	-404.5501	-1.0377	0.0159	1.0129	-0.1318	-0.0015	0.0778	0.0011	0.0158	-1.0377067	2.629	3.6667067
44	-284.391	-0.7396	0.048	1.0705	-0.1661	-0.0295	0.1338	0.0288	0.0589	-0.7396370	2.629	3.3686370
45	-226.2687	-0.5823	0.0299	1.0598	-0.1023	-0.0634	0.0351	0.0631	0.0746	-0.5823235	2.629	3.2113235
46	-190.6165	-0.4864	0.0141	1.0472	-0.0582	0.0098	0.031	-0.01	-0.0051	-0.4863607	2.629	3.1153607
47	-69.4238	-0.1809	0.0568	1.104	-0.0444	-0.0263	-0.0353	0.0264	0.0204	-0.1809071	2.629	2.8099071
48	-187.1863	-0.4777	0.0146	1.0481	-0.0581	0.0155	0.0306	-0.0157	-0.0114	-0.4777087	2.629	3.1067087
49	-191.7987	-0.4926	0.0269	1.0607	-0.0819	-0.0609	0.0007	0.0608	0.0648	-0.4926010	2.629	3.1216010
50	49.1434	0.1268	0.0375	1.0827	0.025	0.0032	0.0213	-0.0033	0.0015	0.1267582	2.629	2.5022418
51	33.1106	0.0843	0.0133	1.0566	0.0098	-0.001	0.0043	0.001	0.0022	0.0843461	2.629	2.5446539
52	-68.8708	-0.1779	0.04	1.0847	-0.0363	-0.0098	0.0266	0.0097	0.0159	-0.1778878	2.629	2.8068878
53	116.2662	0.3	0.0374	1.0791	0.0591	0.0307	0.0458	-0.0309	-0.0225	0.2999926	2.629	2.3290074
54	143.9879	0.3665	0.0101	1.0475	0.0371	-0.0021	-0.0019	0.0021	0.0027	0.3664472	2.629	2.2625528
55	-394.1783	-1.017	0.0276	1.027	-0.1714	0.0597	-0.1063	-0.0593	-0.0809	-1.0169409	2.629	3.6459409
56	-376.7628	-0.9672	0.0189	1.022	-0.1342	0.0905	0.0346	-0.0906	-0.084	-0.9671939	2.629	3.5961939
57	-254.4337	-0.6571	0.0359	1.0621	-0.1267	-0.0534	-0.1	0.0537	0.0404	-0.6571627	2.629	3.2861627
58	-163.7608	-0.4236	0.0414	1.0796	-0.088	0.0286	-0.0626	-0.0283	-0.0417	-0.4236190	2.629	3.0526190
59	-209.4875	-0.5444	0.0491	1.083	-0.1237	-0.0098	0.1026	0.0093	0.0364	-0.5444293	2.629	3.1734293
60	108.8519	0.2777	0.0155	1.0558	0.0349	-0.0189	-0.0128	0.0189	0.0159	0.2777018	2.629	2.3512982
61	-492.317	-1.2707	0.0226	0.9973	-0.193	0.1429	0.035	-0.143	-0.1272	-1.2706993	2.629	3.8996993
62	-368.3907	-0.9436	0.015	1.0199	-0.1166	0.0121	-0.0498	-0.012	-0.0094	-0.9436059	2.629	3.5726059
63	-403.6443	-1.0469	0.0373	1.0346	-0.2061	-0.1432	-0.07	0.1432	0.156	-1.0469326	2.629	3.6759326
64	-332.8449	-0.8638	0.0418	1.0548	-0.1805	-0.0827	-0.1349	0.083	0.0736	-0.8637476	2.629	3.4927476
65	-331.1394	-0.8497	0.0203	1.0326	-0.1222	0.0694	0.0613	-0.0697	-0.0486	-0.8497338	2.629	3.4787338
66	-292.7353	-0.755	0.0317	1.0515	-0.1367	-0.0904	-0.0115	0.0903	0.107	-0.7549941	2.629	3.3839941
67	-148.7802	-0.385	0.0422	1.082	-0.0808	0.0256	-0.0581	-0.0253	-0.0342	-0.3849644	2.629	3.0139644
68	-275.3524	-0.7031	0.013	1.0348	-0.0805	-0.0065	0.0055	0.0064	0.019	-0.7031238	2.629	3.3321238
69	-190.1555	-0.4865	0.0192	1.0527	-0.0681	-0.0305	0.003	0.0303	0.0411	-0.4864444	2.629	3.1154444
70	-128.9954	-0.3296	0.0185	1.0576	-0.0453	-0.0193	0.0021	0.0192	0.0263	-0.3296490	2.629	2.9586490
71	93.9885	0.2429	0.0412	1.0848	0.0503	-0.0303	0.0239	0.0302	0.0342	0.2429517	2.629	2.3860483
72	120.3192	0.3075	0.0189	1.0586	0.0427	0.0151	0.0196	-0.0152	-0.0169	0.3075167	2.629	2.3214833
73	67.0492	0.1717	0.0233	1.0663	0.0265	0.0145	-0.0001	-0.0144	-0.0183	0.1716939	2.629	2.4573061
74	-449.8638	-1.1572	0.0188	1.0048	-0.16	-0.0326	0.013	0.0323	0.0662	-1.1572209	2.629	3.7862209
75	-153.7202	-0.3951	0.0291	1.067	-0.0684	0.0474	0.0329	-0.0475	-0.0344	-0.3950704	2.629	3.0240704
76	-56.1671	-0.1441	0.0276	1.0714	-0.0243	0.016	0.012	-0.016	-0.0112	-0.1441390	2.629	2.7731390
77	14.2537	0.0365	0.0242	1.0686	0.0057	-0.0037	0.001	0.0037	0.0033	0.0365111	2.629	2.5924889
78	-466.2866	-1.2149	0.0422	1.0236	-0.2549	-0.0209	-0.1837	0.0214	0.0238	-1.2148920	2.629	3.8438920
79	-221.0631	-0.5671	0.0239	1.054	-0.0888	0.044	0.0286	-0.0441	-0.0224	-0.5671232	2.629	3.1961232
80	-118.0938	-0.3028	0.0253	1.0657	-0.0487	0.027	-0.0073	-0.027	-0.0207	-0.3028137	2.629	2.9318137
81	-105.4037	-0.2702	0.0249	1.0661	-0.0432	0.0139	-0.0163	-0.0138	-0.0095	-0.2701920	2.629	2.8991920
82	-128.0617	-0.328	0.0231	1.0627	-0.0505	0.0257	-0.0046	-0.0257	-0.0178	-0.3280306	2.629	2.9570306
83	42.2016	0.1091	0.042	1.0879	0.0228	-0.0015	0.0164	0.0014	0.0015	0.1091058	2.629	2.5198942
84	40.2865	0.1033	0.0259	1.07	0.0168	-0.0103	-0.0045	0.0104	0.0067	0.1032896	2.629	2.5257104
85	-295.2347	-0.7597	0.0271	1.0462	-0.1268	0.0322	0.0236	-0.0324	0.0026	-0.7596662	2.629	3.3886662
86	-273.1174	-0.7034	0.0298	1.0527	-0.1233	0.0348	0.0431	-0.035	0.0017	-0.7034305	2.629	3.3324305
87	-172.0145	-0.4416	0.0264	1.0623	-0.0727	0.0115	-0.005	-0.0115	0.0053	-0.4415647	2.629	3.0705647
88	-41.8935	-0.108	0.0366	1.0819	-0.021	0.0046	0.011	-0.0046	0.0021	-0.1080051	2.629	2.7370051
89	-629.9501	-1.6582	0.0492	0.9784	-0.3771	0.0905	0.2079	-0.0917	0.0335	-1.6581795	2.629	4.2871795
90	-505.142	-1.3321	0.0621	1.0325	-0.3428	0.0983	0.231	-0.0995	0.0136	-1.3320732	2.629	3.9610732
91	-234.4714	-0.6064	0.0391	1.0686	-0.1223	-0.0066	0.016	0.0064	0.0421	-0.6064068	2.629	3.2354068
92	-88.9516	-0.2296	0.0382	1.0818	-0.0457	-0.0026	-0.0098	0.0025	0.0125	-0.2295653	2.629	2.8585653
93	135.5931	0.3503	0.0395	1.08	0.071	-0.0021	0.0253	0.0021	-0.0104	0.3503028	2.629	2.2786972
94	180.7964	0.4669	0.0376	1.0736	0.0923	-0.0444	-0.0236	0.0445	0.0187	0.4668571	2.629	2.1621429
95	584.958	1.5247	0.0345	0.9805	0.2883	-0.0909	-0.0434	0.0914	0.0101	1.5246579	2.629	1.1043421
96	543.9095	1.4245	0.0466	1.0051	0.3149	-0.1492	-0.1585	0.1501	0.0529	1.4244679	2.629	1.2045321
97	88.8931	0.23	0.0432	1.0874	0.0489	-0.0111	0.0053	0.0111	-0.0009	0.2300132	2.629	2.3989868
98	543.128	3.9874	0.0499	0.591	0.9137	-0.2495	-0.3423	0.2518	-0.0484	1.4248983	2.629	1.2041017
99	648.25	4.3264	0.0446	0.5316	0.9348	-0.4248	-0.0054	0.4254	0.1938	1.7035644	2.629	0.9254356
100	583.15	4.9964	0.059	0.4358	1.2507	-0.5134	-0.6511	0.517	0.1068	1.5399763	2.629	1.0890237

Table 6. Output Statistics

EXPLORATION OF INTERACTION TERMS:

Table 3 reveals a concerning finding with the maximum Variance Inflation Factor (VIF) reaching 10.65034 (for registered year), surpassing the recommended threshold of 5. This indicates a severe issue of multicollinearity, potentially compromising the stability of coefficient estimates and complicating the interpretation of individual predictor effects. To mitigate this, variables were standardized by centering the mean to 0 and variance to 1. Subsequently, a graph was generated to identify potential interaction terms, aiming to address and reduce the impact of multicollinearity on the model's robustness.

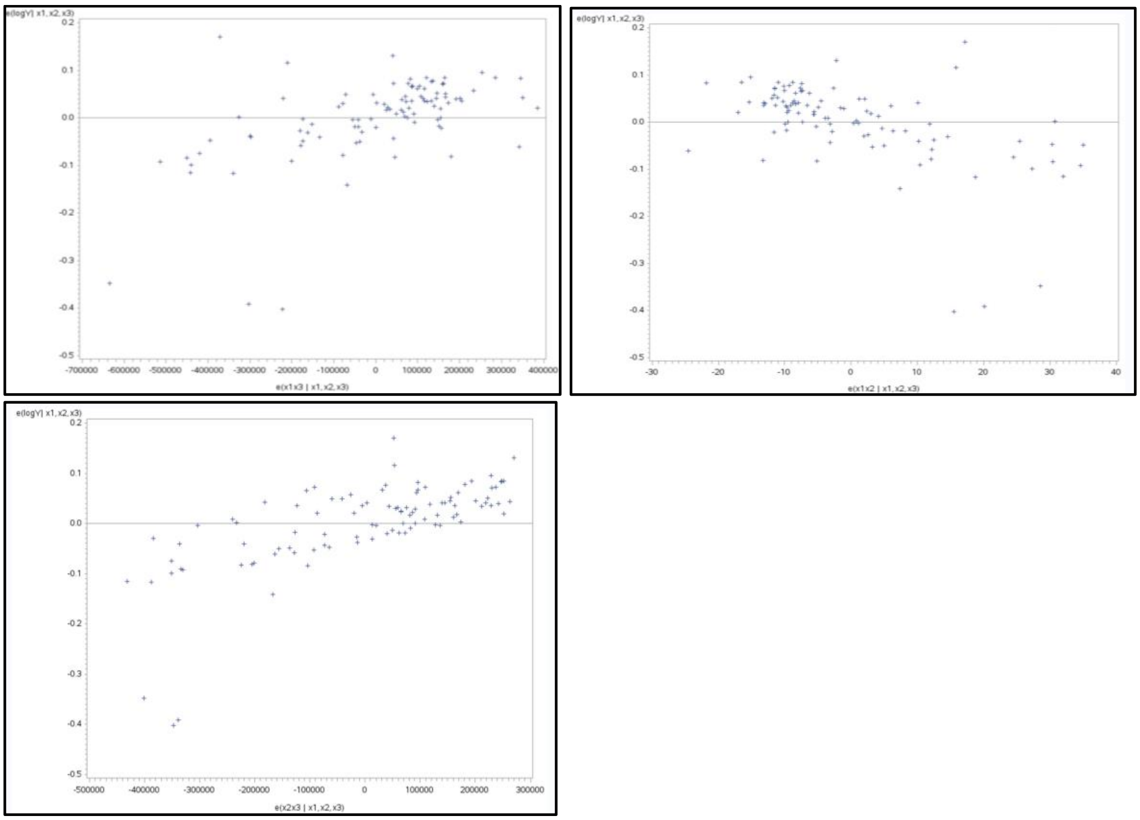

Figure 4. Partial Regression Plots

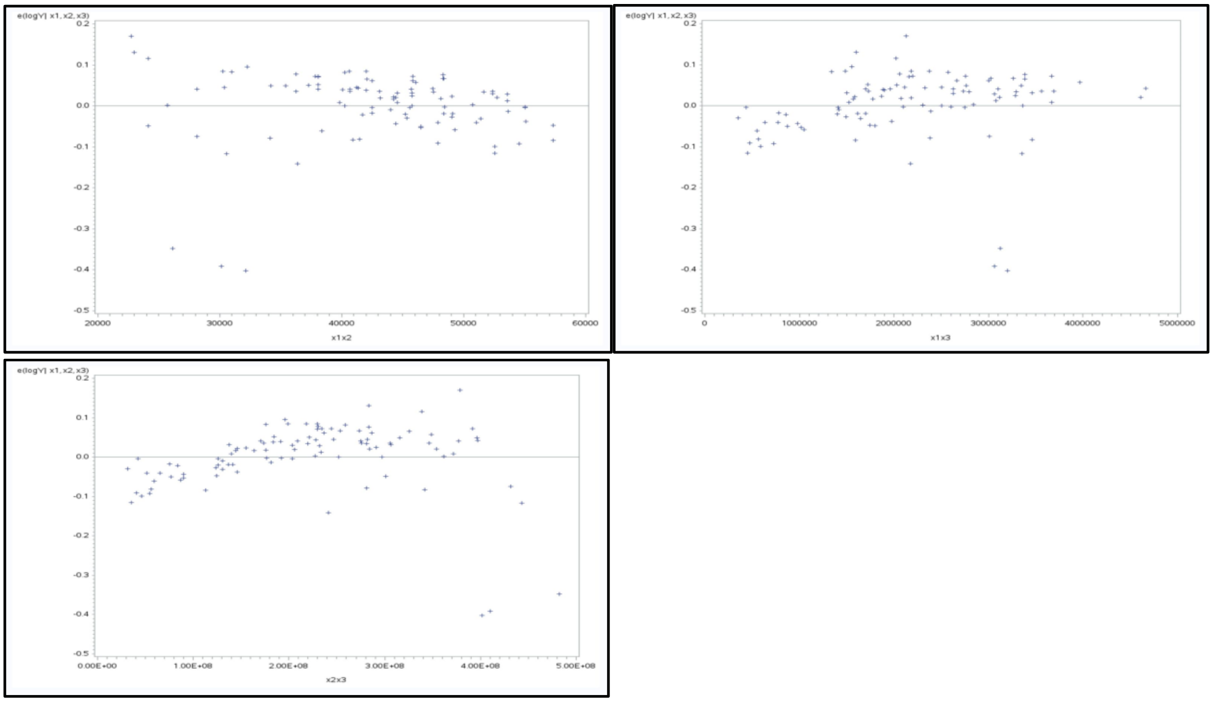

Figure 5. Residuals vs Interactions

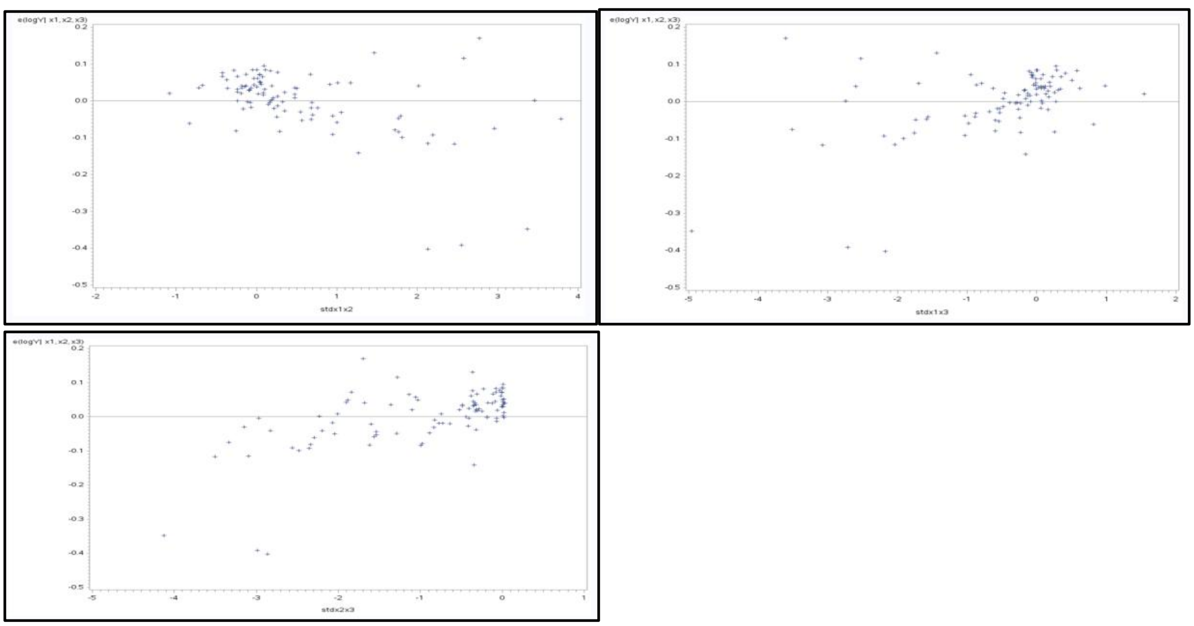

Figure 6. Residuals vs Standardized Interactions

From figure 6, we can see the graph for 3 new standardized interaction terms which are stdx1x2, stdx1x3, stdx2x3.

MODEL SEARCH:

In this sequential modeling approach, we systematically explore models from the smallest to the largest number of predictors, employing criteria such as R^2, Adjusted R^2, C(p), AIC, and SBC. The optimal model is anticipated to exhibit higher R^2 and Adjusted R^2 values, indicative of better explanatory power, along with lower C(p), AIC, and SBC values, suggesting parsimony and reduced complexity. This method ensures a thorough consideration of various criteria to identify the most robust and efficient model.

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
1	0.8798	0.8811	111.0710	-473.3488	-468.13847	registered_year
1	0.8199	0.8218	214.2845	-432.9060	-427.69568	KMs_Driven

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
2	0.9414	0.9426	5.9443	-544.1937	-536.37815	registered_year stdx2x3
2	0.9226	0.9242	38.0287	-516.3531	-508.53755	registered_year stdx1x3

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
3	0.9444	0.9461	1.9132	-548.4179	-537.99722	registered_year stdx1x2 stdx2x3
3	0.9443	0.9460	1.9641	-548.3637	-537.94297	registered_year stdx1x3 stdx2x3

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
4	0.9443	0.9465	3.0650	-547.3252	-534.29937	registered_year KMs_Driven stdx1x3 stdx2x3
4	0.9440	0.9462	3.5765	-546.7771	-533.75125	registered_year stdx1x2 stdx1x3 stdx2x3

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
5	0.9437	0.9466	5.0313	-545.3614	-529.73039	mileage registered_year KMs_Driven stdx1x3 stdx2x3
5	0.9437	0.9466	5.0376	-545.3546	-529.72360	registered_year KMs_Driven stdx1x2 stdx1x3 stdx2x3

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
6	0.9431	0.9466	7.0000	-543.3950	-525.15884	mileage registered_year KMs_Driven stdx1x2 stdx1x3 stdx2x3

Table 7. Best subsets method

Table 8 showcases the optimal models in each subset. Notably, there is a consistent augmentation in both R2 and adjusted R2, reflecting the desired enhancement in explanatory power. Mallow's C(p) values exhibit a pronounced decline as predictors are added, indicating a reduction in bias and assurance that crucial predictors are incorporated. The compelling reduction in Akaike's Information Criterion (AIC) and Schwarz' Bayesian Criterion (SBC) supports our intuition about the 6-predictor model, which outperformed other subsets. The top-performing models emerge as the 6-predictor model (mileage, registered_year, KMs_Driven, stdx1x2, stdx1x3, stdx2x3) and the 5-predictor model (mileage, registered_year, stdx1x3, stdx2x3), solidifying their standing in this method.

STEPWISE SELECTION:

The stepwise selection method systematically constructs regression models by iteratively adding or removing independent variables. This process culminates in the identification of the optimal model. The inclusion or exclusion of predictor variables is contingent on their p-values, with a predefined significance level of $\alpha = 0.10$. If a predictor variable's p-value falls below this threshold, it is added to the model; otherwise, it is removed. This rigorous approach ensures the refinement of the model based on statistical significance, ultimately yielding a parsimonious and robust final regression model.

The REG Procedure
Model: MODEL1
Dependent Variable: logtenY

Number of Observations Read	100
Number of Observations Used	100

Stepwise Selection: Step 1

Variable registered_year Entered: R-Square = 0.8811 and C(p) = 111.0710

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6.25951	6.25951	725.88	<.0001
Error	98	0.84509	0.00862		
Corrected Total	99	7.10460			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-131.82965	5.02730	5.92967	687.63	<.0001
registered_year	0.06721	0.00249	6.25951	725.88	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable stdx2x3 Entered: R-Square = 0.9426 and C(p) = 5.9443

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6.69671	3.34836	796.28	<.0001
Error	97	0.40789	0.00421		
Corrected Total	99	7.10460			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-133.01179	3.51252	6.02992	1433.98	<.0001
registered_year	0.06783	0.00174	6.36724	1514.20	<.0001
stdx2x3	0.06461	0.00634	0.43720	103.97	<.0001

Bounds on condition number: 1.0012, 4.0048

Stepwise Selection: Step 3

Variable stdx1x2 Entered: R-Square = 0.9461 and C(p) = 1.9132

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6.72132	2.24044	561.17	<.0001
Error	96	0.38327	0.00399		
Corrected Total	99	7.10460			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-130.06760	3.62216	5.14802	1289.45	<.0001
registered_year	0.06637	0.00180	5.44284	1363.29	<.0001
stdx1x2	-0.02071	0.00834	0.02461	6.17	0.0148
stdx2x3	0.05282	0.00779	0.18364	46.00	<.0001

Bounds on condition number: 1.6936, 13.224

All variables left in the model are significant at the 0.1000 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	registered_year		registered_year	1	0.8811	0.8811	111.071	725.88	<.0001
2	stdx2x3			2	0.0615	0.9426	5.9443	103.97	<.0001
3	stdx1x2			3	0.0035	0.9461	1.9132	6.17	0.0148

Table 8. Stepwise Selection

From Table 7, our best models for the stepwise regression method are our 5-predictor model (mileage, registered_year, stdx1x3, stdx2x3) and 6-predictor model (mileage, registered_year, KMs_Driven, stdx1x2, stdx1x3, stdx2x3)

BACKWARD ELIMINATION:

Backward elimination, a strategic model selection technique in statistical modeling, systematically prunes less significant independent variables from a regression model. The process initiates with a comprehensive model encompassing all independent variables, progressively discarding one variable at a time based on predefined criteria. This iterative refinement continues until a final model emerges, ensuring the retention of only the most impactful and statistically significant predictors for enhanced model precision.

The REG Procedure
Model: MODEL1
Dependent Variable: logtenY

Number of Observations Read	100
Number of Observations Used	100

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.9466 and C(p) = 7.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	6.72505	1.12084	274.64	<.0001
Error	93	0.37955	0.00408		
Corrected Total	99	7.10460			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-140.05173	13.78739	0.42111	103.18	<.0001
mileage	-0.00040028	0.00206	0.00015343	0.04	0.8467
registered_year	0.07131	0.00682	0.44568	109.21	<.0001
KMs_Driven	3.471669E-7	4.642891E-7	0.00228	0.56	0.4565
stdx1x2	-0.00374	0.02116	0.00012766	0.03	0.8600
stdx1x3	0.01751	0.02022	0.00306	0.75	0.3888
stdx2x3	0.05284	0.00856	0.15562	38.13	<.0001

Bounds on condition number: 15.809, 336.57

Backward Elimination: Step 1

Variable stdx1x2 Removed: R-Square = 0.9466 and C(p) = 5.0313

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6.72492	1.34498	332.99	<.0001
Error	94	0.37967	0.00404		
Corrected Total	99	7.10460			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-141.22788	12.01514	0.55804	138.16	<.0001
mileage	-0.00037814	0.00205	0.00013743	0.03	0.8541
registered_year	0.07189	0.00595	0.58968	145.99	<.0001
KMs_Driven	3.880926E-7	4.004349E-7	0.00379	0.94	0.3349
stdx1x3	0.02075	0.00845	0.02438	6.04	0.0159
stdx2x3	0.05298	0.00848	0.15783	39.08	<.0001

Bounds on condition number: 12.144, 144.22

Backward Elimination: Step 2

Variable mileage Removed: R-Square = 0.9465 and C(p) = 3.0650

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6.72479	1.68120	420.51	<.0001
Error	95	0.37981	0.00400		
Corrected Total	99	7.10460			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-140.55426	11.38835	0.60899	152.32	<.0001
registered_year	0.07155	0.00563	0.64532	161.41	<.0001
KMs_Driven	3.783602E-7	3.949203E-7	0.00367	0.92	0.3405
stdx1x3	0.02038	0.00816	0.02493	6.23	0.0142
stdx2x3	0.05281	0.00838	0.15867	39.69	<.0001

Bounds on condition number: 10.993, 102.64

Backward Elimination: Step 3

Variable KMs_Driven Removed: R-Square = 0.9460 and C(p) = 1.9641

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6.72112	2.24037	560.85	<.0001
Error	96	0.38348	0.00399		
Corrected Total	99	7.10460			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-130.20562	3.60683	5.20571	1303.19	<.0001
registered_year	0.06643	0.00179	5.50232	1377.44	<.0001
stdx1x3	0.02016	0.00815	0.02441	6.11	0.0152
stdx2x3	0.05122	0.00821	0.15535	38.89	<.0001

Bounds on condition number: 1.8599, 14.227

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination								
Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	stdx1x2		5	0.0000	0.9466	5.0313	0.03	0.8600
2	mileage	mileage	4	0.0000	0.9465	3.0650	0.03	0.8541
3	KMs_Driven	KMs_Driven	3	0.0005	0.9460	1.9641	0.92	0.3405

Table 9. Backward Elimination

From table 7, our best models for the backward elimination method are our 5-predictor model and 6-predictor model because the best model is the one which has all the predictor variables lower than the significance level. The best model consists of 5-predictor model (mileage, registered_year, stdx1x3, stdx2x3) and 6-predictor model (mileage, registered_year, KMs_Driven, stdx1x2, stdx1x3, stdx2x3) with an R square value of 0.9466.

From a comparative analysis of the results of the three methods, our Model 1 will be our 6-predictor model (mileage, registered_year, KMs_Driven, stdx1x2, stdx1x3, stdx2x3). Our Model 2 will be our (mileage, registered_year, stdx1x3, stdx2x3).

Model 1 and model 2 will then be investigated for multicollinearity and defined with fitted equations.

The CORR Procedure

7 Variables:	logtenY mileage registered_year KMs_Driven stdx1x2 stdx1x3 stdx2x3

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
logtenY	100	3.61671	0.26789	361.67082	2.77815	4.03342	
mileage	100	21.18220	4.06810	2118	11.30000	28.40000	mileage
registered_year	100	2015	3.74127	201527	2009	2022	registered_year
KMs_Driven	100	109012	53280	10901231	15683	240250	KMs_Driven
stdx1x2	100	0.54899	0.99086	54.89874	-1.07808	3.78272	
stdx1x3	100	-0.51391	1.06231	-51.39106	-4.95419	1.54810	
stdx2x3	100	-0.93965	1.02920	-93.96519	-4.12803	0.03350	

Pearson Correlation Coefficients, N = 100
Prob > |r| under H0: Rho=0

	logtenY	mileage	registered_year	KMs_Driven	stdx1x2	stdx1x3	stdx2x3
logtenY	1.00000	0.59173 <.0001	0.93864 <.0001	-0.90651 <.0001	-0.42254 <.0001	0.41007 <.0001	0.21535 0.0314
mileage mileage	0.59173 <.0001	1.00000	0.55453 <.0001	-0.51910 <.0001	-0.44117 <.0001	0.41746 <.0001	0.22827 0.0224
registered_year registered_year	0.93864 <.0001	0.55453 <.0001	1.00000	-0.94914 <.0001	-0.24582 0.0137	0.22116 0.0270	-0.03470 0.7318
KMs_Driven KMs_Driven	-0.90651 <.0001	-0.51910 <.0001	-0.94914 <.0001	1.00000	0.23710 0.0175	-0.27314 0.0060	-0.05555 0.5830
stdx1x2	-0.42254 <.0001	-0.44117 <.0001	-0.24582 0.0137	0.23710 0.0175	1.00000	-0.93370 <.0001	-0.58198 <.0001
stdx1x3	0.41007 <.0001	0.41746 <.0001	0.22116 0.0270	-0.27314 0.0060	-0.93370 <.0001	1.00000	0.63493 <.0001
stdx2x3	0.21535 0.0314	0.22827 0.0224	-0.03470 0.7318	-0.05555 0.5830	-0.58198 <.0001	0.63493 <.0001	1.00000

The REG Procedure
Model: MODEL1
Dependent Variable: logtenY

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	6.72505	1.12084	274.64	<.0001
Error	93	0.37955	0.00408		
Corrected Total	99	7.10460			

Root MSE	0.06388	R-Square	0.9466
Dependent Mean	3.61671	Adj R-Sq	0.9431
Coeff Var	1.76635		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-140.05173	13.78739	-10.16	<.0001	0
mileage	mileage	1	-0.00040028	0.00206	-0.19	0.8467	1.71102
registered_year	registered_year	1	0.07131	0.00682	10.45	<.0001	15.80856
KMs_Driven	KMs_Driven	1	3.471669E-7	4.642891E-7	0.75	0.4565	14.84431
stdx1x2		1	-0.00374	0.02116	-0.18	0.8600	10.66247
stdx1x3		1	0.01751	0.02022	0.87	0.3888	11.18811
stdx2x3		1	0.05284	0.00856	6.18	<.0001	1.88121

Table 10 Pearson correlation and parameter estimates for Model 1

Model 1 has some VIF values greater than 5 and the average VIF value is 9.34928 also greater than 5 and it is highly correlated, and it has a serious case of multicollinearity. From table 10, the fitted regression model for Model 1 is:

Resale_value_in_Dollars = -140.05173 - 0.00040028(mileage) + 0.07131(registered_year) + 3.471669E-7(KMs_Driven) - 0.00374(stdx1x2) + 0.01751(stdx1x3) + 0.05284(stdx2x3)

The CORR Procedure

6 Variables:	logtenY mileage registered_year KMs_Driven stdx1x3 stdx2x3

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
logtenY	100	3.61671	0.26789	361.67082	2.77815	4.03342	
mileage	100	21.18220	4.06810	2118	11.30000	28.40000	mileage
registered_year	100	2015	3.74127	201527	2009	2022	registered_year
KMs_Driven	100	109012	53280	10901231	15683	240250	KMs_Driven
stdx1x3	100	-0.51391	1.06231	-51.39106	-4.95419	1.54810	
stdx2x3	100	-0.93965	1.02920	-93.96519	-4.12803	0.03350	

Pearson Correlation Coefficients, N = 100
Prob > |r| under H0: Rho=0

	logtenY	mileage	registered_year	KMs_Driven	stdx1x3	stdx2x3
logtenY	1.00000	0.59173 <.0001	0.93864 <.0001	-0.90651 <.0001	0.41007 <.0001	0.21535 0.0314
mileage mileage	0.59173 <.0001	1.00000	0.55453 <.0001	-0.51910 <.0001	0.41746 <.0001	0.22827 0.0224
registered_year registered_year	0.93864 <.0001	0.55453 <.0001	1.00000	-0.94914 <.0001	0.22116 0.0270	-0.03470 0.7318
KMs_Driven KMs_Driven	-0.90651 <.0001	-0.51910 <.0001	-0.94914 <.0001	1.00000	-0.27314 0.0060	-0.05555 0.5830
stdx1x3	0.41007 <.0001	0.41746 <.0001	0.22116 0.0270	-0.27314 0.0060	1.00000	0.63493 <.0001
stdx2x3	0.21535 0.0314	0.22827 0.0224	-0.03470 0.7318	-0.05555 0.5830	0.63493 <.0001	1.00000

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6.72492	1.34498	332.99	<.0001
Error	94	0.37967	0.00404		
Corrected Total	99	7.10460			

Root MSE	0.06355	R-Square	0.9466
Dependent Mean	3.61671	Adj R-Sq	0.9437
Coeff Var	1.75723		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-141.22788	12.01514	-11.75	<.0001	0
mileage	mileage	1	-0.00037814	0.00205	-0.18	0.8541	1.70473
registered_year	registered_year	1	0.07189	0.00595	12.08	<.0001	12.14368
KMs_Driven	KMs_Driven	1	3.880926E-7	4.004349E-7	0.97	0.3349	11.15696
stdx1x3		1	0.02075	0.00845	2.46	0.0159	1.97327
stdx2x3		1	0.05298	0.00848	6.25	<.0001	1.86484

Table 11. Pearson correlation and parameter estimates for Model 2

Model 2 has some VIF values greater than 5 and the average VIF value is 5.395728. It is not highly correlated, and it does not have a serious case of multicollinearity. From table 11, the fitted regression model for Model 2 is:

Resale_value_in_Dollars = -141.22788 - 0.00037814(mileage) + 0.07189(registered_year) + 3.880926E-7(KMs_Driven) + 0.02075(stdx1x3) + 0.05298(stdx2x3)

Using a comparative analysis, we found out that model 1 and original model are highly correlated and have a VIF value as 9.34928 and 7.397. So, we cannot consider these as our best models. Therefore, we have selected model 2 (5-predictor model) which has a very low average VIF value 5.395728 as our best model.

MODEL SELECTION:

	Model 1	**Model 2**	**Original**
R^2	0.9466	0.9466	0.9733
R^2 Adjusted	0.9431	0.9437	0.9724
C(p)	7.0000	5.0313	2.8575
AIC	-543.3950	-545.3614	1197.6862
SBC	-525.15884	-529.73039	1202.89650
VIF MAX	15.80856	12.14368	10.65034
VIF_{avg}	9.34928	5.395728	7.397

Table 12. Model Selection Criteria

Table 13 shows various criteria for two models, indicating diagnostic outcomes and model validations. While model 1 had three extra predictors and model 2 included two additional predictors, our preference leaned towards model 2. This choice was based on its VIF value, which was closest to 5, indicating it as the most suitable among the three models tested. Despite similar metrics for model 1 and 2, we opted against selecting model 1 or the original model as the best fit due to considerably high VIF values.

Bonferroni Joint CI's

Bonferroni Joint CI's give 100(1-α) % confidence in our predictor parameters.

Here, α=0.1,

For calculating two-sided Confidence Interval= $b_k \pm B.s\{b_k\}$

Here, B=t (1-α/2g; n-p). Where g = predictors = 3, p = parameters = 4, n=99,

Using the t-table B= t (1- 0.10/ 2∗3; 96) = 2.082

The CI's values for all the βi's are given below.

C.I. of β_1 = -0.00037 ± (2.082*0.0025) = {0.004835, 0.00575}

C.I. of β_2 = 0.0718 ± (2.082*0.00595) = {-0.08427, 0.0595}

C.I. of β_3 = $3.88(10^{-7}) \pm (2.082*4.456(10^{-7})) = \{(0.122*10^{-7}), 4.456(10^{-7})\}$

Confidence Interval

Confidence Band and Prediction Interval for a specific X_h

We randomly select values for the dataset for predictors Mileage as 17, KMs Driven as 140000, Make Year as 2009, Fifty as 57.

X_h^T = [1 17 140000 2009]

Confidence interval C.I = $\hat{Y} \pm t\,(1 - \alpha/2;\, n - p)$

$s\{\hat{Y}_h\}$ *R*esale Price ̂= -141.22788 - 0.00037814(17) + 0.07189(2009) + 3.880926E-7(140000)

Resale Price = 1481.124

From ANOVA table,

MSE = 12.015

Using Excel,

$X_h^T (X^TX)^{-1}X_h$ = 0.0014754552

Hence, $S\{\hat{Y}_h\}$ = 14.854

C.I = 1481.124 ± t (0.90, 96) * 14.854 = 1481.124 ± 2.082 * 14.854 = {1405.197, 1512.05}

Conclusion: We are 90% confident that the mean resale price lies between 1405.197 USD and 1512.05 USD.

Confidence Band

Confidence band is given by $\hat{Y}_h \pm W^*S\{\hat{Y}_h\}$

$W^2 = p * F(1-\alpha, p, n - p)$ where $F(1-\alpha, p, n - p) = F(0.90, 4, 96)$ =1.9216

W^2 = 4*1.9216 = 7.6864 and W = 2.7724

Confidence Band = 1481.124 ± 2.7724*14.854

(1439.94, 1522.30)

Conclusion: With 90% confidence we can say that the entire regression line falls between the confidence band of 1439.94 USD and 1522.30 USD.

Prediction Interval

The prediction interval is given by the formula. $X_h = \hat{Y}_h \pm t\,(1-\alpha/2, n - p)\, S\{pred\}$

$S\{pred\} = \sqrt{(MSE + S\{\hat{Y}h\}^2)}$

Hence, S{pred} = 15.253

P.I. = 1481.124 ± t (1- 0.1/2 ,96) * 15.253 = 1481.124 ± 1.661* 15.253

P.I. = (1455.7888, 1506.4592)

Conclusion: With 90% confidence we can say that the future resale price will lie between 1455.78880 USD and 1506.4592 USD

FINAL MULTIPLE REGRESSION MODEL:

From the chosen final model, we can see predictors of the model are mileage, registered year, kilometers driven, stdx1x3 and stdx2x3. The response is the main input (y).

Our final fitted regression model is as follows.

Resale_value_in_Dollars = -141.22788 - 0.00037814(mileage) + 0.07189(registered_year) + 3.880926E-7(KMs_Driven) + 0.02075(stdx1x3) + 0.05298(stdx2x3)

From the model, we can infer that in general a unit increase in any of the predictors will cause an increase in the main input. From Table 2. The p-value of <0.0001 gives us confidence in the reliability and significance of our model. All the parameters in the model are also significant at a level of 0.10 with all p-values less than 0.001. The SSR and SSE are 6.72492 and 0.37967 with degrees of freedom of 5 and 94 respectively. These represent the variability of the model and the error respectively. The R^2 for our final model is found in Table 2 as 0.9466. This means that the percentage of variability explained by the model is 94.66%. The adjusted R^2 a(R^2a) is equivalent to R^2. This means that all the variables in the model are contributing to the accuracy and explanation of our model. Moreover, the fact that the average VIF is around 5, the model is not suffering serious multicollinearity making it a usable model. The joint confidence interval for β_1, β_2, β_3, β_4 and β_5 were calculated with a 90% family confidence level.

FINAL DISCUSSION:

In this project, we used a car resale dataset from Kaggle to predict the resale price of cars from the year 2023. We considered 3 predictors: mileage, registered year and kilometers driven. Our Multiple Linear Regression model achieved a high R-Square of 94.66%, indicating good predictive performance.

However, when we tested our model assumptions, we found that the residuals vs. normal scores plot did not show a perfectly straight line, suggesting a violation of normality. Despite this, we proceeded with the model since the residuals vs. response variable plot did not exhibit a funnel shape, indicating constant variance.

We also checked for multicollinearity and found high correlation values between predictor variables. However, the average Variance Inflation Factor (VIF) was 5, suggesting no serious multicollinearity issues in model 2. However, there was evidence of serious multicollinearity in model 1 and the original model.

Furthermore,we standardized variables to address multicollinearity, derived interaction terms, and selected two for inclusion in the model based on their linear association with residuals.

For model selection, we used various methods and ended up with two models. However, one of them had serious multicollinearity issues. In the end, we chose the second model with five predictors as it struck a balance between model complexity, multicollinearity, and still had a high R-Square of 94.66%.

The final model equation is Resale_value_in_Dollars = -141.22788 - 0.00037814(mileage) + 0.07189(registered_year) + 3.880926E-7(KMs_Driven) + 0.02075(stdx1x3) + 0.05298(stdx2x3)

In summary, we believe that the variables “mileage”, "registered year"and "kilometers driven" are excellent choices for predicting a car resale value. This analysis can be valuable for making strategic decisions while purchasing a car.