Beemer Barry Analysis

Team 10

Executive Summary

Beemer Barry runs a used BMW dealership and has recently had trouble turning a profit. In the used car business, predicting what customers value is essential in determining which cars to keep on the lot and how to price them. Therefore, Barry is hoping to use past auction and sales data to inform his future decisions for his business. Moreover, Barry is going to attend the Beaumont Auction in two days. There are five vehicle candidates with detailed specifications to be revealed in the auction and Barry strives to set a bidding strategy to maximize his expected profits. With a budget of $50,000 and a maximum capacity of four cars, Barry wants to know what the optimal bid percentage level for him would be when considering the probability of winning the auction items.

There are several factors Barry must consider when he predicts the resale value of cars and sets the bidding strategy. As the used-car market is more insulated than the new-car market in the US, retailers should determine the resale value based on specifications of cars in demand. They also need to analyze the relationship between specifications under different series of cars because of changing consumer preferences. Furthermore, the mutual relationship between variables should be considered in analyses to mitigate the projection risks.

After conducting regression, simulation, and optimization analyses, Barry can decide on what the projected resale value of each car would be, what the possible percentages of winning for each bid percentage, and which auction items he should bid on to maximize expected profits. From our analysis, it was determined that Barry should bid on all four of the 3 Series cars at a 90% bid level to maximize his profits.

Modelling Approach

To approach Barry's problem, analyzed the data in 5 major steps: Data visualization & cleaning, correlation & regression, winning bid probabilities & expected profits, profit simulation and lastly, bid optimization.

Data visualization involved graphing relevant variables to be used in regression to observe linear relationships and trends. If any variables were deemed non-linear or do not have a relationship to resale value, they were transformed into linear relationships or excluded from the analysis. Additionally, before running a correlation or regression, any categorical variables were converted into dummy variables, and missing data must be identified. Correlation is an essential next step to ensure all the remaining data is suitable and there is no evidence of collinearity. Running a regression enables us to predict resale value of cars for the auction using relevant variables based on historical winning bid data. Using data for past winning bids, we can determine the probabilities of winning at each bid percentage and the expected profits. With expected profits we can determine the average outcome of a bid, however we must simulate profits with probabilities to get a full picture of the possibilities. Next, we determined which cars Barry should bid on to maximize his profits using optimization, keeping in mind the budget and limit of four cars.

Through these models we will be able to report back to Barry and be confident in a recommendation for his used car business.

Analysis

Raw Data Analysis

As Barry noted, his data was not completely suitable for statistical analysis, so our first step was to work through the data and prepare it for regression. We created dummy variables for fuel type, color, and car type and converted Boolean variables into binary variables. The registration date was important to us, as we believed the age of a car will strongly impact its price, so we converted registration date to age in years, noticing this variable was exponentially related to sales price (exhibit 1). To remedy this issue, the log of the original value was taken, providing a linear variable for regression (exhibit 2). Finally, we found that resale values for the 1, 3, and 5 Series did not follow similar trends, therefore we created three separate data sets using these car models.

Regression Analysis

Once the data was prepared, three individual regression analyses were run (one for each series) and all were found significant. Although regression was conducted for the 1 Series model of cars, the model for that set of cars was not relevant in determining prices for the cars at the auction as all five cars at the auction were 3 or 5 Series cars.

As the auction only contained 3 and 5 Series cars, the 3 and 5 Series regression models needed to have predictive power in addition to being a tool for analyzing relationships between variables and the price of cars. Correlation analysis was first conducted on 3 Series car data which revealed that there was high collinearity between “mileage” and “Ln of age” variables, “engine_power” and “sport series” variables, and “engine_power” and “sunroof” variables (exhibit 3). A regression model was built after removing “Ln of age”, “sport series”, and “sunroof” variables as their high correlation with “mileage” and “engine_power” made them redundant. This regression had an R Square value of 0.56 meaning that 56% of the variance in price could be explained by the variables contained in this model (exhibit 4). Unfortunately, plotting the residuals revealed that there was still some collinearity present in the model which meant that although the model could be used for its predictive power, some variables would need to be removed in order to determine the effect of variables on price. By removing “mileage” and “engine_power” the residuals plot revealed drastically reduced collinearity (exhibit 5) at the cost of a lower R squared of the model of just 0.20. As a result, this model was strictly used only to observe the relationship between variables and the price of the car. It was discerned that 3 Series cars which used diesel, had heated seats, were sedans, and painted white were valued higher than cars with other characteristics which was consistent with the findings for 1 Series cars (exhibit 6).

Correlation analysis for 5 Series cars revealed similar insights to 3 Series cars (exhibit 7). The same variables as 3 Series showed high collinearity and as a result, “Ln of age”, “sport series”, and “sunroof” were removed to reduce the collinearity in the model. Removing the aforementioned variables led to a robust regression model with high predictive capacity as evidenced by its R-square value of 0.58. The residual plot also did not reveal any signs of collinearity (exhibit 8). The model revealed that higher mileage had a slightly negative impact on the price of 5 Series cars; higher engine power, diesel fueled cars, white colored cars, and cars with heated seats also tended to be sold at higher prices. A high p-value for hatchbacks made it difficult to discern whether sedans and hatchbacks commanded a premium in the resale market, but estate type cars were generally sold at cheaper prices than these two (exhibit 9).

With regression analyses prepared, resale prices for the upcoming auction could be predicted using the coefficients and vehicle data provided. The biggest indicator of the price of cars wound up being the engine power of the car. The resale price predictions for the five vehicles were: $6,555 for Car #1, $13,035 for Car #2, $10,888 for Car #3, $14,566 for Car #4, and $11,278 for Car #5. A complete breakdown of price based on variable coefficients is as follows:

5 Series Car:

Varaible	*Predicted Price for Vehicle #1*
mileage	(8,934)
engine_power	10,213
diesel	6,064
Black	-
Blue	-
Brown	-
Grey	(2,444)
Silver	-
Estate	(1,694)
Hatchback	-
Heated seats	3,351
	$ 6,555

3 Series Cars:

Variable	*Predicted Price for Vehicle #2*	*Predicted Price for Vehicle #3*	*Predicted Price for Vehicle #4*	*Predicted Price for Vehicle #5*
mileage	(6,148)	(5,079)	(4,485)	(4,229)
engine_power	13,496	9,997	13,496	10,497
diesel	4,648	4,648	4,648	4,648
Black	-	-	-	(677)
Blue	-	(1,569)	-	-
Brown	-	-	-	-
Grey	-	-	(960)	-
Silver	-	-	-	-
Estate	(1,852)	-	-	(1,852)
Hatchback	-	-	(1,024)	-
Heated seats	2,890	2,890	2,890	2,890
	$ 13,035	**$ 10,888**	**$ 14,566**	**$ 11,278**

Profit Simulation Analysis

Resale price predictions in hand, the bid amount maximizing profits can be determined. We can estimate the probability of winning based on different bid percentages of past winning bids. As seen in Figure 1, probabilities of winning based on bid percentage are approximately normally distributed, with bids at 90% of resale value being most likely. Using this knowledge, expected profits at varying bid percentages can be calculated using the NORM.DIST function to determine the probability of winning and corresponding expected profits through multiplication of said probability and the potential profit (predicted resale value – bid amount). Using the predicted resale value for auction car #1, we employ the method described above, presented in Figure 2. As observed for car #1, if Barry wants the highest probability of winning while avoiding losing money, the highest he should bid is at 100%. Since there is a tradeoff between expected profit and bid percentage, the higher Barry bids, the more likely he is to win, however he will cut into his profits. Since Barry is primarily concerned with maximizing profits, he should bid at 90% of resale value, noting that he will have only 58% chance of winning. This same trend applies for each of the cars available at the auction. Expected profits are valuable to gain initial insights to our data, however they report on averages and do not show us all possibilities in profit. For this reason, simulating Barry's potential profits is our next step.

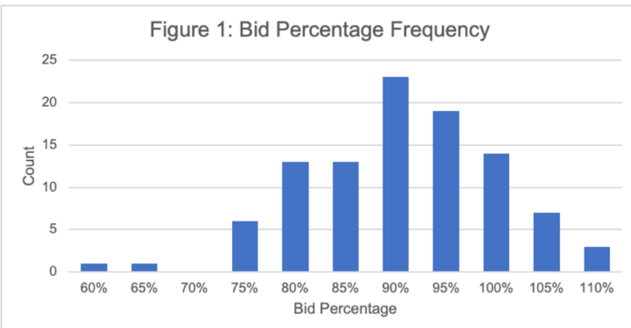


Figure 2: Predict Expected Profit				
Predicted Resale:	**$ 6,555.38**			
Bid percentage	**Bid amount**	**Profit (value saved)**	**Prob of winning**	**Expected profit**
70%	$ 4,588.76	$ 1,966.61	0.026229205	$ 51.58
75%	$ 4,916.53	$ 1,638.84	0.080218549	$ 131.47
80%	$ 5,244.30	$ 1,311.08	0.192737312	$ 252.69
85%	$ 5,572.07	$ 983.31	0.369905486	$ 363.73
90%	$ 5,899.84	$ 655.54	0.580684923	$ 380.66
95%	$ 6,227.61	$ 327.77	0.770166609	$ 252.44
100%	$ 6,555.38	$ -	0.898871094	$ -
105%	$ 6,883.14	$ (327.77)	0.964921445	$ (316.27)
110%	$ 7,210.91	$ (655.54)	0.990528344	$ (649.33)

Figure 3: Simulating Profits at 95% Bid						
Bid # / car	Resale Value	Bid Amt.	Bid percentage	Prob. of winning	Simulate Win	**Profit**
1	$ 6,555.38	$ 6,227.61	95%	77%	1	$ 327.77
2	$ 13,034.97	$ 12,383.23	95%	77%	1	$ 651.75
3	$ 10,887.94	$ 10,343.54	95%	77%	1	$ 544.40
4	$ 14,565.52	$ 13,837.24	95%	77%	0	$ -
5	$ 11,278.10	$ 10,714.19	95%	77%	1	$ 563.90

We decided to simulate profits for both the 90% and 95% bid percentages, as these are Barry's best options for his goals. To start, we created a table (figures 3 and 4) with all the information required. Most notable in this step is the simulation of probability of winning, which was done using the RAND() function. If RAND() returned a value less than the probability of winning at the specified bid percentage, 1 is returned implying Barry won the bid, and if not, 0 is returned, meaning Barry lost the bid. The expected profits are then a product of this binary variable and the difference between resale value and bid amount.

Figure 4: Simulating Profits at 90% Bid						
Bid # / car	Resale Value	Bid Amt.	Bid percentage	Prob of winning	Simulate win	**Profit**
1	$ 6,555.38	$ 5,899.84	90%	58%	0	$ -
2	$ 13,034.97	$ 11,731.48	90%	58%	1	$ 1,303.50
3	$ 10,887.94	$ 9,799.14	90%	58%	1	$ 1,088.79
4	$ 14,565.52	$ 13,108.96	90%	58%	1	$ 1,456.55
5	$ 11,278.10	$ 10,150.29	90%	58%	1	$ 1,127.81

Figure 5: Profit Simulation Summary Statistics					
95%	Car 1	Car 2	Car 3	Car 4	Car 5
Mean	$257.95	$490.12	$421.36	$557.86	$432.52
Min	$ -	$ -	$ -	$ -	$ -
Max	$327.77	$651.75	$544.40	$728.28	$563.90

Figure 6: Profit Simulation Summary Statistics					
90%	Car 1	Car 2	Car 3	Car 4	Car 5
Mean	$367.10	$ 766.46	$ 648.92	$ 838.97	$ 645.11
Min	$ -	$ -	$ -	$ -	$ -
Max	$655.54	$1,303.50	$1,088.79	$1,456.55	$1,127.81

Simulating the probability of winning returns the expected value of profit. In order to get a picture of the full scope of Barry's profit possibilities, we next simulate these profits using data tables. Using the profits in figures 3 and 4, we created data tables using 1000 trials for each car. The summary statistics from each simulation are noted in Figures 5 and 6 for the 95% and 90% bid percentages. One important takeaway from this simulation is that although Barry has less chance of winning with bids at 90%, his profits are greater on average. There is also a possibility for Barry to lose out on all his bids for all cars at both bid percentages. Using the COUNTIF function to count outcomes where profit equals zero and dividing this number by the total number of trials (1000) reveals that there is a 22%-24% chance Barry will not win any of the cars bidding at 95%. At the 90% level we see this range rise to 41%-44%. This analysis illuminates the tradeoff between the probability of winning and potential profits which Barry should consider in his plans to bid.

Bid Optimization

Lastly, we used the predicted resale prices from our regression for the 5 vehicles, the predicted mean profits from simulation analysis, and Barry's constraints to get the optimized bidding portfolio of cars. Two scenarios were considered, one with a higher probability of winning (95% bid), and another with the maximized profit (90% bid). Solver was used by setting the objective to maximize total profits and adding three constraints: total bidding amount within $50,000 budget, the total number of vehicles within 4, and binary decision variables. Solver identified which cars Barry should bid on based on these constraints.

Recommendations and Insights

The decision Barry must make is how much expected profit he is willing to give up in order to ensure he wins his bids. In order to maximize profits, Beemer Barry should bid at 90% of the resale value for cars number two, three, four, and five. He should not bid for car number one. Doing this, Beemer has a 58% chance of winning each of his four bids with an 11% chance of winning all four. His total bid amount would be approximately $45,000 and he can expect to earn around $700 in profit for each successful bid. Another option for Barry is to be more risk-averse and maximize his chances of securing four cars. He can achieve this by bidding at 95% of the resale value which caps his total bid at approximately $47,000 for the same models (cars #2-#5). In this scenario, his profit per car would be lower at around $480 but he would have a 77% chance of securing each car and a 35% chance of winning all four bids. Overall, we believe Barry is more interested in maximizing his profits, therefore we recommend that he bid at 90% for the four cars mentioned above.

Solution Strengths, Weaknesses and Mitigation

A strength of our analysis is that it conducts regression analysis on variables both with and without collinearity. Looking at the regression analyses without collinearity lessens the impact from highly correlated variables and helps us to interpret the relationship between variables better. On the other hand, the regression model with collinearity has more predictive power, which enables us to make more precise predictions (Saslow, 2018).

The first weakness of our model can be attributed to our decision to apply logarithms to the original values for age of cars. There are underlying risks to this approach. After taking the logarithm, the resulting curve minimizes the sum of squares of the logarithms of the errors, and therefore makes the fit tighter for smaller values. A possible mitigating solution is to re-parameterize variables and to run a non-linear regression model in R (Harder & Rodgers, 2020).

Another weakness we have identified with the optimization model is the assumption Barry will bid the same percentage for all cars. This may not be the case, and it may be more desirable to bid at different levels for different cars. We could have further explored this option through optimizing bid percentage levels and probabilities before optimization of which cars to bid on.

Appendix

Exhibit 1: Plotting data of Registration date vs Resale Value reveals exponential relationship

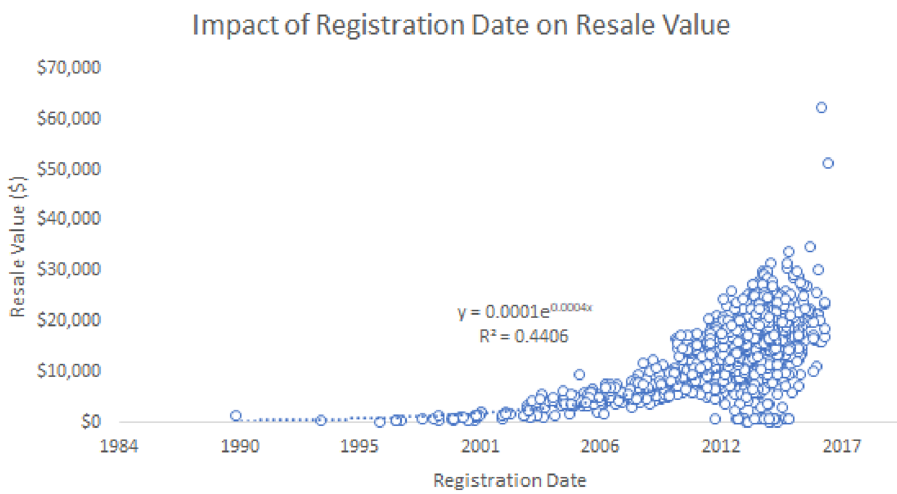


Exhibit 2: Ln of Age of Car was done using Registration Date and plotted against Resale Value to convert the exponential relationship into a linear one

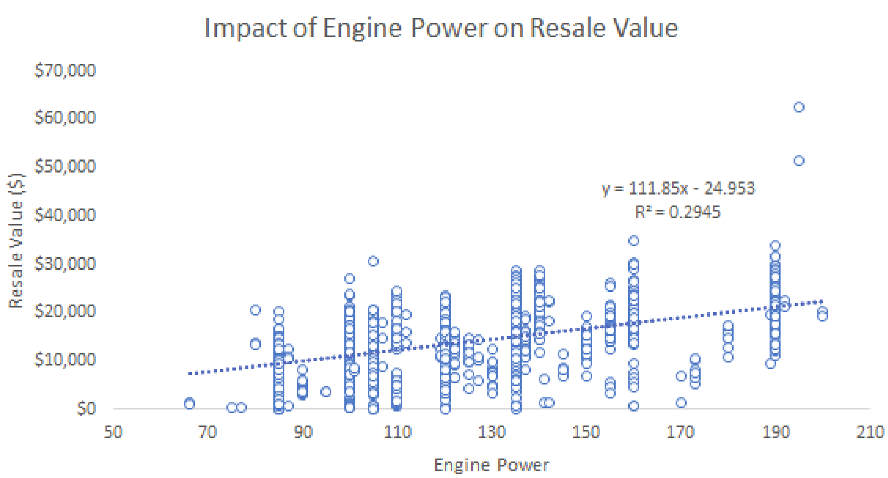


Exhibit 3: Correlation table for 3 Series cars

	mileage	engine_power	Ln of age	diesel	Black	Blue	Brown	Grey	Silver	Estate	Hatchback	Heated seats	Sunroof
mileage													
engine_power	0.0047												
Ln of age	0.4400	-0.1003											
diesel	-0.0213	-0.0062	-0.2020										
Black	0.0252	0.0647	-0.0353	-0.0170									
Blue	0.0092	0.0304	0.0460	0.0022	-0.3268								
Brown	-0.0842	-0.0784	-0.0544	0.0207	-0.2158	-0.1333							
Grey	0.0370	0.0052	0.0841	0.0171	-0.4255	-0.2627	-0.1735						
Silver	0.0325	-0.0192	0.0741	-0.0094	-0.1782	-0.1100	-0.0726	-0.1432					
Estate	0.0750	0.1796	-0.1121	0.1035	0.0140	0.0211	-0.0524	0.0048	0.0017				
Hatchback	-0.0812	-0.3603	0.0102	-0.0508	-0.0119	-0.0014	-0.0231	-0.0181	-0.0233	-0.4449			
Heated seats	-0.0652	0.2102	-0.2431	0.0703	0.0360	-0.0095	0.0305	-0.0016	-0.0252	0.1210	-0.1507		
Sunroof	-0.0507	0.3934	-0.1176	0.0472	-0.0222	0.0297	-0.0352	-0.0268	-0.0013	0.0630	-0.0845	0.0917	
Sport series	-0.0189	0.4728	-0.2016	0.0205	0.0513	-0.0234	-0.0538	-0.0036	-0.0100	0.1460	-0.2238	0.1983	0.2125

Exhibit 4: Regression analysis for 3 Series cars

SUMMARY OUTPUT

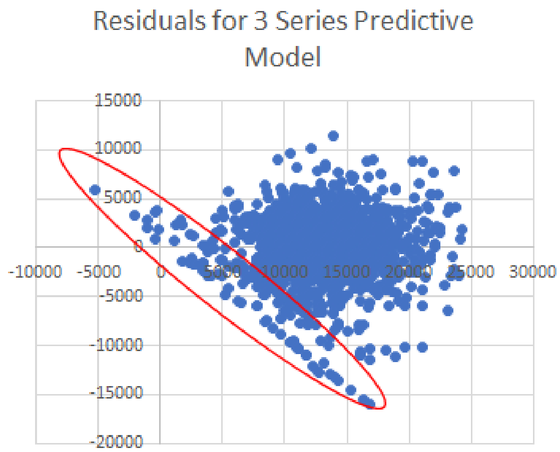
Regression Statistics	
Multiple R	0.7499
R Square	0.5623
Adjusted R Square	0.5594
Standard Error	3243.692
Observations	1644

ANOVA

	df	SS	MS	F	Significance F
Regression	11	22060550389	2005504581	190.6094125	2.8545E-283
Residual	1632	17171153480	10521540.12		
Total	1643	39231703869			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	2078.943	767.326	2.7093	0.006812312	573.90	3583.99	573.90	3583.99
mileage	-0.038	0.002	-21.4579	3.54536E-90	-0.04	-0.03	-0.04	-0.03
engine_power	99.969	3.437	29.0871	3.256E-150	93.23	106.71	93.23	106.71
diesel	4648.073	505.614	9.1929	1.12878E-19	3656.35	5639.79	3656.35	5639.79
Black	-676.811	296.795	-2.2804	0.022712435	-1258.95	-94.67	-1258.95	-94.67
Blue	-1568.878	327.787	-4.7863	1.85256E-06	-2211.81	-925.95	-2211.81	-925.95
Brown	-772.476	387.159	-1.9952	0.046182482	-1531.86	-13.10	-1531.86	-13.10
Grey	-960.232	307.630	-3.1214	0.001831544	-1563.62	-356.84	-1563.62	-356.84
Silver	-1975.444	427.791	-4.6178	4.18255E-06	-2814.52	-1136.37	-2814.52	-1136.37
Estate	-1851.570	180.493	-10.2584	5.69391E-24	-2205.59	-1497.55	-2205.59	-1497.55
Hatchback	-1023.678	256.800	-3.9863	7.00758E-05	-1527.37	-519.99	-1527.37	-519.99
Heated seats	2890.391	222.558	12.9872	8.98152E-37	2453.86	3326.92	2453.86	3326.92

Exhibit 5: Residuals Plots of 3 Series cars with Collinearity and with variables taken out to reduce Collinearity



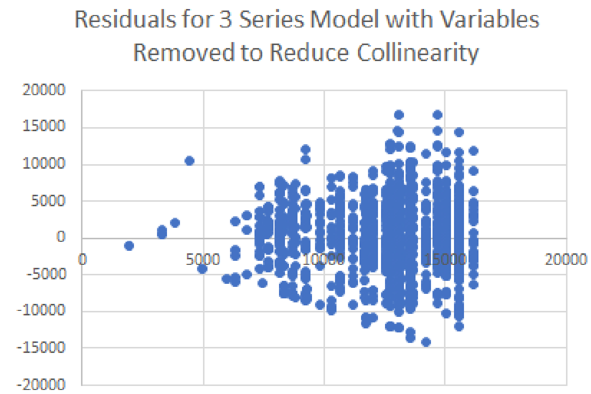


Exhibit 6: Regression analysis for 3 Series cars with variables taken out to reduce collinearity

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.451
R Square	0.204
Adjusted R Square	0.199
Standard Error	4372.913
Observations	1644

ANOVA

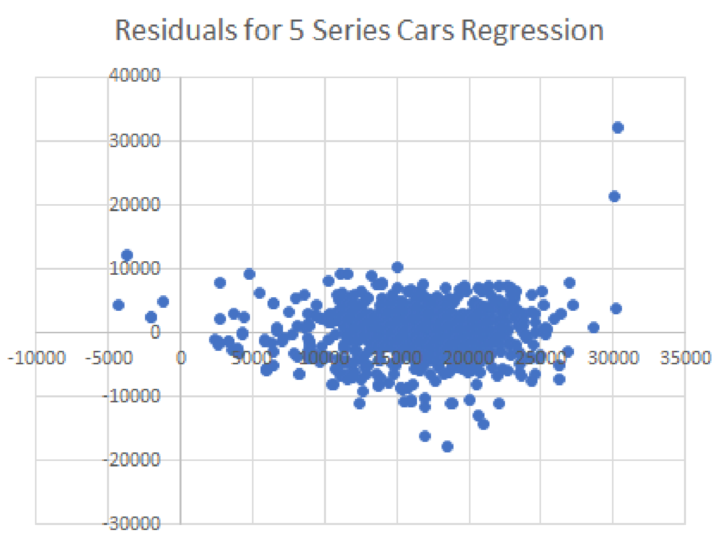
	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	7985755935	887306215	46.401	9.20181E-75
Residual	1634	31245947933	19122367.156		
Total	1643	39231703869			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	7453.562	781.335	9.540	5.003E-21	5921.038	8986.087	5921.038	8986.087
diesel	4365.628	680.969	6.411	1.888E-10	3029.963	5701.293	3029.963	5701.293
Black	-632.950	399.061	-1.586	0.1129102	-1415.675	149.774	-1415.675	149.774
Blue	-1502.967	441.071	-3.408	0.0006714	-2368.091	-637.843	-2368.091	-637.843
Brown	-1158.842	521.142	-2.224	0.0263075	-2181.018	-136.666	-2181.018	-136.666
Grey	-1150.518	413.855	-2.780	0.0054982	-1962.261	-338.776	-1962.261	-338.776
Silver	-2504.649	575.886	-4.349	1.451E-05	-3634.202	-1375.096	-3634.202	-1375.096
Estate	-1983.829	243.044	-8.162	6.489E-16	-2460.539	-1507.118	-2460.539	-1507.118
Hatchback	-3002.560	328.705	-9.135	1.887E-19	-3647.288	-2357.833	-3647.288	-2357.833
Heated seats	4367.772	294.813	14.815	0.000	3789.520	4946.024	3789.520	4946.024

Exhibit 7: Correlation table for 5 Series cars

	mileage	ngine_powe	Ln of age	diesel	Black	Blue	Brown	Grey	Silver	Estate	Hatchback	Heated seats	Sunroof	S
mileage														
engine_power	-0.0933													
Ln of age	0.5652	-0.1701												
diesel	0.0053	-0.0229	-0.1471											
Black	0.0057	0.0457	-0.0535	0.0146										
Blue	0.0986	0.0524	0.1285	-0.0470	-0.3114									
Brown	-0.0348	-0.1040	-0.0620	0.0372	-0.2061	-0.1108								
Grey	-0.0595	-0.0428	0.0011	0.0145	-0.5128	-0.2757	-0.1824							
Silver	0.0424	-0.0353	0.1049	-0.0298	-0.1847	-0.0993	-0.0657	-0.1635						
Estate	0.1038	0.0581	-0.0149	0.0217	0.0341	0.0382	0.0260	-0.0580	-0.0897					
Hatchback	-0.0114	-0.0094	-0.0413	-0.0523	0.0301	-0.0445	-0.0295	0.0050	0.0132	-0.1209				
Heated seats	-0.0590	0.0981	-0.1873	0.0215	0.0239	-0.0475	0.0255	0.0696	-0.1235	0.0122	-0.0439			
Sunroof	-0.0786	0.4332	-0.1453	0.0569	-0.0402	0.0511	-0.0355	-0.0347	0.0441	0.1120	-0.0451	0.0882		
Sport series	-0.1020	0.3528	-0.2139	-0.0345	0.0604	-0.0465	-0.0100	-0.0467	-0.0063	0.0239	-0.0157	0.1475	0.2327	

<u>Exhibit 8:</u> Residuals Plot of 5 Series cars regression



<u>Exhibit 9:</u> Regression analysis for 5 Series cars

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.761
R Square	0.579
Adjusted R Square	0.575
Standard Error	3790.950
Observations	1025

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	11	20036153309	1821468483	126.7434478	1.0117E-181
Residual	1013	14558129872	14371302.93		
Total	1024	34594283180			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	7094.756	1336.814	5.307	1.36744E-07	4471.515	9717.997	4471.515	9717.997
mileage	-0.054	0.002	-26.114	2.3444E-115	-0.058	-0.050	-0.058	-0.050
engine_power	85.107	4.808	17.702	2.66507E-61	75.672	94.541	75.672	94.541
diesel	6063.975	881.403	6.880	1.04732E-11	4334.391	7793.560	4334.391	7793.560
Black	-2067.330	548.988	-3.766	0.000175635	-3144.613	-990.046	-3144.613	-990.046
Blue	-2832.934	602.494	-4.702	2.93234E-06	-4015.213	-1650.655	-4015.213	-1650.655
Brown	-1461.380	688.593	-2.122	0.034057121	-2812.611	-110.149	-2812.611	-110.149
Grey	-2443.690	556.651	-4.390	1.25252E-05	-3536.012	-1351.369	-3536.012	-1351.369
Silver	-2912.983	724.264	-4.022	6.19959E-05	-4334.213	-1491.753	-4334.213	-1491.753
Estate	-1694.368	243.499	-6.958	6.16954E-12	-2172.188	-1216.549	-2172.188	-1216.549
Hatchback	416.618	1113.226	0.374	0.70830163	-1767.876	2601.111	-1767.876	2601.111
Heated seats	3351.063	484.353	6.919	8.0728E-12	2400.612	4301.513	2400.612	4301.513

References

Harder, D., & Rodgers, J. (2020). Theory. Retrieved December 09, 2020, from https://ece.uwaterloo.ca/~dwharder/NumericalAnalysis/06LeastSquares/transformations/complete.html

Saslow, E. (2018). Collinearity. Retrieved December 09, 2020, from https://medium.com/future-vision/collinearity-what-it-means-why-its-bad-and-how-does-it-affect-other-models-94e1db984168%EF%BC%89