Sales Price of Laptops Based on Their Specifications

Hyunwoo Cho

Jay Jung

Gun Hee Lee

Chan Hong Park

Seoyul Um

Mario Wijaya

Team #10

Table of Contents

Introduction:

In the past decade, advancement of technology has been a key focus for many companies as consumer products continue to change due to new emerging technology features; the comparison of the very first iPhone released in 2007 to the latest iPhone 6s in 2016 is a clear evidence of the rapid pace of technology advancement. A laptop, along with a cell phone, is one of the most essential electronic devices. According to Statistics Band, approximately 240 million computers were sold last year. The market for computer laptops continues to remain competitive with major companies adopting technological advances. This revolution has led to short product cycles, continual price erosion, and a barrier for smaller companies to enter the market. The major competitors are faced with a challenging task of correctly pricing their laptop to both remain competitive and incorporate the latest technologies; it is not uncommon to identify laptops incorrectly priced at major retail stores and online. Our team decided to create a regression model to correctly identify laptop prices for consumers.

Problem Statement:

When shopping for a new laptop a consumer may look for certain specifications and features on a budget. College students, who generally are financially unstable, have a limited budget to afford high-end laptops. There are several factors influencing the price of the laptop. Usually high specifications and more features mean more money. The purpose of the paper is to identify the most significant factors of laptop which drives the laptop prices by developing a regression model to forecast the prices of laptops. Using our regression model and analysis, one may be able to identify the correct price of the laptop instead of performing a competitive analysis. For a consumer who lacks knowledge about laptops may find our model to be useful. Especially our model may be helpful to consumers with a limited budget, such as students, since they may be able to predict the price of the laptop given the features and specifications that they want.

Data Description:

Amazon, most popular and reliable online retailer, offers competitive pricing and frequent pricing updates. Thus, the team decided to collect data from Amazon. However, after running regression analysis on the data collected from Amazon, we found that the data did not accurately depict the price because there are too many individual sellers. Therefore, the team recollected the data from BestBuy instead. According to pcworlds.com, BestBuy came in second after Amazon in laptop buying online retailers. [http://www.pcworld.com/article/196696/best_stores_to_buy_laptops.html]

Exactly 200 laptop data points were collected based on popularity and brand. For every data point, the following variables were recorded:

Dependent Variable(y): Sales price of the laptop (in US dollars)

Quantitative Variable:

(1) **Display Size**: Size of screen's display measured in inch

(2) **RAM**: Random access memory (form of computer data storage) measured in gigabyte

(3) **HD**: Hard drive (used for storing information) measured in gigabyte

(4) **SSD**: Solid state disk (used for storing information) measured in gigabyte

(5) **Weight**: Mass of laptop measured in lbs (pounds)

(6) **Processor Speed**: Maximum number of calculations per second the processor can perform, measured in gigahertz (GHz)

Note: Giga indicates 10^6 unit

Qualitative Variable:

(1) **Touch Screen**: Special feature to operate laptop by touching the screen. Laptops are categorized into either touch screen or no touch screen. We coded these into the following categories:

Touch screen = 1, No touch screen = 0

(2) **CPU**: Central processor unit that carries out instructions of a computer program. Types of CPU that we obtained from our samples are AMD A6, AMD A8, AMD FX,

Intel Celeron, Intel Pentium, Intel i3, Intel i5, Intel i7. These types of CPUs were chosen based on the most popular choices of customers.

(3) **Bluetooth**: Wireless technology standard for exchanging data over short distances. One of the most crucial feature to most current laptops. We coded these into the following categories:

Bluetooth = 1, No Bluetooth = 0

(4) **Company**: Manufacturer of the laptop. Manufacturers that we use for our samples are HP, Dell, Lenovo, Asus, Acer and Toshiba. These qualitative and quantitative variables will help us estimate the price of laptops based on certain factors.

During the data collection phase, our group collected samples from Amazon.com. After several analysis, we found that there is a significant number of private sellers who set their own price which is much higher/lower than the market price. After encountering this problem, our team decided to collect data from the Best Buy website because there is only one seller, Best Buy, which is more reliable in terms of stable pricing compared to sporadic pricing on Amazon.

Analysis:

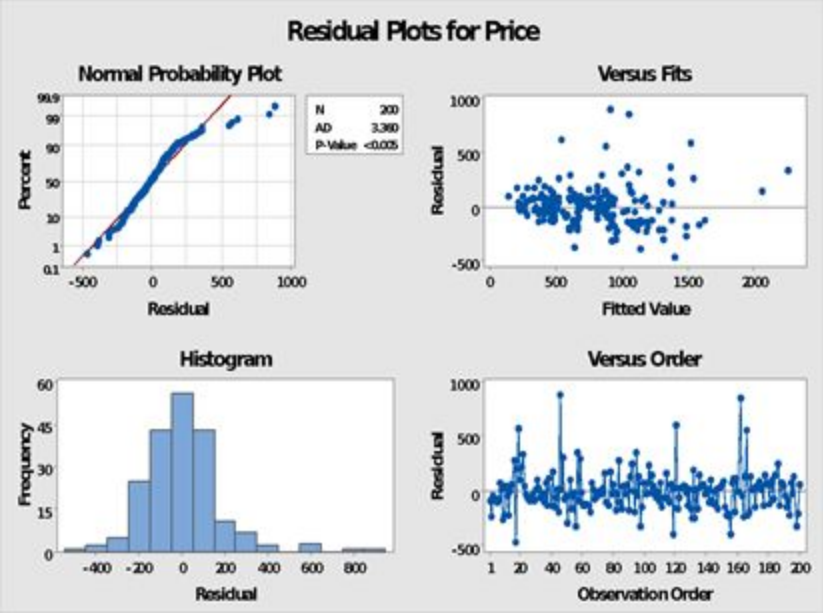
First Order Model

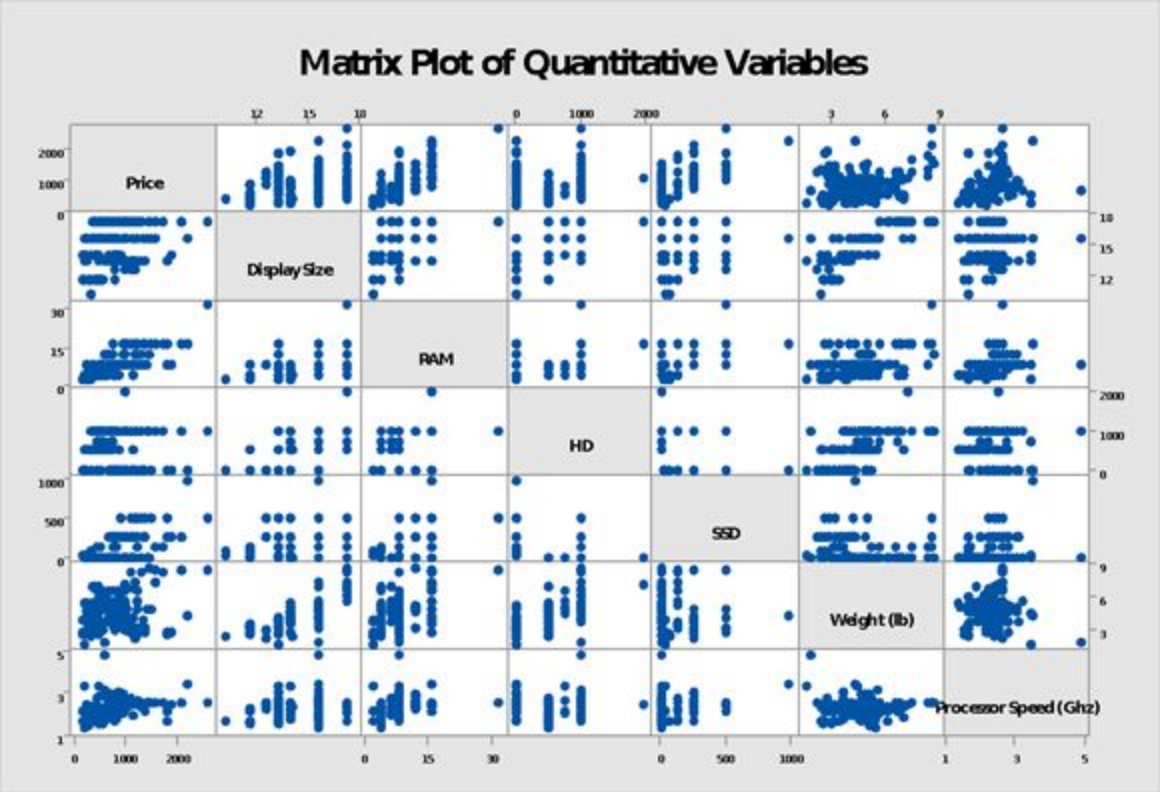
Examining the first model, we identified the regression model with R-sq of 80.12% and R-sq (adj) of 77.90%. Before examining the normal probability, versus fits, histogram, and versus order plots, we kept in mind that some outliers were to be expected as one of the laptop brands is noticeably more expensive compared to others.

The data points of the normal probability plot is not a straight line with a few outliers and the p-value of the Anderson-Darling is less than 0.05; assumption of the normality assumption is severely violated. Also, the constant variance assumption is violated due to the residual versus

fits plot where there exists a severe fan-out pattern with data being clustered. The histogram data is skewed to the right and is not symmetric due to extreme outliers.

Moreover, analyzing the p-values of the variables, we note that every variable except “RAM”, “SSD”, and “Cpu” is not significant due to the value being less than the alpha(0.05). Before assuming the variable’s insignificance, we found out that experimentation of several transformations and deletion of the outliers in different orders are necessary.





Natural Log Transformation Model

We considered 1/y, square root of y, and Ln(y) for transformations and decided to apply Natural log transformation as it had the best versus fit graph. Natural log transformation was performed to validate the assumptions in the regression model and stabilize the variance. After transformation, R-sq and R-sq (adj) improved to 84.22% 82.46% respectively. Also, since p-value of the Anderson Darling Test is 0.23, hypothesis that the data follows a specified distribution for the Anderson Darling test is rejected at alpha = 0.05. The residual versus fits plot displays more random pattern and indicates that there is a lack of fit. Compared to the first order, data is more spread out but there still exists a pattern that hints unusual observation. Additionally, there are many significant outliers. Thus, observing the data of SRES and TRES, we removed the following outliers: 162, 121, 119, 46.

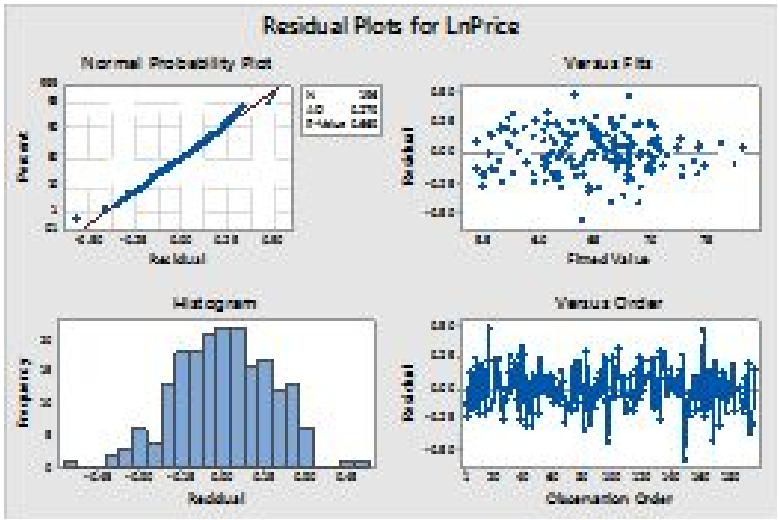
After removing outliers, we identified that our model significantly improved. R-sq and R-sq (adj) improved to 87.49% 86.06% respectively and residual versus fits plot displayed no

pattern. Additionally, the model does not depict multicollinearity as all of VIFs are lower than 10. However, normality probability plot displayed curvature and in the ANOVA, p-values of "HD", "Processor speed", "Bluetooth", and "Company" were less than alpha(0.05), which indicated that they are not significant predictors. This differed with our initial assumption that customers based their purchase heavily on processor speed. In order to ensure this is not a fluke, second order model was applied to confirm this analysis.

Second Order Model with Outliers

In the second order model, stepwise regression and backward elimination was performed but indicated two different results.

The p-values of second order variables were analyzed with the greatest p-value variable being eliminated. This process was repeated until there exists no second order variable with p-value greater than 0.05 and we achieved the following graphs. From the normal probability plot and histogram, we observed some outliers and we wanted an R-sq value of at least 90%. Hence, we checked the unusual observations and removed outliers. The following results was then confirmed with the backward elimination which produced the same results.
Our final variables are "RAM" "HD" "SSD" "Weight" "Touch Screen" "Cpu" "RAM^2" "HD ^2" "Weight^2" "RAM*HD" "RAM*SSD"



Regression Analysis: LnPrice versus RAM, HD, SSD, Weight (lb), Touch Screen 1 No touch 0, Cpu

Method

Categorical predictor coding (1, 0)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	18	49.0636	2.72576	83.24	0.000
RAM	1	0.7322	0.73219	22.36	0.000
HD	1	0.0044	0.00444	0.14	0.713
SSD	1	0.0015	0.00149	0.05	0.831
Weight (lb)	1	0.3943	0.39432	12.04	0.001
Touch Screen 1 No touch 0	1	0.6643	0.66432	20.29	0.000
Cpu	7	6.4161	0.91658	27.99	0.000
RAM*RAM	1	0.5672	0.56718	17.32	0.000
HD*HD	1	0.3861	0.38615	11.79	0.001
SSD*SSD	1	0.2412	0.24116	7.36	0.007
Weight (lb)*Weight (lb)	1	0.5886	0.58859	17.97	0.000
RAM*HD	1	0.3054	0.30538	9.33	0.003
RAM*SSD	1	0.4214	0.42144	12.87	0.000
Error	177	5.7960	0.03275		
Lack-of-Fit	165	5.6531	0.03444	3.66	0.007
Pure Error	12	0.1130	0.00941		
Total	195	54.8597			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.180958	89.43%	88.36%	85.80%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.806	0.188	30.92	0.000	
RAM	0.0728	0.0154	4.73	0.000	24.62
HD	0.000058	0.000158	0.37	0.713	28.14
SSD	-0.000107	0.000501	-0.21	0.831	33.89
Weight (lb)	-0.2014	0.0580	-3.47	0.001	40.75
Touch Screen 1 No touch 0					
1	0.1355	0.0301	4.50	0.000	1.32
Cpu					
AMD A8	0.383	0.128	2.98	0.003	3.40
AMD FX	0.723	0.137	5.29	0.000	2.78
Celeron	-0.083	0.117	-0.71	0.478	7.56
i3	0.402	0.112	3.59	0.000	8.88
i5	0.663	0.112	5.90	0.000	13.75
i7	0.922	0.118	7.82	0.000	19.43
Pentium	0.293	0.116	2.52	0.013	5.38
RAM*RAM	-0.00466	0.00112	-4.16	0.000	74.09
HD*HD	-0.000000	0.000000	-3.43	0.001	16.40
SSD*SSD	-0.000001	0.000001	-2.71	0.007	12.79
Weight (lb)*Weight (lb)	0.02367	0.00558	4.24	0.000	40.60
RAM*HD	0.000045	0.000015	3.05	0.003	38.73
RAM*SSD	0.000168	0.000047	3.59	0.000	61.84

Regression Equation

LnPrice = 5.806 + 0.0728 RAM + 0.000058 HD - 0.000107 SSD - 0.2014 Weight (lb)
+ 0.0 Touch Screen 1 No touch 0_0 + 0.1355 Touch Screen 1 No touch 0_1
+ 0.0 Cpu_AMD A6 + 0.383 Cpu_AMD A8 + 0.723 Cpu_AMD FX - 0.083 Cpu_Celeron
+ 0.402 Cpu_i3 + 0.663 Cpu_i5 + 0.922 Cpu_i7 + 0.293 Cpu_Pentium - 0.00466 RAM*RAM
- 0.000000 HD*HD - 0.000001 SSD*SSD + 0.02367 Weight (lb)*Weight (lb)
+ 0.000045 RAM*HD + 0.000168 RAM*SSD

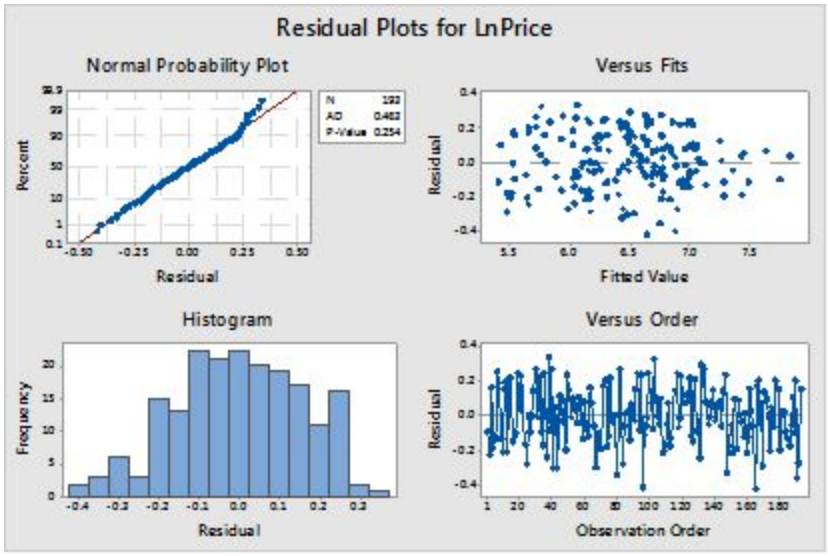
Fits and Diagnostics for Unusual Observations

Obs	LnPrice	Fit	Resid	Std Resid	
16	6.796	6.306	0.490	2.79	R
22	7.063	7.029	0.034	0.57	X
90	7.696	7.627	0.069	1.06	X
93	6.918	6.788	0.130	1.51	X
97	6.477	6.895	-0.418	-2.36	R
102	6.397	6.255	0.142	0.94	X
103	5.561	5.631	-0.071	-0.48	X
104	6.067	5.749	0.317	2.17	R X
132	5.416	5.663	-0.247	-1.67	X
149	5.799	6.170	-0.371	-2.36	R
162	7.265	6.795	0.470	2.65	R
168	6.223	6.639	-0.417	-2.60	R
194	6.397	6.759	-0.362	-2.06	R

R Large residual
X Unusual X

Second Order Model without Outliers

After analyzing SRES, TRES, and Cooks distance, we removed three outliers 16, 149 and 162. With the removal of outliers, the R-sq increased to 90.74% and R-sq (adj) to 89.79% which produced the final model. Our final model data corresponds to the normal probability plot. Additionally, residual versus fits plot displays a random pattern while the histogram is no longer skewed.



Conclusions:

After multiple regression models, our final model is the following:

```
Regression Equation

LnPrice = 5.806 + 0.0728 RAM + 0.000058 HD - 0.000107 SSD - 0.2014 Weight (lb)
          + 0.0 Touch Screen 1 No touch 0_0 + 0.1355 Touch Screen 1 No touch 0_1
          + 0.0 Cpu_AMD A6 + 0.383 Cpu_AMD A8 + 0.723 Cpu_AMD FX - 0.083 Cpu_Celeron
          + 0.402 Cpu_i3 + 0.663 Cpu_i5 + 0.922 Cpu_i7 + 0.293 Cpu_Pentium - 0.00466 RAM*RAM
          - 0.000000 HD*HD - 0.000001 SSD*SSD + 0.02367 Weight (lb)*Weight (lb)
          + 0.000045 RAM*HD + 0.000168 RAM*SSD
```

For instance, this specific laptop found on Bestbuy is 749.99 (899.99 on sale)

The following laptop has the specs : i7, 8gb, 0gb HDD, 256gb SSD, 3.66 pounds, TS 1. According to our final model, the prices come down to $977.23. Through our model, we can say that this product is underpriced.

Appendix:

I. First Order Model Analysis

Figure 1A

Regression Analysis: Price versus Display Size, RAM, HD, SSD, Weight (lb), Processor Sp, ...

Method

Categorical predictor coding (1, 0)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	20	26353204	1317660	36.08	0.000
Display Size	1	88515	88515	2.42	0.121
RAM	1	929413	929413	25.45	0.000
HD	1	20728	20728	0.57	0.452
SSD	1	1458841	1458841	39.94	0.000
Weight (lb)	1	199514	199514	5.46	0.021
Processor Speed (Ghz)	1	10205	10205	0.28	0.598
Touch Screen 1 No touch 0	1	3822	3822	0.10	0.747
Bluetooth 1 No Bluetooth 0	1	41529	41529	1.14	0.288
Company	5	100582	20116	0.55	0.738
Cpu	7	1882384	268912	7.36	0.000
Error	179	6537860	36524		
Lack-of-Fit	167	6456826	38664	5.73	0.001
Pure Error	12	81034	6753		
Total	199	32891064			

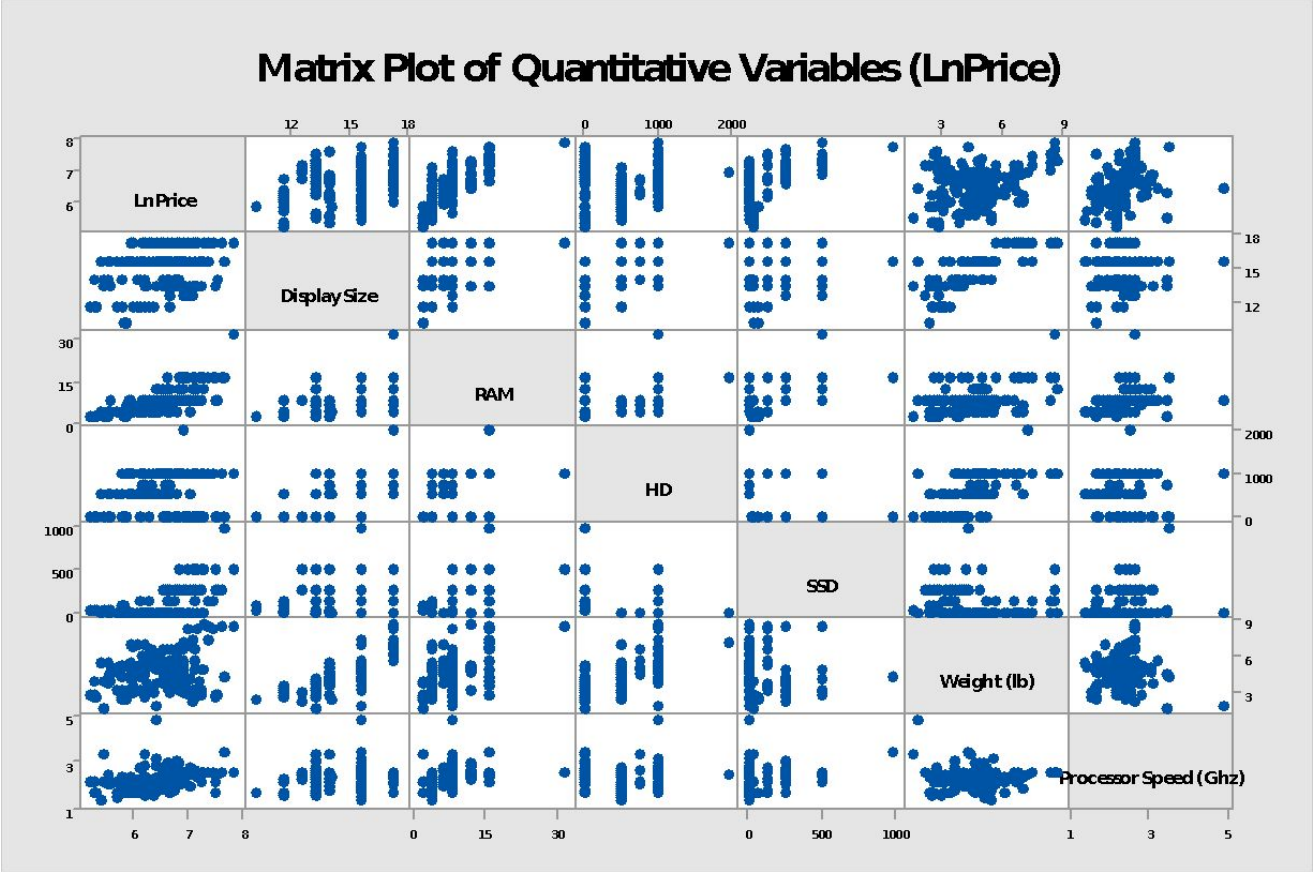
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
191.113	80.12%	77.90%	74.95%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	361	246	1.47	0.144	
Display Size	-27.8	17.9	-1.56	0.121	4.71
RAM	29.26	5.80	5.04	0.000	3.16
HD	0.0462	0.0614	0.75	0.452	3.90
SSD	1.075	0.170	6.32	0.000	3.56
Weight (lb)	49.3	21.1	2.34	0.021	4.93
Processor Speed (Ghz)	-22.4	42.3	-0.53	0.598	1.95
Touch Screen 1 No touch 0					
1	11.0	34.1	0.32	0.747	1.55
Bluetooth 1 No Bluetooth 0					
1	69.9	65.5	1.07	0.288	1.43
Company					
Asus	1.1	52.6	0.02	0.984	2.23
Dell	-39.6	57.5	-0.69	0.492	2.55
HP	11.1	59.2	0.19	0.851	2.64
Lenovo	-3.1	57.2	-0.05	0.957	2.64
Toshiba	-59.5	59.6	-1.00	0.319	2.48
Cpu					
AMD A8	100	143	0.70	0.486	3.79
AMD FX	236	151	1.56	0.120	3.03
Celeron	-21	128	-0.17	0.867	8.02
i3	108	123	0.88	0.382	9.69
i5	237	125	1.90	0.059	15.96
i7	411	130	3.16	0.002	21.55
Pentium	24	133	0.18	0.858	6.29

Figure 1B: Matrix Plot



Correlation: Price, Display Size, RAM, HD, SSD, Weight (lb), Processor Sp, Touch Screen, ...

	Price	Display Size	RAM	HD
Display Size	0.194			
	0.006			
RAM	0.767	0.413		
	0.000	0.000		
HD	-0.043	0.668	0.225	
	0.545	0.000	0.001	
SSD	0.634	-0.254	0.408	-0.597
	0.000	0.000	0.000	0.000
Weight (lb)	0.212	0.845	0.420	0.650
	0.003	0.000	0.000	0.000
Processor Speed	0.415	0.232	0.460	0.143
	0.000	0.001	0.000	0.043
Touch Screen 1 N	0.166	-0.308	0.072	-0.234
	0.019	0.000	0.309	0.001
Bluetooth 1 No B	0.214	-0.098	0.132	-0.093
	0.002	0.169	0.063	0.189
Cpu_AMD A6	-0.137	0.053	-0.112	-0.027
	0.053	0.459	0.114	0.701
Cpu_AMD A8	-0.122	0.109	-0.055	0.114
	0.084	0.123	0.442	0.107
Cpu_AMD FX	-0.024	0.037	-0.022	0.130
	0.741	0.602	0.761	0.066
Cpu_Celeron	-0.394	-0.390	-0.375	-0.323
	0.000	0.000	0.000	0.000
Cpu_i3	-0.274	0.003	-0.232	0.072
	0.000	0.969	0.001	0.308
Cpu_i5	-0.062	-0.037	-0.089	0.025
	0.381	0.602	0.209	0.726
Cpu_i7	0.706	0.224	0.645	0.033
	0.000	0.001	0.000	0.645
Cpu_Pentium	-0.232	-0.032	-0.211	0.029
	0.001	0.651	0.003	0.679
Company_Acer	-0.030	0.067	-0.049	0.016
	0.674	0.349	0.494	0.818
Company_Asus	0.131	0.207	0.216	0.151
	0.064	0.003	0.002	0.033
Company_Dell	-0.063	-0.117	-0.044	-0.070
	0.375	0.100	0.532	0.328
Company_HP	-0.050	-0.054	-0.104	-0.091
	0.482	0.447	0.141	0.202
Company_Lenovo	0.083	-0.097	0.021	-0.059
	0.245	0.170	0.769	0.408
Company_Toshiba	-0.082	-0.007	-0.050	0.052
	0.251	0.925	0.482	0.465

	SSD	Weight (lb)	Processor Speed	Touch Screen 1 N
Weight (lb)	-0.283			
	0.000			
Processor Speed	0.243	0.099		
	0.001	0.163		
Touch Screen 1 N	0.288	-0.330	0.097	
	0.000	0.000	0.173	
Bluetooth 1 No B	0.131	-0.017	0.075	0.143