ADTA 5130 DATA ANALYTICS - 1
FINAL PROJECT

Analyzing Sales Data of Laptops

by

Poojitha Gumberaopeta

INTRODUCTION:

In this project, we will be analyzing sales data of laptops. We will perform ANOVA and build a regression model to analyze our data. We tailored a few research questions that would help us to understand them in detail. We will be using CRISP-DM (Cross Industry Standard Process for Data Mining) methodology for our analysis.

METHODOLOGY:

1. Business understanding:

In our project, we chose a dataset that contains the details of the sales of the laptops. It includes data regarding laptops like its configuration, price, screen size, battery life, RAM, processor speed, hard disk size, whether it has bundled applications and whether it is integrated wireless or not. These features can be called as variables and these variables play a huge part in determining our analysis. We analyze these variables using various methods like ANOVA and Regression models which help us to form a conclusion about the given data. These models help us with providing insights into the data that would help us with improving the sales of the products or with any business aspect. We will try to predict the variables that are acting dependent and independent by which we can draw a conclusion on how a variable is affecting the sale of the product.

2. Data understanding

Our dataset contains 9 variables that are configuration, price, screen size, battery life, RAM, processor speed, hard disk size, bundled applications and integrated wireless. These include both numerical and categorical data.

Data Dictionary:

Variable	**Description**
Configuration	Configuration gives us the hardware and software details of the system of the laptop.
Price	It is the amount of money in dollars required by a customer to buy a laptop.
Screen size	It is the size of the screen of the laptop in inches.

Battery life	It is the capacity of the battery of a laptop which determines how long a laptop can be used, once fully charged. It is measured in hours.
RAM (Random access memory)	It is a short-term memory storage drive which is measured in Gigabytes.
Processor speed	The processor or CPU is the main component of the system, which measured in gigahertz.
Hard disk size	Hard Disk also known as a storage unit, is used to store data in a computer and it is measured in Gigabytes
Bundled applications	Bundled applications in a laptop are the group of similar applications. For example, MS Office (Word, Excel, PowerPoint, etc.)
Integrated wireless	Integrated Wireless in a laptop is whether it has wireless capabilities like Wi-Fi, Bluetooth and NFC, etc.

Descriptive analysis of the variables:

Configuration		*Price*		*Screen Size (Inches)*	
Mean	328.4906149	Mean	1481.551246	Mean	15.5699657
Standard Error	0.693638822	Standard Error	0.435730098	Standard Error	0.002854973
Median	304	Median	1490	Median	15
Mode	61	Mode	1000	Mode	15
Standard Deviation	219.3467584	Standard Deviation	137.7892664	Standard Deviation	0.902817064
Sample Variance	48113.00041	Sample Variance	18985.88193	Sample Variance	0.815078651
Kurtosis	-0.526814974	Kurtosis	4.203014018	Kurtosis	-1.09244368
Skewness	0.577545952	Skewness	-1.439462657	Skewness	0.95266897
Range	863	Range	890	Range	2
Minimum	1	Minimum	1000	Minimum	15
Maximum	864	Maximum	1890	Maximum	17
Sum	32848733	Sum	148153643	Sum	1556981
Count	99999	Count	99999	Count	99999
Confidence Level(95.0%)	1.359523565	Confidence Level(95.0%)	0.854025635	Confidence Level(95.0%)	0.005595711

Battery Life (Hours)		*RAM (GB)*		*Processor Speeds (GHz)*	
Mean	5.02231022	Mean	7.738157	Mean	1.879778
Standard Error	0.00257762	Standard Error	0.013031	Standard Error	0.001076
Median	5	Median	8	Median	2
Mode	6	Mode	8	Mode	2
Standard Deviation	0.81511076	Standard Deviation	4.120614	Standard Deviation	0.340255
Sample Variance	0.66440555	Sample Variance	16.97946	Sample Variance	0.115773
Kurtosis	-1.49378088	Kurtosis	-0.03623	Kurtosis	-1.31435
Skewness	-0.04093737	Skewness	1.045801	Skewness	0.163126
Range	2	Range	12	Range	0.9
Minimum	4	Minimum	4	Minimum	1.5
Maximum	6	Maximum	16	Maximum	2.4
Sum	502226	Sum	773808	Sum	187975.9
Count	99999	Count	99999	Count	99999
Confidence Level(95.0%)	0.0050521	Confidence Level(95.0%)	0.02554	Confidence Level(95.0%)	0.002109

HD Size (GB)	
Mean	137.45537
Standard Error	0.3147379
Median	120
Mode	120
Standard Deviation	99.528363
Sample Variance	9905.8951
Kurtosis	-0.9317101
Skewness	0.8487961
Range	260
Minimum	40
Maximum	300
Sum	13745400
Count	99999
Confidence Level(95.0%)	0.6168824

3. Data preparation

In this stage, we'll modify our data according to our analyzing methods. Since we will be performing ANOVA and Regression, we will choose the data that is numerical. We chose the numerical data and generated a correlation matrix that is used to determine the relations amongst the variables. We have variables configuration, price, screen size, battery life, Ram, processor speed, and hard disk size. The figure below is a correlation matrix. We can see that there is a strong positive correlation of 0.84 between the screen size and the configuration as depicted below with the green color. There is a moderately positive correlation between the battery life and configuration, price and configuration, and between RAM and price. We can see that there are negative correlations for the HD size with most of the variables. Similarly, the variable battery life has mostly the negative correlations.

	Configuration	*Price*	*Screen Size (Inches)*	*Battery Life (Hours)*	*RAM (GB)*	*Processor Speeds (GHz)*	*HD Size (GB)*
Configuration	1						
Price	0.380355364	1					
Screen Size (Inches)	0.841477074	0.268391979	1				
Battery Life (Hours)	0.417955276	0.170909582	-0.110148128	1			
RAM (GB)	0.151998705	0.291611804	0.057364215	-0.100800711	1		
Processor Speeds (GHz)	0.134796962	0.130952858	0.117753987	-0.054986166	0.052108038	1	
HD Size (GB)	-0.142581289	0.178390747	-0.102335503	-0.080799113	-0.097265389	-0.053697173	1

4. Modelling/Analysis

To analyze our data, we will be using ANOVA and Regression methods. We have tailored a few research questions that would give us a better understanding of our dataset.

QUESTION 1:

Perform an ANOVA test on the laptop sales dataset to find the difference in the variability of the Price of the product depending on the RAM, Processor Speed and Hard Disk size.

RESULT:

The below figure shows us the ANOVA test performed on the laptop sales dataset with the given variables. Our null hypothesis (H_0) is that there is no difference amongst the given variables and our alternate hypothesis (H_A) is that there might be one or more variables that might differ from the rest of the variables. Here our confidence level is $\alpha = 0.05$. We can see that our p-value is 0. We can say that the confidence level is greater than our p-value, which means that we reject our null hypothesis (H_0) and accept the alternate hypothesis (H_A). Hence, we can conclude that there is a difference in variability in the price depending on the RAM, processing speed, and the hard disk size of the laptop.

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Price	99999	1.48E+08	1481.551	18985.88		
RAM (GB)	99999	773808	7.738157	16.97946		
Processor Speeds (GHz)	99999	187975.9	1.879778	0.115773		
HD Size (GB)	99999	13745400	137.4554	9905.895		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.55E+11	3	5.17E+10	7152736	0	2.604931
Within Groups	2.89E+09	399992	7227.218			
Total	1.58E+11	399995				

QUESTION 2:

Perform an ANOVA test on the laptop sales dataset to find the difference in the variability of the Screen Size with respective to the Price of the laptop, Configuration, and it's Battery Life.

RESULT:

The below figure shows us the ANOVA test performed on the laptop sales dataset with the given variables. Our null hypothesis (H_0) is that there is no difference amongst the given variables and our alternate hypothesis (H_A) is that there might be one or more variables that might differ from the rest of the variables. Here our confidence level is $\alpha = 0.05$. We can see that our p-value is 0. We can say that the confidence level is greater than our p-value, which means that we reject our null hypothesis (H_0) and accept the alternate hypothesis(H_A) . Hence, we can conclude that there is a difference in variability in the size of the screen with respective to the price, configuration, and battery life of a laptop.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Screen Size (Inches)	99999	1556981	15.56997	0.815079
Price	99999	1.48E+08	1481.551	18985.88
Configuration	99999	32848733	328.4906	48113
Battery Life (Hours)	99999	502226	5.02231	0.664406

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.47E+11	3	4.88E+10	2911753	0	2.604931
Within Groups	6.71E+09	399992	16775.09			
Total	1.53E+11	399995				

QUESTION 3:

Build a regression model that would have Configuration, Hard disk size, Screen size and Processor Speed as independent variables and the Price of the laptop as a dependent variable. Determine the regression equation as well.

RESULT:

The below figure shows us the Regression Statistics on the laptop sales dataset with the given independent and dependent variables respectively. We can see that the R square value is 0.21865 which is equivalent to 21.87%. Since the R value is less than 0.5, we would say that it is a weak model. Therefore, the dependent variable taken, Price is independent of the configuration, hard disk size, screen size and the process speed of the laptop.

The Regression equation (y) is the sum of the coefficients of the independent variables (mx) and the coefficient of the intercept (b).

y = mx+b

y = 1708.316715857 + 0.355195297349509 (Configuration) + 0.337957948415372 (HD size) - 29.4702139265493 (Screen size) + 36.6810863039038 (Processor speed)

SUMMARY OUTPUT

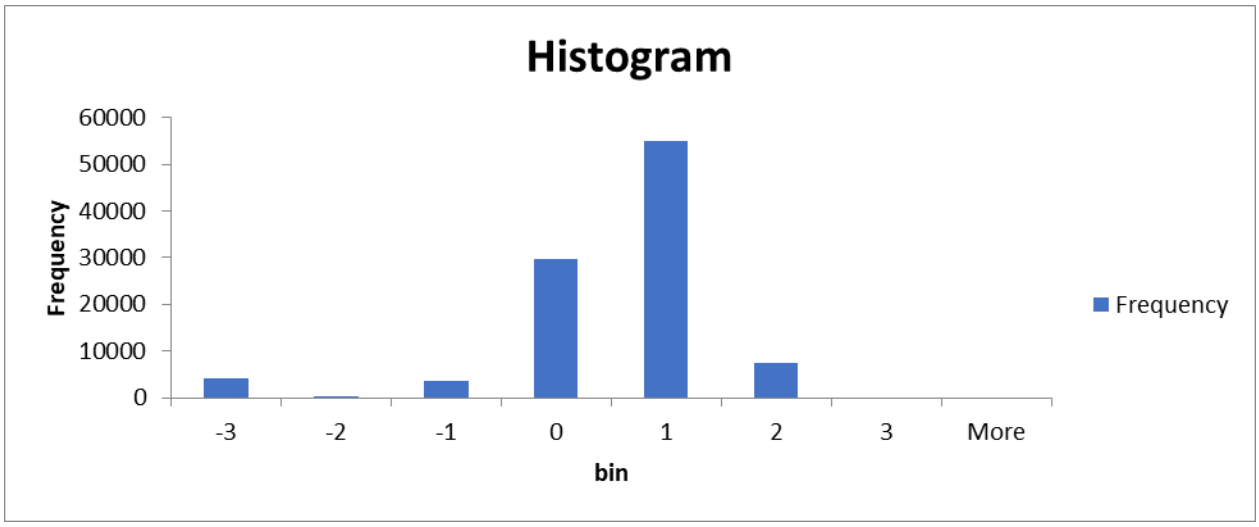
Regression Statistics	
Multiple R	0.467601492
R Square	0.218651155
Adjusted R Square	0.218619899
Standard Error	121.7997961
Observations	99999

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	415120199.1	103780049.8	6995.532	0
Residual	99994	1483430022	14835.19033		
Total	99998	1898550221			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1708.316716	11.59136987	147.3783284	0	1685.597773	1731.03566	1685.597773	1731.035658
Configuration	0.355195297	0.003274549	108.4715095	0	0.348777221	0.36161337	0.348777221	0.361613374
HD Size (GB)	0.337957948	0.003914474	86.33546536	0	0.330285627	0.34563027	0.330285627	0.34563027
Screen Size (Inches)	-29.47021393	0.790090826	-37.29977994	2E-302	-31.01878223	-27.921646	-31.0187822	-27.9216456
Processor Speeds (GHz)	36.6810863	1.143183435	32.08678955	9.4E-225	34.44046082	38.9217118	34.44046082	38.92171179

The histogram below depicts the data of our regression model. We can see that it is not normally distributed.



QUESTION 4:

Build a regression model that would have Price, RAM, Processor Speed and Hard disk size as independent variables and the Configuration of the laptop as a dependent variable. Determine the regression equation as well.

RESULT:

The below figure shows us the Regression Statistics on the laptop sales dataset with the given independent and dependent variables respectively. We can see that the R square value is 0.6098 which is equivalent to 60.98%. Since the R value is greater than 0.5, we can say that it is a strong model. Therefore, the dependent variable taken, configuration is dependent on the price, RAM, processor speed and hard disk size of the laptop.

The Regression equation (y) is the sum of the coefficients of the independent variables (mx) and the coefficient of the intercept (b).

y = mx+b

y = - 2207.89053576653 + 1.81321476503989 (Price) - 10.6937253849496 (RAM) - 58.8510543541409 (Processor Speed) - 0.733713983060879 (Hard Disk Size)

SUMMARY OUTPUT

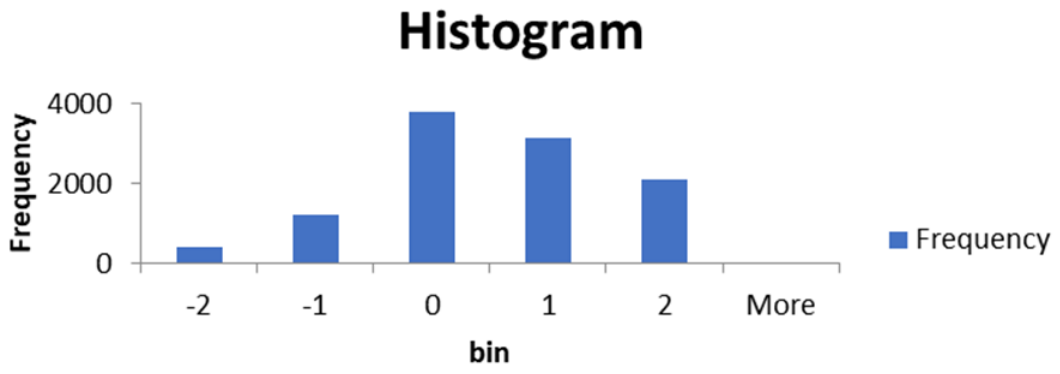
Regression Statistics	
Multiple R	0.780874216
R Square	0.60976454
Adjusted R Square	0.609618028
Standard Error	73.96051466
Observations	10659

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	91064262.97	22766065.74	4161.866416	0
Residual	10654	58279060.45	5470.157729		
Total	10658	149343323.4			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-2207.890536	19.85690881	-111.1900425	0	-2246.813784	-2168.967288	-2246.813784	-2168.967288
Price	1.813214765	0.014530561	124.7862914	0	1.784732154	1.841697376	1.784732154	1.841697376
RAM (GB)	-10.69372538	0.38759394	-27.59002213	6.593E-162	-11.45348186	-9.933968909	-11.45348186	-9.933968909
Processor Speeds (GHz)	-58.85105435	2.929781963	-20.08717887	4.0426E-88	-64.59397392	-53.10813479	-64.59397392	-53.10813479
HD Size (GB)	-0.733713983	0.008344354	-87.92939813	0	-0.750070474	-0.717357492	-0.750070474	-0.717357492

The histogram below depicts the data of our regression model. We can see that it is not normally distributed.



5. Evaluation

To evaluate our models, we have used Excel to perform ANOVA and determine the Regression models. We performed two ANOVA tests and built two Regression models for our analysis. The first ANOVA test that was done to find the difference in the variability of the Price of the product depending on the RAM, Processor Speed and Hard Disk size had a confidence level greater than the p-value, which made us to reject the null (H_0) and accept the alternate hypothesis (H_A) which means that there is a difference in the variability. The second ANOVA test that was done to find the difference in the variability of the Screen Size with respective to the Price of the laptop, Configuration, and it's Battery Life had a confidence level greater than the p-value, which made us to reject the null (H_0) and accept the alternate hypothesis (H_A) which means that there is a difference in the variability. The first regression model that would have configuration, hard disk size, screen size and processor speed as independent variables and the price of the laptop as a dependent variable has a R square value 0.21865 which is equivalent to 21.87% is considered as a weak model and that the dependent variable taken is independent of the other variables. The second regression model that would have price, RAM, processor speed and hard disk size as independent variables and the configuration of the laptop as a dependent variable has a R square value 0.6098 which is equivalent to 60.98% is considered as a moderately strong model and that the dependent variable is dependent on the independent variables.

6. Deployment

The above analysis and models are for the respective dependent and independent variables. The results are subjective with respective to the dependent and independent variables and the business need. One who wants to analyze data and built models needs to deploy the dataset of the sales and perform the analysis in their environment to predict the sales of the laptops.