



AMAZON

Group 93

ABSTRACT

In this report we are working for the office supplies, furniture, and technology department of Amazon to identify what factors lead to the highest profit per item.

By: Jessica Daher, Cameron Byrd, Elliot Tran, Brandon Yang, Brandon Zheng

Table of Contents

Introduction.....	2
Data Exploration.....	2
Statistical Analysis	2
Data Overview and Modifications.....	2
Process of Creating Regression Model	3
Final Population Regression Equation	3
Final Estimated Regression Equation.....	3
Fit of Model	3
Test Joint Significance.....	3
Test for Individual Significance	4
Interpretations of Coefficients.....	4
Residual Plot Interpretation	4
Check for Multicollinearity and Endogeneity.....	4
Relevance of Findings	5
Recommendation.....	5
Conclusion.....	5
Appendix.....	6
Figure 1 – Initial Regression Output	6
Figure 2 – Final Regression Output.....	7
Figure 3 – Revenue Residual Plot.....	7
Figure 4 – Discount % Residual Plot	8
Figure 5 – Correlation Matrix.....	8

Introduction

Amazon sells almost anything you can imagine, in this report they want us to analyze office supplies, furniture, and technology to find what factors lead to highest profit. We ran a regression analysis to see what maximizes profit. We received sales data that was categorized in revenue, discount percentage, region, shipping style, holiday seasons, and some other factors. When running the regression, we found that the discount percentage is hurting Amazon's profit the most. We recommend that Amazon focuses on selling more types of Copiers in the technology category to maximize profits and stay away from discounting anything in the furniture or office supplies category.

Data Exploration

Please click [here](#) to find the Tableau Dashboard for Amazon.

Some key takeaways that can be found in Tableau include the *Profit per State*, the *Profit per Order*, and the *Average Yearly Profit*. First, we can see that in the *Profit per State*, the lighter states have a negative average profit per order which is mostly seen in the western region. Second, we found from the *Profit per Order* that California seems to have the most profit per order, but we also see that Minnesota is darker in color, so they have more profit per capita and therefore have more money to spend on Amazon. Last, we found that in the *Average Yearly Profit*, technology has the highest profit per order.

Statistical Analysis

Data Overview and Modifications

General Order Information

The General Order Information included: order ID, order date, ship date, ship mode, customer ID, name, segment, city, state, postal code, and region.

Modifications:

1. The "Holiday Season" variable was created using the order date. It was defined by whether the order took place during the months of October, November, or December. There was no control for year.
2. Dummy variables for "Shipping Mode" include: "Same Day", "First Class", and "Second Class". The base case is "Standard Class".
3. Dummy variables for "Customer Segment" include: "Consumer" and "Corporate". The base case is "Home Office".
4. Dummy variables for "Region" include: "East", "South", and "West". The base case is "Central".

Numerical Order Information

The Numerical Order Information included: unique order ID, revenue, quantity, discount %, and profit.

Modifications:

1. *Profit per Order* was a numerical variable created. Its calculation was Profit/Quantity.

Categorical Order Information

Categorical Order Information included: unique order ID, product ID, category, and subcategory.

Modifications:

1. Dummy variables for Category include: “Furniture” and “Technology”. The base case is Office Supplies.

Process of Creating Regression Model

First, we created an initial regression model to estimate *Profit per Order* using all variables we believed to be potentially significant. This included: Revenue, Discount (%), Furniture_(d), Technology_(d), First Class_(d), Second Class_(d), Same Day_(d), East_(d), South_(d), West_(d), Holiday Season_(d), Consumer_(d), and Corporate_(d).

After analyzing the Significance F output, we determined that at least one of these independent variables significantly impacts *Profit per Order*. Using Amazon’s significance value of 10%, we discovered that Revenue, Discount %, Furniture_(d), East_(d), South_(d), and West_(d) are all significant.

We then created a final estimated regression model using only the variables that were found to be significant. After comparing the adjusted-R², the final model was determined to have a better fit.

Final Population Regression Equation

$$\widehat{\text{Profit per Order}} = \beta_0 + \beta_1(\text{Revenue}) - \beta_2(\text{Discount}\%) - \beta_3(\text{Furniture}) - \beta_4(\text{East}_{(d)}) - \beta_5(\text{South}_{(d)}) - \beta_6(\text{West}_{(d)}) + \varepsilon$$

Final Estimated Regression Equation

$$\widehat{\text{Profit Per Order}} = 12.83 + 0.04(\text{Revenue}) - 0.60(\text{Dicount } \%) - 12.18(\text{Furniture}_d) - 2.51(\text{East}_d) - 3.05(\text{South}_d) - 3.56(\text{West}_d)$$

Fit of Model

R² = 0.2560 tells us we are 25.60% of the way toward perfectly predicting *Profit per Order* using this model.

The Standard Error of 48.54 tells us that we are, on average, \$48.54 off from predicting *Profit per Order* using this model.

Click [here](#) to see Figure 2 Regression Table.

Test Joint Significance

H_a: At least one independent variable significantly impacts *Profit per Order*

H_0 : None of the independent variables significantly impact *Profit per Order*

Because the p-value of 0 is less than the significance level, we can reject the null hypothesis and conclude that at least one independent variable significantly impacts *Profit per Order*

Test for Individual Significance

Revenue, Discount %, Furniture_(d), East_(d), South_(d), and West_(d) all have p-values less than Amazon's significance value of .1; therefore, we can conclude that these variables significantly impact *Profit per Order*.

Interpretations of Coefficients

As **Revenue** increases by \$100, *Profit per Order* increases by \$4.05, on average and all else constant.

As **Discount %** increases by 10 percentage points, *Profit per Order* decreases by \$6.02, on average and all else constant.

Items sold in the **Furniture_(d)** category have an average *Profit per Order* of \$12.18 less than the items sold in the Office Supplies category, on average and all else constant.

Items sold in the **East_(d)** region have an average *Profit per Order* of \$2.51 less than the items sold in the Central region, on average and all else constant.

Items sold in the **South_(d)** region have an average *Profit per Order* of \$3.05 less than the items sold in the Central region, on average and all else constant.

Items sold in the **West_(d)** region have an average *Profit per Order* of \$3.56 less than the items sold in the Central region, on average and all else constant.

Click [here](#) to see Figure 2 Regression Table.

Residual Plot Interpretation

Since the Revenue residual plot has a funnel shape, there is evidence to suggest heteroskedasticity. This indicates that the standard error may be biased, and the hypothesis tests may be incorrect. To fix this violation, we must use White's standard errors.

All other residual plots besides Revenue have a random shape, which indicates no issues with non-linear patterns or heteroskedasticity. A linear regression would fit this model.

Click [here](#) to see Figure 3 Revenue Residual Plot.

Click [here](#) to see Figure 4 Discount % Residual Plot.

Check for Multicollinearity and Endogeneity

Since none of the independent variables have a correlation outside the range of -0.8 and 0.8 on the correlation matrix, there is no evidence to suggest multicollinearity between the significant variables.

There is also no evidence to suggest there is endogeneity since the correlations between the residuals and significant variables are all very close to zero.

Click [here](#) to see Figure 5 Correlation Matrix.

Relevance of Findings

Some important trends the Amazon leadership should know are the relationship between the discount percentage and profits. We found that the discount percentage was the leading reason why profits are decreasing. As the discount percentage increase profits drop significantly. We know that discounts are an option to entice consumers to buy products that are not selling, but this is significantly hurting Amazons profits. Our team found it surprising that copiers were the highest selling item in the technology category. Copiers make an average profit of \$219.66 per order. This was surprising the amount of profit Amazon gains per copier sold.

The variables my group found to be insignificant were Technology, First Class, Second Class, Same Day, Holiday Season, Customer, and Corporate. Although these variables were insignificant in our regression, we believe that Holiday Season should be mentioned. A Holiday Season could indicate a sale or an influx in profit which could be important.

Recommendation

From our regression analysis, we can make recommendations that Amazon can make in the future to maximize profits. We found that technology has the highest profit per order. To increase these profits Amazon should incorporate more advertising in the technology category. Since copiers are the highest selling item Amazon can work on selling more varieties and brands to increase profit. We also found that the furniture category is making the least amount of profit, so Amazon should decide to either get rid of selling furniture or start advertising more. Finally, Amazon should try to stay away from discounting because it is losing a lot of profits.

Conclusion

After receiving Amazons sales, we were able to run a regression to find what factors maximize profit. When running the regression for Amazon we found that discount percentage was decreasing Amazon's profit the most. We believe that Amazon should not discount any of the office supplies or furniture to ensure their profits in that category maintain positive. Our overall recommendation for Amazon is that they should start carrying more brands and varieties of copiers because of their profit per order. They should continue to focus on selling in their technology category and maybe think of getting rid of furniture for the time being.

Appendix

Figure 1 – Initial Regression Output

Regression Statistics						
Multiple R	0.506302291					
R Square	0.256342009					
Adjusted R Square	0.255363512					
Standard Error	48.54936377					
Observations	9894					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	13	8027319.818	617486.1398	261.9751683	0	
Residual	9880	23287562.33	2357.040722			
Total	9893	31314882.15				
	Coefficients	Standard Error	t Stat	P-value	Lower 90.0%	Upper 90.0%
Intercept	13.59530877	1.682391168	8.080943976	7.18E-16	10.82776206	16.36285548
Revenue	0.040428141	0.000802157	50.39928815	0	0.039108586	0.041747695
Discount (%)	-0.602775657	0.024382847	-24.72129915	5.7003E-131	-0.642885632	-0.562665681
Furniture (d)	-12.12798579	1.247879876	-9.718872806	3.15034E-22	-14.18075801	-10.07521358
Technology (d)	0.155082992	1.326566662	0.11690554	0.906937284	-2.027129606	2.337295589
First Class (d)	1.863116944	1.398056524	1.332647795	0.182678145	-0.436697039	4.162930927
Second Class (d)	-0.74346009	1.276111046	-0.582598272	0.560177064	-2.842672803	1.355752622
Same Day (d)	-0.213152251	2.185333215	-0.097537643	0.92230142	-3.808042587	3.381738085
East (d)	-2.595363402	1.385682096	-1.87298617	0.061099799	-4.874821355	-0.31590545
South (d)	-3.061448216	1.594187182	-1.920381904	0.054838431	-5.683898676	-0.438997757
West (d)	-3.624749082	1.369664125	-2.646451064	0.008147021	-5.877857346	-1.371640818
Holiday Season (d)	0.234197155	1.012685106	0.231263553	0.817114849	-1.431677813	1.900072124
Consumer (d)	-1.426849529	1.340232141	-1.064628646	0.287070029	-3.631541947	0.777842889
Corporate (d)	-0.707349017	1.460306449	-0.484383958	0.628124139	-3.109564617	1.694866582
Group 93						

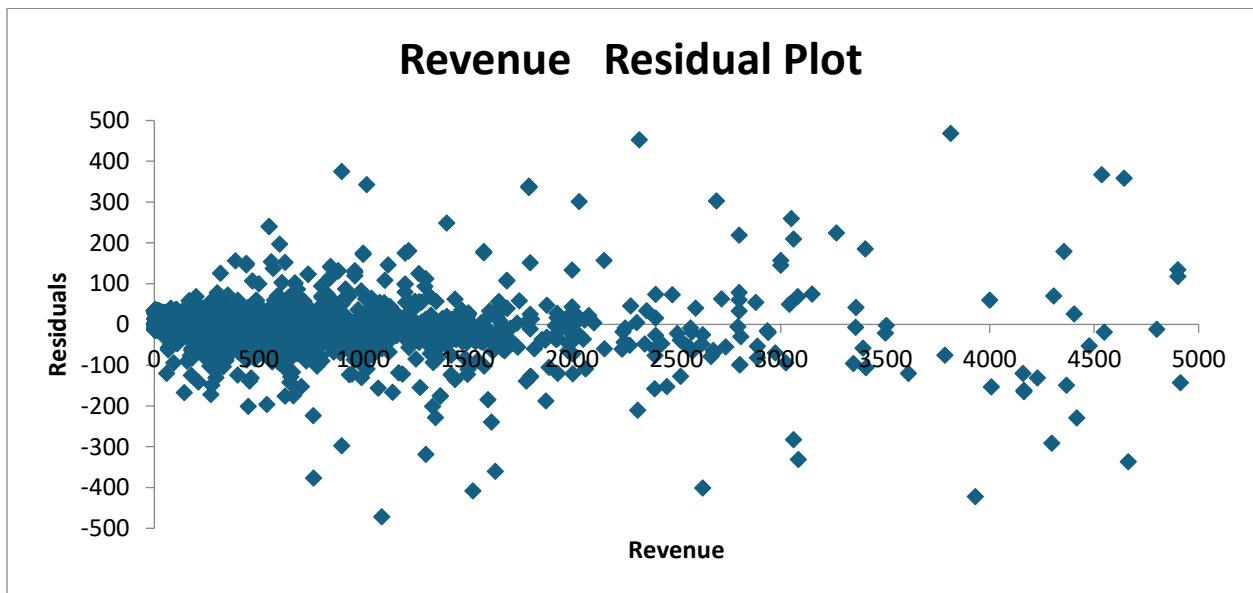
Figure 2 – Final Regression Output

Regression Statistics						
Multiple R	0.506005019					
R Square	0.256041079					
Adjusted R Square	0.255589602					
Standard Error	48.54199282					
Observations	9894					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	6	8017896.217	1336316.036	567.1187117	0	
Residual	9887	23296985.94	2356.325067			
Total	9893	31314882.15				
	Coefficients	Standard Error	t Stat	P-value	Lower 90.0%	Upper 90.0%
Intercept	12.82707067	1.200427681	10.68541727	1.66153E-26	10.85235782	14.80178352
Revenue	0.040456859	0.000785495	51.50491711	0	0.039164713	0.041749004
Discount (%)	-0.60184058	0.024326165	-24.74046305	3.6229E-131	-0.641857309	-0.56182385
Furniture (d)	-12.18325874	1.20158201	-10.13934849	4.84376E-24	-14.15987048	-10.20664701
East (d)	-2.508233161	1.383568807	-1.812872008	0.069881878	-4.784214585	-0.232251737
South (d)	-3.048497083	1.592780715	-1.913946505	0.055655831	-5.66863372	-0.428360447
West (d)	-3.556052787	1.368036478	-2.599384479	0.009352987	-5.806483409	-1.305622165
Group 93						

[Click here to return to R squared and Standard error](#)

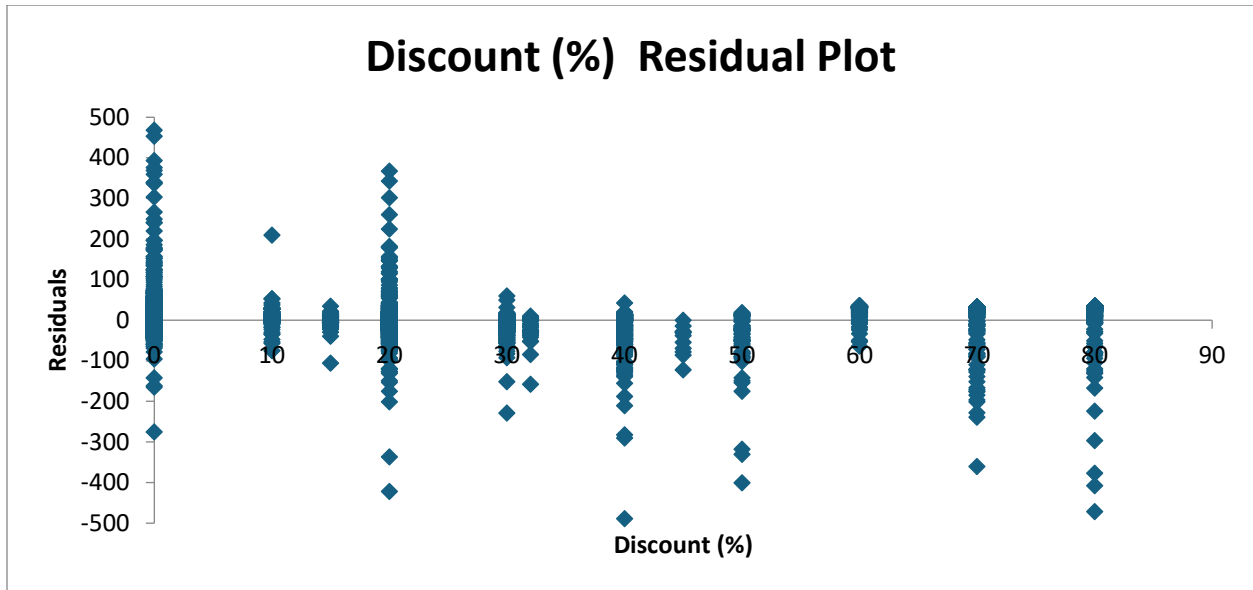
[Click here to return to Coefficient Interpretation](#)

Figure 3 – Revenue Residual Plot



[Return to related section.](#)

Figure 4 – Discount % Residual Plot



[Return to related section.](#)

Figure 5 – Correlation Matrix

	<i>Residuals</i>	<i>Revenue</i>	<i>Discount (%)</i>	<i>Furniture (d)</i>	<i>East (d)</i>	<i>South (d)</i>	<i>West (d)</i>
Residuals	1						
Revenue	-3.90736E-15	1					
Discount (%)	1.11147E-15	-0.028569446	1				
Furniture (d)	3.62878E-16	0.099983534	0.043832373	1			
East (d)	5.60035E-17	0.009498834	-0.033012444	-0.000617207	1		
South (d)	-2.26613E-16	0.008756301	-0.018945442	-0.007914687	-0.278276128	1	
West (d)	-5.04716E-18	-0.004025923	-0.156593159	0.01329699	-0.432485699	-0.302500162	1
Group 93							

[Check for multicollinearity and Endogeneity section](#)