Data Analysis of a Start-Up Business

by
Millicent Cowart
Advisor: Samantha Seals, PhD

An Undergraduate Proseminar
In Partial Fulfillment of the Degree of
Bachelor of Science in Mathematics
The University of West Florida
August 12, 2024

The Proseminar of Millicent Cowart is approved:

Samantha Seals, PhD, Proseminar Advisor	Date
Samantha Seals, PhD, Committee Chair	Date

Accepted for the Department/Division:

Jia Liu, PhD, Chair	Date

Table of Contents

Abstract

Objectives:

This analysis aims to observe and model data collected from a start-up business made using Squarespace. First, a new supplier directory was needed for Milli's glassware. Second, observations of engagement measures could reveal page popularity or website flaws.

Methods:

General descriptive techniques and visualization methods were used to explore data collected on supplier directories. Two Negative Binomial models and one Beta regression model were fit to observe engagement.

Results:

Spocket was chosen as the best supplier for three out of four sections. The sections analyzed product range versus reliability and reputation, the cost of minimum order requirements, eithcs versus reputation, and average shipping time versus cost.

Modeling is based on the current state of engagement and used to make suggestions for adjustment. The higher the bounce rate, the quicker views will decrease as time on page increases. The trend of bounce rate makes sense as a small number of people will view a page for a long period that has a high tendency to induce an exit of the website. On the other hand, pages with a twenty-five percent bounce rate is ideal. These pages have a wide range of average number of seconds on the page, but the views stay the same. Similarly, for a low exit rate of twenty-five percent, views increase quickest as time on page increases. The trend of exit rate indicates that time on page stays the same as views increase for a high exit rate.

The results of a Beta regression model were interpreted using average marginal effects. Beta regression was used to model exit rate as a function of page price, number of seconds spent viewing, sale status, type of page, importance of page, number of weeks the page was on the website, and an interaction between page price and sale status. The resulting average marginal effects provided an interpretation of the relationships. For a one unit increase in page price, the exit rate increases by 2.7 percent. For a one unit increase in time on page, the exit rate increases by 0.2 percent. For a one unit increase in importance of page, the exit rate decreases by 28.1 percent. Additionally, the significant interaction between page price and sale status indicated a significant trend. For a page that was not on sale, as price increases, exit rate exponentially increases. For a page that was on sale, as the price increases, exit rate exponentially decreases.

Conclusions:

These findings indicate that Spocket is the best supplier. Their automated drop shipping system expedites the processing of orders while keeping shipping time and cost low. Additionally, they have around 20,000 reliable suppliers with high ethical standards uncompromising to cost and efficiency. In the dropshipping

world, Spocket has a stellar reputation, yet risk is minimized as they are a U.S. and Europe-based company with its own customer service as well as available communication with individual suppliers. Lastly, returns are accepted and warranties are available.

For modeling, one aim should be to increase the views of pages with bounce rates around twenty-five percent. Another aim should be to analyze pages with a wide range of views but a time on page that stays relatively the same. This would indicate that something on these pages is causing the majority of customers to be uninterested in the content. Considering the results of the Beta regression model, higher ticket items should be placed on sale. This is because an increase in price increases the exit rate, but a page that's on sale has an exponentially decreasing exit rate for higher prices. Additionally, as importance of page increases, the exit rate significantly decreases. This shows that the pages ranked as most important are also viewed that way by customers.

Chapter 1

Introduction

1.1 Statement of Problem

The first business problem to consider was the optimization of sourcing product. The product would be directly shipped to the customer or packaged and shipped by Milli's Glassware. The product and directory should align with ethical, sustainability, and quality standards. Additionally, each directory should be evaluated for risks such as shipping delays or quality issues.

The second business objective was to observe engagement of pages. Depending on the way customers get to a page, this portion aims to observe the trends in leaving the website or viewing additional pages.

1.2 Relevance of Problem

By choosing the best supplier directory, Milli's Glassware will benefit from the best combination of the following attributes: quick shipping times, low shipping costs, an adequate amount of reliable suppliers, and top ethical and sustainability standards.

Next, observing engagement provides a way to adjust themes or business motives and initiatives and update the analysis periodically. For example, what prices should be increased or decreased, what pages are meeting engagement targets, or any malfunctions in pages.

1.3 Literature Review

When exploring a data set two important applications are measures of central tendency and dispersion. Using the mean, two variables can be compared by their averages, or an individual value can be compared to the average [5]. When there are outliers, the median is a better measure of central tendency as it provides information about the central value for a series [5]. Lastly, the mode is best suited for fixed category data. The range can be used to define the boundaries of a data set, and the interquartile range does the same with the four quartiles but is less affected by outliers [5].

Additionally, data visualization is an important part of data exploration. Composition depicts the whole and its segments. Distribution can show the count per segment, the most popular being a histogram [5]. Comparison differentiates data across segments or periods. Lastly, scatter or bubble plots can explore relationships in the dataset.

Moving from exploring to modeling the data, independent integer data recorded over a set interval can reflect a Poisson distribution, especially when concentrated around one bound [3]. Additionally, the Poisson distribution is identified by a single parameter λ represented by a mean that equals its variance [3]. Count data on rare events is usually skewed and non-normal, which violates the assumptions of general linear models: a continuous and normally distributed dependent variable with a mean of zero and constant variance σ^2 [3]. Generalized linear models model some function of the dependent variable, whereas linear regression models the observed value of the dependent variable [3]. Poisson regression models the log-link function which represents the natural log of the dependent variable Y in Equation 1.1 where ϵ_i denotes the difference between the observed value and the value predicted by the model[3]:

$$log(Y) = \beta_0 + \beta_1 X_1 + \epsilon_i \tag{1.1}$$

The model coefficients can be interpreted like a general linear model by exponentiating both sides of the equation as the linearity is with respect to the link function of the observed outcome data [3]. Generalized linear models, including Poisson regression, use maximum likelihood estimation to estimate the regression coefficients [3].

Negative Binomial regression can be used as an alternative when data does not meet the assumptions of Poisson regression. Negative Binomial regression introduces a dispersion parameter that reflects the uncertainty in the true rate of occurrence for individual cases [2]. Assume y_i is generated by a Poisson process and $\hat{\mu_i}$ is a random variable that reflects the unknown true value of the parameter of the Poisson process [2]. Then, $\hat{\mu_i} = d_i\lambda(X_i)_{\epsilon_i}$, and $\lambda(X_i) = \exp(\sum_{j=1}^{J} \beta_j X_{ij})$ where ϵ is an unobserved random variable greater than zero [2]. Epsilon assumes a gamma distribution, so $E(\epsilon_i) = 1$ and $Var(\varepsilon_i) = \theta^{-1}$ [2]. Thus, $E(\hat{\mu_i}) = d_i\lambda(X_i) = \mu_i(X_i, d_i)$ and $Var(\hat{\mu_i}) = \theta^{-1}\mu_i^2$ [2].

Beta regression models are useful for modeling continuous outcomes that assume values in the interval $(0, 1)$. The Beta regression model is based on an alternative parameterization of the Beta density in terms of the variate mean and precision parameters seen in Equation 1.2 with $0 < \mu < 1$ and $\phi > 0$ [1].

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)(\phi-1)}, 0 < y < 1 \tag{1.2}$$

When fitting Beta regression, we write $y_i \sim \beta(\mu, \phi)$ where ϕ is the precision parameter [1]. For a fixed μ, the larger the ϕ, the smaller the variance of y. The beta regression model is defined in Equation 1.3 [1]. In the model, β is the vector of unknown regression parameters, x_i is the vector of predictors or regressors, and g is a link function [1]. The link function is used for two reasons. First, both sides of the equation assume

values on the real line when a link function is applied to μ_i [1]. Second, it adds flexibility as the link function is chosen to yield the best model fit which in this case is the logit function [1].

$$g(\mu_i) = x_i^\top \beta = n_i \tag{1.3}$$

Since the coefficients of the Beta model are on the logit scale, it is not easy to interpret the model. Marginal effects are useful for interpreting models on a scale of interest rather than estimation, and they can be expressed as the instantaneous rate of change of y with respect to x [4]. Average marginal effects provide a summary that reflects the full distribution of x rather than an arbitrary prediction [4]. Marginal effects are different than predictions because they estimate how an outcome changes as an independent variable changes while other covariates are held constant [4].

Chapter 2

Data and Methods

2.1 Data Description

The primary data used in this exploration is an Excel spreadsheet containing eight supplier directories and eighteen attributes of those directories. Product range is reported as the minimum number of suppliers. Directories are categorized by minimum order requirement status: all, some, or no products. Two variables denote the shipping cost, one for automated shipping cost and one for shipping and handling cost. Transaction fee type and percentage as well as each membership fee is recorded for each directory. The high and low end of average shipping time is reported as two separate variables. Furthermore, shipping method and location are provided. Reliability and reputation are evaluated on a scale of one to ten based on consistent product quality, timely shipping, effective communication, accurate order fulfillment, transparent policies, reliable inventory management, scalability, good track record, competitive pricing, and adherence to regulations. Type of directory customer service and communication with individual suppliers is reported. Return policy and warranty status is collected. Another variable denotes if a directory has an automated inventory system. Lastly, ethical and environmental standards are scored on a scale from one to ten. Ten being an outstanding directory that serves as an example for other companies, and one being an exploitative company that participates in environmental degradation and unethical labor practices. Aiming to select the company that best meets a combination of expectations, this first project is separated into four parts, each exploring one variable in union with another. Then, each section will select a company satisfying the conditions of that section. The directory with the most selections is chosen to be the new directory for Milli's Glassware.

The data used for modeling was collected from the Squarespace platform and was collected over a period of fifteen weeks. The data shows for the past thirty days but was collected weekly. The data contains the page, the number of views the page got that month, the average number of seconds viewing, and the bounce and exit rate. The bounce rate is the percentage of customers that enter the site on one page and exit the site without viewing any more pages, whereas the exit rate is the percentage of views to a given page that resulted in an exit of the website.

2.2 Models and Assumptions

All information on supplier directories was obtained from the websites to collect the most relevant data to optimize product sourcing. Although this approach led to some missing data, the data must be independent.

Next, none of the engagement variables are normally distributed. Views and time on page are skewed count variables, and bounce and exit rate represent a bimodal or trimodal distribution with a significant number of observations at the bounds. First, a Negative Binomial model of views as a function of time on page, bounce rate, and exit rate was fit:

$$\ln(\hat{Views}) = 1.204 - 0.004\,Time\ on\ Page + 0.030BR - 0.007ER \tag{2.1}$$

This model was used because the data fit the assumption of Negative Binomial regression with a mean of 7.889 and a variance of 242.447. The significant predictors of this model are bounce and exit rate. The non-significant predictor was time on page. Similarly, a Negative Binomial model of time on page as a function of views, bounce rate, and exit rate was fit as it met the assumptions with a mean of 48.764 and a variance of 2001.444. The significant predictors of this model were bounce and exit rate, and the non-significant predictor was views.

$$\ln(\hat{TimeOnPage}) = 4.517 + 0.013\,Views + 0.014BR - 0.042ER \tag{2.2}$$

Next, an Ordinary Least Squares regression model of exit rate as a function of page price was fit as exit rate is a continuous variable. Figure 2.1 shows heteroscedasticity and non-normal residuals, so the assumptions of Ordinary Least Squares regression were not met.

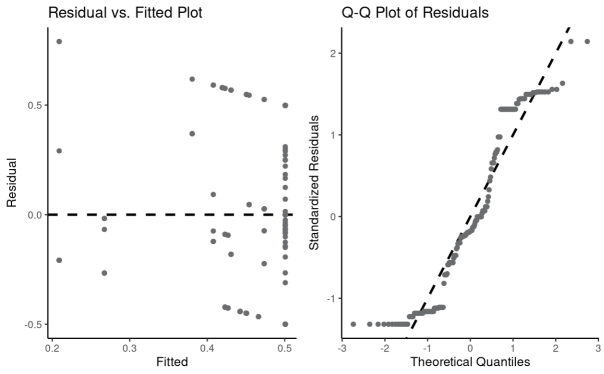

Figure 2.1: OLS Assumptions

Beta regression can handle continuous data that assumes values in the interval $(0, 1)$. Thus, a beta regression model of exit rate as a function of page price was fit. However, the assumptions of this model were also not met which can be visualized in Figure 2.2. The half-normal plot of residuals looks pretty bad, and there are some influential observations in Cook's distance plot. The accuracy of this model fit was only three percent.

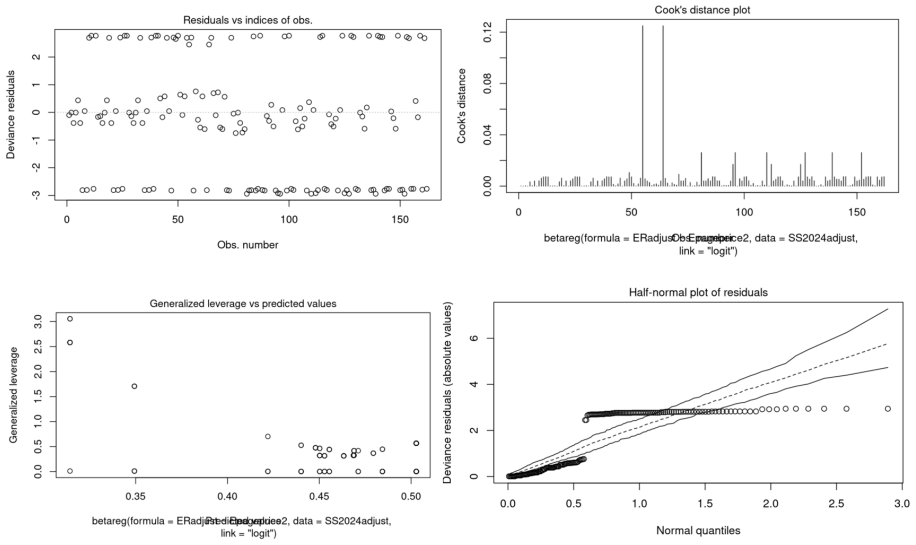

Figure 2.2: Beta Regression Model of Exit Rate as a Function of Page Price

To increase the accuracy, predictors that could influence the exit rate were added, outliers were removed, and interaction terms were considered. The page variable was revalued to create some extra variables: page price, sale page, type of page, number of weeks on the website, and page rank. An interaction was included between page price and sale page. To account for outliers, the page price was adjusted to less than fifty dollars, the exit rate was adjusted to less than or equal to ninety percent, and time on page was adjusted to less than or equal to one hundred seconds. Additionally, the influential observations of Cook's distance were removed. The resulting model:

$$logit(\mu_i) = -11.13 + 0.25\,Price + 0.01\,Seconds + 7.18\,Sale1 + 12.82\,Type3 + 15.90\,Type4 + 11.34\,Type5 + 12.40\,Type6 - 0.03\,Weeks - 4.15\,Rank - 0.52\,Price*Sale1$$

The accuracy of this model fit increased to ninety-five percent. Additionally, all of the predictors were significant except for weeks. The new assumptions check can be viewed in Figure 2.3. Additionally, the average marginal effect estimates are shown in Table 2.1.

Table 2.1: Average Marginal Effects

Term	**Contrast**	**Estimate**	**Std. Error**	**Statistic**	**P-value**
Page Price	Mean (+1)	0.0270948	0.0078668	3.4442038	0.0005727
Time On Page	Mean (+1)	0.0021261	0.0007483	2.8412799	0.0044933
Page Rank	Mean (+1)	-0.2807176	0.0107266	-26.1702632	0.0000000

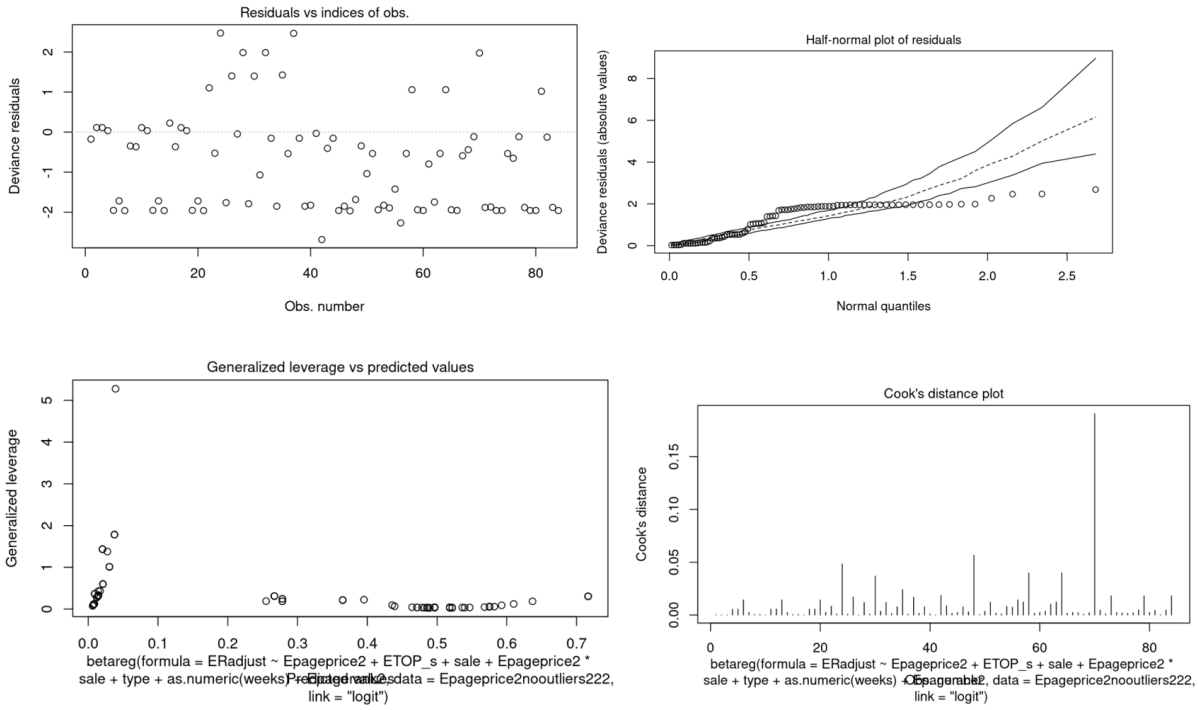

Figure 2.3: Beta Regression Model Assumptions

2.3 Analysis and Results

First, the summary statistics are collected for product range. Then, product range versus reliability and reputation score are visualized in Figure 2.4.

The minimum number of suppliers is one, as it is a direct supplier. The maximum amount of suppliers available in one directory is 300,000. The mean is 66,219 suppliers. Since most directories that have many suppliers source their products internationally from unvetted suppliers, it could be assumed that their reliability and reputation would be scored lower. Based on this scatter plot of these suppliers, this is not true. The smaller suppliers have a wide range of reliability and reputation, and the larger suppliers have a much smaller range that stays between a score of six and eight, which is relatively high. Based on these results, Spocket is in the lead as it has around 20,000 suppliers and a high reliability and reputation score of ten. This amount of suppliers is not comparable to the ones with 200,000 and 300,000. However, is it is closer to the mean than these larger suppliers. It would not be acceptable to choose one of these larger suppliers even though their score is higher than the majority of scores because Spocket has an adequate number of suppliers and products. Next, some supplier directories require purchasing a minimum number of products. In this instance, the bulk order is shipped and the product is then packaged and shipped to

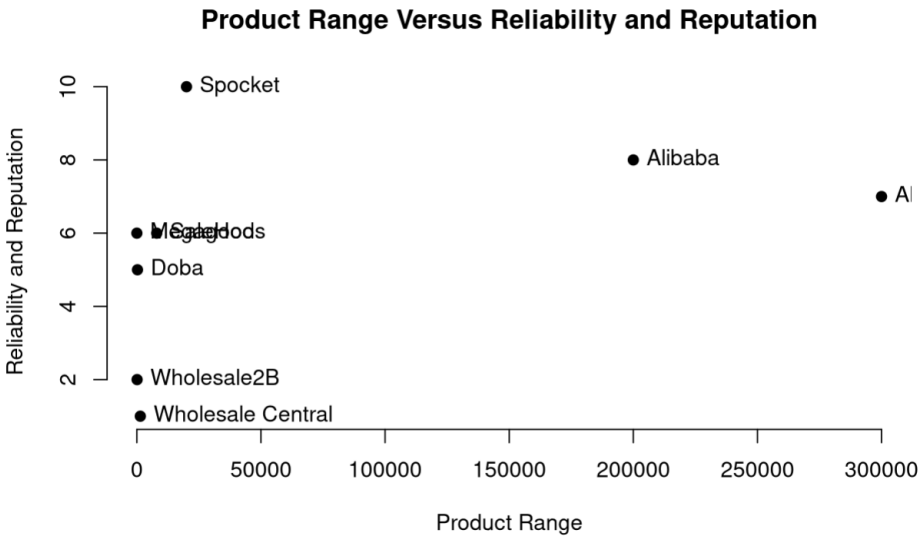

Figure 2.4: Product Range Versus Reliability and Reputation

the customer. One directory has minimum order requirements for all of its products, three have some, and four have none. Some of the directories have automated drop shipping applications, meaning the product bought from Milli's Glassware is automatically shipped to the customer. Due to the convenience of this, the only way packaging and shipping the product myself would be worth the extra time is if fees are lower. This would generate greater profits as bulk orders are already discounted.

The directory that requires minimum order quantities on all of their products did not provide information on transaction fees, despite multiple attempts. However, there is no membership fee, and shipping is expected to cost ten dollars. Since this information is vital and communication is lacking, Wholesale Central has been eliminated. The directories that contain some products with minimum order requirements include payment processing and transaction fees with an average of 3.495 percent. Additionally, these three directories require a membership fee that averages to thirteen dollars and ninety-one cents. The average shipping cost for these products is nine dollars. The directories that do not have minimum order requirements have drop shipping, payment processing, and or transaction fees with an average of 4.33 percent. Additional membership fees average twenty-two dollars and eighty-seven cents. Lastly, shipping averages five dollars.

Based on these results, the transaction fees of companies that don't have minimum order requirements is only 0.838 percent higher on average than those that do. This is an absorbable cost for the automation of shipping. On average, the membership fee for automated companies is eight dollars and ninety-six cents higher. Lastly, the average shipping cost for automated companies is four dollars lower. Since the shipping cost is lower for these automated companies, the higher membership fee is acceptable. The automated companies rank higher. However, these other companies will not be eliminated here because only some of

their products have minimum order requirements. Although, their shipping costs average higher. Now, there are four companies to pick from. Aliexpress has zero transaction fees, no membership fee, and an average shipping cost of three dollars, so this is the selection for this section.

Next, ethics versus reputation can be visualized in figure 2.5. According to this graph, ethics and reputation don't necessarily have a correlation. Spocket is the leader in ethics and reputation, and it maintains an automated inventory system. This system makes the entire process faster while keeping shipping fees low. Customers will be happy because this system is ethical, efficient, and cost effective.

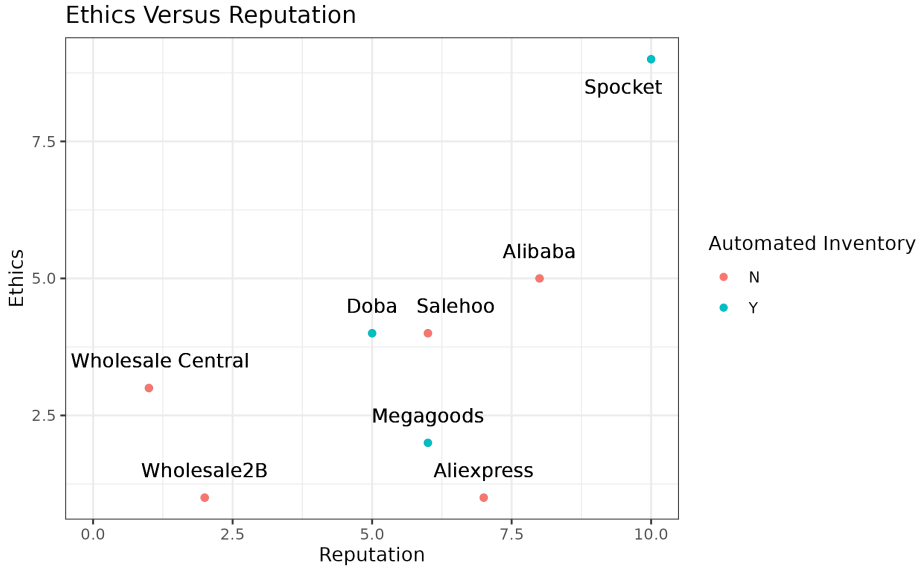

Figure 2.5: Ethics Versus Reputation

Lastly, shipping time and cost can be visualized and separated by ethics score. Figure 2.6 shows the highest shipping time is associated with the lowest shipping cost. However, it also scores the weakest on ethics. On the other hand, Spocket has the highest ethical score with a low shipping time. The only drawback is that the shipping cost is two dollars higher, but this is acceptable based on a much lower shipping time.

These findings indicate that Spocket is the best supplier. Their automated drop shipping system expedites the processing of orders while keeping shipping time and cost low. Additionally, they have around 20,000 reliable suppliers with high ethical standards uncompromising to cost and efficiency. In the dropshipping world, Spocket has a stellar reputation, yet risk is minimized as they are a U.S. and Europe-based company with its own customer service as well as available communication with individual suppliers. Lastly, returns are accepted and warranties are available.

Next, for the skewed count variables, Poisson and Negative Binomial models will be considered. The exit rate will be modeled using Beta regression. The model of views as a function of time on page, bounce rate,

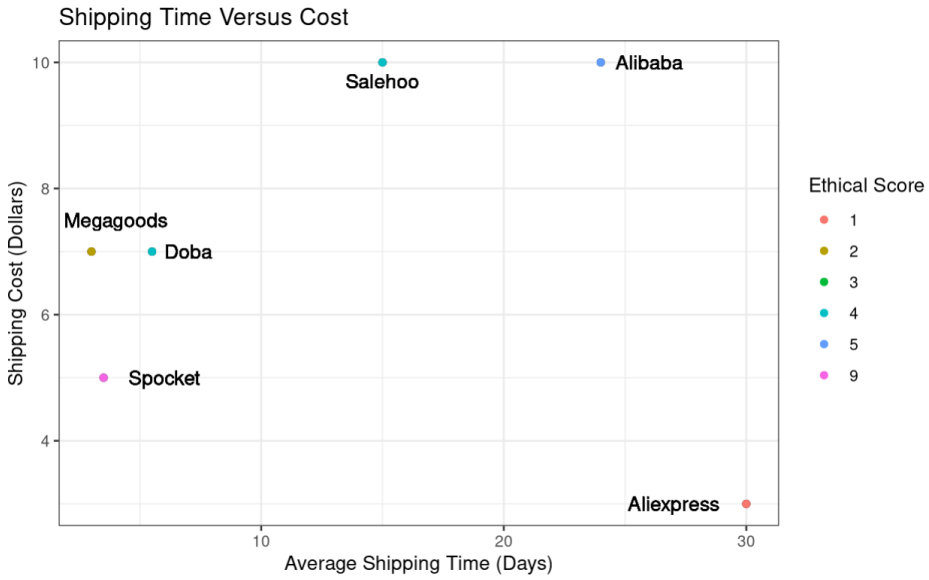

Figure 2.6: Shipping Time Versus Cost

and exit rate can be visualized in Figure 2.7.

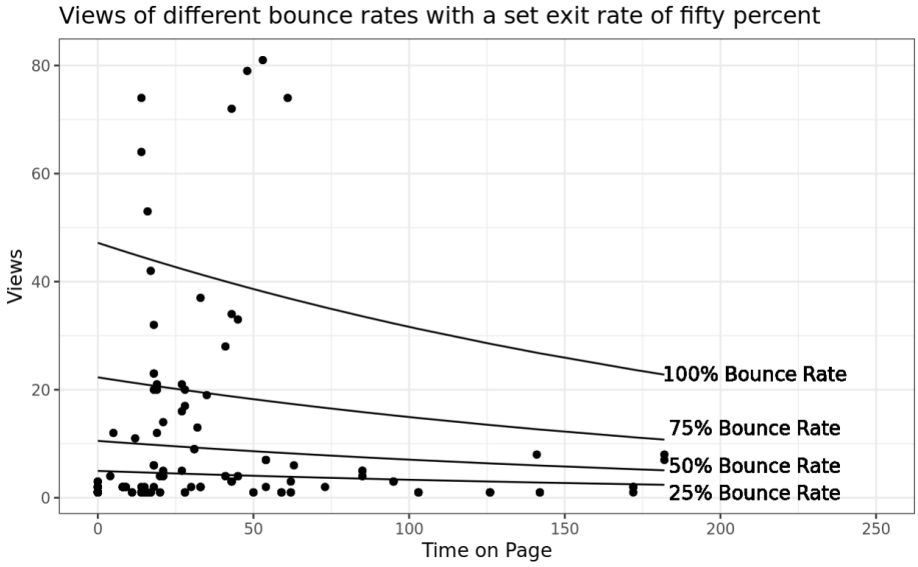

Figure 2.7: Model Views as a Function of Time on Page, Bounce Rate, and Exit Rate

It can be seen that as time on page increases, views decrease. With an increasing bounce rate, the intensity of this trend increases. The trend of bounce rate makes sense as a small number of people will view a page for a long period that has a high tendency to induce an exit of the website. On the other hand, pages with a twenty-five percent bounce rate is ideal. These pages have a wide range of average number of seconds on the page, but the views stay the same. Thus, the aim should be to increase the views of pages with

bounce rates around twenty-five percent. The model of time on page as a function of views, bounce rate, and exit rate can be visualized in Figure 2.8. It can be seen that as views increase, time on page increases. With a decreasing exit rate, the intensity of this trend increases. Similarly, for a low exit rate of twenty-five percent, views increase quickest as time on page increases. The trend of exit rate indicates that time on page stays the same as views increase for a high exit rate. The aim should be to analyze pages with a wide range of views but a time on page that stays relatively the same. This would indicate that something on these pages is causing the majority of customers to be uninterested in the content.

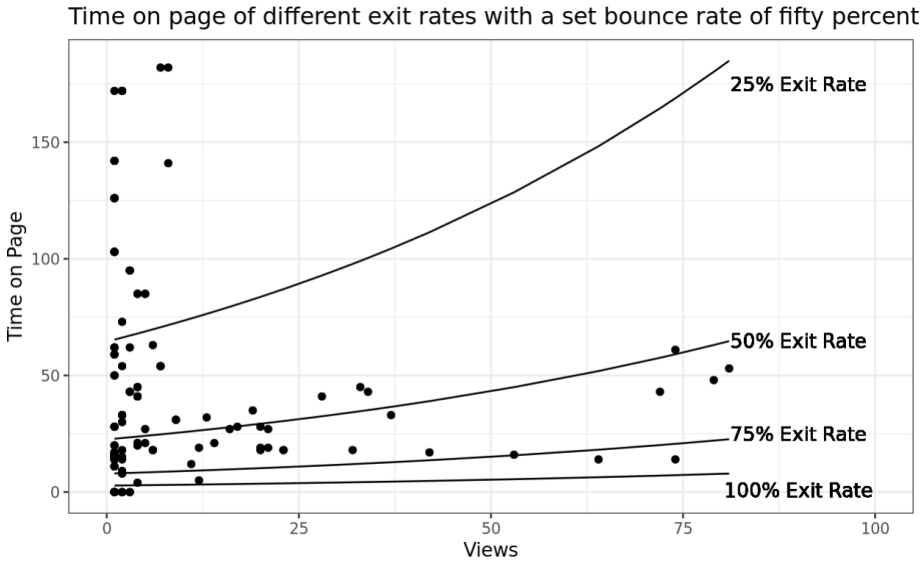

Figure 2.8: Model of Time on Page as a Function of Views, Bounce Rate, and Exit Rate

The first three coefficients of the resulting Beta regression model can be interpreted using average marginal effects. For a one-unit increase in page price, the exit rate increases by 2.7 percent. For a one-unit increase in number of seconds on page, the exit rate increases by 0.2 percent. Lastly, for a one unit increase in page rank, the exit rate decreases by 28.1 percent.

Since there is a significant interaction between page price and if the page was on sale, Figure 2.9 visualizes this portion of the model. For a page that was not on sale, as the price increases, the exit rate exponentially increases. For a page that was on sale, as the price increases, the exit rate exponentially decreases.

Considering the results of the beta regression model, higher ticket items should be placed on sale. This is because an increase in price increases the exit rate, but a page that's on sale has an exponentially decreasing exit rate for higher prices. Additionally, as importance of page increases, the exit rate significantly decreases. This shows that the pages ranked as most important are also viewed that way by customers.

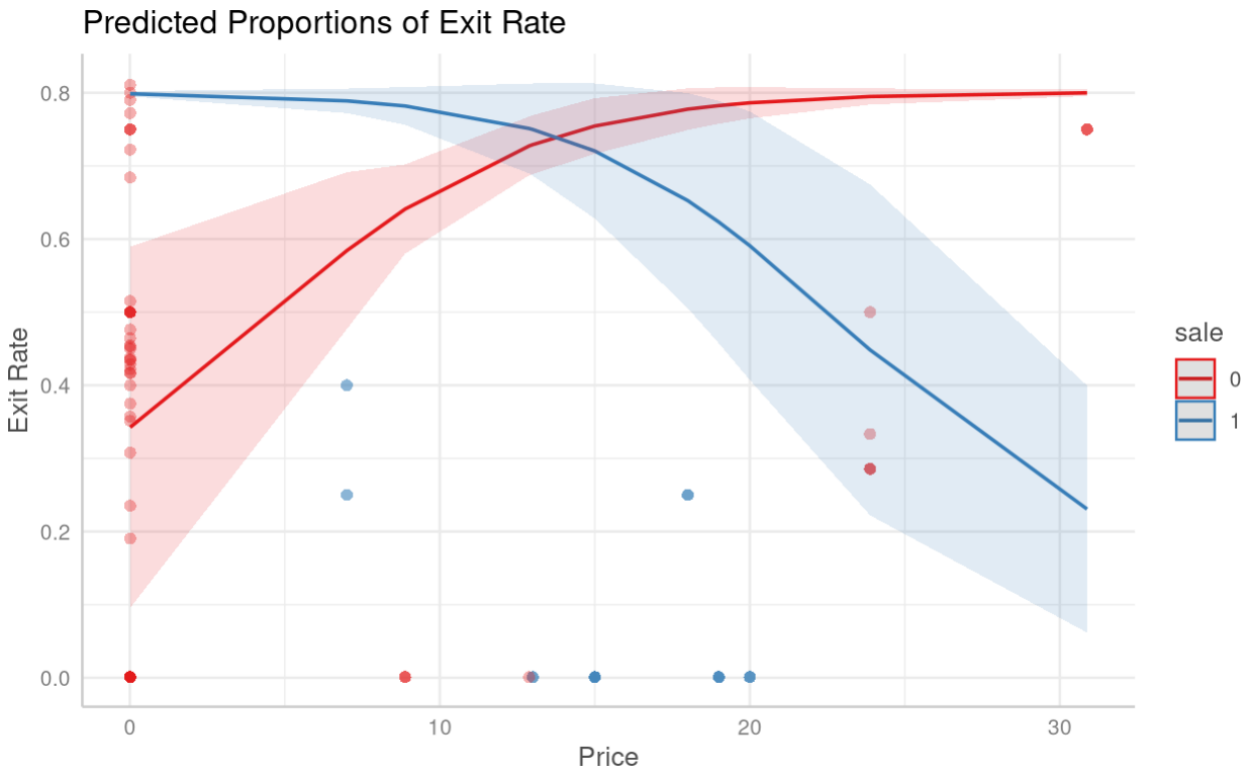

Figure 2.9: Visualization of Interaction Between Price and Sale

Chapter 3

Conclusions

3.1 Summary of Key Findings

These findings indicate that Spocket is the best supplier given the expectations outlined. Their automated drop shipping system expedites the processing of orders while keeping shipping time and cost low. Additionally, they have around 20,000 reliable suppliers with high ethical standards uncompromising to cost and efficiency. In the dropshipping world, Spocket has a stellar reputation, yet risk is minimized as they are a U.S. and Europe-based company with its own customer service as well as available communication with individual suppliers. Lastly, returns are accepted and warranties are available.

Milli's glassware opened using Squarespace on March 6, 2024. Data has been collected from the Squarespace platform since March 19, 2024. All modeling was done using this data spanning over fifteen weeks. For a page that has a twenty-five percent bounce rate, as time on page increases, number of views stays the same. However, for a page that has a one-hundred percent bounce rate, as time on page increases, views decrease. For a page that has a twenty-five percent exit rate, as views increase, time on page increases. Alternatively, for a one-hundred percent exit rate, as views increase, time on page stays the same. Thus, one aim should be to increase the views of pages with bounce rates around twenty-five percent.Another aim should be to analyze pages with a wide range of views but a time on page that stays relatively the same. This would indicate that something on these pages is causing the majority of customers to be uninterested in the content.

Based on the most relevant model results, items that are higher in price are preferred to be on sale rather than those of a lower price. Additionally, for a one unit increase in page rank or the importance of the page, the exit rate decreases by 28.1 percent indicating that customers find important pages the most important.

3.2 Suggestions for Further Study

Zero One Inflated Beta (ZOIB) and Ordered beta are relevant alternatives to beta regression as these models can handle data at the bounds. ZOIB can get complicated with lots of parameters to estimate. Ordered beta is a special case of ZOIB that introduces dependence among the probabilities in ZOIB. Ordered beta combines ordinal regression and beta regression to predict the dependent probabilities of observing exactly zero, exactly one, or something in between. Aditionally, some non-parameteric alternatives for when the beta regression assumptions are not met could be kernel regression or a generalized additive model.

Lastly, the collection of more data and the creation of a shiny application in R are good ideas.

Bibliography

[1] F. Cribari-Neto and A. Zeileis. Beta regression in r. *Journal of Statistical Software*, 34(2), 2010.

[2] W. Gardner, E.P. Mulvey, and E.C. Shaw. Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial mode. *Psychological Bulletin*, 118(3):392–404, 1995a.

[3] M. Hayat and M. Higgins. Understanding poisson regression. *Journal of Nursing Education*, 53(4):207–215, 2014.

[4] T.J. Leeper. Beta regression in r. *The Comprehensive R Archive*, 2024.

[5] Tripathi. *Learn Business Analytics in six steps using SAS and R A practical, step-by-step guide to learning business analytics*. Apress, 2016.