International Journal for Quality Research 18(3) 731–744
ISSN 1800-6450

Kristina Zogović[1]

Article info:
Received 25.08.2023.
Accepted 09.03.2024.

UDC – 005.6
DOI – 10.24874/IJQR18.03-06

UNVEILING UNSEEN INSIGHTS: QUALITY MANAGEMENT AND BUSINESS OPTIMIZATION THROUGH THE ANALYSIS OF PRODUCT SALES PATTERNS WITH CATEGORICAL AND CONTINUOUS PREDICTORS IN A UNIQUE DATASET

Abstract: *This investigation explores e-commerce product sales patterns, integrating Quality Management and Business Optimization perspectives. We analyze product listings, ratings, and sales metrics using a unique dataset from data.world.com, sourced from Wish.com. Our predictive model unveils correlations among categorical and continuous predictors, spotlighting their role in predicting unit sales. Employing robust linear regression, we assess predictor significance via t-tests, hypothesis evaluations, and ANOVA. Model selection, guided by AIC, identifies the optimal fit. Outliers, influential points, and assumptions are evaluated. Employing data visualization, we present results comprehensively. This study empowers stakeholders with insights into e-commerce dynamics, Quality Management, and Business Optimization for informed decision-making.*

Keywords: *E-commerce, sales patterns, quality management, business optimization, dataset analysis, predictive modeling*

1. Introduction

This research is based on applying statistical linear regression to e-commerce, laying the foundation for its investigation. Our investigation explores the intricate connections between variables, encompassing nominal and categorical attributes and the geographical distance from the individual's home where fraudulent transactions occur. The dataset underpinning this study was sourced from data.world.com, featuring a comprehensive collection of 43 columns and 1,573 rows. This expansive dataset enables us to delve into the nuanced relationships that shape the occurrence of fraud transactions based on geographic location.

While previous research has focused mainly on analyzing product listings, this dataset offers a unique blend of information, including product ratings and sales performance metrics. This comprehensive approach empowers us to uncover correlations and patterns associated with product success, transcending the confines of conventional datasets. One intriguing investigation avenue is validating that individuals are sensitive to price drops, specifically when comparing discounted prices to the original retail price.

Corresponding author: Kristina Zogović
Email: zogovic.kristina@gmail.com

Furthermore, our study seeks to identify top product categories, shedding light on the products with the highest sales performance. The relationship between product ratings and success is another captivating aspect we intend to explore, coupled with an assessment of the potential influence of pricing on this relationship. Notably, the dataset is derived from the Wish platform and localized in the French context, adding a layer of specificity to our analysis (https://www.wish.com/).

A significant portion of our methodology revolves around model comparison and selection. Using the AIC as a guiding criterion, we evaluate the fit of different models and compare their explanatory power. The adjusted R-squared values for each model are calculated and contrasted to determine which model best encapsulates the relationships within the dataset. This exploration extends to scrutinizing potential outliers and influential points and assessing multicollinearity and other key assumptions underlying our analysis.

To enhance our presentation of results, we employ data visualization techniques, fostering a clearer understanding of the outcomes. By examining the interplay of variables, interpreting the results of our chosen models, and placing them in the context of AIC-based selection, we aim to contribute to the field's comprehension of the factors influencing fraudulent transactions' geographic patterns. This research aims to illuminate these associations, empowering decision-making, fraud detection, and prevention strategies within an evolving landscape.

2. Research and background

2.1 Logistics level of cross-border e-commerce

In the context of rapid technological advancements, widespread computer adoption, and the increasing interconnection of global economies, the landscape of e-commerce experiences substantial growth. Within this dynamic environment, cross-border transactions have gained prominence, with companies like Wish.com benefiting from technological progress, elevated living standards, and expanding international trade (Fang, 2022). Notably, product quality assurance and adherence to standards becomes critical factor in this cross-border e-commerce arena, significantly influencing consumer trust and regulatory alignment. This study focuses on the pivotal role of e-commerce platform managers as key regulators, exerting a substantial impact on the sector's trajectory. Through the application of multivariate linear models, this research comprehensively explores the factors shaping e-commerce platform development involving various stakeholders. The study presents strategic recommendations to enhance e-commerce platform quality and performance by addressing prevalent challenges and fostering sustained success in this ever-evolving domain. Since the outbreak of the new crown epidemic, global trade has shrunk to a great extent. Although logistics has also been affected to a certain extent, such an epidemic brings excellent prospects for the development of cross-border e-commerce.

To improve the logistics level of cross-border e-commerce, statistics, and big data technology can effectively enhance its logistics level (Wang, 2023).Using existing statistical models adequately and understanding and implementing them properly is one of the challenges that international logistics management faces today.

2.2 Predictive analytics

In e-commerce, the privacy of inquiry responses often confines information to private chat interactions between customer service agents and website visitors, with such data seldom accessible to the public. Publicly accessible content usually revolves

around online consumer reviews, shaping the landscape for discourse analysis in e-commerce research. As invaluable "sales assistants," these reviews help consumers find products that meet their needs (Pan, 2019). Scholarly studies often concentrate on online consumer reviews, recognizing their influence on consumer behavior (Vignardi, 2018). Product reviews have historically wielded more influence on customers than website information (Turkson et al., 2021). In our research, we mainly investigate the crucial association between categorical and continuous predictors and the volume of products sold, homing in on the relationships that substantially impact sales figures. This exploration aims to uncover the nuanced dynamics contributing significantly to product sales.

In the research context, a categorical variable is characterized by its capacity to assume a limited and typically predetermined range of values. This assigns each observation unit to a specific group or nominal category, contingent on qualitative attributes. Velleman and Wilkinson (1993) noted that categorical data analysis encompasses the scrutiny of data wherein the response variable has been partitioned into mutually exclusive, ordered, or unordered categories. Imrey and Koch (2005) further delineate categorical variables with a measurement scale comprising non-numerical categories, organizing observations into common trait-sharing groups. Kadi and Peker (2015) introduced the concept of categorical data, classifying measurement scales into nominal, ordinal, interval, and ratio categories, accompanied by pertinent analysis techniques. Over time, scholars like Wolfinbarger and Gilly (2001), Velleman, Wilkinson, (1993) have proposed hierarchical classifications, including grades, ranks, counted fractions, counts, amounts, and balances, while addressing the adequacy of prescribed analysis methodologies.

Qualitative research, qualitative data, and qualitative variables are primarily concerned with subjective aspects such as opinions, feelings, and experiences, which can vary significantly among individuals. The terms "variable" and "data" are used interchangeably throughout this discourse. Qualitative variables are categorized into two main types: ordinal and nominal. A comprehensive understanding of these ordinal and nominal variable concepts is essential to demystify their application, analysis, interpretation, and derivation of statistical inferences. This understanding also aids researchers in selecting the appropriate statistical analyses based on assigned values.

Advanced tools renowned e-companies employ to enhance sales are fundamentally rooted in statistical models and data science. Predictive analytics, a vital component, empowers businesses to anticipate customer behavior, optimize inventory management, and reduce customer churn (Kuperenko, 2021). Utilizing predictive analyses fueled by data modeling, machine learning, artificial intelligence, data mining, and deep learning algorithms, companies such as Netflix and Amazon revolutionize shopping experiences (Wen, 2022). These predictive capabilities enable tailored recommendations, dynamic pricing, and personalized interactions based on individual user behaviors (Fang & Xiaheng, 2015). Examples from Macy's, Amazon Go, and Airbnb highlight how predictive analytics refines sales strategies and logistics, resulting in substantial sales growth (Oracle, n.d.; Amazon, n.d.; Airbnb, n.d.). Ultimately, the fusion of cutting-edge technology and data-driven insights drives heightened sales performance and customer engagement across the retail landscape (Fang & Xiaheng, 2015; Wang, 2021).

3. Methodological framework: unraveling e-commerce product sales

To understand the intricate dynamics of e-commerce product sales, we embarked on an

analysis rooted in a distinctive dataset obtained from data.world.com. This repository offers an unparalleled amalgamation of product listings, product ratings, and sales performance metrics, painting a comprehensive picture of the e-commerce landscape. This dataset comes from the Wish.com platform and comprises 43 columns and 1,573 rows, providing unprecedented insight. This rich compilation empowers us to venture beyond conventional analyses and unravel the correlations and patterns beneath the surface of product success.

3.1 Essential variables

In the context of our study, essential variables are introduced to provide context for our investigation. Our dataset incorporates five categorical variables (Imrey & Koch, 2005; Cox, 1970) per record, each indicating the fulfillment of specific prerequisites (coded as '1' for satisfied and '0' for unsatisfied). Additionally, six numerical features encapsulate significant attributes of the products, including price (X1), units sold (X2), uses of ad boosts (X3), product rating (X4), rating count (X5), presence of local product badge (X6), product quality badge (X7), product variation inventory (X8), merchant profile picture (X9), shipping express status (X10), and countries shipped to (X11). Our research strategy unfolds through two distinct phases: predictive modeling and subsequent analysis (Yao & Ma, 2023). The predictive model intertwines categorical and continuous predictors, harnessing their collective strength to forecast unit sales (Preissor & Koch, 1997). We leverage the R programming language for modeling, capitalizing on its statistical analysis and hypothesis-testing capabilities.

3.2 Predictive modeling

Guided by the research strategy, our investigation unfolds in two distinct phases: predictive modeling and subsequent analysis (Alkan et al., 2021). This approach enables us to predict product sales and gain deeper insights into the underlying dynamics. The predictive model is meticulously constructed by weaving together categorical and continuous predictors. By harnessing the combined power of these elements, we aim to anticipate the number of units sold, a critical metric in e-commerce.

By meticulously following our experimental protocol, characterized by variable selection, predictive modeling, and comprehensive statistical analyses, we unveil hidden insights driving e-commerce product sales. This methodological rigor allows us to navigate the complexities of online retail and offer nuanced insights into the determinants of success. This approach allows us to validate the consistency of our findings and ensure that the selected model adequately represents the observed patterns. As we navigate our modeling journey, refinement becomes paramount. Through rigorous testing, including individual t-tests and hypothesis assessments, we endeavor to unravel the significance of predictors (Zogovic et al., 2022). This meticulous process sheds light on the potential influence of variables on product sales. Moreover, the role of interaction terms within our model takes center stage, demanding meticulous evaluation through techniques like analysis of variance (ANOVA) and hypothesis testing.

Central to our approach is the utilization of multiple linear regression. We construct and present full regression lines, integrating two categorical predictors, their interaction, and two continuous predictors. This mathematical representation encapsulates the intricate relationship between predictor variables and the dependent variable - units sold. Interaction effects add another layer of complexity, allowing us to explore more nuanced relationships within the dataset.

3.3 Best fit and outliers

The significance of each variable, intersection, and observed model is crucial to our investigation. This involves evaluating various statistical outputs, mainly focusing on the p-values and beta coefficients. A low p-value in the F-test reinforces the relationship between explanatory variables and the response variable, i.e., the number of units sold (https://www.alpharithms.com/correlation-matrix-heatmaps-python-152514/). By scrutinizing these results, we establish the strength of relationships between explanatory variables and the response variable - units sold. These insights contribute to a comprehensive understanding of our model's implications.

To further understand the interplay between variables, we delve into correlation analysis. This includes focusing on the relationship between rating count and units sold, shedding light on potential associations. Additionally, diagnostic checks are meticulously conducted to validate the underlying assumptions of our regression models. We evaluate the linearity of relationships and assess the normal distribution of variables, ensuring the robustness of our analysis.

In our pursuit of robustness and accuracy, we address the potential impact of outliers and influential points on our analysis. A meticulous identification process leads us to pinpoint 66 outliers within our dataset. To further enhance the integrity of our analysis, we curate a refined dataset devoid of these outliers, aptly named "two."

3.4 Visualization and Assessment

The power of visualization is harnessed to enhance the communicative strength of our findings. By integrating the ggplot2 package, we generate compelling scatter plots that visually capture the relationship between units sold and rating count. These visualizations offer a clearer understanding of the data patterns and reinforce the narrative presented by our model.

The integration of visual assessments complements the overall evaluation of our model, enhancing its relevance and reliability. The assessment of model fit is a crucial step in our analysis. We determine which model best captures the relationships within the dataset using the AIC criterion. The adjusted R-square values provide insight into the explanatory power of each model. The comparison between models through AIC and R-square enhances our understanding of the most appropriate model for our research context. This approach allows us to validate the consistency of our findings and ensure that the selected model adequately represents the observed patterns.
Our investigation into e-commerce product sales encompasses a multifaceted approach, combining categorical and continuous predictors within a predictive model. By exploring the interactions between these variables, we aim to unravel the complex fabric of factors influencing sales.

Through systematic analysis, model refinement, and rigorous assessment, we equip ourselves with the tools to comprehend the intricate dance that underlies e-commerce success. This knowledge contributes to the broader landscape of quality management and business optimization within the dynamic realm of online retail. Through the meticulous execution of this methodological framework, fortified by the rigorous selection of variables, predictive modeling, and comprehensive statistical analyses, we embark on a journey to uncover the factors that drive e-commerce product sales.

Our systematic approach equips researchers with the tools to reproduce and extend these findings, fostering a comprehensive exploration of product success within the dynamic realm of online retail.

4. Analyzing results, discussions, and interpretations

4.1 Predictive Model

The predictive model, a pivotal component of our inquiry, uncovers substantial relationships between predictor variables and the response variable: the quantity of units sold (Pardoe, 2021). Our model, methodically crafted by integrating categorical and continuous predictors, showcases a remarkable capacity to elucidate variations in sales performance. We delve into the exploration of the multiple regression model:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5 X_3 X_4 \quad (1)$$

Examining beta coefficients enhances our nuanced comprehension of the effects of variables (Zogovic, et al., 2022). Our findings accentuate the substantial sway that specific predictors, such as price and rating count, exerted on the number of units sold. These coefficients and their corresponding confidence intervals furnish a sturdy framework for interpreting the magnitude and direction of predictor influences. Interestingly, the correlation between the utilization of ad boosts and sales performance unveils a significant positive impact, further affirming the importance of this variable. Introducing a factor variable into a regression introduces distinct intercepts at every level of the factor. In our instance, this signifies our intention to model the number of units sold as follows:

$$799.236 - 50.007X_1 + 4.246X_2 + 416.033X_3 + X_4 - 1407.174X_3{:}X_4. \quad (2)$$

Prior to employing this formula for projecting future sales, it is imperative to ensure the statistical significance of the model. This entails confirming the presence of a substantial statistical relationship between the predictors and the outcome variables, as well as the model's adeptness in fitting the available data. Notably, the results of the F-test provide compelling evidence regarding the overall significance of the model, as evidenced by a p-value of less than $2.2e^{-16}$. This underscores the vital significance of explanatory variables such as price, rating count, uses of ad boosts, and shipping express status. Such outcomes underscore the pivotal role of these variables in shaping product sales dynamics within the realm of the e-commerce platform.

4.2 Interaction Term Evaluation

Our model's interaction term provides valuable insights into its contribution to predictive power. Our rigorous analysis of variance (ANOVA) and hypothesis testing reveals that the interaction term involving ad boosts and shipping express status does not significantly improve the model's explanatory power (Hoffmann, 2021). Consequently, this outcome drives us to streamline the model by emphasizing more influential predictors.

The process of individual t-testing serves to determine the significance of pairwise differences. Initially, we examine whether these coefficients significantly deviate from zero through the partial F-test. After confirming the significance, we proceed to perform pairwise t-tests. The overarching objective of the model is to ascertain whether the evaluated X_i attributes (i=1..., 4) and the interaction term predict the number of units sold. In this context, to assess the significance of the interaction term, we employ a hypothesis test (Namavari, 2019). In cases where the seller does not refrain from boosting their product within the platform, our model simplifies to:

$$\text{units sold} = 799.2 - 50.01 * (\text{price}) + 4.246 * (\text{rating count}).$$

4.3 Possibility of removing both continuous predictors

Our investigation proceeds with a continued focus on equation (2). In statistical analysis,

the t-statistic and its corresponding p-value serve as crucial tools for determining the statistical significance of a given predictor's relationship with the outcome variable, as evidenced by its beta coefficient's significance. Following this line of inquiry, our subsequent step aimed to employ an appropriate hypothesis test to assess the possibility of removing the interaction term represented by equation (1). Upon scrutinizing our findings, the computed p-value of p=0.7657 surpasses the predetermined significance threshold (α=0.05), resulting in the non-rejection of the null hypothesis (H_0: b_5=0). This outcome indicates that the interaction term, which encompasses the interplay between categorical variables uses_ad_boosts and shipping_is_express, lacks substantial statistical significance. Consequently, we eliminate this interaction term from our predictive model, acknowledging its limited contribution to additional explanatory power.

This refinement leads to a more focused and streamlined model, shedding light on the core determinants of e-commerce product sales performance within the dynamic online retail landscape. This reaffirms the pivotal role of the t-statistic and p-value in unraveling intricate predictor-outcome relationships, guiding us toward a more insightful and concise analytical framework (Hoijtink et al., 2019; Rifada et al., 2023). With this understanding, we proceed to equation (3):

units sold=798.692−49.809∗(price)+4.246∗*(rating count)+413.416∗(uses ad boos)−706.492∗(shipping is express)

Subsequently, our investigation delved into the possibility of removing both continuous predictors simultaneously. To assess this, we scrutinized the ANOVA table, comparing the whole model with the reduced one (H_0: b_1=b_2=0). The p-value of <0.001 (below 0.05) led us to reject the null hypothesis. This outcome indicates strong evidence suggesting that the continuous predictors (price and rating count) significantly contribute to predicting the number of units sold, thus warranting their retention within our model. Consequently, equation (3) presents our finalized multiple regression model. In order to identify significant predictors, we utilized t-test values obtained through the summary command in R (H0: bi=0, where i is from the set {1,2,3,4}). The calculated p-value in this context was p=0.06078, which exceeds 0.05, leading us to fail to reject the null hypothesis. This suggests insufficient evidence to claim that price and shipping express status predict the number of sold units. Conversely, we possess enough evidence for other predictors to reject H0 and retain them in our model. Thus, our ultimate model is:

units sold=368.583+4.247∗(rating count)+ +443.802∗(uses ad boosts). (4)

The intercept (b_0) holds a value of 368.583. This figure is the projected number of sold units when the rating count is zero, and no ad boosts are employed. Given that the categorical variable "uses ad boosts" can only assume values of 1 or 0 (Massidda & Marrocu, 2018; Yao & Ma, 2023; Zogovic, 2023), if the seller opts not to utilize ad boosts (uses ad boosts=0) to enhance their product within the platform, our model simplifies to:

units sold=368.583+4.247∗ (rating count) (5)

When there is no rating count, we can expect about 369 units to be sold. We expect the number of sold units to increase by about four (4.25) every time the rating count increases by one (unit).

In the case that the seller pays (utilizes ad boosts =1) to boost his product within the platform, then our model would be as follows: units sold=812.385+4.247∗(rating count).

Moreover, for the rating count, we have interpretation, as well as for the intercept (just a different value).

According to the obtained value from the AIC command, the model without

interaction term (model without continuous predictor) has a slightly better fit than the full model, so we agree that model (4) is the best model.

4.4 Adjusted R-squared (R^2) values analysis and Outliers

4.4.1 Adjusted R-squared (R^2)

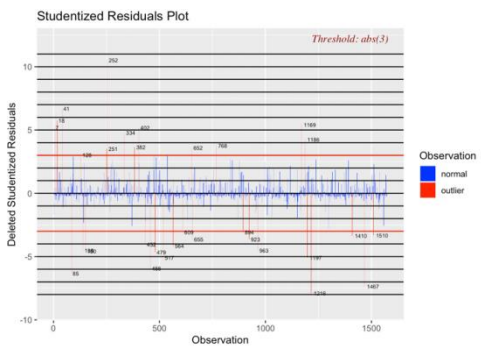
After analyzing the Multiple R-squared (R^2) values, we found that the full and final models had R^2 values of 0.8095 and 0.8096, respectively. As there is no significant difference between the two, the quality of the model has not been compromised.

4.4.2 Reliable predictor

The high values of R^2 indicate that we can expect minimal variation in the model's estimate in practice, making it a reliable predictor. The best R^2 has the model without interaction term (4).

4.4.3 Outliers

The "best model", model (4), as determined above, can be used to identify potential outliers. The Plot 1 illustrates this.



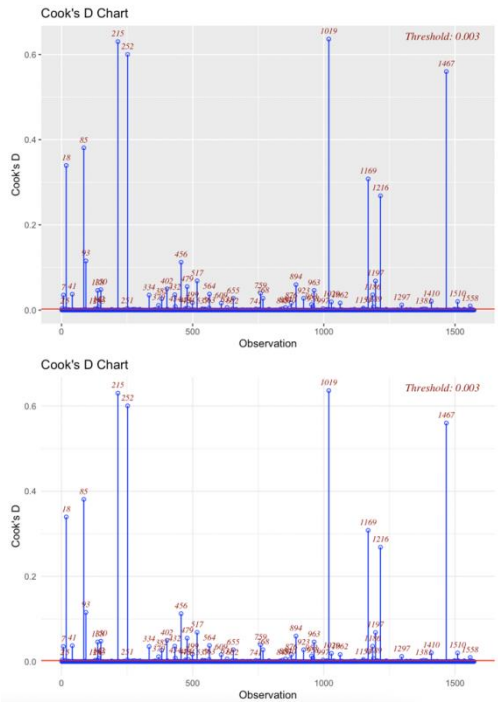
Plot 1. Studentized Residuals Plot

From the diagram above, we can see standard (no outliers) elements from our data set between two red lines and outliers outside of the range. There are 66 outliers.

We created a new dataset without outliers and named it "two."

We used a graphical approach and command:

ols_plot_cooksd_chart(m1) + theme_ minimal() to identify potential influential/leverage points. Output can be seen on the Plot 2.



Plot 2. Cook's D Chart

From Cook's Distance plots (above), we see outliers as "spikes."

4.4.4 Formally assess multicollinearity

Generally, a VIF above 10 indicates a high correlation and is cause for concern. A rule of thumb for interpreting the variance inflation factor: 1. 1 ~ not correlated, 2. Between 1 and 5 = moderately correlated, and 3. Greater than 5 = highly correlated (https://wanderluce.com/news/how-do-you-calculate-vif-in-sas/).

Spearman's correlation coefficient (ρ, Greek later) can take values from +1 to -1. A ρ of

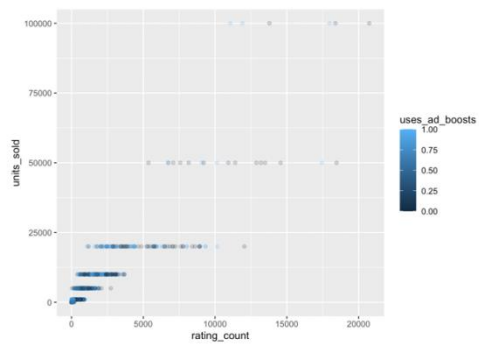
+1 indicates a perfect association of ranks, a ρ of zero indicates no association between ranks and, ρ of -1 indicates a perfect negative association of ranks. The closer ρ is to zero, the weaker the association between the ranks.

Spearman's correlation coefficient measures the strength and direction of association between two ranked variables (https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/labourproductivity/articles/regionalandsubregionalproductivityintheuk/july2022). In all evaluated correlations below, we can conclude a weak (or no) correlation between the two ranks of the evaluated pairs. The strongest correlation of all observed is ρ =-0.042, between ranks of rating_count and uses_ad_boosts. We had positive correlations between price and shipping_is_express (0.129) ranks and price and rating_count (0.035). These ρ were very low/close to zero, so we concluded the absence of correlation.

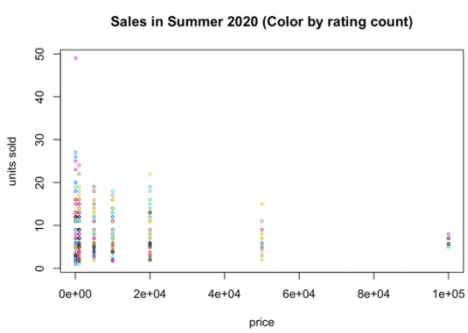
After analyzing the data set without outliers, we got an *"Error in vif. default* ()" message. This error typically occurs when multicollinearity exists in a regression model. Two or more predictor variables in the model are highly (or perfectly) correlated (https://www.statology.org/r-aliased-coefficients-in-the-model/). Obtained VIF value for the new model (created with data set without outliers) indicates a high correlation between categorical variables shipping_is_express and uses_ad_boosts. We needed to fit the regression model again and leave out one of these two variables to fix this error. Another way was to create two models in which we would include one of these variables. We were getting a similar output (warning) with Pearson's product-moment correlation coefficient (0.8994637). Therefore, If we want to build our model over the set, two variables, shipping_is_express, and uses_ad_boosts, cannot be in the same model together. We discussed this with collaborators to see if they want to - (1) drop one of the predictors altogether - (2) report one model with and, separately, one model with uses_ad_boosts.

5. Visualization and assessment of the regression assumptions

Using the *ggplot2* command, diagnostic assessments were made to confirm the validity of our modeling method. Furthermore, examining model assumptions validated our chosen methodology's appropriateness, ensuring our results' reliability. We identified the variables that influence e-commerce product sales and their interrelationships through statistical analyses and meticulous modeling. The visualization of our model using the ggplot command revealed that paid boosting of a product within the platform results in a higher number of units sold on average than when it is not boosted for any rating count value.



Plot 3. Visualization of units_sold vs rating_count



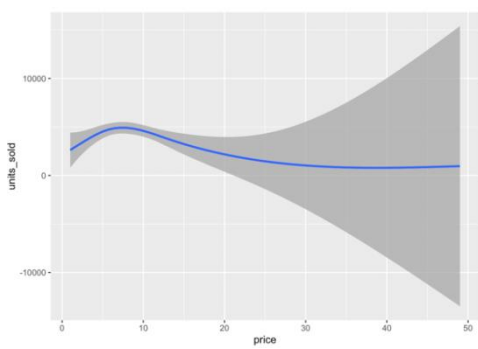
Plot 4. Visualization of Units_sold vs price

We should consider using separate intercepts for different levels of uses_ad_boosts in our model. Plot 2 shows that most units sold are the cheapest products. Pearson's product-moment correlation coefficient reveals a strong positive correlation (0.8994637) between rating count and units sold, emphasizing the importance of maintaining high product ratings for better sales performance. However, we must check if our data satisfies four assumptions before using the coefficient. The correlation between higher ratings and higher sales is evident from the diagram, which shows an increasing trend from zero to around 50, a stable trend until 80, and a decreasing trend after that, with most of the graph being interpreted as a decreasing line.

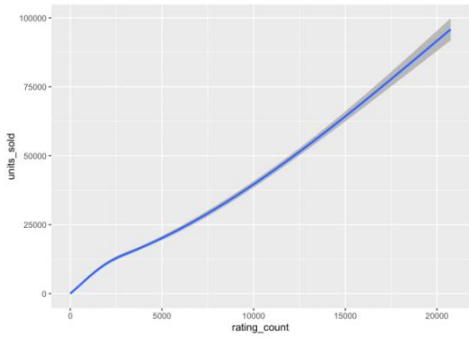
We can see no linear relationship between those two data in the price vs. sold units plot, but another data (rating count vs. units sold) is approximately linear. Since linearity is one of the assumptions of linear regression models, this can pose a problem.

To check for normality, we used the ggplot command. There are some promising trends in the histogram.

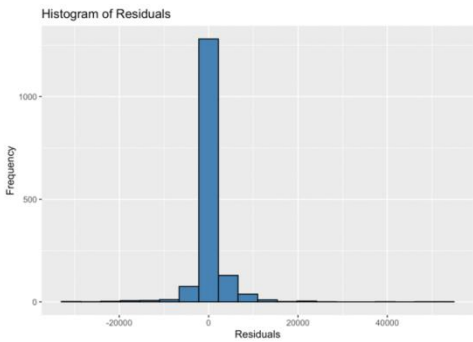
As shown in Plot 6 and Plot 7, we examine the units_sold and summer_products data to see what a more typical analysis of linear model diagnostic plots might reveal.



Plot 5. Check for linearity units_sold vs price

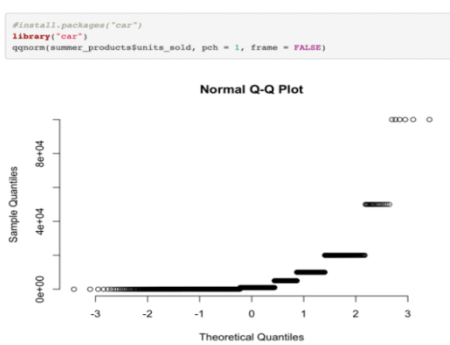


Plot 6. Check for linearity units_sold vs rating_count

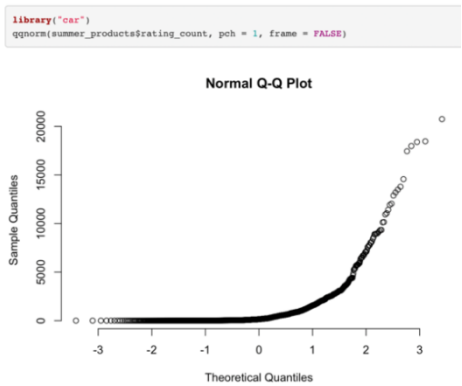


Plot 7. Histogram of Residuals and check for data normality

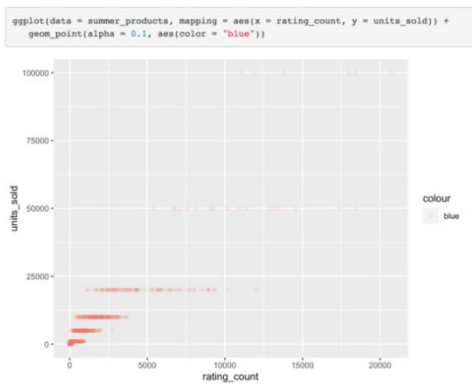
It can be observed from the obtained QQ plots that neither of the investigated predictors comes from a normally distributed dataset. In Plot 8, the geom_point part is included in our ggplot function. Here, it can be seen that many data points are in the interval from zero to 150, which supports our previous findings.



Plot 8. QQ plot for units_sold



Plot 9. QQ plot for rating_count



Plot 10. Visualization of rating_count and uits_sold

6. Conclusion

In conclusion, our rigorous investigation into the multifaceted dynamics of e-commerce product sales has yielded invaluable insights that carry far-reaching implications for the industry. By meticulously examining various variables, we have unveiled the critical determinants that significantly influence product success within online retail. Anchoring our study is a unique dataset sourced from data.world.com, a repository that provides a comprehensive collection of product listings, ratings, and sales performance metrics. This dataset has offered us a distinctive vantage point for analysis, enabling us to paint a comprehensive picture of the interplay between variables and sales outcomes.

We have uncovered intricate correlations and patterns in the sales dynamics by combining categorical and continuous predictors in our predictive model. This model enhances our understanding of the complex relationships between various attributes and sales performance and equips us with a potent tool for predicting future sales with a high degree of accuracy.

Our findings underscore the pivotal roles played by variables such as price, rating count, and the use of ad boosts in driving product sales within the e-commerce landscape. These insights provide actionable guidance to industry practitioners seeking to refine their strategies and enhance decision-making in the competitive online market.

Furthermore, our research has made significant strides in advancing the methodological frontiers of data analysis. We have rigorously assessed the significance of predictors and interactions by leveraging robust statistical techniques, including individual t-tests, hypothesis testing, and ANOVA. By meticulously evaluating the role of the interaction term and subsequently streamlining our model, we contribute to the refinement of predictive modeling practices, setting a precedent for insightful and efficient data analysis.

As we conclude our investigation, it is worth noting that the quality of our model remains uncompromised, as evidenced by the consistent R-squared values across both complete and final models. This underscores the reliability of our model's estimates in practice. Identifying and addressing outliers and influential points has further enhanced the robustness of our analysis.

Our study's contributions echo throughout the e-commerce landscape. Practitioners can leverage our findings to optimize marketing strategies, improve product placement, and make informed decisions that maximize sales potential. For researchers, our study offers an enriched understanding of the intricate relationships between predictor

variables and product sales, paving the way for developing more sophisticated predictive models and methodologies.

In navigating the ever-changing realm of online retail, our research serves as a guiding beacon. It contributes to quality management and Business optimization through analysis, bridging the gap between theoretical knowledge and practical application. By unearthing the underlying mechanisms of e-commerce success, our study fosters innovation, data-driven decision-making, and a deeper comprehension of the dynamic forces shaping the industry's trajectory.

References:

Alkan, O., Kucukoglu, H., & Tutar, G. (2021) Modeling of the Factors Affecting E-Commerce Use in Turkey by Categorical Data Analysis. *International Journal of Advanced Computer Science and Application, 12*, 1-11.

Cox, D. R. (1970). *The Analysis of Binary Data.* Methuen, London.

Creating Correlation Matrices & Heatmaps in Python - αlphαrithms. https://www.alpharithms.com/correlation-matrix-heatmaps-python-152514/

Fang, H. (2022). Analysis of Multiple Linear Regression Algorithm for High Quality Development Factors of Cross-Border E-Commerce. *Journal of Sensors, 2022*. https://doi.org/10.1155/2022/4020607

Fang, J., & Xiaheng, Z. (2015). Cross Border E-commerce Logistics Model Innovation and Development Trend. *China's Circulation Economy*, *6*, 20-26., https://link.springer.com/chapter/10.1007/978-3-662-47721-2_12

Hoffmann, J. P. (2021). *Linear Regression Models: Applications in R*. Chapman and Hall/CRC.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*(5), 539–556.

How do you calculate VIF in SAS? – Wanderluce.com. https://wanderluce.com/news/how-do-you-calculate-vif-in-sas/

How to Fix in R: there are aliased coefficients in the model - Statology. https://www.statology.org/r-aliased-coefficients-in-the-model/

https://www.wish.com/

Imrey, P. B., & Koch, G. (2005). Categorical data analysis. *Encyclopedia of Biostatistics*, *4*. John Wiley & Sons, Hoboken. https://doi.org/10.1002/0470011815.b2a10011

Kadi, F., & Peker, C. (2015). Analyzing the Factors Affecting E-Commerce in Turkey. *International Journal of Management, Accounting and Economics, 2*, 1319-1339.

Kuperenko, V. (2021) Predictive Analytics in Retail & E-commerce, *Big on Data Science & AI*, https://indatalabs.com/blog/predictive-analytics-in-retail-and-e-commerce

Massidda, L., & Marrocu, M. (2018). Quantile Regression Post-Processing of Weather Forecast for Short-Term Solar Power Probabilistic Forecasting. *Energies*, *11*(7), 1763. MDPI AG. Retrieved from http://dx.doi.org/10.3390/en11071763

Namavari, H. (2019). Essays on Objective Procedures for Bayesian Hypothesis Testing [University of Cincinnati / Ohio].

Pan, X. (2019). The network effect and helpfulness of electronic word-of-mouth : understanding the online consumer reviews in social networking sites. https://doi.org/10.48683/1926.00086002

Pardoe, I. (2021). *Applied Regression Modeling*. Wiley.

Preissor, J. S., & Koch, G. G. (1997) Categorical Data Analysis in Public Health Annual Review. *Annual Review of Public Health, 18*, 15-82. https://doi.org/10.1146/annurev.publhealth.18.1.51

Rifada, M., Ratnasari, V., & Purhadi, P. (2023). Parameter Estimation and Hypothesis Testing of The Bivariate Polynomial Ordinal Logistic Regression Model. *Mathematics (2227-7390)*, *11*(3), 579. https://doi.org/10.3390/math11030579

Subregional productivity in the UK - Office for National Statistics. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/labourproductivity/articles/regionalandsubregionalproductivityintheuk/july2022

Turkson, A. J., Addor, J. A., & Kharib, D. Y. (2021). Validating Intrinsic Factors Informing E-Commerce: Categorical Data Analysis Demo. *Open Journal of Statistics*, *11*(5), 737-758. doi: 10.4236/ojs.2021.115044.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, *47*(1), 65-72. https://doi.org/10.1080/00031305.1993.10475938

Vignardi, M. (2018). *Electronic word of mouth communications and consumer experience: an evaluation of pre-and post-purchase responses and reactions to verbal and visual online reviews* (Doctoral dissertation, Kingston University).

Wang, L. (2021). The collaborative development path of cross-border e-commerce and logistics in the data environment. *Journal of Frontiers in Engineering Technology*, *1*(1), 23–26, 2021. https://www.clausiuspress.com/assets/default/article/2021/07/04/article_1625449678.pdf

Wang, S. (2023). A multi-dimensional analysis of CBEC English genre variation in South Asia: Based on Daraz. *PLoS ONE*, *17*(4), 1–19. https://doi.org/10.1371/journal.pone.027971

Wen, L. (2022). Development analysis of cross-border E-commerce logistics based on big data technology under safety law protection. *International Journal of Information Systems in the Service Sector (IJISSS)*, *14*(2), 1-14.

Wolfinbarger, M., & Gilly, M. C. (2001). Shopping Online for Freedom, Control, and Fun. *California Management Review*, 43, 34-55.

Yao, Y. (Angus), & Ma, Z. (2023). Toward a holistic perspective of congruence research with the polynomial regression model. *Journal of Applied Psychology*, *108*(3), 446–465.

Zogovic, K. (2023). Revealing Hidden Trends: Investigating Product Sales Patterns With Categorical And Continuous Predictors In A Distinctive Dataset, *unbublished*

Zogovic, K., et al. (2022). Exploratory research of Covid-19 Vaccination Effects on population in Florida, *MAA-Florida Section and FTYCMA, unbublished*

Kristina Zogović
American College of Education
Miami
USA
zogovic.kristina@gmail.com
ORCID 0009-0005-6119-1845