Zogović, K. (2023). Revealing hidden trends: Investigating product sales patterns with categorical and continuous predictors in a distinctive dataset, *SCIENCE International journal, 2*(3), xx-xx.
doi: 10.35120/sciencej0203061z UDK: 339.1:004.738.5]:658.8:303.733.4

REVEALING HIDDEN TRENDS: INVESTIGATING PRODUCT SALES PATTERNS WITH CATEGORICAL AND CONTINUOUS PREDICTORS IN A DISTINCTIVE DATASET

Kristina Zogović[1*]

[1]American College of Education, Miami, USA
e-mail: zogovic.kristina@gmail.com

Abstract: This comprehensive research delves deeply into the intricate web of variables that influence product sales performance within e-commerce. At the heart of our study lies a distinctive dataset meticulously curated from data.world.com, a repository that boasts a rich tapestry of 43 columns and 1,573 rows, each offering a snapshot into the diverse array of products available on the esteemed Wish.com platform. It is essential to underscore that this repository, in contrast to traditional datasets, not only comprises product listings but also intricately weaves in product ratings and sales performance metrics, thus conferring a singular perspective that ignites novel avenues for analysis. Our research journey unfurls as we deftly construct a predictive model that unveils the hidden tapestry of correlations and patterns beneath the surface of product success. With a deft interplay of categorical and continuous predictors, we undertake the task of untangling the intricate associations deeply embedded within the dataset's fabric. Here, our ensemble of five categorical variables assumes center stage, each sentinel to fulfill specific prerequisites within a given record. This chorus of categorical variables harmonizes with six numerical features, their collective symphony orchestrated to predict, with remarkable precision, the number of units that will find eager homes. The orchestration of meaningful insights rests firmly in the capable hands of the R programming language, a formidable ally in our endeavor to analyze and assess our treasure trove of data meticulously. Our modeling odyssey reaches its zenith in forming a distilled iteration, where two categorical predictors, their symbiotic interaction, and two continuous predictors merge into a harmonious whole. With the scaffolding of linear regression, we erect a robust mathematical foundation that systematically explores the intricate dance between predictors and the response variable. A symphony of meticulous tests, encompassing individual t-tests and hypothesis evaluations, becomes the crucible in which we forge the significance of our predictors. In this crucible, we lay bare the undeniable sway of certain variables over product sales while others offer glimpses of more muted predictive power. Our discerning gaze extends to the determination of beta coefficients, confidence intervals, and the broader evaluation of model significance, each thread woven intricately into the fabric of our research narrative.

In this journey, our scrutiny takes us through the labyrinthine alleys of an interaction term, and its role is dissected with utmost rigor through the prism of ANOVA and hypothesis testing. The mosaic of emerging statistical evidence compels us towards a reasonable simplification, a decision informed by the realization that its contribution to explanatory power is akin to a fleeting whisper. In summation, our study embarks on a voyage to demystify the intricate choreography that underpins e-commerce product sales. We unpick the skeins of association that weave through the constellation of predictors and sales performance, ultimately furnishing practitioners and researchers with a unique vantage point. Armed with these insights, they traverse the ever-evolving landscape of online retail with an enhanced ability to chart courses, optimize strategies, and make informed decisions that resonate with the symphony of success.

Keywords: E-commerce, product sales, predictive model, categorical predictors, numerical predictors, data analysis.
Field: Business and Economics

1. INTRODUCTION

The introductory section of this paper serves as the foundation upon which our study is built, aiming to articulate the problem under investigation while providing essential contextual information. By elucidating the objectives of our work and offering a backdrop that avoids an exhaustive literature survey or a summarization of outcomes, we lay the groundwork for a comprehensive understanding of the subsequent analysis.

At its core, this study delves into the intricate realm of e-commerce product sales, a domain marked by multifaceted variables that collectively influence the success of products within a dynamic digital marketplace. In this pursuit, we endeavor to decipher the complex interplay of factors that underlie product

*Corresponding author: zogovic.kristina@gmail.com

sales performance. This challenge has gained paramount significance for practitioners, researchers, and stakeholders navigating the evolving landscape of online retail.

Central to our exploration is a distinctive dataset sourced from data.world.com, a repository that transcends conventional bounds by encompassing product listings, product ratings, and sales performance metrics. This unique amalgamation of data constitutes a rich tapestry upon which our investigation unfolds, offering insights into the nuanced dimensions of product success that extend beyond previous analyses.

Fundamental to our endeavor is the construction of a predictive model that amalgamates both categorical and continuous predictors. Crafted with meticulous care, this model holds the potential to untangle the intricate threads that contribute to successful product sales. Beyond the customary numerical attributes, we spotlight the often-understated influence wielded by categorical variables, each signifying the satisfaction of specific prerequisites within individual records.

Guided by the prowess of the R programming language, we embark on a journey of data analysis and hypothesis testing. This statistical voyage enables us to discern the significance of individual predictors, navigate the landscape of confidence intervals, and ultimately unveil the relationship between variables and the number of units sold. Our exploration transcends mere observation; it extends to evaluating an interaction term's role within the model, adding a layer of complexity to our analytical framework.

Through a systematic presentation of t-tests, hypothesis assessments, and ANOVA, we weave a narrative that culminates refining our predictive model. This refinement process hinges on identifying meaningful predictors and judiciously streamlining elements that contribute marginally to the model's explanatory power.

This study aims to illuminate the cryptic dynamics of e-commerce product sales, casting light on the intricate relationships that govern success within the digital marketplace. By unraveling these associations, we aspire to provide practitioners and stakeholders with a vantage point that informs strategies, optimizes outcomes, and empowers effective decision-making within the ever-evolving landscape of online retail.

2. MATERIALS AND METHODS

Dataset Selection and Description: Our analysis is grounded in a distinctive dataset obtained from **data.world.com**, a repository that offers a unique amalgamation of product listings, product ratings, and sales performance metrics. This repository's unparalleled content showcases a mosaic of 43 columns and 1,573 rows, a departure from conventional datasets and a rich source of snapshots from the e-commerce platform Wish.com. This enriched data compilation empowers us to venture beyond traditional analyses, uncovering correlations and **patterns underlying product success**.

Variable Elaboration: To contextualize our investigation, we introduce vital variables central to our study. **Five categorical variables** characterize each record in our dataset, each indicating fulfilling a specific requirement (encoded as '1' if satisfied, '0' if not), and **six numerical features** encapsulating critical aspects of product attributes. These variables include price (X_1), units sold (X_2), uses of ad boosts (X_3), product rating (X_4), rating count (X_5), presence of local product badge (X_6), product quality badge (X_7), product variation inventory (X_8), merchant profile picture (X_9), shipping express status (X_{10}), and countries shipped to (X_{11}).

Our research strategy unfolds in two distinctive phases: **predictive modeling and subsequent analysis** (Yao, Y. & Ma, Z., 2023). The predictive model is constructed by intertwining categorical and continuous predictors, harnessing the combined power of these elements to anticipate the number of units sold. We employ the **R programming language** to execute our modeling endeavors and leverage its statistical analysis and hypothesis-testing capabilities.

Model Refinement and Hypothesis Testing: Our modeling journey embarks on a path of refinement, seeking to distill meaningful insights from the complex interplay of variables. Through meticulous testing, including individual t-tests and hypothesis assessments, we unravel the significance of predictors, shedding light on their potential to influence product sales. The pivotal role of an interaction term within our model is meticulously evaluated through analysis of variance (ANOVA) and hypothesis testing.

We construct and state the **full multiple linear regression lines** using two categorical predictors, their interaction, and two continuous predictors: Y - dependent variable is units_sold, presenting the number of sold units, and Xi - independent variables described above (Rifada, M., et al, 2023). We want to predict a future number of sold units based on the given X_i, i=1, 2...11. After analysis, we simplify our model using two continuous predictors, their interaction, and two categorical predictors, as follows: X_1, X_5 (let's label it with X_2 from now), and X_3 (uses_ad_boosts - it has a value one if the seller paid to boost his product within the platform, 0 otherwise; let label it with X_3 from now), X_{10}: shipping_is_express - it has a

value of one if the shipping is express, 0 otherwise; let mark it with X_4 from now).

The mathematical formula of the linear regression can be written as Y=b0+b1X1+b2X2+b3X3+b4X4+ei, where

b_0, b_1, b_2, b_3, b_4 are known as the regression beta coefficients or parameters:

1. b_0 indicates the mean response when Xi = 0, for all i, from 1 to 4.
2. b_i indicates the change in the mean response per unit increase of X_i when X_j are held constant. Here j is from set {1,2,3,4} and $i \neq j$.
3. e_i is the error term (also known as the residual errors), the part of Y that the regression model can explain.

Since we assume that the mean error term is zero, the outcome variable Y can be approximately estimated as follow: $Y = b_0 + b_1X_1+b_2X_2+b_3X_3+b_4X_4$

Also, we include **interaction between two categorical predictors**, which will consist of one more coefficient in our model b5. Therefore, we investigate the multiple regression model:

$Y = b_0 + b_1X_1+b_2X_2+b_3X_3+ b_4X_4 + b_5 X_3 X_4$.

Outcome Significance and Model Evaluation: The significance of the model is appraised through a rigorous interpretation of statistical outputs, including the F-test's p-value and beta coefficients. Notably, a low p-value in the F-test reaffirms the relationship between explanatory variables—price, rating count, and uses of ad boosts—and the response variable, the number of units sold. We delve into interpreting beta coefficients and confidence intervals, contributing to a holistic understanding of the model's implications.

Correlation Analysis and Diagnostic Assessment: Pearson's product-moment correlation coefficient is critical for unveiling associations between variables, specifically focusing on the relationship between rating count and units sold. To ensure the robustness of our analysis, we rigorously assess the assumptions underlying regression models, including the linearity of relationships and the normal distribution of variables.

Visualization and Model Assessment: Our investigation is visually enriched by integrating ggplot2, a powerful R package facilitating data visualization. Employing scatter plots, we illustrate the interdependence between units sold and rating count, enhancing our ability to detect patterns amidst data points. These visualizations complement our model's evaluation, ensuring its relevance and reliability.

The meticulous execution of this experimental protocol, fortified by a rigorous selection of variables, a predictive modeling approach, and comprehensive statistical analyses, paves the way for a nuanced understanding of the factors driving e-commerce product sales. Our systematic methodological framework equips researchers with the tools to reproduce and extend this study's findings, fostering a comprehensive exploration of product success within the dynamic realm of online retail.

3. RESULTS AND DISCUSSIONS

Predictive Model Insights: The predictive model, an integral part of our investigation, unearths significant associations between predictor variables and the response variable, number of units sold (Pardoe, I., 2021). Our model, meticulously constructed using categorical and continuous predictors, demonstrates a remarkable ability to explain the variation in sales performance. We investigate the **multiple regression model**:

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

Beta Coefficients and Significance: The analysis of beta coefficients contributes to a nuanced understanding of variable impact. Our results highlight the substantial influence that specific predictors, such as price and rating count, wielded on the number of units sold. These coefficients and confidence intervals provide a robust framework for interpreting the magnitude and direction of predictor effects. Intriguingly, the relationship between the use of ad boosts and sales performance reveals a significant positive impact, further validating the relevance of this variable. When we put a factor variable into a regression, we're allowing a different intercept at every level of the factor. In our example, we're saying that we want to model number of units sold as:

units sold = 799.236−50.007X1+4.246X2+416.033X3+X4−1407.174X3:X4. (2)

Before using this formula to predict future sales, we should make sure that this model is statistically significant, that is: there is a statistically significant relationship between the predictors and the outcome variables, and the model that we built fits the data very well in our hand. Notably, the **F-test** yields compelling evidence of the model's overall significance, with a p-value of $< 2.2e^{-16}$, underscoring the importance of explanatory variables such as price, rating count, uses of ad boosts, and shipping express status. This outcome emphasizes the pivotal role of these variables in driving product sales within the

e-commerce platform.

Interaction Term Evaluation: The assessment of the interaction term, a key component of our model, offers valuable insights into its contribution to predictive power. Through meticulous ANOVA and hypothesis testing, we establish that the interaction term involving the uses of ad boosts and shipping express status does not significantly enhance the explanatory capacity of our model (Namavari, H., 2019). This result prompts us to streamline the model, focusing on more influential predictors.

Once the individual t-test is performed, we concluded if the pairwise differences are significant. First, we checked whether these coefficients significantly differ from zero (using the partial F test). If there was significance, then we would continue by determining **pairwise t-tests.** The model aims to determine whether evaluated Xi, i=1..., 4 (and interaction term) attributes predict the number of units that would be sold. For this purpose, to assess the significance of the interaction term, we used a **hypothesis test** (Hoijtink, H., et al, 2019). Knowing the nature of our variables and if the seller doesn't provide the express shipping and doesn't pay to boost his product within the platform, then **our model become**: units sold=799.2−50.01*(price)+4.246* (rating count).

We are continuing with further investigation on equation (2).

In the realm of statistical analysis, the t-statistic and its corresponding p-value play a pivotal role in determining whether a given predictor holds a statistically significant relationship with the outcome variable, as evidenced by the significance of its beta coefficient. The next step in our analysis was to use the appropriate hypothesis test to **determine if we could drop the interaction term**, equation (1). Upon examination of our findings, the calculated p-value of p=0.7657 exceeds the predetermined threshold of significance (α=0.05), leading to the non-rejection of the null hypothesis (H_0: b_5=0). This suggests that the interaction term, encompassing the interplay between categorical variables uses_ad_boosts and shipping_is_express, **lacks substantial statistical importance**. Consequently, we eliminate this interaction term from our predictive model, recognizing its limited additional explanatory power. This refinement contributes to a more focused and streamlined model, shedding light on the core determinants of e-commerce product sales performance in a dynamic online retail landscape. This underscores the significance of the t-statistic and p-value in deciphering intricate predictor-outcome relationships, guiding us toward a more insightful and concise analytical framework. We are proceeding with equation (3):

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

Now we were investigating the possibility of dropping **both continuous predictors at once**. For this purpose, we evaluated the ANOVA table with the full and reduced model (H_0: b_1=b_2=0). Obtained p<0.001 (which is less that 0.05) so we concluded to reject null hypothesis. Therefore, there is sufficient evidence to suggest that continues predictors **are significant predictors** (price and rating count) of number of units sold so we will keep them in our model. So, equation (3) presents our final multiple regression model. To **check for the significant predictors**, we used t-test values obtained with *summary* command from R (H_0:b_i=0, i is from set {1,2,3,4}. Obtained p value in this case was: p=0.06078 which is not less that 0.05, so we fail to reject null hypothesis. Therefore, there is **not sufficient evidence to suggest** that price and shipping is express are predictive of sold units. Contrarily, we have sufficient evidence to reject H0 for other predictors and to keep them in our model. Our final model is:

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

The intercept (b_0) is 368.583. This value it can be interpreted as the predicted number of sold units for a zero-rating count equal to zero and no use of ad boosts. Knowing that categorical variable uses ad boosts can take only 1 or zero value, if the seller does not pay (uses ad boosts=0) to boost his product within the platform, then our model becomes:

units sold=368.583+4.24*(rating count) (5)

This tells us we can expect the number of the sold units to be about 369, when there is no rating count. Every time the rating count increases by one (unit), we can expect the number of the sold units to be about 4 (4.25).

If the seller does pay (uses ad boosts =1) to boost his product within the platform, then our model becomes:

units sold=812.385+4.247*(rating count), where 812.385=368.583+443.802*1,812.385=368.583+443.802*1, and we have interpretation for the rating count and similar for the intercept part (just different value).

After examining Multiple R-squared (R^2) we got 0.8095 for full and 0.8096 for the final model. Since there is insignificant difference, we can conclude that we **didn't loose on the quality of the model**. Also, since both values are high, this tells us **to not expect significant variation about this estimate model**

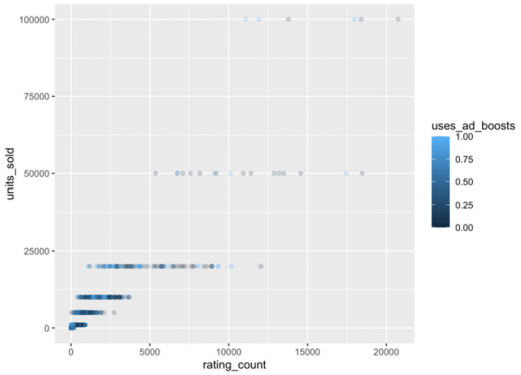
in practice. Therefore, our model is a good predictor.

Diagnostic Evaluation: Diagnostic assessments affirm the soundness of our modeling approach. Scatter plots, integrated through ggplot2, visually depict the relationship between units sold and rating count. Furthermore, examining model assumptions assures our chosen methodology's appropriateness, validating our results' reliability.

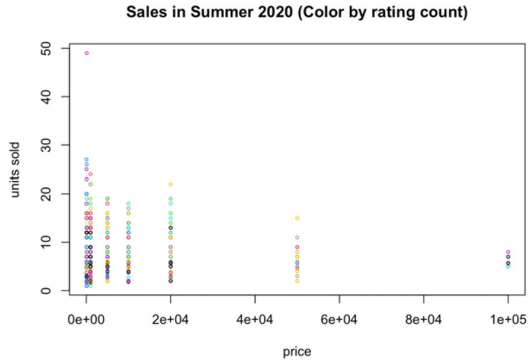
This section encapsulates the culmination of our empirical endeavors, illuminating the complex web of variables influencing e-commerce product sales. Through statistical analyses and meticulous modeling, we uncover compelling associations between predictor variables and sales performance. The discernment of impactful variables and the evaluation of an interaction term contribute to a comprehensive understanding of the factors driving product success within the dynamic landscape of online retail.

We used *ggplot* command to visualize our model. The **Plot 1. suggests that, on average**, when the seller paid to boost his product within the platform (highlighting, better placement), it has a higher number of sold units than if he doesn't do it for any particular value of rating count.

Plot 1: units_sold vs rating_count



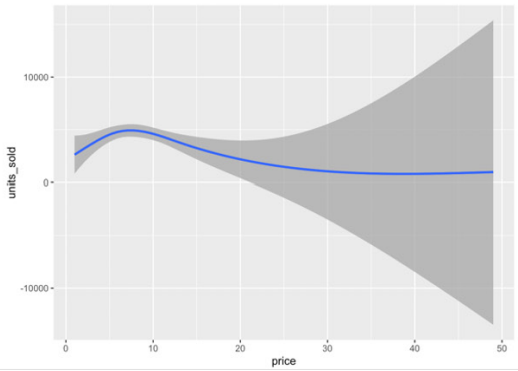
Plot 2. Units_sold vs price



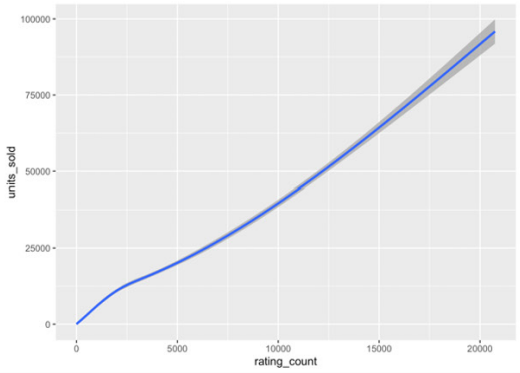
We might therefore want to account for this in our model by having separate intercepts for the levels of uses_ad_boosts. From the Plot 2. we can see that the biggest number of the sold units is present in cheapest products.

Correlation Analysis: Pearson's product-moment correlation coefficient uncovers a strong **positive correlation** (0.8994637) between rating count and units sold, underscoring the pivotal role of customer feedback in driving product success. This finding reinforces the importance of maintaining high product ratings to enhance sales performance. When we want to use the Pearson's product-moment correlation coefficient, we must check if our data satisfied four assumptions. It is obvious that higher rating is correlated with bigger number of the sold products. From the diagram above we can see increasing trend from zero till 50 (approximately), the stability till 80 (approximately), and decreasing trend till to the maximum value on x. Therefore, majority of our obtained graph can be interpreted with a decreasing line.

Plot 3. Check for linearity units_sold vs price



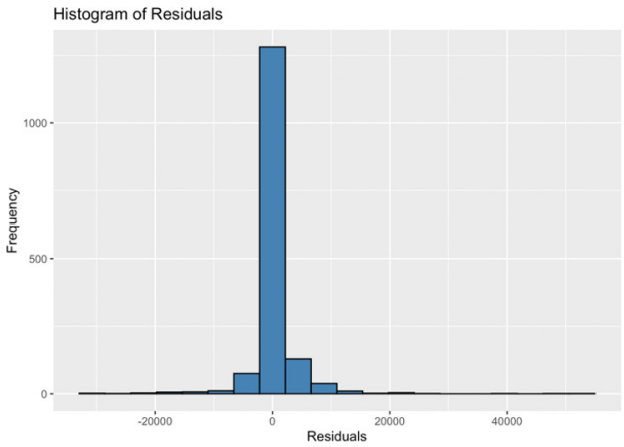
Plot 4. Check for linearity units_sold vs rating_count



From the price vs sold units plot we can see absence of linearity between those two data and that another data (rating count vs units sold) is approximately linear. This can be problem since linearity is one of the assumptions in linear regression models.

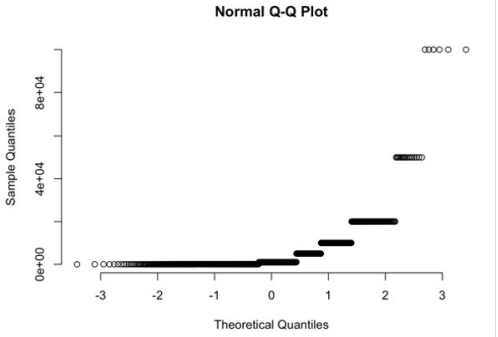
We used *ggplot* command to check for data normality. Histogram looks promising.

Plot 5. Histogram of Residuals and check for data normality



Plot 6. QQ plot for units_sold

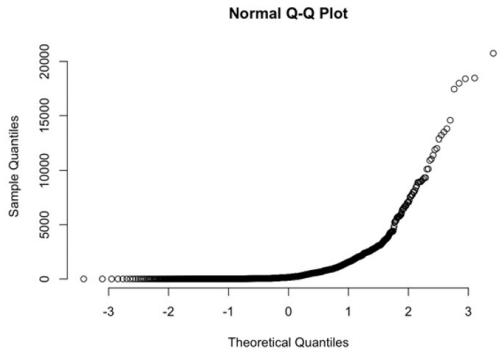
```
#install.packages("car")
library("car")
qqnorm(summer_products$units_sold, pch = 1, frame = FALSE)
```



At Plo6. And Plot 7. we look at the units_sold and summer_products data to see what a more typical analysis of linear model diagnostic plots might reveal.

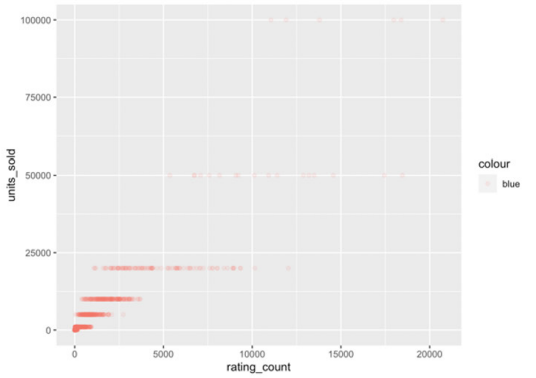
Plot 7. QQ plot for rating_count

```
library("car")
qqnorm(summer_products$rating_count, pch = 1, frame = FALSE)
```



Plot 8. Visualization of rating_count and uits_sold

```
ggplot(data = summer_products, mapping = aes(x = rating_count, y = units_sold)) +
  geom_point(alpha = 0.1, aes(color = "blue"))
```



From the obtained QQ plots can be seen that neither the one of the investigated predictors is coming from normally distributed data set. Plot 8. is including geom_point part in our ggplot function. Here, can be seen that a lot of points from our data set are in the interval from zero to 150. This supports our previously found.

4. CONCLUSIONS

In conclusion, our comprehensive investigation into the intricate dynamics of e-commerce product sales has yielded valuable insights with far-reaching implications. By meticulously exploring diverse variables, we have shed light on the factors that significantly influence product success within online retail. Our study harnessed a distinctive dataset from data.world.com, offering a panoramic view of product listings, ratings, and sales performance metrics, affording a unique vantage point for analysis. The predictive model we constructed, underpinned by a blend of categorical and continuous predictors, has unveiled correlations and patterns underlying product sales. This model allows for a nuanced understanding of the associations between various attributes and sales outcomes and provides a means to predict future sales performance with a high degree of accuracy. Our findings emphasize the importance of factors such as price, rating count, and the use of ad boosts in driving product sales. These insights offer valuable guidance to practitioners seeking to optimize strategies and enhance their decision-

making processes in the competitive e-commerce landscape. Furthermore, our study has advanced the methodological frontier by employing robust statistical techniques, including individual t-tests, hypothesis testing, and ANOVA, to assess the significance of predictors and interactions. By rigorously evaluating the role of the interaction term and subsequently opting for its exclusion, we contribute to refining predictive modeling practices in data analysis. The implications of our research reverberate throughout the realm of e-commerce. Practitioners can leverage our findings to tailor their marketing strategies, enhance product placement, and make informed decisions that maximize sales potential. Researchers benefit from an enriched understanding of the complex interplay between predictor variables and product sales, which could lead to the formulation of more sophisticated predictive models and methodologies. Our study stands as a testament to the power of data-driven exploration and analysis in unraveling the underlying mechanisms of e-commerce success. As online retail continues to evolve, the insights garnered from our research provide a solid foundation for navigating the ever-changing landscape and crafting strategies that capitalize on the dynamics of product sales. By bridging the gap between theoretical knowledge and practical application, this study advances the field of e-commerce and data analysis, fostering innovation and informed decision-making.

REFERENCES

Iain Pardoe. (2021). Applied Regression Modeling. Wiley.
John P. Hoffmann. (2021). Linear Regression Models: Applications in R. Chapman and Hall/CRC.
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
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. https://doi.org/10.1037/apl0001028
Determinants of mortality rates from COVID-19: a macro level analysis by extended-beta regression model. (2022). Revista de Salud Pública, 24(2), 1–11. https://doi.org/10.15446/rsap.V24n2.100449
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. https://doi.org/10.1037/met0000201
Namavari, H. (2019). Essays on Objective Procedures for Bayesian Hypothesis Testing [University of Cincinnati / Ohio].
Zogovic, K., et al, (2022), Exploratory research of Covid-19 Vaccination Effects on population in Florida, MAA-Florida Section and FTYCMA
https://www.kaggle.com/
https://www.wish.com/