RevistaEconomică
Journal of economic-financial theory and practice
Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

DECODING E-COMMERCE FLUCTUATIONS: A MACHINE LEARNING ANALYSIS OF INFLUENTIAL VARIABLES DURING US COVID-19 (2010-2024)

Kevin UNGAR[1]

[1]*Lucian Blaga University of Sibiu, Romania, 0009-0007-3372-9374*

Abstract: *This research examines factors influencing the US e-commerce market during the Covid-19 crisis, investigating consumer behavior across three periods: pre-pandemic (2010–December 2019), pandemic (December 2019–2021), and post-pandemic crisis (2022–2024). Using multiple linear regression analysis in Python and Machine Learning techniques, the study evaluates the impact of key economic indicators (Gross Domestic Product, Unemployment Rate, Consumer Price Index, Internet Penetration Rate, and Consumer Sentiment Index) on e-commerce sales. These variables were used to develop a mathematical model explaining the relationship between economic and sentiment indicators and e-commerce growth. During the pandemic, e-commerce activity surged as lockdowns forced consumers to rely more on online shopping. Post-pandemic, as restrictions eased and confidence recovered, the market exhibited continued growth, surpassing pre-pandemic levels. Despite the initial surge driven by restrictions, e-commerce remained strong even after their removal. The analysis highlights the importance of economic factors in shaping e-commerce trends, with GDP and CPI emerging as particularly influential. Additionally, the study underscores the critical role of internet penetration in sustaining e-commerce, especially during physical distancing measures. These findings provide insights into how economic and technological factors drive long-term changes in consumer behavior within the e-commerce sector.*

Key words: *Consumer Behavior, E-commerce, Crisis Marketing, Multiple Linear Regression, Covid-19 crisis, Digital Transformation*

JEL classification: *D12, D81*

1. Introduction

Understanding consumer behavior during crises like the COVID-19 pandemic is essential for businesses, particularly in the e-commerce sector, due to its significant impact on market dynamics during such periods. The COVID-19 pandemic, which began in December 2019, caused unprecedented disruptions worldwide, significantly altering consumer habits and preferences. Prior research, such as (Loxton et al., 2020), has documented phenomena like panic buying, shifts towards essential goods and the media's role in shaping consumer behavior. However, these studies primarily focused on general consumer reactions without deeply exploring the specific factors driving e-commerce sales during the pandemic.

This study aims to fill this gap by examining the factors influencing e-commerce sales throughout the pandemic. Understanding these drivers is crucial for various stakeholders. Consumers benefit from a market that adapts to their evolving needs, businesses gain insights for developing effective marketing strategies during disruptions, and policymakers can devise regulations to ensure a healthy, competitive e-commerce environment during and after crises.

The COVID-19 pandemic led to a fundamental shift in consumer behavior across many industries, with the e-commerce market experiencing a dramatic surge due to lockdowns, social distancing, and a shift towards online shopping for essentials (UNCTAD, 2020). However, the intricate

[1] *email: kevinungar7@gmail.com*

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

dynamics of how consumer behavior interacted with these factors to influence e-commerce sales remain underexplored.

This research addresses this gap by investigating the factors affecting e-commerce sales during the pandemic using a multiple linear regression model. The study utilizes data from 2010 to 2024, covering pre-pandemic (2010-December 2019), pandemic (December 2019-2021), and post-pandemic (2022-2024) periods. Although 2021 is not the definitive end of the pandemic, it marks a significant turning point with the reopening of many businesses and the widespread rollout of vaccines, indicating a shift in societal management of the virus (The National Bureau of Economic Research, 2021).

Conducted using Python and employing machine learning techniques with libraries such as Scikit-learn (Pedregosa et al., 2011), the study contributes to a deeper understanding of consumer behavior during crises. It offers insights for developing marketing strategies and policies to bolster e-commerce resilience in future disruptions. By utilizing multiple linear regression and machine learning techniques, the study builds a robust model to explain the sustained growth of e-commerce post-crisis, emphasizing the roles of GDP and CPI. These findings provide valuable guidance for practitioners on leveraging economic indicators to predict and enhance e-commerce trends.

The study's model analyzes the influence of economic indicators such as Gross Domestic Product (GDP), unemployment rate, Consumer Price Index (CPI), internet penetration rate, and consumer sentiment index on e-commerce sales. By examining these variables across different timeframes, the research aims to identify the most significant factors influencing e-commerce activity during the pandemic and compare these influences with pre- and post-pandemic trends.

2. Literature review

In recent years, global disruptions (most notably the COVID-19 pandemic) have accelerated digital transformation across industries, making e-commerce a cornerstone of modern economic activity. This shift has spurred a wealth of research examining the interplay between technological innovation, macroeconomic indicators, and consumer behavior. A comprehensive understanding of these dynamics is essential for developing resilient strategies that can not only withstand crises but also drive sustainable growth in a rapidly evolving digital landscape.

Previous studies, such as (Anvari and Norouzi, 2016), have explored the impact of e-commerce and R&D on economic development but did not specifically address how economic indicators affect e-commerce sales during a crisis. Similarly, while (Loxton et al., 2020) provided insights into consumer behavior changes during crises, they did not investigate the long-term economic variables sustaining e-commerce growth beyond the immediate crisis period. This study bridges these gaps by offering a comprehensive evaluation of how key economic and sentiment indicators influence e-commerce sales across different periods.

Recent advancements in machine learning further complement this research. (Bami, Behnampour, and Doosti, 2025) propose a flexible train-test split algorithm that improves model evaluation by accommodating various validation techniques (Hold-out, K-fold, and iterative methods). Their work highlights how subtle variations in model validation can influence predictive accuracy (an insight that is directly applicable to forecasting economic trends in e-commerce).

Similarly, (Liang et al., 2025) extend this discussion by introducing a K-fold cross-validation based frequentist model averaging approach, which addresses issues like nonignorable missing responses in datasets. These methodologies underscore the need for rigorous validation when deploying machine learning techniques in economic forecasting.

The transformative potential of digital technologies is also evident in related fields. (Bouchetara, Zerouti, and Zouambi, 2024) examine how artificial intelligence can revolutionize public sector financial risk management by enhancing decision-making through predictive modeling and scenario analysis. Their findings suggest that similar AI-driven strategies could be applied to e-commerce to better navigate market volatility. In addition, (Mohammedi et al., 2025) explore the impact of the digital economy on

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

sustainable development in the context of geopolitical risks, revealing that digitalization can play a significant role in bolstering both environmental and socio-economic performance.

Lastly, (Li et al., 2025) address the challenges of out-of-distribution generalization in machine learning models, emphasizing that traditional evaluation methods may overestimate model performance when faced with novel data. This insight reinforces the importance of constructing rigorous benchmarks, ensuring that predictive models are truly robust (a consideration that is crucial for accurately forecasting e-commerce trends during and after crises).

In synthesizing these diverse perspectives, the present research integrates traditional economic indicators with advanced machine learning methodologies to provide a nuanced analysis of e-commerce dynamics during the COVID-19 crisis and beyond. This integrated approach not only fills existing research gaps but also offers a more comprehensive framework for understanding and predicting long-term shifts in consumer behavior in the digital economy.

3. Data Collection and Processing

The data utilized for constructing the multiple linear regression model was collected from the reputable source, Statista (https://www.statista.com). The dependent variable, E-commerce sales (y), was collected quarterly from 2010 to 2024. This extended timeframe was chosen to ensure a robust dataset for training the mathematical model, thereby enhancing its predictive performance. However, the availability of other crucial variables such as Gross Domestic Product (GDP), unemployment rate, internet penetration rate, and Consumer Price Index (CPI) was limited to annual frequencies, while the consumer sentiment index was available monthly starting from February 2011. To harmonize the dataset and maintain consistency, several processing steps were undertaken.

Firstly, the consumer sentiment index was transformed from a monthly to quarterly frequency by taking the average. To address the missing values between 2010 and 2011, a method of imputation was applied where the first quarterly value of 2011 was replicated for all four quarters to maintain continuity and preserve the overall trend of the data.

Secondly, for variables with annual frequencies (GDP, unemployment rate, internet penetration rate, CPI), a different approach was adopted. Each annual value was repeated or "broadcasted" across the respective quarters of that year. This method ensured the integrity of the dataset while aligning with the quarterly frequency required for the analysis.

As a result of these processing steps, the original dataset comprising 14 annual values for each variable was expanded into a comprehensive data frame containing 56 observations from January 1, 2010, to December 31, 2023. This harmonized dataset enabled a thorough examination of the relationship between the selected indicators and E-commerce sales over the specified timeframe.

3.1 Normalization of Variables

In the dataset used for the multiple linear regression analysis, each variable was measured on a different scale or unit, ranging from percentages to millions. This discrepancy in scales among the variables can potentially influence the coefficients in a linear regression model. For instance, if one independent variable has a larger scale than others, its coefficient might appear smaller simply because a one-unit change in that variable represents a larger change in the dependent variable compared to a one-unit change in a variable with a smaller scale (James et al., 2013).

To mitigate this issue and ensure that all variables contribute equally to the regression model, normalization was employed. Normalization, also known as z-score normalization, involves scaling the values of each variable so that they have a mean of 0 and a standard deviation of 1. This process standardizes the variables and places them on the same scale, facilitating fair comparison and interpretation of coefficients.

The normalization process involves computing the z-score for each variable. To calculate the z-score, the difference between each data point and the mean of the variable is divided by the standard deviation of the variable. This normalization technique ensures that all variables are treated equally in

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

the regression model and aids in convergence during the optimization process (Montgomery, D. C. 2017). By normalizing the variables, the regression model can effectively analyze the relationship between the independent variables (such as GDP, unemployment rate, internet penetration rate, CPI and consumer sentiment index) and the dependent variable (E-commerce sales) without the bias introduced by varying scales.

This normalization step is crucial for ensuring the accuracy and reliability of the regression analysis, allowing for meaningful interpretation of the coefficients and robust inference about the relationship between the variables.

3.2 Cross-Validation and Model Evaluation

In the process of developing the regression model, a crucial consideration was whether to scale both the independent variables (X) and the dependent variable (Y), or to scale only the independent variables. To address this question, a cross-validation approach was employed using two sets of data: one where both the independent variables (X) and the dependent variable (Y) were scaled, and another where only the independent variables (X) were scaled (James et al., 2013; Montgomery, D. C. 2017).

In this study, a 3-fold cross-validation was performed, dividing the data into three subsets for training and testing. The negative mean squared error (MSE) was calculated as the validation score, with lower MSE values indicating better model performance. Additionally, the validation score was used to evaluate the model's generalizability to unseen data.

The results of the cross-validation are as follows:

Table 1: Results of the cross-validation

Model with both X and Y scaled		Model with only X scaled	
Validation scores	Mean squared error (MSE)	Validation scores	Mean squared error (MSE)
-0.01507865	-0.02301353	-9.20739933e+07	-140526382
-0.03990707		-2.43682554e+08	
-0.01405488		-8.58225999e+07	

Source: Created by the author

The findings in Table 1 underscore that scaling both the independent and dependent variables (X and Y) is critical for creating a uniform feature space. When X and Y are on similar scales, the regression coefficients tend to be more stable and interpretable, which reduces the risk of numerical instability during model optimization (Bami, Behnampour, & Doosti, 2025). This balanced feature enables the model to more effectively learn the relationships among variables. In contrast, when only the independent variables are scaled, the large variance in the magnitude of Y can lead to disproportionate error contributions, resulting in an inflated mean squared error (MSE).

The dramatic difference in MSE values between the two approaches (ranging from -0.023 in the fully scaled model to nearly -140 million when only X is scaled) suggests several underlying issues. One likely explanation is numerical instability: unscaled Y values can trigger computational errors, especially when the data exhibit extreme values or high variance, which in turn manifests as excessively large error metrics (Liang et al., 2025). Moreover, the failure to scale Y creates a scale mismatch. Without proper scaling, the absolute differences between the model's predictions and the actual values are computed on a much larger scale, thereby driving up the MSE. While this does not inherently indicate data leakage, it does highlight the sensitivity of the MSE metric to the target variable's scale. Additionally, this discrepancy may signal the presence of heteroscedasticity (non-constant variance in the model's errors), which can further compromise both model performance and the reliability of error estimates.

Although standardizing Y has proven effective in our study, alternative transformations should be considered. For instance, applying a logarithmic transformation to Y might stabilize the variance and mitigate the influence of outliers, while still preserving the interpretability of the relationships. Future

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

research could explore whether such a log transformation might yield similar or even superior improvements in model performance.

It is also important to note that while MSE is a widely used performance metric, its sensitivity to the scale of the target variable suggests that complementary metrics (such as R-squared or adjusted R-squared) could offer additional insights into the robustness of the model. These metrics, which will be discussed further in a subsequent section, assess the proportion of variance explained by the model and can help validate whether the improvements observed through scaling translate into a more reliable predictive framework.

Overall, the results emphasize the critical importance of scaling both the independent and dependent variables when performing regression analysis. The improved performance of the model that scales both X and Y indicates that a balanced feature is essential for capturing the underlying relationships between variables, reducing numerical issues, and enhancing generalizability to unseen data. Future work should further explore alternative transformations of Y and incorporate additional performance metrics to fully substantiate these conclusions and reinforce the robustness of predictive models in economic applications.

4. Mathematical Model

The mathematical model utilized to analyze the relationship between E-commerce sales (y) and the independent variables is represented by Equation 1:

Equation nr.1:

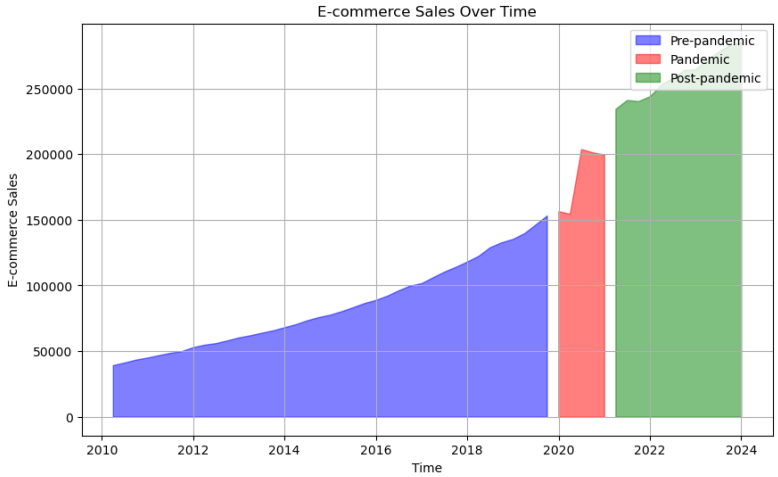
$$y = \beta_0 + \beta_1 * GDP + \beta_2 * \text{Unemployment Rate} + \beta_3 * CPI + \beta_4 * \text{Internet Penetration Rate} + \beta_5 * \text{Consumer Sentiment Index} + \epsilon$$

In this equation:

- "y" represents the E-commerce sales, which is the dependent variable.
- β_0, β_1, β_2, β_3, β_4, and β_5 are the coefficients (parameters) that represent the effect of each independent variable on the dependent variable.
- GDP, Unemployment Rate, CPI, Internet Penetration Rate, and Consumer Sentiment Index are the independent variables represented as x_1, x_2, x_3, x_4, and x_5 respectively.
- ϵ represents the error term, which captures the difference between the observed and predicted values of E-commerce sales that cannot be explained by the independent variables.

This model enables the analysis of how changes in GDP, Unemployment Rate, CPI, Internet Penetration Rate, and Consumer Sentiment Index affect E-commerce sales while controlling for the effects of other variables. For a visual representation of the periods under consideration, a graph illustrating the growth of E-commerce sales over time is presented below:

Figure 1: E-commerce sales over time



Source: Created by the author

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

The figure nr. 1 displays the growth of E-commerce sales over time, covering the period from 2010 to 2024. The x-axis represents time, while the y-axis represents E-commerce sales. The graph is divided into three segments:

- *Pre-pandemic (2010-2020):* The blue segment represents the period before the onset of the COVID-19 pandemic.
- *Pandemic (2020-2021): The red segment covers the duration of the COVID-19 pandemic crisis, from its onset until the end of 2021.*
- *Post-pandemic (2022-2023 December 31)*: The green segment symbolizes the recovery period following the initial crisis, during which the economy and businesses gradually rebound from the impact of COVID-19 restrictions.

These distinct segments allow for a visual understanding of how E-commerce sales evolved throughout different phases, providing context for the analysis conducted using the mathematical model described above.

4.1 Analysis of Coefficients and Mean Squared Error (MSE)

Now that the period has been appropriately segmented, an analysis of the coefficients obtained from the linear regression models for each of the three periods, along with the one covering the entire period range (2010 - 2024), was conducted. This analysis aimed to identify which factors most significantly influenced e-commerce activity during the pandemic and how these influences compared to pre-pandemic and post-pandemic trends. Additionally, analyzing the coefficients from the entire dataset helped improve the model's performance by increasing the number of observations used to estimate the model parameters.

Table 2: Results of Linear Regression for Each Period

Coefficients	GDP	Unemployment Rate	CPI	Internet Penetration Rate	Consumer Sentiment Index	Mean Squared Error:
Whole Dataset	**1.96160945**	**0.23656766**	**-0.8701078**	**0.01724936**	**-0.09095128**	0.01150686847
Pre-pandemic	1.41231862	0.42408446	-0.1011498	0.07732885	0.02109997	0.01230384348
Pandemic	0.21157819	-0.21157819	0.2115781	0.21157819	-0.07909782	0.24861104778
Post-pandemic	0.17539896	0.17539896	0.1753989	0.17539896	0.27361418	0.16713640900

Source: Created by the author

Considering Table 2, when comparing the results obtained from the segmented periods (pre-pandemic, pandemic, and post-pandemic) with those derived from the whole dataset spanning from 2010 to 2024, it becomes evident that various factors must be taken into account to ascertain the approach that best find the genuine influence of the independent variables on e-commerce sales.

Pre-pandemic Phase:

During the pre-pandemic phase, GDP and the Unemployment Rate exhibited relatively high coefficients, signifying a robust positive impact on e-commerce sales. Conversely, other variables displayed smaller coefficients, suggesting weaker influences.

Pandemic Phase:

In contrast, the coefficients for all variables witnessed a significant decrease during the pandemic, indicating a reduction in their influence on e-commerce sales. Notably, the Unemployment Rate turned negative, implying a potentially adverse effect on e-commerce activity during this period.

Post-pandemic Phase:

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

Following the pandemic, the coefficients for all variables remained relatively lower, indicating a moderate influence on e-commerce sales. However, it is noteworthy that the coefficient for the Consumer Sentiment Index saw a substantial increase, suggesting a stronger impact on e-commerce during activity during the recovery phase.

To truly discern the influence of independent variables on e-commerce sales, it is essential to factor in the number of observations available for each period. The significance of this consideration lies in the robustness and reliability of the estimates derived from the regression analysis.

4.1.1 Analyzing Segmented Periods vs. Whole Dataset

Whole Dataset:

With 56 observations per variable, the whole dataset provides a substantial amount of data, ensuring a comprehensive and reliable analysis. The larger sample size enhances the statistical power of the analysis, leading to more accurate estimates of the coefficients and a better understanding of the true influence of the independent variables.

Pre-pandemic Phase:

Despite comprising 40 observations, the pre-pandemic phase offers a substantial amount of data for analysis. However, the smaller sample size compared to the whole dataset may result in slightly less precise estimates of the coefficients. Nonetheless, valuable insights into the influence of independent variables on e-commerce sales during this period can still be gleaned.

Pandemic Phase:

The pandemic phase presents a challenge due to the limited number of observations, with only 8 available for analysis. This smaller sample size may introduce greater variability and uncertainty in the estimates of the coefficients. As a result, the findings derived from this period should be interpreted with caution, recognizing the inherent limitations imposed by the restricted data availability.

Post-pandemic Phase:

Similarly, the post-pandemic phase also comprises only 8 observations, posing similar challenges in terms of data reliability and precision of estimates. Despite the smaller sample size, insights into the influence of independent variables on e-commerce sales during the recovery phase can still be gleaned, albeit with a degree of caution.

In conclusion, while all segmented periods offer valuable insights into the dynamics of e-commerce activity, the whole dataset approach stands out as the most robust and reliable method for discerning the true influence of independent variables on e-commerce sales. The larger sample size ensures more accurate estimates of coefficients and a better understanding of long-term trends and patterns. Nonetheless, the insights gained from each segmented period, when interpreted in conjunction with the number of observations available, contribute to a more comprehensive understanding of the factors driving e-commerce sales dynamics across different phases. Regarding the MSE for each period (lower values are better), it is evident that both the pre-pandemic phase and the whole dataset exhibit lower MSE values, indicating better model performance compared to the pandemic and post-pandemic phases. Therefore, the models trained on the pre-pandemic phase and the whole dataset are considered superior in terms of accuracy. This underscores the significance of the number of observations in accurately capturing the true influence of the independent variables. From these results, it can be discerned that the number of observations plays a pivotal role in determining the reliability and accuracy of the regression model (Belsley et al., 2005). The coefficients being the same for GDP, Unemployment Rate, CPI, and Internet Penetration Rate in both the pandemic and post-pandemic phases could indicate a scenario where the model assigns equal importance or influence to these variables during those respective periods. But, the number of observations is limited, as in the pandemic and post-pandemic phases with only 8 observations each. So, model's ability to accurately estimate the coefficients may be compromised.

Consequently, in the subsequent sections of the article, the *whole dataset* will be utilized for further testing the performance of the mathematical model. This approach ensures a more comprehensive

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

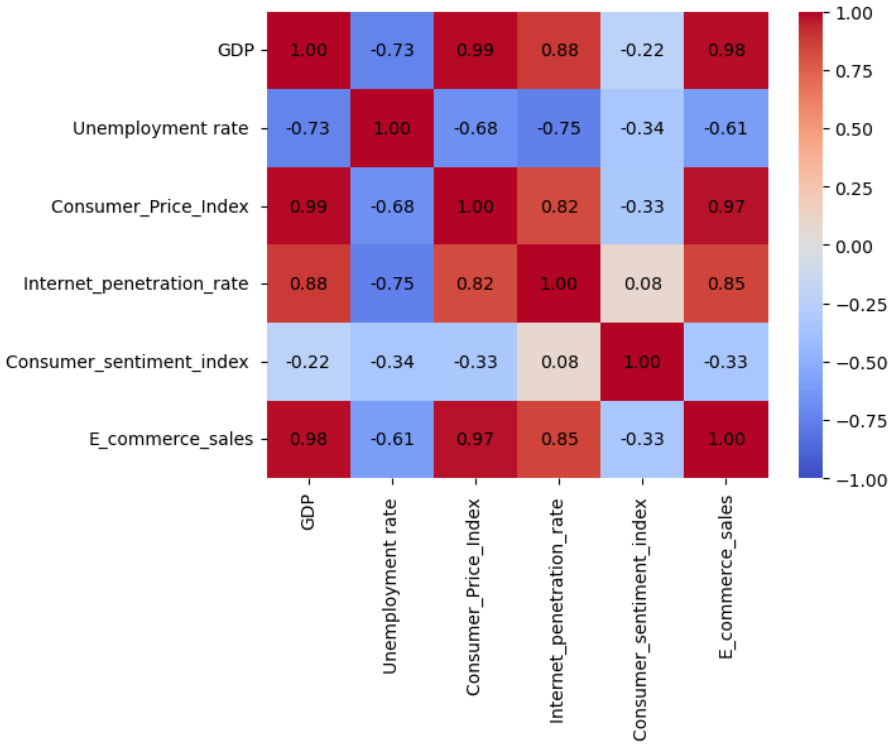
analysis, leveraging the larger sample size to obtain more robust and reliable estimates of the coefficients and model predictions.

5. Analysing Correlation Between Variables through Heatmap

To gain further insights into the relationships between the independent variables and the dependent variable (E-commerce sales), a correlation analysis was conducted. This analysis helps ascertain the degree and direction of association between different variables, providing valuable information for model interpretation and feature selection. But, it's important to mention that correlation doesn't equal causation. Just because two variables are correlated doesn't mean one causes the other to change.

A heatmap visualization was employed to illustrate the correlation matrix, with colors representing the strength and direction of correlation. A color spectrum ranging from cool (blue) to warm (red) hues was utilized, where blue indicates negative correlation, red signifies positive correlation, and darker shades denote stronger correlations.

Figure 2: Heatmap correlation Matrix



Source: Created by the author

The correlation matrix (figure 2) reveals the pairwise correlations between each pair of variables. Here's a brief overview of the correlations observed:

GDP and E-commerce Sales: a strong positive correlation of approximately 0.98 is observed between GDP and E-commerce sales (Hair et al., 2019). This indicates that as GDP increases, E-commerce sales tend to increase as well.

Unemployment Rate and E-commerce Sales: there is a moderately strong negative correlation of around -0.61 between the unemployment rate and E-commerce sales. This suggests that as the unemployment rate rises, E-commerce sales tend to decrease.

Consumer Price Index (CPI) and E-commerce Sales: a strong positive correlation of about 0.97 is found between the CPI and E-commerce sales. This implies that as the CPI increases, E-commerce sales also tend to increase.

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

Internet Penetration Rate and E-commerce Sales: a strong positive correlation of approximately 0.85 is observed between the internet penetration rate and E-commerce sales. This suggests that higher internet penetration rates are associated with higher E-commerce sales.

Consumer Sentiment Index and E-commerce Sales: there is a relatively weak negative correlation of approximately -0.33 between the consumer sentiment index and E-commerce sales. This implies a slight inverse relationship between consumer sentiment and E-commerce sales.

Overall, the correlation analysis provides valuable insights into the relationships between the variables, highlighting potential influential factors on E-commerce sales such as GDP, CPI, and internet penetration rate. The internet penetration rate exhibits a relatively strong positive correlation with E-commerce sales, despite its coefficient value being lower compared to GDP and CPI. This highlights the importance of considering both the magnitude of correlation coefficients and their practical implications. While GDP and CPI may have higher correlation coefficients with E-commerce sales, indicating stronger linear relationships, the internet penetration rate's correlation, although slightly lower in magnitude, still holds practical significance. A correlation coefficient of approximately 0.85 suggests a robust and positive relationship between internet penetration rate and E-commerce sales. This finding underscores the transformative impact of digital technology and online connectivity on consumer behavior and commerce (Anvari, R. D., & Norouzi, D. 2016). Higher internet penetration rates enable greater access to online platforms, facilitating E-commerce transactions and driving sales volumes. As such, even though the coefficient value for internet penetration rate may be lower, its substantial correlation with E-commerce sales underscores its importance as a key determinant of online commerce activity.

6. Preparing the Data for Machine Learning

The data used in the regression model was prepared by partitioning the dataset into training and testing, that serves as a foundational step in the development and assessment of machine learning models. By dividing the dataset, it creates distinct subsets that play unique roles in the model-building process. The data is divided into two distinct subsets: the training set and the testing set. The training set, typically comprising around 75% of the data, is used to train the machine learning model. During this process, the model learns the underlying relationships between the independent variables and the dependent variable. Randomization is employed to shuffle the observations in the dataset before partitioning (James et al., 2013). This ensures that the data is randomly distributed across both training and testing sets, preventing any inherent ordering or bias from influencing the splitting process.

The primary purpose of this process is to evaluate the performance of the machine learning model. By training the model on one subset (training set) and testing it on another unseen subset (testing set), we can assess how well the model generalizes to unseen data and make informed decisions about its effectiveness. Splitting the dataset helps prevent overfitting, a common problem in machine learning where the model learns to memorize the training data rather than generalize to new data (James et al., 2013). By evaluating the model on a separate testing set, we can gauge its ability to generalize and avoid overfitting. This ensures the model doesn't simply become overly specific to the training data and performs poorly on unseen instances.

6.1 Evaluating Model Performance with R-squared and Adjusted R-squared

After building the multiple linear regression model, its performance must by evaluate. One way to assess model performance is through the R-squared (R^2) score and adjusted R-squared. R-squared is a statistical measure that represents the proportion of the variance (the average squared difference from the mean) in the dependent variable (e-commerce sales, in this case) that is explained by the independent variables (such as GDP, unemployment rate, CPI, etc.) included in the model (Hair et al., 2019). It provides insights into how well the independent variables predict the variation in the dependent variable. A higher R-squared value indicates a better fit of the model to the data.

However, R-squared alone may not provide an accurate assessment of model performance, especially when dealing with multiple independent variables. This is where adjusted R-squared comes

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

into play. Adjusted R-squared considers the number of independent variables in the model and penalizes the R-squared value for including unnecessary variables, helping to prevent overfitting (James et al., 2013). By adjusting for model complexity, it provides a more reliable estimate of the model's generalizability. By calculating both R-squared and adjusted R-squared scores, this article provide a comprehensive evaluation of the model's performance, taking into account both its predictive accuracy and its complexity (Cohen et al., 2013). This allows researchers and practitioners to make more informed decisions about the effectiveness of the machine learning model in predicting e-commerce sales.

Table 3: R-squared and Adjusted R-squared

R-squared (R2)	*0.9886608159622904*
Adjusted R-squared	0.9815738259387219

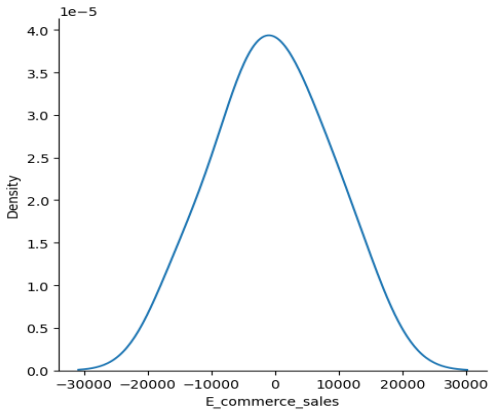
Source: Created by the author

Table 3 presents a R-squared (R2) score of 0.9887, indicating that about 98.87% of the variance in e-commerce sales is explained by the independent variables. The adjusted R-squared, considering model complexity, is slightly lower at 0.9816. These high scores suggest the model effectively captures the relationship between variables and e-commerce sales. Overall, it demonstrates strong predictive performance, showing potential for predicting e-commerce trends based on economic indicators and consumer sentiment.

6.2 Residuals

The residuals represent the discrepancies between the actual e-commerce sales values (actual observations) and the values predicted by the machine learning model (Montgomery & Huang, 2019). Each residual corresponds to a specific observation in the testing dataset and quantifies how much the model's predictions deviate from the actual sales figures. Positive residuals indicate that the model overpredicted sales, while negative residuals indicate underpredictions. Analyzing the residuals helps in evaluating the accuracy of the model's predictions and identifying areas where the model may need improvement. The residuals calculated for this machine learning model are as follows: -13288.630043, -5262.078692, -4618.967570, 3702.077896, 14739.268395, 9447.836888, -15480.977600, 1755.405169, 9773.450146, -482.547797, -4715.097076, -7515.630043, -100.075844, and 4543.820169. These residuals are obtained by subtracting the dependent variable test set values from their corresponding predictions based on the multiple linear regression model output. However, for a better visualization and interpretation, it is necessary to create a kernel density estimation (KDE) plot using the residuals obtained from the machine learning model. The KDE plot visually represents the distribution of the residuals, providing insights into their shape and spread and allowing to assess the accuracy of the model and identify any patterns or deviations in the errors.

Figure 3: Kernel Density Estimation plot (KDE)



Source: Created by the author

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

In Figure 3 the X-axis represents the values of the residuals. It's from negative sales prediction errors (underpredictions) on the left to positive sales prediction errors (overpredictions) on the right. The Y-axis represents the density of residuals. Higher values on the y-axis indicate that more residuals fall within that particular range of the x-axis. The curve seems to be somewhat symmetrical. A good sign that suggests the residuals are not heavily skewed in one direction (positive or negative). This indicates the model might not have a consistent bias towards overpredicting or underpredicting sales and the fact that is centered close to zero, represent a desirable outcome (Hyndman, R. J., 1996). This suggests the residuals are scattered around zero, and the model's predictions are generally close to the actual sales figures.

7. Regression Results (OLS model)

While the initial evaluation of model performance using R-squared and Adjusted R-squared provided an overall measure of fit, a more detailed examination of the relationships between independent variables and e-commerce sales is necessary. To achieve this, an Ordinary Least Squares (OLS) regression model was employed. OLS is a robust statistical method that estimates the linear relationships between variables, ensuring unbiased and efficient parameter estimates under the assumptions of linearity, homoscedasticity, and no multicollinearity. This method allows us to assess the individual impact of each predictor, offering a deeper understanding of the economic and technological factors shaping e-commerce trends. The regression results are summarized in Table 4:

Table 4: Regression Summary

Dep. Variable	Y	**R-squared (uncentred)**	0.987
Mode:	OLS	**Adj. R-squared (uncentred)**	0.985
Method	Least Squares	**F-statistic**	551.6
Date	Sat, 23 Mar 2024	**Prob (F-statistic)**	1.16e-33
Time	18:50:19	**Log-Likelihood**	32.222
No. Observations	42	**AIC**	-52.44
Df Residuals	37	**BIC**	-43.76
Df Model	5		
Covariance type	Nonrobust		

Source: Created by the author

Based on the results in table 4, we can interpret the findings as follows:

R-squared (uncentered): This metric, with a value of 0.987, indicates that approximately 98.7% of the variance in the dependent variable (y) is explained by the independent variables (x1, x2, x3, x4, x5) included in the model. A higher R-squared suggests a better fit of the model to the data.

Adj. R-squared (uncentered): This adjusted version of R-squared, with a value of 0.985, accounts for the number of predictors in the model. It penalizes the inclusion of unnecessary variables and generally provides a more conservative estimate of the goodness of fit compared to the unadjusted R-squared.

F-statistic: The F-statistic, with a value of 551.6, tests the overall significance of the regression model. It assesses whether at least one independent variable is significantly related to the dependent variable. A higher F-statistic and a lower associated p-value (Prob (F-statistic)) suggest a more statistically significant model.

Prob (F-statistic): This p-value, with a value of 1.16e-33 (very close to zero), indicates strong evidence against the null hypothesis, suggesting that the overall regression model is statistically significant. In other words, the independent variables collectively have a significant effect on the dependent variable.

Log-Likelihood: The log-likelihood, with a value of 31.222, measures the goodness of fit of the model. A higher log-likelihood suggests a better fit, as it indicates that the model is more likely to have produced the observed data.

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

AIC and BIC: These information criteria, with values of -52.44 and -43.76 respectively, are used for model selection. Lower values of AIC and BIC indicate a good balance between model fit and complexity. Models with lower AIC and BIC values are generally preferred.

While table 4 provides an overall assessment of the model's fit and significance, a deeper analysis of the individual contributions of each independent variable is necessary. Table 5 below presents the estimated coefficients, allowing us to evaluate the magnitude, direction, and statistical significance of each predictor's impact on e-commerce sales.

Table 5: Coefficients

	Coef	Std err	t	P>ItI	[0.025	0.975]
X1 (GDP)	1.9616	0.256	7.649	0.000	1.442	2.481
X2 (Unemployment rate)	0.2366	0.041	5.773	0.000	0.154	0.320
X3 (CPI)	-0.8701	0.236	-3.681	0.001	-1.349	-0.391
X4 (Internet penetration rate)	0.0172	0.055	0.313	0.756	-0.094	0.129
X5 (Consumer sentiment index)	-0.0910	0.042	-2.183	0.035	-0.175	-0.007

Source: Created by the author

Table 5 offers deeper insights into the contribution of each independent variable, enabling us to assess their individual impact on e-commerce sales:

Coef (Coefficients): These are the estimated coefficients of the independent variables (x1, x2, x3, x4, x5) in the regression equation. They represent the expected change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant. So, for example, for x1 as 1.9616, it means that for every one-unit increase in the independent variable x1, the dependent variable (y) is estimated to increase by approximately 1.9616 units, assuming all other variables in the model remain constant or unchanged.

Std err (Standard Errors): These values represent the standard errors associated with each coefficient estimate. They indicate the precision of the estimated coefficients. Smaller standard errors imply more precise estimates.

t-statistic and P>|t|: These values assess the statistical significance of each coefficient. The fact that there are three 0.000 values in the P>|t| column of the second table indicates that those 3 corresponding independent variables (x1, x2, and x3) have a very strong relationship with the dependent variable (y). The p-value for x1, x2, x3 (0.000) indicates strong evidence against the null hypothesis, suggesting that the coefficient for x is statistically significant. The p-value for x4 (0.756) is relatively high, suggesting that the coefficient for x4 is not statistically significant at conventional levels (since it is greater than 0.05).

[0.025, 0.975] (95% Confidence Interval): These values represent the lower and upper bounds of the 95% confidence interval for each coefficient. They provide a range within which the true population parameter lies. For example: for x1, the 95% confidence interval is [1.442, 2.481], indicating that we are 95% confident that the true coefficient for x1 falls within this range.

Overall, the regression results confirm the strong influence of key economic factors on e-commerce sales. The statistically significant coefficients for variables x1, x2, and x3 suggest that these predictors play a crucial role in shaping online shopping trends, while the high R-squared value indicates that the model effectively captures the relationship between independent variables and e-commerce sales. However, the lack of statistical significance for x4 highlights the possibility that some factors may have a weaker or more complex influence on consumer behavior.

When comparing the OLS model with the initial multiple linear regression model used to evaluate R^2 and adjusted R^2, we observe consistency in the strength of the relationships identified. Both models yield high R^2 values, reinforcing the robustness of the independent variables in explaining e-commerce trends. However, the OLS regression provides a more detailed breakdown of coefficient significance, standard errors, and confidence intervals, allowing for a more nuanced understanding of individual variable contributions. Additionally, the F-statistic and p-values in the OLS model offer

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

further validation of the model's overall significance. This suggests that while the initial regression model provided a strong general evaluation of explanatory power, the OLS model refines this analysis by offering greater interpretability and statistical rigor.

8. Discussions and conclusions

This study provides an in-depth analysis of the U.S. e-commerce market from 2010 to 2024, exploring the impact of various economic and technological factors on consumer behavior. The findings highlight significant relationships between macroeconomic conditions and online shopping trends, reinforcing the idea that e-commerce activity is not only a reflection of technological advancements but also deeply embedded in broader economic dynamics.

A key insight from the study is the strong correlation between GDP and e-commerce sales. The results suggest that as economic growth accelerates, consumers have more disposable income, which translates into increased online spending. This relationship underscores the fundamental role of economic expansion in driving digital commerce. Similarly, inflation, measured through the Consumer Price Index (CPI), exhibits a positive correlation with e-commerce sales, indicating that rising prices influence consumer purchasing patterns. A plausible explanation for this trend is that inflationary pressures may push consumers toward online platforms in search of better deals, discounts, and cost-saving opportunities, particularly during periods of economic uncertainty.

Conversely, the Unemployment Rate shows a moderately strong negative correlation with e-commerce activity, which suggests that as job insecurity rises, consumer spending contracts. The connection between employment stability and e-commerce is crucial, as it highlights the direct impact of labor market conditions on digital consumption. However, despite this negative correlation, e-commerce continued to grow even in periods of economic downturn, indicating that other factors (such as increased reliance on digital platforms and evolving consumer habits) may have mitigated the adverse effects of unemployment on online shopping.

Beyond economic variables, technological infrastructure plays a vital role in shaping the e-commerce market. Internet penetration demonstrates a positive correlation with online sales, reinforcing the notion that access to digital platforms is a prerequisite for participation in e-commerce. However, its relative influence appears to be weaker compared to economic indicators. While high internet penetration rates have made e-commerce widely accessible, the data suggests that consumer spending decisions are still predominantly influenced by economic conditions rather than merely by the availability of online shopping platforms. This finding is significant as it highlights that while digital connectivity is essential for enabling e-commerce, it is not the sole determinant of its growth.

One of the most striking findings of the study is the sustained increase in e-commerce sales even after the gradual lifting of pandemic-related restrictions. During the COVID-19 crisis, lockdowns and physical distancing measures forced consumers to rely heavily on online shopping. However, contrary to expectations, e-commerce did not experience a sharp decline once restrictions were eased. Instead, the market continued to expand, surpassing both pre-pandemic and pandemic-period sales levels. This observation suggests that the pandemic accelerated a fundamental shift in consumer behavior, leading to a long-term preference for digital transactions. The convenience, accessibility, and efficiency of online shopping may have reshaped consumer habits to such an extent that even when physical stores became fully accessible again, many consumers continued to favor e-commerce.

This shift raises important questions about the factors sustaining e-commerce growth in the post-pandemic period. While the initial surge was clearly driven by necessity, the continued expansion indicates that consumers have become more accustomed to the advantages of digital shopping. Factors such as wider product availability, competitive pricing, and personalized shopping experiences facilitated by data analytics and AI-driven recommendations may have contributed to this trend. Additionally, the growing integration of e-commerce with logistical and financial services (such as faster delivery options, seamless payment gateways, and enhanced customer service) has likely reinforced consumer reliance on digital platforms.

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

From a predictive modeling perspective, the study's multiple linear regression model demonstrates strong performance in explaining e-commerce trends. The high R-squared and adjusted R-squared values indicate that the independent variables incorporated in the model account for a substantial proportion of the variance in e-commerce sales. Furthermore, the F-statistic and its corresponding p-value confirm the overall statistical significance of the model, reinforcing its reliability in predicting future e-commerce trends. The model's effectiveness suggests that economic indicators remain powerful tools for understanding and forecasting online shopping behavior, providing valuable insights for policymakers, businesses, and investors.

Furthermore, the cross-validation analysis underscores the importance of appropriate data scaling. The research findings indicate that scaling both the independent and dependent variables yields a significantly more stable and reliable model, minimizing numerical instability and ensuring balanced error contributions. This approach, aligned with recent methodologies (Bami, Behnampour, & Doosti, 2025; Liang et al., 2025), not only enhances the model's predictive performance but also reinforces the validity of the identified relationships between macroeconomic indicators and e-commerce trends.

Additionally, the study underscores the complex interplay of multiple factors in shaping e-commerce trends. While economic variables such as GDP and CPI emerge as the most dominant influences, other factors (including consumer sentiment and internet penetration) also contribute to the broader picture. Although consumer sentiment demonstrates a weaker correlation with e-commerce sales, its statistical significance, with a p-value of 0.035, suggests that psychological and behavioral factors still play a role in digital commerce. This highlights the importance of considering not only hard economic data but also softer variables related to consumer confidence, trust in online platforms, and perceived financial security when analyzing e-commerce trends.

In conclusion, this research identifies GDP, CPI, and the Unemployment Rate as the most influential factors driving consumer behavior in the U.S. e-commerce market. The strong correlations, statistically significant coefficients, and predictive reliability of these variables confirm their central role in shaping digital commerce trends. The findings suggest that during the COVID-19 crisis, consumer behavior (particularly online shopping patterns) was largely driven by macroeconomic conditions. However, the continued growth of e-commerce beyond the pandemic period indicates that structural changes in consumer preferences and technological adoption have played a crucial role in sustaining this expansion. As the e-commerce market continues to evolve, future research should explore additional factors such as digital marketing strategies, shifts in consumer trust, and the role of emerging technologies in further transforming online shopping behaviors.

References

- Anvari, R. D., & Norouzi, D. (2016). The impact of e-commerce and R&D on economic development in some selected countries. Procedia-Social and Behavioral Sciences, 229, 354-362. https://doi.org/10.1016/j.sbspro.2016.07.146
- Bami, Z., Behnampour, A., & Doosti, H. (2025). A New Flexible Train-Test Split Algorithm, an approach for choosing among the Hold-out, K-fold cross-validation, and Hold-out iteration. arXiv preprint arXiv:2501.06492. https://doi.org/10.48550/arXiv.2501.06492
- Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley & Sons.
- Bouchetara, M., Zerouti, M., & Zouambi, A. R. (2024). Leveraging Artificial Intelligence (AI) in Public Sector Financial Risk Management: Innovations, Challenges, and Future Directions. EDPACS, 69(9), 124–144. https://doi.org/10.1080/07366981.2024.2377351
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. (2013). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Third Edition. Taylor and Francis. https://doi.org/10.4324/9780203774441
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Pearson Education Limited.

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0036

- Hyndman, R. J. (1996). Computing and graphing highest density regions. The American Statistician, 50(2), 120-126. https://doi.org/10.1080/00031305.1996.10474359
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, No. 1). Springer. https://doi.org/10.1007/978-3-031-38747-0
- Liang, Z., Cai, L., Wang, S. et al. (2025). K-fold cross-validation based frequentist model averaging for linear models with nonignorable missing responses. Stat Comput, 35, 18. https://doi.org/10.1007/s11222-024-10554-x
- Li, K., Rubungo, A.N., Lei, X. et al. (2025). Probing out-of-distribution generalization in machine learning for materials. Commun Mater, 6, 9. https://doi.org/10.1038/s43246-024-00731-w
- Loxton, M., Truskett, R., Scarf, B., Sindone, L., Baldry, G., & Zhao, Y. (2020). Consumer behaviour during crises: Preliminary research on how coronavirus has manifested consumer panic buying, herd mentality, changing discretionary spending and the role of the media in influencing behaviour. Journal of Risk and Financial Management, 13(8), 166. https://doi.org/10.3390/jrfm13080166
- Montgomery, D. C. (2017). Design and analysis of experiments. John Wiley & Sons.
- Mohammedi, W., Mgadmi, N., Abidi, A. et al. (2025). The impact of the digital economy on sustainable development in the face of geopolitical risks. DESD, 3, 1. https://doi.org/10.1007/s44265-024-00050-5
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.
- The National Bureau of Economic Research. (2021, September). A roadmap for reopening the economy. https://www.nber.org/topics/covid-19
- United Nations Conference on Trade and Development (UNCTAD). (2020, April 21). COVID-19 and e-commerce: Impact on businesses and policy responses.

Data set sources

- E-commerce sales: https://www.statista.com/statistics/187443/quarterly-e-commerce-sales-in-the-the-us/
- GDP: https://www.statista.com/statistics/188105/annual-gdp-of-the-united-states-since-1990/
- Unemployment rate: https://www.statista.com/statistics/263710/unemployment-rate-in-the-united-states/
- Consumer Price Index (CPI): https://www.statista.com/statistics/190974/unadjusted-consumer-price-index-of-all-urban-consumers-in-the-us-since-1992/
- Internet penetration rate: https://www.statista.com/statistics/209117/us-internet-penetration/
- Consumer sentiment index: https://www.statista.com/statistics/216507/monthly-consumer-sentiment-index-for-the-us/