TASHKENT STATE UNIVERSITY OF ECONOMICS

International journal of topics

INNOVATION SCIENCE AND TECHNOLOGY

ISSUE 6, PART I / JUNE 2026

Acceptance of papers

Acceptance of papers

Published monthly

Topics

economics, technology, social sciences

OpenAIRE

Google Scholar

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INTERNATIONAL CENTRE

ISSN 3060-5229

Visit the website
t.me/scopus_IST2100

Electronic publication, Issue 6.
Approved for publication on June, 2026.

THE SCIENTIFIC-POPULAR ELECTRONIC JOURNAL **"INNOVATION SCIENCE AND TECHNOLOGY"** HAS BEEN REGISTERED UNDER THE NUMBER **C-5669633** BY THE AGENCY FOR INFORMATION AND MASS COMMUNICATIONS (AOKA) OF THE REPUBLIC OF UZBEKISTAN, EFFECTIVE FROM OCTOBER 9, 2024.

CONTACTS

Phone: **+998 50 737 87 88**

Website: https://ist-journal.uz

Email: innovationist2025@gmail.com

The scientific electronic journal “Innovation Science and Technology” has been included in the list of scientific publications recommended for the publication of main scientific results of dissertations for the award of PhD and DSc degrees in economics and technical sciences, in accordance with the Resolution No. 370 of the Presidium of the Higher Attestation Commission of the Republic of Uzbekistan, dated May 8, 2025.

CONTENTS

CONTENTS

CONTENTS

CONTENTS

IDENTIFICATION OF FACTORS INFLUENCING RETAIL SALES

Markova Mariya Stanislavovna
Computing Department, School of Law, Technology and Education
Westminster International University in Tashkent

Pooja
Computing Department, School of Law, Technology and Education
Westminster International University in Tashkent
Email: ppooja@wiut.uz

Abstract: The retail sector is an integral part of the U.S. economy that continues to grow. It was reported that sales revenue achieved $7.2 trillion in 2024. This study will use a two-year retail sales dataset including 200,000 transactions across different regions of the United States in order to analyze sales performance and identify the key factors influencing retail revenue. Identification of factors influencing revenue in retail is important. Because it helps to make proper decision-making. To satisfy research objectives, exploratory data analysis (EDA) was applied to understand the structure of data, define relationship patterns, and do data preprocessing to work with cleaned data. Also, hypothesis testing was made to analyze important variations and relationships among selected variables. Moreover, simple and multiple regressions were applied alongside ordinary least squares (OLS) to see the impact results of some variables on the revenue. The regression results show that Unit Price, Quantity and category of products have an impact on the retail revenue. The results also show that usage of multiple predictors suggests greater explanatory power and increased accuracy of prediction compared to models that rely on single predictors. The multiple regression model was chosen as the main approach for predictive analysis and revenue forecasting.

Keywords: retail, hypothesis, forecasting, data-processing, regression.

Annotatsiya: Chakana savdo sektori AQSh iqtisodiyotining oʻsishda davom etayotgan ajralmas qismidir. Maʼlumotlarga koʻra, 2024-yilda savdo daromadi 7,2 trillion dollarga yetgan. Ushbu tadqiqotda AQShning turli hududlari boʻyicha 200 000 ta tranzaksiyani oʻz ichiga olgan ikki yillik chakana savdo maʼlumotlari toʻplamidan foydalanilib, savdo koʻrsatkichlarini tahlil qilish hamda chakana savdo daromadiga taʼsir qiluvchi asosiy omillar aniqlandi. Chakana savdoda daromadga taʼsir qiluvchi omillarni aniqlash muhim ahamiyatga ega, chunki bu toʻgʻri va asoslangan qarorlarni qabul qilishga yordam beradi. Tadqiqot maqsadlariga erishish uchun maʼlumotlar tuzilishini tushunish, yashirin qonuniyatlarni (tendensiyalarni) aniqlash va tozalangan maʼlumotlar bilan ishlash (maʼlumotlarni dastlabki qayta ishlash) uchun boshlangʻich maʼlumotlar tahlili (EDA) qoʻllanildi. Shuningdek, tanlangan oʻzgaruvchilar oʻrtasidagi muhim oʻzgarishlar va munosabatlarni tahlil qilish uchun gipotezalarni sinash amalga oshirildi. Bundan tashqari, baʼzi oʻzgaruvchilarning daromadga taʼsirini baholash uchun oddiy va koʻp oʻzgaruvchan regressiya modellari oddiy eng kichik kvadratlar (OLS) bilan birga qoʻllanildi. Regressiya tahlili natijalari shuni koʻrsatadiki, mahsulot birligining narxi, miqdori va mahsulot kategoriyalari chakana savdo daromadiga sezilarli darajada taʼsir koʻrsatadi. Natijalar shuningdek, bir nechta mustaqil oʻzgaruvchilardan foydalanilgan modellar bitta oʻzgaruvchiga asoslangan modellarga nisbatan yuqoriroq tushuntirish kuchi va prognoz aniqligini taʼminlashini koʻrsatadi. Shu sababli, koʻp oʻzgaruvchili regressiya modeli prediktiv tahlil va daromadlarni bashorat qilish uchun asosiy yondashuv sifatida tanlandi.

Kalit soʻzlar: chakana savdo, gipoteza, prognozlash, maʼlumotlarni qayta ishlash, regressiya.

Аннотация: Розничная торговля является неотъемлемой частью экономики США, которая продолжает расти. Сообщается, что в 2024 году выручка от продаж достигла 7,2 триллиона долларов. В исследовании используется набор данных о розничных продажах в США за два года, включающий 200 000 транзакций, для анализа показателей продаж и определения ключевых факторов, которые влияют на выручку розничной торговли. Выявление факторов, влияющих на выручку в розничной торговле, имеет важное значение, поскольку помогает принимать правильные решения. Для достижения целей исследования был применен разведочный анализ данных (EDA) для понимания структуры данных, определения закономерностей взаимосвязей и предварительной обработки данных для работы с очищенными данными. Также была проведена проверка гипотез для анализа важных вариаций и взаимосвязей между выбранными переменными. Кроме того, были применены простая и множественная регрессия, а также метод наименьших квадратов (OLS) для оценки влияния некоторых переменных на выручку. Результаты регрессионного анализа показали, что цена за единицу товара, количество и категория товаров оказывают значительное влияние на выручку розничной торговли. Результаты также показывают, что использование нескольких независимых переменных обеспечивает более высокую объяснительную способность и точность прогнозирования по сравнению с моделями, основанными на отдельных переменных. В связи с этим модель множественной регрессии была выбрана в качестве основного подхода для прогнозного анализа и прогнозирования выручки.

Ключевые слова: розничная торговля, гипотеза, прогнозирование, обработка данных, регрессия.

INTRODUCTION

Retail is a big sector that suggests a variety of services and products to clients. And, this sector is still continuing to increase. Also, workplaces are created by the retail industry. According to Hasan et al (2025) retail has become an important part of the U.S. economy. Overall sales revenue was reported at $7.2 trillion, in 2024 and worker numbers reached more 32 million. A recent report of Deloitte advises that the total revenue of the top 250 world retailers in 2015 was 4.21 trillion dollars. In the meantime, Walmart, the biggest retailer in the world, has revenue of more than 400 billion dollars, topping the Fortune 500 list. The basic objective in this paper is to detect main patterns that influence revenue in retail transaction dataset. It is helpful to define these patterns due to the reason that retailers will be able to make proper and informed decisions when they know the exact key factors that influence the revenue. As noted by Yeoman (2024) good revenue management relies on analysis of difficult patterns among price, demand, quality and operational constraints. Moreover, Alikhani et al (2019) writes retailers depend on a variety of strategies as revenue management to raise their income to stay competitive.

REVIEW OF LITERATURE ON THE SUBJECT

Retail sales forecasting and factor identification have become increasingly important in modern business analytics due to the growing complexity of consumer markets, competitive pricing, and data-driven decision-making. Researchers have emphasized that retail sales are influenced by a combination of pricing strategies, customer demand patterns, promotional activities, seasonality, and macroeconomic conditions.

Revenue management and pricing strategies are major determinants of retail performance. Hajy Alikhani et al. (2019) demonstrated that agent-based revenue management models improve seller pricing decisions and profitability in retail industries. Similarly, Yeoman (2024) argued that revenue management techniques support strategic business decisions by enabling retailers to optimise inventory, pricing, and demand forecasting. Hasan et al. (2025) further highlighted how investment trends, mergers and acquisitions, and market growth dynamics influence retail stock performance and broader retail market behaviour.

Regression analysis remains one of the most widely used statistical techniques for identifying factors affecting retail sales. Maulud and Abdulazeez (2020) provided a comprehensive review of linear regression applications in machine learning and predictive analytics, emphasizing its usefulness in modelling relationships between dependent and independent variables. Kumar, Shah, and Sahani (2025) also discussed the effectiveness of regression forecasting models in economics, particularly for trend estimation and sales prediction. Ordinary Least Squares (OLS) regression methods, commonly implemented through statistical software such as statsmodels, are widely adopted for estimating retail demand relationships (GeeksforGeeks, 2025).

Several studies have emphasized the importance of exploratory data analysis (EDA) and data preprocessing prior to modelling retail sales data. Shaik, Rao, and Vardhan (2025) identified EDA as a critical step in understanding distributions, identifying anomalies, and detecting hidden patterns in datasets. West (2022) recommended the use of logarithmic transformations for handling skewed business data distributions, while Ally, Mrutu, and Mathew (2025) introduced geometric approaches to skewed lognormal distributions that can improve interpretation of non-normal retail sales data. Alshamrani (2025) further demonstrated that machine learning-enhanced Box–Cox transformations can significantly improve predictive model performance.

Outlier detection has also been recognized as essential in retail sales modelling because extreme values can distort regression estimates and forecasting accuracy. Malpani (2019) proposed the interquartile range (IQR) approach for single-dimensional outlier detection, whereas Dallah et al. (2025) empirically evaluated relative range techniques for identifying outliers in datasets. Kowalskie (2025) examined the impact of outliers on linear regression models and suggested correction strategies to improve model reliability.

Sampling quality and dataset integrity are equally important for reliable retail sales analysis. Liu and Zhang (2024) reviewed sampling approaches for big data profiling, while Yao and Wang (2021) explored optimal subsampling methods for massive datasets. Giri (2025) discussed appropriate sampling techniques and sample size determination for statistical studies, and Tutz (2022) highlighted the integration of probability and non-probability samples in regression modelling. Lock and Ansari (2025), together with Sarracino and Mikucka (2016), warned that duplicated observations may introduce estimation bias and distort analytical findings.

Statistical inference and probability distributions also contribute to accurate retail sales modelling. Musadi and Kurniawati (2023) reviewed normal distribution approximations through binomial and Poisson distributions, while Zhang (2025) discussed applications of Poisson distribution in predictive modelling. Kwak (2023) argued that statistical significance should not rely solely on the conventional $p < 0.05$ threshold, emphasizing broader interpretation of results. Ruxton and Neuhäuser (2010) further explained conditions for using one-tailed hypothesis testing, and Yagin, Pinar, and Fernandes (2024) highlighted the importance of effect sizes in interpreting statistical outcomes. The reviewed studies collectively support the use of regression analysis, exploratory data analysis, data transformation, and robust sampling procedures for identifying and modelling factors influencing retail sales.

RESEARCH METHODOLOGY

The study is based on a retail sales dataset containing transactional data from 2023–2024. Data were collected from a publicly available source and preprocessed through duplicate checking, missing value detection, random sampling, outlier treatment, and Box–Cox transformation. The data were analysed using exploratory data analysis, correlation analysis, hypothesis testing, and simple and multiple regression techniques to identify the key factors influencing retail revenue.

ANALYSIS AND RESULTS

This section aims to explain the data source, preprocessing data, exploratory data analysis and usage of regression technique to define relationship patterns. Each mentioned component before is discussed individually and separately below.

Data overview: The dataset mostly includes data about retail sales in the U.S from 2023 to 2024. It consists of 200,000 rows of data and 14 columns that demonstrate a big choice of information mostly for customer orders, product details and finance metrics. The next table will show the names of columns and their purposes (Table 1).

Table 1. Dataset description[1]

Column name	Purpose
Order_ID	Unique identifier for each order
Order_Date	Data of transaction
Customer_Name	Name of customer
City	City of Customer
State	State of the customer
Region	Region (East, West, South, Centre)
Country	Country (United States)
Category	Broad product category (e.g Accessories, Clothing & Apparel)
Sub_Category	Subdivision of category (e.g., Sportswear, Bags)
Product_Name	Product description
Quantity	Units purchased
Unit_Price	Price per unit (USD)
Revenue	Total sales amount (Quantity × Unit Price)
Profit	Net profit earned from the transaction

The table above represents the basic information about the retail dataset. It includes 5 numeric variables (2 with type of integer and 3 with type of float), 9 categorical variables and 1 date column

Preprocessing Data: Data preprocessing is an important part that ensures that the dataset is clean, reliable and proper for statistical analysis and modelling. In this report, preprocessing focuses on identification of duplicated data, checking for missing values and selection of sample size for future analysis. Using these steps assists to increase the quality of data and ensures that analytical methods are biased on accurate and reliable information. To find duplicates the below code was typed refer to figure 1. Unique data frame was created from

1 Source: developed by the author.

existed and to see difference the length of initial data frame subtracts length of unique data frame. The output of code demonstrates that the dataset lacks duplicated data (Figure 1).

Duplicate Data Handling

```
df_unique = df.drop_duplicates()
```

```
print(f"\nInitial count of dataset: {len(df)}\n")
print(f"Unique count of dataset: {len(df_unique)}\n")
print(f"Difference between Initial and New DataFrames - {len(df)-len(df_unique)}")
```

```
Initial count of dataset: 200000

Unique count of dataset: 200000

Difference between Initial and New DataFrames - 0
```

Figure 1. Duplicate Data Handling[2]

A missing value observation: To define existence of missing values panda's commands as isnull alongside sum were implemented in python by calculating the sum of null values if such exists inside of the dataset. The output of code demonstrates that each column in the dataset does not contain null values.

Sampling Data: It was mentioned in the data introduction part that the dataset has totally 200,000 rows of data. There should be enough computation power to make correct analysis. Although, it can cost a lot. As noted by Liu and Zhang (2020) it is not usually possible to have enough computing power. They also write that using small resources to make computing in demand. Sample data is not effective as a full dataset, but it is not important. Because, a subset of data can help to use regression analysis without issues. Yao and Wang (2021), Dereziński et al (2018) highlights that sampling allows unbiased regression results even with a subset of data. Giri (2024) writes about formulas of sampling for both finite and infinite population of data. Since there are 200,000 rows of transactional retail data and a two years dataset is a finite population. n = 200,000, e = 0.05 (precession), next calculations look like:

$$n = 200{,}000 \;/\; (1 + 200{,}000 * (0.05)\ \hat{}2) = 200{,}000/501 = 399{,}201 = 400. \qquad \text{eq 1}$$

By using this calculation, the sample size of data of 400 rows will be used for hypothesis testing and regression modeling. Sampling is implemented to select a subset of data from a big dataset. It helps to work effectively and without bias. Random sampling is used to select 400 rows. Each value has an equal chance to be chosen in random sampling. It is a known method that is used in regression modeling because of decreasing systematic bias's selection and provides a number of sample data.

Exploratory Data Analysis: According to Shaik et al. (2021), exploratory data analysis is a definite, operative, and practical method that incorporates both science and statistics. It is well known that the EDA process aids in the discovery of unknown information, knowledge, and hidden artifacts.

Correlation Analysis: A linear relationship between two variables is measured by the correlation coefficient, which ranges from -1 to 1. Strong positive correlations are indicated by values near 1, strong negative correlations are indicated by values close to -1, and little to no linear correlation is indicated by values near 0. From figure 2 it can be seen that correlation was implemented only for numeric variables. Internal command pandas command as corr was used and parameters as numeric_only=True. The target variable is Revenue. From correlation Revenue have a strong relationship with Profit which is 0.84 alongside Unit_Price that equals to 0.71 and Quantity with moderate relationship 0.58. On the other hand, Order_ID lacks a helpful relationship. These results demonstrate that Quantity and Unit_Price can be potential predictors of Revenue. In addition, Eta Squared (η^2) will be used to define strong relationships between Category and Revenue, Sub-category and Revenue, Region and Revenue. Because correlation can also lack a proper relationship. Next ongoing images (Figure 3,4) will depict bar charts and pie charts for the above combination of relationships.

The Figure 3 returns a bar chart of the category-wise revenue generated by different product categories. Categories of Electronics and Home & Furniture demonstrate substantial results of revenue than accessories and clothing alongside apparel. It shows customer demand and high sales performance in these product groups. Figure 4 depicts the proportion of revenue among different regions of the USA. It can be observed that East demonstrates a high proportion of revenue that is almost 32%. Meanwhile other regions contribute relatively

2 Source: developed by the author.

similar portions. Meaning that East performs better in terms of retail revenue compared with others. Figure 5 shows the revenue by different product sub-categories. Bedding and Laptops make the highest revenue among all sub-categories. In the meantime, Bags and Wearable Accessories generate lower revenue. This indicates that technology-related and household product categories have stronger retail sales performance and higher customer demand than other sub-categories (Figure 2, 3, 4, 5).

Calculation of correlation metrics

```
correlation = df.corr(numeric_only=True)
```

```
correlation
```

	Order_ID	Quantity	Unit_Price	Revenue	Profit
Order_ID	1.000000	0.001759	-0.002346	0.001494	0.002010
Quantity	0.001759	1.000000	0.007300	0.577178	0.602484
Unit_Price	-0.002346	0.007300	1.000000	0.706247	0.528700
Revenue	0.001494	0.577178	0.706247	1.000000	0.841345
Profit	0.002010	0.602484	0.528700	0.841345	1.000000

Figure 2. Correlation between Order ID, Quantity, unit Price, revenue and Profit[3]

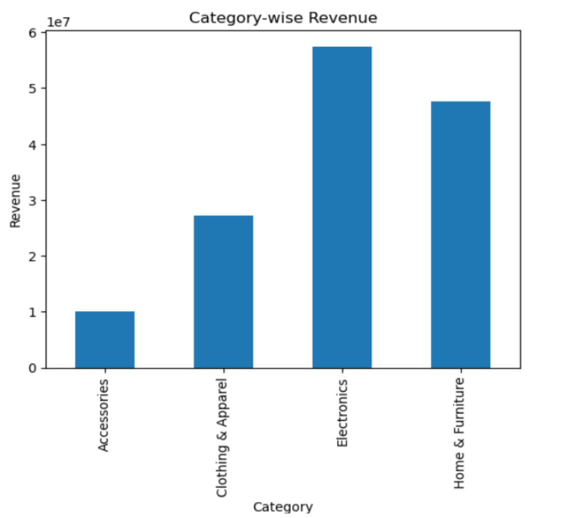

Figure 3. Bar Chart of Category and Revenue[4]

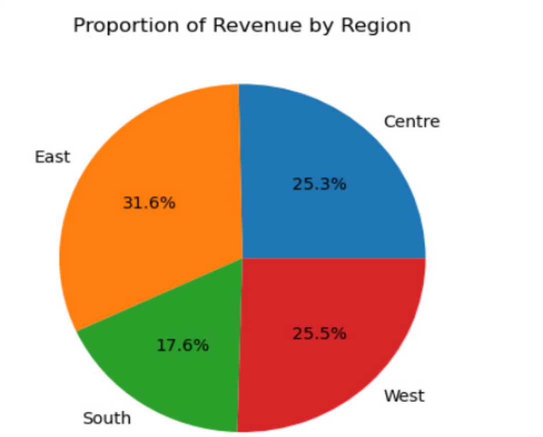

Figure 4. Pie Chart of Sub Categories and Revenue[5]

3 Source: developed by the author.
4 Source: developed by the author.
5 Source: developed by the author.

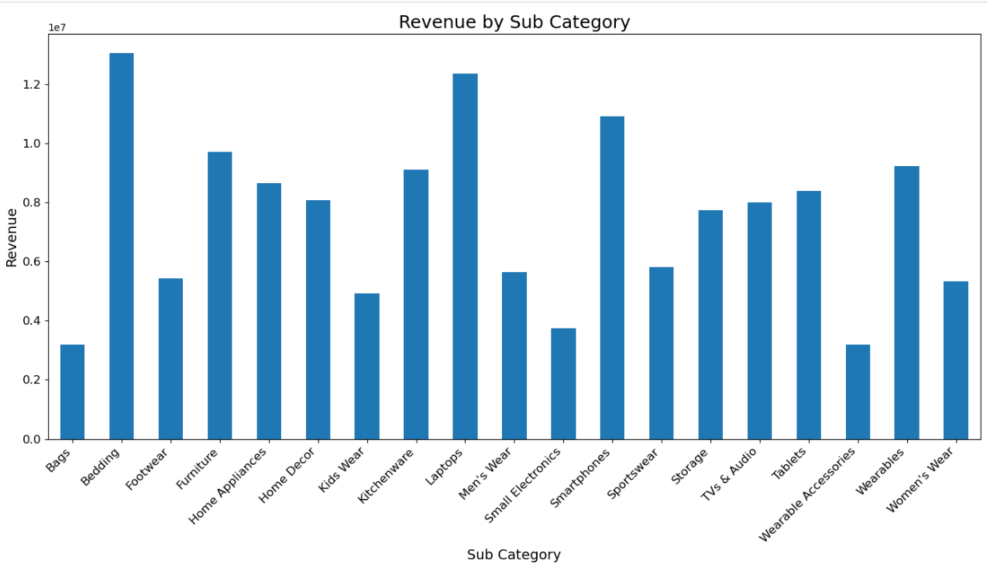

Figure 5. Bar Chart of Sub Categories and Revenue[6]

By using calculation of Eta Squared and significance of effect there were investigations of two types relationships with large effect. Category, $\eta^2 = 0.21$ and Sub-Category, $\eta^2 = 0.24$ both have a large effect, meaning that these two categorical variables explain a great percentage of the variation in Revenue. Region, $\eta^2 = 0.01$ has a very small effect, hence, this variable is considered to have little influence. Such variables should be included as predictors for Revenue in regression models, in addition to numeric variables such as Quantity and Unit_Price.

Outliers' detection and transformation: Extreme values, or outliers, meaningfully skew results on the regression-type statistical methods. When performing an analysis for outliers within a dataset consisting primarily of quantities, revenues and unit prices, Quantities and Revenues had the highest percentage of extreme values at 8.87%, 7.53%, and 6.59%, respectively. This implies that the dataset contains occasional data pertaining to very large transactions or extremely high sales/profits. Conversely, Unit Prices had an extremely low percentage of outliers of only 0.82%. The Box-Cox transformation was applied to reduce the existence of skewness and improve normality of numeric variables in the dataset before making projects analysis. The transformation is good for positively skewed data and can stabilize variance in data source. The box-cox was applied for next numeric variables: Revenue, Unit Price and Profit. Algorithm/pseudocode for box-cox transformation in the figure 9. The method decreases skewness and makes the data distribution more symmetric. In order to see the implementation of transformation the below figures 6, 7, 8 were represents histograms of transformed numeric columns (Figure 6, 7, 8, 9).

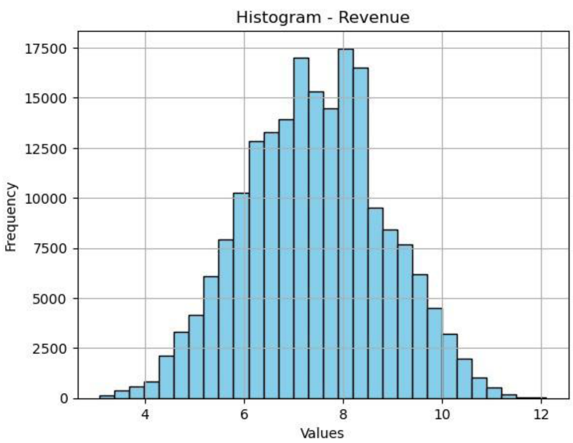

Figure 6. Histogram of transformed Revenue[7]

6 Source: developed by the author.
7 Source: developed by the author.

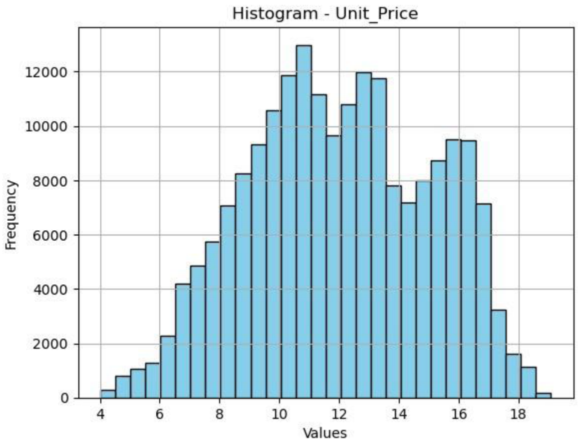

Figure 7. Histogram of transformed Unit_Price[8]

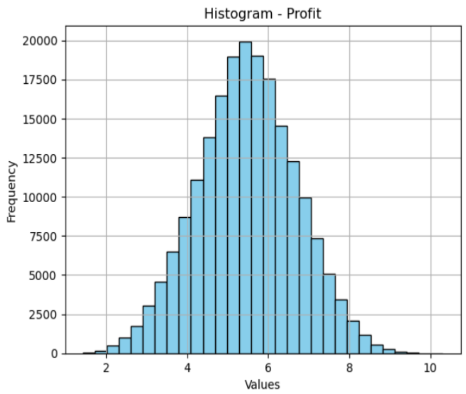

Figure 8. Histogram of transformed Profit[9]

START

1. Select numerical columns
2. FOR each column:
a. Remove missing values
b. Check for zero or negative values
c. IF values ≤ 0:
Add shift constant to make data positive
d. Apply Box–Cox transformation
e. Save transformed column
f. Calculate lambda (λ) and skewness
3. END FOR
END

Figure 9. Algorithm for Box-Cow Transformation[10]

Hypothesis Testing: Statistical testing, or hypothesis testing, is commonly used in quantitative research to determine if differences found between samples are the result of chance or actual effects that can be supported with statistical inference. In this case, significance levels of 0.01, 0.05, and 0.10 will be used for all of the hypotheses evaluated in this study. The statistical test method used in this study is a two-tailed test, which is appropriate when it is possible to have effects in both directions and when detecting a significant difference. The following hypotheses will be tested in this study (Table 2; Figure 10):

8 Source: developed by the author.
9 Source: developed by the author.
10 Source: developed by the author.

Table 2. Research Hypotheses and Statistical Hypothesis Statements[11]

Hypothesis	H0 and H1
Average revenue differs significantly across product categories	H0: There is no difference across average revenue from product categories H1: Average revenue different across different product categories
Unit Price and Quantity are correlated	H0: There is no correlation between Unit Price and Quantity H1: There is correlation between Unit Price and Quantity
Revenue differs significantly between 2023 and 2024	Ho: Revenue do not different between 2023 and 2024 H1: Revenue different between 2023 and 2024
High valued categories generate significantly higher revenue per transaction	Ho: High valued categories do not generate much revenue H1: High valued categories make more revenue

```
F-statistic = 43.0218
P-value = 0.000000

✅ Significant at α = 0.1 → Reject H₀ (revenue differs by category)
✅ Significant at α = 0.05 → Reject H₀ (revenue differs by category)
✅ Significant at α = 0.01 → Reject H₀ (revenue differs by category)
```

Figure 10. Test hypothesis 1 Average revenue differs significantly across product categories[12]

One Way ANOVA that was performed on a random sample of 400. Observations the result of a statistically significant difference will be shown in the average revenue by product category. In the example below the code samples df.sample and then calculates the average revenue across product categories. The `f_oneway` function from the SciPy library was used to perform the ANOVA calculations. It can be seen from the resultant `F` statistic of 43.0218 and a `p` value of 0.000000 which are both less than all the tested levels of significance ($\alpha = 0.10, 0.05, 0.01$) that the null hypothesis stating that average revenue is equal across all product categories can therefore be rejected, further confirming that product categories do indeed exhibit statistically significant differences in terms of revenue performance (Figure 11).

```
Correlation (r) = 0.1104
P-value = 0.027241

✅ Significant at α = 0.1 → Reject H₀ (correlation exists)
✅ Significant at α = 0.05 → Reject H₀ (correlation exists)
❌ Not significant at α = 0.01 → Fail to reject H₀
```

Figure 11. Test Hypothesis 2 Unit Price and Quantity are correlated[13]

Figure 11 provides insight into the extent to which Unit Price is correlated with Quantity based on results of a Pearson correlation test performed on a randomly selected sample of 400 observations created through df.sample. The correlation coefficient produced by the pearsonr function is $r = 0.1104$, indicating a negligible or weak positive association exists between Unit Price and Quantity. The calculated p-value of 0.027241 falls within the significance levels of $\alpha = 0.10$ and $\alpha = 0.05$, indicating that at both of these levels the null hypothesis can be rejected. However, since the estimated p-value does not fall below $\alpha = 0.01$, there is insufficient statistical evidence to reject the hypothesis that there is no correlation between Unit Price and Quantity at a 1% level. Therefore, while there may be a statistically significant level of correlation at moderately high levels of significance, this type of relationship does not demonstrate a strong correlation (Figure 12).

11 Source: developed by the author.
12 Source: developed by the author.
13 Source: developed by the author.

```
T-statistic = -0.4806
P-value = 0.631094

❌ Not significant at α = 0.1 → Fail to reject H₀
❌ Not significant at α = 0.05 → Fail to reject H₀
❌ Not significant at α = 0.01 → Fail to reject H₀
```

Figure 12. Test Hypothesis 3 Revenue differs significantly between 2023 and 2024[14]

Figure 12 shows the results of how to use pandas and scipy to statistically compare the revenue of 2023 and 2024. To do this, an independent t-test comparing the revenue for both years and displays the resulting T-statistic (-0.4806) and P-value (0.631094). It can be observed from the calculated P-value that it is much greater than the common significance levels (0.10, 0.05, 0.01) used to determine significance, and therefore, the conclusion indicated by the red «X» is that there is no significant difference between the revenue of 2023 and 2024 at any of these significance thresholds (Figure 13).

```
T-statistic: 10.5032
P-value: 0.000000

Conclusions at different alpha levels:
✅ At α = 0.1: Reject H₀ → Significant difference
✅ At α = 0.05: Reject H₀ → Significant difference
✅ At α = 0.01: Reject H₀ → Significant difference
```

Figure 13. Test Hypothesis 4 High valued categories generate significantly higher revenue per transaction[15]

Figure 13 represents the outcome of hypothesis 4 to analyze whether High Value Categories (the terms used for Electronics, and Home & Furniture) have substantially greater revenue generating activity than any other category of transaction. In this example, 400 rows of data are sampled and a Boolean column called Is_High_Category is produced to separate the revenue data from the sampled rows into two groups (high revenue versus low revenue). An independent t-test yields a large 'T' statistic with a corresponding small p-value. All three tested alpha levels (0.10, 0.05, and 0.01) showed that the p-value was considerably below the threshold; hence the null hypothesis is rejected at all three alpha levels. The green checkmarks show that there is sufficient statistical evidence that Higher Revenue Transactions (High Value Categories) have greater revenue generating potential than all other Transaction Categories at all three alpha levels.

Simple Regression Analysis: For simple regression each predictor column will be used in pairs with revenue to check how revenue connects with them individually. The first combination is Unit_Price and Revenue. The study used columns after box cox transformation and only 400 randomly selected rows of the dataset. After execution of python code there are the OLS regression outputs displayed on Figure 14 below (Figure 14).

14 Source: developed by the author.
15 Source: developed by the author.

```
                          OLS Regression Results
==============================================================================
Dep. Variable:          Revenue_boxcox   R-squared:                       0.727
Model:                             OLS   Adj. R-squared:                  0.726
Method:                  Least Squares   F-statistic:                     1059.
Date:                 Thu, 04 Dec 2025   Prob (F-statistic):          3.37e-114
Time:                         19:37:00   Log-Likelihood:                -452.17
No. Observations:                  400   AIC:                             908.3
Df Residuals:                      398   BIC:                             916.3
Df Model:                            1
Covariance Type:             nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const                2.4941      0.159     15.727      0.000       2.182       2.806
Unit_Price_boxcox    0.4092      0.013     32.540      0.000       0.385       0.434
==============================================================================
Omnibus:                        60.811   Durbin-Watson:                   2.068
Prob(Omnibus):                   0.000   Jarque-Bera (JB):               33.567
Skew:                            0.561   Prob(JB):                     5.14e-08
Kurtosis:                        2.131   Cond. No.                         53.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 14. Metrics of Simple Regression - Unit_Price and Revenue[16]

The output of code returns OLS Regression results that demonstrate that this regression model fits well due to R-squared equals 0.727. The independent variable Unit_Price_boxcox is significant because p < 0.001 and has a positive effect on the target variable Revenue. The intercept const is statistically significant as p = 0.000 which means it may contribute meaningfully. The Durbin-Watson value is 2.068 no autocorrelation in residuals. The overall model is significant due to the big measure of f statistic and p-value = 0. Second will be the relationship of the Quantity _boxcox variable with Revenue (Figure 15; 16).

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Revenue_boxcox   R-squared:                       0.349
Model:                             OLS   Adj. R-squared:                  0.348
Method:                  Least Squares   F-statistic:                     213.5
Date:                 Thu, 04 Dec 2025   Prob (F-statistic):           5.11e-39
Time:                         19:56:54   Log-Likelihood:                -625.78
No. Observations:                  400   AIC:                             1256.
Df Residuals:                      398   BIC:                             1264.
Df Model:                            1
Covariance Type:             nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const               6.6989      0.080     83.574      0.000       6.541       6.856
Quantity_boxcox     2.6032      0.178     14.612      0.000       2.253       2.953
==============================================================================
Omnibus:                        12.682   Durbin-Watson:                   1.846
Prob(Omnibus):                   0.002   Jarque-Bera (JB):               11.469
Skew:                           -0.354   Prob(JB):                      0.00323
Kurtosis:                        2.567   Cond. No.                         3.40
==============================================================================
```

Figure 15. Metrics of Simple Regression 2 - Quantity with Revenue[17]

16 Source: developed by the author.
17 Source: developed by the author.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         Revenue_boxcox   R-squared:                       0.338
Model:                            OLS   Adj. R-squared:                  0.333
Method:                 Least Squares   F-statistic:                     67.52
Date:                Thu, 04 Dec 2025   Prob (F-statistic):           2.77e-35
Time:                        20:35:28   Log-Likelihood:                -629.06
No. Observations:                 400   AIC:                             1266.
Df Residuals:                     396   BIC:                             1282.
Df Model:                           3
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                      5.9961      0.048    125.186      0.000       5.902       6.090
Cat_Accessories            0.4468      0.117      3.833      0.000       0.218       0.676
Cat_Clothing & Apparel     0.8691      0.093      9.327      0.000       0.686       1.052
Cat_Electronics            2.5868      0.106     24.386      0.000       2.378       2.795
Cat_Home & Furniture       2.0934      0.100     20.862      0.000       1.896       2.291
==============================================================================
Omnibus:                        4.531   Durbin-Watson:                   2.247
Prob(Omnibus):                  0.104   Jarque-Bera (JB):                4.344
Skew:                          -0.251   Prob(JB):                        0.114
Kurtosis:                       3.094   Cond. No.                     4.57e+15
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.42e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Figure 16. Metrics of Simple Regression 3 Category within Revenue[18]

In figure 15 it can be observed that this model fits not as well as the previous one as R squared is only 0.349. In the meantime, Quantity_boxcox has a statistical impact on Revenue (p=0.000). The intercept const is statistically significant as p = 0.00 which means it can contribute meaningfully. The Durbin-Watson value is 1.846 and can have mild positive autocorrelation. The overall model is significant due to big measures of f-statistics as 213.5 and p-value = 0.

The third relationship is the category within revenue. To achieve it, the category was converted to numbers by using the pandas command as get_dummies. Moreover, the logic was applied only for randomly selected 400 rows of sampled data and to each column was added the prefix 'Cat'. In figure 16 which is the output of code, it can be observed that this model is not good due to R-squared being only 0.338. In the meantime, Product categories (Accessories, Clothing & Apparel and everything in between) have statistical impact on Revenue (p=0.000). The intercept const is statistically significant as p = 0.00 which means it can contribute meaningfully. The Durbin-Watson value is 2.247 and can have no autocorrelation. The overall model is significant due to a big measure of f-statistics as 67.52 and p-value = 0.

And lastly the relationship between sub_category and revenue. Similar logic was implemented inside of code as was done to category and revenue relationship. The fit of the model was observed to be less satisfactory than that of the regression because the R2 value of the model is 0.388. Thus, it appears that subcategory accounts for a smaller amount of variance in revenue than the unit price did. Nevertheless, statistically significant subcategories (with a p-value less than 0.05) that have an influence on Revenue_boxcox include: Bags, Bedding, Furniture, Home Appliances, Laptops, Smartphones, TVs and Audio, Tablets and Wearables. Conversely, Footwear (p = 0.733) and Sportswear (p = 0.500) do not have a notable level of impact on revenue in this model. The intercept (with a p-value of 0.000) is statistically significant and should be considered when interpreting the model. Both the overall model (with an F-statistic of 13.44) and the Durbin-Watson statistic of 2.291 (close to 2, indicating that there is no significant autocorrelation in the residuals) indicate that the model is very significant (with an approximate p-value of 0).

Multiple Regression Analysis: To implement multiple regression logic there will be usage of several predictors for revenue to see overall combination of independent values to dependent ones. Predictors as UnitPrice, Quantity and Category will be used for revenue. The main reason to use these predictors is because Unit-Price and Quantity from EDA have good correlation with Revenue. Meanwhile, Categories from Eta Squared calculation have significance with revenue. Additionally, categories include less categorial values than sub category, which make it straightforward to use in multiple regression (Figure 17).

18 Source: developed by the author.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Revenue_boxcox   R-squared:                       0.988
Model:                             OLS   Adj. R-squared:                  0.988
Method:                  Least Squares   F-statistic:                     6428.
Date:                 Thu, 04 Dec 2025   Prob (F-statistic):               0.00
Time:                         21:43:12   Log-Likelihood:                 171.06
No. Observations:                  400   AIC:                            -330.1
Df Residuals:                      394   BIC:                            -306.2
Df Model:                            5
Covariance Type:             nonrobust
==========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                     2.0300      0.040     51.262      0.000       1.952       2.108
Unit_Price_boxcox         0.3902      0.004    105.815      0.000       0.383       0.397
Quantity_boxcox           2.2639      0.025     91.718      0.000       2.215       2.312
Cat_Clothing & Apparel    0.0293      0.024      1.220      0.223      -0.018       0.076
Cat_Electronics          -0.0262      0.032     -0.825      0.410      -0.089       0.036
Cat_Home & Furniture     -0.0355      0.029     -1.213      0.226      -0.093       0.022
==============================================================================
Omnibus:                        42.161   Durbin-Watson:                   1.965
Prob(Omnibus):                   0.000   Jarque-Bera (JB):              110.797
Skew:                            0.504   Prob(JB):                     8.72e-25
Kurtosis:                        5.373   Cond. No.                         72.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 17. Metrics for Multiple Regression[19]

A multiple regression model was developed (shown in Figure 17) which predicts the revenue (as transformed by the Box-Cox method) based on Unit_Price, Quantity and Category. The R-squared (0.988) and Adjusted R-squared (0.988) indicate very high correlations between the predicted and actual revenues. The overall model (as evidenced by the F-statistic = 6428) shows overwhelming significance (Unknown p value), the Durbin-Watson statistic's value (1.965) indicates that the model has not exhibited autocorrelation, and the Condition Number indicates (72.7) that the model does have some level of Multicollinearity.

Justification - The rationale for using a multiple regression analysis approach rather than just using a single regression analysis approach is based on the assumption that real-world business revenues are rarely determined by just one factor; instead, there are typically many interacting variables that jointly influence the ultimate outcomes, such as revenue generation. While Unit_Price explained over 72% of variance in revenue on its own (as illustrated in Figure 24), when combining Unit_Price with both Quantity and Category, the total explanatory power for this combined analysis reached almost 99%, thereby demonstrating how the interaction between these three variables contributes to revenue generation.

Forecasting values of 5, 10 and 30: Regression analysis is a valuable tool that supports the prediction and forecasting of future event outcomes. According to a study by Maulud et al (2020), one of the main purposes of regression is to determine the relationship between the independent and dependent variables, in addition to allowing prediction and forecasting of an event. By using the fitted regression model to estimate future revenues, estimates can be obtained based on combinations of the predictor variables (Unit_Price, Quantity, Category) that we supplied to the model. Each set of generated forecasts can be for any number of future revenues (5, 10, 30, etc.), but they all use the exact same code on python. For the 10 forecasts and the 30 forecasts, the same algorithm structure would be used with the data frame for each set containing more rows of synthetic predictor values.

The results of the model are presented graphically; the blue points on the plot demonstrate how closely the model's future revenue predictions align with the actual revenue recorded for all prior data sets analyzed for trend compliance and the green Xs represent how closely the future revenue predictions based upon the 5 new product scenarios align with the actual predicted revenue. In addition to the blue and green data points,

19 Source: developed by the author.

the dashed line is indicative of perfect prediction accuracy. The line graph in the lower left corner shows the relationship between actual revenue (red) and predicted revenue (green), indicating the accuracy of the prediction and the correlation between predicted and actual revenue (Figure 18).

```
Predicted Revenue (Box-Cox scale):
0     5.976666
1     7.220620
2     8.781803
3    10.155614
4    11.308102
dtype: float64
```

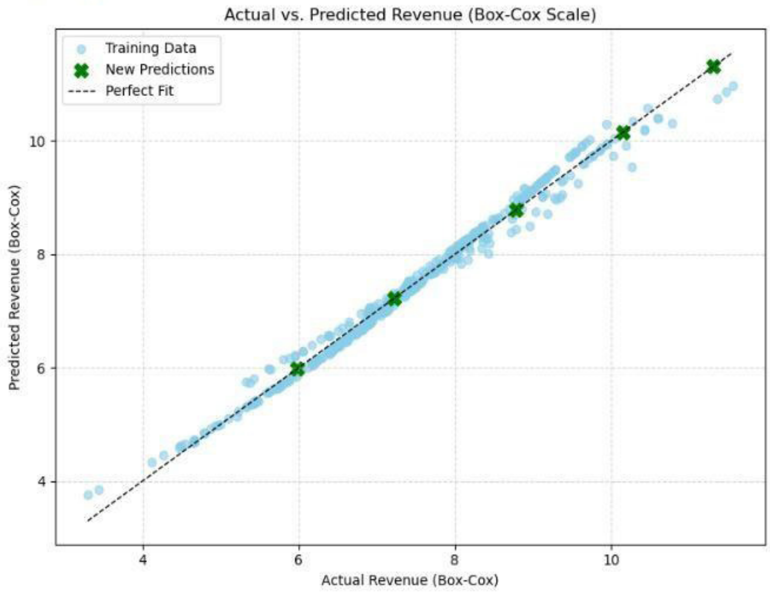

Figure 18. Line Graph of 5 values[20]

The box-Cox scaled revenue line graph depicted in figure 18 demonstrates the accuracy with which the model can predict revenue, as indicated by the distance of the blue dots from the diagonal ‘perfect fit’ line. This indicates the level of agreement between what was anticipated only to happen based upon the previous pattern of the actual data. The five new forecasts are represented by the green X’s which also fall directly within the established pattern and therefore provide confirmation of the reliability and usefulness of these models to estimate wages to be earned in the future.

The analysis of the retail dataset shows that income is mostly driven by Unit Price and Quantity, supported by strong correlations of 0.71 and 0.58, respectively. These results show that both pricing and sales volume are main contributors to revenue in retail transactions. To check the impact of categorical data, the Eta Squared technique was applied. The results define meaningful relationships for both Category (0.21) and Subcategory (0.24) suggesting that product classification has a moderate impact on revenue metric. Electronics and Home & Furniture categories made higher average revenue per transaction among other product groups. Also, hypothesis testing was applied to check statistical differences and relationships among variables. ANOVA testing confirmed statistically differences in average revenue across product categories, with $F \approx 43$ and $p < 0.001$. Pearson correlation analysis shows a weak but statistically positive relationship between Unit Price and Quantity ($r = 0.1104$, $p = 0.027241$) at significance levels $\alpha = 0.05$ and $\alpha = 0.10$. In contrast, the independent t-test comparing revenue metrics between 2023 and 2024 showed no statistically significant difference ($t = -0.4806$, $p = 0.631094$), showing relatively stable revenue among two years. Regression modeling further showed the importance of using multiple predictors. The multiple regression model produced an R^2 value of 0.988, indicating that approximately 98.8% of the variance in Box-Cox-transformed revenue was explained by the selected predictors. This result suggests very strong explanatory power and confirms the suitability of the model for predictive analysis. Forecasting analysis based on the multiple regression model also demonstrated strong predictive performance. Forecasted revenue values for periods 5, 10, and 30 closely followed historical revenue trends and aligned near the regression line. These outputs demonstrate that the regression model can be effectively applied for future revenue estimation and forecasting tasks.

Conclusions and suggestions

This paper depicts that retail gains mostly rely on the basic variables (Unit Price and sales volume). The

20 Source: developed by the author.

high R2 value suppose that other external factors had some impact on the revenue once price and quantity were used. For business, these results imply that revenue optimizations should have a much more focus on high-value category inventory than seasonal shifts from year to year. In the meantime, there are some limitations which should be considered. Firstly, the dataset has only 2 years of information that influences the lack of long-term market trends and seasonal fluctuations. Secondly, analysis mainly focuses on variables such as Unit Price, Quantity, Category, and Subcategory while not considering other variables. Thirdly, there is a linear relationship between variables that may lack to capture more difficult retail behaviour. The future work could involve working with big and diverse datasets and also include other variables to make more accurate predictions. And could use other machine learning techniques to make better forecasting to gain deeper analytic insights.

The dataset used for this work is available on Kaggle by the author Yash Yennewar, available under the Attribution 4.0 International (CC BY 4.0) license. It can be accessed at https://www.kaggle.com/datasets/yashyennewar/product-sales-dataset-2023-2024.

References:

1. Ally, S.J., Mrutu, H.A. & Mathew, R. (2025). Geometric approach to the skewed lognormal distribution. ResearchGate. Available at: https://www.researchgate.net/publication/387628104_Geometric_approach_to_the_ske wed_log-normal_distribution [Accessed 25 November 2025]
2. Alshamrani, S.S. (2025). Machine Learning Techniques Improving the Box–Cox Transformation in Breast Cancer Prediction. Electronics. Available at: https://www.mdpi.com/2079-9292/14/16/3173 [Accessed 25 November 2025]
3. Dallah, D., Sulieman, H., Al Zaatreh, A. & Kamalov, F. (2025). Empirical Evaluation of the Relative Range for Detecting Outliers. National Library of Medicine. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC12295245/ [Accessed 2 November 2025]
4. geekforgeeks (2025). Ordinary Least Squares (OLS) using statsmodels. geekforgeeks. Available at: https://www.geeksforgeeks.org/data-science/ordinary-least-squares-olsusing-statsmodels/ [Accessed 4 December 2025]
5. Giri, O.P. (2025). Choosing Sampling Techniques and Calculating Sample Size. Journal Universitas Pendidikan Indonesia. Available at: https://ejournal.upi.edu/index.php/IJoTis/article/download/75472/28968 [Accessed 28 October 2025]
6. Hajy Alikhani, P., Sadeghi Moghadam, M.R., Razavi, S.M. & Mohaghar, A. (2019). Revenue management and seller pricing decisions in retail industry: An agent-based model. Journal of Industrial Engineering and Management Studies (JIEMS). Available at: https://jiems.icms.ac.ir/article_92257_b5b970adfee266ec0ef810ac219c3c69.pdf [Accessed 17 November 2025]
7. Hasan, M., Khanam, A., Nasrullah, F., Mia, M.M., Jishan, S.S. & Rezvi, R.I. (2025). Investment & Market Growth Analysis: Insights into U.S. Retail Stock Performance,
8. Emerging Players, M&A Trends. Journal of Economics, Finance and Accounting Studies (JEFAS). Available at: https://al-kindipublishers.org/index.php/jefas/article/view/10066 [Accessed 15 November 2025]
9. Kowalskie, A. (2025). The Impact of Outliers on Linear Regression Models: Detection and Correction Strategies. ResearchGate. Available at: https://www.researchgate.net/publication/392609459_The_Impact_of_Outliers_on_Linear_Regression_Models_Detection_and_Correction_Strategies [Accessed 27 October 2025]
10. Kumar, N.K., Shah, R.K. & Sahani, S.K. (2025). Regression Analysis and Forecasting with Regression Model in Economics. ResearchGate. Available at: https://www.researchgate.net/publication/390786451_Regression_Analysis_and_Forec asting_with_Regression_Model_in_Economics [Accessed 2 December 2025]
11. Kwak, S. (2023). Are Only p-Values Less Than 0.05 Significant? A p-Value Greater Than 0.05 Is Also Significant! National Library of Medicine. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC10232224/ [Accessed 25 November 2025]
12. Lock, M. & Ansari, E.W. (2025). New world of big data—new challenges for evidence synthesis: impact of data duplication on estimates generated by meta-analyses. ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/S0895435624003974 [Accessed 2 November 2025]
13. Liu, Z. & Zhang, A. (2024). Sampling for Big Data Profiling: A Survey. ResearchGate. Available at: https://www.researchgate.net/publication/340669716_Sampling_for_Big_Data_Profiling_A_Survey [Accessed 27 October 2025]
14. Malpani, K. (2019). Detecting Outliers for Single Dimensional Data Using Interquartile Range. International Journal of Engineering Research and Applications. Available at: https://www.ijera.com/papers/vol9no9/Series-1/F0909013135.pdf [Accessed 25 November 2025]
15. Maulud, H.D. & Adbulazeez, M.A. (2020). A Review on Linear Regression Comprehensive in Machine Learning. ResearchGate. Available at: https://www.researchgate.net/publication/348111996_A_Review_on_Linear_Regression_Comprehensive_in_Machine_Learning [Accessed 4 December 2025]
16. Musadi, Kurniawati, I. (2023). Normal Distribution approximation through binominal and poison distribution. Journal of Mathematics and Mathematics Education. Available at: https://jurnal.uns.ac.id/jmme/article/download/75051/40975 [Accessed 23 November 2025]
17. Ruxton, D.G. & Neuhäuser, M. (2010). When should we use one-tailed hypothesis testing? Methods in Ecology and Evolution. Available at: https://besjournals.onlinelibrary.wiley.com/doi/epdf/10.1111/j.2041-210X.2010.00014.x [Accessed 28 November 2025]

18. Sarracino, F. & Mikucka, M. (2016). Estimation bias due to duplicated observations: a Monte Carlo simulation. Munich Personal RePEc Archive. Available at: https://mpra.ub.uni-muenchen.de/69064/1/MPRA_paper_69064.pdf [Accessed 2 November 2025]
19. Shaik, H., Rao, A. & Vardhan, B. (2025). Role of Exploratory Data Analysis in Data Science. ResearchGate. Available at: https://www.researchgate.net/publication/353979771_Role_of_Exploratory_Data_Analy sis_in_Data_Science [Accessed 3 December 2025]
20. Tutz, G. (2022). Probability and Non-Probability Samples: Improving Regression Modeling by Using Data from Different Sources. ResearchGate. Available at: https://www.researchgate.net/publication/359728648_Probability_and_Non-Probability_Samples_Improving_Regression_Modeling_by_Using_Data_from_Different _Sources [Accessed 18 November 2025]
21. West, M.R. (2022). Best practice in statistics: The use of log transformation. Sage Journals. Available at: https://journals.sagepub.com/doi/abs/10.1177/00045632211050531 [Accessed 25 November 2025]
22. Yagin, F.H., Pinar, A. & Fernandes, M.S. (2024). Statistical Effect Sizes in Sports Science. Journal of Exercise Sciences & Physical Activity Reviews (JESPAR). Available at: https://e-jespar.com/index.php/jespar/article/download/27/31/110 [Accessed 30 November 2025]
23. Yao, Y. & Wang, Y.H. (2021). A Review on Optimal Subsampling Methods for Massive Datasets. Journal of Data Science. Available at: https://jdsonline.org/journal/JDS/article/102/file/pdf [Accessed 27 October 2025]
24. Yeoman, I. (2024). Using revenue management to make business decisions. Springer Nature Link. Available at: https://link.springer.com/article/10.1057/s41272-024-00473-x [Accessed 17 November 2025]
25. Zhang, Y. (2025). Poisson Distribution and Its Applications. ResearchGate. Available at: https://www.researchgate.net/publication/393590961_Poisson_Distribution_and_Its_Ap plications [Accessed 23 November 2025]

Proofreader: Zokir ALIBEKOV

Layout and Designer: Oloviddin Sobir ugli

2026. № 6

To publish articles in this journal, you may submit articles, advertisements, stories, and other creative materials through the following links. Materials and advertisements are published on a paid basis.

You may subscribe to the journal at any time using the following details. Once subscribed, please send a screenshot or photo of your payment confirmation to our Telegram page @iqtisodiyot_77. Based on this, we will send the latest issue of the journal to your address each month.

“The journal “INNOVATION SCIENCE AND TECHNOLOGY” has been registered by the Agency for Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan from 09.10.2024 under the registration number №390637. License number: C-5669633. PNFL: 30407832680027

Our address: Tashkent city, Yunusobod district, 19th block, House 17.