***Submitted on 17/11/2024 word count: 2057***

**Submitted by,**

**Data Analytics Report on Retail Industry**

**Ecole de Management Léonard De Vinci** 

YEAR 2024-2025

**KIRAN BABU RAJ STUDENT ID 413854 INTERNATIONAL FINANCE AND INVESTMENT**

**SAITTYA PRYA BARUA STUDENT ID 413864 INTERNATIONAL FINANCE AND INVESTMENT**

1. **INTRODUCTION………………………………………………………………………………...1**
2. **DATASET OVERVIEW………………………………………………………………………….2**
3. **KINDS OF ANALYSIS DONE…………………………………………………………………2**
4. **STATISTICAL ANALYSIS………………………………………………………………………3**
5. **CORRELATION ANALYSIS……………………………………………………………………3**
6. **MULTIPLE LINEAR REGRESSION………………………………………………………….4**
7. **HYPOTHESIS TESTING……………………………………………………………………….5**
8. **SALES PATTERN ANALYSIS………………………………………………………………….6**
9. **SPENDING PATTERN ANALYSIS……………………………………………………………7**
10. **CONCLUSION……………………………………………………………………………………9**
11. **REFERENCE…………………………………………………………………...10**

**TABLE OF CONTENT**

**INTRODUCTION**

Retail industry is always changing and have a competitive nature and is always influenced by different factors like the customers wants, the availability of goods, the price of the products and also the seasonal changes. The data set provided here gives an insight about the retail business and consist of information about the customers, the sales and also the customers gender, age and their purchasing behavior. The data analysis report thereby give insight on what factors must be considered to increase the sales and also how the promotions and inventory management must be done in order to meet the customer preferences.

This particular dataset comprises of a thousand transaction records, which include major aspects such as:

Demographics: Information of the customers such as sex and age brackets, which helps to carry out an analysis of purchasing behaviors of the customers.

Product Insights: Types of products purchased, how many were purchased, and at what pricing giving an understanding of the demand of the product.

Sales Information: The amount that was earned from a particular sale and therefore how profitable the sale was and trends in sales over specified periods.

Time Dimensions: The day, month, and quarter when each transaction was made, which makes it possible to assess seasonal effects and trends on the sales.

In view of the above the objectives of this analysis are to find out what purchases a certain customer makes, what aids the sales, and how they can be of benefit to the company. The current analysis will emphasize on illustrating, complementing the graphical pictures, and testing the hypotheses in order to respond to the different possible questions.

1

**DATASET OVERVIEW**

1. Transaction ID: The unique identification number for each transaction
2. Date: Date of each transaction
3. Month: specifies the month of each transaction
4. Customer ID: The unique identification number given to each customer
5. Gender: It specifies the gender of each customer
6. Age and Age group: It is the age and their corresponding age group of the customer
7. Product category: Specifies the category of product purchased
8. Quantity: the number of units purchased by the customer.
9. Price per unit: It is the price of one unit of the product
10. Sales: The data shows the total sales of the products.

**KINDS OF ANALYIS DONE**

1. Statistical Analysis
2. Correlation Analysis
3. Multiple linear Regression
4. Hypothesis Testing
5. Sales Pattern Analysis (pivot table)
6. Spending Pattern Analysis (pivot table and ANOVA)

2

**Statistical Analysis**

The statistical analysis provides an overview on the variables like quantity of item purchases, price per unit and the total amount which is spend on each transaction.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Quantity*** |  | ***Price per Unit*** |  | ***Total Amount*** |  |
|  |  |  |  |  |  |
| **Mean** | 2.514 | **Mean** | 179.89 | **Mean** | 456 |
| **Standard Error** | 0.035820205 | **Standard Error** | 5.998251155 | **Standard Error** | 17.70868 |
| **Median** | 3 | **Median** | 50 | **Median** | 135 |
| **Mode** | 4 | **Mode** | 50 | **Mode** | 50 |
| **Standard Deviation** | 1.132734341 | **Standard Deviation** | 189.6813563 | **Standard Deviation** | 559.9976316 |

Quantity: Here, the average purchase is around 2.5 units with low variability, indicating a consistent purchasing behavior in terms of quantity. The low standard error (0.0358) shows that the sample mean is more equal to the population mean.

Price per unit: The average mean is 179.89 that shows the items sold are not so cheap. the statistical analysis suggests a mix of high cost and low-cost items. The median and mode indicates 50 as the central value

Total amount: The mean value here is 456 which indicates a combined effect of quantity purchased and the price per unit.

**CORRELATION ANALYSIS**

**Correlation Coefficient**: -0.0606: This correlation coefficient indicates a very weak negative relationship between Age and Sales. In practical terms, this means that as Age increases, Sales tend to decrease slightly. However, the effect is extremely small.

**Strength and Significance**: A correlation coefficient close to 0, like -0.0606, suggests that there’s almost no linear relationship between Age and Sales. The impact of Age on Sales is minimal, meaning Age doesn’t have a meaningful influence on sales figures in this dataset.

In conclusion age has significantly no effect on sales, and suggests some other factors affects the sales of different product.

3

**Multiple Linear Regression Analysis**

This summary output pertains to linear regression analysis. Let us analyze the results as follows:

**Multiple R (0.0606):** It refers to how accurately the predicted Sales values concur with the actual Sales figures. When this value is closer to 0 than to 1, it suggests that there is hardly any significant structured relationship between the independent variables and the dependent variable.

**R Square (0.0037):** This indicates that only about 0.37% of the variance in Sales is explained by both Gender and Age. This is a low required value which places Gender and Age as poor predictors of Sales in this dataset.

**Adjusted R Square (0.0017):** The Adjusted required modifies the required value based on the number predictors as a part of the model. In this case, the Adjusted required is close to zero that the model has very little explanatory power.

**statistic (1.8358) and Significance F (0.1600):** The statistic is run to check the goodness of fit of the model. If the p value is substantially high (0.1600), it means the model is not statistically significant using typical alpha levels (p < 0.05). This suggests that Gender and Age don’t separately or cumulatively predict Sales in this case.

**Intercept (558.13):** This is the predicted value of the dependent variable when both Gender and Age are equal to zero. It has been found to be statistically significant (p-value ≈ 0).

**Gender (0.94):** This value represents a co-efficient of 0.94 implying that gender has a very small positive effect on the dependent variable although it’s virtually not important.

**P-value (0.9788)** is above the desired line of 0.05 meaning this predictor is insignificant statistically and we cannot make a reasonable conclusion that it has any real effect.

**Age (-2.48):** A co-efficient value of -2.48 implies that the other independent variable, Age, has a weak moderating relationship with the dependent variable meaning that with an increase in Age, the dependent variable decreases slightly.

**P-value (0.0557):** This figure is very close to 0.05 however it is significant at 5% level. It is on the borderline so Age may have some effect but it is not conclusive with this data.

The regression model does not explain a substantial amount of variance in the dependent variable. In fact, out of the two independent variables, Gender and Age, only one had a significant value but was extremely minimal and the other did not have any significance at all.

4

**HYPOTHESIS TESTING**

Null hypothesis = 0, which means there is no significant difference between male and female.

Alternate null hypothesis ≠0, which means there is significant difference between male and female.

Here the test result shows mean spending of **male =455.43** and **female is 456.55** and **t statistic is -0.0316,** this is very close to zero and there is a very little difference in sample means.

**Degree of freedom(df)=998**

**P value(two-tail) =0.9748,** which is far greater than 0.05 This implies that there is no significant difference in the amount spent by males and females. Thus, we **fail to reject the Null Hypothesis.**

**SALES PATTERN**

Sales breakdowns depiction — and subsequent analysis of the data — helps to identify trends of various natures.

**Sales by Product Category**

The electronics compete for the highest total sales of about 156,905, making there the most preferred category. Clothing calculates the second award of 155,580 in sales were in demand as well. Beauty had the lowest number of sales such that it occupies the last position of all the categories with sales of 143,515. It is worth mentioning that the entire sales from all categories sums up to 4,56,000, this number is within the same range in all 3 Categories.

5

**Sales in different month for different product category**

Sales tends to be higher in the middle of the year and also during the end of the year which shows seasonal and holiday spending. The highly purchased item is electronics which is then followed by clothing and beauty. Therefore, understanding these patterns help to plan promotions, inventory management and marketing strategies in order to increase sales during the peak time.

**Monthly Sales Trend**

The graphic showing the abbreviated period of sales increase is very encouraging.

High Point: The reasonable height of sales in May at 53,150 Troy Ounces and 46,580 Troy Ounces in October stands out. It may indicate that this period records an important level of demand due to seasonality or other marketing efforts.

Low Point: These are the worst sales days at 28,990 in March and 23,620 in September. Such periods may be considered off-peak as the sales focus could be less than in others.

6

**Spending Pattern Analysis**

The spending pattern here analyze the intention of exploring the effect of certain customer statistics such as age and gender on consumer expenditure. This section includes sales data across specific segments of the population by sewers of the society with an intention of finding out the relevance of these patterns with possible statistical test.

**Spending pattern between different Age group**

The analysis here indicates that the spending patterns vary by age group and gender and therefore understanding these trends can help in target the specific category and can include promotional strategies to meet their needs. For instance, focusing on age group like 15-20 which possess the higher spending could a give more sale.

**Product preference and spending pattern between age group**

The analysis suggest that electronics have generally highest sales across most age groups. The age group between 26-30 and 56-60 prefer to spend more on clothing and the beauty product are preferred by age group between 21-25 and 51-55. Thus, this analysis indicates that product preference and there spending pattern varies between age group. Furthermore, by understanding this trends we can do inventory management and promotional strategies in such a way to cater the needs of such age group and increase the sales. Additionally, since here electronics being the highest selling category across all the group can be considered as the focus area of sales. The following graph shows the product preference and spending pattern between different age group

7

**ANOVA Test done to find the spending pattern difference**

ANOVA test is done to analyze if there is any statistically significant difference between age group and sales. Here, the p value is 1.9E-107 which is extremely low than the alpha value which is 0.05which show there is no significant difference between age group and sales. And furthermore, the F-statistic (548.8278) is much higher than the critical value F-Value (3.846) which also supports the rejection of Null Hypothesis. Therefor the ANOVA test shows that the average age and sales shows significant difference.

8

**CONCLUSION**

The retail industry analysis explores transactional data to find sales patterns, spending behaviors, and factors influencing purchases. Statistical analysis revealed that the average transaction involved 2.5 units, with significant variability in both quantity and pricing. Regression and correlation analyses showed that gender and age are weak predictors of sales, with negligible relationships to spending patterns. Hypothesis testing confirmed no significant differences between male and female spending, highlighting a balanced gender influence. Sales patterns indicated electronics and clothing as the most popular categories, with females slightly outspending males and seasonal peaks observed in May and October. Younger customers (15-20) exhibited the highest spending, driven by discretionary purchases, while spending declined in older age groups, reflecting financial commitments or retirement constraints. The findings suggest focusing on high-spending younger demographics, leveraging seasonal trends, and prioritizing product categories like electronics and clothing. Gender-based strategies appear less impactful, while age-targeted approaches could better address spending variations.

9

10

**REFERENCES**

1. Data Analysis with Microsoft Excel by Kenneth Berk and Partrick Carey
2. Automated Data Analysis using excel by Brian D Bissett
3. Excel easy website [*https://www.excel-easy.com/*](https://www.excel-easy.com/)
4. Trump Excel *https://trumpexcel.com/*
5. Excel is Fun (you tube channel)
6. Data source: https*://www.kaggle.com/datasets/mohammadtalib786/retail-sales-dataset*