

Cracking the Code of Product Success: Predicting Success with Data-Driven Insights

Mohammad Wasim Ashraf

Table of Contents

- Introduction
 - Data Source
 - Dataset Details
 - Dataset Features
 - Target Feature
- Goals and Objectives
- Data Cleaning and Preprocessing
 - Dropping Unnecessary Columns
 - Checking for Incorrect Values
 - Checking for Outliers
- Data Exploration and Visualisation
 - One Variable Plots
 - Histogram of Product Price
 - Histogram of Product Rating
 - Bar Chart of Category
 - Histogram of Discount
 - Two Variable Plots
 - Scatter Plot for Price vs Success Percentage
 - Scatter Plot for Rating vs Success Percentage
 - Scatter Plot for Cost of Marketing vs Yearly Sales
 - Box Plot for Yearly Sales by Category
 - Three Variable plots
 - Price vs Discount by Success Percentage
 - Number of Rating vs Success Percentage by Category
 - Marketing Spend vs Yearly Sales by Success Percentage
 - Category vs Price by Success Percentage
- Literature Review
- Summary and Conclusions
- References

Introduction

In the current scenario, where there are thousands of E-commerce platforms with lakhs of products listed in the same category. Categories can be "Electronics", "Fashion", "Beauty

and Health", "Sports and Outdoors" and many more. Enterprise wants to analyze the success of their products in their respective categories. Analyzing the success of their product helps the organisation to recognise the market. This helps the organisation to improve the quality of the product or introduce a new product or take back the product entirely. In today's time it is essential for an organisation to regularly analyse the success of their product because this helps the organisation to reduce the cost of manufacturing and transportation, if the product is not a hit in the market. The success percentage is calculated using various indicators like, cost spent on marketing, the price of the product, customer's rating on the product etcetera. In this report we will build a machine learning model to predict the success percentage of products in their specific category.

Data Source

The dataset that we will be using for this report is named as "Product Sales and Marketing Analytics Dataset". This dataset was procured from a platform called "Kaggle". It is an open platform for data science enthusiasts to publish their work publicly. This is also a platform which provides public datasets for free. Not only that, "Kaggle" provides short courses of programming to the learners from beginner level to advanced level. "Kaggle" also hosts competitions, where participants can show their potential to a much larger audience. Citations for the dataset used in this report can be found in the reference section.


Dataset Details

The dataset "Product Sales and Marketing Analytics Dataset" contains 500 observations and 15 columns. This dataset sheds light on how a product in their respective category performs in the market. This dataset contains 500 different products represented by rows, meaning, each row contains a unique product. To calculate the success percentage of products, this dataset contains various indicators represented in columns for example, seasonality trait index, market trend index, number of ratings for the product etcetera. Analyzing this dataset will help the enterprise to make better business decisions.

```
In [1]: import pandas as pd
df = pd.read_csv("Phase1_Group72.csv")
df.head()
```

Out[1]:

	Unnamed: 0	Product_Name	Category	Sub_category	Price	Rating	No_rating	Dis
0	0	Non-stick Pan	Home & Kitchen	Cookware	669.23	1.6	3682	
1	1	Tent	Sports & Outdoors	Outdoor Gear	67.13	3.2	2827	
2	2	Mascara	Beauty & Health	Makeup	463.25	3.5	4554	
3	3	Cutlery Set	Home & Kitchen	Cookware	1499.18	2.9	4976	
4	4	Blender	Home & Kitchen	Appliances	640.43	2.4	3806	




Above are the first five observation of the dataset.

The 10 random observations from the dataset are as follows:

```
In [5]: import pandas as pd
pd.set_option('display.max_columns', None)
df = pd.read_csv("Phase1_Group72.csv")
df.sample(10)
```

Out[5]:

	Unnamed: 0	Product_Name	Category	Sub_category	Price	Rating	No_rating	
201	201	Boots	Fashion	Footwear	478.59	2.6	1825	
431	431	Sneakers	Fashion	Footwear	1481.57	4.5	3675	
324	324	Jacket	Fashion	Clothing	670.61	3.2	4928	
151	151	Blush	Beauty & Health	Makeup	860.03	1.8	2422	
253	253	Serum	Beauty & Health	Skincare	1007.98	4.8	4353	
139	139	Bed	Home & Kitchen	Furniture	626.51	3.8	4032	
30	30	Sneakers	Fashion	Footwear	1006.63	2.1	206	
371	371	Lenovo ThinkPad	Electronics	Laptops	956.60	4.3	611	
460	460	Yoga Mat	Sports & Outdoors	Fitness Equipment	1329.82	4.1	4461	
464	464	Jacket	Fashion	Clothing	681.73	3.0	453	



Dataset Features

Feature Name	Data Type	Units	Brief Description
Category	Nominal Categorical	NA	The category in which the product belongs
Sub_category	Nominal Categorical	NA	Specific product sub-type
Price	Numeric	Currency	Price of the product
Rating	Numeric	Scale (1–5)	Customer rating
No_rating	Numeric	NA	Number of customer reviews
Discount	Numeric	Percentage (%)	Offered discount on the product
M_Spend	Numeric	Currency	Marketing expenditure
Supply_Chain_E	Numeric	Percentage (%)	Supply chain efficiency
Sales_y	Numeric	NA	Yearly sales
Sales_m	Numeric	NA	Monthly sales
Market_T	Numeric	NA	Market trend index indicates the expenditure for marketing of that product

Target Feature

The target feature for the model will be "Success_Percentage". This will be the feature that we will be predicting in phase-2. The data type of this feature is numerical.

Goals and Objectives

Goals:

- We aim to work on a predictive model that will be able to estimate the success percentage of products seeing their sales metrics, customer feedbacks, marketing efforts, and other trends.
- We need to identify the key factors like price, marketing spend, rating, market trends that are mostly impacting a product's success.
- We need to gain actionable insights for optimising marketing strategies and improving product performance for different categories.

Objectives:

- Analysing the relationship between sales, customer reviews, marketing spend, and success rates is an objective.
- Building the statistical models (like we used the regression models) for predicting product success with the given variables.
- Evaluating the effects of market and seasonality trends on product sales and success.

- Recommend plan of action for improving future marketing allocation and product development based on model results.

Data Cleaning and Preprocessing

Dropping Unnecessary Columns

The first column in the data is unnamed and consists serial numbers. For a supervised machine learning model this is unnecessary as it may lead to overfitting. Therefore, we need to delete this column. The column "Product_Name" also needs to be deleted because each product name is unique and it is acting like ID's and using this column in the machine learning model will lead to overfitting, therefore, dropping this column. first let's check the list of columns in the dataset.

```
In [16]: print(df.columns.tolist())
```

```
['Unnamed: 0', 'Product_Name', 'Category', 'Sub_category', 'Price', 'Rating', 'No_rating', 'Discount', 'M_Spend', 'Supply_Chain_E', 'Sales_y', 'Sales_m', 'Market_T', 'Seasonality_T', 'Success_Percentage']
```

Now dropping the the unnamed column and "Product_Name" column.

```
In [19]: df = df.drop(columns=["Unnamed: 0", "Product_Name"])
```

Prevention is better than cure: After dropping these two columns lets check the list of the columns again to confirm that the two variables have been deleted

```
In [22]: print(df.columns.tolist())
```

```
['Category', 'Sub_category', 'Price', 'Rating', 'No_rating', 'Discount', 'M_Spend', 'Supply_Chain_E', 'Sales_y', 'Sales_m', 'Market_T', 'Seasonality_T', 'Success_Percentage']
```

Checking for Null Values

```
In [25]: df.isnull().sum()
```

```
Out[25]: Category          0
Sub_category          0
Price                 0
Rating                0
No_rating             0
Discount              0
M_Spend               0
Supply_Chain_E        0
Sales_y               0
Sales_m               0
Market_T              0
Seasonality_T         0
Success_Percentage     0
dtype: int64
```

As the output shows all the features that we will be using for the supervised machine learning model have no null values. This concludes that in our data there are no null values. Our dataset is free from null values.

Checking for Incorrect Values

More than often, we come across some datasets, where for some observations there are incorrect or impossible values for example, negative age. The age cannot be negative. therefore, such observation needs to be omitted. In this step, we will check for incorrect values and if there are any such values, we must delete the entire row. we will start by checking the incorrect values for the following features: **Price, Discount, Rating, Success Percentage**. These features cannot have negative values. therefore, it is important to ensure that there is no incorrect values that could hamper the machine learning model.

```
In [30]: import pandas as pd
numeric_cols = df.select_dtypes(include='number').columns
for col in numeric_cols:
    invalid_count = (df[col] <= 0).sum()
    if invalid_count > 0:
        print(f"Invalid values (<= 0) in '{col}':", invalid_count)
```

Invalid values (<= 0) in 'Market_T': 258

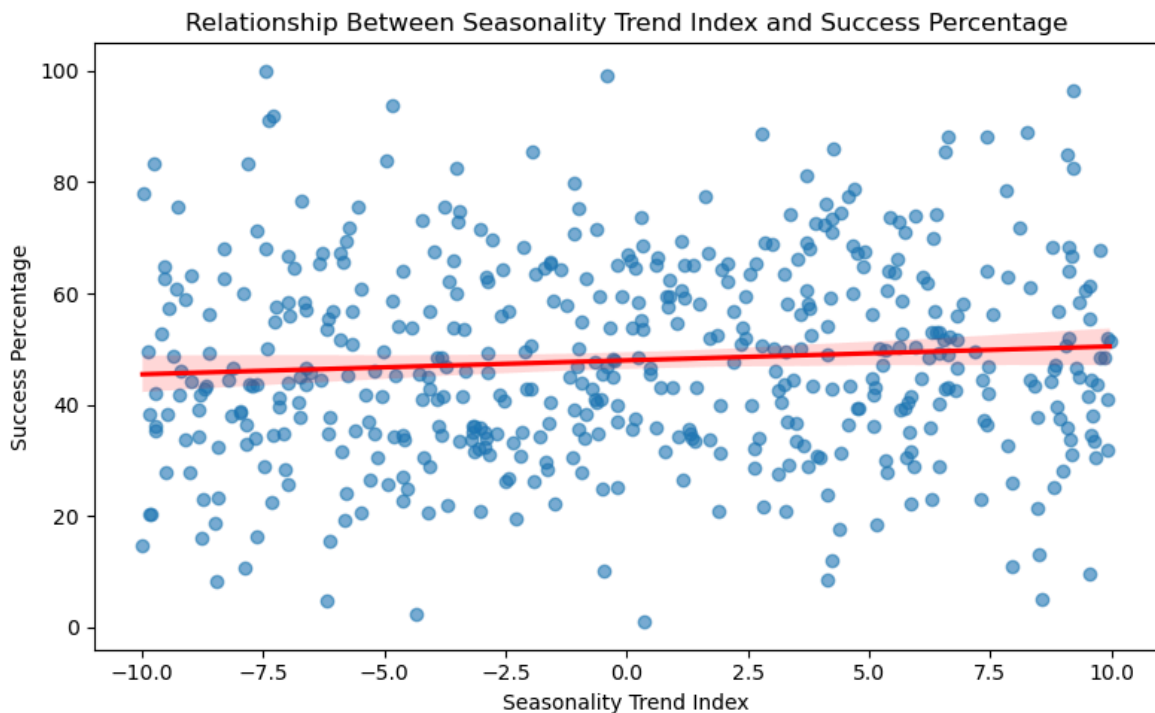
Invalid values (<= 0) in 'Seasonality_T': 251

In our dataset we have 258 negative values for "Marketing trend index" and 251 negative values for "Seasonality trend index" features. the feature "Market_T" captures the market trends. The market trend may be increasing or declining. For example, there is a observation in "Market_T" feature as "-0.25". This means, the market has declined by 25%. Therefore, negative values for "Market_T" feature is possible and should be accepted. The feature "Seasonality_T" captures the seasonality trend index. It is a score that indicates how seasonal fluctuations affect a product. this can be zero but not negative. The negative values for this feature is not acceptable. We are left with two choices first, Delete the observations with negative values of Seasonality_T feature and second, drop this feature from our supervised machine learning model. deleting 251 observations will leave us with only 50 percent of observations. lets make a scatter plot between "Seasonality_T" and our target feature "Success_Percentage".

```
In [33]: import seaborn as sns
import matplotlib.pyplot as plt
plt.figure(figsize=(8, 5))
sns.regplot(
    data=df,
    x='Seasonality_T',
    y='Success_Percentage',
    scatter_kws={'alpha': 0.6},
    line_kws={'color': 'red'}
)

plt.title('Relationship Between Seasonality Trend Index and Success Percentage')
plt.xlabel('Seasonality Trend Index')
plt.ylabel('Success Percentage')
```

```
plt.tight_layout()
plt.show()
```



The scatter plot between "Seasonality_T" and our target feature "Success_Percentage" shows a weak positive relationship. Since the relationship is weak, dropping "Seasonality_T" feature from our supervised machine learning model would be a better decision.

Dropping "Seasonality_T" feature

```
In [37]: df = df.drop(columns=["Seasonality_T"])
```

```
In [39]: print(df.columns.tolist())
```

```
['Category', 'Sub_category', 'Price', 'Rating', 'No_rating', 'Discount', 'M_Spend', 'Supply_Chain_E', 'Sales_y', 'Sales_m', 'Market_T', 'Success_Percentage']
```

Checking for Outliers

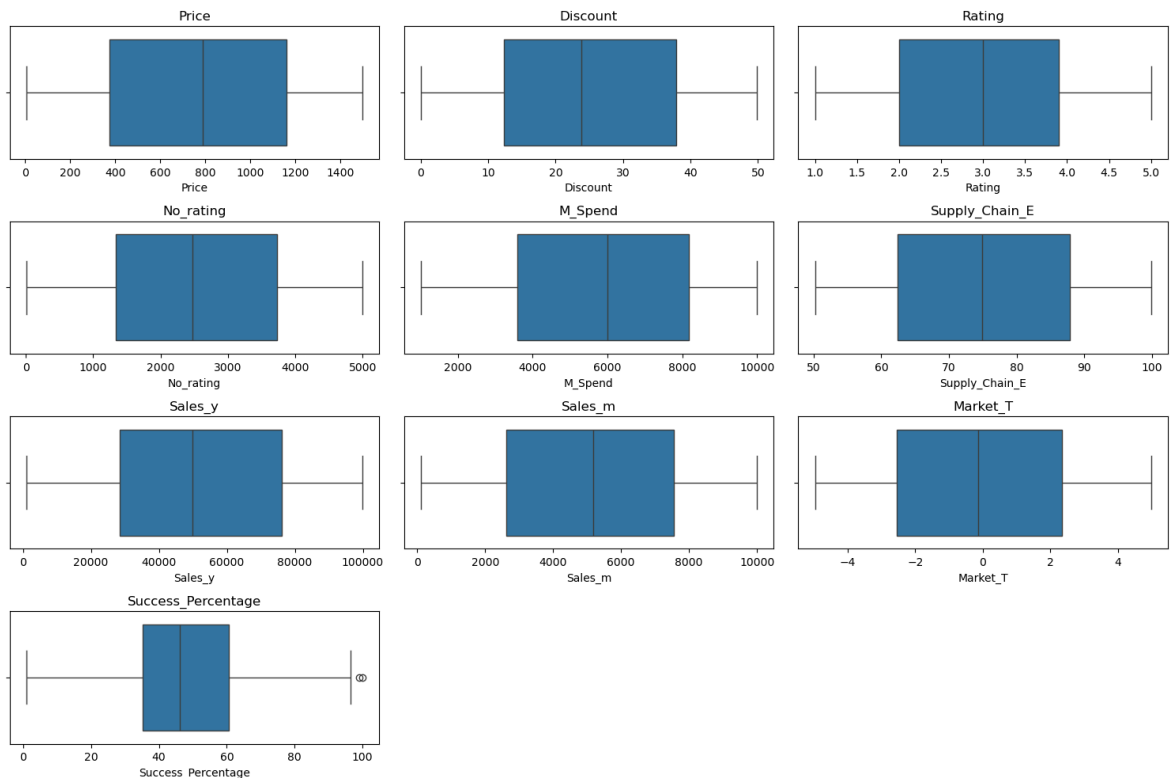
There are two ways we can check outliers for the data. The first way is to generate box plot for each feature to check for outliers and the other way is the Inter-quantile method. first let's check the outliers with the box plot method. In box plot method we will generate box plot for each variable and check for outliers.

```
In [43]: import matplotlib.pyplot as plt
import seaborn as sns

numeric_cols = ['Price', 'Discount', 'Rating', 'No_rating', 'M_Spend', 'Supply_C',
                'Sales_y', 'Sales_m', 'Market_T', 'Success_Percentage']

plt.figure(figsize=(15, 10))
for i, col in enumerate(numeric_cols):
    plt.subplot(4, 3, i+1)
    sns.boxplot(x=df[col])
```

```
plt.title(col)
plt.tight_layout()
plt.show()
```



In box plots the points outside the whiskers are potential outliers. Observing the box plot of all the features we can see there are no points outside the whiskers except for "Success Percentage" feature. There are two outliers in Success Percentage feature" at 100 percent. This is possible, because it is possible that some product have 100 percent success percentage. Therefore, this is acceptable.

Now, let's check outliers using IQR (Interquartile Range) method and validate our results with box plot method.

```
In [45]: def detect_outliers_iqr(data, column):
    Q1 = data[column].quantile(0.25)
    Q3 = data[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = data[(data[column] < lower_bound) | (data[column] > upper_bound)]
    return outliers

numeric_cols = ['Price', 'Discount', 'Rating', 'No_rating', 'M_Spend', 'Supply_C
                'Sales_y', 'Sales_m', 'Market_T', 'Success_Percentage']

for col in numeric_cols:
    outliers = detect_outliers_iqr(df, col)
    print(f"{col}: {len(outliers)} outliers")
```


Price: 0 outliers
 Discount: 0 outliers
 Rating: 0 outliers
 No_rating: 0 outliers
 M_Spend: 0 outliers
 Supply_Chain_E: 0 outliers
 Sales_y: 0 outliers
 Sales_m: 0 outliers
 Market_T: 0 outliers
 Success_Percentage: 2 outliers

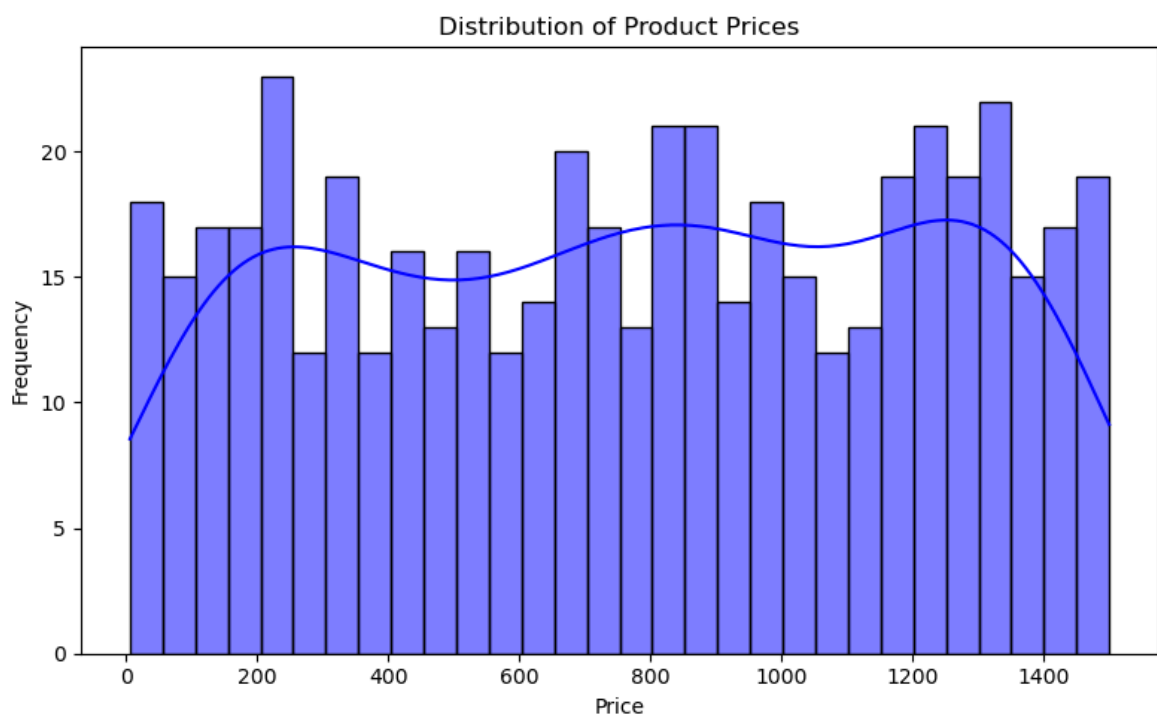
The output shows that all the features have no outliers except success percentage feature. there are two outliers in "Success Percentage" feature. It is the same result we got from box plot method, hence validating our earlier result about outliers.

Data Exploration and Visualisation

One Variable Plots

Histogram of Product Price

```
In [69]: import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(8, 5))
sns.histplot(data=df, x='Price', bins=30, kde=True, color='blue')
plt.title('Distribution of Product Prices')
plt.xlabel('Price')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```

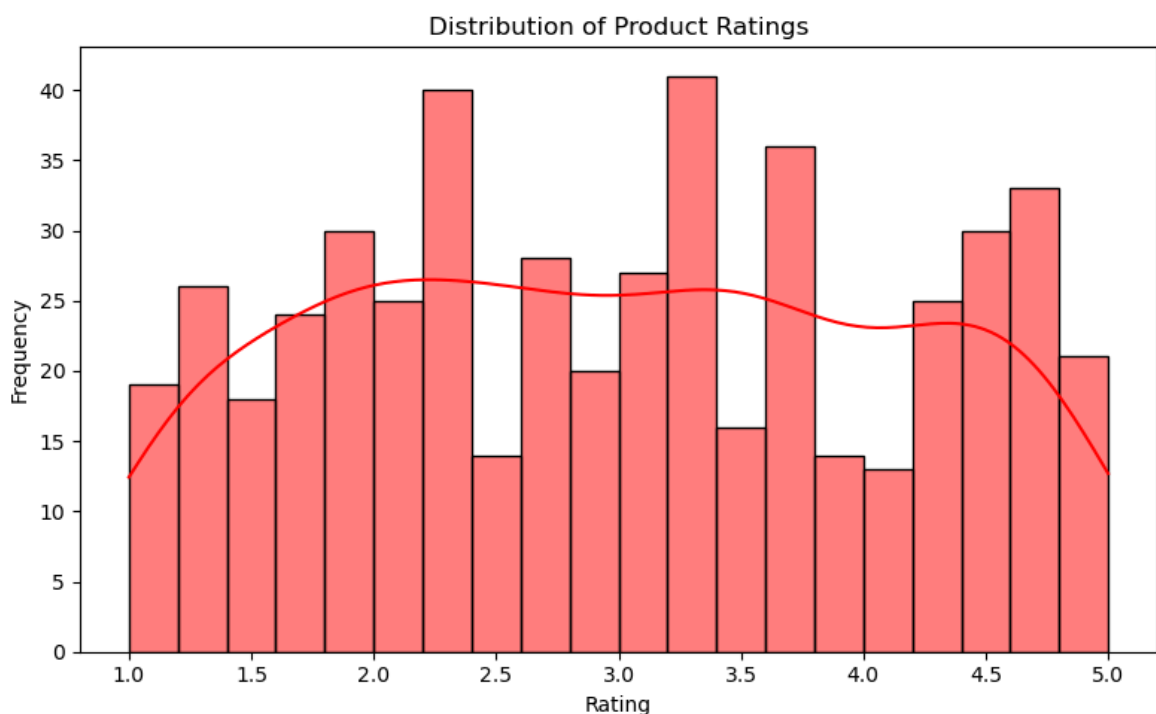


The histogram of price distribution shows the distribution of prices across the dataset. The KDE curve helps us to understand the distribution. Observing the KDE curve we now

know that the prices are evenly spread, there is no skewness towards low or high prices. However, there is more concentration of product prices between 200 and 1400, indicating presence of mid-range products and premium products.

Histogram of Product Rating

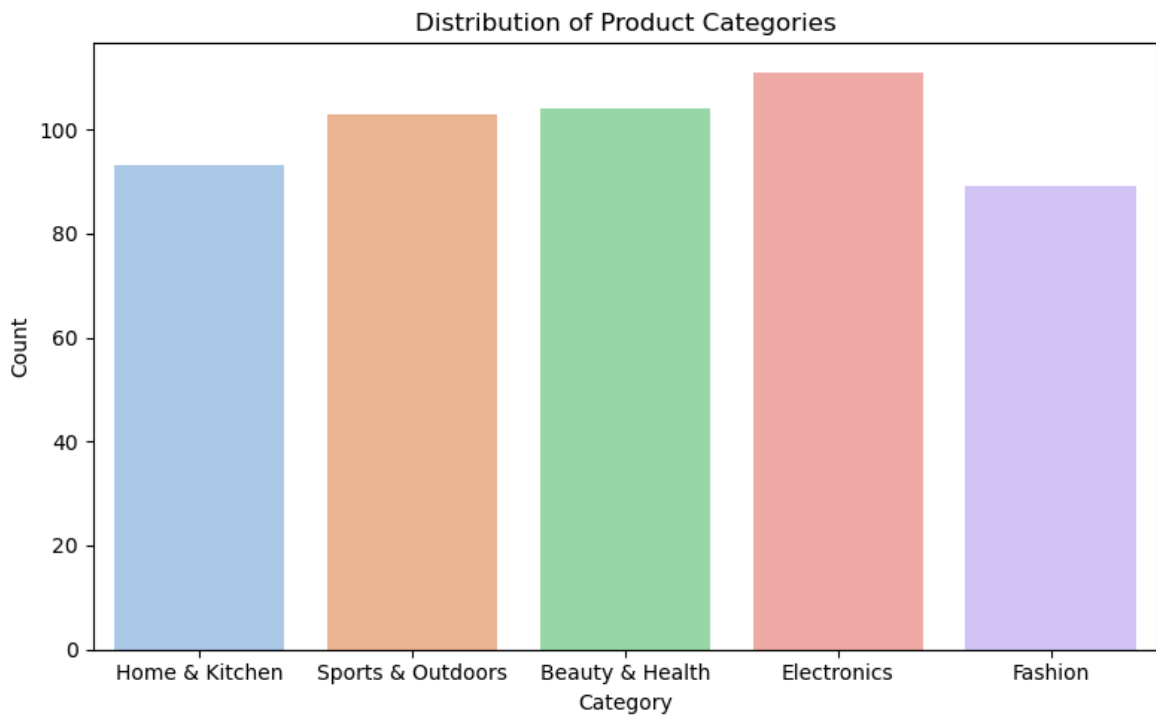
```
In [97]: import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(8, 5))
sns.histplot(data=df, x='Rating', bins=20, kde=True, color='red')
plt.title('Distribution of Product Ratings')
plt.xlabel('Rating')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```



The above histogram shows the distribution of customer's ratings of products after using them. The ratings vary from 1 to 5. The ratings are fairly spreaded across the dataset. here is a cluster at the higher ratings. The histogram shows the diverse satisfaction of customers.

Bar Chart of Category

```
In [88]: import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(8, 5))
sns.countplot(data=df, x='Category', hue='Category', palette='pastel', legend=False)
plt.title('Distribution of Product Categories')
plt.xlabel('Category')
plt.ylabel('Count')
plt.tight_layout()
plt.show()
```



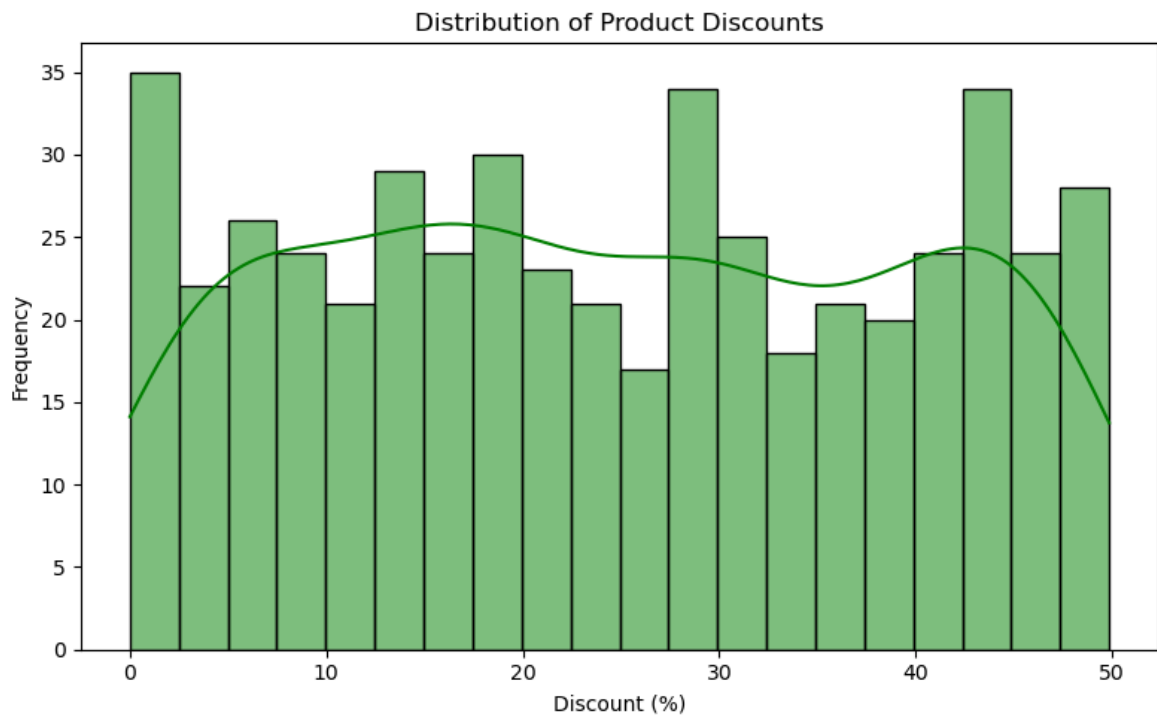
The bar chart for category shows dataset includes products from five major categories:

1. Home and Kitchen
2. Sports and Outdoors
3. Beauty and Health
4. Electronics
5. Fashion

The category Electronics have most products followed by Beauty and Health category. Fashion category has the least number of products.

Histogram of Discount

```
In [102... import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(8, 5))
sns.histplot(data=df, x='Discount', bins=20, kde=True, color='green')
plt.title('Distribution of Product Discounts')
plt.xlabel('Discount (%)')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```

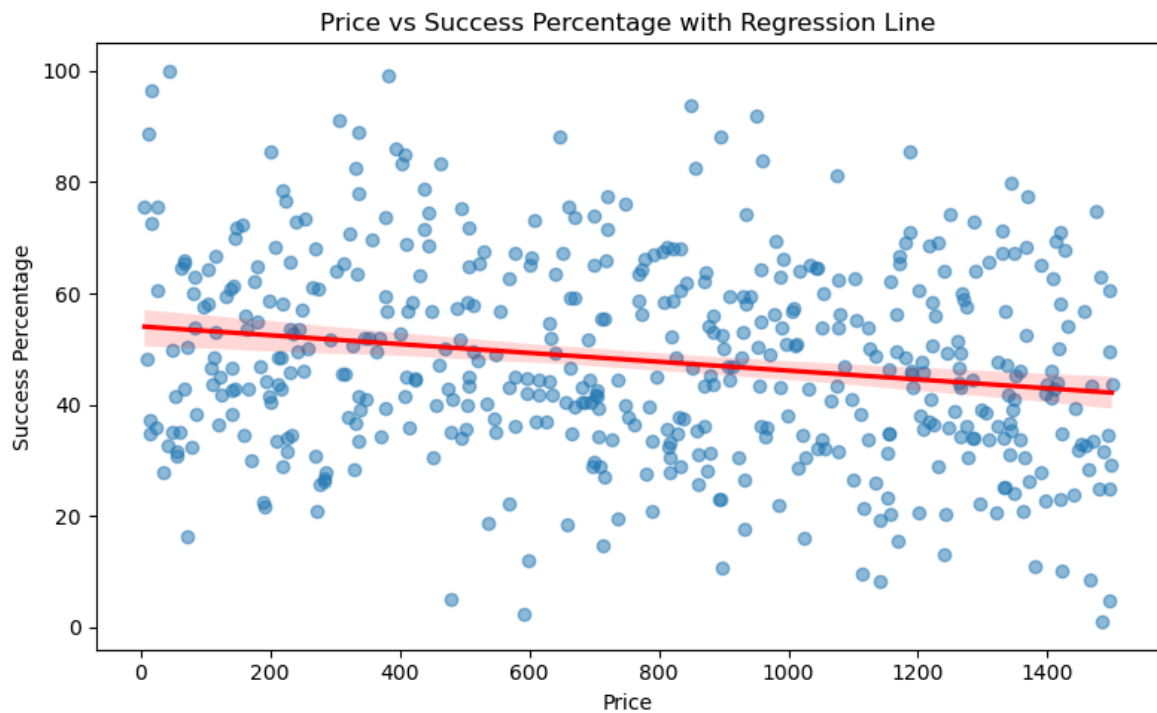


The histogram of discount shows that the discounts are evenly spread across the dataset. However, there is a cluster near 30 and 40 percent. this indicates the products listed in the dataset mostly offers 30 and 40 percent discount.

Two Variable Plots

Scatter Plot for Price vs Success Percentage

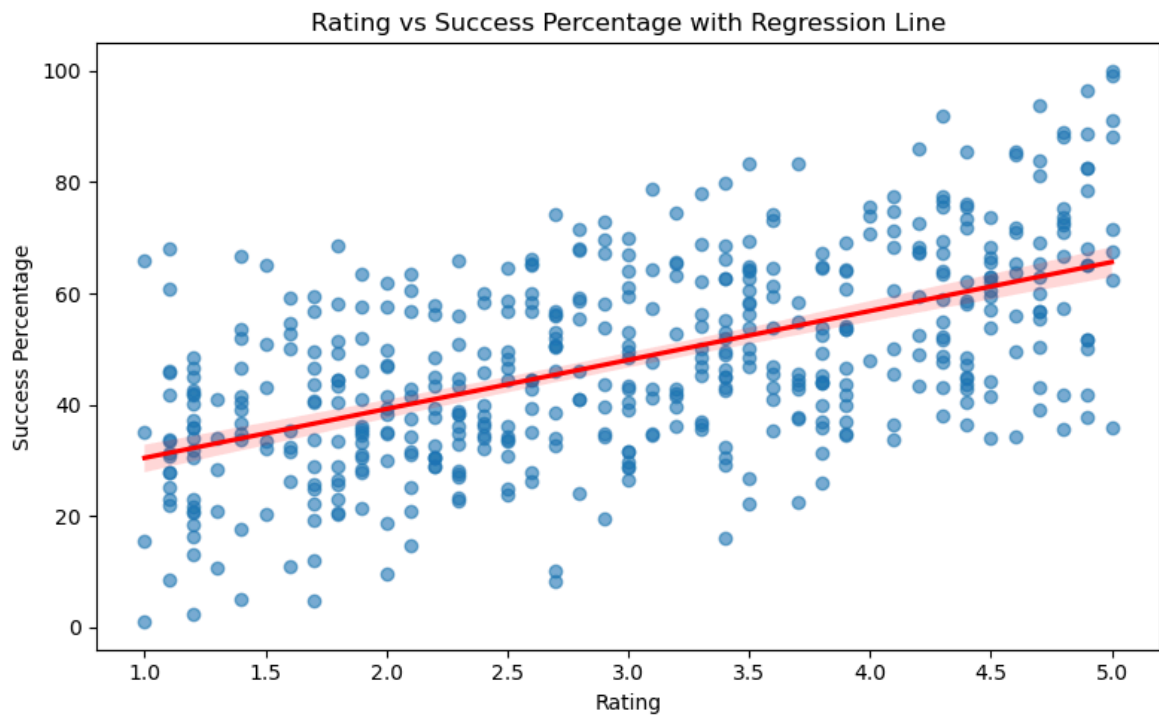
```
In [140... import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(8, 5))
sns.regplot(
    data=df,
    x='Price',
    y='Success_Percentage',
    scatter_kws={'alpha': 0.5},
    line_kws={'color': 'red'}
)
plt.title('Price vs Success Percentage with Regression Line')
plt.xlabel('Price')
plt.ylabel('Success Percentage')
plt.tight_layout()
plt.show()
```



The above scatter plot captures the relationship between price of products and success percentage of products. If we look at the red regression line, it is slightly declined indicating a weak negative relationship. As price of products increases their success percentage decreases. This gives us an important insight about customers. Customers tend to buy products that are pocket friendly, this directly affects the success percentage of that product.

Scatter Plot for Rating vs Success Percentage

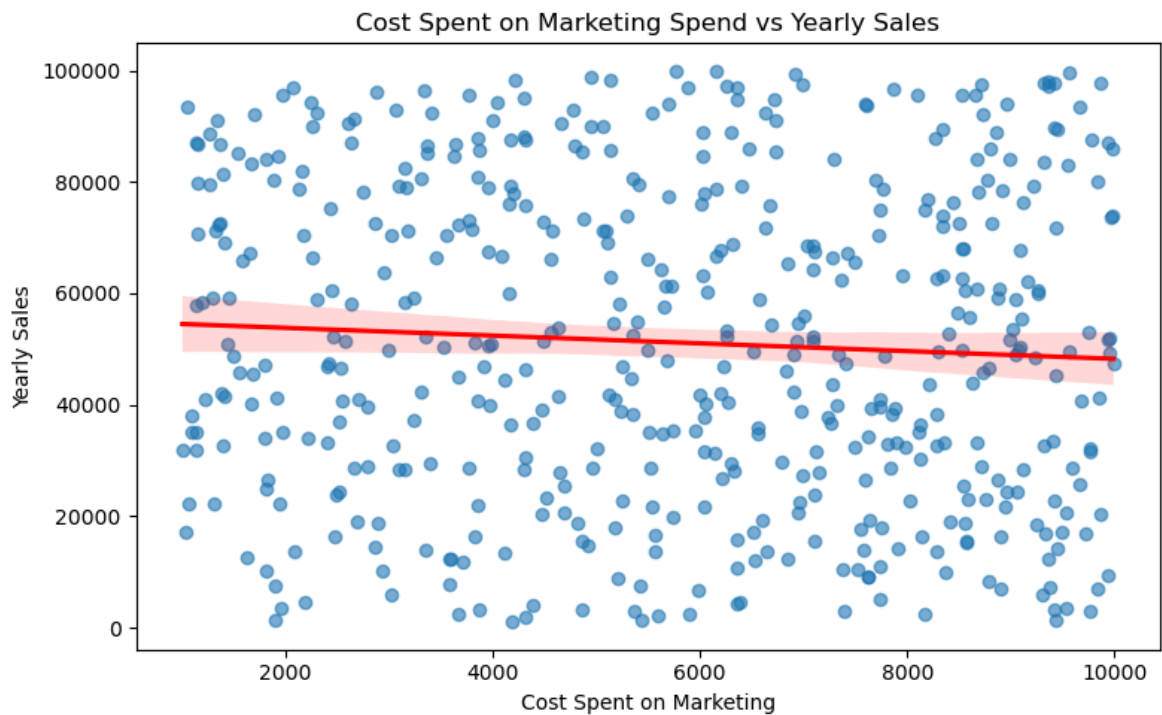
```
In [195... import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(8, 5))
sns.regplot(
    data=df,
    x='Rating',
    y='Success_Percentage',
    scatter_kws={'alpha': 0.6},
    line_kws={'color': 'red'}
)
plt.title('Rating vs Success Percentage with Regression Line')
plt.xlabel('Rating')
plt.ylabel('Success Percentage')
plt.tight_layout()
plt.show()
```



The above scatter plot illustrates the relationship between Ratings and success percentage. This shows how ratings affect the success of a product. From the above scatter plot, as the rating increases, the success percentage also increases. There is a strong positive relationship between ratings and success percentage. This concludes that customers tend to buy products that are highly rated. Before buying a product, checking the ratings of the product is a must for customers.

Scatter Plot for Cost of Marketing vs Yearly Sales

```
In [177... plt.figure(figsize=(8, 5))
sns.regplot(data=df, x='M_Spend', y='Sales_y', scatter_kws={'alpha': 0.6}, line_
plt.title('Cost Spent on Marketing Spend vs Yearly Sales')
plt.xlabel('Cost Spent on Marketing')
plt.ylabel('Yearly Sales')
plt.tight_layout()
plt.show()
```

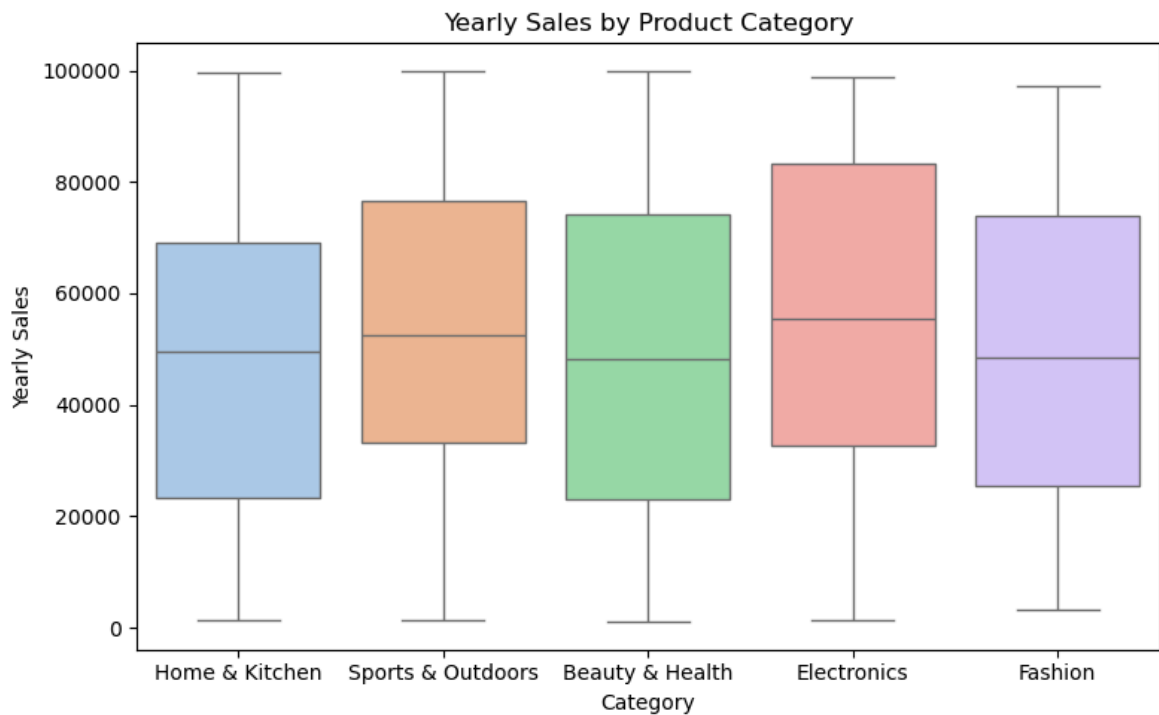


The above scatter plot captures the relationship between cost spent on marketing and yearly sales. Interestingly, there is a weak negative relationship between these two features. This suggests that higher marketing spend does not lead to increased yearly sales in the dataset. The possible reasons can be:

1. Product-market mismatch
2. Possibility that some product categories perform well even with minimal promotion.

Box Plot for Yearly Sales by Category

```
In [183... import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, x='Category', y='Sales_y', hue='Category', palette='pastel')
plt.title('Yearly Sales by Product Category')
plt.xlabel('Category')
plt.ylabel('Yearly Sales')
plt.tight_layout()
plt.show()
```



The box plot shows the yearly sales by product categories. Electronic category have the highest sales annually followed by sports and outdoor category. Home and kitchen category had the least sales.

Three Variable Plots

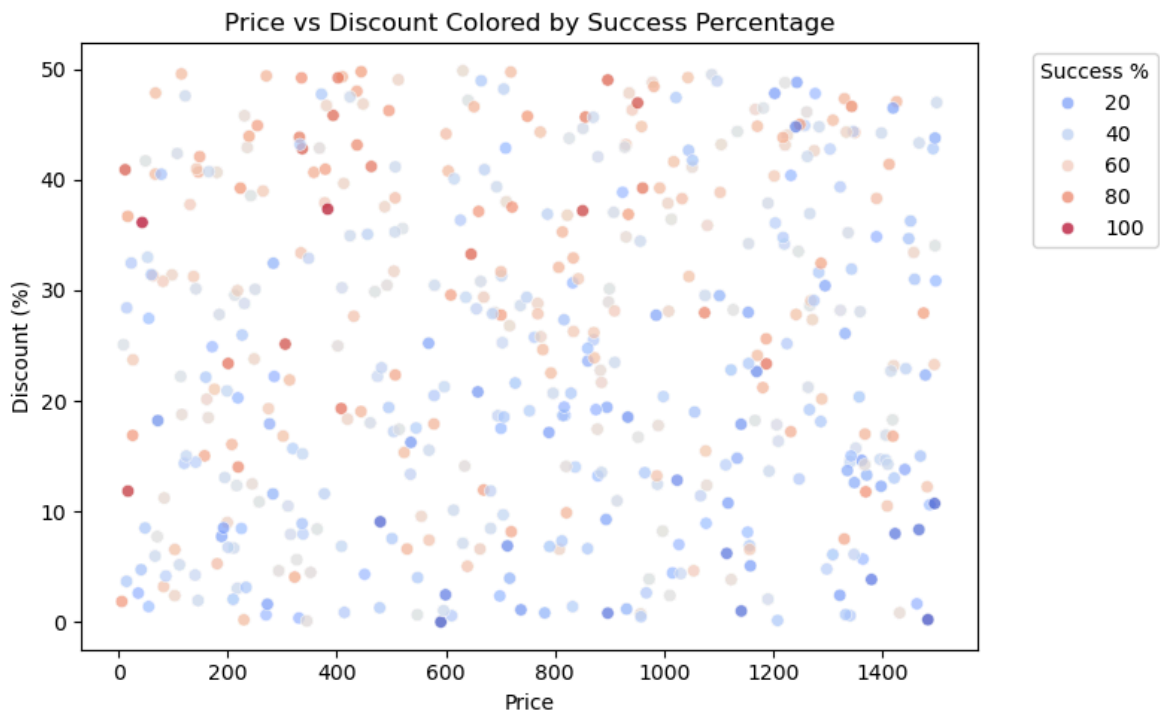
Price vs Discount by Success Percentage

```
In [87]: import matplotlib.pyplot as plt
import seaborn as sns

# Create the scatterplot
plt.figure(figsize=(8, 5))
sns.scatterplot(
    data=df,
    x='Price',
    y='Discount',
    hue='Success_Percentage', # Third variable as color
    palette='coolwarm',      # Color scheme
    alpha=0.7                # Slight transparency for better visibility
)

# Add Labels and title
plt.title('Price vs Discount Colored by Success Percentage')
plt.xlabel('Price')
plt.ylabel('Discount (%)')
plt.legend(title='Success %', bbox_to_anchor=(1.05, 1), loc='upper left')

# Display the plot
plt.tight_layout()
plt.show()
```

To make the best understanding of the plot Marketing Spend vs Yearly Sales coloured by success percentages. I want you to imagine a 4 block in the form of a virtual grid on top of this plot. This grid is dividing the plot into four portions (Top left, Top right, Bottom left and Bottom right). We have Price on the x axis, Discount percentage on the y axis and coloured dots represent the success percentages e.g. blue colours for low percentage and red for high percentage. Common sense says that increase in price means that success percentage will decrease and Increase in discount will increase the success percentage. Looking at the graph and referring to the common sense, we can see the red dots are increasing in the top part because discounts are increasing. Looking at the price and the success percentage we can see the red dots are decreasing with the increase in the price. With this perception, common sense says that the top left block should have the most red dots and they should be dark as well which is clearly visible therefore, increase in the discount and decrease in Price increase the success percentage of the products.

Number of Rating vs Success Percentage by Category

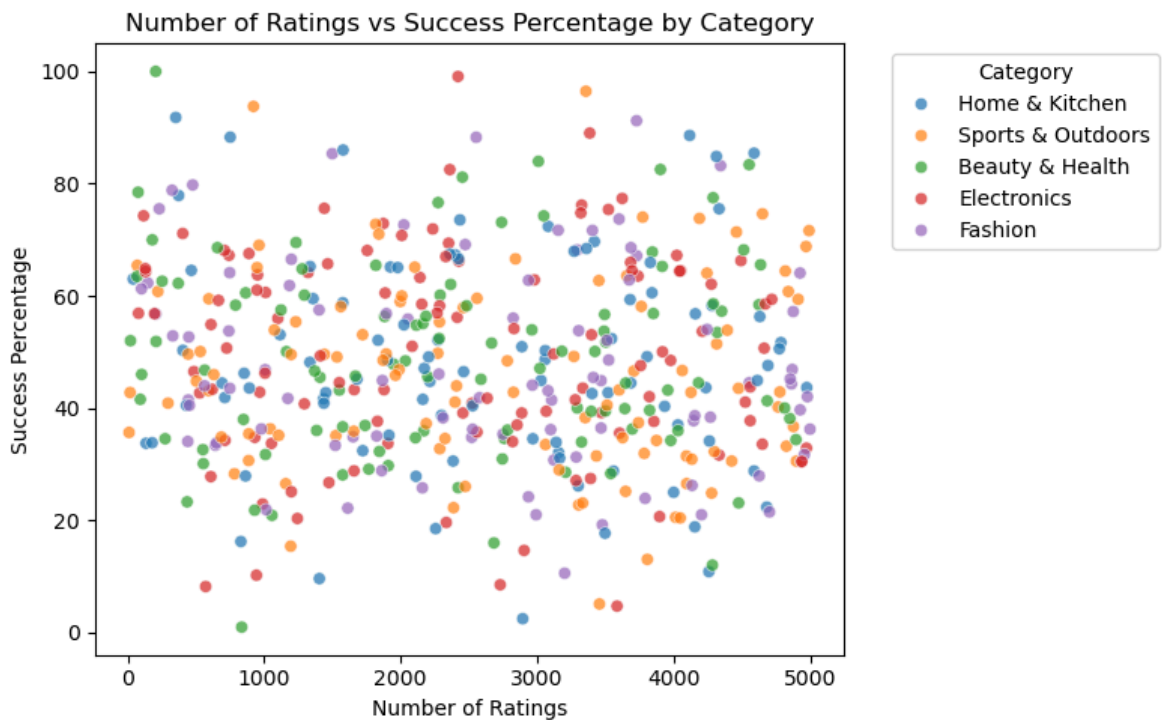
```
In [93]: import matplotlib.pyplot as plt
import seaborn as sns

# Scatterplot using three variables: No_rating, Success_Percentage, and Category
plt.figure(figsize=(8, 5))
sns.scatterplot(
    data=df,
    x='No_rating',
    y='Success_Percentage',
    hue='Category',
    alpha=0.7
)

# Add title and axis labels
plt.title('Number of Ratings vs Success Percentage by Category')
```

```
plt.xlabel('Number of Ratings')
plt.ylabel('Success Percentage')
plt.legend(title='Category', bbox_to_anchor=(1.05, 1), loc='upper left')

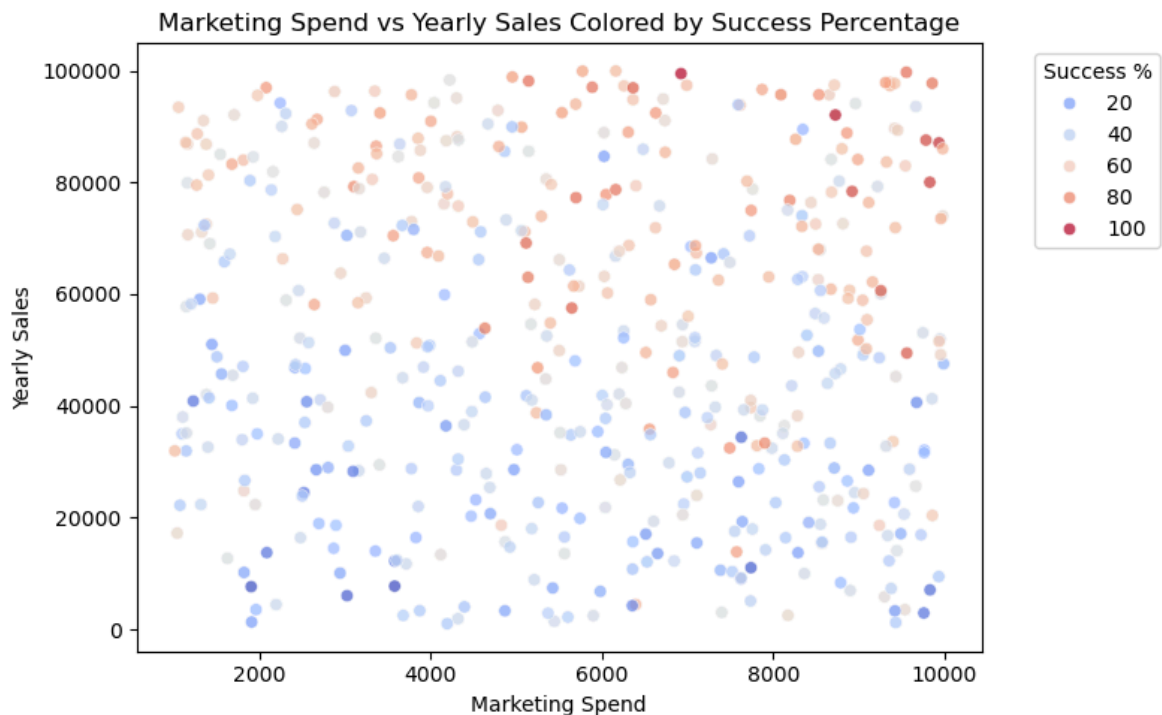
# Show the plot
plt.tight_layout()
plt.show()
```



This plot is a little difficult to interpret, but with a lot of focus and seeing at the success percentage at the Y axis, Number of Ratings at the x axis and Categories in the form of colours, WE can see here that as the number of ratings increase, Fashion, Beauty and health and Home and kitchen success percentage increases. We do not a solid evidence of this finding because the dots for all the categories are well spread. Especially for sports and electronics there are no clear correlations. They are well spread across the grid.

Marketing Spend vs Yearly Sales by Success Percentage

```
In [204... import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(8, 5))
sns.scatterplot(
    data=df,
    x='M_Spend',
    y='Sales_y',
    hue='Success_Percentage',
    palette='coolwarm',
    alpha=0.7
)
plt.title('Marketing Spend vs Yearly Sales Colored by Success Percentage')
plt.xlabel('Marketing Spend')
plt.ylabel('Yearly Sales')
plt.legend(title='Success %', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
```

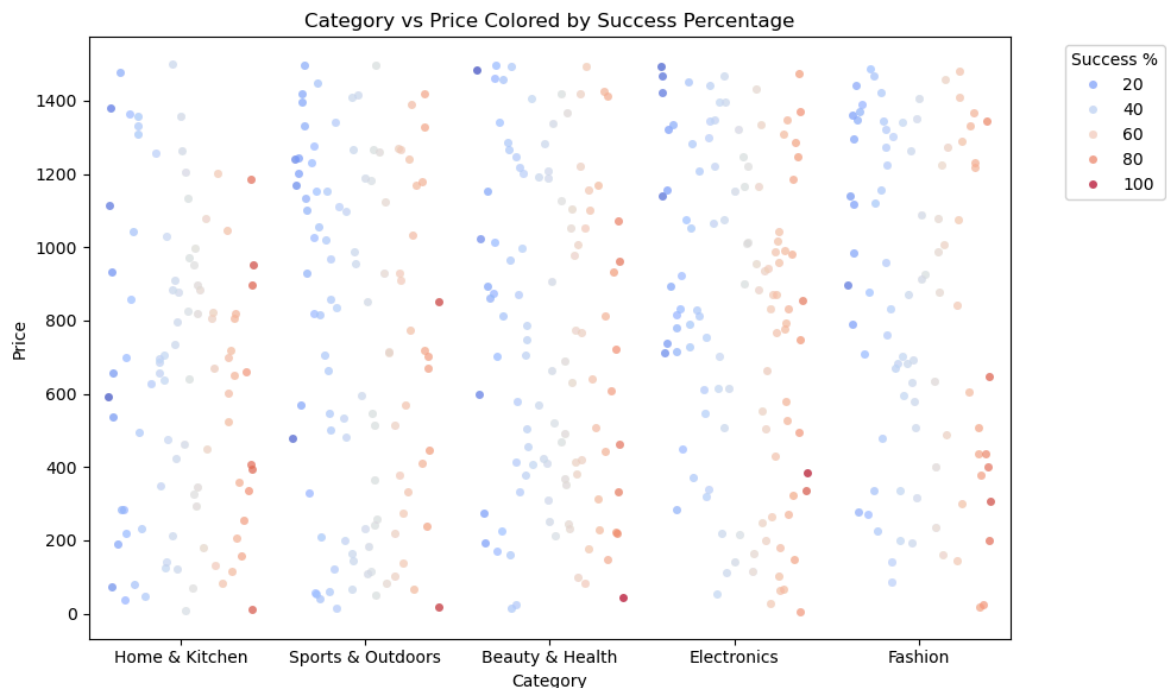


To make the best understanding of the plot Marketing Spend vs Yearly Sales coloured by success percentages. I want you to imagine a 4 block virtual grid on top of this plot. This grid is dividing the plot into four portions (Top left, Top right, Bottom left and Bottom right). We have Marketing spend on the x axis, Yearly sales on the y axis and coloured dots represent the success percentages e.g. blue colours for low percentage and red for high percentage. Common sense says that increase in yearly sales in means that success percentage will increase and Increase in marketing spend will also increase the success percentage. Looking at the graph and referring to the common sense, we can see the red dots are increasing in the top part because yearly sales are increasing. Looking at the marketing spend and the success percentage we can see the red dots are increasing with the increase in the marketing spend. With this perception, common sense says that the top right block should have the most red dots and they should be dark as well which is clearly visible therefore, increase in the yearly sale and marketing spend increase the success percentage of the products.

Category vs Price by Success Percentage

```
In [200... import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(10, 6))
sns.stripplot(
    data=df,
    x='Category',
    y='Price',
    hue='Success_Percentage',
    palette='coolwarm',
    dodge=True,
    alpha=0.7,
    jitter=True
)
plt.title('Category vs Price Colored by Success Percentage')
plt.xlabel('Category')
```

```
plt.ylabel('Price')
plt.legend(title='Success %', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
```



For the Category vs Price coloured by success percentage plot (3 variable plot). We aim to see that how the Price impacts the success percentage of a product based on the category. The colour scheme here is rather a little deceptive (100% should have been blue or green and 20% should have been red) we could not correct it due to time shortage but we will explain what is happening in detail. If we see the legend, blue colour represents 20% and it turns a little darker towards red colour as we move towards the 100%. We can see in this plot that, for home and kitchen, as the Price of the product in increase the success percentage decreases. For home and kitchen we have the highest difference. We can see a small but similar trend in Sports and outdoor categories; the increase in price sees reduction in success percentage. For other categories including, Beauty and health, Electronics and Fashion we have evidences of similar trends but they are not as prominent as home and kitchen and Sports categories.

Literature Review

In today's world, figuring out if a product will succeed in the market or not has become a major area of interest for researchers. For this review, we deep dived into 10 journals articles and 4 conference papers that covered everything including the marketing analytics and customer behaviour and sales forecasting and machine learning. Several researches were done on how a product's perceived worth impacts its success. For instance, Gaur et al. (2020) highlighted how a product's value, its brand reputation, and how much customers are willing to pay all shape its journey. Kumar et al. (2020) added to this by showing how digital innovations and quicker access to customer insights are reshaping how marketers approach their audiences. E-commerce and online selling strategies are another strong point of discussion. Bernal et al. (2021) discussed how

personalised ads and smarter online engagement tactics are changing how shoppers behave. Meanwhile, Chaudhary and Dhar (2021) took a closer look at discounting strategies, concluding that slashing prices isn't always a win because it can sometimes chip away at customer trust. This felt especially relevant to our project, since discount rates are one of the factors we are studying. Customer feedback, especially through online reviews, kept showing up as an impactful predictor. Liu et al. (2017) combined review analysis with sales predictions and found a clear pattern; happier customers usually mean better sales. Shah et al. (2022) showed that tapping into large retail data alongside customer feedback can sharpen predictions even more. Verma et al. (2021) compared different models like Random Forest and XGBoost, finding that using several models often yields stronger results than relying on just one. Jain and Mehta (2023) stressed how important it is to pick the right features and fine tune models if we really want trustworthy forecasts. Other studies, like those by Choudhury et al. (1999) and Rajala and Westerlund (2019), explored how the timing of a launch and the reliability of suppliers can tip the scales for or against a product. Keikhosrokiani et al. (2022) crafted a model that uses subtle patterns in customer behaviour to predict future sales and recommend products. Sharma et al. (2023) demonstrated that tweaking prices dynamically through machine learning doesn't just increase profits but it makes customer buying patterns more predictable too. One paper that really stood out was "Buy When?" (2023), which suggested a fresh way of predicting the best moments for customers to make a purchase by using survival analysis. especially useful when thinking about seasonal spikes. Another, the CausalMMM study (2024), introduced a cway of properly isolating the true effect of different marketing actions, rather than just spotting surface trends.

Summary and Conclusions

In Phase 1 of the project, our focus was upon the understanding the structure, features, and potential applications of the "Product Sales and Marketing Analytics Dataset". This dataset contains 500 entries and 15 variables covering product details, pricing, marketing efforts, customer feedback, sales metrics, and external market trends. The basic goal in this phase was to clearly define the objectives of the modelling task. We identified that predicting the "Success_Percentage" based on the other features would be the primary focus, also we will see relationships between marketing spend, sales performance, customer ratings, and external market factors. The dataset has a lot of information interms of the variables relating to the success of a product in the market, so we can do both predictive modeling and exploratory analysis like understanding market and customer behaviour trends. Variables such as Price, Discount, M_Spend, Rating, Market_T, and Seasonality_T are having plausible impact on the Success_Percentage. The presence of marketing and seasonal indices, external factors also play a reasonable part in defining product success. The data was pre cleaned, but we still did some minor tweaks to set our tone so that we can focus on building models rather than spending significant time on data preprocessing. Our primary focus was to set statistical models to predict Success_Percentage, and explore around to see the key drivers behind product success, and derive conclusions to suggest marketing strategies and product positioning. This

basic understanding prepares us for the next phase, where data exploration, feature selection, and model building will begin.

References

1. Shrivastav, U. (n.d.). Product Sales and Marketing Analytics Dataset. Retrieved April 1, 2025, from <https://www.kaggle.com/datasets/utkarshshrivastav07/product-sales-and-marketing-analytics-dataset>
2. Jing, G., Peng, Q., Zhang, L., Tan, R. and Zhang, J. (2020). Estimation of product success potential using product value. International Journal of Production Research. <https://www.tandfonline.com/doi/full/10.1080/00207543.2020.1788733#abstract>
3. Shah, D., and Murthi, B.P.S. (2021). Marketing in a data-driven digital world: Implications for the role and scope of marketing. Journal of Business Research. <https://www.sciencedirect.com/science/article/abs/pii/S0148296320304355>
4. Rosario, A., and Raimundo., R. (2021). Consumer Marketing Strategy and E-Commerce in the Last Decade: A Literature Review. Journal of Theoretical and Applied Electronic Commerce Research. <https://www.mdpi.com/0718-1876/16/7/164>
5. Tan, W.K. (2023). When do price discounts become attractive? A study comparing discount strategies on consumer perceptions. Asia Pacific Journal of Marketing and Logistics. <https://www.emerald.com/insight/content/doi/10.1108/apjml-06-2021-0456/full/html>
6. Fan, Z.P., Che, Y.J., Chen, Z.Y. (2017). Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. Journal of Business Research. <https://www.sciencedirect.com/science/article/abs/pii/S0148296317300231>
7. Begum, N. (2024). Big Data Analytics and Its Impact on Customer Behavior Prediction in Retail Businesses. Pacific Journal of Business Innovation and Strategy. <https://scienceget.org/index.php/pjbis/article/view/17>
8. Smirnov, P.S., and Sudakov, V.A. (1925). Forecasting new product demand using machine learning. Journal of Physics: Conference Series. <https://iopscience.iop.org/article/10.1088/1742-6596/1925/1/012033/meta>
9. Bataineh, A.Q., Abu-ALSondos, I.A., Frangieh, R.H., Alnajjar, I., and Salameh, A.A. (2024). Predictive Modeling in Marketing Analytics: A Comparative Study of Algorithms and Applications in E-Commerce Sector. Research Gate. https://www.researchgate.net/publication/376851043_Predictive_Modeling_in_Marketing_Commerce_Sector
10. McGinnis, M.A., and Vallopra, R.M. (2006). Purchasing and Supplier Involvement: Issues and Insights Regarding New Product Success. Jornal of Supply Chain Management. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-493X.1999.tb00057.x>
11. Hultink, E.J., and Robben, H.S. (1995). Measuring New Product Success: The Difference that Time Perspective Makes. Journal of Product Innovation Management. <https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-5885.1250392>

12. Sarkar, M., Ayon, E.H., Mia, M.T., Ray, R.k., Chowdhury, M.S., Ghosh, B.P., Al-Imran, M., Islam, M.T., Tayaba, M., and Puja, A.R. (2008, December). Optimizing E-Commerce Profits: A Comprehensive Machine Learning Framework for Dynamic Pricing and Predicting Online Purchases. Journal of Computer Science and Technology Studies. Retrieved from <https://www.neliti.com/publications/589846/optimizing-e-commerce-profits-a-comprehensive-machine-learning-framework-for-dyn>
13. Vallarino, D. (2023, August). Buy when? Survival machine learning model comparison for purchase timing. Cornell University. Retrieved from <https://arxiv.org/abs/2308.14343>
14. Gong, C., Zhang, L., Yao,D., Chen, D., Li, W., Su, Y., and Bi, J. (2024, June). CausalMMM: Learning Causal Structure for Marketing Mix Modeling. Cornell University. Retrieved from <https://arxiv.org/abs/2406.16728>
15. Zhao, X., and Keikhosrokiani, P. (2021, April). Sales Prediction and Product Recommendation Model Through User Behavior Analytics. Computers, Materials and Continua. Retrieved from <https://www.techscience.com/cmc/v70n2/44651>.
16. S Karsli, K Otto, W Li, A Bilton, K Hölttä-Otto. BARRIERS TO PRODUCT REPAIR: EXPLORING MOTIVATIONS AND CAPABILITIES AMONG OPERATORS. University of Melbourne. Retrieved from <https://findanexpert.unimelb.edu.au/scholarlywork/1951860-barriers-to-product-repair--exploring-motivations-and-capabilities-among-operators?cache=1733640278302>

In []: