**Nike Sales Analysis 2024**

1. **Problem Definition**

The primary goal of this data analysis is to uncover insights into the factors influencing sales for Nike company across different countries, product categories and external factors such as Holidays. This involves analyzing trends, identifying drivers of sales performance, recommending strategies for improvement and discover useful sales insights to increase revenue.

**Objectives:**

1. **Identify Primary Factors Influencing Sales:**

* Analyze the relationship between sales metrics such as revenue, units sold, and retail price.
* Evaluate the impact of product attributes (e.g., main category, sub-category, price tier) on sales.
* Understand the role of sales channels (online vs offline) in revenue generation.

1. **Understand Customer Demographics and Behavior Across Locations:**

* Assess sales patterns across countries to identify regional preferences.
* Analyze customer purchasing behavior (e.g., online sales percentage) and product popularity by region.

1. **Evaluate the Impact of External Factors:**

* Quantify the influence of holidays on sales (e.g., HolidayCount and Holiday\_Sales\_Estimate).

1. **Define Key Performance Indicators (KPIs):**

* **Revenue Growth**: Month-over-month changes in total revenue.
* **Sales by Categories**: Performance of main categories and sub-categories in terms of revenue and units sold.
* **Online vs Offline Sales Ratio**: Contribution of each channel to total revenue.
* **Holiday Sales Contribution**: Percentage of total revenue driven by external factor holiday count.

1. **Data Collection**

**Internal Data**

The dataset used for this analysis is sourced from Kaggle public datasets. Specifically, the dataset pertains to Nike’s sales performance across various countries in 2024.

**Description of the Dataset:**

The dataset includes detailed information about Nike’s sales figures, providing insights into product performance, geographical sales distribution, and key metrics that influence business outcomes. Below is a summary of the dataset features:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| **Month** | Indicates the month of sales data. |
| **Country** | Specifies the country where sales were recorded. |
| **Main\_Category** | The primary category of the product (e.g., Apparel, Footwear, Equipment). |
| **Sub\_Category** | The sub-category within the main category (e.g., Tops, Socks, Accessories). |
| **Product\_Line** | Represents the specific product line (e.g., Tech Fleece, Vapor Cricket). |
| **Price\_Tier** | The pricing tier of the product (e.g., Budget, Mid-Range, Premium). |
| **Units\_Sold** | The total number of units sold. |
| **Revenue\_USD** | The revenue generated in USD. |
| **Online\_Sales\_Percentage** | Percentage of total sales made through online channels. |
| **Retail\_Price** | Retail price of the product. |

**External Data**

Holiday data was obtained using the HolidayAPI, which provides comprehensive details about public and regional holidays across different countries.

**Integrating holiday data adds significant value by:**

* Enabling the analysis of the impact of holidays on sales performance.
* Allowing for the creation of features like HolidayCount and Holiday\_Sales\_Estimate, which help quantify holiday-related sales boosts.
* Facilitating deeper insights into seasonal trends and customer purchasing behavior around holidays.

**Data Integration**

To incorporate holiday data into the analysis, we retrieved public holiday information across six countries (United States, India, United Kingdom, Japan, China, and South Korea) for the year 2024. This data was sourced from HolidayAPI and was utilized to understand the impact of holidays on sales.

**Steps Performed**:

1. Used the API key provided by HolidayAPI to fetch holiday data for each country and month for our internal data
2. Queried public holidays for the year 2024 across six countries using python
3. Retrieved details such as the holiday date, name, and whether it was a public holiday.
4. Converted the retrieved data into a structured format using a Python script.
5. Transformed the data to count the number of public holidays per country and month.
6. Added a HolidayCount feature to quantify the holiday influence in each country.
7. Saved the raw holiday data and processed counts into Excel files for further analysis and visualization.

This combination of internal and external data provides a robust framework for analyzing and forecasting sales trends, identifying opportunities, and understanding external factors influencing sales performance across the globe.

1. **Data Preprocessing**

* The dataset was thoroughly checked for missing values. Since there were no missing values detected, no imputation or removal of data points was necessary.
* The data was inspected for duplicate entries to ensure that no records were repeated. No duplicates were found, indicating that the dataset was properly collected and maintained.
* Any inconsistencies in data formats were addressed. For example, all date formats were standardized to ensure uniformity across the dataset. Additionally, categorical variables were normalized to ensure consistent naming conventions (e.g., country names, product categories).

**External Features**

* **Offline Sales Amount**

To capture the full picture of Nike's sales performance, offline sales amount was created as an external feature. This feature was based on historical sales data and online sales percentage available in the dataset, which helped calculate offline sales performance not included in the dataset.

* **Holiday Sales Estimates**

Given that holidays often have a significant impact on retail sales, a holiday sales estimate feature was engineered. This involved identifying assumed hike in sales due to holidays (10% revenue boost per holiday). The feature was designed to help account for seasonal variations in sales.

**4. Exploratory Data Analysis (EDA)**

**Data Preparation**

The dataset was read and pre-processed in R, and relevant features were categorized as factors or numeric variables to aid in the analysis. The data includes variables like Month, Country, Main\_Category, Units\_Sold, Revenue\_USD, and external factors like HolidayCount.

**Correlation Analysis**

Correlation coefficients were calculated to understand the relationships between Holiday\_Sales\_Estimate and potential explanatory variables like Units\_Sold, Online\_Sales\_Percentage, Retail\_Price, Offline.Sales.Amount and Revenue\_USD.

**Key Insights**:

* **Revenue and Offline Sales Amount** are the most influential factors for holiday sales, with moderate positive correlations. These should be prioritized in analysis and modeling.
* **Units Sold and Retail Price** also have some impact but to a lesser degree.
* **Online Sales Percentage** appears to have a negligible negative relationship, indicating that offline sales might dominate during holidays.

**Visualization of Relationships**

Scatter plots with regression lines were used to visualize the relationships between the variables.

**1. Relationships Between Variables:**

* **Offline Sales Amount:** A strong positive linear trend is evident, indicating that higher offline sales are directly correlated with higher revenue.
* **Units Sold:** A positive linear relationship is visible, showing that an increase in units sold contributes to higher revenue.
* **Retail Price:** A weaker positive trend is observed, suggesting that retail price influences revenue, but the impact is not as strong as units sold or offline sales.
* **Offline Sales Amount:** Offline sales correlate positively with both units sold and retail price, implying their combined influence on the revenue.
* **Online Sales Percentage:** The scatterplots involving the online sales percentage appear to lack strong linear relationships with other variables, reflecting its weaker correlation with holiday sales or revenue as observed earlier.
* **Units Sold vs. Retail Price:** A slight negative trend indicates that higher-priced items generally have lower units sold, aligning with typical market behavior.

**2. Density Insights (Diagonal Kernel Density Plots):**

* **Revenue and Offline Sales:** Revenue has a wider distribution with some high peaks, likely contributed by premium product sales or strong-performing markets.Offline sales show a similar density pattern to revenue, reinforcing their close relationship.
* **Units Sold:** Most sales are concentrated in a specific range, suggesting popular product categories.
* **Retail Price:** The distribution of retail price highlights clusters around budget and mid-range pricing tiers, with fewer premium items.

**Additional Notes:**

* The scatterplot matrix visually reinforces that offline sales and units sold are the most influential variables for predicting revenue. This finding should guide the focus of predictive modeling efforts.
* There are weak or negligible correlations involving online sales percentage, which supports the earlier observation that offline sales dominate in holiday performance.

**5. Advanced Analysis**

**Initial Model (LinearModel\_1)**

* A linear regression model (linearModel\_1) was built to predict holiday\_sales using the following independent variables: main\_ctgr, country, product\_line, online\_sales\_percentage, units\_sold, price\_tier, retail\_price, offline\_sales, and holiday.

**Key Findings:**

* The regression summary revealed which variables were significant predictors of holiday sales. Variables such as units\_sold, offline\_sales, and holiday showed a positive and significant relationship with holiday\_sales.
* online\_sales\_percentage showed a weaker or insignificant relationship, which aligns with the correlation results (correlation = -0.072).
* Categories like price\_tier and main\_ctgr likely impacted the model due to their categorical nature, driving segmentation and pricing insights.

**Optimization and Refinement**

* Building on linearModel\_1, further models were developed, leading to the final optimized regression model: Variables like Month, Sub\_Category, and Product\_Line were removed to reduce multicollinearity and improve prediction accuracy.
* Performed log transformation on Holiday Sales Estimate and Offline Sales Amount to reduce the skewness of the data which ultimately improves the prediction accuracy of the model.
* The iterative process reduced RMSE significantly from 0.43899 to 0.41812 highlighting the importance of feature selection and model refinement.

**6. Summary**

**Insights from Regression Analysis**

**Units Sold:** The strongest driver of holiday sales, directly proportional to overall revenue.

**Offline Sales:** A critical revenue channel, with higher offline sales correlating positively with holiday sales.

**Holiday Count:** Seasonality plays a significant role, as holidays amplify sales.

**Retail Price:** Pricing strategy significantly affects sales performance during holidays.

**Online Sales Percentage:** Minimal or negative impact on holiday sales, suggesting that offline sales are more influential during the holiday period.

**Actionable Recommendations:**

* Prioritize inventory and marketing for offline channels during holiday periods.
* Leverage high-performing product categories (main\_ctgr) and focus on regions (country) with stronger holiday sales.
* Reassess online sales strategies, possibly targeting non-holiday periods to complement offline sales performance significantly.

**BONUS TASK:**

For bonus task, I have performed Exponential Smoothing in Excel. It is in the Exponential Smoothing sheet.

* **High Errors:** Periods with high errors (e.g., November and December) suggest that exponential smoothing struggled to predict these months accurately, possibly due to outliers or significant seasonal effects.
* **Seasonality:** If holiday sales or other seasonal factors are influencing the data, simple exponential smoothing may not fully capture these patterns. You may need Holt-Winters triple exponential smoothing for seasonality.
* **Trend Impact:** Exponential smoothing is smoothing out revenue fluctuations, but it doesn’t explicitly account for any upward or downward trends in the data.