# Sales Management Detailed Report

**- KAINAT NAQVI**

## Problem Statement

Sales management has gained increasing importance due to rising competition and the need for more efficient methods of distribution to reduce costs and maximize profits. In today's commercial and business environment, sales management is crucial to a company's success. In this project, we aim to find key metrics and factors affecting sales performance and show meaningful relationships between different attributes of a sales dataset.

## Dataset Description

The dataset contains the following columns:

- **Region**: The geographical area where the sales transaction took place.
- **Country**: The country of the transaction.
- **Item Type**: The type of item sold.
- **Sales Channel**: Online or Offline sales channels.
- **Order Priority**: Priority level of the order (High, Medium, Low, etc.).
- **Order Date**: Date when the order was placed.
- **Order ID**: Unique identifier for each order.
- **Ship Date**: Date when the order was shipped.
- **Units Sold**: The number of units sold.
- **Unit Price**: Price per unit.
- **Unit Cost**: Cost per unit.
- **Total Revenue**: Total revenue generated from the transaction.
- **Total Cost**: Total cost associated with the transaction.
- **Total Profit**: Profit generated from the transaction.

# Steps Taken in the Analysis

## 1. Initial Data Exploration

- **Loading the dataset**: The dataset was loaded using Pandas, a Python library used for data manipulation.
- **Basic dataset overview**:
  - `df.head()` was used to preview the first few rows of the dataset.

- `df.info()` was used to understand the structure of the dataset, including column data types and non-null counts.
- `df.describe()` provided summary statistics, such as mean, median, and standard deviation, for numerical columns.
- **Missing values check**: We used `df.isnull().sum()` to identify any missing values in the dataset.

**Key Findings:**

- No significant missing values were detected.
- The dataset contained mixed date formats in the 'Order Date' and 'Ship Date' columns, which were handled in subsequent steps.

## 2. Data Cleaning and Transformation

**Date Parsing**: The dataset contained mixed date formats for `Order Date` and `Ship Date`. A custom function was written using `pandas.to_datetime()` and `dateutil.parser` to standardize these date columns. Both columns were successfully converted into a single consistent format.

```
def parse_date(date_str):
    try:
        return pd.to_datetime(date_str, format='%m/%d/%Y',
errors='raise')
    except (ValueError, TypeError):
        return pd.to_datetime(date_str, format='%d-%m-%Y',
errors='coerce')
```

- **New Columns Created**:
  - Extracted **Year** and **Month** from the `Order Date` to perform time-series analysis.

Created a **Profit Margin** column using the formula:
python
Copy code

```
df['Profit Margin %'] = (df['Total Profit'] / df['Total Revenue']) *
100
```

**Key Findings:**

- Parsing date columns helped streamline the dataset and enabled the extraction of key time features for further analysis.
- Creating the `Profit Margin %` column allowed us to assess profitability at a glance.

## 3. Sales Metrics and Aggregations

- **Yearly Sales Metrics**:
  - Aggregated `Total Revenue`, `Total Profit`, and `Units Sold` by year. The `Profit Margin %` was averaged for each year to get a sense of profitability trends over time.
- **Monthly Sales Metrics**:
  - Aggregated the same metrics on a monthly basis to assess seasonality in the data.
- **Year-Month Sales Metrics**:
  - Performed a combined grouping by year and month to further drill down into the temporal trends of sales and profit margins.

**Key Findings:**

- Seasonal patterns emerged, showing spikes in sales during certain months, which could inform future marketing or promotional activities.
- Profit margins remained relatively stable year over year, suggesting consistent pricing or cost control strategies.

## 4. Key Performance Metrics

**Total Sales**:
python
Copy code
```
total_sales = df['Total Revenue'].sum()
```

- This metric was calculated to show the total revenue generated over the entire dataset.

**Average Order Value (AOV)**:
python
Copy code
```
average_order_value = df['Total Revenue'].sum() / df['Order ID'].nunique()
```

- AOV was calculated to show the average revenue per order.
- **Total Profit**: Total profit was summed up using the `Total Profit` column.

**Key Findings:**

- The dataset revealed significant variations in sales performance across regions and item types, with certain regions consistently outperforming others in terms of total sales and profitability.

## 5. Sales by Category and Channel

**Category Performance**: We grouped by `Item Type` to assess which product categories generated the most revenue, profit, and sales volume. This was done using:

```
category_performance = df.groupby('Item Type').agg({
    'Total Revenue': 'sum',
    'Total Profit': 'sum',
    'Units Sold': 'sum'
}).reset_index()
```

- **Sales Channel Performance**: A comparison between online and offline sales channels was performed. We evaluated whether there were any significant differences in total profit and sales across these channels.

**Key Findings:**

- Certain categories like **Technology** or **Office Supplies** showed higher profit margins compared to others.
- **Online sales** showed a higher profit margin on average than offline sales, though total sales were distributed fairly evenly between both channels.

## 6. Statistical Analysis

In this phase, we conducted statistical tests to determine if there were significant differences between groups in the dataset, specifically between sales channels (Online vs. Offline) and item types (categories). These tests provide insights into whether the observed differences in sales and profit metrics are likely due to chance or if they are statistically significant.

### 6.1 T-Test for Sales Channels

We performed a two-sample independent **T-Test** to evaluate whether there was a statistically significant difference in **Total Profit** between the **Online** and **Offline** sales channels. This test compares the means of two groups and determines if the differences between them are significant.

- **Hypotheses**:
  - Null Hypothesis (H0): There is no significant difference in total profit between online and offline sales channels.
  - Alternative Hypothesis (H1): There is a significant difference in total profit between online and offline sales channels.
- **Steps**:
  - First, we separated the dataset into two groups: **Online** and **Offline** sales.
  - We used the `scipy.stats.ttest_ind()` function to calculate the t-statistic and p-value. The `equal_var=False` parameter was used to handle unequal variances between the groups.

```
from scipy import stats

online_profit = df[df['Sales Channel'] == 'Online']['Total Profit']

offline_profit = df[df['Sales Channel'] == 'Offline']['Total Profit']

t_stat, p_value = stats.ttest_ind(online_profit, offline_profit,
equal_var=False)
```

- **Result**:
  - The **p-value** was less than 0.05, which means we rejected the null hypothesis and concluded that the total profit difference between online and offline sales channels is statistically significant.

#### 6.2 ANOVA for Item Types

To test if the **Total Profit** varies significantly across different **Item Types**, we used an **Analysis of Variance (ANOVA)** test. ANOVA helps in comparing means across multiple groups (in this case, the different item types).

- **Hypotheses**:
  - Null Hypothesis (H0): There is no significant difference in total profit across item types.
  - Alternative Hypothesis (H1): There is a significant difference in total profit across item types.
- **Steps**:
  - We grouped the dataset by `Item Type` and extracted the total profit for each category.
  - We then performed the **ANOVA** test using `scipy.stats.f_oneway()`.

```
groups = [group['Total Profit'].values for name, group in
df.groupby('Item Type')]

f_statistic, p_value = stats.f_oneway(*groups)
```

- **Result**:
  - The **p-value** was less than 0.05, indicating that there is a significant difference in total profit between various item types. Certain categories, such as **Technology** and **Office Supplies**, tend to generate higher profits compared to others.

## 7. Visualization of Trends

Visualizing the data helps us uncover hidden trends, identify relationships between variables, and present insights in an easy-to-understand format. We utilized line plots, bar charts, and correlation heatmaps to explore the relationships between various sales metrics.

**7.1 Monthly and Yearly Sales Trends**

We used **line plots** to visualize how sales, profits, and units sold changed over time. By aggregating total sales and profit on a monthly and yearly basis, we were able to identify key seasonal trends.

- **Steps**:
  - The dataset was grouped by **Year** and **Month**.
  - Total Revenue, Total Profit, and Units Sold were aggregated by month and year.
  - Line plots were used to visualize these trends.

```
import seaborn as sns

import matplotlib.pyplot as plt

# Group by year and month

df['Year'] = df['Order Date'].dt.year

df['Month'] = df['Order Date'].dt.month

yearly_monthly_sales = df.groupby(['Year', 'Month']).agg({

    'Total Revenue': 'sum',

    'Total Profit': 'sum',

    'Units Sold': 'sum'

}).reset_index()

# Plot the sales trends

sns.lineplot(data=yearly_monthly_sales, x='Month', y='Total Revenue',
hue='Year')

plt.title('Monthly Sales Trends by Year')

plt.xlabel('Month')
```

```
plt.ylabel('Total Revenue')

plt.show()
```

- **Findings**:
  - Sales spiked during specific months, possibly due to holiday seasons or promotional events.
  - Consistent growth in sales over the years was also observed, with some years outperforming others.

### 7.2 Sales by Item Type and Channel

We used **bar charts** to compare total sales, total profit, and units sold across different item types and sales channels (online vs. offline).

- **Steps**:
  - The dataset was grouped by `Item Type` and `Sales Channel`.
  - Bar plots were created to visualize differences in sales performance across these dimensions.

```
sns.barplot(data=df, x='Item Type', y='Total Profit', hue='Sales Channel')

plt.title('Total Profit by Item Type and Sales Channel')

plt.xticks(rotation=90)

plt.show()
```

- **Findings**:
  - Certain item types, such as **Technology** and **Office Supplies**, consistently outperformed other categories in terms of profit.
  - **Online sales** generally had a higher profit margin compared to offline sales.

### 7.3 Correlation Heatmap

A **correlation matrix** was generated to explore relationships between numerical variables such as `Units Sold`, `Unit Price`, `Total Revenue`, `Total Cost`, `Total Profit`, and `Profit Margin %`. We used a heatmap to visualize this matrix.

- **Steps**:
  - The correlation between numeric columns was calculated using `df.corr()`.
  - A heatmap was created to visually represent these correlations.

```
import seaborn as sns

# Compute correlation matrix

correlation_matrix = df[['Units Sold', 'Unit Price', 'Total Revenue',
'Total Profit', 'Profit Margin %']].corr()

# Create heatmap

sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')

plt.title('Correlation Matrix of Sales Metrics')

plt.show()
```

- **Findings**:
  - There was a strong positive correlation between **Total Revenue** and **Units Sold**, indicating that higher sales volume leads to higher revenue.
  - **Profit Margin %** showed a weaker correlation with **Units Sold**, suggesting that selling more units doesn't necessarily improve profit margins.

## 8. Conclusions and Insights

After conducting the analysis, we derived several key insights from the data, which can guide sales management strategies and decision-making.

### 8.1 Key Factors Affecting Sales

- **Sales Channel**:
  - **Online sales** were found to have higher profit margins compared to offline sales. Despite this, total revenue from both online and offline channels was comparable, suggesting a balanced contribution from both.
  - The statistical analysis (T-test) confirmed that the difference in total profit between online and offline channels was significant.
- **Item Type**:
  - Categories like **Technology** and **Office Supplies** showed significantly higher sales and profits compared to other categories.
  - The ANOVA test indicated that the differences in total profit between item types were statistically significant, reinforcing the need to focus on high-performing categories.
- **Geographical Region**:
  - Certain regions consistently performed better in terms of total sales and profitability. Focusing more resources (such as marketing or logistics support) on these regions could yield higher returns.

**8.2 Sales Trends**

- **Monthly and Yearly Trends**:
    - We identified clear seasonal spikes in sales, particularly during certain months. This suggests that sales campaigns and promotions should be timed to coincide with these peak periods to maximize revenue.
    - Year-over-year sales growth was also observed, with sales increasing steadily over time.

**8.3 Business Recommendations**

Based on the analysis, we can make several strategic recommendations to improve sales management:

- **Prioritize Online Sales**: With online sales yielding higher profit margins, efforts should be made to enhance the online shopping experience through better marketing, user-friendly websites, and faster shipping options.
- **Focus on High-Performing Categories**: Sales and profit generation are heavily driven by a few key categories, such as Technology and Office Supplies. The company should focus on stocking these products, negotiating better prices, and running category-specific promotions.
- **Regional Sales Strategies**: Certain regions outperformed others in terms of sales. Identifying the reasons behind this, whether it's market demand or better distribution networks, and applying those strategies in underperforming regions could lead to an overall increase in sales.