ElifSurucu filed path organized 79996a7 · 1 minute ago

897 lines (897 loc) · 35.2 KB

Project Overview

This project aims to analyze e-commerce sales data to uncover insights into sales performance, product category trends, seasonality, and customer preferences. By exploring patterns in order fulfillment, promotions, and geographic sales distribution, the project will provide actionable recommendations to help businesses optimize marketing strategies, enhance customer targeting, and boost sales performance.

Scope of the Project:

The analysis is designed to be exhaustive and insights-driven, covering detailed descriptive and inferential investigations. The goal is to explore the dataset to extract meaningful trends, test hypotheses, and derive data-driven insights that contribute to business decision-making processes.

Key Areas of Focus

Sales Performance Analysis:

- Evaluating total sales, revenue, and order quantity.
- Identifying top-performing product categories, SKUs, and sales channels.
- Measuring average order value and revenue trends.

Seasonality and Time Trends:

- Uncovering monthly and seasonal trends in sales performance.
- Analyzing peak sales periods and high cancellation months.

Customer and Geographic Insights:

- Analyzing customer behavior based on location (city/state).
- Understanding the relationship between shipping service levels and geographic regions.

Promotions and Discounts:

- Evaluating the impact of promotions on order volume and revenue.
- Comparing performance between promoted and non-promoted orders.

Order Fulfillment Insights:

- Assessing the differences in performance between orders fulfilled by Amazon and merchants.
- Analyzing the impact of shipping service levels (Standard vs. Expedited) on

- Analyzing the impact of shipping service levels (Standard vs. Expedited) on sales performance.

Inferential Analysis and Hypothesis Testing:

Testing relationships and significant differences across key variables:

- Promotion effectiveness
- Fulfillment method impact
- Geographic variations in sales and cancellations

Expected Outcomes

By conducting this analysis, the project will deliver:

- Comprehensive insights into sales trends, customer preferences, and product performance.
- Key findings on the effectiveness of promotions, fulfillment strategies, and time-based sales patterns.
- Data-driven recommendations to optimize marketing strategies, reduce cancellations, and improve sales performance.

Business Impact:

The findings will empower businesses to:

- Improve product targeting and inventory management.
- Enhance marketing strategies through insights on seasonality and promotions.
- Optimize fulfillment methods to increase customer satisfaction and reduce cancellations.
- Identify high-performing categories and target locations to maximize revenue growth.

Tools and Techniques

The project will employ:

- Data Analysis: Python (Pandas, NumPy), statistical methods, and hypothesis testing.
- Visualization: Matplotlib, Seaborn for trends and distribution analysis.
- Statistical Tests: Comparative tests, correlation analysis, and significance testing.
- Reporting: Actionable insights with visualized results for clarity and decision-making.

Imports

In [1]:

In [1]:
```
# Standard Data Science Toolkit
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt; plt.style.use("ggplot")
import seaborn as sns

# Inferential Statistical Tests
from scipy.stats import f_oneway
from statsmodels.stats.multicomp import pairwise_tukeyhsd
```

Data Cleaning/Processing

In [2]:
```
file_path = r"c:\Users\Elif Surucu\Documents\Flatiron\Assesments\Capstone\Anal
ecommerce_data = pd.read_csv(file_path)
ecommerce_data.head()
```

Out[2]:

	index	Order ID	Date	Status	Fulfilment	Sales Channel	ship-service-level	Style	
0	1	171-9198151-1101146	2022-04-30	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	JNE3781	JNE3 KR-
1	7	406-7807733-3785945	2022-04-30	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	JNE3405	JNE3
2	12	405-5513694-8146768	2022-04-30	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	JNE3405	JNE3 K
3	14	408-1298370-1920302	2022-04-30	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	J0351	JC S
4	15	403-4965581-9520319	2022-04-30	Shipped - Delivered to Buyer	Merchant	Amazon.in	Standard	PJNE3368	PJNE3 KF

5 rows × 23 columns

In [3]:
```
ecommerce_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32395 entries, 0 to 32394
Data columns (total 23 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   index               32395 non-null  int64  
 1   Order ID            32395 non-null  object 
 2   Date                32395 non-null  object 
 3   Status              32395 non-null  object 
 4   Fulfilment          32395 non-null  object 
 5   Sales Channel       32395 non-null  object 
 6   ship-service-level  32395 non-null  object 
 7   Style               32395 non-null  object 
 8   SKU                 32395 non-null  object 
 9   Category            32395 non-null  object 
 10  Size                32395 non-null  object 
 11  ASIN                32395 non-null  object 
 12  Courier Status      32395 non-null  object 
 13  Qty                 32395 non-null  int64  
 14  currency            32395 non-null  object 
 15  Amount              32395 non-null  float64
 16  ship-city           32395 non-null  object 
 17  ship-state          32395 non-null  object 
 18  ship-postal-code    32395 non-null  float64
 19  ship-country        32395 non-null  object 
 20  promotion-ids       32395 non-null  object 
 21  B2B                 32395 non-null  bool   
 22  fulfilled-by        32395 non-null  object 
dtypes: bool(1), float64(2), int64(2), object(18)
memory usage: 5.5+ MB
```

In [4]:
```
ecommerce_data.describe()
```

Out[4]:

	index	Qty	Amount	ship-postal-code
count	32395.000000	32395.000000	32395.000000	32395.000000
mean	60956.478160	1.004846	650.522920	462097.701096
std	36843.686311	0.085035	284.913465	194276.943115
min	1.000000	1.000000	0.000000	110001.000000
25%	27188.500000	1.000000	459.000000	370001.000000
50%	63461.000000	1.000000	631.000000	500017.000000
75%	91761.500000	1.000000	771.000000	600037.000000
max	128891.000000	5.000000	5495.000000	855115.000000

Unnamed: Deleting the unnecessary column named 22 from the dataset.

In [5]:
```
ecommerce_data = ecommerce_data.drop(columns=['Unnamed: 22'], errors='ignore')
```

Checking the number of missing values (NaN) in each column.

In [6]:
```
missing_values = ecommerce_data.isnull().sum()
```

Converting a date column to date format (datetime)

In [7]:
```
ecommerce_data['Date'] = pd.to_datetime(ecommerce_data['Date'], errors='coerce
```

Convert the values in the ship-postal-code column to string (text).

In [8]:
```
ecommerce_data['ship-postal-code'] = ecommerce_data['ship-postal-code'].astype
```

Removing duplicate rows from a dataset.

In [9]:
```
ecommerce_data = ecommerce_data.drop_duplicates()
```

In [10]:
```
#Summary

cleaned_summary = {
    "missing_values_after_cleaning": missing_values,
    "total_rows_after_cleaning": len(ecommerce_data),
    "duplicates_removed": 128975 - len(ecommerce_data)
}
cleaned_summary
```

Out[10]:
```
{'missing_values_after_cleaning': index                0
 Order ID             0
 Date                 0
 Status               0
 Fulfilment           0
 Sales Channel        0
 ship-service-level   0
 Style                0
 SKU                  0
 Category             0
 Size                 0
 ASIN                 0
 Courier Status       0
 Qty                  0
 currency             0
 Amount               0
 ship-city            0
 ship-state           0
 ship-postal-code     0
 ship-country         0
 promotion-ids        0
```

```
B2B                 0
fulfilled-by        0
dtype: int64,
'total_rows_after_cleaning': 32395,
'duplicates_removed': 96580}
```

In [11]:
```
critical_columns = ['Courier Status', 'fulfilled-by', 'currency', 'Amount',
                    'ship-city', 'ship-state', 'ship-postal-code', 'ship-count
ecommerce_data = ecommerce_data.dropna(subset=critical_columns)
```

In [12]:
```
ecommerce_data['promotion-ids'] = ecommerce_data['promotion-ids'].fillna('No P
```

In [13]:
```
final_summary = {
    "missing_values": ecommerce_data.isnull().sum(),
    "total_rows_after_cleaning": len(ecommerce_data),
    "total_columns": len(ecommerce_data.columns)
}
final_summary
```

Out[13]:
```
{'missing_values': index               0
 Order ID              0
 Date                  0
 Status                0
 Fulfilment            0
 Sales Channel         0
 ship-service-level    0
 Style                 0
 SKU                   0
 Category              0
 Size                  0
 ASIN                  0
 Courier Status        0
 Qty                   0
 currency              0
 Amount                0
 ship-city             0
 ship-state            0
 ship-postal-code      0
 ship-country          0
 promotion-ids         0
 B2B                   0
 fulfilled-by          0
 dtype: int64,
 'total_rows_after_cleaning': 32395,
 'total_columns': 23}
```

Final Cleaning Summary

- Missing Data: All critical columns have been removed from the missing data and no columns are missing anymore.

- Total Row Count: 32,395
- Total Column Count: 23

Dataset is ready for analysis!

In [14]:
```
# Save the cleaned dataset to a new CSV file
cleaned_file_path = r"c:\Users\Elif Surucu\Documents\Flatiron\Assesments\Capst
ecommerce_data.to_csv(cleaned_file_path, index=False)

cleaned_file_path
```

Out[14]:
```
'c:\\Users\\Elif Surucu\\Documents\\Flatiron\\Assesments\\Capstone\\Analyzing_
E_Commerce_SalesPerformance\\Amazon_Sale_Report.csv'
```

In [15]:
```
ecommerce_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32395 entries, 0 to 32394
Data columns (total 23 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   index               32395 non-null  int64
 1   Order ID            32395 non-null  object
 2   Date                32395 non-null  datetime64[ns]
 3   Status              32395 non-null  object
 4   Fulfilment          32395 non-null  object
 5   Sales Channel       32395 non-null  object
 6   ship-service-level  32395 non-null  object
 7   Style               32395 non-null  object
 8   SKU                 32395 non-null  object
 9   Category            32395 non-null  object
 10  Size                32395 non-null  object
 11  ASIN                32395 non-null  object
 12  Courier Status      32395 non-null  object
 13  Qty                 32395 non-null  int64
 14  currency            32395 non-null  object
 15  Amount              32395 non-null  float64
 16  ship-city           32395 non-null  object
 17  ship-state          32395 non-null  object
 18  ship-postal-code    32395 non-null  object
 19  ship-country        32395 non-null  object
 20  promotion-ids       32395 non-null  object
 21  B2B                 32395 non-null  bool
 22  fulfilled-by        32395 non-null  object
dtypes: bool(1), datetime64[ns](1), float64(1), int64(2), object(18)
memory usage: 5.5+ MB
```

The next step:

- We can explore the data with Descriptive Analysis.
- We can perform hypothesis testing with Inferential Analysis.
- We can make the analysis results more understandable with Data Visualization.

Descriptive Analysis Questions

Category	Questions
General Sales Insights	1. What is the total number of orders placed?
	2. What is the total revenue generated?
	3. What is the average order value across all orders?
	4. What are the top 10 best-selling product categories by total sales?
	5. Which SKUs (product codes) have the highest total quantity sold?
	6. Which SKUs generate the highest revenue?
	7. What are the monthly sales trends over time? (group by Date)
	8. Which fulfillment method (Fulfilment) contributes the most to sales?
	9. What is the distribution of Status (shipped, canceled, etc.)?
	10. Which Sales Channel generates the most sales and revenue?
	11. What is the average order quantity (Qty) across different categories?
Seasonality & Time Trends	12. What are the peak sales months and seasons?
	13. Is there a weekly or daily pattern in sales volume?
	14. Which months show the highest cancellation rates?
Customer Location Trends	15. Which ship-city and ship-state have the most orders?
	16. What is the average revenue per shipping state or city?
	17. Which states or cities have the highest cancellation rates?
Promotions & Discounts	18. How many orders included promotion-ids?
	19. What is the average revenue of promoted vs. non-promoted orders?
	20. Which promotions were the most frequently used?
Fulfillment Methods	21. What is the split between orders fulfilled by Amazon and merchants?
	22. What is the average order value for Amazon-fulfilled orders vs. Merchant-fulfilled?
	23. What is the distribution of ship-service-level (Standard vs. Expedited)?

Inferential Analysis Questions

Question	Type of Analysis	Statistical Test
1. Is there a significant difference in average revenue across different product categories?	Compare means	ANOVA
2. Is there a significant difference in sales (revenue) across months for standard shipping orders?	Compare two means	ANOVA
3. Are orders with promotions significantly different in revenue compared to those without promotions?	Compare two means	ANOVA
4. Is there a difference in average Qty sold across		