Est. 1841 YORK ST JOHN UNIVERSITY

The Coversheet

Student Number (as shown on student ID card):	SAMUEL OYEWUNMI
Word Count / Pages / Duration / Other Limits:	15 PAGES
Attempt Number:	1
Date of Submission:	14/01/2025

I have read and understood the Academic Misconduct statement.	Tick to confirm ✓
I have read and understood the Generative Artificial Intelligence use statement.	Tick to confirm ✓
I am satisfied that I have met the Learning Outcomes of this assignment (please check the Assignment Brief if you are unsure)	Met ✓

Self-Assessment – If there are particular aspects of your assignment on which you would like feedback, please indicate below.

Optional for students

Suggested prompt questions-

How have you developed or progressed your learning in this work?

What do you feel is the strongest part of this submission?

What feedback would you give yourself?

What part(s) of this assignment are you still unsure about?

Assessor’s Feedback (may be delivered in line with the submission)	
Were the learning outcomes met?	Yes ✓If not, what was not met:
Assessor’s response to the student’s submission, request for feedback and / or self-assessment (feedback):	

Assessor's Feedback (may be delivered in line with the submission)
What specific actions should the student undertake to progress their learning? (feedforward):
Please take this and other feedback to your next academic tutorial to plan your future work.

ASSESSMENT OBJECTIVE:

- To optimize its sales strategy by analysing historical transaction data.
- To analyze and compare the monthly fluctuations in total revenue and the number of transactions
- To determine which product categories have the highest total revenue and demonstrate consistent revenue growth trends.
- To explore the seasonal variations in sales for different product categories.
- To analyze shifts in customer purchasing behavior across multiple transactions to identify recurring patterns or significant changes in preferences.
- To determine whether these trends can provide actionable insights for enhancing the company's marketing strategy.

ASSESMENT TASK

For this assessment, an exploratory data analysis will be carried out on the dataset gotten from an e-commerce company that wants to optimize its sales strategy by analysing historical transaction data. The company has a database containing details of customer transactions, including customer ID, transaction date, product ID, product category, quantity purchased, and total price.

TOOLS & LIBRARIES USED IN EDA:

Python Software and its Libraries:

- **Pandas**: For data manipulation and cleaning.
- **Matplotlib** & **Seaborn**: For visualizations.
- **NumPy**: For numerical operations.
- **Plotly**: For interactive visualizations.

What is Exploratory Data Analysis (EDA)?

Exploratory Data Analysis (EDA) is a critical step in the data analysis process that involves investigating datasets to summarize their main characteristics, often using statistical and visualization techniques. Introduced by statistician John Tukey in the 1970s, EDA is used to uncover patterns, spot anomalies, test hypotheses, and check assumptions before applying advanced modeling techniques. The goal is to maximize insights and guide further analysis by understanding the structure of the data.

Key components of EDA include:

- **Missing Data Analysis**: Identifying gaps and patterns in missing data.
- **Descriptive Statistics**: Measures such as mean, median, variance, and standard deviation to understand data distribution.
- **Data Visualization**: Tools like histograms, box plots, scatter plots, and heatmaps to identify trends, correlations, and outliers.
- **Feature Relationships**: Analysing relationships between variables through correlation matrices or pairwise plots.

In this data analysis process, exploratory data analysis (EDA) is an essential first step, particularly when working with e-commerce datasets. Prior to modelling or additional analysis, EDA enables analysts to comprehend the structure of the dataset, find patterns, spot anomalies, and test hypotheses. In the context of an e-commerce dataset, the following are the primary steps and procedures used to perform EDA:

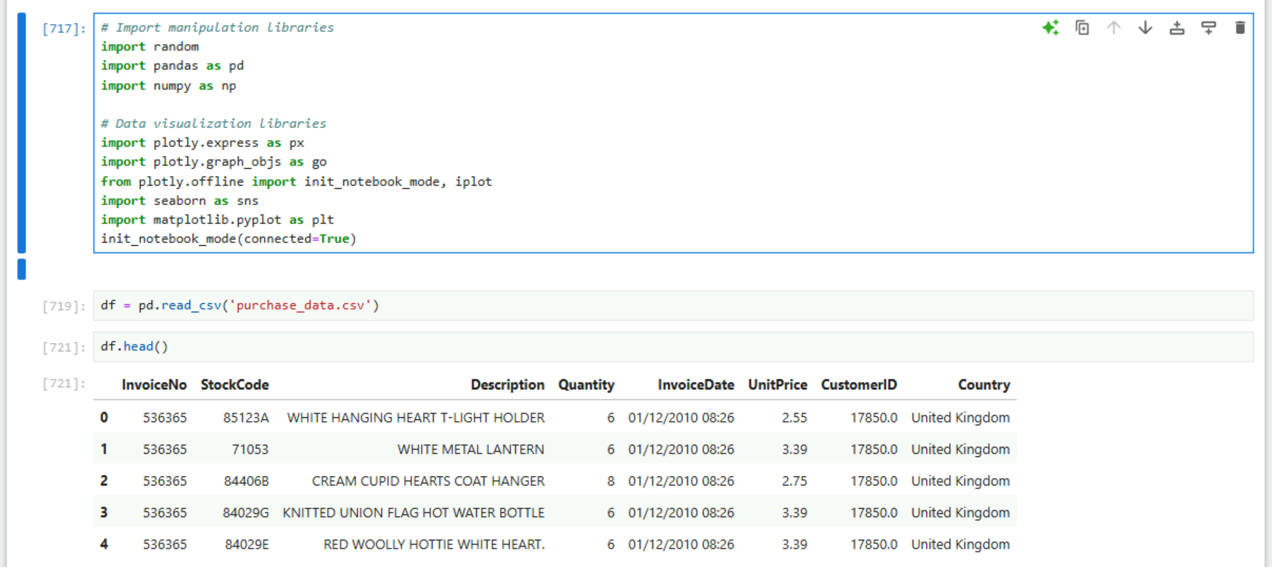
About the dataset

From our initial look at the information, which includes transactions for an e-commerce service, we may infer the following characteristics:

- **InvoiceNo:** Each transaction's unique code is the invoice number ('c') in the beginning indicates that the transaction was cancelled, I suppose.
- **Product Code (StockCode):** A special code assigned to every product.
- **Product Description:** This is the product's name.
- **Quantity:** The total number of items sold during a transaction.
- **InvoiceDate:** The time and date of the transaction are indicated by the invoice date.
- **UnitPrice:** The cost of one unit of the product expressed in currency is known as the unit price .
- **Customer ID**: A special code that is specific to every customer.
- **Country:** The nation in which the client calls home.

Data Collection & Loading:

Firstly, the Dataset used is an ecommerce data(purchase data.csv file) which is imported or loaded of the dataset into the Juptyer notebook using *pd.read(”purchase_data.csv”)* after stating the libraries needed for the EDA as shown below . Then the dataset is examined using the head() or info() functions to look at the first 5 few rows to obtain a general idea of the dataset's structure.



[717]:
```
# Import manipulation libraries
import random
import pandas as pd
import numpy as np

# Data visualization libraries
import plotly.express as px
import plotly.graph_objs as go
from plotly.offline import init_notebook_mode, iplot
import seaborn as sns
import matplotlib.pyplot as plt
init_notebook_mode(connected=True)
```

[719]:
```
df = pd.read_csv('purchase_data.csv')
```

[721]:
```
df.head()
```

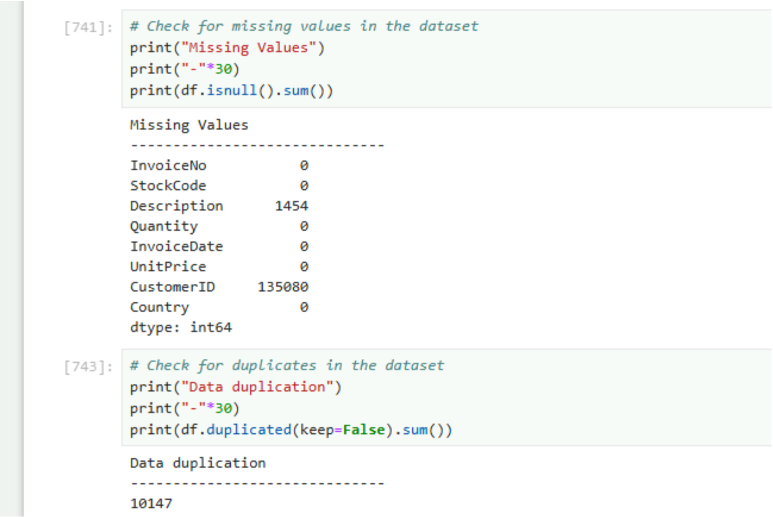
[721]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01/12/2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01/12/2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01/12/2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01/12/2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01/12/2010 08:26	3.39	17850.0	United Kingdom

Data Preprocessing.

Once the dataset is obtained, the subsequent step is data preprocessing, which involves preparing the data for exploration. This process includes tasks such as handling missing values, converting data types, eliminating duplicate entries, and creating new columns.

The raw **Purchase_dataset.csv** contains ***136,534 rows*** with missing values and data duplication of ***10147 data*** which were subsequently removed.



```
[741]: # Check for missing values in the dataset
       print("Missing Values")
       print("-"*30)
       print(df.isnull().sum())
```

```
Missing Values
------------------------------
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

```
[743]: # Check for duplicates in the dataset
       print("Data duplication")
       print("-"*30)
       print(df.duplicated(keep=False).sum())
```

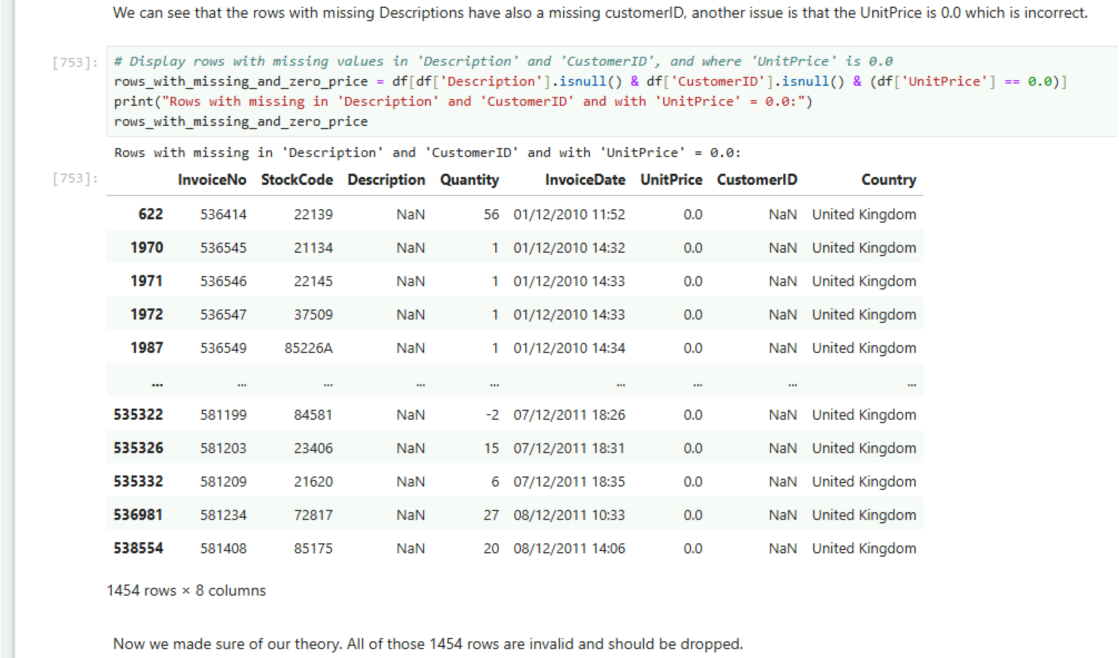
```
Data duplication
------------------------------
10147
```

Fig. 1.0 Missing values in Purchase_dataset.csv codes

From the table above, it is evident that the **Description** and **CustomerID** columns have missing values. Additionally, the **UnitPrice** column contains invalid entries with a value of 0, and the **Quantity** column has negative values, which are incorrect.

To address these issues, the next steps include:

1. Dropping rows with missing values.
2. Removing rows where the **UnitPrice** is 0.
3. Filtering the data to include only rows where the **Quantity** is greater than 0, eliminating negative values in the **Quantity** column.



We can see that the rows with missing Descriptions have also a missing customerID, another issue is that the UnitPrice is 0.0 which is incorrect.

```
[753]: # Display rows with missing values in 'Description' and 'CustomerID', and where 'UnitPrice' is 0.0
       rows_with_missing_and_zero_price = df[df['Description'].isnull() & df['CustomerID'].isnull() & (df['UnitPrice'] == 0.0)]
       print("Rows with missing in 'Description' and 'CustomerID' and with 'UnitPrice' = 0.0:")
       rows_with_missing_and_zero_price
```

```
Rows with missing in 'Description' and 'CustomerID' and with 'UnitPrice' = 0.0:
```

[753]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
622	536414	22139	NaN	56	01/12/2010 11:52	0.0	NaN	United Kingdom
1970	536545	21134	NaN	1	01/12/2010 14:32	0.0	NaN	United Kingdom
1971	536546	22145	NaN	1	01/12/2010 14:33	0.0	NaN	United Kingdom
1972	536547	37509	NaN	1	01/12/2010 14:33	0.0	NaN	United Kingdom
1987	536549	85226A	NaN	1	01/12/2010 14:34	0.0	NaN	United Kingdom
...	...	...	...	...	...	...	...	...
535322	581199	84581	NaN	-2	07/12/2011 18:26	0.0	NaN	United Kingdom
535326	581203	23406	NaN	15	07/12/2011 18:31	0.0	NaN	United Kingdom
535332	581209	21620	NaN	6	07/12/2011 18:35	0.0	NaN	United Kingdom
536981	581234	72817	NaN	27	08/12/2011 10:33	0.0	NaN	United Kingdom
538554	581408	85175	NaN	20	08/12/2011 14:06	0.0	NaN	United Kingdom

1454 rows × 8 columns

Now we made sure of our theory. All of those 1454 rows are invalid and should be dropped.

Fig. 2.0 Missing values in Purchase_dataset.csv codes

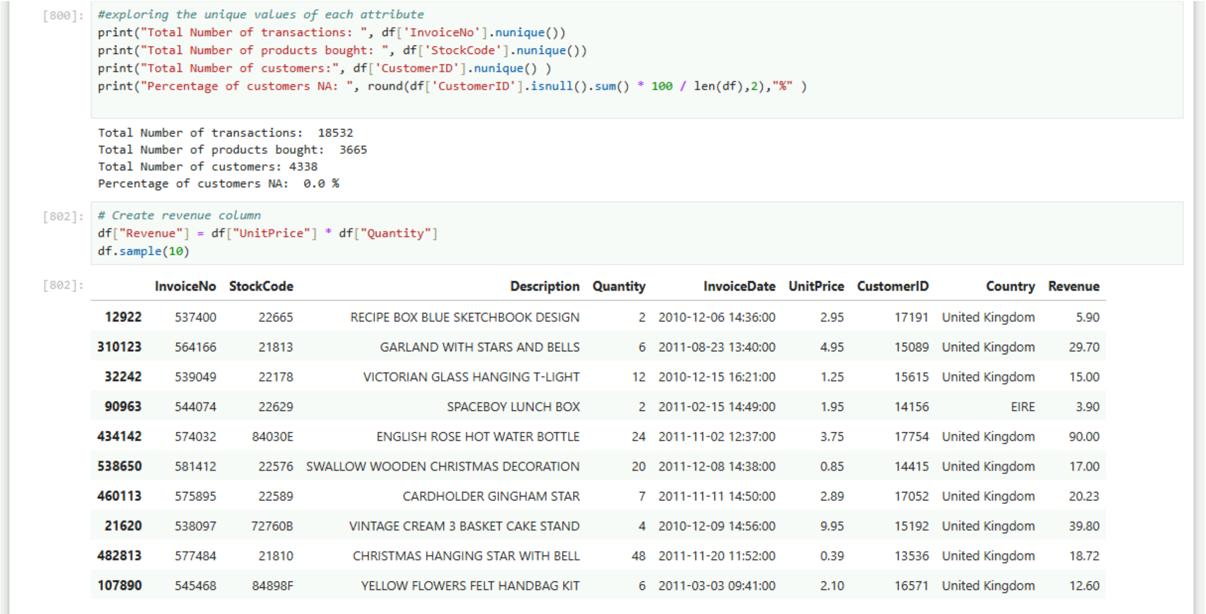
After completing data cleaning, the total number of rows in the dataset decreased from 541,909 to 392,732.

The next steps involve as shown in the image below:

➢ Changing the data types of specific columns:

- **InvoiceDate**: from object to datetime.
- **InvoiceNo**: from object to integer.
- **CustomerID**: from object to integer.

➢ Adding a **Revenue** column to simplify the data exploration process. The **Revenue** column is calculated by multiplying the **Quantity** column by the **UnitPrice** column.



```
[800]: #exploring the unique values of each attribute
       print("Total Number of transactions: ", df['InvoiceNo'].nunique())
       print("Total Number of products bought: ", df['StockCode'].nunique())
       print("Total Number of customers:", df['CustomerID'].nunique() )
       print("Percentage of customers NA: ", round(df['CustomerID'].isnull().sum() * 100 / len(df),2),"%" )

       Total Number of transactions:  18532
       Total Number of products bought:  3665
       Total Number of customers: 4338
       Percentage of customers NA:  0.0 %

[802]: # Create revenue column
       df["Revenue"] = df["UnitPrice"] * df["Quantity"]
       df.sample(10)
```

[802]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Revenue
12922	537400	22665	RECIPE BOX BLUE SKETCHBOOK DESIGN	2	2010-12-06 14:36:00	2.95	17191	United Kingdom	5.90
310123	564166	21813	GARLAND WITH STARS AND BELLS	6	2011-08-23 13:40:00	4.95	15089	United Kingdom	29.70
32242	539049	22178	VICTORIAN GLASS HANGING T-LIGHT	12	2010-12-15 16:21:00	1.25	15615	United Kingdom	15.00
90963	544074	22629	SPACEBOY LUNCH BOX	2	2011-02-15 14:49:00	1.95	14156	EIRE	3.90
434142	574032	84030E	ENGLISH ROSE HOT WATER BOTTLE	24	2011-11-02 12:37:00	3.75	17754	United Kingdom	90.00
538650	581412	22576	SWALLOW WOODEN CHRISTMAS DECORATION	20	2011-12-08 14:38:00	0.85	14415	United Kingdom	17.00
460113	575895	22589	CARDHOLDER GINGHAM STAR	7	2011-11-11 14:50:00	2.89	17052	United Kingdom	20.23
21620	538097	72760B	VINTAGE CREAM 3 BASKET CAKE STAND	4	2010-12-09 14:56:00	9.95	15192	United Kingdom	39.80
482813	577484	21810	CHRISTMAS HANGING STAR WITH BELL	48	2011-11-20 11:52:00	0.39	13536	United Kingdom	18.72
107890	545468	84898F	YELLOW FLOWERS FELT HANDBAG KIT	6	2011-03-03 09:41:00	2.10	16571	United Kingdom	12.60

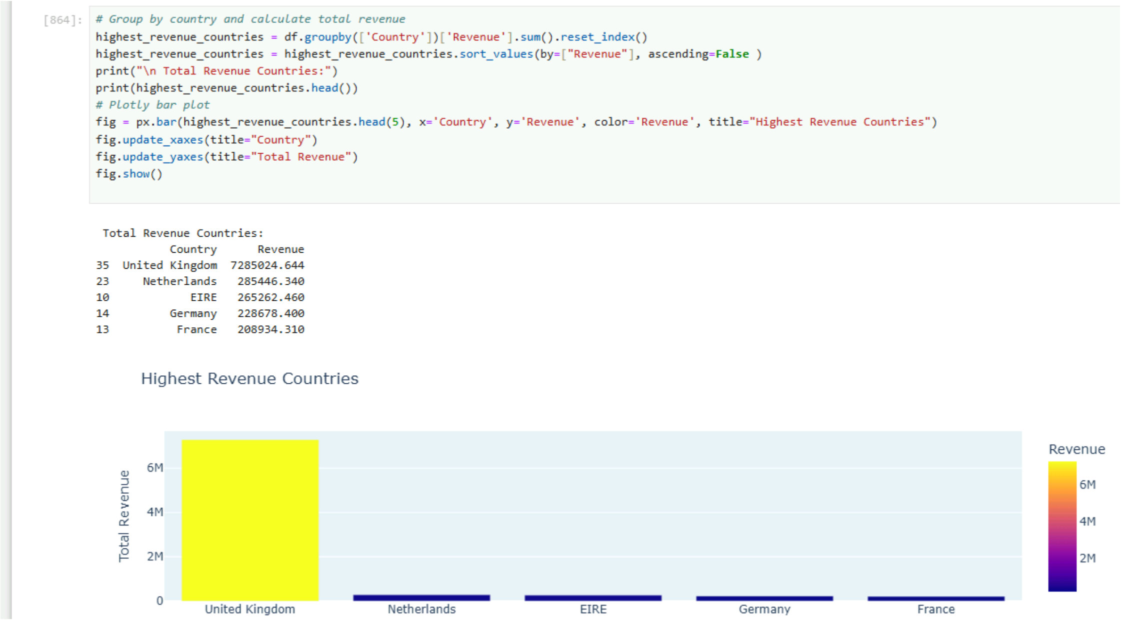
Addition of a new column called Revenue

The data preprocessing stage has resulted in a dataset with 9 columns: **InvoiceNo**, **StockCode**, **Description**, **Quantity**, **InvoiceDate**, **UnitPrice**, **CustomerID**, **Country**, and **Revenue**. The dataset now consists of ***392,732 rows.***

Data Exploration Process

The dataset includes customers from 37 countries. The **United Kingdom (UK)** generated the highest revenue, totaling £7,2M, and accounted for 81.97% of all orders across these countries.

The **Netherlands** and **EIRE** ranked second and third in revenue, with totals of **£285,446.34 and £265,262.46, respectively.**



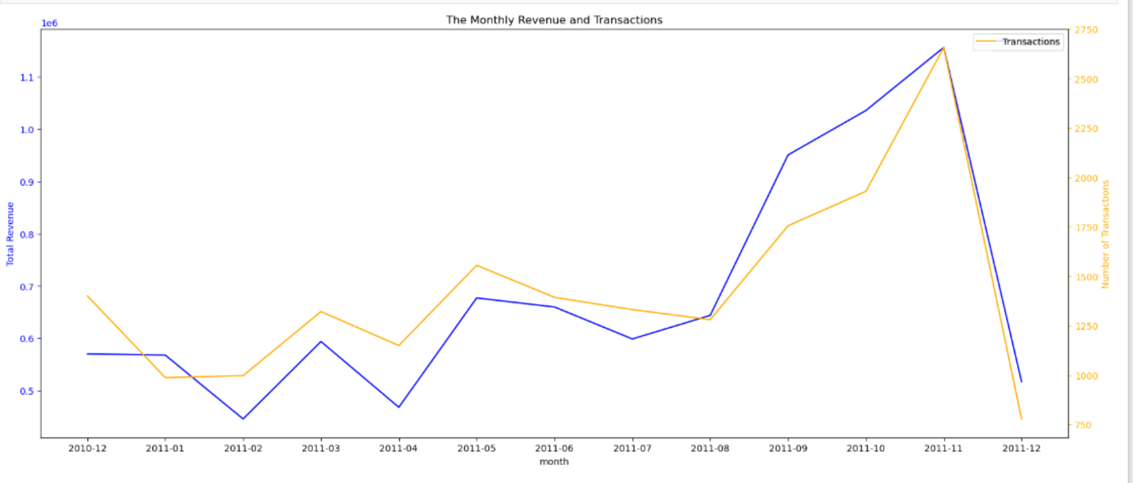
```
[864]: # Group by country and calculate total revenue
highest_revenue_countries = df.groupby(['Country'])['Revenue'].sum().reset_index()
highest_revenue_countries = highest_revenue_countries.sort_values(by=["Revenue"], ascending=False )
print("\n Total Revenue Countries:")
print(highest_revenue_countries.head())
# Plotly bar plot
fig = px.bar(highest_revenue_countries.head(5), x='Country', y='Revenue', color='Revenue', title="Highest Revenue Countries")
fig.update_xaxes(title="Country")
fig.update_yaxes(title="Total Revenue")
fig.show()
```

```
 Total Revenue Countries:
           Country      Revenue
35  United Kingdom  7285024.644
23     Netherlands   285446.340
10            EIRE   265262.460
14         Germany   228678.400
13          France   208934.310
```

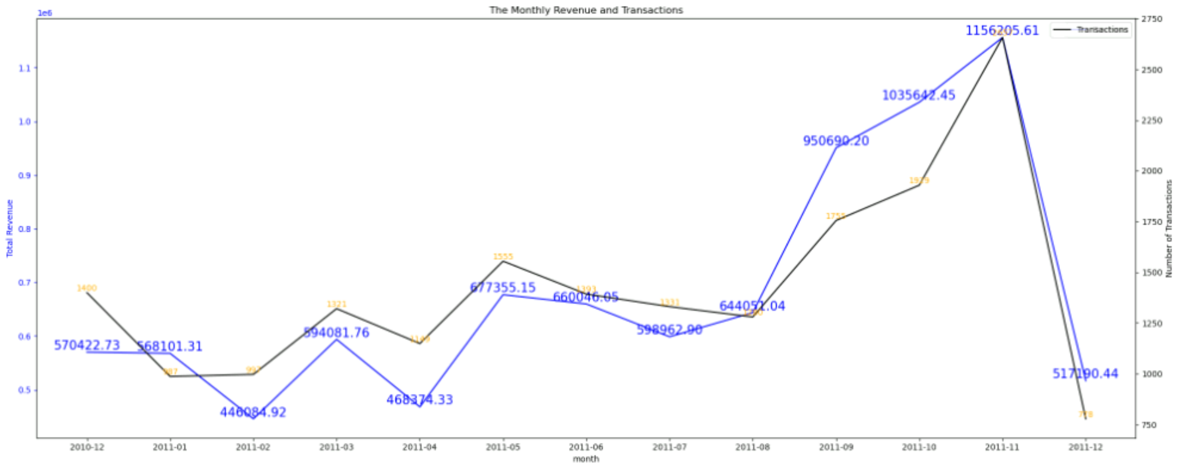
Question 1:

Analyze and compare the monthly fluctuations in total revenue and the number of transactions. Identify any significant anomalies or outliers.

Monthly revenue(the blue line) shows noticeable fluctuations, with February 2011 recording the lowest revenue of £446,084.92 and November 2011 achieving the highest revenue at £1,156,205.61. and the number of transactions (the Orange and Black line) fluctuates every month, with the lowest revenue in December 2012 while the highest revenue occurred in November 2011 which is same period when the total revenue is £1,156,205.61. Then the significant anomalies or outliers in period of lowest revenue might be as result of low marketing seasonal demand or promotion.



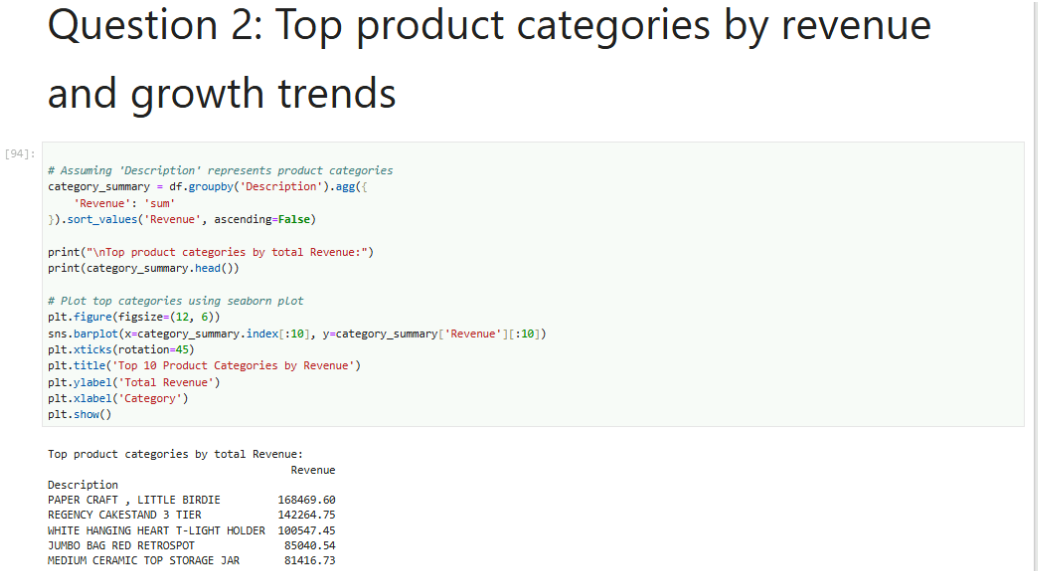
This shows that there is high revenue and high number of transaction in November 2011 than other months



Question 2:

Determine which product categories have the highest total revenue and demonstrate consistent revenue growth trends. Identify any categories with sustained increases in sales.

Checking the Bar chart below, The Top purchased product by Revenue was **Paper Craft Little Birdie**, with a total of £168,469.60 items ordered. Following closely, the **Regency Cakestand 3 Tier** ranked second with £142,264.75 items, while the **White Hanging Heart T-Light Holder** secured the third spot with £100,547.45 items. For the **Sustained Growth,** All three categories show sustained revenue increases, but **Paper Craft Little Birdie** stands out for its significant total revenue and consistent growth.



Question 2: Top product categories by revenue and growth trends

[94]:

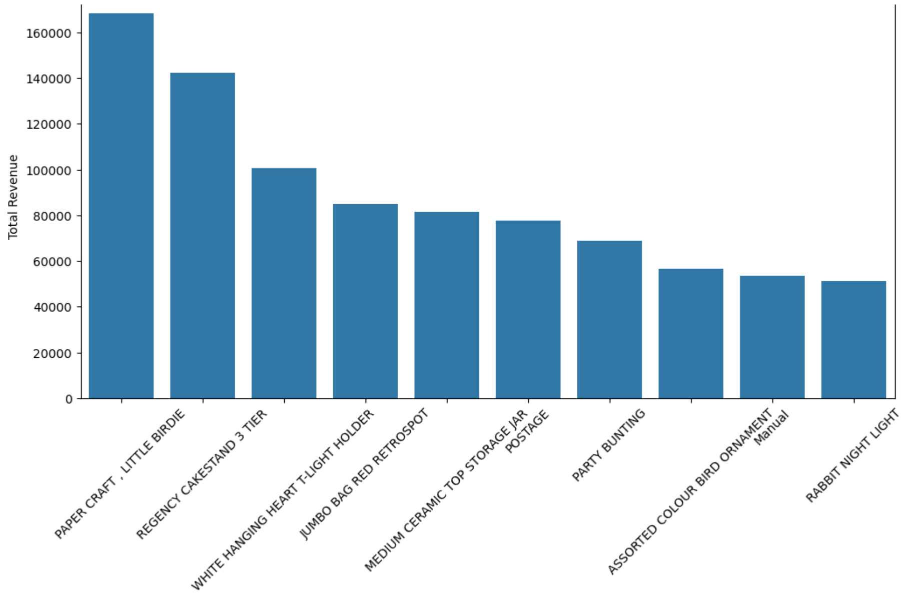
```
# Assuming 'Description' represents product categories
category_summary = df.groupby('Description').agg({
    'Revenue': 'sum'
}).sort_values('Revenue', ascending=False)

print("\nTop product categories by total Revenue:")
print(category_summary.head())

# Plot top categories using seaborn plot
plt.figure(figsize=(12, 6))
sns.barplot(x=category_summary.index[:10], y=category_summary['Revenue'][:10])
plt.xticks(rotation=45)
plt.title('Top 10 Product Categories by Revenue')
plt.ylabel('Total Revenue')
plt.xlabel('Category')
plt.show()
```

```
Top product categories by total Revenue:
                                      Revenue
Description
PAPER CRAFT , LITTLE BIRDIE         168469.60
REGENCY CAKESTAND 3 TIER            142264.75
WHITE HANGING HEART T-LIGHT HOLDER  100547.45
JUMBO BAG RED RETROSPOT              85040.54
MEDIUM CERAMIC TOP STORAGE JAR       81416.73
```

Code for the top Product Categories by Revenue



Bar Chart showing the Top Product Categories by Revenue

Question 3:

Explore the seasonal variations in sales for different product categories. Are there any categories that are sensitive to specific time periods?

The analysis of seasonal sales variations across different product categories reveals the following key insights:

Seasonal Sales Highlights

- Top Performing Month:
  - ✓ January stands out as the month with the highest sales across all product categories.
- Product-Specific Seasonal Variation:
  - ✓ Set 2 Tea Towels "I Love London"
    - ▪ **Highest Seasonal Revenue Variation:** £1,107.20.
    - ▪ **Performance:** Demonstrated the most significant fluctuation in revenue during seasonal changes, indicating strong responsiveness to seasonal demand.

Key Insights

- **January's Dominance:**

✓ The peak in sales during January may be attributed to post-holiday shopping trends, New Year promotions, or seasonal events that drive higher consumer spending.

- **Set 2 Tea Towels "I Love London":**

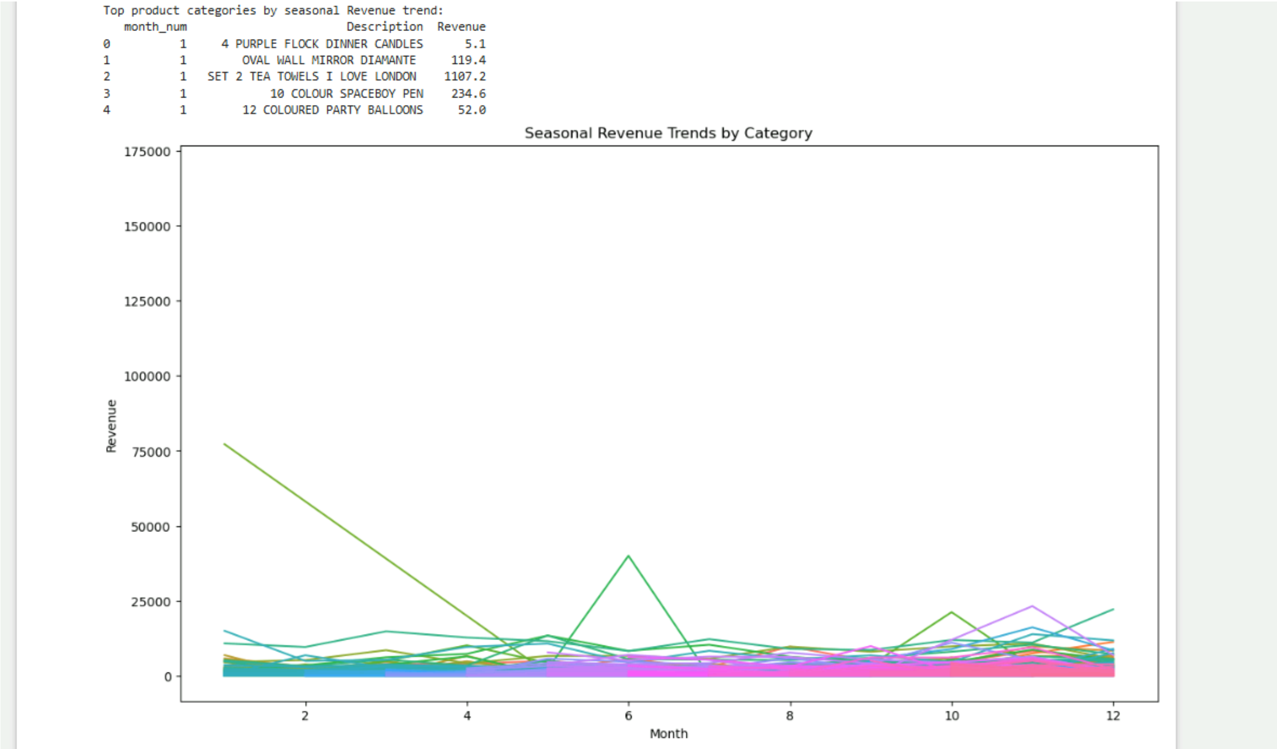
- ✓ The substantial seasonal revenue variation suggests that this product is highly popular during specific times of the year, possibly aligning with seasonal décor trends or promotional campaigns.

Implications for Strategy

- Leveraging Peak Months:
- ✓ Focus marketing and inventory strategies around January to maximize sales during the highest-performing month.
- Capitalizing on High-Variation Products:
- ✓ Implement targeted promotions and stock management for products like Set 2 Tea Towels "I Love London" to enhance revenue during peak seasonal periods and mitigate the impact of lower sales during off-peak times.

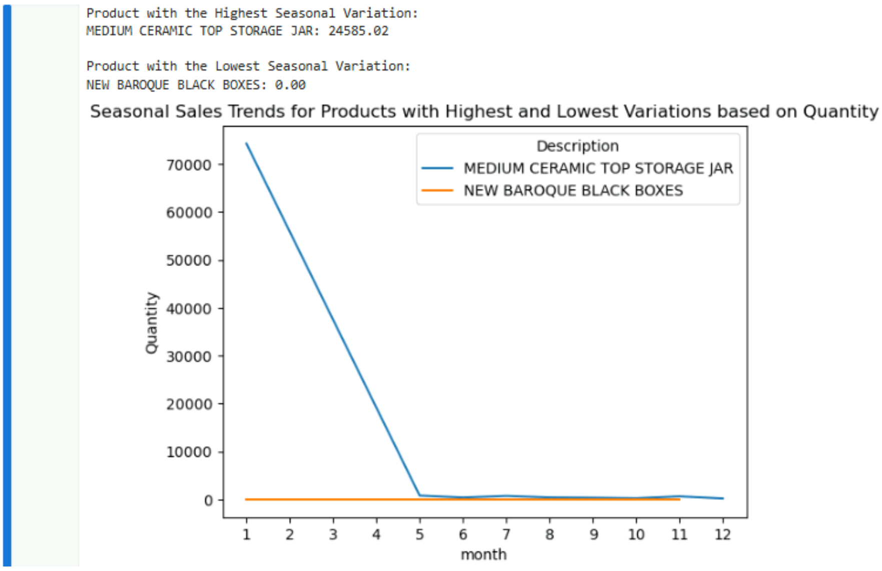
Recommendations

- Enhanced Marketing in January:
  - ✓ Increase advertising efforts, offer special discounts, and introduce new product launches in January to capitalize on the high sales potential.
- Seasonal Promotions for Key Products:
  - ✓ Develop seasonal campaigns specifically for products with significant revenue variations, such as the Set 2 Tea Towels "I Love London," to sustain and boost their performance throughout the year.
- Inventory Optimization:
  - ✓ Align inventory levels with expected seasonal demand patterns to ensure adequate stock during peak months and reduce overstocking during slower periods.
- By understanding and addressing these seasonal sales dynamics, the business can optimize its sales strategies, enhance revenue growth, and improve overall market responsiveness.



```
Top product categories by seasonal Revenue trend:
   month_num                         Description  Revenue
0          1     4 PURPLE FLOCK DINNER CANDLES      5.1
1          1       OVAL WALL MIRROR DIAMANTE     119.4
2          1   SET 2 TEA TOWELS I LOVE LONDON   1107.2
3          1          10 COLOUR SPACEBOY PEN    234.6
4          1        12 COLOURED PARTY BALLOONS     52.0
```

Seasonal Variation in sales for Product Category based on Revenue



```
Product with the Highest Seasonal Variation:
MEDIUM CERAMIC TOP STORAGE JAR: 24585.02

Product with the Lowest Seasonal Variation:
NEW BAROQUE BLACK BOXES: 0.00
```

Seasonal Sales Trends for Products with Highest and Lowest Variations based on Quantity
Description
MEDIUM CERAMIC TOP STORAGE JAR
NEW BAROQUE BLACK BOXES
Quantity
70000
60000
50000
40000
30000
20000
10000
0
1
2
3
4
5
6
7
8
9
10
11
12
month

Seasonal Variation in sales for Product Category based on Quantity

Question 4:

Analyze shifts in customer purchasing behavior across multiple transactions to identify recurring patterns or significant changes in preferences. Determine whether these trends can provide actionable insights for enhancing the company's marketing strategy.

Based on the scatter plot below, the customer purchasing behavior across multiple transactions shows that the number of transactions with the highest revenue made by a customer is between 50 and 100 with over 250,000 total revenue.

Key Insights

1. **Revenue Concentration:**

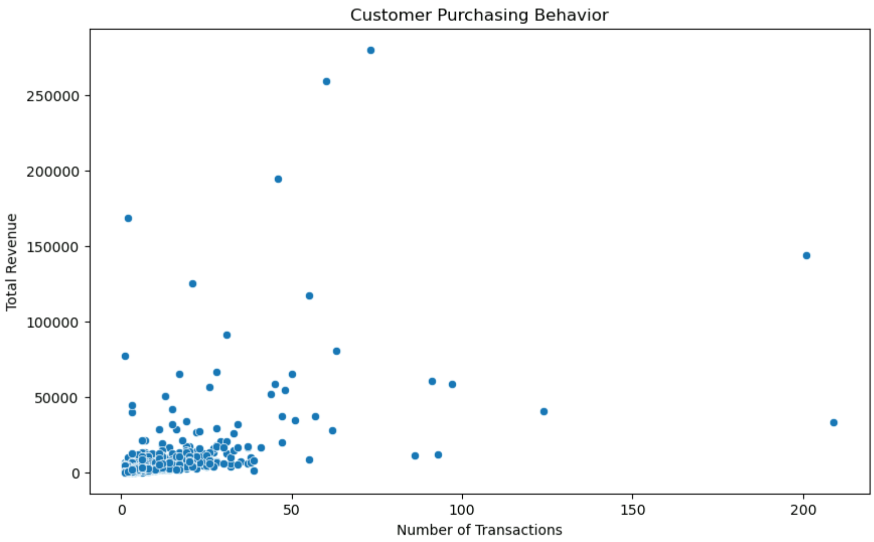
   A substantial portion of total sales (£250,000+) is attributed to customers with 50 to 100 transactions, demonstrating that this group is a key revenue driver.

2. **Customer Loyalty:**

   The high number of transactions suggests a high level of customer satisfaction and loyalty, which are crucial for sustained revenue growth.

3. **Profitability Potential:**

   Focusing on this segment can yield significant returns, as these customers are already showing a propensity to make repeated purchases.



Conclusion

Definitely, this trends can provide actionable insight in order to enhance the company's marketing strategy because there are much number of transactions with less total revenue.

Understanding seasonal sales variations is crucial for optimizing sales strategies and inventory management. January's exceptional performance and the significant revenue variation in specific products like Set 2 Tea Towels "I Love London" highlight areas of opportunity. By implementing targeted marketing efforts and strategic planning, the business can enhance revenue growth and achieve sustained success across all seasons.

REFERENCES

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.