COM618 – Data Science – Week 2 Lab

Week 2 Practical Lab Guide: Exploratory Data Analysis (EDA)

Title: Exploring UK Retail Data **Level:** 6 (UK Higher Education)

Step 1: Download the Dataset

Use a publicly available UK retail dataset such as:

- Online Retail Dataset on GitHub
- Synthetic UK Retail Dataset on Hugging Face

Upload the CSV file to your Jupyter Notebook or Google Colab environment.

Step 2: Load and Inspect the Data

```
import pandas as pd


# Load the dataset
df = pd.read_csv('OnlineRetail.csv', encoding='ISO-8859-1')


# Preview the data
df.head()
```

Step 3: Understand the Structure

```
# Check data types and nulls
df.info()


# Summary statistics
df.describe()


# Count missing values
df.isnull().sum()
```

Tip: Identify which columns have missing values and which are numerical vs categorical.

Step 4: Clean the Data

```
# Drop rows with missing CustomerID
df_clean = df.dropna(subset=['CustomerID'])

# Remove negative quantities (returns)
df_clean = df_clean[df_clean['Quantity'] > 0]

# Create a new column for TotalPrice
df_clean['TotalPrice'] = df_clean['Quantity'] * df_clean['UnitPrice']
```

Step 5: Convert Dates and Extract Features

```
# Convert InvoiceDate to datetime
df_clean['InvoiceDate'] = pd.to_datetime(df_clean['InvoiceDate'])

# Extract Year and Month
df_clean['InvoiceYear'] = df_clean['InvoiceDate'].dt.year
df_clean['InvoiceMonth'] = df_clean['InvoiceDate'].dt.month
```

Step 6: Visualise Key Insights

6.1 Monthly Revenue Trend

```
import matplotlib.pyplot as plt
import seaborn as sns

monthly_sales = df_clean.groupby(['InvoiceYear', 'InvoiceMonth'])['TotalPrice'].sum().reset_index()

plt.figure(figsize=(10,6))
sns.lineplot(data=monthly_sales, x='InvoiceMonth', y='TotalPrice', hue='InvoiceYear')
plt.title('Monthly Revenue Trend')
```

```
plt.xlabel('Month')
plt.ylabel('Total Revenue')
plt.show()
```

6.2 Top 10 Products by Revenue

```
top_products = df_clean.groupby('Description')['TotalPrice'].sum().sort_values(ascending=False).head(10)

plt.figure(figsize=(10,6))
sns.barplot(x=top_products.values, y=top_products.index)
plt.title('Top 10 Products by Revenue')
plt.xlabel('Revenue')
plt.ylabel('Product')
plt.show()
```

6.3 Sales by Country

```
country_sales = df_clean.groupby('Country')['TotalPrice'].sum().sort_values(ascending=False).head(10)

plt.figure(figsize=(10,6))
sns.barplot(x=country_sales.values, y=country_sales.index)
plt.title('Top 10 Countries by Sales')
plt.xlabel('Revenue')
plt.ylabel('Country')
plt.show()
```

Step 7: Reflect and Report

Ask students to answer:

- What patterns did you observe in monthly sales?
- Which products and countries generate the most revenue?

- How did data cleaning affect your results?

Encourage students to save their notebook and prepare a short summary slide or infographic of their findings.

Learning Outcomes

By completing this lab, students will:

- Understand the structure and quality of retail data
- Apply data cleaning and feature engineering techniques
- Use Python to perform EDA and generate visual insights
- Communicate findings clearly and effectively