**E-commerce Sales Analysis Project Notes**

**Project Overview**

I’m working on an E-commerce Sales Analysis project to dig into sales data, explore trends, and segment customers. The goal is to understand key aspects of the business, like which countries and products are performing the best, and how different customer segments behave.

**1. Data Cleaning and Preparation**

**Goal:** Start with cleaning the dataset to make sure it’s ready for analysis.

**Steps:**

* **Loaded the Dataset:** First, I loaded the dataset into Python using Pandas.
* **Handling Missing Values:**
  + The Description and Customer ID columns had some missing values.
  + Decided to fill missing descriptions with "No description" and set missing customer IDs to 0 for anonymous customers.
* **Data Type Conversion:**
  + Converted the InvoiceDate column to a datetime format.
  + Ran into a DtypeWarning because of mixed types in the Invoice column, so I converted the whole column to a string to keep it consistent.
* **Negative Values:**
  + Found some negative values in the Quantity and Price columns, which seemed to be returns or errors.
  + Decided to drop those rows to keep the sales data accurate.
* **Saved the Cleaned Data:** Saved everything as fully\_cleaned\_online\_retail\_II.csv to use in the next steps.

**Challenges:**

* **DtypeWarning in Invoice Column:** Fixed it by converting the entire column to a string.
* **Chained Assignment Warning:** Got this when filling missing values. Changed the approach to avoid chained assignments.

**2. Sales Distribution by Country**

**Goal:** Figure out which countries are driving the most sales.

**Steps:**

* **Calculate Sales by Country:**
  + Calculated total sales by multiplying Quantity by Price for each transaction.
  + Grouped the data by Country to see which countries had the highest sales.
* **Visualization:**
  + Created a bar chart to visualize the top 10 countries by sales. This made it easy to see that the UK is the clear leader.

**Challenges:**

* **matplotlib Installation Issue:** Had trouble installing matplotlib in Visual Studio 2022. Fixed it by running the correct pip install command in Command Prompt and making sure Python was set up correctly.

**3. Product Popularity Analysis**

**Goal:** Identify the most popular products in terms of sales and quantity sold.

**Steps:**

* **Calculate Product Sales:**
  + Grouped the data by StockCode to calculate total sales and total quantity sold for each product.
* **Identify Top Products:**
  + Pulled out the top 10 products by total sales and quantity sold.
* **Visualization:**
  + Created bar charts to show which products were bringing in the most revenue and which were being sold the most.

**Challenges:**

* No major issues here. The calculations and visualizations worked smoothly.

**4. Customer Segmentation**

**Goal:** Segment customers to understand their behavior better.

**Steps:**

* **Calculate Customer Metrics:**
  + For each customer, calculated total spend, number of orders, and average order value.
* **Create Customer Segments:**
  + Used total spend to segment customers into Low-Value, Mid-Value, and High-Value groups.
* **Visualization:**
  + Made a bar chart to visualize how many customers fall into each segment. This helps in identifying who the most valuable customers are.

**Challenges:**

* This step went smoothly, no issues to report.

**Reflection and Next Steps**

This project has been a great exercise in exploring and analyzing e-commerce data. I’ve cleaned up the data, analyzed key aspects like country sales and product popularity, and segmented customers to better understand their value to the business.

**Next Steps:**

1. **Time Series Analysis:** Look at how sales have trended over time to see if there are any patterns or seasonality.
2. **Feature Engineering:** Start creating new features that could be useful for predictive modeling.
3. **Predictive Modeling:** Build models to forecast future sales, predict which customers might churn, or recommend products to customers.
4. **Dashboard Creation:** Use Power BI or Excel to build dashboards that present these insights in a clear, interactive way.
5. **Documenting and Sharing:** Wrap everything up with a report or blog post to share on GitHub or LinkedIn, and make sure it’s all set for my portfolio.

So far, I’ve tackled a few challenges with data types and package installations, but those were good learning experiences. Onward to the next phase!