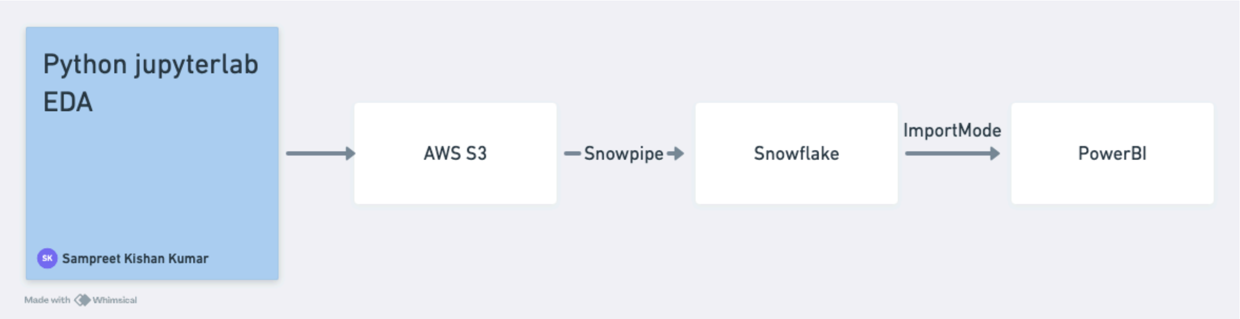
PEI data analyst assessment work

Submitted by: Sampreet Kishan
Modified: Aug 7th, 2024
sampy.prithvi@gmail.com

Task in hand:
Link to the question

Proposed Solution:



EDA:

- Performed simple exploratory data analysis (EDA) on a JupyterLab instance to assess data quality. Verified the absence of null values in all three datasets.
- Ensured unique `Customer_ID` values in the customers table for use as a primary key referencing foreign keys in shipping and orders tables.
- Confirmed 100% match for all foreign keys.
- Validated numerical values within acceptable ranges for all tables.
- Identified a discrepancy between the shipping and orders tables: shipments without corresponding orders.
- For further details, please refer to the Jupyter notebook in the GitHub repository.

AWS S3 (input data source):

Uploaded files to an S3 bucket. Converted the Excel-formatted customers file to CSV using a Lambda function triggered by S3 notifications, as Snowflake does not support Excel. **Note:** A custom layer containing pandas and xlrd was added to the Lambda function due to their absence in the standard environment.

Snowflake data warehousing:

Created a database, schemas, virtual warehouse, and role for the project. Established three tables (customers, orders, shipping) with primary and foreign key relationships. Created stages and file formats to stage datasets on S3. Loaded staged datasets into Snowflake tables. Implemented Snowpipe to automate data ingestion from AWS S3 into Snowflake.

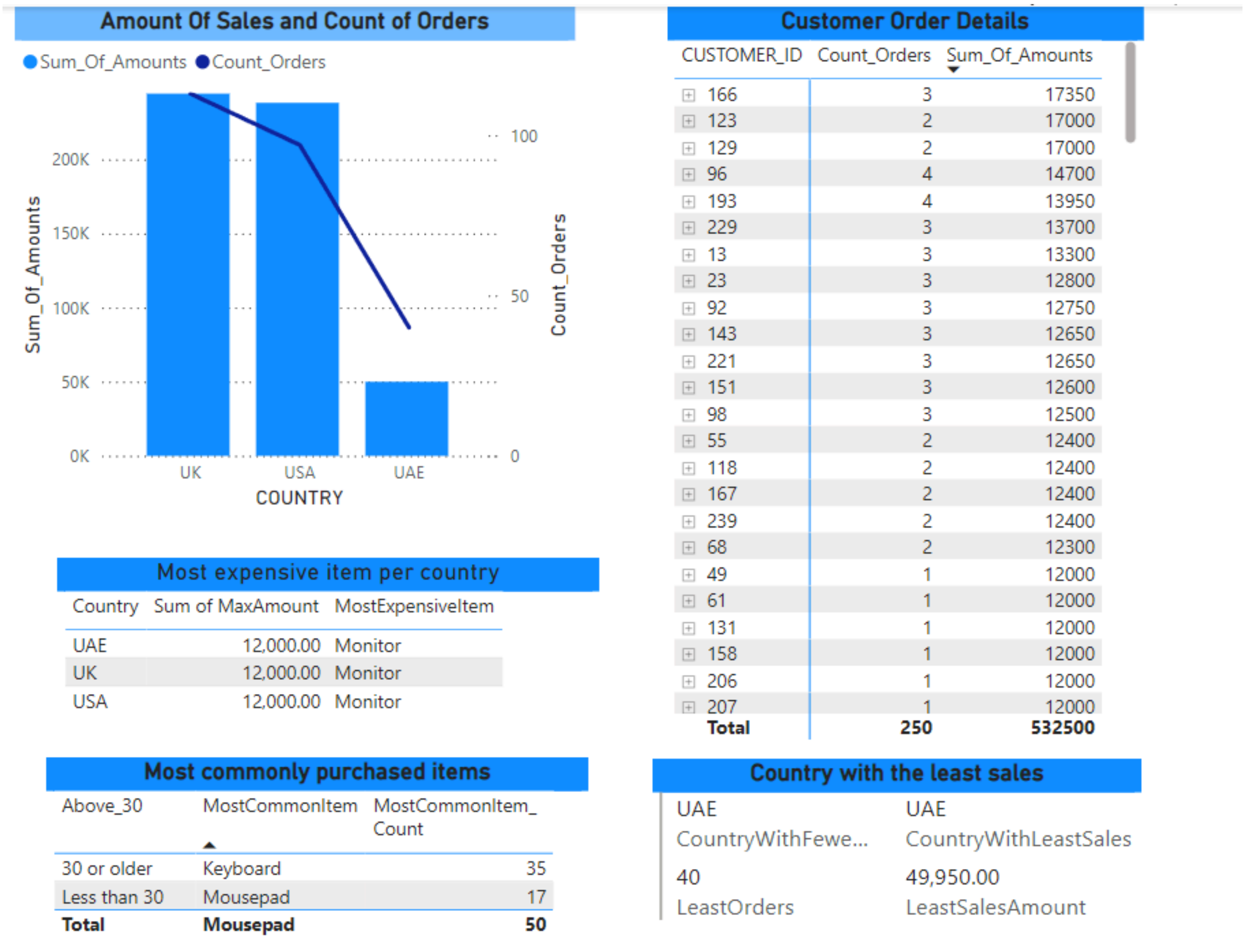
PowerBI (data modeling and data visualization):

Connected PowerBI to the Snowflake account. Performed minimal data type transformations using Power Query. Created a dimension table (customers) and fact tables (shipping, orders) with 1:many relationships in the Power BI model. Developed DAX queries to determine the most expensive items per country, most common item and count per age group, and countries with the least sales items and amounts.

Comments:

The upstream data engineering team should incorporate an order_ID (integer) column into the shipping JSON file. The data engineering team could potentially clean customer names to remove non-alphanumeric characters. Consider pushing data directly into Snowflake tables or continuing to load it into the S3 bucket for ingestion into Power BI using the existing ETL pipeline.

The report:



Amount Of Sales and Count of Orders

Customer Order Details

CUSTOMER_ID	Count_Orders	Sum_Of_Amounts
166	3	17350
123	2	17000
129	2	17000
96	4	14700
193	4	13950
229	3	13700
13	3	13300
23	3	12800
92	3	12750
143	3	12650
221	3	12650
151	3	12600
98	3	12500
55	2	12400
118	2	12400
167	2	12400
239	2	12400
68	2	12300
49	1	12000
61	1	12000
131	1	12000
158	1	12000
206	1	12000
207	1	12000
Total	**250**	**532500**

Most expensive item per country

Country	Sum of MaxAmount	MostExpensiveItem
UAE	12,000.00	Monitor
UK	12,000.00	Monitor
USA	12,000.00	Monitor

Most commonly purchased items

Above_30	MostCommonItem	MostCommonItem_Count
30 or older	Keyboard	35
Less than 30	Mousepad	17
Total	**Mousepad**	**50**

Country with the least sales

UAE CountryWithFewe...	UAE CountryWithLeastSales
40 LeastOrders	49,950.00 LeastSalesAmount

Key Findings:

- **Geographic Revenue:** The United Kingdom and the United States are the primary revenue drivers.
- **Product Performance:** Monitors are the highest-priced item across all countries. Mousepads are popular among younger demographics, while keyboards are preferred by older customers.
- **Customer Behavior:** Customers driving significant sales volumes tend to purchase multiple items.
- **Market Performance:** The United Arab Emirates exhibits the lowest order and sales figures.

Potential Action Points:

- **Market Expansion:** Prioritize market penetration in the UK and US to maximize revenue.

- **Product Strategy:** Explore opportunities to increase the average selling price of mousepads or introduce premium keyboard models.
- **Customer Segmentation:** Implement targeted marketing campaigns based on age demographics and purchase behavior.
- **Market Analysis:** Conduct a deeper analysis of the UAE market to identify potential growth opportunities or challenges.