Google Collab Link

https://colab.research.google.com/drive/1WgPFoY7gXJRsxFHh2On6aujaebc-ykp8?usp=sharing

Prompt

I have a dataset attached that requires transformation using Python. The transformation includes the following criteria:

- Identify top products based on the total transaction revenue per day.
- Detect any anomalies, such as a sharp decrease or increase in the number of transactions for a specific product.
- Identify the most profitable city or province based on the total transaction revenue.

Analysis and Improvement

ChatGPT has provided the results and transformation code based on the given prompt. I will conduct an analysis based on each of the provided criteria. Here is the explanation:

- Top Products by Revenue
    - Analysis
        - In performing data preparation, ChatGPT checks whether there are any null values in the totalTransactionRevenue and transactions columns. These two columns will be used for the transformation in the first criterion.
        - Aggregates the two previously mentioned columns using groupby based on the v2ProductName and productSKU columns. This ensures that the aggregation process is done based on the v2ProductName and productSKU columns.
        - Sorted from the largest to the smallest revenue.
        - Visualization using a bar char.

	v2ProductName	productSKU	total_revenue	total_transactions
0	Google Tote Bag	GGOEGBJC014399	1.862840e+10	47.0
1	Collapsible Shopping Bag	GGOEGBJC019999	1.233000e+10	47.0
2	Sport Bag	GGOEGBMJ013399	9.624530e+09	29.0
3	Google Lunch Bag	GGOEGBCR024399	8.855730e+09	56.0
4	Electronics Accessory Pouch	GGOEGBFC018799	8.597520e+09	47.0
5	Google Canvas Tote Natural/Navy	GGOEGBJL013999	7.532610e+09	52.0

Figure 1. ChatGPT Results for Top Products by Revenue

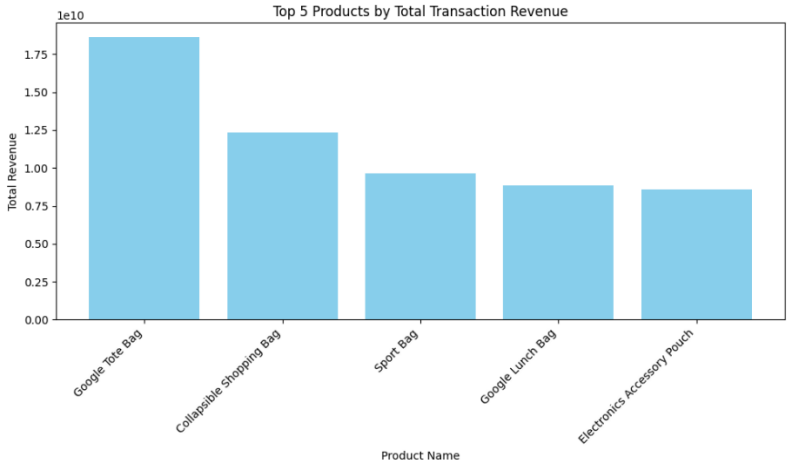

Figure 2. ChatGPT Visualization for Top Products by Revenue

- Improvement
  - Convert the data in the totalTransactionRevenue column to float and the transactions column to int. This is to ensure that the data is aggregated correctly since the source data is provided in CSV format. There is a possibility that the data is treated as a string.
  - Numbers representing total revenue are very difficult to read due to the large number of zeros and the lack of separators. Therefore, I will convert the values into millions and reflect this in the column name. I will also round the values as the client does not require highly detailed figures.
  - Replace null values in the totalTransactionRevenue and transactions columns with 0 instead of dropping them. Although in the given criteria, we are only performing summation, so dropping null values may not have a significant impact, there could be additional aggregation processes later, such as calculating averages. Dropping null values in such cases would certainly have an effect, and the results could become misleading.
  - Removed the aggregation process for the transaction column because it was not required in the given criteria.
  - Since the chart is static, I added numerical details to each bar in the bar chart. Additionally, I will add a comma separator for data details to enhance readability.

	v2ProductName	productSKU	total_revenue_millions	total_revenue_millions_formatted
0	Google Tote Bag	GGOEGBJC014399	18628	18,628
1	Collapsible Shopping Bag	GGOEGBJC019999	12330	12,330
2	Sport Bag	GGOEGBMJ013399	9625	9,625
3	Google Lunch Bag	GGOEGBCR024399	8856	8,856
4	Electronics Accessory Pouch	GGOEGBFC018799	8598	8,598

Figure 3. Improvements for Top Products by Revenue

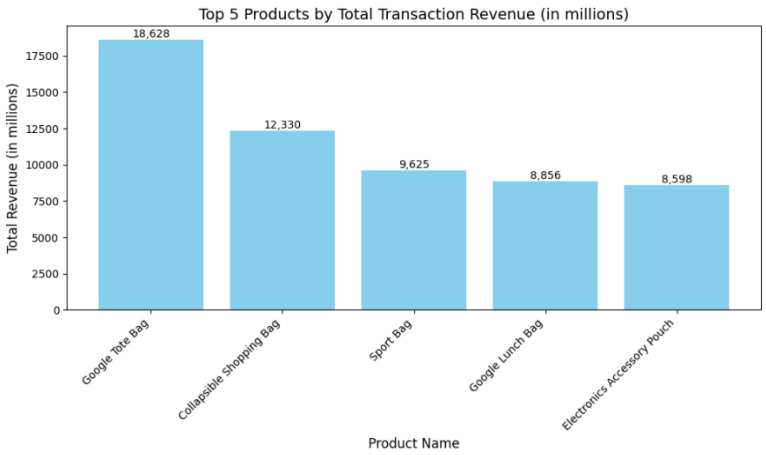

Figure 4. Visualization Improvements for Top Products by Revenue

- Anomalies in Product Transactions and Revenue
  - Analysis
    - Convert the data type of the date column to datetime. This change is made because the transformation requires further analysis on a daily basis.
    - Aggregates totalTransactionRevenue and transactions columns using groupby based on the productSKU and date columns.
    - Calculate the percentage change using pct_change() on totalTransactionRevenue and transactions_change_pct by applying groupby on the productSKU column.
    - Threshold for the percentage change categorized as anomalies was not specified in detail, so using 50% is acceptable.
    - Filter the data considered anomalous with the criteria of percentage change in revenue or transactions greater than 50 percent.
    - Create visualization using a line chart by combining all the data without considering the product type.

	productSKU	date	daily_revenue	daily_transactions	revenue_change_pct	transactions_change_pct
1	9180838	2017-07-27	24710000.0	1.0	-63.430517	0.0
10	GGOEAKDH019899	2016-09-20	64590000.0	1.0	-79.422728	0.0
11	GGOEAKDH019899	2016-09-21	16450000.0	1.0	-74.531661	0.0
12	GGOEAKDH019899	2016-09-22	478120000.0	2.0	2806.504559	100.0
13	GGOEAKDH019899	2016-09-24	128220000.0	1.0	-73.182465	-50.0

Figure 5. ChatGPT Results for Anomalies in Product Transactions and Revenue

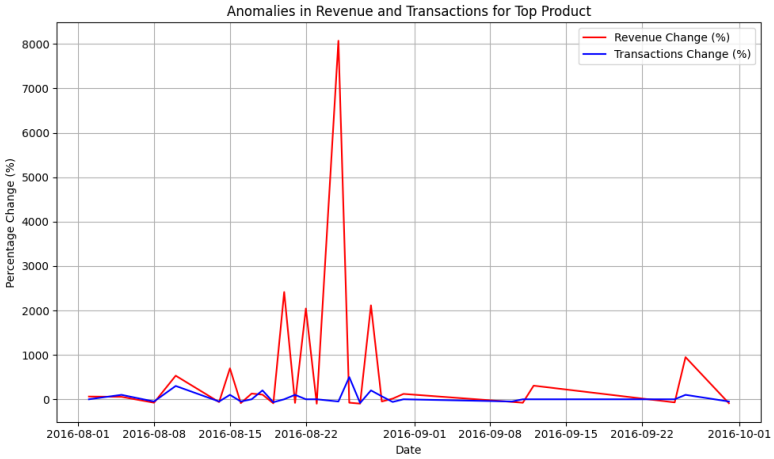

Figure 6. ChatGPT Visualization for Anomalies in Product Transactions and Revenue

- Improvement
  - Assuming that the transformation process for each criterion is independent, when calculating the percentage change, there is a condition that can result in an infinite value. This occurs when the revenue from the previous day is 0 or null, and today's value is, for example, 100. In such a case, the percentage change would be infinite. Therefore, data with a revenue of 0 or null will be dropped, as it holds no meaningful value.
  - Added a new column during the groupby process because one productSKU can have more than one product name. Since the given criteria require analysis on specific products, this adjustment was necessary.
  - Limited the number of decimal places to just 2 because having too many can make the data difficult to read, and the client does not require highly detailed figures.
  - Based on the given criteria, the client wants to see the percentage change in revenue for each product. Therefore, the chart for each product should be separated.

Anomalies in Product Revenue:

	v2ProductName	productSKU	date	daily_revenue	revenue_change_pct
1	1 oz Hand Sanitizer	GGOEGCKQ013199	2016-09-16	7.619400e+08	-66.27
2	1 oz Hand Sanitizer	GGOEGCKQ013199	2016-09-17	3.498000e+07	-95.41
5	1 oz Hand Sanitizer	GGOEGCKQ013199	2016-09-24	1.168400e+08	192.17
6	1 oz Hand Sanitizer	GGOEGCKQ013199	2016-09-26	3.996400e+08	242.04
7	1 oz Hand Sanitizer	GGOEGCKQ013199	2016-10-02	1.524700e+08	-61.85

Figure 7. Improvement Results for Anomalies in Product Transactions and Revenue

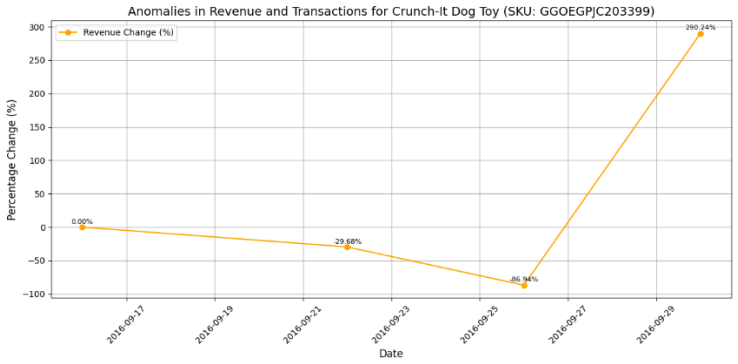

Figure 8. Visualization Improvement for Anomalies in Product Transactions and Revenue

- Most Profitable Cities
  - Analysis
    - ChatGPT only performed aggregation by summing the totalTransactionRevenue and transactions columns using groupby based on the city column. Then, the data was sorted from largest to smallest.

Most Profitable Cities:

	city	total_revenue	total_transactions
0	not available in demo dataset	6.510244e+10	274.0
1	New York	2.125081e+10	117.0
2	San Francisco	7.774340e+09	42.0
3	Mountain View	5.713290e+09	52.0
4	Toronto	5.427980e+09	8.0

Figure 9. ChatGPT Result for Most Profitable Cities

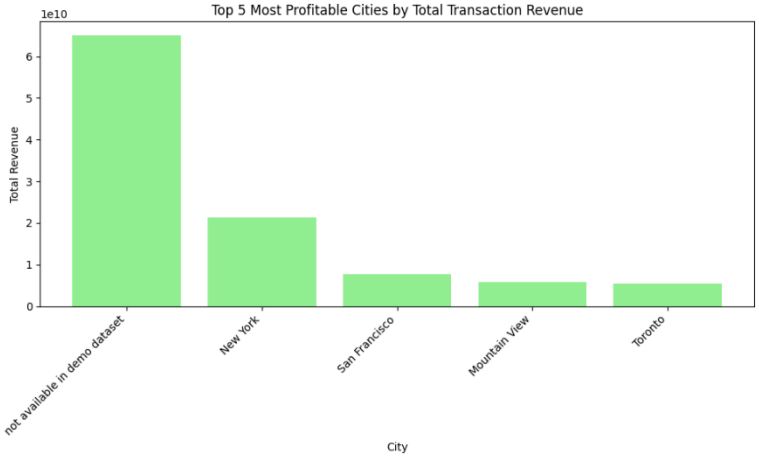

Figure 10. ChatGPT Visualization for Most Profitable Cities

- Improvement
  - Assuming that the transformation for each criterion is separate from one another, the filtering process was performed again by replacing null values with 0 and changing the data type to float.
  - Aggregation process for the transactions column was removed because only totalTransactionRevenue is needed.
  - Total revenue was converted into millions, and a separator was added to improve data readability. Additionally, the client does not require detailed figures.
  - Displaying the visualization using a bar chart and adding revenue details on top of the bars. Additionally, data with the city name 'not available in demo dataset' will not be displayed

	city	total_revenue_millions	total_revenue_millions_formatted
0	not available in demo dataset	65102	65,102
1	New York	21251	21,251
2	San Francisco	7774	7,774
3	Mountain View	5713	5,713
4	Toronto	5428	5,428

Figure 11. Improve Result for Most Profitable Cities

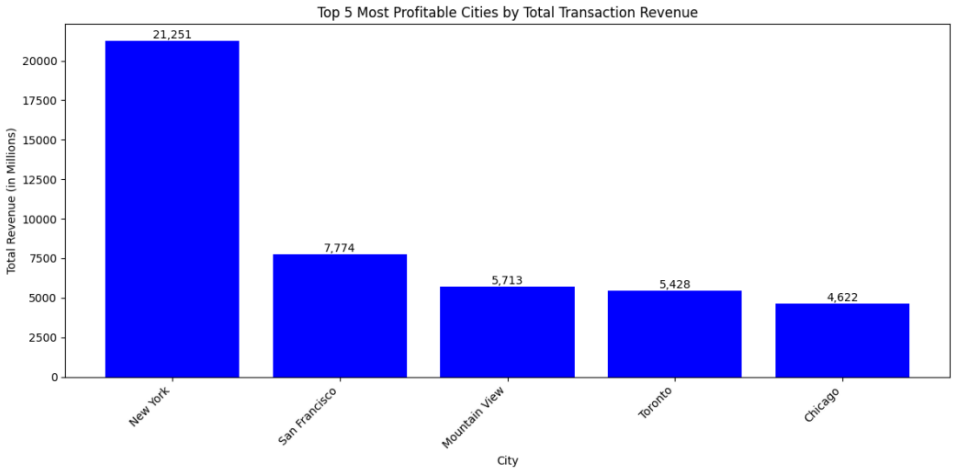

Figure 12. VIsualization Improve for Most Profitable Cities