THE UNIVERSITY OF
WESTERN
AUSTRALIA

Group Assignment: Dashboards and e-Commerce Analysis
(INMT5526: Business Intelligence**)**

Group 35

Nikhil Sawant
Guillermo Celis Prada
Adharsh Sundaram Soudakar

Introduction

The following report lies on the context of sales of clothing items on Amazon India in 2022, using a data source the collection of files provided by Sharma (n.d). To approach a comprehensive data analysis workflow, the following problem statement has been outlined:

The Problem Statement

Amaya Kurtas PTY LTD is a well-established business that manufactures and sells kurtas (traditional Indian shirts) in Perth, Western Australia. The business is contemplating opening a subsidiary in India, but it is uncertain on which state it should be located, which sizes and colours it should focus on, and the appropriate price for its products. The subsidiary is intended to be a brick-and-mortar store but given the popularity of Amazon in India, the company is looking to trade their products in that e-Commerce platform as well. Considering the E-Commerce Sales Dataset (Sharma, n.d), Would it be possible to identify the top 3 states with the highest sales of kurtas, the most popular sizes, and colours, and suggest a price that Amaya Kurtas could potentially sell their products at?

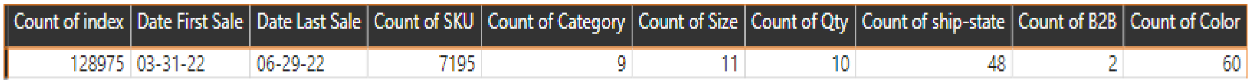
The purpose of this report is to provide an answer to the question above to perform a thorough *Business Intelligence workflow*, which entails Problem Identification, Data Acquisition, Data Preparation, Data Analysis, and Visualisation & Reporting (Reed, 2023, p. 6). This report is structured as follows: Summary of the data set, Analysis Undertaken, Results, and Recommendations and conclusions.

Summary of the dataset

This section of the report presents a summary of the Exploratory Data Analysis (**EDA**) performed to understand the data on hand. Out of the 7 files available in the E-Commerce Sales Dataset (Sharma, n.d), only the “Amazon Sale Report.csv” (**ASR**) and “Sale Report.csv” (**SR**) files were considered to address the problem statement.

Table 1 shows how many unique values of each relevant column were in the ASR, with only the *Color* column being obtained from the SR. It also indicates that the sales occurred between 31/3/2022 and 29/06/2022 (*Date* column in the ASR).

Table 1
Count of unique values of the relevant variables

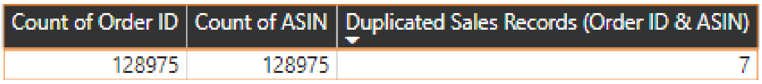


Count of index	Date First Sale	Date Last Sale	Count of SKU	Count of Category	Count of Size	Count of Qty	Count of ship-state	Count of B2B	Count of Color
128975	03-31-22	06-29-22	7195	9	11	10	48	2	60

The *index* column suggested there were 128,975 sales in the ASR, however, it was identified that a unique sales record is determined by the combination of two variables: *Order ID* and *ASIN* (Amazon Standard Identification Number). Using a DAX function in PowerBI, it was found that total of 7 records were duplicated sales within the ASR, as illustrated in Table 2.

Table 2

Total of duplicated sales records found

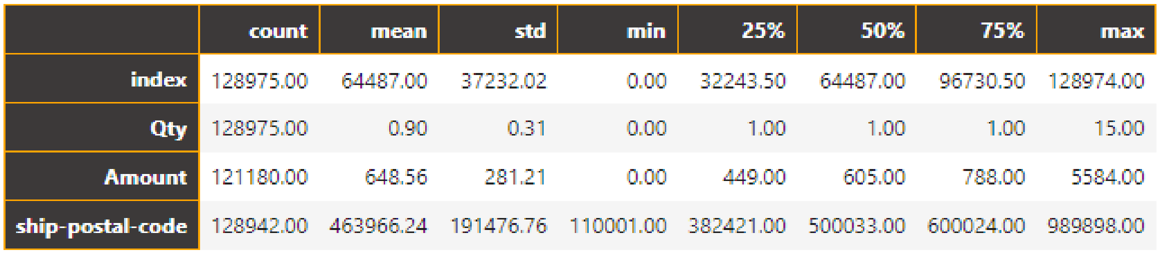


Count of Order ID	Count of ASIN	Duplicated Sales Records (Order ID & ASIN)
128975	128975	7

A descriptive statistics summary was created using Python and its Pandas library, to analyse the numeric variables, as can be seen in Table 3.

Table 3

Descriptive statistics summary on the numeric variables



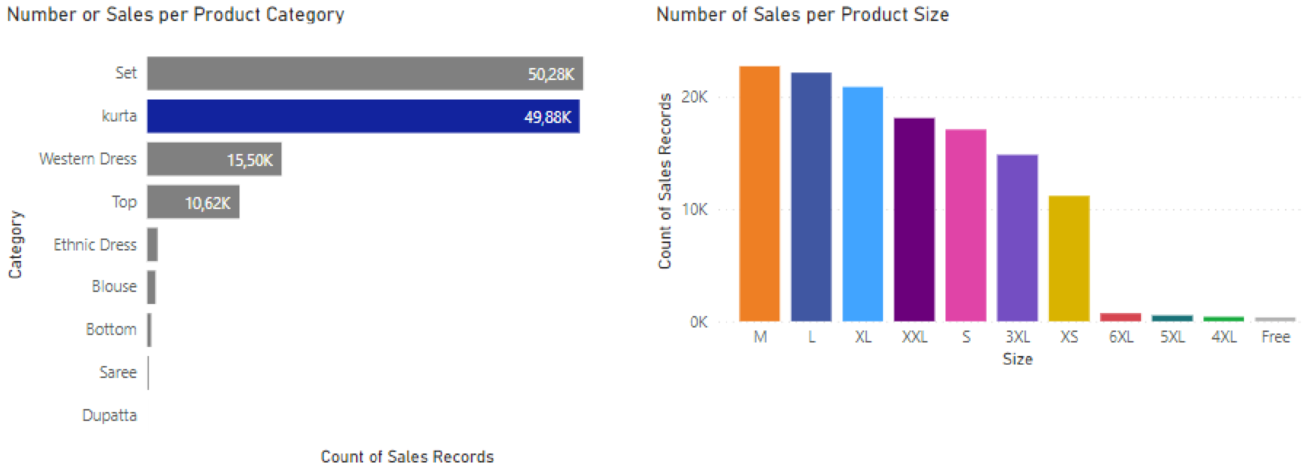
	count	mean	std	min	25%	50%	75%	max
index	128975.00	64487.00	37232.02	0.00	32243.50	64487.00	96730.50	128974.00
Qty	128975.00	0.90	0.31	0.00	1.00	1.00	1.00	15.00
Amount	121180.00	648.56	281.21	0.00	449.00	605.00	788.00	5584.00
ship-postal-code	128942.00	463966.24	191476.76	110001.00	382421.00	500033.00	600024.00	989898.00

This summary suggests that there were sales for 0 INR (*Amount* column), as well as orders with no quantities (*Qty* column) or with up to 15 items in the same order.

On the other hand, the 9 unique product categories suggested on Table 1 are depicted in detail in Figure 1. Kurtas are the second most popular clothing category in the ASR, with over 49800 sales recorded. Moreover, Figure 1 depicts the distribution of number of sales in terms of the product size.

Figure 1

Sales per product category and size



As a preview, Figure 2 shows the outcome of this report after all stages of the *Business Intelligence workflow* were completed. To form this dashboard, several types of visualizations were used, with appropriate titles, legends, and axis names.

Figure 2

Resulting Dashboard created in PowerBI

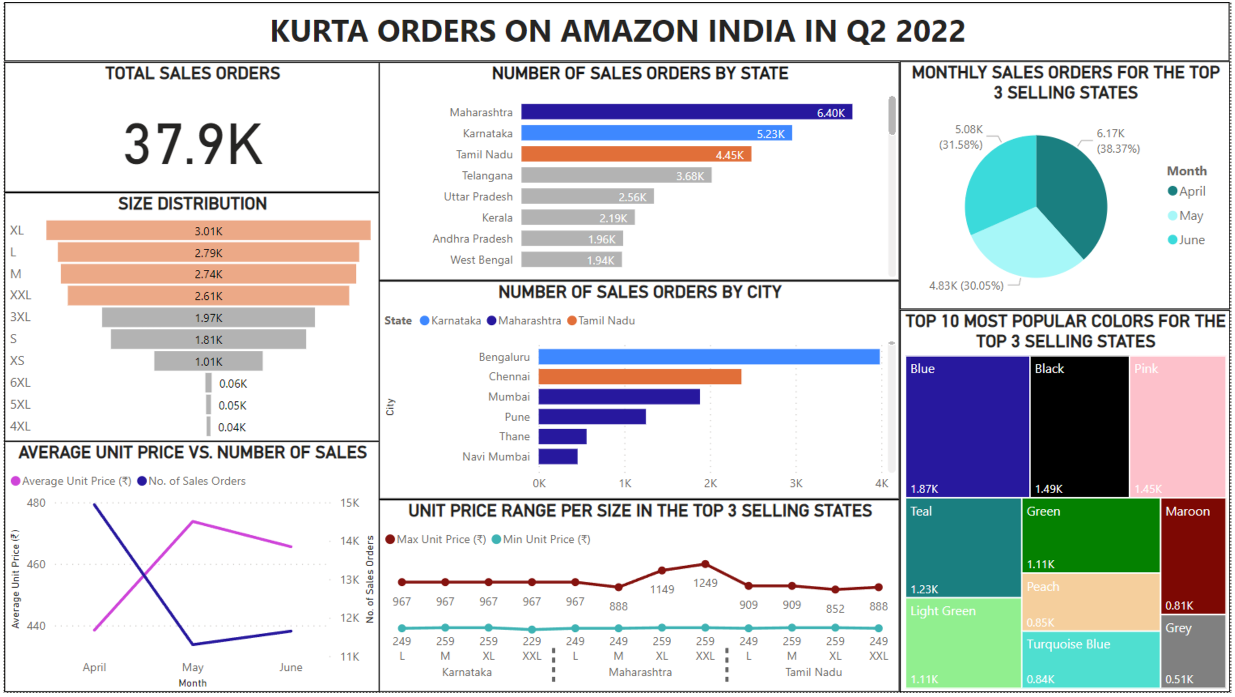


Table 4 provides a summary of the resulting dashboard, which consists of different visualisation used to analyse the data and find answers to the problem statement.

Table 4

Dashboard brief

Visuals used	Purpose
Card	The total number of sales orders, before or after applying a filter, is a valuable information that should be clearly shown. To accomplish this, a Card visualisation was used.
Stacked bar charts	Two sets of stacked bar charts were used to determine the top selling Indian states and their corresponding top selling cities.
Pie chart	The number and percentage of monthly sales orders for the top 3 selling states were represented using a pie chart.
Funnel chart	To analyse the kurta size distribution of the sales orders, a funnel chart was used.
Line charts	Line charts were used to determine the relationship between sizing and pricing of kurtas across the top selling states. Moreover, line charts were also used to establish a correlation between average unit price and numbers of sales order across different months.
Treemap	A treemap was used to discover the most popular colour of kurtas in the top selling states.

Analysis Undertaken

After the *Problem Identification* and *Data Acquisition* stages were completed, we proceeded with the *Data Preparation* step. The EDA allowed us to identify duplicated records (see Table 2), missing values, unnecessary information (columns) and data that needed to be normalised. In addition, some assumptions were made, as specified below.

Data Preparation stage

Assumptions and considerations:

- A unique sales record is determined by the combination of two variables: Order ID and ASIN (Amazon Standard Identification Number).
- The currency for all sales is INR (₹).
- The analysis considered only sales that took place during the second quarter of 2022.
- Sales orders of kurtas from the ASR that were not reported in the SR file (by SKU code) have been excluded from the analysis.
- The analysis excluded Sales that were made to businesses (*B2B* column = False), as the company is seeking to sell their kurtas to individuals only. Moreover, the price that vendors offer to businesses is usually lower, which can affect an accurate pricing estimation.
- The analysis excluded sales with a *Status* of “Cancelled”.
- Sales of products at zero price (*Amount*) or zero quantities (*Qty*) were also excluded from the analysis.

Deleted rows

- A total of 7 rows were deleted since they were duplicates based on their order number (*Order ID*) and product number (*ASIN*).
- A total of 171 rows with a date of the sale (*Date*) ‘03-31-22’ were deleted. Those were the only records outside of the second quarter of 2022, which was the time frame selected for the analysis.

Dropped Columns

- *index*: Irrelevant data.
- *Sales Channel*: Assumed to be Amazon.in for all records.
- *currency*: Assumed to be Indian Rupee (INR) for the few records it was not specified.
- *ship-country*: Assumed to be India (IN) for the few records it was not specified.
- *fulfilled-by*: There was only one shipping company mentioned. Unnecessary data.
- *Unnamed: 22*: unknown/irrelevant data.

Added Columns

- *UnitPrice:* Price per unit. Calculated column based on the quantities sold (*Qty*) and total value (*Amount*) columns.
- *Month:* Name of the month from the *Date* column, to facilitate analysis.

Normalised Data

- *Date*: It was converted from MM-dd-YY format to YYYY-MM-dd format.
- *B2B*: The data in the "business to business" column was converted from False to "Individual" and from True to "Business".
- *promotion-ids*: The empty values were replaced by "No promotion" and any other value replaced by "Promotion Applied".
- *Courier Status*: Missing values were filled with "Unknown" status.
- *Amount*: A total of 7795 rows had a null value, thus it was replaced by 0.
- *ship-state:* Table 1 shows there were 48 unique values found in the ASR, however there are only 28 states in India. Empty values were replaced by "Unknown" whilst state abbreviations and alternative names were normalised accordingly (e.g.: 'RAJASTHAN' , 'Rajshthan', and 'RJ' were converted to 'Rajasthan').

Data Analysis stage

Merging the Amazon Sales Report and the Sale Report files

We merged the ASR and SR files into PowerBI using the *SKU* (Stock Keeping Unit) column as the connecting variable. In the e-Commerce context, an SKU is a unique code assigned to a product to identify it and keep track of it (Decker, 2022). The connection between the two files was done to obtain the *Colour* of the kurtas from the SR, as it was an essential variable to answer the Problem Statement.

Finding the top three states with the highest number of sales

To find the top three states that had the greatest number of sales orders, we used the 'ship-state' column as the Y-axis and the count of Order IDs as the X-axis. As some Indian states have multiple large cities, it was relevant to know how much the sales in each state were spread across their corresponding cities, as this could led to high interstate delivery expenses for the business. Hence, we created an additional chart to determine the amount of sales order in each city of the top selling states. To facilitate the identification of cities for each state, we highlighted the top three states with a distinctive colour and matched that colour for each city name.

Sales distribution across the quarter

Since the analysis only considers the sales in the second quarter of 2022, this distribution had to be the count of different months in the 'Date' field. Using a pie chart, the distribution was visualized.

Size distribution

Using a funnel chart with 'Size' field as category and count of 'Order ID' field as values, we found the most popular sizes of Kurtas ordered. As the problem statement concentrates only on the most popular sizes, the analysis was restricted to the top four sizes.

Correlation between unit price of kurtas and number of sales

We wanted to determine if there was a positive or negative correlation between sales and the price of kurtas. This analysis was done to provide a recommendation to Amaya Kurtas in terms of how they should manage the pricing of their products. A double line chart graph was used, to represent the Average Unit Price and Number of Sales Orders (in the Y-Axis) across the months in which the sales were made (X-Axis). Legends and distinctive colours were used to differentiate the lines.

Range of unit price by size across top 3 selling states

A part of our analysis, we considered appropriate to determine if there was a significant variation in pricing for varied sizes (i.e.: A bigger kurta might be more expensive as it requires more materials). We wanted to know the range of unit price (i.e., the maximum and minimum) for each popular size in all the top three selling states. We used a line chart again with Maximum and Minimum of the 'UnitPrice' field as the Y-axes and 'ship-state' field and 'Size' field as the X-axes.

Popular Kurta colours across the top 3 selling states

Using a treemap with 'Color' field as Category and count of 'Order ID' field as values, we were able understand the top 10 colours of the Kurta orders.

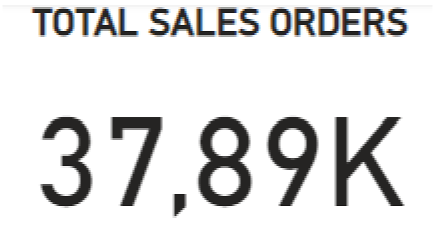
Results

This section of the report is focused on presenting the results obtained from the “Visualisation & Reporting” stage, which was completed in the form of a dashboard.

After applying to the datasets, the *Assumptions and considerations* mentioned in the previous section, we obtained a total of 37893 sales orders that meet our requirements (see Figure 3).

Figure 3

Total sales orders after applying Assumptions and considerations

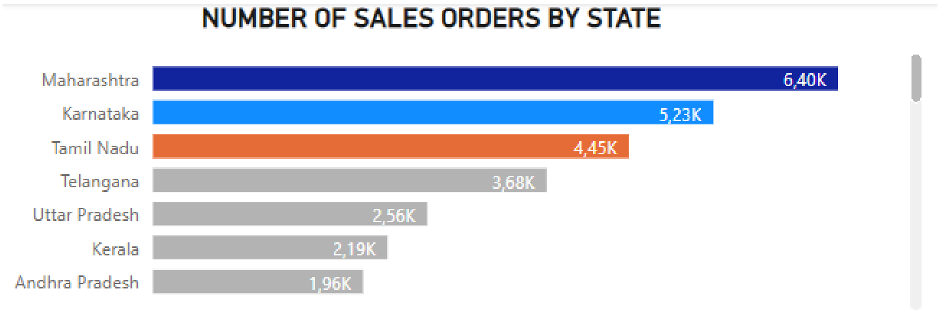


TOTAL SALES ORDERS

37,89K

The analysed sales orders were shipped from Maharashtra, Karnataka, and Tamil Nadu (see Figure 4); which combined made our new data sample, with 16074 sales orders. These 3 states made over 42% of the total sample data of 37893 sales orders.

Figure 4



It was also found that the city of Bengaluru (in Karnataka state) is the top selling city in India for the sample data. On the other hand, shipments in Maharashtra and Tamil Nadu are spread across four and two cities, respectively. Please refer to Figure 5.

Figure 5

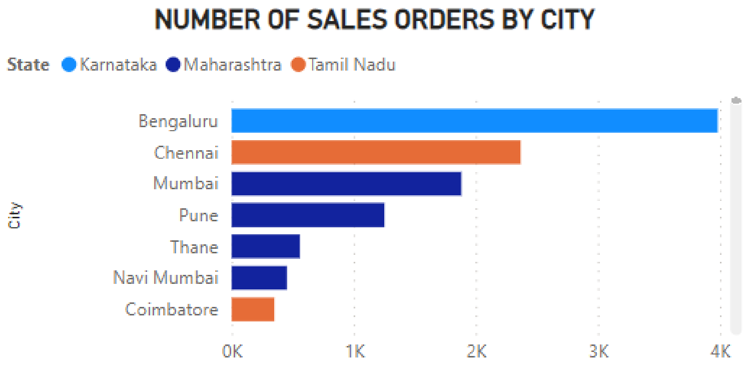
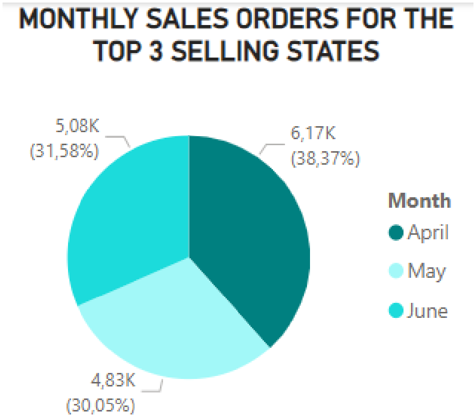


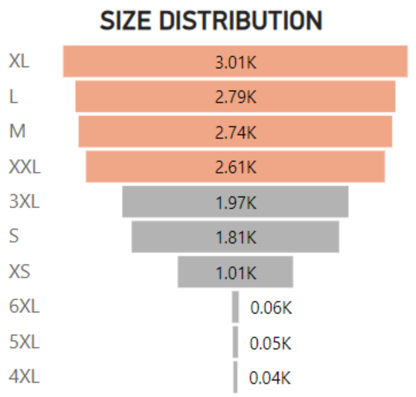
Figure 6 illustrates that the month of April saw the highest sales orders of kurta, with over 6000 orders, followed by June and May. Overall, similar number of sales orders occurred during each month in the top 3 selling states.

Figure 6



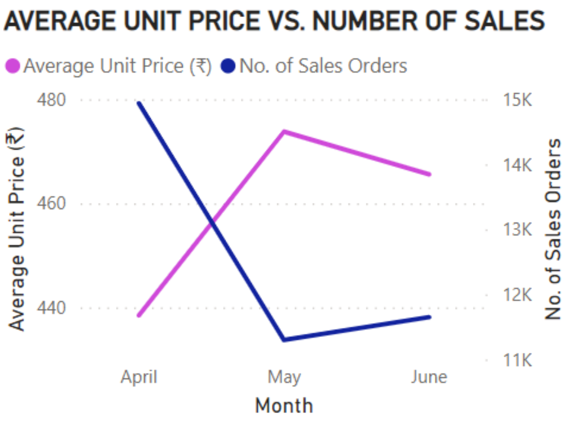
Most of the kurtas sold in the top 3 states were of sizes XL, L, M, and XXL; with over 2600 orders each (see figure 7).

Figure 7



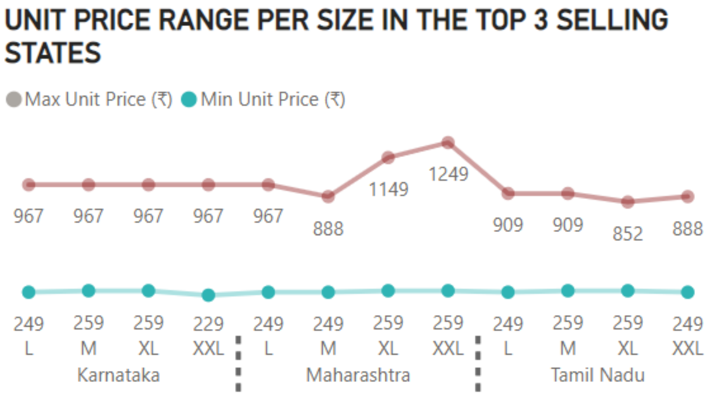
Moreover, it was found that there is inverse correlation between the price of kurtas and its number of sales. When the average price was low (e.g: April), the quantity demanded was high and vice versa (see Figure 8).

Figure 8



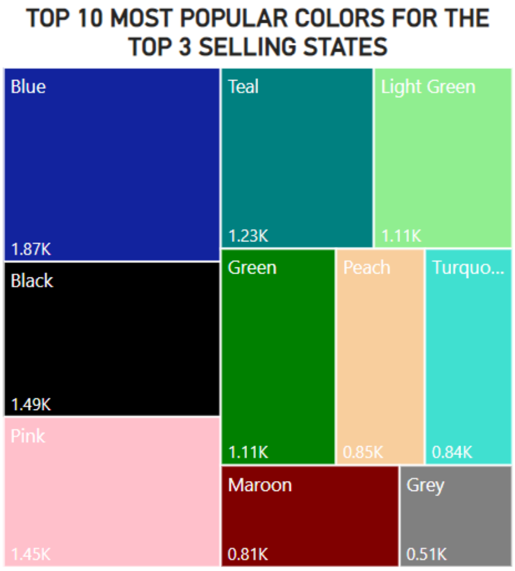
The unit price of a kurta varies in a range from 229 INR up to 1249 INR. Tamil Nadu had the lowest prices per kurta during the second quarter of 2022 in all sizes, except for XXL. Overall, the price per kurta seems to be irrespective of its size, although higher prices for bigger sizes were established in Maharashtra. (See Figure 9).

Figure 9



As can be seen in Figure 10, Blue is the most popular colour for kurtas in the top selling states, with over 1800 orders. Black, Pink, Teal, and Ligh Green were also in high demand, followed by Green, Peach, Turquoise, Maroon, and Grey.

Figure 10



Recommendations and conclusions

Based on the analysis undertaken and the results inferred from it,

- Amaya Kurtas PTY LTD should establish their subsidiary in Bengaluru, Karnataka initially. Although Maharashtra has the greatest number of orders, the orders are distributed across its cities which could end up with the company spending extra on delivering the orders. Whereas with Karnataka, Bengaluru by itself has more orders than three cities (Mumbai, Pune and Thane) of Maharashtra combined.

- As with the Kurtas, they should prioritise on getting sizes XL, L, M and XXL of the top 5 colours and start their sale during the month of April and considering the competition, reducing the price by 5% is recommended to enter the marketplace with a competitive edge. This is considering the variety in Kurtas such as size, type and quality of the fabric which alter the price drastically.

References

Decker, A. (2022, February 4). *What is a SKU Number and How to Use Them.* Shopify. https://www.shopify.com/retail/what-is-a-sku-number

Reed, Tristan W. (2023, October 2). *Week Ten: Online BI & Theory*. The University of Western Australia. https://lms.uwa.edu.au/bbcswebdav/pid-3430860-dt-content-rid-43835026_1/xid-43835026_1

Sharma, A. (n.d.). *E-Commerce Sales Dataset: Analyzing and Maximizing Online Business Performance.* [Data Set]. Kaggle. https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data/data