Data Visualisation Assignment 2

1. Introduction

The data sets chosen deals with online shopping data to discover insights on consumer behaviour and purchasing trends.

2. Problem

As a data analyst working for an online retail company, we are aiming to discover insights into consumer behaviour and purchasing patterns. Analysing customer preferences and trends is crucial to customize products, refine marketing strategies and enhance the overall customer experience and satisfaction.

3. Audience

There are multiple target audiences possible:

- E-commerce Businesses: CEOs of other businesses can use these visuals to gain insights on customer behaviour, popular products, and shopping trends.
- Marketing Teams: The visual analysis can help understand customers preferences and better tailor advertisements to them.
- Product Managers: These visuals can help identify the most popular products and create opportunities for new product development as well as understanding which product categories perform the best.

4. Datasets

Two different data sets were chosen. The first dataset Customer Shopping Trends Dataset, which can be found on Kaggle here https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trendsdataset , is a synthetic dataset of 3900 records created for data analysis purposes.

This dataset has the following columns: Customer ID, Age, Gender, Item Purchased, Category, Purchase Amount in USD, Location, Size, Colour, Season, Review Rating, Subscription Status, Shipping Type, Discount Applied, Promo Code Used, Previous Purchases, Payment Method, Frequency of Purchases.

The second dataset Online Shopping Dataset, which can be found on Kaggle here https://www.kaggle.com/datasets/jacksondivakarr/online-shopping-dataset , also contains information pertaining to customer purchases on an online retail website. This data set has almost 53000 records with similar columns to the previous dataset.

5. User Story

As a data analyst, our task is to visualize and identify the various purchasing trends and the highest selling products among customers. We can also explore how gender and age influence certain customer habits, to better tailor products and create targeted ads for customers.

6. Pre-processing and cleaning

The associated markdown file provides many useful comments and explanations in order to understand the steps taken for the following sections.

For the pre-processing task, we loaded both datasets into two data frames df1 and df2.

For the cleaning, we inspected the data types and transformed our categorial columns (read as chr) to factors.
Then we checked for missing values. Df1 contained no missing values, while Df2 had some. We checked the corresponding columns that had missing values and removed the 31 rows that had numerous columns with Nas. Then we looked at the remaining 400 rows with Nas in the Discount_pct column and decided not to drop those rows given that they only contained Nas in this column.
For Df2 we also dropped the X column as it was redundant and provided no extra information, as well as renamed all the columns to follow Df1's naming.

7. Wrangling

For the wrangling task, many SQL queries and their associated plots were made. The top 5 purchases by category were explored while grouping by gender and age groups. The most popular colours, the season of purchase, and the top colours per season were also inspected in regard to gender and age. The customer reviews were also examined, though there was very little variation for the different groups. The frequency of purchases was also explored varying by age and gender. Finally, the location of the customers was studied.

For more precisions on the wrangling part, the markdown file explains each query and visual and provides the associated insights and findings.

8. Visualisations (and previous iterations)

<u>Visual 1 - What are the top 10 most purchased items?</u>

It is important to know the bestselling products in order to market them to the maximum number of customers. It also allows to gage interest in the different categories of products to know which products are more likely to sell when designing new ones.

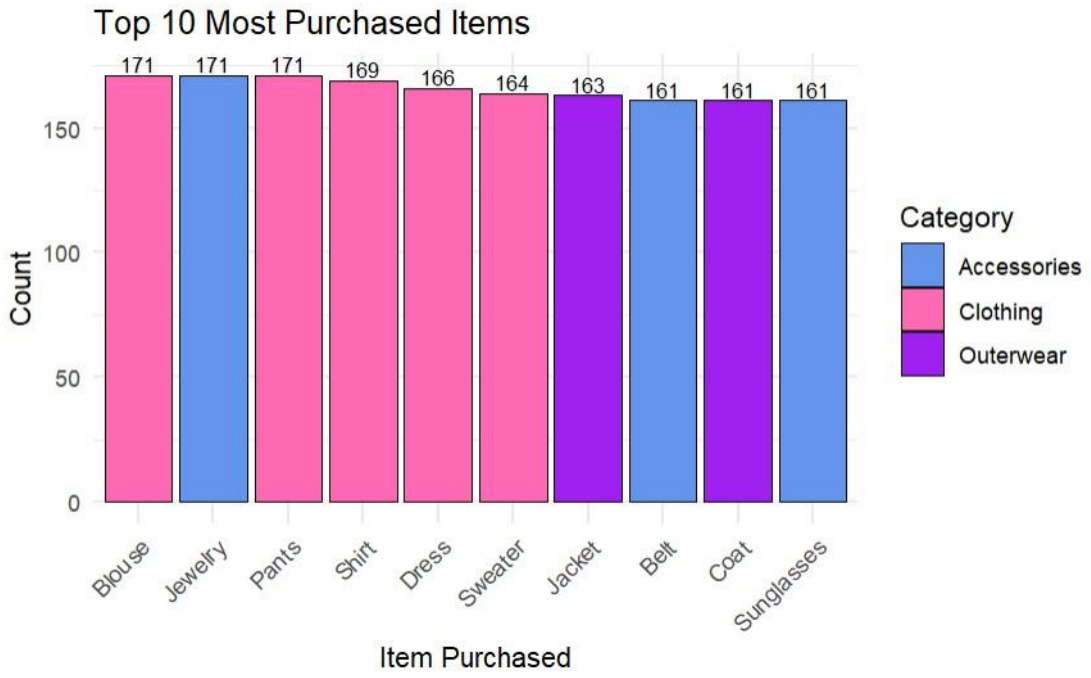

Figure 1: 1st Visual

Here we have a bar chart showing the top 10 most purchased items coloured by category (which also gives insights as to which categories perform better, here the clothing category makes up half of the top 10).

This visual had two previous iterations, which can be seen below:

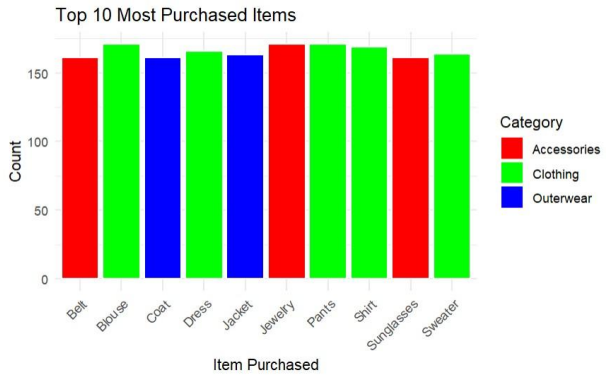

Figure 2: 1st iteration

Figure 3: 2nd iteration

As we can see, the colour theme was improved and the ordering of the items was changed to descending order for more visibility. In the final version, a thicker border was added to distinguish the bars better, as well as the count written on top of each bar.

Visual 2 – Where are the customers located?

Knowing where customers are located can help the business marketing strategy by using targeted advertising, promotions, and events for example that are more likely to resonate with local audiences. It can also be helpful in terms of supply chain optimization as warehouses and distribution centres can be located closer to customers ensuring efficient and timely delivery of products.

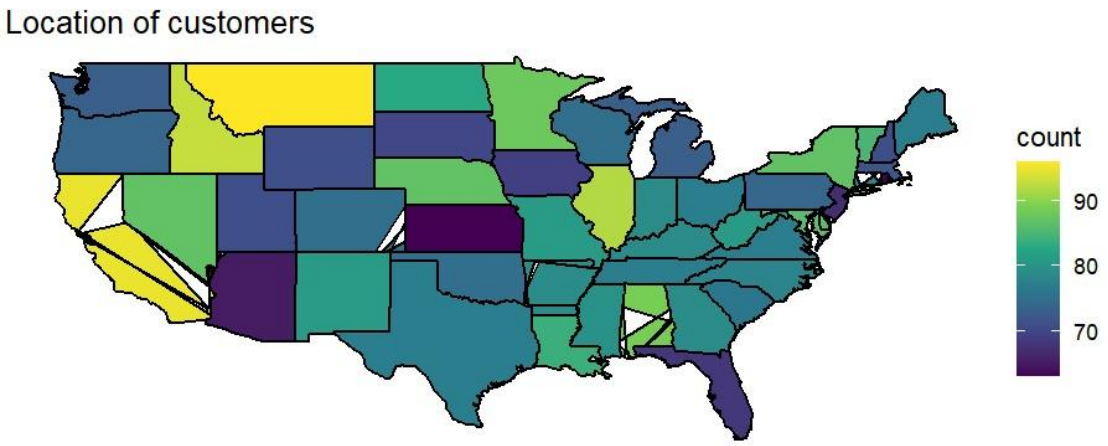

Figure 4: 2nd Visual

Here we have a heat map showing which states have the most customers, in our case it's California and Montana which have the most customers.

This visual had a previous iteration, which can be seen below:

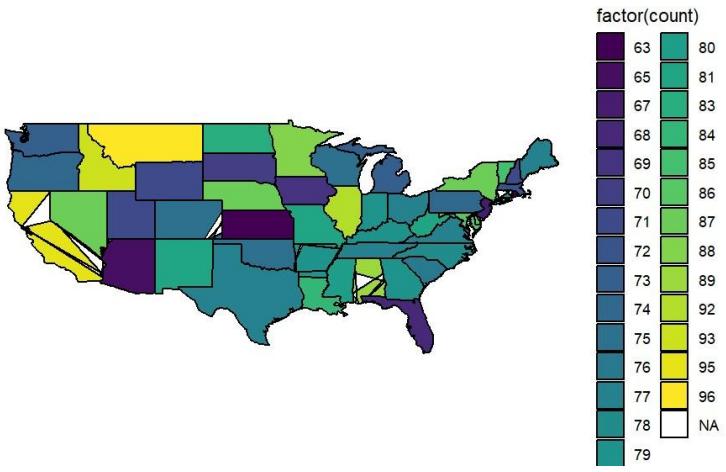

Figure 5: Original graph

The only change was to convert the count to a continuous variable instead of a factor to avoid the unnecessarily long legend.

A few other iterations were attempted; however they were not functioning properly. I tried to make the map interactive using the plotly library and I also tried to overlay bubbles varying in size for the most populated states.

Visual 3 – How frequently do customers purchase items on the website?

Customer traffic is crucial for online retail websites; it is very important to know which customers generate the most sales and how frequently they come back to buy items (the repeat rate).

Figure 6: 3rd Visual

Here we have a line chart showing how many customers purchase items at each frequency. We want to target customers who make weekly, bi-weekly or fortnightly purchases, as those are our main customer base. To increase sales, the retail company could give special sales or promo codes to entice them into buying even more.

This visual had three previous iterations, which can be seen below:

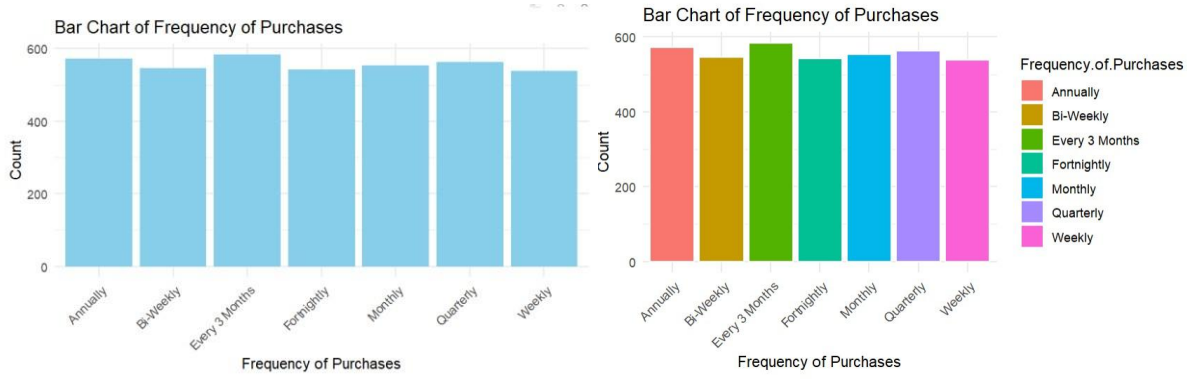

Figure 7: 1st iteration

Figure 8: 2nd iteration

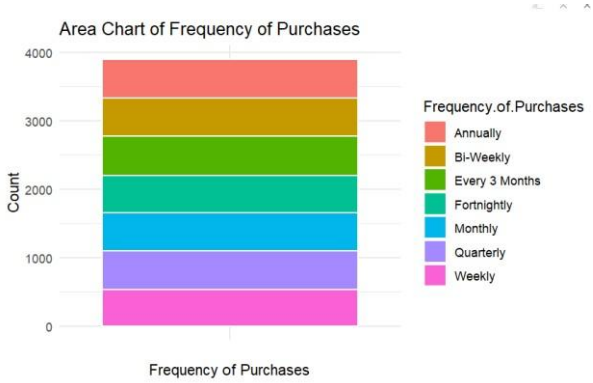

Figure 9: 3rd iteration

The original visual was a bar chart and then an area chart but given that the variation between categories of frequency of purchase was not visible enough, a line chart was chosen to illustrate that variation.