MIT805 Assignment 2

Part 2 – Visualisation report

Stephan Kersop – u16095953

30 October 2022

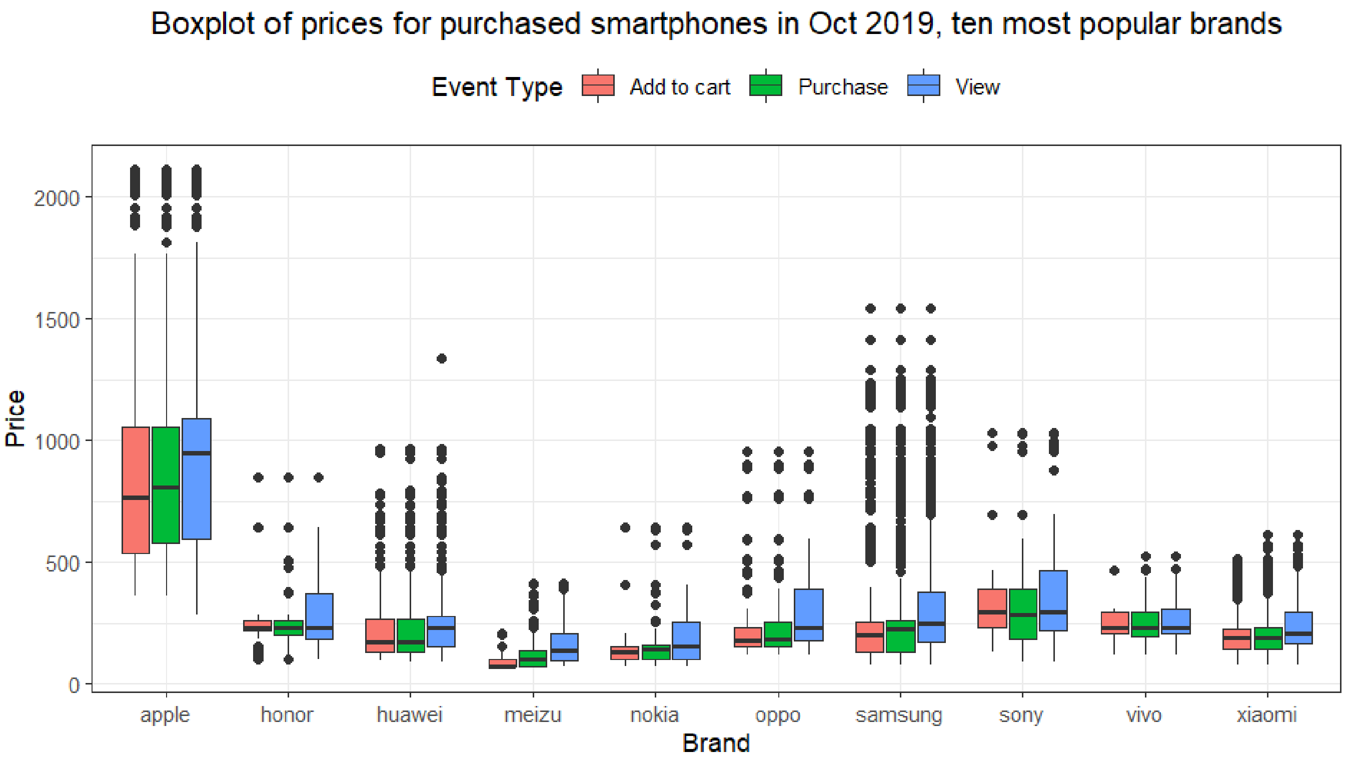
Introduction

For this project, I am working with the *eCommerce behaviour data from multi category store* dataset from kaggle.com [1]. This dataset contains event data from a multi-category online store for the months of October and November 2019. This is part 2 of a two-part assignment, which focuses on using visualisation to uncover hidden information. I will refer to results and discussion from part 1 of the assignment, and therefore recommend that part 1 is read before part 2. Visualisations were created using the *R ggplot2* package [2].

Visualisation and discussion

1) Purchased smartphones boxplot

I start by visualising the purchase prices of the top ten most popular smartphone brands in October 2019, using a boxplot.



This plot provides a more holistic comparison between brand pricing than our MapReduce results from part 1 by containing more contextual information. It gives a five-number summary, that is, it displays the minimum, maximum, median, first quartile, and third quartile for each group. Furthermore, it performs well as a tool to identify and evaluate outliers. In this instance, the groups constitute one brand each, with each brand further split between the different types of events: *view*, *purchase*, and *cart*.

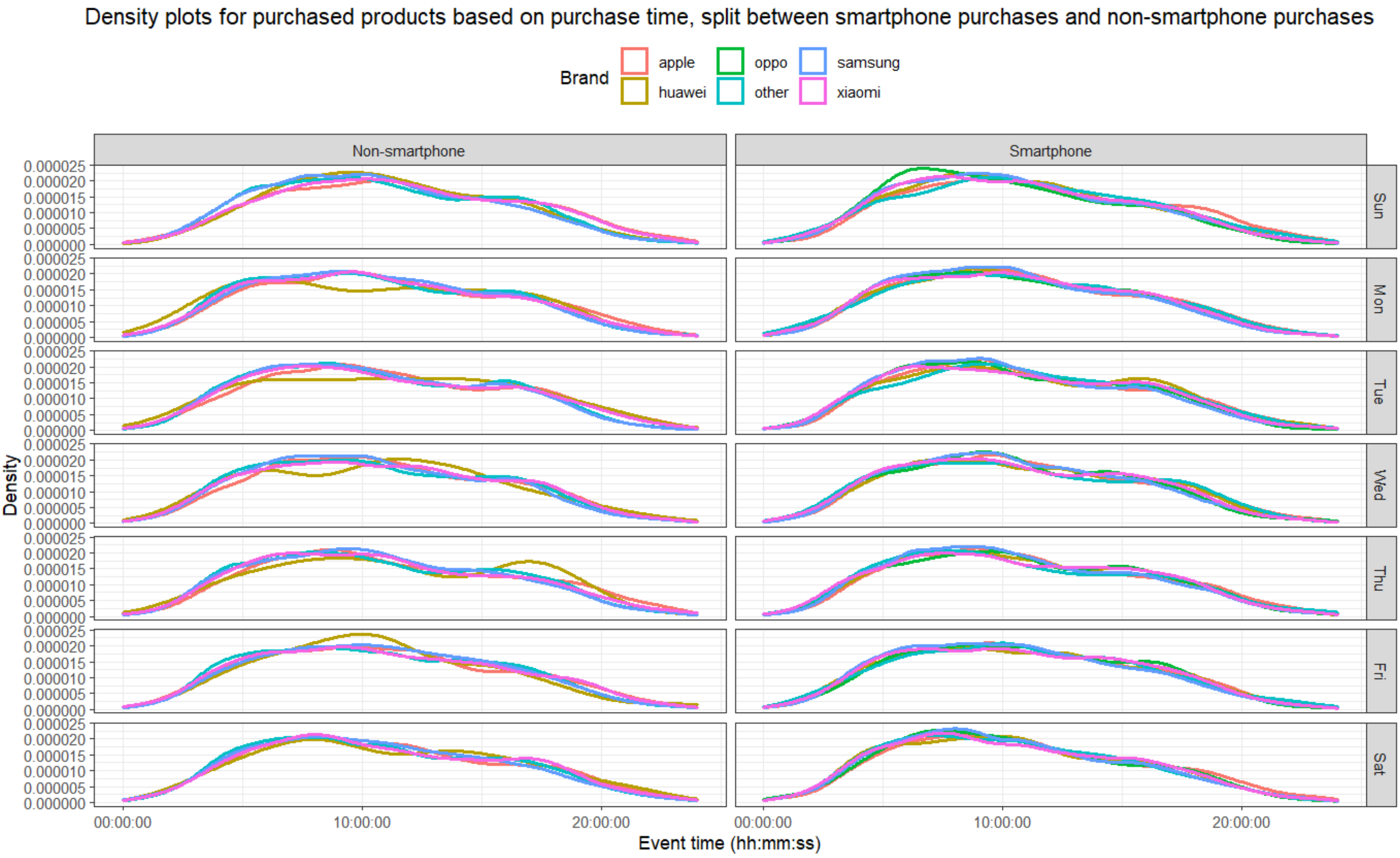
At first glance, we can see that purchased *apple* devices are expensive – the cheapest *apple* devices purchased are more expensive than the median device for any of the other popular brands. However, *apple* has the largest variance in prices, is the most popular *smartphone* brand in the store, and is the only brand to sell devices priced over 1 750. This suggests that the store should recommend *apple* smartphone devices to users who browse smartphone devices, as this brand offers the largest likelihood of selling a higher-priced device.

By examining outliers, we see that *apple*, *huawei*, *xiaomi*, and *samsung* sold comparatively large amounts of high-end devices. This suggests that these brands offer a wider price range of products and that users are willing to pay a premium for these devices, despite the popularity of these brands' lower-end devices. When users are browsing smartphones from these brands, offering within-brand flagship devices may increase sales of their high-end devices.

The plot furthermore shows that the prices of *view* events tend to be higher than the prices of *purchase* events, regardless of brand. Recommending higher-priced devices to users may create a competitive advantage by driving up the user's reference point for device pricing, leading to them purchasing a higher-priced product.

2) Event time density plots

For my second visualisation, I plot the distributions of smartphone and non-smartphone purchases across the time of the day, split by weekday. This provides an overview of when purchases peak for each of our major brands and allows us to identify differences in the timing of *smartphone* purchases and non-*smartphone* purchases. Furthermore, it allows us to evaluate whether peak times differ for different days of the week and helps to identify potential opportunities for cross-advertising between *smartphone* and non-*smartphone* products based on the time of day.



The plot suggests that regardless of day and brand, the largest proportion of purchase events occur between 05h00 and 10h00 UTC, and the second largest proportion of events between 10h00 and 15h00 UTC. Very few purchases are made between 22h00 and 04h00 UTC, which suggests that the store only operates within a small number of bordering time zones or within a single time zone and therefore has no (or very little) traffic during night-time. Based on this information, it may prove beneficial to push advertising and sales campaigns in the mornings to capitalise on the morning market.

The density functions for *smartphone* and non-*smartphone* products are very similar across both the day of the week and the brand. It therefore appears that there is little opportunity for utilisation of differing peak times to advertise categories that are not in their peak times.

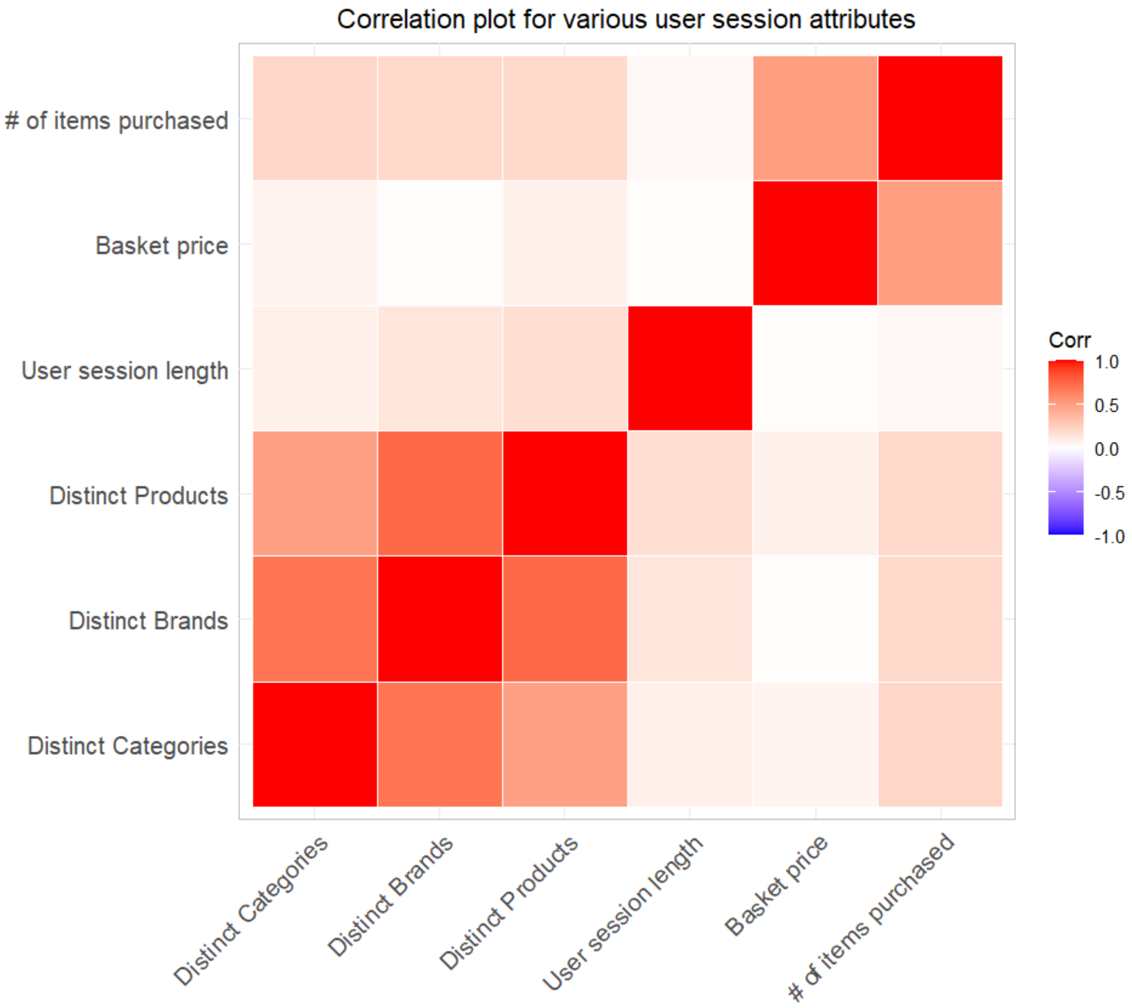
An interesting outlier is the brand *huawei*. A smaller proportion of *huawei*'s non-*smartphone* sales occur in the morning hours between Mondays and Thursdays than for other brands, with a larger proportion of their sales occurring in the afternoon compared to other brands. However, huawei's *smartphone* sales are distributed similarly to the rest of the brands in the store. This suggest that *huawei* may benefit comparatively more from users being shown *huawei* smartphones while browsing other *huawei* products, as they have a longer advertising window that does not overlap with the other major brands.

3) Session characteristic correlation plot

With my final visualisation, I investigate the correlation between various user session characteristics for sessions that resulted in one or more *smartphone purchase* events. This is done using a correlation plot. The characteristics investigated are:

- The number of distinct categories interacted with within the user session
- The number of distinct brands interacted with within the session
- The number of distinct products interacted with within the session
- The length of the user session (in minutes)
- The combined cost of all goods purchased in the session
- The number of items purchased in the session

The resulting graph is:



The characteristics regarding the number of distinct brands, categories, and products interacted with, all exhibit strong correlation with each other. This makes intuitive sense; one would not be able to browse different brands without, for example, browsing different products, or different categories without browsing different brands.

Contrary to my expectations, the user session length is not correlated with any of the other characteristics. This suggests that users spend just as much time investigating a single product, brand, or category as they spend investigating multiple of each. Additionally, session length being uncorrelated with basket price and the number of items purchased suggests that users access the store with a good idea of what they require and what they are willing to pay for it and spending longer on the store is unlikely to alter that idea. Finally, the number of items purchased is strongly correlated with the basket price, and weakly correlated with the characteristics that measure the variety of brands, categories, and products explored.

Conclusion

In this assignment, I used three different visualisations to try and discover new, useful information regarding smartphone sales from the *eCommerce behaviour data from multi category store* dataset. My first visualisation showed that, for smartphones, the product brand is a key factor in its popularity, and that brand popularity can offset high prices. My second visualisation showed that most purchases occur on the mornings and the distribution of purchases are consistent between brands and *smartphone* and non-*smartphone* categories. Finally, my final visualisation showed that the prices of purchased goods in *smartphone*-related user sessions are uncorrelated with the length of the user session and the variety of brands, categories, and products explored

References

[1] “eCommerce behavior data from multi category store.” https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store (accessed Sep. 08, 2022).

[2] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, 2nd ed. Verlag New York: Springer International Publishing, 2016. doi: 10.1007/978-3-319-24277-4.