Hi, everyone. My name is Helena Li, the data analyst in JR Academy. Today I would provide the analysis report in the E-commerce records.

The report has four contents including business purpose, sales data exploration, data clustering, and business recommendation.

First is the business purpose of data analysis. We have done lots of effort to do data preparation, transformation and data segmentation. The final goal of our work is to better understand customer’s behaviour, the past preference and based on his past behaviour to predict his behaviour. In this way we can targeted arrange our inventory and design marketing strategy.

Meanwhile, we can clearly know the hot products, popular departments, and aisles. Based on these evidence, we can adjust the products inventory and price. Through analysing the sales data, we can find the peak order time and ensure adequate inventory in the peak time. Last but not least, we can explore the pattern hidden in the data. For example, the customer’s category, customer features, and so on.

Let’s start our data exploration. The bar chart shows the number of products that people usually buy. We can see that number 6 has the largest purchase amount which means people usually buy 6 products in one order.

The second bar chart shows the purchase amount in different products. We can see that banana which is blue colour has the largest purchase amount. The second largest product is bag of organic banana. The third is organic strawberry.

Next, let’s have a look for the unpopular product. The product 0 calorie strawberry dragon fruit water beverage, 0% fat black cherry Greek yogurt, 0% fat strawberry Greek yogurt have the lowest sales which is only 1 respectively.

The right graph shows the information of aisles. We can see that fresh fruits which is orange colour has the largest purchase amount. The second ranking is fresh vegetables.

This bar chart shows the popular product when customer first add to cart. Bag of organic bananas and banana rank the first and second which are greatly larger that the third product strawberries.

The heatmap shows the peak order time. We can see that at 14 o’clock on Sunday customer order the most.

In terms of department distribution, we can see that the product department has the largest percentage which is blue colour. Produce means agricultural products, collectively, especially vegetables and fruits. Dairy eggs has the second largest percentage.

In terms of best-selling aisles over all departments, we can see that fresh fruits in produce department is on the top and fresh vegetables in produce ranks the second. The purchase amount of the top 1 and 2 are greatly larger that the packaged vegetables fruits which ranks the third.

After we have done the data exploration, we can explore more and do data clustering. Data clustering can segregate groups with similar traits and assign them into clusters. Meanwhile, it can also help understand the customer better so that we can market products better.

In data clustering, we mainly use the two methods: elbow method and silhouette coefficient analysis.

The “elbow” method helps to select the optimal number of clusters by fitting the model with a range of values for K. If the line chart resembles an arm, then the “elbow” (the point of inflection on the curve) is a good indication that the underlying model fits best at that point.

The value of the silhouette ranges between [1, -1], where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters. If most objects have a high value, then the clustering configuration is appropriate.

Compared with the two methods, we can have the clustering number is 4.

After we choose the cluster number is 4, we use Principal Component Analysis (PCA) to do further analysis. PCA can reduce the dimensionality of a multivariate data to two or three principal components. In this project, we choose two principal components. PCA also visualize data with minimal loss of information. In this care, PCA divides the customers into four categories. We can see the different colour in the graph.

This page shows the outcome of doing data clustering. The heatmap shows the features in each category. We can see that cluster 1 and 4 are purely composed of UK customers. We defined 1 as UK in data transformation.

The light colour shows the large amount of purchase sum. We can see cluster 2 has the largest purchase value sum and buys over all 5 products.

Cluster 3 has bought over all 5 products and has the largest purchase max compared with other clusters. Cluster 3 buys the most at one time. The majority of cluster customer are not from UK.

Cluster 4 only bought product 2 with only purchase twice while has a high purchase value sum. It means cluster 4 has bought expensive products.

Based on the above analysis, we provide some business recommendations.