Respected Data Science team leader,

Implemented exploratory data analysis (EDA) on the given ‘Gala Groceries’ dataset and gained these insights for optimizing the required business problem:

* By utilizing the seaborn library, we visualized the distribution of ‘unit\_price’ as positively skewed, which shows there are more sales of products with a low unit price compared to products with a high unit price.
* We found 4 unique values (1, 2, 3, and 4) for ‘quantity’ and they are quite evenly distributed. It’s graphical visualization shows that customers are buying in even quantities across 1 to 4 units.
* We have a total of 300 unique product IDs (or 300 unique products) within the dataset. Output of the helper function shows that product most frequently was sold within this dataset was ‘ecac012c-1dec-41d4-9ebd-56fb7166f6d9’, sold 114 times during the week. Whereas the product least sold was ‘ec0bb9b5-45e3-4de8-963d-e92aa91a201e’ sold just 3 times.
* ‘category’ feature has 22 unique values with ‘fruit’ and ‘vegetables’ being the 2 most frequently purchased product categories while ‘spices’ and ‘herbs’ being the least.
* There are 5 unique values for ‘customer\_type’ (non-member, standard, premium, basic & gold) and they seem to be evenly distributed. According to the bar-plot, ‘non-members’ appear to be the most frequent type of customers, closely followed by ‘standard’ and ‘premium’ customers with ‘gold’ members being the least.
* ‘payment\_type’ feature contains 4 unique values where ‘cash’ seems to be the most frequently used method of payment from this sample of data, with ‘debit cards’ being the least frequent.

Recommendations for the given business problem:

* The given dataset is only from one store and contains only a week worth of data, we require more data (rows) for better analysis of the dataset.
* We require more features (columns) to understand and evaluate the business problem that we are solving for.
* We need to frame the specific problem statement that we want to solve as the current business problem is too broad and we should narrow down the focus in order to deliver a valuable end product like a mobile application for searching available products using k-means clustering.

Best regards,

Mayur Gohane