* Cleaning/Assumptions
  + Filtered out transactions if 0 for either pack size or volume as these transactions we assume to be wrongly input into the system
  + Transactions if spending is 0, would be assumed as a valid good purchases since the assumption is that these goods were given as “free”/”promotional” items
  + Even if a customers bought biscuits (or any product) twice in a transaction, it is said that the person is buying biscuits of two different brands, rather than a duplicate in transaction history
  + We assume this data is from Malaysia, as there is a very high percentage of people of Malay ethnicity in our data, followed by Chinese. (*I think we don't have to assume this as it was alr stated in the slack grp?)*
  + We define monetary spend as the average amount spent per transaction as this will differentiate between high value customers and low value customers.
* Hopkin’s Statistic:
  + Just to prove that clustering our particular dataset would work
  + Not necessary to put in presentation
* K-means Clustering-> optimal number of clusters”
  + If k-means would tell us the optimal number of clusters is 3 or 4, that means on average, the customers can be segmented on 3-4 different clusters across all 3 categories.
  + As a result, the value that K-means would provide would lead us to score our RFM differently, if there should be on average 3 clusters → max score would be 333, if 4 clusters → max score would be 444
  + The main purpose of this, is not to decide the actual number of customer segmentations but more so to give us a rough gauge on how the RFM modelling should be carried out
* RFM modelling
  + Recency scoring is different from the lectures, as we realised that recency values tend to congregate amongst the lower numbers
  + Frequency scoring is based on the number of times a basket of goods were purchases (i.e. grouped by customer ID and date)
  + Monetary scoring is based as per normal: summing across all transactions
* Segmentation
  + There are 6 different segments mainly: Churned, At-risk (of churning) customers, Potential Loyalists, Low-Spending Loyalists, Loyalists, and Best Customers
* Features analysing
  + Churned customers will not be able to do much analysis at this point, but in future if we have time we can choose to look into at which point they churned, why they churned maybe, were they a “best” customer before they churned etc
  + Demographics:
    - Will compare our main target audience (which I predict would be the Potential Loyalists, Low-Spending Loyalists/Loyalists (whichever gives more interesting results or whichever has a significantly higher percentage), Best Customers across the different features → for instance if most of our Best customers tend to majorly come from the North, we should stock up on items that cater to their culture (??)
    - If not much meaningful insights can be drawn from the above, will go into each feature instead, and see the percentage of our target audience within each feature → for instance, if I go into the location, and the North seems to have a very high proportion of people coming from the Potential Loyalists group as compared to the other locations, we can provide some recommendation specific to the North side, to boost these customers up
    - In general, I would prefer the first one to give insights, as we can display it clearly on our presentation, categorised by each target group what our recommendations are based on their demographics
  + Categories of Food
    - For each customer, comparing their average spending on each category with the average price given in dataset “Data Category. If it is beyond average, it means the customer is willing to spend more for higher quality of products and that category is recorded under expensive in Category Table. Otherwise, recorded under column cheap. Premium products can be advertised to them to maximise revenue.
    - Detecting if there is a relationship between other features and category that customers are willing to spend more on. For example, High BMI customers spends more on Sugar. Then, find if any features gather in the same area, so the suggestion of recognising stock can be given to the store in that area.
    - (Extra) Try to fit a model (Desicion Tree) which can predict which type of customer that a new customer likely becomes.

* Impact
  + Compute the expected amount of savings/increase in revenue by keeping customers/ promoting customers; compute by using the average amount spent by the customer historically. Expected amount for the next quarter would be predicted based off the previous quarter for example