The data:

* Which data format to use for clustering?
  + Wide with a column for each individual as we want to cluster on customers?
* Variables to use in dataset
  + Go through data wide spreadsheet, look at variables and evaluate what needs to be done
  + Variable selection:
    - Scale data?
    - Does this variable give information about the individual for clustering?
    - Should we use email as all offers are email. If they are all email, then there is no difference to analyse for clustering
    - Variables that have the same value help with clustering?
    - NAs – check why NAs are in the data. Should they be zero or is there an error in code
    - New variables? Transactions when not in offer period
    - Update any error in code
  + Subset data:
    - Do we want to include people with 0 or 1 offers? Will this affect clustering?
    - Age/income extreme values? I didn’t see any, but it may be apparent in dimension reduction/clustering

Dimension Reduction

* Data to be scaled?
* Methods – what form is our data in?
* PCA – do we have sufficient continuous variables?
* PCA assumptions met?
* Correspondence analysis for count/categorical data – is it needed?

Clustering

* Data to be scaled
* Which method will be best giving data size, shape, etc
* What form is the data?
* Kmeans – continuous
* Hierarchical – continuous or binary
* KModes – categorical or binary (need to check exact rules)
* KPrototypes – mixture of categorical and continuous