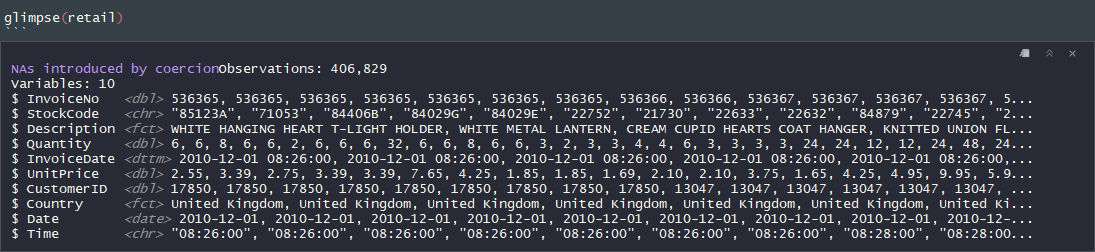
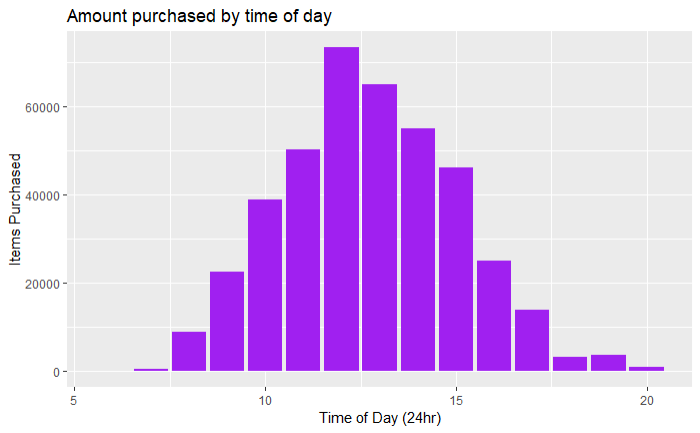
Part A.

1. The Online Retail dataset is a csv of transactional data from a British Online retail store. The dataset contains 10 variables with 406,829 observations. The variables include the invoice number, item stock code, item description, quantity of items purchased, date of invoice, price of the item, an id of the customer, country where the purchase was made, and date/time of the transaction.
2. Descriptive information of the variables –

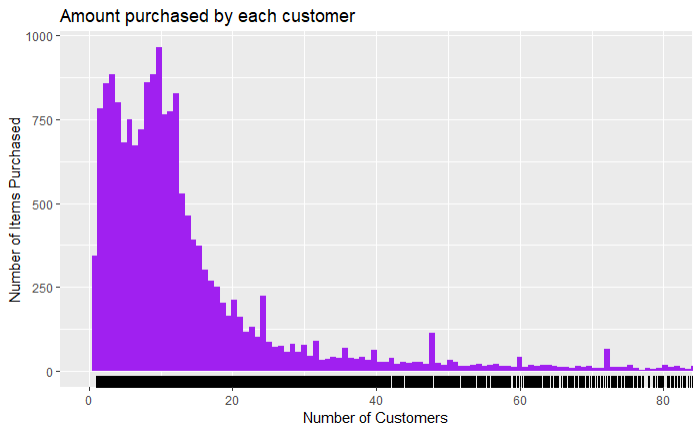


The InvoiceNo is a range of id numbers for transactions, containing 8905 NA values. These transactions are likely returns, purchases by the store, etc. The stock code is a unique identifier for each item. The Description lists the item itself, which includes “WHITE HANGING HEART T-LIGHT HOLDER”, “REGENCY CAKESTAND 3 TIER”, and “JUMBO BAG RED RETROSPOT”. The Quantity is a numerical value of the amount of items purchased, ranging between -80995.00 and 80995.00. The invoice date is a date time data type of the transaction, ranging from 2010-12-01 08:26:00 to 2011-12-09 12:50:00. The UnitPrice is a numerical value, ranging from 0.00 to 38970.00. The CustomerID is a numerical index for each customer associated to the transaction. The Country is a character value of the country where the transaction took place. Date is a date value of the transaction, ranging from 2010-12-01 to 2011-12-09. Time is a character value that contains the time of day the transaction took place.

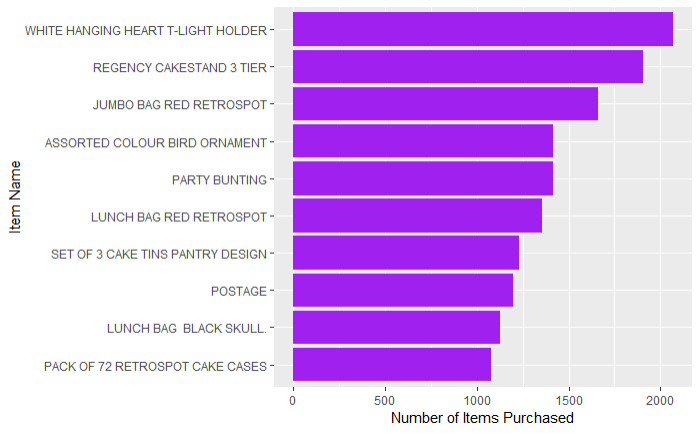
1. Check basic relationship between the variables –
   1. One of the relationships that I checked was the time of day that items were purchased. This would be useful for the company to know so they can purchase ads that would appear around this time of the day. In this case, most purchases took place around lunch time, so customers may be online shopping while on break at work.
   2. I then looked to see what time of the year most items were ordered. I did this by charting the number of items purchased each day. Based on this data, more items are purchased during the holiday season (between October and December). Marketing campaigns could be used here to stimulate increased sales.



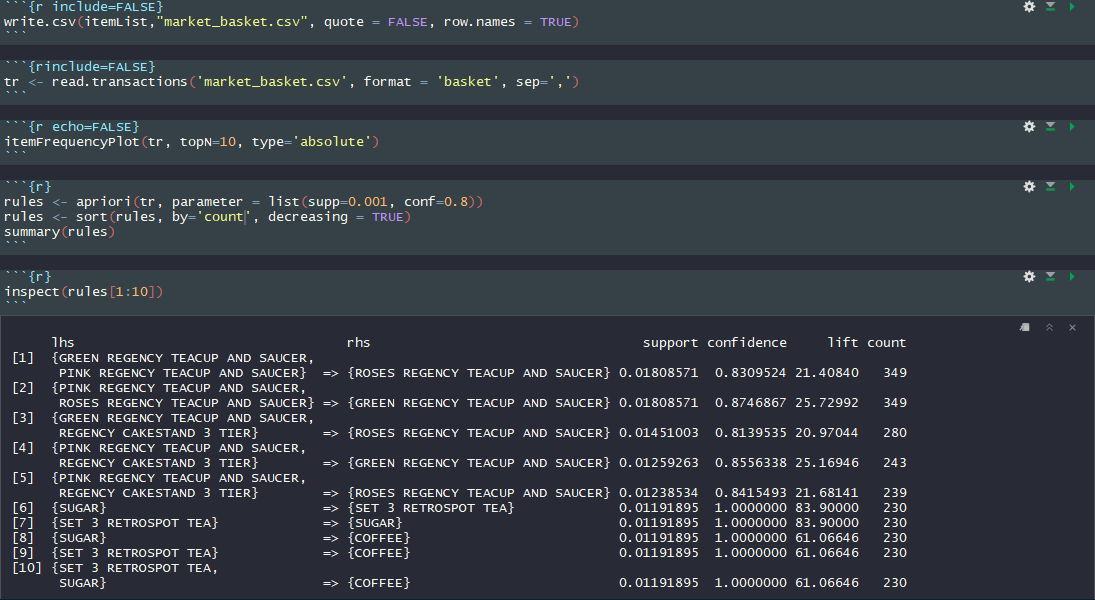
* 1. The next relationship I investigated was the number of items that customers purchased. Based on this, most customers purchased 2-12 items, with the number sliding off after. This would be an opportune area for an incentive program, such as free shipping on orders over $35 or containing more than 15 items.



* 1. I also calculated the top items purchased. Based on this, the most popular items sold appear to be holiday related, such as the WHITE HANGING HEART T-LIGHT HOLDER, a JUMBO BAG RED RETROSPOT, and ASSORTED COLOUR BIRD ORNAMENT. One recommendation here could be to have sales or discounts on bulk purchases of these items.

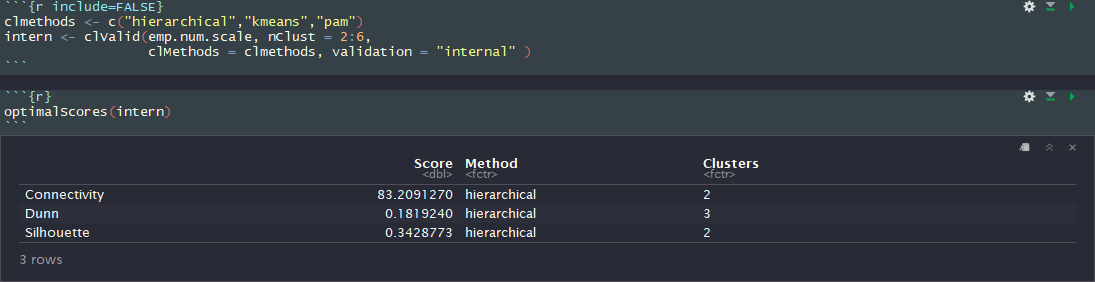


1. My analysis focused on using apriori to identify association rules between the products. Based on this, there appears to be a strong association between “SET 3 RETROSPOT TEA”, “SUGAR”, and “COFFEE”. There is also a high association with REGENCY TEACUP AND SAUCER sets. These all seem to be a very popular items, so advertising potential deals on these would increase sales. My recommendation would be to advertise a deal where purchasing a teacup and saucer includes tea, coffee, or sugar. Running these deals in October, November, and December would likely reach the most customers.

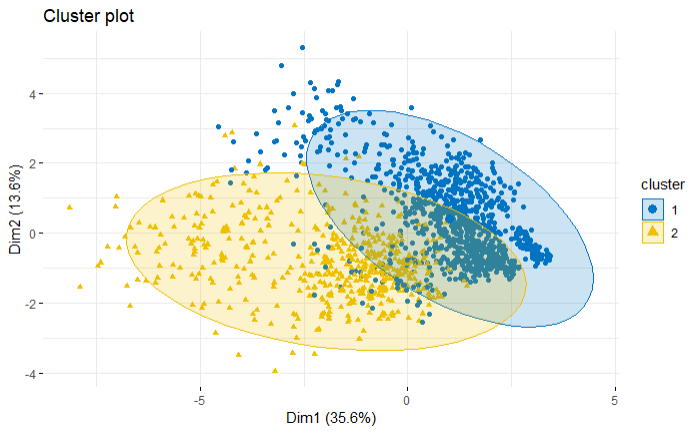


Part B.

* **Cluster size** – I used the clValid package to find the optimal configuration between several different clustering models with 2 – 6 clusters. I landed on the hierarchical method with 2 clusters based on the table below.



Cluster 1 contained 1018 employees and cluster 2 contained 452 employees.



* **Cluster Characteristics** – The defining characteristics in these clusters appear to be related to age and seniority with the company. Based on the code below, cluster 2 contains higher average age, monthly income, total working years, years at company, years in current role, years since last promotion, and years with current manager.

A screenshot of a cell phone

Description automatically generatedA screenshot of a computer screen

Description automatically generatedA close up of a screen

Description automatically generated

These clusters seem to have a loose association to the outcome variables: “Satisfaction on the insurance firm's customer service”, “Satisfaction on the insurance policy”, and “Will consider switch”. Based on the comparisons below, cluster two has a higher percentage of people who will not consider switching. However, the other who outcome variables seem to have very similar concentrations of each level.

A screenshot of a cell phone screen with text

Description automatically generated

* **Business Sense –** This makes business sense because the more tenured employees are likely more comfortable with their insurance plan as they can afford better, more inclusive insurance plans.
* **Surprising factors** – I found it surprising that there was little variation between the clusters in terms of satisfaction with the insurance plan policy and customer service.
* **Conclusions** – Based on the data collected, I would recommend that the insurance company focus on creating more affordable plans and incentives to keep younger, less tenured employee on their policy. One idea would be through gym membership discounts or yearly reduction in premiums the longer the employee keeps the plan.