National College of Ireland

Project Submission Sheet – 2019/2020

Student Name:	Antonio Caruso		
Student ID:	19203608		
Programme:	Higher Diploma in Science in Data Analytics	**Year:**	2020
Module:	Data & Web Mining		
Lecturer:	John Kelly		
Submission Due Date:	24/07/2020		
Project Title:	CA Part 1: Market Basket Analysis		
Word Count:	1021		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: ……

Date: 26/07/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).

2. Projects should be submitted to your Programme Coordinator.

3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.

4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**

5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

1 Table of Contents

Table of Figures

1. Market Basket Analysis

This report covers the principle of tidy data to a dataset titled “Groceries” in order to create a program in R that applies the association rules algorithm and the statistical outcomes of it. The principle of tidy data were extracted from Wickham (2014) in order to clean the dataset beforehand to ensure a correct function of the algorithm.

1.1 Description of the dataset

The groceries dataset contained initially 1499 observation and 35 variables named from V1 to V35. The first 34 columns from V1 to V35 were character data, whereas V35 was logical. The first column V1 presented combined dates and names of product, showing immediately to be the main variable of interest, whereas all the other variables presented sparse repeated names of products, blank spaces and NA. Wickham (2014), provided different methods for cleaning datasets, and since the dates would be related to the products, the main method to apply was to split the first column V1 into two separate columns , so that the lists of groceries would be related to the corresponding dates of purchase.

1.2 Application of tidy data techniques

For the data manipulation, different libraries were applied (Lantz, 2015) such as stringr for the functions and arguments to be consistent, reshape2 for using data tables, and plyr and dplyr respectively for implementing the split-apply combine pattern and manipulate the data.

After reading the file, the set.seed function was used to ensure the algorithm started at the same location in the dataset. An exploratory analysis was conducted to identify the number of rows, variables and structure of the data. The following step was to use the read.table function with a tab separator in order to read the file in a tab delimited table format and obtain a list of products associated with dates (Lantz, 2015)

The following step was to apply the separate function (Wickam, 2014) to create two separate variables named “date” and “product”. Since both the date and product new variables were both characters data types, they were cleaned and converted respectively into dates and factors into a new dataframe (Wickman,2014).

The following step involved the conversion of the cleaned dataframe into transaction. In Order to do that, a new variables named “transactionID” was added with a sequence of number corresponding to 1,499, the same number of observations and was converted into a factor (Chiu, et al.,2002).

The ddply function was used (Chiu, et al., 2002) to combine all products from the transactionID column create and the date as one row, with each item separated by a comma. Subsequently, the date and transactionID variables were set to Null as not relevant for the association rules mining. The column was renamed “Items”.

The new transactionData was then stored into a csv called groceries_market_basket_transaction.csv. Subsequently, the read.transaction function of the arules package was applied to take the transactionData file into a basket format and convert it into an object of the transaction class, and the new file was called “Transactions” (Chiu, et al., 2002).

The new “Transactions” file displayed 1500 rows corresponding to transactions as collections of the products, and 39 columns equalling to products. The outcomes showed a density of 0.37355, which tells the percentage of non zero cells in a sparse matrix. The density allowed to calculate how many items were purchased by using density as below (Chiu, et al., 2002);

- 1500 x 39 x 0.37355556 = 21,853 Items were purchased, as the total number of items that are purchased divided by a possible number of items in that matrix. Th most frequent items are vegetables (1089), poultry (613), waffles (587), dishwashing liquid/detergent (585), icecream (584) and others (18395). This could be visualised with the following plot obtained with the library RcolorBrewer (R Core team, 2013) that shows the relative frequency of items purchased.

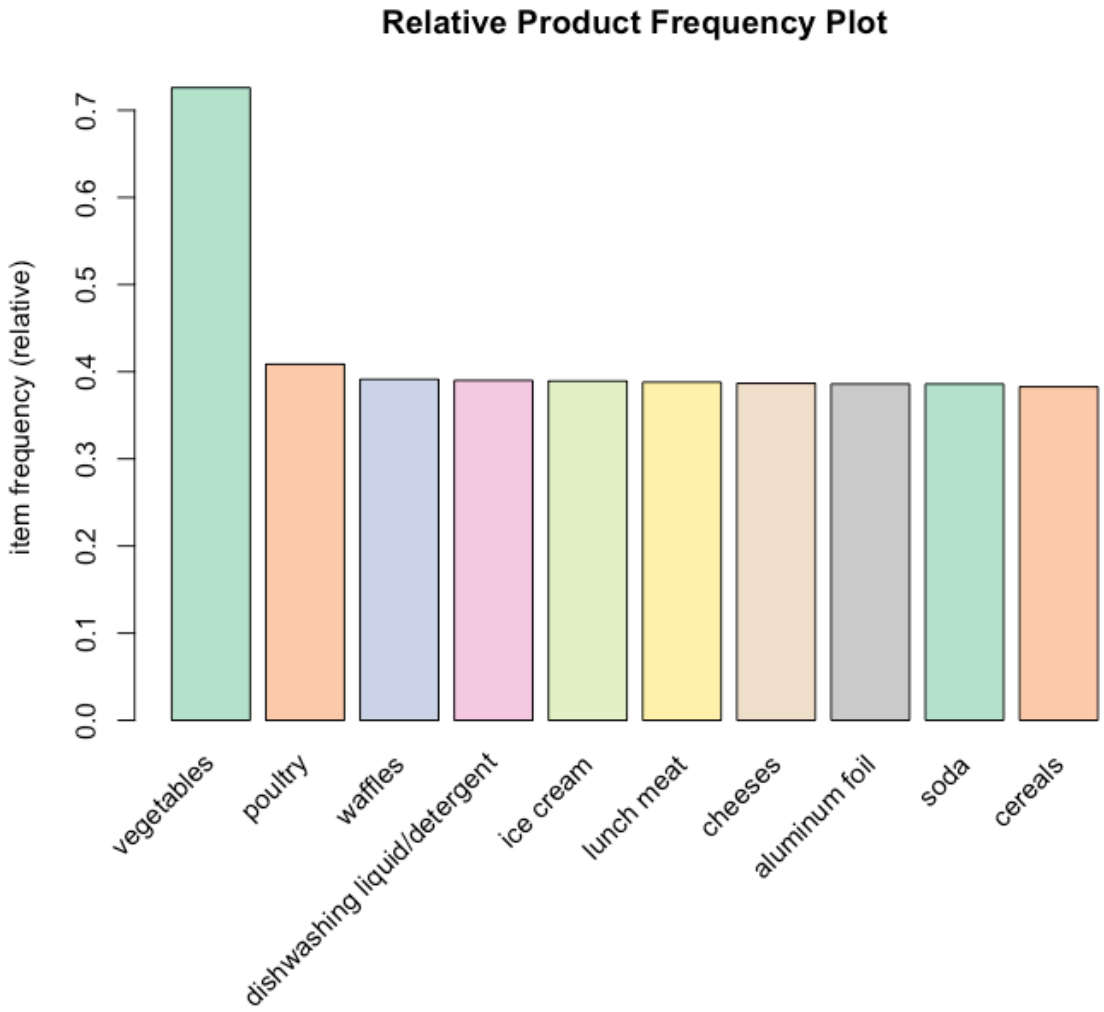

Figure 1 " Relative Frequency Distribution" own results

The plot shows that vegetables were sold the most with a relative frequency slightly over 70%, followed by poultries with 40%, and the other products slightly below the 40% threshold.

1.3 The Association Rules Algorithm

The association rules is a fundamental techniques used to discover patterns and correlations to identify the frequent itemsets in a dataset. The Apriori algorithm (Lantz, 2015) was used for identifying associations in the groceries data by using Quality measures of Support, Confidence and Lift. The scope was to find all subset of items that would occur with a probability greater than Support threshold s. Support measures how often the combination of items A and B co-occur in an observations, whereas confidence measures the probability of B occurring when in a transaction also A occurs. If Lift which is the ratio of probability of A and B occurring together is equal to 1, then the itemsets are independent.

For the groceries dataset, with the Apriori Algorithm the key task when finding association rules was to find all subsets of the items that occur with probability greater than s, where s is the support threshold.

- application with 0.4 confidence and a minimum support of 0.1 and 3 as min and max length of items was generated, resulting in the highest confidence and lift respectively of 0.85, meaning that in a basket of eggs and soda, 85% of chances of a client would result in buying vegetables as well.

The length of 3 items corresponded to 1500 transactions and 1820 rules, obtained thanks to the support and confidence applied. Below are two plot to visualise the conclusive results;

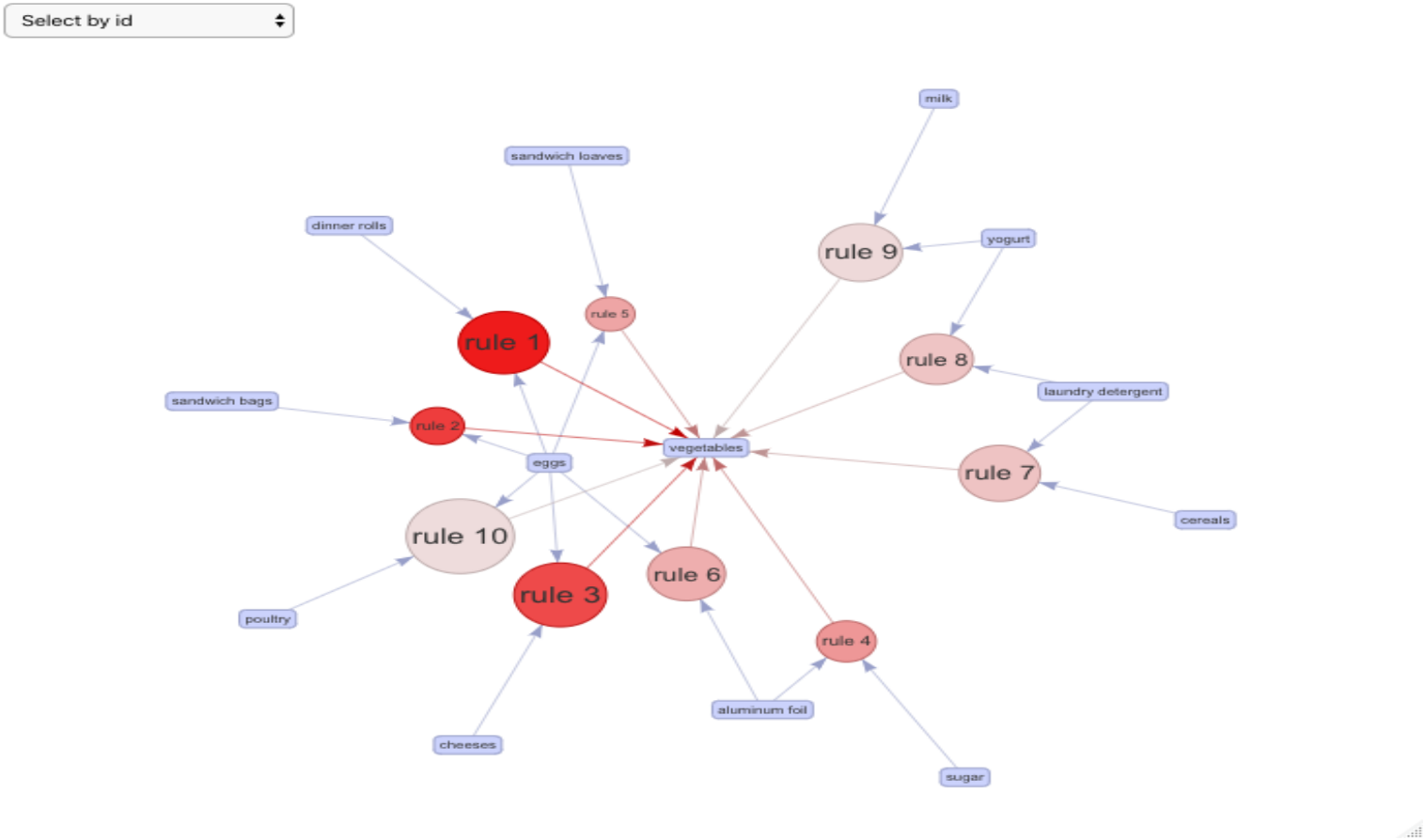

Figure 2 " Apriori Algorithm Plot" own results

The above plot filters the first 10 rules by “confidence”. We can clearly see that vertices were labelled with item names and rules are represented as a second set of vertices. The arrows pointing from items to rules indicate LHS, whereas from a rule to an items are RHS. The size represents

the interest measures. For example, the first with a confidence of rule with the highest confidence shows that a basket of dinner roles and eggs would highly determine the purchase of vegetables.

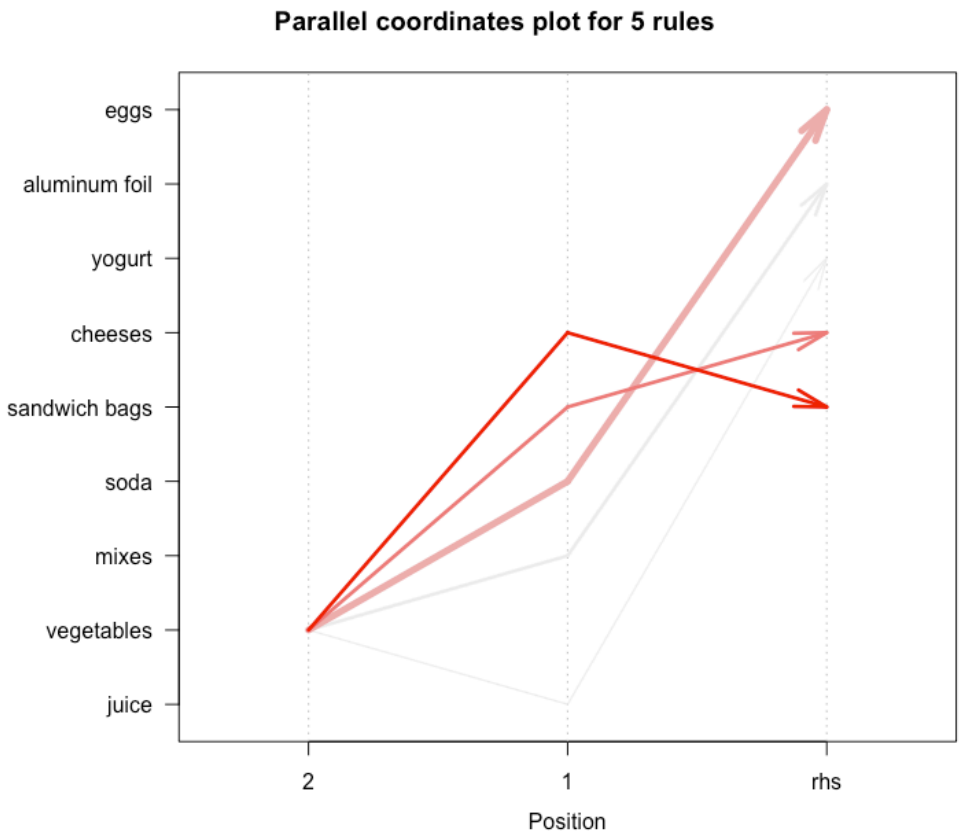

Figure 3"Parallel coordinates plot" own results

The second plot called parallel coordinates plot was filtered by the top 5 rules with the highest lift. It is useful to visualized which products along with which items cause different type of sales. On the x axis we have number 2 that corresponds to the most recent addition to the basket and 1 is the one added previously. The arrows correspond to the highest lifts. The top arrows shows that with a basket of vegetables and soda, somebody would be likely to purchase eggs.

2 References

R Core Team, 2013. *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* [Online]
Available at: https://www.r-project.org/
[Accessed 13 July 2020].

Chiu, K., Luk, R. . W. P., Chan, K. . C. & Chung, F.-L., 2002. *Market-basket analysis with principal component analysis: an exploration.* [Online]
Available at: https://www.researchgate.net/publication/3996638_Market-basket_analysis_with_principal_component_analysis_an_exploration
[Accessed 15 July 2020].

Lantz, B., 2015. *Machine Learning with R.* Second ed. Birmingham: Packt.

Wickham, H., 2014. Tidy Data. *Journal of Statistical Software,* Volume 59, p. 154416.