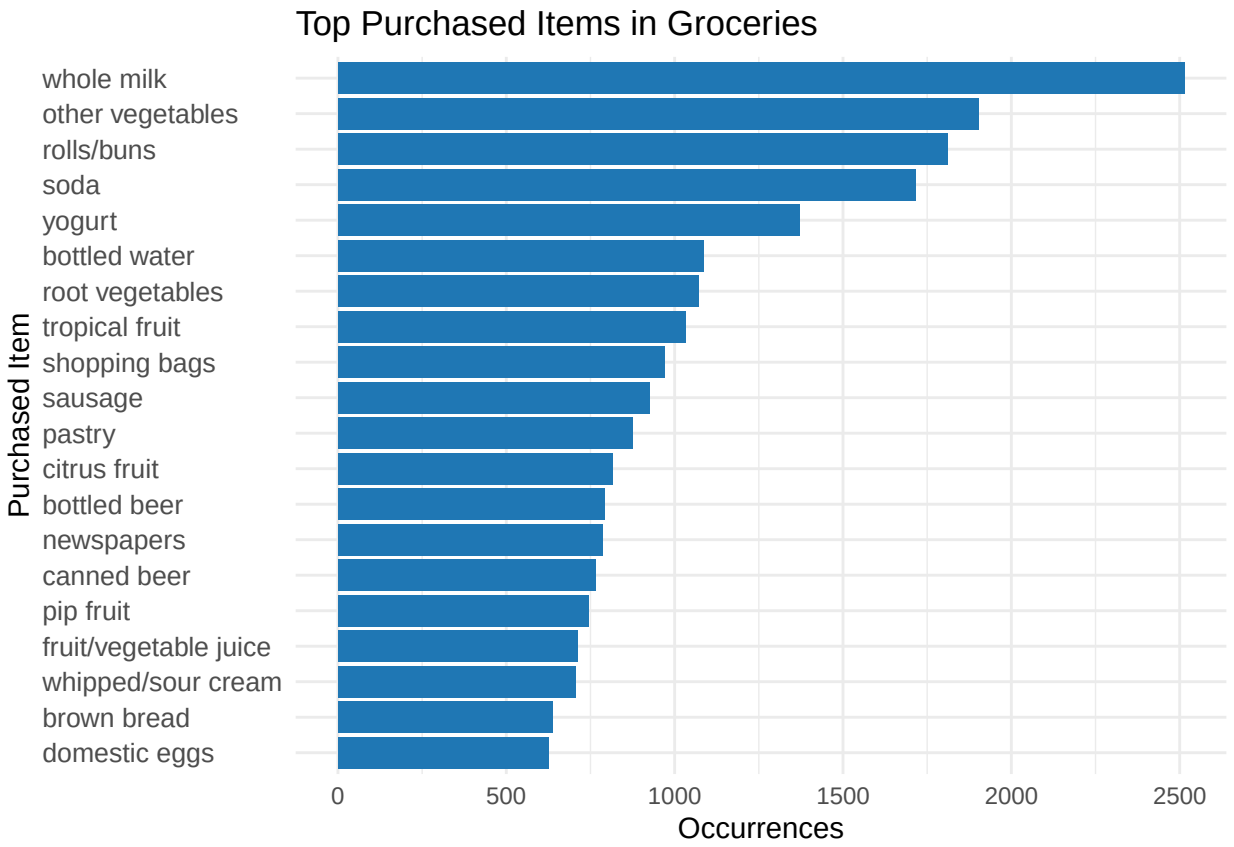
Association rule mining

Exploring the data set

The dataset has a total of 15296 items with a maximum of 4 items in the cart. Below is a snippet of the first few rows of the dataset.

V1	V2	V3	V4
citrus fruit	semi-finished bread	margarine	ready soups
tropical fruit	yogurt	coffee	NA
whole milk	NA	NA	NA
pip fruit	yogurt	cream cheese	meat spreads
other vegetables	whole milk	condensed milk	long life bakery product
whole milk	butter	yogurt	rice

Next, we are going to observe the top 20 items that have been found in the baskets of each of the customs.



Applying the Association Rule

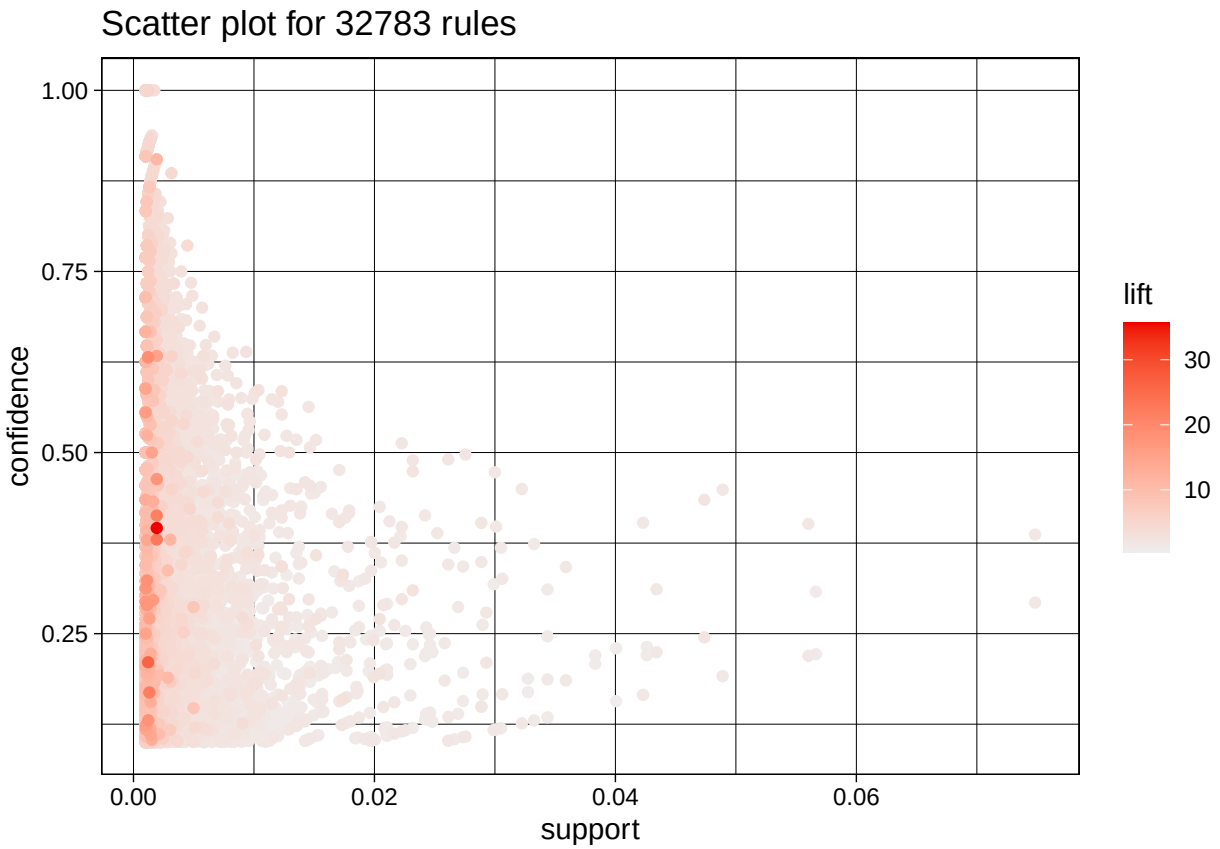
Initial Rule:

Since this is the preliminary analysis, we are establishing the following rules:

- the minimum fraction of carts that contain all of the items referenced in the rules is 0.1% (i.e support = 0.001)
- the minimum confidence that we will find the item in the cart is 10% (conf = 0.1). We are starting with a very low value just to get an idea of what the plot would look like.
- We want to take a look at the combination of at least two items hence, the minlen is 2.

Here's a table demonstrating the results of the rules that we have established.

```
##     lhs                rhs          support     confidence coverage    lift
## [1] {honey}        => {whole milk} 0.001118454 0.7333333  0.001525165 2.870009
## [2] {soap}         => {whole milk} 0.001118454 0.4230769  0.002643620 1.655775
## [3] {tidbits}      => {soda}       0.001016777 0.4347826  0.002338587 2.493345
## [4] {tidbits}      => {rolls/buns} 0.001220132 0.5217391  0.002338587 2.836542
## [5] {cocoa drinks} => {whole milk} 0.001321810 0.5909091  0.002236909 2.312611
##     count
## [1] 11
## [2] 11
## [3] 10
## [4] 12
## [5] 13
```



We found that there were 32783 Rules. That is too high of a number! Let's prune it down and take a look at fewer rules.

Adding Some Restrictions:

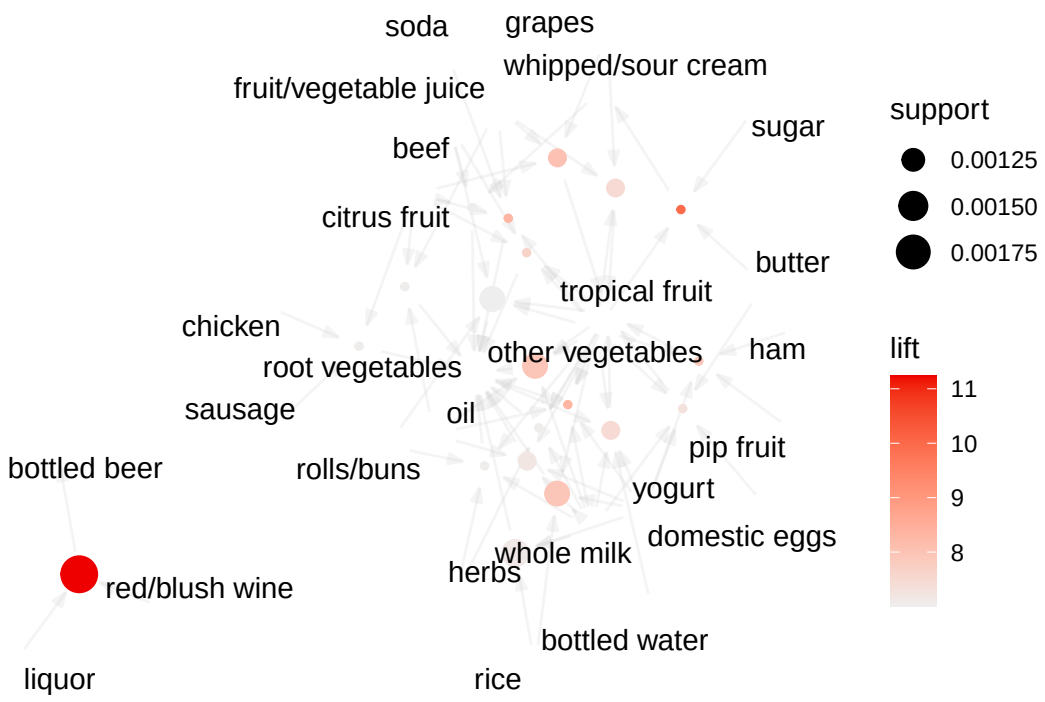
Even after adding limitations where we are looking at lift>1 and confidence>0.2, we still have too many rules. Let's add some more restrictions and prune it further.

Final Analysis:

The new rules are as follows:

- the minimum confidence that we will find the item in the cart is 70% (conf = 0.7). If the confidence is very low (i.e 10%), we are saying that the possibility of someone buying a bread given that they are buying whole milk is only 10%. That wouldn't help us if we were for example reorganizing the store layout and putting items next to each other based on the likelihood of the customer buying the combination of the items.
- The lift will be greater than 8. Meaning we are looking at the factor by which the probability has increased of finding an item on the LHS given that we know the item on the RHS.


Scatter plot for 20 rules
confidence
0.90
0.85
0.80
0.75
0.00100
0.00125
0.00150
0.00175
support
lift
11
10
9
8



Taking a look at the rule with the highest confidence (dark red circle), we can see that if we find liquor and red/blush wine, in the basket, we are likely to find bottled beer (there's a confidence of approximately 90.5%). In the same manner, we can read the rest of the rules where the items at the tail of the arrow (pointing towards the rule) represent items in the LHS while the items at the head of the arrow (pointing out of the rule) represent items on the RHS.