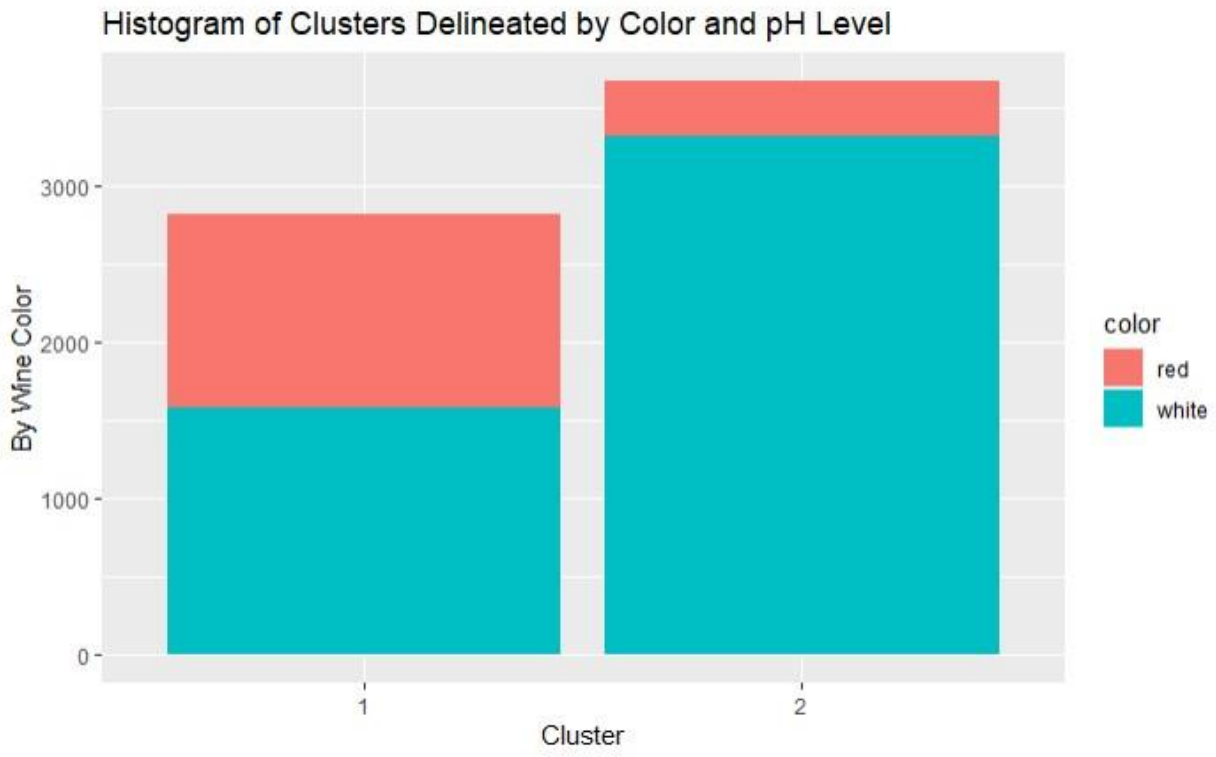
Exercise 4

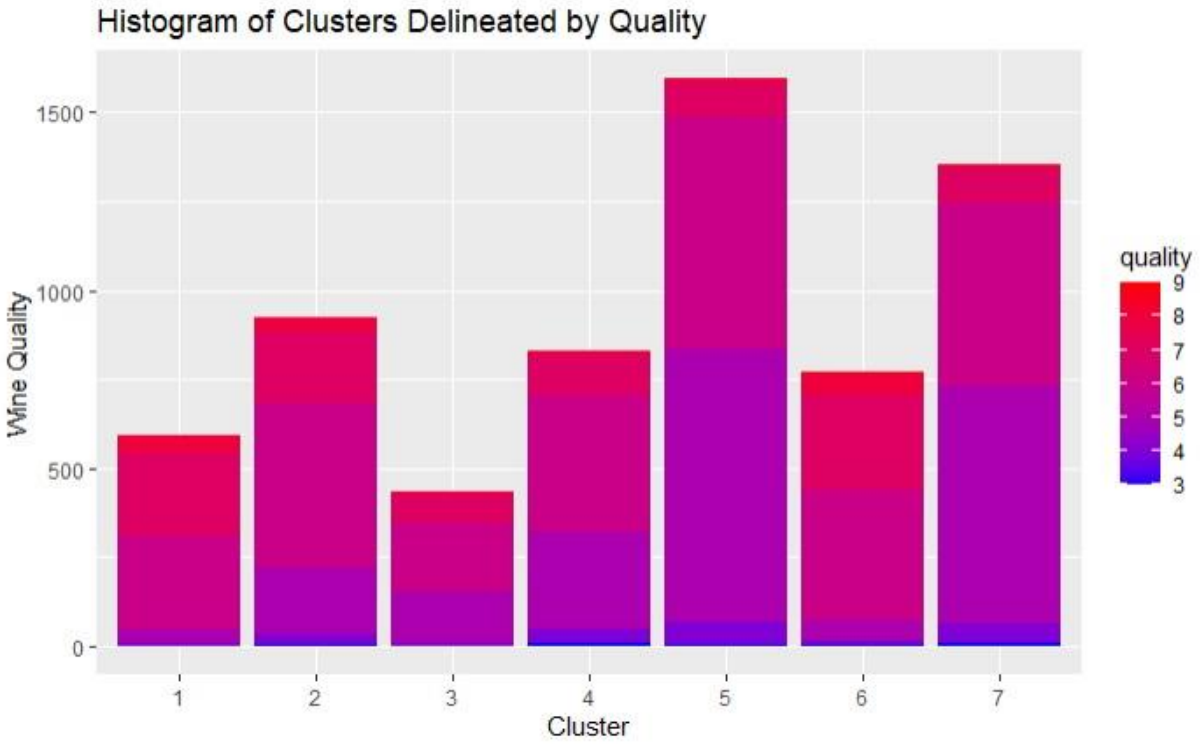
Elliot Spears

Question 1

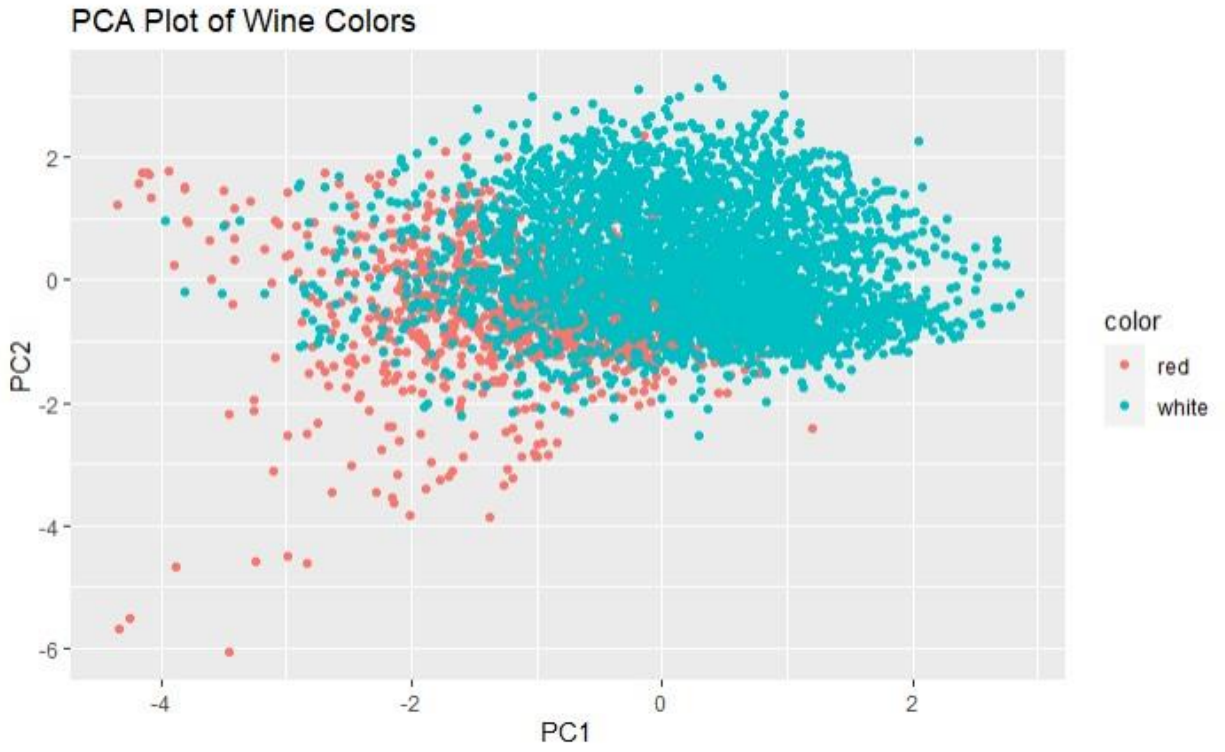
Here we aim to find out if any useful information about wine quality and wine color can be extrapolated by the tools of Principal Components Analysis and Kmeans++. We have data on 6497 different bottles of wine from vinho verde vineyards. We start with Kmeans++ and use it to segregate wines based upon pH levels and color. Then we use it to cluster based upon quality. After that we use PCA in order to test whether it is a better gauge of the relevant relationships. We find that Kmeans++ provides us with more suitable plots for our purposes with this data set.



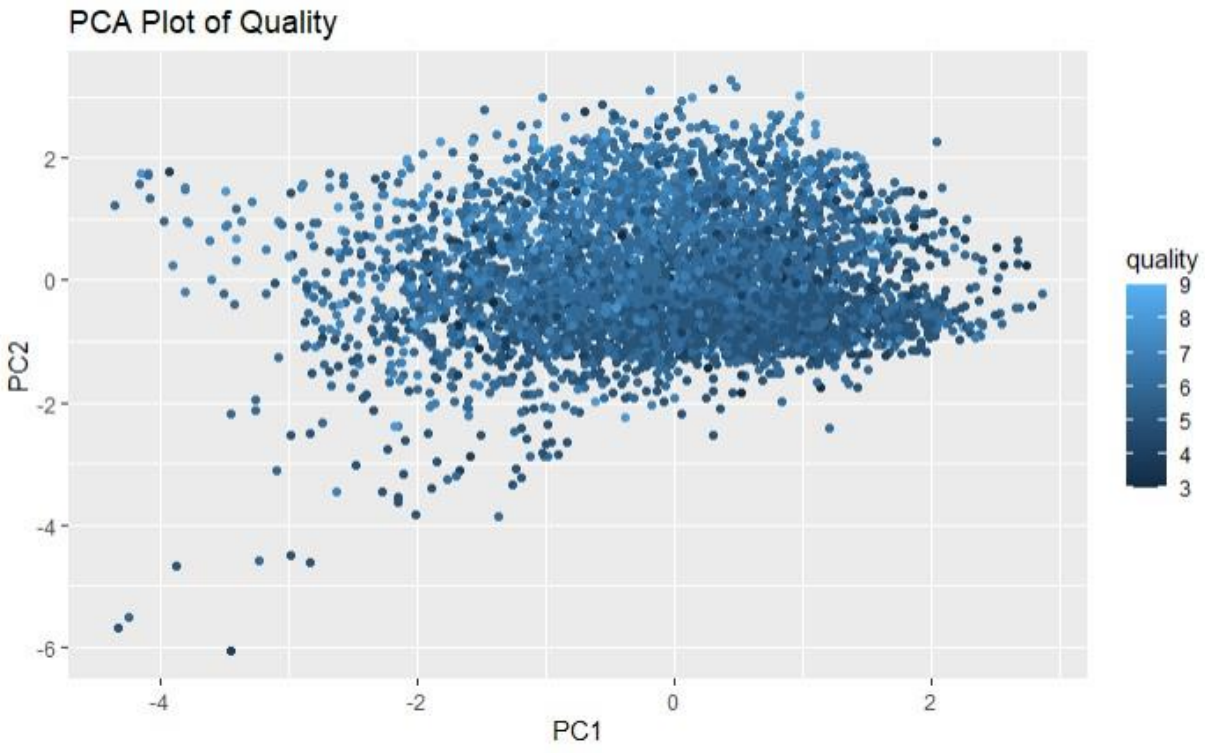
We've sorted two clusters that provide us with information on how the different wines are classified when clustered by pH levels. White wine clearly dominates in cluster two. Cluster two probably consists of the less acidic wines, since pH level are higher in red wines, and pH levels are a measure of acidity.



Again, we have a pretty go segmentation of wines based upon their quality. We have seven clusters. Just by eyeballing, the sixth cluster seems to have the most wins in the 8-9 rating range. Hence, whatever features are used to categorize cluster six would be worth noting to better understand what creates a higher quality wine.



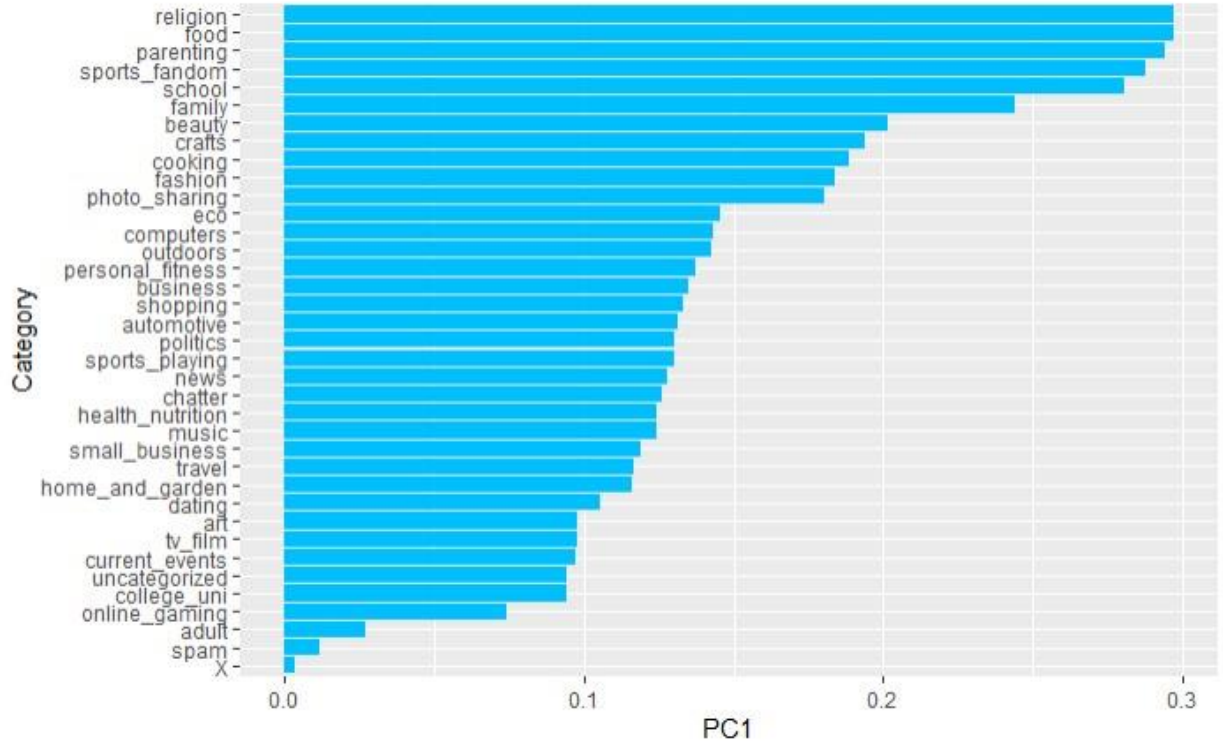
Although there is some differentiation here in terms of the red cluster and the white cluster, this plot is virtually useless as there is tremendous overlap between the two right in the middle of the plot.



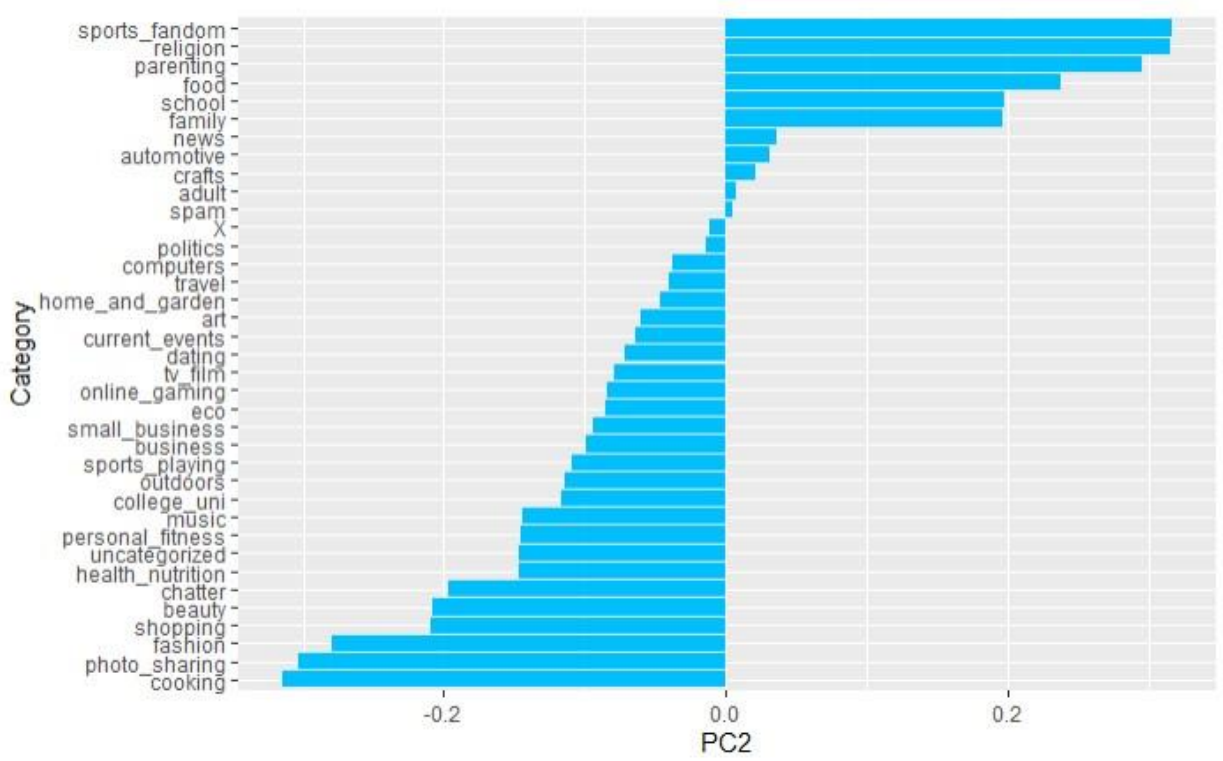
Once again, this graph isn't very helpful in enabling us to decipher anything. The pH levels are what we are using to categorize the PC groupings of wine, but it's difficult to extrapolate anything of substance from this plot about the relationship between pH levels and wine quality. Overall, it seems that Kmeans++ did a better job of providing us with useful plots and information about the relationship between pH levels, wine quality, and color.

Question 2

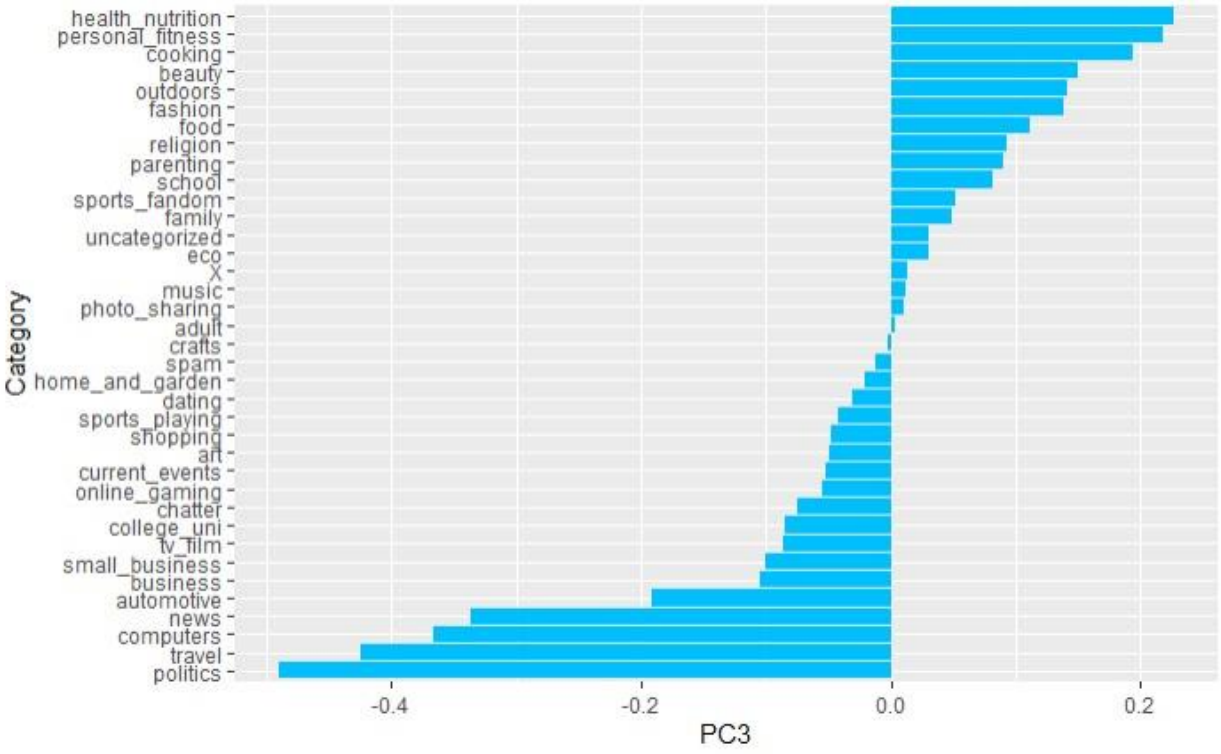
The purpose of this next analysis is to provide a report for NutrientH20, which identifies any interesting market segments that appear to stand out in their social-media audience.



Here we see that variables like spam, adult, and online gaming are on the opposite side of the PC spectrum from variables like religion, parenting, and school. It's difficult to gauge what aspect of the data generated this spread, but perhaps "shareable content for the average user" would be a decent approximation of what created this segmentation.



In this second PCA plot we see that seemingly unrelated categories occupy spaces much closer to one another. Religion and adult are both on the positive end of the spectrum. What could the "ingredient" makeup of this plot be? Well, on the negative side we have categories like: cooking, photo sharing, shopping, fashion, and beauty. On the positive side we have categories like: religion, sports fandom, school, news, and parenting. Perhaps the level of controversy that these tweets stirs up is contributing to the PCA mapping is organized here.



I wanted to produce one more PCA plot. PC3 seems to have more stale topics on the bottom end of the spectrum, whereas more invigorating and fluffy categories makeup the top rungs.

MODEL INFO:
Observations: 7882
Dependent Variable: religion
Type: OLS linear regression

MODEL FIT:
$F(6,7875) = 1347.58$, $p = 0.00$
$R^2 = 0.51$
Adj. $R^2 = 0.51$

Standard errors: OLS

	Est.	S.E.	t val.	p
(Intercept)	-0.00	0.03	-0.18	0.86
family	0.31	0.01	20.80	0.00
politics	0.02	0.01	3.17	0.00
news	-0.06	0.01	-6.39	0.00
current_events	0.03	0.01	2.39	0.02
school	0.36	0.01	23.82	0.00
parenting	0.59	0.01	48.90	0.00

Here I created a linear model to try and predict the expected number of posts about, arguably, the most controversial topic in the data set, using the other most controversial topics in

the data set. Everything with the exception of the "news" category had a positive effect on the predicted value of number of tweets about religion that a particular subject puts out.

MODEL INFO:
Observations: 7882
Dependent variable: adult
Type: OLS linear regression

MODEL FIT:
$F(35,7846) = 31.44$, $p = 0.00$
$R^2 = 0.12$
Adj. $R^2 = 0.12$

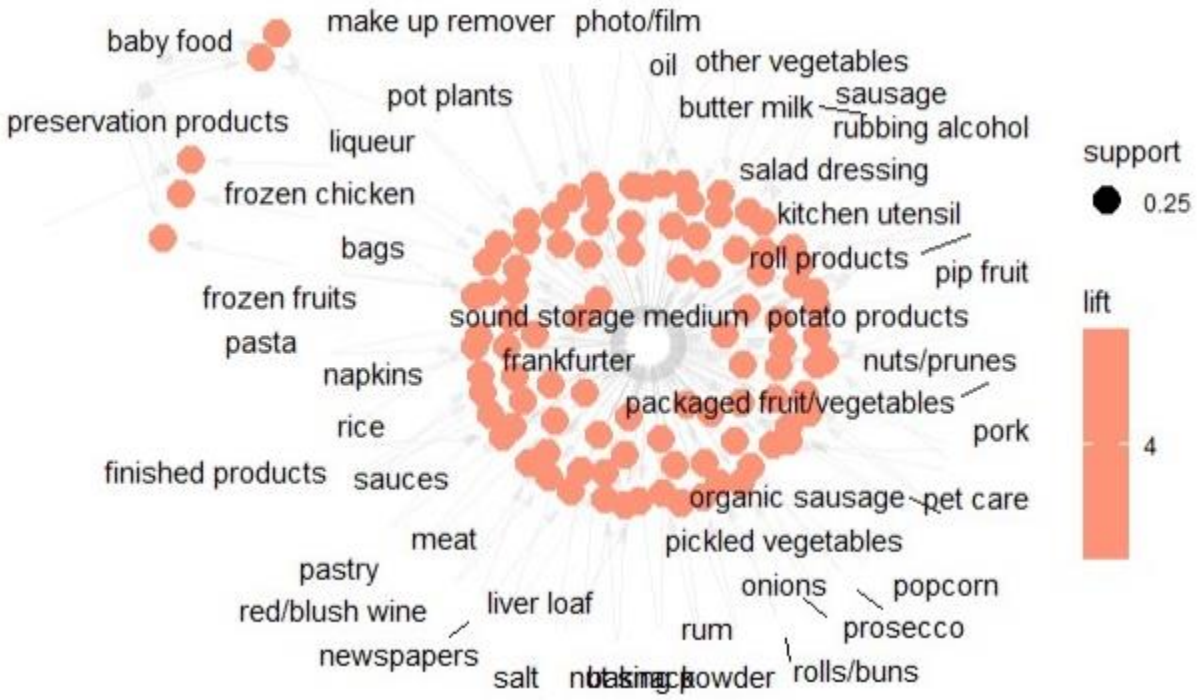
Standard errors: OLS

	Est.	S.E.	t val.	p
(Intercept)	0.15	0.05	3.25	0.00
chatter	0.01	0.01	1.85	0.06
current_events	0.02	0.02	0.99	0.32
travel	0.05	0.01	3.78	0.00
photo_sharing	-0.03	0.01	-3.00	0.00
uncategorized	0.08	0.02	3.73	0.00
tv_film	-0.05	0.01	-3.60	0.00
sports_fandom	-0.03	0.01	-2.53	0.01
politics	-0.07	0.01	-6.77	0.00
food	-0.00	0.02	-0.01	0.99
family	0.05	0.02	2.60	0.01
home_and_garden	0.07	0.03	2.66	0.01
music	-0.00	0.02	-0.16	0.87
news	-0.02	0.01	-1.74	0.08
online_gaming	0.04	0.01	2.92	0.00
shopping	-0.03	0.01	-2.40	0.02
health_nutrition	-0.04	0.01	-4.89	0.00
college_uni	-0.04	0.01	-3.29	0.00
sports_playing	-0.03	0.02	-1.21	0.23
cooking	-0.01	0.01	-1.22	0.22
eco	0.12	0.03	4.48	0.00
computers	0.08	0.02	3.71	0.00
business	-0.02	0.03	-0.68	0.50
outdoors	0.17	0.02	8.37	0.00
crafts	0.04	0.03	1.52	0.13
automotive	0.09	0.02	5.09	0.00
art	0.02	0.01	1.38	0.17
religion	-0.05	0.02	-3.44	0.00
beauty	0.02	0.02	0.92	0.36
parenting	0.05	0.02	2.89	0.00
dating	-0.03	0.01	-2.43	0.01
school	0.05	0.02	2.37	0.02
personal_fitness	0.00	0.01	0.30	0.76
fashion	0.00	0.02	0.05	0.96
small_business	0.23	0.03	7.19	0.00
spam	6.15	0.23	26.59	0.00

I thought it might be of interest which variables do the most to help predict the amount of posts an account makes for the "adult" category. It appears that "spam" is the strongest predictor, which isn't surprising since spam and adult content are very closely intertwined. Interestingly, the outdoors category has a fairly large and positive predictive value at the 5% level for the predicted number of adult tweets. Since posts with "adult" content probably belong to a certain class of user, it might be helpful for NutrientH20 to know how to target that market using this data.

Question 3

The objective of the next case study is to assist a grocery store in finding some interesting association rules through shopping trends based upon customer basket data. We will help the grocery store see which items are usually purchased in tandem with one another in order to provide the grocer with the necessary information to better organize the way in which they stock their shelves, putting food items in close proximity to each other than are demonstrated through this data to be purchased.



Here we see some examples of items that tend to be purchased together. This can be assessed by proximity and the direction of the vectors. For instance, we see baby food and pot plants are often purchased together. It might not make sense to put these items in the same part of the store, but it is useful information, and perhaps some advertising flyers could be posted in the baby food aisle to target that sort of consumer.