Data Mining HW 4 Submission

Patrick Chase

5/3/2021

1.

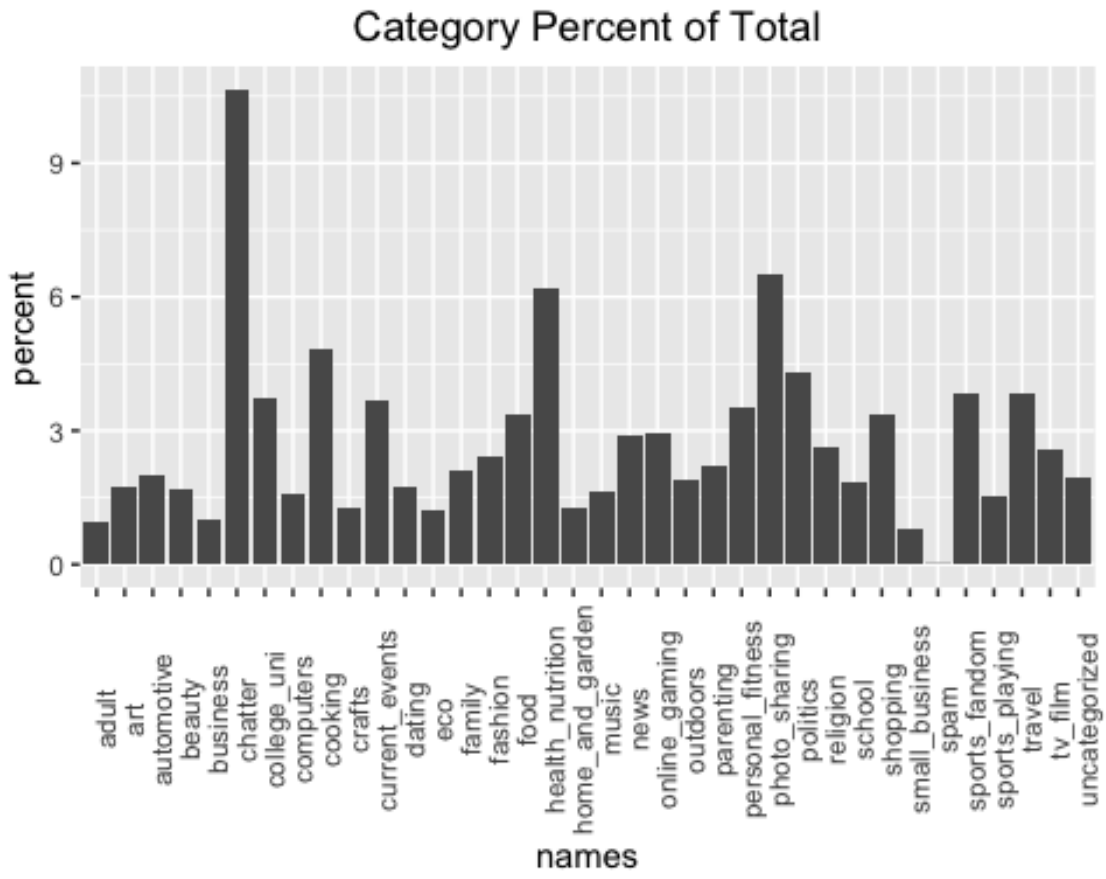
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC
7
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845 0.77930 0.7233
0
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544 0.05521 0.0475
6
## Cumulative Proportion  0.2754 0.5021 0.6436 0.73187 0.79732 0.85253 0.9000
9
##                            PC8     PC9   PC10    PC11
## Standard deviation     0.70817 0.58054 0.4772 0.18119
## Proportion of Variance 0.04559 0.03064 0.0207 0.00298
## Cumulative Proportion  0.94568 0.97632 0.9970 1.00000
```

PCA 1 through PCA 9 capture more than 95% of the variability in our data set. I'll rely on only those for my PCA model.

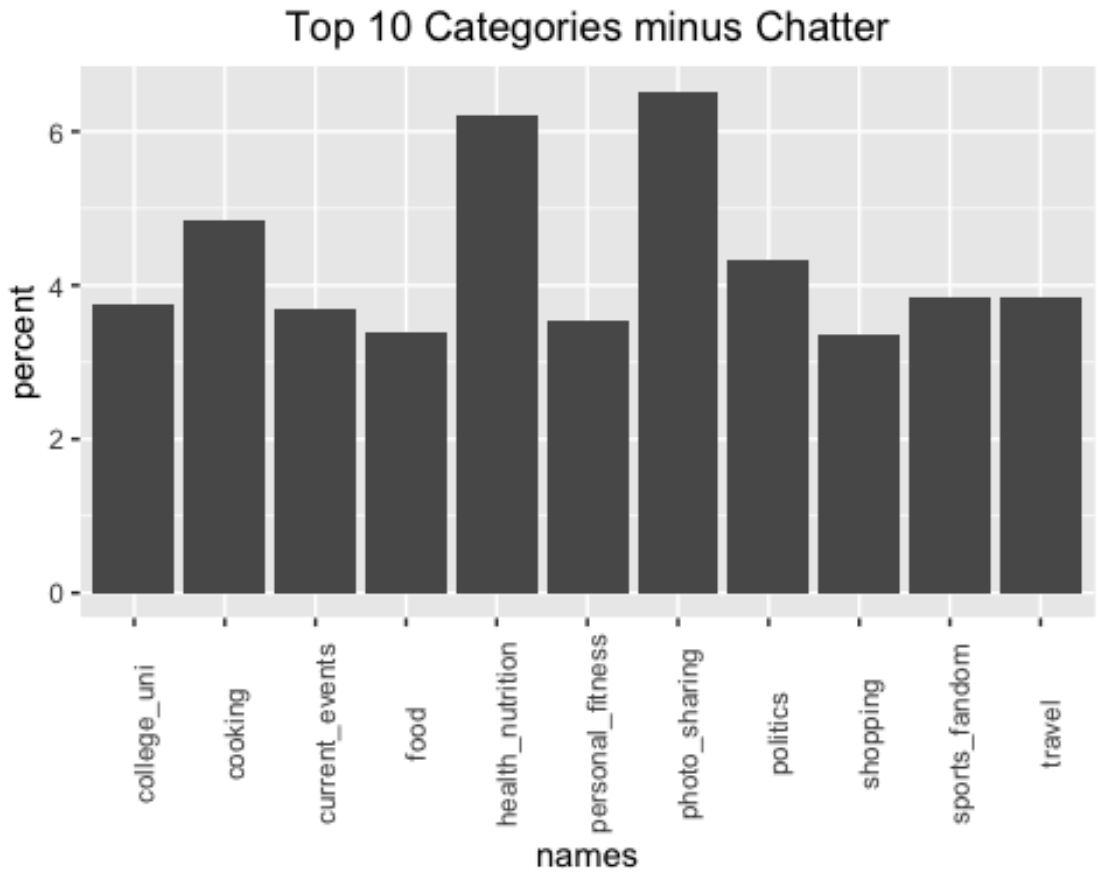
```
## # A tibble: 1 x 6
##   pca_red_accuracy rmse_pca_red pca_qual_accura… rmse_pca_qual cluster_red
l_ac…
##              <dbl>        <dbl>            <dbl>         <dbl>
<dbl>
## 1             99.2         7.81            0.616         0.727
54.9
## # … with 1 more variable: cluster_qual_accuracy <dbl>
```

For this specific task, I think PCA makes a lot more sense, particularly for the binary classification between red and white where I can achieve accuracy of 98%. My PCA model estimating quality is just barely wasn't particularly good. As far as I can tell, the K-means models aren't outputting information that is practically useful. I'm not sure if that's because I made a mistake or if there is something else going on.

2.

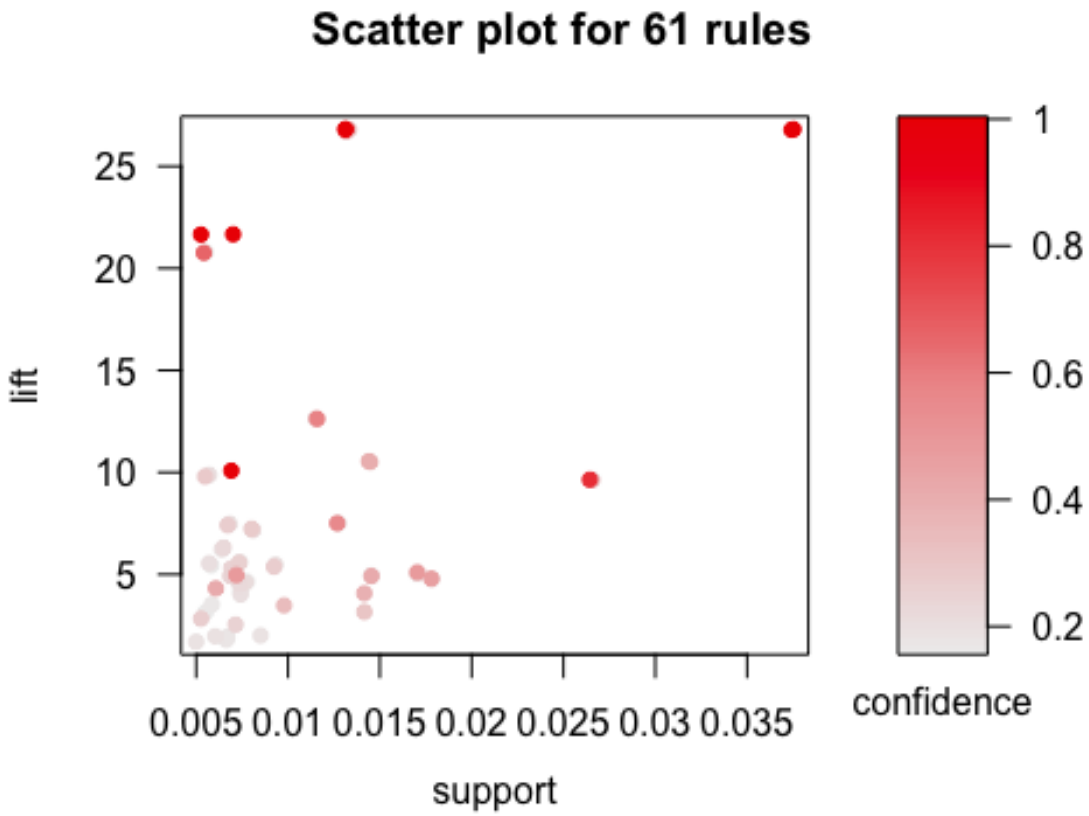


This first graph is showing the percentage breakdown for each category in our data set. Let's take a closer look by focusing on the top 10 categories removing chatter as it is not substantively useful.

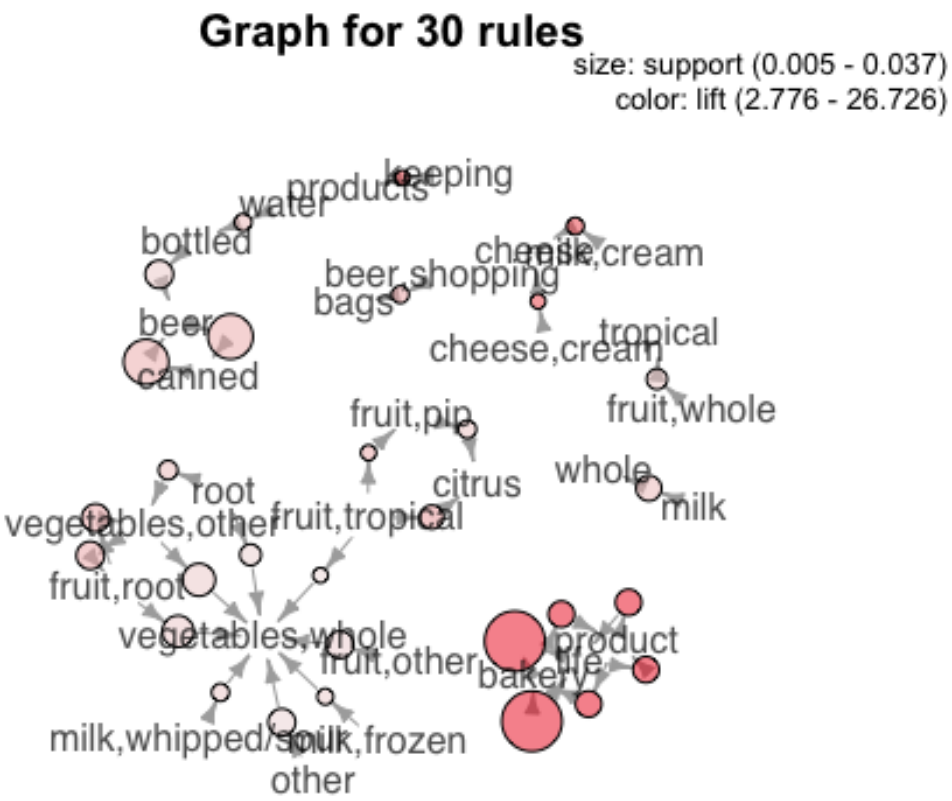


NutrientH20's social media followers seem to be focused on a typical range of what many consumers on twitter are. That said, a point of interest is the relatively large proportion that is related to food in general, when we aggregate between cooking, food, and health_nutrition. This in combination with the large share of photo_sharing gives us an idea of what some followers may be interested. The market segment in this case would be foodies who enjoy sharing their food endevours with their friends through twitter. This could point to a marketing strategy of engaging with chefs turned social media influencers. This would expand reach and give NutrientH20's consumers what they are interested.

3.



I chose a low level for support because I don't think a grocer would be particularly interested in focusing on products that all of their customers already buy. My threshhold for confidence was 15% because I think it would be more useufl to show associations of products in order to inform store organization. I decided to subset my rules to focus on those with the highest confidence in the graphic below.



This graphic shows rules that seem to follow the typical organization of a grocery store. Dairy products are largely purchased in tandem with each other, and most customers seem to be getting whole vegetables along with many other products.

4.

For my data cleaning procedure I basically followed you're work exactly. And even with that I struggled to transition it to a usable format where I could use the principal components I identified to build a classification model. That said, I had intended to build a model where I'd focus on classifying the work of one Mure Dickie. I imported the Document Term Matrix into a normal matrix and normalized it. I had intended to use my PCs as variables in a logit model, with the outcome variable being Author == "MureDickie". I suspect my classification may have been alright, although probably not great. My issue, ultimately was a data processing one. I was ultimately able to get my model to run, but I have no idea what the accuracy of it is.

```
## Error in model.frame.default(Terms, newdata, na.action = na.action, xlev =
object$xlevels): 'data' must be a data.frame, not a matrix or an array
```