ECO395 HW4
Yefu Chen
2021/4/25

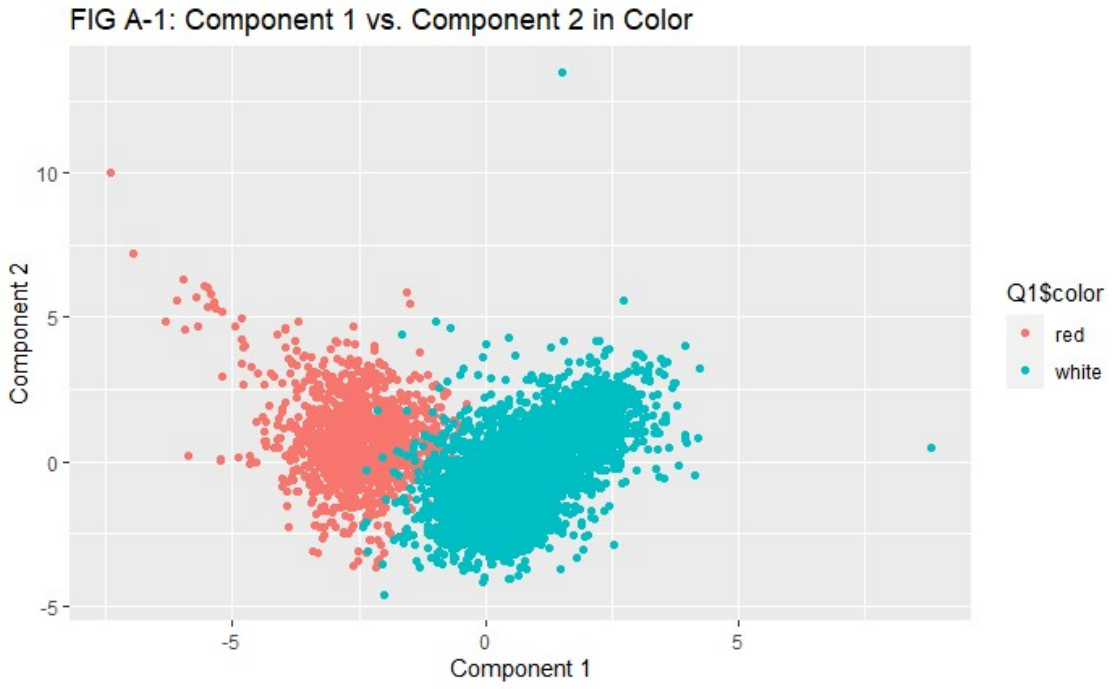
QUESTION 1

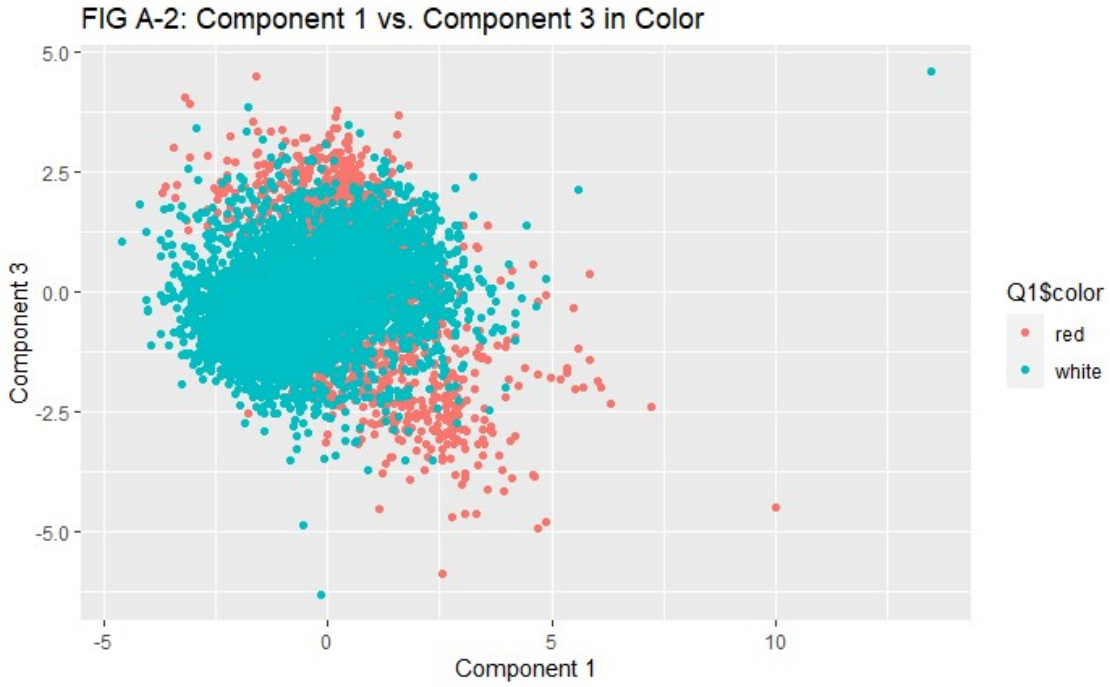
Run both PCA and a clustering algorithm of your choice on the 11 chemical properties and summarize your results.

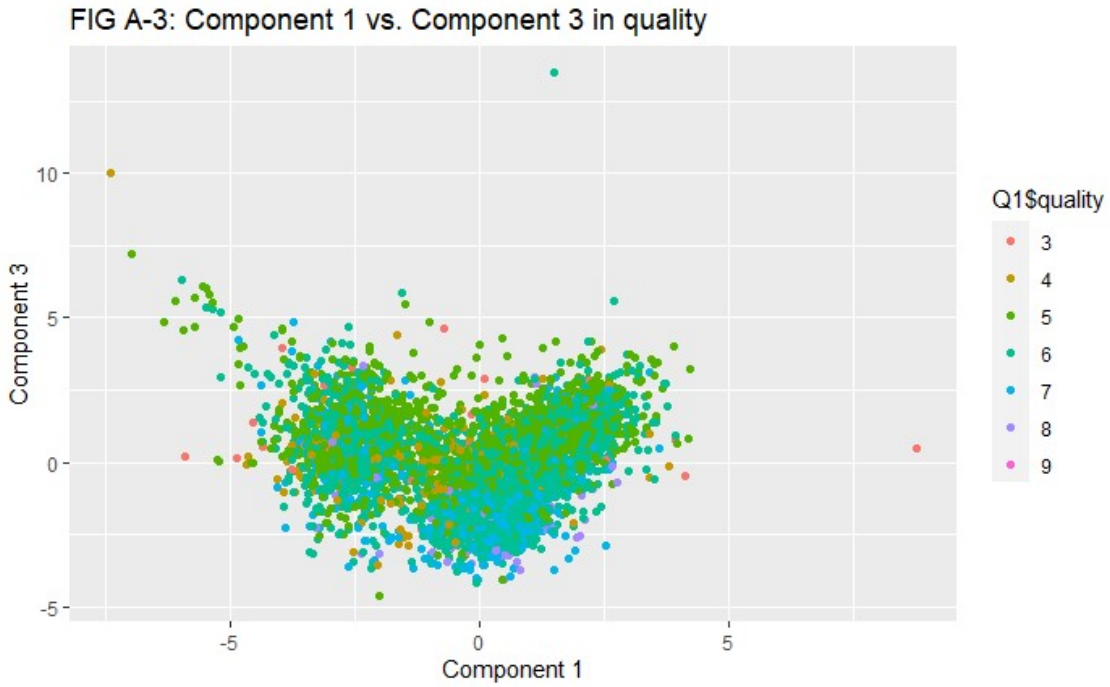
ANSWER:

The following Figures A and B present the clustering of the PCA and K-mean. Overall, the K-mean clustering makes more sense in this analysis since it categorizes elven chemical properties into several groups based on minimums within-cluster variances. PCA creates several components and assigns weights to elven chemical properties for each component. Results are consistent with this claim. Focusing on the results of PCA (Figures A), it can easily distinguish the boundaries in wine color between components 1 and 2 (FIG A-1), but it is hard in components 1 and 3 (FIG A-2) due to overlaps. Also, it is hard to define the quality of the wine based on PCA because the boundaries of different qualities are not clear (FIG A-3). Focusing on the results of K-mean (Figures B), cluster 3 is high-likely to be red wine, while clusters 1 and 2 are white wine (FIG B-1). FIG B-2 demonstrates that the clusters in qualities. We can define cluster 1 as relatively low-quality wine (quality 4 to 6), cluster 3 as medium-quality wine (quality 4 to 7), and cluster 2 as high-quality wine (quality 4-9). Also, cluster 2 wine has various performances in wine quality, needed further clustering.

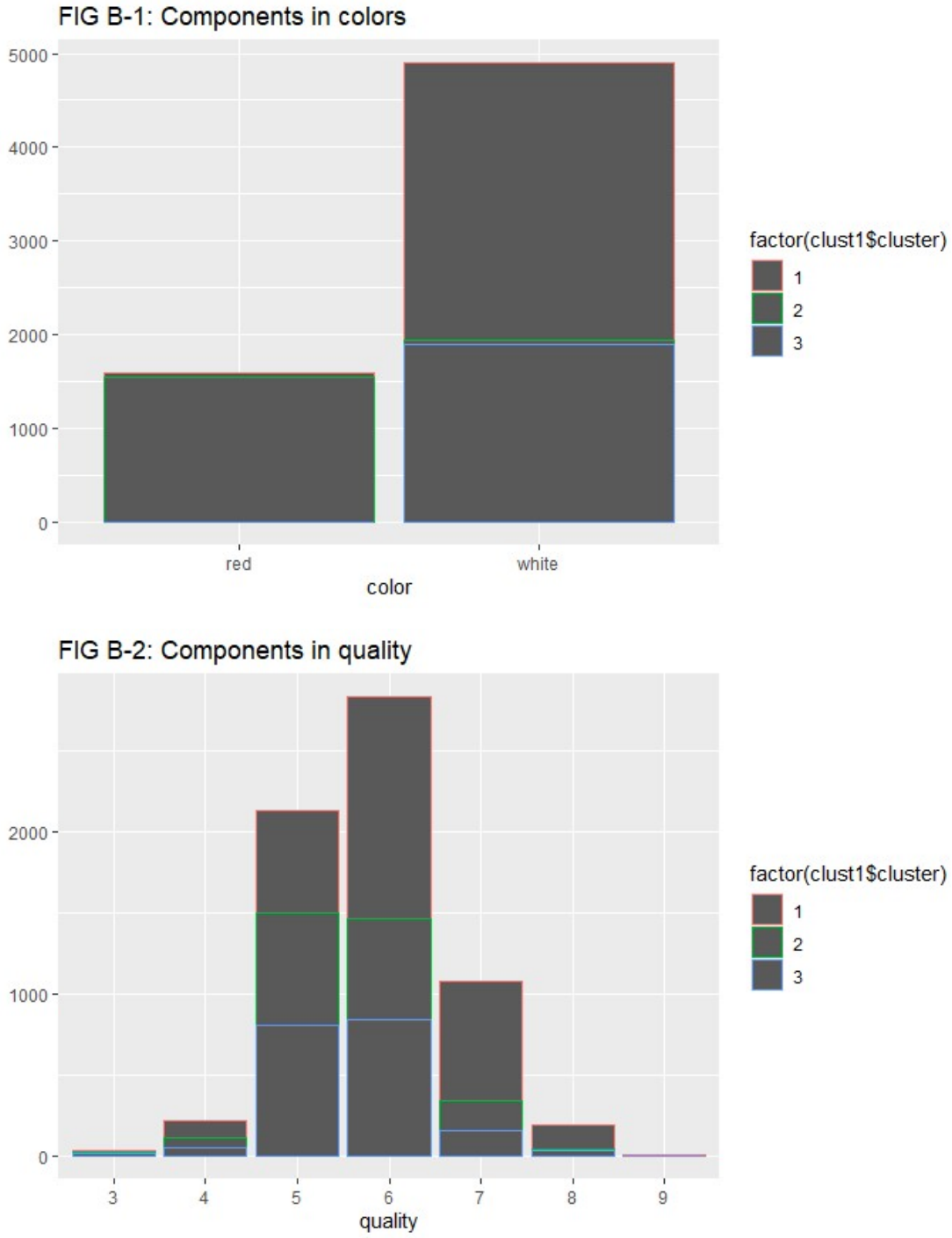
To sum up, in this analysis, the K-mean clustering can make more sense than PCA. The K-mean clustering can distinguish the reds from the whites and the higher from lower quality wines based on unsupervised information.







Figures A. PCA clustering analysis and labels



Figures B. K-mean clustering analysis and labels

QUESTION 2

Analyze this data as you see fit and prepare a short report for NutrientH20 that identifies any interesting market segments that appear to stand out in their social-media audience.

ANSWER

To identify interesting market segments, I apply heatmap, correlation, and K-mean clustering in this analysis. Overall, there are market segments in the NutrientH20 dataset.

Figure 2-1 presents the results of the heatmap. There are eight market segments. The first one includes news, travel, automotive, politics, and computer. The second one includes family, food, religion, school, sports fandom, and parenting. The third one includes outdoors, parenting, and health nutrition. The fourth includes personal fitness, chatter, photo sharing, and shopping. The fifth one includes shopping, fashion, beauty, and cooking. The sixth includes online gaming, sports playing, and college and university. The seventh includes spam and adults. The eighth one includes crats, TV film, art, and dating. The results of Figure 2-2 are consistent with the above claims. Then, I explore the K-mean clustering (choose K=12 as the best parameter) in the NutrientH20 dataset. Table 2-1 presents the clustering results of K-mean analysis.

Based on these results, advertisers can deliver advertisements to those who can be more likely interested. For instance, individuals interested in news and travel may also be interested in automotive, politics, and computers. Those interested in personal fitness can be more interested in chatter, photo sharing, and shopping. Unsurprisingly, the spam and adults are highly correlated in all three analyses. In this case, operators of the website, Twitter, should be aware of this link. By banning the account focusing on either spam or adults, the atmosphere of this social media can be cleaned.

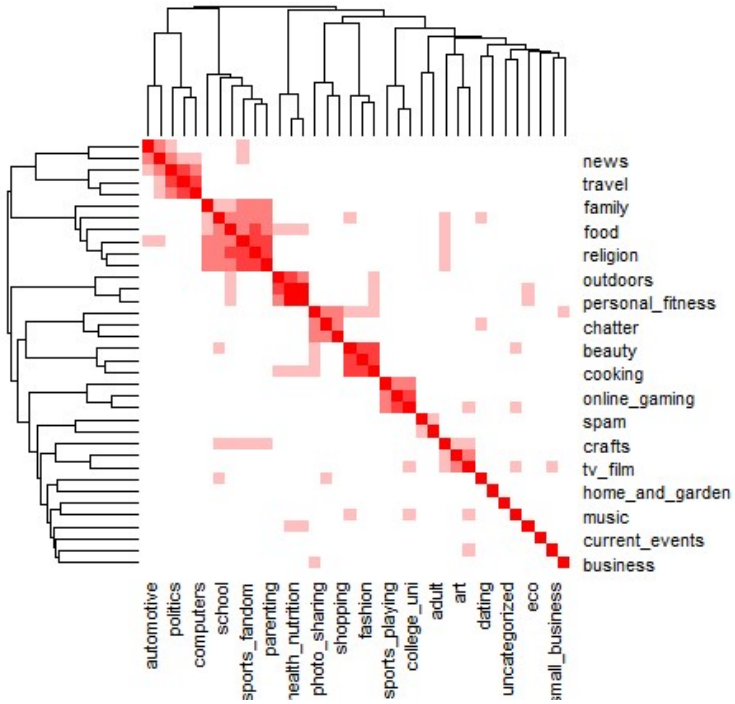


Figure 2-1. Heatmap analysis of interests

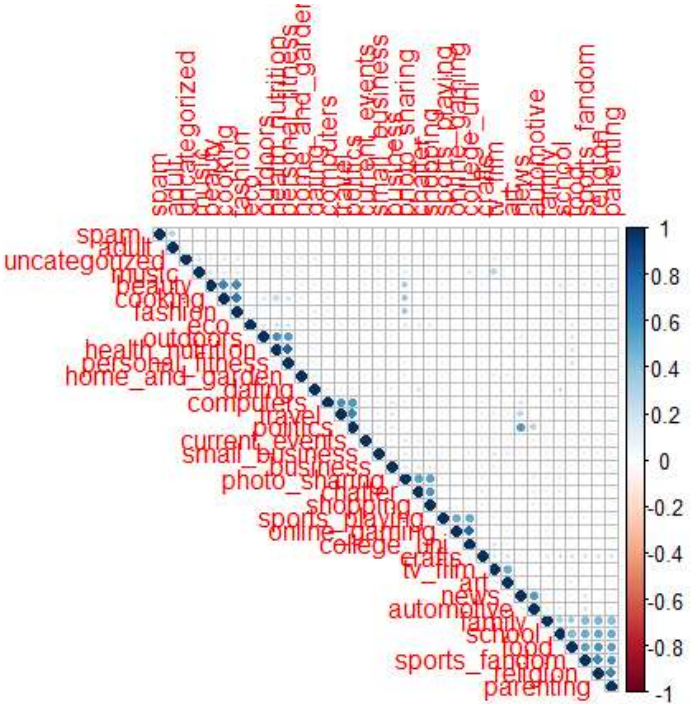


Figure 2-2. Correlation analysis of interests

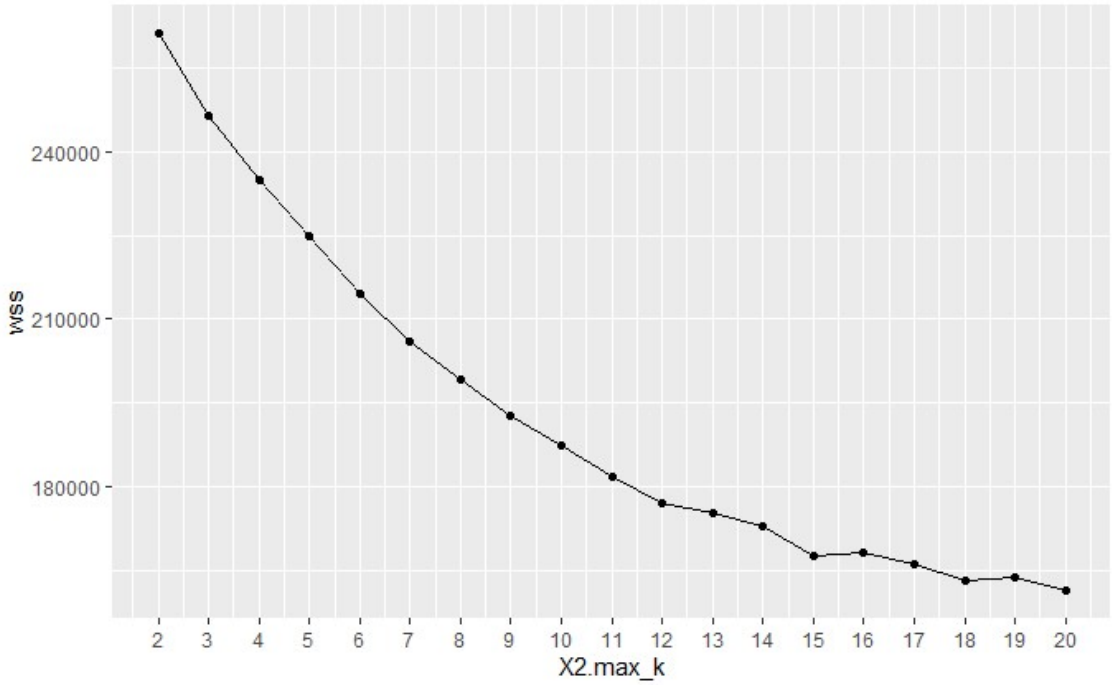


Figure 2-3. Optimal K searching

Tables 2-1. K-mean results

cat	chatter	current_e	travel	photo_sh	uncatego	tv_film	sports_fa	politics	food	family	home_an	adult
1	4.653061	1.877551	2.244898	2.44898	0.918367	0.877551	1.897959	2.244898	1.469388	0.795918	0.693878	7.204082
2	4.696594	1.659443	1.343653	3.28483	1.102167	1.052632	1.541796	1.489164	2.876161	0.956656	0.80805	0.572755
3	9.811752	2.01197	1.10555	6.062024	0.806311	0.850925	1.195865	1.398259	0.852013	0.831338	0.563656	0.355822
4	4.142518	1.619952	1.163895	2.104513	0.72209	1.047506	3.049881	5.513064	1.128266	1.118765	0.638955	0.209026
5	4.051576	1.415473	1.507163	2.653295	0.770774	1.232092	1.303725	1.252149	1.226361	1.097421	0.573066	0.366762
6	3.119921	1.248761	1.086554	1.569871	0.637265	0.707301	0.90783	0.894285	0.714569	0.533201	0.373968	0.401388
7	3.83359	1.639445	1.35131	2.416025	0.699538	0.909091	6.16641	1.101695	4.742681	2.605547	0.639445	0.397535
8	4.156652	1.744635	1.446352	6.042918	1.293991	0.83691	1.139485	1.379828	1.025751	0.896996	0.603004	0.405579
9	4.071006	1.674556	9.142012	2.381657	0.715976	0.970414	1.150888	11.36095	1.695266	0.766272	0.568047	0.139053
10	7.943299	1.634021	1.469072	2.623711	1.525773	0.938144	1.329897	1.314433	1.159794	0.752577	0.943299	0.314433
11	3.199192	1.411844	1.144011	1.689098	0.79004	0.671602	0.896366	0.921938	1.549125	0.554509	0.421265	0.258412
12	3.930693	1.950495	2.108911	2.443069	1.440594	5.618812	1.339109	1.532178	1.633663	0.75	0.757426	0.336634

cat	music	news	online_ga	shopping	health_n	college_u	sports_pl	cooking	eco	computer	business	spam
1	0.693878	1.204082	1.44898	0.959184	2.795918	1.918367	0.530612	1.795918	0.857143	1	0.183673	1.040816
2	0.900929	1.467492	1.111455	1.835913	16.27554	1.216718	0.839009	4.408669	1.318885	0.721362	0.603715	0
3	0.850925	0.645267	0.775843	4.186072	1.448313	1.247008	0.541893	1.214363	0.755169	0.616975	0.646355	0
4	0.589074	6.824228	0.881235	1.054632	1.422803	0.992874	0.548694	1.190024	0.434679	0.420428	0.342043	0
5	0.630372	0.805158	10.91404	1.13467	1.762178	11.11461	2.750716	1.570201	0.464183	0.553009	0.352436	0
6	0.438718	0.543442	0.572184	0.687149	0.690783	0.822266	0.379584	0.799141	0.292699	0.351503	0.262967	0
7	0.701079	0.967643	0.992296	1.338983	1.827427	1.164869	0.731895	1.640986	0.656394	0.750385	0.486903	0
8	1.242489	1	1.143777	1.738197	2.242489	1.497854	0.815451	11.72747	0.5	0.706009	0.564378	0
9	0.636095	3.62426	0.754438	1.254438	1.701183	1.43787	0.671598	1.331361	0.627219	4.121302	0.825444	0
10	0.64433	0.92268	1.036082	1.221649	2.149485	1.417526	0.943299	1.525773	0.613402	0.664948	0.726804	0
11	0.524899	0.802153	0.73755	0.845222	8.430686	0.734859	0.444145	2.379542	0.565276	0.40646	0.290713	0
12	1.700495	1.195545	0.722772	1.425743	1.777228	2.556931	0.762376	1.509901	0.581683	0.477723	0.643564	0

cat	outdoors	crafts	automoti	art	religion	beauty	parenting	dating	school	personal_	fashion	small_business
1	1.142857	0.693878	1	1.265306	1.326531	0.571429	1.204082	0.693878	0.877551	1.755102	0.959184	0.530612
2	3.879257	0.755418	0.736842	0.74613	1.173375	0.585139	1.139319	1.037152	0.752322	8.609907	1.058824	0.328173
3	0.451578	0.538629	0.986942	0.364527	0.574538	0.40914	0.609358	0.439608	0.733406	0.986942	0.71926	0.434168
4	1.142518	0.377672	4.394299	0.463183	0.743468	0.465558	0.983373	0.56057	0.767221	0.890736	0.581948	0.23753
5	0.621777	0.538682	0.928367	1.174785	0.730659	0.39255	0.716332	0.644699	0.501433	1.025788	0.845272	0.404011
6	0.300958	0.271886	0.414272	0.331351	0.527255	0.346878	0.434424	0.319128	0.380575	0.422531	0.454906	0.209779
7	0.676425	1.080123	0.992296	0.682589	5.525424	1.120185	4.249615	0.503852	2.751926	1.181818	0.987673	0.391371
8	0.821888	0.575107	0.843348	0.736052	0.854077	4.208155	0.830472	0.620172	0.916309	1.328326	6.006438	0.433476
9	0.715976	0.686391	0.64497	0.473373	1.328402	0.449704	0.943787	1.112426	0.630178	1.038462	0.66568	0.571006
10	0.85567	0.845361	0.572165	0.695876	1.185567	1.051546	1.092784	9.309278	2.262887	1.350515	2.510309	0.561856
11	1.804845	0.411844	0.456258	0.464334	0.515478	0.333782	0.499327	0.569314	0.382234	4.472409	0.495289	0.195155
12	0.655941	1.118812	0.529703	5.05198	1.118812	0.717822	0.616337	0.450495	0.722772	1.029703	0.910891	0.844059

QUESTION 3

Revisit the notes on association rule mining and the example on music playlists, then use the data on grocery purchases.

ANSWER

The purpose is to explore associations between items that individuals buy from grocery stores. Results indicate that meal purchasing, including meat and vegetables, is the most common daily behavior in this dataset. Also, individuals come to the grocery stores to get fruits, while this purchasing can be separate from meal purchasing.

First, I cleaned and organized the dataset in EXCEL through pivot tables to shape it from wide to long. The cleaned dataset has two columns, including consumer ID and items. Fig 3-1 demonstrates the top 20 popular items. Whole milk is the most popular item. There are around 2500 purchases of whole milk in this dataset. The top 5 popular items are other

vegetables, rolls/buns, soda, and yogurt, besides whole milk.

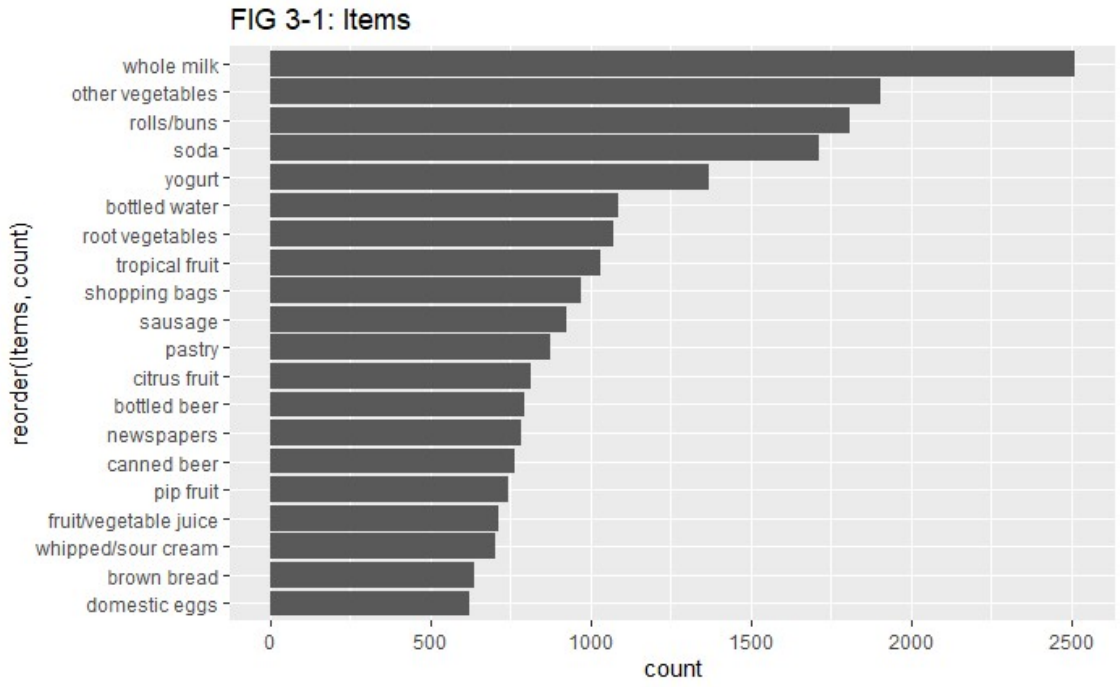


FIG 3-1. Top 20 popular purchase items

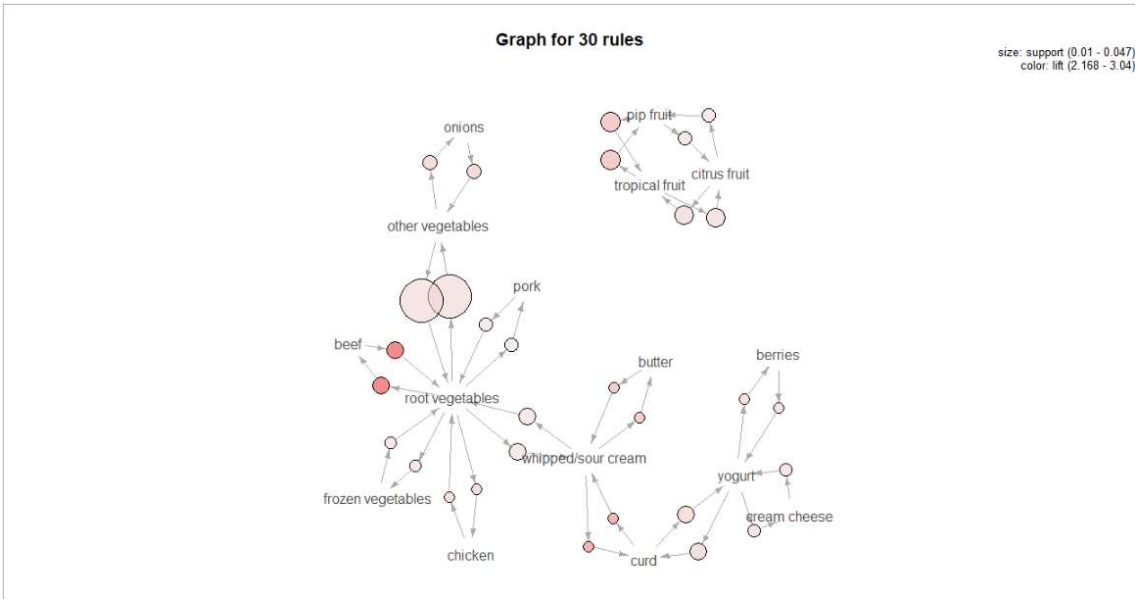


FIG 3-2. Relationship between top 50 transactions

Then, I used “apriori” function to create the rules. Fig 3-2 demonstrates the relationship between top 50 purchases (confidence > 0.05 & support > 0.01). I personally think the confidence should be better than 0.05 in the analysis. After a sort of testing, support > 0.01 performed the best in this modeling. According to Fig 3-2, individuals are most likely buying root vegetables and other vegetables together. Not surprisingly, they tended to buy mean

(e.g., beef and chicken) with vegetables simultaneously. Also, another cluster about fruits (pip fruit, tropical fruit and citrus fruit) is separate from the combinations of meat and vegetable.

This analysis is interesting and makes sense. It is consistent with common knowledge that people have a different shopping list when they visit grocery stores. Mostly, they come here to purchase things to make a meal, such as vegetables and meat. Moreover, they sometimes visit the grocery stores only for fruits.

QUESTION 4

Use this training data (and this data alone) to build the model. Then apply your model to predict the authorship of the articles in the C50test directory, which is about the same size as the training set. Describe your data pre-processing and analysis pipeline in detail.

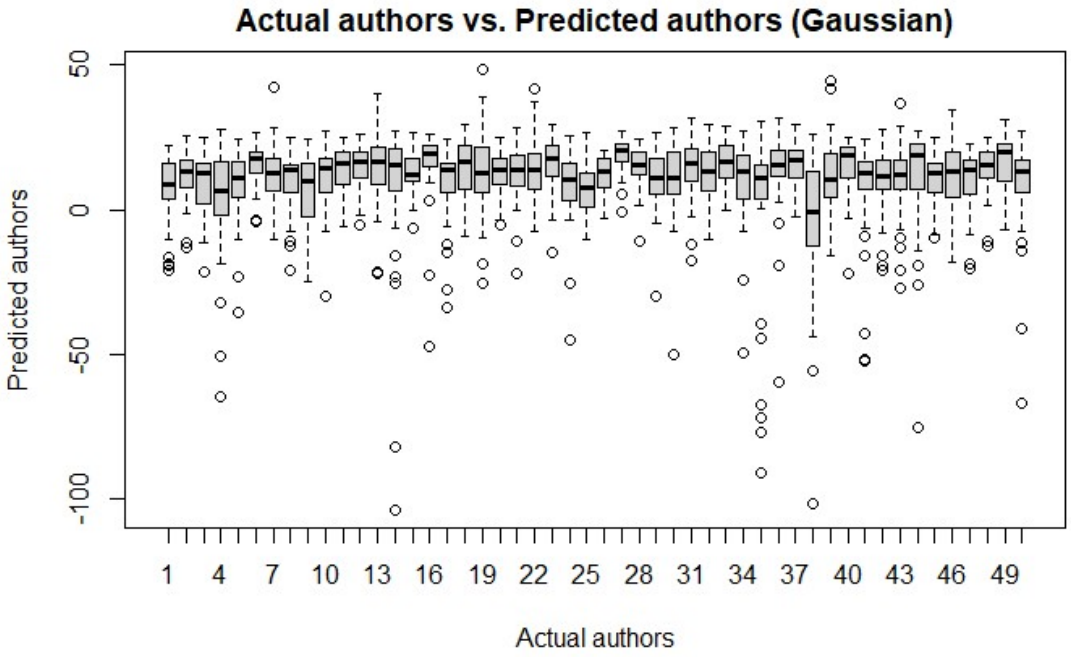
ANSWER

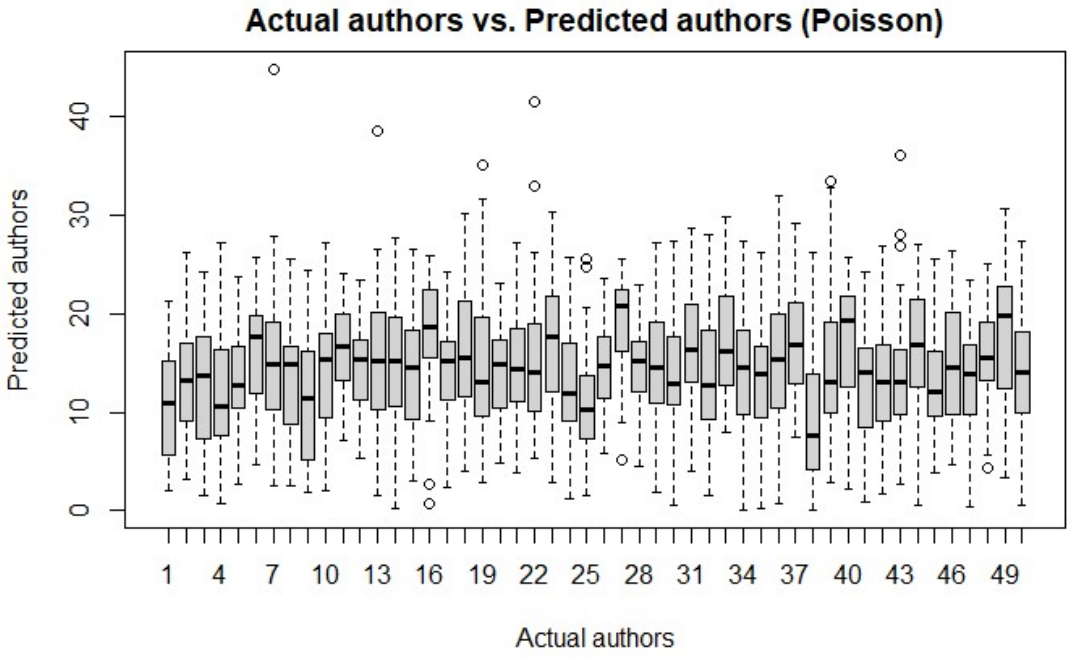
Note: the detailed process is within the Rmd file.

This study aimed to predict authors based on what they wrote. I applied the lasso logistic regression to achieve this purpose. The whole process included three steps. First, I applied a sort of steps to generate the training and test dataset choosing “SMART” as the stop-word. I used a function to script the names of fifty authors and get the texts what they wrote from the training dataset. Then I applied the same operations in the testing dataset. At last, I converted the texts to training and testing feature matrices. The outcomes of this step are two datasets, and both are lists of strings.

Then, I captured the names of authors from both datasets, separately as the train labels and test labels. I noticed that the lasso logistic regression cannot deal with the strings, so I created a table, convert the names to numbers. For instance, “AaronPressman” is 1, and “AlanCrosby” is 2. These numbers then were converted to factor variables. The outcome of this steps are two lists of factor variables as the training dependent variable and testing dependent variables in this study.

At last, I run the lasso logistic regression. I noticed that there were two families in this function, Gaussian and Poisson. I separately ran the regression choosing two families and compared the results. Figures 4-1 demonstrate the predictions versus actual numbers of the two models. Obviously, the Poisson model's performance is much better than the Gaussian model since the variances are much smaller. For instance, when predicting author “1”, the Gaussian model predicted it as numbers from -25 to 20, and the Poisson model predicted it as numbers from 2 to 20.





Figures 4-1. Results of Gaussian model and Poisson model