Matt Girard
CS 690V Visual Analytics
HW 4 (Food Frequency) Writeup

App Layout

The app is structured around 3 plots and a group of dropdowns. The first plot displays all of the original data colored by a binary label of your choice. The labels dropdown consists of the first 25 columns in the dataset which are simply Yes or No. The second plot displays the same data but colored based on the output of a classification model which the user can choose (SGDClassifier and LogisticRegression). The third plot is similar to the second, but displaying the data colored by a clustering algorithm (AgglomerativeClustering).

From the remaining columns, the user can choose to plot any two columns along the X and Y axis. All three plots will update upon selection. The columns selected will be fed into the classification and clustering algorithms (along with the selected label set for classification training).

Notable Predictors

- Diabetes seems to have correlation with GROUP_SWEETS_TOTAL_GRAMS and GROUP_YOGURT_PLAIN_NON_FAT_TOTAL_GRAMS which is not surprising. As both increase, so does the likelihood of a positive diabetes label.

- It seems that if you like watching TV (favCable) and eat lots of peanuts/nuts (GROUP_PEANUTS_OTHER_NUTS_SEEDS_TOTAL_GRAMS) you will have an increased risk of cancer.

- Ice cream sauce and tortilla consumption will increase your risk of heart disease.

Conclusion

While this dataset is extremely wide it also rather sparse. 54 records are hardly enough to build a reliable predictive model.