690V – Visual Analytics
Homework 4
Suhas Keshavamurthy

Question:
Look at dataset 2, a food frequency questionnaire (FFQ) with over 1000 variables for 54 individuals. Analyze the data. Build predictors. Explore your models using interactive visualizations. Express your predictors using sentences we can all understand (like having a cat and eating bagels often decreases your chances of having heart disease). Models you can try are clustering, correlation analyses, regression, ...

Data
https://github.com/SuhasKMurthy/Visual-Analytics/blob/master/HW4/test.csv

Visualization:
Github link –
https://github.com/SuhasKMurthy/Visual-Analytics/blob/master/HW4/vis2.py

How to run the file:

In a command terminal change to the appropriate directory
Then enter 'bokeh serve –show vis2.py' in the terminal
A page in the default browser should open with the visualization

The visualization contains selection for X-Axis and Y-Axis through dropdown.
The smoking habits and other miscellaneous criteria can be chosen through the radiobutton.

Press 'Plot' button to view the visualization.

The diseases are plotted through annular rings.

Black/Grey denotes no disease of that particular type.
Red/Blue/Yellow denotes Cancer/Diabetes/Heart disease respectively.

The K-Means clusters and centroids for '3' clusters are plotted along with this.

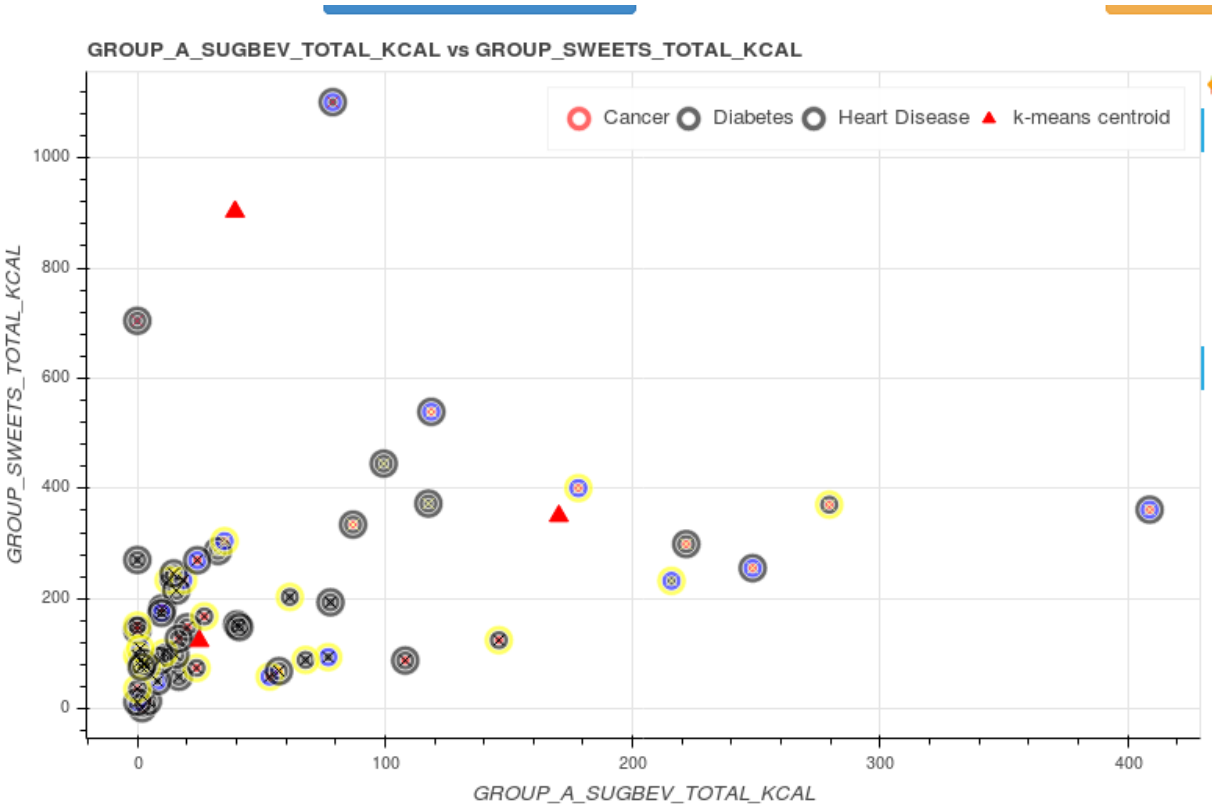
The legend items can be used to hide individual disease

Hover tool provides the coordinates of the K-means cluster centroids.

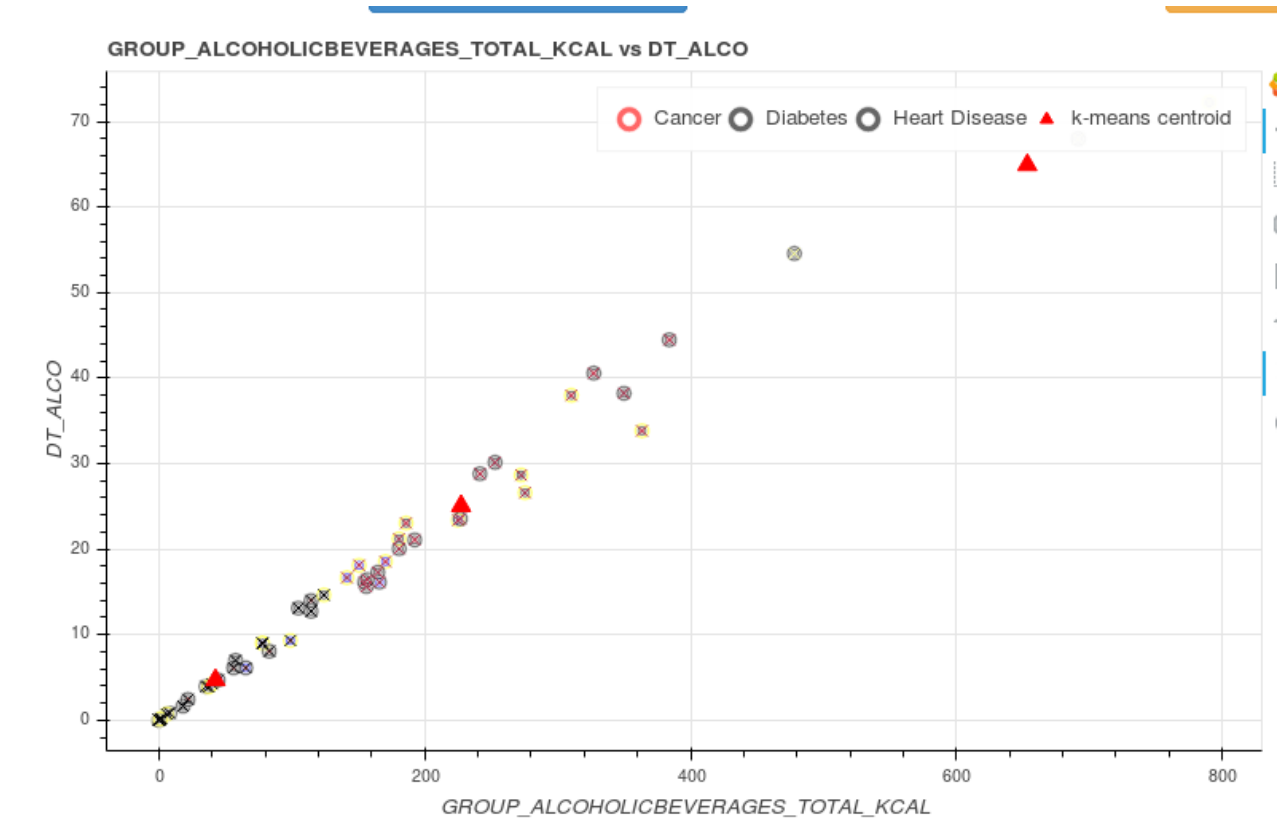
The data has been first sanitized manually. The original dataset which had more than 1000 columns is reduced by removing the sparse columns in the dataset. The new file which is used for the visualization is called test.csv

Some of the observations in the dataset are

There is a higher likelihood of contracting any of the disease (Cancer/diabetes/Heart Disease) as the sugar calories from beverages is consumed.



Alcohol consumed and calories are proportional (This is obvious but the columns in the dataset are cryptic). It also shows that there is no correlation between alcohol consumed and the possibility of any of the above mentioned disease.



The following graph show that the likelihood of diabetes is less if cholesterol is within a certain range.