Snigdha Petluru
spetluru@umd.edu

Processing Document to combine data from Historic Sites and Restaurants

This document describes the process of obtaining the combined file that contains data from historic sites and restaurants in Howard County, i.e., Historic_Sites_View.csv and HH_Restaurants.csv respectively. To combine these datasets, an R script was generated. Each line of code and its function is described below:

1. **setwd([*"Path to the folder that comprises the datasets"*])** – In R, set the working directory to the location where the files are present. This makes it simpler to store an output file which will remain with the remaining datasets.

2. **d=read.csv("Historic_Sites_View.csv") –** Store the contents of the *Historic_Sites_View.csv* file into the data frame *d*.

3. **c = read.csv("HH_Restaurants.csv") –** Store the contents of the *HH_Restaurants.csv* file into the data frame *c*.

4. **e=merge(x=d,y=c,by.x=c("FID","Name","Address","Location","geom"),by.y=c("FID","Name","Address","City","geom"),all="T") –** Here, we are merging both the data frames d and c into a new data frame *e*, such that all the columns within both datasets are preserved. x,y represent the two data frames that need to be merged. Here, by.x represents the columns in data frame d that are to be used as key variables that are common to data frame c. However, similar variables need not be named the same way. Hence, we provide a mapping of which common variables in d correspond to the variables in c. As can be seen, *FID, Name, Address* and *geom* are identical for both data frames. However, *location* and *city* are different column names for the same city data. Hence the *by.x()* and *by.y()* are used. It is essential to follow the same order of elements or the key variables in the *by.x()* and *by.y()* functions. Also, *all="T"* implies a *full outer join*, where all the cases that do not match are appended to the data frame too as NA. Alternatively, all="F" would have provided a natural join, all.x="T" would have provided a left outer join and all.y="T" would have obtained a right outer join.

5. **write.csv(e,"Combined_HistoricSites_Restaurants.csv") –** Write a csv file *"Combined_HistoricSites_Restaurants.csv"* comprising of the data frame e. This is the required file that combines both datasets.