#Get restaurant data from the site

if(!file.exists("/summaryWalkthrough")){dir.create("summaryWalkthrough")}

fileUrl <- "https://data.baltimorecity.gov/api/views/k5ry-ef3g/rows.csv?accessType=DOWNLOAD"

download.file(fileUrl,destfile="summaryWalkthrough/restaurants.csv",method="curl")

restData <- read.csv("summaryWalkthrough/restaurants.csv")

# Look at a bit of the data, get a summary

**head(restData,n=3)**

**tail(restData,n=3)**

**summary(restData)**

# for factor-based variables it will provide a count for each one (eg “McDonalds”:8). Not sure how it chooses which are factors. Seems to just be var-Names that are text instead of numbers.

#how big is the data set

**object.size(fakeData)**

#Get that value in something other than bytes

**print(object.size(fakeData),units="Mb")**

#More in-depth look using str(). Get object type, numrows(obs), cols(variables). The class of each variable (eg ‘Factor w/1277 levels, int)

**str(restData)**

#sample output

'data.frame': 1327 obs. of 6 variables:

$ name : Factor w/ 1277 levels "#1 CHINESE KITCHEN",..: 9 3 992 1 2 4 5 6 7 8 ...

$ zipCode : int 21206 21231 21224 21211 21223 21218 21205 21211 21205 21231 ...

#get quantile or probability information on the data

# for var ‘councileDistrict’ see quantiles

**quantile(restData$councilDistrict,na.rm=TRUE)**

# same var see probability of falling below a certain percent.

quantile(restData$councilDistrict,probs=c(0.5,0.75,0.9))

50% 75% 90%

9 11 12

#By councilDist #9, you’d seen 50% of the districts, by #12 you’d seen 90% of them

#Make a table of a particular variable to see counts of occurreneces for each element. Displays each element of that variable with a count of it’s occureneces beneath it. eg:

1234

14

#zipcode 1234 occurs 14 times in the zipCode var column.

**table(restData$zipCode,useNA="ifany")**

#if there are any missing values, include a count of all of them. Eg Na:37

#Make a 2-dimensional table to get look at variable intersection occurences

**table(restData$councilDistrict,restData$zipCode)**

councilDistrict #’s are row labels, zipCode values are column labels

#Check for missing values (sum,any,all)

#what is the number of NA values for a particular column/variable

**sum(is.na(restData$councilDistrict))**

#Are there any NAs for a particular column/variable

**any(is.na(restData$councilDistrict))**

#Is it true that all values for a particular column/var are greater than some number

**all(restData$zipCode > 0)**

#Are there any NAs in any of my columns, if so, how many in each

**colSums(is.na(restData))**

#If have many columns, can verify that there are no missing values in a data set

**all(colSums(is.na(restData))==0)**

#search for specific values (like == , but can pass a list). How many occurences are there of any of these values

**table(restData$zipCode %in% c("21212","21213"))**

#You can use this to subset your data based on only rows that have desired col values

**restData[restData$zipCode %in% c("21212","21213"),]**

#CROSS TABS

Cross tabs

data(UCBAdmissions)

DF = as.data.frame(UCBAdmissions)

summary(DF)

**xt <- xtabs(Freq ~ Gender + Admit,data=DF)**

Admit

Gender Admitted Rejected

Male 1198 1493

Female 557 1278