Untitled

Taylor Rasley

August 17, 2019

```
library(tidyverse)
library(forcats)
library(RMySQL);
library(ggplot2);
library(DBI);
library("dplyr")
```

Read in files:

```
datapath<-"C:/Users/Taylor/Documents/MSCA/Data Engineering/Final Project/Data Sets"

#Latest business data
License_Dat<-(read.csv(file=paste(datapath,"Business_Licenses_-_Current_Active.csv",sep="/"),hea
der=TRUE,sep=",",as.is=c(10) ))
#License Codes of interest, both good and bad
License_Codes<-(read.csv(file=paste(datapath,"Business License Codes to use.csv",sep="/"),header
=TRUE,sep=","))
#Chicago Zip Codes
Zipcode_values<-(read.csv(file=paste(datapath,"Zipcode Values.csv",sep="/"),header=TRUE,sep="
))
#Zip Code populations
zip_pop<-(read.csv(file=paste(datapath,"PopTable.csv",sep="/"),header=TRUE,sep=","))
```

Set zip-code field to integer

```
License_Dat$ZIP.CODE<-as.integer((License_Dat$ZIP.CODE))
```

create good & bad business dataframes

```
good_license<-filter(License_Codes,Status=="Good")
bad_license<-filter(License_Codes,Status=="Bad")
```

Trim business license files to include only licenses of interest

```
#get business licenses from list
Valid_Licenses<-semi_join(License_Dat,License_Codes,by = c("LICENSE.CODE" = "License.Code"))
```

Calculate business score

```
#Group Licenses by zipcode, license code
zip_license_count<-Valid_Licenses %>% count(ZIP.CODE,LICENSE.CODE)
#can also use:zip_license_count<-summarise(group_by(Valid_Licenses1,ZIP.CODE,LICENSE.CODE),count
=n())

#add column for score for zip, license code groups
zip_license_count$score <- NA

#Create score by zip code/license group
for (i in 1:length(zip_license_count$LICENSE.CODE)) {
      if (zip_license_count$LICENSE.CODE[i] %in%  good_license$License.Code) {
        zip_license_count$score[i]<-zip_license_count$n[i]
      }
      else if (zip_license_count$LICENSE.CODE[i] %in%  bad_license$License.Code) {
          zip_license_count$score[i]<--5*zip_license_count$n[i]
        }

}

#aggregate scores from zip code/license code to just zip code level
zipcode_scores<-aggregate(zip_license_count$score, by=list(Zip_code=zip_license_count$ZIP.CODE),
FUN=sum)

#keep just Chicago zip codes
Zip_Scores_Chicago<-semi_join(zipcode_scores,Zipcode_values,by = c("Zip_code" = "zip_code"))

#normalize by zipcode population
library(purrr)
Zip_Scores_Chicago_wPop<-inner_join(Zip_Scores_Chicago,zip_pop,by=c("Zip_code"="zip_code"))
Zip_Scores_Chicago_wPop$x<-Zip_Scores_Chicago_wPop$x/Zip_Scores_Chicago_wPop$population*100

#Create output file
License_Dat_Out<-select(Zip_Scores_Chicago_wPop,Zip_code,x)
write.table(License_Dat_Out,
            file = "C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/Zip Code Biz License Scores.cs
v",row.names=FALSE,col.names=c("Zip Code","Score"), na="", sep=",")
```