## <<<<<<< HEAD

title: "Pokemon Project R notebook" output: html\_notebook --- The first part is the code about dataset integration and data cleaning.

#read the file  
pokemonRaw <- read.csv("C:\\Users\\liu23\\Documents\\GitHub\\INST737\\dataset\\300k.csv")  
rarity <- read.csv("C:\\Users\\liu23\\Documents\\GitHub\\INST737\\dataset\\pokemonRairty.csv")  
type <-read.csv("C:\\Users\\liu23\\Documents\\GitHub\\INST737\\dataset\\pokemonGOInfo.csv")  
  
#delete all "co-occurance" features and city features  
pokemon=pokemonRaw[c(1:21,23:56,208)]  
  
#integra these three files  
pokemon123=merge.data.frame(pokemon,rarity,by.x="pokemonId",by.y="PID")  
pokemon456=merge.data.frame(pokemon123,type,by.x="pokemonId",by.y="PID")  
  
#clean dupilicate varibles caused by merging and output the dataframe as a csv file  
library(plyr)  
pokemon789=pokemon456[c(1:55,57,58,60,64)]  
pokemonClean=rename(pokemon789,c("Name.x"="Name","Type.1"="Type"))  
  
#comebine four features, "urban","suburban","midurban","rural", as one categorical variable  
pokemonClean$urbanStatus=c(1:nrow(pokemonClean))  
  
for (i in 1:nrow(pokemonClean)){   
 if(pokemonClean[i,39] == "TRUE")  
 {  
 pokemonClean[i,61]= "urban"  
   
 }  
 else if(pokemonClean[i,40] == "TRUE")  
 {  
 pokemonClean[i,61]= "suburban"  
 }  
 else if(pokemonClean[i,41] == "TRUE")  
 {  
 pokemonClean[i,61]= "midurban"  
 }  
 else if(pokemonClean[i,42] == "TRUE")  
 {  
 pokemonClean[i,61]= "rural"  
 }  
}  
  
  
 write.csv(pokemonClean, file="pokemonCleaned.csv")  
 #from the above csv file, we also use excel to split the feature "appearedLocalTime"

The second part is the code about analysis

#by aggregating classID to check whether there are suspicious pokemon that should not be recorded in the dataset  
#also check the distribution of the whole pokemon records and   
library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

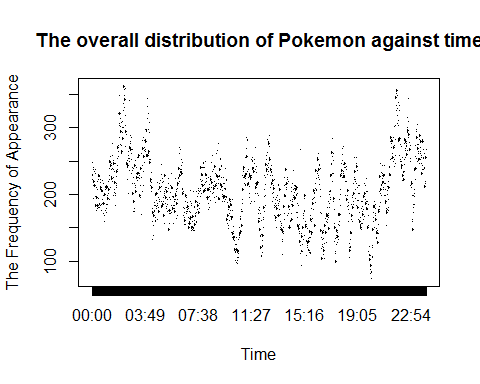
## Loading required package: RSQLite

pokemonClean <- read.csv("C:\\Users\\liu23\\Documents\\GitHub\\INST737\\dataset\\pokemonCleaned.csv")  
m.pokemonClean=as.matrix(pokemonClean)  
classAggregation=sqldf("SELECT pokemonId, Name, count(pokemonId) FROM pokemonClean GROUP BY pokemonId")

## Loading required package: tcltk

## Warning: Quoted identifiers should have class SQL, use DBI::SQL() if the  
## caller performs the quoting.

#check the overall pokemon distribution against time  
  
timeAggregation=sqldf("SELECT LocalTimeMin,count(LocalTimeMin) as number FROM pokemonClean GROUP BY localTimeMin")  
plot((timeAggregation$LocalTimeMin),timeAggregation$number,type="l",xlab="Time",ylab="The Frequency of Appearance",main="The overall distribution of Pokemon against time")



#check the Pidgey distribution against time  
timeAggregation.ID16=sqldf("SELECT LocalTimeMin,count(LocalTimeMin) as number FROM pokemonClean WHERE pokemonID=16 GROUP BY localTimeMin")  
plot((timeAggregation.ID16$LocalTimeMin),timeAggregation.ID16$number,type="l",xlab="Time",ylab="The Frequency of Appearance",main="The distribution of Pidgey against time")  
  
#check the relationship between terrain type that pokemon appears and pokemon type by using chi-square  
obs=table(pokemonClean$Type,pokemonClean$terrainType)  
chisq=chisq.test(obs)

## Warning in chisq.test(obs): Chi-squared approximation may be incorrect

chisq

##   
## Pearson's Chi-squared test  
##   
## data: obs  
## X-squared = 34776, df = 182, p-value < 2.2e-16

#using ANOVA to check the mean differences of pokemon in urban, suburban, midurban and rural area  
obs1=table(pokemonClean$urbanStatus,pokemonClean$Name)  
chisq1=chisq.test(obs1)

## Warning in chisq.test(obs1): Chi-squared approximation may be incorrect

chisq1

##   
## Pearson's Chi-squared test  
##   
## data: obs1  
## X-squared = 6496.2, df = 429, p-value < 2.2e-16

#By using chi-square check the mean differences in rarity groups  
obs2=table(pokemonClean$Rairty,pokemonClean$Name)  
chisq2=chisq.test(obs2)

## Warning in chisq.test(obs2): Chi-squared approximation may be incorrect

chisq2

##   
## Pearson's Chi-squared test  
##   
## data: obs2  
## X-squared = 1184100, df = 572, p-value < 2.2e-16

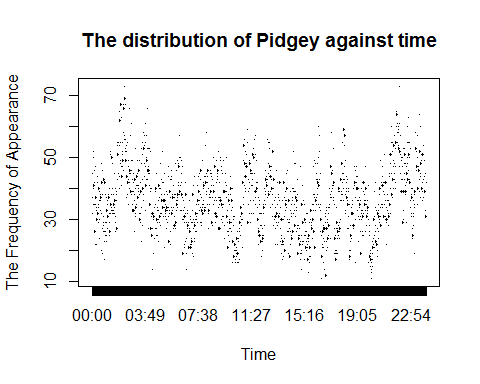
#By using chi square to check the mean differences in type groups  
obs3=table(pokemonClean$Type,pokemonClean$Name)  
chisq3=chisq.test(obs3)

## Warning in chisq.test(obs3): Chi-squared approximation may be incorrect

chisq3

##   
## Pearson's Chi-squared test  
##   
## data: obs3  
## X-squared = 4144300, df = 2002, p-value < 2.2e-16

#check the distribution of pokemon type against long/lat  
library(ggplot2)



g.type=ggplot(pokemonClean,aes(pokemonClean$longitude,pokemonClean$latitude))  
g.type+geom\_point(aes(colour=pokemonClean$Type))

