utility.R

poojasingh

Sun Nov 29 14:50:57 2015

```
suppressWarnings(suppressMessages(library(ggplot2)))

#Contains some utility R -cdoe
#fname = file.choose()
fname = "/Users/poojasingh/Documents/HStatE139/git/statse139-project2/statse139-project/Previous Boston
dfm = read.csv(fname, header=T,sep=" ")
#names(dfm)

#1. Add all times
times = as.matrix(dfm[,7:15], ncol=9)
dfm$totaltime = rowSums(times)

#2. Create factor for gender
#factor(dfm$Gender1F2M, ordered=TRUE) #identify sorted levels in input source
dfm$gender = factor(dfm$Gender1F2M, labels=c("female", "male"))
#levels(dfm$gender) #output levels

#3. Create factor for agegroup
#factor(dfm$Age, ordered=TRUE)  #identify sorted levels in input source
dfm$agrgroup = cut(dfm$Age, breaks=c(15, 25, 35, 45, 55, 65, 75, 85))
#levels(dfm$agrgroup)

#4. Split data by year
dfm.2010 = subset(dfm, (dfm$Year==2010))
dfm.2011 = subset(dfm, (dfm$Year==2011))
dfm.2013 = subset(dfm, (dfm$Year==2013))

#Sanity test for year subset
#length(dfm.2010$BibNum) + length(dfm.2011$BibNum) + length(dfm.2013$BibNum)

#5. Split data by agegroup
dfm.age.18.34 = subset(dfm, (dfm$Age>=18 & dfm$Age<=34))
dfm.age.35.39 = subset(dfm, (dfm$Age>=35 & dfm$Age<=39))
dfm.age.40.44 = subset(dfm, (dfm$Age>=40 & dfm$Age<=44))
dfm.age.45.49 = subset(dfm, (dfm$Age>=45 & dfm$Age<=49))
dfm.age.50.54 = subset(dfm, (dfm$Age>=50 & dfm$Age<=54))
dfm.age.55.59 = subset(dfm, (dfm$Age>=55 & dfm$Age<=59))
dfm.age.60.64 = subset(dfm, (dfm$Age>=60 & dfm$Age<=64))
dfm.age.65.69 = subset(dfm, (dfm$Age>=65 & dfm$Age<=69))
dfm.age.70.74 = subset(dfm, (dfm$Age>=70 & dfm$Age<=74))
dfm.age.75.79 = subset(dfm, (dfm$Age>=75 & dfm$Age<=79))
dfm.age.80.plus = subset(dfm, (dfm$Age>=80))

#Sanity Test for agegroup subgroups - the sum of lengths of subset should equal to length of original d
#length(dfm.age.18.34$BibNum) + length(dfm.age.35.39$BibNum) + length(dfm.age.40.44$BibNum) + length(df
#  length(dfm.age.50.54$BibNum) + length(dfm.age.55.59$BibNum) + length(dfm.age.60.64$BibNum) + length(
#  length(dfm.age.70.74$BibNum) + length(dfm.age.75.79$BibNum) + length(dfm.age.80.plus$BibNum)
```

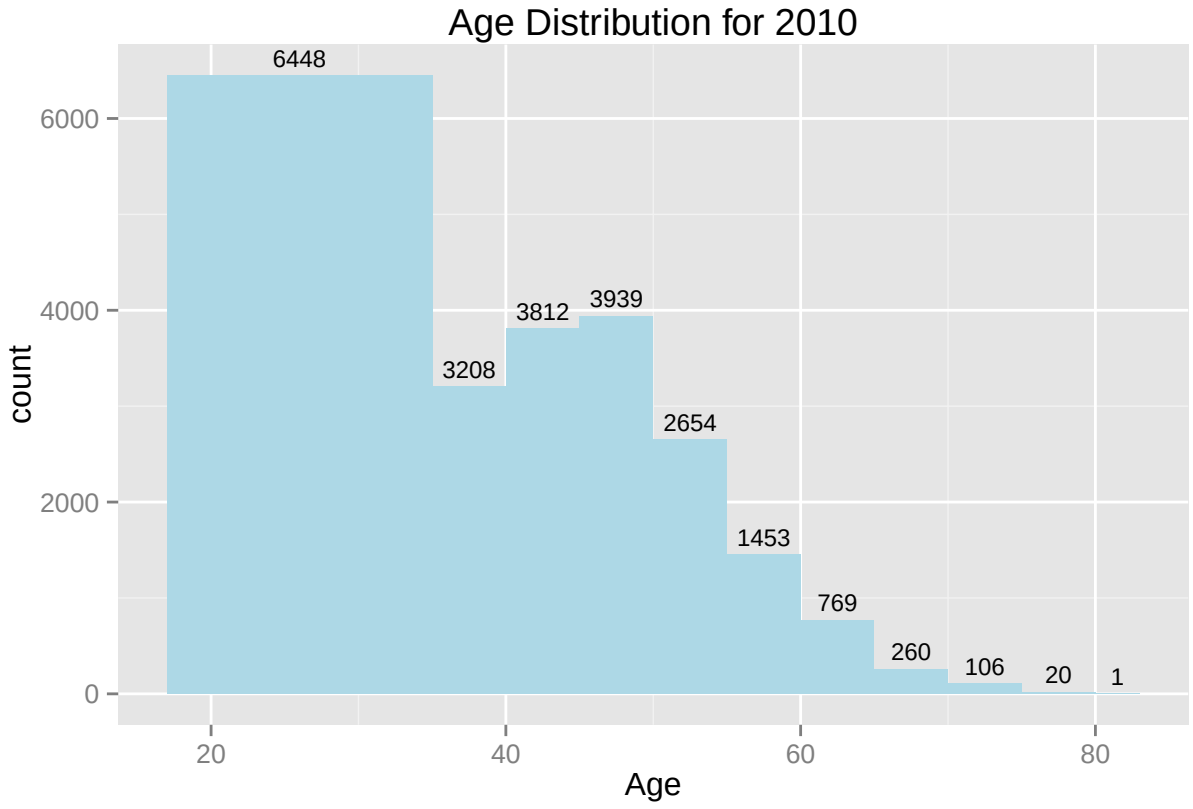
```
#length(dfm$BibNum)



#Plot histogram
#hist(dfm.2010$Age, breaks=c(18, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 83),
#       main="Age Distribution for 2010", xlab="Age", freq=TRUE)

#Plot histogram
plot1 = ggplot(dfm.2010, aes(x=Age)) +
  stat_bin(breaks=c(17, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 83), geom="bar", fill="lightblue") +
  stat_bin(breaks=c(17, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 83), geom="text", cex=3.25, aes(label=.
  labs(title="Age Distribution for 2010")

suppressWarnings(suppressMessages(print(plot1)))
```




```
#Sanity test for histogram count labels
#length(dfm.age.18.34$BibNum[dfm.age.18.34$Year==2010])
#length(dfm.age.75.79$BibNum[dfm.age.75.79$Year==2010])
```