非寿险精算大作业一: 数据汇总和总平均费率厘定

首先利用下面的代码对原始数据进行数据清理.

```
library(data.table)
# Import data. You need to specify your own directory
userlist<-fread("ubi_userlist.txt")
policydata<-fread("ubi_policydata.txt")
claimdata<-fread("ubi_claimdata.txt")
lct <- Sys.getlocale("LC_TIME"); Sys.setlocale("LC_TIME", "C")
    #Solve NA issue in as.Date function
policydata$UNDERWRITEENDDATE<-as.Date(policydata$
    UNDERWRITEENDDATE,format="%d%b%Y:%H:%M:%S")
policydata$KINDSTARTDATE<-as.Date(policydata$KINDSTARTDATE,
    format="%d%b%Y:%H:%M:%S")
policydata$DUEENDDATE<-as.Date(policydata$DUEENDDATE,format="%d
    %b%Y:%H:%M:%S")
claimdata$DAMAGEDATE<-as.Date(claimdata$DAMAGEDATE,format="%d%b
    %Y:%H:%M:%S")
claimdata$REPORTDATE<-as.Date(claimdata$REPORTDATE,format="%d%b
    %Y:%H:%M:%S")
claimdata$ENDCASEDATE<-as.Date(claimdata$ENDCASEDATE,format="%d
    %b%Y:%H:%M:%S")
# Policydata cleaning
policydata<-data.frame(policydata,YEARS=rep(NA,nrow(policydata)
    ),Device_ID=rep(userlist$PN[1],nrow(policydata))) #YEARS is
    the (expected) ultimate earned years; Device_ID is the ID of
     each vehicle
for (i in 1:nrow(policydata)){
policydata$YEARS[i]<-ifelse(as.numeric(difftime(policydata$
    DUEENDDATE[i],policydata$KINDSTARTDATE[i]))+1==366,365,as.
    numeric(difftime(policydata$DUEENDDATE[i],policydata$
    KINDSTARTDATE[i]))+1)/365
policydata$Device_ID[i]<-userlist$PN[userlist$policycode==
    policydata$policycode[i]]
}
policydata<-policydata[-which(policydata$YEARS>2),]
policydata$AGE[policydata$AGE<18|policydata$AGE>100]<-NA
policydata<-policydata[-which(is.na(policydata$AGE)),]
policydata<-policydata[-which(policydata$PREMIUM==0),]
```

```
#Claimdata cleaning
claimdata<-data.frame(claimdata,Delay_report=rep(NA,nrow(claimdata)),Delay_pay=rep(NA,nrow(claimdata)),Device_ID=rep(userlist$PN[1],nrow(claimdata))) #Delay_report is the report delay in days; Delay_pay is the settlement delay in days.
for (i in 1:nrow(claimdata)){
claimdata$Delay_report[i]<-as.numeric(difftime(claimdata$REPORTDATE[i],claimdata$DAMAGEDATE[i],units="days"))
claimdata$Delay_pay[i]<-as.numeric(difftime(claimdata$ENDCASEDATE[i],claimdata$DAMAGEDATE[i],units="days"))
claimdata$Device_ID[i]<-userlist$PN[userlist$policycode==claimdata$policycode[i]]
}
set_diff<-setdiff(unique(claimdata$policycode),unique(policydata$policycode))
print(set_diff) #Two policies in claimdata have been removed from policydata in the data error cleaning procedure.
claimdata<-claimdata[-which(claimdata$policycode%in%set_diff),]
claimdata[is.na(claimdata$ENDCASEDATE),] #Open cases
```

问题:

1. (10 分) 对于交强险和车损险, 按**日历年** 2014, 2015, 2016, 分别汇总其已赚风险单位数和已赚保费.
2. (10 分) 对于交强险和车损险, 分别计算它们的**报案延迟**和**结案延迟**的中位数, 均值, 和 95% 分位数. 基于以上的信息, 在按**事故年**汇总最终索赔次数和最终索赔金额时, 需要注意什么?
3. (10 分) 对于交强险和车损险, 按**事故年** 2014, 2015, 2016, 分别汇总其最终赔款. (注: 不需要使用流量三角形.)
4. (10 分) 对于交强险和车损险, 计算其在经验期 2014, 2015, 2016 年的纯保费和赔付率.
5. (10 分) 对于车损险, 使用 2016 年的纯保费数据进行总平均费率厘定. 假设未来纯保费的增长趋势为 1%(每年), 每个风险单位的固定费用为 200 元, 变动费用为 15%, 利润附加为 5%. 新费率拟在 2020 年 1 月 1 日生效, 所有保单均为一年期保单. 请计算 2020 保单年车损险的平均费率.

注意: 请用文字和数据回答以上问题, 不能直接粘贴 R 的输出结果. 请把相关代码作为附录.

提示: 可使用以下函数计算已到期风险单位数.

```
earned<-function(start,finish,endyear){
#start is the effective date, finish is the expiration date, endyear is the end date of the target calendar year
if(as.numeric(difftime(endyear,start))>0&as.numeric(difftime(endyear,start))<=365) exposure<- min(as.numeric(difftime(endyear,start)),as.numeric(difftime(finish,start)))+1
else if(as.numeric(difftime(endyear,start))<=0) exposure<-0
else if(as.numeric(difftime(endyear,start))>365) exposure<-max(0,365-as.numeric(difftime(endyear,finish)))
exposure
}
earned("2015-4-1","2016-3-31","2014-12-31")/365
earned("2015-4-1","2016-3-31","2015-12-31")/365
earned("2015-4-1","2016-3-31","2016-12-31")/365
```