setwd('C:\\Users\\USER-PC\\Desktop\\Lpoint')

library(dplyr)

library(data.table)

library(ggplot2)

library(googleVis)

library(lubridate)

load(file="product.RData")

load(file="customer.RData")

load(file="master.RData")

load(file="search2.RData")

load(file="search1.RData")

load(file="session.RData")

#대분류 더욱 크게 구분

sport<-c("시즌스포츠","아웃도어/레저","헬스/피트니스","구기/필드스포츠","구기/필드스포츠")

for(i in seq\_along(master)){

master[[i]][master[[i]] %in% sport]<-"스포츠용품"

}

elec<-c("계절가전","냉장/세탁가전","생활/주방가전","영상/음향가전","컴퓨터","모바일")

for(i in seq\_along(master)){

master[[i]][master[[i]] %in% elec]<-"전자기기"

}

sundries<-c("식기/조리기구","주방잡화","출산/육아용품","패션잡화","퍼스널케어","청소/세탁/욕실용품","세제/위생","화장품/뷰티케어")

for(i in seq\_along(master)){

master[[i]][master[[i]] %in% sundries]<-"잡화"

}

eat<-c("건강식품","과일","냉동식품","냉장식품","음료","축산물")

for(i in seq\_along(master)){

master[[i]][master[[i]] %in% eat]<-"식품"

}

home<-c("인테리어/조명","침구/수예","가구")

for(i in seq\_along(master)){

master[[i]][master[[i]] %in% home]<-"가정"

}

etc<-c("자동차용품","원예/애완","문구/사무용품","완구","상품권")

for(i in seq\_along(master)){

master[[i]][master[[i]] %in% etc]<-"기타"

}

#product,master,session,customer 합치기

procust<-inner\_join(product,customer,by=c("CLNT\_ID"="CLNT\_ID"))

master$PD\_C<-as.integer(master$PD\_C)

procustmaster<-inner\_join(procust,master,by=c("PD\_C"="PD\_C"))

procustmaster<-procustmaster[-c(23015,26497),] #결측값 없애기

prosesscustmaster<-inner\_join(procustmaster,session,by=c("CLNT\_ID","SESS\_ID"="CLNT\_ID","SESS\_ID"))

sum(is.na(prosesscustmaster)) #TOT\_PAG\_VIEW\_CT의 결측값 확인후 없애주기

prosesscustmaster<-prosesscustmaster[complete.cases(prosesscustmaster[,c("TOT\_PAG\_VIEW\_CT")]),]

#가격에 ,빼고 Integer형식으로 바꾸기기

prosesscustmaster$PD\_BUY\_AM<-gsub(",","",prosesscustmaster$PD\_BUY\_AM)

prosesscustmaster$PD\_BUY\_AM<-as.integer(prosesscustmaster$PD\_BUY\_AM)

prosesscustmaster$PD\_BUY\_CT<-as.integer(prosesscustmaster$PD\_BUY\_CT)

#날짜데이터형식으로 변경

prosesscustmaster$SESS\_DT<-format(prosesscustmaster$SESS\_DT,format="%Y%m%d")

prosesscustmaster$SESS\_DT<-as.Date(prosesscustmaster$SESS\_DT,format="%Y%m%d")

#나이,성별 factor로 변경

prosesscustmaster$CLNT\_AGE<-as.factor(prosesscustmaster$CLNT\_AGE)

prosesscustmaster$CLNT\_GENDER<-as.factor(prosesscustmaster$CLNT\_GENDER)

save(prosesscustmaster,file="prosesscustmaster.RData")

load(file="prosesscustmaster.RData")

#연령대별 구매가격 평균구하기

price<-procustmaster$PD\_BUY\_AM

count<-procustmaster$PD\_BUY\_CT

age<-procustmaster$CLNT\_AGE

agebuy<-data.frame(age,price,count)

agebuy<-within(agebuy,buyprice<-price\*count)

(table(agebuy$age,agebuy$buyprice))

#연령대별 평균구매액 18page

(tapply(agebuy$buyprice,agebuy$age,mean))

#연령대별 구매빈도 17page

(tapply(agebuy$count,agebuy$age,sum))

#분산분석,튜키검정 - CLNT\_AGE-buyprice

price<-prosesscustmaster$PD\_BUY\_AM

count<-prosesscustmaster$PD\_BUY\_CT

age<-prosesscustmaster$CLNT\_AGE

(tdata2<-data.frame(age,price,count))

(tdata2<-within(tdata2,buyprice<-price\*count))

tdata2$age<-as.factor(tdata2$age)

tapply(tdata2$buyprice,tdata2$age,summary)

#일원분산분석 20page

aov(buyprice~age,data=tdata2)

summary(aov(buyprice~age,data=tdata2))

bartlett.test(buyprice~age,data=tdata2)

#튜키검정 21page

group\_aov<-aov(buyprice~age,data=tdata2)

TukeyHSD(group\_aov)

#50,60대 구매이용 기기빈도 35page

dvc<-prosesscustmaster$DVC\_CTG\_NM

age<-prosesscustmaster$CLNT\_AGE

da<-data.frame(dvc,age)

(da<-subset(da,age>=50 & age<=60))

round(prop.table(table(da$dvc,da$age))\*100,1)

#5060대 품목별 월별 빈도수

pdc<-prosesscustmaster$CLAC1\_NM

age<-prosesscustmaster$CLNT\_AGE

buyct<-prosesscustmaster$PD\_BUY\_CT

month<-month(prosesscustmaster$SESS\_DT)

adata<-data.frame(month,age,pdc,buyct)

(fsdata<-subset(adata,age>=50 & age<=60))

round(prop.table(table(fsdata$pdc))\*100,1)

#5060대 월별구매빈도 23page

round(prop.table(table(fsdata$month))\*100,1)

#5060대 월별상품군구매빈도 25page

round(prop.table(table(fsdata$month,fsdata$pdc))\*100,1)

#5060대 상품군 구매빈도 24page

round(prop.table(table(fsdata$pdc))\*100,1)

#모든 연령대 상품군 구매빈도 24page

round(prop.table(table(adata$pdc))\*100,1)

#회귀분석

chu<-as.Date(c("20180913"),format="%Y%m%d")

day<-prosesscustmaster$SESS\_DT

age<-prosesscustmaster$CLNT\_AGE

buyprice<-prosesscustmaster$PD\_BUY\_AM\*prosesscustmaster$PD\_BUY\_CT

(rdata<-data.frame(day,age,buyprice))

(bdcast<-dcast(rdata,day~.,value.var="buyprice",sum))

(adcast<-dcast(rdata,day~age,value.var="age",length))

#구매금액, 5060대고객수, 추석까지남은날짜로 이루어진 데이터프레임만들기

(rdatat<-data.frame(bdcast$day,bdcast$.))

(rdatat<-within(rdatat,"5060"<-adcast$'50'+adcast$'60'))

(rdatat<-within(rdatat,chuday<-chu-bdcast.day))

#외부데이터중 온도데이터 넣기

temp<-fread("temp.csv")

save(temp,file="temp.RData")

load(file="temp.RData")

temp$일시<-as.Date(temp$일시,format="%Y%m%d")

rdatatt<-cbind(rdatat,temp)

rdatatt<-rdatatt[,-5]

#공휴일 +-3일에 질적변수에 "1"값 넣고 아닌날들에 "0"값 넣기

(rdatatt<-within(rdatatt,holicheck<-"0"))

rdatatt[c(33:39,49:55,64:77,134:140,163:169),6]<-"1"

names(rdatatt)<-c("day","buyprice","fscust","chudday","temp","holicheck")

rdatatt$holicheck<-as.factor(rdatatt$holicheck)

rownames(rdatatt)<-rdatatt$day

rdatatt<-rdatatt[,-1]

rdatatt$chudday<-as.integer(rdatatt$chudday)

rdatatt$temp<-as.numeric(rdatatt$temp)

#데이터 표준화 28page

srdata=rdatatt

for(i in 2:5){

srdata[,i]=scale(srdata[,i])

}

(srdata)

#회귀분석 29page

(reg1<-lm(buyprice~.,data=srdata))

summary(reg1)

#유의하지않았던 holicheck변수 삭제,

#이상값인 4월1일 4월2일 데이터 삭제 후 회귀분석 30,31page

rregt<-srdata[,-c(5)]

rregt<-rregt[-c(1:2),]

pairs.panels(rregt)

reg2<-lm(buyprice~.,data=rregt2)

summary(reg2) #32page

par(mfrow=c(2,2))

plot(reg2) #32page

par(mfrow=c(1,1))

#표준화전 회귀계수로 예측하기 33page

regrd<-rdatatt[,-c(5)]

regrd<-regrd[-c(1:2),]

regr<-lm(buyprice~.,data=regrd)

(lapply(regrd,mean))

#신뢰구간

predict(regr,data.frame("fscust"=3000,"chudday"=73,"temp"=21.3612),interval="confidence")

#예측구간

predict(regr,data.frame("fscust"=3000,"chudday"=73,"temp"=21.3612),interval="prediction")