R Programming Code

```
psidClean_kid = psid[psid$kid<20,]
nrow(psidClean_kid)
psidClean_education = psidClean_kid[psidClean_kid$educatn < 20,]


table(PSIDData)
hist(PSIDData$age) ## which category age peopoe are higher ?

plot(PSIDData$kids, type="o", col="blue", ylim=c(0,12))
lines(PSIDData$married, type="o", pch=22, lty=2, col="red")
title(main="Autos", col.main="red", font.main=4)

barplot(PSIDData$kids,
        main = "Maximum Temperatures in a Week",
        xlab = "Education",
        ylab = "Married",
        names.arg = c(PSIDData$married),
        col = "darkred")

findInterval(PSIDData$age, c(30, 40, 50))

d= density(PSIDData$hours)
plot(d, cex.main = 0.75)

plot(density(PSIDData$age))
plot(density(PSIDData$educatn))
plot(density(PSIDData$married))

kc = kmeans(PSIDData[,2:4] , 2)
kc
par(mfrow=c(1 , 2))
plot(PSIDData[,4:6] , col=kc$cluster)
points(kc$centers[,1:2], col=1:2 , pch=8 , cex=2)




PSID = read.csv("PSID.csv", header =TRUE )
hist(PSID$earnings,cex.main=.75)
par(mfrow=c(1,2))
plot(PSID$educatn,PSID$earnings)
par(mfrow=c(2,4))
plot(PSID$educatn,PSID$earnings)
par(mfrow=c(1,2))
```

```
plot(PSID$hours,PSID$earnings)
plot(PSID)
kc=kmeans(PSID[,2:4],2)
par(mfcol=c(,2))
par(mfcol=c(1,2))
plot(PSID[,5:6])
par(mfcol=c(10,20))
plot(PSID[,5:6])
plot(PSID[,5:6],col=kc$cluster,cex.main=0.75)
points(kc$centers[,1:2],col=5:7,pch=8,cex=2)
plot(PSID[,6:9])
par(mfrow=c(1,2))
plot(PSID$age,PSID$earnings)
plot(PSID[,5:6],col=kc$cluster,cex.main=0.75)
points(kc$centers[,1:2],col=1:2,pch=8,cex=2)
table(PSID$earnings,kc$cluster)
par(mfcol=c(1,2))
plot(hcBrainAve,hang=-1,cex.main=.75,cex.axis=.5)
rect.hclust(hcBrainAve,k=2,border="green")
dis=dis(PSID[2:4],method="euclidean")
dis=dist(PSID[2:4],method="euclidean")
hcBrainAve=hclust(dis,method = "ave")
hcBrainWard=hclust(dis,method = "ward.D")
par(mfcol=c(1,2))
plot(hcBrainAve,hang=-1,cex.main=.75,cex.axis=.5)
rect.hclust(hcBrainAve,k=2,border="green")
dis=dist(PSID[5:6],method="euclidean")
hcBrainAve=hclust(dis,method = "ave")
hcBrainWard=hclust(dis,method = "ward.D")
par(mfcol=c(1,2))
plot(hcBrainAve,hang=-1,cex.main=.75,cex.axis=.5)
rect.hclust(hcBrainAve,k=2,border="green")
dis=dist(PSID[5:6],method="euclidean")
hcBrainAveCut=cutree(hcBrainAve,2)
hcBrainAveCut
hcBrainWardCut=cutree(hcBrainWard,2)
par(mfcol=c(1,2))
plot(PSID[,5:6],col=hcBrainAveCut,cex.main=.75)
plot(PSID[,5:6],col=hcBrainWardCut,cex.main=.75)
par(mfrow=c(1,2))
plot(PSID$educatn,PSID$hours)
pie(PSID$kids)
par(mfrow=c(1,2))
plot(PSID$educatn,PSID$hours)
par(mfrow=c(1,2))
plot(PSID$educatn,PSID$hours)$out
```

```
par(mfrow=c(1,2))
plot(PSID$educatn,PSID$hours,plot=flase)$out
print(outliers)
outliers <- boxplot(PSID$educatn,PSID$hours, plot=FALSE)$out
print(outliers)
PSID[which(PSID$educatn,PSID$hours%in% outliers),]
plot(PSID$earnings)
plot(PSID$earnings)$out
outliers <- boxplot(PSID$earnings, plot=FALSE)$out
print(outliers)
PSID[which(PSID$earnings%in% outliers),]
plot(PSID$earnings)
plot(PSID$educatn)
plot(PSID$educatn)$out
outliers <- boxplot(PSID$educatn, plot=FALSE)$out
print(outliers)
PSID[which(PSID$educatn%in% outliers),]
plot(PSID$educatn)
par(mfrow=c(1,2))
plot(PSID$educatn,PSID$earnings)


psid = read.csv("PSID.csv")
summary(psid)
boxplot(earnings~educatn, data = psid) #Person with high education level has high earnings
boxplot(kids~age, data = psid)
nrow(psid)
min(psid$earnings)
max(psid$earnings)
mean(psid$earnings) #Average earning of a person
boxplot(hours~earnings, data = psid)
skewness(psid$earnings)
plot(psid$educatn, psid$earnings, col="green")
boxplot(earnings~married, data = psid)
plot(psid$married, psid$earnings, col="green")
# how earnings varies accoording to person's marital status
density.default(x=psid$earnings)
par(mfrow=c(1,2))
boxplot(psid$earnings)
boxplot(psid$hours)
plot(ecdf(psid$earnings))

mean(psid$hours)
aggregate(earnings~married,psid, mean)
boxplot(earnings~married,data=psid, cex.axis=0.5)
boxplot(hours~married,data=psid, cex.axis=0.5, col="blue")
```

```
psid = read.csv("PSID.csv")
mean(psid$hours)
sum(is.na(psid))
is.na(psid)
nrow(psid)
data.table
str(psid)
library(data.table)
dtPsid = data.table(psid)
str(dtPsid)
library(e1071)
skewness(psid$earnings)
which(is.na(psid),arr.ind = TRUE)

nrow(psidClean1)
nrow(psid)
data.table(psidClean1)
nrow(psidClean_kid)
data.table(psidClean_kid)
psidClean_kid = psid[psid$kid<20,]
psidClean_education = psidClean_kid[psidClean_kid$educatn < 20,]

psidClean_earning = psidClean_kid[psidClean_kid$hours > 0,]
nrow(psidClean_education)
data.table(psidClean_earning)

plot(psidClean_earning$hours,psidClean_earning$earnings)
fileCleanPsid = data.table(psidClean_earning)
save(fileCleanPsid, "saveddf.RData")
write.csv(fileCleanPsid, "fileCleanPsid.csv")

plot(psidClean_education[,5:7], col = "green")
kc = kmeans(psidClean_earning[,2:4],2)
kc

plot(psidClean_education[,4:6] , col=kc$cluster)
plot(psidClean_education$educatn, psidClean_education$earnings)
summary(psidClean_education)
boxplot(psidClean_education$earnings,psidClean_education$educatn)
boxplot(earnings~educatn, data = psidClean_education)
mean(psidClean_education$hours)
plot(density(psidClean_education$educatn)
```