Appendix 2 Data Analysis

```
setwd("~/Dropbox/Statistics/STA242/Assignment1")
DataforAnalysis=read.table("~/Dropbox/Statistics/STA242/Assignment1/DataforAnalysis")
#
library(ggplot2)
library(plyr)
library(reshape2)

#Response Variable

 #plot race time
  ggplot() +
  geom_histogram(aes(x=DataforAnalysis$Time.mins.,y=..density..),
                 fill="white",colour="black")+
  geom_density(aes(x=DataforAnalysis$Time.mins.),fill="#FFB7C5",alpha=0.5)+
  geom_vline(aes(xintercept=mean(DataforAnalysis$Time.mins.)),
             color="grey30", linetype="dashed", size=1)+
  labs(list(title = "Distribution of Race Time",x="Race Time(mins)"))

library(moments)
library(MASS)
skewness_time=skewness(DataforAnalysis$Time.mins.)   # 0.289 slightly positive skewned
boxcox(lm(DataforAnalysis$Time.mins.~1))+title(main="BoxCox for lm(Race Time ~ 1)")

#As BoxCox suggested, to balance the variance, we should take a sqrt for race time
DataforAnalysis$Response=sqrt(DataforAnalysis$Time.mins.)

#plot sqrt(race time)
ggplot() +
  geom_histogram(aes(x=DataforAnalysis$Response,y=..density..),
                 fill="white",colour="black")+
  geom_density(aes(x=DataforAnalysis$Response),fill="#FFB7C5",alpha=0.5)+
  geom_vline(aes(xintercept=mean(DataforAnalysis$Response)),
             color="grey30", linetype="dashed", size=1)+
  labs(list(title = "Distribution of sqrt(RaceTime)",x="sqrt[Race Time(mins)]"))

skewness_time=skewness(DataforAnalysis$Response)   # 0.011
boxcox(lm(DataforAnalysis$Response~1))+title(main="BoxCox for lm(sqrt(RaceTime) ~ 1)")


#Year
YearAnalysis=DataforAnalysis[c("YEAR","AG","BirthYear","GENDER","Frequency","Experience","Response")]

AgeCut=quantile(YearAnalysis$AG,na.rm=T)[4]
YearAnalysis$AgeCut[YearAnalysis$AG>=AgeCut]= 1
YearAnalysis$AgeCut[YearAnalysis$AG<AgeCut]= 0


Yearmelted = melt(YearAnalysis, id.vars=c("YEAR"))
Yearmeans  = ddply(subset(Yearmelted,variable == "Response"),
                   c("YEAR", "variable"), summarise,
                   mean=mean(value))


ggplot()+
  geom_boxplot(data=subset(Yearmelted,variable == "Response"), aes(factor(YEAR), value),
               fill="#FFB7C5",alpha = 0.5,colour="ivory3",
               outlier.colour="dodgerblue4",outlier.shape = 1,outlier.size = 1)+
  geom_point(data=Yearmeans,aes(factor(YEAR), mean),shape=5)+
  labs(list(title = "Sqrt(RaceTime) over year",x="Year", y="sqrt(RaceTime)"))

 ##Year+Age

YearAgemeans  = ddply(subset(Yearmelted,variable == "AG"|variable == "AgeCut"),
                   c("YEAR", "variable"), summarise,
                   mean=mean(value, na.rm = T))
Yearmeans$Agemean=YearAgemeans$mean[YearAgemeans$variable == "AG"]
Yearmeans$AgeCut=YearAgemeans$mean[YearAgemeans$variable == "AgeCut"]
```

```
remove(YearAgemeans)

ggplot(Yearmeans,aes(x=factor(YEAR),y=mean))+
  geom_point(stat="identity", aes(size=Yearmeans$AgeCut,colour=Yearmeans$Agemean))+
  scale_size(range = c(3, 8))+
  scale_colour_gradient(low="skyblue2", high="dodgerblue4")+
  labs(list(title = "Sqrt(RaceTime) over year: Age Distribution",x="Year", y="sqrt(RaceTime)",colour="Mean(Age)", size="Propotion of
Age>=43"))

##Year+Frequency/Experience

YearFremeans  = ddply(subset(Yearmelted,variable == "Frequency"|variable == "Experience"),
                      c("YEAR", "variable"), summarise,
                      mean=mean(value))
Yearmeans$Frequencymean=YearFremeans$mean[YearFremeans$variable == "Frequency"]
Yearmeans$Experiencemean=YearFremeans$mean[YearFremeans$variable == "Experience"]
remove(YearFremeans)

ggplot(Yearmeans,aes(x=factor(YEAR),y=mean))+
  geom_point(stat="identity", aes(size=Yearmeans$Frequencymean,colour=Yearmeans$Experiencemean))+
  scale_size(range = c(3, 8))+
  scale_colour_gradient(low="#FFB7C5", high="plum4")+
  labs(list(title = "Sqrt(RaceTime) over year: Experience Distribution",x="Year", y="sqrt(RaceTime)",colour="Experience(start counting from
1999)", size="Participation Times(1999-2010)"))

 ##Year+Gender
Yearmeans$GenderDis = ddply(subset(Yearmelted,variable == "GENDER"),
                      c("YEAR", "variable"), summarise,
                      mean=mean(value, na.rm = T))$mean
Yearmeans$GenderCut[Yearmeans$GenderDis<0.5]="more men"
Yearmeans$GenderCut[Yearmeans$GenderDis>0.5]="more women"

ggplot(Yearmeans,aes(x=YEAR,y=mean))+
  geom_point(stat="identity", aes(size=Yearmeans$GenderDis,colour=Yearmeans$GenderCut))+
  scale_size(range = c(2, 6))+
  scale_colour_manual(values = c("dodgerblue4","#FFB7C5"))+
  labs(list(title = "Sqrt(RaceTime) over year: Gender Distribution",
            x="Year", y="sqrt(RaceTime)",
            colour="More men or women", size="Women Propotion"))

   ### Divided by Gender

YearAnalysis$GenderLabel[YearAnalysis$GENDER==0]="men"
YearAnalysis$GenderLabel[YearAnalysis$GENDER==1]="women"

ggplot()+
  geom_boxplot(data=YearAnalysis, aes(factor(YEAR), Response),
               fill="#FFB7C5",alpha = 0.5,colour="ivory3",
               outlier.colour="dodgerblue4",outlier.shape = 1,outlier.size = 1)+
  facet_grid(~GenderLabel)+
  labs(list(title = "Sqrt(RaceTime) over year: Divided by Gender",x="Year", y="sqrt(RaceTime)"))

YearGendermelted = melt(YearAnalysis, id.vars=c("YEAR","GenderLabel"))
YearGenderMean=ddply(subset(YearGendermelted,variable == "Response"),
                     c("YEAR", "GenderLabel","variable"), summarise,
                     mean=mean(value, na.rm = T))
YearGenderAge  = ddply(subset(YearGendermelted,variable == "AG"),
                       c("YEAR", "GenderLabel","variable"), summarise,
                       mean=mean(value, na.rm = T))
YearGenderAgeCut=ddply(subset(YearGendermelted,variable == "AgeCut"),
                       c("YEAR", "GenderLabel","variable"), summarise,
                       mean=mean(value, na.rm = T))
YearGenderExperience  = ddply(subset(YearGendermelted,variable == "Experience"),
                              c("YEAR", "GenderLabel","variable"), summarise,
                              mean=mean(value, na.rm = T))
YearGenderFrequency=ddply(subset(YearGendermelted,variable == "Frequency"),
                          c("YEAR", "GenderLabel","variable"), summarise,
                          mean=mean(value, na.rm = T))

YearGenderMean$AgeCut=YearGenderAgeCut$mean
```

```
YearGenderMean$Age=YearGenderAge$mean
YearGenderMean$Experience=YearGenderExperience$mean
YearGenderMean$Frequency=YearGenderFrequency$mean
remove(YearGenderAge, YearGenderAgeCut, YearGenderExperience, YearGenderFrequency)

ggplot(YearGenderMean,aes(x=factor(YEAR),y=mean))+
  geom_point(stat="identity", aes(size=AgeCut,colour=Age))+
  facet_grid(~GenderLabel)+
  scale_size(range = c(2, 6))+
  scale_colour_gradient(low="skyblue2", high="dodgerblue4")+
  labs(list(title = "Sqrt(RaceTime) over year: Age Distribution",x="Year", y="sqrt(RaceTime)",colour="Mean(Age)", size="Propotion of
Age>=43"))

ggplot(YearGenderMean,aes(x=factor(YEAR),y=mean))+
  geom_point(stat="identity", aes(size=Frequency,colour=Experience))+
  facet_grid(~GenderLabel)+
  scale_size(range = c(2, 6))+
  scale_colour_gradient(low="#FFB7C5", high="plum4")+
  labs(list(title = "Sqrt(RaceTime) over year: Experience Distribution",x="Year", y="sqrt(RaceTime)",colour="Experience(start counting from
1999)", size="Participation Times(1999-2010)"))

#Sqrt(Time(mins)) VS Age


YearAnalysis=subset(YearAnalysis,!is.na(YearAnalysis$AG))
YearAnalysis=subset(YearAnalysis,YearAnalysis$AG>=10)

DensityNoraml=
  function(var)
  {
    x=var
    y=rnorm(length(x),mean(x,na.rm=T),sd(x,na.rm=T))
    d1=density(x,na.rm=T)
    d2=density(y)
    list(d1,d2)
  }

par(mfrow=c(2,3))
mat = matrix(c(1,3,2,4,0,5), 2)
layout(mat,c(2.5,2,0.5), c(1,3))

  #picture1
par(mar=c(0.5, 4.5, 0.5, 0.5))
MenAge=DensityNoraml(YearAnalysis$AG[YearAnalysis$GENDER==0])
plot(MenAge[[1]]$x,MenAge[[2]]$y,type = "l",ann = FALSE, axes = FALSE)
lines(MenAge[[2]]$x,MenAge[[2]]$y,col="red")

  #picture2
par(mar=c(0.5, 0.5, 0.5, 0.5))
WomenAge=DensityNoraml(YearAnalysis$AG[YearAnalysis$GENDER==1])
plot(WomenAge[[1]]$x,WomenAge[[2]]$y,type = "l",ann = FALSE, axes = FALSE)
lines(WomenAge[[2]]$x,WomenAge[[2]]$y,col="red")

  #picture3
par(mar=c(4.5, 4.5, 0.5, 0.5))
menmodel=lm(YearAnalysis$Response[YearAnalysis$GENDER==0]~
          YearAnalysis$AG[YearAnalysis$GENDER==0])
plot(YearAnalysis$AG[YearAnalysis$GENDER==0],
     YearAnalysis$Response[YearAnalysis$GENDER==0],ylim=range(6.5:13.5),
     ann=FALSE,col = "dodgerblue4")
text(72, 7.3, "coef=0.0106***", font=3)
abline(menmodel,col = "#FFB7C5", lty = "dashed",lwd=2)
title(ylab="Sqrt(Time(mins)",xlab = "Age(men)")

  #picture4
par(mar=c(4.5, 0.5, 0.5, 0.5))
womenmodel=lm(YearAnalysis$Response[YearAnalysis$GENDER==1]~
          YearAnalysis$AG[YearAnalysis$GENDER==1])
plot(YearAnalysis$AG[YearAnalysis$GENDER==1],
     YearAnalysis$Response[YearAnalysis$GENDER==1],ylim=range(6.5:13.5),
     ann=FALSE,col="#FFB7C5")
text(72, 7.3, "coef=0.0080***", font=3)
abline(womenmodel,col="dodgerblue4",lty="dashed",lwd=2)
title(xlab = "Age(women)")

  #picture5
par(mar=c(4.5, 0.5, 0.5, 0.5))
Response=DensityNoraml(YearAnalysis$Response)
plot(Response[[1]]$y,Response[[1]]$x,type = "l",ylim=range(6.5:13.5),
     ann = FALSE, axes = FALSE)
lines(Response[[2]]$y,Response[[2]]$x,col = "red")

par(mfrow=c(2,2))
par(mar=c(1, 1, 1, 1))
plot(menmodel)
plot(womenmodel)

par(mfrow=c(1,1))

AgeAnalysis=YearAnalysis[c("AG","GenderLabel","Response")]
Agemelted=melt(AgeAnalysis,id.vars = c("AG","GenderLabel"))
Agemeans  = ddply((Agemelted),
                  c("AG", "GenderLabel","variable"), summarise,
                  mean=mean(value),
                  lower=quantile(value)[2],upper=quantile(value)[4])

ggplot()+
  geom_linerange(data=Agemeans,aes(x=factor(AG),ymin=lower,ymax=upper),colour="wheat3")+
  geom_point(data=Agemeans,aes(x=factor(AG),y=mean,colour=GenderLabel))+
  facet_grid(~GenderLabel)+
  scale_colour_manual(values = c("dodgerblue4","#FFB7C5"))+
  labs(list(title = "Race Performance vs Age",x="Age", y="sqrt(RaceTime)",
            colour="Gender"))

# Race Performance vs Experience
ExperienceAnalysis=DataforAnalysis[c("Identify","GENDER","AG","Frequency","Experience","Response","PLACE")]
ExperienceAnalysis$GenderLabel[ExperienceAnalysis$GENDER==0]="men"
ExperienceAnalysis$GenderLabel[ExperienceAnalysis$GENDER==1]="women"
ExperienceAnalysis=subset(ExperienceAnalysis,AG>=10)
ExperienceAnalysis$ResponseCut[ExperienceAnalysis$Response<=9.5]=0
ExperienceAnalysis$ResponseCut[ExperienceAnalysis$Response>9.5]=1
ExperienceAnalysis$ResponseCut[ExperienceAnalysis$ResponseCut==0]="faster"
ExperienceAnalysis$ResponseCut[ExperienceAnalysis$ResponseCut==1]="slower"
ExperienceAnalysis$PlaceCut[ExperienceAnalysis$PLACE<=500]=0
ExperienceAnalysis$PlaceCut[ExperienceAnalysis$PLACE>500]=1
ExperienceAnalysis$PlaceCut[ExperienceAnalysis$PlaceCut==0]="Top Runner"
ExperienceAnalysis$PlaceCut[ExperienceAnalysis$PlaceCut==1]="Not Top"

ggplot(ExperienceAnalysis,aes(x=AG,y=Experience))+
  geom_point(stat="identity",aes(colour=ResponseCut))+
  facet_grid(~GenderLabel)+
  scale_colour_manual(values = c("dodgerblue4","#FFB7C5"))+
  labs(list(title = "Race Performance(speed) vs Experience",x="Age",
            colour="Race Speed"))

ggplot(ExperienceAnalysis,aes(x=AG,y=Experience))+
  geom_point(stat="identity",aes(colour=PlaceCut))+
  facet_grid(~GenderLabel)+
  scale_colour_manual(values = c("#FFB7C5","dodgerblue4"))+
  labs(list(title = "Race Performance(Place) vs Experience",x="Age",
            colour="Race Place"))

ggplot(ExperienceAnalysis,aes(x=AG,y=Frequency))+
  geom_point(stat="identity",aes(colour=ResponseCut))+
  facet_grid(~GenderLabel)+
  scale_colour_manual(values = c("dodgerblue4","#FFB7C5"))+
  labs(list(title = "Race Performance(speed) vs Frequency",x="Age",
            colour="Race Speed"))

ggplot(ExperienceAnalysis,aes(x=AG,y=Frequency))+
  geom_point(stat="identity",aes(colour=PlaceCut))+
  facet_grid(~GenderLabel)+
  scale_colour_manual(values = c("#FFB7C5","dodgerblue4"))+
  labs(list(title = "Race Performance(Place) vs Frequency",x="Age",
            colour="Race Place"))
# State
library(maps)
states_map = map_data("state")

state.name[51]="district of columbia"
state.abb[51]="DC"

StateAnalysis=DataforAnalysis[c("GENDER","STATE","STATE_CODE","PLACE")]
StateAnalysis$GenderLabel[StateAnalysis$GENDER==0]="men"
StateAnalysis$GenderLabel[StateAnalysis$GENDER==1]="women"
TopState=subset(StateAnalysis,PLACE<=500)
  ## men
STATEFrequencyMan=data.frame(table(StateAnalysis$STATE_CODE[StateAnalysis$GENDER==0]))
TopStateFrequencyMan=data.frame(table(TopState$STATE_CODE[StateAnalysis$GENDER==0]))
STATEFrequencyMan$State=tolower(state.name)[STATEFrequencyMan$Var1]
STATEFrequencyMan$StateAbb=state.abb[STATEFrequencyMan$Var1]
TopStateFrequencyMan$State=tolower(state.name)[TopStateFrequencyMan$Var1]
TopStateFrequencyMan$StateAbb=state.abb[TopStateFrequencyMan$Var1]

ggplot(STATEFrequencyMan, aes(map_id = State)) +
  geom_map(aes(fill = Freq), map = states_map) +
  expand_limits(x = states_map$long, y = states_map$lat)+
  scale_fill_gradient(low="orchid4", high="dodgerblue4")+
  labs(title = "State Pariticipation in Cherry Blossom 10 Mile Run (Men)",
            colour="Overall Participation(1999-2010)")

ggplot(TopStateFrequencyMan, aes(map_id = State)) +
  geom_map(aes(fill = Freq), map = states_map) +
  expand_limits(x = states_map$long, y = states_map$lat)+
  scale_fill_gradient(low="orchid4", high="dodgerblue4")+
  labs(title = "Top 500 Cherry Blossom 10 Mile Men Runners: Distribution among the States")

## women
STATEFrequencyWoman=data.frame(table(StateAnalysis$STATE_CODE[StateAnalysis$GENDER==1]))
TopStateFrequencyWoman=data.frame(table(TopState$STATE_CODE[StateAnalysis$GENDER==1]))
STATEFrequencyWoman$State=tolower(state.name)[STATEFrequencyWoman$Var1]
TopStateFrequencyWoman$State=tolower(state.name)[TopStateFrequencyWoman$Var1]


ggplot(STATEFrequencyWoman, aes(map_id = State)) +
  geom_map(aes(fill = Freq), map = states_map) +
  expand_limits(x = states_map$long, y = states_map$lat)+
  scale_fill_gradient(low="#FFB7C5", high="orchid4")+
  labs(title = "State Pariticipation in Cherry Blossom 10 Mile Run (Women)")

ggplot(TopStateFrequencyWoman, aes(map_id = State)) +
  geom_map(aes(fill = Freq), map = states_map) +
  expand_limits(x = states_map$long, y = states_map$lat)+
  scale_fill_gradient(low="#FFB7C5", high="orchid4")+
  labs(title = "Top 500 Cherry Blossom 10 Mile Women Runners: Distribution among the States")
```