Titanic - Data Exploration

### Variable Descriptions

**survival**: Survival (0 = No; 1 = Yes)

**pclass**: Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)

**name**: Name

**sex**: Sex

**age**: Age

**sibsp**: Number of Siblings/Spouses Aboard

**parch**: Number of Parents/Children Aboard

**ticket**: Ticket Number

**fare**: Passenger Fare

**cabin**: Cabin

**embarked**: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

library(gtools)  
library(ggplot2)  
library(caret)  
library(dplyr)  
library(RCurl)

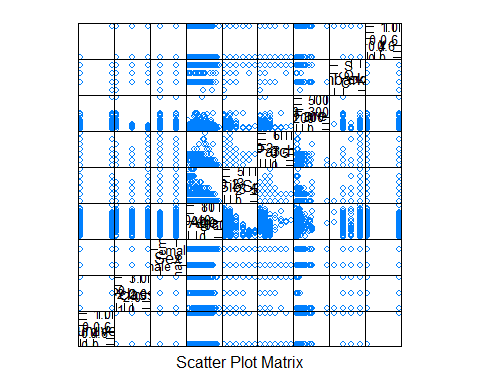
url <- getURL('https://raw.githubusercontent.com/frankwwu/R-Knots/master/Titanic/train.csv')  
train <- read.csv(text = url)   
dim(train)

## [1] 891 12

Removing names and NAs.

train<-train[, !(colnames(train) %in% c('name'))]  
train <-train %>% na.omit()

par(mar = rep(2, 4))  
trainCols <- c("Survived", "Pclass", "Sex", "Age", "SibSp", "Parch", "Fare", "Embarked")  
featurePlot(x=train[,trainCols], y=train$Survived, plot='pairs', par.settings=list(superpose.symbol=list(alpha = rep(1, 9), cex=rep(0, 1), font = rep(1, 2), pch=c(0))))



dev.copy(png, width=900, height=900, file="Titanic-features.png")

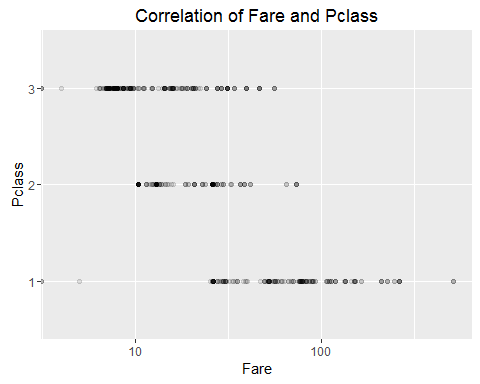
## png   
## 3

dev.off()

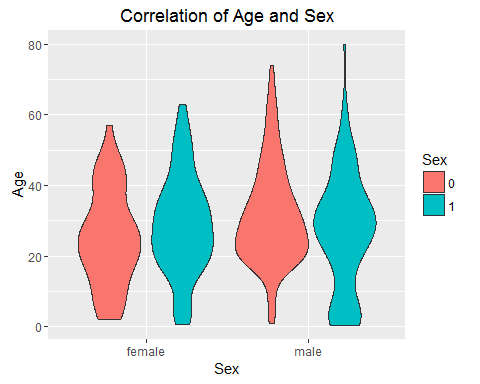
## png   
## 2

train$Survived <- factor(train$Survived)  
train$Sex <- factor(train$Sex)  
train$Pclass <- factor(train$Pclass)

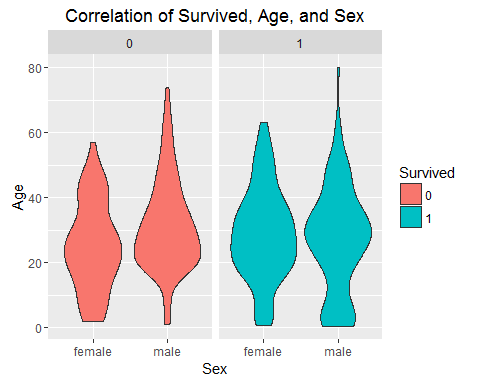
ggplot(train, aes(Fare, Pclass)) +   
 geom\_point(stat="identity", alpha=0.1) +  
 scale\_x\_log10() +  
 ggtitle("Correlation of Fare and Pclass")



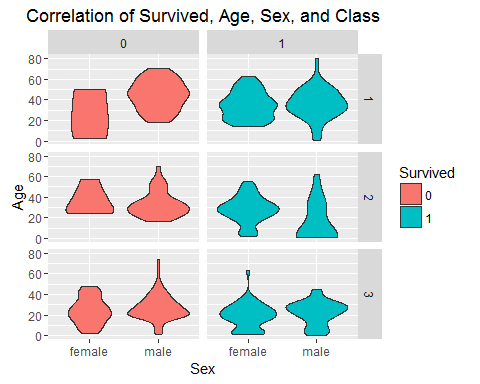
ggplot(data=train, aes(x=Sex, y=Age, fill=Survived)) +  
 geom\_violin() +   
 xlab("Sex") +  
 ylab("Age") +  
 guides(fill=guide\_legend(title="Sex")) +   
 ggtitle("Correlation of Age and Sex")



ggplot(data=train, aes(x=Sex, y=Age, fill=Survived)) +  
 geom\_violin() +   
 facet\_grid(. ~ Survived) +   
 xlab("Sex") +  
 ylab("Age") +  
 guides(fill=guide\_legend(title="Survived")) +   
 ggtitle("Correlation of Survived, Age, and Sex")



ggplot(data=train, aes(x=Sex, y=Age, fill=Survived)) +  
 geom\_violin() +   
 facet\_grid(Pclass ~ Survived) +   
 xlab("Sex") +  
 ylab("Age") +  
 guides(fill=guide\_legend(title="Survived")) +   
 ggtitle("Correlation of Survived, Age, Sex, and Class")



ggplot(train, aes(Age, Fare, color=Survived)) +   
 geom\_point(alpha = 0.5) +  
 facet\_grid(Pclass~Sex) +  
 ggtitle("Training Data")



Suggested training set formula: Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked