```r
# Week 3, Recitation

# Video 2

# Read in data
polling = read.csv("PollingData.csv")
str(polling)
table(polling$Year)
summary(polling)

# Install and load mice package
install.packages("mice")
library(mice)

# Multiple imputation
simple = polling[c("Rasmussen", "SurveyUSA", "PropR", "DiffCount")]
summary(simple)
set.seed(144)
imputed = complete(mice(simple))
summary(imputed)
polling$Rasmussen = imputed$Rasmussen
polling$SurveyUSA = imputed$SurveyUSA
summary(polling)



# Video 3

# Subset data into training set and test set
Train = subset(polling, Year == 2004 | Year == 2008)
Test = subset(polling, Year == 2012)

# Smart Baseline
table(Train$Republican)
sign(20)
sign(-10)
sign(0)
table(sign(Train$Rasmussen))
table(Train$Republican, sign(Train$Rasmussen))



# Video 4

# Multicollinearity
cor(Train)
str(Train)
cor(Train[c("Rasmussen", "SurveyUSA", "PropR", "DiffCount",
```

```
 "Republican")])

# Logistic Regression Model
mod1 = glm(Republican~PropR, data=Train, family="binomial")
summary(mod1)

# Training set predictions
pred1 = predict(mod1, type="response")
table(Train$Republican, pred1 >= 0.5)

# Two-variable model
mod2 = glm(Republican~SurveyUSA+DiffCount, data=Train,
 family="binomial")
pred2 = predict(mod2, type="response")
table(Train$Republican, pred2 >= 0.5)
summary(mod2)



# Video 5

# Smart baseline accuracy
table(Test$Republican, sign(Test$Rasmussen))

# Test set predictions
TestPrediction = predict(mod2, newdata=Test, type="response")
table(Test$Republican, TestPrediction >= 0.5)

# Analyze mistake
subset(Test, TestPrediction >= 0.5 & Republican == 0)
```