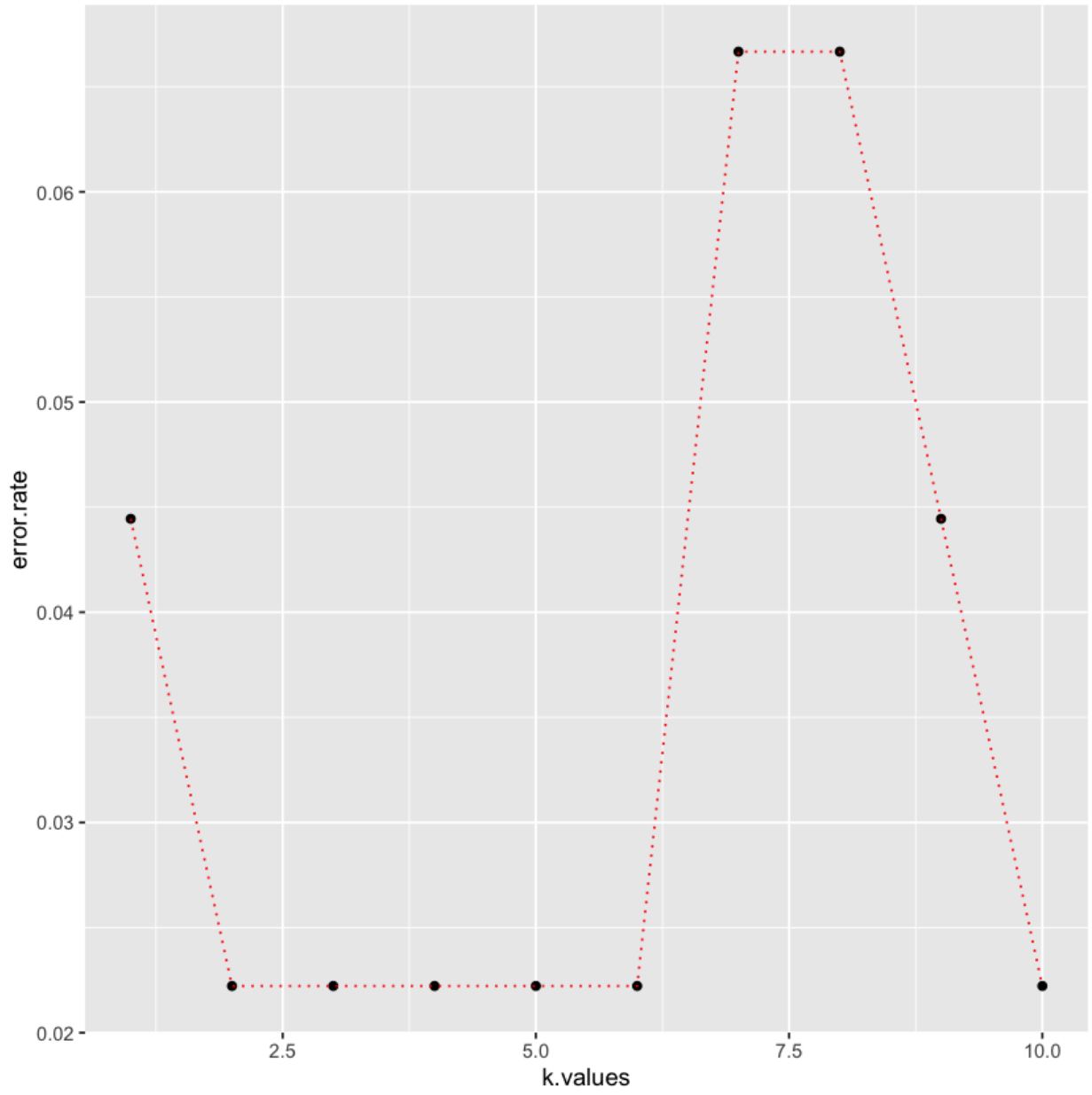
K Nearest Neighbors report

Iris Data Set

```
library(ISLR)
library(ggplot2)
library(caTools)

df <- iris
head(df)
str(df)

# Standardize Data
stand.df <- scale(df[, -5])

# -> Check that the scaling worked by checking the variance of one of the new columns
var(stand.df[,1])
var(stand.df[,2])

# Join the standardized data with the species column
stand.df <- as.data.frame(stand.df)
stand.df$Species <- df$Species
head(stand.df)

# Train and Test Splits
set.seed(24)
split = sample.split(stand.df, SplitRatio = 0.70)
train = subset(stand.df, split == TRUE)
test = subset(stand.df, split == FALSE)

# Build a KNN model
model <- knn(train[1:4],test[1:4],train$Species,k=1)
table(model)

# misclassification rate
mean(test$Species != model)

# Choosing a K Value
model = NULL
error.rate = NULL
for(i in 1:10){
  set.seed(24)
  model = knn(train[1:4],test[1:4],train$Species,k=i)
  error.rate[i] = mean(test$Species != model)
}

error.rate
```

```
k.values <- 1:10
error.df <- data.frame(error.rate,k.values)
ggplot(error.df,aes(x=k.values,y=error.rate)) + geom_point()+ geom_line(lty="dotted",color='red')
```



You should have noticed that the error drops to its lowest for k values between 2-6. Then it begins to jump back up again, this is due to how small the data set it. At k=10 you begin to approach setting k=10% of the data, which is quite large