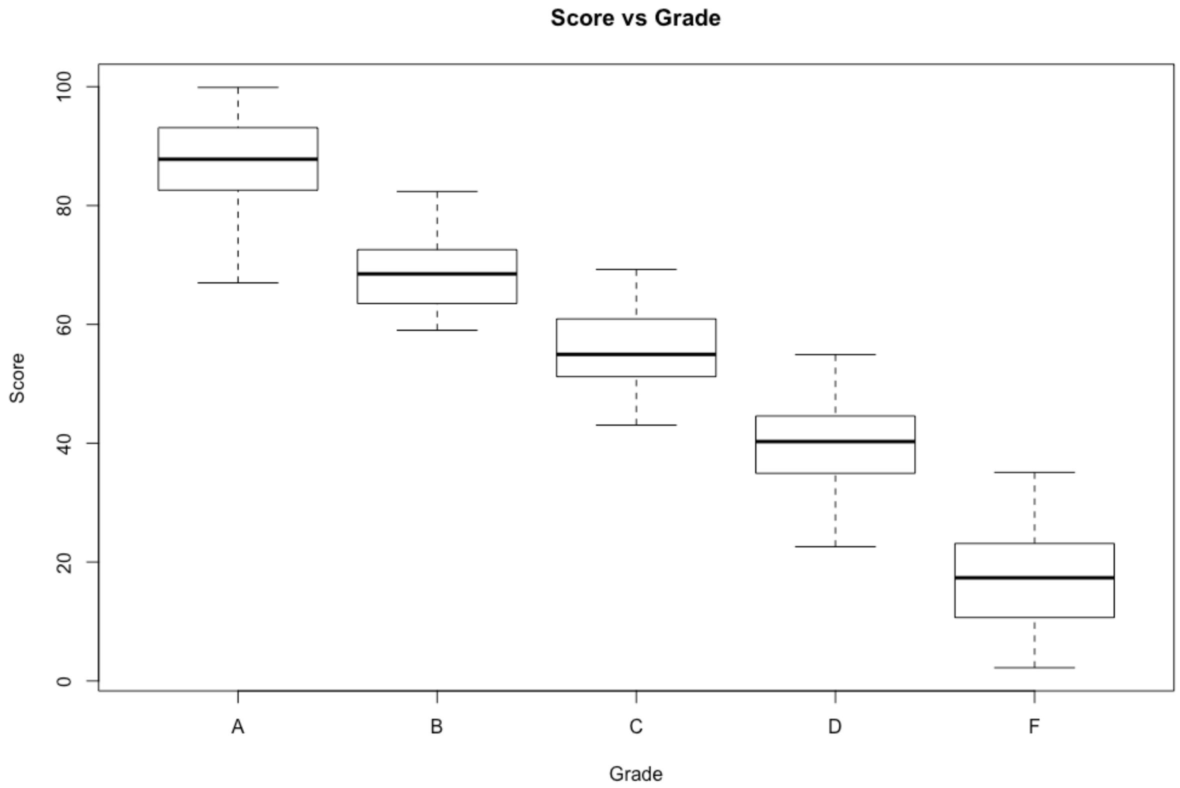
# Prediction Challenge 1

Jack Lin

# First Glance

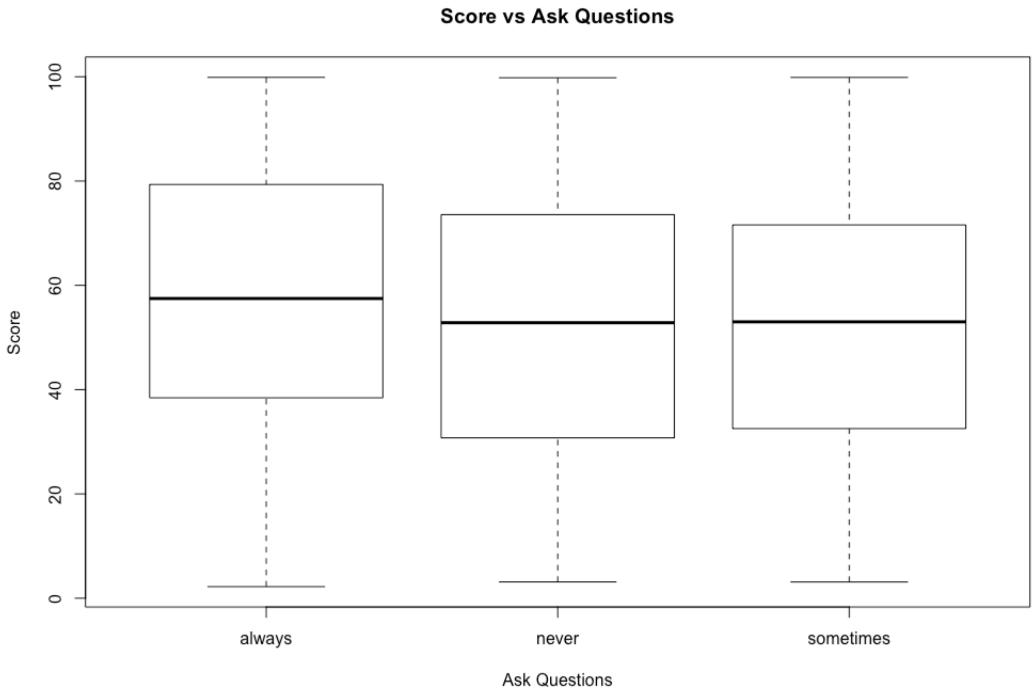


R code:

```
> boxplot(M2018_train$SCORE~M2018_train$GRADE, main = “Score vs Grade”, xlab = "Grade", ylab = "Score")
```

Finding: Apparently raw scores tend to be positively correlated to grades, but there are some overlaps. Therefore, we need to delve into other variables, too….

# First Glance

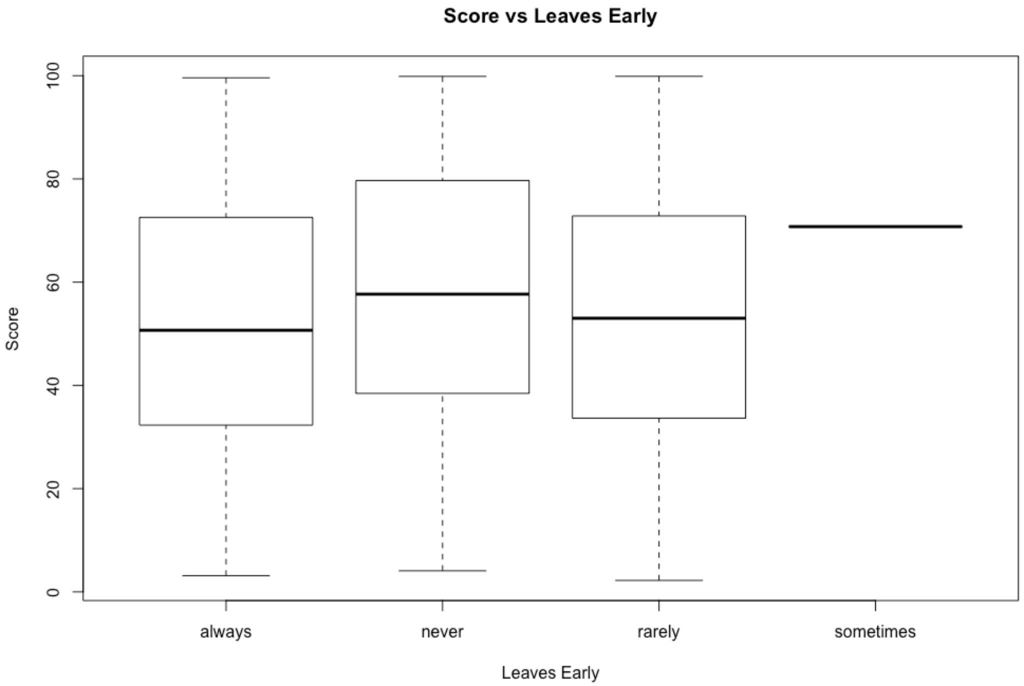


R code:

```
> boxplot(M2018_train$SCORE~M2018_train$ASKS_QUESTIONS, main = "Score vs Ask Questions", xlab = "Ask Questions", ylab = "Score")
```

Finding: Asking questions does not seem to affect your raw score.

# First Glance

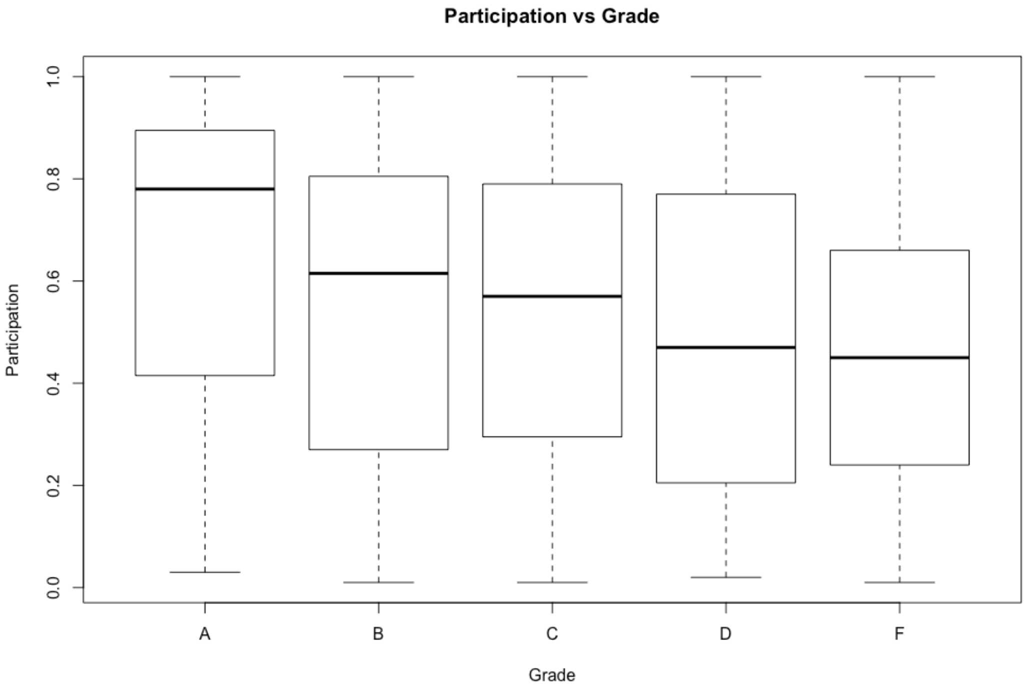


R code:

```
> boxplot(M2018_train$SCORE~M2018_train$LEAVES_EARLY, main = "Score vs Leaves Early", xlab = "Leaves Early", ylab = "Score")
```

Finding: Neither does leaving class early affect your raw score.

# First Glance



R code:

```
> boxplot(M2018_train$PARTICIPATION~M2018_train$GRADE, main = "Participation vs Grade", xlab = "Grade", ylab = "Participation")
```

Finding: Now we see some patterns between participation and grade. At first, it looks like A students tend to score higher in participation. Let's look further…..

# Use Tapply Functions to Data Crunch the Overlapping Score Ranges

```
> tapply(M2018_train$SCORE, M2018_train$GRADE, min)
    A     B     C     D     F 
66.99 59.01 43.04 22.59  2.22 
> tapply(M2018_train$SCORE, M2018_train$GRADE, max)
    A     B     C     D     F 
99.88 82.37 69.25 54.91 35.09 
> tapply(M2018_train$SCORE, M2018_train$GRADE, mean)
       A        B        C        D        F 
87.49059 68.70992 55.69918 39.71797 17.47906 
```

# Use Subset Functions to Obtain the Data Range We Want to Examine

- > AtoC.data <- subset(M2018_train, M2018_train$SCORE >= 66.99)
- > AtoC.data <- subset(AtoC.data, AtoC.data$SCORE <= 82.37)
- > FtoD.data <- subset(M2018_train, M2018_train$SCORE >= 22.59)
- > FtoD.data <- subset(FtoD.data, FtoD.data$SCORE < 35.09)
- > DtoC.data <- subset(M2018_train, M2018_train$SCORE>=43.04)
- > DtoC.data <- subset(DtoC.data, DtoC.data$SCORE<=54.91)
- > CtoB.data <-subset(M2018_train, M2018_train$SCORE >= 59.01)
- > CtoB.data <-subset(CtoB.data, CtoB.data$SCORE <= 69.25)

# Freestyle Prediction

```
> decision <- rep(“F”,nrow(myprediction))
> decision[myprediction$SCORE>=22.59 & myprediction$PARTICIPATION> 0.54] <-“D”
> decision[myprediction$SCORE> 34.94] <-“D"
> decision[myprediction$SCORE>=43.04 & myprediction$PARTICIPATION> 0.55] <-“C"
> decision[myprediction$SCORE> 54.91] <-“C"
> decision[myprediction$SCORE>=59.01 & myprediction$PARTICIPATION>=0.52] <- “B"
> decision[myprediction$SCORE > 69.25] <- “B”
> decision[myprediction$SCORE>=66.99 & myprediction$PARTICIPATION>=0.70] <- “A”
> decision[myprediction$SCORE > 82.37] <- “A”
> myprediction$GRADE <- decision
> error <- mean(M2018_train$GRADE!=myprediction$GRADE)
> error
```

Method: By eyeballing each subset I created, I determined the participation cut-off between grades.

# Apply the Prediction to Test Data

```
> colnames(M2018_test_students)[3] <- "GRADE"
> View(M2018_test_students)
> decision <- rep('F',nrow(myprediction))
> decision[myprediction$SCORE>=22.59 & myprediction$PARTICIPATION> 0.54] <-"D"
> decision[myprediction$SCORE> 34.94] <-“D”
> decision[myprediction$SCORE>=43.04 & myprediction$PARTICIPATION> 0.55] <-“C"
> decision[myprediction$SCORE> 54.91] <-“C”
> decision[myprediction$SCORE>=59.01 & myprediction$PARTICIPATION>=0.52] <- "B"
> decision[myprediction$SCORE > 69.25] <- "B"
> decision[myprediction$SCORE>=66.99 & myprediction$PARTICIPATION>=0.70] <- "A"
> decision[myprediction$SCORE > 82.37] <- "A"
> M2018_test_students$GRADE <- decision
> View(M2018_test_students)
```

Result: I scored 0.86434 on Kaggle, which isn’t bad considered that I did not really overfit the training data. If more time is given, I could have delved further on the high-60s range where you can get anything between A and C.