CS 199

Final Project - MOOC Data Analysis

May 16, 2014

Sam Laane <laane2@illinois.edu>
José Vicente Ruiz <ruizcep2@illinois.edu>
Nick Jeffrey <njeffre2@illinois.edu>

Contents

1 Introduction

This assignment consisted of analysing the provided *MOOC data* and determing whether we could predict final exam scores or determine whether bloom taxology levels actually provided a meaningful differentiation between questions.

For that purpose, a dataset was provided by *Magical Data Fairies*. The programming languages used have been `R` and `Python`, and this report have been typeset using LaTeX.

2 Predicting exam scores

2.1 Implementation

???

3 Finding correlations

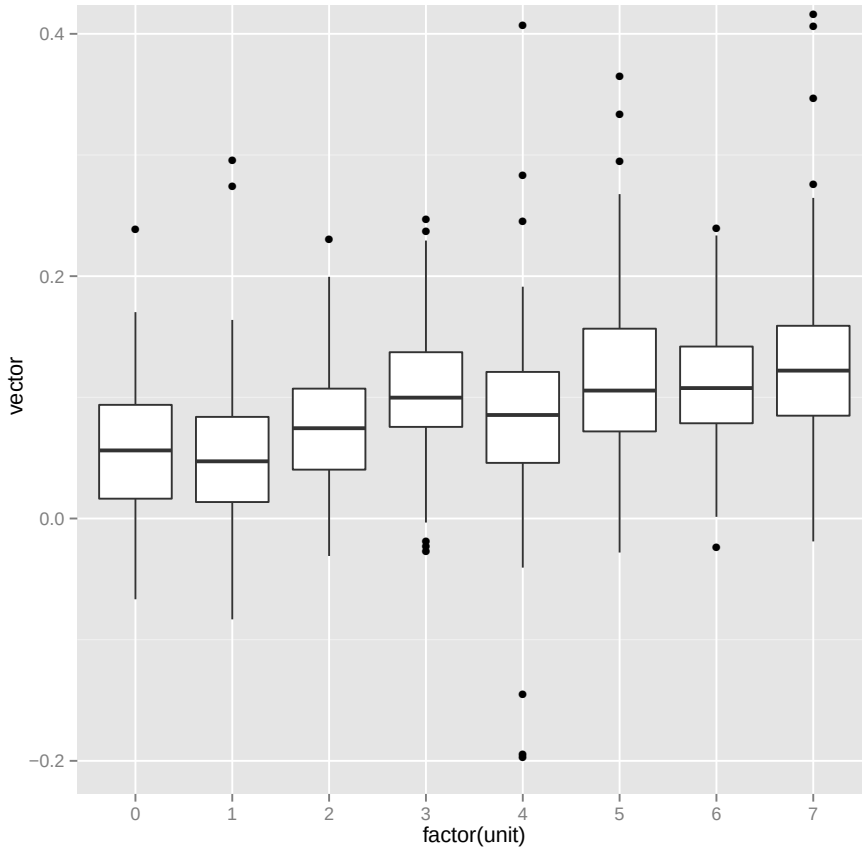
3.1 Implementation

```r
# Define functions.
printf <- function(...) invisible(print(sprintf(...)))
library(ggplot2)

# Read the data.
#toRm = c("Q4Q07_corr", "Q4Q13_corr")
dataWithNames<- read.csv('combinednodropna.csv', h=T) # Headers = True.

#remove user id's
data <- dataWithNames[-1]

# remove values with std of zero

toRm = c()
#check for std of 0
for (d in names(data)){
    s = sd(data[[d]], na.rm = TRUE)
    if (s == 0){
        print(d)
        toRm = c(toRm, d)
    }
}
data <- data[,!(names(data)) %in% toRm]

#quizData <- data[,grep(sprintf("^Q%dQ[0-9][0-9]_corr$",4),names(data))]
#testData <- data[,-(1:(20 * 8-2))]
#ncol(quizData)
#View(quizData)

units = c(0,1,2,3,4,5,6,7)
#units = c(4,5,6)
#quizData <- data[,1:(20*8-2)]
```

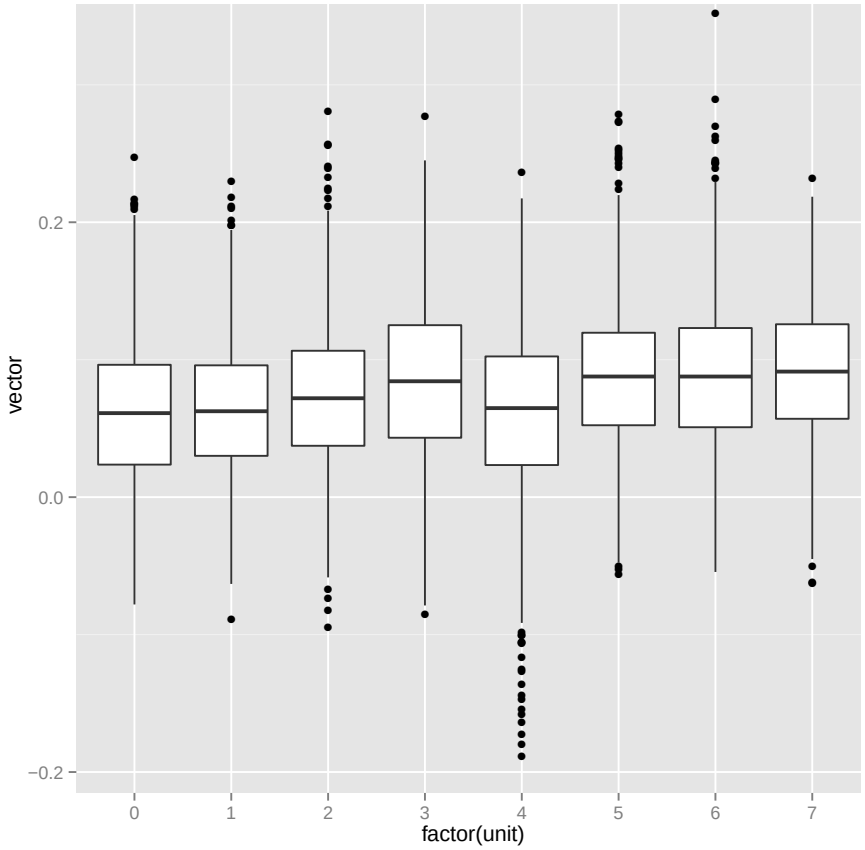
```
#testData <- data[,-(1:(20 * 8-2))]

#calulate Correlation matrix

correlationMatrix = cor(data, use = "na.or.complete")


#find how correlations compare.
correlationSums <- (apply(correlationMatrix, 2, sum )) -1
correlationMeans <- correlationSums / (ncol(data) - 1 )
#View(correlationMeans)
#print(mean(correlationMeans))


#now lets compare the quize to the test
quizCol <- correlationMatrix[,1:(20 * 8-2)]
testRow <- quizCol[-(1:(20 * 8-2)), ]
#View(testRow)
df <- data.frame(vector = numeric(), unit = numeric())
dff <- data.frame(vector = numeric(), unit = numeric())
linedata <- data.frame(unit = numeric(), same = numeric(), others =
   numeric(), diff = numeric())
units = c(0,1,2,3,4,5,6,7)
for (i in units){
    questionsInUnit <- testRow[,grep(sprintf("^Q%dQ[0-9][0-9]_corr$",i)
        ,names(data))]
    t <- questionsInUnit[(1:5) + 5 * i,]
    f <- questionsInUnit[-((1:5) + 5 * i),]
    corInUnit <- mean(t)
    corOutUnit <- mean(f)
    printf("mean correlation of week %d quiz to week %d test questions:
        %f", i, i, corInUnit)
    printf("max  correlation of week %d quiz to week %d test questions:
        %f", i, i, max(t))
    printf("mean correlation of week %d quiz to other test questions: %
       f", i, corOutUnit)
    printf("max  correlation of week %d quiz to other test questions: %
       f", i, max(f))
    printf("the diffence of correlation: %f", corInUnit - corOutUnit)
    #View(t)
    #Sys.sleep(1200)
    df <- rbind(df, data.frame(vector= as.vector(t), unit=i))
    dff <- rbind(dff, data.frame(vector= as.vector(f), unit=i))
    linedata <- rbind(linedata, data.frame(unit = i, same = mean(t),
       other = mean(f), diff = corInUnit - corOutUnit))
}
ggplot(dff, aes(y=vector, x=factor(unit))) + geom_boxplot()
ggsave('unrelatedcorrelationbox.pdf')
```

3.1.1 Results

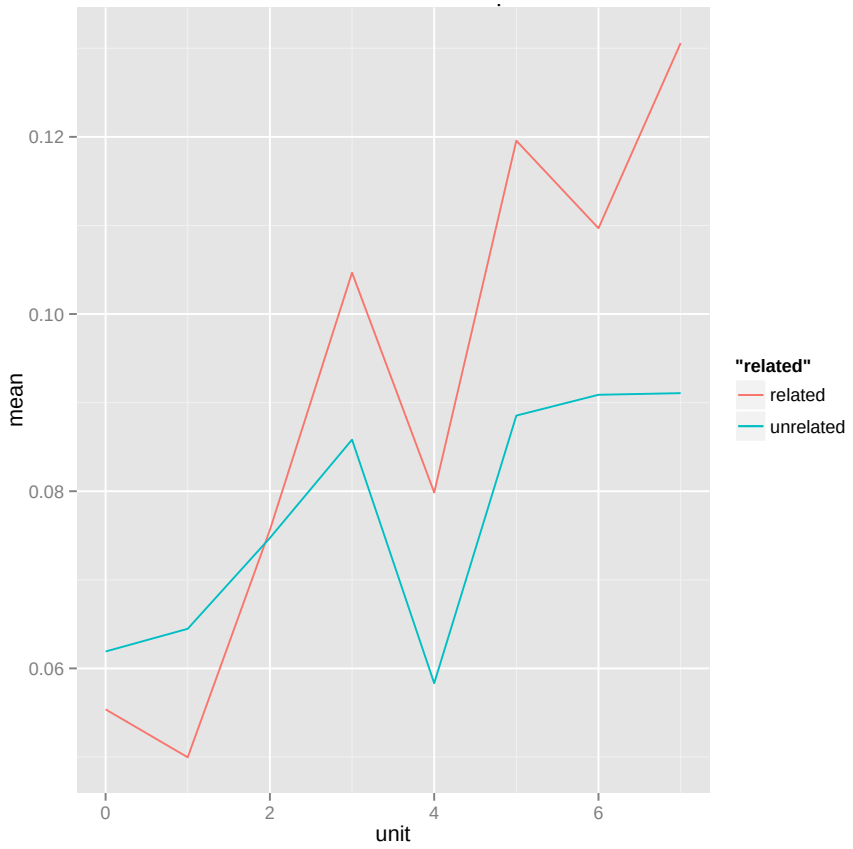
This figure shows a box and whisker plot of correlation for every question on a quiz to every related question on the final.



This figure shows a Box and Whisker plot for every question on a quiz to the unrelated questions on the final.



This figure shows the mean correlations of quiz questions on related and unrelated questions on the final side by side.



3.2 Conclusions

3.2.1 Correlations

Judging from the graphs provided, the correlation of quiz scores with the final had a few interesting trends. One was an outcome we expected, which is that quizzes later in the course more strongly correlate with their questions on the final, most likely due to the simple fact that the material is fresher in the student's minds. This trend can be seen both in the box plot and in the line graph.

Another trend can be seen in the box plots as well: That there are many outliers, expecially later on in the course. The outliers that have much higher correlation are easy to explain because one would expect that questions which are testing the same material woulc have stronger correlations than those testing different material in the same subject. However, there are also outliers in the other direction, even going so far as to have a few questions that are negatively correlated with exam scores. We suspect that this could be due to those being trick questions, that students only studied because they happened to guess wrongly on the quiz, but without having access to the questions themselves, this can at best only be a guess as to why these trends are appearing.