**Multivariate Data Analysis**

**Fall 2018**

**Final Exam**

(Total 95 pts + Extra 5 pts)

**문제풀이를 위해 사용한 모든 R code를 .R 파일로 반드시 함께 제출**

1. Suppose we collect data for a group of students in a statistics class with variables X1=hours studied, X2=undergrad GPA, and Y=receive an A. We fit a logistic regression and produce estimated coefficient,  .
   1. (15 pts) Estimate the probability that a student who studied for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

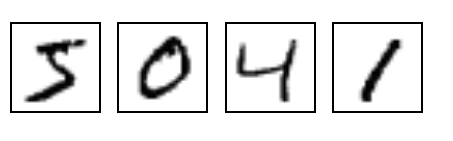
X1 = 40, X2 = 3.5 일때, 40\*0.05 + 3.5\*1 – 6 = -0.5

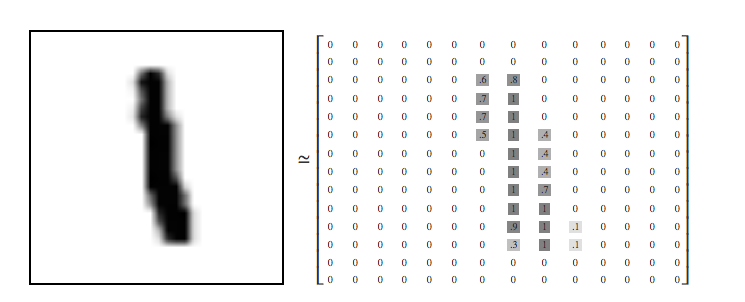
따라서, 학생이 A성적을 받을 확률은 exp(-0.5) = 0.6065307, 60.65307%이다.

* 1. (15 pts) Interpret in terms of the odds of receiving A.

다른 변수들이 동일하다는 가정 하에, 학생이 공부한 시간이 1단위 증가할 때마다, 학생이 A 성적을 받는 odds는 exp(0.05) = 1.051271, 즉 5.1271%만큼 증가한다고 해석할 수 있다.

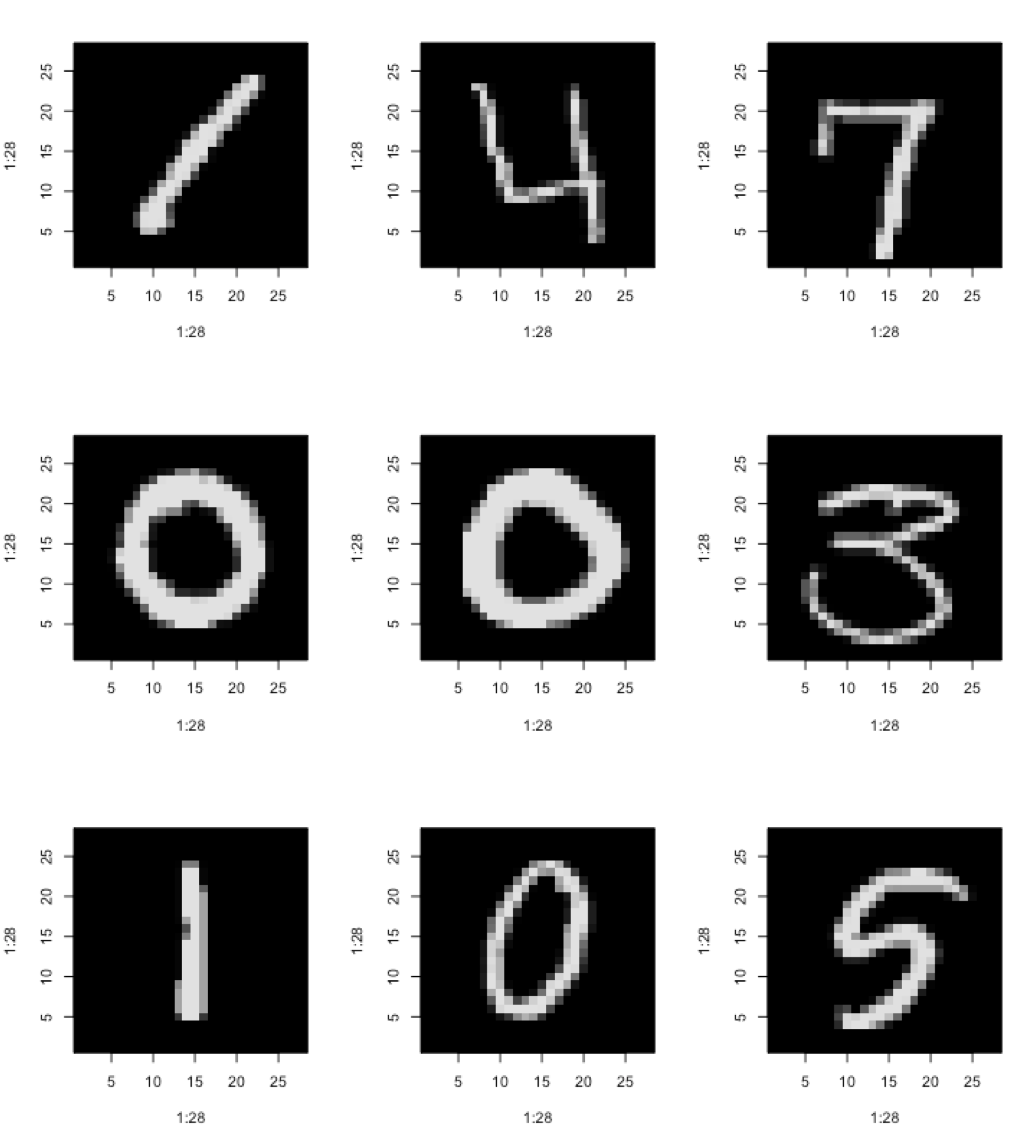
1. MNIST(“mnist\_1000.csv”) is a simple computer vision dataset. It consists of 28x28 pixel images of handwritten digits, such as:

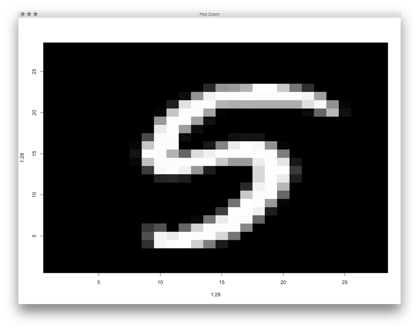
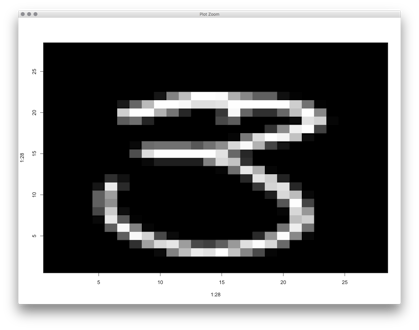
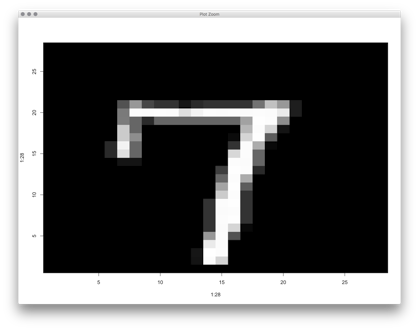
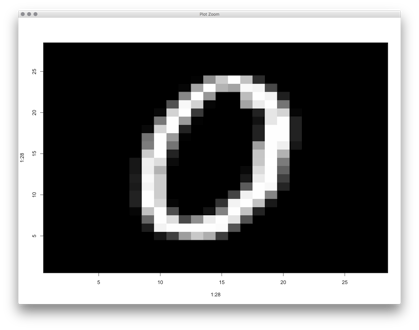
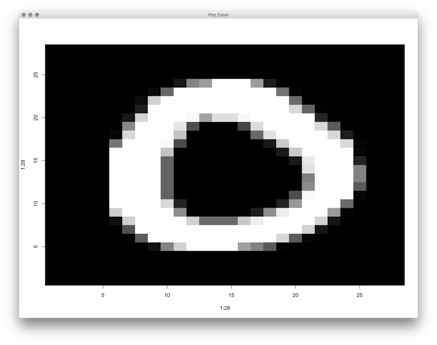
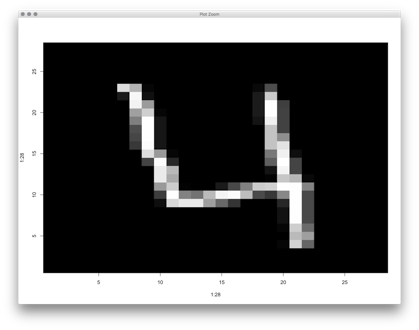
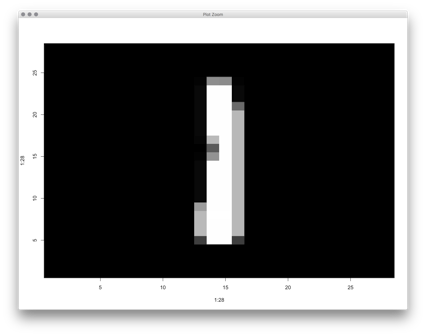
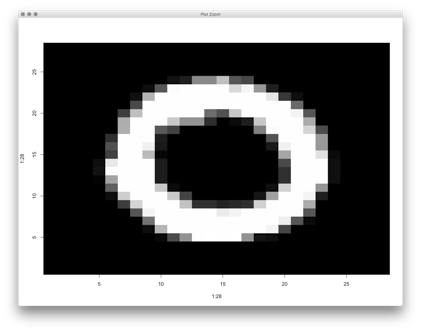
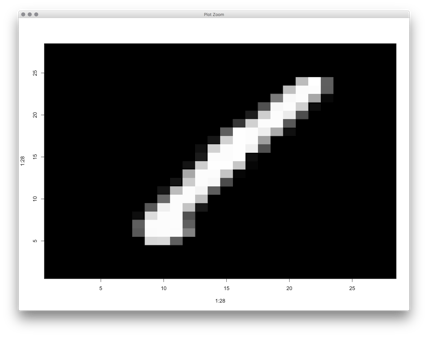
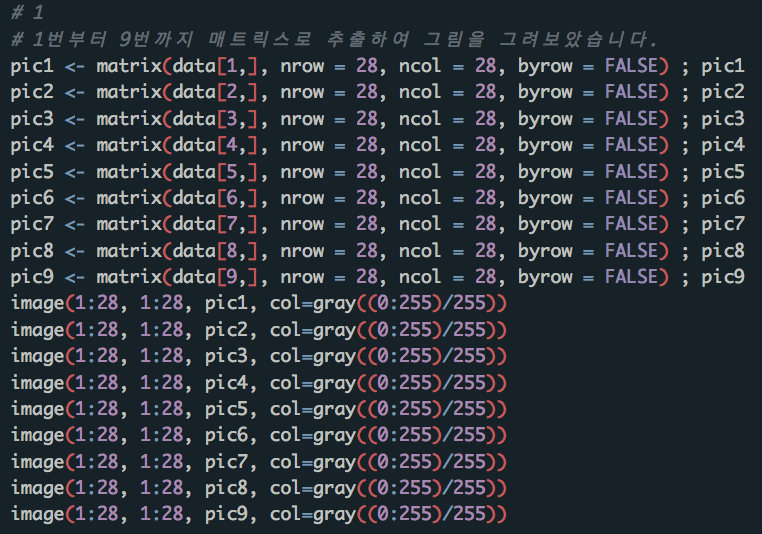


Every MNIST data point, every image, can be thought of as an array of numbers describing how dark each pixel is. For example, each data point is a 784(=28\*28) dimensional vector, which can be converted into 28\*28 matrix as: The 785th column contains the true labels of the handwritten digits.

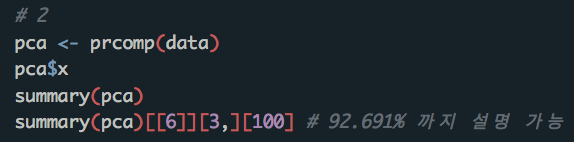
|  |
| --- |
| ## 아래 코드를 통해 전처리 한 후에 아래 문제 진행  data=read.csv("mnist\_1000.csv")  label=data$label # 실제 label 분리 저장  data=as.matrix(data[,-785]) # pixel 자료 분리 저장 |

1. (15 pts) Produce the first 9 images (i=1, 2, …, 9) among 1000 images to explore the data. That is, generate images as something like:

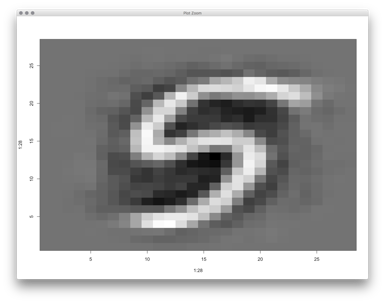
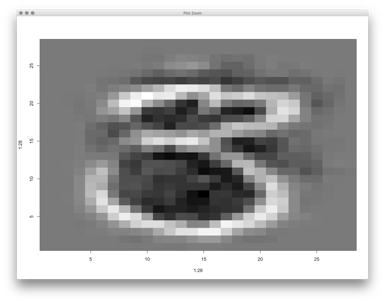
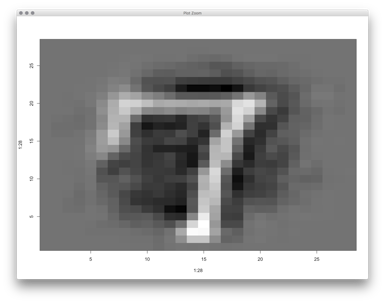
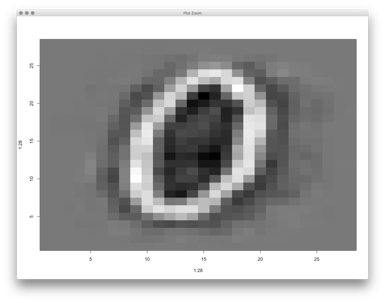
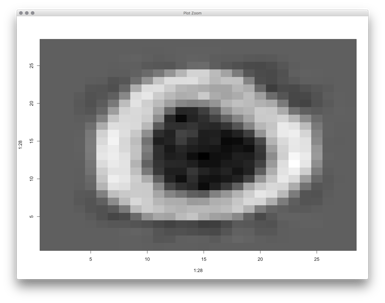
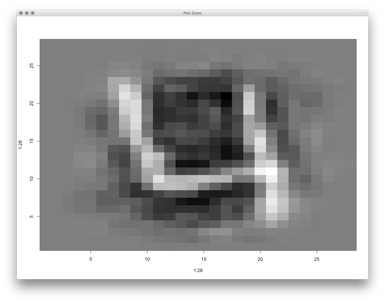
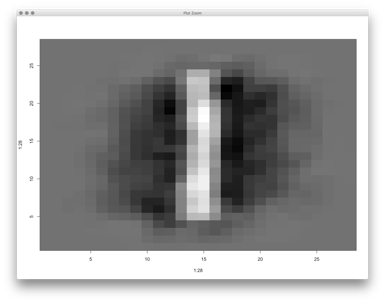
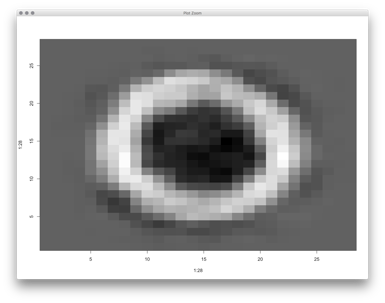
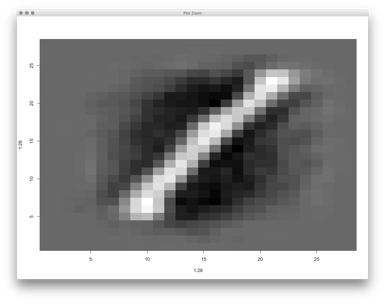




1. (15 pts) Perform principal components analysis using 784 pixels (use the option “scale=T”). If the dimension is reduced from 784 to 100, what percentage of the information is kept?



1. (15 pts) Using the first 100 PCs, reconstruct the first 9 images and compare with the results in problem 1.



1. (20 pts) Using the first 100 PCs, perform K-means clustering with K=10 and nstart=10. ***(Use “set.seed(1)” before running kmeans.)*** Compare the cluster labels and the true label, and describe the result.

