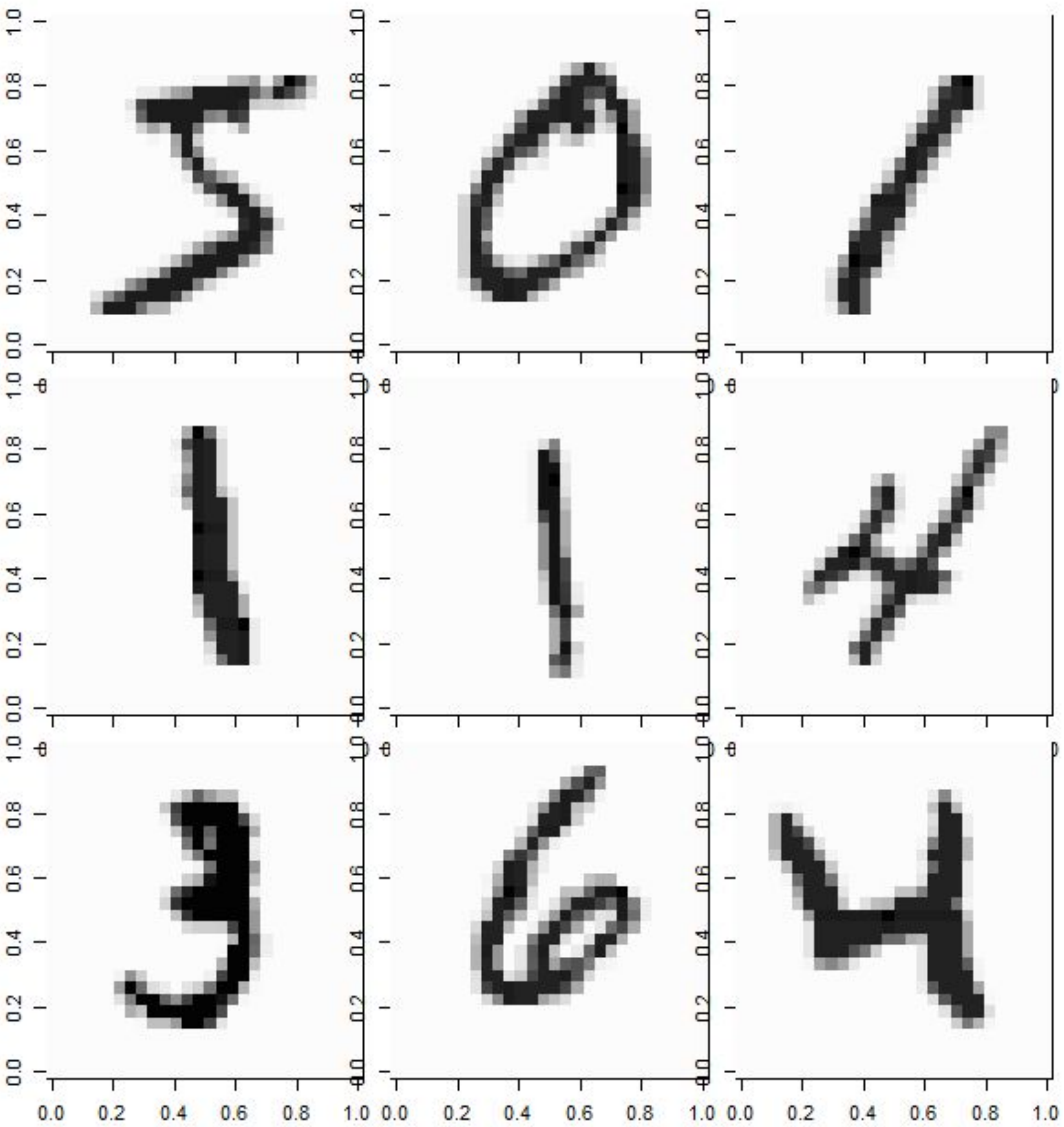
Digits Recoginition

Data discription

In this situation,
we choose a dataset that related to digits recognization.It contains 784 pixels variables. And contain 28000 items which includes all the 0,1,2,3,4,5,6,7,8,9 handwriting digits.

In order to coding 0 and 1, we only choose two type of the digits to fit the GLM model. Here we subset 1s and 8s in both of our traindata(which contains 5081 items) and testdata(which contains 2177 item).

Here are the plots of several items in the data set.



Preprocessing the data

As we can see from the plot of data set, there are many pixels varables are also zero.

So the first things we need to do is to reduce the number of variables which have no help to our prediction So in order to vanish all these things. We first use PCA to reduce the dimension and also eliminate the collinearity.

Then the number of pixels variables we choose is 20 which cover 70 of the variation. The reason we choose 20 is to satisfy the requirement and save the programming time.

our output is here when k=0.05

```
Full model accuracy is 662.3877 .
Removing variable V14 . Accuracy is now 661.4665 .
Removing variable V16 . Accuracy is now 661.1731 .
Removing variable V15 . Accuracy is now 660.652 .
Removing variable V13 . Accuracy is now 660.4225 .
Removing variable V20 . Accuracy is now 660.581 .
Removing variable V18 . Accuracy is now 662.1174 .
Removing variable V9 . Accuracy is now 664.2468 .
Removing variable V17 . Accuracy is now 666.0451 .
Removing variable V8 . Accuracy is now 672.6036 .
Removing variable V11 . Accuracy is now 685.0384 .
Removing variable V3 . Accuracy is now 694.0031 .
Removing variable V19 . Accuracy is now 718.06 .
Removing variable V12 . Accuracy is now 741.5915 .
[1]  1  2  4  5  6  7 10
```

confusion matrix

True\pred	8	1
8	932	65
1	16	1164

our output is here when k=0.01

```
Full model accuracy is 662.3877 .
Removing variable PC14 . Accuracy is now 661.4665 .
Removing variable PC16 . Accuracy is now 661.1731 .
Removing variable PC15 . Accuracy is now 660.652 .
Removing variable PC13 . Accuracy is now 660.4225 .
Removing variable PC20 . Accuracy is now 660.581 .
Removing variable PC18 . Accuracy is now 662.1174 .
Removing variable PC9 . Accuracy is now 664.2468 .
Removing variable PC17 . Accuracy is now 666.0451 .
Removing variable PC8 . Accuracy is now 672.6036 .
[1]  1  2  3  4  5  6  7 10 11 12 19
```

confusion matrix

True\pred	8	1
8	939	58
1	17	1163

PC	1	2	3	4	5
p-value	<2e-16	<2e-16	0.000107	0.001438	<2e-16
PC	6	7	8	9	10
p-value	2.53e-10	<2e-16	0.000210	0.005042	2.53e-14
PC	11	12	13	14	15
p-value	0.000150	1.01e-08	0.27292	0.801610	0.297027
PC	16	17	18	19	20
p-value	0.652590	0.297027	0.000111	5.28e-06	0.000535

By using significance test p=0.05, we will remove PC13, PC14, PC15, PC 16, PC17