For your selected competition, write a few sentences describing the competition problem as you interpreted it.  You want your writeup to be self-contained so your peer-reviewer does not need to go to Kaggle to study the competition description.  Clarity is more important than detail.  What's the overall goal?  What does the data look like?  How will the results be evaluated?  
  
*Example: "The task is to predict whether a given passenger survived the sinking of the Titanic based on various attributes including age, location of the passenger's cabin on the ship, family members, the fare they paid, and other information.  Solutions are evaluated by comparing the percentage of correct answers on a test dataset."*

The model to be developed should be able to recognize hand drawn digits (0-9) from 28x28 Pixel images. The train data was a flat file containing 42000 rows (representing the image) and 28x28+1 columns (+1 for the label) with 8bit gray values (0-255). The test data consisted of 28000 images in a similar format, the first column with the label was missing. The predictions of the labels of this test data is actually graded by Kaggle.

Write a few sentences describing how you approached the problem.  What techniques did you use?   
  
*Example: "I split the data by gender and handled each class separately.  For the females, I trivially classified all of them as "survived."  For the males, I trained a random forest as a classifier.  I ignored the pclass atribute that indicated the location of the passenger's cabin because I didn't think it was relevant."*

I tried to reduce the complexity of the problem by reducing the number of variables. For all images in the test and training data column sums and row sums were determined. This made a total of 57 (28 rows +28 columns+1 label) columns in the new datasets. If successful, this approach should scale very good, since for an e.g. 100x100 pixel image the dimension size would be reduced from 10001 to 201.

Simple Plots for various Digits can be found here:

XXX Github

Write a few sentences describing how you implemented your approach.  What languages and libraries did you use? What challenges did you run into?  
  
*Example: "I partitioned the data by gender manually using Excel.  I used Weka to build the random forest."*

The preparation step of calculating column and row means were done in Python due to much better performance and better debug possibilities. The resulting new train and test data was evaluated with R. Two models were considered, RandomForests and K-nearest Neighbour. Since RandomForests evaluate variables in isolation and pattern of variables appeared more appropriate the k-nearest Neighbour method using the FNN package was investigated further.

Caret package for the evalulation.

Write a few sentences assessing your approach.  Did it work?  What do you think the problems were?  
  
*Example: "My approach did not work so well, achieving a score of 0.65.  This is less than the sample solution.  I suspect I should not have ignored the pclass attribute."*

Benchmark 0.96557

Submission 1: 0.88186

Submission 2: 0. **88657**

Write a few sentences describing how you improved on your solution, and whether or not it worked.  
  
*Example: "I included the pclass attribute and ignored the ticket number attribute.  My score improved to 0.68.*"