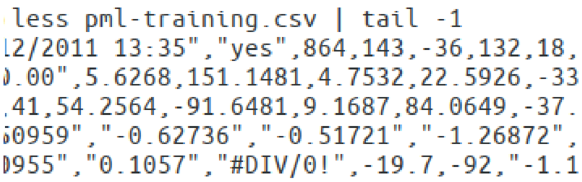
Coursera ML Project

Data files were downloaded in the same directory where the R script was executed:
> wget https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv
> wget https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv

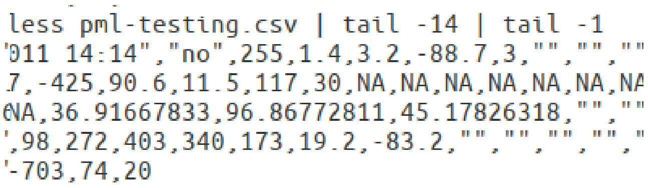
Each point below correspond to one part of the R script:

1) **Get samples and first cleansing**
   Training and testing file are loaded with **read.csv** function.
   By inspecting the raw data for some variables are missing, empty and unphysical, all those fields ("#DIV/0!", "") will be replaced with NA using the **na.strings** function

2) **More cleansing**
   Remove the irrelevant columns (first 7) and NA fields using **colSums** and **is.na** functions



```
less pml-training.csv | tail -1
12/2011 13:35","yes",864,143,-36,132,18,
).00",5.6268,151.1481,4.7532,22.5926,-33
.41,54.2564,-91.6481,9.1687,84.0649,-37.
60959","-0.62736","-0.51721","-1.26872",
0955","0.1057","#DIV/0!",-19.7,-92,"-1.1
```

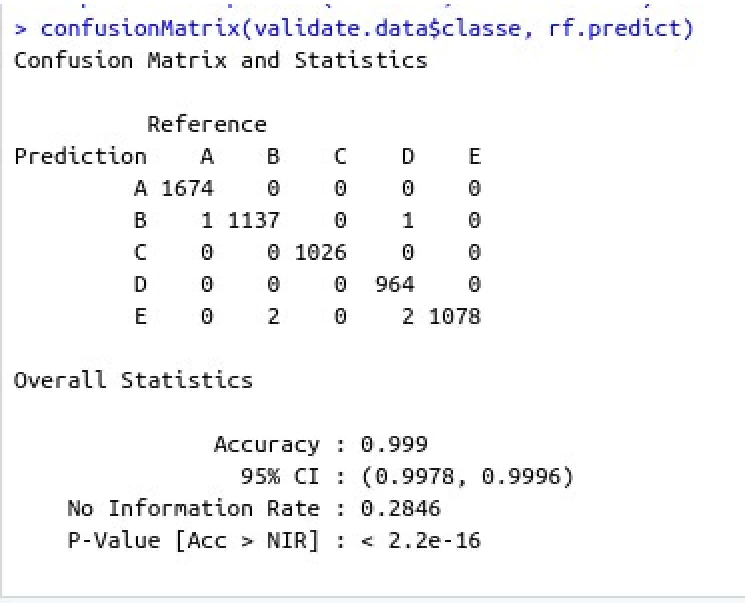


```
less pml-testing.csv | tail -14 | tail -1
'011 14:14","no",255,1.4,3.2,-88.7,3,"","",""
7,-425,90.6,11.5,117,30,NA,NA,NA,NA,NA,NA,NA
NA,36.91667833,96.86772811,45.17826318,"",""
',98,272,403,340,173,19.2,-83.2,"","","","","
'-703,74,20
```

3) **Split cleaned training Data**
   In two parts (70% and 30%) for cross validation with **createDataPartition**

4) **Training random forest model**
   Using cross validation (k-fold = 4)
   Execute training, prediction for "classe" and random-forest method and 250 trees

5) **Check performance on validation data**
   Execute the prediction on validation data with predict function from training and generate the confusion matrix, check out-of-sample errors
   Accuracy estimation with **postResample** function

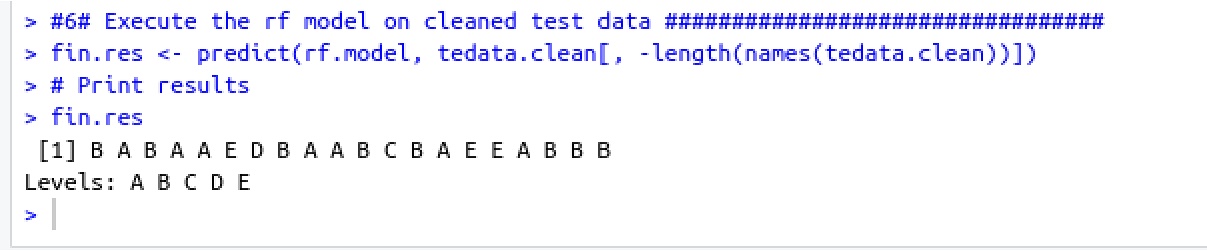


```
> confusionMatrix(validate.data$classe, rf.predict)
Confusion Matrix and Statistics

          Reference
Prediction    A    B    C    D    E
         A 1674    0    0    0    0
         B    1 1137    0    1    0
         C    0    0 1026    0    0
         D    0    0    0  964    0
         E    0    2    0    2 1078

Overall Statistics

               Accuracy : 0.999
                 95% CI : (0.9978, 0.9996)
    No Information Rate : 0.2846
    P-Value [Acc > NIR] : < 2.2e-16
```

The accurancy is of 99.9%, few out-of-samples

6) Classification of test data

The model is finally applied to the cleaned test data and the result is shown below



```
> #6# Execute the rf model on cleaned test data #################################
> fin.res <- predict(rf.model, tedata.clean[, -length(names(tedata.clean))])
> # Print results
> fin.res
 [1] B A B A A E D B A A B C B A E E A B B B
Levels: A B C D E
> 
```