STAT-702-Final-Typing

Prakash Paudyal

5/14/2018

Introduction

The dataset is a record of the time taken for typing the password of a specific system for 51 users.The system can be used by one user at a time and they have access to the password ".tie5Roanl".The data has record for two sessions where the user used the system. A session can be defined as a continious block of time where the user has to enter password multiple times. These multiple entries are recorded as rep in the data set. There are two session names called "sessionIndex 7" and "sessionIndex 8". While checking the data set, I found out that for some users, information from both the sessions have been recorded, while for some only the information from "sessionIndex 8" has been been recorded. The data has a "subject" column, which represents the 51 users for that particular observation row and are named as "soo2", "s003 etc. There are 31 other columns in the data set which are the record of time taken for typing the keys of the password. These 31 columns give the time taken to hit a key and the time interval between hiting two keys. Some of these variables have been explained below: 1) H.period: The amount of time that the"." is held down. 2) DD.period.t: The time between pressing down the "." key and pressing down the "t" key. 3) UD.period.t: The time between the "." key coming up and pressing down the "t" key. Similarly, the times for all the keys of the password ".tie5Roanl" are recorded.

Our goal is to build a classification model for known data where subject for each rows of observation is given and we need to predict the subjects for the 12 sessions of the unknown data.

Data Descriptive Statistics

The known dataset has 35 Colums and 1776 Rows.Here subject is the target variable and the remaining are indepenndent feature variables.The column "x" is the index of the data.Two columns "sessionIndex" and "rep" provide information about when and how many times the users entered the password for the system. There are two sessions where the user used the system and entered password multiple times (which is represented by rep variables). By ploting frequncey count plot for each subject, we can see that some users have participated in both the sesssions while some of them have participated in session 8 only. The number of times each subject entered the password is different e.g. subject "s011" entered the password 11 times and subject "s026 enterd password 53 times. For session 8, all the 51 subjects used the system and entered the password 5 to 50 times. For session 7, only 22 subjects used the sytem and entered the password 1 to 25 times.

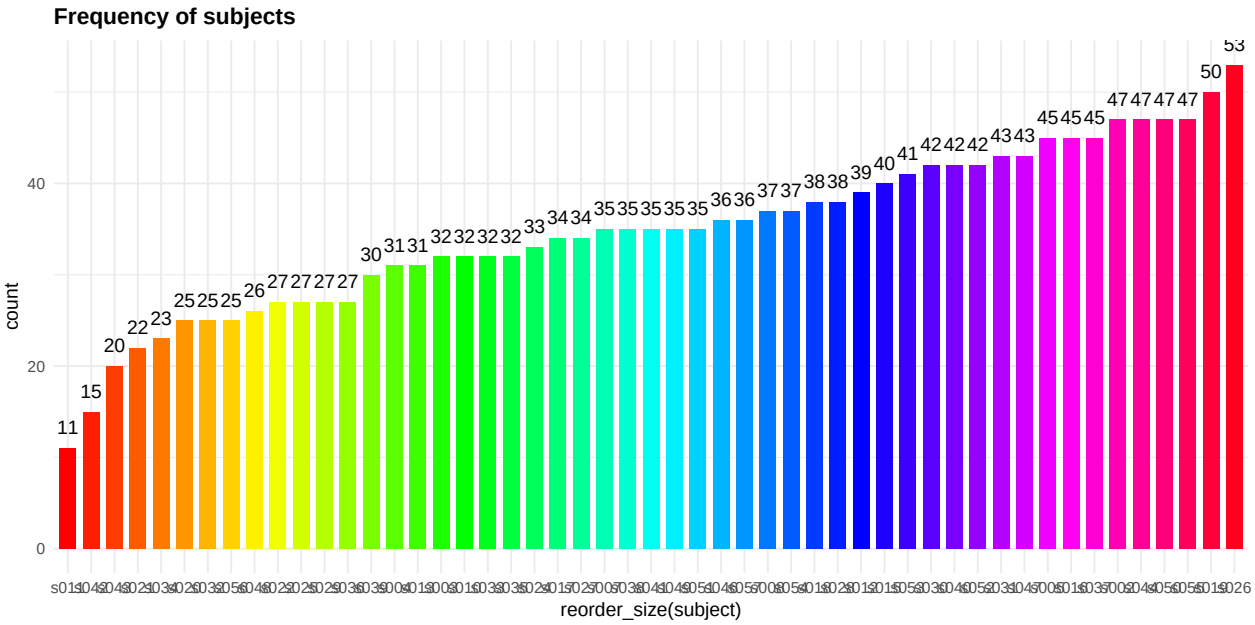
subject	sessionIndex	rep	H.period	DD.period.t	UD.period.t	H.t	DD.t.i	UD.t.i	H.i	DD.i.e	UD.i.
s002	8	1	0.1391	0.3573	0.2182	0.1481	0.2181	0.0700	0.0921	0.1742	0.082
s002	8	2	0.1399	0.2728	0.1329	0.1354	0.1893	0.0539	0.0958	0.1051	0.009
s002	8	3	0.1278	0.1913	0.0635	0.1059	0.1109	0.0050	0.0927	0.1136	0.020
s002	8	4	0.1365	0.4312	0.2947	0.1552	0.1634	0.0082	0.1152	0.1138	-0.001
s002	8	5	0.1428	0.2256	0.0828	0.0961	0.1199	0.0238	0.1154	0.1344	0.019
s002	8	6	0.1734	0.2166	0.0432	0.1454	0.1272	-0.0182	0.1193	0.1452	0.025

```
##     subject      sessionIndex        rep           H.period
##  s026   :  53   Min.   :7.000   Min.   : 1.00   Min.   :0.00690
##  s019   :  50   1st Qu.:8.000   1st Qu.:12.00   1st Qu.:0.07420
##  s002   :  47   Median :8.000   Median :25.00   Median :0.08840
```

```
##  s044   :  47   Mean   :7.803   Mean   :25.15   Mean   :0.09572
##  s050   :  47   3rd Qu.:8.000   3rd Qu.:38.00   3rd Qu.:0.10840
##  s055   :  47   Max.   :8.000   Max.   :50.00   Max.   :0.24920
##  (Other):1485
##   DD.period.t      UD.period.t           H.t              DD.t.i
##  Min.   :0.0474   Min.   :-0.1770   Min.   :0.02010   Min.   :0.0143
##  1st Qu.:0.1326   1st Qu.: 0.0361   1st Qu.:0.06788   1st Qu.:0.1138
##  Median :0.1803   Median : 0.0800   Median :0.08390   Median :0.1404
##  Mean   :0.2272   Mean   : 0.1315   Mean   :0.08900   Mean   :0.1581
##  3rd Qu.:0.2677   3rd Qu.: 0.1774   3rd Qu.:0.10472   3rd Qu.:0.1749
##  Max.   :1.8625   Max.   : 1.7892   Max.   :0.22110   Max.   :1.0841
##
##      UD.t.i              H.i             DD.i.e           UD.i.e
##  Min.   :-0.14410   Min.   :0.0288   Min.   :0.0014   Min.   :-0.12800
##  1st Qu.: 0.02423   1st Qu.:0.0623   1st Qu.:0.0810   1st Qu.:-0.00270
##  Median : 0.05230   Median :0.0795   Median :0.1116   Median : 0.02930
##  Mean   : 0.06907   Mean   :0.0849   Mean   :0.1345   Mean   : 0.04965
##  3rd Qu.: 0.08443   3rd Qu.:0.1008   3rd Qu.:0.1556   3rd Qu.: 0.07460
##  Max.   : 0.92630   Max.   :0.3312   Max.   :1.3439   Max.   : 1.24790
##
##       H.e            DD.e.five        UD.e.five           H.five
##  Min.   :0.00770   Min.   :0.0467   Min.   :-0.1257   Min.   :0.00660
##  1st Qu.:0.07068   1st Qu.:0.1981   1st Qu.: 0.1103   1st Qu.:0.06100
##  Median :0.08500   Median :0.2458   Median : 0.1573   Median :0.07590
##  Mean   :0.09513   Mean   :0.3054   Mean   : 0.2103   Mean   :0.07797
##  3rd Qu.:0.11340   3rd Qu.:0.3511   3rd Qu.: 0.2550   3rd Qu.:0.09210
##  Max.   :0.31540   Max.   :4.9618   Max.   : 4.8827   Max.   :0.18690
##
##  DD.five.Shift.r  UD.five.Shift.r    H.Shift.r        DD.Shift.r.o
##  Min.   :0.1724   Min.   :0.0980   Min.   :0.01140   Min.   :0.0636
##  1st Qu.:0.2829   1st Qu.:0.2075   1st Qu.:0.07225   1st Qu.:0.1496
##  Median :0.3518   Median :0.2768   Median :0.09470   Median :0.1912
##  Mean   :0.3967   Mean   :0.3187   Mean   :0.09819   Mean   :0.2331
##  3rd Qu.:0.4539   3rd Qu.:0.3700   3rd Qu.:0.11910   3rd Qu.:0.2645
##  Max.   :2.9162   Max.   :2.8405   Max.   :0.23900   Max.   :4.1523
##
##   UD.Shift.r.o          H.o              DD.o.a           UD.o.a
##  Min.   :-0.0615   Min.   :0.02460   Min.   :0.0012   Min.   :-0.11310
##  1st Qu.: 0.0425   1st Qu.:0.07225   1st Qu.:0.1019   1st Qu.: 0.01220
##  Median : 0.0899   Median :0.08760   Median :0.1268   Median : 0.03860
##  Mean   : 0.1350   Mean   :0.09017   Mean   :0.1468   Mean   : 0.05663
##  3rd Qu.: 0.1712   3rd Qu.:0.10422   3rd Qu.:0.1631   3rd Qu.: 0.07000
##  Max.   : 4.0120   Max.   :0.23870   Max.   :2.5222   Max.   : 2.43700
##
##       H.a              DD.a.n           UD.a.n              H.n
##  Min.   :0.02010   Min.   :0.0011   Min.   :-0.18290   Min.   :0.01230
##  1st Qu.:0.08607   1st Qu.:0.0904   1st Qu.:-0.01740   1st Qu.:0.06970
##  Median :0.10530   Median :0.1165   Median : 0.01140   Median :0.08865
##  Mean   :0.10925   Mean   :0.1380   Mean   : 0.02871   Mean   :0.09230
##  3rd Qu.:0.12490   3rd Qu.:0.1628   3rd Qu.: 0.05158   3rd Qu.:0.10983
##  Max.   :0.34710   Max.   :2.1814   Max.   : 2.07750   Max.   :0.34520
##
##      DD.n.l           UD.n.l               H.l           DD.l.Return
##  Min.   :0.0013   Min.   :-0.175800   Min.   :0.00610   Min.   :0.0210
```

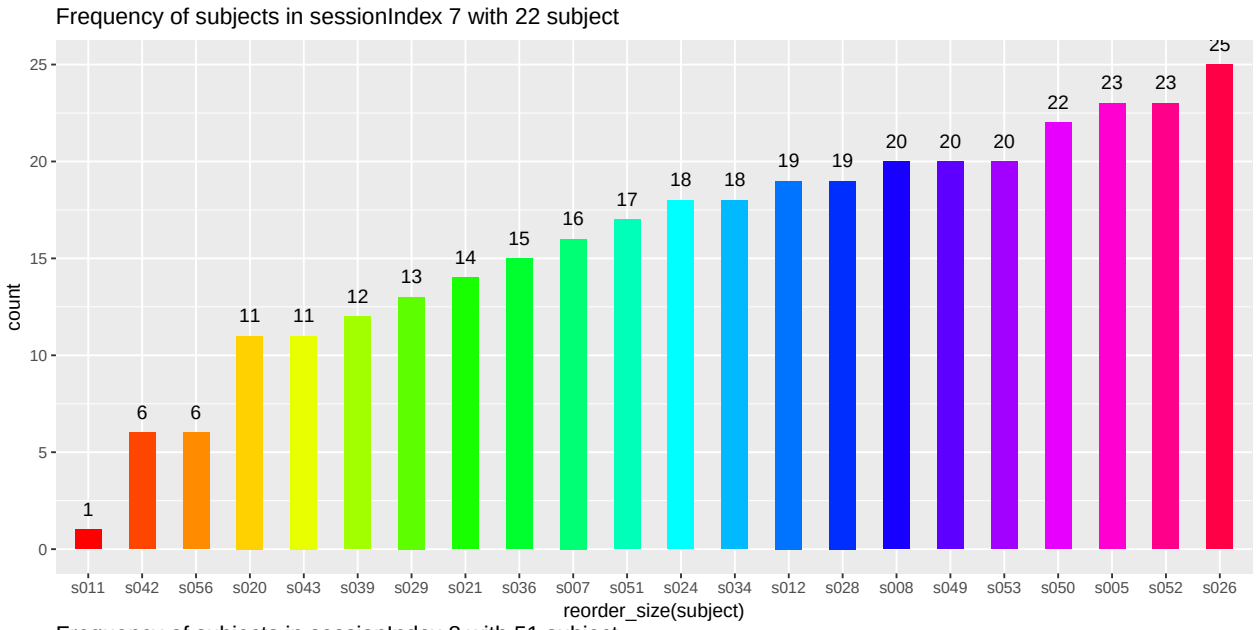
```
##  1st Qu.:0.1154   1st Qu.: 0.008075   1st Qu.:0.08095   1st Qu.:0.2055
##  Median :0.1652   Median : 0.079650   Median :0.09765   Median :0.2488
##  Mean   :0.1791   Mean   : 0.086764   Mean   :0.09861   Mean   :0.2904
##  3rd Qu.:0.2132   3rd Qu.: 0.129800   3rd Qu.:0.11400   3rd Qu.:0.3274
##  Max.   :1.8716   Max.   : 1.799000   Max.   :0.24870   Max.   :1.7679
##
##   UD.l.Return         H.Return
##  Min.   :-0.0646   Min.   :0.01060
##  1st Qu.: 0.1051   1st Qu.:0.07150
##  Median : 0.1455   Median :0.08760
##  Mean   : 0.1918   Mean   :0.09073
##  3rd Qu.: 0.2312   3rd Qu.:0.10530
##  Max.   : 1.6119   Max.   :0.25660
##
```

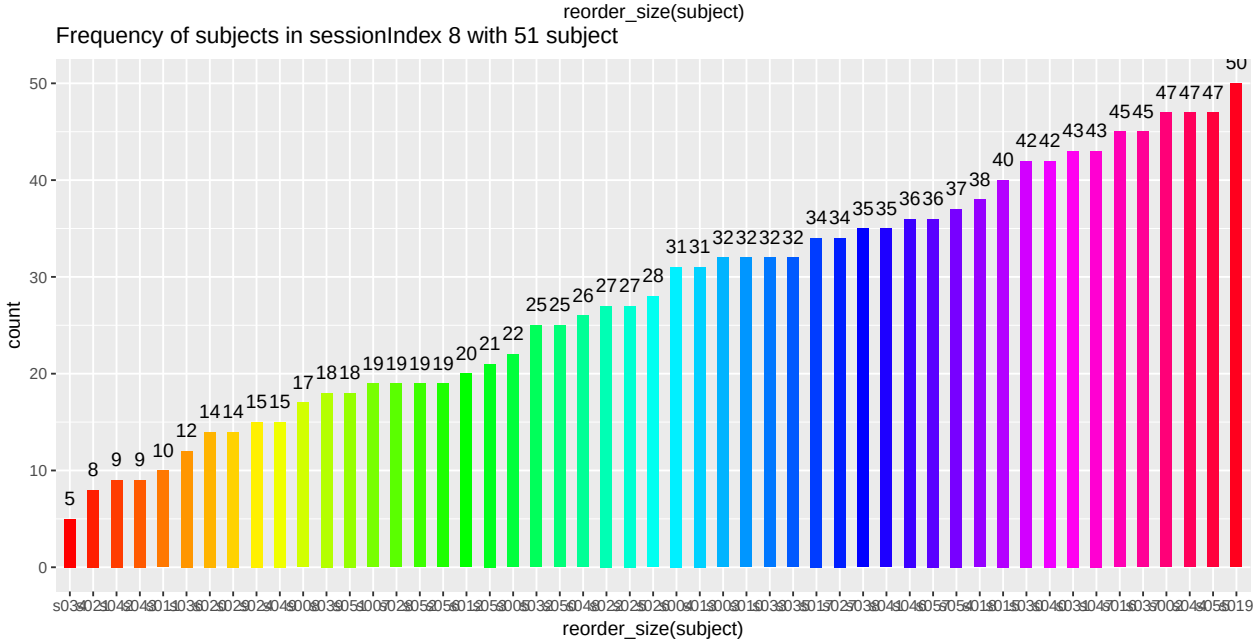
Frequency of subjects



Checking distribution of subjects in each of two session index

	7	8
s002	0	47
s003	0	32
s004	0	31
s005	23	22
s007	16	19
s008	20	17





Exploratory data analysis and feature selection

Checking missing data

There is no missing data.

```
##          subject    sessionIndex             rep       H.period
##                0               0               0              0
##     DD.period.t     UD.period.t             H.t          DD.t.i
##                0               0               0              0
##          UD.t.i             H.i          DD.i.e          UD.i.e
##                0               0               0              0
##             H.e       DD.e.five       UD.e.five         H.five
```

```
##               0               0               0               0
## DD.five.Shift.r UD.five.Shift.r       H.Shift.r    DD.Shift.r.o
##               0               0               0               0
##    UD.Shift.r.o             H.o          DD.o.a          UD.o.a
##               0               0               0               0
##             H.a          DD.a.n          UD.a.n             H.n
##               0               0               0               0
##          DD.n.l          UD.n.l             H.l     DD.l.Return
##               0               0               0               0
##     UD.l.Return        H.Return
##               0               0
```

Correlations:

In statistics, the correlation coefficient r measures the strength and direction of a linear relationship between two variables on a scatterplot. The value of r is always between +1 and –1 The correlations plot and tables shows us that DD. and UD. predictors are higly correlated.

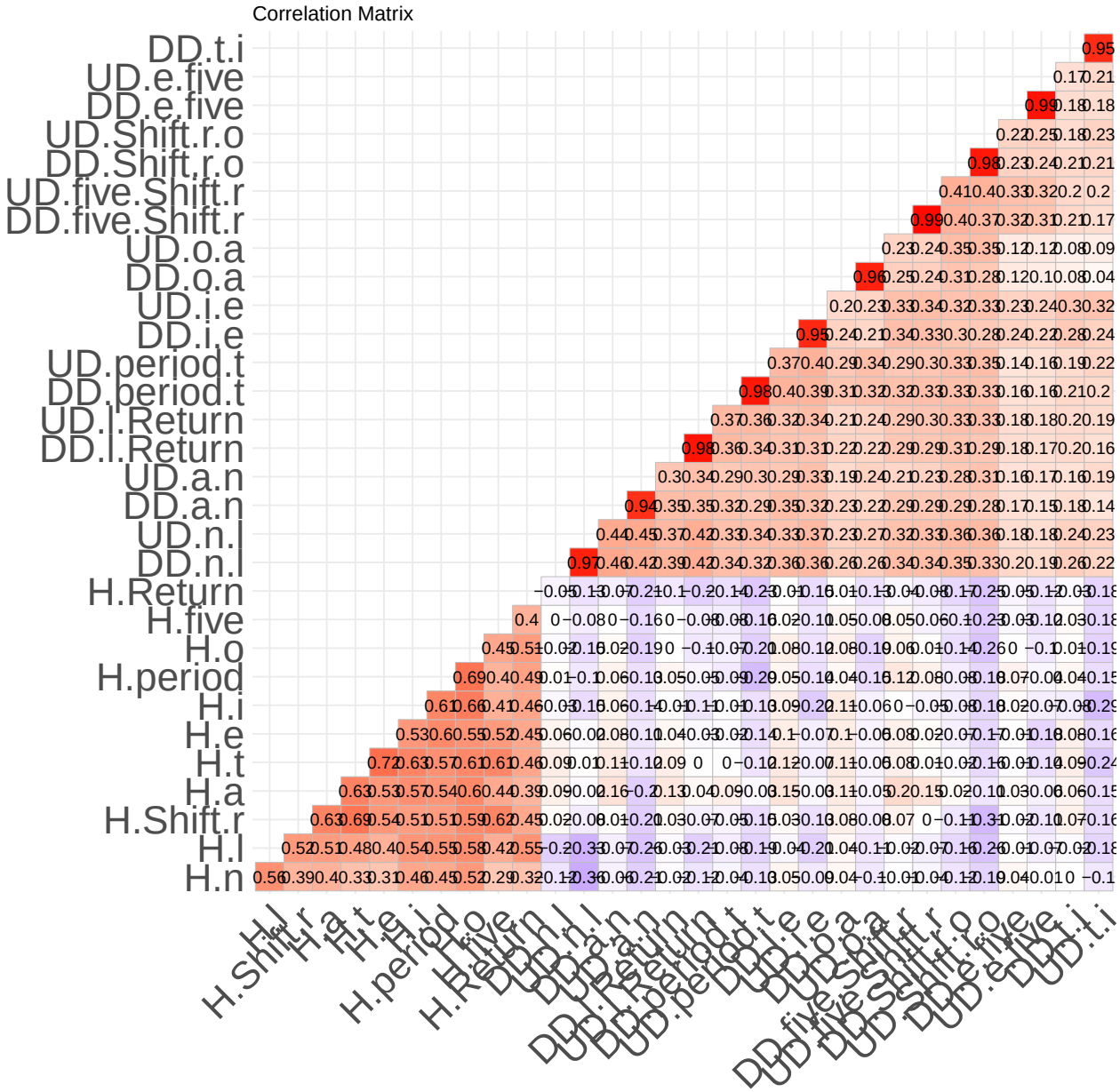
Correlation Table

Correlation table for highly correlaated variables show very strong correlation between the DD and UD variables.

Table 3: Correlation Table

	First.Variable	Second.Variable	Correlation
448	DD.five.Shift.r	UD.five.Shift.r	0.993204
352	DD.e.five	UD.e.five	0.985284
928	DD.l.Return	UD.l.Return	0.984093
544	DD.Shift.r.o	UD.Shift.r.o	0.979723
64	DD.period.t	UD.period.t	0.978750
832	DD.n.l	UD.n.l	0.967695
640	DD.o.a	UD.o.a	0.964335
256	DD.i.e	UD.i.e	0.951695
160	DD.t.i	UD.t.i	0.945435
736	DD.a.n	UD.a.n	0.935253
283	H.t	H.e	0.718687
559	H.period	H.o	0.692425
469	H.t	H.Shift.r	0.689209
565	H.i	H.o	0.656822
655	H.t	H.a	0.629468
190	H.t	H.i	0.627443
667	H.Shift.r	H.a	0.625492
478	H.five	H.Shift.r	0.620931
376	H.t	H.five	0.614709
187	H.period	H.i	0.613479

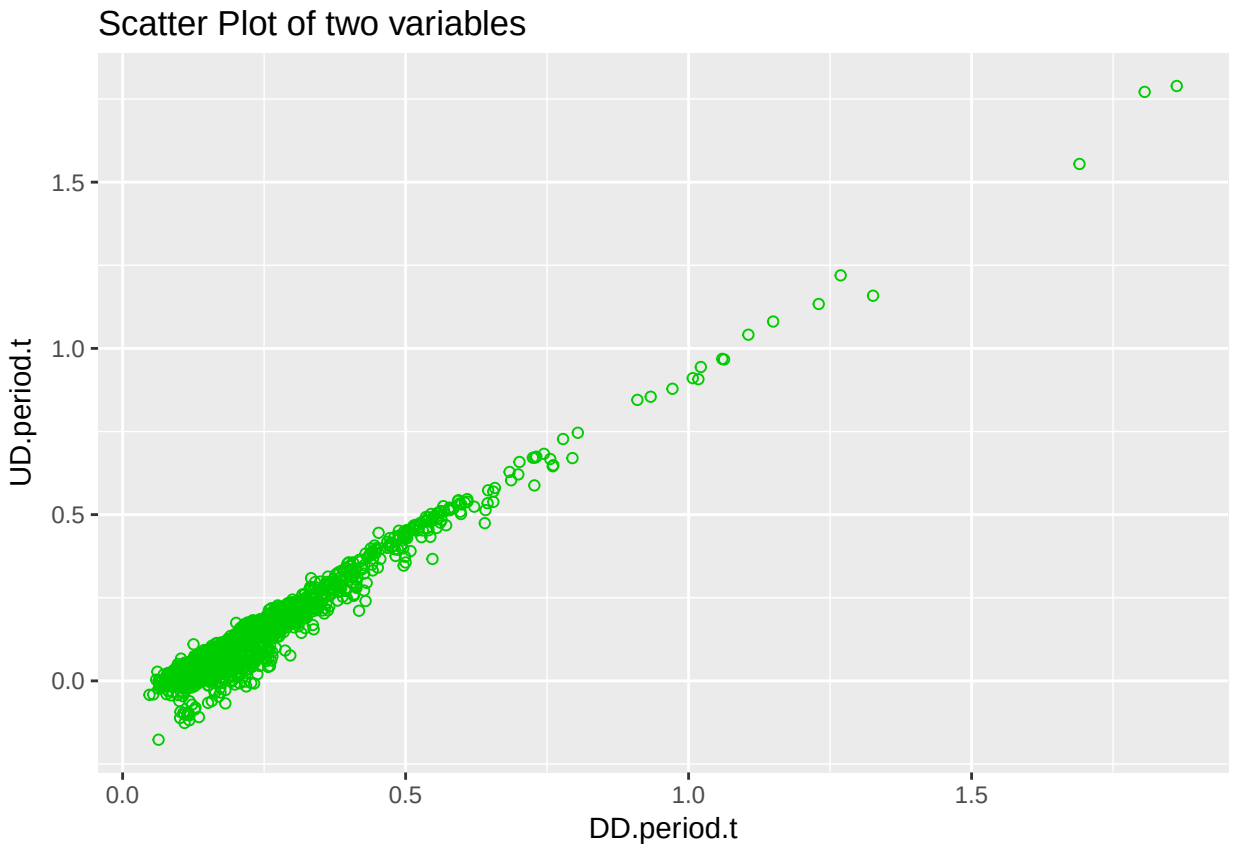
Correlation plot

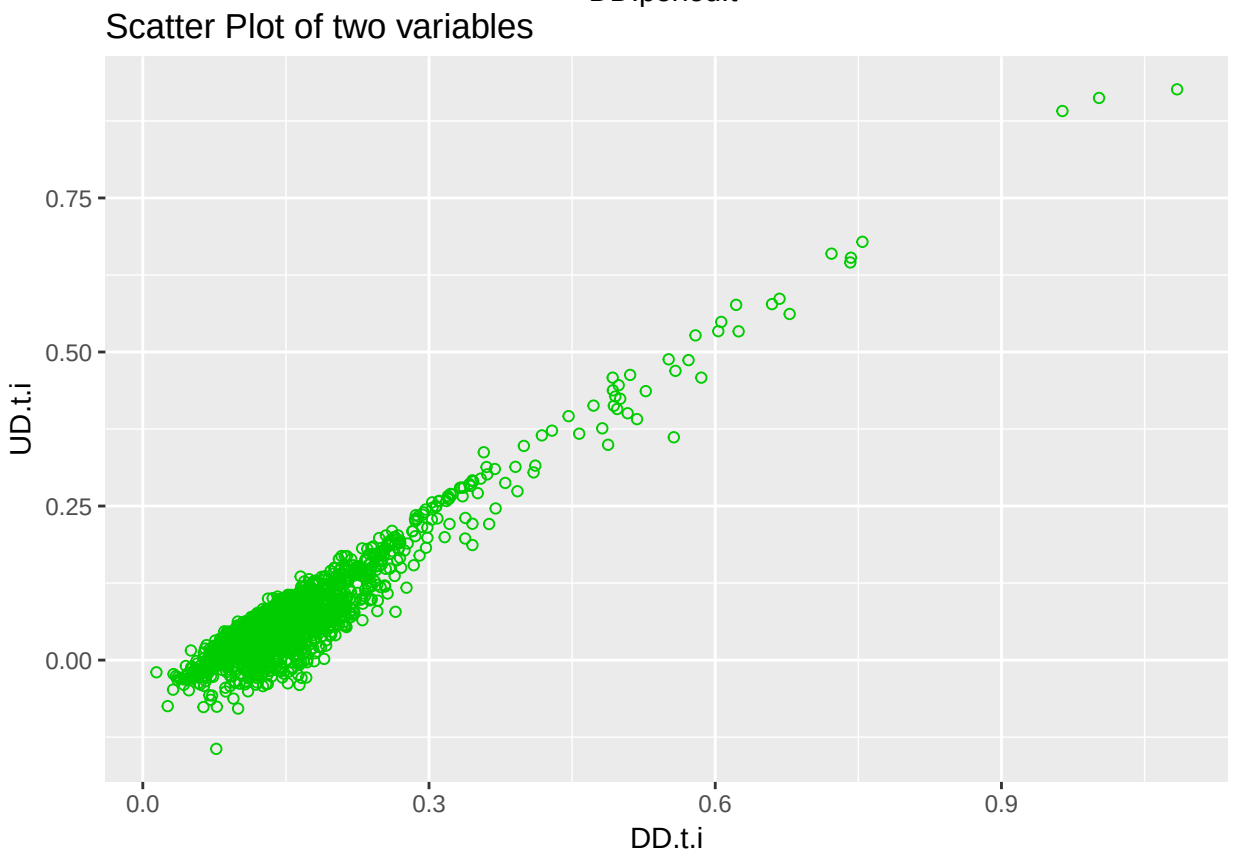


Linear relationship between the variables-correlation

Some of the variables has linear relationship which indicate they are correleted.

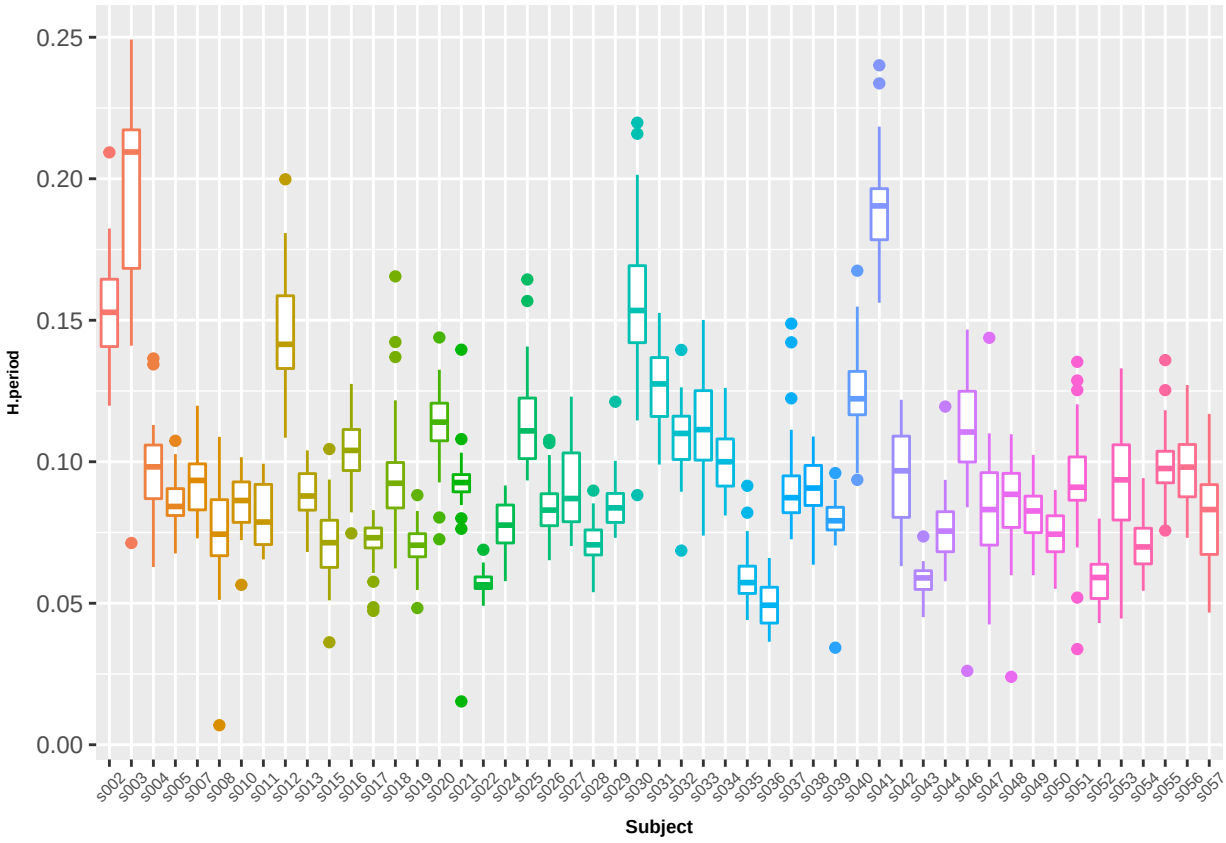
```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

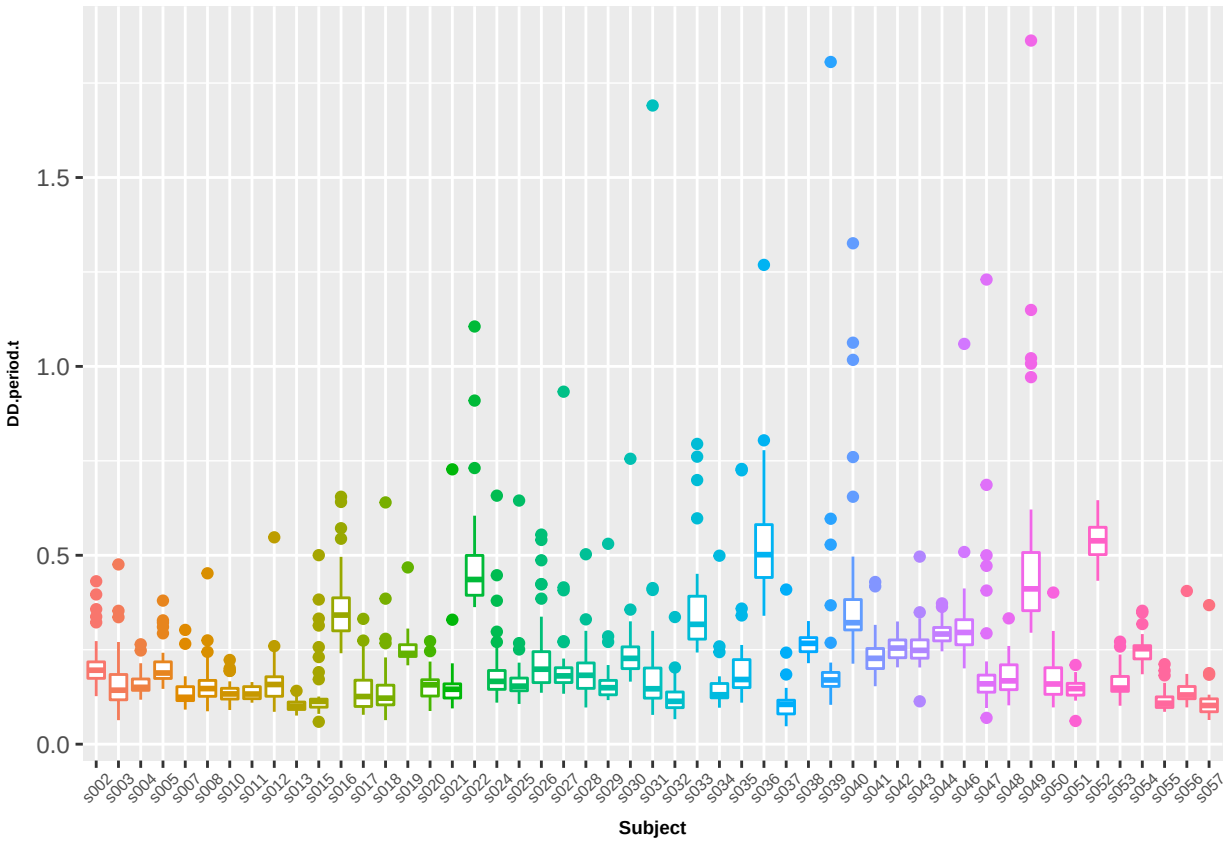

Scatter Plot of two variables
UD.period.t
DD.period.t
0.0
0.5
1.0
1.5

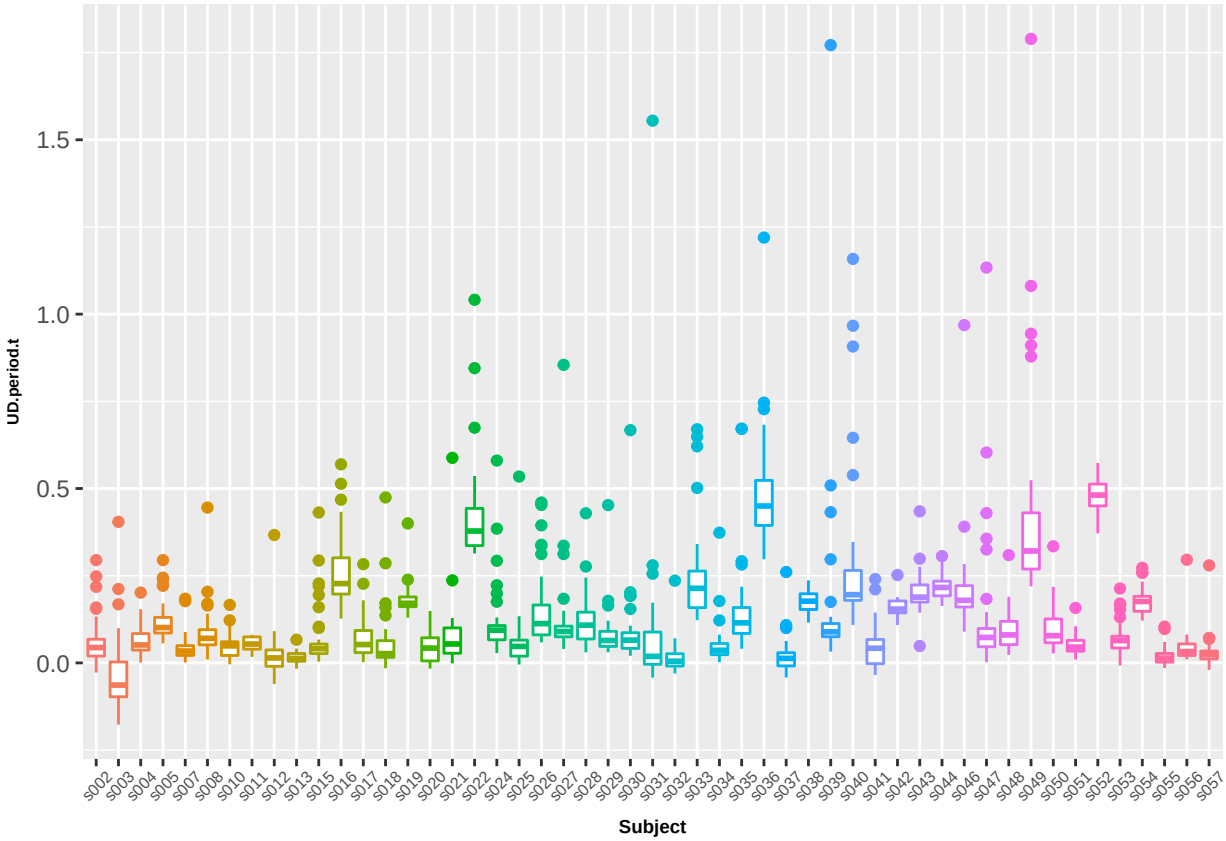

Scatter Plot of two variables
UD.t.i
DD.t.i
0.00
0.25
0.50
0.75
0.0
0.3
0.6
0.9

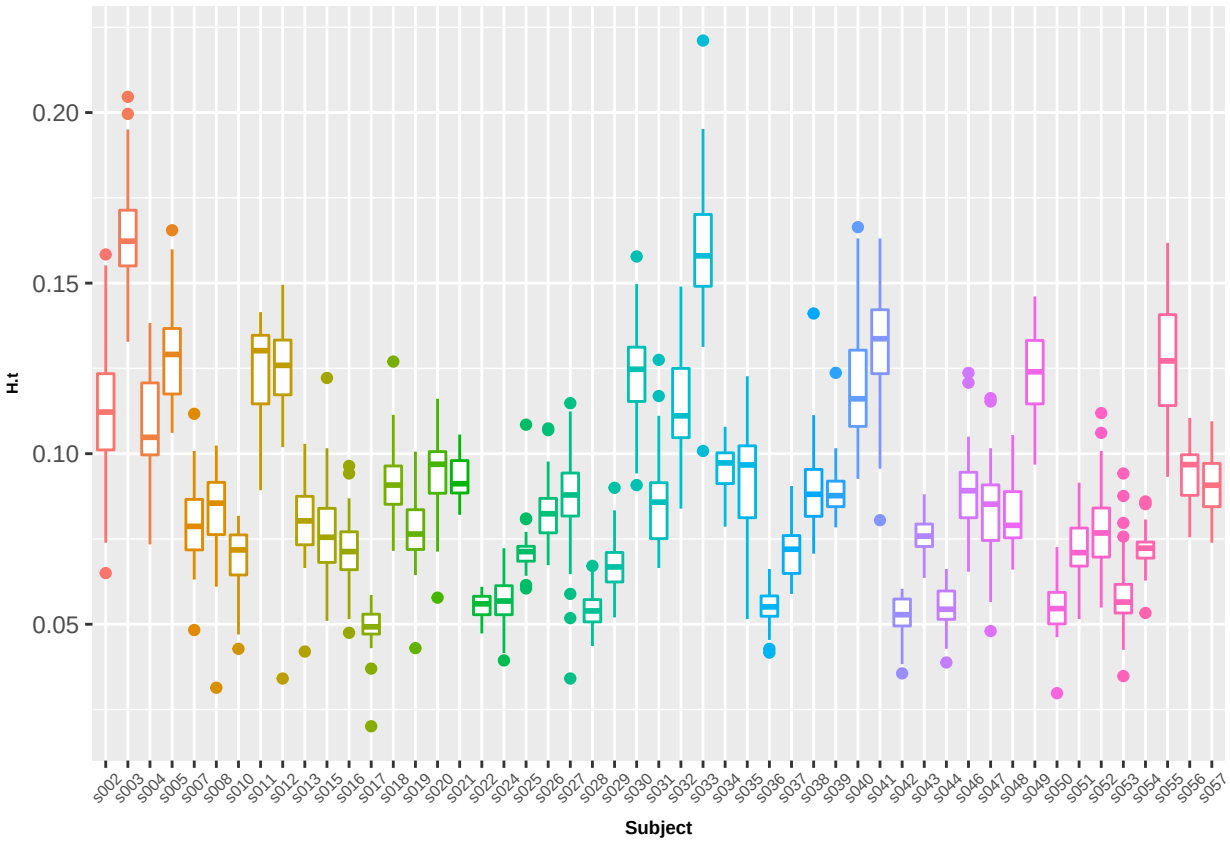
Box plot subject vs all predictor

The following figure shows the box plot between H predictors and all 51 classes. It can be clearly seen that the predictors are seperable in each class. Hence, in early stage of data exploration I noticed thAT these predictors would play significant role in the development of classifiers. ut for DD and UD some of them are ovelaping and some are not overlaping hence we still can say there is difference among the 51 variables.



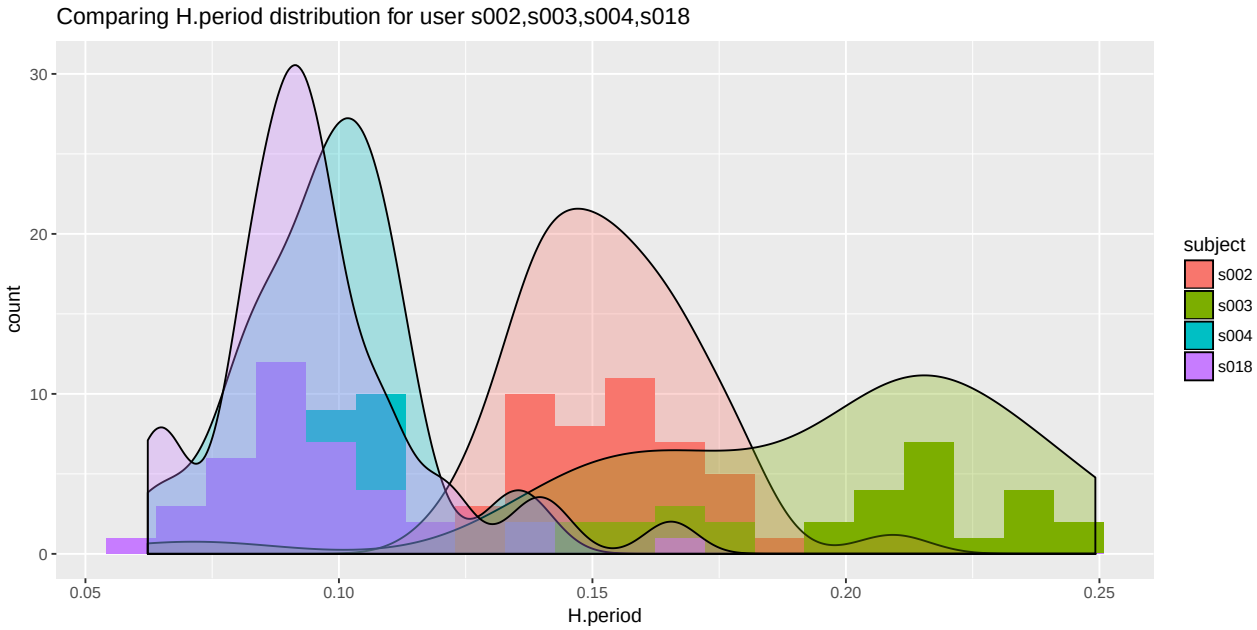

DD.period.t
1.5
1.0
0.5
0.0
s002 s003 s004 s005 s007 s008 s010 s011 s012 s013 s015 s016 s017 s018 s019 s020 s021 s022 s024 s025 s026 s027 s028 s029 s030 s031 s032 s033 s034 s035 s036 s037 s038 s039 s040 s041 s042 s043 s044 s046 s047 s048 s049 s050 s051 s052 s053 s054 s055 s056 s057
Subject


UD.period.t
1.5
1.0
0.5
0.0
s002 s003 s004 s005 s007 s008 s010 s011 s012 s013 s015 s016 s017 s018 s019 s020 s021 s022 s024 s025 s026 s027 s028 s029 s030 s031 s032 s033 s034 s035 s036 s037 s038 s039 s040 s041 s042 s043 s044 s046 s047 s048 s049 s050 s051 s052 s053 s054 s055 s056 s057
Subject



Density Plot for Hold time to compare the subject

Since Hold time looks better predictor among other predictor hence I decided to plot density plot for Hold time to compare the subjects.To build the good classifier we need to know what variables give the correct information about subjects.I ploted density plot for Hold time with subjects to check weather the Hold time is different for subject or not .



Creating Training and Test dataset

I used different methods to split data inorder to make sure good proportion of subject and session in each training and test data. I used sample() and creatPartition() function from r package. The Sample of traing data set is 70% and remaining 30 % is test data.

count the each subject in train and test data

```
## Warning: package 'caret' was built under R version 3.4.4
```

count the each subject in train and test data

```
##
## s002 s003 s004 s005 s007 s008 s010 s011 s012 s013 s015 s016 s017 s018 s019
##   47   32   31   45   35   37   32   11   39   31   40   45   34   38   50
## s020 s021 s022 s024 s025 s026 s027 s028 s029 s030 s031 s032 s033 s034 s035
##   25   22   27   33   27   53   34   38   27   42   43   25   32   23   32
## s036 s037 s038 s039 s040 s041 s042 s043 s044 s046 s047 s048 s049 s050 s051
##   27   45   35   30   42   35   15   20   47   36   43   26   35   47   35
## s052 s053 s054 s055 s056 s057
##   42   41   37   47   25   36
```

```
##
## s002 s003 s004 s005 s007 s008 s010 s011 s012 s013 s015 s016 s017 s018 s019
##   33   23   22   32   25   26   23    8   28   22   28   32   24   27   35
## s020 s021 s022 s024 s025 s026 s027 s028 s029 s030 s031 s032 s033 s034 s035
##   18   16   19   24   19   38   24   27   19   30   31   18   23   17   23
## s036 s037 s038 s039 s040 s041 s042 s043 s044 s046 s047 s048 s049 s050 s051
##   19   32   25   21   30   25   11   14   33   26   31   19   25   33   25
## s052 s053 s054 s055 s056 s057
##   30   29   26   33   18   26
```

```
##
## s002 s003 s004 s005 s007 s008 s010 s011 s012 s013 s015 s016 s017 s018 s019
##   14    9    9   13   10   11    9    3   11    9   12   13   10   11   15
## s020 s021 s022 s024 s025 s026 s027 s028 s029 s030 s031 s032 s033 s034 s035
##    7    6    8    9    8   15   10   11    8   12   12    7    9    6    9
## s036 s037 s038 s039 s040 s041 s042 s043 s044 s046 s047 s048 s049 s050 s051
##    8   13   10    9   12   10    4    6   14   10   12    7   10   14   10
## s052 s053 s054 s055 s056 s057
##   12   12   11   14    7   10
```

```
##
## s002 s003 s004 s005 s007 s008 s010 s011 s012 s013 s015 s016 s017 s018 s019
##   33   23   22   15   13   11   23    8   13   22   28   32   24   27   35
## s020 s021 s022 s024 s025 s026 s027 s028 s029 s030 s031 s032 s033 s034 s035
##   10    4   19   11   19   22   24   12    9   30   31   18   23    4   23
## s036 s037 s038 s039 s040 s041 s042 s043 s044 s046 s047 s048 s049 s050 s051
##    7   32   25   10   30   25    6    8   33   26   31   19   12   17   10
## s052 s053 s054 s055 s056 s057
##   13   14   26   33   14   26
```

```
##
## s002 s003 s004 s005 s007 s008 s010 s011 s012 s013 s015 s016 s017 s018 s019
##    0    0    0   17   12   15    0    0   15    0    0    0    0    0    0
## s020 s021 s022 s024 s025 s026 s027 s028 s029 s030 s031 s032 s033 s034 s035
##    8   12    0   13    0   16    0   15   10    0    0    0    0   13    0
## s036 s037 s038 s039 s040 s041 s042 s043 s044 s046 s047 s048 s049 s050 s051
##   12    0    0   11    0    0    5    6    0    0    0    0   13   16   15
## s052 s053 s054 s055 s056 s057
##   17   15    0    0    4    0
```

```
##
## s002 s003 s004 s005 s007 s008 s010 s011 s012 s013 s015 s016 s017 s018 s019
##   14    9    9    7    6    6    9    2    7    9   12   13   10   11   15
## s020 s021 s022 s024 s025 s026 s027 s028 s029 s030 s031 s032 s033 s034 s035
##    4    4    8    4    8    6   10    7    5   12   12    7    9    1    9
## s036 s037 s038 s039 s040 s041 s042 s043 s044 s046 s047 s048 s049 s050 s051
##    5   13   10    8   12   10    3    1   14   10   12    7    3    8    8
## s052 s053 s054 s055 s056 s057
##    6    7   11   14    5   10
```

```
##
## s002 s003 s004 s005 s007 s008 s010 s011 s012 s013 s015 s016 s017 s018 s019
##    0    0    0    6    4    5    0    1    4    0    0    0    0    0    0
## s020 s021 s022 s024 s025 s026 s027 s028 s029 s030 s031 s032 s033 s034 s035
##    3    2    0    5    0    9    0    4    3    0    0    0    0    5    0
## s036 s037 s038 s039 s040 s041 s042 s043 s044 s046 s047 s048 s049 s050 s051
##    3    0    0    1    0    0    1    5    0    0    0    0    7    6    2
## s052 s053 s054 s055 s056 s057
##    6    5    0    0    2    0
```

Models

1: Multinomial logistics regression

Multinomial logistic regression is used to model nomial outcome variables,in which the log odds of the outcomes are modeled as a linear combination of the predictor variables.Multinomial regression report the odds of being in the different outcome catogories in reference to some base group. Multiple regression model, for k possible outcomes, runnning k-1 independent binary logistic regression,in which one outcome is chossen as a pivot and then k-1 outcome are seperatly regressed against the pivot outcome. Here, I build model with H vairiables and combination of the H and UD variables.I used validation set approach and 5-fold cross validation approach to select the model. Then I calculated the model accuracy for both approachs.The accuracy table is given below.

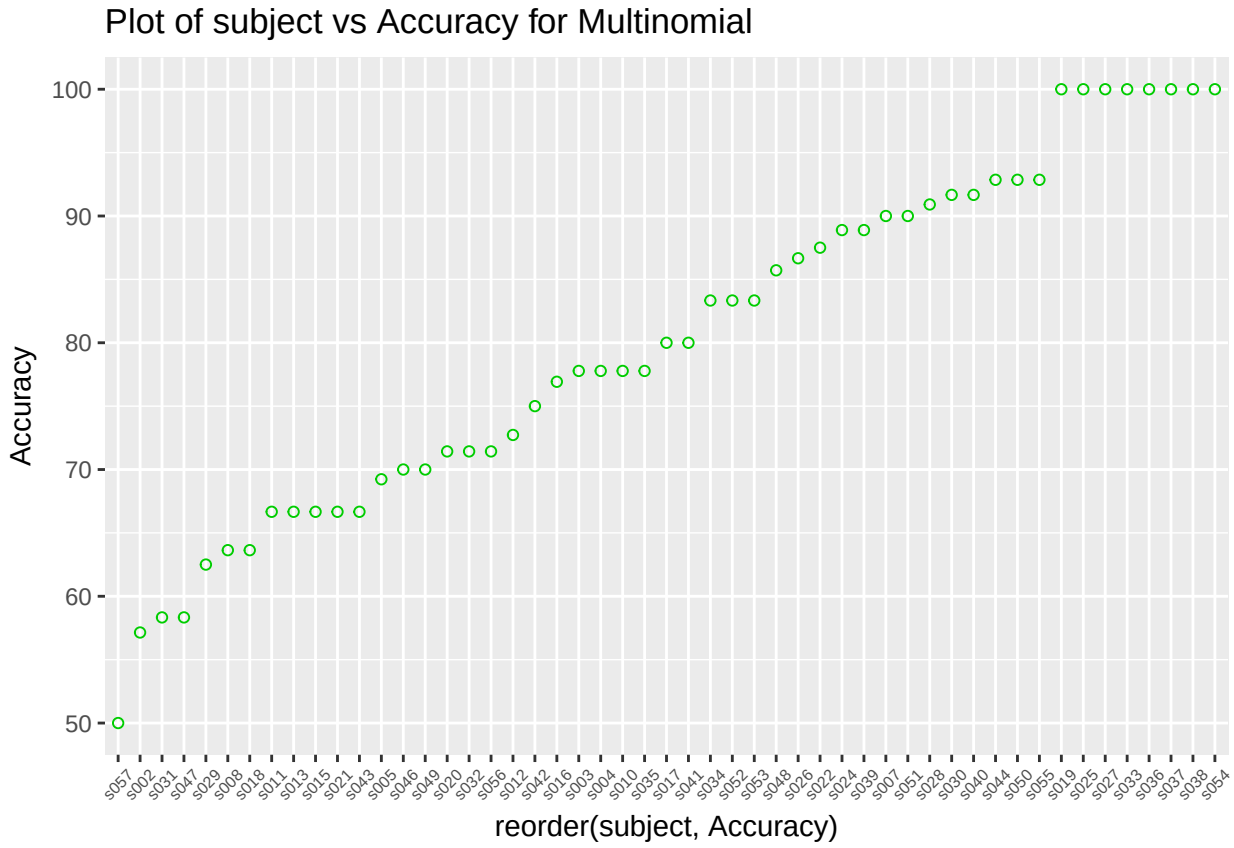
Table 1:Accuracy Table for Multinomial Model

Acuracy table shows that multinomial predict rate is good with training data but there is alot variance for test data .Test accuracy rate is decreased by 17% from the train error rate which is 98%. 5-fold corss validation slightly improved the error rate to 84.57%. subject wise accuracy table and plot are given below. Plot showes that prediction for 15 subjects (s004,s011,s017 ets) are 100% acurate. some subject were poorly predicted (s034,s002,s029 etc) Around 66% of subject has prediction acuracy above 80%.

Model	Data	Accuracy
Multinomial	Train	0.988142
Multinomial	Test	0.808219
Multinomial	5-fold	0.824892

Acuracy of Multinomial regression on test Data for each subject.

	subject	Actual	Predicted	Accuracy
15	s019	15	15	100.00000
20	s025	8	8	100.00000
22	s027	10	10	100.00000
28	s033	9	9	100.00000
31	s036	8	8	100.00000
32	s037	13	13	100.00000
33	s038	10	10	100.00000
48	s054	11	11	100.00000
39	s044	14	13	92.85714
44	s050	14	13	92.85714



2:Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis find the linear combination of original variables that provide the best possible seperation between the group.The besic purpose is to estimate relationship between a single categorical dependent variable and a set of quantitative independent variables. Here I used Lda() from Mass Package to perform this analysis using the H and UD Predictors.

LDA with PCA

```
##  [1] 0.2395465 0.4284722 0.4952757 0.5568685 0.6099270 0.6593981 0.7053277
##  [8] 0.7483145 0.7860626 0.8225439 0.8567681 0.8829390 0.9033983 0.9235181
## [15] 0.9393576 0.9520041 0.9640572 0.9750689 0.9848609 0.9932310 1.0000000
## [22] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## [29] 1.0000000 1.0000000 1.0000000
```

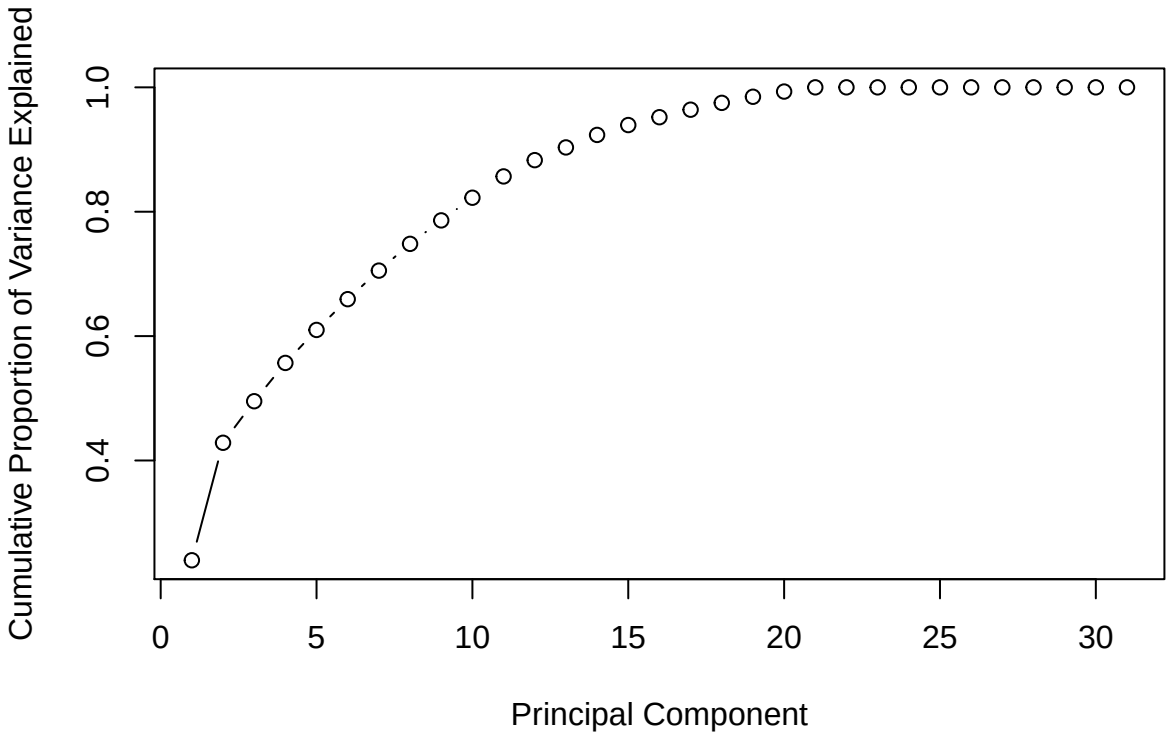


Table 2:Accuracy Table for LDA Model

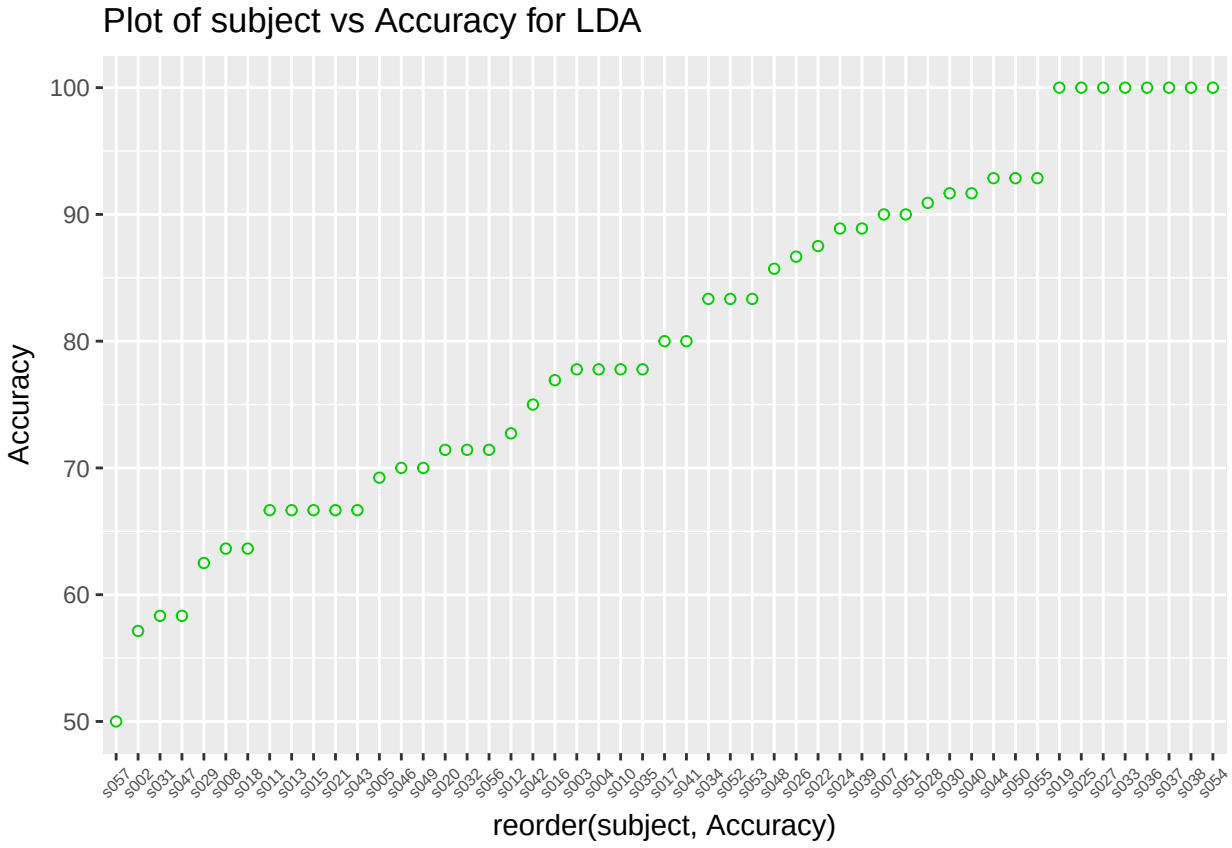
Acuracy table shows that LDA predict rate is good with training data but there is less variance for test data .Train and test accuracy rate are 90.46% and 87.16% for validation set approach.Loocv and 5 -fold model gave similar test errror as in VSA.I implented LDA-PCA to reduce the dimension of the predictior space but it did not improve the test accuracy rate. subject wise accuracy table and plot are given below. Plot showes similar pattern for accuracy as in multinimial.

Model	Data	Accuracy
LDA	Train	0.911462450592885
LDA	Test	0.863013698630137
LDA	LOOCV	0.873873873873874
LDA	5-FOLD	0.875525
LDA-PCA	Train	0.911462450592885
LDA-PCA	Test	0.863013698630137

Accuracy by subject for LDA on test data

```
##  [1] "Table: Table12:LDA"
##  [2] ""
##  [3] "      subject   Actual   Predicted    Accuracy"
##  [4] "---  --------  -------  ----------  ----------"
##  [5] "7        s010        9           9   100.00000"
##  [6] "13       s017       10          10   100.00000"
##  [7] "15       s019       15          15   100.00000"
##  [8] "17       s021        6           6   100.00000"
```

```
##  [9] "18       s022        8           8   100.00000"
## [10] "19       s024        9           9   100.00000"
## [11] "20       s025        8           8   100.00000"
## [12] "22       s027       10          10   100.00000"
## [13] "28       s033        9           9   100.00000"
## [14] "31       s036        8           8   100.00000"
## [15] "32       s037       13          13   100.00000"
## [16] "33       s038       10          10   100.00000"
## [17] "36       s041       10          10   100.00000"
```



Random Forest:

Random Forest build a number of decision trees on bootstraped training samples.When building these trees,each time a split in a tree is considred , a random sample of m predictors is choosen as split candidates from the full set of p predictors.The split is allowed to use only one of those predictors.A fresh sample of m predictors is taken at each split.m=rootsquared p. Random Forest reduce the correlation betweeb trees. Here,I used randomForest funtion from R package to implement the random forest algorithm.I have use all predictors except rep and sessionIndex.

```
## [1] 1
```

```
## [1] 0.953033
```

Accuracy Table for Random Forest

The accuracy table for random forest is given bellow.Random forest provided 100% accuracy for Training data and 95.30% for test accuracy in validation set approach.I used LOOCV model to check if i can get more accuracy but it did not improve the accuracy.5-fold model imroved accuracy very slightly. The subject wise accuracy table given below shows the % of acuracy for each subject.Accuracy plots shows most of the subject were predicted with 100% accuracy.

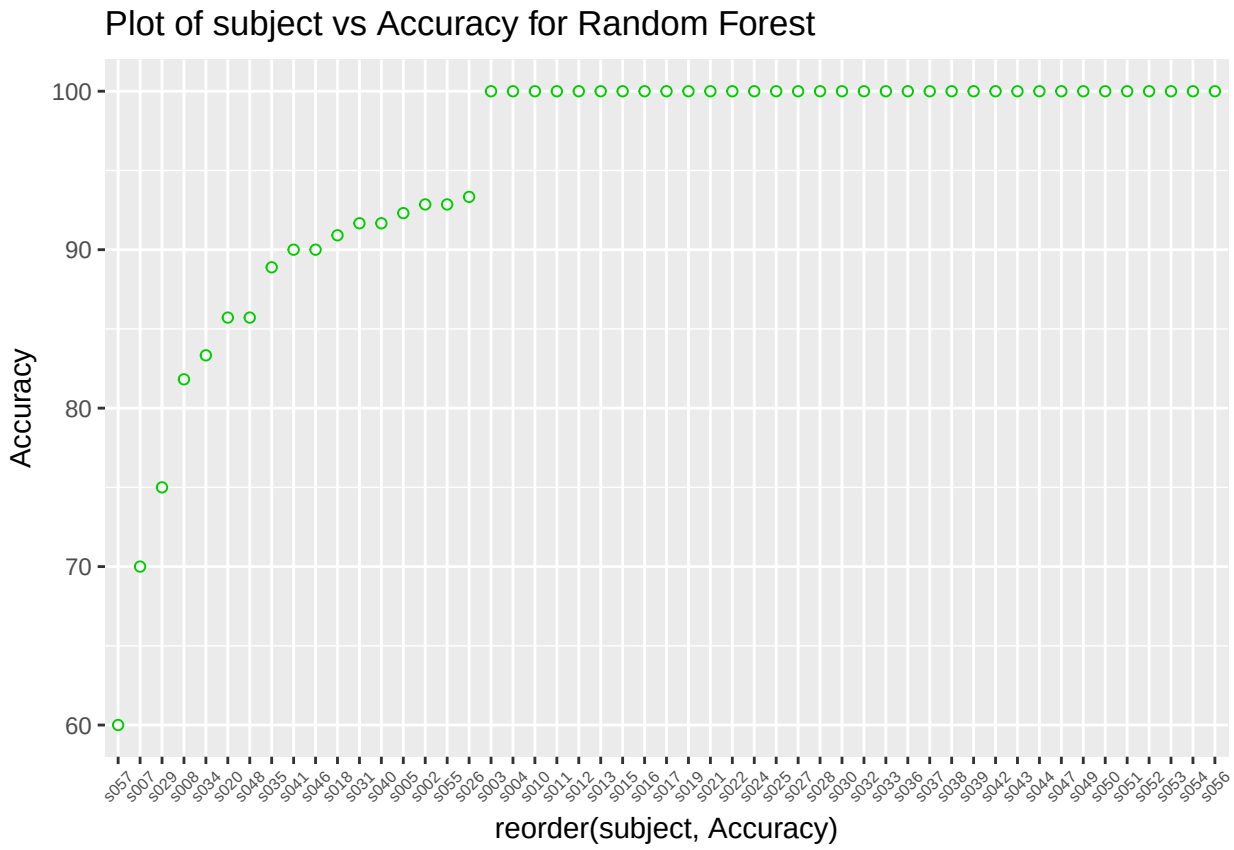
Model	Data	Accuracy
Random Forest	Train	1
Random Forest	Test	0.953033
Random Forest	Loocv	0.95214
Random Forest	5-fold	0.954359

Acuracy of subject for Random Forest

x

Table: Table13:Random Forest

subject	Actual	Predicted	Accuracy
2	s003	9	9 100.00000
3	s004	9	9 100.00000
7	s010	9	9 100.00000
8	s011	3	3 100.00000
9	s012	11	11 100.00000
10	s013	9	9 100.00000
11	s015	12	12 100.00000
12	s016	13	13 100.00000
13	s017	10	10 100.00000
15	s019	15	15 100.00000
17	s021	6	6 100.00000
18	s022	8	8 100.00000
19	s024	9	9 100.00000
20	s025	8	8 100.00000
22	s027	10	10 100.00000
23	s028	11	11 100.00000



Conclusion:

Among the models Random forest provided good accuray for predicting subject, hence decided to use Random Forest to predict final unknown data session.The accuracy rate for random forest is 95.3033 %. The subjectwise accuracy was 100 % for most of the subject and most of other subject has accuracy rate above 80% Hence I decided to use Random forest to predict final data.

Appendix

Table 9: Table11:MUltinom

	subject	Actual	Predicted	Accuracy
15	s019	15	15	100.00000
20	s025	8	8	100.00000
22	s027	10	10	100.00000
28	s033	9	9	100.00000
31	s036	8	8	100.00000
32	s037	13	13	100.00000
33	s038	10	10	100.00000
48	s054	11	11	100.00000
39	s044	14	13	92.85714
44	s050	14	13	92.85714
49	s055	14	13	92.85714

	subject	Actual	Predicted	Accuracy
25	s030	12	11	91.66667
35	s040	12	11	91.66667
23	s028	11	10	90.90909
5	s007	10	9	90.00000
45	s051	10	9	90.00000
19	s024	9	8	88.88889
34	s039	9	8	88.88889
18	s022	8	7	87.50000
21	s026	15	13	86.66667
42	s048	7	6	85.71429
29	s034	6	5	83.33333
46	s052	12	10	83.33333
47	s053	12	10	83.33333
13	s017	10	8	80.00000
36	s041	10	8	80.00000
2	s003	9	7	77.77778
3	s004	9	7	77.77778
7	s010	9	7	77.77778
30	s035	9	7	77.77778
12	s016	13	10	76.92308
37	s042	4	3	75.00000
9	s012	11	8	72.72727
16	s020	7	5	71.42857
27	s032	7	5	71.42857
50	s056	7	5	71.42857
40	s046	10	7	70.00000
43	s049	10	7	70.00000
4	s005	13	9	69.23077
8	s011	3	2	66.66667
10	s013	9	6	66.66667
11	s015	12	8	66.66667
17	s021	6	4	66.66667
38	s043	6	4	66.66667
6	s008	11	7	63.63636
14	s018	11	7	63.63636
24	s029	8	5	62.50000
26	s031	12	7	58.33333
41	s047	12	7	58.33333
1	s002	14	8	57.14286
51	s057	10	5	50.00000

Table 10: Table12:LDA

	subject	Actual	Predicted	Accuracy
7	s010	9	9	100.00000
13	s017	10	10	100.00000
15	s019	15	15	100.00000
17	s021	6	6	100.00000
18	s022	8	8	100.00000
19	s024	9	9	100.00000
20	s025	8	8	100.00000

	subject	Actual	Predicted	Accuracy
22	s027	10	10	100.00000
28	s033	9	9	100.00000
31	s036	8	8	100.00000
32	s037	13	13	100.00000
33	s038	10	10	100.00000
36	s041	10	10	100.00000
37	s042	4	4	100.00000
38	s043	6	6	100.00000
39	s044	14	14	100.00000
46	s052	12	12	100.00000
50	s056	7	7	100.00000
44	s050	14	13	92.85714
49	s055	14	13	92.85714
4	s005	13	12	92.30769
12	s016	13	12	92.30769
25	s030	12	11	91.66667
35	s040	12	11	91.66667
14	s018	11	10	90.90909
23	s028	11	10	90.90909
45	s051	10	9	90.00000
2	s003	9	8	88.88889
10	s013	9	8	88.88889
34	s039	9	8	88.88889
21	s026	15	13	86.66667
27	s032	7	6	85.71429
47	s053	12	10	83.33333
9	s012	11	9	81.81818
3	s004	9	7	77.77778
11	s015	12	9	75.00000
26	s031	12	9	75.00000
48	s054	11	8	72.72727
1	s002	14	10	71.42857
16	s020	7	5	71.42857
5	s007	10	7	70.00000
43	s049	10	7	70.00000
51	s057	10	7	70.00000
8	s011	3	2	66.66667
29	s034	6	4	66.66667
30	s035	9	6	66.66667
41	s047	12	8	66.66667
6	s008	11	7	63.63636
24	s029	8	5	62.50000
40	s046	10	6	60.00000
42	s048	7	3	42.85714

Table 11: Table13:Random Forest

	subject	Actual	Predicted	Accuracy
2	s003	9	9	100.00000
3	s004	9	9	100.00000
7	s010	9	9	100.00000

	subject	Actual	Predicted	Accuracy
8	s011	3	3	100.00000
9	s012	11	11	100.00000
10	s013	9	9	100.00000
11	s015	12	12	100.00000
12	s016	13	13	100.00000
13	s017	10	10	100.00000
15	s019	15	15	100.00000
17	s021	6	6	100.00000
18	s022	8	8	100.00000
19	s024	9	9	100.00000
20	s025	8	8	100.00000
22	s027	10	10	100.00000
23	s028	11	11	100.00000
25	s030	12	12	100.00000
27	s032	7	7	100.00000
28	s033	9	9	100.00000
31	s036	8	8	100.00000
32	s037	13	13	100.00000
33	s038	10	10	100.00000
34	s039	9	9	100.00000
37	s042	4	4	100.00000
38	s043	6	6	100.00000
39	s044	14	14	100.00000
41	s047	12	12	100.00000
43	s049	10	10	100.00000
44	s050	14	14	100.00000
45	s051	10	10	100.00000
46	s052	12	12	100.00000
47	s053	12	12	100.00000
48	s054	11	11	100.00000
50	s056	7	7	100.00000
21	s026	15	14	93.33333
1	s002	14	13	92.85714
49	s055	14	13	92.85714
4	s005	13	12	92.30769
26	s031	12	11	91.66667
35	s040	12	11	91.66667
14	s018	11	10	90.90909
36	s041	10	9	90.00000
40	s046	10	9	90.00000
30	s035	9	8	88.88889
16	s020	7	6	85.71429
42	s048	7	6	85.71429
29	s034	6	5	83.33333
6	s008	11	9	81.81818
24	s029	8	6	75.00000
5	s007	10	7	70.00000
51	s057	10	6	60.00000

Sources

1:Correlations: https://cran.r-project.org/web/packages/ggcorrplot/ggcorrplot.pdf

http://www.sthda.com/english/wiki/ggcorrplot-visualization-of-a-correlation-matrix-using-ggplot2

http://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/

https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guid e.php

2:Multinomial logistic regression

https://en.wikipedia.org/wiki/Multinomial_logistic_regression#Linear_predictor

https://www.analyticsvidhya.com/blog/2016/02/multinomial-ordinal-logistic-regression/

3:creatpartition https://www.rdocumentation.org/packages/caret/versions/6.0-80/topics/createDataPartition

https://arxiv.org/pdf/1204.1177.pdf 4:LDA with PCA

http://little-book-of-r-for-multivariate-analysis.readthedocs.io/en/latest/src/mul tivariateanalysis.html

https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component- analysis-python/
https://towardsdatascience.com/dimensionality-reduction-does-pca-really-improve-class ification-outcome-
6e9ba21f0a32 5:Frequency of subjects https://stackoverflow.com/questions/26553526/how-to-add-frequency-count-labels-to
-the-bars-in-a-bar-graph-using-ggplot2?utm_medium=organic&utm_source=google_rich_q a&utm_campaign=google_rich_
6:box plot http://maths.nayland.school.nz/Year_11/AS1.10_Multivar_data/11_Comparing_Boxplots.htm
http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-dat a-visualization
7:Ridge plot https://cran.r-project.org/web/packages/ggridges/vignettes/introduction.html