Whitepaper STAT-702

Monica, Dheeman, Quinn

05/07/2018

702 Passcode Final Exam

Introduction

The problem we set out to solve was to create a method of predicting who typed in passcodes in a specific system. This system is setup to only allow access to one individual at a time. This specific individual will be asked to re-enter the passcode throughout the session being accessed. These re-entries will be recorded as a different rep value for the individual, but still within the same current session. Conveniently for us, all 51 of the individuals have the same passcode of: .tie5Roan1 . A Known and Unknown dataset were given to train and test our methods for predicting the individuals. These datasets consisted of different recorded values of time for the individual typing in specific keys of the passcode. The two datasets were similar except for Known having the subject that entered the passcode listed, and the Unknown not having the subject listed.

In this report we will discuss how we handled the data that was given, the methods we tried, and the results of the prediction methods. Shortly after the completion of our exploration, a Questioned dataset of 12 unknown sessions will be given to test our best method on. This dataset will be similar to the Unknown, and we will predict the missing individual for each session. From the results of our exploration of this system of passcode predictions we also hope to give our estimation of the error we expect for the Questioned dataset to come.

Data Summary

The known dataset has 1776 entries, where we are trying to predict “subject” against a list of variables. The main task is to predict the accuracy of the 51 individuals who can access using the passcode. The variable session is a block of time where the user has access to the system. The unknown dataset consists of a list of variables where the “subject” are not assigned. This dataset is hard to classify as the labels are missing. But we have made use of this dataset for the variable selection purpose which will be discussed in the later sections.

From the datasets we can see that the columns that consist of the time data are common. Rep and Session were given within both of the 31-time variables that were also given with three main types. The H variables were the recorded time for each key being held. The UD variables were the recorded time from letting up the previous key to pressing down the next key. The DD variables were the recorded time from pressing down the previous key to pressing down the next key.

Normalization Dataset

Before making use of the data, we have processed the dataset. Since, we are not certain about the measuring units of the variables that are used in the dataset, we need to normalize the values. This is done by firstly adding a constant value to each of the value in the known dataset then we have log transferred all the values in the dataset. The log transform is done to normalize the values of each of the data values [1,2,3]. We basically did not plot all the list of variables. But only displayed the UD.period.t variable for the subjects "s002" and "s007". As shown in the figures [3] we can see that after the log transform After normalizing the value, we have included those variables for building the model.

Mutate Variable Creation

Another approach we have used is called mutate of the variables. This is done based on the approach mentioned in the reference in [4]. That project stated that greater impowered can be created by engineering some new features. In the approach they used the standard approach of the keyboard setup, where the hand position of the QWERTY keyboard was used to set up the new set of features. Please have a look at the Appendix [4] to have a good idea of the QWERTY keyboard setup. Based on the QWERTY keyboard and the list of variables we have considered the following list of variables:

- Rep
- Session
- DD and UD variables
- Leftpointer = H.Shift.r + H.t + H.five
- Rightring = H.i + H.o + H.l + H.period
- Lefthand = H.Shirf.r + H.t + H.five + H.a
- Righthand = H.period + H.i + H.o + H.n + H.l + H.return
- Total = sum of all of the time variables except (H.e, DD.e.five, UD.e.five, UD.o.a)

Stepwise selection

Previously, we have talked about the unknown dataset, which is been used for the variable selection purpose. So, we performed hierarchical clustering on the unknown dataset with complete linkage into 73 clusters. This 73 was picked since there were 73 sessions in the unknown so we wanted to cluster each into their own. After assigning a cluster we did a stepwise selection on the dataset. The variables that were found to be best were: rep, H.period, DD.t.i, UD.t.i, UD.i.e, DD.Shift.r.o, H.o, UD.o.a, H.a, DD.n.l, H.l, DD.l.Return, UD.l.Return.

We also did a stepwise selection on the known dataset to see if the suggested model would be different from the clustering stepwise selection on the unknown dataset. To perform stepwise selection in this multiple categorical dataset is not straight forward. We initially, approach in the wrong way by turning the subject names into numeric values from 1 to 51 and performing a similar stepwise selection. Therefore, we have decided to create a loop where the the variables will be compared with one another in "many-to-one" approach, where before creating the model, we will select individual subject and that will be compared

with rest of the other subjects. This approach selected the variables sessionIndex, H.period, UD.period.t, H.t, DD.t.i, UD.t.i, UD.i.e, H.five, DD.Shift.r.o, UD.Shift.r.o, H.o, H.a, H.n, DD.n.l, UD.n.l, DD.l.Return and UD.l.Return based on the Akaike's An Information Criterion (AIC) value of the best model. AIC is a model selection criterion, where the least value of it is stated as the best model.

Methods

Splits

At first, we did hypothesis testing assuming the session.7 == session.8 (we are checking for the variables associated with session.7 is independent to session.8). For that we split the session with 7 and session with 8 and we compared them with subject first, from that we see that some of the subjects in one session is not having in another session. We then selected the subjects that are listed in both sessions and checked if they are equal, but they weren't equal. We performed a Mann-Whitney-Wilcoxon test on each of the 22 selected subjects of both the sessions. The list of diagrams of these 22 subjects are mentioned in [5]. Then the p-value of the time variables of these subjects are adjusted. From, these plots in some of the subjects the variables of the p-values are less than 0.05. Therefore, we reject the null hypothesis of both sessions being not independent. In this revised task we considered the adjusting the p-value using "Holm". This is because Holm is an improved version of Bonferroni (less conservative then higher power and better).

We have tried three different splits for the data including 70:30, 60:40, and 50:50 and fitted models for all of the splits and we have picked the best split based up on the accuracy of the model which has the highest accuracy among all the models. So, for that we have got the model's highest accuracy with 60:40 split which can be seen in table [6].

Variable selection

We have tried four different approaches:

1. One approach is with the original dataset which was provided to us,
2. Second approach is with variable selection, in that we have tried two different approaches for variable selection: We are provided with two datasets "known" and "unknown". So,
    i. The first one is selecting the variables just from the known dataset using stepwise function.
    ii. For the second approach, we have selected the variables from unknown dataset and used the same variables from unknown for known dataset and fitted the models.
3. The last approach that we have tried is mutating, where we have combined the variables based the place where the fingers goes on the keyboard.

Models

The models that we fitted for all of the approaches mentioned above are:

- LDA (Linear Discriminant Analysis)
- Random Forest
- Bagging
- Boosting
- SVM with three different kernels
  - Radial
  - Polynomial
  - Linear
- Model Based Clustering
  - Mclust G = 1
  - Mclust EDDA

We tried QDA too but didn't work for us because of the small training set (because of the more number of variables). We tried KNN but the performance of the model was very poor compared with the other models.

Predictions

Two Prediction Accuracies

For each of the model types and variable subsets we calculated two different accuracy rates. The first rate was an observation to observation rate. We predicted the subject for each observation and checked it against the true subject for that observation. There were just over 700 observations in every testing dataset. So, this first accuracy rate is the number of correctly predicted observations divided by the total number of test observations predicted.

The second rate we gave the name of grouping accuracy. For this accuracy rate the test dataset for each subject were grouped together and then predicted separately. All of the observations of a subject were each predicted. Whichever subject was predicted the most often for that grouping was assigned for that entire group of test subjects. The grouping accuracy was the correctly predicted subject groups divided by total number of 51 subjects. This second method gives an idea of accuracy more similar to how the questioned dataset will be predicted. The first accuracy method gives a better idea of how accurately the models are predicting within the subject groupings.

Results table

Results table [7] shows the observation accuracy rates on our forty percent testing dataset for each of the respective variable subsets. There are nine models and four variable subsets in this results table. The best three overall models in order were Random Forest, Bagging, and Mclust EDDA with average accuracy rates among the four variable subsets of .926, .872, and .857 respectively. Most of the models were in the mid to high eighty percent range. Boosting and Mclust (G = 1) were the two worst performing models both averaging nearly

sixty percent each. The best variable subset was the stepwise selected variables from the known. The subset including the created variables of the mutated H variables did not perform very well. This subset was clearly one of the worst performing and averaged about seven percent worse in accuracy than the best subset. Furthermore, stepwise selection performed much poorly for almost all of the models. One of the reason can be stated that stepwise reduced the number of variables almost into half for the best AIC values. But this reduction did not give any benefit, mainly because a significant number of variables are required for the estimation purpose.

Results table [13] shows the grouping accuracy rates of our predictions on the subject groups within the same forty percent testing dataset. This isn't the most interesting of results since most of the models were able to get 100 %. The following table [14] gives a breakdown of the grouping accuracy with in the SVM polynomial model on the Mutated subset with DD variables. This model and subset somehow got 100% grouping accuracy and only 73.8% observational accuracy. This shows how a group got the correct subject a majority of time while still performing relatively poorly.

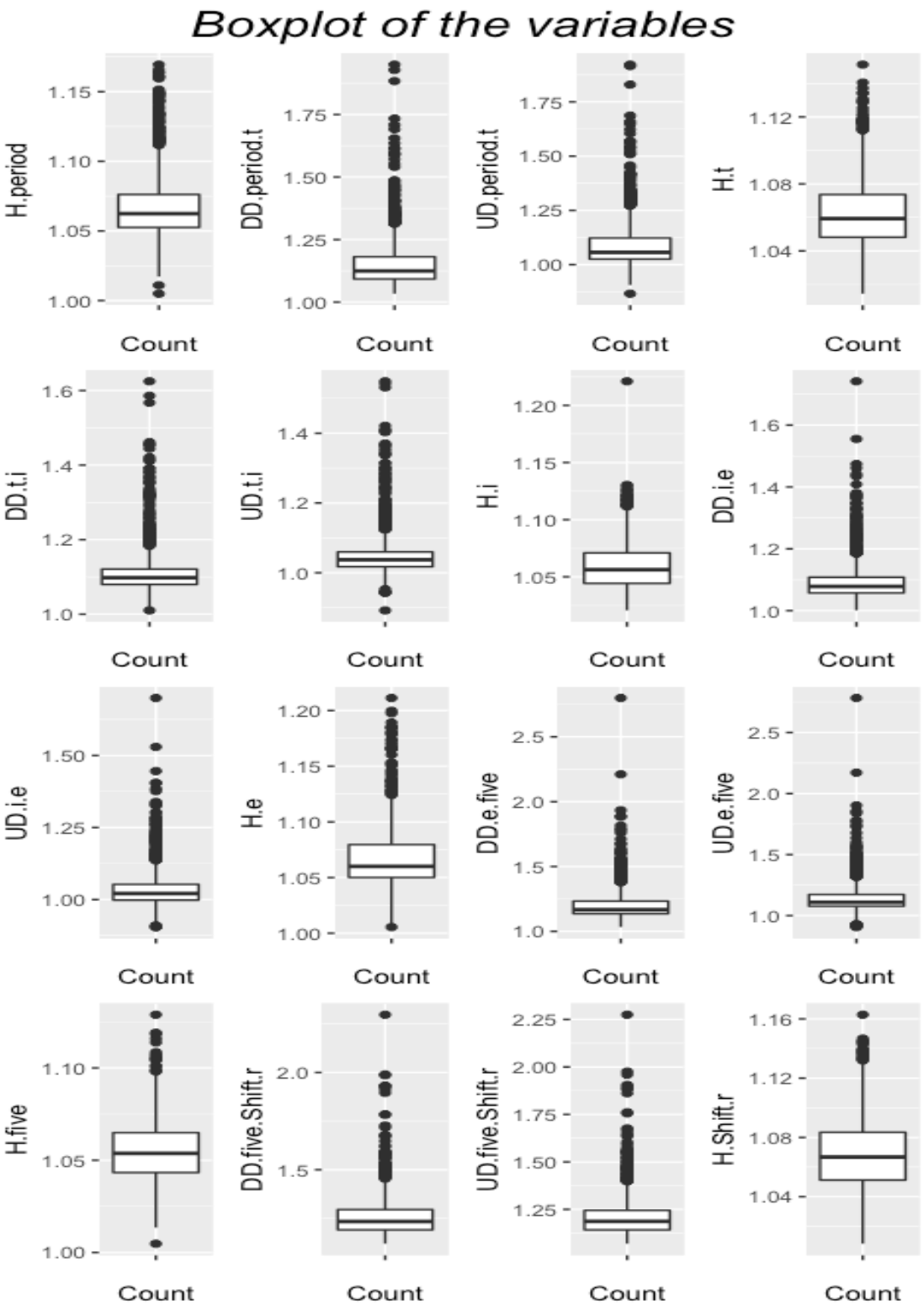
Results table [15] shows the 5-Fold CV that we performed on the known dataset. These results are very similar to the validation set accuracy on the sixty forty percent split from the previous table. The previous best model of Random Forest showed a similar ninety-five percent accuracy in this cross-validation method. This is reassuring that the same per

Prediction Accuracy expected for Questioned

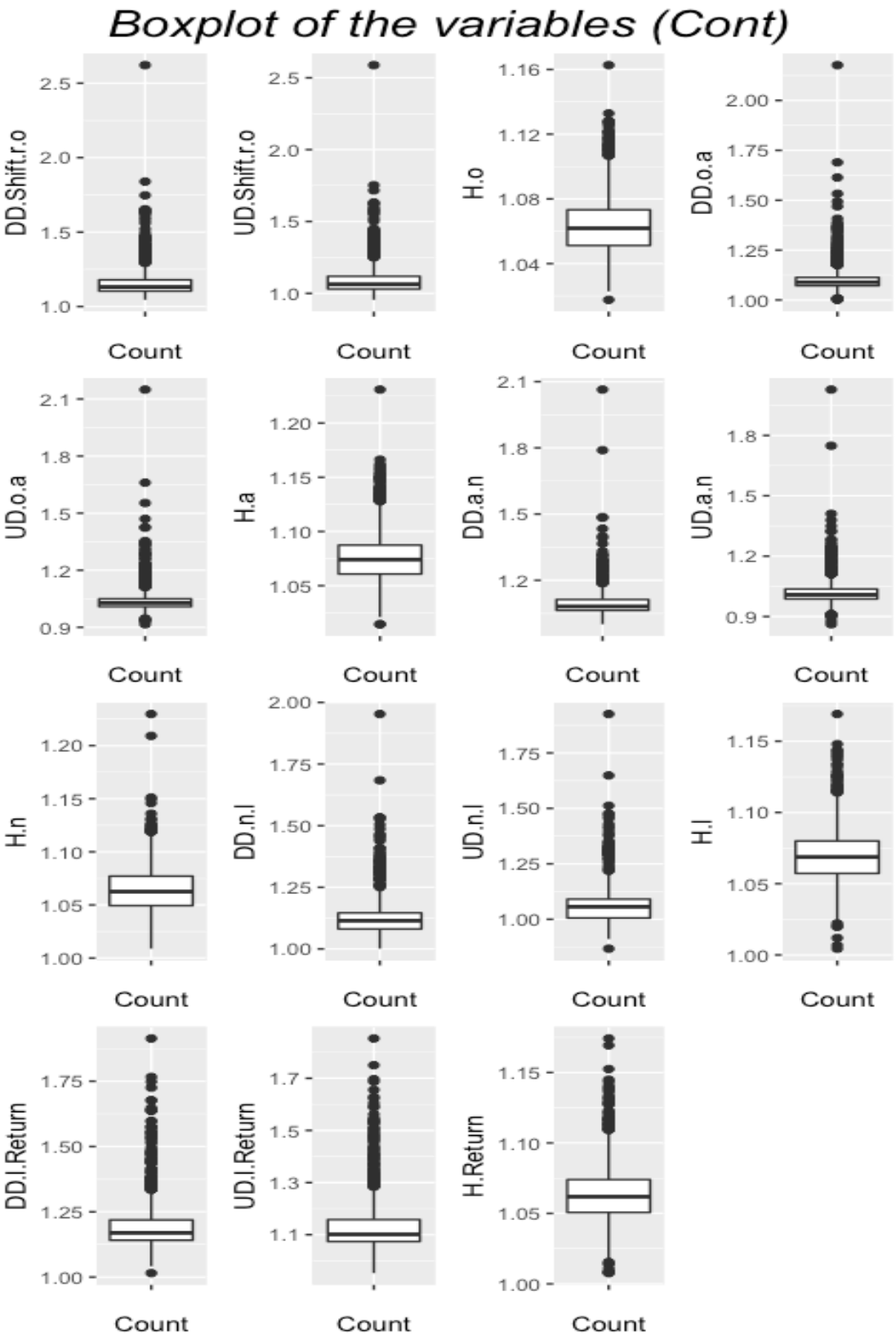
In conclusion, the best model is the Random Forest for the Original transformed dataset without DD variables in table [12]. We predict this Random Forest model to have an accuracy of the observational method to be 0.9479. Since the Questioned data will be just 12 sessions and the method of assigning subjects is a little more forgiving, we expect that accuracy might be even higher. This is since the method involves selecting the most commonly predicted subject from the observational predictions.

Appendix of Images/Plots

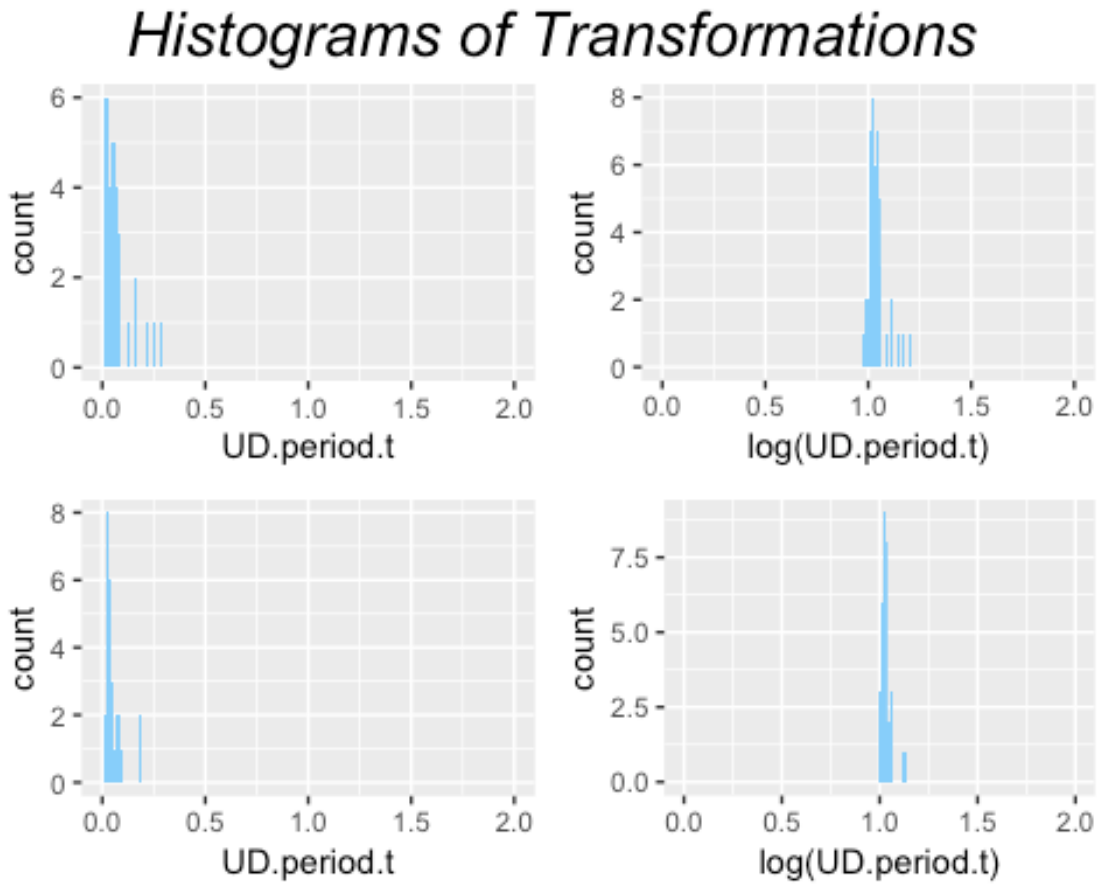
[1] Boxplots



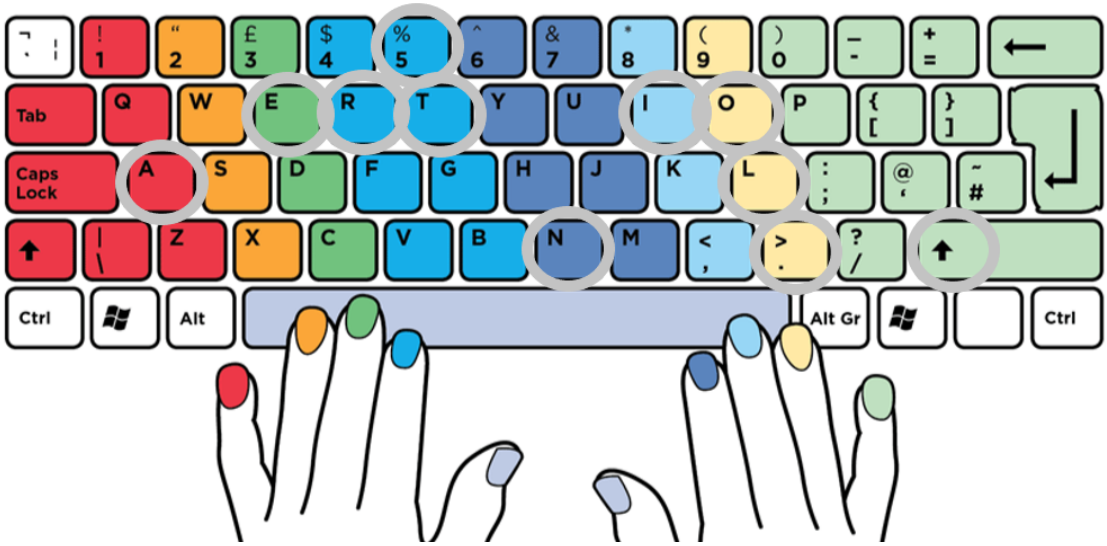
[2] Boxplots 2


Boxplot of the variables (Cont)
DD.Shift.r.o
UD.Shift.r.o
H.o
DD.o.a
UD.o.a
H.a
DD.a.n
UD.a.n
H.n
DD.n.l
UD.n.l
H.l
DD.l.Return
UD.l.Return
H.Return
Count

[3] Histograms

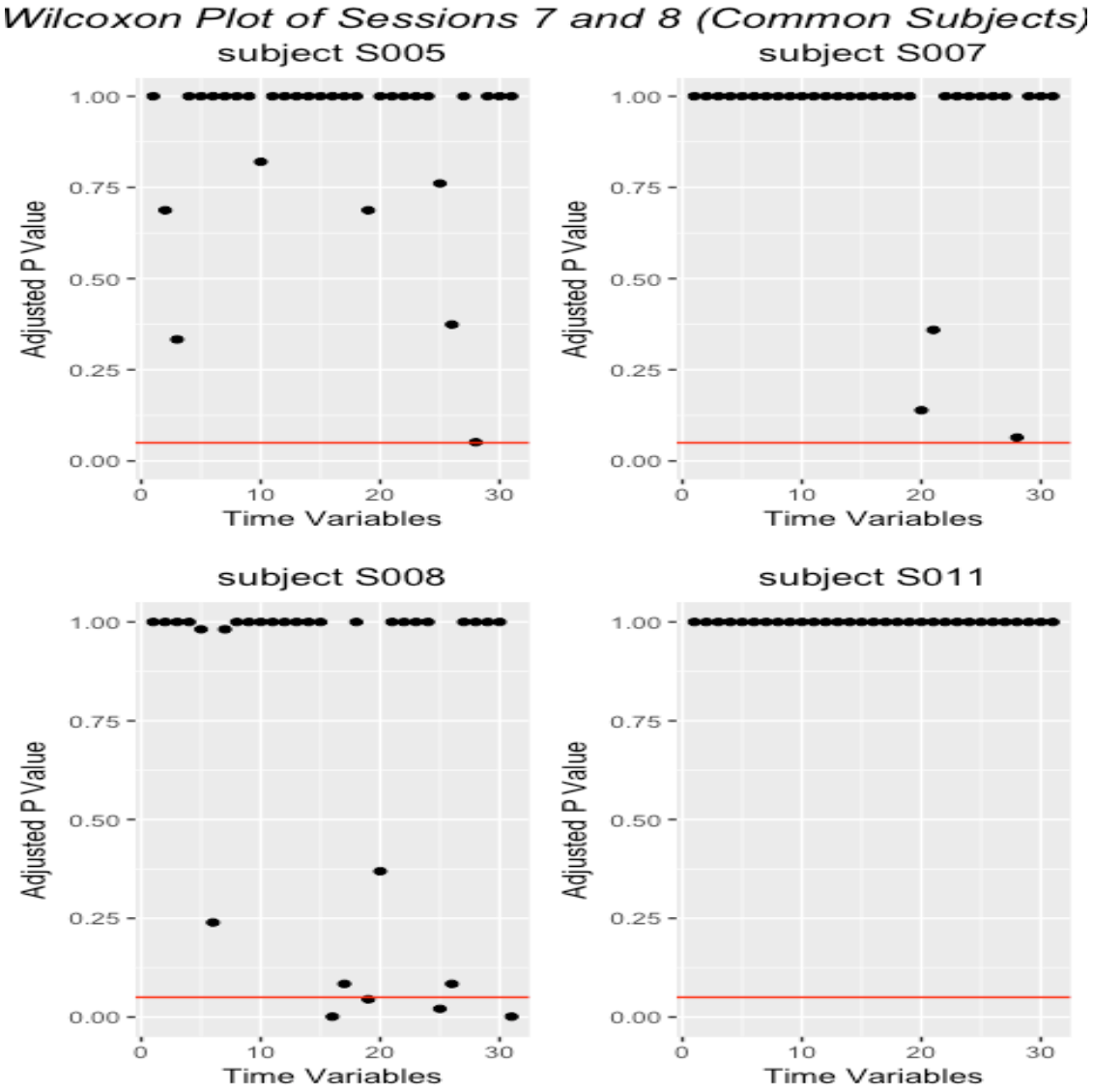


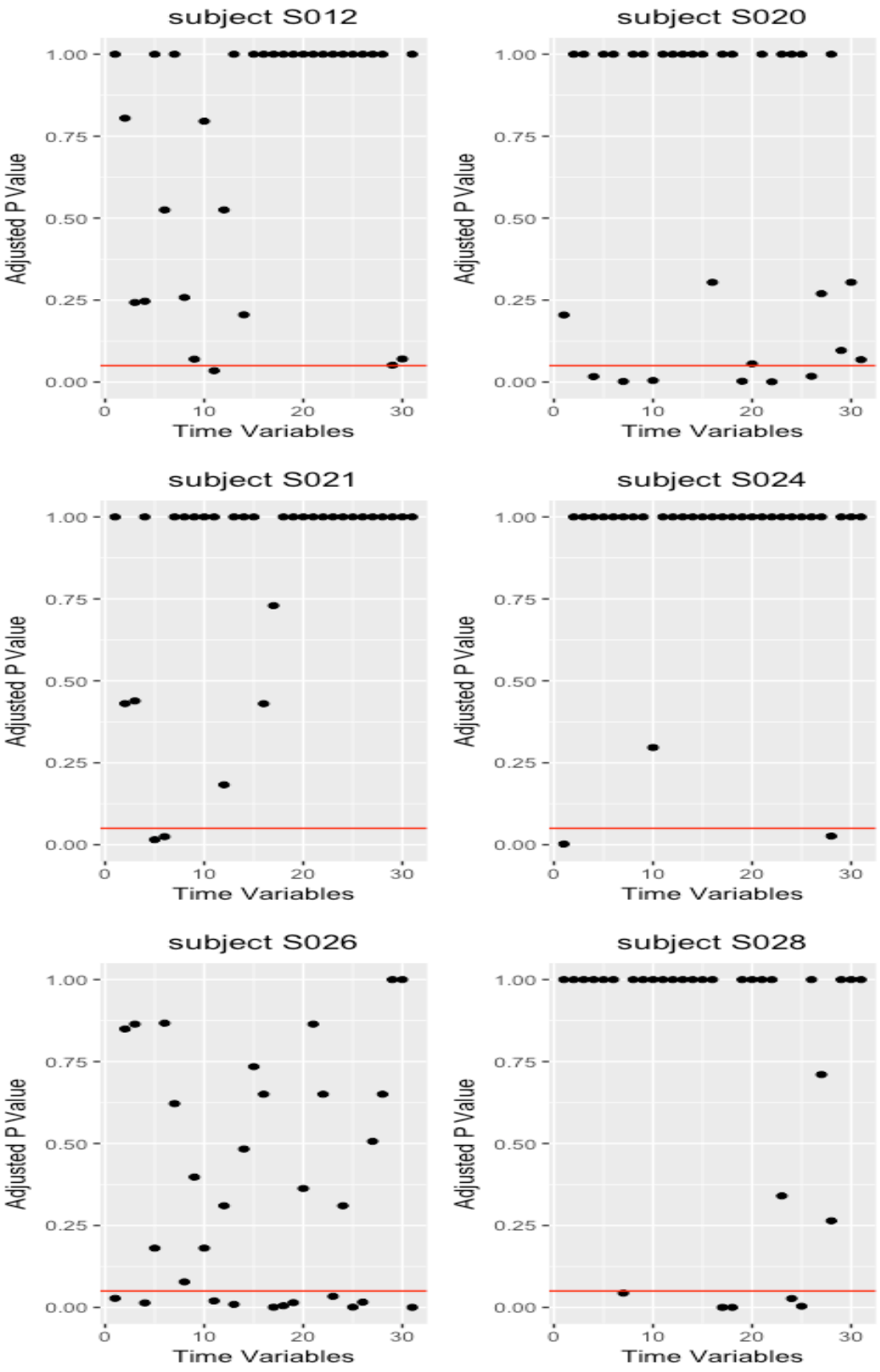
[4] Keyboard Mutated

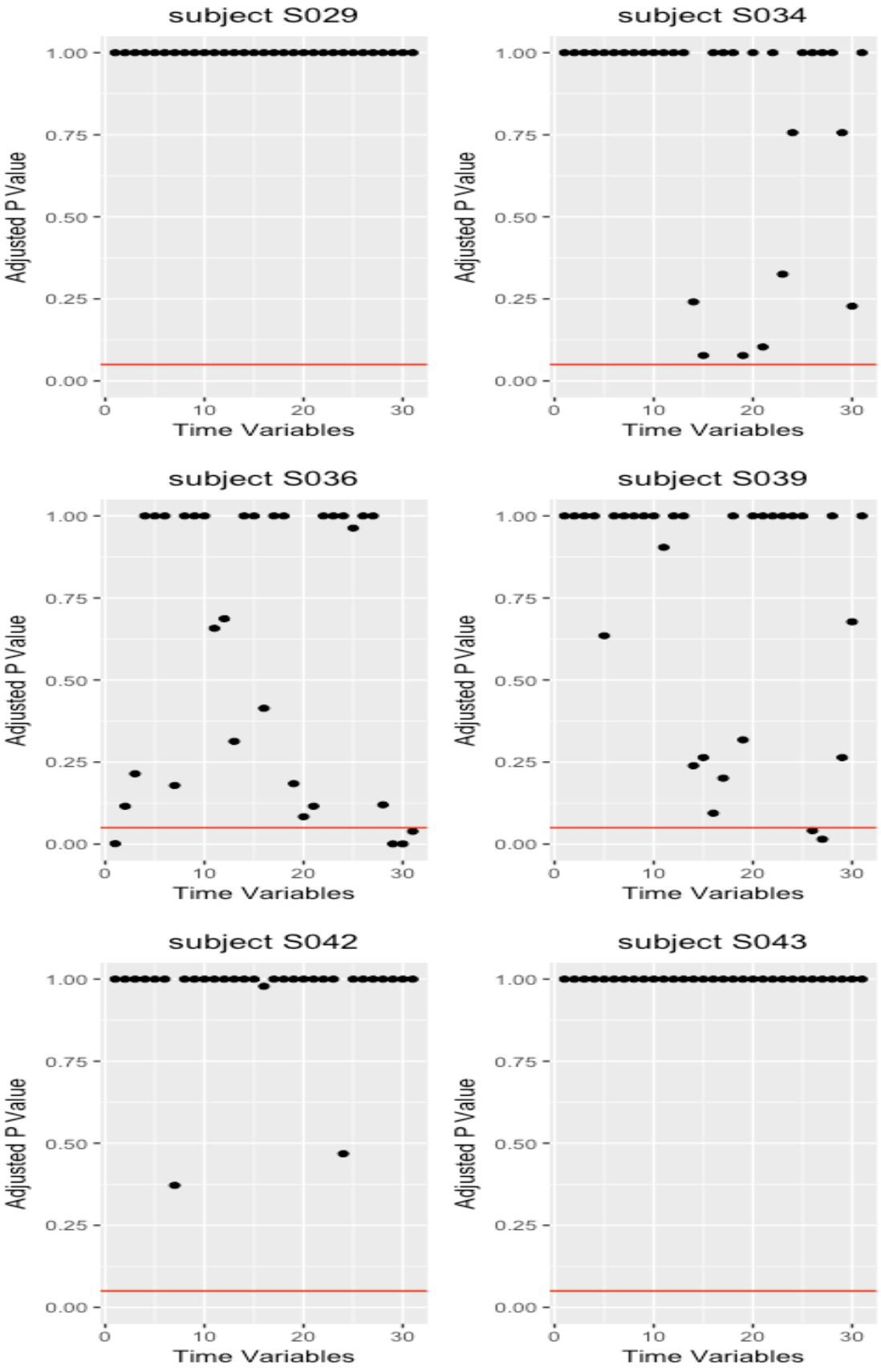


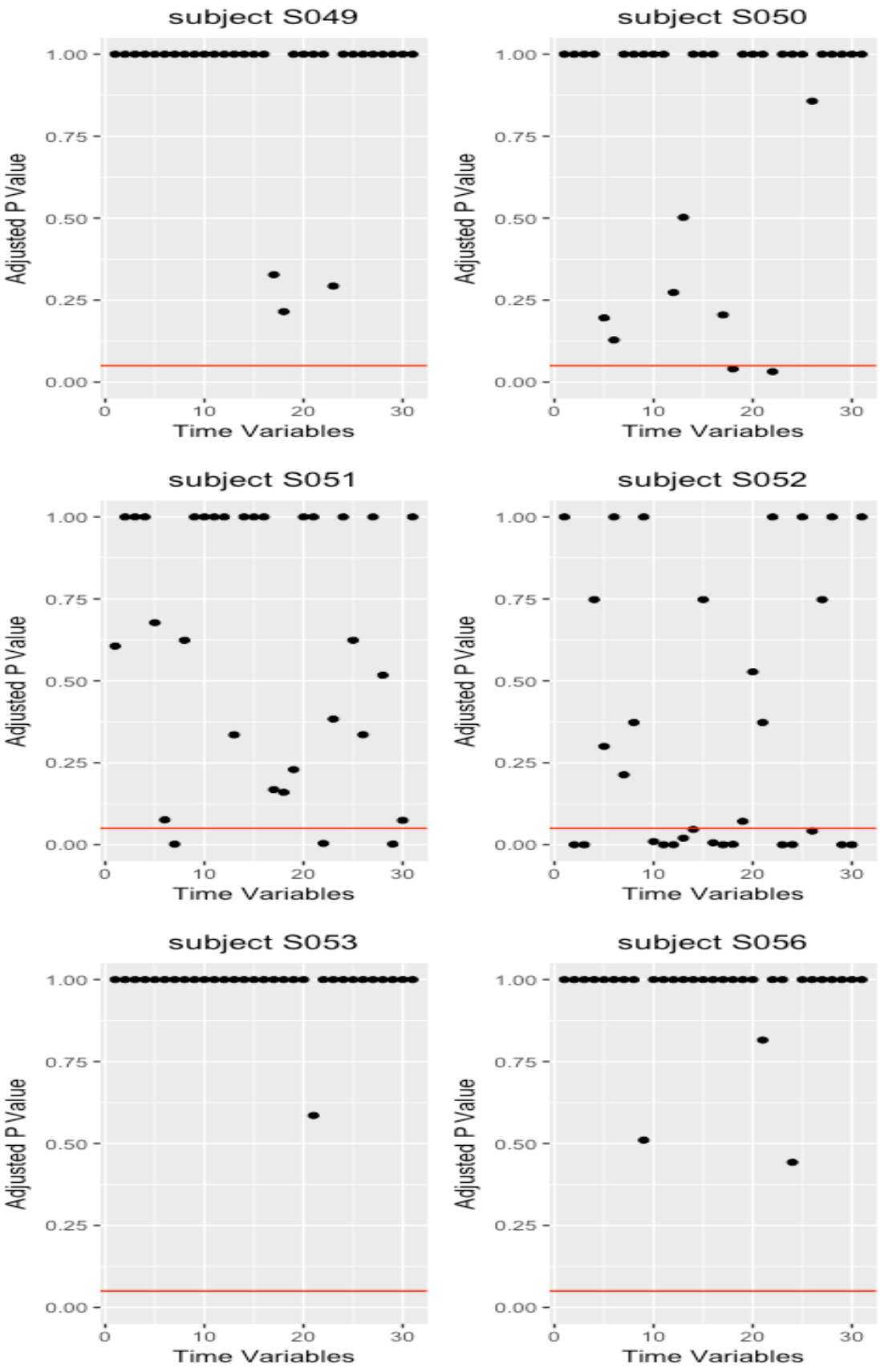
Ref: https://github.com/cjcohan/Springboard/tree/master/Capstone

[5] Hypothesis Session




subject S012
subject S020
subject S021
subject S024
subject S026
subject S028
Adjusted P Value
Time Variables
1.00
0.75
0.50
0.25
0.00
0
10
20
30


subject S029
subject S034
subject S036
subject S039
subject S042
subject S043
Adjusted P Value
Time Variables
1.00
0.75
0.50
0.25
0.00
0
10
20
30


subject S049
subject S050
subject S051
subject S052
subject S053
subject S056
Adjusted P Value
Time Variables
1.00
0.75
0.50
0.25
0.00
0
10
20
30

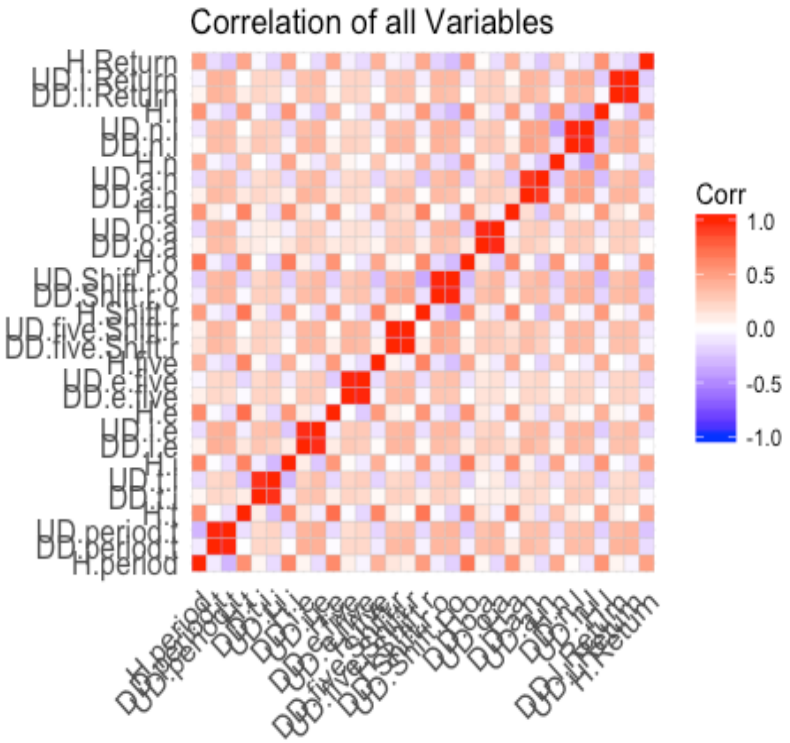
[6] Splits

	7 & 8	50 / 50	60 / 40	70 / 30
Observation	0.8223	0.9369	0.9522	0.9681
Grouping	0.9091	1.0000	1.0000	0.9804

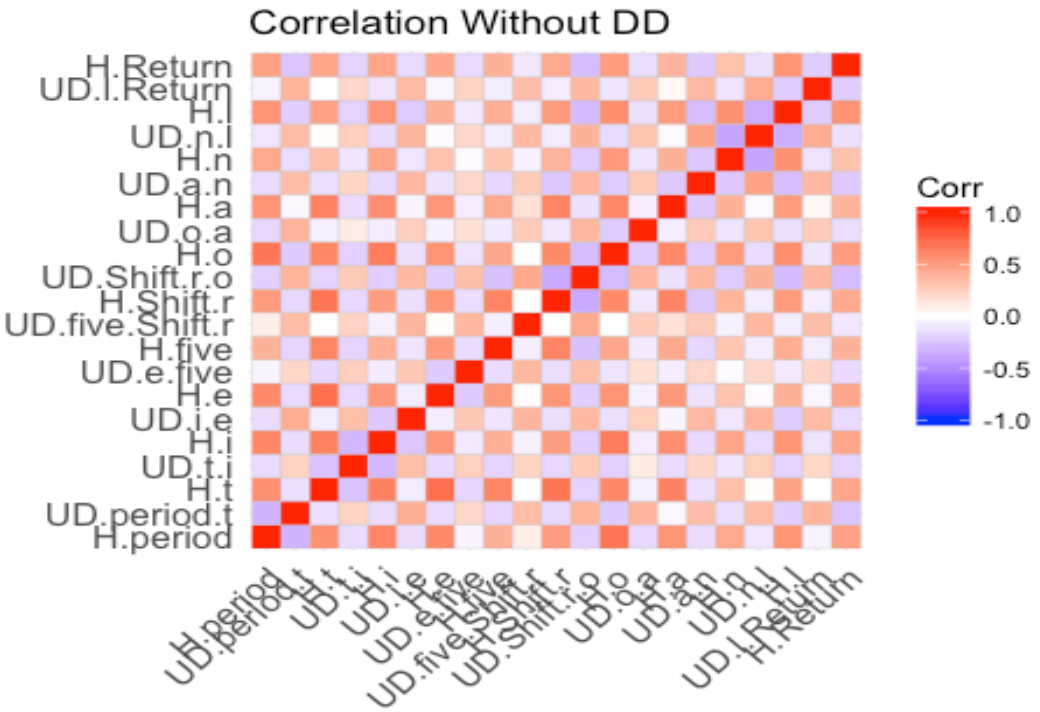
[7] Both UD and DD

	Original	Mutated	Known Stepwise	Unknown Stepwise	PCA	Mean of Model
LDA	0.873	0.864	0.806	0.708	0.873	0.837
Random Forest	0.952	0.906	0.878	0.830	0.851	0.895
Bagging	0.896	0.851	0.810	0.790	0.783	0.841
SVM Radial	0.821	0.782	0.817	0.731	0.727	0.776
SVM Poly	0.834	0.738	0.783	0.617	0.802	0.759
SVM Linear	0.875	0.802	0.802	0.698	0.852	0.815
Mclust EDDA	0.873	0.855	0.473	0.696	0.751	0.809
Average	0.875	0.828	0.806	0.724	0.806	0.819

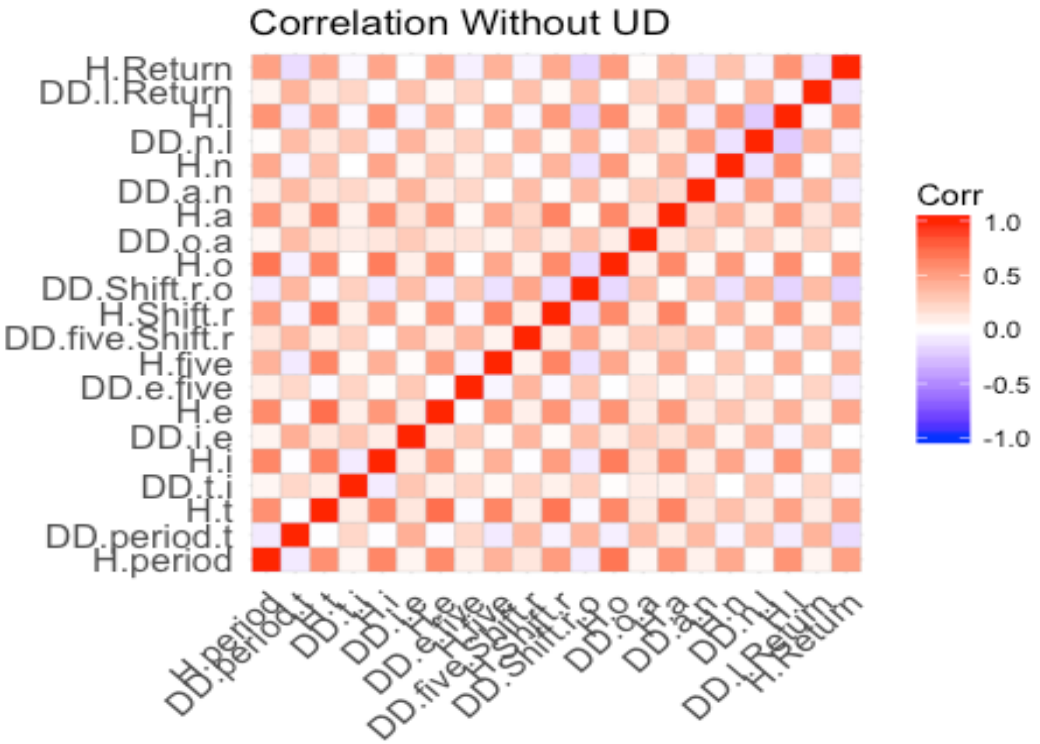
[8] Correlation All Variables



[9] Correlation Without DD



[10] Correlation Without UD



[11] UD vs DD

	DD	UD
LDA	0.467	0.518
Random Forest	0.789	0.798
Bagging	0.743	0.755
SVM Radial	0.665	0.688
SVM Poly	0.615	0.627
SVM Linear	0.622	0.674
Mclust EDDA	0.461	0.458
Average	0.579	0.609

[12] With only UD

	Original	Mutated	Known Stepwise	Unknown Stepwise	PCA	Mean of Model
LDA	0.873	0.760	0.840	0.605	0.873	0.790
Random Forest	0.948	0.892	0.932	0.782	0.851	0.881
Bagging	0.883	0.835	0.866	0.752	0.830	0.834
SVM Radial	0.859	0.788	0.842	0.695	0.783	0.794
SVM Poly	0.842	0.743	0.810	0.588	0.796	0.756
SVM Linear	0.878	0.798	0.854	0.657	0.855	0.808
Mclust EDDA	0.824	0.736	0.769	0.572	0.743	0.729
Average	0.873	0.793	0.845	0.664	0.819	0.841

[13] Grouping

	Original	Mutated	Known Stepwise	Unknown Stepwise	PCA
LDA	1.00	0.980	1.00	0.882	1.00
Random Forest	1.00	1.000	1.00	0.922	1.00
Bagging	1.00	0.961	1.00	0.902	1.00
SVM Radial	1.00	1.000	1.00	0.941	1.00
SVM Poly	1.00	0.961	0.98	0.824	1.00
SVM Linear	1.00	1.000	1.00	0.922	1.00
Mclust EDDA	0.98	0.961	1.00	0.804	0.98

[14] Grouping Example

	Predicted	Actual	Error Rate	Accuracy Rate
s002	11	16	0.3100	0.6900
s003	9	12	0.2500	0.7500
s004	6	12	0.5000	0.5000
s005	18	22	0.1800	0.8200
s007	10	15	0.3300	0.6700
s053	13	15	0.1300	0.8700
s054	10	10	0.0000	1.0000
s055	14	15	0.0700	0.9300
s056	8	13	0.3800	0.6200
s057	10	17	0.4100	0.5900
Totals	525	711	0.2616	0.7384

[15] 5-Fold CV

	VSA	5-Fold CV
LDA	0.873	0.887
Random Forest	0.948	0.960
Bagging	0.883	0.902
SVM Radial	0.859	0.880
SVM Poly	0.842	0.888
SVM Linear	0.878	0.890
Mclust EDDA	0.824	0.842