ICDSS

BEST DATA : Wei Quan, Sophie Lai, Shaun Tan

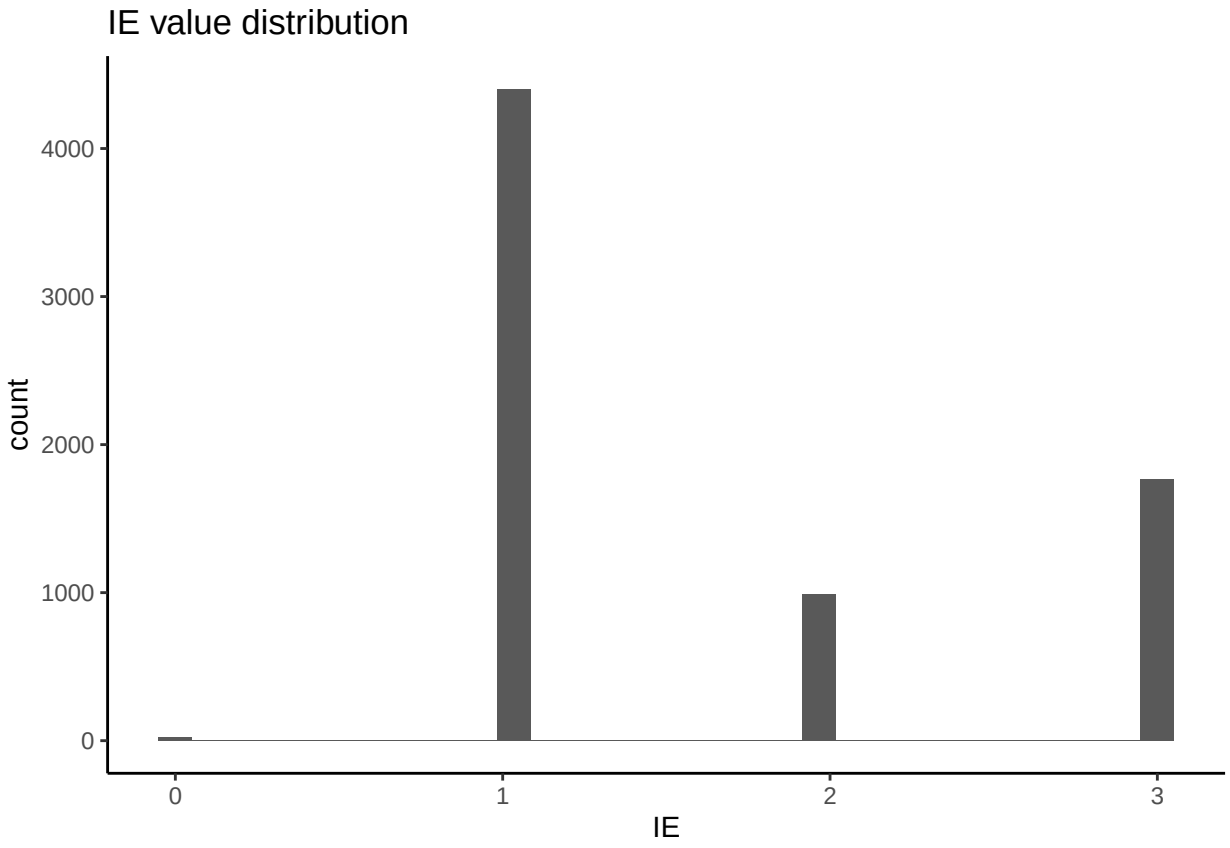
12 December, 2021

Data processing

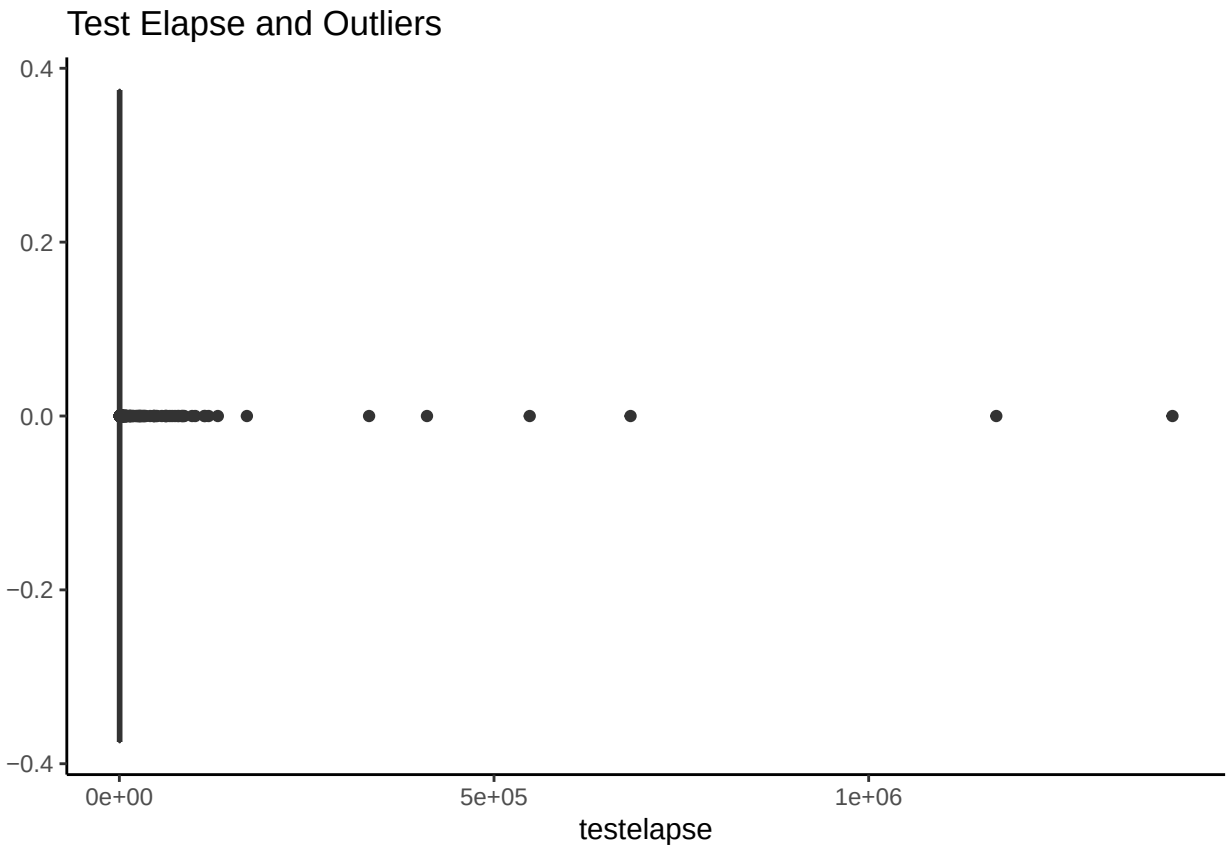
1. filter out IE = 0 (missing value)
2. filter out testelapse value in the top or bottom 1% (remove outliers in people spend too little or too much time on the test)

```
ggplot(data=data,aes(IE)) + geom_histogram() +theme_classic() + ggtitle('IE value distribution')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data=data,aes(testelapse)) + geom_boxplot() +theme_classic() + ggtitle('Test Elapse and Outliers
```




```
# filter with testlapse in 1% and 99% quantile
data_filter <- data %>%
  filter(IE!=0 & testelapse>= quantile(testelapse,0.01) & testelapse <= quantile(testelapse,0.99) )

data_final <- data_filter %>%
  select(ends_with("A"),IE,age,gender)
```

Specify variables of interest

1.independent variables: answers for 91 questions (categorical,1,2,3,4,5)

2.dependent variables: introvert/extrovert (categorical, 1= yes, 0 = no/not sure)

```
data_final <- data_final%>%
  mutate(IE_extrovert=ifelse(IE==2,1,0),
         IE_introvert=ifelse(IE==1,1,0))

head(data_final[,1:90])
```

```
##   Q1A Q2A Q3A Q4A Q5A Q6A Q7A Q8A Q9A Q10A Q11A Q12A Q13A Q14A Q15A Q16A Q17A
## 1   5   3   1   2   3   2   3   3   4    5    1    5    3    5    1    3    3
## 2   5   5   1   5   2   2   5   2   1    3    2    2    3    3    2    2    4
## 3   3   4   5   3   4   5   5   5   5    5    4    5    4    1    1    1    1
## 4   5   2   1   1   5   5   5   4   4    2    5    5    2    1    1    1    2
## 5   1   2   1   1   3   3   5   1   3    4    1    4    5    2    1    5    3
## 6   2   5   5   1   2   4   5   2   4    4    1    4    2    2    1    2    2
##   Q18A Q19A Q20A Q21A Q22A Q23A Q24A Q25A Q26A Q27A Q28A Q29A Q30A Q31A Q32A
## 1    4    1    4    1    3    5    1    4    5    5    4    1    2    2    2
```

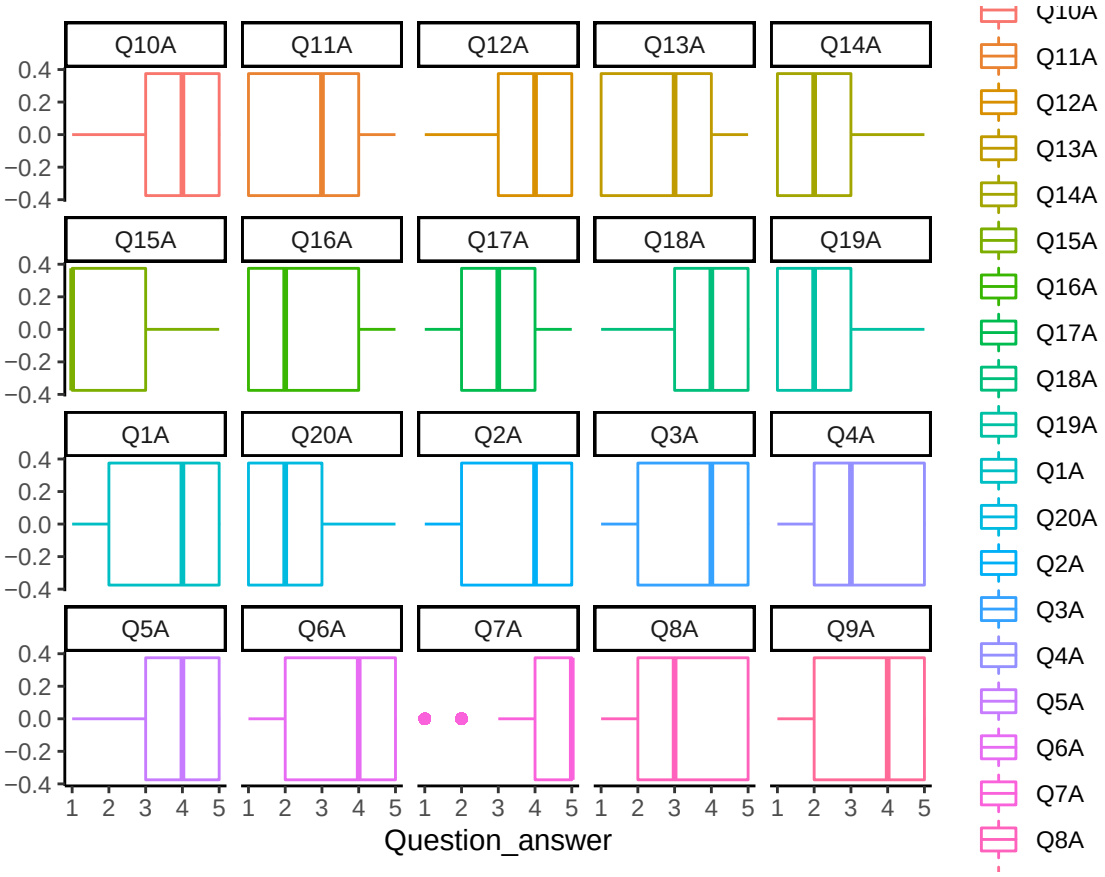
```
## 2    5    2    2    4    4    3    5    5    5    1    5    2    2    3    5
## 3    1    2    1    3    2    5    4    4    4    5    2    4    5    4    5
## 4    4    1    1    1    4    2    2    3    4    3    2    5    5    2    5
## 5    5    5    5    5    4    5    5    5    5    4    2    1    1    4    1
## 6    2    1    1    4    4    2    1    5    3    2    1    5    1    2    1
##   Q33A Q34A Q35A Q36A Q37A Q38A Q39A Q40A Q41A Q42A Q43A Q44A Q45A Q46A Q47A
## 1    4    5    4    2    1    1    5    2    5    4    5    5    1    5    2
## 2    2    2    4    5    4    4    3    3    1    3    4    1    1    4    2
## 3    2    2    3    5    1    4    2    4    4    4    5    5    5    4    3
## 4    4    2    2    3    2    5    1    5    4    5    3    5    2    4    1
## 5    4    4    4    5    2    5    4    3    2    4    5    3    1    5    1
## 6    5    4    4    1    1    4    1    5    5    2    4    3    4    2    1
##   Q48A Q49A Q50A Q51A Q52A Q53A Q54A Q55A Q56A Q57A Q58A Q59A Q60A Q61A Q62A
## 1    4    5    3    1    1    5    3    1    4    4    5    2    4    1    2
## 2    2    4    2    2    1    1    1    3    1    2    1    3    1    1    4
## 3    1    1    5    5    5    4    3    5    4    5    4    5    4    5    3
## 4    4    1    4    5    1    1    5    1    4    4    2    5    5    4    1
## 5    4    3    1    2    1    2    3    1    1    5    1    2    2    5    5
## 6    1    1    4    1    1    1    2    5    5    1    2    5    1    4    2
##   Q63A Q64A Q65A Q66A Q67A Q68A Q69A Q70A Q71A Q72A Q73A Q74A Q75A Q76A Q77A
## 1    2    1    3    1    4    1    1    5    1    5    3    2    5    5    3
## 2    5    1    5    1    1    1    1    5    1    5    3    2    5    2    4
## 3    2    5    2    5    1    4    1    5    1    1    1    3    4    1    1
## 4    1    2    4    2    1    1    5    5    4    1    3    2    4    4    1
## 5    4    1    2    2    5    4    2    5    4    3    1    5    5    5    5
## 6    1    5    5    1    5    4    1    1    1    4    2    1    5    1    4
##   Q78A Q79A Q80A Q81A Q82A Q83A Q84A Q85A Q86A Q87A Q88A Q89A Q90A
## 1    5    3    4    2    1    3    2    1    4    2    5    4    3
## 2    2    5    4    2    1    2    2    2    1    3    4    4    4
## 3    3    4    1    5    5    5    5    5    4    5    3    2    1
## 4    4    4    1    5    5    5    5    5    3    5    4    4    3
## 5    4    4    5    3    2    3    1    1    3    1    2    5    5
## 6    2    2    1    5    2    2    3    1    3    1    2    2    2
```

Data visualisation

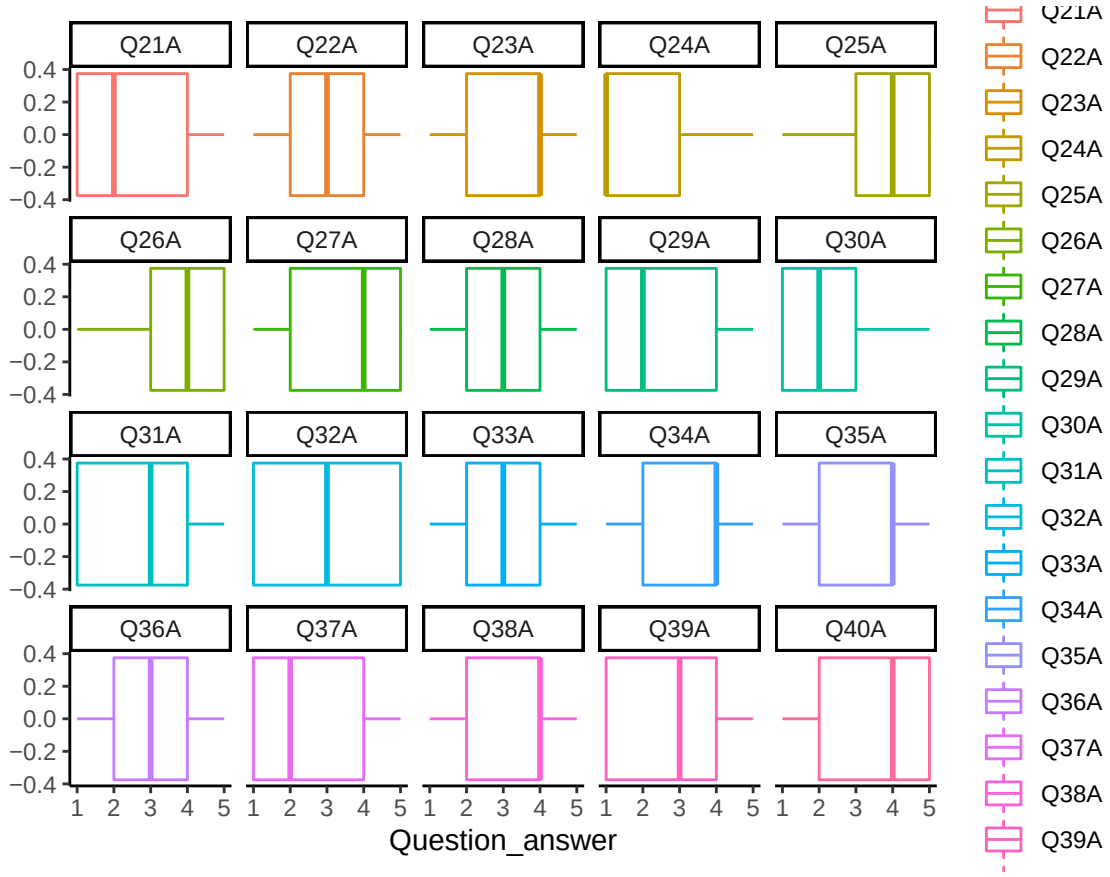
1.Boxplot for values of each question

2.Heatmap for correlation matrix

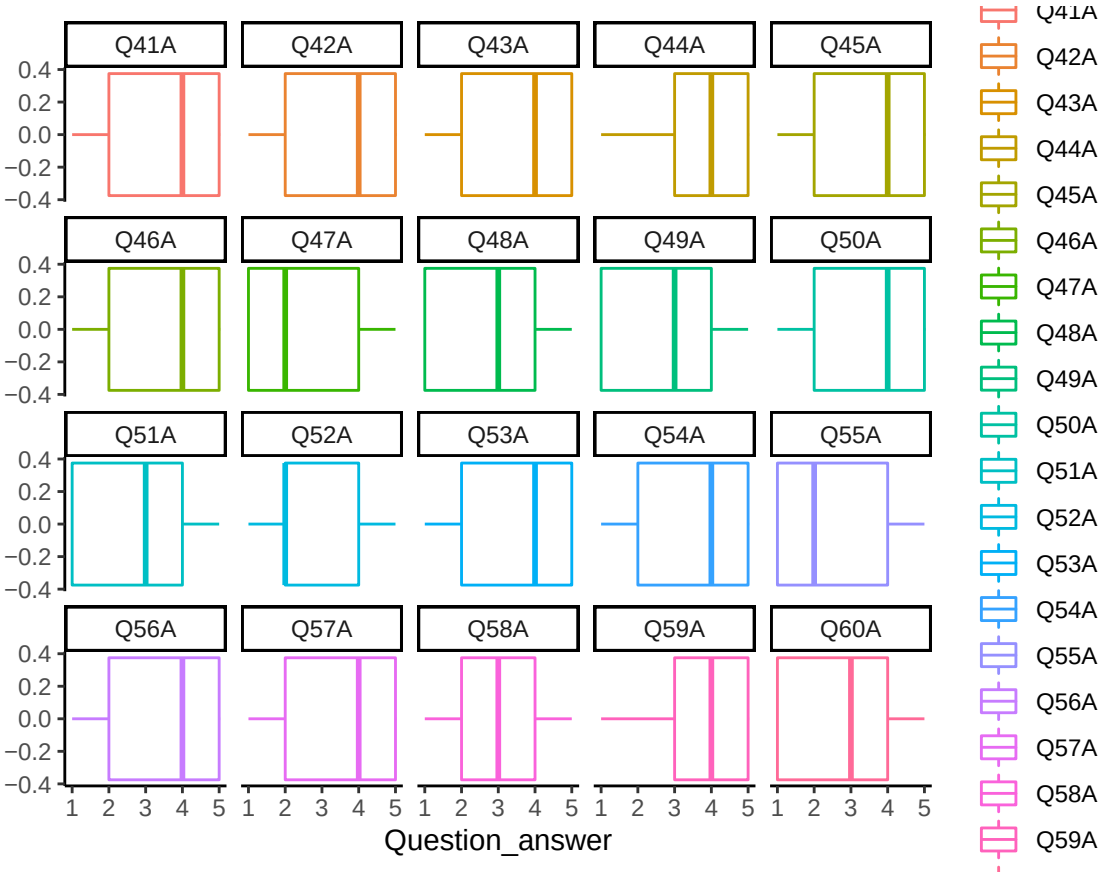
```
# boxplot for each question
data_final %>% gather(Question_number,Question_answer,starts_with('Q')[1:20]) %>%
  ggplot(aes(Question_answer,col=Question_number)) + geom_boxplot() +
  facet_wrap(.~Question_number) + theme_classic()
```



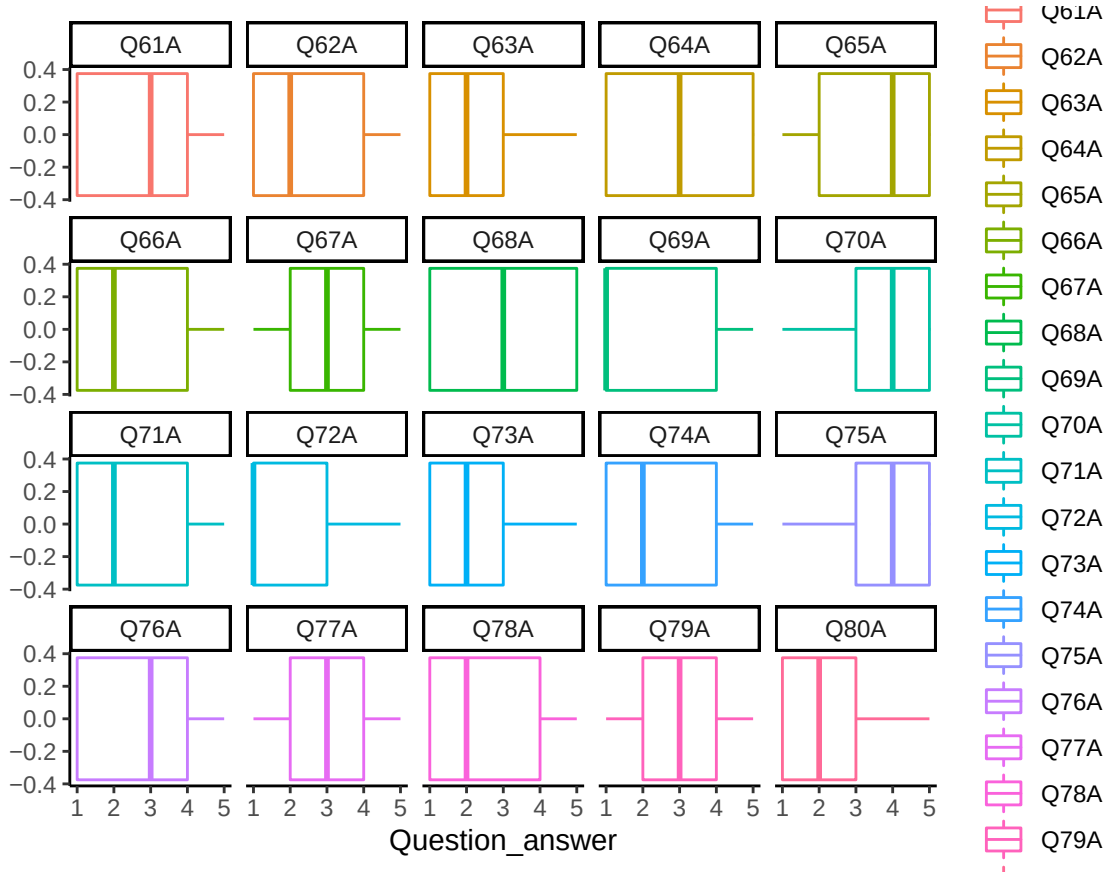

```
data_final %>% gather(Question_number,Question_answer,starts_with('Q')[21:40]) %>%
  ggplot(aes(Question_answer,col=Question_number)) + geom_boxplot() +
  facet_wrap(.~Question_number) + theme_classic()
```



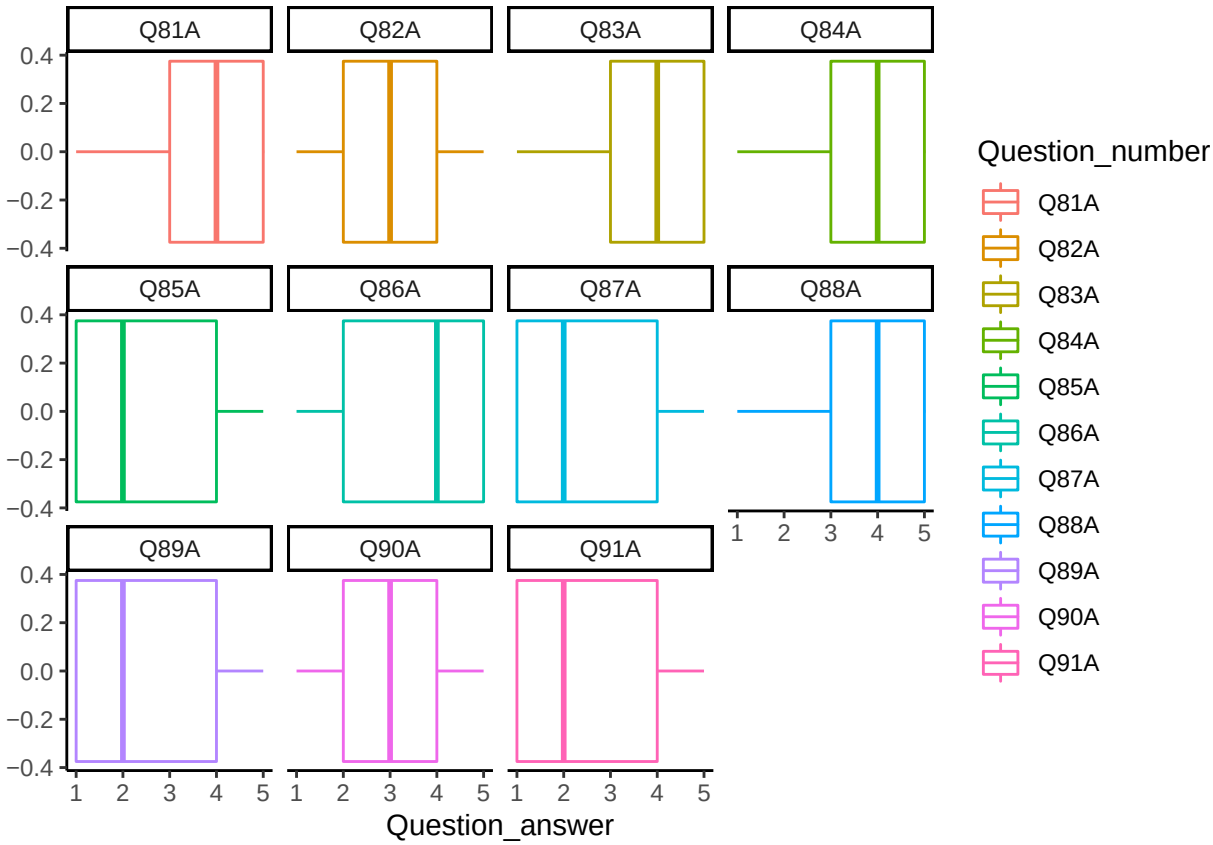

```
data_final %>% gather(Question_number,Question_answer,starts_with('Q')[41:60]) %>%
  ggplot(aes(Question_answer,col=Question_number)) + geom_boxplot() +
  facet_wrap(.~Question_number) + theme_classic()
```



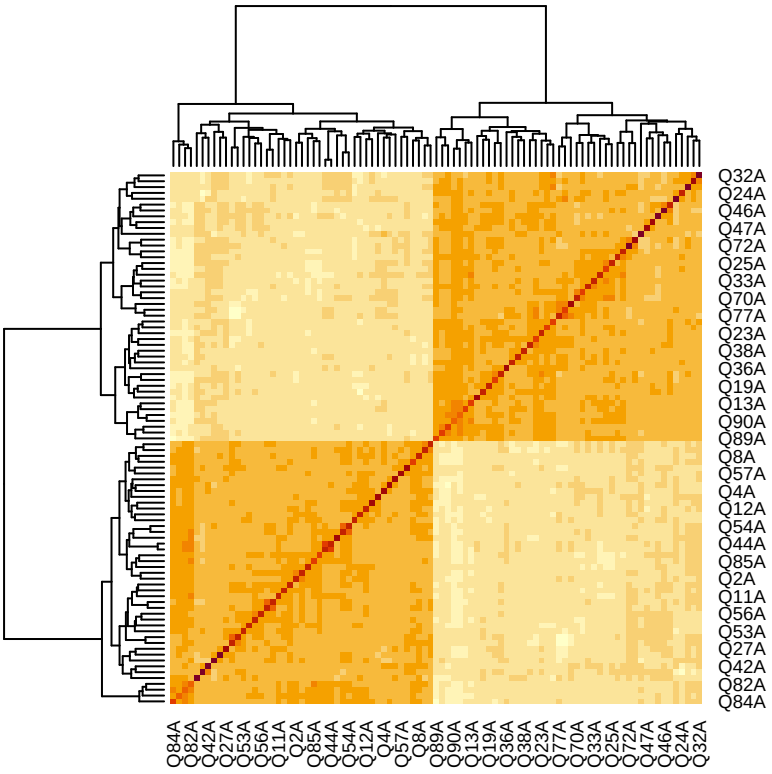

```
data_final %>% gather(Question_number,Question_answer,starts_with('Q')[61:80]) %>%
  ggplot(aes(Question_answer,col=Question_number)) + geom_boxplot() +
  facet_wrap(.~Question_number) + theme_classic()
```




```
data_final %>% gather(Question_number,Question_answer,starts_with('Q')[81:91]) %>%
  ggplot(aes(Question_answer,col=Question_number)) + geom_boxplot() +
  facet_wrap(.~Question_number) + theme_classic()
```




```
# heatmap
cormat <- data_final %>%
  select(starts_with('Q')) %>%
  cor()

heatmap(cormat)
```



Analysis plan

1. Split data to testing data and training data (30%:70%)

```
train <- data_final[1:5000,]
test <- data_final[5000:nrow(data_final),]
```

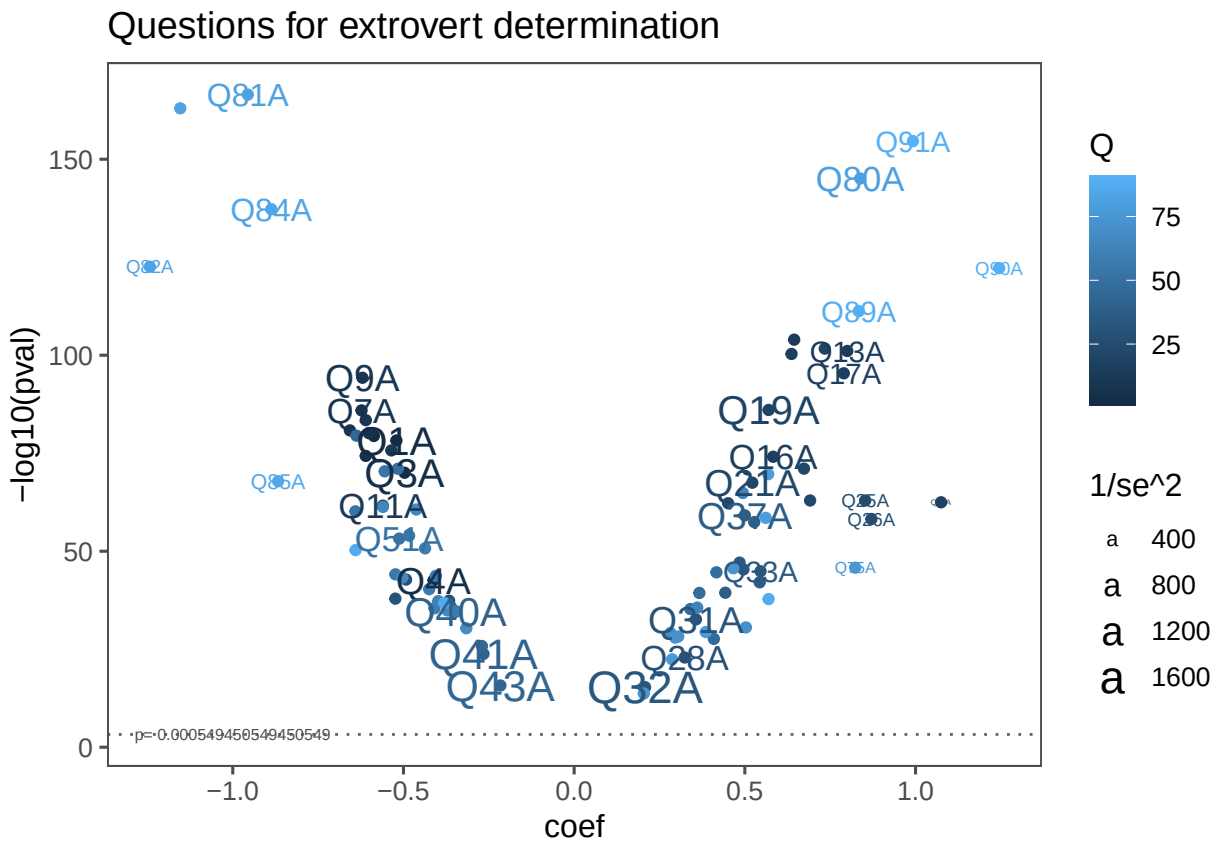
2. Select dominant questions using idea of GWAS(multiple logistic regression) for extrovert and introvert with threshold p=0.05/91 to adjust for multi comparison

```
# multiple logistic regression for extrovert question selection
logistic_ex_reg <- function(x){
  summary(glm(data=train,IE_extrovert~x,family='binomial'))$coefficient[2,]
}

coef <- apply(train[,1:91],2,logistic_ex_reg)[1,]
se <- apply(train[,1:91],2,logistic_ex_reg)[2,]
pval <- apply(train[,1:91],2,logistic_ex_reg)[4,]
results <- data.frame(cbind(coef,se,pval))
results$Q <- 1:91

# multiple logistic regression for introvert question selection
logistic_in_reg <- function(x){
  summary(glm(data=train,IE_introvert~x,family='binomial'))$coefficient[2,]
}

coef_in <- apply(train[,1:91],2,logistic_in_reg)[1,]
se_in <- apply(train[,1:91],2,logistic_in_reg)[2,]
pval_in <- apply(train[,1:91],2,logistic_in_reg)[4,]
results_in <- data.frame(cbind(coef_in,se_in,pval_in))
```

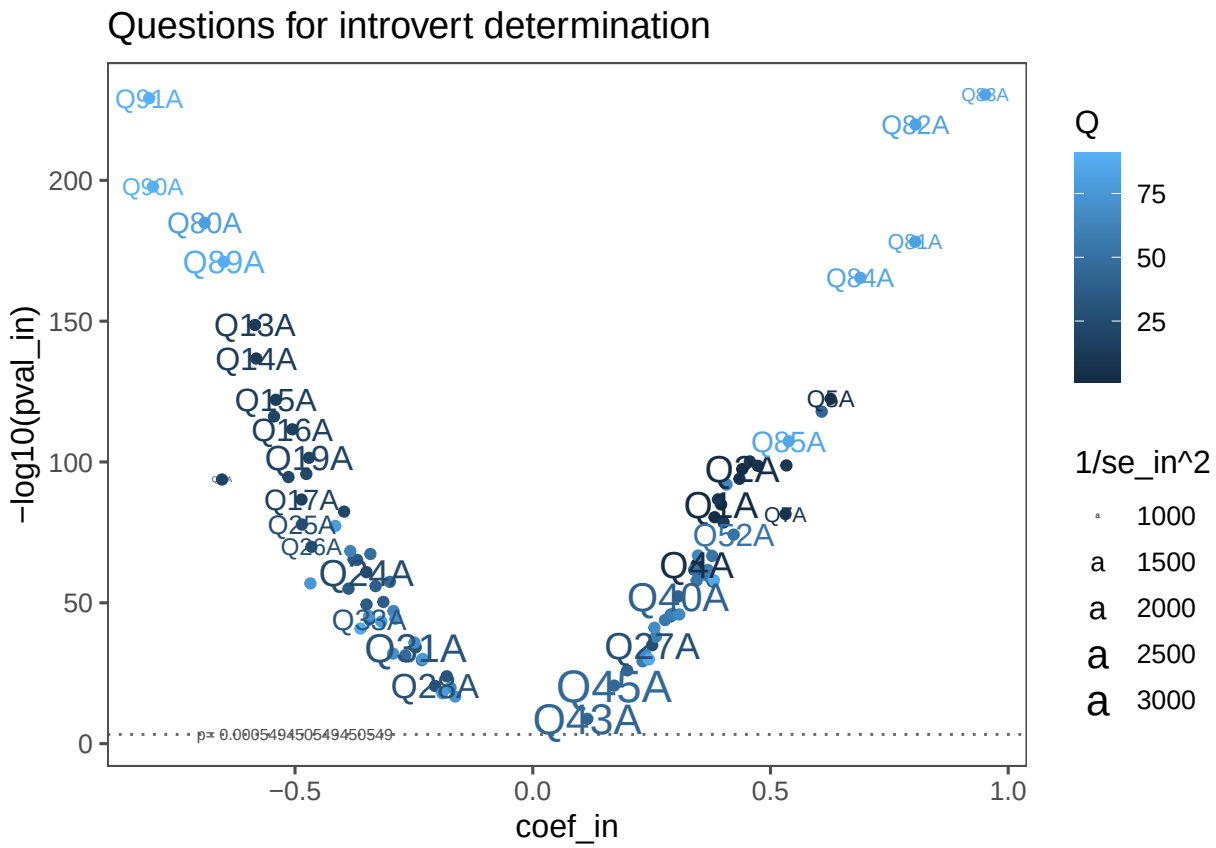
```
results_in$Q <- 1:91
```

3. Volcano plot with x axis=coef and y axis = -log10(pval)

```
# Volcano plot ------------------------------------------------
#x:log(coef)
#y:-log(SE)
#test size:(1/se^2)

pval_bonf = 0.05/nrow(results)

library(ggthemes)

p <- results %>%
  ggplot(aes(x = coef,y = -log10(pval),col=Q,label=rownames(results))) +
  geom_point(size=1.5) +
  geom_hline(yintercept = -log10(pval_bonf),
             col = 'grey40',
             linetype = 'dotted') +
  annotate('text',x=-1.0, y= -log10(pval_bonf + 0.000001),
             label=paste('p=' ,pval_bonf),
             size=2,col='grey30') +
  theme_few()+
  ggtitle('Questions for extrovert determination')

p + geom_text(check_overlap = TRUE,aes(size= 1/se^2))
```



```
p2 <- results_in %>%
  ggplot(aes(x = coef_in,y = -log10(pval_in),col=Q,label=rownames(results_in))) +
  geom_point(size=1.5) +
  geom_hline(yintercept = -log10(pval_bonf),
             col = 'grey40',
             linetype = 'dotted') +
  annotate('text',x=-0.5, y= -log10(pval_bonf + 0.000001),
             label=paste('p=' ,pval_bonf),
             size=2,col='grey30') +
  theme_few()+
  ggtitle('Questions for introvert determination')

p2 + geom_text(check_overlap = TRUE,aes(size= 1/se_in^2))
```



4. Choose significant questions (ideally about 5 questions)

for extrovert: Q80A+Q81A+Q83A+Q84A+Q91A

for introvert: Q80A+Q82A+Q84A+Q89A+Q90A+Q91A

5. Logistic regression: extrovert/introvert ~ Q1+Q2+Q3+age+sex

```
# logistic regression for predicting
model_ex <- glm(data=train, IE_extrovert ~ Q80A+Q81A+Q83A+Q84A+Q91A+age+gender,family='binomial')
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model_ex)
```

```
##
```

```
## Call:
## glm(formula = IE_extrovert ~ Q80A + Q81A + Q83A + Q84A + Q91A +
##     age + gender, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1223  -0.3629  -0.1727  -0.0890   3.3670
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.2457966  0.3343102  -0.735    0.462
## Q80A         0.3360427  0.0429670   7.821 5.24e-15 ***
## Q81A        -0.3625338  0.0456111  -7.948 1.89e-15 ***
## Q83A        -0.5168198  0.0541202  -9.549  < 2e-16 ***
## Q84A        -0.2976905  0.0474874  -6.269 3.64e-10 ***
## Q91A         0.4043185  0.0489608   8.258  < 2e-16 ***
## age         -0.0001158  0.0006991  -0.166    0.868
## gender      -0.1359513  0.0977651  -1.391    0.164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3976.4  on 4999  degrees of freedom
## Residual deviance: 2347.7  on 4992  degrees of freedom
## AIC: 2363.7
##
## Number of Fisher Scoring iterations: 13
```

```
model_in <- glm(data=train, IE_introvert ~ Q80A+Q82A+Q84A+Q89A+Q90A+Q91A+age+gender,family='binomial')
summary(model_in)
```

```
##
## Call:
## glm(formula = IE_introvert ~ Q80A + Q82A + Q84A + Q89A + Q90A +
##     Q91A + age + gender, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5729  -0.6677   0.3404   0.6836   2.4978
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.182e+00  2.824e-01   4.187 2.82e-05 ***
## Q80A        -2.945e-01  3.110e-02  -9.470  < 2e-16 ***
## Q82A         4.440e-01  3.171e-02  14.001  < 2e-16 ***
## Q84A         1.700e-01  3.631e-02   4.682 2.85e-06 ***
## Q89A        -1.610e-01  3.409e-02  -4.724 2.32e-06 ***
## Q90A        -2.202e-01  3.508e-02  -6.278 3.44e-10 ***
## Q91A        -2.946e-01  3.379e-02  -8.719  < 2e-16 ***
## age         -1.054e-06  6.278e-06  -0.168    0.867
## gender      -9.425e-03  6.813e-02  -0.138    0.890
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6664.6  on 4999  degrees of freedom
## Residual deviance: 4532.9  on 4991  degrees of freedom
## AIC: 4550.9
##
## Number of Fisher Scoring iterations: 8
```

```
model_in <- glm(data=train, IE_introvert ~ Q80A+Q82A+Q84A+Q89A+Q90A+Q91A,family='binomial')
summary(model_in)
```

```
##
## Call:
## glm(formula = IE_introvert ~ Q80A + Q82A + Q84A + Q89A + Q90A +
##     Q91A, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5699  -0.6675   0.3401   0.6842   2.4967
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.16378    0.24782   4.696 2.65e-06 ***
## Q80A        -0.29400    0.03110  -9.454  < 2e-16 ***
## Q82A         0.44347    0.03151  14.075  < 2e-16 ***
## Q84A         0.17068    0.03629   4.703 2.56e-06 ***
## Q89A        -0.15918    0.03385  -4.703 2.57e-06 ***
## Q90A        -0.22090    0.03507  -6.299 3.00e-10 ***
## Q91A        -0.29581    0.03377  -8.761  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6664.6  on 4999  degrees of freedom
## Residual deviance: 4535.3  on 4993  degrees of freedom
## AIC: 4549.3
##
## Number of Fisher Scoring iterations: 5
```

Model for probability of extrovert: log(p/1-p) = -0.2457966 + 0.3360427* Q80A -0.3625338*Q81A -0.5168198*Q83A -0.2976905*Q84A + 0.4043185*Q91A -0.0001158*age -0.1359513*gender

Model for probability of introvert: log(p/1-p) = 1.16378 -0.29400* Q80A + 0.44347*Q82A +0.17068*Q84A -0.15918*Q89A -0.22090 *Q90A –0.29581*Q91A

age and gender is not important in determination of introvert!

6. Cross validation

```
test$predict_ex <- ifelse(predict(model_ex,newdata = test,type='response')>0.5,1,0)
test$predict_in <- ifelse(predict(model_in,newdata = test,type='response')>0.5,1,0)

#predict probability
accuracy_extrovert <- mean(test$IE_extrovert == test$predict_ex)
paste('The accuracy of our extrovert predicting model is ',
      round(accuracy_extrovert*100,digits=3),'%',sep='')
```

```
## [1] "The accuracy of our extrovert predicting model is 90.401%"
```

```
accuracy_introvert <- mean(test$IE_introvert == test$predict_in)
paste('The accuracy of our introvert predicting model is ',
      round(accuracy_introvert*100,digits=3),'%',sep='')
```

```
## [1] "The accuracy of our introvert predicting model is 80.604%"
```

```
## parallel connection
test$predict_ex_par <- ifelse(predict(model_ex,newdata = test,type='response')>0.5 |
                              predict(model_in,newdata = test,type='response')<0.5,1,0)
test$predict_in_par <- ifelse(predict(model_ex,newdata = test,type='response')<0.5 |
                              predict(model_in,newdata = test,type='response')>0.5,1,0)

accuracy_extrovert_par <- mean(test$IE_extrovert == test$predict_ex_par)
paste('The accuracy of our parallel connection extrovert predicting model is ',
      round(accuracy_extrovert_par*100,digits=3),'%',sep='')
```

```
## [1] "The accuracy of our parallel connection extrovert predicting model is 76.447%"
```

```
accuracy_introvert_par <- mean(test$IE_introvert == test$predict_in_par)
paste('The accuracy of our parallel connection introvert predicting model is ',
      round(accuracy_introvert_par*100,digits=3),'%',sep='')
```

```
## [1] "The accuracy of our parallel connection introvert predicting model is 71.895%"
```

```
## series connection

test$predict_ex_ser <- ifelse(predict(model_ex,newdata = test,type='response')>0.5 &
                              predict(model_in,newdata = test,type='response')<0.5,1,0)
test$predict_in_ser <- ifelse(predict(model_ex,newdata = test,type='response')<0.5 &
                              predict(model_in,newdata = test,type='response')>0.5,1,0)

accuracy_extrovert_ser <- mean(test$IE_extrovert == test$predict_ex_ser)
paste('The accuracy of our series connection extrovert predicting model is ',
      round(accuracy_extrovert_ser*100,digits=3),'%',sep='')
```

```
## [1] "The accuracy of our series connection extrovert predicting model is 90.401%"
```

```
accuracy_introvert_ser <- mean(test$IE_introvert == test$predict_in_ser)
paste('The accuracy of our series connection introvert predicting model is ',
      round(accuracy_introvert_ser*100,digits=3),'%',sep='')
```

```
## [1] "The accuracy of our series connection introvert predicting model is 80.604%"
```

Conlusion :

Generally, we prefer series models combining the two models. The predicted model has 90.4% accuracy for extrovert and 80.6% for introvert.

Model for probability of extrovert: log(p/1-p) = -0.2457966 + 0.3360427* Q80A -0.3625338*Q81A -0.5168198*Q83A -0.2976905*Q84A + 0.4043185*Q91A -0.0001158*age -0.1359513*gender

Model for probability of introvert: log(p/1-p) = 1.16378 -0.29400* Q80A + 0.44347*Q82A +0.17068*Q84A -0.15918*Q89A -0.22090 *Q90A –0.29581*Q91A

Age and gender is not important in determination of introvert.