ADS2 Coding Challenge 2

0049

2022-6-1

Import the necessary libraries

```
library(tidyverse)
```

1. Turtles

What is the probability that you are on East Beach?

First, I will make this question more clear using conditional probabilities.
P(Green | West) = 0.9
P(Loggerhead | West) = 0.1
P(Green | East) = 0.6
P(Loggerhead | East) = 0.4
I assume that the P(West) = 0.5 and P(East) = 0.5 The data here is: I find a turtle and examine it. It is a **Loggerhead Turtle**
Hypothesis 1: I am on the West Beach.
Hypothesis 2: I am on the East Beach.

P(Loggerhead, West) = 0.1*0.5 = 0.05 P(East, Loggerhead) = 0.4*0.5 = 0.2 P(Loggerhead) = 0.25 P(East | Loggerhead) = 0.2/0.25 = 0.8

The probability that I am on East Beach is 80%.

What additional assumptions do you have to make to arrive at this probability?

First, the real possibilities that I meet this two kinds of turtles on each beach should be the same as the real turtle distribution. That is, I have the same possibility to meet any of the turtle on the beach.

Second, I assume that the prior possibilities for both benches are 0.5. That is, the P(West) = P(East) = 0.5. If the prior possibilities are changed, the result will be changed.

2. Classifying neuron types from electrophysiological recordings

Import the original data and plot it in a useful format to show the original classifications (type)

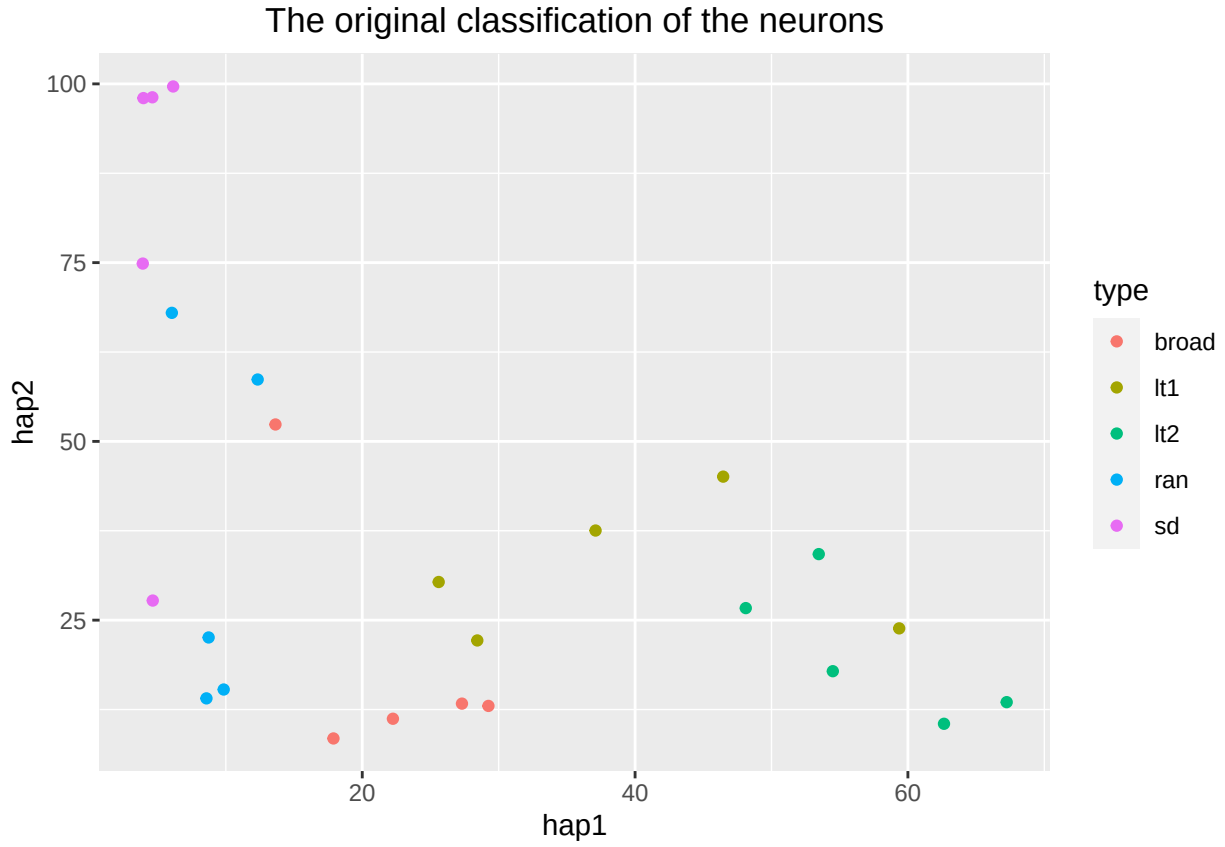
First, import the data.

```
neuron <- read.csv("vmndata.csv")
head(neuron)
```

```
##   type   hap1   hap2
## 1  ran  8.580 14.060
## 2  ran  6.045 67.975
## 3  ran  8.740 22.580
## 4  ran  9.845 15.300
## 5  ran 12.335 58.655
## 6   sd  6.140 99.635
```

Then, plot the data.

```
ggplot(neuron, aes(x = hap1, y = hap2)) +
  geom_point(aes(color = type)) +
  labs(title = "The original classification of the neurons") +
  theme(plot.title = element_text(hjust = 0.5))
```



Use clustering of the model fit data to make your own classification of the recordings (You are allowed to use the *kmeans* function in R)

I will use kmeans function with 5 centers. For the number 1 to 5, each represents the cluster number of the neuron on the same position.

```
cluster <- kmeans(neuron[, 2:3], centers = 5, iter.max = 10)[[1]]
cluster
```

```
##  [1] 5 4 5 5 4 2 2 2 4 5 5 1 3 3 5 3 3 1 1 1 5 4 5 5 5
```

Plot the outcome of this clustering (e.g. by assigning colours by cluster).

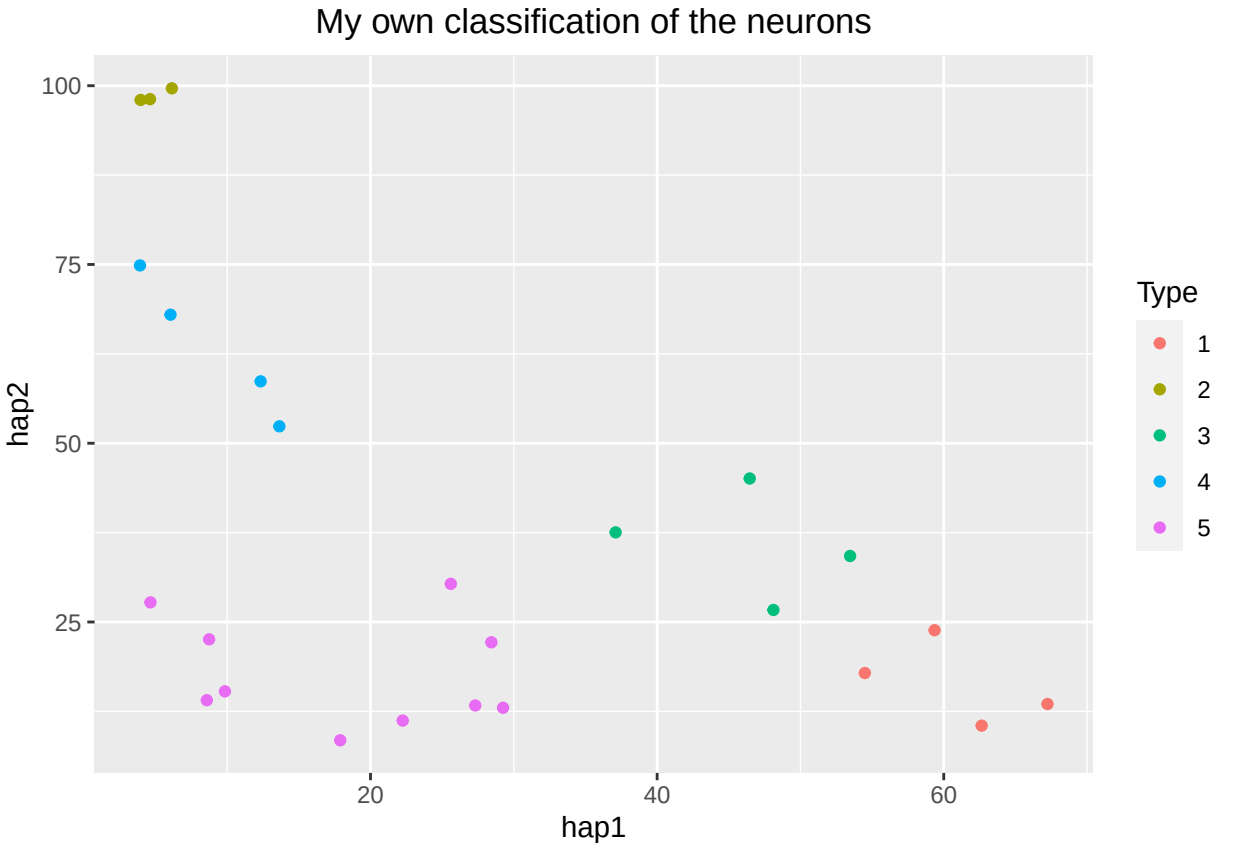
First, I will assign the cluster number to the original data

```
neuron$cluster <- cluster
head(neuron)
```

```
##   type   hap1   hap2 cluster
## 1  ran  8.580 14.060       5
## 2  ran  6.045 67.975       4
## 3  ran  8.740 22.580       5
## 4  ran  9.845 15.300       5
## 5  ran 12.335 58.655       4
## 6   sd  6.140 99.635       2
```

Then, I will plot the outcome of this clustering.

```
ggplot(neuron, aes(x = hap1, y = hap2)) +
  geom_point(aes(color = factor(cluster))) +
  labs(title = "My own classification of the neurons",
       col = "Type") +
  theme(plot.title = element_text(hjust = 0.5))
```



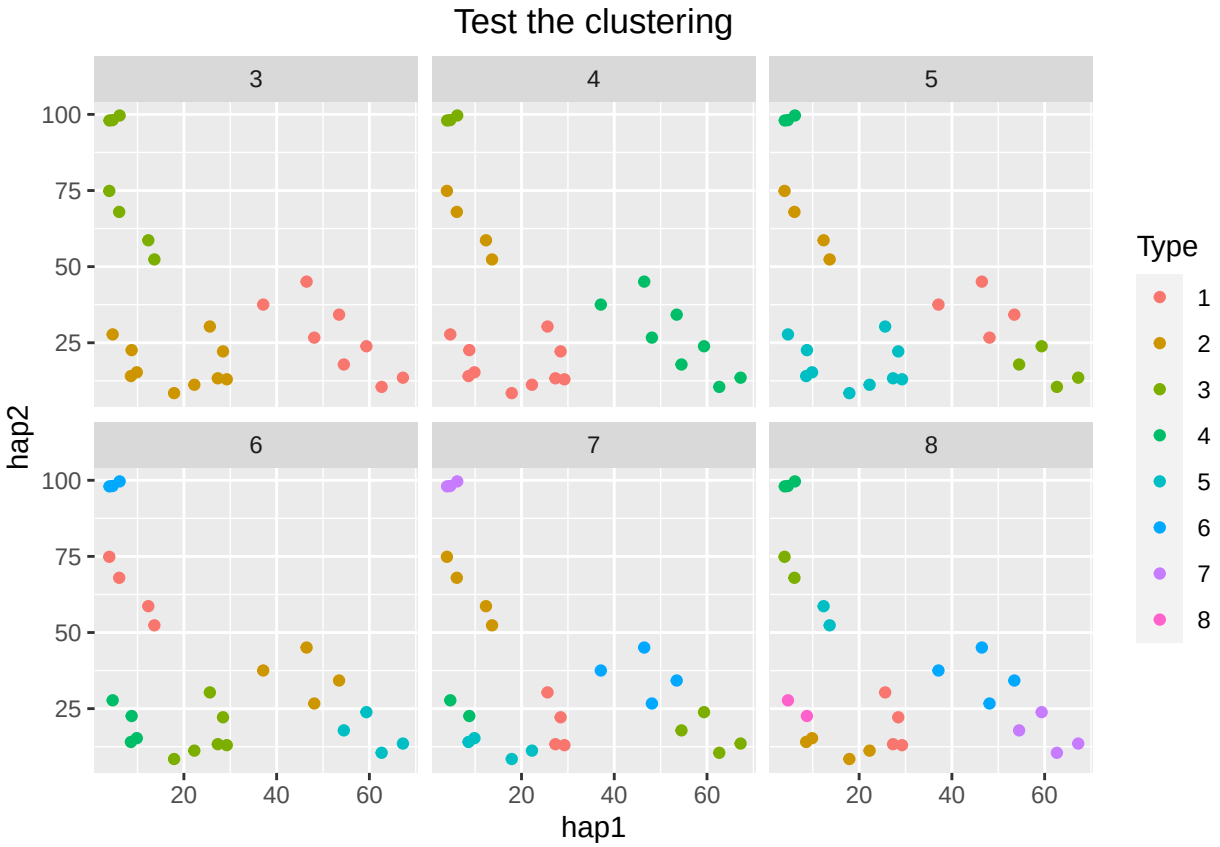
Test the clustering using different subsets of the fit parameters. Describe in words how your clustering compares to the original classifications.

To test the clustering, I will try to use different numbers of clusters (From 3 to 8) and decrease the maximum number of iterations allowed to increase the speed.

```
for(center in 3:8){
  cluster <- kmeans(neuron[, 2:3], centers = center, iter.max = 5)[[1]]
  neuron[, center+2] <- cluster
}

colnames(neuron)[5:10] <- c("3", "4", "5", "6", "7", "8")
test_neuron <- gather(neuron, 5:10, key = "center_number", value = "clusters")

ggplot(test_neuron, aes(x = hap1, y = hap2)) +
  geom_point(aes(color = factor(clusters))) +
  facet_wrap(vars(center_number), nrow = 2) +
  labs(title = "Test the clustering",
       col = "Type") +
  theme(plot.title = element_text(hjust = 0.5))
```



It seems that the clustering performance good when the numbers of centers are 4 or 5.

From the two point plots, my clustering performs much better than the original classifications in this situation. My clustering result has smaller within-cluster variance than the previous one. For details, the points belonging to the same cluster in the original classification are relatively scattered. For example, one purple point representing the 'sd' type keeps away from the other purple points. However in my clustering, all the points in the same type are gathered.

3. Gene knockout in T-cells and PD1 blockade

Import the data, organize it appropriately, and plot these in a useful way.

Import the data.

```
cell <- read.csv("tcells.csv")
head(cell)
```

```
##   Number        Group Stimulation.Index
## 1     43      KO-aPD1               2.49
## 2     44 Control-aPD1               3.11
## 3     46      KO-aPD1               2.21
## 4     47       KO-LCM               2.10
## 5     48       KO-LCM               1.23
## 6     49  Control-LCM               2.19
```

Organize the data.
Since the data has 4 groups, I want to see whether there are some strange patterns.

```
table(cell$Group)
```

```
##
## Control-aPD1  Control-LCM      KO-aPD1       KO-LCM
##            8            8            8            8
```

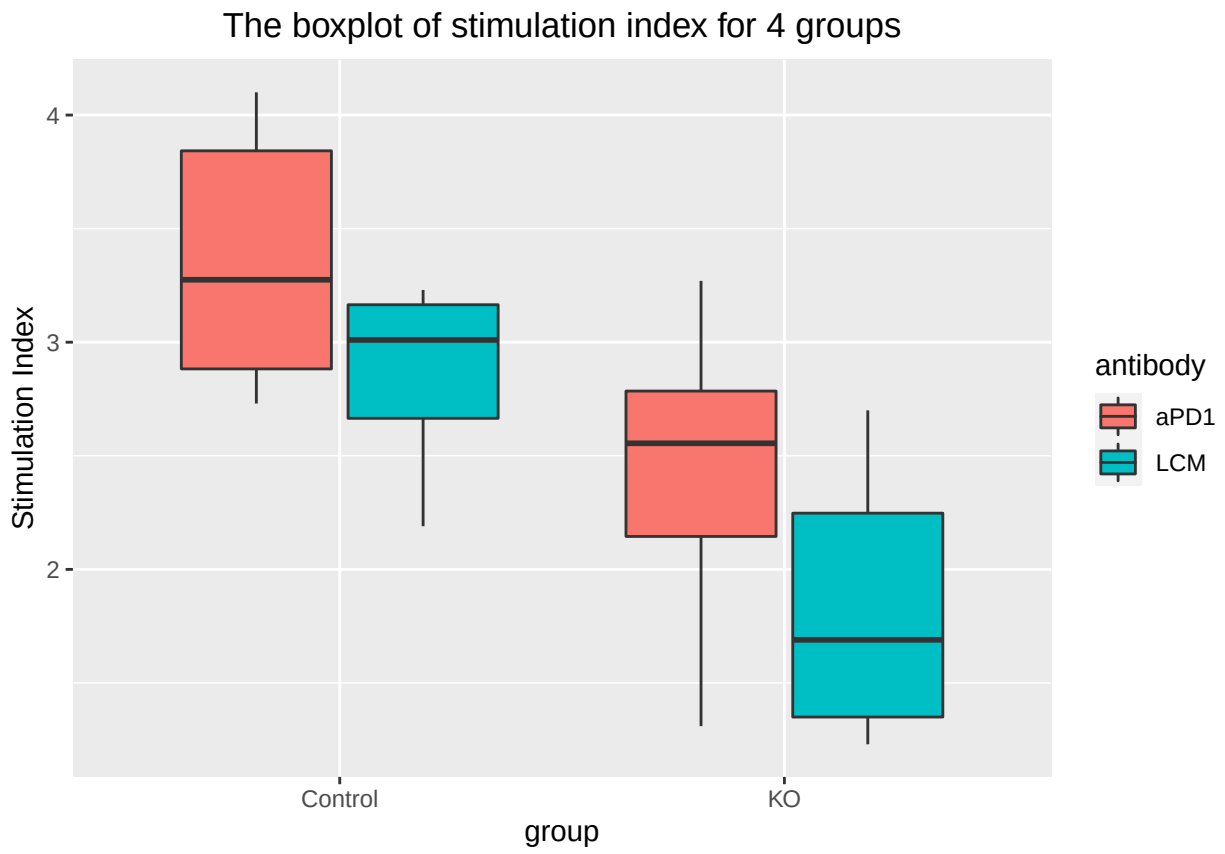
Each group has 8 mice. Then I will divide this four groups into two factors.

```
cell2 <- separate(cell, Group, sep = "-", into = c("group", "antibody"))
head(cell2)
```

```
##   Number   group antibody Stimulation.Index
## 1     43      KO     aPD1              2.49
## 2     44 Control     aPD1              3.11
## 3     46      KO     aPD1              2.21
## 4     47      KO      LCM              2.10
## 5     48      KO      LCM              1.23
## 6     49 Control      LCM              2.19
```

Plot the data.

```
ggplot(cell2, aes(x = group, y = Stimulation.Index)) +
  geom_boxplot(aes(fill = antibody)) +
  labs(title = "The boxplot of stimulation index for 4 groups",
       y = "Stimulation Index") +
  theme(plot.title = element_text(hjust = 0.5))
```



Formulate the correct statistical hypotheses, choose the appropriate statistical test, and check assumptions for this test. Explain your choice briefly.

For the first question: Whether the gene knockout can affect T-cell proliferation. **H0: The gene knockout has no effect on T-cell proliferation.**
HA: The gene knockout can affect T-cell proliferation.
To solve this question, I will select two groups associated with LCM. Since the PD1 factor will affect the result.

```
control <- filter(cell2, group == 'Control', antibody == 'LCM')
KO <- filter(cell2, group == 'KO', antibody == 'LCM')
```

Since I want to compare two groups, I will use t-test and test its assumptions. **H0: The SIs in two groups are normally distributed.**
HA: The SIs are not normally distributed in at least one group.

```
shapiro.test(control$Stimulation.Index)$p
```

```
## [1] 0.1749068
```

```
shapiro.test(KO$Stimulation.Index)$p
```

```
## [1] 0.07216832
```

Both p-values are higher than 0.05. So we do not reject H0. So the assumptions are met and I will use t-test for question 1.

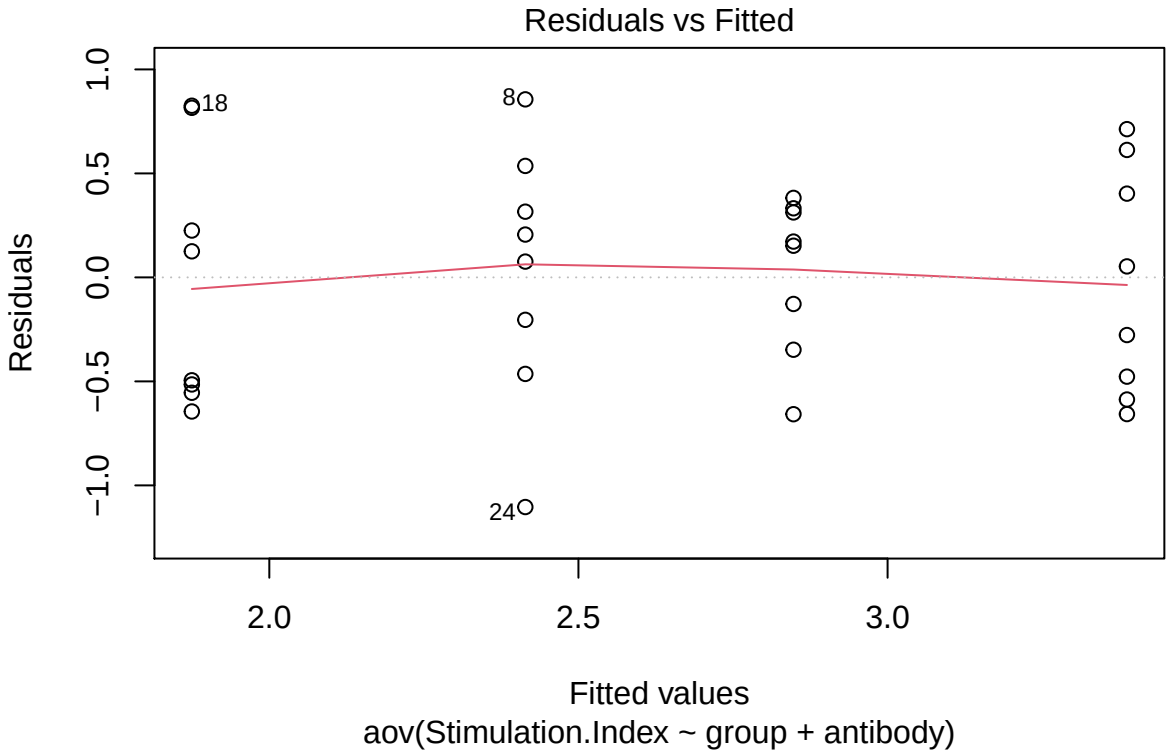
For the second question: If PD1 blockade can rescue the phenotype. I want to compare 4 groups and get the right conclusion. Therefore, I want to use ANOVA to test it. **H0: PD1 blockade can rescue the phenotype.**
HA: PD1 blockade cannot rescue the phenotype.
Then I will test the assumptions of ANOVA.

```
model <- aov(data = cell2, Stimulation.Index~group + antibody)
shapiro.test(resid(model))$p.value
```

```
## [1] 0.2323453
```

The residuals are normally distributed. Next I will test the equality of variances

```
plot(model, 1)
```



The heights of each columns are similar. So I can use ANOVA for question 2.

Identify whether there is any difference between the experimental groups, which factor attributes to the differences, and which factor has a higher effect on T-cell proliferation.

First for question 1: I will test the variance. **H0: The variances of SI in two groups are the same.**
HA: The variances of SI are not different in at least one group.

```
var.test(control$Stimulation.Index, KO$Stimulation.Index)$p.value
```

```
## [1] 0.2103624
```

The p-value is higher than 0.05. So we do not reject H0.

```
t.test(control$Stimulation.Index, KO$Stimulation.Index, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  control$Stimulation.Index and KO$Stimulation.Index
## t = 4.0468, df = 14, p-value = 0.001201
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4829313 1.5720687
## sample estimates:
## mean of x mean of y
##    2.8750    1.8475
```

The p-value is much smaller than 0.05. So we reject H0. **That is, the gene knockout can affect T-cell proliferation.**

Then for question 2:

```
summary(model)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## group        1  7.576   7.576  26.168 1.84e-05 ***
## antibody     1  2.327   2.327   8.039  0.00826 **
## Residuals   29  8.396   0.290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-values are less than 0.05 so there is difference between the experimental groups.

```
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Stimulation.Index ~ group + antibody, data = cell2)
##
## $group
##                 diff       lwr        upr    p adj
## KO-Control -0.973125 -1.362194 -0.5840563 1.84e-05
##
## $antibody
##                diff        lwr        upr     p adj
## LCM-aPD1 -0.539375 -0.9284437 -0.1503063 0.0082565
```

So both the groups and antibodies attribute the differences. And the group (Control vs KO) seems to have a higher effect on T-cell proliferation.

What would you suggest to do next?

I will test whether there may have interactions between different groups (control or experiment) and the antibodies used. Also the number of sample is too small, I will use more mouse models to verify my conclusions.