# Final project Data Analysis Report

Gu Gong

2023/3/14

# 1 Abstract

In the study conducted by Steinmetz et al. (2019), different mice are assigned to various ways of experiments, which are letting them to act under 4 levels of stimuli from right and left screen in order to observe the differences of activity of neurons in the visual cortex. In this project, the main research interests are to find out how the left and right contrast attach influence on the neuron activities. Moreover, based on the model and conclusion, this project is able to predict whether the neuron would react to the different contrasts and how they will react. To investigate the interest of this project, this report mainly focus on the the data of five sessions from two mice, Cori and Frossman. By fitting the data into an ANOVA model with random effects and clustering them into 3 different groups, the report is able to tell that the left contrasts and right contrasts will attach influence on the neurons activity. And the combination of left_contrast of 1 and right_contrast of 1 would have more spikes than other stimuli. After that, the report can predict whether the mouse would react correctly with logistic models. By knowing such conclusions, the readers can design more experiments related with visual nerve stimulation with less preparatory work. And using that conclusions, the related companies could help to improve driving safety, which based on the reaction caused by visual stimuli.

---

# 2 Introduction

From a long time ago, the study of neurons has been a really important area when studying the brains. In 2019, Steinmetz et al conduct an experiment about the working relations between neuron and stimuli. By setting different stimuli randomly with different contrast levels on both sides of screens around the mice, the researchers can observe how the mice react based on such stimuli. Moreover, through recording the activity of the neurons along with the time, the firing rate as the outcome variable in this project can be calculated, which would be analyzed with quantitative methods.

With the data collected by the former paper, this project concentrates on how the neurons would react under different stimuli(left contrast and right contrast in this case), get the fitted model reflecting the Working mechanism behind and how to predict the firing rate based on such fitted model. Since knowing the results of these interests could help to enhance the understanding of neurons.

With the results of this data analysis, the neurosciences can use that to help better to know how entire neurons reaction-systems operate and design more related experiments to investigate the mechanisms behind and that would eventually help the biopharmaceutical companies to design new drugs or to reduce the investigating time of new drugs. Moreover, the results can also help to design new treatments in brain and and psychological areas.

---

# 3 Background

In 2019, Nicholas A Steinmetz published a paper named as “Distributed coding of choice, action, and engagement across the mouse brain” to investigate how the widely distributed neurons in brains corporate with each other to finish complex tasks(Steinmetz, 2019). By that time, the research mainly focus on how the single area such as frontal cortex, parietal cortex and motor cortex, in brains work to face with tasks(Cisek, 2010). And the research conducted by Steinmetz use neuropixel probes, which can record the activity of nearly 1000 neurons in the same animal(Jun, 2017) to record Electrophysiological signals of nearly 30,000 neurons in 42 brain regions of the mouse brain to complete a visual discrimination task.

In this project, the recorded data would be used for statistical analysis to investigate how the visual stimulation affect the neuron activities. In this study, the 5 sessions are selected for better efficiency, and in each session, to measure the neuron activity, this project use the mean firing rate as the statistics. The detailed explanation of variable would be discussed later in the Descriptive analysis part.

Moreover, compare to the research paper “Thirst regulates motivated behavior through modulation of brainwide neural population dynamics”, which use similar tools to investigate activity of nearly 24,000 neurons in 34 brain regions in mice under thirst. we can find out that the visual signals of seeing water would attach influence on the neurons, which would eventually cause the neuron fierce activity. Based on such results, this report aims at repeating the same results that the visual stimulation would cause the neuron activities, by different statistical methods, such as ANOVA model or related clustering methods if needed.

---

# 4 Descriptive analysis

## 4.1 Preprocessing and data cleaning

Firstly, this project load the data with former 5 sessions to the R for further study and analysis.

There are over 5 variables existing in the sessions, and for the research purposes, the variable “spks”(numbers of spikes of neurons in the visual cortex in time bins defined in specialized time), “contrast_left”(contrast of the left stimulus), “contrast_right”(contrast of the right stimulus) and “feedback_type”(type of the feedback, 1 for success and -1 for failure) are selected. And by checking the missing values, we can see that there are no missing values in this dataset.

## 4.2 Choice of the summary measure

For better understanding of the experiment, we should transform the “spks” variable to firing rate. And doing that could eventually generate 1196 firing rates across different sessions along with the time. Choosing such statistics for advanced analysis for mainly 4 reasons:

1. The first one would be the experimental design in the original paper. In the original paper, the analysis window is recorded from 0 to 0.4 s after stimulus onset, which means only the spikes of firing in this segment are recorded.
2. Selecting one indicator, which can express information of the relatively complicated data is more efficient and that would not lose too much information of the original data.
3. The mean help people understand the distribution of the data and the overall performance of the dataset.
4. The mean value can be used for further analysis, such as calculating the variance and standard deviation.

## 4.3 Univariate descriptive analysis

Before further analysis of our research interests, the explanatory data analysis is needed for getting fundamental understanding of the structure, characteristics and properties of dataset. And such analysis could show the readers the pattern in the data and correlation between variables, which would eventually provide how the explaining variables affect the response variables and help to guide the modeling and further analysis.

Firstly, to show the detailed information of the amounts and mean firing rates of various contrasts level, the tables are offered below:

**Table1 Summary contrasts of Mouse Cori**

| | session 1 | | session 2 | | session 3 | |
|---|---|---|---|---|---|---|
| **Contrast.level** | **Left1** | **Right1** | **Left2** | **Right2** | **Left3** | **Right3** |
| 0.00 | 97 | 86 | 133 | 115 | 137 | 109 |
| 0.25 | 46 | 25 | 25 | 41 | 31 | 26 |
| 0.50 | 36 | 41 | 39 | 34 | 28 | 31 |
| 1.00 | 35 | 62 | 54 | 61 | 32 | 62 |

**Table2 Summary contrasts of Mouse Frossman**

| | session 4 | | session 5 | |
|---|---|---|---|---|
| Contrast.level | Left4 | Right4 | Left5 | Right5 |
| 0.00 | 112 | 107 | 112 | 105 |
| 0.25 | 41 | 55 | 46 | 48 |
| 0.50 | 46 | 41 | 43 | 45 |
| 1.00 | 50 | 46 | 53 | 56 |

Based on the table result above, we can see that the left and right stimuli amounts across different sessions. And based on the the original paper of Steinmetz et al. (2019), we can know the amounts of the left and right stimuli are set randomly. And for each session, even if there are different amounts of left contrasts and right contrasts, the total contrasts from different levels and type(left and right) holds.

**Table3 Summary Feedback types and mean firing rate of Mouse Cori**

| | session 1 | | session 2 | | session 3 | |
|---|---|---|---|---|---|---|
| Variable | Success1 | Failure1 | Success2 | Failure2 | Success3 | Failure3 |
| amount | 141.00 | 73.00 | 159.00 | 92.00 | 151.00 | 77.00 |
| mean firing rate | 4.34 | 3.75 | 3.39 | 3.21 | 3.72 | 3.33 |

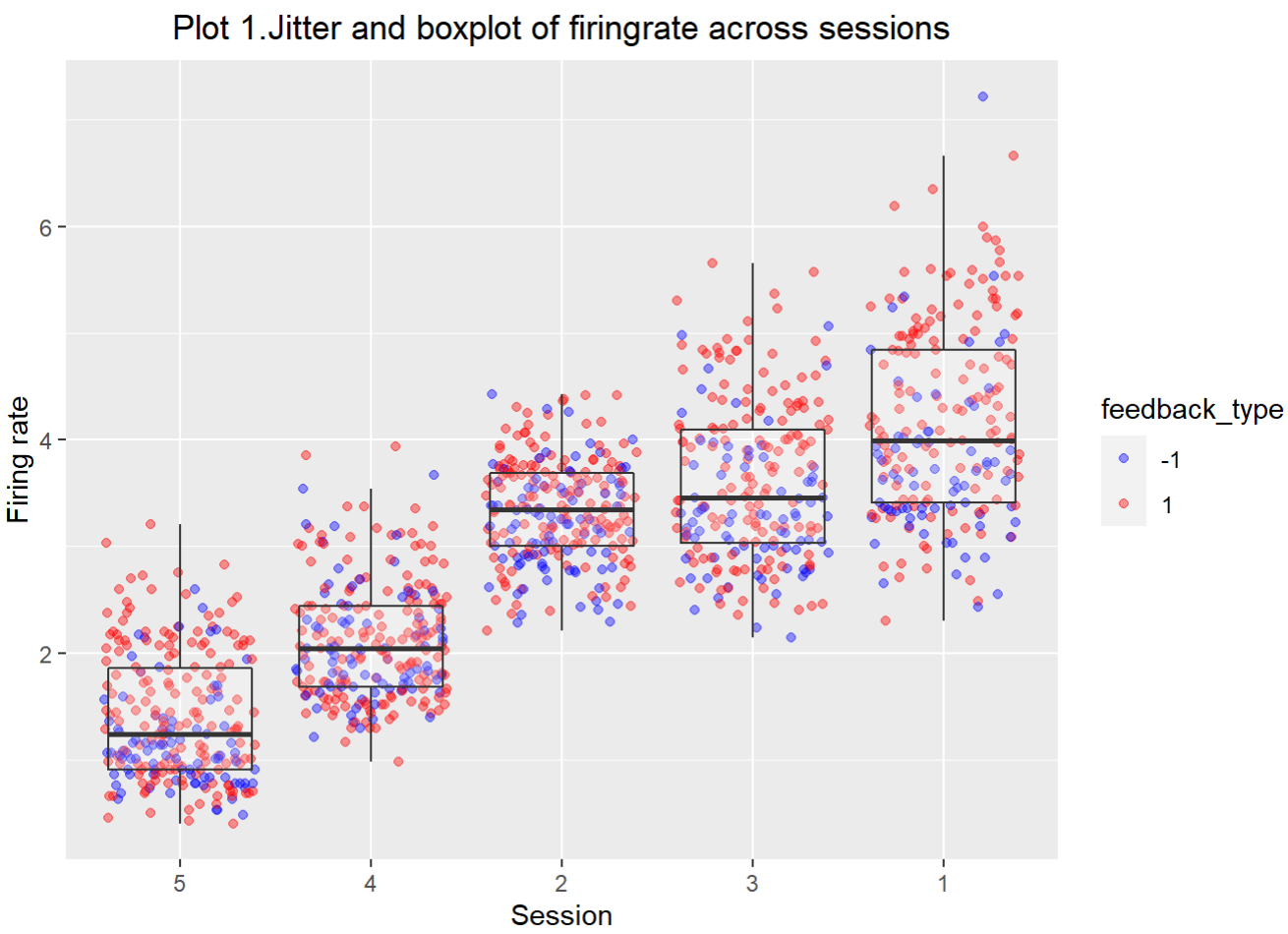
**Table4 Summary Feedback types and mean firing rate of Mouse Frossman**

| | session 4 | | session 5 | |
|---|---|---|---|---|
| Vairable | Success4 | Failure4 | Success5 | Failure5 |
| amount | 166.00 | 83.00 | 168.00 | 86.00 |
| mean firing rate | 2.14 | 2.06 | 1.49 | 1.18 |

Based on the table result above, for different session, the feedback type has a similar trend, which the success feedback is more than the failure type. However, we can see for two different mouse the overall mean firing rate should be close to each other.

# 4.4 Multivariate descriptive analysis

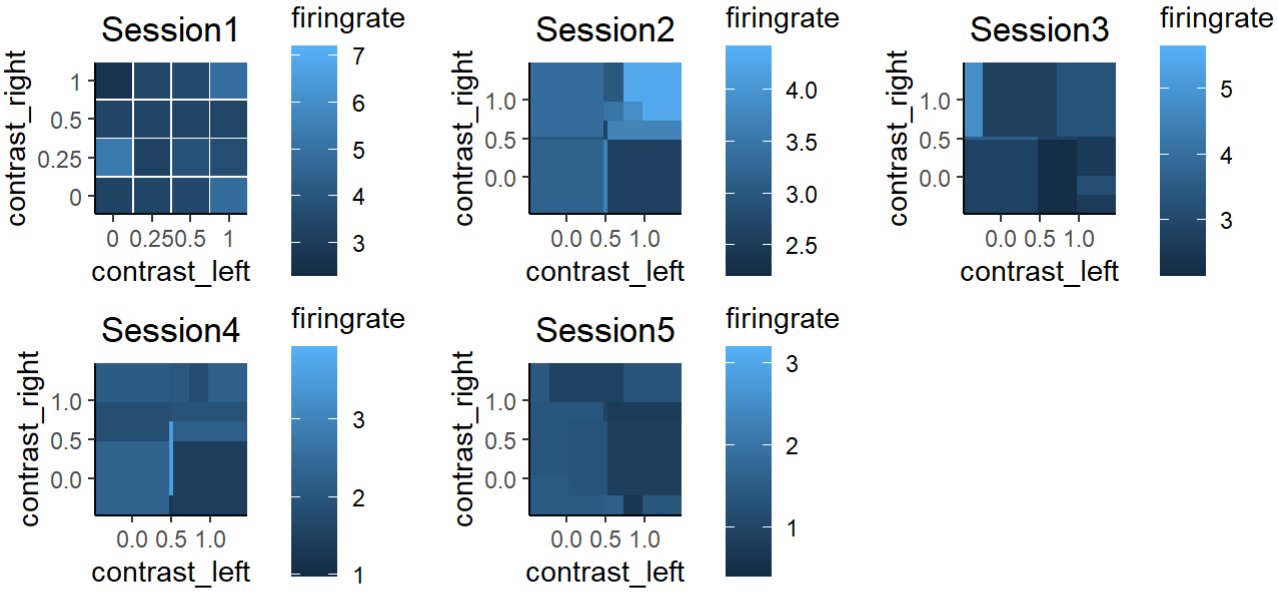
To get more clear description of the dataset, a Jitter-plot is drawn for better explanation as shown below:




Based on the jitter-boxplot shown above, we can see that the session1 have an obvious higher firing rate as well as the most dispersed distribution. And for the mouse Cori, the data is more distributed than the mouse Frossman. Moreover, the mouse Cori have higher firing rate even in different sessions.

Moreover, we can see that the failure feedback type has generally lower firing rate than the success feedback type across different sessions. Therefore, it is reasonable to guess that for failure feedback, the neuron activity would act less compare with the success feedback type. In other words, whenever the mouse get succeed in reacting correctly, they would normally have more neuron activities. However, for the session1, which is the session has higher firing rate, the success and failure feedback would have a more even distribution than what the other sessions.

Based on the analysis above, it is reasonable to guess that the firing rate would differ a lot for different sessions, hence in the subsequent quantitative analysis, it is necessary to set then as random intercept in our model to reduce the influences.




Moreover, we can get how the firing rates change through the different contrasts level as shown above. Based on these plots, we can see that generally with higher contrast levels, the mean firing rate would go higher. To verify such results, it is necessary to set a general anova model based on the dataset and get a more reliable quantitative conclusion.

# 5 Inferential analysis

## 5.1 Overview and parameters explanation

Based on the research interests, an ANOVA model as follows: $Y_{ijkl} = \mu_{...} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + \epsilon_{ijkl}$ is used for Inferential analysis. In this model, the index $i$ represents the fixed effects: left contrasts: 0 ($i = 1$), 0.25 ($i = 2$), 0.5 ( $i = 3$), 1 ($i = 4$), and the index $j$ represents the fixed effects: right contrasts: 0 ($j = 1$), 0.25 ($j = 2$), 0.5 ($j = 3$), 1 ( $j = 4$), $k$ represents the session indicator, which is treated as the random effect in this case since the research interests mainly focus on how the stimuli affect the neuron level and switching the session does affect the model structure, $l$ represents the neuron indicator, $\epsilon_{ijkl}$ represents the unobserved error and $\mu_{...}$ represents the population mean firing rate in this case.

## 5.2 Model choosing

However, in this project, we consider the random effect into the model, here we use the LR test and F test to test if the random effect is significant enough in our model. And based on the R result, we can get the p-value for the result without random effect is shown below:

**Anova Table of final model**

| X. | Df | Sum.Sq | Mean.Sum.Sq | F.value | Pr..F. |
|---|---|---|---|---|---|
| contrast_right | 3 | 3.1 | 1.022 | 0.747 | 5.24e-01 |
| contrast_left | 3 | 54.2 | 18.083 | 13.209 | 0.00e+00 |

| X. | Df | Sum.Sq | Mean.Sum.Sq | F.value | Pr..F. |
|---|---|---|---|---|---|
| Interaction | 9 | 62.0 | 6.884 | 5.029 | 1.10e-06 |
| Residuals | 1180 | 1615.4 | 1.369 | NA | NA |

Therefore, we can get the F-test result as: Test statistic is $MSTR/MSE \sim F(r-1,(n-1)r)$. And P-value is less than 0.05, which means that the random effect exists for this model. And the LR test can the p-value is also less than 0.05, which corss-validate the result from other tests. Hence, we can keep the random effect in our model.

However, to consider whether there is an interaction term existed in the model, the interactions still need to be verified in a statistical way to see how the combination of them would attach effect on the firing rate. Hence the reduced model is established in this case to verify if the reduced model has a better efficiency of analysis and dropping such terms would attach negative influence on the model accuracy. In this project, Based on the R result, we can see that under the confidence level 0.05, the p-value for comparing the two models is way less than 0.05. Hence, we can say that the more complicated model is better than the reduced model, and this interaction term should continue to stay in this model and not to get removed.

Therefore, we should keep the oringial complicated model for further analysis.

# 5.3 assumptions on proposed model

Based on the explanation of each parameters, we can get the constraints on the $\alpha_i$ and $\beta_j$ as $\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{b}\beta_j = 0$ and $\epsilon_{ijk}$ are i.i.d. $N(0,\sigma^2)$.

The assumptions on this model would mainly focus on that $\epsilon_{ijkl}$ are i.i.d. $N(0,\sigma^2)$. Here we can list them out for better understandings:

1. According to the formula above, equal variance of response variables for different factor levels can be represented by error terms having the same variance.
2. The independence of data can be represented in the formula by error terms being independently distributed, meaning they are not affected by or related to each other.
3. For the formula above, the assumption of identical normal distribution of response variables for different factor levels can be represented by error terms having an identical normal distribution with an expected value of 0.

# 5.4 Model fitting

**Anova Table of final model**

| X. | npar | Sum.Sq | Mean.Sum.Sq | F.value |
|---|---|---|---|---|
| contrast_right | 3 | 14.4620 | 4.8207 | 12.0656 |
| contrast_left | 3 | 25.3466 | 8.4489 | 21.1465 |
| Interaction | 9 | 6.9621 | 0.7736 | 1.9361 |

Based on the R result above, we can see that all variables have the large F-value, indicating all variables would have statistically significant impact on the response variable, the mean firing rate. Since in this project, we have the left contrast and right contrast shown as a pair of data, which means that the anova model is a balanced design, here this project does not need to fix the model with different type of sum of squares.

# 5.5 hypotheses test and inference

Based on the research purpose, firstly we can set up the hypotheses test for the contrast_right.

$H_0 : \alpha_i = 0 \ \forall \ i \quad v.s. H_1 : not\ all\ \alpha_i\ are\ zero$

Based on the Anova table in previous part, we can get that P-value is 3.00e-07, which is less than 0.05, the significance level. Therefore, we can conclude that we reject the null hypothesis and thus to get that there is significant difference of firing rate for different left contrasts.

Then we can set up the hypotheses test for different contrast_left.

$H_0 : \beta_j = 0 \ \forall \ j \quad v.s. H_1 : not\ all\ \beta_j\ are\ zero$

Based on the Anova table in previous part, we can get that P-value is 4.54e-02, which is less than 0.05, the significance level. Therefore, we can conclude that we reject the null hypothesis and thus to get that there is significant difference of firing rate for different right contrasts.

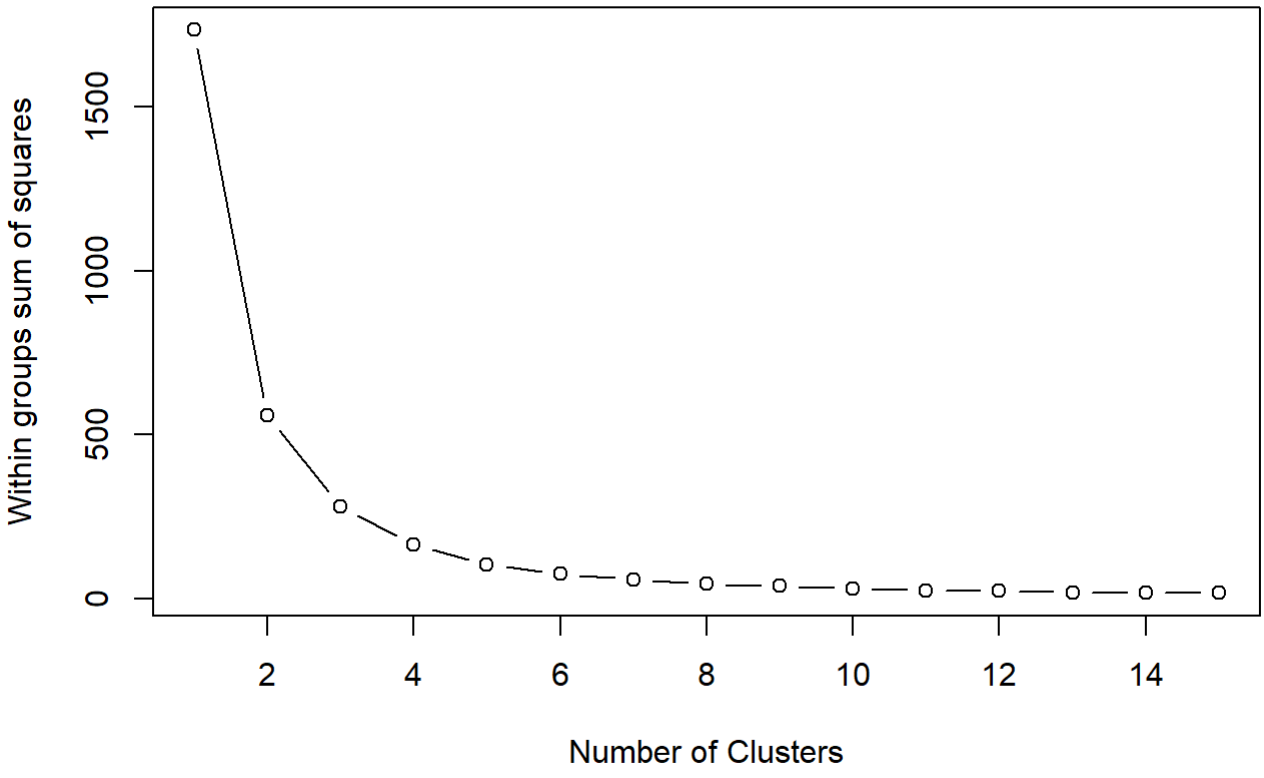
However, even if knowing the variables does attach influence on the response variables, the ANOVA model still is less capable of telling which one of combinations of contrast level would affect more. Therefore, the further analysis is needed.

# 5.6 Classfication

Based on the explanatory data analysis, we can know that for each combination of left and right contrasts of various levels and various firing rate, they would represent different neurons level. Although it may be the same neurons of the mouse act differently across sessions, we can consider them as different neurons to investigate how their firing rates are affected by different combination of contrast levels and types.

Based on different mean firing rate of neurons, we should consider to cluster them into different groups. By investigating such classification, this project is able to tell what kinds of combination of contrasts would have relatively high mean firing rate, which will help the project to cross-validate the anova model results and get a more precise conclusion. However, the application of classification method still need other evidences to be proved that the methods are considered to be valuable to use. Therefore, the density of firing rate across different sessions are plotted below for persuasion.

In this project, the method of classification would mainly use the K-means method to fulfil the classification based on its high-efficiency and interpretable. Moreover, the k-means clustering algorithm is easy to implement, and has good performance when processing large datasets.



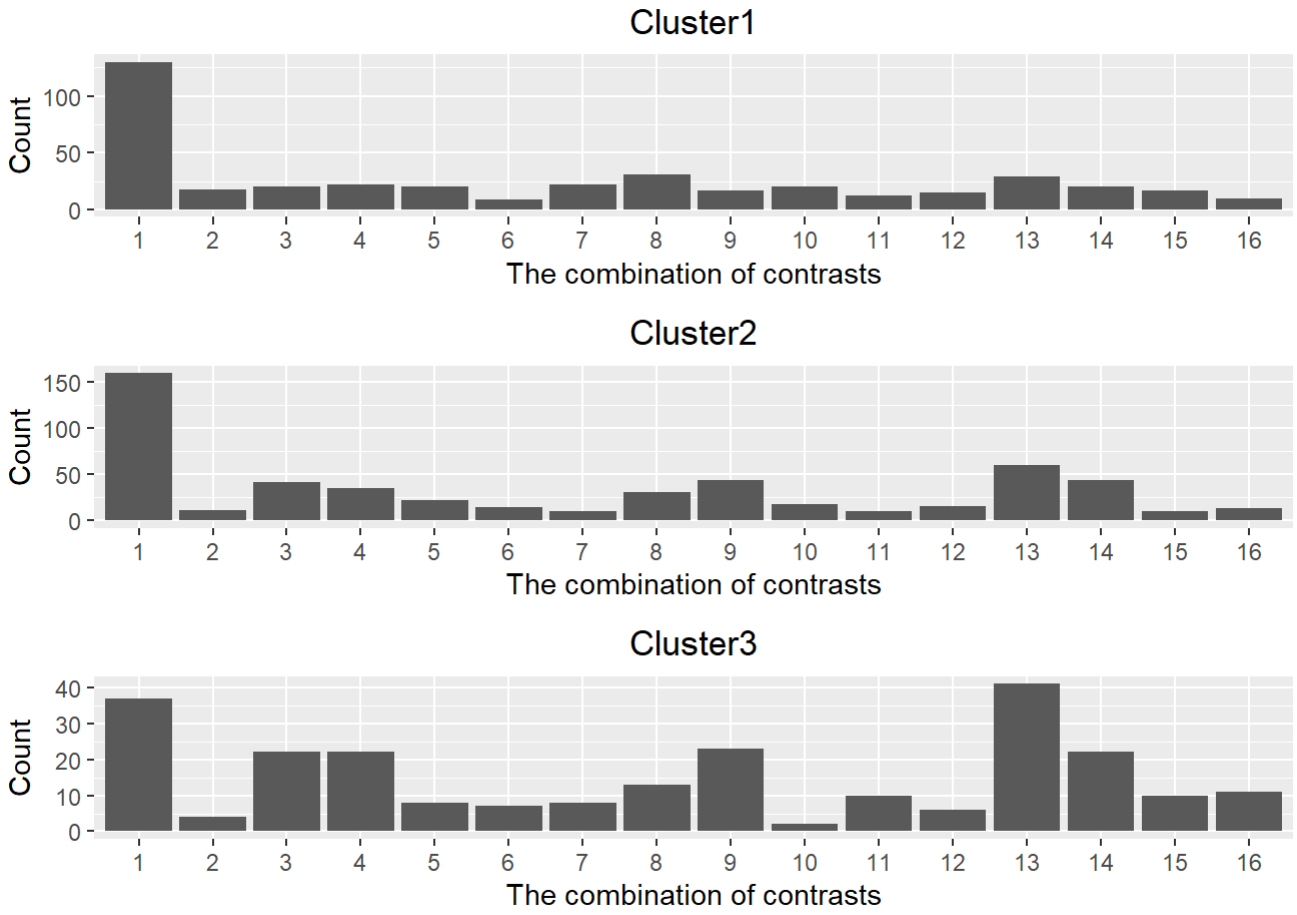
Based on the plot shown above, here we use the Total within sum of squares as the criterion. By showing the goodness of different cluster choosing, we can see the overall goodness would increase with the increasing number of clusters, but the growth trend does not grow very quickly after CLUSTER reaches three. Considering efficiency and explanatoryness, the KMEANS method of choosing 3 clusters is used to classify the data.

Based on the determined number of clusters, the project can draw the plot to show how the combination of left and right contrasts attach influence on the mean firing rate. Based on the clustering results, the project shows that for the first cluster, the mean firing rate is shown below:

**Table6 Summary Means of each cluster**

| Index | Cluster.1 | Cluster.2 | Cluster.3 |
|---|---|---|---|
| Mean | 1.51744 | 4.519926 | 3.129715 |

Moreover, we can plot the bar chart to show the counts of each combination of contrasts in each cluster. Here we use the number indicator to represent the combination of contrast: 1 stands for the left-contrast of 0 with the right-contrast of 0, 2 stands for the left-contrast of 0 with the right-contrast of 0.25, 3 stands for the left-contrast of 0 with the right-contrast of 0.5, 4 stands for the left-contrast of 0 with the right-contrast of 1. 5 stands for the left-contrast of 0.25 with the right-contrast of 0, 6 stands for the left-contrast of 0.25 with the right-contrast of 0.25, 7 stands for the left-contrast of 0.25 with the right-contrast of 0.5, 8 stands for the left-contrast of 0.25 with the right-contrast of 1. 9 stands for the left-contrast of 0.5 with the right-contrast of 0, 10 stands for the left-contrast of 0.5 with the right-contrast of 0.25, 11 stands for the left-contrast of 0.5 with the right-contrast of 0.5, 12 stands for the left-contrast of 0.5 with the right-contrast of 1. 13 stands for the left-contrast of 1 with the right-contrast of 0, 14 stands for the left-contrast of 1 with the right-contrast of 0.25, 15 stands for the left-contrast of 1 with the right-contrast of 0.5, 16 stands for the left-contrast of 1 with the right-contrast of 1.



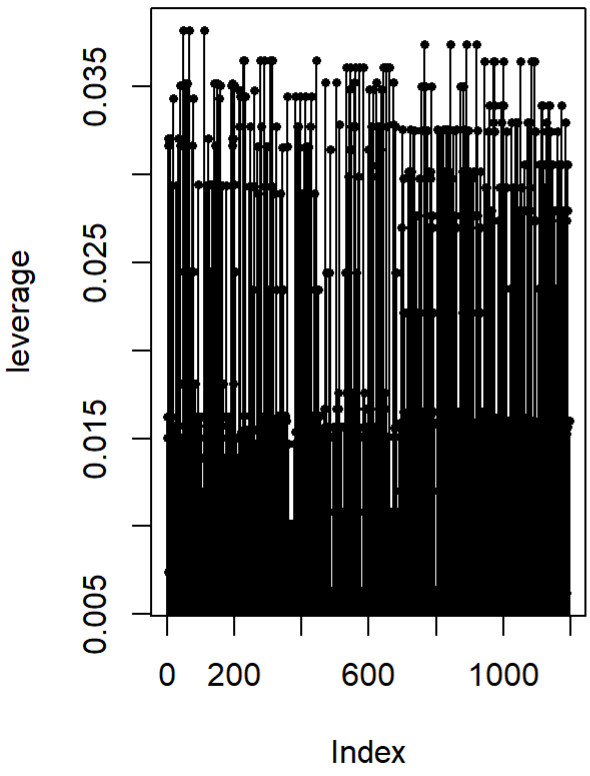

Based on the plots shown above, for any clusters, the combination of lowest left levels and right levels always prevail. Hence it is reasonable to ignore it at the subsequent analysis based on clustering method.

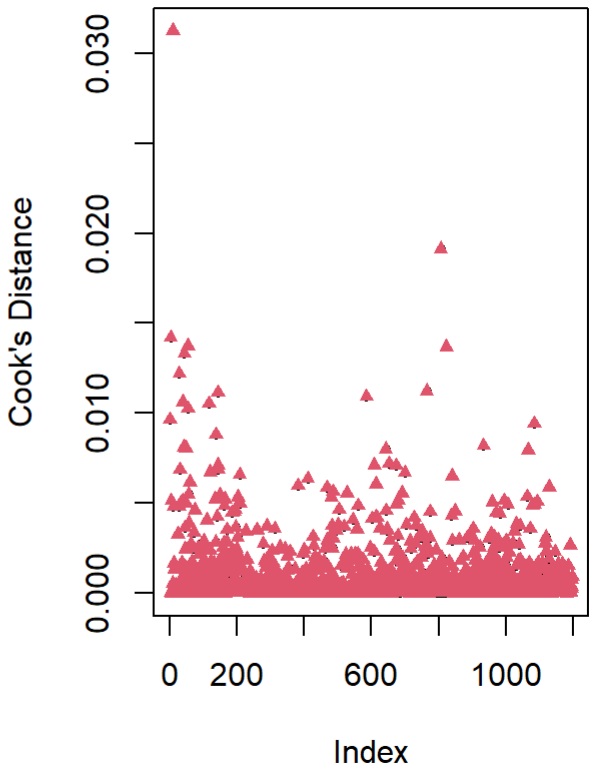
And we can see that when left contrasts hold the highest level(in this case, the combination 13,9), they are clustered into the cluster 2 and 3, which are the clusters that having higher mean firing rate. And that would verify our conclusion drawn from ANOVA model, which is that with higher contrasts, the mice would tend to have more neuron activity.

# 6 Sensitivity analysis

## 6.1 Assumptions test based on plots


From the Pearson residuals and deviance residuals, we can see that the two kinds of residuals are quite similar to each other, the model would not suffer from potential lack-of-fit.



For outliers test, this project use the cooks distance and leverage plot to determine. However, in this project, the outliers are too less and not obvious enough that will have a significant impact on normality and variance homogeneity. It may be beneficial to remove potential outliers, but in this case, the project consider not to remove them for better effiency and interpretation.

# 6.2 Assumptions test based on quantiles

Based on the Shapiro-Wilk normality test result, we can get that the test result shows the evidence of normality violation. And we can get that the Our sample does not come from a normally distributed population. However, in this project the transformation of the data or other ways are not considered to be used for better interpretation and efficiency. Moreover, since we have large sample in this project, which by Central limit theorem, the original distribution of the data would attach negative influence on the analysis.

Based on the R result, we can see that the p-values from Levene's tests is higher than the significance level of 0.05. This indicates that there is evidence of significantly equal variances of error terms. As a result, we can infer that the variations in the different contrast levels and sessions are homogeneous.

Moreover, since the random effect has been discussed in the model fitting part, here we repeat the conclusion that the random effect does exist in the final model.

# 6.3 Alternative methods

To verify our conclusion, another methods are necessary to see if the results are different from the conclusion gained. Here we take two of variable as random effect, which is the mouse and session number. Since taking the mouse as random effect can help us to remove the individual differences across various experimental object. Moreover, if the conclusion still holds, the conclusion drawn would be more reliable.

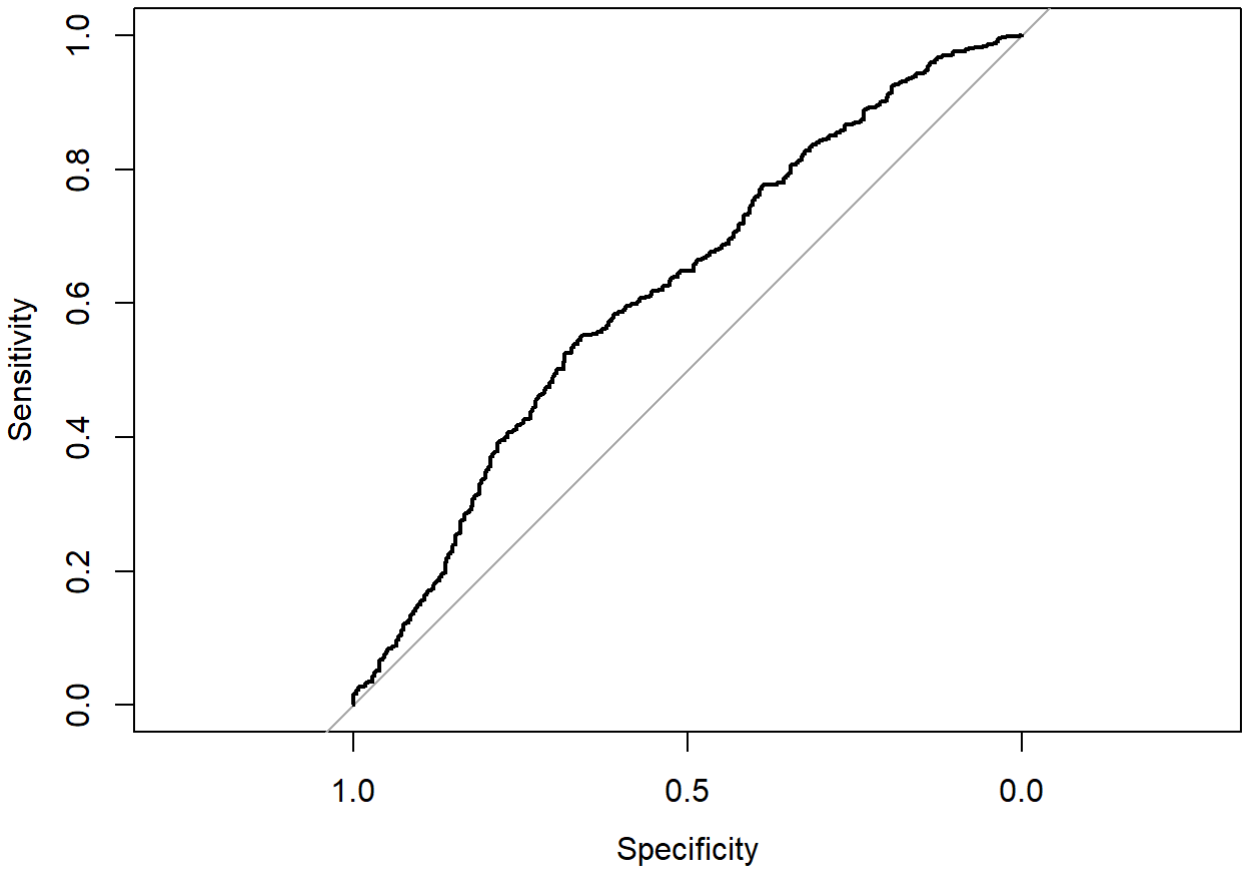
**Anova Table of alternative model**

| X. | Df | Sum.Sq | Mean.Sum.Sq | F.value |
|---|---|---|---|---|
| contrast_right | 3 | 14.4798 | 4.8266 | 12.0804 |
| contrast_left | 3 | 25.3314 | 8.4438 | 21.1338 |
| Interaction | 9 | 6.9568 | 0.7730 | 1.9347 |

Based on the R result above, we can see that all variables have the large F-value, indicating all variables would have statistically significant impact on the response variable, the mean firing rate. And that would help to cross-validate our result and conclude that the left and right contrast would attach influence on the neuron activities.

# 7 Prediction

For the second research interest, this project apply the logistic regression, which take the feedback type as the response variable and the firing rates and contrast levels as the explanatory variables. And here this project uses the ROC curve as the measures of Goodness-of-fit of this logistic regression.




Based on the ROC plot, we can get that the prediction model can be regarded as good, since the area below the curve is relatively bigger, which means higher prediction accuracy. Moreover, we can get the AUC value as the quantitative measure of the prediction accuracy. In this case, we can get the AUC value is 0.62, which is higher 0.5. And that verify our conclusion about the prediction that the prediction accuracy is good.

```
##         actual_values
## pvalues -1  1
##      -1  4  3
##      1  22 71
```

Based on the R result, we can get Sensitivity (True positive rate): the probability of a positive test result, conditioned on the individual truly being positive as below:

$$\frac{TP}{TP+FN} = 71/74 \approx 0.96$$

And the Specificity (True negative rate): the probability of a negative test result, conditioned on the individual truly being negative as below:

$$\frac{TN}{TN+FP} = 4/26 \approx 0.15$$

Based on the low Specificity, the prediction results can result in false positives, which means during the prediction, we may cause the mistake that gather the failure feedback into success feedback.

---

# 8 Discussion

# 8.1 Conclusion

After detailed analysis using ANOVA method and clustering methods as well as using the logistic regression as the prediction method, the report can generate the conclusion that there are significant differences in mean firing rate in different contrast levels and different session, which means different time and mouse, under the confidence level of 0.05, And the combination of contrast levels with high stimulation would generally have the relative higher mean firing rate compare to other situation. That means, with more stimuli, the neuron activity would be more active.

# 8.2 Analysis interpretation

At first, The report use the descriptive analysis to investigate whether there are differences between mean firing rate in different sessions. After that, the further analysis using two-way ANOVA method with random effect also proved that conclusion. Despite the violation of the basic assumption of ANOVA, the conclusion stil holds for the large sample size. Hence the conclusion can be correct taken for further analysis.

Generally, knowing the visual stimulation would attach influence on the neuron activity is very important in the the neuron areas, since such conclusion would help the researchers to set up subsequent and following research to investigate what happen to brains during the stimuli. Especially helpful for those who want to study the on the further analysis on human beings and related research.

# 8.3 Forward-looking and Caveats of the current analysis

By knowing the conclusion, the subsequent researchers can avoid complicated modelling and directly use the conclusion that the contrasts level would affect the neuron activities. Based on that, the related companies could set experiments aiming at helping to design new drugs and reduce the time of trails of new drugs. And it can also attach influence on understanding biological brains, which eventually help artificial intelligence to deal with the input such as visual stimuli. Overall, the conclusion of the project does help the related public reduce their work and become more inspired in such areas.

Despite the correctness of the conclusion, the method using in this report may need further analysis or fix to be more general. For example, using more data from different sessions. Moreover, the next editors or upcoming users of this report should keep in mind that the report does not use the transformation of the original data because the interpretation for that would be much more complicated and thus lead to misunderstanding of important explanations. And the clustering method can be improved to another level, such as using random forest or svm method to cross-validate the research. And the prediction can try more tools such as generalized linear model instead or using the decision tree to verify conclusion or get a more solid prediction result as well.

---

# Acknowledgement

“Yichu Chen”, “Shuang Wu”, “Dawei Wang”

---

# Reference

1.Steinmetz, N.A., Zatka-Haas, P., Carandini, M. et al. Distributed coding of choice, action and engagement across the mouse brain. Nature 576, 266–273 (2019).

2.Cisek, P. & Kalaska, J. F. Neural mechanisms forinteracting with a world full of action choices. Annu. Rev. Neurosci. 33, 269–298 (2010).

3. Jun, J. J. et al. Fully integrated silicon probes forhigh-density recording of neural activity. Nature 551, 232–236 (2017).

4. Allen W.E. et al. Thirst regulates motivated behaviorthrough modulation of brainwide neural population dynamics. Science. 364(6437):253 (2019) ***

# Session info

The detailed code of this project is posted on Github:

```
sessionInfo()
```