How neglect differentially affects sexes: a resilient phenotype or a hidden vulnerability?

Lucia D'Amore, Maurilio Menduni De Rossi

Statistical Learning and Large Data Module 1

Professor Francesca Chiaromonte

Abstract

This study explores the application of various data analysis techniques to behavioral experiment data collected from rats. Unsupervised learning techniques, including Principal Component Analysis (PCA), were employed to uncover underlying patterns in the data. Additionally, supervised learning methods such as Logistic Regression, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA) were utilized to predict the sex of the rats and whether or not they were neglected in their first years. Our analysis demonstrates that there are basal behavioural differences between sexes, while differences between neglected rats and others are present only in female rats. Further research is needed to understand if maltreated females develop a more "resilient" phenotype or if the better performances in the task hide a more subtle vulnerability, like differences in pharmacological treatment outcomes. It would be also important, in order to fully analyze all the longitudinal dataset, to repeat the analysis with functional data analysis techniques.

Contents

1 Introduction

In the following pages we will describe the project we came up with using the software R as means for applying some statistical techniques on data coming from an internship at the department of psychology at the university of Cambridge last summer. The statistical techniques applied vary, from some unsupervised analysis to simple supervised techniques like logistic regressions and linear discriminant analysis.

The aim of the project was to predict the sex of the rat and if it had been exposed to maternal separation early on in its life. We used these techniques to test if there were statistical differences between different groups and to identify the variables that contributed the most. Before delving into our work, let's start from an historical perspective to better understand the consequences of maltreatment on neurodevelopment and describe the experiments carried out last summer from which we collected our data.

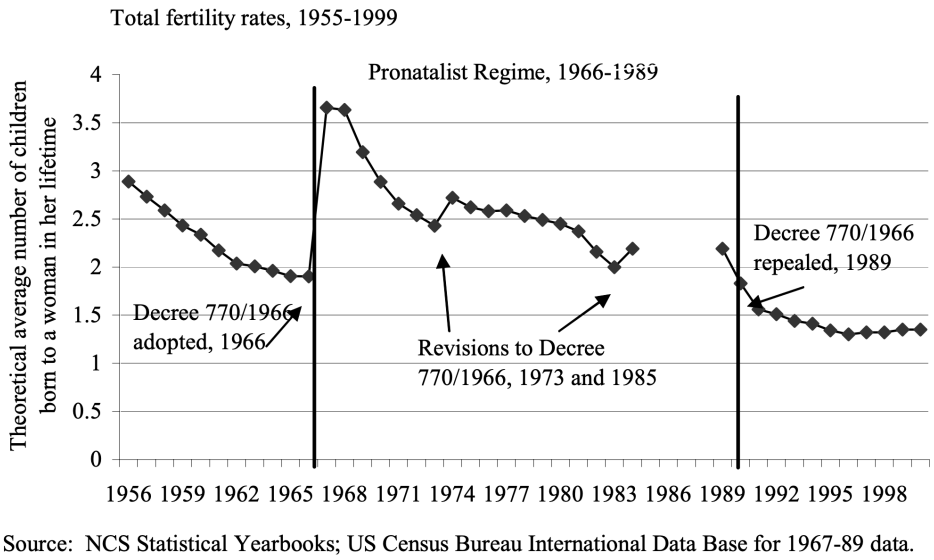


Figure 1: 1.0 Total fertility rates, 1955-1999; *NCS Statistical Yearbooks; US Census Bureau International Data Base for 1967-89 data[Gre06]*

1.1 An historical perspective

In 1966 Nicolae Ceausescu, the communist leader of the socialist republic of Romania, signed decree 770 one of the most coercive anti-abortion policy of the time [Wol01] . This decree represents the beginning of a twenty-year period in Romania (known as the Pronatalist Regime, till Ceaucescu's death in 1989) characterized by governmental pro-natalist strategies meant to increase the labor force in order to rescue the desperate Romanian economy see Fig. 1 [Gre06]. In these same years, governmental choices banned contraceptives, hindered divorce procedures, taxed childless couples and built a new model of child care, detached from the family, whose developmental necessities could be held by the state itself [Wol01]. In the first year after decree 770 the fertility rate of Romania jumped from 1.9 to 3.4, with thousands of children whose development couldn't be afforded by impoverished families [Tei72]. Many families decided to send their infants to state care, with the intention of reclaiming them once their children could be themselves a source of labour. Thousands of infant became "social orphans" who still had their parents but that lived and developed, in their most sensitive years, in state-run institutions. These institutions were underfunded and overcrowded, with very few caregivers (who had no formal education). Numerous studies followed these children longitudinally through their lives, as they were adopted abroad. Severe structural and functional deficits were found in multiple domains of cognitive and socio-emotional functioning, morphological, biological and electrophysiological measures [SFZ+12], [THQ+10], [CE97]. These studies brought plenty of evidence for the striking influence of caring or stressful environment as determinant respectively of better or worse mental and physical

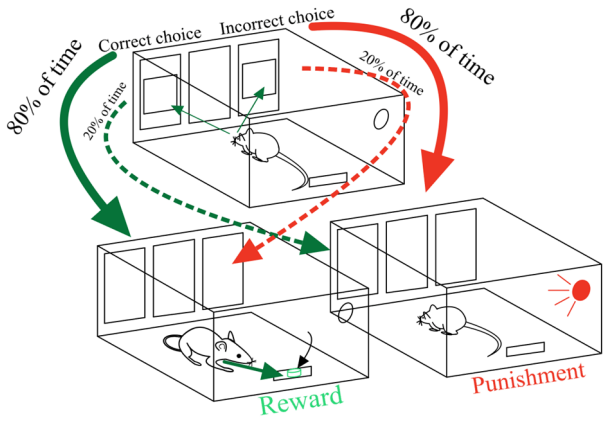


Figure 2: Touch-screen computer-based Spatial Probabilistic Reversal Learning Task

health later in life. This tragic event brought attention to the magnitude of the effects of the early-life stress exposure and sparked multiple lines of research in the field.

2 Dataset description

In the spatial probabilistic reversal learning task (see Figure 2) animals were first required to initiate their own trial by pressing the initiation stimulus presented in the centre window (not shown in fig.1) of the touchscreen before then responding to either a left or right spatial stimulus. There was no time cut-off for animals to start a trial, allowing animals to self-pace within a session. Upon selection of a stimulus by an animal they were rewarded depending upon the reward contingency of the selected option. One spatial stimulus was classed as the "correct" stimulus and was rewarded 80% of the time while the other stimulus had a reward probability of 20% and was called the "incorrect" stimulus. Once a stimulus choice was made animals either had to retrieve a reward pellet from the magazine and could then immediately start the next trial or if they were not rewarded were punished with a timeout of 5s and bright illumination from the house light. Omissions were classed as if animals did not make a stimulus choice within 10s and led to the same "punishment" as a lack of reward. Following 8 consecutive "correct" stimulus choices the contingencies switched so that the spatial location previously associated with the "rich" stimulus was now associated with the "incorrect" stimulus and vice-versa. Animals were allowed to reverse as many times as able within a session which lasted for a maximum of 200 trials or 40 minutes (which ever was reached first). The spatial location of the "correct" stimulus at the start of a session was consistent across sessions and counterbalanced across animals. The experiment involved 64 rats that were observed in 10 sessions before giving them any medicine. We thus have 10 rows for each rat, for a total of 640 initial rows. The rats are equally divided taking into account sex and being exposed or not to early maternal separation. In our dataset there are 17 initial columns, each measuring a specific characteristic of the behaviour of the rats in the experiment. We will deepen the comprehension of the variables that were significant in our analysis as they come up in the results. The last two columns are the two target variables we identified, sex and group (control or maternal separation).

We started with a data panel reflecting the longitudinal nature of the experiment, but we need a different type of dataset to apply the techniques described in the first module of the course. We hence proceed to summarize the info of the ten sessions of each rat in one row only. To do so, we have some different options. The proper ones are specific for longitudinal data, like mixed effect models (that take into account the specific effect each rat has), or for time series, like ARMA or ANIMA. The most complex one is functional data analysis, in which we take into account a little curve for each unit, measured in a discrete way and to which some smoothing techniques are to be applied.

We end up taking into account simpler techniques, taken from the summary of each feature and described in the following lines. We can fix a specific timing and cross check the data with the ones

coming from a different time frame. Another option is to identify some characteristics of the variables that could summarize the ten values we have for each rat. Some interesting statistics are mean in the time frame or the mean and variance in the time frame. We could also use median or two different quantiles to compare. We just need to identify the best statistics to use to summarize the ten rows in one or two numbers. To do so, we create two different datasets: one in which we use the mean of the values to summarize the data (mn) and the other one in which we use the difference between the last and the first session (dif). Dealing with information about how well they learn a task, we think that computing the difference between the first and last value is significant to catch any improvement in the ability of performing the task.
In the end, we found that the mean dataset had the best performance, consistently with our theoretical predictions. Therefore, we will mainly show the results we had with this dataset in all the techniques we apply.

3 Preprocessing steps

As for any type of dataset, we have to modify some columns to achieve a normal distribution and we also do some data wrangling to come up with better measures to work on. At first, we format all the values as numbers and we replace missing values with NA. We also replace categorical variables with dummies and plot the results, as shown in 3.
We can immediately see that some variables are not normally distributed, hence we proceed in applying

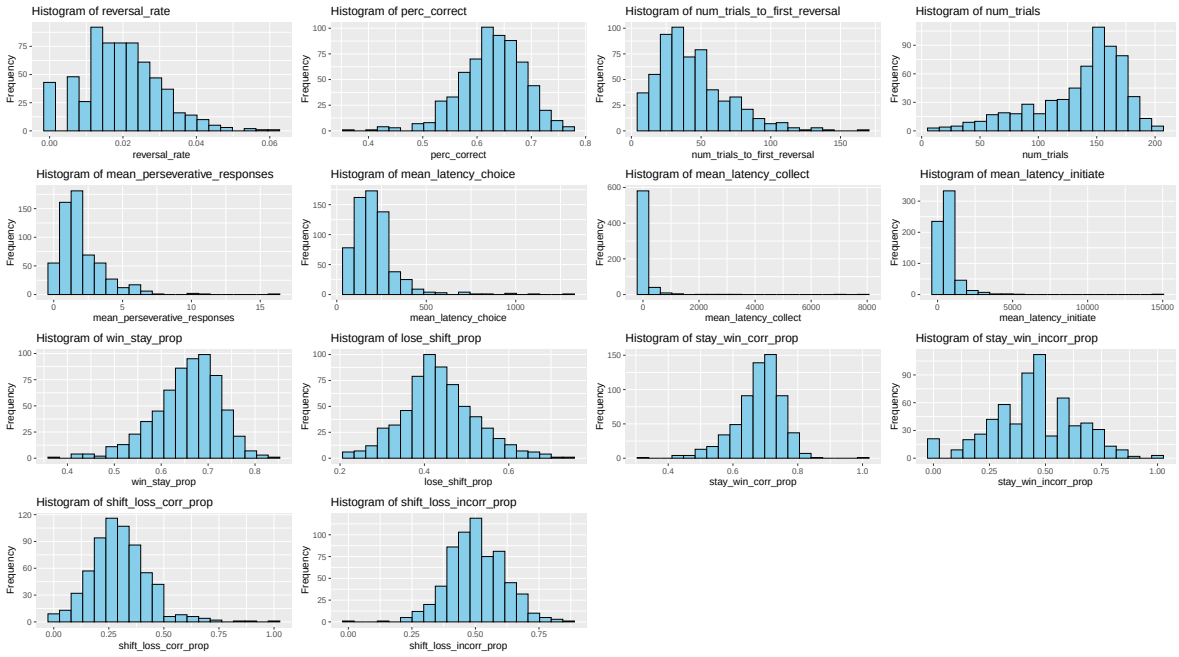


Figure 3: Initial distribution

some transformations. We apply the logarithm function to *mean_latency_choice*, *mean_latency_collect* and *mean_latency_initiate*, three important variables in our dataset, and we apply the square root transformation to *num_trials_to_first_reversal* and to *mean_perseverative_responses*. We will now briefly explain the meaning of those measures. *mean_latency_choice* measures the mean of the time waited before choosing which bottom to push, *mean_latency_collect* measures the time that is spent before collecting the reward/punishment and *mean_latency_initiate* measures the latency to initiate another trial in the same session. Those variables will be very important for the results, especially in supervised analysis.
We can see in 4 the results of the transformation. The variables are way more similar to a normal distribution, but especially in *log_mean_latency_collect* it looks like there are some big outliers that drive the distribution of the data. For further analysis, we could try removing them.

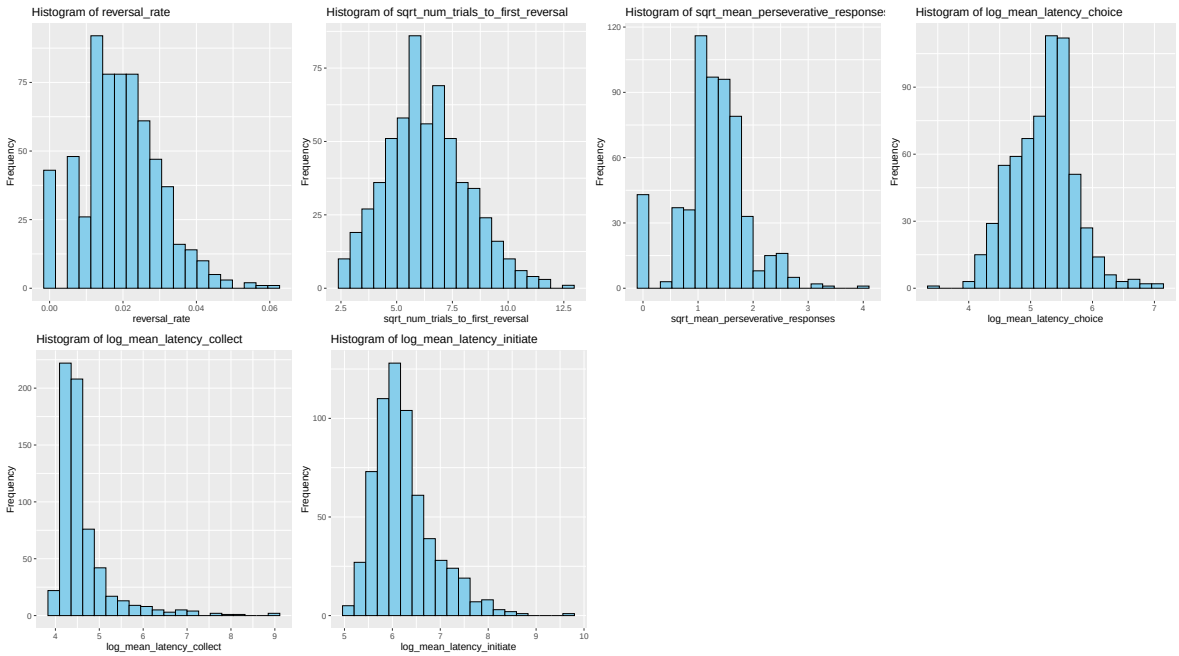


Figure 4: Columns after the transformations

4 Methods

We use the software R to compute some analysis on our data. In particular, we want to predict two binary categories, the group the rats belong to and their sex. To do so, we start by dividing the dataset: for unsupervised analysis we use the variables that encode specific behaviour patterns of the rats, without the target variables, whereas for the logistic regression we have four final dataset: two with the target variable *group*, two with the target variable *sex*; two with the mean dataset and two with the difference one. With Linear Discriminant Analysis and Quadratic Discriminant Analysis we can predict the two target variables together, ending up with four different situations to predict. Let's move on to the display of the actual results.

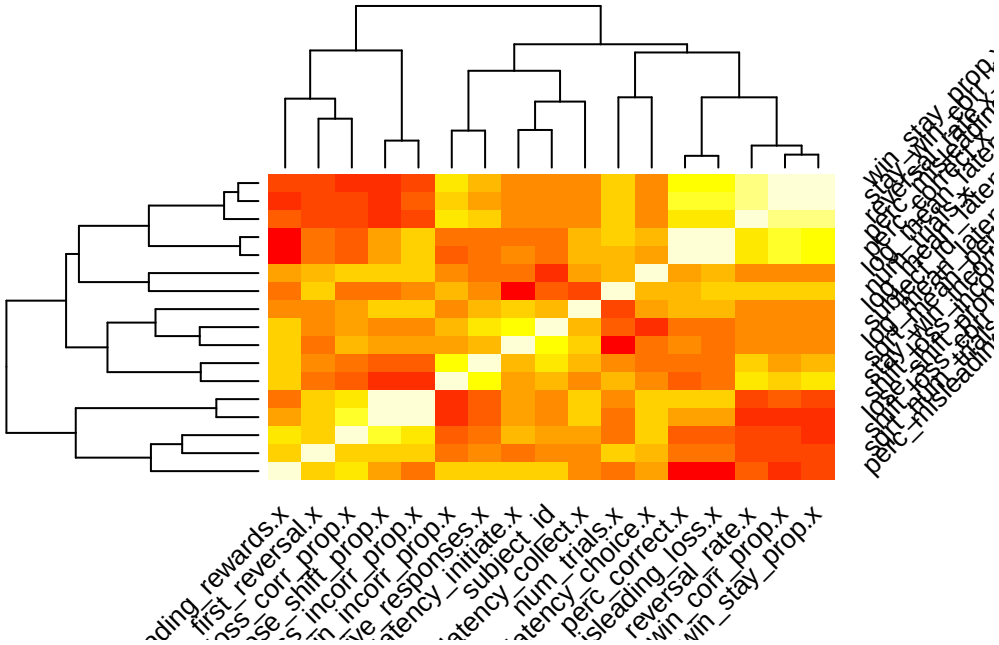


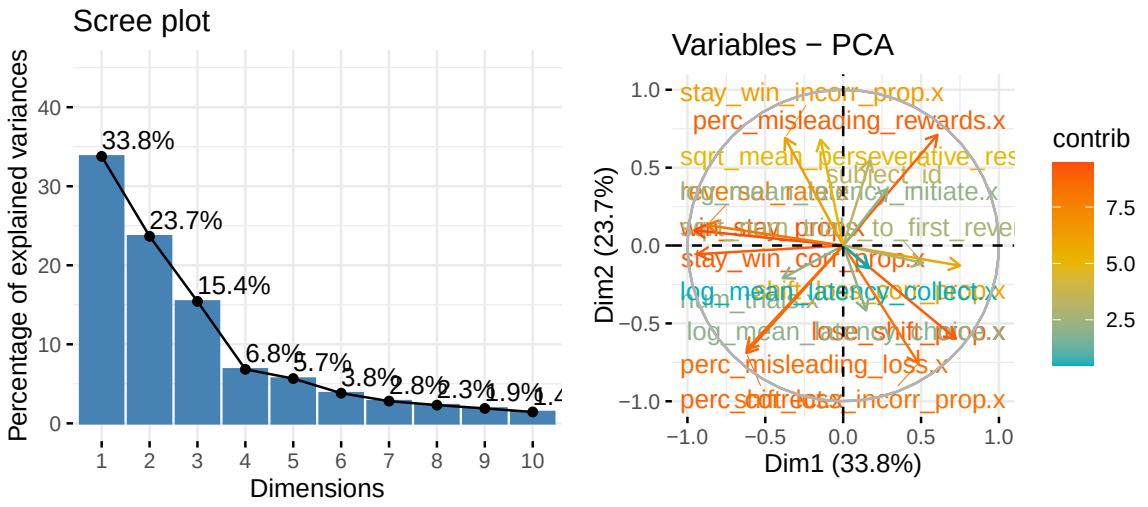
Figure 5: Correlation Plot

5 Some unsupervised analysis

After creating the dif dataset and the mean dataset as described before, we create two other datasets without the target variables to run some unsupervised analysis techniques. After scaling the data, we apply corrplot to see if there are interesting patterns in our columns by displaying the correlation matrix. The results of this are shown in 5.

We tried applying some basic clustering methods, but the results are not interesting. After that, Principal Component Analysis (PCA) helps us identify the first relevant characteristics of our data. PCA is a statistical technique used to reduce the dimensionality of a dataset while preserving as much variability as possible. This is achieved by transforming the original variables into a new set of variables called principal components. These components are orthogonal (uncorrelated) and ordered such that the first few retain most of the variation present in the original dataset. This technique is used to simplify data visualization, to reduce computational complexity and to remove noise. In 6a we can see the most significant principal components and the percentage of the variability of the dataset each one summarizes. We can clearly see an elbow in the fourth dimension: this means that after this dimension adding another one is not so profitable in terms of explained variability, still adding a lot of complexity to the model. Having to choose a number of dimensions to use, four is thus an optimal number.

In 6b we can see even more clearly that the first two dimensions explain more than half the variability of the dataset. This makes sense: we have not a lot of columns so it shouldn't be so difficult to sum up all the important info in not so many dimensions. The columns displayed in red are the most helpful in identifying the principal components: we can see that *stay_win_corr_prop.x* and *win_stay_prop.x* affect almost entirely dimension two when their value changes. The first shows the proportion of trials where the animal selected the same stimulus after they were rewarded on a TRUE correct. The second shows the proportion of all wins (rewarded trials) that were followed by the animal selecting the same stimulus. Hence it makes sense that they are really close in our graphical representation. Another interesting finding is the position of *perc_misleading_rewards.x* in comparison with that of *perc_misleading_loss.x*. They affect both variables with almost the same absolute value, but with opposite sign. It makes sense, given the meaning: they show the percentage of misleading loss or rewards collected. Lastly, we try plotting the position of the individual rats in the two dimensions (7) but we do not find any interesting cluster to highlight.


Scree plot
Percentage of explained variances
Dimensions
33.8%
23.7%
15.4%
6.8%
5.7%
3.8%
2.8%
2.3%
1.9%
Variables − PCA
Dim2 (23.7%)
Dim1 (33.8%)
stay_win_incorr_prop.x
perc_misleading_rewards.x
subject_id
perc_misleading_loss.x
contrib

(a) (b)

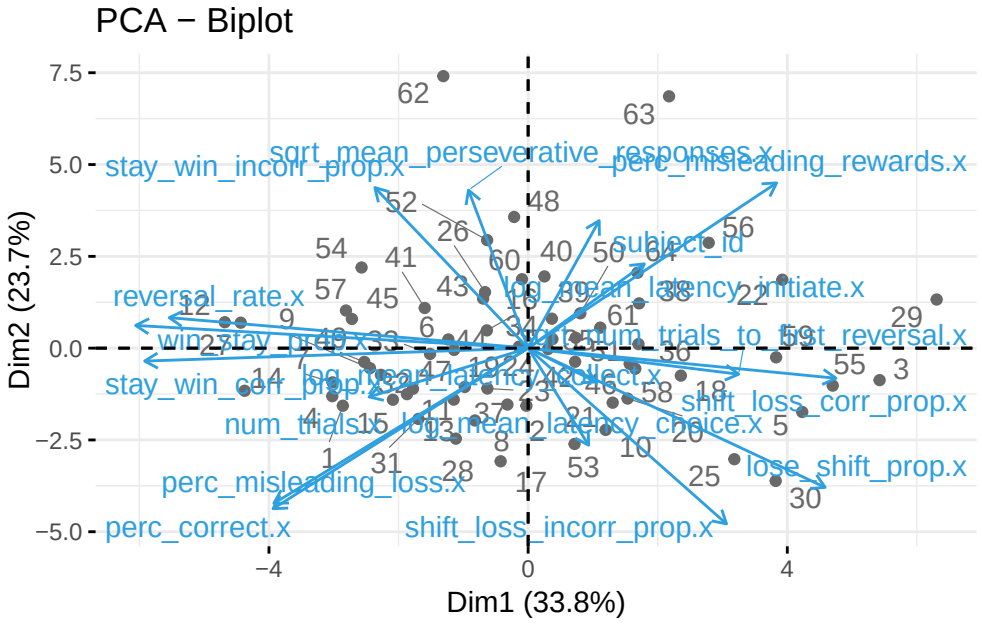

PCA − Biplot
Dim2 (23.7%)
Dim1 (33.8%)
stay_win_incorr_prop.x
sqrt_mean_perseverative_responses.x
perc_misleading_rewards.x
subject_id
reversal_rate.x
log_mean_latency_initiate.x
num_trials_to_first_reversal.x
shift_loss_corr_prop.x
num_trials.x
log_mean_latency_choice.x
lose_shift_prop.x
perc_misleading_loss.x
perc_correct.x
shift_loss_incorr_prop.x

Figure 7

```
$num_trials.x
Effect sizes were labelled following Chen's (2010) recommendations.

very small (Std. beta = 0.12, 95% CI [-0.54, 0.81])
medium (Std. beta = -1.59, 95% CI [-2.69, -0.75])
```

Figure 8: R results with num_trials

6 Supervised analysis

After some initial unsupervised techniques, we move on to the important part of the project: trying to predict the group and sex the rats belonged to. To do so, we use some of the techniques that we learnt in the course, the most suitable for working with this type of data. Initially, we apply the logistic regression technique to both datasets (mean and dif). To do so, we create four additional datasets: two with *sex* as the target variable and two with *group*. For every dataset we divide into train (p=0.8) and test set. We then create some empty list to fill with the results of the for loops we use to compute a different logistic regression for each variable. The formula changes dynamically, and the resulting plots are printed to the pdf file connected to it in the directory. We also store reports and effect sizes, to print if needed.

6.1 Logistic regression: predicting sex

Logistic regression is a statistical method used for modeling the probability of a binary outcome based on one or more predictor variables. It estimates the relationship between the dependent variable (which is categorical) and the independent variables by fitting a logistic function to the data. The output is a probability that ranges from 0 to 1, which can be converted into a binary outcome using a threshold (0.5 in our case). We show the results we had with the mn (mean) dataset and three variables, the ones that had a significant result.

We can see in 8, 10 and 12 that one of the results is significant and respectively negative, positive and negative. We can also visually see the difference in between the sexes, that shows that females (=1 in our experiment) complete more trials in total (9), wait less time before starting a new session (11) and have a bigger latency before choosing a stimulus after they are presented on the screen(11). The first is a rough measure of speed in completing the task, whereas if the last is bigger it may mean that the rat has low attention or is indecisive. For the simple logistic regression models predicting sex the k-fold cross-validation with k=5 had a mean error rate of 0.221, 0.125 and 0.231 respectively for num_trials, log_mean_latency_choice and log_mean_latency_initiate.

We try to predict the sex of the rat also with three types of logistic regressions: a simple one, that has as a regressor only *log_mean_latency_choice.x*; a complete one, that uses all the main predictors; a stepwise one, that iteratively adds or removes predictors to find the best-fitting model. We apply the technique in both directions (forward selection + backward elimination), adding and removing predictors as needed. In the end, the simple model ends up being both simpler and better at predicting, as shown by the Akaike Information Criterion (AIC), a measure used to compare and select models that assesses the relative quality of a model by balancing goodness of fit and model complexity.

6.2 Logistic regression: predicting group

We use the same methodologies to predict the group the rats belong to. Using all the data, there are no stastistically significant results. We then tried splitting the data into an only female dataset and an only male one. Running the for loop on those two dataset, we end up with more interesting results. In particular, we can effectively predict if the female rats are in the maternal separation group or in the control one with two variables: *num_trials.x* and *log_mean_latency_initiate.x*, that were significant also for predicting sex. As we can see in 15 and 16, given that the maternal separation group equals to 1, there is a small but statistically significant and positive effect with the first variable and a negative and statistically significant effect with the second one.

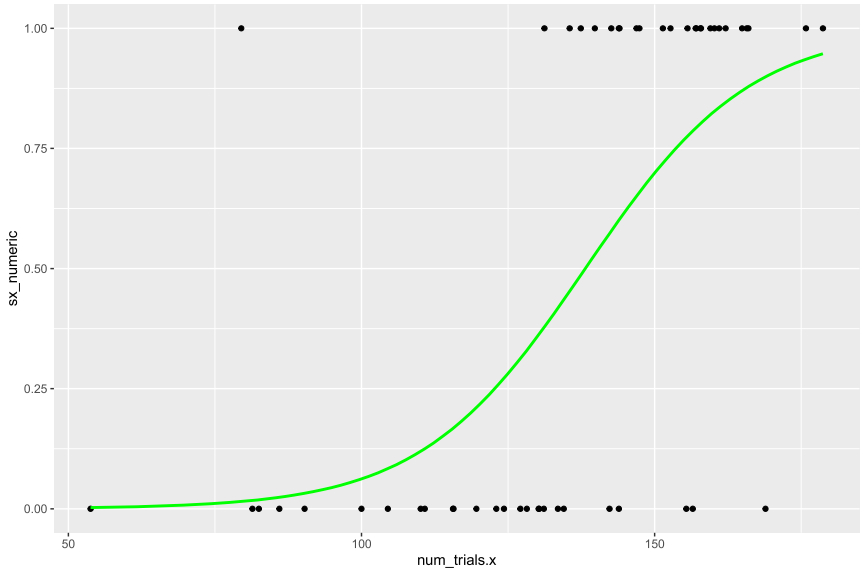


Figure 9: Logistic regression results with num_trials

```
$log_mean_latency_initiate.x
Effect sizes were labelled following Chen's (2010) recommendations.

very small (Std. beta = 0.03, 95% CI [-0.70, 0.78])
large (Std. beta = 2.04, 95% CI [1.11, 3.30])
```

Figure 10: R results with latency_initiate

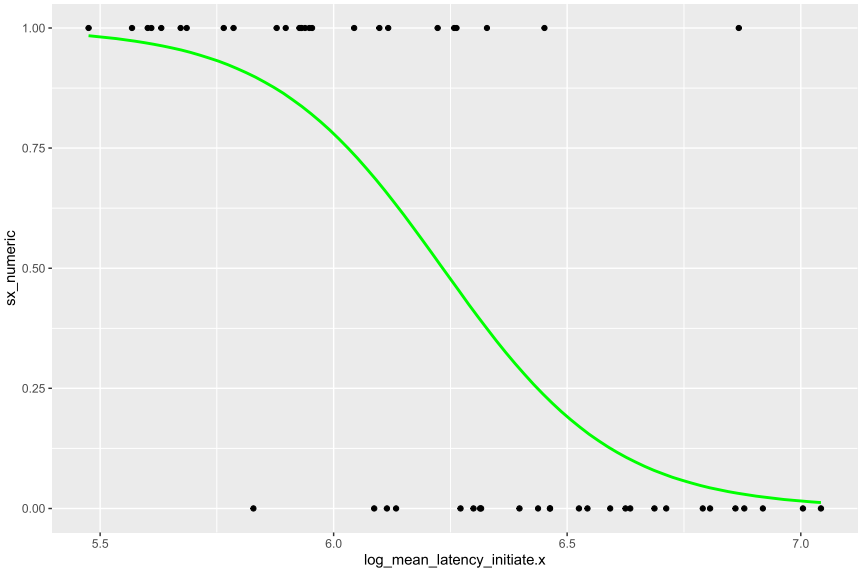


Figure 11: Logistic regression results with latency_initiate

```
$log_mean_latency_choice.x
Effect sizes were labelled following Chen's (2010) recommendations.

very small (Std. beta = 0.11, 95% CI [-0.72, 0.99])
large (Std. beta = -3.14, 95% CI [-5.17, -1.76])
```

Figure 12: R results with latency_choice

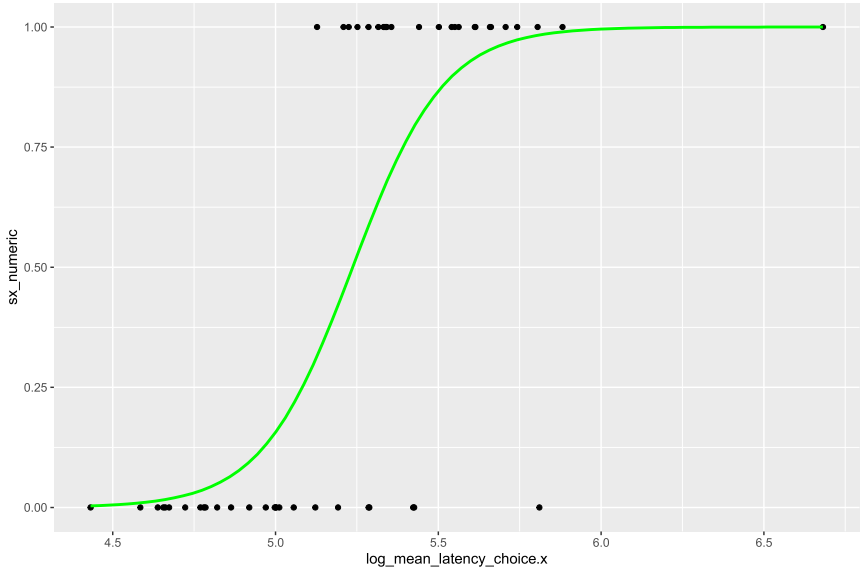


Figure 13: Logistic regression results with latency_choice

```
> AIC(simple_glm1, glm_complete1, glm_stepwise1)
                df       AIC
simple_glm1    2 40.16955
glm_complete1  3 14.48311
glm_stepwise1  3 14.48311
```

Figure 14: Comparison of simple, complete and stepwise regressions in predicting sex

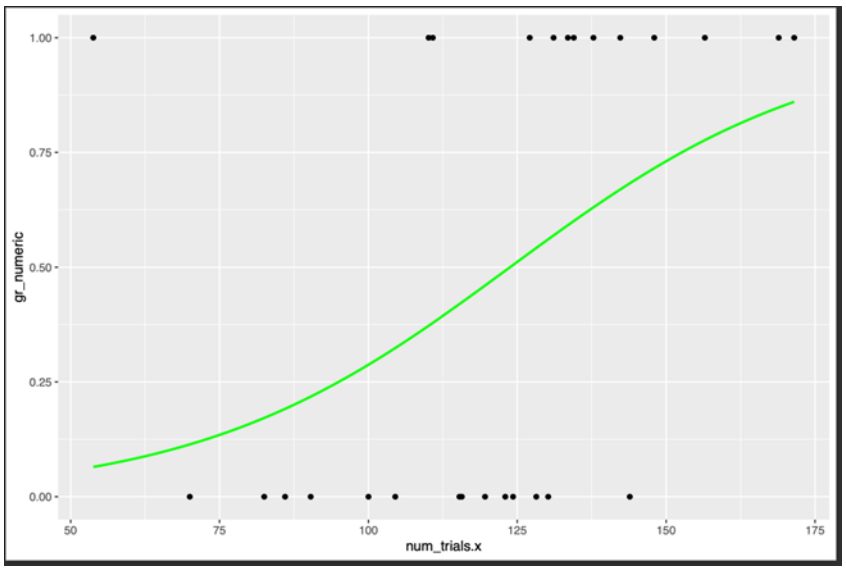


Figure 15: Logistic regression results with num_trials, predicting group

The next step was to run some logistic regressions without the for loop, to compare the different results. We use the same methodology as before: we have a simple linear regression, with *perc_correct.x* as a regressor, a complete one, with all the variables, and a stepwise one. To compute them, we use the dataset with males and females. The results show that the simple logistic regression is the best in terms of speed and predicting power. We can see in 17 that there seem to be a predicting effect connected with the *perc_correct* variable in the complete dataset (and also in the females and males only), but it is not statistically significant. Our interpretation is that it is mostly driven by outliers, so we did not take it into consideration.

6.3 Linear Discriminant analysis

Before delving into our work, we briefly explain what Linear Discriminant Analysis (LDA) is. It is a statistical method used for pattern recognition and machine learning, primarily for dimensionality reduction and classification. It aims to find the linear combinations of features that best separate two or more classes of objects or events. How does it work? LDA is in a way similar to PCA, projecting high-dimensional data onto a lower-dimensional space, but being a supervised learning technique it can use the target variable labels on the train set to divide the data in a more significant way, maximizing the separation between multiple classes. LDA works by finding the directions (linear discriminants) that maximize the ratio of the variance between different classes to the variance within the same class, to ensure that the classes are as distinct as possible in the projected space.

We use this technique repeating the process we did before, to predict distinctively the sex and the group of the rats, and in a compound way: we can set more than to classes, so we create a column with four different variables: female control, female maternal separation, male control, male maternal separation. Hence we try to predict this all, and 18 is the resulting plot in two dimensions. We can see that the fist dimension discriminates pretty effectively by sex, and the second divides the control group from the maternal separation one. We can conclude that using all the variables our model works pretty well.

6.4 Quadratic Discriminant Analysis

We also try Quadratic Discriminant Analysis, to see if non linear boundaries can be better for our purpose. Compared to LDA, it allows for more flexibility by accounting for different covariance structures in each class. As done before, we try to predict both the target variables, sex only and group only.

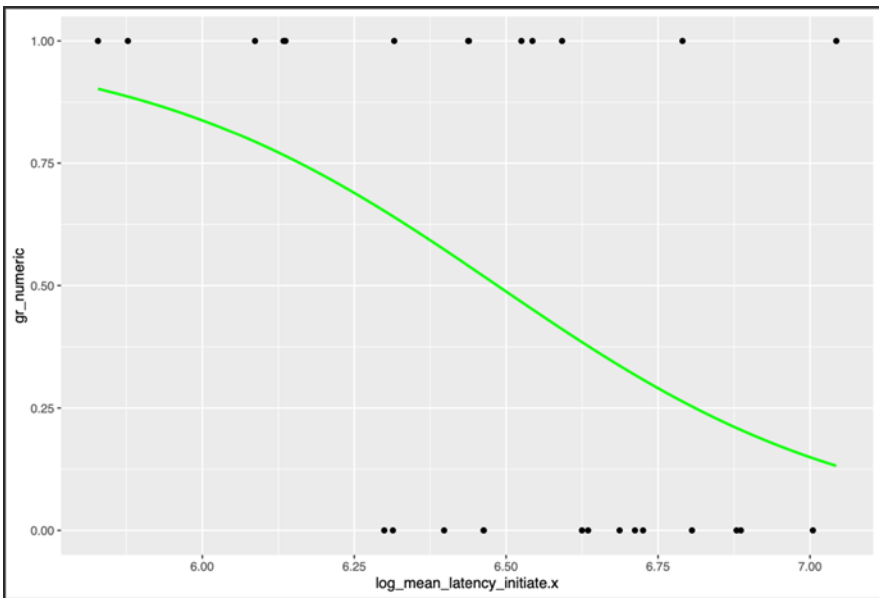


Figure 16: Logistic regression results with latency_initiate, predicting group

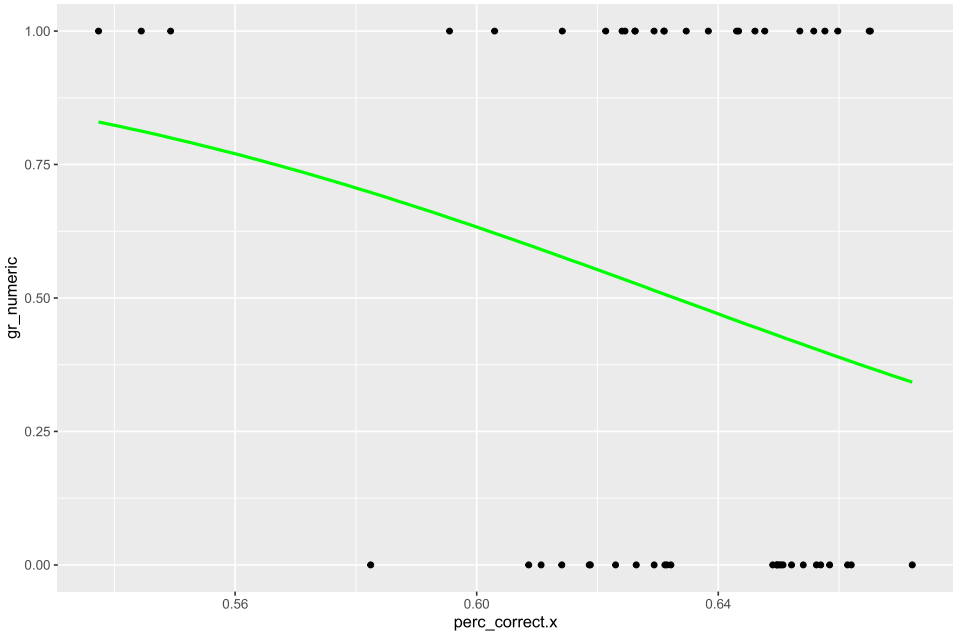


Figure 17: Graph for perc_correct, complete dataset, predicting group

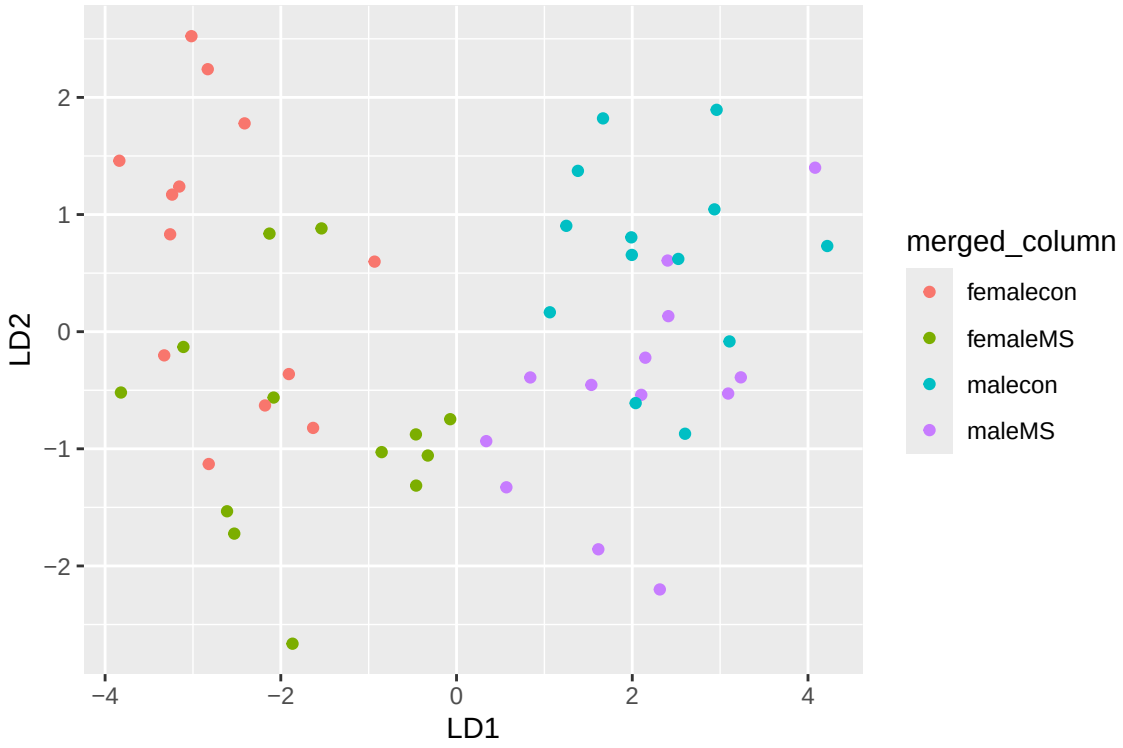


Figure 18: LDA results, predicting four categories

In 21 we can see the resulting scatterplot, after applying the qda function for predicting sex. In those results, the accuracy of the prediction of the sex of the rat using just *log_mean_latency_initiate* and *log_mean_latency_choice* is pretty good. As we can see in 19 (for the R script, we can see 20), only two individuals are not predicted correctly. This means that those two variables can be very effective in predicting this target variable.

In the scatterplot predicting group we find another interesting result, shown in 23. Here the Quadratic Discriminant Analysis ends up parting the space in a very interesting way. That's why we created the linear transposition of the plot (24), where we can see a statistical difference in the last part. We apply a linear regression between those two variables, and we find that even if in the first part the confidence intervals overlap, in the last part they do not. When the control group rats have a higher *Stay_win_incorr_prop*, they have a lower *Stay_win_corr_prop* than maternal separation ones. *Stay_win_incorr_prop* is the proportion of trials where the animal selected the same stimulus after they were rewarded on a FALSE correct. It is expected to be higher if they are more sensitive to reward. *Stay_win_corr_prop* is the proportion of trials where the animal selected the same stimulus after they were rewarded on a TRUE correct. This means that there is a linear relation linking the level of sensitivity to reward (stay_win_incorr_prop) and the correctness of the behavioural strategy in response to a chosen TRUE correct stimuls (stay_win_corr_prop). This linear relation is positive for the maltreated group (MS) and negative for the control group. The model used was: lm ¡- lm(stay_win_corr_prop.yštay_win_incorr_prop.y*group, data = df) and a significant interaction between group and stay_win_incorr_prop.y was found. The last QDA that we run was the one to predict both sex and group, that was not noticeably better than the LDA. We can still see the results in 25.

7 Conclusions and possible further developments

In this study, we apply a range of data analysis techniques to behavioral data from a rat experiment, aiming to predict the group and sex of the subjects. In the experiment, we collect some information about the way the rats respond to the task, and we use some supervised techniques to see if we can

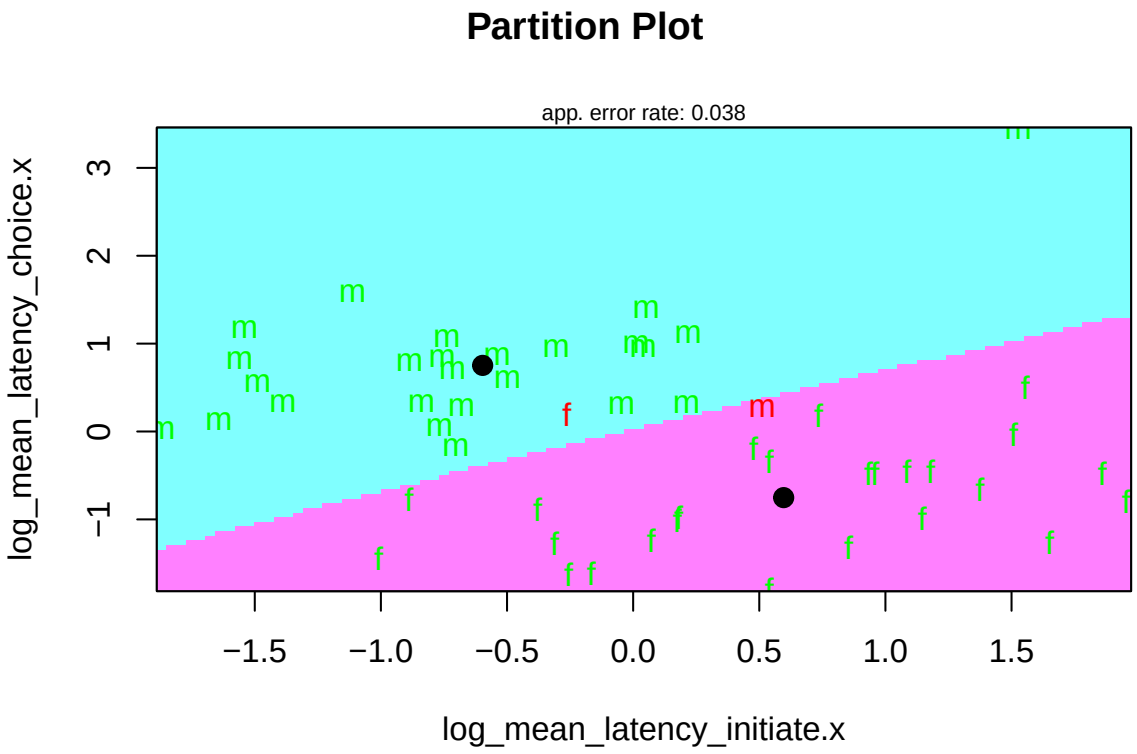


Figure 19: One of the QDA results, predicting sex

```
partimat(factor(sex) ~ log_mean_latency_choice.x + log_mean_latency_initiate.x,
         data=train_transformed3, method = "qda",
         col.correct='green', col.wrong='red')
```

Figure 20: Script of one of the QDA results, predicting sex

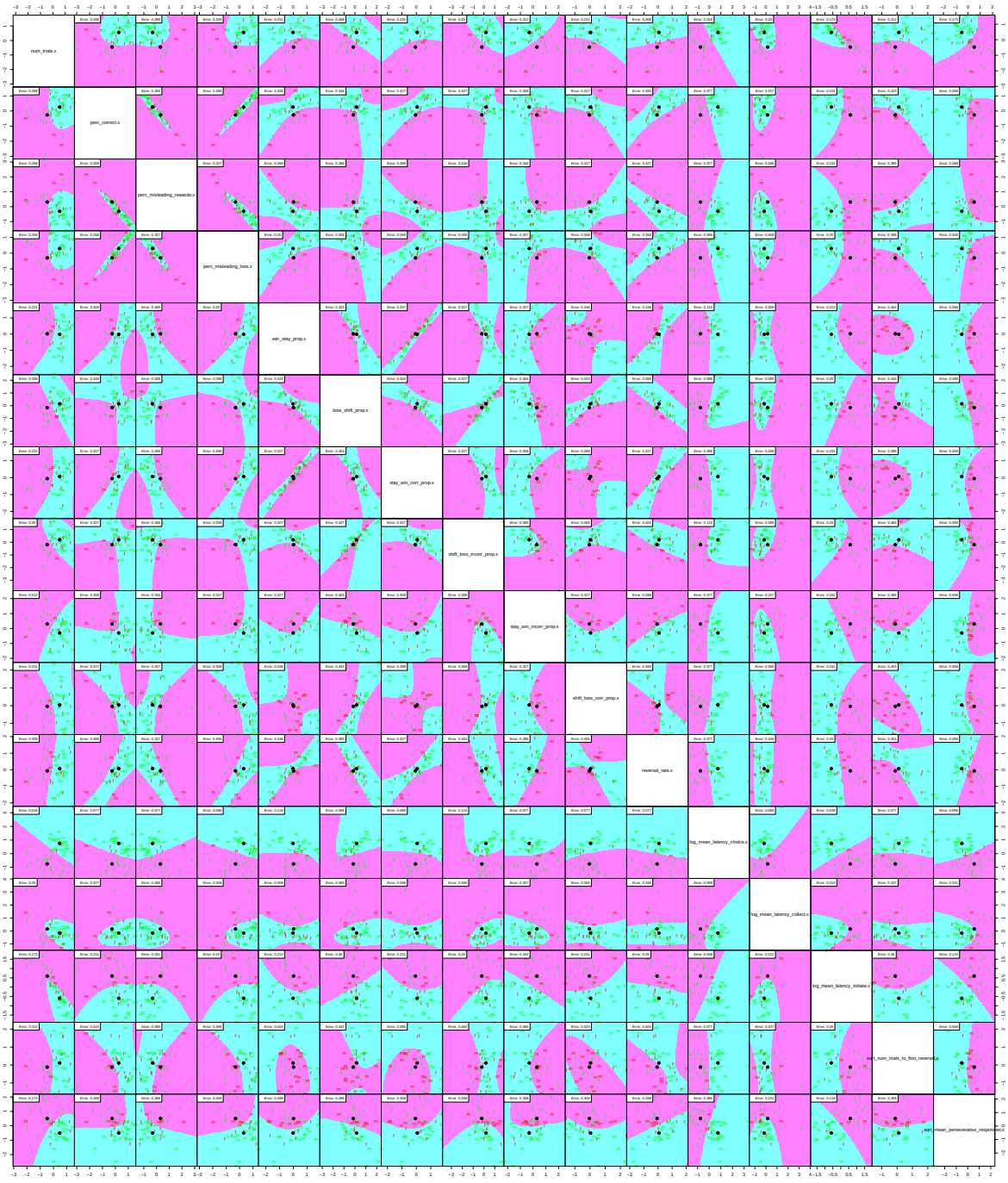


Figure 21: QDA results, predicting sex

```
qda3 <- qda(factor(sex)~  ., data=train_transformed3)
```

Figure 22: Script of QDA results, predicting sex

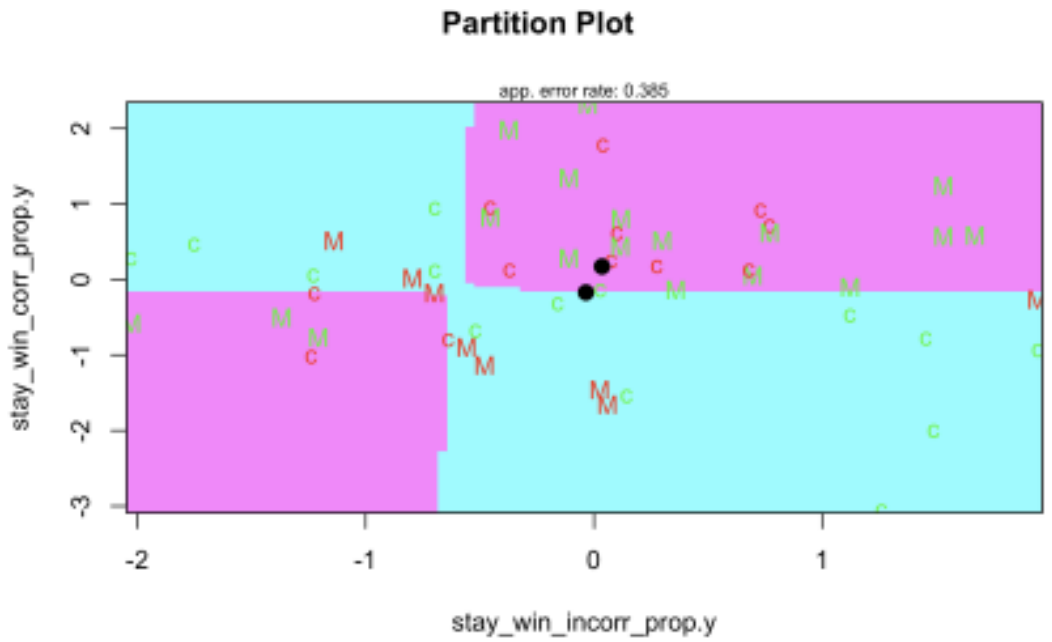


Figure 23

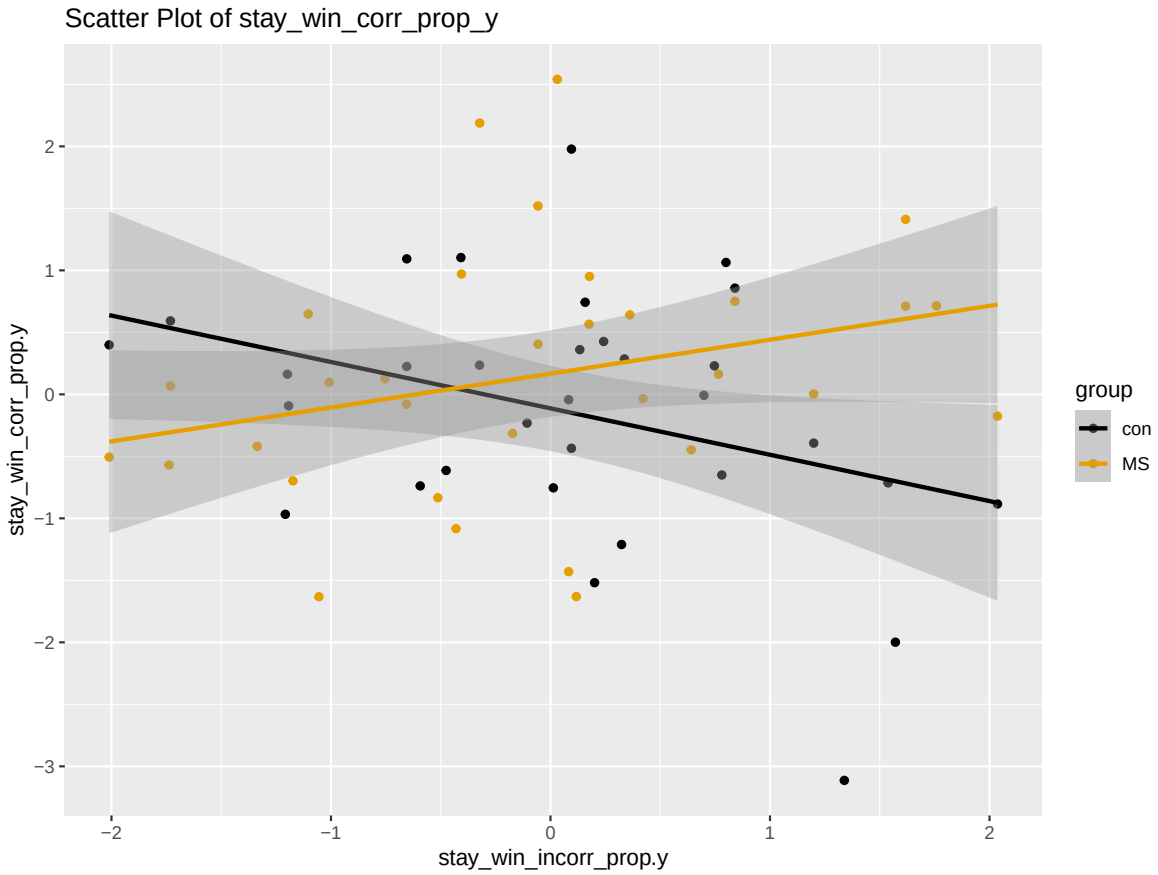


Figure 24: Linear trasposition of the stay_win_incorr_prop.y and stay_win_corr_prop.y relationship

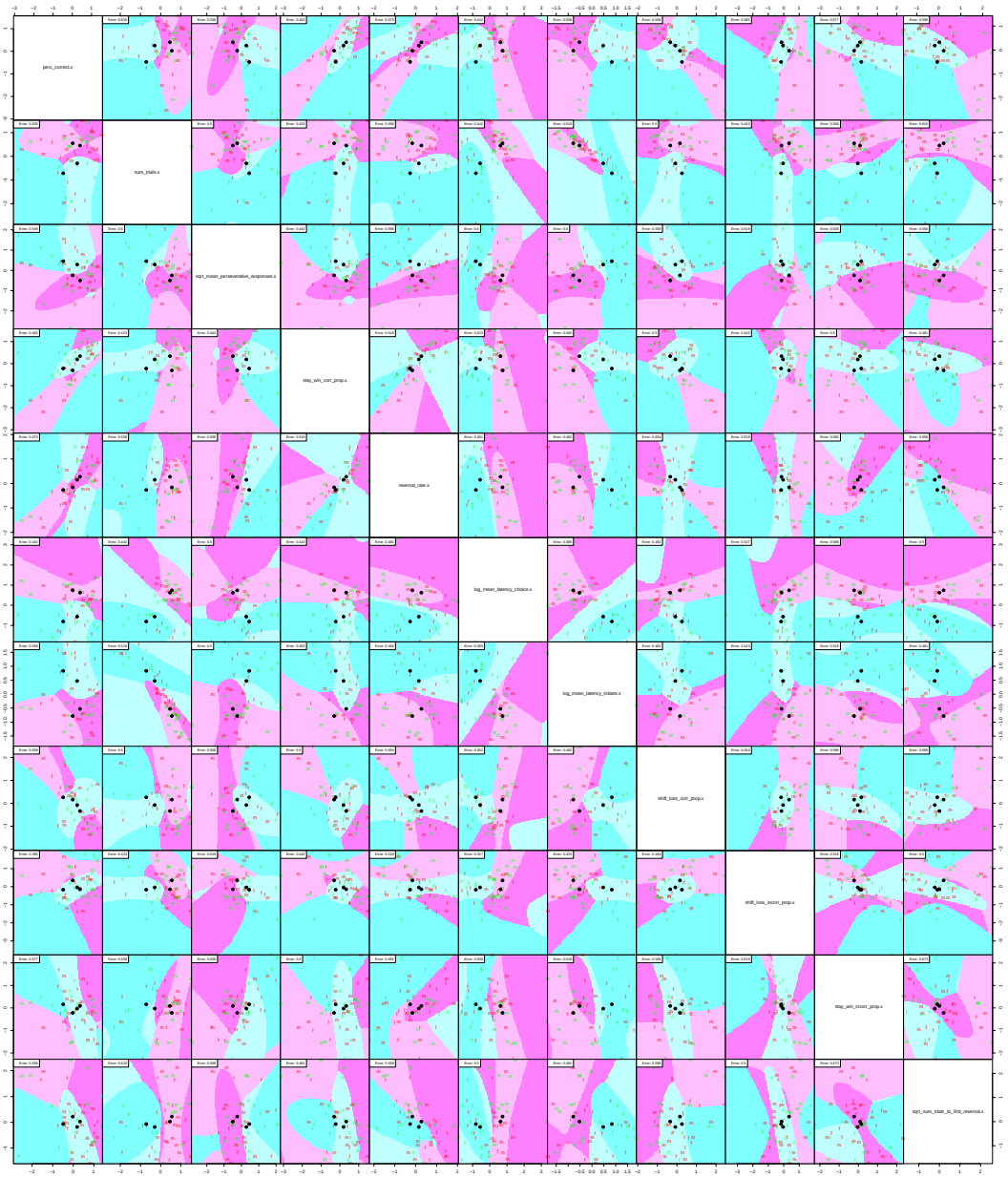


Figure 25: QDA results, predicting sex and group

correctly predict the sex of the rat and if they belong to the group that experienced maternal separation or not.

We find basal behavioural differences between sexes on the PRL task: females complete a higher number of trials in the ten sessions, wait more time before choosing the stimulus but have a lower latency to initiate a new trial after completing the previous one.
Trying to predict if the rats were neglected or not, we found statistically significant results only taking into account female only datasets. This might show a different developmental effect of neglect, that depends on sex. Female rats that were separated from their mothers show a higher responsiveness to stress and different performances compared to their control counterparts.
Lastly, maltreated females achieve higher scores in the test: this could lead to think that they are a more resilient phenotype, in the sense that they can learn faster what to do in behavioural experiments like this one, but this effect could hide a more "subtle" vulnerability.

Overall, this analysis highlights the effectiveness of combining unsupervised and supervised learning techniques to analyze complex behavioral datasets. PCA's role in simplifying the data without significant loss of information, coupled with the classification strengths of LDA and QDA, demonstrates a robust framework for understanding and predicting animal behavior. Future research could explore integrating these methods with more advanced machine learning techniques to further enhance predictive performance and uncover deeper insights into behavioral patterns; it would also be optimum to extend the analysis from a medical point of view. Firstly, it would be optimum to use all the data from the time series we have, using functional data analysis to process all the 640 the original dataset consists of. Secondly, it should be investigated if maltreated females are really that resilient or if there are differences in treatment outcomes and in their performances after pharmacological treatments. To do so, we need data on the following session of the experiment, that were performed after giving the rats different kinds of medicine. Lastly, the analysis could be extended investigating possible brain-wide alterations associated with maltreatment status, in addition to even more behavioural differences, analyzing MRI data.

References

[CE97] MARY CARLSON and FELTON EARLS. Psychological and neuroendocrinological sequelae of early social deprivation in institutionalized children in romania. *Annals of the New York Academy of Sciences*, 807:419–428, 1 1997.

[Gre06] Fern Greenwell. The impact of child welfare reform on child abandonment and deinstitutionalization, romania 1990-2000. *Annales de démographie historique*, 111:133, 2006.

[SFZ+12] Margaret A. Sheridan, Nathan A. Fox, Charles H. Zeanah, Katie A. McLaughlin, and Charles A. Nelson. Variation in neural development as a result of exposure to institutionalization early in childhood. *Proceedings of the National Academy of Sciences*, 109:12927–12932, 8 2012.

[Tei72] Michael S. Teitelbaum. Fertility effects of the abolition of legal abortion in romania. *Population Studies*, 26:405–417, 11 1972.

[THQ+10] Nim Tottenham, Todd A. Hare, Brian T. Quinn, Thomas W. McCarry, Marcella Nurse, Tara Gilhooly, Alexander Millner, Adriana Galvan, Matthew C. Davidson, Inge-Marie Eigsti, Kathleen M. Thomas, Peter J. Freed, Elizabeth S. Booma, Megan R. Gunnar, Margaret Altemus, Jane Aronson, and B.J. Casey. Prolonged institutional rearing is associated with atypically large amygdala volume and difficulties in emotion regulation. *Developmental Science*, 13:46–61, 1 2010.

[Wol01] Sharon L. Wolchik. The politics of duplicity: Controlling reproduction in ceausescu's romania. by gail kligman. berkeley: University of california press, 1998. 358p. 44.95. $American Political Science Review, 95 : 501 - -502, 62001$.