Part 1
Data Analysis Approach

Data Analysis from my perspective is the process which is involved in finding a solution to a real-world problem using data. Data Analysis plays an important role in understanding the data. In order to manipulate or use data for any purpose, the first step is to understand what the given data is about, and what each row or instance of the data is trying to convey.

Data Analysis consists of the following phases:

1. Defining problem statement:

This is the first step in data analysis which includes forming a question. What are we looking to resolve or find a solution to? The question should be concise. Having a clearly defined problem will make the later stages of data analysis easier, for example, In the EEG data analysis, we are interested in comparing patients with schizophrenia to control subjects, so as to find a trend between the corollary discharge process in the nervous system and the symptoms of a schizophrenic illness. With this in mind, we will know which features from the ERPdata.csv will answer our question better and which of those features are to be ignored.

2. Collecting data:

Now we have our problem statement in place, so next, we need to collect data which will answer our problem statement. In real-world scenarios the data very rarely comes from just one source, we need to look for various sources and come up with a methodology to merge data from different sources. More data, especially data from more diverse sources will enable us finding better correlations and finding more actionable insights. For this project, I was given the ERPdata.csv file which contains the averaged, ERP time series for all subjects, conditions, and the 9 electrodes Fz, FCz, Cz, FC3, FC4, C3, C4, CP3, CP4.

3. Data Pre-Processing:

Data pre-processing ensures the data is in proper shape before we can start analyzing it. It includes a variety of processes such as handling missing values, formatting data, Identifying outliers, Identifying errors in data and removing unwanted features if necessary. I've mentioned a brief explanation for each of these pre-processing techniques that can be used, in the context of this project.

- **Handling missing data**

The ERP readings of the electrodes are continuous, so imputing missing values with the median would be a good idea (I could also use mean but outliers have a significant impact on the mean so median is better in this case).

- **Formatting Data**

The ERP readings of electrodes were rounded off to a different decimal value for each row, I made sure that all the ERP readings are rounded off to the same decimal value to maintain consistency.

- **Identifying errors and unusual outliers**

The ERP readings typically ranged from -10 to +10, any values which are significantly higher (for example 117) can be considered as an outlier and can be omitted or imputed with another value.

4.Create Visualizations:

Data can be manipulated in a number of different ways, such as plotting it out and finding correlations between variables, slicing and dicing the data, creating various graphs to understand the way data behaves, etc. As we manipulate data, we start to figure out if we have the proper data we need to answer our problem, else we might need to revise our original question or collect more data.

Once the data was ready and available in proper shape, I started creating data visualization to best depict the data, which I've included in Part 2- EEG Data Analysis

5. Identify interesting observations:

The underlying idea of creating visualizations is to see if there is a trend or a pattern to the data. It is of paramount importance to identify interesting observations from the visuals. Though we identify a trend, we cannot directly conclude that our findings will

hold true for every scenario, so we need to come up with a hypothesis and test the hypothesis to validate our findings. Once I was done with data visualization, I started playing with the graphs to get insights from it. The outcomes of my observation are presented in Part 2- EEG Data Analysis.

6.Interpreting Results:

After analyzing the data and conducting a good number of hypothesis testing for different scenarios, it's time to interpret the results. While Interpreting the results it's important to ask a few questions to make sure the results are providing value.

- Do the results answer the problem statement? How?
- Does the data help defend against any objections? How?
- Are there any limitations on the conclusions, any scenarios that aren't considered or missed?

If the interpretation of the data holds up under all of these questions and considerations, it's likely to have come to a productive conclusion. The only remaining step was to use the results of the data analysis process to decide the best course of action. This brings to the second part of my analysis.

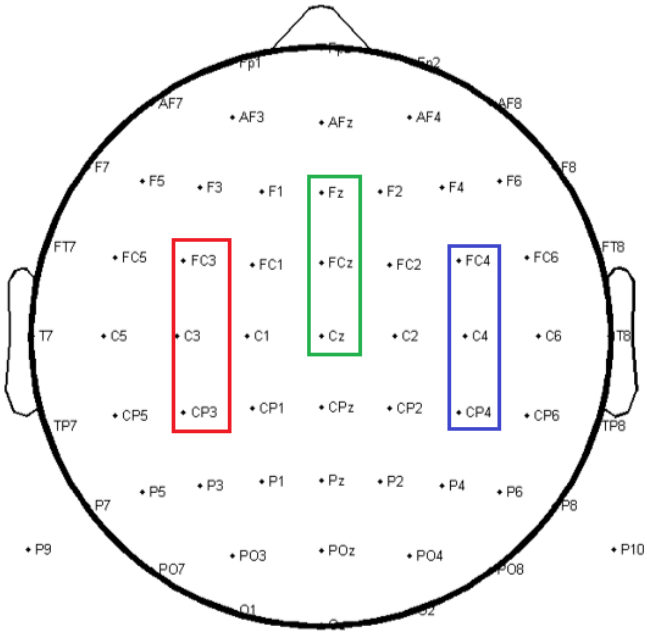
Part 2
EEG Data Analysis

Understanding the data set:

This data set contains the ERP readings for 9 electrodes.
(FC3,C3,CP3,Fz,FCz,Cz,FC4,C4,CP4) for each subject and condition. There are a total of 81 subjects and 3 conditions. The ERP readings for the electrodes are continuous variables which typically ranged from -10 to +10.
The condition variable is categorical. The conditions are events, each condition represents a certain event. Condition1 = button press + tones, Condition 2 = playback tones, Condition 3 = control presses.

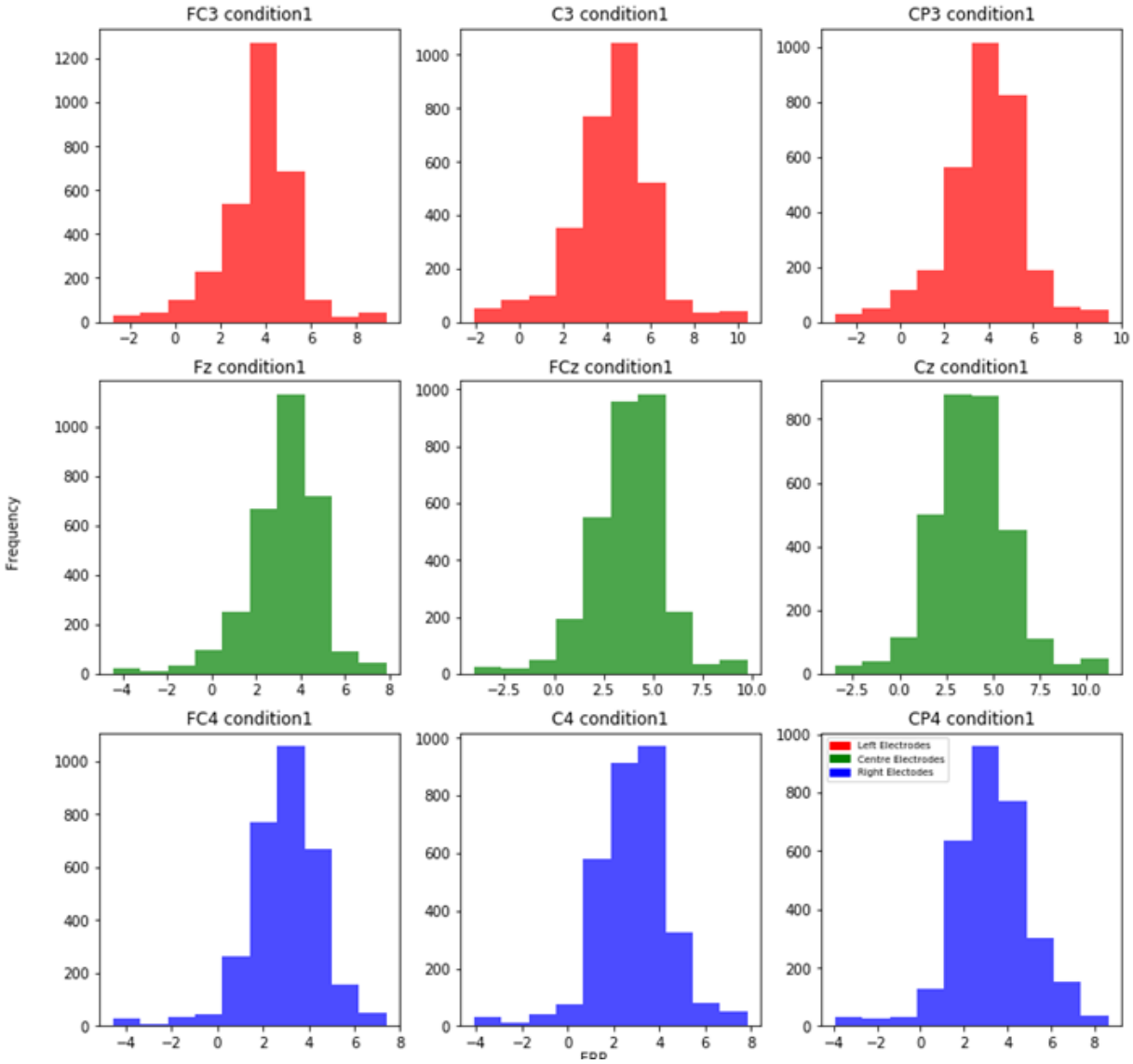
Analysis method:

The analysis I would like to present is a comparison between the different electrodes against conditions and time. Firstly to start with, we can see that the 9 electrodes in the data set are divided positionally.



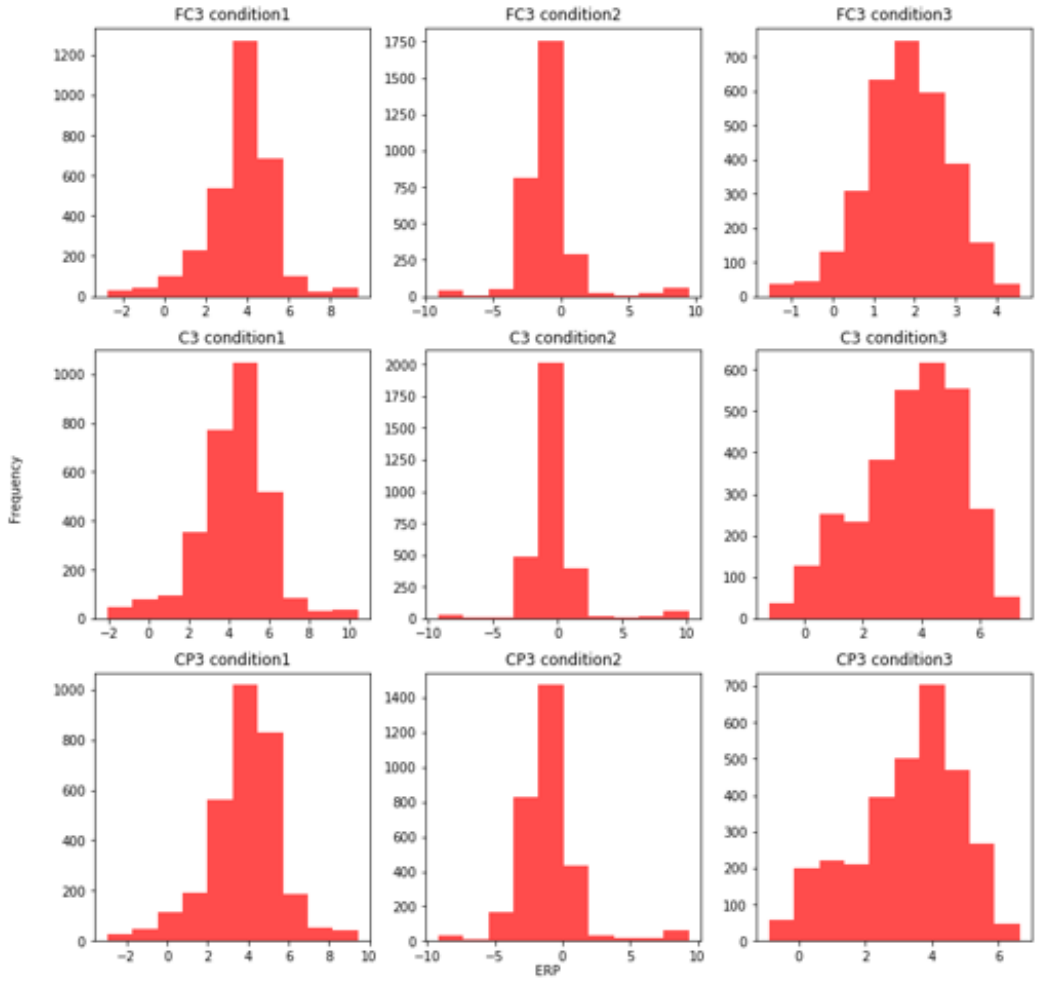
It can be observed that out of the 9 electrodes FC3, C3 & CP3 are towards the left and Fz, FCz & Cz are located in the center and FC4, C4 & CP4 are towards the right. So we can assume that electrodes highlighted in red represent the left hemisphere and electrodes in blue represent the right hemisphere of the brain. Going forward I would like to use the color coding Red,Green & Blue to Visualize the data for left,Centre and Right electrodes.

1. ERP frequency of electrodes (Histograms)



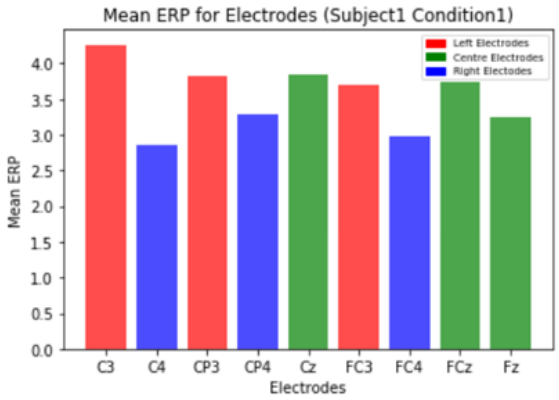
It can be observed that the for the electrodes on the left (red) ERP ranging between 2 to 6 has the highest frequency, also the number of observations having an ERP below zero were less in the left electrodes compared when compared to the center(green) and right(Blue) electrodes. Electrodes in the center tend to have a more uniform distribution compared to the others.

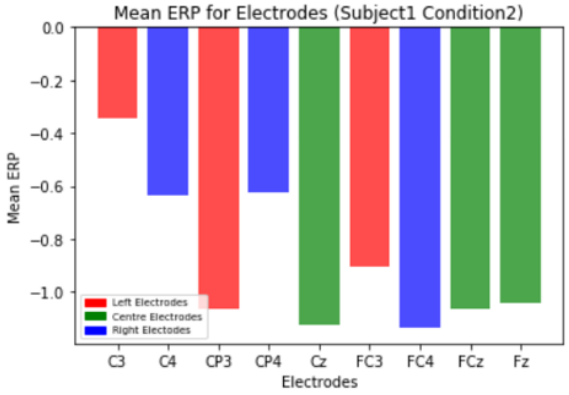
2. ERP Frequency against different conditions

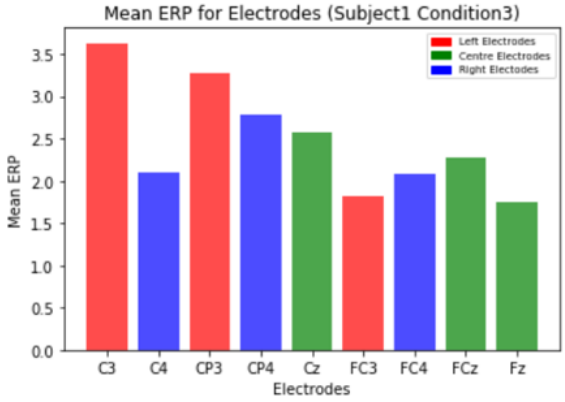


For the electrodes, FC3, C3 & CP3 in condition1 and condition 3, most ERP values are positive and a very small percentage (around 3%) is less than zero. However, the scenario is completely opposite when it comes to condition 2, about 90% of the observations have a an ERP below zero. Compared to other conditions, Condition 1 seems to have a more uniform distribution.

3. Mean ERP of electrodes against different conditions (Bar charts)



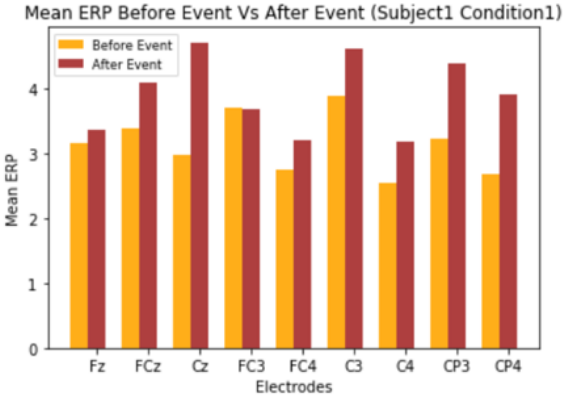


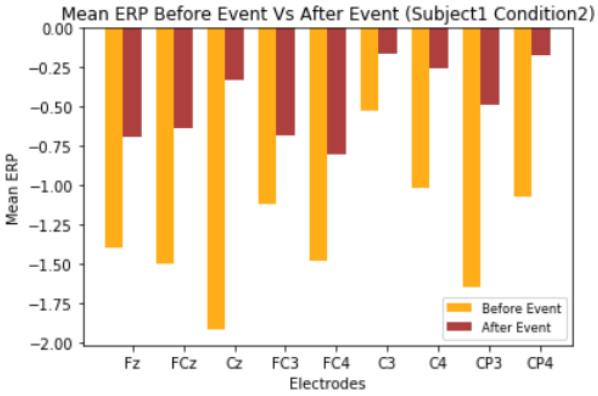


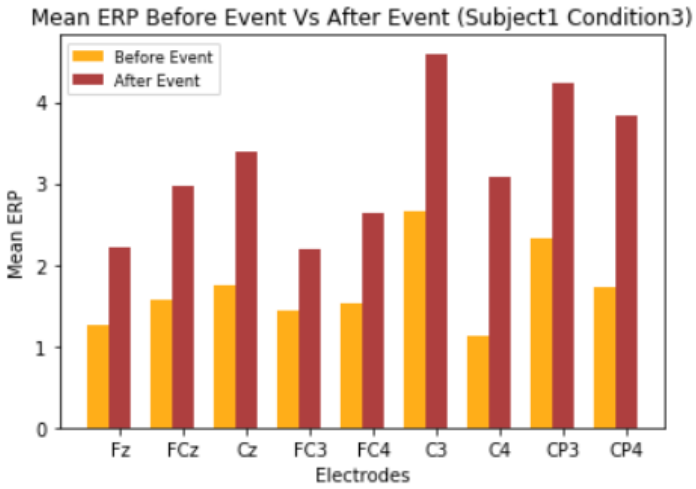
At a glance, it can be cleared noticed that for Condition 1 and 3, electrodes towards the left have the highest average ERP. It should be noticed that for condition 2 the ERP is in a negative scale, the highest average ERP is -0.3 for electrode C3. Electrode C3 has the highest mean ERP across all conditions, while the least average ERP varied across different conditions.

4. Mean ERP Before event Vs After the event

In the above diagrams, we have observed how mean ERP varies for different conditions. It would be interesting to note how the mean ERP would behave before and after the occurrence of an event for different conditions.



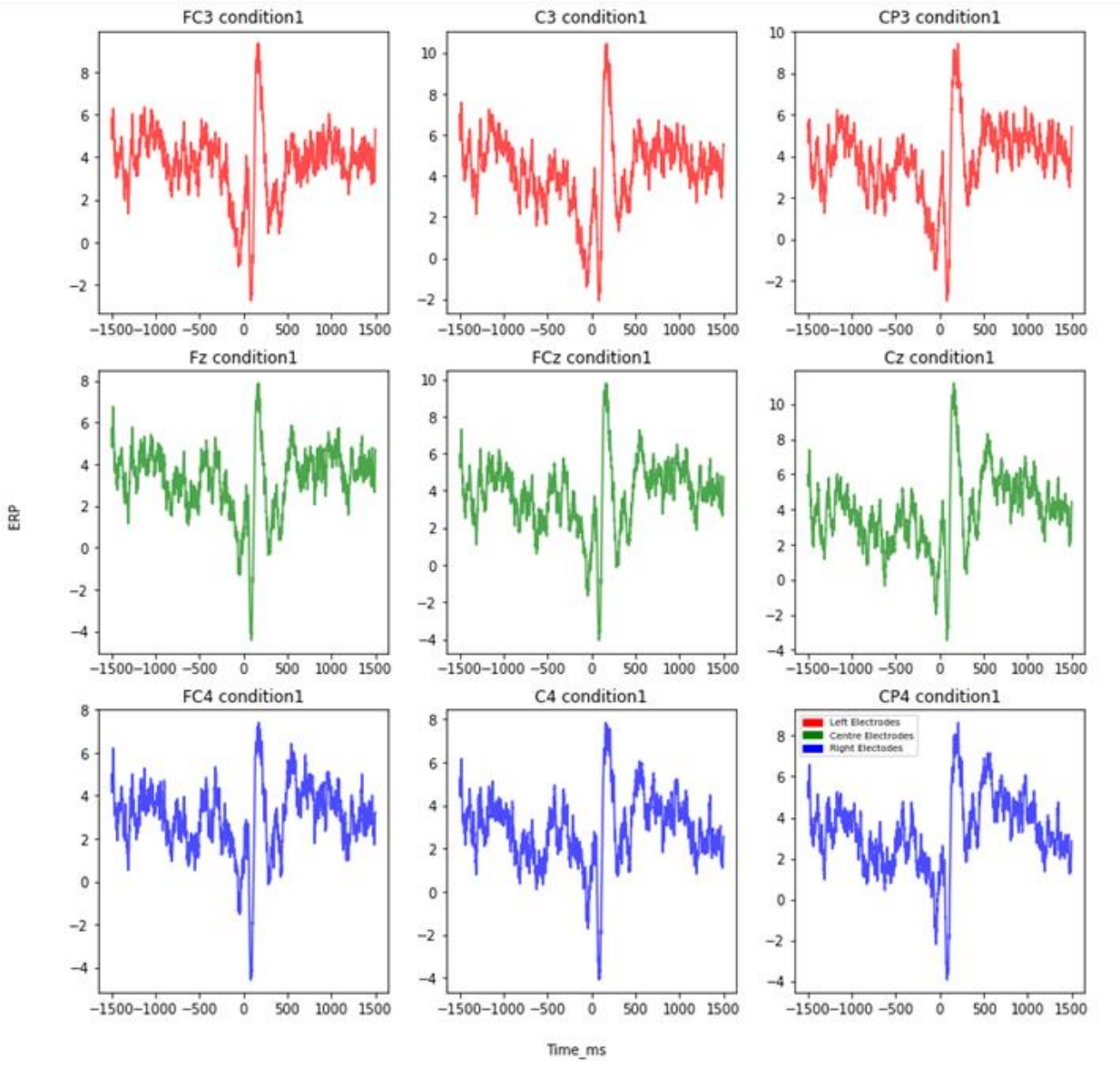




It is evident that the Mean ERP for all the electrodes, across all the conditions, is higher after the occurrence of an event. The percentage change between ‘mean ERP before event occurrence’ and ‘mean ERP after event occurrence’ is the highest in condition 2. Almost all the electrodes have doubled the mean ERP after the occurrence of the event in condition 2.

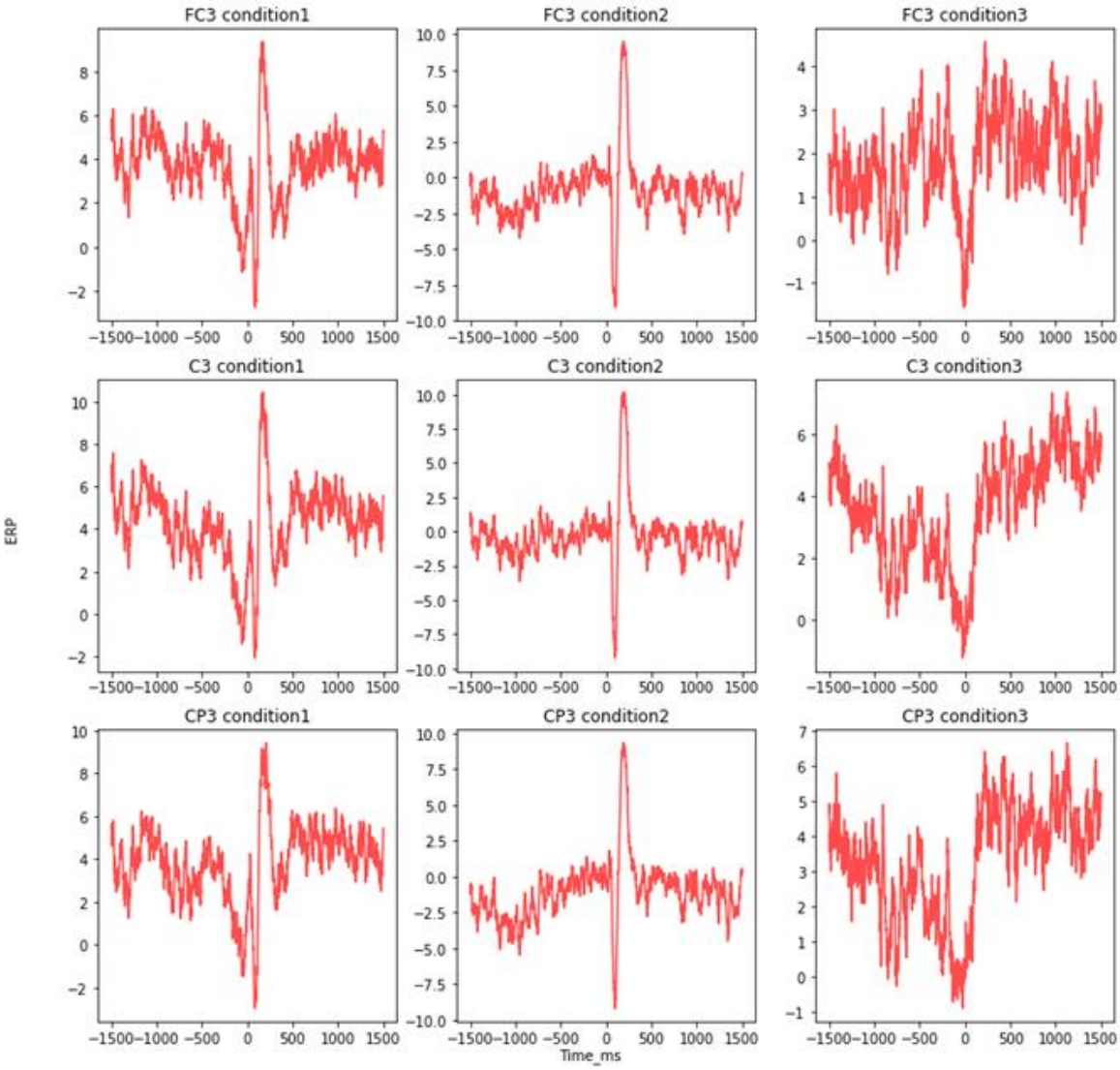
There is a drastic change in the ERP for the Cz electrode in condition 2, the mean ERP has been increased around 4 times over, after the event occurrence. Condition 1 seems to have had a lesser impact compared to the other conditions.

5. ERP vs Time for all electrodes



This diagram describes the fluctuation of ERP for all the electrodes across the time interval, -1500ms to 0 is the time interval before the occurrence if the event and 0 to +1500ms is the time interval after the occurrence of the event. An interesting observation which can be noted is that, in all the electrodes as soon as the event starts there is a sudden dip in the ERP and very quickly it bounces back and the ERP reaches its all-time high. For all the electrodes the highest and the lowest ERP was observed between 0 to 500ms.

6. ERP vs Time across different conditions



For condition 1 and 3, the ERP seems to fluctuate at a significantly higher level after the occurrence of the event. In condition 3, there is a decreasing trend of ERP from -1500ms to 0 and an increasing trend from 0 to +1500ms. However, In condition 2 there just seems to be a sudden dip followed by a spike in ERP as soon as the event occurs but later it seems to get normal. Overall, condition 3 seems to have a huge impact on the ERP after the occurrence of the event.