Parte I: Exploración de la base de datos.

- **Descripción de las características generales del dataset:**

El dataset original consta de 659112 filas y 11 columnas pero para el análisis se parte de un dataset el cual ha sido filtrado y que está compuesto por 7 registros que corresponden a cuatro pacientes distintos (3 registros para el paciente AA, uno para el HA, dos para el JA y uno para el MA), los cuales fueron unificados en único dataset, que consta de 7 columnas (ya que se ha adicionado una columna de referencia al sujeto, y una columna de secuencia temporal, y se han eliminado las columnas no relevantes) y 248983 filas. Definimos utilizar todos los datos juntos en un solo dataset para comodidad en el análisis.

- **Analizar las columnas presentes en el dataset:**

No todas las columnas son relevantes para el análisis, por lo que se decidió seleccionar aquellas columnas que van a ser utilizadas para tal fin, las cuales incluye las columnas: 'SampleIndex', 'Canal 1', 'Canal 2', 'Canal 3', 'Canal 4' y 'Stimulus', y el resto fueron descartadas. La información que nos brindan estas columnas es:

1. SampleIndex: Es el índice de las muestras.
2. Canal 1, 2, 3 y 4: Son las señales captadas por cada canal, las cuales estan expresadas en microvoltios.
3. Stimulus: Son las etiquetas que corresponden a cada estímulo:
    - **99**: NaN (default value).
    - **1**: looking left.
    - **2**: looking right.

La variable Stimulus puede tomar los valores ya mencionados:

- **99**: La cual es información de ruido, momentos en los que el paciente no estaba observando ningún estímulo, por lo que estos satos son irrelevantes para el análisis y posee 432713.
- **1**: La cual corresponde a momentos en el que paciente está observando al estímulo de la izquierda, y posee 110999 valores.
- **2**: La cual corresponde a momentos en el que paciente está observando al estímulo de la derecha, y posee 115400 valores.

Las variables no contienen datos nulos, pero si existen datos dañados, o irrelevantes que corresponden a los valores de 'Stimulo = 99', ya que son valores

que no corresponden a momentos en los que el paciente estaba captando un estímulo, por lo que se toma la decisión de eliminarlos del dataset.

Los datos se adquirieron a una frecuencia de muestreo exacta 200Hz, esto quiere decir que por cada segundo tenemos 200 registros con información. Por lo tanto, si por cada segundo se tienen 200 registros de información, cada registro fue obtenido cada 5 milisegundos, de esta forma se parsearon los datos para ejecutar el análisis.

- **Curación del dataset:**

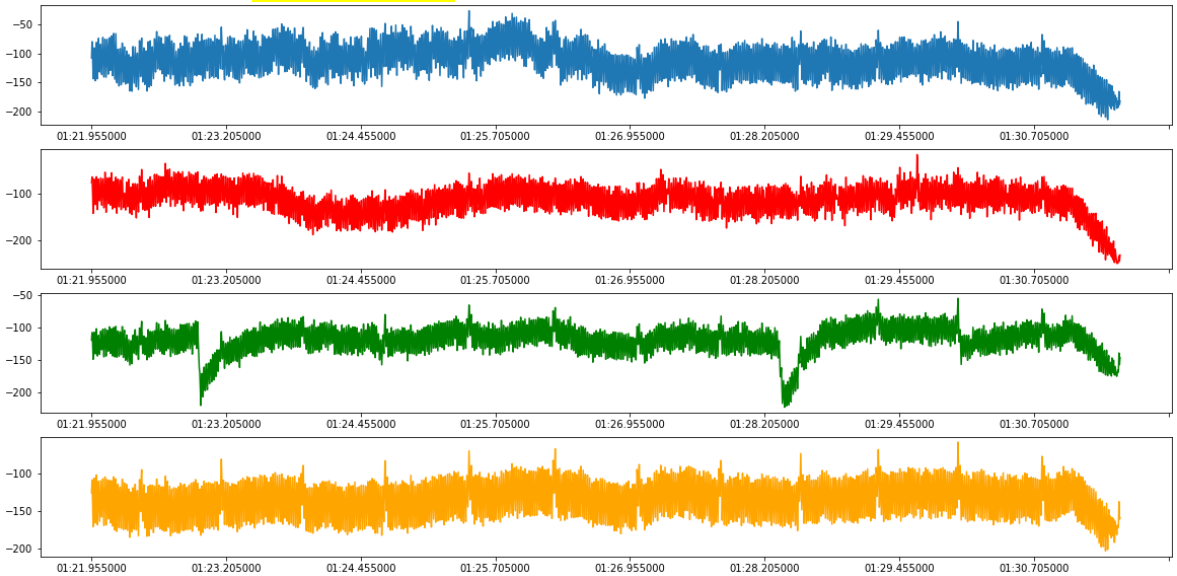
Se decidió trabajar desde un comienzo con el dataset filtrado para ejecutar el análisis ya que este poseía señales ruidosas, por lo que para la curación lo que se realizo fue:

1. Eliminar las columnas no relevantes mencionadas con anterioridad.
2. Restar la media a los canales necesarios, ya que sumimos que si la media está por encima de 300 hay ruido ya que la señal de EEG normal oscila entre los -250 y 250 microvoltios. La idea de este filtrado es eliminar esa componente de continua que sería ruido. Y posteriormente eliminar los valores superiores a 250 e inferiores a -250.

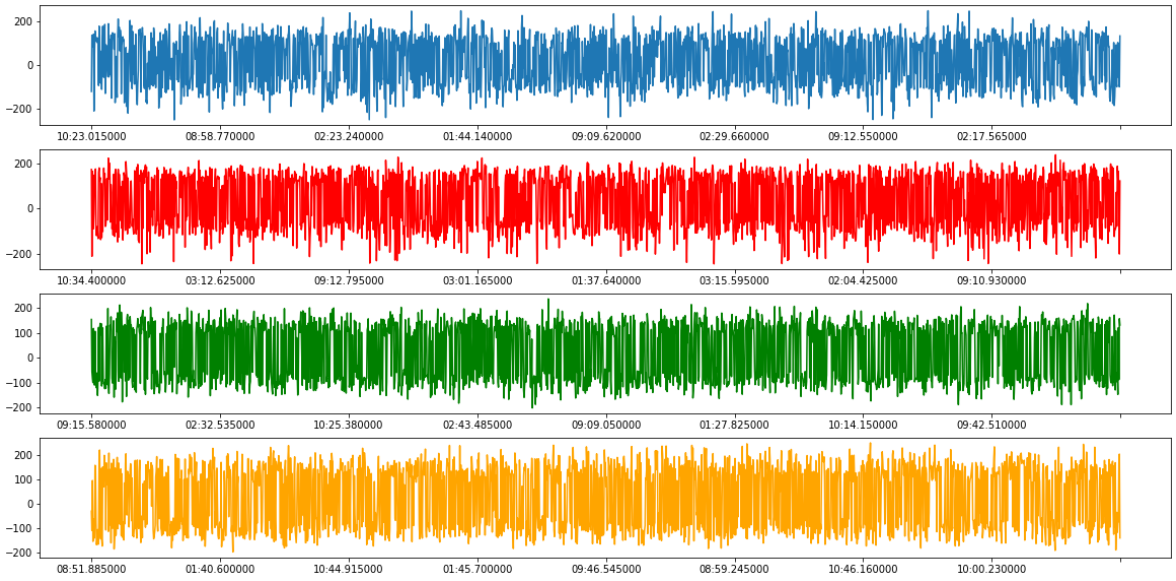
- **Visualizaciones de series temporales:**

○ **Un sujeto - todos los canales:**

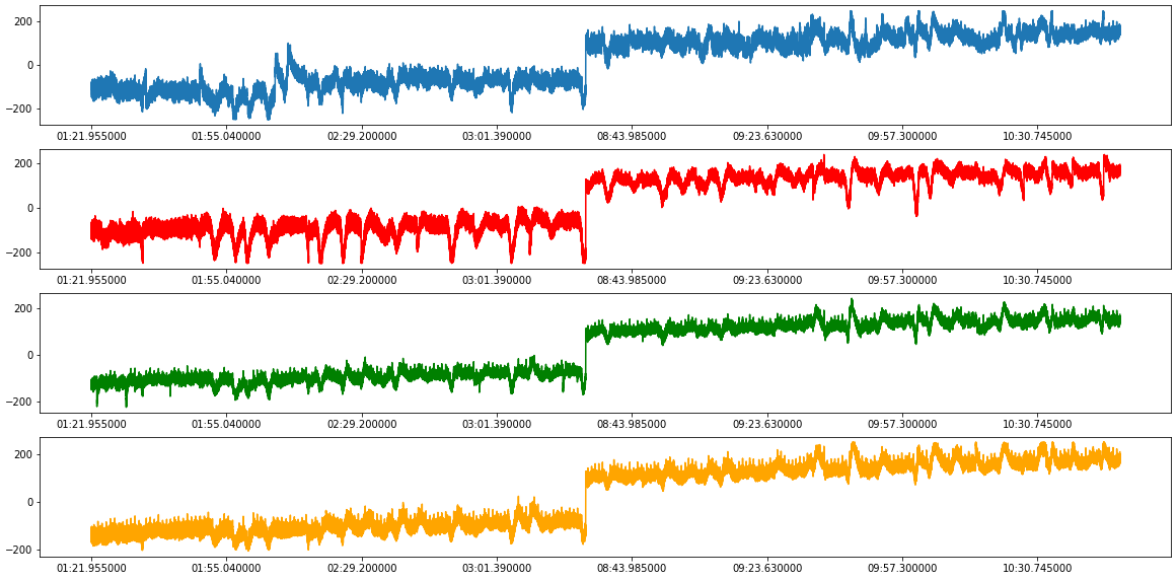
- **Tomando los primeros 10 segundos:**



- **Muestreando 2000 datos (equivalente a analizar 10 segundos):**

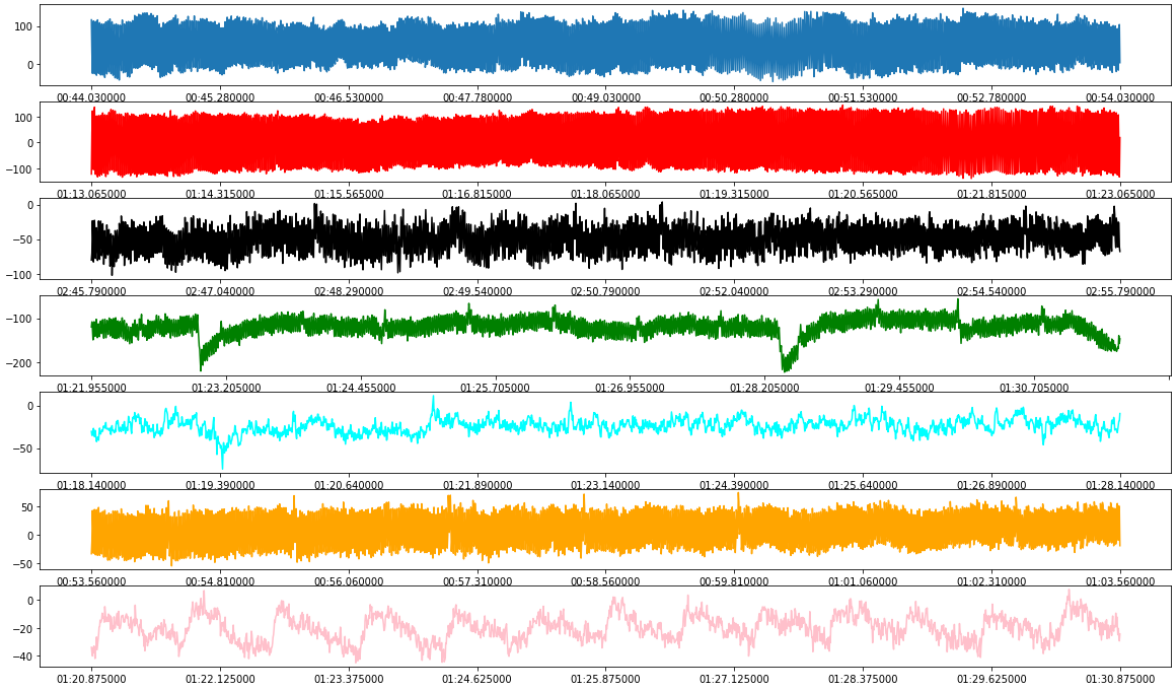


- **Teniendo en cuenta todos los datos:**

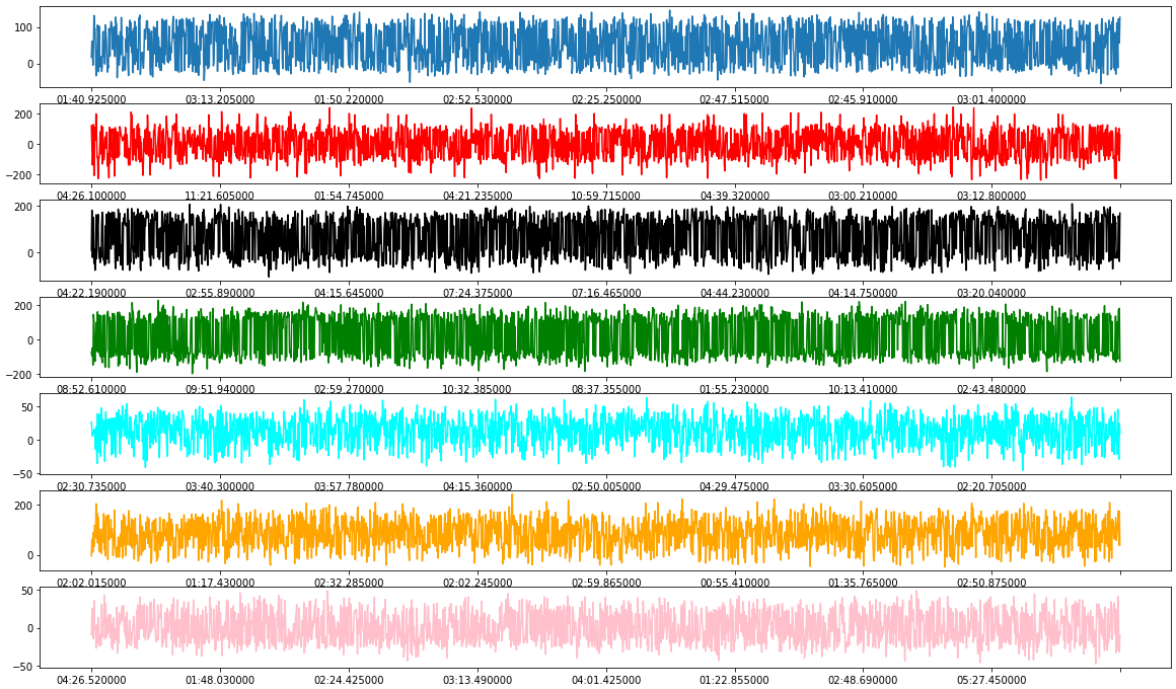


○ Un mismo canal - todos los sujetos

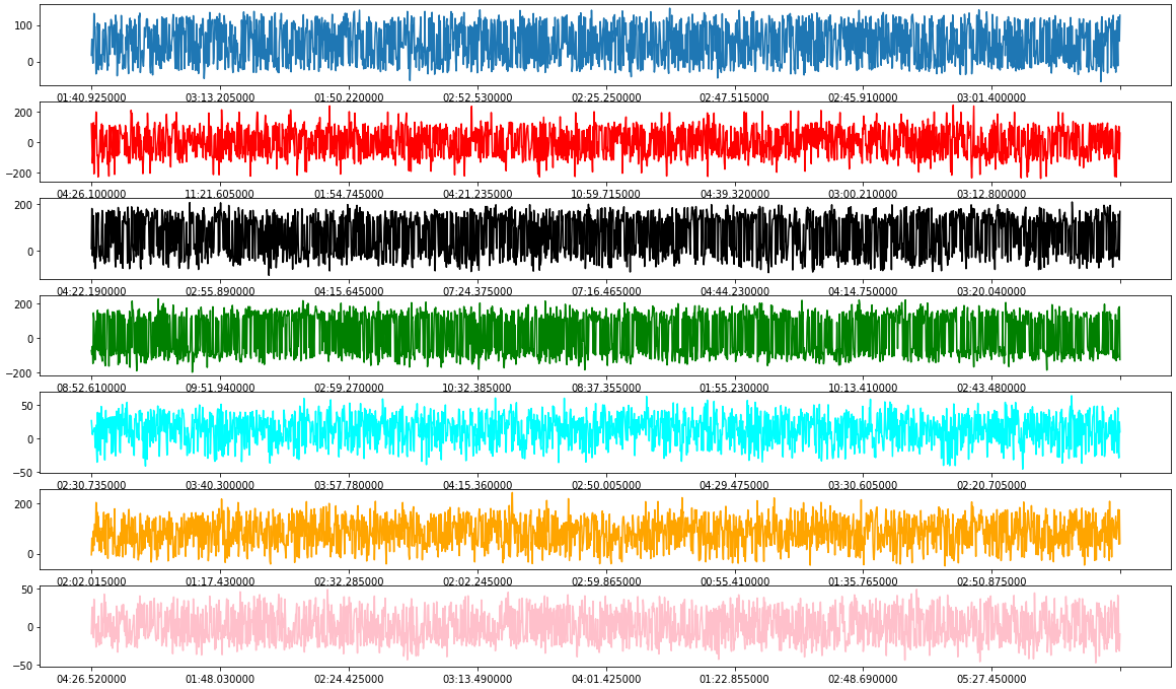
- **Tomando los primeros 10 segundos:**



- **Muestreando 2000 datos (equivalente a analizar 10 segundos):**

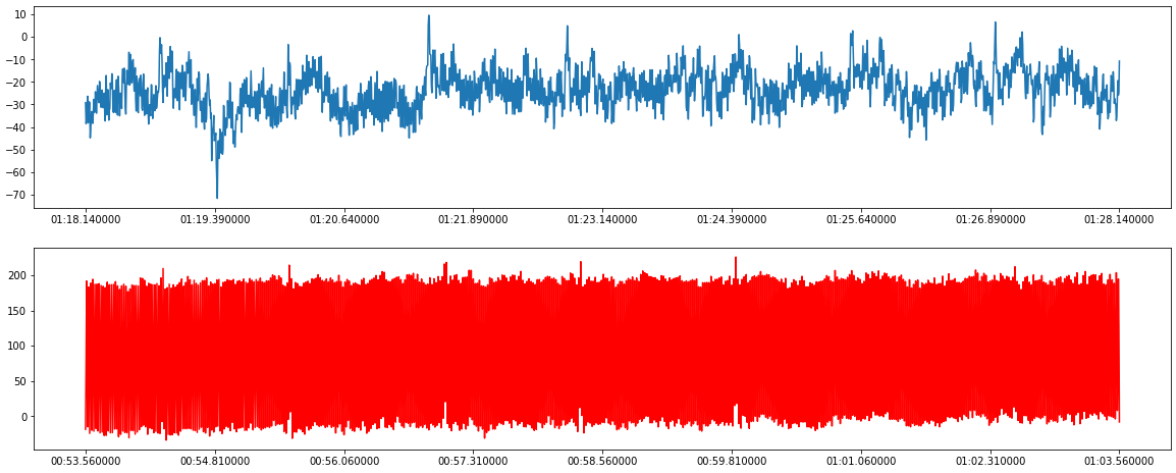


- **Teniendo en cuenta todos los datos:**

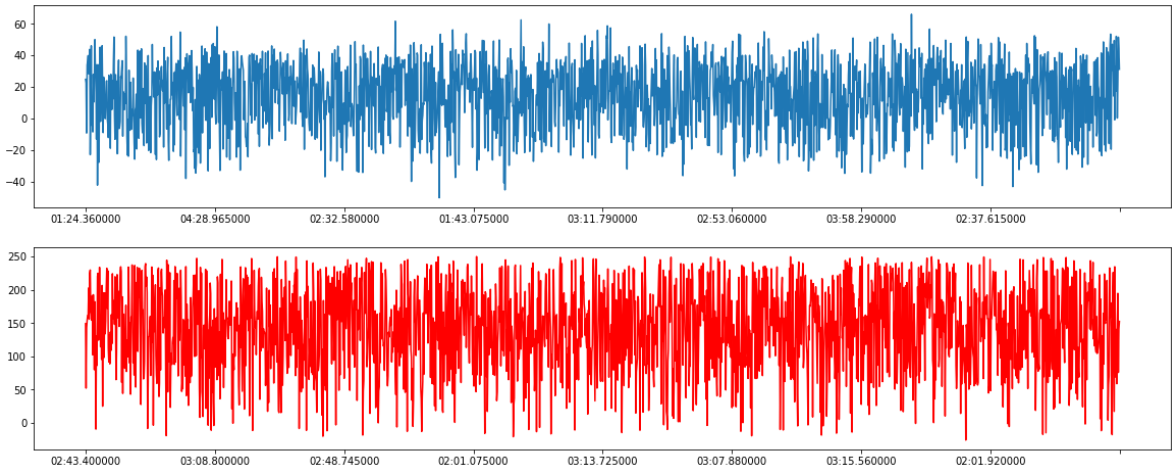


○ Un mismo canal - mismo sujeto en diferentes sesiones:

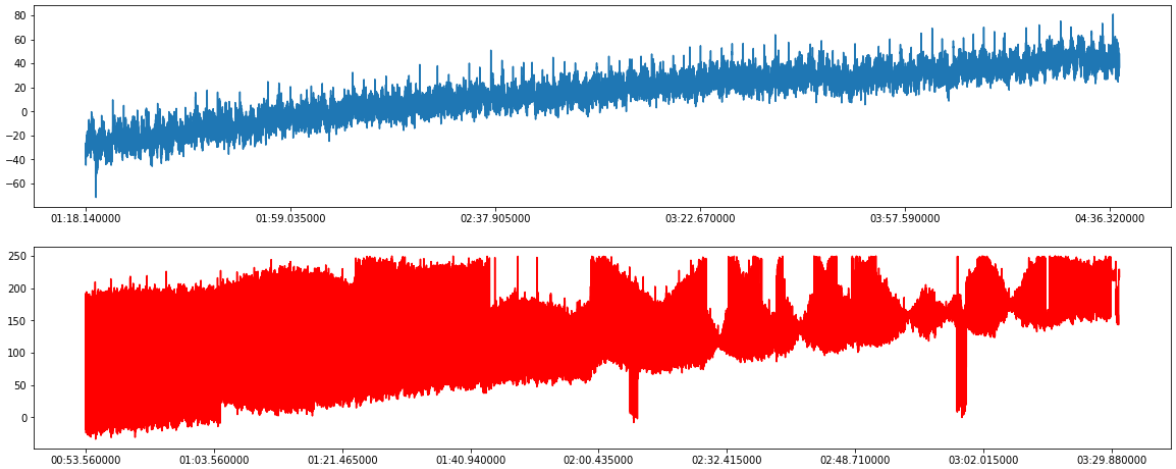
- **Tomando los primeros 10 segundos:**



- **Muestreando 2000 datos(equivalente a analizar 10 segundos):**

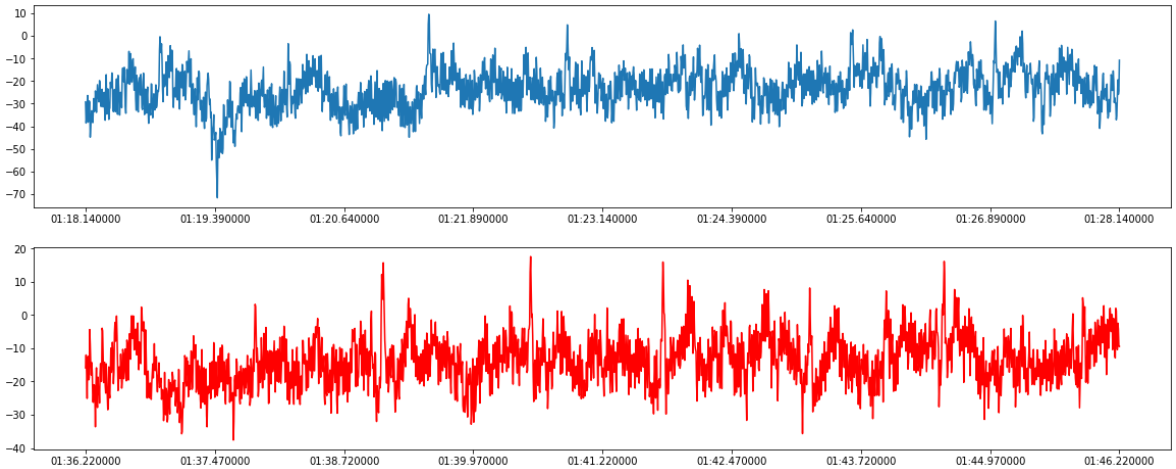


- **Teniendo en cuenta todos los datos:**

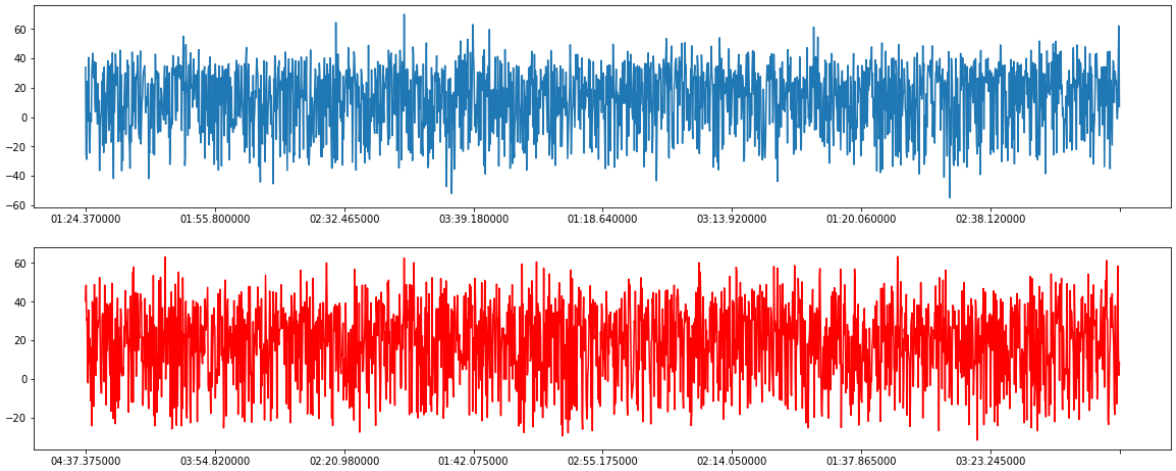


○<u>Un mismo sujeto y canal - diferentes estados:</u>

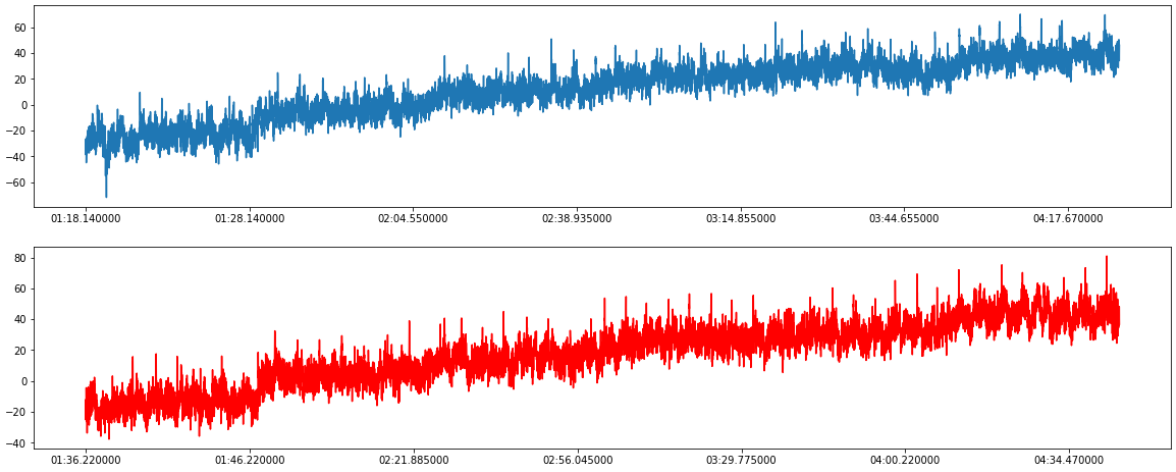
- **Tomando los primeros 10 segundos:**



- **Muestreando 2000 datos(equivalente a analizar 10 segundos):**

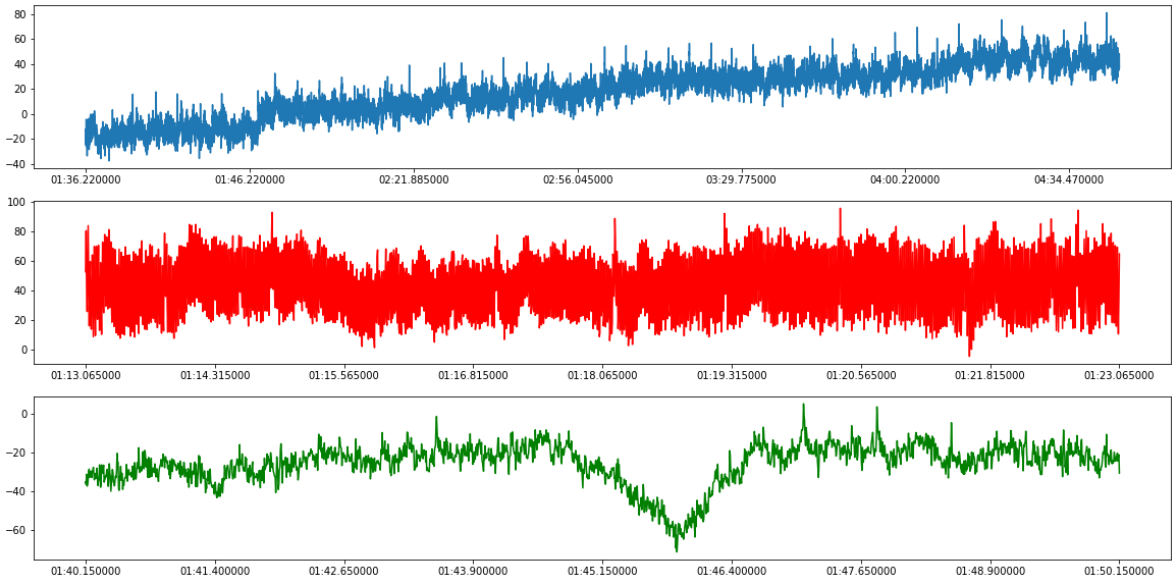


- **Teniendo en cuenta todos los datos:**

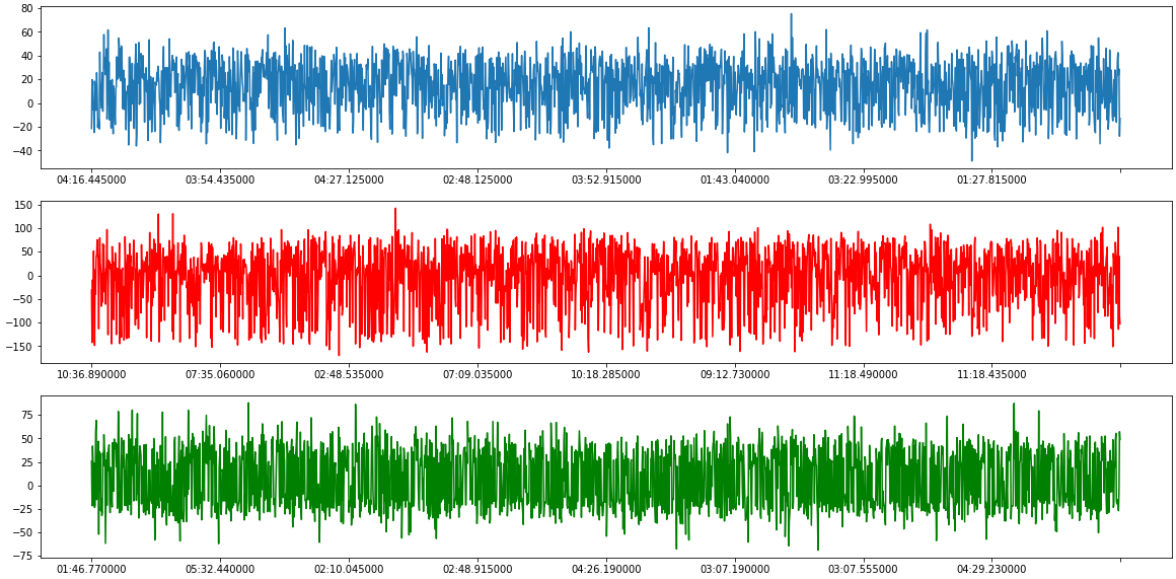


○<u>Mismo estado - Diferentes sujetos.</u>

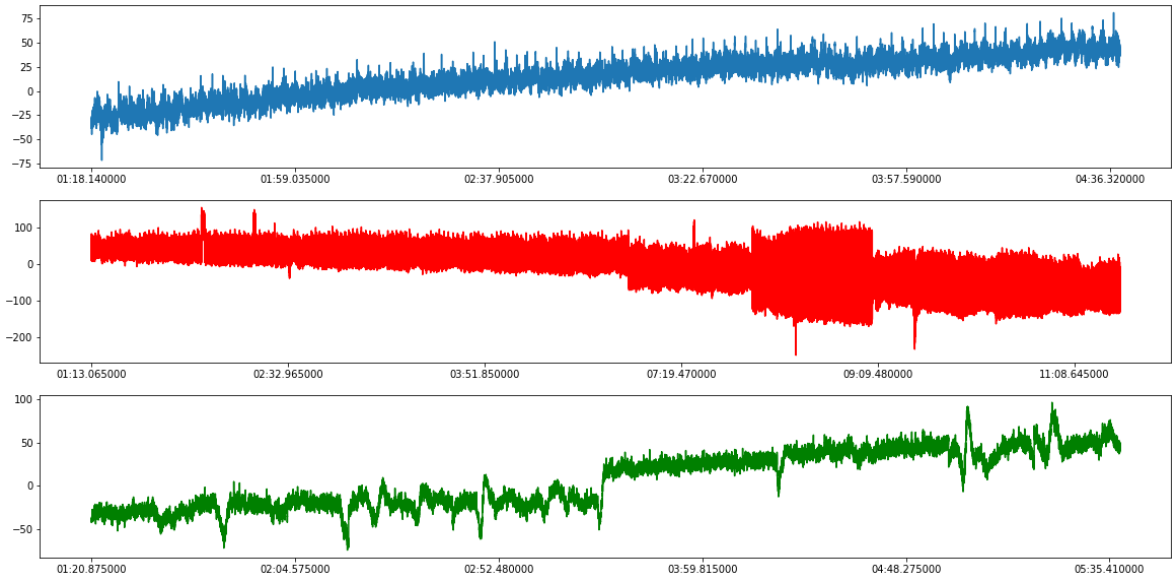
- **Tomando los primeros 10 segundos:**



- **Muestreando 2000 datos(equivalente a analizar 10 segundos):**



- **Teniendo en cuenta todos los datos:**



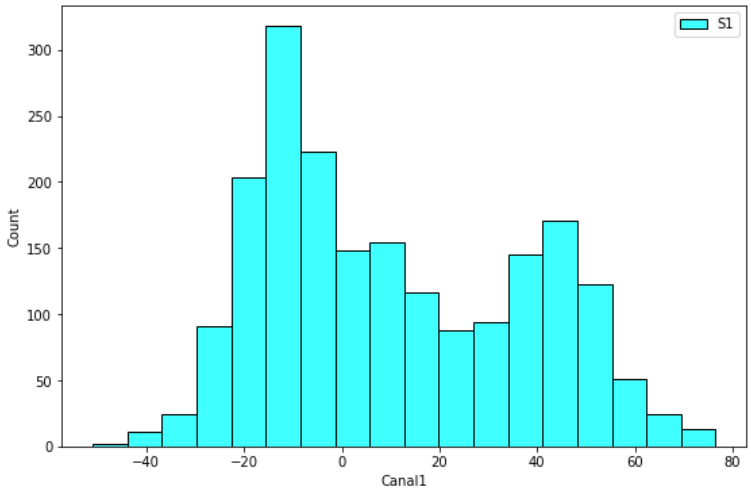
Como se puede observar al analizar estas imágenes, se ve que existe una tendencia de inclinación en la señal a medida que avanza el tiempo, a la vez de que la señal se vuelve más ruidosa.

Parte II: Dominio del tiempo.

- **Nivel Segmento/Estado:**

Se filtraron, para tener en cuenta el análisis, una muestra de 2000 datos para cada estímulo, lo que equivaldría a 10 segundos, correspondientes al canal 1 del paciente MA1.

Se observó gráficamente la distribución de una muestra de 2000 datos del paciente MA1 correspondiente al estímulo 1, y se realizó un test de normalidad, obteniéndose los siguientes resultados:

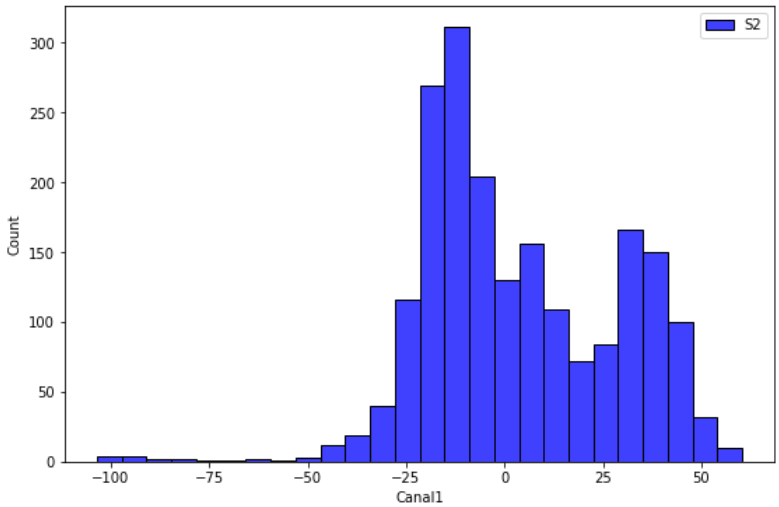


NormaltestResult:

- statistic=14.112413458066792
- pvalue=0.0008620418582425876

En este caso, en el que el paciente estaba captando el estímulo 1, correspondientes al canal 1, aunque gráficamente, los datos parecerían distribuirse como una normal, se puede observar que al realizar el test de hipótesis de normalidad, existen evidencias estadísticamente significativas para rechazar la hipótesis de que los datos se distribuyen de forma normal. Por lo que podría tratarse de una distribución bimodal.

Luego se observó gráficamente la distribución de una muestra de 2000 datos del paciente MA1 correspondiente al estímulo 2, y se realizó un test de normalidad, obteniéndose los siguientes resultados:



NormaltestResult:

- statistic=458.65732226362945

- pvalue=2.5341243861935035e-100

En este caso, en los que el paciente estaba captando el estímulo 2, correspondientes al canal 1, aunque gráficamente, los datos parecerían distribuirse como una normal, al igual que en el caso anterior, se puede observar que al realizar el test de hipótesis de normalidad, existen evidencias estadísticamente significativas para rechazar la hipótesis de que los datos se distribuyen de forma normal. Por lo que podría tratarse también de una distribución bimodal.

Medidas de resumen estadístico:

Stimulus1

- Count= 2000.000000
- Mean= 5.829524
- std = 25.043171
- min= -106.855006
- 25%= -13.480006
- 50%= 2.464994
- 75%= 30.084994
- Max= 60.584994

Stimulus1

- Count= 2000.000000
- Mean= 9.405524
- Std= 26.540701
- Min= -45.605006
- 25%= -12.127506
- 50%= 3.124994
- 75%= 34.372494
- Max= 87.044994

Se decidió tomar como parámetros de centralidad la media de las variables, ya que al tomarse la decisión de trabajar con el dataset curado desde un comienzo no se encuentran outliers a este nivel del análisis, por lo que no hay valores extremos que me alteren el estadístico de centralidad elegido.

Se realizó un test de hipótesis de diferencias de medias para evaluar si existen evidencias estadísticamente significativas entre los estimadores de posición central de los dos estados, lo que arrojó los siguientes resultados:

Ttest_indResult=

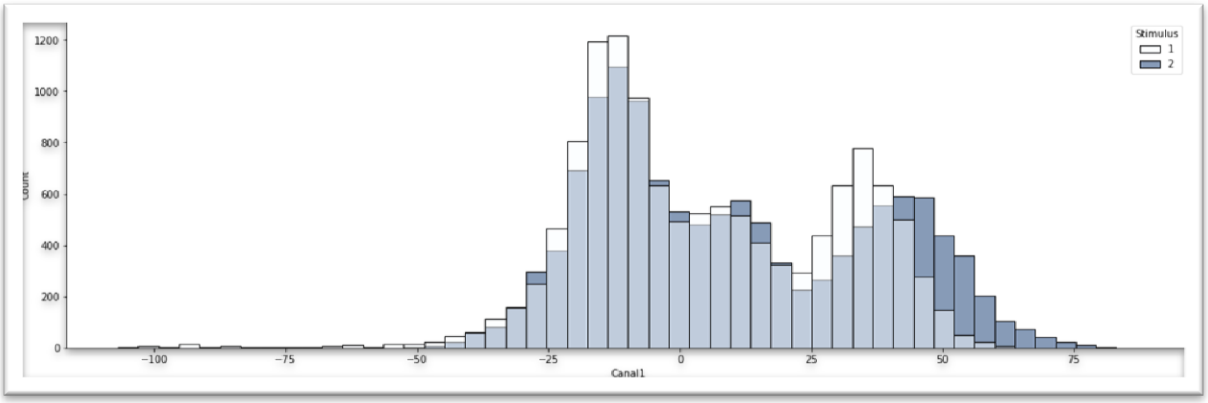
- statistic=-0.5175791823990686
- pvalue=0.6635191968384794

Lo que se puede observar a partir del test de hipótesis realizado es que no existen evidencias estadísticamente significativas para rechazar la hipótesis de que existe una diferencia entre los estimadores de posición central de ambos estados. Por lo que no existiría en principio diferencias entre las señales captadas cuando el paciente observa los diferentes estímulos.

- **Nivel Paciente - un canal:**

Se filtraron, para tener en cuenta el análisis los datos correspondientes al canal 1 del paciente MA1.

Se observó gráficamente la distribución de los seleccionados y se realizó un test de normalidad, obteniéndose los siguientes resultados:



NormaltestResult

- statistic=997.9651282420559
- pvalue=1.9707241394404536e-217)

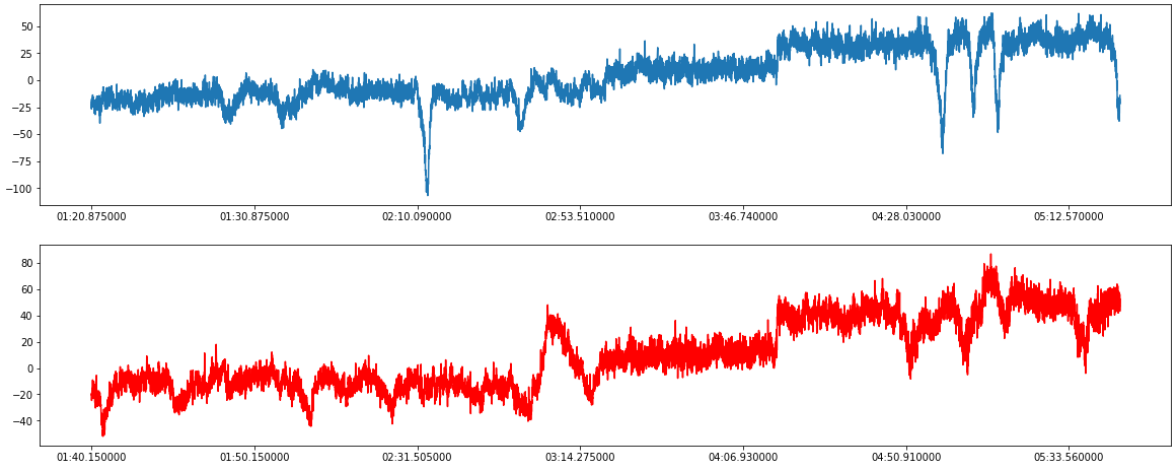
Según lo manifestado en el test de normalidad se muestra evidencias estadísticamente significativas para rechazar la hipótesis de que los datos se distribuyen de forma normal, por lo que en este caso también podría tratarse de una distribución bimodal, como en los casos anteriores.

Test de independencia de variables:

Ttest_indResult

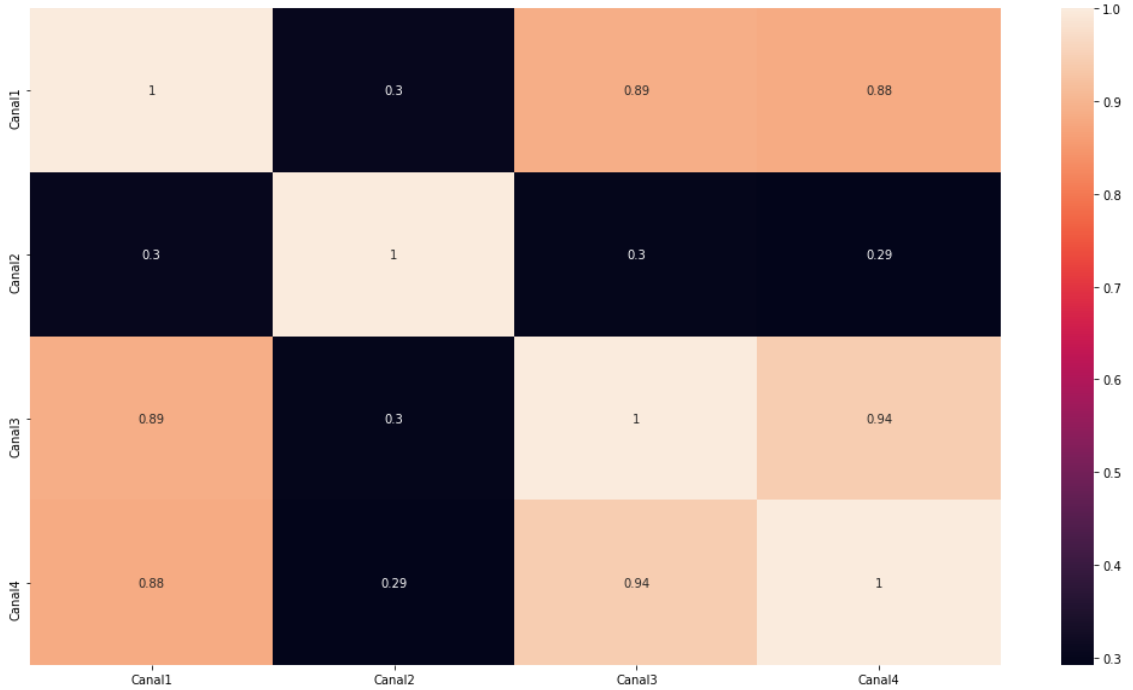
- statistic=-0.6976292070846736
- pvalue=0.5814498209547297

Según lo manifestado en el test de independencia se muestra evidencias estaísticamente significativas para no rechazar la hipótesis de que las señales captadas por los dos estados son independientes entre sí.



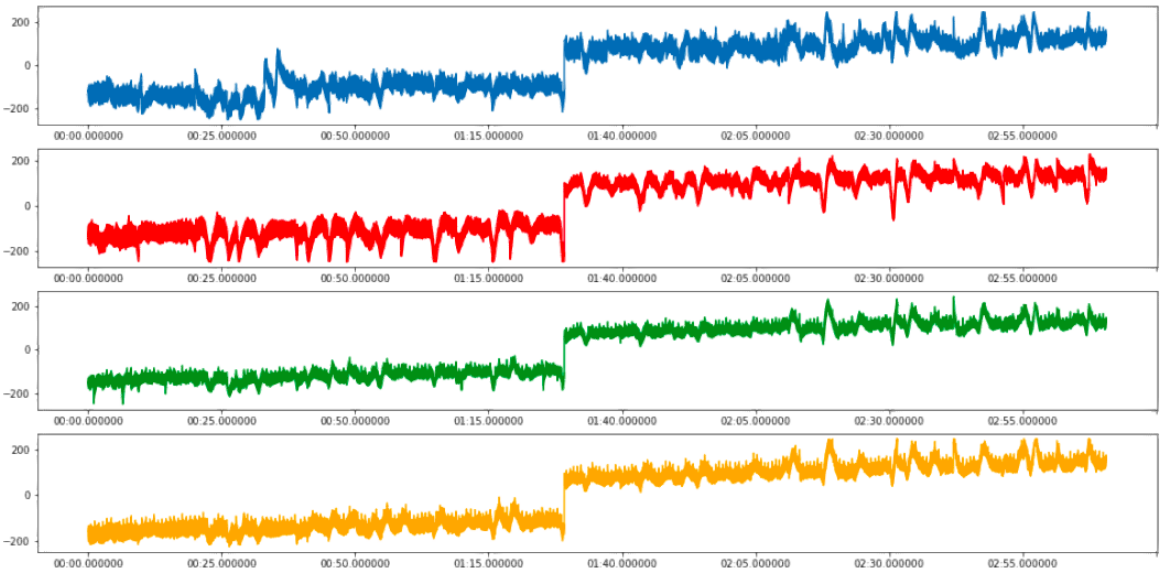
Como se puede observar en el gráfico en el que se representa cada uno de los estados correspondientes al canal 1 del paciente MA1, se puede observar que los voltajes varían a lo largo del tiempo, con una tendencia clara de aumento de la señal.

- **Nivel Paciente - multi canal:**



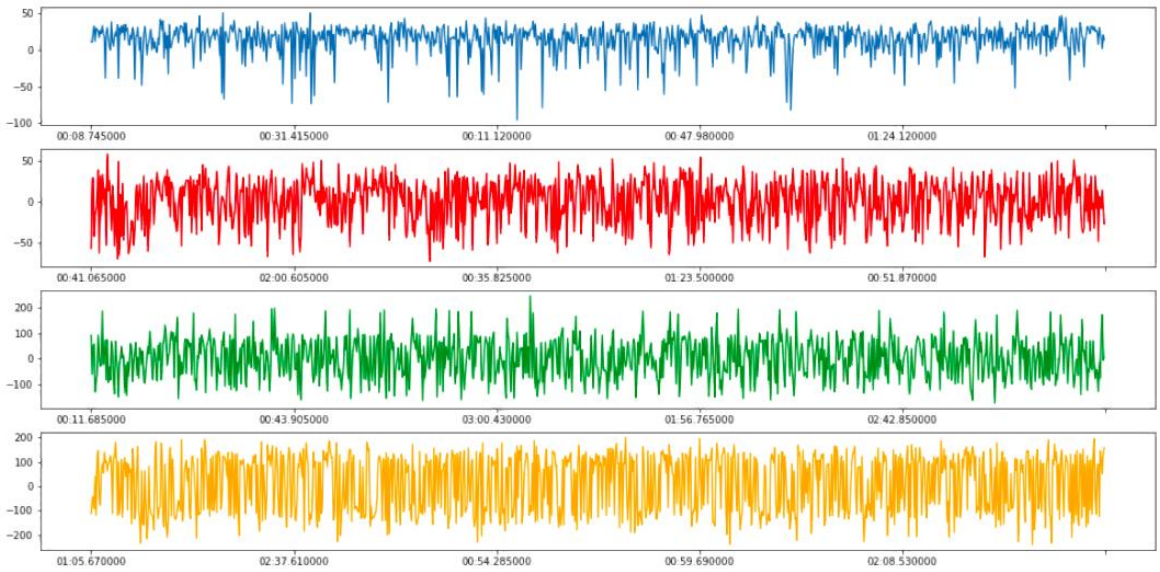
Se puede observar que existe una baja correlación entre el canal 2 y los demás canales, pero entre el resto de los canales, 4, 1 y 3, parecería haber cierta correlación, por lo que no sería necesario trabajar con los cuatro canales, ya que

los canales 4,1 y 3 tendrían una correlación lo suficientemente alta como para considerar utilizar solo uno de ellos, junto con el canal 2.



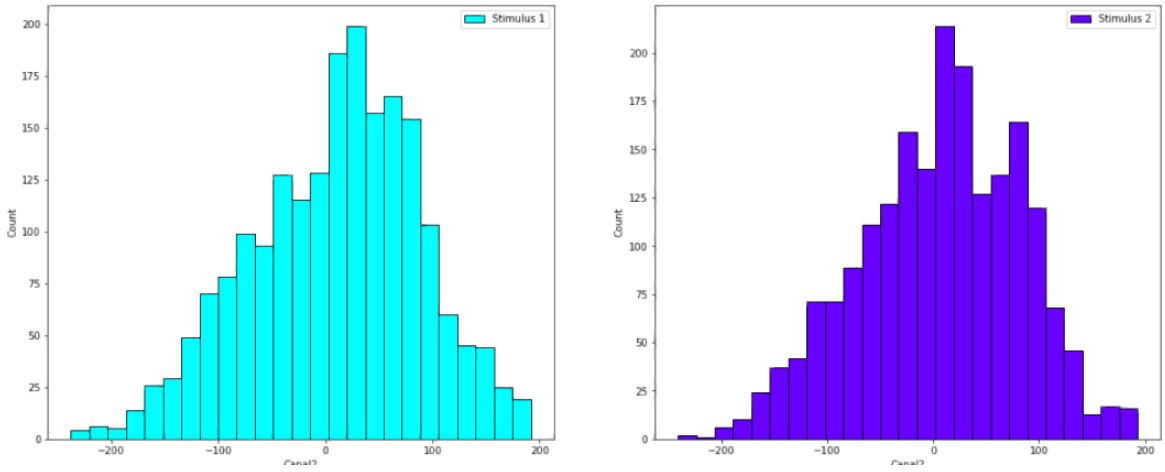
Así mismo se puede observar en el total de la muestra del mismo paciente y todos los canales, la diferencia entre los mismos y la tendencia a aumentar el voltaje con el paso del tiempo.
También podemos ver un salto en la misma sesion en la que los valores si bien siguen con la tendencia en aumento, hay un marcado aumento del volatje.

- **<u>Nivel Multi-Paciente:</u>**



En el grafico podemos observar los distintos pacientes para el canal 2, ya que es el que se observa que tiene menos correlación.

Las diferencias encontradas son notorias, ya que el dato de los pacientes se comporta de distinta manera para el mismo canal.



En el grafico anterior podemos observar la diferencia entre el Stimulus 1 y 2 para canal 2 de los pacientes.

	Canal2	Stimulus
count	52326.0	52326.0
mean	0.0	2.0
std	76.0	0.0
min	-190.0	1.0
25%	-59.0	1.0
50%	1.0	2.0
75%	63.0	2.0
max	250.0	2.0

AA1

	Canal2	Stimulus
count	25241.0	25241.0
mean	0.0	1.0
std	28.0	1.0
min	-87.0	1.0
25%	-21.0	1.0
50%	4.0	1.0
75%	22.0	2.0
max	68.0	2.0

JA1

	Canal2	Stimulus
count	25257.0	25257.0
mean	15.0	1.0
std	19.0	1.0
min	-100.0	1.0
25%	8.0	1.0
50%	19.0	1.0
75%	27.0	2.0
max	56.0	2.0

MA1

	Canal2	Stimulus
count	38135.0	38135.0
mean	5.0	2.0
std	122.0	0.0
min	-250.0	1.0
25%	-110.0	1.0
50%	51.0	2.0
75%	119.0	2.0
max	230.0	2.0

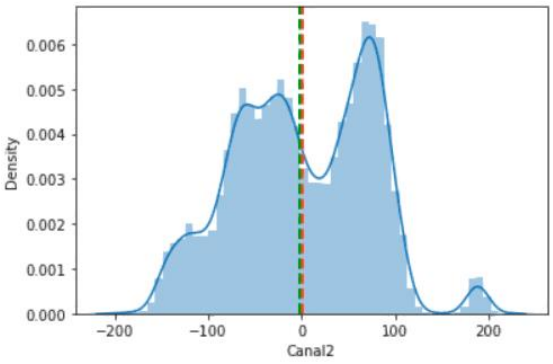
HA1

Como podemos observar para el mismo canal, observamos que las medias de cada paciente son diferentes.

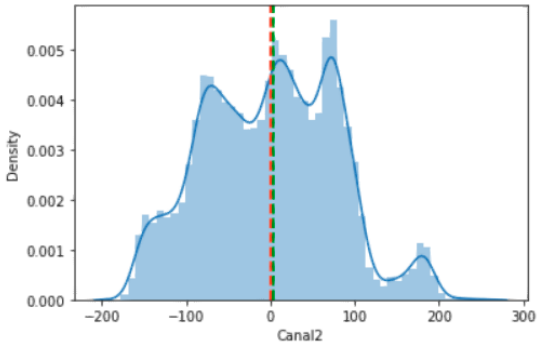
A continuación se grafican las medias y medianas para cada estimulo de las personas, determinando así que todos los pacientes tienen medias diferentes.

AA1

Stimulus 1

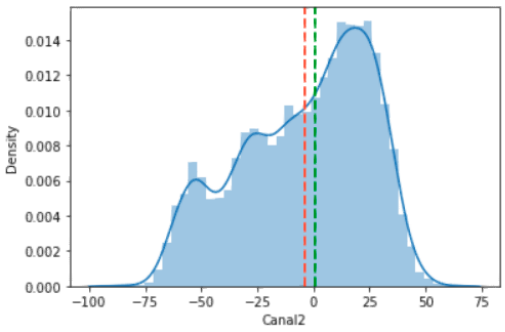


Stimulus 2

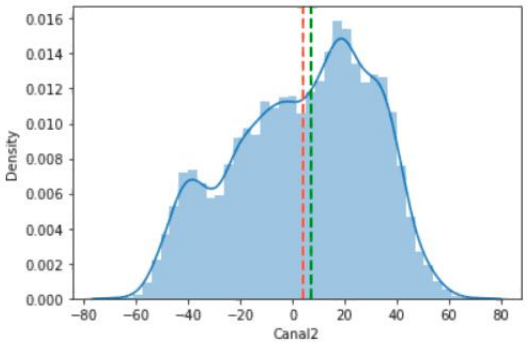


JA1

Stimulus 1

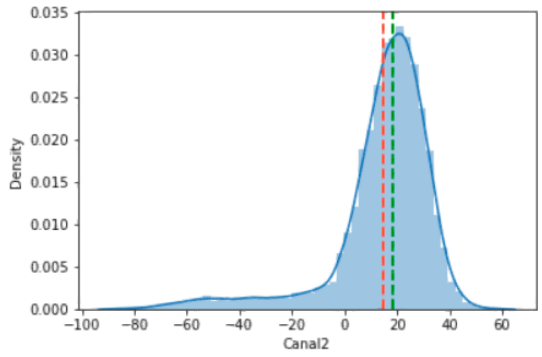


Stimulus 2

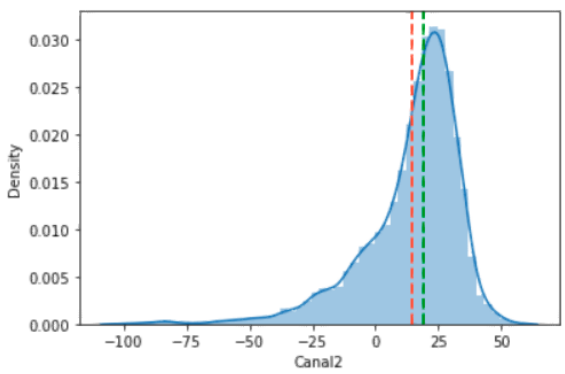


MA1

Stimulus 1

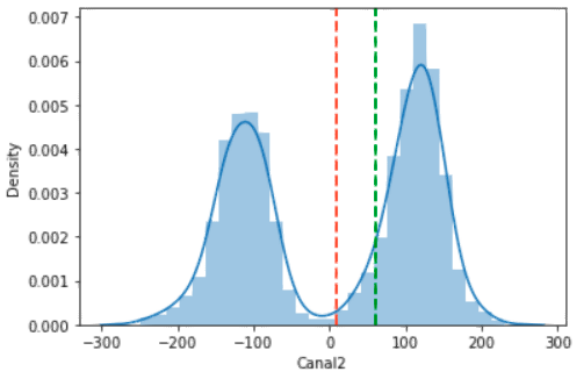


Stimulus 2



HA1

Stimulus 1



Stimulus 2