

Mentoría Data Science aplicado a BCI

Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones

Edición 2021

Integrantes 'Grupo 2'

- Iberra Yanina
- Junco Luis
- Wolfman Gabriel

Contexto de aplicación

En este informe abordaremos el análisis de un dataset de señales de electroencefalograma (EEG).

El dataset consiste en un conjunto de series temporales que reflejan variaciones de voltaje a lo largo del tiempo tomadas a una tasa de muestreo de 200 muestras por segundo, en un contexto experimental en donde los sujetos experimentales fueron estimulados por luces en dos frecuencias específicas: 12,5 Hz y 16.5 Hz.

Contamos con un conjunto de 7 registros, realizados en 4 individuos en 7 sesiones de adquisición utilizando para la toma de las señales la placa Ganglion Board, por medio del software de OpenBCI.

Al presentarle a un potencial usuario un estímulo visual al cual puede dirigir su vista, registrando su actividad cerebral y tras aplicar procesamiento de señales (ciencia de datos) a los registros conseguidos, el cambio de frecuencia de las señales de EEG puede utilizarse como una señal de control binaria. Más aún, si se disponen de diferentes estímulos visuales a diferentes frecuencias, la identificación de cada una de esas frecuencias en la señal eléctrica, permite tener un abanico más complejo de control, que luego puede luego transformarse en un comando o una decisión (la selección de una entre varias opciones, cada una asociada a una luz distinta por ejemplo).

Descripción de cada columna:

- Una primer columna de cuenta de muestras,
- Cuatro columnas con los datos de cada canal correspondientes a su respectivo electrodo de adquisición,
- Tres columnas con las mediciones de acelerómetros presentes en la placa Ganglion (no utilizados en estos estudios),
- Una columna de etiquetas (serán descritas a continuación),

- Una columna de marcas temporales,
- Una última columna de etiquetas temporales utilizadas por el Software OpenBCI para reproducir las señales posteriormente (tampoco utilizadas en este estudio).

Metadatos:

- OpenBCI Raw EEG Data
- Number of channels = 4
- Sample Rate = 200.0 Hz
- First Column = SampleIndex
- Last Column = Timestamp
- Second to last column = stimulus/prediction tags
- TAG CODE:
 - During calibration: 0 --> not looking; 1 --> looking left; 2 --> looking right, 99 --> NaN (default value)
 - During prediction (idem calibration + 10): 10 --> not looking; 11 --> looking left; 12 --> looking right
- Other Columns = EEG data in microvolts followed by Accel Data (in G) interleaved with Aux Data

Parte I: Exploración de la base de datos.

A) Leer los datos, eliminar los metadatos innecesarios.

Eliminamos las primeras 10 líneas de cada archivo, ya que contienen descripciones del hardware y software utilizados, y no las mediciones propiamente dichas.

Cada archivo se incorpora al análisis en una estructura de tabla que en Python se denomina 'dataframe'. Agregamos a cada dataframe dos nuevas columnas:

- 'persona', que indica el individuo al que corresponden las mediciones ('AA', 'HA', 'JA' y 'MA');
- 'sesión', que contiene el número de sesión en la que el individuo se realizó las mediciones (0, 1 ó 2), ya que una misma persona puede tener más de una sesión.

Se tendrán 7 datasets para analizar:

- AA0
- AA1
- AA2
- HA1
- JA1
- JA2
- MA1

B) Describir las características generales del dataset:

Número de registros, diferencias entre los mismos.

El número de registros con el que se cuenta sin realizar ninguna modificación a los dataset es la siguiente:

El archivo AA0 tiene 45.973 muestras.
El archivo AA1 tiene 138.169 muestras.
El archivo AA2 tiene 110.384 muestras.
El archivo HA1 tiene 137.693 muestras.
El archivo JA1 tiene 55.860 muestras.
El archivo JA2 tiene 101.323 muestras.
El archivo MA1 tiene 69.710 muestras.

Muestras totales = 659.112. Todas las muestras contienen 11 columnas.

Graficamos la distribución de frecuencias de las mediciones por personas, por sesión.

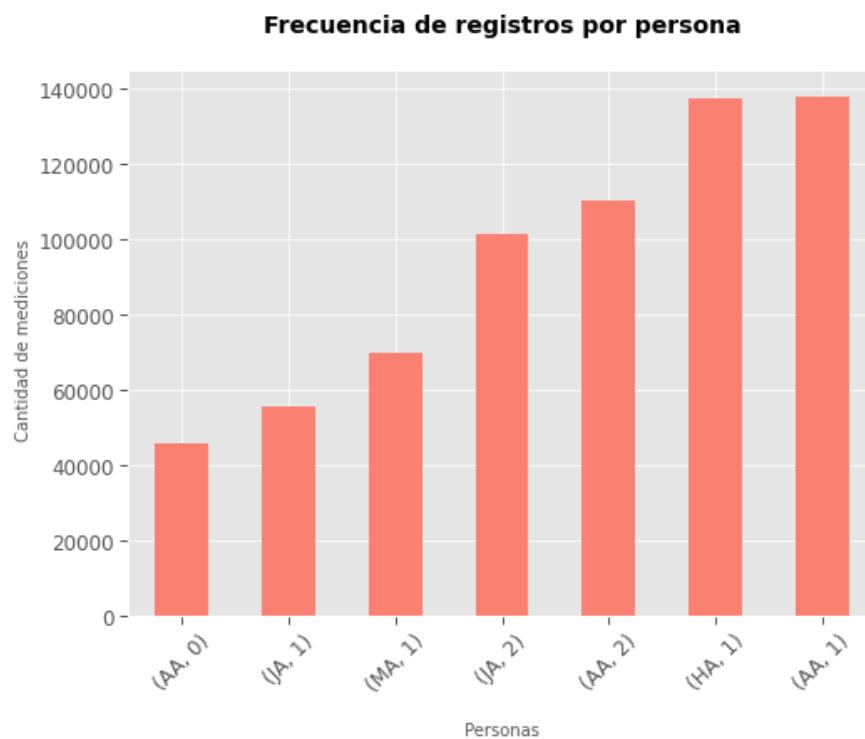


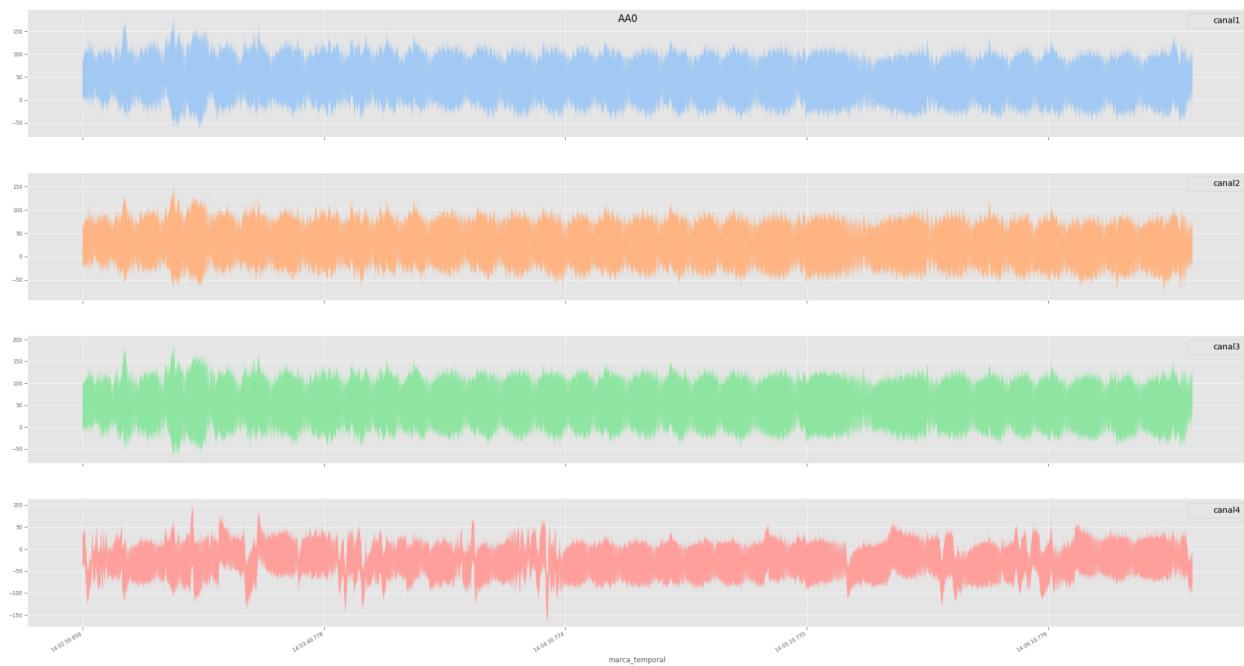
Gráfico 1.1

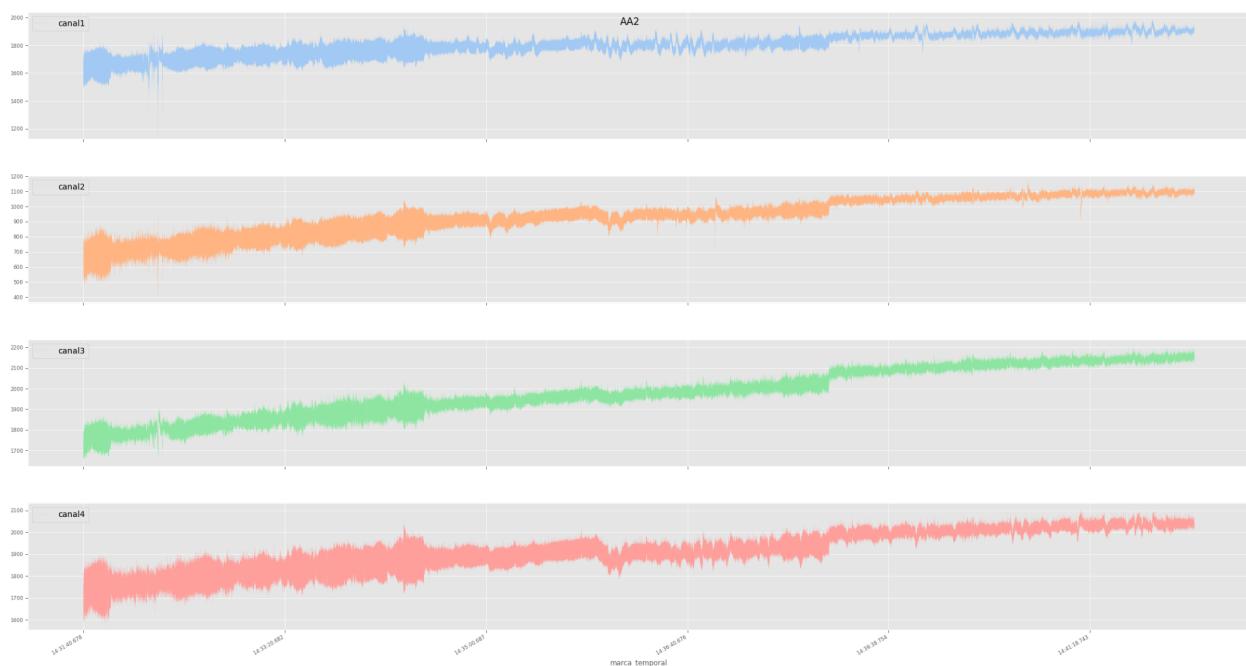
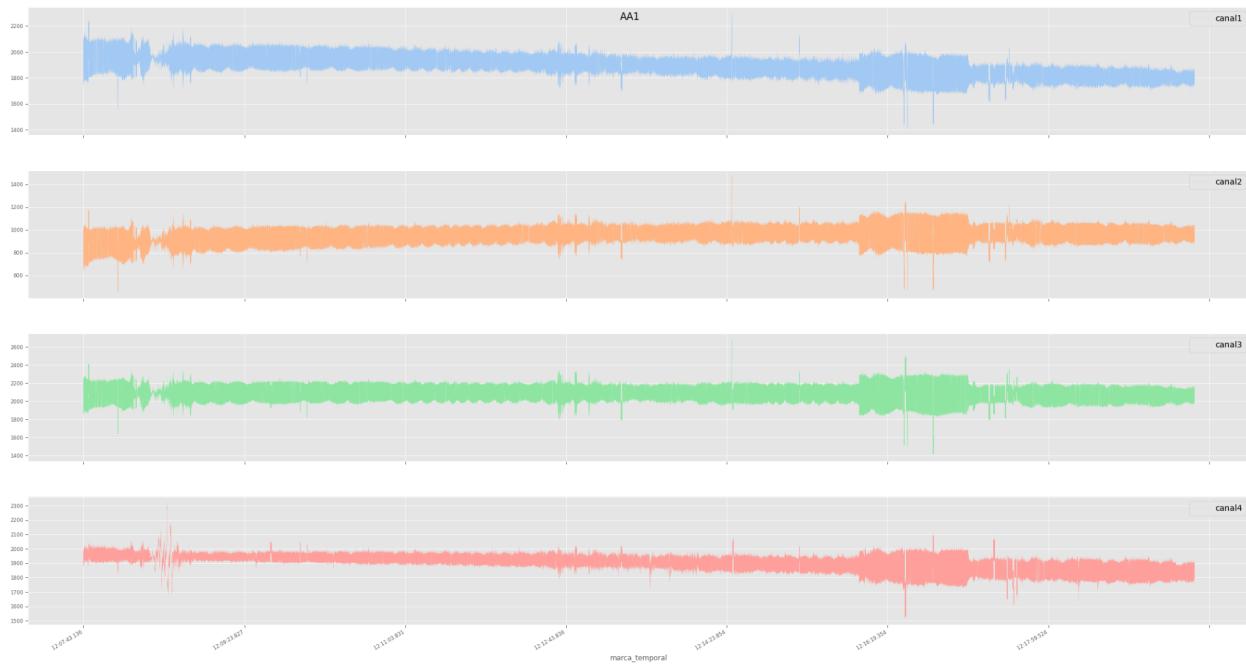
Observaciones:

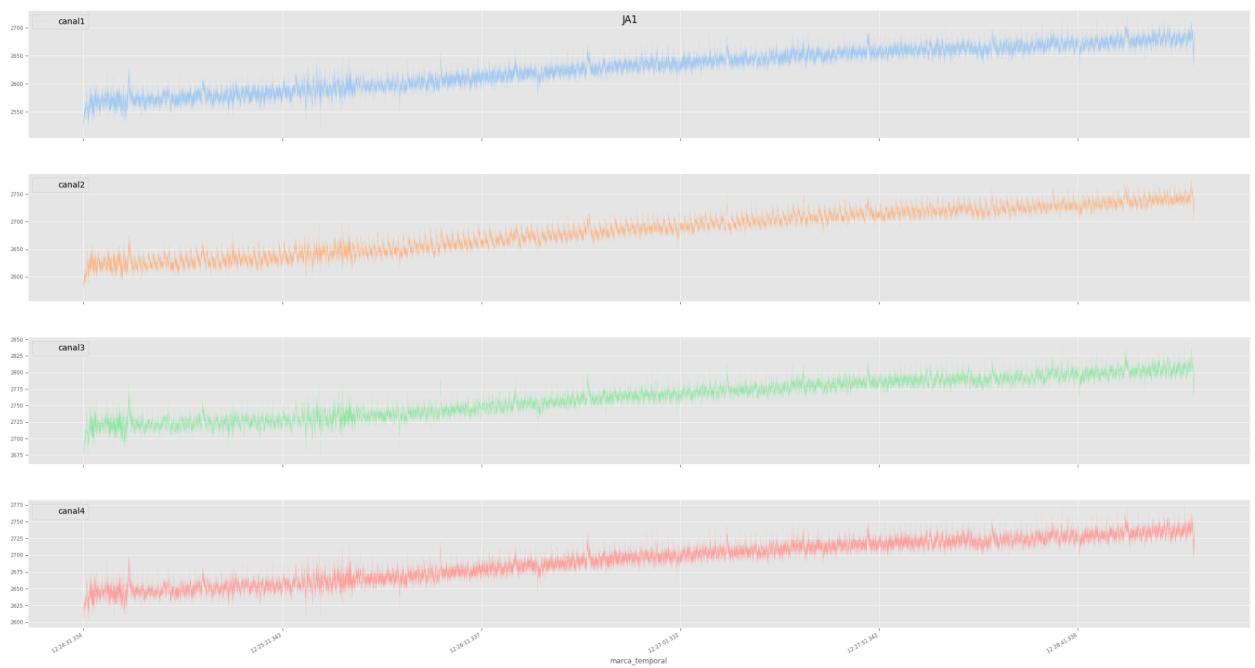
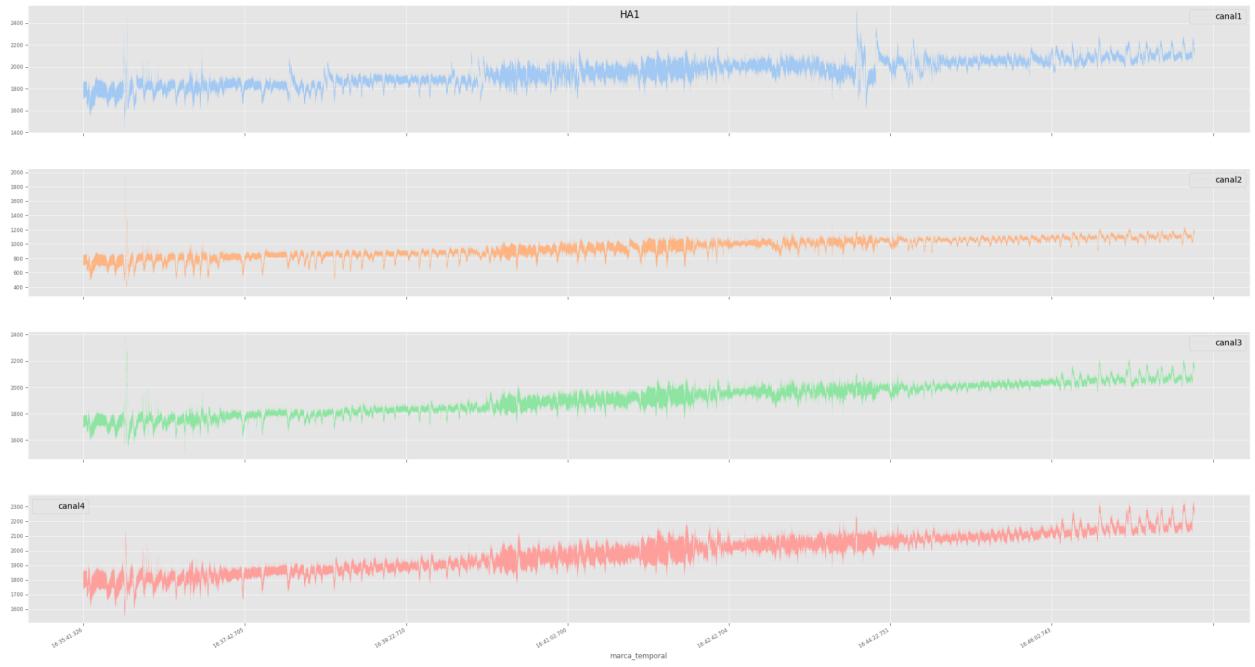
- Notamos cierto desbalance en los datos, ya que la persona 'AA' concentra el 44.7% de los datos, ya que tiene 3 sesiones, en lugar de una o dos como el resto de los individuos.

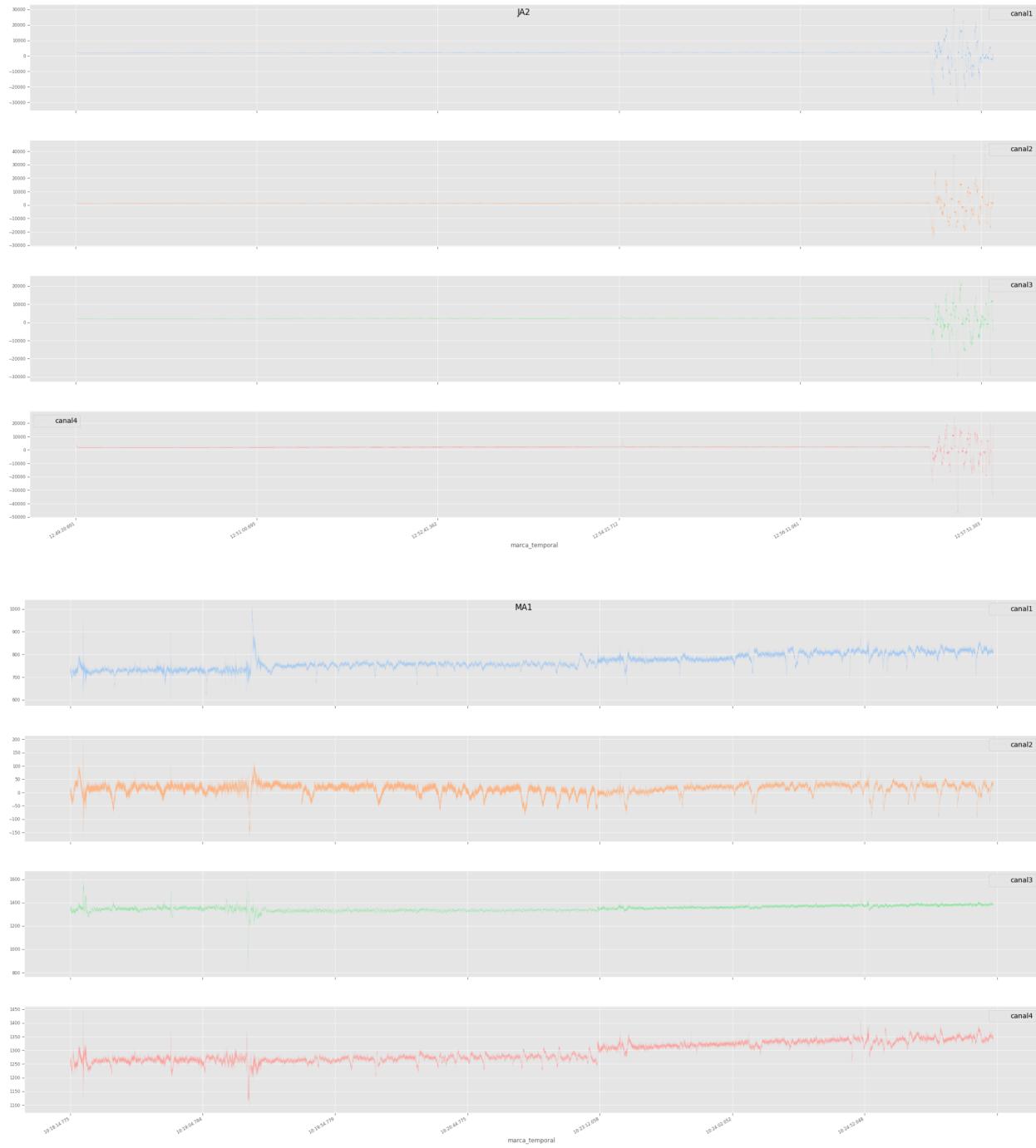
Definir conveniencia de trabajar todos juntos como un solo dataset, o por separado.

Para determinar la conveniencia de trabajar con un único dataframe que unifique todos los registros o los datos por separado para cada individuo y sesión, se grafican los voltajes de cada canal en el tiempo por segmento (persona/sesión).









Observaciones:

- Las mediciones de cada canal tienen rangos variados que oscilan entre los -46913,84 y 44839,75 microvoltios (μ V). Con medias entre 900 y 2000 μ V aproximadamente, lo cual sugiere la presencia de valores extraños o ruido, que se analizarán en las secciones siguientes.

- Se decide trabajar con los datos separados por persona/sesión, ya que se verifican rangos de voltajes distintos en cada caso y por lo tanto será conveniente trabajar cada dataset por separado.

C) Analizar las columnas presentes en el dataset:

¿Todas las columnas son relevantes? ¿Cuáles contienen información útil?

No todas las columnas presentes en los dataset son relevantes para el análisis de éste estudio, por lo tanto se eliminaron las siguientes columnas:

- Columnas con las mediciones de acelerómetros presentes en la placa Ganglion ('acelero1', 'acelero2', 'acelero3').
- Columna de etiquetas temporales utilizadas por el Software OpenBCI para reproducir las señales posteriormente ('marc_temp_sw').

Se conservan las restantes columnas ('num_muestra', 'canal1', 'canal2', 'canal3', 'canal4', 'etiqueta', 'persona' y 'sesion'), ya que contienen información de las mediciones realizadas y los valores de las mismas:

- Se tienen 6 variables numéricas y 2 que son consideradas 'object' por contener strings (incluida la marca temporal que transformamos a un tipo de dato 'datetime 64[ns]' para poder realizar los gráficos de la señal en función del tiempo).
- La columna de tiempo indica en qué momento de tiempo fue realizada cada muestra (en formato HH:MM:SS.MS), cuenta con una sensibilidad de milisegundos.
- La columna 'num_muestra' identifica cada muestra con valores que van desde el 0 al 200 y luego vuelve a reiniciarse (notar que cuentan con 201 valores antes de reiniciar el contador).
- Las 4 columnas de los canales son los provenientes de los electrodos dispuestos en la zona occipital del encéfalo (de manera no invasiva) las cuales expresan la activación de un conjunto de neuronas de la zona con una sensibilidad de micro Volt.
- La variable 'etiqueta', si bien en python se interpreta como entero, la trabajaremos como una variable categórica, ya que tiene dos valores posibles (1 y 2). Nos indica la luz que estuvo encendida durante la prueba, teniendo dos luces de distinta frecuencia. El valor '1' de la etiqueta se asigna a la luz parpadeante a 12,5 Hz y el '2' para la de 16,5 Hz.
- Los registros con etiqueta '99' fueron removidos ya que es cuando el individuo no observa ninguna luz lo cual es considerado un factor de ruido para el análisis.
- Luego de eliminar los registros con etiqueta '99', se observa una disminución en los rangos de voltajes que toman los canales, que van desde un mínimo -146 µV (en el canal 4) a un máximo de 2845 µV (en el canal 3).
- Luego de eliminar los valores '99' quedan 252337. Cantidad de registros eliminados: 406775 del total 659112 (61.72%). Cantidad de registros por persona y sesión:
 - AA0: 29448
 - AA1: 52541
 - AA2: 39959

- HA1: 39965
- JA1: 25241
- JA2: 39926
- MA1: 25257

Frecuencia de registros por persona luego retirar etiqueta

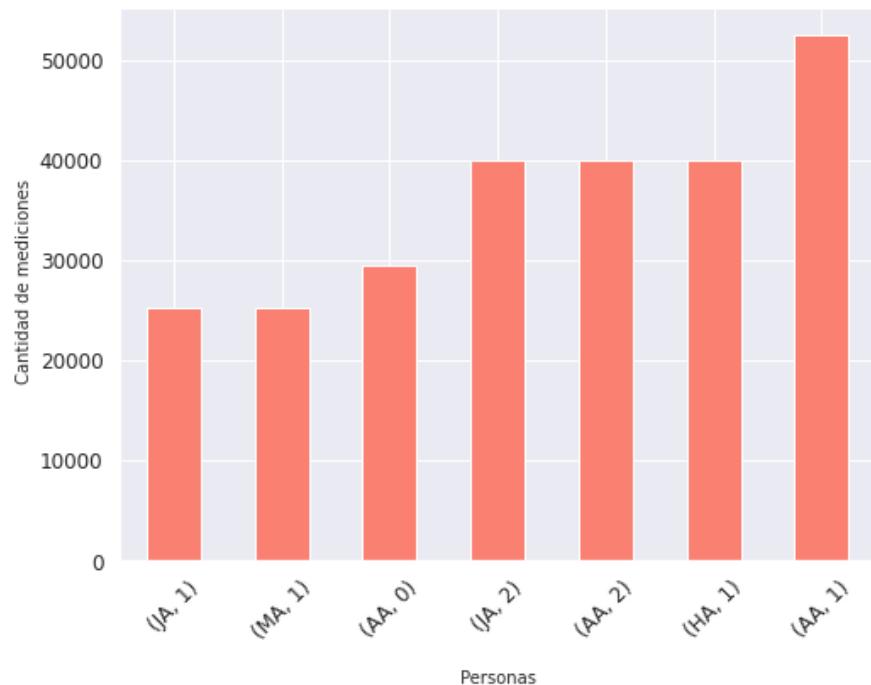
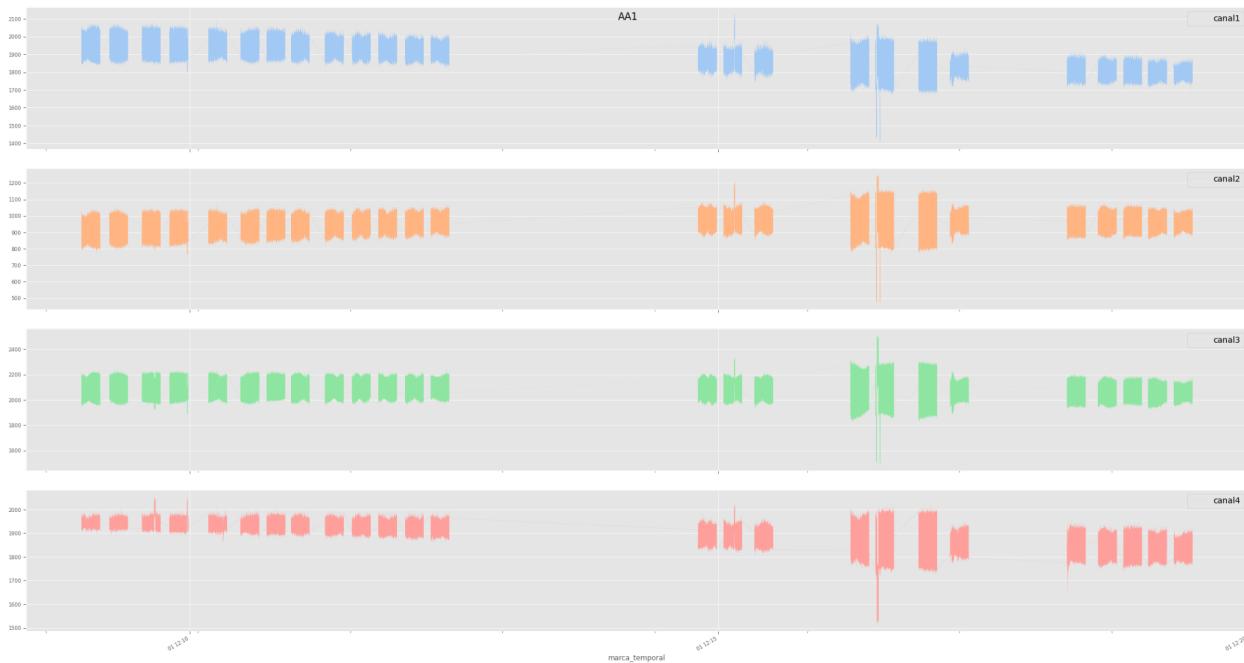
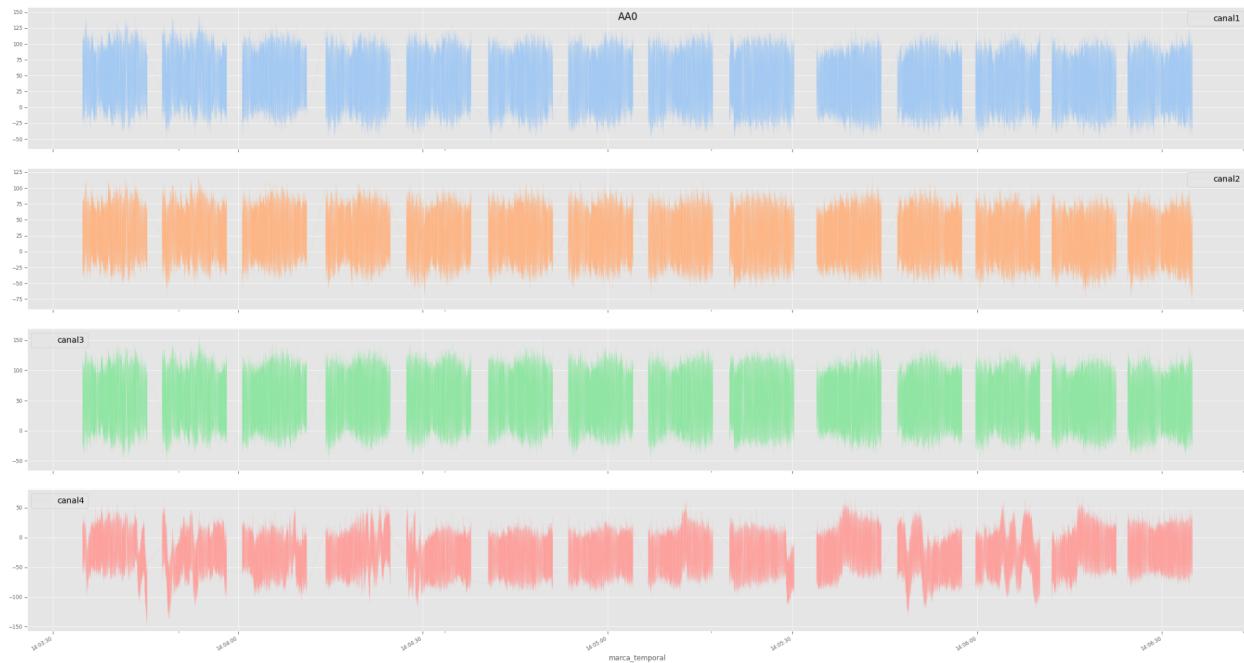
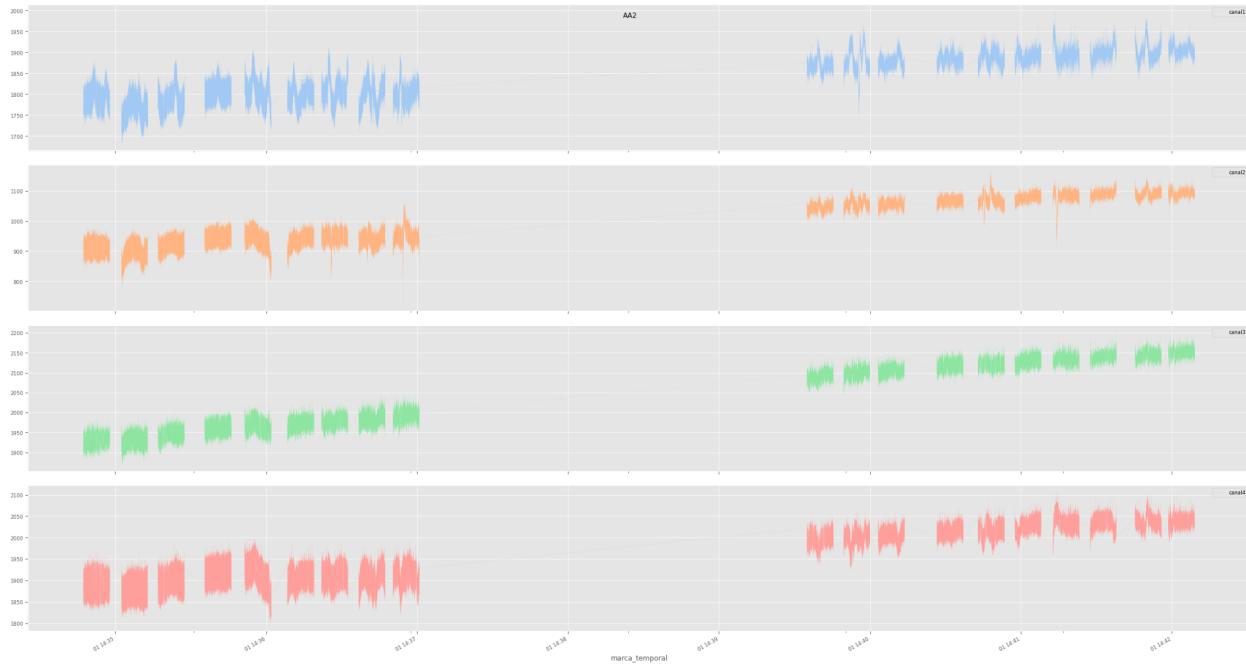


Gráfico 1.2

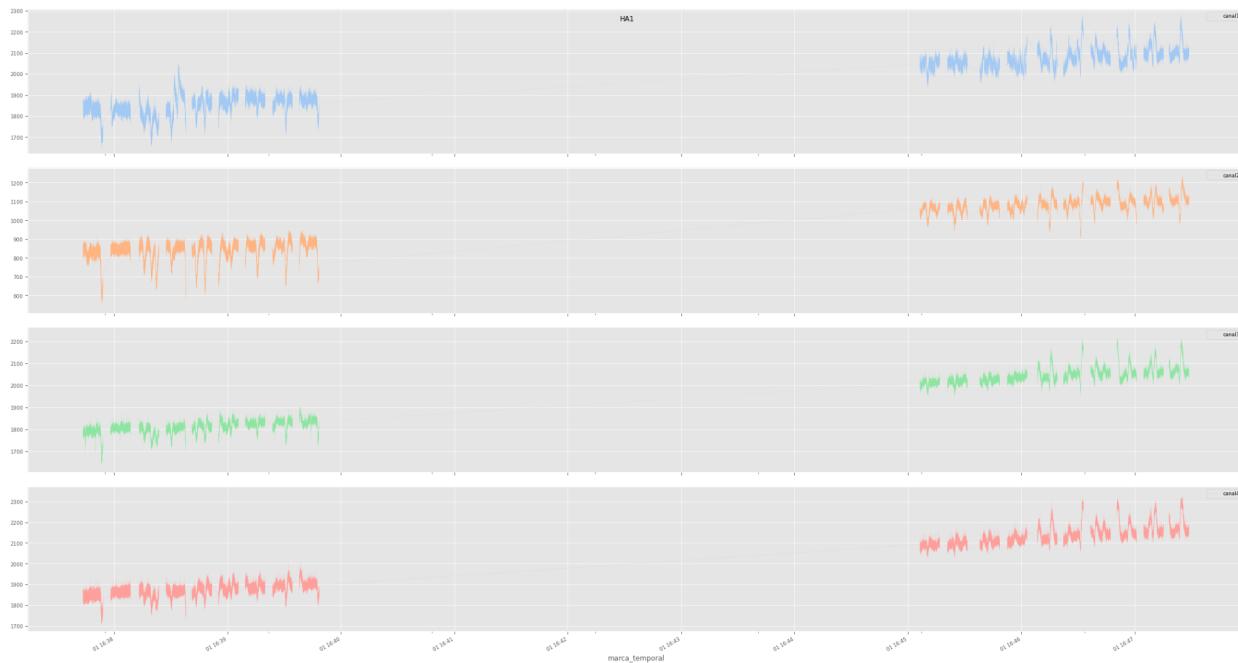
En el siguiente gráfico de voltaje observamos, por ejemplo, para la persona 'AA', sesiones '0' y '1' la diferencia luego de eliminar los registros correspondientes a la etiqueta '99'.

Se puede notar que los rangos de tiempo que corresponden a la etiqueta eliminada se grafican como líneas que unen los voltajes de las etiquetas '1' y '2'.





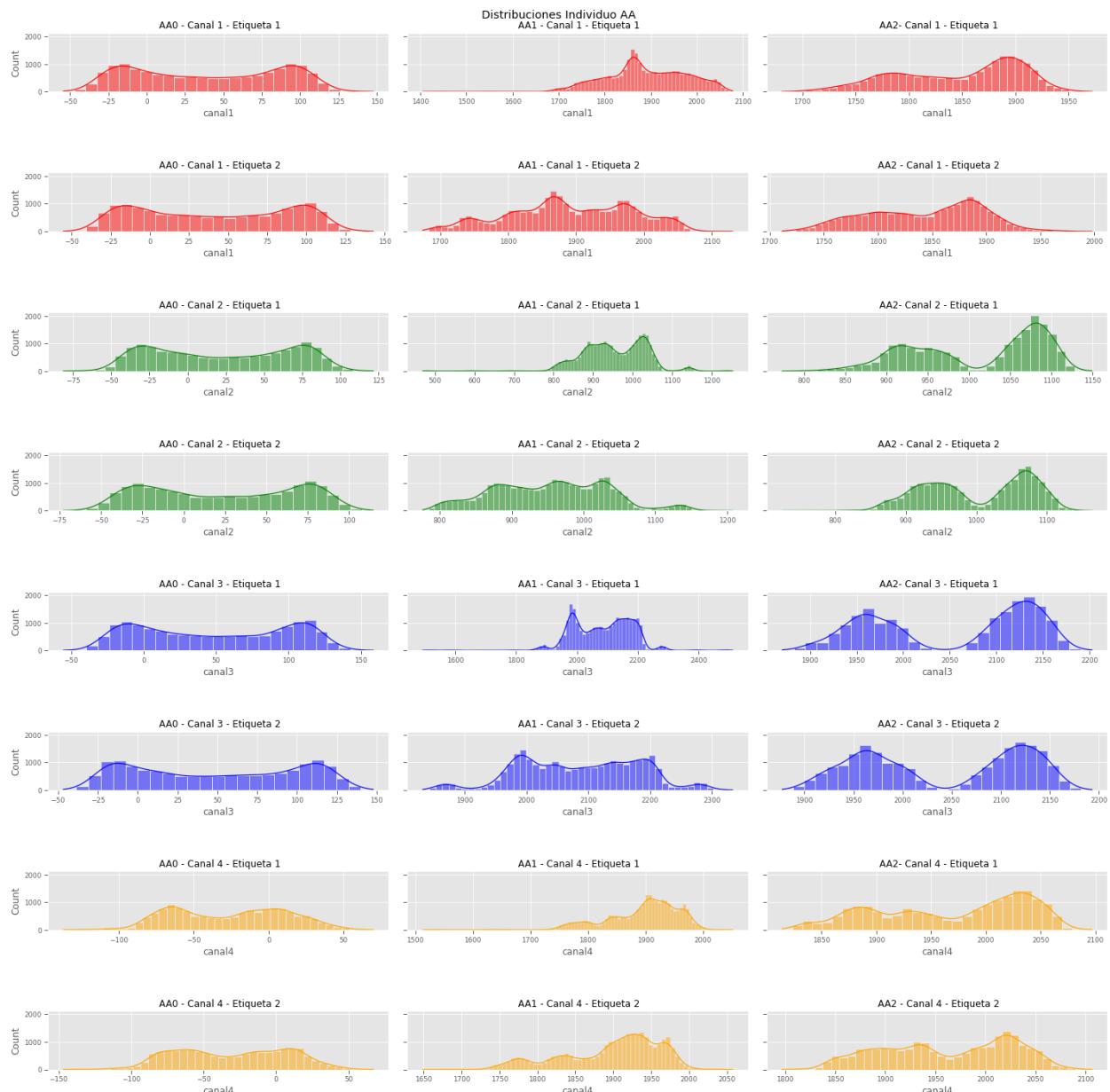
Se puede notar como en el siguiente gráfico, la última señal de color verde (correspondiente a canal 3 del individuo HA) que se produce una deriva en la adquisición de la señal, la cual genera que la gráfica vaya aumentando su valor en el tiempo. Este problema fue resuelto posteriormente quitándole esa componente de continua con su pendiente para poder estandarizar todas las señales.

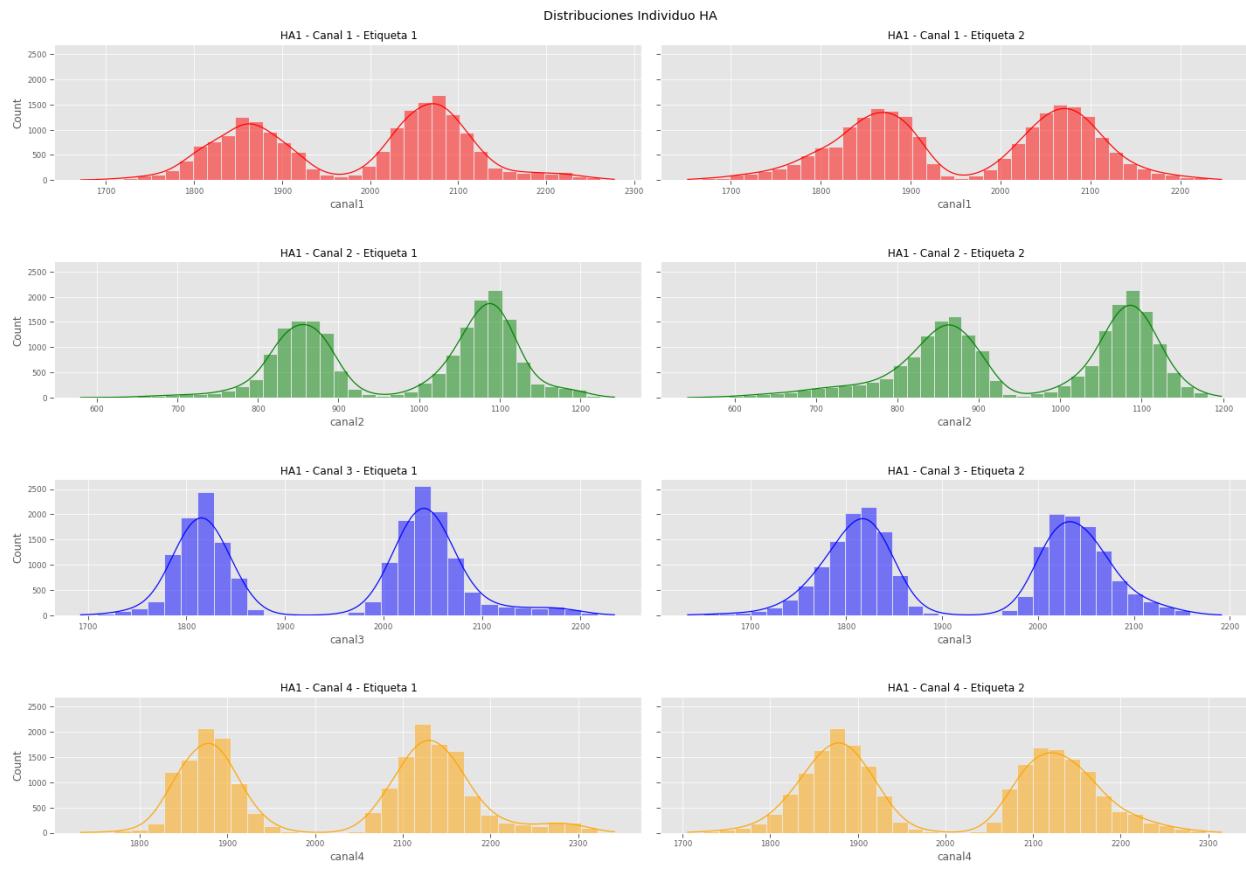


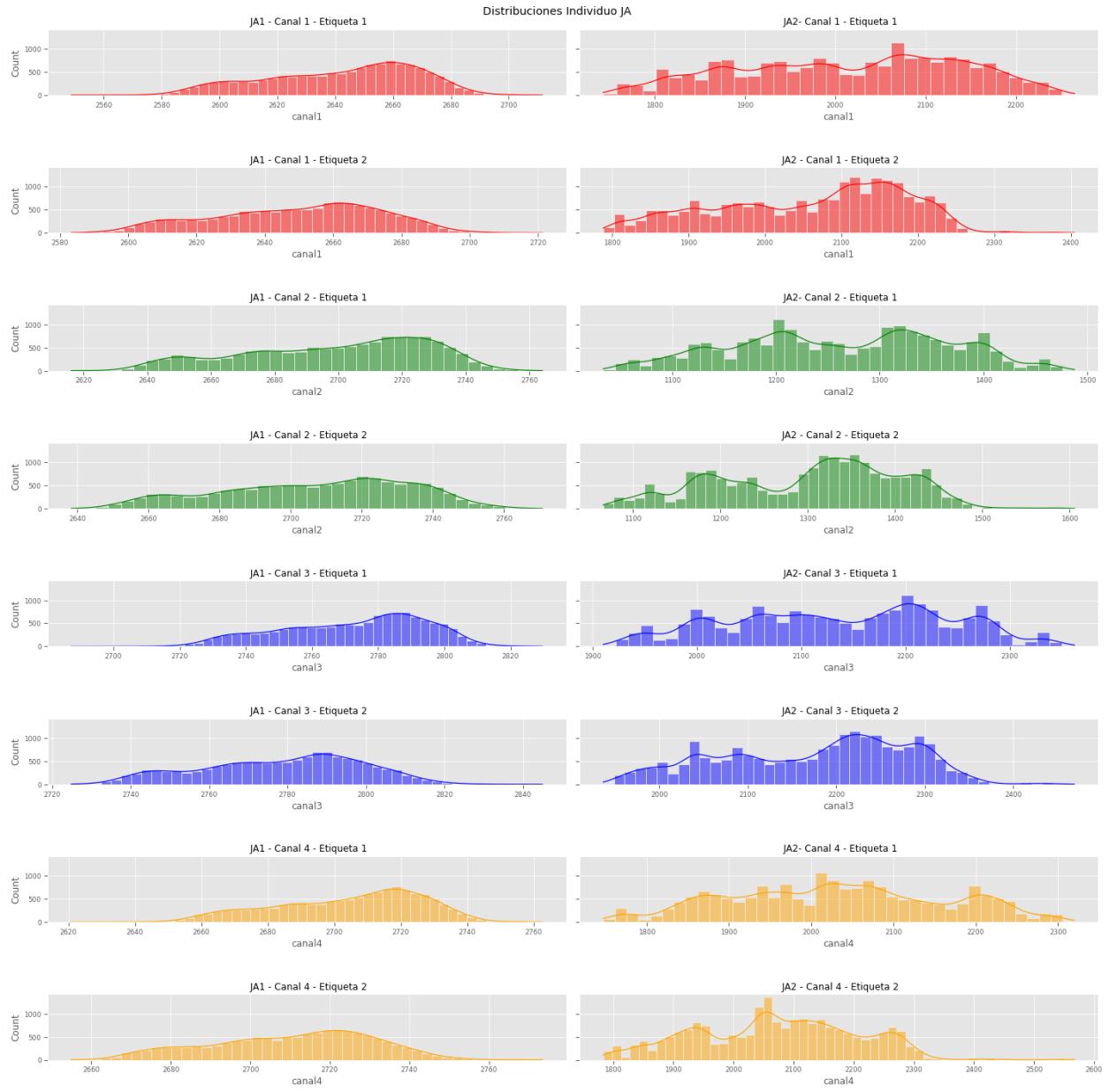
Distribuciones

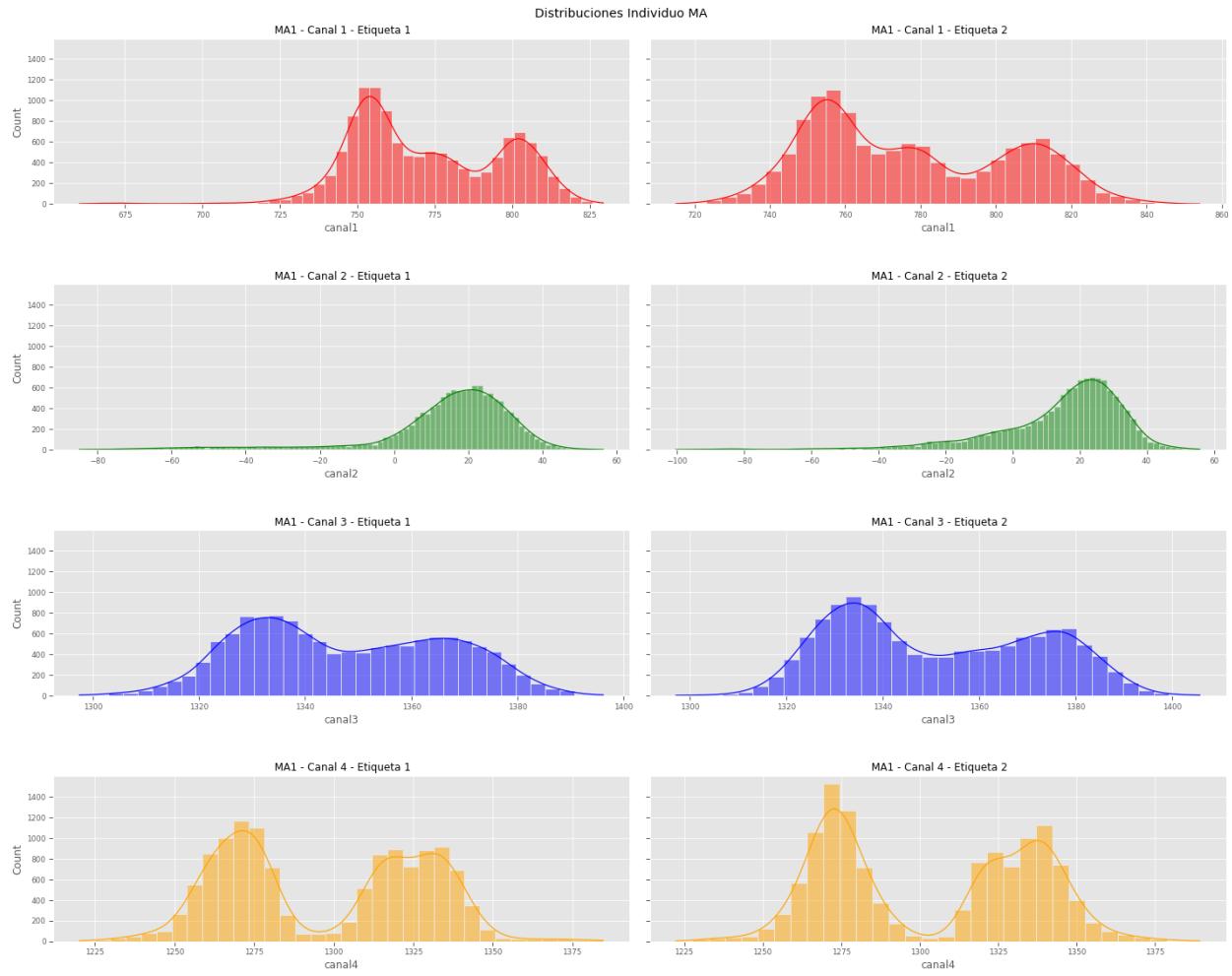
En este punto resulta interesante visualizar las distribuciones de los diferentes dataset, para cada canal de manera completa.

Si bien en su mayoría no parecen presentar distribuciones normales, en algunos casos ('AA2 - Canal 3 - Etiqueta 1' por ejemplo), se ven dos polaridades donde cada una se asemeja a una distribución normal.









Tratamiento de valores extraños

Manteniendo los registros sólo de la etiqueta '1' y '2', se siguen observando rangos de voltajes con valores (min, max) que tienen una gran variabilidad, aún entre mismas personas en diferentes sesiones. También se evidencian diferentes outliers para cada uno de los segmentos mencionados.

Enfocaremos el tratamiento de estos rangos dispares, para cada segmento, mediante:

- Método **"Detrend by Model Fitting"** para llevar los rangos a valores esperados, según el comportamiento de señales EEG (-250, +250).
- **Eliminación de outliers** mediante el método de rangos intercuartílicos.

Método 'Detrend by Model Fitting'

Una tendencia en una serie temporal puede visualizarse fácilmente como una línea que atraviesa las observaciones.

Este enfoque puede ayudar a identificar si una tendencia está presente. Además de utilizarse como herramienta de identificación de tendencias, estos modelos de ajuste también pueden utilizarse para "desdiferenciar" una serie temporal, que es lo que haremos en este caso. Por ejemplo, se puede ajustar un modelo lineal en el índice temporal para predecir la observación. Este conjunto de datos tendría el siguiente aspecto:

X, y
1, obs1
2, obs2
3, obs3
4, obs4
5, obs5

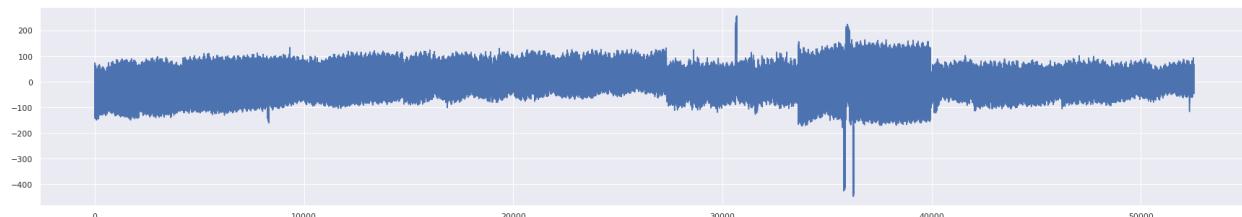
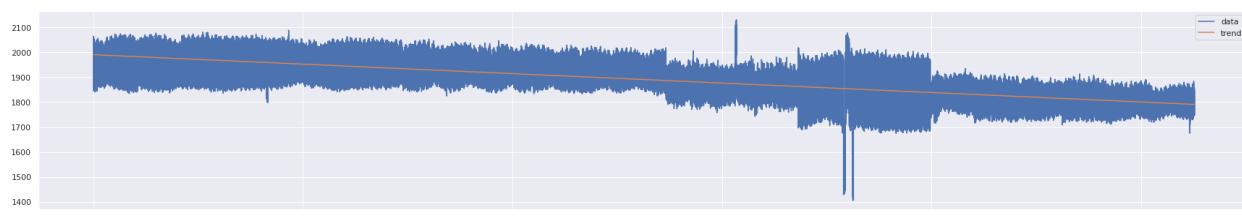
Las predicciones de este modelo formarán una línea recta que puede tomarse como la línea de tendencia del conjunto de datos.

Estas predicciones también pueden restarse de la serie temporal original para obtener una versión sin tendencia del conjunto de datos.

$$\text{value}(t) = \text{observation}(t) - \text{prediction}(t)$$

Los residuos del ajuste del modelo son una forma de eliminación de tendencias del dataset.

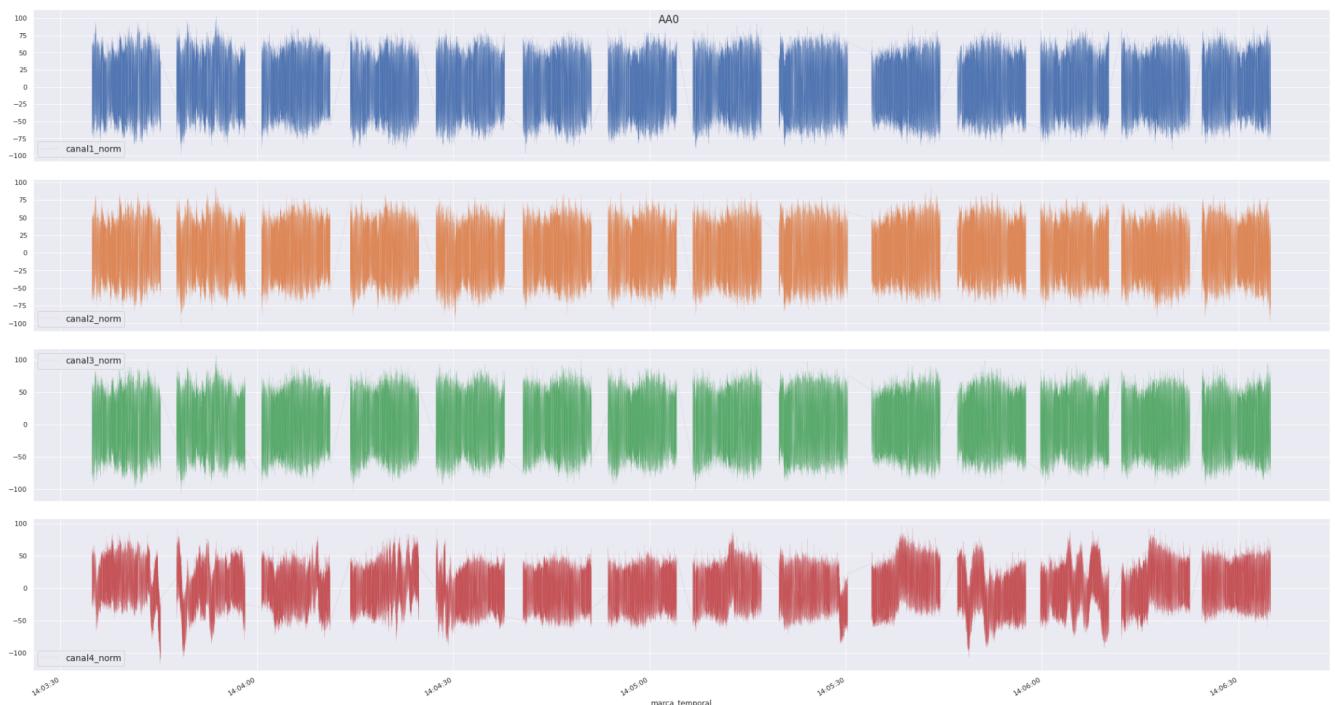
En los siguientes gráficos se ve primero un ejemplo de línea de tendencia para una medición de voltajes y luego la misma señal luego de restarle la línea de tendencia, donde se aprecia que los rangos de voltajes han disminuido significativamente.

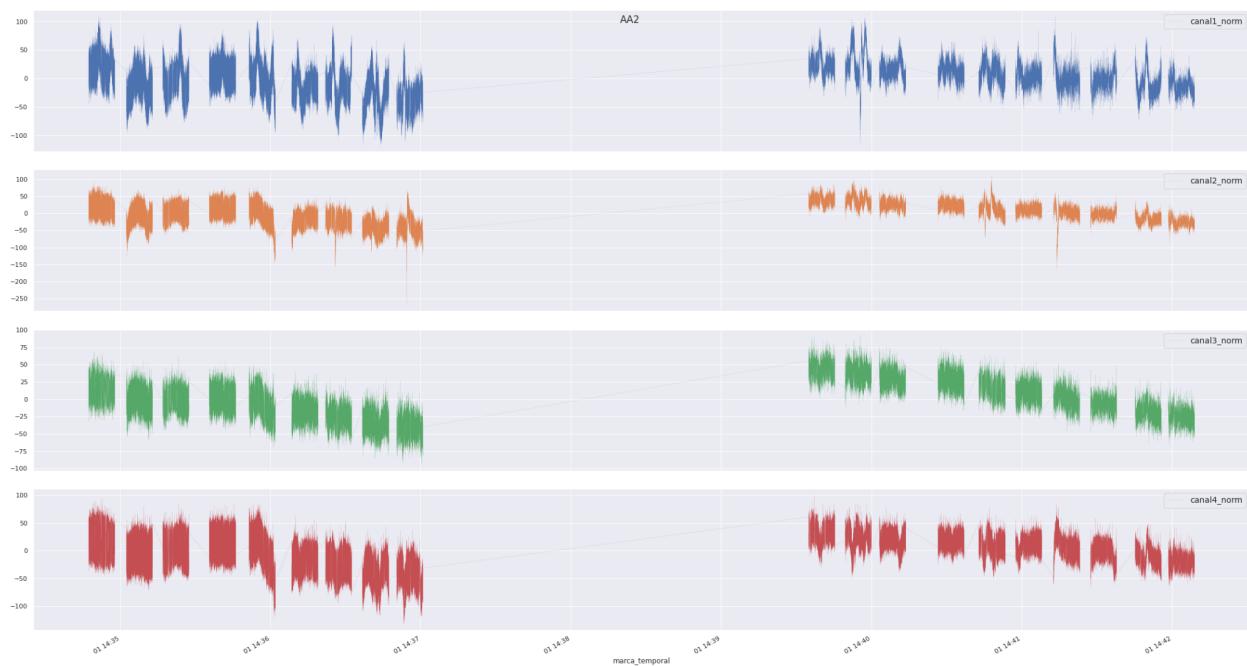
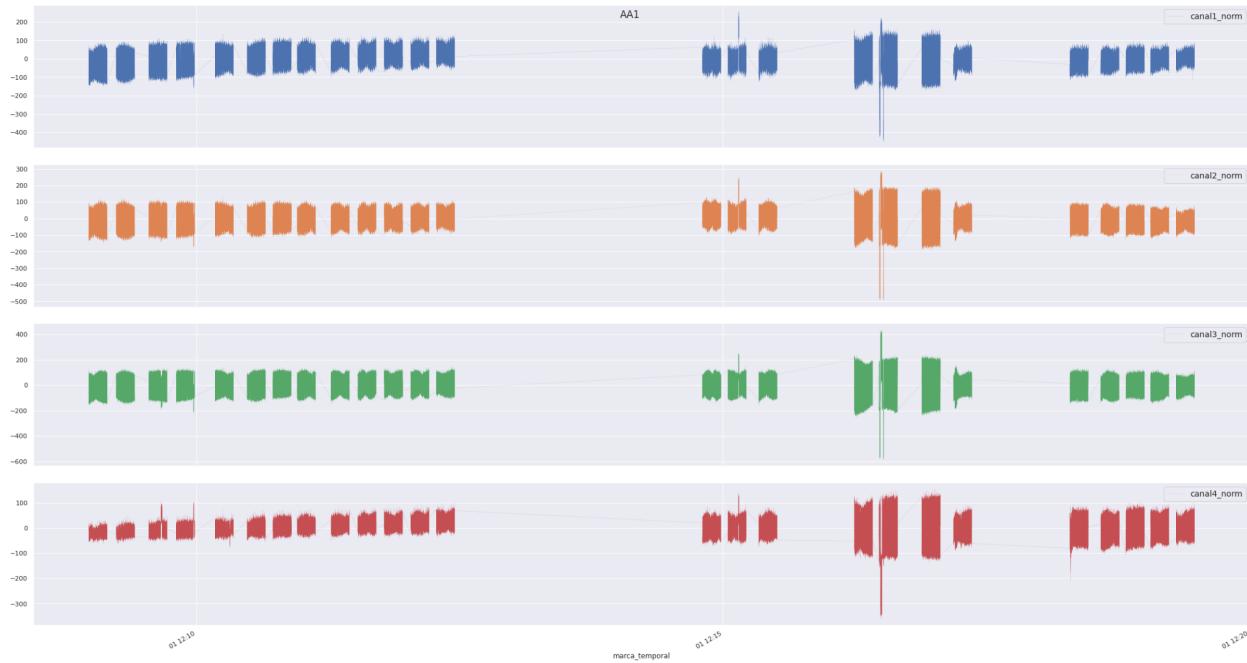


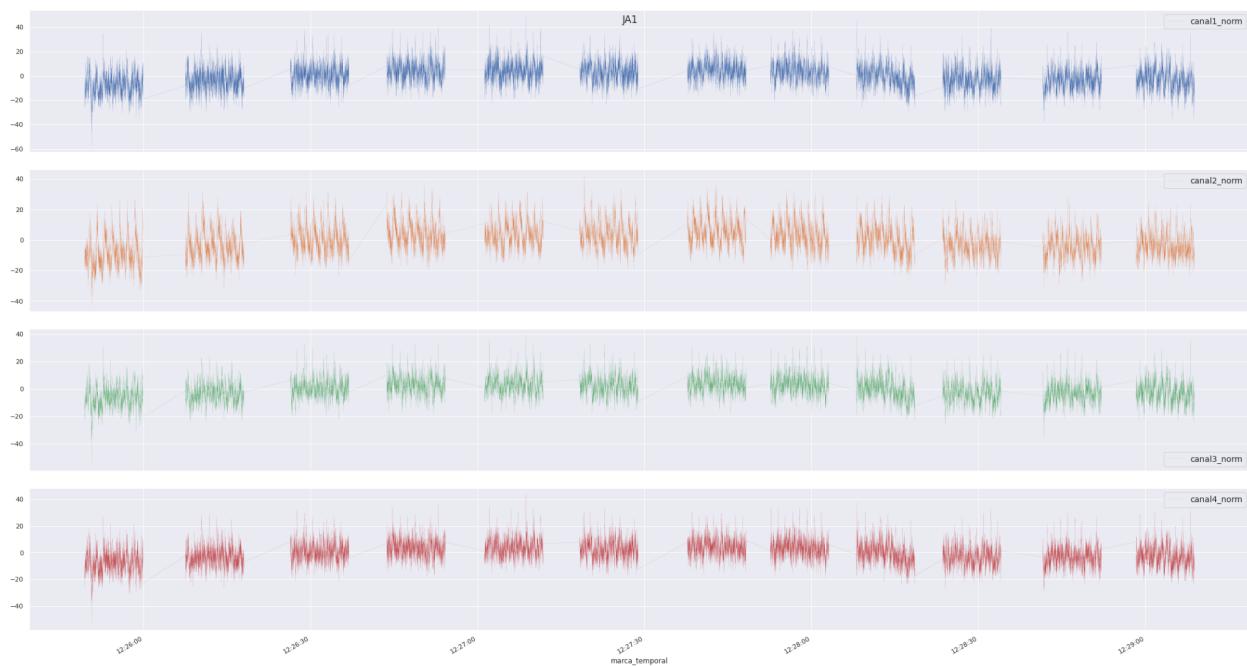
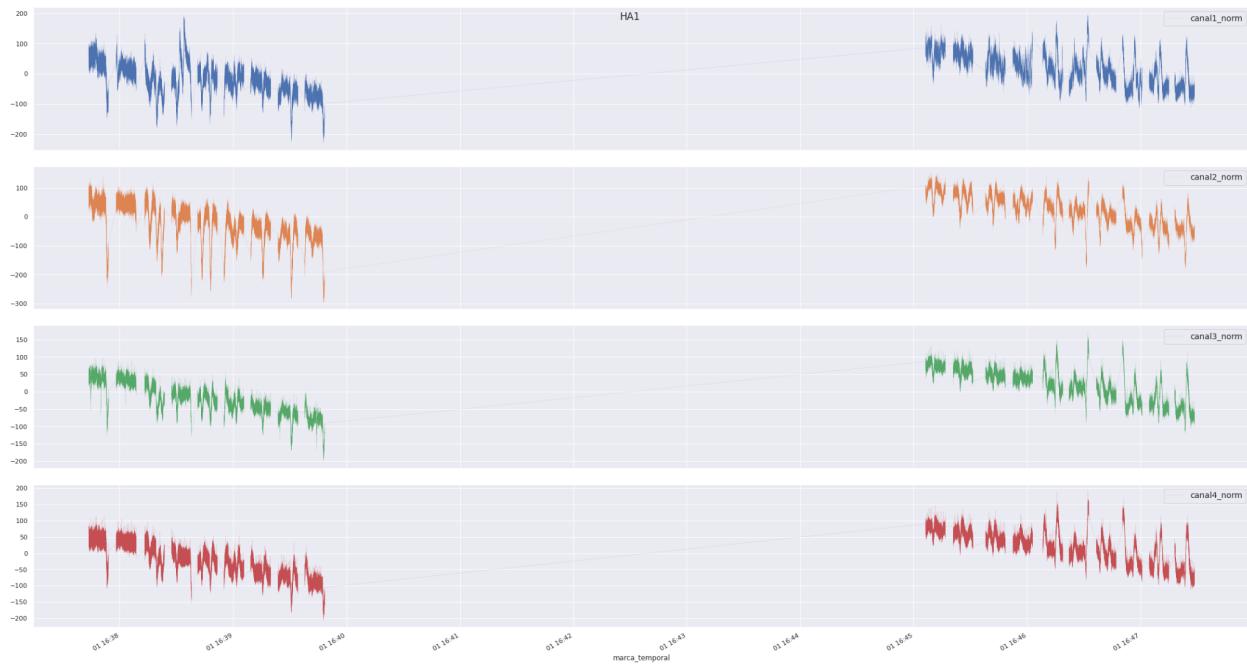
Para los dataset en estudio hemos implementado este método en Python, entrenando un modelo "LinearRegression" de scikit-learn. Obteniendo un nuevo rango de valores para las señales en cada segmento.

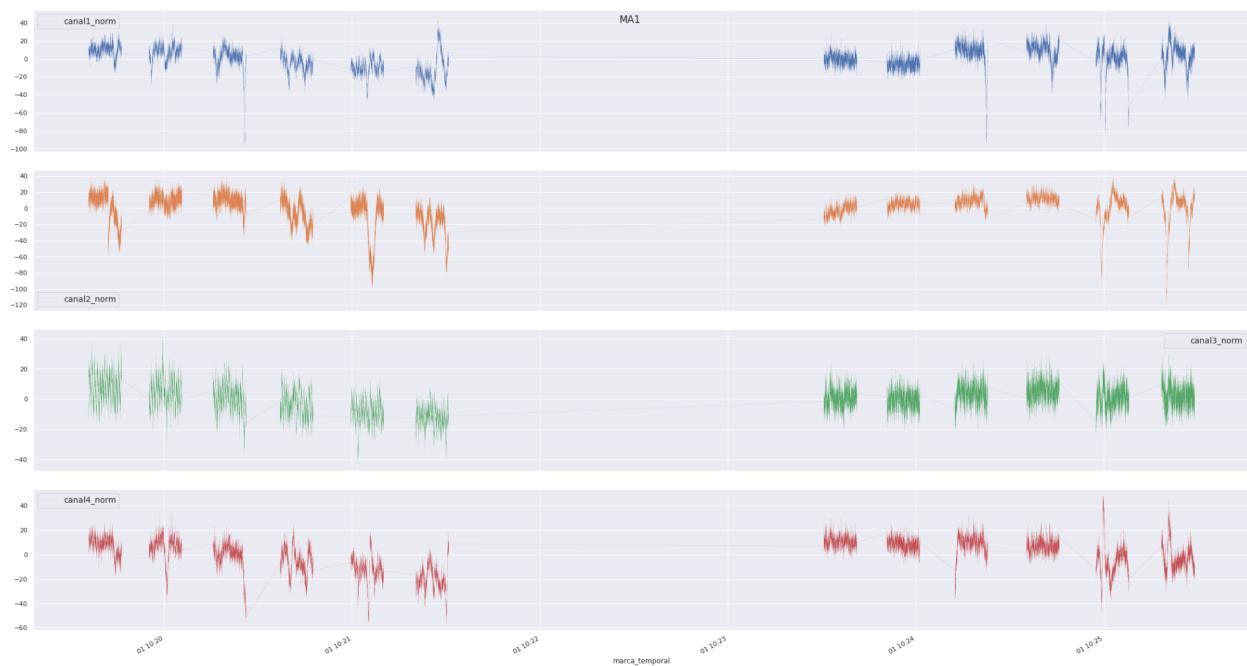
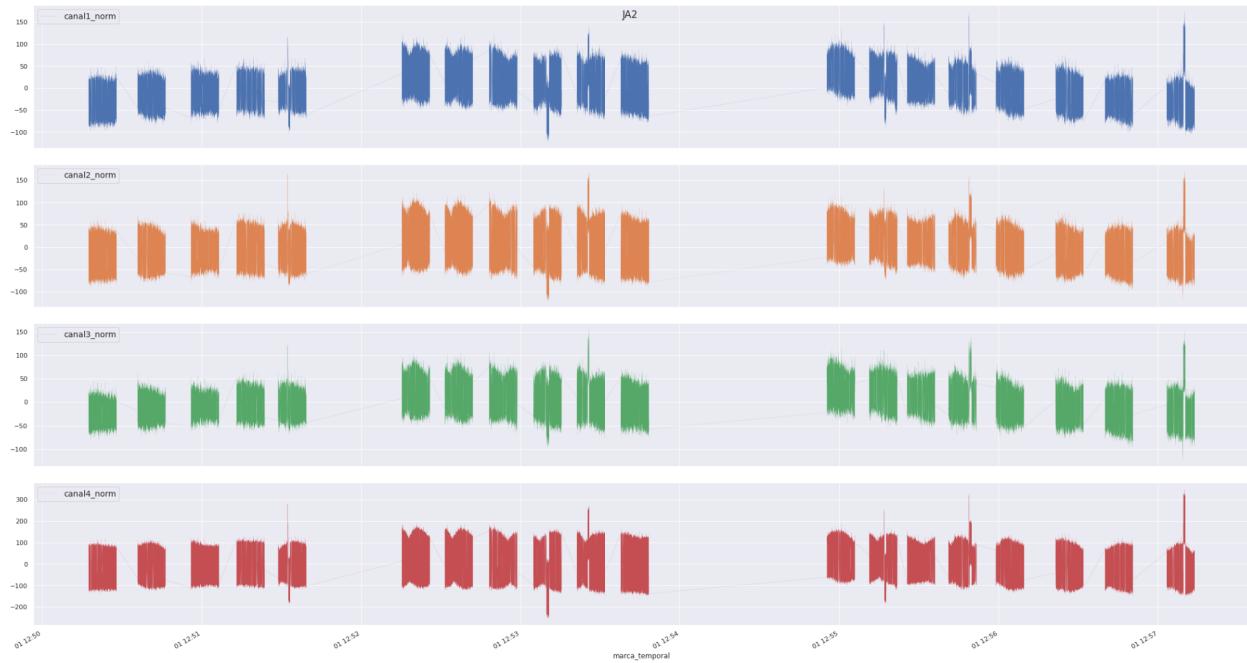
Voltajes de las señales después de restar las líneas de tendencia

A continuación se presentan las gráficas de los voltajes de cada canal en el tiempo por segmento (persona/sesión), después de aplicar la resta de la línea de tendencia. Donde se puede evidenciar la disminución de los rangos de μ V y la sustracción de la deriva mencionada anteriormente.





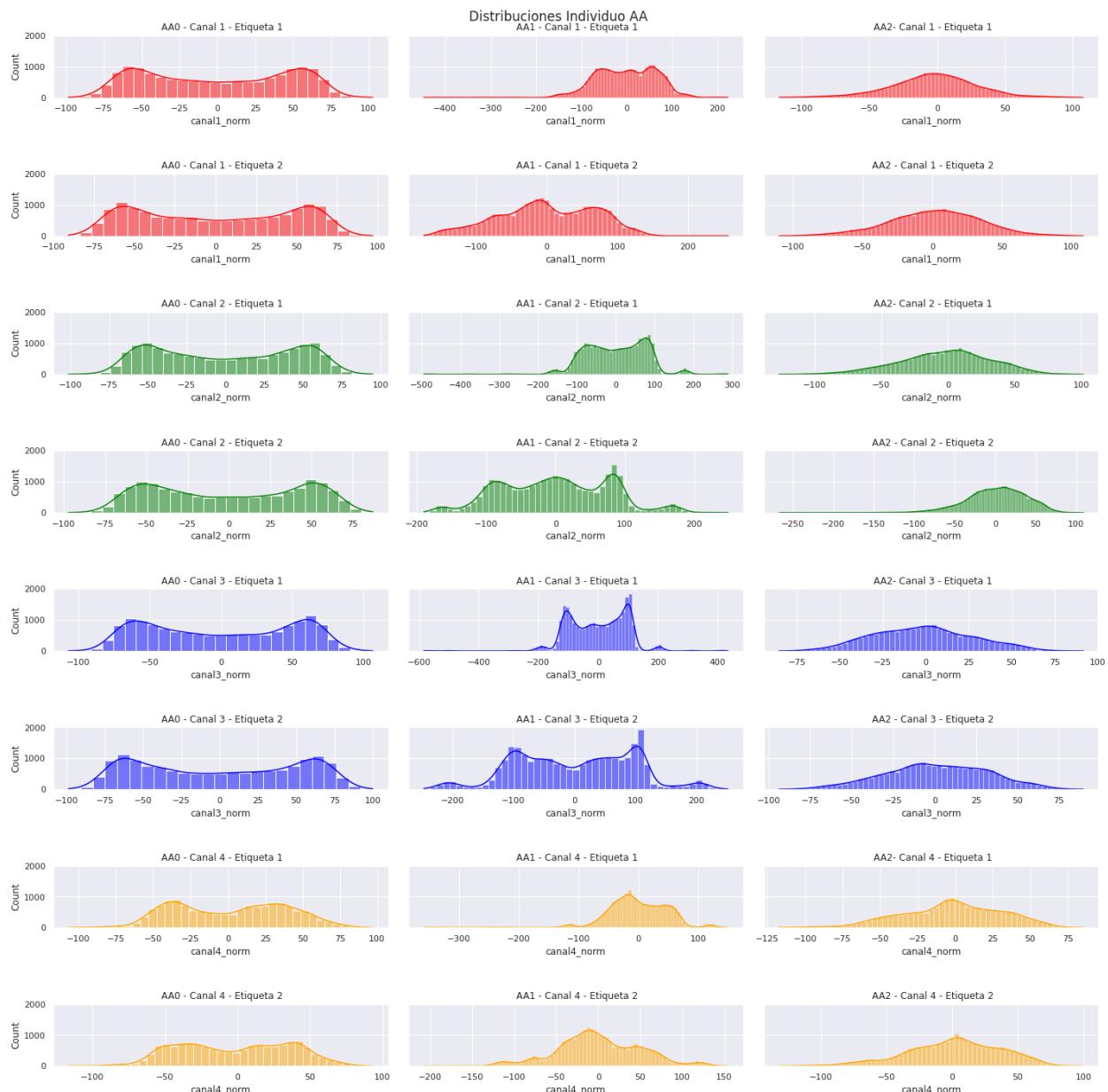


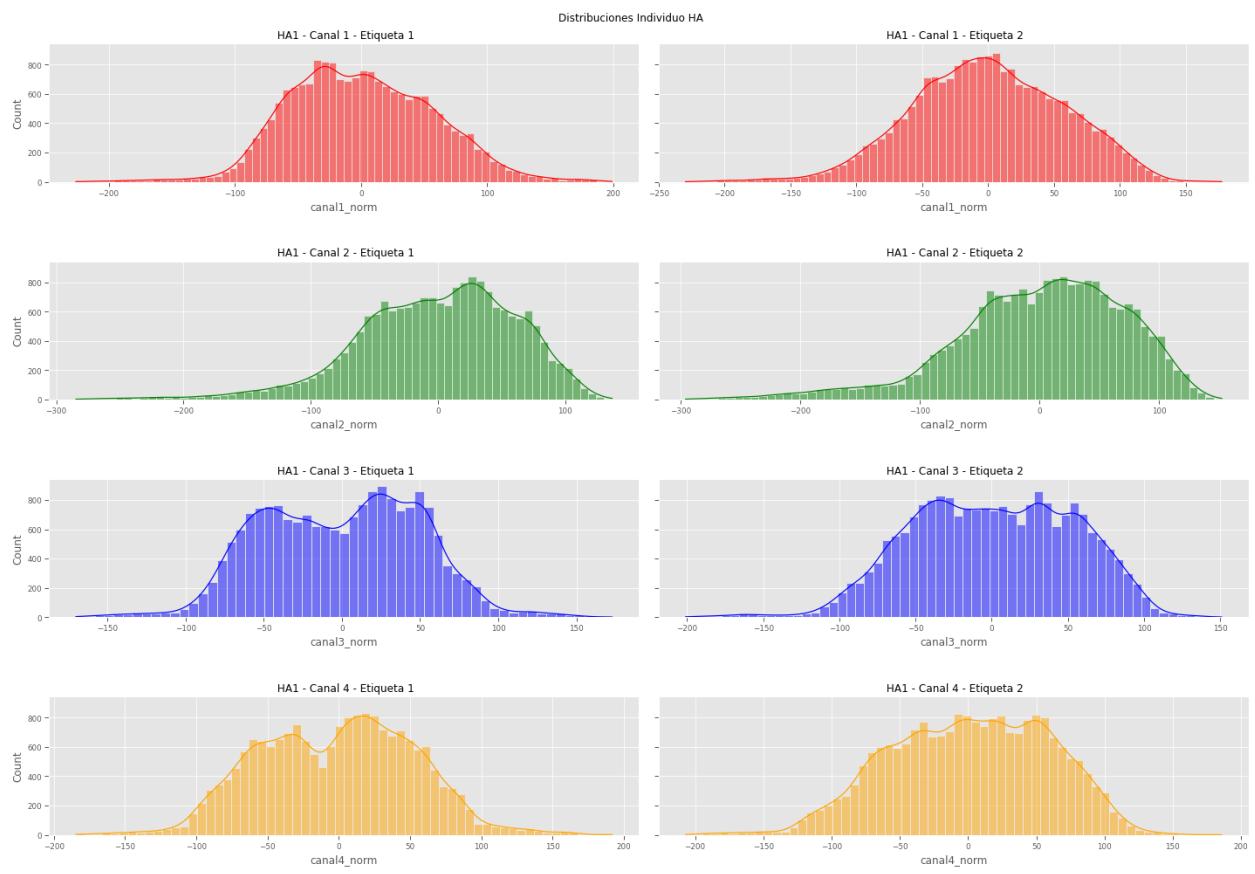


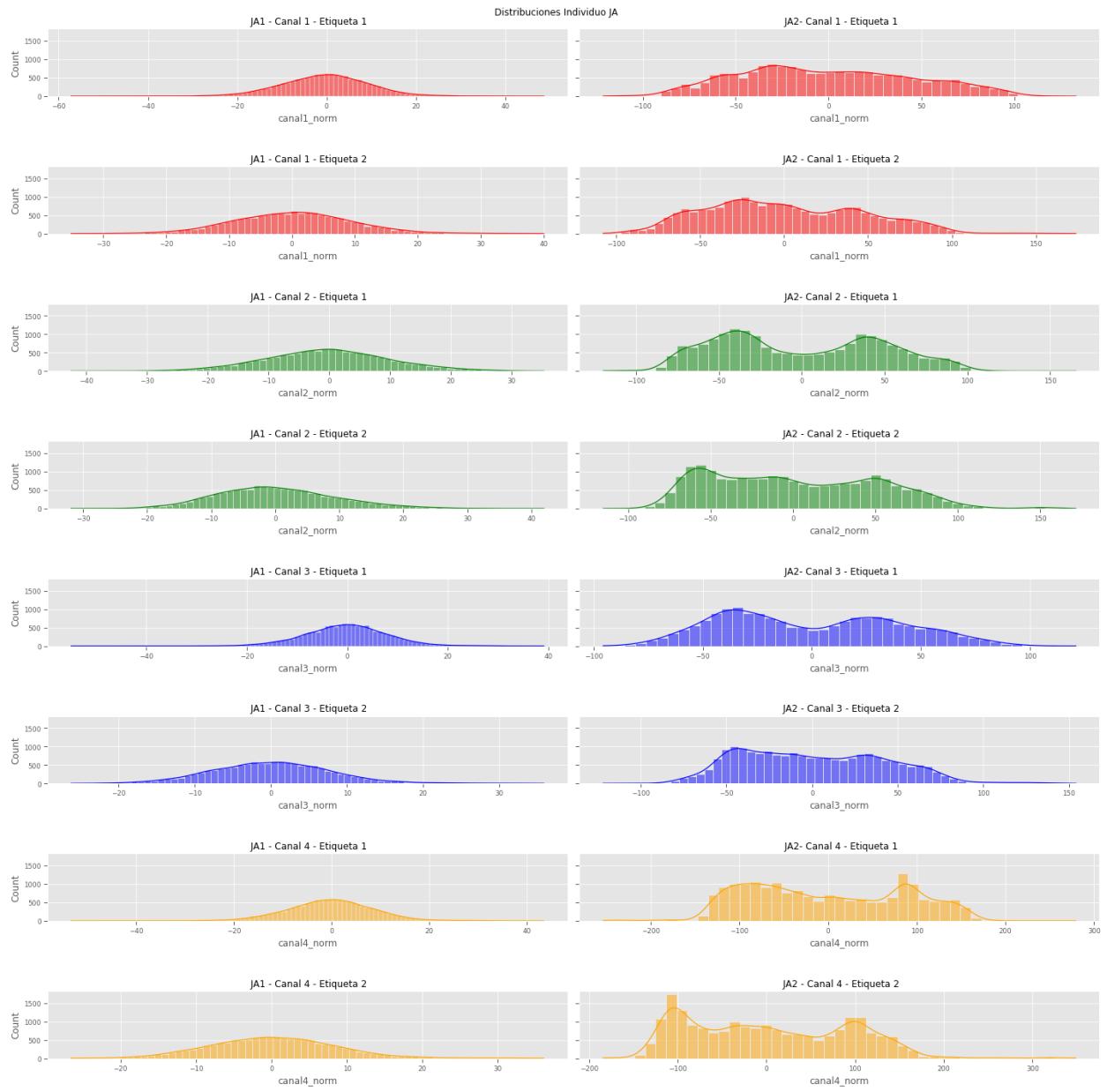
Distribuciones después de restar las líneas de tendencia

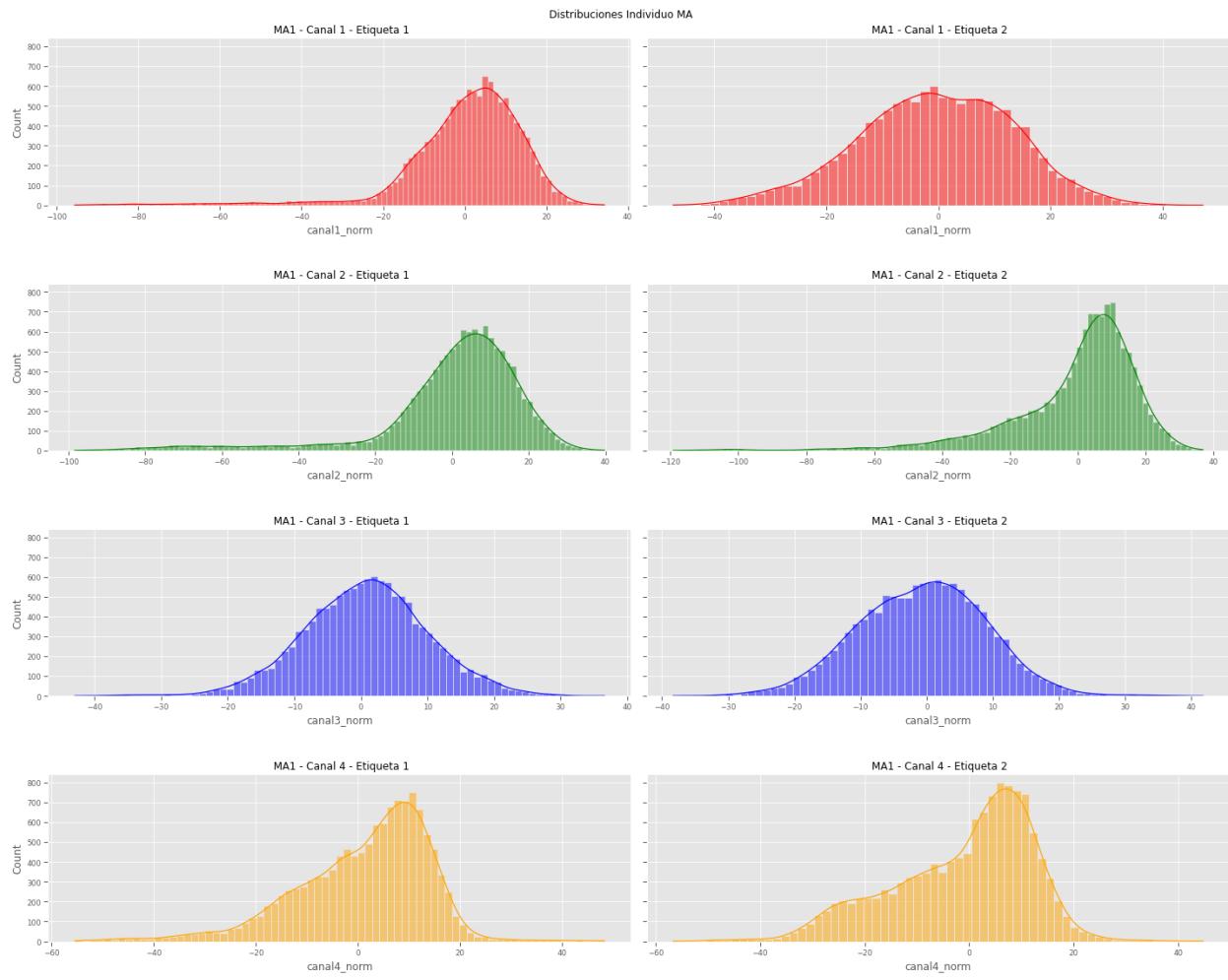
Se grafican a continuación las distribuciones de los valores de voltaje luego de la resta de las líneas de tendencia por canal en cada set de datos.

Como se puede apreciar en los gráficos, las distribuciones ya no se aprecian con dos polaridades con distribuciones casi normales. Ahora se han vuelto algunas de apariencia casi de una distribución normal. Sin embargo cabe destacar que se está visualizando todos los segmentos de los canales, por ende lo qué está pasando es que se produce una mezcla de gaussianas producidas por cada intervalo y que generan las gráficas que se ven a continuación.



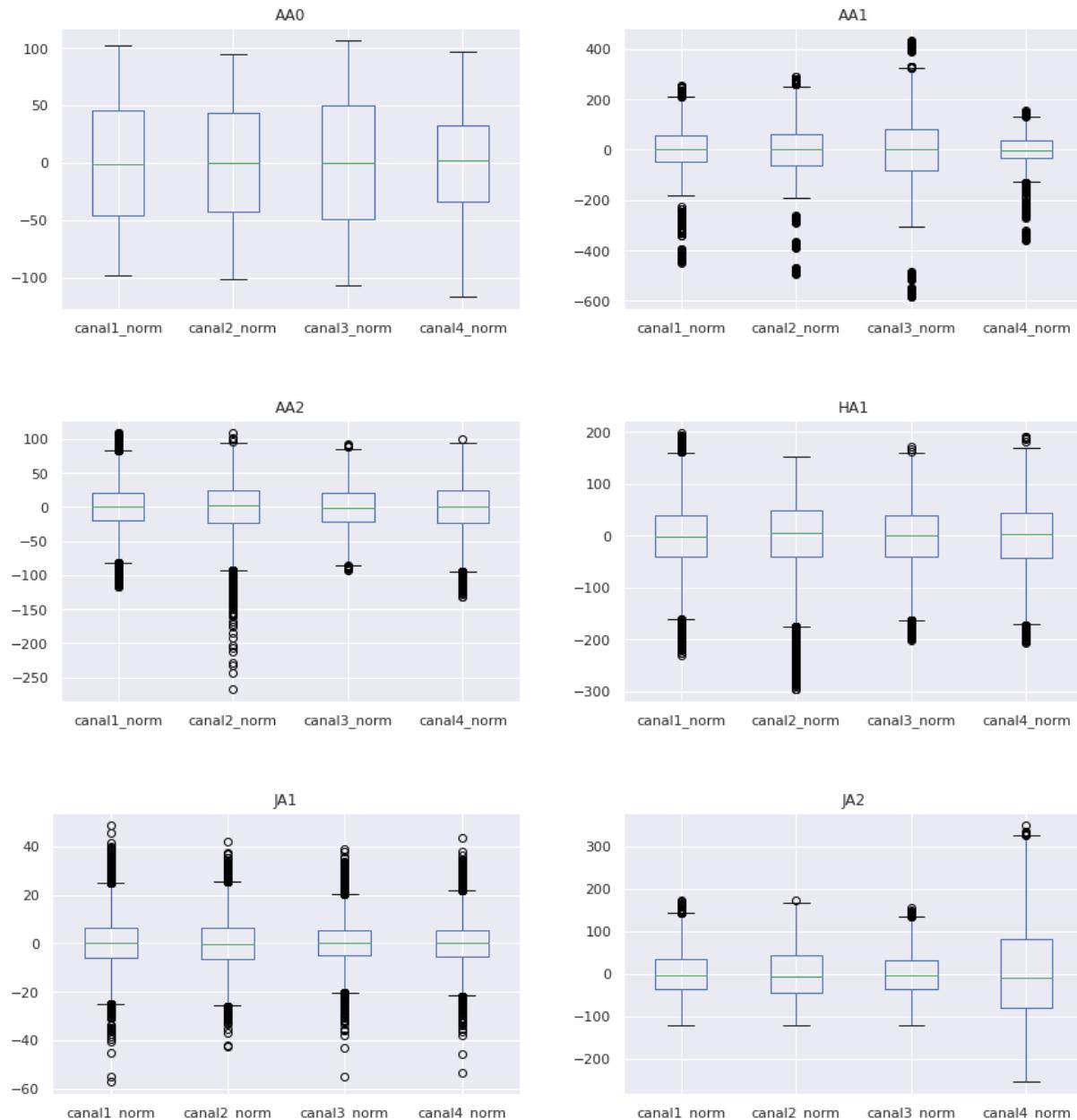


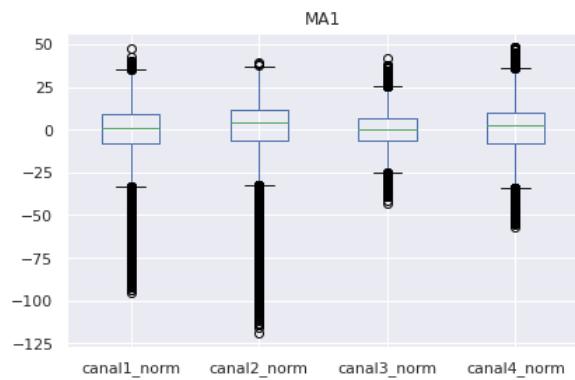




Eliminación de Outliers

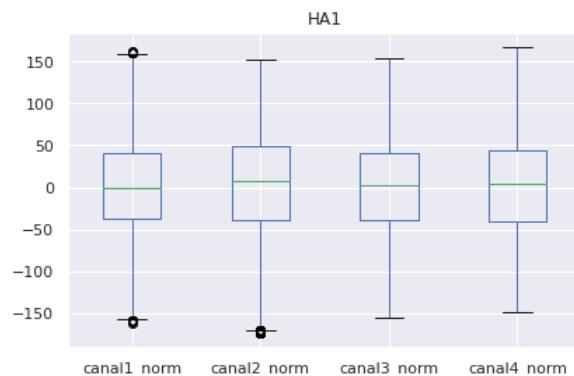
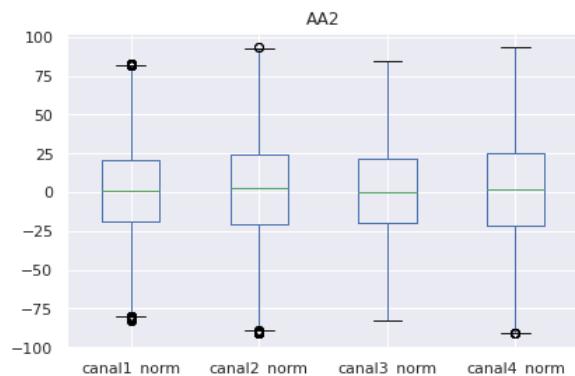
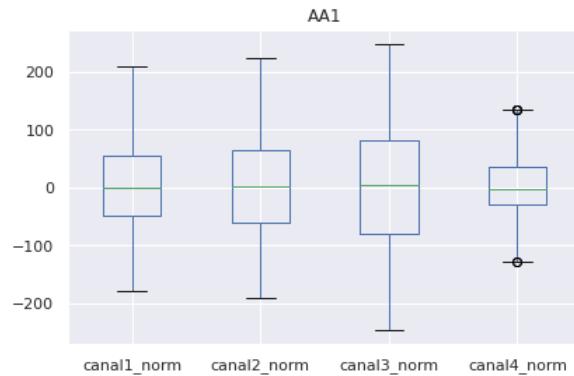
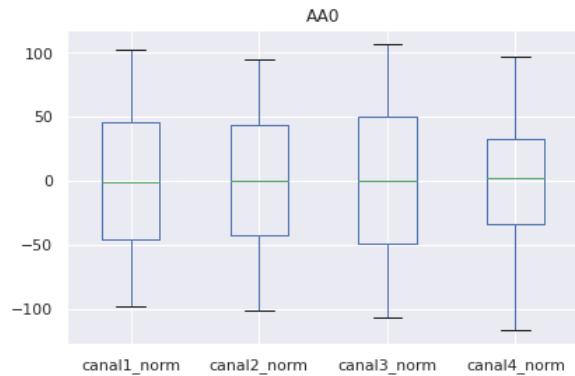
Mediante diagramas ‘boxplots’, graficamos cada dataset, donde podemos ver en forma de círculos negros los valores que se alejan de la concentración de los valores de voltajes, estando por fuera de los rangos intercuatílicos.

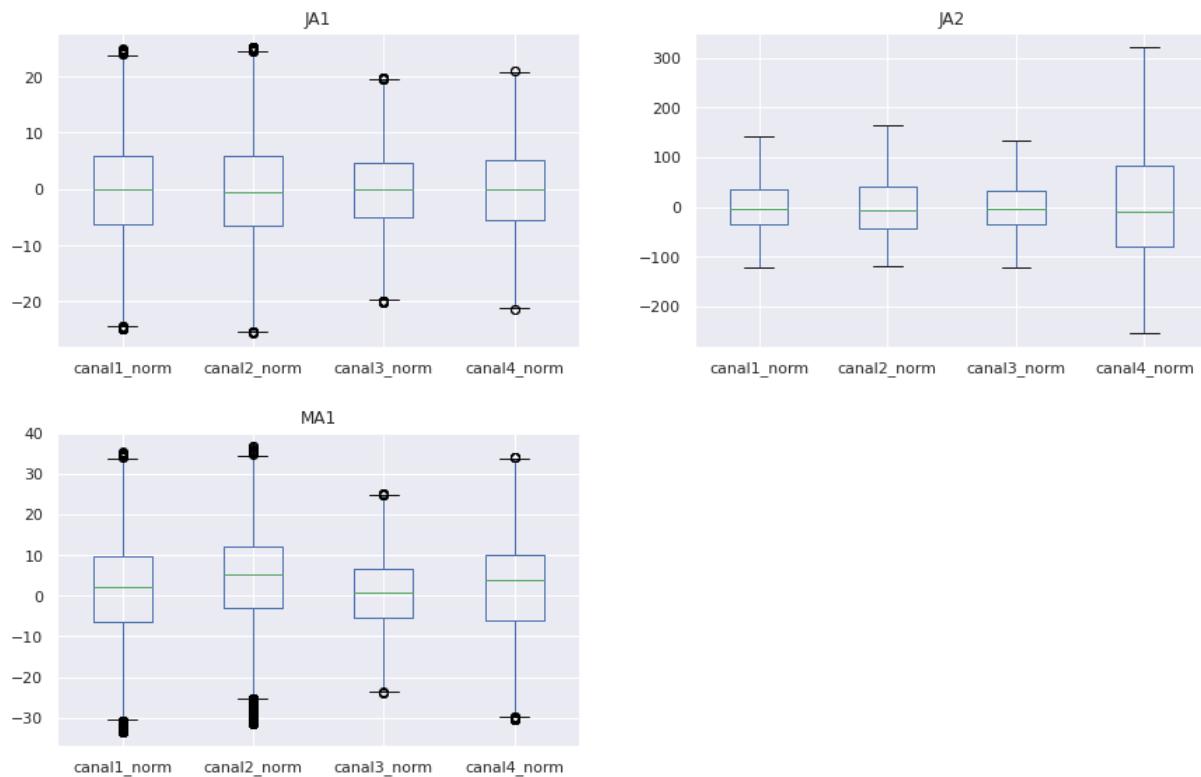




Utilizamos el método de rango intercuartílico (IQR) para eliminar los outliers identificados.

Luego de aplicar la eliminación de outliers, se tienen los siguientes boxplots:





Se verifican que los rangos de voltajes disminuyeron al eliminar outliers (-200 μ V a 300 μ V), lo cual resulta razonable para estudios de EEG.

Luego de eliminar outliers por rango IQR, se tienen las siguientes cantidades de registros:

AA0: 29448 (Registros eliminados 0)
 AA1: 52225 (Registros eliminados 316 => 0.6%)
 AA2: 38967 (Registros eliminados 992 => 2.48%)
 HA1: 39107 (Registros eliminados 858 => 2.14%)
 JA1: 24456 (Registros eliminados 785 => 3.11%)
 JA2: 39881 (Registros eliminados 45 => 0.11%)
 MA1: 22782 (Registros eliminados 2475 => 9.79%)

Quedan 246.866 registros. Cantidad de registros eliminados: 5471 de 252337 (2.16%).

Observaciones:

- Haciendo un análisis de los datos y con conocimiento de dominio evaluamos la posibilidad de que los datos tuvieran "artefactos/ruido" que hacen que estuvieran fuera de los rangos esperados en su mayoría, ya que las señales electroencefalográficas rondan los 200 micro Voltios. Por ende los valores de 2800 son valores que en principio nos llamó la atención pensando qué podía ser ruido. Como se observa en los boxplots,

todo el dataset contiene valores muy por encima de los normales. Por ende se decidió restarle la línea de tendencia de cada canal en particular de cada individuo y sesión, para lograr un análisis más correcto.

- **¿Qué tipo de datos contienen? ¿Qué variables describen las columnas consideradas? ¿Con qué sensibilidad?**

```
num_muestra      int64
canal1           float64
canal2           float64
canal3           float64
canal4           float64
etiqueta         int64
marca_temporal   object
```

Luego se modificó la columna tiempo a formato datetime 64[ns] para poder realizar los gráficos de la señal en función del tiempo.

Las columnas de los canales muestran el voltaje de la señal muestreada con una sensibilidad de micro Volt

La variable tiempo cuenta con una sensibilidad de milisegundos.

- **En el caso de los datos cualitativos. ¿Cuáles son los valores posibles para esta variable? Describa su presencia (frecuencia, intervalos, secuencia, etc.)**

La variable etiqueta tiene datos cualitativos donde tiene dos valores posibles: un 1 para la luz parpadeante a 12,5 Hz o un 2 para la de 16,5 Hz. Los valores con etiqueta 99 fueron removidos ya que es cuando el individuo no observa ninguna luz.

Los datos cualitativos son:

- 'etiqueta' con valores:
 - 1: looking left;
 - 2: looking right;
- 'person'
- 'sesion'

- **Para todas las columnas, ¿hay datos dañados? ¿valores nulos? ¿Qué estrategia considera más pertinente para abordar esos datos? Justifique.**

En nuestro dataset no se encontraron valores nulos en las variables de interés.

En caso de haber casos nulos en las variables de los canales la estrategia más pertinente sería imputarlos con la interpolación entre el valor anterior y el siguiente.

Para el caso de un valor nulo en la etiqueta se podría ver cuál es la etiqueta de la muestra posterior y anterior, si ambas coinciden agregar ese valor y si no coinciden agregar un 99.

- **Suponiendo que los datos se adquieren a una frecuencia de muestreo exacta 200Hz, ¿cómo se manifiesta esta información en el número de muestras presentes en el registro?**

Se va a tener 200 muestras por segundo (1 muestra C/5ms), por ende si fueron estimulados durante 10 segundos seguidos se debe tener 2000 muestras con la misma etiqueta por cada estimulación. Es decir que con la frecuencia de muestreo y la cantidad de muestras se puede determinar cuánto tiempo el individuo fue estimulado con las luces y cuánto tiempo no se registró actividad.

- **Determine la forma más adecuada de parsear los datos temporales para poder graficar las señales en el dominio del tiempo.**

Los datos temporales se pueden parsear de manera por número de muestras teniendo en cuenta la frecuencia de muestreo (200 Hz) multiplicando por la cantidad de segundos que se quieran obtener. El inconveniente de esta técnica es que el timestamp contiene valores repetidos en un número de muestras, por ende no sería tan exacto en términos teóricos aunque en términos prácticos no se observa diferencia.

Otra forma sería poner una muestra cada 5 mS garantizando tomar una sola muestra por cada muestreo sacando el problema del timestamp.

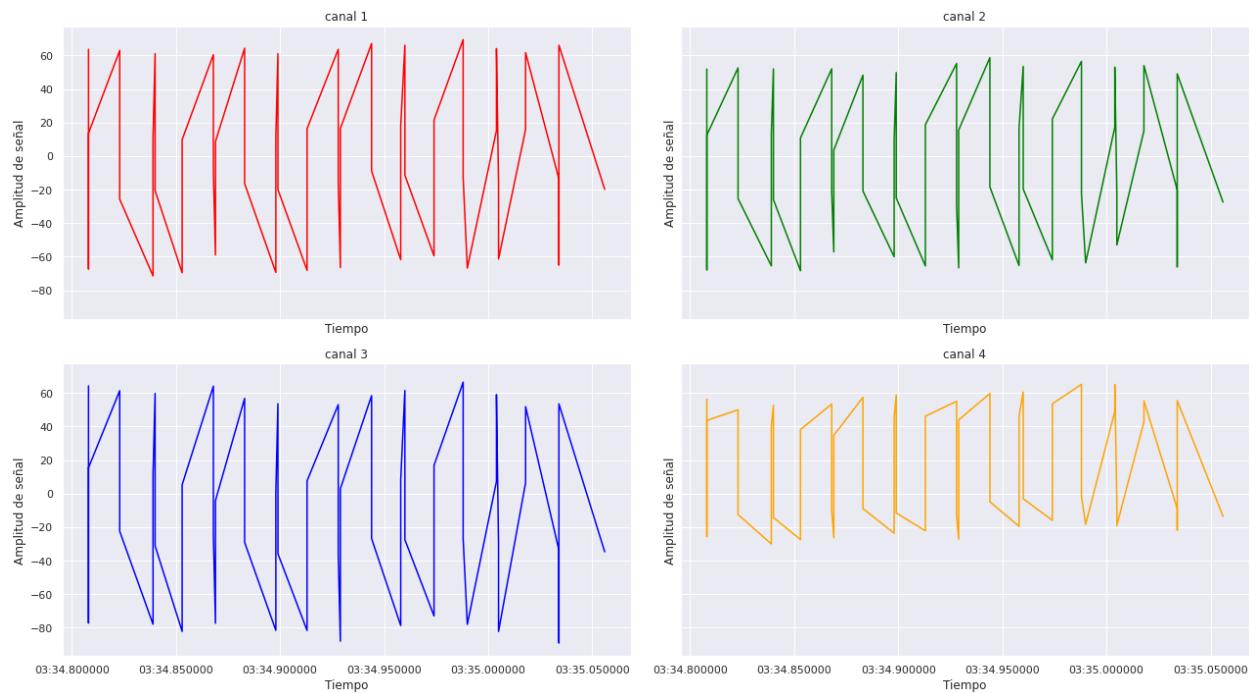
En el análisis propuesto se utilizaron los datos de datetime sin realizar transformaciones.

D) Generar visualizaciones de ejemplo para las series temporales provistas. Determinar los intervalos de tiempo más adecuados para generar visualizaciones claras que permitan comparar las señales en los siguientes escenarios:

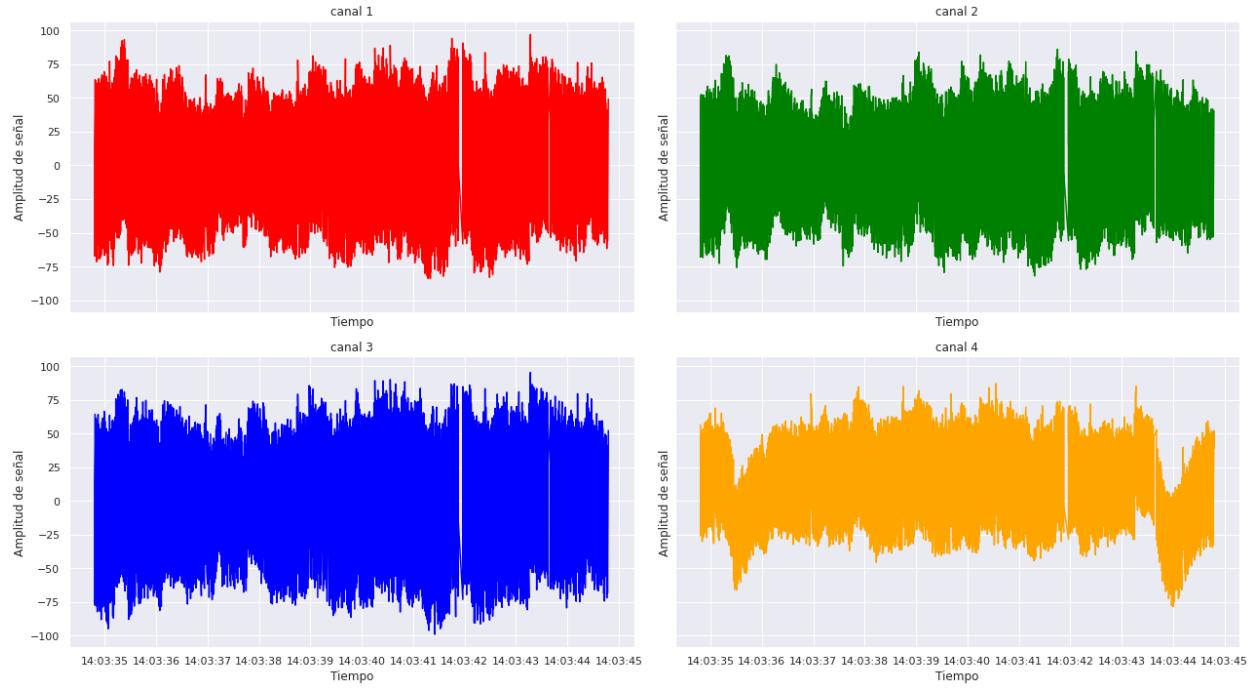
○ **Un sujeto - todos los canales**

Se elige el sujeto "AA" y la sesión '0'.

Usando sólo 50 muestras



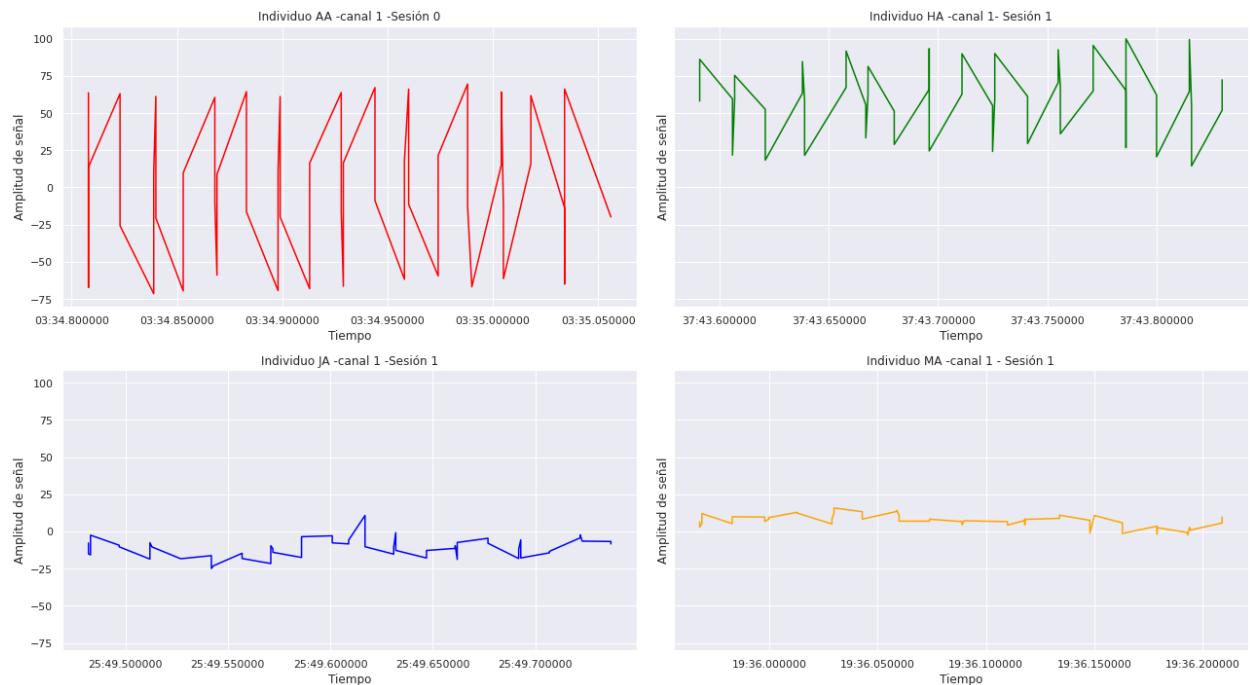
Usando 2000 muestras.



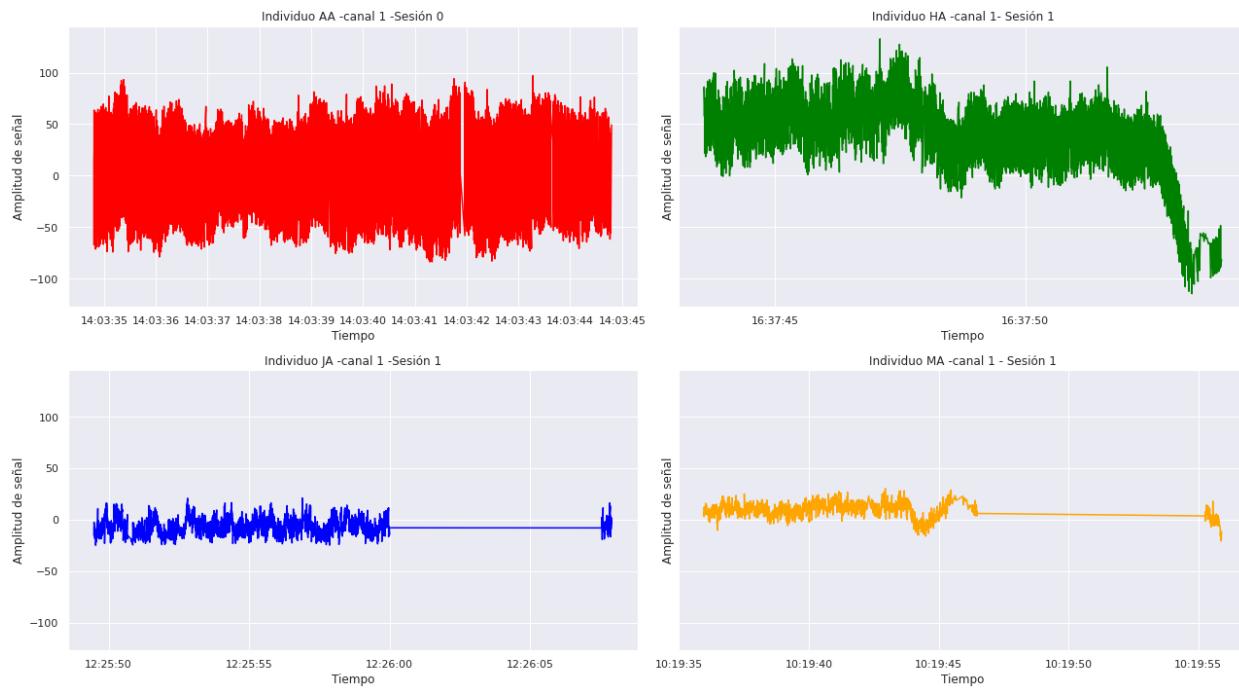
- **Un mismo canal - todos los sujetos**

Se elige el canal 1 para graficar.

Usando 50 muestras:

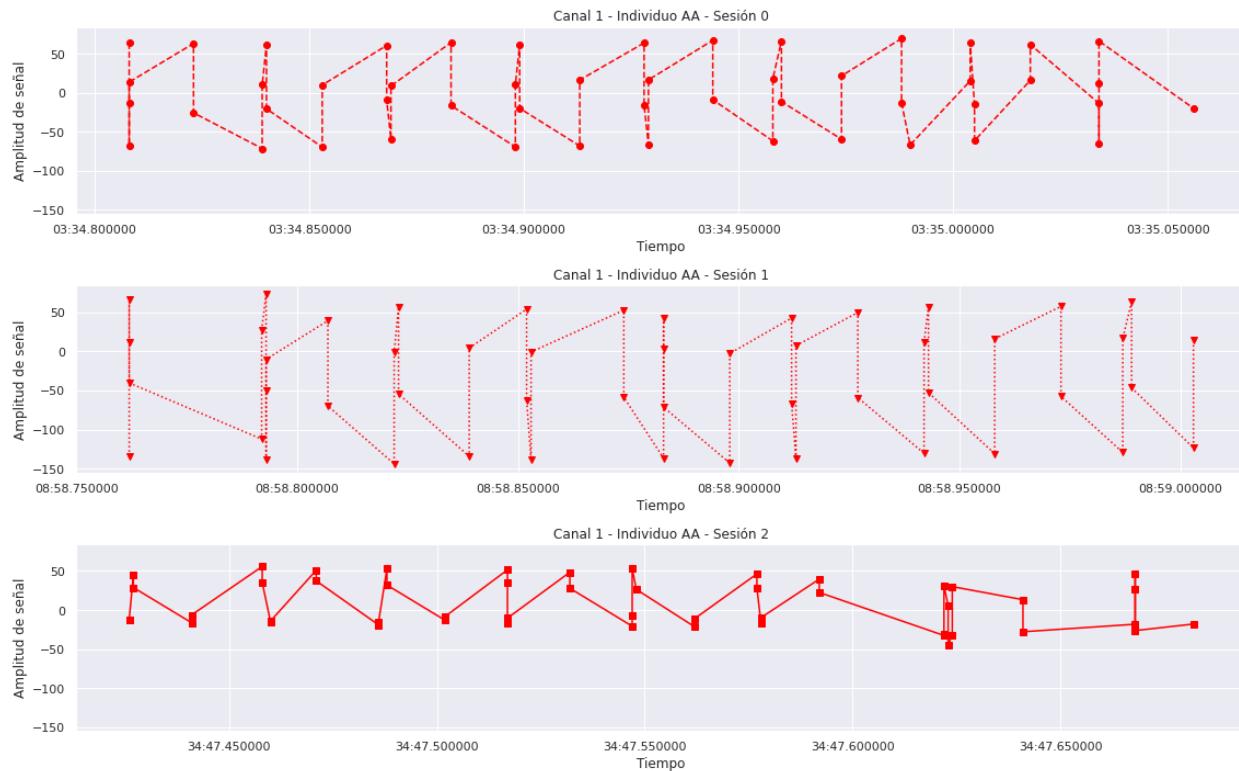


Usando 2000 muestras:



- Un mismo canal - mismo sujeto en diferentes sesiones.

Usando 50 muestras:



Usando 2000 muestras:



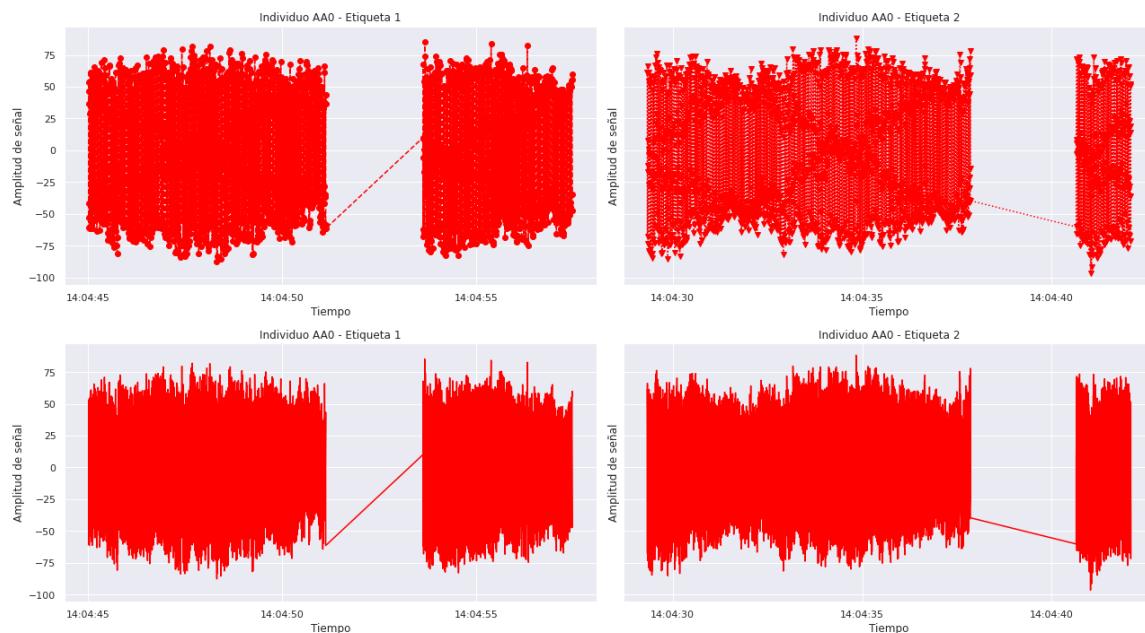
- **Un mismo sujeto y canal - diferentes estados**

Diferentes estados, se refiere a las dos etiquetas posibles:

- 1: para la luz parpadeante a 12,5 Hz;
- 2: para la luz parpadeante a 16,5 Hz.

Se elige graficar el individuo 'AA', sesión '0', canal 1.

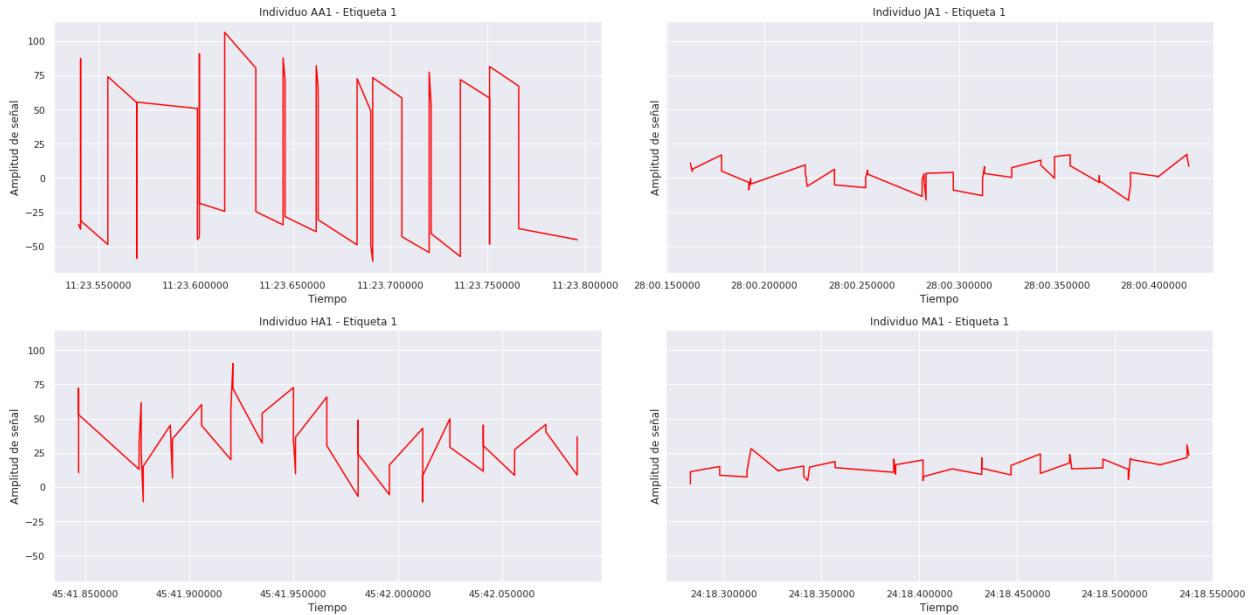
Usando 2000 muestras:



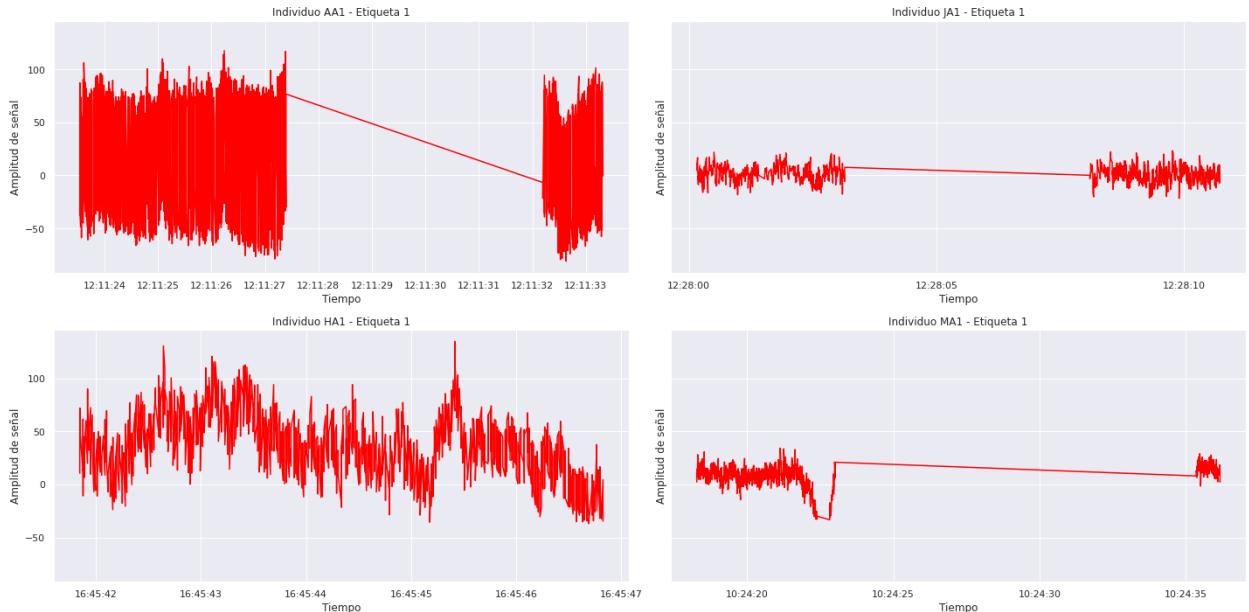
- **Mismo estado - diferentes sujetos**

Se elige graficar la etiqueta 1 para los 4 diferentes sujetos (canal 1).

Usando 50 muestras:



Usando 2000 muestras:



Al analizar estas visualizaciones, ¿extrae alguna información que considere relevante para el problema? ¿Se observa algún fenómeno distinguible a primera vista?

Al analizar estas visualizaciones no se observan grandes patrones para extraer información de las mismas. Las señales en el tiempo muestran la amplitud de la señal y la forma de onda. Con esto podemos determinar la diferencia de amplitud que tienen las distintas personas siendo mayor el individuo AA que el MA por ejemplo. Si la señal es positiva o negativa creemos que no tiene mucha relevancia ya que eso depende de cómo esté dispuesto el electrodo en función de la señal, ya que la dirección de la misma va a depender su signo.

Como conclusión el análisis y exploración en el tiempo, nos brinda información de los outlier, nos permitió detectar que existía un artefacto que impacta en los datos recolectados. También pudimos concluir respecto a la relación que tenían las variables y elegir cómo trabajarlas.

Parte II: Dominio del tiempo.

A) Nivel Segmento/Estado: Seleccione los datos correspondientes a un paciente y un canal, y para él defina un conjunto de señales para cada estado presente en el dataset. Para cada uno de ellos estudie los siguientes elementos y luego compárelos.

a) ¿Presenta los valores de voltaje una distribución normal? Utilizar un criterio gráfico y un test para probarlo. Si la distribución normal no se ajusta, ¿a qué distribución se asemejan?

Se elige el paciente HA (sesión 1), canal 3. Luego se define un conjunto de señales para etiqueta 1 y para etiqueta 2. Se tiene como primicia que cada intervalo de señal con etiqueta 1 y 2 dura aproximadamente 10 segundos. Por ende se utilizan 2000 muestras por intervalo por segmento. Se analizan los siguientes intervalos:

Intervalo 1: HA (sesión 1), canal 3, muestras de la 0 a la 2000.

Intervalo 2: HA (sesión 1), canal 3, muestras de la 2000 a la 4000.

Intervalo 3: HA (sesión 1), canal 3, muestras de la 4000 a la 6000.

Distribuciones Individuo HA



Luego se realizaron dos test para evaluar la normalidad de las distribuciones graficadas.

test de hipótesis para ver normalidad normal_test intervalo 1 etiq 1
NormaltestResult(statistic=24.828621009639097, pvalue=4.060069343488381e-06)

test de hipótesis para ver normalidad normal_test intervalo 1 etiq 2
NormaltestResult(statistic=23.158045516384387, pvalue=9.360397895342006e-06)

test de hipótesis para ver normalidad normal_test intervalo 2 etiq 1
NormaltestResult(statistic=324.5527432070845, pvalue=3.3440077039579404e-71)

test de hipótesis para ver normalidad normal_test intervalo 2 etiq 2
NormaltestResult(statistic=94.0355256368617, pvalue=3.805792067040177e-21)

test de hipótesis para ver normalidad normal_test intervalo 3 etiq 1
NormaltestResult(statistic=90.49376952147769, pvalue=2.2362874425209434e-20)

test de hipótesis para ver normalidad normal_test intervalo 3 etiq 2
NormaltestResult(statistic=360.77927577685807, pvalue=4.547531044480744e-79)

Como conclusión de esto podemos decir que por más de que los test de normalidad nos rechace la hipótesis nula la cual establece normalidad de los histogramas, decimos que de manera gráfica tienden de manera aproximada a la misma, aunque al tener diferencia en sus colas los test suelen rechazarla por su gran sensibilidad.

b) Realice un resumen estadístico de los valores de voltaje en el intervalo de tiempo considerado. ¿Qué estimador de posición central usaría para describir los valores? ¿Y de dispersión?

Cálculo estadísticos de posición central y dispersión para el Intervalo 2: HA (sesión 1), canal 3, muestras de la 2000 a la 4000.

Media de la muestra tomada etiqueta 1: 21.32

Mediana de la muestra tomada etiqueta 1: 18.81

Desvío estándar de la muestra tomada etiqueta 1: 31.28

Media de la muestra tomada etiqueta 2: 36.75

Mediana de la muestra tomada etiqueta 2: 31.38

Desvío estándar de la muestra tomada etiqueta 2: 34.12

Podemos utilizar la media como estimador de posición central ya que al ser eliminados los valores extremos anteriormente este estimador no sufre su desventaja de verse alterados por estos valores.

c) En adición a los datos dañados encontrados en la parte I, ¿Encuentra outliers a este nivel de análisis? ¿Estos outliers deberían ser tratados de forma diferencial? ¿De qué manera?

Los valores outliers fueron eliminados anteriormente, retirando así los diversos artefactos. Y por ende los voltajes se encuentran en un rango normal para las señales EGG (entre $\pm 200 \mu\text{V}$).

d) ¿Existe una diferencia estadísticamente significativa para considerar que los estimadores de posición central son diferentes entre los estados? Use un test de hipótesis para probarlo al menos entre dos estados.

Teníamos como valores de medias del intervalo dos elegido:

Media de la muestra tomada etiqueta 1: 21.32

Media de la muestra tomada etiqueta 2: 36.75

Vemos que existe una diferencia entre la media de los dos grupos.

Diferencia entre las medias: -15.42

Por lo cual realizamos un test de hipótesis para confirmarlo.

Test de diferencia de medias:

```
Ttest_indResult(statistic=-14.137729506484996, pvalue=2.55952239341677e-44)
```

Utilizando el t-test para muestras independientes analizando las 2000 muestras del segundo intervalo, obtuvimos un p-valor cercano a 0, por ende se rechaza la hipótesis nula de que ambas medias son iguales. Es con esto que decimos que la media para cada estado es distinta.

Al realizarlo en el primer intervalo se obtienen resultados similares:

Diferencia entre las medias: -19.36

```
Ttest_indResult(statistic=-33.69010488283825, pvalue=3.403229150304824e-219)
```

e) Resuma las principales conclusiones de este nivel de análisis.

Como conclusión podemos decir primero que por más que algunos intervalos tiendan a parecerse una normal los test nos especifican que no cumple sus requisitos debido a su gran sensibilidad a la hora de realizar los mismos.

Segundo recalcamos la diferencia de medias entre ambas etiquetas. Esto quiere decir que si el encéfalo es estimulado a una frecuencia más baja como la etiqueta 1 su estimulación genera un menor grado de amplitud de voltaje en las señales cerebrales que si lo es estimulado a una frecuencia mayor como la de la etiqueta 2. Sin embargo lo primordial es que a distinta frecuencia el cerebro se estimula de manera diferente.

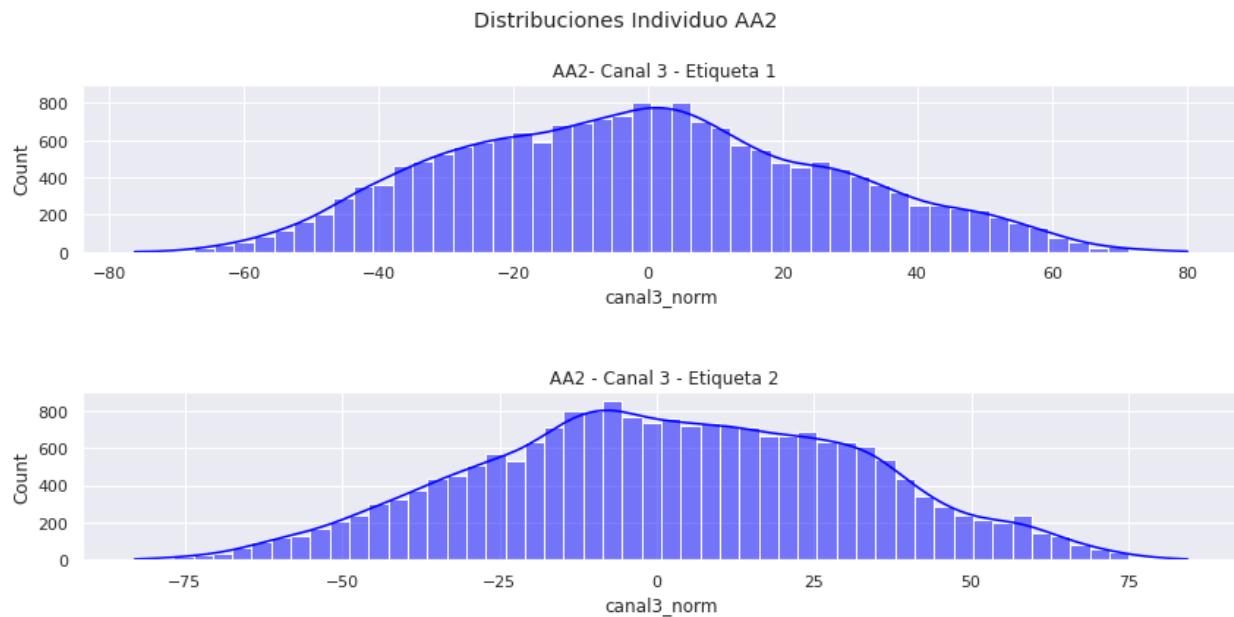
Analizaremos esta premisa posteriormente en el análisis en frecuencia.

B) Nivel Paciente - un canal: Seleccione los datos correspondientes a un paciente y un canal de adquisición y para ese caso estudie los siguientes elementos:

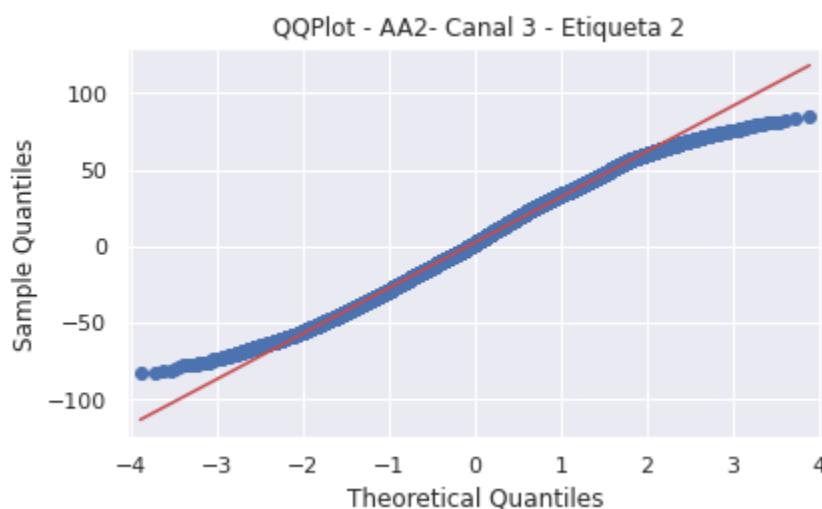
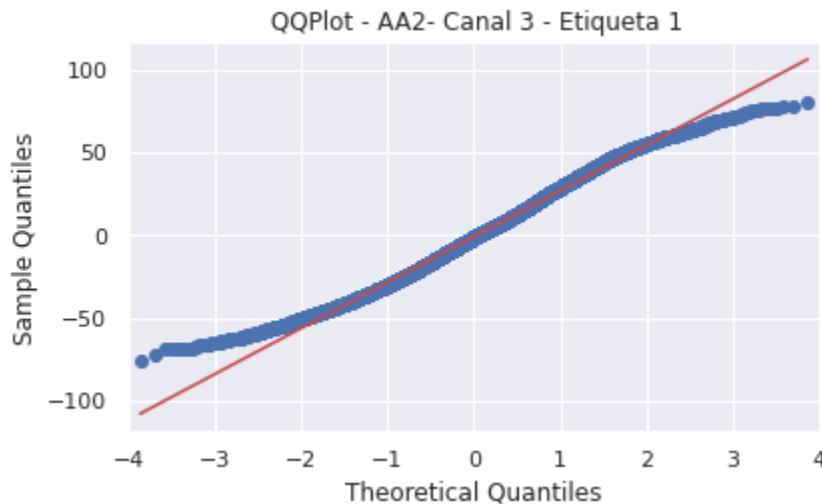
- a) Considere el conjunto completo de valores de voltaje correspondientes a cada uno de los estados a lo largo de todo el registro y repita los elementos del apartado II-A).**

Se considera para este análisis el paciente 'AA2', canal '3'. Cada uno de los estados a lo largo de todo el registro (ambos canales).

Se grafica la distribución de los valores de voltaje



Utilizando como método gráfico la función 'qqplot' podemos ver si nuestra distribución se ajusta a la distribución normal.



Las desviaciones de los puntos azules respecto de la línea roja en ambos gráficos muestran una desviación de la distribución esperada. Por lo tanto nuestras muestras del Paciente "AA2", "Canal3" para ambos estados, no se ajustan a una distribución normal, pero visualmente se aproxima a una distribución normal, por más que presenta colas que difieren de una normal y por lo tanto los tests rechazan la hipótesis de normalidad.

Realizando Test de Hipótesis (Test de DAgostino) obtenemos los siguientes resultados:

Para la etiqueta 1 de la muestra elegida:

Estadístico=522.288, p=0.000

La muestra no parece Gaussiana o Normal(se rechaza la hipótesis nula H0)

Para la etiqueta 2 de la muestra elegida:

Estadístico=415.120, p=0.000

La muestra no parece Gaussiana o Normal(se rechaza la hipótesis nula H0)

De acuerdo a ambos tests de hipótesis, en las muestras elegidas de ambos estados, se obtiene que las muestras no parecen tener una distribución que se ajuste a la normal.

Las distribuciones se parecen a una normal si bien los estadísticos no lo muestran.

Estimadores de posición central (media y mediana)

Etiqueta 1 de la muestra elegida:

Media = -0.94

Mediana = -1.52

Etiqueta 2 de la muestra elegida:

Media = 2.28

Mediana = 1.95

Estimadores de dispersión

Etiqueta 1 de la muestra elegida:

Varianza = 763.21

Desviación Estándar = 27.63

Rango = 156.31

IQR = 39.76

Etiqueta 2 de la muestra elegida:

Varianza = 883.53

Desviación Estándar = 29.72

Rango = 167.37

IQR = 42.93

Medias de asimetría

Para valores cercanos a 0, la variable es simétrica. Si es positiva tiene cola a la derecha y si es negativa tiene cola a la izquierda.

Si la asimetría está entre -0,5 y 0,5, los datos son bastante simétricos

Si la asimetría está entre -1 y -0,5 o entre 0,5 y 1, los datos son moderadamente asimétricos

Si la asimetría es inferior a -1 o superior a 1, los datos están muy sesgados

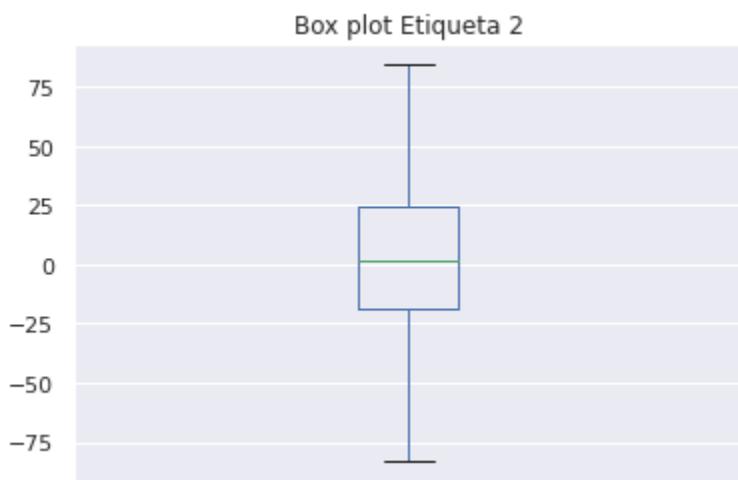
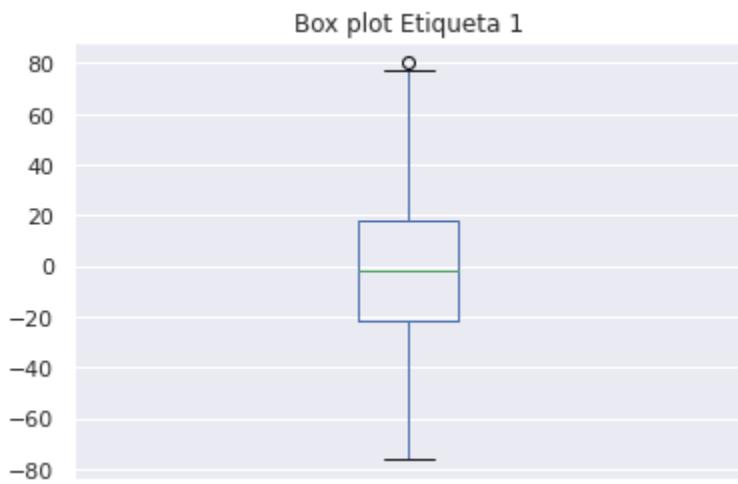
Etiqueta 1 de la muestra elegida:

Asimetría = 0.17

Etiqueta 2 de la muestra elegida:

Asimetría = -0.01

En los datos con los voltajes (ya restada la línea de tendencia), vemos que los valores de voltaje del canal elegido se encuentra comprendido entre -85, +80, aproximadamente. Por lo tanto no presenta outliers.



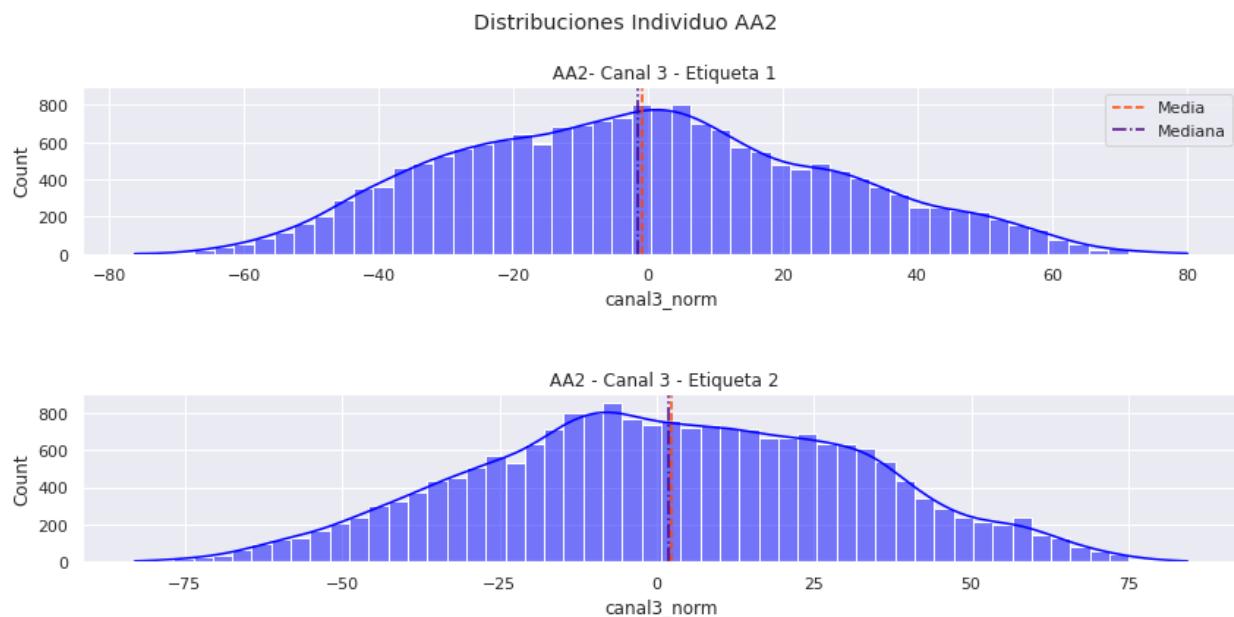
Diferencia de medias

Se grafica la distribución de los valores de voltaje donde se visualiza que si existe una diferencia entre las medidas centrales de media y mediana entre ambas muestras (por estado), aunque ésta no es significativa.

Media etiqueta 1 = -0.94

Media etiqueta 2 = 2.28

Diferencia de medias: -3.22



Realizamos un test de hipótesis (ztest) sobre la diferencia de medias:

hipótesis nula: las medias muestrales son iguales

hipótesis alternativa: las medias muestrales son iguales

Estadístico = -11.080

p-value = 0.000

La media de las muestras no parecen iguales (se rechaza la hipótesis nula H0)

Como el p-valor (=1.56e-28) < alpha (=0.05); hay evidencia muestral suficiente para rechazar la H_0 de que las medias de las muestras son iguales.

b) Ahora que dispone de más datos, ¿son variables independientes el estado registrado de la señal y su voltaje? Use herramientas cuantitativas y cualitativas para justificar su respuesta.

Las variables NO son estadísticamente independientes (o sea, sí hay asociación entre ellas) cuando:

- $Cov(X,Y) \neq 0$
- $Corr(X,Y) \neq 0$
- $Var(X+Y) \neq Var(X)+Var(Y)$
- $E(X.Y) \neq E(X).E(Y)$

La correlación de Pearson asume que las variables aleatorias se distribuyen normalmente, por lo que hay que tenerlo en cuenta a la hora de interpretar los resultados. Alternativamente, puede cambiar la función 'pearson' por la función de correlación de rango de Spearman: 'spearman', que no asume la normalidad de sus variables. En este caso hemos usado la correlación de 'spearman'.

Calculemos estos valores para concluir independencia o no entre las variables, para este punto se utiliza toda la muestra (paciente AA2, Canal3) contemplando la etiqueta 1 y 2.

La covarianza entre ambas variables es: -3.00

Se cumple $Cov(X,Y) \neq 0$

La correlación entre ambas variables es: -0.09

Se cumple que $Corr(X,Y) \neq 0$

Varianza conjunta: 7128.82

Suma de las varianzas: 7134.82

Vemos que $Var(X+Y) \neq Var(X)+Var(Y)$

La esperanza conjunta es de: 3119.68

La multiplicación de las esperanzas es: 3122.68

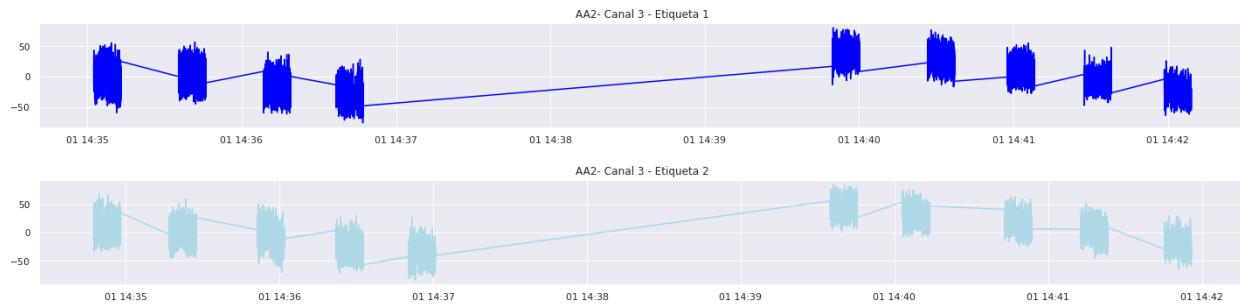
Vemos que $E(X.Y) \neq E(X).E(Y)$

Por lo que concluimos que las variables canal 3 y etiqueta en el conjunto de datos de la persona AA2, no son estadísticamente independientes (o sí tienen asociación).

Pero como vimos la correlación entre ambas variables es cercana a cero, por lo que su asociación es baja (Una correlación cercana a 0 indica que no existe relación lineal entre las variables).

A pesar de estos resultados, a fines prácticos consideraremos estas variables independientes ya que los resultados teóricos calculados arrojan resultados muy similares muy similares a los de la definición de independencia entre variables. Creemos que dicha teoría es muy difícil de cumplir estrictamente en una muestra pequeña, aunque en muchos casos se aproxima como es en este ejemplo.

c) Para cada uno de los estados, los valores de voltaje a lo largo del tiempo, ¿varían con alguna tendencia?



Con los valores de voltaje con su línea de tendencia restada, se observan patrones repetidos entre intervalos de cada etiqueta cuando está encendida la luz de la etiqueta correspondiente. Los intervalos marcados con una línea representan los valores que hemos eliminados por corresponder a etiqueta '99' (intervalos de no estimulación). No se visualiza una tendencia particular en ambos segmentos.

d) Resuma las principales conclusiones de este nivel de análisis.

En base a los análisis realizados, no se logra visualizar una diferencia significativa entre las medias de etiqueta 1 y etiqueta 2 a pesar de que el test rechace esta opinión. Y concluimos que la variable 'canalX' es independiente de la variable 'etiqueta'.

C) Nivel Paciente - multi canal: Seleccione los datos correspondientes a un paciente y para ese caso estudie los siguientes elementos:

a) Las señales de voltaje en función de tiempo para cada canal, ¿son variables independientes entre sí? Use herramientas cuantitativas y cualitativas para justificar su respuesta. (Ejemplo, matriz de correlación)

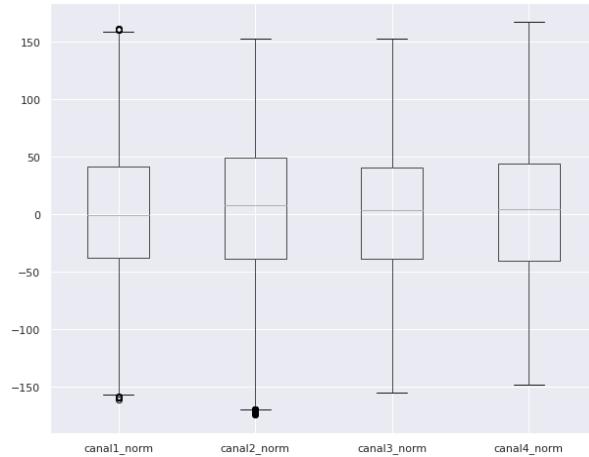
Para responder a la pregunta se realizó una matriz de correlación graficada en un mapa de calor. Se utilizó el individuo HA.



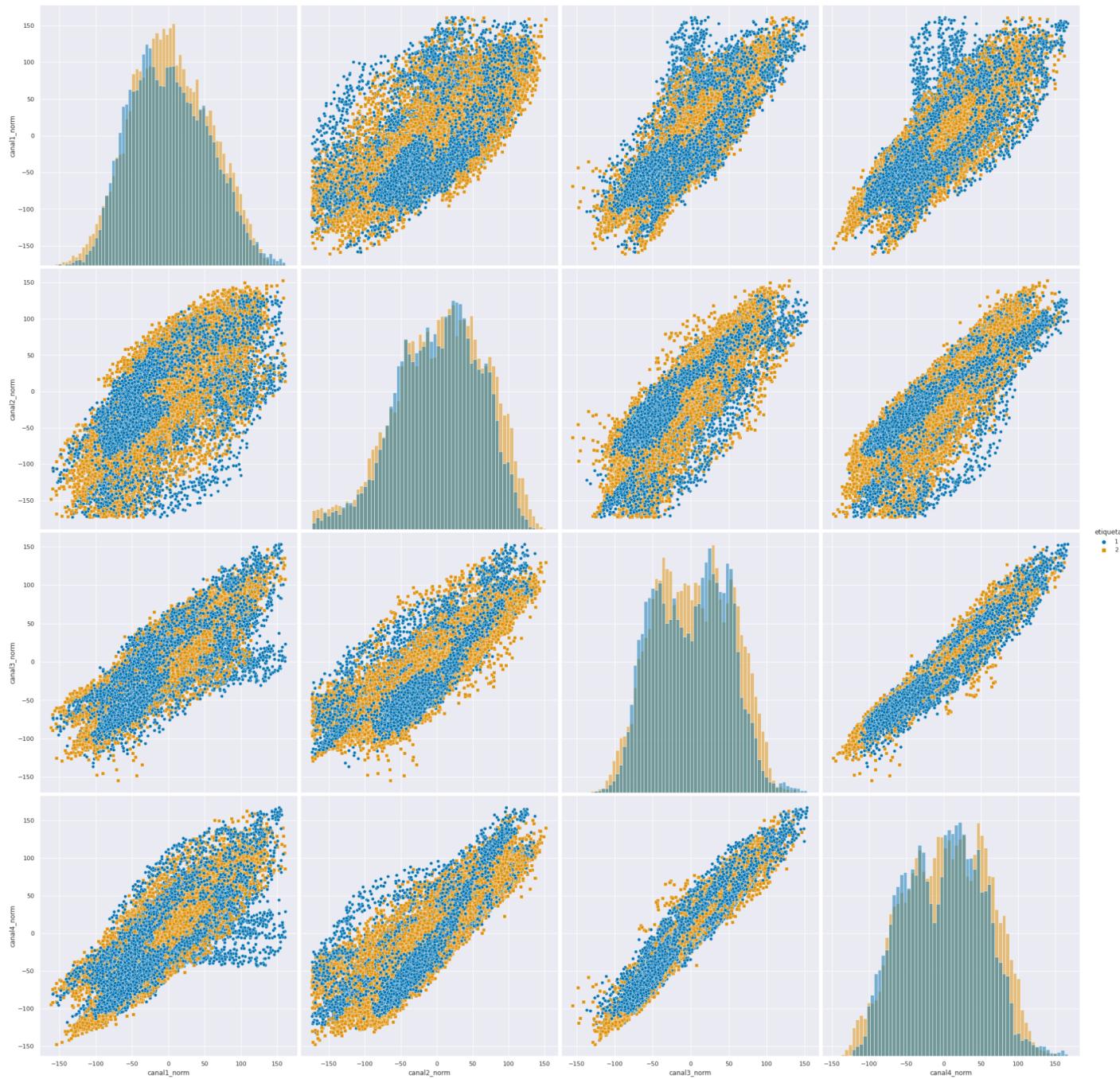
De acuerdo a la información que nos proporciona la matriz de correlación de dichas variables numéricas y el gráfico heatmap, podemos concluir que si existe una relación entre las variables canales del 1 al 4. Las variables no son independientes entre sí. Mientras una aumenta la otra también tiende a aumentar y viceversa, por ende se podría elegir un subconjunto de canales.

b) Tomando los puntos que considere relevantes del apartado II-B) para cada canal y considerando la respuesta anterior. ¿Considera relevante trabajar con todos los canales disponibles o podría quedarse con un subconjunto? Si elige el subconjunto, ¿qué canales elegiría y por qué?

Veamos mediante un diagrama ‘boxplot’, cómo se distribuyen los valores de voltaje.



Graficamos también la densidad conjunta de los canales dividiendo por estados:



Cuando se grafica la distribución de canales de forma conjunta se observa que mantienen una relación lineal en los scatter plot y con distribución estadística aproximadamente similares. Teniendo en cuenta la que existe correlación entre los canales, podríamos trabajar con un subconjunto.

Se decide utilizar sólo un canal, viendo los boxplots, vemos que los canales 3 y 4 no tienen valores por fuera de los rangos intercuartílicos, por lo tanto podríamos quedarnos con cualquiera de esos dos.

Como patrón más general, se analizaron los gráficos de los 4 canales, y el que tenga menos ruido sería el que se utilice como único canal.

**c) Opcional: Tomando un par de canales a elección, analice la distribución conjunta de los valores de voltaje para cada estado de forma cualitativa.
(Ejemplo, Heatmap, scatterplot, 3D-mesh, etc.)**

Se realizaron diversos scatterplots entre las variables discriminando por etiqueta (que se visualizan en el gráfico anterior). Se ven relaciones lineales entre las variables.

d) Resuma las principales conclusiones de este nivel de análisis.

Creemos que se podría trabajar con un solo canal ya que el resto de los canales no aporta información distinta a la que ya aporta uno solo. Esto resulta conveniente para reducir dificultad en el problema como así también su dimensionalidad de variables.

D) Nivel Multi-Paciente.

a) A partir de las conclusiones extraídas de los niveles de análisis anteriores.

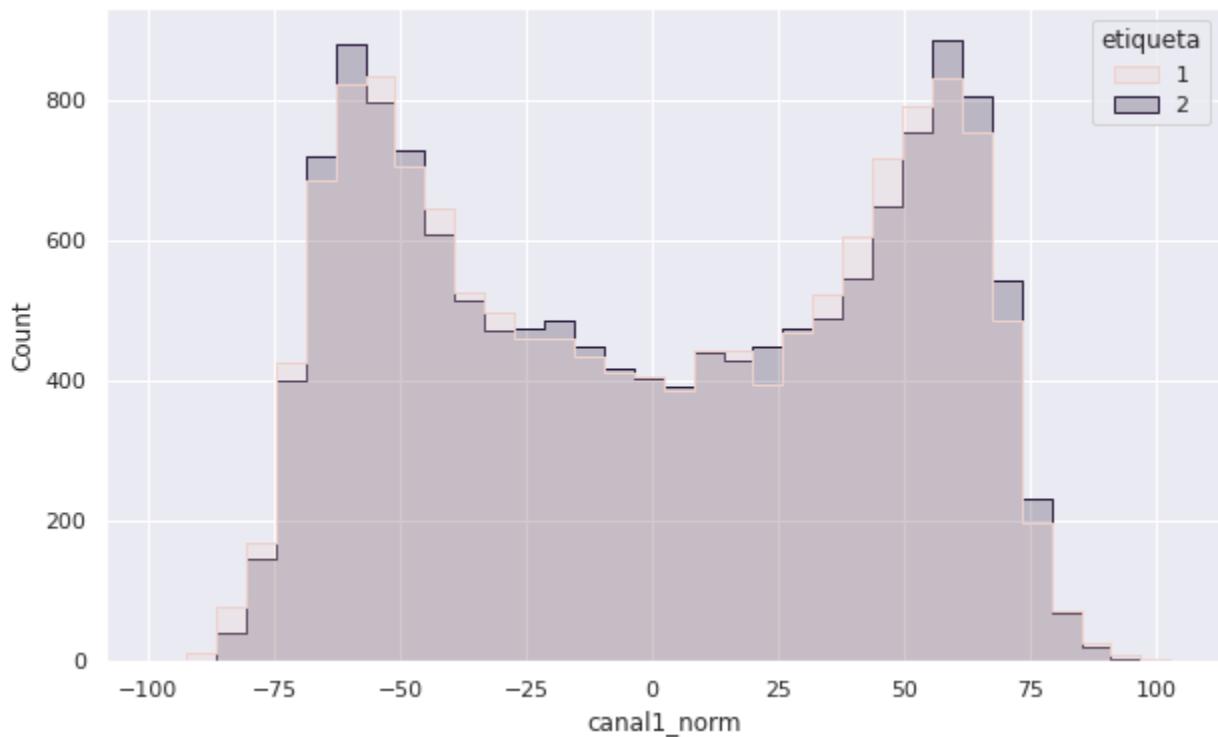
Decida cuáles son los aspectos más importantes a analizar de los registros de un paciente y compárelos entre pacientes. ¿Encuentra diferencias significativas? ¿Qué variables pueden identificar esas diferencias?

A modo de ejemplo: los valores de voltaje medios para cada estado de un paciente, ¿difieren significativamente entre pacientes?

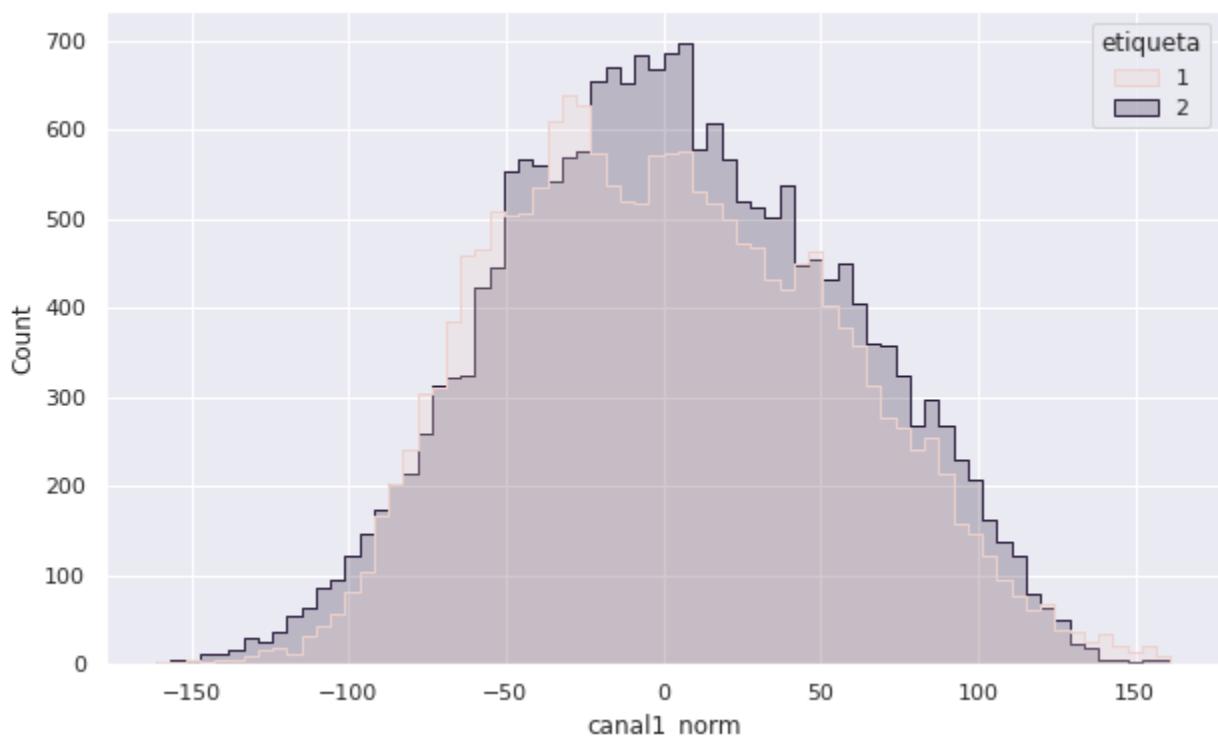
Se seleccionaron dos individuos diferentes HA1 y AA0, canal 1, y ambas etiquetas para el análisis.

A nivel multi paciente, graficamos histogramas de dos personas distintas discriminando por etiqueta:

Persona AA0



Persona HA1



De acuerdo a lo observado en gráficos de caja, histograma y medidas de centralidad, en el dominio del tiempo no encontramos información relevante que nos permita concluir que existe una diferencia entre los registros obtenidos en los diferentes canales e individuos cuando se estimula a un individuo con luces de diferentes frecuencia.

Para el mismo individuo, mismo canal y distinta etiqueta, no existen diferencias significativas en la amplitud de la señal que nos pueda brindar información relevante. Se cree que las diferencias se van a poder visualizar mejor en el dominio de la frecuencia.

Se observa un cambio de amplitud en la señal para individuos diferentes, pero que no están relacionados al experimento de estimular con luces diferente sino a la diferente conductividad que puede llegar a tener su cuero cabelludo o su diferente amplitud de señal por ser diferentes personas.

Además vale aclarar, teniendo en cuenta el conocimiento de dominio, que los niveles de amplitud de señal en el análisis en el dominio del tiempo dependen también de factores externos (como por ejemplo la referencia para las diferencias de potencial varía por cada individuo), por tal motivo no nos brinda demasiada información relevante.

Como conclusión, el análisis y exploración en el tiempo, nos brindó información de outliers y nos permitió detectar que existía un artefacto que impactaba en la amplitud de voltajes de los datos recolectados. También pudimos concluir respecto a la relación que tienen las variables y elegir cómo trabajarlas.