MENTORIA INTERFAZ CEREBRO COMPUTADORA

TP1: ANÁLISIS Y VISUALIZACIÓN DE DATOS

PRIMERA PARTE

INTEGRANTES: Francisco Rua, Clarisa Manzone

Para entender el origen de los datos que analizaremos hay que tener que en cuanta que la BCI (Brain-computer interface) es un tipo tecnología que permite la obtención y procesamiento de señales neuronales. Una forma de medir estas señales es mediante la encefalografía superficial, a través de la colocación de electrodos en la región craneal de interés de los pacientes o sujetos a estudiar. Como nos interesa en estudio de los SSVEP o potenciales evocados visualmente en estado estacionario, los electrodos son ubicados en la región de la corteza visual.

La intención de este estudio es tratar de encontrar algún patrón o semejanza entre la frecuencia de las luces a las cuales los pacientes son expuestos y las ondas cerebrales que se generan cuando estos miran el estímulo luminoso.

PARTE I: EXPLORACIÓN DE LA BASE DE DATOS

Describir las características generales del dataset:

- **Número de registros, diferencias entre los mismos**

Como mencionamos previamente, los datos que analizaremos provienen de mediciones de potenciales eléctricos (voltajes) cerebrales. Contamos con un total de 7 registros, es decir, 7 mediciones a lo largo de un periodo de tiempo de los SSVEP. Estos 7 registros provienen de 4 pacientes diferentes sólo que el número de sesiones que se le realizaron a cada uno de ellos varia de la siguiente forma:

- Paciente AA presenta 3 sesiones
- Paciente JA presenta 2 sesiones
- Paciente HA presenta 1 sesión
- Paciente MA presenta 1 sesión

Las diferencias que podríamos llegar a encontrar en cada sesión dependen de la variación tanto interpaciente como intrapaciente. Recordemos que estamos midiendo potenciales superficiales y por lo tanto el área que abarca la toma de la muestra es mayor a lo deseado provocando que la precisión en la obtención de los datos disminuya. Es por esto que cualquier otro estímulo externo particular que ocurra en cada sesión, además de la frecuencia de parpadeo de las luces, podría provocar cambios en las ondas cerebrales de los pacientes.

- **Definir la conveniencia de trabajar todos juntos como un solo dataset, o por separado**

Si bien se podría trabajar con cada conjunto de datos (cada sesión) independientemente nosotros unificamos las bases de datos en una sola. Para ello usamos el método append de la librería Pandas para juntar los dataframes, previamente añadiendo la columna denominada “name” para identificar los datos de cada sesión y poder agruparlos para realizar los análisis posteriores.

Análisis de las columnas presentes en el dataset:

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

Con respecto al dataset en general, luego de eliminar los metadatos correspondientes a las primeras 10 líneas de los archivos, eliminamos algunas columnas para quedarnos con el conjunto de información que nos interesará posteriormente.

El conjunto de datos originalmente se veía de esta manera:

```
%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
0, 1698.27, 721.51, 1778.40, 1771.82, 0.000, 0.000, 0.000, 99, 14:31:40.678, 1606239100678
1, 1759.80, 775.99, 1817.29, 1823.40, 0.000, 0.000, 0.000, 99, 14:31:40.693, 1606239100693
2, 1625.73, 657.79, 1736.31, 1717.40, 0.000, 0.000, 0.000, 99, 14:31:40.693, 1606239100693
3, 1540.47, 537.86, 1692.34, 1637.11, 0.000, 0.000, 0.000, 99, 14:31:40.707, 1606239100707
4, 1714.80, 727.68, 1792.99, 1786.76, 0.000, 0.000, 0.000, 99, 14:31:40.707, 1606239100707
5, 1767.88, 753.35, 1821.33, 1835.18, 0.000, 0.000, 0.000, 99, 14:31:40.708, 1606239100708
6, 1632.67, 660.23, 1743.84, 1719.60, 0.000, 0.000, 0.000, 99, 14:31:40.708, 1606239100708
7, 1536.94, 582.71, 1680.36, 1625.90, 0.000, 0.000, 0.000, 99, 14:31:40.723, 1606239100723
8, 1707.39, 721.68, 1782.64, 1793.22, 0.000, 0.000, 0.000, 99, 14:31:40.723, 1606239100723
9, 1756.60, 777.47, 1806.60, 1807.12, 0.000, 0.000, 0.000, 99, 14:31:40.738, 1606239100738
10, 1619.96, 600.47, 1724.60, 1683.40, 0.000, 0.000, 0.000, 99, 14:31:40.738, 1606239100738
11, 1520.94, 562.91, 1665.74, 1608.43, 0.000, 0.000, 0.000, 99, 14:31:40.738, 1606239100738
12, 1684.90, 706.23, 1759.63, 1758.27, 0.000, 0.000, 0.000, 99, 14:31:40.738, 1606239100738
13, 1743.45, 740.70, 1790.01, 1803.61, 0.000, 0.000, 0.000, 99, 14:31:40.753, 1606239100753
14, 1611.32, 653.33, 1718.11, 1680.95, 0.000, 0.000, 0.000, 99, 14:31:40.753, 1606239100753
15, 1515.23, 533.27, 1661.83, 1612.00, 0.000, 0.000, 0.000, 99, 14:31:40.767, 1606239100767
16, 1681.30, 691.95, 1759.04, 1761.34, 0.000, 0.000, 0.000, 99, 14:31:40.767, 1606239100767
17, 1745.47, 799.80, 1795.41, 1812.23, 0.000, 0.000, 0.000, 99, 14:31:40.768, 1606239100768
18, 1596.28, 659.27, 1706.58, 1675.83, 0.000, 0.000, 0.000, 99, 14:31:40.768, 1606239100768
```

1 2 3 4 5

Información de cada columna

1: Número de muestra/índice

2: Mediciones temporales (voltajes)

3: Acelerómetros

4: Etiquetas SSVEP

5: Etiquetas temporales

Por lo tanto, eliminamos algunas columnas y nos quedamos con las que contienen los siguientes datos:

- Canal 0
- Canal 1
- Canal 2

Mediciones temporales de los voltajes

- Canal 3
- Etiquetas SSEVP
- Etiquetas temporales (sólo utilizamos la primera de ellas ya que la segunda brinda la misma información, pero en otro formato utilizado por el Software OpenBCI)

Se puede notar que no trabajamos con las tres columnas de los acelerómetros ya que su valor es cero en todos los casos porque no fueron utilizados para la recolección de los resultados.

¿Qué tipo de datos contienen? ¿Qué variables describen las columnas consideradas? ¿Con qué sensibilidad? En el caso de los datos cualitativos. ¿Cuáles son los valores posibles para esta variable? Describa su presencia (frecuencia, intervalos, secuencia, etc.)

Con las modificaciones el data frame nos quedó de esta manera

	ch0	ch1	ch2	ch3	label	time	Name
0	-1.86	-12.95	-9.18	-41.36	99	0.000	AA0
1	10.77	3.47	6.87	-15.13	99	0.005	AA0
2	87.61	65.61	88.78	32.23	99	0.010	AA0
3	83.04	50.88	77.30	4.49	99	0.015	AA0
4	8.07	-20.55	-6.68	-38.09	99	0.020	AA0
...	...	...	...	...	...	...	...
659107	805.10	36.59	1370.61	1342.16	99	348.525	MA1
659108	809.05	27.30	1378.69	1340.12	99	348.530	MA1
659109	818.05	27.28	1384.75	1345.36	99	348.535	MA1
659110	810.54	31.55	1373.65	1334.40	99	348.540	MA1
659111	799.01	29.78	1368.15	1336.88	99	348.545	MA1

659112 rows × 7 columns

Con respecto a los tipos de datos que contiene cada columna podemos usar el comando df.dtypes que nos ofrece Pandas para verlos todos juntos

```
1 df.dtypes
```

```
ch0       float64
ch1       float64
ch2       float64
ch3       float64
label       int64
time      float64
Name       object
dtype: object
```

Podemos decir que la mayoría de las columnas contienen datos enteros o de punto flotante excepto la columna “Name” que contiene strings.

Por otro lado, en cuanto a los tipos de variables de cada columna, aquellas que señalan los **canales** (ch0, ch1, ch2 y ch3) corresponden a variables de tipo numéricas continuas. Estas columnas almacenan la información de la medición de voltajes de cada uno de los electrodos de los que se tomaron los datos. La unidad de la medición es el microvoltio y la sensibilidad con la que captan la diferencia de potencial es de 0.01uV.

Luego tenemos la columna ***“label”*** cuya variable es de tipo categórica. Se le asigna el valor 1 cuando el individuo está frente al estímulo de la luz parpadeante a 12,5Hz y 2 frente a la luz con frecuencia de 16,5 Hz. Ambos estímulos duran un periodo aproximado de 10 segundos. Cuando los sujetos en estudio no están frente a los estímulos controlados se le asigna a la columna el valor arbitrario de 99.

La columna que indica el ***tiempo relativo*** en el que se recolectan los datos de diferencias de potenciales muestra con una sensibilidad de 0.005 segundos el cambio de voltaje que se produce en cada uno de los registros de los canales. El tiempo puede ser considerada como una variable numérica discreta en este caso, si tomamos los intervalos de 0.005s.

Por último, sólo mencionar la columna “***Name***” la cual agregamos con el fin de identificar cada sujeto y cada sesión. Los datos que almacena son de tipo categórico.

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

No encontramos en ninguna de las columnas valores nulos o datos dañados.

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?

Si suponemos que las muestras fueron tomadas con una frecuencia de 200 Hz (200 mediciones de voltaje por segundo) podemos obtener la diferencia de tiempo entre la toma de una

muestra y la siguiente dividiendo 1/200 Hz = 0.005s. Con este Δ de tiempo podemos calcular el numero total de muestras sabiendo la duración de cada sesión de la siguiente manera:

Numero de muestras = T/dt

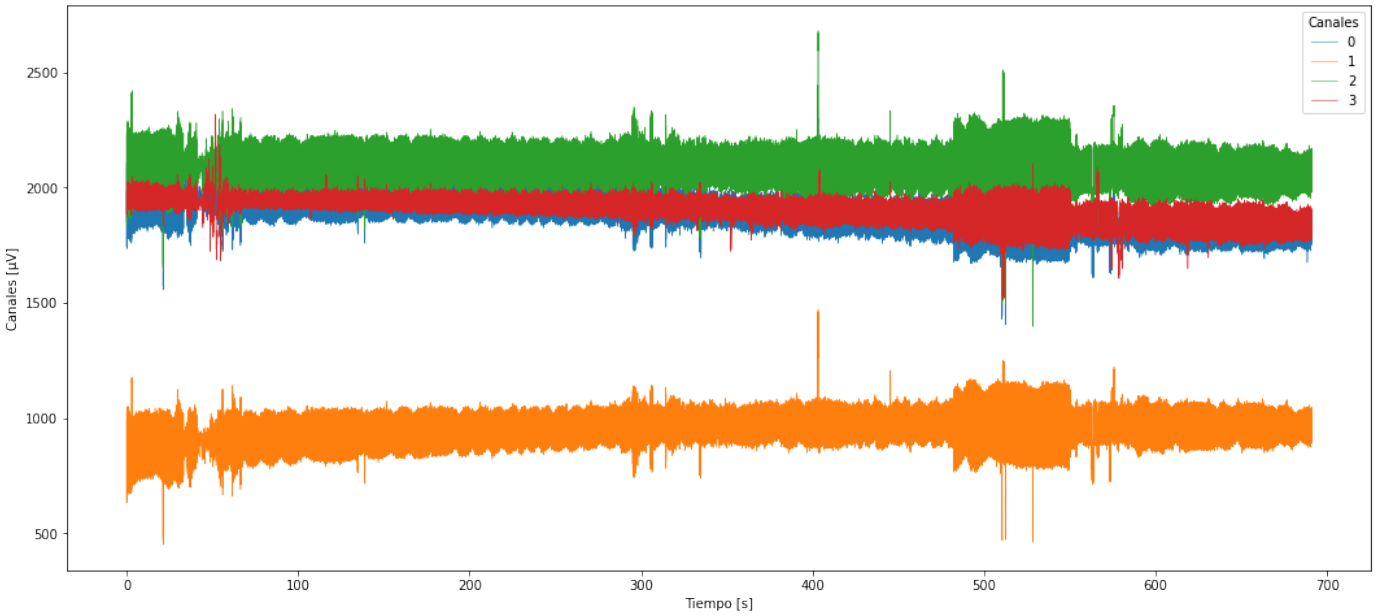
Siendo T el tiempo total de una sesión y dt el tiempo entre cada toma de muestra.

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

Como se puede notar hicimos una transformación en la columna del tiempo, ya que no conservamos los datos originales de la hora de la medición, sino que colocamos el tiempo relativo entre la toma de cada muestra o voltaje lo cual nos es más útil a la hora de analizar y graficar los resultados. De esta manera la primera muestra se registra como 0.000 segundos y se corresponde con el inicio de la sesión. Luego dividimos los datos en intervalos de 0.005 segundos ya que, como se especifica que la frecuencia de muestreo fue de 200 Hz, significa que cada muestra individual fue adquirida cada 0.005 segundos en promedio.

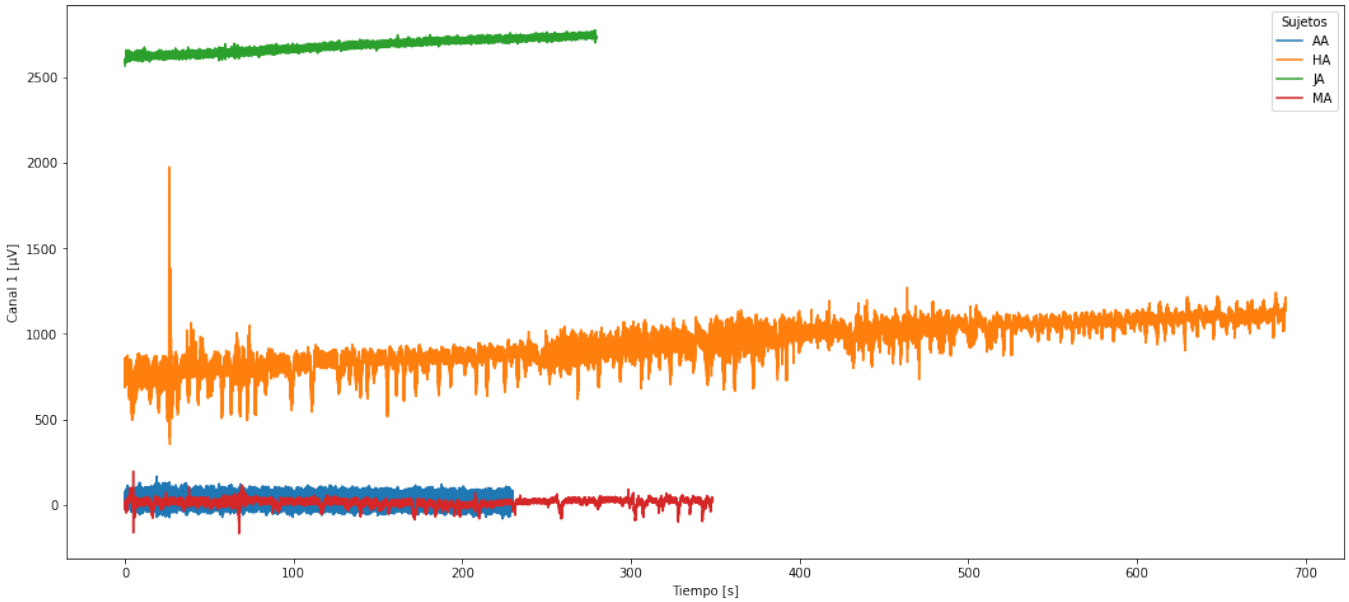
Generar visualizaciones de ejemplo para las series temporales provistas. Determinar los intervalos de tiempo más adecuados para generar visualizaciones claras que permitan comparar las señales en los siguientes escenarios:

- **Un sujeto (AA1) - todos los canales**



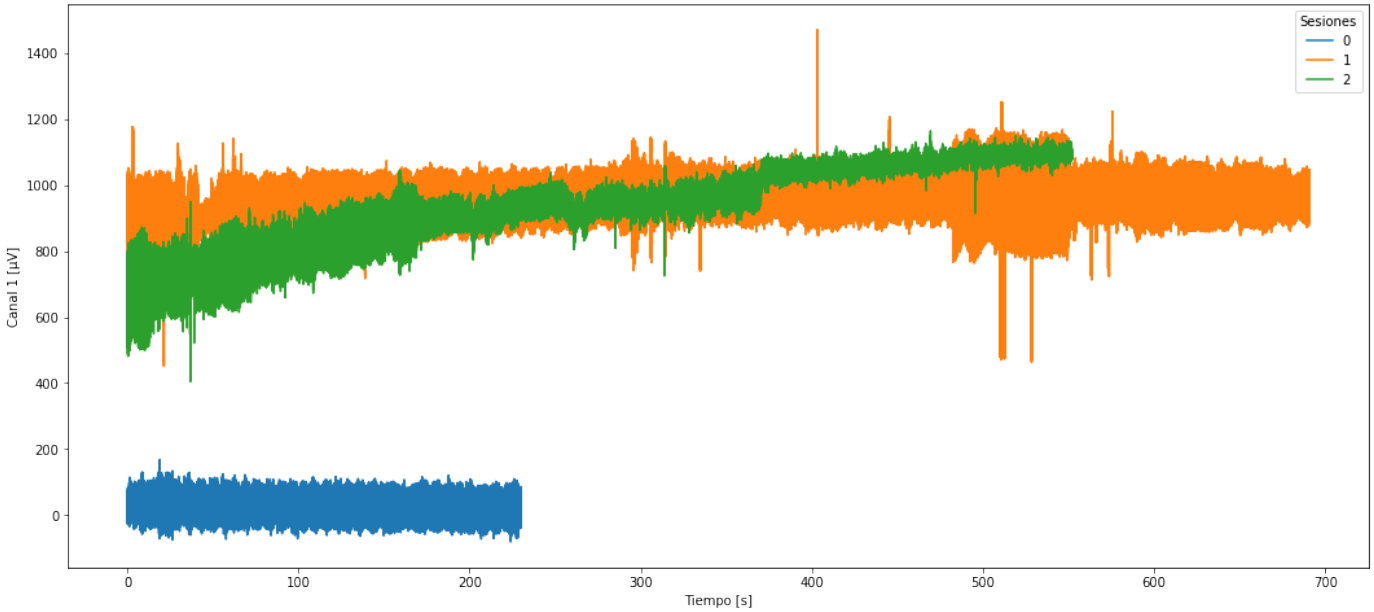
En este primer gráfico podemos observar que en el sujeto AA sesión 1 los canales 0, 2 y 3 presentan una oscilación de voltaje alrededor de 2000/ 2200 uV mientras que el canal 1 se encuentra entre mediciones de 700 uV aproximadamente. Podemos ver una leve tendencia en los 3 canales que se ubican en la parte superior del gráfico a disminuir sus valores de diferencia de potencial en función del tiempo, sin embargo, el canal 1 parece mantener su voltaje.

- **Un mismo canal (Ch1) - todos los sujetos**



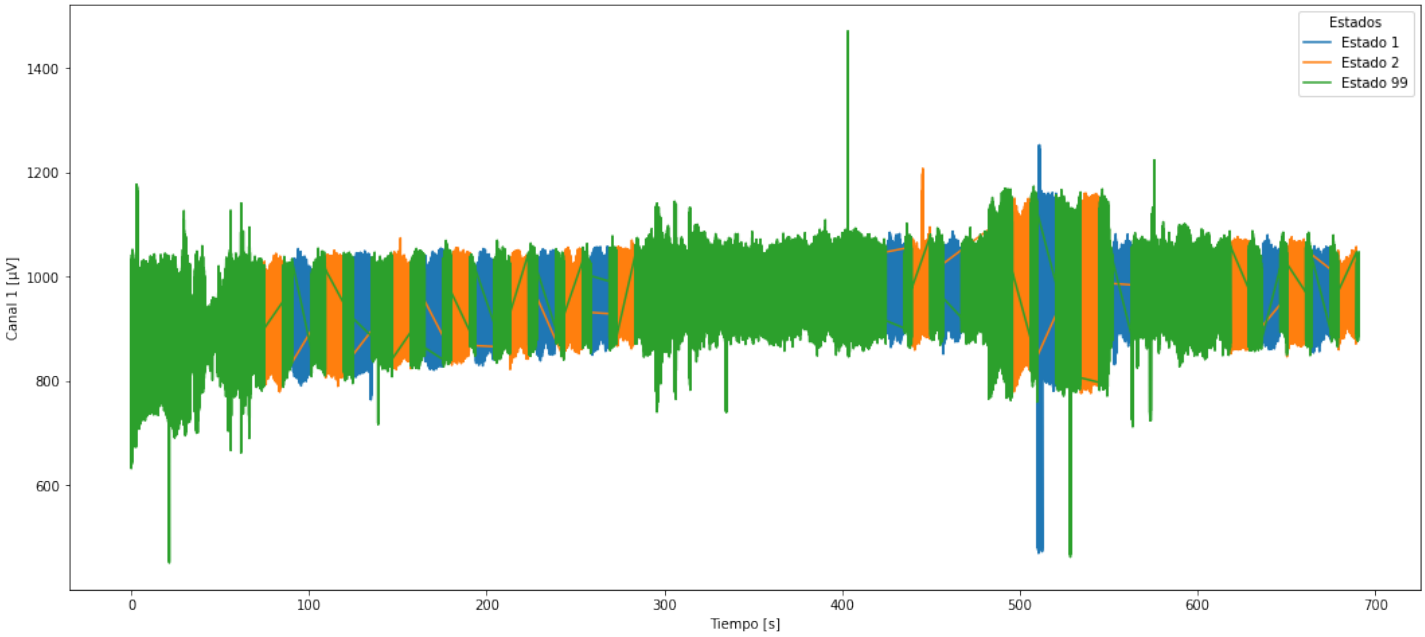
Cuando visualizamos en canal ch1 para los diferentes sujetos e algunas de sus sesiones notamos que los que se corresponden con los pacientes AA y HA además de tener voltajes medios mayores, éstos aumentan con el tiempo. Los sujetos AA y MA presentan valores de diferencia de potenciales alrededor de 0 Um y los mismos parecen mantenerse con el paso de los segundos.

- **Un mismo canal (Ch1) - mismo sujeto en diferentes sesiones (AA0, AA1, AA2)**



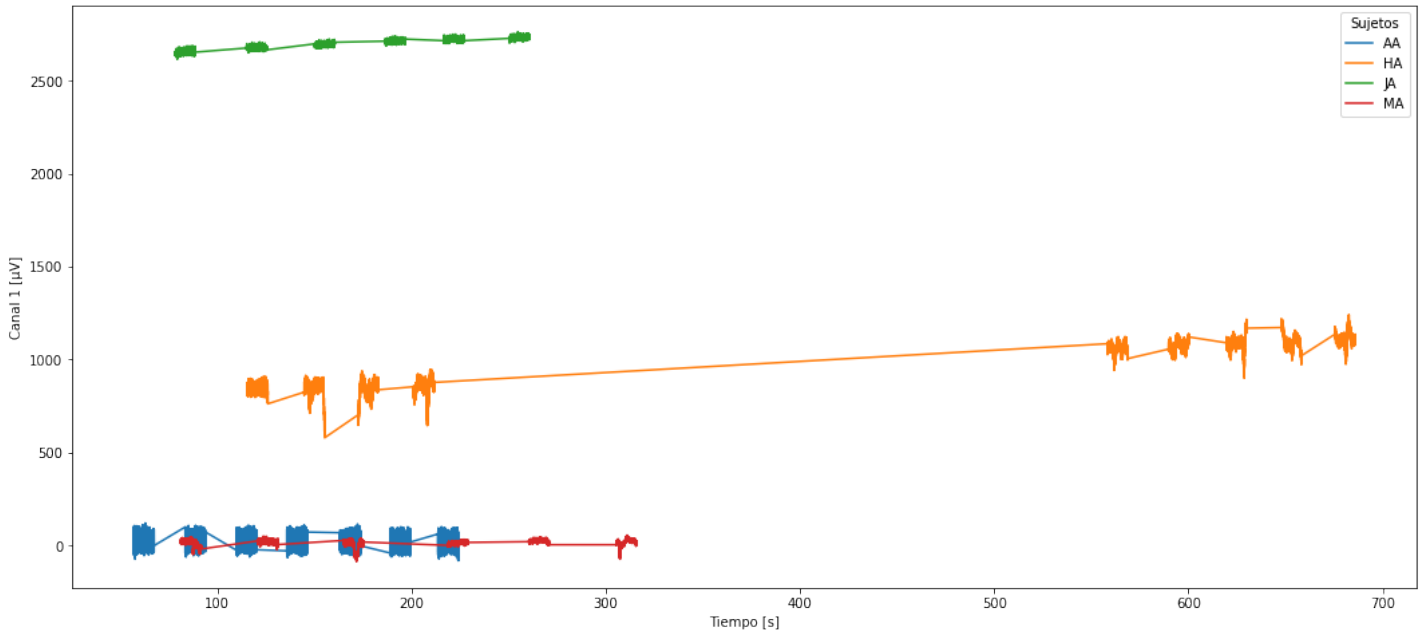
Aquí analizamos el sujeto AA para el canal ch1 en sus 3 sesiones. Vemos como para el mismo paciente los voltajes medios se ubican alrededor de diferentes valores en las distintas sesiones. En este gráfico podemos notar como en la sesión 2 los potenciales aumenten visiblemente en función del tiempo, efecto no tan notorio para las sesiones 0 y 1.

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



Aquí graficamos el sujeto AA1 para el canal 1 diferenciando los diferentes estados (1 y 2 cuando observa las luces a una determinada frecuencia y 99 cuando no está frente a estos estímulos). Este gráfico con estos intervalos de tiempo prolongados no nos permite apreciar si existen o no diferencias en las ondas cerebrales del paciente a medida que el estímulo cambia.

- **Mismo estado (estado 1) - diferentes sujetos**



Al filtrar los valores de diferencia de potencial para un determinado estado, en este caso el 1 que se corresponde con la luz de frecuencia de 12,5 Hz vemos que para los individuos JA y HA los voltajes medios aumentan en función del tiempo, siendo particularmente notorio para el caso HA. Los otros dos pacientes, AA y MA presentan voltajes medios alrededor de 0 uV los cuales parecen mantenerse a lo largo del tiempo. Es una interpretación similar a la que

observamos para el gráfico de un canal – todos los sujetos ya que utilizamos el mismo canal para graficar.

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

En general podemos observar que en algunos casos el voltaje medio aumenta en función del tiempo lo que podría deberse a algún error durante el proceso de medición. Para el resto de las situaciones no encontramos alguna correlación que nos permita sacar conclusiones a simple vista, más allá de lo que describimos para cada gráfico en particular.