MENTORIA INTERFAZ CEREBRO COMPUTADORA

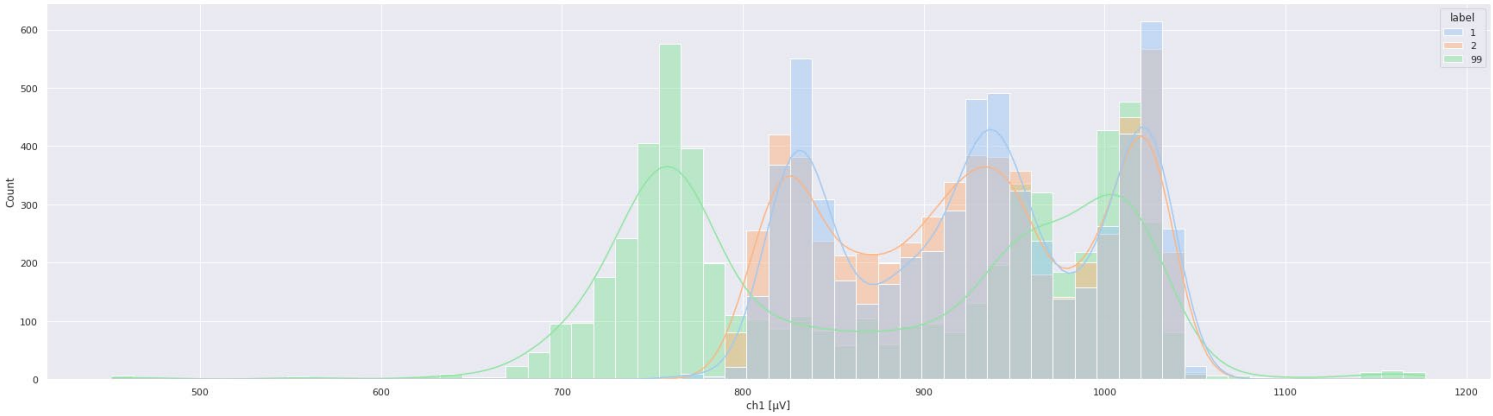
TP1: ANÁLISIS Y VISUALIZACIÓN DE DATOS

Dominio del tiempo.

INTEGRANTE: Clarisa Manzone

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?



Una forma de visualizar los datos para ver si se ajustan a una distribución normal es mediante un histograma donde se grafican los conteos de los eventos en función de los posibles valores que puede tomar la variable. Para las distribuciones que se asemejan a una normal el histograma debería ser simétrico, centrado en la media. Además, la mediana y la moda también coinciden con la media de los registros.

Podemos notar que los histogramas generados a partir de los registros de los diferentes estados para el paciente AA1, canal 1 no se asemejan a lo descripto para una distribución normal. El intervalo de tiempo que tomamos corresponde a los primeros 30 segundos, es decir, teniendo en cuanta la frecuencia de muestreo de 200 Hz serían los primeros 6000 registros de cada estímulo. Observamos en general más de un pico en el gráfico, por lo que existen varios valores con elevada frecuencia y una pérdida de la simetría. Esto se corresponde más bien a una distribución polimodal.

Si bien puede ser un poco difícil discernir en el gráfico el histograma correspondiente a cada estado, elegimos superponerlos para evaluar la coincidencia entre los diferentes estímulos. Podemos notar que los estados 1 y 2 presentan una distribución similar.

No obstante, para estar más seguros que los datos de estas muestras no provienen de una distribución normal, se debería realizar algún test estadístico que acompañe las conclusiones que obtuvimos mediante las visualizaciones de los histogramas. Para ello analizamos los datos con el test no paramétrico de Kolmogorov-Smirnov que toma la media y desviación estándar

de la muestra de datos simulando que provienen de una distribución normal. Compara la función de densidad de la supuesta distribución normal con la función de densidad obtenida y de allí se obtiene un p-valor para el punto de mayor discrepancia entre los resultados de ambas funciones.

Las hipótesis que el test plantea, por lo tanto, son las siguientes:

H_0 = la muestra proviene de una distribución normal

H_1 = la muestra no proviene de una distribución normal

Establecimos un p-valor de 0.005 para rechazar H_0.

Los estadísticos obtenidos fueron:

Estado1 = ks = 0.0938739209756628, p-valor = 1.817423487125089e-46

Estado 2 = ks = 0.08334994661732054, p-valor = 1.038602226808831e-36

Estado 99 = ks = 0.14950981731985363, p-valor = 1.5196592902492923e-117

Para los 3 estados es p-valor es menor a 0.005 por lo que no encontramos evidencia suficiente para aceptar H_0 según este test.

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?

Con el método .describe() podemos obtener los principales valores de medidas de centralidad (media, mediana) y dispersión (desviación estándar, valores máximos y mínimos, rango intercuartil) para el conjunto de datos indicado.

ESTADO 1

```
1 df_AA1_ch1_1.describe()
```

	ch1	time	label
count	6000.000000	6000.000000	6000.0
mean	928.546978	130.022540	1.0
std	72.780084	29.759003	0.0
min	763.970000	90.710000	1.0
25%	855.612500	98.208750	1.0
50%	933.450000	129.387500	1.0
75%	999.372500	166.646250	1.0
max	1053.750000	174.145000	1.0

ESTADO 2

```
1 df_AA1_ch1_2.describe()
```

	ch1	time	label
count	6000.000000	6000.000000	6000.0
mean	924.508878	113.468873	2.0
std	73.487398	28.755372	0.0
min	779.610000	74.945000	2.0
25%	858.020000	82.443750	2.0
50%	927.570000	113.642500	2.0
75%	995.330000	148.386250	2.0
max	1074.310000	155.885000	2.0

ESTADO 99

```
1 df_AA1_ch1_99.describe()
```

	ch1	time	label
count	6000.000000	6000.000000	6000.0
mean	873.997347	14.997500	99.0
std	117.612500	8.660976	0.0
min	451.280000	0.000000	99.0
25%	761.547500	7.498750	99.0
50%	883.350000	14.997500	99.0
75%	986.917500	22.496250	99.0
max	1177.090000	29.995000	99.0

Como estimador de medida de dispersión de los datos podemos usar el valor dado por la desviación estándar que toma la distancia promedio de cada punto muestral respecto a la media.

Por otra parte, podemos utilizar la media como estimador de centralidad, sin embargo, como vimos en el histograma, existen valores extremos y más de un pico para la frecuencia de ciertos voltajes. Estas faltas de simetría en la distribución de los resultados hacen que el valor de la media tienda a desviarse hacia los valores extremos, por eso consideramos que la mediana, el valor del 50% de los datos ordenados de menor a mayor es un mejor estimador de la posición central.

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?

En la primera parte del análisis y exploración de los datos no encontramos datos dañados o faltantes aplicando al conjunto total de datos de todos los pacientes y todas las sesiones el método isna().

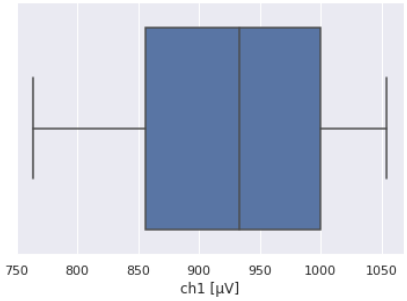
```
1 df_total.isna().sum()
```

```
ch0      0
ch1      0
ch2      0
ch3      0
label    0
time     0
Name     0
dtype: int64
```

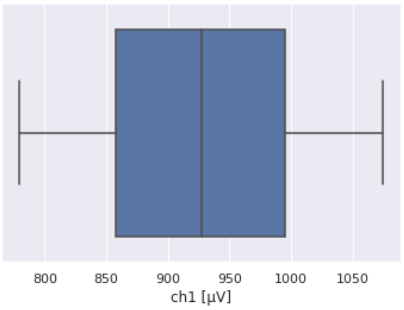
Sin embargo, decidimos realizar un gráfico de cajas para cada estado del paciente AA sesión 1, ch1 para ver si encontrábamos datos que superaran el 1.5 del rango intercuartil. Esta es una de las tantas formas en las que se pueden detectar valores atípicos en un conjunto de datos, aunque no es aconsejable utilizarlas para detectar outliers en series temporales ya que al no ser datos estáticos sus distribuciones se pueden modificar con el tiempo.

Realizamos gráficos de cajas para observar lo expuesto anteriormente.

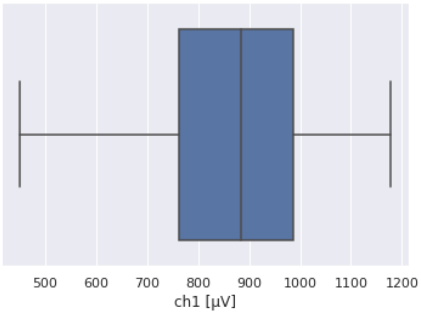
Estado 1



Estado 2



Estado 99



Analizando los boxplots no encontramos outliers en este nivel de análisis (para este conjunto de datos en esta fracción de tiempo).

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.

Decidimos utilizar el test t de student para determinar si existe diferencia estadísticamente significativa entre los valores de las medias del conjunto voltajes analizados para paciente AA sesión 1, canal 1.

El test plantea las siguientes hipótesis:

Ho= las medias son iguales

H1= las medias son diferentes

Definimos que dos muestras son estadísticamente diferentes si su p-valor es menor a 0,001 y utilizamos el módulo stats de la librería scipy para hacer los tests.

Los resultados que obtuvimos al comparar son los siguientes:

Comparación canal 1 estados 1 y 2

```
(statistic=3.024230086927614, pvalue=0.002497938989936647)
```

Comparación canal 1 estados 1 y 99

```
(statistic=30.550198782511902, pvalue=1.829466998169589e-197)
```

Comparación canal 1 estados 2 y 99

```
(statistic=28.212498127472703, pvalue=1.3153702295850544e-169)
```

De acuerdo con este test no se encontró evidencia suficiente para rechazar la hipótesis H0 cuando comparamos los estados 1 y 2, por lo que sus medias pertenecerían a la misma distribución de probabilidad. Por otra parte, se encontró con una evidencia del 99,9% que las medias de los 1 y 2 difieren estadísticamente de la media del estado 99, es decir cuando el paciente no está recibiendo ninguna frecuencia de estímulo.

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

Las principales conclusiones de estos análisis son que los valores de voltaje no presentan una distribución gaussiana para el paciente AA, sesión 1 canal 1 en el intervalo de tiempo analizado (30 segundos) para cada estado de estimulación o no estimulado. Utilizamos el test de

Kolmogorov-Smirnov para probar lo anterior. La distribución de probabilidad de los voltajes para cada uno de los estados se asemeja más bien a una distribución polimodal.

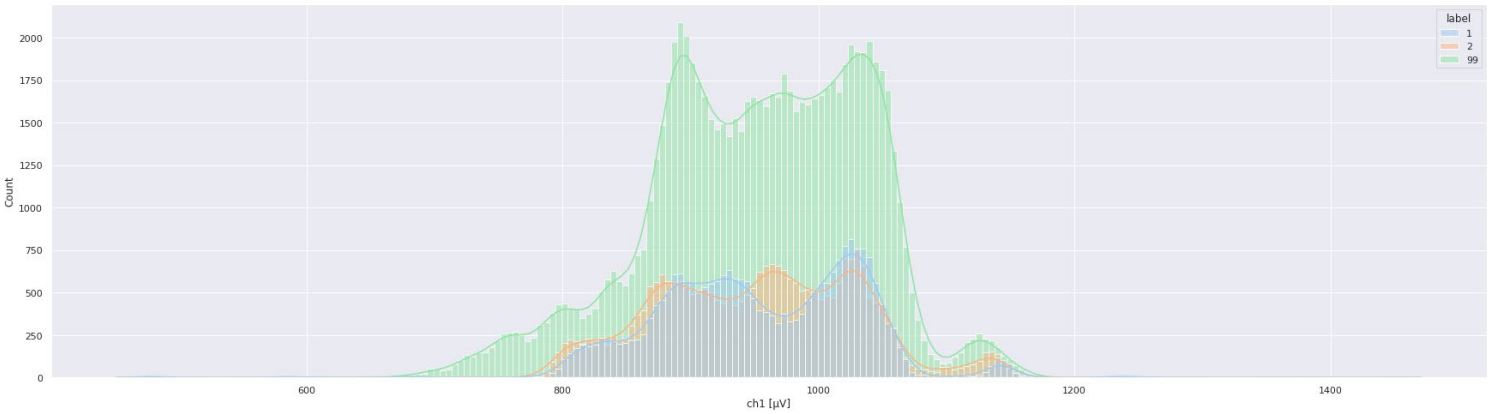
Además, no encontramos diferencia estadísticamente significativa entre las medias de los voltajes de los estímulos 1 y 2, pero sí entre cualquiera de estos estados de estimulación con respecto al control no estimulado etiquetado como 99. Si bien con este análisis no podemos distinguir diferencias entre los estímulos, por lo menos nos indica que al recibir cualquiera de ellos el paciente responde de manera diferente a cuando no esté en presencia de las luces.

Finalmente podemos decir que no encontramos datos faltantes ni outliers con los métodos empleados para este conjunto de datos.

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).

Histograma de distribución de valores de voltaje total para el paciente AA1 canal ch1



En el histograma mostrado podemos ver cómo se distribuyen los valores de voltajes para los distintos estados del paciente AA sesión 1 esta vez tomando el intervalo de tiempo completo de duración de la sesión. Observando las formas de las distribuciones notamos que no se asemejan a una normal. Otra vez encontramos más de un pico de elevada frecuencia y además las curvas están ladeadas hacia la izquierda. Para confirmar que la distribución no es gaussiana realizamos nuevamente el test de normalidad de Kolmogorov-Smirnov plateando las siguientes hipótesis:

H_0 = la muestra proviene de una distribución normal

H_1 = la muestra no proviene de una distribución normal

Establecimos un p-valor de 0.005 para rechazar H_0.

Los estadísticos obtenidos fueron:

Estado1 = ks = 0.06592516356030198, p-valor = 1.0542109040048149e-95

Estado 2 = ks =0.03892888953542273, p-valor = 2.0037299448272755e-36

Estado 99 = ks =0.04250755058493105, p-valor = 7.022657776100944e-135

Para los 3 estados es p-valor es menor a 0.005 por lo que no encontramos evidencia suficiente para aceptar H_0 según este test.

Para analizar el conjunto total de datos, aplicamos nuevamente el método describe() para cada estado del paciente para ver las medidas de centralidad como la media y la moda y de desviación (rango intercuartil y desviación estándar)

Los valores arrojados fueron los siguientes:

ESTADO 1

```
1 df_AA1_ch1_1.describe()
```

	ch1	time	label
count	25210.000000	25210.000000	25210.0
mean	954.391747	363.986869	1.0
std	77.301691	196.409915	0.0
min	469.690000	90.710000	1.0
25%	897.642500	193.741250	1.0
50%	951.330000	268.972500	1.0
75%	1018.260000	520.323750	1.0
max	1252.590000	674.910000	1.0

ESTADO 2

```
1 df_AA1_ch1_2.describe()
```

	ch1	time	label
count	27331.000000	27331.000000	27331.0
mean	953.729099	363.502762	2.0
std	78.625753	209.907422	0.0
min	776.390000	74.945000	2.0
25%	891.490000	182.582500	2.0
50%	956.880000	278.285000	2.0
75%	1015.480000	542.142500	2.0
max	1207.520000	689.495000	2.0

ESTADO 99

```
1 df_AA1_ch1_99.describe()
```

	ch1	time	label
count	85628.000000	85628.000000	85628.0
mean	951.981044	334.181962	99.0
std	85.349934	196.021652	0.0
min	451.280000	0.000000	99.0
25%	894.620000	159.593750	99.0
50%	957.720000	350.687500	99.0
75%	1018.890000	489.236250	99.0
max	1471.020000	690.840000	99.0

Nuevamente consideramos que la mediana sería un mejor estimador de la posición central que la media de las distribuciones de los valores de voltajes para los diferentes estados ya que

las mismas presentan curvas no simétricas y con valores extremos hacia la izquierda que tienden a desviar la media hacia esa dirección.

Luego evaluamos mediante test de student si existían diferencias estadísticamente significativas entre los valores de las medias de cada estado.

Las hipótesis planteadas son:

Ho= las medias son iguales

H1= las medias son diferentes

Definimos que dos muestras son estadísticamente diferentes si su p-valor es menor a 0,001 y utilizamos el módulo stats de la librería scipy para hacer los tests.

Los resultados que obtuvimos al comparar son los siguientes:

Comparación canal 1 estados 1 y 2

```
(statistic=0.9729509715440818, pvalue=0.3305821493762561)
```

Comparación canal 1 estados 1 y 99

```
(statistic=4.024874946717955, pvalue=5.704343319624697e-05)
```

Comparación canal 1 estados 2 y 99

```
(statistic=3.003508558407704, pvalue=0.0026694447237093984)
```

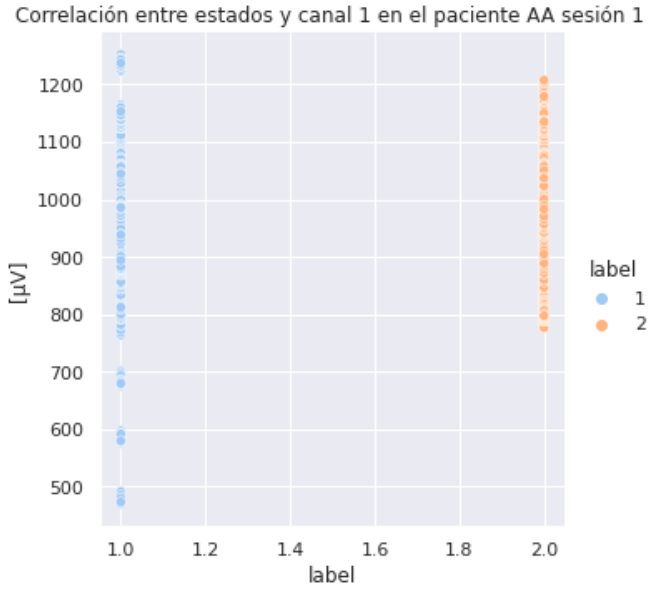
De acuerdo con este test no se encontró evidencia suficiente para rechazar la hipótesis H0 cuando comparamos los estados 1 y 2, por lo que sus medias pertenecerían a la misma distribución de probabilidad. De igual manera tampoco se encontró evidencia suficiente para rechazar H0 entre las medias de los estados 2 y 99. Por otra parte, se encontró con una evidencia del 99,9% que las medias de los 1 y 99 difieren estadísticamente entre sí.

Al analizar la totalidad de los datos para este paciente con el canal ch1 no obtuvimos conclusiones diferentes de las mencionadas para el análisis de los datos acotados a los primeros 30 segundos del experimento.

La única diferencia que pudimos notar es que al comparar las medias con el test de student cuando tomamos la totalidad de los datos, los estados 2 y 99 no difieren estadísticamente como lo hacían cuando analizábamos una sección de los voltajes. Nuevamente esta discrepancia puede tener origen en el hecho de que son datos pueden variar su distribución de probabilidad con el tiempo.

> 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.

Para visualizar la dependencia entre dos variables elegimos los gráficos de dispersión de puntos



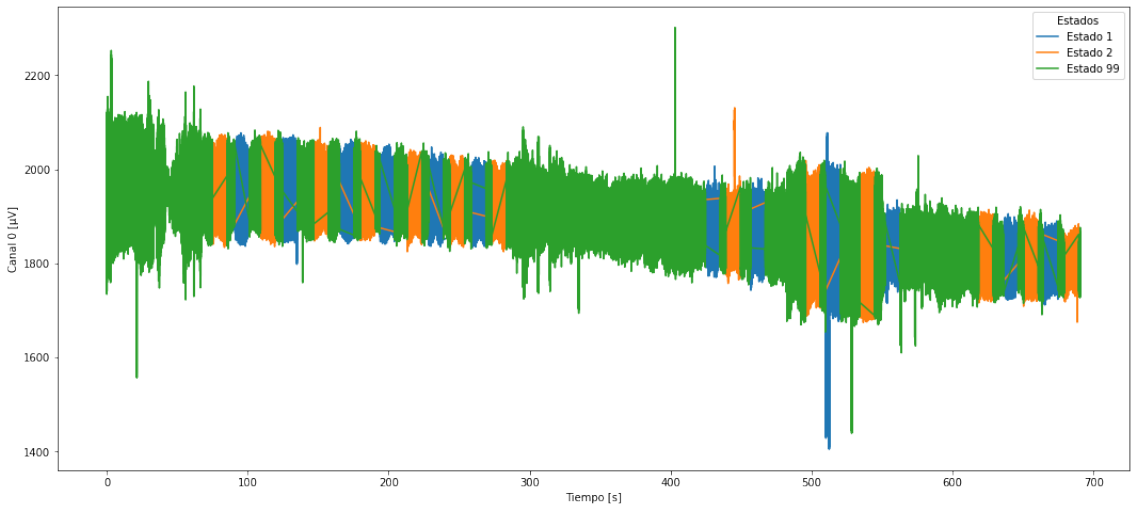
Con este primer dot plot quisimos ver cómo se posicionaban los estados 1 y 2 con respecto a los valores de voltaje para el canal 1 del paciente AA1. No nos muestra mucha más información comparada con los histogramas ya que vemos que en rango aproximado de valores de voltajes entre 800 y 1200 microvoltios hay una superposición de ambos estados.

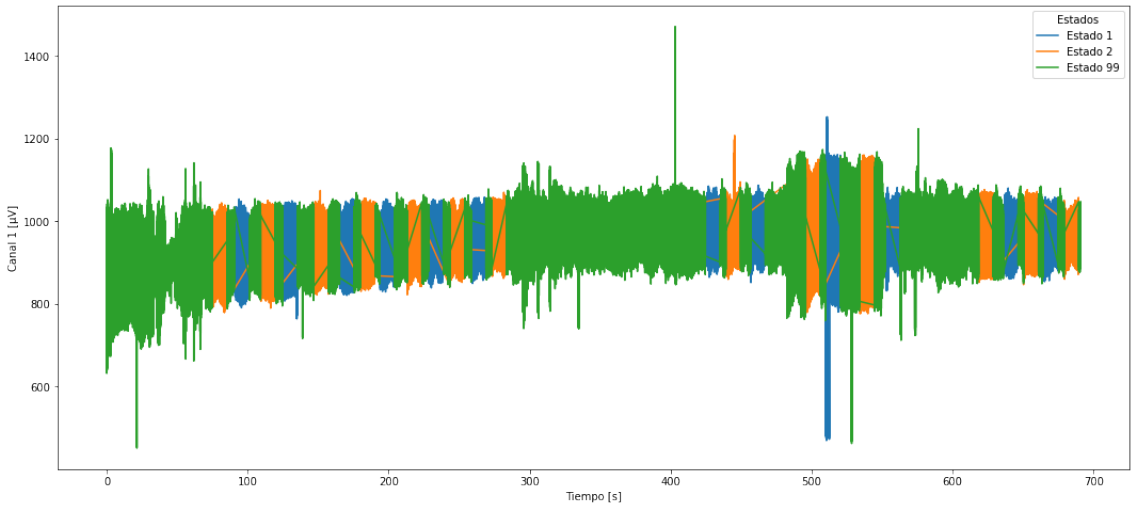
Por otra parte, decimos realizar un análisis mas amplio tomando los valores medios de los voltajes de todos los canales de todos los pacientes y sesiones para ver si el estado 1 se correlacionaba con un rango particular de microvoltios y el estado 2 con otro. Como podemos apreciar en el gráfico de dispersión no existe ningún patrón que indique algún tipo de correlación entre la señal registrada y el estímulo. Vemos que hay una dispersión similar de todos los voltajes medios de los diferentes canales para los dos estados.

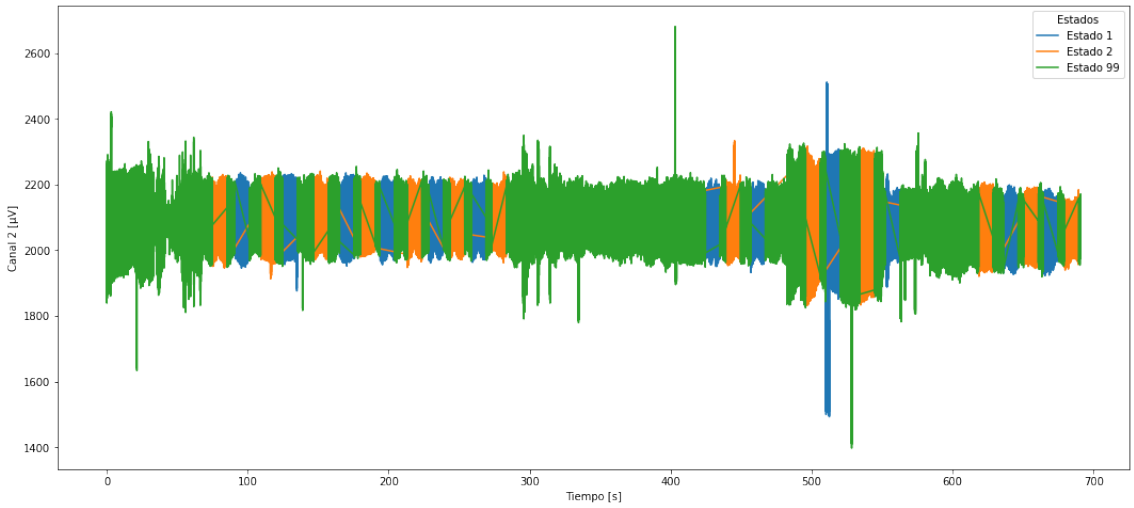
c) Para cada uno de los estados, los valores de voltaje a lo largo del tiempo, ¿varían con alguna tendencia?

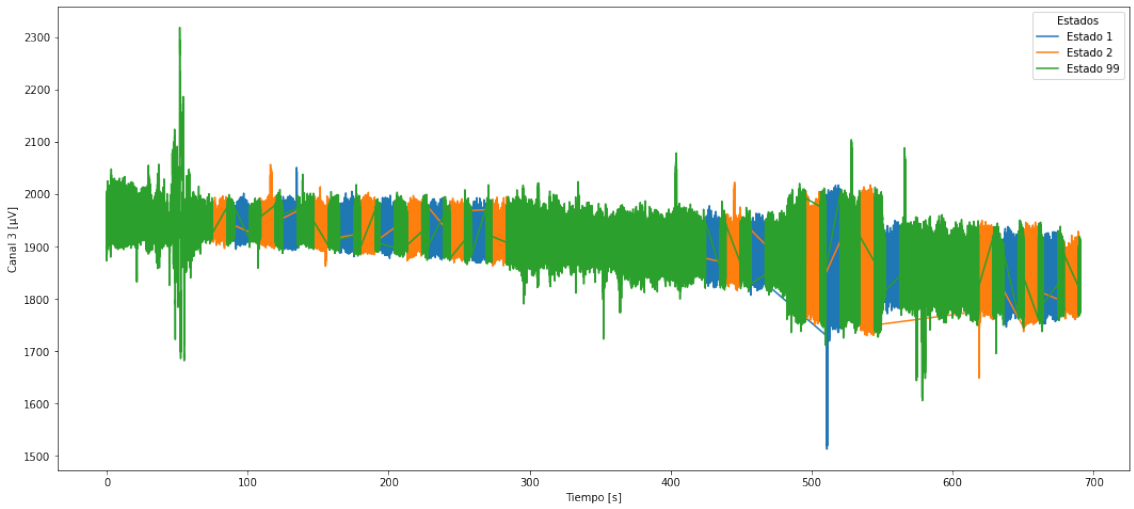
Seguimos analizando el caso particular del paciente AA1 y mostramos para cada uno de los 4 canales cómo varía el voltaje en función del tiempo. Podemos notar que tanto en el canal 0 como en el 3 los voltajes tienden a decrecer a medida que pasan los segundos, pero en los canales 1 y 2 los voltajes parecen fluctuar alrededor de una media constante. Las tendencias

hacia la disminución del voltaje observada en algunos canales pueden deberse a algún error sistemático en los instrumentos de medición que generan un sesgo en la determinación de los microvoltios.









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

Como mencionamos previamente, no se encontró evidencia suficiente para determinar que las distribuciones de probabilidad de los voltajes pertenezcan a una normal.

No encontramos correlación entre los valores de voltaje medios y los estados de estimulación (frecuencias 1 y 2) para todo el conjunto de individuos estudiados.

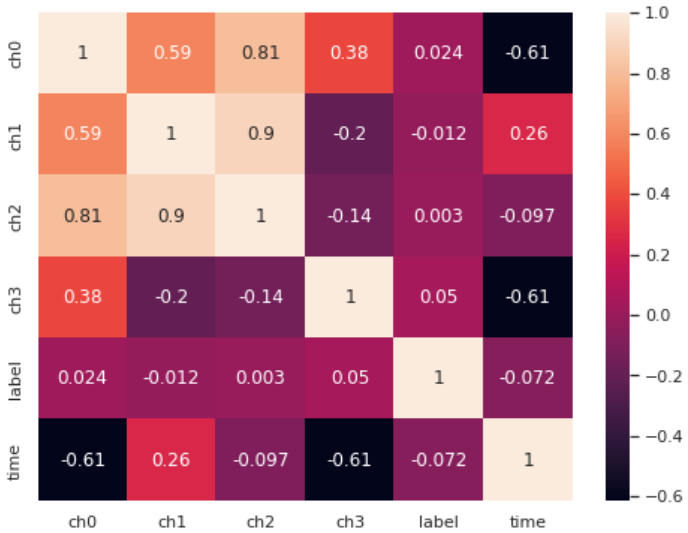
Los valores de voltaje en función del tiempo de duración de la sesión tienden a disminuir para ciertos canales en el paciente AA sesión 1, pero para otros canales se mantienen constantes.

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

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

La siguiente matriz de correlación para muestra en cada casilla el coeficiente de Pearson correspondiente a cada relación.

PACIENTE AA1

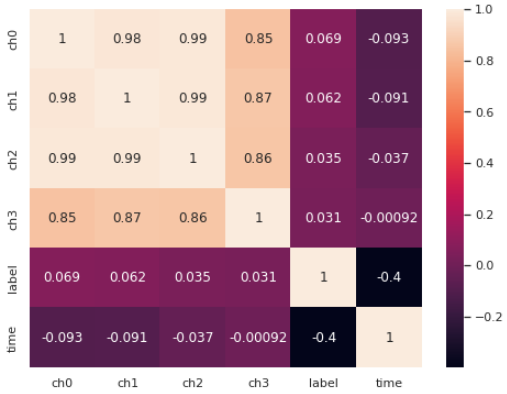


El coeficiente de correlación de Pearson es una medida de relación lineal entre dos variables independientes. Para calcularse tiene en cuenta las varianzas y la covarianza de las variables involucradas. Toma un rango de valores que va de -1 a 1. Los valores que rondan al cero indican una relación lineal débil, mientras que los cercanos a los extremos positivo y negativo indican una relación lineal más fuerte entre las variables.

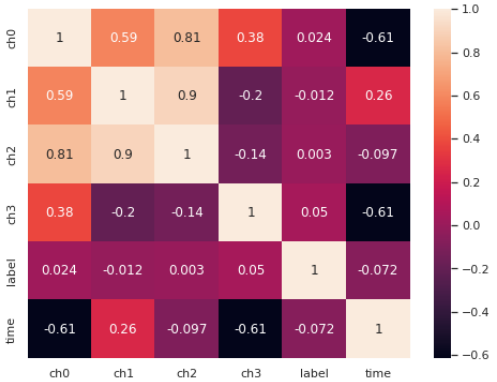
Como podemos apreciar en la matriz de correlación las variables que se corresponden con los canales 0 y 3 tienen un coeficiente de -0,61 con respecto al tiempo mientras que los canales 1 y 2 tienen coeficientes más cercanos al cero. Esto tiene relación con lo observado en los gráficos de voltaje en función del tiempo para cada canal. Allí vimos que los canales 0 y 3 tienen a disminuir sus valores en función del tiempo y por lo tanto en la matriz de correlación obtenemos coeficientes negativos y de valor absoluto más grandes que para los canales 1 y 2 cuyos voltajes no se modificaban notoriamente con los segundos.

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?

PACIENTE AA0

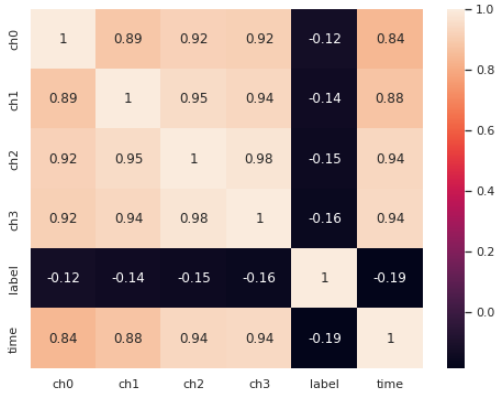


PACIENTE AA1

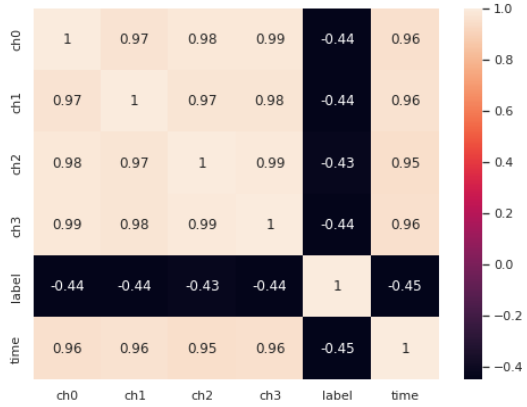


PACIENTE AA2

PACIENTE HA1

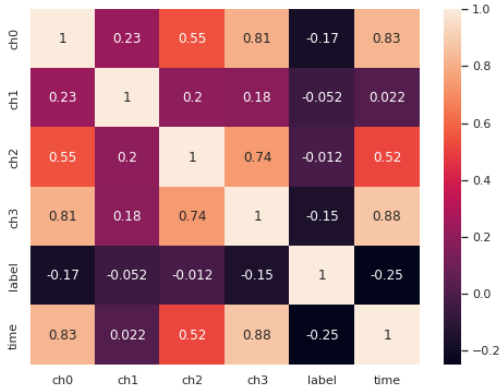


PACIENTE JA1



PACIENTE JA2

PACIENTE MA1



Al aplicar la matriz de correlación en todos los pacientes y sesiones vemos que hay casos donde hay una correlación lineal mas fuerte expresada a través del coeficiente de Pearson entre los voltajes de los canales y el tiempo. Esos casos corresponden a los pacientes AA sesión 2, HA1 y JA sesión 1. En otras sesiones de los mismos pacientes dicha correlación no es tan pronunciada, como el caso del paciente AA sesión 0 donde en todos los canales hay una correlación lineal débil. Para el resto de los casos vemos que hay canales dentro de una sesión que tienen mayor correlación con el tiempo comparados con otros.

Como conclusión, para trabajar, no es necesario tener en cuenta todos los canales, podemos tomar alguno de los que tienen correlación con el tiempo y otro que no la tenga y evaluar su comportamiento.

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?

Elegimos representar la distribución de las mediciones de voltaje de los canales ch1 para los diferentes paciente y sesiones. Con solo ver las distribuciones notamos que no hay una consistencia entre los diferentes pacientes y sesione para un mismo estado.

Tomamos al azar algunas de estas distribuciones para comparar las medias entre los pacientes para un mismo estado usando el test de student y definiendo el p-valor menor o igual a 0.001 para que la diferencia sea significativa.

Recordemos que las hipótesis a tener en cuenta son:

Ho= las medias son iguales

H1= las medias son diferentes

Obtuvimos los siguientes resultados al comparar los pacientes AA1 y HA1

ESTADO 1 CANAL 1

```
(statistic=-24.534790726616343, pvalue=4.8361200439161636e-132)
```

ESTADO 2 CANAL 1

```
(statistic=-4.116062873692886, pvalue=3.86037655558061e-05)
```

ESTADO 2 CANAL 99

```
(statistic=44.85081655464578, pvalue=0.0)
```

Vemos que hay comparaciones que son estadísticamente significativas entre ellas, como las del estado 1 y 99 y otras en otras no encontramos evidencia suficiente para rechazar H0, como en el caso del estado 2 para el canal 1.

El test de student realizado da un soporte estadístico a la conclusión obtenida al inspeccionar visualmente los histogramas, es decir, hay medias en las que no encontramos evidencias suficientes para decir que difieren entre los pacientes para un mismo estado y canal y otras que sí.