Actividad 7

Visualización de datos con la biblioteca Seaborn.

Jorge Benz Olguín Aguilar

División de Ciencias Exactas, Departamento de Física

Universidad de Sonora

16 de mayo de 2019

La biblioteca de visualización de datos Seaborn, está basada en la biblioteca Matplotlib para Python. Seaborn está fuertemente ligada a Pandas y a sus estructuras de datos, y se pueden producir gráficas de forma más sencilla y más atractivas visualmente. Seaborn ha sido desarrollada por Michael Waskom de la Universidad de Nueva York. [1]

La práctica solicita que obtengamos dos gráficas, una con la biblioteca pandas y otra con matplotlib. Que comparemos cual de las dos formas de trabajar es más comoda y eficiente. Lo primero que haremos será descargar los datos utilizando el método read_csv(), convertir los datos obtenidos a un Data Frame el cual podamos manipular con métodos propios de él. Una vez realizado lo anterior procederemos a la 'limpieza' de Data Frame, una actividad que por si sola llega a representar la mayoría del tiempo destinado al analisis de datos. Como los datos que obtuvimos están prácticamente listos para utilizarse no fue necesario tal cantidad de tiempo. Utilizamos el método drop() de pandas para eliminar las columnas que no necesitamos; así como, el primer renglón, que representa las unidades de las columnas.

Ahora bien, como ya tenemos listo el Data Frame para trabajar, utilizaremos la función corr() para crear una matriz de correlación y poder analizar la dependencia que existe entre las variables. Pero de que estamos hablando cuando decimos correlación; en probabilidad y estadística, la correlación indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos

variables estadísticas. Se considera que dos variables cuantitativas están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los valores homónimos de la otra: si tenemos dos variables (A y B) existe correlación entre ellas si al disminuir los valores de A lo hacen también los de B y viceversa. La correlación entre dos variables no implica, por sí misma, ninguna relación de causalidad[3]

La relación entre dos variables cuantitativas queda representada mediante la línea de mejor ajuste, trazada a partir de la nube de puntos. Los principales componentes elementales de una línea de ajuste y, por lo tanto, de una correlación, son la fuerza, el sentido y la forma:

1. La fuerza extrema según el caso, mide el grado en que la línea representa a la nube de puntos: si la nube es estrecha y alargada, se representa por una línea recta, lo que indica que la relación es fuerte; si la nube de puntos tiene una tendencia elíptica o circular, la relación es débil.

2. El sentido mide la variación de los valores de B con respecto a A: si al crecer los valores de A lo hacen los de B, la relación es directa (pendiente positiva); si al crecer los valores de A disminuyen los de B, la relación es inversa (pendiente negativa).

3. La forma establece el tipo de línea que define el mejor ajuste: la línea recta, la curva monotónica o la curva no monotónica

Existen diversos coeficientes que miden el grado de correlación, adaptados a la naturaleza de los datos. El más conocido es el coeficiente de correlación de Pearson (introducido en realidad por Francis Galton), que se obtiene dividiendo la covarianza de dos variables entre el producto de sus desviaciones estándar. En estadística, el coeficiente de correlación de Pearson es una medida lineal entre dos variables aleatorias cuantitativas. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables.[4]

En las siguientes dos gráficas, la primera hecha con Seaborn y la segunda con Matplotlib, podemos hacer un comparativo de la dificultad que represento elaborar una y otra. A simple vista parecen iguales, nos brindan la misma información y son muy buenas si hablamos de su estética, pero, realizar la gráfica con Seaborn fue sumamente fácil mientras que con matplotlib fue un proceso un poco mas complicado. Para la primera figura básicamente solo tuvimos que utilizar la función heatmap() de la biblioteca de Seaborn y listo. La segunda figura: determinamos los valores de los ejes, se agregaron los nombres a las las variables en los ejes, se roto 90 grados las

etiquetas en el eje x y por último se gráfico en una matriz de correlación la dependencia entre las variables, como ya habiamos mencionado.

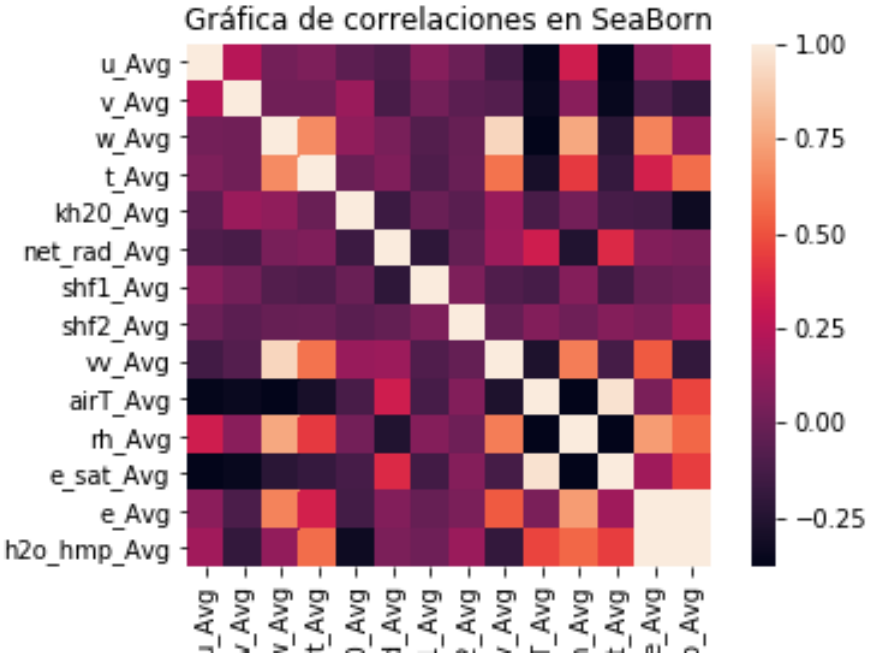

Figura 1:

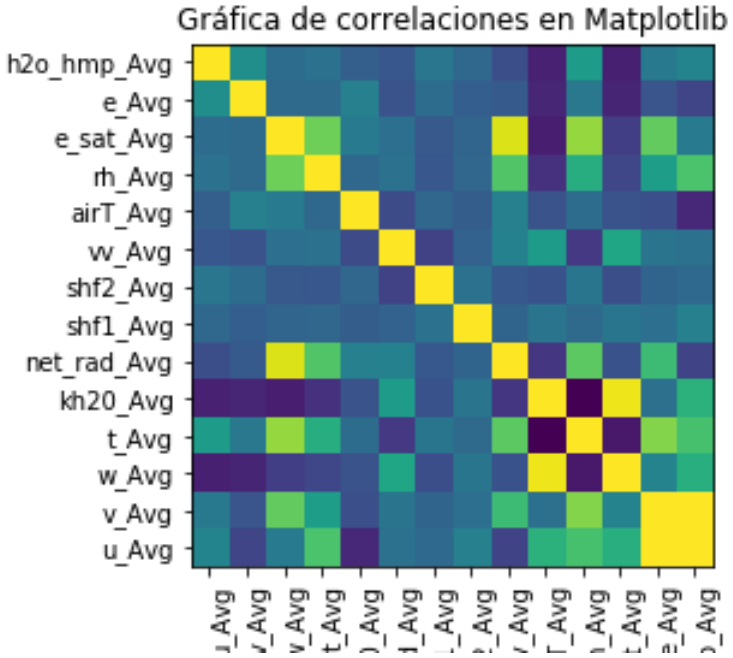

Figura 2:

El valor del índice de correlación varía en el intervalo [-1,1], indicando el signo el sentido de la relación:

1. Si r = 1, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada relación directa: cuando una de ellas aumenta, la otra también lo hace en proporción constante.

2. Si 0 ¡r ¡1, existe una correlación positiva.

3. Si r = 0, no existe relación lineal. Pero esto no necesariamente implica que las variables son independientes: pueden existir todavía relaciones no lineales entre las dos variables.

4. Si -1 ¡r ¡0, existe una correlación negativa.

5. Si r = -1, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada relación inversa: cuando una de ellas aumenta, la otra disminuye en proporción constante

Con la información antes mencionada podemos hacer un breve análisis de las siguientes relaciones entre variables que presentan un coeficiente de correlación mayor a 0.6. En la figura 3 podemos observar que conforme el coeficiente de correlación se acerca a cero, en este caso es 0.667, la relación entre las variables empieza a desaparecer y lo observamos en la falta de linealidad de los puntos gráficados y lo mismo ocurre para la figura 4.

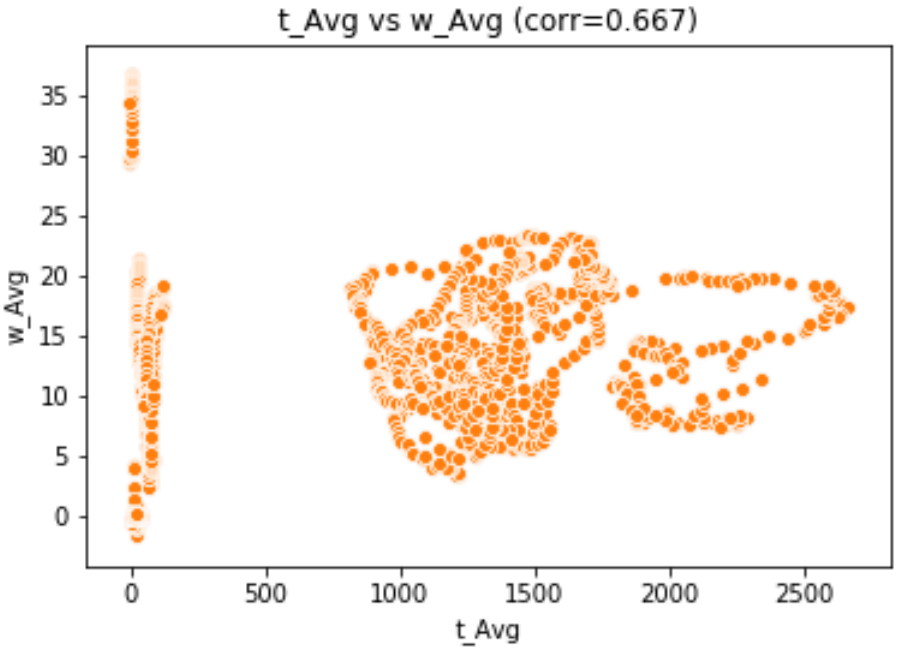

Figura 3:

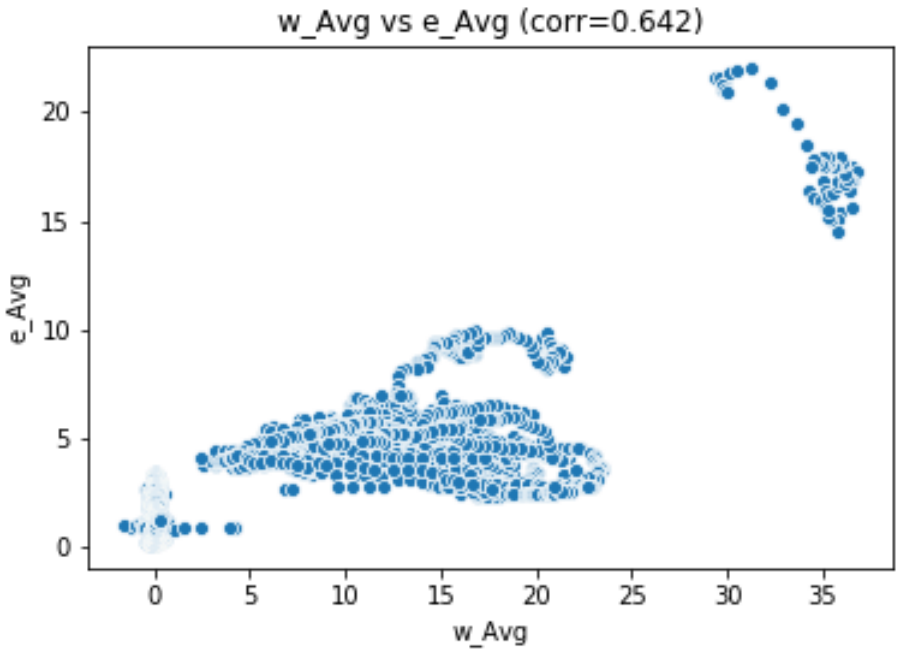

Figura 4:

El coeficiente de la figura 5 0.924 nos proporciona una forma mas parecida a una linea recta, un claro indicativo que la dependencia entre las dos variables es buena y positiva (pendiente positiva)

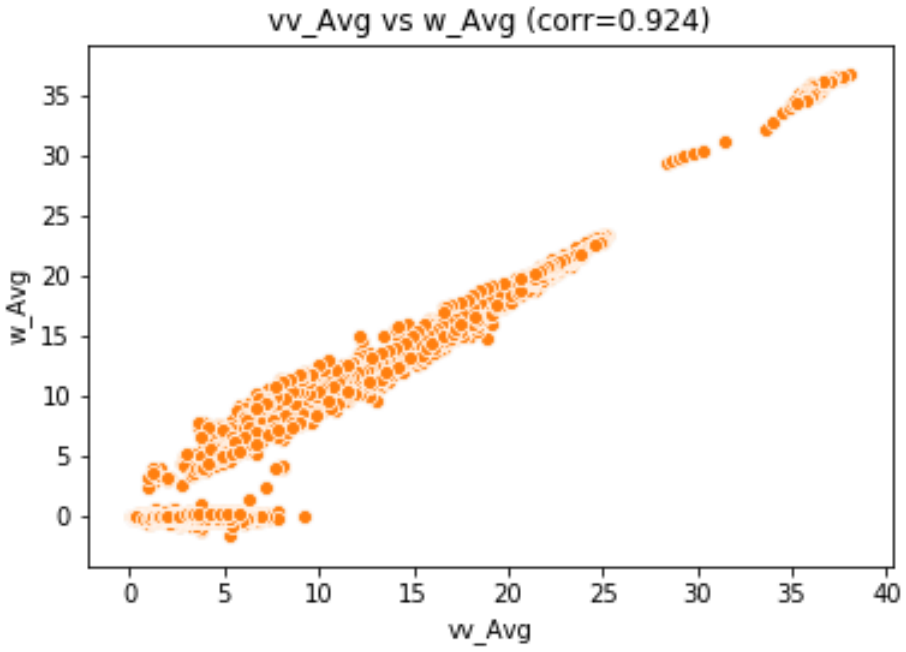

Figura 5:

Para las siguientes dos figuras es interesante notar como la forma que nos da la distribución de

los puntos, correlación 0.963 para la figura 6 y 0.760 para la figura 7, nos permite identificar si la dependencia es alta o baja sin prestarse a confusión por la curva que presentan.

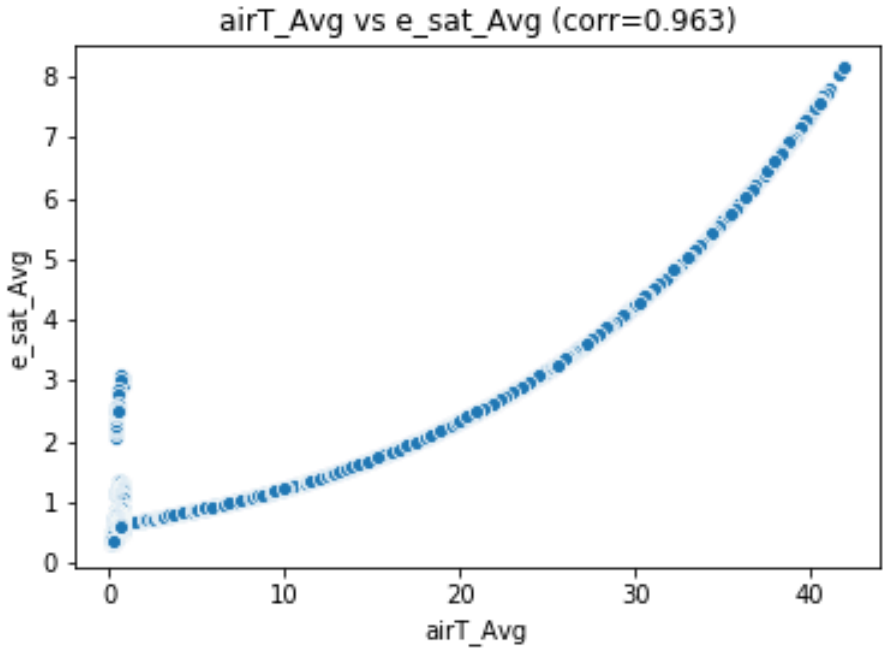

Figura 6:

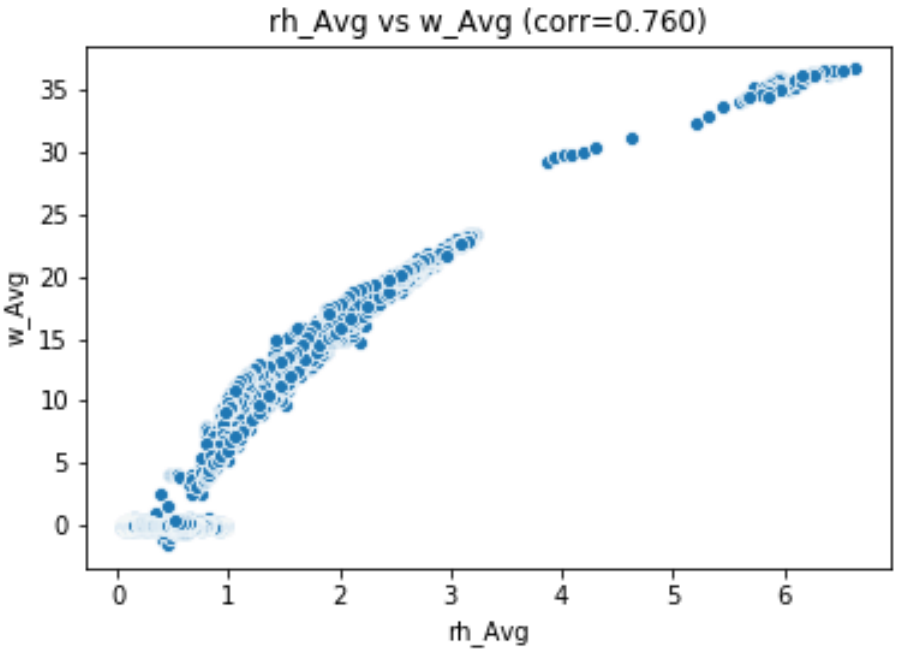

Figura 7:

Esta gráfica encaja perfectamente en la explicación que dimos sobre las dos primeras figuras.

.90.0.9

Figura 8:

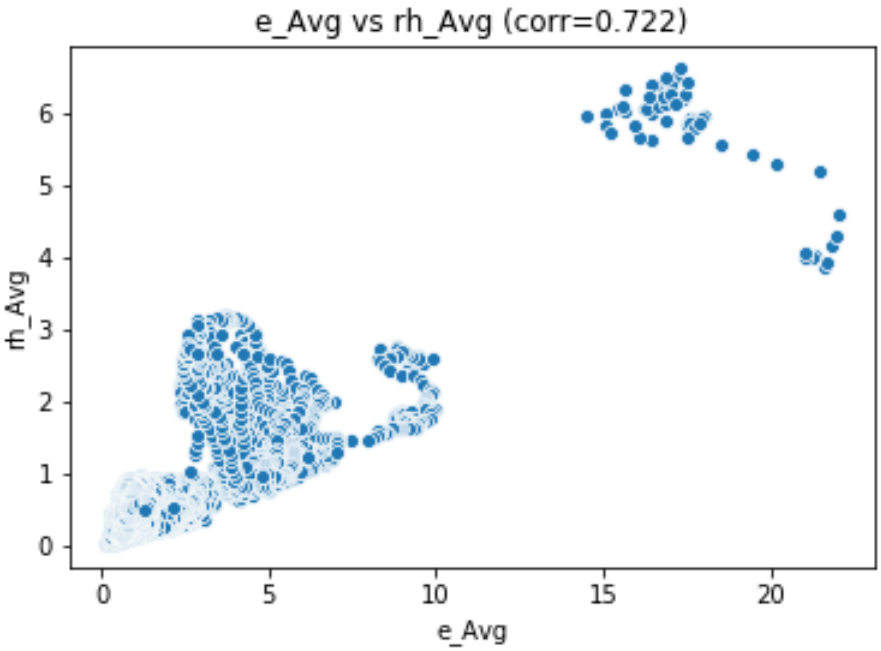

Figura 9:

La figura 10, con un coeficiente de correlación de 0.999, con una linea recta casi perfecta, es el mejor ejemplo de la dependencia entre variables es casi absoluta; esto es, cuando una crece la otra lo hace en la misma proporción.

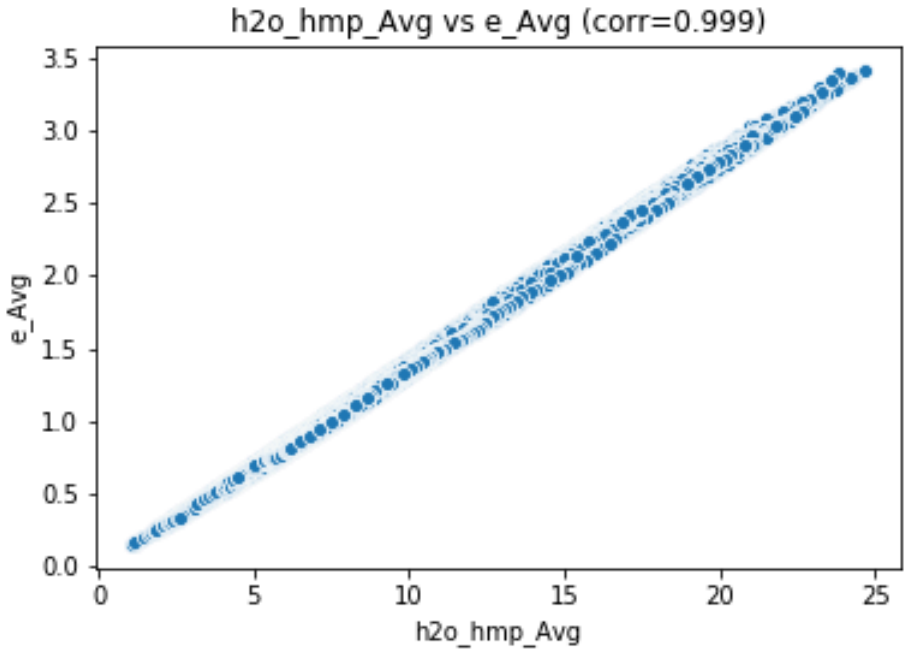

Figura 10:

Conclusiones

El manejo de grandes cantidades de datos se ha convertido en una herramienta básica de trabajo. El análisis de ellos no sería posible sin la utilización de los distintos tipos de gráficas que las bibliotecas especializadas nos permiten obtener. Mientras más especializada sea la biblioteca más fácil será construir los distintos tipos de gráficas, es por esta razón que para análisis estadístico la biblioteca de Seaborn es superior a la de matplotlib en lo que a eficiencia se refiere.

Referencias

[1] http://computacional1.pbworks.com/w/page/132439518/Actividad7

[2] http://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html?highlight=c

[3] Correlación. (2019, 13 de febrero). Wikipedia, La enciclopedia libre.

[4] Coeficiente de correlación de Pearson. (2019, 22 de abril). Wikipedia, La enciclopedia libre.

[5] https://seaborn.pydata.org/generated/seaborn.lmplot.html?highlight=show