Reporte de la Actividad 7

Marco Antonio Cabello López
Grupo 1

Domingo 17 de Marzo del 2019

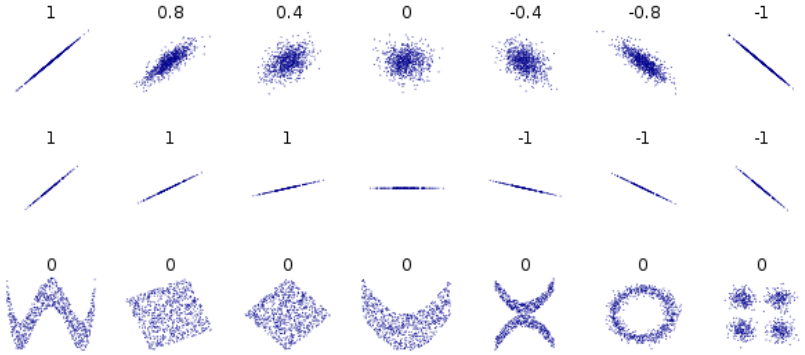
1 Introducción

Primeramente y como introducción, en esta actividad analizamos datos del año 2009 de una estación de meteorología ubicada en un campo de Nogal, para visualizar las correlaciones entre distintos parámetros medidos. Se comparó el uso de las bibliotecas SeaBorn y Matplotlib para la elaboración de gráficos, y se graficaron las variables que mostraban un índice de correlación mayor que 0.5.

2 Desarrollo de la actividad

2.1 Correlación

Consideramos la medida de dependencia entre dos cantidades es el "Coeficiente de correlación", Se obtiene dividiendo la covarianza de las dos variables por el producto de sus desviaciones estándar. El coeficiente de correlación es +1 en el caso de una relación lineal (correlación) es directa y perfecta (creciente), y -1 en el caso de una relación lineal (anticorrelación) es inversa y perfecta (decreciente), y algún valor en el intervalo abierto (-1, +1) en todos los demás casos, lo que indica el grado de dependencia lineal entre las variables. A medida que se acerca a cero, hay menos de una relación más debil será la correlación. Cuanto más cerca esté el coeficiente de -1 o +1, más fuerte será la correlación entre las variables.



2.2 Heat Map

Es una representación gráfica de datos donde los valores individuales contenidos en una matriz se representan como colores con sus respectivas variaciones de intensidad. En esta actividad el uso de Heat Map será útil para ubicar las variables con mayor correlación del data frame, entonces de la gráfica de correlaciones, podremos observar entre que variables hay alguna relación lineal entre ellas (sin incluir la relación con ellas mismas).

2.3 Metodología

Comenzamos a trabajar con el archivo, primero fue necesario descargar librerías para la visualización y el análisis de datos :

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import math
```

El archivo a leer contiene caracteres especiales que no puede reconocer la librería de Pandas, utilizamos el lector de Python:

```
df = pd.DataFrame(pd.read_csv("meteo-nogal-09.csv", engine="python"))
```

El documento contaba con varias columnas nombradas "unnamed" que no contenían datos relevantes, así que las eliminamos para quedarnos con un data frame más pequeño:

```
df=df.drop(df.columns[df.columns.str.contains('unnamed:',case = False)],axis = 1)
```

El archivo no contaba con una variable de datetime, entonces la creamos a partir de dos columnas que contienen fechas y horas:

```
df["FECHA"] = df["DATE"] + " "+ df["TIME"]
```

Convertimos las variables object a float64 para poder operar con los datos de forma numérica:

```
df[df.columns[0:14]] = df[df.columns[0:14]].apply(pd.to_numeric, errors='coerce')
```

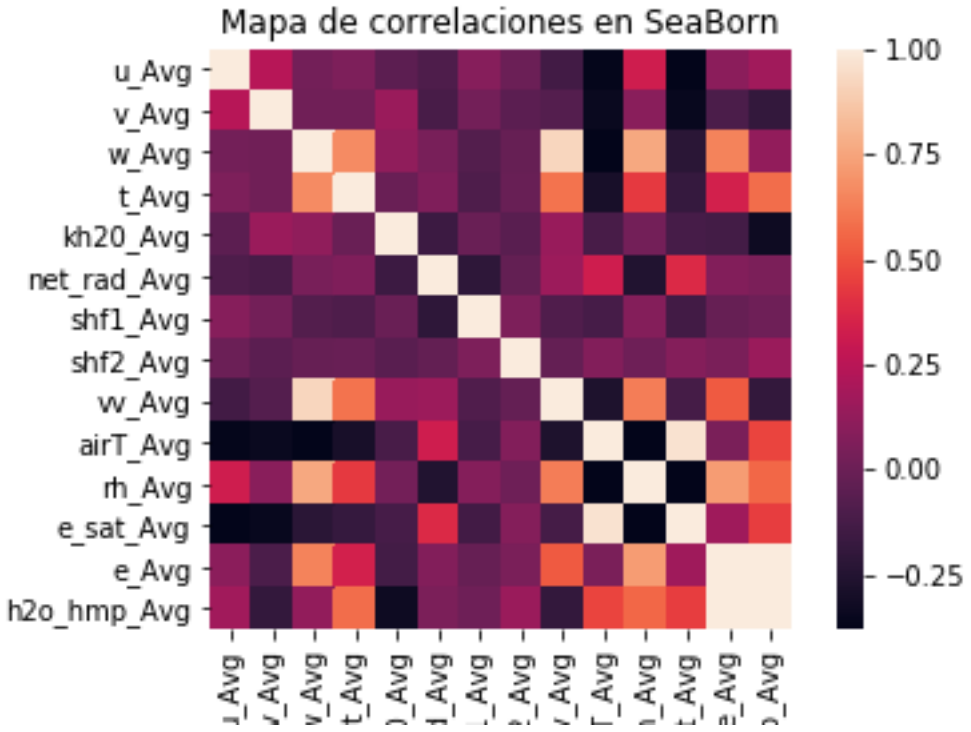
Utilizamos la función corr para encontrar las correlaciones entre las variables:

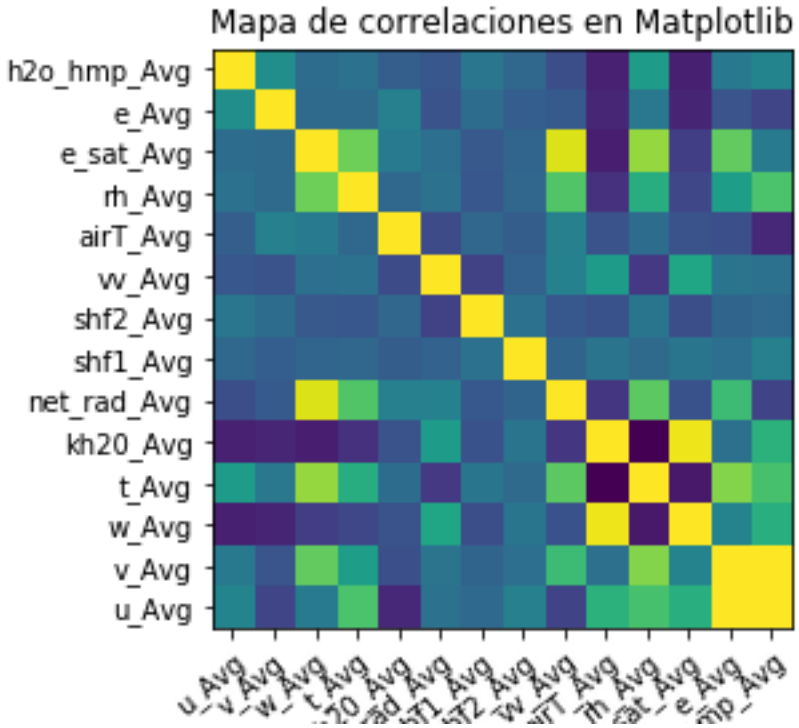
```
df_corr = df.corr(method='pearson', min_periods=1))
```

Posteriormente, creamos dos gráficas del tipo Heat Map, una mediante la bilbioteca Seaborn y otra mediante la biblioteca Matplotlib.
Finalmente, graficamos las correlaciones cuyo valor absoluto era mayor a 0.5.

2.4 Resultados

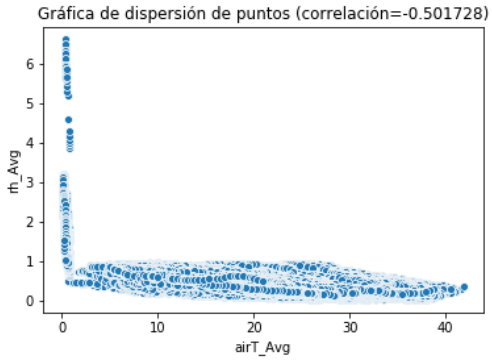
Gráficas de Heat Map elaboradas con Seaborn y Matplotlib:

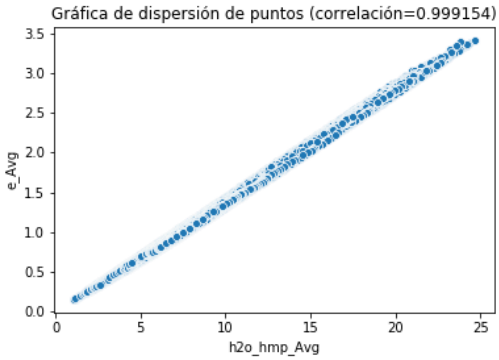




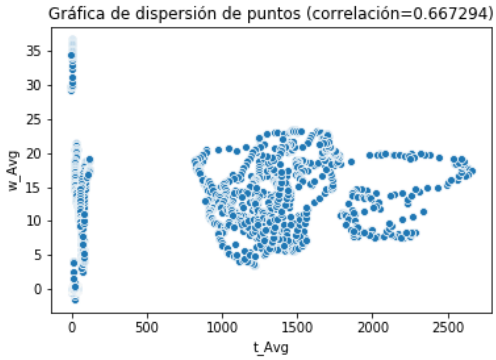
Podemos observar que las dos gráficas son similares en apariencia, el diagrama de Seaborn cuenta con una barra (a la derecha) que indica la correlación correspondiente a cada color, lo cual no se puede lograr en Matplotlib. Por este hecho la de Seaborn es mejor.

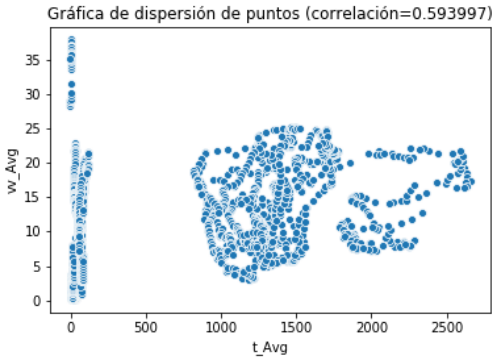
Las correlaciones cuyo valor absoluto era mayor a 0.5 fueron 13 en total. Entonces las gráficas de dispersión de puntos con sus respectivas correlaciones, una en el que la anticorrelación es muy cercana a -0.5 (izquierda) y otra en el que la correlación es muy cercana a 1 (derecha).

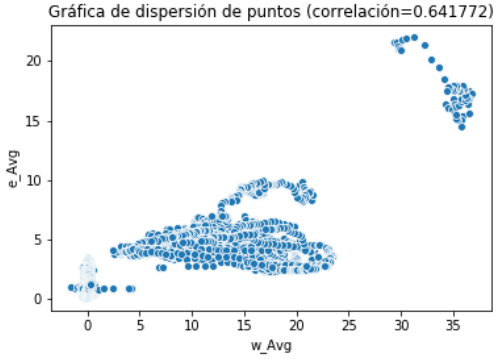


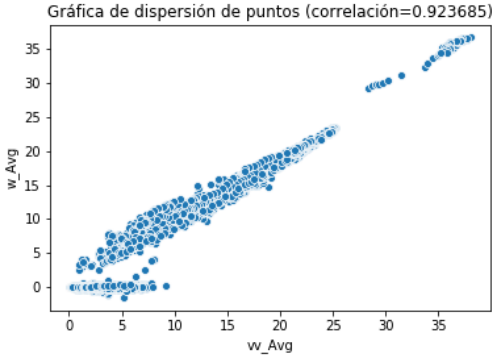


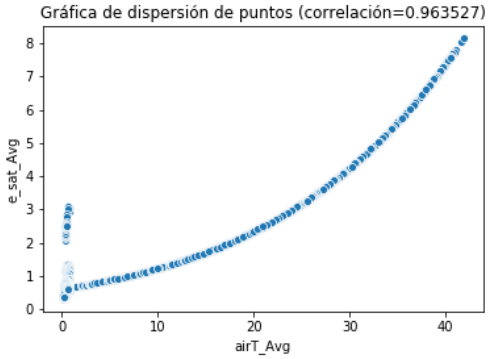
Gráficas de dispersión de puntos.

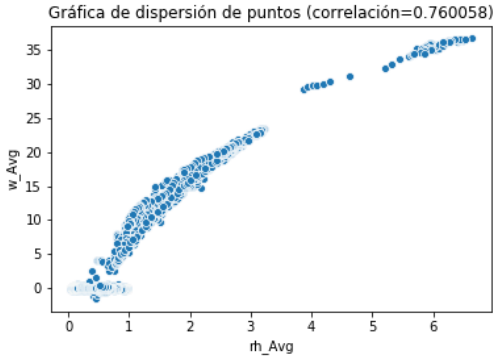


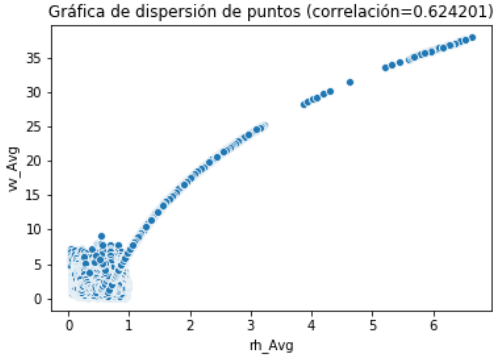


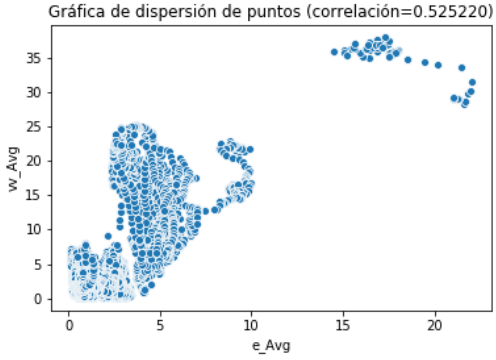


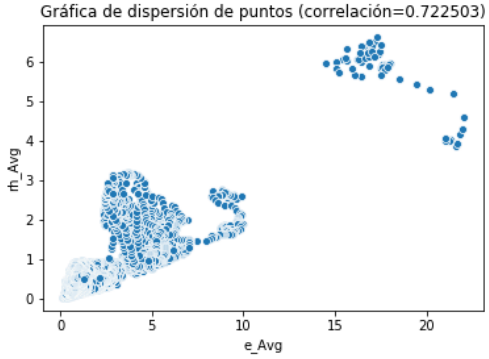


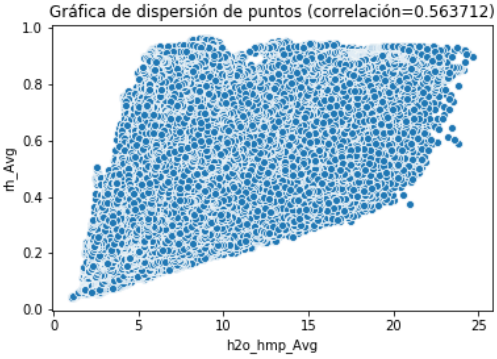

Gráfica de dispersión de puntos (correlación=0.963527)
e_sat_Avg
airT_Avg

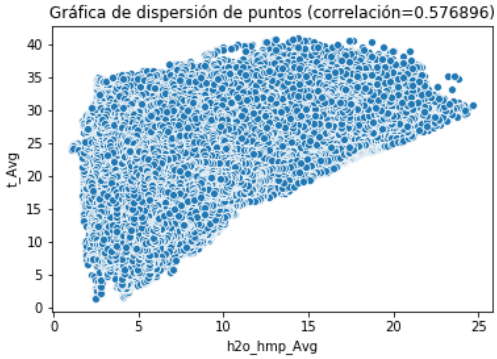

Gráfica de dispersión de puntos (correlación=0.760058)
w_Avg
rh_Avg


Gráfica de dispersión de puntos (correlación=0.624201)
vv_Avg
rh_Avg


Gráfica de dispersión de puntos (correlación=0.525220)
vv_Avg
e_Avg


Gráfica de dispersión de puntos (correlación=0.722503)
rh_Avg
e_Avg


Gráfica de dispersión de puntos (correlación=0.563712)
rh_Avg
h2o_hmp_Avg


Gráfica de dispersión de puntos (correlación=0.576896)
t_Avg
h2o_hmp_Avg

3 Conclusiones

Podemos observar en las gráficas de dispersión de puntos, que a medida que la correlación se aproxima más a -1 o +1 la relación entre las variables es mayor, mientras que al aproximarse al valor de 0 las variables estan menos relacionadas (por lo que podemos discernir la linealidad entre las distintas gráficas que obtuvimos), tambien utilizamos las bibliotecas Seaborn y Matplotlib al momento de graficar los Heat Map, en estas gráficas notamos inmediatamente la facilidad que tiene SeaBorn en comparación a Matplotlib, ya que ocupamos muchas más líneas de código para elaborar el Heat Map en Matplotlib, y además no contenía la barra lateral derecha que proporciona SeaBorn.

4 Referencias

- Correlación. Recuperado de:
  https://en.wikipedia.org/wiki/Correlation_and_dependence
  https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html
- Matplotlib. Recuperado de:
  https://matplotlib.org/
- Seaborn; Scatter Plot. Recuperado de:
  https://seaborn.pydata.org/
  https://python-graph-gallery.com/seaborn/