ANÁLISIS EXPLORATORIO DE DATOS USANDO DATOS DEL NDBC (*NATIONAL DATA BUOY CENTER*) DE ESTADOS UNIDOS

Cristian F. Zapata; Belky Alejandra

Cfzapatag@uqvirtual.edu.co; belkya.marulandac@uqvirtual.edu.co

RESUMEN

En este trabajo se hizo un análisis exploratorio de datos utilizando datos del NDBC de Estados Unidos, en particular a los datos de la boya de la estación 44025 ubicada en Long Island (sur de Islip) Nueva York. El análisis pretendía descubrir cuáles de las variables físicas (entre las que puede tener información la boya), incidían más en la creación de olas. En este trabajo se muestra cómo se abordó la actividad, y se explica, utilizando conceptos físicos, los resultados obtenidos. Finalmente se concluyó (entre otras cosas) que la variable que poseía mayor correlación con la creación de olas era la velocidad del viento. El lector puede tener acceso a los códigos utilizados para este trabajo, accediendo al repositorio en GitHub.

INTRODUCCIÓN

Antaño, los reyes tenían súbditos que enlistaban el *tiempo meteorológico* para definir el *clima* de un país. De esta manera, se podía saber cuándo sembrar, guardar reservas alimentarias, y cuándo consumirlas. Y no es muy diferente a lo que hoy día se hace, incluso comúnmente las personas consultan el tiempo meteorológico en internet para así discernir si al momento de salir de casa, llevan paraguas o se ponen la chaqueta que más combina con sus zapatos. Entonces, *la climatología* es una de las actividades más antiguas (e importantes) para el ser humano. Y su razón de ser es casi que inmediata: estudiarla permite predecir acontecimientos, con lo cual se puede tomar medidas de precaución frente a esos posibles eventos.

En este trabajo se hizo un análisis exploratorio de datos, a los datos obtenidos de la boya de la estación 44025, ubicada en Long Island (al sur de Islip) en Nueva York. Además, se explicó el proceso y las pautas con las que se abordó la actividad.

PROCEDIMIENTO

La pregunta que se quiso responder con este análisis exploratorio de datos fue: *¿cuáles de los parámetros que mide la boya de la estación 44025, ubicada en Long Island (al sur de Islip) NY, presenta mayor correlación con las características de las olas?* En otras palabras: qué parámetros físicos inciden (y cuáles no) en la formación de olas.

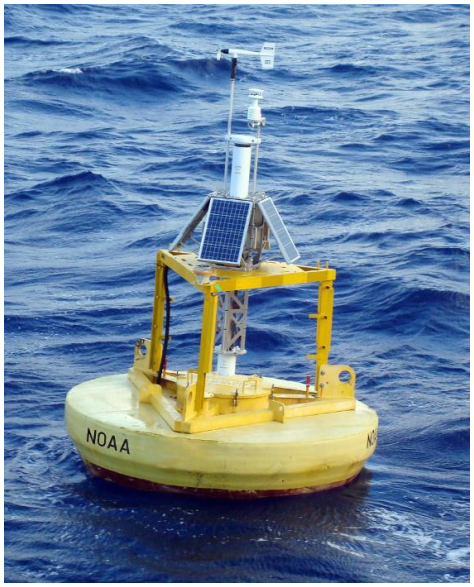

Imagen de la boya de la estación 44025

Los datos se extrajeron del *National Data Buoy Center (NDBC)* de Estados Unidos. En particular se eligió como objeto de estudio la estación 44025 porque se consideró que era la que mejores registros presentaba, a saberse, tenía registros de grandes lapsos; lo que significa una mayor cantidad de datos.

El análisis exploratorio de datos, y con él, la resolución de la pregunta incipiente, comienza con ver la información general del DataSet:

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 34506 entries, 2016-01-01 00:50:00 to 2019-12-31 22:50:00
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   latitude                34506 non-null  float64
 1   longitude               34506 non-null  float64
 2   wind_dir                34479 non-null  float64
 3   wind_spd                34506 non-null  float32
 4   gust                    34506 non-null  float32
 5   wave_height             34488 non-null  float32
 6   dominant_wpd            34458 non-null  timedelta64[ns]
 7   average_wpd             34488 non-null  timedelta64[ns]
 8   mean_wave_dir           34458 non-null  float64
 9   air_pressure            34506 non-null  float32
 10  air_temperature         34505 non-null  float32
 11  sea_surface_temperature  34506 non-null  float32
 12  dewpt_temperature       819 non-null    float32
 13  visibility              0 non-null      float32
 14  water_level             0 non-null      float32
dtypes: float32(9), float64(4), timedelta64[ns](2)
memory usage: 3.0 MB
```

Imagen 1: Información general de los datos.

En principio, el DataSet tenía: 34506 **filas** y 15 **columnas.**

A continuación, se definen las variables (columnas) del DataSet:

- [0] & [1]: **Latitud y Longitud** del sitio donde se encuentra la boya.
- [2]: **Dirección de la que proviene el viento** (en grados; [°]) en el sentido de las agujas del reloj desde el Norte verdadero (Norte Magnético; Polo sur). Medida en un periodo de ocho minutos.
- [3]: **Velocidad del viento** [m/s]. Medida en un periodo de ocho minutos.
- [4]: **Velocidad de la ráfaga** [m/s]. Medida en un periodo de ocho minutos.
- [5]: **Altura de las olas** [m]. El promedio de altura de las olas para un periodo de veinte minutos.
- [6]: **Periodo de ola dominante** [s]. Es el periodo con la energía de ola máxima.
- [7]: **Periodo de ola promedio** [s]. Periodo promedio de todas las olas en un intervalo de veinte minutos.
- [8]: Dirección de donde provienen las olas en el periodo dominante [°], el norte verdadero corresponde a 0° y el Este a 90°.
- [9]: **Presión del aire** [hPa].
- [10]: **Temperatura del aire** [°C].
- [11]: **Temperatura de la superficie del mar** [°C].
- [12]: **Tendencia de la Presión** [hPa]. Da (más o menos) la dirección y la cantidad de cambio de presión durante un periodo de tres horas.
- [13]: **Visibilidad de la estación** [millas náuticas]. El rango para las boyas es 0 a 1,6 millas náuticas.
- [14]: **Nivel del mar** [m].

Obsérvese que casi todos los datos (filas) son del tipo *float*. Lo que significa que: todos los datos son numéricos. Ya que la boya arroja *mediciones.*

En la **imagen 1** se logra divisar que las columnas *'visibility'* y *'water_level'* no tienen datos con valores diferentes de cero, al no ser dinámicos, no aportan información. De manera levemente similar, la columna *'dewpt_temperature' tiene ¡33687 datos nulos!* Entonces se procedió a eliminar esas columnas.

Las otras columnas, aunque en menor medida, también tenían datos nulos. Estos datos nulos fueron cambiados por la media. Bien pudo habérseles asignado el valor de cero, pero se consideró que tal asignación incurriría a un sesgo considerablemente grande.

Para conocer los valores 'normales', se les aplicó estadística a los datos. De esa manera se encontraron las siguientes medidas de tendencia central y de variabilidad:

	latitude	longitude	wind_dir	wind_spd	gust	wave_height	dominant_wpd	average_wpd	mean_wave_dir	air_press
count	34506.000000	34506.000000	34506.000000	34506.000000	34506.000000	34506.000000	34506	34506	34506.000000	34506.000
mean	40.250999	-73.164001	196.409293	6.793682	8.288418	1.311205	0 days 00:00:07.432988620	0 days 00:00:05.129636687	164.275089	1016.716
std	0.000000	0.000000	100.122819	3.536670	4.325649	0.741839	0 days 00:00:02.552718263	0 days 00:00:01.055101823	69.480957	8.077
min	40.250999	-73.164001	1.000000	0.000000	0.000000	0.200000	0 days 00:00:02.470000029	0 days 00:00:02.829999924	1.000000	977.099
25%	40.250999	-73.164001	107.000000	4.200000	5.000000	0.800000	0 days 00:00:05.559999943	0 days 00:00:04.369999886	113.000000	1011.900
50%	40.250999	-73.164001	212.000000	6.300000	7.600000	1.110000	0 days 00:00:07.139999866	0 days 00:00:04.960000038	153.000000	1016.799
75%	40.250999	-73.164001	282.000000	9.100000	11.000000	1.640000	0 days 00:00:09.090000153	0 days 00:00:05.719999790	198.000000	1021.799
max	40.250999	-73.164001	360.000000	23.200001	27.900000	7.010000	0 days 00:00:17.389999390	0 days 00:00:10.939999580	360.000000	1043.800

Imagen 2: Medidas de tendencia central y de variabilidad.

Aunque las medidas de tendencia central y de variabilidad dan una leve visión del comportamiento de los datos, no es suficiente. Por ello se hizo: análisis univariado; bivariado y multivariado. En la sección de *Análisis y Resultados* se discuten los resultados de dichos análisis.

- **Análisis univariado:**

  De las características menos técnicas que describen las olas, está su altura. Por eso se procedió a realizar un histograma que relacionaba la altura de las olas y el tiempo de medida. Esto para un intervalo de 4 años: se tomaron datos desde enero del 2016, hasta datos de enero de enero del 2020. Nótese que el tiempo no es una variable, sino un parámetro. Es por ello que comparar la altura de las olas en función del tiempo cabe dentro de un análisis univariado.

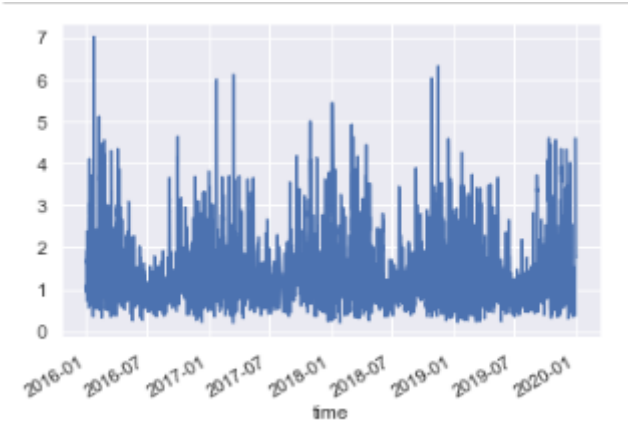

Imagen 3: Histograma de la altura de las olas.

Aunque el histograma ayudó a tener una mejor idea sobre la altura de las olas, se consideró que un diagrama de boxplot también sería útil.

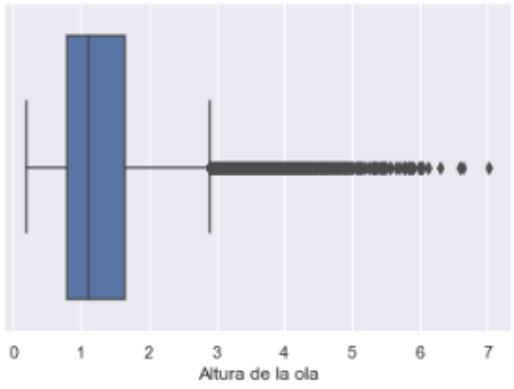

Imagen 4: Boxplot altura promedio de las olas.

También se hizo una boxplot para la temperatura del aire.

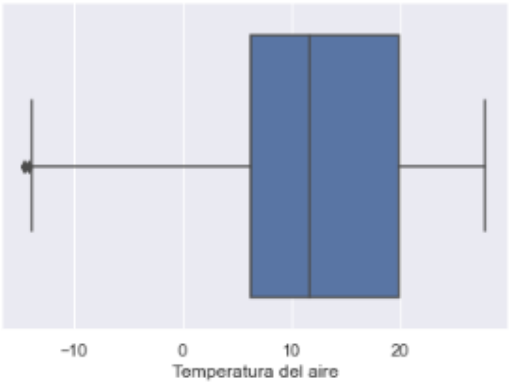

Imagen 5: Boxplot de la Temperatura del aire.

Además, se hizo un gráfico de *Rosa de los Vientos,* aprovechando que en metrología se usan con la finalidad de estudiar la dirección del viento:

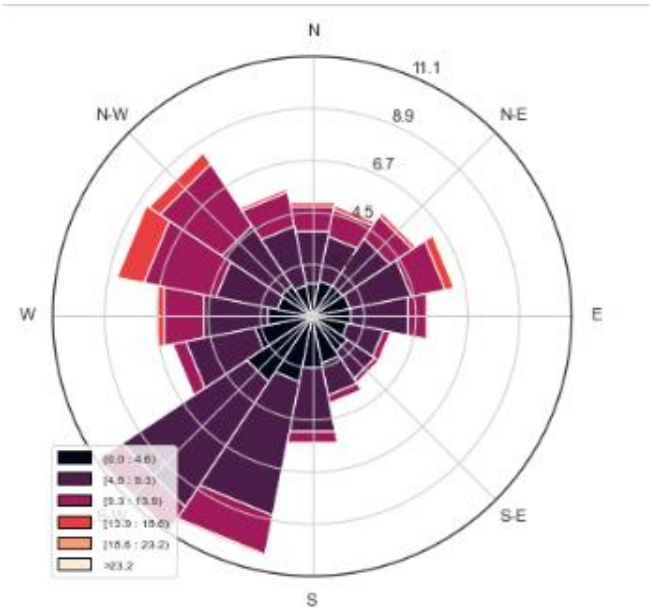

Imagen 6: Gráfico Rosa de los Vientos.

- **Análisis bivariado:**

  Se hizo dos diagramas de dispersión, con la intención de verificar si existe alguna correlación entre dos variables. Esto es, ver si existe alguna tendencia lineal.

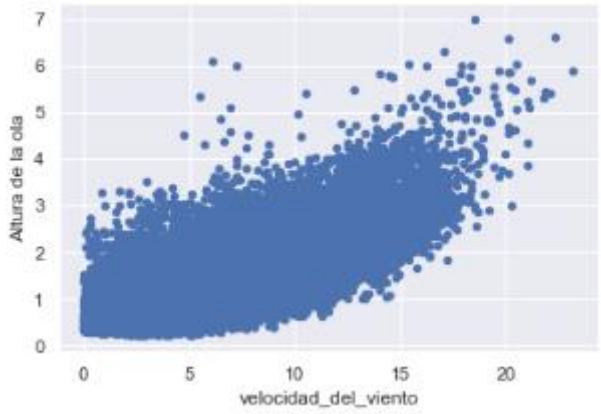

Imagen 7: Diagrama de dispersión de Altura de la ola y la Velocidad del Viento.

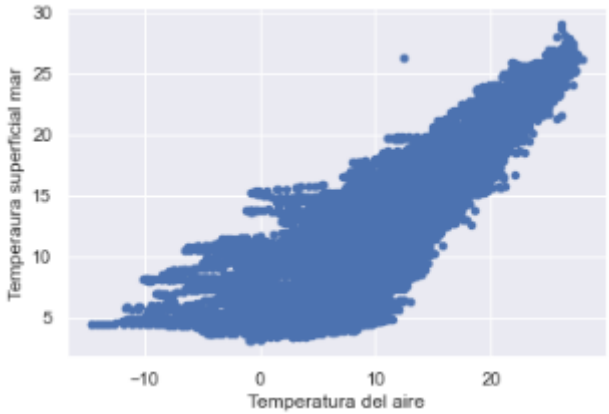

Imagen 8: Diagrama de dispersión de las Temperaturas de la superficie del mar y el aire.

- **Análisis multivariado:** en este se compara todos los posibles pares de variables con la finalidad de encontrar alguna correlación. Los números de cada cuadro equivale al índice de correlación de cada par de variables.

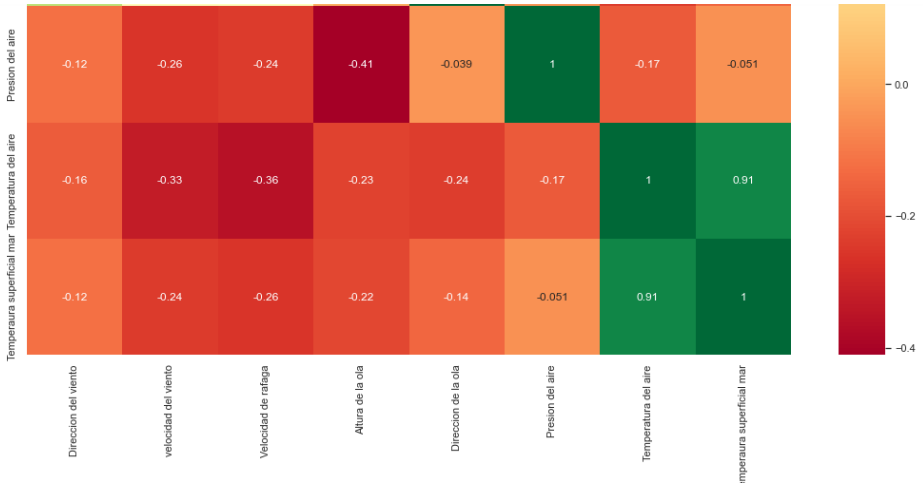

Imagen 9: Se ve el fragmento de la **Matriz de correlaciones** para los datos.

ANÁLISIS Y RESULTADOS

Antes de hacer el análisis con los datos, se hicieron hipótesis. Después, se aplicaron pruebas tipo A/B, para comprobar la veracidad de dichas hipótesis. Las hipótesis y los resultados se muestran a continuación:

1. Las olas que tienen altura mayor a 1.31 m (1.31 es la altura media) son producidas por vientos mayores que las olas con altura menor a 1.31 m. **Hipótesis aprobada.**
2. Las olas dominantes con periodo menor a la media, se producen por vientos mayores que las que tienen un periodo mayor. **Hipótesis aprobada.**
3. Cuando la presión es mayor a la media, la velocidad del viento es mayor que la velocidad del viento a una presión menor. **Hipótesis desmentida.**
4. La temperatura del aire no afecta el periodo promedio de las olas. **Hipótesis aprobada.**
5. La temperatura de la superficie del agua, afecta de manera inversa al periodo de la ola promedio. **Hipótesis desmentida.**
6. Las olas dominantes con mayor periodo (dominante) tienen menor altura. **Hipótesis aprobada.**

Histograma de altura de olas con respecto al tiempo: Si se toma un fragmento de un año en el histograma, se puede ver que en principio se presenta un comportamiento Gaussiano, cuyo máximo aparece (en cada año) a inicios de año. Mientras que los mínimos ocurren a mediados de año. Este comportamiento se logró comprobar con los datos meteorológicos de Nueva York. Según la hipótesis 1, a medida que los vientos tienen velocidades mayores, se presentan olas más grandes. Y según los datos meteorológicos de NY, las épocas de vientos menos veloces son a mediados de año. Entonces, tanto la velocidad de los vientos (como la alturade las olas), es un comportamiento periódico, en donde se espera que, a inicios de año en NY, haya olas más grandes, como vientos más rápidos.

Boxplot de altura de las olas: este gráfico muestra que el rango de la altura de las olas es (0-3) m (exclusive). Es decir, qué tamaño se consideraría como el tamaño 'normal' de una ola. Además,

muestra que la mediana corresponde a un valor levemente mayor a 1 metro. Pero se evidencia que NO coincide con el valor medio, ya que este es aproximadamente 1.3 metros. Por ese lado se puede pensar que: si bien es cierto que el comportamiento de la altura de las olas no es gaussiano, no se aleja mucho de él. Y presenta un leve sesgo. Lo que, si es cierto, es que los datos presentan muchos valores atípicos y extremadamente atípicos, como la medida de 7 metros.

Boxplot de la temperatura del aire: en este gráfico se muestra que el rango de temperatura del aire es aproximadamente (-13, 27) °C. Donde la mitad de los datos toman valores de –13 a 11 °C, y la otra mitad de 11-27 °C. Hay más dispersión para las temperaturas menores a 11°C. Además, se logra apreciar que no tiene tantos datos atípicos. Quizá la causa de esto sea que, medir temperaturas no sea tan complicado como medir la altura de las olas.

Gráfico de Rosa de los Vientos: esta gráfica muestra cuál es la dirección de la que proviene el viento, como sus intensidades. Esto para los 4 años que se estudiaron. Así la dirección de la que provino la mayoría de vientos en el periodo 2016-2020 en el lugar donde está la boya, es el sur-oeste. Y la dirección de menos procedencia fue el sur-este. Sin embargo, la dirección cuyos vientos poseían velocidades mayores era el noreste. Pareciera que en la latitud norte (norte magnético terrestre) del planeta, se crean vientos más veloces. Mientras que, en la latitud sur, se produce mayor cantidad de viento, pero con velocidades más pequeñas. En adición: parta (imaginariamente) el circulo con rectas, tal que pasen por la mitad de todos los triángulos (de modo que al final quede una estructura como los radios de la rueda de una bicicleta), pareciera que, si un triángulo es grande, el triángulo opuesto por el vértice a él es pequeño. Es decir, pareciera que, si en una dirección el viento fluye más continuamente, se espera que la dirección que 'al frente' ocurriese todo lo contrario, que no fluyese viento.

Diagrama de dispersión de Altura de la ola y la Velocidad del Viento: se ve que se tiene una correspondencia casi lineal entre las dos variables. Aunque se presenta mucha dispersión, esto podría deberse a la cuantiosa cantidad de datos atípicos que se presentaban para los valores de las alturas de las olas.

Diagrama de dispersión de las Temperaturas de la superficie del mar y el aire: este diagrama tiene un mejor comportamiento que el anterior, también se presenta una fuerte correlación. Y para este caso, no hay tanta dispersión.

Matriz de correlaciones: a continuación, se va a enlistar las variables que poseen una correlación mayor a 0.5:

- **Velocidad del viento y velocidad de la ráfaga [correlación de 0.99]:** como las ráfagas son vientos muy veloces, pero de poca duración, resulta casi evidente que entre más rápido sean los vientos, mayores ráfagas hay. De manera inversa, si no hay viento (velocidad cero del movimiento del aire), no van a haber ráfagas.

- **Temperatura superficial del mar y temperatura del aire [0.91]:** es de esperarse que la temperatura superficial del mar coincida con la del aire, puesto que las agitadas moléculas del aire, le imparte energía a las moléculas que permanecen en la superficie del agua. No se espera un índice de correlación de 1, puesto que el mar es un reservorio térmico. Entonces, la superficie del mar (siendo la parte más externa de la capa superficial) actúa como una interfaz en donde la temperatura del aire, y la acción del viento intentan 'contaminar' de calor al mar, cuyos esfuerzos se van desvaneciendo con la profundidad.

- **Velocidad de la ráfaga y altura de la ola [0.7]:** la forma en que se crea una ola es la siguiente: Cuando el viento se mueve en la superficie del agua, hace fricción sobre ella. Haciendo que la estructura lisa del agua se convierta en un campo de pequeñas rizaduras u ondulitas llamadas *ondas capilares*. Estas son del orden de los milímetros. Pero con la creación de las ondas capilares, se intensifica la fricción, haciendo que se creen *ondas de gravedad* que van desde los pocos centímetros hasta las decenas de los metros. Por otro lado, las ráfagas son vientos de menos de 20 segundos, pero con velocidades grandísimas. Además, se sabe que, si un cuerpo se mueve dentro de un fluido, la fricción existente entre la interfaz fluido-cuerpo aumenta, conforme crece la velocidad relativa del cuerpo al fluido. Es así como las moléculas de la superficie del agua experimentan una fricción mayor con las ráfagas, es decir, con vientos de velocidades mayores.

- **Velocidad del viento y altura de la ola [0.69]:** la explicación a esta correlación es idéntica que la correlación anterior, salvo que es menor puesto que las velocidades promedio de los vientos son menores que las velocidades de la ráfaga.

- **Dirección del viento y dirección de la ola [0.51]:** se podría pensar que, lo más razonable es que la dirección del viento coincida con la dirección de las olas. Pero eso solo funciona en los casos ideales: puede ocurrir que una ola sea creada por un cierto viento, pero que, en su viaje, sea desviada de la dirección original por otro (u otros vientos), dando razón al por qué de la presente correlación.

CONCLUSIONES

- Se encontró que el parámetro que más incide en las olas es la velocidad del viento.
- Mediante el análisis se pudo concluir que, la velocidad de los vientos (y con ellos, la altura de las olas) al sur de Islip, son mayores en épocas de inicios de año; y son menores a mitad de año.
- No necesariamente, la dirección del viento coincide con la dirección de las olas.
- Aunque no era algo que se estaba buscando, se halló que la temperatura del aire está fuertemente relacionada con la temperatura de la superficie del mar.
- La presión del aire, no tiene relación con la velocidad del viento.
- Los vientos en la región estudiada, en el intérvalo 2016-2020, provinieron mayormente de la dirección suroeste.
- La medición de la altura de las olas es una tarea demandante.

REFERENCIAS

[1] Introduction to Python-Part 2. Sage Lichtenwalner. Ocean Data Lab

[2] ¿Cómo hacer al análisis exploratorio de datos? Guía paso a paso. Miguel Sotaquirá. Junio 11 2021. Codificando Bits.

[3] Datos climáticos y meteorológicos históricos simulados para Nueva York. Meteoblue.

[4] Lectura de Diagramas de Caja. KhanAcademyEspañol

[5] Fuerzas de rozamiento. Departamento de Física Aplicada III.

[6] Ráfaga (viento). (2022, 1 de agosto). Wikipedia, La enciclopedia libre. Fecha de consulta: 04:50, octubre 23, 2022