## Línea horizontal



Trabajo Final- Data Analytics

20.07.2023

**─**

Manuel López Vargas

Curso Data Analytics - Comisión 42265 - CoderHouse  
La Plata, Buenos Aires, Argentina

# Índice

1. Introducción……………………………………………………………………………………………………….. **2**
2. Alcance……………………………………………………………………………………………………………….. **3**
3. Hipótesis…………………………………………………………………………………………………………….. **3**
4. Dataset……………………………………………………………………………………………………………….. **3**
5. Diagrama Entidad-Relación………………………………………………………………………………… **5**
6. Listado de Tablas………………………………………………………………………………………………… **6**
7. Listado de columnas por tablas………………………………………………………………………….. **6**
8. Transformaciones y Segmentaciones………………...………………………………………………. **8**
9. Modelo relacional de Power-BI…………………………………………………………………………… **9**
10. Medidas calculadas…………………………………………………………………………………………….. **10**
11. Visualización de los datos…………………………………………………………………………………… **13**
    1. Página de “Promedios por estación y horas”.................................................. **13**
    2. Página de “Promedios por mes y máximos”................................................... **14**
    3. Página de “Correlación índice UV-Humedad relativa”.................................... **15**
    4. Página de “Correlación índice UV- Temperatura”........................................... **17**
    5. Página de “Correlación índice UV- Presión Atmosférica”............................... **17**
    6. Página de “Correlación índice UV- Velocidad del viento”............................... **18**
    7. Página de “Matriz de confusión”........................................................................ **18**
    8. Página de “Mapa y TMIs”..................................................................................... **19**
12. Conclusión…………………………………………………………………………………………………………… **20**
13. Bibliografía…………………………………………………………………………………………………………... **21**

# Introducción

En el siguiente trabajo se extrajeron los datos meteorológicos provenientes de las estaciones automáticas Davis Vantage Pro 2 calibradas, en la Ciudad Autónoma de Buenos Aires (CABA). Junto con esto tenemos datos de la ubicación geográfica de las Torres de Monitoreo Inteligente (TMI) equipadas con las estaciones meteorológicas.

La estación guarda los datos de Radiación UV según el índice de la siguiente tabla:

| **Valor Índice** | **Categoría de exposición** |
| --- | --- |
| **0-2** | **Mínimo** |
| **3-4** | **Bajo** |
| **5-6** | **Moderado** |
| **7-9** | **Alto** |
| **10+** | **Extremo** |

**Tabla 1**

Este índice es el adoptado por la Organización Mundial de Meteorología.

(<http://www.wmo.int/pages/index_es.html>)

Con los datos obtenidos, utilizando SQL y Power-BI para realizar un tablero, analizaremos los mismos para ver si existen tendencias, en qué épocas del año la radiación UV es más alta y en qué periodos diurnos esta es más nociva.

# Alcance

Este trabajo se encuentra orientado para ser utilizado por un nivel táctico como el Ministerio de Salud de CABA. Esto, junto con la tabla extraída de la O.M.M. podrán saber en qué épocas del año y en qué periodos del día la radiación UV es nociva para el Humano, así como también de que variables éste depende. Con eso se podrán tomar medidas de concientización a la población de CABA para poder reducir las tasas de cáncer de piel por exposición al Sol.

# Hipótesis

Ver si la radiación UV medida tiene alguna correlación con alguno de los otros datos, p.j.

velocidad del viento, presión atmosférica, etc. Por otro lado también, ver si esta cambia en

los distintos puntos de la ciudad, ver el gradiente de la radiación y si hay alguna tendencia.

Por último, también calcular el valor medio de la radiación en CABA en las distintas

estaciones y cómo cambia a lo largo del día así como también las mediciones máximas y

en qué horarios se dan generalmente, y con esto ver si presenta algún riesgo para la

población.

# Dataset

En el presente trabajo utilizamos 3 tablas extraídas de la página de la ciudad de Buenos Aires, y estas constan en:

* **Tabla de TMIs:**

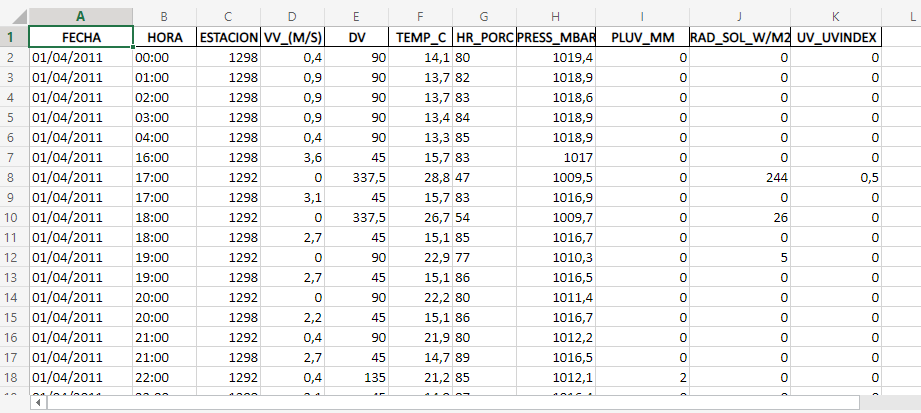
Esta contiene identificadas todas las torres de monitoreo inteligente junto con su dirección y localización geográfica.



A esta tabla, y como veremos más adelante, a las demás también, le hicimos una limpieza de datos utilizando SQL. (Ver sección 8)

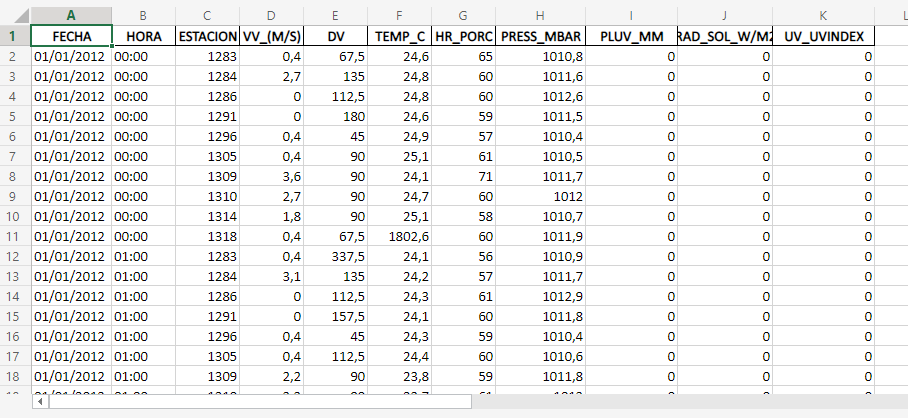
* **Tabla de datos meteorológicos 2011:**

Esta tabla contiene todas las mediciones de los distintos parámetros, así como también la estación o TMI y la fecha donde se realizó esa medida, durante el año 2011.



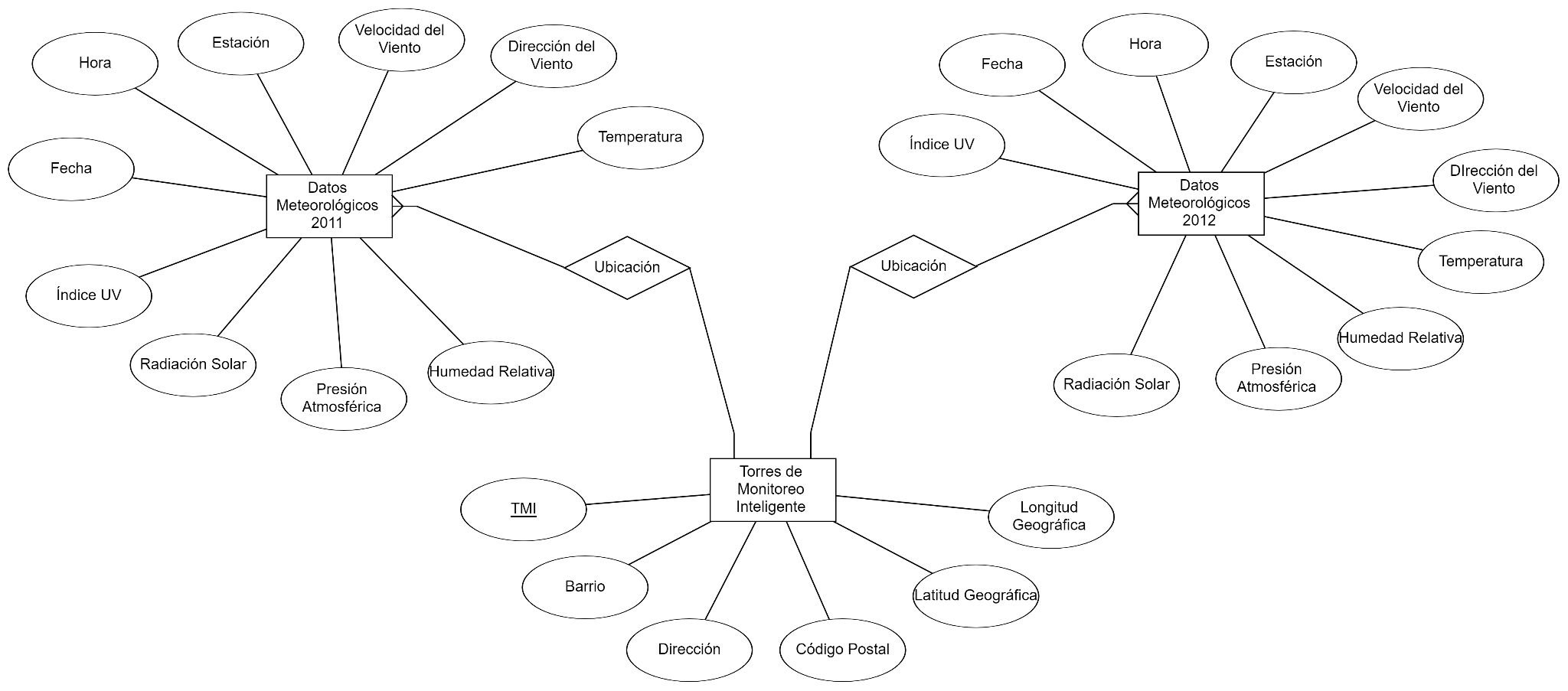
* **Tabla de datos meteorológicos 2012:**

Esta tabla contiene todas las mediciones de los distintos parámetros, así como también la estación o TMI y la fecha donde se realizó esa medida, durante el año 2012.



# Diagrama Entidad-Relación

A continuación se muestra el diagrama entidad-relación creado para este proyecto:



# Listado de tablas

Aquí mencionaremos las tablas utilizadas junto con una breve descripción de las

mismas, definiendo las Primary Key (PK) y las Foreign Key (FK).

* **Datos Meteorológicos 2011:**

Contiene todas las mediciones meteorológicas de todas las torres de monitoreo inteligente (TMI), discriminado por día y horario, tomando medidas cada 1hs durante el año 2011.

* FK: Estaciones
* **Datos Meteorológicos 2012:**

Contiene todas las mediciones meteorológicas de todas las torres de monitoreo inteligente (TMI), discriminado por día y horario, tomando medidas cada 1hs durante el año 2012.

* FK: Estaciones
* **Torres de Monitoreo Inteligente:**

Contiene la dirección en CABA de cada una de las TMI y también la ubicación geográfica.

* PK: TMI

# Listado de columnas por tablas

A continuación, se mencionan las columnas de cada tabla como así también su tipo de campo y clave.

* **Torres de monitoreo inteligente**

|  |  |  |
| --- | --- | --- |
| **Campo** | **Tipo de Campo** | **Tipo de Clave** |
| **TMI** | **Smallint** | **PK** |
| **Barrio** | **nvarchar** | **-** |
| **Dirección** | **nvarchar** | **-** |
| **Código Postal** | **Smallint** | **-** |
| **Latitud** | **Decimal** | **-** |
| **Longitud** | **Decimal** | **-** |

* **Datos Meteorológicos 2011**

|  |  |  |
| --- | --- | --- |
| **Campo** | **Tipo de Campo** | **Tipo de Clave** |
| **Fecha** | **Date** | **-** |
| **Hora** | **Time** | **-** |
| **Estación** | **Smallint** | **FK** |
| **Velocidad del Viento [m/s]** | **Decimal** | **-** |
| **Dirección del Viento** | **Decimal** | **-** |
| **Temperatura [°C]** | **Decimal** | **-** |
| **Humedad Relativa [%]** | **Tinyint** | **-** |
| **Presión Atmosférica [mbar]** | **Decimal** | **-** |
| **Radiación Solar [W/m2]** | **Samllint** | **-** |
| **Índice UV** | **Decimal** | **-** |

* **Datos meteorológicos 2012**

| **Campo** | **Tipo de Campo** | **Tipo de Clave** |
| --- | --- | --- |
| **Fecha** | **Date** | **-** |
| **Hora** | **Time** | **-** |
| **Estación** | **Smallint** | **FK** |
| **Velocidad del Viento [m/s]** | **Decimal** | **-** |
| **Dirección del Viento** | **Decimal** | **-** |
| **Temperatura [°C]** | **Decimal** | **-** |
| **Humedad Relativa [%]** | **Tinyint** | **-** |
| **Presión Atmosférica [mbar]** | **Decimal** | **-** |
| **Radiación Solar [W/m2]** | **Samllint** | **-** |
| **Índice UV** | **Decimal** | **-** |

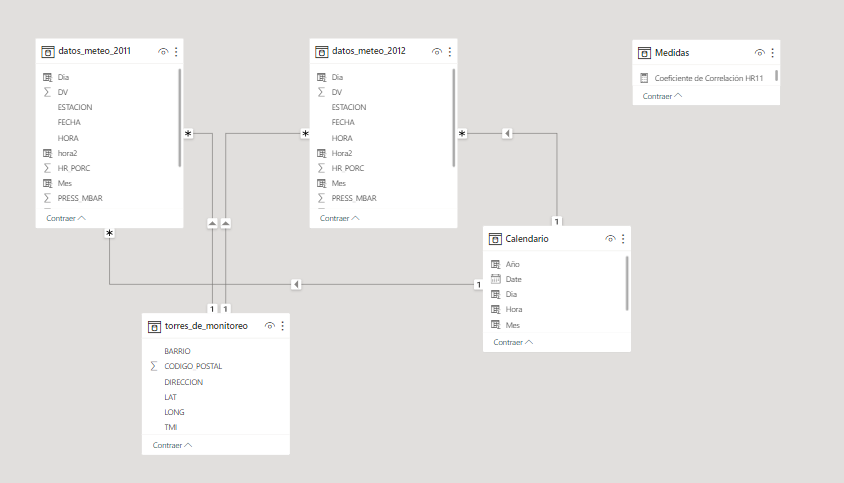
# Transformaciones y segmentaciones

Como mencionamos anteriormente, tuvimos que hacer una limpieza a nuestras tablas. Lo primero que hicimos fue eliminar algunos campos que eran redundantes o no nos interesaban. En la tabla de TMIs eliminamos los campos “DIRECCIÓN” , “CALLE” , “ALTURA” , “BARRIO 1” y “ARG\_CODIGO\_POSTAL”. De las tablas de datos meteorológicos 2011 y 2012 eliminamos el campo “PLUV\_MM”.

Luego, para poder tener nuestras tablas acordes para trabajar en Power BI y realizar medidas y cálculos y para también poder relacionar nuestras tablas hicimos algunas transformaciones a las mismas. Primero, como había medidas de algunas torres de monitoreo inteligente (TMI) que no estaban en nuestra tabla de estaciones meteorológicas, eliminamos todos los registros de las correspondientes TMIs de las tablas de mediciones, datos meteorológicos 2011 y 2012. Luego, elegimos medidas máximas y mínimas “coherentes” para los distintos datos medidos y rellenamos con un “0” dichas medidas. Y por último, tuvimos que modificar los campos latitud y longitud de la tabla de estaciones meteorológicas para que sean dimensiones y no medidas, así como también el tipo de dato cambiandolo a decimal (18,3).

# Modelo relacional de Power-BI

Una vez cargados los datos a Power-BI creamos una tabla de calendario que nos va a ser útil para calcular medidas y realizar gráficos y una tabla de medidas donde se encuentran todas las medidas calculadas. Con esto, obtuvimos lo siguiente



# 

# Medidas calculadas

Las medidas calculadas fueron las siguientes

**Promedios por estación:** Realizamos una medida que calcula el promedio del índice UV en cada estación (en las que había datos) para ambos años. Esto lo hicimos de la siguiente manera:

Promedio estación 201\* = CALCULATE(AVERAGE(datos\_meteo\_201\*[UV\_UVINDEX]), datos\_meteo\_201\*[Mes] >= a, datos\_meteo\_201\*[Mes] <= b, datos\_meteo\_201\*[hora2] > 10, datos\_meteo\_201\*[hora2] < 16, datos\_meteo\_201\*[UV\_UVINDEX] > 0)

Filtramos por meses y por horas, donde a es el mes inicial y b el mes final. Por ejemplo, para calcular el promedio en verano a=1, b=3 (haciendo una estimación a primer orden de las estaciones). Luego, el filtro en horas lo hicimos teniendo en cuenta que en estas horas es cuando la radiación del Sol es más significativa. Por último, aplicamos el filtro para el índice UV mayor a cero para desestimar los outliers que antes transformamos en ceros, y con esto también no tomamos en cuenta las medidas con lluvia o nubosidad alta. (Los “ \* ” indican que pueden tomar los valores 1 o 2, correspondientes a los dos años de las mediciones).

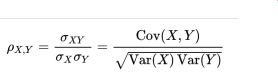
**Promedio por hora:** Realizamos una medida que calcula el promedio del índice UV por hora en verano e invierno, que es cuando el índice UV es más elevado. Esto lo hicimos de la siguiente manera:

Rad Xhs estación = CALCULATE(AVERAGE(datos\_meteo\_201\*[UV\_UVINDEX]), datos\_meteo\_201\*[Mes] > 10 , datos\_meteo\_201\*[hora2] = X ,

datos\_meteo\_201\*[UV\_UVINDEX] > 0 )

Al igual que antes, filtramos por estación y por índice UV mayor a cero, y también filtramos la hora en la que queremos la medida, tomando de 7hs a 20hs. Luego, con estas medidas, mostramos cómo varía el índice UV en este periodo de horas del día en verano y primavera.

**Coeficiente de correlación de Pearson**: Realizamos esta medida para saber qué correlación tienen las diferentes variables medidas con el índice UV. Las variables utilizadas fueron: Presión, Humedad Relativa, Velocidad del Viento y Temperatura, para ambos años. Para esto tuvimos que aplicar la definición del coeficiente de correlación, que está dado por



Lo primero que hicimos fue calcular la covarianza entre el índice UV y las demás variables, utilizando su definición, de la siguiente manera.



Covarianza(variable)\* =

VAR \_CONTEO =COUNTROWS(datos\_meteo\_201\*)

VAR \_Avg1 = AVERAGEX(datos\_meteo\_201\*, datos\_meteo\_2011[UV\_UVINDEX]) VAR \_Avg2 = AVERAGEX(datos\_meteo\_201\*, datos\_meteo\_2011[variable]) VAR \_TablaIntermedia =

ADDCOLUMNS(

datos\_meteo\_201\*,

"Covarianza",

DIVIDE(

(datos\_meteo\_201\*[UV\_UVINDEX]-\_Avg1)\*

(datos\_meteo\_201\*[variable]-\_Avg2),

\_CONTEO))

RETURN

SUMX(\_TablaIntermedia,[Covarianza])

Por último, teniendo la covarianza, calculamos el coeficiente de correlación de Pearson como sigue

Coeficiente de Correlación variable 1\* =

DIVIDE(

[Covarianza(variable)1\*],

STDEV.P(datos\_meteo\_201\*[variable])\*

STDEV.P(datos\_meteo\_201\*[UV\_UVINDEX]),

0)

Este coeficiente es una medida de la dependencia lineal de dos variables aleatorias cuantitativas. Toma valores entre -1 y 1 en donde valores cercanos a -1 indica una fuerte correlación negativa (dependencia lineal con la inversa), valores cercanos a 1 indica una fuerte correlación positiva y valores cercanos a 0 indica que no hay una correlación entre las variables.

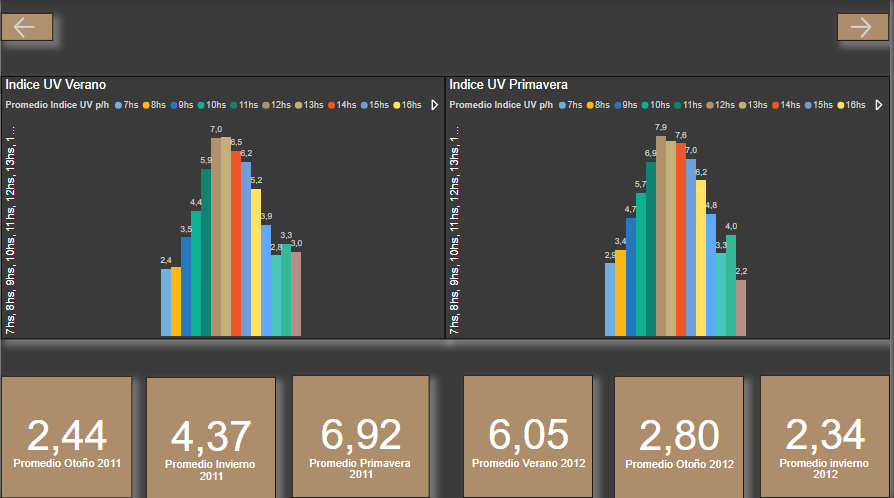
# Visualización de datos

A continuación presentaremos un resumen de cada una de las páginas del tablero creado en Power-BI.

* 1. **Página de “Promedios por estación y horas”**

En esta solapa lo que mostramos en la parte inferior en tarjetas son los promedios de Índice UV (IUV) por estación, utilizando las medidas antes calculadas. Como mencionamos anteriormente, esto lo hicimos tomando los datos a partir de las 10hs hasta las 16hs y descartando los datos que son nulos.

En la parte superior realizamos dos gráficos de barras para mostrar cómo varía el IUV a lo largo del día durante las dos estaciones con mayor promedio de IUV.

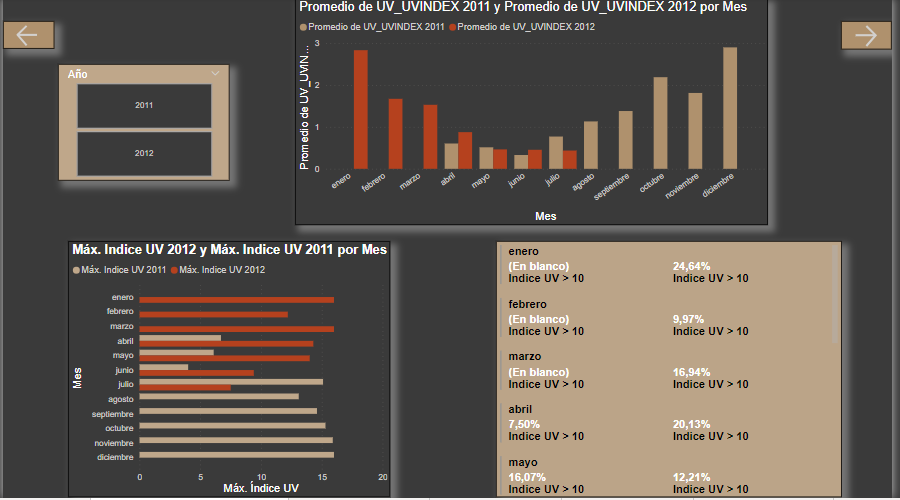


* 1. **Página de “Promedios por mes y máximos”**

Aquí lo que mostramos en el gráfico superior es el promedio de IUV por mes en ambos años, 2011 y 2012.

En el gráfico inferior izquierdo se muestra el valor máximo medido del IUV por cada mes de ambos años. Y en el gráfico inferior derecho se muestra el porcentaje de medidas que son superiores a 10 ( que corresponde a una categoría de exposición “extrema”, según la Tabla 1) por mes, en forma de tarjeta de varias filas. Las medidas “ En Blanco” corresponden a los meses de cada año en los que no hay datos.

Por último, en esta solapa realizamos una segmentación por año, donde se puede escoger utilizando los botones del segmentador si ver los datos del año 2011, 2012 o ambos.



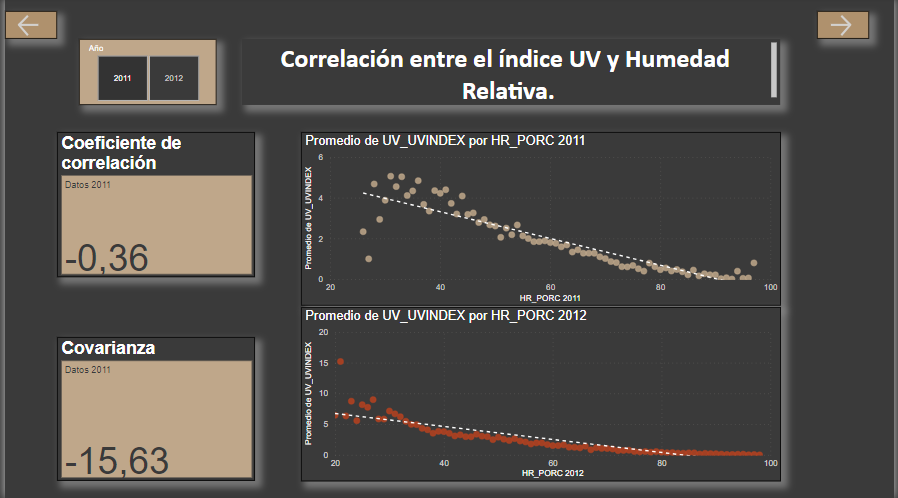
* 1. **Página de “Correlación índice UV-Humedad relativa”**

En la presente página del tablero estudiamos la dependencia estadística entre el IUV y la humedad relativa. Para esto primero calculamos la covarianza entre estas dos variables y con esto el coeficiente de correlación, que nos da una medida de la dependencia lineal entre dos variables. Esto se muestra en las dos tarjetas a la izquierda.

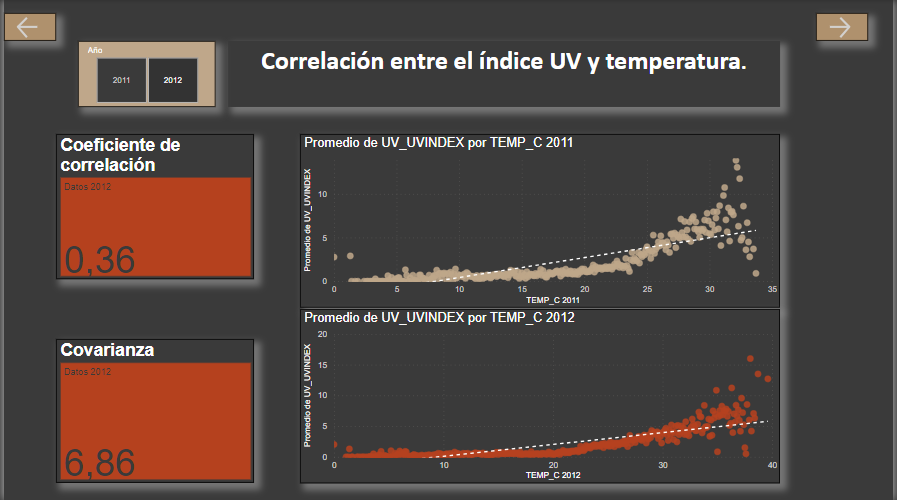
Luego, también hicimos un gráfico de dispersión donde ploteamos humedad relativa vs promedio de IUV, donde incluimos una línea de tendencia en blanco a trazos.

Aquí también realizamos una segmentación por año, en donde solo se ven afectadas las tarjetas de covarianza y coeficiente de correlación, tal que se pueda elegir ver los datos del 2011 o 2012.

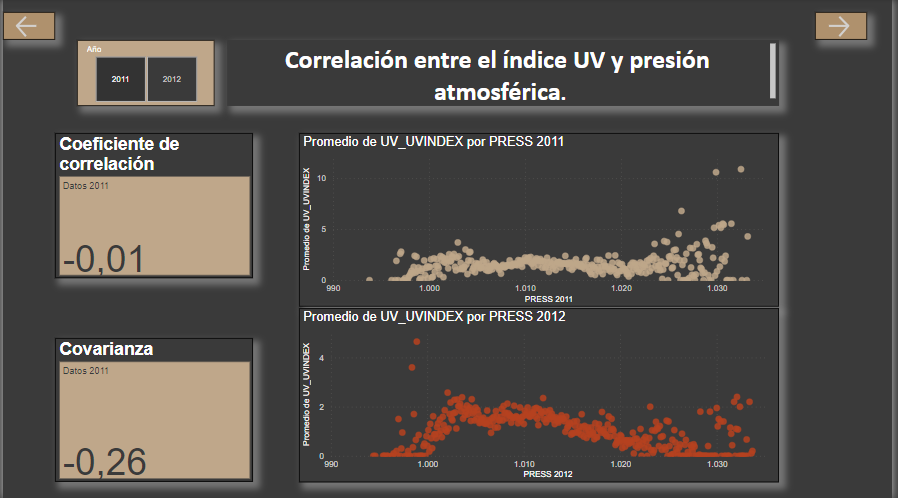
En las siguientes subsecciones 11.4 11.5 11.6 se hará el mismo tratamiento pero con diferentes variables, que aclararemos en cada subsección.



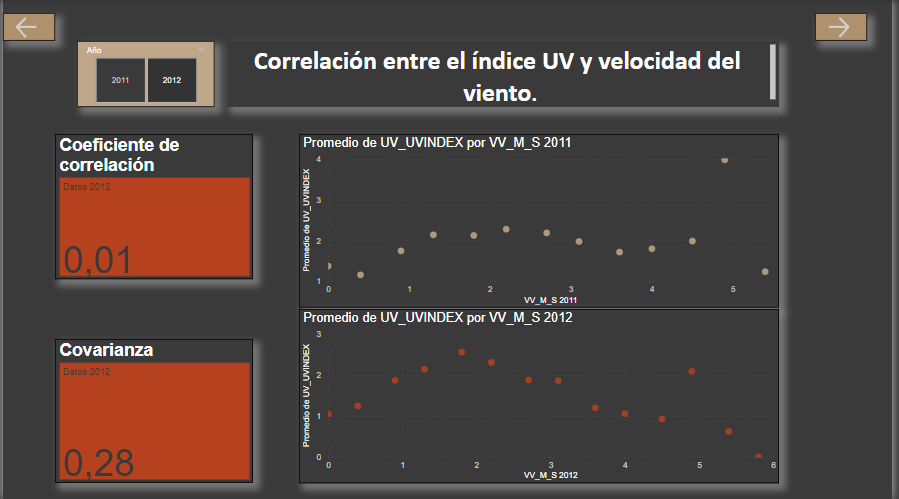
* 1. **Página de “Correlación índice UV-Temperatura”**



* 1. **Página de “Correlación índice UV-Presión atmosférica”**

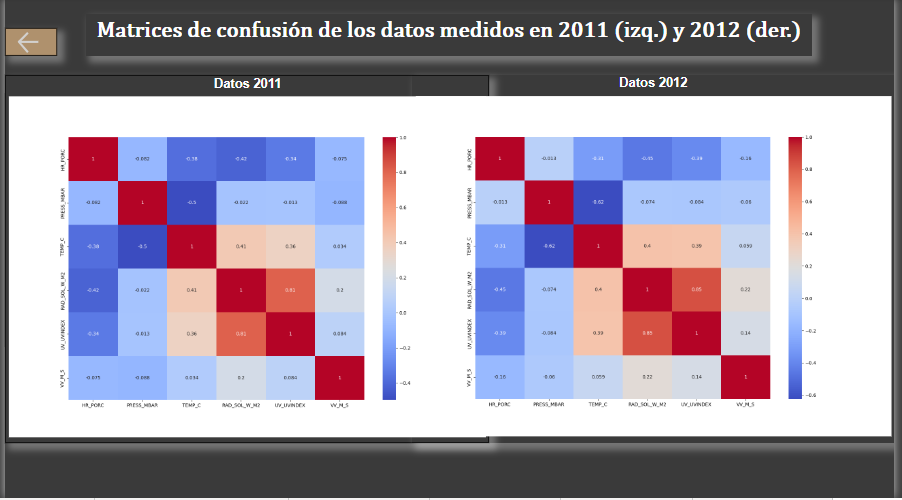


* 1. **Página de “Correlación índice UV-Velocidad del viento”**

****

* 1. **Página de “Matriz de confusión”**

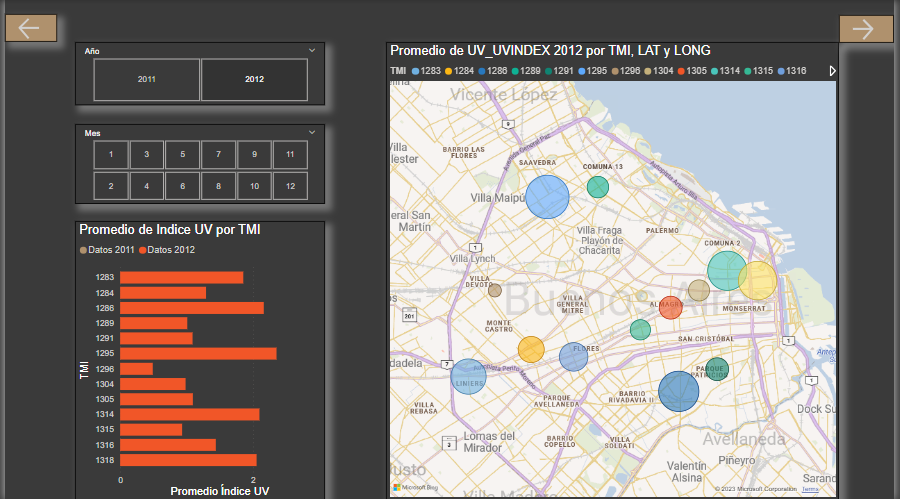
En esta solapa simplemente lo que mostramos fue un resumen de las anteriores realizando una matriz de confusión, que muestra la correlación entre las variables, utilizando todas las variables, para ambos años.

****

* 1. **Página de “Mapa y TMIs”**

Por último, en esta página se muestra el promedio de IUV por TMIs y a la derecha un mapa que muestra la localización de cada TMI y el tamaño de la burbuja indica el promedio de IUV.

A esta le realizamos dos segmentaciones, una por año ( la cual no afecta al mapa) y otra por mes, con las cuales se puede elegir qué año visualizar y que mes utilizando los botones del segmentador. Los meses en los cuales el mapa y el gráfico de barras quedan en blancobson aquellos en los que no hay datos disponibles.



# Conclusiones

Se observó que las estaciones con medidas de IUV más altas en promedio son el verano y primavera, así como también que las medidas más altas en promedio ocurren cerca de las 12hs-13hs del mediodía, como era de esperarse.

Por otro lado se vió que en varios meses de los que hay medidas, la medida máxima alcanza un IUV > 15, la cual es extremadamente alta, así como también que en la mayoría de los meses el IUV alcanza medidas > 10, es decir, que casi todos los meses del año los IUV alcanzan medidas extremas. Además, vimos que en los meses de diciembre y enero más del 24% de las medidas superan un IUV de 10 y la mayoría de los meses dichas medidas alcanzan o superan el 10% del total.

Luego, en cuanto a las correlaciones entre el IUV y otros observables, vimos que ésta con la humedad relativa tienen un coeficiente de correlación de apox. **-0.36 – -0.39**, es decir que hay una correlación negativa no despreciable entre dichas variables. Por otro lado, el coeficiente de correlación entre IUV y la temperatura es de **0.30 – 0.36**, lo que muestra una correlación positiva no despreciable. Esta última no es concluyente ya que no es claro si existe dicha correlación realmente o es indirecta/casual debido a que las medidas más altas son en las estaciones más calurosas y sea realmente causa de un fenómeno geométrico- físico-astronómico. Con respecto a la correlación con la presión atmosférica- IUV y la velocidad del viento-IUV, se ve que no hay una correlación entre las variables.

Por último, se puede ver que los promedios de IUVs en las distintas TMIs cambian significativamente. Esto puede deberse a distintos factores climáticos, humanos o estructurales/arquitectónicos donde están ubicadas las distintas TMIs. Los datos disponibles no son suficientes para dar una conclusión acerca de si realmente el IUV promedio cambia a lo largo de la ciudad o no.

# Bibliografía

(<http://www.wmo.int/pages/index_es.html>)

<https://data.buenosaires.gob.ar/dataset/informacion-meteorologica/resource/juqdkmgo-1224-resource>