PROYECTO FINAL: CREACIÓN DE UN DASHBOARD

Seguimiento del Proyecto

Descripción breve

Se realiza el seguimiento de los pasos realizados durante el desarrollo del trabajo final requerido por el curso de Data Analytics de CoderHouse.

Victoria Gardella, Eduardo Gonik, Lautaro Manuel Torres

Tabla de contenido

Introducción

Se dispone a realizar el análisis del dataset 'Hazardous Air Pollutants' de la EPA que presenta información comprehensiva sobre la presencia de contaminantes en el aire en los Estados Unidos. Se informan más de 2 millones de medidas de aire geolocalizadas de forma precisa, y complementadas con información sobre poblaciones cercanas, instrumentación analítica y parámetros estadísticos específicos de la medida.
Con la idea de presentar un panorama claro sobre la situación que podría ser empleado por agencias gubernamentales para implementar planes de saneamiento ambiental, decidimos que nuestro análisis se debería enfocar en poder discriminar cuales son los contaminantes que aquejan a las distintas áreas geográficas del país, estado a estado.
El análisis se dio de forma desagregada, con las áreas discriminadas de análisis resultando:

- Introducción
- Descripción general del proyecto
- Análisis de la distribución geográfica de los contaminantes
- Comparación con los parámetros regulatorios establecidos por EPA.
- Descripción cuantitativa de las medidas.
- Descripción de las técnicas y equipamientos analíticos empleados

Atributos:

Se realiza un listado de los atributos inicialmente incluidos en la base de datos. Se adjunta una pequeña expicación y el nombre utilizado en la base de datos para cada uno de los atributos.

1. Código de la Medida (sample_id): número único identificatorio de cada medida realizada.
2. Codigo de Estado (state_id): El código FIPS (correspondiente a las siglas de The Federal Information Processing Standards) del estado donde reside el punto de monitoreo.
3. Codigo de Condado (county_id): El código FIPS del condado donde reside el punto de monitoreo.
4. Codigo de Sitio (site_id): Un número único que identifica un sitio particular dentro de un condado.
5. Código de Parámetro (parameter_id): El código AQS (siglas de Air Qualty System) correspondiente al parámetro medido por el monitor.
6. Latitude (latitude): La distancia angular del sitio de monitoreo al norte del ecuador medido en decimas de grado.
7. Longitud (longitude): La distancia angular del sitio de monitoreo al este del primer meridiano medido en decimas de grado.
8. Nombre del Parámetro (parameter_name): El nombre o la descripción asignada en el AQS al parámetro medido por el monitor. Puede referir a contaminantes o no contaminantes.
9. Duración de la medida (sample_duration): El tiempo que transcurre mientras que el aire pasa por el instrumento de medida antes de ser analizado (es decir, de que se realice la medida). Para monitoreos continuos puede usarse para representar el tiempo promedio de varias medidas (por ejemplo, un valor de 1 hora puede usarse para representar el promedio de cuatro medidas de un minuto tomadas cada 15 minutos).
10. Fecha de la medida (sample_date): La fecha correspondiente a la toma de la medida. Todas se anotan con respecto al uso horario de la zona de monitoreo.
11. Unidades de Medida (units_of_measure): La unidad de medida para el parámetro.
12. Tipo de Evento (event_type): Indica si algún dato incluido fue medido durante eventos excepcionales (algún evento que afecte la calidad del aire del cual se

carece el control, como un evento atmosférico natural o una catástrofe no controlable)

1. ‘No Events’ significa que no ocurrió ningún evento;
2. ‘Events Included’ significa que ocurrió algún evento y los datos anotados corresponden a este;
3. ‘Events Excluded’ significa que ocurrió un evento, pero se excluyeron los datos medidos durante este;
4. ‘Concurred Events Excluded’ significa que ocurrió un evento, pero solo algunos datos fueron incluidos.
   Si un evento que afecte el parámetro en cuestión ocurrió, los datos tendrán múltiples medidas de cada punto de monitoreo.
13. Conteo de Observaciones (obs_count): El número de observaciones (medidas) tomadas durante el día.
14. Media aritmética (arithmetic_mean): El promedio o media aritmética de los valores obtenidos en el día para un parámetro.
15. Código de Método (method_id): Un código interno del sistema que indica el método (proceso, equipo y protocolo) usados para realizar la medida.
16. Nombre del Método (method_name): Una corta descripción del proceso, equipo y protocolo usados para realizar una medida.
17. Nombre del Sitio (site_name): El nombre del sitio (si es que lo tiene) dado por el estado o por la agencia de monitoreo que lo opera.
18. Dirección del Sitio (site_address): La dirección aproximada de la calle donde reside el punto de monitoreo.
19. Nombre del Estado (state_name): Nombre del estado donde se localiza el punto de monitoreo.
20. Nombre del Condado (county_name): Nombre del condado donde se localiza el punto de monitoreo.
21. Nombre de la Ciudad (city_name): El nombre de la ciudad donde está ubicado el punto de monitoreo. Representa los límites legales de las ciudades y no de las zonas urbanas.
22. Nombre CBSA (CBSA_name): El nombre del área estadística básica (en inglés, core based statistical área) metropolitana donde esta ubicado el punto de monitoreo.
23. Código Alfabético (alpha_code): código de letras usado para representar cada uno de los estados de forma univoca.

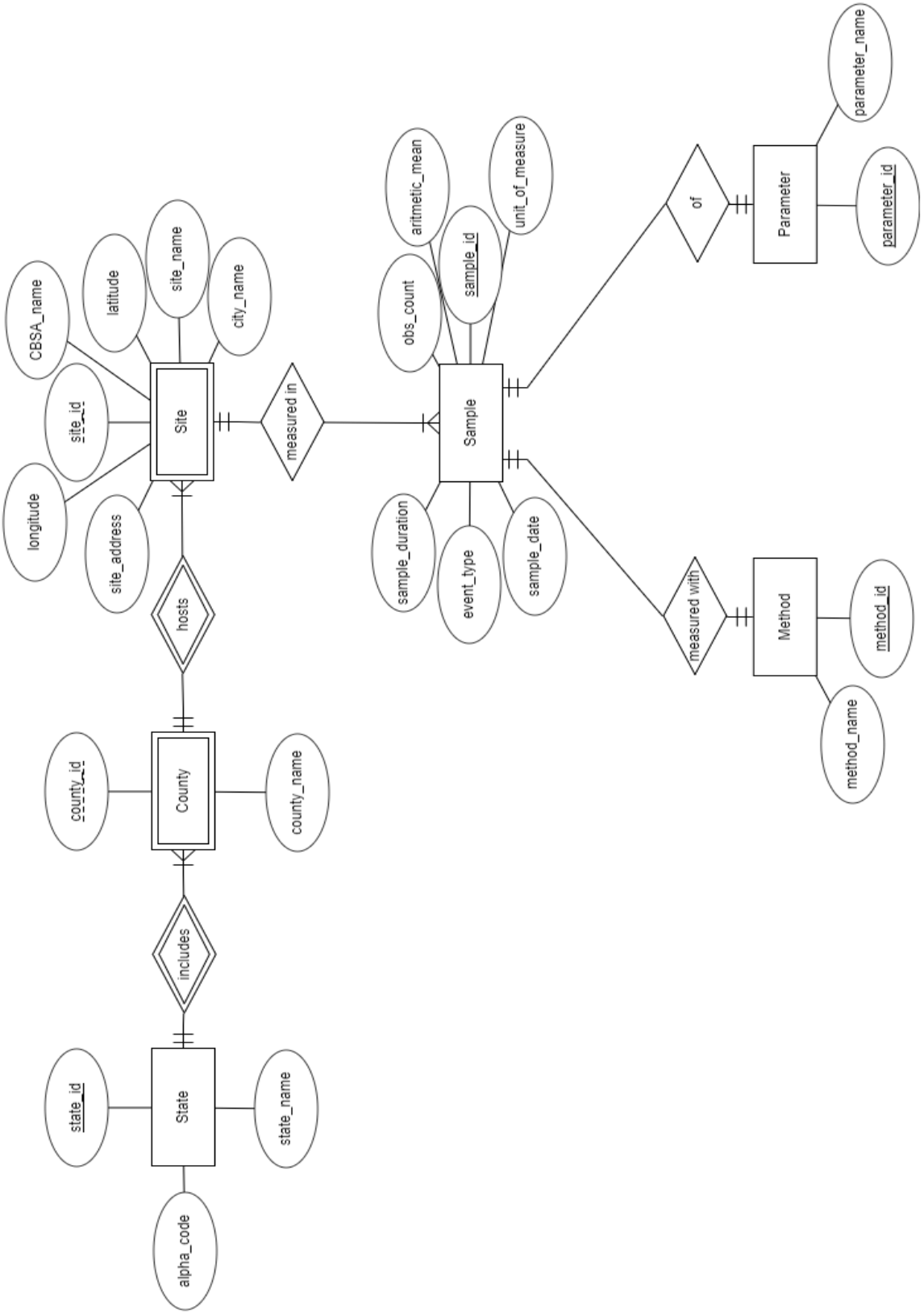
Reconocimiento de Tablas:

En este apartado se incluyen las tablas inicialmente incluidas en la base de datos, aclarando los atributos o campos que lo componen, el tipo de datos usados en cada atributo, las claves primarias (PK), las claves foráneas (FK) y las claves candidatas (CK).

Base de datos: Hazardous Air Pollutants

Tabla	Campos	PK	CK	FK	Tipo
State	state_id	Si			Int
	state_name				Text
	alpha_code		Si		Text
County	county_id	Si			Int
	county_name				Text
	state_id			Si	Int
Site	site_id	Si			Int
	site_name				Varchar
	site_address				Varchar
	latitude				Decimal
	longitude				Decimal
	CBSA_name				Text
	county_id			Si	Int
Sample	sample_id	Si			Int
	sample_date				Datetime
	sample_duration				Varchar
	event_type				Text
	obs_count				Int
	aritmetic_mean				Decimal
	unit_of_measure				Varchar
	method_id			Si	Int
	parameter_id			Si	Int
Method	method_id	Si			Int
	method_name		Si		Varchar
Parameter	parameter_id	Si			Int
	parameter_name		Si		Varchar

Diagrama de Entidad-Relación.



Modificaciones:

Preprocesamiento de los datos:

Debido a que el dataset original que empleamos es muy extenso, y por tanto difícil de manipular manualmente en programas como Microsoft Excel, se decidió hacer un pre-procesamiento de los datos. Para ello se empleó el lenguaje de programación Python 3, en el entorno Google Colaboratory. Colaboratory aporta gran velocidad para el procesamiento de datos (por hacerse en VPS alojados en la infraestructura de Google) a comparación de correr de forma local en un entorno como JupyterLab. El notebook se encuentra alojado en GitHub, en el siguiente link : https://colab.research.google.com/github/egonik-unlp/random_projects/blob/master/data_analytics/dataset_kaggle_DA_metadata.ipynb. Se resumen los pasos realizados en el siguiente punteo:

1. Se renombraron las columnas de acuerdo con el siguiente conjunto de claves -> valores:
   ```
   {
      'state_code': 'state_id',
      'county_code':'county_id',
      'parameter_code':'parameter_id',
      'method_code':'method_id',
      'site_num': 'site_id',
      'date_local':'sample_date',
   }
   ```
2. El dataset original tiene aproximadamente 8x10^6 instancias, que lo hacen muy dificil de usar. Desde ahora en adelante usaremos una fracción del dataset. Para ello elegimos rows al azar (usando el generador de números aleatorios MT19937. Se muestra el script empleado a continuación:

   ```
   def remove_n_rows(N:float, dataframe:pd.core.frame.DataFrame) -> pd.core.frame.DataFrame:
     '''
     N representa la fraccion en la que quiero reducir el dataframe
     '''

     number_of_rows = len(dataframe) -  int(len(dataframe)/N)
     drop_indices = np.random.choice(dataframe.index ,number_of_rows,replace = False)
     d(md(f'El dataframe original tenia {len(dataframe)} rows, se redujeron a {len(dataframe) - len(drop_indices)}' ))
     return dataframe.drop(drop_indices)

   df2 = remove_n_rows(15.01,df)
   ```

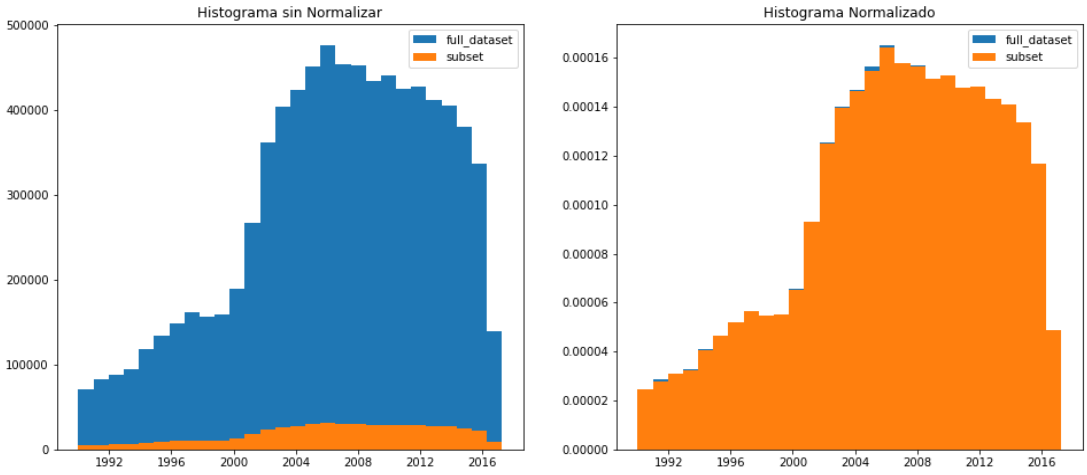
3. Debido a que county y site son entidades débiles, para identificarlas de forma única, se generaron ids únicos mediante la concatenación de state_id y county_id para county, y county_id y site_id para site.
4. Se separaron las tablas del dataset original en tablas separadas por entidad (según el diagrama ER mostrado antes).
5. Debido a que el dataset original está estructurado en forma de medidas individuales, todas las tablas tendrán valores repetidos, correspondientes a la cada una de las medidas (a excepción de la tabla medidas. Por ejemplo, la tabla states (donde cada instancia debería representar un estado) tiene tantas

entradas como medidas se realizaron. Esto no tiene sentido, por ello nos quedaremos solo con los valores únicos de cada tabla, identificados por el ID.

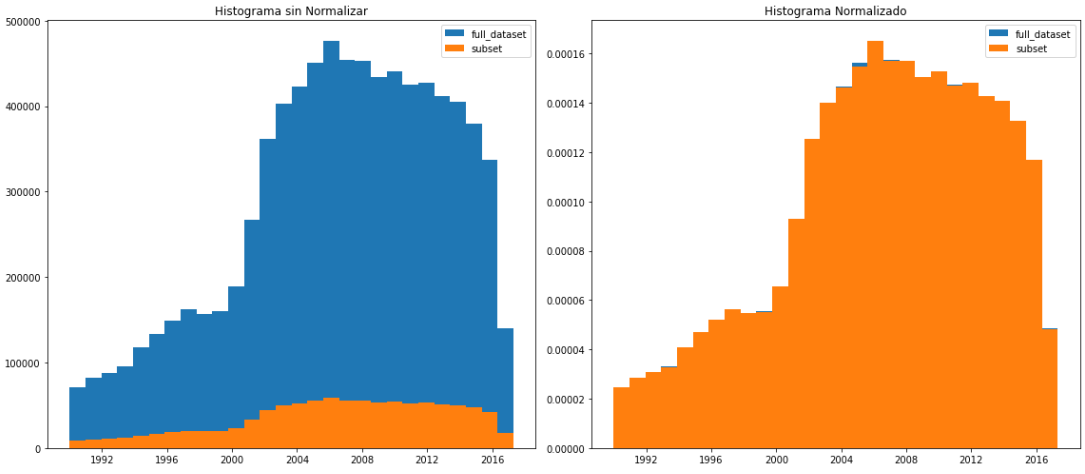
6. Cada una de las tablas generados en el paso 4 se guardó como una pestaña en un archivo xlsx. Este último es el dataset final.

Aclaraciones:

En los siguientes histogramas se muestra la distribución de fechas en el dataset original y en el muestreo, de los cuales se infiere que el muestreo parece ser representativo.



Mas adelante, usando el mismo mecanismo de sampleo se aumento el número de medidas seleccionadas a 1 millon aproximadamente. Al hacer nuevamente un análisis mediante histogramas, se obtuvieron los siguientes resultados:



Una vez que se estableció la base de datos a utilizar, se realizaron las siguientes modificaciones:

Entrega: Modelo Relacional.

- Tabla: ‘site_location’. Se incorporó la información de latitud y longitud para cada sitio (o site) como una tabla con relación 1:1 con la tabla ‘site’. Esta información se adjunta en la entrega en el archivo site_location_table.csv.

- Tabla: ‘parameter_category’. Se incorporó una tabla que categoriza los parámetros medidos, en categorías relacionadas a su comportamiento fisicoquímico. Se confeccionó manualmente el archivo fuente ‘cats.json’, que mapea ‘parameter_id’, con las categorías. También se adjunta.
  Debido a que la interpretación de PowerBI del archivo previamente mencionado no fueron totalmente correctos, se debió transponer la tabla interpretada, y se eligió la primera fila como etiqueta.

Entrega: Columnas y medidas calculadas.

- Se agregó la tabla calendario ‘sample_date’

Entrega: Medidas calculadas avanzadas.

- En la tabla ‘sample’ se definió la columna ‘year’, de forma de facilitar la implementación posterior de ciertos filtros y visuaizaciones.

Transformaciones de datos:

Entrega: Extracción de Información.

- Tabla: ‘country’.
  - Cambio de nombre de la columna: ‘county_unique_id’ a ‘county_id’.
  - Cambio de tipo de datos: números a texto (‘county_id’ y ‘state_id’).

- Tabla: ‘method’.
  - Cambio de tipo de datos: números a texto (‘method_id’).

- Tabla: ‘parameter’.
  - Cambio de tipo de datos: números a texto (‘parameter_id’).

- Tabla: ‘sample’.
  - Cambio de nombre de columna: ‘site_unique_id’ a ‘site_id’.
  - Cambio de tipo de datos: número a texto (‘sample_id’, ‘site_id’, ‘parameter_id’, ‘method_id’).

- Tabla: ‘site’.
  - Cambio de nombre de columna: ‘county_unique_id’ a ‘county_id’, ‘site_unique_id’ a ‘site_id’.
  - Cambio de tipo de datos: número a texto (‘county_id’, ‘site_id’).

- Tabla: ‘state’.
  - Cambio de tipo de datos: número a texto (‘state_id’).

Entrega: Columnas y medidas calculadas.

- En la tabla ‘sample’, la columna ‘sample_duration’ se tranformó en medida. Los valores originales tenían formato del estilo ‘24 HOURS’, ‘3 HOURS’, ‘15 MINUTES’. Los valores en minutos fueron reemplazados por sus equivalentes en horas. Luego la columna fue separada por el divisor ’ ‘ de forma de separar los valores numéricos del descriptor ‘HOURS’, los cuales posteriormente fueron descartados. La nueva columna con valores numéricos fue denominada ‘sample_duration_in_hours’. Vale aclarar que había datos con el título COMPOSITE DATA; esos datos fueron ignorados en este análisis y fueron reemplazados por 0.

Columnas y medidas calculadas:

Entrega: Modelo Relacional.

- Tabla: ‘method’. Debido a que el campo ‘method_name’ incorporaba información de la técnica de medida y el equipo empleado separadas por ‘ - ‘, se dividió esa columna empleando este separador en otras dos nuevas (‘method_name’ y ‘technique_name’).

- Tabla: ‘site’. En el atributo ‘city_name’ se indica ciudades cercanas a puntos de medida, en caso de no haber ninguna se indica ‘Not in a city’. En consecuencia, generamos una nueva columna que indicase si el punto de medida se encontraba o no en un contexto urbano (‘is_urban’)

Entrega: Columnas y medidas calculadas.

- En la tabla ‘sample’, se calculó la cantidad ‘measurement_total’ a partir de la multiplicación de ‘arithmetic_mean’ y ‘obs_count’. Este valor representa la medida total en cada proceso de muestreo.

- En la tabla ‘sample, se generaron las columnas calculadas ‘sample_duration_in_minutes’ y ‘sample_duration_in_hours’. La generación de la columna ‘sample_duration_in_hours’ fue descripta en el apartado de ‘Transformación de datos’, mientras que la columna ‘sample_duration_in_minutes’ se obtuvo multiplicando los valores de ‘sample_duration_in_hours’*60.

- En la tabla ‘sample’, para obtener la duración en minutos de cada proceso de muestreo individual, se calculó la columna cociente ‘[sample_duration_in_minutes]/[obs_count]’ que se denominó ‘sample_duration_per_obs’.

Entrega: Visualizaciones y filtros.

- En la tabla ‘parameter’ se incorporó una columna ‘avg_measurements’ que, para cada parámetro evaluado, muestra su valor promedio. Se agregó por considerar que podría ser útil para gráficos en el futuro. Dicha columna se obtuvo al hacer un

‘left join’ con la tabla ‘sample’, recuperando el campo ‘arithmetic_mean’ y luego agrupando por ‘parameter_name’ y ‘parameter_id’, con promedio como variable de sumarización.

Entrega: Medidas calculadas avanzadas.

- En la tabla ‘site’ se incorporó la columna ‘sample_by_site’ la cual resume el número de medidas realizadas por sitio de monitoreo. Para lograr este análisis se utilizó la medida ‘*CALCULATE(COUNT('sample'[sample_id]),site[site_id])*’. Aunque no es necesaria para esta entrega, esta columna se añadió por interés de los integrantes del grupo, pues puede servirnos posteriormente para hacer análisis estadísticos.

- En la tabla ‘county’ se incorporó la columna ‘sample_by_county’ la cual indica el número de medidas realizadas por condado en total. Para lograr este análisis se definió la variable ‘VAR medidas = CALCULATE(COUNT('sample'[sample_id]), site[site_id])’ y luego la medida ‘RETURN CALCULATE(medidas, county[county_id])’.

- En la tabla ‘state’ se incorporó la medida ‘samples_by_state_00-10’, la cual indica el número total de medidas realizadas por estado entre los años 2000 y 2010, expresándolos como un porcentaje del total de las medidas realizadas. Se considera que este tipo de análisis puede generar resultados estadísticamente interesantes que seguiremos desarrollando mas adelante, siendo este específicamente desarrollado para esta entrega solamente. Para lograr este análisis se definieron las variables ‘VAR total = CALCULATE(COUNT('sample'[sample_id]), state[state_id])’ y ‘VAR decada00 = DATESBETWEEN ('sample'[sample_date], DATE(2000, 01, 01), DATE(2009, 12, 31))’. Posteriormente se definió la función ‘RETURN TOTALYTD (CALCULATE(COUNT('sample'[sample_id]), state[state_id]), decada00)*100/total’.

- Se definió el parámetro ‘Porcentaje’, el cual consiste en valores entre 0 y 1 definidos con aumentos de 0,1 unidades. En la tabla ‘state’ se incorporó la medida ‘sample_percentage’ la cual nos permite visualizar las medidas con un filtro del 0% (correspondiente al 0 del slider) al 100% (correspondiente al 1). Esta medida fue definida específicamente para esta entrega. Se definió la variable ‘VAR medidasporestado = CALCULATE(COUNT('sample'[sample_id]), state[state_id])’ y posteriormente la función ‘RETURN medidasporestado*Porcentaje[Valor Porcentaje]’

- Se definió la medida ‘arit_mean_by_parameter’ en la tabla ‘sample’, la cual sirve para calcular la media aritmética de los valores expresados en la comuna ‘aritmetic_mean’ según utilizando como filto la columna ‘parameter_id’. Esta medida fue definida como ‘arit_mean_per_parameter = CALCULATE(AVERAGE('sample'[arithmetic_mean]),'sample'[parameter_id])’

- En la tabla ‘sample’ se definió la columna ‘year’, de forma de facilitar la implementación posterior de ciertos filtros y visuaizaciones.