## DESCRIPCIÓN DEL CONJUNTO DE DATOS A TRABAJAR EN EL PROYECTO DE TRATAMIENTO DE DATOS

MINERÍA DE DATOS

08 de mayo de 2023

|  |  |
| --- | --- |
| **Grupo:** | 3CV15 |

|  |
| --- |
| Alumnos: |
| Flores Ponce Alan Marcelo |
| García Cruz Octavio Arturo |
| Sampayo Hernández Mauro |

# Proyecto No. 1. Tratamiento de datos

## Generar un reporte en BIRT que considere los aspectos que se describen a continuación.

1. **Descripción del conjunto de datos:**

Progreso mundial de vacunación contra COVID-19 *COVID-19 World Vaccination Progress*

Vacunación diaria y total contra COVID-19 en el mundo de Our World in Data Enlace de acceso:

<https://www.kaggle.com/gpreda/covid-world-vaccination-progress>

Los datos se recopilan diariamente del repositorio Our World in Data GitHub para covid-19, se fusionan y se cargan. Los datos de vacunación a nivel de país se recopilan y reúnen en un solo archivo. Luego, este archivo de datos se fusiona con el archivo de datos de ubicaciones para incluir información sobre las fuentes de vacunación. Se incluye un segundo archivo, con información de los fabricantes.

Créditos:

Gabriel Preda. Científico de datos

# Primer archivo

Los datos (vacunaciones del país) contienen la siguiente información:

|  |  |  |  |
| --- | --- | --- | --- |
| **Atributo** | **Descripción** | **Tipo** | **Dominio** |
| country | Este es el país para el que se proporciona la información de vacunación | Categórico | 222 valores, nombres de países. |
| iso\_code | Código ISO de cada país | Categórico | 222 valores, un  código por país |
| date | Fecha para la entrada de datos; para algunas de las fechas solo se tienen las vacunas diarias, para otras, solo el total (acumulativo). | Numérico | 01/12/2020 a  17/09/2021 |
| total\_vaccinations | Es el número absoluto de inmunizaciones totales en el país | Numérico | (0, ∞) |
| people\_vaccinated | Una persona, según el esquema de inmunización, recibirá una o más (normalmente 2) vacunas; en un momento determinado, el número de vacunaciones puede ser mayor que el número de personas. | Numérico | (0, ∞) |
| people\_fully\_vaccinated | Este es el número de personas que recibieron el conjunto completo de inmunización de acuerdo con el esquema de inmunización (típicamente 2), en un momento determinado, puede haber un cierto número de personas que recibieron una vacuna y otro número (menor) de personas que recibieron todas las vacunas del esquema. | Numérico | (0, ∞) |
| daily\_vaccinations\_raw | Para una determinada entrada de datos, el número de vacunaciones para esa fecha/país. | Numérico | (0, ∞) |
| daily\_vaccinations | Para una determinada entrada de datos, el número de vacunaciones para esa fecha/país. | Numérico | (0, ∞) |
| total\_vaccinations\_per\_ hundred | Relación (en porcentaje) entre el número de vacunaciones y la población total hasta la fecha en el país. | Numérico | (0, 100) |
| people\_vaccinated\_per\_ hundred | Relación (en porcentaje) entre la población inmunizada y la población total hasta la fecha en el país. | Numérico | (0, 100) |
| people\_fully\_vaccinated  \_per\_hundred | Relación (en porcentaje) entre la población totalmente inmunizada y la población total hasta la fecha en el país. | Numérico | (0, 100) |
| daily\_vaccinations\_per\_ million | Relación (en ppm) entre el número de vacunación y la población total para la fecha  actual en el país | Numérico | (0, ∞) |
| vaccines | Número total de vacunas utilizadas en el país (actualizadas). | Categórico | CanSino, Covaxin, Moderna, Oxford/AstraZen eca, Pfizer/BioNTech, Sinopharm/Beiji ng, Sinovac, Sputnik V, Abdala, Soberana02, QazVac, Sinopharm/Haya tVax, Johnson&Johnso n. |
| source\_name | Fuente de la información (autoridad nacional, organización internacional, organización local, etc.) | Nominal | Diferentes instituciones gubernamentales. (Ministry of Health, World Health Organization, COVID19 Vaccine  Information platform, etc.) |
| source\_website | Sitio web de la fuente de información | Nominal | Cadena de texto, URL. |

# Segundo archivo

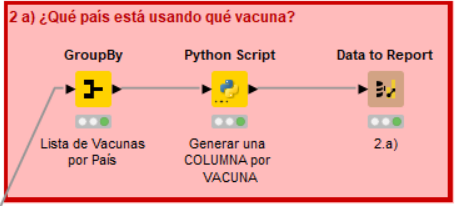
Hay un segundo archivo agregado recientemente (vacunas aplicadas de países por fabricante), con las siguientes columnas:

|  |  |  |  |
| --- | --- | --- | --- |
| **Atributo** | **Descripción** | **Tipo** | **Dominio** |
| location | país | Categórico | 222 valores, nombres de países. |
| date | fecha | Numérico | 01/12/2020 a 17/09/2021 |
| vaccine | Tipo de vacuna | Categórico | CanSino, Covaxin, Moderna, Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing, Sinovac, Sputnik V, Abdala, Soberana02,  QazVac, Sinopharm/HayatV ax, Johnson&Johnson. |
| Total\_vaccinations | Número total de vacunas | Numérico | número total de vacunas/a la fecha y tipo de vacuna |

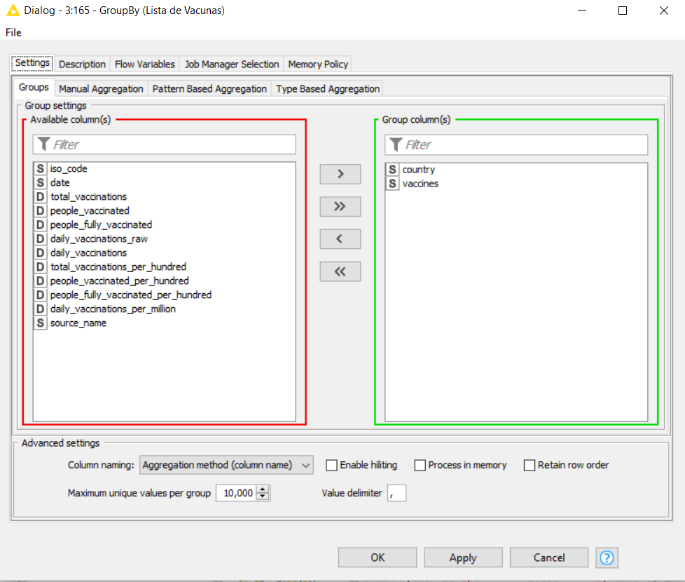
# Responder las tareas propuestas por el autor:

Realice un seguimiento de la vacunación contra COVID-19 en el mundo, responda las siguientes preguntas:

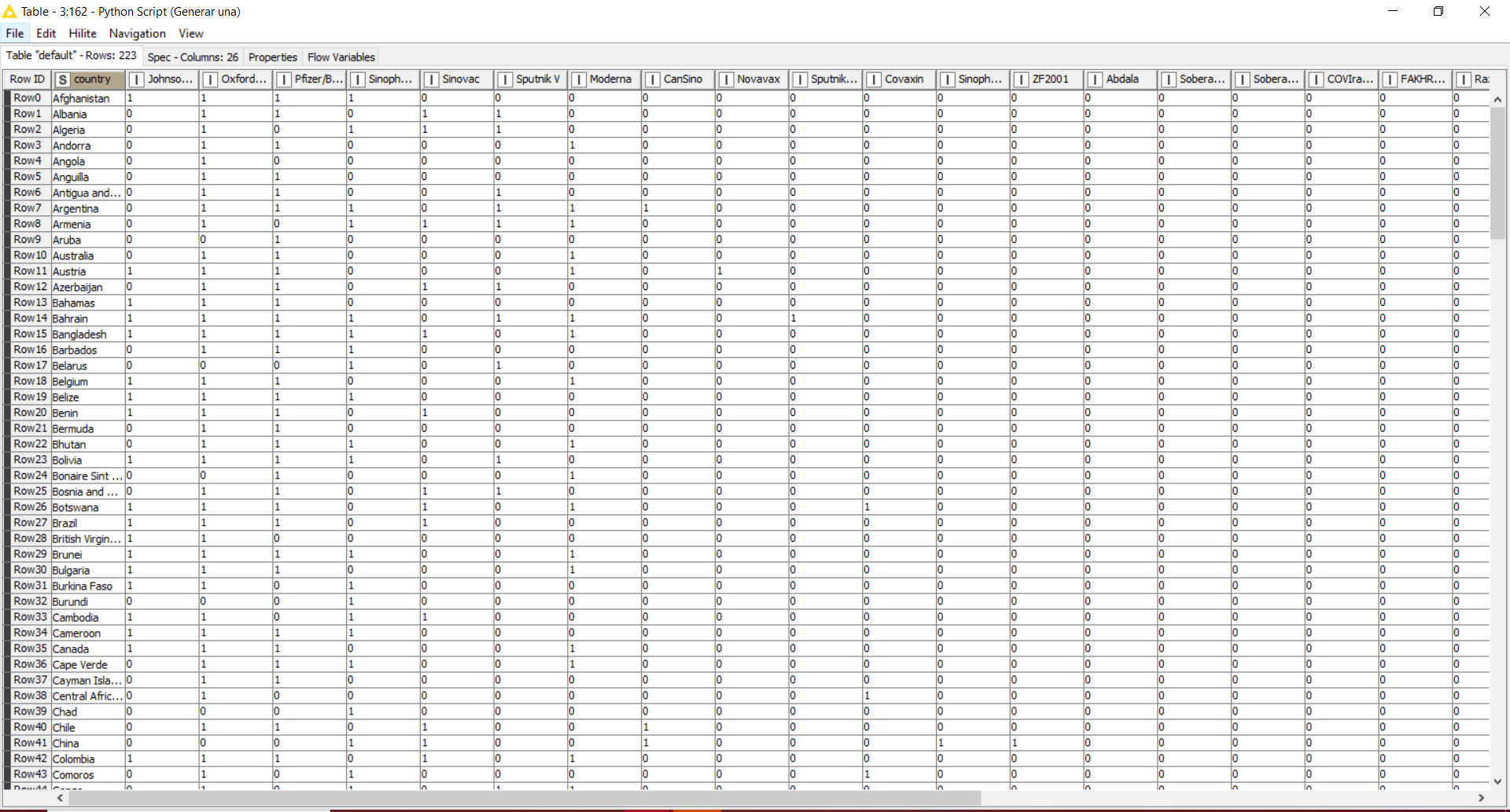
* 1. ¿Qué país está usando qué vacuna?



**Configuración “GroupBy”:**



**Tabla Resultante:**



**Código del Nodo de Python:**

El siguiente código se encarga de generar una COLUMNA por cada una de las VACUNAS enlistadas en todos registros presentes en la columna “vaccines”, sin REPETIR; y se le asigna un ‘1’ en caso de que dicho país haya hecho uso de dicha vacuna, y un ‘0’ en caso contrario.

**import** knime**.**scripting**.**io **as** knio

**import** numpy **as** np

# VARIABLES

vaccineName **=** "" # Almacena el nombre de una VACUNA

vaccines **=** **[]** # Almacena el nombre de TODAS las VACUNAS

vaccValues **=** **[]** # Almacena los valores de los registros de cada COLUMNA que sera generada

# por cada VACUNA presente en la Tabla de Entrada

indexCountry **=** 0 # Indice de registro de la TABLA

indexCol **=** 2 # Indice de columna de la TABLA

# Convirtiendo la tabla de ENTRADA a dataFrame

df **=** knio**.**input\_tables**[**0**].**to\_pandas**()**

# Obteniendo la longitud de la TABLA (cant. total de registros)

numRows **=** **len(**df**.**index**)**

# Iterando cada uno de los REGISTROS presentes en la Tabla

**for** index**,** row **in** df**.**iterrows**():**

# Iterando en cada caracter presente en un REGISTRO

**for** c **in** row**[**'vaccines'**]:**

# -------------------------------------------------------------------------------------

# CASO: Si se lee una coma ',', significa que hemos llegado al final del nombre

# de una VACUNA

**if** c **==** ','**:**

# ---------------------------------------------------------------------------------

# CASO: Si el nombre de la VACUNA aun no se encuentra en la LISTA DE VACUNAS

**if** **not(**vaccineName **in** vaccines**):**

# Se agrega el nombre de la VACUNA a la LISTA DE VACUNAS

vaccines**.**append**(**vaccineName**)**

# Se crea una lista para almacenar los valores de cada uno de los registros

# con los que contara la COLUMNA de la VACUNA registrada en la LISTA DE

# VACUNAS. La lista se llenara con ceros '0'

vaccValues**.**append**(**np**.**zeros**(**numRows**,** dtype **=** **int))**

# ---------------------------------------------------------------------------------

# Se le asigna el valor de uno '1' a el registro del pais que esta haciendo uso

# de la VACUNA, cuyo nombre fue leido

vaccValues**[**vaccines**.**index**(**vaccineName**)][**indexCountry**]** **=** 1

# Se vacia la variable de NOMBRE DE VACUNA para almacenar los nombres de las

# vacunas posteriores

vaccineName **=** ""

# -------------------------------------------------------------------------------------

# CASO: Si se lee un solo ESPACIO VACIO dentro de la variable del NOMBRE DE LA VACUNA,

# este se elimina

**elif** vaccineName **==** " "**:**

# Se vacia la variable de NOMBRE DE VACUNA

vaccineName **=** ""

# Se lee un caracter

vaccineName **=** vaccineName **+** c

# -------------------------------------------------------------------------------------

# CASO POR DEFECTO

**else:**

# Se lee un caracter

vaccineName **=** vaccineName **+** c

# -------------------------------------------------------------------------------------

# -----------------------------------------------------------------------------------------

# NOTA: Se agrega esta validacion, debido a que al finalizar la lectura de un REGISTRO, no

# hay un caracter que defina el final del nombre de la ultima VACUNA ENLISTADA

# CASO: Si el nombre de la VACUNA aun no se encuentra en la LISTA DE VACUNAS

**if** **not(**vaccineName **in** vaccines**):**

# Se agrega el nombre de la VACUNA a la LISTA DE VACUNAS

vaccines**.**append**(**vaccineName**)**

# Se crea una lista para almacenar los valores de cada uno de los registros con los

# que contara la COLUMNA de la VACUNA registrada en la LISTA DE VACUNAS. La lista se

# llenara con ceros '0'

vaccValues**.**append**(**np**.**zeros**(**numRows**,** dtype **=** **int))**

# Se le asigna el valor de uno '1' a el registro del pais que esta haciendo uso de la VACUNA,

# cuyo nombre fue leido

vaccValues**[**vaccines**.**index**(**vaccineName**)][**indexCountry**]** **=** 1

# -----------------------------------------------------------------------------------------

# Se aumenta el valor del Indice de REGISTROS por 1

indexCountry **+=** 1

# Se vacia la variable de NOMBRE DE VACUNA

vaccineName **=** ""

# Se generan las COLUMNAS para cada VACUNA de la LISTA DE VACUNAS

**for** vacc **in** vaccines**:**

# Insercion de la COLUMNA

df**.**insert**(**indexCol**,** vacc**,** vaccValues**[**vaccines**.**index**(**vacc**)])**

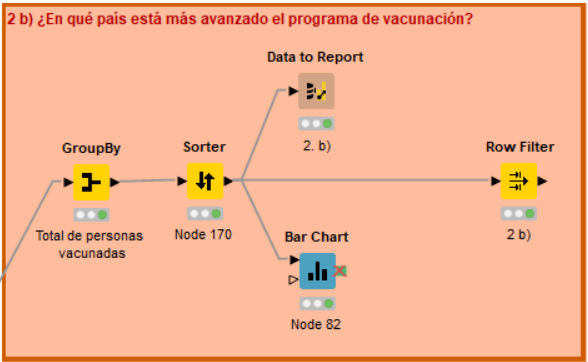
# Se aumenta el valor del Indice de COLUMNAS por 1

indexCol **+=** 1

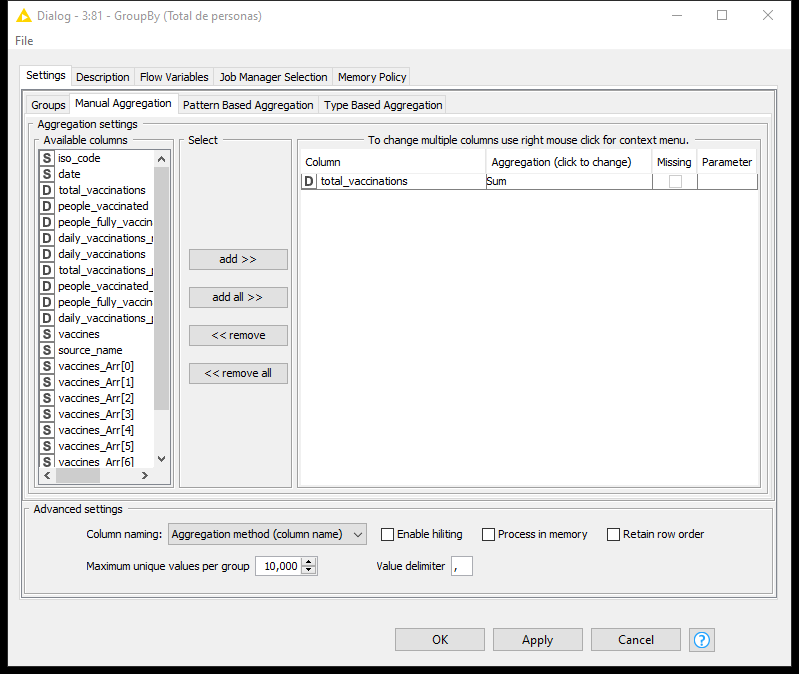
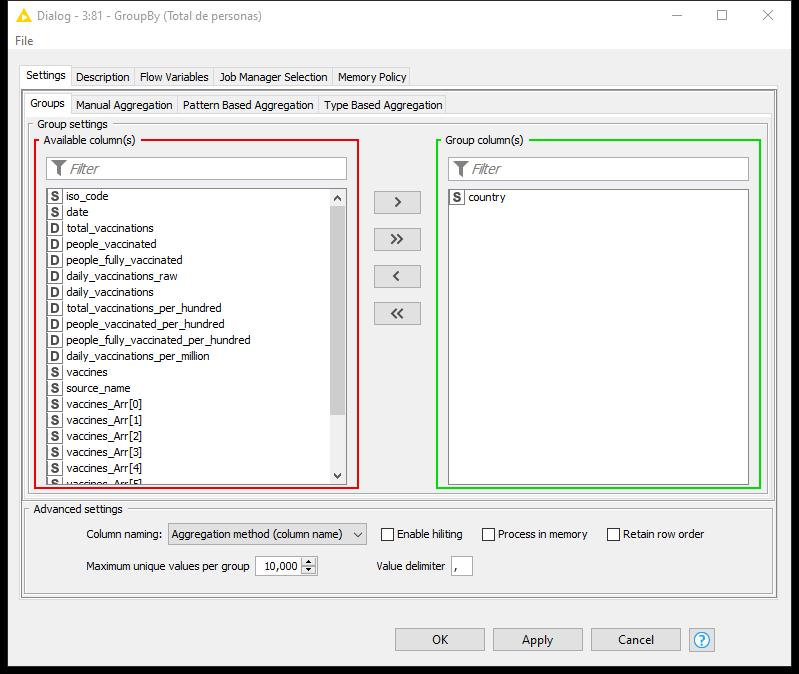
# Convirtiendo el dataFrame a un dataTable e igualandolo a la tabla de SALIDA

knio**.**output\_tables**[**0**]** **=** knio**.**Table**.**from\_pandas**(**df**)**

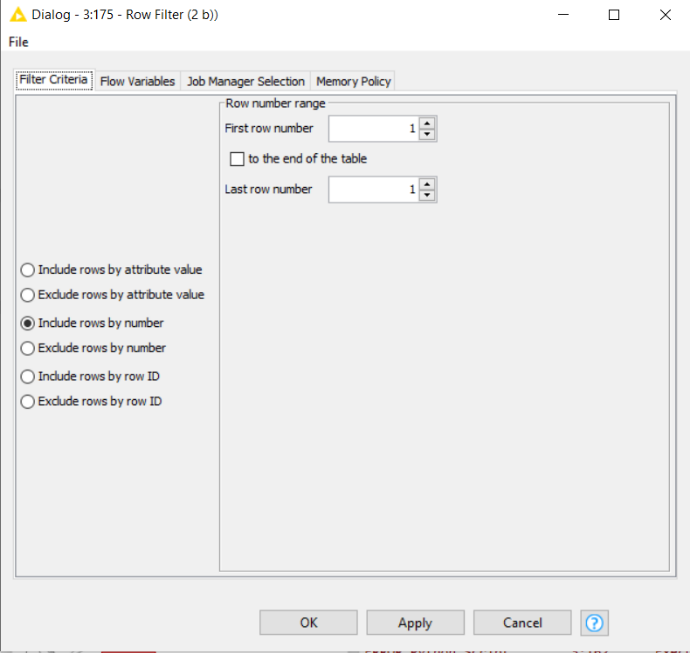
* 1. ¿En qué país está más avanzado el programa de vacunación?



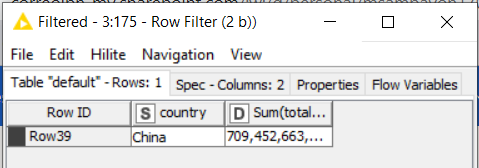
**Configuración “GroupBy”:**



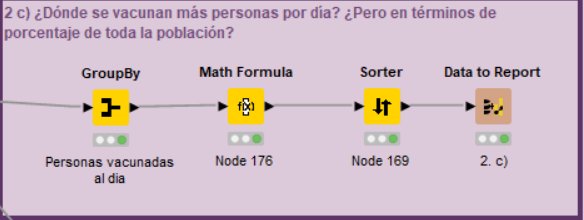
**Configuración “Row Filter”:**



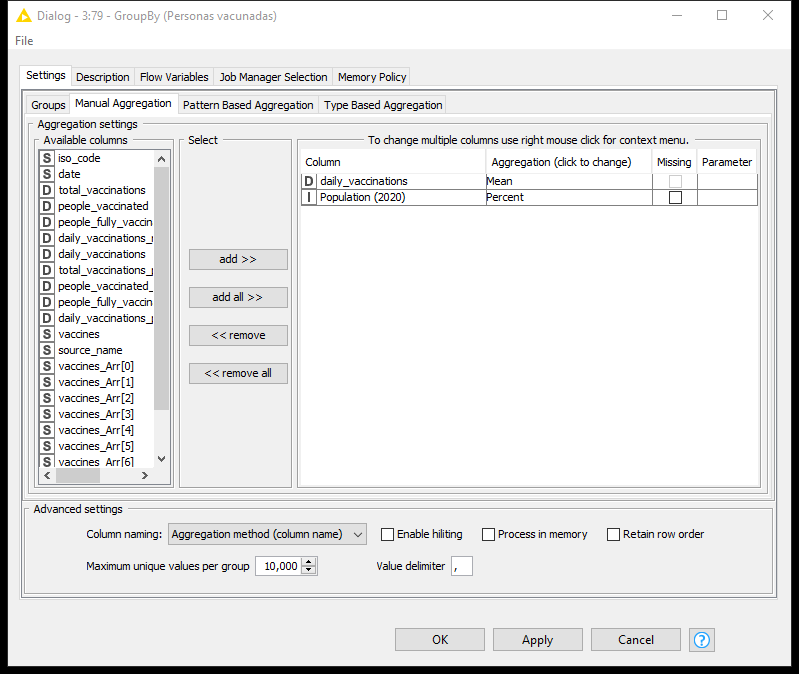
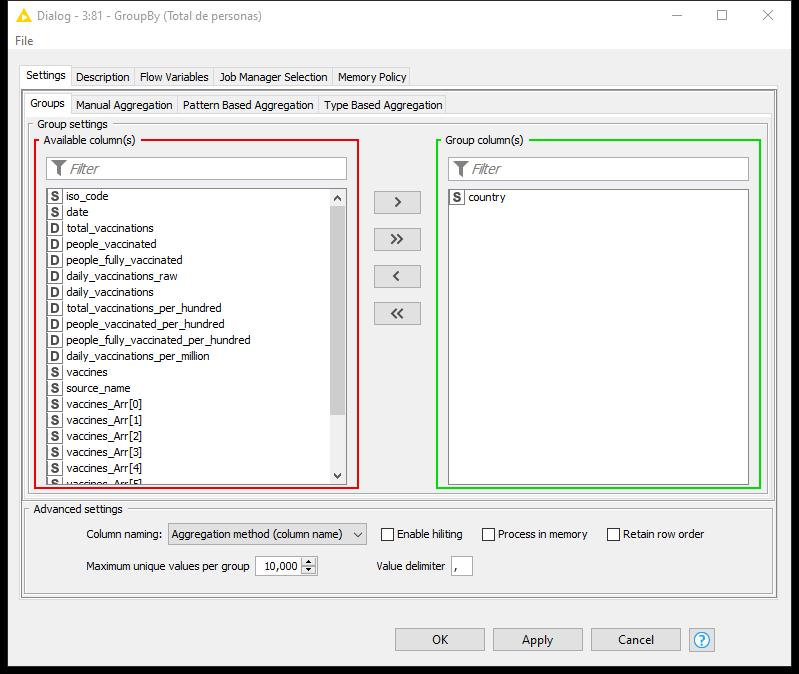
**Tabla Resultante:**



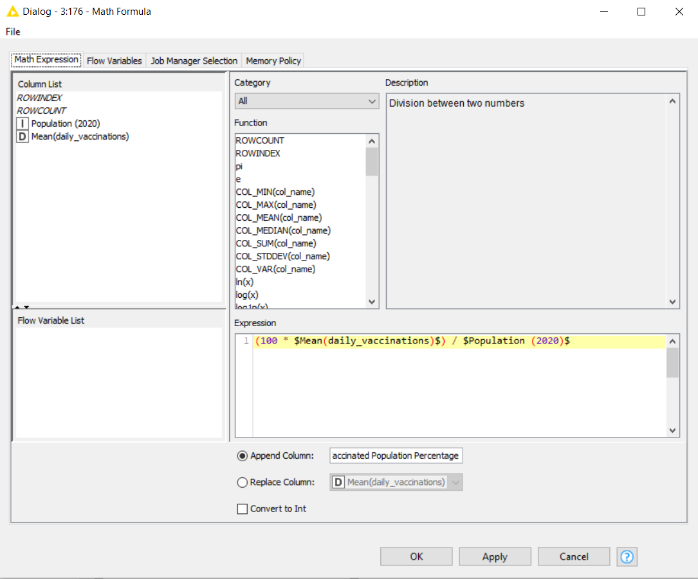
* 1. ¿Dónde se vacunan más personas por día? ¿Pero en términos de porcentaje de toda la población?



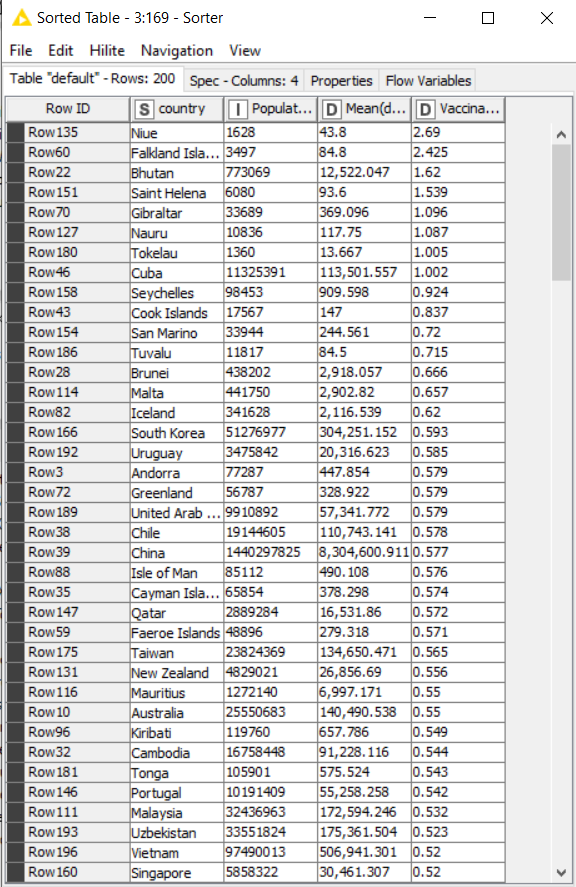
**Configuración “GroupBy”:**



**Configuración “Math Formula”:**



**Tabla Resultante:**



# Problemas iniciales:

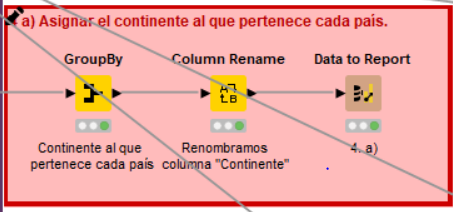
* Un problema que presenta es que los nombres de las vacunas están en forma de lista.

Hay que analizar la forma de identificar la cantidad de cada tipo de vacunas que se están aplicando en total en el conjunto de datos

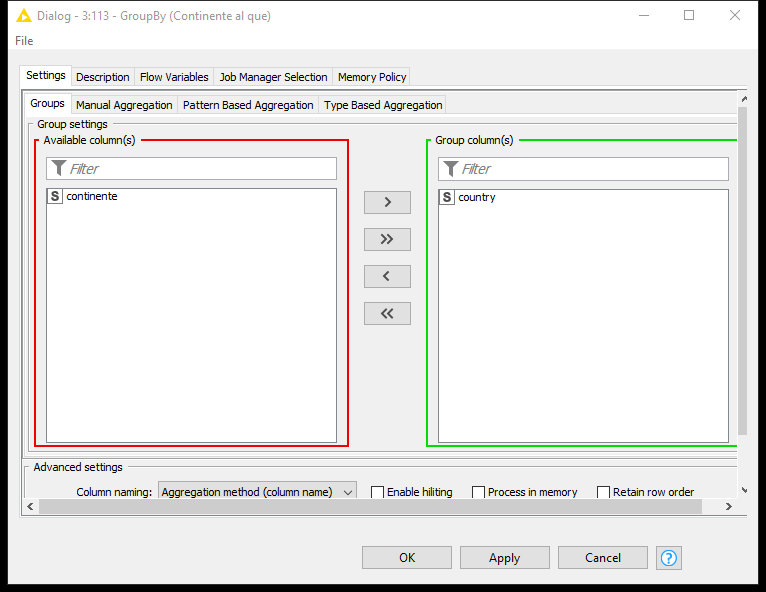
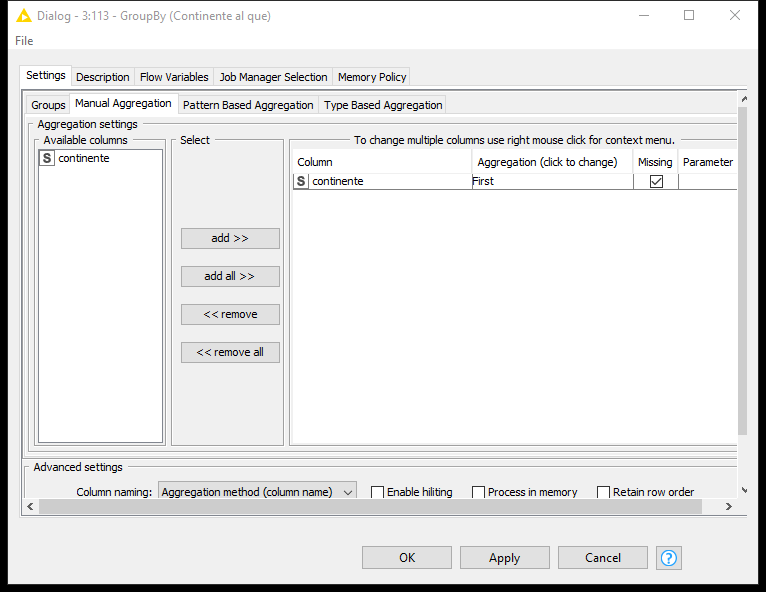
* Algunos países no reportan cifras de vacunación todos los días, hay que ver cuáles son.
* Hay que tener cuidado, ya que la cantidad de vacunas se van acumulando
* Se pueden eliminar los valores faltantes ya que algunos países no reportan cifras de vacunación todos los días, quizá filtrar a partir de un umbral los países de los cuales no se tengan suficientes datos si es que los hay. También sería recomendable un binning para pasar el valor de "vacunas diarias" de numérico a categórico.

# Acciones a realizar:

1. Asignar el continente al que pertenece cada país. Asignar el continente a todos los países. (<https://gist.github.com/kintero/7d1db891401f56256c79>)



**Configuración “GroupBy”:**

**Tabla Resultante:**

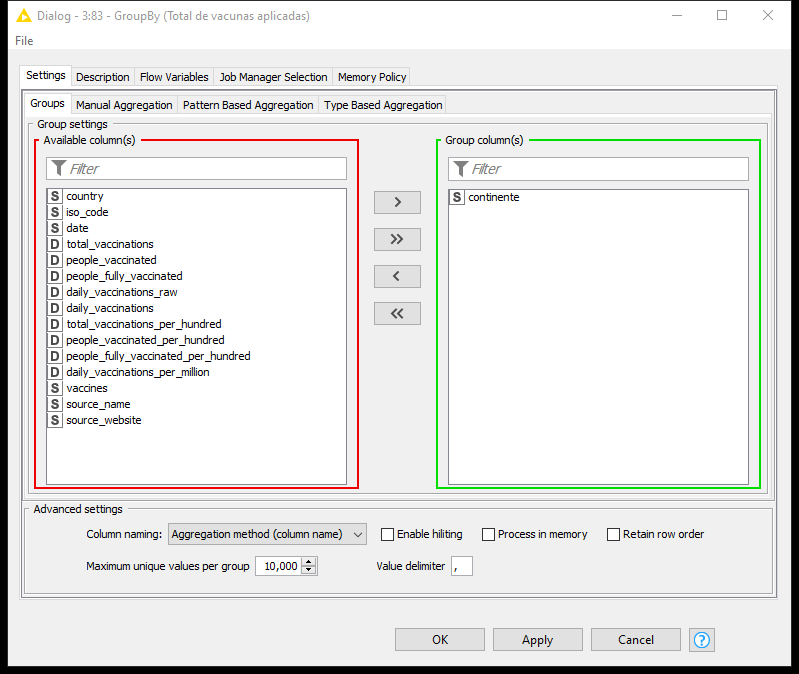
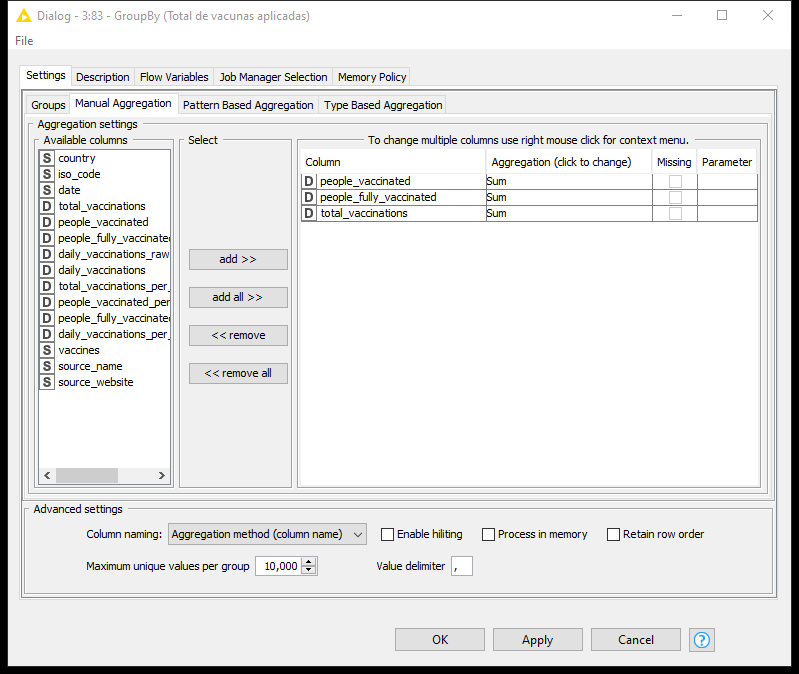


1. Reportar la cantidad total de vacunas que se aplicó en cada continente (Recategorizar)

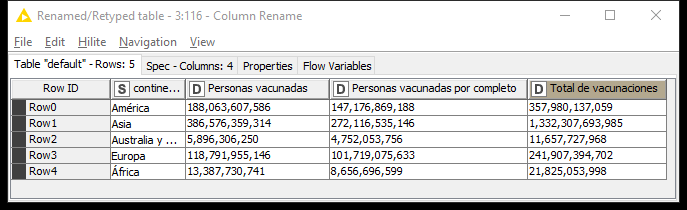
A picture containing text, screenshot, diagram, font

Description automatically generated

**Configuración “GroupBy”:**

**Tabla Resultante:**



1. Elegir cinco países de cada continente. Elija a México y muestre los tipos y la cantidad de vacunas que se aplicaron en cada país (sugerencia: usar country\_vaccinations\_by\_manufacturer)

A screenshot of a computer

Description automatically generated with medium confidence

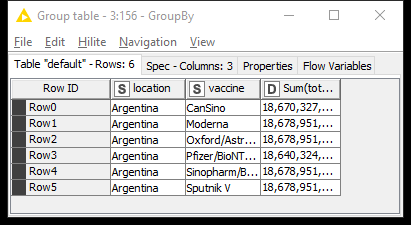
**Configuración “GroupBy”:**

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamenteInterfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Tabla Resultante:**

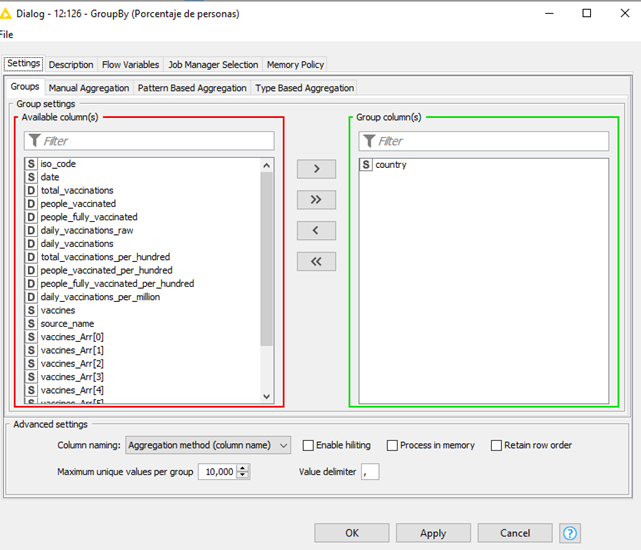
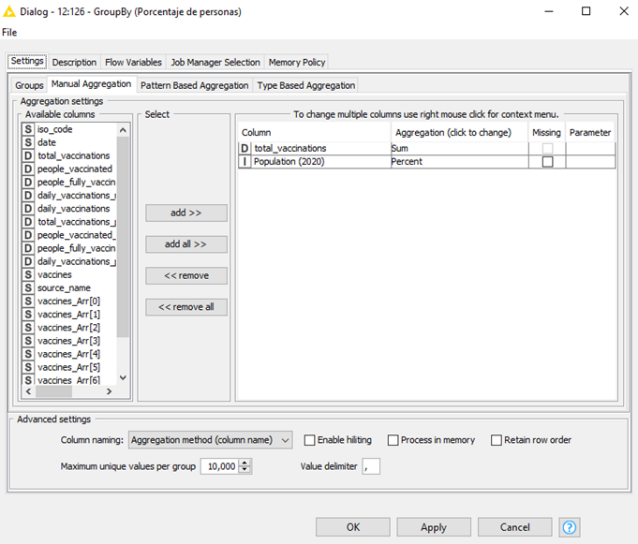


1. Reportar el porcentaje de la población que ha sido vacunada considerando la población del país.

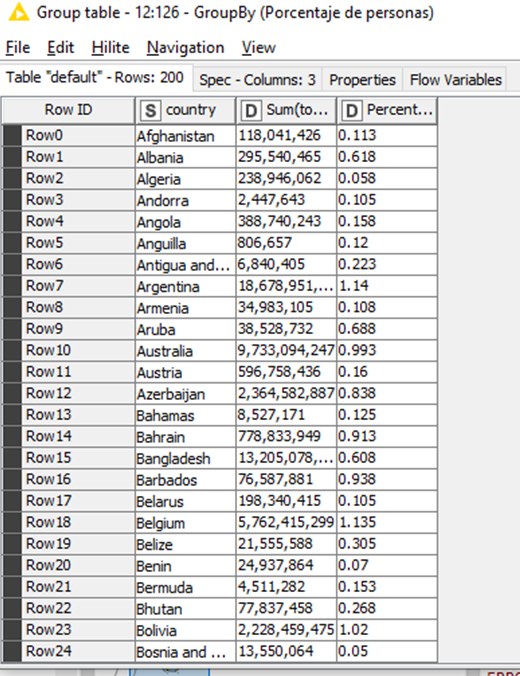
A picture containing text, screenshot, font, line

Description automatically generated

**Configuración “GroupBy”:**



**Tabla Resultante:**



1. Qué países son los que vacunaron más personas en cada continente.

A picture containing text, screenshot, diagram, line

Description automatically generated

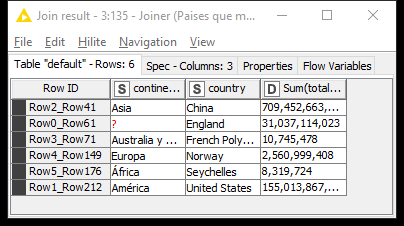
**Configuración “Joiner”:**

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamenteInterfaz de usuario gráfica

Descripción generada automáticamente

**Tabla Resultante:**



1. Crear un ciclo anidado en el cual reporte las vacunas por trimestre que se aplicaron en cada uno de los países (los cinco elegidos) de tres continentes

Escala de tiempo

Descripción generada automáticamente

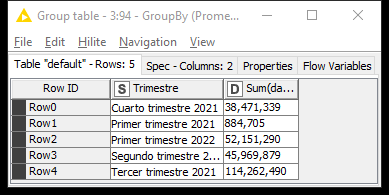
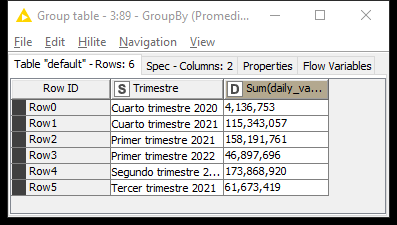
**Configuración “GroupBy”:**

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamenteInterfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

**Tabla Resultante:**



1. Reportar qué lugar ocupa México en América y a nivel mundial en cantidad de vacunas aplicadas, considerando la cantidad de la población.

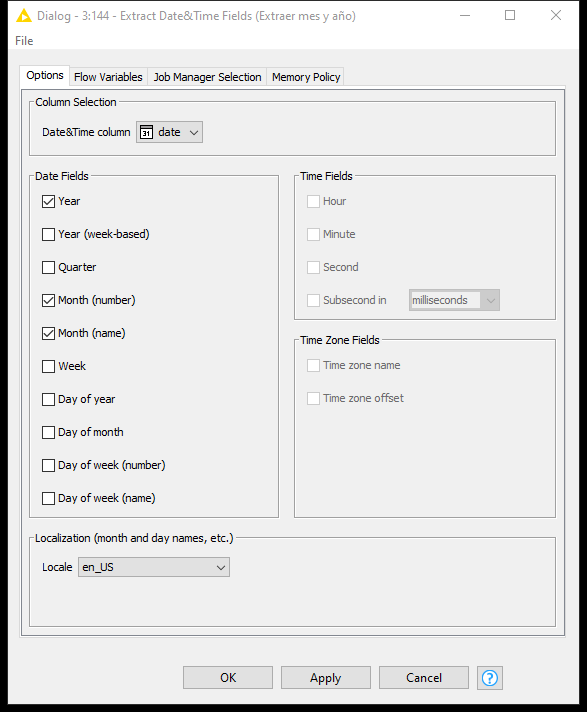
*Debido a que México no se encuentra dentro de la base de datos “vacunations by manufacturer”, no se ha podido realizar este inciso. Las posiciones de otros países ya han sido incluidas en capturas anteriores.*

# Elegir un país y realice lo siguiente:

1. Elegir un país y ver el total de vacunas aplicadas por mes, para hacer esto se tiene que filtrar el país y transformar la columna "date" para que esté en meses en lugar de días, al hacer esto se tendrían que sumar las demás columnas de "vacunas diarias".



**Configuración “Extract Date&Time Fields”:**



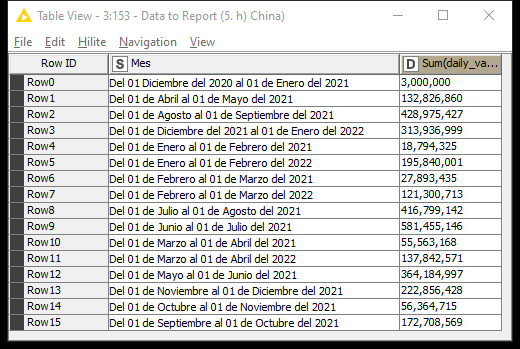
1. Se podría tomar un intervalo de fecha, por ejemplo, del 01 de enero de 2021 al 01 de marzo de 2021 (a elegir), y dentro de ese intervalo ver que vacuna fue la más aplicada

**Configuración “RuleEngine”:**

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Tabla Resultante:**

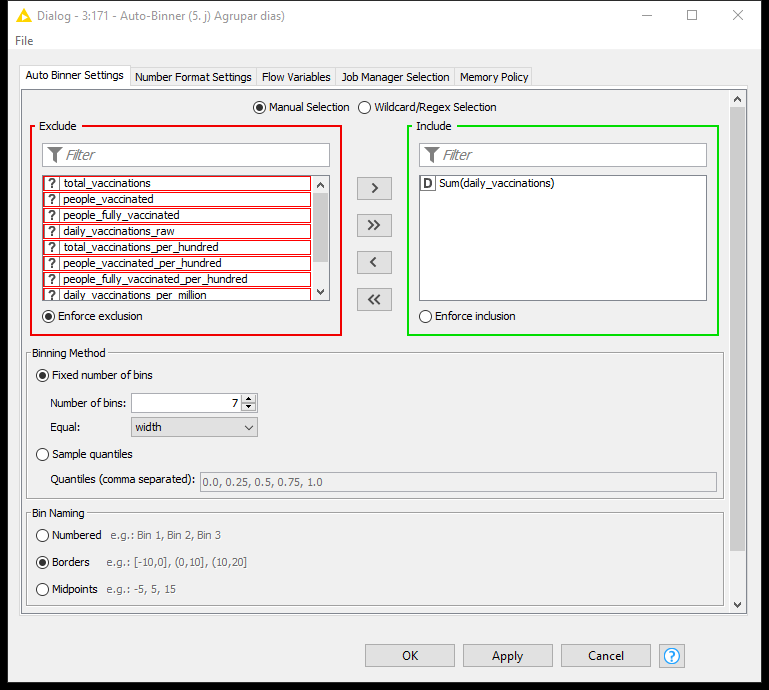


1. Crear binners de las vacunas aplicadas por día.

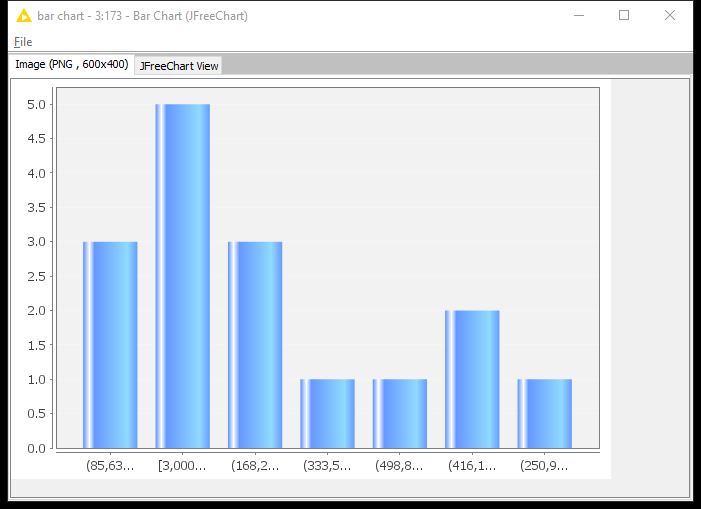
Diagrama

Descripción generada automáticamente

**Configuración “Auto-Binner”:**



1. Analizar las variables que tiene el conjunto de datos y presente aspectos descriptivos que es necesario resaltar. Use gráficas y tablas de frecuencias.



Revisar los siguientes países, al parecer son los que actualizan los datos más seguido: UK, USA, México, United Arab Emirates