DESCRIPCIÓN DEL CONJUNTO DE DATOS A TRABAJAR EN EL PROYECTO DETRATAMIENTO 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 DataEnlace 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 unsolo 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	$(0, \infty)$
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	$(0, \infty)$
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	$(0, \infty)$
daily_vaccinations_raw	Para una determinada entrada de datos, el número de vacunaciones para esa fecha/país.	Numérico	$(0, \infty)$
daily_vaccinations	Para una determinada entrada de datos, el número de vacunaciones para esa fecha/país.	Numérico	$(0, \infty)$
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	$(0, \infty)$
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

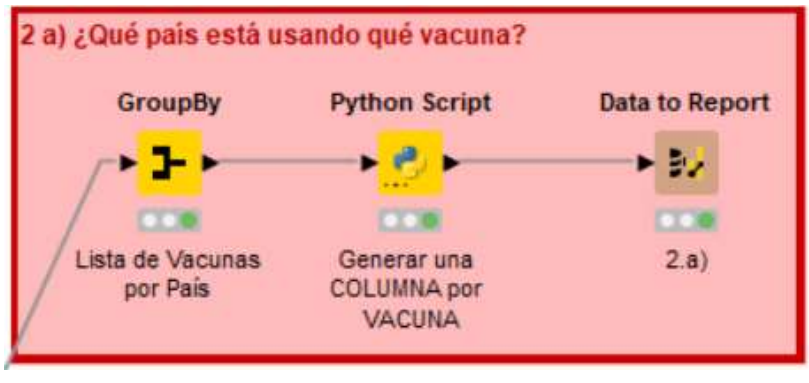
Hay un segundo archivo agregado recientemente (vacunas aplicadas de países por fabricante), con las siguientes columnas:

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/HayatVax, Johnson&Johnson.
Total_vaccinations	Número total de vacunas	Numérico	número total de vacunas/a la fecha y tipo de vacuna

2) Responder las tareas propuestas por el autor:

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

a) ¿Qué país está usando qué vacuna?



Configuración "GroupBy":

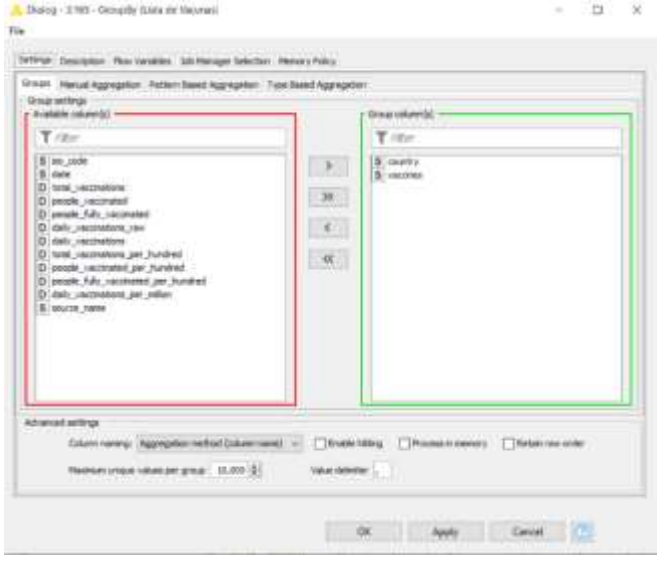
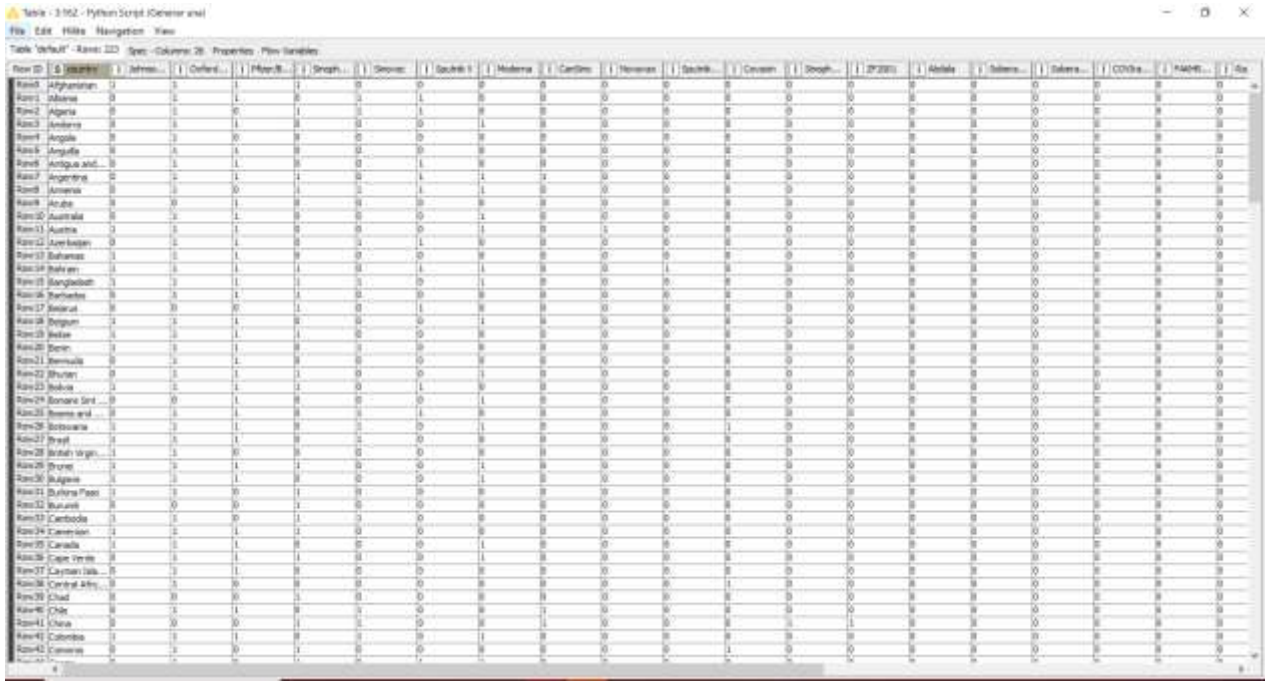


Tabla Resultante:



Código del Nodo de Python:

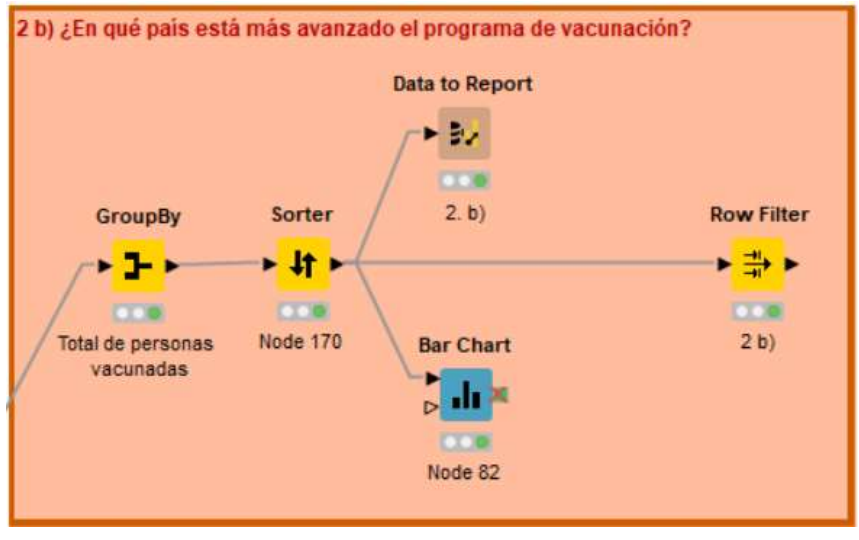
El siguiente código se encarga de generar una COLUMNA por cada una de las VACUNAS enlistadas en todos registros presentes en la columna "vaccines", sin REPETIR; y se le asigna un '1' en caso de que dicho país haya hecho uso de dicha vacuna, y un '0' en caso contrario.

```
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 == ',':
        # ------------------------------------------------------------------------
--------------
```

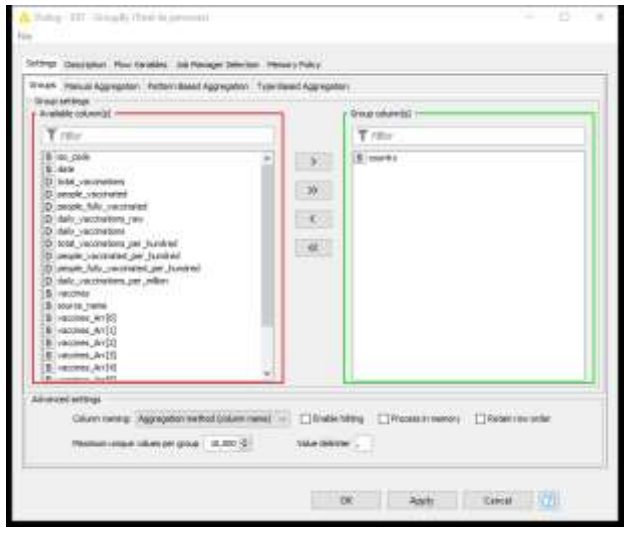
```
            # CASO: Si el nombre de la VACUNA aun no se encuentra en la LISTA
DE VACUNAS
            if not(vaccineName in vaccines):
                # Se agrega el nombre de la VACUNA a la LISTA DE VACUNAS
                vaccines.append(vaccineName)
                # Se crea una lista para almacenar 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
```

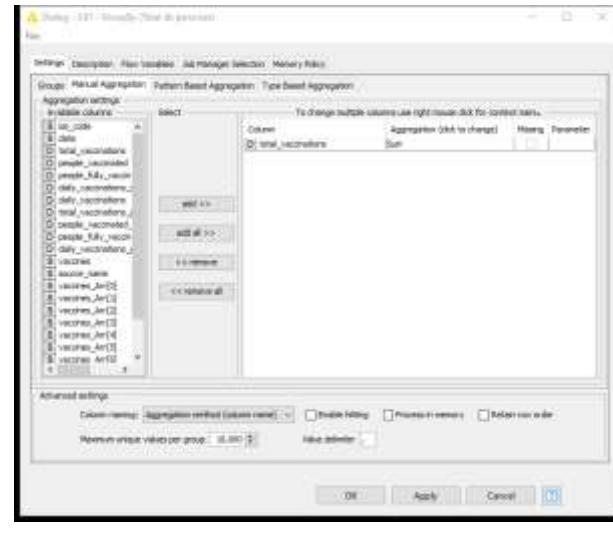
```
    vaccValues[vaccines.index(vaccineName)][indexCountry] = 1
# ------------------------------------------------------------------------
--------------

    # 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)
```

b) ¿En qué país está más avanzado el programa de vacunación?



Configuración "GroupBy":





Configuración "Row Filter":

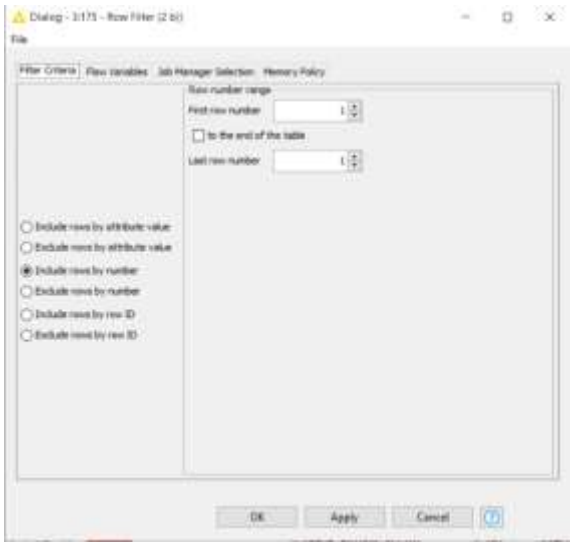
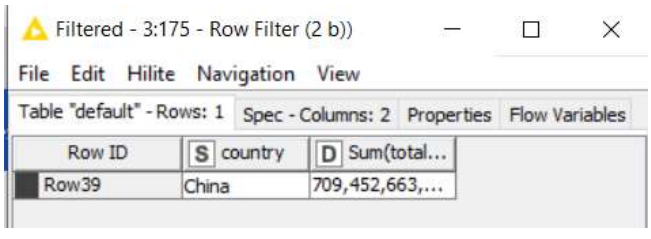
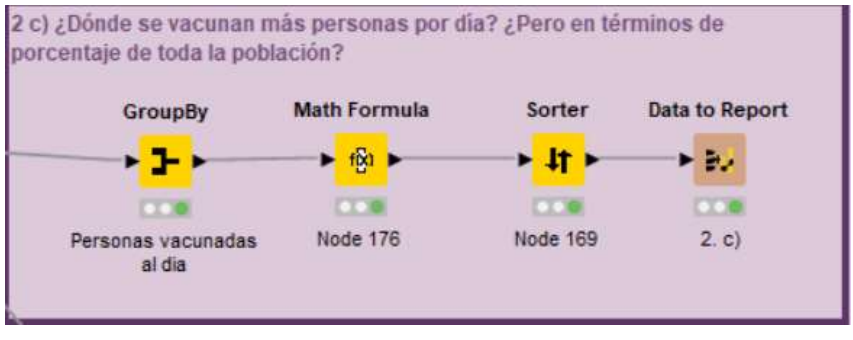


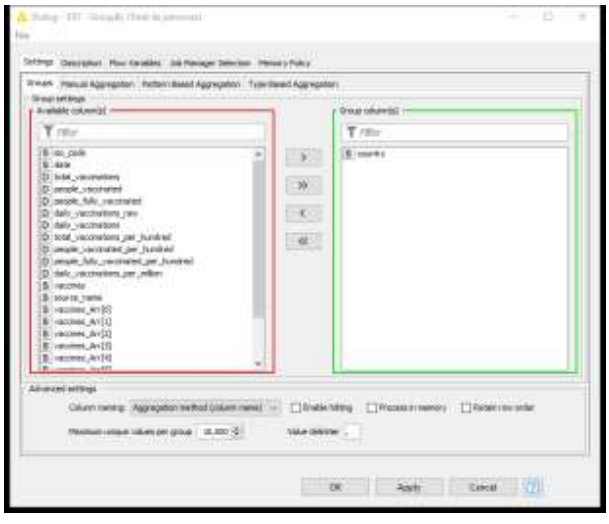
Tabla Resultante:

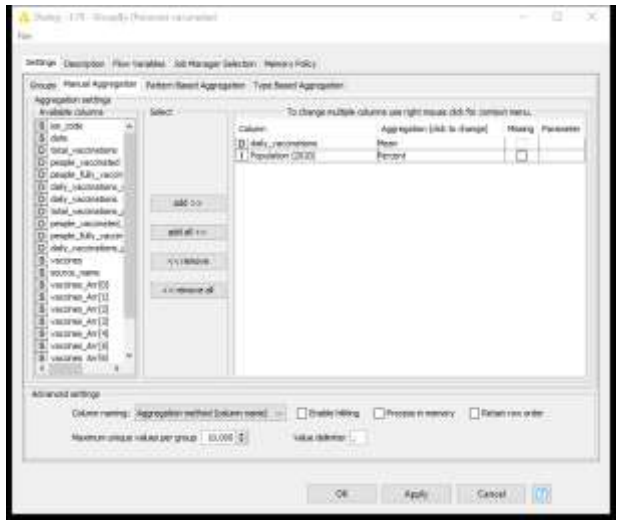


c) ¿Dónde se vacunan más personas por día? ¿Pero en términos de porcentaje de toda la población?



Configuración "GroupBy":





Configuración “Math Formula”:

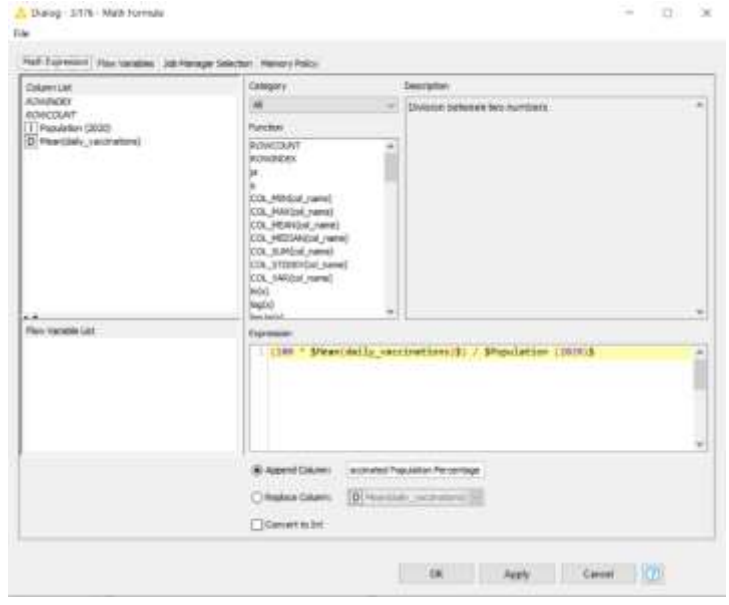


Tabla Resultante:

Sorted Table - 3:169 - Sorter

File Edit Hilite Navigation View

Table "default" - Rows: 200 Spec - Columns: 4 Properties Flow Variables

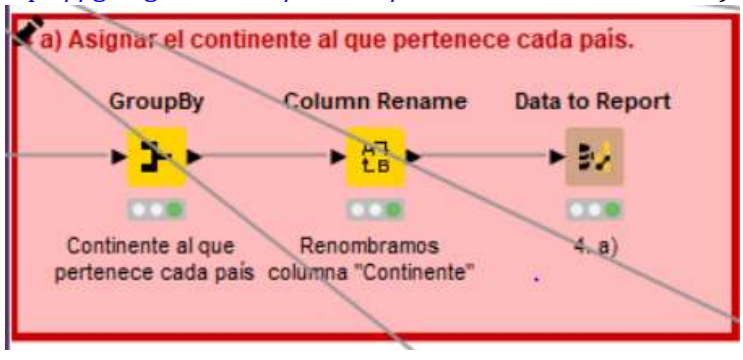
Row ID	country	Populat...	Mean(d...	Vaccina...
Row135	Niue	1628	43.8	2.69
Row60	Falkland Isla...	3497	84.8	2.425
Row22	Bhutan	773069	12,522.047	1.62
Row151	Saint Helena	6080	93.6	1.539
Row70	Gibraltar	33689	369.096	1.096
Row127	Nauru	10836	117.75	1.087
Row180	Tokelau	1360	13.667	1.005
Row46	Cuba	11325391	113,501.557	1.002
Row158	Seychelles	98453	909.598	0.924
Row43	Cook Islands	17567	147	0.837
Row154	San Marino	33944	244.561	0.72
Row186	Tuvalu	11817	84.5	0.715
Row28	Brunei	438202	2,918.057	0.666
Row114	Malta	441750	2,902.82	0.657
Row82	Iceland	341628	2,116.539	0.62
Row166	South Korea	51276977	304,251.152	0.593
Row192	Uruguay	3475842	20,316.623	0.585
Row3	Andorra	77287	447.854	0.579
Row72	Greenland	56787	328.922	0.579
Row189	United Arab ...	9910892	57,341.772	0.579
Row38	Chile	19144605	110,743.141	0.578
Row39	China	1440297825	8,304,600.911	0.577
Row88	Isle of Man	85112	490.108	0.576
Row35	Cayman Isla...	65854	378.298	0.574
Row147	Qatar	2889284	16,531.86	0.572
Row59	Faeroe Islands	48896	279.318	0.571
Row175	Taiwan	23824369	134,650.471	0.565
Row131	New Zealand	4829021	26,856.69	0.556
Row116	Mauritius	1272140	6,997.171	0.55
Row10	Australia	25550683	140,490.538	0.55
Row96	Kiribati	119760	657.786	0.549
Row32	Cambodia	16758448	91,228.116	0.544
Row181	Tonga	105901	575.524	0.543
Row146	Portugal	10191409	55,258.258	0.542
Row111	Malaysia	32436963	172,594.246	0.532
Row193	Uzbekistan	33551824	175,361.504	0.523
Row196	Vietnam	97490013	506,941.301	0.52
Row160	Singapore	5858322	30,461.307	0.52

3) Problemas iniciales:

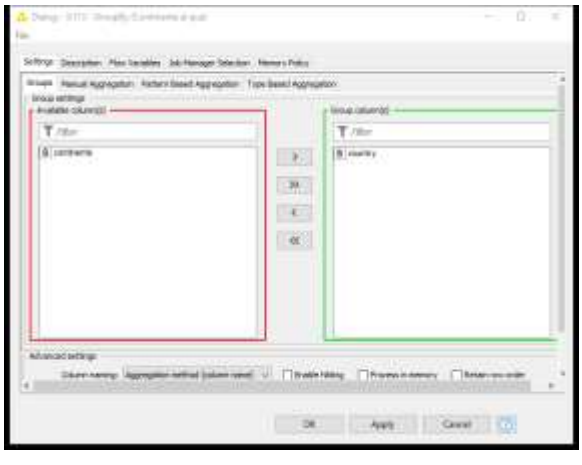
- Un problema que presenta es que los nombres de las vacunas están en forma de lista. Hay que analizar la forma de identificar la cantidad de cada tipo de vacunas que se están aplicando en total en el conjunto de datos
- Algunos países no reportan cifras de vacunación todos los días, hay que ver cuáles 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.

4) Acciones a realizar:

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



Configuración "GroupBy":



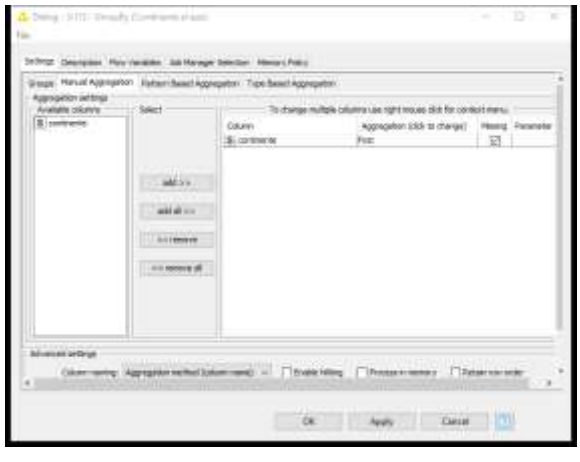
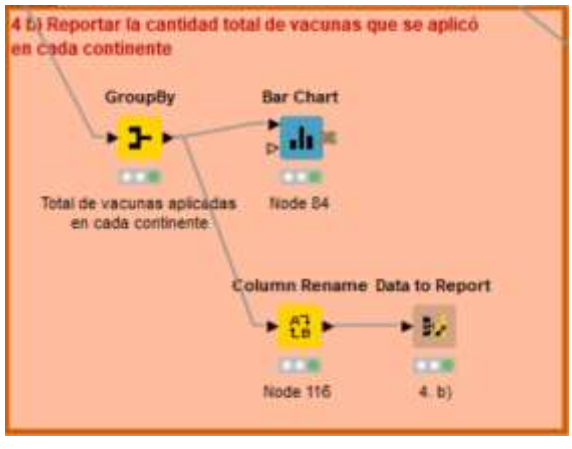


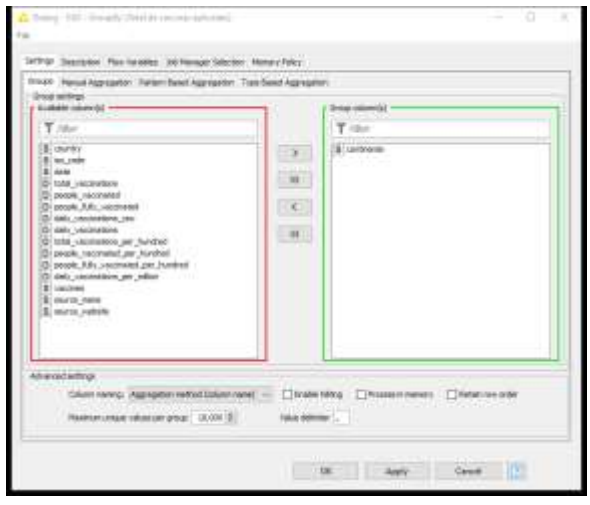
Tabla Resultante:

Row ID	country	Contine...
Row0	Afghanistan	Asia
Row1	Albania	Europa
Row2	Algeria	África
Row3	Andorra	Europa
Row4	Angola	África
Row5	Anguilla	América
Row6	Antigua and...	América
Row7	Argentina	América
Row8	Armenia	Asia
Row9	Aruba	América
Row10	Australia	Australia y ...
Row11	Austria	Europa
Row12	Azerbaijan	Asia
Row13	Bahamas	América
Row14	Bahrain	Asia
Row15	Bangladesh	Asia
Row16	Barbados	América
Row17	Belarus	Europa
Row18	Belgium	Europa
Row19	Belize	América
Row20	Benin	África
Row21	Bermuda	América
Row22	Bhutan	Asia
Row23	Bolivia	América
Row24	Bonaire Sint ...	?
Row25	Bosnia and ...	Europa
Row26	Botswana	África
Row27	Brazil	América
Row28	British Virgin...	?
Row29	Brunei	Asia
Row30	Bulgaria	Europa
Row31	Burkina Faso	África
Row32	Burundi	África
Row33	Cambodia	Asia
Row34	Cameroon	África
Row35	Canada	América
Row36	Cape Verde	África
Row37	Cayman Isla...	América
Row38	Central Afric...	África
Row39	Chad	África

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



Configuración "GroupBy":



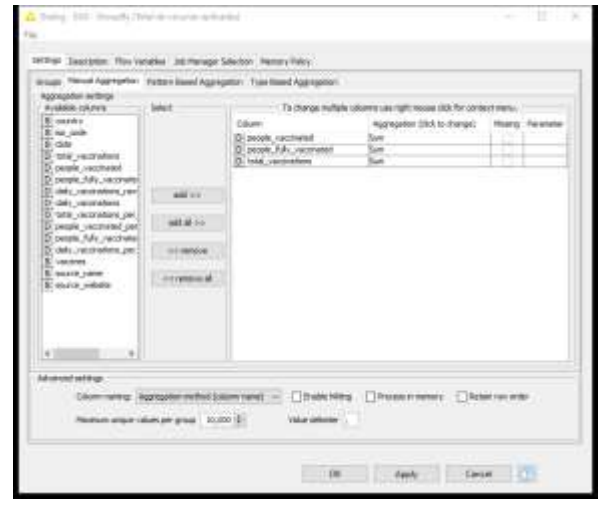


Tabla Resultante:

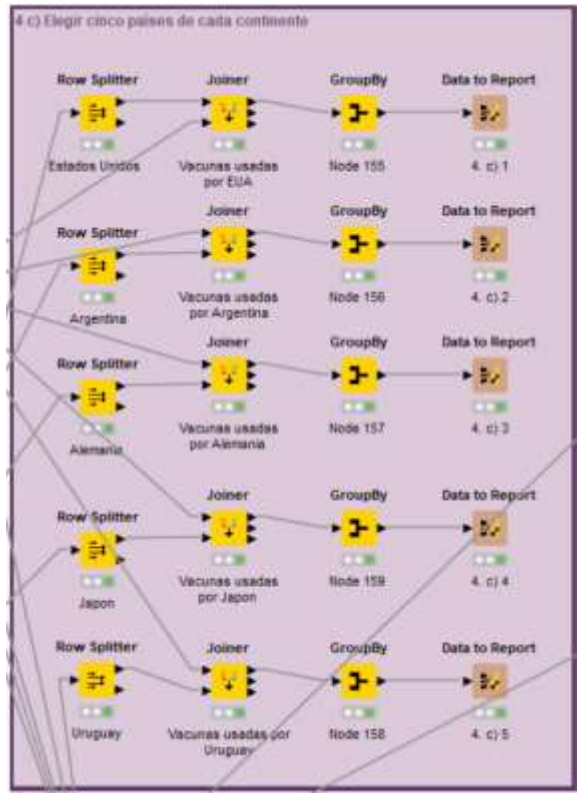
Renamed/Retyped table - 3:116 - Column Rename

File Edit Hilite Navigation View

Table "default" - Rows: 5 | Spec - Columns: 4 | Properties | Flow Variables

Row ID	S contine...	D Personas vacunadas	D Personas vacunadas por completo	D Total de vacunaciones
Row0	América	188,063,607,586	147,176,869,188	357,980,137,059
Row1	Asia	386,576,359,314	272,116,535,146	1,332,307,693,985
Row2	Australia y ...	5,896,306,250	4,752,053,756	11,657,727,968
Row3	Europa	118,791,955,146	101,719,075,633	241,907,394,702
Row4	África	13,387,730,741	8,656,696,599	21,825,053,998

c) 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)



Configuración "GroupBy":

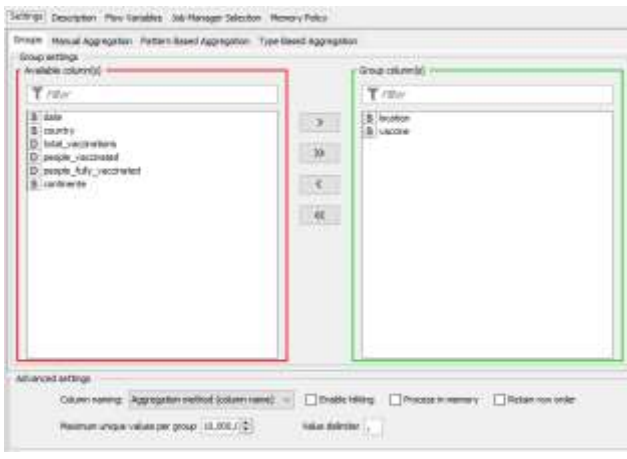


Tabla Resultante:

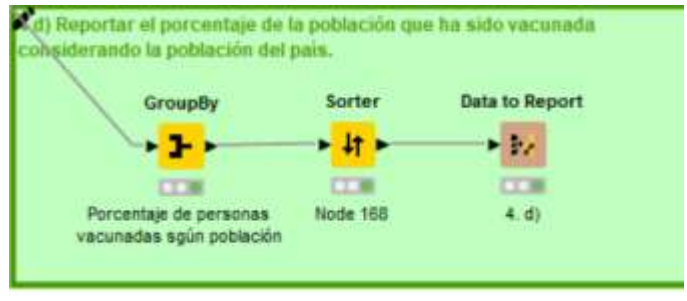
Group table - 3:156 - GroupBy

File Edit Hilite Navigation View

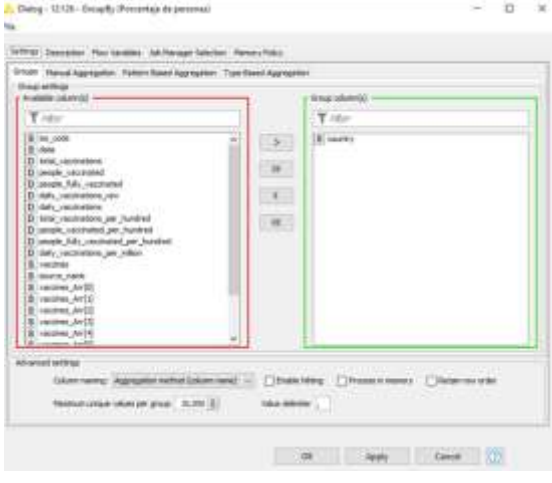
Table "default" - Rows: 6 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	location	vaccine	Sum(tot...
Row0	Argentina	CanSino	18,670,327,...
Row1	Argentina	Moderna	18,678,951,...
Row2	Argentina	Oxford/Astr...	18,678,951,...
Row3	Argentina	Pfizer/BioNT...	18,640,324,...
Row4	Argentina	Sinopharm/B...	18,678,951,...
Row5	Argentina	Sputnik V	18,678,951,...

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



Configuración "GroupBy":



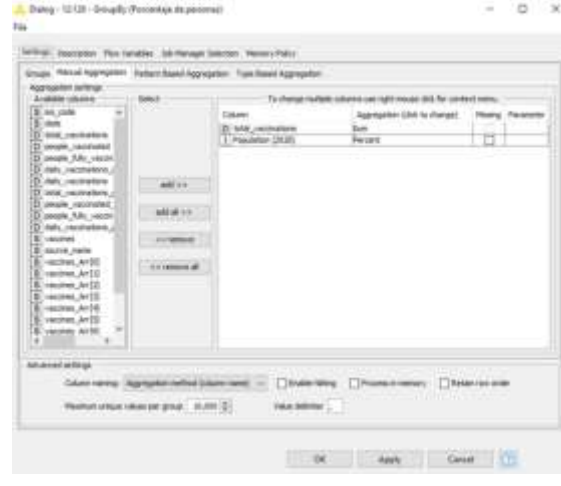


Tabla Resultante:

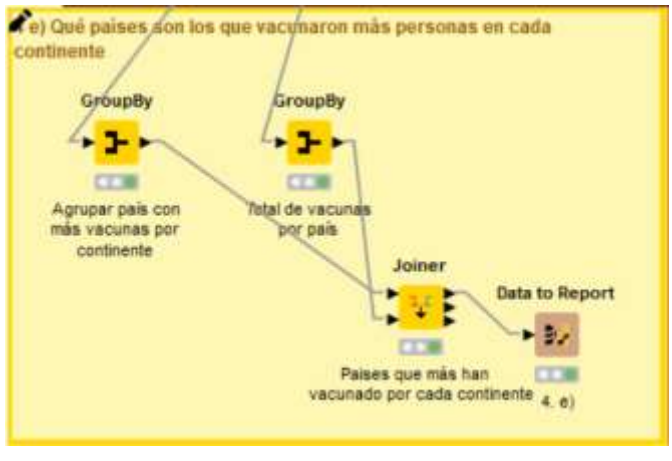
Group table - 12:126 - GroupBy (Porcentaje de personas)

File Edit Hilite Navigation View

Table "default" - Rows: 200 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	S country	D Sum(to...	D Percent...
Row0	Afghanistan	118,041,426	0.113
Row1	Albania	295,540,465	0.618
Row2	Algeria	238,946,062	0.058
Row3	Andorra	2,447,643	0.105
Row4	Angola	388,740,243	0.158
Row5	Anguilla	806,657	0.12
Row6	Antigua and...	6,840,405	0.223
Row7	Argentina	18,678,951,...	1.14
Row8	Armenia	34,983,105	0.108
Row9	Aruba	38,528,732	0.688
Row10	Australia	9,733,094,247	0.993
Row11	Austria	596,758,436	0.16
Row12	Azerbaijan	2,364,582,887	0.838
Row13	Bahamas	8,527,171	0.125
Row14	Bahrain	778,833,949	0.913
Row15	Bangladesh	13,205,078,...	0.608
Row16	Barbados	76,587,881	0.938
Row17	Belarus	198,340,415	0.105
Row18	Belgium	5,762,415,299	1.135
Row19	Belize	21,555,588	0.305
Row20	Benin	24,937,864	0.07
Row21	Bermuda	4,511,282	0.153
Row22	Bhutan	77,837,458	0.268
Row23	Bolivia	2,228,459,475	1.02
Row24	Bosnia and ...	13,550,064	0.05

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



Configuración "Joiner":

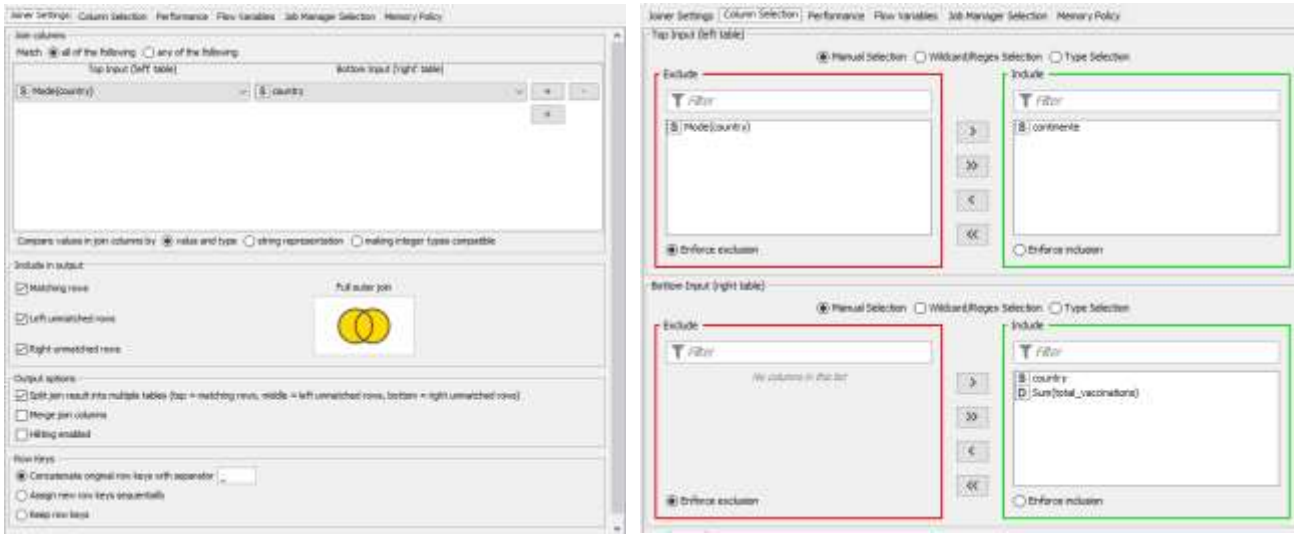
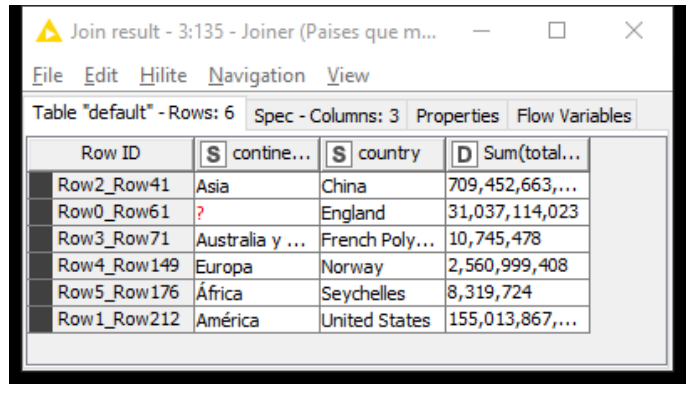
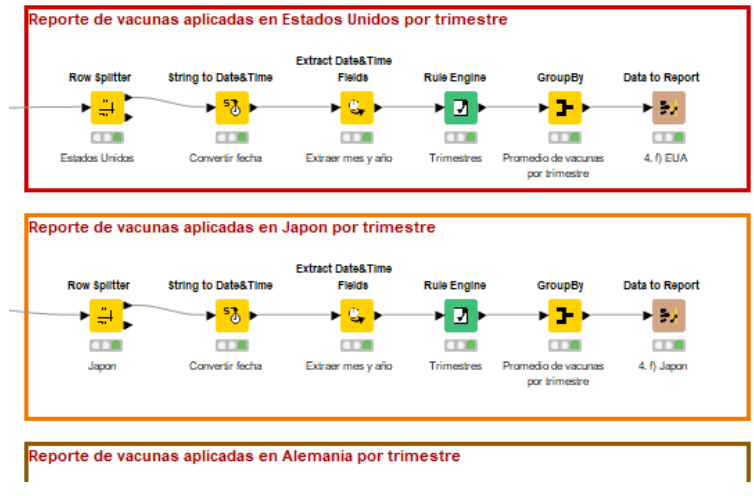


Tabla Resultante:

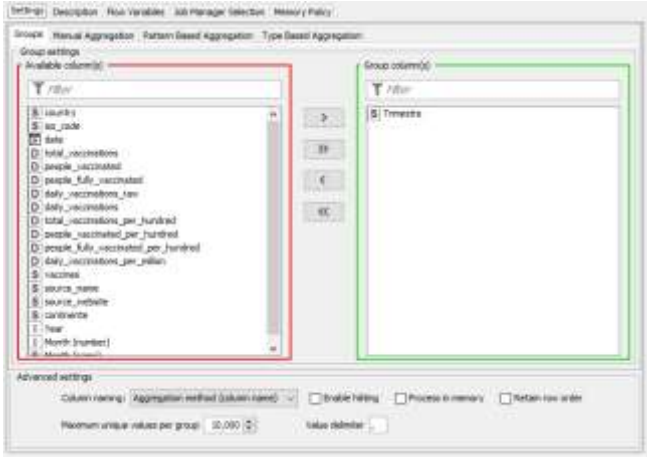


Row ID	S contine...	S country	D Sum(total...
Row2_Row41	Asia	China	709,452,663,...
Row0_Row61	?	England	31,037,114,023
Row3_Row71	Australia y ...	French Poly...	10,745,478
Row4_Row149	Europa	Norway	2,560,999,408
Row5_Row176	África	Seychelles	8,319,724
Row1_Row212	América	United States	155,013,867,...

f) 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



Configuración "GroupBy":



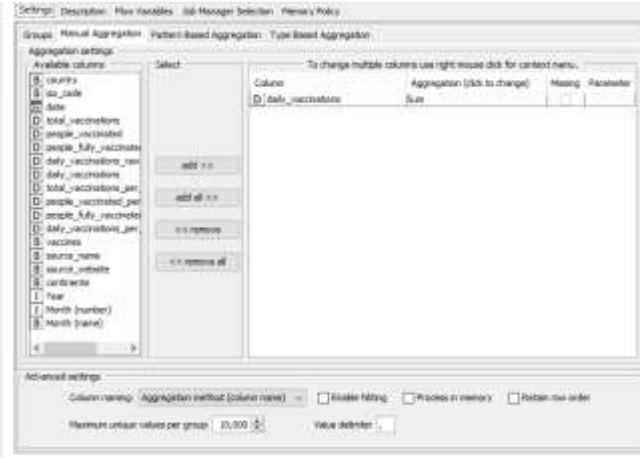


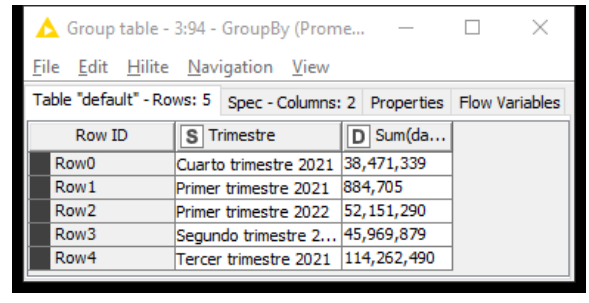
Tabla Resultante:

Group table - 3:39 - GroupBy (Promedi...

File Edit Hilite Navigation View

Table "default" - Rows: 6 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	Trimestre	Sum(daily_va...
Row0	Cuarto trimestre 2020	4,136,753
Row1	Cuarto trimestre 2021	115,343,057
Row2	Primer trimestre 2021	158,191,761
Row3	Primer trimestre 2022	46,897,696
Row4	Segundo trimestre 2...	175,868,920
Row5	Tercer trimestre 2021	61,673,419



Group table - 3:94 - GroupBy (Prome...

File Edit Hilite Navigation View

Table "default" - Rows: 5 | Spec - Columns: 2 | Properties | Flow Variables

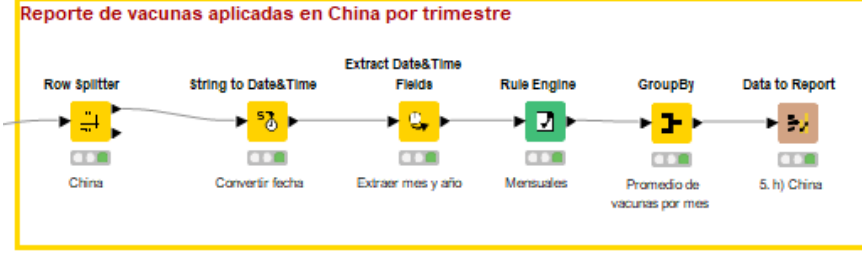
Row ID	Trimestre	Sum(da...
Row0	Cuarto trimestre 2021	38,471,339
Row1	Primer trimestre 2021	884,705
Row2	Primer trimestre 2022	52,151,290
Row3	Segundo trimestre 2...	45,969,879
Row4	Tercer trimestre 2021	114,262,490

g) 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.

5) Elegir un país y realice lo siguiente:

h) 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":

i) 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":

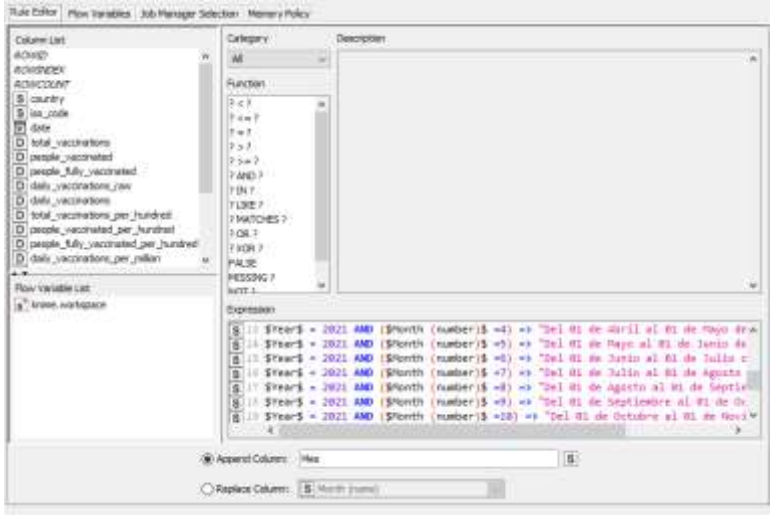


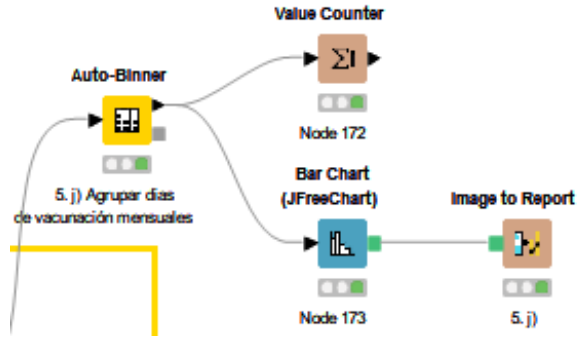
Tabla Resultante:

Table View - 3:153 - Data to Report (5. h) China)

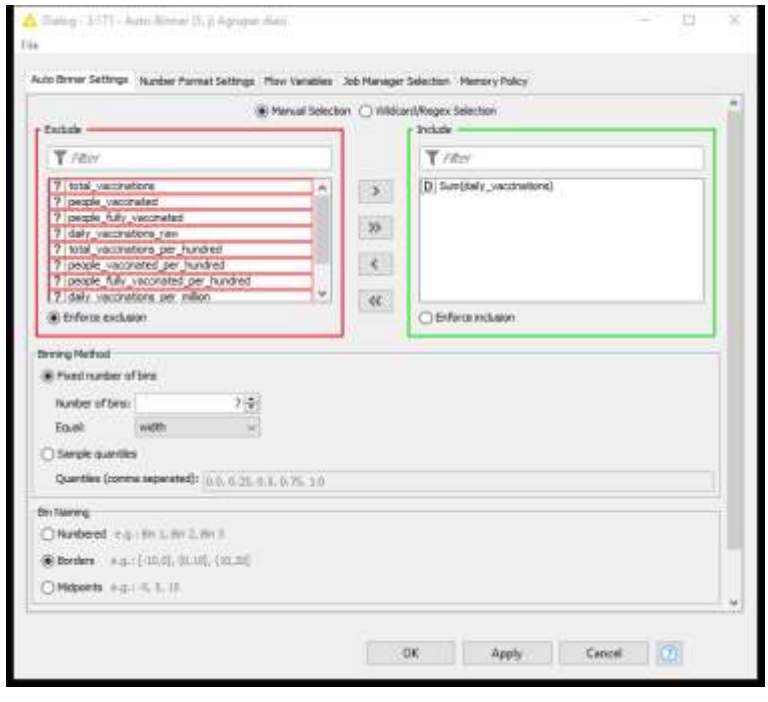
File Edit Hilite Navigation View

Row ID	S Mes	D Sum(daily_va...
Row0	Del 01 Diciembre del 2020 al 01 de Enero del 2021	3,000,000
Row1	Del 01 de Abril al 01 de Mayo del 2021	132,826,860
Row2	Del 01 de Agosto al 01 de Septiembre del 2021	428,975,427
Row3	Del 01 de Diciembre del 2021 al 01 de Enero del 2022	313,936,999
Row4	Del 01 de Enero al 01 de Febrero del 2021	18,794,325
Row5	Del 01 de Enero al 01 de Febrero del 2022	195,840,001
Row6	Del 01 de Febrero al 01 de Marzo del 2021	27,893,435
Row7	Del 01 de Febrero al 01 de Marzo del 2022	121,300,713
Row8	Del 01 de Julio al 01 de Agosto del 2021	416,799,142
Row9	Del 01 de Junio al 01 de Julio del 2021	581,455,146
Row10	Del 01 de Marzo al 01 de Abril del 2021	55,563,168
Row11	Del 01 de Marzo al 01 de Abril del 2022	137,842,571
Row12	Del 01 de Mayo al 01 de Junio del 2021	364,184,997
Row13	Del 01 de Noviembre al 01 de Diciembre del 2021	222,856,428
Row14	Del 01 de Octubre al 01 de Noviembre del 2021	56,364,715
Row15	Del 01 de Septiembre al 01 de Octubre del 2021	172,708,569

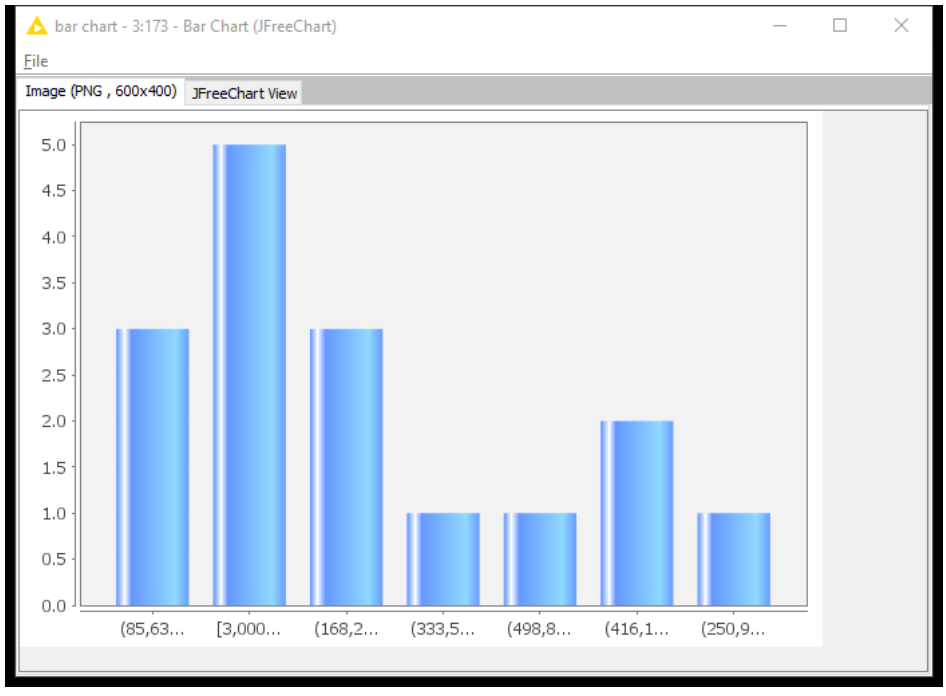
j) Crear binners de las vacunas aplicadas por día.



Configuración "Auto-Binner":



k) 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