Carrera: Data Analíticos

Módulo 4

farmacéutica BIOGENESYS

Nombre del autor: Daniel Stivens Suarez Cabrejo

Email: dssuarezc@hotmail.es

Cohorte: da-pt08

Fecha de entrega: 25/08/25

Institución:

Introducción

El principal propósito del proyecto es según el análisis, los gráficos y los insights descubiertos, encontrar las diferentes áreas prioritarias que necesiten de nuestros servicios para tener la posibilidad de expandirnos con el objetivo de bajar los casos confirmados y aumentar el número de casos recuperados de los países más afectados por esta pandemia.

Objetivos específicos:

- Aplicar técnicas de carga, filtrado, limpieza y transformación de datos utilizando Python.
- Aplicar herramientas avanzadas de análisis estadístico y visualización, tales como histogramas, gráficos de barras y mapas de calor para realizar análisis exploratorio.
- Emplear técnicas avanzadas de pandas y numpy, explorando series temporales, identificando tendencias a largo plazo como pueden ser patrones estacionales.
- Por ultimo entender como exportar los datos a la herramienta power bi, donde se diseñaran dashboard interactivos, con el fin de que el cliente pueda comprender mejor la información y pueda interactuar con ella, consiguiendo una

experiencia agradable al usuario y poder mostrar los insgiths más valiosos de la problemática planteada.

Desarrollo del proyecto

En primera medida se obtuvieron los datos del dataset, verificando que se hayan importado todos los datos comparando el número de filas y columnas.

```
#mostramos las dimenciones del dataframe generado con la funcion shape
print(df_data_latinoamerica.shape)
```

```
(12216057, 50)
```

Seleccionamos solo los países que se requiere que realicemos el estudio, para esta ocasión seria Brasil, México, Colombia, Argentina, Perú y Chile. Con esta instrucción se redujeron las filas.

```
#se imprimen las dimenciones del dataframe para verificar
print(df_datos_filtrados.shape)
```

```
(11970289, 50)
```

Se filtran los datos mayores a la fecha 2021-01-01, en esta instrucción se reducen bastantes filas y esto nos ayudara bastante a la hora de generar los gráficos.

```
# mas adelante ayudando a la optimización de
print(df_datos_filtrados.shape)
```

```
(7537296, 50)
```

Verificando el dataset nos damos cuenta que tenemos todavía hay bastantes filas y esto se debe a que por cada país hay datos más específicos por subregión pero estos datos se resumen al principio de cada país es decir si dejamos estas filas estaríamos repitiendo información y los cálculos quedarían mal.

```
#verificamos las dimenciones del nuevo dataset
df_datos_filtrados.shape
```

```
(3744, 49)
```

Adicional eliminamos la columna location_key ya que como solo vamos a colocar la información general por país, esta columna se repetiría con la columna country_code y seria duplicar información, por esta razón quedan 49 columnas.

Hay bastantes campos vacíos y esto puede afectar al análisis más adelante, por esta razón cambiamos estos valores con la mediana de dicha columna para no afectar los cálculos posteriormente.

Hay un caso en particular donde en toda la columna hay campos vacíos que son en los países Perú, Chile, Argentina y México, para estos casos en particular no se puede colocar la mediana porque no existe, entonces se reemplazarían estos valores vacíos por cero.

Por ultimo cambiamos los tipos de datos donde encontramos datos tipo fecha, texto, decimal y entero.

```
dtypes: Int64(27), datetime64[ns](1), float64(19), string(2)
```

Conclusiones de la limpieza de datos

La limpieza y transformación de datos es importante en cualquier analista de datos porque es el punto de partida de que los cálculos, visualizaciones e insights sean los correctos, una mala limpieza de datos o colocar el tipo de dato incorrecto puede cambiar la información y mostrar datos que no corresponden con la brindada inicialmente.

Consignas 11 y 12 de DAM4L5

11 ¿Qué implican estas métricas y cómo pueden ayudar en el análisis de datos?

Estad medidas nos ayudan a saber que tan dispersos están los datos y si tienen sesgo a la izquierda o la derecha, también nos puede ayudar a cuál es el dato que más se repite o más común como es la moda y nos sirve para comparar en las diferentes columnas y poder generar conclusiones e insgiths valiosos, según el promedio podemos verificar si hay datos atípicos o que no corresponde al resto de grupo de datos.

La mediana es un valor que nos puede ayudar a identificar valores atípicos tanto muy altos como muy bajos.

¿Se muestran todas las estadísticas en todas las columnas durante el análisis?

no se muestran todas las estadísticas descriptivas de todas las columnas ya que hay columnas que no son numéricas como pueden ser la fecha, el código del país, el nombre, etc. Estas estadísticas se pueden visualizar mejor con la extensión jupiter donde al ver el dataset como tabla al pararnos en cada columna al lado izquierdo aparecen estas estadísticas.

Con ayuda del for podemos recorrer todas las columnas numéricas y nos dan una idea de que datos pueden haber en estas columnas verificando por ejemplo el mínimo, el máximo, el promedio, la varianza, etc.

¿Cuál es la razón de la respuesta anterior y cómo podría afectar la interpretación de los resultados obtenidos?

Porque hay columnas que no son numéricas y los datos mínimos que se pueden sacar puede ser la frecuencia con la que se repite un dato.

La exclusión de estas columnas Puede llevar a una comprensión incompleta o sesgada de los datos si no se consideran estas columnas.

12 ¿Qué representa la mediana?

Este representa el valor central de los datos ordenados de menor a mayor, este valor no es sensible con datos atipicos.

¿Cómo varía la dispersión de los datos en el conjunto de datos analizado, en términos de la varianza y el rango?

La varianza indica que tan dispersos están los datos con respecto al promedio, una varianza alta significa que los valores son muy diferentes entre sí mientras que una varianza baja indica que los valores son cercanos entre si y muy cerca a la media

En el tema del rango signica que tanto varian los datos según l intervalo del rango, parecido a la varianza entre mayor sea el rango mayo es la dispersión de los datos y viceversa entre más bajo el rango menos variabilidad hay en los datos.

Según los datos analizados en la mayoría de las columnas la desviación estándar es alta significa que los datos varían mucho individualmente, los datos se alejan significativamente del promedio donde es más difícil sacar interpretaciones o tendencias.

¿Qué nos puede indicar esto sobre la consistencia o la variabilidad de los datos en relación con la mediana?

Son medidas totalmente diferentes las cuales no se relacionan entre si es decir no depende una de la otra y ninguna es susceptible entre si

EDA e insights

En primera medida se puede evidenciar que en los únicos países que hubieron casos recuperados fueron en los países de Brasil y Colombia según el grafico 2.

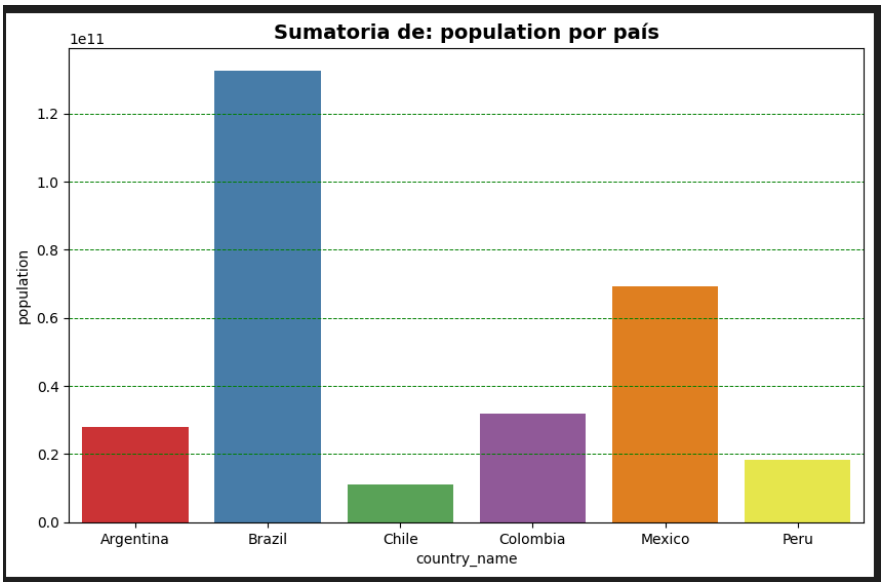

Fig 1.

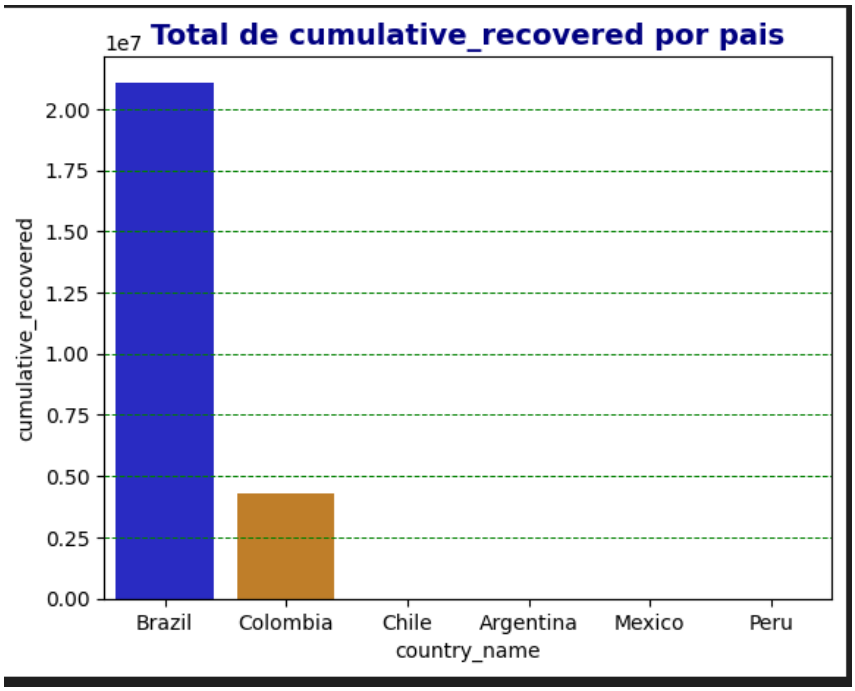

Fig 2.

Brasil tiene el doble de población con respecto a los demás paises como se puede evidenciar en el grafico 1. Por ende las estadisticas van hacer mucho mayores que los demás paises. En algunos gráficos se normalizaron los datos mas que todo en los de mortalidad para comparar mejor entre paises.

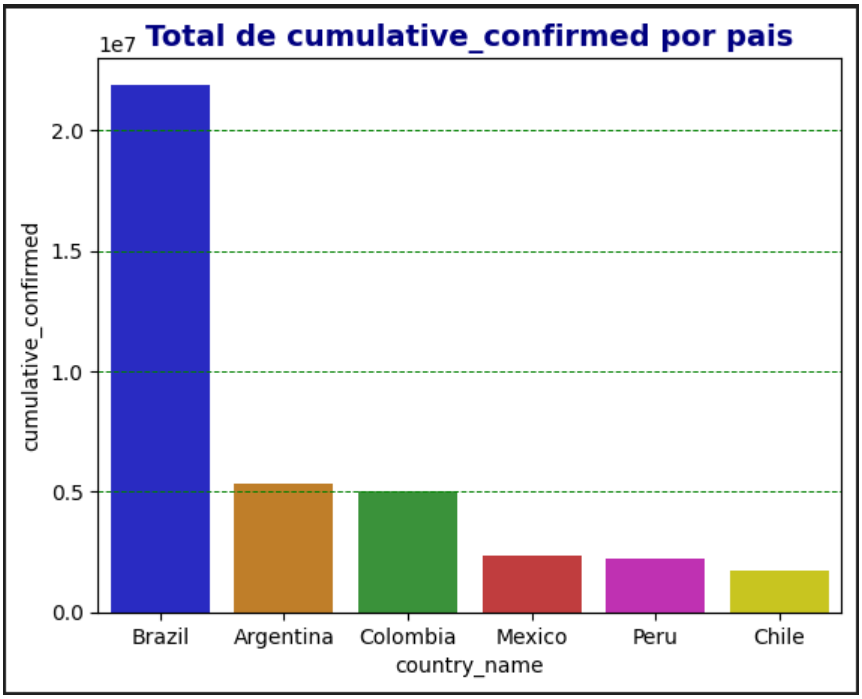

Fig 3.

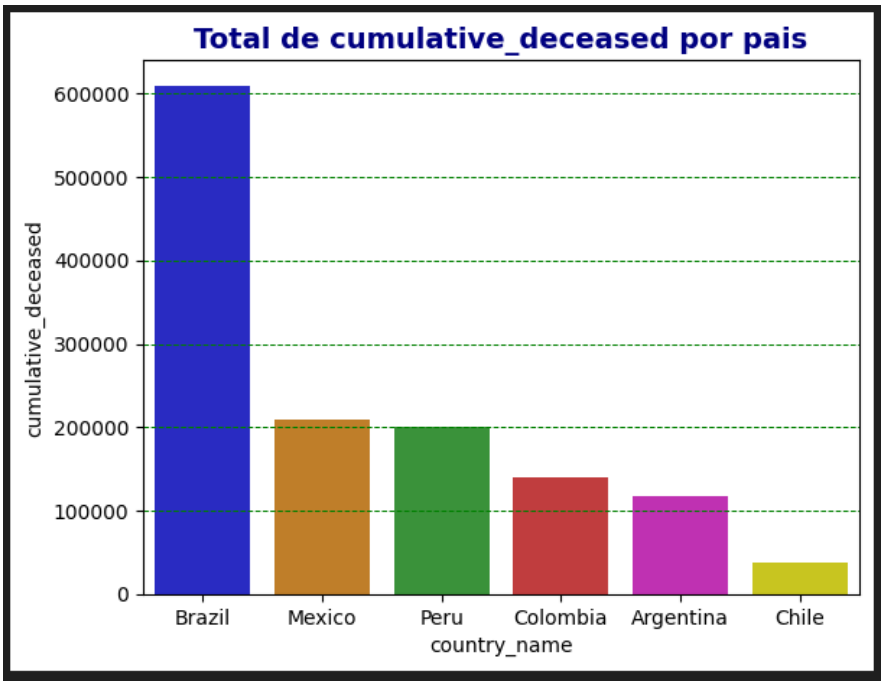

Fig 4.

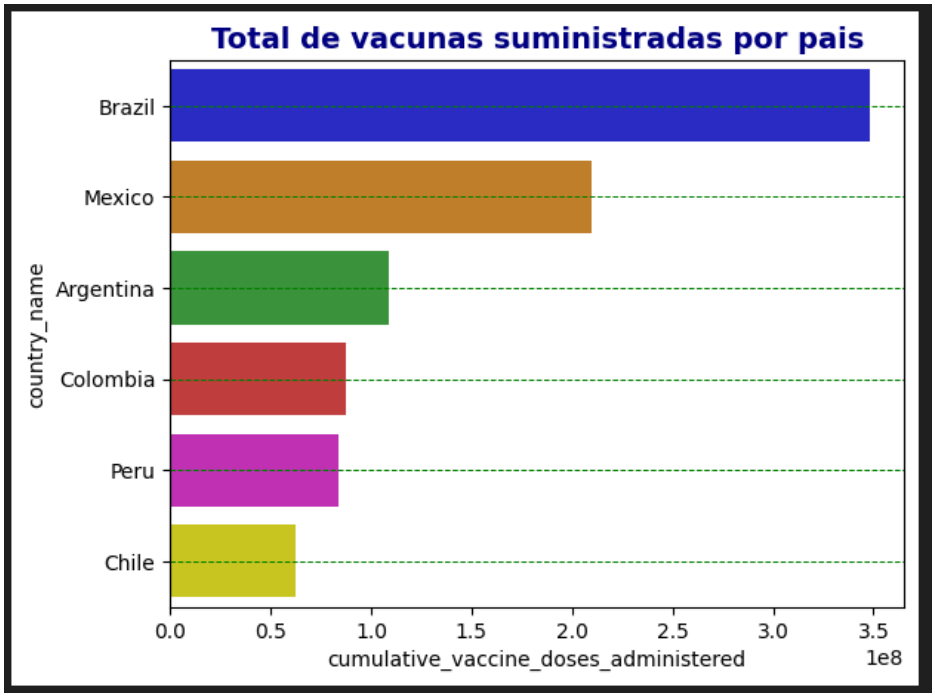

Fig 5.

En la figura 3 se puede visualizar que en argentina y Colombia hubieron más casos confirmados que México y Perú, pero al visualizar la figura 4 nos podemos dar cuenta que México y Perú tuvieron más personas fallecidas, se podría deducir que en estos países les falto suministrar más vacunas o intervinieron otros factores que más adelante confirmaremos.

Por ejemplo en la figura 5 podemos visualizar que Perú fue el penúltimo país que recibió menos vacunas y la consecuencia de esto fueron personas fallecidas.

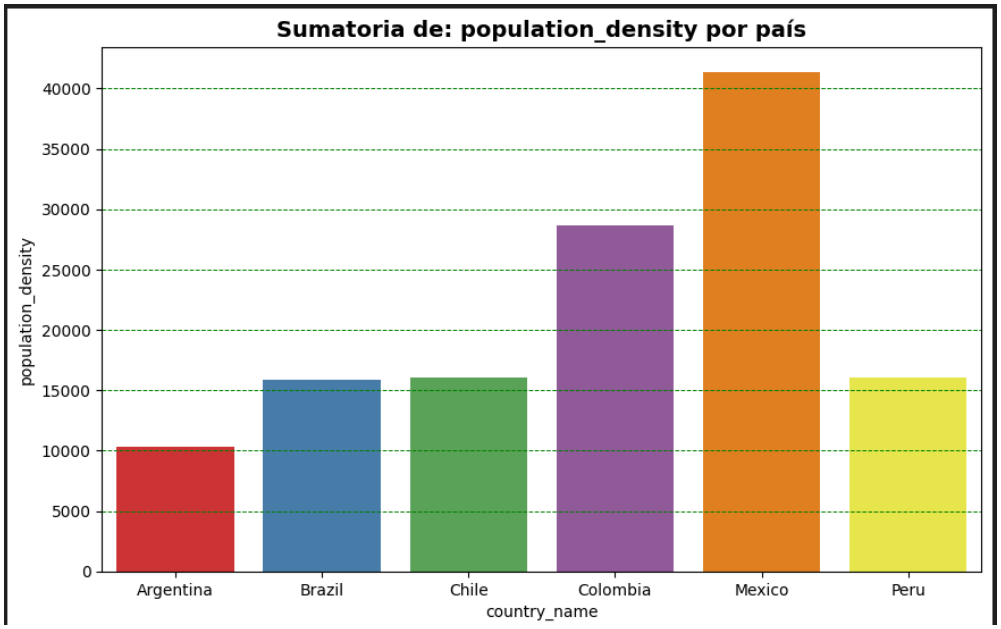

Fig 6.

La figura 6 nos puede dar la razón de porque México hubieron más personas fallecidas, ya que a comparación de los otros países, tiene mayor densidad poblacional es decir hay mayor población por km2, a pesar de ser el segundo país con mas vacunas suministradas no significa que van haber mas casos recuperados como se puede observar en la figura 2.

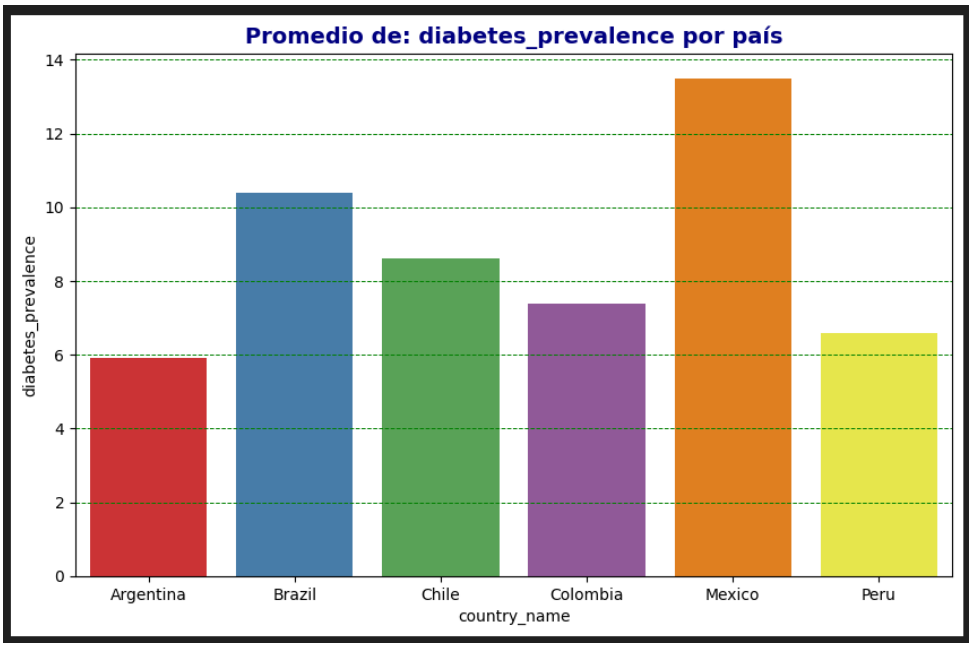

Fig 7.

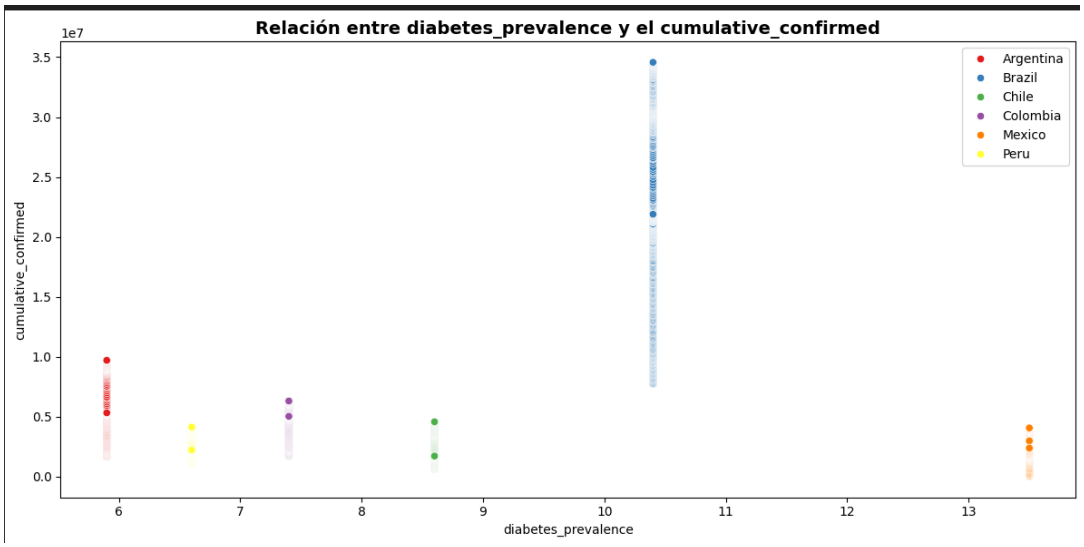

Fig 8.

las personas diabéticas son más susceptibles a contagiarse y posteriormente fallecer por COVID ya que México y Brasil son los paises que lideran esta estadistica según la fig 7 y 8

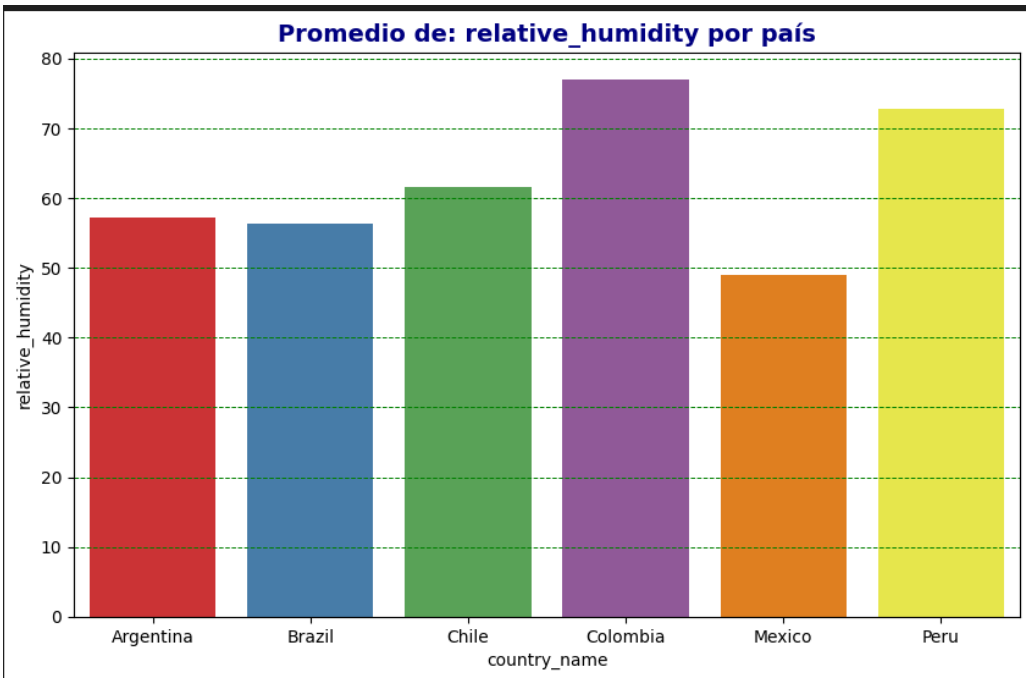

Fig 9.

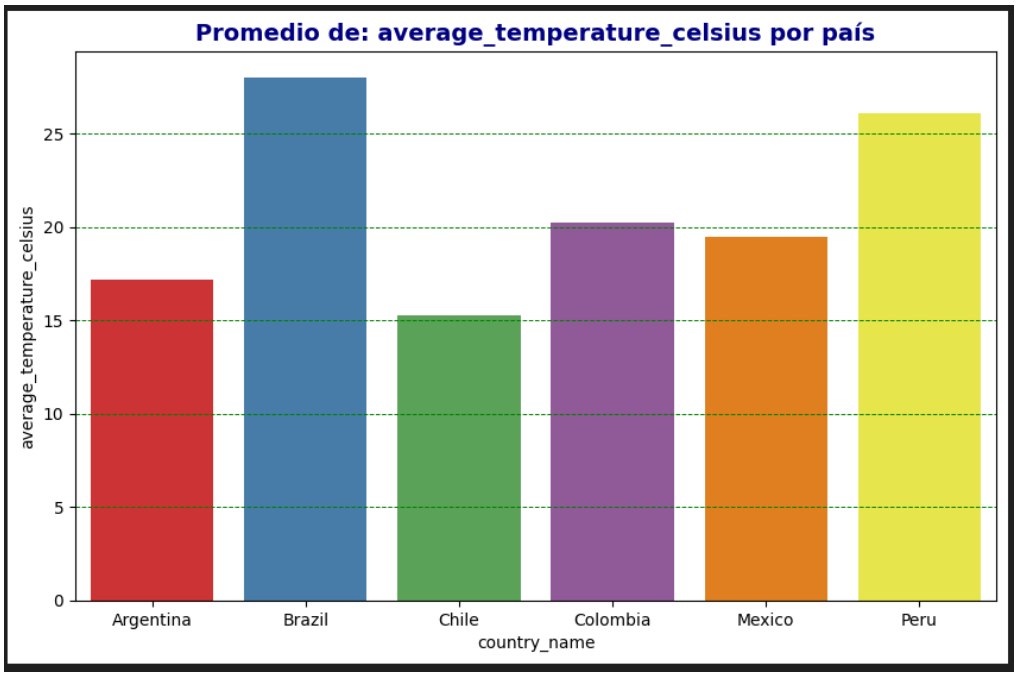

Fig 10..

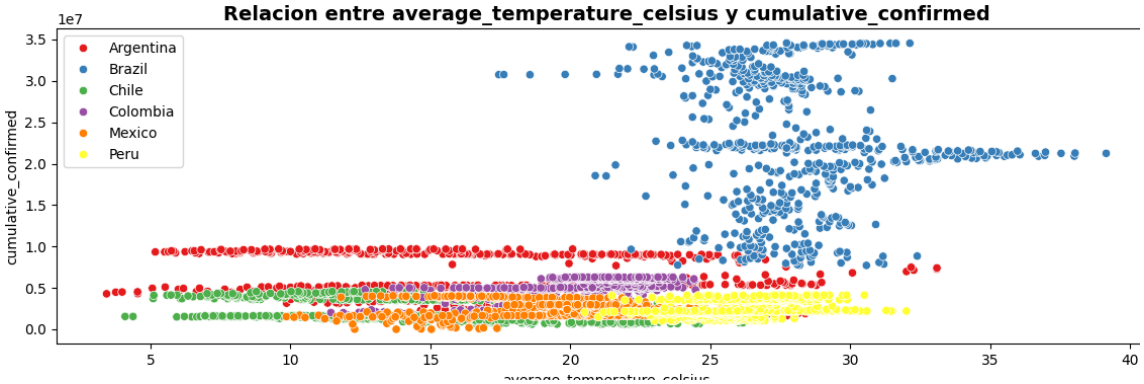

Fig 11.

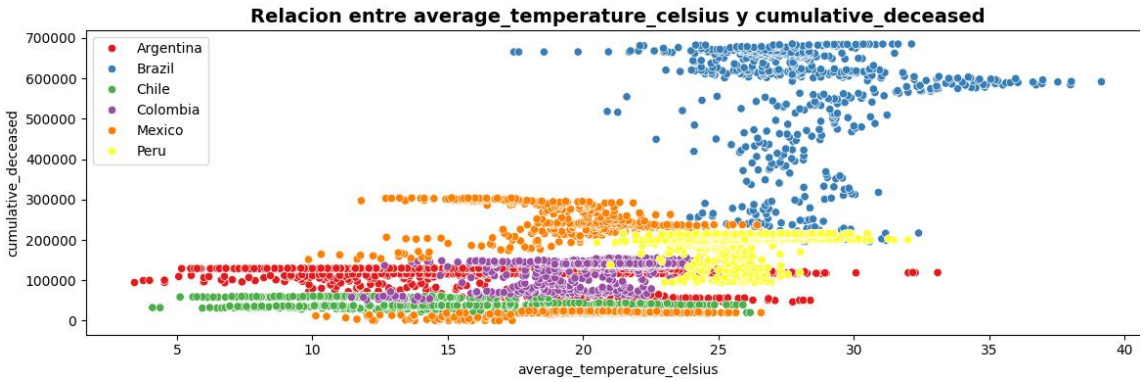

Fig 12.

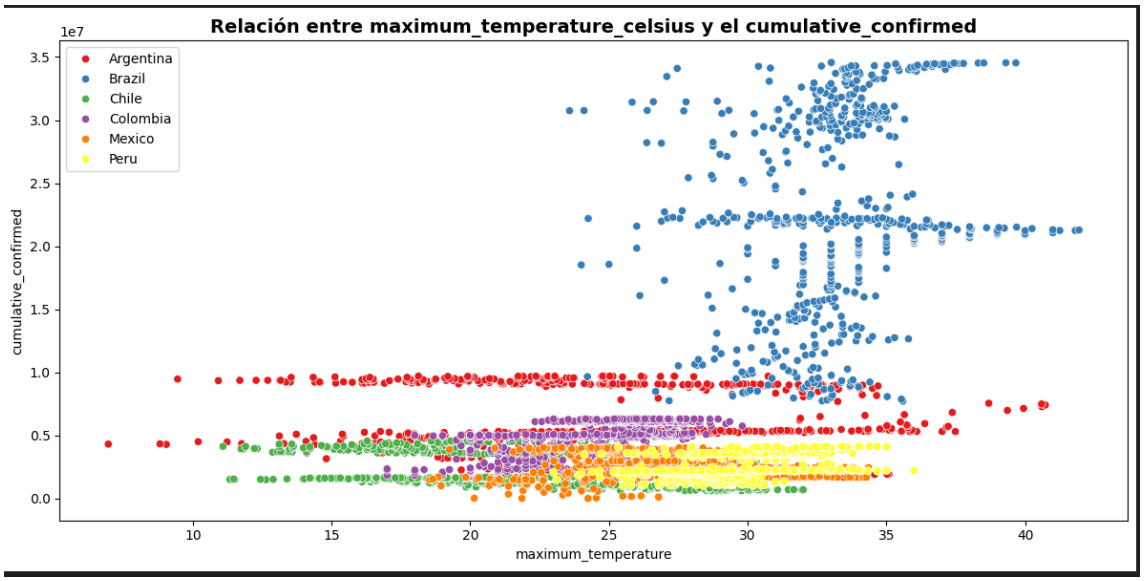

Fig 13.

Las figuras 9 a la 13 están relacionados al clima para verificar si el clima es un factor determinante para que hayan más contagios. Por ejemplo en la figura 9 Perú y Colombia son los países que tienen mayor humedad relativa a comparación de los demás y en la figura 10 los países con mayor temperatura promedio son Brasil y Perú.

En las figuras 11 y 12 nos podemos dar cuenta que México y Perú sobresalen en las personas fallecidas con respecto a los demás países. En la figura 13 se podría decir que la temperatura es un factor determinante para que hayan casos confirmados pero esta hipotesis es incorrecta porque como Brasil es el país con mas casos confirmados y fallecidos pero tenemos que tener en cuenta que por lo general brasil tiene clima cálido. si veirifcamos los otros paises nos damos cuenta que los casos confirmados y fallecidos están distribuidos por igual a lo largo del eje x.

Para estas épocas del año se podrían implementar medidas más rigurosas, más jornadas de vacunación, restricción de movilidad, mayor personal médico calificado.

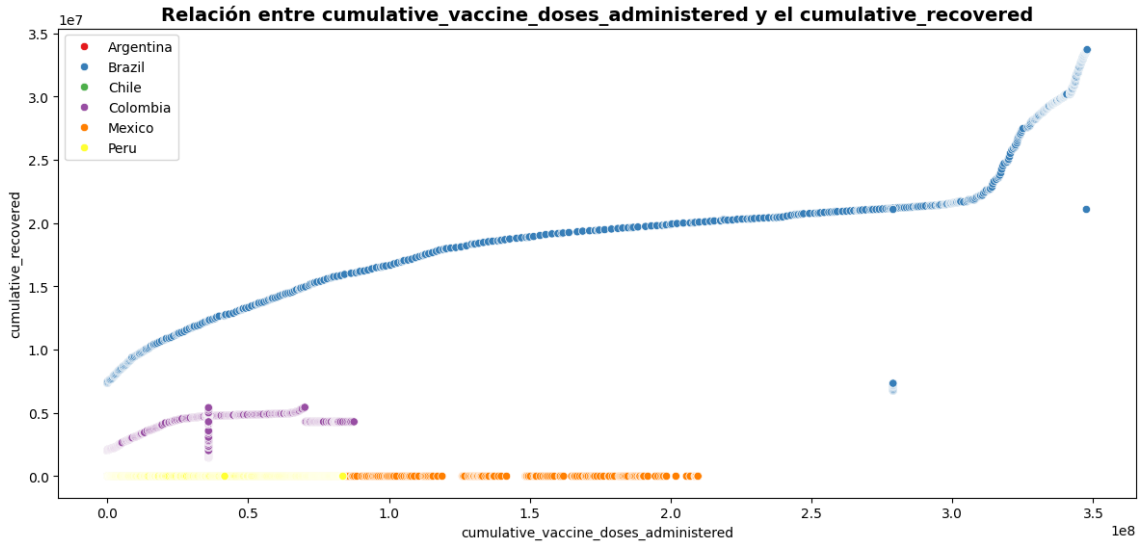

Fig 14.

En la figura 14 podemos evidenciar que las vacunas obtuvieron más casos recuperados en Brasil y Colombia mientras que en los otros países no surtió mucho efecto.

la vacunación fue más efectiva para reducir mortalidad y prevenir contagios graves, en lugar de reflejarse en la métrica de recuperados. (Chile, Perú, México, argentina) la vacunación tuvo más impacto en reducir mortalidad que en aumentar el conteo de recuperados.

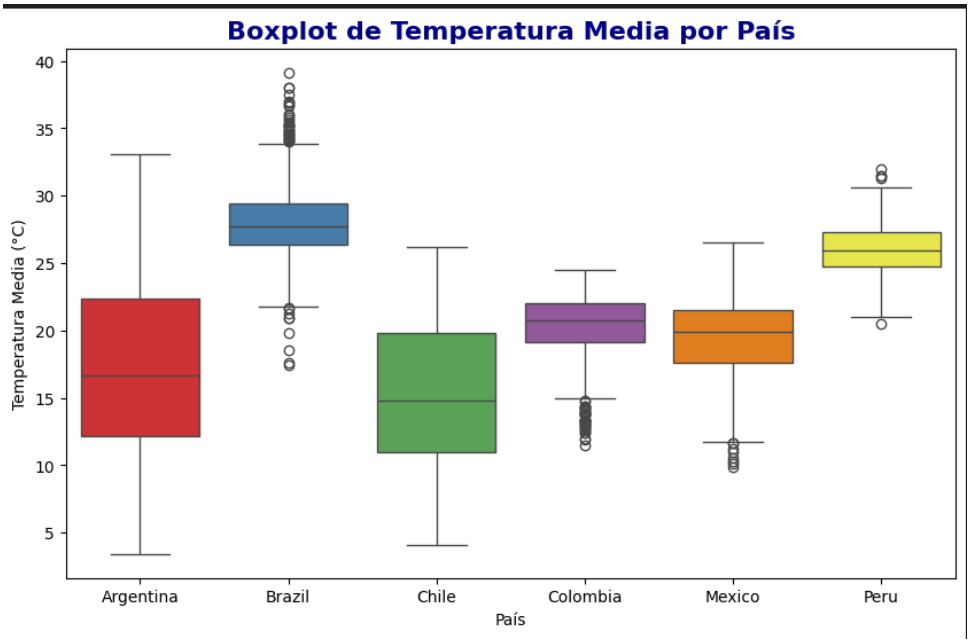

Fig 15.

En la figura 15 se puede visualizar que Brasil y Perú tiene temperaturas más altas que los otros paises, donde Brasil tiene valores atipicos bastante altos, argentina y chile tiene temperaturas más frías y Colombia se podría decir que tiene una temperatura más estable, se podría llegar a la conclusión que en verano es donde mas casos confirmados y fallecidos hay y esta hipótesis la podemos confirmar en los siguientes gráficos.

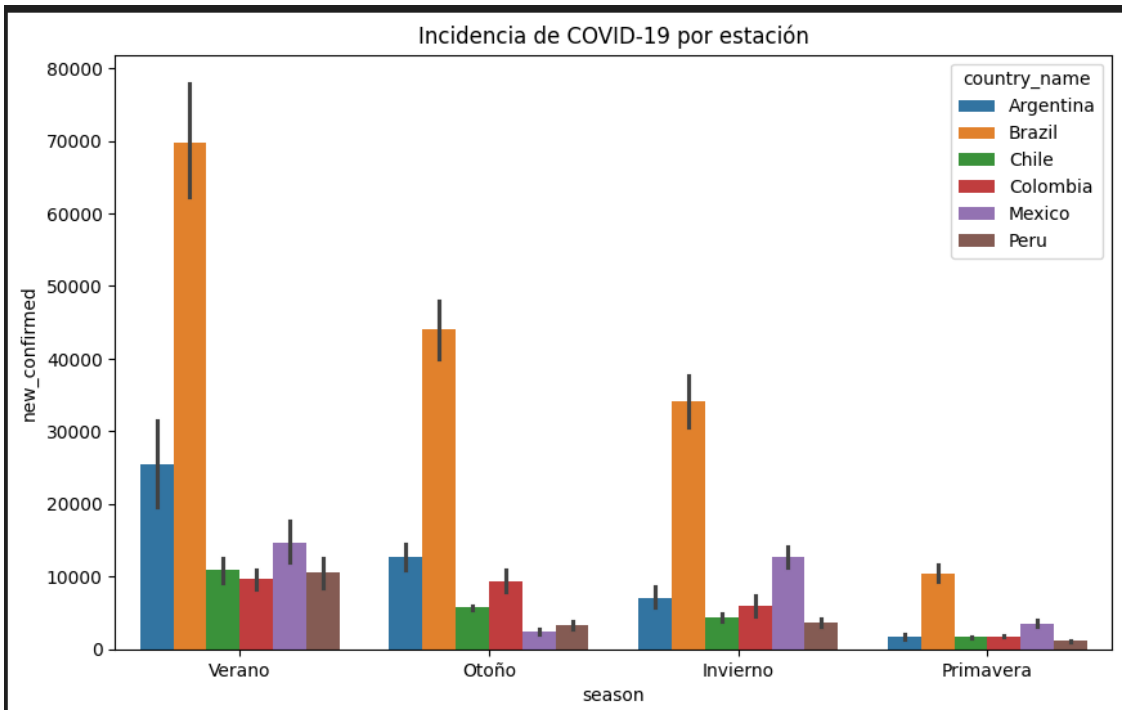
Incidencia de COVID-19 por estación
country_name
Argentina
Brazil
Chile
Colombia
Mexico
Peru
new_confirmed
80000
70000
60000
50000
40000
30000
20000
10000
0
Verano
Otoño
Invierno
Primavera
season

Fig 16.

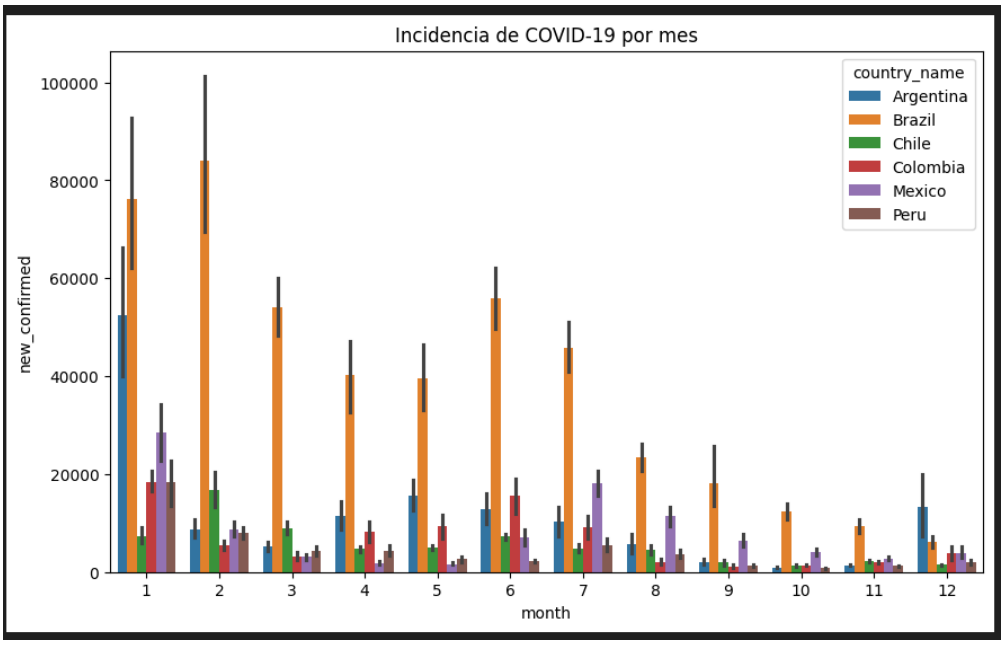
Incidencia de COVID-19 por mes
country_name
Argentina
Brazil
Chile
Colombia
Mexico
Peru
new_confirmed
100000
80000
60000
40000
20000
0
1
2
3
4
5
6
7
8
9
10
11
12
month

Fig 17.

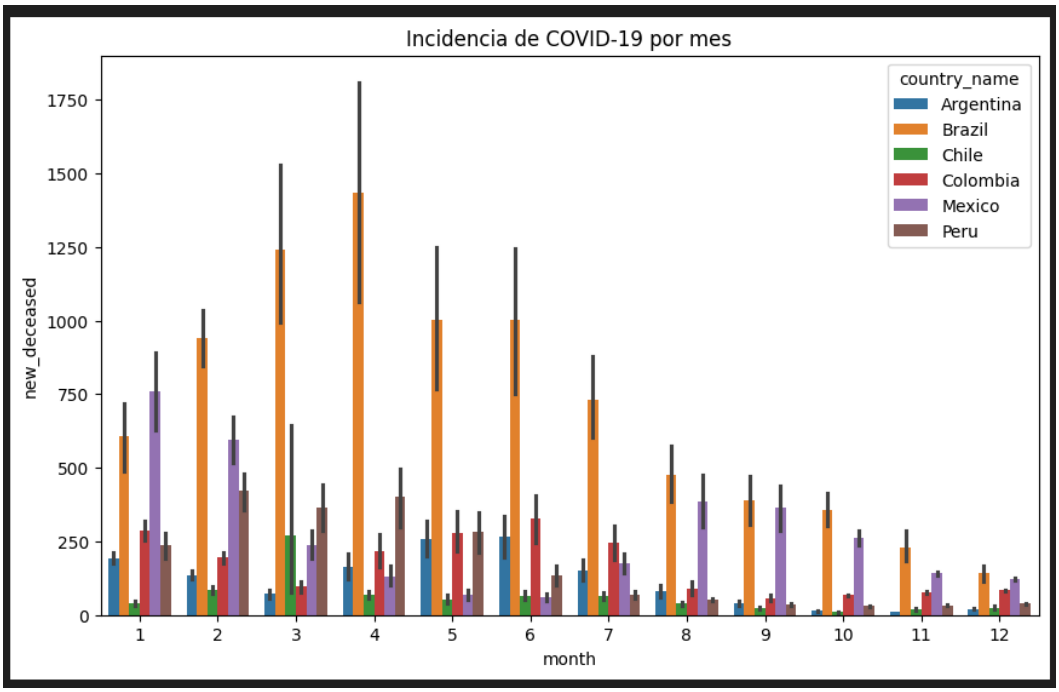
Incidencia de COVID-19 por mes
country_name
Argentina
Brazil
Chile
Colombia
Mexico
Peru
new_deceased
0
250
500
750
1000
1250
1500
1750
1
2
3
4
5
6
7
8
9
10
11
12
month

Fig 18.

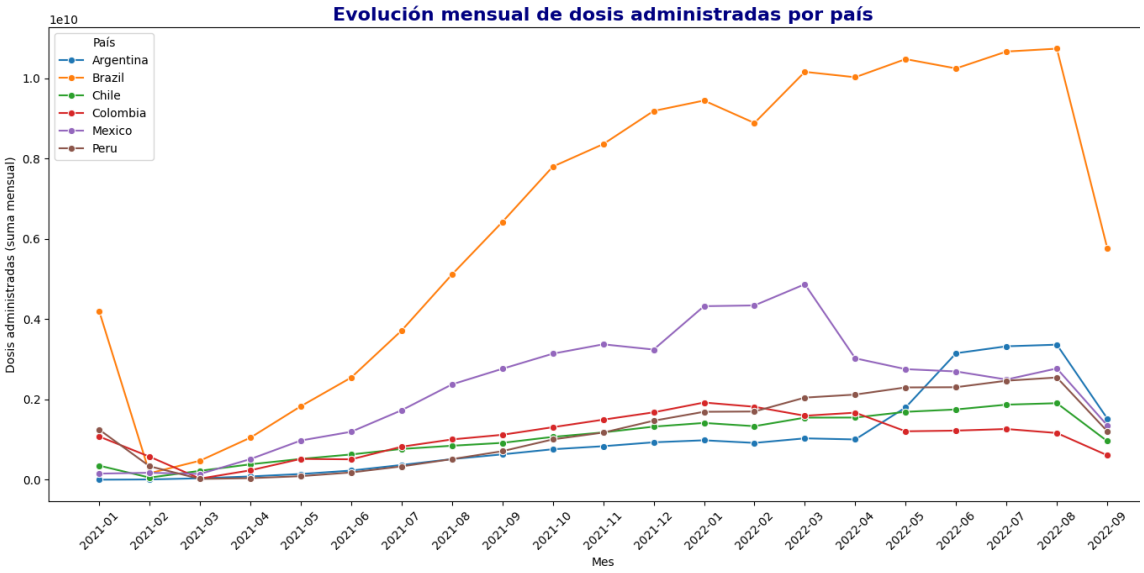
Evolución mensual de dosis administradas por país
1e10
País
Argentina
Brazil
Chile
Colombia
Mexico
Peru
Dosis administradas (suma mensual)
0.0
0.2
0.4
0.6
0.8
1.0
2021-01
2021-02
2021-03
2021-04
2021-05
2021-06
2021-07
2021-08
2021-09
2021-10
2021-11
2021-12
2022-01
2022-02
2022-03
2022-04
2022-05
2022-06
2022-07
2022-08
2022-09
Mes

Fig 19.

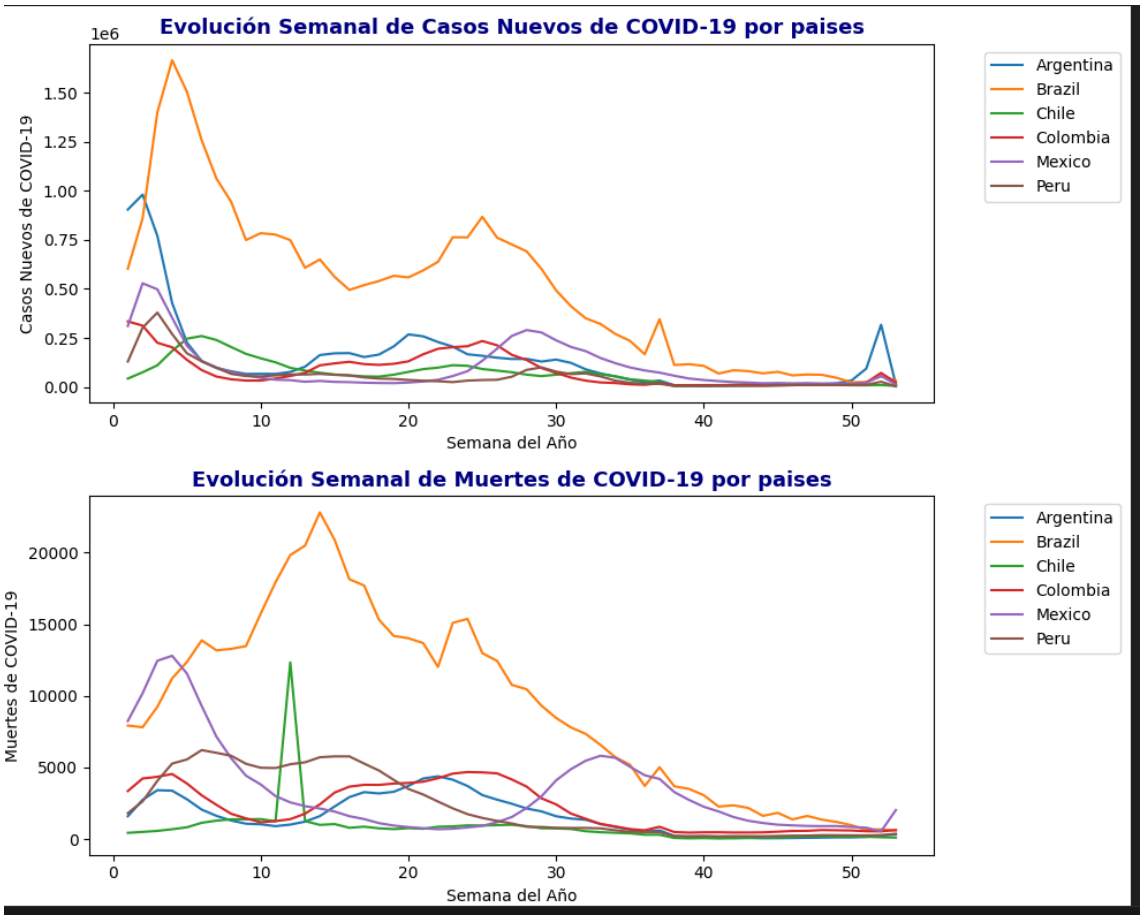

Fig 20.

Los gráficos del 16 al 20 tienen en común que son series temporales los cuales nos van a ayudar a verificar patrones y poder predecir cual es el comportamiento de los próximos meses y evitar tener picos muy altos.

Por ejemplo en la fig 16 se puede evidenciar a simple vista que los casos confirmados van en descendencia cronológicamente por estación, obteniendo menos casos en primavera. Con este patrón podemos llegar a la conclusión que verano es la estación donde más casos confirmados y fallecidos hay.

Se sugiere que a finales del año que es la temporada donde no hay tantos casos, se suministren más vacunas, se promuevan más campañas de prevención, mayor control de movilidad.

Por ejemplo en el grafico 19 las dosis suministradas a lo largo del tiempo no van en aumento como Brasil y esto puedo ayudar considerablemente para que no hayan casos a futuro.

En la figura 20 se puede evidenciar un gráfico con más detalle a lo largo del tiempo, comparando casos confirmados vs los fallecidos, donde Brasil sobresale con respecto a los demás países, visualizando un pico de fallecidos de la semana del 10 al 20, después de la semana 25 los fallecidos se fueron reduciendo progresivamente, dando a entender que se tomaron más medidas.

En el caso de argentina se puede verificar un pico de confirmados alto en la primer semana pero no tantos muertos como México y Perú que a lo largo del año sobresalieron con respecto a los demás países.

Para bajar estos números en el caso de México y Perú se podrían implementar las medidas que tuvo argentina que pudieron ser mayores medidas sanitarias, menor interacción social, campañas de prevención, varias jornadas de vacunación.

Comparando la figura 16 con la 19 podemos concluir que hay una correlación negativa entre casos confirmados vs vacunas suministradas, significa que entre mas vacunas suministradas menos casos confirmados van haber. Observando la fig 19 La vacunación masiva y temprana (Brasil, México) tuvo un impacto directo en la reducción de mortalidad, mientras que los países con campañas más lentas o tardías (Perú, Colombia, Chile) experimentaron descensos más graduales.

Los primeros meses fueron los más críticos, para esta época se hubiesen implementado cierres tempranos, control fronterizo, mas restricciones, protocolos de movilidad.

Brasil necesita un enfoque distinto a los demás países dada su magnitud (más recursos hospitalarios, campañas focalizadas en estados críticos,

Brasil, México y Perú muestran que la mortalidad fue muy alta comparada con contagios. Para bajar estos números se puede implementar la necesidad de más UCI, oxígeno, medicamentos y profesionales capacitados.

La mortalidad refleja tanto la magnitud de los contagios como la capacidad de respuesta de los sistemas de salud. **Prevenir el colapso hospitalario y vacunar rápidamente a los más vulnerables** son las acciones que más impacto hubieran tenido para tener más recuperados.

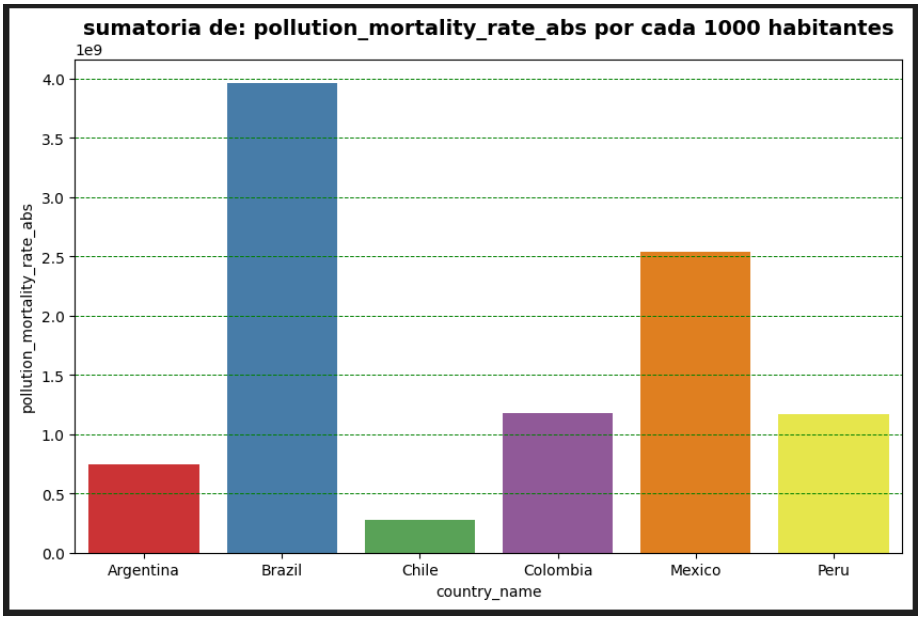

Fig 21.

Del grafico 21 podemos deducir que Brasil, México y Perú tienen más mortalidad por contaminación a comparación de los otros países, esto se puede deber a la población o la densidad poblacional siendo un factor determinante para que hayan más casos confirmados.

Para minimizar los casos en estos países se debe priorizar políticas ambientales (reducción de emisiones, transporte limpio, control de industrias).

Carrera: Data Analíticos

Módulo 4

Análisis del dashboard

El lienzo tiene unas medidas de 1080*1920 con un fondo de color verde y gráficos de color amarillo para que genere un contraste agradable a la vista. Como el tema es de un virus se utilizaron estos colores para darle una idea al lector de que va a tratar el tema.

En las diferentes paginas tenemos unos botones para navegar entre las diferentes páginas y la página principal.

En el power bi podemos encontrar 4 paginas que están distribuidas de la siguiente manera:

1 Menú principal

Esta es la primer página que encontraremos en nuestro dashboard con un menú de 3 opciones que corresponden a acumulados, evolución temporal e incidencia covid-19.

2. acumulados

Esta segunda página se encontrara 2 apartados, el primero con unos KPI en la parte superior donde se evidenciaran medidas que considero que son importantes para comparar como son (total confirmados, total fallecidos, total recuperados, temperatura promedio, PIB, promedio de lluvia, etc.).

En la parte izquierda encontraremos un segmentador para categorizar la información en país y fecha(año, trimestre, mes).

En la parte inferior habrán 4 gráficos de barras mostrando acumulados de los diferentes países y en la parte derecha esta un mapa para evidenciar cada país cuando se seleccione, para darnos una idea de donde esta ubicado y su área.

3. Evolución temporal

En la parte superior habrán diferentes gráficos de dispersión relacionando pares de variables verificando si están correlacionadas o no.

En la parte inferior tenemos un gráfico de líneas verificando el comportamiento a lo largo del tiempo de los diferentes países entre total confirmados, total fallecidos y recuperados.

Por ultimo se puede evidenciar un gráfico de torta mostrando la distribución de edades de los diferentes países.

4. Incidencia de covid-19

La ultima pagina se pueden visualizar dos gráficos de barras agrupados comparando los casos confirmados vs los casos fallecidos para verificar si las vacunas están surtiendo efecto o si se están tomando las medidas adecuadas para disminuir los casos y poder bajar los picos en los próximos años.

Carrera: Data Analíticos

Módulo 4

Conclusiones y Recomendaciones

Conclusión general

Tanto en el notebook como en el dashboard y analizando todos los gráficos generados podemos llegar a la conclusión que los países que más necesitan los centros de laboratorios farmacéuticos son **Brasil**, México y Perú por varios factores como son densidad poblacional, mortalidad por contaminación, clima, etc. **Brasil como tal necesita una prioridad mayor con respecto a los otros dos países porque es un país que supera en varias estadísticas a los demás países como pueden ser (población, área, PIB, clima, mortalidad (infantil, masculina, femenina, contaminación), etc.).**

Conclusiones estratégicas

1. Se recomienda implementar campañas de concientización, restricciones focalizadas y refuerzo del sistema hospitalario antes de verano que es la estación de picos más altos.
2. Se sugiere más asignación de camas UCI, personal medico calificado, enfermeras y bastantes suministros médicos para inicios de año.
3. Se recomienda priorizar la vacunación en regiones con alta letalidad para reducir el impacto sanitario en estos casos serian para Brasil, México y Perú.
4. La vacunación debe ir acompañada de inversión continua en atención primaria, educación comunitaria y vigilancia epidemiológica.
5. En zonas urbanas con alto grado de población por ejemplo lima Perú, la transmisión comunitaria se intensifica, las estrategias pueden ser mas centros de salud en estas zonas.
6. En contextos con alta pobreza y desempleo, las restricciones de movilidad y cuarentenas estrictas generan resistencia social. Las medidas deben ser acompañadas por apoyos económicos a las personas más vulnerables.

Reflexión personal

En este proyecto aprendí a integrar Python con power bi que son dos herramientas muy útiles a la hora de limpieza, transformación y manipulación de datos. Los mapas de calor son de gran utilidad porque nos ayuda a verificar que variables están relacionadas tanto positivo como negativamente y esto ayuda a sacar insights muy valiosos. Los gráficos de violin plot es un grafico que es muy fácil de entender y resume muchos datos en un solo grafico y se pueden sacar muchas conclusiones a simple vista.

Como analista de datos aprendí a deducir o predecir los datos que pueden haber los próximos meses o años gracias a las series temporales, sacando el provecho a los gráficos de líneas, identificando patrones y tomar las medidas necesarias para bajar estas estadísticas.

La limpieza es el primer paso para trabajar un dataset como analista de datos y es un factor demasiado importante porque se puede pasar de un dataset de 2 gb a uno que no supera las 2 Mb de peso y sin perder información elemental, solo hay que analizar bien los datos y saber que datos nos están suministrando para realizar una correcta limpieza. Esto ayuda mucho tanto en el rendimiento como en la generación de gráficos, mostrando información verídica.

Respondiendo la pregunta me hubiera gustado trabajar con el dataset original de 22 millones de registros y 707 columnas para tener en cuenta mas variables que hubiesen podido ayudar a generar mas conclusiones e insights valiosos para las expansiones de los centros farmacéuticos de byogenesis.