**Expansión Estratégica de Biogenesys con Python**

**Nombre del autor:** Marcelo Fabián López

**Email:** mfl201410@gmail.com

**Cohorte:** da-pt01

**Fecha de entrega:** 29/07/2024

**A logo with black and blue dots

Description automatically generated**

**Biogenesis Laboratories** es una empresa en crecimiento que busca expandirse en el mercado latinoamericano, específicamente en seis países:

* Argentina
* Brasil
* Chile
* Colombia
* Perú
* México

# Introducción

La empresa farmacéutica **BIOGENESYS**busca identificar las ubicaciones óptimas para la expansión de laboratorios farmacéuticos, basándose en el análisis de datos de incidencia de COVID-19, tasas de vacunación, y la disponibilidad de infraestructuras sanitarias. La meta es **optimizar la respuesta a los efectos de la pandemia y postpandemia con el fin de mejorar el acceso a las vacunas**.

# Desarrollo del proyecto:

**Avance 1:**

Comenzamos cargando un archivo CSV (‘data\_latinoamerica’) en Visual Studio Code para visualizarlo y trabajarlo con el lenguaje Python.

Instalamos e importamos librerías tales como Numpy, Pandas, Matplotib y Seaborn, que nos resultaron muy útiles para realizar cálculos, transformaciones y gráficos.

También creamos un archivo requirements.txt con una lista de todas las librerías instaladas y sus versiones.

El dataframe creado a partir del archivo CSV es un archivo de Jupyter Notebook y tiene originalmente 12216057 filas (o registros) y 50 columnas.

Como nos centramos en solo 6 países, hicimos una lista que solo incluyó a estos (Argentina, Brasil, Chile, Colombia, Perú y México) y filtramos el dataset, que pasó a ser conformado solo por estos países. La razón por la cual filtramos por estos 6 países es porque son el centro de nuestro objetivo de expansión.

El nuevo dataframe quedó constituido por 11970289 registros.

Entre las visualizaciones preliminares que realizamos, comprobamos la cantidad de filas y columnas, el nombre de éstas últimas, y algo muy relevante: la cantidad de faltantes o nulos.

En primer lugar, hicimos una comprobación para ver cuáles eran las columnas con más de 2.5 millones de datos faltantes por columna y vimos esto:

A screenshot of a computer program

Description automatically generated

Esta gran cantidad de datos nulos nos obligó a buscar una estrategia para que nuestro estudio tuviera un mejor aprovechamiento, y pensamos en reemplazar dicho faltantes.

Pero antes decidimos descartar aquellas filas que tenían una location\_key diferente de la adoptada para los países del estudio (‘AR’, ‘BR’, ‘CL’, ‘CO’, ‘PE’ y ‘MX’). Registros con location\_key tales como AR\_A, AR\_A\_014, por ejemplo, quedaron fuera del dataset, debido a que se trataría de registros repetidos o zonales.

Luego filtramos el dataframe para que sus registros incluyan solo datos desde el primero de enero de 2021.

De esta manera, trabajando con los 6 países seleccionados y desde el 01/01/2021, nuestro dataset se redujo a 3750 filas.

Es aquí cuando decidimos ocuparnos de los valores faltantes o nulos y primero analizamos su composición por columna y cantidad:

A screenshot of a computer program

Description automatically generated

Observamos columnas con un faltante de más del 50% de los datos, como ‘new\_recovered’ y ‘cumulative\_recovered’).

En las columnas con mayor cantidad de datos faltantes, como las citadas (más ‘cumulative\_vaccine\_doses\_administered’ reemplazamos por cero, mientras que en las demás reemplazamos por la media, mediante códigos de Python.

Esto nos proporcionó un dataset listo para realizar cálculos y estadísticas, que estudiamos de manera preliminar con funciones como .head().

A continuamos visualizamos medidas estadísticas con un blucle FOR y la función .describe() y obtuvimos las medidas de COUNT, UNIQUE, TOP Y FREC para las columnas. También creamos una función para obtener la mediana, varianza, rango y moda para cada columna con valores numéricos.

A continuación, imprimimos algunos datos descriptivos de las columnas, tales como COUNT, MEAN, STD, MIN, 25%, 50%, 75% Y MAX.

Respondiendo a la pregunta realizada en la consigna de este avance, debo decir que no todas las estadísticas se muestran para todas las columnas, ya que algunas como ‘location\_key’ o ‘country’ no son del tipo de datos numéricos. De todos modos, estas primeras aproximaciones pueden sernos muy útiles para un futuro análisis.

También analizamos el tipo de dato para cada columna mediante la función .info().

Luego guardamos el dataset en un nuevo CSV llamado DatosFinalesFiltrado.

Analizando el proceso, debo aclarar que podríamos haber eliminado los registros con valores faltantes (una de las posibilidades evaluadas), pero preferimos tratarlas como se ha descripto. Esta es otra alternativa posible y nos permitió ejercitarnos en los códigos necesarios para efectuarlas.

Como respuesta a las preguntas de la consigna, debo decir que la media, también conocida como promedio aritmético, es una medida de tendencia central que se calcula sumando todos los valores de un conjunto de datos y dividiendo el resultado entre el número total de valores.

En cuanto a medidas tales como varianza y rango, observamos en muchos casos (como el representado abajo) una muy alta dispersión de valores, por lo que un análisis debe ser muy cuidadoso de cada columna a evaluar, a la hora de obtener insights productivos.

A screenshot of a computer

Description automatically generated

La combinación de una varianza extremadamente alta y un rango amplio junto con el uso de la mediana como una medida de tendencia central puede indicar una alta variabilidad en los datos, inconsistencias significativas y/o presencia de outliers.

El objetivo de este avance fue aplicar técnicas de limpieza de datos para asegurar su calidad, facilitar su análisis y poder tomar decisiones estratégicas confiables.

**Avance 2:**

# EDA e insights:

Teniendo nuestros datos ‘limpios’, profundizamos en el análisis exploratorio de datos (EDA) para extraer insights que nos orienten en la planificación estratégica para la expansión de la empresa. Buscamos a través de visualizaciones claras y detalladas**, descubrir patrones, tendencias y anomalías** en los datos de incidencia de COVID-19, tasas de vacunación y disponibilidad de infraestructura sanitaria. El objetivo es crear un cuadro completo que permita a los directivos **tomar decisiones informadas**sobre dónde ubicar los nuevos laboratorios y centros de vacunación para maximizar el impacto positivo en la salud pública.

Comenzamos poniendo la columna de fechas ‘date’ como índice para que nos permita realizar un análisis de los datos a lo largo del tiempo (y que esté como tipo de datos de fecha).

Luego realizamos una matriz de correlación por medio de un heatmat, y vimos una mayoría de correlaciones fuertes entre ellas:

* cumulative\_confirmed y cumulative\_deceased
* cumulative\_recovered y cumulative\_confirmed
* gpd\_per\_capita\_usd y human\_development\_index
* area\_urban\_square\_kilometer y population\_largest\_city

A red and blue squares

Description automatically generated

Construimos gráficos de barras para analizar columnas tales como: new\_confirmed’, ‘new\_deceased’, ‘population’, ‘population\_male’, ‘population\_female’, ‘population\_rural’, 'population\_urban’, que dieron cuadros parecidos a éste:

A graph with blue bars

Description automatically generated Con Brasil sobresaliendo debido a su mayor población.

Gráficos de population de varios rangos etarios, como éste (siempre con la misma distribución atada a la población total):

A graph with blue bars

Description automatically generated

Gráficos de gdp\_uss\_per\_capita y país:

A screenshot of a graph

Description automatically generated

De países con ciudades más pobladas:

De superficie rural y urbana, expectativa de vida y tasa de mortalidad de adultos hombres y mujeres y de tasa de muertes por polución y comorbilidad.

De estas gráficas pudimos ver que Brasil y México son los países de mayor población y que podrían representar un mercado prometedor, que México y Colombia son los países con mayor densidad de población.

Por otra parte, Chile sobresale con su PBI per cápita, por lo que no debería dejar de ser tenido en cuenta, lo mismo que Argentina, con su alta tasa de médicos cada 1000 habitantes.

También generamos gráficos histplot para analizar algunas columnas, como ‘cumulative\_confirmed’, que nos muestra una distribución altamente asimétrica con una gran dispersión y la presencia de varios outliers.

A graph of a number of columns

Description automatically generated

También hicimos gráficos de scatterplot para analizar la relación entre la temperatura y nuevos casos confirmados, que no parecen determinar una correlación fuerte.

A graph of different colored dots

Description automatically generated

Realizamos análisis a lo largo del tiempo, como el de la evolución de las dosis administradas por mes en cada país:

A graph with different colored lines

Description automatically generated

Y de casos confirmados, muertes, recuperaciones, y de comparación de nuevos casos por país, de lo que deducimos que Chile fue el país que mejor manejó el tema de la vacunación y prevención.

A graph of different colored lines

Description automatically generated

Hicimos un gráfico boxplot de temperaturas promedio:

A chart with different colored squares

Description automatically generated

También una serie de violinplots, como éste, que indican una alta variabilidad en Brasil y Argentina. La alta dispersión y las largas colas en los gráficos de Brasil y Argentina indican que estos países han tenido días con un número extremadamente alto de casos nuevos, lo que sugiere posibles brotes o picos en ciertas épocas.

A chart of different colors of different sizes

Description automatically generated

Hicimos un gráfico de barras para estudiar la distribución general de la población por rangos etarios:

A graph of a number of people

Description automatically generated

Generamos gráficos para analizar la vacunación por mes por rangos etarios, y nos dio gráficos como éste:

A graph with different colored lines

Description automatically generated

También hicimos un barplot para comparar la tasa de mortalidad masculina y femenina:

A graph of different colored squares

Description automatically generated

Observamos que la tasa es más alta en hombres.

# Generamos una matriz de correlación de indicadores demográficos y de salud, donde vemos que el fumar y la diabetes son las enfermedades preexistentes más presentes, por lo que podrían influir en las muertes por COVID agravando los cuadros.

A screenshot of a graph

Description automatically generated

Generamos barplots para analizar la relación entre el fumar y la tasa de muertes por COVID, y de la relación entre la presencia de diabetes y la tasa de muertes por COVID:

A graph of a number of people with diabetes

Description automatically generated with medium confidence

Aquí vemos el de diabetes, por ejemplo.

Como resultado del análisis de todos estos gráficos, creemos que Brasil, México y Colombia son los países que necesitan más ayuda, ya que son los países que más casos de confirmados y muertes han tenido, además de ser los que menos vacunas con relación a la población han suministrado.

**Avance 3**

**EDA con Numpy y Pandas**

Aquí procuramos un análisis más profundo y sofisticado de los datos relacionados con la incidencia de COVID-19. El objetivo en este avance fue pulir y preparar nuestros datos para una visualización avanzada que permita identificar con precisión las ubicaciones más estratégicas para la expansión de los laboratorios farmacéuticos.

Mediante la función .isocalendar() creamos nuevas columnas, como ‘week’ en este caso, para hacer análisis en el tiempo:

A computer screen with text and numbers

Description automatically generated

De esta manera podemos hacer gráficos con variaciones semanales, como éste de nuevos casos por semana:

A graph of a virus

Description automatically generated with medium confidence

También podemos hacer lo mismo por mes:

A graph of a number of people

Description automatically generated

Generamos gráficos que analizan 2 variables, como la temperatura promedio y los casos nuevos confirmados:

A graph with red and blue lines

Description automatically generated

Vemos que con la disminución de la temperatura se dio un aumento en los casos, pero para Marzo de 2022hubo un gran aumento, que disminuyó bruscamente, posiblemente por la mayor aportación de vacunas en Brasil.

También analizamos la tasa de urbanización y los casos confirmados:

A screenshot of a graph

Description automatically generated

Donde vemos que en Chile se da un mayor impacto en la propagación, extraño fenómeno teniendo en cuenta que fue uno de los países que mejor manejó la pandemia.

Preparamos gráficos para ver la distribución de la población en los países seleccionados, por ejemplo, Argentina:

A pie chart with different colored circles

Description automatically generated

Una gráfica muy interesante es la que nos muestra el número de dosis por habitante, donde vemos que Chile sobresale en primer lugar y Brasil es el país con menos dosis:

A graph with green bars

Description automatically generated with medium confidence

Analizamos también la incidencia de fumar de y diabetes en los países de estudio, y vimos que la diabetes afecta mucho más a México y a Brasil. Chile y Argentina, por otra parte, son los más afectados por el fumar.

Sin embargo, mientras Brasil aparece entre los países con mayor mortalidad, México aparece entre los menores. Lo mismo sucede entre Chile (mayor) y Argentina (menor), por lo que es difícil establecer una correlación entre estos factores y la fatalidad por COVID.

A screenshot of a graph

Description automatically generated

También calculamos la tasa de mortandad por COVID (muertes/población) y vimos resultados similares en los países:

A graph of red and white bars

Description automatically generated

Otra gráfica esclarecedora es la que nos muestra la relación entre la población vacunada en el tiempo y la reducción de casos, donde vemos que, a mayor vacunación, menor cantidad de nuevos casos, en general:

A graph with lines and numbers

Description automatically generated

Con otro gráfico vimos que la tasa de mortalidad fue disminuyendo con el correr del tiempo y la vacunación de la población:

A graph showing a line of a virus

Description automatically generated with medium confidence

Analizamos la cantidad acumulada de casos activos y recuperados, y vimos que ambas fueron creciendo en el tiempo.

A graph with blue and orange lines

Description automatically generated

Finalmente graficamos en un mapa de calor para casos confirmados, muertes, vacunas administradas y recuperados:

A map of the south america with blue dots

Description automatically generated

# Análisis del dashboard

Realicé un dashboard en Power BI que consta de una carátula, 6 botones de navegación, que se repiten en todos los paneles, y 6 paneles.

A screenshot of a computer

Description automatically generated

En el panel ‘General’ puse arriba y en todos los paneles, tarjetas con métricas que me parecen muy importantes, como la cantidad de fallecidos, la cantidad de confirmados y la población total. Estos indicadores son relevantes para tener una idea clara de la magnitud de la pandemia.

A screenshot of a graph

Description automatically generated

Incluí en todos los paneles segmentadores para países y años (nos enfocamos en los años 2021 y 2022).

Incluí 4 gráficos en ‘General’.

El primero desde la izquierda (de líneas) es uno sobre nuevos casos confirmados por mes y por año (para los 6 países), que considero importante a la hora de evaluar cómo se fue desarrollando la pandemia.

A su derecha hay otro (de líneas) sobre la cantidad de fallecidos por mes y por año, con el mismo objeto: mostrar la tendencia y la cantidad de fallecidos.

Abajo a la izquierda tenemos un gráfico de líneas que muestra la evolución de la vacunación por mes y por año de los países en estudio.

Y finalmente a su derecha un gráfico de barras para mostrar la cantidad total de dosis aplicadas por país.

En el panel ‘Clima’ incluí 3 gráficos (realizados en Python).

A screenshot of a computer

Description automatically generated

El primero a la izquierda (scatterplot) es uno sobre la relación entre la temperatura promedio y el número de casos confirmados, para analizar si realmente, como se pensaba, había una mayor probabilidad de contagios con bajas temperaturas.

A su derecha tenemos uno de líneas que muestra la evolución de casos activos versus casos recuperados y abajo otro de líneas que nos muestra la relación entre nuevos casos y temperatura promedio.

El cuarto panel se llama ‘Geografía’ y allí tenemos 5 gráficos que nos permiten conocer a los países más en detalle según algunos indicadores elegidos:

A screenshot of a graph

Description automatically generated

Todos son gráficos de barras, y el primero de la izquierda es uno que muestra la población total por país. Le sigue uno con el PBI per cápita y otro con el índice de desarrollo humano por país.

Abajo tenemos uno de densidad de población y otro que muestra porcentajes de población urbana y rural.

El quinto panel se llama ‘Correlaciones’:

A screenshot of a computer

Description automatically generated

Incluye 3 gráficos realizados en Python y 2 e Power BI.

Los dos primeros son de barras y nos muestran la incidencia de enfermedades preexistentes (tabaquismo y diabetes) con las muertes por COVID.

Abajo tenemos una tabla con indicadores básicos y relevantes, para una idea rápida de las dosis administradas, casos confirmados y cantidad de fallecidos por país y por año.

Y a la derecha un gráfico de correlaciones (heatmap) con un análisis de múltiples variables, tomadas de varias columnas del dataset.

Por último, tenemos un panel llamado ‘Conclusiones’, donde aconsejamos a los directivos de la empresa a seguir profundizando en las investigaciones, como paso siguiente a la expansión definitiva.

Aquí hacemos hincapié en priorizar áreas con alta incidencia de Covid-19, evaluar la infraestructura sanitaria, análisis demográficos, condiciones socioeconómicas y tener en cuenta los factores ambientales y de salud pública.

A screenshot of a computer

Description automatically generated

# Conclusiones y Recomendaciones

Para abordar esta tarea, vamos a sintetizar los resultados obtenidos a partir del análisis de los datos y las visualizaciones. La síntesis se enfoca en los factores clave que afectan la expansión de los laboratorios farmacéuticos en los países en estudio: Brasil, Argentina, Chile, Perú, Colombia y México.

Síntesis de Resultados:

**1. Incidencia de COVID-19:**

* Brasil y México presentan las cifras más altas de casos confirmados y fallecimientos por COVID-19, indicando una alta necesidad de intervención y recursos.
* Chile y Argentina muestran un mejor manejo de la pandemia con tasas de vacunación relativamente altas.

**2. Tasas de Vacunación:**

* Chile se destaca por una alta tasa de vacunación, lo cual indica un sistema de salud eficiente y receptivo.
* Perú y Colombia muestran una necesidad moderada de mejora en sus programas de vacunación.

**3. Disponibilidad de Infraestructura Sanitaria:**

* Brasil y México tienen una infraestructura sanitaria más desarrollada en términos de número de hospitales y personal médico, pero esta puede estar sobrecargada debido a la alta incidencia de COVID-19.
* Perú y Colombia presentan deficiencias en la disponibilidad de personal médico y camas hospitalarias, destacando áreas críticas para inversión.

**4. Datos Demográficos:**

* Brasil y México tienen grandes poblaciones urbanas y rurales, sugiriendo una necesidad diversa de servicios de salud tanto en áreas metropolitanas como en zonas remotas.
* Chile y Argentina tienen poblaciones más concentradas en áreas urbanas, lo que facilita la distribución de recursos.

**5. Indicadores Socioeconómicos:**

* Chile y Argentina tienen altos índices de desarrollo humano y PIB per cápita, lo que sugiere una mayor capacidad de inversión en salud.
* Perú y Colombia muestran desafíos socioeconómicos más pronunciados, lo que requiere estrategias de inversión más enfocadas y sostenibles.

**6. Factores Ambientales y de Salud Pública:**

* Brasil y México tienen altas prevalencias de enfermedades crónicas como diabetes y problemas relacionados con la contaminación, aumentando la complejidad de la gestión de la salud pública.
* Chile presenta menores tasas de comorbilidad, lo cual facilita la intervención sanitaria.

**Conclusiones Estratégicas**

1. Brasil y México deben ser prioritarios debido a su alta incidencia de COVID-19 y grandes poblaciones, pero las estrategias deben incluir la expansión de infraestructura sanitaria y programas de salud pública enfocados en enfermedades crónicas.
2. Chile y Argentina son óptimos para la inversión debido a su alto desarrollo socioeconómico y buenas tasas de vacunación. Aquí, la expansión puede enfocarse en mejorar la capacidad de respuesta ante futuras crisis sanitarias y en innovaciones en salud pública.
3. Perú y Colombia necesitan inversiones específicas para mejorar la infraestructura sanitaria y aumentar las tasas de vacunación. Estos países presentan oportunidades para el desarrollo de nuevos centros de vacunación y laboratorios en áreas subatendidas.
4. Áreas Urbanas vs. Rurales: En todos los países, las zonas urbanas deberían ser el foco inicial de expansión debido a la alta concentración de la población y la mayor facilidad de acceso a servicios. Sin embargo, es crucial desarrollar estrategias para las áreas rurales para asegurar una cobertura completa.

**Identificación de Ubicaciones Óptimas para la Expansión de Laboratorios Farmacéuticos:**

1. **Brasil**:
   * Ciudades clave: São Paulo, Rio de Janeiro, Brasilia.
   * Razón: Alta incidencia de COVID-19, grandes poblaciones urbanas, y la necesidad de infraestructura sanitaria mejorada.
2. **México**:
   * Ciudades clave: Ciudad de México, Guadalajara, Monterrey.
   * Razón: Alta incidencia de COVID-19, desafíos socioeconómicos y de salud pública, y grandes poblaciones urbanas.
3. **Chile**:
   * Ciudades clave: Santiago, Valparaíso, Concepción.
   * Razón: Alta tasa de vacunación, alto índice de desarrollo humano, y buena infraestructura sanitaria.
4. **Argentina**:
   * Ciudades clave: Buenos Aires, Córdoba, Rosario.
   * Razón: Buena tasa de vacunación, alto desarrollo socioeconómico, y concentración urbana.
5. **Perú**:
   * Ciudades clave: Lima, Arequipa, Trujillo.
   * Razón: Necesidad de mejorar infraestructura sanitaria, moderada tasa de vacunación, y desafíos socioeconómicos.
6. **Colombia**:
   * Ciudades clave: Bogotá, Medellín, Cali.
   * Razón: Necesidad de mejorar infraestructura sanitaria, moderada tasa de vacunación, y desafíos socioeconómicos.

Estas ubicaciones representan puntos estratégicos que permitirán a BIOGENESYS maximizar el impacto de su expansión en América Latina, mejorando el acceso a vacunas y la capacidad de respuesta ante futuras crisis sanitarias.

# Reflexión personal

A lo largo de este proyecto aprendí a transformar y analizar una base de datos con distintas librerías de Python. También a realizar gráficos propios de estas librerías en particular.

Aprendí a trabajar en Visual Studio Code con archivos de Jupyter Notebook.

Aprendí que para trabajar con grandes volúmenes de registros un lenguaje como Python es sumamente ventajoso, por su capacidad para manejar grandes datasets.

Aprendí a incluir información obtenida en Python, como gráficos, en un dashboard de Power BI, lo que enriquece una presentación.

También aprendí sobre la manera en que un analista de datos debe trabajar cuando una empresa le solicita consejos para evaluar alguna acción específica, como lo fue en este caso la posible expansión de un laboratorio en países de América.

Por último, aprendí de la importancia de la colaboración con compañeros e instructores, ya que, si bien este es un trabajo individual, sería muy difícil resolver todas las inquietudes que nos surgen si no consultáramos con nuestros pares o profesores, muchas veces acrecentando nuestros conocimientos (y viceversa).

Si tuviera que volver a empezar en este proyecto me gustaría analizar más en profundidad la base de datos. Tuve un inconveniente (aparición de datos con exceso de decimales) en algunos registros y al corregirlos me di cuenta de que hay errores en la base, como por ejemplo valores iguales para diferentes países en una misma columna. También me hubiera gustado tener más tiempo para llegar a conclusiones más certeras y no solamente proponer continuar profundizando en los análisis.

# 