**Análisis de COVID-19🚀**

**Nombre del autor: Alejo Nicolas Peralta Falconi**

**Email: peraltaalejo966@gmail.com**

**Cohorte: DA-FT008**

**Fecha de entrega:** 18/11/2024

**Institución:** Empresa Farmacéutica BIOGENESYS

Logotipo, nombre de la empresa

Descripción generada automáticamente

# Introducción

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

Como Data Analista de este proyecto, debo realizar un estudio que ayudará a la estrategia de expansión en Latinoamérica, específicamente en los países de Argentina, Brasil, Chile, Colombia, México y Perú. La dirección ha propuesto hacer esta inversión para ubicar regiones y recolectar datos que sean de importancia para la toma de decisiones informadas.

Los objetivos son concretos y medibles:

* Realizar un análisis exploratorio de datos sobre la incidencia de COVID-19 y otros factores relevantes, identificando tendencias y oportunidades mediante estadísticas, mediciones y visualizaciones.
* Aplicar técnicas de limpieza de datos para asegurar la calidad de los datos, facilitando análisis y decisiones estratégicas confiables.
* Mejorar el acceso a los datos mediante operaciones eficientes de extracción, transformación y carga (ETL), aumentando la eficacia del análisis y la toma de decisiones.
* Desarrollar dashboards interactivos con visualizaciones eficientes, permitiendo explorar datos desde múltiples perspectivas para una toma de decisiones informada y estratégica.

Se utilizarán los datos proporcionados para ofrecer insights valiosos sobre posibles ubicaciones estratégicas para futuros laboratorios y centros de vacunación. Este análisis de tendencias y demografía será fundamental para comprender el entorno del mercado en los países de interés.

**Informe de Avance 1: Carga y Transformación de los Datos**

**Introducción**

El primer avance del proyecto se centró en la carga y transformación de un dataset extenso, adaptado para analizar los datos de incidencia de COVID-19 en países de Latinoamérica. Este paso inicial fue esencial para garantizar que los datos estuvieran limpios, organizados y listos para el análisis posterior. A través de un proceso sistemático, se seleccionaron datos clave para reducir la complejidad del archivo original, facilitando así su manipulación y análisis.

**Objetivos**

1. Revisión del Dataset Original: Entender la estructura y el contenido del archivo proporcionado, identificando las columnas y registros relevantes.
2. Filtrado por País y Fecha: Aislar la información correspondiente a los países de interés en Latinoamérica y a un período posterior al 1 de enero de 2021.
3. Limpieza y Preparación Inicial: Gestionar valores nulos, transformar tipos de datos y asegurar la consistencia del dataset para futuras etapas del análisis.

**Desarrollo**

1. Análisis Inicial del Dataset

El archivo original contenía aproximadamente 22 millones de registros y 707 columnas. Dada su magnitud, se decidió trabajar con una versión optimizada preparada previamente por el equipo de ingeniería de datos. Este nuevo archivo, llamado "data\_latinoamerica.csv", se centró exclusivamente en los países de Latinoamérica, con un tamaño significativamente reducido (12,216,057 filas y 50 columnas).

Adicionalmente, un archivo "readme.txt" proporcionó detalles sobre las columnas incluidas en este nuevo dataset, lo que facilitó la comprensión de los datos y permitió identificar variables clave para el análisis.

**2. Filtrado por Países de Interés**

Se seleccionaron los siguientes países para el análisis, dada su relevancia en la región:

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

Este enfoque permitió priorizar las áreas de mayor interés estratégico para la expansión de laboratorios farmacéuticos.

**3. Filtrado Temporal**

El análisis se restringió a datos registrados a partir del 1 de enero de 2021, eliminando información previa que no era relevante para el contexto actual del proyecto. Este filtro temporal permitió centrar la atención en patrones más recientes y alineados con los objetivos del análisis.

**4. Gestión de Valores Faltantes**

La limpieza preliminar incluyó:

* Relleno de valores nulos: Se aplicaron estrategias como el uso de valores promedio, datos anteriores o siguientes, según correspondiera.
* Eliminación de registros incompletos: En casos donde los valores faltantes eran significativos, se descartaron para garantizar la calidad del análisis.
* Transformación de datos: Se corrigieron los tipos de datos para asegurar la compatibilidad con herramientas analíticas.

**5. Análisis de Variables Clave**

Se realizó una exploración inicial para comprender la distribución y características de las variables principales, como:

* Incidencia de COVID-19: Número de casos confirmados por país y período.
* Tasas de vacunación: Dosis administradas acumuladas y cobertura porcentual de la población.
* Factores contextuales: Infraestructura sanitaria y datos demográficos relevantes.

**6. Almacenamiento de Datos Filtrados**

Los datos resultantes de este proceso fueron guardados en un archivo llamado "DatosFinalesFiltrado.csv". Este archivo optimizado servirá como base para las siguientes etapas del proyecto, evitando la necesidad de repetir el proceso de limpieza y filtrado.

**Conclusiones**

1. Importancia de la Limpieza de Datos: La transformación del dataset fue fundamental para reducir su complejidad y mejorar su manejabilidad.
2. Enfoque Regional y Temporal: El filtrado por países y fechas relevantes asegura que el análisis se alinee con los objetivos estratégicos.
3. Base Sólida para el Análisis: El archivo resultante está listo para su uso en análisis exploratorios y avanzados, garantizando datos consistentes y confiables.

**Informe de Avance 2: Análisis Exploratorio – Visualización**

**Introducción**

En esta etapa, nos enfocamos en el análisis exploratorio de los datos para obtener información clave que respalde las decisiones estratégicas de expansión de los laboratorios farmacéuticos en los países seleccionados: Colombia, Argentina, Chile, México, Perú y Brasil. A través de la exploración estadística y la creación de visualizaciones, buscamos identificar patrones, tendencias y anomalías relacionadas con la incidencia de COVID-19, las tasas de vacunación y factores demográficos y sanitarios.

**Objetivos**

1. Explorar las propiedades estadísticas del dataset para comprender mejor la situación de cada país.
2. Crear visualizaciones que permitan identificar patrones y relaciones clave entre variables.
3. Identificar tendencias estacionales, geográficas o demográficas relacionadas con la propagación del COVID-19 y la vacunación.
4. Personalizar gráficos y presentaciones para una comunicación efectiva de los hallazgos.

**Desarrollo**

**1. Análisis Estadístico**

Se realizó un análisis estadístico utilizando medidas de tendencia central (media, mediana) y dispersión (desviación estándar, varianza) para evaluar la distribución de las variables clave. Este análisis permitió identificar las características más relevantes, tales como:

* Distribución de los casos confirmados y fallecimientos por país.
* Tasas de vacunación y su variabilidad entre países.
* Relación entre indicadores demográficos (como densidad de población) y variables de salud (como incidencia de casos).

Además, se calculó la correlación entre variables para identificar relaciones significativas. Por ejemplo, se analizaron correlaciones entre las tasas de vacunación, la incidencia de casos y factores como la temperatura media y la densidad poblacional.

**2. Visualización de Hallazgos Clave**

Las visualizaciones se realizaron para representar gráficamente las observaciones obtenidas durante el análisis estadístico. Entre las más destacadas se encuentran:

* **Histogramas y gráficos de densidad**: Para explorar la distribución de los casos confirmados y las tasas de vacunación en los países seleccionados.
* **Gráficos de barras**: Comparando casos confirmados, muertes acumuladas y tasas de vacunación entre países, permitiendo identificar países con mayor impacto del COVID-19 o mayor cobertura de vacunación.
* **Mapas de calor**: Para representar gráficamente las correlaciones entre variables, destacando relaciones significativas que podrían influir en las estrategias de expansión.
* **Diagramas de dispersión**: Analizando la relación entre la temperatura media y la incidencia del COVID-19, así como entre la temperatura y los fallecimientos, lo que permite explorar posibles patrones estacionales.
* **Evolución temporal**: Se analizaron las dosis de vacunas administradas, casos confirmados y muertes reportadas a lo largo del tiempo, observando variaciones mensuales por país.

**3. Identificación de Tendencias y Patrones**

El análisis exploratorio permitió detectar varias tendencias y patrones importantes:

1. **Patrones Temporales**: Se observaron fluctuaciones en los casos confirmados y la administración de vacunas en diferentes meses, sugiriendo estacionalidad y variabilidad en las campañas de vacunación.
2. **Diferencias Geográficas**: Algunos países, como Brasil y México, presentan tasas más altas de incidencia y mortalidad, mientras que otros muestran mejores resultados en términos de cobertura de vacunación.
3. **Factores Ambientales**: Se identificaron posibles influencias de la temperatura media sobre la propagación del COVID-19, aunque se requieren análisis más profundos para confirmar estas observaciones.

**4. Personalización de Visualizaciones**

Para mejorar la presentación de los hallazgos, las visualizaciones fueron personalizadas con:

* Paletas de colores diferenciadas por país.
* Etiquetas descriptivas y leyendas claras para facilitar la interpretación.
* Títulos informativos que destacan el propósito de cada gráfico.
* Tamaños ajustados para mejorar la legibilidad.

**Conclusiones**

1. El análisis estadístico y visual permitió identificar patrones clave en la incidencia del COVID-19 y las tasas de vacunación en los países seleccionados.
2. Las visualizaciones destacan diferencias significativas entre los países, lo que ayudará a priorizar áreas para la expansión de los laboratorios farmacéuticos.
3. La personalización de los gráficos mejoró la claridad y la comunicación de los resultados.

**Informe de Avance 3: EDA con Numpy y Pandas**

**Introducción**

En esta tercera etapa del proyecto, profundizamos en el análisis exploratorio de datos (EDA) utilizando herramientas avanzadas de Pandas y Numpy. Este análisis tiene como objetivo identificar patrones temporales, tendencias y correlaciones que respalden la toma de decisiones para la expansión de laboratorios farmacéuticos en América Latina. Enfocándonos en los datos de incidencia de COVID-19, tasas de vacunación y características demográficas, buscamos preparar el terreno para visualizaciones avanzadas y recomendaciones estratégicas basadas en datos sólidos.

**Objetivos**

1. **Explorar series temporales**: Analizar la evolución de casos, muertes y tasas de vacunación, identificando tendencias, estacionalidad y patrones temporales relevantes.
2. **Investigar correlaciones**: Examinar las relaciones entre variables clave como urbanización, temperatura, incidencia de casos y vacunación.
3. **Generar gráficos avanzados**: Representar visualmente los hallazgos clave a través de gráficos dinámicos y descriptivos.
4. **Preparar los datos**: Pulir el dataset y aplicar funciones personalizadas para optimizar las columnas de análisis.

**Desarrollo**

**1. Análisis Exploratorio de Series Temporales**

Se analizó la evolución de elementos clave del conjunto de datos, aplicando técnicas avanzadas como:

* **Identificación de tendencias**: Examinar la evolución mensual y anual de casos confirmados, recuperados y fallecimientos para cada país.
* **Estacionalidad y patrones**: Evaluar la periodicidad en el incremento de casos, vinculándola con eventos relevantes (temporadas frías, campañas de vacunación, etc.).
* **Análisis de autocorrelación y descomposición temporal**: Determinar relaciones entre datos pasados y actuales, descomponiendo las series en componentes como tendencia, estacionalidad y ruido.

**2. Visualización de Hallazgos**

Se crearon gráficos para representar los resultados más destacados del análisis temporal y exploratorio:

1. **Evolución de Casos Activos vs. Recuperados**:
   * Se observó el progreso en la recuperación de casos, mostrando países con mayor éxito en la contención del virus.
2. **Tasa de Crecimiento (%)**:
   * Comparación del crecimiento semanal y mensual de casos confirmados, identificando picos significativos.
3. **Relación entre la Cobertura de Vacunación y la Reducción de Casos**:
   * Gráficos que demuestran cómo la expansión de la vacunación influye en la disminución de casos y mortalidad.
4. **Progreso de la Vacunación por País**:
   * Gráficos acumulativos que ilustran las dosis administradas por país a lo largo del tiempo.
5. **Impacto de la Urbanización en la Propagación del COVID-19**:
   * Análisis que relaciona tasas de incidencia con densidad poblacional y grado de urbanización.
6. **Evolución Semanal y Anual de Casos Nuevos**:
   * Visualización de picos y disminuciones en los casos confirmados en diferentes períodos temporales.
7. **Variación Mensual de Casos y Muertes**:
   * Comparaciones de variaciones estacionales por país, destacando patrones recurrentes.
8. **Distribución de la Población por Edad**:
   * Análisis demográfico para identificar grupos de mayor riesgo en cada país.
9. **Prevalencia de Condiciones Preexistentes en Países con Altas y Bajas Tasas de Mortalidad**:
   * Relación entre factores de riesgo (como diabetes y enfermedades cardiovasculares) y mortalidad.
10. **Comparación de Estrategias de Vacunación**:
    * Análisis de la eficacia y rapidez en la distribución de vacunas en los países estudiados.

**3. Investigación de Correlaciones**

Se estudiaron correlaciones significativas entre las variables del dataset, destacando:

* Relación entre tasas de vacunación y disminución de la mortalidad.
* Impacto de la temperatura media en la propagación del virus.
* Influencia de la densidad poblacional y urbanización en la incidencia de casos.
* Correlaciones entre variables demográficas y tasas de mortalidad (como proporción de adultos mayores en la población).

**4. Aplicación de Funciones Personalizadas**

Se implementaron funciones creadas en avances previos para calcular estadísticas clave (mediana, varianza, rango) y aplicarlas sobre columnas seleccionadas. Esto permitió:

* Analizar la dispersión de los datos y la consistencia entre países.
* Comparar métricas como la mortalidad y la incidencia de casos entre distintas regiones.
* Identificar outliers y corregir posibles inconsistencias.

**Conclusiones**

1. **Tendencias Identificadas**: Se observaron patrones claros de estacionalidad y variabilidad en los datos, con países como Brasil y México mostrando mayores desafíos en la contención del virus.
2. **Relaciones Significativas**: Las correlaciones destacan el impacto positivo de la vacunación y la influencia de factores demográficos y ambientales en la incidencia del COVID-19.
3. **Preparación de los Datos**: Los datos han sido pulidos y transformados, listos para ser utilizados en visualizaciones avanzadas en el siguiente avance.

**Informe de Avance 4: Aplicaciones Prácticas - Integración en Power BI**

**Introducción**

En esta etapa final del proyecto, integramos el análisis realizado en fases previas en una plataforma visual interactiva utilizando Power BI. Este proceso transforma los datos analíticos en dashboards que permiten a los directivos explorar información clave para la toma de decisiones estratégicas. El objetivo es priorizar áreas de expansión para laboratorios y centros de vacunación basados en indicadores de incidencia de COVID-19 y cobertura de vacunación.

**Objetivos**

1. **Importar datos preparados a Power BI**: Utilizar los datos limpios y filtrados del archivo "DatosFinalesFiltrado.csv".
2. **Crear dashboards interactivos**: Diseñar visualizaciones que faciliten la comprensión de los hallazgos analíticos y permitan explorar los datos de forma dinámica.
3. **Comparar visualizaciones estáticas e interactivas**: Destacar cómo cada enfoque aporta valor en diferentes contextos de comunicación y toma de decisiones.

**Desarrollo**

**1. Conexión de Python con Power BI**

El primer paso consistió en conectar el dataset preparado con Power BI. Este proceso incluyó:

* Cargar el archivo **"DatosFinalesFiltrado.csv"** directamente en Power BI.
* Configurar las relaciones entre tablas en caso de utilizar datos adicionales o externos para enriquecer los análisis.
* Explorar la posibilidad de utilizar scripts de Python en Power BI para análisis adicionales o procesamiento dinámico.

Esta integración permite utilizar la capacidad de Power BI para combinar análisis técnico avanzado con visualización interactiva.

**2. Creación de Dashboards Interactivos**

Se diseñaron dashboards en Power BI que sintetizan los hallazgos clave del análisis. Los elementos más relevantes incluyen:

1. **Indicadores Clave de Rendimiento (KPIs)**:
   * Tasas de vacunación por país.
   * Incidencia acumulada de casos por cada 100,000 habitantes.
   * Mortalidad relacionada con el COVID-19.
2. **Mapa interactivo**:
   * Muestra la distribución geográfica de la incidencia de COVID-19 y las tasas de vacunación en los países analizados.
3. **Gráficos Interactivos**:
   * **Series temporales**: Evolución de casos y vacunación por país.
   * **Gráficos de dispersión**: Relación entre cobertura de vacunación y reducción de casos.
   * **Histogramas y boxplots**: Variabilidad en la incidencia entre países y su relación con factores demográficos.
   * **Mapas de calor**: Identificación de correlaciones entre variables clave.
4. **Filtros dinámicos**:
   * Por país, período de tiempo y métricas específicas (por ejemplo, incidencia semanal o mensual).
5. **Tablas resumidas**:
   * Resúmenes de datos clave para cada país, incluyendo porcentajes de población vacunada y casos activos.

Estos elementos interactivos proporcionan a los directivos la flexibilidad para explorar áreas específicas de interés y tomar decisiones fundamentadas.

**3. Visualizaciones Estáticas vs. Interactivas**

**Visualizaciones Estáticas**:

* **Ventajas**:
  + Claridad y simplicidad para reportes impresos.
  + Ideales para presentaciones en contextos formales.
* **Limitaciones**:
  + Falta de dinamismo y capacidad de profundización.

**Visualizaciones Interactivas**:

* **Ventajas**:
  + Permiten explorar múltiples capas de información.
  + Adaptables a diferentes audiencias y contextos.
  + Favorecen una comprensión más profunda a través de filtros y desagregaciones.
* **Limitaciones**:
  + Requieren acceso a plataformas digitales y conocimientos básicos de navegación en Power BI.

Power BI también ofrece la posibilidad de ejecutar scripts de Python, lo que permite integrar análisis más sofisticados directamente en los dashboards, aprovechando la sinergia entre ambas herramientas.

**Conclusiones**

1. **Utilidad de Power BI**: La plataforma facilita la visualización interactiva de grandes volúmenes de datos, permitiendo una exploración intuitiva por parte de los tomadores de decisiones.
2. **Dashboards como herramienta clave**: Las visualizaciones dinámicas destacan patrones y áreas prioritarias, maximizando el impacto del análisis previo.
3. **Estática vs. Interactiva**: Si bien las visualizaciones estáticas son útiles para reportes tradicionales, las interactivas ofrecen un nivel de detalle y adaptabilidad superior.

**Conclusión Final del Proyecto: Análisis y Expansión de Laboratorios Farmacéuticos en Latinoamérica**

A lo largo de este proyecto, se abordó un análisis exhaustivo y estructurado que permitió transformar un vasto conjunto de datos en una herramienta poderosa para la toma de decisiones estratégicas. Cada uno de los avances realizados contribuyó de manera significativa a los objetivos finales, generando insights clave que guiarán la expansión de laboratorios farmacéuticos en América Latina. A continuación, se destacan los aprendizajes más relevantes de cada fase:

**1. Carga y Transformación de los Datos**

El primer avance estableció las bases sólidas para el análisis. Se aprendió que:

* **La limpieza y estructuración de datos son esenciales** para trabajar con grandes volúmenes de información.
* Filtrar por regiones específicas y fechas relevantes no solo mejora la eficiencia del análisis, sino que también asegura que los resultados sean contextualmente pertinentes.
* La gestión adecuada de valores faltantes y la corrección de tipos de datos son pasos críticos para evitar sesgos y garantizar consistencia.

**2. Análisis Exploratorio y Visualización**

La segunda fase permitió profundizar en el conocimiento del dataset y generar representaciones visuales que facilitaron la identificación de patrones y tendencias. Se aprendió que:

* **La visualización de datos es una herramienta poderosa** para comunicar insights de manera clara y efectiva.
* Gráficos como histogramas, mapas de calor y diagramas de dispersión no solo ilustran las relaciones entre variables, sino que también permiten detectar anomalías y áreas de oportunidad.
* Las tasas de vacunación, la incidencia de casos y la infraestructura sanitaria se destacan como indicadores clave para priorizar regiones.

**3. Análisis Exploratorio Detallado**

El análisis avanzado con Pandas y Numpy permitió explorar con mayor precisión las dinámicas temporales y las correlaciones entre variables. Los aprendizajes incluyen:

* Las **series temporales revelan tendencias importantes**, como la estacionalidad en la incidencia de casos y la evolución de las tasas de vacunación.
* La relación entre la cobertura de vacunación y la reducción de casos confirma el impacto positivo de las campañas de vacunación en la contención de la pandemia.
* El análisis detallado de la mortalidad y otros indicadores socioeconómicos sugiere que factores como la urbanización y las condiciones preexistentes influyen significativamente en los resultados de salud.

**4. Aplicaciones Prácticas con Power BI**

La integración en Power BI permitió sintetizar el análisis en dashboards interactivos que facilitan la toma de decisiones. Se aprendió que:

* **Las herramientas de visualización interactiva son fundamentales** para convertir datos complejos en información accesible y útil para los directivos.
* Comparar visualizaciones estáticas con interactivas demuestra que la exploración dinámica de datos mejora la comprensión y la capacidad de respuesta estratégica.
* Presentar insights accionables, como la priorización de regiones con alta incidencia y baja cobertura de vacunación, es clave para el impacto empresarial y social.

**Conclusiones Generales**

1. **La importancia de los datos limpios y organizados**: Desde el inicio del proyecto, quedó claro que una preparación adecuada del dataset es esencial para garantizar la calidad del análisis.
2. **La relevancia de las visualizaciones y la narrativa de datos**: A lo largo del proyecto, se demostró cómo gráficos bien diseñados y análisis visuales pueden simplificar la comprensión de información compleja.
3. **Insights clave para la expansión estratégica**: Identificar regiones prioritarias basadas en datos objetivos ayuda a maximizar el impacto positivo en la salud pública, optimizando los recursos disponibles.
4. **El valor del análisis colaborativo**: Este proyecto destacó la necesidad de una colaboración interdisciplinaria entre analistas de datos, expertos en salud pública y estrategas empresariales.

**Impacto y Aplicaciones Futuras**

Este proyecto no solo permitió responder preguntas críticas sobre la expansión de laboratorios farmacéuticos, sino que también estableció un modelo replicable para futuros análisis en áreas relacionadas con la salud pública, la distribución de recursos y la planificación estratégica en contextos de alta incertidumbre.

Con las herramientas desarrolladas y los hallazgos obtenidos, la empresa farmacéutica cuenta con una base sólida para tomar decisiones informadas que maximicen tanto el impacto social como el retorno de la inversión.