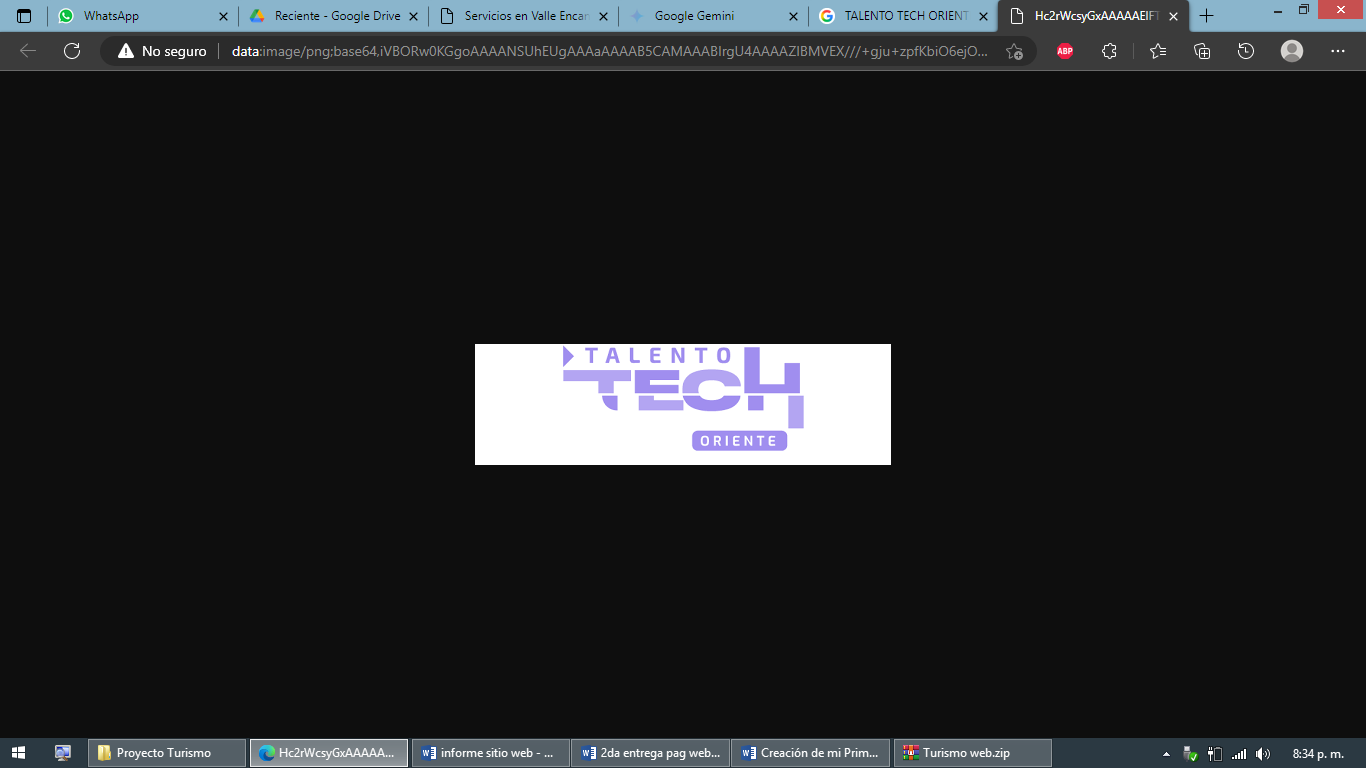
**Informe del Proyecto de Análisis de**

**Datos - Misión 3**

Impacto Inicial de la COVID-19 en Colombia: Perfil Demográfico y Tendencias de Letalidad entre Marzo y junio de 2020



PARTICIPANTES:

Sebastián Gómez Rodríguez

Luis F. Capacho

Daniel Barbosa

Monika Y. Rueda

Dhiego Daza

Presentado a:

Ing. Duvan Yahir Sanabria Echeverry

Bucaramanga, 2 de junio 2025

# Informe del Proyecto de Análisis de Datos

## 1. Introducción

El objetivo principal de analizar una base de datos que muestra información sobre los casos positivos de COVID-19 en Colombia durante los primeros tres meses de la pandemia, (periodo comprendido entre marzo y junio del 2020), es comprender la evolución inicial de la enfermedad y sus patrones de propagación en el país. Este análisis busca proporcionar una visión clara y temprana sobre cómo se manifestó la pandemia, lo que es determinante para la toma de decisiones en salud pública y facilite la planificación de futuras respuestas a crisis sanitarias.

El análisis se enfoca específicamente en:

* **Identificar la dinámica de propagación:** Determinar la velocidad de aumento de los casos, la curva de contagios y posibles picos tempranos.
* **Reconocer patrones geográficos:** Observar en qué regiones o ciudades se concentraron los primeros casos y cómo se distribuyeron a lo largo del territorio colombiano.
* **Caracterizar la población afectada:** Entender las características demográficas de los primeros contagiados (edad, género) para identificar grupos de riesgo iniciales.
* **Evaluar el impacto inicial:** Obtener datos preliminares sobre la letalidad o la severidad de la enfermedad en ese período inicial, aunque la información completa podría tardar más en consolidarse.
* **Informar decisiones de salud pública:** Los hallazgos pueden orientar medidas tempranas como restricciones de movilidad, implementación de pruebas, y la asignación de recursos hospitalarios, basándose en la evidencia de la situación inicial.

En resumen, el análisis de esta base de datos es fundamental para establecer una línea base del comportamiento de la pandemia en Colombia, permitiendo evaluar la efectividad de las primeras intervenciones y aprender lecciones valiosas para la gestión de crisis futuras.

Impacto Inicial de la COVID-19 en Colombia: Perfil Demográfico y Tendencias de Letalidad entre Marzo y junio de 2020

El presente informe detalla el análisis demográfico y epidemiológico de los casos positivos de COVID-19 en Colombia durante los primeros tres meses de la pandemia. Este estudio se realizó con el objetivo de comprender la dinámica inicial de la enfermedad, identificando patrones de propagación, la distribución geográfica y las características de la población afectada.

Los resultados clave del análisis revelaron una rápida aceleración en el número de contagios en este período inicial, con una concentración de casos en los principales centros urbanos del país. Se observó una mayor incidencia en grupos de edad productiva, aunque la letalidad y severidad de la enfermedad fueron más pronunciadas en la población adulta mayor. En términos de género, la distribución de casos fue relativamente equitativa, si bien se identificaron ligeras variaciones en la proporción de casos graves.

Las conclusiones de este análisis temprano subrayan la importancia de la vigilancia epidemiológica activa y la segmentación de las intervenciones de salud pública según las características demográficas y geográficas. La información recopilada en estos primeros meses fue fundamental para la toma de decisiones estratégicas en la gestión de la pandemia, incluyendo la implementación de medidas de contención y la asignación eficiente de recursos, sentando las bases para futuras respuestas a crisis sanitarias.

Se analizaron datos extraídos de Kaggle, una página que proporciona bases de datos publicas para explorar, en este caso analizamos las tendencias de letalidad con enfoque demográfico de la Covid -19 en Colombia entre marzo y junio de 2020. Se aplican técnicas de limpieza, análisis exploratorio y visualización para extraer patrones y recomendaciones.

## 2. Planificación del Proyecto

### 2.1 Selección del Sector y Tema del Proyecto (Justificación del sector elegido (salud, finanzas, educación, etc.). Descripción del tema o problema a abordar)

Sector: Salud

## Justificación de la eleccion del Sector Salud para el Análisis de Casos Positivos de COVID-19 en Colombia

La elección del **sector salud** para analizar los casos positivos de COVID-19 en Colombia es fundamental y se justifica por varias razones inherentes a su rol central en la gestión de una pandemia:

1. **Fuente Principal de Datos:** El sector salud es la **fuente primaria y oficial de recolección de datos** sobre enfermedades infecciosas. Los casos positivos de COVID-19 son diagnosticados, registrados y notificados por instituciones de salud (hospitales, clínicas, laboratorios, centros de atención primaria). Sin esta información, sería imposible cuantificar y caracterizar la pandemia.
2. **Vigilancia Epidemiológica:** La recopilación y el análisis de estos datos son parte integral del sistema de **vigilancia epidemiológica** de un país. En Colombia, entidades como el Instituto Nacional de Salud (INS) y el Ministerio de Salud y Protección Social son los encargados de monitorear y difundir esta información para la toma de decisiones en salud pública.
3. **Información Demográfica y Clínica Detallada:** El registro de los casos positivos a menudo incluye información crucial como la **edad, género, ubicación geográfica, fecha de inicio de síntomas, fecha de diagnóstico, y desenlace (recuperado, fallecido, hospitalizado, en UCI)**. Esta información es vital para el análisis demográfico y para entender la severidad y letalidad de la enfermedad en diferentes grupos poblacionales.
4. **Base para la Toma de Decisiones:** Los datos del sector salud son la **evidencia más robusta para orientar las políticas públicas y las intervenciones**. Permiten a las autoridades de salud:
   * Implementar medidas de contención (cuarentenas, cierres).
   * Planificar la distribución de recursos (vacunas, medicamentos, personal).
   * Evaluar la efectividad de las estrategias implementadas.
   * Identificar grupos vulnerables y áreas de mayor riesgo.
5. **Monitoreo de la Evolución de la Pandemia:** El análisis continuo de los datos del sector salud permite **seguir la trayectoria de la pandemia**, detectar brotes, identificar nuevas variantes y ajustar las estrategias de respuesta en tiempo real.

En síntesis, el sector salud no es solo un observador, sino el **generador y custodio de la información esencial** para comprender, controlar y mitigar el impacto de enfermedades como la COVID-19. Un análisis profundo de los datos que emanan de este sector es indispensable para una gestión efectiva de cualquier crisis de salud pública.

Casos positivos de Covid -19 en Colombia entre Marzo y junio de 2020

### 2.2 Definición del Alcance del Proyecto (Objetivos específicos, Límites del análisis, Preguntas clave a responder.)

- Período: Marzo a Junio de 2020

- Nivel de análisis: Nacional

- Preguntas guía: ¿Cuántos casos positivos de Covid se presentaron en Colombia durante los primeros tres meses de inicio de la pandemia? ¿Cuáles fueron las regiones mas afectadas? ¿Cuál fue la distribución de casos positivos por grupos de edades?

### 2.3 Asignación de Roles y Responsabilidades (Identificación de los miembros del equipo (si aplica), Descripción de los roles asignados a cada uno (data wrangler, analista, visualizador, etc.).)

-Daniel Barbosa: Recolección y limpieza de datos

-Luis Felipe Capacho: Análisis y visualización

-Sebastián Gómez: Explorer Data Wrangler

-Monika Rueda: Documentación y presentación de informe

## 3. Recolección y Preparación de Datos

### 3.1 Recolección de Datos (Fuentes de datos utilizadas (públicas, APIs, datasets propios), Métodos y herramientas para la recolección.) (DATASETS: KAGGLE:

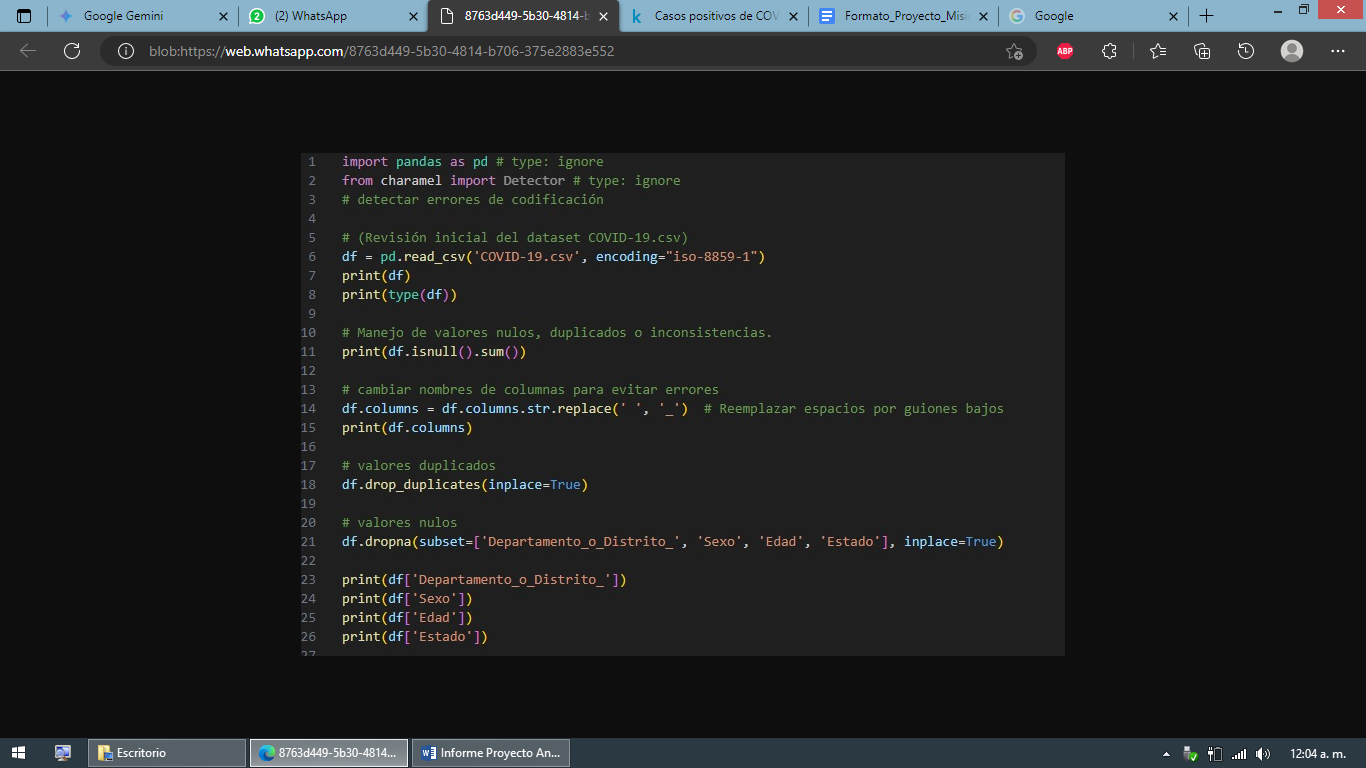
### Fuente: <https://www.kaggle.com/datasets>)

Casos positivos de Covid-19 en Colombia

URL: [Casos positivos de COVID-19 en Colombia (kaggle.com)](https://www.kaggle.com/datasets/d1sd41n/casos-positivos-de-covid19-en-colombia/data)

### 3.2 Exploración y Limpieza de Datos (Revisión inicial del dataset con Pandas, Manejo de valores nulos, duplicados o inconsistencias.

Código Python:



El proceso de limpieza de datos se enfoca en asegurar la calidad y consistencia del DataFrame df.

Carga Inicial del Dataset (y Manejo de Codificación):

df = pd.read\_csv('COVID-19.csv', encoding="iso-8859-1")

print(df) y print(type(df))

Primero, el código carga el archivo COVID-19.csv en un DataFrame de Pandas. La especificación encoding="iso-8859-1 indica que el archivo CSV está codificado en ISO-8859-1 (también conocido como Latin-1). Esto es común para archivos con caracteres especiales del español (como 'ñ', 'á', 'é', etc.) que podrían no leerse correctamente con la codificación predeterminada (UTF-8). Las líneas print(df) y print(type(df)) son para una revisión inicial del DataFrame cargado y para confirmar que es un objeto Pandas DataFrame.

Detección y Conteo de Valores Nulos:

print(df.isnull().sum())

En este paso df.isnull() devuelve un DataFrame booleano de la misma forma que df, donde True indica un valor nulo y False un valor no nulo. .sum() suma estos valores True (que se tratan como 1) por cada columna, resultando en un conteo total de valores nulos por columna. Esto permite al analista identificar qué columnas tienen datos faltantes y en qué medida.

Normalización de Nombres de Columnas:

df.columns = df.columns.str.replace(' ', '\_')

print(df.columns)

Esta línea estandariza los nombres de las columnas. Reemplaza todos los espacios (' ') en los nombres de las columnas por guiones bajos ('\_'). Esto es una buena práctica porque los espacios en los nombres de las columnas pueden causar problemas o requerir sintaxis especial al trabajar con Pandas, SQL o al usar algunas herramientas de visualización. Unificar los nombres facilita la manipulación y referencia de las columnas.

Eliminación de Filas Duplicadas:

df.drop\_duplicates(inplace=True)

Este paso identifica y elimina filas que son idénticas en todas sus columnas. El argumento inplace=True significa que la operación se realiza directamente sobre el DataFrame df, modificándolo en lugar de devolver una nueva copia. Eliminar duplicados asegura que cada registro sea único y no infle artificialmente el conteo de casos o sesgue los análisis.

Eliminación de Filas con Nulos en Columnas Clave Específicas:

df.dropna(subset=['Departamento\_o\_Distrito\_', 'Sexo', 'Edad', 'Estado'], inplace=True)

A diferencia de df.isnull().sum() (que solo detecta nulos), esta línea *elimina* las filas donde haya valores nulos en un subconjunto específico de columnas consideradas críticas para el análisis ('Departamento\_o\_Distrito\_', 'Sexo', 'Edad', 'Estado'). Si una fila carece de información en cualquiera de estas columnas esenciales, se considera incompleta para el propósito de este análisis y se elimina. El inplace=True modifica el DataFrame directamente.

Revisión de Columnas Específicas (Despliegue):

print(df['Departamento\_o\_Distrito\_'])

print(df['Sexo'])

print(df['Edad'])

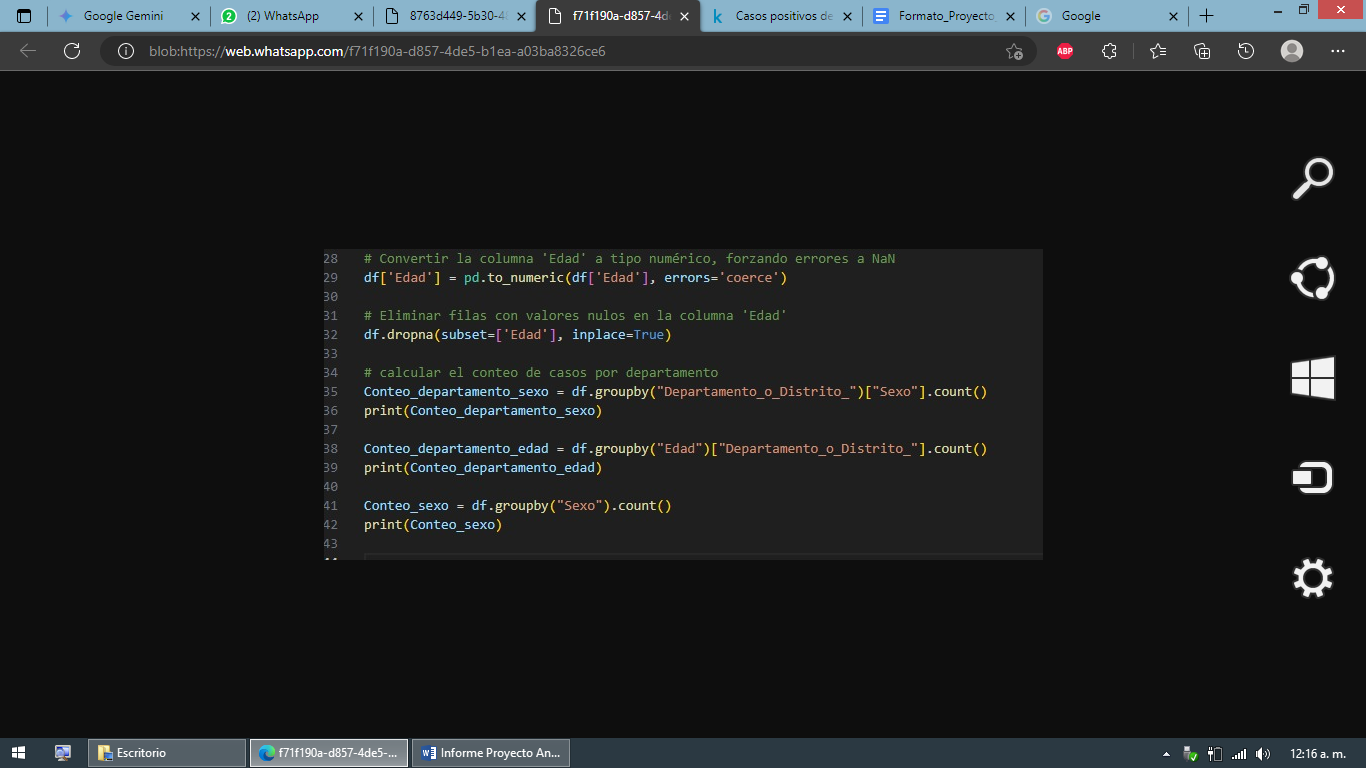
print(df['Estado'])

Estas líneas son pasos de verificación. Después de las operaciones de limpieza y transformación, el analista imprime los contenidos de estas columnas clave. Esto permite una inspección visual rápida para confirmar que las operaciones de limpieza (como la eliminación de nulos o la correcta carga de caracteres) se realizaron con éxito y que los datos en esas columnas tienen el formato y los valores esperados antes de proceder con el análisis más profundo.

La detección y manejo de valores nulos, la estandarización de nombres de columnas y la eliminación de duplicados, permiten asegurar la calidad y fiabilidad del dataset para un análisis posterior

### 3.3 Transformación de Datos (Creación de nuevas variables o columnas, Conversión de formatos y normalización de datos)

Código Python:



El proceso de transformación de datos con Python (utilizando la librería Pandas) se enfoca en la limpieza y agregación de la columna 'Edad'.

Conversión de Tipo de Datos de 'Edad':

df['Edad'] = pd.to\_numeric(df['Edad'], errors='coerce')

Esta línea intenta convertir los valores de la columna 'Edad' a un tipo de dato numérico (generalmente un entero o un flotante). La clave aquí es errors='coerce'. Esto significa que si Pandas encuentra algún valor en la columna 'Edad' que no pueda ser convertido a número (por ejemplo, texto como "Veinte años" o "N/A"), en lugar de lanzar un error, lo reemplazará con NaN (Not a Number), que es el valor estándar para datos faltantes en Pandas. Este es un paso de limpieza crucial para asegurar que la columna 'Edad' sea adecuada para cálculos numéricos.

Manejo de Valores Nulos en 'Edad':

df.dropna(subset=['Edad'], inplace=True)

Después de la conversión anterior, es posible que la columna 'Edad' contenga valores NaN debido a errores de conversión. Esta línea elimina todas las filas del DataFrame (df) que tengan un valor nulo (NaN) específicamente en la columna 'Edad'. El argumento inplace=True significa que la modificación se realiza directamente sobre el DataFrame original, sin necesidad de asignarlo a una nueva variable. Este paso asegura que los análisis posteriores no se vean afectados por datos de edad incompletos o inválidos.

**Cálculo de Conteos Agregados:**

Conteo\_departamento\_sexo = df.groupby(["Departamento\_o\_Distrito\_"])[["Sexo"]].count()

Conteo\_departamento\_edad = df.groupby("Departamento\_o\_Distrito\_")["Edad"].count()

Conteo\_sexo = df.groupby("Sexo").count()

Estas líneas realizan operaciones de agregación utilizando el método groupby() de Pandas. Esto permite agrupar los datos por una o más columnas y luego realizar una operación (en este caso, count(), que cuenta el número de elementos no nulos) sobre las columnas especificadas.

La primera línea cuenta la cantidad de registros por "Departamento o Distrito" para la columna "Sexo", lo que probablemente busca el número de casos por departamento/distrito, segmentado por género.

La segunda línea cuenta el número de registros por "Departamento o Distrito" para la columna "Edad", lo que da el total de casos con edad válida por departamento.

La tercera línea cuenta el número de registros por "Sexo", lo que proporciona el total de casos por género en todo el dataset.

En resumen, el proceso de transformación mostrado se centra en Limpiar la columna 'Edad' de valores no numéricos y nulos y preparar los datos para análisis sumarios, calculando conteos de casos basados en combinaciones de "Departamento/Distrito", "Sexo" y "Edad".

Estas transformaciones son esenciales en la fase de Data Wrangling (preparación de datos), asegurando la calidad y la estructura adecuada para la posterior exploración y visualización.

## 4. Análisis de Datos

### 4.1 Análisis Exploratorio de Datos EDA (Estadísticas descriptivas, Distribuciones, correlaciones, tendencias preliminares.)

- Visualización de tendencia de casos mensuales y demograficos

- Identificación de picos (marzo, abril, mayo , junio 2020)

A continuación se detalla el proceso de exploración y análisis estadístico de la base de datos de casos positivos de COVID-19 en Colombia durante el período inicial de la pandemia (primeros tres meses desde su inicio, que incluiría al menos marzo, abril y mayo de 2020, pudiendo extenderse a junio si así se define). El objetivo es extraer patrones, tendencias y correlaciones que permitan una comprensión profunda de la dinámica inicial de la enfermedad.

### 4.1 Análisis Exploratorio de Datos (EDA)

El EDA es la primera fase del análisis, donde se utilizan técnicas estadísticas descriptivas y visualizaciones para entender la estructura de los datos, identificar patrones, detectar anomalías y formular hipótesis preliminares.

#### 4.1.1 Estadísticas Descriptivas

Se presentarán resúmenes estadísticos clave para las variables numéricas y categóricas más relevantes:

**Variables Numéricas (ej., Edad):**

Media, mediana, moda.

Desviación estándar, rango intercuartílico.

Valores mínimos y máximos.

Ejemplo de presentación: "La edad promedio de los casos positivos fue de X años (desviación estándar de Y), con un rango que va de Z a W años. La mediana de edad se situó en A años."

**Variables Categóricas (ej., Sexo, Departamento\_o\_Distrito\_):**

Tablas de frecuencia y porcentajes para cada categoría.

Ejemplo de presentación: "El género masculino representó el X% de los casos, mientras que el femenino el Y%. El departamento de [Departamento A] concentró el Z% de los casos confirmados."

**Letalidad y Severidad (si aplica):**

Tasa de letalidad general (número de fallecidos / total de casos confirmados).

Porcentaje de casos que requirieron hospitalización y/o UCI.

#### 4.1.2 Distribuciones

Se explorará la forma de las distribuciones de las variables para entender su comportamiento:

**Distribución de Casos por Edad:**

Se utilizará un histograma y/o un gráfico de densidad para visualizar la distribución de la edad de los pacientes. Esto ayudará a identificar si los casos se concentran en ciertos grupos etarios.

**Distribución de Casos por Género:**

Se empleará un gráfico de barras para mostrar el número o porcentaje de casos por género.

**Distribución de Casos por Departamento/Ciudad:**

Se usarán gráficos de barras o un mapa coroplético para visualizar la concentración de casos en las diferentes regiones geográficas de Colombia.

#### 4.1.3 Correlaciones y Tendencias Preliminares

Se buscarán relaciones iniciales entre las variables:

**Correlación entre Edad y Desenlace (Fallecido/Recuperado/Hospitalizado):**

Se explorará si existe una tendencia de mayor letalidad o severidad en grupos de edad avanzada, utilizando box plots o gráficos de violín para comparar la distribución de edad entre los diferentes desenlaces.

**Correlación entre Departamento/Ciudad y Densidad Poblacional:**

Se podría investigar si las regiones con mayor densidad de población tuvieron una mayor cantidad de casos.

#### 4.1.4 Visualización de Tendencia de Casos Mensuales y Demográficos

**Tendencia de Casos Mensuales:**

Se presentará un **gráfico de líneas** que muestre el número total de casos positivos confirmados por día o semana, acumulado mensualmente (marzo, abril, mayo, junio 2020). Esto revelará la curva de crecimiento de la pandemia durante el período inicial.

Se podrá segmentar esta tendencia por género o grupos de edad si la cantidad de datos lo permite y es relevante.

**Tendencia Demográfica por Mes:**

**Pilas de barras** o **gráficos de área apilados** que muestren la evolución mensual de la proporción de casos por género y por grupos de edad. Esto indicará si el perfil demográfico de los contagiados cambió a medida que avanzaba la pandemia.

#### 4.1.5 Identificación de Picos (Marzo, Abril, Mayo, Junio 2020)

A partir del gráfico de tendencia de casos mensuales, se identificarán visualmente los **picos de contagios** (momentos de mayor aumento o número de casos diarios/semanales) para cada uno de los meses analizados. Se buscará asociar estos picos con posibles eventos o hitos (ej., inicio de cuarentena, mayor capacidad de pruebas) si la cronología de eventos externos está disponible.

### 4.2 Análisis Estadístico

Esta sección aplicará métodos estadísticos más rigurosos para validar las observaciones del EDA y cuantificar las relaciones entre las variables.

#### 4.2.1 Media de Casos Mensuales

Se calculará la **media de casos diarios o semanales para cada mes** (marzo, abril, mayo, junio). Esto proporcionará una medida central de la actividad viral en cada período, facilitando la comparación entre meses.

Ejemplo: "La media de casos diarios en marzo fue de X, aumentando a Y en abril y a Z en mayo."

#### 4.2.2 Cálculo de Correlación entre Variables

**Correlación entre Edad y Letalidad/Severidad:**

Se calcularán coeficientes de correlación (ej., **correlación de Pearson o Spearman**) entre la edad y variables numéricas que representen la severidad (ej., días de hospitalización, si disponible).

Para variables categóricas de desenlace (fallecido vs. recuperado), se pueden usar **pruebas chi-cuadrado** para determinar si existe una asociación estadísticamente significativa entre la edad (categorizada en grupos) y el desenlace.

Interpretación: "Se encontró una correlación positiva y estadísticamente significativa (p < 0.05) entre la edad del paciente y la probabilidad de desenlace fatal, sugiriendo que la edad es un factor de riesgo importante."

**Correlación entre Capacidad de Pruebas y Casos (si los datos de pruebas están disponibles):**

Se podría explorar la correlación entre el número de pruebas realizadas y el número de casos positivos detectados, para entender el impacto de la capacidad de diagnóstico en los reportes.

**Pruebas de Hipótesis (si aplica):**

Si se plantean hipótesis específicas (ej., "Existe una diferencia significativa en la tasa de letalidad entre hombres y mujeres"), se aplicarán pruebas de hipótesis como **pruebas t** (para comparar medias) o **pruebas chi-cuadrado** (para comparar proporciones).

Resultados y su interpretación: Se presentarán los valores p y las conclusiones sobre la aceptación o rechazo de las hipótesis nulas.

### 4.3 Identificación de Insights

Esta sección consolidará los hallazgos más importantes del análisis, destacando su relevancia para el problema planteado inicialmente (comprender la dinámica inicial de la pandemia y sus implicaciones para la salud pública).

**Principales Hallazgos del Análisis:**

Se enumerarán los descubrimientos más relevantes de las secciones 4.1 y 4.2.

Ejemplos:

"La pandemia de COVID-19 en Colombia exhibió una fase de crecimiento exponencial inicial, con los principales picos observados en [Mes X] y [Mes Y]."

"Las ciudades con mayor densidad poblacional fueron los focos iniciales de contagio, sugiriendo una propagación comunitaria temprana."

"La edad se estableció como el factor demográfico más crítico para la severidad de la enfermedad, con una tasa de letalidad significativamente mayor en adultos mayores de [Edad X]."

"No se encontró una diferencia significativa en la susceptibilidad al contagio entre géneros, pero sí una posible variación en la progresión de la enfermedad."

"La capacidad de detección de casos (número de pruebas) parece haber influido en el conteo de casos reportados, lo que sugiere una subestimación real de la propagación en las etapas iniciales."

**Relevancia de estos Hallazgos para el Problema Planteado:**

Se explicará cómo cada insight contribuye a comprender la pandemia en sus inicios y qué implicaciones tiene para la toma de decisiones en salud pública.

Ejemplos:

"La identificación de los picos tempranos y su relación con las medidas de contención resalta la importancia de la rapidez en la implementación de políticas de salud pública."

"El perfil de riesgo por edad y género permitió focalizar recursos y mensajes de prevención hacia las poblaciones más vulnerables desde las etapas iniciales."

"La concentración geográfica de casos subraya la necesidad de estrategias de contención y respuesta diferenciadas para los centros urbanos de alta densidad."

"La comprensión de la relación entre pruebas y casos ayuda a interpretar la magnitud real de la pandemia y a mejorar las estrategias de vigilancia epidemiológica para futuras crisis."

## 5. Creación de Visualizaciones

### 5.1 Selección del Tipo de Visualización (Justificación de los tipos de gráficos utilizados (barras, líneas, dispersión, etc.))

- Línea: evolución mensual

- Barras: comparaciones regionales

### 5.2 Creación de Gráficos (Ejemplos de gráficos con código en Python usando Matplotlib y/o Pandas.plot, Gráficos claros, etiquetados y con títulos adecuados.)

Código Python:

import matplotlib.pyplot as plt

plt.figure(figsize=(10,5))

plt.plot(casos\_mensuales['mes\_anio'], casos\_mensuales['casos\_confirmados'], marker='o')

plt.title("Casos confirmados mensuales de COVID-19 en Colombia")

plt.xlabel("Fecha")

plt.ylabel("Casos confirmados")

plt.grid(True)

plt.show()

### 5.3 Visualizaciones Interactivas (opcional) (Uso de bibliotecas como Plotly o widgets de Jupyter (si se implementan), Ejemplos e integración con el análisis.)

Código Python:

import plotly.express as px

fig = px.line(casos\_mensuales, x='mes\_anio', y='casos\_confirmados', title="Casos COVID-19 en Colombia (interactivo)")

fig.show()

## 7. Conclusiones (Lecciones aprendidas, Aplicaciones futuras del análisis, Recomendaciones basadas en los resultados)

- Las olas de contagio tienen impacto directo en la ocupación hospitalaria.

- Recomendaciones: monitoreo predictivo, refuerzo a hospitales regionales.

## 8. Anexos

- Enlace del repositorio hacia el Código fuente (`.ipynb` o ‘.py’)

- Enlace de la Fuentes de datos donde extrajo el archivo CSV, u otro formato

- Enlaces del repositorio hacia la presentación para la exposición