UN RECORRIDO POR LA EVOLUCIÓN TEMPORAL Y GEOGRÁFICA DE LA PANDEMIA (2020-2021)

Análisis Exploratorio de Datos (EDA)

Covid-19 en Estados Unidos

Descripción del proyecto

Este proyecto realiza un análisis exploratorio de datos sobre la evolución del Covid-19 en Estados Unidos. El objetivo es limpiar, explorar y analizar los datos para obtener conclusiones claras y comprensibles que permitan interpretar el comportamiento de la pandemia a lo largo del tiempo.

Objetivos

- Comprender la estructura y calidad del dataset
- Realizar la limpieza y el preprocesado de los datos
- Analizar la evolución temporal de los casos y fallecimientos
- Generar visualizaciones que ayuden a comunicar los resultados obtenidos

1. Carga del dataset

Importamos pandas y cargamos el dataset

2. Exploración inicial del dataset

Se visualizan las primeras filas del dataset para comprender su estructura y las variables disponibles.

	date	state	death	deathConfirmed	deathIncrease	deathProbable	hosp
0	2021-03-07	AK	305.0	NaN	0	NaN	
1	2021-03-07	AL	10148.0	7963.0	-1	2185.0	
2	2021-03-07	AR	5319.0	4308.0	22	1011.0	
3	2021-03-07	AS	0.0	NaN	0	NaN	
4	2021-03-07	AZ	16328.0	14403.0	5	1925.0	
5	2021-03-07	CA	54124.0	NaN	258	NaN	
6	2021-03-07	CO	5989.0	5251.0	3	735.0	
7	2021-03-07	CT	7704.0	6327.0	0	1377.0	
8	2021-03-07	DC	1030.0	NaN	0	NaN	
9	2021-03-07	DE	1473.0	1337.0	9	136.0	

10 rows × 41 columns

Se analiza el tamaño del dataset cuantas filas tiene y cuantas columnas

```
(20780, 41)
```

Esta inspección permite identificar tipos de datos y posibles valores nulos.

```
<class 'pandas.DataFrame'>
RangeIndex: 20780 entries, 0 to 20779
Data columns (total 41 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   date                                20780 non-null  str
 1   state                               20780 non-null  str
 2   death                               19930 non-null  float64
 3   deathConfirmed                      9422 non-null   float64
 4   deathIncrease                       20780 non-null  int64
 5   deathProbable                       7593 non-null   float64
 6   hospitalized                        12382 non-null  float64
 7   hospitalizedCumulative              12382 non-null  float64
 8   hospitalizedCurrently               17339 non-null  float64
 9   hospitalizedIncrease                20780 non-null  int64
 10  inIcuCumulative                     3789 non-null   float64
 11  inIcuCurrently                      11636 non-null  float64
 12  negative                            13290 non-null  float64
 13  negativeIncrease                    20780 non-null  int64
 14  negativeTestsAntibody               1458 non-null   float64
 15  negativeTestsPeopleAntibody         972 non-null    float64
 16  negativeTestsViral                  5024 non-null   float64
 17  onVentilatorCumulative              1290 non-null   float64
 18  onVentilatorCurrently               9126 non-null   float64
 19  positive                            20592 non-null  float64
 20  positiveCasesViral                  14246 non-null  float64
 21  positiveIncrease                    20780 non-null  int64
 22  positiveScore                       20780 non-null  int64
 23  positiveTestsAntibody               3346 non-null   float64
 24  positiveTestsAntigen                2233 non-null   float64
 25  positiveTestsPeopleAntibody         1094 non-null   float64
 26  positiveTestsPeopleAntigen          633 non-null    float64
 27  positiveTestsViral                  8958 non-null   float64
 28  recovered                           12003 non-null  float64
 29  totalTestEncountersViral            5231 non-null   float64
 30  totalTestEncountersViralIncrease    20780 non-null  int64
 31  totalTestResults                    20614 non-null  float64
 32  totalTestResultsIncrease            20780 non-null  int64
 33  totalTestsAntibody                  4789 non-null   float64
 34  totalTestsAntigen                   3421 non-null   float64
 35  totalTestsPeopleAntibody            2200 non-null   float64
 36  totalTestsPeopleAntigen             999 non-null    float64
 37  totalTestsPeopleViral               9197 non-null   float64
 38  totalTestsPeopleViralIncrease       20780 non-null  int64
 39  totalTestsViral                     14516 non-null  float64
 40  totalTestsViralIncrease             20780 non-null  int64
dtypes: float64(30), int64(9), str(2)
memory usage: 6.5 MB
```

Resume las columnas númericas

	death	deathConfirmed	deathIncrease	deathProbable	hospitalized
count	19930.000000	9422.000000	20780.000000	7593.000000	12382.000000
mean	3682.216859	3770.182764	24.790712	417.291321	9262.762478
std	6281.366321	4157.640633	60.162742	537.625982	12620.544081
min	0.000000	0.000000	-201.000000	0.000000	1.000000
25%	161.250000	607.000000	0.000000	79.000000	985.250000
50%	1108.000000	2409.500000	6.000000	216.000000	4472.000000
75%	4387.500000	5462.000000	24.000000	460.000000	12248.500000
max	54124.000000	21177.000000	2559.000000	2594.000000	82237.000000

8 rows × 39 columns

Esto ayuda a saber que variables hay y elegir cual vamos analizar

```
Index(['date', 'state', 'death', 'deathConfirmed', 'deathIncrease',
       'deathProbable', 'hospitalized', 'hospitalizedCumulative',
       'hospitalizedCurrently', 'hospitalizedIncrease', 'inIcuCumulative',
       'inIcuCurrently', 'negative', 'negativeIncrease',
       'negativeTestsAntibody', 'negativeTestsPeopleAntibody',
       'negativeTestsViral', 'onVentilatorCumulative', 'onVentilatorCurrently',
       'positive', 'positiveCasesViral', 'positiveIncrease', 'positiveScore',
       'positiveTestsAntibody', 'positiveTestsAntigen',
       'positiveTestsPeopleAntibody', 'positiveTestsPeopleAntigen',
       'positiveTestsViral', 'recovered', 'totalTestEncountersViral',
       'totalTestEncountersViralIncrease', 'totalTestResults',
       'totalTestResultsIncrease', 'totalTestsAntibody', 'totalTestsAntigen',
       'totalTestsPeopleAntibody', 'totalTestsPeopleAntigen',
       'totalTestsPeopleViral', 'totalTestsPeopleViralIncrease',
       'totalTestsViral', 'totalTestsViralIncrease'],
      dtype='str')
```

3. Preparación y transformación de datos

Miramos que tipo de dato es `date` aunque ya sabemos que es string tenemos que convertir a número.

```
<StringDtype(storage='python', na_value=nan)>
```

La columna `date`, que inicialmente estaba almacenada como texto, se convierte al tipo datetime para permitir un análisis temporal correcto. Aqui decimos que interprete la columna date con fechas reales

Se valida que la conversión de la columna `date` se ha realizado correctamente.

```
dtype('<M8[us]')
```

Se ordenan los registros cronológicamente en función de la columna `date` para garantizar la coherencia del análisis temporal.

Se verifican las primeras filas del dataset para confirmar que la ordenación se ha realizado correctamente.

	date	state	death	deathConfirmed	deathIncrease	deathProbable	hc
20779	2020-01-13	WA	NaN	NaN	0	NaN	
20778	2020-01-14	WA	NaN	NaN	0	NaN	
20777	2020-01-15	WA	NaN	NaN	0	NaN	
20776	2020-01-16	WA	NaN	NaN	0	NaN	
20775	2020-01-17	WA	NaN	NaN	0	NaN	

5 rows × 41 columns

Se verifican las últimas filas del dataset para confirmar que la ordenación se ha realizado correctamente.

	date	state	death	deathConfirmed	deathIncrease	deathProbable	hos
32	2021-03-07	NE	2113.0	NaN	0	NaN	
31	2021-03-07	ND	1478.0	NaN	0	NaN	
30	2021-03-07	NC	11502.0	10169.0	0	1333.0	
28	2021-03-07	MS	6808.0	4737.0	3	2071.0	
0	2021-03-07	AK	305.0	NaN	0	NaN	

5 rows × 41 columns

4. Selección de variables relevantes

Se revisan nuevamente los nombres de las columnas con el objetivo de seleccionar las variables relevantes para el análisis.

```
Index(['date', 'state', 'death', 'deathConfirmed', 'deathIncrease',
       'deathProbable', 'hospitalized', 'hospitalizedCumulative',
       'hospitalizedCurrently', 'hospitalizedIncrease', 'inIcuCumulative',
       'inIcuCurrently', 'negative', 'negativeIncrease',
       'negativeTestsAntibody', 'negativeTestsPeopleAntibody',
       'negativeTestsViral', 'onVentilatorCumulative', 'onVentilatorCurrently',
       'positive', 'positiveCasesViral', 'positiveIncrease', 'positiveScore',
       'positiveTestsAntibody', 'positiveTestsAntigen',
       'positiveTestsPeopleAntibody', 'positiveTestsPeopleAntigen',
       'positiveTestsViral', 'recovered', 'totalTestEncountersViral',
       'totalTestEncountersViralIncrease', 'totalTestResults',
       'totalTestResultsIncrease', 'totalTestsAntibody', 'totalTestsAntigen',
       'totalTestsPeopleAntibody', 'totalTestsPeopleAntigen',
       'totalTestsPeopleViral', 'totalTestsPeopleViralIncrease',
       'totalTestsViral', 'totalTestsViralIncrease'],
      dtype='str')
```

5. Ingeniería de variables temporales

Se crean variables temporales (año y mes) a partir de la fecha con el objetivo de facilitar el análisis temporal y la construcción de visualizaciones que permitan identificar tendencias a lo largo del tiempo.

Creamos una nueva columna `year` usando el año que hay dentro de la columna `date`

Creamos una nueva columna `month` usando el mes que hay dentro de la columna `date`

Se seleccionan las variables principales relacionadas con la evolución de los casos, fallecimientos y hospitalizaciones por Covid-19.

6. Tratamiento de valores nulos

Se analiza la presencia de valores nulos en cada columna del dataset seleccionado con el objetivo de evaluar la calidad de los datos antes de su tratamiento.

```
date                      0
state                     0
positive                188
positiveIncrease          0
death                   850
deathIncrease             0
hospitalizedCurrently  3441
year                      0
month                     0
dtype: int64
```

Se identifican valores nulos en las variables acumulativas (casos positivos, fallecimientos). Dado que estos datos son históricos, los nulos se interpretan como una ausencia de actualización en el reporte de ese día.

Por tanto, para mantener la coherencia de la serie temporal, se procede a imputar los valores faltantes propagando el último dato conocido (método forward fill), asumiendo que la cifra acumulada se mantiene si no hay nuevo reporte. Los valores nulos restantes (al inicio de la serie) se rellenan con cero

Verificamos que todas las columnas ya no tengan valores nulos

```
date                    0
state                   0
positive                0
positiveIncrease        0
death                   0
deathIncrease           0
hospitalizedCurrently   0
year                    0
month                   0
dtype: int64
```

7. Validación del dataset preparado para análisis

Se visualizan las columnas `date` , `year` y `month` para verificar que las variables temporales se han generado correctamente.

	date	year	month
20779	2020-01-13	2020	1
20778	2020-01-14	2020	1
20777	2020-01-15	2020	1
20776	2020-01-16	2020	1
20775	2020-01-17	2020	1

Se visualizan las primeras filas del dataset de análisis para comprobar que la selección de variables y la limpieza de datos se han realizado correctamente.

	date	state	positive	positiveIncrease	death	deathIncrease	hospital
20779	2020-01-13	WA	0.0	0	0.0	0	
20778	2020-01-14	WA	0.0	0	0.0	0	
20777	2020-01-15	WA	0.0	0	0.0	0	
20776	2020-01-16	WA	0.0	0	0.0	0	
20775	2020-01-17	WA	0.0	0	0.0	0	
20774	2020-01-18	WA	0.0	0	0.0	0	
20773	2020-01-19	WA	1.0	1	0.0	0	
20772	2020-01-20	WA	1.0	0	0.0	0	
20771	2020-01-21	WA	2.0	1	0.0	0	
20770	2020-01-22	WA	2.0	0	0.0	0	

8. Análisis exploratorio y visualización de datos

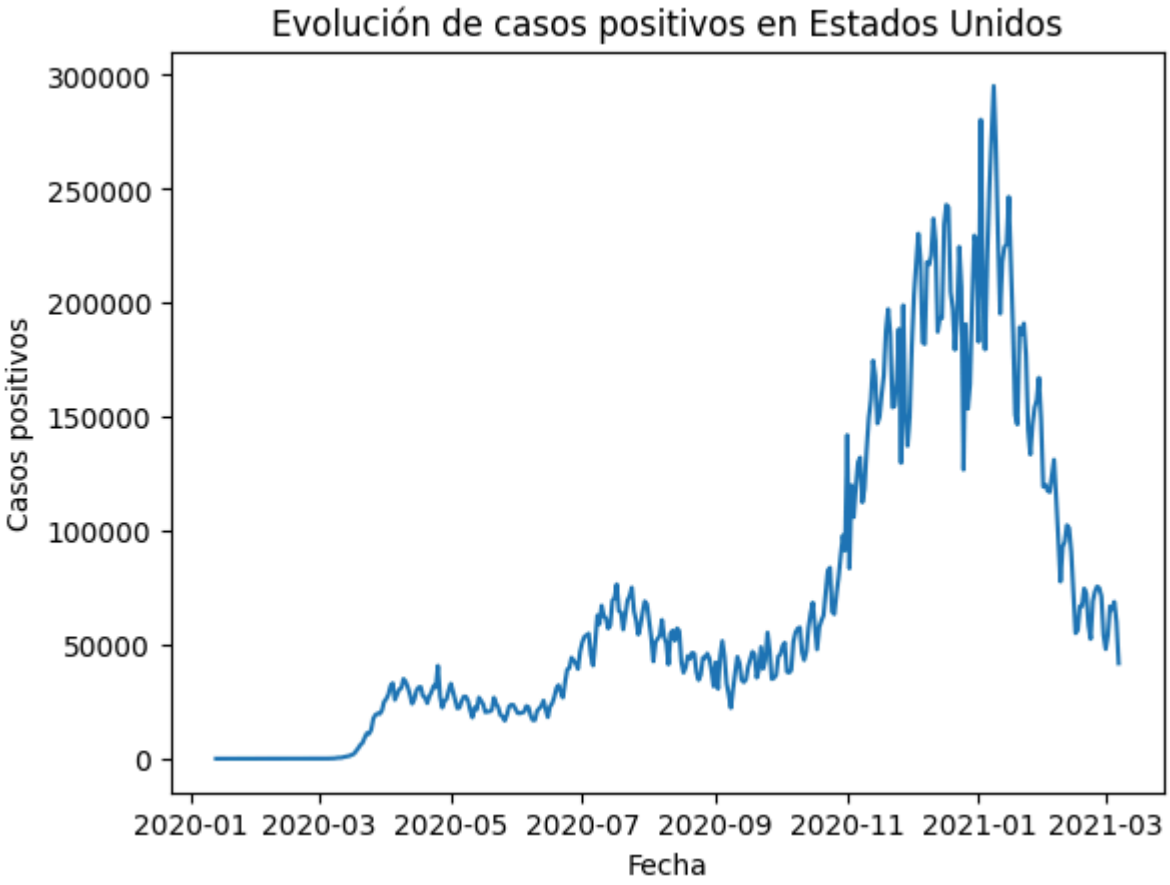
En esta sección se exploran las relaciones y tendencias principales del dataset mediante visualizaciones que permitan identificar patrones en la evolución temporal de la pandemia.

8.1 Evolución temporal de los casos positivos

Se analiza la evolución temporal del número de casos positivos con el objetivo de identificar tendencias, picos de contagio y periodos de mayor incidencia de la pandemia.

Se agrupan los casos positivos por fecha con el objetivo de analizar la evolución temporal del número total de contagios en Estados Unidos.

	date	positiveIncrease
0	2020-01-13	0
1	2020-01-14	0
2	2020-01-15	0
3	2020-01-16	0
4	2020-01-17	0
5	2020-01-18	0
6	2020-01-19	1
7	2020-01-20	0
8	2020-01-21	1
9	2020-01-22	0

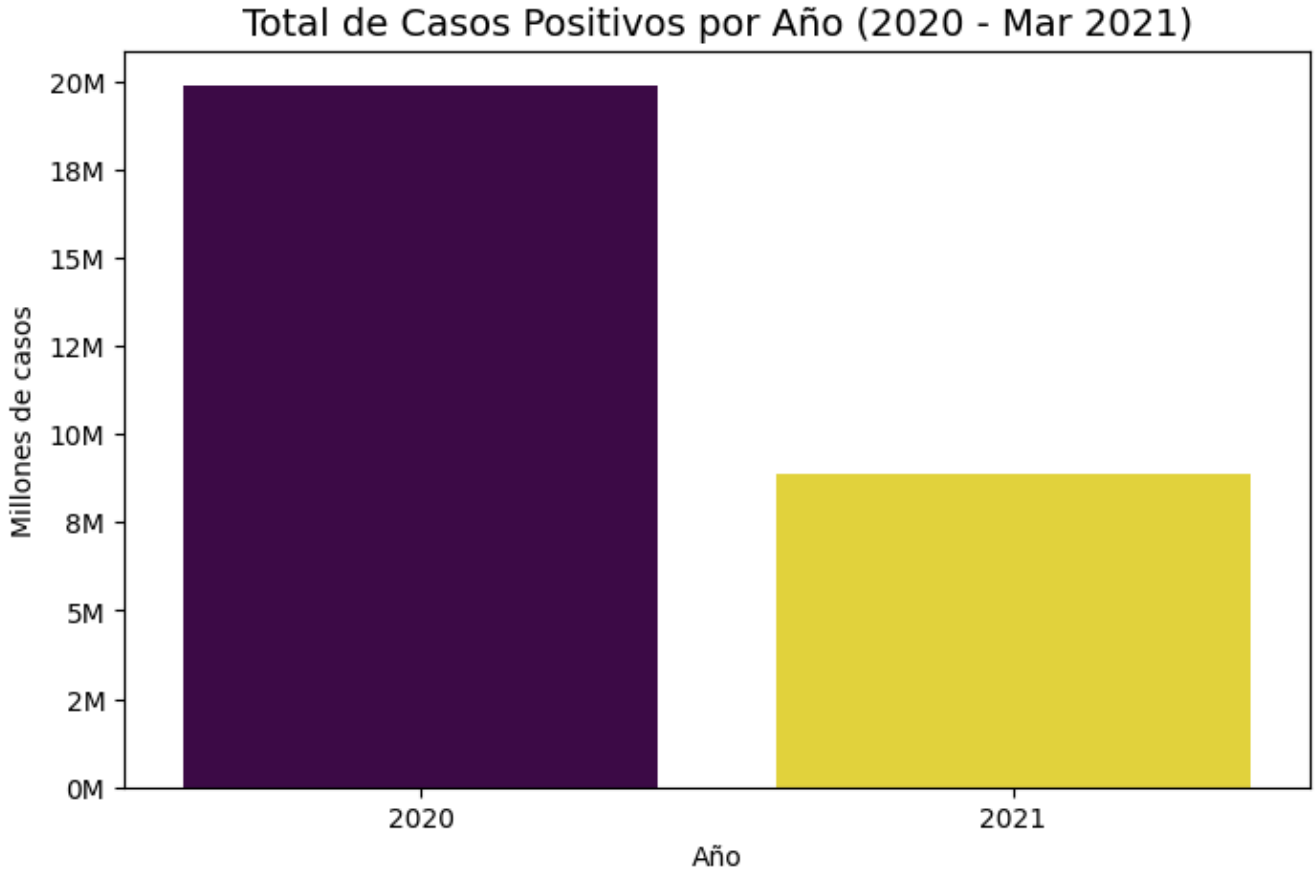


8.1.1 Análisis comparativo de casos por año

Para comprender la magnitud global de la pandemia en los distintos periodos, agrupamos los datos anualmente.

> **Nota técnica:** Es importante recordar que el dataset finaliza en **marzo de 2021**, por lo que los datos de ese año no representan el año completo, sino solo el primer trimestre. Además, utilizamos la variable `positiveIncrease` (casos nuevos) en lugar de la acumulada para obtener la suma real de contagios en cada periodo.

	year	positiveIncrease
0	2020	19864278
1	2021	8892115

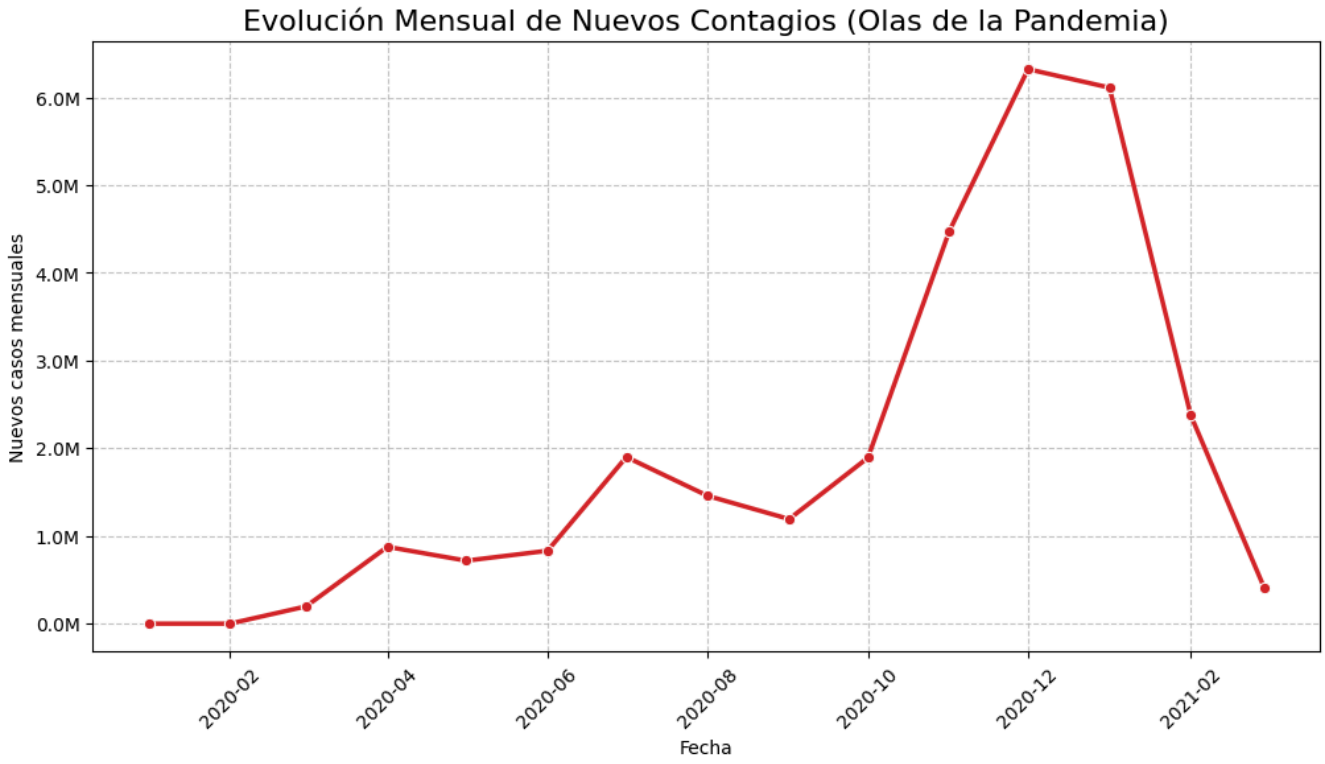


8.1.2 Evolución mensual: La curva epidémica

El análisis anual nos ofrece una visión general, pero oculta la dinámica real de los contagios. Para identificar patrones estacionales y las diferentes "olas" de la pandemia, es necesario aumentar la granularidad del análisis.

En esta sección, agrupamos los nuevos casos (`positiveIncrease`) por mes y año. Esto nos permitirá visualizar la velocidad de transmisión del virus a lo largo del tiempo.

	year	month	positiveIncrease
0	2020	1	2
1	2020	2	16
2	2020	3	196851
3	2020	4	876279
4	2020	5	718205
5	2020	6	831597
6	2020	7	1900180
7	2020	8	1457213
8	2020	9	1192663
9	2020	10	1892015
10	2020	11	4475991
11	2020	12	6323266
12	2021	1	6112572
13	2021	2	2374243
14	2021	3	405300



8.2 Análisis Geográfico: Impacto por Estado

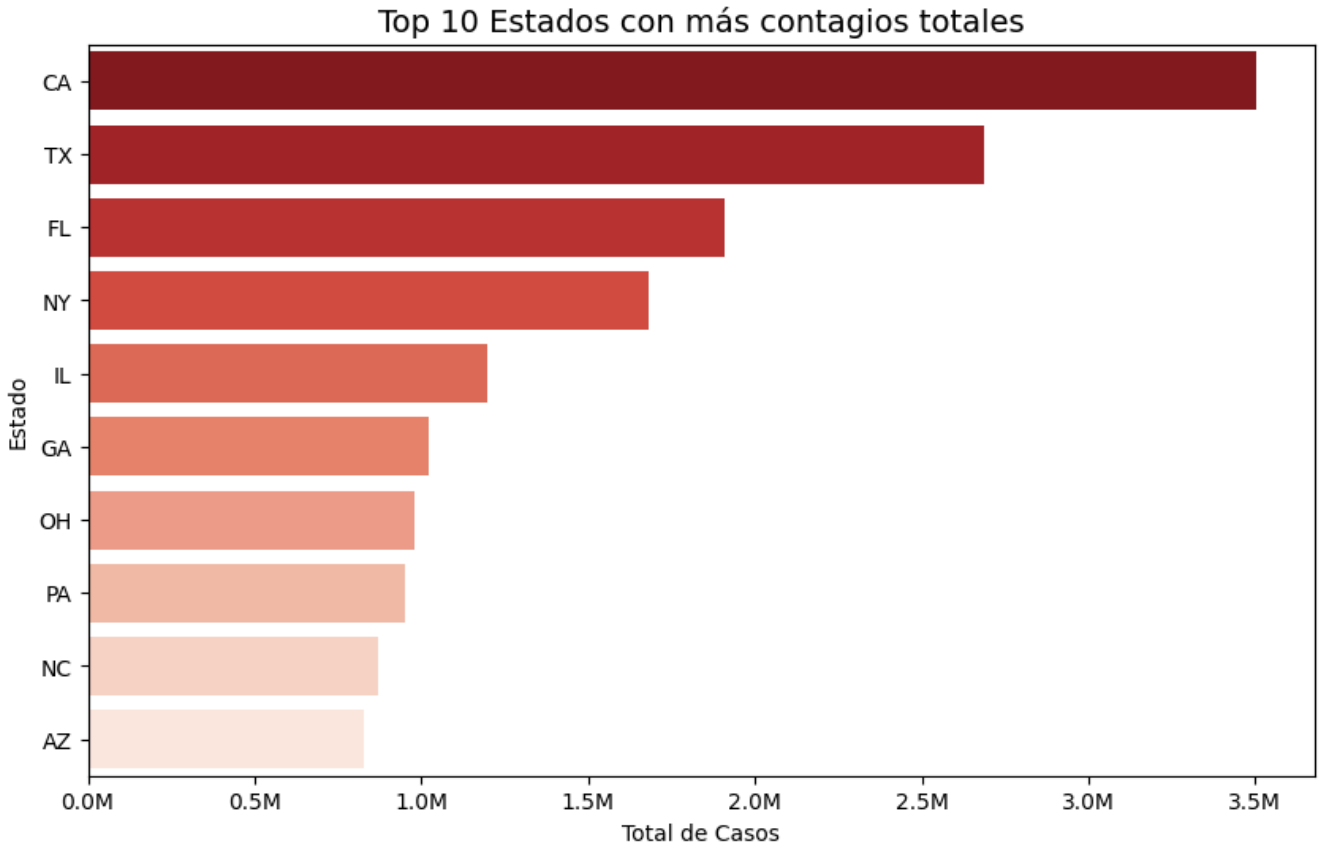
Estados Unidos es un país enorme y la pandemia no afectó a todos los estados por igual. Para visualizar la distribución geográfica del virus, utilizaremos un mapa de coropletas (mapa de calor geográfico).

Analizaremos el total de casos acumulados por estado para identificar las "zonas calientes" del país.

	state	positiveIncrease
0	AK	56886
1	AL	499819
2	AR	324818
3	AS	0
4	AZ	826452
5	CA	3501341
6	CO	436600
7	CT	285330
8	DC	41419
9	DE	88354

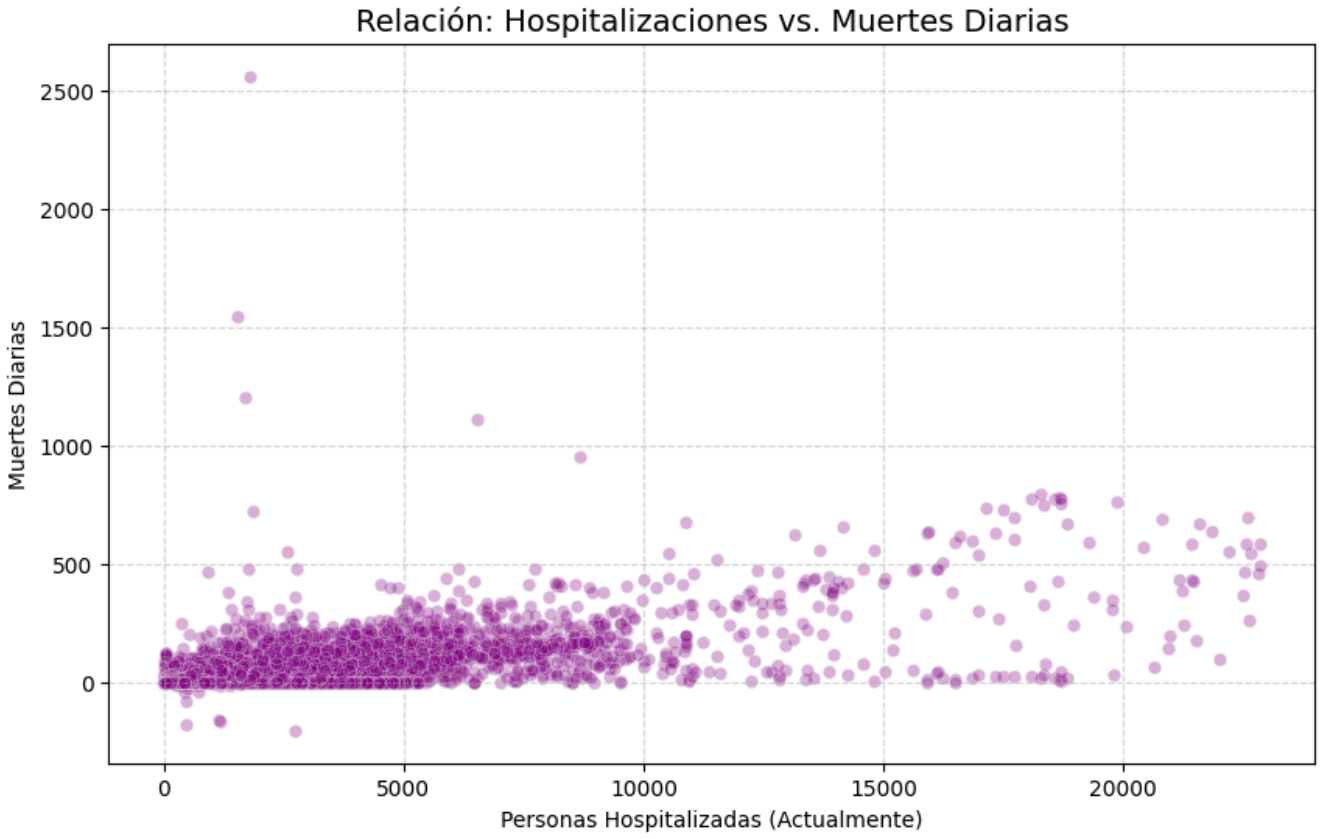
8.3 Top 10 Estados con mayor incidencia

Complementando el mapa, generamos un ranking de los 10 estados con mayor número absoluto de contagios. Esto nos ayuda a identificar dónde se concentró la mayor carga sanitaria.



8.4 Relación entre Hospitalizaciones y Fallecimientos

¿Una mayor saturación hospitalaria implicó siempre una mayor mortalidad? Mediante un gráfico de dispersión (Scatter Plot), analizamos la correlación entre las personas hospitalizadas actualmente y el incremento diario de muertes.



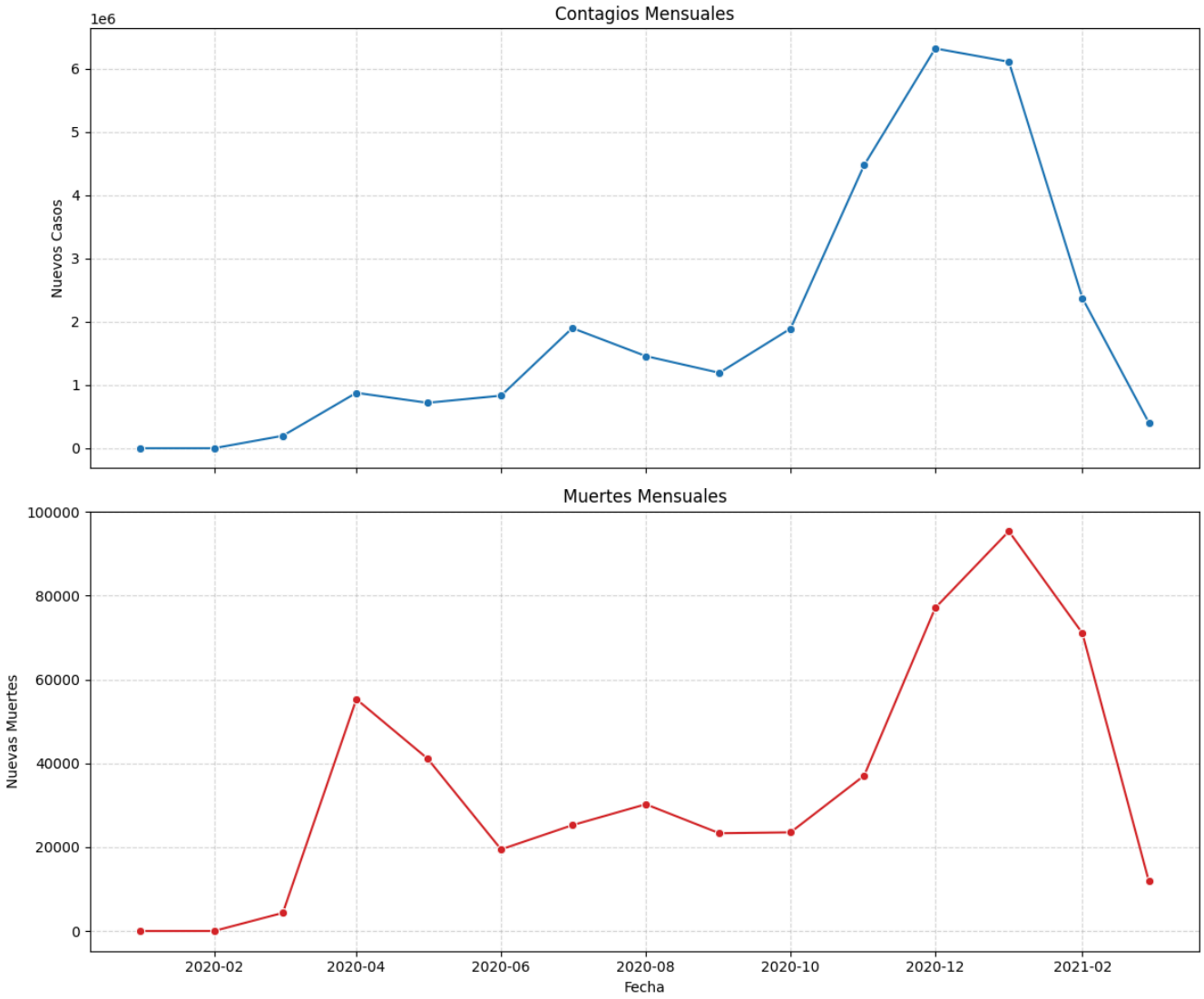
8.4 Análisis de Mortalidad: Contagios vs. Fallecimientos

Analizar solo los contagios nos cuenta una parte incompleta de la historia. Para comprender la severidad real de la pandemia en sus distintas fases, es necesario comparar la evolución de los nuevos casos (`positiveIncrease`) con la de los nuevos fallecimientos (`deathIncrease`).

En esta sección, generamos una visualización comparativa utilizando **subplots** (dos gráficos alineados verticalmente). Esto nos permite:

1. Observar si los picos de contagios se traducen inmediatamente en picos de mortalidad.
2. Identificar el **desfase temporal (lag)**: habitualmente, el aumento de muertes ocurre semanas después del aumento de casos.
3. Comparar tendencias: ¿Hubo momentos con pocos casos pero alta mortalidad? (Común al inicio de la pandemia).

	year	month	deathIncrease	periodo
0	2020	1	0	2020-01-01
1	2020	2	5	2020-02-01
2	2020	3	4326	2020-03-01
3	2020	4	55315	2020-04-01
4	2020	5	41137	2020-05-01
5	2020	6	19475	2020-06-01
6	2020	7	25249	2020-07-01
7	2020	8	30244	2020-08-01
8	2020	9	23329	2020-09-01
9	2020	10	23545	2020-10-01


Contagios Mensuales
1e6
Nuevos Casos
0
1
2
3
4
5
6
Muertes Mensuales
Nuevas Muertes
0
20000
40000
60000
80000
100000
2020-02
2020-04
2020-06
2020-08
2020-10
2020-12
2021-02
Fecha