Análisis Exploratorio de Datos 2023

Unidad 3 - Medidas de tendencia central

3A. Exploración de datos con Matplotlib

Nombre: David Aaron Ramirez Olmeda
Programa: Maestría en Ciencia de Datos e Información

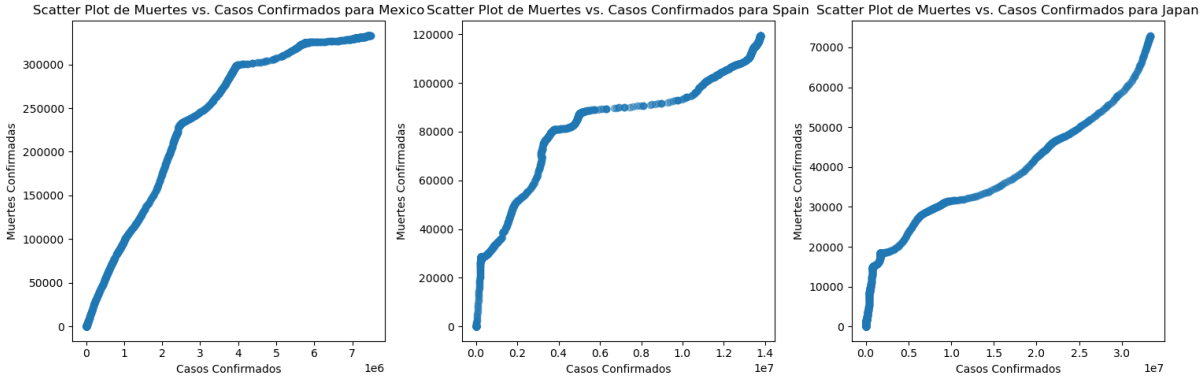
Introducción:

En este análisis de datos, hemos explorado la evolución de la pandemia de COVID-19 en tres países seleccionados: México, España y Japón. Utilizando conjuntos de datos que contienen información sobre casos confirmados y muertes confirmadas, realizamos una serie de pasos para comprender mejor el impacto de la pandemia en cada uno de estos países. Esto incluyó la exploración de estadísticas descriptivas, identificación de outliers, ajuste de distribuciones y la creación de visualizaciones relevantes.

In [1]:
```
import pandas as pd
import numpy as np
from scipy import stats

# Especifica las rutas
ruta_confirmados = '/Users/aaron/Documentos/MCDI/Semestre 2/Análisis Expl
ruta_muertes = '/Users/aaron/Documentos/MCDI/Semestre 2/Análisis Explorat
ruta_recuperados = '/Users/aaron/Documentos/MCDI/Semestre 2/Análisis Expl

# Carga los datos
df_confirmados = pd.read_csv(ruta_confirmados)
df_muertes = pd.read_csv(ruta_muertes)
```

In [2]:
```
paises_interes = ['Mexico', 'Spain', 'Japan']

# Filtrar los datos para los países de interés
df_paises = df_confirmados[df_confirmados['Country/Region'].isin(paises_i
df_muertes_paises = df_muertes[df_muertes['Country/Region'].isin(paises_i

# Transponer los datos para tener fechas como filas
df_paises_transpuesto = df_paises.melt(id_vars=['Province/State', 'Countr
                                       var_name='Fecha',
                                       value_name='CasosConfirmados')

df_muertes_transpuesto = df_muertes_paises.melt(id_vars=['Province/State'
                                                var_name='Fecha',
                                                value_name='MuertesConfirm

# Combinar los DataFrames de casos confirmados y muertes confirmadas
df_combined = df_paises_transpuesto.merge(df_muertes_transpuesto, on=['Pr

# Renombrar las columnas para mayor claridad
df_combined.rename(columns={'Country/Region': 'Country'}, inplace=True)
```

In [3]:
```
import matplotlib.pyplot as plt

# Crear subplots para cada país
fig, axes = plt.subplots(nrows=1, ncols=len(paises_interes), figsize=(15,

# Configurar los títulos de los subplots
for i, pais in enumerate(paises_interes):
    df_pais = df_combined[df_combined['Country'] == pais]
    axes[i].set_title(f'Scatter Plot de Muertes vs. Casos Confirmados par
    axes[i].scatter(df_pais['CasosConfirmados'], df_pais['MuertesConfirma
    axes[i].set_xlabel('Casos Confirmados')
    axes[i].set_ylabel('Muertes Confirmadas')

plt.tight_layout()
plt.show()
```



Se muestra la relación entre las variables de muertes y casos confirmados. Visualmente no se observa claramente si existe algún patrón o tendencia en los datos que sugiera una posible distribución subyacente, por lo que es mejor proceder con alguna prueba estadística o similar para ajustar los daatos a una distribución. Por ahora pasaremos a calcular datos que nos

In [10]:
```
# Calcular y mostrar estadísticas descriptivas para cada país
for pais in paises_interes:
    df_pais = df_paises_transpuesto[df_paises_transpuesto['Country/Region

    media = df_pais['CasosConfirmados'].mean()
    mediana = df_pais['CasosConfirmados'].median()
    moda = stats.mode(df_pais['CasosConfirmados'], axis=None)[0][0]
    varianza = df_pais['CasosConfirmados'].var()
    asimetria = df_pais['CasosConfirmados'].skew()
    curtosis = df_pais['CasosConfirmados'].kurtosis()

    print(f'Estadísticas descriptivas para {pais}:')
    print(f'Media: {media}')
    print(f'Mediana: {mediana}')
    print(f'Moda: {moda}')
    print(f'Varianza: {varianza}')
    print(f'Asimetría: {asimetria}')
    print(f'Curtosis: {curtosis}')
    print('\n')
```

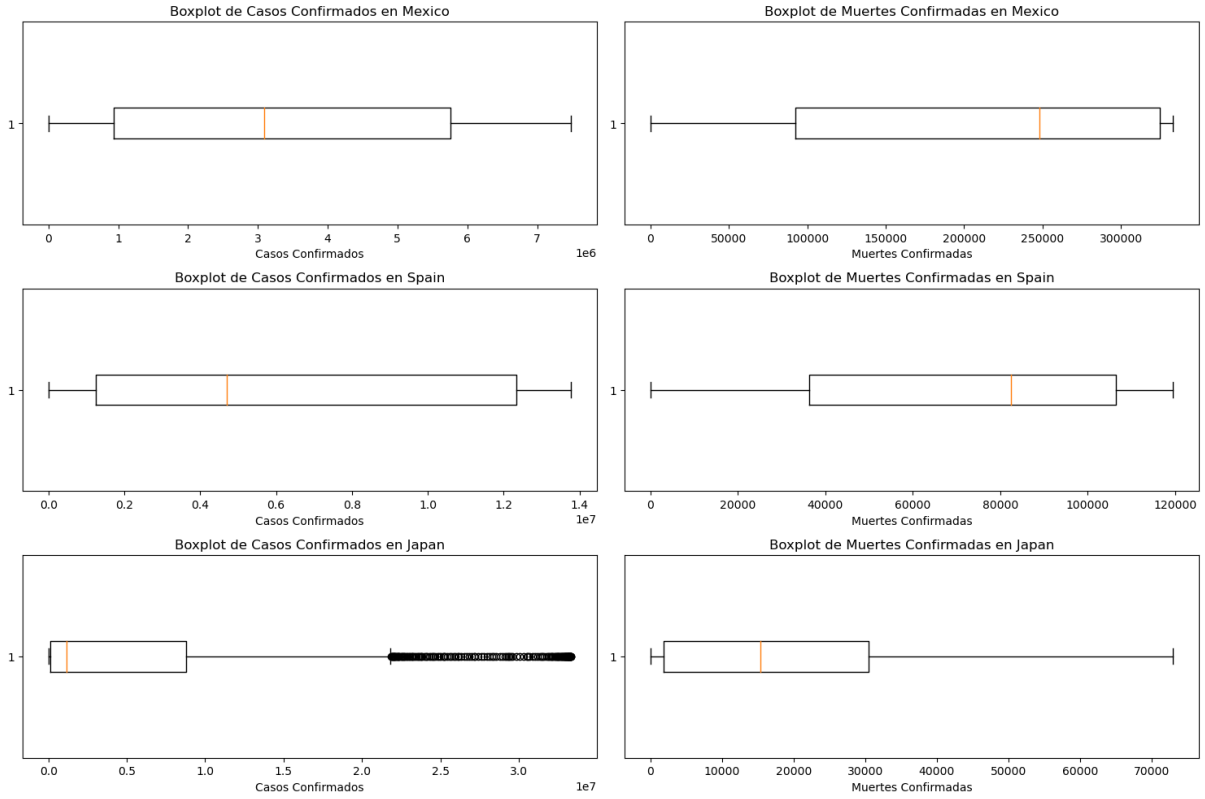
```
Estadísticas descriptivas para Mexico:
Media: 3450663.179352581
Mediana: 3091971.0
Moda: 0
Varianza: 6710816427353.128
Asimetría: 0.17023767702812947
Curtosis: -1.412098491522878


Estadísticas descriptivas para Spain:
Media: 6186592.4304461945
Mediana: 4693540.0
Moda: 2
Varianza: 27407292972312.88
Asimetría: 0.33680155703882075
Curtosis: -1.533452331934623


Estadísticas descriptivas para Japan:
Media: 6420132.660542432
Mediana: 1149874.0
Moda: 2
Varianza: 96314692731574.27
Asimetría: 1.5553087822817666
Curtosis: 1.0654874011291984
```

Estadísticas Descriptivas:

- **Media**: La media es aproximadamente 3,450,663, lo que significa que, en promedio, hubo alrededor de 3,450,663 casos confirmados en México durante el período de tiempo analizado.

- **Mediana**: La mediana es el valor medio cuando los datos se ordenan de menor a mayor, para México es 3,091,971, lo que indica que la mitad de los días tuvieron menos de esos casos confirmados y la otra mitad tuvo más.
- **Moda**: La moda es el valor que aparece con mayor frecuencia en los datos. En México, la moda es 0, lo que sugiere que en muchos días no se reportaron casos confirmados.
- **Varianza**: La varianza mide la dispersión de los datos. Una varianza alta indica que los valores están dispersos, mientras que una baja indica que están agrupados cerca de la media. En México, la varianza es alta, lo que significa que los datos de casos confirmados tienden a estar dispersos.
- **Asimetría**: La asimetría indica la simetría de la distribución de datos. Un valor positivo de asimetría (0.17 en este caso) sugiere que la distribución tiene una cola larga hacia la derecha, lo que significa que hay algunos días con un alto número de casos confirmados que afectan la asimetría.
- **Curtosis**: La curtosis mide la forma de la distribución. Un valor negativo de curtosis (-1.41 en este caso) indica una distribución achatada y con colas ligeramente más delgadas de lo normal.

In [11]:
```
import matplotlib.pyplot as plt

# Crear subplots para cada país
fig, axes = plt.subplots(nrows=len(paises_interes), ncols=2, figsize=(15,

# Configurar los títulos de los subplots
for i, pais in enumerate(paises_interes):
    df_pais = df_combined[df_combined['Country'] == pais]
    axes[i, 0].set_title(f'Boxplot de Casos Confirmados en {pais}')
    axes[i, 0].boxplot(df_pais['CasosConfirmados'], vert=False)
    axes[i, 0].set_xlabel('Casos Confirmados')

    axes[i, 1].set_title(f'Boxplot de Muertes Confirmadas en {pais}')
    axes[i, 1].boxplot(df_pais['MuertesConfirmadas'], vert=False)
    axes[i, 1].set_xlabel('Muertes Confirmadas')

plt.tight_layout()
plt.show()
```

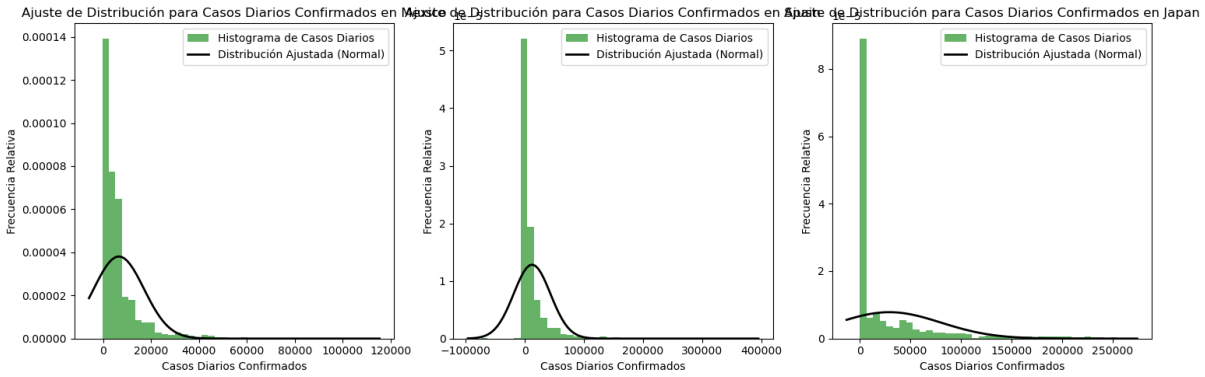


Boxplots para ambas variables, muertes y casos confirmados, para cada país. Visualmente se observan outliers o puntos que se encuentran significativamente a la derecha en el boxplot de casos confirmados para Japón. Un enfoque común para identificar outliers es utilizar el rango intercuartil (IQR)

In [12]:
```
df_japon = df_combined[df_combined['Country'] == 'Japan']

# Extrae la columna de fechas y casos confirmados para Japón
fechas_japon = df_japon['Fecha']
casos_japon = df_japon['CasosConfirmados']

# Calcula el primer cuartil (Q1)
Q1 = np.percentile(casos_japon, 25)

# Calcula el tercer cuartil (Q3)
Q3 = np.percentile(casos_japon, 75)

# Calcula el rango intercuartil (IQR)
IQR = Q3 - Q1

# Calcula el umbral superior para outliers
umbral_superior = Q3 + (1.5 * IQR)

# Identifica los outliers
outliers = df_japon[df_japon['CasosConfirmados'] > umbral_superior]

# Imprime la cuenta de todos los outliers
print(f"Cantidad de outliers para casos confirmados en Japón: {len(outlie

# Crea un DataFrame de outliers con fechas y valores
df_outliers = pd.DataFrame({
    'Fecha': outliers['Fecha'],
    'CasosConfirmados': outliers['CasosConfirmados']
})

# Imprime el DataFrame de outliers
print("\nDataFrame de outliers:")
print(df_outliers)
```

```
Cantidad de outliers para casos confirmados en Japón: 143

DataFrame de outliers:
         Fecha  CasosConfirmados
3000  10/18/22          21843970
3003  10/19/22          21887525
3006  10/20/22          21923635
3009  10/21/22          21955228
3012  10/22/22          21989401
...        ...               ...
3414    3/5/23          33282370
3417    3/6/23          33286633
3420    3/7/23          33298799
3423    3/8/23          33310604
3426    3/9/23          33320438

[143 rows x 2 columns]
```

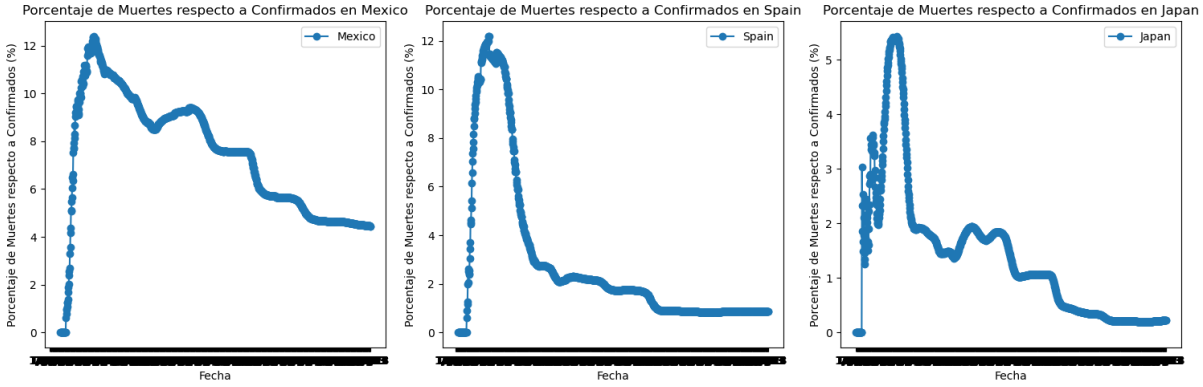
In [14]:
```
import matplotlib.pyplot as plt
from scipy.stats import norm

# Crear una nueva columna para casos diarios confirmados
df_combined['CasosDiariosConfirmados'] = df_combined.groupby('Country')['

# Configurar subplots
fig, axes = plt.subplots(nrows=1, ncols=len(paises_interes), figsize=(15,

# Iterar a través de los países y graficar las distribuciones ajustadas
for i, pais in enumerate(paises_interes):
    casos_diarios_pais = df_combined[df_combined['Country'] == pais]['Cas
    mu, std = norm.fit(casos_diarios_pais)

    # Histograma de casos diarios confirmados
    axes[i].hist(casos_diarios_pais, bins=40, density=True, alpha=0.6, co

    # Distribución ajustada (normal)
    xmin, xmax = axes[i].get_xlim()
    x = np.linspace(xmin, xmax, 100)
    p = norm.pdf(x, mu, std)
    axes[i].plot(x, p, 'k', linewidth=2, label='Distribución Ajustada (No

    axes[i].set_xlabel('Casos Diarios Confirmados')
    axes[i].set_ylabel('Frecuencia Relativa')
    axes[i].set_title(f'Ajuste de Distribución para Casos Diarios Confirm
    axes[i].legend()

plt.tight_layout()
plt.show()
```



Como se explicó en un principio, podríamos hacer alguna prueba estadística para determinar que distribución se puede ajustar más a nuestros datos. Ajustar nuestras visualizaciones (bins), reducir el periodo de tiempo a analizar, probar otros métodos, etc. pueden ser otros métodos eficientes para llegar a resultados aún más claros, para fines ilustrativos, me parece que esto es un buen primer acercamiento.

Ahora, para una comparación del impacto de la pandemia en los países seleccionados podríamos, por ejemplo, comparar el porcentaje de muertes respecto a las personas confirmadas. Esto nos puedo dar un entendimiento de, por ejemplo, el sistema de salud del país en general.

In [15]:
```
import matplotlib.pyplot as plt

# Filtrar los datos para los países de interés
df_paises = df_combined[df_combined['Country'].isin(paises_interes)]

# Calcular el porcentaje de muertes respecto a casos confirmados
df_paises['PorcentajeMuertes'] = (df_paises['MuertesConfirmadas'] / df_pa

# Configurar subplots
fig, axes = plt.subplots(nrows=1, ncols=len(paises_interes), figsize=(15,

# Iterar a través de los países y crear gráficas de porcentaje de muertes
for i, pais in enumerate(paises_interes):
    df_pais = df_paises[df_paises['Country'] == pais]

    axes[i].plot(df_pais['Fecha'], df_pais['PorcentajeMuertes'], label=f'

    axes[i].set_xlabel('Fecha')
    axes[i].set_ylabel('Porcentaje de Muertes respecto a Confirmados (%)'
    axes[i].set_title(f'Porcentaje de Muertes respecto a Confirmados en {
    axes[i].legend()

plt.tight_layout()
plt.show()
```



Y finalmente es fácil darse cuenta como algunos países trataban de manera más eficiente a pacientes confirmados que otros.

Conclusiones:

- A lo largo de nuestro análisis, hemos observado diferencias significativas en la evolución de la pandemia en México, España y Japón. Cada país ha experimentado variaciones en la magnitud y velocidad de propagación de la enfermedad, así como en la letalidad relativa de la misma.
- Las estadísticas descriptivas nos han proporcionado una comprensión cuantitativa de la dispersión de casos confirmados en cada país. Observamos diferencias en la media, mediana, varianza y asimetría de los casos confirmados, lo que sugiere variaciones en la propagación y gestión de la pandemia.
- La identificación de outliers nos permitió destacar valores atípicos en los datos, lo que puede indicar eventos especiales.

- Hemos explorado la posibilidad de ajustar las distribuciones de casos diarios confirmados a una distribución normal, lo que podría proporcionar información sobre la dinámica de la enfermedad en cada país. Sin embargo, los resultados sugieren que en algunos casos, como Japón, las distribuciones no se ajustan bien a una distribución normal para el periodo de tiempo comprendido por los datos.
- Finalmente, visualizaciones muestran el porcentaje de muertes respecto a los casos confirmados a lo largo del tiempo para cada país. Estas visualizaciones resaltan las diferencias en la letalidad.

Referencias:

https://github.com/CSSEGISandData/COVID-19/tree/master