Proyecto de Ciencia de Datos

COVID19_Analytics

Elaborado por: Jhonatan Rodrigues Custodio – Data Science Junior

correo: jhonatan9494c@gmail.com

1.-Problema De Negocio

Una entidad gubernamental responsable de la gestión de la salud en un país (PERÚ) enfrenta el desafío de comprender y analizar la propagación del COVID-19 para tomar decisiones informadas y eficaces en la gestión de la pandemia. Como científico de datos, tu tarea es analizar los datos relacionados con el COVID-19 y presentar insights a través de visualizaciones que respondan a las siguientes preguntas clave:

2.-Preguntas:

- ¿Cómo ha evolucionado el Covid-19 en el país en comparación con el impacto observado a nivel global?
- ¿Cuál ha sido la evolución de los nuevos casos diarios reportados de Covid-19 en el país a lo largo del tiempo?
- ¿Cuál es la evolución del índice de letalidad del Covid-19 en el país, comparado con los países con los índices históricos más elevados?
- Desde una perspectiva demográfica, ¿cuáles son las características que tienen un mayor impacto en el índice de letalidad de un país?
- ¿Cómo ha evolucionado el número de muertes por COVID-19 a lo largo del tiempo?

3.- Obtención, Tratamiento y Análisis Exploratorio (EDA)

2.1.- Cargando Base de Datos

Obtenemos dos dataset una para el covid y la otra para la población

```
df_covid = pd.read_csv(StringIO(requests.get("https://covid19.who.int/WHO-COVID-19-global-data.csv").text))
df_population = pd.read_excel('https://raw.githubusercontent.com/ElProfeAlejo/Bootcamp_Databases/main/WPP2022_GEN_F01_DEMOGRAPHIC_INDICATORS_COMPACT_REV1.xlsx',
                              sheet_name=0, skiprows=16)
df_population = df_population[df_population['Year'] == 2019]
```

2.2.- EDA Base Covid World Health Organization

```
df_covid.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50880 entries, 0 to 50879
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Date_reported      50880 non-null  object
 1   Country_code       49396 non-null  object
 2   Country            49608 non-null  object
 3   WHO_region         45792 non-null  object
 4   New_cases          36622 non-null  float64
 5   Cumulative_cases   50880 non-null  int64
 6   New_deaths         24216 non-null  float64
 7   Cumulative_deaths  50880 non-null  int64
dtypes: float64(2), int64(2), object(4)
memory usage: 3.1+ MB
```

En el paso de Análisis Exploratorio de Datos (EDA) para el conjunto de datos df_covid, se realizaron las siguientes modificaciones y tratamientos:

- Eliminación de registros con valores nulos en la columna 'Country_code'.

```
df_covid_limpio = df_covid.dropna(subset=['Country_code'])
```

- Selección y retención de las columnas específicas: ['Date_reported', 'Country_code', 'Country', 'New_cases', 'Cumulative_cases', 'New_deaths', 'Cumulative_deaths'].

```
df_covid_limpio = df_covid_limpio[['Date_reported','Country_code','Country','New_cases',
                                   'Cumulative_cases','New_deaths','Cumulative_deaths']]
```

- Sustitución de valores nulos por cero en todo el dataframe.

```
#sustituir valores nulos por 0
df_covid_limpio=df_covid_limpio.fillna(0)
```

- Conversión del formato de la columna 'Date_reported' a datetime.

```
df_covid_limpio['Date_reported']=pd.to_datetime(df_covid_limpio['Date_reported'], format='%Y-%m-%d')
```

- Conversión del formato de las columnas ['New_cases', 'New_deaths'] a int64.

```
df_covid_limpio['New_cases'] = pd.to_numeric(df_covid_limpio['New_cases'], errors='coerce').astype(int)
df_covid_limpio['New_deaths'] = pd.to_numeric(df_covid_limpio['New_deaths'], errors='coerce').astype(int)
```

- Creación de una nueva columna 'lethality_rate' utilizando la fórmula: 'Cumulative_deaths'/'Cumulative_cases'*100.

```
df_covid_limpio['lethality_rate'] = (df_covid_limpio['Cumulative_deaths'] / df_covid_limpio['Cumulative_cases']) * 100
```

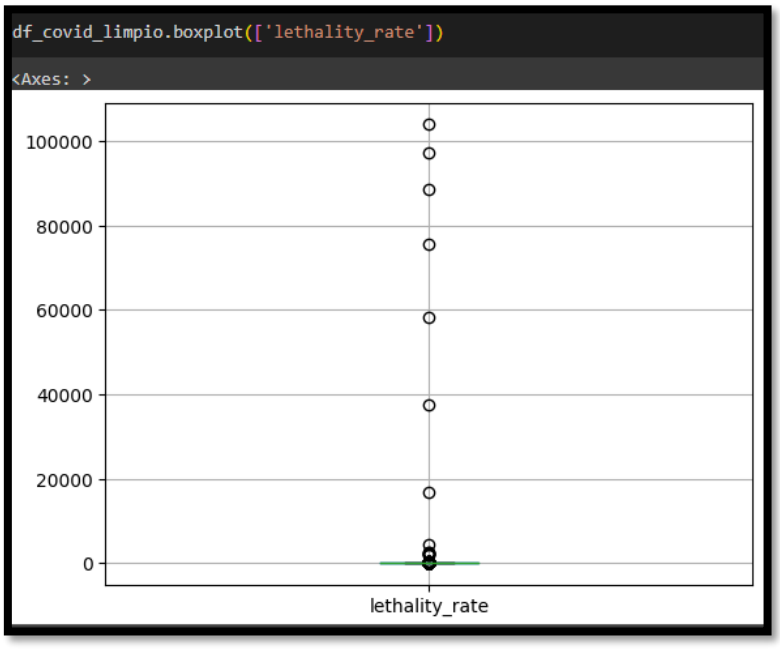
- Filtrado de registros, manteniendo solo aquellos cuyo valor en la columna 'lethality_rate' se encuentra entre los cuantiles 0 a 0.99 para eliminar outliers.

```
df_covid_limpio.describe()
```

VALORES INF

	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths	lethality_rate
count	4.939600e+04	4.939600e+04	49396.000000	4.939600e+04	4.530700e+04
mean	1.567324e+04	1.643477e+06	142.040530	1.928293e+04	inf
std	2.408294e+05	7.274286e+06	875.025827	7.909933e+04	NaN
min	-6.507900e+04	0.000000e+00	-3432.000000	0.000000e+00	0.000000e+00
25%	0.000000e+00	3.846750e+03	0.000000	2.900000e+01	5.128205e-01
50%	1.020000e+02	4.283250e+04	0.000000	4.980000e+02	1.181060e+00
75%	1.999250e+03	4.998840e+05	20.000000	6.630250e+03	2.160936e+00
max	4.047548e+07	1.034368e+08	47687.000000	1.165780e+06	inf

```
df_covid_limpio.boxplot(['lethality_rate'])
```



```
q_low = df_covid_limpio['lethality_rate'].quantile(0)
q_high = df_covid_limpio['lethality_rate'].quantile(0.99)
df_covid_limpio = df_covid_limpio[(df_covid_limpio['lethality_rate'] >= q_low) & (df_covid_limpio['lethality_rate'] <= q_high)]
```

- Reinicio del índice del dataframe final df_covid_limpio.

```
#Finalmente reiniciaremos el índice del dataframe final df_covid_limpio
df_covid_limpio.reset_index(drop=True, inplace=True)
```

```
df_covid_limpio.describe()
```

Ya no presentan valores inf

	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths	lethality_rate
count	4.485300e+04	4.485300e+04	44853.000000	4.485300e+04	44853.000000
mean	1.722128e+04	1.809406e+06	151.147125	2.114609e+04	1.595293
std	2.526782e+05	7.614161e+06	905.888414	8.276044e+04	1.666892
min	-6.507900e+04	1.000000e+00	-3432.000000	0.000000e+00	0.000000
25%	5.000000e+00	8.173000e+03	0.000000	7.300000e+01	0.504360
50%	1.720000e+02	6.399300e+04	1.000000	8.030000e+02	1.164988
75%	2.624000e+03	6.263340e+05	26.000000	7.975000e+03	2.126546
max	4.047548e+07	1.034368e+08	47687.000000	1.165780e+06	12.776413

2.3.- EDA Base Population United Nations

En el marco del Análisis Exploratorio de Datos (EDA) para el conjunto de datos denominado df_population, se llevaron a cabo diversas transformaciones y tratamientos:

- ✓ Se procedió a seleccionar exclusivamente las columnas necesarias, conservando únicamente las siguientes: ['ISO2 Alpha-code','Total Population, as of 1 July (thousands)','Male Population, as of 1 July (thousands)','Female Population, as of 1 July (thousands)','Population Density, as of 1 July (persons per square km)','Life Expectancy at Birth, both sexes (years)'].

```
#Mantener en el dataframe sólo algunas columnas
df_population_limpio=df_population[['ISO2 Alpha-code','Total Population, as of 1 July (thousands)',
                                    'Male Population, as of 1 July (thousands)',
                                    'Female Population, as of 1 July (thousands)',
                                    'Population Density, as of 1 July (persons per square km)',
                                    'Life Expectancy at Birth, both sexes (years)']]
```

- ✓ Se realizaron ajustes en la nomenclatura de las columnas para simplificar su denominación, renombrándolas de la siguiente manera:
  - 'ISO2 Alpha-code': 'Country_code'
  - 'Total Population, as of 1 July (thousands)': 'Total_Population'
  - 'Male Population, as of 1 July (thousands)': 'Male_Population'
  - 'Female Population, as of 1 July (thousands)': 'Female_Population'
  - 'Population Density, as of 1 July (persons per square km)': 'Population_Density'
  - 'Life Expectancy at Birth, both sexes (years)': 'Life_Expectancy'

```
#Renombrar columnas
df_population_limpio.rename(columns={
    'ISO2 Alpha-code': 'Country_code',
    'Total Population, as of 1 July (thousands)': 'Total_Population',
    'Male Population, as of 1 July (thousands)': 'Male_Population',
    'Female Population, as of 1 July (thousands)': 'Female_Population',
    'Population Density, as of 1 July (persons per square km)': 'Population_Density',
    'Life Expectancy at Birth, both sexes (years)': 'Life_Expectancy'
}, inplace=True)
```

- ✓ Se procedió a la eliminación de registros que contenían valores nulos.

```
#Eliminar valores nulos
df_population_limpio.dropna(inplace=True)
```

- ✓ Se realizó la conversión de las columnas ['Total_Population','Male_Population','Female_Population','Population_Density', 'Life_Expectancy'] al formato float.

```
columnas=['Total_Population','Male_Population','Female_Population','Population_Density','Life_Expectancy']
df_population_limpio[columnas] = df_population_limpio[columnas].apply(pd.to_numeric, errors='coerce').astype(float)
```

- ✓ Verificamos si quitamos los outlier con respecto al impacto en general y por el país que analizamos que es Perú

```
#calcular el porcentaje de poblacion total que perderia si eliminaba esos registros outliers_df
percentage_lost = (len(outliers_df) / len(df_population_limpio)) * 100
print(f"Porcentaje de población total perdida al eliminar valores atípicos: {percentage_lost:.2f}%")
```

```
Porcentaje de población total perdida al eliminar valores atípicos: 21.28%
```

```
#calcular el porcentaje de poblacion total que perderia si eliminaba esos registros outliers_df en el Perú
percentage_lost = (len(outliers_df[outliers_df['Country_code']=='PE']) / len(df_population_limpio)) * 100
print(f"Porcentaje de población total perdida al eliminar valores atípicos en Perú: {percentage_lost:.2f}%")
```

```
Porcentaje de población total perdida al eliminar valores atípicos en Perú: 0.00%
```

Decisión: Decidimos no quitar los outliers

- ✓ Se multiplicaron por 1000 los valores actuales de las columnas 'Total_Population','Male_Population','Female_Population', y se sobreescribieron dichas columnas con los resultados obtenidos.

```
#Multiplicar por 1000 el valor actual de las siguientes columnas 'Total_Population','Male_Population','Female_Population' y sobreescribir
df_population_limpio['Total_Population'] = df_population_limpio['Total_Population'] * 1000
df_population_limpio['Male_Population'] = df_population_limpio['Male_Population'] * 1000
df_population_limpio['Female_Population'] = df_population_limpio['Female_Population'] * 1000
df_population_limpio.head()
```

- ✓ Se llevó a cabo nuevamente la eliminación de registros que contenían valores nulos.

```
#Eliminar valores nulos
df_population_limpio.dropna(inplace=True)
```

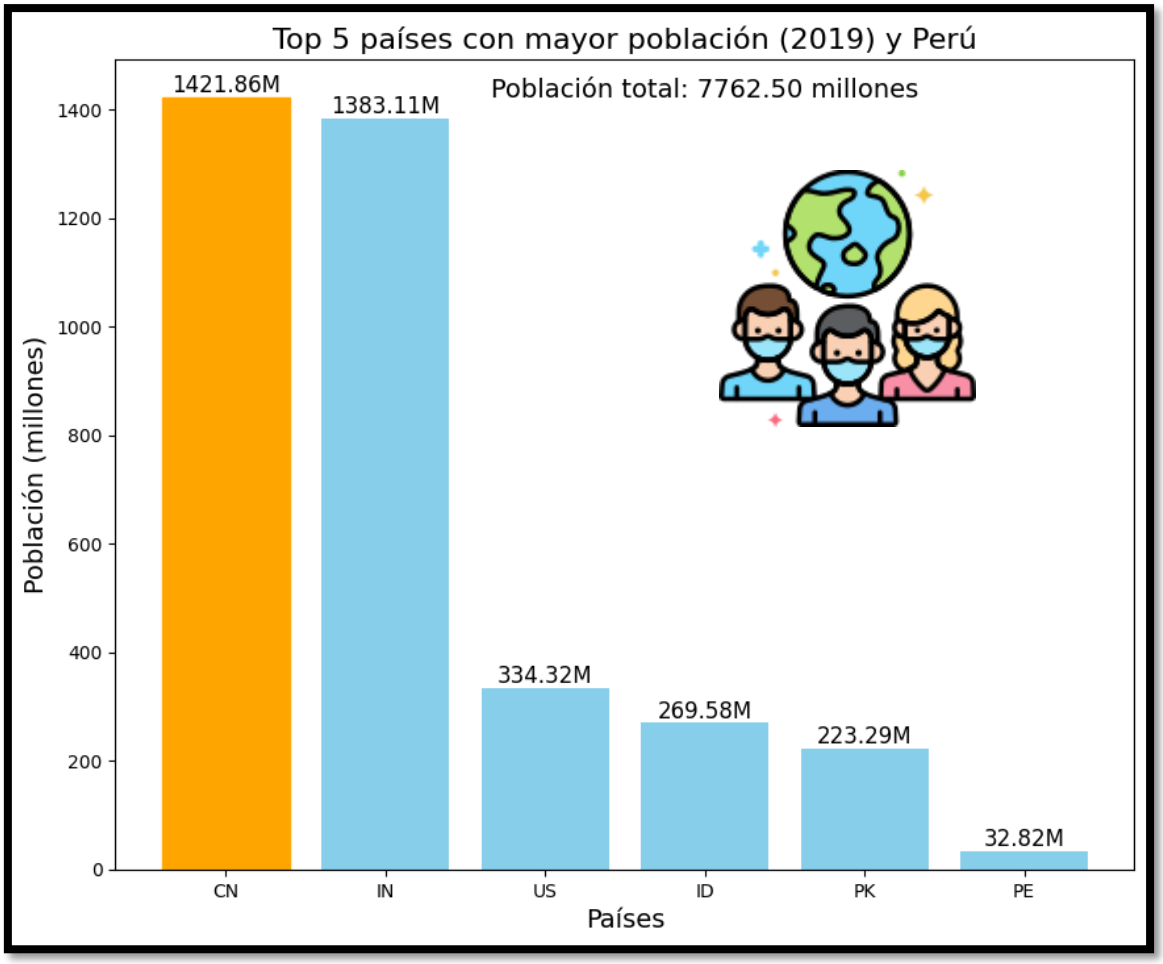
- ✓ Para finalizar, se reinició el índice del dataframe resultante, el cual fue denominado df_population_limpio.

```
#Reiniciar el indice
df_population_limpio.reset_index(drop=True, inplace=True)
```

```
df_population_limpio.describe()
```

	Total_Population	Male_Population	Female_Population	Population_Density	Life_Expectancy
count	2.350000e+02	2.350000e+02	2.350000e+02	235.000000	235.000000
mean	3.303193e+07	1.661798e+07	1.641395e+07	459.047102	73.489502
std	1.344877e+08	6.889130e+07	6.560933e+07	2222.151677	7.360874
min	1.752000e+03	8.790000e+02	8.740000e+02	0.136000	52.910000
25%	4.000210e+05	1.948405e+05	2.109970e+05	38.731000	68.524500
50%	5.453924e+06	2.738222e+06	2.767844e+06	95.237000	75.057000
75%	2.075953e+07	1.040942e+07	1.035011e+07	239.942000	79.100500
max	1.421864e+09	7.267819e+08	6.950821e+08	24855.034000	86.542000

2.4.- Top 5 Países con más población



Poblaciones:

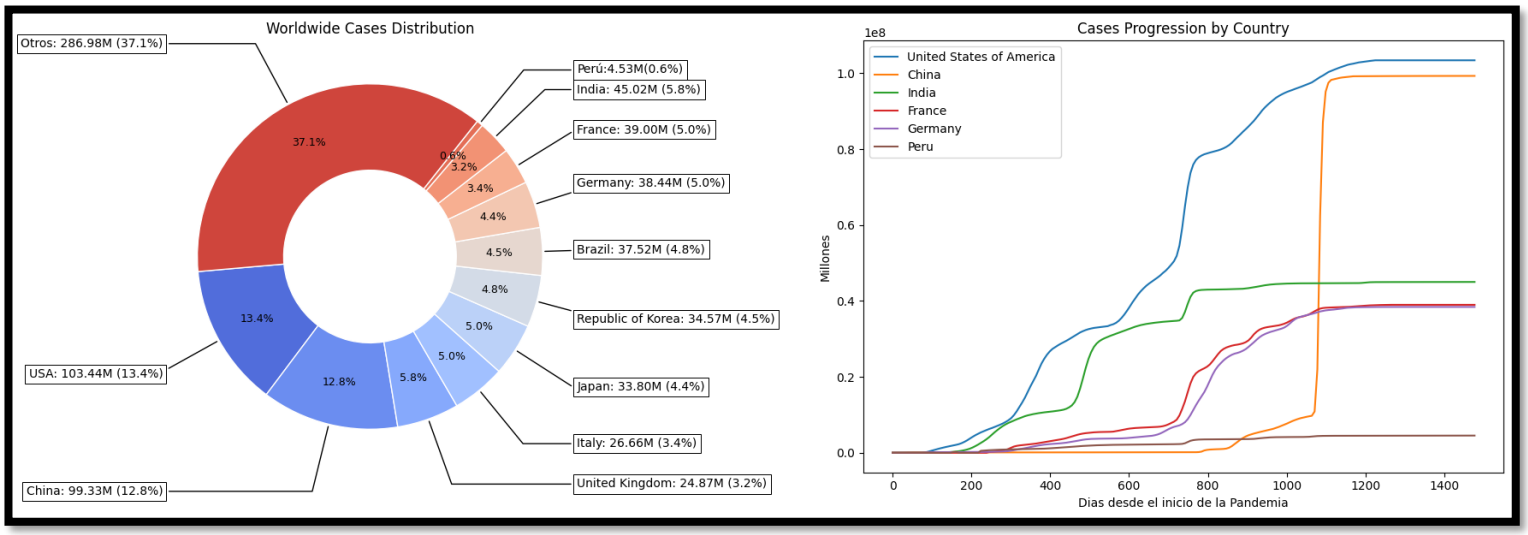
- ✓ China: 1421.86M
- ✓ India: 1383.11M
- ✓ Estados Unidos: 334.32M
- ✓ Indonesia: 269.58M
- ✓ Pakistán: 223.29M
- ✓ Perú: 32.82M

Población total: 7762.50 millones

Se puede observar que China e India tienen las poblaciones más grandes, seguidas por Estados Unidos, Indonesia y Pakistán. Perú tiene una población significativamente menor en comparación.

2.5.- Respondiendo los insight

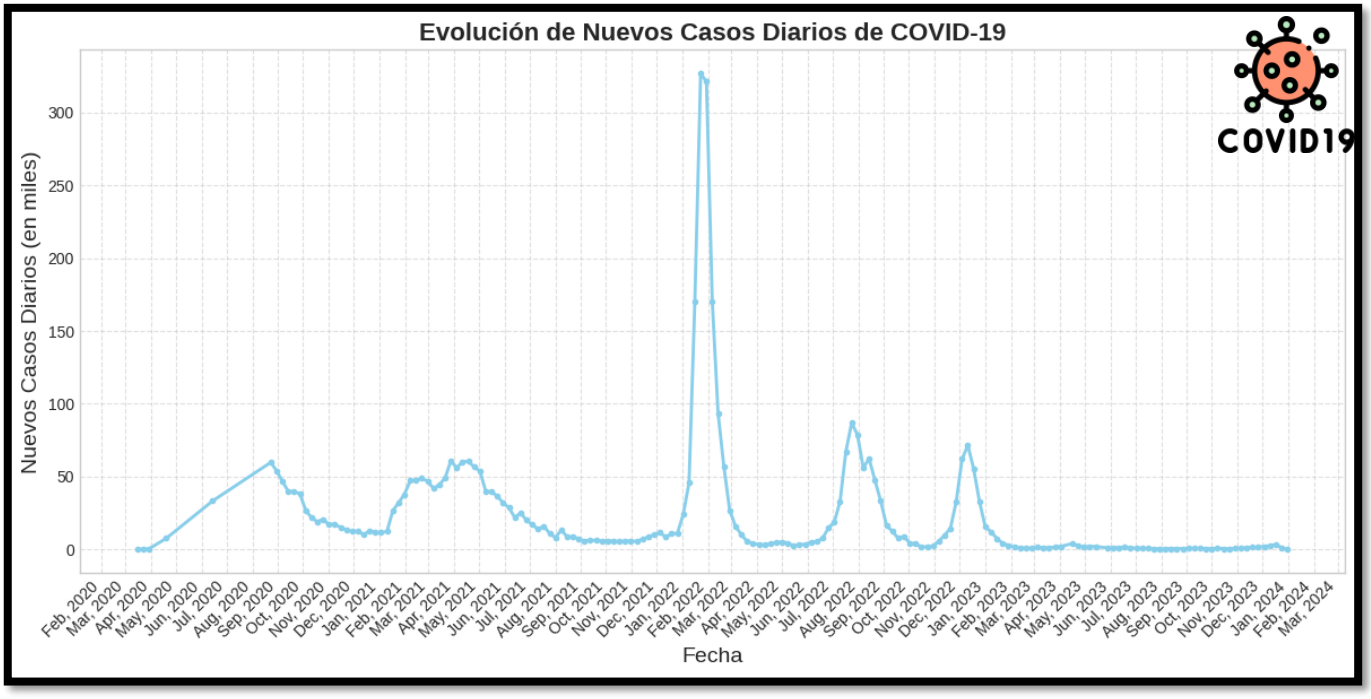
- ✓ **¿Cómo ha evolucionado el Covid-19 en el país en comparación con el impacto observado a nivel global?**



Distribución mundial de casos: El gráfico circular muestra que Perú tiene 4.53M de casos, lo que representa el 6% del total mundial.

Progresión de casos en el Perú: El gráfico lineal muestra la progresión de los casos a lo largo del tiempo en varios países, incluido Perú. La línea de Perú muestra una tendencia ascendente moderada, lo que indica un aumento constante en el número de casos.

- ✓ **¿Cuál ha sido la evolución de los nuevos casos diarios reportados de Covid-19 en el país a lo largo del tiempo?**



El gráfico muestra la evolución de los nuevos casos diarios de COVID-19 desde febrero de 2020 hasta febrero de 2024.

Análisis del gráfico: La línea azul Representa los nuevos casos diarios de COVID-19 en el Perú se observa un patrón oscilante con un pico muy pronunciado alrededor de enero del 2022, lo que indica un aumento significativo en los casos nuevos durante ese período hasta abril del 2022.

El aumento significativo de casos de COVID-19 en Perú entre enero y abril de 2022 podría deberse a varios factores.

Variantes del virus: Las nuevas variantes del virus, como la variante Ómicron, pueden ser más transmisibles y causar un aumento en el número de casos.

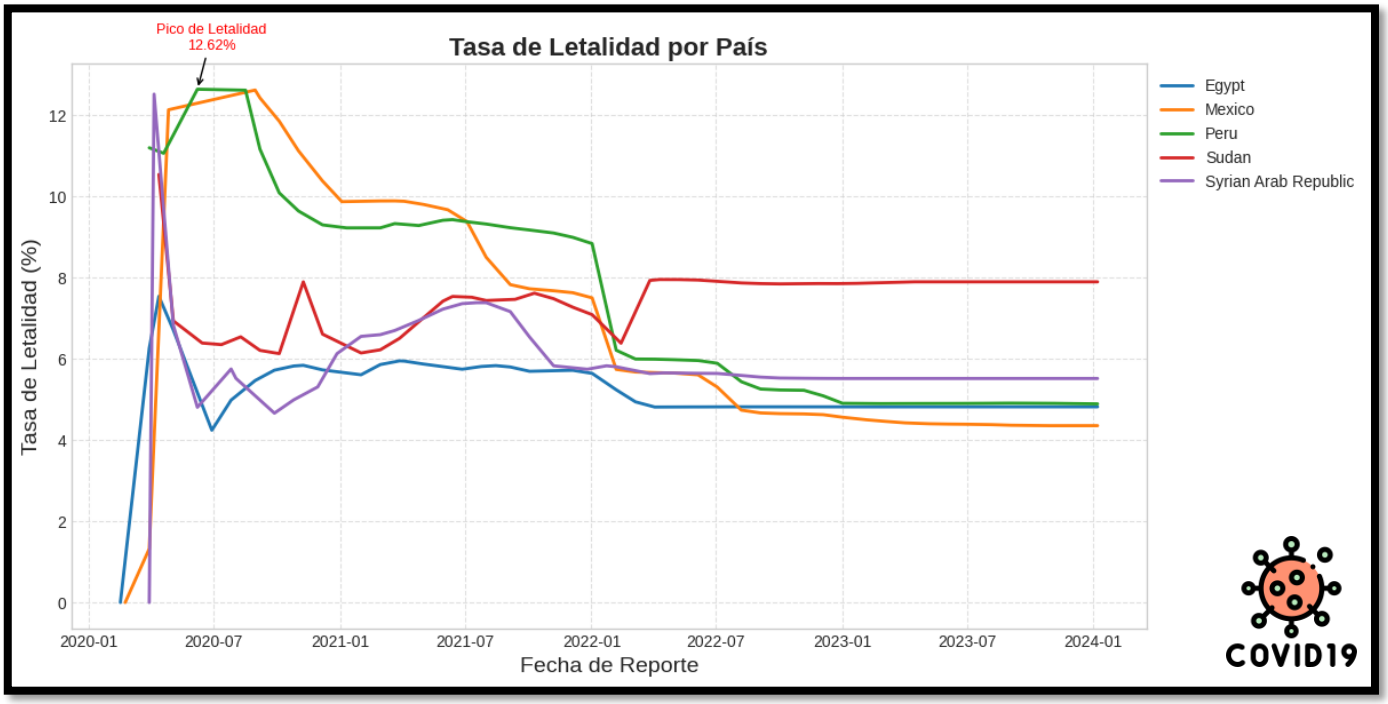
La variante Ómicron del COVID-19 fue confirmada en Perú el 19 de diciembre de 2021

Variante Ómicron llegó al Perú: 12 casos identificados, restricciones y adelanto de la dosis de refuerzo - Infobae

Esto produjo como consecuencia que se incluyera una vacuna de refuerzo es por eso que en abril comienza a bajar los casos de coronavirus

Factores socioeconómicos: Los factores sociales y económicos, como la economía informal, pueden dificultar la contención de la propagación del virus.

- ✓ **¿Cuál es la evolución del índice de letalidad del Covid-19 en el país, comparado con los países con los índices históricos más elevados?**



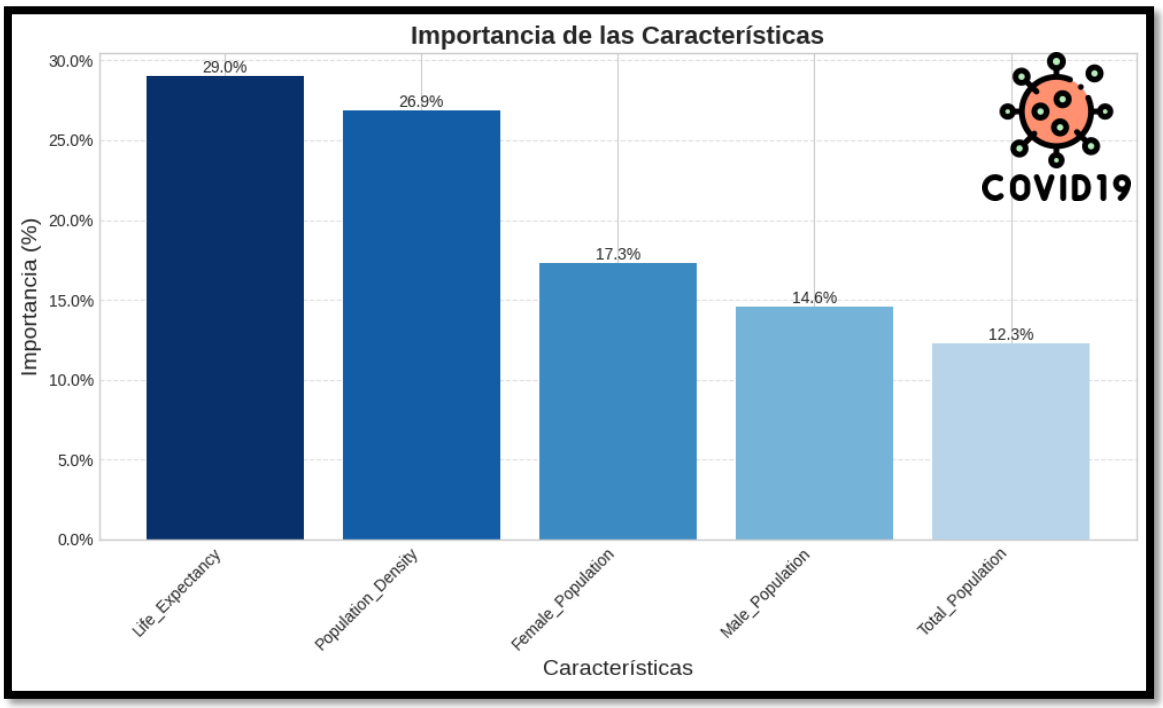
Análisis:

Pico de Letalidad: La tasa de letalidad en Perú, representada por la línea verde, alcanzó su pico alrededor de junio de 2020 con una tasa cercana al 12.62%.

Tendencia a la Baja: Desde el pico, la tasa de letalidad ha mostrado una tendencia general a la baja.

Estabilización: La tasa de letalidad se estabilizó alrededor del 6% hacia enero de 2022 y se mantuvo relativamente constante hasta enero de 2024.

- ✓ **Desde una perspectiva demográfica, ¿cuáles son las características que tienen un mayor impacto en el índice de letalidad de un país?**



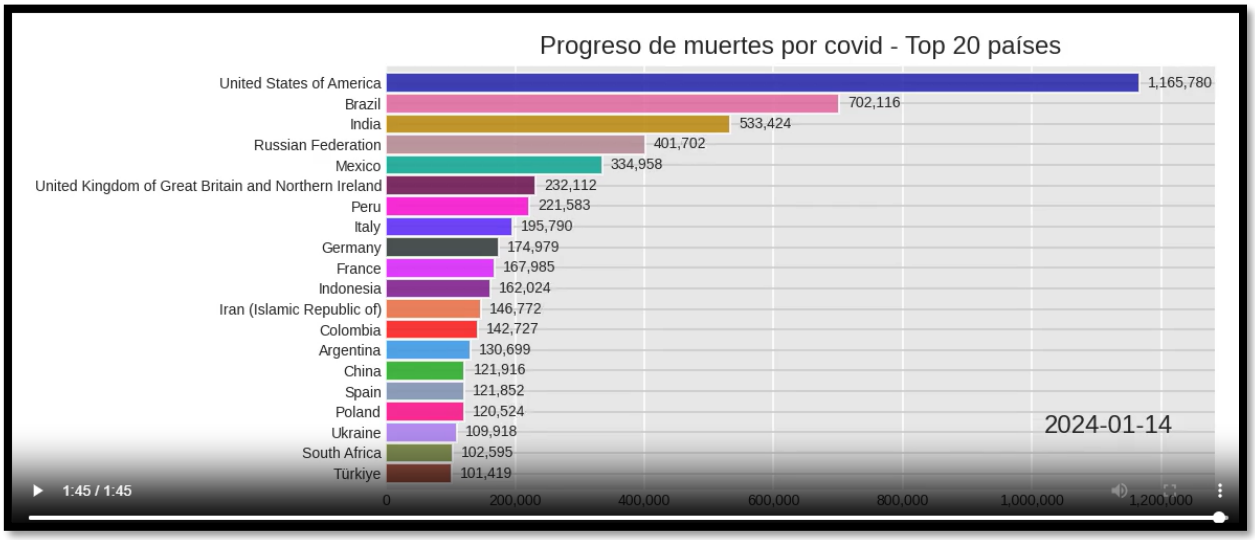
Análisis:

Expectativa de Vida: Esta característica tiene la mayor importancia con un 29.0%. Esto sugiere que la expectativa de vida es un factor muy relevante en el conjunto de datos analizado.

Densidad Poblacional: La densidad poblacional tiene una importancia del 26.9%, lo que indica que también es un factor significativo.

Esto sugiere que la expectativa de vida y la concentración de población son relevantes en la incidencia de la letalidad del Covid19.

✓ **¿Cómo ha evolucionado el número de muertes por COVID-19 a lo largo del tiempo?**



Análisis:

En los primeros días de la pandemia, China experimentó un alto número de muertes, pero a medida que avanzaba el tiempo, Estados Unidos superó a China en abril del 2020 y se convirtió en el país con la mayor cantidad de muertes por COVID-19, manteniendo esa posición hasta el final del periodo analizado. Estados Unidos registró el mayor número de fallecimientos a nivel mundial. En el caso de Perú, nuestro análisis revela que se ubicó en el séptimo lugar del mundo en términos de mortalidad, considerando todo el periodo hasta enero de 2024.