covid19_analysis

May 14, 2025

[8]:
```
import pandas as pd

df = pd.read_csv('owid-covid-data.csv')

print("Shape of dataset:", df.shape)
```

```
Shape of dataset: (350085, 67)
```

[9]:
```
print("Columns:")
print(df.columns.tolist())
```

```
Columns:
['iso_code', 'continent', 'location', 'date', 'total_cases', 'new_cases',
'new_cases_smoothed', 'total_deaths', 'new_deaths', 'new_deaths_smoothed',
'total_cases_per_million', 'new_cases_per_million',
'new_cases_smoothed_per_million', 'total_deaths_per_million',
'new_deaths_per_million', 'new_deaths_smoothed_per_million',
'reproduction_rate', 'icu_patients', 'icu_patients_per_million',
'hosp_patients', 'hosp_patients_per_million', 'weekly_icu_admissions',
'weekly_icu_admissions_per_million', 'weekly_hosp_admissions',
'weekly_hosp_admissions_per_million', 'total_tests', 'new_tests',
'total_tests_per_thousand', 'new_tests_per_thousand', 'new_tests_smoothed',
'new_tests_smoothed_per_thousand', 'positive_rate', 'tests_per_case',
'tests_units', 'total_vaccinations', 'people_vaccinated',
'people_fully_vaccinated', 'total_boosters', 'new_vaccinations',
'new_vaccinations_smoothed', 'total_vaccinations_per_hundred',
'people_vaccinated_per_hundred', 'people_fully_vaccinated_per_hundred',
'total_boosters_per_hundred', 'new_vaccinations_smoothed_per_million',
'new_people_vaccinated_smoothed', 'new_people_vaccinated_smoothed_per_hundred',
'stringency_index', 'population_density', 'median_age', 'aged_65_older',
'aged_70_older', 'gdp_per_capita', 'extreme_poverty', 'cardiovasc_death_rate',
'diabetes_prevalence', 'female_smokers', 'male_smokers',
'handwashing_facilities', 'hospital_beds_per_thousand', 'life_expectancy',
'human_development_index', 'population', 'excess_mortality_cumulative_absolute',
'excess_mortality_cumulative', 'excess_mortality',
'excess_mortality_cumulative_per_million']
```

[10]:
```
df.head()
```

[10]:
```
  iso_code continent     location        date  total_cases  new_cases  \
0      AFG      Asia  Afghanistan  2020-01-03          NaN        0.0
1      AFG      Asia  Afghanistan  2020-01-04          NaN        0.0
2      AFG      Asia  Afghanistan  2020-01-05          NaN        0.0
3      AFG      Asia  Afghanistan  2020-01-06          NaN        0.0
4      AFG      Asia  Afghanistan  2020-01-07          NaN        0.0

   new_cases_smoothed  total_deaths  new_deaths  new_deaths_smoothed  …  \
0                 NaN           NaN         0.0                  NaN  …
1                 NaN           NaN         0.0                  NaN  …
2                 NaN           NaN         0.0                  NaN  …
3                 NaN           NaN         0.0                  NaN  …
4                 NaN           NaN         0.0                  NaN  …

   male_smokers  handwashing_facilities  hospital_beds_per_thousand  \
0           NaN                  37.746                         0.5
1           NaN                  37.746                         0.5
2           NaN                  37.746                         0.5
3           NaN                  37.746                         0.5
4           NaN                  37.746                         0.5

   life_expectancy  human_development_index  population  \
0            64.83                    0.511  41128772.0
1            64.83                    0.511  41128772.0
2            64.83                    0.511  41128772.0
3            64.83                    0.511  41128772.0
4            64.83                    0.511  41128772.0

   excess_mortality_cumulative_absolute  excess_mortality_cumulative  \
0                                   NaN                          NaN
1                                   NaN                          NaN
2                                   NaN                          NaN
3                                   NaN                          NaN
4                                   NaN                          NaN

   excess_mortality  excess_mortality_cumulative_per_million
0               NaN                                     NaN
1               NaN                                     NaN
2               NaN                                     NaN
3               NaN                                     NaN
4               NaN                                     NaN

[5 rows x 67 columns]
```

[11]:
```
# Checking for missing values
missing_values = df.isnull().sum().sort_values(ascending=False)
print("Missing values (top 10):")
```

```
print(missing_values.head(10))
```

```
Missing values (top 10):
weekly_icu_admissions                          339880
weekly_icu_admissions_per_million              339880
excess_mortality_cumulative_per_million        337901
excess_mortality                               337901
excess_mortality_cumulative                    337901
excess_mortality_cumulative_absolute           337901
weekly_hosp_admissions                         326832
weekly_hosp_admissions_per_million             326832
icu_patients_per_million                       312470
icu_patients                                   312470
dtype: int64
```

[12]:
```
# defining countries of interest: Kenya, Nigeria and South Africa
countries = ['Kenya', 'Nigeria', 'South Africa']

#Filtering the dataframe to only 3 countries(Kenya, Nigeria and South Africa)
df = df[df['location'].isin(countries)]
```

[13]:
```
# Dropping rows with missing critical values
df = df.dropna(subset=['total_cases', 'total_deaths'])
```

[15]:
```
# Filling missing values with 0
df.fillna(0, inplace=True)
```

[16]:
```
# Convert date to datetime format
df['date'] = pd.to_datetime(df['date'])
```

[18]:
```
# Checking data after cleaning
print(df.isnull().sum().sort_values(ascending=False).head(10))
df.info()
```

```
iso_code                                0
aged_65_older                           0
people_fully_vaccinated                 0
total_boosters                          0
new_vaccinations                        0
new_vaccinations_smoothed               0
total_vaccinations_per_hundred          0
people_vaccinated_per_hundred           0
people_fully_vaccinated_per_hundred     0
total_boosters_per_hundred              0
dtype: int64
<class 'pandas.core.frame.DataFrame'>
Index: 3906 entries, 157871 to 290726
```

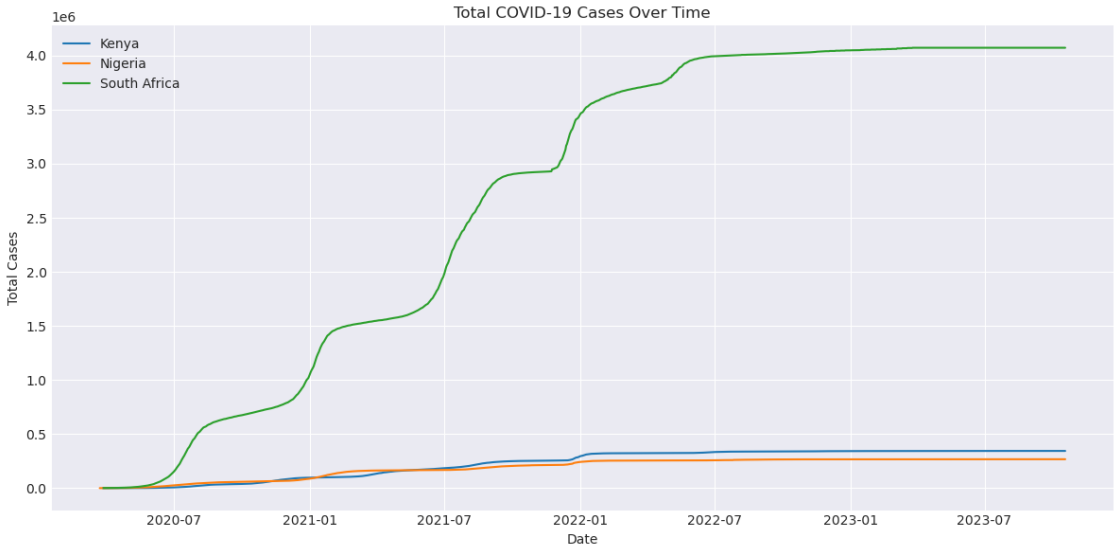
```
Data columns (total 67 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   iso_code                              3906 non-null   object
 1   continent                             3906 non-null   object
 2   location                              3906 non-null   object
 3   date                                  3906 non-null   datetime64[ns]
 4   total_cases                           3906 non-null   float64
 5   new_cases                             3906 non-null   float64
 6   new_cases_smoothed                    3906 non-null   float64
 7   total_deaths                          3906 non-null   float64
 8   new_deaths                            3906 non-null   float64
 9   new_deaths_smoothed                   3906 non-null   float64
 10  total_cases_per_million               3906 non-null   float64
 11  new_cases_per_million                 3906 non-null   float64
 12  new_cases_smoothed_per_million        3906 non-null   float64
 13  total_deaths_per_million              3906 non-null   float64
 14  new_deaths_per_million                3906 non-null   float64
 15  new_deaths_smoothed_per_million       3906 non-null   float64
 16  reproduction_rate                     3906 non-null   float64
 17  icu_patients                          3906 non-null   float64
 18  icu_patients_per_million              3906 non-null   float64
 19  hosp_patients                         3906 non-null   float64
 20  hosp_patients_per_million             3906 non-null   float64
 21  weekly_icu_admissions                 3906 non-null   float64
 22  weekly_icu_admissions_per_million     3906 non-null   float64
 23  weekly_hosp_admissions                3906 non-null   float64
 24  weekly_hosp_admissions_per_million    3906 non-null   float64
 25  total_tests                           3906 non-null   float64
 26  new_tests                             3906 non-null   float64
 27  total_tests_per_thousand              3906 non-null   float64
 28  new_tests_per_thousand                3906 non-null   float64
 29  new_tests_smoothed                    3906 non-null   float64
 30  new_tests_smoothed_per_thousand       3906 non-null   float64
 31  positive_rate                         3906 non-null   float64
 32  tests_per_case                        3906 non-null   float64
 33  tests_units                           3906 non-null   object
 34  total_vaccinations                    3906 non-null   float64
 35  people_vaccinated                     3906 non-null   float64
 36  people_fully_vaccinated               3906 non-null   float64
 37  total_boosters                        3906 non-null   float64
 38  new_vaccinations                      3906 non-null   float64
 39  new_vaccinations_smoothed             3906 non-null   float64
 40  total_vaccinations_per_hundred        3906 non-null   float64
 41  people_vaccinated_per_hundred         3906 non-null   float64
 42  people_fully_vaccinated_per_hundred   3906 non-null   float64
 43  total_boosters_per_hundred            3906 non-null   float64
 44  new_vaccinations_smoothed_per_million 3906 non-null   float64
```

```
 45  new_people_vaccinated_smoothed               3906 non-null   float64
 46  new_people_vaccinated_smoothed_per_hundred   3906 non-null   float64
 47  stringency_index                             3906 non-null   float64
 48  population_density                           3906 non-null   float64
 49  median_age                                   3906 non-null   float64
 50  aged_65_older                                3906 non-null   float64
 51  aged_70_older                                3906 non-null   float64
 52  gdp_per_capita                               3906 non-null   float64
 53  extreme_poverty                              3906 non-null   float64
 54  cardiovasc_death_rate                        3906 non-null   float64
 55  diabetes_prevalence                          3906 non-null   float64
 56  female_smokers                               3906 non-null   float64
 57  male_smokers                                 3906 non-null   float64
 58  handwashing_facilities                       3906 non-null   float64
 59  hospital_beds_per_thousand                   3906 non-null   float64
 60  life_expectancy                              3906 non-null   float64
 61  human_development_index                      3906 non-null   float64
 62  population                                   3906 non-null   float64
 63  excess_mortality_cumulative_absolute         3906 non-null   float64
 64  excess_mortality_cumulative                  3906 non-null   float64
 65  excess_mortality                             3906 non-null   float64
 66  excess_mortality_cumulative_per_million      3906 non-null   float64
dtypes: datetime64[ns](1), float64(62), object(4)
memory usage: 2.0+ MB
```

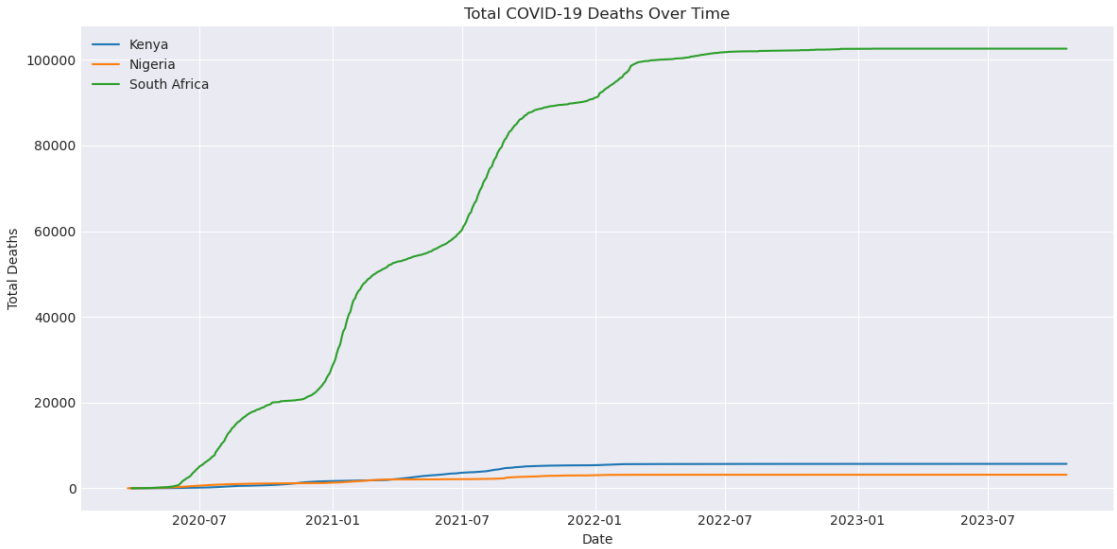
[19]:
```python
# Exploratory Data Analysis
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('seaborn-darkgrid')

#1. Plot of Total Covid-19 cases over time
plt.figure(figsize=(12, 6))

for country in countries:
    subset = df[df['location'] == country]
    plt.plot(subset['date'], subset['total_cases'], label=country)

plt.title("Total COVID-19 Cases Over Time")
plt.xlabel("Date")
plt.ylabel("Total Cases")
plt.legend()
plt.tight_layout()
plt.show()
```

```
/tmp/ipykernel_176/2708895129.py:4: MatplotlibDeprecationWarning: The seaborn
styles shipped by Matplotlib are deprecated since 3.6, as they no longer
correspond to the styles shipped by seaborn. However, they will remain available
as 'seaborn-v0_8-<style>'. Alternatively, directly use the seaborn API instead.
```

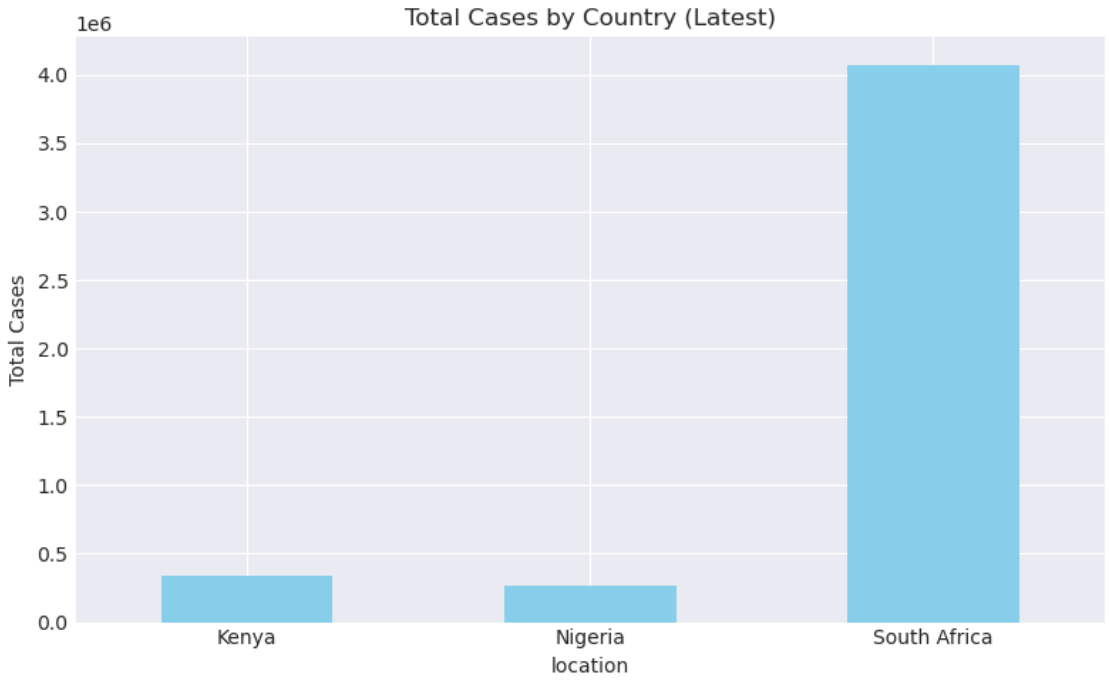
```
plt.style.use('seaborn-darkgrid')
```



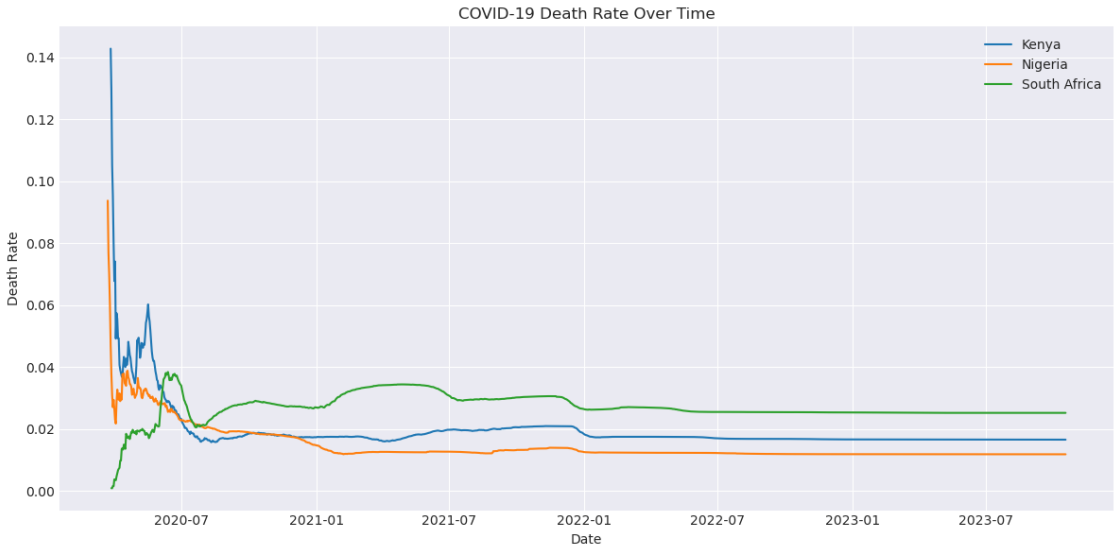
[20]:
```
# Plot of Total deaths over time
plt.figure(figsize=(12, 6))

for country in countries:
    subset = df[df['location'] == country]
    plt.plot(subset['date'], subset['total_deaths'], label=country)

plt.title("Total COVID-19 Deaths Over Time")
plt.xlabel("Date")
plt.ylabel("Total Deaths")
plt.legend()
plt.tight_layout()
plt.show()
```



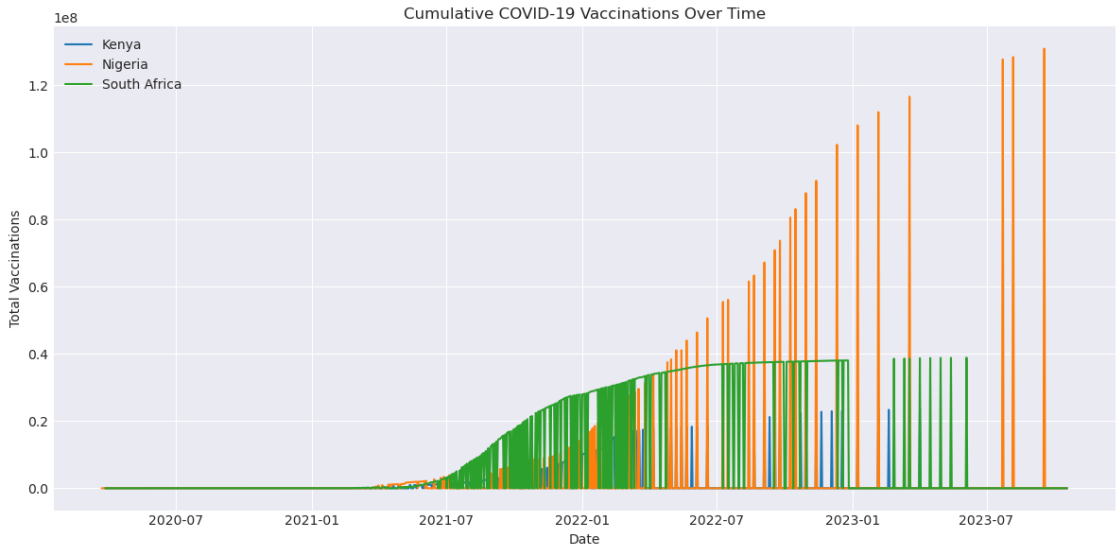
[21]:
```
# Bar Chart of latest cases per country
latest = df[df['date'] == df['date'].max()]
latest_cases = latest.groupby('location')['total_cases'].max()

latest_cases.plot(kind='bar', color='skyblue', figsize=(8, 5))
plt.title("Total Cases by Country (Latest)")
plt.ylabel("Total Cases")
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```



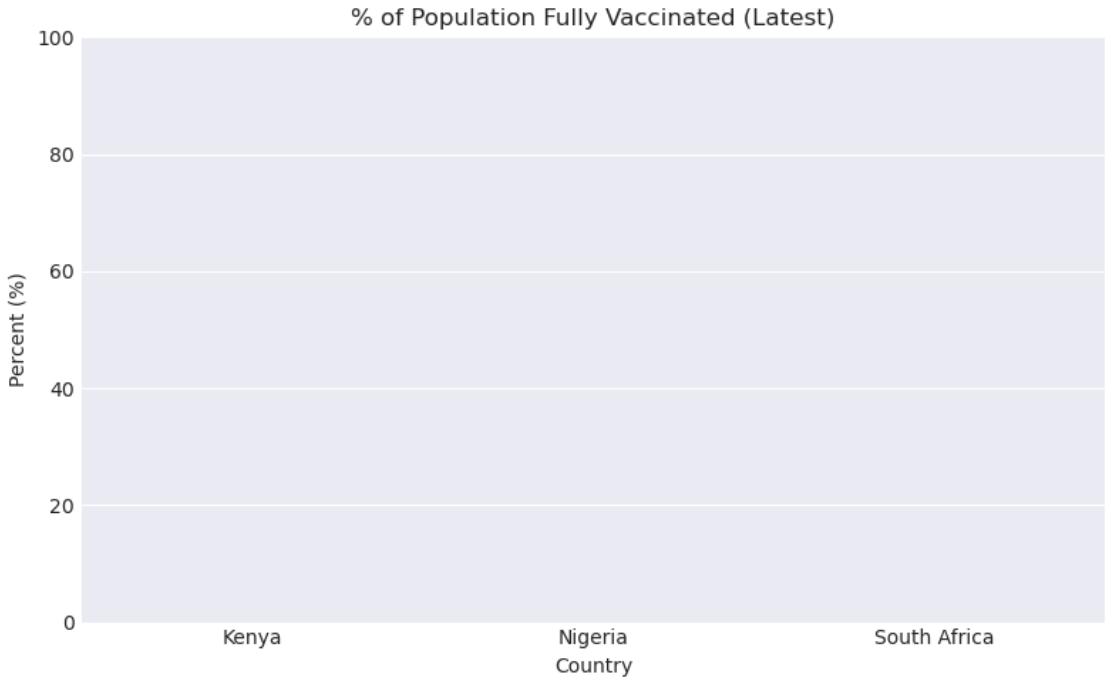
[22]:
```
# Calculation and plotting of death rate over time
df['death_rate'] = df['total_deaths'] / df['total_cases']

plt.figure(figsize=(12, 6))
for country in countries:
    subset = df[df['location'] == country]
    plt.plot(subset['date'], subset['death_rate'], label=country)

plt.title("COVID-19 Death Rate Over Time")
plt.xlabel("Date")
plt.ylabel("Death Rate")
plt.legend()
plt.tight_layout()
plt.show()
```



[23]:
```
# Plot of cumulative vaccinations over time
plt.figure(figsize=(12, 6))

for country in countries:
    subset = df[df['location'] == country]
    plt.plot(subset['date'], subset['total_vaccinations'], label=country)

plt.title("Cumulative COVID-19 Vaccinations Over Time")
plt.xlabel("Date")
plt.ylabel("Total Vaccinations")
plt.legend()
plt.tight_layout()
plt.show()
```



[25]:
```
# Bar Chart of percentage population fully vaccinated
latest = df[df['date'] == df['date'].max()].copy()
latest['percent_fully_vaccinated'] = (latest['people_fully_vaccinated'] /␣
 ↪latest['population']) * 100
# Plot
plt.figure(figsize=(8, 5))
sns.barplot(x='location', y='percent_fully_vaccinated', data=latest,␣
 ↪palette='viridis')
plt.title("% of Population Fully Vaccinated (Latest)")
plt.ylabel("Percent (%)")
plt.xlabel("Country")
plt.ylim(0, 100)
plt.tight_layout()
plt.show()
```



[ ]:
```
## Key Insights

- **South Africa** had the highest number of total cases and deaths, but also led in vaccination coverage.
- **Kenya** showed multiple case spikes but maintained a relatively low death rate.
- **Nigeria** had the flattest curve in total cases and vaccinations, possibly due to underreporting or logistical challenges.
- Death rates across the three countries decreased significantly following vaccine rollouts.
```