Date - 25/10/2023

Team ID - 3881

Project Title - Covid 19 Vaccines Analysis

Importing Dependencies

In [7]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Loading Dataset

In [8]:
```
dataset = pd.read_csv("C:\\Users\\yuvar\\Documents\\country_vaccinations.csv")
```

Data Exploration

In [9]:
```
dataset
```

Out[9]:

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccina
0	Afghanistan	AFG	2021-02-22	0.0	0.0	NaN	NaN	NaN	
1	Afghanistan	AFG	2021-02-23	NaN	NaN	NaN	NaN	1367.0	
2	Afghanistan	AFG	2021-02-24	NaN	NaN	NaN	NaN	1367.0	
3	Afghanistan	AFG	2021-02-25	NaN	NaN	NaN	NaN	1367.0	
4	Afghanistan	AFG	2021-02-26	NaN	NaN	NaN	NaN	1367.0	
...	...	...	...	...	...	...	...	...	
86507	Zimbabwe	ZWE	2022-03-25	8691642.0	4814582.0	3473523.0	139213.0	69579.0	
86508	Zimbabwe	ZWE	2022-03-26	8791728.0	4886242.0	3487962.0	100086.0	83429.0	
86509	Zimbabwe	ZWE	2022-03-27	8845039.0	4918147.0	3493763.0	53311.0	90629.0	
86510	Zimbabwe	ZWE	2022-03-28	8934360.0	4975433.0	3501493.0	89321.0	100614.0	
86511	Zimbabwe	ZWE	2022-03-29	9039729.0	5053114.0	3510256.0	105369.0	103751.0	

86512 rows × 15 columns

In [10]:
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 86512 entries, 0 to 86511
Data columns (total 15 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   country                             86512 non-null  object
 1   iso_code                            86512 non-null  object
 2   date                                86512 non-null  object
 3   total_vaccinations                  43607 non-null  float64
 4   people_vaccinated                   41294 non-null  float64
 5   people_fully_vaccinated             38802 non-null  float64
 6   daily_vaccinations_raw              35362 non-null  float64
 7   daily_vaccinations                  86213 non-null  float64
 8   total_vaccinations_per_hundred      43607 non-null  float64
 9   people_vaccinated_per_hundred       41294 non-null  float64
 10  people_fully_vaccinated_per_hundred 38802 non-null  float64
 11  daily_vaccinations_per_million      86213 non-null  float64
 12  vaccines                            86512 non-null  object
 13  source_name                         86512 non-null  object
 14  source_website                      86512 non-null  object
dtypes: float64(9), object(6)
memory usage: 9.9+ MB
```

In [11]:
```
dataset.describe()
```

Out[11]:

	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_per_hundred	people_va
count	4.360700e+04	4.129400e+04	3.880200e+04	3.536200e+04	8.621300e+04	43607.000000	
mean	4.592964e+07	1.770508e+07	1.413830e+07	2.705996e+05	1.313055e+05	80.188543	
std	2.246004e+08	7.078731e+07	5.713920e+07	1.212427e+06	7.682388e+05	67.913577	
min	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000	
25%	5.264100e+05	3.494642e+05	2.439622e+05	4.668000e+03	9.000000e+02	16.050000	
50%	3.590096e+06	2.187310e+06	1.722140e+06	2.530900e+04	7.343000e+03	67.520000	
75%	1.701230e+07	9.152520e+06	7.559870e+06	1.234925e+05	4.409800e+04	132.735000	
max	3.263129e+09	1.275541e+09	1.240777e+09	2.474100e+07	2.242429e+07	345.370000	

In [12]:
```
dataset.columns
```

Out[12]:
```
Index(['country', 'iso_code', 'date', 'total_vaccinations',
       'people_vaccinated', 'people_fully_vaccinated',
       'daily_vaccinations_raw', 'daily_vaccinations',
       'total_vaccinations_per_hundred', 'people_vaccinated_per_hundred',
       'people_fully_vaccinated_per_hundred', 'daily_vaccinations_per_million',
       'vaccines', 'source_name', 'source_website'],
      dtype='object')
```

Data Pre-Processing

Check for missing values

In [13]:
```
dataset.isnull()
```

Out[13]:

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinatio
0	False	False	False	False	False	True	True	True	
1	False	False	False	True	True	True	True	False	
2	False	False	False	True	True	True	True	False	
3	False	False	False	True	True	True	True	False	
4	False	False	False	True	True	True	True	False	
...	...	...	...	...	...	...	...	...	...
86507	False	False	False	False	False	False	False	False	
86508	False	False	False	False	False	False	False	False	
86509	False	False	False	False	False	False	False	False	
86510	False	False	False	False	False	False	False	False	
86511	False	False	False	False	False	False	False	False	

86512 rows × 15 columns

In [14]:
```
dataset.isnull().sum()
```

Out[14]:
```
country                                0
iso_code                               0
date                                   0
total_vaccinations                 42905
people_vaccinated                  45218
people_fully_vaccinated            47710
daily_vaccinations_raw             51150
daily_vaccinations                   299
total_vaccinations_per_hundred     42905
people_vaccinated_per_hundred      45218
people_fully_vaccinated_per_hundred    47710
daily_vaccinations_per_million       299
vaccines                               0
source_name                            0
source_website                         0
dtype: int64
```

In [15]:
```
dataset.dropna(inplace=True)
```

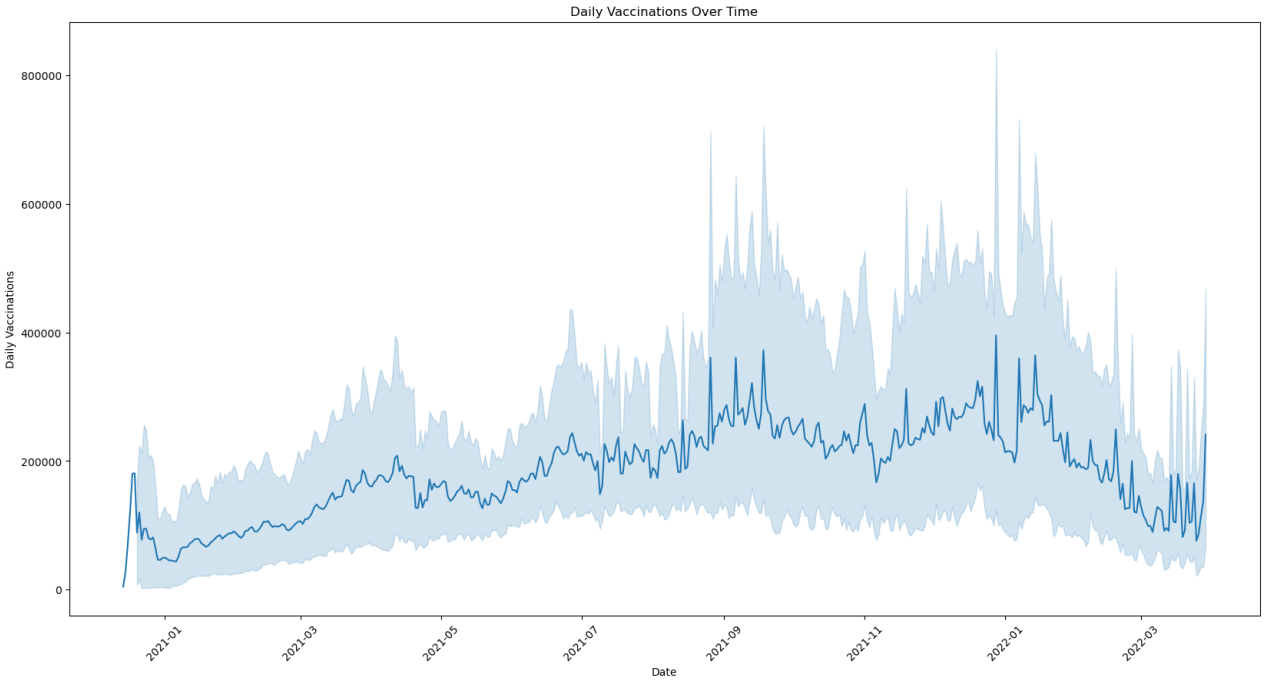
In [16]:
```
dataset.isnull().sum()
```

Out[16]:
```
country                                0
iso_code                               0
date                                   0
total_vaccinations                     0
people_vaccinated                      0
people_fully_vaccinated                0
daily_vaccinations_raw                 0
daily_vaccinations                     0
total_vaccinations_per_hundred         0
people_vaccinated_per_hundred          0
people_fully_vaccinated_per_hundred    0
daily_vaccinations_per_million         0
vaccines                               0
source_name                            0
source_website                         0
dtype: int64
```

Data Visualization

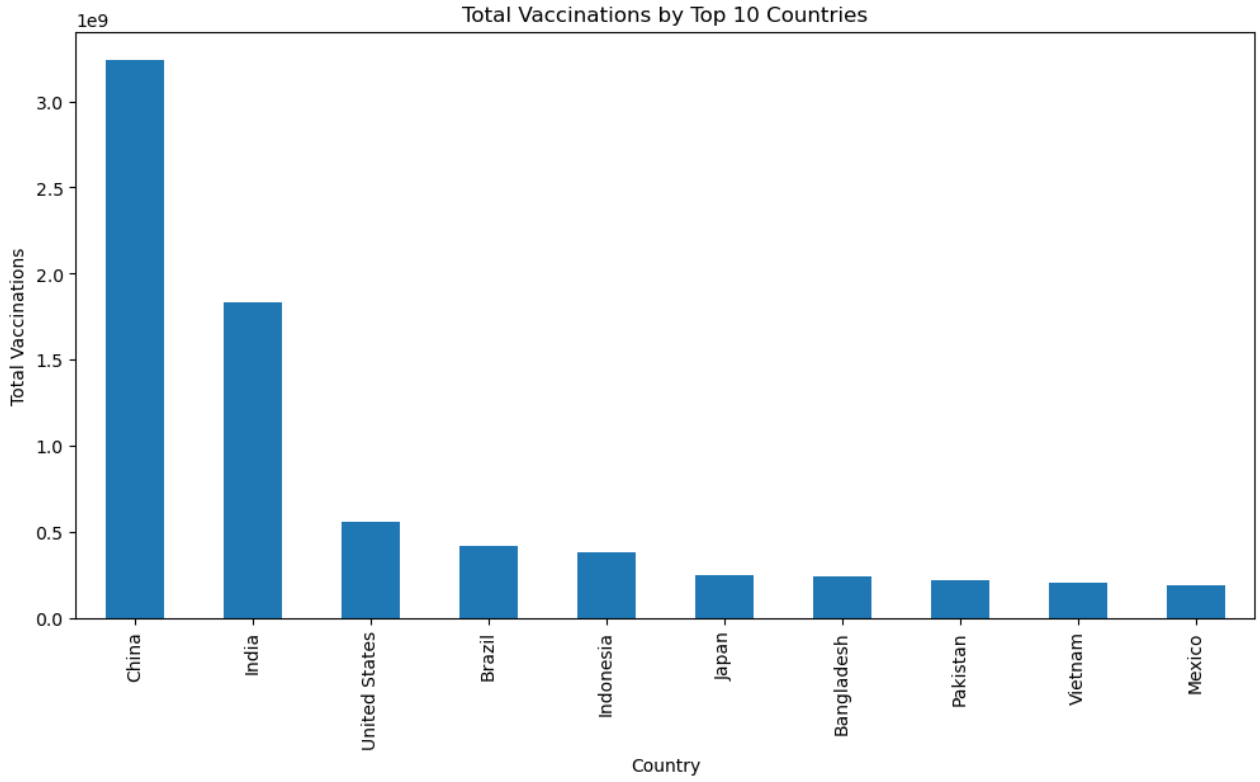
Plot a line chart for daily vaccinations over time

In [34]:
```
plt.figure(figsize=(20,10))
sns.lineplot(x='date', y='daily_vaccinations', data=dataset)
plt.title('Daily Vaccinations Over Time')
plt.xticks(rotation=45)
plt.xlabel('Date')
plt.ylabel('Daily Vaccinations')
plt.show()
```



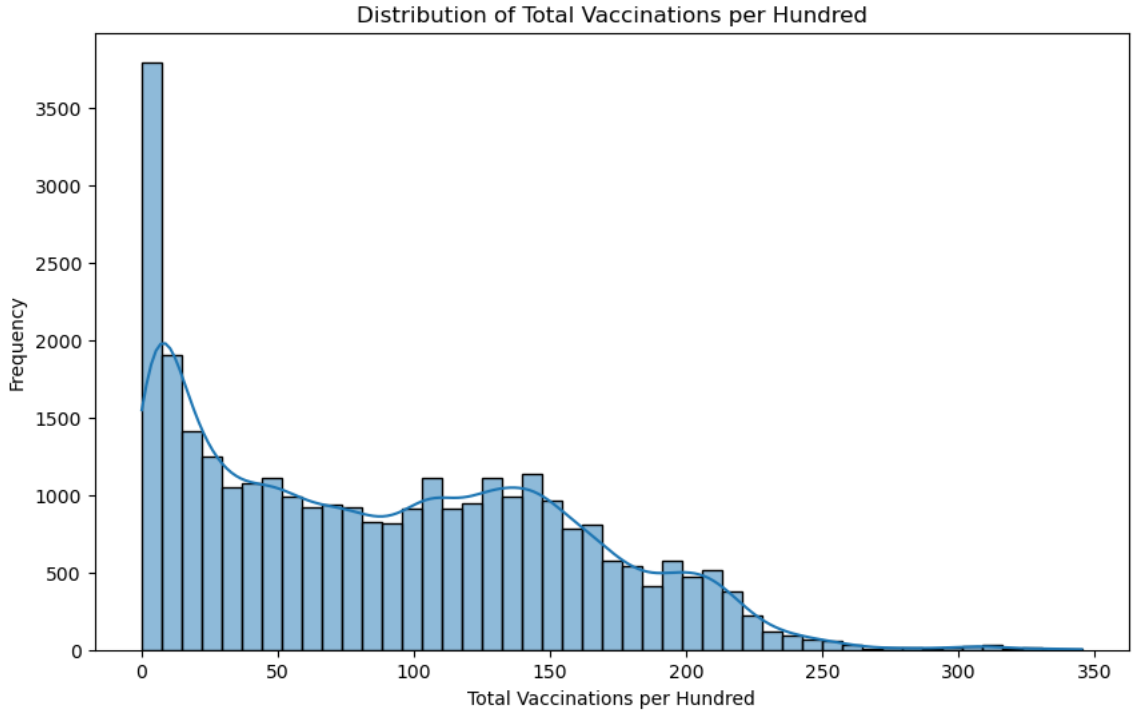
Create a bar chart to show total vaccinations by country

In [18]:
```
plt.figure(figsize=(12, 6))
total_vaccinations_by_country = dataset.groupby('country')['total_vaccinations'].max().sort_values(ascending=False)
total_vaccinations_by_country[:10].plot(kind='bar')
plt.title('Total Vaccinations by Top 10 Countries')
plt.xlabel('Country')
plt.ylabel('Total Vaccinations')
plt.show()
```



Create a histogram for total vaccinations per hundred

In [19]:
```
plt.figure(figsize=(10, 6))
sns.histplot(dataset['total_vaccinations_per_hundred'], kde=True)
plt.title('Distribution of Total Vaccinations per Hundred')
plt.xlabel('Total Vaccinations per Hundred')
plt.ylabel('Frequency')
plt.show()
```



Initial Data Exploration

Display basic statistics of the data

In [29]:
```
dataset.info()
dataset.head()
dataset.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 30847 entries, 2021-05-27 to 2022-03-29
Data columns (total 14 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   country                              30847 non-null  object
 1   iso_code                             30847 non-null  object
 2   total_vaccinations                   30847 non-null  float64
 3   people_vaccinated                    30847 non-null  float64
 4   people_fully_vaccinated              30847 non-null  float64
 5   daily_vaccinations_raw               30847 non-null  float64
 6   daily_vaccinations                   30847 non-null  float64
 7   total_vaccinations_per_hundred       30847 non-null  float64
 8   people_vaccinated_per_hundred        30847 non-null  float64
 9   people_fully_vaccinated_per_hundred  30847 non-null  float64
 10  daily_vaccinations_per_million       30847 non-null  float64
 11  vaccines                             30847 non-null  object
 12  source_name                          30847 non-null  object
 13  source_website                       30847 non-null  object
dtypes: float64(9), object(5)
memory usage: 3.5+ MB
```

Out[29]:

	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_per_hundred	people_va
count	3.084700e+04	3.084700e+04	3.084700e+04	3.084700e+04	3.084700e+04	30847.000000	
mean	3.980375e+07	2.177533e+07	1.579596e+07	2.021875e+05	1.975297e+05	88.609156	
std	1.451667e+08	8.053173e+07	5.898165e+07	7.041931e+05	6.400504e+05	67.492111	
min	3.000000e+00	3.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000	
25%	1.153332e+06	7.339795e+05	3.704450e+05	5.498000e+03	7.329500e+03	25.475000	
50%	6.335305e+06	3.688092e+06	2.211035e+06	2.908100e+04	3.247200e+04	81.470000	
75%	2.520629e+07	1.440668e+07	9.121526e+06	1.344580e+05	1.402915e+05	140.745000	
max	3.243599e+09	1.275541e+09	1.240777e+09	1.862727e+07	1.307071e+07	345.370000	

Statistical Analysis

Calculate summary statistics

In [23]:
```
summary_stats = dataset.describe()
```

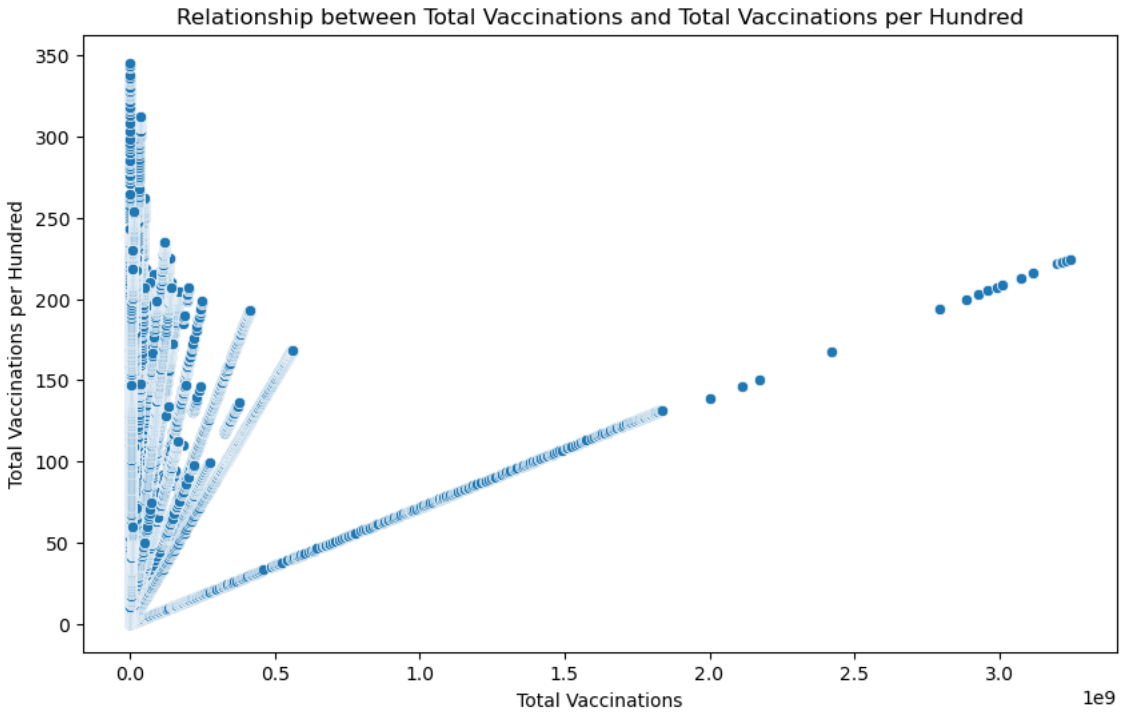
Visualize the summary statistics

In [33]:
```
plt.figure(figsize=(20,10))
sns.heatmap(summary_stats, annot=True,fmt='.0f')
plt.title('Summary Statistics')
plt.show()
```

Explore relationships between columns

Example: Relationship between 'total_vaccinations' and 'total_vaccinations_per_hundred

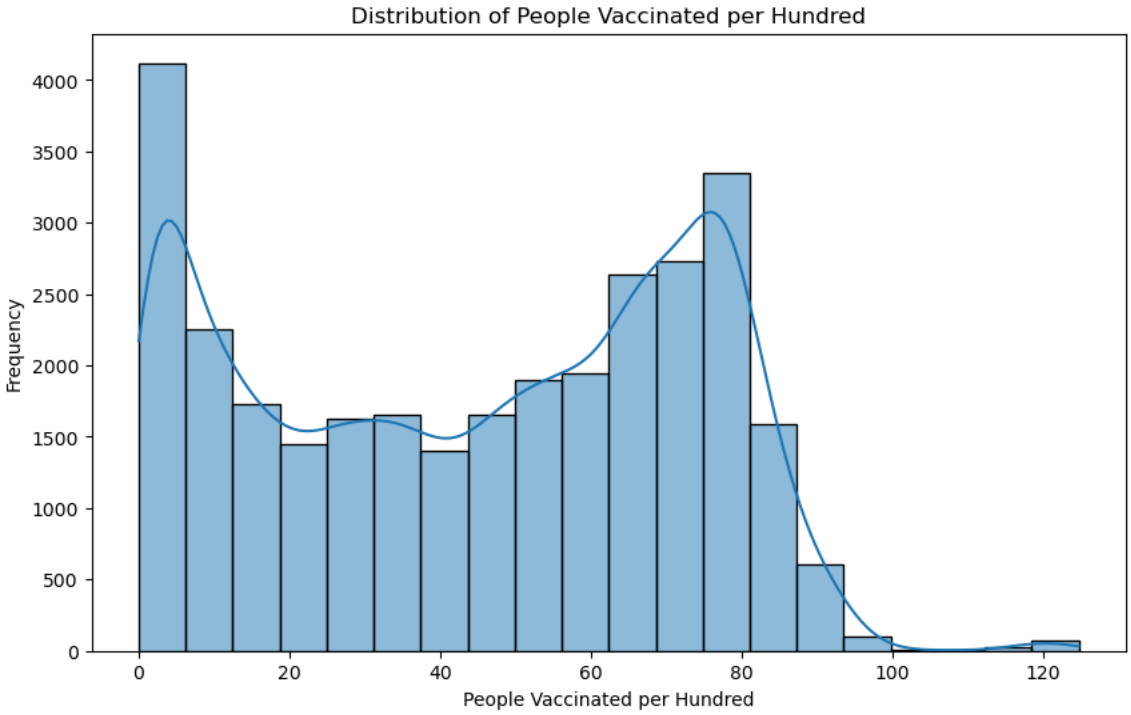
In [44]:
```
plt.figure(figsize=(10,6))
sns.scatterplot(x='total_vaccinations', y='total_vaccinations_per_hundred', data=dataset)
plt.xlabel('Total Vaccinations')
plt.ylabel('Total Vaccinations per Hundred')
plt.title('Relationship between Total Vaccinations and Total Vaccinations per Hundred')
plt.show()
```



Visualize distribution of key variables

Example: Distribution of 'people_vaccinated_per_hundred'

In [26]:
```
plt.figure(figsize=(10, 6))
sns.histplot(dataset['people_vaccinated_per_hundred'], bins=20, kde=True)
plt.xlabel('People Vaccinated per Hundred')
plt.ylabel('Frequency')
plt.title('Distribution of People Vaccinated per Hundred')
plt.show()
```

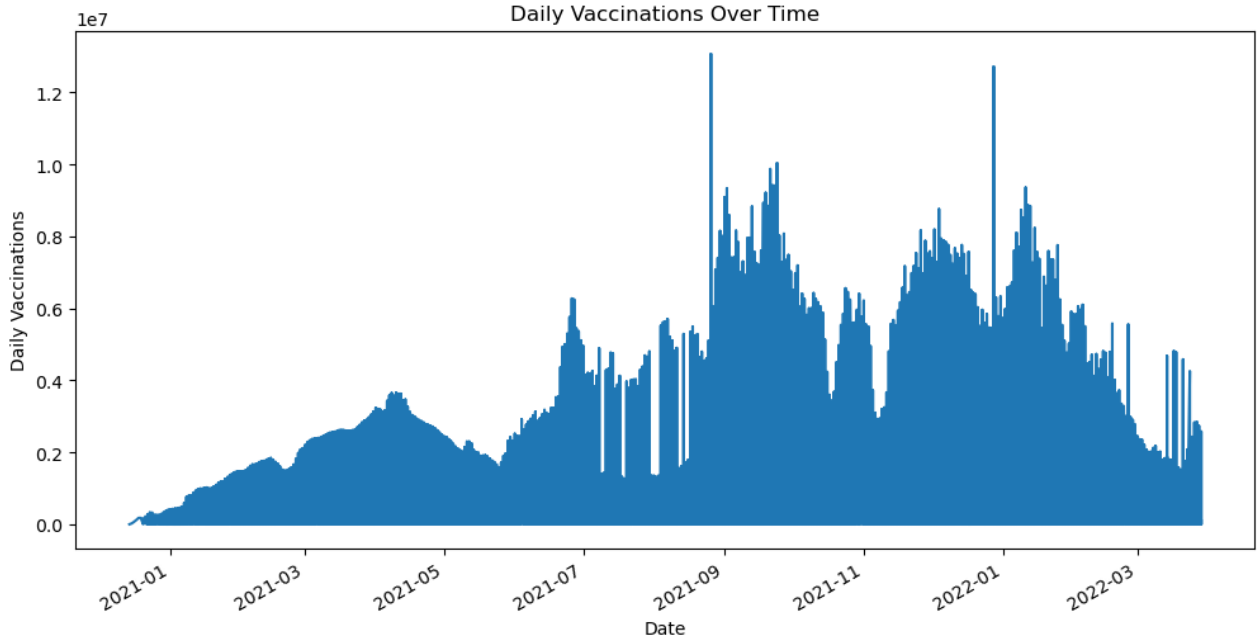


Time Series Analysis

In [27]:
```
dataset['date'] = pd.to_datetime(dataset['date'])
dataset.set_index('date', inplace=True)
```

Example: Daily Vaccinations over time

In [28]:
```
plt.figure(figsize=(12, 6))
dataset['daily_vaccinations'].plot()
plt.xlabel('Date')
plt.ylabel('Daily Vaccinations')
plt.title('Daily Vaccinations Over Time')
plt.show()
```



In [ ]: