COVID Data Analysis

BY:

- RANJANI ANJUR VENKATRAMAN
- CODY GARDNER
- SHARON COLSON

Agenda

1. General Information
2. Exploratory Data
3. Visualizations/Data Trends
4. Linear Regression
5. Hypothesis Tests
6. Conclusion
7. Questions

General Information

Background:

- Analyzing the trends in COVID data for different countries in the world

Resources/APIs:

- Historical API :
  - Using Historical API retrieved 30 days data of Cases, Deaths, and Recovered for multiple countries across world
  - Link : https://disease.sh/v3/covid-19/historical/
- Vaccination API :
  - Using Vaccination API retrieved vaccination information for multiple countries across world
  - Link : https://disease.sh/v3/covid-19/vaccine/coverage/countries?lastdays=1
- Daily data API :
  - Using Daily data API retrieved present day covid information
  - Link : https://disease.sh/v3/covid-19/countries/

Main COVID Data Frame

	Country	Cases	Deaths	Recovered	Vaccination_count	Latitude	Longitude	Population	Continent	Unvaccination_count
0	Afghanistan	137853.0	5983.0	82586.0	1046423.0	33.0	65.0	39869084	Asia	38822661.0
1	Albania	132638.0	2456.0	130064.0	1056075.0	41.0	20.0	2874408	Europe	1818333.0
2	Algeria	150504.5	3902.5	104103.0	3087405.0	28.0	3.0	44701619	Africa	41614214.0
3	Andorra	14256.0	127.0	13836.5	72315.0	42.5	1.6	77398	Europe	5083.0
4	Angola	40580.5	951.5	34672.0	1578873.0	-12.5	18.5	33961015	Africa	32382142.0

Group by Covid data

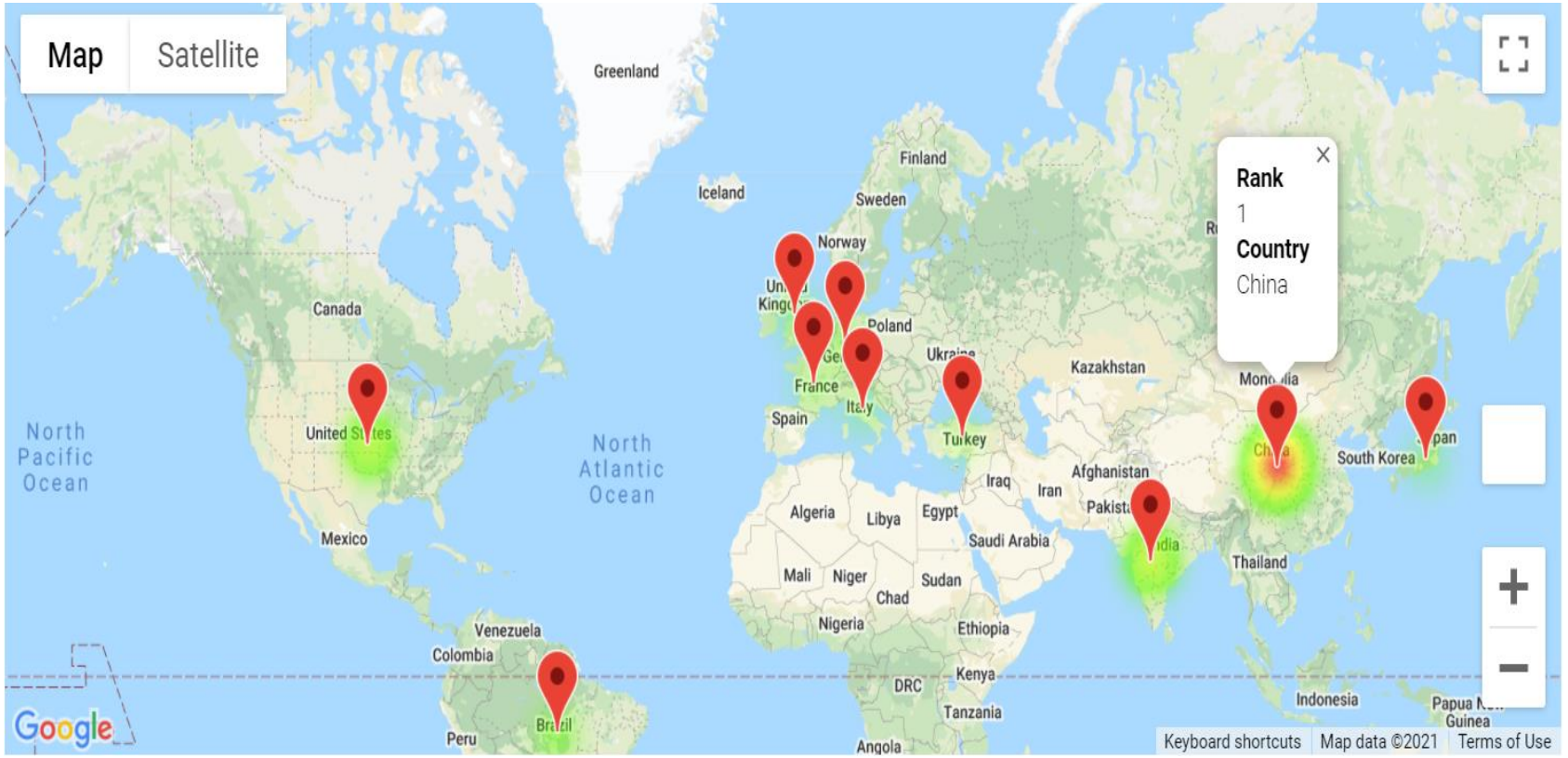
```
covid_groupby = covid_df.groupby('Country').median().reset_index()
covid_groupby.head(10)
```

	Country	Cases	Deaths	Recovered	Vaccination_count
0	Afghanistan	137853.0	5983.0	82586.0	1054347.0
1	Albania	132647.0	2456.0	130067.0	1060042.0
2	Algeria	151103.0	3910.0	104397.0	3097862.0
3	Andorra	14273.0	127.0	13844.0	72844.0
4	Angola	40631.0	952.0	34724.0	1581644.0
5	Antigua and Barbuda	1268.0	42.0	1224.0	65834.0
6	Argentina	4737213.0	101158.0	4363105.0	26511672.0
7	Armenia	227111.0	4558.0	218128.0	121954.0
8	Australia	20767.0	820.0	19893.0	9806809.0
9	Austria	653001.0	10728.0	639352.0	8983841.0

```
subset_today = today_df[['Country','Latitude','Longitude','Population','Continent']]
subset_today.head()
```

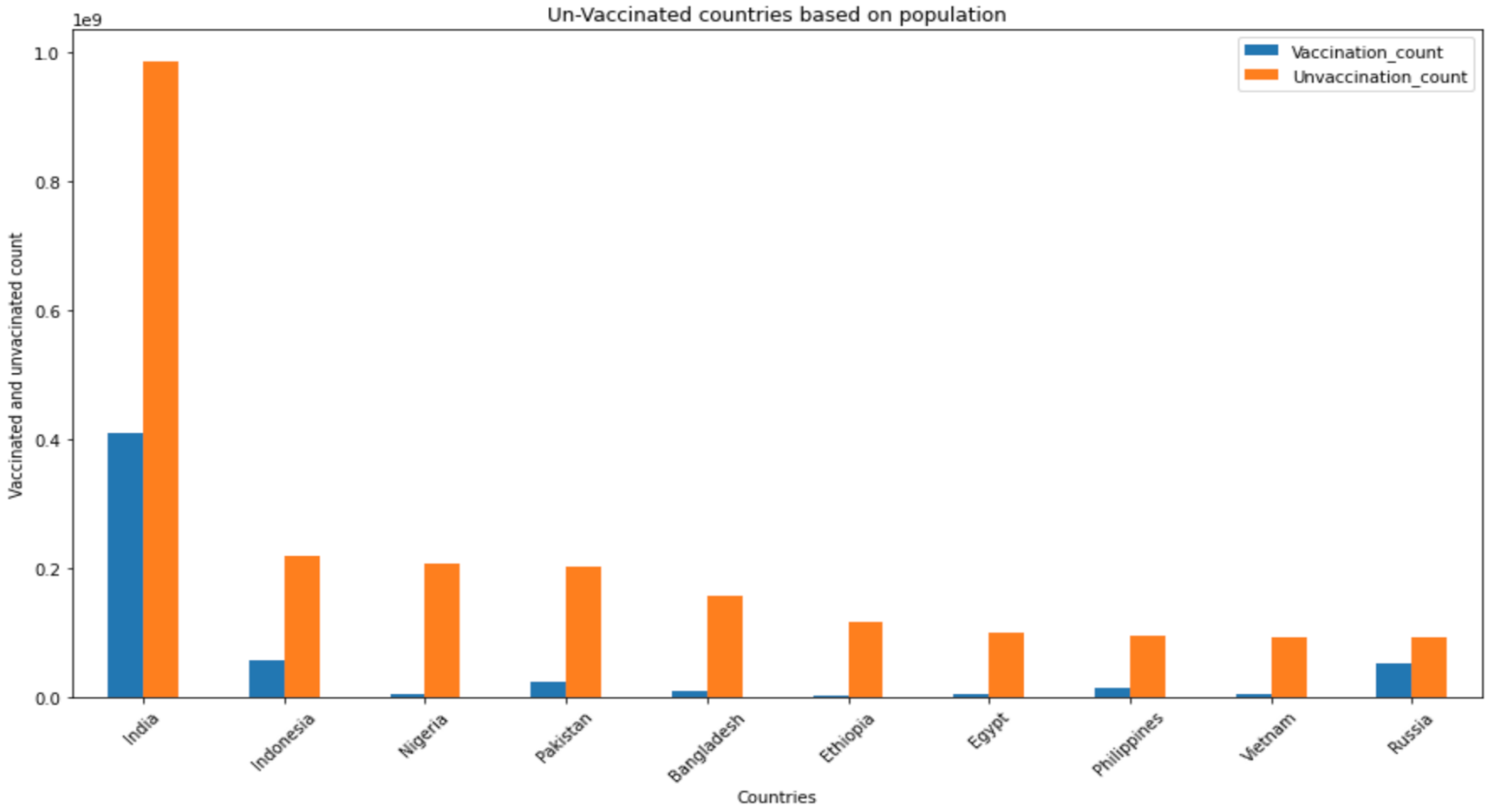
	Country	Latitude	Longitude	Population	Continent
0	Afghanistan	33.0	65.0	39869084	Asia
1	Albania	41.0	20.0	2874408	Europe
2	Algeria	28.0	3.0	44701619	Africa
3	Andorra	42.5	1.6	77398	Europe
4	Angola	-12.5	18.5	33961015	Africa

Top 10 Vaccinated Countries

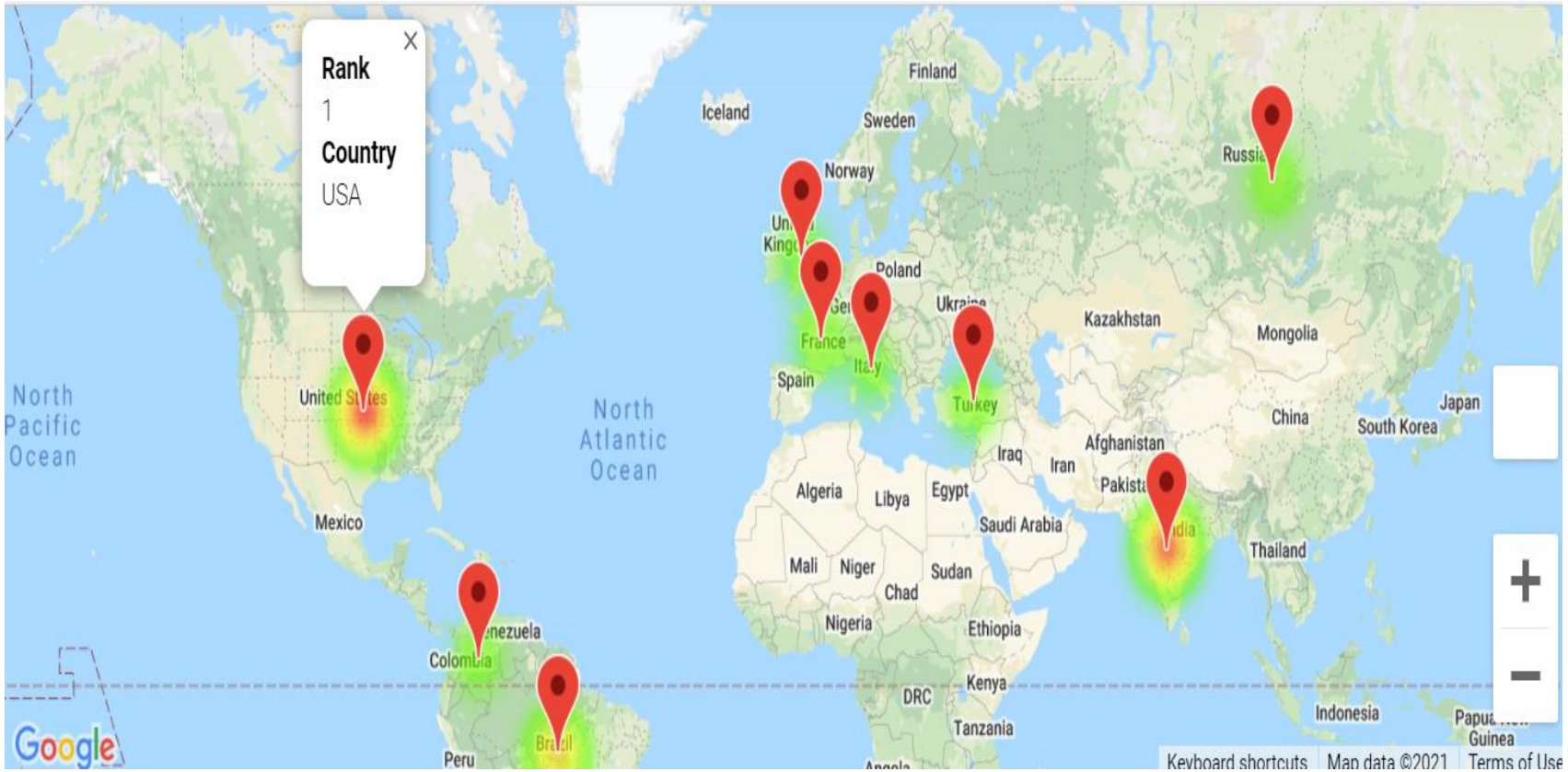


Rank	Country	Latitude	Longitude
1	China	35.0000	105.0000
2	India	20.0000	77.0000
3	USA	38.0000	-97.0000
4	Brazil	-10.0000	-55.0000
5	Germany	51.0000	9.0000
6	UK	54.0000	-2.0000
7	Japan	36.0000	138.0000
8	France	46.0000	2.0000
9	Turkey	39.0000	35.0000
10	Italy	42.8333	12.8333

Top 10 Unvaccinated Countries

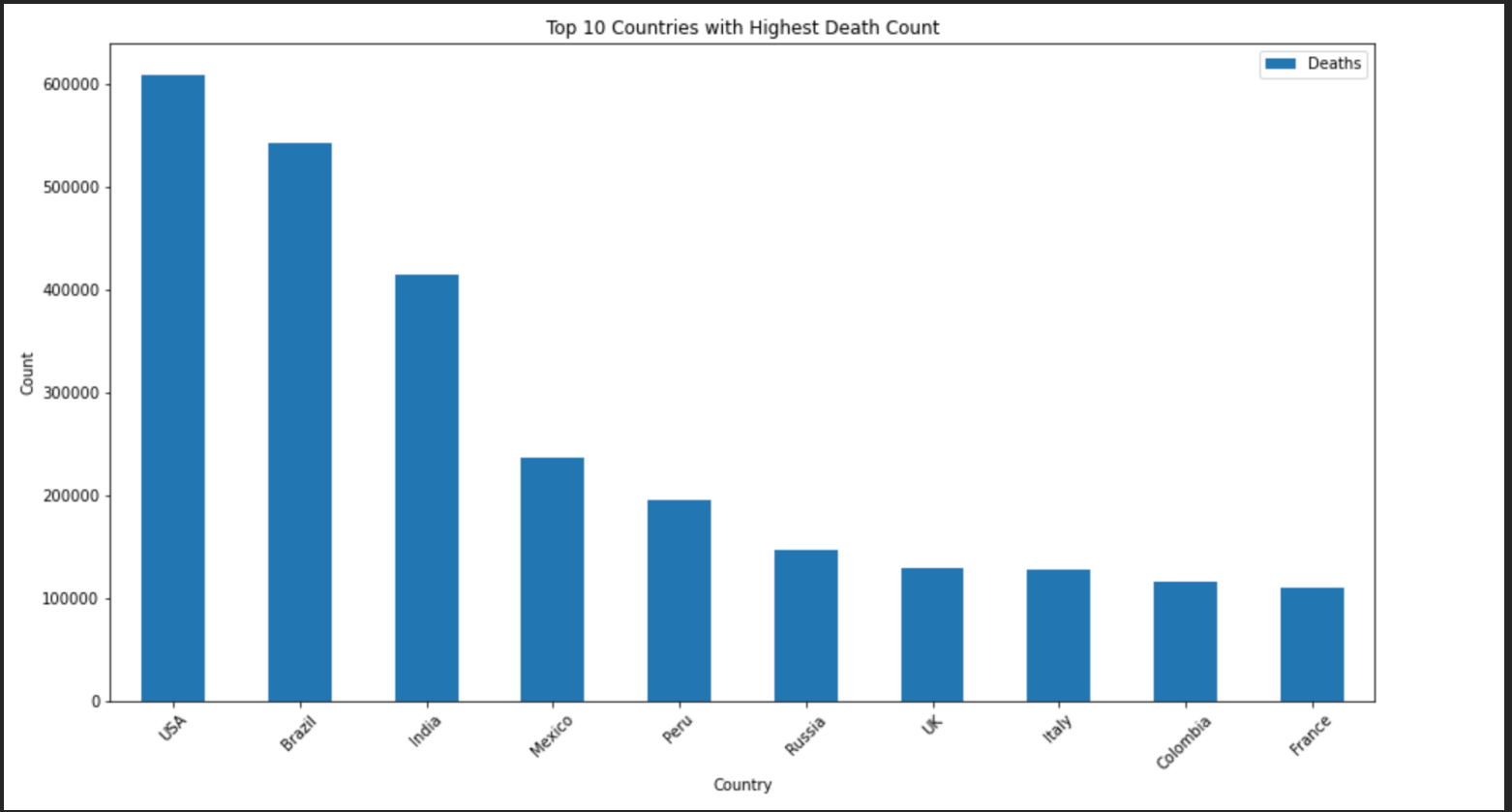


Top 10 Countries with Highest COVID Cases

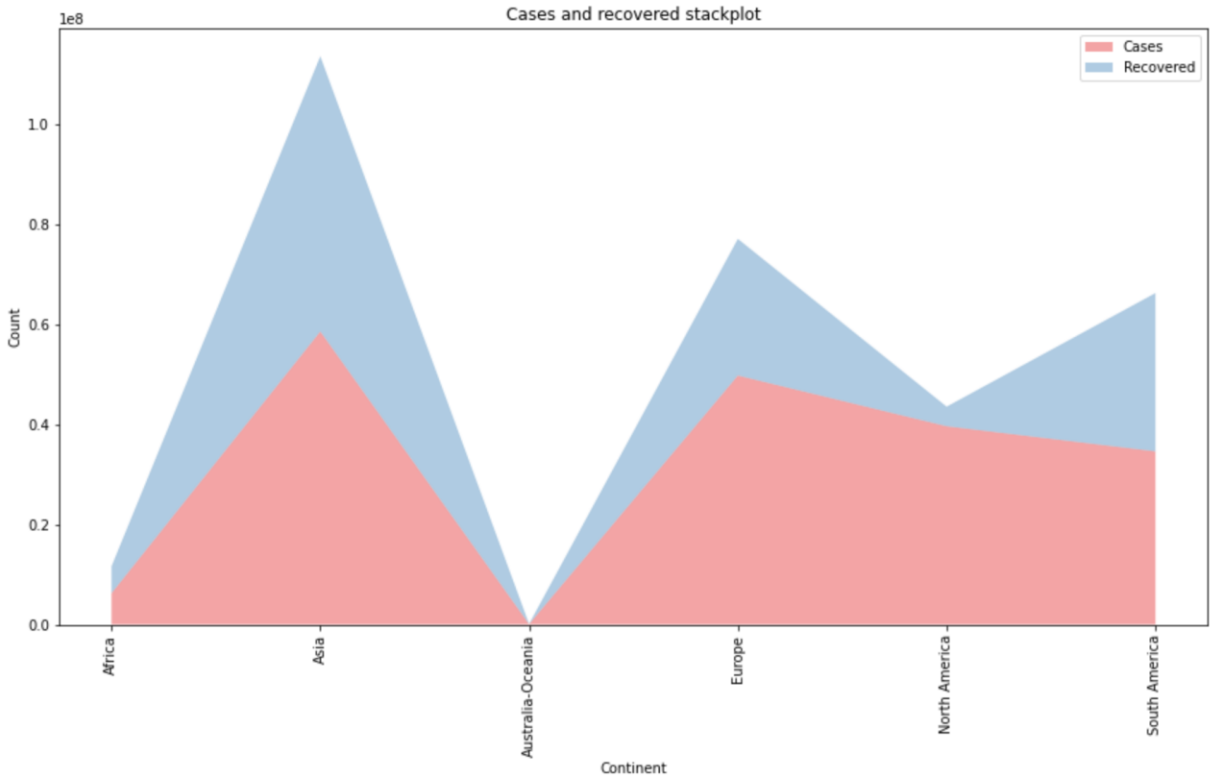


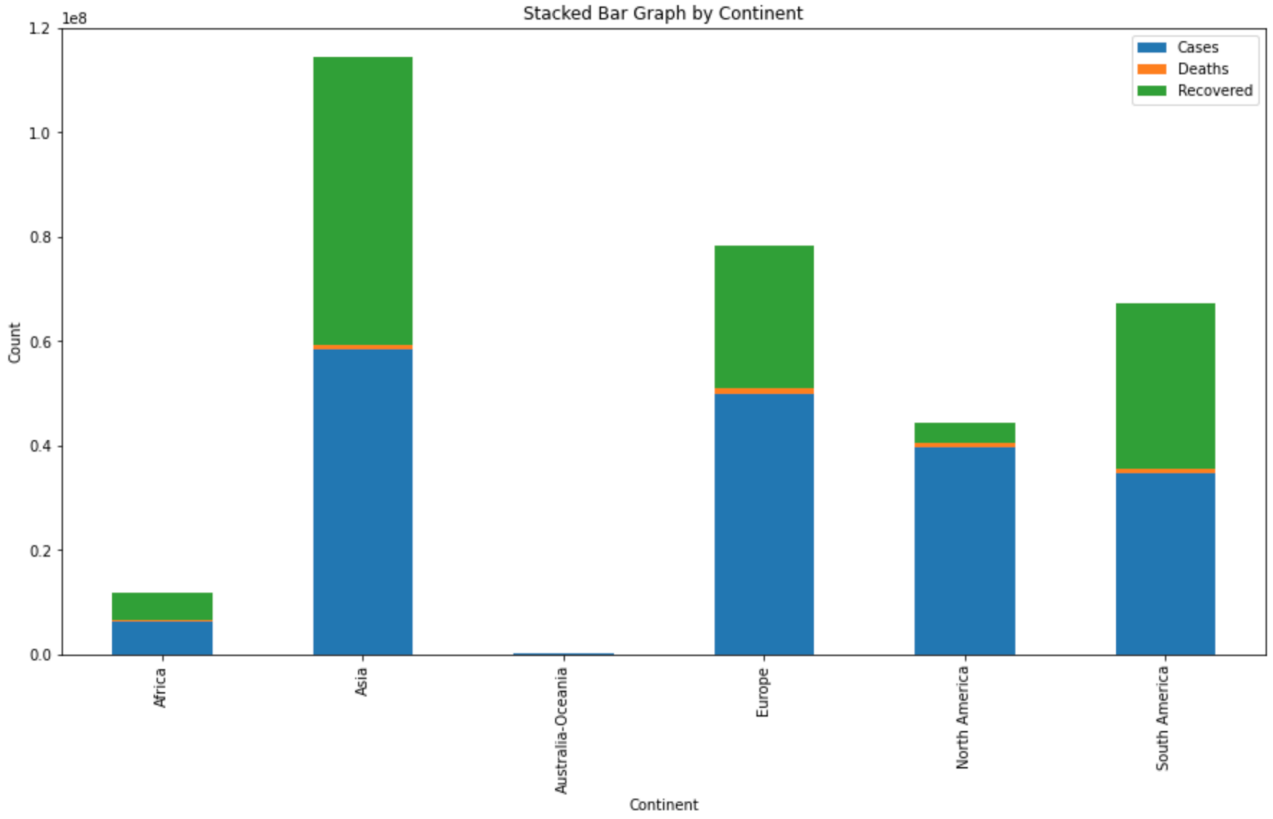
Rank	Country	Latitude	Longitude
1	USA	38.0000	-97.0000
2	India	20.0000	77.0000
3	Brazil	-10.0000	-55.0000
4	Russia	60.0000	100.0000
5	France	46.0000	2.0000
6	Turkey	39.0000	35.0000
7	UK	54.0000	-2.0000
8	Argentina	-34.0000	-64.0000
9	Colombia	4.0000	-72.0000
10	Italy	42.8333	12.8333

Top 10 Countries by Death Count



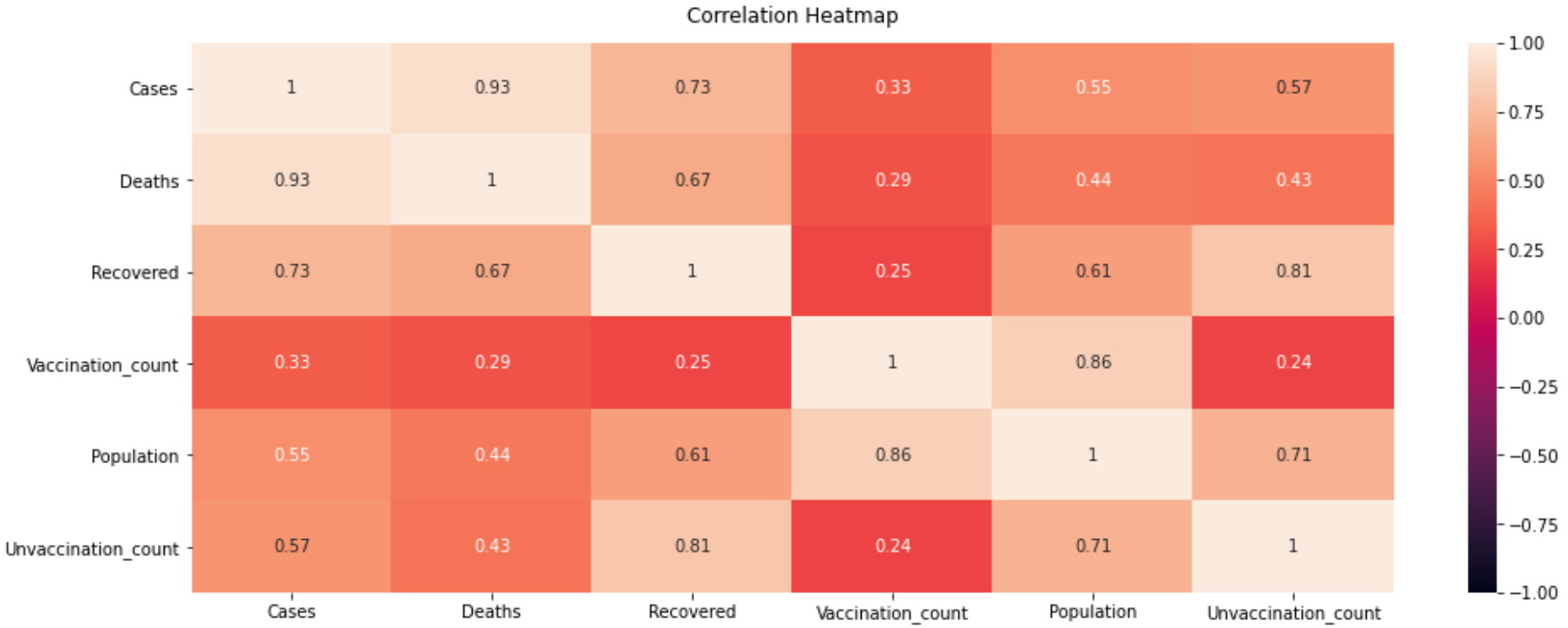
Cases and Recoveries





COVID Summary by Continent

Correlation Matrix



Pearson's Correlation Coefficient

- **Hypothesis:**
  - H0: Cases and Vaccination are independent
  - H1: There is a dependency between Cases and Vaccination
- **Inference:**
  - Since p-value is less than 0.05 we can reject null hypothesis and accept the alternate hypothesis that there is a dependency between cases and vaccination

- H0: Cases and Vaccination are independent

- H1: There is a dependency between Cases and Vaccination

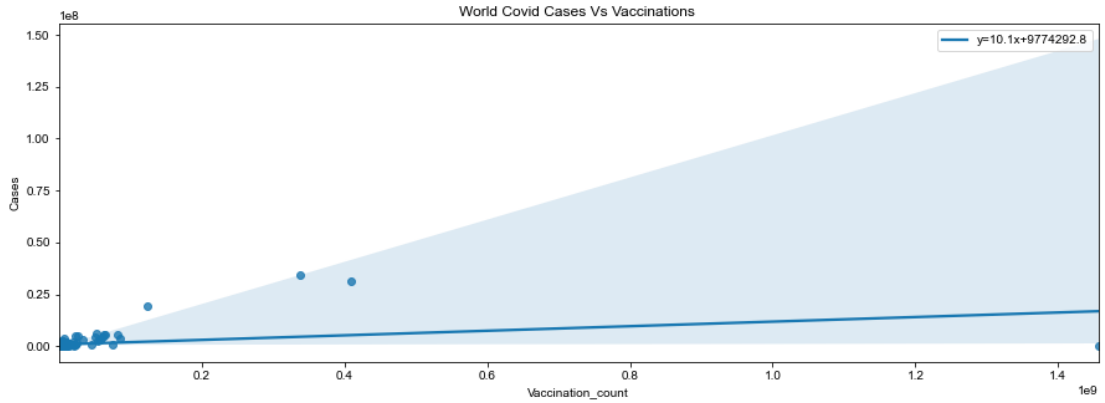
```
import scipy.stats as st
from scipy.stats import linregress

correlation=round(st.pearsonr(main_covid_df['Vaccination_count'],main_covid_df['Cases'])[0],2)
p_value = st.pearsonr(main_covid_df['Vaccination_count'],main_covid_df['Cases'])[1]
print(f"The correlation between cases and vaccination is {correlation}")
print(f"The p-value for cases and vaccination is {p_value}")
```

```
The correlation between cases and vaccination is 0.33
The p-value for cases and vaccination is 4.6514891598922895e-06
```

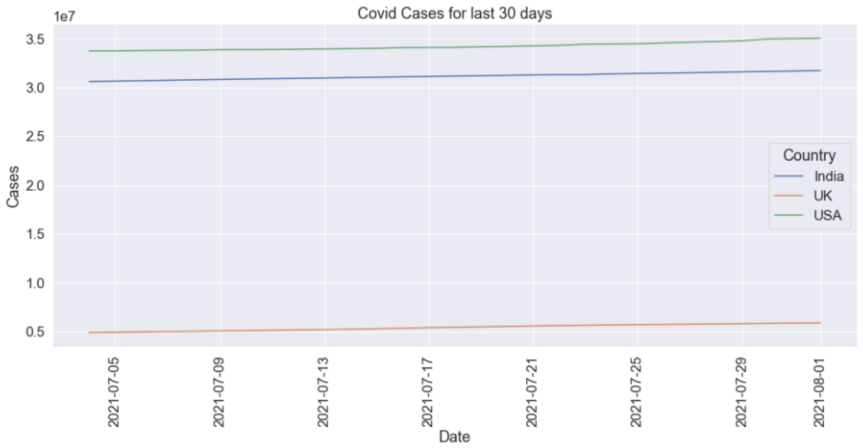
```
if p_value > 0.05:
        print('Probably cases and vaccination are not correlated')
else:
        print('Probably cases and vaccination have correlation')
```

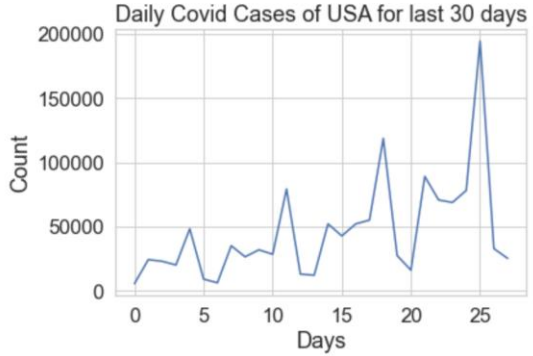
```
Probably cases and vaccination have correlation
```



Augmented Dickey Fuller Test

- **Hypothesis:**
  - H0: a unit root is present (series is non-stationary).
  - H1: a unit root is not present (series is stationary).
- **Inference:**
  - As p-value is greater than 0.05 and ADF statistic > greater than critical value we fail to reject null hypothesis. So, the time series data is non-stationary





In [49]:
```
X = usa_t
result = adfuller(X)
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
print('Critical Values:')
for key, value in result[4].items():
    print('\t%s: %.3f' % (key, value))

if result[0] < result[4]["5%"]:
    print ("Reject Ho - Time Series is Stationary")
else:
    print ("Failed to Reject Ho - Time Series is Non-Stationary")
```

```
ADF Statistic: 5.206378
p-value: 1.000000
Critical Values:
        1%: -3.809
        5%: -3.022
        10%: -2.651
Failed to Reject Ho - Time Series is Non-Stationary
```

Conclusion and Takeaways

Conclusion:

- Based on the analysis, we are seeing an upward trend and seasonality in COVID cases

Takeaways:

- The suggestion of this analysis to any business owner or manager would be to prepare an action plan for further restrictions due to rise in Covid-19 cases.
- Further analysis based on a larger time frame (say 90 or 180 days) might give additional information
- It could be that vaccination counts have not yet come to a number that would cause Covid-19 cases to fall or that the new Delta (or any subsequent) variant will cause Covid-19 cases to continue to rise despite increasing vaccination counts.

THANK YOU !!

ANY QUESTIONS??