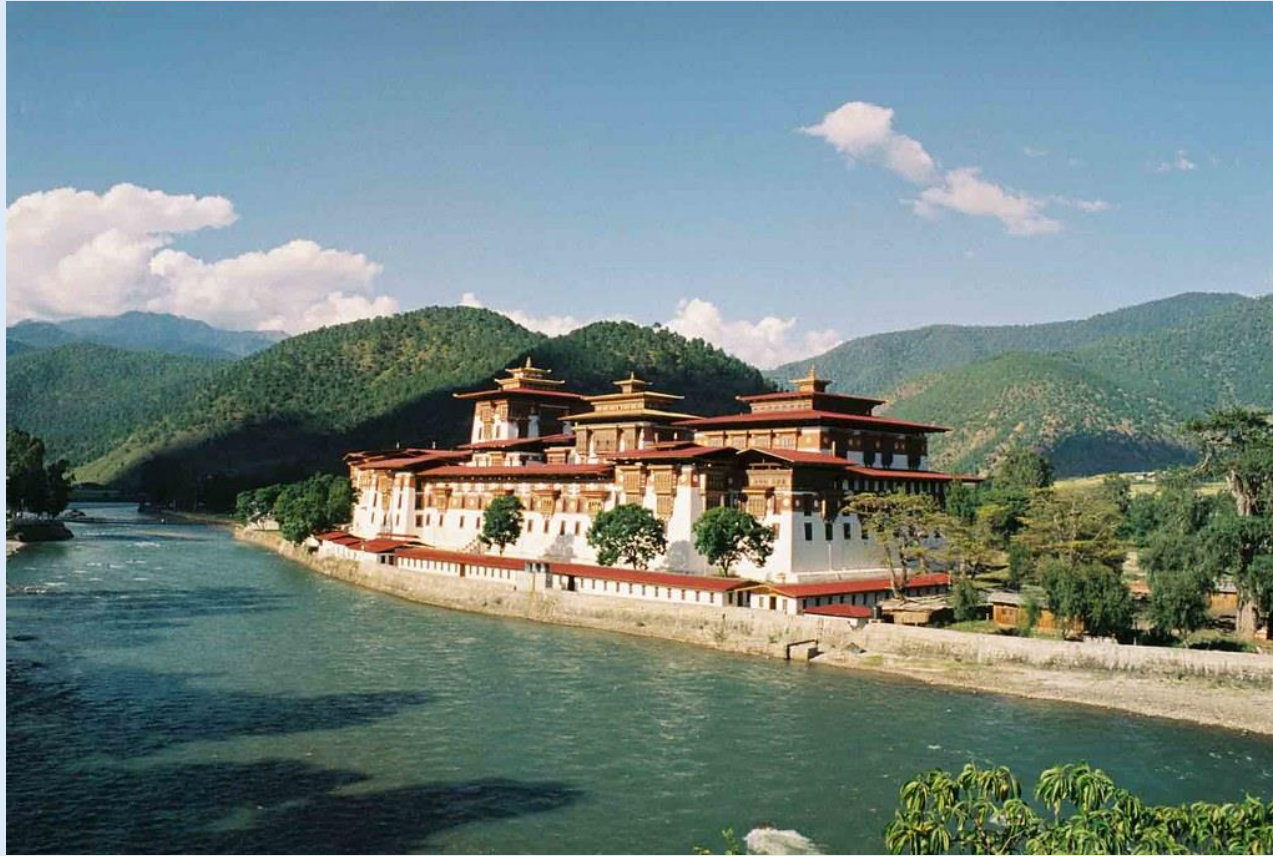


Exploratory Analysis of South Asian Countries



Marissa Hollins

12/30/2024

Overview

- Introduction
- Executive Summary
- Methodology
- Results
- Conclusion

Introduction

South Asia is a region comprised of 8 countries consisting of Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka. The economic traits of this collection of countries are diverse regarding emissions levels, sanitation services, and life expectancy. Factors such as geographical location, urban population, and political influence among others are the driving shift that leads to the resulting traits. The goal of the analysis I have performed is to discover eras of economic growth and decline, as well as which countries take the lead in development and academics. To achieve this, I have carried out times series and comparison analysis on a Kaggle dataset containing South Asian economic data.

Executive Summary

Using data obtained via Kaggle.com my analysis seeks to gain insight into South Asia's economic patterns. Focus was placed on, but not limited to, inflation's correlation to poverty and unemployment rates. As well as the correlation between urban population percentages and academic performance.

The data was cleaned and preprocessed using a combination of Google Sheets and Python; specifically, the Pandas and Numpy libraries. Google Sheets was used to remove special characters that aren't readable to Pandas, while Pandas and Numpy were used to locate and replace null values within a field and replace them with the field's mean value.

Findings resulting from my comparative analysis show that Maldives, Sri Lanka, and Bangladesh take the lead in development, low mortality rates, and academics. Times series analysis revealed a spike in unemployment rates during times that inflation has had a significant increase, particularly in 2005, 2017 and 2020. The results of my analysis were stored into a Pandas dataframe and later visualized using Tableau.

Methodology

Executive Summary

- **Data Collection:**
 - Dataset downloaded via Kaggle.com
- **Data Wrangling:**
 - Used Google Sheets to remove duplicate rows and special characters unreadable to Pandas.
 - Used Pandas and Numpy to correct data types and calculate mean values of columns to fill in null values.
 - Utilized Find and Replace in Google Sheets to plug in mean values.
- **Data Profiling:**
 - Performed exploratory analysis within Python kernel using Pandas.
 - Visualized results of analysis within Tableau.

Found null values and converted them to a numeric datatype using Numpy. Followed by filling the nulls with 0 to calculate the mean value of the columns with missing values.

```
In [147]: import pandas as pd
import numpy as np

OriginalSouthAsianDF = pd.read_csv('OriginalSouthAsianDataset.csv')
OriginalSouthAsianDF

OriginalSouthAsianDF = OriginalSouthAsianDF.iloc[0:192, :]
OriginalSouthAsianDF
OriginalSouthAsianDF.isnull().sum()
OriginalSouthAsianDF.replace('NaN', np.nan)
OriginalSouthAsianDF = OriginalSouthAsianDF.fillna(0)
OriginalSouthAsianDF
```

Out[147]:

	Country	Year	GDP (current US\$)	GDP growth (annual %)	GDP per capita (current US\$)	Unemployment, total (% of total labor force) (modeled ILO estimate)	Inflation, consumer prices (annual %)	Foreign direct investment, net inflows (% of GDP)	Trade (% of GDP)	Gini Index	Population, total	Population growth (annual %)	Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)	ex
0	Afghanistan	2000.0	3.521418e+09	0.00	180.19	7.96	0.00	0.00	0.00	0.0	1.954298e+07	1.44	0.0	
1	Afghanistan	2001.0	2.813572e+09	-9.43	142.90	7.96	0.00	0.02	0.00	0.0	1.968863e+07	0.74	0.0	
2	Afghanistan	2002.0	3.825701e+09	28.60	182.17	7.94	0.00	1.31	0.00	0.0	2.100026e+07	6.45	0.0	
3	Afghanistan	2003.0	4.520947e+09	8.83	199.64	7.92	0.00	1.28	0.00	0.0	2.264513e+07	7.54	0.0	
4	Afghanistan	2004.0	5.224897e+09	1.41	221.83	7.91	0.00	3.58	0.00	0.0	2.355355e+07	3.93	0.0	
5	Afghanistan	2005.0	6.203257e+09	11.23	254.12	7.91	12.69	3.58	0.00	0.0	2.441119e+07	3.58	0.0	

A combination of Pandas and Numpy was used to return the mean of all columns from column index 2 to 33. These values were used to replace the 0 values in their respective categorical column.

```
In [42]: OriginalSouthAsianDF.iloc[:,2:33].apply(np.mean)

Out[42]: GDP (current US$)                287938602444.25
GDP growth (annual %)                    4.99
GDP per capita (current US$)             2138.91
Unemployment, total (% of total labor force) (modeled ILO estimate)    6.29
Inflation, consumer prices (annual %)    6.31
Foreign direct investment, net inflows (% of GDP)    1.86
Trade (% of GDP)                          36.37
Gini index                               7.50
Population, total                        211093808.79
Population growth (annual %)             1.69
Poverty headcount ratio at $2.15 a day (2017 PPP) (% of population)    2.74
Life expectancy at birth, total (years)   65.53
Mortality rate, infant (per 1,000 live births)    37.49
Literacy rate, adult total (% of people ages 15 and above)    22.21
School enrollment, primary (% gross)       88.12
Urban population (% of total population)    28.86
Access to electricity (% of population)    73.20
People using at least basic drinking water services (% of population)    82.30
People using at least basic sanitation services (% of population)    55.23
Carbon dioxide (CO2) emissions excluding LULUCF per capita (t CO2e/capita)    1.05
PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)    40.46
Renewable energy consumption (% of total final energy consumption)    44.01
Forest area (% of land area)              22.14
Control of Corruption: Percentile Rank     31.12
Political Stability and Absence of Violence/Terrorism: Estimate    -0.88
Regulatory Quality: Estimate              -0.58
Rule of Law: Estimate                    -0.44
Voice and Accountability: Estimate        -0.46
Individuals using the Internet (% of population)    15.77
Research and development expenditure (% of GDP)    0.11
High-technology exports (% of manufactured exports)    0.99
dtype: float64
```

Pandas was used to check the data types of each column. This was followed by revising the data types as needed prior to analysis.

```
In [18]: OriginalSouthAsianDF['Year'] = OriginalSouthAsianDF['Year'].astype('int64')
OriginalSouthAsianDF.dtypes
```

```
Out[18]: Country                object
Year                int64
GDP (current US$)    float64
GDP growth (annual %) float64
GDP per capita (current US$) float64
Unemployment, total (% of total labor force) (modeled ILO estimate) float64
Inflation, consumer prices (annual %) float64
Foreign direct investment, net inflows (% of GDP) float64
Trade (% of GDP) float64
Gini index float64
Population, total float64
Population growth (annual %) float64
Poverty headcount ratio at $2.15 a day (2017 PPP) (% of population) float64
Life expectancy at birth, total (years) float64
Mortality rate, infant (per 1,000 live births) float64
Literacy rate, adult total (% of people ages 15 and above) float64
School enrollment, primary (% gross) float64
Urban population (% of total population) float64
Access to electricity (% of population) float64
People using at least basic drinking water services (% of population) float64
People using at least basic sanitation services (% of population) float64
Carbon dioxide (CO2) emissions excluding LULUCF per capita (t CO2e/capita) float64
PM2.5 air pollution, mean annual exposure (micrograms per cubic meter) float64
Renewable energy consumption (% of total final energy consumption) float64
Forest area (% of land area) float64
Control of Corruption: Percentile Rank float64
Political Stability and Absence of Violence/Terrorism: Estimate float64
Regulatory Quality: Estimate float64
Rule of Law: Estimate float64
Voice and Accountability: Estimate float64
Individuals using the Internet (% of population) float64
Research and development expenditure (% of GDP) float64
High-technology exports (% of manufactured exports) float64
dtype: object
```


Python was utilized to discover insights into the correlation between South Asia’s life expectancy and mortality rates by country and throughout the years. The results were stored into a Pandas dataframe and brought to life in Tableau.

```
In [17]: df_copy = pd.DataFrame(df.groupby('Year')['Life expectancy at birth, total (years)'].mean())
df_copy['Life Expectancy'] = df.groupby('Year')['Life expectancy at birth, total (years)'].mean()
df_copy['Mortality Rate'] = df.groupby('Year')['Mortality rate, infant (per 1,000 live births)'].mean()
df_copy.drop('Life expectancy at birth, total (years)', axis=1, inplace=True)
df_copy
```

```
Out[17]:
```

	Life Expectancy	Mortality Rate
Year		
2000	63.96375	58.6625
2001	64.65000	56.3000
2002	65.13250	54.0125
2003	65.67750	51.8250
2004	65.42750	50.5125
2005	66.40875	47.7875
2006	66.79875	45.8500
2007	67.06000	44.0375
2008	67.41000	42.3000
2009	67.41625	41.0875
2010	68.37000	39.0875
2011	68.74125	37.5675
2012	69.19125	36.1125
2013	69.57250	34.7375
2014	69.88750	33.4000
2015	70.11500	32.1375
2016	70.60125	30.9250
2017	70.83500	29.7625
2018	71.08875	28.7000
2019	71.39125	27.6500
2020	71.01250	26.6375
2021	70.53625	25.7125
2022	71.36375	24.8625
2023	65.79125	37.5000

```
In [20]: df_copy2 = pd.DataFrame(df.groupby('Country')['Life expectancy at birth, total (years)'].mean())
df_copy2['AVG Life Expectancy'] = df.groupby('Country')['Life expectancy at birth, total (years)'].mean()
df_copy2['AVG Mortality Rate'] = df.groupby('Country')['Mortality rate, infant (per 1,000 live births)'].mean()
df_copy2.drop('Life expectancy at birth, total (years)', axis=1, inplace=True)
df_copy2
```

```
Out[20]:
```

	AVG Life Expectancy	AVG Mortality Rate
Country		
Afghanistan	60.699583	63.733333
Bangladesh	69.134167	39.416667
Bhutan	68.163333	34.408333
India	67.043750	44.687500
Maldives	76.745000	14.504167
Nepal	66.838333	37.887500
Pakistan	64.625833	66.775000
Sri Lanka	72.897500	10.983333

Time series analysis showed a gradual uptick in life expectancy across the region from 2000-2022 before suddenly dropping 0.35% in 2023. On the flipside, the mortality rate is seen decreasing until having a spike of 1.35% in 2023. Bangladesh, Maldives, and Sri Lanka lead the pack with the best life expectancy and mortality rates.



Python was utilized to discover insights into inflation's impact on poverty and GDP's impact on inflation. This was done by creating a time series plot. The results were stored into a Pandas dataframe and brought to life in Tableau.

```
In [27]: df_copy9 = pd.DataFrame(df.groupby('Year')['Inflation, consumer prices (annual %)'].mean())
df_copy9['AVG Inflation %'] = df.groupby('Year')['Inflation, consumer prices (annual %)'].mean()
df_copy9['AVG Poverty Headcount'] = df.groupby('Year')['Poverty headcount ratio at $2.15 a day (2017 PPP) (% of population)'].mean()
df_copy9.drop('Inflation, consumer prices (annual %)', axis=1, inplace=True)
df_copy9
```

Out[27]:

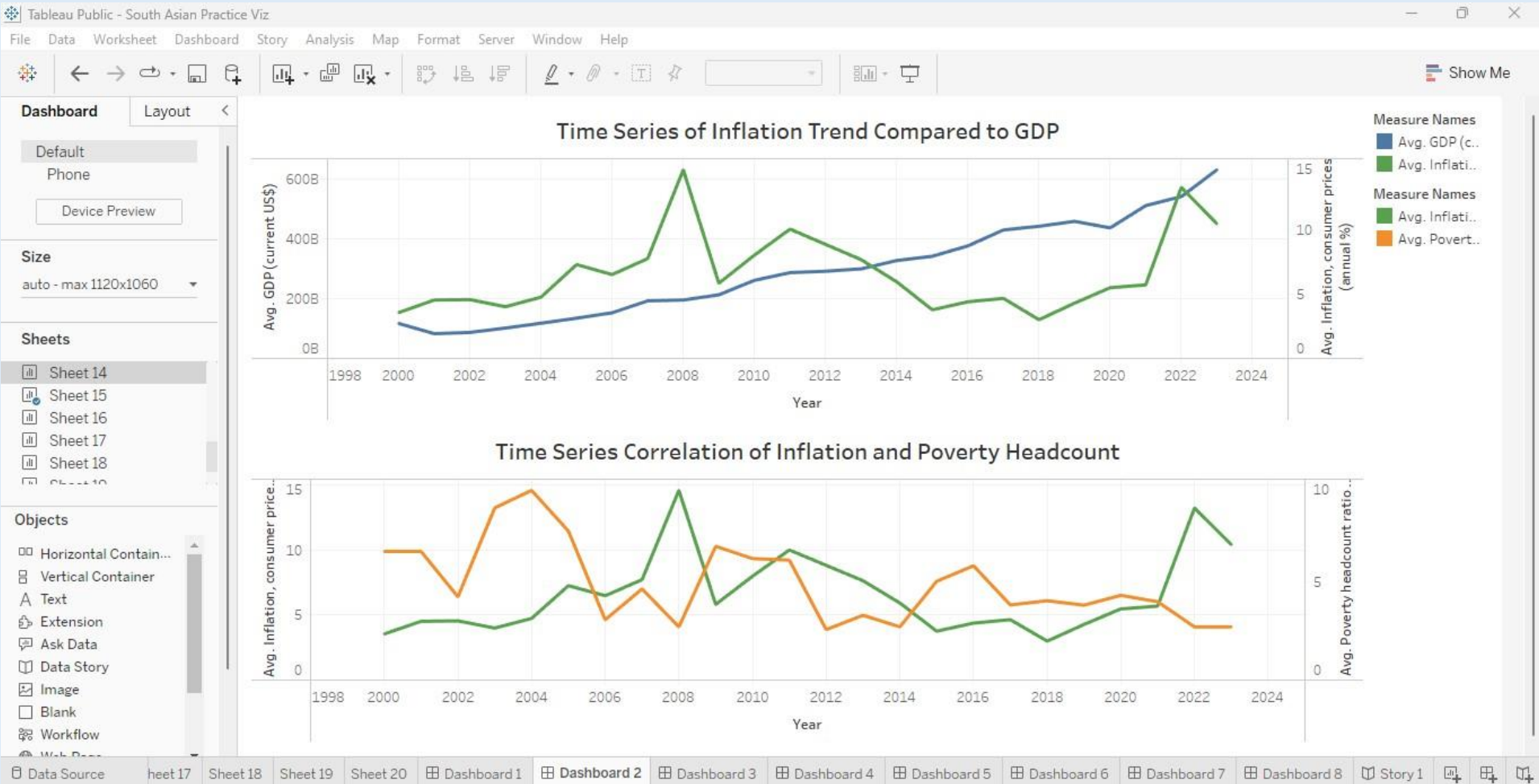
	AVG Inflation %	AVG Poverty Headcount
Year		
2000	3.55000	6.5250
2001	4.52250	6.5250
2002	4.55625	4.2250
2003	4.00375	8.7375
2004	4.74375	9.6250
2005	7.26750	7.5625
2006	6.49375	3.0625
2007	7.72750	4.6250
2008	14.60000	2.7000
2009	5.82250	6.7875
2010	7.99750	6.1625
2011	10.01250	6.0875
2012	8.82750	2.5625
2013	7.64625	3.2875
2014	5.93125	2.7000
2015	3.76375	5.0125
2016	4.38500	5.8000
2017	4.65125	3.8125
2018	2.99750	4.0250
2019	4.26125	3.8000
2020	5.47750	4.3000
2021	5.69250	3.9750
2022	13.24000	2.7000
2023	10.42750	2.7000

```
In [38]: df_copy3_1 = pd.DataFrame(df.groupby('Year')['Inflation, consumer prices (annual %)'].mean())
df_copy3_1['AVG Inflation %'] = pd.DataFrame(df.groupby('Year')['Inflation, consumer prices (annual %)'].mean())
df_copy3_1.drop('Inflation, consumer prices (annual %)', axis=1, inplace=True)
df_copy3_1['AVG GDP'] = df.groupby('Year')['GDP (current US$)'].mean()
df_copy3_1
```

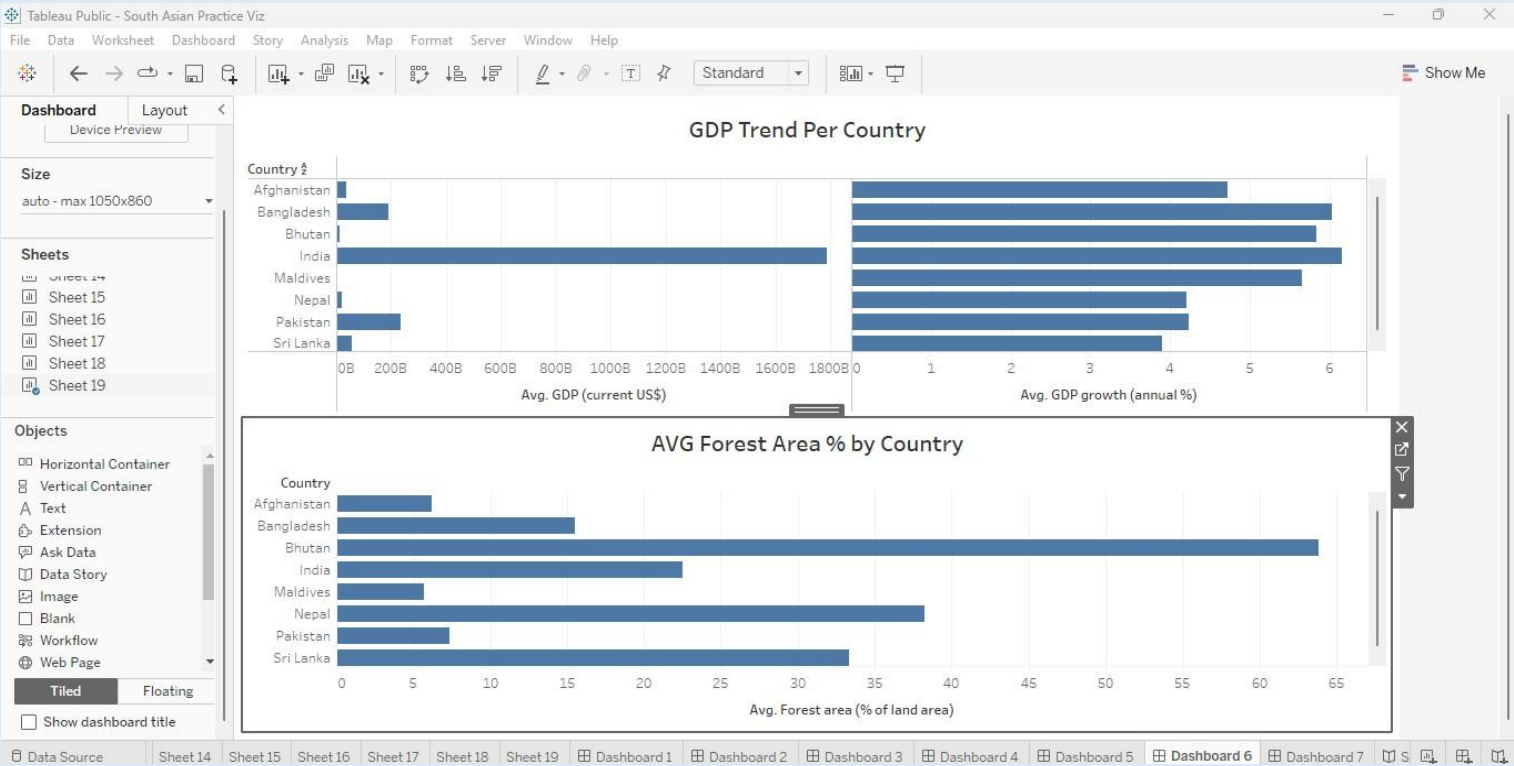
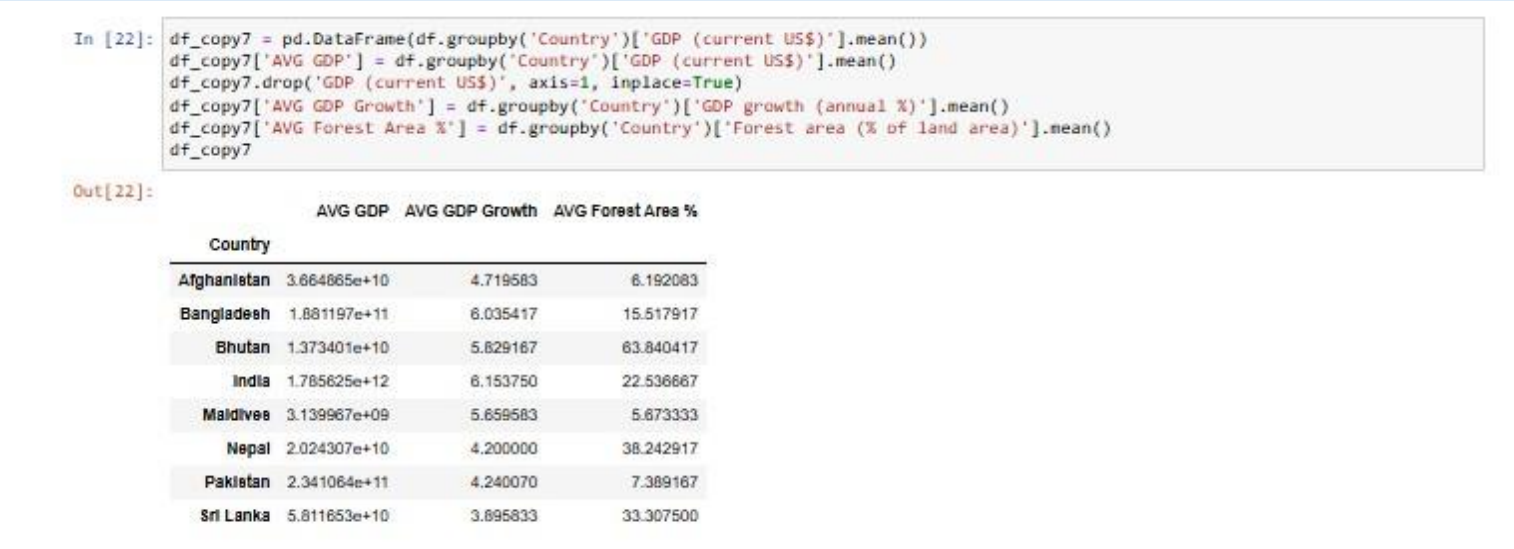
Out[38]:

	AVG Inflation %	AVG GDP
Year		
2000	3.55000	1.168418e+11
2001	4.52250	8.275918e+10
2002	4.55625	8.693961e+10
2003	4.00375	1.014495e+11
2004	4.74375	1.176540e+11
2005	7.26750	1.344049e+11
2006	6.49375	1.525761e+11
2007	7.72750	1.923949e+11
2008	14.60000	1.950743e+11
2009	5.82250	2.125018e+11
2010	7.99750	2.608446e+11
2011	10.01250	2.864806e+11
2012	8.82750	2.912354e+11
2013	7.64625	2.991897e+11
2014	5.93125	3.269390e+11
2015	3.76375	3.412401e+11
2016	4.38500	3.757982e+11
2017	4.65125	4.290636e+11
2018	2.99750	4.413308e+11
2019	4.26125	4.578078e+11
2020	5.47750	4.359830e+11
2021	5.69250	5.103530e+11
2022	13.24000	5.404874e+11
2023	10.42750	6.296506e+11

Time series analysis revealed the highest spikes in inflation occurred from 2007-2008 at 4.3% and 2021-2022 at 4.8%. The longest period of decreased inflation was from 2011-2015, showing a 3.8% decrease. In comparison, South Asia's GDP shows a consistent upward trajectory from 2000-2023, despite 2 periods of decline from 2000-2001 at 0.49% and 2019-2020 at 0.31%. A slight correlation found between inflation and poverty showed eras where poverty increased as inflation increased. The most significant increase in poverty was revealed to be 3.85% from 2002-2003.



Bar charts showed that India brings the highest GDP amount at a staggering 76% compared to the other countries. Bhutan and Maldives had by far the lowest GDP amounts, with Bhutan at 0.59% and Maldives at 0.13%. Despite this, Bhutan and Maldives showed significant GDP growth with Bhutan at 14.3% and Maldives at 13.9%. When it comes to the percentage forest area, the bar charts show a trend of GDP amounts being lowest in countries with the highest percentage of forest area; in this case Bhutan, Nepal, and Sri Lanka.

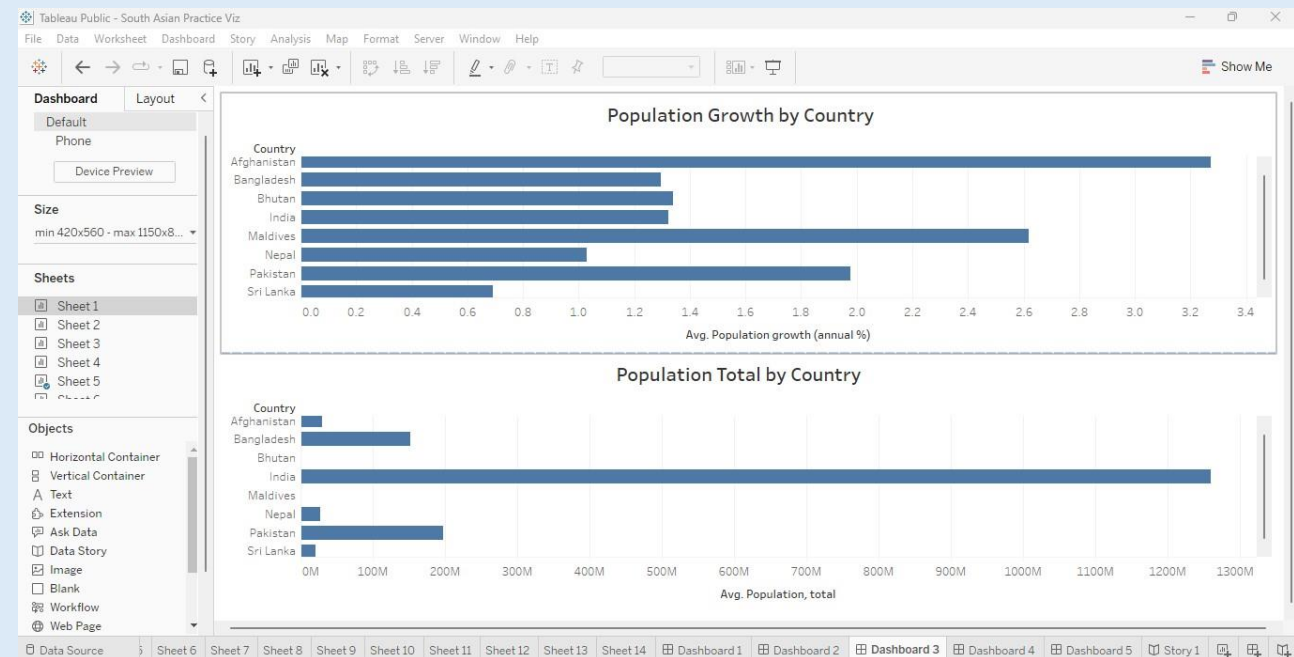



```
In [23]: df_copy5 = pd.DataFrame(df.groupby('Country')['Population, total'].mean()).sort_values(by='Population, total')
df_copy5['Population Growth %'] = df.groupby('Country')['Population growth (annual %)'].mean()
df_copy5
```

Out[23]:

Country	Population, total	Population Growth %
Maldives	3.946752e+05	2.616667
Bhutan	7.084806e+05	1.338750
Sri Lanka	2.070600e+07	0.691667
Nepal	2.743271e+07	1.026750
Afghanistan	3.037887e+07	3.271250
Bangladesh	1.516797e+08	1.295000
Pakistan	1.980191e+08	1.975833
India	1.259431e+09	1.321667

India dwarfs the other south Asian countries making up 74.6% of the population; while Afghanistan takes the lead in population growth by 24.15%.



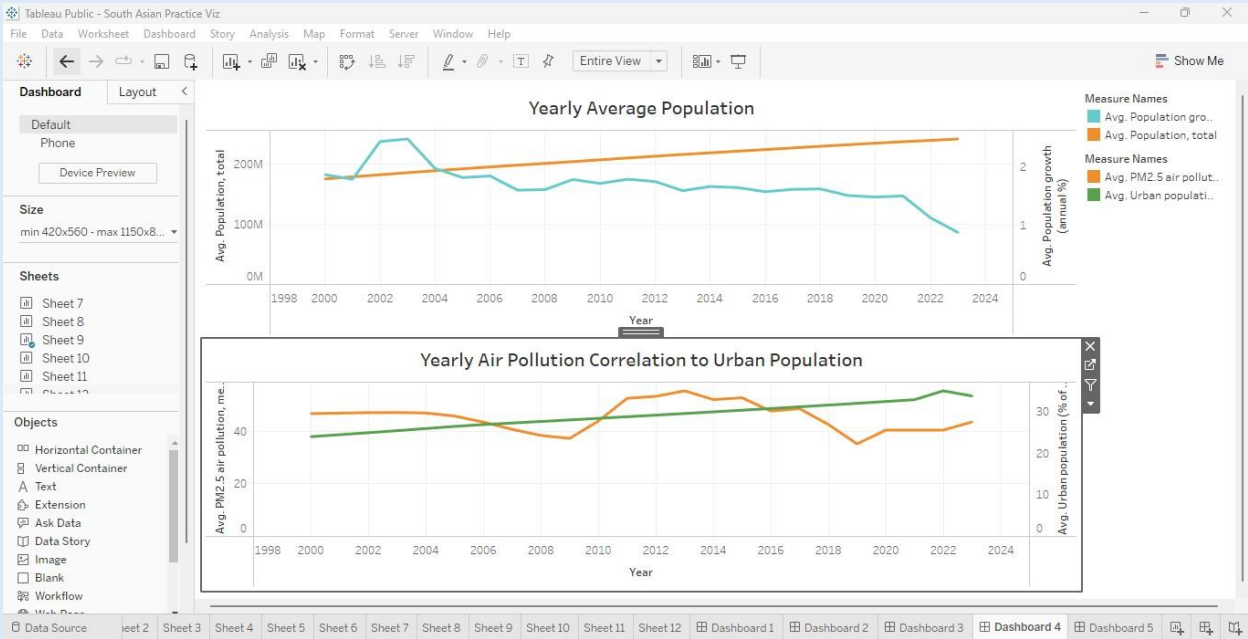
The population total maintained a consistent upward trajectory from 2000-2023. Meanwhile the population growth shows only 1 era of significant increase at 1.55% from 2000-2001, before continuing a trend of decline from 2003-2023. Although the population growth had increased numbers during it’s timeline, it was never enough to compensate for the dip in growth. The highest percent of decrease are shown 2003-2004 with 3.69% and 2021-2023 with 1.53% . The urban population maintained an upward trend from 2000-2022 before taking a 0.16% dip in 2023. The urban population doesn’t appear to have any impact on pollution based on the pattern of decreased pollution levels as the urban population has increased. For example, there was a 0.89% drop from 2004-2009 and an even greater decrease of 1.85% from 2013-2019.

In [24]:

```
df_copy6 = pd.DataFrame(df.groupby('Year')['Population, total'].mean())
df_copy6['Yearly Urban Population %'] = df.groupby('Year')['Urban population (% of total population)'].mean()
df_copy6['AVG Annual Pollution Exposure'] = df.groupby('Year')['PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)'].mean()
df_copy6
```

Out[24]:

Year	Population, total	Yearly Urban Population %	AVG Annual Pollution Exposure
2000	1.758682e+08	23.903750	46.72500
2001	1.792893e+08	24.378750	46.92500
2002	1.827588e+08	24.855000	47.09125
2003	1.861698e+08	25.342500	47.12875
2004	1.894629e+08	25.847500	46.95750
2005	1.926580e+08	26.362500	45.80000
2006	1.957368e+08	26.803750	43.46375
2007	1.986818e+08	27.171250	40.73875
2008	2.015886e+08	27.543750	38.41500
2009	2.045515e+08	27.915000	37.26125
2010	2.075683e+08	28.293750	43.89625
2011	2.106123e+08	28.677500	52.54625
2012	2.135883e+08	29.061250	53.29875
2013	2.164605e+08	29.450000	55.39375
2014	2.192538e+08	29.845000	52.07125
2015	2.219431e+08	30.243750	52.75125
2016	2.246341e+08	30.642500	47.68000
2017	2.273684e+08	31.052500	48.61125
2018	2.300668e+08	31.470000	42.60750
2019	2.326998e+08	31.891250	36.18750
2020	2.353165e+08	32.323750	40.46000
2021	2.377390e+08	32.758750	40.46000
2022	2.399185e+08	34.884286	40.46000
2023	2.423187e+08	33.653750	43.53625



```
In [26]: df_copy8 = pd.DataFrame(df.groupby('Country')['Urban population (% of total population)'].mean())
df_copy8['Urban Population %'] = df.groupby('Country')['Urban population (% of total population)'].mean()
df_copy8.drop('Urban population (% of total population)', axis=1, inplace=True)
df_copy8['AVG Access to Basic Sanitation Services'] = df.groupby('Country')['People using at least basic sanitation services (%)'].mean()
df_copy8['AVG Access to Basic Drinking Water Services'] = df.groupby('Country')['People using at least basic drinking water services (%)'].mean()
df_copy8['AVG Access to Electricity'] = df.groupby('Country')['Access to electricity (% of population)'].mean()
df_copy8
```

Out[26]:

	Urban Population %	AVG Access to Basic Sanitation Services	AVG Access to Basic Drinking Water Services	AVG Access to Electricity
Country				
Afghanistan	24.164583	38.192500	53.682500	57.762500
Bangladesh	31.742500	42.677083	96.026667	63.650000
Bhutan	35.682500	63.172500	92.431667	76.245833
India	31.653750	46.311667	86.513333	79.412500
Maldives	36.427500	87.283750	96.715833	94.870833
Nepal	17.280870	46.840000	85.811250	66.404167
Pakistan	35.355417	51.273333	88.731250	84.966667
Sri Lanka	18.416667	84.522083	85.885833	86.716667

```
In [28]: df_copy10 = pd.DataFrame(df.groupby('Country')['Urban population (% of total population)'].mean())
df_copy10['AVG Urban Population %'] = df.groupby('Country')['Urban population (% of total population)'].mean()
df_copy10['AVG School Enrollment %'] = df.groupby('Country')['School enrollment, primary (% gross)'].mean()
df_copy10['AVG Literacy Rate %'] = df.groupby('Country')['Literacy rate, adult total (% of people ages 15 and above)'].mean()
df_copy10.drop('Urban population (% of total population)', axis=1, inplace=True)
df_copy10
```

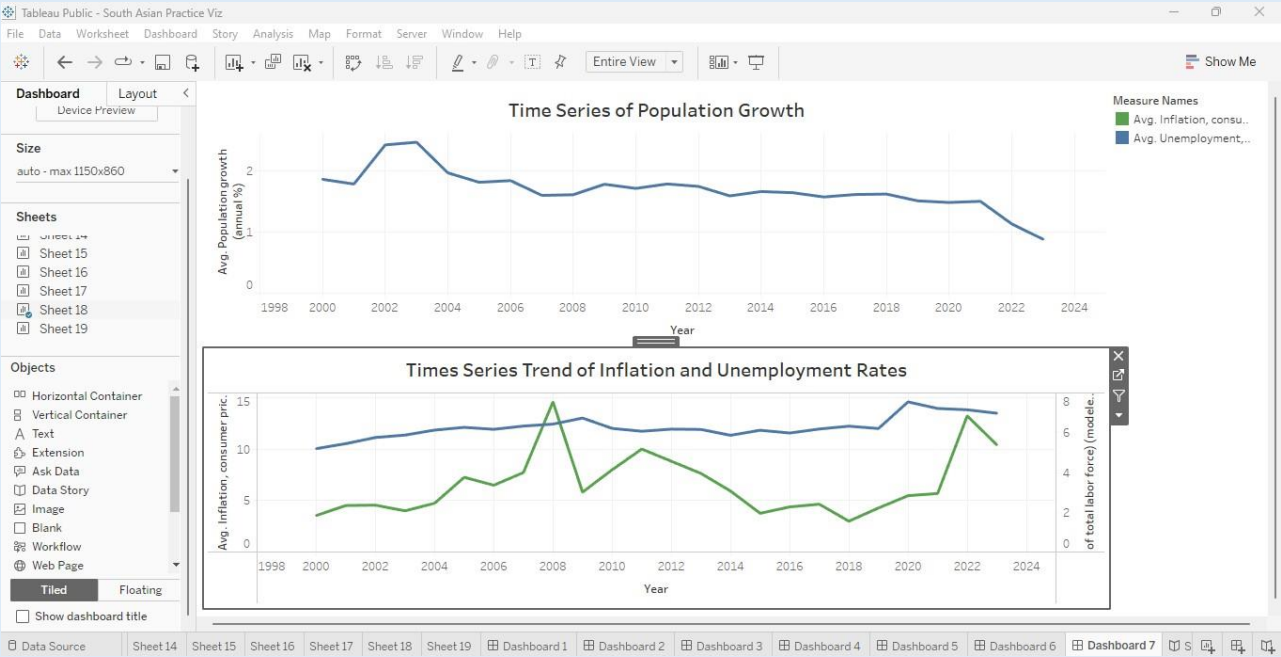
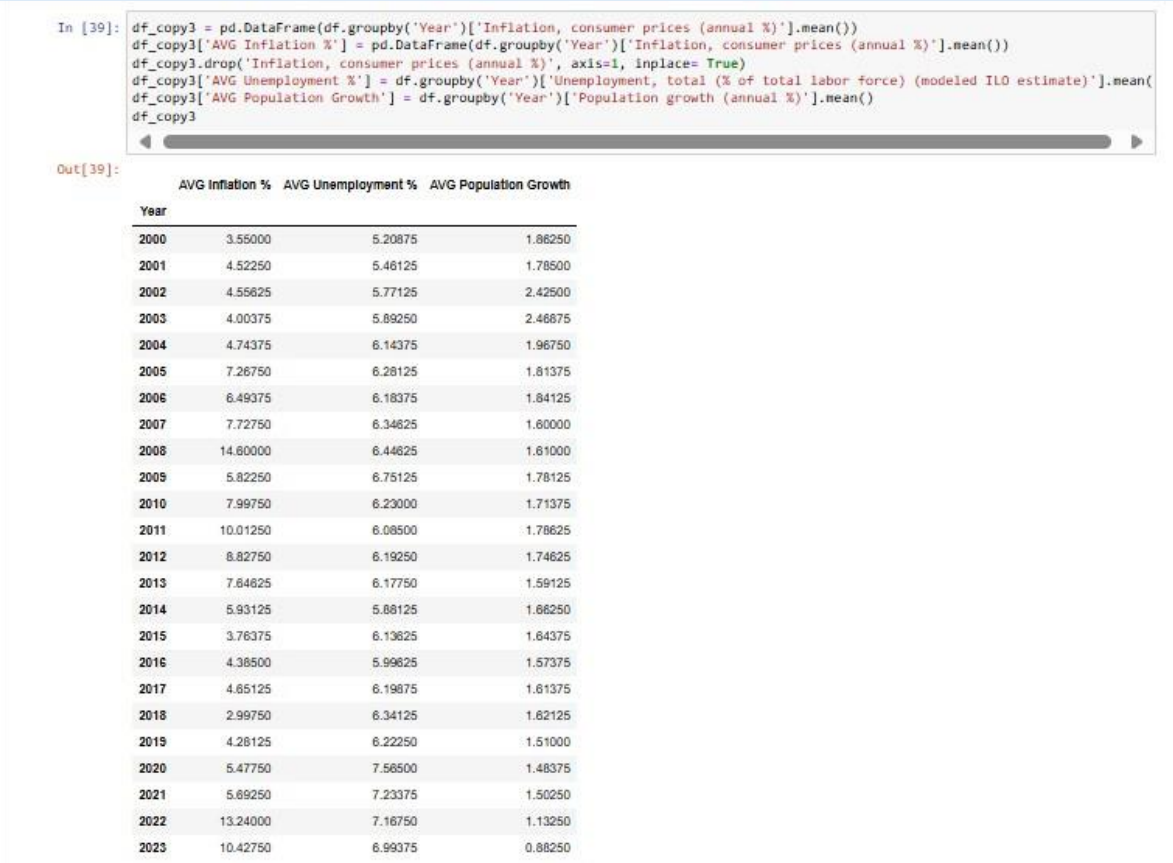
Out[28]:

	AVG Urban Population %	AVG School Enrollment %	AVG Literacy Rate %
Country			
Afghanistan	24.164583	93.620000	23.500000
Bangladesh	31.742500	97.197500	45.250000
Bhutan	35.682500	99.013333	28.625000
India	31.653750	105.837500	31.833333
Maldives	36.427500	105.010417	44.000000
Nepal	17.280870	128.047083	30.541667
Pakistan	35.355417	74.769583	40.125000
Sri Lanka	18.416667	96.929167	53.916667

Bar charts show that Maldives, Bhutan, and Sri Lanka are the most developed countries based on the abundance of essential resources available to their population; resources such as electricity, sanitation, and basic water services. In comparison, Afghanistan is the least developed as they have the smallest urban population, hold the smallest percentage of basic resources available, and have the lowest literacy rate. In contrast, Bangladesh, Maldives, and Sri Lanka have the highest literacy rates, most likely because of their larger urban population.



Time series analysis showed the unemployment rate increasing as inflation increased. The unemployment rate raised 0.97% from 2000-2009 before dropping 0.33% in 2010. From 2010-2019, the unemployment rate plateaued while the inflation % steadily dropped 3.8%. However, the unemployment rate increased 0.89% in 2019 before declining from 2020-2024.



Conclusion

Patterns detected in my visualizations reveal that Bangladesh, Bhutan, Maldives, and Sri Lanka are the most developed and advanced countries when it comes to resources, education, and life expectancy; unlike Afghanistan, who came last in these categories. Despite India making up 3/4 of the south Asian population and having the highest GDP, they are not as ahead as the more developed countries due to limited essential resources amongst their vast population.

Times series plots showed a correlation between GDP and inflation where inflation increased during periods when the GDP declines. With bar charts, I was able to discover high forest area percentages as a factor that negatively impacts GDP. The countries with the highest forest area percentage had the lowest GDP...with an exception of India.

After reviewing the trajectory of inflation, I was able to pinpoint patterns that showed unemployment and poverty rates following inflation's trend over the years, with an exception of the unemployment % hitting a plateau from 2010-2019 despite inflation percentage.