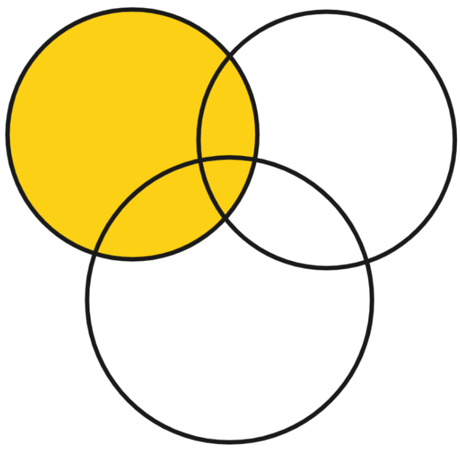
Analyzing Key Metrics of Nepal: A Data Analytics Approach



- MAHIM DHUNGEL

Table of Contents:

Introduction:

Nepal, a South Asian country nestled between the majestic Himalayan Mountains and the lush plains of the Terai, is a land of cultural diversity and natural beauty. As a developing country, Nepal has been the subject of numerous studies and research projects that aim to understand the social, economic, and environmental challenges facing the country. In this study, I collected data from four different online sources to examine key socio-economic indicators in Nepal.

All the sources of data come from the World Bank's Open Data and the worlddata info Platforms, which provides access to a wide range of socio-economic indicators for countries around the world. The socio-economic indicators I chose for analysis were:

1) Unemployment, total (% of total labour force)
   https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?locations=NP
2) GDP (current US$) - Nepal
   https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=NP
3) Final consumption expenditure (% of GDP) - Nepal
   https://data.worldbank.org/indicator/NE.CON.TOTL.ZS?locations=NP
4) Historical inflation rates
   https://www.worlddata.info/asia/nepal/inflation-rates.php

By gathering data from these four sources, the analysis aims to provide an overview of key socio-economic indicators in Nepal. Specifically, I will examine trends in the aforementioned indicators over time. The analysis will help shed light on the past and the current state of development in Nepal.

100 %

Results | Messages

	FACT_ID	CONSUMPTION_ID	GDP_ID	INFLATION_ID	UNEMP_ID	CONSUMPTION	INFLATION	Unemployment	Employment	GDP	YEAR_FIELD
16	fact1975	CONS1975	GDP1975	INF1975	UNEMP1975	89.9704308	7.585985214	0	100	1575789...	1975-01-01
17	fact1976	CONS1976	GDP1976	INF1976	UNEMP1976	88.2718167	-3.113235733	0	100	1452792...	1976-01-01
18	fact1977	CONS1977	GDP1977	INF1977	UNEMP1977	86.5046296	9.898852843	0	100	1382400...	1977-01-01
19	fact1978	CONS1978	GDP1978	INF1978	UNEMP1978	86.6207177	7.346091251	0	100	1604162...	1978-01-01
20	fact1979	CONS1979	GDP1979	INF1979	UNEMP1979	88.3637192	3.565453785	0	100	1851250...	1979-01-01
21	fact1980	CONS1980	GDP1980	INF1980	UNEMP1980	88.9041146	14.68450278	0	100	1945916...	1980-01-01
22	fact1981	CONS1981	GDP1981	INF1981	UNEMP1981	89.10902	11.14485673	0	100	2275583...	1981-01-01
23	fact1982	CONS1982	GDP1982	INF1982	UNEMP1982	90.0671222	11.69854659	0	100	2395429...	1982-01-01
24	fact1983	CONS1983	GDP1983	INF1983	UNEMP1983	91.451423	12.37723808	0	100	2447174...	1983-01-01
25	fact1984	CONS1984	GDP1984	INF1984	UNEMP1984	90.1345519	2.845784957	0	100	2581207...	1984-01-01
26	fact1985	CONS1985	GDP1985	INF1985	UNEMP1985	85.914224	8.052641266	0	100	2619913...	1985-01-01
27	fact1986	CONS1986	GDP1986	INF1986	UNEMP1986	88.3762831	18.99894955	0	100	2850784...	1986-01-01
28	fact1987	CONS1987	GDP1987	INF1987	UNEMP1987	87.9045628	10.75032804	0	100	2957255...	1987-01-01
29	fact1988	CONS1988	GDP1988	INF1988	UNEMP1988	88.6425506	8.983003382	0	100	3487009...	1988-01-01
30	fact1989	CONS1989	GDP1989	INF1989	UNEMP1989	89.7088127	8.846886912	0	100	3525228...	1989-01-01

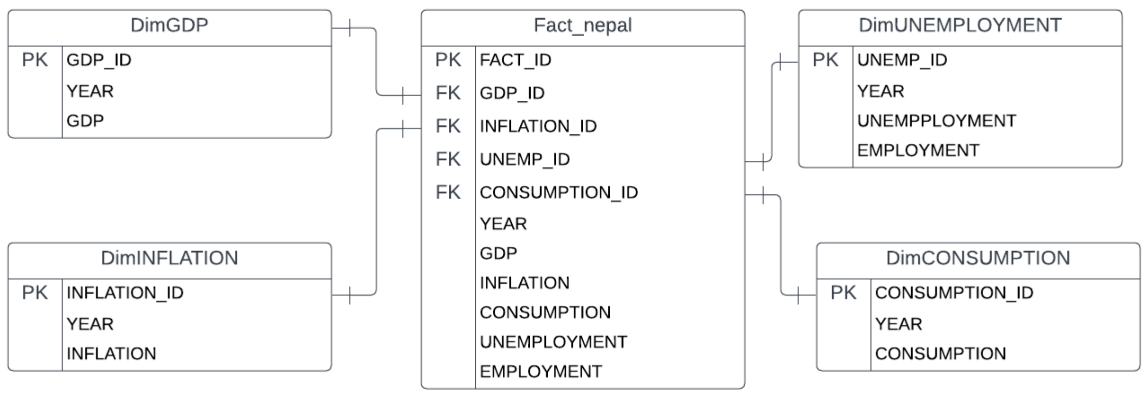
Query executed successfully. | RAS-RDSH-04\SQLEXPRESS (15.... | CLARKU\MDhungel (52) | mdhungel db | 00:00:00 | 62 row

Overview of the schema

The logical schema is an essential component of any database that defines the structure and organisation of data within the system. In this schema, there are four dimension tables, named Consumption, Inflation, GDP, and Unemployment, which provide context and meaning to the data. Each dimension table contains attributes related to its respective subject area.

Additionally, there is one fact table that consolidates all the metrics from the dimension tables in one place. This fact table contains the measures or metrics that will be analysed, such as the total consumption, inflation rate, GDP growth, and unemployment rate among others.

By combining the metrics from the fact table with the attributes from the dimension tables, users can gain valuable insights into the relationships between various factors and make informed decisions based on the data.



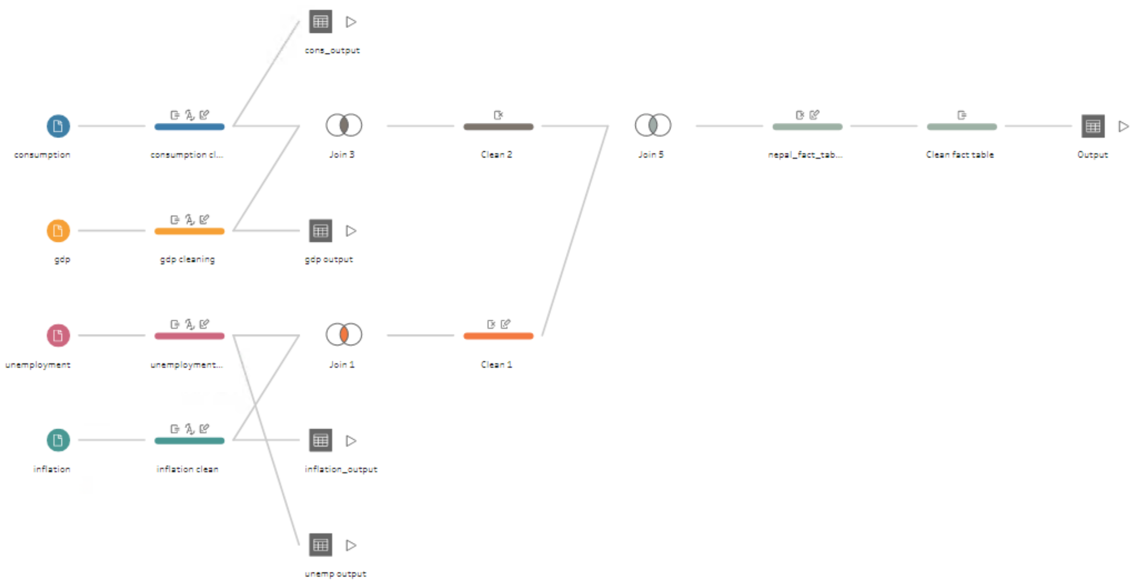
Data cleaning and pipelining in Tableau:

Data cleaning and pipelining are critical processes in Tableau that help to ensure accurate and reliable data analysis. Data cleaning involves the identification and removal of errors, inconsistencies, and outliers in the data, while pipelining refers to the process of transforming, filtering, and aggregating data to prepare it for analysis.

In Tableau, data cleaning and pipelining can be accomplished through various techniques and tools. For example, Tableau's data preparation tools, such as the data interpreter, can automatically detect and correct common data issues, such as misspellings and formatting errors. Additionally, Tableau's data cleansing tools, such as the cluster analysis and outlier detection, can be used to identify and remove data anomalies that can skew analysis results.

Once the data has been cleaned, pipelining can be used to transform, filter, and aggregate data to prepare it for analysis. Tableau's data blending and joining features can be used to combine data from multiple sources and create meaningful relationships between different datasets. Data blending allows data to be combined on the fly, while data joining allows data to be joined based on common fields or keys.

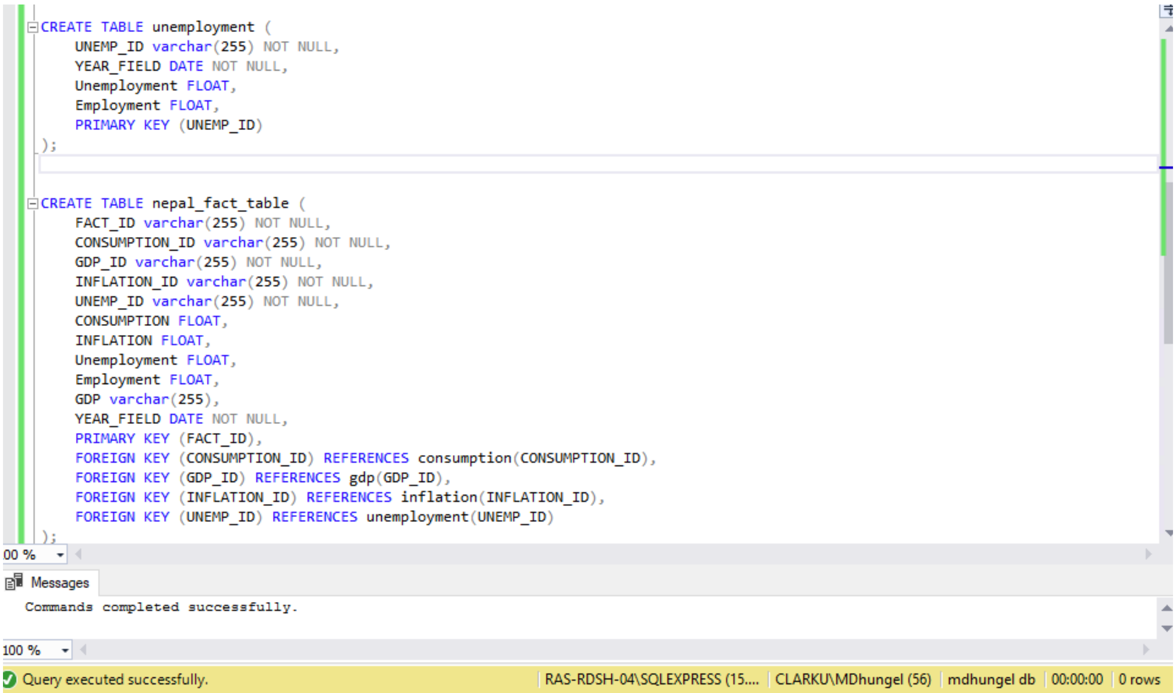
Overall, data cleaning and pipelining are essential components of the data analysis process in Tableau. By ensuring that data is accurate, reliable, and properly prepared, data cleaning and pipelining can help analysts to derive meaningful insights and make informed decisions based on the data.



(TABLEAU FLOW FILE ATTACHED IN GITHUB)

Data management in MICROSOFT SQL SERVER

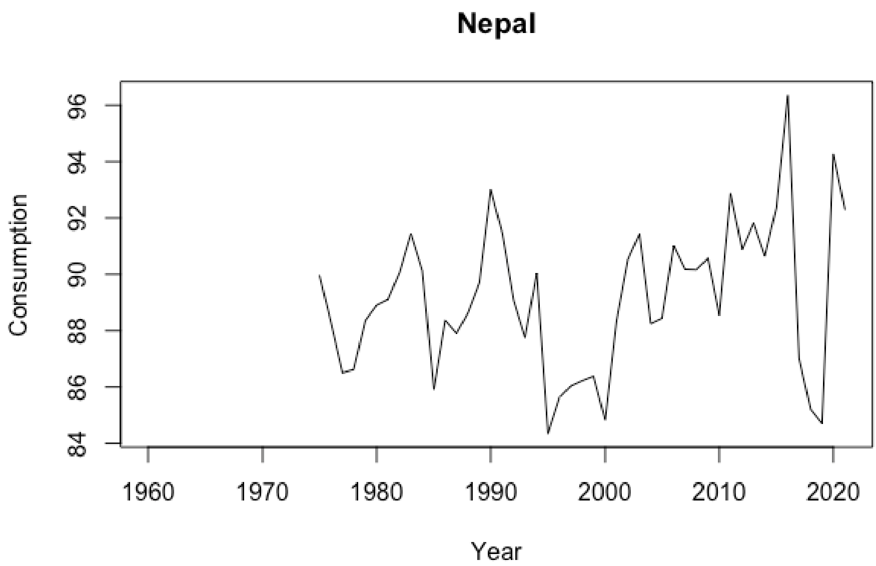
Microsoft SQL Server is a robust data management system that provides tools and features for organising, storing, and retrieving data in a secure and efficient manner. SQL Server supports data modelling, data integration, data warehousing, and data analysis, as well as features for managing data integrity, consolidation, and security.

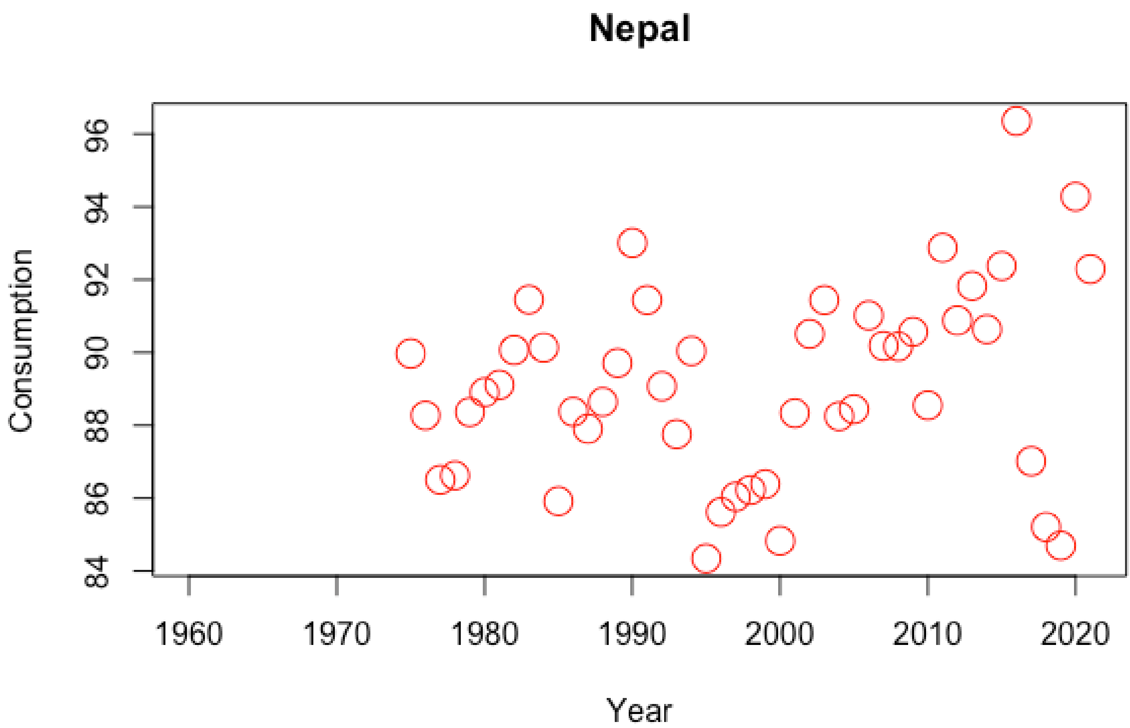


(TABLE CREATION IN THE DATABASE)
(Table creation code included in Github)

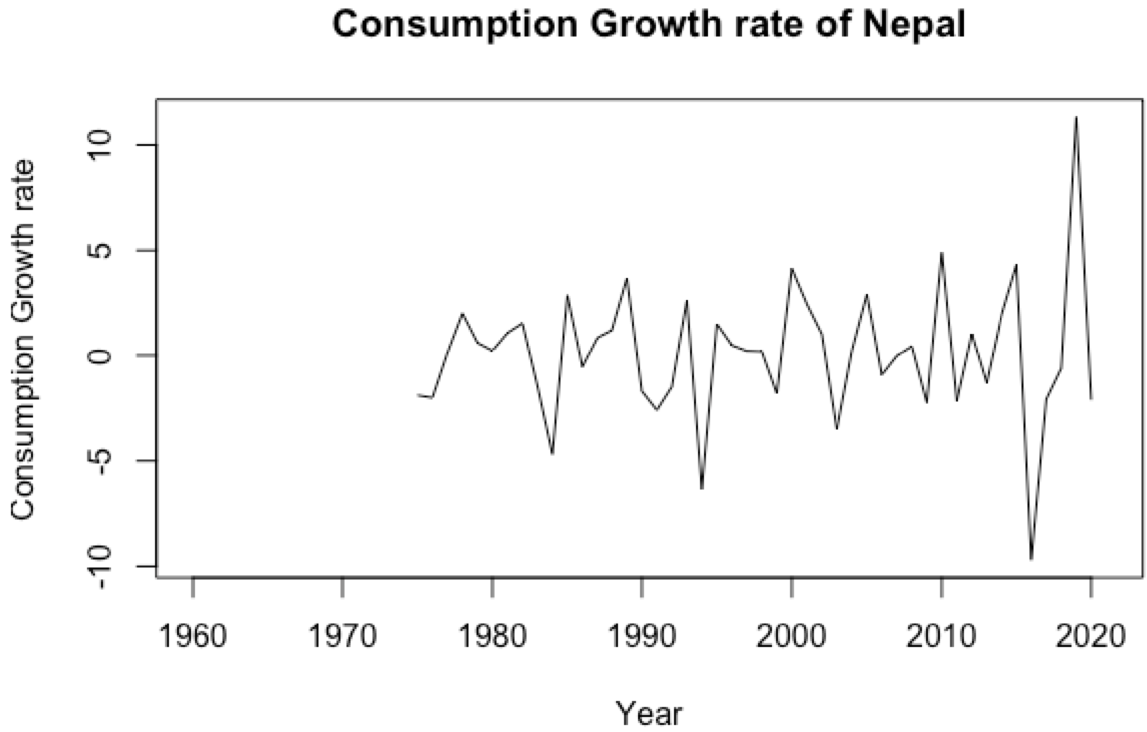
Key Performance Indicators:

1) **How has the consumption fared in Nepal over the years?**

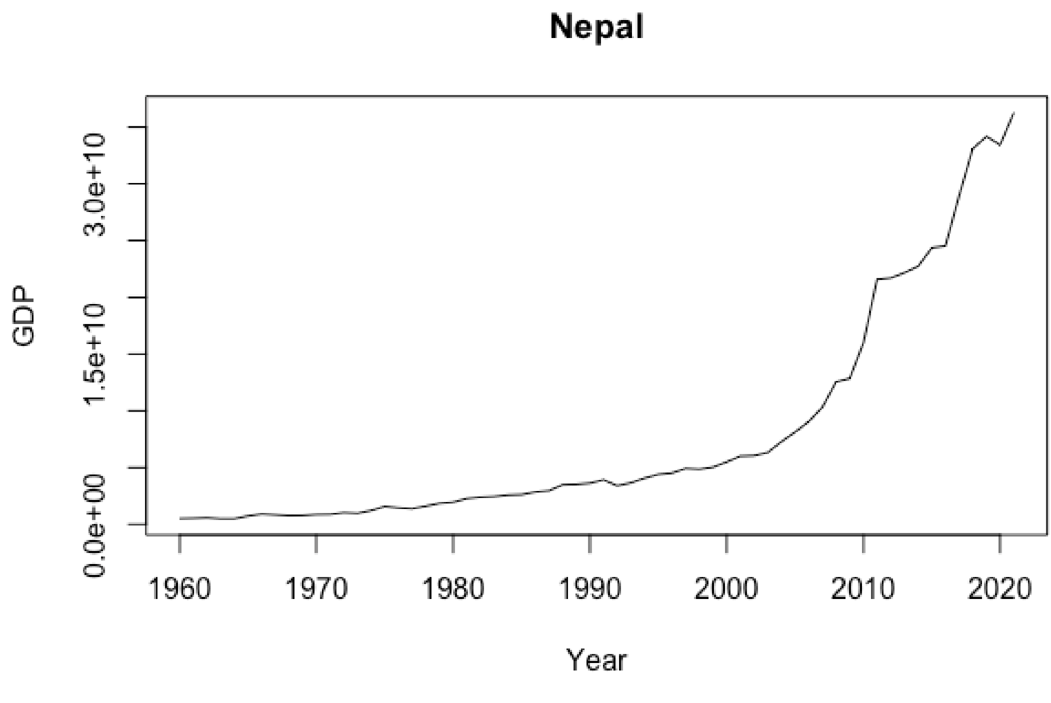




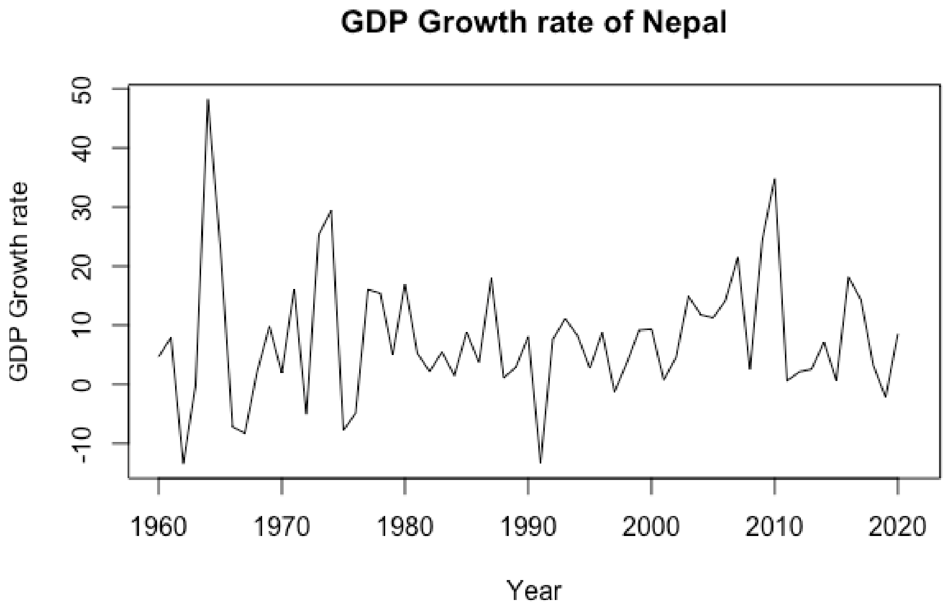
2) **What is the consumption growth rate of Nepal?**



3) **What kind of growth do we see in the GDP of Nepal? Is it steady or being a developing nation, is it exponential?**



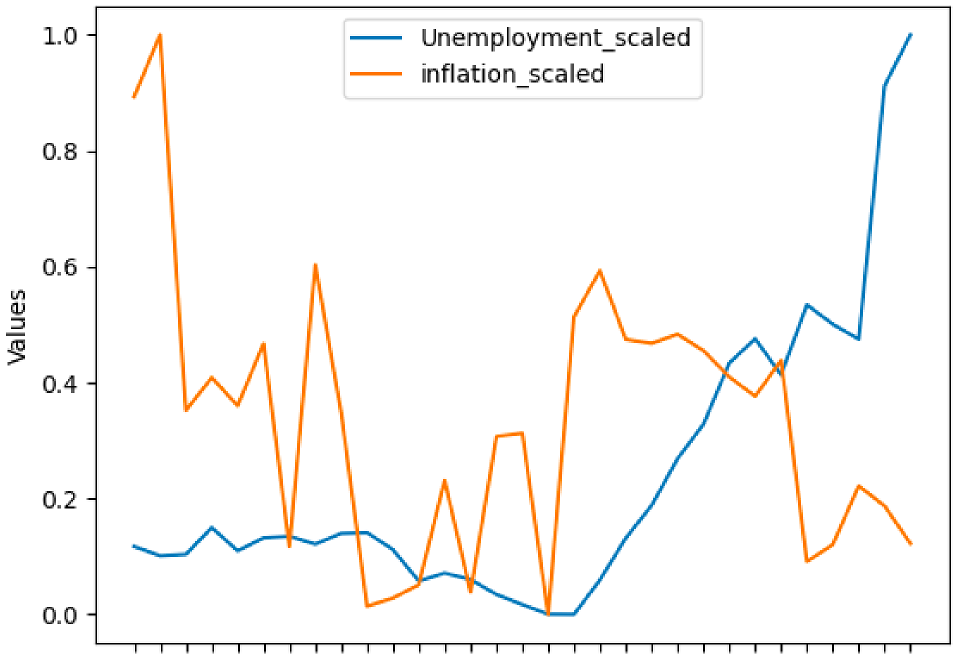
4) What is the GDP growth rate of Nepal? Has it been increasing or decreasing?



5) **Phillips curve in Nepal:** According to this theory, when GDP is high, there is low unemployment, which leads to higher wages and increased demand for goods and services. This increased demand, in turn, can lead to higher prices and inflation. Conversely, when GDP is low and unemployment is high, there is less demand for goods and services, which can lead to lower prices and deflation.

Since our datasets are dealing with vastly different value ranges, it is often necessary to scale them to enable meaningful comparison and analysis. One way to do this is to normalise the data using a common scaling technique such as **Min-Max normalisation.**

We also need to ignore the first 31 values of unemployment because they do not exist.



Checking the Phillips curve. The data don't seem to agree with the theory. There is no correlation.

Summary statistics between GDP and Consumption in R:

```
#summary statistics of gdp
summary(gdp)
sd(gdp)
range(gdp)
skewness(gdp)
kurtosis(gdp)

# slicing the first 16 rows because the data is not av
consumption <- data %>%
  slice(16:n()) %>%
  select(CONSUMPTION) %>%
  unlist()
#summary statistics of consumption
summary(consumption)
sd(consumption)
range(consumption)
skewness(consumption)
kurtosis(consumption)
```

```
> summary(gdp)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
4.961e+08 1.400e+09 3.576e+09 7.937e+09 8.815e+09 3.629e+10
> sd(gdp)
[1] 9985856837
> range(gdp)
[1]   496098775 36289000000
> skewness(gdp)
[1] 1.536747
> kurtosis(gdp)
[1] 1.074782
>
```

```
> summary(consumption)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  84.34   87.38   89.07   89.16   90.76   96.36
> sd(consumption)
[1] 2.639537
> range(consumption)
[1] 84.34470 96.35714
> skewness(consumption)
[1] 0.2367494
> kurtosis(consumption)
[1] -0.2384927
>
```

Running a simple OLS linear regression in R and interpreting its coefficient estimates.

```r
#O.LS linear regression
lin_reg <- lm(consumption ~ gdp, data=data)
```

```
lin_reg          List of  13
```

```
Call:
lm(formula = consumption ~ gdp, data = data)

Coefficients:
(Intercept)          gdp
  8.844e+01    7.093e-11
```

Interpreting goodness-of-fit metrics for OLS linear regression.

```r
# run this to calculate b0hat because the dataframes are not of the same size
# for ques no 1.f
gdp <- data %>%
  slice(16:n()) %>%
  select(GDP) %>%
  unlist()
consumption <- data %>%
  slice(16:n()) %>%
  select(CONSUMPTION) %>%
  unlist()
# Interpret goodness-of-fit metrics for OLS linear regression.
( b0hat <- cov(gdp,consumption)/var(gdp) )
( b1hat <- mean(consumption) - b0hat*mean(gdp) )
```

```
> ( b0hat <- cov(gdp,consumption)/var(gdp) )
[1] 7.092657e-11
> ( b1hat <- mean(consumption) - b0hat*mean(gdp) )
[1] 88.43656
>
```

Relating R2 with correlation coefficient.

```
# Relate R2 with correlation coefficient.
y_hat<-fitted(lin_reg)
u<-resid(lin_reg)

var(y_hat)/var(consumption)
1-var(u)/var(consumption)
cor(consumption,y_hat)^2
```

```
> var(y_hat)/var(consumption)
[1] 0.07953421
> 1-var(u)/var(consumption)
[1] 0.07953421
> cor(consumption,y_hat)^2
[1] 0.07953421
>
```

We can see that 7% of the variation in consumption of Nepal is explained by the country's GDP.

Monte-Carlo Simulation:

R ▾ | Global Environment ▾

Data	
data	64 obs. of 1 variable
Values	
b0	25
b0hat	num [1:100] 21.7 25.4 28.9 25.4 28.4 ...
b1	0.5
b1hat	num [1:100] 0.546 0.477 0.425 0.479 0.464 ...
bhat	Named num [1:2] 24.181 0.504
j	100L
n	64
r	100
se	3
u	num [1:64] -4.06 8.12 1.26 6.47 -9.76 ...
x	int [1:64] 37 22 16 65 76 26 45 29 11 55 ...
y	num [1:64] 39.4 44.1 34.3 64 53.2 ...

We can see that by simulating using Monte Carlo simulation, the parameters from the sample regression function are close to that of the true population parameters.