UNIVERSITY OF CALCUTTA

LADY BRABOURNE COLLEGE

B.Sc. SEMESTER–VI(HONOURS)EXAMINATION-2023(UNDER CBCS)

SUBJECT: STATISTICS(STSA)

Paper: DSE B2- PROJECT WORK

PROJECT TITLE: ***STATISTICAL ANALYSIS OF THE IMPACT OF COFFEE PLANTATIONS ON THE GDP OF SOME PROMINENT COFFEE EXPORTING COUNTRIES FOR THE PERIOD 2000-2019.***

SUBMITTED BY

CU ROLL NUMBER: 203031-11-0077

CU REGISTRATION NUMBER: 031-1211-0413-20

SUPERVISOR'S CERTIFICATE

This is to certify that **Ms DEBOMITA PAUL(University Registration No.: 031-1211-0413-20, University Roll No.: 203031-11-0077)** a student of B.Sc. Honours in Statistics of Lady Brabourne College under the University of Calcutta has worked under my supervision and guidance for her Project Work and prepared a Project Report with the title **STATISTICAL ANALYSIS OF THE IMPACT OF COFFEE PLANTATIONS ON THE GDP OF SOME PROMINENT COFFEE EXPORTING COUNTRIES FOR THE PERIOD 2000-2019** which she is submitting, is her genuine and original work to the best of my knowledge.

Moutushi Chatterjee 24/7/23

Signature:

Name: Dr. Moutushi Chatterjee

Designation: Assistant Professor

Name of the College: SQC and OR unit,

Indian Statistical Institute, Bangalore.

Place: Indian Statistical Institute, Bangalore

Date: 24th July, 2023

STUDENT'S DECLARATION

I hereby declare that the Project Work with the title **STATISTICAL ANALYSIS OF THE IMPACT OF COFFEE PLANTATIONS ON THE GDP OF SOME PROMINENT COFFEE EXPORTING COUNTRIES FOR THE PERIOD 2000-2019** submitted by me for the partial fulfilment of the degree of B.Sc. Honours in Statistics under the University of Calcutta is my original work and has not been submitted earlier to any other University or Institution for the fulfilment of the requirement for any course of study. I also declare that no chapter of this manuscript in whole or in part has been incorporated in this report from any earlier work done by others or by me. However, extracts of any literature which has been used for this report has been duly acknowledged providing details of such literature in the references.

Signature:

Name: DEBOMITA PAUL

Address: 61/5, ROBERTSON ROAD,
NAIHATI, NORTH 24 PARGANAS – 743165,
WEST BENGAL, INDIA

University Registration No.: 031-1211-0413-20

University Roll No.: 203031-11-0077

Place: Lady Brabourne College.

Date: 31st July, 2023.

ACKNOWLEDGEMENT

I would like to express my gratitude to several individuals and organizations for supporting me throughout my final semester project. First, I wish to express my sincere gratitude to my supervisor, Professor Dr. Moutushi Chatterjee for her exemplary guidance, enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that have helped me tremendously at all times in my research and doing this project. I am also extremely grateful to all my other professors, Dr. Bratati Chakraborty and Prof. Ipshita Samanta to help me finish this project in the best possible way. Their immense knowledge, profound experience and professional expertise have enabled me to complete this research successfully. I also wish to express my sincere thanks to the University of Calcutta and Lady Brabourne College for giving me the opportunity for doing this project. I would also like to express my gratitude for our principal ma'am for providing the means and support to complete the project. Thanks for all the encouragement!

Last but not the least; I want to thank my family and my friends for their constant encouragement and for exchanging interesting ideas throughout the course of this project.

TABLE OF CONTENT

ABSTRACT

This project presents four datasets for Brazil, Vietnam, Ethiopia and India. Each dataset describes the amount of coffee produced, exported and consumed domestically, gross domestic product (GDP) and population from 2000-2019. Each dataset contains five variables and twenty observations. The variables are coffee production, coffee consumption, coffee export, GDP (in billions of US$) and population. Furthermore, all the variables contain numerical data.

The primary goal for this project is to use the available variables to predict the GDP of each country. At first, exploratory data analysis is carried out. Since, the data sets include information regarding coffee production, export and consumed domestically in the years 2000-2019, this makes the data suitable for forecasting and time series analysis.

Furthermore, one simple linear regression model to predict coffee export based on coffee production is constructed for each country. Then, multiple linear regression models are constructed for each country to predict the GDP of the respective country.

Lastly, Principal Component Analysis is carried out to affectively reduce the dimension of the dataset while maintaining a significant amount of the total variability and getting rid of the problem of multicollinearity.

INTRODUCTION

Coffee, being the second-largest traded commodity in the world, has successfully converted approximately 30–40% of the world's population into coffee aficionados. On a global scale, coffee has an ever-growing interlinkage with a country's economy. The majority of the countries that produce coffee have greatly benefited from this socioeconomic crop, which has had a significant influence on their economies.

Gross Domestic Product (GDP) is the total value of all the finished goods and services produced in a country in a year. An increasing value of GDP indicates economic growth in terms of an increase in the quantity of different goods and services. For this project, my objective was to apply the techniques I have learnt throughout my bachelor's degree to predict coffee production, coffee consumption, and coffee export based on the previous year's data and to predict GDP based on population, coffee consumption and coffee export.

The top 10 coffee-producing nations are: Brazil, Vietnam, Columbia, Indonesia, Ethiopia, Honduras, India, Uganda, Mexico and Guatemala. I have included four of the top 10 coffee-producing nations in my project: Brazil, Vietnam, Ethiopia, and India. Brazil, a country in South America, is the largest coffee producer in the world and therefore plays a significant role in this project. Vietnam, a country in south-east Asia, is the second largest coffee producer in the world. The third country is Ethiopia, which is in Africa. Ethiopia has been chosen over Columbia and Indonesia to avoid narrowing down this project to certain continents. Lastly, India has been included as it is my native country. Moreover, India is the seventh largest coffee producer in the world and therefore has global significance with respect to coffee.

Here, is a description of the variables in the datasets:

t: year

prd_b: coffee production of Brazil

dc_b: domestic coffee consumption of Brazil

exp_b: coffee export of Brazil

GDP_b: GDP of Brazil

ppl_b: population of Brazil

prd_e: coffee production of Ethiopia

dc_e: domestic coffee consumption of Ethiopia

exp_e: coffee export of Ethiopia
GDP_e: GDP of Ethiopia

ppl_e: population of Ethiopia

prd_v: coffee production of Vietnam

dc_v: domestic coffee consumption of Vietnam

exp_v: coffee export of Vietnam

GDP_v: GDP of Vietnam

ppl_v: population of Vietnam

prd_i: coffee production of India

dc_i: domestic coffee consumption of India

exp_i: coffee export of India

GDP_i: GDP of India

ppl_i: population of India

METHODOLOGY

LINE DIAGRAM: In a simple line diagram, we plot each pair of values in the xy plane. The plotted points are then joined successively by line segments and the resulting chart is known as a line diagram.

MULTIPLE LINE DIAGRAM: In a multiple line graph, more than one dependent variable is charted on the graph and compared over a single independent variable (often time). Different dependent variables are often given different coloured lines to distinguish between each data set.

MULTIPLE BAR DIAGRAM: A multiple bar diagram is a bar chart in which multiple data sets are represented by drawing the bars side by side in a cluster.

COMPONENT BAR DIAGRAM: A component bar chart is used to represent data in which the total magnitude is divided into different components.

SCATTER PLOT: Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.

REGRESSION PLOTS: Regression plots as the name suggests creates a regression line between 2 parameters and helps to visualize their linear relationships.

TREND: Trend is a pattern in data that shows the movement of a series to relatively higher or lower values over a long period of time. In other words, a trend is observed when there is an increasing or decreasing slope in the time series. The use of a trend line can help predict future or unknown values.

Formula: $Y = a + b*t$,

SIMPLE LINEAR REGRESSION: In statistics, simple linear regression is a linear regression model with a single explanatory variable. That is, it concerns two-dimensional sample points with one independent variable and one dependent variable (conventionally, the x and y coordinates in a Cartesian coordinate system) and finds a linear function (a non-vertical straight line) that, as accurately as possible, predicts the dependent variable values as a function of the independent variable. The adjective simple refers to the fact that the outcome variable is related to a single predictor.

Formula: $Y=\beta_0 + \beta_1 * X_1 + \epsilon$

MULTIPLE LINEAR REGRESSION: Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple regressions are an extension of linear (OLS) regression that uses just one explanatory variable.

Formula: $Y=\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots + \beta_p * X_p$

POLYNOMIAL REGRESSION: Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x.
The appropriate model is $f(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$ where c_i are the parameters of the model for all i =1(1)n.

MULTICOLINEARITY: In statistics, multicollinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.

PRINCIPAL COMPONENT ANALYSIS: Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Let c be a k by one column vector. Usually we normalize c so that c'c = 1. The matrix product Xc is a linear combination of columns of X.

$$Xc = (x1 \quad x2 \quad \cdots \quad x3) \begin{pmatrix} c1 \\ c2 \\ \vdots \\ ck \end{pmatrix} = c_1x_1 + c_2x_2 + \ldots c_kx_k$$

where x_i is the i-th column of X, and c_i is the i-th entity of c. Note Xc is m by one.

Intuitively we hope Xc preserves the information in X to the largest extent possible. Thus the goal of PCA is to find an optimal c that maximizes the variation of Xc :

$$\max (Xc)'(Xc) = \max c' X'Xc \text{ (subject to } c'c = 1) \qquad (1)$$

Note that the k by k square matrix X'X is symmetric. As a result, we can apply a special spectral decomposition

$$X'X = V\Lambda V^{-1} = V\Lambda V' \qquad (2)$$

where Λ is the diagonal matrix of eigenvalues, and the columns of V are normalized eigen-vectors (i.e., $v'_i v_i = 1, \forall i = 1, \ldots, k$). We can show $V^{-1} = V'$ because the eigen-vectors of a symmetric matrix are orthogonal $v'_i v_j = 0, (\forall i \neq j)$. . It follows that

$$c'X' Xc = c'V\Lambda V'c = (c'v_1, \ldots c'v_k) \begin{pmatrix} \lambda 1 & \cdots & 0 \\ \vdots & \lambda i & \vdots \\ 0 & \cdots & \lambda k \end{pmatrix} (v'_1 c \ldots v'_k c) = \lambda_1 c'v_1 v'_1 c + \ldots + \lambda_k c'v_k v'_k c \qquad (3)$$

Suppose the eigenvalues are sorted in descending order:

$$\lambda_1 \geq \lambda_2 \ldots \geq \lambda_k$$

Because each $c'v_i v'_i c = (c'v_i)^2$ is a square term and non-negative, it follows that

$$\lambda_1 c'v_1 v'_1 c + \ldots + \lambda_k c'v_k v'_k c \leq \lambda_1 (c'v_1 v'_1 c + \ldots + c'v_k v'_k c') = \lambda_1 c'VV'c = \lambda_1 c'VV^{-1}c = \lambda_1 c'c = \lambda_1 \quad (4)$$

The equality in (4) holds only when $c = v_1$.

To summarize, the first principal component is given by Xv_1, where the weighting vector v_1 is the eigenvector belonging to the largest eigenvalue λ_1 of X'X. The variation of the first principal component is λ_1.

SCREE PLOT: The scree plot plots the variances of the PCs along the new feature axes. To determine the appropriate number of components, we look for a bend in the scree plot.

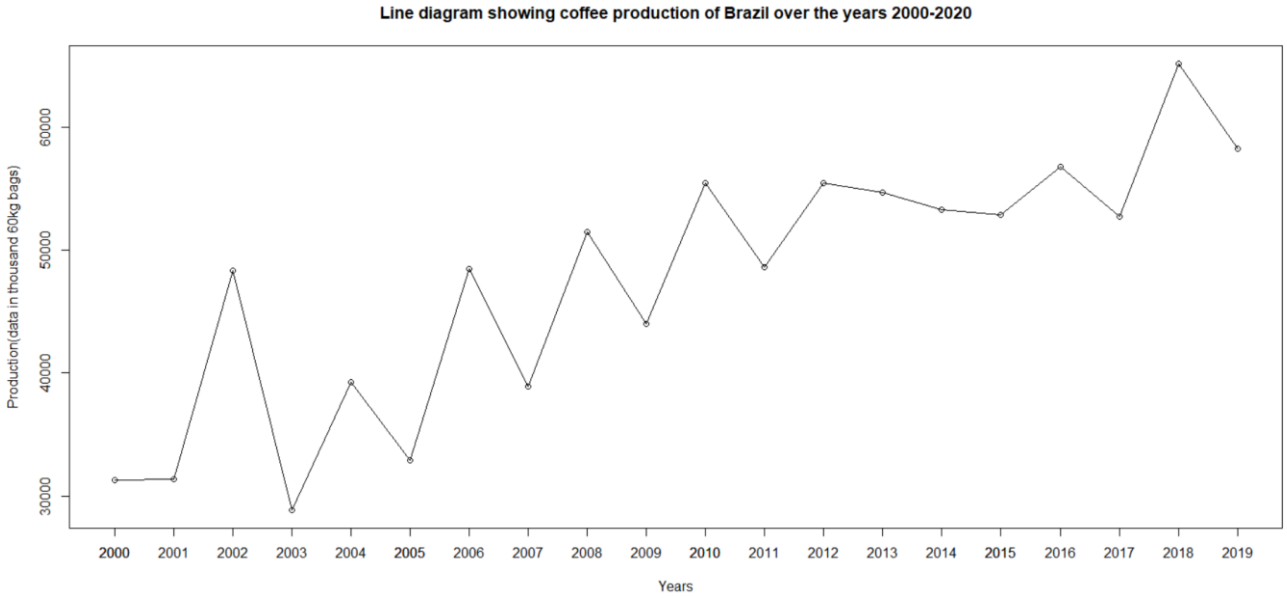
STATISTICAL ANALYSIS

In this study, I have taken a dataset that documents the amount of coffee produced, consumed and exported(all data in thousand 60kg bags), population and GDP of Brazil, Vietnam, Ethiopia and India for the period 2000-2019. Now, my primary task is to conduct Exploratory Data Analysis.

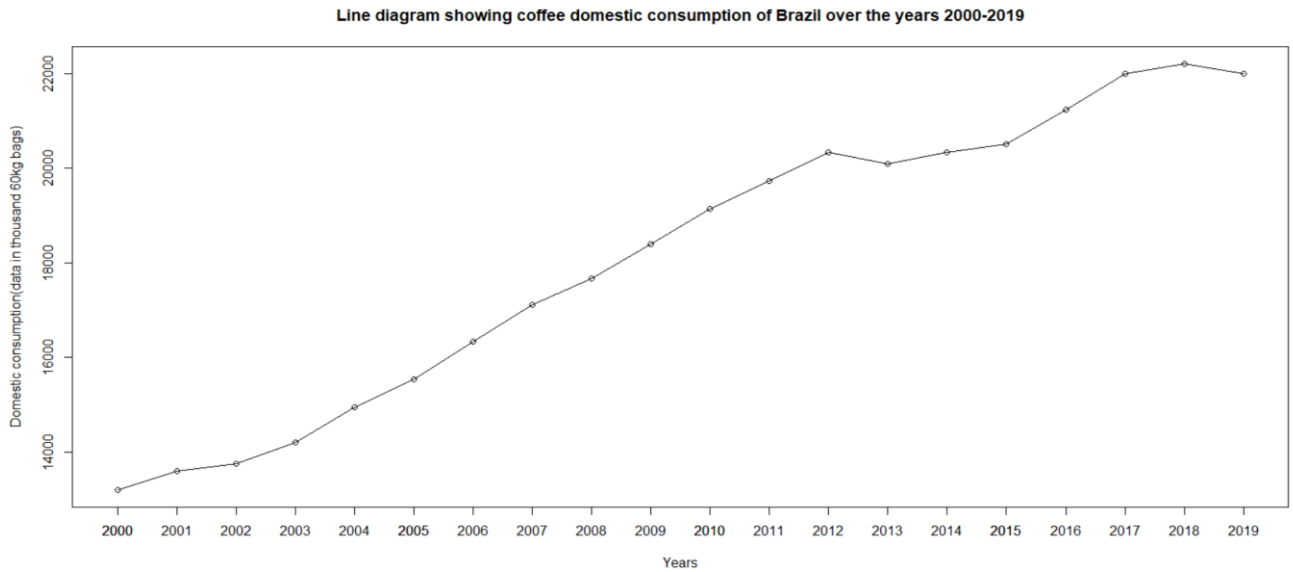
EXPLORATORY DATA ANALYSIS:

LINE DIAGRAMS:

1. BRAZIL

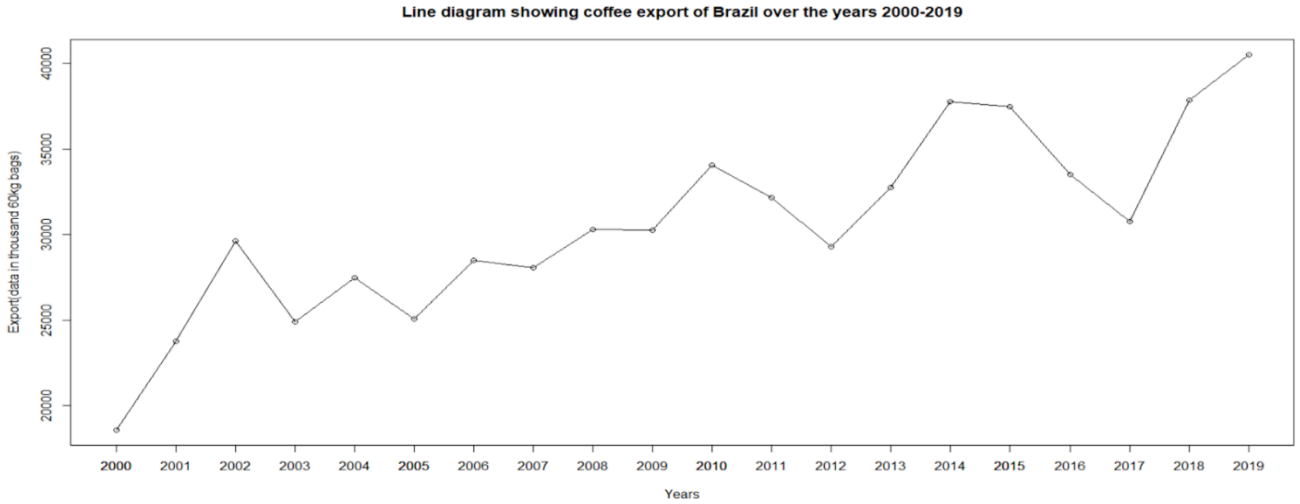


It is clear from the graph above that the data exhibits many ups and downs. This indicates that the data shows cyclical fluctuations, which is however beyond the scope of the syllabus. We can also notice an overall upward trend in the graph.



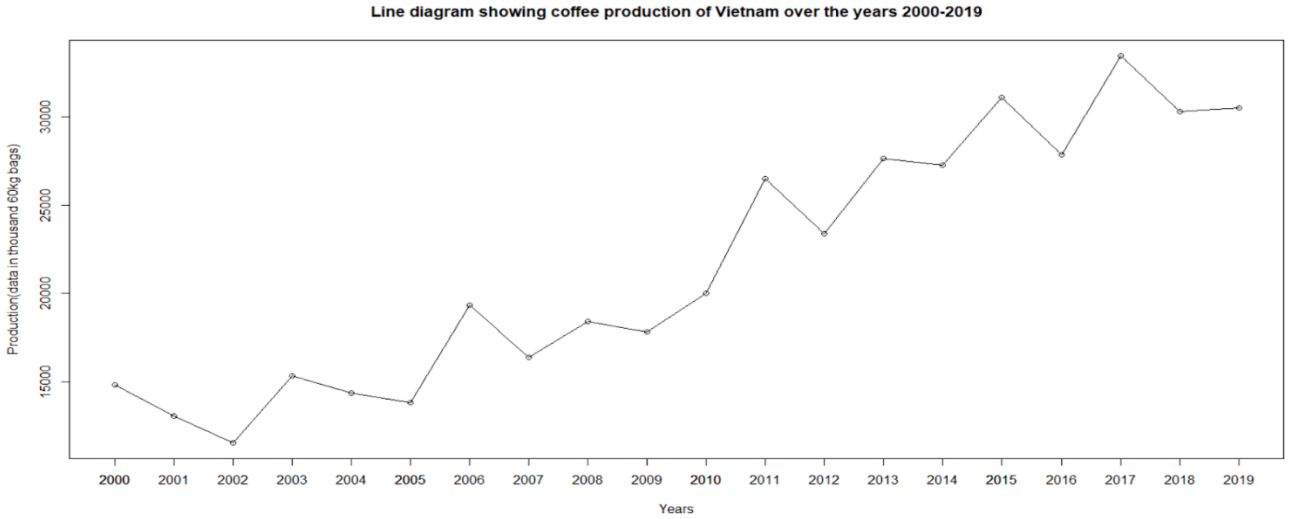
The graph illustrates an upward trend from 2000 to 2012. However, coffee domestic consumption

slightly dropped in 2013, and again, an upward trend can be seen from 2015 onwards. Hence, there is an overall upward trend in the graph.

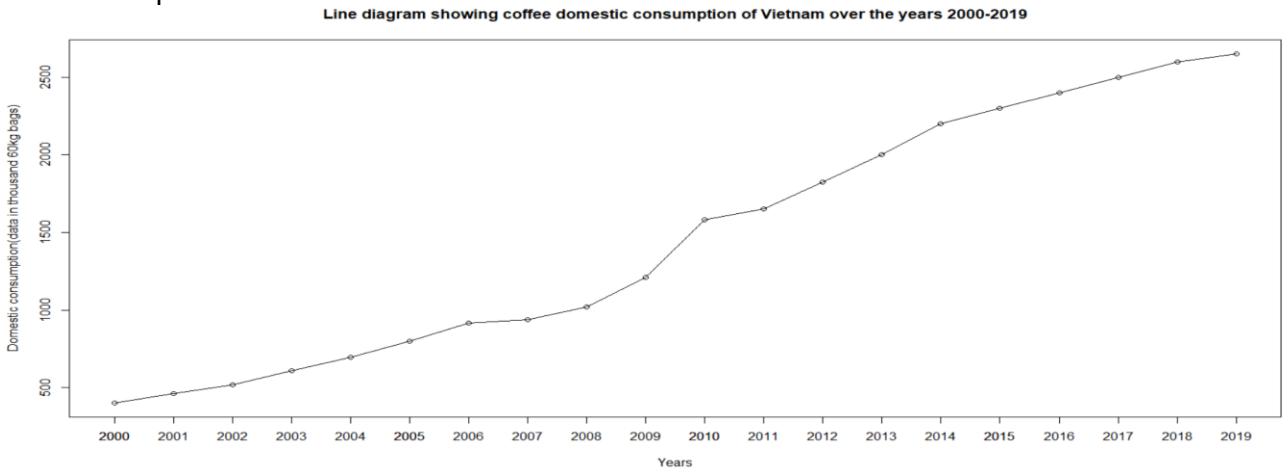


The graph illustrates a strictly upward trend from 2000 to 2002. From 2003 onwards, numerous oscillations can be seen, which indicate the presence of cyclical fluctuations in the data. Moreover, an overall upward trend is noticed.

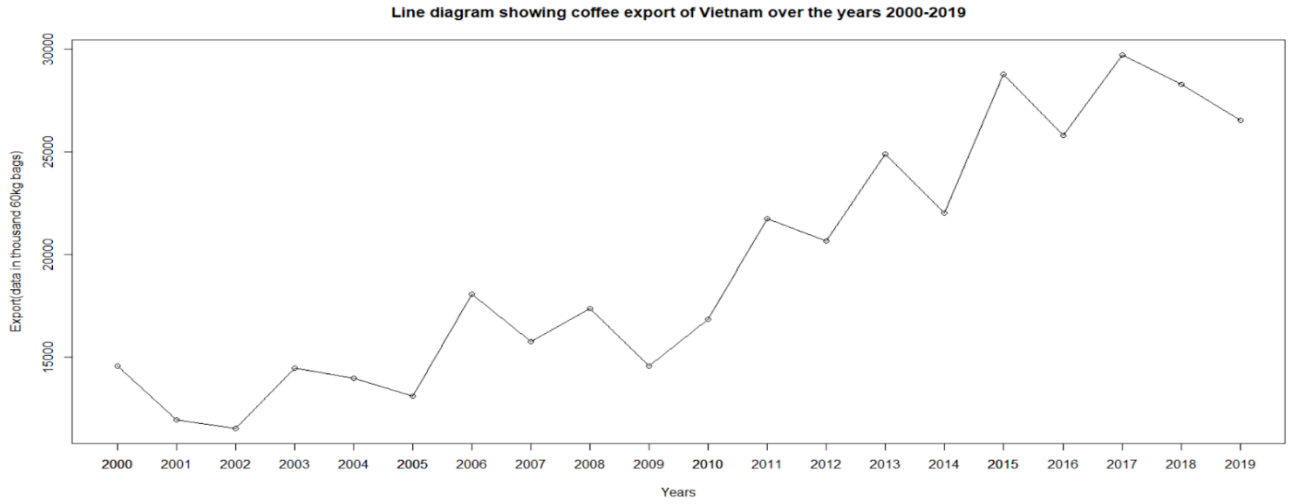
2. VIETNAM



The graph illustrates a strictly downward trend from 2000 to 2002. From 2003 onwards, numerous oscillations can be seen, which indicate the presence of cyclical fluctuations in the data. Moreover, an overall upward trend is noticed.

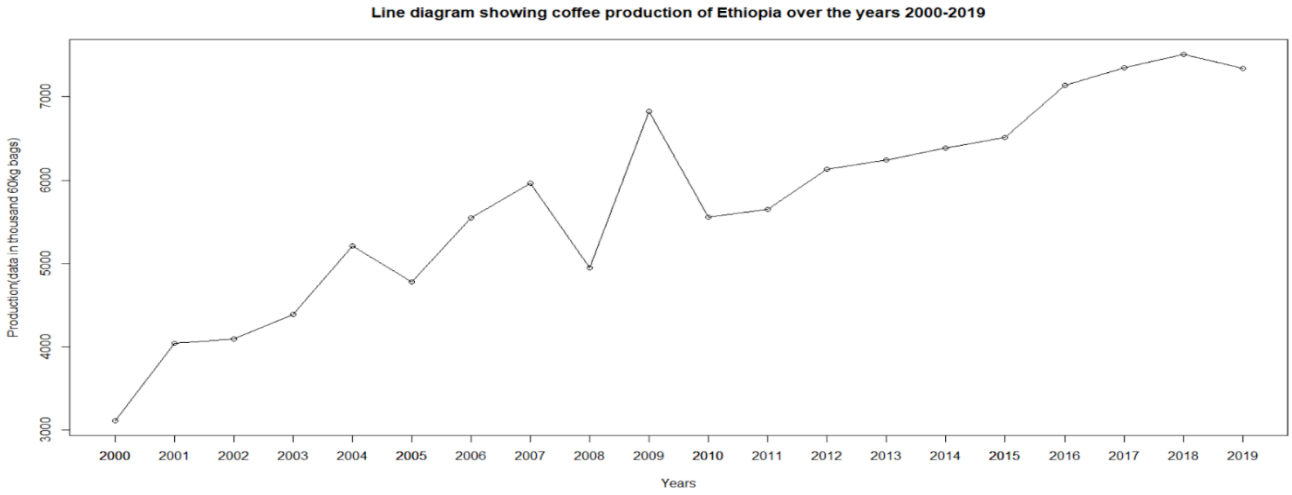


It is clear from the graph above that the data shows a strictly increasing linear trend over the years.

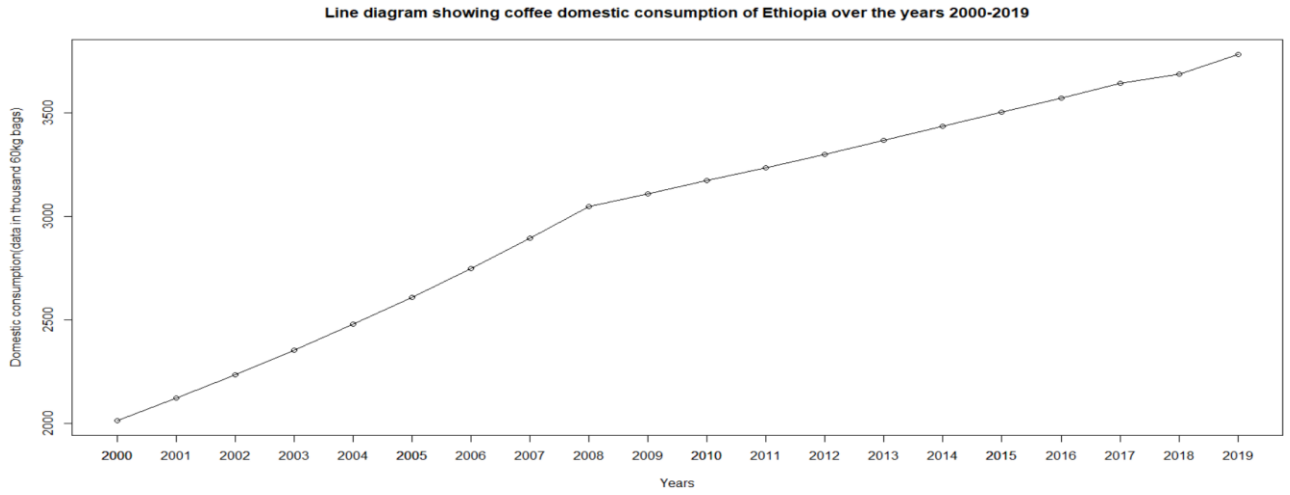


The graph illustrates a downward trend from 2000 to 2002. From 2002 onwards, numerous oscillations can be seen, which indicate the presence of cyclical fluctuations in the data. Moreover, an overall upward trend is noticed.

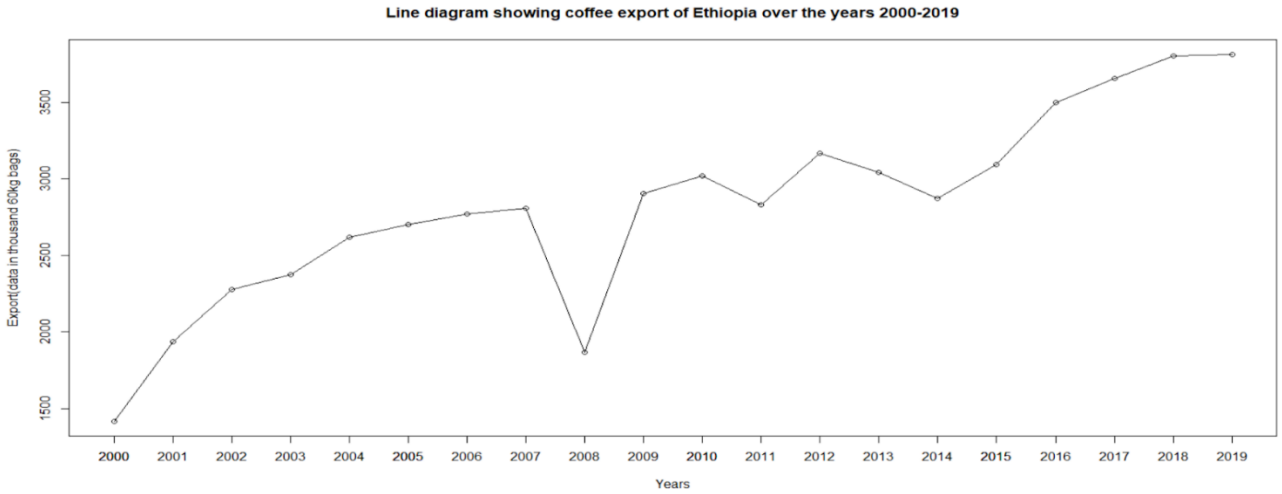
3. ETHIOPIA



It is clear from the graph that there is initially an increasing trend. Some fluctuations are visible between 2004 and 2010, but from 2010, a strictly increasing trend can be noted.

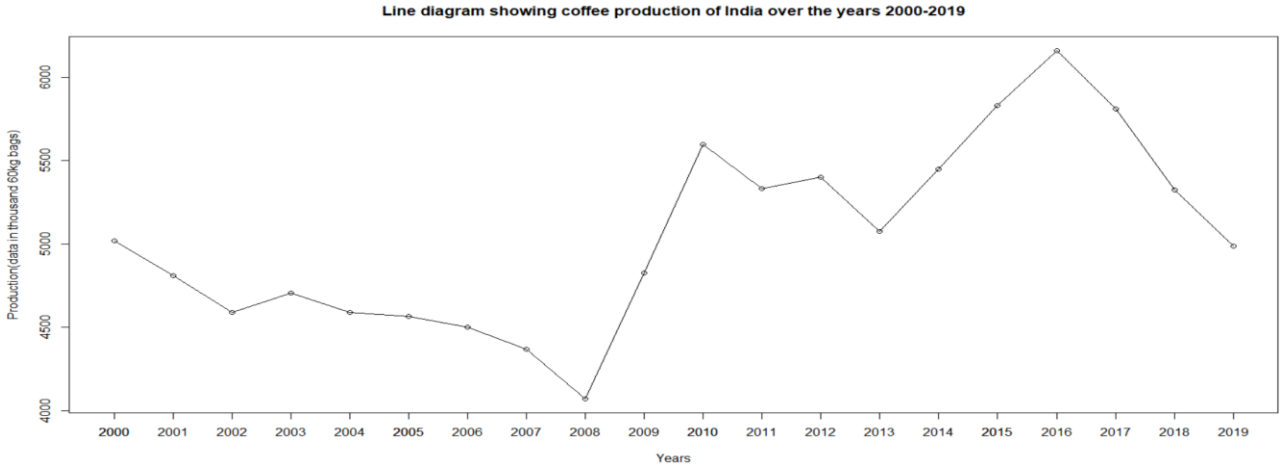


It is clear from the graph above that the data shows a strictly increasing linear trend over the years.

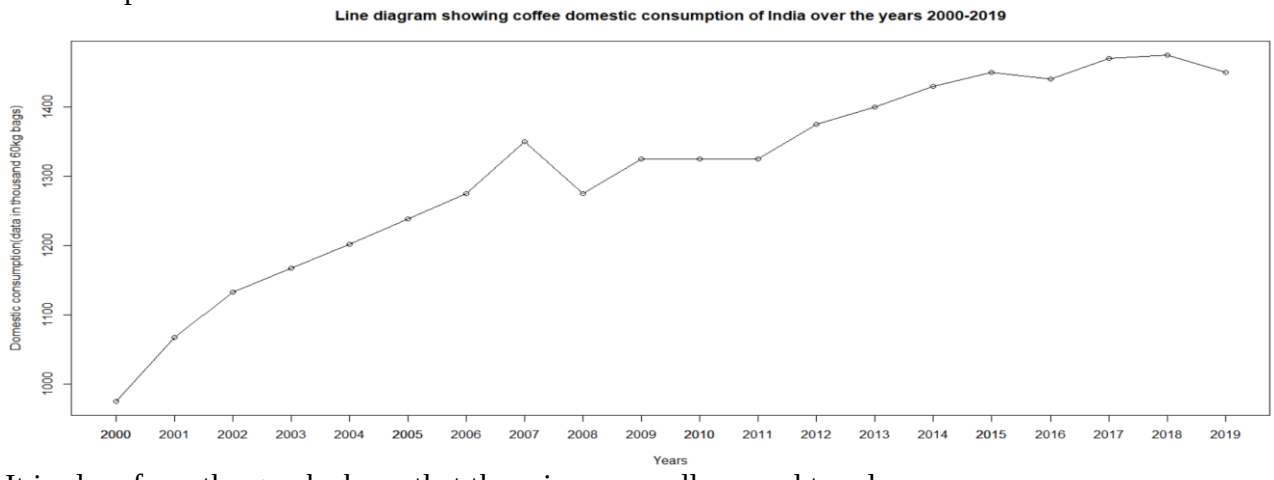


From 2000 to 2007, the graph shows an increasing trend. In 2008, a significant decrease in coffee export is noticed. A few oscillations are observed after 2010. In addition, an overall upward trend can be seen.

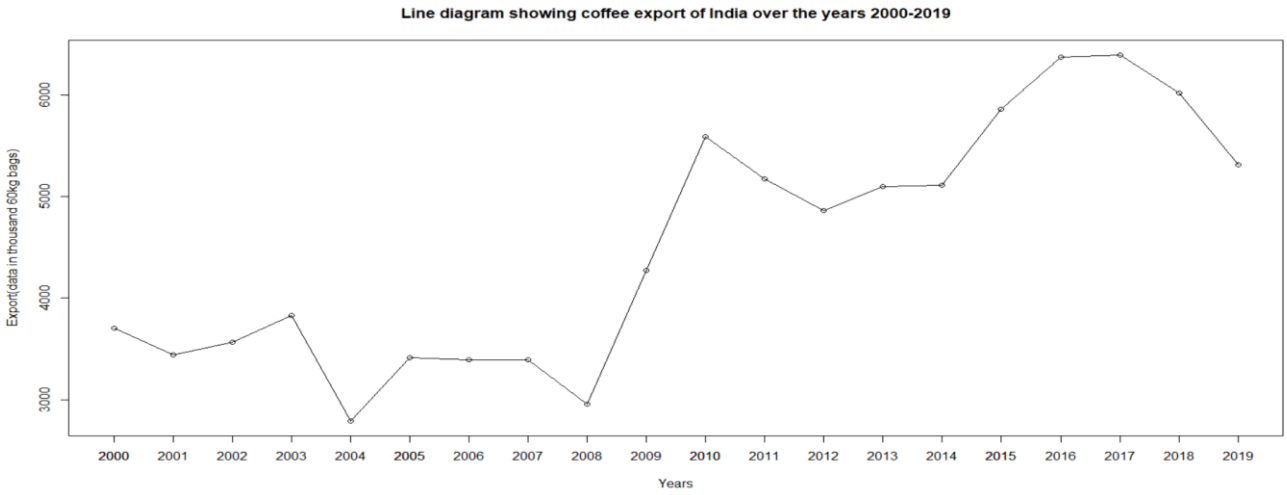
4. INDIA



The graph illustrates a downward trend from 2000 to 2008. Between 2008 and 2010, there was a significant rise in coffee production. A few oscillations can be seen after 2010. However, the overall trend is upward.



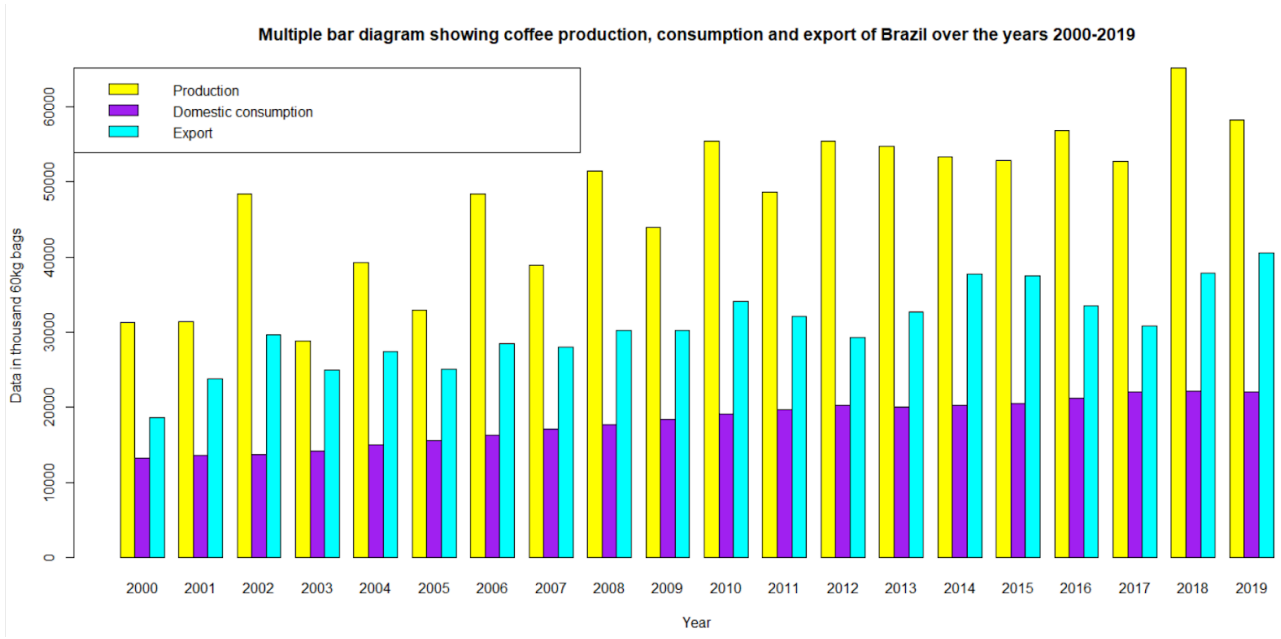
It is clear from the graph above that there is an overall upward trend.



The graph illustrates an upward trend from 2000 to 2003. From 2003 onwards, numerous oscillations can be seen, which indicate the presence of cyclical fluctuations in the data. Moreover, an overall upward trend is noticed.

MULTIPLE BAR DIAGRAMS:

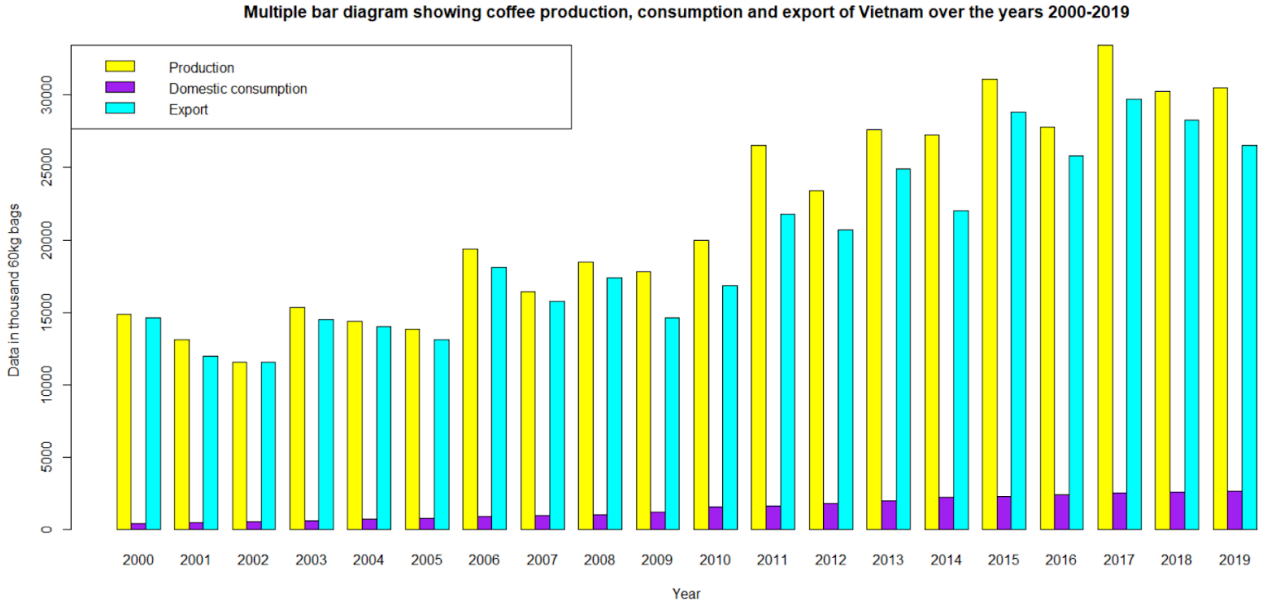
1. BRAZIL



From the graph above, we can observe the following:

- Production has significantly increased over the years. The quantity of coffee produced in 2018 was more than twice the quantity produced in 2000.
- The quantity of coffee exported has always been more than the quantity domestically consumed.

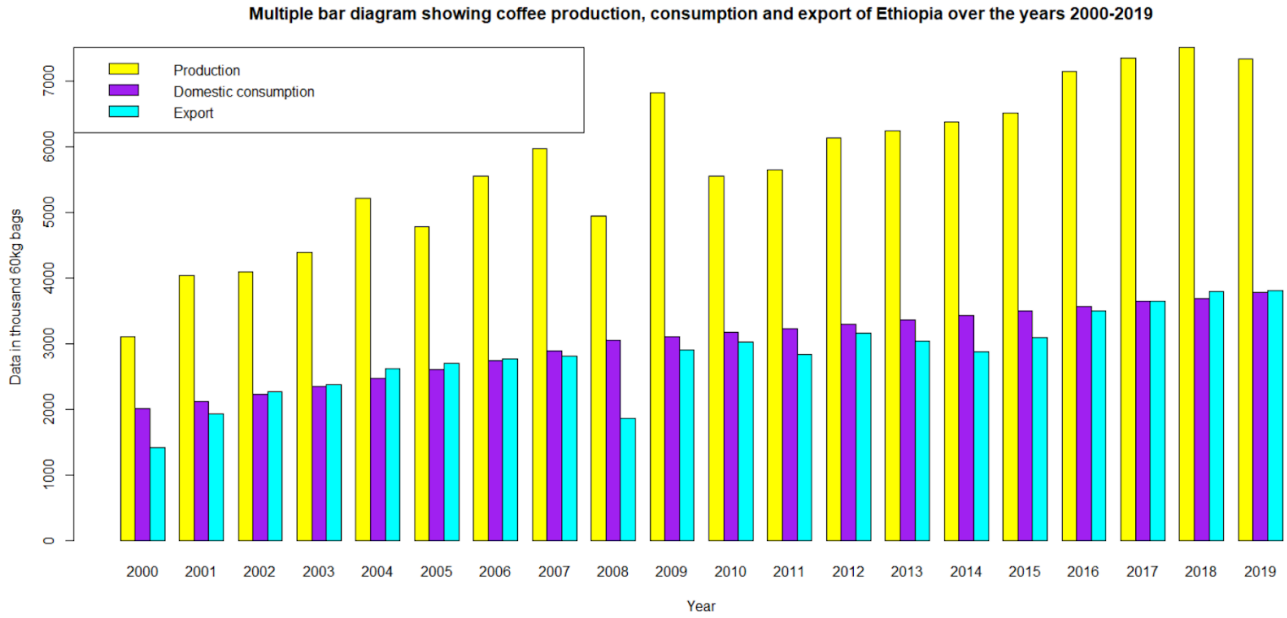
2. VIETNAM



From the graph above, we can observe the following:

- Production has significantly increased over the years. The quantity of coffee produced in 2019 was more than twice the quantity produced in 2000.
- A very large proportion of the coffee production is exported.
- When compared to coffee exports, domestic coffee consumption has been very minimal.

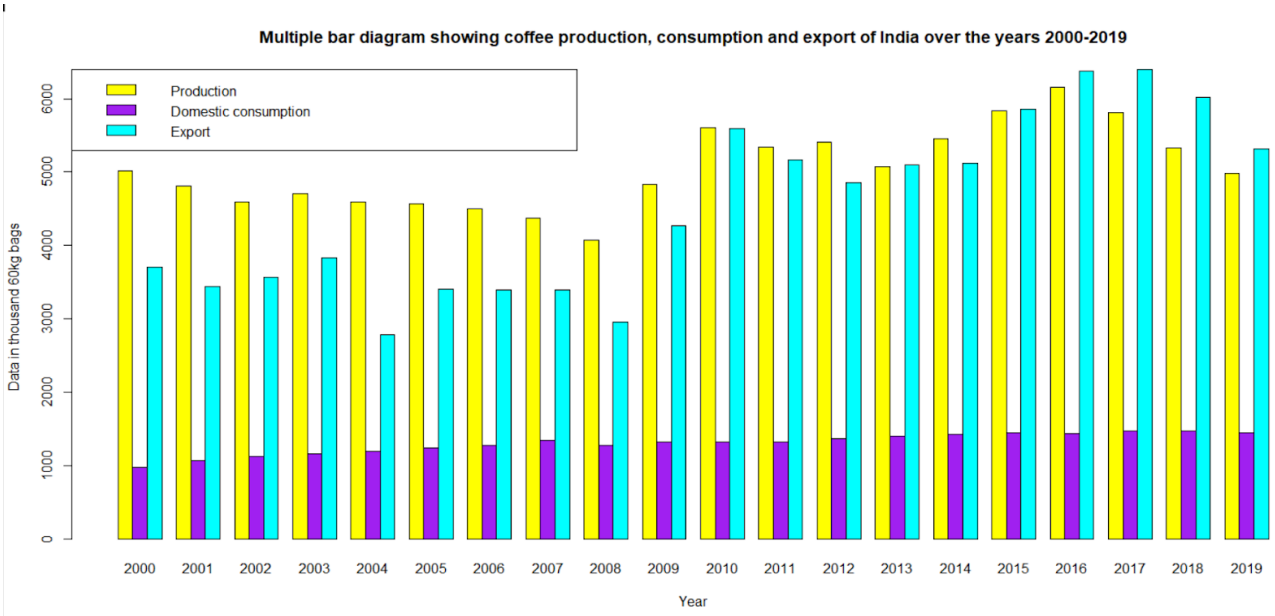
3. ETHIOPIA



From the graph above, we can observe the following:

- Despite various ups and downs throughout the period, coffee production has increased.
- Prior to 2017, more coffee was domestically consumed than exported. From 2017, there has been a little surplus in exporting coffee compared to domestic consumption.
- Both export and domestic consumption of coffee have increased over time

4. INDIA

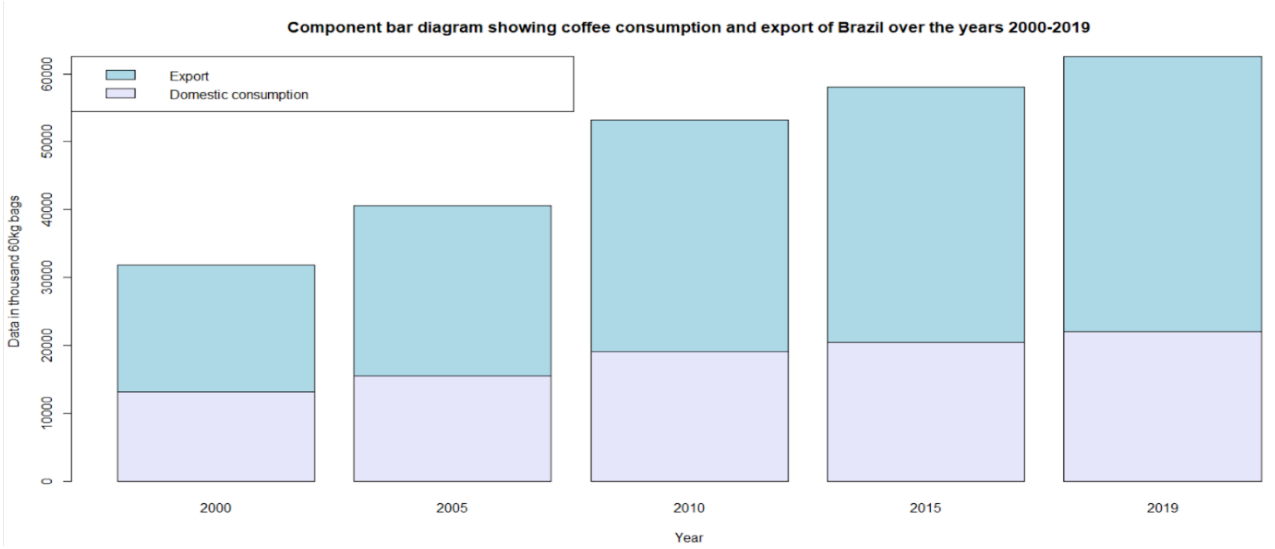


From the graph, we can observe the following:

- The amount of coffee produced in 2000 and 2019 is almost equivalent. Production hasn't really increased significantly.
- Domestic consumption has increased over time.
- The amount of coffee exported from 2015 has exceeded the amount produced. This phenomenon results from the lengthy procedure required to prepare the coffee cherries for export or consumption. As a result, not all of the coffee that is produced may be processed in a single year and thus, cannot be exported. The remaining coffee is exported the next year, bringing the total to a net export that exceeds production.

COMPONENT BAR DIAGRAMS:

1. BRAZIL

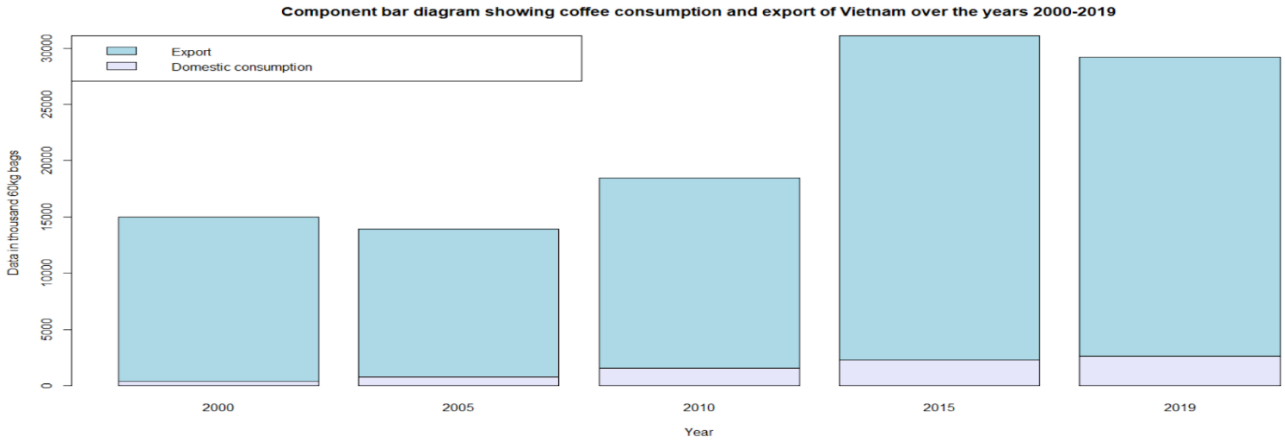


From the graph above, we can observe the following:

- Coffee export has increased more rapidly than domestic consumption over the years.
- In 2000, the quantity of coffee exported and consumed domestically were equivalent. In contrast, the amount of coffee exported in 2019 is almost twice as much as the amount consumed domestically.

- The production(domestic consumption + export) has also gone up. The amount of coffee exported in 2019 exceeds the amount produced in 2000.

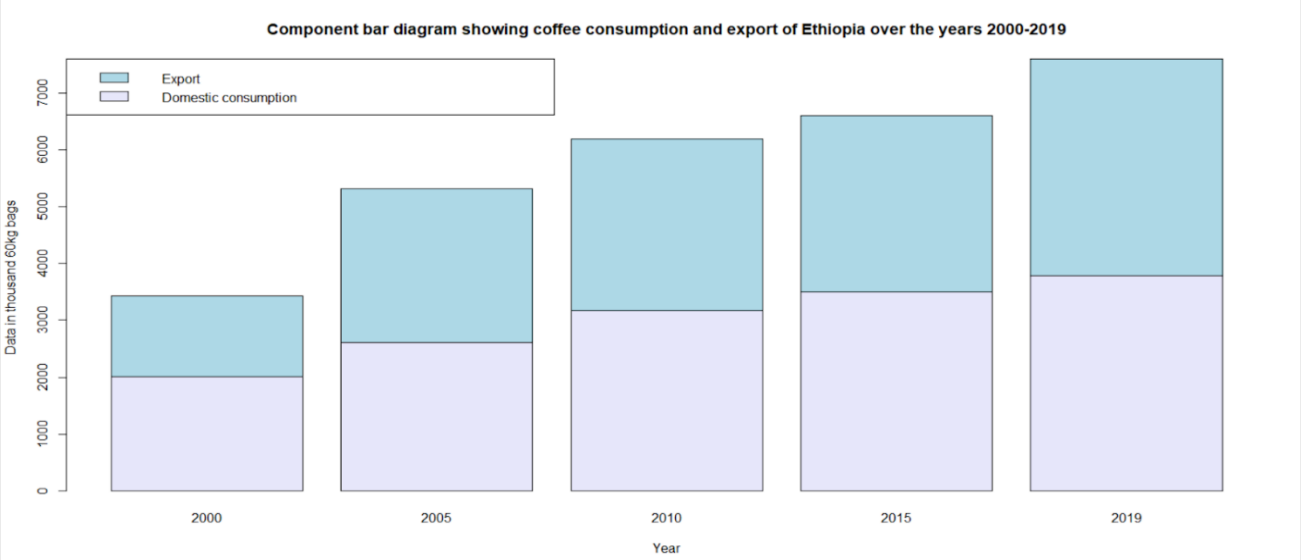
2. VIETNAM



From the graph, we can observe the following:

- The amount of coffee consumed domestically is almost negligible compared to the amount exported over the years.
- Coffee export has increased over the years. The amount of coffee exported in 2019 is almost twice the amount exported in 2000.

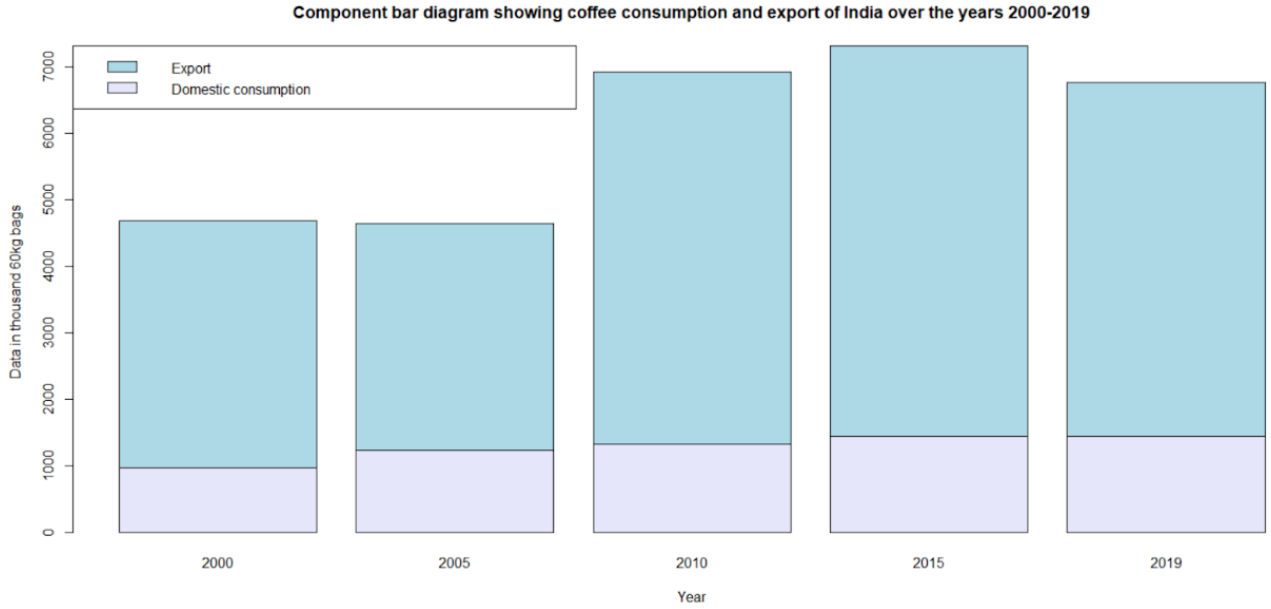
3. ETHIOPIA



From the graph, we can observe the following:

- In most of the cases, domestic consumption is more than export.
- Both coffee export and domestic consumption has increased over the years.
- The production(domestic consumption + export) has also gone up. The amount of coffee exported in 2019 exceeds the amount of produced in 2000.

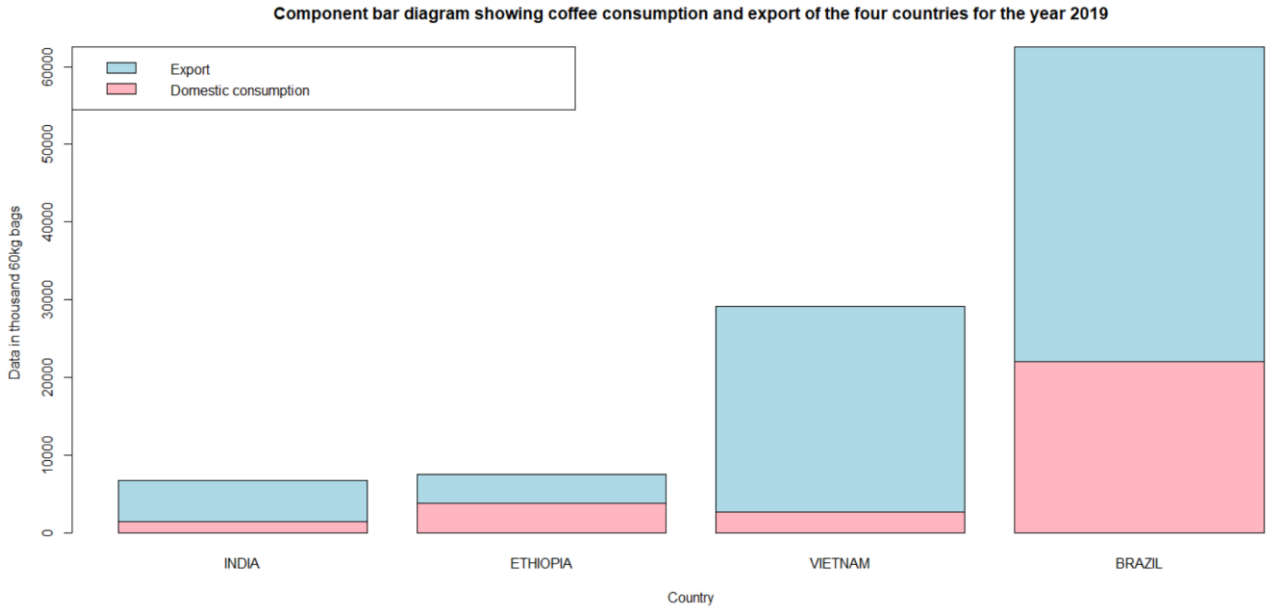
4. INDIA



From the graph above, we can observe the following:

- Coffee export has increased more than domestic consumption over the years.
- A large proportion of coffee production is exported rather than consumed domestically.

5. All countries together for a particular year – 2019



From the graph, we can observe the following:

- India has the least amount of coffee domestically consumed among the 4 countries whereas Brazil has the most. Arranging the countries in ascending order of their domestic consumption: India, Vietnam, Ethiopia and Brazil.
- Ethiopia has the least amount of coffee export whereas Brazil has the most. Ethiopia, India, Vietnam, and Brazil are the nations in ascending order of coffee export.
- Brazil produces the maximum amount of coffee, while India produces the least. India, Ethiopia, Vietnam, and Brazil are the nations in ascending order of coffee production.
- It is evident that Brazil is the world's largest coffee producer.

MODELLING:

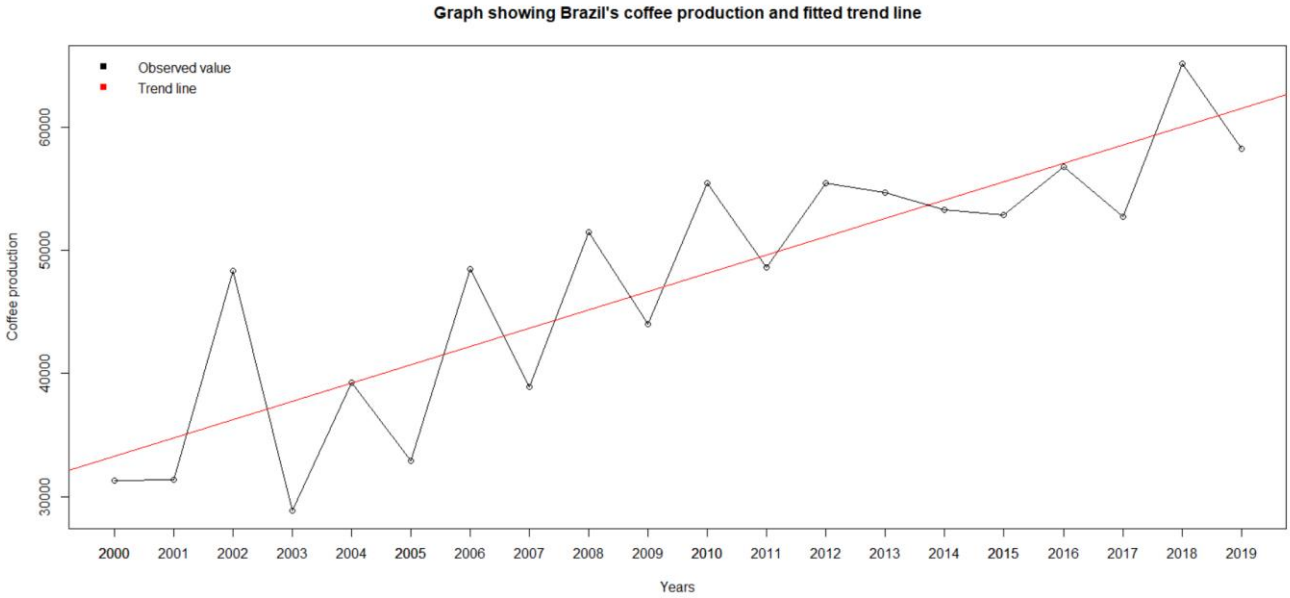
There are four datasets considered under this study, one for each country. These datasets will be used to train the prediction models for each country. The data for the year 2020 will be used to test the efficiency of the models.

FITTING TREND EQUATION AND FORECASTING USING TIME SERIES ANALYSIS:

Here, I have used data on coffee production, domestic consumption and export of Brazil, Vietnam, Ethiopia and India to draw time series plot, fit trend equation and forecast the future values.

Fitting trend equation on coffee production

1. BRAZIL



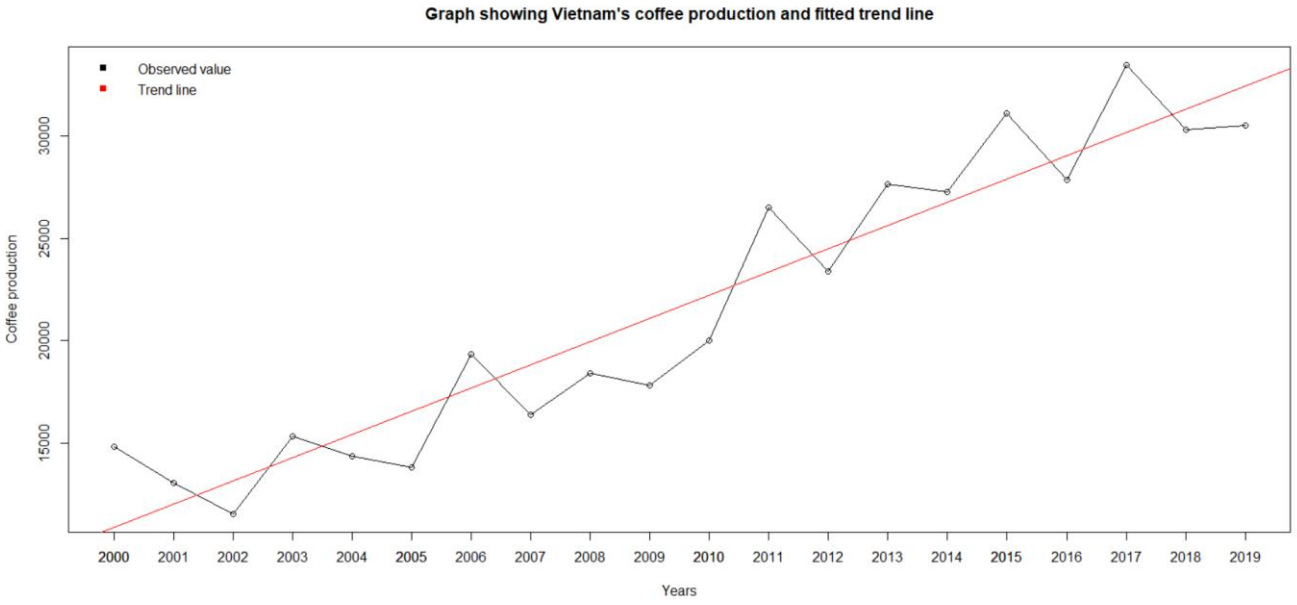
Here, a linear trend equation has been fitted.
The trend equation obtained is prd_b = -2937808.6 + 1485.6*t
The value of adjusted R-squared is 0.7064 which implies 70.64% of the total variability can be explained by this trend equation.

Table showing the observed and fitted values of coffee production in Brazil

Year	Observed values	Fitted values
2000	31310	33292.17
2001	31365	34777.72
2002	48352	36263.27
2003	28873	37748.82
2004	39281	39234.37
2005	32933	40719.92
2006	48432	42205.47
2007	38911	43691.02
2008	51491	45176.57
2009	43977	46662.12
2010	55428	48147.68

2011	48592	49633.23
2012	55418	51118.78
2013	54689	52604.33
2014	53305	54089.88
2015	52871	55575.43
2016	56788	57060.98
2017	52740	58546.53
2018	65131	60032.08
2019	58211	61517.63

2. VIETNAM



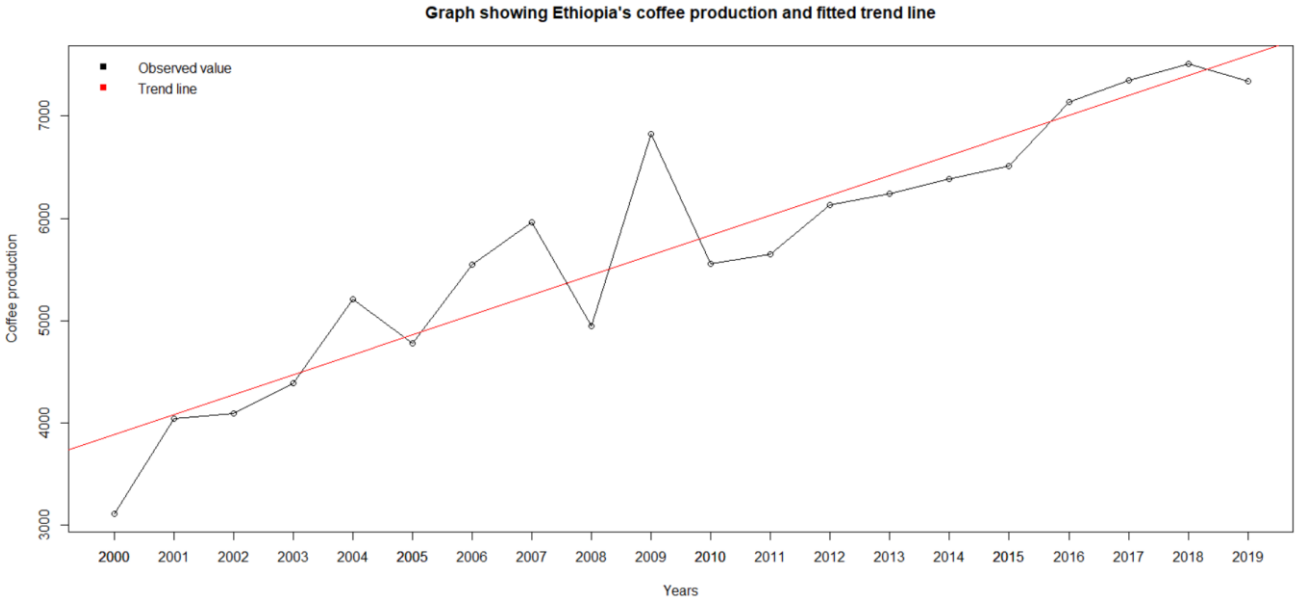
Here, a linear trend equation has been fitted.
The trend equation obtained is prd_v = -2.250e+06 + 1.130e+03*t
The value of adjusted R-squared is 0.8911 which implies 89.11% of the total variability can be explained by this trend equation.

Table showing the observed and fitted values of coffee production in Vietnam

Year	Observed values	Fitted values
2000	14841	10907.97
2001	13093	12038.34
2002	11574	13168.70
2003	15337	14299.07
2004	14370	15429.44
2005	13842	16559.80
2006	19340	17690.17
2007	16405	18820.53
2008	18438	19950.90
2009	17825	21081.27
2010	20000	22211.63
2011	26500	23342.00
2012	23402	24472.37

2013	27610	25602.73
2014	27241	26733.10
2015	31090	27863.46
2016	27819	28993.83
2017	33432	30124.20
2018	30283	31254.56
2019	30487	32384.93

3. ETHIOPIA



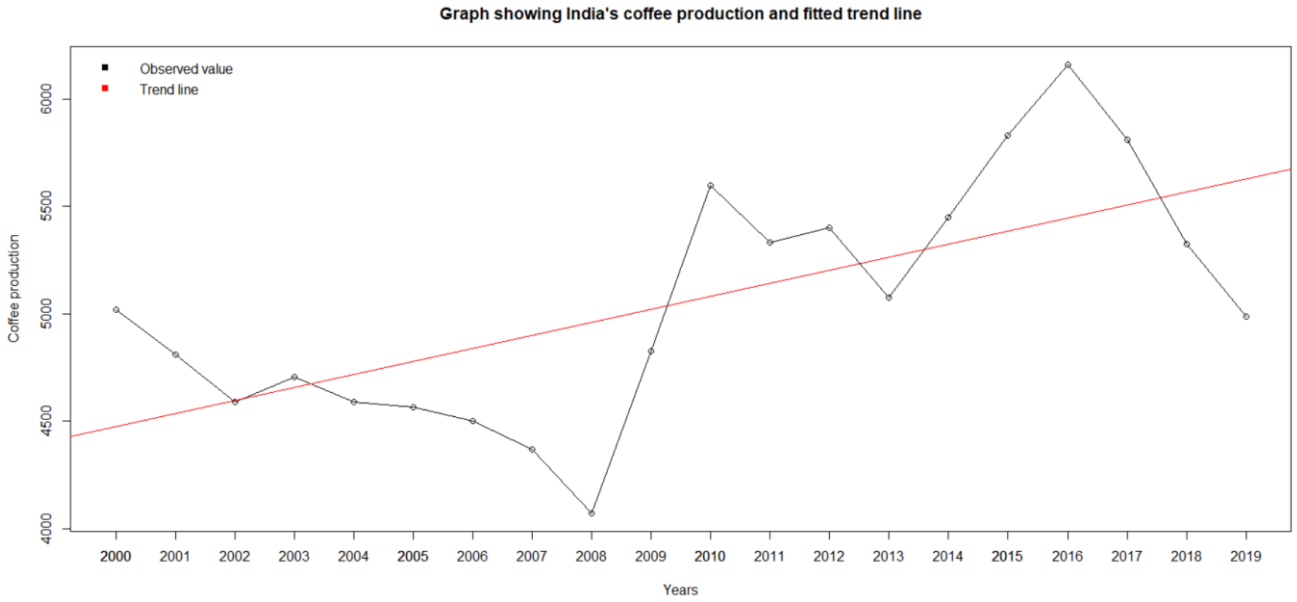
Here, a linear trend equation has been fitted.
The trend equation obtained is prd_e = -385670.2 + 194.8*t
The value of adjusted R-squared is 0.8594 which implies 85.94% of the total variability can be explained by this trend equation.

Table showing the observed and fitted values of coffee production in Ethiopia

Year	Observed values	Fitted values
2000	3115	3887.700
2001	4044	4082.479
2002	4094	4277.258
2003	4394	4472.037
2004	5213	4666.816
2005	4779	4861.595
2006	5551	5056.374
2007	5967	5251.153
2008	4949	5445.932
2009	6830	5640.711
2010	5560	5835.489
2011	5650	6030.268
2012	6132	6225.047
2013	6242	6419.826

2014	6383	6614.605
2015	6515	6809.384
2016	7143	7004.163
2017	7347	7198.942
2018	7511	7393.721
2019	7343	7588.500

4. INDIA



Here, a linear trend equation has been fitted.
The trend equation obtained is prd_i = -116969.07 + 60.72*t
The value of adjusted R-squared is 0.3921 which implies 39.21% of the total variability can be explained by this trend equation.

Table showing the observed and fitted values of coffee production in India

Year	Observed values	Fitted values
2000	5020	4474.543
2001	4810	4535.265
2002	4588	4595.986
2003	4708	4656.708
2004	4591	4717.430
2005	4566	4778.152
2006	4500	4838.874
2007	4367	4899.595
2008	4072	4960.317
2009	4827	5021.039
2010	5600	5081.761
2011	5334	5142.483
2012	5403	5203.205
2013	5075	5263.926
2014	5450	5324.648
2015	5830	5385.370

2016	6161	5446.092
2017	5813	5506.814
2018	5325	5567.535
2019	4988	5628.257

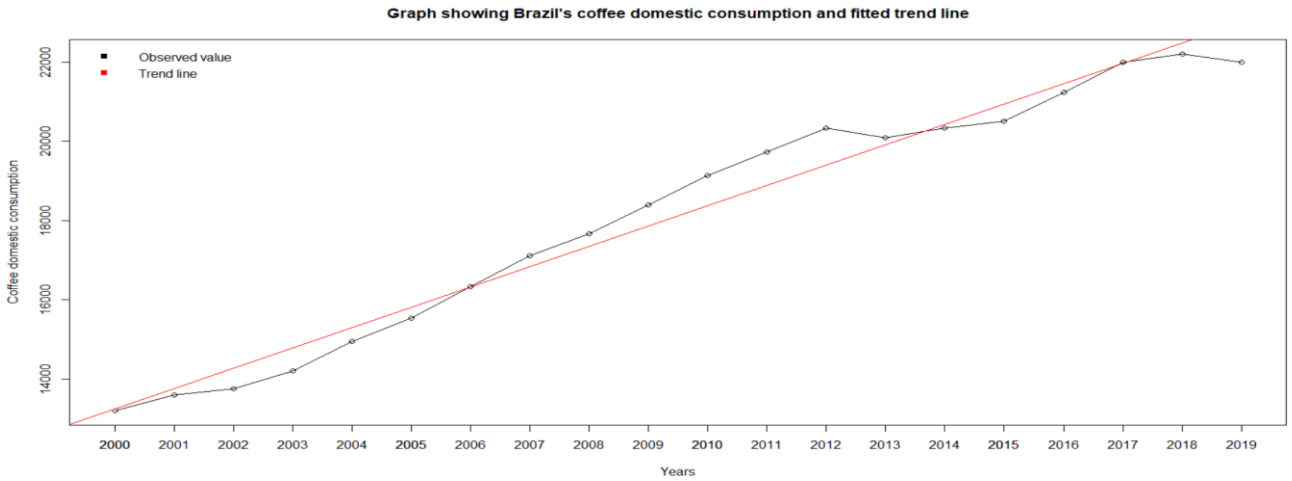
Forecasting: Here, coffee production for the year 2020 is being predicted using the trend equations obtained for each of the four countries.

Table showing observed and predicted coffee production for the year 2020

COUNTRY	PREDICTED COFFEE PRODUCTION	COFFEE PRODUCTION (DATA IN 1000 60KG BAG)
BRAZIL (prd_b = -2937808.6 + 1485.6*t)	63003.18	69000
VIETNAM (prd_v = -2.250e+06 + 1.130e+03*t)	33515.29	29000
ETHIOPIA (prd_e = -385670.2 + 194.8*t)	7783.279	7375
INDIA (prd_i = -116969.07 + 60.72*t)	5688.979	5700

Fitting trend equation on coffee domestic consumption

1. BRAZIL



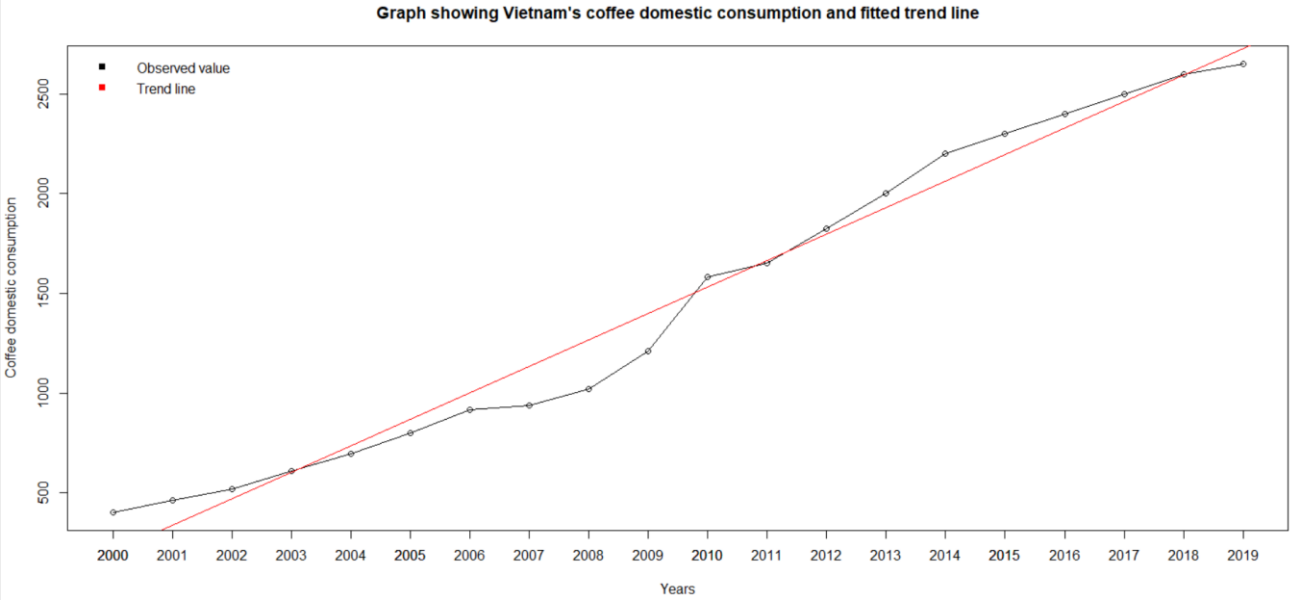
Here, a linear trend equation has been fitted.
The trend equation obtained is dc_b = -1.013e+06 + 5.130e+02*t
The value of adjusted R-squared is 0.9721 which implies 97.21% of the total variability can be explained by this trend equation.

Table showing the observed and fitted values of coffee domestic consumption in Brazil

Year	Observed values	Fitted values
2000	13200	132390.13

2001	13590	13752.09
2002	13750	14265.05
2003	14200	14778.01
2004	14946	15290.97
2005	15538	15803.93
2006	16331	16316.89
2007	17110	16829.85
2008	17660	17342.81
2009	18390	17855.77
2010	19132	18368.73
2011	19720	18881.69
2012	20330	19394.65
2013	20085	19907.61
2014	20333	20420.57
2015	20508	20933.53
2016	21225	21446.49
2017	21997	21959.45
2018	22200	22472.41
2019	22000	22985.37

2. VIETNAM



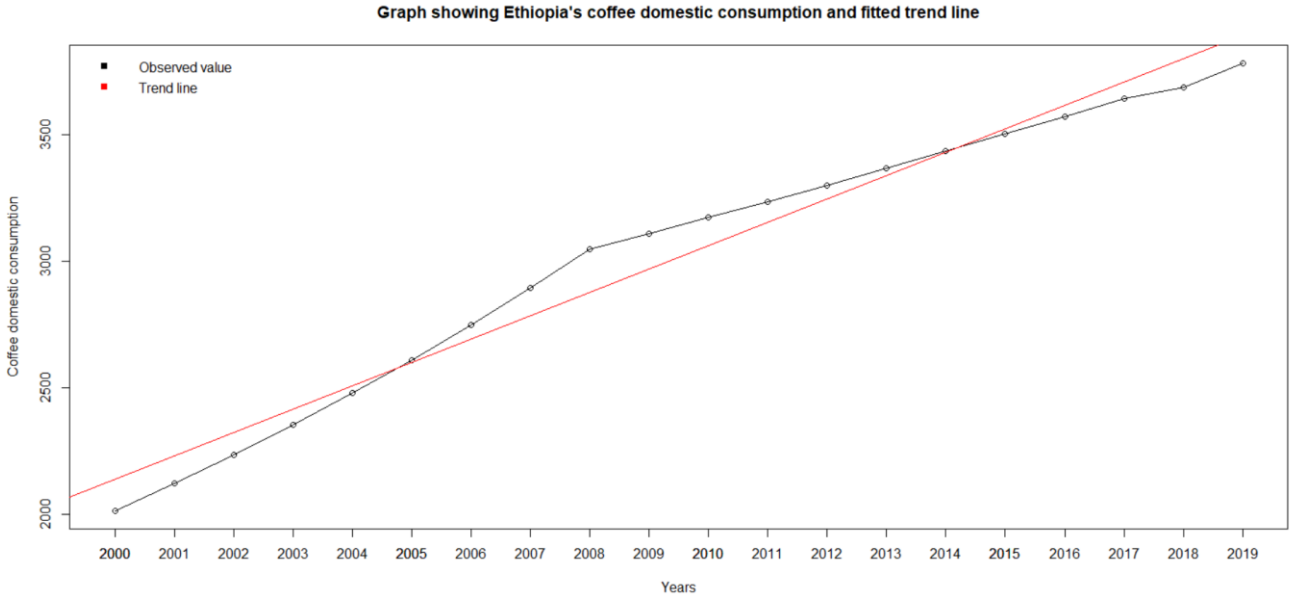
Here, a linear trend equation has been fitted.
The trend equation obtained is dc_v = -2.654e+05 + 1.328e+02*t
The value of adjusted R-squared is 0.9775 which implies 97.75% of the total variability can be explained by this trend equation.

Table showing the observed and fitted values of coffee domestic consumption in Brazil

Year	Observed values	Fitted values
2000	402	202.2714
2001	461	335.0692
2002	519	467.8669

2003	607	600.6647
2004	696	733.4624
2005	800	866.2602
2006	917	999.0579
2007	938	1131.8556
2008	1021	1264.6534
2009	1208	1397.4511
2010	1583	1530.2489
2011	1650	1663.0466
2012	1825	1795.8444
2013	2000	1928.6421
2014	2200	2061.4398
2015	2300	2194.2376
2016	2400	2327.0353
2017	2500	2459.8331
2018	2600	2592.6308
2019	2650	2725.4286

3. ETHIOPIA



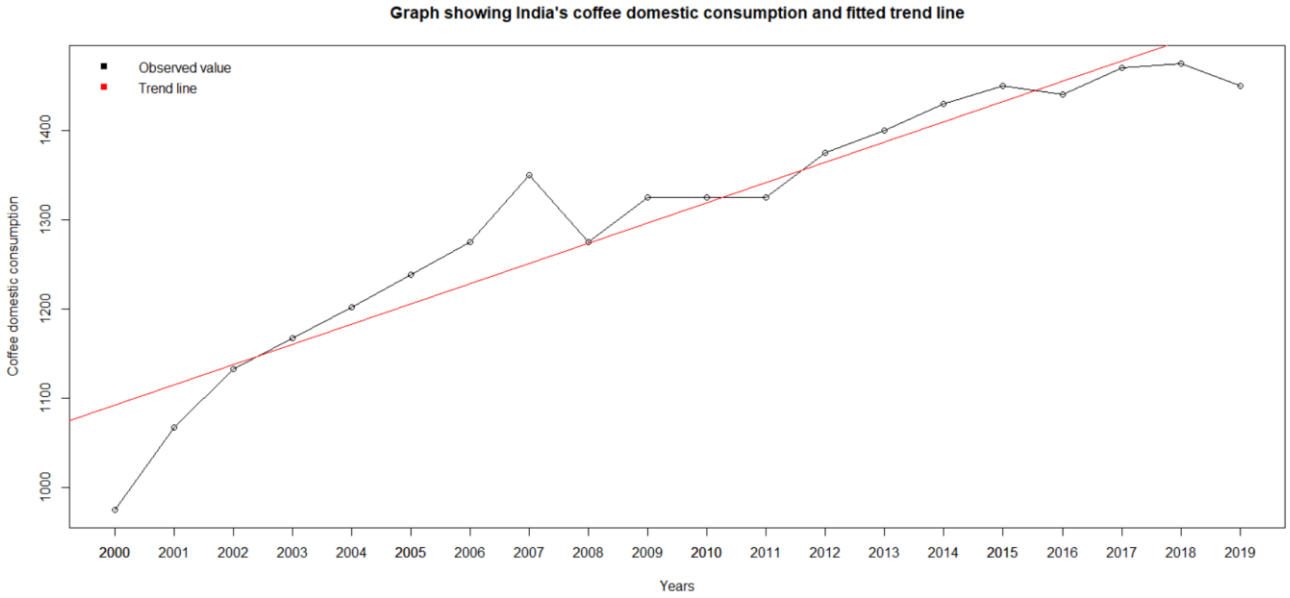
Here, a linear trend equation has been fitted.
The trend equation obtained is dc_e = -1.825e+05 + 9.231e+01*t
The value of adjusted R-squared is 0.9713 which implies 97.13% of the total variability can be explained by this trend equation.

Table showing the observed and fitted values of coffee domestic consumption in Ethiopia

Year	Observed values	Fitted values
2000	2014	2137.700
2001	2121	2230.005
2002	2234	2322.311
2003	2353	2414.616
2004	2478	2506.921

2005	2609	2599.226
2006	2748	2691.532
2007	2894	2783.837
2008	3048	2876.142
2009	3109	2968.447
2010	3171	3060.753
2011	3235	3153.058
2012	3299	3245.363
2013	3365	3337.668
2014	3433	3429.974
2015	3501	3522.279
2016	3571	3614.584
2017	3643	3706.889
2018	3685	3799.195
2019	3781	3891.500

4. INDIA



Here, a linear trend equation has been fitted.
The trend equation obtained is dc_i = -44278.08 + 22.68*t
The value of adjusted R-squared is 0.8947 which implies 89.47% of the total variability can be explained by this trend equation.

Table showing the observed and fitted values of coffee domestic consumption in India

Year	Observed values	Fitted values
2000	975	1091.843
2001	1067	1114.528
2002	1133	1137.213
2003	1167	1159.898
2004	1202	1182.583
2005	1238	1205.268
2006	1275	1227.953

2007	1350	1250.638
2008	1275	1273.323
2009	1325	1296.008
2010	1325	1318.692
2011	1325	1341.377
2012	1375	1364.062
2013	1400	1386.747
2014	1430	1409.432
2015	1450	1432.117
2016	1440	1454.802
2017	1470	1477.487
2018	1475	1500.172
2019	1450	1522.857

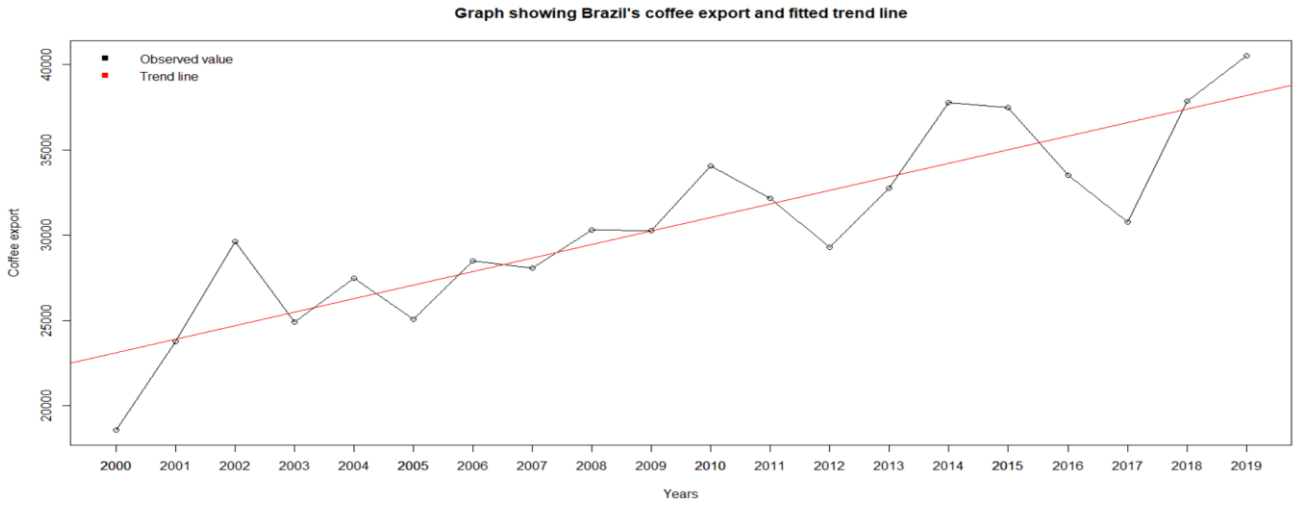
Forecasting: Here, coffee consumed domestically for the year 2020 is being predicted using the trend equations obtained for each of the four countries.

Table showing observed and predicted coffee domestic consumption for the year 2020

COUNTRY	PREDICTED COFFEE DOMESTIC CONSUMPTION	COFFEE DOMESTIC CONSUMPTION (DATA IN 1000 60KG BAG)
BRAZIL (dc_b = -1.013e+06 + 5.130e+02*t)	23498.33	22300
VIETNAM (dc_v = -2.654e+05 + 1.328e+02*t)	2858.226	2800
ETHIOPIA (dc_e = -1.825e+05 + 9.231e+01*t)	3983.805	3900
INDIA (dc_i = -44278.08 + 22.68*t)	1545.542	1600

Fitting trend equation on coffee export

1. BRAZIL

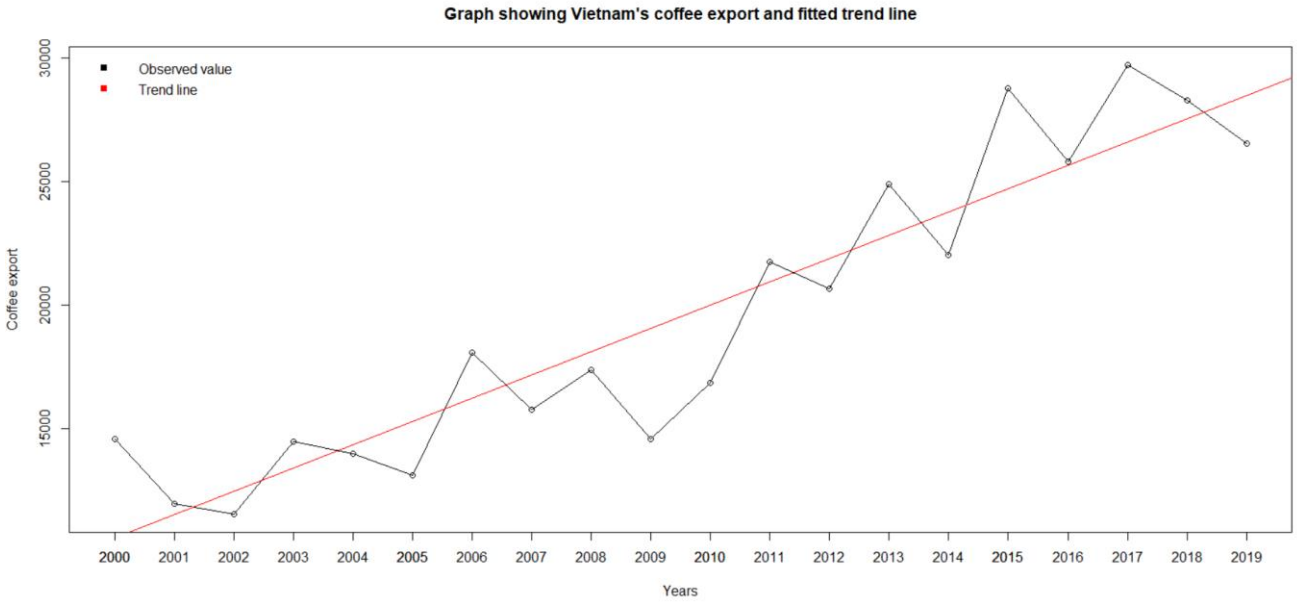


Here, a linear trend equation has been fitted.
The trend equation obtained is exp_b = -1564021.2 + 793.6*t
The value of adjusted R-squared is 0.7408 which implies 74.08% of the total variability can be explained by this trend equation.

Table showing the observed and fitted values of coffee export in Brazil

Year	Observed values	Fitted values
2000	18577	23093.06
2001	23767	23886.61
2002	29613	24680.17
2003	24909	25473.73
2004	27468	26267.29
2005	25078	27060.84
2006	28486	27854.40
2007	28044	28647.96
2008	30292	29441.51
2009	30255	30235.07
2010	34054	31028.63
2011	32149	31822.19
2012	29283	32615.74
2013	32752	33409.30
2014	37782	34202.86
2015	37473	34996.41
2016	33491	35789.97
2017	30783	36583.53
2018	37870	37377.09
2019	40511	38170.64

2. VIETNAM



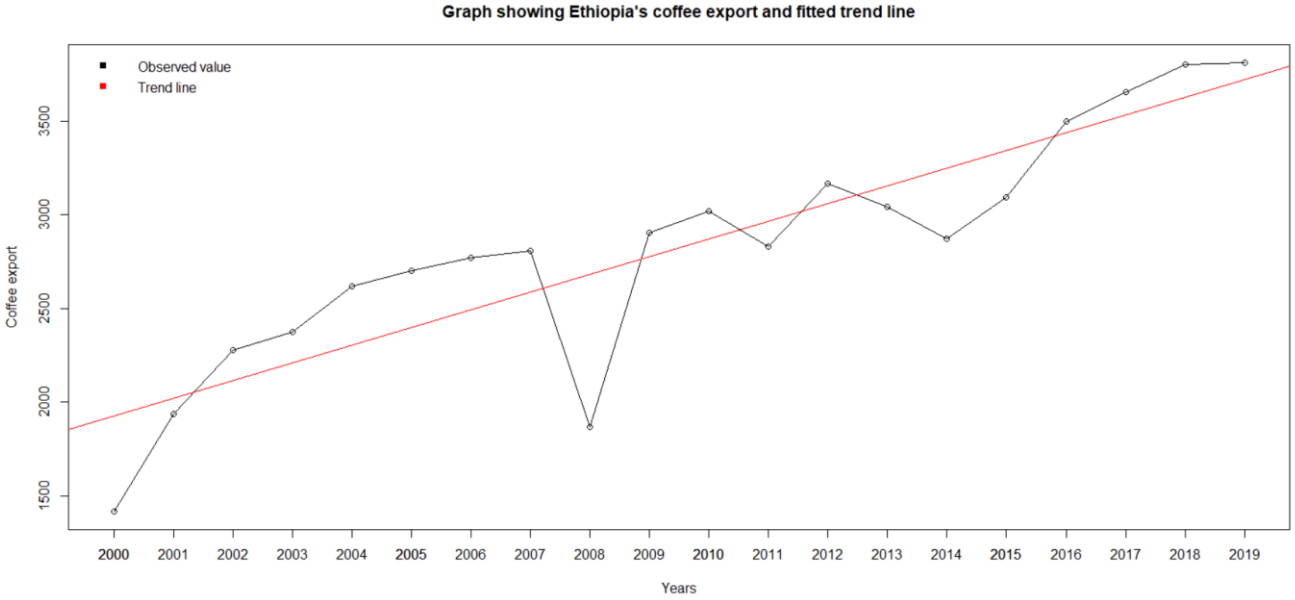
Here, a linear trend equation has been fitted.
The trend equation obtained is exp_v = -1.876e+06 + 9432e+02*t

The value of adjusted R-squared is 0.8482 which implies 84.82% of the total variability can be explained by this trend equation.

Table showing the observed and fitted values of coffee export in Vietnam

Year	Observed values	Fitted values
2000	14606	13825.14
2001	11966	12355.28
2002	11555	11077.98
2003	14497	14242.21
2004	13994	13429.08
2005	13122	12985.10
2006	18090	17608.25
2007	15774	15140.27
2008	17386	16849.78
2009	14591	16334.32
2010	16850	18163.23
2011	21760	23628.95
2012	20665	21023.91
2013	24902	24562.33
2014	22035	24525.04
2015	28790	27488.59
2016	25819	24738.07
2017	29732	29457.93
2018	28283	26810.00
2019	26537	26981.54

3. ETHIOPIA

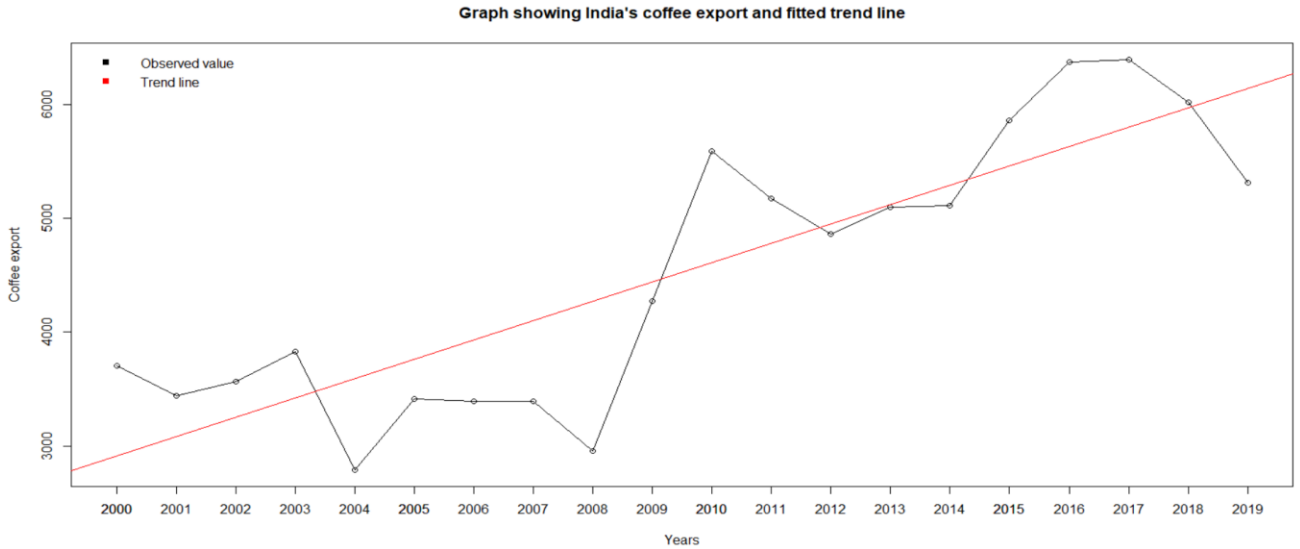


Here, a linear trend equation has been fitted.
The trend equation obtained is exp_e = -187314.48 + 94.62*t
The value of adjusted R-squared is 0.7721 which implies 77.21% of the total variability can be explained by this trend equation.

Table showing the observed and fitted values of coffee export in Ethiopia

Year	Observed values	Fitted values
2000	1418	1924.614
2001	1939	2019.234
2002	2277	2133.853
2003	2374	2208.473
2004	2620	2303.092
2005	2702	2397.712
2006	2770	2492.332
2007	2806	2586.951
2008	1868	2681.571
2009	2904	2776.190
2010	3022	2870.810
2011	2832	2965.429
2012	3166	3060.049
2013	3044	3154.668
2014	2872	3249.288
2015	3092	3343.098
2016	3497	3438.527
2017	3654	3533.147
2018	3801	3627.766
2019	3812	3722.386

4. INDIA



Here, a linear trend equation has been fitted.
The trend equation obtained is exp_i = -337415.52 + 170.16*t
The value of adjusted R-squared is 0.7112 which implies 71.12% of the total variability can be explained by this trend equation.

Table showing the observed and fitted values of coffee export in India

Year	Observed values	Fitted values
2000	3705	2910.800

2001	3441	3080.963
2002	3567	3251.126
2003	3826	3421.289
2004	2790	3591.453
2005	3410	3761.616
2006	3393	3931.779
2007	3389	4101.942
2008	2954	4272.105
2009	4274	4442.268
2010	5594	4612.432
2011	5172	4782.595
2012	4859	4952.758
2013	5095	5122.921
2014	5115	5293.084
2015	5861	5463.247
2016	6371	5633.411
2017	6395	5803.574
2018	6022	5973.737
2019	5314	6143.900

Forecasting: Here, coffee export for the year 2020 is being predicted using the trend equations obtained for each of the four countries.

Table showing observed and predicted coffee export for the year 2020

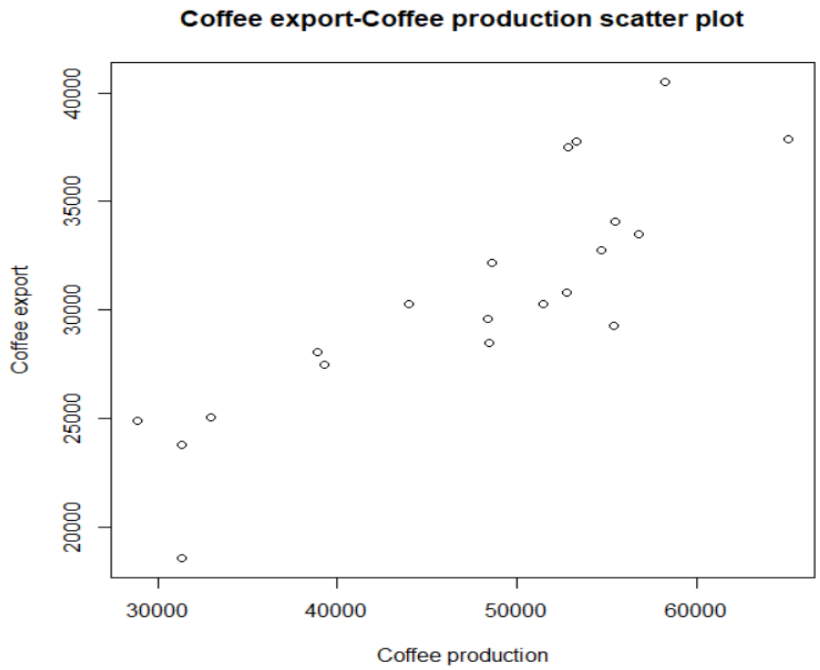
COUNTRY	PREDICTED COFFEE EXPORT	COFFEE EXPORT (DATA IN 1000 60KG BAG)
BRAZIL (exp_b = -1564021.2 + 793.6*t)	38964.2	44848
VIETNAM (exp_v = -1.876e+06 + 9432e+02*t)	29450.95	25625
ETHIOPIA (exp_e = -187314.48 + 94.62*t)	3817.005	3443
INDIA (exp_i = -337415.52 + 170.16*t)	6314.063	5213

SIMPLE LINEAR REGRESSION:

Here, I have used data on coffee production and coffee export of Brazil, Vietnam, Ethiopia and India to check whether these variables are related or not. Therefore, fit a regression equation.

1. BRAZIL

First, we plot a scatter plot of coffee export-coffee production.



From the scatter plot, we can observe that coffee export and coffee production are linearly related. Hence, a simple linear regression equation must be fitted.

The output is as follows:

```
Call:
lm(formula = exp_b ~ prd_b)

Residuals:
     Min       1Q   Median       3Q      Max
-4940.2 -1625.3    75.2  1174.4  5036.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.386e+03  3.069e+03   3.058  0.00677 **
prd_b       4.482e-01  6.332e-02   7.078 1.34e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2855 on 18 degrees of freedom
Multiple R-squared:  0.7357,    Adjusted R-squared:  0.721
F-statistic:  50.1 on 1 and 18 DF,  p-value: 1.339e-06
```

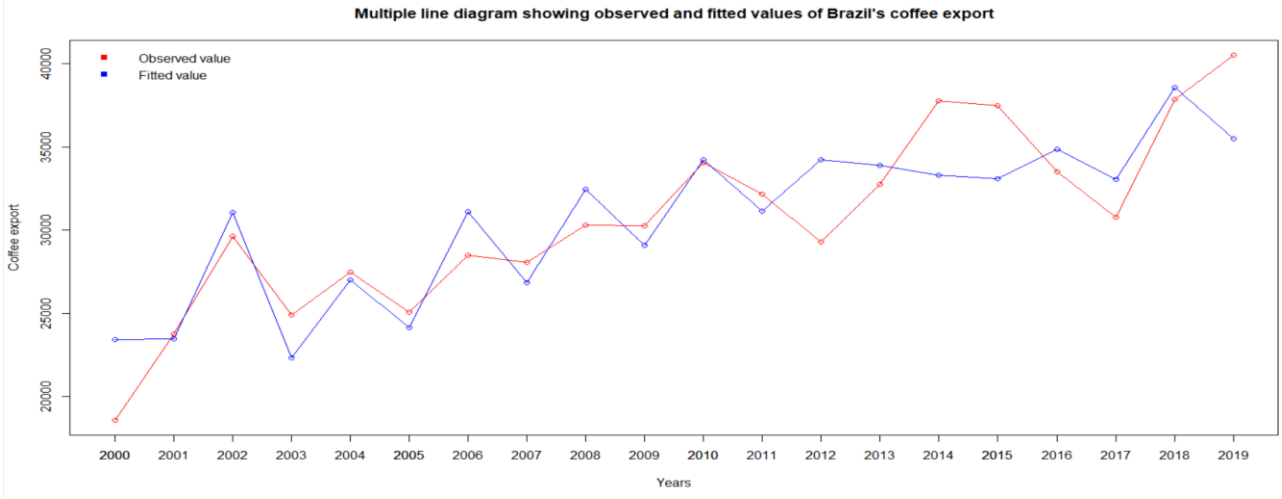
This indicates that our prediction model is: exp_b = 9.386e+03 + 4.482e-01*prd_b
The value of adjusted R-squared is 0.721 which implies 72.1% of the total variability can be explained by this regression equation.

Table showing the observed and predicted values of coffee export in Brazil

Year	Observed values	Predicted values
2000	18577	23418.33
2001	23767	23442.98
2002	29613	31056.33
2003	24909	22326.09
2004	27468	26990.82
2005	25078	24145.73
2006	28486	31092.18
2007	28044	26824.99
2008	30292	32463.19
2009	30255	29095.51
2010	34054	34227.70
2011	32149	31163.89
2012	29283	34223.22
2013	32752	33896.49
2014	37782	33276.20

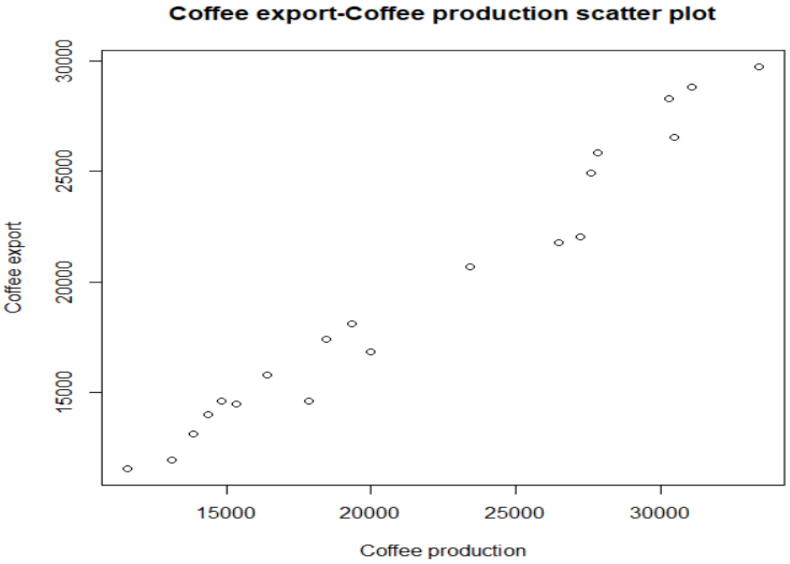
2015	37473	33081.68
2016	33491	34837.23
2017	30783	33022.97
2018	37870	38576.46
2019	40511	35475.00

Here, is a graph showing the observed and the fitted values of coffee export.



2. VIETNAM

Here, we plot a scatter plot of coffee export-coffee production.



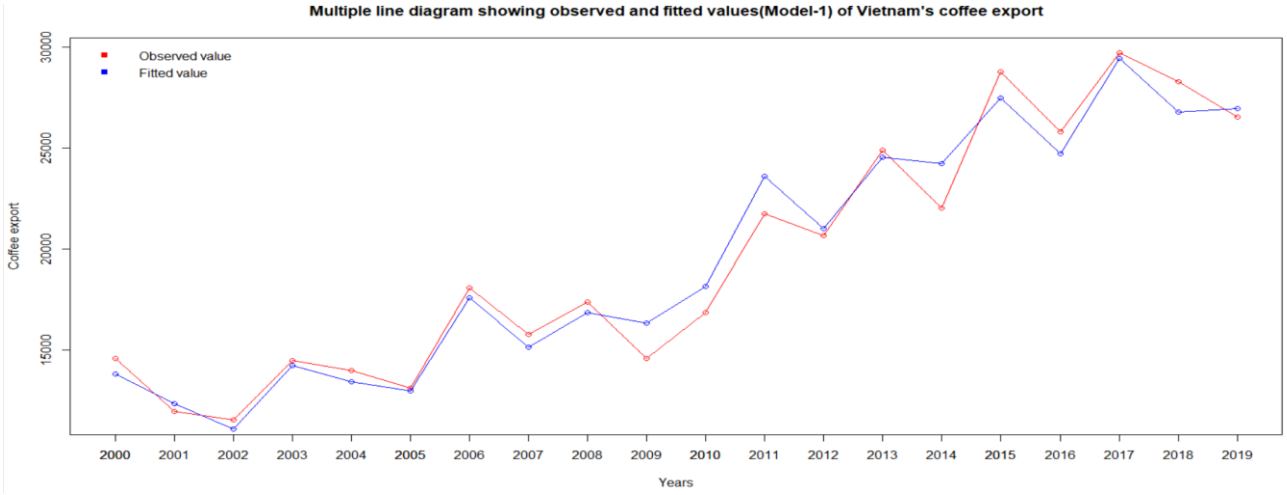
From the scatter plot, we can observe that coffee export and coffee production are related. To check whether production and export are linearly related, we fit both a linear and a polynomial regression and see which gives better result.

<u>Model-1:</u> Fitting a linear regression

The output is as follows:

```
Call:
lm(formula = exp_v ~ prd_v)

Residuals:
     Min       1Q   Median       3Q      Max
-2217.0   -403.1    306.9    582.1   1473.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.346e+03  7.977e+02   1.687    0.109
prd_v       8.409e-01  3.512e-02  23.943 4.22e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1081 on 18 degrees of freedom
Multiple R-squared:  0.9696,    Adjusted R-squared:  0.9679
F-statistic: 573.3 on 1 and 18 DF,  p-value: 4.223e-15
```

This indicates that our prediction model is: exp_v = 1.346e+03 + 8.409e-01*prd_v
The value of adjusted R-squared is 0.9679 which implies 96.79% of the total variability can be explained by this regression equation.
Here, is a graph showing the observed and the fitted values of coffee export.



Model-2: Fitting a second-degree polynomial regression
The output is as follows:

```
Call:
lm(formula = exp_v ~ poly(prd_v, 2, raw = TRUE))

Residuals:
     Min       1Q   Median       3Q      Max
-1952.0   -699.7    208.6    814.1   1291.5

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  6.446e+03  3.097e+03   2.082   0.0528 .
poly(prd_v, 2, raw = TRUE)1 3.347e-01  2.998e-01   1.116   0.2798
poly(prd_v, 2, raw = TRUE)2 1.135e-05  6.681e-06   1.699   0.1076
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1028 on 17 degrees of freedom
Multiple R-squared:   0.974,      Adjusted R-squared:   0.9709
F-statistic: 318.1 on 2 and 17 DF,  p-value: 3.394e-14
```

This indicates that our prediction model is:
exp_v=6.446e+03 + 3.347e-01*prd_v + 1.135e-05*(prd_v^2)
The value of adjusted R-squared is 0.9709 which implies 97.09% of the total variability can be explained by this regression equation.
Here, is a graph showing the observed and the fitted values of coffee export under Model-2.

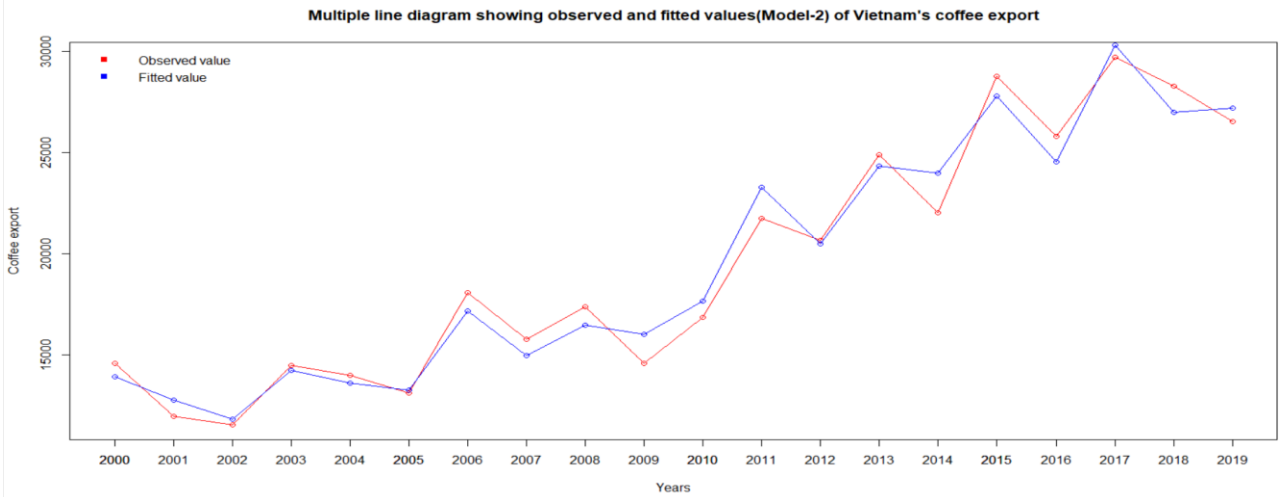


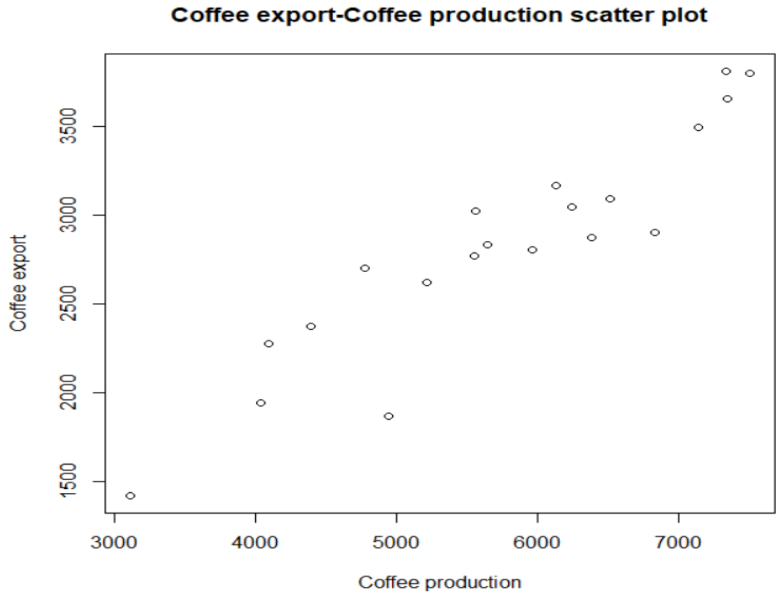
Table showing the observed and fitted values of coffee export in Vietnam

Year	Observed values	Fitted values of Model-1 (Adj R^2 = 0.9679)	Fitted values of Model-2 (Adj R^2 = 0.9709)
2000	14606	10587.61	13913.60
2001	11966	11530.78	12774.29
2002	11555	12473.95	11840.57
2003	14497	13417.12	14249.52
2004	13994	14360.28	13599.79
2005	13122	15303.45	13253.98
2006	18090	16246.62	17164.98
2007	15774	17189.78	14991.79
2008	17386	18132.95	16476.29
2009	14591	19076.12	16018.79
2010	16850	20019.28	17680.61
2011	21760	20962.45	23286.99
2012	20665	21905.62	20495.27
2013	24902	22848.78	24340.27
2014	22035	23791.95	23987.02
2015	28790	24735.12	27823.75
2016	25819	25678.28	24541.72
2017	29732	26621.45	30322.86
2018	28283	27564.62	26991.46
2019	26537	28507.79	27200.45

Hence, it can be concluded that model-2 is better than model-1 as the value of adjusted R-squared is more for model-2.

3. Ethiopia

Here, we plot a scatter plot of coffee export-coffee production.



From the scatter plot, we can observe that coffee export and coffee production are linearly related. Hence, a simple linear regression equation must be fitted.

The output is as follows:

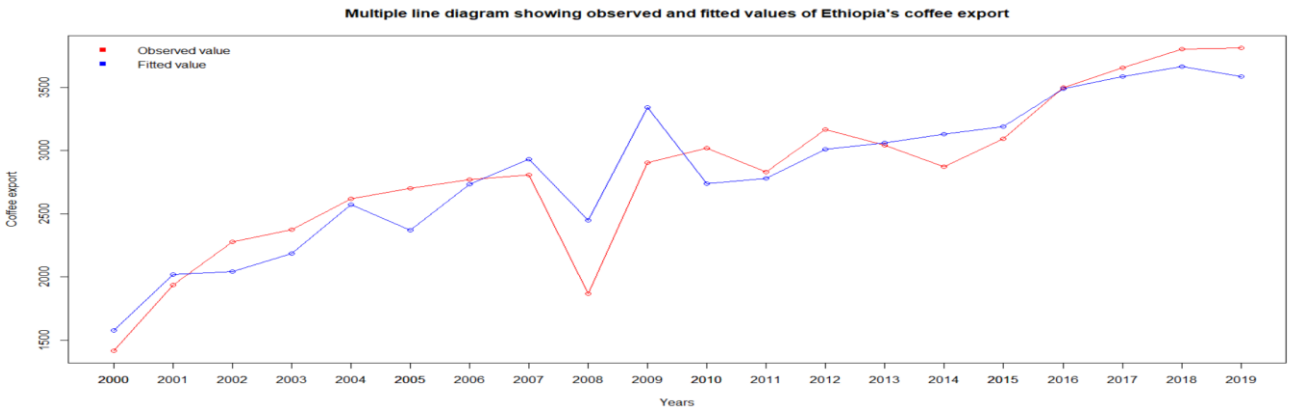
```
Call:
lm(formula = exp_e ~ prd_e)

Residuals:
    Min      1Q  Median      3Q     Max
-581.09 -106.62   40.46  163.76  333.57

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.90745  261.26603   0.386    0.704
prd_e         0.47448    0.04456  10.649 3.37e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 240.4 on 18 degrees of freedom
Multiple R-squared:  0.863,     Adjusted R-squared:  0.8554
F-statistic: 113.4 on 1 and 18 DF,  p-value: 3.367e-09
```

This indicates that our prediction model is: exp_e = 100.907 + 0.474*prd_e
The value of adjusted R-squared is 0.8554 which implies 85.54% of the total variability can be explained by this regression equation.

Table showing the observed and fitted values of coffee export in Ethiopia

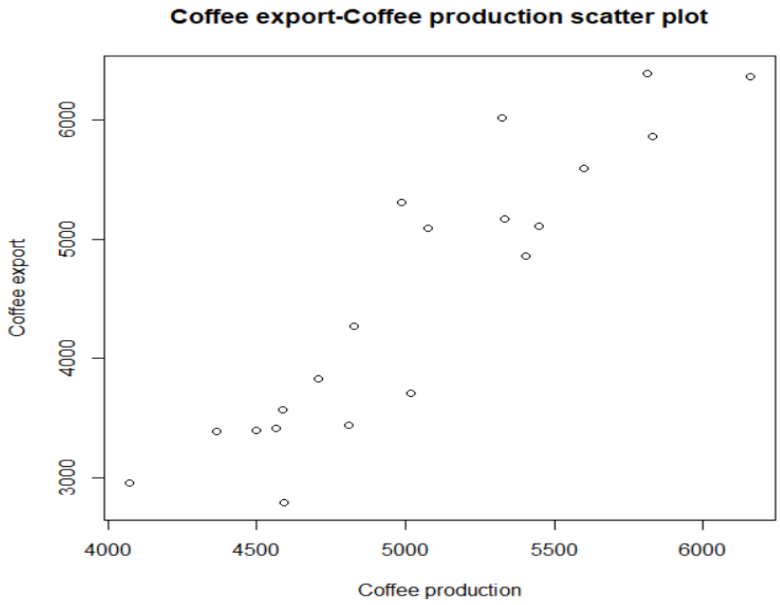
Year	Observed values	Predicted values
2000	1418	1578.901
2001	1939	2019.690
2002	2277	2043.413
2003	2374	2185.756
2004	2620	2574.352
2005	2702	2368.430
2006	2770	2734.725
2007	2806	2932.108
2008	1868	2449.091
2009	2904	3341.581
2010	3022	2738.996
2011	2832	2781.699
2012	3166	3010.396
2013	3044	3062.589
2014	2872	3129.490
2015	3092	3192.121
2016	3497	3490.092
2017	3654	3586.885
2018	3801	3664.699
2019	3812	3584.987

Here, is a graph showing the observed and the fitted values of coffee export.



4. INDIA

Here, we plot a scatter plot of coffee export-coffee production.



From the scatter plot, we can observe that coffee export and coffee production are linearly related. Hence, a simple linear regression equation must be fitted.

The output is as follows:

```
Call:
lm(formula = exp_i ~ prd_i)

Residuals:
    Min      1Q  Median      3Q     Max
-836.29 -226.34  -41.34  236.70  959.18

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5358.9188  1040.0979  -5.152 6.68e-05 ***
prd_i           1.9571     0.2047   9.559 1.78e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 492.3 on 18 degrees of freedom
Multiple R-squared:  0.8354,    Adjusted R-squared:  0.8263
F-statistic: 91.37 on 1 and 18 DF,  p-value: 1.78e-08
```

This indicates that our prediction model is: exp_i = -5358.9188 + 1.9571*prd_i

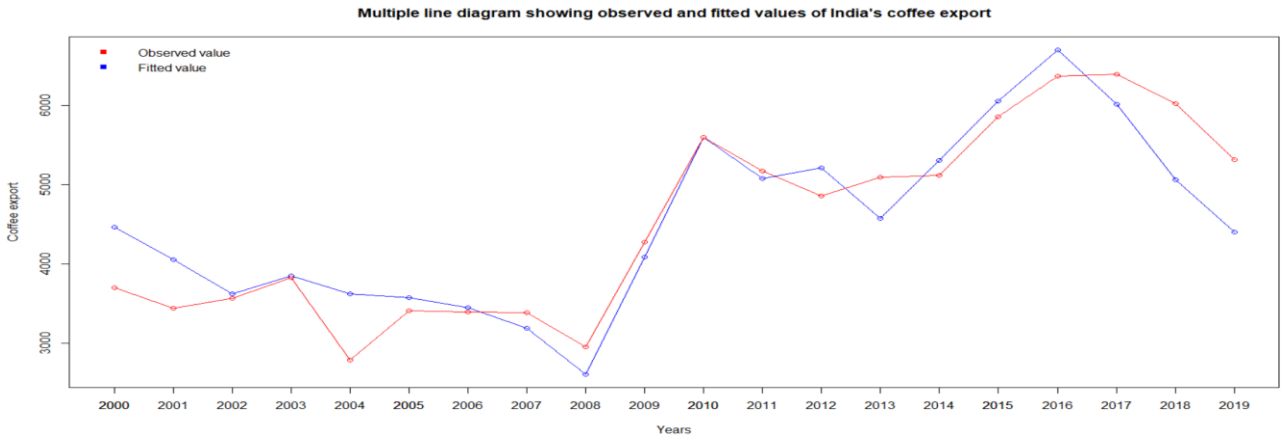
The value of adjusted R-squared is 0.8263 which implies 82.63% of the total variability can be explained by this regression equation.

Table showing the observed and fitted values of coffee export in Ethiopia

Year	Observed values	Predicted values
2000	3705	4465.896
2001	3441	4054.898
2002	3567	3620.414
2003	3826	3855.270
2004	2790	3626.285
2005	3410	3577.357
2006	3393	3448.186
2007	3389	3187.887
2008	2954	2610.533
2009	4274	4088.169
2010	5594	5601.034
2011	5172	5080.436

2012	4859	5215.478
2013	5095	4573.538
2014	5115	5307.464
2015	5861	6051.175
2016	6371	6698.986
2017	6395	6017.904
2018	6022	5062.822
2019	5314	4403.268

Here, is a graph showing the observed and the fitted values of coffee export.



Forecasting: Here, coffee export for the year 2020 is being predicted using the regression equations obtained for each of the four countries.

Table showing observed and predicted coffee export for the year 2020

COUNTRY	COFFEE PRODUCTION (DATA IN 1000 60KG BAGS)	PREDICTED COFFEE EXPORT	COFFEE EXPORT (DATA IN 1000 60KG BAGS)
BRAZIL (exp_b = 9.386e+03 +4.482e-01*prd_b)	69000	40310.49	44848
VIETNAM (exp_v=6.446e+03 + 3.347e-01*prd_v + 1.135e-05*(prd_v^2))	29000	25698.68	25625
ETHIOPIA (exp_e = 100.907 + 0.474*prd_e)	7375	3600.17	3443
INDIA (exp_i = -5358.9188 + 1.9571*prd_i)	5700	5796.747	5213

MULTIPLE LINEAR REGRESSION:

Here, at first, I have constructed a multiple linear regression model using all the variables except coffee production to predict the GDP of the four countries. This is because both coffee export and domestic consumption directly depend on coffee production.

1. BRAZIL

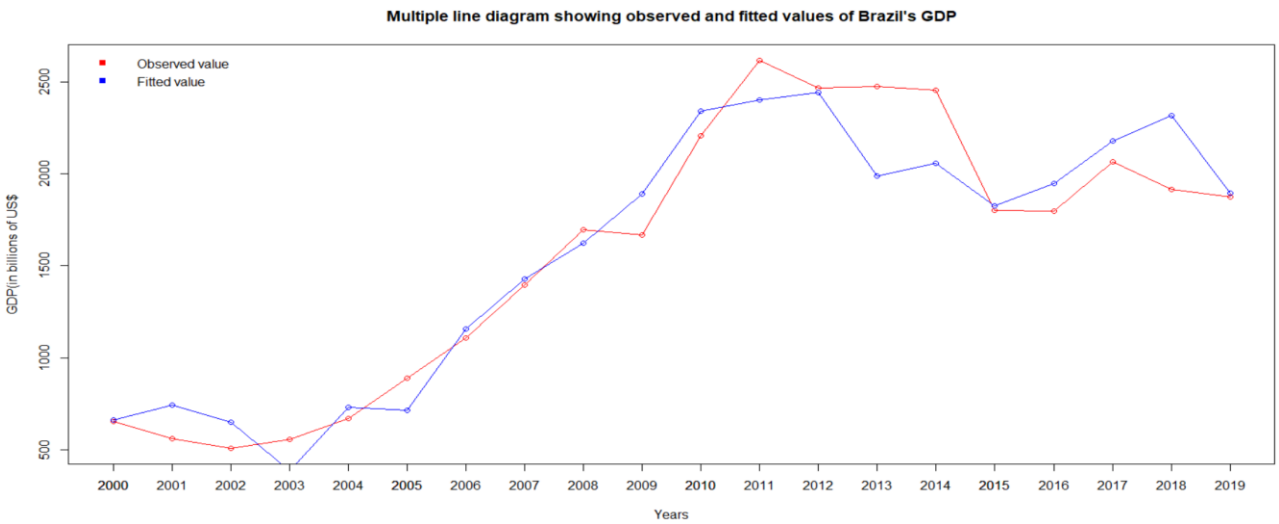
Model-1: The output is given below,

```
Call:
lm(formula = GDP_b ~ ppl_b + exp_b + dc_b)

Residuals:
    Min      1Q  Median      3Q     Max
-399.53 -135.84  -27.93   98.54  485.70

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.628e+04  5.325e+03   4.934 0.000149 ***
ppl_b       -2.198e-04  4.119e-05  -5.336 6.68e-05 ***
exp_b        4.443e-02  2.136e-02   2.080 0.053944 .
dc_b         9.260e-01  1.333e-01   6.944 3.31e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 228.1 on 16 degrees of freedom
Multiple R-squared:  0.9169,    Adjusted R-squared:  0.9013
F-statistic: 58.82 on 3 and 16 DF,  p-value: 7.329e-09
```

This indicates our prediction model is:
GDP_b = 2.628e+04 - 2.198e-04*ppl_b + 4.443e-02*exp_b + 9.26e-01*dc_b

Here, the intercept is positive which means that if all the values corresponding to independent variables would be 0, the GDP would have been positive.
The value of adjusted R-squared is 0.9013 which implies 90.13% of the total variability can be explained by this regression equation.

Here, is a graph showing the observed and the fitted values of GDP of Brazil.



From the graph, it is evident that the fitted values are substantially different from the observed values. Hence, the fit is not good inspite of the value of adjusted R-squared being high. Therefore we need to check for multicollinearity.
"Multicollinearity in regression analysis occurs when two or more predictor variables are highly correlated to each other". This means that they do not provide unique or independent information in the regression model and can cause problems while interpreting it.

Checking multicollinearity:

```
Call:
imcdiag(mod = fit)


All Individual Multicollinearity Diagnostics Result

          VIF    TOL       Wi        Fi Leamer    CVIF Klein   IND1   IND2
ppl_b 76.2845 0.0131 639.9180 1355.1205 0.1145 -7.0176     1 0.0015 1.0707
exp_b  4.8683 0.2054  32.8802   69.6287 0.4532 -0.4478     0 0.0242 0.8621
dc_b  61.4556 0.0163 513.8723 1088.2001 0.1276 -5.6534     1 0.0019 1.0673

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

exp_b , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.9169

* use method argument to check which regressors may be the reason of collinearity
===================================
```

From the above output, we see that the test shows there is multicollinearity in the model and either the variable 'ppl_b' or 'dc_b' is the root cause of multicollinearity.

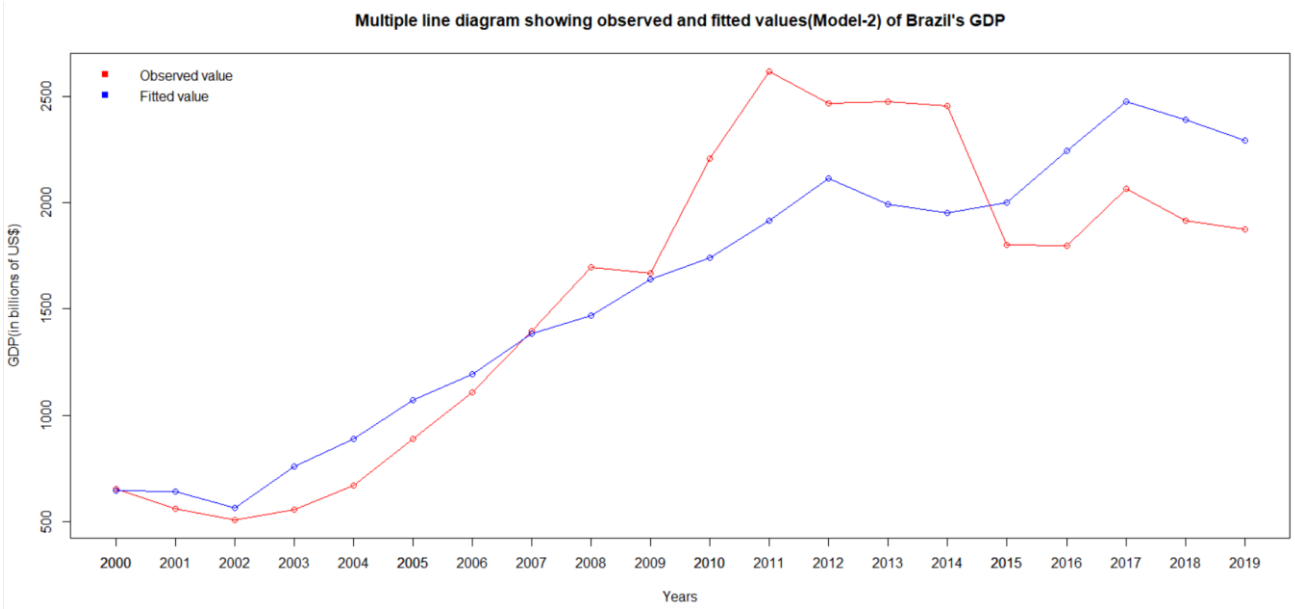
Model-2: Removing ppl_b from Model-1
The output is as follows:

```
Call:
lm(formula = GDP_b ~ exp_b + dc_b)

Residuals:
   Min     1Q Median     3Q    Max
-471.7 -207.5  -69.2  257.4  699.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.089e+03  5.177e+02  -4.036 0.000858 ***
exp_b       -1.898e-02  2.871e-02  -0.661 0.517353
dc_b         2.341e-01  5.045e-02   4.640 0.000234 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 368.9 on 17 degrees of freedom
Multiple R-squared:  0.7689,    Adjusted R-squared:  0.7417
F-statistic: 28.28 on 2 and 17 DF,  p-value: 3.91e-06
```

This indicates our prediction model is:
GDP_b = -2.089e+03 – 1.898e-02*exp_b + 2.341e-01*dc_b

The value of adjusted R-squared is 0.7417 which implies 74.17% of the total variability can be explained by this regression equation.

Here, is a graph showing the observed and the fitted values of GDP of Brazil under Model-2.



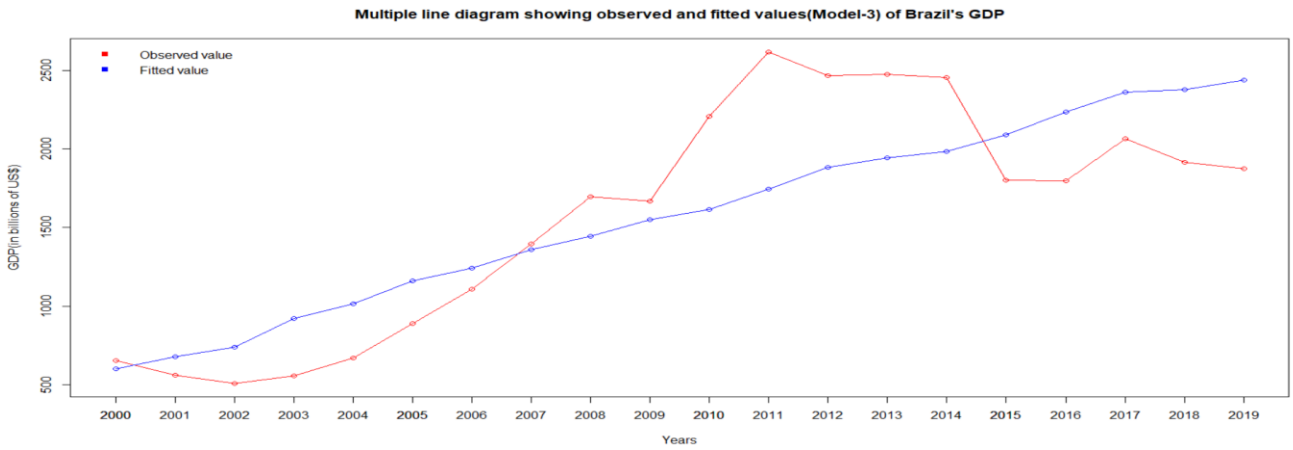
Model-3: Removing dc_b from Model-1
The output is as follows:

```
Call:
lm(formula = GDP_b ~ exp_b + ppl_b)

Residuals:
   Min     1Q Median     3Q    Max
-565.8 -309.3 -125.1  306.4  872.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.433e+03  2.687e+03  -3.510  0.00268 **
exp_b       -1.163e-02  3.843e-02  -0.303  0.76576
ppl_b        5.828e-05  1.872e-05   3.113  0.00633 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 443.2 on 17 degrees of freedom
Multiple R-squared:  0.6663,    Adjusted R-squared:  0.6271
F-statistic: 16.98 on 2 and 17 DF,  p-value: 8.873e-05
```

This indicates our prediction model is:
GDP_b = -9.433e+03 – 1.163e-02*exp_b + 5.828e-05*ppl_b

The value of adjusted R-squared is 0.6271 which implies 62.71% of the total variability can be explained by this regression equation.

Here, is a graph showing the observed and the fitted values of GDP of Brazil under Model-3.



On comparing model-2 and model-3, we can say that model-2 is better than model-3. Model-2 has a greater adjusted R-squared value compared to model-3. Additionally, model-2's graph shows better-fitted values than model-3's.

2. VIETNAM

Model-1: The output is given below,

```
Call:
lm(formula = GDP_v ~ ppl_v + exp_v + dc_v)

Residuals:
     Min       1Q   Median       3Q      Max
-17.655   -7.169   -2.811    5.127   33.309

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.878e+02  3.738e+02  -0.502  0.62221
ppl_v        1.850e-06  4.797e-06   0.386  0.70488
exp_v        6.346e-04  1.592e-03   0.399  0.69541
dc_v         1.113e-01  3.491e-02   3.188  0.00572 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.82 on 16 degrees of freedom
Multiple R-squared:  0.9847,     Adjusted R-squared:  0.9818
F-statistic: 342.7 on 3 and 16 DF,  p-value: 1.009e-14
```

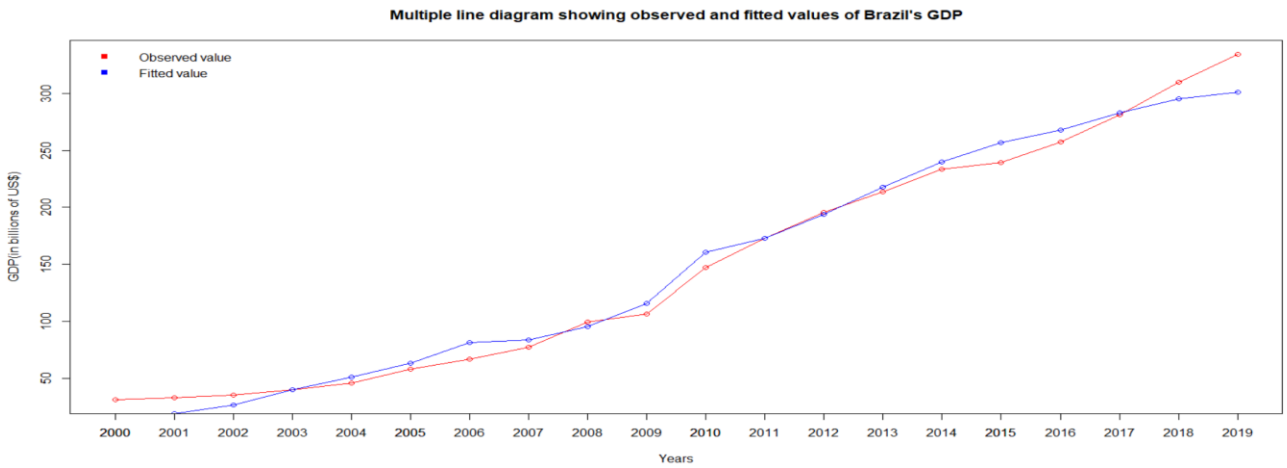
This indicates our prediction model is:

GDP_v = -1.878e+02 – 1.850e-06*ppl_v + 6.346e-04*exp_v + 1.113e-01*dc_v

Here, the intercept is negative which means that if all the values corresponding to independent variables would be 0, the GDP would have been negative.

The value of adjusted R-squared is 0.9818 which implies 98.18% of the total variability can be explained by this regression equation.

Here, is a graph showing the observed and the fitted values of GDP of Vietnam.



From the graph, it is evident that the fitted values are very close to the observed values.
Still, we would check for multicollinearity.

Checking multicollinearity:

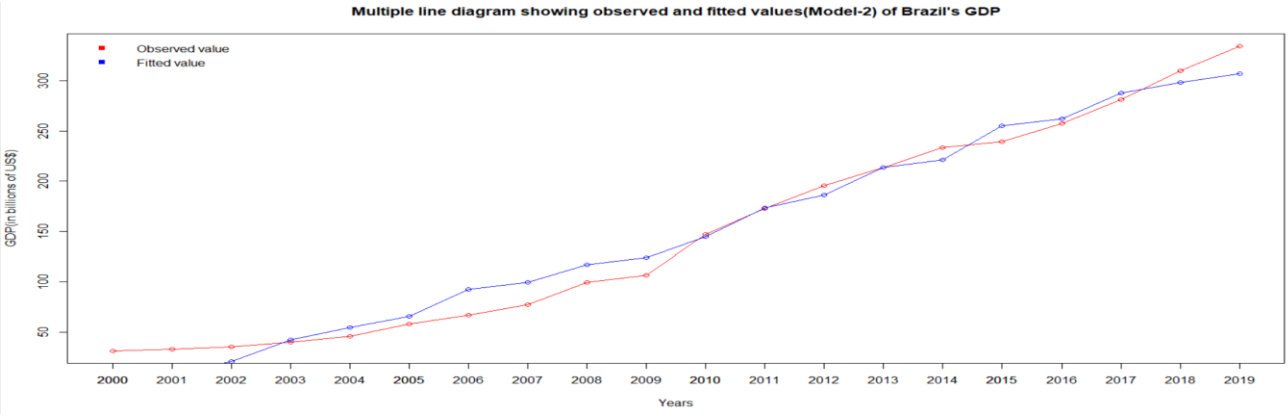
```
Call:
imcdiag(mod = fit)


All Individual Multicollinearity Diagnostics Result

          VIF    TOL       Wi        Fi Leamer    CVIF Klein   IND1   IND2
ppl_v 63.6792 0.0157 532.7731 1128.2253 0.1253 -0.5308     0 0.0018 1.0317
exp_v  9.1639 0.1091  69.3932  146.9504 0.3303 -0.0764     0 0.0128 0.9338
dc_v  76.4199 0.0131 641.0693 1357.5585 0.1144 -0.6370     1 0.0015 1.0345

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

ppl_v , exp_v , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.9847

* use method argument to check which regressors may be the reason of collinearity
===================================
```

From the above output, we see that the test shows there is multicollinearity in the model and the variable 'dc_v' is the root cause of multicollinearity.

Model-2: Removing dc_v from Model-1
The output is as follows:

```
Call:
lm(formula = GDP_v ~ exp_v + ppl_v)

Residuals:
    Min      1Q  Median      3Q     Max
-25.898 -10.727  -1.867  12.114  28.742

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.318e+03  1.467e+02  -8.986 7.25e-08 ***
exp_v        2.759e-03  1.793e-03   1.539    0.142
ppl_v        1.621e-05  2.050e-06   7.905 4.29e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.15 on 17 degrees of freedom
Multiple R-squared:  0.9749,    Adjusted R-squared:  0.972
F-statistic: 330.7 on 2 and 17 DF,  p-value: 2.462e-14
```

This indicates our prediction model is:
GDP_v = -1.318e+03 + 2.759e-03*exp_v + 1.621e-05*ppl_v
The value of adjusted R-squared is 0.9720 which implies 97.20% of the total variability can be explained by this regression equation.
Here, is a graph showing the observed and the fitted values of GDP of Vietnam under Model-2.



3. ETHIOPIA

Model-1: The output is given below,

```
Call:
lm(formula = GDP_e ~ ppl_e + exp_e + dc_e)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9489 -1.4902 -0.6305  2.4047  7.1910

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.474e+02  5.441e+00 -27.096 8.49e-15 ***
ppl_e        3.710e-06  2.690e-07  13.796 2.66e-10 ***
exp_e       -4.820e-03  2.716e-03  -1.775    0.095 .
dc_e        -4.350e-02  6.604e-03  -6.588 6.25e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.503 on 16 degrees of freedom
Multiple R-squared:  0.9877,    Adjusted R-squared:  0.9854
F-statistic: 427.9 on 3 and 16 DF,  p-value: 1.752e-15
```

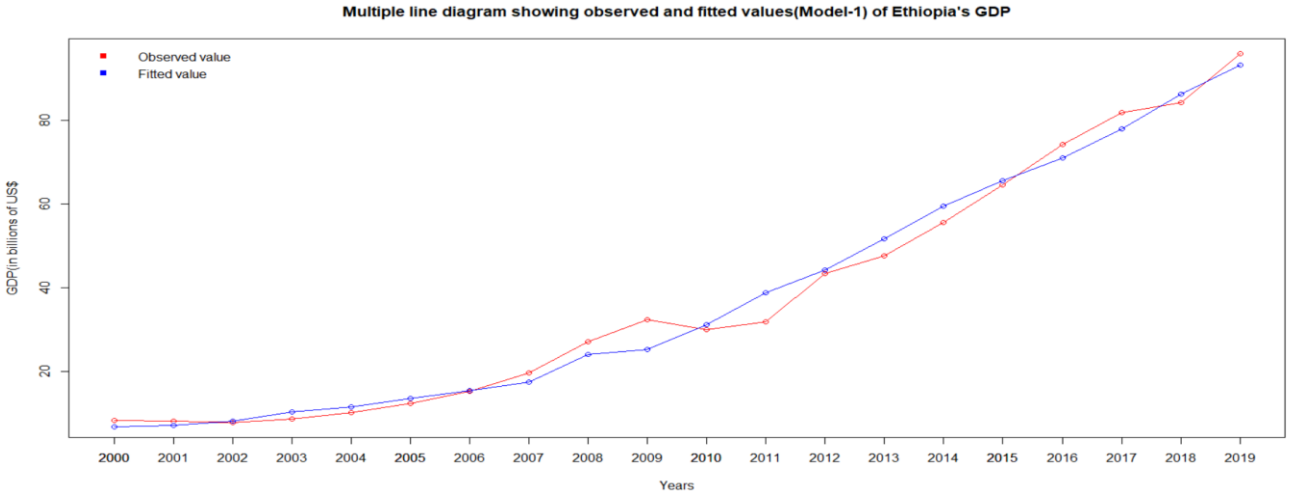
This indicates our prediction model is:

GDP_e = -1.474e+02 + 3.710e-06*ppl_e – 4.820e-03*exp_e – 4.350e-02*dc_v

Here, the intercept is negative which means that if all the values corresponding to independent variables would be 0, the GDP would have been negative.

The value of adjusted R-squared is 0.9854 which implies 98.54% of the total variability can be explained by this regression equation.

Here, is a graph showing the observed and the fitted values of GDP of Ethiopia.



From the graph, it is evident that the fitted values are close to the observed values. Still, we would check for multicollinearity.

Checking multicollinearity:

```
Call:
imcdiag(mod = fit)


All Individual Multicollinearity Diagnostics Result

          VIF    TOL       Wi       Fi Leamer    CVIF Klein   IND1   IND2
ppl_e 24.1231 0.0415 196.5462 416.2154 0.2036 -0.1999     0 0.0049 1.0685
exp_e  4.5663 0.2190  30.3132  64.1926 0.4680 -0.0378     0 0.0258 0.8706
dc_e  20.7019 0.0483 167.4664 354.6348 0.2198 -0.1716     0 0.0057 1.0609

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

exp_e , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.9877

* use method argument to check which regressors may be the reason of collinearity
===================================
```

From the above output, we see that the test shows there is no multicollinearity in the model. Hence, there is no need to remove any variable from model-1.

4. INDIA

Model-1: The output is given below,

```
Call:
lm(formula = GDP_i ~ ppl_i + exp_i + dc_i)

Residuals:
    Min      1Q  Median      3Q     Max
-117.34  -53.05  -11.68   42.80  211.98

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.128e+03  4.847e+02 -16.767 1.42e-11 ***
ppl_i        9.466e-06  1.006e-06   9.410 6.36e-08 ***
exp_i        5.058e-02  3.721e-02   1.359  0.19298
dc_i        -1.704e+00  5.751e-01  -2.964  0.00914 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 93.12 on 16 degrees of freedom
Multiple R-squared:  0.9881,    Adjusted R-squared:  0.9859
F-statistic: 442.9 on 3 and 16 DF,  p-value: 1.335e-15
```

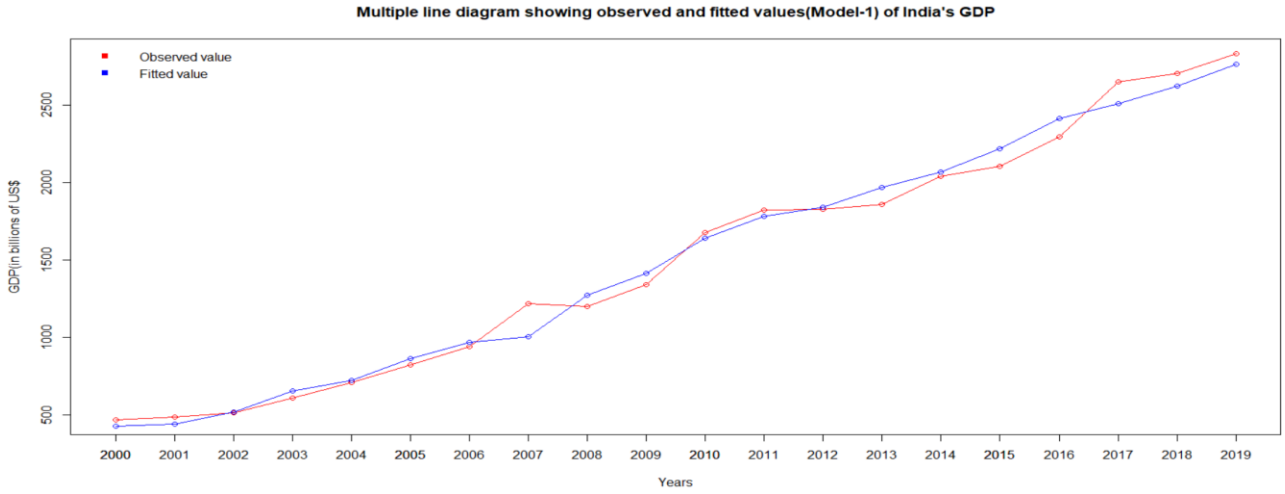
This indicates our prediction model is:

GDP_i = -8.128e+03 + 9.466e-06*ppl_i + 5.058e-02*exp_i – 1.704e+00*dc_i

Here, the intercept is negative which means that if all the values corresponding to independent variables would be 0, the GDP would have been negative.

The value of adjusted R-squared is 0.9859 which implies 98.59% of the total variability can be explained by this regression equation.

Here, is a graph showing the observed and the fitted values of GDP of India.



From the graph, it is evident that the fitted values are close to the observed values. Still, we would check for multicollinearity.

Checking multicollinearity:

```
Call:
imcdiag(mod = fit)


All Individual Multicollinearity Diagnostics Result

          VIF     TOL       Wi       Fi Leamer    CVIF Klein   IND1   IND2
ppl_i 22.5267 0.0444 182.9769 387.4805 0.2107 -0.1693     0 0.0052 1.0816
exp_i  4.2341 0.2362  27.4902  58.2146 0.4860 -0.0318     0 0.0278 0.8646
dc_i  14.4986 0.0690 114.7378 242.9742 0.2626 -0.1090     0 0.0081 1.0538

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

exp_i , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.9881

* use method argument to check which regressors may be the reason of collinearity
===================================
```

From the above output, we see that the test shows there is no multicollinearity in the model. Hence, there is no need to remove any variable from model-1.

Forecasting: Here, GDP for the year 2020 is being predicted using the multiple linear regression equations obtained for each of the four countries.

Table showing observed and predicted GDP for the year 2020

COUNTRY	MULTIPLE REGRESSION EQUATION (AFTER REMOVING MULTICOLLINEARITY)	PREDICTED GDP	GDP (IN BILLIONS OF US$)
BRAZIL	GDP_b = -2.089e+03 – 1.898e-02*exp_b + 2.341e-01*dc_b (Adjusted R-squared: 0.7417)	2279.601	1448.56
VIETNAM	GDP_v = -1.318e+03 + 2.759e-03*exp_v + 1.621e-05*ppl_v (Adjusted R-squared: 0.972)	318.8165	346.62
ETHIOPIA	GDP_e = -1.474e+02 + 3.710e-06*ppl_e – 4.820e-03*exp_e – 4.350e-02*dc_v (Adjusted R-squared: 0.9854)	101.1728	107.66
INDIA	GDP_i = -8.128e+03 + 9.466e-06*ppl_i + 5.058e-02*exp_i – 1.704e+00*dc_i (Adjusted R-squared: 0.9859)	2627.552	2667.69

PRINCIPAL COMPONENT ANALYSIS:

Principal component(PC) is the linear combination of the variables after standardizing. The variance-covariance matrix of the standardized variables is obtained and then its eigen values are computed. The elements in the eigen vector corresponding to the largest eigen value is the coefficients of first PC. The elements in the eigen vector corresponding to the second largest eigen value is the coefficients of second PC and so on.

1. BRAZIL

```
Standard deviations (1, .., p=5):
[1] 2.09003739 0.58249991 0.41364226 0.34592177 0.04093605

Rotation (n x k) = (5 x 5):
            PC1          PC2         PC3        PC4         PC5
prd_b 0.4429480  0.312141156 -0.70588581  0.4535688 -0.04863716
dc_b  0.4678680 -0.175515887  0.35707775  0.2979544  0.73076148
exp_b 0.4361418  0.575928486  0.08514392 -0.6796250  0.09458913
GDP_b 0.4225495 -0.734861565 -0.31355729 -0.4085620 -0.12723721
ppl_b 0.4648967  0.006850679  0.51831751  0.2769214 -0.66218250
```

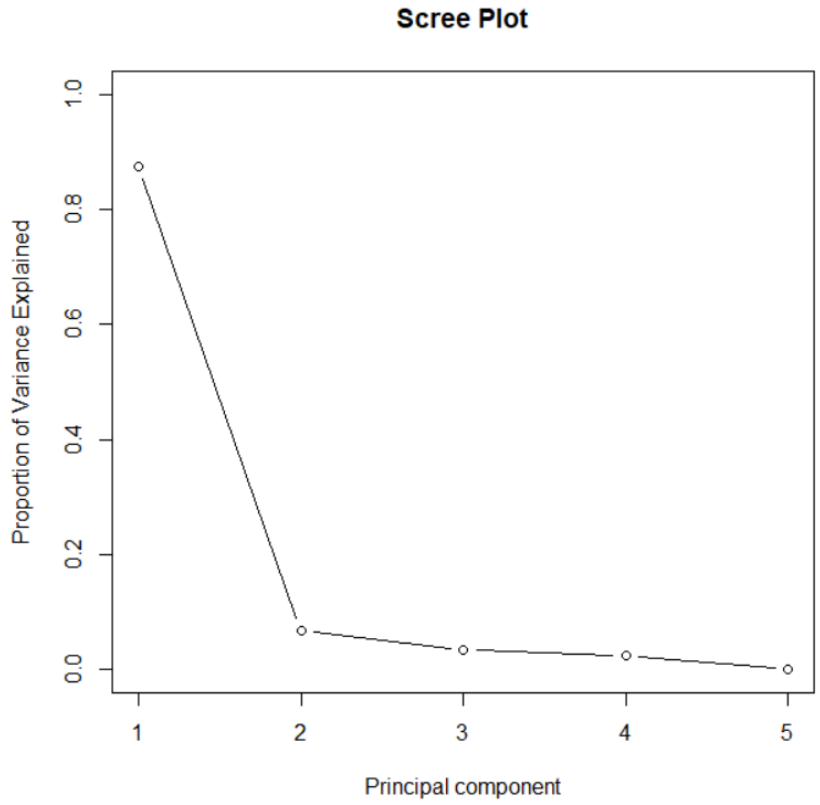
Here, we can see the coefficients of the five variables in each PCs.

The first PC is given as,

PC1 = 0.4429480*prd_b + 0.4678680*dc_b + 0.4361418*exp_b + 0.4225495*GDP_b + 0.4648967*ppl_b.
Similarly, the four PCs can be obtained.

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5
Standard deviation     2.0900 0.58250 0.41364 0.34592 0.04094
Proportion of Variance 0.8737 0.06786 0.03422 0.02393 0.00034
Cumulative Proportion  0.8737 0.94151 0.97573 0.99966 1.00000
```

Here, we can say that instead of using five different variables, we can use first PC, which explains 87.37% of the total variability. As a result, it is easier to work with multivariate data since it can be effectively reduced to univariate data.



From the scree plot, we can see there is a sharp bend at the point 2, so we can use both first and second PCs or only the first PC will also suffice.

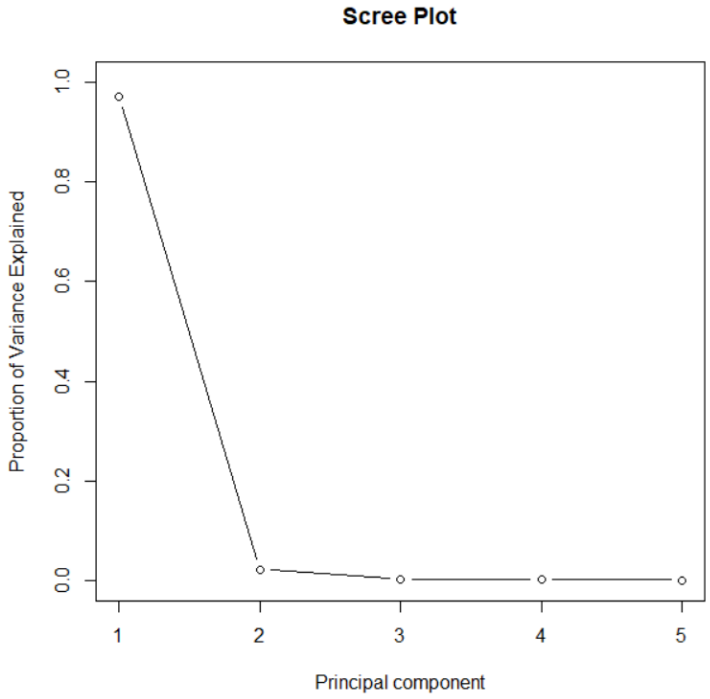
2. VIETNAM

```
Standard deviations (1, .., p=5):
[1] 2.20380202 0.33301965 0.11857678 0.11661139 0.06852679

Rotation (n x k) = (5 x 5):
            PC1         PC2         PC3         PC4         PC5
prd_v 0.4472184  0.4332465 -0.13981897  0.7308923  0.2419510
dc_v  0.4504938 -0.3118494 -0.03108677  0.1739357 -0.8176706
exp_v 0.4418220  0.6502871  0.05596301 -0.6005868 -0.1344757
GDP_v 0.4489133 -0.3386009  0.75313662 -0.0420782  0.3388818
ppl_v 0.4475726 -0.4213354 -0.63963868 -0.2703104  0.3740988
```

Here, we can see the coefficients of the five variables in each PCs.
The first PC is given as,
PC1 = 0.4472184*prd_v + 0.4504938*dc_v + 0.4418220*exp_v + 0.4489133*GDP_v + 0.4475726*ppl_v

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5
Standard deviation     2.2038 0.33302 0.11858 0.11661 0.06853
Proportion of Variance 0.9714 0.02218 0.00281 0.00272 0.00094
Cumulative Proportion  0.9714 0.99353 0.99634 0.99906 1.00000
```

Here, we can say that instead of using five different variables, we can use first PC, which explains 97.14% of the total variability.



From the scree plot, we can see there is a sharp bend at the point 2, so we can use both first and second PCs or only the first PC will also suffice.

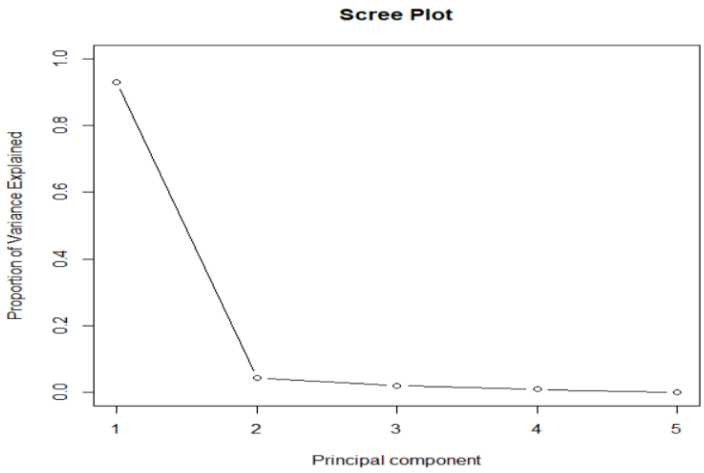
3. ETHIOPIA

```
Standard deviations (1, .., p=5):
[1] 2.1554226 0.4580089 0.3143296 0.2104838 0.0357049

Rotation (n x k) = (5 x 5):
            PC1        PC2         PC3        PC4        PC5
prd_e 0.4488783  0.3584386 -0.37896114 -0.7164838 -0.1143213
dc_e  0.4513379 -0.1938958 -0.61140999  0.4392441  0.4381107
exp_e 0.4341723  0.6873525  0.44045485  0.3662514  0.1044054
GDP_e 0.4431116 -0.5102127  0.53584094 -0.3232077  0.3895463
ppl_e 0.4582018 -0.3180498  0.03795121  0.2347600 -0.7951999
```

Here, we can see the coefficients of the five variables in each PCs.
The first PC is given as,
PC1 = 0.4488783*prd_e + 0.4513379*dc_e + 0.4341723*exp_e + 0.4431116*GDP_e + 0.4582018*ppl_e

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5
Standard deviation     2.1554 0.45801 0.31433 0.21048 0.03570
Proportion of Variance 0.9292 0.04195 0.01976 0.00886 0.00025
Cumulative Proportion  0.9292 0.97112 0.99088 0.99975 1.00000
```

Here, we can say that instead of using five different variables, we can use first PC, which explains 92.92% of the total variability.



From the scree plot, we can see there is a sharp bend at the point 2, so we can use both first and second PCs or only the first PC will also suffice.

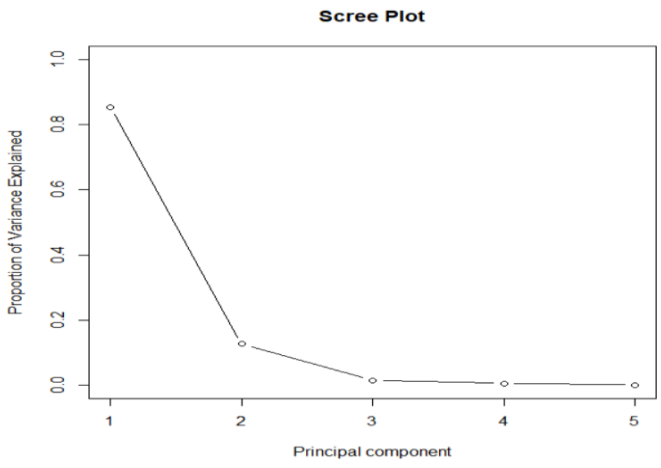
4. INDIA

```
Standard deviations (1, .., p=5):
[1] 2.06435401 0.79406470 0.27706997 0.16474044 0.06321867

Rotation (n x k) = (5 x 5):
            PC1        PC2        PC3        PC4         PC5
prd_i 0.3925610  0.7215313 -0.3547399 -0.4436786 -0.05096446
dc_i  0.4403476 -0.4624094 -0.6967272  0.2222196 -0.23971074
exp_i 0.4591499  0.3568519  0.2660648  0.7635577  0.08959464
GDP_i 0.4706332 -0.2158931  0.5452940 -0.3034498 -0.58520698
ppl_i 0.4686956 -0.3026830  0.1435091 -0.2804738  0.76775450
```

Here, we can see the coefficients of the five variables in each PCs.
The first PC is given as,
PC1 = 0.3925610*prd_i + 0.4403476*dc_i + 0.4591499*exp_i + 0.4706332*GDP_i + 0.4686956*ppl_i

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     2.0644 0.7941 0.27707 0.16474 0.06322
Proportion of Variance 0.8523 0.1261 0.01535 0.00543 0.00080
Cumulative Proportion  0.8523 0.9784 0.99377 0.99920 1.00000
```

Here, we can say that instead of using five different variables, we can use first PC, which explains 85.23% of the total variability.



From the scree plot, we can see there is a sharp bend at the point 2, so we can use both first and second PCs or only the first PC will also suffice.

CONCLUSION

I would like to end the project by recapitulating the highlights of the most significant prediction models.

- For each country, a trend equation on coffee production has been fitted. The best one is for Vietnam which explains 89.11% of the total variability.
- A trend equation for coffee consumed domestically has been fitted for each country. The best one is for Vietnam which explains 97.75% of the total variability. Moreover, the trend equation for Brazil, Ethiopia and India explains 97.21%, 97.13% and 89.47% of the total variability respectively.
- For each country, an export trend equation for coffee has been fitted. The best one is for Ethiopia which explains 85.54% of the total variability.
- A simple linear regression equation has been fitted for Brazil, Ethiopia, and India using coffee export as the dependent variable and coffee production as the independent variable. Explaining 85.54% of the total variability, the best one is for Ethiopia.
- A polynomial regression equation has been fitted for Vietnam using coffee export as the dependent variable and coffee production as the independent variable. Here, the regression equation explains 97.09% of the total variability.
- A multiple linear regression equation(MLR) has been fitted for each country using GDP as the independent variable and coffee export, coffee consumed domestically and population as the dependent variables.

  After checking for multicollinearity, we observe the following:
  - In case of Brazil, multicollinearity arises due to population, which is why it has been removed. Therefore, GDP depends on coffee export and coffee consumed domestically. Thus the new MLR equation obtained explains 74.17% of the total variability.
  - For, Vietnam, multicollinearity arises due to coffee consumed domestically, which is thus removed. Hence, GDP depends on coffee export and population. Thus the new MLR equation obtained explains 97.20% of the total variability.
  - For Ethiopia, multicollinearity does not arise. Hence, the original MLR equation is kept intact and it explains 98.54% of the total variability.
  - For India, multicollinearity does not arise. Hence, the original MLR equation is kept intact and it explains 98.59% of the total variability.

Note that, a prediction model is considered superior to the rest if it explains a greater percent of the total variability of the dependent variables.

REFERENCE

- https://www.ecotactbags.com/blog/top-10-coffee-producing-countries-in-2022
- https://www.worldstopexports.com/coffee-exports-country/
- https://www.projectpro.io/recipes/check-multicollinearity-r
- https://datascienceplus.com/multicollinearity-in-r/
- https://www.britannica.com/topic/coffee-production
- https://www.turing.com/kb/guide-to-principal-component-analysis
- https://www.geeksforgeeks.org/principal-component-analysis-with-r-programming/

APPENDIX

- SECONDARY COFFEE DATASET FROM INTERNATIONAL COFFEE ORGANISATION(ICO): https://www.ico.org/new_historical.asp

- SECONDARY GDP DATA FROM https://www.macrotrends.net/countries/BRA/brazil/gdp-gross-domestic-product

- SECONDARY POPULATION DATA FROM https://population.un.org/dataportal/data/indicators/49/locations/76,231,704,876,732,887,894,716,356/start/1999/end/2023/table/pivotbylocation

The first 5 data points for the 4 datasets are as follows:

1. BRAZIL

Year	prd_b	dc_b	exp_b	GDP_b	ppl_b
2000	31310	13200	18577	655.45	175873720
2001	31365	13590	23767	559.99	178211882
2002	48352	13750	29613	509.79	180476686
2003	28873	14200	24909	558.23	182629278
2004	39281	14946	27468	669.29	184722043
2005	32933	15538	25078	891.63	186797334

2. VIETNAM

Year	prd_v	dc_v	exp_v	GDP_v	ppl_v
2000	14841	402	14606	31.17	79001142
2001	13093	461	11966	32.69	79817777
2002	11574	519	11555	35.06	80642308
2003	15337	607	14497	39.55	81475825
2004	14370	696	13994	45.43	82311227
2005	13842	800	13122	57.63	83142095

3. ETHIOPIA

Year	prd_e	dc_e	exp_e	GDP_e	ppl_e
2000	3115	2014	1418	8.24	67031867
2001	4044	2121	1939	8.23	69018933
2002	4094	2234	2277	7.85	71073215
2003	4394	2353	2374	8.62	73168839
2004	5213	2478	2620	10.13	75301026
2005	4779	2609	2702	12.4	77469941

4. INDIA

Year	prd_i	dc_i	exp_i	GDP_i	ppl_i
2000	5020	975	3705	468.39	1059633675
2001	4810	1067	3441	485.44	1078970907
2002	4588	1133	3567	514.94	1098313039
2003	4708	1167	3826	607.7	1117415124
2004	4591	1202	2790	709.15	1136264583
2005	4566	1238	3410	820.38	1154638713

- Code for the entire project is as follows:
  (Please, note that only R programming was used here)

```
#To plot time-series graphs

plot(t, prd_b, main="Line diagram showing coffee production of Brazil over the years 2000-
2019", xlab="Years", ylab="Production(data in thousand 60kg bags)",type="o")

#To plot multiple bar diagram showing coffee production, domestic consumption and export

barplot(height=rbind(production,consumption,export),
      beside = TRUE,names.arg=t,
      main="Multiple bar diagram showing coffee production, consumption and export of Brazil
over the years 2000-2019",
      xlab="Year",ylab="Data in thousand 60kg bags",
      col=c("yellow","purple","cyan"),
      legend.text = c("Production" , "Domestic consumption" , "Export"),
      args.legend = list(x = "topleft"))

# To plot component bar diagram showing coffee domestic consumption and export

d<-rbind(d1[1,],d1[6,],d1[11,],d1[16,],d1[20,])  #d1 contains the dataset for BRAZIL
t1<-d[,1]     #t1 contains year
barplot(height=rbind(d$dc_b,d$exp_b),
      names.arg=t1,
      main="Component bar diagram showing coffee consumption and export of Brazil over the
years 2000-2019",
      xlab="Year",ylab="Data in thousand 60kg bags",
      col=c("lavender","lightblue"),
      legend.text = c("Domestic consumption" , "Export"),
      args.legend = list(x = "topleft"))

#fitting trend equation

fit<-lm(exp_b~t)
plot(t, exp_b, main="Graph showing Brazil's coffee export and fitted trend line",
xlab="Years", ylab="Coffee export", type="o")
abline(fit, col='red')
legend("topleft", legend = c("Observed value", "Trend line"),
      pch=15,bty="n" , col=c("black", "red"))
axis(1,d1$Year)

#To plot Coffee export-Coffee production scatter plot
plot(prd_b, exp_b, main="Coffee export-Coffee production scatter plot", xlab="Coffee
production", ylab="Coffee export")
```

```
#Fitting linear regression

fit_sr<-lm(exp_b~prd_b)
summary(fit_sr)

#To plot multiple line diagram showing observed and fitted values

p<-predict(fit_sr, newdata = data.frame(prd_b=prd_b))  #storing the fitted values
plot(t, p, main="Multiple line diagram showing observed and fitted values of India's coffee
export", xlab="Years", ylab="Coffee export", col="blue", type="o")
lines(t, exp_b, col="red", type = "o")
legend("topleft", legend = c("Observed value","Fitted value"),
       pch=15,bty="n" , col=c("red","blue"))
axis(1,t)

#Fitting second order polynomial regression
fit_pr<-lm(exp_v~poly(prd_v,2,raw=TRUE))

#Fitting multiple linear regressionfit_mlr<-lm(GDP_v~ppl_v+exp_v+dc_v)
summary(fit_mlr)

#Checking multicollinearity
library(mctest)
imcdiag(fit_mlr)

#Finding principal components using PCA
ds<-d1[,2:6]   #ds contain a subset of the dataset
head(ds)
d_nm<-scale(ds)   #standardizing ds
head(d_nm)
dpca<-prcomp(d_nm)       #to find PCs
dpca   #shows the PCs
summary(dpca)  #shows variance and cumulative proportion of the PCs

#To plot scree plot
dpca$sdev    # Compute standard deviation
dpca.var <- dpca$sdev ^ 2    # Compute variance
propve <- dpca.var / sum(dpca.var)    # Proportion of variance for a scree plot
plot(propve, xlab = "Principal component",
          ylab = "Proportion of Variance Explained",
          ylim = c(0, 1), type = "b",
          main = "Scree Plot")
```