Introduction - Motivation

There is a debate among economists and environmentalists that economic development comes at the cost of deforestation. Over the last 10000 years, we have lost one-third of the world's forest cover, half of which happened in the last 100 years. However, not all kind of economic advancements may be linked with deforestation. [3],[13],[16].

Problem Definition

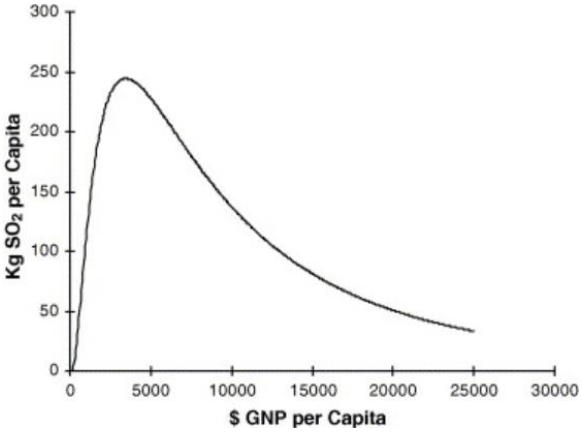
The objective is to determine what correlations exist between deforestation and socio-economic factors like per capita GDP, inflation, infant mortality etc. By bringing multiple such factors together, we can do meaningful analysis of the net effect of these factors on deforestation [12] across countries.

Our aim is to orchestrate a data pipeline, train analytical models and build interactive visualizations that would help users garner critical information on this important topic. We will make use of data from reliable sources like worldbank.org, undp.org and unicef.org for this project.

Survey

Research has linked deforestation [11] with trade-liberalization [1], unemployment [2], pollution and carbon emission [5], infant mortality [6], colonization [18], population [7],[14], GDP [8] and corruption [15],[19].

Most of the current studies are based on a single factor and have not been inclusive of other factors. For example - the Environmental Kuznets curve (EKC) hypothesizes the relationship between various indicators of environmental degradation like deforestation and income per capita [9].



Even though the existence and relevance of the first half of the EKC curve is well established [10], we feel more analysis needs to be done with multiple factors.

Proposed Method

Intuition

Deforestation is a complex topic and there could be many localized factors at play. Factors relevant for one country may not be applicable to another.

We analyzed many research papers on this topic and realized the need for

- Analyzing different factors collectively to gain insights on the causes of deforestation
- An interactive visualization that would allow the user to slice and dice the information and derive insights from it.

Our project aims to address these two requirements in an innovative way. We intend to -

- Analyze each of these factors against deforestation data.
- Identify the most relevant factors by ranking them by correlation score.
- Identify cluster of countries based on all relevant data points.
- Predict future deforestation rate for a country / region.

The intuition behind our approach is to provide a unique and creative platform to analyze and identify the socio-economic factors that are most relevant to the deforestation rate globally / country / region. This platform can be utilized by government agencies and independent researchers to derive policy decisions. Once the most relevant factors are identified, a deeper study can be done on the shortlisted factors. This will help us derive inferences about the strength of the relationship of these factors with deforestation.

Additionally, we will rank the factors in order of their importance and bring all meaningful comparisons under one hood. User can interact with the data and compare various factors against deforestation statistics across individual countries or collectively.

Data

The factors that we have considered are –

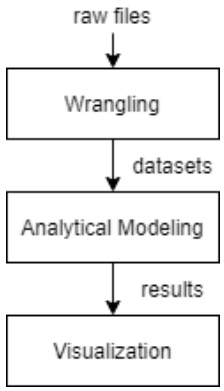
i. Forest area (% of land area)
ii. GDP per capita
iii. GNI per capita
iv. Inflation %
v. Population density
vi. Human Development Index
vii. Unemployment %
viii. Infant mortality %

Data collection was done by downloading and web scraping. Raw data was obtained from different websites - worldbank.org, undp.org, unicef.org and un.org as flat files. Income classification for countries was derived using web scraping.

Approach

The project is divided into three sub-systems:

1. Wrangling – cleanse, transform, and integrate the disparate data files and populate Country, Measurements and Relationships Table
2. Analytical Modeling – prep and train the models with the data to generate the results
3. Visualization – Dashboards to expose the results to user



Wrangling

The datasets are cleansed, transformed, aggregated to obtain a yearly measure by each country for the last 30 years.

Data cleaning and analysis was implemented using Python (pandas, sklearn, statsmodels, urllib, regex, matplotlib).

Data cleaning involved eliminating countries where factor measurements were missing or the datapoints were outliers. Cleansed data was stored in following 4 flat files.

Table	Fields
Country	Country name, continent, income status and development status
Measurements	Country id, Year, Factor measurements, Actual vs Predicted
Relationships	Correlation measure & classification
Clustering	Country id, Factor measurements and cluster id

A representative data in the Measurements table is as mentioned below -

Field	Example Value
year	2000
hdi (YoY change)	-0.0045
% Inflation	0.0338
infant_mortality (YoY change)	-0.0154
gdp (YoY change)	0.0501
gni (YoY change)	0.0637
% forest cover	0.3313
population (YoY change)	0.0107
% unemployment	0.0399
country_id	116
Info	actual

Analytical Modelling

This layer calculates the YoY change for every factor and obtains the analytical models results. It performs the following –

- Calculation of correlation score between forest cover and factors using Pearson's correlation coefficient.
- Identification of top ranked factors using classification and ranking.
- Principal Components Analysis for dimensionality reduction and K-means clustering on the principal components. This was done to derive insights which will not be available via classification.
- Prediction of Future Forest cover using Holt Winter method.

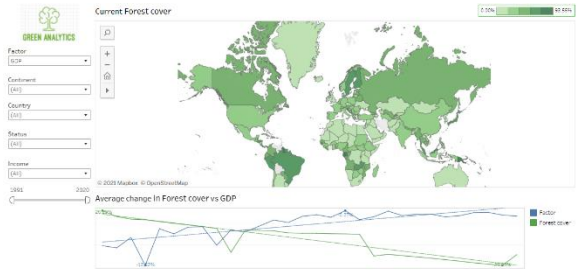
Visualization

Web interface of the Tableau dashboard is published using public Tableau at -

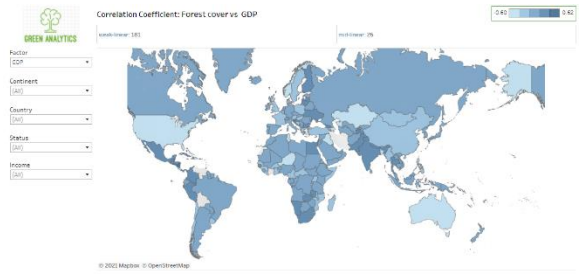
https://public.tableau.com/app/profile/subh6914/viz/GreenAnalytics_16375909366540/Dashboard1

The visualization subsystem consists of seven interactive Tableau dashboards –

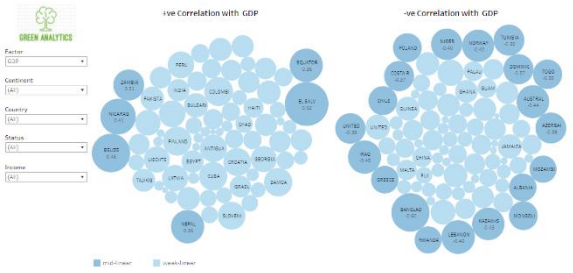
Dashboard #1
Filters: Geography (Continent, Income level, Country), Development Status, Socio-economic factor, Time range.
Layout: Choropleth with current forest cover and trendlines depicting the change in forest cover versus the selected socio -economic factor.



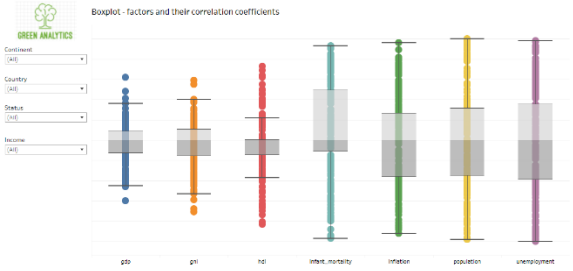
Dashboard #2
Filters: Geography (Continent, Income level, Country), Development Status, Socio-economic factor
Layout: Choropleth depicting correlation coefficient of Forest cover versus selected Socio-economic factor



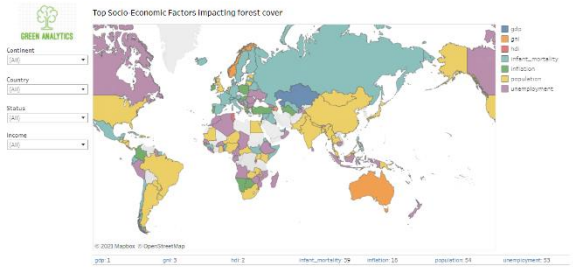
Dashboard #3
Filters: Geography (Continent, Income level, Country), Development Status, Socio-economic factor
Layout: Bubble chart depicting countries having positive and negative correlation against selected socio-economic factor



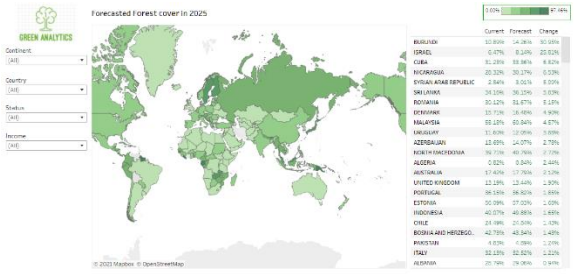
Dashboard #4
Filters: Geography (Continent, Income level, Country), Development Status
Layout: Box plot of correlations between all socio-economic factors and forest cover for the selected region.



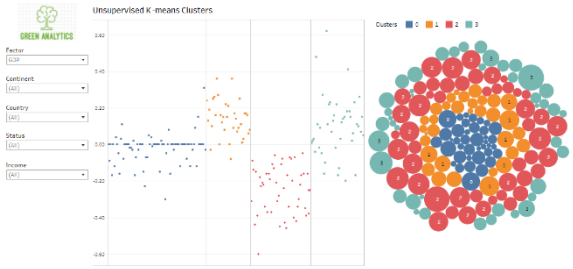
Dashboard #5
Filters: Geography (Continent, Income level, Country), Development Status
Layout: Choropleth depicting top socio-economic factor impacting forest cover for the selected region



Dashboard #6
Filters: Geography (Continent, Income level, Country), Development Status
Layout: Choropleth depicting forecasted forest cover in 2025 for the selected region



Dashboard #7
Filters: Geography (Continent, Income level, Country), Development Status, Socio-economic factor
Layout: Scatter plot and Bubble chart depicting clusters derived through K-means for selected socio-economic factor



Design of Experiments/ Evaluation

Testbed consists of following components:
OS: Windows 10
Browser: Google Chrome
Python: Python 3, Pandas, Numpy, Sklearn
Tableau: Tableau Desktop 2021
IDE: Visual Studio Code, Jupyter notebook
Version Control: GitHub

The questions the Tableau dashboards will be designed to answer are as follows

Dashboard #1
- What is the current forest cover for the selected region?
- What is the trend and average / yearly changes in socio-economic factor and forest cover?

Dashboard #2
- What is the strength of correlation between forest cover and socio-economic factor?

- How do countries in a region fare against each other in terms of impact of selected factor against forest cover?

Dashboard #3

- Which countries have a positive or negative correlation for a selected factor?
- How do countries compare against each other in these two categories?

Dashboard #4

- Which countries are the outliers for correlation between selected factor and forest cover?

Dashboard #5

- For a selected region, what is the top socio-economic factor for individual countries impacting forest cover?
- Is there a dominant socio-economic factor impacting forest cover in the selected region?

Dashboard #6

- What is the forecasted forest cover for 2025 for a selected region?
- Which countries are likely to have increased forest cover in future?
- Which countries are likely to have reduced forest cover in future?

Dashboard #7

- What is the forecasted forest cover for 2025 for a selected region?

Observations

Using the dashboards, we were able to answer the following questions with just few clicks

1. *Which socio-economic factor is ranked the highest overall, by continent?*
   For Africa, unemployment is the top factor correlating to forest cover change for 18 countries (Dashboard -5)

2. *Which socio-economic factor is ranked the highest overall, by income group?*
   For "Low Middle Income" group, population is the top factor correlating to forest cover change for 19 countries (Dashboard -5)

3. *Which countries have performed better than their neighbors?*
   - Amongst European countries, Malta has the highest correlation (0.9) of population with forest cover. This is contrary to popular belief. (Dashboard -3)

4. *Which countries should be analyzed further to understand how deforestation can be slowed down without impacting the socio-economic factors?*
   - Portugal has a mid linear correlation score between HDI and forest cover. This implies that the country has improved its HDI without compromising its forest cover. (Dashboard -3). Portugal could be a model country to study in detail.

5. *What is the predicted level of deforestation?*
   - Amongst developed countries, United States will increase its forest cover by 0.53% by 2025 (Dashboard -6)

6. *For developed countries, which countries are outliers while analyzing the impact of unemployment on forest cover?*
   - Canada, Greece and Switzerland are outliers when we analyze the impact of unemployment with forest cover. (Dashboard – 4)

7. *After 2000, what is the overall trend for inflation and forest cover for Asian countries?*
    - Inflation has been more or less flat in general for the 20 year period from 2001 to 2020. During the same period, forest cover has shown a slight upward trend from 3.82% to 3.86%. (Dashboard -1)

8. *For Upper Middle-Income countries, which countries have the strongest correlation between inflation and forest cover?*
    - Turkmenistan, Paraguay, Namibia, Tuvalu, Guyana, Columbia, Botswana has a strong positive correlation with forest cover whereas Turkey has a strong negative correlative with forest cover. (Dashboard-2 and Dashboard-3)

9. *Among the High-Income countries of Asia which factor has most -ve impact on Forest cover over past 3 decades?*
    - Population (Dashboard-5)

10. *Among the North and South Americas which countries have a strong decline in Forest cover?*
    - Brazil, El–Salvador, Nicaragua (Dashboard-1 and Dashboard-7)

11. Which countries can be grouped together on all factors combined and how their cluster distribution appear?
    - Dashboard 7

Conclusions and discussion

We have adopted a more holistic approach towards understanding the potential causes of change in forest cover. Based on the inferences derived above, we can conclude that not all socio-economic factors have a negative impact on forest cover.

In the long term, we hope these results would help in conducting deeper studies on the higher ranked factors. Also, we would like to showcase the results to government and non-government agencies, which will help them draft policies for the future.

Distribution of team member efforts

The project team is a mix of Project Management, Functional and Technical experts.

All team members have contributed similar amount of effort in the following activities:

- exploring project problem statements
- conducting literature survey
- identifying possible data sources
- discussing data transformation and high-level design
- project deliverables planning and coordination

Workload was equally divided amongst the team members as follows.

Responsibility	Champion(s)
Project Lead	Subhabrata Chaudhuri
Wrangling	Piyush Jain, Sahil Poonatar, Suneet Taparia, Sanjay Naik
Analytical Modeling	Suneet Taparia, Sanjay Naik, Saurabh Sinha
Visualization	Subhabrata Chaudhuri, Piyush Jain, Saurabh Sinha, Sahil Poonatar

References:

1. Juan Robalino and Luis Diego Herrera, 2009. Trade and Deforestation. World Trade Organization, p. 30
2. Muhammad Tariq, 2015. An Overview of Deforestation Causes and Its Environmental Hazards in Khyber Pukhtunkhwa. Journal of Natural Sciences Research, p. 4
3. Robert Walker, 1993. Deforestation and economic development. Canadian Journal of Regional Science, p.483
4. EK Yiridoe, DM Nanang, 2001. An econometric analysis of the causes of tropical deforestation: Ghana. American Agricultural Economics Association Conference
5. Bo Pieter Johannes Andrée, Andres Chamorro, Phoebe Spencer, Eric Koomen, Harun Dogo, 2019, Revisiting the relation between economic growth and the environment; a global assessment of deforestation, pollution and carbon emission, Renewable and Sustainable Energy Reviews, Elsevier
6. A Chakrabarti, 2021, Deforestation and infant mortality: Evidence from Indonesia, Economics & Human Biology, Elsevier
7. M Bhattarai and M Hamming, 2001. Institutions and the Environmental Kuznets Curve for deforestation, World Development Vol. 29, p. 5
8. Bolarinwa A. Ajanaku and Alan R. Collins, 2020. Economic growth and deforestation in African countries: Is the Environmental Kuznets Curve hypothesis still applicable? Agricultural & Applied Economics Association Annual Meeting, Kansas City, p. 26
9. David I. Stern, June 2003. The Environmental Kuznets Curve
10. Jeremy Irvin, Hao Sheng, Neel Ramachandran, Sonja Johnson-Yu, Sharon Zhou, Kyle Story, Rose Rustowicz, Cooper Elsworth, Kemen Austin, Andrew Y. Ng, Nov 2020. ForestNet: Classifying Drivers of Deforestation in Indonesia using Deep learning on Satellite Imagery
11. Jesús Crespo Cuaresma, Olha Danylo, Steffen Fritz, Ian McCallum, Michael Obersteiner, Linda See & Brian Walsh, 2017. Economic development and forest cover: Evidence from satellite data
12. Shichao Gao, 2019. Deforestration prediction using Time Series and LSTM
13. G. Cornelis van Kooten and Sen Wang, 2003. Institutional, social and economic factors behind deforestation: a cross-country examination
14. K Pahari, S Murai, 1999, Modelling for prediction of global deforestation based on the growth of human population, ISPRS journal of photogrammetry and remote sensing, Elsevier, p.320
15. Cuneyt Koyuncu and Rasim Yilmaz, 2009, The Impact of Corruption on Deforestation: A Cross-Country Evidence, The Journal of Developing Areas, Vol. 42, No. 2, pp. 213-222
16. Lykke E. Andersen, 1997, Modelling the Relationship between Government Policy, Economic Growth and Deforestation in the Brazilian Amazon
17. Huirong Feng, C. W. Lim, Liqun Chen, Xinnian Zhou, Chengjun Zhou, Yi Lin, "Sustainable Deforestation Evaluation Model and System Dynamics Analysis", The Scientific World Journal, vol. 2014, Article ID 106209, 14 pages, 2014. https://doi.org/10.1155/2014/106209

18. Haeuber, Richard. “Development and Deforestation: Indian Forestry in Perspective.” The Journal of Developing Areas 27, no. 4 (1993): 485–514. http://www.jstor.org/stable/4192258 .
19. Atrayee Banerjee and Chowdhury Madhurima, Vol. 5(8), pp. 122-129, September 2013, ‘Forest degradation and livelihood of local communities in India: A human rights approach’ https://academicjournals.org/journal/JHF/article-full-text-pdf/488F2773349