**CSE 545 Project Report : Sustainable Cities and Communities**

Pulkit Varshney, Pratik Thorwe, Purva Makarand Mhasakar, Saurabh Parekh

Introduction

Sustainable cities and communities is one of the seventeen sustainable development goals (SDG) defined by the United Nations General Assembly. This SDG Goal 11 fosters sustainable cities and communities[1]. As a part of this project, we utilized Big Data tools and technologies/concepts (similarity search, hypothesis testing, recommendation system) to implement solutions targeted towards addressing the needs of SDG goal 11. To make cities and communities inclusive, sustainable and resilient, we aim towards addressing the following targets of the goal [1]:

* Enhance inclusive and sustainable urbanization and capacity for sustainable human settlement
* Reduce the adverse per capita environmental impact of cities by paying attention to air quality
* Provide universal access to safe, inclusive and accessible, green and public spaces
* Support least developed countries in sustainable and resilient development though financial and technical assistance

The world is rapidly urbanizing. Currently around 3.9 billion people live in the cities and this number is expected to grow. By 2030, there can be 5 billion people living in the cities. [2] Cities are the powerhouses of economic growth, however it is critical that they are sustainable and resilient. Cities which occupy just 3% of the earth's land, account for nearly 80% of energy consumption. They contribute to about 70% of resource use and are a major contributor to global carbon emissions. As the urbanization progresses rapidly, it will lead to overburdened resources, services and infrastructure in the cities.

We focus on addressing these concerns by generating intelligent analytics using Big data frameworks (spark and Tensorflow). We use unstructured as well as structured data to gather a wide range of features such as urban population growth, Carbon dioxide emission metrics and forest area coverage etc. Our primary aim is to analyze the dependency of developing countries on the developed countries and how to reduce it by considering various attributes related to urban sustainability such as trend in trades, population in urban areas, Environmental conditions, consumption of renewable/non renewable resources, mortality rate and its causes, economy factors such as GDP etc.

Background

To build solutions targeted towards sustainable urbanization and development, we require a rich feature set consisting of features such as economic growth, carbon emission, GDP rate etc. Moreover it is significant that the data should be available for a wide range of time frames across multiple regions. Researchers have been using big data analytics to achieve the sustainable development goals. [Jonathan D.Moyer](https://www.sciencedirect.com/science/article/pii/S0305750X19303985#!) et al. [7] and [Charru Malhotra](https://link.springer.com/chapter/10.1007/978-981-10-7515-5_19#auth-Charru-Malhotra) et al.[8] in their work use analytics to study the current scenario of development goals. Kunal Pritwani et al.[9] have advocated the use of Spark framework to analyze the World Development Indicators which in turn help in understanding the development trends of cities. Researchers have also been widely using huge volumes of unstructured data such as news, satellite images, videos to gather data and development trends. We study the work presented by [Giorgio Pasquali](https://sciprofiles.com/profile/861832) et al. [3], where they used time series based satellite images for building footprint extraction which can be used to analyze urban development. They use the Multi-Temporal Urban Development SpaceNet Dataset presented by Adam Van Etten et al. [4]. These satellite images can be used for tracking human development activities, population statistics and forest area coverages. We draw inspiration from works such as Waleed Alsabhan et al.[5], [Giorgio Pasquali](https://sciprofiles.com/profile/861832) et al. [3] and [Mina Talal](https://ieeexplore.ieee.org/author/37086637746) et al.[6] who have leveraged tensorflow based image segmentation models to extract intelligent insights from satellite images. We built upon this by enlarging our focus area on forest and land area coverages and then performing the prediction of trends for upcoming time periods. In our project, we use big data to extract, load and transform our data and further apply concepts like recommendation systems and similarity search to present our big data analytics to achieve the sustainable development goal 11.

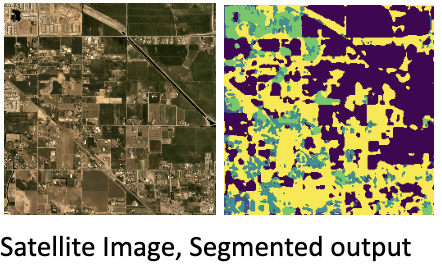
Data

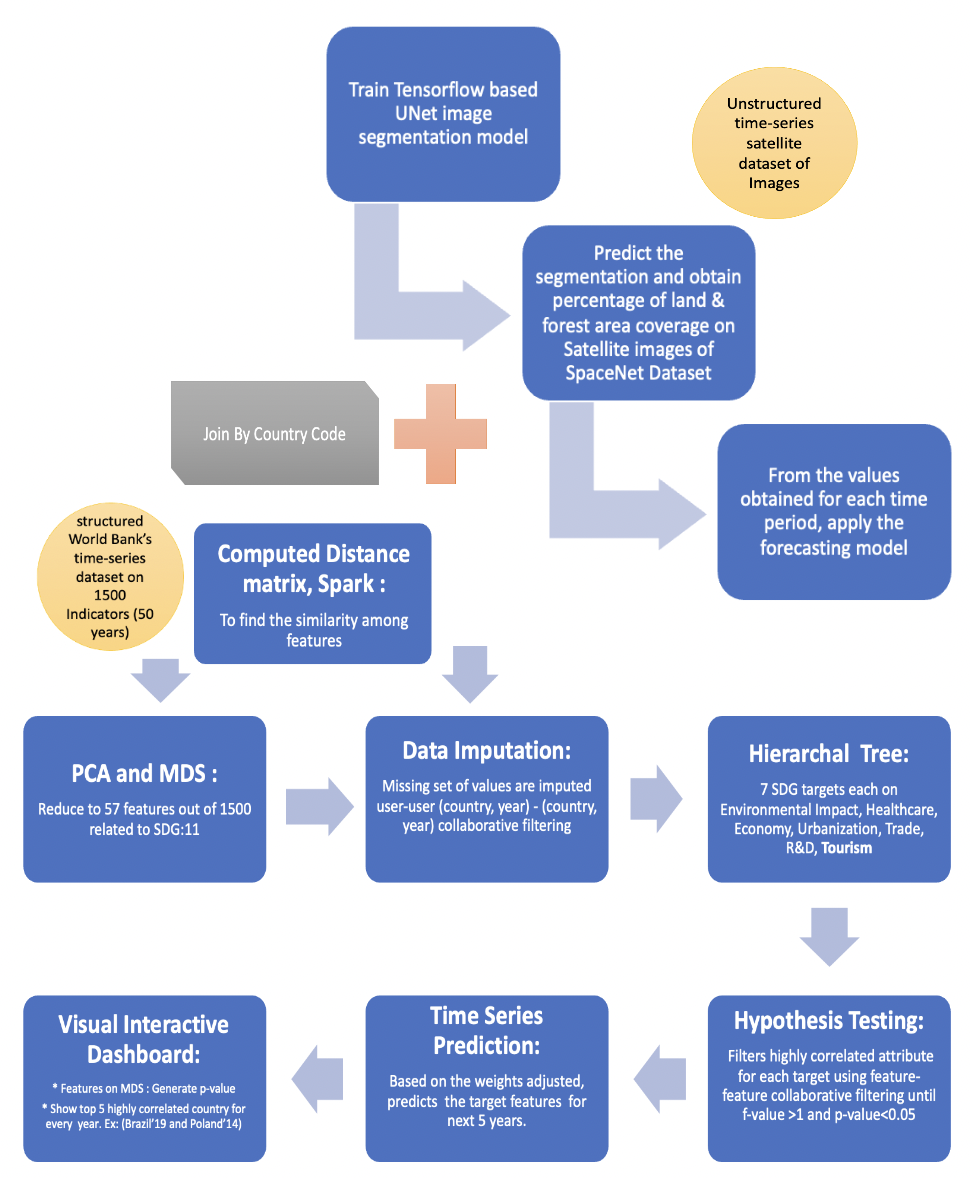
Our focus was to gather structured as well as unstructured data which helps in formulating a rich feature set. We describe our datasets in detail followed by their usage in the next section.

# 

Methodology

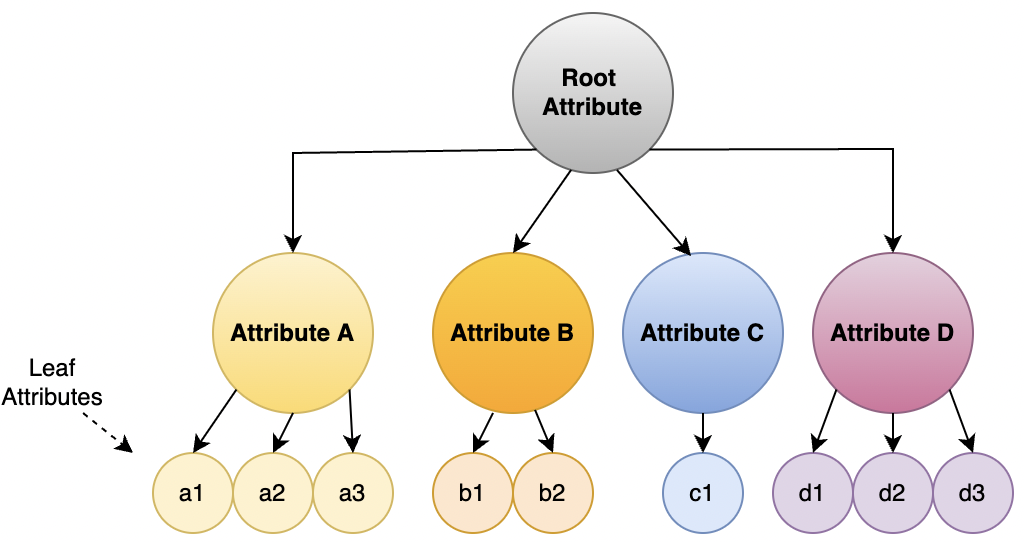
To achieve our aim of creating a sustainable world by creating sustainable cities across the world. We have come up with a recommendation system, which helps recommend the trends on the independent attributes to control the dependent attributes in such a way that we achieve our desired goal. For example: we suggest that GDP is a dependent variable which is dependent on various factors such as trade, expenditure on health care, tourism, Urban population etc. Our aim is to recommend the changes required in the independent attributes required over the years in order to obtain the optimum GDP growth in these years. To obtain the desired inferences we have structured our architecture in the following manner:

**Convert unstructured data to structured data using TensorFlow:** For each location and each month within the location, we apply image segmentation on the satellite images of the Multi-Temporal Urban Development SpaceNet[4] dataset. Firstly, we use the Semantic segmentation of aerial imagery[12] dataset to train the UNet architecture model. After the model is trained, given an input satellite image, our model segments the image into 6 classes. From these we calculate the percentage of greenery and urban land area coverage. We can observe the segmentation output in the figure. Moreover, we also have the data of the number of buildings per satellite image from the SpaceNet dataset. Once we obtain time series based structured data from the unstructured data, we perform forecasting using Prophet[13] forecasting procedure. We forecast the green area and urban area for each region. Next, we add these insights to the structured dataset. We further discuss the results in the inferences section. 



**Find highly correlated countries with time and impute missing values by Similarity Search:**1. [**Spark, similarity search**]: We standardize the data using min-max scalar, finds all-pair cosine similarity, filters top 100 highly correlated users(countries with year), imputes the missing values, converts rdd to csv again for hypothesis testing. Next, for the selected set of features, one good analysis comes out- we can compare countries' growth in a year wrt to another country’s growth in a different year. Ex: Brazil’19 correlates with Poland 14**.**

**Generate root Hypothesis and test it :**

1. Our primary hypothesis (root hypothesis of the hypothesis tree) is that over the coming years, the dependency of the developing and under developed countries over the developed countries will reduce. Thus, we considered our root attributes to be factors that show dependency of a country over another (for eg: Trade (% of GDP), Cost of business startup procedures (% of GNI per capita) etc)

2. Here, the null hypothesis states that the root attribute is equally dependent on all of the attributes. Thus, we use multivariate linear regression to check the attributes that prove the null hypothesis wrong. For obtaining such a combination of attributes we keep fitting the model until we generate a F-value > 1 and P-value < 0.05/h here h is the number of attributes that satisfy the mentioned condition this is used to obtain Bonferroni correction.

Note: Each Attribute under root attribute defines the sub targets of our SDG and the root target defines the main hypothesis.

**Generate a Hypothesis tree to test all the hypotheses :** Once we test the root hypothesis we obtain the attributes on which the root hypothesis is depended. Based on these attributes we recursively perform hypothesis testing to find the child attributes on which the concerned parent depends.

Next, we reach the leaf nodes of our hypothesis tree which contains the base attributes that govern our root attribute. 

(Sample output of hypothesis testing can be seen on the figure in left)

**Recommend the changes for the coming years in the attributes to govern the required change in the root hypothesis :**

Based on the hypothesis tree, we predict the values of the attributes for the next 10 years using the Auto Regressive Integrated Moving Average(ARIMA) model. The first step is to check if the series is stationary or not using the [Augmented Dickey Fuller(ADF) test](https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/). If the time series has a trend or seasonality component, then it is made stationary through differencing. Then, we use the ACF and PACF plots to decide whether to include an AR term(p), MA term(q), or both. Finally, we build the model and forecast the values for the next 10 years.

**Evaluation/Results**

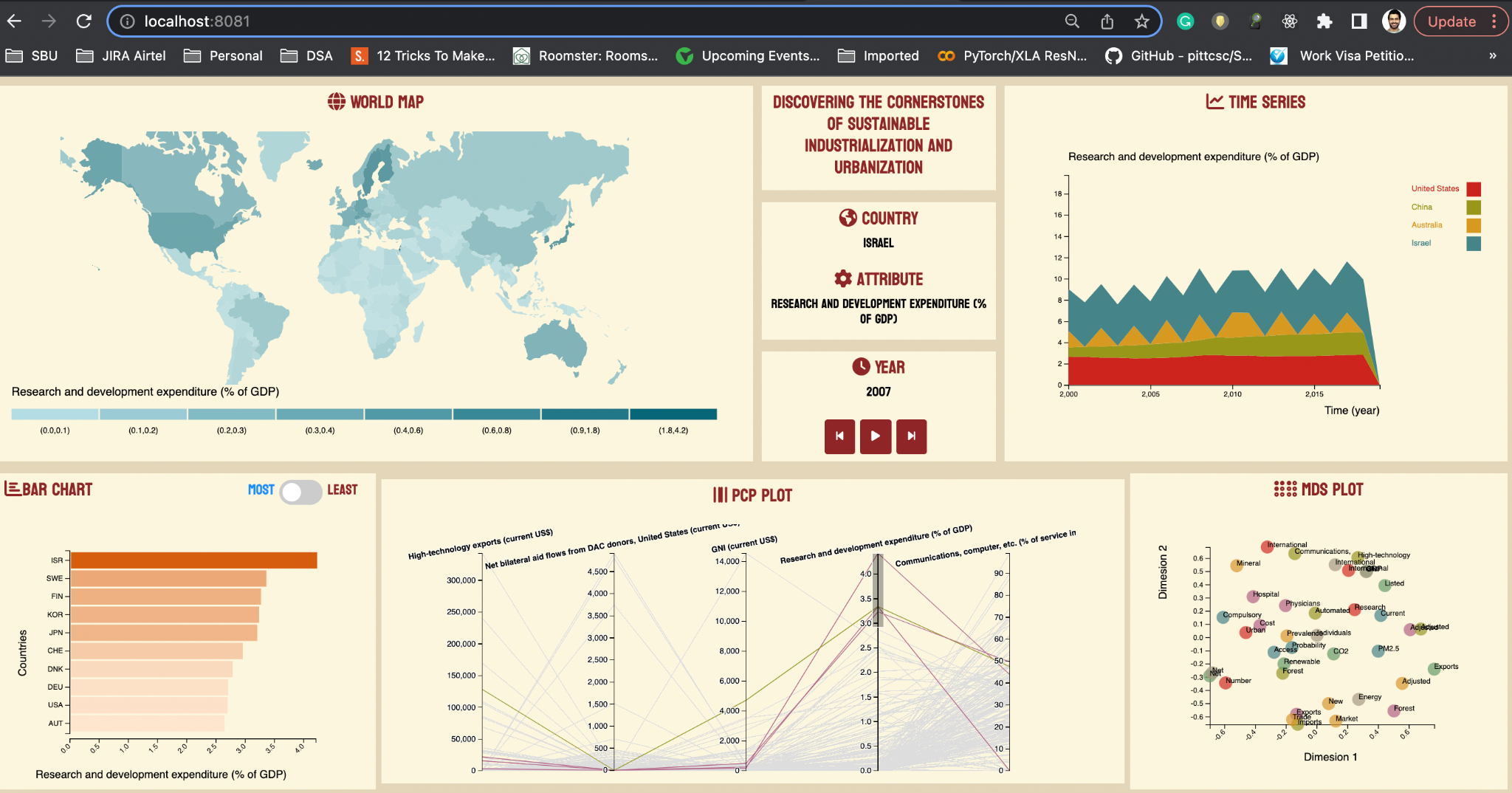


Fig: You see China’s spend into R&D has surpassed US overtime but Israel is leading globally. One hypothesis is Israel used to get a lot of aids from US initially, and it spends heavily in Military.(verify p-val)

(image :<https://drive.google.com/drive/folders/1eMtVMAsF61T_Ygou7iZY0veRLvikat30?usp=sharing> )

**Interactive Visual Dashboard**

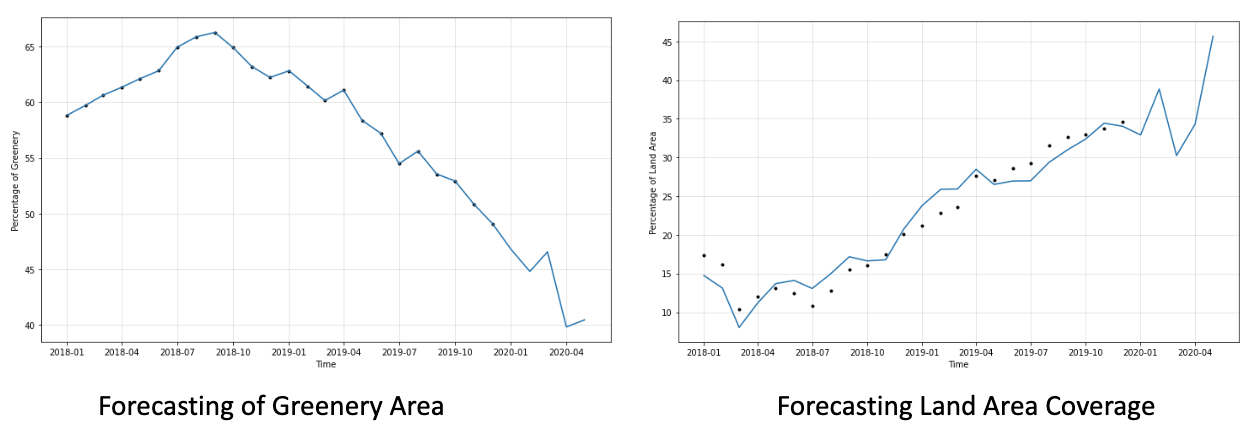
Firstly, we select features from MDS that will change BarChart, GeoMap, TimeSeries. Next, selecting any set of features will update PCP. Also, you can see the p-value of these sets of attributes on our dependent attributes(7 SDG:11 Targets) to check the hypothesis. Additionally selecting a country and a year will show highly correlated 5 countries for that year and time. Ex: Top 5 countries with similar trends to the USA in 2020.

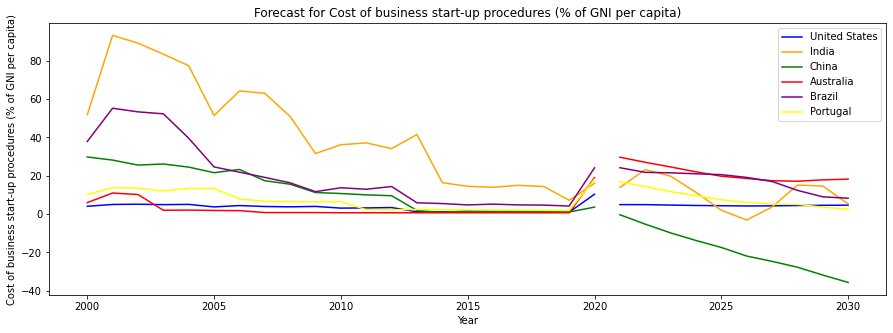
**Forecasting forest and land area**

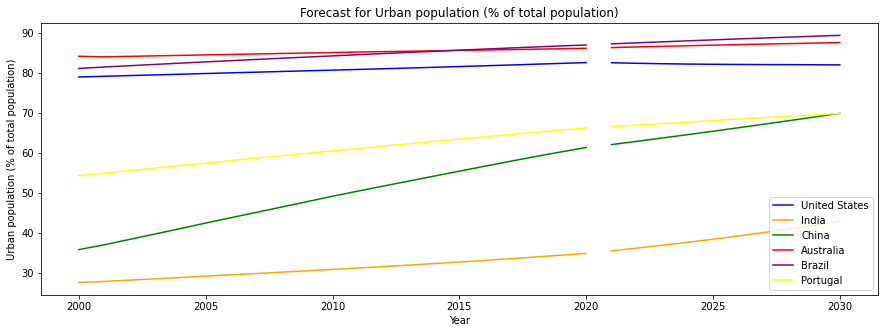
We predict greenery ,land area coverage of areas of interest from the satellite images. In the figure we see the forecast of a region of Louisiana in the USA. We observe that the green area relatively declined over time, while the urban land area coverage has been increasing steadily. As seen in the 1st Figure.

**Forecasting trends for business start-ups & urban population**

We predict the cost of business start-up procedures for 5 countries and urban population for the next 10 years from 2021 to 2030 based on past 20 years and other exogenous attributes recommended by hypothesis testing. As seen in below 2nd and 3rd Figures.







**Conclusion:**

Considering the similarity search that we obtained the missing trends in the past data and then we used hypothesis testing to understand the relation between various attributes using which we predicted the trends of the attributes affecting the growth of a country given that we maintain the sustainability in the cities. We successfully obtained trends required in various parameters mentioned above in the upcoming years, to reach our goal of creating sustainable countries by developing sustainable cities.

References

[1]<https://sdgs.un.org/goals/goal11>

[2]<https://www.un.org/sustainabledevelopment/cities/>

[3]<https://www.mdpi.com/2072-4292/11/23/2803>

[4]<https://openaccess.thecvf.com/content/CVPR2021/papers/Van_Etten_The_Multi-Temporal_Urban_Development_SpaceNet_Dataset_CVPR_2021_paper.pdf>

[5]<https://www.hindawi.com/journals/cin/2022/5008854/>

[6]<https://ieeexplore.ieee.org/abstract/document/8642743>

[7]https://www.sciencedirect.com/science/article/pii/S0305750X19303985

[8]<https://link.springer.com/chapter/10.1007/978-981-10-7515-5_19>

[9]<https://computerresearch.org/index.php/computer/article/view/1771>

[10]<https://databank.worldbank.org/source/world-development-indicators>

[11]<https://humansintheloop.org/resources/datasets/semantic-segmentation-dataset/>

[12]<https://www.kaggle.com/datasets/humansintheloop/semantic-segmentation-of-aerial-imagery>

[13]<https://facebook.github.io/prophet/>

[14]<https://towardsdatascience.com/u-net-for-semantic-segmentation-on-unbalanced-aerial-imagery-3474fa1d3e56>

[15]<https://github.com/amirhosseinh77/UNet-AerialSegmentation>

[16]<https://github.com/dead-pool-kit/CSE-545-final-project>