Engineering Clinics Project Report

‘ VISION INDIA 2020 ’

Team Members :

Tamil Selvan V - 18BIS009
Harshavardhan T - 18BIS034
Pawan Kumar S - 18BIS042
Sheik Harris F - 18BIS048
Guruprasath M - 18BIS057
Dhinesh Babu R - 18BIS059

ABSTRACT

Our Honourable ex president Dr. APJ Abdul Kalam with a team of 500 experts planned a mission “Vision India 2020:A Vision for the new millennium” ,which mainly focused on Doubling the production of Agriculture and food processing,and providing good infrastructure to the rural areas,Increase in literacy and health care,growth of critical technologies and statistical industries and development in Information and Communication technology.After his death is it possible to achieve his dream? We use certain Machine Learning algorithms such as Linear Regression,Polynomial Regression to predict the growth rate of GDP at 2020.The growth rate of India at 2015 was 6.01676143 and the growth rate of India at 2020 will be 5.9359773 .From our prediction we can see that ,the growth rate of India was decreasing since 2015 rather than to be increasing according to Dr. APJ Abdul Kalam.Therefore we conclude that, Dr. APJ Abdul Kalam’s mission “Vision 2020 “ can be a failure.

INTRODUCTION

The report is on “Vision India 2020”. Dr. APJ Abdul Kalam sir’s main aim in this mission is to double the growth rate of India's GDP by 2020. The main aim of our project is to conclude whether ‘Vision India 2020’ is a Success or Failure then to deliver the numerical insights on economy which influences development and work on finding the possible year in which ‘Vision India 2020’ could succeed.

OBJECTIVE:

This project is to predict whether Vision India 2020 is possible or not by using machine learning.

LIBRARIES REQUIRED:

1. Numpy
2. Pandas
3. Matplotlib
4. seaborn
5. Scikit

RESOURCES:

We have collected various data sets from our government database which is based on the ideas given by Dr. A. P. J. Abdul Kalam.

The Official Site Link is https://data.gov.in/

PROPOSED WORK and RELATED WORK:

- Problem statement analysis
- Phase of data solving
- Algorithm Selection

Problem statement analysis :

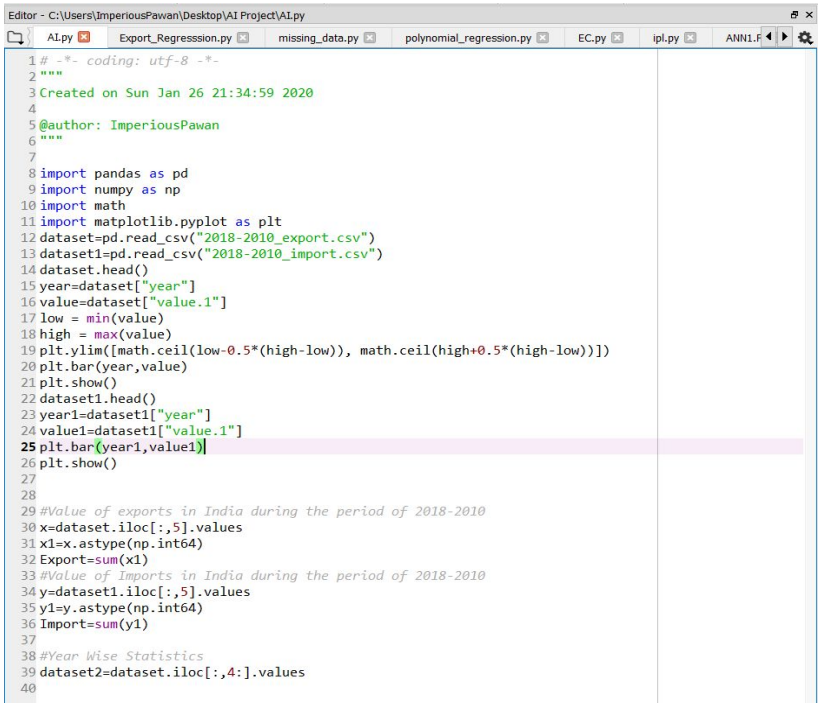
The initiative is to check whether our honourable ex-president Dr. APJ Abdul Kalam sir's mission of "Vision India 2020" is a success or not.He mainly focused on Doubling the production of Agriculture and food processing,and providing good infrastructure to the rural areas,Increase in literacy and health care,growth of critical technologies and statistical industries and development in Information and Communication technology as the key factors to achieve the mission.Our mission is to say whether "Vision India 2020" is a success or failure and to predict the possible year when this mission could get fulfilled.So this project can be divided into two phases.The first phase is to say whether "Vision India 2020" is a success or failure.The next phase is to find(or)predict the possible year in which the principles provided by Dr. APJ Abdul Kalam will be achieved so that India could be framed as a developed nation.

Phase of Data Solving :

One of the most important tasks in solving a data science problem is to frame up the required impactful data related to the problem statement. Let me tell the data solving relative to our two phases of our project.First is for "The success of Vision India 2020" ,in

that according to check the objectives of the accomplishment of Kalam sir’s objective,First we framed up a dataset with yearly numbers of the following subjects of our nation.Agriculture forestry and fishing , Mining and Quarrying , Manufacturing , Electricity Gas and water supply , Construction , Trade Hotels and Restaurant ,Transport Storage and Communication , Financing Insurance and Real estate and Business Services , Community Social and Services ,GDP at factor cost ,Consumption of Fixed capital , Net domestic product(NDP)at factor cost , Net factor income from abroad , gross National Income(GNI) at factor cost , Net National Income(NNI)at factor cost , Population(in millions) , Per capita income(Rs.) , Net Indirect taxes , Indirect taxes , Subsidies , Gross Domestic Product at market prices , Government final Capital Expenditure(GFCE) , Private final capital expenditure(PFCE) , Gross fixed capital formation(GFCF) , CIS , Valuables , Exports of goods and services , Less imports of goods and services , Discrepancies , Expenditure on Gross Domestic Product(GDP) are the vast ranged parameters which could show a nation’s progress to us.But the beauty of effective data analysis lies in finding the most influential parameters and use them to solve our parameters which nearly becomes equivalent to using all the possible features of impact.We could find the correlation between these features and the impact of each feature over the result.Using them we can process the necessary dataset for efficient problem solving.To be more clear over a factor let me take up the example of Import export data we have collected a dataset over a lakh of data for specific import and specific export to several countries over the years.We just

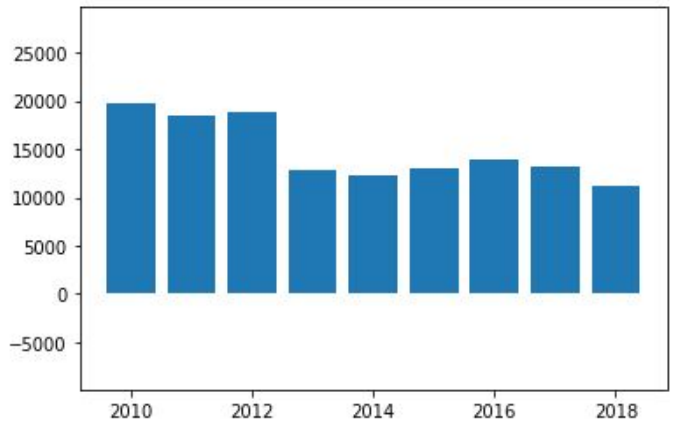
plotted two bar charts Export value(in.rs) vs Year and Import value(in.rs) vs Year.



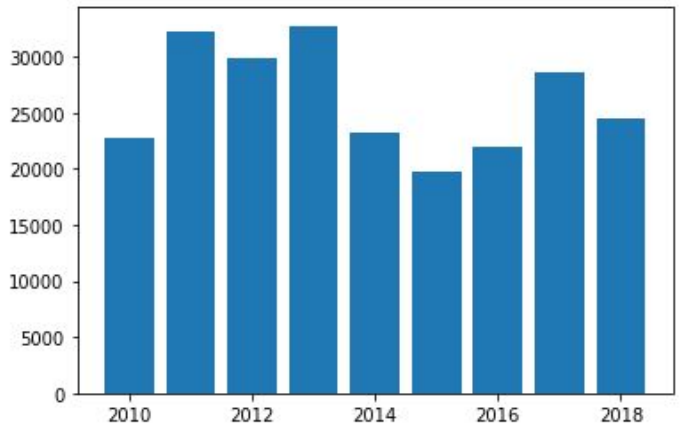

```
# -*- coding: utf-8 -*-
"""
Created on Sun Jan 26 21:34:59 2020

@author: ImperiousPawan
"""

import pandas as pd
import numpy as np
import math
import matplotlib.pyplot as plt
dataset=pd.read_csv("2018-2010_export.csv")
dataset1=pd.read_csv("2018-2010_import.csv")
dataset.head()
year=dataset["year"]
value=dataset["value.1"]
low = min(value)
high = max(value)
plt.ylim([math.ceil(low-0.5*(high-low)), math.ceil(high+0.5*(high-low))])
plt.bar(year,value)
plt.show()
dataset1.head()
year1=dataset1["year"]
value1=dataset1["value.1"]
plt.bar(year1,value1)
plt.show()


#Value of exports in India during the period of 2018-2010
x=dataset.iloc[:,5].values
x1=x.astype(np.int64)
Export=sum(x1)
#Value of Imports in India during the period of 2018-2010
y=dataset1.iloc[:,5].values
y1=y.astype(np.int64)
Import=sum(y1)

#Year Wise Statistics
dataset2=dataset.iloc[:,4:].values

```

Output :

Export->

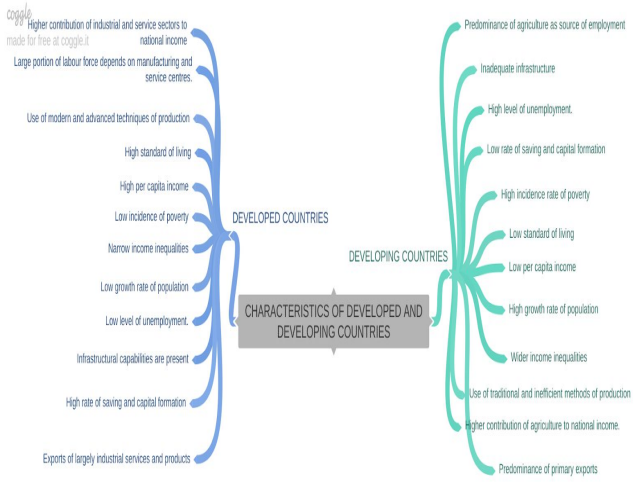


Import - >



What is the inference we have made from this is that in the year 2013 the Import-Export tradeoff is too high (i.e Import-Export ratio is comparatively higher than other years) Interestingly,in the same year 2013,India faced its record low GDP less than 4.5% . So we planned to find the effect if Import-Export tradeoff and its effect on growth,where it became evident that Import and Export ratio is inversely proportional to the growth.It showed us obvious difference on the outcome.From this we can conclude that the Import-Export ratio has much more impact in the growth of a nation.Likewise we sorted features such as agriculture forestry and fishing , Mining and Quarrying , Manufacturing , Electricity Gas and water supply , Construction , Trade Hotels and Restaurant ,Transport Storage and Communication , Financing Insurance and Real estate and Business Services , Community Social and Services to be used in our model.Using this data we could be able to find out whether “Vision India 2020” is a success or failure.But for phase 2 (i.e ‘To say by which year this mission could get fulfilled’) which is a tricky one ,there requires much more research on the parameters to be set as goal state.After

those level framing we can train our model on the existing previous year data,tune it on validation and finally use the goal state as test input to check the year in which Kalam sir's mission could possibly succeed.I will clear one of the challenge by one of our experiences.After training our model on the finalized dataset,we used the current state of USA in those parameters to check when India may reach the current state of USA.It gave us a pitiful result of a year over 2100.From this what became evident is each nation has its own specific goal state to say it a developed nation using numbers.USA has a different geographical size,different count of population,different kind of need etc.So it will not be right if we use the numbers of a developed nation to say whether our country reached the state of developed countries.From the picture given below we can infer the possible qualities of developed countries.We can use those factors,but the challenge is to find the adequate numbers for 'our country'.It requires research work with more data and case studies.Phase 2 of our project is still in the research phase.Phase 1 has a clear and finalized pavement to acquire the output we wish for.



Algorithm Selection :

To say whether ‘Vision India 2020’ is a success or failure we framed up a process of finding yearly growth rates and predict the growth rate of 2020.Developed countries may have growth rates lesser than before since their factors become different in progress.But,since India is a developing country the growth rate should be increasing over 2020 more than any other previous year to at least say India is progressing towards the goal.Using growth rate formula (which will be provided with a walkthrough in ‘Exploratory data analysis’) we could compute the growth rate over the years.Now using them as labelled inputs we can train our model.Since we use only one factor of impact after processing here,the model algorithms of our selection were ‘linear and polynomial regression’ . But since the growth rates could be much scattered over the axes linear regression may not be much accurate,it may have higher loss value and lesser accuracy.Comparing to linear regression,polynomial regression could give much lesser loss and higher accuracy.The challenge though,polynomial regression usually overfits on training data if hyperparameters are fixed in favor of training data.So implementing polynomial regression is a test and try process,start with the least degree model(i.e 2) observe the results and decide on tuning the hyperparameters.Through this process we can have a controlled model implementation to prevent overfitting.Another possible algorithm which we could have tried is ‘Multiple Linear Regression’ to use all the areas in the finalized dataset after the phase of data solving.But the problem in that is for 2020 we have to generate all the numerics which we cant be sure of.So it was

not used,same applies for neural networks.But if we could find the goal state numerics,we are sure that our model will provide a efficient result on the possible year of ‘Vision India 2020’s’ fulfillment.

Related work(summarized):

There hasn’t been any authenticated data science based works on ‘Vision India 2020’ till now.There has been researches on when India could reach developed state but still there is no proper source of outcome.So hopefully our project becomes fulfilled with an conclusion to the status of ‘Vision India 2020’ and with another phase well versed to show potential results on the year of fulfillment.This project not only shows the outcome of ‘Vision India 2020’ but has some more abstractions to it.From the correlations and Impact numbers we could be able to find the response of one field over another, the field to be concentrated and its effects,most influential field over development etc.Using this there could be a better possible learning on the system and what could be done to uplift and reform.

EXPLORATORY ANALYSIS AND PROPOSED SYSTEM

WHAT IS DATA CLEANING?

- Data cleaning means filtering and modifying your data such that it is easier to explore, understand, and model.

- Filtering out the parts you don't want or need so that you don't need to look at or process them.

<u>THE UNCLEARED DATA:</u>

Indicator (Rs in Crores)	2004	2005	2006	2007	2008	2009	2010	2011
Agriculture, Forestry & Fishing	565426	NaN	619190	655080	655689	662509	709103	728667
Mining & Quarrying	85028	86141	92578	95997	98055	104225	109421	108469
Manufacturing	453225	499020	570458	629073	656302	719728	774162	793468
Electricity, Gas & Water Supply	62675	67123	73362	79430	83050	88266	90944	98105
Construction	2,28,855	2,58,129	2,84,806	3,15,495	3,32,329	3,55,717	3,84,199	4,04,617
Trade, Hotels & Restaurant	4,77,303	5,35,397	5,94,918	6,55,013	6,92,224	7,46,178	8,13,079	9,04,884
Transport, Storage & Communication	2,50,417	2,80,010	3,15,166	3,54,507	3,92,901	4,51,035	5,17,376	5,57,888
Financing, Ins., Real Estate & Business Services	4,37,174	4,92,340	5,61,063	6,28,124	7,03,629	7,69,883	8,49,995	9,31,714
Community, Social & Pers. Services	4,11,361	4,40,426	4,52,823	4,83,917	5,44,497	6,10,096	6,37,675	6,74,703
Gross Domestic Product (GDP) at Factor Cost	29,71,464	32,53,073	35,64,364	38,96,636	41,58,676	45,07,637	48,85,954	52,02,514

TO CLEAN THE DATA,WE ARE DOING :

- Transposing the data frames.
- Removing all commas and parentheses ().
- Converting into integer (int) datatype.
- Handling missing data.
- Filtering necessary features.
- Categorizing dependent & independent variables .
- Normalization of data

1.TRANSPOSING THE DATAFRAMES:

Interchanging each row and the corresponding column in data frames.

BEFORE TRANSPOSE:

Indicator (Rs in Crores)	2004	2005	2006	2007	2008	2009	2010	2011
Agriculture, Forestry & Fishing	565426	NaN	619190	655080	655689	662509	709103	728667
Mining & Quarrying	85028	86141	92578	95997	98055	104225	109421	108469
Manufacturing	453225	499020	570458	629073	656302	719728	774162	793468
Electricity, Gas & Water Supply	62675	67123	73362	79430	83050	88266	90944	98105
Construction	2,28,855	2,58,129	2,84,806	3,15,495	3,32,329	3,55,717	3,84,199	4,04,617
Trade, Hotels & Restaurant	4,77,303	5,35,397	5,94,918	6,55,013	6,92,224	7,46,178	8,13,079	9,04,884
Transport, Storage & Communication	2,50,417	2,80,010	3,15,166	3,54,507	3,92,901	4,51,035	5,17,376	5,57,888
Financing, Ins., Real Estate & Business Services	4,37,174	4,92,340	5,61,063	6,28,124	7,03,629	7,69,883	8,49,995	9,31,714
Community, Social & Pers. Services	4,11,361	4,40,426	4,52,823	4,83,917	5,44,497	6,10,096	6,37,675	6,74,703
Gross Domestic Product (GDP) at Factor Cost	29,71,464	32,53,073	35,64,364	38,96,636	41,58,676	45,07,637	48,85,954	52,02,514

CODE TO TRANSPOSE:

```
In [12]:  1  df = df.transpose()
```

AFTER TRANSPOSE:

Out[12]:

Indicator (Rs in Crores)	Agriculture, Forestry & Fishing	Mining & Quarrying	Manufacturing	Electricity, Gas & Water Supply	Construction	Trade, Hotels & Restaurant	Transport, Storage & Communication	Financing, Ins., Real Estate & Business Services	Community, Social & Pers. Services	Gross Domestic Product (GDP) at Factor Cost	...	Gross Domestic Product (GDP) at Market Prices
2004	565426	85028	453225	62675	2,28,855	4,77,303	2,50,417	4,37,174	4,11,361	29,71,464	...	32,42,209
2005	NaN	86141	499020	67123	2,58,129	5,35,397	2,80,010	4,92,340	4,40,426	32,53,073	...	35,43,244
2006	619190	92578	570458	73362	2,84,806	5,94,918	3,15,166	5,61,063	4,52,823	35,64,364	...	38,71,489
2007	655080	95997	629073	79430	3,15,495	6,55,013	3,54,507	6,28,124	4,83,917	38,96,636	...	42,50,947
2008	655689	98055	656302	83050	3,32,329	6,92,224	3,92,901	7,03,629	5,44,497	41,58,676	...	44,16,350
2009	662509	104225	719728	88266	3,55,717	7,46,178	4,51,035	7,69,883	6,10,096	45,07,637	...	47,80,179
2010	709103	109421	774162	90944	3,84,199	8,13,079	5,17,376	8,49,995	6,37,675	48,85,954	...	52,36,823
2011	728667	108469	793468	98105	4,04,617	9,04,884	5,57,888	9,31,714	6,74,703	52,02,514	...	55,95,856

8 rows × 30 columns

2. REMOVING ALL COMMAS AND PARENTHESES ():

- Since the data is in string datatype it must be converted to integer datatype to do operation in data.
- Before converting into integer datatype the commas and parentheses have to be removed in order to convert the string data type into integer data type.

WITH BRACKETS AND COMMA:

(25,154)
(37,288)
(54,997)
(1,16,472)
(28,777)
(95,097)
(1,32,049)
16,720

6	Constructi	2,28,855	2,58,129	2,84,806	3,15,495	3,32,329	3,55,717	3,84,199	4,04,617
7	Trade, Hot	4,77,303	5,35,397	5,94,918	6,55,013	6,92,224	7,46,178	8,13,079	9,04,884
8	Transport,	2,50,417	2,80,010	3,15,166	3,54,507	3,92,901	4,51,035	5,17,376	5,57,888
9	Financing,	4,37,174	4,92,340	5,61,063	6,28,124	7,03,629	7,69,883	8,49,995	9,31,714
10	Communit	4,11,361	4,40,426	4,52,823	4,83,917	5,44,497	6,10,096	6,37,675	6,74,703
11	Gross Dom	29,71,464	32,53,073	35,64,364	38,96,636	41,58,676	45,07,637	48,85,954	52,02,514
12	Consumpti	3,19,891	3,50,894	3,85,699	4,27,630	4,68,903	5,20,320	5,64,463	6,01,034
13	Net Dome	26,51,573	29,02,179	31,78,665	34,69,006	36,89,773	39,87,317	43,21,491	46,01,480
14	Net Factor	-22,375	-24,896	-29,515	-17,179	-25,384	-27,664	-52,776	-51,828
15	Gross Nati	29,49,089	32,28,177	35,34,849	38,79,457	41,33,292	44,79,973	48,33,178	51,50,686
16	Net Natior	26,29,198	28,77,283	31,49,150	34,51,827	36,64,389	39,59,653	42,68,715	45,49,652

CODE FOR REMOVING BRACKETS AND COMMA:

```
In [ ]:
#removing commas
for i in df.columns:
    df[i] = df[i].str.replace(',','')
    df[i] = df[i].str.replace(')','')
    df[i] = df[i].str.replace('(','')
```

AFTER REMOVING BRACKETS AND COMMA:

Indicator (Rs in Crores)	Agriculture, Forestry & Fishing	Mining & Quarrying	Manufacturing	Electricity, Gas & Water Supply	Construction	Trade, Hotels & Restaurant
2004	565426	85028	453225	62675	228855	477303
2005	NaN	86141	499020	67123	258129	535397
2006	619190	92578	570458	73362	284806	594918
2007	655080	95997	629073	79430	315495	655013
2008	655689	98055	656302	83050	332329	692224
2009	662509	104225	719728	88266	355717	746178
2010	709103	109421	774162	90944	384199	813079
2011	728667	108469	793468	98105	404617	904884

Discrepancies
25154
37288
54997
116472
28777
95097
132049
16720

3.HANDLING MISSING DATA.

The missing data can be handled in two ways:

1.removing the entire columns/rows.

2.replacing a value to that missing data.

We are following the second way,since we can't remove that entire row.

We are replacing the Num value with the mean of that entire row.

BEFORE HANDLING MISSING DATA

Indicator (Rs in Crores)	Agriculture, Forestry & Fishing	Mining & Quarrying	Manufacturing	Electricity, Gas & Water Supply	Co
2004	565426	85028	453225	62675	
2005	NaN	86141	499020	67123	
2006	619190	92578	570458	73362	
2007	655080	95997	629073	79430	
2008	655689	98055	656302	83050	
2009	662509	104225	719728	88266	
2010	709103	109421	774162	90944	
2011	728667	108469	793468	98105	

CODE FOR HANDLING MISSING DATA

In [17]:
```
#handling the missing data

D = df['Agriculture, Forestry & Fishing'].replace(np.nan,0)
D = D.astype(int)
mean = D.mean()
df['Agriculture, Forestry & Fishing'].replace(np.nan,mean,inplace = True)
df['Agriculture, Forestry & Fishing'].astype(int)
df
```

AFTER HANDLING MISSING DATA

Indicator (Rs in Crores)	Agriculture, Forestry & Fishing	Mining & Quarrying	Manufacturing
2004	565426	85028	453225
2005	574458	86141	499020
2006	619190	92578	570458
2007	655080	95997	629073
2008	655689	98055	656302
2009	662509	104225	719728
2010	709103	109421	774162
2011	728667	108469	793468

4. CONVERTING INTO INTEGER(INT) DATATYPE.

Since we can’t do arithmetic operations like addition,multiplication and etc when the data is in string/object data type it must be converted into integer datatype to do arithmetic operations.

BEFORE CONVERTING INTO INTEGER (INT) DATATYPE

```
In [19]:  1 df.dtypes
```

```
Out[19]: Indicator (Rs in Crores)
         Agriculture, Forestry & Fishing                    object
          Mining & Quarrying                                object
         Manufacturing                                      object
         Electricity, Gas & Water Supply                    object
         Construction                                       object
         Trade, Hotels & Restaurant                         object
         Transport, Storage & Communication                 object
         Financing, Ins., Real Estate & Business Services   object
```

CODE FOR CONVERTING INTO INTEGER(INT) DATATYPE

```
In [20]:  1 #INT COVENTING
          2 for i in df.columns:
          3     df[i] = df[i].astype(int)
```

AFTER CONVERTING INTO INTEGER(INT) DATATYPE

```
In [21]:  1 df.dtypes
```

```
Out[21]: Indicator (Rs in Crores)
         Agriculture, Forestry & Fishing                    int32
          Mining & Quarrying                                int32
         Manufacturing                                      int32
         Electricity, Gas & Water Supply                    int32
         Construction                                       int32
         Trade, Hotels & Restaurant                         int32
         Transport, Storage & Communication                 int32
         Financing, Ins., Real Estate & Business Services   int32
         Community, Social & Pers. Services                 int32
         Gross Domestic Product (GDP) at Factor Cost        int32
```

5.FILTERING NECESSARY FEATURES

The main aim of vision 2020 is to double the growth rate of india's gdp. Dr APJ. Abdul Kalam had already given in which sectors the impact must be seen .

The sectors are:

•Agriculture and food processing.

•Infrastructure with reliable electric power.

•Education and Healthcare.

•Information and Communication Technology.

•Critical technologies and strategic industries.

So,all other rows except these five rows are removed from the dataset

.

BEFORE FILTERING NECESSARY FEATURES

```
In [24]:  1 df.columns

Out[24]: Index(['Agriculture, Forestry & Fishing', ' Mining & Quarrying',
                'Manufacturing', 'Electricity, Gas & Water Supply', 'Construction',
                'Trade, Hotels & Restaurant', 'Transport, Storage & Communication',
                'Financing, Ins., Real Estate & Business Services',
                'Community, Social & Pers. Services',
                'Gross Domestic Product (GDP) at Factor Cost',
                'Consumption of Fixed Capital',
                'Net Domestic Product (NDP) at Factor Cost',
                'Net Factor Income from abroad',
                'Gross National Income (GNI) at Factor Cost',
                'Net National Income (NNI) at Factor Cost', 'Population (in Mn.)',
                'Per Capita Income (Rs.)', 'Net Indirect Taxes', 'Indirect Taxes',
                'Subsidies', 'Gross Domestic Product (GDP) at Market Prices',
                'Government Final Capital Expenditure (GFCE)',
                'Private Final Consumption Expenditure (PFCE)',
                'Gross Fixed Capital Formation (GFCF)', 'CIS', 'Valuables',
                'Exports of Goods and Services', 'Less Imports of Goods and Services',
                'Discrepancies ', 'Expenditure on Gross Domestic Product (GDP)'],
               dtype='object', name='Indicator (Rs in Crores)')
```

AFTER FILTERING NECESSARY FEATURES

```
In [5]:  1  data.iloc[:,1:10].columns

Out[5]: Index(['Agriculture, Forestry & Fishing', ' Mining & Quarrying',
               'Manufacturing', 'Electricity, Gas & Water Supply', 'Construction',
               'Trade, Hotels & Restaurant', 'Transport, Storage & Communication',
               'Financing, Ins., Real Estate & Business Services',
               'Community, Social & Pers. Services'],
              dtype='object')
```

6.CATEGORIZING DEPENDENT AND INDEPENDENT VARIABLES.

Agriculture, Forestry & Fishing	Mining & Quarrying	Manufacturing	Electricity, Gas & Water Supply	Construction	Trade, Hotels & Restaurant	Transport, Storage & Communication
565426	85028	453225	62675	228855	477303	250417
574458	86141	499020	67123	258129	535397	280010
619190	92578	570458	73362	284806	594918	315166
655080	95997	629073	79430	315495	655013	354507
655689	98055	656302	83050	332329	692224	392901
662509	104225	719728	88266	355717	746178	451035
709103	109421	774162	90944	384199	813079	517376

The independent variables are above listed five variables and the dependent variable is gdp rate.

7.NORMALIZATION OF DATA

The goal of normalization is to change the values of numeric columns in the dataset to a common scale and to make data as origin centred.

BEFORE NORMALIZATION:

Agriculture, Forestry & Fishing	Mining & Quarrying	Manufacturing	Electricity, Gas & Water Supply	Construction	Trade, Hotels & Restaurant	Transport, Storage & Communication
565426	85028	453225	62675	228855	477303	250417
574458	86141	499020	67123	258129	535397	280010
619190	92578	570458	73362	284806	594918	315166
655080	95997	629073	79430	315495	655013	354507
655689	98055	656302	83050	332329	692224	392901
662509	104225	719728	88266	355717	746178	451035
709103	109421	774162	90944	384199	813079	517376

CODE FOR NORMALIZATION:

```
col=data.iloc[:,1:10].columns
temp=[]
templ=[]
for i in col:
    t=data[i].values
    sd=np.std(t)
    mean=np.mean(t)
    templ=[]
    for j in t:
        d=(j-mean)/sd
        templ.append(round(d,4))
    temp.append(templ)

pd1=pd.DataFrame(data=stu, columns=col)
pd1
```

AFTER NORMALIZATION:

Agriculture, Forestry & Fishing	Mining & Quarrying	Manufacturing	Electricity, Gas & Water Supply	Construction	Trade, Hotels & Restaurant	Transport, Storage & Communication
-1.4886	-1.3223	-0.4986	0.1623	0.1735	0.2991	1.1571
-1.4166	-1.2901	-0.5583	-0.1696	0.0643	0.7657	1.3564
-1.5824	-1.1880	-0.5726	-0.0677	0.1669	0.7132	1.1821
-1.5560	-1.1649	-0.6162	-0.0826	0.2357	0.6944	0.9299
-1.6048	-1.0923	-0.6252	-0.0879	0.2068	0.6162	1.1149
-1.4985	-1.0634	-0.6176	-0.1675	0.1112	0.5153	1.0164
-1.3409	-1.0564	-0.7185	-0.3403	0.0287	0.5875	1.2252
-1.4431	-1.1037	-0.6809	-0.2683	0.1962	0.6038	1.0966

EXPLORATORY DATA ANALYSIS :

CORRELATION :

Correlation is a term that is a measure of the strength of a linear relationship between two quantitative variables.To know how the features are dependent on themselves, there is a need for correlation. Correlation can also be visualized using heat map.A **heat map** (or **heatmap**) is a data visualization technique that shows the magnitude of a phenomenon as color in two dimensions.

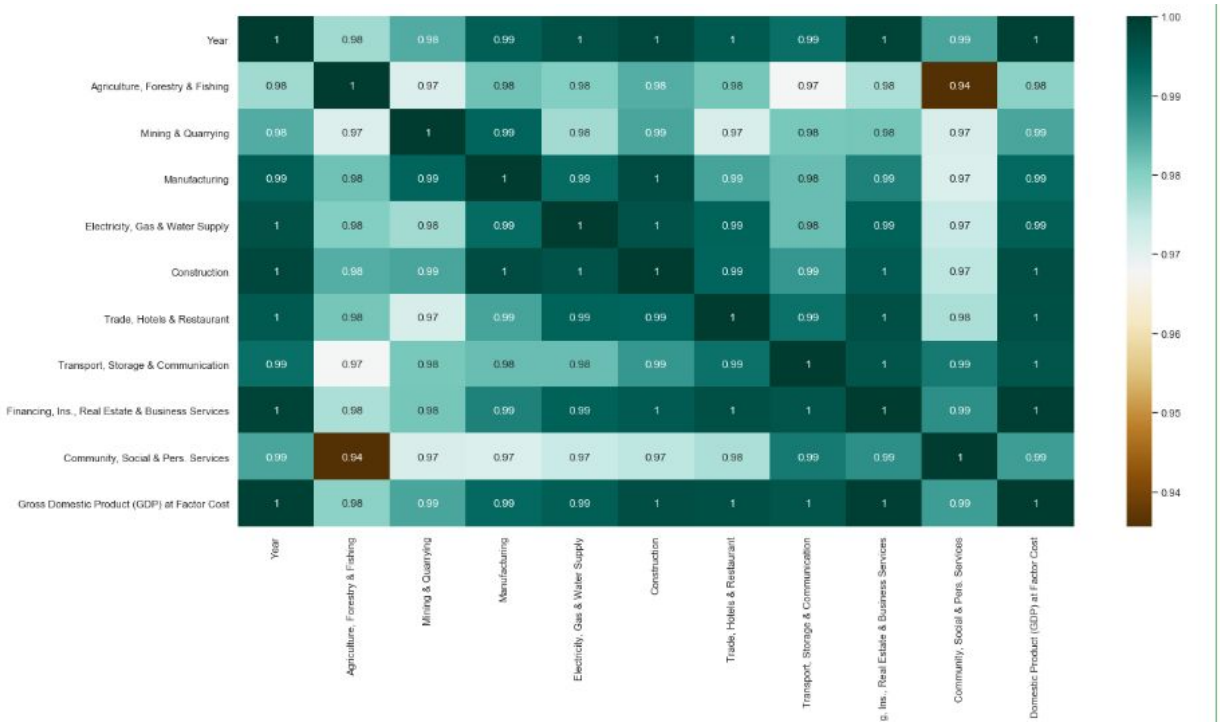
CODE FOR HEAT MAP:

```
In [3]: plt.figure(figsize=(20,10))
        c= df.corr()
        sns.heatmap(c,cmap='BrBG',annot=True)
        c
```

OUTPUT :

Out[3]:

	Year	Agriculture, Forestry & Fishing	Mining & Quarrying	Manufacturing	Electricity, Gas & Water Supply	Construction	Trade, Hotels & Restaurant	Transport, Storage & Communication	Financing, Ins., Real Estate & Business Services	Community, Social & Pers. Services	Gross Domestic Product (GDP) at Factor Cost
Year	1.000000	0.977688	0.984501	0.994486	0.996561	0.998186	0.995429	0.992436	0.998828	0.985279	0.999177
Agriculture, Forestry & Fishing	0.977688	1.000000	0.971137	0.982130	0.980207	0.984341	0.981543	0.968015	0.976733	0.935610	0.979406
Mining & Quarrying	0.984501	0.971137	1.000000	0.993611	0.977536	0.985193	0.971882	0.981273	0.981585	0.971595	0.985329
Manufacturing	0.994486	0.982130	0.993611	1.000000	0.992708	0.997268	0.985581	0.982570	0.989897	0.971519	0.993110
Electricity, Gas & Water Supply	0.996561	0.980207	0.977536	0.992708	1.000000	0.996466	0.993956	0.982536	0.994104	0.973891	0.994503
Construction	0.998186	0.984341	0.985193	0.997268	0.996466	1.000000	0.993676	0.986983	0.995084	0.974963	0.997023
Trade, Hotels & Restaurant	0.995429	0.981543	0.971882	0.985581	0.993956	0.993676	1.000000	0.991965	0.996794	0.976742	0.996836
Transport, Storage & Communication	0.992436	0.968015	0.981273	0.982570	0.982536	0.986983	0.991965	1.000000	0.995978	0.990943	0.996095
Financing, Ins., Real Estate & Business Services	0.998828	0.976733	0.981585	0.989897	0.994104	0.995084	0.996794	0.995978	1.000000	0.988033	0.999382
Community, Social & Pers. Services	0.985279	0.935610	0.971595	0.971519	0.973891	0.974963	0.976742	0.990943	0.988033	1.000000	0.986345
Gross Domestic Product (GDP) at Factor Cost	0.999177	0.979406	0.985329	0.993110	0.994503	0.997023	0.996836	0.996095	0.999382	0.986345	1.000000



PLOTTING BAR GRAPH AND LINE GRAPH AGAINST YEAR AND FEATURES('AGRICULTURE ETC...')

BAR GRAPH :

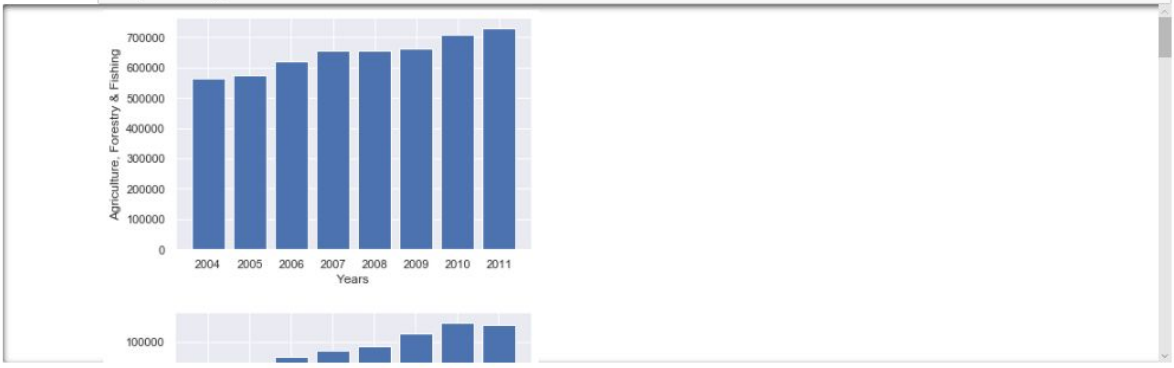
A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

LINE GRAPH :

A line graph (also known as a line plot or line chart) is a graph which uses lines to connect individual data points that display quantitative values over a specified time interval.

CODE FOR GRAPHS AND SAMPLE OUTPUT :

```
In [4]: for i in x.columns:
            plt.bar(y,x[i])
            plt.xlabel('Years')
            plt.ylabel(i)
            plt.show()
```




```
In [5]: for i in x.columns:
            plt.plot(x[i],y)
            plt.ylabel('Years')
            plt.xlabel(i)
            plt.show()
```

BUILDING UP THE MODEL :

model() - This function implements linear regression.
model2() - This function implements polynomial regression of degree ‘deg’ passed as argument in function call.
visualize_model() - This function plots the regression line on the scatter graph of the passed values.

CODING :

```
In [6]: def model(X,y):
            from sklearn.model_selection import train_test_split
            X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.28)
            from sklearn.linear_model import LinearRegression
            lin = LinearRegression()
            lin.fit(X_train, y_train)
            print("score = ",lin.score(X_test,y_test))
            return lin


        def visualize_model(lm,X,y,X1):
            plt.scatter(X, y, color = 'blue')
            plt.plot(X, lm.predict(X1), color = 'red')
            plt.title('Linear Regression')
            plt.show()
```

```
In [12]: def model2(X,y,deg):
             from sklearn.model_selection import train_test_split
             X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.28)
             from sklearn.preprocessing import PolynomialFeatures
             from sklearn.linear_model import LinearRegression
             poly = PolynomialFeatures(degree = deg)
             X_poly = poly.fit_transform(X_train)

             poly.fit(X_poly, y_train)
             lin2 = LinearRegression()
             lin2.fit(X_poly, y_train)
             print("score = ",lin2.score(poly.fit_transform(X_test),y_test))
             return lin2,poly
```

CREATING DATAFRAMES AND CALCULATING GROWTH RATE :

x - All features from agriculture to gdp.

gr = Growth rate for every feature(except gdp) over years.

GDP = Growth rate for only gdp.

GROWTH RATE FORMULA :

((GDP OF YEAR 2/GDP OF YEAR 1)-1)*100

```
In [7]: z = []
        col = x.columns
        gr = pd.DataFrame();
        for i in range(len(col)):
            for j in range(1,8):
                z.append(((df[col[i]][j]/df[col[i]][j-1])-1)*100)
            gr[col[i]] = z
            z = []
        GDP = gr['Gross Domestic Product (GDP) at Factor Cost']
        gr=gr.drop(['Gross Domestic Product (GDP) at Factor Cost'],axis = 1)
```

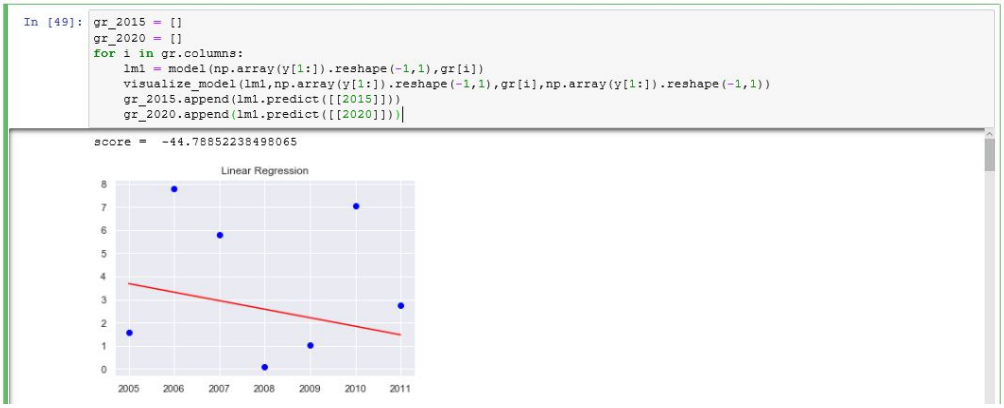
USING MODEL FUNCTION FINDING GROWTH RATE OF INDIVIDUAL FEATURE OVER THE YEAR 2015 AND 2020.

BUILDING THE MODEL AGAINST EACH FEATURE WITH YEARS(eg: agriculture vs year etc...)

gr_2015 = ARRAY WHICH CONTAINS PREDICTED GROWTH RATE OF EACH FEATURE WITH YEAR 2015.(eg : first element is growth rate of agriculture on 2015 etc...)

gr_2020 = SIMILAR TO ARRAY gr_2015 FOR THE YEAR 2020

CODING:



exp_gr = array which contains double the values of gr_2015. This array is what Dr. abdul kalam expected in 2020.

```
In [50]: exp_gr = []
         for i in gr_2015:
             exp_gr.append(2*i)
```

BUILDING THE MODEL FOR GROWTH RATE OF ALL THE FEATURES WITH GROWTH RATE OF GDP

BUILDING POLYNOMIAL REGRESSION WITH DEGREE 3.
CONVERTING ALL ARRAYS TO DATAFRAME:

DATAFRAME || ARRAY
q_2015 = gr_2015
r_2020 = gr_2020
exp = exp_gr

```
In [57]: lm,poly = model2(gr,GDP,3)

         col = gr.columns
         q_2015 = pd.DataFrame();
         r_2020 = pd.DataFrame();
         exp = pd.DataFrame();
         GDP_df = pd.DataFrame(GDP);
         for i in range(len(col)):
             q_2015[col[i]] = [gr_2015[i]]
             r_2020[col[i]] = [gr_2020[i]]
             exp[col[i]] = [exp_gr[i]]

         score =  0.9635169848746701
```

In [20]:
```
gr_2015
```

Out[20]:
```
[array([1.46487233]),
 array([0.12457763]),
 array([1.04356603]),
 array([4.01998476]),
 array([-0.64781741]),
 array([7.2520322]),
 array([9.48744006]),
 array([7.24562153]),
 array([6.87147683])]
```

In [23]:
```
q_2015
```

Out[23]:

	Agriculture, Forestry & Fishing	Mining & Quarrying	Manufacturing	Electricity, Gas & Water Supply	Construction	Trade, Hotels & Restaurant	Transport, Storage & Communication	Fina E
0	[1.4648723308575882]	[0.12457763140264433]	[1.0435660346115583]	[4.019984757725524]	[-0.6478174127041711]	[7.2520321950134985]	[9.487440062607561]	[7.2

In [22]:
```
gr_2020
```

Out[22]:
```
[array([-0.98113436]),
 array([-2.09308357]),
 array([-4.39015141]),
 array([2.03755499]),
 array([-6.76630737]),
 array([5.61389348]),
 array([8.01533974]),
 array([4.44205887]),
 array([7.14876124])]
```

In [24]:
```
r_2020
```

Out[24]:

	Agriculture, Forestry & Fishing	Mining & Quarrying	Manufacturing	Electricity, Gas & Water Supply	Construction	Trade, Hotels & Restaurant	Transport, Storage & Communication	Finar Est
0	[-0.9811343618727051]	[-2.0930835675177377]	[-4.3901514128187955]	[2.0375549856737507]	[-6.766307365302055]	[5.613893484687765]	[8.01533974005929]	[4.44

In [25]:
```
exp_gr
```

Out[25]:
```
[array([2.92974466]),
 array([0.24915526]),
 array([2.08713207]),
 array([8.03996952]),
 array([-1.29563483]),
 array([14.50406439]),
 array([18.97488013]),
 array([14.49124305]),
 array([13.74295365])]
```

In [26]:
```
exp
```

Out[26]:

	Agriculture, Forestry & Fishing	Mining & Quarrying	Manufacturing	Electricity, Gas & Water Supply	Construction	Trade, Hotels & Restaurant	Transport, Storage & Communication	F
0	[2.9297446617151763]	[0.24915526280528866]	[2.0871320692231166]	[8.039969515451048]	[-1.2956348254083423]	[14.504064390026997]	[18.974880125215122]	[14

THE PREDICTED GDP OF THE YEAR 2015 WITH q_2015(GROWTH RATE OF ALL FEATURES IN 2015) IS :

```
In [19]: gr_2015
         lm.predict(poly.fit_transform(q_2015))

Out[19]: array([6.01676143])
```

THE PREDICTED GDP OF THE YEAR 2015 WITH r_2020(GROWTH RATE OF ALL FEATURES IN 2020) IS :

```
In [18]: gr_2020
         lm.predict(poly.fit_transform(r_2020))

Out[18]: array([5.9359773])
```

THE PREDICTED GDP OF THE YEAR 2015 WITH exP(GROWTH RATE OF ALL FEATURES EXPECTED BY DR. ABDUL KALAM IN 2020) IS :

```
In [17]: exp_gr
         lm.predict(poly.fit_transform(exp))

Out[17]: array([6.54294282])
```

RESULT :

It is clear that the GDP of 2015 is greater than GDP OF 2020. So, we may say that GDP of India has been decreasing since 2015.

And it is also clear that expected GDP OF 2020 is not equal to the actual GDP of India in 2020 So, we conclude that India will not be a developed country in 2020.

VISION INDIA 2020 WILL BE A FAILURE.

REFERENCES:

EDUCATION:

Enrollments in schools have increased significantly but only a meagre amount of students reach for higher education.There are 993 universities over India.But National Knowledge Commission has estimated that it needs at least 1500 universities to educate its youth is out of range.

HEALTH:

“Health for all” is what Dr.Abdul Kalam aimed for 2020.”But with huge supply side problems,this seems to be far from our vision.With 17.5% of world population Indians hold up for 20% of world’s global burden of disease..An estimate of 60 millions Indians are pushed into poverty for their medical needs.There is also shortage of doctors and beds across India such as 1 bed for 614 people in Goa to 1 bed for 8789 people in Bihar.Some 6 out of 10 hospitals in poor states do not have ICU and one fourth have abysmal sanitation facilities.

FOOD PROCESSING AND PRODUCTION:

“Dr APJ Abdul Kalam dreamt to increase the food production by a factor of 2 by 2020.In 2012 there were 217 million malnourished people in India which is an absolute zero of the Vision 2020.India tops with the hunger chart around 200 million citizens sleeping hungry each night.In 2019,India stands 102 in the Global Hunger Index (GHI) among 117 qualifying

countries.Also 38% of children in 2016 below 5 years were stunted which is 31% in urban and 41% in rural areas. But with huge supply side problems,this seems to be far from our vision.

Dataset:https://data.gov.in/

Article:https://www.livemint.com/news/india/india-s-vision-2020-meets-the-real-world-11575308659670.html

Courses : https://www.coursera.org/learn/machine-learning